Practice Tasks for Weeks 10 & 11

Download and open the file called *unemployed.csv*! The data originate from 1982 and cover 252 American households. The variables are defined as follows:

UCOMP: Amount of unemployment compensation received (in USD)

UHOURS: Number of hours the household head spent unemployed

HEADY: Earnings of the household head (in USD)

EDUC: Number of years of schooling completed by the household head

MALE: 1 if the household head is male, 0 if female

MARRIED: Marital status of the household head: 1 if married, 0 otherwise

FAMSIZE: Family size (number of individuals)

WHITE: 1 if the household head is white, 0 otherwise

SPOUSEY: Earnings of the spouse (in USD)

Estimate a linear regression model using OLS, where the dependent variable is UCOMP, and the explanatory variables are the remaining eight listed above.

1. Report the estimated coefficient of the MARRIED variable, and interpret its coefficient.
2. Identify the variable with the strongest and weakest multicollinearity. For these variables, specify what percentage of their variance cannot be explained by the other explanatory variables.
3. List the non-significant variables at the $\alpha = 1\%$ significance level. What kind of statistical measure was applied to identify these variables and why?
4. Using at least two different methods, examine whether the non-significant variables identified in task 3 can be jointly excluded from the model.
5. Test the hypothesis that the sum of the coefficients of the MALE and MARRIED variables is equal to zero at the $\alpha = 5\%$ level. Clearly state the null (H_0) and alternative hypotheses (H_1), provide the value of the test statistic, the p-value, and the decision. Explain the economic interpretation of the hypothesis test result.
6. Extend the model by including the square of the UHOURS variable, and the interaction between UHOURS and MALE. In this new model, determine the marginal effect of the number of hours the household head spends unemployed on unemployment compensation for a female household head with an average number of unemployed hours. Interpret the result.
7. Examine with the appropriate statistical test and with graphical tools whether the extension of the model in task 6 is justified!

The file *statistics_scores.xlsx* contains data on 166 Business Informatics students participating in the course Statistics II. Each student can be clearly classified into exactly one of the four seminar attendance categories listed below. The variables are defined as follows:

- **Score**: The student's score on the third Statistics II midterm exam (maximum score: 10 points)
- **PrevGrade**: The student's grade in the Statistics I course
- **D_Always**: Dummy variable equal to 1 if the student always attended seminars, 0 otherwise
- **D_Mostly**: Dummy variable equal to 1 if the student missed no more than four seminars, 0 otherwise
- **D_Sometimes**: Dummy variable equal to 1 if the student missed 5–10 seminars, 0 otherwise
- **D_Never**: Dummy variable equal to 1 if the student missed 11–12 seminars (i.e., never or only once attended seminar, given the 12-week semester), 0 otherwise

Estimate a linear regression model using OLS, where the dependent variable is Score, and the explanatory variables include: PrevGrade, the dummy variables describing seminar attendance and the interaction terms between PrevGrade and the dummy variables. Use students who never attended class as the reference category.

1. For two students who mostly attend seminars, how much higher is the expected midterm score for the student whose Statistics I grade is higher by one?
2. Interpret the interaction coefficient between PrevGrade and D_Sometimes.
3. Test whether the interaction terms are jointly significant at the $\alpha = 1\%$ level. State the null (H_0) and alternative hypotheses (H_1). Provide the value of the test statistic, the p-value, and your conclusion.
4. In the reduced model (from task 3), test whether the D_Sometimes variable is significant at the $\alpha = 5\%$ level. What does the result imply? Is it worthwhile to attend seminars only sometimes in terms of the midterm exam score?
5. Use the simpler version of the White test to examine the presence of heteroskedasticity in the original (full) model at the $\alpha = 5\%$ level. If heteroskedasticity is detected, address it using the Generalized Least Squares (GLS) method. If heteroskedasticity is not detected, retain the original model.
6. What alternative method can be used to address heteroskedasticity? How do the results differ compared to the GLS method?

Download the file employee.xlsx to your computer and open it in R. The dataset contains information on 474 employees, and the data was collected in 2000. The variables are defined as follows:

- **CurrentSalary**: The employee’s annual salary at the time of the survey, measured in thousands of dollars
- **StartSalary**: The employee’s initial annual salary (in thousands of dollars); assume that no employee ever changed jobs
- **Settlement**: The location of the employee's workplace (Budapest, Town, or Village)

Estimate a linear model using OLS, where the dependent variable is CurrentSalary, and the explanatory variables are StartSalary, Settlement and their interaction. Use Village as the reference category for the settlement variable.

1. What is the marginal effect of StartSalary for an employee working in Budapest? Interpret the result!
2. Use the appropriate version of the Breusch–Pagan test (with or without Koenker correction) to test for heteroskedasticity in the residuals at the $\alpha = 5\%$ level. Write down the null hypothesis (H_0) and alternative hypothesis (H_1), the p-value, and your decision for all the tests performed in this task.
3. If heteroskedasticity is detected, address it using White’s robust standard errors. If no heteroskedasticity is detected, continue using the original model.
4. What alternative methods can be used to handle heteroskedasticity? How do the results differ from those obtained using White’s method?
5. Estimate a log-lin model, where the dependent variable is the logarithm of CurrentSalary, and the explanatory variables are StartSalary and Settlement. Interpret the coefficient of StartSalary.
6. Estimate a log-log model, where both the dependent variable (CurrentSalary) and the independent variable (StartSalary) appear in logarithmic form, and include Settlement as well. Interpret the coefficient of StartSalary.
7. Test the hypothesis that the coefficient of StartSalary is significantly lower than 1 in the log-log model. State the null (H_0) and alternative hypothesis (H_1), the test statistic and p-value, and your decision. What is the economic interpretation of the result?
8. Which model specification is more appropriate — log-log or log-linear? Support your answer graphically and by conducting the appropriate hypothesis test.