**Model Building and Evaluation**

**John Lance Abbott**

**Data Structure**

The data is presented as information on past loan applicants. Most of the data is categorical, but some of it is numeric. The logistic regression can handle both, but some consistency may be desired. Variables that are identifiers like OBS and TELEPHONE are excluded from the analysis except as identifiers.

**Data Quality**







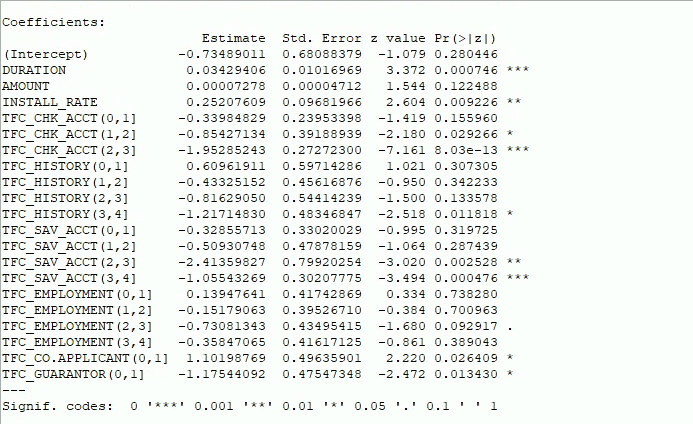
First I checked the some correlations between variables that might contaminate one another (shown above). Logistic regression assumes all variables are independent, and based on the correlation coefficients above, none of the correlations I was concerned with are particularly strong.

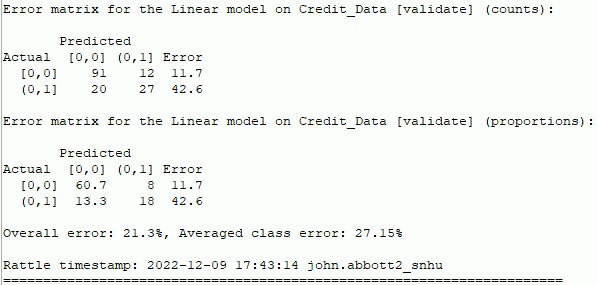
Next I had Rattle recode the variables that are meant to be categoric, which it had tried to identify as numeric, and then partitioned it as 70/15/15 training/testing/validation.

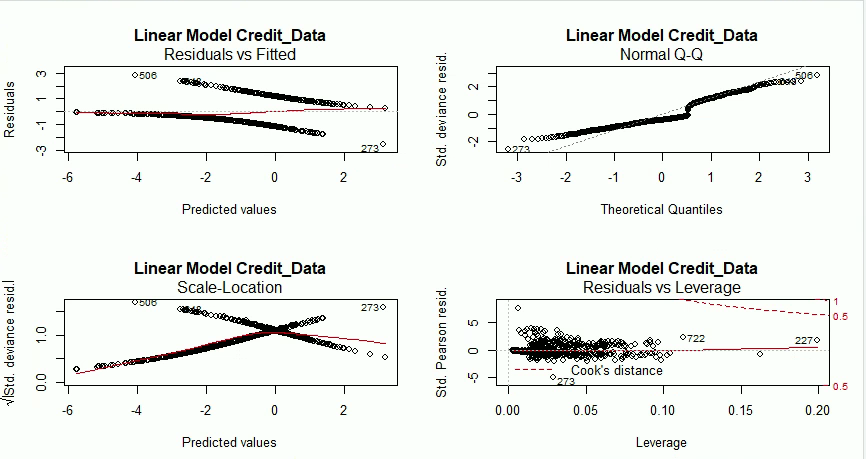
**Iteration 1**

-Variables: CHK\_ACCT, SAV\_ACCT, DURATION, HISTORY, AMOUNT, EMPLOYMENT, INSTALL\_RATE, CO.APPLICANT, GUARANTOR

-Partition: 70/15/15







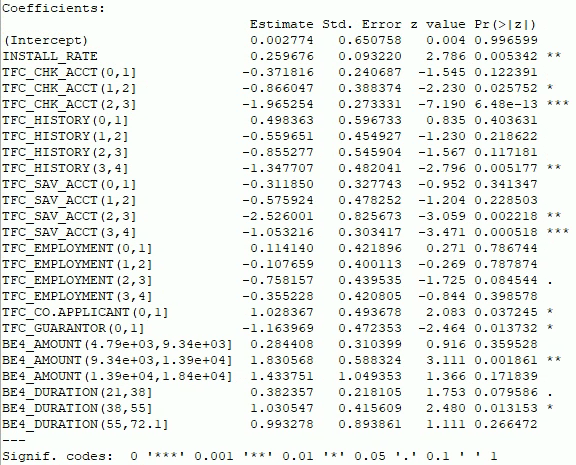
In this first iteration, the model does fairly well, but there are some improvements that could be made. The error rate of 21.3% is good, and the deadly “false negative” of 13.3% is good, but with further iterations we might improve it. I want to pay close attention to the Leverage vs Std. Pearson Residual, which assesses the goodness of fit to the model for each observation. Anything below 5 on the residual is a good fit and above 5 can be considered an outlier. The leverage indicates how effectual that observation is to the model, to as the observation travels to the right on that plot, I want to see it closer to the red line.

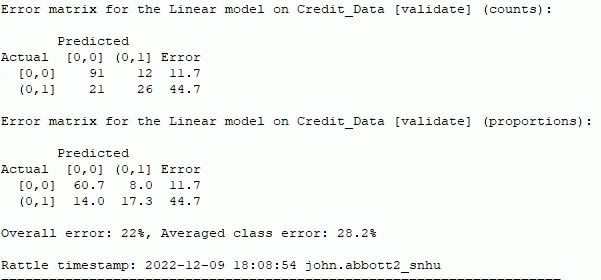
**Iteration 2**

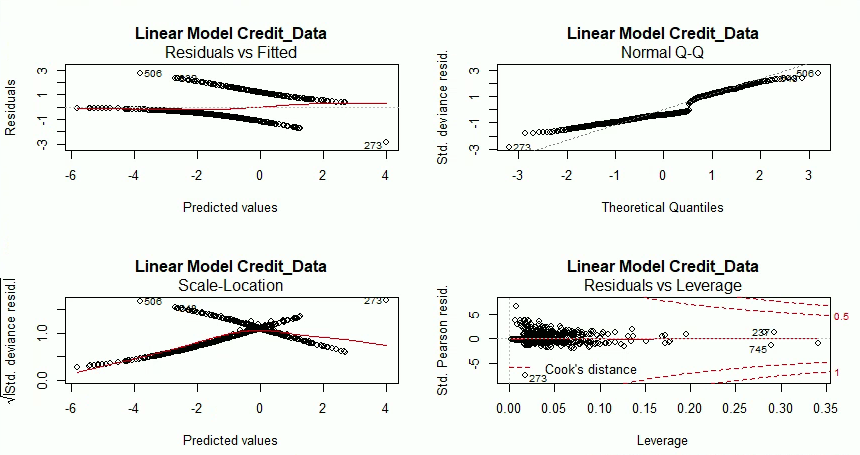
AMMOUNT, DURATION binned as Equal Width

-Variables: CHK\_ACCT, SAV\_ACCT, DURATION, HISTORY, AMOUNT, EMPLOYMENT, INSTALL\_RATE, CO.APPLICANT, GUARANTOR

-Partition 70/15/15



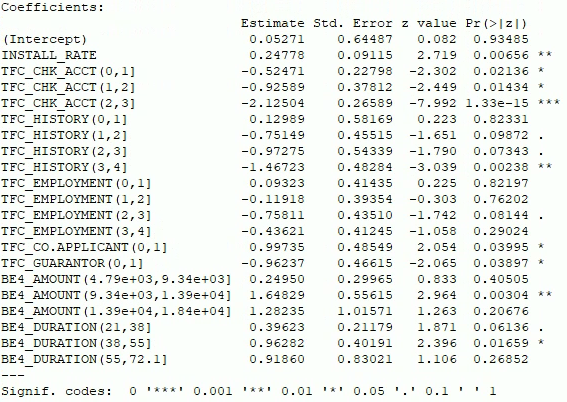


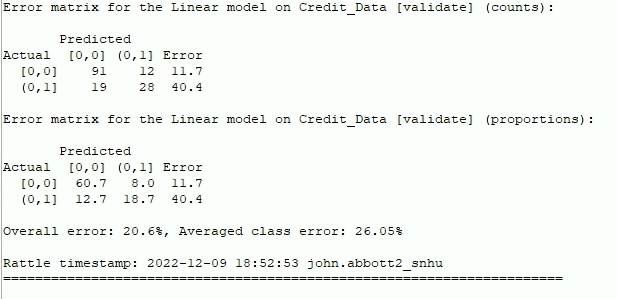


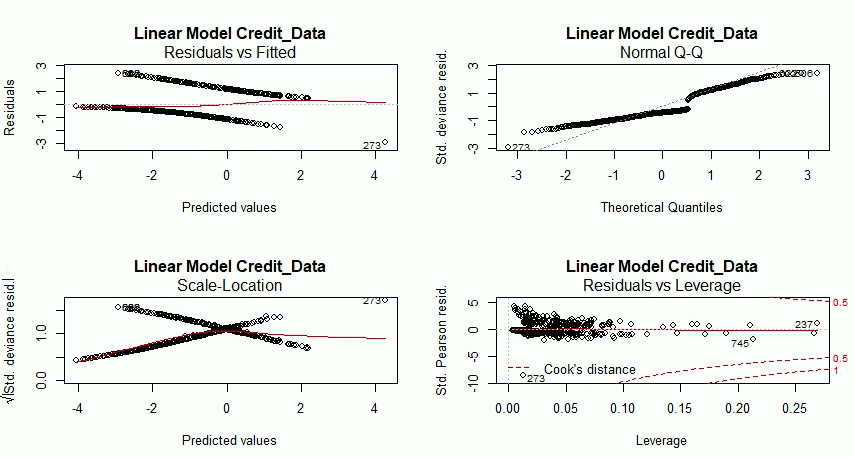
This iteration made every variable categoric by binning numeric variables, and it turned out to be a less effective model. The error was worse and the residuals were more scattered.

**Iteration 3**

Same as 2 but removing SAV\_ACCT



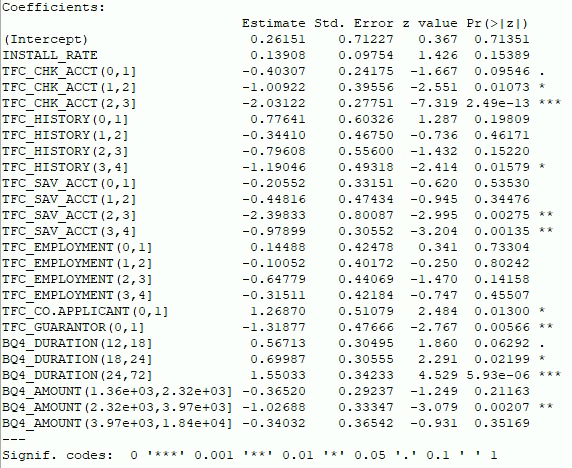


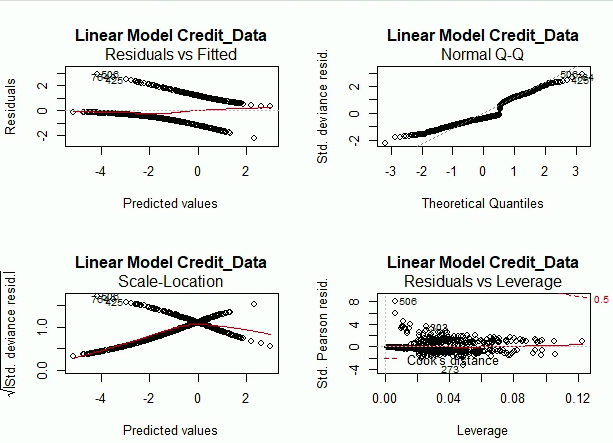


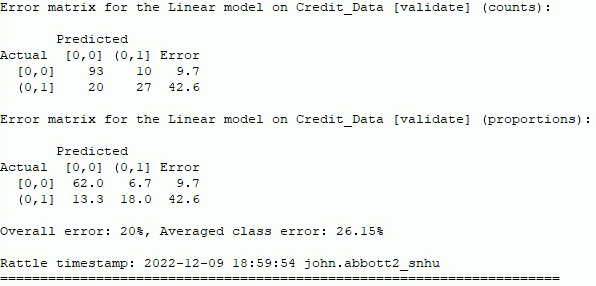
In this iteration I kept the same binning to see if removing SAV\_ACCT would improve the model, since it’s possible that it could be somewhat redundant in the presence of CHK\_ACCT, but that turned out not to be the case, and the model became worse for it. The error was reduced, but the residuals were worse.

**Iteration 4**

Same as 2 but using quantiles instead of equal width bins, SAV\_ACCT reincluded







This is the best iteration I attained. I reinstated SAV\_ACCT, and re-binned AMMOUNT and DURATION as quantiles. This allowed the bins to be based more on their statistics rather than set lengths. The residuals are very tight with only one outlier, and the error is all the way down to 20%. The false negatives are slightly higher than the previous iteration, but the false positives are much lower to compensate.

**Evaluation and Concerns**

The model overall did well. In practice it would manage to avoid a default nearly 90% of the time, and consistently loan to reliable applicants well over 90% of the time. This is far and away vastly better than taking on a 30% default rate as the descriptive statistics would show.

Consistently the residual plot identified observation 506 to be an outlier. However, looking at the data entry, there’s no explicit numeric extreme in entry 506. It could be that 506 was an anomaly, either strongly indicating a default, and then not defaulting, or the reverse. The fact that this only happened once is good news, but I might consider just removing it from the data because of its strangeness.

I would prefer to have gotten the false negative rate down below 10% if possible. Recall that a default on a loan incurs a 150% loss to the company, whereas a false positive (not loaning when the company should have) is a 100% loss of opportunity. I feel that there is a way to bring the false negative down below 10% but I wasn’t able to achieve it.