Introduction to Multivariate Data Analysis

Final Project Report

José Antonio Álvarez Ocete - 917933752
jocete@ucdavis.edu

June 9, 2020

Contents

1 Multiple Linear Regression

1.1 Introduction

In this first section, we will conduct a multiple linear regression following question 1 of the 5th homework assignment. We will estimate the regression coefficients (betas) following different several methods seen during the lectures and we will provide an estimation for a new response.

1.2 Summary

For this first analysis, I've selected the Battery Failure dataset. In this example, we want to predict the cycles of life of a certain battery before it fails. We are provided the following variables: Charge rate (amps), discharge rate (amps), depth of discharge (% of rated ampere-hours), temperature (ºC), and end of charge voltage (volts). These are the first three rows of data:

Z_1 Charge rate (amps)	Z_2 Discharge Rate (amps)	Z_3 Depth of Discharge (% of rated ampere-hours)	Z_4 Temperature (ºC)	Z_5 End of Charge Voltage (volts)	Y Cicles to failure
0.375	3.13	60.0	40	2.00	101
1.000	3.13	76.8	30	1.99	141
1.000	3.13	60.0	20	2.00	96

1.3 Analysis

(1) Find the least square estimate beta hat

We obtain the least square estimate beta hat following out notes:

$$\hat{\vec{\beta}} = (Z^T \cdot Z)^{-1} \cdot Z^T \cdot \vec{Y}$$

Obtaining:

$$\hat{\vec{\beta}} = \begin{pmatrix} -2937.7571 \\ -33.7934 \\ -0.1798 \\ -1.7397 \\ 7.0627 \\ 1529.2897 \end{pmatrix}$$

(2) Find the R^2 statistic

We use:

$$R^2 = \frac{||\hat{\vec{Y}} - \bar{Y} \cdot \vec{1_n}||^2}{||\vec{Y} - \bar{Y} \cdot \vec{1_n}||^2}$$

Obtaining $R^2 = 0.4799$.

(3) Find sigma_hat_square and estimated Cov(beta square)

We use:

$$\hat{\sigma}^2 = \frac{1}{n-r-1}||\hat{\vec{\epsilon}}||^2$$

and

$$\hat{Cov}(\hat{\vec{\beta}}) = \hat{\sigma}^2(Z^TZ)^{-1}$$

We obtain the following:

$$\hat{\sigma}^2 = 7138.186$$

$$\hat{Cov}(\hat{\vec{\beta}}) = \begin{pmatrix} 1.633e+07 & -2933.74 & 2980.4460 & -34.78143 & -991.73637 & -8.160e+06 \\ -2.934e+03 & 1880.55 & 18.5503 & 17.34897 & 10.28445 & -1.764e+02 \\ 2.980e+03 & 18.55 & 193.4117 & -3.23257 & 0.34449 & -1.696e+03 \\ -3.478e+01 & 17.35 & -3.2326 & 1.79944 & -0.08092 & -4.242e+01 \\ -9.917e+02 & 10.28 & 0.3445 & -0.08092 & 3.89193 & 4.549e+02 \\ -8.160e+06 & -176.39 & -1695.7845 & -42.42251 & 454.86652 & 4.081e+06 \end{pmatrix}$$

(4) 95% confidence interval for each β_j

We use one at a time confidence intervals for the betas:

$$\beta_j \in [\hat{\beta}_j \pm \hat{\sigma} \cdot \sqrt{\omega_{jj}} \cdot t_{n-r-1}(\frac{\alpha}{2})]$$

Obtaining:

$$\begin{gathered} \beta_0 \in [-11604, 5729] \\ \beta_1 \in [-126.8, 59.22] \\ \beta_2 \in [-30.01, 29.65] \\ \beta_3 \in [-4.617, 1.137] \\ \beta_4 \in [2.831, 11.29] \\ \beta_5 \in [-2804, 5862] \end{gathered}$$

(5) 95% simultaneous confidence intervals for all betas based on the confidence region

Using the formula from the notes:

$$\beta_j \in [\hat{\beta}_j \pm \hat{\sigma} \cdot \sqrt{\omega_{jj}} \cdot \sqrt{(r+1) \cdot F_{r+1,n-r-1}(\alpha)}]$$

We obtain:

$$\begin{gathered} \beta_0 \in [-19640, 13764] \\ \beta_1 \in [-213, 145.5] \\ \beta_2 \in [-57.67, 57.31] \\ \beta_3 \in [-7.285, 3.805] \\ \beta_4 \in [-1.092, 15.22] \\ \beta_5 \in [-6822, 9880] \end{gathered}$$

(6) 95% simultaneous confidence intervals for all betas based on the Bonferroni correction

We compute a final set of intervals for the ebtas using the Bonferroni correction:

$$\beta_j \in [\hat{\beta}_j \pm \hat{\sigma} \cdot \sqrt{\omega_{jj}} \cdot t_{n-r-1}(\frac{\alpha}{2(r+1)})]$$

We obtain:

$$\begin{aligned}
\beta_0 &\in [-15338, 9462] \\
\beta_1 &\in [-166.9, 99.29] \\
\beta_2 &\in [-42.86, 42.5] \\
\beta_3 &\in [-5.856, 2.377] \\
\beta_4 &\in [1.009, 13.12] \\
\beta_5 &\in [-4670, 7729]
\end{aligned}$$

(7) Test $H_0 : \beta_1 = \beta_2 = 0$ at significance level $\alpha = 0.05$

Using this matrix for the linear transformation:

$$C = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

We can compute the F-test statistic:

$$\vec{\beta}_{(2)}^T \Omega_{22}^{-1} \vec{\beta}_{(2)} = 108296$$

And compare it to:

$$(r-q) \cdot \hat{\sigma}^2 \cdot F_{r-q,n-r-1}(\alpha) = 133445$$

Since $108296 < 133445$, we don't have sufficient evidence to assure that $\beta_1 = \beta_2 = 0$.

(8) 95% confidence interval for the mean response given $\mathbb{E}(Y_0) = \beta_0 + \sum_{i=1}^{5} \beta_i \cdot \bar{z}_i$, where $\bar{z}_i$ is the sample mean of $z_{i,j}$ for $i \in \{1, ..., n\}$

First, compute $\vec{z_0}$:

$$\vec{z_0} = \begin{pmatrix} 1.000 \\ 1.031 \\ 3.034 \\ 62.840 \\ 19.500 \\ 1.999 \end{pmatrix}$$

And now compute the confidence intervals for it's associated value using the formula in the class notes:

$$\vec{z_0}^T \vec{\beta} \in [\vec{z_0}^T \hat{\vec{\beta}} \pm \hat{\sigma} \cdot t_{n-r-1}(\frac{\alpha}{2}) \sqrt{\vec{z_0}^T (Z^T Z)^{-1} \vec{z_0}}]$$

Obtaining the following interval: $\vec{z_0}^T \vec{\beta} \in [71.78, 152.8]$

(9) 95% confidence interval for a new response Y_0 given $\vec{z_0}$

Using a similiar formula:

$$\vec{z_0}^T\vec{\beta} \in [\vec{z_0}^T\hat{\vec{\beta}} \pm \hat{\sigma}\cdot t_{n-r-1}(\frac{\alpha}{2})\sqrt{1+\vec{z_0}^T(Z^TZ)^{-1}\vec{z_0}}]$$

And using that $Y_0 = \vec{z_0}^T\vec{\beta} + \epsilon_0$ we obtain:

$$Y_0 \in [-73.38, 298]$$

A substantialy bigger interval than the over in **(8)**, which makes sense since we are including the error now.

2 Two-Sample test and LDA

2.1 Introduction

Briefly summarize the goal of the analysis in your own words

2.2 Summary

Summarize your data by plots or sample estimates

2.3 Analysis

Implement the analysis based on what you have done in homework

3 PCA

3.1 Introduction

Briefly summarize the goal of the analysis in your own words

3.2 Summary

Summarize your data by plots or sample estimates

3.3 Analysis

Implement the analysis based on what you have done in homework