# HOMEWORK 6
# GRA6036 ECONOMETRICS WITH PROGRAMMING
# AUTUMN 2020

EMIL A. STOLTENBERG

**Exercise 1.** It's your turn to derive the least squares estimators. We have data

$$(Y_1, x_1), \ldots, (Y_n, x_n)$$

that come from the model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed (i.i.d.) random variables with expectation $\mathrm{E}[\varepsilon_1] = 0$ and variance $\mathrm{Var}(\varepsilon_1) = \sigma^2$; $x_1, \ldots, x_n$ are fixed numbers (not random variables), and we assume that they are not all equal, so that $\sum_{i=1}^n (x_i - \bar{x}_n)^2 > 0$. We'll take $\beta_0, \beta_1$ and $\sigma^2$ to be unknown parameters. The least squares estimators for $\beta_0$ and $\beta_1$ are the minimisers of the function

$$g(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2.$$

We'll denote the minimisers of $g(\beta_0, \beta_1)$ by $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

(a) Show that

$$\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_i, \quad \text{and} \quad \mathrm{Var}(Y_i) = \sigma^2,$$

for each $i$. What is $\mathrm{E}[(Y_i - \beta_0 - \beta_1 x_i)^2]$ equal to?

(b) Why must it be true that $g(\widehat{\beta}_0, \widehat{\beta}_1) \leq g(1.23, 4.56)$?

(c) Explain why it is a good idea to use the minimisers $\widehat{\beta}_0$ and $\widehat{\beta}_1$ as our estimators for the unknown parameters $\beta_0$ and $\beta_1$.

(d) Use what you know about sums to show that

$$\frac{\partial}{\partial \beta_0} g(\beta_0, \beta_1) = -2n\big(\bar{Y}_n - \beta_0 - \beta_1 \bar{x}_n\big),$$
$$\frac{\partial}{\partial \beta_1} g(\beta_0, \beta_1) = -2\big(\sum_{i=1}^n Y_i x_i - n\beta_0 \bar{x}_n - \beta_1 \sum_{i=1}^n x_i^2\big).$$

where $\bar{Y}_n = (1/n)\sum_{i=1}^n Y_i$ and $\bar{x}_n = (1/n)\sum_{i=1}^n x_i$.

(e) Set

$$\frac{\partial}{\partial \beta_0} g(\beta_0, \beta_1) = 0, \quad \text{and} \quad \frac{\partial}{\partial \beta_1} g(\beta_0, \beta_1) = 0,$$

*Date*: September 28, 2020.

and solve for $\beta_0$ and $\beta_1$. Show that the solution is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(Y_i - \bar{Y}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}, \quad \text{and} \quad \widehat{\beta}_0 = \bar{Y}_n - \widehat{\beta}_1 \bar{x}_n. \tag{1}$$

These are the least squares estimators.

(f) Show that we may write

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}_n) Y_i}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}, \quad \text{and} \quad \widehat{\beta}_0 = \sum_{i=1}^n \big(\frac{1}{n} - \frac{(x_i - \bar{x}_n)\bar{x}_n}{\sum_{j=1}^n (x_j - \bar{x}_n)^2}\big) Y_i.$$

(g) The expressions from (f) makes it easier to show that

$$\mathrm{E}\,[\widehat{\beta}_0] = \beta_0, \quad \text{and} \quad \mathrm{E}\,[\widehat{\beta}_1] = \beta_1,$$

and that

$$\mathrm{Var}(\widehat{\beta}_0) = \big(\frac{1}{n} + \frac{\bar{x}_n^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\big)\sigma^2, \quad \text{and} \quad \mathrm{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}.$$

Check that these expressions are correct.

(h) As you can see from the expressions in (g), in order to say something about the uncertainty (the variance) of our estimators we need to estimate the unknown $\sigma^2$. Consider the estimator given by

$$\widehat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2. \tag{2}$$

Explain why this estimator seems reasonable at doing what we want it to do, namely estimate $\sigma^2$.

**Exercise 2.** In this exercise we take a closer look at the estimator $\widehat{\sigma}_n^2$ given in (2), that is we continue work on the model introduced and studied in Ex. 1. The goal of the exercise is to show that $\widehat{\sigma}_n^2$ is biased for $\sigma^2$, and consequently construct an unbiased estimator. Recall that, in general, an estimator $\widehat{\theta}_n$ for $\theta$ is *unbiased* for $\theta$ if $\mathrm{E}\,[\widehat{\theta}_n] = \theta$, and is *biased* for $\theta$ if $\mathrm{E}\,[\widehat{\theta}_n] \neq \theta$.

(a) When evaluated in the minimiser $\widehat{\beta}_0$ and $\widehat{\beta}_1$ of $g(\beta_0, \beta_1)$, the partial derivatives of $g(\beta_0, \beta_1)$ equals zero, i.e.

$$\frac{\partial}{\partial \beta_0} g(\widehat{\beta}_0, \widehat{\beta}_1) = 0, \quad \text{and} \quad \frac{\partial}{\partial \beta_1} g(\widehat{\beta}_0, \widehat{\beta}_1) = 0.$$

This fact is helpful when deriving that

$$\widehat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 = \frac{1}{n} \sum_{i=1}^n Y_i^2 - \bar{Y}_n^2 - \frac{1}{n} (\widehat{\beta}_1)^2 \sum_{i=1}^n (x_i - \bar{x}_n)^2.$$

Derive the right hand side expression yourself. If you don't manage this, just skip to (b).

(b) Show that

$$\mathrm{E}\,[Y_i^2] = \sigma^2 + (\beta_0 + \beta_1 x_i)^2, \quad \text{for } i = 1, \ldots, n,$$

and that

$$\mathrm{E}\,[\bar{Y}_n^2] = \frac{\sigma^2}{n} + (\beta_0 + \beta_1 \bar{x}_n)^2, \quad \text{and} \quad \mathrm{E}\,[(\widehat{\beta}_1)^2] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} + \beta_1^2.$$

*Hint*: Recall that $\mathrm{Var}(Z) = \mathrm{E}\,[Z^2] - (\mathrm{E}\,[Z])^2$.

(c) Show that we can write

$$\mathrm{E}\,\big[\sum_{i=1}^n Y_i^2\big] = n\sigma^2 + n(\beta_0 + \beta_1 \bar{x}_n)^2 + \beta_1^2 \sum_{i=1}^n (x_i - \bar{x}_n)^2.$$

*Hint*: Use that $\beta_0 + \beta_1 x_i = \beta_0 + \beta_1 \bar{x}_n + \beta_1 (x_i - \bar{x}_n)$.

(d) Combine the results from (a), (b), and (c) to show that $\widehat{\sigma}_n^2$ is biased for $\sigma^2$, in fact

$$\mathrm{E}\,[\widehat{\sigma}_n^2] = \frac{n-2}{n}\sigma^2.$$

(e) Find an unbiased estimator for $\sigma^2$.

**Exercise 3.** A good way to get to know a model, as well as the behaviour of the estimators of the unknown parameters of the model, is to simulate data from it. Consider the model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, n, \tag{3}$$

where the $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. normal random variables with expectation 0 and variance $\sigma^2$. To simulate from this model we need to insert some actual numbers for the unknown parameters. Set $\beta_0 = -0.543$, $\beta_1 = 1.234$, and $\sigma^2 = 2.345$. For the covariates $x_1, \ldots, x_n$ we'll use

$$x_i = i/n, \quad \text{for } i = 1, \ldots, n,$$

and we set $n = 100$. In the Matlab script below I simulate a dataset

$$Y_1, \ldots, Y_n,$$

from this model.

```
n = 100;
beta0 = -0.543; beta1 = 2.345;
sigma2 = 1.234;
x = linspace(1/n,1,n);
eps = normrnd(0,sqrt(sigma2),1,n);
y = beta0 + beta1.*x + eps;
```

(a) Simulate a dataset using the code above and estimate the parameters $\beta_0$ and $\beta_1$. Make a scatter plot of you simulated data, and add the true regression line $(x, \beta_0 + \beta_1 x)$ and the estimated regression line $(x, \widehat{\beta}_0 + \widehat{\beta}_1 x)$ to the scatter plot. One such plot is in Figure 1. Run the script a few times.

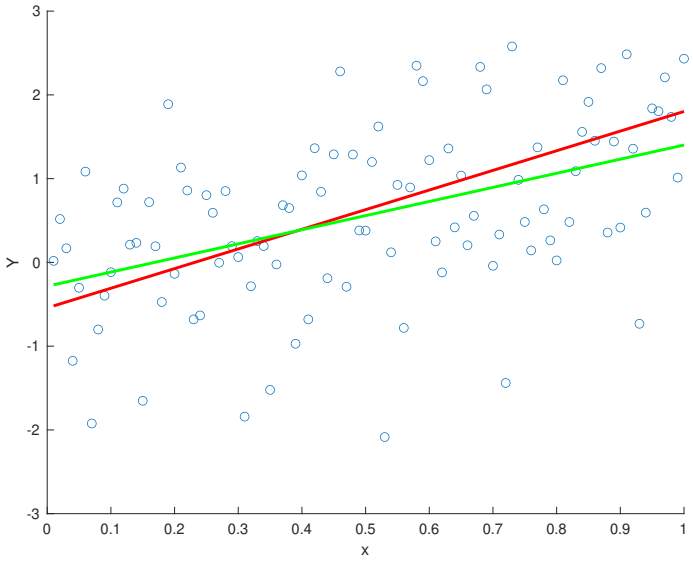


FIGURE 1. The plot described in Ex. 3(a). The red line is the true regression line, the green line is the estimated regression line.

(b) We now want to simulate 1000 datasets from the model in (3), for each dataset we estimate $\beta_1$ using the least squares estimator $\widehat{\beta}_1$, and save our estimate. To do that we wrap parts of the script above into a for-loop. Here is code with some details that you must fill in (note also that the script builds on the script above).

```
sims = 1000;
hats = zeros(1,n);
for u = 1:sims
    eps = normrnd(0,sqrt(sigma2),1,n);
    y = beta0 + beta1.*x + eps;
    beta1hat = % fill in
    hats(u) = beta1hat;
end
```

(c) Make a histogram of the estimates of $\beta_1$, and add the pdf of the normal distribution with expectation $\mathrm{E}\,[\widehat{\beta}_1]$ and $\mathrm{Var}(\widehat{\beta}_1)$. Your plot should resemble the plot in Figure 2

(d) We call $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ the fitted values. Show that we can write

$$\widehat{Y}_i = \sum_{j=1}^{n} \{\frac{1}{n} + \frac{(x_j - \bar{x}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\} Y_j, \quad \text{for } i = 1, \ldots, n,$$

and use this expression to show that the variance of the fitted values are

$$\mathrm{Var}(\widehat{Y}_i) = \{\frac{1}{n} + \frac{(x_i - \bar{x}_n)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\} \sigma^2 \quad \text{for } i = 1, \ldots, n.$$

(e) The residuals $u_1, \ldots, u_n$ are defined by

$$u_i = Y_i - \widehat{Y}_i, \quad \text{for } i = 1, \ldots, n. \tag{4}$$

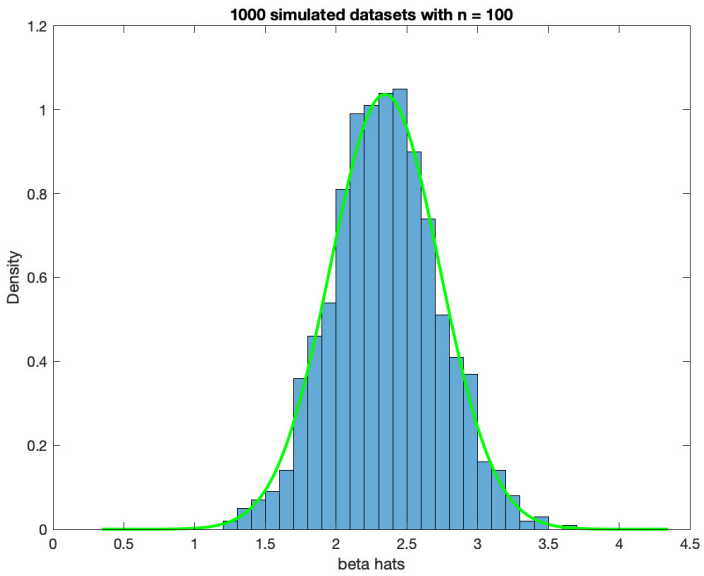


FIGURE 2. The histogram described in Ex. 3(c).

They give the deviance from the observation $Y_i$ to the fitted line $\widehat{\beta}_0 - \beta_1 x_i$ at each point $x_i$ in the data. Show that $\mathrm{E}\,[u_i] = 0$. For each $i$, we can write

$$u_i = \varepsilon_i - \sum_{j=1}^{n} \Big\{ \frac{1}{n} + \frac{(x_j - \bar{x}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} \Big\} \varepsilon_j.$$

for $i = 1, \ldots, n$. Try to show this, and use it to find that the variance of the $i$th residual is

$$\mathrm{Var}(u_i) = \sigma^2 \big(1 - \frac{1}{n} - \frac{(x_i - \bar{x}_n)^2}{\sum_{j=1}^{n}(x_j - \bar{x}_n)^2}\big).$$

So when $n$ and $\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ are big, then $\mathrm{Var}(u_i) \approx \sigma^2$. Our $n$ equals 100. and with our covariates $x_i = i/n$ for $i = 1, \ldots, n$ we have

$$\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = \frac{(n+1)(n-1)}{12n}, \quad \text{and} \quad \max_{i \le n}(x_i - \bar{x}_n)^2 = \frac{(n-1)^2}{4n^2}.$$

You do not need to show this, but if you want to, it helps to know that $\sum_{i=1}^{n} i = n(n+1)/2$, and that $\sum_{i=1}^{n} i^2 = n(n+1)(2n+1)/6$.

(f) In Figure 3 I have made a histogram of the residuals from fitting the model in (3) on simulated data, and added the pdf of a normal distribution. Reproduce the figure.

**Exercise 4.** In this exercise we continue our analysis of basketball players in the National Basketball Association (NBA) during the 2019–2020 season. The NBA is the men's professional basketball league in the US. You can find the dataset for this exercise on Itslearning,

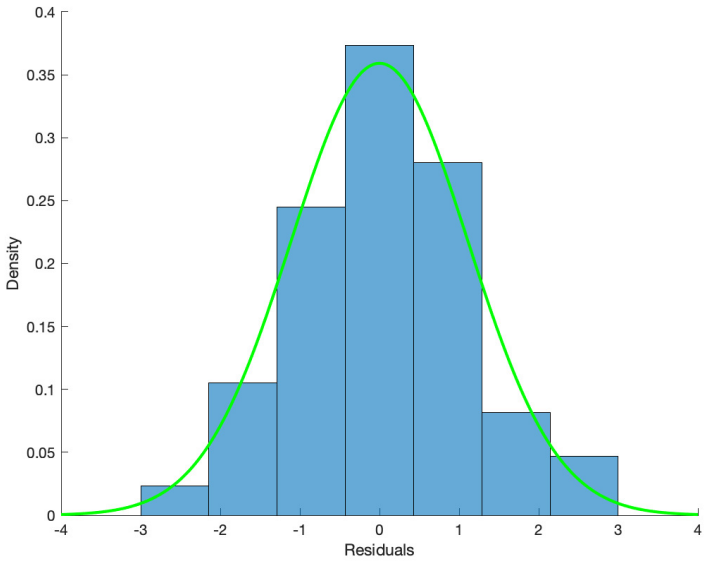


FIGURE 3. The histogram described in Ex. 3(f).

it is called `nba_20192020_lec6.txt`. I retrieved these data from https://www.basketball-reference.com/.

In this exercise we look at the relation between the efficiency of a player and his salary. The efficiency, which we denote `EFF`, is defined by

$$\texttt{EFF} = \texttt{PTS} + \texttt{TRB} + \texttt{AST} + \texttt{STL} + \texttt{BLK} - \texttt{missFG} - \texttt{missFT} - \texttt{TOV}, \tag{5}$$

where

- `PTS`= Points scored per game;
- `TRB`= Rebounds per game;
- `AST`= Assists per game;
- `STL`= Steals per game;
- `BLK`= Blocks per game;
- `missFG`= Missed field goals (i.e. shot attempts) per game;
- `missFT`= Missed free throws per game;
- `TOV`= Turnovers (losing the ball to the opposing team) per game;

all based on the 2019–2020 (regular) season.

To compute these quantities we need the following columns from the `nba_20192020_lec6.txt` dataset, `PTS`, `TRB`, `AST`, `STL`, `BLK`, `FG`, `FGA`, `FT`, `FTA`, `TOV`. All of these are *per game* statistics. Most of these are self explanatory in view of the list above, except `FG`= Made field goals per game; `FGA`= Field goal attempts per game; `FT`= Free throws per game; and `FTA`= Free throw attempts per game. Thus,

$$\texttt{missFG} = \texttt{FGA} - \texttt{FG}, \quad \text{and} \quad \texttt{missFT} = \texttt{FTA} - \texttt{FT}.$$

The first few rows of the dataset look like this

(a) Compute the efficiency of each player using the formula in (5).

(b) Make a histogram of the efficiency of players who played 20 games or more, and played 10 minutes or more per game (a NBA game lasts for 48 minutes). Among these players, find the name of the player with the lowest efficiency, and the name of the player with the highest efficiency.

(c) Make a scatter plot of efficiency and salary of the players who played 20 games or more, and played 10 minutes or more per game. Add the line $(x, -3519495.44 + 963183.84\,x)$ to your scatter plot.

(d) We would like to fit a model of the type

$$\texttt{Salary}_i = \beta_0 + \beta_1 \texttt{Eff}_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. normally distributed random variables with expectation 0 and variance $\sigma^2$ to the data. Looking at the scatter plot you made in (c), can you see a problem with fitting this model to the data, that is, an assumption of this model that seems to be broken in the data? *Hint*: Look at the spread of the blue dots around the green line as the efficiency increases.

(e) Still, just considering the players who played 20 games or more, and played 10 minutes or more per game (I find 308 such), define

$$Y_i = \log(\texttt{Salary}_i), \quad \text{and} \quad x_i = \log(\texttt{Eff}_i), \quad \text{for } i = 1, \ldots, 308.$$

Make a scatter plot with of the pairs $(x_i, Y_i)$ for $i = 1, \ldots, 308$.

(f) Consider the model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{for } i = 1, \ldots, 308, \tag{6}$$

where $\varepsilon_1, \ldots, \varepsilon_{308}$ are independent normally distributed random variables with expectation 0 and variance $\sigma^2$. Use the least squares estimators to estimate $\beta_0$ and $\beta_1$. What are your estimates? Add the fitted line

$$(x, \widehat{\beta}_0 + \widehat{\beta}_1 x),$$

to the scatter plot you made in (f).

(g) Use the estimator $\widehat{\sigma}^2$ given in (2), and estimate $\sigma^2$.

(h) Under the assumption that $\varepsilon_1, \ldots, \varepsilon_n$ are independent normals,

$$\frac{\widehat{\beta}_1 - \beta_1}{\{\mathrm{Var}(\widehat{\beta}_1)\}^{1/2}} \sim \mathrm{N}(0, 1).$$

Construct an interval that contains $\beta_1$ with probability 0.95. Use you estimates to compute a realisation of this interval. See hw4. Ex. 4(j).

(i) Compute the residuals as defined in (4), and make a histogram of these.

Department of Economics, BI Norwegian Business School
*Email address*: emil.a.stoltenberg@bi.no