Geneva Graduate Institute, Econometrics I

Problem Set 3 Solutions

Francesco Casalena

Fall 2024

Problem 1

You can find the data set for this question and a description of the variables on Moodle. The data spreadsheet contains four variables: average hourly earnings *ahe*, age *age*, gender *female*, and education *bachelor*. To answer the questions, use the asymptotic distribution of the OLS/ML estimator to conduct hypothesis tests or generate 95% confidence intervals. In your quantitative statements, be mindful of the units (e.g. dollars vs. percentages).

1. Do males on average earn more than females? Do individuals with a college degree earn on average more than individuals without? How large are the wage differentials?

Solution:

```
rm(list=ls())
data <- read.csv("dat_CPS08.csv", header = TRUE)

mandata <- data[ which(data$female==0), ]
womandata <- data[ which(data$female==1), ]

coldata <- data[ which(data$bachelor==1), ]
nocoldata <- data[ which(data$bachelor==0), ]

mean(mandata$ahe)
```

```
## [1] 20.11387
```

```
mean(womandata$ahe)
```

```
## [1] 17.48396
```

```
mean(mandata$ahe) - mean(womandata$ahe)#difference btw genders
```

```
## [1] 2.629912
```

```
mean(coldata$ahe)
```

```
## [1] 22.90834
```

```
mean(nocoldata$ahe)
```

```
## [1] 15.33174
```

```
mean(coldata$ahe)-mean(nocoldata$ahe) #difference btw degrees
```

```
## [1] 7.576594
```

On average, males earn more than females by $ 2.63 and individuals with a college degree earn more than individuals without by $ 7.58. To see whether the differences are significant (i.e. whether indeed we can reject the null hypothesis that the true averages in the population are the same across two groups), we can use the t-test, which is derived in the following. The analysis talks about earnings of males vs females, but the analogous applies for individuals with a college degree vs without.

Let $x_i \sim N(\mathbb{E}[x_i], \mathbb{V}[x_i])$ be earnings of males. We thus have the finite-sample distribution of the estimator for the mean:

$$\bar{X} \equiv \frac{1}{n_x}\sum_{i=1}^{n_x} x_i \sim N(\mu_x, v_x)\ , \quad \mu_x = \mathbb{E}[x_i]\ ,\ v_x = \frac{1}{n_x}\mathbb{V}[x_i]\ ,\ n_x = \#\text{ of males in sample}$$

Hence:

$$\bar{X} \overset{approx.}{\sim} N(\mu_X, \hat{v}_x) \ , \quad \hat{v}_x = \frac{1}{n_x}\widehat{\mathbb{V}[x_i]} \ , \ \widehat{\mathbb{V}[x_i]} = \frac{1}{n_x}\sum_{i=1}^{n_x}(x_i - \bar{x})^2 \ .$$

Analogously, we have $\bar{Y} \equiv \frac{1}{n_y}\sum_{i=1}^{n_y} y_i \sim N(\mu_y, v_y)$ for the earnings of females, which we can denote by y_i.

Using this result, we could test $\mathcal{H}_0 : \mu_x = \mu_{x,0}$ (some specific value, e.g. 0, or 20) against $\mathcal{H}_1 : \mu_X \neq \mu_{x,0}$. For example, we could do that using the t-test (two-sided):

$$t_x = \frac{\bar{X} - \mu_{x,0}}{\sqrt{\frac{1}{n_x}\widehat{\mathbb{V}[x_i]}}} \overset{approx.}{\sim} N(0,1) \qquad \text{(becomes exact as } \ n_X \to \infty) \ ,$$

and we could compute p-values as $2(1 - \Phi(t_X))$, where Φ is the cdf of $N(0,1)$.

However, here we want to test whether the average difference in earnings between males and females in the relevant population, $\mu_X - \mu_Y = \mathbb{E}[x_i] - \mathbb{E}[y_i]$, is zero. Thus we proceed with analogous steps, just using the distribution of the difference in sample means, $\bar{X} - \bar{Y}$, instead of that of a single sample mean, e.g. $\bar{X}$:

$$\bar{X} - \bar{Y} \sim N(\mu_{x-y}, v_{x-y}) \ , \quad \mu_{x-y} = \mu_X - \mu_Y \ , \ v_{x-y} = \frac{1}{n_x}\mathbb{V}[x_i] + \frac{1}{n_y}\mathbb{V}[y_i] \ .$$

In turn, we get

$$t_{x-y} = \frac{\bar{X} - \bar{Y} - 0}{\sqrt{\hat{v}_{x-y}}} \overset{approx.}{\sim} N(0,1) \ ,$$

where $\hat{v}_{x-y} = \frac{1}{n_x}\widehat{\mathbb{V}[x_i]} + \frac{1}{n_y}\widehat{\mathbb{V}[y_i]}$.

```
# Conduct t-tests manually:

diff1 <- mean(mandata$ahe) - mean(womandata$ahe)
varDiff1 <- var(mandata$ahe)/length(mandata$ahe) + var(womandata$ahe)/length(womandata$ahe)
tStat1 <- diff1 / sqrt(varDiff1)
pVal1 <- 2*(1 - pnorm(tStat1))
tStat1
```

```
## [1] 11.61044
```

```
pVal1
```

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

```
diff2 <- mean(coldata$ahe) - mean(nocoldata$ahe)
varDiff2 <- var(coldata$ahe)/length(coldata$ahe) + var(nocoldata$ahe)/length(nocoldata$ahe)
tStat2 <- diff2 / sqrt(varDiff2)
pVal2 <- 2*(1 - pnorm(tStat2))
tStat2
```

```
## [1] 34.88667
```

```
pVal2
```

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

```
# Conduct t-tests using built-in function in R:

t.test(mandata$ahe,womandata$ahe)
```

```
##
##  Welch Two Sample t-test
##
## data:  mandata$ahe and womandata$ahe
## t = 11.61, df = 7599.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.185884 3.073939
## sample estimates:
## mean of x mean of y
##  20.11387  17.48396
```

```
t.test(coldata$ahe,nocoldata$ahe)
```

```
##
##  Welch Two Sample t-test
##
## data:  coldata$ahe and nocoldata$ahe
## t = 34.887, df = 6599.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  7.150856 8.002332
## sample estimates:
## mean of x mean of y
##  22.90834  15.33174
```

From above, we can see the averages are significantly different both for different genders and for different education levels.

2. Run a regression of earnings on age, gender, and education. If age increases from 28 to 29, how are earnings expected to change? If age increases from 37 to 38, how are earnings expected to change?

Solution:

```
# Compute the OLS estimator manually:
# Generate constant term
data$const <- rep(1)
# Generate y vector
y <- data$ahe
# Generate X matrix
X <- cbind(data$const, data$age, data$female, data$bachelor)
# Compute beta_hat = (X'X)^{-1}X'Y
beta_hat <- solve( t(X) %*% X ) %*% t(X) %*% y

beta_hat
```

```
##             [,1]
## [1,] -0.6356977
## [2,]  0.5852144
## [3,] -3.6640258
## [4,]  8.0830009
```

```
# Use the built-in lm() function
myOLSb <- lm(ahe ~ age+female+bachelor, data=data)
summary(myOLSb)
```

```
##
## Call:
```

```
## lm(formula = ahe ~ age + female + bachelor, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.139  -5.773  -1.509   4.112  57.414
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.6357     1.0854  -0.586    0.558
## age           0.5852     0.0362  16.165   <2e-16 ***
## female       -3.6640     0.2107 -17.391   <2e-16 ***
## bachelor      8.0830     0.2088  38.709   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.072 on 7707 degrees of freedom
## Multiple R-squared:  0.1998, Adjusted R-squared:  0.1995
## F-statistic: 641.5 on 3 and 7707 DF,  p-value: < 2.2e-16
```

The model is:

$$y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_k x_{i,k} + u_i$$

The marginal effects are given by:

$$\frac{\partial y_i}{\partial x_{i,c}} = \beta_c \qquad \text{for any covariate } \; c = 1, 2, ..., k$$

This means that $\frac{\Delta y_i}{\Delta x_{i,c}} \approx \beta_c$, or

$$\underbrace{\Delta y_i}_{\text{level change in } y_i} \approx \beta_c \quad \underbrace{\Delta x_{i,c}}_{\text{level change in } x_i}$$

for small Δx_{i_c}, i.e. small changes in x_{i_c}. Therefore, $x_{i,c}$ is going up by 1 *unit* ($\Delta x_{i,c} = 1$) increases $y_{i,c}$ by β_c *units*. The relationship/effect is independent of the level of $x_{i,c}$. For example, age going up by 1 **year* increases earnings by 0.59 *dollars*, no matter whether age increases from 28 to 29 or from 37 to 38.

Sidenote: the relationship is not only positive in our sample, but it is significantly different from zero, i.e. we can reject the null hypothesis that the true relationship/effect is zero at any common significance level, as indicated by the tiny p-value for a t-test, shown by the lm function.

3. Run a regression of the logarithm of earnings on age, gender, and education. If age increases from 28 to 29, how are earnings expected to change? If age increases from 37 to 38, how are earnings expected to change?

Solution:

```r
# Compute the OLS estimator manually:
# Generate logs
data$log_ahe <- log(data$ahe)
# Generate y vector
y <- data$log_ahe
# Generate X matrix
X <- cbind(data$const, data$age, data$female, data$bachelor)
# Compute beta_hat = (X'X)^{-1}X'Y
beta_hat <- solve( t(X) %*% X ) %*% t(X) %*% y

beta_hat
```

```
##             [,1]
## [1,]  1.87634048
## [2,]  0.02732698
## [3,] -0.18592385
## [4,]  0.42812744
```

```r
# Use the built-in lm() function
myOLSc <- lm(log(ahe) ~ age+female+bachelor, data=data)
summary(myOLSc)
```

```
##
## Call:
## lm(formula = log(ahe) ~ age + female + bachelor, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34755 -0.27810  0.01842  0.30954  1.66410
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.876340   0.056160   33.41   <2e-16 ***
## age          0.027327   0.001873   14.59   <2e-16 ***
## female      -0.185924   0.010901  -17.06   <2e-16 ***
## bachelor     0.428127   0.010804   39.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4694 on 7707 degrees of freedom
## Multiple R-squared:  0.2007, Adjusted R-squared:  0.2003
## F-statistic: 644.9 on 3 and 7707 DF,  p-value: < 2.2e-16
```

Still $\frac{\partial y_i}{\partial x_{i,c}} = \beta_c$, as we use the same linear regression model. However, y_i is equal to the log of the variable we are actually interested in, average hourly earnings: $y_i = log(\tilde{y}_i)$, or $\tilde{y}_i = \exp\{y_i\}$. This means:

$$\frac{\partial \tilde{y}_i}{\partial x_{i,c}} = \frac{\partial \tilde{y}_i}{\partial y_i}\frac{\partial y_i}{\partial x_{i,c}} = \tilde{y}_i \beta_c \,,$$

and therefore

$$\frac{\Delta \tilde{y}_i}{\Delta x_{i,c}} \approx \tilde{y}_i \beta_c \quad \Longleftrightarrow \quad \underbrace{\frac{\Delta \tilde{y}_i}{\tilde{y}_i}}_{\text{percentage change in } \tilde{y}_i} \approx \beta_c \underbrace{\Delta x_{i,c}}_{\text{dis. change in } x_{i,c}} .$$

Hence, $x_{i,c}$ going up by 1 *unit* ($\Delta x_{i,c} = 1$) increases $\tilde{y}_i$ by β_c *percent* ($\frac{\Delta \tilde{y}_i}{\tilde{y}_i} = \beta_c$). For example, age going up by 1 *year* increases earnings by 2.73 *percent*. Again, this percentage effect is independent of the level of $x_{i,c}$. However, note that this means that the absolute effect (in units of $\tilde{y}_i$) depends on the level of y_i. Given that (based on our regression results) an older individual earns on average (everything else equal) more than a younger individual, the expected level change in earnings measured in dollars is higher for an individual going from 37 to 38 years than from 28 to 29. If age increases from 28 to 29 (37 to 38), wage increases by approximately \$0.77 (\$1.01).

4. Run a regression of the logarithm of earnings on gender, education, and the logarithm of age. If age increases from 28 to 29, how are earnings expected to change? If age increases from 37 to 38, how are earnings expected to change?

Solution:

```r
# Compute the OLS estimator manually:
# Generate logs
data$log_age <- log(data$age)
# Generate y vector
y <- data$log_ahe
# Generate X matrix
X <- cbind(data$const, data$log_age, data$female, data$bachelor)
# Compute beta_hat = (X'X)^{-1}X'Y
beta_hat <- solve( t(X) %*% X ) %*% t(X) %*% y

beta_hat
```

```
##             [,1]
## [1,] -0.03452566
## [2,]  0.80390509
## [3,] -0.18588958
## [4,]  0.42825412
```

```r
# Use the built-in lm() function
myOLSd <- lm(log(ahe) ~ log(age)+female+bachelor, data=data)
summary(myOLSd)
```

```
##
## Call:
## lm(formula = log(ahe) ~ log(age) + female + bachelor, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34852 -0.27913  0.02117  0.30921  1.66325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03453    0.18622  -0.185    0.853
## log(age)     0.80391    0.05496  14.626   <2e-16 ***
## female      -0.18589    0.01090 -17.054   <2e-16 ***
## bachelor     0.42825    0.01080  39.641   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4694 on 7707 degrees of freedom
## Multiple R-squared:  0.2008, Adjusted R-squared:  0.2005
## F-statistic: 645.3 on 3 and 7707 DF,  p-value: < 2.2e-16
```

Still $\frac{\partial y_i}{\partial x_{i,c}} = \beta_c$. However, if $x_{i,c} = \log(\tilde{x}_{i,c})$, then :

$$\frac{\partial y_i}{\partial \tilde{x}_{i,c}} = \frac{\partial y_i}{\partial x_{i,c}} \frac{\partial x_{i,c}}{\partial \tilde{x}_{i,c}} = \beta_c \frac{1}{\tilde{x}_{i,c}}$$

This means that:

$$\frac{\Delta y_i}{\Delta \tilde{x}_{i,c}} \approx \beta_c \frac{1}{\tilde{x}_{i,c}} \quad \iff \quad \underbrace{\Delta y_i}_{\text{level change in } y_i} = \beta_c \underbrace{\frac{\Delta \tilde{x}_{i,c}}{\tilde{x}_{i,c}}}_{\text{percentage change in } \tilde{x}_{i,c}}$$

If also $y_i = \log(\tilde{y}_i)$, then, putting the two pieces from the last and present exercises together:

$$\frac{\partial \tilde{y}_i}{\partial \tilde{x}_{i,c}} = \frac{\partial \tilde{y}_i}{\partial y_i}\frac{\partial y_i}{\partial x_{i,c}}\frac{\partial x_{i,c}}{\partial \tilde{x}_{i,c}} = \tilde{y}_i \beta_c \frac{1}{\tilde{x}_{i,c}} \, .$$

This means that:

$$\frac{\partial \tilde{y}_i}{\partial \tilde{x}_{i,c}} \approx \tilde{y}_i \beta_c \frac{1}{\tilde{x}_{i,c}} \quad \Longleftrightarrow \quad \underbrace{\frac{\Delta \tilde{y}_i}{y_i}}_{\text{percentage change in } \tilde{y}_i} \approx \beta_c \underbrace{\frac{\Delta \tilde{x}_{i,c}}{\tilde{x}_{i,c}}}_{\text{percentage change in } \tilde{x}_{i,c}} \, .$$

Hence, if $\tilde{x}_{i,c}$ increases by 1 *percent*, $\tilde{y}_i$ increases by β_c *percent*. In our case, age increasing by 1% increases $\tilde{y}_i$ by 0.8%. Therefore, when age increases from 28 to 29 (37 to 38), this is an increase of 3.57% (2.70%), and so the wage is expected to increase by 2.87% (2.17%).

5. Run a regression of the logarithm of earnings on age, age^2, gender, and education. If age increases from 28 to 29, how are earnings expected to change? If age increases from 37 to 38, how are earnings expected to change?

Solution:

```
# Compute the OLS estimator manually:
# Generate squares
data$agesq <- data$age^2
# Generate y vector
y <- data$log_ahe
# Generate X matrix
X <- cbind(data$const, data$age, data$agesq, data$female, data$bachelor)
# Compute beta_hat = (X'X)^{-1}X'Y
beta_hat <- solve( t(X) %*% X ) %*% t(X) %*% y

beta_hat
```

```
##               [,1]
## [1,]  1.0854297723
## [2,]  0.0813724732
## [3,] -0.0009148162
## [4,] -0.1858687321
## [5,]  0.4283779959
```

```
# Use the built-in lm() function
myOLSe <- lm(log(ahe) ~ age+I(age^2)+female+bachelor, data=data)
summary(myOLSe)
```

```
##
## Call:
## lm(formula = log(ahe) ~ age + I(age^2) + female + bachelor, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34922 -0.27960  0.02046  0.30927  1.66268
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.0854298  0.6382725   1.701   0.0891 .
## age          0.0813725  0.0434864   1.871   0.0614 .
## I(age^2)    -0.0009148  0.0007354  -1.244   0.2135
## female      -0.1858687  0.0109006 -17.051   <2e-16 ***
```

```
## bachelor     0.4283780  0.0108057  39.644   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4694 on 7706 degrees of freedom
## Multiple R-squared:  0.2008, Adjusted R-squared:  0.2004
## F-statistic: 484.1 on 4 and 7706 DF,  p-value: < 2.2e-16
```

Still, $\frac{\partial y_i}{\partial x_{i,c}} = \beta_c$ for any covariate c.

Now, we have the case where one of the covariates is age and one is age squared. Denoting the former covariate by $x_{i,c}$, the latter by $x_{i,d}$ and denoting age by $\tilde{x}_i$, we have $x_{i,c} = \tilde{x}_i$ and $x_{i,d} = \tilde{x}^2$. As a result:

$$\frac{\partial y_i}{\partial \tilde{x}_i} = \frac{\partial y_i}{\partial x_{i,c}}\frac{\partial x_{i,c}}{\partial \tilde{x}_i} + \frac{\partial y_i}{\partial x_{i,d}}\frac{\partial x_{i,c}}{\partial \tilde{x}_i} = \beta_c \times 1 + \beta_d \times 2\tilde{x}_i = \beta_c + 2\beta_d \tilde{x}_i \ .$$

If also $y_i = \log(\tilde{y}_i)$, following the derivations above, we have:

$$\frac{\Delta \tilde{y}_i}{\tilde{y}_i} \approx (\beta_c + 2\beta_d \tilde{x}_i)\Delta \tilde{x}_i \ ,$$

i.e., when $\tilde{x}_i$ increases by one *unit*, from $\tilde{x}_i$ to $\tilde{x}_i + 1$, then $\tilde{y}_i$ increases by $(\beta_c + \beta_d \times 2\tilde{x}_i)$ *percent*. Note that this effect depends on the level of $\tilde{x}_i$.

Therefore, when age increases from 28 to 29 (from 37 to 38) earnings are expected to increase by approximately $(\hat{\beta}_c + 2\hat{\beta}_d 28)1 = 3\%$ $((\hat{\beta}_c + 2\hat{\beta}_d 37)1 = 1.3\%)$.

6. Plot the regression relation (the so-called age-earnings profile) between *age* (on the x-axis) and log *ahe* (on the y-axis) for the age range 20-65 using the estimates from (e) for males with a bachelor degree. At what age does the age-earnings profile peak?

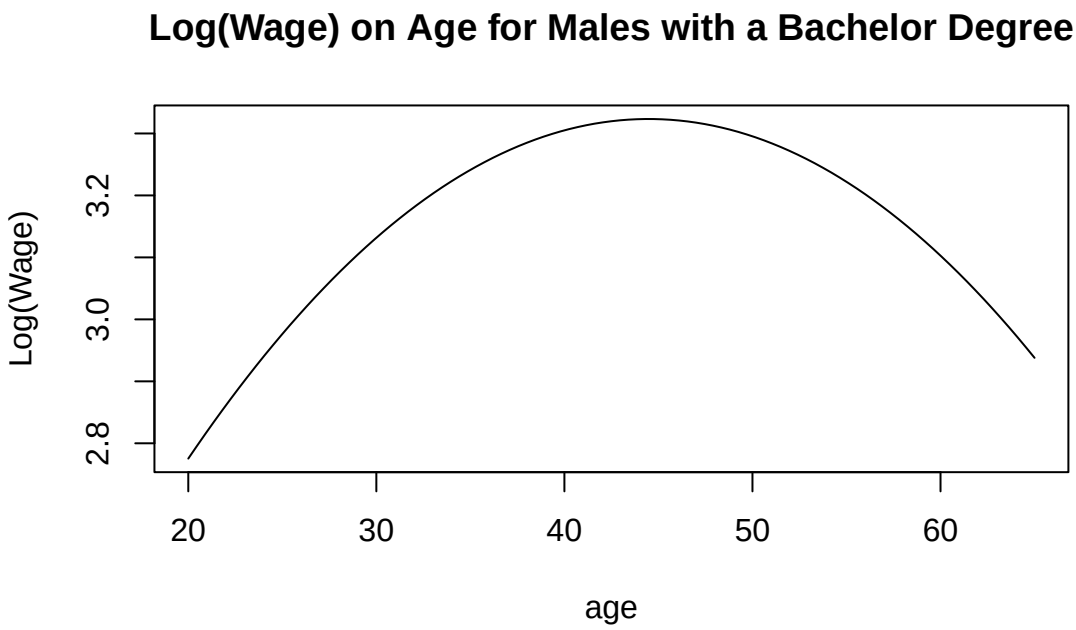
Solution:

We take the estimated regression model for males (i.e. $female_i = 0$) with a bachelor degree (i.e. $bachelor_i = 1$):

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{\beta}_3 x_i^2 + \hat{\beta}_5$$

Where $\hat{y}_i$ are the estimated log average hourly earnings, depending on the age x_i and the estimated coefficients $\hat{\beta}_j$ of the $\hat{\beta}$ vector. We construct a line plot of this function, for all the values of age between 20 to 65 on the x-axis.

```
beta_e <- myOLSe$coefficients
curve(beta_e[1]+beta_e[2]*x+beta_e[3]*x^2+beta_e[4]*0+beta_e[5]*1,from=20,to=65,
xlab="age", ylab="Log(Wage)",
main="Log(Wage) on Age for Males with a Bachelor Degree")
```



Since this is a globally concave function, we can find its maximum by taking the first order condition and solving for x_i:

$$\frac{\partial \hat{y}_i}{\partial x_i} = \hat{\beta}_2 + 2\hat{\beta}_3 x_i = 0 \iff x_i^{\max} = -\frac{\hat{\beta}_2}{2\hat{\beta}_3}$$

```
peak_age <- -beta_e[2]/(2*beta_e[3])
peak_age
```

```
##      age
## 44.47477
```

We can see the relation between log(wage) and age is reverse U-shaped and peaks at age 44-45.

7. Is the effect of age on earnings different for males than for females? Specify and estimate a regression that you can use to answer this question. You can suppose that the relationship between age and log-earnings is linear for both males and females.

Hint: construct a covariate as the interaction $female * age$.

Solution:

Recall the model is $y_i = x_i'\beta + u_i$. If $x_{i,c} = \text{age}$ and $x_{i,d} = \text{age} \times \text{female} = \text{age} \times \mathbf{1}\{\text{i is female}\}$, then:

$$\frac{\partial y_i}{\partial \text{age}_i} = \beta_c + \beta_d \times \mathbf{1}\{\text{i is female}\}$$

If y_i denotes the log of earnings $\tilde{y}_i$, following the derivations from the previous exercises, we get

$$\frac{\Delta \tilde{y}_i}{\tilde{y}_i} \approx (\beta_c + \beta_d \mathbf{1}\{\text{i is female}\})\Delta\text{age}_i$$

i.e. for males ageing by one year, earnings change by β_c percent, while for females the effect is $\beta_c + \beta_d$. Testing $\mathcal{H}_0 : \beta_d = 0$ directly tests whether the effect of age on (log-)earnings is different for males and females.

Let's build the test statistic for a t-test. We have

$$\mathbb{E}[\hat{\beta}] = \beta, \qquad \mathbb{V}[\hat{\beta}] = \sigma^2 \mathbb{E}[X'X]^{-1} = \frac{\sigma^2}{n} \mathbb{E}[x_i x_i']^{-1}$$

(where we used the fact that $X'X = \sum_{i=1}^{n} x_i x_i'$) and therefore

$$\mathbb{E}[\hat{\beta}_d] = \beta_d \quad \text{and} \quad \mathbb{V}[\hat{\beta}_d] = (\sigma^2 \mathbb{E}[X'X]^{-1})_{dd} \, .$$

(Note that the same expression for $\mathbb{V}[\hat{\beta}]$ is obtained when approximating it via the asymptotic distribution of $\hat{\beta}$.) We can estimate $\mathbb{V}[\hat{\beta}]$ and therefore $\mathbb{V}[\hat{\beta}_d]$ by

$$\hat{\mathbb{V}}[\hat{\beta}] = \frac{\hat{\sigma}^2}{n} (\frac{1}{n} \sum_{i=1}^{n} x_i x_i')^{-1} = \hat{\sigma}^2 (\sum_{i=1}^{n} x_i x_i')^{-1} = \hat{\sigma}^2 (X'X)^{-1} \, .$$

We thus have the test statistic:

$$t = \frac{\hat{\beta}_d - 0}{\sqrt{\hat{\mathbb{V}}[\hat{\beta}_d]}} = \frac{\hat{\beta}_d}{\sqrt{\hat{\sigma}^2 [(X'X)^{-1}]_{dd}}} \xrightarrow{d} N(0,1) \, .$$

We can compute the p-value as usual as : $p = 2(1 - \Phi(t))$.

```
# Compute the OLS estimator manually:
# Generate interaction terms
data$age_female <- data$age*data$female
data$agesq_female <- data$agesq*data$female
# Generate y vector
y <- data$log_ahe
# Generate X matrix
X <- cbind(data$const, data$age, data$age_female, data$female, data$bachelor)
# Compute beta_hat = (X'X)^{-1}X'Y
beta_hat <- solve( t(X) %*% X ) %*% t(X) %*% y

beta_hat
```

```
##              [,1]
## [1,]  1.66019772
## [2,]  0.03463648
## [3,] -0.01676664
## [4,]  0.30980214
## [5,]  0.42667287
```

```
# Compute test statistic manually
# Generate predicted values
y_hat <- X %*% beta_hat
# Generate residuals
u_hat <- y - y_hat
# Compute sigma_hat
sig_hat <- (t(u_hat) %*% u_hat)/(nrow(X)-ncol(X))
# Compute (X'X)^{-1}
XXinv <- solve( t(X) %*% X )
# Compute s.e. of beta_hat for the interaction term
se_bet_d <- sqrt(sig_hat * XXinv[3,3])
# Compute test statistic
t_stat <- beta_hat[3]/se_bet_d

t_stat
```

```
##           [,1]
## [1,] -4.442041
```

```r
# Use the built-in lm() function
myOLSh <- lm(log(ahe) ~ age+I(age*female)+female+bachelor,
data=data)
summary(myOLSh)
```

```
##
## Call:
## lm(formula = log(ahe) ~ age + I(age * female) + female + bachelor,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32731 -0.27812  0.01866  0.30528  1.68288
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.660198   0.074256  22.358  < 2e-16 ***
## age               0.034636   0.002492  13.902  < 2e-16 ***
## I(age * female) -0.016767   0.003775  -4.442 9.04e-06 ***
## female            0.309802   0.112129   2.763  0.00574 **
## bachelor          0.426673   0.010796  39.521  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4688 on 7706 degrees of freedom
## Multiple R-squared:  0.2027, Adjusted R-squared:  0.2023
## F-statistic: 489.8 on 4 and 7706 DF,  p-value: < 2.2e-16
```

We obtain a t-statistic of -4.44, which means that the negative effect of $female * age$ is significantly different from zero, i.e. the effect of age on earnings is higher for males than for females and this difference is statistically significant.