TU DORTMUND
Fakultät Statistik
Prof. Dr. Sonja Kuhnt
Dr. Paul Wiemann
Dr. Birte Hellwig
M.Sc. Hendrik Dohme

Winter semester 2021/22

Project III
Regression Analysis

The given dataset (`vwcars.csv`) provides information on cars sold on a used car platform (Exchange and Mart) in the United Kingdom in the year 2020. It contains listings of the models *Up*, *Passat*, and *T-Roc* by the manufacturer Volkswagen (VW). It is an extract of a larger data set available on `kaggle.com`.

The data set contains the following nine variables:

price - the price of the cars, in 1000 GBP (£)
year - the year that the car was first registered in
mileage - the total distance (in 1000 miles) the car has been driven
mpg - the distance (in miles) the car can travel with one gallon (uk) of fuel
fuelType - the type of fuel the car consumes
engineSize - the size of the car's engine, in liters
tax - the amount of the annual tax (Vehicle Excise Duty) to be paid for the car
transmission - the type of gearbox the car has

Tasks 1: Data preparation

i. Compute the log of the car price (`logprice`).

ii. Use in this project as the fuel consumption measure liters per 100 kilometers (`lp100`) instead of miles per gallon (`mpg`). The formula for the conversion is given by

$$\texttt{lp100} = \frac{282.48}{\texttt{mpg}}.$$

iii. Use the variable `year` to calculate the cars' age and use the new variable in the following.

Tasks 2: Linear regression

i. Decide whether to use the raw price `price` or the log-transformed price `logprice` as the response variable in the linear regression analysis. To make your choice it is advisable to estimate the full model (e.g., use all suitable explanatory variables) for both candidates. Consult model diagnostic tools to choose which one is more in line with the assumptions made on the linear model.

ii. Once you have settled on the transformation of the response variable, find the "best" set of explanatory variables for the price using Best Subset Selection. Use the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) as the selection criteria. Compare the included variables of the two models.

iii. Estimate the “best” linear model for the dependent variable w.r.t. the BIC from ii. Interpret the coefficients of the model and their statistical significance, provide confidence intervals for the regression parameters and evaluate the goodness of fit.

Submission

Submission of the report and the corresponding (executable and commented) program code until *Friday, January 28, 2022, 08:30 am*, in Moodle.