

Statistical Learning

Exercises sheet 2 – Linear Regression

1. This question involves the use of simple linear regression on the **Auto** data set.
 - (a) Use the `lm()` function to perform a simple linear regression with **mpg** as the response and **horsepower** as the predictor. Use the `summary()` function to print the results. Comment on the output. For example:
 - i. Is there a relationship between the predictor and the response?
 - ii. How strong is the relationship between the predictor and the response?
 - iii. Is the relationship between the predictor and the response positive or negative?
 - iv. What is the predicted **mpg** associated with a **horsepower** of 98? What are the associated 95% prediction intervals?
 - (b) Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.
 - (c) Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.
2. This problem involves the **Boston** data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.
 - (a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.
 - (b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?
 - (c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.
 - (d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X , fit a model of the form
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon.$$
3. This question should be answered using the **Carseats** data set.
 - (a) Fit a multiple regression model to predict **Sales** using **Price**, **Urban**, and **US**.
 - (b) Provide an interpretation of each coefficient in the model. Be careful, some of the variables in the model are qualitative!
 - (c) Write out the model in equation form, being careful to handle the qualitative variables properly.
 - (d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?
 - (e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.
 - (f) How well do the models in (a) and (e) fit the data?
 - (g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).
 - (h) Is there evidence of outliers or high leverage observations in the model from (e)?