California Lutheran University

Project 2

ECON-562: Advanced Analytics

Andrew Frohner
4-14-2025

Regression Analysis in R

Introduction:

The data set under analysis is the "Diabetes" data set from the LARS package in software program R.

This data set is popular in "Least Angle" regression analysis. This involves using Linear Regression in a high dimensional setting and adjusting the coefficients in a special way to achieve an "equiangular" effect in the dimensional space.

That is not necessarily what we're here to do, but we will use this data for regression analysis across multiple settings to test prediction accuracy.

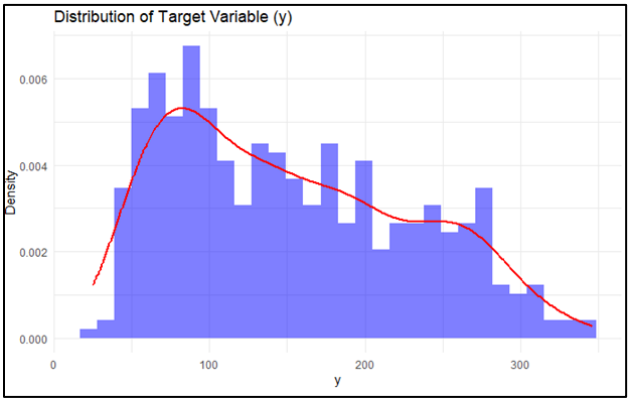
Data Description & Data Quality

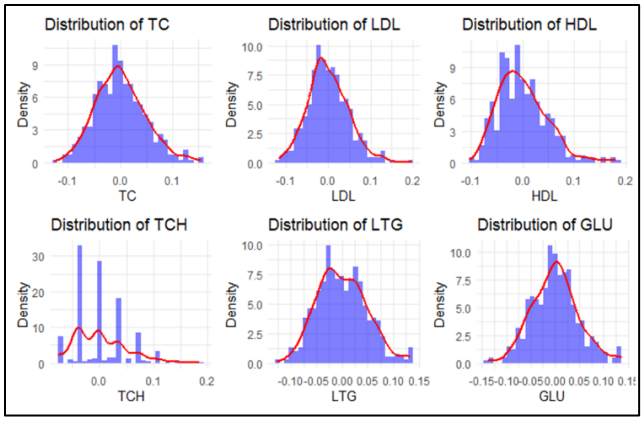
Variable	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
age	442	0.00	0.05	0.01	0.00	0.05	-0.11	0.11	0.22	-0.23	-0.69	0.00
sex	442	0.00	0.05	-0.04	0.00	0.00	-0.04	0.05	0.10	0.13	-1.99	0.00
bmi	442	0.00	0.05	-0.01	0.00	0.05	-0.09	0.17	0.26	0.59	0.07	0.00
map	442	0.00	0.05	-0.01	0.00	0.05	-0.11	0.13	0.24	0.29	-0.55	0.00
tc	442	0.00	0.05	0.00	0.00	0.05	-0.13	0.15	0.28	0.38	0.20	0.00
ldl	442	0.00	0.05	0.00	0.00	0.04	-0.12	0.20	0.31	0.43	0.56	0.00
hdl	442	0.00	0.05	-0.01	0.00	0.05	-0.10	0.18	0.28	0.79	0.94	0.00
tch	442	0.00	0.05	0.00	0.00	0.05	-0.08	0.19	0.26	0.73	0.41	0.00
ltg	442	0.00	0.05	0.00	0.00	0.05	-0.13	0.13	0.26	0.29	-0.16	0.00
glu	442	0.00	0.05	0.00	0.00	0.04	-0.14	0.14	0.27	0.21	0.21	0.00
Outcome	442	152.13	77.09	140.50	147.54	88.21	25.00	346.00	321.00	0.44	-0.90	3.67

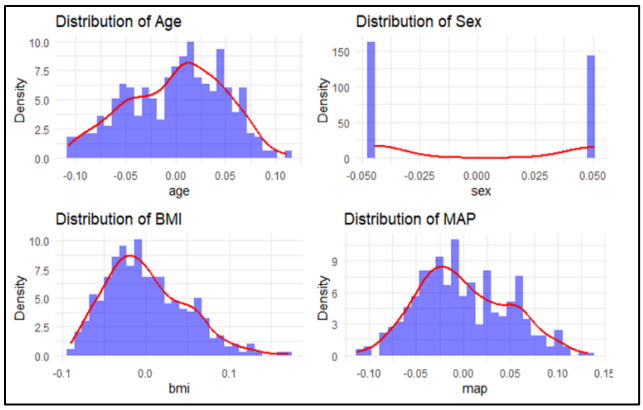
Summary statistics of the raw data are listed above.
The data set totals 442 individual patient observations, and it comes standardized as shown by the mean and "sd" (standard deviation) metric of the predictors.

There are also no Missing values in the data set, and all the predictors are assigned the "numeric" class in R (though sex only takes 2 numeric values -.04 and .05 – the latter denoting female status).

Distributions shown below:







Many of the variables in the data do not follow the normal distribution.

The *Yeo Johnson* transformation was performed to all variables excluding our response variable y and the binary indicator sex.

The linear regression, and the regularized regression settings assumes that residuals are normally distributed.
Without making this transformation our coefficient estimates, and their standard errors will be inaccurate.

Variable	Outliers_Count
age	0
sex	0
bmi	0
map	0
tc	**2**
ldl	**2**
hdl	**5**
tch	0
ltg	**9**
glu	**1**
y	0

The outlier exploration process consisted of counting the number of observations for each variable with a Z-Score above +/- 3.
Much of the data did not lie in outlier territory. The 19 total observations were capped at the 90th percentile value of their respective variable.
This process acknowledges that they are much higher/lower than the average observation but will help our model resist fitting a relationship to these abnormally high values that do not represent the true population.

Analysis

Linear Model

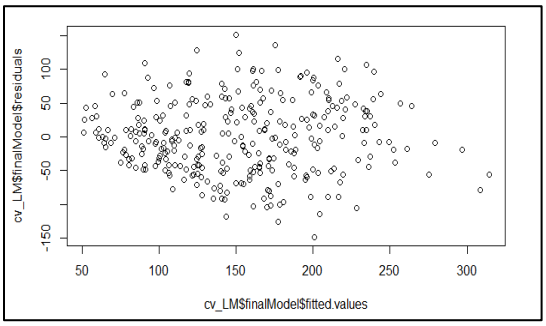
The simplest analysis we can (and should) perform with our data is the linear regression using all the predictors. This model can serve as a benchmark for performance when we're testing more sophisticated analytics.

Results

Model	Mean_Prediction_Error	MSE	RSS	MAD	Rsquared	BIC	LR_stat
Linear Model	46.04479	3179.576	416524.4	57.33114	0.4864485	3444.914	202.5953

Summary:

```
Raw Data Regression Summary
================================================
                        Dependent variable:
                    ----------------------------
                               .outcome
------------------------------------------------
age                          -0.675 (75.550)
sex                        -168.264** (76.216)
bmi                        572.804*** (85.578)
map                        312.143*** (82.831)
tc                      -19,056.170 (32,415.590)
ldl                      16,607.990 (28,486.800)
hdl                      6,994.647 (12,119.430)
tch                         193.973 (245.846)
ltg                      6,224.260 (10,114.880)
glu                         95.446 (80.175)
Constant                    70.847 (143.835)
------------------------------------------------
Observations                   311
R2                            0.479
Adjusted R2                   0.461
Residual Std. Error      56.080 (df = 300)
F Statistic          27.549*** (df = 10; 300)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```



The Linear Model found significance on sex, bmi, map, ltg. The coefficients are significant at the 5% and 1% level – provided strong evidence that their effect is truly different than zero, but their standard errors are also very high. These results were generated after performing 10-fold cross validation and fitting the model on the entire data set.

The Mean prediction error is lower than the MAD, suggestive that the model is predicting diabetes outcomes that are *lower* than the true outcome.

The LR statistic helps us compare a more complex model to a simpler one. Here I used the LR stat to compare to an "intercept-only" model. The reading of 210 provides strong evidence that the predictor set provides much stronger predictive power than just using a vector of 1's. This is good to know – but it doesn't provide great comfort in the face of abnormally large estimates and large standard errors.

The large estimates could be a result of multi-collinearity in our data. 6 of our predictors are derived from a patient's blood sample.

Best-Subset Model

Another simple alternative to running a linear model with all predictors is to run all possible models and select the model with the best performance on a specified metric. Here, selected the model that generated the *smallest* BIC metric.

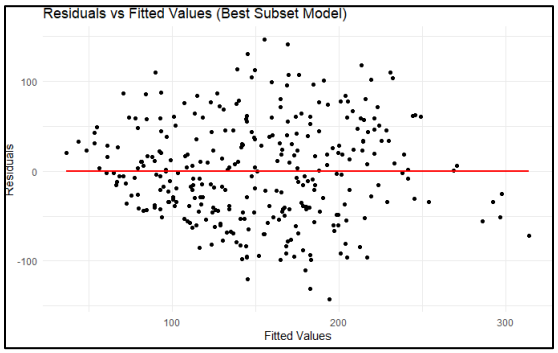
Minimizing prediction errors likely would have generated the same 10 predictor model we just assessed. That model predicts well because it is using "full information" to generate a prediction of diabetes outcome. Which is great for this data, but we want to deploy this model on new data. BIC is a favorable metric to optimize upon because it will penalize adding complexity (more predictors) to our model.

Results:

Model	Mean_Prediction_Error	MSE	MAD	RSS	Rsquared	BIC	LR_stat
Best Subset Selection Model	46.1128	3207.282	56.69052	420153.9	0.4819736	3419.105	2.889991

Summary:

```
Best Subset Selection Model Summary
================================================
                        Dependent variable:
                    ---------------------------
                                y
------------------------------------------------
sex                     -157.589** (74.750)
bmi                     581.135*** (82.365)
map                     326.797*** (79.342)
hdl                    -231.336*** (77.956)
ltg                     326.349*** (59.644)
Constant                153.205*** (3.323)
------------------------------------------------
Observations                  311
R2                           0.474
Adjusted R2                  0.465
Residual Std. Error     55.877 (df = 305)
F Statistic          54.933*** (df = 5; 305)
================================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```



The Best Subset is aiming to capture the "true data generating process" within the data. It is not penalizing any terms in the model, but it excluding them if they do not meaningfully add to the model's fit after considering model complexity it adds.

The Best Subset of this data was found with sex, bmi, map, hdl, and ltg predictors. All the predictors were found statistically significant as expected, and the standard errors on hdl & ltg are significantly lower than the linear model.

The LR statistic (here comparing Best subset to the linear model) is 2.88. Providing strong evidence that the Best subset predicts almost as well as the full model.

This reduction in complexity does not come at an unreasonably high cost to prediction accuracy as the Mean Prediction error increased less than one-tenth of a percentage point using the best subset.

This model looks favorable to the linear model because of its simplicity and more realistic coefficient estimates. Suggestive that a best-subset reduces collinearity present in our data.

Ridge Regression Model

The Ridge Regression is an obvious candidate for the diabetes data set. Here we have 6 predictors that are generated from a single blood sample. There is a chance that these predictors may exhibit collinearity, particularly the cholesterol metrics. Ridge regression's process of shrinking coefficient helps reduce the variability of estimates and model complexity, and aid in interpretability.

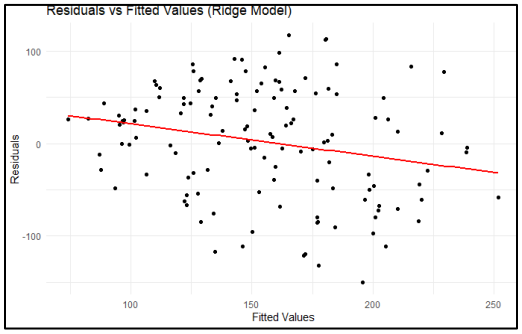
Results

Model	Mean_Prediction_Error	MSE	MAD	RSS	Rsquared	BIC	LR_stat
Ridge Regression Model	50.19658	3623.696	65.41329	474704.2	0.4147161	2342.971	-206.0514

Summary:

Ridge Regression Model Coefficients

term	step	estimate	lambda	dev.ratio
(Intercept)	1	153.83315	66.66497	0.4316289
age	1	41.87957	66.66497	0.4316289
sex	1	-55.45510	66.66497	0.4316289
bmi	1	325.57157	66.66497	0.4316289
map	1	210.69518	66.66497	0.4316289
tc	1	27.42247	66.66497	0.4316289
ldl	1	-16.39153	66.66497	0.4316289
hdl	1	-127.98615	66.66497	0.4316289
tch	1	133.21700	66.66497	0.4316289
ltg	1	181.46599	66.66497	0.4316289
glu	1	118.50454	66.66497	0.4316289



The coefficient estimates from the model reflect this. The highest magnitude predictors from the linear model (age, sex, bmi, map, ltg, and tc) are scaled down dramatically from their OLS estimates.

What's more important is the sign changes on variables age, tc, and hdl (they reversed in the ridge regression!).

They reversed to the theoretically correct sign as well. It is more intuitive that as age and total cholesterol increase a patient is at a higher risk of receiving a diabetes diagnosis or outcome. The linear model did not catch that – it was fitting the pattern it saw in this particular data set.

The BIC with Ridge also prints significantly lower than OLS (at the cost of some prediction accuracy).

The pattern in the Resids vs Fitted values is not perfect but appears to have slightly less pronounced vertical spread than OLS.

The output of Ridge regression provides evidence to support we're effectively mitigating collinearity, and capturing more "general" patterns in our data

LASSO Regression Model

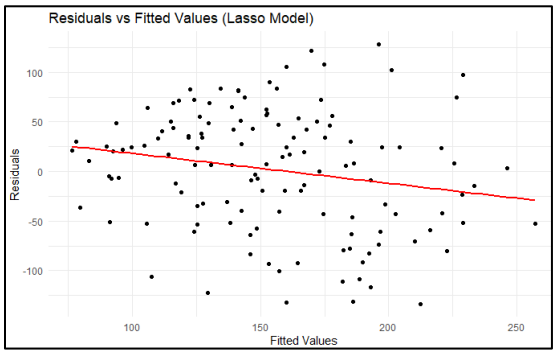
An alternative regularization method we can employ in the regression setting is LASSO. Unlike Ridge, where we mitigated collinearity by shrinking coefficients but keeping them in the data set – LASSO will shrink some of the coefficients all the way to zero. This is a more aggressive method of reducing model complexity, as collinear predictors will be entirely removed from our model.

Results

Model	Mean_Prediction_Error	MSE	MAD	RSS	Rsquared	BIC	LR_stat
Lasso Regression Model	50.35466	3651.921	63.76617	478401.7	0.4101573	2345.384	-211.2103

Summary:

term	step	estimate	lambda	dev.ratio
(Intercept)	1	152.73296	9.232341	0.4371433
bmi	1	532.54926	9.232341	0.4371433
map	1	172.34703	9.232341	0.4371433
hdl	1	-54.04956	9.232341	0.4371433
tch	1	1.87760	9.232341	0.4371433
ltg	1	258.87678	9.232341	0.4371433



LASSO regularization left the model with 5 predictors 4 of which appeared in the best subset selection (bmi, map, hdl, ltg in the best subset). This is good signal that the linear regression approach is in fact capturing the meaningful predictors in our data.

Again, we see the opposite sign on the hdl predictor – reflecting the theoretically correct relationship good cholesterol has with diabetes.

The small estimate on tch (relative to the higher coeff on hdl) confirms that we were modeling we were not mitigating collinearity in the non-regularized regression setting, as tch is a ratio of Total cholesterol / HDL cholesterol.

The Coefficient estimate on ltg (glucose) level is closer to the estimate in the Best Subset model – suggesting that LASSO is more effectively capturing general data patterns in the data rather than fitting to the specific pattern in this single data set.

As expected, we give up some pure prediction accuracy noting the Mean prediction error is 22% higher than the linear model, but we reduce the BIC significantly and use 5 fewer predictors.

Conclusion

The linear regression setting provides a great framework for estimating a relationship between a set of predictors and a target variable. In addition to that this modeling framework allows the modelers to interpret how the predictors impact the target. This provides great benefit in the context of a study on diabetes outcomes.
The simplest modeling approach allows us to find an "accurate" model. But it comes with coefficient estimates that do not make sense.

This doesn't negate us from using linear regression, but it does propel us to test different techniques to uncover a model that will not only predict accurately but will provide a more realistic interpretation of this particular data and will perform similarly if we deploy it on new data.

Model	Mean_Prediction_Error	MSE	RSS	MAD	Rsquared	BIC	LR_stat
Linear Model	46.04479	3179.576	416524.4	57.33114	0.4864485	3444.914	202.595312
Best Subset Selection Model	46.11280	3207.282	420153.9	56.69052	0.4819736	3419.105	2.889991
Ridge Regression Model	50.19658	3623.696	474704.2	65.41329	0.4147161	2342.971	-213.623301
Lasso Regression Model	50.35466	3651.921	478401.7	63.76617	0.4101573	2345.384	-211.210268

Using regularization and predictor subset selections we were able to alter the functional form of the model and find a data generating process that predicts *comparatively* well to the fully specified linear model with significantly reduced complexity.

Regularization and subset selection are not always this effective. These are relatively sophisticated statistical techniques that require careful selection of regularization parameters and awareness of performance metrics that a model is supposed to target. We were able to find a strong subset model and regularized models through partitioning a large data set and 10-fold cross-validation.

In the context of best subsets – we tested all possible model combinations across 10 splits of training data before selecting a model an optimal predictor set to use for performance measurement on test data.

In the context of LASSO and Ridge, we cross-validated the hyperparameter lambda to tune it towards the value that results in the lowest prediction error – then selected the hyperparameter that reduces overall variance and model complexity.

The diabetes data set is but a sample of 442 people. Deploying the linear model assumes that the entire population exhibits the same diabetes outcomes – they do not. The alternate model we've developed are designed more generally, keeping in mind that the populations diabetes outcomes may be different.

Appendix

ECON-562-PROJ2-AF.html