Heart Failure Explorations in R

First few code chunks are a linear model analysis. In conducting a regression analysis, I perform an initial summary and analysis, do variable selection with AIC, run diagnostics to consider whether any data transformations are necessary, and then interpret the model given the data.

After a linear model analysis, we look towards other methods to understand the dataset further.

Dataset comes from Kaggle: https://www.kaggle.com/andrewmvd/heart-failure-clinical-data

```r
h_fail <- read.csv("datasets_727551_1263738_heart_failure_clinical_records_dataset.csv")

lmod <- lm(DEATH_EVENT ~ ., h_fail) ## Full variable model
summary(lmod)
```

```
##
## Call:
## lm(formula = DEATH_EVENT ~ ., data = h_fail)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80866 -0.28041 -0.04205  0.24742  0.96983
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.664e+00  6.954e-01   2.392  0.01738 *
## age                       5.767e-03  1.867e-03   3.088  0.00221 **
## anaemia                  -2.766e-03  4.438e-02  -0.062  0.95035
## creatinine_phosphokinase  3.427e-05  2.247e-05   1.525  0.12840
## diabetes                  1.928e-02  4.410e-02   0.437  0.66236
## ejection_fraction        -9.834e-03  1.844e-03  -5.333 1.96e-07 ***
## high_blood_pressure      -1.430e-02  4.565e-02  -0.313  0.75438
## platelets                -8.370e-08  2.208e-07  -0.379  0.70492
## serum_creatinine          8.527e-02  2.123e-02   4.017 7.54e-05 ***
## serum_sodium             -7.599e-03  5.024e-03  -1.513  0.13149
## sex                      -6.369e-02  5.108e-02  -1.247  0.21353
## smoking                  -5.733e-03  5.119e-02  -0.112  0.91091
## time                     -2.733e-03  2.903e-04  -9.415  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3646 on 286 degrees of freedom
## Multiple R-squared:  0.4168, Adjusted R-squared:  0.3924
## F-statistic: 17.04 on 12 and 286 DF,  p-value: < 2.2e-16
```

From the summary alone on the full model, we observe that some variables are more significant than others. To account for this, we can instead do variable selection to find the optimal model. Rather than traditional backward elimination, we use the AIC (Akaike information criterion) to find the best model.

```r
mod1 <- step(lmod)
```

```
## Start:  AIC=-590.72
## DEATH_EVENT ~ age + anaemia + creatinine_phosphokinase + diabetes +
##     ejection_fraction + high_blood_pressure + platelets + serum_creatinine +
##     serum_sodium + sex + smoking + time
## 
##                            Df Sum of Sq    RSS     AIC
## - anaemia                   1    0.0005 38.010 -592.72
## - smoking                   1    0.0017 38.011 -592.71
## - high_blood_pressure       1    0.0130 38.022 -592.62
## - platelets                 1    0.0191 38.028 -592.57
## - diabetes                  1    0.0254 38.034 -592.52
## - sex                       1    0.2066 38.216 -591.10
## <none>                                38.009 -590.72
## - serum_sodium              1    0.3041 38.313 -590.34
## - creatinine_phosphokinase  1    0.3090 38.318 -590.30
## - age                       1    1.2676 39.277 -582.91
## - serum_creatinine          1    2.1446 40.154 -576.31
## - ejection_fraction         1    3.7801 41.789 -564.37
## - time                      1   11.7810 49.790 -512.00
## 
## Step:  AIC=-592.72
## DEATH_EVENT ~ age + creatinine_phosphokinase + diabetes + ejection_fraction + 
##     high_blood_pressure + platelets + serum_creatinine + serum_sodium + 
##     sex + smoking + time
## 
##                            Df Sum of Sq    RSS     AIC
## - smoking                   1    0.0015 38.011 -594.71
## - high_blood_pressure       1    0.0129 38.022 -594.62
## - platelets                 1    0.0189 38.028 -594.57
## - diabetes                  1    0.0255 38.035 -594.52
## - sex                       1    0.2060 38.216 -593.10
## <none>                                38.010 -592.72
## - serum_sodium              1    0.3075 38.317 -592.31
## - creatinine_phosphokinase  1    0.3256 38.335 -592.17
## - age                       1    1.2677 39.277 -584.91
## - serum_creatinine          1    2.1446 40.154 -578.31
## - ejection_fraction         1    3.7804 41.790 -566.37
## - time                      1   11.9876 49.997 -512.75
## 
## Step:  AIC=-594.71
## DEATH_EVENT ~ age + creatinine_phosphokinase + diabetes + ejection_fraction + 
##     high_blood_pressure + platelets + serum_creatinine + serum_sodium + 
##     sex + time
## 
##                            Df Sum of Sq    RSS     AIC
## - high_blood_pressure       1    0.0127 38.024 -596.61
## - platelets                 1    0.0202 38.031 -596.55
## - diabetes                  1    0.0270 38.038 -596.50
## <none>                                38.011 -594.71
## - sex                       1    0.2733 38.284 -594.57
## - serum_sodium              1    0.3077 38.319 -594.30
## - creatinine_phosphokinase  1    0.3283 38.339 -594.14
## - age                       1    1.2696 39.281 -586.88
## - serum_creatinine          1    2.1513 40.162 -580.25
## - ejection_fraction         1    3.7793 41.790 -568.37
## - time                      1   11.9897 50.001 -514.73
## 
```

```
## Step:  AIC=-596.61
## DEATH_EVENT ~ age + creatinine_phosphokinase + diabetes + ejection_fraction +
##     platelets + serum_creatinine + serum_sodium + sex + time
##
##                            Df Sum of Sq    RSS     AIC
## - platelets                 1    0.0216 38.045 -598.44
## - diabetes                  1    0.0277 38.051 -598.39
## <none>                                  38.024 -596.61
## - sex                       1    0.2641 38.288 -596.54
## - serum_sodium              1    0.3143 38.338 -596.15
## - creatinine_phosphokinase  1    0.3382 38.362 -595.96
## - age                       1    1.2591 39.283 -588.87
## - serum_creatinine          1    2.1649 40.189 -582.05
## - ejection_fraction         1    3.7790 41.803 -570.28
## - time                      1   12.2858 50.310 -514.89
##
## Step:  AIC=-598.44
## DEATH_EVENT ~ age + creatinine_phosphokinase + diabetes + ejection_fraction +
##     serum_creatinine + serum_sodium + sex + time
##
##                            Df Sum of Sq    RSS     AIC
## - diabetes                  1    0.0242 38.070 -600.25
## - sex                       1    0.2516 38.297 -598.47
## <none>                                  38.045 -598.44
## - serum_sodium              1    0.3232 38.369 -597.91
## - creatinine_phosphokinase  1    0.3335 38.379 -597.83
## - age                       1    1.2718 39.317 -590.61
## - serum_creatinine          1    2.1755 40.221 -583.81
## - ejection_fraction         1    3.8181 41.863 -571.84
## - time                      1   12.2760 50.321 -516.82
##
## Step:  AIC=-600.25
## DEATH_EVENT ~ age + creatinine_phosphokinase + ejection_fraction +
##     serum_creatinine + serum_sodium + sex + time
##
##                            Df Sum of Sq    RSS     AIC
## <none>                                  38.070 -600.25
## - sex                       1    0.2830 38.353 -600.03
## - creatinine_phosphokinase  1    0.3337 38.403 -599.64
## - serum_sodium              1    0.3460 38.416 -599.54
## - age                       1    1.2516 39.321 -592.58
## - serum_creatinine          1    2.1579 40.228 -585.76
## - ejection_fraction         1    3.8212 41.891 -573.65
## - time                      1   12.2628 50.332 -518.76
```

```r
summary(mod1)
```

```
##
## Call:
## lm(formula = DEATH_EVENT ~ age + creatinine_phosphokinase + ejection_fraction +
##     serum_creatinine + serum_sodium + sex + time, data = h_fail)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7966 -0.2792 -0.0428  0.2440  0.9754
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.703e+00  6.809e-01   2.501  0.01292 *
## age                       5.692e-03  1.840e-03   3.093  0.00217 **
## creatinine_phosphokinase  3.484e-05  2.181e-05   1.597  0.11130
## ejection_fraction        -9.874e-03  1.827e-03  -5.405 1.35e-07 ***
## serum_creatinine          8.526e-02  2.099e-02   4.061 6.28e-05 ***
## serum_sodium             -8.021e-03  4.932e-03  -1.626  0.10498
## sex                      -6.558e-02  4.459e-02  -1.471  0.14239
## time                     -2.710e-03  2.799e-04  -9.682  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3617 on 291 degrees of freedom
## Multiple R-squared:  0.4159, Adjusted R-squared:  0.4019
## F-statistic:  29.6 on 7 and 291 DF,  p-value: < 2.2e-16
```

After variable selection, we see the best predictors for the response are age, creatinine_phosphokinase, ejection_fraction, serum_creatinine, serum_sodium, sex, and time. From the summary, we note that there are still variables that are insignificant at the 0.1 level. Under traditional backward selection, we would remove those variables and have an even *smaller* model. Instead, we can trust variable selection with minimal AIC as it is often a good estimator for the test error of the model. More info on the AIC criterion can be found in Hastie, Tibshirani, Friedman's *Elements of Statistical Learning* Section 7.5.