Linear Modelling Summary Sheet

Compiled by: Shravan Vasishth (vasishth@uni-potsdam.de)

Version dated: March 8, 2013

These notes summarize the lecture notes from the Linear Modelling course at Sheffield's School of Mathematics and Statistics, MSc degree programme. The original notes were written by Dr. Kevin Walters and Dr. Jeremy Oakley. This summary is completely derived from these notes and from other MSc sources. Any errors are most probably mine. Everything is in matrix form unless a lower case letter with a subscript (such as x_i) is used (even there, I might deviate from this convention if I need to index sub-matrices; it's best to look at the context to decide what is meant).

Background

Some key distributional results

to-do

Some very basic matrix algebra facts

to-do

Basic facts

$$y = X\beta + \epsilon \tag{1}$$

$E(y) = X\beta = \mu \quad E(\epsilon) = 0$

$Var(y) = \sigma^2 I_n \quad Var(\epsilon) = \sigma^2 I_n$

$$y = X\hat{\beta} + e \tag{2}$$

Results for $\hat{\beta}$	Results for e
$E(\hat{\beta}) = \beta$	$E(e) = 0$
$Var(\hat{\beta}) = \sigma^2(X^TX)^{-1} = \frac{\sigma^2}{S_{xx}}$	$Var(e) = \sigma^2 M$
$\hat{\beta} \sim N_p(\beta, \sigma^2(X^TX)^{-1})$	$Var(e_i) = \sigma^2 m_{ii}$
	$E(e_i^2) = \sigma^2 m_{ii}$
$\hat{\beta} = (X^TX)^{-1}X^Ty$, X has full rank	$E(\sum e_i^2) = \sigma^2(n-p)$

Sum of Squares:

$$S(\hat{\beta}) = \sum e_i^2 = e^Te = (y - X\hat{\beta})^T(y - X\hat{\beta}) = y^Ty - y^TX\hat{\beta} = S_r \tag{3}$$

Alternatively: $S_r = y^Ty - \hat{\beta}^TX^TX\hat{\beta} = y^Ty - \hat{\beta}^TX^Ty$ (see review exercises 2).

Estimation of error variance: $e = My$

$$e = y - X\hat{\beta} = y - X(X^TX)^{-1}X^Ty = My \tag{4}$$

where

$$M = I_n - X(X^TX)^{-1}X^T \quad \text{M is symmetric, idempotent } n \times n \tag{5}$$

Note that $MX = 0$, which means that

$$E(e) = E(My) = ME(y) = MX\beta = 0 \tag{6}$$

Also, $Var(e) = Var(My) = MVar(y)M^T = \sigma^2 I_n M$.

Important properties of M:

- M is singular because every idempotent matrix except I_n is singular.
- $trace(M) = rank(M) = n - p$.

Residual mean square:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-p} \quad E(\hat{\sigma}^2) = \sigma^2 \tag{7}$$

The square root of $\hat{\sigma}^2$, $\hat{\sigma}$ is the **residual standard error**. Note: The phrase "standard error" here should not be misinterpreted to mean standard error in the sense of "SE".

Variance-covariance matrix:

In a model like

```
fm<-lm(Maint ~ Age, data = data)
```

, the variance-covariance matrix is:

$$\begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{pmatrix} \tag{8}$$

The correlation between the two parameter estimates is therefore:

$$Corr(\hat{\beta}_0, \hat{\beta}_1) = \frac{Cov(\hat{\beta}_0, \hat{\beta}_1)}{SE(\hat{\beta}_0)SE(\hat{\beta}_1)} \tag{9}$$

Example (tractor data):

```
> vcov(fm)
            (Intercept)     Age
(Intercept)       21591 -4624.0
Age               -4624  1267.9
```

We can check the correlation calculation using

```
> cov2cor(vcov(fm))
            (Intercept)      Age
(Intercept)     1.00000 -0.88378
Age            -0.88378  1.00000
```

Some short-cuts for hand-calculations

$S_{xx} = \sum(x_i - \bar{x})^2 \quad = \sum x_i^2 - n\bar{x}^2$

$S_{yy} = \sum(y_i - \bar{y})^2 \quad = \sum y_i^2 - n\bar{y}^2$

$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) \quad = \sum x_iy_i - n\bar{x}\bar{y}$

$$\hat{\beta} = (X^TX)^{-1}X^Ty = \begin{pmatrix} \bar{y} - \bar{x}\frac{S_{xy}}{S_{xx}} \\ \frac{S_{xy}}{S_{xx}} \end{pmatrix} \tag{10}$$

$$X^TX = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \tag{11}$$

$$(X^TX)^{-1} = \frac{1}{nS_{xx}} \begin{pmatrix} S_{xx} + n\bar{x}^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \tag{12}$$

Note that $\sum_{i=1}^n x_i = n\bar{x}$.

$$X^Ty = \begin{pmatrix} n\bar{y} \\ S_{xy} + n\bar{x}\bar{y} \end{pmatrix} \tag{13}$$

See [1, 25] for a full exposition.

Gauss-Markov conditions

This imposes distributional assumptions on $\epsilon = y - X\beta$.

$E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 I_n$,

Gauss-Markov theorem

Let a be any $p \times 1$ vector and suppose that X has rank p. Of all estimators of $\theta = a^T\beta$ that are unbiased and linear functions of y, the estimator $\hat{\theta} = a^T\hat{\beta}$ has minimum variance. Note that θ is a scalar.

Note: no normality assumption required! But if $\epsilon \sim N(0, \sigma^2)$, $\hat{\beta}$ have smaller variances than any other estimators.

Minimum variance unbiased linear estimators: to-do

R^2 or Coefficient of determination

$S_{TOTAL} = (y - \bar{y})^T(y - \bar{y}) = y^Ty - n\bar{y}^2$

$S_{REG} = (X\hat{\beta} - \bar{y})^T(X\hat{\beta} - \bar{y})$

$S_r = \sum e_i^2 = (y - X\hat{\beta})^T(y - X\hat{\beta})$

$$S_{TOTAL} = S_{REG} + S_r \tag{14}$$

$$R^2 = \frac{S_{TOTAL} - S_r}{S_{TOTAL}} = \frac{S_{REG}}{S_{TOTAL}} \tag{15}$$

For $y = 1_n\beta_0 + \epsilon$, then $R^2 = \frac{S_{REG}}{S_{TOTAL}} = 0$ because $X\hat{\beta} = \bar{y}$. So $S_{REG} = (X\hat{\beta} - \bar{y})^T(X\hat{\beta} - \bar{y}) = 0$.

In simple linear regression, $R^2 = r^2$. R^2 is a generalization of r^2.

Adjusted $R^2 = R^2_{Adj}$. $R^2_{Adj} = 1 - \frac{S_r/(n-p)}{S_{TOTAL}/(n-1)}$.

R^2 increases with increasing numbers of explanatory variables, therefore R^2_{Adj} is better.

Hypothesis testing

Some theoretical background

Multivariate normal:
Let $X^T =< X_1, \ldots, X_p >$, where X_i are univariate random variables.
X has a multivariate normal distribution if and only if every component of X has a univariate normal distribution.
Linear transformations:
Let A, b be constants. Then, $Ax + b \sim N_q(A\mu + b, A\Sigma A^T)$.
Standardization:
Note that Σ is positive definite (it's a variance covariance matrix), so $\Sigma = CC^T$. C is like a square root (not necessarily unique).
It follows "immediately" that

$$C^{-1}(X-\mu) \sim N_p(0_p, I_p) \tag{16}$$

If Σ is a diagonal matrix, then $X_1, \ldots, X_n$ are independent and uncorrelated.
Quadratic forms:
Recall distributional result: If we have n independent standard normal random variables, their sum of squares is χ^2_n.
Lt $z = C^{-1}(X-\mu)$, and $\Sigma = CC^T$. The sum of squares $z^T z$ is:

$$\begin{aligned} z^T z &= [C^{-1}(X-\mu)]^T [C^{-1}(X-\mu)] \\ &= (X-\mu)^T [C^{-1}]^T [C^{-1}](X-\mu) \quad \ldots (AB)^T = B^T A^T \end{aligned} \tag{17}$$

Note that $[C^{-1}]^T = [C^T]^{-1}$. Therefore,

$$\begin{aligned} [C^{-1}]^T [C^{-1}] &= [C^T]^{-1}[C^{-1}] \\ &= (C^T C)^{-1} \\ &= (CC^T)^{-1} \\ &= \Sigma^{-1} \end{aligned} \tag{18}$$

Therefore: $z^T z = (X-\mu)^T \Sigma^{-1}(X-\mu) \sim \chi^2_p$, where p is the number of parameters.
Quadratic expressions involving idempotent matrices
Given a matrix K that is idempotent, symmetric. Then:

$$x^T K x = x^T K^2 x = x^T K^T K x \tag{19}$$

Let $x \sim N_n(\mu, \sigma^2 I_n)$, and let K be a symmetric, idempotent $n \times n$ matrix such that $K\mu = 0$. Let r be the rank or trace of K. Then we have the **sum of squares property**:

$$x^T K x \sim \sigma^2 \chi^2_r \tag{20}$$

The above generalizes the fact that if we have n independent standard normal random variables, their sum of squares is χ^2_n.
Two points about the sum of squares property:

- Recall that the expectation of a chi-squared random variable is its degrees of freedom. It follows that:

  $$E(x^T K x) = \sigma^2 r \tag{21}$$

  If $K\mu \neq 0$, $E(x^T K x) = \sigma^2 r + \mu^T K \mu$.
- If K is idempotent, so is $I - K$. This allows us to split $x^T x$ into two components sums of squares:

  $$x^T x = x^T K x + x^T (I-K)x \tag{22}$$

Partition sum of squares:
Let $K_1, K_2, \ldots, K_q$ be symmetric idempotent $n \times n$ matrices such that $\sum K_i = I_n$ and $K_i K_j = 0$, for all $i \neq j$. Let $x \sim N_n(\mu, \sigma^2)$. Then we have the following partitioning into independent sums of squares:

$$x^T x = \sum x^T K_i x \tag{23}$$

If $K_i\mu = 0$, then $x^T K_i x \sim \sigma^2 \chi^2_{r_i}$, where r_i is the rank of K_i.

Confidence intervals for $\hat{\beta}$

Note that $\hat{\beta} \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$, and that $\frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2_{n-p}}{n-p}$.
From distributional theory we know that $T = \frac{X}{\sqrt{Y/v}}$, when $X \sim N(0,1)$ and $Y \sim \chi^2_v$.
Let x_i be a column vector containing the values of the explanatory/regressor variables for a new observation i. Then if we define:

$$X = \frac{x_i^T \hat{\beta} - x_i^T \beta}{\sqrt{\sigma^2 x_i^T (X^T X)^{-1} x_i}} \sim N(0,1) \tag{24}$$

and

$$Y = \frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2_{n-p}}{n-p} \tag{25}$$

It follows that $T = \frac{X}{\sqrt{Y/v}}$:

$$T = \frac{x_i^T \hat{\beta} - x_i^T \beta}{\sqrt{\hat{\sigma}^2 x_i^T (X^T X)^{-1} x_i}} = \frac{\frac{x_i^T \hat{\beta} - x_i^T \beta}{\sqrt{\sigma^2 x_i^T (X^T X)^{-1} x_i}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} \sim t_{n-p} \tag{26}$$

I.e., a 95% CI:

$$x_i^T \hat{\beta} \pm t_{n-p,1-\alpha/2} \sqrt{\hat{\sigma}^2 x_i^T (X^T X)^{-1} x_i} \tag{27}$$

Cf. a prediction interval:

$$x_i^T \hat{\beta} \pm t_{n-p,1-\alpha/2} \sqrt{\hat{\sigma}^2 (1 + x_i^T (X^T X)^{-1} x_i)} \tag{28}$$

Note that

1. A prediction interval will be wider about the edges; this is because the term $\hat{\sigma}^2(1 + x_i^T (X^T X)^{-1} x_i)$ in the prediction interval formula is minimized at the mean value of the predictor variable. When $x_i = \bar{x}$ we have the smallest value for the term, and so the further away the x_i value from $\bar{x}$, the larger the interval.
2. The width of the prediction interval stays much more constant around the range of observed values. This is because 1 is much larger than $x_i^T (X^T X)^{-1} x_i$); so if x_i is near the mean value for x then this term will not change much.

Distributions of estimators and residuals

Covar$(\hat{\beta}, e) = 0$:
$\text{Var}\begin{pmatrix} \hat{\beta} \\ e \end{pmatrix} = \begin{pmatrix} Var(\hat{\beta}) & 0 \\ 0 & Var(e) \end{pmatrix} = \begin{pmatrix} \sigma^2 (X^T X)^{-1} & 0 \\ 0 & \sigma^2 M \end{pmatrix}$.
Confidence intervals for components of β
Let $G = (X^T X)^{-1}$, and g_{ii} the i-th diagonal element.

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 g_{ii}) \tag{29}$$

Since $\hat{\beta}$ and S_r are independent, we have:

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{g_{ii}}} \sim t_{n-p} \tag{30}$$

The 95% CI:

$$\hat{\beta}_i \pm t_{n-p,(1-\alpha)/2} \hat{\sigma} \sqrt{g_{ii}} \tag{31}$$

Maximum likelihood estimators

to-do

Hypothesis testing

A general format for specifying null hypotheses: $H_0 : C\beta = c$, where C is a $q \times p$ matrix and c is a $q \times 1$ vector of known constants. The matrix C effectively asserts specific values for q linear functions of β. In other words, it asserts q null hypotheses stated in terms of (components of) the parameter vector β.
E.g., given:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \tag{32}$$

we can test $H_0 : \beta_1 = 1, \beta_2 = 2$ by setting
$C = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ and $c = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$.
The alternative is usually the negation of the null, i.e., $H_1 : C\beta \neq c$, which means that at least one of the q linear functions does not take its hypothesized value.
Constructing a test:

$$C\hat{\beta} \sim N_q(c, \sigma^2 C(X^T X)^{-1} C^T) \tag{33}$$

So, if H_0 is true, by sum of squares property:

$$(C\hat{\beta} - c)^T C (X^T X)^{-1} C^T (C\hat{\beta} - c) \sim \sigma^2 \chi^2_q \tag{34}$$

In other words:

$$\frac{(C\hat{\beta}-c)^T C(X^TX)^{-1}C^T(C\hat{\beta}-c)}{\sigma^2} \sim \chi^2_q \tag{35}$$

Note that $\hat{\beta}$ is independent of $\hat{\sigma}^2$, and recall that

$$\frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2_{n-p}}{n-p} \Leftrightarrow \frac{\hat{\sigma}^2(n-p)}{\sigma^2} \sim \chi^2_{n-p} \tag{36}$$

Recall distributional result: if $X \sim \chi^2_v, Y \sim \chi^2_w$ and X, Y independent then $\frac{X/v}{Y/w} \sim F, v, w$.
It follows that if H_0 is true, and setting $X = \frac{(C\hat{\beta}-c)^T C(X^TX)^{-1}C^T(C\hat{\beta}-c)}{\sigma^2}$, $Y = \frac{\hat{\sigma}^2(n-p)}{\sigma^2}$, and setting the degrees of freedom to $v = q$ and $w = n-p$:

$$\frac{X/v}{Y/w} = \frac{\frac{(C\hat{\beta}-c)^T C(X^TX)^{-1}C^T(C\hat{\beta}-c)}{\sigma^2}/q}{\frac{\hat{\sigma}^2(n-p)}{\sigma^2}/(n-p)} \tag{37}$$

Simplifying:

$$\frac{(C\hat{\beta}-c)^T C(X^TX)^{-1}C^T(C\hat{\beta}-c)}{q\hat{\sigma}^2} \sim F_{q,n-p} \tag{38}$$

This is a **one-sided test** even though the original alternative was two-sided.
Special cases of hypothesis tests:
When q is 1, we have only one hypothesis to test, the i-th element of β. Given:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \tag{39}$$

we can test $H_0 : \beta_1 = 0$ by setting
$C = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$ and $c = 0$.
Using the fact that $X \sim t(v) \Leftrightarrow X^2 \sim F(1, v)$, we have

$$\frac{\hat{\beta}_i - c_i}{\hat{\sigma}\sqrt{g_{ii}}} \sim t_{n-p} \tag{40}$$

Sum of squares

This is a very important section!

> Recall: If K is idempotent, so is $I - K$. This allows us to split $x^T x$ into two components sums of squares:
>
> $$x^T x = x^T K x + x^T (I-K)x \tag{41}$$
>
> Let $K_1, K_2, \dots, K_q$ be symmetric idempotent $n \times n$ matrices such that $\sum K_i = I_n$ and $K_i K_j = 0$, for all $i \neq j$. Let $x \sim N_n(\mu, \sigma^2)$. Then we have the following partitioning into independent sums of squares:
>
> $$x^T x = \sum x^T K_i x \tag{42}$$
>
> If $K_i\mu = 0$, then $x^T K_i x \sim \sigma^2 \chi^2_{r_i}$, where r_i is the rank of K_i.

We can use the sum of squares property just in case K is idempotent, and $K\mu = 0$. Below, $K = M$ and $\mu = E(y) = X\beta$.
Consider the sum of squares partition:

$$y^T y = \underbrace{y^T M y}_{\uparrow\; S_r = e^T e} + \underbrace{y^T (I-M) y}_{\uparrow\; \hat{\beta}^T (X^T X)\hat{\beta}} \tag{43}$$

Note that the preconditions for sums of squares partitioning are satisfied:

1. M is idempotent (and symmetric), rank=trace=$n-p$.
2. $I - M$ is idempotent (and symmetric), rank=trace=p.
3. $ME(y) = 0$ because $ME(y) = MX\beta$ and $MX = 0$.

We can therefore partition the sum of squares into two independent sums of squares:

$$y^T y = \underbrace{y^T M y}_{\uparrow\; e^T e \sim \sigma^2 \chi^2_{n-p}} \quad + \quad \underbrace{y^T (I-M) y}_{\uparrow\; \sim \sigma^2 \chi^2_p \text{ iff } X\beta = 0, i.e., \beta = 0} \tag{44}$$

So, iff we have $H_0 : \beta = 0$, we can partition sum of squares as above. Saying that $\beta = 0$ is equivalent to saying that X has rank p and $X\beta = 0$.

Testing the effect of a subset of regressor variables

Let:

$$C = (0_{p-q} I_q) \quad c = 0, \text{ and } \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \tag{45}$$

Here, $\beta_{1,2}$ are vectors (sub-vectors?), not components of the β vector. Then, $C \times \beta = \beta_2$ and $H_0 : \beta_2 = 0$. Note that order of elements in β is arbitrary; i.e., any subset of β can be tested.
Since $C \times \beta = \beta_2$ and $c = 0$, we can construct a sum of squares:

$$(C\hat{\beta} - c)^T C(X^TX)^{-1}C^T(C\hat{\beta} - c) \sim \sigma^2 \chi^2_q \tag{46}$$

This becomes (since $C\beta = \hat{\beta}_2$):

$$\hat{\beta}_2^T C(X^TX)^{-1}C^T \hat{\beta}_2 \sim \sigma^2 \chi^2_q \tag{47}$$

We can rewrite this as: $\hat{\beta}_2^T G_{qq} \hat{\beta}_2$, where
$G_{qq} = C(X^TX)^{-1}C^T$ (G_{qq} should not be confused with g_{ii}) is a $q \times q$ submatrix of $G = (X^TX)^{-1}$.
Note that $\hat{\beta}$ is independent of $\hat{\sigma}^2$, and recall that $\frac{\hat{\sigma}^2(n-p)}{\sigma^2} \sim \chi^2_{n-p}$. We can now construct the F-test as before:

$$\frac{\hat{\beta}_2^T C(X^TX)^{-1}C^T \hat{\beta}_2}{q\hat{\sigma}^2} = \frac{\hat{\beta}_2^T G \hat{\beta}_2}{q\hat{\sigma}^2} \sim F_{q,n-p} \tag{48}$$

Sums of squares:
We can construct three idempotent matrices:

- $M = I_n - X(X^TX)^{-1}X^T$
- $M_1 = X(X^TX)^{-1}X^T - [X(X^TX)^{-1}C^T][\underbrace{C(X^TX)^{-1}C^T}_{\uparrow\; G}]^{-1}[C(X^TX)^{-1}X^T]$

  (that is: $M_1 = X(X^TX)^{-1}X^T - M_2$)
- $M_2 = [X(X^TX)^{-1}C^T][\underbrace{C(X^TX)^{-1}C^T}_{\uparrow\; G}]^{-1}[C(X^TX)^{-1}X^T]$

Note that $M + M_1 + M_2 = I_n$ and $MM_1 = MM_2 = M_1M_2 = 0$. I.e., sum of squares partition property applies. We have three independent sums of squares:

1. $S_r = y^T M y$
2. $S_1 = y^T M_1 y = \hat{\beta}^T X^T X \hat{\beta} - \hat{\beta}_2^T G_{qq}^{-1} \hat{\beta}_2$
3. $S_2 = y^T M_2 y = \hat{\beta}_2^T G_{qq}^{-1} \hat{\beta}_2$

So: $y^T y = S_r + S_1 + S_2$. Then:

- It is unconditionally true that $S_r \sim \sigma^2 \chi^2_{n-p}$.
- If $H_0 : \beta = 0$ is true, then $E(\hat{\beta}_2) = \beta_2 = 0$. It follows from the sum of squares property that $S_2 \sim \sigma^2 \chi^2_q$.
- Regarding S_1: We can prove that $M_1 = X_1(X_1^T X_1)^{-1} X_1^T$, where X_1 contains the first $p-q$ columns of X. It follows that:
  $S_1 = y^T M_1 y = y^T X_1 (X_1^T X_1)^{-1} X_1^T y$
  Note that $X_1(X_1^T X_1)^{-1} X_1^T$ is idempotent. If $\beta = 0$, i.e., if $E(y) = X\beta = 0$, we can use the sum of squares property and conclude that
  $S_1 \sim \sigma^2 \chi^2_{p-q}$
  The degrees of freedom are $p - q$ because the rank=trace of $X_1(X_1^T X_1)^{-1} X_1^T$ is $n - p$.
  Thus, S_1 is testing $\beta_1 = 0$ but under the assumption that $\beta_2 = 0$.

Analysis of variance

Sources of variation	SS	df	MS	MS ratio
Due to X_1 if $\beta_2 = 0$ d	S_1	$p-q$	$S_1/(p-q)$	F_1 $F_{p-q,n-p}$
Due to X_2	S_2	q	S_2/q	F_2 $F_{q,n-p}$
Residuals	S_r	$n-p$	$\hat{\sigma}^2$	
Total	$y^T y$	n		

Note:

1. The ANOVA tests are **performed in order**: First we test $H_0 : \beta_2 = 0$. Then, if this test does not reject the null, we test $H_0 : \beta_1 = 0$ **on the assumption (which may or may not be true)** that $\beta_2 = 0$.
2. What happens if we reject the first hypothesis?

The null or minimal model (constant term)
We can set $C = I_p$ and $c = 0$. This tests whether all coefficients are zero. But this states that $E(y) = 0$, whereas it should have a non-zero value (e.g., reading times). We include the constant term to accommodate this desire to have $E(y) = \mu =\neq 0$. In matrix format: let β be the parameter vector; then, $\beta_1 = \mu$ is the first, constant, term, and the rest of the parameters are the vector β_2 $(p-1 \times 1)$. The first column of X will be $X_1 = 1_n$.

1. $S_1 = y^T(X_1^T X_1)^{-1} X_1^T y = (\sum y)^2/n = n\bar{y}^2$
2. $S_r = y^T y - \hat{\beta}^T X^T X \hat{\beta}$
3. $S_2 = y^T y - S_1 - S_r = \hat{\beta}^T X^T X \hat{\beta} - n\bar{y}^2$

It is normal to omit the row in the ANOVA table corresponding to the constant term.

Testing whether all predictors (besides the constant term) are zero

To test whether p predictor variables have any effect on y,we set $q = p - 1$, and our anova table looks like this:

Sources of variation	SS	df	MS	MS ratio
Due to regressors	S_2	$p-1$	$\frac{S_2}{(p-1)}$	F_2 $F_{p-1,n-p}$
Residuals	S_r	$n-p$	$\hat{\sigma}^2$	
Total (**adjusted**)	$S_{yy} = (y-\bar{y})^T(y-\bar{y}) = y^T y - n\bar{y}^2$	n-1		

Note that $S_{yy} = \sum(y_i - \bar{y})^2$ is the residual sum of squares that we get after fitting the constant $\hat{\mu} = \bar{y}$.

Testing a subset of predictors β_2

Sources of variation	SS	df	MS	MS ratio
Due to X_1 if $\beta_2 = 0$ (test of β_1)	S_1	$p-q-1$	$\frac{S_1}{(p-q-1)}$	(F_1) $F_{p-q-1,n-p}$
Due to X_2 (test of β_2)	S_2	q	$\frac{S_2}{q}$	F_2 $F_{q,n-p}$
Residuals	S_r	$n-p$	$\hat{\sigma}^2$	
Total	S_{yy}	n-1		

Checking model assumptions

Standardized residuals (`studres` in R)

Recall that $Var(e) = \sigma^2 M$, where $M = I_n - X(X^T X)^{-1} X^T$ M is symmetric, idempotent $n \times n$. The diagonals of M are all less than 1, and are not all equal (i.e., not equal variance), and off-diagonals are not 0 (i.e., the residuals are correlated). Correcting for unequal variance is done by the **scaled residual**:

$$e_i* = \frac{e_i}{\sqrt{m_{ii}}} \tag{49}$$

Note: $Var(e_i*) = \sigma^2$. The **standardized residuals** are

$$s_i = \frac{e_i*}{\hat{\sigma}} \tag{50}$$

This is approximately t_{n-p} (approximately because e_i* and $\hat{\sigma}$ are not independent). Since $s_i \sim t_{n-p}$, we can designate a residual as an outlier if $| s_i |> t_{crit}$ where t_{crit} is the critical t-value.

Standardized deletion residuals

This is a more exact way to test for outliers than the above discussion. Define:

$$\hat{\beta}_{-i} = (X_{-i}^T X_{-i})^{-1} X_{-i}^T y_{-i} \tag{51}$$

where the $-i$ refers to removing data point i. Standardized deletion residuals are

$$s_{-i} = \frac{e_i}{\hat{\sigma}_{-i}\sqrt{m_{ii}}} \tag{52}$$

We can compute s_{-i} from s_i:

$$s_{-i} = \frac{s_i\sqrt{n-p-1}}{\sqrt{n-p-s_i^2}} \sim t_{n-p-1} \tag{53}$$

If n is large, $s_{-i} \approx s_i$.

Correcting for multiple testing

Šidák correction: "suppose we are performing n tests and in each test we specify the probability of making a type I error to be β (note: don't confuse this as type II error). Then, if the tests are independent, the probability of at least one false positive claim in the n tests is given by

$$1 - (1-\beta)^n = \alpha \Leftrightarrow \beta = 1 - (1-\alpha)^{1/n} \tag{54}$$

This correction "has a stronger bound [than the Bonferroni] and so has greater statistical power."

Checks

1. Normality: qqnorm etc. Hist is a useful addition to qqplot in large samples.
2. Independence: index-plots: residuals against observation number. Not useful for small samples. Or: compute correlation between e_i, e_{i+1} pairs of residuals.
3. Homoscedasticity: residuals against fitted. Fan out suggests violation. A quadratic trend in a plot of residuals against predictor x could suggest that a quadratic predictor term is needed; note that $X^T e = 0$. (review exercises 3), so we will never have a perfect straight line in such a plot. Alternative: Bartlett's test.

Formal tests of normality

Komogorov-Smirnov and Shapiro-Wilk. Only useful for large samples ; not very powerful and not much better than diagnostic plots. Tests may be useful as follow-ups if non-normality is suspected.

Influence and leverage (`lm.influence$hat` in R)

A point can influence the parameter estimates without being an exceptional outlier. Influence does not depend on "outlyingness". Potential to influence (e.g., by being an extreme x value) is called leverage; once the y value is also extreme, we have influence. I.e., it takes an extreme x and y value to be influential, and it takes only an extreme x value to have leverage.

Leverage more formally defined: recall that $M = I_n - X(X^T X)^{-1} X^T$. Define a hat matrix $H = I - M = X(X^T X)^{-1} X^T$. It's called a hat matrix because it puts a hat on y: $\hat{y} = X\hat{\beta} = Hy$. Since x_i^T is the i-th row of X, we have $h_{ii} = x_i^T (X^T X)^{-1} x_i$. The measure for leverage is:

$$h_{ii} = 1 - m_{ii} \tag{55}$$

Notice that h_{ii} is a scalar, so trace($h_{ii} = h_{ii}$. So (because for a square matrix A,B, tr(AB)= tr(BA)):

$$h_{ii} = tr(x_i^T (X^T X)^{-1} x_i) = tr(x_i^T x_i (X^T X)^{-1}) \tag{56}$$

Since $X^T X = \sum_{i=1}^n x_i x_i^T$, h_{ii} represents the magnitude of $x_i x_i^T$ relative to the sum of the values for all observations. Note that h_{ii} only depends on X.

Also note that

$$\sum_{i=1}^n h_{ii} = tr(X^T X (X^T X)^{-1}) = tr(I_p) = p \quad mean(h_{ii}) = p/n \tag{57}$$

h_{ii} measures leverage because $Var(e_i) = \sigma^2 m_{ii} = \sigma^2(1-h_{ii})$ and $Var(\hat{y}_i) = \sigma^2 h_{ii}$. Therefore h_{ii} has to lie between 0 and 1. When it is close to one, the fitted value will be close to the actual value of y_i—signalling potential for leverage (aside by SV: the explanation sounds circular to me—this statement says it has leverage by definition. Also, I don't know why I should care that a data point has *potential* to influence the estimates).

A cutoff one can use to identify high leverage points is $h_{ii} > 2p/n$ or $h_{ii} > 3p/n$.

The leverage of a data point is directly related to how far away it is from the mean:

$$h_{ii} = n^{-1} + \frac{(x_i - \bar{x})^2}{S_{xx}} \tag{58}$$

In `lm.influence`, "coefficients is the matrix whose i-th row contains the change in the estimated coefficients which results when the i-th case is dropped from the regression. sigma is a vector whose i-th element contains the estimate of the residual standard error obtained when the i-th case is dropped from the regression" (p. 71 of lecture notes).

Cook's distance D: A measure of influence

Let s_i be the i-th standardized residual, $\hat{\beta}_{-i}$ the estimate of the vector of parameters with the i-th row removed.

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{-i})^T (X^T X)^{-1} (\hat{\beta} - \hat{\beta}_{-i})}{p\hat{\sigma}^2} = \frac{s_i^2 h_{ii}}{p(1-h_{ii})} \tag{59}$$

A data point is influential if it is outlying as well as high leverage. Cutoff for Cook's distance is $\frac{4}{n}$.

Procedure for checking model fit: to-do, see p 73

Transformations

Suppose Y is a random variable whose variance depends on its mean. I.e., $E(y) = \mu, Var(y) = g(\mu)$. The function $g(\cdot)$ is known.

We seek a transformation from y to $z = f(y)$ such that the variance of z is (approximately) constant. [Some important details skipped–to-do]

Box-Cox family:

$$f_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases} \tag{60}$$

We assume that $f_\lambda(y) \sim N(x_i^T \beta, \sigma^2)$. So we have to just estimate λ by MLE, along with β.

Maximum likelihood estimation of λ: to-do (see p 78)

References

[1] Norman R. Draper and Harry Smith. *Applied Regression Analysis*. Wiley, New York, 1998.

Cheat sheet template taken from Winston Chang: http://www.stdout.org/~winston/latex/