PSA Screening for Prostate Cancer Diagnosis and Treatment

Shibi He

March 12, 2020

Research question

In this project, I analyzed the data at hand to examine how useful is PSA screening for the diagnosis and treatment of prostate cancer. I started with the transformation of the variables, and then used model selection method to choose an appropriate model. Next, I checked the assumptions of the model, such as linearity, homosckedasticity, and normality. I also checked for the problem of multicolinearity. Finally, I performed a series of regression diagnostics, such as the detection for outliers and inflential observations. My major findings are as follows.

1.a. Transformation

I consider "PSA" as the response and all other variables as the potential predictors. The scatterplot matrix (Appendix, page 1) of these variables shows that many data points concentrate at the lower left part of the graphs and the relationships between PSA and many other variables do not seem to be linear, indicating transformation may be needed. Applying Box-Cox method, I take log transformation for the variable "cancer_volume", "prostage_weight", "BPH","pct_G", square transformation for "age", and -0.5 power transformation for "capsular_penetration". I also use log transformation for the response "PSA". The result of Likelihood Ratio test suggests that these transformations seem to be appropriate. The scatterplot matrix after transformation (Appendix, page 4) also shows some improvement. Therefore, the candidate model 1 is as follows:

$$\begin{aligned} log(PSA) \sim SVI + log(cancer\_volume) + I(capsular\_penetration^{-0.5}) \\ + log(prostate\_weight) + log(BPH) + GS + log(pct\_G) + I(age^2). \end{aligned}$$

1.b. Model selection

Both backward elimination and bidirectional stepwise method suggest to drop the regressor "capsular_penetration" and give the lowest AIC of -101.66 (Appendix, page 4-6). Therefore, I consider the following model 2:

$$\begin{aligned} log(PSA) \sim & SVI + log(cancer_{v}olume) + log(prostate_{w}eight) \\ & + log(BPH) + GS + log(pct_{G}) + I(age^{2}). \end{aligned}$$

Since the purpose of this project is to examine whether PSA screening is useful for diagnosing prostate cancer, I would keep the variables that are closedly related to prostate cancer in the model, such as "cancer_volume", "GS" and "pct_G".

1.c. Check assumptions of the model

I first checked the residual plots of model 2. The residual plots (except for "prostate_weight" and age) look like null plots and do not have any visual evidence for curvature. Moreover, the Tukey test gives a p-value of 0.2, providing no evidence against the mean function (Appendix, page 6). Next, I use the test for non-constant variance to further check for homoskedasticity. The test gives a p-value of 0.92, suggesting the model satisfies the constant variance assumption (Appendix, page7). The QQ plot is very close to a straight line, suggesting there is no violation to the normality assumption (Appendix, page 8). At last, I check whether the model suffers from multicolinearity problem. I found all the VIF are samll (<5), suggesting there is no multicolinearity problem in our model (Appendix, page 9). It seems like model 2 satisfies all the assumptions, and therefore no correction is needed.

According to the summary output of model2/my.model (Appendix, page 5), the F-test gives an extremely small p-value, suggesting the overall model is significant. Moreover, the Adjusted R-squared is 0.6661, suggesting 66.6% of the variation in PSA can be explained by this model. Therefore, I would say this model did a fairly good job explaining the variation in PSA. Moreover, the summary output shows that log(cancer_volume), log(prostate_weight), and log(pct_G) all have significantly positive effects on the response log(PSA). The effect of SVI is also positive and statistically significant, suggesting that the expected log(PSA) is higher for the people who have Seminal Vesicle Invasion than those who do not have this invasion.

2. Answer the research question

The results I get so far help answer the research question. The variables that are closely related to cancer, such as cancer_volume, prostate_weight, and pct_G all have significantly positive effects on the PSA level. The average PSA level is also significantly higher for people who have Seminal Vesicle Invasion than those who do not have this invasion. These results suggest that people who have more severe prostate cancer tend to have higher level of PSA. On the other hand, BPH, i.e. the benign prostatic hyperlasia amount, does not have significant effect on PSA level, further confirming that only the malignant organs are associated with the increased PSA level. Therefore, the PSA screening is very helpful for diagnosing the prostate cancer, especially for the more severe ones. To reduce mortality de to prostate cancer, I would suggest the patients to take the PSA screening.

3. Influence analysis

The outlier test suggests no Studentized residuals with Bonferroni $p < 0.05$, so there is no outlier in the dataset (Appendix, page 9). The Cook's distances show that observation 52 and 108 have the largest values, so they are the relatively more influential observations (Appendix, page10). Next, I remove these two influential observations and fit the model2 again to see if they have any impact on the estimated coefficients.

After removing the two influential observations, I find the effect of "BPH" becomes significant, and the effect of "prostate_weight" becomes less significant. The results imply that even the benigh prostatic hyperplasia could also be associated with higher level of PSA. Therefore, when we interpret the results of model2, we should be very cautious as the results we see can be largely attributed to the two influential observations. In conclusion, I still suggest the patients who potentially suffer from the prostate cancer to take the PSA screening as PSA is a reliable indicator of prostate cancer. However, the PSA screening should not be the only one test or measure for diagnosis as the benigh prostatic hyperplasia could also possibly increase the PSA level.