

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	

---

1. The dataset **trips** is based on the 2016 NYC Yellow Cab trip record data and provides a sample of 4985 taxi trips. For each trip you are given the following measurements:

**trip\_duration** : Trip duration in seconds.

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

**passenger\_count** : Number of passengers on the trip.

- (a) (1 point) Compute the average number of passengers per trip and write down the *result* (not the R-code)!

- (b) (2 points) Create a new variable in the dataset called **workday** that equals **TRUE** (logical) if the trip took place on a workday and **FALSE** (logical) otherwise.

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

$$\text{trip\_duration}_i = \beta_0 + \beta_1 \text{workday}_i + u_i \quad (1)$$

with  $i = 1, \dots, n$  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*. Pay attention to the scales of the variables.

- (f) (6 points) A taxi driver claims that trips on weekends are shorter on average than trips during the week. Express this hypothesis in terms of the model coefficients and test it at a 95% significance level (5% 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 it 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.

**Promo** (0/1): Equals 1 if there was a promotion on that day in the store and is 0 otherwise.

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

- (a) (1 point) Create a new variable in the dataset called **Sales1000** that equals **Sales** divided by 1000 and fit the model

$$\text{Sales1000}_i = \beta_0 + \beta_1 \text{Promo}_i + \beta_2 \text{Customers}_i + u_i \quad (2)$$

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

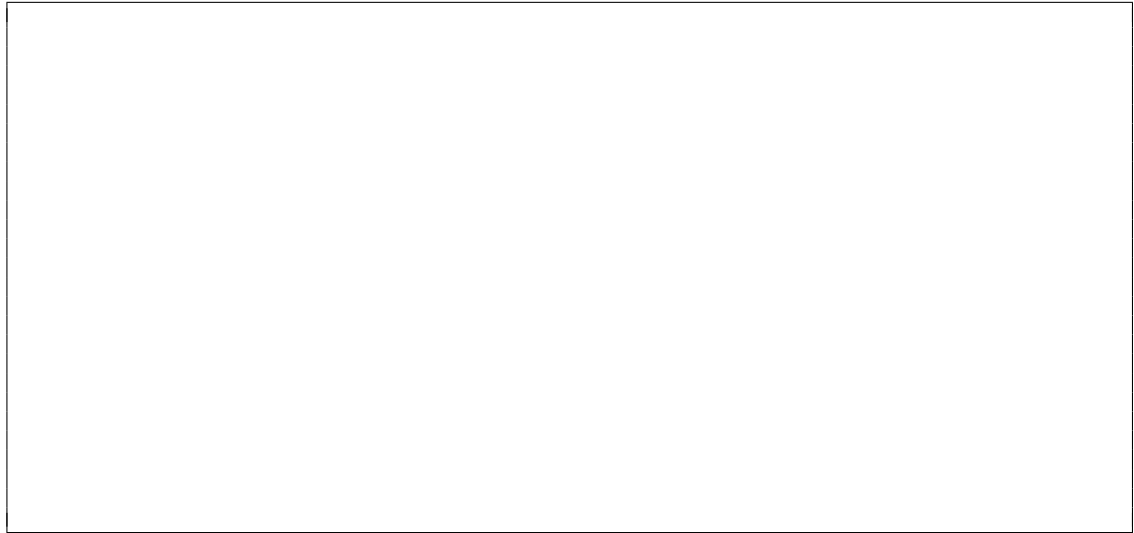
- (b) (2 points) Write down the estimated regression *equation*.

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

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

$$\text{Sales1000}_i = \beta_0 + \beta_1 \text{Promo}_i + u_i \quad (3)$$

with the same assumptions about  $u_i$  as in (2) and finds a much higher estimated coefficient for **Promo**. Estimate the model in (3) and compare the coefficient for **Promo** with the corresponding coefficient from (2). How would you explain the difference between the two coefficients?



- (e) (2 points) Create a new variable called **Customers300** that equals the number of customers minus 300 and fit the model:

$$\text{Sales1000}_i = \beta_0 + \beta_1 \text{Promo}_i + \beta_2 \text{Customers300}_i + u_i$$

- (f) (5 points) Then manager asks you to estimate the sales for the store on a non-promotion day with 300 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.

