CHAPTER - IV
REGRESSION MODELS

4.1 Introduction

Regression is a statistical measure that attempts to determine the strength of the relationship between one dependent variable and a series of other changing variables.

A regression model is the functional relationship between the response variable and the explanatory variable or independent variable.

4.2 Multiple linear Regression Model

A regression model with one explanatory variable $X$ that has a linear relationship with a response $Y$ is called simple linear regression model. The simple linear regression model of the form is

$$Y_i = \beta + \beta X_i + \epsilon_i$$

Where $Y_i$ denotes the dependents variable, the $X_i$ is explanatory variable or independent variable, $\beta$ is the intercept, $\beta$ is the regression coefficient and $\epsilon_i$ is error term.

A simple linear regression model is extended with more than one explanatory variable $X_1, X_2, \ldots, X_k$ which is linearly related with the response variable $Y$ is called multiple regression model. The functional form of the relationship is

$$Y = f(X_1, X_2, \ldots, X_k)$$
This relationship shows that \( Y \) is a function of \( k \) explanatory variables. It is described in a simple way by a multiple linear regression equation of the form

\[
Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i
\]

Where \( Y_i \) denotes the dependent variable, the \( X_i \) are explanatory variables, the subscript \( i \) refers to the \( i^{th} \) observation and first subscript used in describing the explanatory variables identifies the variable in model, the \( \epsilon_i \) is a stochastic disturbance term, \( \beta \) is the intercept of the model and \( \beta_1, \beta_2, \ldots, \beta_k \) are the regression coefficients.

Assumptions of the multiple linear Regression Model are stated as follows,

- \( u_i \) is normally distributed
- \( E(u_i) = 0 \)
- \( E(u_i^2) = \sigma^2 \)
- \( E(u_i u_j) = 0 \) for \( i \neq j \)
- Each of the explanatory variables is non-stochastic with fixed values in repeated samples, and such that for any sample size \( \sum_{i=1}^{n} X_{ki} - \bar{X}_k/n \) is a finite number different from zero.
- The number of observations exceeds the number of coefficients to be estimated.
- No exact linear relationship exists between any of the explanatory variables.
The multiple linear regression model is also represented in the matrix form as

\[ Y = \beta \cdot X \]

Where

\[
\begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_n
\end{bmatrix} =
\begin{bmatrix}
1 & X_{11} & X_{21} & \cdots & X_{k1} \\
1 & X_{12} & X_{22} & \cdots & X_{k2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & X_{1n} & X_{2n} & \cdots & X_{kn}
\end{bmatrix}
\]

\[
\beta =
\begin{bmatrix}
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_k
\end{bmatrix}
\]

\[
u =
\begin{bmatrix}
u_1 \\
u_2 \\
\vdots \\
u_n
\end{bmatrix}
\]

In general, \( Y \) is an \( n \times 1 \) vector of the observations, \( X \) is an \( n \times k \) matrix of the levels of the regressor variables, \( \beta \) is a \( k \times 1 \) vector of the regression coefficients, and \( u \) is an \( n \times 1 \) vector of random errors. The following assumptions are used in the representation of regression model in the matrix form.

1. \( u \sim N(0, \Sigma) \) where \( 0 \) is a column vector of zeros and \( \Sigma \) is a covariance matrix.
2. \( \Sigma = \sigma^2 I_n \) where \( I_n \) is an identity matrix of order \( n \times n \) with units in the principal diagonal and zeros everywhere else. \( X \) is a set of fixed number. It means that elements of the matrix are non-stochastic with values fixed in repeated samples and the matrix \( \frac{1}{n} \Sigma' \Sigma \) is non-singular and such that for any sample size its elements are finite.
3. \( X \) has rank \( k < n \). It means that the number of observations must exceed the number of parameters to be estimated and that no exact linear relationship exists between any of the \( X \) variables.
4.2.1 Least square method of Estimation

Least square method is used to estimate the regression coefficients. Least square principles are also applied for the regression model which is expressed in matrix. As the multiple regression model is given as,

\[ Y = \beta \cdot X \]

Where

\[
Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1k} \\ 1 & X_{12} & X_{22} & \ldots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \ldots & X_{kn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}
\]

The sum of the residual square is obtained as,

\[
\sum_{i=1}^{n} u_i^2 = u'u = \sum_{i=1}^{n} u_i^2
\]

\[
= I - \hat{\beta} \cdot \hat{\beta} = \hat{\beta} \cdot \hat{\beta}
\]

\[
= \hat{r}' - \hat{\beta}' \cdot \hat{\beta} + \hat{\beta}' \cdot \hat{\beta} + \hat{\beta}' \cdot \hat{\beta}
\]

\[
= \hat{r}' - \hat{\beta}' \cdot \hat{\beta} + \hat{\beta}' \cdot \hat{\beta}
\]

(Since \( \hat{\beta}' \cdot \hat{\beta} \) is a scalar, thus it will equal to \( Y' \cdot \hat{\beta} \))

\[
\therefore \sum_{i=1}^{n} u_i^2 = \hat{r}' - \hat{\beta}' X' + \hat{\beta}' \cdot \hat{\beta}
\]

\[
\frac{\partial \sum_{i=1}^{n} u_i^2}{\partial \hat{\beta}} = \hat{r}' - X' \cdot \hat{\beta}
\]

\[
\therefore 2X' \cdot \hat{\beta} - X'Y = 0
\]

\[
\hat{\beta}' \cdot \hat{\beta} = X' \cdot \hat{\beta}
\]
Premultiplying by $\mathbf{X}'\mathbf{X}^{-1}$ to both sides we have,

$$
\mathbf{X}'\mathbf{X}^{-1} \mathbf{X}' \beta = \mathbf{X}'\mathbf{X}^{-1} X''
$$

$$
\therefore \beta = \mathbf{X}'\mathbf{X}^{-1} X''
$$

Where $X'$ is a column vector, $X$ is a row vector. This is the fundamental result for the least square estimators.

The least square estimator $\beta = \mathbf{X}'\mathbf{X}^{-1} X''$ is the best linear unbiased estimator of $\beta$, having minimum variance in the class of linear unbiased estimators.

### 4.2.2 Test for Adequacy of Regression model

The test of adequacy of regression model is a test is used to check the linear relationship between the response variable $y$ and any of the regressor variables $x_1, x_2, \cdots, x_k$. the test procedure is described as follows:

The null hypothesis is stated as

$$
H_0 : \beta = \beta = \beta = \cdots = \beta = \beta = \cdots
$$

and the alternative hypothesis is also stated as

$$
H_1 : \beta \neq \beta \text{ for atleast one } j
$$

$$
j = 2, \cdots, k
$$

To test, the said null hypothesis, the analysis of variance has been used as in the simple regression model.

The total sum of squares (SST) is divided into a sum of squares due to regression (SSR), and a residual sum of squares (SSRes). Thus

$$
SS_T = SS_R + SS_{Res}
$$
The computational formula for $SS_R$ is found by starting with

$$SS_R = \hat{y} - \frac{\sum_{i=1}^{n}y_i}{n}.$$  

(4.2.2)

Therefore, the regression sum of squares is

$$SS_R = \beta \hat{y} - \frac{\sum_{i=1}^{n}y_i}{n}.$$  

(4.2.2a)

From the equation (4.2.2) and (4.2.2a), the residual sum of squares can be written as

$$SS_{Res} = SS_T - SS_R$$

$$SS_{Res} = \sum_{i=1}^{n}y_i^2 - \beta \left( \frac{\sum_{i=1}^{n}y_i}{n} \right)^2 - \left( \hat{y} - \frac{\sum_{i=1}^{n}y_i}{n} \right)^2.$$  

The residual sum of squares is

$$SS_{Res} = \hat{y} - \hat{\beta} \hat{y}.$$  

By the definition of an F statistic,

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

if the observed value of $F_0$ is large than the expected $F_{k,n-k-1}$ statistic value, then $H_0$ is rejected. i.e. accept $H_1: \beta \neq \gamma$, $j = 2, \ldots, k$.

**4.2.3 Multiple correlation coefficient**

The multiple correlation coefficient measures the strength of the association between $Y$ and the best fitting linear combination of the $X$’s, which is linear combination of $\hat{y}_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$. It is always nonnegative.
In fact, no other linear combination of the X’s will have as great a correlation with Y. It is denoted by $R_{y|x_1, x_2, \ldots, x_k}$.

The multiple correlation coefficient is a generalization of the simple correlation coefficient $r$ to the case of several independent variables. Two computational formula $(\hat{Y})$ (David G. Kleinbaum et al., 2009) provide useful interpretations of the multiple correlation coefficient and its square:

$$R_{y|x_1, x_2, \ldots, x_k} = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})(\hat{Y}_i - \overline{\hat{Y}})}{\sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2 \sum_{i=1}^{n} (\hat{Y}_i - \overline{\hat{Y}})^2}}$$  \hspace{1cm} (4.2.3)

$$R^2_{y|x_1, x_2, \ldots, x_k} = \frac{\sum_{i=1}^{n} \hat{Y}_i^2 - \overline{\hat{Y}}^2 - \sum_{i=1}^{n} Y_i^2 - \overline{Y}^2}{\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2} = \frac{SS_Y - \overline{SS}_Y}{SS_Y}$$ \hspace{1cm} (4.2.3a)

Where $\hat{Y}_i = \beta_0 + \beta_1 Y_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$ (the predicted value for the $i^{th}$ individual) and $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$. The eqn.(4.2.3a), is most useful for assessing the fit of the regression model. The eqn.(4.2.3) indicates that $R_{y|x_1, x_2, \ldots, x_k}$ is the simple linear correlation between the observed values ($Y$) and the predicted values.

### 4.2.4 Test on individual regression coefficients

When one of the regressor not equal to zero in overall model adequacy test, then it can be find out which regressor variable is contributes significantly in the model. In this situation individual regression coefficients test can be applied. The
statement of hypotheses for testing the significance of any individual regression coefficient is

\[ H_0 : \beta_j = 0 \]

\[ H_1 : \beta_j \neq 0 \]

If \( H_0 : \beta_j = 0 \) is accepted, then it can be inferred that the regressor \( x_j \) can be deleted from the model. The test statistic of the hypothesis is

\[ t_0 = \frac{\hat{\beta}_j}{\sqrt{\sigma^2 C_{jj}^{-1}}} = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \sim t \text{ distribution with } n - d.f \]

Where \( C_{jj} \) is the diagonal element of \( \mathbf{X}^T \mathbf{X}^{-1} \) corresponding to \( \beta_j \). The null hypothesis \( H_0 : \beta_j = 1 \) is rejected if \( |t_0| > t_{\alpha/2, n-k} \). Note that this is really a partial or marginal test because the regression coefficient \( \hat{\beta}_j \) depends on all the other regressor variables \( x_j(i \neq i) \) that are in the model. Hence it is a test of the contribution of \( x_j \) given the other regressors in the model.

4.2.5 Multicollinearity

It is the existence of the strong relationship between two or more regressors. Two independent variables are said to be complete collinearity if their correlation coefficient is 1, and no collinearity means that their correlation is 0. Multicollinearity exists when any one independent variable is highly correlated with a set of other independent variables.

Sources of Multicollinearity

1. The data collection method employed
2. Constraints on the model or in the population
3. Model specification
4. An overdefined model
4.2.6 Forward addition and backward elimination method

The procedures of forward addition and backward elimination are largely trial and error processes to fit the best model. The forward addition method starts with the assumption that there are no regressors in the model other than the intercept. An optimal subset can be made by inserting the variable once at time. The first independent variable is selected for entry into the equation is one that has the largest correlation with the response variable. This entering variable is called the regressor, which will produce the largest $t$ statistic for testing the significance of regression. This regressor is entered if the $t$ statistic is greater than the $t$ table value. In the similar manner the second independent variable selected for entry into the equation after adjusting the effect of first entered independent variable. In the same way, the regressors with large coefficients are entered into the equation to fit a good model.

The backward elimination method is working opposite to the forward method. That is, this method starts with all variables in the model. The variable which has the least partial F statistic compared with the F table value is to be eliminated. In the similar manner the variables are eliminated until get the largest $t$ statistic than the $t$ table value. Note that the $t$ test is equal to the $F$ test when test for significance of single variable.

4.2.7 Multicollinearity Diagnostics

Several techniques are available to detecting the multicollinearity. Here the Variance inflation factor is used to detect the multicollinearity.
4.2.8 Variance Inflation Factors

The variance inflation factor (VIF) is the measure, which can be used to detect multicollinearity. The variance inflation factor for each term in the model measures the combined effect of the dependencies among the regressors on the variance of that term. It is denoted by \( C_{ij} \). The diagonal elements of the \( C = (X'X)^{-1} \) matrix are very useful in detecting multicollinearity. The \( C_{jj} \) is the \( j^{th} \) diagonal element of \( C \), it can be written as \( C_{jj} = \left( -R_j^2 \right)^{-1} \), Where \( R_j^2 \) is the coefficient of the determination obtained when \( x_j \) is regressed on the remaining \((k-1)\) regressors. If \( x_j \) is nearly orthogonal to the remaining regressors, \( R_j^2 \) is small and \( C_{jj} \) is near to unity, while if \( x_j \) is nearly linearly dependent on some subset of the remaining regressors, \( R_j^2 \) is near unity and \( C_{jj} \) is large. Since the variance of the \( j^{th} \) regression coefficients is \( C_{jj} \sigma \), and it can be view \( C_{jj} \) as the factor by which the variance of \( \hat{\beta}_j \) is increased due to linear dependencies among the regressors.

4.2.9 Partial correlation coefficient

The partial correlation coefficient is a measure, which is used to find the strength of the linear relationship between two variables after controlling the effects of remaining variables. The partial correlation between the variables \( Y \) and \( X \) after the control variables \( Z_1, Z_2, \ldots, Z_p \), then the corresponding partial correlation coefficient is denoted by \( r_{yx|z_1,z_2,\ldots,z_p} \). The order of the partial correlation depends on the number of controlled variables. Now the first order partial correlation of the form is \( r_{yx|z} \), second order partial correlation of the form is \( r_{yx|z_1,z_2} \), and in general, \( p^{th} \) order partial correlation of the form is \( r_{yx|z_1,z_2,\ldots,z_p} \) (David G. Kleinbaum et.al., 2009).
The partial correlation between the variables Y and X holding constant the variable Z is computed by the following formula

\[
r_{YX.Z} = \frac{r_{YX} - r_{YZ}r_{XZ}}{\sqrt{1 - r_{YZ}^2} \sqrt{1 - r_{XZ}^2}}
\]

where \( r_{YXZ} \) is the first order partial correlation coefficient and \( r_{YX}, r_{YZ}, r_{XZ} \) are zero order correlation coefficients.

The partial correlation between the variables Y and X holding constant the variables \( Z_1 \) and \( Z_2 \) is computed by the following formula:

\[
r_{YX.Z_1Z_2} = \frac{r_{YX,Z_1} - r_{YZ_2,Z_1}r_{XZ_2,Z_1}}{\sqrt{1 - r_{YZ_2,Z_1}^2} \sqrt{1 - r_{XZ_2,Z_1}^2}}
\]

where \( r_{YX,Z_1Z_2} \) is the second order partial correlation coefficient and \( r_{YX,Z_1}, r_{YZ_2,Z_1}, r_{XZ_2,Z_1} \) are the first order partial correlation coefficients.

4.3 Poisson Regression Model

Poisson regression analysis is a regression technique, which is applied for modelling dependent variables that describe count data. The basic Poisson regression model relates the probability function of a dependent variable \( y_i \) to a vector of independent variables \( x_i \). Let \( k \) be the number of regressors, \( x_i \) is then a column vector of dimension \( (k \times 1) \) and \( n \) is the number of observations in the sample. The standard univariate Poisson regression model have the following three assumptions:
Assumption

- \( f(y | \lambda) = \frac{e^{-\lambda} \lambda^y}{y!} \quad y = 0, 1, 2, \ldots \)

where \( f(y | \lambda) \) is the conditional probability function of \( y \) given \( \lambda \) and the parameter \( \lambda > 0 \).

The mean and variance of the Poisson distribution is can be derived as

\[
E(x) = \sum_{x=0}^{\infty} xp(x) \\
= \sum_{x=0}^{\infty} x e^{-\lambda} \frac{\lambda^x}{x!} \\
= 0 + \sum_{x=1}^{\infty} x e^{-\lambda} \frac{\lambda^x}{x!} \\
= \sum_{x=1}^{\infty} x e^{-\lambda} \frac{\lambda^x}{x(x-1)!} \\
= \sum_{x=1}^{\infty} e^{-\lambda} \frac{\lambda^x}{(x-1)!}
\]

Multiply and divide by \( \lambda \)

\[
= \sum_{x=1}^{\infty} \frac{\lambda e^{-\lambda} \lambda^x}{\lambda (x-1)!} \\
= e^{-\lambda} \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\
= e^{-\lambda} \lambda \left[ e^\lambda - 1 \right] \\
= e^{-\lambda} e^\lambda \\
= 1
\]
Variance

\[ V(x) = \mathbb{E}(x^2) - \mathbb{E}(x)^2 \]

\[ E(x^2) = \sum_{x=0}^{\infty} x^2 p(x) \]

\[ = \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda} \lambda^x}{x!} \]

\[ = \sum_{x=0}^{\infty} x(x-1) + x \frac{e^{-\lambda} \lambda^x}{x!} \]

\[ = \sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda} \lambda^x}{x!} + \sum_{x=0}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!} \]

\[ = \sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda} \lambda^x}{x!} + \mathbb{E}(x) \]

\[ = \sum_{x=0}^{\infty} \frac{e^{-\lambda} \lambda^x}{(x-2)!} + \mathbb{E}(x) \]

Multiply divide by \( \lambda \)

\[ = \sum_{x=0}^{\infty} \frac{\lambda^x}{(x-2)!} + \frac{\lambda}{x!} \]

\[ = e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^x}{(x-2)!} + \frac{\lambda}{x!} \]

\[ = e^{-\lambda} \left[ 1 + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} \right] + \frac{\lambda}{x!} \]

\[ = e^{-\lambda} \lambda^2 e^\lambda + \frac{\lambda}{x!} \]

\[ E(x^2) = \mathbb{E}(x^2) = \mathbb{E}(x)^2 + \frac{\lambda}{x!} \]

\[ V(x) = \mathbb{E}(x^2) - \mathbb{E}(x)^2 \]

\[ = \frac{\lambda}{x!} \]
$V(x) = I$

- $\lambda = \exp(x'\beta)$

where $\beta$ is a $k \times 1$ vector of the parameters, and $x$ is a $k \times 1$ vector of regressors, including constant.

- Observation pairs $(x_i, y_i), i = 1, 2, \ldots, n$ are independently distributed.

The conditional probability function $f(y | \lambda)$ can be obtained by combining the assumptions (1) and (2). That is,

$$f(y | \lambda) = \frac{\exp(-\exp(x'\beta)) \exp(yx'\beta)}{y!} \quad y = 1, 2, \ldots$$

The Poisson distribution has only one parameter that simultaneously determines the conditional mean and variance. The Poisson regression model as defined by the above assumptions, which implies an exponential mean function,

$$E(y | \lambda) = I = \exp(x'\lambda)$$

and, the exponential conditional variance function $Var(y | \lambda) = I = \exp(x'\lambda)$.

It is true the conditional mean and conditional variance are equal in Poisson regression model.

### 4.3.1 Maximum likelihood estimation of the parameter

The regression coefficient $\beta$, is estimated by using the maximum likelihood method. The independent sample of $n$ pairs of observations $(x_i, y_i), i = 1, 2, \ldots, n$ is the joint probability distribution of the sample which is the product of the individual conditional probability distributions:
\[ f(y_1, \ldots, y_n \mid x_1, \ldots, x_n; \beta) = \prod_{i=1}^{n} (y_i \mid x_i; \beta) \quad (4.3.1) \]

The function of the parameters, (4.3.1) is called likelihood function, and it can be written as

\[ L = \ell (\beta, y_1, \ldots, y_n, x_1, \ldots, x_n) \]

\( \hat{\beta} \) is a M.L.E of \( \beta \) when it maximize the likelihood equation.

\[(i.e.) \beta = \max_{\beta} L(\beta, y_1, \ldots, y_n, x_1, \ldots, x_n) \]

The maximization of likelihood function is equivalent to maximization of the logarithmic, or log-likelihood function \( l = \log L \). The log-likelihood function of the Poisson regression model is

\[ l(\beta, y, x) = \sum_{i=1}^{n} \log f(y_i \mid x_i; \beta) \]

\[ = \sum_{i=1}^{n} \log \left( \exp(-\exp(x'_i \beta) \exp(y'_i \beta)) \right) \frac{1}{y'_i!} \]

\[ l(\beta, y, x) = \sum_{i=1}^{n} \exp(x'_i \beta + y'_i \beta - \log(y'_i!)) \]

The maximum likelihood estimates of the model parameters are obtained by the first derivative of the log-likelihood with respect to \( \beta \), the first order derivatives are setting equal to zero. In the Poisson regression model, there are \( k \) such derivatives with respect to \( \beta \), and so on. The vector that collects the \( k \) first order derivatives, which is alternatively denoted as gradient vector. The gradient vector can be written as,
\[ s_n(\beta, y, x) = \frac{\partial}{\partial \beta} \left[ \sum_{i=1}^{n} \exp(x_i' \beta + y_i x_i') \cdot \log(y_i) \right] = \sum_{i=1}^{n} -x_i' \exp(x_i' \beta + y_i x_i') \]

\[ = \sum_{i=1}^{n} -\exp(x_i' \beta + y_i x_i') \]

\[ s_n(\beta, y, x) = \sum_{i=1}^{n} -\exp(x_i' \beta) \]

The subscript \( n \) of \( s_n \) is used as a reminder that the score depends on the sample size.

The maximum likelihood estimator \( \hat{\beta} \) is the value of \( \beta \) that solves the first order conditions for a maximum. i.e.

\[ \begin{align*}
\sum_{i=1}^{n} y_i &= \exp(x_i' \hat{\beta}) \\
\sum_{i=1}^{n} y_i &= \sum_{i=1}^{n} \exp(x_i' \beta) \\
n\bar{y} &= \sum_{i=1}^{n} \exp(x_i' \beta) \\
\bar{y} &= \frac{\sum_{i=1}^{n} \exp(x_i' \beta)}{n}
\end{align*} \]

By Jensen's inequality

\[ \bar{x} \beta \leq \log \bar{y} \]

\[ \beta \leq \frac{\log \bar{y}}{\bar{x}} \]

The improved solution of \( \beta \) may be obtained by using the Newton Raphson iterative method.
The Newton Raphson method for the Poisson regression to obtain the $\beta$ is as,

$$\beta_{t+1} = \beta_t - \hat{g}_r \hat{g}_r$$

where $g = \frac{\partial \log L}{\partial \beta}$ and $H = \frac{\partial^2 \log L}{\partial \beta \partial \beta'}$, which are evaluated at $\beta_t$.

$$\beta_{t+1} = \beta_t + \left[ \sum_{i=1}^{n} \mu_i x_i \right] \sum_{i=1}^{n} x_i \phi_i - \mu$$

where $\mu = \exp(x' \beta)$

In addition that the matrix of second order derivatives is denoted as the Hessian matrix and is negative definite for all values of $\beta$, the solution to (4.3.1) is called the maximum likelihood estimator. The Hessian matrix of the log-likelihood function of the Poisson regression model is given by

$$H_n(\beta; y, x) = \frac{\partial^2 l(\beta; y, x)}{\partial \beta \partial \beta'}$$

$$H_n(\beta; y, x) = \frac{\partial^2 l(\beta; y, x)}{\partial \beta \partial \beta'} = \frac{\partial s_\beta y, x}{\partial \beta'}$$

$$= \frac{\partial}{\partial \beta} \left[ \sum_{i=1}^{n} y_i - \exp(x_i' \beta \phi_i) \right]$$

$$H_n(\beta; y, x) = \sum_{i=1}^{n} \exp(x_i' \beta \phi_i)$$

The Hessian matrix $H_n$ is negative definite, the log-likelihood function of the Poisson regression model is concave, and the set of parameters are solving by the first-order conditions which are the unique maximum likelihood estimators.
4.3.2 Likelihood Ratio Test

The likelihood ratio is used to test the validity of the null hypothesis. The null hypothesis can be stated as, \( H_0 : R\beta = I \)

Under \( H_0 \) the Likelihood ratio test is

\[
LR = -2 \left( -\hat{\chi} - \chi \right),
\]

Where \( \hat{\chi} \) denote the value of the restricted log-likelihood function, \( \chi \) denote the value of the unrestricted log-likelihood function, \( k \) denote the number of restrictions and \( \chi \) is a chi-squared distribution with \( k \) degrees of freedom. This test is based on the difference of two log-likelihood values and is named as the log likelihood-ratio test.

4.3.3 Goodness of fit measures

Goodness of fit measures are considered to the overall performance of the model. In general linear models (GLMs), the common goodness-of-fit measures are the Pearson and deviance statistic, which are weighted sums of residuals.

4.3.4 Pearson Statistic

Pearson statistic is a standard measure of goodness of fit for any model of \( y_i \) with mean \( \mu \) and variance \( \omega \). The Pearson statistic can be written as,

\[
P = \sum_{i=1}^{n} \frac{\hat{y}_i - \mu}{\omega}^2
\]

where \( \hat{\mu} \) and \( \hat{\omega} \) are estimates of \( \mu \) and \( \omega \). If the mean and variance are correctly specified then

\[
E \left[ \sum_{i=1}^{n} \frac{\hat{y}_i - \mu}{\omega}^2 \right] = 0.
\]

62
In practice $P$ is compared with $(n - 1)$ degrees of freedom correction due to estimation of $\mu$

The Pearson statistic for the Poisson regression model is obtained by setting $\omega = \mu$, so that

$$
P_p = \sum_{i=1}^{n} \frac{\hat{y}_i - \mu_i}{\mu_i^3}
$$

In the general linear model literature, if $P_p > 1$ is the evidence of overdispersion, that is, the true variance exceeds the mean, which implies

$$
E\left[ \sum_{i=1}^{n} \left( \frac{\hat{y}_i - \mu_i}{\mu_i} \right)^2 \right] > 1; \quad P_p < 1 \quad \text{indicates the underdispersion.}
$$

4.3.5 Deviance Statistic

The deviance statistic is another measure of goodness of fit to find the overall performance of the model. Let $L(\mu) = y L(\mu)$ denote the log-likelihood function, where $\mu$ is the $n \times 1$ vector with $i^{th}$ entry $\mu_i$, then the fitted log-likelihood is $L(\hat{\mu})$, and is the maximum log-likelihood achievable is $L(y)$, and it is in the full model with $n$ parameters, where $\hat{\mu}$ and $y$ are the $n \times 1$ vectors with $i^{th}$ entries $\hat{\mu}_i$ and $y_i$. The deviance is defined to be

$$
D(y, \mu) = -2 \log(L(y)) - \log(L(\hat{\mu}))
$$

it is twice the difference between the maximum log-likelihood achievable and the log-likelihood of the fitted model.
For the Poisson model the deviance statistic is

\[ D_p = \sum_{i=1}^{n} \left\{ y_i \ln \left( \frac{y_i}{\mu_i} \right) - y_i + \mu_i \right\}. \]

The sum of the right second of the equation is zero if an intercept is included and the exponential mean function is used. The Deviance can more easily be calculated as

\[ D_p = \sum_{i=1}^{n} y_i \ln \left( \frac{y_i}{\mu} \right) \]

The deviance measures the difference between the maximum log-likelihood achievable and the log-likelihood achieved by the model under scrutiny. The deviance has an approximate chi-square distribution with n-k degrees of freedom.

4.3.6 Wald Test

The Wald test is the testing of significance of particular independent variables in a statistical model. If the Wald test is significant for a particular independent variable or group of independent variables, then it can be concluded that the parameters are associated with the variables are not zero, so that the variables are included in the model. The independent variables can be omitted from the model if the Wald test is not significant. Wald test is the asymptotic distribution of the maximum likelihood estimator in the unrestricted model.

\[ \theta \sim N(\theta , \hat{\text{Var}}(\theta)) \]

where \( \hat{\theta} \) consists of the estimated regression coefficients plus any additional parameters, such as \( \sigma \), for any linear combination of the parameter vector

\[ R\theta - r \sim N(R\theta - r, R\text{Var}(\theta) R') \]

64
Under $H_0 : R\theta - q = 0$ and therefore

$$W = R\theta - \left( \begin{array}{c} \hat{\theta} \\ Var(\hat{\theta}) \end{array} \right) (R\theta - )$$

has a chi-squared distribution with degrees of freedom equal to the number of restrictions if the null hypothesis true. The Wald statistic simplifies to the squared “ $t$ ” statistic if the number of restrictions is one. Dividing the $w$ by the degrees of freedom produces an “ $F$ ”-statistic.

### 4.4 Negative Binomial Regression Model

Negative binomial regression is a regression for modelling count variables and it is usually used for overdispersed count outcome variables. The derivation of the Negative Binomial model as either a Poisson–gamma mixture model, or a member of the exponential family of distributions which serve as the basis of generalized linear models. In deriving the Negative Binomial as a general linear model, first the canonical negative binomial can be derived. The negative binomial model is created by converting the canonical link and inverse link in the general linear model derived Negative binomial canonical to log form.

### 4.5 Poisson–gamma mixture model

The negative binomial PDF is derived from the specification of a count outcome which is characterized by

$$f(y; \lambda, u) = \frac{e^{-\lambda u} (\lambda u)^y}{y!}$$

Where $u$ is exposure and $\lambda$ is parameter. Eqn.(4.5.1) can be thought of as a Poisson model with gamma heterogeneity where the gamma noise has a mean of 1. The gamma mixture accommodates overdispersed or correlated Poisson counts.
The distribution of $y$ is conditioned on $x$ and $u$ is Poisson with the conditioned mean and variance given by $\mu$ (4.5.1). The conditional mean of $y$ under gamma heterogeneity is thereby expressed as $\lambda$ rather than as only $\lambda$. The unconditional distribution of $y$ is derived from the following expression:

$$f(y; \lambda, u) = \int_0^\infty \frac{(-\lambda)^y_i}{y_i!} (\lambda)^{\gamma_i} \gamma_i \frac{\Gamma(\gamma_i)}{\Gamma(\gamma)} u_i^{\gamma_i} e^{-u_i} \partial_i$$

(4.5.2)

The unconditional distribution of $y$ is specified by $g(u)$. For this model a gamma distribution is given $u = \exp(\nu)$. Assigning a mean of 1 to the gamma distribution, we have

$$f(y; \lambda, u) = \int_0^\infty \frac{(-\lambda)^y_i}{y_i!} (\lambda)^{\gamma_i} \gamma_i \frac{\Gamma(\gamma_i)}{\Gamma(\gamma)} u_i^{\gamma_i} e^{-u_i} \partial_i$$

$$= \frac{\lambda^{\gamma_i}}{y!} \Gamma(\gamma) e^{-\lambda^{y_i}} \int_0^\infty e^{-u_i} u_i^{y_i} u_i^{\gamma_i - 1} \partial_i$$

$$= \frac{\lambda^{\gamma_i}}{y!} \Gamma(\gamma) e^{-\lambda^{y_i}} \int_0^\infty e^{-u_i} u_i^{y_i} u_i^{\gamma_i - 1} \partial_i$$

$$= \frac{\lambda^{\gamma_i}}{\lambda^{y_i} + \gamma_i} \Gamma(\gamma) e^{-\lambda^{y_i}} \int_0^\infty e^{-u_i} u_i^{y_i} u_i^{\gamma_i - 1} \partial_i$$

(4.5.3)

Where $y_i = y_i + \gamma_i$, the equation (4.5.3) can be derived further

$$f(y; \lambda, u) = \frac{\lambda^{\gamma_i}}{\lambda^{y_i} + \gamma_i} \Gamma(\gamma) e^{-\lambda^{y_i}} \int_0^\infty e^{-u_i} u_i^{y_i} u_i^{\gamma_i - 1} \partial_i$$

(4.5.4)
\[
\int_0^\infty e^{-\left(\lambda_i + \mu_i\right) y_i + 1} \partial_i = \frac{\Gamma(v_i + 1)}{(\lambda_i + 1)^{y_i + 1}}.
\]

Then the equation (4.5.4) can be continued as

\[
f(y; \lambda, \mu) = \frac{\Gamma(v_i + 1)}{\Gamma(v_i + 1) \Gamma(\gamma)} \frac{\lambda_i^{y_i}}{(\lambda_i + 1)^{y_i + 1}} \gamma^v \left(\frac{1}{\lambda_i + 1}\right)^{y_i + 1} \left(1 - \frac{\lambda_i}{\lambda_i + 1}\right)^{y_i + 1}.
\]

Inverting \( v \), the gamma scale parameter yields \( \alpha \) which is the negative binomial heterogeneity or overdispersion parameter. \( \lambda \) and \( \mu \) are equated and then recognized the resulting negative binomial probability mass function

\[
f(y; \lambda, u) = \frac{\Gamma(v_i + 1)}{\Gamma(v_i + 1) \Gamma(\gamma)} \gamma^v \left(\frac{1}{\lambda_i + 1}\right)^{y_i + 1} \left(1 - \frac{\lambda_i}{\lambda_i + 1}\right)^{y_i + 1}.
\]

Equation 4.5.8 is a commonly observed form of the negative binomial probability mass function.

Given \( y \), it follows that \( \Gamma(v + 1) = \lambda_i \), \( \Gamma(v + 1) = \gamma^v \left(\frac{1}{\alpha}\right)^{y_i + 1} \), and \( \Gamma(1) = 1 \).

\[
\frac{\Gamma(v_i + 1)}{\Gamma(y_i + 1)} \left(\lambda_i + \mu_i\right) = \frac{\Gamma(1)}{\Gamma(y_i + 1)} \left(\frac{1}{\alpha}\right)^{y_i + 1} \left(1 - \frac{1}{\alpha}\right)^{y_i + 1} = \frac{1}{\alpha}.
\]
The equation (4.5.7) can be expressed in the popular form of negative binomial probability distribution.

\[
f(y; \lambda, \alpha) = \binom{y_i + 1}{\alpha - 1} \left( \frac{\lambda}{\lambda + \alpha \mu} \right)^{\alpha - 1} \left( \frac{\lambda + \alpha \mu}{\lambda + \alpha \mu} \right)^{y_i} \tag{4.5.9}
\]

The mean and variance of the Negative Binomial is can written as

\[
E(y) = \sum_{y=0}^{\infty} \frac{y}{1 + \alpha \mu} \left( \frac{1}{\alpha} \right)^{y+1} \left( \frac{1}{1 + \alpha \mu} \right)^{1+y_i} \left( \frac{1}{1 + \alpha \mu} \right)^{y_i+1}
\]

\[
= \sum_{y=0}^{\infty} \frac{y}{1 + \alpha \mu} \left( \frac{1}{\alpha} \right)^{y+1} \left( \frac{1}{1 + \alpha \mu} \right)^{1+y_i} \left( \frac{1}{1 + \alpha \mu} \right)^{y_i+1}
\]

\[
= \sum_{y=0}^{\infty} \frac{y}{1 + \alpha \mu} \left( \frac{1}{\alpha} \right)^{y+1} \left( \frac{1}{1 + \alpha \mu} \right)^{1+y_i} \left( \frac{1}{1 + \alpha \mu} \right)^{y_i+1}
\]

\[
= \sum_{y=0}^{\infty} \frac{y}{1 + \alpha \mu} \left( \frac{1}{\alpha} \right)^{y+1} \left( \frac{1}{1 + \alpha \mu} \right)^{1+y_i} \left( \frac{1}{1 + \alpha \mu} \right)^{y_i+1}
\]

let \( s = \frac{1}{\alpha} \) and \( z = y - \) inside the summation

\[
= \mu \sum_{y=0}^{\infty} \frac{z + s - 1}{z!} \left( \frac{1}{\alpha} \right)^{s} \left( \frac{1}{1 + \alpha \mu} \right)^{z}
\]
\[
E(y) = \mu \sum_{s=0}^{\infty} \left( \frac{z + s - 1}{s - 1} \right) \left( \frac{1}{1 + \alpha \mu} \right)^s \left( 1 - \frac{1}{1 + \alpha \mu} \right)^{z - s} 
\]

\[
E(y^2) = \mu \sum_{s=0}^{\infty} y_i^2 \left( \frac{y_i + 1}{\alpha} \right) \left( \frac{1}{1 + \alpha \mu} \right)^y \left( 1 - \frac{1}{1 + \alpha \mu} \right) 
\]

\[
= \mu \sum_{s=0}^{\infty} y_i \left( \frac{y_i + 1}{\alpha} \right) \left( \frac{1}{1 + \alpha \mu} \right)^y \left( 1 - \frac{1}{1 + \alpha \mu} \right) 
\]

let \( s = r + \) and \( z = y - \) inside the summation

\[
= \left( 1 + \alpha \mu \right) \sum_{s=0}^{\infty} \frac{s}{z + s} \left( \frac{1}{1 + \alpha \mu} \right)^s \left( 1 - \frac{1}{1 + \alpha \mu} \right)^{z - s} 
\]

\[
= \left( 1 + \alpha \mu \right) \sum_{s=0}^{\infty} \frac{s}{z + s} \left( \frac{1}{1 + \alpha \mu} \right)^s \left( 1 - \frac{1}{1 + \alpha \mu} \right)^{z - s} 
\]

\[
= \left( 1 + \alpha \mu \right) \sum_{s=0}^{\infty} \frac{s}{z + s} \left( \frac{1}{1 + \alpha \mu} \right)^s \left( 1 - \frac{1}{1 + \alpha \mu} \right)^{z - s} 
\]
\[ E(y^2) = \mu + \lambda + \mu \]

\[ V(y) = \mathbb{E}(y^2) - \mathbb{E}(y)^2 = \mu + \lambda + \mu - \mu \]

\[ V(y) = \mu + \mu \]

4.5.10

The negative binomial regression is obtained by setting \( \mu = \exp(x' \beta) \) in the negative binomial distribution.

\[ \text{Parameter Estimation} \]

The maximum likelihood estimation method is used to obtain the parameters in this model. The likelihood function is defined as,

\[
l(\mu, y, \alpha) = \prod_{i=1}^{n} \left( \frac{y_i + \frac{1}{\alpha} - 1}{1 - \frac{1}{1 + \alpha \mu_i}} \right)^{y_i} \left( \frac{1}{1 + \alpha \mu_i} \right)^{\frac{1}{\alpha} - 1} \left( 1 - \frac{1}{1 + \alpha \mu_i} \right)^{y_i}
\]

\[
= \prod_{i=1}^{n} \left[ \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(y_i + 1) \Gamma(\frac{1}{\alpha})} \right] \left( \frac{1}{1 + \alpha \mu_i} \right)^{\frac{1}{\alpha} - 1} \left( \frac{\Gamma\left(y_i + \frac{1}{\alpha} \right)}{\Gamma\left(y_i + 1\right) \Gamma\left(\frac{1}{\alpha}\right)} \right)^{y_i} \left( 1 - \frac{1}{1 + \alpha \mu_i} \right)^{y_i}
\]

\[
= \prod_{i=1}^{n} \left[ \frac{\Gamma\left(y_i + \frac{1}{\alpha} \right)}{\Gamma\left(y_i + 1\right) \Gamma\left(\frac{1}{\alpha}\right)} \right] \left( \frac{1}{1 + \alpha \mu_i} \right)^{\frac{1}{\alpha} - 1} \left( \frac{1}{1 + \alpha \mu_i} \right)^{\alpha \mu_i} \left( 1 - \frac{1}{1 + \alpha \mu_i} \right)^{y_i}
\]
The log-likelihood is obtained by taking the natural log of both sides of the equation.

\[ L = \log L(\mu, \alpha) = \sum_{i=1}^{n} y_i \ln \left( \frac{\alpha \mu_i}{1 + \alpha \mu_i} \right) - \frac{1}{\alpha} \ln \left( + \frac{\alpha \mu_i}{1 + \alpha \mu_i} + \ln \Gamma \left( \frac{y_i + 1}{\alpha} \right) - \ln \Gamma \left( \frac{y_i}{\alpha} \right) + 1 - \ln \Gamma \left( \frac{1}{\alpha} \right) \right) \]

The negative binomial log-likelihood parameterized in terms of \( \beta \), the model coefficients can be expressed as:

\[ L(\beta, \gamma, \alpha) = \sum_{i=1}^{n} y_i \ln \left( \frac{\alpha \cdot cp(x' \beta)}{1 + \alpha \cdot cp(x' \beta)} \right) - \frac{1}{\alpha} \ln \left( + \frac{\alpha \cdot cp(x' \beta)}{1 + \alpha \cdot cp(x' \beta)} + \ln \Gamma \left( \frac{y_i + 1}{\alpha} \right) - \ln \Gamma \left( \frac{y_i}{\alpha} \right) + 1 - \ln \Gamma \left( \frac{1}{\alpha} \right) \right) \]

The maximum likelihood estimates of the model parameters are obtained by the first order derivative of the log-likelihood with respect to \( \beta \), the gradient and setting them equal to zero. The gradient is the vector of first order partial derivatives of the log-likelihood, and it is more accurate representation for estimating more than one parameter. The negative binomial heterogeneity parameter is also estimated by the derivative of the log-likelihood with respect to \( \alpha \) and setting them equal to zero.

With \( \mu = xp(x' \beta) \) the first order partial derivatives of the log-likelihood function with respect to \( \beta \) and \( \alpha \) are given in the following equations

**NB Gradient - \( \beta \)**

\[ \frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} \frac{x_i(y_i - \mu)}{1 + \alpha \mu_i} \]

Owing to the importance of the parameterization, the gradient in terms of \( \mu \)

\[ \frac{\partial L}{\partial \mu} = \frac{\partial}{\partial \mu} \left[ \sum_{i=1}^{n} y_i \ln \left( \frac{\alpha \mu_i}{1 + \alpha \mu_i} \right) - \frac{1}{\alpha} \ln \left( + \frac{\alpha \mu_i}{1 + \alpha \mu_i} + \ln \Gamma \left( \frac{y_i + 1}{\alpha} \right) - \ln \Gamma \left( \frac{y_i}{\alpha} \right) + 1 - \ln \Gamma \left( \frac{1}{\alpha} \right) \right) \right] \]

\[ = \sum_{i=1}^{n} \left[ \frac{y_i}{\alpha \mu_i} - \frac{y_i}{1 + \alpha \mu_i} \cdot \frac{1}{\alpha + \alpha \mu_i} \right] \]
\[
\begin{align*}
\sum_{i=1}^{n} \frac{y_i}{\mu_i + \alpha \mu} &= \sum_{i=1}^{n} \frac{\mu_i}{\mu_i + \alpha \mu} \\
\sum_{i=1}^{n} y_i &= \sum_{i=1}^{n} \frac{1}{\mu_i + \alpha \mu} \\
\sum_{i=1}^{n} y_i &= \sum_{i=1}^{n} \mu_i \\
\sum_{i=1}^{n} y_i &= \sum_{i=1}^{n} \mu_i \\
\exp(x' \beta) &= \exp(y_i \mu_i) \\
\beta &= \frac{\log y_i}{x'} \\
\end{align*}
\]

where \( \mu = \exp(x' \beta) \)

Taking log on both sides
\[
\begin{align*}
x' \beta &= \log y_i \\
\beta &= \frac{\log y_i}{x'} \\
\end{align*}
\]
the improved solution for $\beta$ may be obtained by using the Newton Raphson iterative method.

The Newton Raphson method for the Negative binomial regression, to obtain the $\beta$ is as,

$$\beta_{n+1} = \beta - \hat{I}_{r}^{-1} \hat{g}_{r}$$

where $g = \frac{\gamma \log L}{\partial \beta}$ and $H = \frac{\gamma^2 \log L}{\partial \beta \partial \beta}$ which are evaluated at $\beta_{n}$.

$$\beta_{n+1} = \beta + \frac{1}{n} \sum_{i=1}^{n} \frac{\mu_{ir}(1+\alpha)}{(1+\alpha \mu_{ir})^2} x_{i} x'_{i} \frac{x_{i}(y_{i} - \mu_{i})}{\hat{r}}$$

$$\beta_{n+1} = \beta + \left[ \frac{1}{n} \sum_{i=1}^{n} \frac{\mu_{ir}(1+\alpha)}{(1+\alpha \mu_{ir})^2} x_{i} x'_{i} \right]^{-1} \sum_{i=1}^{n} \frac{x_{i}(y_{i} - \mu_{i})}{\hat{r}}$$

where $\mu = x'p$.

NB Gradient – $\alpha$

$$\frac{\partial L}{\partial \alpha} = \sum_{i=1}^{n} \left[ \frac{1}{\alpha} \left( \ln(1+\alpha \mu_{i}) + \frac{\alpha}{1+\alpha \mu_{i}} \right) + \psi' \left( \frac{y_{i}}{\alpha} + 1 \right) - \psi' \left( \frac{1}{\alpha} \right) \right]$$

The estimator of $\alpha$ is also obtained by the method of moments by using the equation (4.3.10), which implies

$$V(y) = \mu + \eta \mu$$

$$E(y - \mu)^2 = \mu + \eta \mu$$

$$\frac{E(y_{i} - \mu_{i})^2}{\mu_{i}} = \gamma$$
\[\alpha = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i - \mu}{\mu}^2 - \frac{\mu}{\mu} \]  

\[\hat{\alpha} = \frac{1}{n-k} \sum_{i=1}^{n} \frac{(y_i - \mu)^2 - \mu}{\mu} \]  

The matrix of second order partial derivatives of the log-likelihood is commonly known as the Hessian. The observed information matrix is defined as the negative of the Hessian, evaluated at the maximum likelihood estimate.

The second order partial derivatives of the log-likelihood function with respect to \(\beta\) and \(\alpha\) are given in the following equations

**NB – Hessian – \(\beta\)**

\[\frac{\partial^2 L}{\partial \beta' \beta} = n \sum_{i=1}^{n} \frac{\mu (1 + \alpha)}{1 + \alpha \mu} \cdot \frac{x_i x_i'}{1 + \alpha \mu} \]

**NB – Hessian – \(\alpha\)**

\[\frac{\partial^2 L}{\partial \alpha} = n \sum_{i=1}^{n} \left[ 1 - \frac{1}{\alpha} \left( \frac{\alpha + 2\alpha \mu_i (y_i - \mu_i) - \alpha \mu_i 1 + \alpha \mu_i}{(1 + \alpha \mu_i)^2} + 2 \ln(1 + \alpha \mu_i) \right) + \psi(y_i + 1) - \psi(\frac{1}{\alpha}) \right] \]

The parameter \(\alpha\) and \(\beta\) can be obtained through the maximization.

**4.6 Goodness of fit statistic**

In general liner model, the overall performance of model can be obtained by the common goodness of fit measures such as Pearson and Deviance statistic are used.

**4.6.1 Pearson statistic and Deviance statistic**

Pearson statistic is a standard measure of goodness of fit for the Negative binomial regression model of \(y_i\) with mean \(\mu\) and variance \(\omega\). The Pearson statistic can be written as,
\[ P = \sum_{i=1}^{n} \frac{\hat{y}_i - \mu}{\omega} \]

where \( \hat{\mu} \) and \( \hat{\omega} \) are estimates of \( \mu \) and \( \omega \). The Pearson statistic of the Negative Binomial regression is,

\[ P_p = \sum_{i=1}^{n} \frac{\hat{y}_i - \mu}{\mu + \alpha \mu} \]

Where if \( \omega = \mu + \eta \mu \), if \( P_p > 1 \) is the evidence of overdispersion, that is, the true variance exceeds the mean, which implies

\[ E\left[ \sum_{i=1}^{n} \frac{\hat{y}_i - \mu}{\mu} \right] > 1; \quad P_p < 1 \] indicates the under dispersion.

Deviance is the another measure of goodness of fit to find the overall performance of the Negative binomial regression. The Deviance measure is defined as,

\[ D(y, \mu) = 2 \left( \log L(y) - \log L(y|\mu) \right) \]

it is twice the difference between the maximum log-likelihood achievable and the log-likelihood of the fitted model.

In Negative Binomial regression model additionally introduce a dispersion parameter \( \phi \), with variance scaled by \( a \phi \). Then the log-likelihood is \( L(\mu, \phi) \) and the scaled deviance is defined to be

\[ SD(y, \mu, \phi) = \left( \log L(y, \phi) - \log L(y, \mu) \right) \]

(4.6.1)
For general linear model densities $SD(y, \mu, \phi)$ equals a function of $y$ and $\mu$ divided by $a \phi$. It is convenient to multiply $SD(y, \mu, \phi)$ by the dispersion factor $a \phi$ and the deviance is defined as

$$D(y, \mu, \phi) = \ln(a(\phi)) E(y, \phi) - \ln(\mu, \phi).$$

(4.6.2)

In the linear regression model, the deviance equals the residual sum of squares $\sum_{i=1}^{n} y_i - \mu \sim \chi^2$ under normality. This has lead to the deviance being used in the generalised linear model framework as a generalization of the sum of squares. This provides the motivation for the deviance residual. The analysis of deviance generalizes the analysis of variance to compare sequences of nested general linear models. In Negative Binomial regression model the deviance statistic is calculated when $\alpha$ is known. It is can be written as

$$D = \sum_{i=1}^{n} \left\{ y_i \ln \left( \frac{y_i}{\mu} \right) - \alpha \left[ \frac{y_i + \alpha}{\mu + \alpha} \right] \right\}.$$

Which is approximated distributed as chi-square with n-k degrees of freedom.