1. Categorical variables

A categorical variable is a variable that can take on one of a limited number of values, assigning that observation to a particular group or category based on a qualitative property. For example, *student* could be a simple categorical variable with two categories: student and non-student. If we assign "student" a value of 0 and "non-student" a value of 1, we can directly include *student* in our regression.

Suppose you wanted to estimate the price of a car in dollars, and one of the variables that you had in your dataset was on the color of the car (*color*). *color* can take on values "red", "green", "blue", and "gray"—all of the car colors you observe in your data. In order to run a regression, one might be tempted to estimate something like the following:

$$price = \beta_0 + \beta_1 color + \beta_2 mileage + \beta_3 volume + u$$

where *mileage* is the car's mileage, *volume* is the cubic foot volume of the car, and *color* takes on values 1, 2, 3, or 4 for whether the car is red, green, blue, or gray, respectively. The problem in this set up is that $\hat{\beta}_1$ does not make sense for interpretation: it is not clear what $\hat{\beta}_1$=34.49 would mean, for example. With a categorical variable, the category numbers are just representations of the different categories, with no ordinal meaning. There is no "1 unit increase" in a categorical variable.

In order to have this categorical variable make sense, we can use lessons learned from modules covering dummy (or indicator, or 0-1) variables. We create $g - 1$ dummy variables, where g is the number of values that the categorical variable takes on. This is important: if we develop dummy variables for all values of the categorical variable and include them in our regression model, then we will have perfect collinearity—one of your dummies can be predicted perfectly with a combination of all the other category dummies. If we create a dummy variable for *green*, then the variable for an observation takes on value 1 if the car color is green, and 0 otherwise. A similar logical holds for *blue* and *gray*. Afterwards, we can design the following regression:

$$price = \beta_0 + \delta_0 blue + \delta_1 green + \delta_2 gray + \beta_1 mileage + \beta_2 volume + u$$

Question: How do we interpret the δ parameters?

Answer: It is important to know the **reference group**, i.e. the group that has been left out. In the above regression, we do not have a dummy variable for red cars, which should tell us that our reference group are red cars. Thus:

1. δ_0 : the average difference in price between red and blue cars, holding mileage and volume constant.
2. δ_1 : the average difference in price between red and green cars, holding mileage and volume constant.
3. δ_2 : the average difference in price between red and gray cars, holding mileage and volume constant.

Note also that with categorical variables in a regression, the intercept term provides information about the reference group. In this case, β_0 gives the average price for red cars with 0 mileage and 0 volume. We can interpret the δ coefficients as intercept shifters for different car colors. This is similar to what we saw when interpreting regression coefficients with a single dummy variable.

Given that we translated our categorical variable into dummy variables, we are now able to do the same sort of analysis that we had done previously, including the use of interactions.

2. Bad controls

(Note: this section is adapted from a blog post on bad controls:
http://www.g-feed.com/2012/10/bad-control.html)

Often, a researcher will be concerned about omitted variable bias (where the omitted variable is related to one of the independent variables in the regression model as well as the dependent variable), and so will want to control for potential omitted variables, assuming one is able to get data for them. However, an omitted variable may not always be a desirable variable to include as a control in the regression model, particularly when the variable considered for inclusion is itself affected by the independent variable of interest.

The author in the above post, Marshall Burke, considers the case of the effect of temperature on conflict. Temperature is likely to affect conflict, but is probably correlated with many other things that are also likely to affect conflict, such as per capita GDP levels. Thus, one regresses conflict (however this is measured) on temperature and per capita GDP, and finds that the estimated β associated with temperature is not statistically significant at conventional levels, while the effect of GDP is large and significant. While it is tempting to conclude that the effect of temperature is not robust to controlling for GDP, one can see that temperature *also affects* economic productivity, such that per capita GDP is really an outcome variable. Thus, it doesn't make sense to "hold economic productivity constant" when estimating the relationship between temperature and conflict, when at least some part of temperature's effect is through income—we can't hold GDP fully constant when considering an effect of temperature if changing temperature also changes GDP.

What can one do in the face of such "*bad controls*"? Burke provides two options: one option is to show the reduced form relationship between temperature and conflict without any controls, which can tell you as much as you'll likely want to know if temperature is as good as randomly assigned. The second option is to just be clear about the (lack of) relationships between variables in your model and argue that these variables do not present concerns. Still, it could be useful to include per capita GDP in a regression model if one wants to explore how temperature might affect conflict through some channel other than per capita GDP.

In general, it is important to think about how your independent variables might be related to each other, particularly if one variable causally affects another. If one of your independent variables affects your outcome variable partly through its effect on a second independent variable, your estimated β for that first variable will not capture its full impact on the outcome.

3. Confidence Intervals for Predictions

You've already had to "predict" a value of the dependent variable, y, given certain values of the independent variables. Multiply each variable's value by the estimated β for that variable, and take the sum of those values plus the intercept to generate the estimated y. But this prediction is just a guess (since our β estimates are uncertain), and we can construct a confidence interval to give a range of possible values for this prediction, to demonstrate this uncertainty. There are two kinds of predictions we can make:

- **A confidence interval for the average y given $x_1, ...x_k$**
- **A confidence interval for a particular y given $x_1, ..., x_k$**

We will use Wooldridge's birth weight data to construct both kinds of confidence intervals to demonstrate the (subtle) difference between them.

$$bwght = \beta_0 + \beta_1 lfaminc + \beta_2 motheduc + \beta_3 parity + u$$

where *bwght* is birthweight in ounces, $lfaminc$ is the log of family income in $1000s, *motheduc* is the education of the mother in years, and *parity* is the birth order of the child.

Estimating this equation in R, we get the following results:

```
Call:
lm(formula = bwght ~ lfaminc + motheduc + parity, data = bwght)

Residuals:
    Min      1Q  Median      3Q     Max
-94.533 -11.888   0.779  13.136 151.477

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 105.5652     3.3666  31.356  < 2e-16 ***
lfaminc       2.1313     0.6506   3.276  0.00108 **
motheduc      0.3172     0.2520   1.259  0.20829
parity        1.5261     0.6119   2.494  0.01275 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.21 on 1383 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.01633,   Adjusted R-squared:  0.0142
F-statistic: 7.654 on 3 and 1383 DF,  p-value: 4.482e-05
```

A. Confidence interval for average birthweight

Recall that our model gives us the expected value[1]:

$$\mathbb{E}[\text{bwght} \mid \text{lfaminc, motheduc, parity}] = \beta_0 + \beta_1 ln(faminc) + \beta_2 motheduc + \beta_3 parity$$

Our regression gives us an estimate of this:

$$\widehat{\mathbb{E}}[\text{bwght} \mid \text{faminc, motheduc, parity}] = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 ln(faminc) + \hat{\beta}_2 motheduc + \hat{\beta}_3 parity$$

In words: when we plug in particular values of the independent variables, we obtain a prediction for y, which is an estimate for the expected value of y given the particular values for the explanatory variables.

Say we're interested in a confidence interval for the average birthweight (or expected birthweight) for babies with family income of \$14,500 ($ln(14.5) = 2.674$), mothers with 12 years of education, and with 2 older siblings ($parity = 3$). In other words, we're interested in:

$$\begin{aligned}\widehat{\mathbb{E}}[\text{bwght} \mid \text{faminc=14.5, motheduc=12, parity=3}] &= 105.57 + 2.13 ln(faminc) + 0.317 motheduc + 1.53 parity \\ \hat{y}_{faminc=14.5, motheduc=12, parity=3} &= 105.57 + 2.13(2.674) + .317(12) + 1.53(3) \\ &= 119.66 \text{ ounces}\end{aligned}$$

How do we find a standard error for our estimate of the expected value of y for these particular values of the explanatory variables? Calculating this standard error is complicated because $\widehat{bwght}$ is a function of our $\hat{\beta}$s, which are all random variables. To avoid this computation, we want to **transform our data**. Before proceeding with the formal steps, recall that we have the following regression in mind:

$$bwght = \beta_0 + \beta_1 lfaminc + \beta_2 motheduc + \beta_3 parity$$

Then, recalling what the intercept represents, we can say that

$$\hat{\beta}_0 = \hat{E}(bwght | lfaminc = 0, motheduc = 0, parity = 0)$$

If we modify the regression by subtracting our particular values (specified above) for each of the independent variables, then we get the following regression

$$bwght = \beta_0 + \beta_1 (lfaminc - 2.674) + \beta_2 (motheduc - 12) + \beta_3 (parity - 3)$$

Then under this transformation, the definition of the intercept gives us

$$\hat{\beta}_0 = \hat{E}(bwght | lfminc = 2.674, meduc = 12, parity = 3)$$

In other words, the new intercept is the predicted birthweight for babies with family income of \$14,500 ($ln(14.5) = 2.674$), mothers with 12 years of education, and with 2 older siblings ($parity = 3$), since in the transformed model it is for these values that the independent variables take a value of 0. That's perfect! If we run a regression, R always outputs a standard error for the intercept coefficient. We can then take the standard error for our average predicted value from there.

Steps in computing a confidence interval for the predicted average y when $x_j = \alpha_j$ (when our explanatory variables are equal to some particular values α):

[1] The error term u, which we do not measure, is what takes us from the expected value under the model to the actual observations y.

1. Generate new variables: $\tilde{x}_j = x_j - \alpha_j$.
2. Run the regression of: $y = \tilde{\beta}_0 + \tilde{\beta}_1\tilde{x}_1 + ... + \tilde{\beta}_k\tilde{x}_k$
3. Then $\hat{\mathbb{E}}[y|x_1 = \alpha_1, ..., x_k = \alpha_k] = \tilde{\beta}_0$ and the standard error for this estimate is $SE(\tilde{\beta}_0)$.
4. Plug these values into the formula for confidence intervals and interpret.

Below is the output from Step 2 (what variables did I have to create in R?):

```
Call:
lm(formula = bwght ~ lfaminc_0 + motheduc_0 + parity_0, data = bwght)

Residuals:
    Min      1Q  Median      3Q     Max
-94.533 -11.888   0.779  13.136 151.477

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 119.6491     1.0066 118.864  < 2e-16 ***
lfaminc_0     2.1313     0.6506   3.276  0.00108 **
motheduc_0    0.3172     0.2520   1.259  0.20829
parity_0      1.5261     0.6119   2.494  0.01275 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.21 on 1383 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.01633,   Adjusted R-squared:  0.0142
F-statistic: 7.654 on 3 and 1383 DF,  p-value: 4.482e-05
```

Using this output, the 95% confidence interval for the average birthweight for babies given family income of \$14,500 ($ln(14.5) = 2.674$), mothers with 12 years of education, and with 2 older siblings ($parity = 3$) is:

$$\big[119.65 - 1.96(1.007), 119.65 + 1.96(1.007)\big] = \big[117.68, 121.62\big]$$

B. Confidence interval for a particular birthweight

I am taking from Woolridge p. 208:

"The previous method allows us to put a confidence interval around the OLS estimate of $E[y|x_1, \cdots, x_k]$ for any values of the explanatory variables. In other words, we obtain a confidence interval for the *average* value of y for the subpopulation with a given set of covariates. But a confidence interval for the average person in the subpopulation is not the same as a confidence interval for a particular unit (individual, family, firm, and so on) from the population. In forming

a **confidence interval for a particular unit**, we must account for another very important source of variation: the **variance in the unobserved error**, which measures our ignorance of the unobserved factors that affect y".

Let $bwght^0$ denote the value for which we would like to construct a confidence interval (representing the value for the observation labeled 0 in the data):

$$bwght^0 = \beta_0 + \beta_1 lfaminc^0 + \beta_2 meduc^0 + \beta_3 parity^0 + u^0$$

Our best prediction of $bwght^0$ is $\widehat{bwght^0}$, where

$$\widehat{bwght}^0 = \hat{\beta}_0 + \hat{\beta}_1 lfaminc^0 + \hat{\beta}_2 meduc^0 + \hat{\beta}_3 parity^0$$

Now there is some error associated with using $\widehat{bwght}^0$ to predict $bwght^0$ (think back to the picture we drew in Section 2 notes and think about the distance between our actual value of y and the predicted value of $\hat{y}$ representing the residuals, denoted by $\hat{u}_i$ in those notes). Then

$$\begin{aligned}\hat{u}^0 = bwght^0 - \widehat{bwght}^0 &= \beta_0 + \beta_1 lfaminc^0 + \beta_2 meduc^0 + \beta_3 parity^0 + u^0 \\ &\quad - (\hat{\beta}_0 + \hat{\beta}_1 lfaminc + \hat{\beta}_2 meduc + \hat{\beta}_3 parity)\end{aligned}$$

Taking the expected value and remembering our regression assumptions we get:

$$\begin{aligned}E[\hat{u}^0] = E[bwght^0 - \widehat{bwght}^0] &= \beta_0 + \beta_1 lfaminc^0 + \beta_2 meduc^0 + \beta_3 parity^0 + E[u^0] \\ &\quad - (E[\hat{\beta}_0] + E[\hat{\beta}_1] lfaminc + E[\hat{\beta}_2] meduc + E[\hat{\beta}_3] parity) \\ &= 0\end{aligned}$$

Then for the variance (recalling rules about variance operations):

$$\begin{aligned}Var(\hat{u}^0) &= Var(bwght^0 - \widehat{bwght}^0) = Var(\beta_0 + \beta_1 lfaminc^0 + \beta_2 meduc^0 + \beta_3 parity^0 + u^0 - \widehat{bwght}^0) \\ &= Var(u^0 - \widehat{bwght}^0) = Var(\widehat{bwght}^0) + Var(u^0) \\ &= Var(\widehat{bwght}^0) + \sigma^2 \\ \widehat{Var(\hat{u}^0)} &= Var(\widehat{bwght}^0) + \hat{\sigma}^2 \\ &= Var(\widehat{bwght}^0) + \frac{\sum \hat{u}_i^2}{n-k-1} = Var(\widehat{bwght}^0) + \frac{SSR}{n-k-1}\end{aligned}$$

So you should see that there are two sources of variation in $\hat{u}^0$. First we have the sampling error in $\widehat{bwght}^0$ which arises because we have estimated the population parameters (β). Second, we have the variance of the error in the population u^0—since we do not measure this variance σ^2, we estimate it with $\hat{\sigma}^2$.

Now we can compute the $Var(\widehat{bwght}^0)$ exactly the way we did before by subtracting the specific values we are interested in, and re-running the regression, and looking at SE for the intercept term (recall that the SE is the square root of the estimator for the variance). We can also compute

$\hat{\sigma}^2 = \frac{SSR}{n-k-1}$ from our regression output. We can now calculate $se(\hat{u}^0) = \sqrt{Var(\widehat{bwght}^0) + \hat{\sigma}^2}$. Then the 95% confidence interval for $bwght^0$ is

$$\widehat{bwght}^0 \pm 1.96 \cdot se(\hat{u}^0)$$

Steps in computing a confidence interval for a particular y when $x_j = \alpha_j$:

1. Generate new variables: $\tilde{x}_j = x_j - \alpha_j$.
2. Run the regression of: $y = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{x}_1 + ... + \tilde{\beta}_k \tilde{x}_k + \tilde{u}$
3. Then $\hat{\mathbb{E}}[y|x_1 = \alpha_1, ..., x_k = \alpha_k] = \tilde{\beta}_0$ and the standard error for this estimate is $SE(\tilde{\beta}_0)$.
4. Get an estimate for the variance of $\hat{u} = \hat{\sigma}^2$ from the R output (see code below). This can be done by calling the "sigma" object from your regression summary.
5. Compute the standard error: $\sqrt{SE(\tilde{\beta}_0)^2 + \hat{\sigma}^2}$.
6. Plug these values into the formula for confidence intervals and interpret.

Let's first get our sigma (squared) from our regression results:

```
> summary(lm(bwght~ lfaminc_0+ motheduc_0 + parity_0, data=bwght))$sigma^2
[1] 408.5987
```

Using the same output as above, the confidence interval for a particular baby's birthweight with given family income of \$14,500 ($ln(14.5) = 2.674$), mothers with 12 years of education, and with 2 older siblings ($parity = 3$) is:

$$SE = \sqrt{SE\left(\hat{\beta}_0\right)^2 + \hat{\sigma}^2} = \sqrt{(1.007^2) + 408.59} = 20.239.$$

$$\begin{aligned} &\left[119.64 - 1.96(20.239) \quad , \quad 119.64 + 1.96(20.239)\right] \\ &\left[79.972 \quad , \quad 159.308\right] \end{aligned}$$

Observe that this is a much wider confidence interval that for our predicated average y for particular values of our x variables. This reflects the additional source of uncertainty we have when making a prediction about a particular observation. That uncertainty doesn't affect us when considering predicted average y because of our assumption that u is mean 0.

4. Chow Test

A Chow test is a shorter way to perform a specific type of F-test. We use a **Chow test to check if our regression parameters are different between two different groups**. In Section 7, we saw that interacting dummy variables with a) other dummies and b) continuous variables allows us to test whether different groups have different intercepts and different slopes, respectively. We may also wish to test the null that two groups follow the same regression function, against the alternative that one or more of the slope or intercept parameters differs across groups.

We will work through an example that uses sleep75 data set from the Woodridge text that we are already familiar with. Suppose we are interested in seeing whether age and total hours worked affect time slept (in minutes per week). Given this, we're interested in the following regression:

$$sleep = \beta_0 + \beta_2 age + \beta_4 totwrk + u$$

You suspect that the relationship between *sleep* and *age* and *totwrk* is different if you have young kids vs. if you do not have young kids (defined in the data as kids under the age of 3), as individuals with young kids have more time-consuming childcare responsibilities.

Question: What regression would you run as the unrestricted model to test your hypothesis that people with vs. without young kids get different amounts of sleep?

Answer: We can rewrite restricted and unrestricted regressions as:

$$\begin{aligned} \text{Unrestricted}: sleep &= \beta_0 + \beta_1 yngkids + \beta_2 age + \beta_3 yngkids * age \\ &\quad + \beta_4 totwrk + \beta_5 yngkids * totwork + e \\ \text{Restricted}: sleep &= \beta_0 + \beta_2 age + \beta_4 totwrk + e \end{aligned}$$

Before we proceed, let's practice interpreting the coefficients in the unrestricted model. Write down the interpretation of each coefficient below (answers on next page):

- β_1:
- β_2:
- β_3:
- β_4:
- β_5:

Question 1: What is the average number of minutes slept per week for individuals who are 50 years old, work 40 hours per week, and do not have young kids? (answer below)

Question 2: What is the average number of minutes slept per week for individuals who are 30 years old, work 45 hours per week, and have a young child? (answer below)

- β_1: the average difference in minutes of sleep for people with young kids relative to those without young kids, holding constant age and total hours worked.[2]
- β_2: being one year older is associated with a β_2 change in minutes of sleep, holding constant total hours worked and whether the respondent has young kids.
- β_3: the difference in the change in minutes slept associated with being one year older for people who have young kids vs those who don't, holding total hours worked constant.
- β_4: working one more hour per week is associated with a β_4 change in minutes of sleep per week, holding constant age and whether the respondent has young kids.
- β_5: the difference in the change in minutes slept associated with working one more hour per week for people who have young kids relative to those who don't, holding age constant.

Answer 1: $\beta_0 + \beta_2(50) + \beta_4(40)$. For these individuals, $yngkids = 0$.

Answer 2: $\beta_0 + \beta_1 + \beta_2(30) + \beta_3(30) + \beta_4(45) + \beta_5(45)$.

Now back to our hypothesis test! There are two ways you could formally test this hypothesis:

A. F-test

If you suspect that this whole regression might be different if we ran it for only people with young kids, that's equivalent to saying that each of the βs is different depending on whether the respondent has young kids.

The F-test that will tell us whether there is a significant difference between these two models:

$$H_0 : \beta_1, \beta_3, \beta_5 = 0$$
$$H_1 : \text{not } H_0$$

And our F stat would be:

$$F = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(n - k_{UR} - 1)}$$

B. Chow-Test

The Chow test is just a way to complete that F-test *without running the UR regression with all of those pesky interactions*. What Chow realized is that we can get everything we need to compute the F-stat from running the following regressions for different subsamples:

$$\text{(A) } sleep = \beta_0 + \beta_1 age + \beta_2 totwrk + u \quad \textit{Have young kids only}$$
$$\text{(B) } sleep = \beta_0 + \beta_1 age + \beta_2 totwork + u \quad \textit{Do not have young kids only}$$

[2] Any time we include a variable in an interaction in our model, we also want to include it on its own. This is important to be able to correctly interpret the coefficient on the interaction term.

Chow noticed that:

1. $SSR_{UR} = SSR_A + SSR_B$
2. $q = k + 1$ the hypothesis that each beta is the same across the two groups involves $k + 1$ restrictions, since we are testing whether the intercept and each slope term may differ by group.
3. The unrestricted model, which we can think of as having a group dummy variable and k interaction terms in addition to the intercept and variables themselves, has $n - 2(k + 1)$ degrees of freedom

We can use these three facts to rewrite our F-statistic in a way so that we only need to run (1) Restricted Model, (2) Model A and (3) Model B instead of the usual restricted and unrestricted regressions:

$$F = \frac{\Big(SSR_{pooled} - (SSR_A + SSR_B)\Big)/q}{(SSR_A + SSR_B)/(n - 2(k+1))}$$

What this means is that we can calculate the F-statistic that tests whether or not each parameter in our original model (1) is different for individuals with vs without young kids without actually running the unrestricted model. From here, the Chow test is the same as the usual F-tests.

Bottom-line: The Chow test is just an F-test for a specific situation: when you want to see if the regression is totally different (every parameter) between different groups, but you want to avoid running the unrestricted regression. This is helpful since it's usually easier to run regressions for sub-samples that to run regressions that require creating new interaction terms.

Let's try this test with data. To compute the F/Chow-statistic, in R we need to run:

- Estimate equation (1) for all people together (call it reg_restricted):

```
Call:
lm(formula = sleep ~ age + totwrk, data = sleep75)

Residuals:
     Min       1Q   Median       3Q      Max
-2384.63  -242.13     7.81   262.45  1302.19

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3469.20059   68.11787  50.929   <2e-16 ***
age            2.92388    1.39671   2.093   0.0367 *
totwrk        -0.14901    0.01672  -8.912   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 420.1 on 703 degrees of freedom
Multiple R-squared:  0.1088,    Adjusted R-squared:  0.1063
F-statistic: 42.93 on 2 and 703 DF,  p-value: < 2.2e-16
```

- Estimate equation (1) for only people with young kids (call it A_reg):

```
Call:
lm(formula = sleep ~ age + totwrk, data = sleep75[which(sleep75$yngkid ==
    1), ])

Residuals:
     Min       1Q   Median       3Q      Max
-1210.02  -258.66   -16.41   290.26  1062.36

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3153.34833  331.43332   9.514 3.58e-15 ***
age            7.45240   11.40556   0.653    0.515
totwrk        -0.05659    0.05146  -1.100    0.274
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 442.2 on 88 degrees of freedom
Multiple R-squared:  0.01544,   Adjusted R-squared:  -0.006935
F-statistic: 0.6901 on 2 and 88 DF,  p-value: 0.5042
```

- Estimate equation (1) for only people without young kids (call it B_reg):

```
Call:
lm(formula = sleep ~ age + totwrk, data = sleep75[which(sleep75$yngkid ==
    0), ])

Residuals:
     Min       1Q   Median       3Q      Max
-2376.03  -228.19     9.76   241.80  1297.21

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3498.7306    74.1757  47.168   <2e-16 ***
age            2.8570     1.4759   1.936   0.0534 .
totwrk        -0.1628     0.0177  -9.194   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 416.1 on 612 degrees of freedom
Multiple R-squared:  0.1294,    Adjusted R-squared:  0.1266
F-statistic: 45.49 on 2 and 612 DF,  p-value: < 2.2e-16
```

To compute the Chow statistic we use:

- $SSR_{pooled} = 124084606$

```
> sum(reg_restricted$residuals^2)
[1] 124084606
```

- $SSR_A = 17207661$

```
> sum(A_reg$residuals^2)
[1] 17207661
```

- $SSR_B = 105982703$

```
> sum(B_reg$residuals^2)
[1] 105982703
```

- $q = 3$ restrictions we are testing
- $k = 2$ variables in the restricted model
- $n = 706$ sample size
- $\Rightarrow F = \frac{(SSR_{pooled} - (SSR_A + SSR_B))/q}{SSR_A + SSR_B/(n-2(k+1))} = \frac{124084606 - (17207661 + 105982703)/3}{(17207661 + 105982703)/700} = 1.69378$

The critical value for $F_{3,700}$ and $\alpha = 0.05$ is about 2.63. Hence we fail to reject the critical value at the 5% significance level. This might seem surprising initially, but we notice that there are only about 90 people (out of around 700) in the dataset who have small kids. It could just be that we don't have enough observations to be picking up the effect of having small kids once we control for age and total hours worked.

5. Linear Probability Model

A. Introduction

Up until this point all the models we have seen have had continuous dependent variables: the y variable was some quantitative amount that took on a range of possible values (a test score, a percentage,...). Sometimes though **y can be a dummy variable**: it is defined to take on two values, 0 and 1. What changes now that we have a binary variable on the left hand side?

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$$

- It no longer makes a lot of sense to interpret β_j as the unit change in y given a one-unit increase in x_j holding all other factors fixed. Indeed, y either changes from $0 \rightarrow 1$, $1 \rightarrow 0$ or doesn't change.
- The β_j's still have a useful interpretation though: each β_j measures the change (an increase or decrease) in the probability that $y = 1$ when x_j changes by one unit holding all other factors fixed.
- Similarly the regression estimates $\hat{\beta}_j$ measure the *predicted* change in the probability that $y = 1$ when x_j inceases by one unit.

How to see this? Think about the following two expressions:

$$\begin{aligned} E(y|x) &= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \\ Pr(y = 1|x) &= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \\ E(y|x) &= Pr(y = 1|x) \end{aligned}$$

In words, when y is a binary variable taking on values zero and one, it is always true that $E(y|x) = Pr(y = 1|x)$: the expected value of y is the probability that $y = 1$. Then looking back to what we covered earlier in the semester:

$$\Delta P(y = 1|x) = \beta_j \Delta x_j \Longrightarrow \beta_j = \frac{\Delta P(y = 1|x)}{\Delta x_j}$$

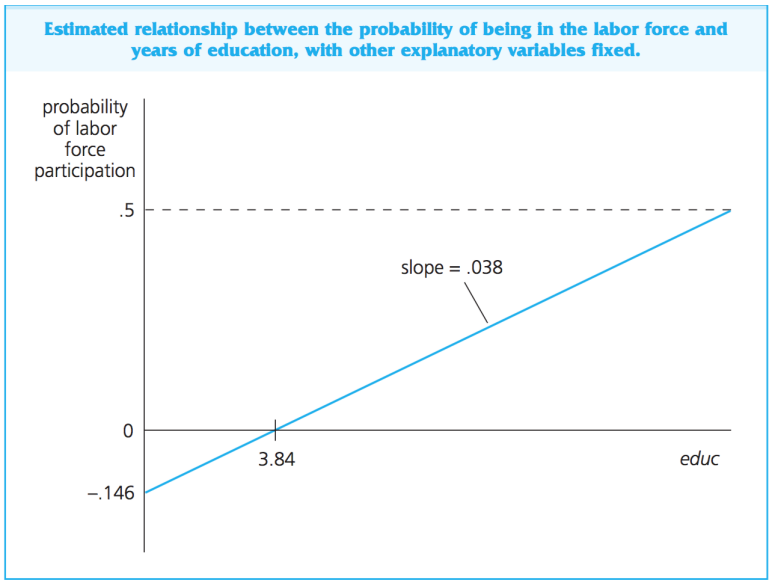
We therefore call models with a dummy outcome variable "linear probability models."

B. Example

From Wooldridge p.247: Let $inlf$ ("in the labor force") be a binary variables indicating labor force participation by a married women during 1975: $inlf = 1$ if the woman reports working for a wage outside the home at some point during the year, and zero otherwise. Our independent variables might include: husband's earnings ($nwifeinc$, measured in thousand of dollars), years of education ($educ$), past years of labor market experience ($exper$), age (age), number of children less than six years old ($kidslt6$), and number of kids between 6 and 18 years of age ($kidsge6$). Using data, we estimate the following linear probability model

$$\begin{aligned} \widehat{inlf} = 0.586 &- 0.0034 nwifeinc + 0.038 educ + 0.039 exper - 0.00060 exper^2 - 0.017 age \\ &- 0.262 kindslt6 + 0.0130 kidsge6 \end{aligned}$$

Let's start by interpreting the coefficient on education: holding all else fixed, another year of education increases the predicted probability of labor force participation by 0.038 (or 3.8%). The graph below depicts the probability of labor force participation and $educ$. The other independent variables are fixed at the values $nwifeinc = 50$, $exper = 5$, $age = 30$, $kidslt6 = 1$, and $kidsge6 = 3.84$ (we picked these for the purposes of the example, you could have picked any other values for these variables. The key is that they remain fixed as we vary education and assess the impact on the probability of participating in the labor force).



Because we are trying to fit a line to the data, we can get negative probabilities. As we can see the predicted probability is negative until education equals 3.84 years. This isn't too worrisome because in the sample not many women will only have 3.84 years of education. The marginal effect of another year of education on the probability of labor force participation is given by the slope (it's always 0.038).

C. Drawbacks

1. The predicted probabilities from our regression **aren't bounded** between zero and one. A related problem is that a "probability cannot be linearly related to the independent variables". In the previous example, an additional child reduces the probability of working by 0.262. It follows that going from zero to four children reduces the probability by $0.262 * 4 = 1.048$, which is impossible.
2. This model uses the idea that the probability that the dummy is equal to one is actually a function of our x's, which means the variance of the dummy is a function of our x's. Indeed, when y is a binary variable

   $$Var(y|x) = p(x)[1 - p(x)]$$

   where $p(x)$ is a shorthand for the probability that $y = 1$, $p(x) = \beta_0 + \beta_1 x + \cdots \beta_k x_k$. This means that there must be heteroskedasticity in the linear probability model. This violates our assumption of homoskedasticity:

   $$Var(u|x) = Var(u) = \sigma^2$$

Several other models are applied when the outcome variable is a dummy, or indeed for continuous variables that are bounded in some way. Logit, probit, and tobit regressions are some examples that help to deal with above drawbacks of using OLS with a dummy outcome variables. Nevertheless, despite these limitations OLS is commonly used in such situations because of the ease of interpretation and because the estimated results are usually qualitatively quite similar to what is obtained using other approaches.