Week 2 Code

Ryan Martin

January 10, 2018

Discussion Section Questions

Note, R displays decimals in an unfamiliar way. 1.08e+3 is R's way of saying 1.08×10^3

3.8

The file `br2.dat` contains data on 1080 houses sold in Baton Rouge, Louisiana during mid-2005. The data include sale price and the house size in square feet. Also included is an indicator variable $TRADITIONAL$ indicating whether the house style is traditional or not.

a

For the traditional-style houses estimate the linear regression model $PRICE = \beta_1 + \beta_2 SQFT + e$. Test the null hypothesis that the slope is zero against the alternative that it is positive, using the $\alpha = .01$ level of significance. Follow and show all the test steps described in Chapter 3.4.

Solution: Testing β_2 is zero is done for us in the regresssion (for a two sided test). And we can actually divide the two-sided p-value in half to get the one sided p-value; be cautious, this fact only works because the T distribution (of whatever df) is symmetric around 0 and we are testing around 0. But, I will also follow the steps.

Recall the notation, $t_{a,m}$ where a is the probability (area) in the left tail of the t-distribution of degrees of freedom m's density curve up until the line $t_{a,m}$. That is, $t_{a,m}$ is defined as the solution (in t) to the equation

$$P(T_m < t) = a$$

, with T_m a T-distributed random variable with degrees of freedom equal to m.

1. Statement.
$$\begin{cases} H_0 : \beta_2 \leq 0 \\ H_1 : \beta_2 > 0 \end{cases}$$
2. the test statistic is $t = \frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} \sim T_{(N-2)}$. Note here that N is the number of rows of the data set, 582.
3. $t_{.99,582-2}$ is our critical value; rejecting to the far right (vs $t_{.01,580}$)
4. check if $\frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} < t_{.99,582-2}$. From our regression, we see $\frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} = 32.1$. The code below shows $t_{.99,582-2} = 2.33$. So, our test statistic is not within the critical value bounds. Said another way, our test statistic is in the critical (or rejection) region.
5. Reject the null

```
my_wd <- "C:/Users/ryanj/Dropbox/TA/Econ 103/Winter 2018/Data/s4poe_statadata"
my_file <- paste(my_wd, "br2.dta", sep = "/")

library(haven)
dat <- read_stata(my_file)
```

```
    #could also use
    #haven::read_stata(my_file)
    #for questions, type ?read_stata or ?haven::read_stata
nrow(dat)
```

```
## [1] 1080
```

```
#Take a look
#View(dat)

#note, we only want the traditional houses
dat <- dat[dat$traditional==1,]

attach(dat)
reg_out <- lm(price ~ sqft)
s <- summary(reg_out)
s
```

```
##
## Call:
## lm(formula = price ~ sqft)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -274691  -22349   -1486   17950  511700
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28407.559   5728.161  -4.959 9.31e-07 ***
## sqft           73.772      2.301  32.061  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48680 on 580 degrees of freedom
## Multiple R-squared:  0.6393, Adjusted R-squared:  0.6387
## F-statistic:  1028 on 1 and 580 DF,  p-value: < 2.2e-16
```

```
#Note in the regression that they did the math for us
#The t.value
coef(s)[2,1]/coef(s)[2,2]
```

```
## [1] 32.06124
```

```
coef(s)[2,3] #same, because null is equals 0.
```

```
## [1] 32.06124
```

```
#if wanted to calculate critical values ourself
t.crit <- qt(p=.99, df = nrow(dat) - 2) #critical value
t.crit
```

```
## [1] 2.332794
```

```
#since
coef(s)[2,3] > t.crit #TRUE means reject null for out one sided test at .01 sig level
```

```
## [1] TRUE
```

b

Using the linear model in (a), test the null hypothesis (H_0) that the expected price of a house of 2000 square feet is equal to, or less than, \$120,000. What is the appropriate alternative hypothesis? Use the $\alpha = .01$ level of significance. Obtain the p-value of the test and show its value on a sketch. What is your conclusion?

Solution:

The appropriate alternative is

$$H_1 : \text{Expected price of a house of 2000 square feet is greater than \$120,000}$$

Note that the expected price of a house of $x_0 = 2000$ square feet is $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x_0$ in our model. So, we can rephrase our hypothesis test as follows:

$$\begin{cases} H_0 : \beta_1 + \beta_2 2000 \leq 120,000 \\ H_1 : \beta_1 + \beta_2 2000 > 120,000 \end{cases}$$

As described on p.114 of the textbook, our test statistic is

$$(\hat{\beta}_1 + \hat{\beta}_2 2000 - 120,000)/se(\hat{\beta}_1 + \hat{\beta}_2 2000)$$

where

$$se(\hat{\beta}_1 + \hat{\beta}_2 2000) = \sqrt{var(\hat{\beta}_1) + 2,000^2 var(\beta_2) + 2 \times 2,000 cov(\beta_1, \beta_2)}$$

This would be a lot of terms to do by hand. Luckily the computer does all this for us!

```
#calculating covariance matrix
#upper left is var of \beta_1, lower right is var of \beta_2.
#square root of these are standard errors in regression output.
#off-diag are covariances
my_cov_mat <- vcov(reg_out)
my_cov_mat
```

```
##             (Intercept)          sqft
## (Intercept) 32811823.40 -12335.340644
## sqft          -12335.34      5.294462
```

```
my_se <- sqrt(my_cov_mat[1,1] + 2000^2*my_cov_mat[2,2] + 4000 *
                my_cov_mat[1,2])

s <- summary(reg_out)
test_stat = (coef(s)[1,1] + 2000*coef(s)[2,1]-120000)/my_se
t_c = qt(.99, nrow(dat) - 2)
test_stat
```

```
## [1] -0.4005866
```

```
t_c #same as before
```

```
## [1] 2.332794
```

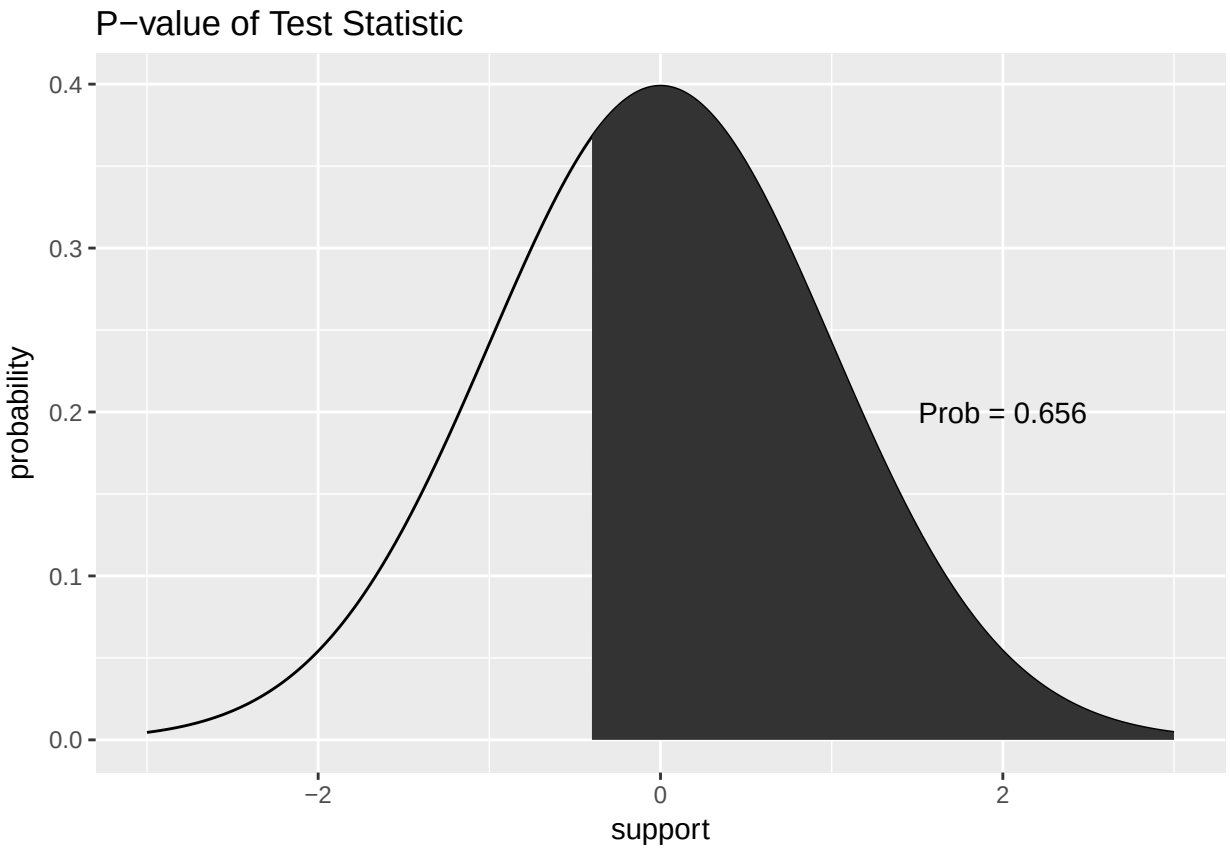
```
test_stat < t_c #fail to reject null
```

```
## [1] TRUE
```

```r
#p value is probability of observing our observation
#or more extreme, under the null.
#Since it's a right tail test, just the right tail.
#But a two-sided test should include probability of
#more estreme on both sides! be careful. See book 111-112
p_value <- 1 - pt(test_stat, nrow(dat)-2)
p_value
```

```
## [1] 0.655564
```

```r
#Then use this new data.frame with geom_polygon
support = -300:300/100
plot_data <- as.data.frame(cbind(support,
                  probability = dt(support, nrow(dat)-2)))
shade <- as.data.frame(rbind(c(test_stat,0),
               subset(plot_data, support > test_stat),
               c(plot_data[nrow(plot_data), "support"], 0)))
names(shade) <- c("x","y")
library(ggplot2)
ggplot(data = plot_data, aes(x = support, y= probability))  +
  geom_line() +
  annotate("text", x = 2, y = .2, label =
             paste("Prob = ", round(p_value,3), sep = "")) +
    #rounded so didn't display 10 digits
  geom_polygon(data = shade, aes(x,y )) +
  ggtitle("P-value of Test Statistic")
```



c

Based on the estimated results from part (a), construct a 95% interval estimate of the expected price of a house of 2000 square feet.

Solution Code below. To do it by hand, use $\hat{\beta}_1 + \hat{\beta}_2 2000 \pm t_{.975,N-2} se(\hat{\beta}_1 + \hat{\beta}_2 2000)$. Very similar to (b).

```
#R has a built in method
new.dat <- data.frame(sqft=2000)
predict.lm(reg_out, newdata = new.dat, interval = "confidence",
        level = .95)
```

```
##        fit      lwr      upr
## 1 119136.3 114901.8 123370.8
```

d

For the traditional-style houses, estimate the quadratic regression model $PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$. Test the null hypothesis that the marginal effect of an additional square foot of living area in a home with 2000 square feet of living space is \$75 against the alternative that the effect is less than \$75. Use the $\alpha = .01$ level of significance. Repeat the same test for a home of 4000 square feet of living space. Discuss your conclusions.

Solution For the 2000 SQFT case, we have

$$\begin{cases} H_0 : 2\alpha_2 2000 \geq 75 \\ H_1 : 2\alpha_2 2000 < 75 \end{cases}$$

Test statistics is $\frac{\hat{\alpha_2} - 75/4000}{se(\hat{\alpha_2})} \sim t_{N-2}$

For the 4000 square foot case, the test statistics is $\frac{\hat{\alpha_2} - 75/8000}{se(\hat{\alpha_2})} \sim t_{N-2}$

See 3.12 (c) and (d) for details

```
#for 2000
reg_sq <- lm(price ~ I(sqft^2))
s <- summary(reg_sq)
test_stat <- (coef(s)[2,1] - 75/4000)/coef(s)[2,2]
pt(test_stat, df = nrow(dat) - 2, lower.tail = T) #reject null
```

```
## [1] 7.717093e-65
```

```
#for 4000
test_stat <- (coef(s)[2,1] - 75/8000)/coef(s)[2,2]
pt(test_stat, df = nrow(dat) - 2, lower.tail = T) #accept null
```

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

e

For the traditional-style houses, estimate the log-linear regression model $\log(PRICE) = \gamma_1 + \gamma_2 SQFT + e$. Test the null hypothesis that the marginal effect of an additional square foot of living area in a home with 2000 square feet of living space is \$75 against the alternative that the effect is less than \$75. Use the $\alpha = .01$ level of significance. Repeat the same test for a home of 40000 square feet of living space. Discuss your conclusions.

Solution: Note, this is actually much harder than d!

Recall that the marginal effect of an additional square foot of living area in a home on price is found through implicit differentiation

$$\frac{1}{\widehat{PRICE}}\frac{dPRICE}{dSQFT} = \hat{\gamma}_2 \Rightarrow \frac{dPRICE}{dSQFT} = \hat{\gamma}_2\widehat{PRICE} = \hat{\gamma}_2 \exp(\hat{\gamma}_1 + \hat{\gamma}_2 SQFT)$$ [1]

Our test:

$\begin{cases} H_0 : \gamma_2 \exp(\gamma_1 + \gamma_2 2000) \geq 75 \\ H_1 : \gamma_2 \exp(\gamma_1 + \gamma_2 2000) < 75 \end{cases}$ Our test statistic is $\frac{\hat{\gamma}_2 \exp(\hat{\gamma}_1 + \hat{\gamma}_2 2000) - 75}{se(\hat{\gamma}_2 \exp(\hat{\gamma}_1 + \hat{\gamma}_2 2000))} \sim T_{N-2}$ Note that the variance of $\hat{\gamma}_2 \exp(\hat{\gamma}_1 + \hat{\gamma}_2 2000))$ is not a trivial thing to calculate! Most people would just bootstrap the value or use R's predict function. If you really wanted to do it by hand, you will need more than the matrices I calculated before! You would want to use SR6 (normality of the errors), to determine that $\binom{\hat{\gamma}_1}{\hat{\gamma}_2} \sim N(\text{estimated value, covariance matrix})$, where the covariance matrix is as calculated in part b of this problem. Suppose $\hat{\mu}$ is your estimate (vector) and $\hat{\Sigma}$ is your covariance matrix. Then the pdf of $\binom{\hat{\gamma}_1}{\hat{\gamma}_2}$ is exaclty $f(x,y) := \frac{1}{2\pi\sqrt{|det(\hat{\Sigma})|}} \exp(-\frac{1}{2}(\binom{x}{y} - \hat{\mu})'\hat{\Sigma}^{-1}((\binom{x}{y} - \hat{\mu}))$. Then, you can calculate,

$$E(\hat{\gamma}_2 \exp(\hat{\gamma}_1 + \hat{\gamma}_2 2000)) = \int y \exp(x + y2000) f(x,y) dxdy \qquad \text{(Expectation)}$$

$$Var(\hat{\gamma}_2 \exp(\hat{\gamma}_1 + \hat{\gamma}_2 2000)) = \int (y \exp(x + y2000) - E(\hat{\gamma}_2 \exp(\hat{\gamma}_1 + \hat{\gamma}_2 2000))^2 f(x,y) dxdy$$

Even with a computer, this is a little cumbersome to calculate; even computers struggle with integrals! It turns out a reasonable approximation to this integral is just taking the variance of many simulations of $\hat{\gamma}_2 \exp(\hat{\gamma}_1 + \hat{\gamma}_2 2000)$ where $\binom{\hat{\gamma}_1}{\hat{\gamma}_2}$ are simulated according to the normal distribution above.

```
###### 2000 SQFT
attach(dat)
log_reg <- lm(I(log(price)) ~ sqft)
logs <- summary(log_reg)
logs
```

```
##
## Call:
## lm(formula = I(log(price)) ~ sqft)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46087 -0.12238  0.02825  0.16430  0.90762
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.080e+01  3.467e-02  311.47   <2e-16 ***
## sqft        4.132e-04  1.393e-05   29.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2947 on 580 degrees of freedom
## Multiple R-squared:  0.6028, Adjusted R-squared:  0.6022
## F-statistic: 880.4 on 1 and 580 DF,  p-value: < 2.2e-16
```

```
my_cov_mat <- vcov(log_reg)
library(MASS) #no need to install.packages(MASS), just lib call
temp <-  mvrnorm(n = 1000, mu = logs$coefficients[1:2,1], Sigma=my_cov_mat)
dim(temp) #each row is a draw from the joint normal of our coefficients
```

```
## [1] 1000    2
```

```
sd_est <- sd(temp[,2] * exp(temp[,1] + 2000 * temp[,2]) )
estimated_marg <- coef(logs)[2,1] * exp(coef(logs)[1,1] + 2000 * coef(logs)[2,1])
estimated_marg #estimated marginal effect on price at 2000 sqft.
```

```
## [1] 46.2433
```

```
  # i.e., one more sqft at 2000 sqft is worth around 46 to price.
  # this is lower than 75. Expect to reject the null
test_stat <- (estimated_marg - 75)/sd_est
test_stat < t_c #TRUE, so reject null
```

```
## [1] TRUE
```

```
############################################################
#4000 sqft


sd_est <- sd(temp[,2] * exp(temp[,1] + 4000 * temp[,2]) )
estimated_marg <- coef(logs)[2,1] * exp(coef(logs)[1,1] + 4000 * coef(logs)[2,1])
estimated_marg #106. Much larger marginal effect, as expected by format!
```

```
## [1] 105.677
```

```
    #Since it's much larger than 75, expect to fail to reject null
    #note, when y is loged and x is linear, have increasing marginal effect of x on y!
    #if y vs log(x), then have decreasing marginal effect of x on y
test_stat <- (estimated_marg - 75)/sd_est
test_stat < - t_c #fail to reject null
```

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

```
########################################
#Sanity check
new.dat2 <- data.frame(sqft=c(2000,4000))
predict_win <- predict.lm(log_reg, newdata = new.dat2, interval = "confidence",
        level = .99)
predict_win
```

```
##        fit      lwr      upr
## 1 11.62541 11.59169 11.65913
## 2 12.45188 12.38399 12.51978
```

```
#lower 2000
coef(logs)[1] *exp(predict_win[1,2])
```

```
## [1] 1168386
```

```
#upper 2000
coef(logs)[1] *exp(predict_win[1,3])
```

```
## [1] 1249910
```

```
#lower 4000
coef(logs)[1] *exp(predict_win[2,2])
```

```
## [1] 2580344
```

```
#upper 4000
coef(logs)[1] *exp(predict_win[2,3])
```

```
## [1] 2955634
```

3.12

Is the relationship between experience and wages constant over one's lifetime? To investigate we will fit a quadratic model using the data file `cps4_small.dat`, which contains 1,000 observations on hourly wage rates, experience and other variables from the 2008 CPS.

a

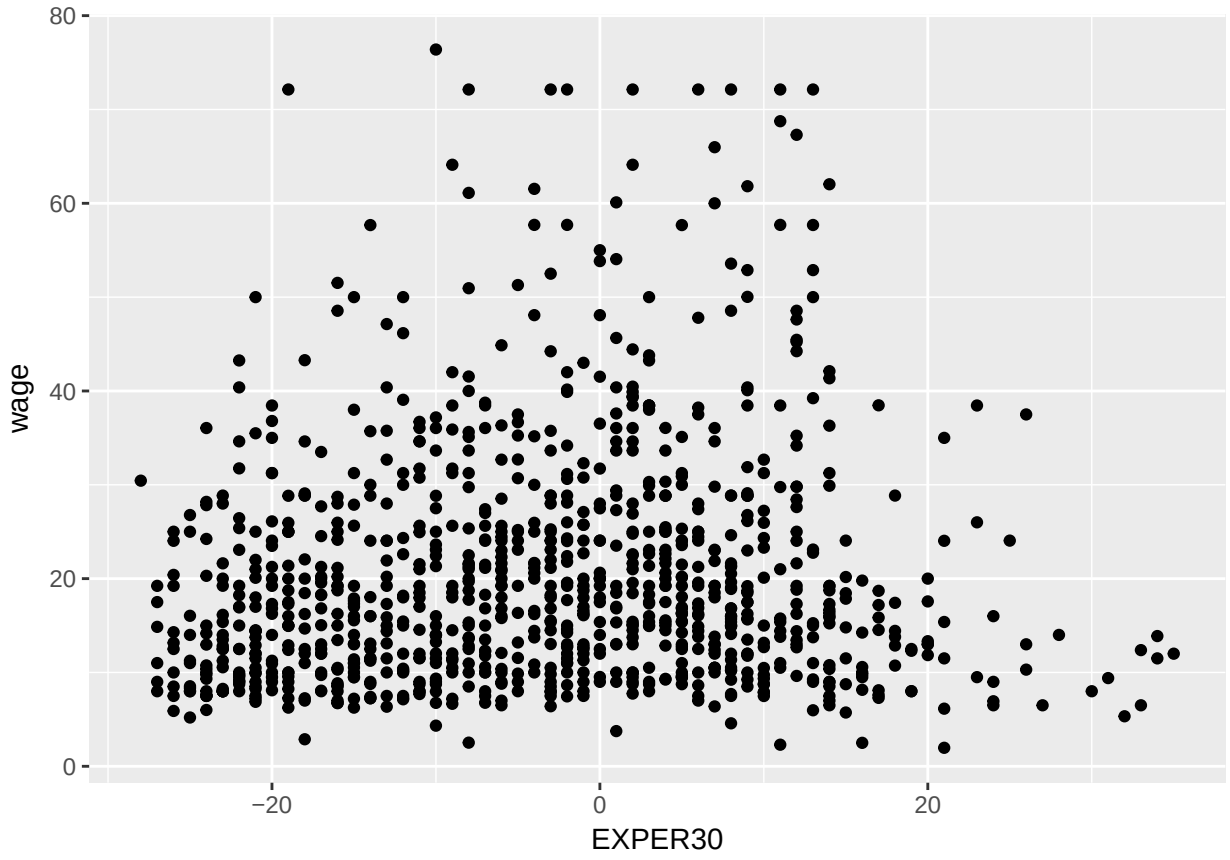
Create a new variable called EXPER30 = EXPER - 30. Construct a scatter diagram with WAGE on the vertical axis and EXPER30 on the horizontal axis. Are any patterns evident?

Solution: Code is below. The only pattern I notice is that wages look a little more spread out around when expr30 is closest to 0. This makes sense. Young and old people tend to be paid the least. Anyway, we don't really expect linear returns to experience from this plot. It looks increasing then decreasing, on average.

```
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)
    #could also use
    #haven::read_stata(my_file)
    #for questions, type ?read_stata or ?haven::read_stata
dim(dat)
```

```
## [1] 1000   12
```

```
#View(dat)
EXPER30 = dat$exper - 30
dat2 <- cbind(dat, EXPER30)

library(ggplot2)
ggplot(data = dat2, aes(x = EXPER30, y = wage)) +
  geom_point()
```



b

Estimate by least squares the quadratic model $WAGE = \gamma_1 + \gamma_2(EXPER30)^2 + e$. Are the coefficient estimates statistically significant? Test the null hypothesis that $\gamma_2 \geq 0$ against the alternative that $\gamma_2 < 0$ at the $\alpha = .05$ level of significance. What conclusions do you draw?

Solution Coefficient estimates are statistically significant at .05 level (for two sided test); get this for free from the regression. We see that the estimate is negative and thus know the one-sided test must also reject the null! Nevertheless, I do the test below. Our test statistic is $\hat{\gamma}_2/se(\hat{\gamma}_2)$. Test against $t_{.05,998}$. Reject the null.

```
reg_exp30_sq <- lm(data = dat2, wage ~ I(EXPER30^2))


s <- summary(reg_exp30_sq)
s
```

```
##
## Call:
## lm(formula = wage ~ I(EXPER30^2), data = dat2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.662  -8.515  -3.196   5.463  54.706
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  23.066945   0.526596  43.804  < 2e-16 ***
## I(EXPER30^2) -0.013828   0.001956  -7.068 2.94e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.53 on 998 degrees of freedom
## Multiple R-squared:  0.04767,    Adjusted R-squared:  0.04672
## F-statistic: 49.96 on 1 and 998 DF,  p-value: 2.943e-12
```

```
test_stat = coef(s)[2,1]/coef(s)[2,2]
test_stat == coef(s)[2,3] #same
```

```
## [1] TRUE
```

```
test_stat
```

```
## [1] -7.068312
```

```
qt(.05,998)
```

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

```
test_stat < qt(.05,998) #true, so reject null
```

```
## [1] TRUE
```

```
#can look at all of coef(s) by just:
coef(s)
```

```
##                 Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept)  23.06694471 0.526596225 43.803855 1.181290e-234
## I(EXPER30^2) -0.01382828 0.001956377 -7.068312  2.942799e-12
```

```
#calculating pvalue
#alternative way to perform test
p_value <- pt(q = coef(s)[2, 3], s$df[2], lower=T)
  # or lower = F if wanted to test a right tail test
  #s$df[2] #the df of the regression
  #this is faster than calculating everything by hand
p_value
```

```
## [1] 1.471399e-12
```

```
# This is less than .05 so reject null
# note this is half the value of Pr(>|t|) in summary
# because t-distributions are symmetric.
```

c

Using the estimation in part (b), compute the estimated marginal effect of experience upon wage for a person with 10 years' experience, 30 years' experience, and 50 years' experience. Are these slopes significantly different from zero at the $\alpha = .05$ level of significance?

Solution Since it's squared, the slope is $2 \times \hat{\gamma_2} \times EXPR30$. For the significance tests, we use 3.13 from page 117 of the text.

$$t = \frac{2 \times EXPER30 \times \hat{\gamma_2} - 0}{se(2 \times EXPER30 \times \hat{\gamma_2})} \sim t_{N-2}$$

If $EXPER30 \neq 0$, then

$$\frac{2 \times EXPER30 \times \hat{\gamma_2} - 0}{se(2 \times EXPER30 \times \hat{\gamma_2})} = \frac{\hat{\gamma_2} - 0}{se(\hat{\gamma_2})}$$

, and the significance tests are the same as the coefficients being different from 0. Thus, these are significant for 10 and 50 years of experience.

```
####################################
#slopes

#10 years
2*s$coef[2,1]*(10-30)
```

```
## [1] 0.5531314
```

```
#30 years
2*s$coef[2,1]*(30-30) #0 by design, wage maximized here
```

```
## [1] 0
```

```
#50 years
2*s$coef[2,1]*(50-30)
```

```
## [1] -0.5531314
```

```
####################################
#Significance tests

#30 years must be 0, by assumption

#10 and 50 years have same test as coefficient gamma_2.
#same as part b.
```

d

Construct 95% interval estimates of each of the slopes in part (c). How precisely are we estimating these values? *Note, precision and accuracy are not the same thing!*

Solution. Following pattern on page 116 for CI. That is, $\hat{\lambda} \pm t_c se(\hat{\lambda})$. Note, $se(2a\hat{\gamma_2}) = \sqrt{var(2a\hat{\gamma_2})} = \sqrt{4a^2 var(\hat{\gamma_2})} = 2a \times se(\hat{\gamma_2})$

```
t_95 <- qt(.975,df= nrow(dat)-2)


#10 years
2*s$coef[2,1]*(10-30) + t_95*2*s$coef[2,2]*(10-30)
```

```
## [1] 0.399568
```

```
2*s$coef[2,1]*(10-30) - t_95*2*s$coef[2,2]*(10-30)
```

```
## [1] 0.7066948
```

```
#30 years 0 by design

#50 years
2*s$coef[2,1]*(50-30) + t_95*2*s$coef[2,2]*(50-30)
```

```
## [1] -0.399568
```

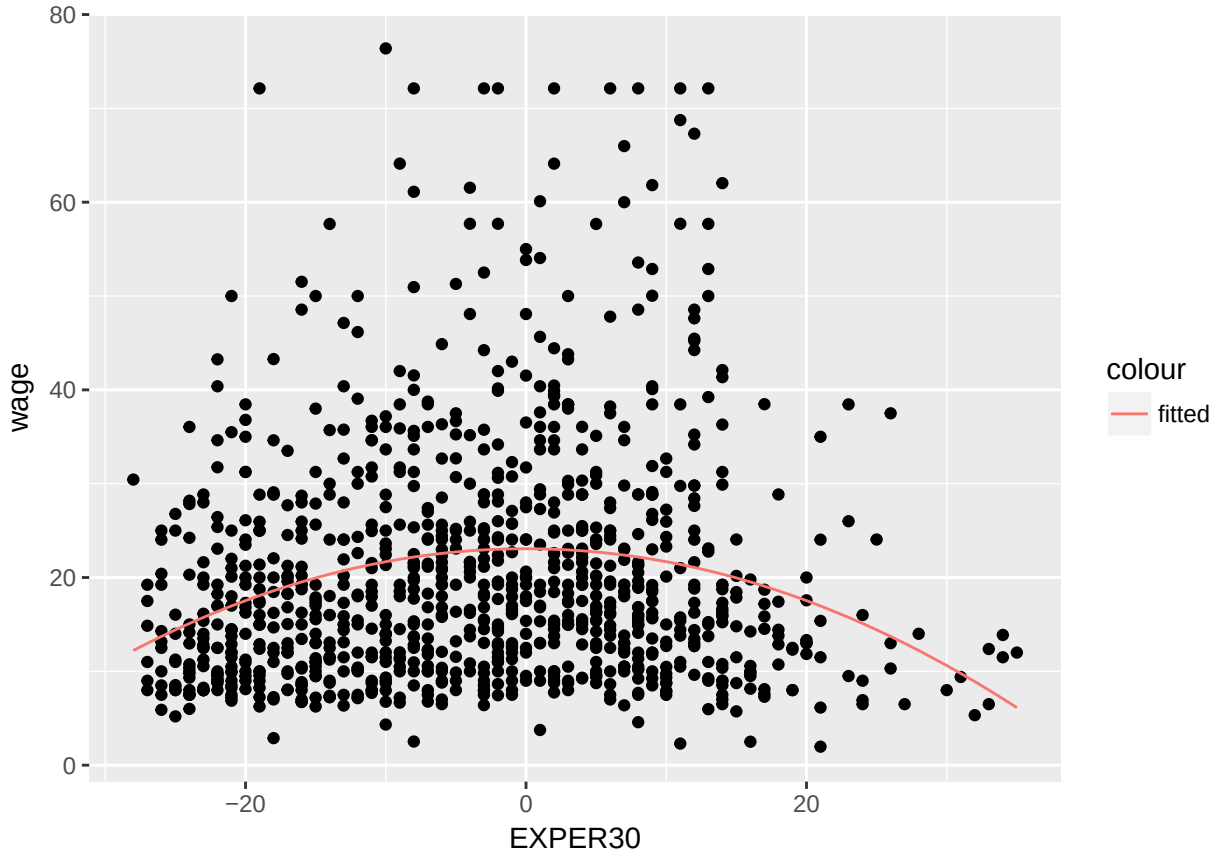
```
2*s$coef[2,1]*(50-30) - t_95*2*s$coef[2,2]*(50-30)
```

```
## [1] -0.7066948
```

e

Using the estimation result from part (b) create the fitted values $W\hat{A}GE = \hat{\gamma}_1 + gam\hat{m}a_2(EXPER30)^2$, where the hat denotes the least squares estimates. Plot these fitted values and $WAGE$ on the vertical axis of the same graph against EXPER30 on the horizontal axis. Are the estimates in part (c) consistent with the graph?

```
library(ggplot2)

my_predicted <- predict(reg_exp30_sq, dat2)
dat3 <- cbind(dat2, my_predicted)
ggplot(data = dat2, aes(x = EXPER30, y = wage)) +
  geom_point() +
 geom_line(data = dat3, mapping =
      aes(x = EXPER30, y = my_predicted, col = 'fitted') )
```



f

Estimate the linear regression $WAGE = \beta_1 + \beta_2 EXPER30 + e$ and the linear regression $WAGE = \alpha_1 + \alpha_2 EXPER + e$. What differences do you observe between these regressions and why do they occur?

What is the estimated marginal effect of experience on wage from these regressions? Based on your work in parts (b)-(d), is the assumption of constant slope in this model a good one? Explain.

Solution As is shown below, the regressions only differ in the values of the intercept. (Note, of course, this wouldn't be the case if we compared $WAGE = \alpha_1 + \alpha_2 EXPER^2 + e$ vs $WAGE = \gamma_1 + \gamma_2 EXPER30^2 + e$. It's a special result because of the *linearity* of estimate and *linearity* of transform EXPER30 from EXPER).

The estimated marginal effect of experience on wage is .08895. The fact that it's positive (and statistically significant) suggests the upward trend for first 30 years is stronger (on average) than the post-30 downward trent. Note that it's between the very positive slope estimate at 10 years and negative slope estimate at 50 years in part c.

Constant slope does not appear to match with the scatter plots. Note, that this did not stop from getting a statistically significant answer, however!

```
lin_reg_exper30 <- lm(data = dat2, wage ~ EXPER30)
lin_reg_exper <- lm(data = dat2, wage ~ exper)

summary(lin_reg_exper)
```

```
##
## Call:
## lm(formula = wage ~ exper, data = dat2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.824  -9.224  -3.429   5.422  56.353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.25768    0.92733  19.688  < 2e-16 ***
## exper        0.08895    0.03148   2.826  0.00481 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.79 on 998 degrees of freedom
## Multiple R-squared:  0.007937,   Adjusted R-squared:  0.006943
## F-statistic: 7.985 on 1 and 998 DF,  p-value: 0.004812
```

```
summary(lin_reg_exper30)
```

```
##
## Call:
## lm(formula = wage ~ EXPER30, data = dat2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.824  -9.224  -3.429   5.422  56.353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.92629    0.41913  49.928  < 2e-16 ***
## EXPER30      0.08895    0.03148   2.826  0.00481 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.79 on 998 degrees of freedom
```

```
## Multiple R-squared:  0.007937,   Adjusted R-squared:  0.006943
## F-statistic: 7.985 on 1 and 998 DF,  p-value: 0.004812
```

g

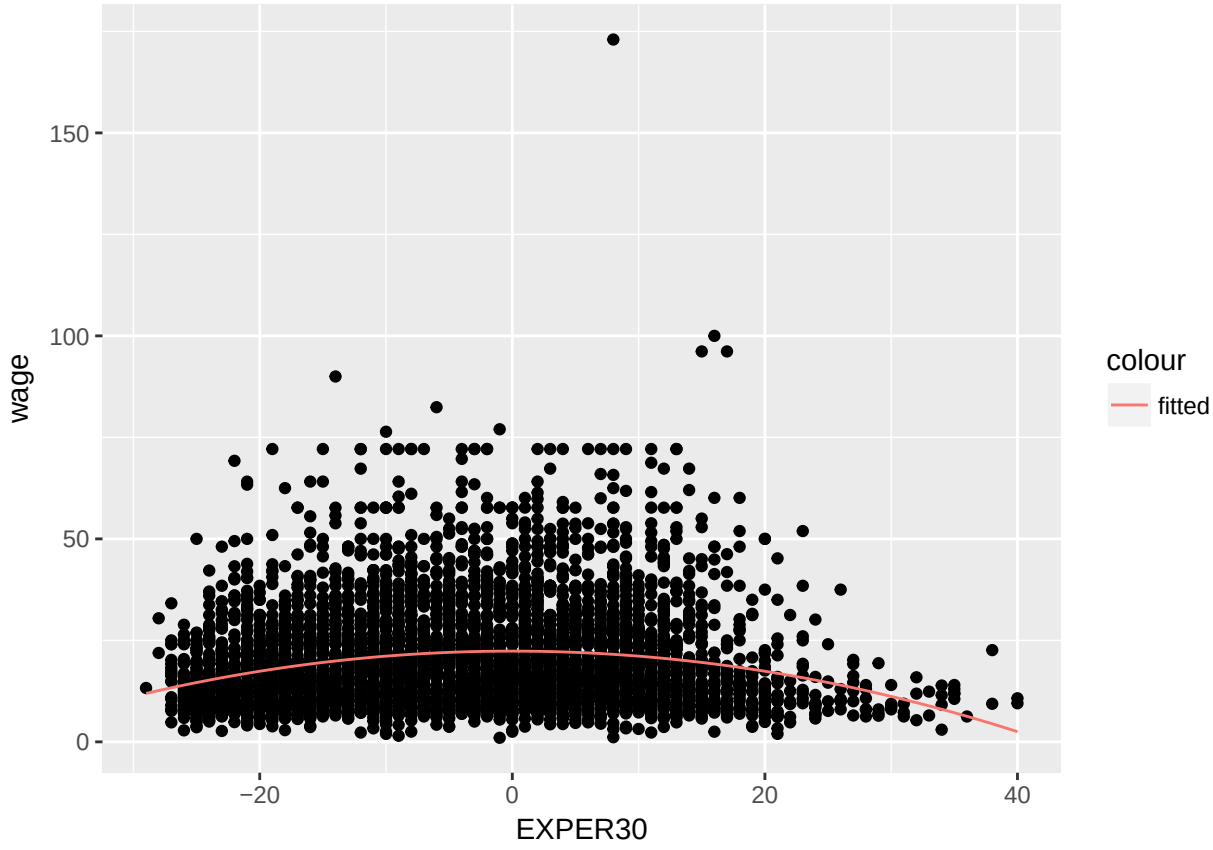
Use the larger data `cps4.dat` (4838 observations) to repeat parts (b), (c) and (d). How much has the larger sample improved the precision of the interval estimates in part (d)?

Solution: This is easy with copy and paste!

```
################################################
## reading in bigger dataset
my_wd <- "C:/Users/ryanj/Dropbox/TA/Econ 103/Winter 2018/Data/s4poe_statadata"
my_file <- paste(my_wd, "cps4.dta", sep = "/")

#library(haven)
dat <- read_stata(my_file) #got the big dataset and
        #didn't even have to pay
EXPER30 = dat$exper - 30
dat2 <- cbind(dat, EXPER30)

library(ggplot2)
pp <- ggplot(data = dat2, aes(x = EXPER30, y = wage)) +
  geom_point() ##saved as pp to reuse later

################################################
## redo b

reg_exp30_sq <- lm(data = dat2, wage ~ I(EXPER30^2))


s <- summary(reg_exp30_sq)
p_value <- pt(q = coef(s)[2, 3], s$df[2], lower=FALSE)
p_value
```

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

```
################################################
#have a look
my_predicted <- predict(reg_exp30_sq, dat2)
dat3 <- cbind(dat2, my_predicted)
pp + geom_line(data = dat3, mapping =
                 aes(x = EXPER30, y = my_predicted, col = 'fitted') )
```




```
#looks good!

#####################################################
## redo c

#10 years
2*s$coef[2,1]*(10-30)
```

```
## [1] 0.4957226
```

```
#30 years
2*s$coef[2,1]*(30-30) #0 by design, wage maximized here
```

```
## [1] 0
```

```
#50 years
2*s$coef[2,1]*(50-30)
```

```
## [1] -0.4957226
```

```
#####################################################
### redo d

t_95 <- qt(.975,df= nrow(dat)-2)


#10 years
2*s$coef[2,1]*(10-30) + t_95*2*s$coef[2,2]*(10-30)
```

```
## [1] 0.4267858
```

```
2*s$coef[2,1]*(10-30) - t_95*2*s$coef[2,2]*(10-30)
```

```
## [1] 0.5646594
```

```
#30 years 0 by design

#50 years
2*s$coef[2,1]*(50-30) + t_95*2*s$coef[2,2]*(50-30)
```

```
## [1] -0.4267858
```

```
2*s$coef[2,1]*(50-30) - t_95*2*s$coef[2,2]*(50-30)
```

```
## [1] -0.5646594
```