0.1 Regression

The basic form of a formula is "response $\sim$ model".

0.2 Analysis of Variance

Analysis of variance (ANOVA) is a collection of statistical models, and their associated procedures, in which the observed variance in a particular variable is partitioned into components attributable to different sources of variation.

0.2.1 Multiple Linear Regression

The basic model for multiple regression analysis is

$$y = b_0 + b_1 x_1 + \cdots + b_k x_k + e$$

0.2.2 Example: MTCars

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (197374 models).

A data frame with 32 observations on 11 variables. Let us assume that the variable mpg is the response variable, with the other ten being predictor variables.

- **mpg** Miles/(US) gallon
- **cyl** Number of cylinders
- **disp** Displacement (cu.in.)
- **hp** Gross horsepower
- **drat** Rear axle ratio
- **wt** Weight (lb/1000)
- **qsec** 1/4 mile time
- **vs** V/S
- **am** Transmission (0 = automatic, 1 = manual)
- **gear** Number of forward gears
- **carb** Number of carburetors

0.2.3 Model specification and output

Specification of a multiple regression analysis is done by setting up a model formula with + between the explanatory variables:

```
lm(mpg~cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb, data=mtcars)
```

which is meant to be read as "mpg is described using a model that is additive in cyl, disp, and so forth. The output is as follows:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared: 0.869,      Adjusted R-squared: 0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Notice that none of the predictor variables is significant. The only one that comes even close is "wt".

```
       mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
```

```
am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

In many cases there is a high degree of correlation between two predictor variables. The variable "disp" has correlation coefficients of -0.85, 0.79 and 0.89 and with "cyl", "hp" and "wt" respectively.

0.2.4 Multicollinearity

0.2.5 Variable Selection Procedures

There are three types of variable selection procedure.

- Forward Selection
- Backward Elimination
- Stepwise selection

The `R` command we use to perform variable selection procedures is `step()`

direction - the mode of stepwise search, can be one of "both", "backward", or "forward", with a default of "both". If the scope argument is missing the default for direction is "backward".

0.2.6 Coefficient of Determination

```
Analysis of Variance Table

Response: mpg
          Df Sum Sq Mean Sq  F value    Pr(>F)
cyl        1 817.71  817.71 116.4245 5.034e-10 ***
disp       1  37.59   37.59   5.3526  0.030911 *
hp         1   9.37    9.37   1.3342  0.261031
drat       1  16.47   16.47   2.3446  0.140644
wt         1  77.48   77.48  11.0309  0.003244 **
qsec       1   3.95    3.95   0.5623  0.461656
vs         1   0.13    0.13   0.0185  0.893173
am         1  14.47   14.47   2.0608  0.165858
gear       1   0.97    0.97   0.1384  0.713653
carb       1   0.41    0.41   0.0579  0.812179
Residuals 21 147.49    7.02
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

0.2.7 Backward Elimination

Our initial model includes all the predictor variables.

```
> step(fit.all)
Start:  AIC=70.9
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq    RSS    AIC
- cyl   1    0.0799 147.57 68.915
- vs    1    0.1601 147.66 68.932
- carb  1    0.4067 147.90 68.986
- gear  1    1.3531 148.85 69.190
- drat  1    1.6270 149.12 69.249
- disp  1    3.9167 151.41 69.736
- hp    1    6.8399 154.33 70.348
- qsec  1    8.8641 156.36 70.765
<none>              147.49 70.898
- am    1   10.5467 158.04 71.108
- wt    1   27.0144 174.51 74.280
```

This tables tells us the effect or removing each predictor variable individually, in terms of the AIC. Consider the first row. This tells us the AIC value of a model fitted without the "cyl" variable would be 68.915. Included in the table is effect of not removing any variables. If the "wt" variable was to be removed, the AIC value would increase to 74.280.

```
..
- cyl   1    0.0799 147.57 68.915
..
<none>              147.49 70.898
..
- wt    1   27.0144 174.51 74.280
```

The procedure removes variables as appropriate, until it found that removing anymore variables would increase the AIC.

```
Step:  AIC=61.31
mpg ~ wt + qsec + am

       Df Sum of Sq    RSS    AIC
<none>              169.29 61.307
```

```
- am    1    26.178 195.46 63.908
- qsec  1   109.034 278.32 75.217
- wt    1   183.347 352.63 82.790
```

The outcome of this procedure is that "mpg" is best explained as a linear combination of the "am", "qsec" and "wt" variables.

```
Coefficients:
(Intercept)           wt         qsec           am
      9.618       -3.917        1.226        2.936
```