Mary Washington

MATH 440

Dr. James Young

03 December 2018

AIC vs BIC model selection on Heart Disease prediction and cross validation

The article I looked at is called "Comparison of Akaike information criteria (AIC) and Bayesian information criteria (BIC) in selection of an asymmetric price relationship" by Henry de-Graft Acquah. Asymmetrical price transmission (APT) is the phenomenon of when prices of ready products increase promptly whenever prices of inputs increase but take time to decrease after input price decreases. This article looked at model selection using both AIC and BIC and makes the argument that BIC outperforms AIC in selecting the price relationship in large samples. Monte Carlo experimentation was used. According to MIT Monte Carol simulations (experimentations) are a statistical technique used to model probabilistic systems and establish the odds for a variety of outcomes.

The purpose of APT is to find a single model that best captures the underlying data generating process for derivation of policy conclusions. Both AIC and BIC have been used to address the model selection process because they are both information-theoretic. Both have different purposes. Akaike's Information Criteria (AIC) is a model selection that minimizes the negative likelihood penalized by the number of parameters. Bayesian Information Criteria (BIC) is derived within the Bayesian framework but does not use the prior and posterior as traditional Bayes. AIC is used to find the best model to the unknown data generating process, while BIC is used to approximate the true model. Because of this, AIC does not directly depend on the sample size. BIC reflects sample size and has properties of asymptotic consistency which AIC lacks. BIC applies a larger penalty than AIC, for reasonable sample sizes, and by ceteris paribus, it selects a simpler model. Though numerous model selection criteria have been extended off the idea of AIC and BIC, such as Consistent Akaike Information Criteria (CAIC) (Burhnam and Anderson, 1998) and Draper's Information Criteria (DIC)(Draper, 1995), AIC and BIC are easier to compute using most standard software, and thus are more extensively used in empirical analysis. Previous studies have mostly focused on comparing the two using generalized linear models. The article focuses on using the two to empirically evaluate the performances for testing asymmetry. This was a supervised learning experiment and the true data generating process was known for all experiments and the Monte Carlo simulations were essential in deriving the model recovery rates of the true model. Below is the Granger and Lee asymmetric Error Correction Model data generating process (DGP) (equation 1):

$$\Delta \boldsymbol{P}_{\boldsymbol{A},\boldsymbol{t}} = \boldsymbol{\beta}_{\mathbf{1}} \Delta \boldsymbol{P}_{\boldsymbol{B},\boldsymbol{t}} + \boldsymbol{\beta}_{\mathbf{2}}^{+} \boldsymbol{ECT}_{\boldsymbol{t-1}}^{+} + \boldsymbol{\beta}_{\mathbf{2}}^{-} \boldsymbol{ECT}_{\boldsymbol{t-1}}^{-} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\delta}_{\boldsymbol{\epsilon}}^{\mathbf{2}})$$

Various samples sizes were used, P_A and P_B are non-stationary variables and there is an equilibrium relationship between the two which produce a stationary series. The equilibrium equation is estimated by least squares and lagged deviation, which is denoted by the ECT_{t-1} term. The ECT is decomposed into both positive and negative deviations using Wolfram segmentation

(Granger and Lee, 1989) and plugged into the previous equation and the hypothesis test for asymmetry is H_0: $\beta_2^+ = \beta_2^-$.

$$\boldsymbol{Where}: \boldsymbol{ECT} = \boldsymbol{P_A} - \boldsymbol{\beta_1 P_{Bt}}$$

$$\boldsymbol{ECT_{t-1}^+} = \boldsymbol{ECT_{t-1}}\ \boldsymbol{if}\ \boldsymbol{ECT_{t-1}} > \mathbf{0}\ \boldsymbol{and}\ \mathbf{0}\ \boldsymbol{otherwise}$$

$$\boldsymbol{ECT_{t-1}^-} = \boldsymbol{ECT_{t-1}}\ \boldsymbol{if}\ \boldsymbol{ECT_{t-1}} < \mathbf{0}\ \boldsymbol{and}\ \mathbf{0}\ \boldsymbol{otherwise}$$

The von Cramon-Taubadel and Loy (1996) asymmetric Error Correction Model (equation 2) can be written as follows:

$$\Delta \boldsymbol{P_{A,t}} = \boldsymbol{\beta_1^+} \Delta \boldsymbol{P_{B,t}^+} + \boldsymbol{\beta_1^-} \Delta \boldsymbol{P_{B,t}^-} + \boldsymbol{\beta_2^+ ECT_{t-1}^+} + \boldsymbol{\beta_2^- ECT_{t-1}^-} + \boldsymbol{\epsilon}\ \ \boldsymbol{\epsilon} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\delta_\epsilon^2})$$

Where $\Delta P_{B,t}^+$ and $\Delta P_{B,t}^-$ are the positive and negative changes and the remaining variables are defined as in equation 1. Von Cramon-Taubadel and Loy (1996) suggested that $\Delta P_{A,t}$ can be broken up into both positive and negative components as well to allow for more complex dynamics. The hypothesis test for asymmetry are H_0: $\beta_1^+ = \beta_1^-$ and $\beta_2^+ = \beta_2^-$. Since equation 2 uses a linear combination of coefficients, and joint F-test can be used to determine the symmetry or asymmetry of the price transmission process.

Lastly a contrasting model by Houck(1977), proposes that does not take into account adjustments to the equilibrium level but the direct impact of price increases and decreases are affected by asymmetries specified. Model variables are defined as in equation 2 and the hypothesis test for asymmetry is H_0: $\beta_1^+ = \beta_1^-$

$$\Delta \boldsymbol{P_{A,t}} = \boldsymbol{\beta_1^+} \Delta \boldsymbol{P_{B,t}^+} + \boldsymbol{\beta_1^-} \Delta \boldsymbol{P_{B,t}^-} + \boldsymbol{V_t}\ \ \boldsymbol{V_t} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\delta_v^2})$$

AIC = -2Log(p(L)) + 2p and **BIC = -2Log(p(L))+plog(n)** where p = number of parameters and L = Likelihood under the fitted model. AIC draws from (Akaike, 1973; Bozdogan, 1987; Zucchini, 2000) and BIC has the factors for two competing models (Schwarz, 1978; Kass and Rafftery, 1995). Clearly the difference is the last term which depends on the sample size. The purpose of the simulation study is to see which one does a better job of identifying the true model. The two models are fitted to the simulated data and their ability to recover the true models were measured. The recovery rates were derived using 1000 Monte Carlo simulations with sample sizes of 50, 150, and 500. The relative performances are compared by their ability to recover the true DGP across various model recovery rates (sample sizes) as illustrated in the table below: CECM = complex asymmetric error correction model, HKD = Houck's model, and SECM = standard asymmetric error correction model. As the sample size increased, the model selection methods had and increased ability to recover the true asymmetric DGP. AIC is inconsistent and doesn't improve in performance in large samples while BIC is both consistent and improves in performance in large sample size.

Table 1. Relative performance of the model selection methods across sample size.

Experiment criterion	Model fitted			
	Methods	CECM (%)	HKD (%)	SECM (DGP) (%)
$n=50$ $\sigma=1$	AIC	17.0	4.8	78.2
	BIC	6.3	11.9	81.8
$n=150$ $\sigma=1$	AIC	17.5	0.0	82.5
	BIC	3.0	0.1	96.9
$n=500$ $\sigma=1$	AIC	16.8	0.0	83.2
	BIC	1.6	0.0	98.4

Note: Recovery rates based on 1000 replications.

Next the effects of noise level on the model was simulated by using three error sizes ranging from small to large. Again 1000 Monte Carlo simulations were performed with a sample size of 150. This is compared to the true model error rate in the DGP and was increased systematically. The performance was deteriorated with increasing amounts of noise in SECM. BIC outperformed AIC for smaller noise but lost at higher noise levels, see table below.

Table 2. Relative performance of the selection methods across error size.

Experiment criterion	Model fitted			
	Methods	CECM (%)	HKD (%)	SECM (DGP) (%)
$n=150$ $\sigma=3$	AIC	12.3	22.4	65.3
	BIC	1.2	52.1	46.7
$n=150$ $\sigma=2$	AIC	17.3	15.1	77.7
	BIC	1.8	18.7	79.5
$n=150$ $\sigma=1$	AIC	18.3	0.0	81.7
	BIC	2.4	0.1	97.5

Note: Recovery rates percentages based on 1000 replications.

Small error and large sample sizes improves the recovery rate of the true asymmetric data generating process and vice versa. See the table below

Table 3. Effects of sample size and stochastic variance on model recovery.

Experiment criterion	Model fitted			
	Methods	CECM (%)	HKD (%)	SECM (DGP) (%)
$\sigma=2$ $n=50$	AIC	9.9	35.4	54.7
	BIC	2.8	55.9	41.3
$n=150$ $\sigma=0.5$	AIC	18.3	0.0	81.7
	BIC	2.5	0.0	97.5

Note: Recovery rates based on 1000 replications.

An increase in the difference between the asymmetric adjustments parameters from 0.25 and 0.50 led to an improvement in the model recovery rates of the try asymmetric data generating process by the model selection methods. See table below

Table 4. Effects of the level of asymmetry on model recovery.

Experiment criterion	Model Fitted			
	Methods	CECM (%)	HKD (%)	SECM (DGP) (%)
$\beta_2^+ - \beta_2^- = 0.25$	AIC	16.35	0.22	83.43
	BIC	2.71	1.72	95.57
$\beta_2^+ - \beta_2^- = 0.50$	AIC	16.5	0.00	83.5
	BIC	2.83	0.03	97.14

Note: Recovery rates based on 1000 replications.

This echoed existing theoretical and empirical work on the performance of model selection methods in other applications. The findings of the current study reinforced the importance of design and informativeness in conducting APT analysis. Overall the current results suggest that AIC should be preferred in smaller samples while BIC should be preferred in larger models, in the ATP context. On the level of asymmetry, these results indicated that performance of the model selection methods depend on the difference in asymmetric adjustments parameters and speeds. Adjusting the parameters had an improved effect on the testing of asymmetry. BIC should be preferred to AIC in applications in which the data has strong levels of asymmetry. This study shed light empirically on the relative performance of try model selection algorithms and established that BIC correctly identified the try asymmetric data generating process.

I also looked at the paper "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion" by M. Stone. Dr. Stone states "when maximum likelihood estimation is used in a model, logarithmic assessment of the performance of predicting density is found to lead to asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Akaike recognized that the method of choice for models that have a large difference in their parametric dimensionality, is unsatisfactory because of the unreserved maximization of likelihood.

Stone adopted the notation of $S= \{(x_i,y_i), I = 1,\ldots,n\}$ and supposing that S is a database with n number of items, and the problems is the choice of predicting density for y given x from a class of formal predicting densities $\{f(y|x,\alpha,S),\alpha \varepsilon A \}$, whose members are indeed by the choice parameter α. The choice of α specifies a predicting density of *y* for each *x* , whose form depends in a prescribed way of S. This notation is not intended to carry any other probabilistic interpretation. There are two complementary cases:

Case 1: $f(y|x,\alpha,S) = f(y|x, \alpha)$ is an independent set of S

Case 2: $f(y|x, \alpha,S)$ properly dependent set of S

Case 1 is just the formal equivalent to a statistical model with a conventional α. The paper focuses on case 2, which is a general example and called A after Akaike. Now the formula is

$f(y|x, \alpha,S)=f_\alpha(y|x,\theta\hat{}_\alpha(S))$, where $f_\alpha(y|x,\theta_\alpha)$, $\theta_\alpha \varepsilon \Theta_\alpha$ and are the densities for a conventional parametric model α and $\theta\hat{}_\alpha(S)$ is the supposed unique likelihood estimator maximizing $L(\alpha, \theta_\alpha) = \Sigma \log f_\alpha(y|x,\theta_\alpha)$.

For case 1 the choice of α to maximize $A(\alpha)$ would be equivalent to MLE of α for the "log-likelihood". Case 1 introduces no new innovations. However, case 2 uses cross-validation of $f^{(i)}(y)=f(y|x_i, \alpha,S_{-i})$ where $S_{-I} = S\ (x_i,y_i)$. This gives us $A(\alpha)= \Sigma\log f(y_i|x_i, \alpha,S_{-i})$. This is asymptotically equivalent, under weak conditions, to AIC, which "corrects" maximum likelihood as a method of choice of model. Now suppose that θ^ and θ_{-I} are unique solutions for $L'(\theta) = 0$ and $L'(\theta) – l'(y_i|x_i,\theta) = 0$ respectively. Also suppose L'' is invertible.

The key assumption for the asymptotic equivalence with Akaike's criterion is: *The conditional distribution of y given x in the distribution P is f(y|x,θ*) for some unique θ*ε Θ, that is, the conventional model {f(y|x, θ), θ ε Θ is true.* This gives us general equivalence, but weaker assumptions will suffice for particular choices of $\{f_\alpha(y|x, \theta_\alpha), \theta_\alpha\ \varepsilon\ \Theta_\alpha\}$. If we consider two models α_1, α_2 and of type $f_\alpha(y|x,\theta_\alpha)$, $\theta_\alpha\ \varepsilon\ \Theta_\alpha$ with $\Theta_{\alpha1}\ c\ \Theta_{\alpha2}$ and suppose both are true, then we know that under regularity conditions,

$2\{L(\alpha_2, \theta\text{^}_{\alpha2})-L(\alpha_1, \theta\text{^}_{\alpha1})\}$ is asymptotically x^2 with $d = p_{\alpha2}-p_{\alpha1}$ degrees of freedom. Hence, by $A(\alpha_2)-A(\alpha_1)$ is asymptotically $½x^2_d – d$. This shows how the simpler model will be favored by the choice criterion $A(\alpha)$.

For the first article I enjoyed learning what an asymmetric price relation is. I have an interest in economics, so this article was quite interesting. I also learned what Monte Carlo experimentation is and how many industries do this when they do a simulated study. Since I have taken a few algorithms class and understand both time and space complexity of I understand the need for efficiency. If an algorithm just performs but doesn't perform well, what is the point? When you have smaller datasets, of course the time doesn't matter. When we are talking about datasets that can take minutes, hours, or even days to compute, you want something that not only does the job, but one that does it in a timely manner.

The process of trying to find out the true data-generating process sounds a lot like reverse engineering. One has the final output and must backtrack to figure out what produced it. Both AIC and BIC have their purpose. This is a great concept when you are decided which of the two to use. Do you need to know the best approximate model, or do you need to know the data-generating process? The tool you pick depends on the job at hand. The bottom line is that in most situations BIC is the model to use, if you have small amounts of data but if you get larger sets, then AIC is the one to choose.

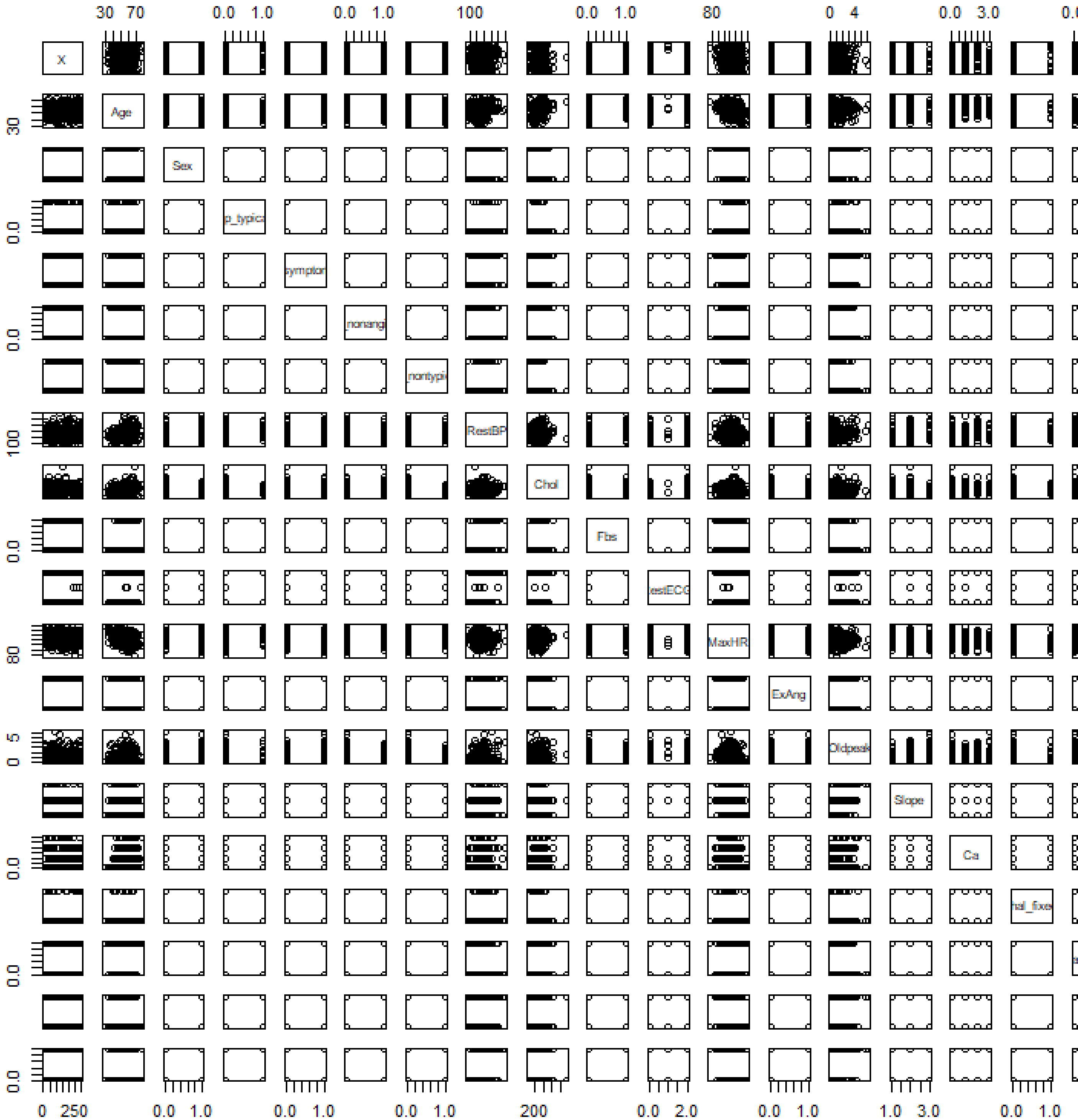
The second article was a trickier because it had a lot to digest. Most of it I didn't understand, and I had to go back and read it again. It did a lot of proving equivalence between different formulas and equations. I have to take it as given because I don't know enough to disprove the theories. I have always been told that cross-validation is a good technique because it limits the biases.

I also decided to code and do an analysis of my own. The dataset that I looked was in the Introduction to Statistical Learning with r (ISLR) package and it is heart disease. The goal was to predict if someone had Atherosclerotic Heart Disease (AHD). I did a hard code as indicator variables for the original categorical variables. I dropped the first column because it was just the case number, 1-303. Cp_ is the chest pain with the type of chest pain as an indicator variable.

RestBP= resting blood pressure, chol = cholesterol measurement, Fbs = fasting blood glucose, MaxHR = maximum heart rate, Thal_ is Thallium stress test, with indicator variables for the type. The model before any improvement is below

```
lm(formula = AHD ~ Age + Sex + cp_typical + cp_asymptomatic +
    cp_nonanginal + cp_nontypical + RestBP + Chol + Fbs + RestECG +
    MaxHR + ExAng + Oldpeak + Slope + Ca + Thal_fixed + Thal_normal +
    Thal_reversable)

Residuals:
     Min       1Q   Median       3Q      Max
-0.94392 -0.22064 -0.03192  0.16968  0.98703

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.2253135  0.3813955   0.591  0.55516
Age             -0.0005914  0.0027128  -0.218  0.82758
Sex              0.1563337  0.0489936   3.191  0.00158 **
cp_typical      -0.0935401  0.0912863  -1.025  0.30639
cp_asymptomatic  0.1803204  0.0660498   2.730  0.00673 **
cp_nonanginal   -0.0628984  0.0636698  -0.988  0.32406
cp_nontypical    0.0833978  0.0560044   1.489  0.13758
RestBP           0.0019420  0.0012474   1.557  0.12064
Chol             0.0003809  0.0004150   0.918  0.35945
Fbs             -0.0597135  0.0593020  -1.007  0.31483
RestECG          0.0255107  0.0210155   1.214  0.22581
MaxHR           -0.0023506  0.0011143  -2.109  0.03579 *
ExAng            0.1089087  0.0509239   2.139  0.03333 *
Oldpeak          0.0350288  0.0231243   1.515  0.13095
Slope            0.0677117  0.0423170   1.600  0.11070

Ca               0.1362083  0.0249158   5.467 1.01e-07 ***
Thal_fixed      -0.2088678  0.2617825  -0.798  0.42562
Thal_normal     -0.2803070  0.2487226  -1.127  0.26071
Thal_reversable -0.0592977  0.2508420  -0.236  0.81330
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3447 on 280 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.5524,	Adjusted R-squared:  0.5236
F-statistic:  19.2 on 18 and 280 DF,  p-value: < 2.2e-16
```

And the pairs correlation plot is below



I noticed some non-linear relationships, so I did a power transformation to fix it, here are the lamdas

Estimated transformation parameters

```
         Age          Sex     cp_typical cp_asymptomatic   cp_nonanginal   cp_nontypical        R
estBP
   1.190235775    0.494834792   -2.816744364   -0.056589193   -0.650041950   -1.33195284
3   -0.517623372
        Chol          Fbs       RestECG        MaxHR        ExAng       Oldpeak        Slope
   0.010841820   -1.460692221    0.009378732    2.454337568   -0.483809071    0.224381709
-0.358223860
          Ca     Thal_fixed    Thal_normal Thal_reversable          AHD
  -0.218157590   -3.599306509    0.126809431   -0.310100663   -0.100361945
```

Since some were close to zero, I just did a log transformation of them and them put them back into the model.

```
lm(formula = tAHD ~ tAge + tSex + tcp_typical + tcp_asymptomatic +
    tcp_nonanginal + tcp_nontypical + tRestBP + tChol + tFbs +
    tRestECG + tMaxHR + tExAng + tOldpeak + tSlope + tCa + tThal_fixed +
    tThal_normal + tThal_reversable)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60914 -0.11068  0.01323  0.12548  0.63116

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.661e-01  4.294e-01   1.319  0.18839
tAge              4.854e-04  6.321e-04   0.768  0.44316
tSex             -1.012e-01  3.128e-02  -3.237  0.00135 **
tcp_typical      -1.222e-07  1.239e-07  -0.986  0.32481
tcp_asymptomatic -2.443e-02  8.293e-03  -2.946  0.00349 **
tcp_nonanginal   -2.029e-03  1.956e-03  -1.038  0.30034
tcp_nontypical    9.090e-05  7.029e-05   1.293  0.19701
tRestBP           4.941e+00  2.340e+00   2.112  0.03560 *
tChol            -6.499e-02  6.078e-02  -1.069  0.28590
tFbs             -3.969e-05  4.086e-05  -0.971  0.33219
tRestECG         -5.194e-03  4.587e-03  -1.132  0.25845
tMaxHR            3.385e-07  1.948e-07   1.738  0.08337 .
tExAng            6.233e-03  3.561e-03   1.751  0.08110 .
tOldpeak         -2.887e-02  4.310e-02  -0.670  0.50342
tSlope            3.375e-01  1.274e-01   2.649  0.00852 **
tCa               9.928e-02  1.495e-02   6.643  1.6e-10 ***
tThal_fixed      -8.966e-09  9.600e-09  -0.934  0.35114
tThal_normal      4.020e-01  3.251e-01   1.236  0.21733
tThal_reversable -1.652e-02  4.582e-02  -0.360  0.71877
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.1996 on 280 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.5665,	Adjusted R-squared:  0.5386
F-statistic: 20.33 on 18 and 280 DF,  p-value: < 2.2e-16
```

It is somewhat improved. I did models for each variable and ended up with 18 equations. I extracted AIC, corrected AIC, and BIC for each one, but I will show the first three

```
> #Subset size=1
> n <- length(om1$residuals)
> npar <- length(om1$coefficients) +1
> #Calculate AIC
> extractAIC(om1,k=2)
[1]    2.0000 -754.5967
> #Calculate AICc
> extractAIC(om1,k=2)+2*npar*(npar+1)/(n-npar-1)
[1]    2.080268 -754.516427
> #Calculate BIC
> extractAIC(om1,k=log(n))
[1]    2.0000 -747.1692
>
> #Subset size=2
> npar <- length(om2$coefficients) +1
> #Calculate AIC
> extractAIC(om2,k=2)
[1]    3.0000 -782.7554
> #Calculate AICc
> extractAIC(om2,k=2)+2*npar*(npar+1)/(n-npar-1)
[1]    3.134228 -782.621211
> #Calculate BIC
> extractAIC(om2,k=log(n))
[1]    3.0000 -771.6142
>
> #Subset size=3
> npar <- length(om3$coefficients) +1
> #Calculate AIC
> extractAIC(om3,k=2)
[1]    4.0000 -786.6089
> #Calculate AICc
> extractAIC(om3,k=2)+2*npar*(npar+1)/(n-npar-1)
[1]    4.20202 -786.40690
> #Calculate BIC
> extractAIC(om3,k=log(n))
[1]    4.000 -771.754
```

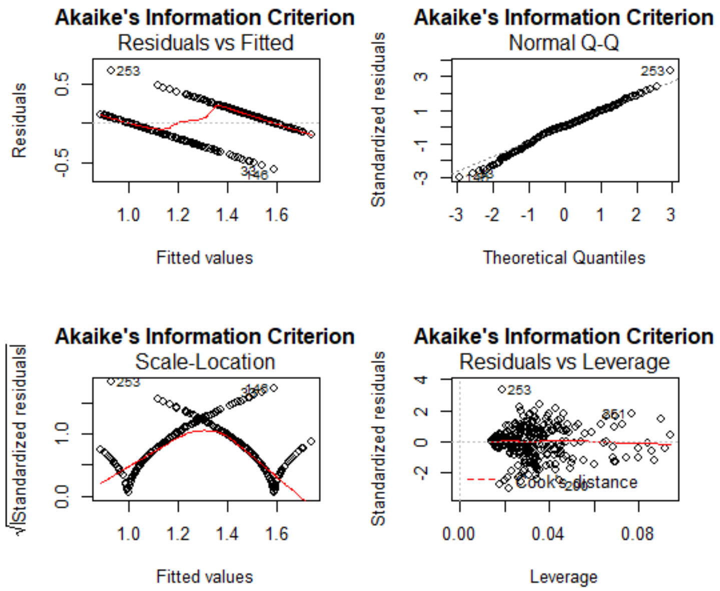
When doing backwards selection for AIC this was the final model

```
Step:  AIC=-955.45
tAHD ~ tSex + tcp_asymptomatic + tRestBP + tMaxHR + tExAng +
    tSlope + tCa + tThal_fixed + tThal_normal

                   Df Sum of Sq    RSS     AIC
<none>                          11.451 -955.45
- tMaxHR            1   0.10328 11.554 -954.77
- tThal_fixed       1   0.13634 11.587 -953.92
- tExAng            1   0.14972 11.600 -953.57
- tRestBP           1   0.18890 11.640 -952.56
- tSex              1   0.40443 11.855 -947.08
- tSlope            1   0.47341 11.924 -945.34
- tThal_normal      1   0.77492 12.226 -937.88
- tcp_asymptomatic  1   1.11673 12.567 -929.63
- tCa               1   1.93685 13.388 -910.73
```

Backwards BIC this was the final model

```
Step:  AIC=-925.35
tAHD ~ tSex + tcp_asymptomatic + tExAng + tSlope + tCa + tThal_normal

                   Df Sum of Sq    RSS     AIC
<none>                          11.845 -925.35
- tExAng            1   0.23455 12.079 -925.20
- tSex              1   0.30142 12.146 -923.55
- tSlope            1   0.66347 12.508 -914.77
- tThal_normal      1   0.84239 12.687 -910.52
- tcp_asymptomatic  1   1.21506 13.060 -901.86
- tCa               1   2.22840 14.073 -879.52
```

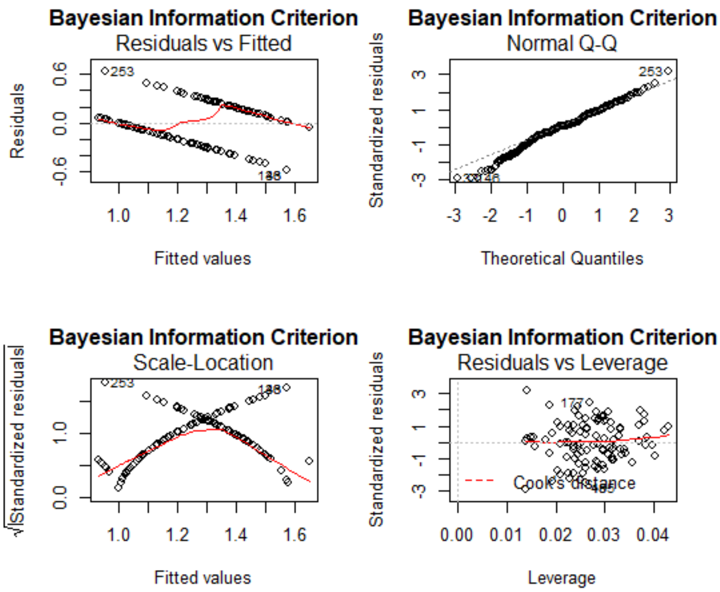
From above one can see that the BIC model does not include tRestBP, tMaxHR, tExANG, tSlope, or tThal_fixed. AIC is more robust, but BIC is a little easier to interpret. I plotted the two and did an output of their summary.



The variance isn't constant, and the scale-location looks quadradic in nature, but data looks to be normally distributed.

```
lm(formula = tAHD ~ tSex + tcp_asymptomatic + tRestBP + tMaxHR +
    tExAng + tSlope + tCa + tThal_fixed + tThal_normal)

Residuals:
     Min       1Q   Median       3Q      Max
-0.58700 -0.10682  0.00757  0.12755  0.65812

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.490e-01  1.865e-01   0.799 0.425100
tSex             -9.613e-02  3.009e-02  -3.195 0.001554 **
tcp_asymptomatic -3.153e-02  5.940e-03  -5.309 2.20e-07 ***
tRestBP           4.729e+00  2.166e+00   2.184 0.029803 *
tMaxHR            2.888e-07  1.788e-07   1.615 0.107505
tExAng            6.850e-03  3.524e-03   1.944 0.052882 .
tSlope            3.810e-01  1.102e-01   3.457 0.000629 ***
tCa               9.700e-02  1.387e-02   6.992 1.89e-11 ***
tThal_fixed      -5.992e-09  3.230e-09  -1.855 0.064611 .
tThal_normal      2.888e-01  6.531e-02   4.422 1.38e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1991 on 289 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.5551,	Adjusted R-squared:  0.5413
F-statistic: 40.07 on 9 and 289 DF,  p-value: < 2.2e-16
```

Looking at this summary tMaxHR is insignificant and both tExAng and tThal_fixed are only suggestively significant. Still an improvement on the original model. Next up is BIC

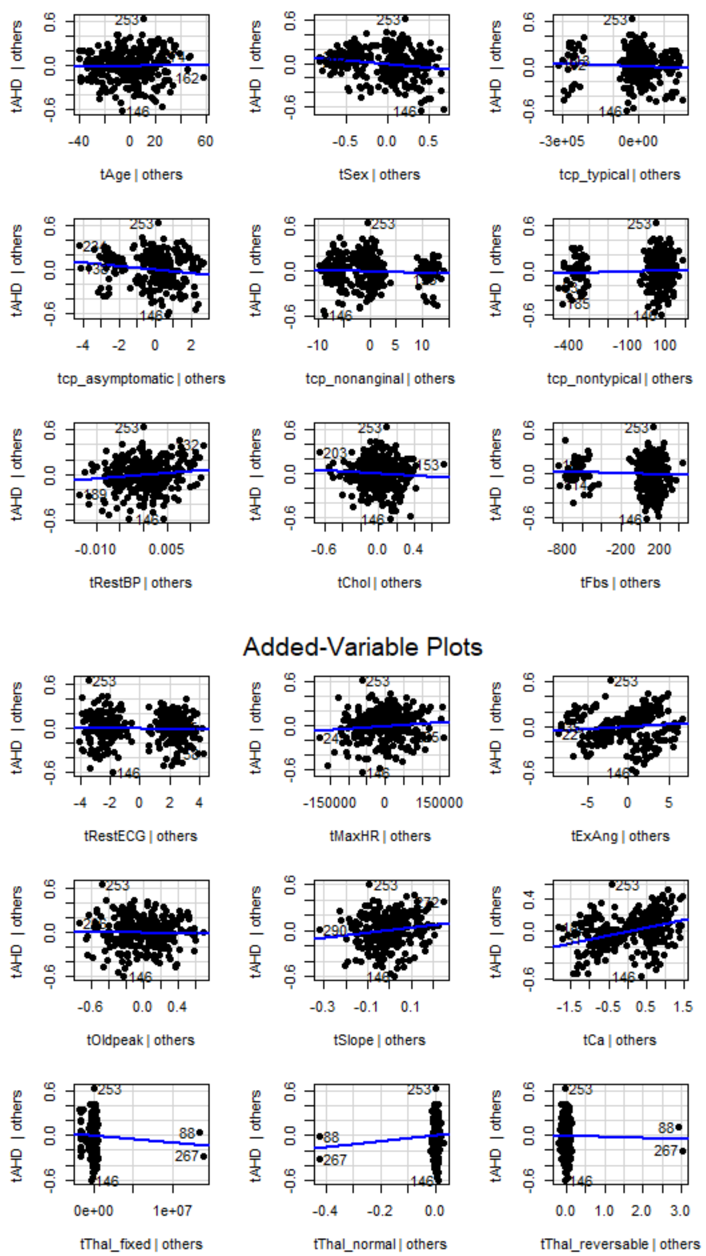


This model looks to have the same exact problems as before when it was plotted. The summary however looks a whole lot better

```
lm(formula = tAHD ~ tSex + tcp_asymptomatic + tExAng + tSlope +
    tCa + tThal_normal)

Residuals:
     Min       1Q   Median       3Q      Max
-0.57468 -0.08731  0.01386  0.13711  0.63311

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.432348   0.095792   4.513 9.26e-06 ***
tSex             -0.082042   0.030097  -2.726   0.0068 **
tcp_asymptomatic -0.032452   0.005929  -5.473 9.53e-08 ***
tExAng            0.008402   0.003494   2.405   0.0168 *
tSlope            0.425272   0.105153   4.044 6.72e-05 ***
tCa               0.101973   0.013758   7.412 1.35e-12 ***
tThal_normal      0.292229   0.064126   4.557 7.63e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2014 on 292 degrees of freedom
  (4 observations deleted due to missingness)
```

Multiple R-squared: 0.5398,	Adjusted R-squared: 0.5304
F-statistic: 57.09 on 6 and 292 DF, p-value: < 2.2e-16

Every predictor is significant to almost the 1% significance level. This seems to be the preferred model for predicting AHD. Overall while the AIC model had a better adjusted R^2 that the BIC model, the BIC model outperformed AIC when it came to variable significance. I would use this model to do model selection (I also did an added-variable plot to see which ones added anything extra to the model and the BIC aligned pretty well with this also).



References:

Acquah, Henry de-Graft. “Comparison of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) in Selection of an Asymmetric Price Relationship.” *Journal of Development and Agricultural Economics,* vol. 2, no. 1, Jan. 2010, pp. 001–006.

Stone, M. “An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion.” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, 1 Jan. 1977, pp. 44–47.