Notes on regressions

J. W. Mason

November 23, 2021

The most current version of this document can be found at `jwmason.org/teaching`.

A regression is a tool to find the best fitting linear, or straight-line, relationship between a dependent variable and one or more independent variables.

The dependent variable is the outcome we are trying to explain.

Let's say we have 10 observations of two variables. y is the dependent variable - the one we are trying to explain. x is the independent variable – the one that we think might explain it.[1]

[1] These numbers don't represent anything in particular, they were just generated to use as an example.

Observation	y	x
1	14	8
2	5	-2
3	12	14
4	11	9
5	23	12
6	6	3
7	15	13
8	8	11
9	7	0
10	19	15

Looking at the numbers, it seems like there might be a relationship between the variables - the high values of x mostly go with the high values of y, and the low values of x mostly go with the low values of y. To get a better sense of the relationship between the two variables, we can plot them in a graph. (This kind of graph is called a scatterplot.)

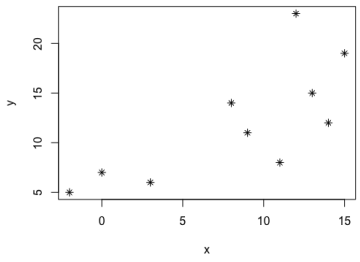


Figure 1: Scatterplot of x and y.

Visually we now have a strong impression of a relationship, the points seem to fall roughly along an upward-sloping straight line, though some are closer to it than others.Any statistical software or spreadsheet program will let you calculate a regression line, or trend line for the points. This is the straight line that gives the best fit to the points in the scatterplot. We can add that next.

Regressions are nothing more than a set of techniques for picking the right line to draw here, and evaluating how well it fits.

The idea of a regression is to come up with our best guess for the parameters β_0 and β in the equation

$$y = \beta_0 + \beta x + e$$

When we calculate the best β to fit our data, we call this doing a

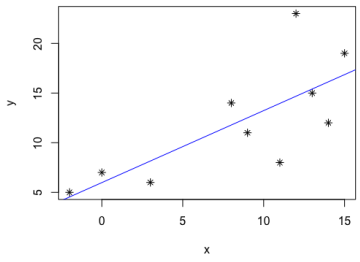


Figure 2: Scatterplot of x and y with regression line.

regression of y on x. Here e is a random error or disturbance term. If this equation were the true process generating y, then e would literally be a random variable. In practice, it represents all the influences on y other than x – all the things affecting x that our regression does not capture.

If we have more than one independent variable, we are looking for our best choice for β_1, β_2 and so on in the equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + e$$

Again y is the dependent variable and the xs are the independent variables. We also often call y the left-hand side variable and the xs the right-hand side variables.

The "best" values here simply means the ones that fit the the observed values of the independent variable most closely. With more than one independent variable, we can't directly visualize the relationship in a graph. But we can still think of it as fitting a line through a scatterplot – now the horizontal axis is just a weighted average of the various independent variables.

To decide how good a fit our line gives to the data, we divide the deviation or difference form the mean of each observation into the part explained by the line and the part that is left unexplained.

The variation in the dependent variable that we are hoping to explain is the deviation, or difference, of the observed values from the average value. If all the values were the same, there would be no variation, and nothing for a regression to explain.

While we could simply add up the differences of the values from the average, it's more common to use the squares of the differences. The average squared difference between the value and the mean is the variance; the square root of this number is the standard deviation.

In the case of our data above, the average value is 12 If you look at the differences between the observed values and 12 and square them, we get the following values:

$(14 - 12)^2 = 4$
$(5 - 12)^2 = 49$
$(12 - 12)^2 = 0$
$(11 - 12)^2 = 1$
$(23 - 12)^2 = 121$
$(6 - 12)^2 = 36$
$(15 - 12)^2 = 9$
$(8 - 12)^2 = 16$
$(7 - 12)^2 = 25$
$(19 - 12)^2 = 49$

These ten numbers add up to 310. Take the average and you get 31. So the *variance* of y is 31.[2] The square root of this is the *standard deviation* – in this case, 5.6. We think of this as the "typical" difference between an observed value and the mean.

Notice that the variance and standard deviation are most strongly affected by values that are far from the mean. In this case, the fifth observation, which is 11 away from the mean, accounts for nearly half the variance.

You might think that it would be simpler to call the typical deviation the average difference between an observation and the mean, without bothering with the squares and square roots. This value is called the mean absolute deviation, and it is occasionally reported. But in the vast majority of cases, people describe the variation, or spread, of the observed values in terms of the variance or standard deviation. Why we use this measure is used is a complex question - to some extent it's historical accident, to some extent it's for mathematical convenience, and to some extent it is a genuinely more meaningful value. For our purposes, it's enough to know that variance and standard deviation – the square root of the average squared deviation – is the usual measures of how much different observations of some value vary from each other.

The question then becomes, how much of this variation is explained by our regression. To answer this, we can divide the deviation of each observed value from the mean into two parts. There is the part that is explained by the regression – the deviation we would predict on the basis of the independent variable(s). And then there is the unexplained, or residual variation that is not accounted for by our regression.

This is shown in the next figure. One of our observations has an x value of 12 and a y value of 23. Our regression, reflected in the trend line, implies that when the independent variable is 12 the dependent variable should be around 15. So of the 11 deviation from the mean (11 = 23 -12) in the y value for this observation, 3 (15 - 12) is explained by the regression and the remaining 8 is the residual or unexplained part. (Note that the explained part can be negative if the regression predicts that the value should be less than the mean but it is actually greater, or vice versa.) The regression line is precisely the line that minimizes the squares of the residuals.

Picking the straight line that minimizes the sum of squared residuals is an *ordinary least squares (OLS)* regression. There are many other ways to estimate the relationship between variables, but OLS is the simplest and most widely used, and the starting point for most other approaches, so we will focus on that here.

[2] In many statistical calculations, we instead use the sum of the squared deviations divided by the number of observations minus 1. In this case, that would be 310/9 = 34.4. With 10 observations, the difference is noticeable. With larger samples, as is usual, this value will be essentially the variance.

The variance of x is given by

$$var(x) = \frac{\Sigma(x - \bar{x})^2}{n}$$

where $\bar{x}$ is the average value of x and n is the number of observations. The standard deviation of x is just the square root of the variance:

$$st.\ dev.(x) = \sqrt{var(x)}$$

.

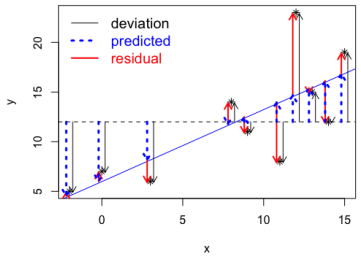


Figure 3: Predicted and residual variation.

There are three questions we ask about a regression line: What is the line itself, how much of the variation in the dependent variable does it explain, and how much better is it than other possible lines?

Once we have drawn our regression line, we can ask three questions about it. First, what is the line itself – where is it, and how steep is it? Second, how good a job does it do describing the dependent variable? How much of the variation in that variable is explained by the regression, and how much is left unexplained? Third, how much better is this line than other possible lines? Are we confident that this line is better than one that is steeper or flatter and, in particular, that it is better than simply drawing a horizontal line at the average value? The statistics reported in a regression table are answers to these questions.

All these questions have answers based on the sample data we have. We often want to interpret them in terms of the underlying population from which the sample is supposed to be drawn. The numbers reported in our regression just describe the data we actually see. But we usually think of them as an estimate, or guess, about a relationship existing out in the world. We imagine that we could draw another sample, or set of observations from the same population or economic process that the first one came from, and see more or less similar results.

The answer to the first question is the line we drew. We can also think of it as an equation. In our example here, the line is approximately:

$$y = 6 + 0.7x$$

You can easily see that by looking at the line – where x is 0, the regression line is around 6; where x is 5, the line is around 10, and so on. The equation and the line are two equivalent ways of describing the same relationship. If we have more than one right-hand variable, we can't represent as a line on a flat paper. But we could imagine an equation with two right-hand variables as a plane through a three-dimensional space, and so on. Or we could draw a line with the horizontal axis as a weighted average of all the right-hand side variables.

In terms of the second question, we want to know how much variation is left after we account for the variation described by the line. In other words, if we took the sum of squares of the deviations from the regression line (the red arrows in the figure), how much smaller would that be than the sum of squares of the deviations from the average value (the black arrows in the figure)? Another way of thinking of this is, If we were to guess the values of the dependent

variable based on the regression, how much more accurate would our guess be than if we just guessed the average.

For example, if you had to guess the height of a child picked at random from a room full of children, the best guess you could make would be the average height of all the children. But if you also knew the ages of the children, you could make a more accurate guess, since children consistently get taller as they get older. On the other hand, if it was a roomful of adults, knowing their ages would not help you guess their heights. So a regression of height on age will explain a lot of the variation in heigh for a room full of children, but very little of it for a room full of adults. In this case, your best guess for the height of any individual would simply be the average height of the people in the room.

Remember, we are judging best fit by the sum of squared residuals. If we just wanted to minimize our average error, the best guess will be the median. (If there are an even number of observations, any value between the middle two will work as well.) But if we want to minimize the average *squared* error we will do better with the mean, since that reduces the chance of a really big miss on either side.

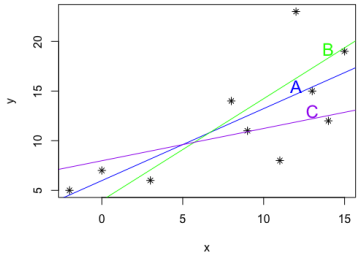


Figure 4: Regression line (A) versus other possible relationships between x and y.

For the third question, we want to know how precise our estimate of the best line is. Would a line that was steeper or flatter, or higher or lower, fit the data just about as well? Or will any good-fitting line have to be very close to the one we drew? Lines B or C in Figure 4 also seem to give a reasonably close fit for our ten observations. But are they really almost as good as line A? In other words, are the squared residuals much higher than for the regression line, or are they about the same? How much would we have to change our line before we could confidently say that it did *not* describe the relationship between the variables?

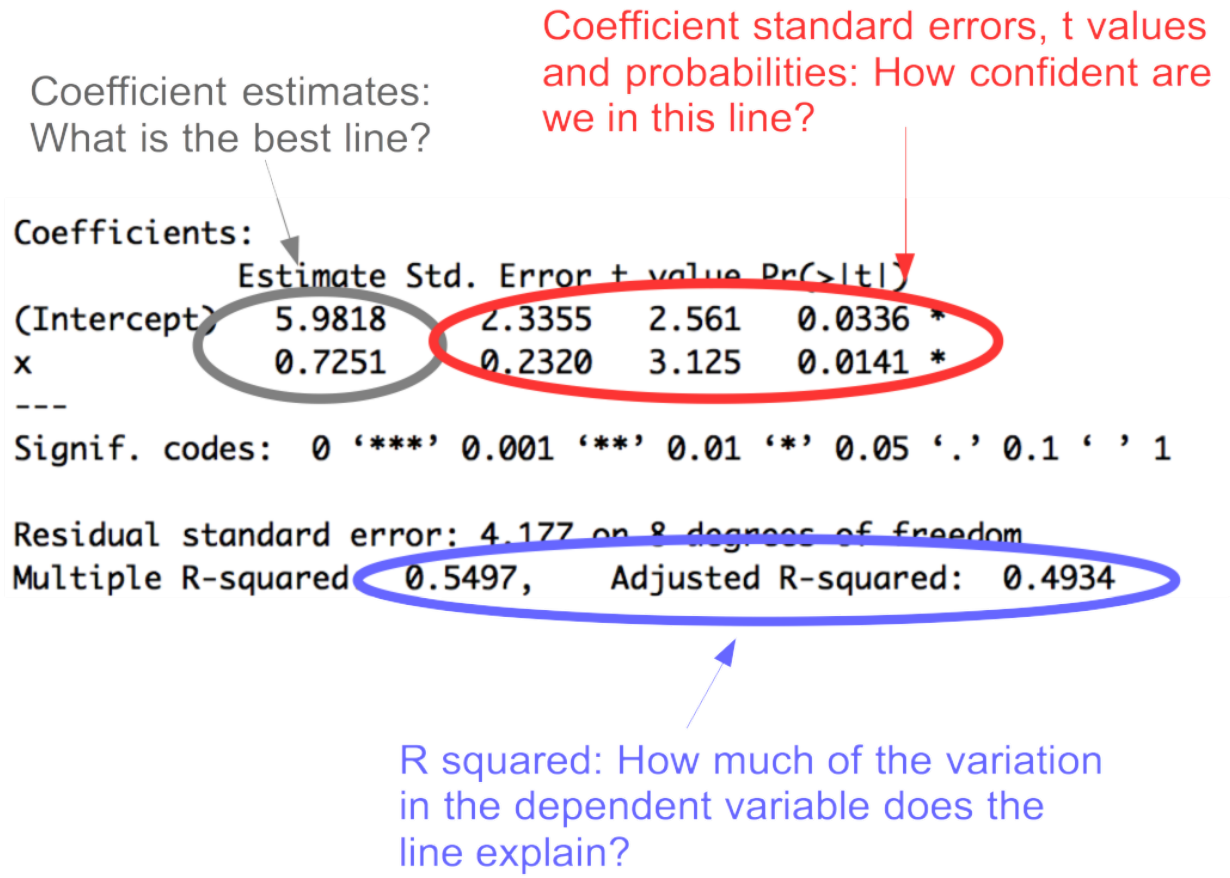
All three of these questions – the shape of the line, how much of the variation it describes, and our confidence in it – are important for deciding whether the regression is telling us something useful. If we are not at all confident about the results – if we think there's a good chance the true relationship is very different from our estimates – then we probably don't want to rely on them. But even if we are confident in the line, it may still not be useful if the effect is very small, or if it explains only a small part of the variation in the dependent variable. To take a familiar example, if you hear about some new health research showing that eating less of some food is good for your health, you will want to know how confident scientists are in the result – how much data is based on, how consistent are the results across different studies, and so on. But you will also want to know how big the effect is, how much of a change it will produce compared with other things that affect health. It's quite possible for

an effect to be very precisely estimated, but also too small to worry about.

A typical regression table shows the parameter estimates (including intercept), the standard errors and/or t-values for the the parameters, some indication of whether the estimate is "significant", and an r-squared and/or adjusted r-squared. These numbers are used to answer the three questions above.

If you run a regression using some kind of statistical software, the numbers you get are the answers to the three questions above. You may get other information as well, but the most prominent (and familiar) regression outputs are intended to answer those questions.

Let's try this with our example values. If we run a regression of y on x in R, we will get the following output. (R also gives us a few other values, which we can ignore for now.) Other software will give the same values, in a similar table.



The coefficients or parameter estimates tell us what the best-fitting line is. Our confidence in this line is shown by the standard errors, t-values and probabilities. The r-squared describes how much of the variation in the dependent variable is explained by the line.

The coefficients or parameter estimates, plus the intercept or constant term, describe the line itself.

The coefficients or parameter estimates tell us what the line itself is. The intercept or constant tells us where the line crosses the y axis – what the expected value of the dependent variable would be if all the independent variables were zero. This is usually not interesting, although sometimes it is. The parameter estimate tells us the slope of the line – how much we expect the dependent variable to increase if the independent variable increases by one unit.

In this case, the coefficient estimates are 5.98 for the intercept and 0.725 for x. In other words, this is telling us that the best line to draw through the points – in terms of minimizing the squared errors – is

$$y = 5.98 + 0.725x$$

So for example if x is 10, we would expect y to be around 13.25. The intercept is the expected value of y when $x = 0$. In terms of the line, it's the height of the line at the point where it crosses the vertical axis. The coefficient is the slope of the line – the expected increase in y when x increases by 1.

To interpret the coefficients, think of them as describing the relationship between the variables. To do this, we need to pay attention to the units of the variables.

The interpretation of this coefficient is, if x is one higher in our second observation than in our first, we would expect y to be 0.7 higher. But: 1 what? 0.7 what? In this case, we are simply working with an abstract example, so the numbers aren't one or 0.7 anything. But in any real world regression, the variables will have units, and those units will determine what the results actually mean. The coefficient is always in units of the dependent variable per unit of the independent variable. Imagine you see a regression of annual earnings on years of education, and the estimated coefficient is 5. The units of the right hand side variable are obviously years, but what about the left-hand side? It makes a big difference if income is five dollars higher per year of schooling, or five thousand dollars higher per year of schooling, or five percent higher per year of schooling. Sometimes a coefficient will seem to have a very high or very low value simply

because of the units being used. The regression of income on years of schooling would have a much higher coefficient (1,000 times higher, to be exact) if the left-hand side variable were in dollars rather than thousands of dollars. But the substantive relationship would be exactly the same.

When you read regression results, you should always think of the parameter estimates in terms of "we expect so much more of *this* for each additional *that*." The numbers don't mean anything unless you. know exactly what this and that are.

Sometimes a right-hand side variable is an indicator, or dummy, variable, which takes on a value of either 0 or 1 In this case, the coefficient shows how much higher we expect the dependent variable to be when the indicator is one – when whatever condition it represents is true. For instance, if our regression for earnings also had a variable on the right side for sex, then coefficient would tell us how much higher earnings were for one sex than the other. In this case, we need to be sure we know which value gets a 1 and which value gets a zero. Sometimes – say if the variable were college graduate – this is obvious, but in other cases, like sex, the choice is arbitrary.

Sometimes it is the dependent variable that is categorical - something that is either true or false, rather than having a continuous numerical value. In that case, the regression is run slightly differently. (A *logistic* or *probit* regression is normally used) Assuming the correct regression was run, we can interpret the coefficient in this case as the percent increase in the probability of the dependent variable being true that results from a one-unit increase in the value of the independent variable.

To interpret the coefficient estimates, it's also helpful to know something about the variation in each of the variables. In general, we can think of a one standard deviation difference as being a typical or normal difference between one observation and another. If a one standard deviation change in an independent variable produces a one standard deviation change in the dependent variable, then the independent variable explains all of the variation in the dependent variable. We will almost never see a value of 1 in real data, but the closer we get, the more important this independent variable is in explaining variation in the dependent variable. Sometimes regression results include *standardized coefficients*. These are the estimated coefficient divided by the standard deviation of the dependent variable, multiplied by the standard deviation of the independent variable.In other words, the standardized coefficient is the number of standard deviations of change in the left-hand side variable we expect from a one standard deviation change in the right-hand side variable. With a single variable on the right-hand side, the highest possible value

The standardized coefficient of y on x is equal to the coefficient times $\frac{st.\ dev(y)}{st.\ dev(x)}$.

for a standardized coefficient is 1. This gives a useful benchmark for whether an effect is "big". If the standardized coefficient has a value of 0.1 or 0.2 or so, we can say that that independent variable is an important source of variation in the dependent variable. If the standardized coefficient is very small, like 0.01, then we can say that while there may be a relationship between the variables, this independent variable is not a major or important explanation for the variation in the dependent variable. It would take an extremely large change in the independent variable to produce a substantial change in the dependent variable.

Standardized coefficients are only occasionally reported in published regression results. But there often is a table of descriptive statistics that gives the standard deviations of the variables. If you are wondering whether an effect is large enough to matter, it can be useful to use these to calculate the standardized coefficient yourself. If there is only one variable on the right side, there's no need to do this – you already get the same information from r squared. (With just one independent variable, r-squared will be just equal to the square of the standard coefficient.)

The r-squared describes how much of the variation in the dependent variable is explained by the line.

The share of the variation in the dependent variable that is explained by the regression is shown by r-squared or adjusted r-squared. The simplest way of thinking about this is the fraction of variation accounted for by the regression – the proportion of the total variance that is explained rather than residual.

To see how this is calculated, let's go back to our original x and y values, and combine them with our equation to come up with predictions for y based on x.

Observation	x	Actual y	Predicted y: $6 + 0.7x$	Residual: y - predicted y
1	8	14	11.8	2.2
2	-2	5	4.5	0.5
3	14	12	16.1	-4.1
4	9	11	12.5	-1.5
5	12	23	14.7	8.3
6	3	6	8.2	-2.2
7	13	15	15.4	-0.4
8	11	8	14.0	-6.0
9	0	7	6.0	1.0
10	15	19	16.9	2.1

The first two columns of the table just show the observed values of x and y The third column shows us our prediction of y given the observed x. This is the same as the regression line we drew earlier. The final column shows the residuals - the red lines from Figure 3. This is the error in our prediction – the difference between the actual value of y and what we would predict based on our regression. You can check the values in the last two columns yourself.

To get r-squared from this table, you first take the squares of all the residuals: $2.2^2 = 4.84$, $0.5^2 = 0.25$, $-4.1^2 = 16.81$, and so on. Add up all these squares and you get 139.45. This is the sum of squared residuals, or *residual sum of squares* or *error sum of squares*. Then divide this by the sum of squared deviations of y that that we calculated earlier.[3] In this case, that gives $139.45/310 = 0.45$. This is the fraction of the total variance that the regression has *not* explained. The remainder, or $1 - 0.45 = 0.55$, is the fraction it *has* explained. If you look up at the regression results, you will see that the first r-squared results, "multiple r-squared", is indeed almost exactly 0.55. (It is not quite the same because of rounding.)

[3] We could equally well divide the mean squared error by the variance.

A second way to calculate r-squared is to subtract the mean of y from the predicted value, and square the result. Since the mean of y is 12, this gives us $(11.8 - 12)^2 = .04$, $(4.5 - 12)^2 = 56.25$, $(16.1 - 12)^2 = 16.81$, and so on. Add up all these squares and you get 170.65. This is the *regression sum of squares*. If we divide this by 310, we get 0.55 – again, the same number reported in the regression as r squared.

Note: While the first way of calculating r-squared works for any regression line, or indeed for any set of forecasted values produced by whatever method, the second way works only for an OLS regression. Only in the case of an OLS regression will the regression sum of squares and the error sum of squares add up to exactly the sum of squared deviations of the dependent variable.

Regression tables often, as here, also report an *adjusted r-squared*. This is computed using a slightly more complex formula that compensates for the fact that adding additional variables to a regression will always improve the fit. (In the extreme case, where the number of right-hand side variables is just one less than the number of observations, the fit will always be perfect - r - squared will always be 1.) The adjustment reduces r-square slightly for each additional variable that is included on the right-hand side. If, as here, there are one tenth as many independent variables as observations, the adjustment increases the unexplained part of the variation (1 - r-squared) by one tenth. (You can verify this is the case in the regression results here.) It will only be noticeably different from the default r-squared when the number of observations is low, as here, or when there is an

The formula for r-squared is

$$r^2 = 1 - \frac{ESS}{TSS}$$

where ESS is the error sum of squares, or sum of squared residuals, and TSS is the total sum of squares, or the sum of the squared difference between the dependent variable and its average value.

An equivalent way of expressing this is

$$r^2 = 1 - \frac{SE^2}{var(x)}$$

where SE is the standard error of the regression and var(x) is the variance of the dependent variable.

The formula for adjusted r-squared is:

$$adj.\ r^2 = r^2 - (1 - r^2)(\frac{p}{n - p - 1})$$

where p is the number of independent variables and n is the number of observations.

exceptionally large number of independent variables. If the number of independent variables is much smaller than the number of observations, as it usually is, then r-squared and adjusted r-squared will be almost identical.

Note that while there are separate parameter estimates, standard errors, and t- and p-values for each independent variable, there is only a single r-squared (and a single adjusted r-squared) for the whole regression.

How closely our regression line fits the data is an important question, and r-squared is an important statistic for answering it. But there are a number of reasons to be cautious in how we use it.

1. While a higher r-squared is often taken to be a "good" result, we shouldn't assume this - and we certainly shouldn't make changes to our regression just to get a higher r-squared. Finding that some variables are not related to each other is not intrinsically any less interesting or important that finding that they are.

2. A high r-squared shows that our regression equation does a good job capturing the variation in the dependent variable. But an extremely high r-squared is also usually a sign we've done something wrong. Complex social phenomena of the sort economics studies almost always are subject to multiple influences, not all of which can be readily identified or measured reliably (or at all). So we would be very surprised if we were able to predict 90 or 95 or 99 percent of the variation in an interesting economic variable on the basis of other genuinely independent variables. When we see an r-squared like that, it is often because the right-hand side of the regression contains something that is really just an alternative measure of the dependent variable. For example, if you regressed the growth in GDP on employment growth and labor productivity growth, you would get an r-squared close to one. But that wouldn't mean anything, because by definition output is equal to employment times labor productivity.

3. Related to this is a third concern - r-squared can be very misleading if the variables are *non-stationary*. This means that the mean values change over time - the expected value of a variable observed at a later date is higher or lower than the expected value of the same variable observed at an earlier date. This is true of many macroeconomic variables. If you regress a non-stationary variable on one or more other non-stationary variables, then the r-squared of the regression will always get close to one if your sample covers a long enough period. This apparent close fit doesn't imply any genuine relationship between the variables. It just reflects that time

is passing for all of them, so things that grow or shrink over time are all growing or shrinking together. The usual solution to this is to look at the changes, or *first differences*, in the variables rather than their levels, or to otherwise modify the variables to make them stationary. If this isn't done, the r-squared of a regression involving non-stationary variables will be meaningless.

4. The r-squared statistic applies to the regression as a whole. If the right-hand side variables are all of genuine interest, this isn't a problem. But in many cases, the regression includes a number of *control* variables that are of no interest themselves, but are included only to try to eliminate extraneous influences on the dependent variable. (Often these controls and their coefficients are not even listed individually in the regression results.) The r-squared of the regression will reflect the contributions of these control variables as well as the variables the research is actually focused on. For example, a regression of individual income might include a dummy variable for each state, to eliminate, or control for, geographic variation in income that is not relevant for the effect being studied. There is nothing wrong with this in principle, but it's important to realize that the r-squared for the regression will include the contribution of these state dummies as well as of whatever variables the paper is actually about. Unless variation in income across states is what we are studying, the fact that knowing the state someone lives in can help predict their income is not very interesting.

In general r-squared is a useful measure when our regression includes a small number of independent variables that are all of economic interest and that we are confident are stationary and are not mechanically linked to the dependent variable. When these conditions aren't met, a large r-squared is not informative. A very low r-squared is still informative in this case – it tells us that the variables in our regression explain very little of the variation in the dependent variable

The standard errors of the estimates tell us how much better these parameters fit the data than other parameters would.

Our confidence in this line is shown by the standard errors, t-values and probabilities. In published regression results, either the standard error or the t-value is often reported in parentheses below the parameter estimate. All three of these statistics are measures of how much uncertainty there is in our parameter estimates.

The standard error is a measure of how much the coefficient es-

timate could change before the fit got much worse. In general, an estimate one standard error from the one we got would also be reasonable given the data, an estimate two standard errors away would be unlikely but not crazy, and an estimate much more than two standard errors away is clearly a bad fit for this data. One way to think of this is to imagine drawing another sample from the same population and run another regression on it, we would not be at all surprised if the new parameter estimate was one standard error from the first one, and only moderately surprised if it was two standard errors away. But if it there was a four or five standard error difference between the parameter estimates, we would strongly suspect we were not actually looking at the same population or had done something else wrong.

Suppose there really is a true value of the coefficient in the population or relationship we are interested in. If we took a random sample from the population, we wouldn't expect to see exactly the true relationship - there is always some randomness in the data, whether from other variables we are not observing or pure random noise. But we would expect that most of the time, our estimated relationship would not be too far from the true relationship. Specifically, we would expect

- half the parameter estimates to be within 0.67 standard errors of the true value.
- 80 percent of the estimates to be within 1.3 standard errors of the true value.
- 90 percent of the estimates to be within 1.64 standard errors of the true value.
- 95 percent to be within 1.96 standard errors of the true value.
- 99 percent to be within 2.6 standard errors of the true value.
- 99.9 percent to be within 3.3 standard errors of the true standard value.

And so on. These numbers are called the critical values of the t distribution.

Again, what this means is that you would not be at all surprised if the true coefficient was one standard error more or less than what you estimated, but quite surprised if it was three standard errors more or less, and amazed if it was four standard errors more or less – unless, again, you did something wrong, or there is not really a stable underlying population or the assumptions of the regression were violated in some other way.

An easy way to think about this is to imagine you go back and get another sample from wherever you got this one from. Even if your new data was from the exact same population, or form the exact same economic process, you wouldn't expect your regression results to be exactly the same. There are always going to be other variables you haven't accounted for, as well as measurement error and pure random noise. On the other hand, if the data was really coming from the same place, you wouldn't expect it to be too different either. The relationship between x and y in this sample is strong enough that you would expect it to reflect something about the underlying process produces the data, that will be present in another sample too. So on balance, if you drew another sample, you'd expect the new coefficients you estimate to be somewhat different from these, but not too different. The standard errors of the coefficients (and the t-statistics and probabilities derived from them) are telling how different you should expect your the results from your next sample to be from the results from your last one.

In this case, the estimated coefficient of y on x is 0.73, and the standard error of the coefficient is 0.23. This means that while our best guess is that an increase of 1 in y will be associated with an increase of about 0.7 in x, but we would not be surprised if with more data we found a coefficient closer to 0.5 or 1. On the other hand, based on this data, we would feel reasonably confident that the true parameter is probably not more than two standard errors from our estimate, or in other words somewhere between 0.25 and 1.3. Similarly, we estimated an intercept of close to 6, with a standard error of 2.3. So we are reasonably sure that if we drew more data from the same source, we would find an intercept is somewhere between 3.5 and 8, roughly. (With this range of uncertainties, it would be entirely reasonable to write the equation as $y = 6 + 0.7x$. Given the standard errors, it's silly to pretend we have more than one significant figure.)

We can see this visually in Figure 5. This shows the regression line and lines one and two standard errors above and below it. The inner dashed lines are one standard error away from the estimate – we would not be at all surprised if we drew another sample from the same population and found that the best line was anywhere in this range. The outer dotted lines are two standard errors away. We would only be moderately surprised if a line through our next sample was somewhere in this range. Beyond the outer dotted line, we would be more surprised. If the relationship in the larger population lies out of that range, then we should see a relationship as far from it as our estimate in less than 5 percent of the samples we draw from it.

To put it another way: Any line that falls entirely within the one-standard-error bands fits our data almost as well as the regression

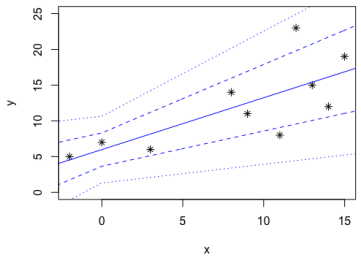


Figure 5: Regression line with one- and two-standard error bands.

line. Any line inside the two-standard-error bands fits only moderately less well. But a line that goes outside the two-standard-error band fits our data much less well than our regression did - the difference is big enough in this case that we can say the regression is reasonable evidence against that line describing the population our sample comes from.[4]

[4] Because the two-standard error band excludes a horizontal line, we say that our estimate is significant at the 5 percent level. See below.

The t-value is just the parameter estimate divided by the standard error. In other words, it says how many standard errors the estimate is from zero. In this case, with a parameter estimate of 0.73 and a standard error of 0.23, the t-value will be around 3.2. Similarly, the t-value of the intercept will be $6/2.3 \approx 2.6$.

The t-value doesn't give you any new information if you already have the coefficient estimate and its standard error. But it may be a more convenient way of representing it. One advantage of the t value is that, unlike the standard error, the it doesn't have any units. A standard error of 0.5 may be very large or very small, depending on the units of the variables. But t statistics of 0.1 or 0.5 or 1 imply the same relative degrees of precision in any regression, regardless of the variables that are being observed.

(One source of confusion in reading regression results is that when you see a number in parentheses under the coefficient estimate, it might be either the standard error of the coefficient or the t-statistic. In many cases the table will say, but in other cases, you will need to use context to figure out which it is.)

The standard errors, and the t-statistics and probabilities derived from them, assume that the underlying data is normally distributed. If it is not, the values in the regression table may be misleading – in general, they will suggest more confidence in the results than is actually warranted.

The t values and probability measure tell us how likely we would be to get these estimates if the true value of the parameter were zero.

Another way of thinking about the precision of the estimate is to ask how likely we would have been to find a coefficient this large, if the value in the larger population were zero - if there were no true relationship between the variables at all. Looking up at the critical values listed above, we can see that in 95 percent of samples, the parameter estimate should be within 1.96 standard errors of the true parameter. That means that if the t-value of our estimate is less than 1.96, there is a 5 percent or greater chance that we could have found an effect this large in our sample even if there is no relationship between the variables in the larger population.

The probability value is derived from t value and the critical val-

ues - it indicates how often we would get a parameter estimate this large by chance if the true parameter was zero. For example, in this regression the probability value is listed as 0.0141. This means that if we had a larger population with the same degree of random noise we see in the sample (the same standard deviations of x and y) and we repeatedly took 10 observations at random from it, we would get an apparent relationship at least as strong as the one we see here in about 1.4 percent of those samples.

By convention, a value of less than 5 percent here is often referred to as "significant", though there is nothing special about the 5 percent threshold. Significance levels are often indicated by stars next to the parameter estimate, with more stars indicating a higher level of significance – that is, a lower probability that we would see a parameter estimate this large by chance if the true parameter were zero. The default in R, as in many programs, is to print one star for a probability below 5 percent, two stars for a probability below 1 percent, and three stars for a probability below 0.1 percent. Since 0.141 is lower than 0.05 but not lower than 0.01, there is one star here.

The most common measure of "statistical significance" is a p value of less than 0.05, or a t-value greater than two. Statistical significance is often misunderstood.

Most regression tables show the significance of each parameter estimate, often by placing one or more stars next to it. In common use, an estimate is "statistically significant" if the t-statistic is greater than 1.96 – that is, if the estimate is more than about two standard errors from zero.

Statistical significance in this sense is often misunderstood, and there is a good deal of controversy about how useful or meaningful it is. Nonetheless, it is almost always reported in regression results.

Because there is so much confusion about statistical significance, we need to be clear first on what it does *not* mean.

- Significance at the 5 percent level does *not* mean that there is a 95 percent chance that our estimate is correct. In fact, no estimate is ever exactly correct. In general, there is no "true" parameter out there in the world to find, so to say an estimate is correct is meaningless. But even if we more realistically imagine a very large population with a certain relationship between the two variables x and y, we would not expect any finite sample drawn from that population to show exactly the same relationship as the population as a whole. In that sense, the chance that our estimate is exactly correct is always 0.

- Significance at the 5 percent level does not mean that you would find a zero relationship in fewer than 5 percent of samples. You will *never* find exactly zero relationship between two variables with any random component, for the same reason. By chance alone, they will always have a slight positive or slight negative correlation.
- Lack of significance at the 5 percent level does *not* mean that the true value is probably near zero. The t-statistics are just measuring the uncertainty of our estimate. If you get a t-statistic of 1.96, that implies you would see a relationship as strong as this in 5 percent of samples from a population where the true relationship was zero. But it equally well means you would see a relationship as weak as this in 5 percent of samples from a population where the true relationship was twice as strong as your estimate.
- Lack of significance at the 5 percent level does *not* mean that the estimate is totally uninformative or worthless. Lack of significance means that regression can't be used as evidence against a zero relationship between the variables. But it still can be used as evidence against a very strong positive or negative relationship - one with a coefficient more than two standard errors from the one we estimated.
- Significance at the 5 percent level does not *quite* mean that if the true relationship were zero, you would see an effect this strong in fewer than 5 percent of samples. Significance at the 5 percent level means that if the true relationship were zero, *and* if the random errors follow a normal distribution, you would see an effect this strong in fewer than 5 percent of samples. But the true distribution of errors, like the true coefficient, is something we can never know, and in most cases doesn't exist even in principle.

Here is what statistical significance does mean: Significance at the 5 percent level means that if there were an underlying population with variables normally distributed and perfectly uncorrelated (i.e. a true coefficient of exactly zero), then we would see a relationship this strong in fewer than 5 percent of samples this size drawn from that population.

This is an imaginary, hypothetical case – we can never know the true population, in many cases it does not make sense to think of a true population or data-generating process even in principle, and to the extent one does exist there is no reason to think the distribution of variables is perfectly normal. People often interpret significance at the 5 percent or 1 percent, etc. level as meaning that there is a 5 percent, or 1 percent etc. chance that something is or is not true in

the real world. But it cannot be interpreted that way. It describes the chance of seeing this data in an imaginary hypothetical case.

A better way of thinking about significance is simply as a measure of how well this line fits the particular data we are looking at. The standard error of the regression asks how much we can change our line before we get a noticeably worse fit. Statistical significance asks the same question a different way. It asks: How much better is our fit than if we didn't use x at all? In the case of a bivariate regression, not using x just means drawing a horizontal line at the mean of y. For a regression with more than one variable on the right-hand side, it means doing a regression with all our other independent variables except for x.

What significance is telling us is precisely how much better our regression line fits the observed data than the one without x does. Notice that this is a question strictly about the observations in our sample – it does not involve any hypothetical population from which they are drawn.

Where does the 5 percent come from, then? That is simply a way of deciding how much better the fit is, and in particular, whether it is sufficiently better to justify including x in the regression. The usual definition of significance – how likely you would be to find a relationship this strong in a population with perfectly uncorrelated normally distributed variables – is just a convenient benchmark for deciding how much better our regression line fits than one that was drawn without using x.

In other words, if your estimate for the coefficient on a variable is statistically significant, that it means that a regression using the variable fits the observed data better than one without the variable. This fact is worth knowing. But without more information, you can't use it to draw any conclusions about the probability of any claims about the world.

The logic of significance tests is that a coefficient two or fewer standard errors from the one we have would fit the data about as well, a coefficient two to three standard errors away would fit substantially worse, and a coefficient more than three standard errors away would fit very much worse. The language of how likely you would be to get this line by chance, is just a way of quantifying how much worse the fit is.

As is often the case in statistics, negative claims are more straightforward than positive ones. What you can say is that if the estimate is not significant at conventional levels, you shouldn't be confident that the true relationship, if any, has the same sign as your estimated one. A positive coefficient with a probability of 20 or 30 or 40 percent (or in other words a t-statistic of 1 or 1.5) is not giving you any use-

ful evidence that there actually is a positive relationship between the variables. If your regression line slopes upward but is not significant at the 5 percent level, what that means is that it doesn't fit the data much better than a flat line (or one that sloped slightly downward) would.

The fact that a non-significant coefficient isn't giving you useful information on whether the relationship is positive or negative, does not mean that it is giving you no information at all. Remember, all significance is telling you is whether your fit is better than you would get from a regression that did not use the variable at all. But not using the variable may not be the relevant counterfactual. A value that is not significantly different from zero, is still significantly different from very large positive or negative values.

Suppose our coefficient estimate were positive 3 with a standard error of 2. This would not be significantly different from zero. So we could not use this regression as evidence that there was a positive relationship between the variables - our results are consistent with no relationship or a weak negative one. But our results are *not* consistent with an extremely strong relationship. They do give us evidence that if we could draw another sample from the same population, we would not find a coefficient greater than 10 or less than -5. Since these values are three standard errors away from our estimate, the regression does constitute meaningful evidence against them. (This assumes, again, that we have an underlying population with more or less normally distributed variables). If the regression is being used to investigate a hypothesis of a specific relationship between the variables – i.e. if our null hypothesis is something other than a coefficient of zero – then estimates that are not statistically significant in the conventional sense may still be very informative

The use of 5 percent as the cutoff for significance is a historical accident. There is no reason why it is more intrinsically appropriate than 1 percent or some other threshold. It continues to be used today partly out of inertia, and partly out of a kind of collective judgement about how much precision it is reasonable to demand from economic relationships.[5]

For all these reasons, it is probably best not to focus too much on significance levels. The standard error of the coefficient or t-statistic give the same information about the precision of the estimate, in a more meaningful way.

[5] In physics, a t-statistic of 3 (significant at the 0.3 percent level) is often required before a result can be considered evidence, and a t-statistic of 5 (significant at the 0.00003 percent level) is required before it can be considered a discovery.

A large coefficient means a strong effect. A small standard error or large t-statistic means a precisely estimated effect. A large r-squared means the regression as a whole does a good job describing the dependent variable.

These three rules cover most of what you need to know in reading a regression table.

A regression with multiple independent variables is similar to one with just one variable. We can think of it as fitting the best line on a graph with the dependent variable on the vertical axis and a weighted average of the independent variables on the horizontal axis.

A regression with multiple right-hand side variables is conceptually the same as one with just one right-hand side variable. We are still trying to find the values of β_1, β_2, etc. in the equation

We refer to an equation with more than one independent variable as a *multivariate* regression, while one with just one independent variable is a *bivariate* regression.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + e$$

that minimize the sum of squared residuals. The interpretation of the coefficients is the same: Suppose we estimated the equation

$$y = 3 + 2x_1 - 0.5x_2$$

That would mean that for every one-unit increase in x_1, we would expect y to be 2 higher, for every one-unit increase in x_2 we would expect y to be 0.5 lower, and when x and y are both 0 we would expect y to equal 3.

The standard errors of the coefficients are also the same: An equation with a coefficient one standard error above or below the one we estimated would fit the data almost as well, but a coefficient more than two standard errors above or below would fit the data definitely worse. Or to put it another way, if we could draw another sample from the same population and run the regression again, we would not be at all surprised if the new coefficients were one standard error away from our first estimate, we would be moderately surprised if they were two standard errors away, and we would feel that something was almost certainly wrong if they were more than three standard errors away.

Note that no matter how many variables we have, we still only have one intercept. This is the predicted value of y when all of the independent variables are zero. In most cases with many variables on the right-hand side this is economically meaningless and can safely be ignored.

R-squared also has the same meaning with multiple independent variables. It is the fraction of the variance in the dependent variable that is explained by the regression. The only thing to note is

that r-squared is calculated for the regression as a whole – it doesn't tell you anything about which particular right-hand side variables explain the variance. This is especially a concern if not all of the right-hand side variables are economically interesting.

Standard errors of the coefficients, t-statistics and p-values are calculated for each independent variable and have the same interpretation as in the bivariate case. The only difference is that the significance test is based on the null of the same regression without that one variable, as opposed to the null of y equal to its mean value as in the bivariate case. So for x_n to be insignificant, in a multivariate equation, means that the regression fits about as well without x_n as with it, and so gives you no reason to think that x_n is related to y.

If the independent variables are more or less uncorrelated with each other, then a multivariate regression will give more or less the same coefficients (except for the intercept) as you would get from regressing the dependent variable on each of the independent variables separately. If any of the independent variables are strongly correlated with each other, however, then the estimated coefficients may be quite different from what you would get regressing the dependent variable on them separately, and the confidence intervals (or standard errors of the coefficients) will always be larger. Adding a variable will always raise r-squared, but it may or may not raise adjusted r-squared.

A problem for regressions with more than one independent variable is that the independent variables may be correlated with each other. This makes it hard to estimate their separate effects.

Suppose we have a sample of ten observations as follows. Again, y is the dependent variable; x_1 and x_2 are two independent variables. We think that x_1 and x_2 might help explain, or predict, the values of y.

Observation	y	x_1	x_2
1	3	2	0
2	7	2	8
3	9	9	6
4	28	11	12
5	14	4	-1
6	15	9	3
7	-5	0	-9
8	4	1	4
9	9	9	1
10	8	6	1

It certainly seems that there is some relationship between the variables – observation 4, with the highest value of y, also has the

highest values of x_1 and x_2, observation 7, with the lowest value of y, has the lowest values of x_1 and x_2, and so on. If we run regressions of y on each of the two independent variables, our sense that there is a strong relationship will be confirmed. The results of regressing y on x_1 are shown below.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1742     3.1297   0.056  0.95698
x1            1.7030     0.4801   3.547  0.00754 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.763 on 8 degrees of freedom
Multiple R-squared:  0.6113,    Adjusted R-squared:  0.5628
```

As you can see, we estimate a coefficient of x_1 of 1.7, meaning that we think an increase of 1 in x_1 will be associated with an increase of 1.7 in y. Based on the standard error, if we drew another sample from the same population we would expect to find a coefficient between 1.2 and 2.2, and would be quite surprised if we found a coefficient less than 0.7 or more than 2.7. Because the lower bound of this confidence interval is greater than zero (i.e. the t-statistic is greater than 2), we say the estimate is statistically significant. In fact, with a t-statistic of 3.5 it is significant at the 1 percent level – in a population with normally distributed variables, we would expect a sample of ten observations to show a relationship this strong purely by chance less than 1 percent of the time. Finally, we see the r-squared is 0.61 – that means that if we know x_1 we can predict about 60 percent of the variation in y.

Based on this regression our best prediction for y would be given by the equation:

$$y = 0.2 + 1.7x_1$$

Here is the regression of y on x_2:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.4031     2.2047   2.904   0.0198 *
x2            1.1188     0.3711   3.015   0.0167 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.325 on 8 degrees of freedom
Multiple R-squared:  0.5319,    Adjusted R-squared:  0.4734
```

The results for x_2 are similar. The estimated coefficient is smaller, although based on the standard errors, we aren't especially confident

about this.[6] The coefficient on x_2 has a t-statistic of 3, meaning it also passes the conventional test for statistical significance. And the r-squared of this regression is 0.53, meaning that knowing x_2 allows us to predict about half the variation in y.

Based on the second regression our best prediction for y would be:

$$y = 6.4 + 1.1x_2$$

The two regressions are supposed to be telling us the effects of x_1 and x_2 on y. So we might think that to estimate their combined effects, we should just add up the individual effects. In other words, we might think that to predict y on the basis of both x_1 and x_2, we should add the two equations together, perhaps using the average of the intercepts. This would give us something like:

$$y = 3.3 + 1.7x_1 + 1.1x_2$$

But this is wrong!

As the figure shows, this prediction does not fit the observed values of y at all. The horizontal axis shows the predicted values of y on the basis of the previous equation and the observed values of x_1 and x_2 – for example, for observation 1, we would predict that $y = 3.3 + 1.7 \cdot 2 + 1.1 \cdot 0 = 6.7$. The vertical axis is the actual values of y (e.g. 3 for observation 1). And the diagonal line shows the prediction. No one would pick that line as the best fit to those points. This prediction is no better than not looking at x_1 and x_2 at all – the mean squared error is approximately equal to the variance of y, so we would do just as well just drawing a horizontal line through the mean value of y.

We find another puzzle when we run a regression with both x_1 and x_2 on the right-hand side. Here is what we get:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2334     2.7828   0.443   0.6710
x1            1.1911     0.4986   2.389   0.0483 *
x2            0.6616     0.3512   1.884   0.1016
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.018 on 7 degrees of freedom
Multiple R-squared:  0.7421,    Adjusted R-squared:  0.6684
```

R-squared is higher in the new regression: 0.74, compared with 0.61 and 0.53 in the earlier ones. Even adjusted r-squared, with penalizes us for adding another independent variable, is higher. Using both variables allows us to predict the values of y more accurately

[6] As a rough rule of thumb, the difference in the coefficients should be greater than the sum of their standard errors for us to believe that coefficients would be different in the population the sample is drawn from.

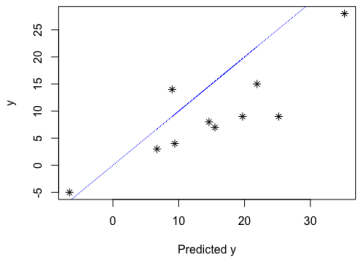


Figure 6: Actual y and a bad prediction.

than we can using either x_1 or x_2 alone. But our parameter estimates have changed – they are smaller, and much less precise. Compared with the regressions on each of them separately, the t-statistics for x_1 and x_2 have fallen from 3.5 and 3 to 2.4 and 1.9 respectively. In other words, combining the two variables makes us much less certain about the effects of either one of them. In the case of x_2, we are no longer confident that it has an effect on y at all. This is somewhat paradoxical – using both variables makes us more confident in our predictions of the value of y, but less confident about how it is affected by either of the independent variables.

The explanation for both these puzzles is the same: x_1 and x_2 are correlated with each other. In this case, they have a correlation coefficient of 0.55, meaning that about half the variation in each variable is shared with the other one. If we want to estimate the effect of each variable on its own, we can only use the variation that is not shared with the other variable – when y, x_1 and x_2 are all high, as in observation 4, there's no way to know if the high value of y is due to the high value of x_1 or the high value of x_2. The shared variation of the two variables can't be used to estimate the effects of either one. This means that when two independent variables are highly correlated, we don't have much information about their individual effects, so our estimate cannot be very precise. This is illustrated in Figure 7 – only about half the variation of x_1 and x_2 is available for estimating their separate effects on y. For predicting the dependent variable, however, this is not a problem – when both x_1 and x_2 are high we can be confident that y will also be high, even if we don't know which of them is responsible.

On the other hand, when we run the regressions separately, our results seem more precise, but are they biased. Remember, x_2 is often high when x_1 is high. So a regression of y on x_1 will show a stronger effect, because some of the apparent effect of x_1 will reflect the influence of x_2, which we haven't included. In terms of Figure 7, using the whole circle for x_1 will mean using the shared variation as well, which may be telling us about the effect of x_2 rather than x_1. This upward bias in the separate regressions is why the line in Figure 6 is too steep. (If the independent variables were negatively correlated, the bias would go the other way, and the line would be too shallow.)

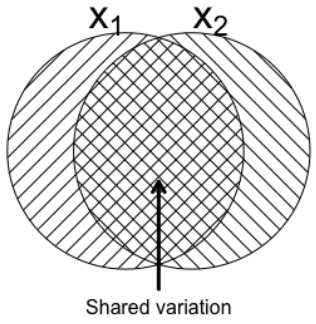


Figure 7: Coefficient estimates can use only variation that is not shared with any other independent variable. So when two independent variables are correlated, there is less information available to estimate their effects.

To see the problem more clearly, imagine the data looked like this:

Observation	y	x_1	x_2
1	3	2	1
2	6	4	2
3	0	0	0
4	12	8	4
5	-3	-2	-1

In this case, the relationship between two variables is obvious. $y = x_1 + x_2$ will allow us to predict y perfectly. But so will $y = 2x_2 - x_1$, or $y = 0.5x_1 + 2x_2$, or $y = -2x_1 + 7x_2$, or any of an infinite number of other equations. In this case, we have complete confidence in our predictions of y. But because all the variation in x_1 and x_2 is shared, we have no idea what the independent effect of either one is.

Only if the independent variables are completely uncorrelated with each other will the coefficient estimates and standard errors be the same in the multivariate regression as in the separate bivariate ones.

Regression results are normally presented in tables.

For the remaining assignments in this class — and more importantly, in your future empirical work — you will need to present the results of regressions. Here are a few guidelines and suggestions.

The image nearby is a typical regression table from a published paper. We won't worry for now about the context of the table or what specifically it's showing. Instead let's just look at a few features of it as a table.

Table 2. --- Difference in Differences Estimates of the Impact of Wrongful Discharge Laws on Private-Sector Real Hourly Wages by Major Industry Categories using CPS-MORG files: 1979-2013

	Total Private		Priv. Manufacturing		Priv. Nonmanufacturing	
	D-D w/state, month-year, demo FEs, regXyr,	Col. 1 w/ stateXt, stateXt$_2$, demoXt	Same as Col. 1	Same as Col. 2	Same as Col. 1	Same as Col. 2
	(1)	(2)	(3)	(4)	(5)	(6)
Implied Contract	0.33	0.27	-0.14	0.46	0.51	0.25
	(1.14)	(0.95)	(1.29)	(0.97)	(1.24)	(0.94)
Public Policy	-0.23	-0.61	0.38	0.44	-0.89	-1.08
	(1.00)	(0.76)	(1.08)	(0.81)	(1.08)	(0.79)
Good Faith	2.00**	2.34**	2.73*	3.33***	1.69+	1.96*
	(0.64)	(0.86)	(1.36)	(0.85)	(0.86)	(0.85)
R-Sq	0.30	0.30	0.30	0.30	0.30	0.30
N	4,652,387	4,652,387	1,375,973	1,375,973	3,276,414	3,276,414

*** p < 0.001, ** p < 0.01, * p < 0.05, + p < 0.10.

Note: Each entry shows output from analysis using specification 1, which is a difference and differences regression in which the dependent variable is 100 times the natural logarithm of real U.S. year 2000 hourly wages of private-sector wage and salary employees in 50 U.S. states. Each entry in columns 1 and 2 show output for all private-sector workers, columns 3 and 4 for private-sector manuafcturing workers, and columns 5 and 6 for private-sector nonmanufacturing workers. The three main rows of the table show, respectively, the estimates for the regression coefficients β_{IC}, β_{PP}, and β_{GF}. The real U.S. year 2000 hourly wage data are taken from Curret Population Survey Monthly Outgoing Rotation Group monthly files for the years 1979 to 2013. All models include state main effects and indicators for each month-year in the sample, indicators for individuals' demographic characteristics, as well as interactions between four Census-region dummies and calendar year dummies. Models in columns 2, 4, and 6 also include interactions between state and year linear and quadratic time trends, and interactions between workers' demographic characteristics and a linear time trend. Models are weighted using CPS earnings weights. Huber-White robust standard errors are used to allow for unrestricted error correlations across observations within states.

1. It has an informative title, which tells you exactly what the coefficients represent.

2. If the models use different dependent variables, these are indicated for each column with descriptive names, rather than whatever variables were used internally.

3. One table reports multiple models, or specifications, which are numbered. This is standard — it both saves space, and makes easy to compare the coefficients. Often, though not in this case, the different models will include different independent variables. In

that case, the space for variables that aren't part of that model are left blank.

4. The variables have informative, grammatical names — again, they aren't just the variables used by the software.

5. The standard errors of the estimates are in parentheses, under the coefficients. Practice is not consistent here — some people put the t value (the coefficient divided by the standard error) in this place instead. You need to pay attention to which it is when looking at a regression table. One easy way to check: standard errors are always positive, while t values have the same sign as the coefficient. In my opinion, standard errors are more informative.

6. Conventional significance levels are indicated with asterisks. Usually, one star means significance at the 10 percent level, two stars means significance at the 5 percent level, and three stars means significance at the 1 percent level. Sometimes, as here, a different convention is used — this should always be given in the notes to the table. In my opinion significance levels are not really needed. All the information in them is already given by the standard errors, and people focus too much on conventional significance levels. That said, the vast majority of regression tables do indicate significance levels, either with stars or in some other way.

7. R-squared and number of observations are at the bottom. You may see r-squared, adjusted r-squared, or both. (Adjusted r-squared is reduced slightly based the number of independent variables the model includes.) It doesn't really matter which one you give; unless the model has almost as many variables as observations, they will very close. Giving both is redundant. N is helpful if the different models have very different numbers of observations, for instance because a variable used in some of the models is missing from many observations, or because the models are being estimated for different populations, as here. If N is similar across the models, you can safely leave it out.

8. Other statistics, such as residual standard error and F-statistic, are not reported here. Many statistics packages will give these by default, but they are almost never needed and should be omitted unless you have a specific reason for including them.

9. The regression includes a number of fixed effect, or dummy, variables which are included only as controls. These are listed in the notes but their coefficients are not given in the table. If there are dummies included in some but not all models, this is typically indicated with a line in the regression table. E.g. you might have a

line labeled "State dummies?" under the coefficients, with a "Yes" or "No" to indicate whether the model includes state fixed effects. Other times, as here, the notes say which dummies are included in which model.

Producing regression tables in stargazer

Stargazer is a widely used R package for producing attractive regression tables without a lot of work. If you give it one or more regression objects, stargazer will produce tables similar to the one shown here. Here are a few suggestions for using it.

Simply calling stargazer will output a table. Alternatively, you can assign the results to an object:

```
reg.table <- stargazer(model)
```

You can then call the object (in this case `reg.table`) at some other point in your code to output the table. This can be useful if, for example, you want to do all your regressions, including producing the tables, in one place in your code, but present them somewhere else.

The basic syntax of stargazer is:

```
stargazer(model1, model2 ..., type=, options)
```

Here the first set of arguments is any number of models that you want to report in the table, `type` is the form of the output (usually either `'html'` or `'latex'`) and options are a variety of options that control the layout and appearance of the table.

There are many options, which you can read about under help for stargazer. A few that you will often want to use (including for assignments in this class) are `covariate.labels`, to give informative names to the independent variables; `dep.var.labels`, to give informative names to the dependent variables; `omit`, to leave some independent variables out of the table; `omit.stat`, to leave out unwanted statistics; and `add.line`, to add additional lines to the regression table. Unless you have some reason to do otherwise, I'd suggest using

```
omit.stat = c('adj.rsq', 'f', 'n', 'ser')
```

in the options whenever you use stargazer. Note that if you omit variables form some specifications but not others, you will need to indicate this manually; the `omit.labels` option in stargazer does not work reliably.

To get stargazer tables to appear properly in a knitted file, you may need to specify `results='asis'` in the chunk options.

These notes are intended to give a very basic introduction to time series data for people doing data work in policy settings. They introduce some basic terminology; explain why serial correlation and nonstationarity in time series data present problems fro conventional statistical analysis; and introduce some basic techniques for diagnosing and dealing with these problems.

If you want to explore time-series eocnometrics in more depth, a good recent textbooks are Enders 2015, *Applied Econometric Time Series*. An older but still useful one is Hamilton 1994, *Time Series Analysis*.

Time series data is data where each observation represents a different moment in time.

Time series data is data indexed over time - that is, where each observation is associated a different point in time. Most macroeconomic data is time series data. The frequency of the data refers to the time between observations.n macroeconomics, time series data is most often monthly, quarterly or annual, but in other settings other frequencies are common.

Some of the same issues may occur with data that is indexed in some other way, for instance geographically. Panel data is data that is indexed both over time and some other identifier. For instance, we may have observations of various countries at various dates. In these notes we will deal with only time series data.

If the values of the variable are not affected by the time they take place, then we can work with time series data the same way we would work with any other data. More often though, the variation in time series variables has a time dimension - the value of the variable is in some way related to the date of the observation. Two central concepts used to describe the time dimension of variation are serial correlation and nonstationarity.

Data that is not indexed over time is called *cross-sectional* data. While most macroeconomic data is time-series, most microeconomic data is cross-sectional, since there will be many distinct individual cases at any given moment.

With time series data, *lagged* observations are observations from earlier periods. A model that includes, for example,“four lags", includes the values of the variable in the four preceding periods. Similarly, *leads* are the values of the variable in subsequent periods.

First-differencing a variable means looking at the *change* in its value between periods rather than its *level* in a given period. First-differencing variables is a common transformation with time-series data. Where the data reflects a process of growth, it is more appropriate to look at the percentage change in the variable or, equivalently, the first difference of its log value.

Time series data often displays serial correlation: The value of a variable in one period tends to be similar to the value in nearby periods.

In thinking about the problems of time series data, we need to keep in mind three related but distinct concepts: serial correlation, autocorrelation and autoregression.

Serial correlation means that the values of a variable are correlated with the values that come before and after them. Serial correlation is a sample property – that is, it is something we directly observe (or don't observe) in the data in front of us. Let's call our variable x. Positive serial correlation means that high values of x are likely to be followed by high values, and low values of x are likely to be followed by low values. (Negative serial correlation means that high values are likely to be followed by low values – this is uncommon in real data.) We can calculate it simply as the correlation of the variable with its own values shifted one period forward or backward.

Autocorrelation is a property of a model. It means that after we have written down a regression with independent variables we think are likely to explain our variable, the remaining variation still displays serial correlation. Formally, serial correlation means that if we estimate

$$x_t = \beta_0 + \beta_1 y_{1,t} + \beta_2 y_{2,t} + ... + \beta_n y_{n,t} + e_t$$

the residuals from the model (the e terms) show serial correlation. A value of y that is higher than the model predicts on the basis of the y terms will be followed by another one that is also higher, and a value that is lower than predicted will be followed by another one that is also lower.

Autocorrelation is different from serial correlation in that serial correlation is a property of the sample – the values we actually observe. Autocorrelation is a property of the sample *plus* a model that we've used to explain it (along with our observations of the independent variables in the model). Once we have estimated a model, we can definitely say whether it displays autocorrelation - we just look at the residuals and see if they are serially correlated.

If the right-hand side variables (the ys in the equation above) do not themselves have any serial correlation, the model's autocorrelation will be the same as the serial correlation in the dependent variable.

Autoregression is a population concept. It describes a relationship we suppose or imagine exists in the "true" underlying population or process our sample draws from. Autoregression means a data generating process in which the previous values of the variable have a causal effect on the current value. In other words, when we speak

of an autoregressive process, we are imagining a process that is described by

$$x_t = \beta_0 + \beta_1 y_1 + ... + \beta_n y_n + \beta_{n+1} x_{t-1} + e$$

In other words, to say the process is autoregressive is to say that we think the value of the variable in one period has a causal effect on the value in the nest period. This is not something we can ever directly see in the data. Rather, it describes our belief about the real economic or social process that generates the data.

In summary: x shows serial correlation if there is a statistical relationship between the variable's value in one period and its value in the next period. A model of x show autocorrelation if it fails to fully account for the serial correlation of x – if there is still serial correlation in the residuals, the variation in x that is left over after we've accounted for the effects of the independent variable(s). We say the process generating x is autoregressive if we think earlier values of x really do have an effect on later values. The first describes a sample, the second describes a model based on the sample, and the third describes the population or data generating process that the sample is supposed to come from.[7]

As a concrete example, think about two variables we might observe: traffic fatalities by day, and the number of housing units in a city by year. In both cases, we might well find serial correlation: days with traffic fatalities lower or higher than average might well be followed by other days with fatalities lower or higher than average, and years with higher or lower number of housing units than the average (of all the years in the sample) might well be followed by other years with a higher or lower number of housing units. So both variables would display serial correlation.

in the case of the auto traffic fatalities, we might construct a model in which we estimate fatalities as a function of total miles driven, the time of year, etc. We might well find that after incorporating these variables into our model, there was no longer any serial correlation in the remaining variation in traffic fatalities (the residual). Or we might find that we had eliminated some but not all of the serial correlation. In this case, our model would display autocorrelation – higher-than-predicted observations of fatalities would often be followed by other higher-than-predicted observations. In this case, however, we might well believe that this is simply because we have left out some important serially-correlated variable that influences traffic fatalities. We are not likely to believe that the "true" process generating fatalities is autoregressive (unless we can come up with some story where fatalities today cause more fatalities tomorrow). For the housing stock, by contrast, we do believe that the underlying process is autoregres-

[7] While the definitions here are standard ones, usage is not entirely consistent. In particular, many textbooks are focused almost entirely on regression models. Because of this they don't discuss serial correlation as a descriptive term for a variable itself, but instead use it as a synonym for autocorrelation. But it's better to keep the concepts distinct.

sive – since housing units last for many years, the number of units in existence last year has a strong influence on the number of units in existence this year. In this case, if we came up with a model that did not display autocorrelation, that would almost certainly mean we'd done something wrong.

Serial correlation means we see similar values in observations close to each other in time. Autocorrelation is serial correlation our model has not gotten rid of. An autoregressive process is one with serial correlation that a good model should not get rid of.

What about a process that is not autoregressive, but does display serial correlation? – that is, one where values near each other in time are correlated, but future values are *not* affected by current values. In econometrics, we usually describe this as a *moving average* process. This is the case when there is some other variable that affects the one we are looking at, which acts over several periods. In the case of the traffic-accident example above, we would expect traffic fatalities per day to follow a moving-average process because seasonal effects and weather conditions continue over multiple days.

All time series variables will show some degree of serial correlation. But when serial correlation is large, standard regression results become misleading.

Any sample of time series will show some positive or negative serial correlation simply by chance. The vast majority of real-world time series data will show some positive serial correlation beyond what we would expect from chance, because real social and economic processes don't line up exactly with the periods we are observing, and almost always have some persistence over time. For an underlying process to have no serial correlation, it would have to describe some concrete development whose full effects are entirely contained in the period of time covered by a single observation. Not many things that happen in the world fit this description.

While any time series data will display some serial correlation, a modest amount of serial correlation is not an issue. If looking at the previous observation helps us only a little in predicting the current observation, then we don't need to worry about making any adjustments for serial correlation. But if most of the variation is shared with the observations before and after, our regression results will be misleading. A high value of a variable is not very informative if we already knew it was likely to be high, based on the variable's previous value.

For example, if we look at the US unemployment rate since 1945, the mean value is 5.8 and the standard deviation is 1.6. So if our next

observation of the unemployment rate is 3.6, that will be about 1.5 standard deviations below the mean – around the 5th percentile. Such an extreme value would, normally, be quite informative for any statistical model of unemployment. It could be taken as evidence for a causal relationship between unemployment and any other variable that was also unusually low or high. But suppose the current unemployment rate is 3.7. In this case, an observation of 3.6 would be giving us little new information. The unemployment rate has been 3.7 in the past few months, and seldom changes by much from month to month. So if it went down to 3.6 next month that would give us only a little new information about the factors influencing it.

Another way of thinking about this is that when there is a high degree of serial correlation in your variable, you "really" have fewer observations. The sample shown in Figure consists of 100 observations. But visually, it doesn't seem to consist of 100 independent values, but a half dozen or so distinct peaks and troughs. The large peak in the center consists of many observations, but we might suspect that it really reflects a single underlying event.[8] This means that if you run a regression on serially correlated data without making any correction, you will be more confident in your parameter estimates than is justified by the data.

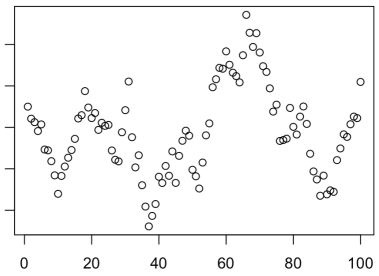


Figure 8: Serially correlated values.

[8] In fact the data was generated randomly and doesn't represent anything.

For example, suppose we have 100 observations of two variables – a small number of observations for cross-sectional data, but a reasonably large one for time-series data. If the two variables are actually independent of each other and are normally distributed, we would expect that 5 percent of the time we would find, just by chance, a relationship between them at appeared significant at the 5 percent level. (That's what "significant at the 5 percent level" means.) But now suppose that the variables were still completely independent of each other, but were serially correlated, so that each variable's current value was related to its own previous value. Specifically, suppose they are described by processes like this:

$$y_t = \beta X_t + c(y_{t-1} - \beta X_{t-1}) + e_t \tag{1}$$

Here y is the variable itself, X is a set of other observable variables that influence it, e is the error term, meaning the combined influence of everything that we don't observe, and c is the autocorrelation term, showing how persistent y is over time. (In real world cases, c would normally be between 0 and 1.)

The value of c is telling us how much an unusually high or low value of y in one period affects the predicted value of y in the next period. For example, if $c = 0.5$, that means that if we found y was 10 units higher than we would predict on the basis of other observable variables in one period, we would expect it to be 5 units higher than

we would otherwise predict in the next period. In other words, $c = 0.5$ means that half the variation of y persists from one period to the next.

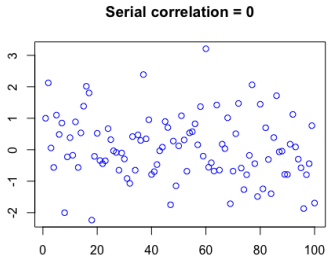
The figures nearby show randomly generated variables with different degrees of autocorrelation. As you can see, even fairly high levels of autocorrelation are not easily visible to the eye. It's not until we reach 0.75 or so that the data shows clear structure. But even levels of serial correlation well below this can make conventional significance tests misleading.

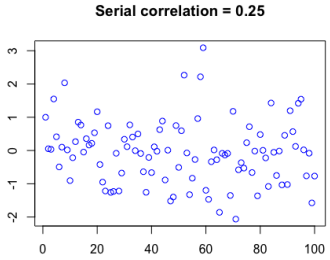
Again, with normally distributed variables with no relationship, we would expect to to find a "significant" relationship by chance 5 percent of the time. But if the variables, while independent, are generated by an autoregressive process like that of Equation 1, then a regression of one variable on the other will pass conventional significance tests more often,, even though there is still no relationship between them. Specifically, with $c = 0.25$, we will find a signifiant relationship about 6 percent of the time; with $c = 0.5$, we will find one about 10 percent of the time; with $c = 0.75$ we will find one about 25 percent of the time; with $c = 0.9$ we will find one about half the time; with $c = 0.95$ we will find one about two-thirds of the time; and with $c = 0.99$ we will find an apparently significant correlation between unrelated variables about three-quarters of the time. (There is no formula for these values, they're just how it works out.)

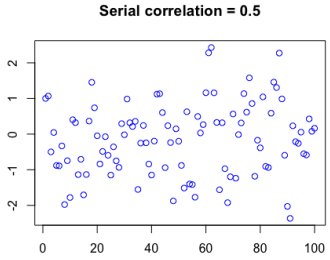
Clearly, using conventional significance tests with serially correlated data can be misleading, especially when the degree of serial correlation is high.

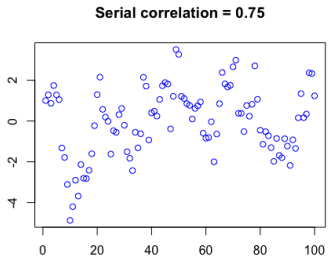
One solution is simply to do a conventional regression, but keep in mind that the appropriate standard errors on the coefficient estimates are wider than the ones the regression reports. You can find the serial correlation in the residuals of the regression, and then increase your standard errors by a corresponding amount. This approach is formalized by using *robust* standard errors, which are calculated to give a sense of the uncertainty of the estimate given the serial correlation in the residuals.[9] Some people argue that, given that well-behaved normal distributions are rare in economic data, robust standard errors should be used by default in any kind of regression. But the case for using them is strongest with time series data that displays a high degree of serial correlation.

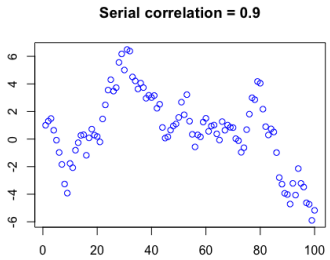
There are a number of different ways to calculate robust standard errors, but they normally give similar results. (And, we should always keep in mind, there are no "true" standard errors - their purpose is just to give us a sense of how confident we should be in our exact estimate.) We don't need to go into the details of how robust errors are calculated here. The general idea is that conventional stan-

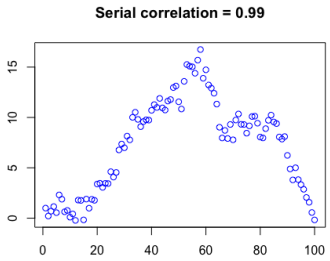












dard errors assume the error terms are independent of each other and of the dependent variable, and are normally distributed. (This is sometimes called "white noise" residuals.) Since we can observe the residuals, we don't have to make this assumption. Robust standard errors take into account both the correlation between error terms and any greater (or lesser) proportion of very large errors compared with what we would expect under a normal distribution.

In R, you can calculate robust standard errors using the `sandwich` and `lmtest` packages. Then you write:

```
m <- lm(...)
coeftest(m, vcov.  = vcovHC(m, type = 'HC1'))
```

where the `lm` in first step is the regression you want to do.

Other statistics software will also have ways to produce robust standard errors. The result will be a set of regression coefficients with robust standard errors. (The coefficients themselves should be the same as in the original regression.)

Instead of (or in addition to) using robust standard errors, another solution is to transform the variable(s) to reduce or eliminate the serial correlation, as discussed further below. This is essential if the series is nonstationary, as discussed in the following section.

Time series may be stationary or nonstationary. In a stationary series, the expected value is the same for later observations as for earlier ones; in a nonstationary series, the expected value changes over time.

We do not know very much of the future
Except that from generation to generation
The same things happen again and again. Stationary
Men learn little from others' experience.
But in the life of one man, never
The same time returns. Sever Nonstationary
The cord, shed the scale. Only
The fool, fixed in his folly, may think
He can turn the wheel on which he turns.

(from T. S. Eliot, Murder in the Cathedral)

Figure 10: T. S. Eliot on the difference between stationary and nonstationary processes.

We think of the process generating time series data as either stationary or nonstationary. This is a population concept - it is a property of the underlying process we imagine our sample represents, it is not something we can ever directly observe in data. Deciding whether the underlying process is stationary is very important for making statistical inferences from the data.

Stationary processes are also sometimes referred to as *ergodic*. This means that the mean and other distribution moments converge to a single value as the sample size grows.

A stationary process is when the expected value of the variable is the same when we take observations from earlier or from later periods. A nonstationary process is one where the expected value depends on the period it comes from. Imagine you pick a sample of observations of the variable over 20 consecutive periods. Then imagine you look at another sample of 20 observations from a much later period. If the average value of the second sample is generally close to the average from the first sample, the variable is stationary. If the average of the second sample is systematically different form the

average from the earlier period (that is, the average in the later period is almost always larger, or almost always smaller) then the series is nonstationary.

Conventional statistical analysis assumes there is a population with a true mean, standard deviation and so on, which samples will cluster around; the larger the sample, the closer your estimates will get to the true value. With nonstationary data, there are no such values. Later samples will systematically different from earlier ones and, in general, the means will not converge to any particular value as the sample gets larger. Standard regression results will be misleading or meaningless when applied to nonstationary data. And unlike with stationary data, the problem gets worse rather than better as the sample size rises.

Serial correlation is evident with just a few observations, but stationarity and nonstationarity can only be distinguished if sample is large enough. Samples from stationary data will eventually converge to the same value; but in the presence of serial correlation, it will take some amount of time for this convergence to happen. To take a simple example: Many variables we observe are seasonal, with systmeatically different values in the winter and the summer. If we observed such a series over less than a year, such variables would appear to be nonstationary, but with data covering several years they might turn out to be stationary.

A process may be nonstationary for one or more of three reasons: It may have a trend; it may have a structural break; or it may have a unit root.

- A series with a trend is described by a process like $x_t = at + \beta_1 y_1 .... e$. In other words, on average it increases or decreases by a certain amount each period, in addition to the effects of other independent variables and random or unexplained variation.
- A series with a unit root is described by a process like $x_t = x_{t-1} + \beta_1 y_1 .... e$. In other words, the value in each period is equal to the value in the previous period, plus the effects of other independent variables and random variation.
- A series with a structural break is described by a process like $x_t = a_t A + \beta_1 y_1 .... e$, where a is zero in earlier periods and 1 in later periods.

A trend and a structural break both describe a process that changes over time; in the one case, the change is continuous, while in the other it is a once and for all shift at a particular time. The difference between a trend and a unit root may be less obvious at first, but it is very important. The question is whether the variable is changing over

time for some reason independent of its own values, or whether the variable itself is conserve over time. Both trend and unit root imply a continuous change in the mean value of samples over time. But with a trend, an exceptionally high or low value in one period does not tell us anything about the expected value in one period, while with a unit root, if the variable is exceptionally high or low in one period, we should expect to see it similarly high or low in the next period.

In the case of a variable with a unit root or a trend or both, the difference between expected values grows without limit as the dates become more separated in time. In the case of a series with a break, the expected difference between later observations and earlier observations will be equal to A. In both cases, no matter how large our samples are, if they are widely separated in time their means will not be the same.

One important difference between trend and unit-root processes is how we should adjust our forecast when the current observation departs strongly from the past trend. With a variable that follows a trend, we would expect it to gradually return to its old path. With a unit-root proess, we would expect the old rate of growth to continue from the current value. This is illustrated in Figure 11.

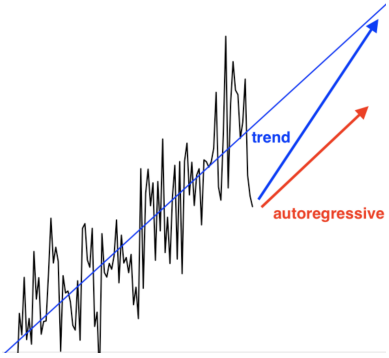


Figure 11: Forecasts for a trend versus autorergessive process.

Nonstationary variables with a trend or unit root or both have some further properties which create large problems for standard statistical tools. First, the variance and standard deviation of a sample grows without limit as the sample grows larger. With a stationary variable, a larger sample means more precise estimates of the population variance. But for a variable with a trend or unit root, the population variance is infinite. So the larger the sample you have, the larger the variance you will observe.[10]

If a a variable with a trend or unit root is regressed on another variable with a trend or unit root, then as the sample gets larger, the r-squared of the regression will go to one and the t-statistic will go to infinity. In other words, if you look at a larger enough sample, any nonstationary variable will appear to explain essentially all the variation in any other nonstationary variable, and the relationship between them will appear more and more precisely estimated as the sample grows. This is true regardless of whether there is any other relationship between the variables. Conversely, if a variable with a trend or unit root is regressed on one or more variables all of which are stationary, the r-squared of the regression will go to zero as the sample side gets larger, and with a large enough sample, none of the right-hand side variables will be significant. Again, this is true regardless of any relationship between the economic processes generating the variables.

The reason for this behavior is simple: A nonstationary variable

[10] This is also true for certain stationary distributions, such as power laws. But in the vast majority of cases where a stationary process is assumed, including any variable without a time component, it is safe to also assume a finite population variance.

changes over time. The farther into the future you take your next observation from, the farther it is likely to be from the observations you've taken so far. With the passage of enough time, this variation will dominate any other source of variation in the variable - whether the value is high or low will come to depend entirely on whether the observation is an early or a late one. This means that any two nonstationary variables will be perfectly correlated if your sample covers a long enough period, because time is passing for both of them. And conversely, any relationship with a stationary variable that you might find in a shorter sample will disappear as variation over time comes to account for almost all the variation.

For example, Figure 12 is a scatterplot showing the global price of beef and total employment in Oklahoma. Visually, there seems to be a definite relationship between them. And if we do a regression of the log of Oklahoma employment on the log of the beef price, as shown in Figure 13, we will find what looks like a very strong relationship – a coefficient of 0.12, with a t-value of 25. This means that for each 10 percent increase in the price of beef, employment in Oklahoma increases by a precisely estimated 1.2 percent.

But before we start spinning out theories about Oklahoma's dependence on global agriculture markets and the significance of this for its political culture, we should note that both these series are nonstationary. The apparent relationship is simply a result of the fact that both of these series tend to rise over time. We can see this clearly if we look at the residuals from the regression, as shown in Figure 14.[11] These are clearly not white noise. The correlation between adjacent residuals is 0.95 – much too high. That tells us that we cannot trust the result of the regression. Whenever you find a strong relationship in time series data, you should check the residuals. Unless they look reasonably close to white noise, you cannot trust your results.

The bottom line is that you cannot do statistical analysis on nonstationary data. You can do analysis on autocorrelated data, but you have to be cautious – your standard errors will be too small.

There are variety of formal tests for autocorrelation and for nonstationarity, but in most cases it is straightforward to decide whether they apply.

Serial correlation in the most basic sense simply means the correlation of a variable with its own values shifted in time. In practice we are usually interested in the relationship of the variable with its own values one period earlier. So the simplest test for serial correlation is to regress the variable on its own values shifted one period back-

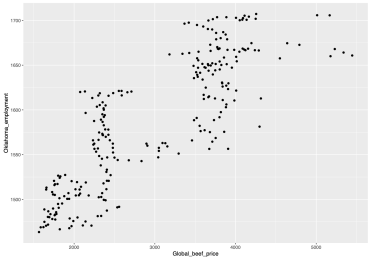


Figure 12: The global price of beef (x) and total employment in Oklahoma (y).

```
##
## Call:
## lm(formula = log.OK.employment ~ log.beef.price)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.057339 -0.015311  0.001294  0.016223  0.058907
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.41147    0.03723  172.20   <2e-16 ***
## log.beef.price  0.12043    0.00468   25.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
##
## Residual standard error: 0.02386 on 252 degrees of free
## Multiple R-squared:  0.7243, Adjusted R-squared:  0.723
## F-statistic: 662.1 on 1 and 252 DF,  p-value: < 2.2e-16
```

Figure 13: A regression of Oklahoma employment on the global beef price.

[11] In R, you can get the residuals of a regression with `x$residuals`, where x is a regression object.

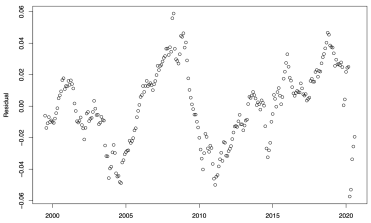


Figure 14: Residuals from the regression of Oklahoma employment on the global beef price.

ward (or forward). If the coefficient is significantly different from zero, there is serial correlation. If the coefficient is not significantly different from one, then we should treat the series as nonstationary.

A more general version of this is the autocorrelation function (ACF), which is simply the correlation across many lags – that is the, the correlation of the current value with the value one period ago, two periods ago, three periods ago, etc. Statistics packages can straightforwardly calculate and plot these – the example in Figure was generated in R using the `acf` function from the `tseries` package. Autocorrelation at 0 lags is always equal to one; in series without serial correlation, the value at all other lags should be close to zero. For a series that is stationary but shows serial correlation, the values will be high at low lags but will descend toward zero as the lag length gets longer. For nonstationary series, the values will remain close to one even after many lags.

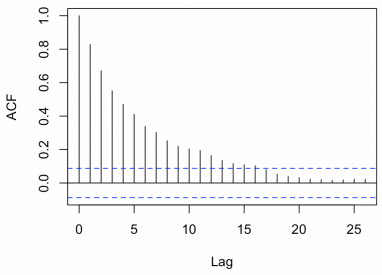


Figure 15: Output of an autocorrrelation function.

Note that there is no critical values or definite tests here – judgement is required. But if a plot like Figure does not show a clear downward slope, it's best to treat the series as nonstationary. If the plot shows a downward slope but early values are above 0.25 or so, it's a good idea to use robust standard errors and/or transform the data to reduce the serial correlation.

Again, when the sample is small, it can be difficult or impossible to distinguish between a series that is nonstationary and one that is stationary but strongly serially correlated. With a large enough, sample, the serial correlation in a series with a trend or unit root will always approach 1. One hundred observations is enough to bring serial correlation in a series with a unit root close to one; with a trend, it depends on how big the trend is relative to other sources of variation. With a series with a structural break, serial correlation will approach some value that will be somewhere between zero and one depending on how big the shift at the break is compared with other sources of variation.

A more formal test for a unit root is the augmented Dickey-Fuller test. Essentially, this tests the hypothesis that the coefficient on the previous period's value is equal to one. If we fail to reject this null, we conclude that the series has a unit root. In R, you can do this with the `adf.test` function from the `tseries` package. Again, the hypothesis being tested here is that there *is* a unit root, so if the p-value is large, we fail to reject the null and conclude that a unit root is present.

As discussed above, unit roots are only one of the ways in which a series can be nonstationary. If a series has a trend but is otherwise stationary (what is called *trend-stationary*), the Dickey-Fuller test will reject the null. So it can't be used by itself to decide if a series is

stationary.

Series that have a unit root will follow a *random walk*. That is, if you take one sample and then a second sample at a different time, the farther they are separated in time, the farther you expect their means to be. But unlike with a trend, you can't predict in which direction the difference will be.

In general, while formal tests are useful and convenient, they are no substitutes for common sense. If you suspect a series is nonstationary, the first thing to do is to look at the plot and see if there is any obvious structure to it, or if later values are systematically different from earlier ones. You should also think about what substantive economic, social or biophysical process lies behind the data – is there reason to think that it would or would not change systematically over time?

Various transformations can turn nonstationary series into stationary ones and reduce or eliminate serial correlation.

The main ways we deal with a series that appears nonstationary are to transform it by converting levels to changes; detrending it; or normalizing it by some other variable. For stationary data with a high degree of serial correlation these are also options; we also might change the frequency of the data, by using annual rather than quarterly or monthly data for example.

If our series are nonstationary but we are convinced that the economically relevant relationship is between levels, so that using changes would be inappropriate, we can instead estimate an *error correction model*. The details of this are beyond the scope of these notes, but it's something you may want to learn about if you continue to work with time series data.

Many series have a high degree of serial correlation in their levels, but a much lower one in their changes. For example, if this month's unemployment rate is 3.7 percent, I can be quite confident that next month's unemployment rate will be between 3.5 and 4 percent. Prior to the pandemic, it was essentially unheard of for the unemployment rate to change by more than a few tenths of a percent from one month to the next. The *change* in the unemployment rate, on the other hand, is much harder to predict – it is very common for a month with rising unemployment to be followed by one with falling unemployment, in a pattern that in most periods looks basically random.

If the variable typically changes by a percent of its current value, rather than by a fixed amount, then we should use percentage changes rather than absolute changes. This is the same as using the change in the log of the variable. This is appropriate for things we think of as

growing at a steady percentage rate, such as GDP. Either way, using changes rather than levels is appropriate whenever you suspect the series has a unit root.

To *detrend* a variable, first regress its value on time (i.e. a variable that takes a value of 1 in the first period, 2 in the second period, and so on.) Then use the residual from this regression in place of the original variable. This is appropriate for variables we think change steadily over time, but not in a way that depends on their own past values.

To *normalize* a variable, you simply divide it by some other variable. For this to make sense, the two variables should (a) have similar time series properties and (2) be related to each other in an economically meaningful way. For example, many macroeconomic variables can be sensibly divided by GDP or by population.

For data that shows serial correlation over shorter but not longer lags, reducing the frequency of your observations can get rid of the serial correlation without losing much information. For example, if you have quarterly data that shows autocorrelation at less than four lags, or monthly data that shows autocorrelation at less than 12 lags, you could convert it to annual data by picking the average value for each year (for a flow) or the last value of each year (for a stock).

In any of these cases, you will want to confirm that your transformed variable appears stationary and does not have an excessive degree of serial correlation.

In some cases, we have two (or more) series that appear to be nonstationary, but we are convinced the relevant relationship is between their levels, not changes, and there is no suitable third variable to normalize them by. We describe these series as being *cointegrated*. In that case, we can estimate an error correction model. The basic idea of this is that while the individual variables are nonstationary, the relationship between them may be stationary. This is sometimes described as being like a person rambling through a park while holding a dog on a leash. While their individual paths may follow an (in this case literal) random walk, the distance between them is limited by a leash. In this case there may be no relationship between the direction the person and dog are walking at any given moment, but there will be a strong relationship between their positions.[12]

In economics, we might think of the relationship between the exchange rate and trade flows as being a good candidate for this kind of models. Both the exchange rate and the balance of trade between two countries may be nonstationary, but the balance of trade may nonetheless be strongly related to the level of the exchange rate. In this case, an error correction model would be appropriate.

The details of how this is done are beyond the scope of these

[12] This metaphor is taken from Murray 1994, "A Drunk and Her Dog: An Illustration of Cointegration and Error Correction.".

notes. In the great majority of real world cases, the right response to a nonstationary series is to transform it to make it stationary.