Homework1 - Linear Regression

Lu Haibo

Mondy, Mar 4, 2019

Deadline: Mar 11, 2019

1. In this exercise you will create some simulated data and will fit simple linear regression models to it.

(a) create a tibble, which has 2 variables and 100 observations: $x \sim Uniform(0,1)$, and $y = -1 + 0.5x + \epsilon$, where $\epsilon \sim N(0, 0.25)$.

(b) Using `ggplot`, create a scatterplot displaying the relationship between x and y.

(c) Using the `lm()` function and the `broom` package, fit a least squares linear model to predict y using x. Comment on the model obtained. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1?

(d) Display the least squares line on the scatterplot using `ggplot`.

(e) Now fit a polynomial regression model that predicts y using x and x^2. Is there evidence that the quadratic term improves the model fit? Explain your answer. (put the result with (c) using package `stargazer`)

(f) Repeat (a)–(e) after modifying the data generation process in such a way that there is more (or less) noise in the data. The model should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term ϵ. Describe your results. (Hint: you can use the `purrr` package)

(g) What are the confidence intervals for β_0 and β_1 based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.

2. This problem focuses on the *collinearity* problem. First, perform the following commands in `R`:

(a) What is the correlation between `x1` and `x2`? Create a scatterplot displaying the relationship between the variables.

(b) Using this data, fit a least squares regression to predict `y` using `x1` and `x2` . Describe the results obtained. What are $\hat{\beta}_0, \hat{\beta}_1$, and $\hat{\beta}_2$? How do these relate to the true β_0, β_1 and β_2? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

(c) Now fit a least squares regression to predict `y` using only `x1`. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

(d) Now fit a least squares regression to predict `y` using only `x2`. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

(e) Do the results obtained in (c)–(e) contradict each other? Explain your answer.

(f) Now suppose we obtain one additional observation, which was unfortunately mismeasured.

Re-fit the linear models from (b) to (d) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

3. This problem involves the `Boston` data set, which is inclued in the `MASS` library. We will now try to predict per capita crime rate (`crim`) 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

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$