DSC 423: Data Analysis and Regression Assignment 09

Saransh Thakur

Id: 2020070

Problem 1

2. When multicollinearity exists, least squares estimates are unbiased and their variances are large, therefore they are far from the true value.

Ridge regression reduces standard errors by introducing some bias into the regression estimates. Ridge regression estimates are usually stable, which means they are not impacted by slight changes in the data on which the fitted regression is based. Ridge regression models produce more exact predictions for the new observations; thus, it is employed to tackle multicollinearity.

Graphical user interface, chart, histogram

Description automatically generated

Text

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

1. **Residual** is the difference between an observed value of the response variable and the value of the response variable predicted from the regression line.

Graphical user interface, application, table, Excel

Description automatically generated

Graphical user interface, application, table, Excel

Description automatically generated

Text

Description automatically generated

The sum of the residuals should be Zero.

Chart, histogram

Description automatically generated

The histogram for the residuals plot is normally distributed.

Graphical user interface, chart, scatter chart

Description automatically generated

The residual plot is heteroscedasticity, which implies it has a different variance and the plot mimics a football shape. Means variance on the end differs from the middle, and 95 percent of values are between two standard deviations, for any points above or below two standard deviations considered outliers. In this case, we have outliers.

Chart, line chart

Description automatically generated

The distribution is normal because most of the observations fall on a straight line and the plot consists of outliers and influential points.

1. Lasso regression is a type of continuous feature selection that is used to pick the features.

When we run Lasso, we strive to reduce the variance, and because of the Lasso equation, many of the Beta values will become "Zero," therefore the values associated with betas can be deleted from the model. As a result, Lasso can be thought of as a continuous kind of feature selection.

Graphical user interface

Description automatically generated with low confidence

Text

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

1. The ridge regression and Lasso models are not the same since many betas become zero in the Lasso model, letting the value associated with the betas be omitted from the model.

Looking at the coefficient of Lasso, we can see that students in English, schoolHasLibrary, fatherBornUS, Selborne, and preschool were removed by Lasso.

If our primary aim is prediction and accumulating information about all the features, we may not need to use feature selection at all and instead rely on ridge regression to keep track of all the predictors in the model. If we need to limit the number of predictors for practical reasons, LASSO is a good option.

Problem 2

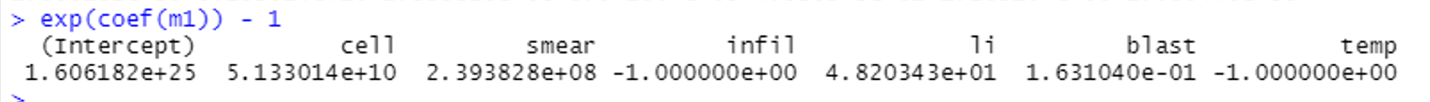
Table

Description automatically generated

The p-value for the model does not fall underneath the significant value of 0.05, as shown by the above model. As a result, we need to prune our model to ensure that our value is less than 0.05.

A screenshot of a computer

Description automatically generated with low confidence



To get the final model I remove the variable with the worst p-value, so after pruning the final model is

A screenshot of a computer

Description automatically generated with low confidence

Text

Description automatically generated

1. Glm is used to fit generalized linear models, which are defined by providing a symbolic description of the linear predictor as well as a description of the error distribution.

**Glm** — Used to fit generalized linear models.

Syntax: -

**glm (formula, family=gaussian, data...)**

where:

formula: The linear model's formula (e.g., y x1 + x2).

family: The statistical family to be used to fit the model. The default is gaussian, but other alternatives include binomial, Gamma, and Poisson.

data: The name of the data frame that includes the data.

Lm is a function that is used to fit linear models. It can do regression, single stratum analysis of variance, and analysis of covariance.

**Lm** — Used to fit linear models.

Syntax: -

**lm (formula, data)**

where:

formula: The linear model data formula (e.g., y x1 + x2)

Data: The name of the data frame that contains the data.

The only difference between these two functions is the **family argument** in the glm () function.

We will get the same results if we use lm () or glm () to fit a linear regression model.

The glm () function can also be used to fit more complex models, such as:

Logistic regression (family=binomial)

Poisson regression (family=Poisson)

1. The initial full model is

M1 <- glm (formula = remiss ~ cell, smear, infil, li, blast, temp, family = "binomial", data = remission)

The final model after pruning is

M2 <- glm (formula = remiss ~ li, family = "binomial", data = d)

A picture containing text

Description automatically generated

For the final model after pruning, we are only left with one independent variable that is li.

We can de-log the coefficient using exp(coef(m2)) – 1.

For every unit change in li, the probability of remiss changes by 17.1244863