Simple Regression Theory II

Assessing how good the regression equation is likely to be

Assignment 1A gets into drawing inferences about how close the regression line might be to the true line. We make these inferences by examining how close the data points are to the regression line we draw. If the data points line up well, we infer that our regression line is likely to be close to the true line. If our data points are widely scattered around our regression line, we consider ourselves less certain about where the true line is.

If we are unlucky, the data points will fool us by lining up very well in a wrong direction. Other times, again if we are unlucky, the data points will fail to line up very well, even though there really is a relationship between our X and Y variables. If those things happen, we will come to an incorrect conclusion about the relationship between X and Y. Hypothesis testing, which we discuss soon, is how we try to sort this out.

In Simple Regression Theory I, I discussed the assumption that the observed points come from a true line with a random error added or subtracted. I called this Assumption 1, rather than just The Assumption, because there we will need some more assumptions if we want to make inferences about how good our regression line is for explaining and predicting.

Let's review the notation that we are using: Each data point is produced by this "true" equation: $Y_i=\alpha+\beta X_i+e_i$. This equation says that the *i*th data point's Y value is the true line's intercept, plus the *i*th X value times the true line's slope, plus a random error e that is different for each data point. That is Assumption 1 expressed with algebra.

Once we accept Assumption 1, there are two parameters, the true slope β and the true intercept α, that we want to estimate, based on our data.

An **estimate** is a specific numerical guess as to what some unknown parameter is. In the Assignment 1A instructions, page 2, the example spreadsheet shows 0.067857 as an estimate of the true slope β. The word "estimate" in that sentence is important. 0.0678857 is *not* β. It is an *estimate* of β.

An **estimator** is a recipe, or a method, for obtaining an estimate. You have used two estimators of β already. One is the eyeball estimator (plot the points, draw the line, inspect the graph, calculate the slope). The other is the least squares estimator (type the data into a spreadsheet, implement the least squares formula, report the result).

Here's an idea that can take some getting used to: Estimates are random variables. This may seem like a funny idea, because you get your estimate from your data. However, if you use the theory we have been developing here, each Y value in your data is a random variable. That is because each Y value includes the random error that is causing it to be above or below the true line. Your estimates of the intercept and slope, $\hat{\alpha}$ and $\hat{\beta}$ ("alpha-hat" and "beta hat"), derive from these Y values. The $\hat{\alpha}$ and $\hat{\beta}$ that you get from any particular data set depends on what the errors e_i were when those data were generated. This makes your estimates of the parameters random variables.

In assignment 1, everybody in the class had data with the same true slope and intercept, but different errors. As a result of the different errors, everybody got different values for $\hat{\alpha}$ and $\hat{\beta}$. That is what it means to say that the estimates of α and β are random variables.

Because each estimate is a random variable, each estimator (recipe for getting the estimate) has what is called a **sampling distribution**. This means that each estimator, when used in a particular situation, has an expected value and a variance – expected spread. A good estimator – a good way of making an estimate – would be one that is likely to give an estimate that is close to the true value. A good estimator would have an expected value that is near the true value and a variance that is small.

As mentioned, **Assumption 1** is that there is a **true line**, and that the observed data points scatter around that line due to **random error**.

By adding more assumptions, we can use the sampling distribution idea to move toward assessing how good an estimator the least squares method is. We can also use this idea to assess how good an estimator the eyeball method is.

Assumption 2: The expected value of any error is 0.

We can write this assumption in algebraic terms as:

Expected(e_i) = 0 for all each observation *i*. (i numbers the observations from 1 to N),

This means that we assume that the points we observe are not systematically above or below the true line. If, in practice, there are more points, say, above the true line than below it, that is entirely due to random error.

If this assumption is true, then the expected values of the least squares line's slope and intercept are the true line's slope and intercept. Algebraically, we can write: Expected($\hat{\alpha}$) = α and Expected($\hat{\beta}$) = β.

Those formulas mean that the least squares regression line is just as likely to be above the true line as below it, and just as likely to be too steep as to be not steep enough. The slope and intercept estimates are aimed at their targets. The jargon term for this is that the estimators are **unbiased**.

The next two assumptions are about the variances and covariances of the observations' errors (e_i).

Assumption 3: All the errors have the same variance. Expected(e_i^2) = σ^2 for all i

Expected(e_i^2) is the variance of *i*th data point's error e_i . The variance formula usually involves subtracting the mean, but here the mean of each error is 0. That is assumption 2, that the expected value of e_i is 0.

Notice that, in the formula Expected(e_i^2) = σ^2, there is a subscript *i* on the left side of the equals sign, but no *i* subscript on the right side. This expresses the idea that all the errors' variances are the same.

I use σ^2 ("sigma-squared") here for the variance of the error, because that is the textbook convention.

All the errors having the same variance means that all of the observations are equally likely to be far from or near to the true line. No one observation is more reliable than any other, deserving more weight.

In almost any data set, some data points will be closer to the true line than others. This, we assume, is entirely by luck. If one or two points stick way out, like a basketball player in a room full of jockeys, we have doubts about assumption 3.

Assumption 4: All the errors are independent of each other.
Expected(e_ie_j) = 0 for any two observations such that i ≠ j

Expected(e_ie_j) is the covariance of the two random variables e_i and e_j . The assumption that this is 0 for any pair of observations means this: If the one point happens to be above the true line, it is not any more or less likely that the next point will also be above the true line.

A comment on making these assumptions

These assumptions 1 through 4 are not made just from the data. Rather, these assumptions must come primarily from our general understanding of how the data were generated. We can look at the data and get an idea about how reasonable these assumptions are. (Assignment 2 does this.) Usually, we pick the assumptions we make based on a combination of the look of the data and our understanding of where the data came from.

When you fit a least squares line to some points, you are implicitly making assumptions 1 through 4. When you draw a straight line by eye through a bunch of points, you are also implicitly making assumptions 1 through 4.

What if you are not comfortable with making those assumptions? Later in the course, we will get into some of the alternative models that you can try when one of more of these assumptions do not hold. For example, we will talk about non-linear models. For now, let us stick with the linear model, which means that we accept those assumptions.

Variances of the least squares intercept and slope estimates

Earlier, it was pointed out that the estimates for the slope and intercept, $\hat{\alpha}$ and $\hat{\beta}$, are random variables, so each has a mean, or expected value, and a variance. From assumptions 1 through 4, one can derive formulas for the variances of $\hat{\alpha}$ and $\hat{\beta}$ when they are estimated using least squares. (You cannot derive formulas for the variances of $\hat{\alpha}$ and $\hat{\beta}$ when they are estimated using the draw-by-eye method. To estimate those variances, you could ask a number of people to draw regression lines by eye. That is what this class does for Assignment 1!)

Here are formulas for the variances of the least squares estimators of $\hat{\alpha}$ and $\hat{\beta}$. (If you would like to see how these formulas are derived, please consult a statistics textbook.)

$$Variance(\hat{\alpha}) = \sigma^2[\frac{1}{N} + \frac{\overline{X}}{\sum_{i=1}^{N}(X_i - \overline{X})^2}] \qquad Variance(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

In these formulas, σ^2 is the variance of the errors. Asumption 3 says all errors have the same variance, which is σ^2.

One of the themes of this course is that you can learn something from formulas like these, if you take the trouble to examine them. Don’t let them intimidate you!

Each formula has σ^2 multiplied or divided by something. This means that the variances of our regression line parameters are both proportional to σ^2, the variance of the errors from the true equation. An implication is that the smaller the errors are, the less randomness there is in our $\hat{\alpha}$ and $\hat{\beta}$ values, and the closer our estimated parameters are likely to be to the true parameters.

In the formula for the variance of $\hat{\beta}$, the denominator gets bigger when there are more X's and when those X's are more spread out. A bigger denominator makes a fraction smaller. This tells us something about designing a study or experiment: If you want your estimates to come out close to the truth, get a lot of observations that are spread out over a big range of values of the independent variable.

Those formulas have a major drawback: We cannot directly use them! That is because we don't know what σ^2 is.

What we can do is estimate σ^2. We call our estimate s^2, and calculate it from the residuals like this:

$$s^2 = \frac{\sum_{i=1}^{N} u_i^2}{N - 2}$$

In that expression, the residuals are designated by u_i. The numerator says to square each residual and then add them all up, giving you the sum of the squares of the residuals. (The least squares line is the line that makes this sum the smallest.)

To get the estimate of the variance of the errors, we divide the sum of squared residuals by N–2. That makes the estimated variance like an average squared residual. I say “like” an average squared residual because we divide by N–2, rather than by N. Why divide by N-2, instead of N? Using N-2 allows for the fact that the least squares line will fit the data better than the true line does. The least squares line is, by the definition of “least squares,” the line with the smallest sum of squared residuals. Any other line, including the true line, will have a larger sum of squared residuals.

N-2 is the **degrees of freedom**. In your introductory statistics course, you used models in which the degrees of freedom is N-1. For instance, when you estimated a population variance based on the data in a sample of N from that population, the degrees of freedom was N-1.

Why did you subtract 1 from N then, and why are we subtracting 2 now? Then, to estimate the variance of a population around the population’s mean, you used one statistic, the mean of the sample. Now, we are using predicted values calculated from two statistics, the slope and the intercept of the regression line. It’s easier to get a predicted value that is close to your observation when you have two estimated parameters, instead of just one. Using N–2 allows for that.

In general, the number of degrees of freedom is N-P, where P is the number of parameters used to get the number(s) that you subtract from each observation's value.

To repeat, the reason you subtract P from N in these situations is that the estimated mean and the estimated regression line fit the data better than the actual population mean or the true line would. The more parameters you have, the easier it is to fit your model to the data, even if the model really is not any good. You have to make the degrees of freedom smaller to make the estimated variance come out big enough to correct for this.

Substituting s^2 for σ^2 in the formulas for the variances of $\hat{\alpha}$ and $\hat{\beta}$ gives estimated variances for $\hat{\alpha}$ and $\hat{\beta}$. You just change every "σ" into an "s".

$$Estimated\,Variance(\hat{\alpha}) = s^2[\frac{1}{N} + \frac{\overline{X}}{\sum_{i=1}^{N}(X_i - \overline{X})^2}]$$

$$Estimated\ Variance(\hat{\beta}) = \frac{s^2}{\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

Standard Error

The **standard error** of something is the square root of its estimated variance.

s, the square root of s^2, is called the **standard error of the regression**. Here is the formula:

$$s = \sqrt{\frac{\sum_{i=1}^{N} u_i^2}{N - 2}}$$

$\hat{\beta}$ has a standard error, too. It's the square root of the estimated variance of $\hat{\beta}$::

$$Std.Error\,of\ \hat{\beta} = \frac{s}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2}}$$

This measures of how much risk there is in using $\hat{\beta}$ as an estimate of β.

You can deduce from the formula that the risk in the estimate of β is smaller if:

the actual points lie close to the regression line (so that s is small),
or
the X's are far apart (so that the sum of squared X deviations is big).

Hypothesis tests on the slope parameter

To do conventional hypothesis testing and confidence intervals, we must make another assumption about the errors, one that is even more restrictive:

Assumption 5 (needed for hypothesis testing): Each error has the normal distribution. Each error e_i has the normal distribution with a mean of 0 and a variance of σ^2.

The normal distribution should be familiar from introductory statistics. The normal distribution's density function has a symmetrical bell shape. A normal random variable will be within one standard deviation of its mean about 68.26% of the time. It will be within two standard deviations 99.54% of the time. This means that, on average, 99.54% of your data points should be closer to the true line than 2 times σ^2. The normal distribution is tight, with very few outliers.

If the errors are normally distributed, as assumption 5 states, then the following expression has the t distribution with (Number of observations - 2) degrees of freedom.

$$t = \frac{\hat{\beta}-\beta}{\mathit{Standard\ Error}(\hat{\beta})}$$

If you would like to see this derived, please consult a statistics textbook. The general idea is that we use the t distribution for this rather than the normal or z distribution to allow for the fact that the regression line generally fits the data better than the true line. The regression line fits too well, in other words. Using the wider t distribution, rather than the narrower normal distribution, corrects for this.

A branching place in our discussion. At this point, we can either go through the mechanics of hypothesis testing, or we can discuss the basic philosophy of hypothesis testing. Which is best for you depends on what you already know about statistics and how you learn.

1. For the philosophy discussion, please see the downloadable Philosophy of Hypothesis Testing file. Then come back here to see the mechanics.
2. If you want to see the mechanics first, continue reading here.

Hypothesis testing mechanics for simple regression

To use the t formula above to test hypotheses about what the true β is, do this:

1. Plug your hypothesized value for β in where the β is in this expression:

   $$\frac{\hat{\beta}-\beta}{\mathit{Standard\ Error}(\hat{\beta})}$$

   Usually, the hypothesized value is 0, but not always.

2. Plug in the estimated coefficient where the $\hat{\beta}$ is, and put the standard error of the coefficient in the denominator of the fraction.

3. Evaluate the expression. This is your t value.

4. Find the critical value from the t table (there is a t-table is in the downloadable file of tables) by first picking a significance level. The most common significance level is 5%, or 0.05. That tells you which column in the table to use Use the row in the t table that corresponds to the number of

degrees of freedom you have. In general, the degrees for freedom is the number of observations minus the number of parameters that you are estimating. For simple regression, the number of parameters is 2 (the parameters are the slope and the intercept), so use the row for (Number of observations - 2) degrees of freedom.

5. The column and the row give you a "critical t value." If your calculated t value (ignore any minus sign) is bigger than this critical value from the t table number, reject the hypothesized value for β. Otherwise, you don't reject it.

In Assignment 1A, you will expand the spreadsheet from Assignment 1 to include a cell that calculates the value of the t fraction. This "t value," as we will call it, will be only for testing the hypothesis that the true β is 0. This will do steps 1, 2, and 3 for you.

In step 5 above, you are supposed to ignore any minus sign. That is because we are doing a "two-tailed" test. We do this because we want to reject hypothesized values for the true slope that are either too high or too low to be reasonable, given our data. The column headings in the t table in the downloadable file show significance levels for a two-tailed test.

Some books present their t tables differently. They base their tables on a one-tailed test, so they have you use the column headed 0.025 to get the critical value for a two-tailed 0.05-significance-level test. The critical value you get is the same, because the t table value at the 0.05 significance level for a two-tailed test is equal to the t table value at the 0.025 significance level for a one-tailed test. When you use other books' t tables, read the fine print so you know which column to use.

Significance levels and types of errors

Why use a significance level of 0.05? Only because it's the most common choice. You can choose any level you want. In choosing a significance level, you are trading off two types of possible error:

1. Type 1 error: Rejecting an hypothesis that's true
2. Type 2 error: Refusing to reject an hypothesis that's false

If you are testing the hypothesis that the true slope is 0, which is what you usually do, then these become:

1. Type 1 error: The true slope is 0, but you say that there is a slope. In other words, there really is no relationship between X and Y, but you fool yourself into thinking that there is a relationship.

2. Type 2 error: The true slope is not 0, but you say that the true slope might be 0. In other words, there really is a relationship between X and Y, but you say that you are not sure that there is a relationship.

Smaller significance levels, like 0.01 or 0.001, make Type 1 errors less likely. Actually, the significance level *is* the probability of making a Type 1 error. At a 0.001 significance level, if you do find that your estimate is significant, you can be very confident that the true value is different from the hypothesized value. You will only be wrong one time in a thousand. If your hypothesized value is 0, which it usually is, you can be very confident that the true parameter is not 0. If what you are testing is your estimate of the slope, which it usually is, then you can be very confident that the slope is not 0 and that there really is a relationship between your X and your Y variables.

The drawback of a small significance level is that it increases the probability of a Type 2 error. With a small significance level, it is more likely that you will fail to detect a true relationship between X and Y. With a small significance level, you are demanding overwhelming evidence of a relationship being there.

The ideal way to pick a significance level is to weigh the consequences of each type of error, and pick a significance level that best balances the costs. What researchers often do, however, is pick .05, because they know that this significance level is generally accepted.

Watch out for confusing or contradictory terminology:

1. “α-level” of significance. The significance level is sometimes called the α-level. Try not to confuse this with the intercept α of a linear regression equation.

2. “High” significance level. A low “alpha” number is a “high” level of significance. If you reject the hypothesis that a coefficient is 0 at the 0.001 level, that coefficient is “highly” significant.

3. 95% significance or 0.05 significance. These terms are interchangeable. The same goes for 99% significance or 0.01 significance.

Usually, the context will enable you keep these straight.

Confidence intervals for equation parameters

An alternative way to test hypotheses about β is to calculate a confidence interval for β and then see if your hypothesized value is inside it. The 95% confidence interval (two-tailed test) for β is:

95% confidence interval
for the slope estimate

$$\hat{\beta} \pm t_{0.05}\, \mathit{Standard\ Error}(\hat{\beta}) \quad = \quad \hat{\beta} \pm t_{0.05} \frac{s}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}}$$

The $t_{0.05}$ in the formulas above means the value from the *t* table in the column for the 0.05 significance level for a two-tailed test and the row for N-2 degrees of freedom. For a 99% confidence interval, you would use $t_{0.01}$.

The ± means that you evaluate the expression once with a + to get the top of the confidence interval, then you evaluate it with a - to get the bottom.

This confidence interval is called “two-tailed” because it has a high end and a low end that are equidistant from the estimated value. That is what the ± does. (A one-tailed confidence interval would have one end that is above the estimated value and go off to minus infinity to the left, or it could have one end that is below the estimated value and go off to plus infinity to the right.)

Look now at the right hand version of the confidence interval formula above. Let’s explore the underlying relationships. The s in the numerator tells us that the confidence intervals for the coefficients get bigger if the residuals are bigger. The denominator tells us that having more X values and having them more spread out from their mean makes the confidence interval smaller. Again, this tells us that we gain confidence in our estimate if the residuals are small and if we have a lot of spread-out X values.

For the intercept, α, the 95% confidence interval is:

$$\hat{a} \pm t_{0.05}\ s \sqrt{\frac{1}{N} + \frac{\overline{X}}{\sum_{i=1}^{N}(X_i-\overline{X})^2}}$$

Confidence interval for the prediction of Y

We can also calculate confidence intervals for a prediction. Let X_0 be the X value for which we want the predicted Y value. We'll call that predicted Y value $\hat{Y}_0$. $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}X_0$.

The 95% confidence interval for the prediction is:

$$\hat{Y}_0 \pm t_{0.05}\ s \sqrt{\frac{1}{N} + \frac{(X_0-\overline{X})^2}{\sum_{i=1}^{N}(X_i-\overline{X})^2} + 1}$$

Here's what we can see in this expression: The confidence interval is the big expression to the right of the ±. The width of this confidence interval depends on the sum of squared residuals ***s*** and depends inversely on $\sum(X_i-\overline{X})^2$, the sum of the squared deviations of the X values from their mean. Big residuals, relative to the spread of the X's, make s big, which makes the confidence intervals wide. More X values, or more spread-out X values, make $\sum(X_i-\overline{X})^2$ expression larger, which makes the confidence interval narrower. The $(X_0-\overline{X})^2$ expression in the numerator under the square root sign tells us that confidence interval gets wider as the X_0 you choose gets further from the mean of the X's.

R^2, a measure of fit

An overall measure of how well the regression line fits the points is R^2 (read "R squared"). R^2 is always between 0 (no fit) and 1 (perfect fit). R^2 shows on how big the residuals are in relation to the deviations of the Y's from their mean. It is customary to say that the R^2 tells you how much of the variation in the Y's is "explained" by the regression line.

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}$$

Notice that the only difference between the top and the bottom of the fraction is that the top has y-hats and the bottom has y-bar. The y-hats are the predicted values for Y from the regression equation. Y-bar is the average of the Y values.

Here is the R^2 formula in words:

$$1 - \frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2} = 1 - \frac{\textit{Sum of the squares of the residuals}}{\textit{Sum of the squares of the distances of the Y values from their mean}}$$

The R-Squared tells you how much your ability to predict is improved by using the regression line, compared with not using it.

The least possible improvement is 0. This is if the regression line is no help at all. You might as well use the mean of the Y values for your prediction.

The most possible improvement is 1. This is if the regression line fits the data perfectly.

That is why the R-squared is always between 0 and 1. The regression line is never worse than worthless (0), and it can't be better than perfect (1).

> Some statistical software reports an "adjusted" R-squared. This allows for the fact that an X variable that is really completely unrelated to your Y variable will probably have some relationship to Y in your data just by luck. The adjusted R-squared reduces the R-squared by how much fit would probably happen just by luck. Sometimes this reduction is more than the calculated R-squared, so you can have an adjusted R-squared that is less than 0.

All conclusions from the R-squared are based on the assumptions behind using least squares being true. If those assumptions are not true, then it is possible that using the regression line to predict would be worse than worthless.

Correlation

“Correlation” is a measure of how much two variables are linearly related to each other. If two variables, X and Y, go up and down together in more or less a straight line way, they are “positively correlated.” If Y goes down when X goes up, they are “negatively correlated.”

In a simple regression, the correlation is the square root of the regression’s R^2. For this reason, the correlation (also called “the correlation coefficient”) is designated as r.

An alternative formula for r is:

The correlation coefficient

$$r = \frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i-\bar{Y})^2}}$$

r can range from -1 to +1

With some algebra, you can show that r and $\hat{\beta}$ are related. Multiply the top and bottom of the r formula by the square root of the sum of the squares of the x deviations. You get this:

$$r = \frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})\sqrt{\sum_{i=1}^{N}(X_i-\bar{X})^2}}{\sum_{i=1}^{N}(X_i-\bar{X})^2\sqrt{\sum_{i=1}^{N}(Y_i-\bar{Y})^2}}$$

Look at the parts of this fraction that are not under square root symbols. These are the left halves of the numerator and denominator. Do you recognize them as the same as the formula for the least squares $\hat{\beta}$?

$$r = \hat{\beta}\frac{\sqrt{\sum_{i=1}^{N}(X_i-\bar{X})^2}}{\sqrt{\sum_{i=1}^{N}(Y_i-\bar{Y})^2}}$$

The higher the slope of the regression line is (higher absolute value of $\hat{\beta}$), the higher r is. When X and Y are not correlated, r = 0. According to the formula above, $\hat{\beta}$ must be 0 as well. So, if X and Y are unrelated, the least squares line is horizontal.

Durbin-Watson Statistic

In Assignment 2 you will see that the regular goodness of fit statistics (R^2 and t) cannot detect situations where a linear least squares model is not appropriate. The Durbin-Watson statistic can detect some of those situations. In particular, the Durbin-Watson statistic tests for serial (as in "series") correlation of the residuals.

Here is the Durbin-Watson statistic formula:

$$DW = \frac{\sum_{i=2}^{N} (u_i - u_{i-1})^2}{\sum_{i=1}^{N} u_i^2}$$ (The *u*'s in this formula are the residuals.)

A rule-of-thumb for interpretation of DW:

DW < 1 indicates residuals track each other. A positive residual tends to be followed by another positive residual. A negative residual tends to be followed by another negative residual.

DW near 2 indicates no serial correlation.

DW > 3 indicates residuals alternate, positive-negative-positive-negative..

For a more formal test, see the Durbin-Watson table in the downloadable file of tables.

Serial correlation of residuals indicates that you can do better than your current model at predicting. Least Squares assumes that the next residual will be 0. If there is serial correlation, that means that you can partly predict a residual from the one that came before.

If the Durbin-Watson test finds serial correlation, it may indicate that a curved model would be better than a straight line model. Another possibility is that the true relationship is a line, but the error in one observation is affecting the error in the next observation. If this is so, you can get better predictions with a more elaborate model that takes this effect into account.