Statistics and Data Analysis

Phu Sakulwongtana

1 Introduction

Definition 1.1. **(Linear Model)** Given the explainatory variable x, the model is

$$Y_i = \beta_i + \beta_i x_i + e_i$$

for $i = 1, 2, \dots, N$ as Y_i denotes the i-th observation and Y corresponds to value x_i, where we have:

- Explanatory variable x
- e_i is the error associated with i-th observation.
- β_0, β_1 are unknown variable.

The model can be defined as $\mathbb{E}[Y_i] = \beta_0 + \beta_1 x_i$ where have $i = 1, \dots, N$, as we can define: $e_i = Y_i - \mathbb{E}[Y_i]$. Furthermore, we can write the prediction $\mathbb{E}[Y_i | X_i = x_i]$ in order to make clear that the x-value are random variable. We are interested in how Y depends on x.

Remark 1. **(General Linear Regression)** In general, there are m explainatory variables labelled $x_1, \dots, x_m$ by consider the following model:

$$Y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_{im} + c_i \qquad \text{where} \qquad i = 1, \dots, M$$

If some of the explainatory variables are discrete and some are continuous then we have a general linear regression, as we can describe the model in compacted manner $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + c_i$

Remark 2. **(Error in General Linear Regression)** The error usually assume to be independent, normally distribution with 0 means and constant variance σ, thus Y_i is independent normally distributed random variable:

$$\mathbb{E}[Y_i] \boldsymbol{x}_i^T \boldsymbol{\beta} \qquad \text{var}(Y_i) = \sigma^2$$

Remark 3. **(Example Model)** The model might compare the mean of 2 groups i.e a 2 sample problem. For j-th observation in the i-th group, we have:

$$Y_{ij} = \nu_i + e_{ij} \qquad i = 1, 2 \quad j = 1, \dots, n$$

with the assuption about the error, then Y_{ij} is independent of $\mathcal{N}(\mu_i, \sigma^2)$. The above model is:

$$Y_{ij} = x_{ij1}\mu_1 + x_{ij2}\mu_2 + e_{ij} \qquad \text{where} \qquad x_{ijk} = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{if } k \neq i \end{cases}$$

x_{ijk} is called the dummy variable or indicator variable. Furthermore, we can extend the problem into $I \geq 2$ groups where:

$$Y_{ij} = x_{ij1}\mu_1 + \cdots + x_{ijI}\mu_I + e_{ij}$$

2 Inference for Linear Model

Definition 2.1. (Least Square Estimation) The least square established of the element $\boldsymbol{\beta}$ minimizer the following sum of square:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2$$

with respected to elements of β. Note that the method can be written in the matrix of the form $Y = X\boldsymbol{\beta} + \boldsymbol{e}$, where we have:

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix} \qquad X = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix} \in \mathbb{R}^{N \times p} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \qquad \boldsymbol{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_N \end{pmatrix}$$

as we have X is known as design matrix in the context of design experiment. The optimization objective can be written as:

$$S(\boldsymbol{\beta}) = \boldsymbol{e}^T \boldsymbol{e} = (\boldsymbol{Y} - X\boldsymbol{\beta})^T (\boldsymbol{Y} - X\boldsymbol{\beta})$$

Remark 4. If we consider the differentiate with respected to $\boldsymbol{\beta}$ and we have $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T Y$ assuming $p < N$ and so $X^T X$ is full rank. We can show that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator with $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$. The covariance matrix of $\hat{\beta}$ if in addition $e_i : i = 1, \dots, N$ are independent with constant variance σ^2:

$$\mathrm{v}(\hat{\boldsymbol{\beta}}) = \sigma^2 (X^T X)^{-1}$$

and so we can assume the error are independent normally distribution with mean of 0 and constant variance of σ^2, then we have: $\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$

Theorem 2.1. ***(Gauss-Markov)*** *If $\psi = \boldsymbol{c}^T \boldsymbol{\beta}$ is an estimatable function, there exists the unique a linear unbiased estimator of it which has minimal variance, which is equal to $\hat{\psi} = \boldsymbol{c}^T \hat{\boldsymbol{\beta}}$*

Remark 5. The least square estimator is MLE of a data that is normall distributed, we can see that the normal distribution is:

$$L(\boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N \exp\left(-\frac{S(\boldsymbol{\beta})}{2\sigma^2} \right)$$

from which it can be seen that maximization of $L(\boldsymbol{\beta}, \sigma^2)$.

Remark 6. The estimation of σ^2. For the residuals, the fitted values are the estimator of the mean response for each observation: for the i-th observation of the fitted value $\hat{\mu}_i = \boldsymbol{x}_i \hat{\beta}$ with the following residual:

$$\hat{\boldsymbol{e}} = \boldsymbol{Y} - X\hat{\boldsymbol{\beta}} = \boldsymbol{Y} - X(X^T X)^{-1} X^T \boldsymbol{Y}$$

where $\hat{e} = (I_N - H)\boldsymbol{Y}$ where $H = X(X^T X)^{-1} X^T$.

Definition 2.2. (RSS) The sum of the squared where the residual sum of square (RSS) and is the minimal of $S(\boldsymbol{\beta})$ and so:

$$\mathrm{RSS} = \hat{\boldsymbol{e}}^T \boldsymbol{e} = \sum_{i=1}^{n} \hat{\boldsymbol{e}}^2 = \sum_{i=1}^{N} (Y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})^2$$

We can so that the expected value of RSS:

$$\mathbb{E}[\mathrm{RSS}] = \mathbb{E}[\hat{\boldsymbol{e}}^T \hat{\boldsymbol{e}}] = \mathbb{E}[Y^T (I_n - H) Y] = (N - p)\sigma^2$$

Remark 7. The unbiased estimator of $\hat{\sigma}^2 = \mathrm{RSS}/(N - p)$ under the assuption of independent $e_i \sim \mathcal{N}(0, \sigma^2)$, it can be shown that $\mathrm{RSS}/\sigma^2 \sim \mathcal{X}_{N-p}^2$ independent of $\hat{\boldsymbol{\beta}}$.

Definition 2.3. (Weight Least Square) Weighted Least Square is used when we know that the error isn't constantly the weighted least square estimation of $\boldsymbol{\beta}_j$ minimizes

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} w_i(Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$$

Remark 8. The greater the weight, the more reliable i.e having small variance. Ideally, we would like to put $w_i = 1/\operatorname{var}(Y_i)$. This is what we obtained from normal criterion with $Y_i\sqrt{w_i}$, which has constant variance. Furthermore, the matrix form is:

$$S(\boldsymbol{\beta}) = (\boldsymbol{Y} - X\boldsymbol{\beta})^T V^{-1}(\boldsymbol{Y} - X\boldsymbol{\beta})$$

where $\boldsymbol{V}$ is $N \times N$ diagonal matrix with $\operatorname{var}(Y_i)$ is the diagonal. The estimator of $\hat{\boldsymbol{\beta}}$ is:

$$\hat{\boldsymbol{\beta}} = (X^TV^{-1}X)^{-1}X^TV^{-1}\boldsymbol{Y}$$

Note that V can be genearlized to a correlated errors, while the solution is scale invariance for the value V.

Remark 9. However, we don't really know V in practice and there are 2 ways we can fix this:

- It is sometimes to assume that the variance of Y_i are propotional to $f(\boldsymbol{x}_i)$ as $\boldsymbol{w}_i = 1/f(\boldsymbol{x}_i)$ even if $\operatorname{var}(\boldsymbol{x}_i) = cf(\boldsymbol{x}_i)$ and still works.
- The weighted least square can be used iteratively as we can guess $\boldsymbol{V}$ and re-estimate it.

3 Confidence Interval Test

Remark 10. **(Test For Regression Parameter)** We can derive that $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 v_j)$ where v_j is the (j,j)-th diagonal $(X^TX)^{-1}$ as we have $\text{RSS}/\sigma^2 \sim \mathcal{X}^2_{N-p}$ as we have:

$$t = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 v_j}}$$

where we have $\hat{\sigma}^2 = \text{RSS}/(N-p)$ an exact $100(1-\alpha)$ percent for β_j has limit:

$$\hat{\beta}_j \pm t_{N-p,1/(2\alpha)} \operatorname{se}(\hat{\beta}_j)$$

where $\operatorname{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 v_j}$ and $t_{N-p,1/(2\alpha)}$ is upperbound $100 \cdot 1/2\alpha$ point of t_{N-p} distribution. It can be used to test hypothesis of the form:

- $H_0 : \beta_j = \beta^*$ for given j
- $H_0 : \beta_j = 0$ for a given j, where we obtain the p-value the usual way.

Remark 11. **(CI Test for Linear Combination of Parameters)** Inference about $\psi = \boldsymbol{c}^T\boldsymbol{\beta}$ an estimated expected response. Let $\hat{\psi} = \boldsymbol{c}^T\hat{\boldsymbol{\beta}}$, which we have:

$$\frac{\hat{\psi} - \psi}{\sqrt{\hat{\sigma}^2 v}} \sim t_{N-p}$$

where $c^T(X^TX)^{-1}c = v$, we can construct $100(1-\alpha)$-percent interval and we have the test start for $H_0 : \psi = \psi^*$.

Remark 12. **(Test About Multiple Parameter)** Consider testing null hypothesis H_0 that a subset of $p-q$ parameter out of p are 0, leaning q non-zero parameter. Let H_1 denotes an alternative hypothesis that all p are not 0.

- We let RSS_0 and RSS denote the residual sum of square under H_0 and H_1 respectively.
- It can be shown that the difference between $\text{RSS}_0 - \text{RSS}$ is independent of RSS and so:

  $$\frac{\text{RSS}_0 - \text{RSS}}{\sigma^2} \sim \mathcal{X}^2_{p-q}$$

  under H_0
- We obtain F-test, as we have:

  $$\frac{(\text{RSS}_0 - \text{RSS})/(p-q)}{\text{RSS}/(N-p)} \sim F_{p-q,N-p}$$

  under H_0, in case where $p - q = 1$ and $F = t^2$ where t is the t-statistics given above.

4 Multiple Linear Regression

Definition 4.1. (Multiple-Linear Regression) We consider the following model as we have:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + e_i \qquad \text{for} \qquad i = 1, \dots, N$$

where Y_i is the response $\boldsymbol{Y}$ corresponds to $x_{i1}, \dots, x_{im}$.

Remark 13. Given the normally distributed errors $e_1, \dots, e_N$ with constant variance σ^2. We can see that Y_i is independent $\mathcal{N}(\mu_i, \sigma^2)$ where:

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}$$

The matrix form of the model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{e}$ where i-th row of X is $(1, x_{i1}, \dots, x_{im})$ and $\boldsymbol{\beta} = (\beta_0, \dots, \beta_m)^T$. We will assume that $X^T X$ is invertible.

Remark 14. We have the following results on the multiple linear regression models:

- Least Square Estimate $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}$
- Sampling Distribution of $\hat{\boldsymbol{\beta}}$ where we have $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$
- Residual Sum of Square is

  $$\begin{aligned} \text{RSS} &= (\boldsymbol{Y} - X\hat{\boldsymbol{\beta}})^T (\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}) = \boldsymbol{Y}^T\boldsymbol{Y} - 2\hat{\boldsymbol{\beta}}^T X^T \boldsymbol{Y} + \hat{\boldsymbol{\beta}}^T X^T X \hat{\boldsymbol{\beta}} \\ &= \boldsymbol{Y}^T\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^T X^T \boldsymbol{Y} \end{aligned}$$

- Unbiased Estimator of σ^2: $\hat{\sigma}^2 = (\text{RSS}) \,/\, (N - m - 1)$
- T-test for $H_0 : \beta_j = 0$ as we have $t = \hat{\beta}_j / \text{se}(\hat{\beta}_j) \sim t_{N-m-1}$ under H_0.
- F-test for H_0 : ν of regression parameter $\beta_1, \dots, \beta_m$ are 0 i.e testing for omission of ν explainatory variable where $\nu \le m$ as we have:

  $$F = \frac{(\text{RSS}_0 - \text{RSS})/\nu}{\text{RSS}/(N-m-1)} \sim F_{\nu, N-m-1} \qquad \text{under } H_0$$

- Special Case: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0$ under H_0. The least square estimator of β_0 is the sample mean:

  $$\text{RSS}_0 = \sum_i (Y_i - \bar{Y})^2 = \text{ correlated total sum of square}$$

The results are represented in a table of analysis of variance table:

Source of Variation	**Sum of Square**	**Degree of Freedom**	**Mean-Square**	**F**
Regression	SS(reg)	m	SS(reg)$/m$	$\dfrac{\text{SS(reg)}/m}{\text{RSS}\,/(N-m-1)}$
Residual	RSS	$N-m-1$	$\dfrac{\text{RSS}}{N-m-1}$	–
Total	CTSS	$N-1$		

where we have SS(reg) $= \sum_{i=1}^{N}(\hat{\mu}_i - \bar{Y})^2 = \text{CTSS} - \text{RSS}$ where $\hat{\mu}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ are fitted value.

5 Interpretation of Regression Parameter

Remark 15. **(Partial Regression Coefficient)** The coefficient β_j of an explainatory variable x_j in a multiple linear regression that includes other explainatory variable called partial regression coefficient. It makes sure that the rate of change of the mean response with x_j while holding constant the values of other explainatory variable in model.

Remark 16. **(Total Regression Coefficient)** The coefficient x_j in the simple linear regression of the respons variable x_j on its own is called total regression coefficient. It measures the rate of the mean response with x_j ignorning the value of the other.

Remark 17. **(Checking Model Adequacy)** Assessment of the model assumptions, we have:

- Linearity of the relationship between predictor and Y.
- Normality of e_i: Though the normal distribution theoretically can guarantee generting the arbitary real number, very extreme values occurs under normal distribution with small probability.
  - The outier have strong influence of least square regression as the detection of it is the most important task.
  - Another possible deviation could be skewness of the distributed shape.
  - Some deviation from normality are less dangerous: Limited/Restricted Value Range.
  - Normality is an idealization that never holds precisely in practice, we have to find whether the deviation that gives us misleading conclusion about the data.
- Homogeneocity of the variance of e_i: The variance don't depends on any of the predictor variable. (oppose to heteroscedastic)
- Independent of e_i of each other.

Remark 18. **(Matrix Plot)** All scratter plot of any pair of predictor variables and response arranged in matrix form can be used without having fitted the linear regression or access linearlity, outlier and heteroscedastic.

The danger of over-interpreting as it doesn't give a full impression. Non-linearity shape of the plot of single predictor vs response is caued by the value of other predictor than the real violation of linearity. It may reveal co-linearity and leverate points, which are not violation but problematic.

Remark 19. **(Residual and Standardized Residual)** Residuals are the deviation of the observation of their ideal. It can be interpreted as the estimation of error e_i (denoted by $\hat{e}_i$). Standardized residual are residual divded by their estimated standard-deviation so it have $\sigma = 1$. This helps us access their size.

- Too mant standardized with manitude of greater than 2 suggests that the eror distribution have heavier tail but around 5-percent of the standardized residual should be expected to be > 2

- The covariance of residual is $V(\hat{\boldsymbol{e}}) = \sigma^2[I - H]$ where $H = X(X^TX)^{-1}X^T$ called hat matrix, and so we have:
$$\text{var}(\hat{e}_i) = \sigma^2(1 - h_{ii}) \qquad \text{cov}(\hat{e}_i, \hat{e}_j) = -\sigma^2 h_{ij} \text{ for } i \neq j$$
while the error are assumed to be uncorrelated, we just show that the residual are correlated. For i-th observation, the standard deviation residual is given by:
$$r_i = \frac{\hat{e}_i}{\sigma\sqrt{1 - h_{ii}}}$$

Remark 20. **(Residual Plots)** Makes sense to use standardized residual for the plot but a raw plot but other kinds of plots may works well as we have the following kinds of plots:

- *Predictor vs Residual*: Error is assumed to be independent from the predictor variable.
  - The plot should be randomly scrattered.
  - Plot reveals non-linearity, heteroscedastic, auto-correlation of residual with neighbouring values of outier.
- *Fitted Values vs Residuals*: If the model is true, the correlation between the residual are fitted value is zero. If any problem, it will have similar problem with predictor and residual plot.
- *Observation vs Residual*: If the observation order is informative, the error should be iid and plot should randomly scrattered. It may reveal the autocorrelation and heteroscedasity as well.
- *Normal Probability Plot of Residual*: The normal probability plot the sorted residual (standardized) $r_{(i)}$ (i smallest residual) against theoretically quantity of normal distribution $\Phi^{-1}\left((i - 0.5)/n\right)$
  - Ideal local sorted realization of standard normal distribution.
  - They should looks like a straight line.
  - It also indicates a deviation from normality, including outlier too.

Remark 21. **(Remedies for Violate Model Assumption)**

- *Non-Linearity*: It helps to transform one or more predictors and/or response. A linear nodel may holds some non-linear function of observed variable.
- *Non-Normality* of error distribution. Robust linear regression may help with outlier. Skewness can be helped using a transformation. Some transformation to response and predictor.
- *Heteroscedasity*: Weighted square, transformation, and robust regression.
- *Dependence of Errors*: Not affect regression parameter estimator but it affects standard deviation and confidence interval. If assuption can be made, time series may apply.

Definition 5.1. (Coefficient of Determination) R^2 is defined by:
$$R^2 = \frac{\sum(\hat{\mu}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\text{SS(reg)}}{\text{CTSS}} = 1 - \frac{\text{RSS}}{\text{CTSS}}$$

Propotion of the total variations explained by regression model. It is also a square of the correlation between the observed and filled values. This is known as multiple correlation coefficient.

- Measures how the models accounts for data. This isn't directly related to be model assuption.
- Small value of R^2 means that the assuption are violated or some crutial information is missing in the data or model is fine but the error variance is large.

The model violation still possible if R^2 is relatively high:

- True relationship is strong and monotone but slightly non-linear
- The case of heteroscedasity with small error variances a linear may yield a good fit with high R^2.

Definition 5.2. (Regression Outlier) Observation with unusual y-values compared to other observation with small x-values. If there are small number of regression outlier, these may show up in residual plot with large residual.

Definition 5.3. (Leverage Points) Observation with usual x-values compared to the bulk of data. Linear regression don't assume normality for the predictors therefore leverage points don't violate the model. However, this cause instability of the regression as if there is a small change in the data, it may lead to large change in the least square regression estimator. It can be distinguished between 2 points:

- Good Leverage Points: If the y-values are inline with the other y-value. The fitted $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ is similar to the observation value y_i, as the omitting the observation won't change the fitting.
- Bad Leverage Points: y-value is unusual, then $\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ is difference from y_i depending on the extended of the effected.

Leverage describes the potential for affecting the model fit. We introduce the Mahalanobis distance MD_i, as we have:

$$\text{MD}_i = \sqrt{(x_i - \bar{x})\hat{\Sigma}_x^{-1}(x_i - \bar{x})}$$

where $\hat{\Sigma}_X$ is empirical covariance matrix of x-observation. There is ono-to-one relation with diagonal element of Hat, as we have:

$$\text{MD}_i^2 = (N-1)\left(h_{ii} - \frac{1}{N}\right)$$

We see that $\text{MD}_i^2 \sim \mathcal{X}_{p-1}^2$ which can be used to access whether a distance is unusually large. Note that the leverage points can be prevented in design experiment.

Remark 22. Checking for the outlier can be given by:

- It is good to check before fitting the data. In higher dimension, it is hard to spot such unusual observation: residual plot can help but as it is derived from fitted model, which might be affected by outlier.
- Another possibility is Cook's statistics, fit the model repeatedly by omitting one observation and see how the fit value change. The change is given as $X(\hat{\beta} - \hat{\beta}_{(i)})$ where $\hat{\beta}_{(i)}$ is the least square without the i-th observation. It is given by:

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

  The largest value of D_i indicates the i-th observation is influential. It can be shown that:

  $$D_i = \frac{1}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right) r_i^2$$

  There is no need to fit at all.

Both Mahalanobis distance and Cook's statistics can't reliabily finds all the points. As there can be masking effect, when there are > 1 outlier points prevents the effect of each other.

Remark 23. **(Collinearity)** Strict collinearity means that these is linear dependence among the predictor variables:

- X^TX isn't invertible and least square estimator doesn't exists.

- This may happen if the number of predictors is large compared to the observation.

The approximation collinearity is when predictors are linear dependents. This Happens when some of predictions are strongly correlated. This can be detected from matrix plot.

- Although the least square can be computed, X^TX is close to singularity, this means that some of regression parameter estimator may be unstable and should be interpreted with care.
- $\sigma^2(X^TX)^{-1}$ convariance matrix of β might have large variance entry.

6 Robust Regression

Remark 24. **(Problems with Least Square Error)** The least square error can be sensitive to recall that the estimator is found by minimizing

$$S(\boldsymbol{\beta}) = \sum_{i=1}^N e_i(\boldsymbol{\beta})^2 = \sum_{i=1}^N (Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2$$

and so any residual that already has high value will have high contribution to the sum.

Remark 25. They should be a good under the normal model (efficiency) as we have: Under linear regression with other distribution of error term (espescially with the heavy tail). They should not also be sensitivity to outlier.

Remark 26. **(Constructing a Robust Classifier)** The square error can be seen as minimizes the error variance estimator $\hat{\sigma^2}$. This means that constructing a linear regression is to minimize the scale estimator (function that is propotional to the variation of the residual around hyperplane). The way to estimate the size of the residual leads to difference regression estimator. There are 2 ways to make LS robust:

- Instead of using square the residual, use another function to reflect distance on Y_i, we have:

  $$\sum_{i=1}^N |e_i(\beta)| = \sum_{i=1}^N \left|Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right|$$

  The minimize of β is known as L_1-estimator. The generalization is known as M-estimation. It can be shown to be MLE of the double exponential. For $m = 0$ can be L_1-estimator is β_0 is median.
- We could minimize the median of the residual as we have:

  $$\text{MED}\left\{e_i(\boldsymbol{\beta})^2 : i = 1,\dots,N\right\} = \text{MED}\left\{(Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_i)^2 : i = 1,\dots,N\right\}$$

  This is known as least median of the squares.

Remark 27. **(Effect of Leverage Points)** The deficiency of least square estimate and it applied to regression outliers, but least square is also affected by leverage points. Consider a simple linear regression: least square estimator for the slope $\boldsymbol{\beta}_1$ is

$$\beta_1 = \frac{\sum_i (x_i - \bar{x})y_i}{\sum_k (x_k - \bar{x})^2} = \sum_i v_i y_i$$

with $v_i = (x_i - \bar{x})/\left\{\sum_k (x_k - \bar{x})^2\right\}$. Hence β_i is the weighted sum of the y_i where the large weight is given by observation with large $(x_i - \bar{x})$, which are the leverage points.

Remark 28. It can be shown that R^2 is affeced by leverage poitns in the sense that good leverage points imposing the fit (increase $\mathbb{R}^2$) and vice versa. We will see not far later that M-estimator aren't much better than the least square in dealing with leverage points but least median square-estimator is.

Remark 29. **(Comments on the Least Square)**

- We have discussed earlier. Our old methods works best with one outlier. We have to use the residual from the fitted model to find outlier.
- In term of efficiency, robust methods are designed to only slightly affected by small deviations from model assuption and not be catastrophically affected by large deviations.
- Some robust models are not as efficient as least square. Their asymptotics variance is larger than the one that the least square estimator holds.

Definition 6.1. (M-Estimators) Instead of minimizing on square residual, another function is the be used with the following criterior for $\boldsymbol{u} \in \mathbb{R}$ as we have:

- Positivity $\rho(\boldsymbol{u}) \ge 0$
- Zero for zero residual $\rho(0) = 0$
- Symmetric $\rho(\boldsymbol{u}) = \rho(-\boldsymbol{u})$
- Increase for increasing the residual $\rho(\boldsymbol{u}) \ge \rho(\boldsymbol{u}')$ if $|u| \ge |u'|$

Assume we know σ, the M-estimator $\hat{\boldsymbol{\beta}}_M$ is defined as $\hat{\boldsymbol{\beta}}$ that minimizes:

$$\sum_{i=1}^{N} \left(\frac{e_i(\boldsymbol{\beta})}{\sigma}\right) = \sum_{i=1}^{N} \rho\left(\frac{Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\right)$$

in $\boldsymbol{\beta}$. Note that the least square and L_1-estimator are M-estimators.

Remark 30. The differentiate with respected to $\boldsymbol{\beta}$ and setting the zero yields:

$$\sum_{i=1}^{N} \boldsymbol{x}_i^T \psi\left(\frac{Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\sigma}\right) = \mathbf{0}$$

where we have $\psi(\cdot) = \rho'(\cdot)$. The equation can't be solved analytically. By setting $u_i = (Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})/\sigma$ as we can rewrite:

$$\sum_{i=1}^{N} \boldsymbol{x}_i^T \boldsymbol{w}_i \omega_i = 0$$

where $w_i = \psi(\boldsymbol{u}_i)/u_i$. This is like weighted w_i, which we can approximate the $\hat{\beta}_M$ as solution of:

$$\hat{\boldsymbol{\beta}}_M = (X^TWX)^{-1}X^TW\boldsymbol{Y}$$

where $W = \text{diag}(w_1, \dots, w_N)$ as w_i depends on $\hat{e}_i$ and hence $\hat{\boldsymbol{\beta}}_M$ and we use the iteration to be:

$$\hat{\boldsymbol{\beta}}_M^{(k+1)} = (X^TW^{(k)}X)^{-1}X^TW^{(k)}\boldsymbol{Y}$$

until convergence. The value $\boldsymbol{w}_1, \cdots, \boldsymbol{w}_N$ can be interpreted as robustness weight and can be used as outlier identification. The observation is an outlier by M-estimator if $\boldsymbol{w}_i$ is small.

- The residual $\hat{e}_i(\boldsymbol{\beta})$ only enter through $\psi(\boldsymbol{u}_i)$. If ψ is bounded, the influence of large residual on regression is bounded as well.
- The influence on leverage point isn't bounded because of the factor $\boldsymbol{x}_i$, unless $\psi(u) = 0$ or $\psi(u)$ very small for those observation with large $\boldsymbol{x}_i$

Remark 31. Under fiarly general condition, M-estimator can be shown to be consistent. A necessary condition is that $\mathbb{E}[\psi(Z)] = 0$ where $Z \sim F$ as F is the error distribution holds for all symmetric error if ψ is bounded. Furthermore,

$$\sqrt{n}(\hat{\beta}_M - \beta) \overset{n\to\infty}{\sim} \mathcal{N}(0, V(\psi, F)L^{-1})$$

where $V(\psi, F)$ is matrix that depends on the influenced function ψ, and true error distribution F with $L = (1/N)\lim X^TX$, which can be used to compute the test and confidence interval.

Remark 32. Under normality, covariance matrix of the lease square error is $C = \sigma^2(X^TX)^{-1}$. For M estimators with many other ρ-functions, the covariance matrix is shown to be:

$$V(\psi, F)L^{-1} = bC$$

for some constant as $1/b$ is called efficiency of the estimator compared to least square estimator. Assuming that the asymptotics holds approximately for finite sample, b can be interpreted as the factor with which the number of observation has to be multiplied to arrive at same precision (with efficiency 0.5 we need 2 as much observation)

Definition 6.2. (Bi-Square Objective Function) L_1 isn't robust against leverage point as we have the alternative to be:

$$\rho_B(u) = \begin{cases} \dfrac{1}{6}\left\{1 - \left[1 - \left(\dfrac{u}{c}\right)^2\right]^3\right\} & \text{if } |u| \le c \\ \dfrac{1}{6} & \text{if } |u| > c \end{cases}$$

wherer c is tuning constant, which we also have:

$$\psi_B(u) = \begin{cases} \dfrac{u}{c^2}\left[1 - \left(\dfrac{u}{c}\right)^2\right]^2 & \text{if } |u| \le c \\ 0 & \text{if } |u| > 0 \end{cases}$$

c can be used to tune the robustness and efficiency. For small $|u|$, $\psi_B(u)$ is like a line, which is propotional to ψ-function for the least square estimator.

- If c is large, most u are small and most residual are treated in smaller way as LS-estimator. For $c \to \infty$, the efficiency of Bi-square M-estimator converges to 1.
- If c shouldn't too large to correct for extreme outlier. The suggested value $\sigma^2 = 1$ and $c = 4.658$, the usually suggested value.

7 Variable Selection

Remark 33. **(Variable Selection)** Consider the regression, as we have:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + e_i$$

for we have $i = 1, \dots, N$. Variable selection is where we select the variables that are relevant: This is the same as finding coefficient β_i that are zero. Very small with almost zero contribution

Remark 34. **Pros and Cons** There are various reasons to use variable selection as we have:

- If number of variable p is large compared to number of observation N as X^TX are closed to collinearity and estimation can be unstable.
- Simple communication is better and same observation are expensive to get the data.

However, there are various argument against using variable selection:

- As long as enough observation are avaliable, it is usually much worst to leave out, the variable that are important.
- As β_i will be zero, it is better to leave out as it is unless there is a reason to do it.

- Some of variable are highly correlated, usually not all of these variables are needed, but the decision about which model should be kept can be arbitary.
- It is often unstable and have to be taken with care concerning explaination and causual inference.
- It can't be taken to granted that a variable left out is unimportant as it may be represented by another variable.

Remark 35. **(Reviewing Old Methods)** Let's consider some of the results presented in eariler parts:

- The F-test in anova table is for testing none of the explainatory variables affect the response.
- Another F-test is used to testing whether j-th explanatory variable x_j doesn't affect the response given the other $m-1$ variables in the model, so that the regression coefficient is 0.
- Computer output usually gives t-statistics and associated p-values against each of the regression coefficient. We shouldn't deduce that we can remove several variables. If any of them gives non-significant, we can only select one to drop with highest p-value (most non-significant)
- Reason why not dropping several: If there are 2 are highly correlated one of them is needed into model but every single pair can be dropped as non-significant.
- If we drop one of the coefficient, the others coefficient will change. Similarly as we added variable, the estimator remain unchanged is where there is another orthogonal columns in the X.

Remark 36. **(Best Subset Selection)** With m explainatory, there are 2^m possible regression model:

- Starting with the best model for each explainatory variable $k = 1, \dots, m$. For a fixed k, the best model it can be defined as model with smallest RSS or model that minimize p-value of F-test by the null hypothesis that $\beta_1 = \cdots = \beta_m = 0$ against at least one of the parameter of the choosen model to be non-zero.
- Adding more variable always decrease RSS implied that increases R^2.
- The best model with $l > k$ variable doesn't necessary contain all k variables of the best model with k-variables. t and F-test can't be used to compare the best subset of difference size.

Remark 37. **(Akaike Information Criterior (AIC))** For quite general model by maximum likelihood as we have:

$$\text{AIC} = -2\hat{l}(\text{model}) + 2p$$

as we have $\hat{l}$ is the maximum of likelihood, while p is the number of parameter, where we have:

$$\hat{l}(\text{model}) = -\frac{N}{2}\log(\sigma^2) - \frac{\text{RSS}}{2\sigma^2} + \text{ const}$$

Since σ is usually unknown, we use $\hat{\sigma}^2_{\text{ML}} = \text{RSS}/N$ instead of:

$$\text{AIC} = N\log\left(\frac{\text{RSS}}{N}\right) + 2p$$

Remark 38. **(Other function of RSS)** Penalizing large models have been suggested such that mallow C_p and so-called "adjusted R^2", which delivers number between 0 and 1 like R^2 but is maximized when $\hat{\sigma}^2$ is maximized. However, this is soft criterion.

Remark 39. **(Leave-One-Out Cross Validation)** General method is fitted on the remaining $N-1$ points. There are advantages and disadvantage of LOO-CV and we have:

- Advantages: It doesn't base on any model assuption. Even if the fitted model is wrong.
- Disadvantage: Model has to be fitted N-times which gives us computationally demanding if N is large.

This is the same as any Cross-Validation Scheme in ML.

Remark 40. **(Stepwise Methods - Backward Elimiation)** The best subset selection has major disadvantage but we had to fit all 2^m possible model. The stepwise selection is a general principle that can be applied to wide range of selection problem, where are 2 basic approaches starting with Backward Elimiation: Start by fitting full model with all variables. Call this backward m-model as we set $k = m$

- Fit all k models with $k-1$ variables.
- Choose best of these model with minimal RSS as backward $k-1$ model
- Set $k = k-1$
- Repeat

The total number of models to be fitted is:

$$m + (m-1) + \cdots + 1 = \frac{m(m+1)}{2}$$

Because backward elimiation procedure a nested sequence of models as we have the smaller models are always sub-models of the larger one, it is possible to use t and F test to compared models.

p-value for the test comparing the backward k-models (H_1) with backward $k-1$-models (H_0) is computed and the algorithm is stopped with the backward $k-1$-models.

Remark 41. **(Stepwise Methods - Forward Selection)** We can have the forward model with smaller procedure, which we have:

- Start with $k = 0$, only fitting the mean to the data.
- Fit all $m-k$ models with $k+1$ variables.
- Choose one with best of these model, minimal RSS as the forward $k+1$-model.
- Set $k = k+1$ and go back unless $k > m$.

Remark 42. **(General Notes on Stepwise Model)**

- Backward/Forward model aren't guarantee to arrive at the same sequence of model nor the best model.
- In experiences, Backward model generally performs better.
- Steppwise methof can be unstable as a small change in the model causes the best model to change.
- For small observation, we usually use the forward model instead.

8 LASSO

Remark 43. It deals with when there is too large data, and it is more stable than the methods mensioned. However, it introduces some biases (which are moderate) and can be out-performed in clear cut situation (if all regression are close to 0 or all large).

Remark 44. For computing LASSO, the original explainatory variables $z_1, \dots, z_m$ are transformed to new explainatory variable:

$$x_{ij} = \frac{z_{ij} - \bar{z}_j}{s_j} \qquad i = 1, \dots, N \quad j = 1, \dots, m$$

where we have:

$$\bar{z}_j = \frac{1}{N}\sum_{i=1}^{N} z_{ij} \qquad S_j = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(z_{ij} - \bar{z}_j)^2}$$

This is a general practice too as now the LASSO estimator $\hat{\boldsymbol{\beta}}_L$ of $\boldsymbol{\beta}$ is defined by minimizing the sum of square error:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N}(Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_m x_{im})^2 \quad \text{such that} \quad \sum_{i=1}^{m} |\beta_i| \leq t$$

where t is pre-choosen. It is recommended to choose t by LOO-CV with $t = ct_0$ where $c = 0.1, 0.2, \cdots, 1$. It can be shown that in the model where the variables are much less important thant some other. The estimated LASSO regression parameter $\hat{\beta}_{L_i}$ are reduced to zero.

Remark 45. β_0 can be estimated by least-square independently of slope parameter due to the constraint In most cases, there is either no reason why β_0 should be estimated by 0 or β_0 is known and doesn't have to be estimated.

Remark 46. LASSO is known as shrinkage method as it forces the parameter to be smaller than the unconstrainted least square. Furthermore:

- It can be shown that the stepwise and best subset tends to choose variable with models of which β_i is estimate to have a large value.
- There is no theoretical work on how to choose it the value t, however.

9 ANOVA Model

Remark 47. Theory of linear model applied to situation in which response Y_i is modelled as dependent on catagorical or a group member. These model are often called ANOVA model:

Remark 48. The ANOVA model is often refer to breakdown of the corrected total sum of square (CTSS) into the sum of thE RSS of the full number + sum of square contribution explained by some of the regression parameter.

- Sum of squares can be seen as quantifying variation and the usual variation and the usual variance estimators under normalizing are actually sum of square divided some cosntant.
- Such decomposition play a stronger role for ANOVA models with catagorical predictors.

Definition 9.1. (One Way Layout) We have the following model:

$$Y_{ij} = \mu_i + e_{ij}$$

for $i = 1, \ldots, I$ as we have $j = 1, \cdots, n_i$. It is the easiest ANOVA model.

Remark 49. It has been shown that this can be written down by the use of indicator variable as the explainatory variables. The indicator of the first group is 1 for all observation in the first group and 0 for all other observation. We have the following notation $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}$ as we have:

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_{11} \\ \vdots \\ y_{In_I} \end{pmatrix} \qquad X = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_I \end{pmatrix} \qquad \boldsymbol{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_N \end{pmatrix}$$

The column of X denoting the I indicator variable and the roes of the n-observation.

- This notaiton makes it possible to apply the theory above.
- It can be shown that the least square estimator above can be used to calculate the group means.
- The F-test in section above is usually applied to test the $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$ using $\text{RSS}_0 = \text{CTSS}$ against a model where at least 2 of the group has differ mean.

Definition 9.2. (2-Way Layout) The observation are classified according to 2 factors. In this case, we have the notation:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

for $i = 1, \cdots, I$ and $j = 1, \dots, J$ and $k = 1, \cdots, n_{ij}$ where we have:

- μ is interpreted as the overall expected value.
- α_i is the effect of level i of the first factor.
- β_j is the effect of level j of the second factor.
- γ_{ij} is the interaction effect.

Remark 50. The simple model can be applied again.

- The relationship suggest an X-matrix with the first column, which corresponds to the $\boldsymbol{\mu}$ contains only ones.
- I columns corresponding to indicator variable for the level of the first factor.
- J columns corresponding to level of second factor.
- IJ columns for the interaction.

And, so we have the following $1 + I + J + IJ$ columns. This could be multiplied by the following vectors:

$$\boldsymbol{\beta} = (\boldsymbol{\mu}, \alpha_1, \cdots, \alpha_I, \beta_1, \cdots, \beta_J, \gamma_{11}, \cdots, \gamma_{IJ})^T$$

The simplest case $I = J = n_{ij} = 2$ for each i and j this giving us the following matrix:

$$\boldsymbol{y} = \begin{pmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \end{pmatrix} \qquad X = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

However, one can show that $X^T X$ isn't invertible as there are too many parameter. The problem where isn't too small of observation but the fact that there are only IJ means. (this can be estimated and therefore more than IJ parameter are not supported)

Remark 51. This problem of non-singularity can be solved by using the contraints is given by:

- $\sum_{i=1}^{I} \alpha_i = 0$
- $\sum_{j=1}^{J} \beta_j = 0$
- $\sum_{i=1}^{I} \gamma_{ij} = 0$ for $j = 1, \cdots, J$
- $\sum_{j=1}^{J} \gamma_{ij} = 0$ for $i = 1, \cdots, I$

as we have $I + J$ contraints of which one is redundance because if all but one of these sums are 0, it can be shown that the last one has to be zero as well.

Remark 52. **(Matrix Algebra)** We will have consider the matrix algebra related to the problem:

- First constraints means that $\alpha_1 = -\sum_{i=2}^{I} \alpha_i$
- The column belong to level 1 is omitted from X and α_1 is omitted from $\boldsymbol{\beta}$ and for the observation of level 1 of the first factor because it is replaced by $-\sum_{i=2}^{I} \alpha_i$
- There are matrix entries -1 in the column corresponding to $\alpha_2, \cdots, \alpha_I$
- With other constraints as X matrix are no longer simple indicator vectors, while it could be achieved by contraining some parameter to be zero instead.

The matrix with

$$1 + I + J + IJ - 2 - (I + J - 1) = IJ$$

columns result so that $X^T X$ to invertible and the least square estimator, which can be computed:

$$X = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_2 \\ \beta_2 \\ \gamma_{22} \end{pmatrix}$$

All other parameter can be obtained from the constraint.

Remark 53. **(ANOVA Table)** All the theories above can be obtained. F-test are particularly used in many application to find out whether:

- All interaction could be ignorned (with RSS_0 computed from a model when all $\gamma_{ij} = 0$)
- All $\alpha_i = 0$ is compatible with the data
- All $\beta_j = 0$ is the compatible with the data

The ANOVA table decomposes the CTSS (which is computed from a model in which any the overall mean μ is filled by $\bar{Y}$)

- into RSS, the full model puts the sum of square explained by the first factor, second factor, interactions.
- Unless all n_{ij} are equal (balanced design), this depends on the order of factors.

Remark 54. For variable selection, most methology can be applied, though in most application there is a particular order in which terms are removed in the stepwise backward fashion: It is reasonable to have the original factor in the model if there are interaction in the model involve these factors. For example, in 2-ways layout with backward selection usually:

- It is checked whether removal if all interaction terms improves the mdodel. If not, the model isn't reduced.
- After removal of all interactions, it is checked removal of which of the 2 factors (all parameter belonging to that factor) is better, and whether this improves the model.
- If so, after removal of this factor, it is checked whether removal of the other factor impress the model even further.

This check can be done by using F-tests, AIC or LOO-CV.

10 Generalized Linear Model

Definition 10.1. (GLM) It extends the idea underlying the linear model to situation when the reponse is binomial, poisson, gamma and other distribution that belong to exponential family of distribution. It consists of 3 pairs of components:

- Random Component: Independent observation $Y_1, \dots, Y_N$
- Systematic Component: Linear predictor by η_i for i-th observation for $i = 1, \dots, N$
- Link Between Random Component and Systematic Component through the use of *link function* g:

$$g(\mu_i) = \eta_i \qquad \mu_i = \mathbb{E}[Y_i]$$

Using the eariler notation $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$. The function g is called the link function as it describes how the expected response is linked to explainatory variable off factors. The link function itself is assumed to be a monotonics and differentiable function.

Remark 55. There are some special cases/examples of the link function related to the distribution:

- Linear Model: Linear Model discussed in previous section is the special case $g(\mu) = \mu$ calling identity link function.
- Binary data: We consider linear logistic model. Suppose, the i-th observation consists of a Bernoulli with outcome of $Y_i = 1$ (success) or $Y_i = 0$ (failure). If π_i is the associated probability of success. then the under logistic regression model given by:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

  In this case: $\mu_i = \mathbb{E}[Y_i] = \pi_i$ and so the left-handside of the equation has the link function to be: $\log(1/(1-\pi))$ is called logit of π and the link function in this case is called logit link. For $0 < \pi < 1$, then $-\infty < \log(1/(1-\pi)) < \infty$ with the consquence that there are no constant on the unknown parameter $\boldsymbol{\beta}$ (which simplify the estimation)

- Binomial Data: Suppose that the i-th observation corresponds to specified number of n_i (≥ 1) of independent Bernoulli trails with the same x_i. The distribution of the number of success Y_i is $\text{Bin}(n_i, \pi_i)$ where $\mu = n\pi_i$, then the linear logistic model given by equation is still appropriate. The link function:

$$g(\mu) = \log\left(\frac{\mu}{n-\mu}\right)$$

- Poisson data: Log-Linear Model: Now suppose the i-th observation consists of a Poisson counts Y_i. If $\mathbb{E}[Y_i] = \mu_i$ a log-linear regression model has:

$$\log(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

  where $g(\mu) = \log(\mu)$ is called log-link.

Remark 56. **(Exponential Families of Distribution)** We will assume that the response variable is a random Y whose pdf or pmf depends on the parameter θ and ϕ has the form:

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

where a, b, c are known function, as we assume $a(\phi) = \phi/w$ where w is known weight and ϕ is a dispersion parameter or scale parameter, which for some distribution is known and some other is unknown, and $a(\phi) > 0$

Remark 57. Let's consider the example: the common distribution, we have for $w = 1$ and $a(\phi) = \phi$, and so the following distributions are:

Distribution	θ	ϕ	$b(\theta)$	$c(y, \phi)$
Poisson(μ)	$\log(\mu)$	1	$\exp(\theta)$	$-\log y!$
Bin(n, π)	$\log\left(\frac{\pi}{1-\pi}\right)$	1	$n\log(1 + \exp(\theta))$	$\log\begin{pmatrix} n \\ y \end{pmatrix}$
$\mathcal{N}(\mu, \sigma^2)$	μ	σ^2	$\frac{1}{2}\theta^2$	$-\frac{1}{2}\left[\frac{y^2}{\phi} + \log(2\pi\phi)\right]$

Proposition 10.1. ***(Mean of Exponential Family)*** *The mean of the expectation is* $\mathbb{E}[Y] = b'(\theta)$

Proof. In the following $f(y; \theta, \phi)$ is abbreviated to $f(y)$. As we have $1 = \int_{-\infty}^{\infty} f(y) \text{ d}y$ and we have:

$$0 = \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f(y) \text{ d}y = \int_{-\infty}^{\infty} \frac{\partial f(y)}{\partial\theta} \text{ d}y = \int_{-\infty}^{\infty} \frac{y - b'(\theta)}{a(\phi)} f(y) \text{ d}y$$

Assuming $a(\phi) \neq 0$, then we have:

$$0 = \int_{-\infty}^{\infty} yf(y) \text{ d} - b'(\theta) = \mathbb{E} - b'(\theta)$$

Gives the result. □

Proposition 10.2. ***(Variance)*** *Given the variance* $\text{var}(Y) = b''(\theta)a(\phi)$

Proof. We consider the derivative of the above again and we have:

$$0 = \int_{-\infty}^{\infty} \left\{\frac{[y - b'(\theta)]^2}{a(\phi)} f(y) - b''(\theta)f(y)\right\} \text{ d}y = \frac{\text{var}(Y)}{a(\phi)} - b''(\theta)$$

□

Definition 10.2. (Variance Function) From the previous result, the variance of Y can be written as:

$$V(\mu)a(\phi) \qquad \text{and} \qquad V(\mu)\frac{\phi}{w}$$

where $V(\mu)$ is called variance function. a can be any function of ϕ, and there would not be any difficulty in dealing with any form of a, when ϕ is known. On the other hand, when ϕ is unknown matter are awkward, unless we write $a(\phi) = \phi/w$ where $w = 1$. As we have the following variance:

Distribution	$V(\mu)$
Poisson(μ)	μ
Bin(n, π)	$\mu(n - \mu)/n$
$\mathcal{N}(n, \sigma^2)$	1

11 Some GLM Theory

Remark 58. Assume that the response $Y_1, \dots, Y_N$ are independent from distribution with pdf given by:

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}$$

We are assuming a common parameter ϕ for all observation.

Proposition 11.1. *If l denotes the log-likelihood function given the data $Y_1, \cdots, Y_N$ then the likelihood equation are:*

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^N \frac{y_i - \mu_i}{V(\mu_i)} \frac{d\mu_i}{d\eta_i} x_{ij} = 0$$

for $j \in [p]$. Note that absent of ϕ in the likelihood equation.

Proof. The unknown parameter $\boldsymbol{\beta}$ and ϕ. Let l denotes the resulting likelihood function given the data $Y_1, \cdots, Y_N$ as we have:

$$l = \sum_{i=1}^N l_i \qquad \text{where} \qquad l_i = \log f(y_i; \theta_i, \phi) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)$$

The following shows the step in obey the the likelihood equation:

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^N \frac{\partial l_i}{\partial \beta_j} = 0 \qquad \text{for} \qquad j = 1, \cdots, p$$

Firstly, for $i = 1, \cdots, N$ and we have:

$$\begin{aligned} \frac{\partial l_i}{\partial \theta_i} &= \frac{y_i - b'(\theta_i)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)} \\ \frac{\partial l_i}{\partial \mu_i} &= \frac{\partial l}{\partial \theta_i} \frac{d\theta_i}{d\mu_i} = \frac{y_i - \mu_i}{a_i(\phi)} \frac{1}{b''(\theta_i)} = \frac{y_i - \mu_i}{\text{var}(y_i)} \\ \frac{\partial l_i}{\partial \eta_i} &= \frac{\partial l}{\partial \mu_i} \frac{d\mu_i}{d\eta_i} = \frac{y_i - \mu_i}{\text{var}(y_i)} \frac{d\mu_i}{d\eta_i} \end{aligned}$$

Putting back to the likelihood function, and we have:

$$\sum_{i=1}^N \frac{\partial l}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^N \frac{y_i - \mu_i}{V(\mu_i)} \frac{d\mu_i}{d\eta_i} x_{ij}$$

□

Remark 59. We can't solve the β algebratically. If the weight $V(\mu_i)$ where known and independent $\boldsymbol{\beta}$, then we can solve the non-linear equation to find $\boldsymbol{\beta}$:

- Starting with $\hat{\boldsymbol{\beta}}_0$ at the MLE $\hat{\boldsymbol{\beta}}$ and then we repleat the process until the convergence. The iteration can be shown as:

  $$(X^TWX)^{(s-1)} \hat{\boldsymbol{\beta}}^{(s)} = (X^TW\boldsymbol{z})^{(s-1)}$$

  where $W \in \mathbb{R}^{N \times N}$ diagonal matrix with (i,i)-th claims:

  $$W_{ii} = \frac{1}{V(\mu_i)} \left(\frac{d\mu_i}{d\eta_i}\right)^2$$

  and z that the i-th element as we have $z_i = \eta_i + (y_i - \mu_i)\left(\frac{d\eta_i}{d\mu_i}\right)$.
- The equation looks like the equation above for weighted least square estimation but with weights is given by w_{ii} and an adjusted response variable by z_i for the i-th observation.
- The iterative procedure can start with setting $\mu_i^{(0)} = y_i$ instead of random guess $\hat{\boldsymbol{\beta}}^{(0)}$.
- This is known as Iteratively Re-Weighted Least Square. At convergence, $\hat{\boldsymbol{\beta}} = (X^TWX)^{-1}X^TW\boldsymbol{z}$ as a minimizer of $\left\| \sqrt{W}(\boldsymbol{z} - X\boldsymbol{\beta}) \right\|^2$

- In this case of normal distributed errp, we have the link function to be:

$$\frac{d\eta_i}{d\mu_i} = 1 \qquad z_i = y_i$$

  and the above procedure reduces to non-iterative normal equation.

Remark 60. The ML estimator of ϕ is biased. The value itself can be obtained from the Peason's statistics as we have:

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

The quantity X^2/ϕ is the sume of square of zero mean and unit variance of random variable with $N-p$ degree of freedom. If the model is adequate, then approximately $X^2/\phi \sim \mathcal{X}^2_{N-p}$ and so we have:

$$\hat{\phi} = \frac{\hat{X}^2}{N-p}$$

Please note that $X^2 = \left\|\sqrt{W}(\boldsymbol{z} - X\hat{\boldsymbol{\beta}})\right\|^2$ at convergence of W and $\boldsymbol{z}$.

Remark 61. **(Large Sample Distribution of $\hat{\boldsymbol{\beta}}$)** To obtain the large sample Distribution of $\hat{\boldsymbol{\beta}}$, we use a Taylor expansion of the log-likelihood around the parameter $\boldsymbol{\beta}_0$ and evaluate this at $\hat{\boldsymbol{\beta}}$:

$$\left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} \approx \left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_0} + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\left.\frac{\partial^2 l}{\partial \boldsymbol{\beta}^2}\right|_{\boldsymbol{\beta}_0}$$

We reduce to the following ratio:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \approx \left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}_0} \Big/ \left.\frac{\partial l}{\partial \boldsymbol{\beta}^2}\right|_{\beta_0}$$

with equality in the large sample limit. The numerator has expeced value equal to zero and variance $\mathcal{I}$ is made up of the sum iid random variable l_i. The large sample limit $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ follows a $\mathcal{N}(\boldsymbol{0}, \mathcal{I})$ random variable divided by $\mathcal{I}$. This implies that as $n \to \infty$:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \sim \mathcal{N}(\boldsymbol{0}, \mathcal{I}^{-1})$$

Generated to parameter vectors $\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \mathcal{I}^{-1})$. This result is exact in case of normally distributed error.

Remark 62. **(Covariance of $\hat{\boldsymbol{\beta}}$)** The fisher information matrix is used to calculate the covariance matrix associated with ML-estimates. The (j,k)-th element of information matrix can be written as:

$$\mathbb{E}\left(\frac{\partial l}{\partial \beta_j}\frac{\partial l}{\partial \boldsymbol{\beta}_k}\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{\mathbb{E}[(Y_i - \mu_i)(Y_{i'} - \mu_{i'})]}{V_i V_{i'}}\right)\frac{d\mu_i}{d\eta_i}\frac{d\mu_{i'}}{d\eta_{i'}} x_{ij}x_{i'k}$$

where $V_i = \text{var}(Y_i)$. Let's consider the fact that we have:

$$\mathbb{E}[(Y_i - \mu_i)(Y_{i'} - \mu_{i'})] = \begin{cases} \text{var}(Y_i) = V_i & \text{for } i = i' \\ \text{cov}(Y_i, Y_{i'}) & \text{otherwise} \end{cases}$$

and, so we have

$$\mathbb{E}\left(\frac{\partial l}{\partial \beta_j}\frac{\partial l}{\partial \boldsymbol{\beta}_k}\right) = \sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{V_i}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

which follows that $\mathcal{I} = X^T W X/\phi$. As this value $\mathcal{I}$ is defined as the negative of the expected value of Hessian, it can be seen as a measure of the curvature of log-likelihood near the ML estimate of $\boldsymbol{\beta}$:

- Flat likelihood: Low negative expected second derivative implies a low information.
- Sharp likelihood: High negative expected second derivative implies a high information.

12 Confidence Interval For Model Parameter

Remark 63. In the case of unknown ϕ, we follows the same construction as before but when we want to estimate ϕ, we will have to use appropriate T-test.

Remark 64. **(CI of Parameter)** We have the following distribution:

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \mathcal{I}^{-1})$$

Now, the standard deviation of β_j is the (j, j)-th element of $\mathcal{I}^{-1}$ and, we can approximate the $100(1-\alpha)$ percent confidence interval for β_j is given as:

$$\hat{\beta}_j \pm z_{\alpha/2}\,\text{se}(\hat{\beta}_j)$$

The result can be exact if σ^2 can be known.

Remark 65. Let $\hat{\psi} = \boldsymbol{c}^T\hat{\boldsymbol{\beta}}$, then we have $\hat{\psi} \sim \mathcal{N}(\psi, \boldsymbol{c}^T\mathcal{I}^{-1}\boldsymbol{c})$, and we can construct the $100(1-\alpha)$ percent confidence interval of ψ.

Remark 66. To test the Null hypothesis, i.e $H_0 : \beta_j = 0$ for some j. We can the fact that:

$$\frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \sim \mathcal{N}(0, 1) \qquad \text{under } H_0$$

we now obtai nthe p-value the usual way. with the usually way of applying when the ϕ is unknown

Remark 67. **(Model Comparision)** We want to compare 2 models: M_0 and M where M_0 is a special case of M and let $l(\hat{\boldsymbol{\beta}}_0)$ and $l(\hat{\boldsymbol{\beta}})$ be maximum likelihood of the 2 models.

- Null Hypothesis: The subset of $p-q$ parameter out of p-parameter in linear predictor are all 0, while H_1 being an alternative hypothesis that all p are not 0.
- H_0 can be tested using a likelihood ratio test. If H_0 is true then in lage sample limit we have:

  $$2[l(\hat{\boldsymbol{\beta}}) - l(\hat{\boldsymbol{\beta}}_0)] \sim \chi^2_{p-q}$$

  If H_0 is false then M will mostly likely have higher likelihood than M_0, hence the log-likelihood ratio would be too large to be consistent with χ^2 distribution.

Definition 12.1. (Deviance) Fitting GLMs, it is useful to have a quantity that is smaller to the residual sum of squares in a linear model context. This is the deviate and defined as:

$$D = 2[l(\hat{\boldsymbol{\beta}}_{\text{sat}}) - l(\hat{\boldsymbol{\beta}})]/\phi$$

where $l(\hat{\boldsymbol{\beta}}_{\text{sat}})$ denotes the maximum likelihood of saturated model as we have 1 parameter per datum, which is based on setting $\hat{\mu}_i = y_i$, which is the highest value of likelihood that can possibility have.

Definition 12.2. (Scaled Deviance) The scaled variance is defined as $D^* = D/\phi$, which depends on the dispersion parameter. For binomial and poisson distributions the scale deviance and the deviance are the same, and we have $D^* \sim \chi^2_{N-p}$.

Remark 68. There are difference value of deviances for each kind of distribution, which are denoted as:

Distribution	Deviance
Normal	$\frac{1}{\sigma^2}\sum_{i=1}^N (Y_i - \hat{\mu}_i)^2$
Poisson	$2\sum_{i=1}^N [Y_i \log(Y_i/\hat{\mu}_i) - (Y_i - \hat{\mu}_i)]$
Poisson (Constant included)	$2\sum_{i=1}^N Y_i \log(Y_i/\hat{\mu}_i)$
Binomial	$2\sum_{i=1}^N [Y_i \log(Y_i/\hat{\mu}_i) + (n_i - Y_i)\log[(n_i - Y_i)/(n_i - \hat{\mu}_i)]]$

Remark 69. In the Binomial and Poisson case, the scaled deviance may be used to test the goodness-of-fit as we have: $D^* \sim \chi^2_{N-p}$ under proposed model. Given the definition of deviance, under H_0, likelihood ratio test can be express as:

$$D^*_0 - D^* \sim \chi^2_{p-q}$$

The dispersion parameter has to be known so that the deviance can be calculated.

Remark 70. **(Model Comparision with $\boldsymbol{\phi}$)** Under H_0 we knew that $D^*_0 - D^* \sim \chi^2_{p-q}$ and $D^* \sim \chi^2_{N-p}$. If $D^*_0 - D^*$ and D^* are treated as asymptotics independent. Under null and in the large sample limit, we have:

$$F = \frac{(D^*_0 - D^*)/p - q}{D^*/(N-p)} \sim F_{p-q,N-p}$$

This is equivalent to:

$$F = \frac{(D_0 - D)/p - q}{D/(N-p)} \sim F_{p-q,N-p}$$

Hence allow for model comparison, when ϕ is unknown.

Remark 71. **(Other Statistics)** The scores statistic $U_1, \cdots, U_p$ are defined as:

$$U_j = \frac{\partial l}{\partial \beta_j}$$

for $j = 1, \cdots, p$ where l is the log-likelihood function. Let $\boldsymbol{U} = (U_1, \cdots, U_p)^T$, where we have the following properties of the vectors:

- Expectation: $\mathbb{E}[\boldsymbol{U}] = 0$
- Covariance Matrix: $V(\boldsymbol{U}) = \mathcal{I}$
- Asymptotics Sampling Distribution: $\boldsymbol{U} \sim \mathcal{N}(\boldsymbol{0}, \mathcal{I})$ and $\boldsymbol{U}^T\mathcal{I}^{-1}\boldsymbol{U} \sim \chi^2_p$

Remark 72. **(Wald Statistics)** If $\hat{\boldsymbol{\beta}}$ is maximum likelihood estimator of $\boldsymbol{\beta}$ that it can be shown asymptotically as:

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathcal{I}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_p$$

where $\mathcal{I}(\hat{\boldsymbol{\beta}})$ is the information matrix evaluate at $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$. Since result holds for subset of parameter, which can be done using the submatrix. Score statistics can be used as alternative to likelihood ratio test for multiple parameter hypothesis testing.

13 Binomial Data and Logistic Regression

Remark 73. Suppose there are N response $Y_1, \cdots, Y_N$ are independent such that $Y_i \sim \text{Bin}(n_i, \pi_i)$ where we have $g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ for some link function g and $\mu_i = \mathbb{E}[Y_i] = n_i\pi_i$ for $i = 1, \cdots, N$. This includes binary case for $n_i = 1$ for all i. The link function is logit link, which gives the logistic model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

Other models can be used such as:

- Probit: $\Phi^{-1}(\pi_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ where it links cumulative normal distribution from $\mathcal{N}(0,1)$.
- Complementary Log-Log: As we have $\log[-\log(1-\pi_i)] = \boldsymbol{x}_i^T\boldsymbol{\beta}$

The expression for the probability of success π is obtained by inverting the equation for the model, which are given as:

- Logistic: $\pi = 1/(1+\exp(-\eta))$ which is a cdf of logistic distribution.
- Probit: $\pi = \Phi(\eta)$ which is a cdf of standard deviation.
- Complementary Log-Log: $\pi = 1 - \exp[-\exp(\eta)]$, which is extream value distribution.

Remark 74. In the current contex, $\eta = \boldsymbol{x}^T\boldsymbol{\beta}$ denotes the linear predictor for any observation:

- After $\boldsymbol{\beta}$ has been estimated by $\hat{\boldsymbol{\beta}}$ plugging it in to $\hat{\eta} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ giving the estimated probability that a new observation of Y_{N+1} is 1 given X_{N+1}.
- The choices ensure that $0 \le \pi \le 1$ as required. This results in the function that has properties that $-\infty < g(y) < \infty$ with the consquence that there are no constraints on the unknown parameter in the linear predictor.

Remark 75. **(Interpretation of Logistic Regression Parameter)** If π denotes the probability of success, when the logit link function given by $\log(\pi/(1-\pi))$ is the log of odds on success. So the coefficient β_j of the explaination variable x_j in the logistic regression model means that the rate of change of odd with x_j given the other explainatory variable to be cosntant.

Remark 76. Suppose that x_i is indicator variable with just 2 levels: 0 and 1 then at $x_1 = 0$, we have:

$$\text{logit}(\pi) = \beta_0 + \beta_2 x_2 + \cdots + \beta_m x_m$$

On the other hand at $x_1 = 1$, we have

$$\text{logit}(\pi') = \beta_0 + \beta_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

Let's subtract both and we have:

$$\begin{aligned}\beta_i &= \text{logit}(\pi') - \text{logit}(\pi) \\ &= \log\left\{\frac{\pi'}{1-\pi'}\right\} - \log\left\{\frac{\pi}{1-\pi}\right\} \\ &= \log\left\{\frac{\pi'/(1-\pi')}{\pi/(1-\pi)}\right\}\end{aligned}$$

The log of ratio of the odds on the success of 2 values of x_1 or the log-odd-ratio. The odd on a success when $x_1 = 1$ is $\exp(\beta_1)$ times the odd on success when $x_1 = 0$ given other values being constant.

Remark 77. **(Likelihood Equation)** We have the general equation, which our setting follows, but for logistic model, they reduced to:

$$\sum_{i=1}^{N}(y_i - \hat{\mu}_i)x_{ij} = 0$$

for $j = 1, \cdots, p$

Remark 78. **(Maximum Likelihood Estimate)** For the iterative procedure, we have:

$$(X^TWX)^{(s-1)}\hat{\boldsymbol{\beta}}^{(s)} = (X^TWZ)^{(s-1)}$$

For the logistic model the (i,i)-th element of the diagonal matrix is given by:

$$w_{ii} = \frac{1}{V_i}\left(\frac{d\mu_i}{d\eta_i}\right)^2$$

where $V_i = \text{var}(Y_i) = V(\mu_i) = n_i\pi_i(1-\pi_i)$, which we also have:

$$\frac{d\mu_i}{d\eta_i} = n_i\pi_i(1-\pi_i)$$

Hence, we have $w_{ii} = n_i\pi_i(1-\pi_i)$. For the $\boldsymbol{z}$, we have its i-th element to be

$$z_i = \eta_i + (y_i - \mu_i)\frac{d\eta_i}{d\mu_i} \qquad \text{where} \qquad \frac{d\eta_i}{d\mu_i} = \frac{1}{n_i\pi_i(1-\pi_i)}$$

Remark 79. **(Sampling Distribution)** We have the following sampling distribution of $\hat{\boldsymbol{\beta}}$ as we have:

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, (\boldsymbol{x}^T W \boldsymbol{x})^{-1})$$

Remark 80. **(Deviance)** For testing goodness of fit of a particular model is given by:

$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{Y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right)\right]$$

where $\hat{\mu}_i$ are the fitted $\hat{\mu}_i$ under the model. If model is true, then $D \sim \chi^2_{N-p}$ which depends provides a test statistics for goodness-of-fit test.

Remark 81. The test for $H_0 = \beta_j = 0$, which we have under H_0.

$$\frac{\hat{\beta}_j}{\operatorname{se}(\hat{\beta}_j)} \sim \mathcal{N}(0, 1)$$

Remark 82. **(Testing)** The test for omission of the j-th explainatory variable, given other explainatory variable in the model.

- Given the test set H_0 as ν of regression parameter $\beta_1, \cdots, \beta_m$ are 0. We are assuming that the linear predictor consists of cosntant terms and terms from m explainatory variable and H_0 test for the omission of ν variable where $\nu \leq m$.
- Let D_0 and D denotes the deviate under H_0 and the full model, respectively. The likelihood ratio test is: under H_0

$$D - D_0 \sim \chi^2_\nu$$

- In the special case, we have $H_0 : \beta_0 = \beta_1 = \cdots = \beta_m = 0$. Under H_0 the β_i's are all equal to and the MLE of the common probability of an success is the observed propotion of success.

This resulting deviate be denoted by C, as the analysis of deviance table as we have:

Source of Variation	**Deviance**	**Df**
Regression	$C - D$	m
Residual	D	$N - m - 1$
Total	C	$N - 1$

Remark 83. **(Peason Chi-Squared Statistics)** Given the alternative test of goodness of it. The statistic is denoted by X^2, which is based on the following table of observed and fitting frequences:

Observation	1	2	$\cdots$	N
Num of Success	Y_1	Y_2	$\cdots$	Y_N
Num of Failure	$n_1 - Y_1$	$n_2 - Y_2$	$\cdots$	$n_N - Y_N$

with the similar table of fitted value in which Y_i is replaced by $\hat{\mu}_i = n_i\hat{\pi}_i$. Using a common relation o for observed frequency and e for fitted (expected) frequency. D and X^2 has the form:

$$D = 2\sum o \log\left(\frac{o}{e}\right) \qquad X^2 = \sum \frac{(o-e)^2}{e}$$

This for of deviance D is denoted by G^2. The Peason chi-square statistic is after some algebra given by:

$$X^2 = \sum_{i=1}^{N} \frac{(Y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1 - \hat{\pi}_i)}$$

(if the model is true) has same large sample distribution and the deviance D, which is χ^2_{N-p}. Using Taylor series expansion of $s\log(s/t)$ about $s=t$ up to quadratic term, it can be shown that $D \sim \chi^2$. (It likely to be poor if any of fitted values in $2N$ are small).

Remark 84. **(Binary Data)** If each observation has difference pattern of the explainatory variable. The result for the estimation still holds. The deviance can be shown to be depend on the binary observation through the fitted value and so isn't for assessting goodness of fit.

Remark 85. **(Hosmer and Lemeshow)** They proposed a test obtained by groupping observations into about $g \approx 10$ groups of observation with some number per group according to their predicted probability.

- We are given $2 \times g$ table for which χ^2 is calculated.
- Large sample distribution of resulting statistic if the model is the suggested to approximately be χ^2_{g-2}.
- This is the same for binomial but it doesn't relative in the certain cases.

Remark 86. **(Checking Model Adequecy)**

- Raw Residual:
$$\hat{e}_i = Y_i - n_i\hat{\pi}_i$$
How well the raw data is fitted
- Peason Chi-Square:
$$X_i = \frac{\hat{e}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$
So that the chi-square statistics $X^2 = \sum_{i=1}^N X_i^2$. These standardized by estimated standard deviation of Y_i making them the compatible in size.
- Standardized Peason Residual:
$$r_{p_i} = \frac{X_i}{\sqrt{1-h_{ii}}}$$
where h_{ii} is the diagonal of the hat matrix given by $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$. The variance of this is 1 and comparable in size of X-space. For the mathematical comparison, these are better than X_i. However, X_i is more naturally interpreted in terms of which points are welled-fitted.
- Deviance Residual: And, we have:
$$d_i = \text{sign}(\hat{e}_i)\left\{2\left[y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i-y_i)\log\left(\frac{n_i-y_i}{n_i-\hat{\mu}_i}\right)\right]\right\}^{1/2}$$
so that the deviance is $\sum_{i=1}^N d_i^2$, the $\text{sign}(\hat{e}_i)$ gives d_i same sign as $\hat{e}_i$.
  - The interpretation is formalized how strongly observation contribute to the deviance (The standard way of measuring the quality of the overfit).
  - They show to what extent the observation indicates that the model is violated and rather saturated model is needed.
- Standardized Deviance Residual:
$$r_{D_i} = \frac{d_i}{\sqrt{1-h_{ii}}}$$
Interpretation: This make the d_i directly naturally comparable by unifying their variance and adjust for location in x-space.

- Cook Statistics:
$$D_i = \frac{1}{p}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T X^T W X(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})$$
The numerator of D_i is the weighted sum of the square difference of the fitted logic of the π_i's with and without i-th observation with w_{ii} as weight. The alternative D_i's are calculate in the following form:
$$D_i = \frac{1}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right) r_{P_i}^2$$
This quantity the effect on $\hat{\boldsymbol{\beta}}$ of omitting observation of the i-th.

Remark 87. **(Model Selection)** There are many kinds of model selection as we have:

- Akaike Information Criterion:
$$\text{AIC} = D + 2p + \text{const}$$
where D is deviance statistic and p is number of parameter in linear predictor as LOO-CV and we can be used as well.
- 2 models are compared by the difference of deviance: Suppose model M_1 with p_1 regression parameter is submodel of M_2 with p_2 parameter. Let $l(\hat{\beta}_j)$ be maximum value of the log likelihood under M_j and D_j denotes the deviate of model M_j. We want to test:
$$H_0 : M_1 \qquad \text{vs} \qquad H_1 : M_2$$
Then the likelihood test statistic of $-2\log$ likelihood ratio is:
$$2[l(\hat{\beta}_2) - l(\hat{\beta}_1)] = D_1 - D_2$$
Under null hypothesis has approximate $\chi^2_{p_2-p_1}$ distribution.

Remark 88. **(Analysis of Variance)** The regression deviance $C - D$ can be partitioned in same way as for normal linear model, where RSS is replaced by deviance D.

14 Contingency Tables

Remark 89. We have the following construction of table:

- Let's the row variable called A with I possible categories and the column variable be called B with J possible categories.
- Now we consider $I \times J$ contingency table that has been obtained by allocating a random sample of N observation on the pair of variable A and B to IJ possible combination.
- For cell (i, j) of contingency table as we have $i \in [I]$ and $j \in [J]$ as we let:
  - π_{ij} is the probability that an observation being to the cell.
  - μ_{ij} is the expeced frequency.
- Please Recall that $\mu_{ij} = N\pi_{ij}$. If the row variable A is independent of column variable B then
$$\pi_{ij} = \pi_{i+}\pi_{j+}$$
where we have π_{i+} is the probability observation belong to row i and the sub-script indicates summation over the corresponding subscript. Similar for column j.

- If A is independent of B, we have
$$\mu_{ij} = \frac{\mu_{i+}\mu_{+j}}{N}$$
where $\mu_{i+} = N\pi_{i+}$ and its expected frequency of the row i. Similar for the row j.

Remark 90. **(Model Under Null-Hypothesis)** Typical null hypothesis is H_0 assuming that A and B are independent. Taking the logarithm to get:

$$\log \mu_{ij} = \log \mu_{i+} + \log \mu_{+j} - \log N$$

We rewite as the sum of 3 terms, one depending on i and on j and a constant. If the variable A and B aren't independent then the equality doesn't hold. Let ϕ_{ij} then denote the difference between LHS and RHS, then the alternative is:

$$\log \mu_{ij} = \lambda + \alpha_i + \beta_j + \phi_{ij}$$

Remark 91. **(Constriants on Parameter)** The linear predictor in alternative hypothesis contains more parameter than cell in the contingency table. So we need to have a constraint, which can be set by either:

- Set the parameter for one specific outcome to zero:
$$\alpha_1 = 0 \qquad \beta_1 = 0 \qquad \phi_{1j} = 0 \qquad \phi_{i1} = 0$$
This means that for every variable categories 1 is regared as reference category.
- Make parameter sum to zero as we have:
$$\sum_i \alpha_i = 0 \qquad \text{and} \qquad \sum_j \phi_{ij} = \sum_i \phi_{ij} = 0$$

The model of alternative hypothesis is a saturated log-linear model for 2-ways contingency table. There are IJ effective parameter (number of parameter - number of effective constraints). This is equal to the number of frequency in the table so every frequency can be fitted perfectly:

- The submodel (null hypothesis) is unsatuarated.
- The estimation of alternative hypothesis gives $\hat{\mu}_{ij} = n_{ij}$, which is a perfect fit and $G^2 = 0$

Compare the alternative hypothesis with model of ANOVA 2-ways, there are similar in character with logarithm of expected response on the LHS instead of expected response. So the model is log-linear model.

Remark 92. **(Parameter)** The parameter in log-linear model $(\lambda, \alpha_i, \beta_j, \phi_{ij})$ doesn't have straightfoward interpretation. It is mainly of interest whether certain parameter vanishes or not as this will imply certain independent.

Remark 93. **(Interaction)** The extra term introduced is called interaction term. For testing hypothesis is based on goodness-of-fit statistic as we have:

$$G^2 = 2\sum o \log\left(\frac{o}{e}\right)$$

where o and e denote observed and fitted expected frequency. We use it rather Pearson Chi-Square statistics.

Remark 94. **(Fitting Model)** The models are fitting using maximum likelihood which requires the specification of the joint distribution of the observation (observation frequency) as the joint distribution is multinomial distribution.

- It can be shown that joint distribution of independent Poisson random variable conditioned on their sum is a multinomial distribution.
- In practice log-linear model for contingency table data are fitted as if the observed frequency are independent Poisson.

Remark 95. **(Goodness-of-Fit and Model Checking)**

- Goodness of fit is done as for the Poisson data with G^2
- Nested model compared by difference of their G^2 values.
- Standardized Residual: Similar to case of logistic regression for 2-ways contingency table as we have:
  - Raw-Residual as we have $\hat{e}_{ij} = n_{ij} - \hat{\mu}_{ij}$
  - Peason or χ^2-residual:
  $$X_{ij} = \frac{\hat{e}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$$
  So that chi-squared statistics $X^2 = \sum_i \sum_j X_{ij}^2$
  - Standardized Peason Residual:
  $$r_{P_{ij}} = \frac{X_{ij}}{\sqrt{1 - h_{(ij)}}}$$
  where $h_{(ij)}$ is the leverage for the other with combination (ij) of 2 factors.
  - Deviate residual can be found:
  $$d_{ij} = \text{sign}(\hat{e}_{ij}) \left\{ 2 \left[y_{ij} \log\left(\frac{y_{ij}}{\hat{\mu}_{ij}}\right) - (y_{ij} - \hat{\mu}_{ij}) \right] \right\}^{1/2}$$
  note that $D = \sum_i \sum_j d_{ij}^2$
  - Standardized Deviance Residual:
  $$r_{D_{ij}} = \frac{d_{ij}}{\sqrt{1 - h_{(ij)}}}$$

15 Generalized Additive Model

Definition 15.1. GAM The generalized additive model as we have:

$$g\left\{\mathbb{E}[Y_i]\right\} = \eta_i = X_i^* \theta + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}, x_{i4}) + \cdots$$

where X_i^* is the i-th row of X^*, which is the model matrix for any parameter model components with parameter vector θ and f_i is smooth function over covariate x_{ij}. It is subjected to a constraint that $\sum_i f_j(x_{ij}) = 0$ for each j.

Remark 96. The model GAM can flexibly determine the function value of the relationship between response and some explainatory variable avoid the drawnback of modeling using parameter relationship. One can model the discrete and continuous variable.

Remark 97. Smooth term can be represented by regression spine. Linear combination of basis function $b_{jk(x_j)}$ and regression parameter β_{jk} as we have:

$$f_j(\boldsymbol{x}_j) = \sum_{k=1}^{q_j} \beta_{jk} b_{jk}(\boldsymbol{x}_j)$$

as we have j is smooth term for j-th explainatory variable. The regression spline of 2 covariance, which can be written as:

$$f_{jp}(\boldsymbol{x}_j, \boldsymbol{x}_p) = \sum_{k=1}^{q_j} \beta_{jp,k} b_{ip,k}(\boldsymbol{x}_j, \boldsymbol{x}_p)$$

Remark 98. **(Problem With Polynomial Basis)**

- As the number of polynomial above, the increasingly colinear.
- Highly correlated parameter estimator leads to high estimator variance and numerical problem.
- We can use orthogonal basis as we still give the problem over a domain (but useful in a single point).
- Practical solution is the use the continuous variable can be categorized into groups based on interval and frequency.
- Another problem comes from the cut points, in which the relationship between a response variable and set of covariates is flat in the interval (by the assumptions).
- To overcome all issue, we can use the spine bound are typically used to determine flexibly the relationship between the continuous predictor and the outcome of interest, which avoid the disadvantage of categorization, which are not as correlated as polynomial basis function.
- Common choices for responsibility smooth function includes smoothing spine as we can place the knots at every data point, and referred to as full rank smoother because the size of spine basis is equal to number of observation.
- However, this leads to as many parameter as there are data which result in expensive computation regression.

Remark 99. **(Parameter Estimation)** The regerssion can overfit which we can consider the model to maximize the following function:

$$l(\beta) - \frac{1}{2}\sum_j \lambda_j \int \left\{ f_j^{d_j}(x_j) \right\}^2 \, \mathrm{d}x_j$$

Remark 100. **(Regularization)** This can be written as the quadratic form $\boldsymbol{\beta}$ with known coefficient matrix S_j. Let's consider $d_j = 2$ and for regression spine basis in one 1D as we have:

$$\begin{aligned}
\int \left\{ f_j^{d_j}(x_j) \right\}^2 \, \mathrm{d}x_j &= \int \left\{ \frac{\partial^2 f_j(x_j)}{\partial x_j^2} \right\}^2 \, \mathrm{d}x_j \\
&= \int \left\{ \frac{\partial^2 \sum_{k=1}^{q_j} \beta_{jk} b_{jk}(x_k)}{\partial^2 x_j} \right\}^2 \, \mathrm{d}x_j \\
&= \int \left\{ \boldsymbol{\beta}^T b_j''(x_j) \right\}^2 \, \mathrm{d}x_j \\
&= \int \boldsymbol{\beta}^T b_j''(x_j) b_j''(x_j)^T \boldsymbol{\beta} \, \mathrm{d}x_j \\
&= \beta^T \left\{ \int \boldsymbol{b}_j''(x_j) b_j''(x_j)^T \, \mathrm{d}x_j \right\} \boldsymbol{\beta} \\
&= \boldsymbol{\beta}^T S_j \boldsymbol{\beta}
\end{aligned}$$

The estimator of $\boldsymbol{\beta}$ is given by:

$$\hat{\boldsymbol{\beta}} = (X^T W X + S)^{-1} X^T W \boldsymbol{z}$$

where $S = \sum_j \lambda_j S_j$ as $\boldsymbol{\beta}$ is bias due to the penalty. The value of λ_j is done using Cross-Validation or generalized AIC.

Remark 101. **(Inference)** Let's consider the genertic smooth model component $f(x_j)$ as the interval can be constructed by seeking some constant C_i and A such that

$$\mathrm{ACP} = \frac{1}{2}\mathbb{E}\left\{ \sum_i \mathbb{I}\left(\left| \hat{f}(x_i) - f(x_i) \right| \le q_{\alpha/2} A / \sqrt{C_i} \right) \right\} = 1 - \alpha$$

As we have $\alpha \in (0, 1)$ and $q_{\alpha/2}$ is the $\alpha/2$ critical point from standard normalize distribution.

Remark 102. Defining $b(x) = \mathbb{E}[\hat{f}(x)] - f(x)$ and $v(x) = \hat{f}(x) - \mathbb{E}[\hat{f}(x)]$ and so $\hat{f} - f = b + v$, and having I be random variable unifying distribution on $\{1, 2, \cdots, n\}$ as we have:

$$\mathbb{P}\left(|B + V| \leq q_{\alpha/2} A\right)$$

as we have $B = \sqrt{C_I} b(x_I)$ and $V = \sqrt{C_I} v(x_I)$. It is necessary to find a distribution of $B + V$ and value of C_i and A so that the requirement is met.

Remark 103. The condition above the approximately met with posterior distribution:

$$\boldsymbol{\beta} | \boldsymbol{y} \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, (\mathcal{I} + \mathcal{S})^{-1})$$

Confidence Interval can be easily obtained. Any strictly parameter model component to obtain confidence interval is equivalent to using classical likelihood results. This is because it isn't penalized.