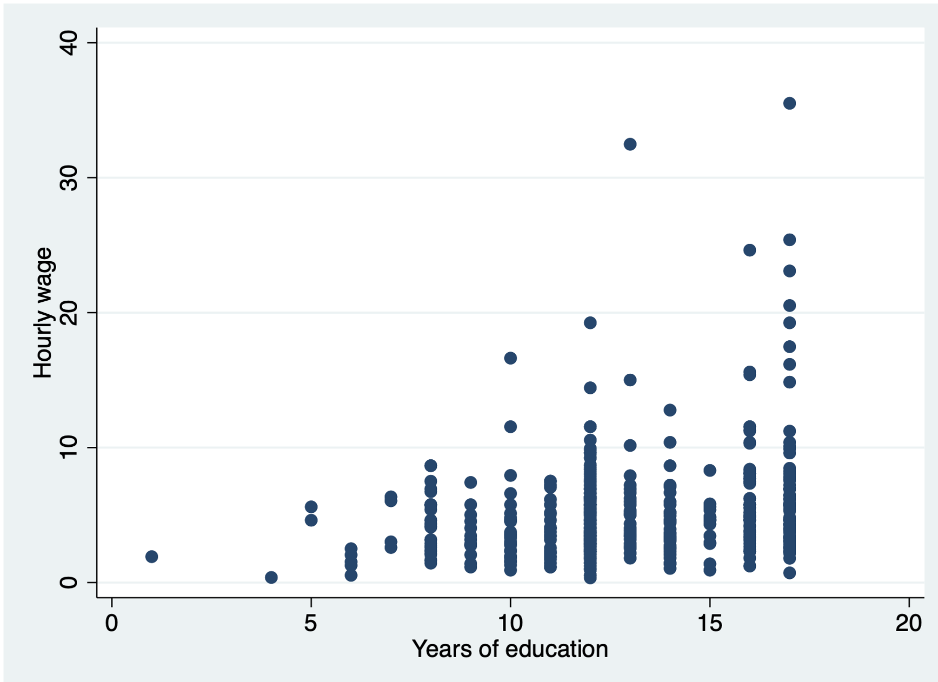
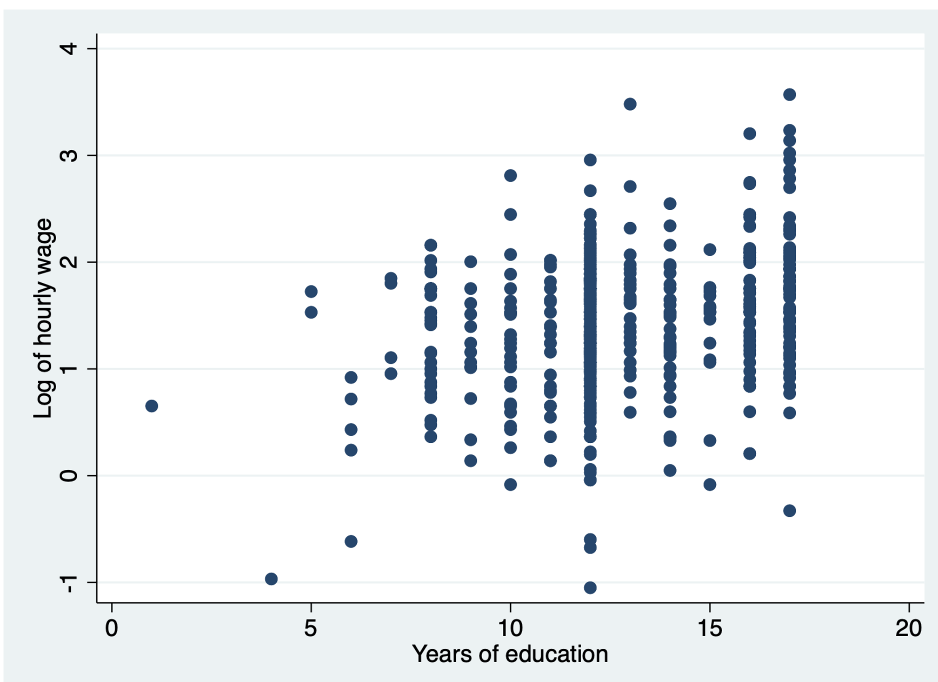
Dis 5: Nonlinear Regression; Dummy and Interaction

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 $\ln x_i$:

$$\frac{\partial \ln \hat{y}_i}{\partial \ln x_i} = \hat{\beta}_1 \tag{$*$}$$

    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?
    $\rightarrow$ **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:

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

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

   (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 ________ change in education is associated with a ________ ________ change in expected wage.
$\text{wage}_i = \beta_0 + \beta_1 \ln \text{educ}_i + u_i$	One ________ change in education is associated with a ________ ________ change in expected wage.
$\ln \text{wage}_i = \beta_0 + \beta_1 \ln \text{educ}_i + u_i$	One ________ change in education is associated with a ________ ________ change in expected wage.

   (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?

(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?

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

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