Chapter 4 Problem Set

Tianwei Liu

10/4/2019

```
load("Ch4_Exercise3_Presidents_and_Economy.RData")
#load("Ch4_Exercise5_Height_and_Wages_UK.RData") ## we will load the other dataset later
```

3. Voters care about the economy, often more than any other issue. It is not surprising, then, that politicians invariably argue that their party is best for the economy. Who is right? In this exercise, we'll look at the U.S economic and presidential party data in PresPartyEconGrowth.dta to test if there is any difference in economic performance between Republican and Democratic presidents. We will use two different dependent variables: • ChangeGDPpc is the change in real per capita GDP in each year from 1962 to 2013 (in inflation-adjusted U.S. dollars, available from the World Bank). • Unemployment is the unemployment rate each year from 1947 to 2013 (available from the U.S. Bureau of Labor Statistics). Our independent variable is LagDemPres. This variable equals 1 if the president in the previous year was a Democrat and 0 if the president in the previous year was a Republican. The idea is that the president's policies do not take effect immediately, so the economic growth in a given year may be influenced by who was president the year before.

(a) Estimate a model with Unemployment as the dependent variable and LagDemPres as the independent variable. Interpret the coefficients.

```
reg1 <- lm (Unemployment ~ LagDemPresident, data = dta)
summary (reg1)
```

```
##
## Call:
## lm(formula = Unemployment ~ LagDemPresident, data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3806 -1.2084 -0.3806  0.7916  4.3194
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.2361     0.2639  23.630   <2e-16 ***
## LagDemPresident  -0.9555     0.3880  -2.463   0.0164 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.583 on 65 degrees of freedom
## Multiple R-squared:  0.08534,    Adjusted R-squared:  0.07127
## F-statistic: 6.065 on 1 and 65 DF,  p-value: 0.01645
```

The coefficient of the intercept is 6.236, meaning that when the president previous year is republican, the unemployment rate is 6.236, and this coefficient is very statistically significant ($p<0.001$). The coefficient of LagDemPresident is -0.955, meaning that if the president previous year is democrat, the unemployment rate is 0.955 less than if the president was a republican, so at 5.281, and the statistic is significant ($p<0.05$).

(b) Estimate a model with ChangeGDPpc as the dependent variable and LagDemPres as the independent variable. Interpret the coefficients. Explain why the sample size differs from the first model.

```r
reg2 <- lm(ChangeGDPpc ~ LagDemPresident, data = dta)
summary(reg2)
```

```
##
## Call:
## lm(formula = ChangeGDPpc ~ LagDemPresident, data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2119.3  -217.0   110.3   403.1  1209.9
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)        481.1      111.4   4.318 7.44e-05 ***
## LagDemPresident    220.0      164.0   1.341    0.186
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 589.6 on 50 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.03474,    Adjusted R-squared:  0.01543
## F-statistic: 1.799 on 1 and 50 DF,  p-value: 0.1858
```

The coefficient of the intercept is 481, meaning that when the president previous year was republican, the change in GDP per capita is 481; this coefficient is very statsitically significant with p value less than 0.001. The coefficient of LagDemPresident on ChangeGDPpc is 220, meaning that if the previous year president was democrat, GDP per capica will be 220 more than if the previous president was repubilc, so at 701. However, the coefficient is not staitically significant.

(c) Choose an alpha level and alternative hypothesis, and indicate for each model above whether you accept or reject the null hypothesis. alpha = 0.05

Hnull: There is no effect (Beta_1 = 0) Halternative: There is an effect (Beta_1 is not equal to 0)

In this first model (Unemployment ~ LagDemPresident), we reject the null as the p-value of the estimated Beta_1 is 0.016 < 0.05. We reject the null hypothesis that Beta_1 = 0 and accept the alternative. In the second model (ChangeGDPpc ~ LagDemPresident), we do not reject the null as the p-value of the estimated Beta_1 is 0.19 > 0.05.

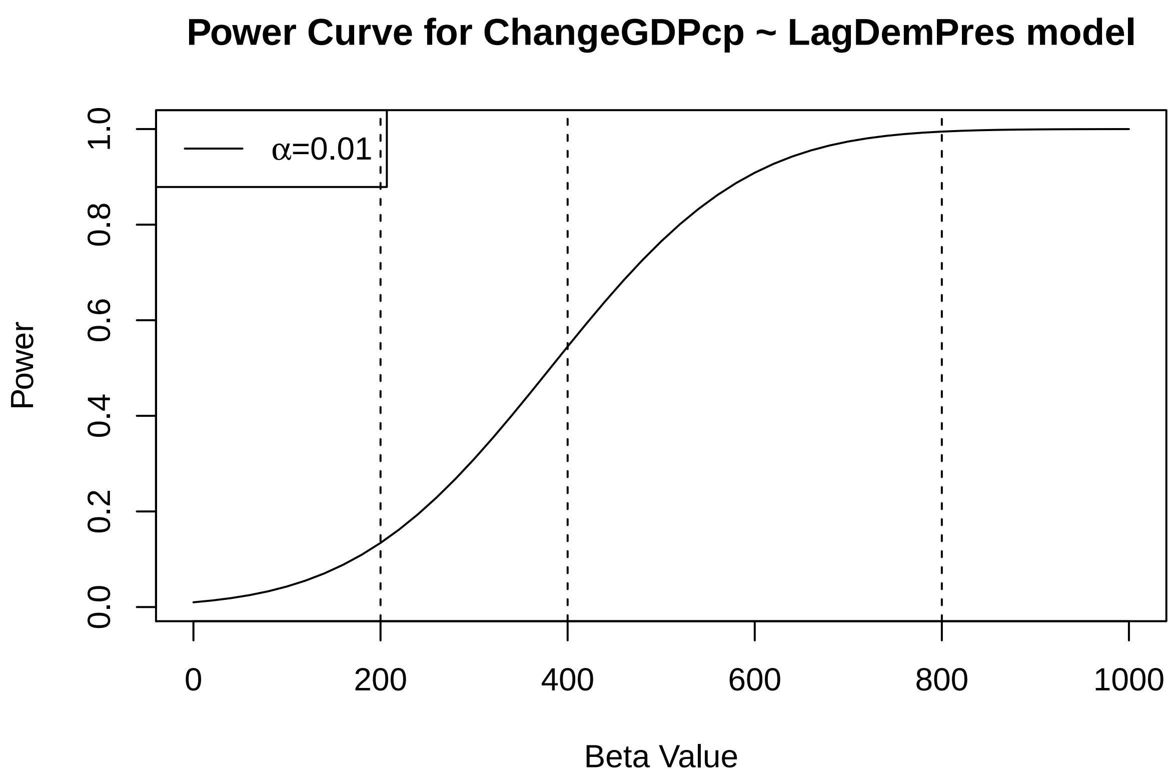
(d) Explain in your own words what the p value means for the LagDemPres variable in each model.

The p value refers to the probability as extreme as we can actually observe if the null hypothesis is true. In the first model (Unemployment ~ LagDemPresident), the largest probability we can actually observe if Beta_1 = 0 is 0.016. In the second model (ChangeGDPpc ~ LagDemPresident), the largest probability we can actually observe if Beta_1 = 0 is 0.19.

(e) Create a power curve for the model with ChangeGDPpc as the dependent variable for alpha = 0.01 and a one-sided alternative hypothesis. Explain what the power curve means by indicating what the curve means for true Beta1 = 200, 400, and 800. Use the code in the Computing Corner, but with the actual standard error of from the regression output.

```
Betarange = seq(0, 1000, 20)
se <- sqrt(vcov(reg2)[2, 2])
power_0.01 <- 1 - pnorm(qnorm(0.99, mean = 0, sd = 1), mean = Betarange/se, sd = 1)
plot (Betarange, power_0.01, main="Power Curve for ChangeGDPcp ~ LagDemPres model", xlab = "Beta Value"
legend ("topleft", c(expression(paste(alpha, "=0.01"))), lty = 1)

abline (v = 200, lty = 2)
abline (v = 400, lty = 2)
abline (v = 800, lty = 2)
```



When true Beta_1 is 200, we have roughly 0.1 probability committing a type I error (reject the null when it's in fact true). When true Beta_1 is 400, the power rises and we have roughly 0.5 probability committing a type I error. When true Beta_1 is 800, the power converges to 1 and we almost always reject the null (if we don't, we commit a type I error).

(f) Discuss the implications of the power curve for the interpretation of the results for the model in which ChangeGDPpc was the dependent variable.

The estimated Beta_1 we got from our model is 220, and the p-value is 0.19. Therefore, if Beta_1 is indeed 220, the power is 0.19, meaning that we have a probability of 0.19 committing a type I error. Therefore, we do not rejec the null. However, this left us with a probability of 0.81 committing a type II error, as power = 1 - pr(typeII error). An insignificant with a lower power is not particularly interesting, but an insignificant statistic with a large power would potentially be important finding.

4. Run the simulation code in the initial part of the education and salary question from the Exercises in Chapter 3 (page 87).

(a) Generate t statistics for the coefficient on education for each simulation. What are the minimal and maximal values of these t statistics?

```
## Set model and simulation parameters
set.seed(50) ## Save the results when the reps is equal to 50
Obs        = 100     # Number of observations in each simulation
Reps       = 50      # Number of times we run the simulation
TrueBeta0   = 12000 # "True" beta0 for the simulated
TrueBeta1   = 1000  # "True" beta1 for the simulated
SD          = 10000 # The standard deviation of the error. The bigger this is, the larger the average v
Ed = 16 * runif(Obs)# Simulate years of education as being between 0 and 16
  # "runif" is a uniform random variable between 0 and 1, with all values having equal probability
CoefMatrix  = matrix(NA, Reps, 2)   # Matrix to store our results.

# Loop: repeat the commands between the brackets multiple times
for (ii in 1:Reps) {
  Salary    = TrueBeta0+ TrueBeta1* Ed + SD*rnorm(Obs)
  OLS.result = lm(Salary ~ Ed) # Run a regression using simulated values of Y
  CoefMatrix[ii,1]  = summary(OLS.result)$coefficients[,3][2]    # Put OLS.result t statistics on educa
  CoefMatrix[ii,2]  = summary(OLS.result)$coefficients[,4][2]  # Put OLS.result p-value on education in
}                             # This closes the "loop"

c(mean(CoefMatrix[,1]), min(CoefMatrix[,1]), max(CoefMatrix[,1]))
```

```
## [1] 4.449030 2.391940 6.789105
```

```
# Average, min and max of t statistics
```

The minimum of the t-statistic is 2.39 and the maximum is 6.79.

(b) Generate two-sided p values for the coefficient on education for each simulation. What are the minimal and maximal values of these p values?

```
c(mean(CoefMatrix[,2]), min(CoefMatrix[,2]), max(CoefMatrix[,2]))
```

```
## [1] 1.031400e-03 8.722912e-10 1.866717e-02
```

The minimum p-value is 8.72 * 10ˆ-10 and the maximum is 1.87*10ˆ-2.

(c) In what percent of the simulations do we reject the null hypothesis that Beta_Education = 0 at the alpha = 0.05 level with a two-sided alternative hypothesis?

We reject 100% of the simulations based on alpha = 0.05, because the maximum of the p-value from the previous question is less than 0.05.

(d) Re-run the simulations, but set the true value of Beta_Education to zero. Do this for 500 simulations, and report what percent of time we reject the null at the alpha = 0.05 level with a two-sided alternative hypothesis. The code provided in Chapter 3 provides tips on how to do this.

```r
set.seed(500)
Obs       = 100     # Number of observations in each simulation
Reps      = 500     # Number of times we run the simulation
TrueBeta0   = 12000 # "True" beta0 for the simulated
TrueBeta1   = 0 # "True" beta1 for the simulated
SD          = 10000 # The standard deviation of the error. The bigger this is, the larger the average v
Ed = 16 * runif(Obs)# Simulate years of education as being between 0 and 16
  # "runif" is a uniform random variable between 0 and 1, with all values having equal probability
CoefMatrix2 = matrix(NA, Reps, 2)   # Matrix to store our results.

# Loop: repeat the commands between the brackets multiple times
for (ii in 1:Reps) {
  Salary    = TrueBeta0+ TrueBeta1* Ed + SD*rnorm(Obs)
  OLS.result = lm(Salary ~ Ed) # Run a regression using simulated values of Y
  CoefMatrix2[ii,1] = summary(OLS.result)$coefficients[,3][2]    # Put OLS.result t statistics on educa
  CoefMatrix2[ii,2]  = summary(OLS.result)$coefficients[,4][2]  # Put OLS.result p-value on education i
}                             # This closes the "loop"

c(mean(CoefMatrix2[,2]), min(CoefMatrix2[,2]), max(CoefMatrix2[,2]))
```

```
## [1] 4.990349e-01 9.265290e-06 9.944119e-01
```

```r
CoefMatrix2_condition <- CoefMatrix2[,2]<0.05
length(CoefMatrix2_condition[CoefMatrix2_condition == TRUE])/Reps
```

```
## [1] 0.056
```

In this case, 5.6 percent of time we reject the null at alpha = 0.05.

5. We will continue the analysis of height and wages in Britain from the Exercises in Chapter 3 (page 88).

(a) Estimate the model with income at age 33 as the dependent variable and height at age 33 as the independent variable. (Exclude observations with wages above 400 British pounds per hour and height less than 40 inches.) Interpret the t statistics on the coefficients.

```r
load("Ch4_Exercise5_Height_and_Wages_UK.RData")
```

```r
dta_subset <- subset (dta, (dta$gwage33 <= 400) & (dta$height33>=40))
reg3 <- lm (gwage33 ~ height33, data = dta_subset)
summary(reg3)
```

```
##
## Call:
## lm(formula = gwage33 ~ height33, data = dta_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -9.632  -3.835  -2.014   0.356 157.690
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.34591    5.01799  -1.862 0.062616 .
## height33     0.26810    0.07199   3.724 0.000199 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.04 on 3667 degrees of freedom
## Multiple R-squared:  0.003768,   Adjusted R-squared:  0.003496
## F-statistic: 13.87 on 1 and 3667 DF,  p-value: 0.0001989
```

Given the rule of the thumb, the absolute value of a t statistic larger than 2 will be considered significant. In this case, the t statistic on the intercept coefficient is -1.86, therefore not significant. In contrast, the t statistic on height33 is 3.72, greater than 2, therefore significant.

(b) Explain the p values for the two estimated coefficients. P value for the intercept coefficient is 0.0626 > 0.05, so not significant. P value for the height coefficient is 0.0002 < 0.001 so it is very significant.

(c) Show how to calculate the 95 percent confidence interval for the coefficient on height.

```
center <- summary(reg3)$coefficients[2,1]
n <- 3669
error <- qnorm(0.975)*summary(reg3)$coefficients[2,2]
left <- center - error
right <- center + error
left
```

```
## [1] 0.1270013
```

```
right
```

```
## [1] 0.4091956
```

The confidence interval for the coeffieicnet on height is [0.127, 0.409].

(d) Do we accept or reject the null hypothesis that Beta1 = 0 for alpha = 0.01 and a two-sided alternative? Explain your answer.

We reject the null hypothesis because the p value is 0.0002 for the coefficient on height which is smaller than alpha = 0.01.

(e) Do we accept or reject the null hypothesis that Beta0 = 0 (the constant) for alpha = 0.01 and a two-sided alternative? Explain.

We do not reject the null that Beta0 = 0 becasue the p value for the coefficient on intercept is 0.0626 which is greater than alpha = 0.01.

(f) Limit the sample size to the first 800 observations. Do we accept or reject the null hypothesis that Beta1 = 0 for alpha = 0.01 and a two-sided alternative? Explain if/how/why this answer differs from the earlier hypothesis test about Beta1.

```
reg4 <- lm (gwage33 ~ height33, data = dta[1:800,])
summary(reg4)
```

```
##
## Call:
## lm(formula = gwage33 ~ height33, data = dta[1:800, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -14.25   -7.19   -4.68   -2.24 2484.08
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -56.6775    56.2209  -1.008    0.314
## height33      0.9914     0.8110   1.222    0.222
##
## Residual standard error: 88.76 on 798 degrees of freedom
## Multiple R-squared:  0.001869,   Adjusted R-squared:  0.0006182
## F-statistic: 1.494 on 1 and 798 DF,  p-value: 0.2219
```

If we limit the sample size to the first 800 observations, we do not reject the null hypothesis for both Beta0 and Beta1 because the p values of both coefficients are greater than 0.01.