Kidney Transplant success model

Alexander James Ryan

6 May 2018

Contents

Analysis Renal success

The goal of this analysis is to model the probability of graft failure in kidney transplants, using a dataset of 1158 patients. The predictor variables include the binary variables, sex of the patient (male), presence of cardio-vascular problems (cardio). The continuous predictor variables include the age of the patient (age), haematocrit level before the transplant (HC).

Descriptive Statistics

Reject is the outcome binary variable and indicates whether there are symptoms of graft failure of the kidney transplant. The mean age of patients in the dataset is 46.43, with the youngest person being 15 and the oldest 76. The mean haematocrit level is 31.86, with the smallest being 14 and the largest 60.

A total 366 people had symptoms of graft failure and 207 people showed signs of cardio-vascular problems. There were 59 people with both cardio-vascular problems and symptoms of graft rejection.

Of the 664 men, 196 had symptoms of graft rejection (prop=0.295), 123 showed signs of cardio-vascular problems (prop=0.185) Of the 494 women, 170 had symptoms of graft rejection (prop=0.344), 84 showed signs of cardio-vascular problems (prop=0.170). The sum function was used.

Model Selection

Link Function Investigation

Link functions logit, probit, cloglog and cauchy were all considered. All functions produce roughly the same logistic response graph, with a similar AIC score (difference is less than 5). The link function that produces the best fit statistic (Using the Hosmer-Lemeshow statistic) is the probit link function. Although the use of

logit is preferred because it is easier to interpret than probit, a one unit change in X1 is associated with a B1 change in the log odds of 'success', ceteris paribus. Thus the logit link function will be used.

Variable Selection

The R function used to model the logistic is the glm function. At the 5% level of significance, the only covariates that are significant are the age of the patient at the time of the transplant (male).

Despite the haemocritic level variable (HC) not being statistically significant in the model, it will still be included. The medical understanding of rejection of kidney transplants, is the stronger the immune system, the higher the probability of rejection of the graft by the body. The haemocritic level is a measurement of the proportion of red blood cells over total blood in the body. Low HC levels may mean that there are higher number of white blood cells, and a sign of other serious disorders, diseases or cancers. The haemocritic level is thus used like a latent variable for the strength of the immune system and the overall health of the patient.

Cardio is not statistically significant. This variable will not be included in the model. Elevated HCT levels may be positively associated with cardiovascular risk factors (Jin et al. 2015), and therefore by including both HC and cardio in the model, we run the risk of multicollinearity. The continuous variable HC should contain the information found in cardio.

According studies done by Lau et al. (2018) on kidney transplants, they found that female organs were more often rejected than male ones, a trend higher among male patients. This highlights an important missing variable, the sex of the donor who provided the kidney. Without this information, it seems like we are missing a critical factor in the success probability of the kidney graft. Thus, despite not being statistically significant, the variable male, will be included in the final model.

Interactions

All possible interaction terms are considered between the variables. When a full logistic model is generated, none of the coefficients are significant. When a StepAIC function is used, all the interaction terms are removed. The only terms left are age and male.

Polynomial terms

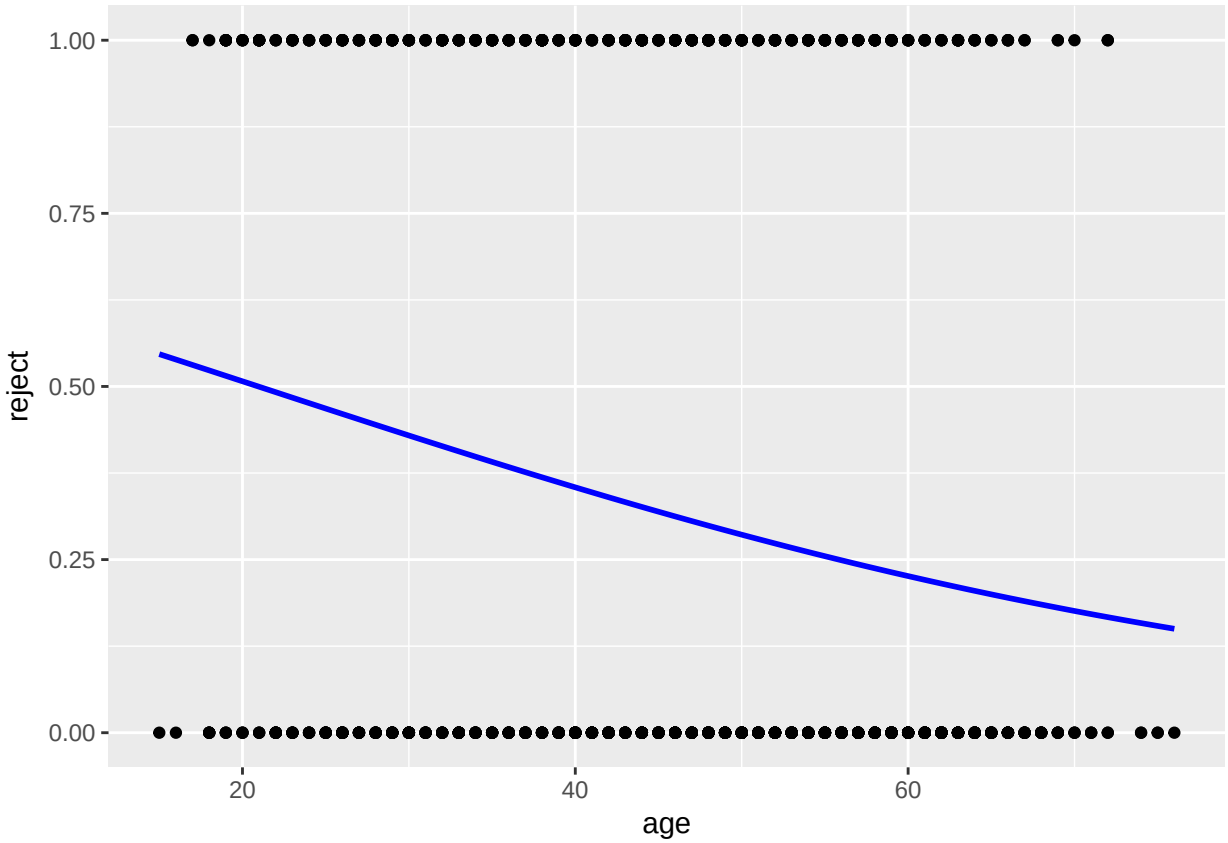
Polynomial terms for binary variables does not make sense, so we will not consider polynomial terms for male and cardio. When including quadratic and cubic terms for age and HC in the model, the coefficients are not significant. When a stepAIC function is used, those polynomial terms are subsequently removed. The only terms left are age and male.

Final Logistic model

$$\pi(reject) = \frac{exp(\beta_0 + \beta_1 age + \beta_2 male + \beta_3 HC)}{1 + exp(\beta_0 + \beta_1 age + \beta_2 male + \beta_3 HC)}$$

The intercept is (0.72). The coefficient for age (-0.03), implies that age has a negative influence on the probability of graft failure. That is, the older the patient, the higher the probability of success of the transplantation. One major factor on the success of a kidney transplant is the strength of the immune system. A younger person on average will have a stronger immune system than an older person, and so it makes sense that a younger patient's body will have a higher probability of rejecting the new kidney. Although there are limitations to this idea. At a certain higher age, the probability of rejection should increase, as the problems associated with post surgery are more difficult for much older people.

Plot of probility of rejection, by age.



The coefficient for male is (-0.22), which implies that being male means that on average you have a lower probability of graft rejection. The coefficient for HC is (0.0025) means that haemocritic level has a much smaller impact on the probability of rejection than the age of the patient. Importantly, the higher the haematocrit level, the higher the probability of graft rejection.

Goodness of Fit

The Pearson statistic and the Deviance statistic will not be used as Goodness of Fit indicators, because they should only be used when there are only categorical regressors in the model.

The Hosmer-Lemeshow statistic is used to test the Goodness of fit of the model. The code used is from Generalised Linear Models slides. Hosmer and Lemeshow recommend using g > p + 1. Using g=4, the code generates a p value of (0.64), and then g=10 generates p value of (0.57), which means there is no evidence of a poor fit.

Predictive Quality

When applied to a logistic model, even with a perfect fit, R^2 can never reach 1. Since the traditional method verifying predictive quality, the R^2 metric, can no longer be used effectively, we must use alternatives.

Nagelkerke's R^2

Nagelkerke's R^2 is a metric designed to imitate the R^2 coefficient of determination. The result we found was (~0.05), implying that the model might have poor predictive quality. This might also be an indication of

missing variables in the dataset. One example of a missing variable is the sex of the patient who provided the kidney that has been transplanted.

Concordance measure

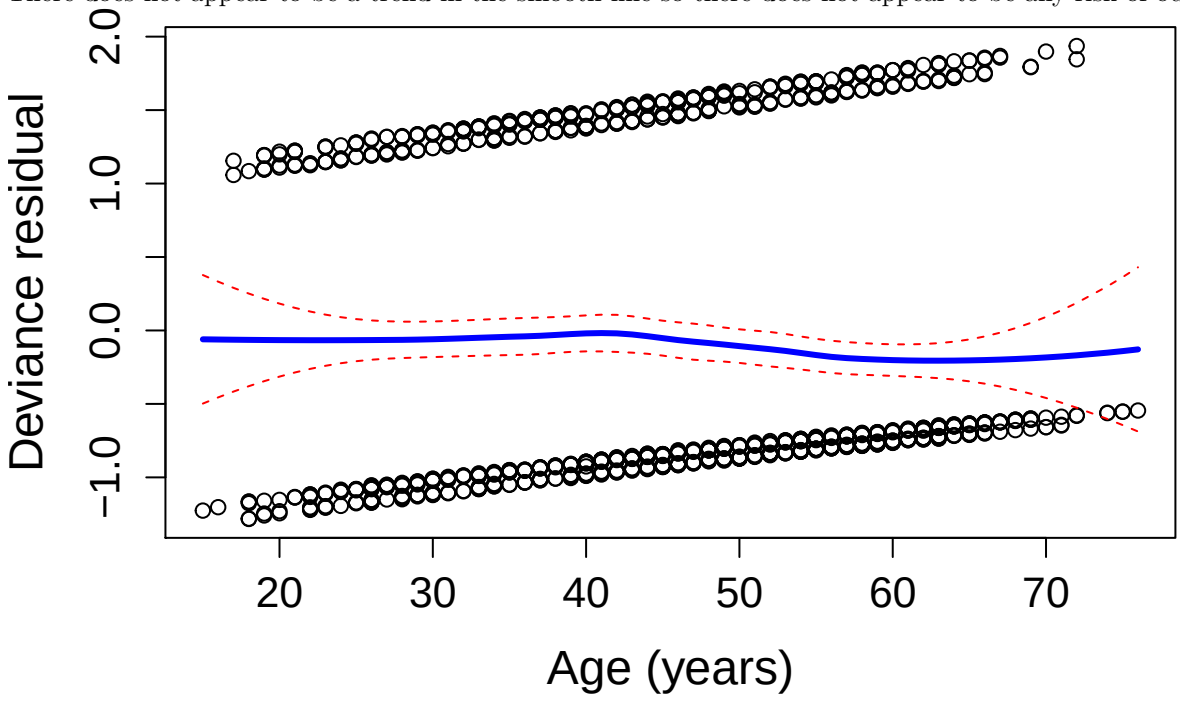
Concordance is a measurement of the association between actual variables and fitted values in the model, in percentage terms. A concordance range between (60-70%) is considered a well fitted model, while a range between (85-95%) suggests that there might be overfit with the model.

The measured Concordance for the model is 62.08%. This is in the range that implies that we have a well-fitting model.

Residual Diagnostics

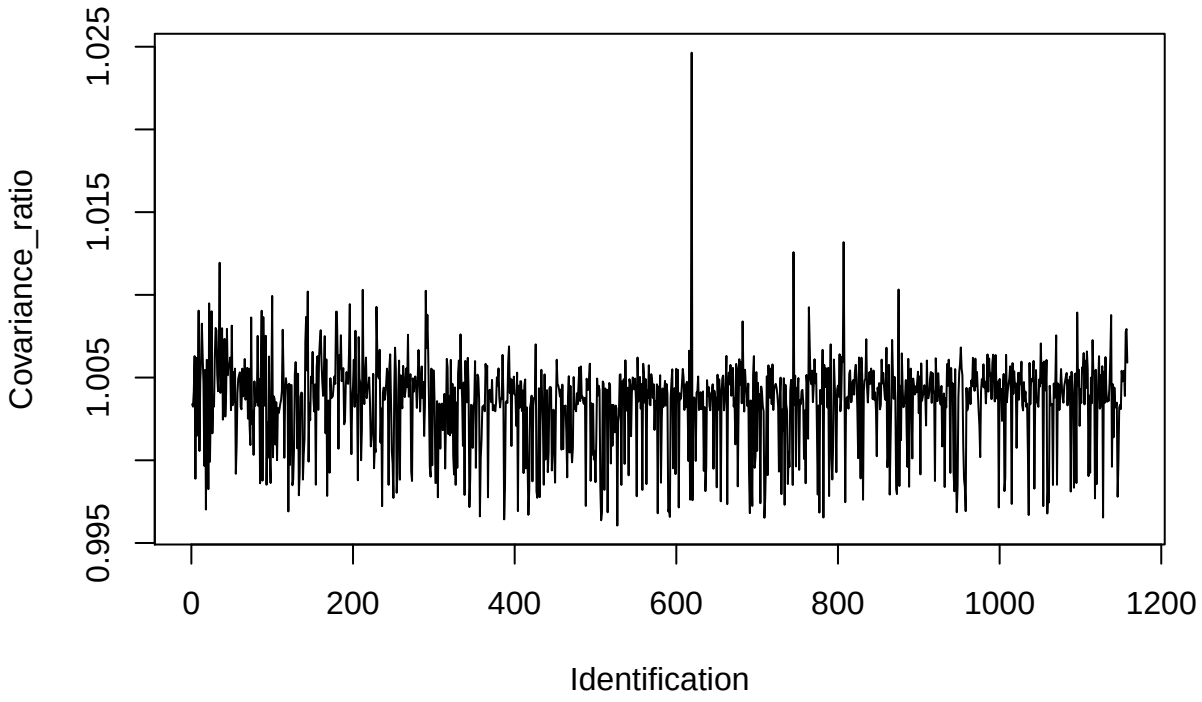
Outliers

There does not appear to be a trend in the smooth line so there does not appear to be any risk of outliers.



Influential Observations

A plot of the Studentized residuals reveals one value. A plot of the Dffits reveals nothing. A plot of the Covariance ratio reveals that there may be one influential observation, patient number roughly 615. This shows individuals whose removal causes the greatest change in the covariance matrix of regression coefficients.



Baysian Approach

The coefficient for age (-0.03), for male (-0.22), for HC (0.002) all appear to be roughly the same as the model generated in the frequentist approach.

Analysis Summary

As indicated by the poor predictive quality of the final model (Nagelkerke's R^2=~0.05), it is recommended to include more information in the model. The first variable that should be included is the sex of the patient who provided the transplanted kidney. Without this variable, the sex of the patient is only half the story. Core factors that influence graft rejection are the age of the patient: The older the patient, the higher the probability of success of the transplant.