L 4: Model comparison and selection[1]

With the `cars` data consider testing the null model

$$\mathtt{distance}_i = \beta\mathtt{speed}_i^2 + \epsilon_i$$

against the full model.
We can use the ‘anova‘ function to perform the appropriate F-test:

```
> b <- lm(dist˜speed+I(speedˆ2),data=cars)
> b0 <-lm(dist˜I(speedˆ2) - 1,data=cars)
> anova(b,b0)
Analysis of Variance Table

Model 1: dist ˜ speed + I(speedˆ2)
Model 2: dist ˜ I(speedˆ2) - 1
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     47 10825
2     49 11936 -2   -1111.2 2.4123 0.1006
```

So some evidence in favour of the null model

[1] Faraway, J.J. Linear models with R. CRC Press. Chapter 10.

anova() with single argument vs drop1()

```
> anova(b)
Analysis of Variance Table

Response: dist
           Df  Sum Sq Mean Sq  F value    Pr(>F)
speed       1 1609.18 1609.18 114.0024 3.704e-14 ***
I(speed^2)  1   34.44   34.44   2.4402     0.125
Residuals  47  663.42   14.12
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> drop1(b, test="F")
Single term deletions

Model:
dist ~ speed + I(speed^2)
           Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>                  663.42 135.27
speed       1    12.717 676.14 134.22  0.9009 0.3474
I(speed^2)  1    34.445 697.86 135.80  2.4402 0.1250
>
```

`anova` with single argument: Table based on the sequence of ever simpler models obtained by removing terms sequentially from the model. Each row in the model tests one model in the sequence against the closest more complicated model in the sequence. Not necessarily helpful in presence of confounding.

`drop1`: Table obtained by F-ratio test comparison of the full model with each of the models produced by dropping a single effect from the full model.

AIC

Hypothesis testing: We are asking the question ”what is the simplest model that is defensible for these data” since hypothesis testing always sticks with the simplest model unless the data provide strong evidence that it is less adequate than the alternative.

Akaike Information Criterion (AIC): Want to find the model that will be as ”close” as possible to the truth, in terms of smallest error when the model is used for prediction.

If l is the maximized log likelihood of the model and p the number of parameters then

$$AIC = -2l + 2p.$$

Attempt to estimate a particular measure of the ”distance” between the fitted model, and the underlying ”true model” that generated the data.

The ‘AIC‘ function will compute the AIC of a fitted model object in R. e.g.

```
> AIC(b0, b)
   df      AIC
b0  2 414.8026
b   4 412.2635
```

Which is the better model?

Models with lower AIC are considered to be better than models with higher AIC

Model selection strategies: backward selection

Situation: We have a large number of possible prediction terms in a model, automatic model selection methods are often used to try and sort through the model space of possible models to find one that is ”best” in some sense.

Backward selection is often used. It starts with the largest possible model and consists of repeatedly deleting the model term with the highest p-value (as reported by ‘drop1‘) and refitting, until all p-values are below some threshold.

Example: Swiss data

Standardised fertility measure and socio economic indicators for 47 french-speaking provinces of Switzerland at 1888.

```
> summary(swiss)
   Fertility      Agriculture     Examination      Education
 Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00
 1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00
 Median :70.40   Median :54.10   Median :16.00   Median : 8.00
 Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98
 3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00
 Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00
    Catholic       Infant.Mortality
 Min.   :  2.150   Min.   :10.80
 1st Qu.:  5.195   1st Qu.:18.15
 Median : 15.140   Median :20.00
 Mean   : 41.144   Mean   :19.94
 3rd Qu.: 93.125   3rd Qu.:21.70
 Max.   :100.000   Max.   :26.60
```

Example: Swiss data

```
> summary(lm1 <- lm(Fertility ˜ ., data = swiss))

Call:
lm(formula = Fertility ˜ ., data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.2743  -5.2617   0.5032   4.1198  15.3213 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *  
Examination      -0.25801    0.25388  -1.016  0.31546    
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 ** 
Infant.Mortality  1.07705    0.38172   2.822  0.00734 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,Adjusted R-squared:  0.671 
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

Backward selection using drop1()

```
> drop1(lm1,test="F")
Single term deletions

Model:
Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality
                 Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                        2105.0 190.69
Agriculture       1    307.72 2412.8 195.10  5.9934  0.018727 *
Examination       1     53.03 2158.1 189.86  1.0328  0.315462
Education         1   1162.56 3267.6 209.36 22.6432 2.431e-05 ***
Catholic          1    447.71 2552.8 197.75  8.7200  0.005190 **
Infant.Mortality  1    408.75 2513.8 197.03  7.9612  0.007336 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> drop1(lm(Fertility ~ .- Examination, data = swiss),test="F")
Single term deletions

Model:
Fertility ~ (Agriculture + Examination + Education + Catholic +
    Infant.Mortality) - Examination
                 Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                        2158.1 189.86
Agriculture       1    264.18 2422.2 193.29  5.1413   0.02857 *
Education         1   2249.97 4408.0 221.43 43.7886 5.140e-08 ***
Catholic          1    956.57 3114.6 205.10 18.6165 9.503e-05 ***
Infant.Mortality  1    409.81 2567.9 196.03  7.9757   0.00722 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

step()

Function step() does automatic backward selection, based on AIC.

```
> slm1 <- step(lm1)
Start:  AIC=190.69
Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality

                   Df Sum of Sq    RSS    AIC
- Examination       1     53.03 2158.1 189.86
<none>                          2105.0 190.69
- Agriculture       1    307.72 2412.8 195.10
- Infant.Mortality  1    408.75 2513.8 197.03
- Catholic          1    447.71 2552.8 197.75
- Education         1   1162.56 3267.6 209.36

Step:  AIC=189.86
Fertility ~ Agriculture + Education + Catholic + Infant.Mortality

                   Df Sum of Sq    RSS    AIC
<none>                          2158.1 189.86
- Agriculture       1    264.18 2422.2 193.29
- Infant.Mortality  1    409.81 2567.9 196.03
- Catholic          1    956.57 3114.6 205.10
- Education         1   2249.97 4408.0 221.43
```

End up with same model as from backward selection using p-values. But this is not always the case.

Purpose of model selection

Model selection for prediction: Linear models are used for two main purposes: prediction and explanation. If you want to build a model for predicting future response values, model selection can be helpful.

- Collecting data can be expensive so we would like to drop any redundant predictors that do not help the accuracy of prediction.
- Unnecessary predictors (or combinations of predictors) add noise to the estimation of the parameters and can make the prediction worse.
- Depending on the application it may be preferable to have a simple model with fewer predictors.

Purpose of model selection

Model selection for explanation: Can we use the automatic model selection procedure (like step() or backward selection using p-values) to determine which predictors explain the response and which do not?

Purpose of model selection

Model selection for explanation: Can we use the automatic model selection procedure (like step() or backward selection using p-values) to determine which predictors explain the response and which do not?

Ans: Not really.

- Excluded predictors may still have relationship to response.
- It is preferable to formulate a small number of questions as hypotheses and test them (e.g. assess the effect of one predictor conditional on a small set of covariates).
- It is a bad idea to choose a model based on a large number of hypothesis tests. If you start doing lots of tests, the chance that you reject a null that is really true will be greater than the assumed significance level. Criterion based methods preferable.
- You need to interpret the parameter estimates, so more careful model selection is needed.
- Association can be mistaken for causation in the case of observational data.

Cross Validation

For prediction, when you only want to test how well your predictors predict, *k*-fold cross validation can be useful:

- Randomly split the data into k groups.
    - Take each group as the test set in turn, using the remaining $k - 1$ groups as the training set.
    - Fit to the training set, use the fitted model to predict the test set and obtain the prediction errors (observed minus fitted).
    - Compute a summary of the prediction errors, e.g. the mean squared-error.
- The mean of the summary statistic across the k folds can then be used to compare different models

Warning - you may need to think carefully about how to split the data, e.g. if ordered in time.

CV in R

Use cv package with function cv(). Default is mean squared-error and 10-fold cross validation.

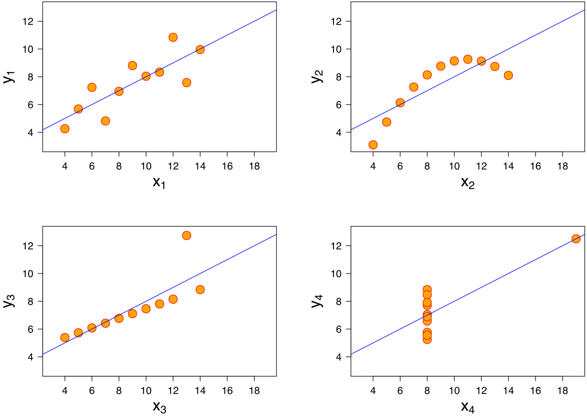
```
>b0 <- lm(dist~I(speed^2)-1,data=cars)
>b1<- lm(dist~speed+I(speed^2)-1, data=cars)
>summary(cv(b0))

criterion: mse
cross-validation criterion = 251.8584
bias-adjusted cross-validation criterion = 251.1677
full-sample criterion = 238.7174
>summary(cv(b1))

criterion: mse
cross-validation criterion = 242.8862
bias-adjusted cross-validation criterion = 241.4613
full-sample criterion = 216.6223
```

Don’t be afraid of subjectivity

- There is usually not an obvious best model.
- May be several models which fit similarly.
  - Do the models have similar qualitative consequences?
  - Do they make similar predictions?
  - What cost/ effort is needed to measure the predictors?
  - Is there a difference in model diagnostics?
  - Are there any other (non-statistical!) reasons to prefer one model over another?
- If models which seem comparable and which fit well give different answers to your question, then it may be that the question cannot be answered with the dataset.

Other issues 1 (6.6 Weisberg)

Important to look at the data (plots/ graphics/ summary statistics). Tests can miss things.



From Anscombe (1973)

Other issues 2 (6.6 Weisberg)

- Relevance of data. Sample vs population (more later). No point coming up with a detailed model if the inference is swamped by other uncertainties.
- Using the data twice. Defining variables and then testing.
- Multiple testing
- The lab is not the real world. How realistic are the datasets?

Summary

- The aim is to construct a model that predicts well or explains the relationships in the data.
- Automatic variable selections are not guaranteed to be consistent with these goals.
- Use these methods as a guide only.
- Try out model selection with the example data.