Assignment 5

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 – NBA salaries

This question uses data on player statistics for 269 players in the National Basketball Association (NBA). Consider the model that predicts the number of points scored in a game (`points`) as a function of years in the league (`exper`), age, and years played in college (`coll`). The first version of this model (1) is estimated below and includes the variable experience squared (defined for you already as `expersq`).

```
> lm.points = lm( points ~ exper  + expersq + age + coll, data.nba)
> summary(lm.points)

Call:
lm(formula = points ~ exper + expersq + age + coll, data = data.nba)

Residuals:
     Min       1Q   Median       3Q      Max
-11.7979  -3.8132  -0.8622   3.4938  21.1195

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.98355    6.97093   5.018 9.58e-07 ***
exper        2.33702    0.40458   5.776 2.14e-08 ***
expersq     -0.07577    0.02343  -3.234 0.001377 **
age         -1.06174    0.29440  -3.606 0.000371 ***
coll        -1.27589    0.44957  -2.838 0.004892 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.497 on 264 degrees of freedom
Multiple R-squared:  0.1398,    Adjusted R-squared:  0.1268
F-statistic: 10.73 on 4 and 264 DF,  p-value: 4.536e-08
```

1. Model fit

a. Using the output above, conduct a test on the goodness-of-fit of the model, making sure to explicitly state the null hypothesis.

b. Is the model a good fit? At what significance level?

2. Quadratics

Holding age and years in college fixed, at what level of experience (years in the league) does one additional year in the league lower the number of points scored in a game? Is this reasonable?

3. Why might the variable for number of years in college have a negative, statistically significant, coefficient? (hint: the NBA can draft players before they graduate from college, and even right out of high school)

4. The second version of the model (2) includes the variable age squared (agesq).

```
> lm.points2 = lm( points ~ exper  + expersq + age + agesq + coll, data.nba)
> summary(lm.points2)

Call:
lm(formula = points ~ exper + expersq + age + agesq + coll, data = data.nba)

Residuals:
     Min       1Q   Median       3Q      Max
-11.5734  -4.0028  -0.7963   3.5126  20.9717

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 73.86287   35.84962   2.060  0.04035 *
exper        2.84383    0.61130   4.652 5.21e-06 ***
expersq     -0.12749    0.05232  -2.437  0.01548 *
age         -4.00994    2.68281  -1.495  0.13620
agesq        0.05426    0.04908   1.106  0.26991
coll        -1.30258    0.45003  -2.894  0.00412 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.495 on 263 degrees of freedom
Multiple R-squared:  0.1438,    Adjusted R-squared:  0.1275
F-statistic: 8.832 on 5 and 263 DF,  p-value: 9.161e-08
```

a. Is it necessary to include age squared in the model?

b. What does it tell us about the effect that age has given that we control for years in the league (exper) and years played in college (coll)?

5. The output below shows the regression of log(wage) (lwage) on points, exper, expersq, age, and coll.

```
> lm.wage = lm( lwage ~ exper + expersq + points + age + coll, data.nba)
> summary(lm.wage)

Call:
lm(formula = lwage ~ exper + expersq + points + age + coll, data = data.nba)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9442 -0.2891  0.1172  0.3943  1.7094

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.792880   0.845662   8.033 3.22e-14 ***
exper        0.219415   0.049770   4.409 1.52e-05 ***
expersq     -0.007166   0.002769  -2.588   0.0102 *
points       0.077893   0.007134  10.919  < 2e-16 ***
age         -0.048981   0.034955  -1.401   0.1623
coll        -0.040560   0.052899  -0.767   0.4439
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6372 on 263 degrees of freedom
Multiple R-squared:  0.4872,    Adjusted R-squared:  0.4775
F-statistic: 49.98 on 5 and 263 DF,  p-value: < 2.2e-16
```

a. Are age and coll statistically significant in the wage regression?

b. Given the results from the first model (1) what might we be concerned about with respect to the inclusion of experience and points in the wage equation?

Part B – Post-secondary education

Use the dataset **mosad_v2026.xlsx** for this question (Mosad means institution in Hebrew). It is a sample dataset from the Israeli central bureau of statistics. The survey contains 31,000 individuals who completed their undergraduate degrees in the early 2000's. 46% studied at universities and the rest were at colleges ("mechlalot") which are either government subsidized or private.

1. Data exploration
- **a.** Create a dummy variable for each field of study (social science, business, law, engineering, computer science, or design) using the categorical variable `thum`. (*hint*: use either as.numeric() or ifelse() and make sure that your dummy variables take values of 0 or 1 and not TRUE or FALSE)
- **b.** Create dummy variables for each type of degree-granting institution (university, subsidized college, and non-subsidized college) using the variable `mosad_type4`
- **c.** Create the outcome variable `log_wageafter3`: the log of the average monthly wage three years after completing the degree (wageafter3). Provide a statistical description of the sample distribution for this variable.
- **d.** How does the distribution of `log_wageafter3` change for each type of degree-granting institution?
- **e.** Create a variable `experience` measuring the total number of months worked for the first three years after graduation (including year of degree completion) (*hint*: unlike in part c, you should add up all four variables on number of months worked)
- **f.** How does the distribution of experience (from 1e), change for each type of degree-granting institution?
- **g.** What is your opinion? Are students who graduate university vs. those from colleges similar in how much they earn three years out?

2. Wage regression
- **a.** Run a log-level regression using your outcome variable (log monthly wage 3 years after completing the degree). The explanatory variables are: degree-granting institution (use private as the base category), field of study (use computer science as the base category), gender, and experience. Use the respective dummy variables you created in 1. (*hint*: your final model should have 10 parameters including the intercept)
- **b.** Interpret the coefficient estimates and report on their significance
- **c.** Is there a significant difference between the three types of degree-granting institutions? Conduct the appropriate hypothesis test.

3. Omitted variables
- **a.** Rerun the regression in (2a) this time adding the explanatory variable for IQ proxy
- **b.** How does inclusion of IQ proxy affect the estimation results?
- **c.** A colleague of yours claims that the difference in wages between university and college graduates are not due to gaps in quality of institution but rather because better students opt to attend university. Do your regression results support this claim?

Questions 4 – 6 can be answered after lecture on interaction terms

4. Interactions
- **a.** Use interaction terms to run a regression to check if the rate of returns of wage from experience is different between university and colleges. (*hint*: create new variables representing the interaction terms between your dummies on degree-granting institution and experience)

b. Use an F-test to check the if experience has a different rate of return depending on the institution type

5. Separate regressions
- **a.** run the regression from (3a) separately for each type of degree-granting institution: university, subsidized college, and non-subsidized college.
- **b.** Use the results from (5a) to compare the predicted wage between someone who studied at a non-subsidized college versus a university. The person is a woman with 35 months of experience and a psychometric entrance exam score of 710 who studied business.
- **c.** Repeat (5b) above except this time use the regression results from (3a)
- **d.** Explain the differences between the predictions in 5b and 5c.

6. Conduct a Chow-Fisher test of structural change between the three degree-granting institutions and address the significance of the test. You can use either version of the test (optional: try both!)