Model Selection and Validation for Predicting PSA level of Prostate Cancer

Abstract

For this case study, we attempt to find the suitable model for investigating the relationship between PSA levels and several clinical measures using model selection and validation methods. The data set we used here is from Kutner et al.(2004) with the sample of 97 men who had prostate cancer. We conclude that cancer volume, prostatic hyperplasia, capsular penetration and Gleason score can affect PSA level significantly among all clinical measures. We also suggest that medical researchers could concentrate on those relevant measures to prevent the prostate cancer.

Background and Significance

Prostate cancer is one of the most common form of cancer affecting men for the last twenty years, especially given that there are countless possibilities and elements that can lead to it. (Kim, S., Jang, K., Park, W., Kwon, D., Kang, W., Lim, H., & Moon, J., 2014). Due to advances in treatment uncovered through previous researches, the rate of mortality has dropped significantly. Moreover, the PSA test is one of the most crucial test to diagnose the cancer. (Beebe-Dimmer,J.L., Faerber, G. J., Morgenstern, H., Werny, D., Wojno, K., Halstead-Nussloch, B., & Cooney, K. A.,2008). In this case study, we use the data provided by Michael. H Kutner (2004) on 97 men with severe prostate cancer. The data include the PSA level and several measurements. Thus, this case study attempts to find out the correlation between level of PSA and the risk factors, which can help us advance in early detection and future clinical treatment.

Exploratory Data Analysis

In this case study of PSA level in men, we want to produce a correlation matrix for the concerned predictor variables. Taking note that X7 (Gleason Score has 3 levels in the given data set) is a qualitative variable, it was transformed into a categorical variable using two indicator variables (X8 and X9).

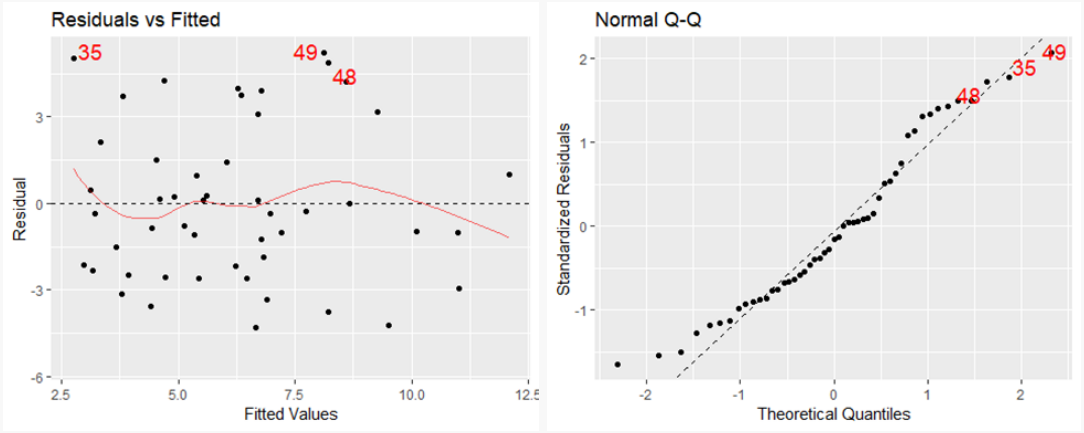
	X1	X2	X3	X4	X5	X6	X8	X9
X1	1.00	0.01	0.04	-0.13	0.58	0.69	-0.26	-0.22
X2	0.01	1.00	0.16	0.32	0.00	0.00	0.07	-0.10
X3	0.04	0.16	1.00	0.37	0.12	0.10	-0.22	0.09
X4	-0.13	0.32	0.37	1.00	-0.12	-0.08	-0.09	0.14
X5	0.58	0.00	0.12	-0.12	1.00	0.68	-0.32	-0.02
X6	0.69	0.00	0.10	-0.08	0.68	1.00	-0.31	-0.08
X8	-0.26	0.07	-0.22	-0.09	-0.32	-0.31	1.00	-0.64
X9	-0.22	-0.10	0.09	0.14	-0.02	-0.08	-0.64	1.00

From the correlation matrix above, we see that the pair of variables with the highest strongly positive association is between cancer volume (X1) and capsular penetration (X6) (r = 0.69). This indicates that men with a high estimate of prostate cancer volume are also likely to have high degree of capsular penetration.

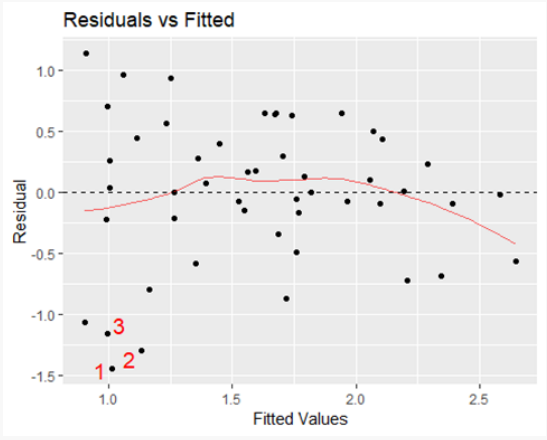
The table also shows that there is moderately positive association with two other pairs of variables - one pair being cancer volume (X1) and seminal vesicle invasion (X5), and the other being between seminal vesicle invasion (X5) and capsular penetration (X6). Similarly, we also see moderately negative association between the two indicator variables of Gleason score (X8 and X9).

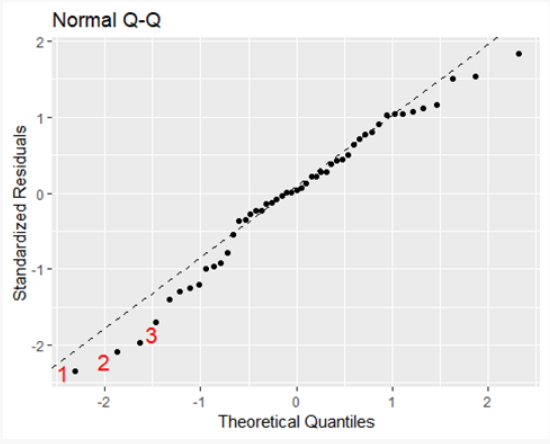
It is important to note that there is absolutely no association between prostate weight (X2) and seminal vesicle invasion (X5), also none between prostate weight and capsular penetration. Prostate weight is also seen to have very weak association with both cancer volume and the Gleason score indicators. This shows that prostate size has relatively little to no effect on most of the other variables.

Model



The two plots provided above are stem from the initial linear model Y~ X1 + X2 + X3 + X4 + X5 + X6 + X8 + X9, where we find some non-constant error and some curvature shown by the residual plot (by "mplot(model, which = 1)"), so it indicates of linearity. Also, from Normal Q-Q plot (by "mplot()"), we see some departures from normality especially the points in the middle, which violates the normality assumption.





Thus, we apply the variable transformation to make the errors more normally distributed and reduce the curvature. When we fit the model with lnY, we obtain a random scatter plot and normally distributed error in the normal Q-Q plot shown on the graphs above.

Model Selection:

For more precise results, we use the first 49 out of 97 of our observations as our model building data set and the latter half for our validation data set. It is worth noting that we regard X7 (the Gleason score) as categorical variable which has 3 levels, so we use Rcode to introduce 2 indicator variables via mutate(data set, categorical variable= as.numeric()) to replace it. Once this is done, we proceed with the actual model selection. stepAIC(selected model, direction = "both"), which is a function from the MASS package (from 'dplyr') and performs stepwise selection, is used to determine that lnY ~ X1 + X4 + X6 + X8 + X9 is our final model. As an additional step, we also used the function select_criteria(lm(model), n, s),) on all 32 possible combinations of the 5 variables to compare the four criteria to find the most suitable model.

Here we state our deciding factors for each criterion:

R^2: This criterion is met when adding more explanatory variables leads to a very minimal (or no) increase in R^2. We may take into consideration our Model 31 (lnY~X4+X6+X8+X9) because when adding another explanatory variable (X1), R^2 increased at most by 0.1. Maybe even Model 32(lnY~X1+X4+X6+X8+X9) is preferred because adding another explanatory variable (X2) increased the R^2 by at most 0.12.

Adjusted coefficient of determination: For this criteria, the simplest model with R^2 (adj) closest to the upper limit is chosen. The upper limit here is 0.53 which is attained with model 29(lnY~X1+X4+X8+X9). But since the R^2 (adj) of model 32 (lnY~X1+X4+X6+X8+X9) is 0.52, which is also close to the upper limit, we need to take model 32 into consideration as well.

Mallows' C_p Criterion: For this test, all models with small C_p and with C_p close to p' are considered for further study. Model 29 and 32 are preferred because they have a small C_p and a C_p close to p'.

Akaike's Information Criterion: This criterion requires the selection of the model with the smallest AIC, which in this case is the model 29, having -39.83 for AIC.

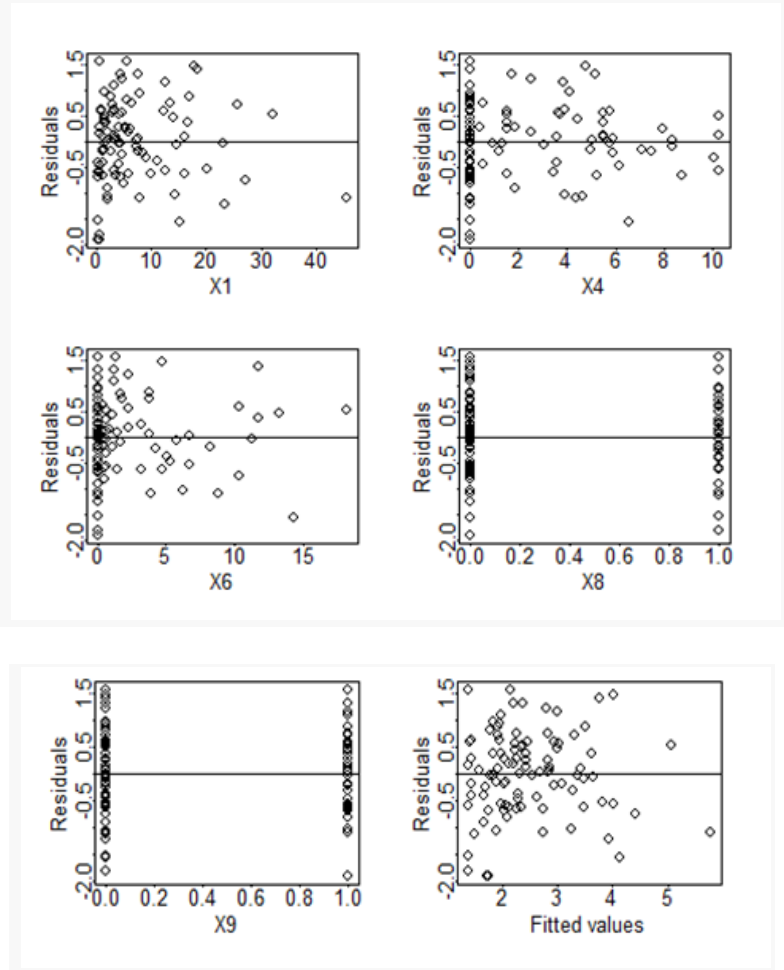
Based on the 4 criteria above and the stepAIC output, we can conclude that lnY~X1+X4+X6+X8+X9 is the suitable model.

Model validation:

As mentioned, we use the second half data set to validate the selected model. To do this we calculate the MSPE, which is the sum((Y_obs-Y_pred)^2/n_star) (around 1.67), and we calculate the MSRes (around 0.41). Because the two values are close based on the regression fit, it is an appropriate indication of the predictive ability of the model.

Model diagnostics:

To check whether the functional form of the model is adequate, we install the package "car" in order to call plot(), a function that we used to derive the plots of residual against the selected 5 explanatory variables separately. All six plots below show the random patterns which indicate a proper functional form.

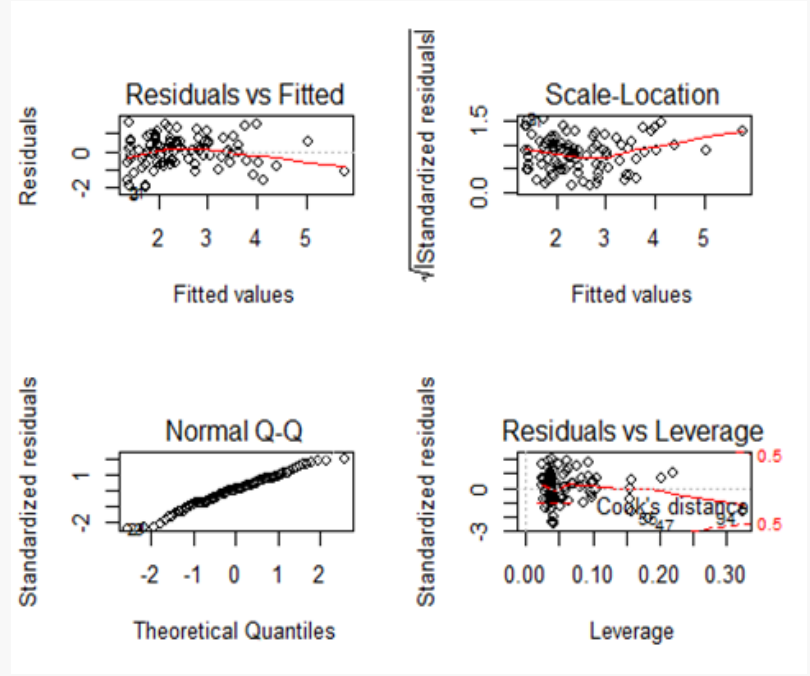


Then, we test whether there are outlying X observations and outlying Y observations. We

know observation i is outlying X observation if $P_{ii} > \frac{2p'}{n}$ or $P_{ii} > 0.5$ where P_{ii} is leverage value(by calling hatvalues()). By these rules, we conclude observations 47, 55, 76, 78, 79, 91, 94 and 97 are potentially outlying X observations. For outlying Y observations, the critical value is t1-α/(2n),n-p'-1 = 3.601304. Since the $|t_i|$ values are all smaller than the critical value, there are no outlying Y observations. This is proved by the R command which(abs(t) > t_crit), which gives us the outcome 'named integer(0)'.

Notice that there are 48 observations in the validation data set which can be considered as a medium data set. No observation is influential because their corresponding DFFITS are not larger than 1. Moreover, by using R to find VIF and the mean of the VIF, we find that the largest VIF (roughly 2.47) is not larger than 10 (i.e. none of the VIFs is larger than 10), and the mean VIF (around 2.03) is not considerably larger than 1. Therefore, there is no indication of multicollinearity.

Diagnostic Plots



By calling layout(matrix()), we get four diagnostic plots. Based on these plots, we can test whether the multiple linear regression model with normal errors is appropriate or not. We can conclude that there is a linear relationship since there is no pattern in the residual plot. Also since the errors in the residual vs. fitted values graph seem to be centered around 0 and their common variances in the plot show no clear increase/ decrease, we can summarize that the mean of errors is 0. What's more, errors are seen to be pairwise independent because the patients with prostate cancer whose data are collected for this study are randomly selected. When we look at the Normal Q-Q plot, errors are normally distributed since points are close to the line. For Residuals versus Leverage plot, we do not have any influential observation outside the bonds, so it has low Cook's distance. For Scale-Location plot, it appears

to be a straight line with randomly spread points, so there is no indication of heteroscedasticity (i.e. it is homoscedasticity). All these factors thus conclude that our model is adequate.

Discussion/Conclusion:

The goal is to detect the correlation of PSA level and seven possible clinical measures based on the prostate cancer data set. In terms of the above model selection and validation, we find cancer volume, prostate hyperplasia, seminal vesicle invasion and Gleason score have closer relation with PSA level compared with the rest possible measures. Moreover, our finding may exert influence on related medical fields. For instance, researchers could do further investigation about the subclasses of the above four clinical measures. However, there also exist some limitations of our case study. To illustrate, the sample size of the data set is not large and the data only concentrate on the elder men whereas we need data from young men to make the comparison as well.

References

[1] Kim, S., Jang, K., Park, W., Kwon, D., Kang, W., Lim, H., & Moon, J. (2014). Serum prostate-specific antigen levels and type of work in tire manufacturing workers. *Annals of Occupational and Environmental Medicine, 26*(1). doi:10.1186/s40557-014-0050-z

[2] Beebe-Dimmer,J.L., Faerber, G. J., Morgenstern, H., Werny, D., Wojno, K., Halstead-Nussloch, B., & Cooney, K. A. (2008). Body Composition and Serum Prostate-Specific Antigen: Review and Findings from Flint Men's Health Study. Urology, 71(4), 554-560, doi: [10.1016/j.urology.2007.11.049]