Regression Analysis I

Simple Linear Regression

Hossein Moradi Rekabdarkolaee

South Dakota State University

email: *hossein.moradirekabdarkolaee@sdstate.edu*.

Office: Arch, Math, and Engineering Building, *Room: 254*.

Definition

- Regression analysis is a statistical methodology that utilizes the relation between two or more quantitative/qualitative variables so that a response or outcome variable can be predicted from the other(s).

- The assumption is that there is a functional relationship among variables and our goal is to mathematically model this relationship.

- **Applications**:
  - Business,
  - Medical Sciences,
  - social and behavioral sciences,
  - biological sciences.

Example

- Sales of a product can be predicted by utilizing the relationship between sales and amount of advertising expenditures.
- The performance of an employee on a job can be predicted by utilizing the relationship between performance and a battery of aptitude tests.
- The size of the vocabulary of a child can be predicted by utilizing the relationship between size of vocabulary and age of the child and amount of education of the parents.
- The length of hospital stay of a surgical patient can be predicted by utilizing the relationshipbetween the time in the hospital and the severity of the operation.

Definition

- The term regression comes from a Latin root meaning “going back”.
- This term was originally used by Sir Francis Galton (1822 - 1911)
- **Types of Regression (model):**
  - Functional versus Statistical relationship,
  - Classification depending on the number of dependent and independent variables
    - **Simple regression:** One dependent and One independent.
    - **Multiple regression:** One dependent and two or more independent variables
    - **Multivariate regression:** Two or more dependent variables
  - Depending on the type of relationship (linear versus non-linear).
  - Depending on the nature of the dependent variable.

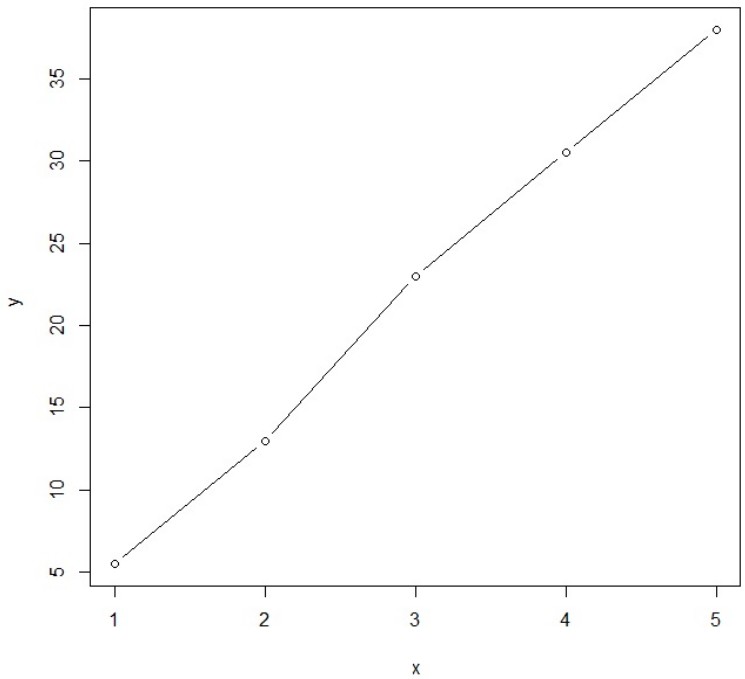
Relations between Variables

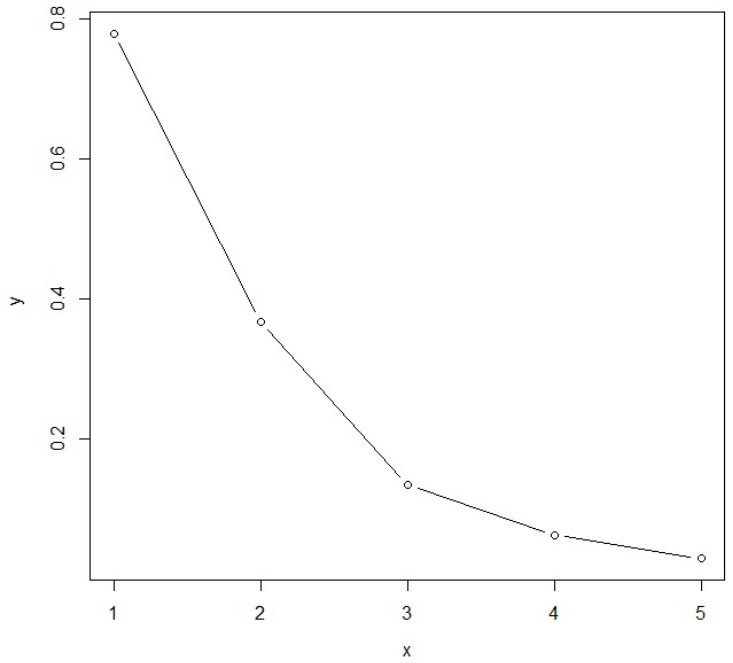
- **Functional Relation between Two Variables**:
  - A functional relation between two variables is expressed by a mathematical formula.
  - **example**: Consider **X** as independent variable and Y as dependent variable

$$\begin{aligned} Y &= 5 + 3\mathbf{X} \\ Y &= \exp\{-\mathbf{X}/2\} \end{aligned}$$

- It is a **perfect** relationship.
- Given a particular value of **X**, the function f indicates the corresponding value of Y.

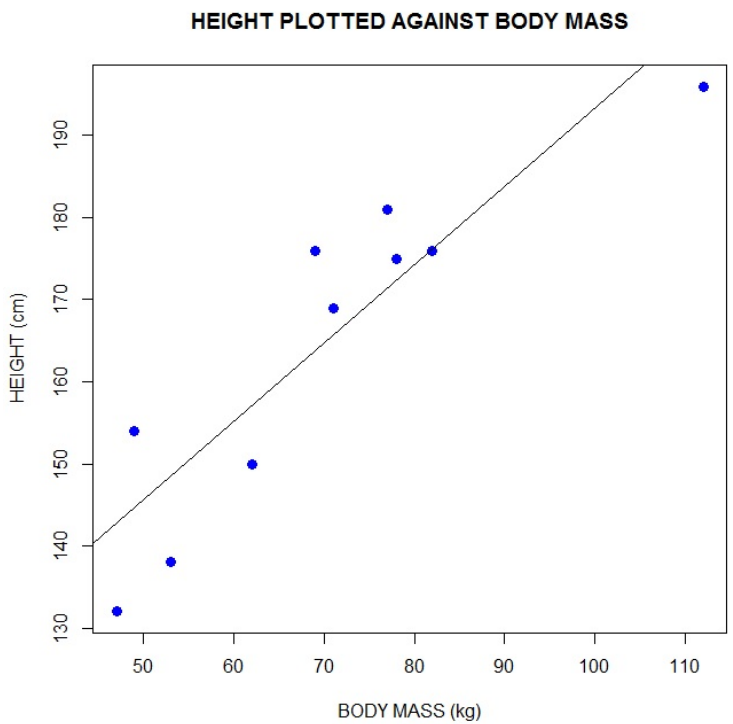
Functional Relation between Two Variables

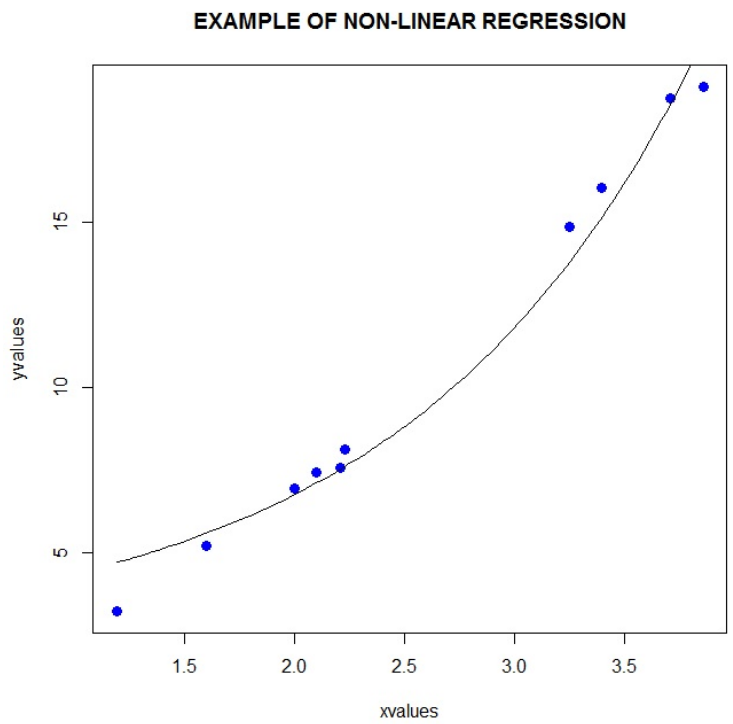




Relations between Variables

- **Statistical Relation between Two Variables**:
  - A statistical relation is not a perfect relation. In general, the observations for a statistical relation do not fall directly on-the curve of relationship.





Usage of Regression Analysis

Regression analysis is used for following:

- Model specification.
- Parameter estimation.
- Variable screening.
- Prediction.
- Model evaluation.

Basic Concepts

- A regression model is a formal means of expressing the two essential ingredients of a statistical relation:
  1. A tendency of the response variable Y to vary with the predictor variable X in a systematic fashion.
  2. A scattering of points around the curve of statistical relationship.
- These two characteristics are embodied in a regression model by postulating that:
  1. There is a probability distribution of Y for each level of X.
  2. The means of these probability distributions vary in some systematic fashion with X.

Basic Concepts

- Regression with one independent variable is called **simple regression**.
- Regression with more than one independent variable is called **multiple regression**.
- Different assumption can be made for the **functional form of the relationship**.
  - Functional relationship can be **linear** based on the regression coefficients,
  - or **non-linear** based on the regression coefficients.

Basic Concepts

Linear Model

$$\begin{aligned} Y &= \beta_0 + \beta_1 \mathbf{X}, \\ Y &= \beta_0 + \beta_1 \mathbf{X}^2, \\ Y &= \beta_0 + \beta_1 \log(\mathbf{X}). \end{aligned}$$

Non-linear Model

$$\begin{aligned} Y &= \beta_0 + \beta_1^2 \mathbf{X}, \\ Y &= \beta_0 + \log(\beta_1) \mathbf{X}^2, \\ Y &= \beta_0 * \beta_1 \log(\mathbf{X}). \end{aligned}$$

Model and Assumption for Simple Linear Regression

- Suppose, X denotes the predictor (independent, covariate) variable and y shows the response (dependent) variable.
- Let $i = 1, 2, \ldots, n$ denotes samples.
- We assume:
  1. x_i are nonrandom and observed with negligible error (hence can be treated as constant).
  2. The model is linear in terms of the coefficients i.e. it has the form

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where β_0 is the intercept, β_1 is the slope, and ϵ is the error (residual).

Model and Assumption for Simple Linear Regression

- Assumptions for the error
  1. It is independent from the independent variable.
  2. ϵ_i are uncorrelated random variables,
  3. ϵ_i have mean 0,
  4. ϵ_i have homogeneous variance
- In mathematical terms

$$\begin{aligned} E(\epsilon_i) &= 0, \\ Var(\epsilon_i) &= E(\epsilon_i^2) = \sigma^2. \end{aligned}$$

Model and Assumption for Simple Linear Regression

- At each level of X, y is a random variable with a mean and variance, and distributions.
- Mean of response

$$\begin{aligned} E(y_i) &= E(\beta_0 + \beta_1 x_i + \epsilon_i) \\ &= \beta_0 + \beta_1 E(x_i) + E(\epsilon_i) \\ &= \beta_0 + \beta_1 x_i. \end{aligned}$$

- Variance of response

$$\begin{aligned} Var(y_i) &= Var(\beta_0 + \beta_1 x_i + \epsilon_i) \\ &= Var(\beta_0) + Var(\beta_1 x_i) + Var(\epsilon_i) \\ &= Var(\epsilon_i) = \sigma^2. \end{aligned}$$

Model and Assumption for Simple Linear Regression

- Saying that ϵ_i are uncorrelated implies that the observations are independent of each other.
- This assumption can only be checked by thinking about the problem.
- Although it is not necessary to assume a specific distribution, we often assume that the ϵ_i (and hence the y_i) are normally distributed i.e.

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

Therefore,

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

Meaning of Regression Parameters

- The parameters β_0 and β_1, in regression model are called regression coefficients.
- β_1 is the slope of the regression line.
  - It indicates the change in the mean of the probability distribution of Y per unit increase in X.
- β_0 is the Y intercept of the regression model.
  - When the scope of the model includes X = 0, β_0 gives the mean of the probability distribution of Y at X = 0.
  - When the scope of the model does not cover X = 0, β_0 does not have any particular meaning as a separate term in the regression model.

Estimation of Regression Function

We need to estimate the unknown parameters β_0 and β_1 of the regression model and use these estimates to estimate the mean response.

- Least Squares Method.

- Maximum Likelihood Estimation.

Least Squares Method

- Let partition the total variability in the response, i.e.

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

- Let $\hat{y}_i$ denote the predicted values.

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&\Longrightarrow \\
SS_{TOTAL} &= SS_{MODEL} + SS_{ERROR}
\end{aligned}$$

Least Squares Method

- $e_i = y_i - \hat{y}_i$ is called the i^{th} residual.
- Since we want a good prediction, the residual must be small.
- In other word, we want to minimize the error.
- Since $E(\epsilon_i) = 0$, therefore $\sum_{i=1}^{n} \epsilon_i = 0$. (No minimization here, sorry)
- What should we do??

Least Squares Method

- Ordinary Least Squares (OLS) method seeks to find the estimate for β_0 and β_1 such that minimize the $\sum_{i=1}^{n} e_i^2$.
- In other words, OLS attempts to minimize the sum of squared residuals.

$$
\begin{aligned}
min_{\beta_0,\beta_1} \sum_{i=1}^{n} e_i^2 &= min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2
\end{aligned}
$$

Least Squares Method

- Take the partial derivative of the objective function with respect to β_0 and β_1.
- Set the partial derivatives equal to zero.
- Solve for β_0 and β_1.

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

and

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

Least Squares Method

- Therefore, the OLS estimates for the regression coefficients are:

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

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- Then the prediction equation is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Least Squares Method

- Let

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

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

$$S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}).$$

- Then

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

Least Squares Method

- The sum of squares statistics can also be written as

$$SS_{TOTAL} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = S_{yy},$$

$$SS_{MODEL} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 S_{xx} = \hat{\beta}_1 S_{xy},$$

and

$$SS_{ERROR} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = S_{yy} - \hat{\beta}_1 S_{xy}.$$

Least Squares Method

- The associated with each sum of squares is a degree of freedom (df).

$$df_{TOTAL} = n - 1,$$

$$df_{MODEL} = df_{REG} = \text{number of X variables in the model},$$

and

$$df_{ERROR} = n - \text{number of parameter in the model}.$$

Least Squares Method

- Given the sum of squares and degrees of freedom, we can then compute the mean squares to be the ration of the sum of squares divided by the sum of squares divided by the degrees of freedom.

$$MS_{MODEL} = MS_{REG} = \frac{SS_{MODEL}}{df_{MODEL}},$$

and

$$MS_{ERROR} = MS_{RES} = \frac{SS_{ERROR}}{df_{ERROR}}.$$

- A third parameter to estimate is the assumed to be constant variance σ^2.
- An unbiased estimate of σ^2 is $\hat{\sigma}^2 = MS_{ERROR}$.

Least Squares Method

- Properties of Least Squares Estimators:
- Under the conditions of regression model, the OLS estimators β_0 and β_1 are unbiased, i.e.

$$E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad E(\hat{\beta}_1) = \beta_1.$$

- In other words, the OLS estimator tends to neither overestimate or underestimate systematically.
- Under the conditions of regression model, the OLS estimators have minimum variance among all unbiased linear estimators.

Least Squares Method

- **Properties of Fitted Regression line**:
  - The sum of the residuals is zero:

$$\sum_{i=1}^{n} e_i = 0$$

  - The sum of the squared residuals, $\sum_{i=1}^{n} e_i^2$, is a minimum.
  - The sum of the observed values y_i equals the sum of the fitted values $\hat{y}_i$:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i.$$

Least Squares Method

- **Properties of Fitted Regression line**:
  - The sum of the weighted residuals is zero when the residual in the i^{th} trial is weighted by the level of the predictor variable in the i^{th} trial:
$$\sum_{i=1}^{n} x_i e_i = 0$$
  - The sum of the weighted residuals is zero when the residual in the i^{th} trial is weighted by the fitted value of the response variable for the i^{th} trial:
$$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$
  - The regression line always goes through the point $(\bar{x}, \bar{y})$.

Least Squares Method

- The simple linear regression model can be expressed in several forms.
- One is so-called centered model.

$$y_i = \beta_0^* + \beta_1(x_i - \bar{x}) + \epsilon_i,$$

- This creates a new axes that instead of being center at $(0, 0)$ are now centered at $(\bar{x}, \bar{y})$.
- The estimator for β_1 is exactly the same.
- The estimator of β_0^* is $\hat{\beta}_0^* = \bar{y}$.

Inferences in Regression and Correlation Analysis

- A measure of the quality of the fit (prediction) of our estimated model is the coefficient of determination, R^2.

$$R^2 = \frac{SS_{MODEL}}{SS_{TOTAL}} = 1 - \frac{SS_{ERROR}}{SS_{TOTAL}} = \frac{b_1 S_{xy}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}}.$$

- R^2 ranges from 0 to $+1$ and measures the fraction of the variability in the y data explained by the regression model(by x variable).
- **Interpretation**:
  - If $R^2 = 1$, then all points fall on a straight line.
  - $R^2 = 0$, then either the points are highly scattered or have nonlinear pattern.
  - This indicates that the linear regression model is poor.

Inferences in Regression and Correlation Analysis

- The closer R^2 is to $+1$, the better the prediction capability of the model.
- The relationship between x and y can be explained by the Pearson correlation coefficient.
- Let ρ denotes the Pearson correlation between x and y.

$$\rho = \frac{cov(x, y)}{\sqrt{var(x)var(y)}}$$

- ρ is between -1 to +1.
- The closer $|\rho|$ to +1, the stronger the relationship between x and y.

- For simple linear regression, ρ can be estimated as

$$\hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = (sign\ of\ \hat{\beta}_1)\sqrt{R^2}.$$

Inferences in Regression and Correlation Analysis

- To test for a significance correlation between x and y, we test $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$.
- The test statistic is

$$t_{obs} = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}},$$

- Under the standard assumptions, the test statistic follows a t-distribution with n-2 degrees of freedom.

Inferences in Regression and Correlation Analysis

- Under the assumption that $E(\epsilon_i) = 0$, the sampling distribution of the β_1 is

$$\begin{aligned} E(\hat{\beta}_1) &= \beta_1, \\ Var(\hat{\beta}_1) &= \frac{\sigma^2}{S_{xx}}. \end{aligned}$$

- By replacing the σ^2 with its unbiased estimator, we have an estimate for the $Var(\hat{\beta}_1)$.
- Then, we can test the $\beta_1 = 0$.
- The reason for interest in testing whether or not $\beta_1 = 0$ is that, when $\beta_1 = 0$, there is no linear association between y and x.

Inferences in Regression and Correlation Analysis

- If we assume that the errors have a normal distribution, then

$$\beta_1 \sim N\left(\hat{\beta}_1, \frac{\sigma^2}{S_{xx}}\right).$$

- Then by a Z-score transformation, we have

$$Z = \frac{\beta_1 - \hat{\beta}_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}}.$$

- Since σ^2 is unknown, we use its unbiased estimate, which is *MSE*.
- Then the statistic follow a t distribution with degree of freedom $n-2$, i.e.

$$t = \frac{\beta_1 - \hat{\beta}_1}{\sqrt{\frac{MSE}{S_{xx}}}}.$$

- Since $\frac{\beta_1 - \hat{\beta}_1}{\sqrt{\frac{MSE}{S_{xx}}}}$ follows a t distribution, we can make the following probability statement:

$$P\left\{ t(\alpha/2, n-2) \leq \frac{\beta_1 - \hat{\beta}_1}{\sqrt{\frac{MSE}{S_{xx}}}} \leq t(1-\alpha/2, n-2) \right\} = 1 - \alpha.$$

- Since t-distribution is a symmetric distribution, we have

$$t(1-\alpha/2, n-2) = -t(\alpha/2, n-2).$$

- Therefore, the confidence interval for β_1:

$$\left(\hat{\beta}_1 - t(\alpha/2, n-2)\sqrt{\frac{MSE}{S_{xx}}}, \hat{\beta}_1 + t(\alpha/2, n-2)\sqrt{\frac{MSE}{S_{xx}}} \right)$$

Inferences in Regression and Correlation Analysis

- Under the assumption that $E(\epsilon_i) = 0$, the sampling distribution of the β_0 is

$$\begin{aligned} E(\hat{\beta}_0) &= \beta_0, \\ Var(\hat{\beta}_0) &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right). \end{aligned}$$

- By replacing the σ^2 with its unbiased estimator, we have an estimate for the $Var(\hat{\beta}_0)$.
- Then, we can test the $\beta_0 = 0$.

Inferences in Regression and Correlation Analysis

- If we assume that the errors have a normal distribution, then

$$\beta_0 \sim N\left(\hat{\beta}_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right).$$

- Then by a Z-score transformation, we have

$$Z = \frac{\beta_0 - \hat{\beta}_0}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}.$$

- Since σ^2 is unknown, we use its unbiased estimate, which is MSE.
- Then the statistic follow a t distribution with degree of freedom $n-2$, i.e.

$$t = \frac{\beta_0 - \hat{\beta}_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}.$$

Inferences in Regression and Correlation Analysis

- Since $\frac{\beta_0 - \hat{\beta}_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$ follows a t distribution, we can make the following probability statement:

$$P\left\{ t(\alpha/2, n-2) \leq \frac{\beta_0 - \hat{\beta}_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \leq t(1-\alpha/2, n-2) \right\} = 1-\alpha.$$

- Since t-distribution is a symmetric distribution, we have

$$t(1-\alpha/2, n-2) = -t(\alpha/2, n-2).$$

- Therefore, the confidence interval for β_0:

$$\left(\hat{\beta}_0 - t(\alpha/2, n-2)\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}, \hat{\beta}_0 + t(\alpha/2, n-2)\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} \right)$$

- Under the assumption that $E(\epsilon_i) = 0$, the sampling distribution of the mean of the predicted response, $\hat{y}_i$ is

$$\begin{aligned} E(\hat{y}_i) &= y_i, \\ Var(\hat{y}_i) &= \sigma^2 \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right). \end{aligned}$$

- By replacing the σ^2 with its unbiased estimator, we have an estimate for the $Var(\hat{y}_i)$.
- Then, we can test the provide an interval estimate for y_i, which is

$$\left(\hat{y}_i - t_{\alpha/2,n-2} \sqrt{MSE \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right)}, \hat{y}_i + t_{\alpha/2,n-2} \sqrt{MSE \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right)} \right)$$

Inferences in Regression and Correlation Analysis

- Under the assumption that $E(\epsilon_i) = 0$, the sampling distribution of the mean of the prediction of new observation, $\hat{y}_{i,new}$ is

$$\begin{aligned} E(\hat{y}_{i,new}) &= y_{i,new}, \\ Var(\hat{y}_{i,new}) &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{i,new} - \bar{x})^2}{S_{xx}}\right). \end{aligned}$$

- By replacing the σ^2 with its unbiased estimator, we have an estimate for the $Var(\hat{y}_{i,new})$.
- Then, we can test the provide an interval estimate for $y_{i,new}$, which is

$$\left(\hat{y}_{i,new} - t_{\alpha/2,n-2}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_{i,new} - \bar{x})^2}{S_{xx}}\right)}, \hat{y}_{i,new} + t_{\alpha/2,n-2}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_{i,new} - \bar{x})^2}{S_{xx}}\right)}\right)$$

Inferences in Regression and Correlation Analysis

- Under the assumption that $E(\epsilon_i) = 0$, the sampling distribution of the prediction of the mean of m new observations for given $\hat{x}_i$ is

$$\begin{aligned} E(\hat{y}_i) &= y_i, \\ Var(\hat{y}_i) &= \sigma^2 \left(\frac{1}{m} + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right). \end{aligned}$$

- By replacing the σ^2 with its unbiased estimator, we have an estimate for the $Var(\hat{y}_i)$.

- Then, we can test the provide an interval estimate for the prediction of the mean of m new observations for given $\hat{x}_i$, which is

$$\left(\hat{y}_i - t_{\alpha/2,n-2} \sqrt{MSE \left(\frac{1}{m} + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right)}, \hat{y}_i + t_{\alpha/2,n-2} \sqrt{MSE \left(\frac{1}{m} + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right)} \right)$$

Analysis of Variance Approach to Regression Analysis

- The partitioning of the total variability and degrees of freedom can be displayed in an analysis of variance (ANOVA) table.
- An ANOVA table displays the sources of variation, the sum of squares, the degrees of freedom, and the mean squares.

Table: analysis of variance table for regression.

Source	df	Sum of Squares (SS)	Mean of SS
Model	1	$\hat{\beta}_1 S_{xy}$	SS_{MODEL}
Error	n-2	$S_{yy} - \hat{\beta}_1 S_{xy}$	$\frac{SS_{ERROR}}{n-2}$
Total	n-1	S_{yy}	

Analysis of Variance Approach to Regression Analysis

- The ANOVA table can be used to test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.
- The test statistics is

$$F_{obs} = \frac{MS_{MODEL}}{MS_{ERROR}} \sim F_{1,n-2},$$

- For a two-sided test, this test statistic is based on a F distribution.
- This test statistic is equivalent to the t-test statistic given earlier because $F_{1,n-2} = t^2_{n-2}$.
- Therefore, the F statistics can be also used to determine if x belongs in the model.

Maximum Likelihood Estimation

- We assumed y_i are normally distributed, i.e.

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

- The likelihood function is

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right\}.$$

- We can find the ML estimates for β_0 and β_1.

Departures from Model Based on Residuals

- The regression function is not linear.
- The error terms do not have constant variance.
- The error terms are not independent.
- The model fits all but one or a few outlier observations.
- The error terms are not normally distributed.

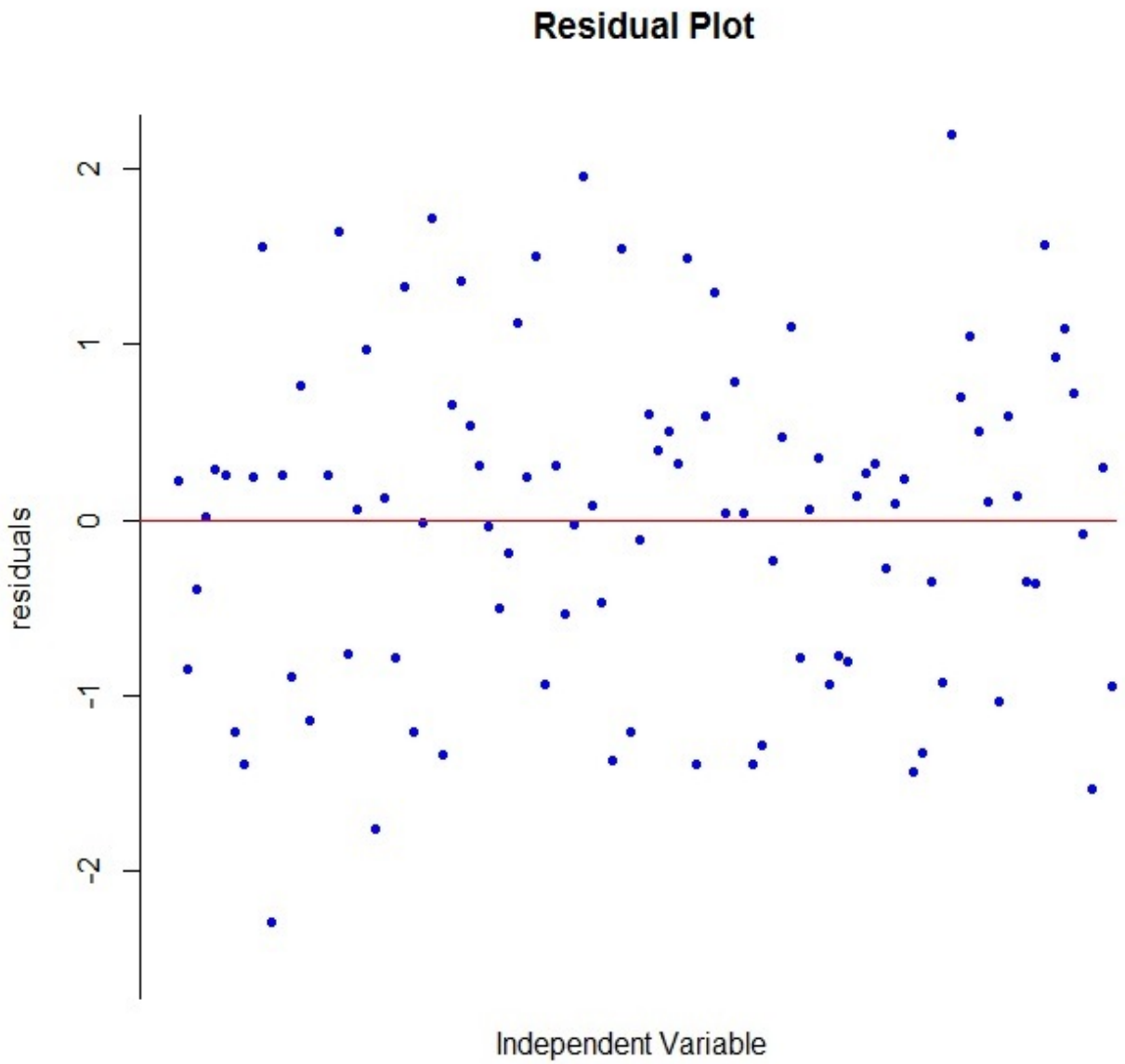
Diagnostics and Remedial Measures

- Recall that residual is the difference between an observed response and the predicted response at the given value of covariate: $e_i = y_i - \hat{y}_i$.
- An analysis of the residual can be used to inspect the adequacy of the model.
- In addition, we can check for any possible violations of our assumptions.
- One way to do this is to inspect a plot of residuals versus the independent variable values.

Diagnostics and Remedial Measures

- One way to do this is to inspect a plot of residuals versus the independent variable values.
- As mentioned earlier, the OLS procedure guarantees that $\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$.
- Hence, if the model is appropriate, a plot of the residuals versus x_i should be centered at 0, and scattered about the line.
- The absolute size of the residuals should be small depending on the size of the residuals when compared to the observed response data.

Diagnostics and Remedial Measures

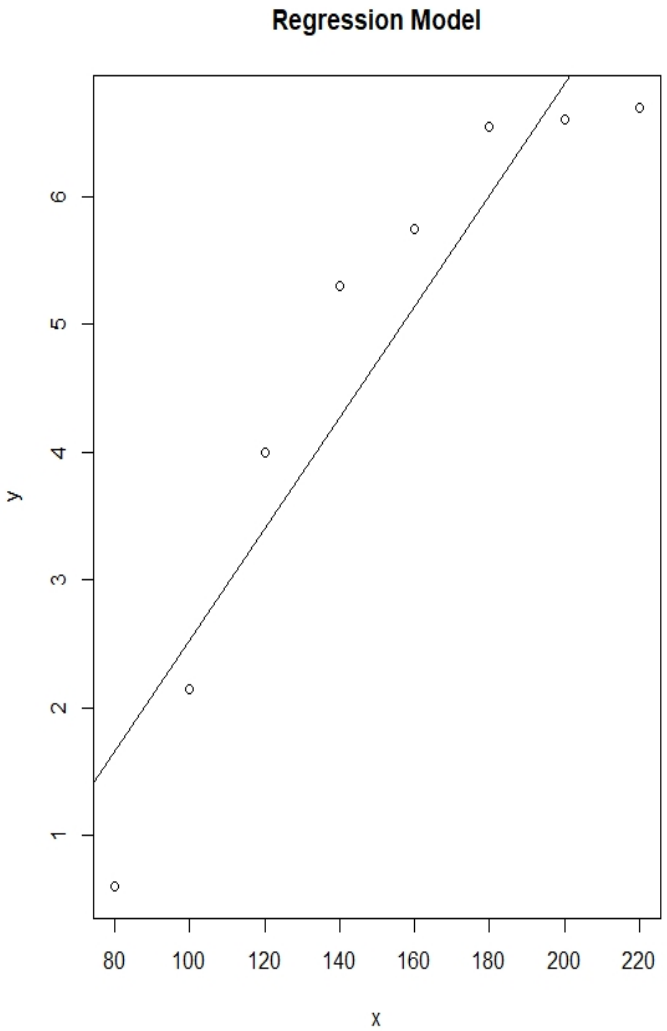


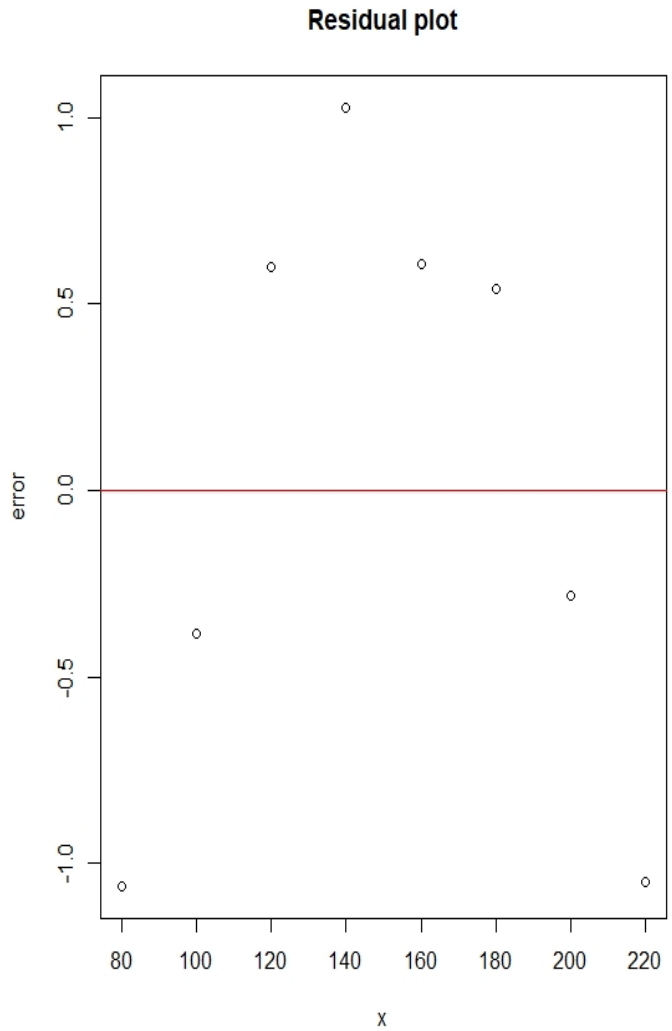
Diagnostics and Remedial Measures

- Plot of residuals against predictor variable.
- Plot of residuals against fitted values.
- Box plot of residuals.
- Normal probability plot of residuals

Diagnostics and Remedial Measures

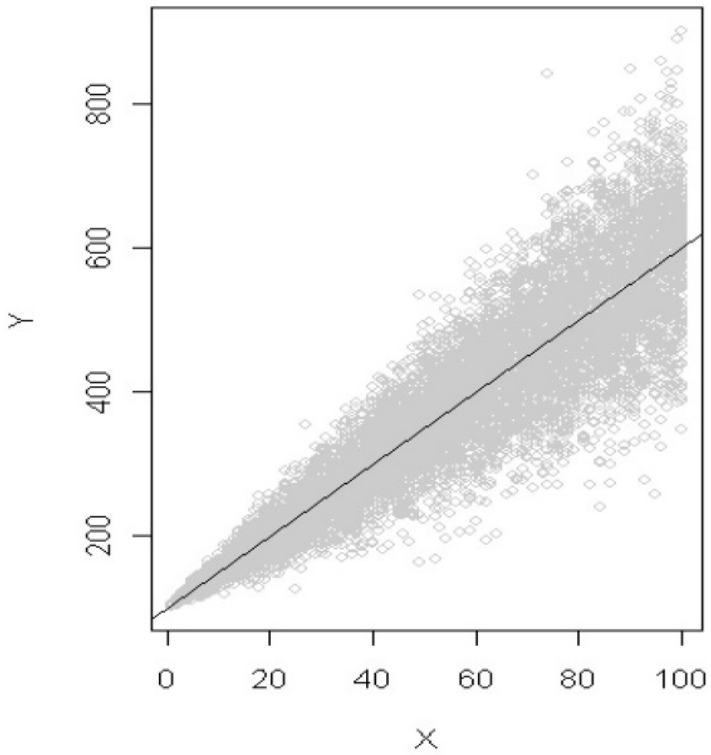
- Study of the relation between maps distributed and bus ridership in eight test cities.

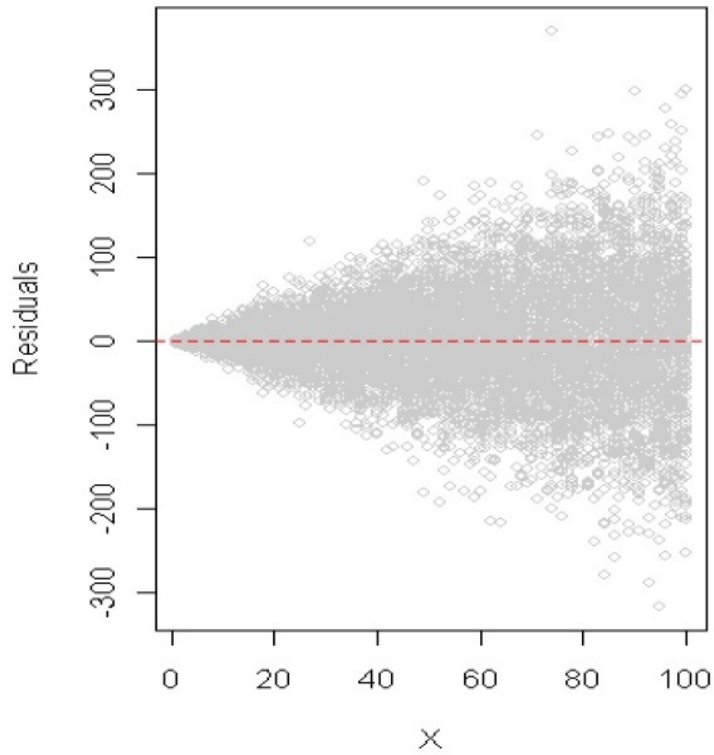




Diagnostics and Remedial Measures

- The plot of residuals against a predictor variable or against the fitted values can also be used to check the constancy of variance assumption.





$$y_i \sim N(100 + 5x, i^2); \quad i = 1, 2, \ldots, 100.$$

Tests for Constancy of Error Variance (Brown-Forsythe Test)

- The Brown-Forsythe test, a modification of the Levene test.
- This test is robust against serious departures from normality.
- The test is based on the variability of the residuals.
- To conduct the Brown-Forsythe test:
  - Divide the data set into two groups, according to the level of X.
  - One group consists of cases where the X level is comparatively low and the other group consists of cases where the X level is comparatively high.
  - Calculate the absolute deviations of residuals around their group median.
  - Run a two-sample t test.
  - Make a decision.

Tests for Constancy of Error Variance (Brown-Forsythe Test)

- Divide the sample to two groups

$$n = n_1 + n_2.$$

- Calculate the absolute deviations of residuals around their group median:

$$d_{i,1} = |e_{i,1} - \tilde{e}_1|, \quad d_{i,2} = |e_{i,2} - \tilde{e}_2|.$$

- Run a two-sample t test:

$$t^*_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n-2},$$

where

$$s^2 = \frac{\sum(d_{i,1} - \bar{d}_1)^2 + \sum(d_{i,2} - \bar{d}_2)^2}{n-2}.$$

Tests for Constancy of Error Variance (Breusch-Pagan Test)

- This test, a large-sample test, assumes that the error terms are independent and normally distributed.
- In addition,

$$ln(\sigma_i^2) = \gamma_0 + \gamma_1 X_i,$$

- Constancy of error variance corresponds to $\gamma_1 = 0$.
- To test $\gamma_1 = 0$ vs. $\gamma_1 \neq 0$, the test statistic is:

$$\chi^*_{BP} = \frac{\frac{SSR^*}{2}}{\left(\frac{SSE}{n}\right)^2} \sim \chi^2_1$$

- where SSR^* is the regression sum of squares when regressing e^2 on X and SSE is the error sum of squares when regressing Y on X.

Simultaneous Inferences

- Simultaneous statistical inferences are concerned with making inferences for several parameters or hypotheses simultaneously.
- Examples of families of hypotheses (or parameters)
  - All pair-wise comparisons in one-way ANOVA model
  - Comparisons with a control in a one-way ANOVA model
  - Regression coefficients and their functions (e.g., mean response)
- Two ways of making inferences about family of hypotheses:

  (i) Multiple testing (MT) - Controls the Family Wise Error Rate (FWER): is the probability of making one or more false discoveries, or type I errors when performing multiple hypotheses tests.

  (ii) Simultaneous confidence intervals (SCI) – Controls the family-wise confidence coefficient (simultaneous coverage probability).

Simultaneous Inferences (MCP)

- Multiple comparison procedures (MCP)
- **Bonferroni MCP**
  - Appropriate for any fixed number (g) of pre-planned comparisons
  - General procedure
  - Easy to calculate critical points (or the adjusted p-values)
  - Conservative
- **Tukey MCP**
  - Appropriate for all pair-wise comparisons For 4 treatments (A, B, C, and D), all pairs: AB, AC, AD, BC, BD, CD
- **Dunnett MCP**
  - Appropriate for comparisons with a control
  - Critical point and p-values calculated from multivariate t-distribution
- **Scheffe**
  - Contains infinitely many contrasts

Simultaneous Inferences (The Bonferroni MCP)

- Based on Bonferroni inequality (Boole's inequality).
- Let $A_1, A_2, \ldots, A_n$ be any n events, then

$$p(A_1 \cap A_2 \cap \ldots \cap A_n) = p\left(\cap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} p(A_i) - n + 1.$$

Proof???

Simultaneous Inferences (The Bonferroni MCP)

- Bonferroni simultaneous confidence intervals for β_0 and β_1:

$$\left(\hat{\beta}_1 - t(\alpha/4, n-2)\sqrt{\frac{MSE}{S_{xx}}}, \hat{\beta}_1 + t(\alpha/4, n-2)\sqrt{\frac{MSE}{S_{xx}}}\right),$$

$$\left(\hat{\beta}_0 - t(\alpha/4, n-2)\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}, \hat{\beta}_0 + t(\alpha/4, n-2)\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}\right)$$

- Define:

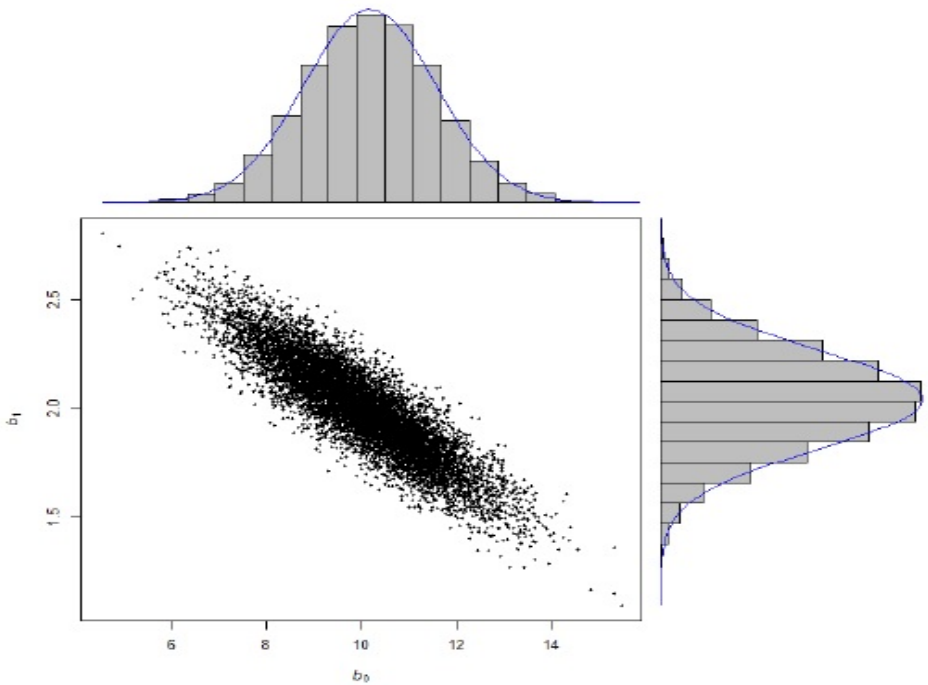
  A_1 to be the event the CI for β_0,
  A_2 to be the event the CI for β_1.

- Then, we have $p(A_1) = 1 - \alpha^*$ and $p(A_2) = 1 - \alpha^*$.
- Thus, $p(A_1 \cap A_2)$ is the probability that the two confidence intervals simultaneously cover their respective parameters.
- This can be adjusted to get a desirable joint coverage.

Simultaneous Inferences (The Bonferroni MCP)

- Example: For Toluca data, construct a 95% Bonferroni SCI for β_0 and β_1:
- Construct each confidence interval at 1- 0.05/2 = 0.975 (97.5%)

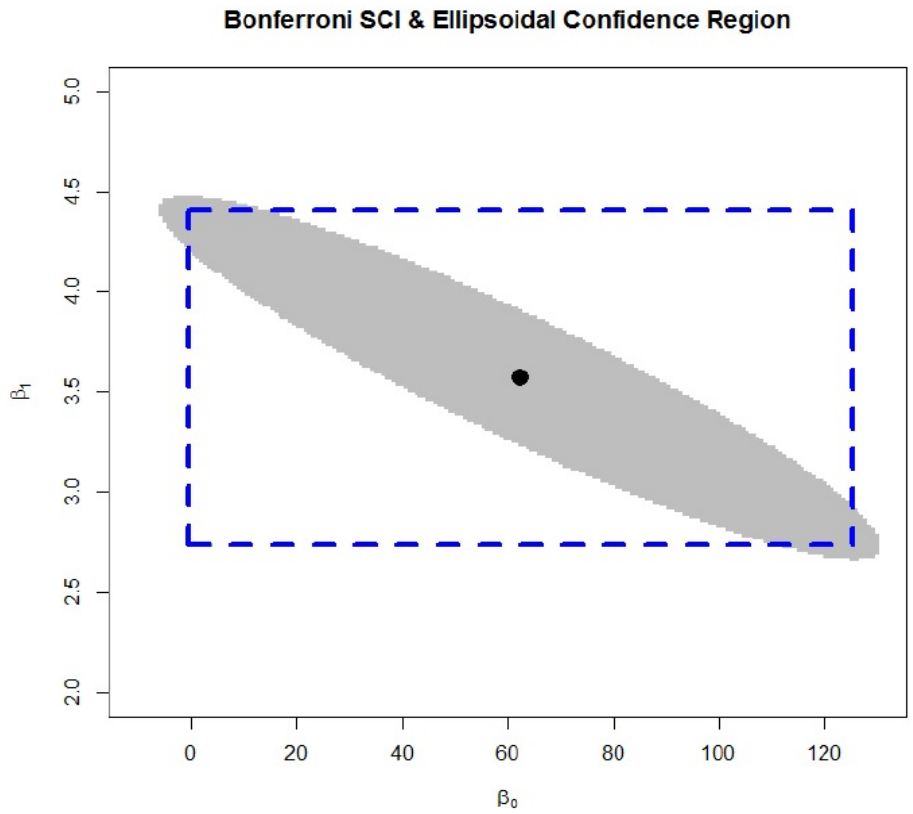
$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \sim N\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \begin{bmatrix} \sigma^2_{b_0} & -\overline{X}\sigma^2_{b_1} \\ -\overline{X}\sigma^2_{b_1} & \sigma^2_{b_1} \end{bmatrix} \right)$$



Simultaneous Inferences

- A $(1-\alpha)\%$ **joint confidence region** for $\beta = (\beta_0, \beta_1)^T$ is given by

$$\left\{ \beta : \frac{(\hat{\beta}-\beta)^T \mathbf{X}^T \mathbf{X} (\hat{\beta}-\beta)}{pMSE} \leq F_{1-\alpha}(p, n-p) \right\}$$



Simultaneous Inferences

- Simultaneous Estimation of Mean Responses
  - Estimate m mean responses with a given family confidence coefficient:

    Bonferroni procedure:

    $$\hat{y} \pm BS_{\hat{y}_h}$$

    where $B = t_{1-\frac{\alpha}{2m},n-2}$ is the Bonferroni adjusted critical point.

    Working-Hotelling: based on the simultaneous confidence band for the entire regression

    $$\hat{y} \pm WS_{\hat{y}_h}$$

    where $W = 2F_{1-\alpha,2,n-2}$.
  - Calculate both B and W multiples to determine which procedure leads to tighter confidence limits.

Inverse Predictions

- Sometimes, a regression model of y on x is used to make a prediction of the value of x which gave rise to a new observation y.

- This is known as an inverse prediction.
- This approach is also called calibration method.
- In inverse predictions, regression model is assumed as before:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

- The estimated regression function based on n observations is obtained as usual.

Inverse Predictions

- A new observation $y_{h(new)}$ becomes available, and it is desired to estimate the level $x_{h(new)}$ that gave rise to this new observation.
- Natural point estimation is:

$$\hat{x}_{h(new)} = \frac{y_{h(new)} - \hat{\beta}_0}{\hat{\beta}_1}; \quad \hat{\beta}_1 \neq 0.$$

- Approximate $1 - \alpha$ confidence limits for $x_{h(new)}$ is:

$$\hat{x}_{h(new)} \pm t(\alpha/2, n-2)\sqrt{var(\hat{x})},$$

- where

$$var(\hat{x}) = \frac{MSE}{\hat{\beta}_1^2}\left[1 + \frac{1}{n} + \frac{(\hat{x}_{h(new)} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

Inverse Predictions

- The approximate confidence interval is appropriate if

$$\frac{MSE[t(\alpha/2, n-2)]^2}{\hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  is small.

- The inverse prediction problem has aroused controversy among statisticians. Some statisticians have suggested that inverse predictions should be made in direct fashion by regressing x on y.

- This regression is called inverse regression.

Other Topics

- Design Aspects of a Simple Linear Regression
  - Choice of X levels
  - Power and sample size calculations

Choice of x levels

- Optimum choice of the x levels depends on the objective of the regression analysis (i.e., inferences about $\beta_0, \beta_1, \beta_0 + \beta_1 x$, or prediction).
- Many of the variances in simple linear regression model have the term $\sum_{i=1}^{n}(x_i - \bar{x})^2$ in their denominators. This suggests that to obtain small variances and hence narrow confidence intervals, it is desirable to select x values that are spread out.
- Power to detect a significant intercept or a slope can be increased with widely spaced X values.

Power and sample size calculations

- Assume equally spaced x values with equal number of observations at each x (generally recommended).
- To calculate the number of observations needed at each level of x, it is important to specify the desired power, values of x, the slope to be detected, and the common variance (σ^2).
- Example:

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}$$

- Let $X = 0, 1, 2, 3, 4$, $\sigma^2 = 1$ and m replicates at each x i.e. total sample size n = 5m.
- We would like to detect a specified slope under the alternative (say 1) with a given power.

Power and sample size calculations

$$Power = \Pr\{t_{n-2} > t_{0.95,n-2} \mid \beta_1 = 1\} = \Pr\{t_{5m-2,\sqrt{10m}} > t_{0.95,5m-2}\}$$