Assignment 4

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

- **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. In each of the following situations will the OLS estimator be biased? Explain why or why not.

a. The variance of the random errors is not constant

b. A variable that affects the outcome variable is missing from the model

c. Two of the explanatory variables have a high correlation of 0.95

2. Consider the following estimated regression on $N = 706$ observations.

$$\widehat{sleep} = 3{,}638.25 - 0.148\,totwrk - 11.13\,educ + 2.20\,age$$
$$(112.28) \quad (0.017) \quad (5.88) \quad (1.45)$$

The numbers in parentheses are the standard errors, and the goodness-of-fit is $R^2 = 0.113$.

a. Are `educ` and `age` each statistically significantly different from zero at $\alpha = 5\%$? (make sure to show how you calculate the t-statistic)

b. Test the hypothesis that the effect size of `totwrk` is positive (at $\alpha = 5\%$)

c. Running the regression omitting `educ` and `age` we estimate the following, with $R^2 = 0.103$

$$\widehat{sleep} = 3{,}586.38 - 0.151\,totwrk$$
$$(38.91) \quad (0.017)$$

Are `educ` and `age` jointly statistically significant in the original estimation at $\alpha = 5\%$? (*hint*: when the unrestricted and restricted model both have the same outcome variables, there are two ways to calculate the F-statistic: $F = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(N-M)} = \frac{(R^2_{UR} - R^2_R)/q}{(1-R^2_{UR})/(N-M)}$)

d. Does inclusion of `educ` and `age` have an impact on the marginal effect of `totwrk` on `sleep`?

Questions 3 – 6 are based on the dataset BWGHT2_v2026 on child birthweight from HW2. Refer to **BWGHT2DES.txt** for variable names and descriptions. ***There is no need to use R for these questions.***

3. All-controls Regression

The following results show the estimated effect of the average number of cigarettes smoked on child birthweight, controlling for the following variables: `npvis, mage, drink`.
The variables `magesq` is the mother's age squared and the variable `npvissq` is the number of physician visits squared.

```
> lm.BW.control <- lm( bwght ~ cigs + npvis + npvissq + mage + magesq + drink, data.BW)
> summary(lm.BW.control)

Call:
lm(formula = bwght ~ cigs + npvis + npvissq + mage + magesq +
    drink, data = data.BW)

Residuals:
     Min       1Q   Median       3Q      Max
-2960.02  -333.55     6.88   343.38  1747.17

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2188.3556   459.9037   4.758 2.17e-06 ***
cigs         -10.9866     3.9199  -2.803  0.00514 **
npvis         30.9516    11.4014   2.715  0.00672 **
npvissq       -0.5838     0.3595  -1.624  0.10465
mage          62.7780    31.0093   2.024  0.04312 *
magesq        -0.9919     0.5199  -1.908  0.05660 .
drink         -5.1426    56.5638  -0.091  0.92757
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 567.7 on 1325 degrees of freedom
  (133 observations deleted due to missingness)
Multiple R-squared:  0.02363,   Adjusted R-squared:  0.01921
F-statistic: 5.344 on 6 and 1325 DF,  p-value: 1.836e-05
```

a. Test the significance of the control variables in the regression- report which ones are statistically significant at $\alpha = 5\%$ and which ones are not.

b. For any variables that are statistically significant, interpret the coefficient estimates.

c. What is the predicted birthweight for a woman aged 30, who smokes 20 cigarettes a day, doesn't drink, and visited the physician 5 times during her pregnancy? What is the birthweight of a baby born to a woman with identical characteristics except she doesn't smoke?

d. What is the predicted difference in birthweight of a baby whose mother never visited the doctor and who smoked 20 cigarettes a day, and one whose mother visited the doctor 8 times and who never smoked? Assume both mothers are the same age and drink the same amount.

4. Quadratic terms

a. Consider the variable measuring the squared number of visits to the physician (`npvissq`). Why might we want to include this variable in our model? Given the level of significance of this variable, what can we say about the relationship between child birthweight and number of physician visits?

b. What is the optimal age to give birth if the goal is to reach the highest birthweight?

5. Model Specification

The researcher wishes to test the significance of all the controls jointly

```
> lm.BW <- lm( bwght ~ cigs, data.BW)
> anova(lm.BW.control, lm.BW)
Error in anova.lmlist(object, ...) :
  models were not all fitted to the same size of dataset
> nrow(lm.BW$model)
[1] 1383
> nrow(lm.BW.control$model)
[1] 1332
```

a. Explain the error above. Why does it occur?

b. Consider the output below. What null hypothesis being tested? What is the alternative hypothesis? What is the test statistic? Can we reject the null hypothesis at the 5% significance level?

```
> lm.BW.R <- lm( bwght ~ cigs, lm.BW.control$model)
> anova(lm.BW.control, lm.BW.R)
Analysis of Variance Table

Model 1: bwght ~ cigs + npvis + npvissq + mage + magesq + drink
Model 2: bwght ~ cigs
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1   1325 426969123
2   1330 433648214 -5  -6679091 4.1454 0.0009681 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. The coefficient on number of cigarettes smoked each day in `lm.BW.R` is -12.906. What can we learn from the comparison of this with the coefficient from `lm.BW.control`? (hint: in your answer refer to A2 in the Gauss-Markov assumptions).

Part B Multivariate regression – Discrimination Dataset

Use the file **DISCRIM_v2026** to answer the following questions. This dataset was taken from research on discrimination in the United States. The claim was that fast-food restaurants charge higher prices in neighborhoods where a higher percentage of the population was black. Each observation i represents a different zipcode in New Jersey. The variables measure prices for different items at the restaurants and demographic characteristics of the populations living in the area.

6. Data exploration

a. Calculate the sample averages and standard deviations for the variables: `prpblck`, `psoda`, and `density`.

b. Create a scatterplot comparing price of soda (y-axis) against the proportion of black people.

c. Create a scatterplot comparing price of soda (y-axis) against income.

d. What can we learn from the scatterplots about the correct model and the relation between these variables?

(continued)

7. Regression Model

Consider the log-log model that explains the price of soda as a function of the proportion of black people living in the zipcode, and population density. Model (1):

$$\log(psoda_i) = \beta_0 + \beta_1 \log(prpblck_i) + \beta_2 \log(density_i) + u_i$$

- **a.** Report the results of the OLS estimation of this regression in the form of a formula. (hint: use the variables `lpsoda`, `lprpblck`, `ldensity`)
- **b.** Test if the estimated parameters are **each** (separately) statistically significant (at $\alpha = 10\%$) , and report the minimal statistical significance level for each variable.
- **c.** Interpret the estimated coefficient of `lprpblck`. Discuss whether or not you think it is economically significant (meaningful).

8. Restricted Model

- **a.** Run the single-variate regression of `lpsoda` (outcome) on `lprpblck`. Model (2):

$$\log(psoda_i) = \beta_0 + \beta_1 \log(prpblck_i) + u_i$$

- **b.** Compare the estimated $\hat{\beta}_1$ in Model (1) to that from Model (2). Is the effect of discrimination larger or smaller when we control for population density?
- **c.** Use an F-test to test the null hypothesis that $\beta_2 = 0$ in model (1)

9. Control Model

Add another explanatory variable to create Model (3):

$$\log(psoda_i) = \beta_0 + \beta_1 \log(prpblck_i) + \beta_2 \log(density_i) + \beta_3 \log(hseval_i) + \beta_4 \log(income_i) + u_i$$

- **a.** Report the results of adding the variables `lhseval` and `lincome` to the model. Write the new estimated regression model in the form of a formula.
- **b.** Test the hypothesis that these two newly added variables do not belong in the model.
- **c.** Compare $\hat{\beta}_1$, the estimated coefficient of `lprpblck`, to your estimate from Model (1).

10. Multicollinearity

- **a.** What is the linear correlation between the variables `lprpblck` and `ldensity`? What are the implications of this correlation (if any) on your OLS estimates for Model (1)?
- **b.** What is the linear correlation between the variables `lhseval` and `lincome`? What are the implications of this correlation (if any) on your OLS estimates for Model (3)?