Linear Modelling Summary Sheet
Compiled by: Shravan Vasishth
(vasishth@uni-potsdam.de)
Version dated: May 15, 2013

Background

Derivations of combinations of functions

$$(uv)' = uv' + vu' \tag{1}$$

$$(u/v)' = \frac{vu' - uv'}{v^2} \tag{2}$$

Some key distributional results

Some very basic matrix algebra facts

Inverse (2x2):
$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

Inverse of non-singular matrices. If A and B are non-singular matrices then $(AB)^{?1} = B^{-1}A^{-1}$.

Symmetric square matrix: $A = A^T$.

If a symmetric matrix A is non-singular then A^{-1} is also symmetric.

$AA^{-1} = A^{-1}A = I$ given A is square and invertible.

If the symmetric matrix A is non-singular then $A^{?1}$ is also symmetric.

Multiplication is distributive: Multiplication is distributive over addition and subtraction, so $(A-B)(C-D) = AC - BC - AD + BD$.

Transpose: $(A+B)^T = A^T + B^T$ and $(AB)^T = B^T A^T$

Sum of squares: $\sum x_i^2 = \mathbf{x}^T\mathbf{x}$.

Symmetry under multiplication: If A is $n \times p$, then AA^T and A^TA are symmetric.

Trace of a square matrix:

1. $tr(A) = \sum a_{ii}$
2. $tr(A+B) = tr(A) + tr(B)$
3. $tr(cA) = ctr(A)$
4. $tr(AB) = tr(BA)$

Idempotent: $A^2 = AA = A$. Example: $A = I_n$; this is the only non-singular idempotent matrix.

If A is idempotent and if $A \neq I_n$, then A is singular and its trace is equal to its rank $n-p$, for some $p > 0$.

Inverse of a matrix product: If A and B are non-singular matrices then $(AB)^{-1} = B^{-1}A^{-1}$.

Rank: the number of linearly independent columns or rows of A.

How to determine linear independence:

Basic facts

$$y = X\beta + \epsilon \tag{3}$$

$E(y) = X\beta = \mu \quad E(\epsilon) = 0$
$Var(y) = \sigma^2 I_n \quad Var(\epsilon) = \sigma^2 I_n$

$$y = X\hat{\beta} + e \tag{4}$$

Note that $S_{xx} = \frac{1}{(X'X)^{-1}}$, and $G = (X'X)^{-1}$, so that $S_{xx} = \frac{1}{G}$.

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

Sum of Squares:

$$S(\hat{\beta}) = \sum e_i^2 = e^Te = (y - X\hat{\beta})^T(y - X\hat{\beta}) = y^Ty - y^TX\hat{\beta} = S_r \tag{5}$$

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: $e = My$

$$e = y - X\hat{\beta} = y - X(X^TX)^{-1}X^Ty = My \tag{6}$$

where

$$M = I_n - X(X^TX)^{-1}X^T \tag{7}$$

M is symmetric, idempotent $n \times n$.
Note that $MX = 0$, which means that

$$E(e) = E(My) = ME(y) = MX\beta = 0 \tag{8}$$

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:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-p} \quad E(\hat{\sigma}^2) = \sigma^2 \tag{9}$$

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:

$$\begin{pmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{pmatrix} \tag{10}$$

The correlation between the two parameter estimates is therefore:

$$Corr(\hat{\beta}_0, \hat{\beta}_1) = \frac{Cov(\hat{\beta}_0, \hat{\beta}_1)}{SE(\hat{\beta}_0)SE(\hat{\beta}_1)} \tag{11}$$

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

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

$$\hat{\beta} = (X^TX)^{-1}X^Ty = \begin{pmatrix} \bar{y} - \bar{x}\frac{S_{xy}}{S_{xx}} \\ \frac{S_{xy}}{S_{xx}} \end{pmatrix} \tag{12}$$

$$X^TX = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \tag{13}$$

$$(X^TX)^{-1} = \frac{1}{nS_{xx}} \begin{pmatrix} S_{xx} + n\bar{x}^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \tag{14}$$

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

$$X^Ty = \begin{pmatrix} n\bar{y} \\ S_{xy} + n\bar{x}\bar{y} \end{pmatrix} \tag{15}$$

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

$S_{TOTAL} = (y - \bar{y})^T(y - \bar{y}) = y^T y - n\bar{y}^2$
$S_{REG} = (X\hat{\beta} - \bar{y})^T(X\hat{\beta} - \bar{y})$
$S_r = \sum e_i^2 = (y - X\hat{\beta})^T(y - X\hat{\beta})$

$$S_{TOTAL} = S_{REG} + S_r \tag{16}$$

$$R^2 = \frac{S_{TOTAL} - S_r}{S_{TOTAL}} = \frac{S_{REG}}{S_{TOTAL}} \tag{17}$$

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

$$C^{-1}(X - \mu) \sim N_p(0_p, I_p) \tag{18}$$

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:

$$\begin{aligned} z^T z &= [C^{-1}(X-\mu)]^T[C^{-1}(X-\mu)] \\ &= (X-\mu)^T[C^{-1}]^T[C^{-1}](X-\mu) \quad \ldots (AB)^T = B^T A^T \end{aligned} \tag{19}$$

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

$$\begin{aligned} [C^{-1}]^T[C^{-1}] &= [C^T]^{-1}[C^{-1}] \\ &= (C^T C)^{-1} \\ &= (CC^T)^{-1} \\ &= \Sigma^{-1} \end{aligned} \tag{20}$$

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:

$$x^T K x = x^T K^2 x = x^T K^T K x \tag{21}$$

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:

$$x^T K x \sim \sigma^2\chi^2_r \tag{22}$$

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:

$$E(x^T K x) = \sigma^2 r \tag{23}$$

  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:

$$x^T x = x^T K x + x^T(I - K)x \tag{24}$$

Partition sum of squares:

[helps prove independence of SSs]

1. Let $K_1, K_2, \ldots, K_q$ be **symmetric idempotent** $n \times n$ **matrices** such that
2. $\sum K_i = I_n$ and
3. $K_i K_j = 0$, for all $i \neq j$.
4. Let $x \sim N_n(\mu, \sigma^2 I_n)$.

Then we have the following partitioning into **independent** sums of squares:

$$x^T x = \sum x^T K_i x \tag{25}$$

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.

Example:

$$y^T y = y^T M y + y^T(I - M)y \tag{26}$$

Let $K_1 = M$ and $K_2 = (I - M)$. It is easy to check that all four conditions above are satisfied; therefore the sums of squares are independent.

Note that

$$y^T M y = e^T e \sim \chi^2_{n-p} \tag{27}$$

and

$$y^T(I - My) = \hat{\beta}^T(X^T X)\hat{\beta} \sim \chi^2_p \tag{28}$$

Recall distributional result: if $X \sim \chi^2_v, Y \sim \chi^2_w$ and X, Y independent then $\frac{X/v}{Y/w} \sim F_{v,w}$.

Therefore, $\frac{\frac{y^T(I-My)}{n-p}}{\frac{y^T M y}{p}} \sim F_{p,n-p}$.

Confidence intervals for $\hat\beta$

Note that $\hat\beta \sim N_p(\beta, \sigma^2(X^TX)^{-1})$, 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:

$$X = \frac{x_i^T\hat\beta - x_i^T\beta}{\sqrt{\sigma^2 x_i^T(X^TX)^{-1}x_i}} \sim N(0,1) \tag{29}$$

and

$$Y = \frac{\hat\sigma^2}{\sigma^2} \sim \frac{\chi^2_{n-p}}{n-p} \tag{30}$$

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

$$T = \frac{x_i^T\hat\beta - x_i^T\beta}{\sqrt{\hat\sigma^2 x_i^T(X^TX)^{-1}x_i}} = \frac{\frac{x_i^T\hat\beta - x_i^T\beta}{\sqrt{\sigma^2 x_i^T(X^TX)^{-1}x_i}}}{\sqrt{\frac{\hat\sigma^2}{\sigma^2}}} \sim t_{n-p} \tag{31}$$

I.e., a 95% CI:

$$x_i^T\hat\beta \pm t_{n-p,1-\alpha/2}\sqrt{\hat\sigma^2 x_i^T(X^TX)^{-1}x_i} \tag{32}$$

Cf. a prediction interval:

$$x_i^T\hat\beta \pm t_{n-p,1-\alpha/2}\sqrt{\hat\sigma^2(1 + x_i^T(X^TX)^{-1}x_i)} \tag{33}$$

Note that

1. A prediction interval will be wider about the edges; this is because the term $\hat\sigma^2(1 + x_i^T(X^TX)^{-1}x_i)$ 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 $x_i^T(X^TX)^{-1}x_i)$; so if x_i is near the mean value for x then this term will not change much.

Distributions of estimators and residuals

$\text{Covar}(\hat\beta, e) = 0$:
$\text{Var}\begin{pmatrix}\hat\beta \\ e\end{pmatrix} = \begin{pmatrix}Var(\hat\beta) & 0 \\ 0 & Var(e)\end{pmatrix} = \begin{pmatrix}\sigma^2(X^TX)^{-1} & 0 \\ 0 & \sigma^2 M\end{pmatrix}$.

Confidence intervals for components of β

Let $G = (X^TX)^{-1}$, and g_{ii} the i-th diagonal element.

$$\hat\beta_i \sim N(\beta_i, \sigma^2 g_{ii}) \tag{34}$$

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

$$\frac{\hat\beta_i - \beta_i}{\hat\sigma\sqrt{g_{ii}}} \sim t_{n-p} \tag{35}$$

The 95% CI:

$$\hat\beta_i \pm t_{n-p,(1-\alpha)/2}\hat\sigma\sqrt{g_{ii}} \tag{36}$$

Maximum likelihood estimators

For σ^2:
Let X_i, $i = 1, \dots, n$ be a random variable with PDF $f(x;\sigma) = \frac{1}{2\sigma}exp(-\frac{|x|}{\sigma})$. Find $\hat\sigma$, the MLE of σ.

$$L(\sigma) = \prod f(x_i;\sigma) = \frac{1}{(2\sigma)^n}exp(-\sum\frac{(x_i-\mu)^2}{\sigma^2}) \tag{37}$$

Let ℓ be log likelihood. Then:

$$\ell(x;\sigma) = -n\log 2 - n\log\sigma - \sum(x_i-\mu)^2/\sigma^2 \tag{38}$$

Differentiating and equating to zero to find maximum:

$$\ell'(\sigma) = -\frac{n}{\sigma} + \sum(x_i-\mu)^2/\sigma^3 = 0 \tag{39}$$

Rearranging the above, the MLE for σ is:

$$\hat\sigma^2 = \sum(x_i-\mu)^2/n \tag{40}$$

Since $S_r \sim \chi^2_{n-p}$, $E(S_r) = \sigma^2(n-p)$. So we need to correct S_r as $S_r/n-p$ to get $E(S_r) = \sigma^2$.

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:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \tag{41}$$

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:

$$C\hat\beta \sim N_q(c, \sigma^2 C(X^TX)^{-1}C^T) \tag{42}$$

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

$$(C\hat\beta - c)^T[C(X^TX)^{-1}C^T]^{-1}(C\hat\beta - c) \sim \sigma^2\chi^2_q \tag{43}$$

In other words:

$$\frac{(C\hat\beta - c)^T[C(X^TX)^{-1}C^T]^{-1}(C\hat\beta - c)}{\sigma^2} \sim \chi^2_q \tag{44}$$

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

$$\frac{\hat\sigma^2}{\sigma^2} \sim \frac{\chi^2_{n-p}}{n-p} \Leftrightarrow \frac{\hat\sigma^2(n-p)}{\sigma^2} \sim \chi^2_{n-p} \tag{45}$$

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 $X = \frac{(C\hat\beta-c)^T[C(X^TX)^{-1}C^T]^{-1}(C\hat\beta-c)}{\sigma^2}$, $Y = \frac{\hat\sigma^2(n-p)}{\sigma^2}$, and setting the degrees of freedom to $v = q$ and $w = n-p$:

$$\frac{X/v}{Y/w} = \frac{\frac{(C\hat\beta-c)^T[C(X^TX)^{-1}C^T]^{-1}(C\hat\beta-c)}{\sigma^2}/q}{\frac{\hat\sigma^2(n-p)}{\sigma^2}/(n-p)} \tag{46}$$

Simplifying:

$$\frac{(C\hat\beta - c)^T[C(X^TX)^{-1}C^T]^{-1}(C\hat\beta - c)}{q\hat\sigma^2} \sim F_{q,n-p} \tag{47}$$

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:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \tag{48}$$

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

$$\frac{\hat{\beta}_i - c_i}{\hat{\sigma}\sqrt{g_{ii}}} \sim t_{n-p} \tag{49}$$

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:

$$x^T x = x^T K x + x^T (I - K) x \tag{50}$$

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:

$$x^T x = \sum x^T K_i x \tag{51}$$

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:

$$y^T y = \underset{S_r = e^T e}{\underset{\uparrow}{\underline{y^T M y}}} + \underset{\hat{\beta}^T (X^T X)\hat{\beta}}{\underset{\uparrow}{\underline{y^T (I - M) y}}} \tag{52}$$

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:

$$y^T y = \underset{e^T e \sim \sigma^2 \chi^2_{n-p}}{\underset{\uparrow}{\underline{y^T M y}}} \qquad + \qquad \underset{\sim \sigma^2 \chi^2_p \text{ iff } X\beta = 0, i.e., \beta = 0}{\underset{\uparrow}{\underline{y^T (I - M) y}}} \tag{53}$$

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:

$$C = (0_{p-q} I_q) \quad c = 0, \text{ and } \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \tag{54}$$

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:

$$(C\hat{\beta} - c)^T [C(X^T X)^{-1} C^T]^{-1} (C\hat{\beta} - c) \sim \sigma^2 \chi^2_q \tag{55}$$

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

$$\hat{\beta}_2^T [C(X^T X)^{-1} C^T]^{-1} \hat{\beta}_2 \sim \sigma^2 \chi^2_q \tag{56}$$

We can rewrite this as: $\hat{\beta}_2^T G_{qq}^{-1} \hat{\beta}_2$, where $G_{qq} = C(X^T X)^{-1} C^T$ (G_{qq} should not be confused with g_{ii}) is a $q \times q$ submatrix of $G = (X^T X)^{-1}$.
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:

$$\frac{\hat{\beta}_2^T C(X^T X)^{-1} C^T \hat{\beta}_2}{q\hat{\sigma}^2} = \frac{\hat{\beta}_2^T G \hat{\beta}_2}{q\hat{\sigma}^2} \sim F_{q,n-p} \tag{57}$$

Sums of squares:
We can construct three idempotent matrices:

- $M = I_n - X(X^T X)^{-1} X^T$
- $M_1 = X(X^T X)^{-1} X^T - [X(X^T X)^{-1} C^T] \underset{G}{\underset{\uparrow}{\underline{[C(X^T X)^{-1} C^T]}}}^{-1} [C(X^T X)^{-1} X^T]$

  (that is: $M_1 = X(X^T X)^{-1} X^T - M_2$)
- $M_2 = [X(X^T X)^{-1} C^T] \underset{G}{\underset{\uparrow}{\underline{[C(X^T X)^{-1} C^T]}}}^{-1} [C(X^T X)^{-1} X^T]$

Note that $M + M_1 + M_2 = I_n$ and $MM_1 = MM_2 = M_1 M_2 = 0$. 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. $S_1 = y^T(X_1^TX_1)^{-1}X_1^Ty = (\sum y)^2/n = n\bar{y}^2$
2. $S_r = y^Ty - \hat{\beta}^TX^TX\hat{\beta}$
3. $S_2 = y^Ty - S_1 - S_r = \hat{\beta}^TX^TX\hat{\beta} - n\bar{y}^2$

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**)	$S_{yy} = (y-\bar{y})^T(y-\bar{y}) = y^Ty - n\bar{y}^2$	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 (`stdres` in R)

Recall that $Var(e) = \sigma^2 M$, where $M = I_n - X(X^TX)^{-1}X^T$ 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**:

$$e_i* = \frac{e_i}{\sqrt{m_{ii}}} \tag{58}$$

Note: $Var(e_i*) = \sigma^2$ because $e_i \sim N(0, \sigma^2 m_{ii})$, therefore $e_i = \frac{e_i}{\sqrt{m_{ii}}}* \sim N(0,\sigma^2)$.
The **standardized residuals** are

$$s_i = \frac{e_i*}{\hat{\sigma}} \tag{59}$$

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 $\mid s_i \mid > t_{crit}$ where t_{crit} is the critical t-value.

Standardized deletion residuals (`studres` in R)

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

$$\hat{\beta}_{-i} = (X_{-i}^TX_{-i})^{-1}X_{-i}^Ty_{-i} \tag{60}$$

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

$$s_{-i} = \frac{e_i}{\hat{\sigma}_{-i}\sqrt{m_{ii}}} \tag{61}$$

We can compute s_{-i} from s_i:

$$s_{-i} = \frac{s_i\sqrt{n-p-1}}{\sqrt{n-p-s_i^2}} \sim t_{n-p-1} \tag{62}$$

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

$$1-(1-\beta)^n = \alpha \Leftrightarrow \beta = 1-(1-\alpha)^{1/n} \tag{63}$$

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. For small samples, use scaled or standardized residuals if sample size is small (not sure why).
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 $X^Te = 0$. (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 $M = I_n - X(X^TX)^{-1}X^T$. Define a hat matrix $H = I - M = X(X^TX)^{-1}X^T$. 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 $h_{ii} = x_i^T(X^TX)^{-1}x_i$. The measure for leverage is:

$$h_{ii} = 1 - m_{ii} \tag{64}$$

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)):

$$h_{ii} = tr(x_i^T(X^TX)^{-1}x_i) = tr(x_i^Tx_i(X^TX)^{-1}) \tag{65}$$

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

$$\sum_{i=1}^n h_{ii} = tr(X^TX(X^TX)^{-1}) = tr(I_p) = p \quad mean(h_{ii}) = p/n \tag{66}$$

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:

$$h_{ii} = n^{-1} + \frac{(x_i - \bar{x})^2}{S_{xx}} \tag{67}$$

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.

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{-i})^T (X^T X)^{-1} (\hat{\beta} - \hat{\beta}_{-i})}{p\hat{\sigma}^2} = \frac{s_i^2 h_{ii}}{p(1-h_{ii})} \tag{68}$$

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 $Var(z)$ is (approximately) constant.
Expand $f(\cdot)$ in a Taylor series expansion, keeping only the first-order term:

$$z = f(y) \approx f(\mu) + (y - \mu) f'(\mu) \tag{69}$$

Then: $E(z) = f(\mu)$ and $Var(z) = g(\mu) f'(\mu)^2$. The variance needs to be constant at, say, k^2:

$$k^2 = g(\mu) f'(\mu)^2 \Rightarrow f'(\mu) = \frac{k}{\sqrt{g(\mu)}} \tag{70}$$

So,

$$f(\mu) = \int f'(\mu) = k \int \frac{1}{\sqrt{g(\mu)}} \tag{71}$$

$$\boxed{f(\mu) = k \int [\sqrt{g(\mu)}]^{-1/2} \, d\mu} \tag{72}$$

Example 1: Let $g(\mu) = a\mu$; then $f(\mu) = 2k\sqrt{\frac{\mu}{a}}$. So, $z = \sqrt{\mu}$.
Example 2: Let $g(\mu) = a\mu^2$; then $f(\mu) = k\sqrt{\frac{1}{a}} \log \mu$. So, $z = \log \mu$.
Estimating a transformation: $\exists \lambda$ such that

$$f_\lambda(y_i) = x_i^T \beta + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \tag{73}$$

We use maximum likelihood estimation to estimate λ. Note that
$L(\beta_\lambda, \sigma_\lambda^2, \lambda; y) \propto$

$$(\frac{1}{\sigma})^n \exp[-\frac{1}{2\sigma^2} \sum [f_\lambda(y_i) - x_i^T \beta]^2] \underbrace{[\prod f'_\lambda(y_i)]}_{\text{Jacobian}} \tag{74}$$

For fixed λ, we estimate $\hat{\beta}$ and $\hat{\sigma}^2$ in the usual MLE way, and then we turn our attention to λ:

$$L(\hat{\beta}_\lambda, \hat{\sigma}_\lambda^2, \lambda; y) = S_\lambda^{-n/2} \prod f'_\lambda(y_i) \tag{75}$$

Taking logs:

$$\ell = c - \frac{n}{2} \log S_\lambda + \sum \log f'_\lambda(y_i) \tag{76}$$

Box-Cox family:

$$f_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases} \tag{77}$$

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.
Box-Cox by hand:
Since $f_\lambda = \frac{y^\lambda - 1}{\lambda}$, it follows that $f'_\lambda(y) = y^{\lambda - 1}$.
Now, for different λ you can figure out the log likelihoods by hand by solving this equation:

$$\ell = c - \frac{n}{2} \log \underbrace{S_\lambda}_{\text{Residual sum of squares}} + (\lambda - 1) \sum \log(y_i) \tag{78}$$

Factors

Overcoming multicollinearity through parameterization

to-do: multicollinearity explanation
If the model matrix X is not full rank (this is true when we include a column for the intercept), then we can put constraints on the predictors (through parameterization). E.g., treatment contrasts (corner-point constraints), sum contrasts, etc.
to-do constraints on p. 94-95

Model selection

S_r and R^2 can't be used for model selection: "S_r will always decrease when we add more regressor variables, so the best fitting model is always the full model which contains all the possible regressor variables - so S_r is not a good model selection tool. For the same reason, the coefficient of determination R^2 is not a useful measure in model selection. R^2 will not decrease as the number of parameters increases (i.e. it is a non-decreasing function of the number of parameters in the model)."
Penalized likelihood methods for model comparison:
We can compare models using log likelihood:

$$\ell = n \log \hat{\sigma}^2 + z(p) \tag{79}$$

where z is some penalty function. "Then we declare that the optimal model is that which minimizes ℓ. We can think of z as an ad hoc adjustment that tries to give simpler models credit for having fewer regressor variables."
AIC etc cannot be used to compare across datasets, but can be used to compare non-nested models (cf. ANOVA, which allows only nested models to be compared).
AIC: here, $z(p) = 2p$. To calculate AIC:

$$AIC = 2p + n \log \frac{S_r}{n} \tag{80}$$

[Note: does not match up with the AIC function output in R.]
Where does AIC come from? From the fact that maximum likelihood of $\hat{\sigma}^2$ is:

$$L(\sigma) \propto (\hat{\sigma}^2)^{-n/2} \tag{81}$$

$-2 \times loglik = n \log \hat{\sigma}^2 = n \log \frac{S_r}{n}$ is a good model selection tool: smaller values (smaller S_r) will mean better fit.

BIC: $z(p) = p \log n$. This penalty will be large for large n, compared to AIC.
Mallow's C_p:

$$C_p = \frac{S_r}{\hat{\sigma}_f^2} - n + 2p_r \quad (82)$$

$\hat{\sigma}_f^2$ is the residual mean square of the full model, S_r is residual sums of squares of reduced model, p_r is the number of regressors in the reduced model.
We want a small C_p and $C_p \approx p$.
Best subsets method: in leaps library, regsubsets command.

```
b<-regsubsets(model specification)
summary(b)$rsq, $cp, $bic
```

Backward elimination: fit full model, and remove the t-value that's smallest, and so on. Forward elimination goes in the other direction.
Stepwise selection: step function in MASS. Incrementally add a predictor as above, but once one is added, try removing other predictors with smallest t-value. Repeat until nothing can be added or deleted.

```
step(fm,scope=list(upper/lower=formula),
        direction="forward/backward/both")
```

Generalized least squares

Let $Var(\epsilon) = \sigma^2\Sigma$, where Σ is known and non-singular. If $\Sigma \neq I_n$ then we either have correlation or non-equal variance or both. If Σ is known, we only need to estimate σ and we are back in least squares theory, with some modification:

$$y \sim N(X\beta, \sigma^2\Sigma) \quad (83)$$

Likelihood $L(\beta, \sigma^2; y)$ is now:

$$(2\pi)^{-n/2} \mid \sigma^2\Sigma \mid^{-1/2} \exp[-\frac{1}{2\sigma^2}(y - X\beta)^T\Sigma^{-1}(y - X\beta)] \quad (84)$$

The MLE of β minimizes:

$$S = (y - X\beta)^T\Sigma^{-1}(y - X\beta) \quad (85)$$

instead of $(y - X\beta)^T(y - X\beta)$.
Least squares estimators:
$\hat{\beta} : (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y$
$E(\hat{\beta}) = \beta$
$Var(\hat{\beta}) = \sigma^2(X^T\Sigma^{-1}X)^{-1}$
Estimator of σ^2 is $\frac{S_v}{n-p}$, where
$S_v = y^T\Sigma^{-1}y - \hat{\beta}^TX^T\Sigma^{-1}X\hat{\beta} = y^T\Sigma^{-1}y - \hat{\beta}^TX^T\Sigma^{-1}y$.

Weighted least squares

Suppose $\Sigma = diag(c1, \ldots, cn)$ (uncorrelated, but not homoscedastic) and so $\Sigma^{-1} = diag(1/c1, \ldots, 1/cn)$. Let $w_i = 1/c_i$.
The sum of squares will be $S = \sum_{i=1}^n w_i(y_i - x_i^T\beta)^2$. So each squared residual is weighted by that observation's variance; observations with large variance are less reliable, and are down-weighted.
Compared to $X^TX = \sum x_ix_i^T$ in WLS we have $X^T\Sigma^{-1}X = \sum w_ix_ix_i^T$. And instead of $X^Ty = \sum x_iy_i$ in WLS we have $X^T\Sigma^{-1}y = \sum w_ix_iy_i$.
So
$\hat{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y = (\sum w_ix_ix_i^T)^{-1}(\sum w_ix_iy_i)$
"Weighted LS is appealing because it allows us to adjust for different (known) variances in the observations in an intuitive way that is easy to implement. The variances of the different observations might be known from pilot or previous studies" (or are estimated from data).
[SV: I don't get this "known variance" business. How can we ever **know** what the variance is? Pilot or previous data will only yield estimates.]
The main disadvantage of WLS is that weights have to specified in advance.

OLS vs WLS

With dataset wls1.txt, if we fit:

```
summary(m0<-lm(y~X.x,data))
summary(m0.wls<-lm(y~X.x,data,
                   weights=I(1/X.x^2)))
```

The coefs will be the same in each, but SEs of coefs will be smaller in WLS fit (because S_r will be down-weighted.
Effect of scaling weights:
Multiplying the weights by some constant will change residual standard error, but leaves SEs and coefs unchanged. This is because $Var(\hat{\beta}) = \sigma^2(X^T\Sigma^{-1}X)^{-1}$, so whatever factor σ^2 gets multiplied by, it will be cancelled out because it also appears in Σ^{-1}.
Differently put:
"Let w_i' be the scaled weights, $\hat{\sigma}'$ be the residual standard error in the analysis with the scaled weights, S_r' be the residual sum of squares for the scaled analysis and $(\Sigma^{-1})'$ be the weight matrix for the scaled analysis. Then if $w_i' = 16w_i$, we see that $\hat{\sigma}' = 4\hat{\sigma}$ since $S_r' = 16S_r$."
The SEs will not change because: "If $(\Sigma^{-1})' = 16\Sigma^{-1}$ and $\hat{\sigma}' = 4\hat{\sigma}$ then
$Var(\hat{\beta}') = (\hat{\sigma}^2)'(X^T(\Sigma^{-1})'X)^{-1} = (\hat{\sigma}^2)(X^T(\Sigma^{-1})X)^{-1}$.
So the standard errors dont change."

```
summary(m0.wls<-lm(y~X.x,data,
                   weights=I(16*1/X.x^2)))
```

1. If you get the weights wrong, SEs will increase. So, if unsure about weights use OLS.

2. "If looking at the standardized residuals shows that there are observations that may be outliers and they reside in regions that will be given high weight then it may be safer to use OLS rather than WLS."

3. One can estimate the weights from the data, but one needs lots of replicates for this, with fewer replicates the SEs will increase.

4. If you don't have enough replicates, you can group x values close to each other.

5. Outliers can dramatically influence estimates in WLS.

Conclusion: WLS is a powerful tool, if weights are known, but outliers must be studied carefully.

Using group means in WLS with replicated data

Without replicates in the data, we have:

$$X^Ty = \sum_i^n x_iy_i \quad (86)$$

If we have k replicates:

$$X^Ty = \sum_{i=1}^k\sum_{j=1}^{n_i} x_iy_{ij} = \sum_{i=1}^k x_i\sum_{j=1}^{n_i} y_{ij} = \sum_{i=1}^k n_ix_i\bar{y}_i \quad (87)$$

This is equivalent to having $\bar{y}_i$ as observations for the replicate sets, and n_i as weights in WLS. If we had $\bar{y}_i$ as observations, then their variances would be σ^2/n_i, so we would have unequal variances and would use WLS with $w_i = n_i$. **Example**: tractor data.

Replication

Define replicates here as repeated measurements that are mutually independent (cf. replicates which are not independent, as in linear mixed model theory). Since all x_i^T within a replicate set come from the same distribution, their variance should be an estimate of σ^2.

Let y_{ij} be the jth observation in the ith replicate set, where $j = 1, \ldots, n_i$ and n_i is the size of the ith replicate set $(i = 1, \ldots, k)$. When $n_i = 1$ we have no replication, and for higher values we have replication. *Within* each replicate, we can produce an estimator:

$$\hat{\sigma}^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = (n_i - 1)^{-1} S_i \qquad (88)$$

$\bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$ is the mean of the replicate set.

$S_i \sim \sigma^2 \chi^2_{n_i - 1}$.

In a one-factor model, $S_r = \sum_{i=1}^k S_i$ and dfs are $n - k = \sum_{i=1}^k (n_i - 1)$. So, in the general case, $df_r = n - k$, $S_r = \sum_{i=1}^k S_i \sim \chi^2_{df_R}$. The ratio $\hat{\sigma}^2 / df_R$ is an unbiased estimator of σ^2.

The distributional fact being used here is that the sum of independent chi-squared distributions has a chi-squared distribution, and the degrees of freedom is the sum of dfs of the RVs being summed.

The replication estimator: $\hat{\sigma}^2_R = \frac{S_R}{df_R}$ is the replication estimator of σ^2.

"The $\hat{\sigma}^2_R$ is independent of the particular form that we have used for the model. We obtain the same replication sum of squares with the same degrees of freedom whether we postulate a linear, quadratic, or some other relationship between Maintenance cost and Age." (to-do: don't get this). This is the strength of the replication sum of squares.

The S_R is in general not equal to S_r.

Partitioning replication sum of squares

to-do

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/