Week 13 Live Session - Instructor Version

w203 Instructional Team

March 9, 2018

Announcements

HW 13 is due 24 hours before next live session

Peer feedback distribution

Reminder: Quiz 2 is available and due before the week 14 live session.

|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|
|

Interpreting Specifications

What is the interpretation of β_0 and β_1 in each of the following specifications?

Log-level: $\log y = \beta_0 + \beta_1 x + u$

> *A unit change in x is associated with a $100 \cdot \beta_1$ percent change in y*

Level-log: $y = \beta_0 + \beta_1 \log x + u$

> *A 1 percent increase in x is associated with a change in y of $\beta/100$ units.*

Log-log: $\log y = \beta_0 + \beta_1 \log x + u$

A 1 percent increase in x is associated with a β_1 percent change in y

Added indicator: $y = \beta_0 + \beta_1 x + \beta_2 I(x = 0) + u$

Due to the presence of $I(x = 0)$ the interpretation of β_1 depends on the value of x. If $x \notin \{-1, 0\}$ then a unit increase of x is associated with a change in y of β_1 units. If not the change in y associated with a unit increase in x depends on both β_1 and β_2.

No intercept: $y = \beta_1 x + u$

Here β_1 has an interpretation identical to normal level-level form of the regression only if it is plausible that $y = 0$ whenever $x = 0$.

Issues with MLR: Using Logarithms

Using logarithms for the dependent or independent variables is one method used by statisticians to allow nonlinear relationships between the explained and explanatory variables.

Another potential benefit of using logs is that taking the log of a variable often narrows its range, which is useful when working with variables that are large monetary values.

Be careful not to use log transformation indiscriminantly - in some cases this can create extreme values. For example, when a variable y is between zero and one and takes on values close to zero, log(y) can be very large in magnitude whereas the original variable, y, is bounded between zero and one.

Hypothesis tests in MLR

Testing Hypotheses about a Single Population Parameter: The t Test

Hypothesis testing for a single coefficient is identical to the bivariate regression case with the t test statistic. The t statistic associated with any OLS coefficient can be used to test whether the corresponding unknown parameter in the population is equal to any given constant.

In most applications, our primary interest lies in testing the null hypothesis H_0: $\beta_j = 0$, where j corresponds to any of the k independent variables. The statistic we use to test (against any alternative) is called "the" t statistic or "the" t ratio of $\hat{\beta}_j$ and is defined as $t_{\hat{\beta}_j} = \hat{\beta}_j / se(\hat{\beta}_j)$

Since β_j measures the partial effect of x_j on (the expected value of) y, after controlling for all other independent variables, this hypothesis is saying that once $x_1, x_2, ., x_{j-1}, x_j, ., x_k$ have been accounted for, x_j has no effect on the expected value of y.

Note: We cannot state the null hypothesis as "x_j does have a partial effect on y" because this is true for any value of β_j other than zero.

Testing Hypotheses about a Single Linear Combination of the Parameters

In 4.4, Wooldridge shows how to test hypotheses about a single linear combination of the b_j by rearranging the equation and running a regression using transformed variables.

Remember to pay attention to the magnitude of the coefficient estimates in addition to the size of the t statistics.

Q. What is the difference between economical (or practical) and statistical significance?

The statistical significance of a variable x_j is determined entirely by the size of $t_{\hat{\beta}_j}$.The economic or practical significance of a variable is related to the size (and sign) of $\hat{\beta}_j$.

Testing Multiple Linear Restrictions: The F Test

To test multiple hypotheses about the underlying parameters $\beta_0, \beta_1, ..., \beta_k$, we can use multiple restrictions to test whether a set of independent variables has no partial effect on a dependent variable.

It is often useful to test joint hypotheses together rather than use independent tests of the coefficients. For instance, the joint test that math and verbal SAT scores have no effect on W203 grades against the alternative that one or both scores has an effect.

Tests of joint hypotheses have test statistics that are distributed according to either the F or χ^2 distributions. These tests are often called Wald tests and may be quoted either as F or as χ^2 statistics.

When computing an F statistic, the numerator df is the number of restrictions being tested, while the denominator df is the degrees of freedom in the unrestricted model. If there is only one numerator degree of freedom, we are testing only a single hypothesis and it seems like this should be equivalent to the usual t test. If a random variable t follows the t_{N-K} distribution, then its square t^2 follows the $F_{(1,N-K)}$ distribution.

Issues with MLR: Using Logarithms

Using logarithms for the dependent or independent variables is one method used by econometricians to allow nonlinear relationships between the explained and explanatory variables.

Another potential benefit of using logs is that taking the log of a variable often narrows its range, whifh is useful when working with variables that are large monetary values.

Be careful not to use log transformation indiscriminantly - in some cases this can create extreme values. For example, when a variable y is between zero and one and takes on values close to zero, log(y) can be very large in magnitude whereas the original variable, y, is bounded between zero and one.

R Exercise

Qualitative Data: Using Dummy Variables

In the folowing identify the reference group, explain why the indicator variables have been included in the following models, and interpret all coefficients.

$wage = \beta_0 + \beta_1 educ + \beta_2 I(educ \geq 12)$

> *Including $I(educ \geq 12)$ constitutes an intercept shift representing a graduation affect. i.e. it allows the regression visually it will be a double kink in the regression line at educ = 12.*

$wage = \beta_0 + \beta_1 educ + \beta_2 female$

> *Including an indicator for female is an intercept shift that makes the reference group male, so that β_0 is now the intercept for males and $\beta_0 + \beta_2$ is now the intercept for females. Additionally the marginal returns to education for all genders is β_1.*

$wage = \beta_0 + \beta_1 educ + \beta_2 female + \beta_3 educ * female$

> *Including an interaction allow for both different intercepts and differing slopes (or marginal returns to education) for males and females, so that the marginal returns to education for males is β_1 and marginal returns to education for females is $\beta_1 + \beta_3$.*

$wage = \beta_0 + \beta_1 female + \beta_2 I(educ = 2) + \beta_3 I(educ = 3) + ... + \beta_{20} I(educ = 20)$

> *Including an indicator for all education levels allows for different marginal returns to every education level, and removing interactions implies that marginal returns to education for males and females are the same. Note the reference group here is males with one year of education.*

R Exercise

The file, engin.RData contains data from the Material Requirement Planning Survey carried out in Thailand during 1998. It was collected by Thada Chaisawangwong, a former graduate student at MSU. These data are for engineers in Thailand, and represents a more homogeneous group than data sets that consist of people across a variety of occupations.

```
library(car)
library(lmtest)
```

```
## Loading required package: zoo
```

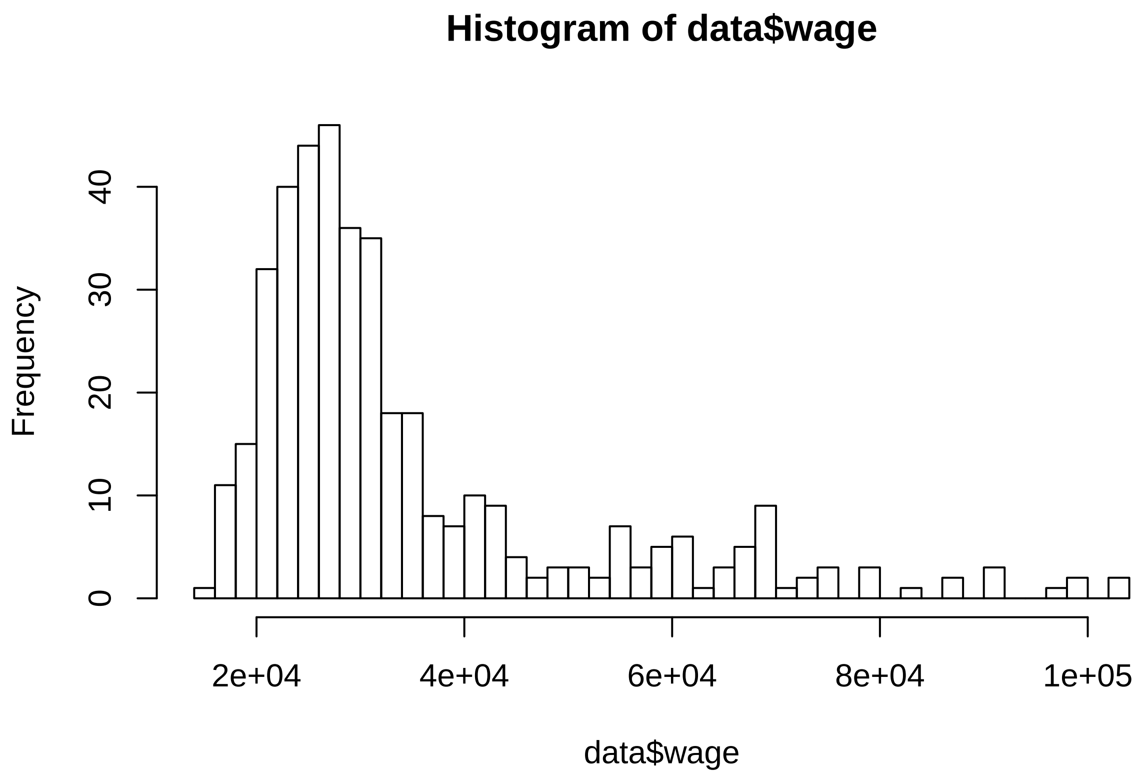
```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

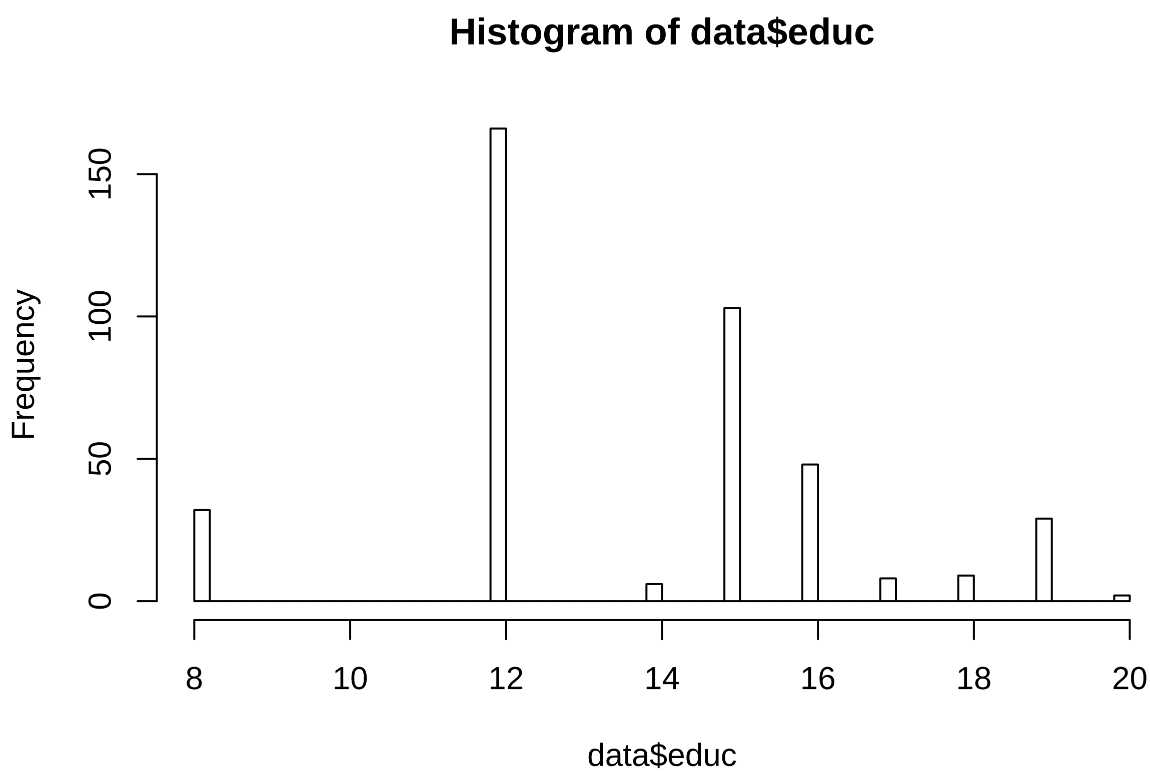
```
library(sandwich)
load("engin.RData")
```

1. Use visualizations to investigate the bivariate relationship between wage and educ. Based on this analysis, what transformation, if any, would you apply to wage?

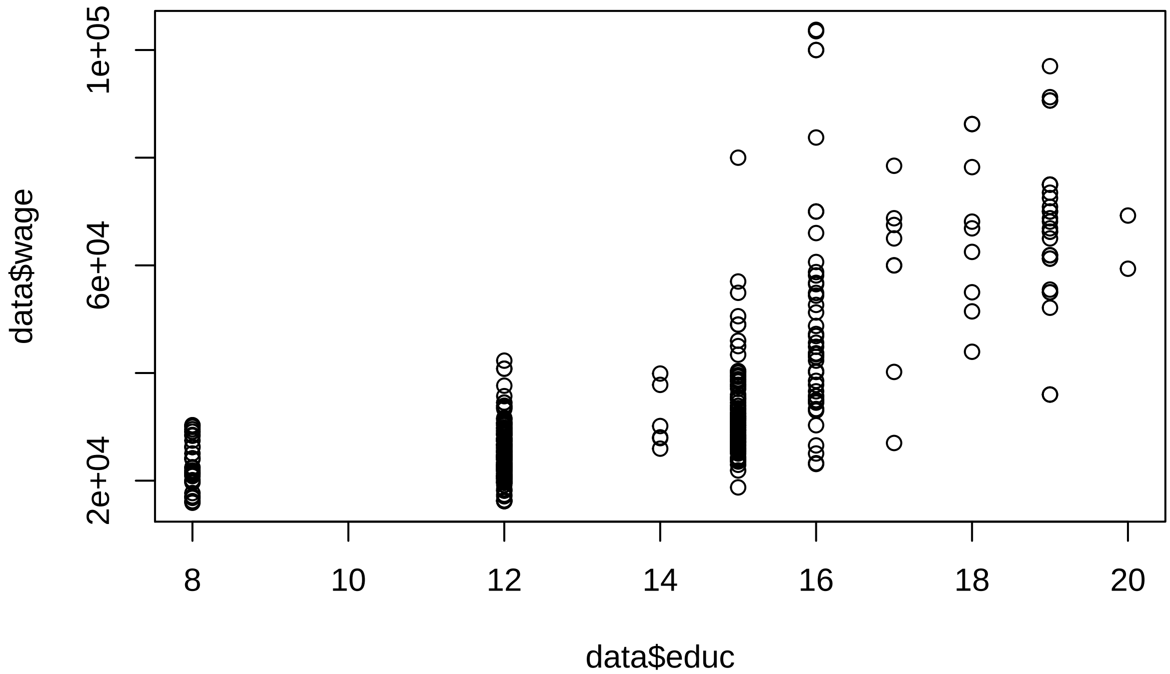
```
hist(data$wage, breaks = 50)
```



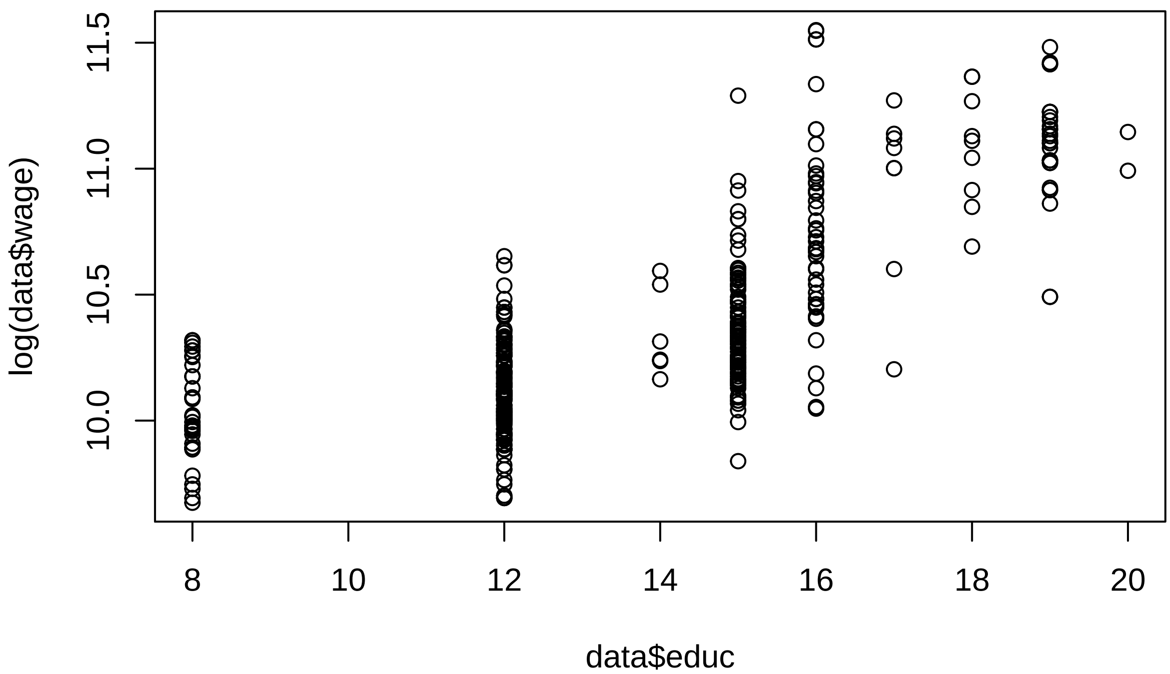

```
hist(data$educ, breaks = 50)
```




```
plot(data$educ, data$wage)
```




```
plot(data$educ, log(data$wage))
```




```
# the log transform looks more linear.
```

2. Create a linear model, model1, with just male and educ on the right hand side. Show how you would test the hypothesis that males and females have the same wages for all levels of education.

```
model1 = lm(lwage ~ male + educ, data = data)
coeftest(model1, vcov = vcovHC)
```

```
##
## t test of coefficients:
```

```
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 9.0422962  0.0697548 129.6297 < 2.2e-16 ***
## male        0.2209357  0.0270824   8.1579  4.46e-15 ***
## educ        0.0876873  0.0055327  15.8489 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The reference group here is female whose intercept is the "Intercept" coefficient and the male intercept is the sum of the "Intercept" and "male" coefficients thus testing whether females and males have the same wages is the same as testing whether the coefficient on "male" is different from zero. Here we see that the coefficient on "male" is highly statistically significant thus we reject the null that they are the same.

3. You are considering adding two variables representing experience to the model, exper and pexper. Show how you would test whether these variables are jointly significant.

```r
model2 = lm(lwage ~ male + educ + exper + pexper, data = data)
coeftest(model2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  8.9757428  0.1250991 71.7490 < 2.2e-16 ***
## male         0.2314251  0.0274754  8.4230 6.747e-16 ***
## educ         0.0864418  0.0056158 15.3925 < 2.2e-16 ***
## exper        0.0075838  0.0077168  0.9828   0.32632
## pexper      -0.0021549  0.0011883 -1.8134   0.07052 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
linearHypothesis(model2, c("exper = 0", "pexper = 0"), vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## exper = 0
## pexper = 0
##
## Model 1: restricted model
## Model 2: lwage ~ male + educ + exper + pexper
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1    400
## 2    398  2 1.9304 0.1464
```

Consequently we cannot reject the Null hypothesis that exper and pexper are jointly irrelevant to wages.

4. You are considering adding a variable, swage, representing starting wage to the right hand side. Explain how this would affect your ability to understand the effects of gender.

Previously we were comparing the difference is marginal effects between males and females, and also the difference in wages between males and females with the same education level. Now we if we include "swage" we would be comparing the the difference in wages between males and females *with the same level of education AND the same starting salary.*

5. Now show how you would alter your model to test whether marginal returns to education are the same for both men and for women.

```
model3 = lm(lwage ~ male + educ + male*educ, data = data)
coeftest(model3, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  9.3875129  0.1047858 89.5877 < 2.2e-16 ***
## male        -0.3563124  0.1553999 -2.2929 0.0223747 *
## educ         0.0595607  0.0087278  6.8243 3.296e-11 ***
## male:educ    0.0435615  0.0115654  3.7665 0.0001904 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here the reference group is females whose marginal effect of education is the coefficient on "educ" and marginal effect of education for mend is the sum of the coefficients on "educ" and "male:educ", thus testing where their marginal effects differ is the same as testing whether the coefficient on "male:educ" is different from zero, it highly statistically significant thus we can reject the hypothesis that they are the same.

6. As time allows, continue trying different model specifications, with the goal of understanding what effect gender has on wages.