Assignment 3

Directions

Submit two files on Moodle before the deadline:

[1] your **solution set** as a *.pdf [2] the **R script** as a *.R type file

For both, the first line should include your student ID and the assignment number

If you worked with one or two other students write "group:" and then their student ID numbers

[1] the **solution set** should include for each answer:

- the question number/part
- your answer in words
- the command in R that was used (if relevant)
- the output from Rstudio (if relevant)

[2] the **R script** should include:

- all the commands you used to generate the solutions
- individually written comments that explain your code to the grader

Additional notes on reporting results in this assignment

- **Confidence intervals**: report confidence intervals as a range [L, U] where L is a number representing the lower bound and U is a number representing the upper bound.
- **Hypothesis testing**: always include all four steps: (1) write the null and alternative hypothesis you are testing; (2) calculate the test statistic; (3) provide the critical value you are using; (4) state the conclusion of your hypothesis test based on the appropriate rejection regions.
- **Significance tests:** You can use the t-statistics and p-values provided by the output in R to make statements about significance (without writing out all four steps as above)

Part A

1. Change of Variables: Income-Consumption model

$$\text{Consumption}_i = \beta_0 + \beta_1 \text{GDP}_i + u$$

is estimated first for observations collected in Israel and then for a dataset in the US. The estimates for the two datasets are:

Estimates	Israel	United States
intercept β_0	-21.894	515.784
slope β_1	0.815	0.645

The variables for consumption and GDP in Israel are measured in units of millions of NIS while in the US they are measured in units of billions of dollars. Assume that the exchange rate is 3.4NIS to $1 and translate the slope and intercept parameters of from Israel into identical units as in the US.

(continued)

2. Change of Variables: Wage model

$$\ln(\text{wage})_i = \beta_0 + \beta_1 \cdot \text{education}_i + u_i$$

is estimated for women who work aged 25-65, where the wage is measured in units of dollars and education is measured in years of education. Translate the results of the model below from dollars to NIS: what is the estimate for the intercept and what is the estimate for the slope? (*hint*: use the fact that $\ln(cY) = \ln(c) + \ln(y)$ and assume the exchange rate is 3.4NIS to 1 USD)

```
Call:
lm(formula = log_wage_per_hour ~ educ, data = myregdata)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0906 -0.3654 -0.0343  0.3430  2.8511

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.496805   0.020433  122.19   <2e-16 ***
educ        0.078791   0.001436   54.88   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5524 on 13432 degrees of freedom
  (8767 observations deleted due to missingness)
Multiple R-squared:  0.1831,    Adjusted R-squared:  0.1831
F-statistic:  3012 on 1 and 13432 DF,  p-value: < 2.2e-16
```

3. Regression through the origin: the Income-consumption model revisited

A colleague estimates the Income-consumption model

$$\text{Consumption}_i = \beta_0 + \beta_1 \text{GDP}_i + u$$

and sends you the following output from R:

```
Call:
lm(formula = consumption ~ gdp, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-2364.2 -1132.4   174.6   959.2  2197.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.189e+04  1.377e+03   -15.9   <2e-16 ***
gdp          8.155e-01  1.914e-02    42.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1325 on 32 degrees of freedom
Multiple R-squared:  0.9827,    Adjusted R-squared:  0.9821
F-statistic:  1815 on 1 and 32 DF,  p-value: < 2.2e-16
```

Your colleague claims it is not reasonable for the regression's intercept to be negative, since it is impossible for there to be negative consumption. As such she decides to estimate a model without an intercept (a regression through the origin):

$$\text{Consumption}_i = \beta_1 \text{GDP}_i + u$$

She also sends you the results of this second estimation in R:

```
Call:
lm(formula = consumption ~ gdp + 0, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
 -7666  -3784   -742   2509   4548

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
gdp  0.51524    0.00928   55.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3892 on 33 degrees of freedom
Multiple R-squared:  0.9894,    Adjusted R-squared:  0.9891
F-statistic:  3083 on 1 and 33 DF,  p-value: < 2.2e-16
```

a) Is the intercept of the first model significant?

b) Given the estimated results, what is the correct model to use in this context?

Part B Applying everything you learned so far

We are interested in modeling the housing market. We start with model (1) below:

$$\text{price}_i = \beta_0 + \beta_1 \text{bdrms}_i + \beta_2 \text{lotsize}_i + u_i$$

Load the data in in the file HPRICE1_v2026.xlsx in R. The relevant variables are:

- price: house sale price in USD
- bdrms: number of bedrooms in the house
- lotsize: size of the house's lot (land the house is built on) in square feet
- lprice, llotsize: the log of the price variable and the log of the lot size respectively

4. <u>The Classical Assumptions</u>

a) Write the four classical assumptions A1-A4 in terms of this model

b) What are some other variables that might affect housing price, aside from size and number of bedrooms?

c) For each assumption A1-A4, explain in a sentence or two whether you think the assumption is reasonable in the context of housing prices (*hint*: recall that any variables that are not included in the model but that affect price will be part of the random error u)

5. <u>Model Estimation I & II (regression through the origin)</u>

a) Estimate model (1) using the `lm` command in R and write the results as a formula

b) Your colleague points out that when the lot size and the number of bedrooms are zero, the price should also be zero. Test this hypothesis using the estimation results in **5a**.

c) Estimate model (2) below (*hint*: in `lm` use the formula: "`price ~ 0 + bdrms + lotsize`"):

$$\text{price}_i = \beta_1 \text{bdrms}_i + \beta_2 \text{lotsize}_i + u_i$$

6. Linearity of the Model

a) Create a scatterplot between price and lot size (put price on the y-axis)

b) Create a scatterplot between log price and log lot size (put log price on the y-axis)

c) Which of the two relationships (**6a** and **6b**) seems more linear?

For the remainder of this problem (parts 4 through 7) we consider the model (3):

$$\mathbf{log(price)}_i = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\mathbf{bdrms}_i + \boldsymbol{\beta_2}\,\mathbf{log(lotsize)}_i + \boldsymbol{u_i}$$

7. Model Estimation III

a) Estimate model (3) above and report the results as a formula.

b) By how much does the log price of a house change if the log lot size increases by 4 (all else equal)?

c) By how much does the log price of an apartment change if the log lot size increases by 4 and the number of bedrooms increases by 2? compare this to the result in **7b.**

d) A house has 4 bedrooms and a lot size of 6,000. What is the predicted price of this house?

8. Confidence Intervals

a) Calculate the 95% confidence interval for β_2 and explain in a sentence or two the economic interpretation for this range.

b) Calculate the 90% and 99% confidence intervals for β_2 and in a couple sentences explain which one is wider and why. Make sure to relate your answer to the definition of a confidence interval.

9. Hypothesis Testing

a) Test the hypothesis that increasing the log lot-size by 1 (holding the number of bedrooms fixed) will not change the house price (at 5% significance).

b) What is the lowest significance level for which you would reject the null hypothesis in **9a**?

c) Test the hypothesis (at 5% significance) that increasing the number of bedrooms by one increases the log price by 1.

10. Change of Variables

a) Suppose you want to translate the lot size from square feet to meters squared. What will be the value of β_2 in the new model where the apartment size and lot size are measured in meters squared? (*hint*: use the fact that 1 square foot = 0.093 square meters)

b) Confirm your answer in R: create a new variable: `llotsize.m` which measures the log lot size in meters squared. Estimate model (3) using these new variables for apartment and lot size. Compare the results to the estimation in **7a**.