Econometrics 2018, Final Exam, Time to complete : 60 minutes
Minimum points required for a positive grade: 20
Name and student ID ____
Signature ________________________

This exam contains 4 pages (including this cover page) and 2 questions.
Total of points is 40.

Grade Table (for teacher use only)

Question	Points	Score
Problem 1	20	
Problem 2	20	
Total:	40	

`https://github.com/feb-uni-sofia/econometrics2018-exam-2018-06-20`

The dataset `trips` is based on the 2016 NYC Yellow Cab trip record data and provides a sample of 988 taxi trips. For each trip you are given the following measurements:

1. **trip_duration** : Trip duration in seconds.

   day : Day of week with two categories: 'workday' and 'weekend'.

   (a) (1 point) Compute the average trip duration and write down the *result* (not the R-code)!

   (b) (2 points) Create a new variable in the dataset called `trip_duration_minutes` that contains the trip duration in minutes. Create another variable called `weekend` that is TRUE (logical) if the trip took place on a weekend and FALSE (logical) otherwise.

   (c) (1 point) Fit the linear regression model :

$$\text{trip\_duration\_minutes}_i = \beta_0 + \beta_1 \text{weekend}_i + u_i$$

   with $i = 1, \ldots, n$ and where u_i are independent random terms with zero mean and constant variance.

   (d) (2 points) Write down the estimated regression *equation.*

(e) (5 points) Explain the *meaning* of the estimated regression coefficients *in relation to the data*.

(f) (6 points) A taxi driver claims that trips on workdays take longer on average than trips during the week. Express this hypothesis in terms of the model coefficients and test it at a 90% significance level (10% error probability) using the output from the model fitted in (c) . Write down the test-statistic and the p-value of the test and explain your decision to reject or not to reject the hypothesis.

(g) (3 points) The taxi driver confronts you with the claim that your test is invalid because is assumes that the trip duration is normally distributed in both groups (weekends/workdays). Respond to that claim.

2. The dataset `store` contains data about a Rossmann drug store located in Germany. Each observation corresponds to one of 942 days from 2013-01-01 to 2015-07-31 and consists of the following measurements:

   Sales (numeric): Store sales in EUR.

   SchoolHoliday (0/1): Equals `1` if the day was a school holiday in the state where the store is located and `0` otherwise.

   Customers (numeric): Number of customers for the day.

   (a) (1 point) Fit the linear regression model:

   $$\text{Sales}_i = \beta_0 + \beta_1 \text{SchoolHoliday}_i + \beta_2 \text{Customers}_i + u_i \tag{1}$$

   with $i = 1, \ldots, n$ and where u_i are independent random terms with zero mean and constant variance.

   (b) (2 points) Write down the estimated regression *equation*.

   (c) (5 points) Explain the *meaning* of the regression coefficients *in relation* to the data (generic answers bring no points).

   (d) (5 points) The store manager has estimated the following model:

   $$\text{Sales}_i = \beta_0 + \beta_1 \text{SchoolHoliday}_i + u_i \tag{2}$$

   with the same assumptions about u_i as in (1) and finds a much higher estimated coefficient for `SchoolHoliday`. Estimate the model in (2) and compare the coefficient for `SchoolHoliday` with the corresponding coefficient from (1). How would you explain the difference between the two coefficients?

(e) (2 points) Create a new variable called `CustomersCentered` that equals the number of customers minus the average number of customers per day and fit the model:

$$\text{Sales}_i = \beta_0 + \beta_1 \text{SchoolHoliday}_i + \beta_2 \text{CustomersCentered}_i + u_i$$

(f) (5 points) The manager asks you to estimate the sales for the store on a non-school-holiday day with the usual (average) number of customers. Write down your estimated sales amount together with an approximate 95% confidence interval and explain to the manager why you are calculating this interval and what it means.