

# Assignment 1

The purpose of this assignment is for you to examine the determinants of the price of second-hand cars. A dataset of  $n = 804$  cars sold in 2014 is available. It records the following characteristics of each car  $i$ , where  $i = 1, \dots, n$ :

$price_i$	selling <b>price</b> of car $i$
$mileage_i$	number of <b>miles</b> car $i$ has been driven
$cylinder_i$	number of <b>cylinders</b> of car $i$
$liter_i$	<b>cylinder volume</b> of car $i$
$cruise_i$	= 1 if car $i$ has <b>cruise control</b> , = 0 if not
$sound_i$	= 1 if car $i$ has <b>quality loudspeakers</b> , = 0 if not
$leather_i$	= 1 if car $i$ has <b>leather seats</b> , = 0 if not

- (a) Plot a **histogram** of **price** and compute the **mean** as well as the **median** of **price**. Discuss the latter two in the light of the histogram's shape.
- (b) Denote the **population mean price** of cars with a **cylinder** volume of more than 3 liters by  $\mu_p^{large}$  and that of cars with a cylinder volume of less than 3 liters by  $\mu_p^{small}$ . Test the null hypothesis

$$H_0: \mu_p^{large} = \mu_p^{small}$$

at a significance level of  $\alpha = 0.01$ . Hint: see Stock & Watson (2015, Section 3.4).

- (c) Consider the linear regression model

$$price_i = \beta_0 + \beta_1 liter_i + u_i, \quad (1)$$

for  $i = 1, \dots, n$ , making the usual three least squares assumptions (LSA's).

Estimate the model in (1) by OLS, computing heteroscedasticity-robust standard errors in the process and summarising the results in an output table. Interpret the estimated coefficient  $\hat{\beta}_1$ .

- (d) Create a scatter plots of **price** vs. **liter** and add to it the estimated sample regression line. Also add to the plot the following two estimated regression lines:

- (i) one with smallest intercept and steepest slope,
- (ii) the other with largest intercept and shallowest slope

contained in the 95% heteroscedasticity-robust confidence intervals of the coefficients.

- (e) Consider now the linear regression model

$$price_i = \beta_1 liter_i + u_i, \quad (2)$$

for  $i = 1, \dots, n$ , again making the usual three least squares assumptions (LSA's).

Estimate the model in (2) by OLS, computing heteroscedasticity-robust standard errors in the process and summarising the results in an output table. Interpret the estimated coefficient  $\hat{\beta}_1$ .

- (f) Create a new scatter plot of *price* vs. *liter* that does not use the default layout but has both axis begin at the origin. Add to this plot the sample regression of (2). Interpret the graph.
- (g) Interpret the OLS fit of the model in (1). Why do the values of  $R^2$  and  $SER$  make little sense in model (2)?
- (h) Consider now the extended model

$$price_i = \beta_0 + \beta_1 liter_i + \beta_2 mileage_i + \beta_3 cruise_i + \beta_4 sound_i + \beta_5 leather_i + u_i, \quad (3)$$

for  $i = 1, \dots, n$ . Assume, here and in all subsequent parts below, homoscedasticity. What could be good reasons for including the extra explanatory variables?

- (i) Estimate the model in (3) by OLS and display your results. Has the effect of *liter* on *price* changed relative to the model in (1)? What might have been a good reason for not adding *cylinder* as an explanatory variable to the model, too?
- (j) Create a dummy variable  $D$  for those cars which were driven for more than 30,000 miles. Use  $D$  to illustrate the dummy variable trap in a regression model for *price*. Explain carefully your reasoning.

## Points

question	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	compile2pdf	total
points	2	2	2	2	2	2	3	3	3	2	2	25