Multiple linear regression

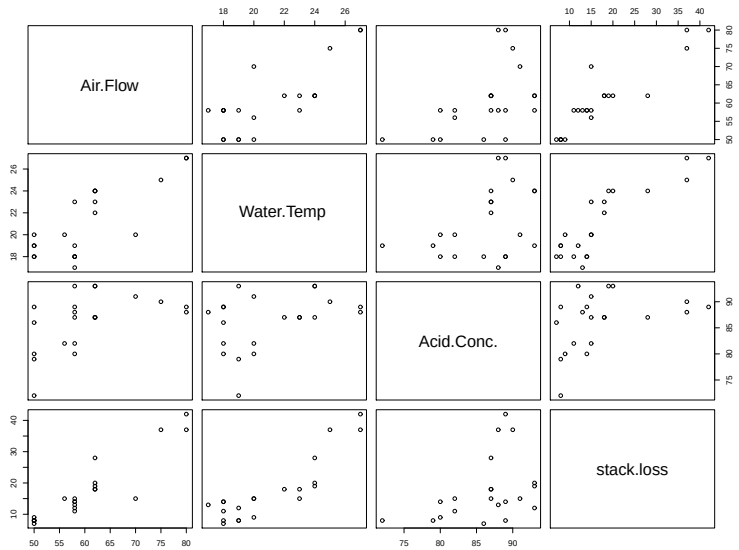
(STATS544.2: Applied multiple linear regression)

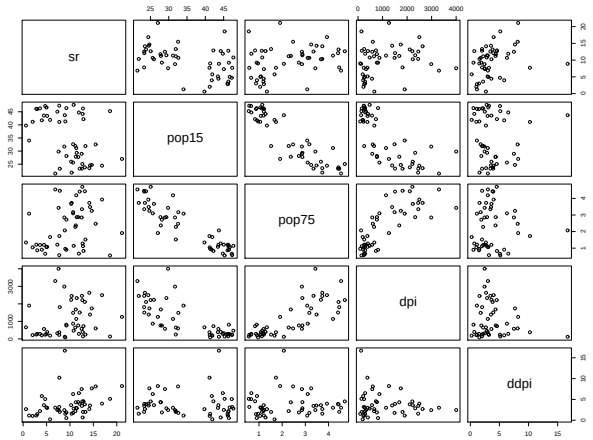
Gunnar Stefansson

August 8, 2013

Datasets

We will mostly be working with two datasets in the lecture, **stackloss** and **LifeCycleSavings**, both part of the **datasets** package that comes with your installation of R.





Point estimation (with R)

Example (chemistry)
We get the data by writing:

```
data(stackloss)
```

We look at the data using a pairs plot:

```
pairs(stackloss)
```

We fit a model with **stack.flow** as the dependent variable and the others as independent variables using:

```
fit.stack <- lm(stack.loss~Air.Flow + Water.Temp + Acid.Conc.
```

Writing:

```
summary(fit.stack)
```

returns the the point estimates of the parameters among other things

Interpreting coefficients

```
> data(LifeCycleSavings)
> fit.life <- lm(sr ~ pop15 + pop75 + dpi + ddpi ,data=LifeCycleSavings)
> summary(fit.life)

Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2422 -2.6857 -0.2488  2.4280  9.7509

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared: 0.3385, Adjusted R-squared: 0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

R^2

The summary command gives the R^2 value.

It can also be extracted directly, for use in computations.

As usual, this is $1 - SSE/SSTOT$.

It is also the squared correlation between y and $\hat{y}$.

Example (chemistry)

We continue working with the **stackloss** data.

We can get the R^2 along with other things writing:

```
summary(fit.stack)
```

The R^2 can be extracted with

```
summary(fit.stack)$r.squared
```

Example (economics)

We continue working with the **LifeCycleSavings data.**

We can get the R^2 along with other things writing:

```
summary(fit.life)
```

The R^2 can be extracted with

```
summary(fit.life)$r.squared
```

P-values

Example (chemistry)
We continue working with the **stackloss** data.
By writing

```
summary(fit.stack)
```

we get some statistics and informations about out model. In the last column of the Coefficients table, marked $Pr(>|t|)$ we get the p-values for the tests of the individual coefficients.
On the last line of the output we get the p-value for the overal test that all the coefficients of the model are equal to zero.p
In general we reject the null hypothesis that the coefficient(s) is equal to zero if the p-value is smaller than the α-value used.
Example (economics)
We continue working with the **LifeCycleSavings data.**
By writing

```
summary(fit.life)
```

we get some statistics and informations about out model. In the last column of the Coefficients table, marked $Pr(>|t|)$ we get the p-values for the tests of the individual coefficients.
On the last line of the output we get the p-value for the overal test that all the coefficients of the model are equal to zero.
In general we reject the null hypothesis that the coefficient(s) is equal to zero if the p-value is smaller than the α-value used.

Building a model

We can build models by

- specifying all the terms and fitting a single model
- starting from a null model, adding one variable at a time
- starting from a full model with all available variables and dropping one at a time

Stepwise linear regression

Can formalise the model building process
Forward stepwise regression
Backwards stepwise regression
Need to choose a criterion: P-value vs AIC etc

Thou shalt not seek a model with a shotgun (the multiplicity issue)

It is quite common that very many potential descriptors exist

Testing nonsense will eventually yield a significant result

Example (biology): Consider again the Icelandic ecosystem data. Suppose we want to search for significance among the variables in the data set but also have 100 more variables - all of which are noise.

```
b<-read.table("http://tgax14.rhi.hi.is/html/data/biol/borecol.txt",header=T)
n<-nrows(b)
bad100<-matrix(rnorm(n*100),nrow=n)
newb<-as.data.frame(cbind(b,bad100))
```

Now check, which variable is most highly correlated with the growth, in variable G.

Multiplicity corrections

Can correct for multiplicity: Bonferroni, Scheffe...
Simplest: Bonferroni
Better: Holm

Comparing nested models

If one model is a **submodel** of another, then an F-test is used to compare the models.

H_0 : submodel is correct

$$F = \frac{\frac{SSE(R)-SSE(F)}{df(R)-df(F)}}{\frac{SSE(F)}{df(F)}} \sim F_{df(R)-df(F),df(F)}, \text{ if } H_0 \text{ is true.}$$

The usual t-test for dropping one variable is a special case ($t^2 = F$).

Example: Check to see whether one or two straight lines are needed to explain a data set.

Comparing non-nested models

Try to make them nested in a supermodel

Try to avoid the comparison

Worst-case: Use the AIC or similar criterion

Example (biology): The case of natural mortality vs serial correlation in fish stock assessments (Myers et al).

Example (general): Variable intercept vs variable slope.