Geneva Graduate Institute, Econometrics I

Problem Set 4 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.

1. Run a regression of the logarithm of earnings on age, age^2, $female$, and $bachelor$. Based on your results, what are the predicted log-earnings of a 30 year old female with a bachelor degree? Note that you can write the quantity of interest as

$$\mathbb{E}[y_i|age = 30, female = 1, bachelor = 1] = \tilde{x}_i'\beta\ , \quad \text{where} \quad \tilde{x}_i = [1, 30, 30^2, 1, 1]'\ .$$

Solution:

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

# Compute OLS manually
y <- log(data$ahe)
X <- cbind(rep(1), data$age, data$age^2, data$female, data$bachelor)
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 built-in function
myOLSa <- lm(log(ahe) ~ age + I(age^2) + female + bachelor, data = data)
summary(myOLSa)
```

```
## 
## 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
```

```
# Compute predicted value
x_fem30 <- c(1, 30, 900, 1, 1)
y_fem30 <- t(x_fem30) %*% beta_hat
y_fem30
```

```
##          [,1]
## [1,] 2.945779
```

The log average hourly wage for 30-year old females with a bachelor's degree is 2.95. This amounts to $19.1.

2. Using a t-test and a significance level of $\alpha = 0.05$, can you reject the null hypothesis that the expected hourly earnings of a 30 year old female with a bachelor degree are equal to 20 dollars per hour (i.e. that the expected log-earnings are equal to $\ln 20 \approx 2.99$)?

Hint: Note that we can write $\mathcal{H}_0 : \tilde{x}_i'\beta = 2.99$, with $\tilde{x}_i$ as defined above. Based on the (asymptotic) distribution of β, you can find that of $\tilde{x}_i'\beta$, which allows you to construct a t-test for that quantity.

Solution:

To test $\mathcal{H}_0 : \hat{y}_{fem30} = 2.99$ against the alternative $\mathcal{H}_1 : \hat{y}_{fem30} \neq 2.99$, we use the test statistic:

$$T(X) = \left| \frac{\hat{y}_{fem30} - 2.99}{SE(\hat{y}_{fem30})} \right| \sim N(0,1)$$

The standard error of a predicted value in a linear regression model is given by:

$$SE(\hat{y}_i) = SE(\tilde{x}_i'\hat{\beta}) = \sqrt{\widehat{\mathbb{V}[\tilde{x}_i'\hat{\beta}]}} = \sqrt{\tilde{x}_i'\widehat{var[\hat{\beta}]}\tilde{x}_i} = \sqrt{\tilde{x}_i'\hat{\sigma}^2(X'X)^{-1}\tilde{x}_i} \,,$$

where $\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}\hat{u}_i^2$.

```
# Construct prediction error for y_fem30
y_hat <- X %*% beta_hat # Predicted values
u_hat <- y - y_hat # Sample residuals
sigma2 <- (t(u_hat) %*% u_hat)/(nrow(X)-ncol(X)) # Variance of error term
SE_y_fem30 <- sqrt(t(x_fem30) %*% solve(t(X) %*% X) %*% x_fem30) * sqrt(sigma2)

# Construct test statistic
t_stat <- abs((y_fem30-2.99)/SE_y_fem30)
t_stat
```

```
##          [,1]
## [1,] 3.953751
```

The value of the test statistic is above the critical value of 1.96, so we reject the null hypothesis at the 5% significance level that the average hourly wage for women with a bachelor's degree is equal $20.

3. Using your t-test, construct a 95%-confidence interval for the expected log-earnings of a 30 year old female with a bachelor degree.

Solution:

Given the t-test with size $\alpha = 0.05$:

$$\varphi(X) = \mathbf{1}\left\{ \left| \frac{\hat{y}_{fem30} - y}{SE(\hat{y}_{fem30})} \right| \leq 1.96 \right\}$$

We can construct the 95% CI by solving for all the values of y that allow us to accept the null hypothesis $\mathcal{H}_0 : y = y_{fem30}$:

$$\begin{aligned}
\left|\frac{y - \hat{y}_{fem30}}{SE(\hat{y}_{fem30})}\right| &\leq 1.96 \\
-1.96 \leq \frac{y - \hat{y}_{fem30}}{SE((\hat{y}_{fem30}))} &\leq 1.96 \\
-1.96 \times SE(\hat{y}_{fem30}) \leq y - \hat{y}_{fem30} &\leq 1.96 \times SE(\hat{y}_{fem30}) \\
\hat{y}_{fem30} - 1.96 \times SE(\hat{y}_{fem30}) \leq y &\leq \hat{y}_{fem30} + 1.96 \times SE(\hat{y}_{fem30})
\end{aligned}$$

```
# 95% CI for 30-y females with bachelor
CI_fem30 <- c(y_fem30-1.96*SE_y_fem30 , y_fem30+1.96*SE_y_fem30)
CI_fem30
```

```
## [1] 2.923857 2.967701
```

We thus have that: $CI^{fem30}_{95\%} = [2.924, 2.968] = [\$18.62, \$19.45]$.

4. Redo exercises (1) and (3) as a function of age. Concretely, plot the regression relation (the so-called age-earnings profile) between expected *age* (on the x-axis) and log *ahe* (on the y-axis) for the age range 20-65 for females with a bachelor degree, i.e. plot

$$\mathbb{E}[\log\ ahe \mid age,\ male,\ bachelor]$$

as a function of *age*. Also, overlay confidence bands around the age-earnings profile by plotting the 95% confidence interval for the above quantity as a function of *age*.

Solution:

So far we have considered point- and interval-estimation of the log average hourly earnings for 30-year old females with a bachelor, i.e.

$$\hat{y}_i = \tilde{x}_i'\hat{\beta}$$

for a specific $\tilde{x}_i = [1, 30, 30^2, 1, 1]$. Now we want to conduct our estimation for many different ages. To do this, we consider the point estimate

$$\hat{y}_i = \tilde{x}_i'\hat{\beta}, \qquad \tilde{x}_i \equiv [1, age, age^2, 1, 1]'$$

for many different values of *age*, and the confidence interval

$$[\hat{y}_i - 1.96 \times SE(\hat{y}_i)\ ,\ \hat{y}_i + 1.96 \times SE(\hat{y}_i)]\ ,$$

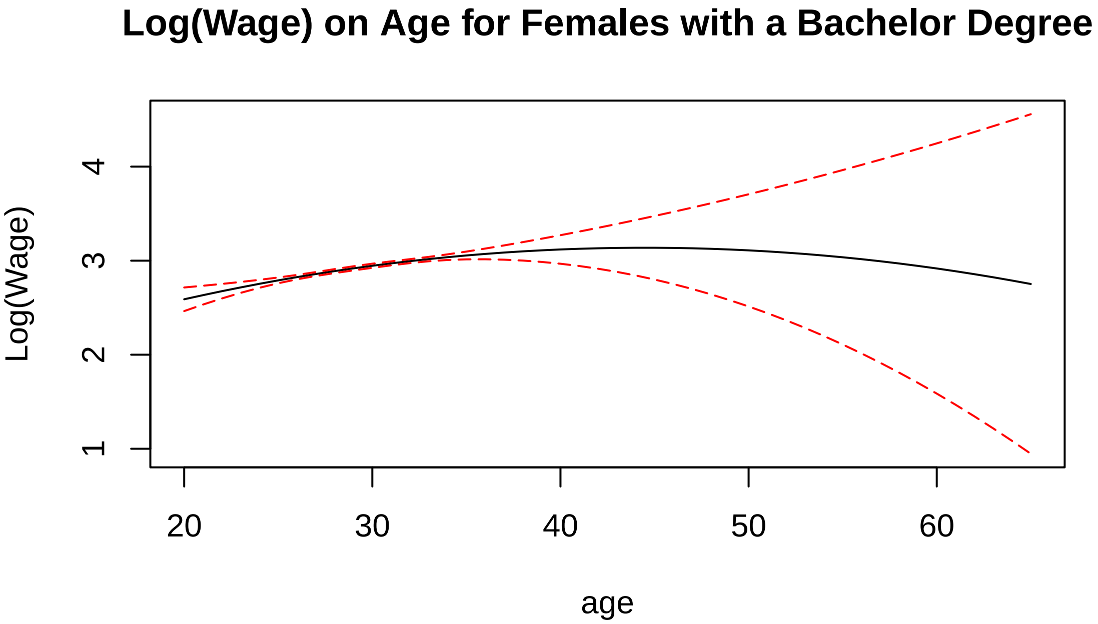
whereby

$$SE(\hat{y}_i) = \sqrt{\tilde{x}_i'\hat{\sigma}^2(X'X)^{-1}\tilde{x}_i}$$

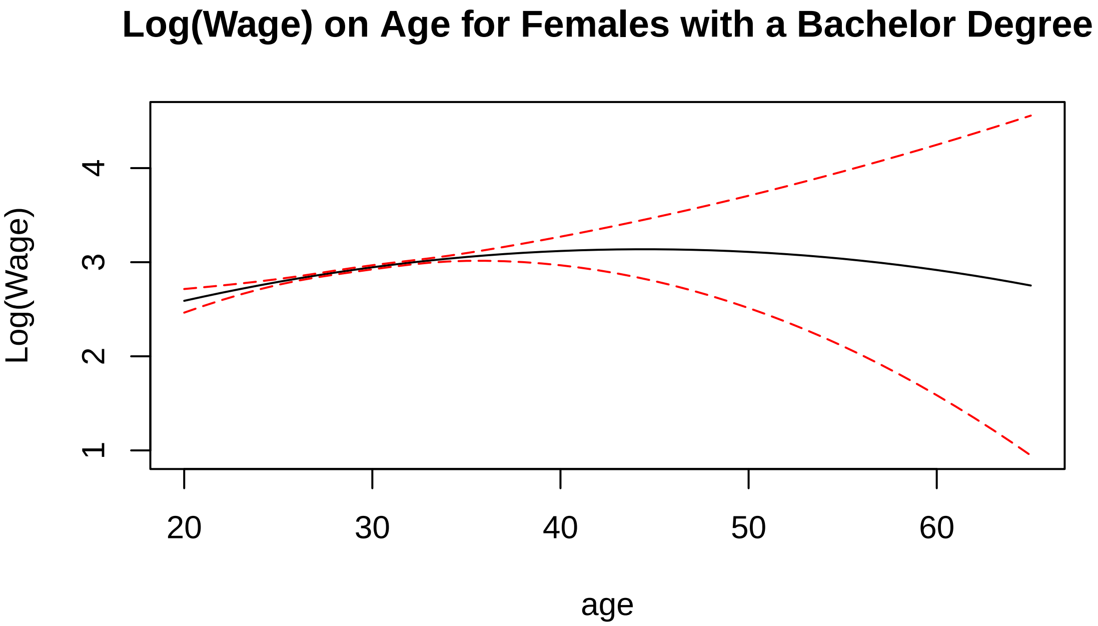
for $x_i = [1, age, age^2, 1, 1]'$, as defined above. Hence, both the point estimate and the confidence interval will be a function of age, and we are predicting the wage of females with a bachelor for all ages from 20 to 65 years.

```r
# Compute prediction manually
age <- seq(20,65,by=1) # Generate ages from 20 to 65
x_ages <- cbind(rep(1, length(age)), age, age^2,
                    rep(1, length(age)), rep(1, length(age)))
y_ages <- x_ages %*% beta_hat # Compute point estimate
# Compute SE
se_ages <- sqrt(diag(x_ages%*%solve(t(X)%*%X)%*%t(x_ages)))*sqrt(sigma2)
y_ages_LB <- y_ages-1.96*se_ages
y_ages_UB <- y_ages+1.96*se_ages

# Plot the predition with 95% CI
plot(age,y_ages,xlab="age", ylab="Log(Wage)", type="l",
ylim=c(min(y_ages_LB),max(y_ages_UB)),
main="Log(Wage) on Age for Females with a Bachelor Degree")
lines(age,y_ages_LB,lty=5,col="red")
lines(age,y_ages_UB,lty=5,col="red")
```



```r
# Could also have used the predict() function
data1 <- data.frame(age=age,female=rep(1,length(age)),
                    bachelor=rep(1,length(age)))
y_ages <- predict(myOLSa, newdata = data1, interval = 'confidence')
plot(age,y_ages[,1],xlab="age", ylab="Log(Wage)", type="l",
ylim=c(min(y_ages[,2]),max(y_ages[,3])),
main="Log(Wage) on Age for Females with a Bachelor Degree")
lines(age,y_ages[,2],lty=5,col="red")
lines(age,y_ages[,3],lty=5,col="red")
```



5. Can you interpret the coefficient in front of *bachelor* as the causal effect of obtaining a bachelor degree on earnings? Discuss.

Solution:

No because omitted variables like ability are correlated with the regressors. Therefore, A3 fails and we have omitted variable bias, which leads to inconsistent estimators. More precisely, if our estimated model is

$$y_i = x_i'\beta + u_i, \qquad x_i = [1, age_i, age_i^2, female_i, bachelor_i]' \ ,$$

but the true Data Generating Process (DGP) is

$$y_i = x_i'\beta + z_i'\delta + v_i, \qquad z_i = [ability_i] \ ,$$

then the error term in our regression is $u_i = z_i'\delta + v_i$, and it is correlated with our regressors:

$$\mathbb{E}[x_i'u_i] = \mathbb{E}[x_i'(z_i'\delta)] = \mathbb{E}[x_i'z_i']\delta \neq 0 \ ,$$

because ability is correlated with whether or not someone gets a bachelor's degree. To solve this issue, we could try to find an IV to isolate the exogenous variation in education (e.g. quarter of birth as in Angrist øKrueger, 1991).

6. By virtue of including both *age* and age^2, the regression you interpreted so far assumes a non-linear relationship between age and log-earnings, and this relationship is assumed to be the same for males and females. Keeping the assumption of such a non-linear relationship between age and log-earnings, test whether this relationship is different for males and females.

Hint: construct two covariates as the interactions $female * age$ and $female * age^2$, and test whether they are jointly (!) significantly different from zero.

Solution:

To test the joint hypothesis we use the Wald test. We rely on the asymptotic distribution of the OLS estimator:

$$\hat{\beta} \overset{approx.}{\sim} N\left(\beta, \frac{1}{N}\hat{\mathbb{V}}[\beta]\right), \qquad \mathbb{V}[\beta] = \hat{\sigma}^2\hat{Q}^{-1}$$

Under the null hypothesis that $\hat{\beta}_4 = \hat{\beta}_5 = 0$, the linear combination $g(\hat{\beta})$ of the estimator $\hat{\beta}$ will be asymptotically distributed as:

$$g(\hat{\beta}) \overset{approx.}{\sim} N\left(g(\beta), \frac{1}{N}\hat{\mathbb{V}}[g(\beta)]\right) \; ,$$

where

$$\hat{\mathbb{V}}[g(\hat{\beta})] \approx \frac{1}{N}G(\hat{\beta})\hat{\sigma}^2\hat{Q}^{-1}G(\hat{\beta})'$$

and

$$g(\hat{\beta}) = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \beta - \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \text{and } \; G(\hat{\beta}) = \frac{\partial g(\hat{\beta})}{\partial \beta'} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

The Wald test statistic will therefore be chi-squared distributed with degrees of freedom equal to 2, the number of tested restrictions:

$$T_W = Ng(\hat{\beta})'[G(\hat{\beta})\hat{\sigma}^2\hat{Q}^{-1}G(\hat{\beta})']^{-1}g(\hat{\beta}) \sim \chi^2_m$$

```r
# Run OLS manually with interaction terms
X <- cbind(rep(1), data$age, data$age^2,
           data$age*data$female, data$age^2*data$female,
           data$female, data$bachelor)
beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
sigma2 <- (t(y-X%*%beta_hat) %*% (y-X%*%beta_hat))/(nrow(X)-ncol(X))

# Wald test statistic for joint hypothesis
G <- matrix(c(0, 0, 0, 1, 0, 0, 0,
              0, 0, 0, 0, 1, 0, 0),
              nrow = 2, byrow = TRUE) # Matrix of restrictions
W <- (t(G%*%beta_hat)%*%solve(G%*%solve(t(X) %*% X)%*%t(G))%*%(G%*%beta_hat))/sigma2
W
```

```
##          [,1]
## [1,] 20.74468
```

```r
# Implement this by using the car package
library(car)
myOLSb <- lm(log(ahe) ~ age * female + I(age^2) * female + female + bachelor, data = data)
summary(myOLSb)
```

```
## 
## Call:
## lm(formula = log(ahe) ~ age * female + I(age^2) * female + female + 
##     bachelor, data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.32985 -0.27819  0.02009  0.30440  1.68064
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.315174   0.853681   0.369   0.7120
## age               0.126467   0.058118   2.176   0.0296 *
## female            1.417720   1.283361   1.105   0.2693
## I(age^2)         -0.001553   0.000982  -1.582   0.1138
## bachelor          0.426972   0.010797  39.544   <2e-16 ***
## age:female       -0.092385   0.087482  -1.056   0.2910
## female:I(age^2)  0.001278   0.001480   0.864   0.3877
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4688 on 7704 degrees of freedom
## Multiple R-squared:  0.203,  Adjusted R-squared:  0.2023
## F-statistic:   327 on 6 and 7704 DF,  p-value: < 2.2e-16
```

```
linearHypothesis(myOLSb, c("age:female = 0", "female:I(age^2) = 0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## age:female = 0
## female:I(age^2) = 0
##
## Model 1: restricted model
## Model 2: log(ahe) ~ age * female + I(age^2) * female + female + bachelor
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   7706 1697.7
## 2   7704 1693.1  2    4.5591 10.372 3.173e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Wald test statistic equals 20.745. Since the test statistic is higher than the critical value, we reject the null hypothesis that the non-linear relationship between age and log-earnings is the same for males and females at the 1% significance level. (The F-test statistic, which equals the Wald test statistic divided by the number of degrees of freedom, is equal to $20.745/2 = 10.372$. It is distributed as a Chi-squared RV with one degree of freedom, and – unsurprisingly – it leads to the same conclusion about our hypothesis as the Wald test.)