PREDICTIVE ANALYTICS

Problem Set 2: Regression

1 Problem to demonstrate that the population regression line is fixed, but least square regression line varies

Suppose the population regression line is given by $Y = 2 + 3x$, while the data comes from the model $y = 2 + 3x + \epsilon$.

Step 1: For x in the range [5,10] graph the population regression line.

Step 2: Generate $x_i (i = 1, 2, .., n)$ from $Uniform(5, 10)$ and $\epsilon_i (i = 1, 2, .., n)$ from $N(0, 4^2)$. Hence, compute $y_1, y_2, .., y_n$.

Step 3: On the basis of the data $(x_i, y_i)(i = 1, 2, .., n)$ generated in Step 2, report the least squares regression line. Graph the least-squares regression line over the population regression line.

Step 4: Repeat steps 2-3 five times.

Interpret the findings.

Take $n = 50$. Set the seed as seed=123.

2 Problem to demonstrate that $\hat{\beta}_0$ and $\hat{\beta}$ infact minimises RSS

Step 1: Generate x_i from Uniform(5, 10) and mean centre the values. Generate ϵ_i from $N(0, 1)$. Calculate $y_i = 2 + 3x_i + \epsilon_i$, i = 1,2,.., n. Take n=50 and seed=123.

Step 2: Now imagine that you only have the data on $(x_i, y_i), i = 1, 2, .., n$, without knowing the mechanism that was used to generate the data in step 1. Assuming a linear regression of the type $y_i = \beta_0 + \beta x_i + \epsilon_i$, and based on these data $(x_i, y_i), i = 1, 2, .., n$, obtain the least squares estimates of β_0 and β.

Step 3: Take a large number of grid values of (β_0, β) that also include the least squares estimates obtained from step 2. Compute the RSS for each parametric choice of (β_0, β), where $RSS = (y_1 - \beta_0 - \beta x_1)^2 + (y_2 - \beta_0 - \beta x_2)^2 + ...(y_n - \beta_0 - \beta x_n)^2$. Find out for which combination of (β_0, β), RSS is minimum.

3 Problem to demonstrate that least square estimators are unbiased

Step 1: Generate $x_i(i = 1, 2, .., n)$ from $Uniform(0, 1)$, $\epsilon_i(i = 1, 2, .., n)$ from $N(0, 1)$ and hence generate y using $y_i = \beta_0 + \beta x_i + \epsilon_i$. (Take $\beta_0 = 2, \beta = 3$).

Step 2: On the basis of the data $(x_i, y_i)(i = 1, 2, .., n)$ generated in Step 1, obtain the least square estimates of β_0 and β.

Repeat Steps 1-2, $R = 100$ times. In each simulation obtain $\hat{\beta}_0$ and $\hat{\beta}$. Finally, the least-square estimates will be given by the average of these estimated values. Compare these with the true β_0 and β and comment.

Take $n = 50$ and seed=123.

4 Problem to demonstrate the utility of multiple linear regression

Attach "Carseats" data from ISLR library in R.

(a) Select the quantitative predictors from the data set "Carseats". Run the linear regression equations of "sales" on each of the quantitative predictors separately.

(b) Run multiple linear regression of sales on the quantitative predictors.

(c) Provide the output of (a) and (b) in a single table using stargazer.

(d) Compare the results in (a) and (b) and comment.

(e) Comment on whether at least one predictor is useful in predicting the response.

(f) Comment on the goodness of fit of the multiple linear regression model.

(g) Find the confidence interval of average sales of carseats corresponding to the average values of all the quantitative predictors.

(h) Find the prediction interval of Sales of carseats corresponding to the given values of the quantitative predictors for store 1 .

5 Problem to demonstrate the role of qualitative (nominal) predictors in addition to quantitative predictors in multiple linear regression

Attach "Credits" data from R. Identify the nature of all the variables that appear in the data set.

Regress "balance" on

(a) "gender" only.

(b) "gender" and "ethnicity" .

(c) "gender", "ethnicity", "income".

(d) Output all the regressions in (a)-(c) in a single table using stargazer. Comment on the significant coefficients in each of the models.
(e) Explain how gender affects "balance" in each of the models (a)- (c) .
(f) Compare the average credit card balance of a male African with a male Caucasian on the basis of model (b).
(g) Compare the average credit card balance of a male African with a male Caucasian when each earns 100,000 dollars. For comparison, use the model in (c).
(h) Compare and comment on the answers in (f) and (g) (i) Predict the credit card balance of a female Asian whose income is 2000,000 dollars.
(j) Check the goodness of fit of the different models in (a) -(c) in terms of AIC, BIC and adjusted R^2. Which model would you prefer?

6 Problem to demonstrate the role of qualitative (ordinal) predictors in addition to quantitative predictors in multiple linear regression

Consider "diamonds" data set in R. It is in the ggplot2 package. Make a list of all the ordinal categorical variables. Identify the response and predictors.
(a) Run a linear regression of the response on the quality of cut. Write the fitted regression model.
(b) Test whether the expected price of diamond with premium cut is significantly different from that of the ideal cut.
(c) What is the expected price of a diamond of ideal cut?
(d) Modify the regression model in (i) by incorporating the predictor table. Write the fitted regression model.
(e) Test for the significance of "table" in predicting the price of diamond.
(f) Find the expected price of a diamond with an average table value and which is of fair cut.

7 Problem to demonstrate the impact of ignoring interaction term in multiple linear regression

Consider a simulation setting where the data is generated as follows:

Step 1: Generate x_{1i} from Normal(0,1) distribution, $i = 1, 2, .., n$

Step 2: Generate x_{2i} from Bernoulli (0.3) distribution, $i = 1, 2, .., n$

Step 3: Generate ϵ_i from Normal(0,1) and hence generate the response $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 (x_{1i} \times x_{2i}) + \epsilon_i$, $i = 1, 2, , , n$.

Step 4: Run two separate multiple linear regressions (i) using the model in Step 3 and (ii) using the model in Step 3 without the interaction term.

Repeat Steps 1-4 , $R = 1000$ times. At each simulation compute the MSE for the correct model (i.e. model with the interaction term) and the naive model (i.e. the model without the interaction term). Finally find the average MSE's for each model. From the output, demonstrate the impact of ignoring the interaction term.

Carry out the analysis for $n = 100$ and the following parametric configurations: $(\beta_0, \beta_1, \beta_2, \beta_3) = (-2.5, 1.2, 2.3, 0.001)$, (-2.5, 1.2. 2.3, 3.1). Set seed as 123.

8 Problem to demonstrate the utility of non-linear regression over linear regression

Get the fgl data set from "MASS" library.

(a) Considering the refractive index (RI) of "Vehicle Window glass" as the variable of interest and assuming linearity of regression, run multiple linear regression of RI on different metallic oxides. From the p value, report which metallic oxide best explains the refractive index.

(b) Can you further improve the regression of the refractive index of "Vehicle Window glass" on the predictor chosen by you in part (a)?

9 Problem to demonstrate multicollinearity

Consider the Credit data in the ISLR library. Choose balance as the response and Age, Limit and Rating as the predictors.
(a) Make a scatter plot of (i) Age versus Limit and (ii) Rating Versus Limit. Comment on the scatter plot.
(b) Run three separate regressions: (i) Balance on Age and Limit (ii) Balance on Age, Rating and Limit (iii) Balance on Rating and Limit. Present all the regression output in a single table using stargazer. What is the marked difference that you can observe from the output?
(c) Calculate the variance inflation factor (VIF) and comment on multicollinearity.

10 Problem to demonstrate the detection of outlier, leverage and influential points

With reference to Problem 4, detect outliers, leverage points and influential points if any.

11 Problem to demonstrate the utility of K nearest neighbour regression over least squares regression

Consider a setting with $n = 1000$ observations. Generate
(i) x_{1i} from $N(0, 2^2)$ and x_{2i} from Poisson($\lambda = 1.5$).
(ii) ϵ_i from $N(0, 1)$.
(iii) $y_i = -2 + 1.4x_{1i} - 2.6x_{2i} + \epsilon_i$.
Split the data into train and test sets. Keep the first 800 observations as training data and the remaining as test data. Work out the following:

1. Fit a least squares regression of y on x_1 and x_2. Calculate test MSE.

2. Fit a KNN model with $k = 1, 2, 5, 9, 15$. Calculate test MSE for each choice of k.

Suppose the data in Step 3 is generated as :
$y_i = -2 + 1.4x_{1i} - 2.6x_{2i} + 2.9x_{1i}^2 + 3.1\exp(x_{2i}) - 1.5x_{1i}x_{2i}^2$.
Work out the problems in (1) and (2). Compare and comment on the results.