Linear Regression Models

AIC

The Akaike information criterion (AIC) is a measure of the relative quality of a statistical model, for a given set of data. As such, AIC provides a means for model selection. Given a set of candidate models for the data, the preferred model is the ***one with the minimum AIC value***.

AIC does not provide a test of a model in the sense of testing a null hypothesis; i.e. AIC can tell nothing about the quality of the model in an absolute sense. If all the candidate models fit poorly, AIC will not give any indication of that.

Question

Load the school absenteeism data set and fit a linear model relating the log of the number of days absent to the other variables with the commands:

```
require(MASS)
data(quine)
lm1 = lm(log(Days + 2.5) ~ . , data=quine)
```

Use the `step()` function in `R` to perform model selection using default parameters. What variables remain in the model after model selection?

```
step(lm1)
```

In the first phase, Removing sex as a predictor variable results in the largest decrease in AIC value. Removing age and Ethnicity actually increas the AIC value, which is undesirable. Hence we have the final model.

```
> step(lm1)
Start:  AIC=-54.92
log(Days + 2.5) ~ Eth + Sex + Age + Lrn

       Df Sum of Sq     RSS     AIC
- Sex   1    0.4379  91.502 -56.218
- Lrn   1    0.6529  91.717 -55.875
<none>               91.064 -54.918
- Age   3    4.4012  95.465 -54.027
- Eth   1   10.1349 101.199 -41.512
```

In the second phase, sex has been removed as a predictor variable results. Now removing learner variable results in a decrease of AIC. Removing age and Ethnicity actually increas the AIC value, which is undesirable.

```
Step:  AIC=-56.22
log(Days + 2.5) ~ Eth + Age + Lrn

       Df Sum of Sq     RSS     AIC
- Lrn   1    0.5230  92.025 -57.386
<none>               91.502 -56.218
- Age   3    4.6031  96.105 -55.052
- Eth   1   10.0639 101.566 -42.983
```

In the third phase: Removing age and Ethnicity actually increas the AIC value, which is undesirable. The best approach is not to remove any variables.

```
Step:  AIC=-57.39
log(Days + 2.5) ~ Eth + Age

       Df Sum of Sq     RSS     AIC
<none>               92.025 -57.386
- Age   3    4.0801  96.105 -57.052
- Eth   1   10.0152 102.040 -44.303

Call:
```

```
lm(formula = log(Days + 2.5) ~ Eth + Age, data = quine)

Coefficients:
(Intercept)          EthN         AgeF1         AgeF2
     2.8204      -0.5254      -0.1648       0.2025
      AgeF3
     0.2215
```