Week 5 Code

Ryan Martin

January 31, 2018

Chapter 5 Notes - Multiple Regression

Key Ideas

Coefficients to linear regressions interpreted as partial derivatives.

explanatory variables are the x's.

Assume. - MR1. linear model is correct - MR2. mean 0 errors - MR3. homoskedastic errors - MR4. uncorrelated errors - MR5. no multicollinearity in x's - MR6. (Sometimes) Errors are $N(0, \sigma^2)$

Consequences 1. $E(y|x)$ is everything but e 2. Var(y|x) = var(e) 3. $cov(y_i, y_j|x) = 0$ 4. $y|x \sim N(\beta_0 \sum \beta_i x_i, \sigma^2)$

Interpolation vs Extrapolation 1. Extrapolation is often unlikely to be accurate 2. Interpolation is safer, but may not be sensible if high degree polynomial model or if overfit.

$$\hat{\sigma}^2 = \frac{1}{N-K} \sum \hat{e_i}^2$$

where K is the number of β parameters

Gauss-Markov - for MR model, if MR1-MR5 hold then OLS is BLUE

Recall variance covariance matrices

$se(b_k) := \sqrt{(\widehat{var(b_k)})}$

same rules for linear combos as before

Delta Method Mentioned in 5.6!

- Delta method is an approximation formula for complicated coefficients (e.g. things that we can't easily compute the variance for)
- If $\lambda = f(\beta_1, \beta_2)$ then

$$var(\lambda) \approx (\frac{\partial f}{\partial \beta_1})^2 var(\beta_1) + (\frac{\partial f}{\partial \beta_2})^2 var(\beta_2) + 2(\frac{\partial f}{\partial \beta_1})(\frac{\partial f}{\partial \beta_1}) cov(\beta_1, \beta_2)$$

See appendix 5b.5 for more.

5.7 - Interaction Terms

- So partial of 1 variable may depend on value of another
- log linear models still have same interpretation as before

5.8 Measuring Goodness of Fit - R^2

$R^2 = SSR/SST = \frac{(\sum \hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum \hat{e_i}^2}{\sum (y_i - \bar{y})^2}$ where the last equality comes from adding and subtracting y_i from the sum for all i and then a little algebra (called decomposing the square). Note, the last equality depends on an intercept term being in the regression! If don't have an intercept, also the below interpretation is not correct! If no intercept, don't discuss the R^2 (go back to plots)

Whenever we have an R^2, the interpretation is "R^2 percent of the variation in y is explained by the variation in the x variables"

A small but worthwile point to make is that explaining variation is not the same as explaining! I may be able to explain a good deal of variation in car max speed with variation in price, but that doesn't mean price explains max speed.

Questions

5.19

Use the data in `cps4_small.dat` to estimate the following wage equation:

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 HRSWK + e$$

a

Report the regression results. Interpret the estimates for non-intercept terms. Are they significantly different from 0?

Solution: The regression is below. The p-values for all non-intercept terms are significant; that is, all the p-values are less than .05. Recall the interpretation of log linear regression is expected (or average) percent change in y per unit change in the corresponding x, holding all else fixed. So, I interpret the coefficient as follows

- 1 year increase in education leads to an expected 9% growth in wages.
- 1 year increase in experience leads to an expected .6% increase in wages
- an employee who works 1 hour more per week (on average) is expected to earn, on average (all else fixed) .9% more in wages.

```
my_wd <- "C:/Users/ryanj/Dropbox/TA/Econ 103/Winter 2018/Data/s4poe_statadata"
my_file <- paste(my_wd, "cps4_small.dta", sep = "/")
library(haven)
dat <- read_stata(my_file)

reg_out <- lm(data = dat, I(log(wage)) ~ educ + exper + hrswk)
summary(reg_out)
```

```
##
## Call:
## lm(formula = I(log(wage)) ~ educ + exper + hrswk, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.65551 -0.36387 -0.01818  0.34827  1.53364
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.100540   0.109548  10.046  < 2e-16 ***
## educ        0.090306   0.006078  14.858  < 2e-16 ***
## exper       0.005776   0.001275   4.531 6.58e-06 ***
## hrswk       0.008941   0.001581   5.654 2.05e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5137 on 996 degrees of freedom
## Multiple R-squared:  0.2197, Adjusted R-squared:  0.2173
## F-statistic: 93.46 on 3 and 996 DF,  p-value: < 2.2e-16
```

b

Test the hypothesis that an extra year of education increases the wage rate by at least 10% against the alternative that is less than 10%

Solution We are testing $H_0 : 100 \times \beta_2 \geq 10$ $H_1 : 100 \times \beta_2 < 10$ This is a one sided test, to the left. Since the level isn't specified, I will test at the .05 level. We see below that the p-value is .056, so we fail to reject the null. Note that the degrees of freedom of our t-statistic is 1000 - 4, since the data have 1000 rows but we have 4 terms to estimate in our regression

```
s <- summary(reg_out)
s
```

```
##
## Call:
## lm(formula = I(log(wage)) ~ educ + exper + hrswk, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.65551 -0.36387 -0.01818  0.34827  1.53364
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.100540   0.109548  10.046  < 2e-16 ***
## educ        0.090306   0.006078  14.858  < 2e-16 ***
## exper       0.005776   0.001275   4.531 6.58e-06 ***
## hrswk       0.008941   0.001581   5.654 2.05e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5137 on 996 degrees of freedom
## Multiple R-squared:  0.2197, Adjusted R-squared:  0.2173
## F-statistic: 93.46 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
t_stat <- (coef(s)[2,1] - .10)/coef(s)[2,2]
t_stat
```

```
## [1] -1.594978
```

```
t.crit <- qt(.05, nrow(dat)-2)
t.crit
```

```
## [1] -1.646382
```

```
t_stat< t.crit #FALSE means fail to reject null
```

```
## [1] FALSE
```

```
p_value <-  pt(t_stat, s$df[2]) #now p_value just to the left!
      #note, df is now 1000 -4 because 4 terms to regression!
p_value
```

```
## [1] 0.05551723
```

```
#We see it's just larger than .05
```

c

Find a 90% interval estimate for the percentage increase in wage from working an addtional hour per week

Solution The confidence interval is $100 \times b_2 \pm t_{.95,1000-4} se(100 \times b_2) = 100 \times b_2 \pm t_{.95,1000-4} 100 se(\times b_2)$. I get [8.03, 10.03] as my interval (which includes 10, as it should since we failed to reject the null earlier)

```
t.crit <- qt(.05, s$df[2])
coef(s)[2,1]* 100 + 100 * t.crit * coef(s)[2,2]
```

```
## [1] 8.029877
```

```
coef(s)[2,1]* 100 - 100 * t.crit * coef(s)[2,2]
```

```
## [1] 10.03125
```

```
#Note the CI includes 10 (just barely), another reason
#to fail to reject the null.
```

d

Re-estimate the model with the additional variables $EDUC \times EXPER$, $EDUC^2$ and $EXPER^2$. Report the results. Are the estimated coefficients significantly different from zero?

Solution Note, that education was significant before, but it is no longer! This is a tricky feature of regression. The useful terms depend on what else is (or isn't) in the model! The interaction term is also not significant at the 5 percent level. All the other terms are significant at the .05 level.

```
reg_out <- lm(data = dat, I(log(wage)) ~ educ +
          exper + hrswk + I(educ^2) +
          I(exper^2) + I(educ* exper))

# alternative way could specify interaction term
#lm(data = dat, I(log(wage)) ~ educ + exper + hrswk + educ:exper +
#    I(educ^2) + I(exper^2))

#alternative 2 - note the times term means individuals and interaction
#not just times!
#lm(data = dat, I(log(wage)) ~ educ*exper + hrswk  +
#    I(educ^2) + I(exper^2))

s <- summary(reg_out)
s
```

```
##
```

```
## Call:
## lm(formula = I(log(wage)) ~ educ + exper + hrswk + I(educ^2) +
##     I(exper^2) + I(educ * exper), data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.33817 -0.32561 -0.02568  0.32695  1.47365
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.266e-01  3.404e-01   2.722   0.0066 **
## educ             4.903e-02  3.663e-02   1.339   0.1810
## exper            5.274e-02  9.749e-03   5.410 7.89e-08 ***
## hrswk            6.693e-03  1.568e-03   4.268 2.16e-05 ***
## I(educ^2)        2.365e-03  1.105e-03   2.141   0.0325 *
## I(exper^2)      -6.287e-04  8.881e-05  -7.080 2.73e-12 ***
## I(educ * exper) -9.238e-04  5.054e-04  -1.828   0.0679 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5002 on 993 degrees of freedom
## Multiple R-squared:  0.2624, Adjusted R-squared:  0.2579
## F-statistic: 58.87 on 6 and 993 DF,  p-value: < 2.2e-16
```

e

For the new model, find expressions for the marginal effects $\frac{\partial \ln(WAGE)}{\partial EDUC}$ and $\frac{\partial \ln(WAGE)}{\partial EXPER}$

Solution This is just a hand problem. Note that the textbook author is using the term "marginal effect" even though he wants the marginal in (x, ln(WAGE)) space rather than (x, WAGE) space. This is different from how we usually use that word in class. It's fine, because he also gives us the derivative form he wants. Anyway, our full regression is

$$\begin{aligned}\ln(WAGE) = \beta_1 &+ \beta_2 EDUC + \beta_3 EXPER + \beta_4 HRSWK \\ &+ \beta_5 EDUC^2 + \beta_6 EXPER^2 + \beta_7 EDUC \times EXPER + e\end{aligned}$$

Thus,

$$\frac{\partial \ln(WAGE)}{\partial EDUC} = \beta_2 + 2\beta_5 EDUC + \beta_7 EXPER$$

while

$$\frac{\partial \ln(WAGE)}{\partial EXPER} = \beta_3 + 2\beta_6 EXPER + \beta_7 EDUC$$

f

Estimate the marginal effects $\frac{\partial \ln(WAGE)}{\partial EDUC}$ for two workers Jill and Wendy. Jill has 16 years of education and 10 years of experience while Wendy has 12 years of education and 10 years of experience. What can you say about the marginal effect of education as education increases?

Solution We just plug our estimates into the formulas above.

```
#Jill:
coef(s)[2,1] + 2*coef(s)[5,1]*16 + coef(s)[7,1]*10
```

```
## [1] 0.1154664
```

```
#Wendy
coef(s)[2,1] + 2*coef(s)[5,1]*12 + coef(s)[7,1]*10
```

```
## [1] 0.09654733
```

The interpretation is that a 1 unit increase in education for Jill, holding her experience level fixed at 10, would increase her wages by 11.5%. For Wendy, 9.7%.

g

Test, as an alternative hypothesis, that Jill's marginal effect of education is greater than that of Wendy. Use a 5% significance level.

Solution This sounds complicated! But it's actually easy, which we see as soon as we write down the hypothesis:

$$H_0 : \beta_2 + 2\beta_5 16 + \beta_7 10 \geq \beta_2 + 2\beta_5 12 + \beta_7 10$$
$$\Leftrightarrow H_0 : 8\beta_5 \geq 0$$
$$\Leftrightarrow H_0 : \beta_5 \geq 0$$

So, our t-statistic is simply $\beta_5/se(\beta_5)$. We see below that our test statistic is more extreme than our critical value. So, reject the null. It appears Jill's marginal effect is larger than Wendy's at the .95 level.

```
coef(s)[5,3] #test-stat
```

```
## [1] 2.140631
```

```
t.crit <- qt(.95, 1000-7)
t.crit
```

```
## [1] 1.64639
```

h

Estimate the marginal effects $\frac{\partial \ln(WAGE)}{\partial EXPER}$ for two workers Chris and Dave. Chris has 16 years of education and 20 years of experience while Dave has 16 years of education and 30 years of experience. What can you say about the marginal effect of experience as experience increases?

Solution: First note that $\frac{\partial}{\partial EXPER}\frac{\partial \ln(WAGE)}{\partial EXPER} = \frac{\partial^2 \ln(WAGE)}{\partial EXPER^2} = 2\beta_6$. So, if β_6 is positive, the marginal effect of experience is increasing as experience increases. If β_6 is negative, it's decreasing. The estimate is negative, so it appears we are in the second case: the more experience you have, the less (on average) it helps.

```
#Chris:
coef(s)[3,1] + 2*coef(s)[6,1]*20 + coef(s)[7,1]*16
```

```
## [1] 0.01281497
```

```
#Dave
coef(s)[3,1] + 2*coef(s)[6,1]*30 + coef(s)[7,1]*16
```

```
## [1] 0.00024038
```

i

For someone with 16 years of education, find a 95% interval estimate for the number of years of experience after which the marginal effect of experience becomes negative.

$$\begin{aligned}\frac{\partial \ln(WAGE)}{\partial EXPER} &= \beta_3 + 2\beta_6 EXPER + \beta_7 EDUC \\ \Rightarrow \frac{\partial \ln(WAGE)}{\partial EXPER} &\leq 0 \\ \Leftrightarrow \beta_3 + 2\beta_6 EXPER + \beta_7 EDUC &\leq 0 \\ \Leftrightarrow -2\beta_6 EXPER &\geq \beta_3 + \beta_7 EDUC \\ \Leftrightarrow \begin{cases} EXPER \geq (\beta_3 + \beta_7 EDUC)/(-2\beta_6) \text{ if } \beta_6 < 0 \\ EXPER \leq (\beta_3 + \beta_7 EDUC)/(-2\beta_6) \text{ if } \beta_6 > 0 \end{cases}\end{aligned}$$

In any case, our estimate for the point of change, when $EDUC = 16$ is

$$\hat{\lambda} = -(b_3 + b_7 \times 16)/(2b_6)$$

Note that this formula involves dividing by a coefficient! Whenever you have a coefficient in the bottom, the correct/easiest way to get a formula for the distribution of the term is to use the delta method (although if you were in a stats class, you would probably just simulate it!).

Note, however, that the delta method is given for a two-term problem. We have a 3 term problem. We need to use a mathematical substitution to make it all work. Call $z = b_3 + 16 \times b_7$. Then $\lambda = -z/(2b_6)$ Next, note that $var(-\lambda) = (-1)^2 var(\lambda) = var(\lambda)$, so we can just find the variance of $\lambda^\star = z/(2b_6$

Applying the delta method gets us

$$\begin{aligned}\hat{var}(\lambda^\star) &\approx (\frac{\partial \lambda^\star}{\partial z})^2 var(z) + (\frac{\partial \lambda^\star}{\partial b_6})^2 var(b_6) + 2(\frac{\partial \lambda^\star}{\partial z})(\frac{\partial \lambda^\star}{\partial b_6}) covar(z, b_6) \\ &= (\frac{1}{2b_6})^2 [var(b_3) + 16^2 var(b_7) + 2 \times 16 cov(b_3, b_7)] + (-\frac{b_3 + 16b_7}{2b_6^2})^2 var(b_6) \\ &\quad + 2(\frac{1}{2b_6})(-\frac{b_3 + 16b_7}{2b_6^2}) covar(b_3 + 16b_7, b_6) \\ &= (\frac{1}{2b_6})^2 [var(b_3) + 16^2 var(b_7) + 2 \times 16 cov(b_3, b_7)] + (-\frac{b_3 + 16b_7}{2b_6^2})^2 var(b_6) \\ &\quad + 2(\frac{1}{2b_6})(-\frac{b_3 + 16b_7}{2b_6^2})[covar(b_3, b_6) + 16 covar(b_7, b_6)]\end{aligned}$$

Finally, our confidence interval is

$$\hat{\lambda} \pm t_{.975,1000-7} \sqrt{\hat{var}(\lambda)}$$

Which is complicated, but we can do it. For an answer, I get [27.2, 33.2]

```
cov_mat <- vcov(reg_out) #has all covariances for us

lamb.var.hat = (1/(2*coef(s)[6,1]))^2*(cov_mat[3,3] +
    16^2 *cov_mat[7,7] + 2*16*cov_mat[3,7]) +
    (-(coef(s)[3,1] + 16*coef(s)[7,1])/(2 *coef(s)[6,1]^2))^2*
    cov_mat[6,6] + 2*1/(2*coef(s)[6,1]) * (-(coef(s)[3,1] +
    16*coef(s)[7,1])/(2 *coef(s)[6,1]^2))*(cov_mat[3,6] +
```

```
                                        16*cov_mat[6,7])

lamb.hat = -(coef(s)[3,1] + 16*coef(s)[7,1])/(2 *coef(s)[6,1])
lamb.hat
```

```
## [1] 30.19116
```

```
sqrt(lamb.var.hat)
```

```
## [1] 1.516336
```

```
lamb.hat + sqrt(lamb.var.hat)*qt(.975, 1000-7)
```

```
## [1] 33.16675
```

```
lamb.hat - sqrt(lamb.var.hat)*qt(.975, 1000-7)
```

```
## [1] 27.21557
```