Economics 250 — Regression notes

Regressions

Often the questions we care about in economics isn't the mean of variables, but rather the relationship between variables. For example: "How much does an extra year of education affect wages?" and "Do countries with less pollution have higher life expectancy?" are both economics questions about relationships rather than means.

In the language of algebra, what we're saying is that we care not just about the level of the variable, but the slope of the line describing the relationship between variables. In the language of calculus, we care about not just the value but also the derivative. The first example question above is asking "What is the slope of the relationship?" while the second is a hypothesis test of the slope: "Is there evidence that the slope is negative?"

A regression lets us estimate of how one variable is affected by another variable (or several other variables). It lets us predict the level of the relationship between variables ("What is the average wage of someone with a Bachelor's degree?") and the slope of that relationship ("What is the average increase in wage of an increase of 1 year of education?"). Just like the sample means we have talked about so much in Economics 250, we create it from the information in a population sample.

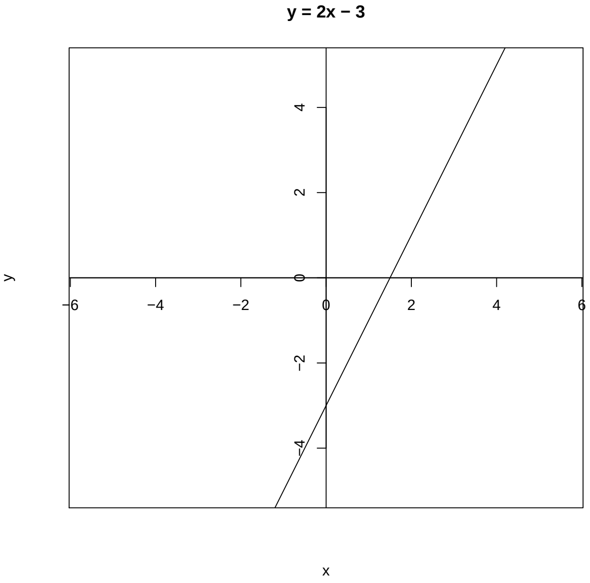
Recall the equation for a line:

$$y = mx + b$$

where m is the slope of the line and b is the vertical intercept (where it intersects the y-axis). For example:

$$y = 2x - 3$$

is the straight line depicted to the right.



What we're trying to accomplish with a regression is to get a line such as $y = 2x - 3$ but using our data to determine the values 2 and -3.

Because we will soon want to allow for the possibility of multiple x values, we adopt a slightly different notation: we'll write a regression line as:

$$y = \beta_1 + \beta_2 x$$

This is just the same as the straight line equation with β_1 replacing b and β_2 replacing m.

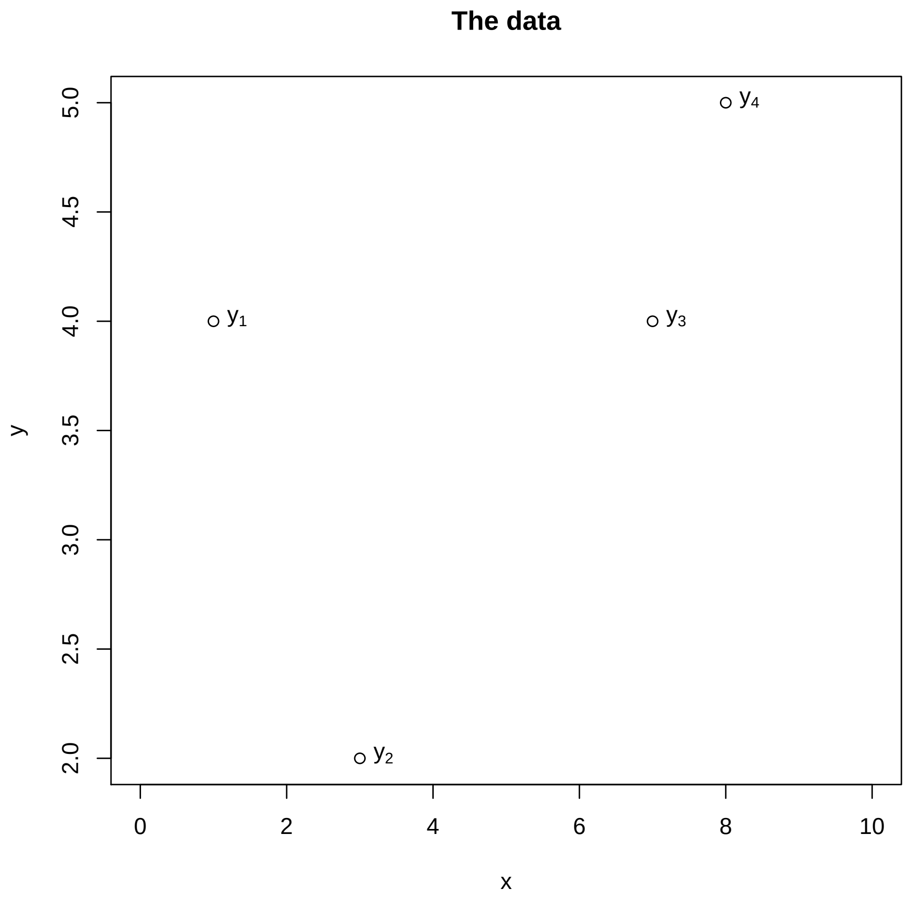
Because statistics deals with the real world where we accept that relationships and variables are never perfect matches, we are also going to add another term, u, called an "error term":

$$y = \beta_1 + \beta_2 x + u$$

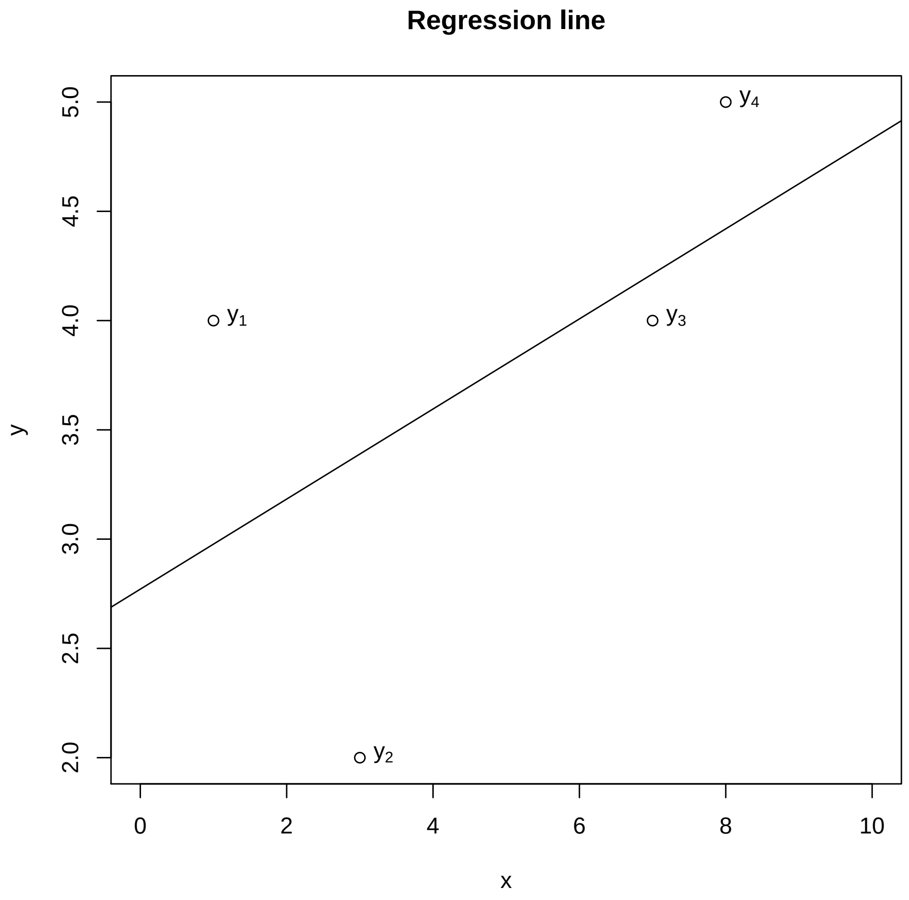
The idea behind u is that we think that there is a linear relationship between y and x, but that there's also some randomness, perhaps due to other factors, measurement error, or some other random component in how x affects y. Essentially what we're saying is that y is determined by x but also some "noise." In future Econometrics courses you'll spend a lot of time discussing the assumptions and requirements that we need u to satisfy: for now, we'll just accept it as a technical detail and move on.

Finding a regression line

There are may different ways to find such a line, but the easiest and most common is to use Ordinary Least Squares (OLS). Consider a simple data set of just 4 (x, y) pairs: $(1, 4), (3, 2), (7, 4), (8, 5)$. Graphically, those points look like this:

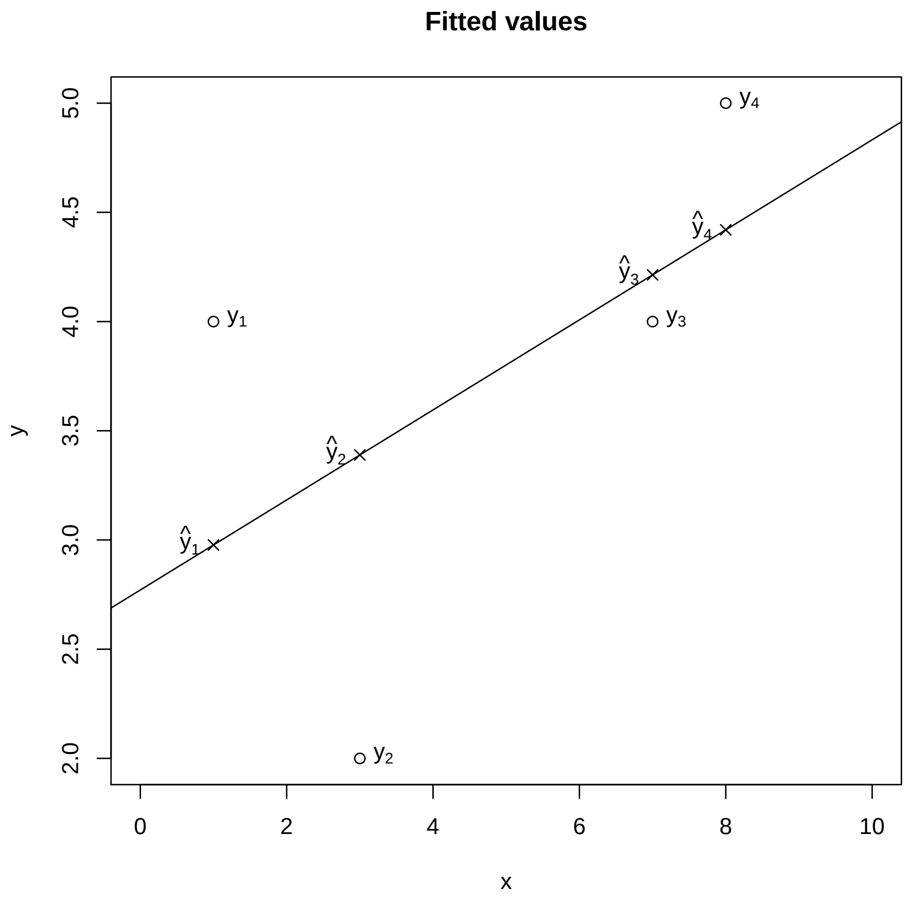


What we want to do, using our sample, is to find the best straight line through the graph, such as in the following:



There are, of course, many such lines. What OLS does is pick the line that minimizes the "sum of squared residuals" (SSR). So what's a residual?

If we take any line (it doesn't have to be the best one), we can figure out the predicted values of y along that line by just plugging in the values of x from our data and seeing what value of y the line gives us. We denote these predicted values of y as $\widehat{y}$. The following graph adds the $\widehat{y}$ values:

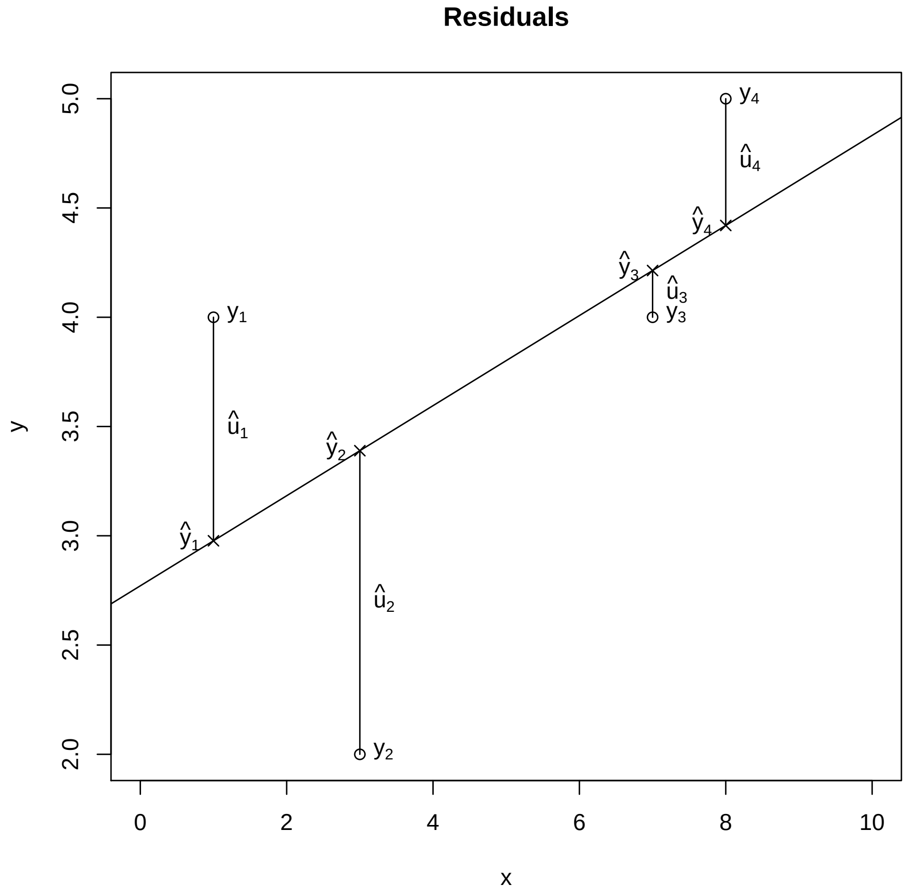


Notice that they are all on the regression line: this isn't coincidence: a regression line can *only* make predictions of points along that line.

We can then see how much the *actual* y values in the data differ from the *predicted* $\widehat{y}$ values. This distance, denoted $\widehat{u}$, is called a residual:

$$\widehat{u} = y - \widehat{y}$$

Graphically, the residuals are the vertical lines shown below:



The line picked by OLS is simply the line with values of β_1 and β_2 that make the following expression as small as possible:

$$\sum_{t=1}^{n} \widehat{u}_t^2 = \sum_{t=1}^{n} (y_t - \widehat{y}_t)^2$$

where the t subscript denotes the value of y, $\widehat{y}$, and $\widehat{u}$ for the tth observation in the data.

The essential idea behind using squares is to "penalize" choices of the line that make significant errors in prediction. Squaring also has the benefit of removing negatives, since negative errors are just as bad as positive errors.

Once we have these estimates, we put a hat over these values to indicate that they are estimates from the data: $\widehat{\beta}_1$, $\widehat{\beta}_2$.

In the straight line case (one constant plus one variable), it is possible to calculate the values of $\widehat{\beta}_1$ and $\widehat{\beta}_2$; you can see the details in the textbook: I won't ask it. In the more general case, where we have more than one variable on the right-hand side, the calculations involve matrix algebra and/or calculus, which is far beyond this course.

In practice, we simply use a regression package such as Gretl (free software) or STATA (commercial) to calculate our $\widehat{\beta}$ estimates. Excel (with the Data Analysis toolpak) can do some basic regressions as well, but is quite limited.

Using regression results

Suppose we have obtained a set of data to help us analyse an interesting question. I'll continue with the wage example we covered in class (this data is also linked from the course website so that you can run the regressions yourself).

Suppose the model we want to estimate is the following:

$$wage = \beta_1 + \beta_2 educ + u$$

That is, wage is a linear function of years of education (plus some random noise). We can load this into our regression program, ask it to run a regression (in Gretl: Model -> Ordinary Least Squares, then select "wage" as the dependent variable and add "educ" as the regressors (in addition to the "const" variable, which is there by default).

Gretl produces the following (the output from STATA is similar):

```
Model 1: OLS, using observations 1-526
Dependent variable: wage

             coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  const       -0.904852    0.684968     -1.321    0.1871
  educ         0.541359    0.0532480    10.17     2.78e-22 ***

Mean dependent var   5.896103   S.D. dependent var   3.693086
Sum squared resid    5980.682   S.E. of regression   3.378390
R-squared            0.164758   Adjusted R-squared   0.163164
F(1, 524)            103.3627   P-value(F)           2.78e-22
Log-likelihood      -1385.712   Akaike criterion     2775.423
Schwarz criterion    2783.954   Hannan-Quinn         2778.764
```

There are lots of numbers here, most of which we aren't going to discuss in this course. The important ones for us are the following:

- "const" and "educ" coefficients: these are the estimates $\widehat{\beta}_1$ and $\widehat{\beta}_2$
- Next to those are the standard errors, which are important for hypothesis testing (and confidence intervals) for our β values.
- The t-ratio and p-value are simply the t statistic and associated p-value for testing the null hypothesis $\beta_i = 0$ against the alternative $\beta_i \neq 0$.

- Finally we want to take note of the “R-squared” value (0.164758).

The first thing to discuss is the coefficients. $\widehat{\beta}_1 = -0.904852$ is the estimate of the vertical intercept of our line. Its interpretation is that it is the expected wage of someone who has 0 years of education. Constant coefficients only sometimes have a useful interpretation in regression results: if we aren't likely to see someone with $educ = 0$, the interpretation isn't particularly useful, as in this case.

$\widehat{\beta}_2 = 0.541359$ is much more important to us: it is the *slope* of the regression line: it tells us how much we would expect to see wage increase, on average, for someone with one extra year of education.

Hypothesis testing and confidence intervals

Testing a hypothesis about one of the values of β is not much different than testing the hypothesis about a sample mean such as $\overline{x}$. Recall that for a sample mean, we would build a t statistic by calculating:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

Often we simply called the $(s/\sqrt{n})$ term the “standard error” of $\overline{x}$. In our regression, the standard error is given to us by the regression software, so we can just plug it in: no dividing by $\sqrt{n}$ needed. So our t statistic for testing values of β_2 is just:

$$t = \frac{\widehat{\beta}_2 - \beta_{20}}{SE(\widehat{\beta}_2)}$$

where β_{20} is our null hypothesis value (the equivalent of μ for a mean test), and $SE(\widehat{\beta}_2)$ is just the standard error value given by the regression program.

For example, to test the following:

$$H_0 : \beta_2 = 0.5$$
$$H_a : \beta_2 > 0.5$$

we would calculate:

$$t = \frac{\widehat{\beta}_2 - 0.5}{SE(\widehat{\beta}_2)} = \frac{0.541359 - 0.5}{0.053248} = 0.77$$

then we would look for p values in the t statistic table. There is one difference here from calculating means, however: with calculating means we used $df = n - 1$ to figure out which table row to look at. Here, we're going to subtract the number of β coefficients being estimated at once: 2. So in this case, with 526 observations, we would have 524 degrees of freedom.

The regression results above (and those of most regression programs) automatically perform a t test for us, to test the hypotheses:

$$\begin{aligned} H_0 :& \beta_i = 0 \\ H_a :& \beta_i \neq 0 \end{aligned}$$

for each $\widehat{\beta}_i$ (2 in this case). Thus at a glance we can tell whether our included variables have coefficients that are statistically different from 0.[1] You can also calculate these t statistics yourself: they are simply $\widehat{\beta}/SE(\widehat{\beta})$ for each coefficient.

To calculate a confidence interval for the value of any of the β values, we simply calculate:

$$\left[\widehat{\beta}_i - t^* SE(\widehat{\beta}_i), \widehat{\beta}_i + t^* SE(\widehat{\beta}_i)\right]$$

which, other than using $SE(\widehat{\beta}_i)$ instead of $s/\sqrt{n}$, is just like the confidence intervals we've seen all along.

For example, a 95% confidence interval for β_2 from the results above is

$$[0.541359 - 1.984(0.0532480), 0.541359 + 1.984(0.0532480)] = [0.4357, 0.6470]$$

where 1.984 is the critical value for $df = 100$, which is the next-smallest df value in our textbook's Table D to the actual $df = 524$. You could, using a computer, calculate the more precise $t^*_{524} = 1.964502$, but this will make only a slight difference to the resulting interval.

Fitted values, and goodness of fit

Another thing we might do in a regression is to make predictions: given x, what does the model predict for the value of y? For example, in the model above, what is the prediction for an individual with 16 years of school? This is just a matter of plugging our $educ = 16$ value into the equation using our $\widehat{\beta}$ estimates:

$$\begin{aligned} \widehat{y} &= -0.904852 + 0.541359(educ) \\ &= -0.904852 + 0.541359(16) = 7.756892 \end{aligned}$$

So our model predicts a wage of $7.76 for someone with 16 years of school.

[1] Most regression programs go a step further by including asterisks (*): * indicates weak evidence (p between 0.1 and 0.05), ** indicates stronger evidence (p between 0.05 and 0.01), and *** indicates very strong evidence (p smaller than 0.01). You can see the results above have no asterisks for the constant (we cannot reject that it equals 0, even at the weak $\alpha = 0.1$ level), but has three for *educ*, for which the test very strongly rejects the hypothesis that the coefficient on *educ* equals 0.

Once we have a fitted value, we can very easily calculate a residual, as well. For example, the first data point in the sample is: $wage = 3.10, educ = 11$. The fitted value $\widehat{y}$ is 5.05. The residual, then, is:

$$\begin{aligned}\widehat{u} &= y - \widehat{y}\\ &= 3.10 - 5.05\\ &= -1.95\end{aligned}$$

which either means our model overpredicted the wage for this data point, or, if we believe that the model accurately represents reality, this individual was underpaid.

One other interesting output of any regression package is the R^2 value. This has a nice interpretation: it tells us the proportion of the variation in y that can be explained by the variation x. In the example above, $R^2 = 0.165$ so about 16.5% of the variation in *wage* can be explained by variation in *educ*.

Multiple linear regression

We can expand our model to have multiple x variables: for example, including both *educ* and *exper* (years of experience). This is easy enough to do in Gretl or STATA: we just include another variable and run the regression. For the wage data set, the (Gretl) output is the following:

```
Model 2: OLS, using observations 1-526
Dependent variable: wage

             coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  const      -3.39054      0.766566     -4.423    1.18e-05 ***
  educ        0.644272     0.0538061    11.97     2.28e-29 ***
  exper       0.0700954    0.0109776     6.385    3.78e-10 ***

Mean dependent var   5.896103   S.D. dependent var   3.693086
Sum squared resid    5548.160   S.E. of regression   3.257044
R-squared            0.225162   Adjusted R-squared   0.222199
F(2, 523)            75.98998   P-value(F)           1.07e-29
Log-likelihood      -1365.969   Akaike criterion     2737.937
Schwarz criterion    2750.733   Hannan-Quinn         2742.948
```

The first thing to note is that the R^2 value has increased significantly: including both variables lets us explain 22.5% of the variation in *wage*.

The only real difference in this regression is that we can't depict it as a straight line anymore. Our results are actually a *plane* in the three-dimensional space of *wage*, *educ*, and *exper*.

The interpretation of the coefficients is now slightly different than the regression line case: they are now the slopes (or the effect of a change of 1 in one of the right-hand side variables) *holding all other values constant*. Thus the coefficient of $\widehat{\beta}_2 = 0.644272$ tells us that someone with the same years of experience but one extra year of education will have a wage that is, on average, 64 cents higher.

For the purposes of hypothesis testing on any of the β_i values, and finding confidence intervals for the β_i values, very little has changed: the calculations are exactly the same as in one variable case described above. There is one minor difference, however: since we are now estimating three coefficients, the degrees of freedom has changed from $n-2$ to $n-3$. Since $n = 526$ here, that is a very small difference indeed.

Finding fitted values works just like before as well, just with one more term. For example, the expected wage for someone who has 12 years of education and 5 years of experience is:

$$\begin{aligned}
\widehat{y} &= -3.39054 + 0.644272(educ) + 0.0700954(exper) \\
&= -3.39054 + 0.644272(12) + 0.0700954(5) \\
&= 4.69
\end{aligned}$$

Calculating the residual is no different than in the one-variable case.

Dummy variables

We can also expand our model to categorical data by adding a “dummy” variable: a variable that takes on the value 0 or 1 depending on whether the observation belongs to a particular category or group. Very commonly we use a gender variable as a dummy to allow for differences between men and women.

For example, the “wage” data set includes several such dummies, including $female$. We can run a regression by including both our other variables and the $female$ dummy:

```
Model 3: OLS, using observations 1-526
Dependent variable: wage

             coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  const        0.622817    0.672533      0.9261   0.3548
  educ         0.506452    0.0503906    10.05     7.56e-22 ***
  female      -2.27336     0.279044     -8.147    2.76e-15 ***

Mean dependent var    5.896103   S.D. dependent var   3.693086
Sum squared resid     5307.161   S.E. of regression   3.185520
R-squared             0.258819   Adjusted R-squared   0.255985
F(2, 523)             91.31542   P-value(F)           9.66e-35
Log-likelihood       -1354.289   Akaike criterion     2714.578
Schwarz criterion     2727.374   Hannan-Quinn         2719.588
```

This is still a multiple linear regression, so our interpretation of variables still requires the *holding other variables constant* caveat, but now our "female" coefficient allows the regression line to be higher for women than men.

Both fitted regression lines still have the same *slope* (0.506452), but our best fit regression line for women is \$2.27 lower than the line for men, at any given level of education (in other words, holding education constant).

Summary

The important things to know (i.e. for studying for the exam) for dealing with regressions are:

- Interpreting coefficients
- Interpreting R^2
- Conduct a t-test for a hypothesis involving one of the β_i values, given $\widehat{\beta}_i$ and $SE(\widehat{\beta}_i)$
- Knowing the degrees of freedom for that t test
- Construct a confidence interval for one of the β values, given $\widehat{\beta}_i$ and $SE(\widehat{\beta}_i)$
- Calculating a fitted value given the $\widehat{\beta}_i$ estimates
- Calculating a residual

We also discussed in class a few tricks we can use such as adding squared terms or logarithms to attempt to match some common non-linear relationships. While these are helpful tricks for practically using regressions to explain data, they do complicated the interpretation of coefficients, and aren't something I expect you to know how to deal with on the Economics 250 exam. They await you in Economics 351.