Assignment 1

Use R's `lm()` function for OLS estimation of all the regressions in Qn. 1 to Qn. 5. Assume homoskedasticity unless otherwise instructed.

Question 1

(a) Suppose $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ with $\beta_2 > 0$ and $E(\epsilon \mid X) = 0$, so that the conditional mean $E(Y \mid X)$ is a "U-shaped" quadratic function in X, initially decreasing in X, before becoming increasing in X. At what value of X is $E(Y \mid X)$ predicted to start increasing (the "turning point")? What if $\beta_1 = 0$ and $\beta_2 > 0$?

(b) Using the data in `earnings2019.csv`, estimate the following two regressions:

i. $\ln earn = \beta_0 + \beta_1 educ + \beta_2 educ^2 + \epsilon$

ii. $\ln earn = \beta_0 + \beta_1 educ^2 + \epsilon$

For each model, find out the level of `educ` at which $E(\ln Y \mid X)$ is predicted to stop decreasing and start increasing. Which empirical model, in your opinion, makes more sense, and why?

Question 2

(The Background) Suppose $y = f(x)$, and let Δy be the change in y associated with a change in x by Δx. Then in percentage terms, the ratio of the change in y to the change in x is

$$\frac{\Delta y/y}{\Delta x/x} = \frac{\Delta y}{\Delta x}\frac{x}{y}$$

The **elasticity** of y with respect to x is the limit of the above as $\Delta x \to 0$, i.e.,

$$El_y(x) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}\frac{x}{y} = f'(x)\frac{x}{y}\,.$$

Furthermore, since $\ln y = \ln f(x)$ and $x = \exp\{\ln x\}$, we have

$$\frac{d\ln y}{d\ln x} = \left(\frac{d\ln y}{dx}\right)\left(\frac{dx}{d\ln x}\right) = \left(\frac{1}{f(x)}f'(x)\right)(x) = f'(x)\frac{x}{y} = El_y(x)\,.$$

(The Question) Using the data in `rdchem`, which contains data sampled from a population of firms in a certain chemical industry, estimate a regression of $\ln rd$ (annual expenditure on research) on an intercept and $\ln sales$. Does this regression fit the data well? What is your estimated elasticity of rd with respect to $sales$?

To get the data, load the `wooldridge` package using the `library()` command. The `wooldridge` package has to be installed on your `R` setup first.

Question 3

The data set `ceosal2` from the `wooldridge` library contains data on 177 CEOs.

(a) Use R's `summary()` and `cor()` functions to generate summary statistics and the correlation matrix for the variables `lsalary` (log salary), `lsales` (log sales), `ceoten` (CEO tenure at current firm), `lmktval` (log market value of the firm), and firm `profits`.

(b) Estimate the regressions

$$\text{(A)} \quad lsalary_i = \beta_0 + \beta_1 lmktval_i + \epsilon_i \,,\; i = 1, \ldots, 177$$

and

$$\text{(B)} \quad lsalary_i = \beta_0 + \beta_1 lmktval_i + \beta_2 profits + \epsilon_i \,,\; i = 1, \ldots, 177$$

and present the estimates, standard errors, t-stats and corresponding p-values for the two models. Explain carefully why the coefficient on `lmktval` drops and its standard error increases substantially when `profit` is included in the model. Finally, explain why we cannot apply the log tranformation to `profits`.

(c) Present the R^2 for regression (A) and show that the square of the correlation between `lsalary` and the fitted `lsalary`, and the square of the correlation between `lsalary` and `lmktval` all have the same value.

(d) Now estimate the regression

$$lsalary_i = \beta_0 + \beta_1 lmktval_i + \beta_2 profits + \beta_3 ceoten_i + \beta_4 lsales_i + \epsilon_i \,,\; i = 1, \ldots, 177 \,.$$

What is the estimated percentage increase in CEO salary for another year of CEO tenure?

Question 4

The dataset `beauty` (from the `wooldridge` package) contains labor market data on 1260 individuals. These data were collected by interviewers who visited the respondents in their homes, and who (presumably without the knowledge of the respondents) rated their physical appearance on a five-point scale, from "Strikingly Handsome/Beautiful" to "Homely". These ratings were used to generate the dummy variables `abvavg` and `belavg`.[1] This is data from 1977.

(a) What percentage of the sample were rated "below average" and what percentage were rated "above average"?

(b) Estimate the following regression using OLS, assuming homoskedasticity:

$$lwage = \beta_0 + \beta_1 belavg + \beta_2 abvavg + \beta_3 female + \beta_4 educ + \beta_5 exper + \beta_6 exper^2 + \epsilon$$

and report your result using the command `coeftest(modelname)` where `modelname` is whatever you named your model. Interpret your results pertaining to the variables `belavg`, `abvavg` and `female`.

[1] Hamermesh, D.S. and J.E. Biddle (1994), "Beauty and the Labor Market", *American Economic Review*, 84, 5, 1174-1194.

(c) Estimate the following regression using OLS, assuming homoskedasticity:

$$\begin{aligned} lwage = \beta_0 + \beta_1 belavg + \beta_2 abvavg + \beta_3 female + \beta_4 educ + \beta_5 exper + \beta_6 exper^2 + \\ \beta_7 female \cdot belavg + \beta_8 female \cdot abvavg + \beta_9 female \cdot educ + \\ \beta_{10} female \cdot exper + \beta_{11} female \cdot exper^2 + \epsilon \end{aligned}$$

and report your result using the command `coeftest(modelname)` where `modelname` is whatever you named your model. What are your main takeaways from this regression? (Hint: you can use use the formula `lwage ~ female * (belavg + abvavg + educ + exper + I(exper^2))` to obtain all the interaction terms.)

(d) Compute heteroskedasticity-robust standard errors for the regression in (c). Use the "HC1" version. Are there any major changes in the results?

Question 5

The dataset `kielmc` (from the `wooldridge` library) contains data sampled from a population of homes in 1978 and in 1981 (different homs are included in the 1978 and 1981 samples, i.e., we have a "pooled cross sectional dataset"). Sometime between 1978 and 1981, a rumor began circulating that a garbage incinerator would be built within the vicinity of the homes. Construction of the incinerator actually began after 1981, at the rumored location.

(a) Estimate the following simple linear regression

$$rprice = \beta_0 + \beta_1 nearinc + \epsilon$$

twice, once using only 1978 data and another time using only 1981 data. Report your results, including number of observations and R^2. The dummy variable `nearinc` indicates if a home is within 3 miles of the rumored location of the incinerator. The dummy variable `y81` indicates whether an observation is 1981 data or not. Interpret the estimates of both the intercept and the coefficient on `nearinc`. Do they indicate that the rumor of an impending incinerator caused prices of houses near the rumored location of the incinerator to fall? If yes, test whether the effect of the rumor is statistically significant. *Hint: to carry out this test, run the following regression on the full data set:*

$$rprice = \beta_0 + \delta_0 y81 + \beta_1 nearinc + \delta_1 y81 \cdot nearinc + \epsilon\,.$$

(b) Repeat the exercise in part (a), this time including the controls `age`, `agesq`, `intst` (distance to interstate highway), `land` (land area of home), `area` (house area), `rooms` (number of rooms), `bath` (number of bathrooms). What changes?

Question 6

In any (simple or multiple) linear regression, the square of the correlation of the regressand Y_i and the fitted values $\hat{Y}_i$ is

$$Cor(Y_i, \hat{Y}_i^{ols})^2 = \frac{\left(\sum_{i=1}^n (Y_i - \overline{Y})(\hat{Y}_i^{ols} - \overline{\hat{Y}^{ols}})\right)^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2 \sum_{i=1}^n (\hat{Y}_i^{ols} - \overline{\hat{Y}^{ols}})^2}\,.$$

Show that in any regression with an intercept: (a) this is equal to the R^2 of the regression, and (b) this is equal to the square of the correlation between Y_i and the regressor X_i in the case of a simple linear regression.

The next four questions explore various "mechanical" aspects of least squares estimation of regression lines.

Question 7

The file `anscombe.xlsx` contains 11 observations on four pairs of variables (X_1, Y_1), (X_2, Y_2), (X_3, Y_3) and (X_4, Y_4). For each pair (X_j, Y_j), $j = 1, 2, 3, 4$, estimate the regression

$$Y_j = \beta_0 + \beta_1 X_j + \epsilon$$

and produce a scatterplot of the Y observations against the X observations, together with the fitted regression line. Compare your plots and the estimation results for the four regressions. What lessons do you learn from this exercise?

Question 8

(a) Estimate the regression of `y` on `x` for each of the three data sets below. Use the `lm()` and `summary()` functions as shown below, replacing `dat` with `dat1`, `dat2` and `dat3` for the three regressions. What R^2 value is shown in each case? Can you trust these values? Why or why not?

```
dat1 <- data.frame(y=c(1,1,1,1,1), x=c(4,5,3,1,8))    # y = 1 for all observations
dat2 <- data.frame(y=c(1,2,3,4,5), x=c(1,2,3,4,5))    # y = x for all observations
dat3 <- data.frame(y=c(4,1,0,1,4), x=c(-2,-1,0,1,2))  # y = x^2 for all observations
mdl <- lm(y~x, data=dat1)
sum.mdl <- summary(mdl)
sum.mdl$coefficients
cat("R-squared:", sum.mdl$r.squared)
```

(b) Estimate the regression of `y` on `x` for each of the three data sets below, and examine the estimates of the slope coefficient? Can you explain the differences (or lack thereof) in the three estimates?

```
dat1 <- data.frame(y=c(0,4,2,6,5), x=c(1,2,3,4,5))    # y = 1 for all observations
dat2 <- data.frame(y=c(0,4,2+100,6,5), x=c(1,2,3,4,5))    # y = x for all observations
dat3 <- data.frame(y=c(0,4,2,6,5+100), x=c(1,2,3,4,5))    # y = x for all observations
```