Multiple Comparison:

Usage

We want to find out which of the contrasts of a factor is significantly non zero

Cases:

1) **No pre-hypothesized contrasts** specified: We will use general exploring to investigate all contrasts
2) Certain scientifically important contrasts which are **pre-specified** for investigation: Only need to find out whether those pre-specified contrasts are significant

Individual vs Family Wise Type I Error Rate:

Suppose that we have a test statistic T_j and critical value c_j, we reject the null hypothesis when $T_j \geq c_j$

Individual Type I error rate at α, each c_j such that

$$P(T_j \geq c_j) \leq \alpha$$

Family Wise Type I error rate at α, each c_j such that the probability of the union is less than α

$$P\left(\cup_{i=1}^{m} \{T_j \geq c_j\}\right) \leq \alpha$$

General Exploration:

$$H_0: \sum_{k=1}^{a} c_k\mu_k = 0 \ \forall c \in \mathcal{R}^a \ \ vs. H_1: At\ least\ 1\ c_k\mu_k \neq 0$$

What contrasts to pick for general exploration:

- For general exploration, we will look at the estimates that we get from the linear model and see which are the contrasts that we can take that could have a big difference that is significant.
- We can also pick contrasts that are of interest to us

Sheffe's Solution:

- **Critical Value Choice**:

$$c_\alpha = \sqrt{(a-1)F_{(a-1),n_{ERR}}(\alpha)}$$

n_{ERR} – df $of\ \hat{\sigma}^2$
a – Number of levels in the factor

- **Test Statistics**:
  Note that this test statistics will be used to test each of the levels
  Note that $T_c \sim t_{n_{ERR}}$

$$T_C = \frac{|\sum_{k=1}^{a} c_k\bar{Y}_k|}{\sqrt{\hat{\sigma}^2 \sum_{k=1}^{a} \frac{c_k^2}{n_k}}} = \frac{\hat{C}}{se(\hat{C})}$$

c_k – Contrast for treatment level (k)
$\bar{Y}_k$ – Sample mean at treatment level (k)
n_k – Sample size of treatment k
$\hat{\sigma}^2$ – Estimate of the error variance σ^2 which is also **MSE**
$\hat{C}$ - $|\sum_{k=1}^{a} c_k\bar{Y}_k|$
$se(\hat{C})$ - $\sqrt{\hat{\sigma}^2 \sum_{k=1}^{a} \frac{c_k^2}{n_k}}$

- **p-value**:

$$p = P\left(F_{(a-1),n_{ERR}} \geq \frac{T_C^2}{a-1}\right)$$

n_{ERR} – df $of\ \hat{\sigma}^2$
a – Number of levels in the factor

This is the p-value for 2 sided test and if we want one-sided just need to compare with $\frac{\alpha}{2}$ instead of α
We will reject if the p-value is smaller than our level of significance

- **Confidence Intervals**:
  The confidence interval for the contrast $\sum_{k=1}^{a} c_k\mu_k$

$$\hat{C} \pm se(\hat{C})\sqrt{(a-1)F_{(a-1),n_{ERR}}(\alpha)}$$

n_{ERR} – df $of\ \hat{\sigma}^2$
a – Number of levels in the factor
c_k – Contrast for treatment level (k)
$\bar{Y}_k$ – Sample mean at treatment level (k)
n_k – Sample size of treatment k
$\hat{\sigma}^2$ – Estimate of the error variance σ^2 which is also **MSE**
$\hat{C}$ - $|\sum_{k=1}^{a} c_k\bar{Y}_k|$
$se(\hat{C})$ - $\sqrt{\hat{\sigma}^2 \sum_{k=1}^{a} \frac{c_k^2}{n_k}}$

Pairwise Contrasts:

If we only want to find out if there are differences in any pairs of treatment
Only pairwise contrasts: $\mu_k - \mu_j, 1 \leq k < j \leq a$ are of interests

We want to find c_α such that:

$$P\left(\max_{k,j} \frac{|\bar{Y}_k - \bar{Y}_j|}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n_k} + \frac{1}{n_j}\right)}} \geq c_\alpha\right) \leq \alpha$$

n_{ERR} – df $of\ \hat{\sigma}^2$
$\bar{Y}_k$ – Sample mean at treatment level (k)
n_k – Sample size of treatment k
$\hat{\sigma}^2$ – Estimate of the error variance σ^2 which is also **MSE**

Tukey's Criterion:

$$q_{a,n_{ERR}} = \left(\frac{\sqrt{n}(\max \bar{Y}_i - \min \bar{Y}_i)}{\hat{\sigma}}\right)$$

$\bar{Y}_i$ – Sample mean of the sample i
$\hat{\sigma}$ – Estimate of error variance σ^2 which is also MSE
n_{ERR} – df $of\ \hat{\sigma}^2$
a – Number of levels in the factor

Assumptions:

1) Size of the samples are the same $n_i = n$
2) Tukey's Criterion is under the assumption that $\mu_1 = \cdots = \mu_a$

Studentised Range Distribution:

$q_{a,n_{ERR}}(\alpha)$ – Upper α – quantile of the Studentised Range Distribution which is also the Tukey's Criterion for pairwise comparison at level α

Note that the values can be read off with the degrees of freedom using a distribution table of the Studentised Range Distribution

Q-Statistic:

For the $\mu_i - \mu_j$ contrast:

$$Q_{ij} = \begin{cases} \frac{\sqrt{n}|\bar{Y}_i - \bar{Y}_j|}{\hat{\sigma}}, & if\ n_1 = \cdots = n_a = n \\ \frac{\sqrt{\tilde{n}_{ij}}|\bar{Y}_i - \bar{Y}_j|}{\hat{\sigma}}, & otherwise \end{cases}$$

$\bar{Y}_i$ – Sample mean of the sample i
n – Sample size of the sample i (Where in this case all the samples have the same size)
$\hat{\sigma}^2$ – Estimate of the error variance σ^2 which is also **MSE**
$\hat{\sigma}$ - $\sqrt{\hat{\sigma}^2}$
$\tilde{n}_{ij} = \frac{2n_in_j}{n_i+n_j}$ – Use this when the sample sizes are not the same

Relationship Between T Statistics and Q Statistics:

$$|T_{ij}| = \frac{Q_{ij}}{\sqrt{2}}$$

t-statistics for contrast $\mu_i - \mu_j$:

$$T_{ij} = \begin{cases} \frac{\sqrt{n}}{2}\frac{|\bar{Y}_i - \bar{Y}_j|}{\hat{\sigma}}, & if\ n_1 = \cdots = n_a = n \\ \sqrt{\frac{n_in_j}{n_i + n_j}}\frac{|\bar{Y}_i - \bar{Y}_j|}{\hat{\sigma}}, & otherwise \end{cases}$$

$\bar{Y}_i$ – Sample mean of the sample i
n_i – Sample size of the sample i
$\hat{\sigma}^2$ – Estimate of the error variance σ^2 which is also **MSE**
$\hat{\sigma}$ - $\sqrt{\hat{\sigma}^2}$

In regression coefficients:

$$T_{ij} = \begin{cases} \frac{\hat{\beta}_i}{sd(\hat{\beta}_i)}, & if\ j = 1 \\ \frac{\hat{\beta}_i - \hat{\beta}_j}{[Var(\hat{\beta}_i) + Var(\hat{\beta}_j) - 2Cov(\hat{\beta}_i, \hat{\beta}_j)]^{\frac{1}{2}}}, & otherwise \end{cases}$$

$\hat{\beta}_i$: $\mu_i - \mu_1$ when $i = 2, \cdots, a$

Significance of Pairwise Comparison:

The contrast $\mu_i - \mu_j$ is significant at level α if

$$Q_{ij} > q_{a,n_{(ERR)}}(\alpha) \text{ or } |T_{ij}| > \frac{q_{a,n_{(ERR)}}(\alpha)}{\sqrt{2}}$$

Confidence Interval for Tukey's Criteria:

For $\mu_i - \mu_1$:

$$\hat{\beta}_i \pm sd(\hat{\beta}_i)\frac{q_{a,n_{(ERR)}}(\alpha)}{\sqrt{2}}$$

For $\mu_i - \mu_j\ (i, j > 1)$:

$$(\hat{\beta}_i - \hat{\beta}_j) \pm [Var(\hat{\beta}_i) + Var(\hat{\beta}_j) - 2Cov(\hat{\beta}_i, \hat{\beta}_j)]^{\frac{1}{2}}\frac{q_{a,n_{(ERR)}}(\alpha)}{\sqrt{2}}$$

n_{ERR} – df $of\ \hat{\sigma}^2$
a – Number of levels in the factor
$\hat{\beta}_i : \mu_i - \mu_1$ when $i = 2, \cdots, a$

Bonferroni Method:

Suppose that we have k prespecified contrasts which are only ones of concern
We can control the Overall Type I Error Rate α for the k contrasts with the Bonferroni's Method

Adjustment for Individual Error Rate:

$$\alpha_j = \frac{\alpha}{k}$$

α – Overall Error Rate
k – Number of prespecified contrasts

Individual Tests:
We will compute the $T - Statistics$ for the contrasts that has been pre-specified
Note that the t-statics that we are computing is for 2-sided test so if we use R we will need to multiply by 2 if we are computing for 1 side

We can do either of the comparison, where will just scale up the p-value or we just scale down the significance level to be that of the individual error rate:

Compare the following with α:

$$p_j = k \times P(T_j \geq T_j^0)$$

p_j – p-value for the jth test
T_j – Test Statistics for the jth test
T_j^0 – Observed Test Statistics for jth test
k – Number of prespecified contrasts

Compare the following with $\frac{\alpha}{k}$:

$$p_j = P(T_j \geq T_j^0)$$

p_j – p-value for the jth test
T_j – Test Statistics for the jth test
T_j^0 – Observed Test Statistics for jth test
k – Number of prespecified contrasts

Efficient Criteria: When the test value is smaller than the other value. It means that at that level of significance, it can accept more values so we should use it

Since the value is smaller, it is able to accept a larger set of values while maintaining the family wise Type I Error rate at the same level

level as the Sheffe's Criteria

- In general, Tukey better than Scheffe
- Scheffe is only for general exploration
- Bonferroni is good when we are only testing a few contrasts

Other Special Multiple LRM:

Analysis of Covariance (ANOCOV):

Terms:

- Concomitant Variable – Factors that affect our response but we do not have control over

ANOCOV Model is a special multiple LRM that has both factor and quantitative predictors

Uses:

1) Comparison of treatment effects of factor predictors, we will need to adjust for the effects of the concomitant variables which have effect on the response as well so that the comparison is more efficient
2) When the response and quantitative predictors is studied in different categories. The regression functions are to be compared

$$y_i = \beta_0 + \sum_{j=2}^{k} \alpha_j u_{ij} + \gamma x_i + \epsilon_i$$

u_i – Dummy variables representing the factor variables
x_i – Represents the quantitative variable value
α_i – Regression coefficient for Factor Variable
γ – Regression coefficient for quantitative variable

Adjusting for effects of concomitant variable:

Suppose that we have a concomitant variable (X) that is a quantitative variable and we want to compare the treatment effects of 2 factors A and B

The model that we fit will be as follows:

$$y_i = \beta_0 + \alpha_2 u_{i2} + \lambda_2 v_{i2} + \gamma_{22}(u_{i2})(v_{i2}) + \beta_1 x_i + \epsilon_i$$

α – Regression Coefficient for Factor A
u – Indicator (Dummy) variable for Factor A
λ – Regression Coefficient for Factor B
v – Indicator (Dummy) variable for Factor B
γ – Regression Coefficient for Interaction term of Factor A and B
β_1 – Regression Coefficient for concomitant variable
x_i – Value for concomitant variable

Note that we fit x the first into the *lm* function when we are using the R code. It helps us to give a fairer comparison between the variables as it will help to explain for so of the regression that is caused by the concomitant variable so that we know what is the actual effect of the factor variable

Comparison of Regression Lines:

When the relation between a response variable and a quantitative variable is to be studied across different categories

We can use this method to do the following

1. Whether the regression lines have the same intercept
2. Whether the regression lines have the same slope

We can fit the following model:

$$Y = \beta_0 + \sum_{j=2}^{k} \alpha_j U_j + \beta X + \sum_{j=2}^{k} \xi_j U_j X + \epsilon$$

α – Regression Coefficient of categorical variable
U – Indicator (dummy) variable of categorical variable
β – Regression Coefficient of quantitative variable
X – Value of quantitative variable
ξ – Interaction term between factor variable and quantitative variable

We can compare the regression functions using the following models:

$$Y = \beta_0 + \beta X + \epsilon, \qquad for\ category\ 1$$
$$Y = (\beta_0 + \alpha_j) + (\beta + \xi_j)X + \epsilon, \qquad for\ category\ j\ (\geq 2)$$

α – Regression Coefficient of categorical variable
β – Regression Coefficient of quantitative variable
X – Value of quantitative variable
ξ – Interaction term between factor variable and quantitative variable

After we fit the linear models, we can get the coefficients and get the coefficients for the regression lines of different categories

Non-Linear Predictor Terms:

Polynomial Regression Models:
This is when our predictor variables does not have to conform to a linear relationship with the response variable

We could have the following model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{2i}^2 + \epsilon, \qquad i = 1, \cdots, n$$

$\beta -$ Regression Coefficients of the variable
x – Values of the quantitative variable

We should centralise the variables to reduce the collinearity

$$y_i = \beta_0 + \beta_1(x_{1i} - \overline{x_1}) + \beta_2(x_{2i} - \overline{x_2}) + \beta_3(x_{1i} - \bar{x}_1)^2 + \beta_4(x_{2i} - \bar{x}_2)^2 + \epsilon$$

$\beta -$ Regression Coefficients of the variable
x – Values of the quantitative variable
$\bar{x}$ – Mean of the ith quantitative variable

Steps:

1) Plot the scatterplot of the response against the predictor to see what is the relationship
2) Fit the relationship (i.e. cubic) into the *lm* model and specifying higher order with $I()$ function
3) Fit another relationship into the lm model and centralising it by finding the mean and fit under $I()$ function

Transformations that we can consider:

1) Polynomial relationships $(x^2, x^{\frac{1}{2}}, \dots)$ (Any powers of $\mathbb{Q}$)
2) Inverse Transformation
3) Log Transformation (Makes bigger values smaller)

We can compare the adjusted R^2 to check if the transformed model is a better fit

Piece-Wise Linear Model:

We can introduce an auxiliary X truncated at a point $X = c$ when the gradient of the line changes

$$\tilde{X} = (X - c)^+ = \begin{cases} X - c, & if\ X - c \geq 0 \\ 0, & otherwise \end{cases}$$

We can fit the following model:
Original: $Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i$
Augmented: $Y_i = \beta_0 + \beta_1 X_1 + \alpha_1 \tilde{X}_i + \epsilon_i$

β – Regression Coefficient of Quantitative Variable
α – Regression Coefficient to account to account for change in slope
X – Value of Quantitative Variable
$\tilde{X}$ – Value of the auxiliary variable

This allows the gradient to change at the point $X = c$ from β_1 to $\beta_1 + \alpha_1$

We can use this to test if there is a change in slope at a specified point say c by testing $H_0: \alpha_1 = 0$

Variable Selection: Criteria

We will use different criteria to assess if the predictor variable is relevant or irrelevant to our analysis

Terms:

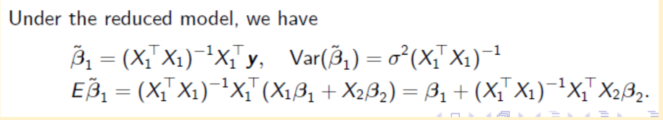
- **Relevant Predictor**: Can explain a proportion of the variation of the response variable which cannot otherwise be explained by other predictors
- **Causal Variable**: Relevant Predictor and it is the variable that is directly causing the variation of the predictor
- **Surrogate Variable**: There is some underlying casual variable that is unobserved that is causing the variation of both the response and the surrogate variable. This will cause there to be a correlation between the surrogates and the response variable
- **Irrelevant Variable**: Neither casual nor surrogates of casual variables. But it could have correlations with the response due to their correlation with relevant variables
- **Collinearity:** Phenomena that some predictors are highly or moderately correlated among themselves

Purpose:
Out of a large number of candidate predictors, we want to select predictors for the following goals:

- Identify relevant predictors (E.g. QTL Mapping in Genetic Studies)
  - Focuses on the accuracy of the selection
- Build a model for prediction (E.g. Disease Diagnostics)
  - Focuses of the accuracy of prediction

Under-Fitting:

When the model that we have selected does not contain all the relevant predictors



Under the reduced model, we have

$$\tilde{\beta}_1 = (X_1^\top X_1)^{-1} X_1^\top \boldsymbol{y}, \quad \mathrm{Var}(\tilde{\beta}_1) = \sigma^2 (X_1^\top X_1)^{-1}$$

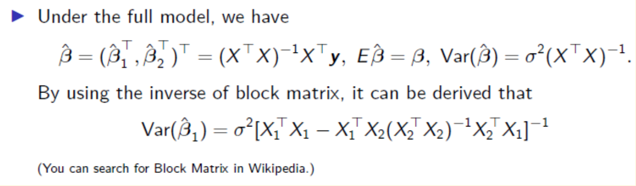
$$E\tilde{\beta}_1 = (X_1^\top X_1)^{-1} X_1^\top (X_1\beta_1 + X_2\beta_2) = \beta_1 + (X_1^\top X_1)^{-1} X_1^\top X_2\beta_2.$$

Issue: Biased Estimator of LSE of β_1
There will be a bias of $(X_1^T X_1)^{-1}\ X_1^T X_2 \beta_2$

It will have a lower variance as compared to an over-fitted model but there is higher bias

Over-Fitting:

When the model that we have selected contains irrelevant predictors in addition to all the relevant predictors



- Under the full model, we have

$$\hat{\beta} = (\hat{\beta}_1^\top, \hat{\beta}_2^\top)^\top = (X^\top X)^{-1} X^\top \boldsymbol{y},\ E\hat{\beta} = \beta,\ \mathrm{Var}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}.$$

By using the inverse of block matrix, it can be derived that

$$\mathrm{Var}(\hat{\beta}_1) = \sigma^2 [X_1^\top X_1 - X_1^\top X_2 (X_2^\top X_2)^{-1} X_2^\top X_1]^{-1}$$

(You can search for Block Matrix in Wikipedia.)

Issue: Higher Variance

It will always have an unbiased estimate of LSE of β but it will have higher variance

Effects of Over-Fitting and Under-Fitting of Model:

Under-Fitting increases the bias of the prediction
Over-Fitting increases the variance of the prediction

Suppose we have a model M

$$\hat{y}_m = X_M \hat{\beta}_M = X_M (X_M^T X_M)^{-1} X_M^T y$$

X_M – Design Matrix
β_M – Parameter Matrix
$E(y) = \mu$

Properties:

$$E(\hat{y}_M) = \ X_M \hat{\beta}_M = X_M (X_M^T X_M)^{-1} X_M^T \mu$$

$$Var(\hat{y}_M) = \sigma^2 X_M (X_M^T X_M)^{-1} X_M^T$$

$$\sum_{i=1}^{n} Var(\hat{y}_{iM}) = \sigma^2 tr(X_M (X_M^T X_M)^{-1} X_M^T) = |M|\sigma^2$$

Note that tr takes the diagonal elements of the matrix where $X_M (X_M^T X_M)^{-1} X_M^T$ is the covariance matrix

σ^2 – Variance of the model ($E(\epsilon^2)$)
$|M|$ = Number of predictors in Model M + 1 (For the σ^2 estimate)
X_M – Design Matrix
β_M – Parameter Matrix
$E(y_i) = \mu_i$
$\hat{y}_{iM}$ – Prediction of ith value under Model M

Sum of Prediction Squared Error (SPSE):
If we make use of Model M to predict n unobserved $y's$ with the same design matrix X_M

$$SPSE = \sum_{i=1}^{n} E(y_{n+i} - \hat{y}_{iM})^2 = \sum_{i=1}^{n} E(y_{n+i} - \mu_i)^2 + \sum_{i=1}^{n} E(\mu_i - \ \hat{y}_{iM})^2$$

$$= n\sigma^2 + |M|\sigma^2 + \sum_{i=1}^{n} (\mu_i - \mu_{iM})^2$$

Intuition: We will check the expected deviation for the unobserved value of y_{n+i} and its predicted value

y_{n+i} – Unobserved response with the same value of predictor x_i as that of y_i
$\hat{y}_{iM}$- Predicted value for the new value using model M
$|M|\sigma^2$ – Variance Term
$\sum_{i=1}^{n} (\mu_i - \mu_{iM})^2$ – Bias Term

Principle of Variable Selection:

Accuracy of the prediction is measured by the **SPSE** which consists of a variance component and bias component

Accuracy can also be measured by the **MSE** which also has a variance and bias component

$$MSE = Var(\hat{\mu}) + Bias^2(\hat{\mu})$$

We will want to strike a balance between the variance and bias during variable selection so that SPSE or MSE is minimised

However it is hard to compute SPSE or MSE in practical case so we can make use of model selection criteria to assess the accuracy instead

Selection Criteria:

R^2 & R_a^2:
Suitable for models with the same number of predictors

$$R^2 = \frac{SSR}{SST}$$

$$R_a^2 = R^2 - \frac{p}{n - p - 1}(1 - R^2)$$

Estimates the proportion of the population variance that is explained by the regression model

Criteria: Select the one with the largest value

Issues:
R^2 – Always increases as the number of predictors increases, the criterion will always select the model containing all predictors. Does not strike a balance between variance and bias

R_a^2 – Even though it is not strictly increasing as the number of predictors increases, it still increases when a predictor have even a small contribution to the explained variation of the model

It is not good to discern when we have different number of predictors as it as selects the full model

Mallow's C_p (Complexity Parameter):

Depends on a good estimate of σ^2

Make use of $\sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2$ to estimate $SPSE$ instead since we do not know $\sum_{i=1}^{n}(y_{n+i} - \hat{y}_{iM})^2$

$$\widehat{SPSE} = \sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2 + 2|M|\sigma^2$$

$$C_p = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \hat{y}_{iM})^2 - n + 2|M|$$

$\hat{\sigma}^2$ – Unbiased estimate of σ^2

σ^2 – Variance of the model $(E(\epsilon^2))$

$|M| =$ Number of predictors in Model M + 1 (For the σ^2 estimate)

y_i – Observed value for ith y value

$\hat{y}_{iM}$ – Prediction of ith value under Model M

Criteria:

If model M is the correct model then $E[C_p] \approx |M|$

Choose the model with the smallest value of C_p which is closest to $|M|$ (The one that is closest to $|M|$)

Akaike's Information Criterion (AIC):

Developed by minimizing the Kullback-Leibler distance between the current model and "true" model

Kullback-Leibler distance:

$$I(f, g_M) = \int f(y)\log\left(\frac{f(y)}{g_M(y)}\right) dy$$

$g_M(y)$ – Model M

$f(y)$ – True Model approximated by the empirical distribution

AIC:

$$AIC = \ -2\log L(\hat{\beta}_M, \hat{\sigma}_M) + 2|M|$$

$|M| =$ Number of predictors in Model M + 1 (For the σ^2 estimate)

L – Likelihood Function of Normal Distribution

$\hat{\beta}_M$ – MLE estimate of β under model M, with likelihood function L

$\hat{\sigma}_M$ – MLE estimate of σ under model M, with likelihood function L

For Multiple LRM with normality assumption:

$$AIC = \ -2\log L(\hat{\sigma}_M) + 2|M| + C$$

$|M| =$ Number of predictors in Model M + 1 (For the σ^2 estimate)

L – Likelihood Function of Normal Distribution

$\hat{\sigma}_M$ – MLE estimate of σ under model M, with likelihood function L

$$\hat{\sigma}_M^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \ \hat{y}_{iM})^2$$

$$C = n(\ln(2\pi) + 1)$$

Log-Likelihood Function $\log(L)$:

$$L = -\frac{n}{2}\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \ \hat{y}_{iM})^2$$

Estimation of $\hat{\beta}_M$ & $\hat{\sigma}_M$ from LSE:

$$\hat{\sigma}_M^2 = \frac{df_{LSE}}{df_{MLE}}\hat{\sigma}_{LSE}^2 = \frac{n_{ERR}}{n}\hat{\sigma}^2$$

$$\hat{\beta}_M = \hat{\beta}_{LSE}$$

n_{ERR} – Degrees of Freedom of the error term

n – Total number of observations

Criteria:

Select the Model with the lowest AIC

Bayesian Information Criteria (BIC):

Developed under Bayesian Framework

Marginal Density of data y given model M:

$$m(y|M) = \int L(y; \beta_M)\pi(\beta_M)\, d\beta_M$$

M – Model

$\pi(\beta_M)$ – Prior Distribution on Parameters of Model M

β_M – Regression Coefficient on Model M

L – Likelihood Function based on Normal Distribution

Posterior Probability:

$$p(M|y) = \frac{m(y|M)p(M)}{\sum_{M\in\mathcal{M}} p(M)\, m(y|M)}$$

$\mathcal{M}$ - Set of all possible models

$p(M)$ – Prior Distribution on Model $M \in \mathcal{M}$

M – Model

BIC:

$$BIC = -2\log L(\hat{\beta}_M, \hat{\sigma}_M) + |M|\ln(n)$$

$|M| =$ Number of predictors in Model M + 1 (For the σ^2 estimate)

L – Likelihood Function of Normal Distribution

$\hat{\beta}_M$ – MLE estimate of β under model M, with likelihood function L

$\hat{\sigma}_M$ – MLE estimate of σ under model M, with likelihood function L

n – Number of observations

Note that it is quite similar to AIC just that $\ln(n)$ is 2 under AIC

Estimation of $\hat{\beta}_M$ & $\hat{\sigma}_M$ from LSE:

$$\hat{\sigma}_M^2 = \frac{df_{LSE}}{df_{MLE}}\hat{\sigma}_{LSE}^2 = \frac{n_{ERR}}{n}\hat{\sigma}^2$$

$$\hat{\beta}_M = \hat{\beta}_{LSE}$$

n_{ERR} – Degrees of Freedom of the error term

n – Total number of observations

Criteria:

Select the Model with the lowest BIC

Extended Bayesian Information Criterion (EBIC):

For models with huge number of predictors and takes into account the model class complexity

$$EBIC = \ -2\log L(\hat{\beta}_M, \hat{\sigma}_M) + |M|\ln(n) + 2\gamma\ln\binom{p}{|M|}$$

$|M| =$ Number of predictors in Model M + 1 (For the σ^2 estimate)

L – Likelihood Function of Normal Distribution

$\hat{\beta}_M$ – MLE estimate of β under model M, with likelihood function L

$\hat{\sigma}_M$ – MLE estimate of σ under model M, with likelihood function L

n – Number of observations

$$\gamma = \left(1 - \left(\frac{\ln n}{2\ln p}\right)\right)$$

p – Number of predictor variables

Estimation of $\hat{\beta}_M$ & $\hat{\sigma}_M$ from LSE:

$$\hat{\sigma}_M^2 = \frac{df_{LSE}}{df_{MLE}}\hat{\sigma}_{LSE}^2 = \frac{n_{ERR}}{n}\hat{\sigma}^2$$

$$\hat{\beta}_M = \hat{\beta}_{LSE}$$

n_{ERR} – Degrees of Freedom of the error term

n – Total number of observations

Criteria:

Select the Model with the lowest EBIC

Cross Validation (CV):
Based on the estimated prediction error

Idea:
Suppose we have a data set, we divide the data into different parts.
Each part is alternatively reserved as testing data for the assessment of the model
Remaining parts are used as training data to estimate the model
We will then compute a score which approximates the predictor error

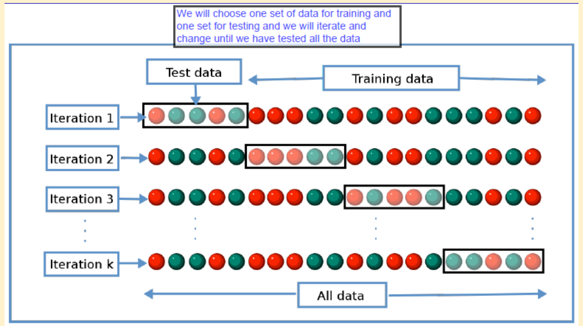

We will choose one set of data for training and one set for testing and we will iterate and change until we have tested all the data

Figure 1 Illustration of Cross Validation

Leave-out-one Cross Validation:
Procedure:
Test data consists of only 1 observation
Training data consists of the remaining (n-1) observations
We will repeat this n times

k-fold Cross Validation:
Preferred over leave-out-one CV as it has better theoretical properties and less computation demands

Procedure:
Data is divided into k parts
Each time one part is taken as the test data
Remaining k-1 parts are used as training data
We will repeat this k times

Scores:
For a model: $y = x^T\beta + \epsilon$

Leave-out-one CV Score:

$$CV = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - x_i^T\hat{\beta}^{-i}\right]^2$$

$\hat{\beta}^{-i}$ – Estimate of β obtained by leaving out the i^{th} data point (y_i, x_i)
n – Number of observations

k-fold CV Score:

$$CV = \frac{1}{k}\sum_{j=1}^{k}\left|\left|y_j - X_j\hat{\beta}^{-j}\right|\right|^2$$

$\hat{\beta}^{-j}$ – Estimate obtained by using the remaining data after omitting the jth group's data points (y_j, X_j)
k – Number of groups that we are breaking it up into

Note that we could have random samples with replacement from the data for k-fold validation so each time we pick $\frac{n}{k}$ number of points. This is what is used in R. The CV score could change as well since we are taking random samples.
We could also have equally divided fixed parts where every value will be used

Criteria:
We will select the model with the lowest CV Score which corresponds to the one with the smallest prediction error

Variable Selection: Procedure

Terms:

- **Full Model**: Contains all the predictor variables
- **Null Model**: Contains none of the predictor variables ($lm(y \sim 1)$)

General Model Selection Process:

1) By using certain mechanism, generate a small number of candidate models (relatively compared with 2^p (total number of possible models))
2) Using a certain selection criteria, select among the candidate models

Selection by Removing Redundant Predictors:

Can be used when p is not very large

Options:

1) **Removing all at once**:
Fit the full model.
Remove all the predictors with p-value bigger than a certain level α (E.g. 0.05)

2) **Sequentially removing**:
Fit the full model
Remove the predictor with the largest p-value which is bigger than α
Refit the model with the remaining predictors
Repeat this procedure until none of the predictors have a p-value bigger than α

After removing the redundant predictors, the remaining model will be the final model

Sequential Procedures:

Starts with a Null or Full Model and add/remove predictors based on whether they whether the reduction of the residual sum of squares

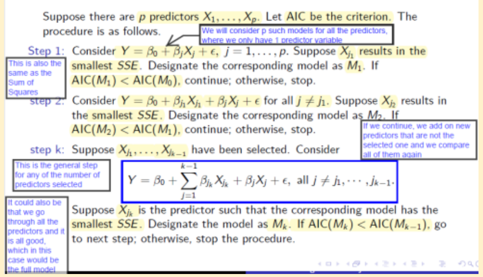
Criterion: Can use anything other than R^2 or R_a^2 as the will always select the full model. (AIC, BIC, EBIC, CV etc.) We choose AIC for this course

Forward Selection:

Procedure:

1) Start with Null Model with no predictors
2) Augment the model by including additional predictors one at a time, each time, the predictor with the largest contribution to reducing the residual sum of squares is added
3) The augmented model is compared with the previous model by some criterion. If the augmented model is favoured by the criterion, the process continue;
4) Otherwise, it is stopped if the previous model is favoured and the previous model is taken as the final selected model

Detailed Procedure (Chap 7 Slide 12):



Suppose there are p predictors $X_1, \ldots, X_p$. Let AIC be the criterion. The procedure is as follows.

We will consider p such models for all the predictors, where we only have 1 predictor variable

Step 1: Consider $Y = \beta_0 + \beta_j X_j + \epsilon$, $j = 1, \ldots, p$. Suppose X_{j_1} results in the smallest SSE. Designate the corresponding model as M_1. If $\text{AIC}(M_1) < \text{AIC}(M_0)$, continue; otherwise, stop.

This is also the same as the Sum of Squares

step 2: Consider $Y = \beta_0 + \beta_{j_1} X_{j_1} + \beta_j X_j + \epsilon$ for all $j \neq j_1$. Suppose X_{j_2} results in the smallest SSE. Designate the corresponding model as M_2. If $\text{AIC}(M_2) < \text{AIC}(M_1)$, continue; otherwise, stop.

If we continue, we add on new predictors that are not the selected one and we compare all of them again

step k: Suppose $X_{j_1}, \ldots, X_{j_{k-1}}$ have been selected. Consider

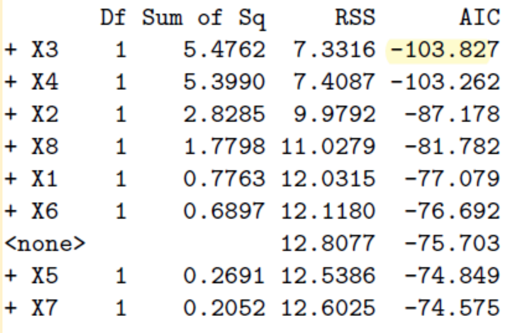
This is the general step for any of the number of predictors selected

$$\boxed{Y = \beta_0 + \sum_{j=1}^{k-1} \beta_{j_k} X_{j_k} + \beta_j X_j + \epsilon, \text{ all } j \neq j_1, \cdots, j_{k-1}.}$$

It could also be that we go through all the predictors and it is all good, which in this case would be the full model

Suppose X_{j_k} is the predictor such that the corresponding model has the smallest SSE. Designate the model as M_k. If $\text{AIC}(M_k) < \text{AIC}(M_{k-1})$, go to next step; otherwise, stop the procedure.

Meaning of Results in Each Forward Step:



```
       Df Sum of Sq     RSS      AIC
+ X3    1    5.4762  7.3316 -103.827
+ X4    1    5.3990  7.4087 -103.262
+ X2    1    2.8285  9.9792  -87.178
+ X8    1    1.7798 11.0279  -81.782
+ X1    1    0.7763 12.0315  -77.079
+ X6    1    0.6897 12.1180  -76.692
<none>              12.8077  -75.703
+ X5    1    0.2691 12.5386  -74.849
+ X7    1    0.2052 12.6025  -74.575
```

Figure 2 Example Results

- **Sum of Sq**: The amount of sum of square increased by adding the corresponding variable
- **RSS**: Residual Sum of Squares of the model including all the variables in the current model with the corresponding variable added
  - Note that SSE is a decreasing function with the number of variables that are added. Therefore, we could have a higher RSS for the previous model but we should compare the AIC
- **AIC**: AIC of the current model computed by the formula $n\ln(\hat{\sigma}^2_{MLE}) + 2|M|$ where $|M| = p + 1$ and p is the number of predictor variables

We will stop once the previous model has the smallest AIC

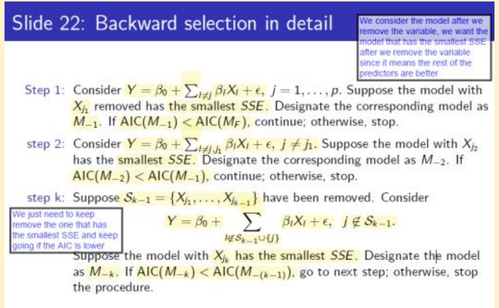
Backward Selection:

Opposite of Forward Selection and we will start with the Full Model instead

Procedure:

1) Start with Full Model with all predictors
2) Augment the model by removing predictors one at a time, each time, the model that after removing the predictor and have the lowest RSS (removing it reduces the RSS the most) will be removed.
3) The reduced model is compared with the previous model by some criterion. If the reduced model is favoured by the criterion, the process continue;
4) Otherwise, it is stopped if the previous model is favoured and the previous model is taken as the final selected model

Detailed Procedure (Chap 7 Slide 22):



Slide 22: Backward selection in detail

We consider the model after we remove the variable, we want the model that has the smallest SSE after we remove the variable since it means the rest of the predictors are better

Step 1: Consider $Y = \beta_0 + \sum_{l \neq j} \beta_l X_l + \epsilon$, $j = 1, \ldots, p$. Suppose the model with X_{j_1} removed has the smallest SSE. Designate the corresponding model as M_{-1}. If $\text{AIC}(M_{-1}) < \text{AIC}(M_F)$, continue; otherwise, stop.

step 2: Consider $Y = \beta_0 + \sum_{l \neq j, j_1} \beta_l X_l + \epsilon$, $j \neq j_1$. Suppose the model with X_{j_2} has the smallest SSE. Designate the corresponding model as M_{-2}. If $\text{AIC}(M_{-2}) < \text{AIC}(M_{-1})$, continue; otherwise, stop.

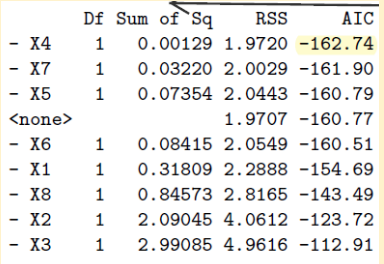
step k: Suppose $\mathcal{S}_{k-1} = \{X_{j_1}, \ldots, X_{j_{k-1}}\}$ have been removed. Consider

$$Y = \beta_0 + \sum_{l \notin \mathcal{S}_{k-1} \cup \{j\}} \beta_l X_l + \epsilon, \quad j \notin \mathcal{S}_{k-1}.$$

We just need to keep remove the one that has the smallest SSE and keep going if the AIC is lower

Suppose the model with X_{j_k} has the smallest SSE. Designate the model as M_{-k}. If $\text{AIC}(M_{-k}) < \text{AIC}(M_{-(k-1)})$, go to next step; otherwise, stop the procedure.

Meaning of Results in Each Backward Step:



```
       Df Sum of Sq    RSS     AIC
- X4    1   0.00129 1.9720 -162.74
- X7    1   0.03220 2.0029 -161.90
- X5    1   0.07354 2.0443 -160.79
<none>              1.9707 -160.77
- X6    1   0.08415 2.0549 -160.51
- X1    1   0.31809 2.2888 -154.69
- X8    1   0.84573 2.8165 -143.49
- X2    1   2.09045 4.0612 -123.72
- X3    1   2.99085 4.9616 -112.91
```

Figure 3 Example Results

- **Sum of Sq**: The amount of sum of square reduced by deleting the corresponding variable
  - Find the one with the smallest sum of square since it means that the variable does not really explain the model that well
- **RSS**: The residual sum of square of the model eliminating all the deleted variables so far including the current corresponding variable
- **AIC**: AIC of the current model computed by the formula $n\ln(\hat{\sigma}^2_{MLE}) + 2|M|$ where $|M| = p + 1$ and p is the number of predictor variables

We will stop once the previous model has the smallest AIC

Stepwise Selection:

Upwards Stepwise Selection:

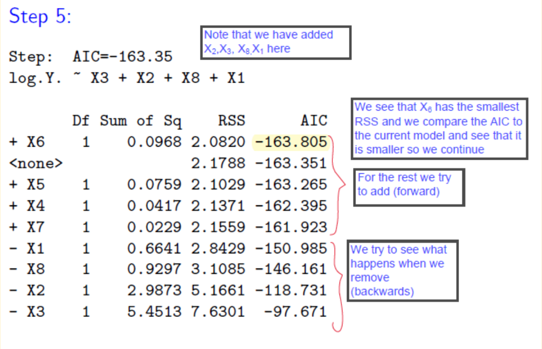
1) Start with Null Model
2) In the forward selection procedure, once a new predictor is included in the model
3) We will apply the backward procedure to the augmented model time until no predictors can be removed, then the procedure proceeds to the next forward step

Downwards Stepwise Selection:

1) Start with Full Model
2) In the backwards selection procedure, once a predictor is removed from the current model
3) We will apply the forward procedure to the reduced model time until no new predictors can be added, then the procedure proceeds to the next backward step

Actual Selection:

Note that during the actual selection, we can perform the addition and removal at the same time



Step 5:

```
Step:  AIC=-163.35
log.Y. ~ X3 + X2 + X8 + X1

       Df Sum of Sq    RSS      AIC
+ X6    1    0.0968 2.0820 -163.805
<none>              2.1788 -163.351
+ X5    1    0.0759 2.1029 -163.265
+ X4    1    0.0417 2.1371 -162.395
+ X7    1    0.0229 2.1559 -161.923
- X1    1    0.6641 2.8429 -150.985
- X8    1    0.9297 3.1085 -146.161
- X2    1    2.9873 5.1661 -118.731
- X3    1    5.4513 7.6301  -97.671
```

Note that we have added X_2,X_3, X_8,X_1 here

We see that X_6 has the smallest RSS and we compare the AIC to the current model and see that it is smaller so we continue

For the rest we try to add (forward)

We try to see what happens when we remove (backwards)

For instance, we have selected $X_1 + X_2 + X_3 + X_8$

We can consider the AIC for the model for removing the selected variables and adding new predictor variables at the same time and selecting the one with the lowest AIC

Penalised Likelihood Approach:

Least Absolute Shrinkage and Selection Operator (LASSO):

We will try to reduce the variance by adding a little bias so that our model does not have as high of a variability.

We will reduce the slope of the model and the β values and by doing so, we will cause some of the β values to be 0 and it will let us know which are variables that are important

For LRM: $y = X\beta + \epsilon$

We will reduce the following term with the addition of $\lambda||\beta||_1$ to our least squares cost function

$$\frac{1}{n}||y - X\beta||_2^2 + \lambda||\beta||_1$$

$||\cdot||_2$ - L_2 norm, the normal least squares that we use

$||\cdot||_1$ - L_1 norm, just take the modulus of the β estimates

λ – Penalty Parameter to be chosen through the LASSO process and we assess it through the CV Score

$y - X\beta$ – Residuals of the estimates

$||\beta||_1 = \sum_{j=1}^{p}|\beta_j|$ – If β has dimensions p

Choice for λ:

If $\lambda = 0$, LASSO estimator is the same as the LSE since the second part is not there

If $\lambda = \infty$, all the components of β are estimated as 0

For certain nonzero λ, some components will be non-zero and some will be zero. The variables with non-zero coefficients are the selected variables

Criterion for λ:

Model for Prediction: CV

Identify Important Variables: BIC, AIC

Number of p is huge: EBIC

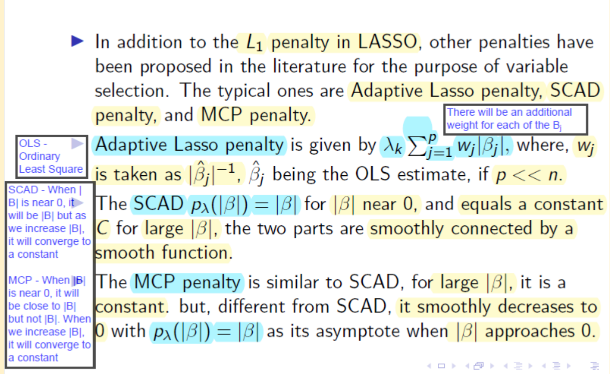
Procedure:

1) Specify a sequence of λ values
2) At each value, carry out the penalized minimization
3) This gives us a model with certain selected variables for each value
4) Use the selection criterion to determine the best choice for λ
5) Find out the model that is selected with that choice of λ and the variables with non-zero coefficients are selected

Note:

LASSO is not very good as it does not have selection consistency. (When $n \to \infty$, the number of relevant predictors selected is not the actual number of predicted variables and does not have a probability of 1)

Other Penalty Functions:



- In addition to the L_1 penalty in LASSO, other penalties have been proposed in the literature for the purpose of variable selection. The typical ones are Adaptive Lasso penalty, SCAD penalty, and MCP penalty.

Adaptive Lasso penalty is given by $\lambda_k \sum_{j=1}^{p} w_j|\beta_j|$, where, w_j is taken as $|\hat{\beta}_j|^{-1}$, $\hat{\beta}_j$ being the OLS estimate, if $p << n$.

The SCAD $p_\lambda(|\beta|) = |\beta|$ for $|\beta|$ near 0, and equals a constant C for large $|\beta|$, the two parts are smoothly connected by a smooth function.

The MCP penalty is similar to SCAD, for large $|\beta|$, it is a constant. but, different from SCAD, it smoothly decreases to 0 with $p_\lambda(|\beta|) = |\beta|$ as its asymptote when $|\beta|$ approaches 0.

There will be an additional weight for each of the B_j

OLS - Ordinary Least Square

SCAD - When |B| is near 0, it will be |B| but as we increase |B|, it will converge to a constant

MCP - When |B| is near 0, it will be close to |B| but not |B|. When we increase |B|, it will converge to a constant

Model Diagnostics:

When we want to check for any model inadequacies

Raw Materials for Diagnostics:

Fitted Values:

$$\hat{y}_i = x_i^T\hat{\beta}\,, \qquad i = 1, \cdots, n$$

Hat Matrix:

$$H = X(X^TX)^{-1}X^T$$

X – Design Matrix of the model $y = X\beta + \epsilon$

Hat Values:

The i^{th} diagonal element of H

The following is the hat value of the ith observation

$$h_{ii} = x_i^T(X^TX)^{-1}x_i$$

X – Design Matrix of the model $y = X\beta + \epsilon$

x_i^T – ith row of X

h_{ii} is the squared Mahalanobis distance of the ith point to the centroid

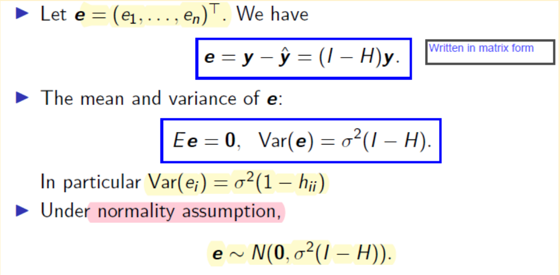
Mahalanobis Distance:

$$(\mathbf{x}_i - \bar{\mathbf{x}})^\top \Sigma_{\mathbf{X}}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) = (\mathbf{x}_i^\top - \mathbf{1}^\top X)[X^\top(I - \frac{\mathbf{1}\mathbf{1}^\top}{n})X]^{-1}(\mathbf{x}_i - X^\top\mathbf{1})$$

Pearson's Residuals:

Difference between the observed values and the fitted values

$$e_i = y_i - \hat{y}_i, \qquad i = 1, \cdots, n$$



- Let $\mathbf{e} = (e_1, \ldots, e_n)^\top$. We have

$$\boxed{\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (I - H)\mathbf{y}.}$$

Written in matrix form

- The mean and variance of $\mathbf{e}$:

$$\boxed{E\mathbf{e} = \mathbf{0}, \quad \mathrm{Var}(\mathbf{e}) = \sigma^2(I - H).}$$

In particular $\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$

- Under normality assumption,

$$\mathbf{e} \sim N(\mathbf{0}, \sigma^2(I - H)).$$

Studentised Residuals:

Standardising the Pearson Residuals

$$r_i' = \frac{e_i}{sd(e_i)} = \frac{(y_i - \hat{y}_i)}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

e_i – Pearson Residual

$sd(e_i)$ – Standard Deviation of Pearson Residual

h_{ii} – Hat value of the ith observation

Studentised Deleted Residuals:

We delete the ith observation and see how much is the deviation from the new fitted model and standardise it

$$r_i^* = \frac{y_i - \hat{y}_{i(i)}}{sd(y_i - \hat{y}_{i(i)})} = \frac{(y_i - \hat{y}_i)}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}$$

h_{ii} – Hat value of the ith observation

$\hat{y}_{i(i)}$ – Predicted value of the fitted model with the ith observation deleted

$\hat{\sigma}_{(i)}$ – Counterpart of $\hat{\sigma}$ when the ith observation is deleted

Cook's Distance:

Checks whether the regression coefficient increases a lot after deleting the ith observation

$$d_i = (\hat{\beta}_{(i)} - \hat{\beta})^T(X^TX)(\hat{\beta}_{(i)} - \hat{\beta}) \,/\, (p\hat{\sigma}^2)$$

$\hat{\beta}$ – Estimated Regression Coefficient

$\hat{\beta}_{(i)}$ – Estimated Regression Coefficient when the ith observation is deleted from the data

X – Design Matrix

p – Number of predictor variables

$\hat{\sigma}^2$ – MSE of the model

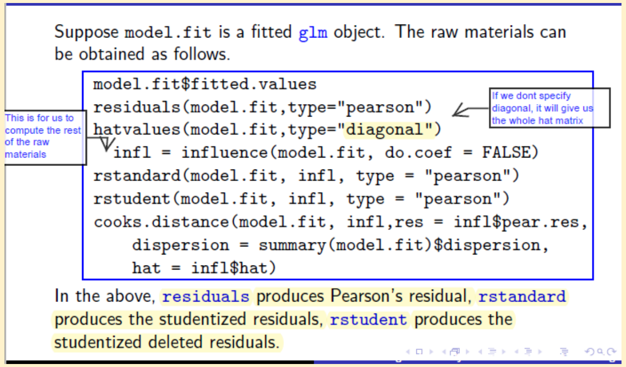
Variance Inflation Factor:

$X_1, \cdots, X_p$ is a regression model with p predictors

R_k^2 – Coefficient of determination of the model after we delete the kth predictor variable and regress on X_k such that $X_k = \beta_0 + \sum_{j \neq k} \beta_j X_j + \epsilon$

$$VIF_k = \frac{1}{1 - R_k^2}$$

R Code for Raw Materials:



Suppose `model.fit` is a fitted `glm` object. The raw materials can be obtained as follows.

```
model.fit$fitted.values
residuals(model.fit,type="pearson")
hatvalues(model.fit,type="diagonal")
 infl = influence(model.fit, do.coef = FALSE)
rstandard(model.fit, infl, type = "pearson")
rstudent(model.fit, infl, type = "pearson")
cooks.distance(model.fit, infl,res = infl$pear.res,
    dispersion = summary(model.fit)$dispersion,
    hat = infl$hat)
```

If we dont specify diagonal, it will give us the whole hat matrix

This is for us to compute the rest of the raw materials

In the above, `residuals` produces Pearson's residual, `rstandard` produces the studentized residuals, `rstudent` produces the studentized deleted residuals.

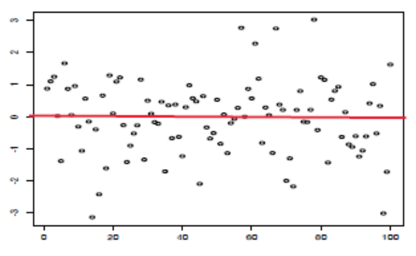
Systematic Discrepancies:

We will need to check these to ensure that our model is adequate for prediction else we cannot make any inference from the model

Null Pattern of Residual Plots:

The points will be scattered around randomly and would be scattered evenly within a horizontal band around 0.

This happen when there is no discrepancies at all



1) **Non-Linearity**. The regression function is not linear (in terms of the predictors)

   Ways to check:
   Plot of Pearson Residuals against fitted values
   Plot of Pearson Residuals against predictor variables
   Scatter plot of response against predictor variables

   Analysis:
   If there are any non-linear trend, it indicates that there is non-linearity

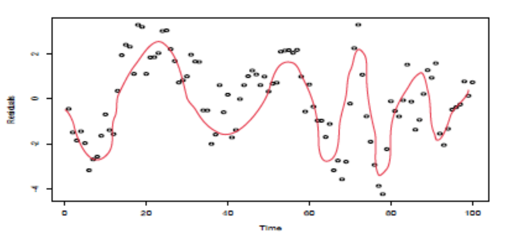
2) **Non-Homogeneity**. The error terms do not have constant variance

   Ways to check:
   Plot of Pearson Residuals against fitted values
   Plot of Pearson Residuals against predictor variables

   Analysis:
   If the variances are not a constant, the vertical range of the residuals will have an obvious change along the x-axis. If we draw a straight line down for the ranges, we will see that the values are not constant

3) Error terms are **not independent**

Ways to check:
Plot of Pearson Residuals against time/space at which the observations are obtained. (Note that this can only be checked when we have the time/spatial order



Analysis:
So long as there is a pattern that is non null, then there is a correlation of the values over time

4) **Non-Normality**. The error terms do not have normal distribution

Ways to check:
Distribution plot of the residuals: Boxplot, histogram, etc.
Normal Probability plot of the residuals

Analysis:
If normality holds, the points of the normal probability plot (or Q-Q plot) should roughly fall along a straight line

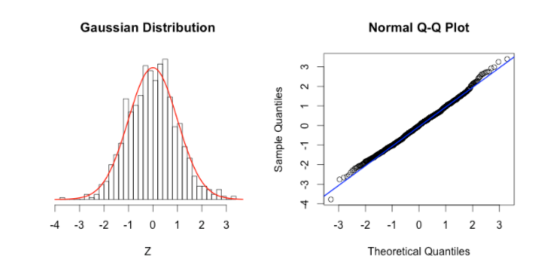


Figure 4 Normally Distributed

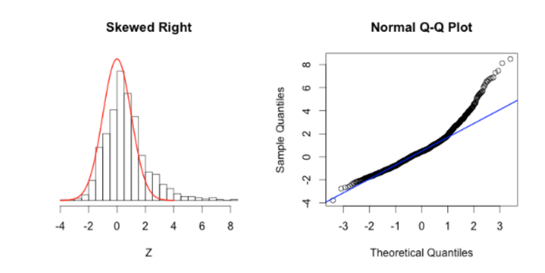


Figure 5 Right Skewed

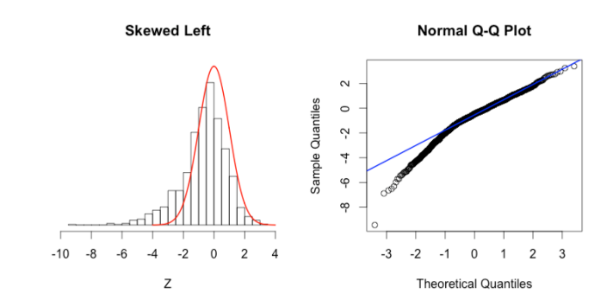


Figure 6 Left Skewed

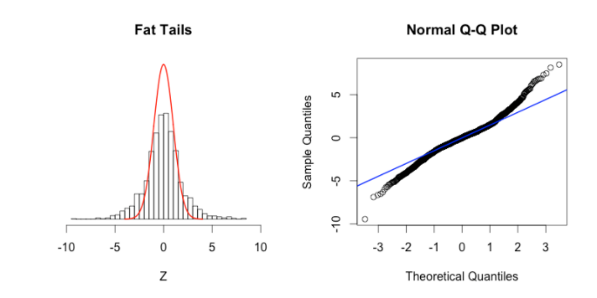


Figure 7 Fat Tails

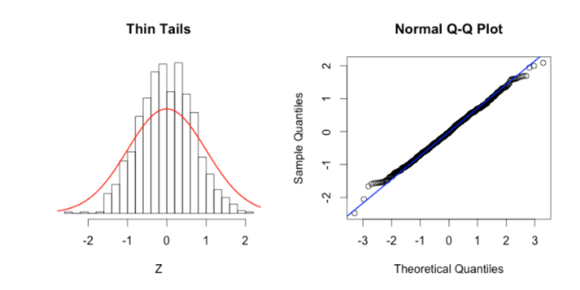


Figure 8 Thin Tails

5) **Missing Predictors**. Some Important predictors are omitted from the model

Ways to check:
Residual Plots against other predictors that are not included in the model

Analysis:
If there is a linear trend for the plot, it will mean that the predictor is missing

Detection of Outliers:

Nature of Outliers:

1) **Leverage**: Whether a points is far away from the major cluster in the x – space

Measurement: Hat Values
Criteria: A point is considered high leverage if $h_{ii} > 2p/n$
Comparisons: Find the largest value

2) **Consistency**: Whether a point is consistent in terms of fitting in the (x, y) space (Whether the point will be close to the new fitted line)

Measurement: Studentised Deletion Residual
It can be used because it helps us to measure how far away the point is from the fitted value if we regress on the values without the stated value
Criteria:
Comparisons: Find the largest absolute value

3) **Influence**: Whether a point highly affects the fitting of the model

Measurement: Cook's Distance
It measures how much the change in the slope when we remove the point and the larger the deviation, the larger the distance
Criteria:
Comparisons: Find the largest value

Assessments for Outliers:

Informal:
Plot normal probability plots of studentized deletion residual, hat values and Cook's distance

Points at the extremes of the plot that deviate from the major trend are considered as outliers
We can roughly gauge which are the ones that are extreme and we will check to see which are the values that we want to use the formal test for

Formal:
Introduce an indicator variable:

$$u = \begin{cases} 1, & for\ the\ ith\ unit \\ 0, & otherwise \end{cases}$$

Where this indicator variable is 1 it is the ith value

Test:
If the p-value of the coefficient of u in the linear predictor is less than α then it shows that the ith value is an outlier else it is not

Sequential: We can do it sequentially where we
1) Plot normal probability plots of the outlier measures and identify one outlier at a time,
2) Test the significance,
3) Remove it if necessary
4) Refit the model
5) Repeat 1 – 4 until we find an outlier that is not significant and we stop

All at once:

1) Plot normal probability plots of the outlier measures and identify all the potential outliers
2) Test the significance,
3) Remove it if necessary
4) Refit the model

Note:

- Every model must be diagnosed
- Model assumption, fitting and diagnostics are recursive processes.
- The diagnostics either confirms the adequacy of the current model or points the direction for model modification
- Any systematic discrepancies must be remedied.
- Influential outliers should be excluded from the fitting of the model. Other outliers must be investigated to find out the possible causes
- Models can only be used if no discrepancy can be found by all possible diagnostic techniques

Weighted LSE & Multicollinearity:

Weighted Least Square Estimation:

Remedy when the assumption of common variance of ϵ_i is not satisfied
When we have constant variance: $\epsilon \sim N(0, \sigma^2 I)$

Consider an unequal variance model: $y = X\beta + \epsilon$

The variance matrix of ϵ is given by:

$$\Sigma = \sigma^2 \begin{pmatrix} w_1^{-1} & 0 & \cdots & 0 \\ 0 & w_2^{-1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & w_n^{-1} \end{pmatrix}$$

w_i – Weight of each of the error terms
Note $w_i's$ are unequal weights

Remedy when weights are known:

$$y = X\beta + \epsilon$$

$w_i's$ are known

$$W = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}$$

Multiply by $W^{\frac{1}{2}}$ throughout

$$W^{\frac{1}{2}} = W^{\frac{1}{2}} X\beta + W^{\frac{1}{2}}\epsilon$$

Let $\tilde{y} = W^{\frac{1}{2}} y$, $\tilde{X} = W^{\frac{1}{2}} X$ and $\tilde{\epsilon} = W^{\frac{1}{2}}\epsilon$

$$Var(\tilde{\epsilon}) = W^{\frac{1}{2}} \Sigma W^{\frac{1}{2}} = \sigma^2 I$$

We can see that if we perform that transformation, it will be a constant variance model. The following will be a model with constant variance

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

Weighted Least Square Estimates (WLSE):

Minimise $\left|\left|\tilde{y} - \tilde{X}\beta\right|\right|_2^2$ so that we can get a weighted least squares estimate. Note that the $\hat{\beta}_W \neq \beta$

$$\hat{\beta}_W = \left(\tilde{X}^T \tilde{X}\right)^{-1} \tilde{X}^T \tilde{y} = (X^T W X)^{-1} X^T W y$$

Expression of Target Function as Sum of Squares:

$$\left|\left|\tilde{y} - \tilde{X}\beta\right|\right|_2^2 = \sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2$$

We are trying to find the difference between the observed value and the fitted values

Relationship between weights and variance:
We note that w_i will be smaller if there is a higher variance for the ith term because we will want to reduce the term so if $(y_i - x_i^T \beta)^2$ is large, w_i will be small.
Intuition: We want to give smaller weights for those that causes larger variance so that the data will not vary that much when we have those values. It will increase the bias but reduce the variability of the model

Distribution of WLSE:

Properties:

$$E(\hat{\beta}_W) = \beta$$
$$Var(\hat{\beta}_W) = \sigma^2 (X^T W X)^{-1}$$
$$\hat{\beta}_W \sim N\left(\beta, \sigma^2 (X^T W X)^{-1}\right)$$

Estimation of σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \left|\left|\tilde{y} - \tilde{X}\beta\right|\right|_2^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2$$

n – Number of observations
p – Number of predictors
w_i – Weight of observation i

We can make similar inference on β as a constant variance LRM

Estimation of Unknown Weights:

Variance is usually a function of the mean, we can estimate it using the residuals and fitted values

1) Fit the regression model by unweighted least squares and obtain the residuals r and the fitted value $\hat{y}$
2) Regress the log of the absolute residual on the log of the fitted values, e.g.

$$\ln|r_i| = \alpha_0 + \alpha_1 \ln \hat{y}_i + e_i$$

r_i – Residual value of the ith value
$\hat{y}_i$ – Fitted value of y
α – Regression Coefficient for fitted values of y

If smallest $\hat{y}_i < 0$, replace $\hat{y}_i = \hat{y}_i - \min \hat{y}_i + c$, where c is some positive constant and we can take $c = 1$

3) Estimate the standard deviation and weights as

$$s_i = e^{\hat{\alpha}_0} \hat{y}_i^{\hat{\alpha}_1}$$
$$s_i = e^{\ln|r_i|}$$
$$w_i = \frac{1}{s_i^2}$$

s_i – Standard Deviation of the ith value
w_i – Weight of the ith value
$\hat{y}_i$ – Fitted value of y
α – Regression Coefficient for fitted values of y

Note that s_i can also be computed as the exponent of the fitted values for the model: $\ln|r_i| = \alpha_0 + \alpha_1 \ln \hat{y}_i + e_i$

4) Repeat the procedure if necessary

When there are replicate of predictor values:

Make use of the sample variance for estimation of the weights

Fit the following model:

$$\ln s_i = \alpha_0 + \alpha_1 \ln \bar{y}_i + e_i$$

s_i – Sample Deviation of the values at the ith position
$\bar{y}_i$ – Sample Mean of the values at the ith position
α – Regression Coefficient for fitted values of y

Standard Deviation of the ith predictor values can be estimated as:

$$s_i = e^{\hat{\alpha}_0} \bar{y}_i^{\hat{\alpha}_1}$$
$$s_i = e^{\ln s_i}$$
$$w_i = \frac{1}{s_i^2}$$

s_i – Standard Deviation of the ith predictor value
w_i – Weight of the ith predictor value
$\bar{y}_i$ – Sample Mean of the values at the ith predictor value
α – Regression Coefficient for sample mean of the predictor values

Procedure for Fitting:
1) Fit an unweighted regression line and analyze the residual to identify non-constant variance
2) Use the residual to estimate the weights
3) Fit a weighted regression line using the estimated weights
4) Compare the unweighted and weighted regression models

Diagnostics for weighted model:

Results from Weighted LSE:

$$\hat{y} = X\hat{\beta}_W \;\&\; e = y - X\hat{\beta}_W$$

For the transformed model:

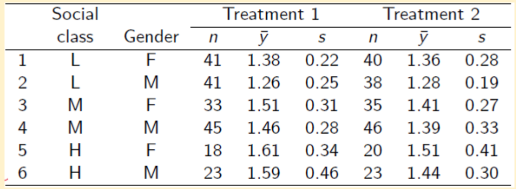
$$\hat{y}_W = W^{\frac{1}{2}} \hat{y}; \quad e_w = W^{\frac{1}{2}} e$$

Note that the Weighted LSE result will only give us a model that adjusts to give use the least square estimate. However, we will still need to multiply the weights to the fitted and residuals to check for the constancy of the variance

Inference of Models:
However, we will be able to make predictions with the current weighted model
Since the previous model does not have constant variance, we cannot use it for any inference and we can only use the newly fitted model for inference

When we have grouped values (categorical variables):
Suppose we have Factor 1 with α levels and Factor 2 with β levels

We will have 2 way ANOVA and for each cell we have
n_{ij} – Size of the cell
$\bar{y}_{ij}$ – Mean of the ij combination
s_{ij} – Standard Deviation for the ij combination



	Social class	Gender	Treatment 1			Treatment 2		
			n	$\bar{y}$	s	n	$\bar{y}$	s
1	L	F	41	1.38	0.22	40	1.36	0.28
2	L	M	41	1.26	0.25	38	1.28	0.19
3	M	F	33	1.51	0.31	35	1.41	0.27
4	M	M	45	1.46	0.28	46	1.39	0.33
5	H	F	18	1.61	0.34	20	1.51	0.41
6	H	M	23	1.59	0.46	23	1.44	0.30

Figure 9 Example of the Table

Consider each $\bar{y}_{ij}$ to come from the ij population with mean μ_{ij} and variance σ_i^2
Variance of $\bar{y}_i = \frac{\sigma_i^2}{n}$ (Variance of sample mean)

Approaches:
1) Assume σ_i^2 are equal across groups, which should be subjected to checking, then we can take the weights as $w_i = n_i$ since we still need to consider the variance caused by the sample size
2) If σ_i^2 are not equal, the weights can be taken as $w_i = n_i/s_i^2$ so that the variance becomes a constant
3) If there is a relationship between the σ_i^2 and μ_i suppose $\sigma_i^2 = \gamma(\mu_i)$, the weights can be taken as $w_i = n_i/\hat{\gamma}(\bar{y}_i)$ using the following method

Fit the following model:

$$\ln s_i = \alpha_0 + \alpha_1 \ln \bar{y}_i + e_i$$

s_i – Sample Deviation of the values at the ith position
$\bar{y}_i$ – Sample Mean of the values at the ith position
α – Regression Coefficient for fitted values of y

Standard Deviation of the ith predictor values can be estimated as:

$$s_i = e^{\hat{\alpha}_0}\bar{y}_i^{\hat{\alpha}_1}$$
$$s_i = e^{\ln s_i}$$
$$w_i = \frac{n}{s_i^2}$$

s_i – Standard Deviation of the ith predictor value
w_i – Weight of the ith predictor value
$\bar{y}_i$ – Sample Mean of the values at the ith predictor value
α – Regression Coefficient for sample mean of the predictor values

Multicollinearity:

Where there is the existence of correlation among predictor variables such as pairwise correlation, linear dependence of one predictor variable on other predictor variables

Issues:
- If multicollinearity is serious, it could cause serious problems in regression analysis
- If one predictor is perfectly linearly dependent on the other predictor, it causes X^TX to be singular and we will not be able to compute the LSE
- High Multicollinearity will cause X^TX to be nearly singular and makes the LSE extremely unstable, a slight change in the observations could change the LSE significantly. We will have high variance for LSE which makes it inaccurate and useless

Informal Indications of Multicollinearity:
1) Large changes in the estimated regression coefficients when a predictor is added or deleted, or an observation is altered or deleted
   a. Try to delete some variables and see if there is a significant change
2) Nonsignificant results in individual test on the regression coefficient for important predictor variables which we know beforehand that should be significant
3) Estimated regression coefficient with an algebraic sign that is opposite of that expected from theoretical consideration or prior experience
4) Large coefficients of sample correlation between pairs of predictors in the correlation matrix
   a. Find Pearson Correlation Coefficient between pairs or plot a Scatterplot

Variance Inflation Factors (VIF):

LSE of β_j:

$$\hat{\beta}_j = \frac{x_j^T(I - H_{-j})y}{x_j^T(I - H_{-j})x_j}, \qquad Var(\hat{\beta}) = \frac{\sigma^2}{x_j^T(I - H_{-j})x_j}$$

H_{-j} – Projection Matrix of X_{-j}
X_{-j} – Submatrix of X without its jth column
$\hat{\beta}_j$ – Regression Coefficient of the jth variable

$SSE_j = x_j^T(I - H_{-j})x_j$ – SSE when x_j is regressed on X_{-j}

Manipulation of SSE_j:

$$SSE_j = SST_j - SSR_j = SST_j\left(1 - \frac{SSR_j}{SST_j}\right) = SST_j(1 - R_j^2)$$

$$SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$$

R_j^2 – Coefficient of multiple determination of x_j regressed on X_{-j}

Analysis:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SSE_j} = \frac{\sigma^2}{SST_j(1 - R_j^2)} = \frac{\sigma^2}{SST_j}\left(\frac{1}{1 - R_j^2}\right)$$

- If x_j is uncorrelated with X_{-j}, $R_j^2 = 0 \Rightarrow Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j}\left(\frac{1}{1-0}\right) = \frac{\sigma^2}{SST_j}$
- If x_j is correlated with X_{-j}, the variance is inflated by a factor

$$VIF_j = \frac{1}{1 - R_j^2}$$

R_k^2 – Coefficient of determination of the model after we delete the kth predictor variable and regress on X_k such that $X_k = \beta_0 + \sum_{j\neq k}\beta_j X_j + \epsilon$

VIF_j is called the variance inflation factor and if x_j can be explained by the other values, then the variance will be inflated since $1/0.c = d$ which is a positive number and will increase the variance of β_j
Alternative Formula:

$$VIF_j = \frac{(n-1)\widehat{Var}(X_j)\widehat{Var}(\beta_j)}{\hat{\sigma}^2}$$

$\hat{\sigma}^2$- Mean Squared Error of the model
n – Number of observations
$\widehat{Var}(\beta_j)$ – Variance of the fitted β
$\widehat{Var}(X_j)$ – Variance of X_j against other X_i (The x-values that we have from the data)

Using VIF to detect Multicollinearity:
1) Compute the VIF for the various variables
2) Find values of VIF that are above 5 (which indicates a considerate multicollinearity)
3) Remove the predictor with the largest VIF above 5
4) Repeat Step 1-3 until we no longer have a VIF above 5 and the final variables

Ridge Regression:

Used to solve the problem of X^TX being nearly singular, similar to LASSO regression where we modify the slope

Solution: We add a diagonal matrix to X^TX, $X^TX + \lambda I$

Ridge Regression Estimator:
$\hat{\beta}_{RIG} = (X^TX + \lambda I)^{-1}X^Ty$ where $\lambda > 0$ is a parameter to be chosen

Penalised Sum of Squares:
The ridge regression estimator is the minimiser of the penalised sum of squares

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}(x_{ij}\beta_j)\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$
$$= ||y - X\beta||^2 + \lambda||\beta||^2$$

p – Number of predictor variables
n – Number of observations
X – Design Matrix
β – Regression Coefficients
λ – Parameter that should be determined

Properties:

Mean of $\hat{\beta}_{RIG}$:

$$E\hat{\beta}_{\text{RIG}} = (X^\top X + \lambda I)^{-1} X^\top X\beta = \beta - \lambda(X'X + \lambda I)^{-1}\beta.$$

Biased:
The estimator is not unbiased

$$\beta - E\hat{\beta}_{\text{RIG}} = \lambda(X^\top X + \lambda I)^{-1}\beta.$$

λ increase, Bias increase

- Let Q be the orthogonal matrix such that $X^\top X = Q\Delta Q^\top$, where $\Delta = \underbrace{\text{Diag}(\tau_0, \tau_1, \ldots, \tau_p)}_{>0}$. Thus

$$\beta - E\hat{\beta}_{\text{RIG}} = Q\begin{pmatrix} \frac{\lambda}{\tau_0+\lambda} & 0 & \cdots & 0 \\ 0 & \frac{\lambda}{\tau_1+\lambda} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{\lambda}{\tau_p+\lambda} \end{pmatrix} Q^\top\beta.$$

- Let $\tilde{\beta}_j$ denote the jth component of $Q^\top\beta$. We have

$$\|\beta - E\hat{\beta}_{\text{RIG}}\|^2 = \sum_{j=0}^{p}\left(\frac{\lambda}{\tau_j+\lambda}\right)^2 \tilde{\beta}_j^2$$

Lambda increase, bias increase

λ increase, Variance Decreases:

- The variance matrix of $\hat{\beta}_{\text{RIG}}$ is given by

$$\begin{aligned} \text{Var}(\hat{\beta}_{\text{RIG}}) &= \sigma^2(X^\top X + \lambda I)^{-1}X^\top X(X^\top X + \lambda I)^{-1} \\ &= \sigma^2 Q\begin{pmatrix} \frac{\tau_0}{(\tau_0+\lambda)^2} & 0 & \cdots & 0 \\ 0 & \frac{\tau_1}{(\tau_1+\lambda)^2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{\tau_p}{(\tau_p+\lambda)^2} \end{pmatrix} Q^\top \end{aligned}$$

Thus,

$$\text{TR}(\text{Var}(\hat{\beta}_{\text{RIG}})) = \sigma^2 \sum_{j=0}^{p} \frac{\tau_j}{(\tau_j+\lambda)^2}.$$

lambda increase, the variance decreases

Balance between Variance and Bias:

$$\begin{aligned} \text{MSE} &= \sum_{j=0}^{p} \text{MSE}(\hat{\beta}_j) \\ &= \overset{\text{Bias}}{\|\beta - E\hat{\beta}_{\text{RIG}}\|^2} + \overset{\text{Variance}}{\text{TR}(\text{Var}(\hat{\beta}_{\text{RIG}}))}. \end{aligned}$$

We need to strike a balance between the variance and the bias to get the minimum MSE

Selection of λ:
Make use of Cross-Validation (CV) as MSE cannot be readily used as a criterion and CV helps to check the prediction error

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[y_i - x_{(i)}^T \beta_{-i}(\lambda)]^2$$

$\hat{\beta}_{-i}(\lambda)$ – Ridge Regression Estimate of β with parameter λ by deleting the ith observation
n – Number of observations
$x_{(i)}^T$- ith row vector of the design matrix X

Choice: The best λ is the minimizer of $CV(\lambda)$

R Code:
Make use of lm.ridge(formula, data, lambda, …)

- formula – Expression for the regression model
- data – Data frame used for the model
- lambda – Scalar or vector of ridge constants. Supply a sequence of values that we want to test for the lambda values

Outputs:

- lambda – Vector of lambda values
- GCV – Vector of GCV values

To get the best lambda, get the GCV scores and find the index that has the smallest CV score then use that to find the value of lambda

Fit the final model with the best value of λ using lm.ridge again and set lambda to only be the best one. We will be able to get the coefficients from the model from there

Remarks:

- Mainly used for building model for predictions
- Cannot be used to assess the importance or effects of the predictor variables
- Ridge regression estimates cannot be used to construct confidence interval and conduct hypothesis testing
- If we need to make inference on the effects of the predictors, we can adopt the strategy of removing predictors having large VIF (Anything to do with inference, use VIF)

Variable Transformation:

Can be used as a remedy for non-normality. The violation of normality usually also goes together with the violation of constancy of variance (because variance could be dependent on mean and there it violates normality and variance is not constant so it violates homogeneity) and we can solve both at the same time

Difference from Weighted LSE:
Weighted LSE: If the variance is proportional to some constant, transform both X and y
Variance Stabilisation Transformation: If the variance depends on the mean, transform the response variable

Variance Stabilization Transformation:
Help to rectify simultaneously the discrepancy in both normality and constancy of variance

Suppose we have a random variable with variance σ^2 depending on its mean μ: $\sigma^2 = V(\mu)$ where it is a function of μ

Transformation:
$h(X)$ be the transformation such that the variance becomes approximately independent of its mean

Expand at μ (Using Taylor Series):

$$h(X) \approx h(\mu) + h'(\mu)(X - \mu)$$

Treating $h(\mu)$ as the mean of $h(X)$

$$Var\big(h(X)\big) \approx E[h(X) - h(\mu)]^2 \approx [h'(\mu)]^2 V(\mu)$$

We want $Var\big(h(X)\big)$ to be a constant, we just set $[h'(\mu)]^2 V(\mu) = 1$ for convenience:

$$h'(\mu) = \frac{1}{\sqrt{V(\mu)}} \Rightarrow h(\mu) = \int \frac{1}{\sqrt{V(\mu)}}\, d\mu$$

Sub back into $h(X)$ to find the transformation

Common Transformation:
Based on the nature of the data, the residual plot may not give an idea of the relationship

1) **Proportion Response**: $h(X) = \sin^{-1}\sqrt{X}$
   a. Has similar nature to Binomial Random Variable
   b. Variance depends on mean
   c. Variance is smaller if the mean is closer to either zero or 1, larger if mean is closer to 0.5
2) Count Response: $h(X) = \sqrt{X}$
   a. Can be approximated by Poisson Random Variable
   b. Poisson Random Variable, $V(\mu) = \mu$ and for general case, the variance is proportional to the mean $V(\mu) = c\mu$

Steps:
1) Plot out the residual plot against fitted values and identify that there is non-constant variance and if the variance increases as mean increases, we can think of a transformation
2) Think of the relation between the variance and mean
3) Use the relation to derive a transformation using the variance stabilisation transformation
4) Plot the new residual plot against fitted values with the transformation and see if the variance is roughly constant
5) If variance is roughly constant, then we can accept the transformation and use the transformation instead

Two Sample t-test:
To check if the 2 groups have significantly different values

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

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

$\bar{X}_i$ – Mean of the ith group
n_i – Number of observations for the ith group
s_i – Standard deviation for the ith group

Box-Cox Transformation:

For certain non-normal continuous random variables, a transformation cannot be determined solely by the data type. We can carry out a preliminary analysis to determine the transformation

Transformation:

$$h(Y) = \frac{Y^\lambda - 1}{\lambda}$$

λ – Can be determined from the data
Note when $\lambda = 0$, $h(Y) = \ln(Y),\ since \lim_{\lambda \to 0} \frac{Y^{\lambda} - 1}{\lambda} = \ln(Y)$

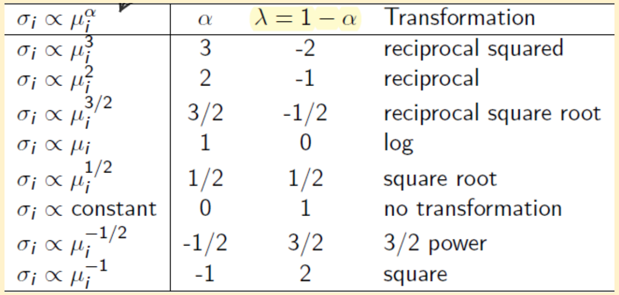
Determining λ from power variance function:
Consider a power function of σ_i and μ_i

$$\sigma_i \propto \mu_i^\alpha$$

σ_i – Standard deviation of the ith treatment
μ_i – Mean of the ith treatment

$$\lambda = 1 - \alpha$$

We determine the relationship and find the power then we compute the value of λ



$\sigma_i \propto \mu_i^\alpha$	α	$\lambda = 1 - \alpha$	Transformation
$\sigma_i \propto \mu_i^3$	3	-2	reciprocal squared
$\sigma_i \propto \mu_i^2$	2	-1	reciprocal
$\sigma_i \propto \mu_i^{3/2}$	3/2	-1/2	reciprocal square root
$\sigma_i \propto \mu_i$	1	0	log
$\sigma_i \propto \mu_i^{1/2}$	1/2	1/2	square root
$\sigma_i \propto$ constant	0	1	no transformation
$\sigma_i \propto \mu_i^{-1/2}$	-1/2	3/2	3/2 power
$\sigma_i \propto \mu_i^{-1}$	-1	2	square

Determining of α:

Method I: If the observations are **grouped**, for each group, compute s_i and $\bar{y}_i$. Fit the regression model:

$$\ln s_i = \beta_0 + \beta_1 \ln \bar{y}_i + \epsilon_i$$

s_i – Standard deviation of the ith group
$\bar{y}_i$ – Mean of the ith group
β – Regression Coefficient

$$s_i = e^{\beta_0} \bar{y}_i^{\beta_1}$$
$$\Rightarrow \sigma_i = \mu_i^{\beta_1}$$

s_i – Standard deviation of the ith group
$\bar{y}_i$ – Mean of the ith group
β – Regression Coefficient

Note that $e^{\hat{\beta}_0}$ is a constant so we derive the above equation and $\hat{\beta}_1$ is the estimate for α. We can take the rough value of $\hat{\beta}_1$ round off to nearest 0.5 to take as α estimate and compute λ from there

If the observations **ungrouped**, compute the r_i and $\hat{y}_i$ value for each of the observations. Fit the regression model:

$$\ln |r_i| = \beta_0 + \beta_1 \ln \hat{y}_i + \epsilon_i$$

r_i – Residual of the ith value
$\hat{y}_i$ – Fitted value of the ith value
β – Regression coefficient

$$r_i = e^{\beta_0} \hat{y}_i^{\beta_1}$$
$$\Rightarrow \sigma_i = \mu_i^{\beta_1}$$

r_i – Residual of the ith value
$\hat{y}_i$ – Fitted value of the ith value
β – Regression coefficient

Note that $e^{\hat{\beta}_0}$ is a constant so we derive the above equation and $\hat{\beta}_1$ is the estimate for α. We can take the rough value of $\hat{\beta}_1$ round off to nearest 0.5 to take as α estimate and compute λ from there

Method II: (Only works for grouped observations)
Select a few number of α values, say $\alpha_k, k = 1, \cdots, K$
For each k,

$$R_k = \frac{\max_i \frac{s_i}{\bar{y}_{i.}^{\alpha_k}}}{\min_i \frac{s_i}{\bar{y}_{i.}^{\alpha_k}}}$$

s_i – Standard deviation of the ith group
$\bar{y}_i$ – Mean of the ith group
$\frac{s_i}{\bar{y}_{i.}^{\alpha k}}$ – Ratio of standard deviation of ith group and mean of ith value where the mean is raised to the power of k

Selection: We will select the α_k with the smallest R_k.We find the ratio of the maximum over the minimum so that if the variance becomes constant then the value R_k should be small since both values should be quite similar

Direct Determination of λ:

1) For grouped observations:

- Select a few number of λ values, e.g. $\lambda_k,\ k = 1, \cdots, K$
- For each k, make the transformation $y_{ij} \rightarrow y_{ij}^{\lambda_k}$
- With the transformed data, compute $s^2_{\lambda_k\, i}, i = 1, \cdots, g$
- Select λ_k such that the following value is closest to 1

$$\frac{\max s^2_{\lambda_k\, i}}{\min s^2_{\lambda_k\, i}}$$

We find the one with the ratio of max/min to be 1 so that the variance of the transformation will be constant throughout since the largest and smallest value have a ratio of 1

2) For ungrouped observations:

- Select a few number of λ values, e.g. $\lambda_k,\ k = 1, \cdots, K$
- For each k, make the transformation $y_i \rightarrow y_i^{\lambda_k}$. Analyse the regression models with $y_{i_k}^{\lambda_k}$ as the response variable
- Select the λ_k such that the residual plot for checking constancy of variance exhibits a random pattern

Make the transformation, regress on the transformed values, check the residual against fitted values plot