Multiple regression I

Suppose we have

- a single response variable y
- several predictor/explanatory variables $x_1, \ldots, x_p$

Data for multiple linear regression consist of the values of y and $x_1, \ldots, x_p$ for n individuals. We write the data in the form:

Individual	Predictors				Response
i	x_1	x_2	$\cdots$	x_p	Y
1	x_{11}	x_{21}	$\cdots$	x_{p1}	Y_1
2	x_{12}	x_{22}	$\cdots$	x_{p2}	Y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
n	x_{1n}	x_{2n}	$\cdots$	x_{pn}	Y_n

Multiple regression II

Data set on size of prostate tumor as function of several predictors.

```
##       lcavol lweight     lpsa age
## 1 -0.5798185  2.7695 -0.43078  50
## 2 -0.9942523  3.3196 -0.16252  58
## 3 -0.5108256  2.6912 -0.16252  74
## 4 -1.2039728  3.2828 -0.16252  58
## 5  0.7514161  3.4324  0.37156  62
## 6 -1.0498221  3.2288  0.76547  50
```

Multiple regression III

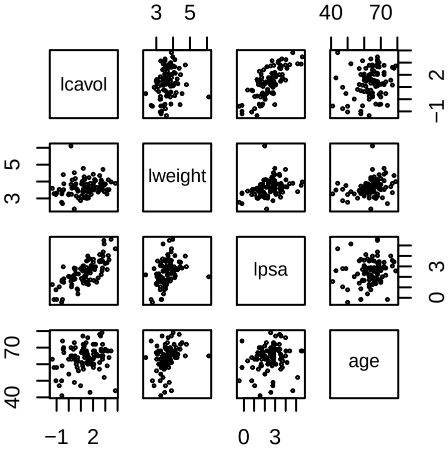
Plot variables against one another and compute correlations between variables to see what's going on...

```
cor(newp)
```

```
##            lcavol   lweight      lpsa       age
## lcavol  1.0000000 0.1941284 0.7344603 0.2249999
## lweight 0.1941284 1.0000000 0.3541218 0.3075247
## lpsa    0.7344603 0.3541218 1.0000000 0.1695929
## age     0.2249999 0.3075247 0.1695929 1.0000000
```

```
plot(newp,cex=.3)
```

Multiple regression IV



Multiple regression V

The multiple regression linear model posits the following relationship between Y and $x_1, \ldots, x_p$:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

where

- $\epsilon \sim N(0, \sigma)$ is a random variable
- The ϵ_i's corresponding to observations $(Y_i; X_{1i}, X_{2i}, \cdots, X_{pi})$ on different individuals are independent of each other
- β_j is the change in Y for each unit change in X_j *when holding all other predictors constant*

Estimating the Regression Parameters I

The true population parameters $\beta_0, \beta_1, \ldots, \beta_p$ and σ are estimated from the data by the least squares method. That is, we minimize the *residual sum of squares*

$$\begin{aligned}
\mathrm{SSE} &= \sum_{i=1}^{n} (r_i)^2 \\
&= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\
&= \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - \cdots - b_p X_{pi})^2
\end{aligned}$$

Estimating the Regression Parameters II

The estimator of σ^2 is

$$S^2 = \frac{\mathsf{SSE}}{n-p-1} = \frac{\sum(r_i)^2}{n-p-1}$$

where $n-p-1$ is the number of degrees of freedom. (Why $n-p-1$?)

Estimating the Regression Parameters III

Normal equations:

Define $\bar{Y}, \bar{x}_j, j = 1, \ldots, p$ as the averages and

$\tilde{Y}_i = Y_i - \bar{Y}, \tilde{x}_{ji} = x_{ji} - \bar{x}_j$, as the centered variables.

For any two vectors define $v \cdot w = \sum_{i=1}^{n} v_i w_i$.

Then: $b_0 = \bar{Y} - \sum b_j \bar{x}_j$, and p-linear equations for the p unknowns: $b_1, \ldots, b_p$:

$$\tilde{Y} \cdot \tilde{x}_1 = b_1 \tilde{x}_1 \cdot \tilde{x}_1 + b_2 \tilde{x}_1 \cdot \tilde{x}_2 + \ldots + b_p \tilde{x}_1 \cdot \tilde{x}_p.$$

$$\tilde{Y} \cdot \tilde{x}_2 = b_1 \tilde{x}_2 \cdot \tilde{x}_1 + b_2 \tilde{x}_2 \cdot \tilde{x}_2 + \ldots + b_p \tilde{x}_2 \cdot \tilde{x}_p.$$

$$\vdots$$

$$\tilde{Y} \cdot \tilde{x}_p = b_1 \tilde{x}_p \cdot \tilde{x}_1 + b_2 \tilde{x}_p \cdot \tilde{x}_2 + \ldots + b_p \tilde{x}_p \cdot \tilde{x}_p.$$

Estimating the Regression Parameters IV

As with simple linear regression, we need to check that the model assumptions are met:

- The sample is a SRS from the population

  This can't be checked; this needs to be taken care of when the sample is drawn.

- There is a linear relationship in the population

  Checking this isn't as straightforward as with simple linear regression, but we should draw a plot of *residuals vs. fitted values* and check for any patterns.

- The standard deviation of the residuals is constant.

  Using the same plot as above, check for non-uniformity in the spread of residuals around the center line.

- The response varies Normally about the population regression line.

  Check with a *Normal quantile plot* of the residuals.

Estimating the Regression Parameters V

```
mod_full=lm(lcavol~lpsa+lweight+age,data=prostate)
summary(mod_full)


##
## Call:
## lm(formula = lcavol ~ lpsa + lweight + age, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.07322 -0.54594 -0.01828  0.56280  1.69501
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.91121    0.81073  -1.124   0.2639
## lpsa         0.76782    0.07526  10.202   <2e-16 ***
## lweight     -0.26773    0.18118  -1.478   0.1429
## age          0.02092    0.01147   1.824   0.0713 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7942 on 93 degrees of freedom
## Multiple R-squared:  0.5601,Adjusted R-squared:  0.546
## F-statistic: 39.48 on 3 and 93 DF,  p-value: < 2.2e-16
```

Inference for Regression Coefficients I

The estimates b_j are againm linear functions of the Y's and we can compute their variance.

A 95% confidence interval for β_j is

$$b_j \pm t^* \mathrm{SE}(b_j)$$

where t^* is the number such that 95% of the area of the t_{n-p-1} distribution falls between $-t^*$ and t^*
To test the hypothesis

$$H_0 : \beta_j = 0 \qquad (\beta_i \ \text{ arbitrary for } i \neq j)$$

compute the t-statistic

$$T = \frac{b_j}{\mathrm{SE}(b_j)}$$

Inference for Regression Coefficients II

- the p-value for this test statistic is computed from the t_{n-p-1} distribution
  $\rightarrow$ for $H_a : \beta_j > 0$, p-value is $P(t_{n-p-1} > T)$
  $\rightarrow$ for $H_a : \beta_j < 0$, p-value is $P(t_{n-p-1} < T)$
  $\rightarrow$ for $H_a : \beta_j \neq 0$, p-value is $2P(t_{n-p-1} > |T|)$
- if the regression model assumptions are true, testing $H_0 : \beta_j = 0$ corresponds to testing whether or not X_j is a significant predictor of Y, *assuming all the other predictors are already in the model*.

ANOVA table for multiple regression I

The basic ideas of the regression ANOVA table are the same in simple and multiple regression.

ANOVA expresses variation in the form of sums of squares. It breaks the total variation into two parts: SSR and SSE:

Source	SS	df
Regression (SSR)	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	p
Residual (SSE)	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$n - p - 1$
Total	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$n - 1$

$$\text{SST} = \text{SSR} + \text{SSE}$$

ANOVA table for multiple regression II

The statistic

$$R^2 = 1 - \frac{\mathsf{SSE}}{\mathsf{SST}} = \frac{\mathsf{SSR}}{\mathsf{SST}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

is the proportion of the variation of the response variable Y that is explained by the explanatory variables $X_1, X_2, \cdots, X_p$. R^2 is called the **multiple correlation coefficient**.

Forward selection; add variables one at a time I

```
options(digits=3)
mod_psa=lm(lcavol~lpsa,data=prostate);sll=summary(mod_psa)
c(sll$r.squared,sll$adj.r.squared,sll$sigma,sll$coefficients[2,4])
```

```
## [1] 5.39e-01 5.35e-01 8.04e-01 1.12e-17
```

```
mod_psa_weight=lm(lcavol~lpsa+lweight,data=prostate); sll=summary(mod_psa_weight)
c(sll$r.squared,sll$adj.r.squared,sll$sigma,sll$coefficients[3,4])
```

```
## [1] 0.544 0.535 0.804 0.314
```

```
sll=summary(mod_full)
c(sll$r.squared,sll$adj.r.squared,sll$sigma,sll$coefficients[4,4])
```

```
## [1] 0.5601 0.5460 0.7942 0.0713
```

Forward selection; add variables one at a time II

The R^2 increases with every additional predictor. This is a mathematical fact. But some predictors may not be particularly useful in the regression.

One possibility is to use Adjusted-R^2 :

$$R^2_{adj} = 1 - \frac{\mathrm{SSE}/(n-p-1)}{\mathrm{SST}/(n-1)} = 1 - \frac{S^2}{\mathrm{Var}(Y)}$$

Adjusted R^2 does not necessarily increase with more predictors.

The adjusted R^2 compares the estimated sigmas - the numerator in the fraction is S^2. The denominator is fixed to variance of $Y_i, i = 1, \ldots, n$. So if S is smaller a model is better.

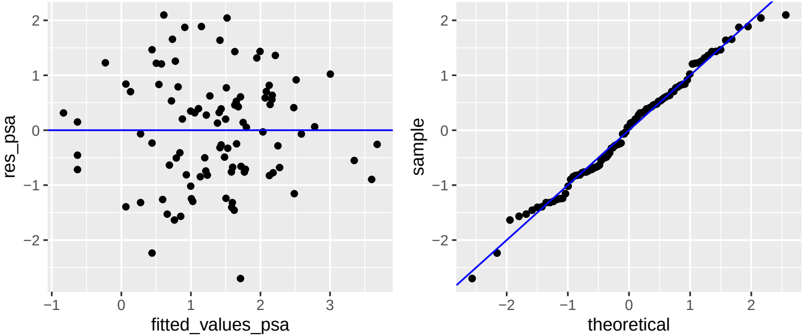
Forward selection; add variables one at a time III

But... this does not give us a sense of how much the decrease is really meaningful.
So ... you can use p-values of the new coefficient, to decide if to add it.
Or ... use analysis of variance F-tests.

Forward selection; add variables one at a time IV

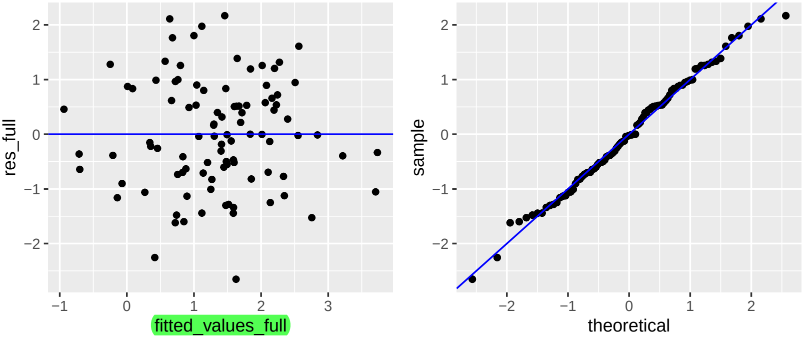
Residual Plots:

Regressing on lpsa only, plot against fitted value:



Forward selection; add variables one at a time V

Regressing on on lpsa, weight and age - plot against fitted values:



Comparing models I

To test the hypotheses

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$
$$H_a : \exists j \in \{1, \ldots, p\} \text{ such that } \beta_j \neq 0$$

NOTE: $\text{SST} = \text{SSE}_{H_0}$.

Under H_0 the variation in Y is explained by the mean $b_0 = \bar{Y}$.

Calculate the F statistic

$$\begin{aligned} F = \frac{\text{MSR}}{\text{MSE}} &= \frac{\text{SSR}/p}{\text{SSE}/(n-p-1)} \\ &= \frac{(\text{SSE}_{H_0} - \text{SSE}_{H_a})/(n-1-(n-p-1))}{\text{SSE}_{H_a}/(n-p-1)} \end{aligned}$$

Under H_0,

$$F \sim F_{p,n-p-1}$$

Comparing models II

To test the hypotheses

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

(saying nothing about $\beta_0, \beta_2, \beta_3, \ldots, \beta_p$)

- Regress Y on $x_2, x_3, \ldots, x_p$ and get residual sum of squares SSE_{H_0}.
- Regress Y on $x_1, x_2, \ldots, x_p$ and get residual sum of squares SSE_{H_a}.

Comparing models III

Then calculate the F statistic:

$$F = \frac{(\mathrm{SSE}_{H_0} - \mathrm{SSE}_{H_a}) \,/\, ((n-p) - (n-p-1))}{\mathrm{SSE}_{H_a}/(n-p-1)}$$
$$= \frac{\mathrm{SSE}_{H_0} - \mathrm{SSE}_{H_a}}{\mathrm{SSE}_{H_a}/(n-p-1)}$$

It can be shown that

$$F = \frac{b_1^2}{\mathrm{SE}(b_1)^2} = T^2$$

Under H_0,

$$F \sim F_{1,n-p-1}$$

Comparing models IV

More generally, to test the hypothesis

H_0: q specific explanatory variables all have zero coefficients

H_a: at least one of the q has a nonzero coefficient

- Regress Y on all predictor variables *except* the q variables of interest, and get the residual sum of squares SSE_{H_0}.
- Regress Y on *all* predictor variables and get the residual sum of squares SSE_{H_a}.

Comparing models V

Calculate the F statistic:

$$
\begin{aligned}
F &= \frac{(\mathsf{SSE}_{H_0} - \mathsf{SSE}_{H_a})/(\mathsf{df}_{H_0} - \mathsf{df}_{H_a})}{\mathsf{SSE}_{H_a}/\mathsf{df}_{H_a}} \\
&= \frac{(\mathsf{SSE}_{H_0} - \mathsf{SSE}_{H_a})/q}{\mathsf{SSE}_{H_a}/(n-p-1)}
\end{aligned}
$$

Under H_0,

$$F \sim F_{q,n-p-1}$$

Comparing models VI

Example: Compare simple regression model of prostate cancer against psa vs. multiple regression model against all three predictors.

```
anova(mod_psa,mod_full)
```

```
## Analysis of Variance Table
##
## Model 1: lcavol ~ lpsa
## Model 2: lcavol ~ lpsa + lweight + age
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     95 61.4
## 2     93 58.7  2      2.76 2.19   0.12
```

```
SSEa=sigma(mod_full)^2*mod_full$df.residual
SSE0=sigma(mod_psa)^2*mod_psa$df.residual
deltadf=mod_psa$df.residual-mod_full$df.residual
F=(SSE0-SSEa)/(deltadf)/(SSEa/mod_full$df.residual)
print(c(SSEa,SSE0,F,1-pf(F,deltadf,mod_full$df.residual)))
```

```
## [1] 58.658 61.421  2.190  0.118
```

Comparing models VII

Warning: the RSS in this output is what we call SSE, not SSR.

Based on the F-test there is no reason to reject the Null that the variables weight and age are not needed to predict cancer.

Binary predictors - dummy variables I

One of the predictors x could be binary. For example male/female.

Imagine two predicors x_1, x_2 - one continuous and one binary.

The linear model is:

$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$.

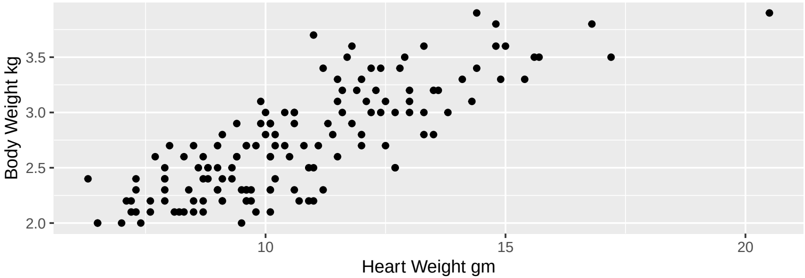
We can write this:

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_{1i} + \epsilon_i & \text{if } x_{2i} = 0 \\ \beta_0 + \beta_2 + \beta_1 x_{1i} & \text{if } x_{2i} = 1 \end{cases}$$

This produces two parallel lines one for each level of x_2.

Binary predictors - dummy variables II

Predict body weight based on heart weight of cats



Binary predictors - dummy variables III

```
cca=lm(Bwt~Hwt,data=cats)
ssa=summary(cca)
ssa$coefficients
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.02    0.10843     9.4 1.26e-16
## Hwt             0.16    0.00994    16.1 6.97e-34
```

```
c(ssa$r.squared,ssa$adj.r.squared,ssa$sigma)
```

```
## [1] 0.647 0.644 0.290
```

Binary predictors - dummy variables IV

Add dummy variable of SEX:

```
cca1=lm(Bwt~Hwt+Sex,data=cats)
ssa1=summary(cca1)
ssa1$coefficients
```

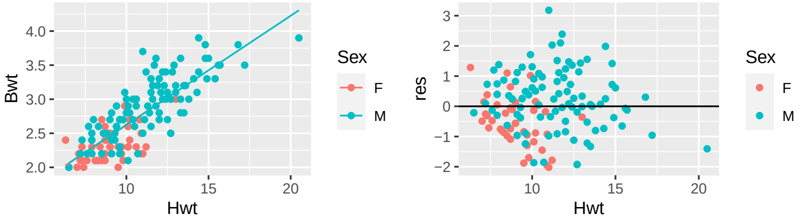
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.060     0.1020   10.40 3.78e-19
## Hwt            0.141     0.0102   13.83 5.12e-28
## SexM           0.241     0.0528    4.56 1.10e-05
```

```
c(ssa1$r.squared,ssa1$adj.r.squared,ssa1$sigma)
```

```
## [1] 0.692 0.688 0.271
```

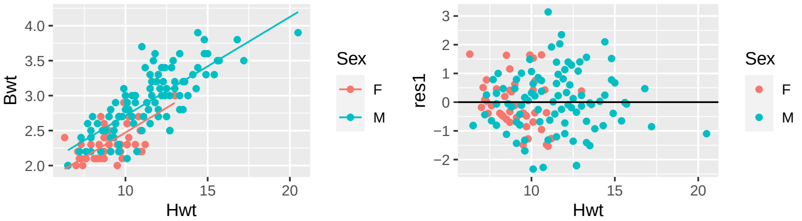
Binary predictors - dummy variables V

```
p1=qplot(Hwt,Bwt,geom=c("point"),data=cats,color=Sex)+
  geom_line(aes(y=predict(cca)))
cats$res=(cca$residuals-mean(cca$residuals))/sd(cca$residuals)
p2=qplot(Hwt,res,data=cats,color=Sex)+geom_hline(yintercept=0)
grid.arrange(p1, p2,  ncol=2)
```



Binary predictors - dummy variables VI

```
p1=qplot(Hwt,Bwt,geom=c("point"),data=cats,color=Sex)+geom_line(aes(y=predict(cca1)))
cats$res1=(cca1$residuals-mean(cca1$residuals))/sd(cca1$residuals)
p2=qplot(Hwt,res1,data=cats,color=Sex)+geom_hline(yintercept=0)
grid.arrange(p1, p2,  ncol=2)
```


Bwt
Hwt
Sex
F
M
res1
Hwt
Sex
F
M