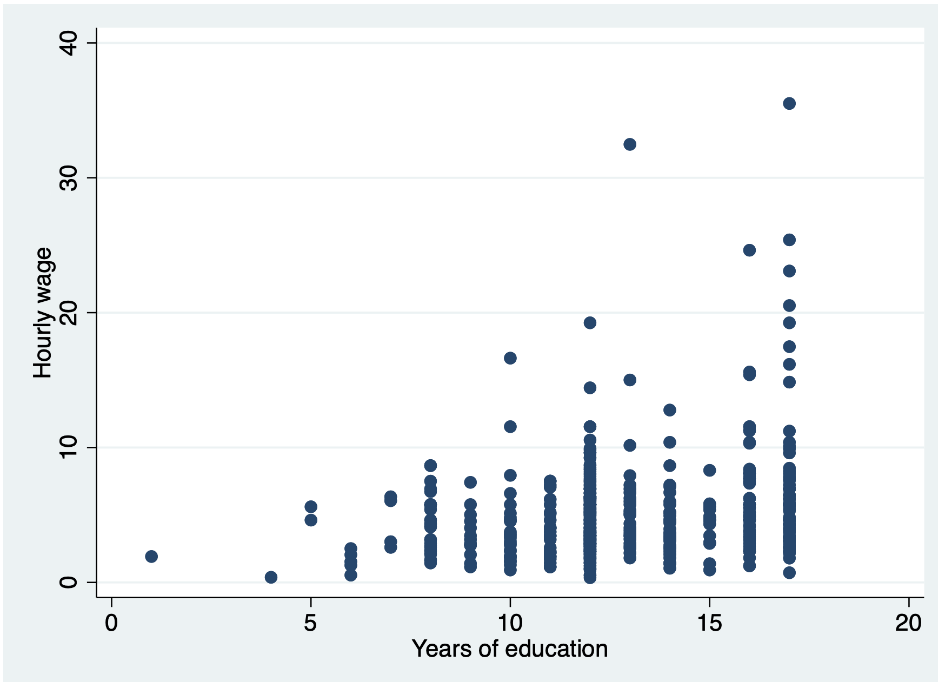
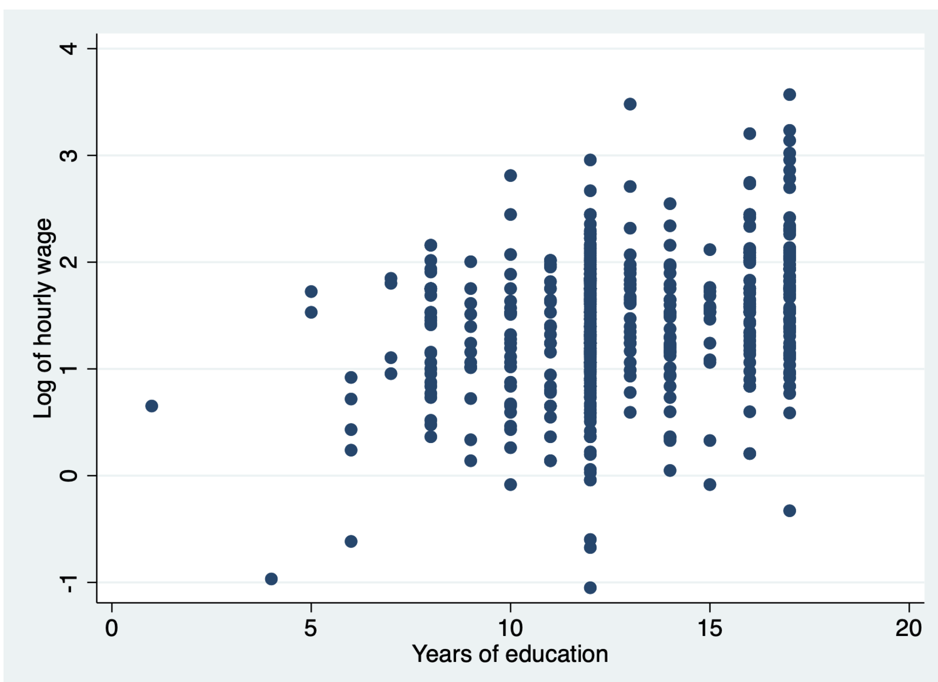
Dis 5: Nonlinear Regression; Dummy and Interaction

1 Log transformation of variables

- Some variables don't seem to grow linearly ...



- ... unless they've been transformed in some way



- But transforming variables alters their interpretation:
  - Consider an estimated line: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
    Here, $\hat{\beta}_1$ can be interpretated as rate of change from x into y. In other words, $\hat{\beta}_1$ reflects how much change of x is estimated to reflect on change in y:
$$\frac{\partial \hat{y}_i}{\partial x_i} = \hat{\beta}_1$$
  - Suppose that we transform both y and x by taking the log: $\ln \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \ln x_i$
    Let's take a similar approach by taking the derivative of $\ln \hat{y}_i$ with respect to $\ln x_i$:
$$\frac{\partial \ln \hat{y}_i}{\partial \ln x_i} = \hat{\beta}_1 \qquad (*)$$
    But ideally, we'd like to know how change in x directly reflects change in y. In order to do so, notice that
$$\frac{\partial \ln z}{\partial z} = \frac{1}{z} \quad \Rightarrow \quad \partial \ln z = \frac{\partial z}{z}$$
    This means that equation $(*)$ can be expressed as
$$\frac{\partial \ln \hat{y}_i}{\partial \ln x_i} = \underbrace{\frac{\partial \hat{y}_i / \hat{y}_i}{\partial x_i / x_i}}_{\text{elasticity}} = \hat{\beta}_1 \quad \Rightarrow \quad \frac{\%\Delta \hat{y}_i / 100}{\%\Delta x_i / 100} = \frac{\%\Delta \hat{y}_i}{\%\Delta x_i} = \hat{\beta}_1$$
    This means that when x increases by 1%, y is predicted to change by $\hat{\beta}_1$ percent.
  - Suppose that we only transform y by taking the log: $\ln \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
    In this case,
$$\frac{\partial \ln \hat{y}_i}{\partial x_i} = \frac{\partial \hat{y}_i / \hat{y}_i}{\partial x_i} = \hat{\beta}_1 \quad \Rightarrow \quad \frac{\%\Delta \hat{y}_i / 100}{\partial x_i} = \hat{\beta}_1$$
$$\frac{\%\Delta \hat{y}_i}{\partial x_i} = \hat{\beta}_1 \times 100$$
    This means that when x increases by 1 unit, y is predicted to change by $\hat{\beta}_1 \times 100$ percent.
  - Similar exercise can be done for only transforming x by taking its log. To summarize:

Model	**Regressand**	**Regressor**	**Interpretation of β_1**
Level-Level (Linear-Linear)	y	x	$\beta_1 = \frac{\Delta y}{\Delta x}$
Log-Log	$\ln y$	$\ln x$	$\beta_1 = \frac{\%\Delta y}{\%\Delta x}$
Log-Level (Log-Linear)	$\ln y$	x	$\beta_1 \times 100 = \frac{\%\Delta y}{\Delta x}$
Level-Log (Linear-Log)	y	$\ln x$	$\frac{\beta_1}{100} = \frac{\Delta y}{\%\Delta x}$

2 Dummy variables and interaction terms

- **Dummy variables**: Variables that are binary (record only 0 or 1).

  Ex. A variable recording sex (female = 1 if the observation is a female; female = 0 if the observation is a male)

  Ex. A variable recording the enactment of a policy (= 1 if the policy is in effect; = 0 if not)

  - Consider the following regression model:

$$\text{wage}_i = \beta_0 + \beta_1 \text{female}_i + u_i$$

  - What's the expected wage for male and female?
    * For male:

$$\begin{aligned} E[\text{wage}|\text{female} = 0] &= E[\beta_0 + \beta_1 \text{female}_i + u_i|\text{female} = 0] \\ &= \beta_0 + \beta_1 E[\text{female}_i|\text{female} = 0] + E[u_i|\text{female} = 0] \\ &= \beta_0 \end{aligned}$$

    * For female:

$$\begin{aligned} E[\text{wage}|\text{female} = 1] &= E[\beta_0 + \beta_1 \text{female}_i + u_i|\text{female} = 1] \\ &= \beta_0 + \beta_1 E[\text{female}_i|\text{female} = 1] + E[u_i|\text{female} = 1] \\ &= \beta_0 + \beta_1 \end{aligned}$$

  - What does this tell us about the coefficient interpretation?
    * β_0: Expected (average) wage for male.
      (i.e. Intercept of the model for male observations)
    * $\beta_0 + \beta_1$: Expected wage for female.
      (i.e. Intercept of the model for female observations)
    * β_1: Change in expected wage due to the observation being female.
  - **Dummy variable trap**

    Can you include both a female and a male dummy variable into the wage regression model?
    → **No, because of perfect colinearity**: male + female = 1

Recall why perfect colinarity is an issue. Say that we include both male and female dummies:

$$\begin{aligned} \text{wage}_i &= \beta_0 + \beta_1 \text{female}_i + \beta_2 \text{male}_i + u_i \\ &= \beta_0 + \beta_1 \text{female}_i + \beta_2 (1 - \text{female}_i) + u_i \\ &= (\beta_0 + \beta_2) + (\beta_1 - \beta_2) \text{female}_i + u_i \end{aligned}$$

This is equivalent to running

$$\text{wage}_i = \gamma_0 + \gamma_1 \text{female}_i + u_i$$

where

$$\begin{cases} \gamma_0 = \beta_0 + \beta_2 \\ \gamma_1 = \beta_1 - \beta_2 \end{cases}$$

However, this gives us two equations with three unknown βs, so the βs are not uniquely identified, which is why we cannot include variables that are perfectly colinear.

- **Interaction terms**: Products of two (or more) variables, when it's usually one (or more) is a dummy variable.

  Ex. female $\times$ educ (= 0 if observation is male; = educ if observation is female)

  Ex. policy_in_place $\times$ first_year ($= 0$ if policy is not in place, or the observation is not in the first year of the policy; $= 1$ if this is the first year that a policy is in place)

  - Consider the following regression model:

  $$\text{wage}_i = \beta_0 + \beta_1 \text{female}_i + \beta_2 \text{educ}_i + \beta_3 \text{female}_i \times \text{educ}_i + u_i$$

  - What's the change in wage with respect to change in years of education for male and female?
    * In general:

    $$\frac{\partial \text{wage}_i}{\partial \text{educ}_i} = \beta_2 + \beta_3 \text{female}_i$$

    * For male:

    $$\left.\frac{\partial \text{wage}_i}{\partial \text{educ}_i}\right|_{\text{female}=0} = \beta_2$$

    * For female:

    $$\left.\frac{\partial \text{wage}_i}{\partial \text{educ}_i}\right|_{\text{female}=1} = \beta_2 + \beta_3$$

  - What does this tell us about the coefficient interpretation?
    * β_2: *For male*, increase in one year of education is correlated with β_2 unit increase in wage.
    * $\beta_2 + \beta_3$: *For female*, increase in one year of education is correlated with $\beta_2 + \beta_3$ unit increase in wage.
    * β_3: Change in effect of education on wage due to the observation being female.

- To summarize:
  - Include dummy variable in your regression model changes intercept
  - Include interaction term in your regression model changes slope
  - Beware of dummy variable trap (for including either just dummy variable or interaction terms)

- Do things in Stata:
  - If x_1 is categorical (say, x_1 records three categories: "low", "medium", "high"), and you want to include all possible dummies, attach `i.` in front of the variable name when running regression:

```
. reg y i.x1

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(2, 27)        =      2.51
       Model |  131.608198         2  65.8040989   Prob > F        =    0.1000
    Residual |  707.653795        27  26.2093998   R-squared       =    0.1568
-------------+----------------------------------   Adj R-squared   =    0.0944
       Total |  839.261993        29  28.9400687   Root MSE        =    5.1195

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |
     medium  |   5.648803   2.522923     2.24   0.034     .4721923    10.82541
       high  |   3.150673   2.400064     1.31   0.200    -1.773852    8.075197
             |
       _cons |   7.499662   1.934994     3.88   0.001     3.529384    11.46994
------------------------------------------------------------------------------
```

    (Stata is smart enough to avoid include all three dummies to avoid perfect colinarity issue.)

  - If x_1 is categorical, x_2 is a continuous variable, and you want to include the interaction term $x_1 \times x_2$, use # to indicate multiplicative product, and attach `c.` in front of the continuous variable:

```
. reg y i.x1 i.x1#c.x2

      Source |       SS           df       MS      Number of obs   =        30
-------------+----------------------------------   F(5, 24)        =      1.29
       Model |  177.610807         5  35.5221615   Prob > F        =    0.3015
    Residual |  661.651185        24  27.5687994   R-squared       =    0.2116
-------------+----------------------------------   Adj R-squared   =    0.0474
       Total |  839.261993        29  28.9400687   Root MSE        =    5.2506

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |
     medium  |   12.80461   28.63537     0.45   0.659     -46.2959    71.90512
       high  |   29.46032   26.36466     1.12   0.275    -24.95367    83.87431
             |
     x1#c.x2 |
        low  |   3.483692   3.609395     0.97   0.344    -3.965734    10.93312
     medium  |   2.384018   3.045411     0.78   0.441    -3.901402    8.669438
       high  |  -.8345443   2.367309    -0.35   0.728     -5.72043    4.051341
             |
       _cons |  -13.79825   22.15547    -0.62   0.539    -59.52489     31.9284
------------------------------------------------------------------------------
```

  - Alternatively, you could also just generate interaction terms on your own. Say x_3 is a dummy variable, x_4 is another variable, you can generate the interaction term between x_3 and x_4 (call it x3x4) by running

    `gen x3x4 = x3 * x4`

    You can then include `x3x4` as a variable in your `regress` command.

3 Problems

1. Load the dataset from http://fmwww.bc.edu/ec-p/data/wooldridge/wage2.dta into Stata (don't forget to first change your working directory).

   Dataset codebook is available at http://fmwww.bc.edu/ec-p/data/wooldridge/wage2.des

   (a) Start off by estimating the following regression model:

$$\text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + u_i$$

Running the following commands in Stata:

```
use "http://fmwww.bc.edu/ec-p/data/wooldridge/wage2.dta", clear
reg wage educ
```

The regression output looks like the following:

```
. reg wage educ

      Source |       SS           df       MS      Number of obs   =       935
-------------+----------------------------------   F(1, 933)       =    111.79
       Model |  16340644.5         1  16340644.5   Prob > F        =    0.0000
    Residual |   136375524       933  146168.836   R-squared       =    0.1070
-------------+----------------------------------   Adj R-squared   =    0.1060
       Total |   152716168       934  163507.675   Root MSE        =    382.32

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   60.21428   5.694982    10.57   0.000     49.03783    71.39074
       _cons |   146.9524   77.71496     1.89   0.059     -5.56393    299.4688
------------------------------------------------------------------------------
```

Based on the Coef. column, our estimated linear line looks like the following:

$$\begin{aligned}\widehat{\text{wage}}_i &= \hat{\beta}_0 + \hat{\beta}_1 \text{educ}_i \\ &= 146.952 + 60.214 \text{educ}_i\end{aligned}$$

   (b) Does this model suffer from omitted variable bias? Explain.

Yes. We can come up with multiple stories about variables that correlate with educ and also contributes to explaining wage. One such variable is years of working experience:

- At the same age, the more years of education a person has, the less years of working experience this person has. So years of experience is correlated with years of education.
- The more experienced one is, the higher their wage is going to be. So years of experience helps explain wage.

Thus, exper in this dataset is a omitted variable.

The more explicit way of showing this is by including exper in your regression model, and compare the coefficient on educ between the two models. If the coefficient on educ has changed greatly with respect to its standard error, then we know that there was bias in estimating this coefficient when we omitted exper.

So from our model in (a), coefficient on educ is estimated to be 60.214, and its standard error is 5.695.

Now estimate the model including exper:

$$\text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + u_i$$

Running it in Stata:

```
reg wage educ exper
```

The regression output looks like the following:

```
. reg wage educ exper

      Source |       SS           df       MS      Number of obs   =       935
-------------+----------------------------------   F(2, 932)       =     73.26
       Model |  20747023.1         2  10373511.5   Prob > F        =    0.0000
    Residual |   131969145       932  141597.795   R-squared       =    0.1359
-------------+----------------------------------   Adj R-squared   =    0.1340
       Total |   152716168       934  163507.675   Root MSE        =    376.29

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   76.21639   6.296604    12.10   0.000     63.85922    88.57355
       exper |   17.63777   3.161775     5.58   0.000     11.43275    23.84279
       _cons |  -272.5279   107.2627    -2.54   0.011    -483.0323   -62.02344
------------------------------------------------------------------------------
```

With exper included, coefficient estimate on educ increases to 76.216, and its standard error is now 6.297. This is an increase of $76.216 - 60.214 = 16.002$, which is greater than either standard error we have.

This means that coefficient estimate on educ has changed when we included exper in the regression model, indicating that educ was suffering from omitted variable bias in our model proposed in (a).

(c) Consider the following alternative model. What's the interpretation of β_1 in each model?

Model	**Interpretation on β_1**
$\ln \text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + u_i$	One unit change in education is associated with a $\beta_1 \times 100$ percent change in expected wage.
$\text{wage}_i = \beta_0 + \beta_1 \ln \text{educ}_i + u_i$	One percent change in education is associated with a $\frac{\beta_1}{100}$ unit change in expected wage.
$\ln \text{wage}_i = \beta_0 + \beta_1 \ln \text{educ}_i + u_i$	One percent change in education is associated with a β_1 percent change in expected wage.

(Refer to the first part of this handout for explanation.)

(d) Say that we want to estimate a model that satisfies the following criterion:

- Both educ and exper are included as explanatory variables

- We think exper matters a lot, so let's also include the squared exper
- Changes are reflected in percentage for the response variable
- Consider a different intercept and slope for people living in the south

What does this regression model look like?

The first two bullet points tell us that educ, exper, exper^2 are all going to be included in our regression model.

For the third bullet point, to interpret the coefficients as how much change in an explanatory variable is reflected percentage wise onto the response variable, we need to take the natural log of the response variable.

For the last bullet point, having different intercept and slope for people living in the south means that we need to include both a south dummy variable, and interaction terms between south and all other regressors (educ, exper, exper^2).

The final model looks like the following:

$$
\begin{aligned}
\ln \text{wage}_i = {} & \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 \\
& + \beta_4 \text{south}_i + \beta_5 \text{south}_i \times \text{educ}_i + \beta_6 \text{south}_i \times \text{exper}_i + \beta_7 \text{south}_i \times \text{exper}_i^2 + u_i
\end{aligned}
$$

where

- β_0: Intercept for all non-southerner
- β_1: Slope coefficient on educ for all non-southerner
- β_2: Slope coefficient on exper for all non-southerner
- β_3: Slope coefficient on exper squared for all non-southerner
- $\beta_0 + \beta_4$: Intercept for all southerner
- $\beta_1 + \beta_5$: Slope coefficient on educ for all southerner
- $\beta_2 + \beta_6$: Slope coefficient on exper for all southerner
- $\beta_3 + \beta_7$: Slope coefficient on exper squared for all southerner

(e) Estimate the regression model you proposed in (d). How can you tell if people living in the south actually don't have a separate intercept, or a separate slope for some variable?

Running the following commands in Stata:

```
gen log_wage = log(wage)
gen exper_squared = exper^2
reg log_wage educ exper exper_squared south south#c.educ south#c.exper
south#c.exper_squared
```

(The last two lines are supposed to be all in one line of code)

The regression output looks like the following:

```
. reg log_wage educ exper exper_squared south south#c.educ south#c.exper south#c.exper_squared

      Source |       SS           df       MS      Number of obs   =       935
-------------+----------------------------------   F(7, 927)       =     26.55
       Model |  27.6630193         7  3.9518599   Prob > F        =    0.0000
    Residual |  137.993264       927  .148860047   R-squared       =    0.1670
-------------+----------------------------------   Adj R-squared   =    0.1607
       Total |  165.656283       934  .177362188   Root MSE        =    .38582

------------------------------------------------------------------------------
          log_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
              educ |   .0634528    .007826     8.11   0.000     .0480942    .0788115
             exper |  -.0027977    .016306    -0.17   0.864    -.0347986    .0292032
     exper_squared |   .0007443   .0006871     1.08   0.279    -.0006041    .0020927
             south |  -1.032999   .2653163    -3.89   0.000    -1.553689   -.5123085
                   |
      south#c.educ |
                 1 |   .0375999   .0142058     2.65   0.008     .0097207    .0654792
                   |
     south#c.exper |
                 1 |   .0567574   .0281586     2.02   0.044     .0014955    .1120194
                   |
south#c.exper_squared |
                 1 |  -.0017509   .0011722    -1.49   0.136    -.0040514    .0005495
                   |
             _cons |   5.893468   .1482967    39.74   0.000     5.602432    6.184504
------------------------------------------------------------------------------
```

which means our linear line estimate is the following:

$$\begin{aligned}
\widehat{\ln \text{wage}}_i &= \hat{\beta}_0 + \hat{\beta}_1 \text{educ}_i + \hat{\beta}_2 \text{exper}_i + \hat{\beta}_3 \text{exper}_i^2 \\
&\quad + \hat{\beta}_4 \text{south}_i + \hat{\beta}_5 \text{south}_i \times \text{educ}_i + \hat{\beta}_6 \text{south}_i \times \text{exper}_i + \hat{\beta}_7 \text{south}_i \times \text{exper}_i^2 \\
&= 5.893 + .063 \text{educ}_i - .003 \text{exper}_i + .001 \text{exper}_i^2 \\
&\quad - 1.033 \text{south}_i + .038 \text{south}_i \times \text{educ}_i + .057 \text{south}_i \times \text{exper}_i - .002 \text{south}_i \times \text{exper}_i^2
\end{aligned}$$

Now, if people living in the south don't have a separate intercept or slope in this model, it's easy to account for that by looking at the t-statistic or p-value for these coefficient estimates (namely, $\hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6, \hat{\beta}_7$).

Recall that the t-statistic and p-value reported in the regression output table is testing the null hypothesis of $\beta_j = 0$. Thus, if we fail to reject the null at specified significance level (usually 5%), then we cannot claim that the coefficient is statistically significantly different from 0. In other words, including the corresponding x_j variable did nothing in further explaining y, since its β_j is basically 0 (i.e. as if x_j has never entered the model).

(Sidenote: If you're interested in knowing whether southerners potentially have **any** different intercept or slope considering all variables instead of just some, then we can perform a F-test with the null hypothesis

$$H_0 : \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$

This differs from looking at t-statistic reported in the regression output table. The t-statistic looks at just one β_j, and try to say something about whether southerners have a different slope for **just that one specific corresponding** x_j. When doing a joint F-test, you are testing that whether there's any evidence to support that the southerners have **potentially just a different intercept, just a different slope on one variable, just a different slope on two variables, or more, or some**

combination of all the aforementioned scenarios.

Doing this using Stata's `test` command is a bit tricky. Stata doesn't naturally recognize interaction terms generated using # as a variable. So you need to in the first place generate a separate interaction variable, and then run regression using this interaction variable you generated instead of using variables interacted with #.)

(f) If a non-southerner's experience increases from 10 to 11 years, how does that affect estimates of wage?

Since we are looking at non-southerner, south $= 0$, so we can ignore all the dummy and interaction terms for now.

The tricky bit of this problem is then that exper enters the equation in two different ways: both through exper directly, and through exper^2.

Let's look at when experience increases from 10 to 11 years, how exper and exper^2 are changing:

- exper changes by $11 - 10 = 1$
- exper^2 changes by $11^2 - 10^2 = 21$

So in terms of effect on predicted **ln(wage)**, increase experience from 10 to 11 years causes predicted **ln(wage)** to increase by

$$\Delta(\widehat{\ln(\text{wage})}) = \hat{\beta}_2 \Delta(\text{exper}) + \hat{\beta}_3 \Delta(\text{exper}^2) = -.003 * 1 + .001 * 21 = .018$$

The last bit is to translate change in **ln(wage)** to change in **wage**. Since when our response variable is measured in natural log, the interpretation from change in levels in x onto y is to say that y has changed by $\beta \times 100$ percent, this means that we can consider the .018 we found as an "effective" coefficient on exper in total, and say that predicted **wage** has increased by $.018 \times 100 = 1.8\%$

Thus, increasing non-southerner's years of experience from 10 to 11 is associated with 1.8% change in their wage.

(g) Use your regression model in (d) to predict the relationship between wage level and years of edcuation, for

- people living in the south with 10 years of experience, and
- people not living in the south with 10 years of experience

We need to create a new set of data with the same variables as what we used in the regression analysis, and since we are interested in the relationship between wage level and years of education, we need to find wage level at all possible years of education in our data. To do this, run the following command in Stata:

```
sum educ
```

(`sum` is short for `summarize`)

The output looks like the following:

```
. sum educ
```

Variable	Obs	Mean	Std. Dev.	Min	Max
educ	935	13.46845	2.196654	9	18

This tells us that education ranges from 9 to 18. Since we are measuring education in years, educ as a variable only records integer, so there should be in total $18 - 9 + 1 = 10$ values of education that we need to fill in (9, 10, 11, . . ., 18).

To create the new set of data to predict onto, we need to tell Stata explicitly how many observations there are by using `set obs`. The Stata commands look like the following:

```
// Predict relationship between wage and educ
summarize educ
local n_obs = 18 - 9 + 1
clear all
set obs ‘n_obs’ // should be back quotation followed by a single quotation mark

// Generate dataset to predict onto (make sure your explanatory variables are
// named the same as in the original dataset)
egen educ = seq(), from(9) to(18)
gen exper = 10
gen exper_squared = exper^2
gen south = 1
save "400_sp21_dis-5_predicted-data.dta", replace
```

And remember that before using the `predict` command to predict the values of ln(wage) ($\rightarrow$ the regression model's response variable), we need to rerun the regression model again:

```
// Before predict, need to run the original regression model again
use "400_sp21_dis-5_wage-data-for-reg-model.dta", clear
quietly reg log_wage educ exper exper_squared south south#c.educ
  south#c.exper south#c.exper_squared

use "400_sp21_dis-5_predicted-data.dta", clear
predict south_10_exper_log

replace south = 0 // now for non-southerner with 10 years of experience
predict nonsouth_10_exper_log

// Recall that our original response variable was log_wage: need to
// convert to wage level by taking exponential
gen south_10_exper_level = exp(south_10_exper_log)
gen nonsouth_10_exper_level = exp(nonsouth_10_exper_log)
save "400_sp21_dis-5_predicted-data.dta", replace
```

A Do-file solution is provided for this question.

(h) Plot your relationship between predicted wage level and years of edcuation for two cases outlined in (g).

Stata commands look like the following:

```
// Plot the relationship between predicted wage level and educ
label variable south_10_exper_level "south = 1, exper = 10"
label variable nonsouth_10_exper_level "south = 0, exper = 10"
```

```
line south_10_exper_level nonsouth_10_exper_level educ, legend(size(medsmall))
  ytitle("Monthly wage ($)") xtitle("Years of education")
graph export "400_sp21_dis-5_1g.png", replace
```

A Do-file solution is provided for this question. The resulting plot should look like the following: