Project Summary

Pratheek Eravelli & Kanav Bhatnagar

3/28/2022

Summary

1. Abstract

This case study is focused on examining the correlation between level of PSA and a number of clinical measures for 97 men with advanced prostate cancer who were about to undergo a radical prostatectomy. The main objective here was to understand the data while using the statistical techniques and tools to construct predictive models to describe the association between the level of PSA and the clinical measures.

We begin with an initial linear model which uses 6 different clinical measures, and we proceed by evaluating each variable and removing those that are not a good statistical fit for the model, effectively performing model selection manually. Once we have this chosen model, we analyze any unusual observations, as well as check if all of the regular linear regression model assumptions are satisfied. Finally, we check to see if any remedial measures using variable transformations are necessary, and implement them if need be.

2. Introduction

The dataset has 97 observations and 9 variables (one index, one response, and 25 predictor variables). Our response variable is the level of prostate-specific antigen (`psalevel`). The initial model has 6 predictor variables, namely

- X_1: cancer volume (`cancervolume`)
- X_2: patient age (`age`)
- X_3: amount of benign prostatic hyperplasia (`hyperplasia`)
- X_4: seminal vesicle invasion (`svi`)
- X_5: capsular penetration (`capsular`)
- X_6: Gleason score (`gleason`)

3. Results

Statistical results from our analysis can be broken down into the following categories:

3.1. Model Selection

Using the p-values for individual t-tests for the predictor variables in the original model, we were able to determine which variables were significant in their contribution to explaining the variation in the response, and which were not. We proceeded by removing the predictor variable with the highest p-value until we had a final model with all significant predictor variables. The models selected were (in order)

- `psalevel` $=-14.7460$ $+2.0375$ `cancervolume`-0.5327 `age`$+1.3518$ `hyperplasia`$+19.6441$ `svi`$+1.0974$ `capsular`$+6.9942$ `gleason`

In the linear model given above, it was noticed that capsular penetration, i.e. `capsular` was the least significant variable with a slope p-value of 0.4103. We fit a linear model removing this variable and confirmed using an ANOVA test that the smaller model was a better fit.

- `psalevel` $=-18.4353+2.2595$ `cancervolume`-0.5261 `age`$+1.3714$ `hyperplasia`$+23.6477$ `svi`$+7.4688$ `gleason`

In the linear model given above, it was noticed that patient age, i.e. `age` was the least significant variable with a slope p-value of 0.2674. We fit a linear model removing this variable and confirmed using an ANOVA test that the smaller model was a better fit.

- `psalevel` $=-43.4846+2.2995$ `cancervolume`$+0.9001$ `hyperplasia`$+22.5019$ `svi`$+6.3942$ `gleason`

In the linear model given above, it was noticed that the amount of benign prostatic hyperplasia, i.e. `hyperplasia` was the least significant variable with a slope p-value of 0.3987. We fit a linear model removing this variable and confirmed using an ANOVA test that the smaller model was a better fit.

- `psalevel` $=-44.1849+2.2496$ `cancervolume`$+21.8808$ `svi`$+6.8982$ `gleason`

In the linear model given above, it was noticed that the gleason score, i.e. `gleason` was the least significant variable with a slope p-value of 0.1693. We fit a linear model removing this variable and confirmed using an ANOVA test that the smaller model was a better fit.

- `psalevel` $=1.060+2.477$ `cancervolume`$+24.647$ `svi`

Both of the predictors in the given model, i.e. `psalevel` $=1.060+2.477$ `cancervolume`$+24.647$ `svi` are statistically significant for the regression. Thus, this was chosen as the final model

All future analysis is performed on the chosen model.

3.2. Unusual Observations

High Leverage Points

Using the hat matrix for the chosen model, we found that there were 12 data points with high leverages. The points are: $55, 62, 64, 71, 73, 74, 86, 88, 90, 91, 94, 97$.

Once we had these data points, we used the interquartile range of the data to find which of the high leverage points can be considered “bad” high leverage points and which can be considered “good”. None of the high leverage points were outside the upper and lower limits imposed using the interquartistudentized Breusch-Pagan testle ranges, and thus we had 0 “bad” high leverage points.

Outliers

In order to find outliers, we found the studentized residuals for the model. Then, we computed the bonferroni critical value using the number of observations and the number of parameters. Finally, we checked to see if any of the residuals were greater than the absolute values of the bonferroni critical value, which would mean that those residuals corresponded to data points which are outliers.

Using this method, we found two outliers, points 96, and 97.

Influential Points

Using Cook's distance as a measure, we searched for influential points in the model. Point 97 was the only point with a Cook's distance greater than 1, i.e. the only highly influential point.

Point 97 is the only point which is a high leverage point, an outlier, and a highly influential point

3.3. Model Assumptions

Homoscedasticity

The plot given in the appendix for the chosen model's residuals showed that the values of the residuals did not appear to be constant along the 0 line. Moreover, the studentized Breusch-Pagan test resulted in a p-value of 0.0000196, which is much smaller than the threshold of 0.05. Thus, we rejected the null hypothesis and concluded that the assumption of constant variance for the data is not met.

Normality

Using the Q-Q plot given in the appendix, we can see that the points did not seem to be distributed normally. Additionally, a one-sample Kolmogorov-Smirnov test resulted in a p-value of $3.542e-14$, which is much smaller than the threshold of 0.05. Thus, we rejected the null hypothesis and concluded that the the distribution of sample means (across independent samples) is not normal.

3.4. Remedial Measures

Since both the normality and the constant variance assumptions were not met, we wanted to use data transformation as a remedial measure for the model. Using the boxcox plot in the appendix, we can see that the optimal lambda value for a data transformation is near 0. Thus, a log transform for the data should be sufficient.

We performed a log transform, and performed tests to check if the new model satisfied the model assumptions. The resultant linear model meets both the constant variance and normality assumptions, with p-values greater than the threshold of 0.05 for both the studentized Breusch-Pagan test and a one-sample Kolmogorov-Smirnov test. The relevant graphs for the new model are also available in the appendix.

4. Conclusion

Appendix