Revision notes - ST3131

Ma Hongqiang

November 18, 2018

Contents

1 Simple Linear Regression

1.1 Simple Linear Regression Model

Definition 1.1 (Simple Linear Regression Model).
Suppose x, y are two variables. The simple linear regression model makes the following assumption on the relation of y with respect to x:

1. $y = \beta_0 + \beta_1 x + \varepsilon$
2. $E(y \mid x) = \beta_0 + \beta_1 x$
3. $\mathrm{Var}(y \mid x) = \mathrm{Var}(\beta_0 + \beta_1 x + \varepsilon) = \mathrm{Var}(\varepsilon) = \sigma^2$

1.2 Least Square Estimation of Parameters

Suppose we have data $\{(y_i, x_i)\}$ for $i = 1, 2, \ldots, n$. The method of least square estimation estimates β_0 and β_1 such that the sum of squares of the differences between the observations y_i and the striaght line is minimum.
Essentially, the sum can be expressed as

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

and we need to minimise the sum with respect to β_0, β_1, so we have the following result

$$\frac{\partial S}{\partial \beta_0}|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

and

$$\frac{\partial S}{\partial \beta_1}|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

where $\hat{\beta}_0, \hat{\beta}_1$ are estimator of β_0, β_1. Notably, after rearranging the first equation, we have

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y}$$

Solving the two equations, we have

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

where

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

and

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = \sum_{i=1}^{n} y_i (x_i - \bar{x})$$

from which, $\hat{\beta}_0$ can be easily obtained.

Definition 1.2 (Residual).
The ith residual of the data e_i is defined as

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Theorem 1.1 (Unbiasness of $\hat{\beta}_0$ and $\hat{\beta}_1$).
Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimator.
This can be easily proven by writing $\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$ where $c_i = \frac{x_i - \bar{x}}{S_{xx}}$, and take expectation.
The unbiasness of $\hat{\beta}_0$ follows immediately.
Notably, some important intermediate results are

- $\sum_{i=1}^{n} c_i = 0$
- $\sum_{i=1}^{n} c_i x_i = 1$

Note, for any $\tilde{\beta}_1 = \mathbf{c} \cdot \mathbf{y}$, as long as $\mathbf{c} \cdot \mathbf{1} = 0$ and $\mathbf{c} \cdot \mathbf{x} = 1$, we will have $\tilde{\beta}_1$ to be unbiased.

Theorem 1.2 (Variance of $\hat{\beta}_0$ and $\hat{\beta}_1$).
We have

- $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$
- $\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$

Theorem 1.3 (Gauss-Markov Theorem).
Gauss-Markov Theorem suggusts that for the simple linear regression model mentioned before, with assumption $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$ and uncorrelated errors, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and have minimum variance whenm compared with all other unbiased estimators that are linear combinations of the y_i.

Therefore, least square estimators are the best linear unbiased estimators.

Theorem 1.4 (Properties of Least Square Fit).

1. $\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0$ (from $\frac{\partial S}{\partial \beta_0}$)
2. The least-square regression line always passes through the centroid, i.e., $(\bar{y}, \bar{x})$ of the data.
3. $\sum_{i=1}^{n} x_i e_i = 0$ (from $\frac{\partial S}{\partial \beta_1}$)
4. $\sum_{i=1}^{n} \hat{y}_i e_i = 0$ (from (1) and (3))

1.3 Esimtation of σ^2

Definition 1.3 (Estimation of σ^2)**.**
We define sum of squares of the residue, SS_{Res} to be

$$SS_{\text{Res}} := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Using $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, we can arrive at[1]

$$SS_{\text{Res}} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy}$$

Here, we define corrected sum of squares of response, SS_{T} to be

$$SS_T := \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

Therefore,

$$SS_{\text{Res}} = SS_{\text{T}} - \hat{\beta}_1 S_{xy}$$

It is known *from textbook* or here that the expected value of SS_{Res} is

$$E(SS_{\text{Res}}) = (n-2)\sigma^2$$

Therefore, **residual mean square** MS_{Res}

$$MS_{\text{Res}} := \frac{SS_{\text{Res}}}{n-2} := \hat{\sigma}^2$$

is an **unbiased estimator** of σ^2. where as $\hat{\sigma}$ is called the **standard error of regression**.

Remark: $\hat{\sigma}^2$ is model dependent due to the term $\hat{\beta}_1 S_{xy}$.

Definition 1.4 (Alternative Form of Simple Linear Regression)**.**
Simple arithmetic gives that the original linear regression model is equivalent to

$$y_i = \beta_0' + \beta_1(x_i - \bar{x}) + \varepsilon_i$$

where $\beta_0' = \beta_0 + \beta_1\bar{x} = \bar{y}$.
The fitted model is

$$\hat{y} = \bar{y} + \hat{\beta}_1(x - \bar{x})$$

A nice property of this model is that

$$\text{cov}(\hat{\beta}_0', \hat{\beta}_1) = 0$$

[1] via $(y_i - \hat{y}_i)^2 = ((y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x}))^2$

1.4 Hypothesis Testing on Slope and Intercept

In this subsection, we further assume: errors are normally and independently distributed with mean 0 and variance σ^2, i.e.,

$$\varepsilon_i \sim N(0, \sigma^2)$$

Theorem 1.5 (Testing $\hat{\beta}_1$)**.**
Here, we are testing null hypothesis

$$H_0 : \beta_1 = \beta_{10}$$

against alternative

$$H_1 : \beta_1 \neq \beta_{10}$$

where $y_i \sim NID(\beta_0 + \beta_1 x_i, \sigma^2)$.
Since $\hat{\beta}_1$ is a linear combination of y_i's, $\hat{\beta}_1$ is also normally distributed with mean β_1 and variance $\frac{\sigma^2}{S_{xx}}$.
After normalizing, $Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0, 1)$ if **the null hypothesis** $H_0 : \beta_1 = \beta_{10}$ is **true**.

However, the exact σ^2 is not available. Therefore, we make use of MS_{Res}, which is an unbiased estimator of σ^2.
It is known *from textbook* that $\frac{(n-2)MS_{\text{Res}}}{\sigma^2}$ follows a χ^2_{n-2} distribution. Also, MS_{Res} and β_1 are independent. Therefore,

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MS_{\text{Res}}}{S_{xx}}}} \sim t_{n-2}$$

follows a t_{n-2} distribution if the **null hypothesis** $H_0 : \beta_1 = \beta_{10}$ is true.
Here, the standard error of the slope is $\text{se}(\hat{\beta}_1) = \sqrt{\frac{MS_{\text{Res}}}{S_{xx}}}$, so we can write $t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\text{se}(\hat{\beta}_1)}$.
We should reject the null hypothesis if $|t_0| > t_{\frac{\alpha}{2}, n-2}$

Theorem 1.6 (Testing of $\hat{\beta}_0$)**.**
Here, we are testing null hypothesis

$$H_0 : \beta_0 = \beta_{00}$$

against alternative hypothesis

$$H_1 : \beta_0 \neq \beta_{00}$$

Similarly, we can formulate the test statistics:

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{\text{Res}}(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}} = \frac{\hat{\beta}_0 - \beta_{00}}{\text{se}(\hat{\beta}_0)}$$

where the stanard error of the intercept is $\text{se}(\hat{\beta}_0) = \sqrt{MS_{\text{Res}}(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}$.
We reject the null hypothesis H_0 if $|t_0| > t_{\frac{\alpha}{2}, n-2}$.

There is an important special case where $H_0 : \beta_1 = 0$. Failing to reject this H_0 is equivalent to saying there is *no* linear relationship between y and x.

1.5 Analysis of Variance(ANOVA)

Definition 1.5 (Model Sum of Squares SS_{R})**.**
We identify the corrected value of response $y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$. Therefore,

$$SS_{\mathrm{T}} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

Furthermore, the third term is 0[2]
We arrive at

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

which is

$$SS_{\mathrm{T}} = SS_{\mathrm{R}} + SS_{\mathrm{Res}}$$

where $SS_{\mathrm{R}} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is known as the **regression/model sum of squares**.
From previous result, we have

$$SS_{\mathrm{R}} = \hat{\beta}_1 S_{xy}$$

Furthermore, we have the following results

Theorem 1.7.

- $SS_{\mathrm{Res}} = (n-2)MS_{\mathrm{Res}}$ follows a χ^2_{n-2} distribution.
- If the null hypothesis $H_0 : \beta_1 = 0$ is true, then $\frac{SS_{\mathrm{R}}}{\sigma^2}$ follows a χ^2_1 distribution.
- SS_{Res} and SS_{R} are independent.

Therefore, under H_0,

$$F_0 = \frac{SS_{\mathrm{R}}/df_{\mathrm{R}}}{SS_{\mathrm{Res}}/df_{\mathrm{Res}}} = \frac{MS_{\mathrm{R}}}{MS_{\mathrm{Res}}}$$

follows the $F_{1,n-2}$ distribution. Therefore, we reject H_0 if $F_0 > F_{\alpha,1,n-2}$. Furthermore, we have

$$E(MS_{\mathrm{Res}}) = \sigma^2$$

and

$$E(MS_{\mathrm{R}}) = \sigma^2 + \beta_1^2 S_{xx}$$

So, suppose the H_0 is rejected, i.e., $\beta_1 \neq 0$, F_0 will follow a noncentral F distribution with 1 and $n-2$ degress of freedom and a non-centrality parameter $\lambda = \frac{\beta_1^2 S_{xx}}{\sigma^2}$.

Theorem 1.8 (Confidence Intervals on statistics)**.**
$100(1-\alpha)$ percent confidence interval(CI) on the slope β_1

$$\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\mathrm{se}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\mathrm{se}(\hat{\beta}_1)$$

[2] by breaking down the first bracket into 2 terms and use existing result in Theorem 1.4

$100(1-\alpha)$ percent CI on the intercept β_0 is

$$\hat{\beta}_0 - t_{\frac{\alpha}{2},n-2}\mathrm{se}(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\frac{\alpha}{2},n-2}\mathrm{se}(\hat{\beta}_0)$$

$100(1-\alpha)$ percent CI on σ^2

$$\frac{(n-2)MS_{\text{Res}}}{\chi^2_{\frac{\alpha}{2},n-2}} \leq \sigma^2 \leq \frac{(n-2)MS_{\text{Res}}}{\chi^2_{1-\frac{\alpha}{2},n-2}}$$

Theorem 1.9 (Interval Estimation of Mean Response $E(y)$)**.**
We have

$$\hat{E(y \mid x_0)} := \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$\mathrm{Var}(\hat{\mu}_{y|x_0}) = \mathrm{Var}(\bar{y} + \hat{\beta}_1(x_0 - \bar{x})) = \sigma^2[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}]$$

Furthermore, $\hat{\mu}_{y|x_0}$ is a normally distributed random variable as it is a linear combination of observation y_i. As a result

$$\frac{\hat{\mu}_{y|x_0} - E(y \mid x_0)}{\sqrt{MS_{\text{Res}}(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}})}}$$

is t with $n-2$ degrees of freedom.
$100(1-\alpha)$ percent CI on the mean response at the point $x = x_0$ is

$$\hat{\mu}_{y|x_0} - t_{\frac{\alpha}{2},n-2}\sqrt{MS_{\text{Res}}(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}})} \leq E(y \mid x_0) \leq \hat{\mu}_{y|x_0} + t_{\frac{\alpha}{2},n-2}\sqrt{MS_{\text{Res}}(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}})}$$

1.6 Prediction of New Observations

Theorem 1.10 (Prediction Interval for Future Observation y_0)**.**
$\psi = y_0 - \hat{y_0}$ is normally distributed with mean 0 and variance

$$\sigma^2[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}]$$

And $100(1-\alpha)$ percent prediction interval on a future y_0 is

$$\hat{\mu}_{y|x_0} - t_{\frac{\alpha}{2},n-2}\sqrt{MS_{\text{Res}}(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}})} \leq E(y \mid x_0) \leq \hat{\mu}_{y|x_0} + t_{\frac{\alpha}{2},n-2}\sqrt{MS_{\text{Res}}(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}})}$$

Furthermore, a $100(1-\alpha)$ percent prediction interval on the **mean** of m future observations, $\bar{y}_0$ on the response at $x = x_0$ is

$$\hat{\mu}_{y|x_0} - t_{\frac{\alpha}{2},n-2}\sqrt{MS_{\text{Res}}(\frac{1}{m} + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}})} \leq E(y \mid x_0) \leq \hat{\mu}_{y|x_0} + t_{\frac{\alpha}{2},n-2}\sqrt{MS_{\text{Res}}(\frac{1}{m} + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}})}$$

1.7 Coefficient of Determination

Definition 1.6 (Coefficient of Determination)**.**
The coefficient of determination $R^2$is defined as

$$R^2 = \frac{SS_{\mathrm{R}}}{SS_{\mathrm{T}}} = 1 - \frac{SS_{\mathrm{Res}}}{SS_{\mathrm{T}}}$$

We can argue that R^2 is between 0 and 1 as $\hat{\beta}_0$, $\hat{\beta}_1$ are chosen to minimise SS_{Res}, whereas SS_{T} is the SS_{Res} when the regression line is $\hat{y} = \bar{y}$, an unoptimal case.
Here SS_{T} is teh measure of the variability in y without considering the effect of the regressor variable x.
SS_{R} is a measure of the variability in y after x has been considered.
Therefore, R^2 is often called the proportion of variation explained by teh reression x.

However, R^2 statistic should be used with caution.

1.8 Considerations in Use of Regression

- Regression models are intended as interpolation equations over the range of the regressor variable used to fit the model.
- The disposition of the x values plays an important role in the least-square fit.
- **Outliers** are observations that differ considerably from the rest of the data.
- Strong relationship between two variables does not imply that the variables are related in any *causal* sense. Casuality implies necessary correlation.
- In some application of regression, the value of the regressor variable x is required to predict y is unknown.

1.9 Regression Through the Origin

Here, we introdue a special case of simple linear regression, the **no-intercept** model.

$$y_i = \beta_1 x + \epsilon$$

By solving normal equation, we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

$\hat{\beta}_1$ is stil an unbiased estimator for β.
Also, note the change in numerator in the estimator of σ^2.

$$\hat{\sigma}^2 = MS_{\mathrm{Res}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-1} = \frac{\sum_{i=1} n y_i^2 - \hat{\beta}_1 \sum_{i=1}^{n} y_i x_i}{n-1}$$

and $100(1-\alpha)$ percent CI on β_1 is

$$\hat{\beta}_1 - t_{\frac{\alpha}{2},n-1}\sqrt{\frac{MS_{\text{Res}}}{\sum_{i=1}^n x_i^2}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\frac{\alpha}{2},n-1}\sqrt{\frac{MS_{\text{Res}}}{\sum_{i=1}^n x_i^2}}$$

$100(1-\alpha)$ percent CI on $E(y \mid x_0)$, the mean response at $x = x_0$ is

$$\hat{\mu}_{y|x_0} - t_{\frac{\alpha}{2},n-1}\sqrt{\frac{x_0^2 MS_{\text{Res}}}{\sum_{i=1}^n x_i^2}} \leq E(y \mid x_0) \leq \hat{\mu}_{y|x_0} + t_{\frac{\alpha}{2},n-1}\sqrt{\frac{x_0^2 MS_{\text{Res}}}{\sum_{i=1}^n x_i^2}}$$

$100(1-\alpha)$ percent prediction interval on a future observation at $x = x_0$ is

$$\hat{y_0} - t_{\frac{\alpha}{2},n-1}\sqrt{MS_{\text{Res}}(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2})} \leq E(y \mid x_0) \leq \hat{y_0} + t_{\frac{\alpha}{2},n-1}\sqrt{MS_{\text{Res}}(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2})}$$

Coefficient of determination R_0^2 is

$$R_0^2 = \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2}$$

R_0^2 indicates the proportion of variability around the origin accounted by the regression.

1.10 Estimation of Statistics by Maximum Likelihood

Here, we assume that ε^2 is normally distributed. Therefore, for n data points (y_i, x_i), the likelihood function is

$$L(y_i, x_i, \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_1 x_i)^2} = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\beta_0-\beta_1 x_i)^2}$$

Suppose the maximum likelihood estimator takes the value $\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\sigma}^2$, the we need the partial differentials $\frac{\partial \ln L}{\partial \beta_0}$, $\frac{\partial \ln L}{\partial \beta_1}$, $\frac{\partial \ln L}{\partial \sigma^2}$ to be zero at these values. Solving, we have

$$\tilde{\beta}_0 = b\hat{e}ta_0, \qquad \tilde{\beta}_1 = \hat{\beta}_1$$

However, the estimator of σ^2 is biased:

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2}{n} = \frac{n-2}{n}\hat{\sigma}^2$$

1.11 Case where regressor x is Random

Suppose that x and y are jointly distributed random variables but the form of joint distribution is unknown. It can be shown that all of our previous regression results hold if the following conditions are satisfied:

1. The conditional distribution of y given x is normal with conditional mean $\beta_0 + \beta_1 x$ and conditional variance σ^2.

2. x's are independent random variables whose probability distribution does not involve β_0, β_1 and σ^2.

A special case is that x and y are jointly distributed according to the bivariate normal distribution with density

$$f(y,x) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\{-\frac{1}{2(1-\rho^2)}[(\frac{y-\mu_1}{\sigma_1})^2 + (\frac{x-\mu_2}{\sigma_2})^2 - 2\rho(\frac{y-\mu_1}{\sigma_1})(\frac{y-\mu_2}{\sigma_2})]\}$$

where $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$.
The conditional distribution of y given x is

$$f(y \mid x) = \frac{1}{\sqrt{2\pi}\sigma_{1,2}} \exp[-\frac{1}{2}(\frac{y-\beta_0-\beta_1 x}{\sigma_{12}})^2]$$

We have

$$E(y \mid x) = \beta_0 + \beta_1 x$$

where

$$\beta_0 = \mu_1 - \mu_2\rho\frac{\sigma_1}{\sigma_2}$$

$$\beta_1 = \frac{\sigma_1}{\sigma_2}\rho$$

and

$$\sigma_{1.2}^2 = \sigma_1^2(1-\rho^2)$$

The maximum likelihood estimator is still the same, where

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

and

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Definition 1.7 (Sample Correlation Coefficient)**.**
We define the sample correlation coefficient

$$r = \frac{\sum_{i=1}^n y_i(x_i-\bar{x})}{[\sum_{i=1}^n (x_i-\bar{x})^2 \sum_{i=1}^n (y_i-\bar{y})^2]^{\frac{1}{2}}} = \frac{S_{xy}}{\sqrt{S_{xx}SS_{\mathrm{T}}}}$$

Note, $\hat{\beta}_1 = (\frac{SS_{\mathrm{T}}}{S_{xx}})^{\frac{1}{2}} r$.

A coincidence is that $r^2 = R^2$.

Theorem 1.11 (Testing of ρ)**.**
Suppose we want to test $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$, the test statistic is

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

which follows t_{n-2} distribution if $H_)$ is true.
We reject the null hypothesis H_0 if $|t_0| > t_{\frac{\alpha}{2},n-2}$.

More generally, suppose we want to test $H_0 : \rho = \rho_0$, against $H_1 : \rho \neq \rho_0$, then for *moderately large examples*$(n \geq 25)$, test statistics

$$Z_0 = \frac{\operatorname{arctanh} r - \operatorname{arctanh} \rho_0}{(}n - 3)^{\frac{1}{2}}$$

follows standard normal distribution. H_0 is rejected if $|Z_0| > Z_{\frac{\alpha}{2}}$.
$100(1-\alpha)$ percent CI is

$$\tanh(\operatorname{arctanh} r - \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n-3}}) \leq \rho \leq \tanh(\operatorname{arctanh} r + \frac{Z_{\frac{\alpha}{2}}}{\sqrt{n-3}})$$

2 Multi-linear Regression

Definition 2.1 (Multiple Linear Regression Model)**.**
The multiple linear regression model with k regressor is of the form

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

where β_j, $j = 1, \ldots, k$ are called **partial regression coefficients**.

Sometimes, there are $x_i^{p_i} x_j^{p_j}$ terms inside the model. We call such term "interaction terms".

2.1 Estimation of $\boldsymbol{\beta}$

Suppose we have n data point, we denote $x_i j$ to be the jth regressor variable of the ith experiment and y_i the response. Therefore, we can write the model as such

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y} \in \mathbb{R}^n$, $X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \in \mathbb{R}^{n\times(k+1)}$ where k is the number of regressor variable, $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$ and $\epsilon \in \mathbb{R}^n$.

Essentially, we want to minimize $S(\boldsymbol{\beta}) = \sum_{i=1}^n \varepsilon_i^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Differentiating and solving, we have

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

We can then write the fitted regression model as

$$\hat{y} = \boldsymbol{x}'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_j$$

For all the fitted values $\hat{\boldsymbol{y}}$, we have

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} := \boldsymbol{H}\boldsymbol{y}$$

Here $\boldsymbol{H} := \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is called the **hat matrix**.
This hat matrix is idempotent and symmetric. Also, $I - H$ is also idempotent and symmetric.

As a result, the residual vector $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$.
We can easily show that

Theorem 2.1 (Unbiased Estimation of $\hat{\beta}$)**.**
$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

Therefore, $\hat{\boldsymbol{\beta}}$ is an **unbiased estimator** of $\boldsymbol{\beta}$ if the model is correct.
Also, the covariance matrix of $\hat{\boldsymbol{\beta}}$, a $p \times p$ matrix, is

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Theorem 2.2 (Gauss-Markov Theorem)**.**
$\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$, and also the maximum likelihood estimator.

2.2 Esimation of $\bar{\sigma}^2$

For multiple linear regression, we have following formula for SS_{Res}:

$$SS_{\text{Res}} = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'y$$

It can be shown that SS_{Res} has $n-p$ degrees of freedom where p is the number of parameter estimated in regression model.
Therefore, residual mean square MS_{Res} has the following value

$$MS_{\text{Res}} = \frac{SS_{\text{Res}}}{n-p} := \hat{\sigma}^2$$

It can be shown that $E(\hat{\sigma}^2) = \sigma^2$.
If we further assume $\boldsymbol{\varepsilon}$ is distributed as $N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then

$$L(\boldsymbol{\varepsilon}, \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp(-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})$$

Substituting $\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$, we have

$$\ln L(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

Therefore, we want to minimise $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$, which in turn gives the same $\hat{\boldsymbol{\beta}}$ as the least square estimator, but a biased $\tilde{\sigma}^2$:

$$\tilde{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n}$$

2.3 Hypothesis Testing in Multiple Linear Regression

Here, we assume that our random errors to be independent and follow a normal distribution with mean 0 and variance σ^2.

Theorem 2.3 (Test for Significance of Overall Regression)**.**
Suppose we want to test the null hypothesis that the whole regression does not make sense:

$$H_0 := \beta_1 = \cdots \beta_k = 0$$

against

$$H_1 : \beta_j \neq 0 \text{ for at least one } j$$

It is known that, **if null hypothesis is true**, $\frac{SS_{\text{R}}}{\sigma^2}$ follows a χ^2_k distribution, whereas $\frac{SS_{\text{Res}}}{\sigma^2}$ follows a χ^2_{n-k-1} distribution. Also, as SS_{Res} and SS_{R} are independent.
Therefore, we test

$$F_0 = \frac{SS_{\text{R}}/k}{SS_{\text{Res}}/(n-k-1)} = \frac{MS_{\text{R}}}{MS_{\text{Res}}} \sim F_{k,n-k-1}$$

and reject H_0 if $F_0 > F_{\alpha,k,n-k-1}$.

It is worth noting that, in general

$$E(MS_{\text{R}}) = \sigma^2 + \frac{\boldsymbol{\beta}^{*\prime}\boldsymbol{X}_c'\boldsymbol{X}_c\boldsymbol{\beta}^*}{k\sigma^2}$$

and

$$E(MS_{\text{Res}}) = \sigma^2$$

where $\boldsymbol{\beta}^* = (\beta_1, \dots, \beta_k)'$, a truncated $\boldsymbol{\beta}$ with intercept β_0 removed, and $\boldsymbol{X}_c := \begin{pmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1k} - \bar{x}_k \\ \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{nk} - \bar{x}_k \end{pmatrix}$.
Therefore, if F_0 is large, it is likely that at least one $\beta_j \neq 0$. In this case, F_0 will follow a noncentral F distribution with k and $n-k-1$ degrees of freedom and noncentrality parameter of $\lambda = \frac{\boldsymbol{\beta}^{*\prime}\boldsymbol{X}_c'\boldsymbol{X}_c\boldsymbol{\beta}^*}{\sigma^2}$.
In general, we have the following ANOVA table:

Theorem 2.4 (ANOVA for Significance of Regression in Multiple Regression)**.**
Here,

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_{R}	k	MS_{R}	$\frac{MS_{\text{R}}}{MS_{\text{Res}}}$
Residual	SS_{Res}	$n-k-1$	MS_{Res}	
Total	SS_{T}	$n-1$		

- $SS_{\text{Res}} = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}\boldsymbol{X}'\boldsymbol{y}$
- $SS_{\text{T}} = \boldsymbol{y}'\boldsymbol{y} - \frac{(\sum_{i=1}^n y_i)^2}{n}$.
- $SS_{\text{R}} = SS_{\text{T}} - SS_{\text{Res}} = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \frac{(\sum_{i=1}^n y_i)^2}{n}$.

Definition 2.2 (Adjusted R^2)**.**
In simple linear regression, we have

$$R^2 = \frac{SS_{\text{R}}}{SS_{\text{T}}} = 1 - \frac{SS_{\text{Res}}}{SS_{\text{T}}}$$

Adjusted R^2 is defined in a similar manner:

$$R^2_{\text{Adj}} = 1 - \frac{SS_{\text{Res}}/(n-p)}{SS_{\text{T}}/(n-1)}$$

It is worth noting the numerator is the residual mean square.

2.4 Test on Individual and Subset of Regression Coefficients

Theorem 2.5 (Test on Individual Regression Coefficients)**.**
Suppose we want to test the significance of one particular regressor x_j, we will have null hypothesis

$$H_0 : \beta_j = 0$$

against alternative $H_1 : \beta + j \neq 0$.
The test statistic is

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{\mathrm{se}(\hat{\beta}_j)} \sim t_{n-k-1}$$

where C_{jj} is the diagonal element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ corresponding to $\hat{\beta}_j$.
The null hypothesis H_0 is rejected if $|t_0| > t_{\frac{\alpha}{2},n-k-1}$.

Remark: This is a test of **contribution** of x_j **given that other regressors in the model**. This is a **partial**, or **marginal** test because the regression coefficient $\hat{\beta}_j$ depends on all of the other regressor variables $x_i (i \neq j)$ that are in the model.

2.5 Extra-Sum-Of-Squares Method

Consider the regression model with k regressors: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. We partition $\boldsymbol{\beta} := \begin{pmatrix} \boldsymbol{\beta}_1 \\ - \\ \boldsymbol{\beta}_2 \end{pmatrix}$ such that $\boldsymbol{\beta}_1$ has $(p-r)$ rows, and $\boldsymbol{\beta}_2$ has r rows.
Let null hypothesis H_0 be

$$H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$$

against alternative $H_1 : \boldsymbol{\beta}_2 \neq \boldsymbol{0}$ for the full model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$. The full model has the following properties:

- $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$
- $SS_{\mathrm{R}}(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$ (p degrees of freedom)
- $MS_{\mathrm{Res}} = \frac{\boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}}{n-p}$.

Suppose null hypothesis is true, then we will have a reduced model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$, which has the following properties:

- $\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}$
- $SS_{\mathrm{R}}(\boldsymbol{\beta}_1) = \hat{\boldsymbol{\beta}}_1'\boldsymbol{X}_1'\boldsymbol{y}$ ($p - r$ degree of freedom)

We further define the **regression sum of squares**(SS_{R}) due to $\boldsymbol{\beta}_2$ given that $\boldsymbol{\beta}_1$ is already in the model to be

$$SS_{\mathrm{R}}(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1) = SS_{\mathrm{R}}(\boldsymbol{\beta}) - SS_{\mathrm{R}}(\boldsymbol{\beta}_1)$$

with r degrees of freedom.
This sum of squares is called **extra sum of squares due to** $\boldsymbol{\beta}_2$ because it measures in

the increase of regression sum of squares that results from adding new regressors to a model with existing regressors.
With the above definition, we proceed to test H_0. Since $SS_{\mathrm{R}}(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1)$ is independent of MS_{Res}, the test statistic is

$$F_0 = \frac{SS_{\mathrm{R}}(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1)/r}{MS_{\mathrm{Res}}}$$

And we reject H_0 if $F_0 \geq F_{\alpha,r,n-p}$, concluding that at least one of the parameters in β_2 is not zero.
Remark: If $\boldsymbol{\beta}_2 \neq 0$, then F_0 follows a noncentral F distribution with a noncentrality parameter of

$$\lambda = \frac{1}{\sigma^2}\boldsymbol{\beta}_2'\boldsymbol{X}_2'[\boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\mathbf{1}']\boldsymbol{X}_2\boldsymbol{\beta}_2$$

Here, the maximal power for this test occurs when $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal to one another.

2.6 Special Case of Orthogonal Columns in $\boldsymbol{X}$

Suppose we have $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, if columns of $\boldsymbol{X}_1$ are orthogonal to columns in $\boldsymbol{X}_2$, then the normal equation is reduced to

$$\boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1'\boldsymbol{y} \text{ and } \boldsymbol{X}_2'\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_2'\boldsymbol{y}$$

Therefore

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y} \text{ and } \hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y}$$

The full model will have a SS_{R} of

$$\begin{aligned} SS_{\mathrm{R}}(\boldsymbol{\beta}) &= \hat{\boldsymbol{\beta}}_1'\boldsymbol{X}_1'\boldsymbol{y} + \hat{\boldsymbol{\beta}}_2'\boldsymbol{X}_2'\boldsymbol{y} \\ &= \boldsymbol{y}'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y} + \boldsymbol{y}'\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y} \\ &= SS_{\mathrm{R}}(\boldsymbol{\beta}_1) + SS_{\mathrm{R}}(\boldsymbol{\beta}_2) \end{aligned}$$

2.7 Testing General Linear Hypothesis

Here, let null hypothesis $H_0 : \boldsymbol{T\beta} = \mathbf{0}$, where $\boldsymbol{T}$ is a $m \times p$ matrix of constants, such that only r of the m equations in $\boldsymbol{T\beta} = \mathbf{0}$ are independent.
The full model originally is $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, with $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ and $SS_{\mathrm{Res}}(FM) = \boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}$ with $n-p$ degrees of freedom.
The reduced model is generated by using r independent equations to solve for r of the regression coefficients in the full model in terms of the remaining $p-r$ regression coefficients.
The reduced model is $\boldsymbol{y} = \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}$ where $\boldsymbol{Z} \in \mathbb{R}^{n\times(p-r)}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{(\shortmid-\diagdown)\times\nVdash}$.
Similarly, $\hat{\gamma} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$ and $SS_{\mathrm{Res}}(RM) = \boldsymbol{y}'\boldsymbol{y} - \hat{\gamma}'\boldsymbol{Z}'\boldsymbol{y}$, with $n-p+r$ degrees of freedom.
Due to reduction of regressor variables, certainly we have $SS_{\mathrm{Res}}(RM) \geq SS_{\mathrm{Res}}(FM)$.
To test the hypothesis H_0, let $SS_{\mathrm{H}} = SS_{\mathrm{Res}}(RM) - SS_{\mathrm{Res}}(FM)$ with r degree of freedom.
Therefore, the F statistics is

$$F_0 = \frac{SS_{\mathrm{H}}/r}{SS_{\mathrm{Res}}(FM)/(n-p)}$$

We reject H_0 if $F_0 > F_{\alpha,r,n-p}$.

2.8 Prediction of New Observations

Let $\boldsymbol{x}_0' = [1, x_{01}, \ldots, x_{0k}]$. The **point estimate of the future observation** y_0 at the point $x_{01}, \ldots, x_{0k}$ is $\hat{y}_0 = \boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$.
A $100(1-\alpha)$ percent prediction interval for this future observataion is

$$\hat{y}_0 - t_{\frac{\alpha}{2},n-p}\sqrt{\hat{\sigma}^2(1+\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)} \leq y_0 \leq \hat{y}_0 + t_{\frac{\alpha}{2},n-p}\sqrt{\hat{\sigma}^2(1+\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)}$$

2.9 Hidden Extrapolation in Multiple Regression

Definition 2.3 (Regressor Variable Hull(RVH))**.**
Regressor Variable Hull is the smallest convex set containing all of the original n data points $x_{i1}, \ldots, x_{ik}$, $i = 1, 2, \ldots, n$.

If a point lies inside or on the boundary of the RVH, prediction or estimation involves interpolation, whereas point lying outside RVH requires extrapolation.
Let $h_{\max} = \max h_{ii} \mid \boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Then the set of point $\boldsymbol{x}$ satisfying

$$\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{x}' \leq h_{\max}$$

is an ellipsoid enclosing all points inside the RVH. Therefore, $h = \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{x}' > h_{\max}$ given by $\boldsymbol{x}$ will be outside the ellipsoid enclosing RVH, whereas $h \leq h_{\max}$ will be inside the ellipsoid and *possibly* inside the RVH.

2.10 Standardized Regression Coefficients

We introduce two ways to scale the regression coefficients, namely unit normal scaling and unit length scaling.

2.10.1 Unit Normal Scaling

Denote $s_y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$ and $s_j^2 = \frac{1}{n-1}\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$. Then, we define the transformed variable

$$y_i^* = \frac{y_i - \bar{y}}{s_y}$$

and

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

This transformatino guarantees **all** of the scaled regressor and scaled responses have sample mean equal to 0 and sample variance equal to 1.
Using these new variables, regression model becomes

$$y_i^* = b_1 z_{i1} + b_2 z_{i2} + \cdots + b_k z_{ik} + \varepsilon_i$$

Here, the least square estimate of $b_0 = \bar{y}^* = 0$.
The least square estimator of $\boldsymbol{b}$ is $\hat{\boldsymbol{b}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}^*$.

2.10.2 Unit Length Scaling

Here, we define the transformed variable as

$$y_i^0 = \frac{y_i - \bar{y}}{SS_{\mathrm{T}}^{\frac{1}{2}}}$$

and

$$w_{ij} = \frac{x_{ij} - \bar{x}_j}{s_{jj}^{\frac{1}{2}}}$$

This transformation guarantees **each** new regressow w_j has mean $\bar{w}_j = 0$ and length $\sqrt{\sum_{i=1}^{n}(w_{ij} - \bar{w}_j)^2} = 1$.

The regression model is

$$y_i^0 = b_1 w_{i1} + b_2 w_{i2} + \cdots + b_k w_{ik} + \varepsilon_i$$

and the least square regression coefficient $\hat{\boldsymbol{b}} = (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{y}^0$.

Do note

$$\boldsymbol{W}'\boldsymbol{W} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{12} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1k} & r_{2k} & \cdots & 1 \end{pmatrix}$$

where $r_{ij} = \frac{S_{ij}}{(S_{ii}S_{jj})^{\frac{1}{2}}}$ is the simple correlation between regressor x_i and x_j.

Also, $\boldsymbol{W}'\boldsymbol{y}^0 = \begin{pmatrix} r_{1y} \\ r_{2y} \\ \vdots \\ r_{ky} \end{pmatrix}$, where $r_{jy} = \frac{S_{jy}}{(S_{jj}SS_{\mathrm{T}})^{\frac{1}{2}}}$.

Note, $\boldsymbol{Z}'\boldsymbol{Z} = (n-1)\boldsymbol{W}'\boldsymbol{W}$. Therefore, both methods produce the **same** coefficients. $\hat{\boldsymbol{b}}$ is usually called **standardized regression coefficients**.

The relationship between original and standardized regression coefficients are

$$\hat{\beta}_j = \hat{b}_j\left(\frac{SS_{\mathrm{T}}}{S_{jj}}\right)^{\frac{1}{2}}$$

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{k} \hat{\beta}_j \bar{x}_j$$

2.11 Multicollinearity

Multicollinearity is the near-linear dependence among the regression variables.

Definition 2.4 (Variance Inflation Factors)**.**
The main diagonal elements of the inverse of the $\boldsymbol{X}'\boldsymbol{X}$ matrix in correlation form are often called **variance inflation factors**.

Do note, $\mathrm{VIF}_j = \frac{1}{1-R_j^2}$, where R_j^2 is the coefficient of multiple determination obtained from regressing x_j on the other regressor variables.

2.12 Wrong Sign of Regression Coefficients

The following problems are possible reasons:

1. Range of some regressors is too small.
2. Important regressors have not been included in the model.
3. Multicollinearity is present.
4. Computation Error

3 Model Adequacy Checking

In this section, we assume

- Relationship between response y and regressors is linear, at least approximately.
- Error terms has 0 mean.
- Error term has constant σ^2 variance.
- Errors are uncorelated.
- Errors are normally distributed.

Definition 3.1 (Raw Residuals).
Rar residuals is defined as

$$e_i = y_i - \hat{y}_i$$

Recall, the MS_{Res} is calculated by

$$MS_{\text{Res}} = \frac{SS_{\text{Res}}}{n-p} = \frac{sum_{i=1}^{n}(e_i - \bar{e})^2}{n-p} = \frac{sum_{i=1}^{n} e_i^2}{n-p}$$

Here, the residuals are not independent, since n residuals only have $n-p$ degrees of freedom associated with them.
With residual, we can make plots of them: However, the scale of residuals are largely dependent on the unit of the response. Therefore, we need to scaling them.

Definition 3.2 (Standardized Residuals).
Standardized residuals d_i are defined from e_i by

$$d_i = \frac{e_i}{\sqrt{MS_{\text{Res}}}}$$

Here, the standardized residuals will have mean 0 and *approximately* unit variance. A large standardized residuals($d_j > 3$) could potentially indicates outlier.
Since $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$, where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, by substituting $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, we will have

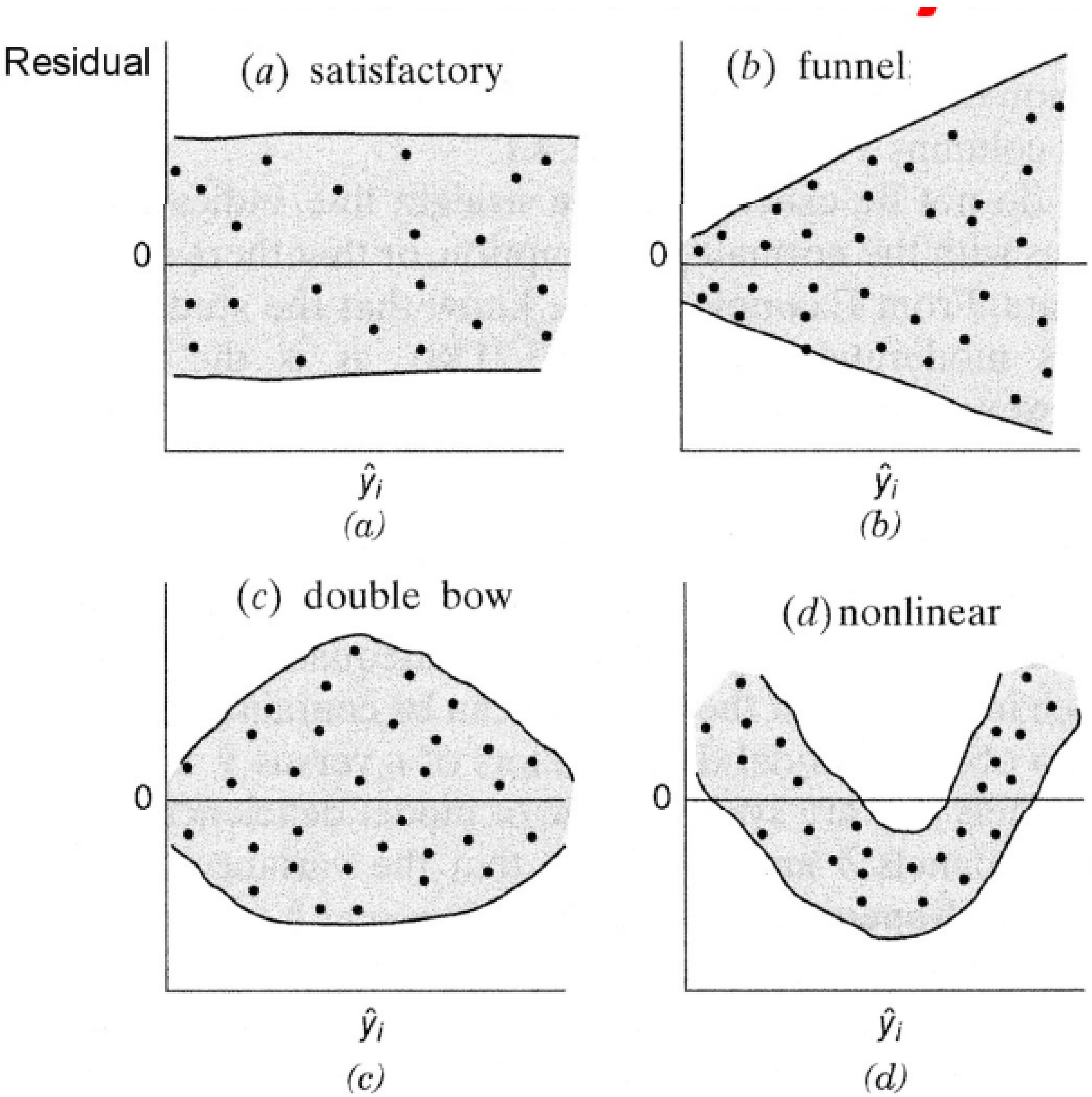
$$\boldsymbol{e} = (\boldsymbol{I} = \boldsymbol{H})\varepsilon$$

and covariance matrix of residuals

$$\text{Var}(\boldsymbol{e}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$

Definition 3.3 (Studentized Residuals).
Studentized residuals is defined as

$$\frac{e_i}{\sqrt{MS_{\text{Res}}}(1 - h_{ii})}$$



Definition 3.4 (PRESS Residual).
PRESS Residual $e_{(i)}$ is defined as

$$e_{(i)} = y_i - \hat{y}_{(i)}$$

where $\widehat{y_{(i)}}$ is the fitted value of the ith response based on all observations except the ith one. We have

$$e_{(i)} = \frac{e_i}{1 - h_{ii}}$$

The variance of $e_{(i)}$ is

$$\mathrm{Var}[e_{(i)}] = \frac{\sigma^2}{1 - h_{ii}}$$

Therefore, we can also define standardized PRESS Residual as

$$\frac{e_{(i)}}{\sqrt{\text{Var}[e_{(i)}]}} = \frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$$

Definition 3.5 (R-Student)**.**
R-Student Residual t_j is defined as

$$t_i = \frac{e_i}{\sqrt{S_{(i)}^2(1-h_{ii})}}$$

where $S_{(i)}^2$ is the estimat of σ^2 based on a data set with ith observation removed, i.e.

$$S_{(i)}^2 = \frac{(n-p)MS_{\text{Res}} - e_i^2/(1-h_{ii})}{n-p-1}$$

3.1 Normal Probability Plot

Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, let the CDF of X_i be $F(x)$. Then, we have

$$\frac{X_i - \mu}{\sigma} \sim N(0,1)$$

and

$$F(\frac{X_i - \mu}{\sigma}) \sim U(0,1)$$

Suppose we order them in increasing order with $F(\frac{X_{(1)}-\mu}{\sigma})$ being the smallest, then

$$F(\frac{X_{(i)} - \mu}{\sigma}) \sim \text{Beta}(i, n+1-i)$$

where

$$E[F(\frac{X_{(i)} - \mu}{\sigma})] = \frac{i}{n+1}$$

Therefore, heuristically we have

$$F(\frac{X_{(i)} - \mu}{\sigma}) \sim \frac{i}{n+1}$$

in turn, we hrrive at

$$\frac{X_{(i)} - \mu}{\sigma} \sim F^{-1}(\frac{i}{n+1})$$

so

$$X_{(i)} \sim \sigma F^{-1}(\frac{i}{n+1}) + \mu$$

Therefore, a **normal Q-Q** probability plot is obtained by plotting $X_{(i)}$ against $F^{-1}(\frac{i}{n+1})$. A **normal P-P** probability plot is obtained by plotting $F(\frac{X_{(i)}-\mu}{\sigma})$ against $\frac{i}{n+1}$. The plot is expected to be linear if the data set comes from a normal distribution.

3.2 PRESS Statistic

Recall that PRESS residuals $e_{(i)} = y_i - \hat{y}_{(i)}$, where $\hat{y}_{(i)}$ is the predicted value of the ith observed response based on a model fit to the remaining $n-1$ sample points.

Definition 3.6 (PRESS Statistics)**.**
The PRESS statistics

$$\text{PRESS} = \sum_{i=1}^{n}[y_i - \hat{y}_{(i)}]^2 = \sum_{i=1}^{n}(\frac{e_i}{1-h_{ii}})^2$$

PRESS statistic can be used as measure of model quality, where a model with a small value of PRESS is desired.

3.3 Detection and Treatment of Outliers

Residual plots against $\hat{y}_i$ and the normal probability plot are helpful in identifying outliers.

3.4 Lack of Fit of Regression Model

A lack of fit test requires replicated observations on the response y for at least one level of x. These replicated observations are then used to obtain a **model independent estimate** of σ^2.
Suppose we have n_i observations on the response at the ith level of the regressor x_i, where $i = 1, \ldots, m$. Let y_{ij} denote the jth observation on the response at x_i, where $j = 1, \ldots, n_i$. There are $n = \sum_{i=1}^{m} n_i$ total observations.
We can decompose SS_{Res} as

$$SS_{\text{Res}} = SS_{\text{PE}} + SS_{\text{LOF}}$$

where SS_{PE} is the sum of squares due to **pure error** and SS_{LOF} is the sum of squares due to **lack of fit**.
In particular the above equation corresponds to the equation below

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2$$

Note, here SS_{PE} is **model independent** estimate of σ^2.
Therefore, the test statistics is

$$F_0 = \frac{SS_{\text{LOF}}/(m-2)}{SS_{\text{PE}}/(n-m)} = \frac{MS_{\text{LOF}}}{MS_{\text{PE}}}$$

where $E(MS_{\text{LOF}}) = \sigma^2 + \frac{\sum_{i=1}^{m} n_i[E(y_i)-\beta_0-\beta_1 x_i]^2}{m-2}$, which will be close to σ^2 if there is no lack of fit.
We conclude that regression function is not linear if $F_0 > F_{\alpha,m-2,n-m}$, and conclude there is no strong evidence of lack of fit otherwise. In the case of no lack of fit, both MS_{PE} and MS_{LOF} will be combined to estimate σ^2.

4 Transformation and Weighting to Correct Model Inadequacies

In this section, we assume

1. The model errors have mean zero and constant variance and are uncorrelated.
2. The model errors have a normal distribution.
3. The form of the model.

The useful variable stabilizing transformations can be found in the table below: Sometimes, the model itself is not linear, therefore, we can linearise the model so that $y' = \beta_0 + \beta_1 x'$ for some transformed y and x.

Relationship of σ^2 to $E(y)$	Transformation
$\sigma^2 \propto$ constant	$y' = y$
$\sigma^2 \propto E(y)$	$y' = \sqrt{y}$
$\sigma^2 \propto E(y)[1 - E(y)]$	$y' = \sin^{-1}(\sqrt{y})$
$\sigma^2 \propto E(y)^2$	$y' = \ln(y)$
$\sigma^2 \propto E(y)^3$	$y' = y^{-\frac{1}{2}}$
$\sigma^2 \propto E(y)^4$	$y' = y^{-1}$

4.1 Transformation on y: Box-Cox Method

Theorem 4.1 (Box Cox Method)**.**
We try to transform $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ to $\boldsymbol{y}^{(\lambda)} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ where

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda \bar{y}^{\lambda - 1}}, & \lambda \neq 0 \\ \bar{y} \ln y, & \lambda = 0 \end{cases}$$

where $\bar{y} = \frac{1}{\ln(\frac{1}{n}\sum_{i=1}^n \ln y_i)}$. We select λ to minimize the residual sum of squares from the regression of y^λ on x.

Box-Cox Method transforms on y. We can also do transformation on x such that $y = \beta_0 + \beta_1 x^\alpha + \epsilon$. To obtain the α, we need to do the following iterations:

- Let $\alpha_0 = 0$
- Denote $\xi = \begin{cases} x^\alpha, & \alpha \neq 0 \\ \ln x, & \alpha = 0 \end{cases}$.
- Taylor series $E(y) = f(\alpha) + (\alpha - \alpha_0) f'(\alpha_0) = \beta_0 + \beta_1 x + (\alpha - 1) f'(\alpha_0)$
- Therefore, $E(y) = \beta_0 + \beta_1 x + \underbrace{(\alpha - 1)\beta_1}_{\gamma} x^{\alpha_0} \ln x$.
- Use linear regression to find the estimators of above coefficients $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\gamma}$.
- Then we have $\alpha_1 = \frac{\hat{\gamma}}{\hat{\beta}_1} + 1$

4.2 Generalized and Weighted Least Squares

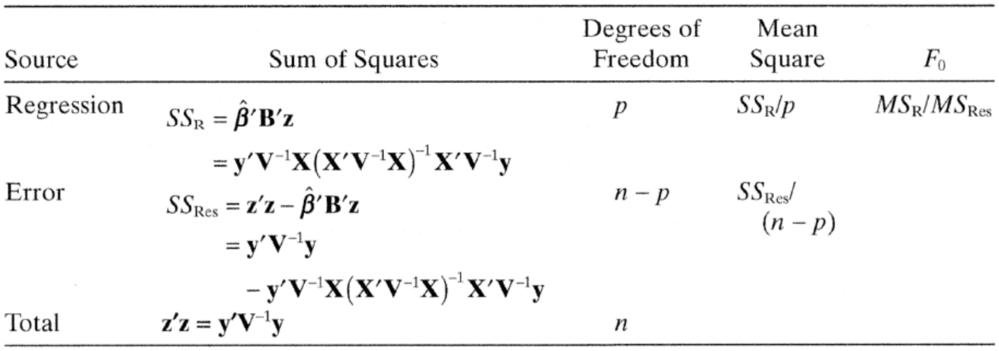
For **generalized least squares**, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = 0$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$, where this V is non-singular and positive definite.
We have a non-singular symmetric matrix $\boldsymbol{K}$, where $\boldsymbol{KK} = \boldsymbol{V}$.
Define $\boldsymbol{z} = \boldsymbol{K}^{-1}\boldsymbol{y}$, $\boldsymbol{B} = \boldsymbol{K}^{-1}\boldsymbol{X}$ and $\boldsymbol{g} = \boldsymbol{K}^{-1}\boldsymbol{\epsilon}$. This gives

$$\boldsymbol{z} = \boldsymbol{B\beta} + \boldsymbol{g}$$

with $E(\boldsymbol{g}) = \boldsymbol{0}$ and $\mathrm{Var}\boldsymbol{g} = \sigma^2\boldsymbol{I}$.
Therefore, $S(\boldsymbol{\beta}) = (\boldsymbol{y}-\boldsymbol{X\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})$, and least square normal equation is $(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$, so

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$$

and $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.
We have $\mathrm{Var}\hat{\boldsymbol{\beta}} = \sigma^2(\boldsymbol{B}'\boldsymbol{B})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}$.
The ANOVA table is as followed



Source	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	$SS_{\mathrm{R}} = \hat{\boldsymbol{\beta}}'\mathbf{B}'\mathbf{z}$ $= \mathbf{y}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$	p	SS_{R}/p	$MS_{\mathrm{R}}/MS_{\mathrm{Res}}$
Error	$SS_{\mathrm{Res}} = \mathbf{z}'\mathbf{z} - \hat{\boldsymbol{\beta}}'\mathbf{B}'\mathbf{z}$ $= \mathbf{y}'\mathbf{V}^{-1}\mathbf{y}$ $- \mathbf{y}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$	$n-p$	$SS_{\mathrm{Res}}/(n-p)$	
Total	$\mathbf{z}'\mathbf{z} = \mathbf{y}'\mathbf{V}^{-1}\mathbf{y}$	n		

Theorem 4.2 (Weighted Least Square)**.**
Weighted Least Square is the special case of generalized least square where

$$\sigma^2\boldsymbol{V} = \sigma^2\begin{pmatrix} \frac{1}{w_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{w_n} \end{pmatrix}$$

So $\boldsymbol{W} = \boldsymbol{V}^{-1}$. Therefore,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}$$

The transformed matrix $\boldsymbol{B} = \begin{pmatrix} 1\sqrt{w_1} & x_{11}\sqrt{w_1}\cdots x_{1k}\sqrt{w_1} & & \\ 1\sqrt{w_2} & x_{21}\sqrt{w_2}\cdots x_{2k}\sqrt{w_2} & & \\ \cdots & \cdots & \ddots & \cdots \\ 1\sqrt{w_n} & x_{n1}\sqrt{w_n}\cdots x_{nk}\sqrt{w_n} & & \end{pmatrix}$ and $\boldsymbol{Z} = \begin{pmatrix} y_1\sqrt{w_1} \\ y_2\sqrt{w_2} \\ \vdots \\ y_n\sqrt{w_n} \end{pmatrix}$.

5 Diagonistics for Leverage and Influence

Definition 5.1.
Leverage is a measure of the effect of an observation i on the predicted value.
The **leverage** of the observation i is h_{ii} where $h_{ii} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$ where $\boldsymbol{x}_i'$ is the ith row of the $\boldsymbol{X}$ matrix.
Also, we note that $\bar{h} = \frac{1}{n}\sum_{i=1}^n h_{ii} = \frac{p}{n}$.
Any observation i for which h_{ii} exceeds $2\bar{h} = 2\frac{p}{n}$ is remote enough from the rest of the data and is considered a **leverage point**.

Remark: Not all leverage points are going to be influential on regression coefficients. Usually, a point is influential if it deteriorates MS_{Res} a lot.
More precisely, we use Cook's Distance as a measure of **influence**.

Definition 5.2 (Cook's Distance)**.**
Let $\hat{\boldsymbol{\beta}}$ be the estimate based on all n points and $\hat{\boldsymbol{\beta}}_{(i)}$ be the estimate obtained by deleting the ith point. Then the Cook's distance for the ith point is defined to be

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{pMS_{\text{Res}}}$$

Points with large values of D have considerable influence on the least-square estimates $\hat{\boldsymbol{\beta}}$. The magnitude of D_i is usually assessed by comparing it to $F_{\alpha,p,n-p}$ and since $F_{0.5,p,n-p} \approx 1$, we usually consider points for which $D_i > 1$ to be influential.

Cook's distance can be calculated via alternative formula:

$$D_i = \frac{r_i^2}{p}\frac{\text{Var}(\hat{y}_i)}{\text{Var}(e_i)} = \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}}$$

where r_i is the ith studentized residual.
Also, we can write D_i as

$$D_i = \frac{(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})'(\hat{\boldsymbol{y}}_{(i)} - \hat{\boldsymbol{y}})}{pMS_{\text{Res}}}$$

Measure of influence can also include DFFITS and DFBETAS.

Definition 5.3 (DFBETAS)**.**
We define $DFBETAS_{j,i}$ on the jth regression coefficient with respect to ith observation.

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S_{(i)}^2 C_{jj}}}$$

where C_{jj} is the jth diagonal elements of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, $S_{(i)}^2$ is the estimate of σ^2 based on a data set with ith observation removed, and $\hat{\beta}_{j(i)}$ is the jth regression coefficient computed without use of ith observation.
A large value of $DFBETAS_{j,i}$ indicates that observation i has considerable influence on the jth regression coefficients.

Similarly, we define $DFFITS_i$.

Definition 5.4 (DFFITS)**.**
We define $DFFITS_i$ on the ith predicted y_i.

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S^2_{(i)} h_{ii}}}$$

where $\hat{y}_{(i)}$ is the fitted value of y_i obtained without the use of ith observation.
$DFFITS_i$ is the number of standard deviations that the fitted value $\hat{y}_i$ **changes** if observation i is removed. (Note: $\text{Var}(\hat{y}_i)) = \sigma^2 h_{ii})$

An alternative formula for $DFFITS_i$ is

$$DFFITS_i = \frac{h_{ii}}{1 - h_{ii}}^{\frac{1}{2}} t_i$$

where t_i is the R-student residual.
Therefore, if the data point is an outlier, then R-student will be large in magnitude while if data has high leverage, h_{ii} will be close to unity. In either case $DFFITS_i$ is large.
The cutoff value suggested for $DFFBETAS_{j,i}$ and $DFFITS_i$ is

- $|DFBETAS_{j,i}| > \frac{2}{\sqrt{n}}$
- $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$

5.1 A Measure of Model Performance

Definition 5.5 (Generalised Variance)**.**
The overall precision of estimation can be measured using **generalized variance** GV:

$$GV(\hat{\boldsymbol{\beta}}) = |\text{Var}(\hat{\boldsymbol{\beta}})| = |\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}|$$

To express the role of ith observation on the preision of estimation, define $COVRATIO_i$ to be

$$COVRATIO_i = \frac{|(\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1} S^2_{(i)}|}{(\boldsymbol{X}'\boldsymbol{X})^{-1} MS_{\text{Res}}}$$

If $COVRATIO_i > 1$, the ith observation improves the precision of estimation, whereas a less than 1 $COVRATIO_i$ suggesta that inclusion of ith point degrades precision.

The suggested cutoff values for $COVRATIO_i$: if $COVRATIO_i > 1 + \frac{3p}{n}$ or if $COVRATIO_i < 1 - \frac{3p}{n}$, ten the ith point should be considered influential. The lower bound is appropriate if $n > 3p$. These cutoffs are only recommended for large samples.

6 Polynomial Regression Models

6.1 Polynomial Models

kth order polynomial model in one variable is in the form of

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta^k x^k + \varepsilon$$

The definition can extend to kth order polynomial model in many variables. For example, the second-order polynomial in 2 variables is in the form of

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

There are a few important considerations when fitting a polynomial in one variable:

1. **Order of the model** should be kept as **low** as possible.

2. **Model-Building Strategy**: forward selection or backward elimination.

3. **Extrapolation** should be avoided

4. **Ill-COnditioning I**. As order of polynomial increases, the $\boldsymbol{X'X}$ matrix becomes ill-conditioned.
   Solution: Non-essential ill-conditioning caused by arbitrary choice of origin can be removed by first centering the regressor variables, i.e. $x \to x - \bar{x}$.

5. **Ill-Conditioning II**. If values of x are limited to a narrow range, there can be significant ill-conditioning and multicollinearity in the columns of $\boldsymbol{X}$ matrix.

6. **Hierarchy**. A model is **hierarchical** if the power of varaible is continuous.
   Only hierarchical models are invariant under linear transformation of x, which means both models will produce the same predicted values, residuals, R^2, etc.

6.2 Piecewise Polynomial Fitting(Splines)

Splines are piecewise polynomial of order k. The joint points of the pieces are called **knots**.

Definition 6.1 (Cubic Spline)**.**
A cubic spline with h knots, $t_1 < \cdots < t_h$ with continuous first and second derivatives can be written as

$$S(x) = \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{i=1}^{h} \beta_i (x - t_i)_+^3$$

where

$$(x - t_i)_+ = \begin{cases} (x - t_i) & \text{if } x - t_i > 0 \\ 0 & \text{if } x - t_i \leq 0 \end{cases}$$

Generally, we require function values and $k-1$ derivatives to agree on the knots, so that the spline is a continous functin with $k-1$ continuous derivatives. To do this, we need to assume the position of the knots are known.

Definition 6.2 (Cubic Spline with No Continuity Restriction)**.**

$$S(x) = \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{i=1}^{h} \sum_{k=0}^{3} \beta_{ij} (x - t_i)_+^j$$

Theorem 6.1 (Hypothesis Testing of Cubic Spline)**.**

- $H_0 : \beta_0 = 0$ tests continuity of $S(x)$.
- $H_0 : \beta_0 = \beta_1 = 0$ tests continuity of $S(x)$ and $S'(x)$.
- $H_0 : \beta_0 = \beta_1 = \beta_2 = 0$ tests continuity of $S(x)$, $S'(x)$ and $S''(x)$.
- $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$ tests single cubic polynomial fits data better than a cubic spline over range of x.

Cubic Spline model is good as fitting it can be treated as fitting a mutiple linear regression model.
In general, we can have linear spline just by changing the formula's power, for both the continuous and discontinuous case.

6.3 Non-parametric Regression

In paramatric model, we have a proposed equation and we try to find the corresponding unknown coefficient(s). In Non-parametric model, we do not have a proposed equation.

Definition 6.3 (Kernel Regression)**.**
We estimate the ith response in this way (called kernel smoother estimation)

$$\tilde{y}_i = \sum_{j=1}^{n} w_{ij} y_j$$

where

$$w_{ij} = \frac{K(\frac{x_i - x_j}{b})}{\sum_{k=1}^{n} K(\frac{x_i - x_k}{b})}$$

and $K(t) = \begin{cases} 1, & |t| \leq 0.5 \\ 0, & |t| > 0.5 \end{cases}$.
Essentially, K will be 0 if and only if for that particular i and J, $x_i - x_j \leq \frac{b}{2}$. This is known as Box kernel.

The properties of kernel smoother depend much more on choice of bandwidth b than actual kernerl function. Other kernel functions can be found in lecture notes.

6.4 Locally Weighted Regression(LOESS)

Let x_0 be specific location of interest. We pick a span d, and let the $\delta(x_0)$ be the distance of the farthest point in the neighbourhood $[x_0 - d, x_0 + d]$. For another point x_j, the weight is $W(\frac{|x_0 - x_j}{\delta(x_0)})$ where $W(t) = \begin{cases} (1-t^3)^3 & \text{for } 0 \leq < 1 \\ 0 & \text{elsewhere} \end{cases}$. We then use all points in the neighbourhood to generate a **weighted least-squares estimate** of the specific response x_0.

6.5 Estimating σ^2

Here estimator $\tilde{\boldsymbol{y}} = \boldsymbol{Sy}$ is a linear function in $\boldsymbol{y}$, regardless whether our model is parametric or not. For example, in multiple linear regression, we have $\hat{\boldsymbol{y}} = \boldsymbol{Hy}$.
Therefore, we have

$$SS_{\text{Res}} = (\boldsymbol{y} - \boldsymbol{Sy})'(\boldsymbol{y} - \boldsymbol{Sy}) = \boldsymbol{y}'[\boldsymbol{I} - \boldsymbol{S}' - \boldsymbol{S} + \boldsymbol{S}'\boldsymbol{S}]\boldsymbol{y}$$

And

$$E(SS_{\text{Res}}) = \text{tr}[(\boldsymbol{I} - \boldsymbol{S}' - \boldsymbol{S} + \boldsymbol{S}'\boldsymbol{S})\boldsymbol{\sigma}^2\boldsymbol{I}] = \sigma^2[n - 2\text{tr}(\boldsymbol{S}) + \text{tr}(\boldsymbol{S}'\boldsymbol{S})]$$

Therefore, an estimator of $\boldsymbol{\sigma}^2$ is

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}(y_i - \tilde{y}_i)^2}{n - 2\text{tr}(\boldsymbol{S}) + \text{tr}(\boldsymbol{S}'\boldsymbol{S})}$$

and also

$$R^2 = \frac{SS_{\text{T}} - SS_{\text{Res}}}{SS_{\text{T}}}$$

6.6 Polynomial Models in Two or More Variables

An example of a two-variable quadratic model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$.
We can use repeated experiment to get an independent estimate of σ^2 so as to test lack of fit.
We can transform the regressor variable to an indicator variable. This reduces the complexity of $\boldsymbol{X}$ matrix.
To fit a quadratic model, we usually requires four points in a rectangular fashion. We can add some more experiments in the cross manner.

6.7 Rotability

$\sqrt{\text{Var}[\hat{y}(\boldsymbol{x}_0)]}$ is the same for all points $\boldsymbol{x}_0$ that are the same distance from the center of the design. This is desirable, and this is how we introduce the points in experiment.

6.8 Orthogonal Polynomials

Suppose the model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \epsilon_i$. Generally, the column of $\boldsymbol{X}$ matrix will not be orthogonal.
Now suppose that we fit the model

$$y_i = \alpha_0 P_0(x_i) + \alpha_1 P_1(x_i) + \cdots + \alpha_k P_k(x_i) + \epsilon$$

where $P_u(x_i)$ is a uth order orthogonal polynomial defined such that

$$\sum_{i=1}^{n} P_r(x_i) P_s(x_i) = 0, r \neq s \forall r, s = 0, \ldots, k$$

and $P_0(x_i) = 1$.

Then the model becomes $\boldsymbol{y} = \boldsymbol{X\alpha} + \boldsymbol{\epsilon}$, where $\boldsymbol{X} = \begin{pmatrix} P_0(x_1) \cdots P_k(x_1) & & \\ \vdots & \ddots & \vdots \\ P_0(x_n) & \cdots & P_k(x_n) \end{pmatrix}$.

Due to orthogonality, we have the nice property:

$$\boldsymbol{X'X} = \text{diag}[\sum_{i=1}^{n} P_0^2(x_i), \ldots, \sum_{i=1}^{n} P_k^2(x_i)]$$

The least square estimators of α are found from $(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$:

$$\hat{\alpha}_j = \frac{\sum_{i=1}^{n} P_j(x_i) y_i}{\sum_{i=1}^{n} P_j^2(x_i)}$$

Since, we set $P_0(x_i) = 1$, we have $\hat{\alpha}_0 = \bar{y}$.
The residual sum of squares is

$$SS_{\text{Res}}(k) = SS_{\text{T}} - \sum_{j=1}^{k} \hat{\alpha}_j [\sum_{i=1}^{n} P_j(x_i) y_i]$$

The regression sum of sequares for any model parameter does not depend on the other parameters in teh model. The regression sum of squares is

$$SS_{\text{R}}(\alpha_j) := SS_{\text{R}}(\alpha_j \mid \alpha_0, \cdots, \alpha_{j-1}, \alpha_{j+1}, \cdots, \alpha_k) = \hat{\alpha}_j \sum_{i=1}^{n} P_j(x_i) y_i := SS_{\text{R}}(\alpha_j \mid \alpha_0)$$

The last equality comes from orthogonality.
To test $H_0 = \alpha_k = 0$, we need

$$F_0 = \frac{SS_{\text{R}}(\alpha_k)}{SS_{\text{Res}}(k)/(n-k-1)}$$

The orthogonal polynomials $P_j(x_i)$ can be easily constructed if the levels of x are equally spaced. For explicit formula, look into lecture notes.

7 Indicator Variables

There are two types of variables: quantitative variables and qualitative/categorical variables.
A qualitative variable has no natural scale of measurement.
We can use indicator variables to indicate whether a variable should be considered in the model.
We can use the indicator variable in or not in interaction terms.
If there is no interaction, we only need to test the coefficient of indicator.
If there is interaction term, we also need to test for interaction term.

7.1 Special Case Testing

We are interested in testing 3 cases, namely parallel lines, concurrent lines and coincident lines.

7.1.1 Parallel Lines

Two lines are parallel if their slopes are equal and we allow the y-intercepts to differ.
For example, for the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$, $x_2 \in \{0, 1\}$ is an indicator variable that allows both slope and y intercept to differ.

- Suppose $x_2 = 0$, we have $y = \beta_0 + \beta_1 x_1 + \epsilon$
- Suppose $x_2 = 1$, we have $y = (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) x_1 + \epsilon$

To test for parallel lines, we test $H_0 : \beta_{12} = 0$.
For multiple indicator variables $x_2, \ldots, x_n$, the genral model with interaction terms is

$$y = \beta_0 + \beta_1 x_1 + \sum_{i=2}^{n} (\beta_i x_i + \beta_{1i} x_1 x_i) + \epsilon$$

and we test $H_0 = \beta_2 = \cdots = \beta_n = 0$

7.1.2 Concurrent Lines

Concurrent lines are lines where their y-intercepts are the same and we allow the slopes to differ.
Using the model above, we are testing $H_0 : \beta_2 = \cdots = \beta_n = 0$.

7.1.3 Coincident Lines

Coincident lines are lines where they have the same slope as well as the same y-intercept.
Using the model above, we are testing $H_0 : \beta_2 = \cdots = \beta_n = 0$, and $\beta_{12} = \cdots = \beta_{1n} = 0$.

7.2 Indicator Variable versus Regression on Allocated Codes

The difference between indicator variable against allocated codes is that

- Each indicator variable is only binary, i.e., $\{0, 1\}$, one level is represented by a 1 for that particular variable and 0 for anything else, and each variable will become a regressor variable.
- Allocated codes is just different assignement of values in **one** variable to represent different levels, and turns to only one regressor variable.

However, allocated codes do not work well. This is because, $E(y \mid x_1, x_{i+1}) - E(y \mid x_1, x_i) = \beta$ for all i, which doesn't make sense, as it imposes a particular metric on the levels of the qualitative factor.

7.3 Indicator Variable as a Substitute for a Quantitative Regressor

We can represent a range of quantitative variable as a set of indicator variables. However, it has the disadvantage that more parameters are required to represent the information content.
However, an advantage of the indicator variable approach is that it *does not* require the analyst to make any prior assumptions about the **functional form of the relationship** between the response and the regressor variable.

7.4 Pooled Two-sample t Procedure

Suppose that an Simple Random Sample of size n_1 is drawn from a Normal population with unknown mean μ_1 and that an independent SRS of size n_2 is drawn from another Normal population with unknown mean μ_2. Suppose also that the two population have the **same standard deviation**.
To test the hypothesis that $H_0 : \mu_1 = \mu_2$ against alternative $H_1 : \mu_1 \neq \mu_2$, we use test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where s_p^2 is the pooled estimators of σ^2, in the form of

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Here, we can treat the two variables coming from two binary choices of the same categorical variable.
One way ANOVA is a generalization of the pooled two-sample t procedure.
In one-way ANOVA, let k denote number of levels and n denote number of observations per level.
For the model for one-way classification analysis of variance we have

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \text{ where } i = 1, \ldots, k, \quad j = 1, \ldots, n$$

and the regression model is

$$y_{ij} = \beta_0 + \sum_{p=1}^{k-1} \beta_p x_{pj} + \epsilon_{ij}, \text{ where } i = 1, \dots, k, \quad j = 1, \dots, n$$

Here, $x_1, \dots, x_{k-1}$ are $k-1$ indicator variables, used to represent k levels.
The connection between the two models is that they share the same ANOVA table.
For the experimental design model, μ is the common mean, τ_i is the binary variable, which equals 1 if it comes from the ith level and 0 otherwise.
For the regression model, we can create a matrix $X := \mathbb{R}^{kn\times(1+k)}$ where $M = (1, 1_{\tau_1}, \dots, 1_{\tau_k})$. However, the matrix is firstly, not linearly independently on a column basis, which makes $X'X$ singular. Therefore, we impose an additional constraint:

$$\sum_{i=1}^{k} \tau_i = 0 \quad (*)$$

And the matrix $X = (1, 1_{\tau_1} - 1_{\tau_k}, \cdots, 1_{\tau_{k_1}} - 1_{\tau_k})$.
After we get this X, we can solve the actual value μ and τ by solving

$$(X'X) \begin{pmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_k \end{pmatrix} = X'y$$

The ANOVA table obtained will be We can use the regression model too, where we arrive

Degrees of Variation	Sum. Squares	Degrees of Freedom	Mean Square	F_0
Treatment	$n\sum_{i=1}^{k}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$k-1$	$\frac{SS_{\text{Treatment}}}{k-1}$	$\frac{MS_{\text{Treatments}}}{MS_{\text{Res}}}$
Error	$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i\cdot})^2$	$k(n-1)$	$\frac{SS_{\text{Res}}}{k(n-1)}$	
Total	$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{\cdot\cdot})^2$	$kn-1$		

at the same regressor and ANOVA.
We can use the ANOVA table's F_0 to test

$$H_0 : \tau_1 = \cdots = \tau_k = 0 \text{ against } H_1 : \tau_i \neq 0 \text{ for at least one } i$$

The relationship between parameters of regression model and paramaters of Analysis of Variance model is that

$$\beta_0 = \mu_k, \beta_i = \mu_i - \mu_k, i = 1, \dots, k-1$$

8 Multicollinearity

If there is no linear relationship between regressors, they are said to be **orthogonal**. When there are **near-linear** dependencies among the regressors, the problem of **multicollinearity** is said to exist. To observe multicollinearity, we need to use unit-length scaling on X: $y \to \frac{y-\mu_y}{\sigma_y\sqrt{n-1}}$. We then compute the correlation coefficients via $\boldsymbol{X}'\boldsymbol{X}$.

Theorem 8.1 (Eigenvalue Decomposition)**.**
We can decompose A to be $A = T\Lambda T'$ where $T = (t_1, \ldots, t_p)$ are p orthogonal eigenvectors of A, and $\Lambda = \text{diag}[\lambda_1, \ldots, \lambda_p]$ where $\lambda_1, \ldots, \lambda_p$ are p *positive* eignevalues of A.
With this setup, we have

- $\text{tr}(A) = \sum_{i=1}^{p} \lambda_i$
- $\text{tr}(A') = \sum_{i=1}^{p} 1/\lambda_i$.

8.1 Source of Multicollinearity

Given the multiple regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, if we have scaled the regressor variables and response to unit length, then $\boldsymbol{X}'\boldsymbol{X}$ is a matrix of correlations between regressors, whereas $\boldsymbol{X}'\boldsymbol{y}$ is a vector of correlation between regressors and the response.
Let jth column of $\boldsymbol{X}$ matrix be $\boldsymbol{X}_j$. Then we say $X_1, \ldots, X_p$ will be linearly dependent if there is a set of constants $t_1, \ldots, t_p$ not all zero, such that $\sum_{j=1}^{p} t_j \boldsymbol{X}_j = \mathbf{0}$.
The sum above is a vector of constant $\boldsymbol{m}$ if regressor are not centered.
If the above equation holds exactly for a subset of columns of $\boldsymbol{X}$, then the rank of $\boldsymbol{X}'\boldsymbol{X}$ matrix is less than p and inverse does not exist.
If above equation is approximately true for some subsets of columns of $\boldsymbol{X}$, then tehre will be a near-linear dependency in $\boldsymbol{X}'\boldsymbol{X}$ and problem of multicollinearity is said to exist.
Therefore, we can see multicolinerality as a form of ill-conditioning in the $\boldsymbol{X}'\boldsymbol{X}$ matrix.
There are four primary sources of multicollinearity:

1. The data collection method employed, where the data is only sampled from a subspace of region of regressors.
2. Constraints on the model or in the population
3. Model Specification, whether we center the data
4. An overdefined model, which has more regressor variables than observations.

8.2 Effect of Multicollinearity

Suppose there are only two regressor variable x_1, x_2 and model

$$y = \beta_1 x_1 + \beta_2 x_2$$

have x_1, x_2, y scaled to unit length. We have, for $(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$,

$$\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} r_{1y} \\ r_{2y} \end{pmatrix}$$

where r_{12} is the simple correlation between x_1 and x_2, and r_{iy} is the simple correlation between x_i and y.
Then we have

$$\boldsymbol{C} := (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{pmatrix}$$

and $\hat{\beta}_1 = \frac{r_{1y}-r_{12}r_{2y}}{1-r_{12}^2}$, $\hat{\beta}_2 = \frac{r_{2y}-r_{12}r_{1y}}{1-r_{12}^2}$. Strong multicolliearity between x_1 and x_2 will cause r_{12} to be large. This results, the variance of $\hat{\beta}_j$ and $\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$ is very large.

When there are more than two regressor variables, diagonal elements of $\mathbf{C}$ has $C_{jj} = \frac{1}{1-R_j^2}$ and $\text{Var}(\hat{\beta}_j) = C_{jj}\sigma^2$.
Where R_j^2 is the coefficient of multiple determination from the regression of x_j on the remaining $p-1$ regressors.
If there is strong multicollinearity between x_i and any subset of the other $p-1$ regressors, then the value of R_i^2 will be close to unity.
Generaly, the covariance between two β will also be large if the corresponding regressor are involved in a multicollinear relationship.
Multicollinearity also tends to produce least-square estimates $\hat{\beta}_j$ that are twoo large in absolute value.
The square distance L_1^2 is defined as

$$L_1^2 = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

The expected squared distance $E(L_1^2) = \sigma^2 \text{tr}(\boldsymbol{X}'\boldsymbol{X})^{-1}$.
Here, $E(L_1^2) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}$. When the multicollinearity present, some of the eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ will be small, which makes this expectation large. So we just need to check eigenvalues.
We have

$$E(\hat{\boldsymbol{\beta}'}\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 \text{tr}(\boldsymbol{X}'\boldsymbol{X})$$

8.3 Multicollinearity Diagonostics

Techniques for detecting multicollinearity include

- Examination of Correlation Matrix, which detects only correlation between 2 variables.
- Variance Inflation Factors
- Eigensystem Analysis of $\boldsymbol{X}'\boldsymbol{X}$

8.3.1 Variance Inflation Factors

We look at matrix $\mathbf{C} := (\boldsymbol{X}'\boldsymbol{X})^{-1}$, which gives the $\text{Var}\hat{\beta} := \sigma^2\mathbf{C}$. The variance inflation factor is defined by diagonal element of this matrix. If the regressor variable is **centered**, we have

$$VIF_j = C_{jj} = (1 - R_j^2)^{-1}$$

If x_j is nearly orthogonal to remaining regressors, R_j^2 is small and C_jj is close to unity. However, if x_j is nearly dependent on some subset of remaining regressors, R_j^2 is nearly unity and C_jj is large. We can view C_{jj} as the factor by which the variance of $\hat{\beta}_1$ is increased due to near-linear dependencies among the regressors.

8.3.2 Eigensystem analysis of $\boldsymbol{X}'\boldsymbol{X}$

The characteristic roots, or **eigenvalues** of $\boldsymbol{X}'\boldsymbol{X}$, say $\lambda_1, \ldots, \lambda_p$, can be used to measure the extent of multicollinearity in the data.

If there are one or more near-linear dependencies in the data, then one or more of the eignevalues will be small. One or more small eignevalues imply that there are near-linear dependencies among the $\boldsymbol{X}$.

Specifically, we compute the **condition number** of $\boldsymbol{X}'\boldsymbol{X}$

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$$

If the condition number is less than 100, there is no serious problems with multicollinearity. Condition numbers betwween 100 and 1000 implies moderate to strong multicollinearity and if κ exceeds 1000, severe multicollinearity is indicated.

The **condition incides** of the $\boldsymbol{X}'\boldsymbol{X}$ matrix are

$$\kappa_j = \frac{\lambda_{\max}}{\lambda_j}, \quad j = 1, \ldots, p$$

8.3.3 Singular-Value Decomposition

The $n \times p$ $\boldsymbol{X}$ matrix may be decomposed as

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{T}'$$

where $\boldsymbol{U}$ is $n \times p$, $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}$. $\boldsymbol{T}$ is $p \times p$, where $\boldsymbol{T}$ is the matrix of eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$, and $\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}$, and $\boldsymbol{D}$ is a $p \times p$ diagonal matrix with nonnegative diagonal elements μ_j. The μ_j's are called **singular values** of $\boldsymbol{X}$.

Note, the squares of singular values of $\boldsymbol{X}$ are the eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$.

8.4 Methods for Dealing with Multicollinearity

We can deal with multicollinearity by

- Collecting additional data
- Model Respecification(centering)
- Ridge regression
- Pricipal component regression

8.4.1 Ridge Regression

To understand ridge regression, we need to understand mean squared error. Suppose $\hat{\theta}$ is an estimator of θ, then

- $\bar{(}\hat{\theta} - \theta) = E[(\hat{\theta} - \theta)^2] - [E(\hat{\theta} - \theta)]^2$.
- $\text{Var}\hat{\theta} = E[(\hat{\theta} - \theta)^2] - [E(\hat{\theta}) - \theta]^2$.
- $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2$
- If $E(\hat{\theta}) = \theta$, $MSE(\hat{\theta}) = \text{Var}(\hat{\theta})$

Suppose a vector $\hat{\boldsymbol{\theta}}$ is an estiamtor of $\boldsymbol{\theta}$, then

$$\begin{aligned} MSE(\hat{\boldsymbol{\theta}}) &= \text{tr}(E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})']) \\ &= E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \\ &= \sum_{i=1}^{p} E[(\hat{\beta}_i - \beta_i)^2] \\ &= \sum_{i=1}^{p} (\text{Var}(\hat{\beta}_i - \beta_i) + [E(\hat{\beta}_i - \beta_i)]^2) \\ &= \sum_{i=1}^{p} \text{Var}(\hat{\beta}_i) + \sum_{i=1}^{p} [E(\hat{\beta}_i) - \beta_i]^2 \end{aligned}$$

The problem with method of least squares is the requirement that $\hat{\boldsymbol{\beta}}$ is an **unbiased estimator** of $\boldsymbol{\beta}$. The Gauss-Markov property assures us the least-squares estimator has minimum variance in the class of unbiased linear estimators, but there is no guarantee that this variance will be small.
Therefore, we perform **ridge regression** as below:

$$(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})\hat{\boldsymbol{\beta}}_R = \boldsymbol{X}'\boldsymbol{y}, \quad k \geq 0$$

Therefore, we get the biased estimator $\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})\boldsymbol{X}'\boldsymbol{y}$.
The property of Ridge regression include:

- Ridge estimator is a linear transformation of the least square estimator:

  $$\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{Z}_k\hat{\boldsymbol{\beta}}$$

- $E(\hat{\boldsymbol{\beta}}_R) = E(\boldsymbol{Z}_k\hat{\boldsymbol{\beta}}) = \boldsymbol{Z}_k\boldsymbol{\beta}$, therefore, ridge estimator is a biased estimator of $\boldsymbol{\beta}$. Here, $\boldsymbol{Z}_k$ is called the **biasing parameter**.
- Covariance matrix of $\hat{\boldsymbol{\beta}}_R$ is given by

  $$\text{Var}(\hat{\boldsymbol{\beta}}_R) = \sigma^2(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I})^{-1}$$

- Mean square error of the ridge estimator is

$$\begin{aligned} MSE(\hat{\boldsymbol{\beta}}_R) &= \sigma^2 \text{tr}[(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-1}] + k^2\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}-k\boldsymbol{I})^{-2}\boldsymbol{\beta} \\ &= \sigma^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+k)^2} + k^2\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I})^{-2}\boldsymbol{\beta} \end{aligned}$$

  where $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$.
  Note the first term is the variance of $\hat{\boldsymbol{\beta}}_R$ whereas the second term is bias in $\hat{\boldsymbol{\beta}}_R$. Their magnitude is inversely related via k.

- $$\begin{aligned} SS_{\text{Res}} &= (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_R)'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_R) \\ &= (\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\hat{\beta}_R-\hat{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\beta}_R-\hat{\beta}) \end{aligned}$$

  The first term is usual SS_{Res}, whereas the second term is some positive value.

- The ridge regression estimates may be computed by using an ordinary least-squares computer program and augmenting the standardized data as follows:

$$\boldsymbol{X}_A = \begin{pmatrix} \boldsymbol{X} \\ \sqrt{k}\boldsymbol{I}_p \end{pmatrix} \quad \boldsymbol{y}_A = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0}_p \end{pmatrix}$$

  and $\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}_A'\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A'\boldsymbol{y}_A = (\boldsymbol{X}'\boldsymbol{X}+k\boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{y}$.

Therefore, we need choose biasing constant carefully. The ridge trace is a plot of elements of $\hat{\boldsymbol{\beta}}_R$ versus k for k usually in the interval 0 to 1. As k increases, some ridge estimates will vary dramatically. At some value of k, the ridge estimates $\hat{\boldsymbol{\beta}}_R$ stablize. We just select a reasonably small value of k at which ridge estimates $\hat{\boldsymbol{\beta}}_R$ are stable.

8.4.2 Principal Component Regression

In multiple linear regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$, we can state

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{\beta}+\boldsymbol{\epsilon}$$

where we denote $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$ a $p \times p$ diagonal matrix of eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{T}$ a $p \times p$ orthogonal matrix whose columns are the eigenvectors associated with $\lambda_1, \ldots, \lambda_p$. Then, we can write $\boldsymbol{y}$ as

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\alpha}+\boldsymbol{\epsilon}$$

where

- $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{T}$
- $\boldsymbol{\alpha} = \boldsymbol{T}'\boldsymbol{\beta}$
- $\boldsymbol{T}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{T} = \boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{\Lambda}$

The columns of $\boldsymbol{Z} = [Z_1, \ldots, Z_p]$, which defines a new set of orthogonal, are referred to as **principal components**.
The least square estimator of $\boldsymbol{\alpha}$ is

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y} = \boldsymbol{\Lambda}^{-1}\boldsymbol{Z}'\boldsymbol{y}$$

and covariance matrix of $\hat{\boldsymbol{\alpha}}$ is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1} = \sigma^2\boldsymbol{\Lambda}^{-1}$$

The samll eigenvalues of $\boldsymbol{X}'\boldsymbol{X}$ means the variance of corresponding orthogonal regression coefficient will be large.
Let λ_j as the vairance of the jth principal component, then

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathrm{Var}(\boldsymbol{T}\hat{\boldsymbol{\alpha}}) = \sigma^2\boldsymbol{T}\boldsymbol{\Lambda}^{-1}\boldsymbol{T}'$$

and therefore $\mathrm{Var}(\hat{\beta}_j) = \sigma^2\frac{\sum_{i=1}^p t_{ij}^2}{\lambda_j}$. The variance willbe large if any λ_j is small.
One should realise $\boldsymbol{Z} = (\sum_{i=1}^p \boldsymbol{X}_i t_{i1}, \ldots, \sum_{i=1}^p \boldsymbol{X}_i t_{ip})$, so the columns of $\boldsymbol{Z}$ is just a linear combination of $\boldsymbol{X}$'s columns. If some $\lambda_j \approx 0$, then some columns of $\boldsymbol{Z}$ will be close to $\mathbf{0}$.
Let's consider the regression $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, where $\lambda_{p-s+1}, \ldots, \lambda_p$ are close to 0, then we understand the corresponding columns of $\boldsymbol{Z}$ will close to 0, so we can safely set $\alpha_{p-s+1} = \cdots = \alpha_p = 0$, and the model becomes

$$\boldsymbol{y} = \underbrace{(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{p-s})}_{\boldsymbol{Z}_{pc}}\underbrace{(\alpha_1, \ldots, \alpha_{p-s})' + \boldsymbol{\epsilon}}_{\boldsymbol{\alpha}_{pc}}$$

And the principal component regression coefficient $\hat{\beta}_{pc} = \boldsymbol{T}\hat{\boldsymbol{\alpha}}_{pc}$.

9 Variable Selectino and Model Building

In most practical problems, the analyst has a rather large pool of possible **candidate regressors**, of which only a few are likely to be important. Therefore, we need to find an appropriate subset of regressor for the model, which is known as the *variable selection problem.* None of the variable selection procedures are guaranteed to produce the best regression equation for a given data set.

9.1 Consequence of Model Misspecification

Assume that there are K candidate regressors $x_1, \ldots, x_K$ and $n \geq K + 1 - r$ observations on these regressors and the response y. The full model containing all K regressors is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Let r be the number of regressors deleted from the previous model. So the number of variables that are retained is $p = k + 1 - r$. The full model can be represented as

$$\boldsymbol{y} = \boldsymbol{X}_p\boldsymbol{\beta}_p + \boldsymbol{X}_r\boldsymbol{\beta}_r + \boldsymbol{\epsilon}$$

and for the full model the least-square estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, and an estimate of the residual variance σ^2 is

$$\hat{\sigma}^2_* = \frac{\boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}^{*\prime}\boldsymbol{X}'\boldsymbol{y}}{n - K - 1} = \frac{\boldsymbol{y}'[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{y}}{n - K - 1}$$

For the subset model $\boldsymbol{y} = \boldsymbol{X}_p\boldsymbol{\beta}_p + \boldsymbol{\epsilon}$, the least square estimate of $\boldsymbol{\beta}_p$ is

$$\hat{\boldsymbol{\beta}}_p = (\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p'\boldsymbol{y}$$

and estimate of the residual variance is

$$\hat{\sigma}^2 = \frac{\boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}_p'\boldsymbol{X}_p'\boldsymbol{y}}{n - K - 1} = \frac{\boldsymbol{y}'[\boldsymbol{I} - \boldsymbol{X}_p(\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p']\boldsymbol{y}}{n - K - 1}$$

The properties of the estimates $\hat{\beta}_p$ and $\hat{\beta}^2$ from teh subset model

1. $E(\hat{\boldsymbol{\beta}}_p) = \boldsymbol{\beta}_p + (\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p'\boldsymbol{X}_r\boldsymbol{\beta}_r = \boldsymbol{\beta}_p + \boldsymbol{A}\boldsymbol{\beta}_r$, so the estimate $\hat{\boldsymbol{\beta}}_p$ is a biased estimate of $\hat{\boldsymbol{\beta}}_p$.

2. $\mathrm{Var}(\hat{\boldsymbol{\beta}}_p) = \sigma^2(\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}$, whereas $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, so matrix $\mathrm{Var}(\hat{\boldsymbol{\beta}}_p^*) - \mathrm{Var}(\hat{\boldsymbol{\beta}}_p)$ is positive semidefinite.
   The variance of the least -squares estimates of the parameters of the parameters in the full model are greater than or equal to the variances of the corresponding parameters in the subset model.

3. Since $\hat{\boldsymbol{\beta}}_p$ is biased estimate, we compare the precision of parameter estimate in terms of MSE. The mean square error of $\hat{\boldsymbol{\beta}}_p$ is

   $$MSE(\hat{\boldsymbol{\beta}}_p) = \sigma^2(\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1} + \boldsymbol{A}\boldsymbol{\beta}_r\boldsymbol{\beta}_r'\boldsymbol{A}'$$

   where $\boldsymbol{A} = (\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p'\boldsymbol{X}_r$. If the matrix $\text{Var}(\hat{\boldsymbol{\beta}}_r^*) - \boldsymbol{\beta}_r\boldsymbol{\beta}_r'$ is positive semidefinite, the matrix $\text{Var}(\hat{\boldsymbol{\beta}}_p^*) - MSE(\hat{\boldsymbol{\beta}}_p)$ is positive semidefinite.

4. The parameters $\hat{\sigma}_*^2$ from full model is an unbiased estimate of σ^2. For the subset model, $E(\hat{\sigma}^2) = \sigma^2 + \frac{\boldsymbol{\beta}_r'\boldsymbol{X}_r'[\boldsymbol{I} - \boldsymbol{X}_p(\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p']\boldsymbol{X}_r\boldsymbol{\beta}_{\boldsymbol{r}}}{n-p}$, which is biased.

5. Suppose we wish to predict the response at point $\boldsymbol{x}' = [\boldsymbol{x}_p', \boldsymbol{x}_r']$, the full model states a predicted value of $\hat{y}^* = \boldsymbol{x}'\hat{\boldsymbol{\beta}}^*$ with mean $\boldsymbol{x}'\boldsymbol{\beta}$, and a prediction variance $\text{Var}(\hat{y}^*) = \sigma^2[1 + \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}]$.
   The subset model will have $\hat{y} = \boldsymbol{x}_p'\hat{\boldsymbol{\beta}}_p$ with $E(\hat{y}) = \boldsymbol{x}_p'\boldsymbol{\beta}_p + \boldsymbol{x}_p'\boldsymbol{A}\boldsymbol{\beta}_r$. The prediction mean square error $MSE(\hat{y}) = \sigma^2[1 + \boldsymbol{x}_p'(\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}\boldsymbol{x}_p] + (\boldsymbol{x}_p'\boldsymbol{A}\boldsymbol{\beta}_r - \boldsymbol{x}_r'\boldsymbol{\beta}_r)^2$. So $\hat{y}$ will be a biased estimate of y unless $\boldsymbol{x}_p'\boldsymbol{A}\boldsymbol{\beta}_r = 0$.
   However, $\text{Var}(\hat{y}^*) \geq MSE(\hat{y})$ if $\text{Var}(\hat{\boldsymbol{\beta}}_r^*) - \boldsymbol{\beta}_r\boldsymbol{\beta}_r'$ is positive definite.

9.2 Criteria for Evaluation Subset Regression Models

There are a few criteria:

- Coefficient of Multiple determination R^2
- Adjusted R^2
- Residual mean square
- Mallows's C_p statistic $E[\hat{y}_i - E(y_i)]^2 = [E(y_i) - E(\hat{y}_i)]^2 + \text{Var}(\hat{y}_i)$
- AIC
- PRESS

9.3 All Possible Regression

Fit all regression equations involving one candidate regressor, two candidate regressors and so on. The equations are evaluated according to some suitable criterion and best regression model is selected.
If we assume that intercept β_0 is included and if there are K candidate regressors, there are 2^K total equations to be estimated.

9.4 Stepwise Regression Methods

We will look at forward selection, backward elimination and stepwise regression.

9.4.1 Forward Selection

This procedure beigns with the assumption that there are **no regressors** in the model other than intercept.
The first regressor is the regressor that will produce the largest value of the F statistics for testing significance of regression, where $F = \frac{SS_R(x_1)}{MS_{\text{Res}}(x_1)}$. Otherwise, we choose the variable with the smallest p value that is smaller than some threshold α_{IN}.
The second regressor chosen for entry is the one with the largest partial F statistics is

$$F = \frac{SS_R(x_2 \mid x_1)}{MS_{\text{Res}}(x_1, x_2)}$$

with the same threshold α_{IN} or F_{IN}.
This procedure terminates either when partial F statistics at a step does not exceed F_{IN} or the last candidate is added.

9.4.2 Backward Elimination

Backward elimination attempts to find a good model by working in the opposite direction. We begin with a model that includes all K candidate regressors, and the partial F statistics is computed for each regressor as if it were the last variable to enter the model. The smallest of these partial F statistics is compared with a preselected value F_{OUT} and if the smallest partial F value is less than F_{OUT} or largest p-value is larger than some α_{OUT}, the regressor is removed from the model.
We continue untill the smallest partial F value is not less than preselected cutoff value F_{OUT}.

9.4.3 Stepwise Regression

Stepwise regression is the modification of forward selection, in which at each step all regressors entered in the model previously are reassessed via their partial F statistics.
A regressor added at an earliear step may now be redundant, so if the partial F statistic for a variable is less than F_{OUT}, that variable will be dropped. Stepwise regression then requires two cutoff values, one for entering variables and one for removing.
Frequently we choose $F_{IN} > F_{OUT}$, making it relatively more difficult to add a regressor than to delete one.