Week 2: Linear Models

2.1 Simple Linear Regression (SLR) analysis

- **Introduction**

X predictor Y response variable

$Y = \beta_0 + \beta_1 X$

Definition

Simple linear regression (SLR) is a method to explain the relationship between **two quantitative variables** using a **straight line**. One variable is a response variable Y and the other one is a predictor variable X.

Lets represent data as n pairs of observations: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
We are going to discuss

- how to calculate the intercept and slope estimates in the simple linear regression problem?
- how to assess the accuracy of the parameter estimates?
- how to assess the accuracy of the SLR model?
- what are some potential problems arising in linear regression in general?

2.1 Simple Linear Regression (SLR) analysis

- **Boston dataset in MASS package in R** (Housing Values in Suburbs of Boston, 506 rows (observations) and 14 columns (variables)).
  Fit a simple linear regression model using **medv** (median house value) as **response variable** and **lstat** (per cent of households with low socioeconomic status) as **predictor variable**.

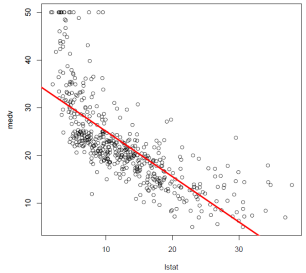


Figure: Scater plot of lstat versus medv in Boston data set.

2.1 Simple Linear Regression (SLR) analysis

- **Steps for simple linear regression analysis**
  - **Step 1**: Inspect, summarise and visualise your dataset.
  - **Step 2**: Produce a **scatter plot** of the response variable versus the explanatory variable. What is the relationship?
  - **Step 3**: **Fit** the SLR model using the **lm() function** in R. Write down the resulting regression equation. What does this equation tell you?
  - **Step 4**: Assess the **accuracy of the coefficient** estimates using the R output.
  - **Step 5**: Assess the **accuracy of the SLR model**.
  - **Step 6**: Identify any potential problems in your analysis by using **diagnostic plots**.
  - **Step 7**: Use the regression equation to **predict**

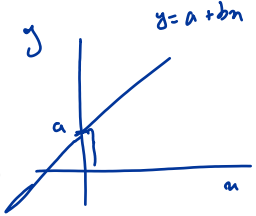
Steps 1 and 2 just the code

2.1 Simple Linear Regression (SLR) analysis

- **Step 3: Fitting the SLR**
  - SLR model:one independent variable X.
  - The relationship between $E(Y_i)$ and X_i is a straight line:

$$E(Y_i) = \beta_0 + \beta_1 X_i, \quad \text{for} \quad i = 1, \ldots, n,$$

intercept slope of the line



where

- β_0 - intercept of the line - the value of $E(Y_i)$ when $X = 0$;
- β_1 - slope of the line - the rate of change in $E(Y_i)$ per unit change in X.

  - random error ε_i : deviation of the observation Y_i from its population mean $E(Y_i)$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{for} \quad i = 1, \ldots, n,$$

Important assumptions in SLR analysis

  - X_i are measured without error (fixed constants)
  - $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

2.1 Simple Linear Regression (SLR) analysis

- **Cont. Step 3: Fitting the SLR**
  If $\hat{\beta}_0$ and $\hat{\beta}_1$ are the empirical estimates of β_0 and β_1, then

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

  is the estimated mean of Y_i, or prediction of Y_i, when $X_i = x_i$, for each $i = 1, \ldots, n$.

- **Question**: How to estimate $\hat{\beta}_0$ and $\hat{\beta}_1$??
  - Define the residuals as $e_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$.
  - The estimates of β_0 and β_1 are obtained by minimizing the residual sum of squares (RSS), given by

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2.$$

2.1 Simple Linear Regression (SLR) analysis

- **Cont. Step 3: Fitting the SLR**
  - **Minimizing RSS**
    - The derivatives of *RSS* with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ are set to zero.

$$n(\hat{\beta}_0) + (\sum_{i=1}^{n} x_i)\hat{\beta}_1 = \sum_{i=1}^{n} y_i \quad \text{and} \quad (\sum_{i=1}^{n} x_i)\hat{\beta}_0 + (\sum_{i=1}^{n} x_i^2)\hat{\beta}_1 = \sum_{i=1}^{n} x_i y_i. \qquad (2.1.1)$$

      These equations are called **normal equations**.
    - Solving the above equations gives the least squares estimates for the slope and intercept:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad (2.1.2)$$

      where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the sample means.
  - **The estimates form (2.1.2) give the equation of the best fitting line:**

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

2.1 Simple Linear Regression (SLR) analysis

- **Cont. Step 3: Fitting the SLR**

  To make sure $\hat{\beta}_0$ and $\hat{\beta}_1$ really minimize RSS:
  - **Calculate the second derivatives** of RSS with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$, called **Hessian matrix** (matrix of second derivatives)

$$H = 2\begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}.$$

  - **Show that H is positive definite**. Since $n > 0$ and

$$\det(H) = 4\left(n(\sum x_i^2) - (\sum x_i)^2\right) > 0,$$

  H is positive definite and therefore $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize *RSS*.

Note: For $c \neq 0$ and $A_{p\times p}$, $|cA| = c^p|A|$ and

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

R Code

$\widehat{medv} = 34.55 - 0.95\,lstat$

2.1 Simple Linear Regression (SLR) analysis

- **Step 4: Assessing the accuracy of the estimated coefficients**
  - The coefficient estimates are unbiased,

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

  - $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$ can be computed as:

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \quad \text{and} \quad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

    where $\sigma^2 = \text{Var}(\varepsilon)$.
  - Generally, σ is unknown, but can be estimated by **the residual standard error**:

$$\hat{\sigma} = RSE = \sqrt{\frac{RSS}{n-2}}.$$

    When σ is estimated, we should write $\widehat{SE}(\hat{\beta}_0)$ and $\widehat{SE}(\hat{\beta}_1)$.
    For simplicity, we will **not use** this extra "hat" in our notations.

R Code

2.1 Simple Linear Regression (SLR) analysis

- **Cont. Step 4: Assessing the accuracy of the estimated coefficients**
  - **Confidence Interval**

    Standard errors can be used to compute a $(1-\alpha)100\%$ confidence intervals for β_0 and β_1 as:

    $$[\hat{\beta}_k - t_{\alpha/2,n-2} SE(\hat{\beta}_k), \hat{\beta}_k + t_{\alpha/2,n-2} SE(\hat{\beta}_k)],$$

    $k = 0, 1$, where $t_{\alpha/2,n-2}$ is $\alpha/2$ critical value of a Student-t distribution with $n-2$ degrees of freedom.

2.1 Simple Linear Regression (SLR) analysis

- **Cont. Step 4: Assessing the accuracy of the estimated coefficients**
  - **Hypothesis tests on the coefficients**

    We want to perform hypothesis tests on the coefficients

$$\begin{cases} H_0 : \beta_1 = 0 (\text{ there is no relationship between } X \text{ and } Y) \\ H_1 : \beta_1 \neq 0 (\text{ there is some relationship between } X \text{ and } Y) \end{cases}.$$

- Compute a t-statistic, given by

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \sim t_{n-2}, \qquad \text{Under } H_0.$$

- If $|t| < t_{\alpha/2, n-2}$, we **cannot reject the** H_0 at the α level of significance.
- Equivalent to the decision based on a p-value:

  we reject H_0 if p-value is small enough (p-value $< \alpha$).

R Code. What is P-value??

2.1 Simple Linear Regression (SLR) analysis

- **Step 5: Assessing the accuracy of the SLR model**
  The quality of a linear regression fit is typically assessed using RSE and the R^2 statistic.
  - **Residual Standard Error**
    - Recall:

$$RSE = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

    - The RSE is considered a measure of lack of fit of the model.
    - Useful when fits from two different models are compared.
    - The RSS will be small for the model which fits the data well.

R Code.

2.1 Simple Linear Regression (SLR) analysis

- **Cont. Step 5: Assessing the accuracy of the SLR model**
  - **Coefficient of Determination R^2**
    - R^2 : A measure of the contribution of the independent variable(s) in the model

$$R^2 = \frac{TSS - RSS}{TSS},$$

where **the total sum of squares (TSS)** and **the residual sum of squares** are

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad \text{and} \quad RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

- TSS: variability inherent in the response **before** the regression is performed;
- RSS: variability left **unexplained** after performing the regression;
- TSS-RSS: variability in the response **explained** by performing the regression;
- R^2 is **the proportion of variability in Y that can be explained using X**;

$$0 \leq R^2 \leq 1.$$

- In the **simple linear regression** setting, $R^2 = r^2$, where r is the **correlation coefficient**.

R Code.

2.1 Simple Linear Regression (SLR) analysis

- **Step 6: Diagnostic plots**
  Some potential problems may arise in linear regression. Below is the list of possible issues and some diagnostic plots to identify them.
  - **Non-linearity of the response-predictor relationship**: **Residual plots** - we plot **residuals** e_i **versus** x_i **or** $\hat{y}_i$. Non-linearity can be seen in the presence of a pattern, such as U-shape.
  - **Correlation of error terms**: If there is a time component in the data, we plot the residuals as a function of time (when data is time dependent).
  - **Non-constant variance of error terms**: **Residual plots** - heteroscedasticity can be seen in the form of a funnel shape in the e_i **versus** $\hat{y}_i$ plot.
  - **Outliers**: observations where y_i is unusually far from $\hat{y}_i$.
    Use **plot of studentized residuals**, computed by dividing each residual by its estimated standard error (RSE). Observations whose studentized residuals are greater than 3 in absolute value are possible outliers.

$$r_D = \frac{r}{RSE}$$

2.1 Simple Linear Regression (SLR) analysis

- **Cont. Step 6: Diagnostic plots**
  - **High-leverage points**: Observations with high-leverage have an unusual value for x_i.
    Plot **studentized residuals versus the leverage statistic** defined by

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2}.$$

    The leverage statistic has values between $1/n$ and 1, with average $2/n$. If given observation has h_i that exceeds 2 or 3 times the average $2/n$, then that point has high leverage.
  - **Collinearity**: Collinearity refers to the situation in which two or more predictor variables are closely related to one another (not the case in SLR).

R Code.

2.1 Simple Linear Regression (SLR) analysis

- **Step 7: Prediction**

  Prediction interval or confidence interval?
  - A prediction interval reflects the uncertainty around a single value, while a confidence interval reflects the uncertainty around the mean prediction values.
  - A prediction interval will be generally much wider than a confidence interval for the same value.

  Which one should we use?
  - The answer to this question depends on the context and the purpose of the analysis.
  - Generally, we are interested in specific individual predictions, so a prediction interval would be more appropriate.
  - Using a confidence interval when you should be using a prediction interval will greatly underestimate the uncertainty in a given predicted value.

For more information, refer to "Introduction to Linear Regression Analysis", D.C. Montgomery et al. Page 30-34

`https://www.statology.org/confidence-interval-vs-prediction-interval/`

R Code.

2.2 Linear Models (LM)

- **Introduction to Linear Gaussian Models**
  The basis model for analysis of continuous data is

$$\mathbb{E}(Y_i) = \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta} \quad , \; Y_i \text{ are independent and } \quad Y_i \sim N(\mu_i, \sigma^2),$$

  These models are called **general linear models** and there are three main models of this form:
  - **Multiple linear regression**: association between a continuous response and several explanatory variables
  - **Analysis of variance (ANOVA)**: comparisons of the means of three or more groups (response variable is continuous and explanatory variables are categorical or qualitative and they are called factors).
  - **Analysis of covariance (ANCOVA)**: comparisons of the means, but also includes one or more covariates. (Similar to ANOVA, but at least one of the explanatory variables is continuous).

2.2 Linear Models (LM)

- **Cont. Introduction to Linear Gaussian Models**
  - These **general linear models** are usually written as

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{2.2.1}$$

where $\mathbf{y}^\top = [Y_1, \ldots, Y_N]$, $\beta^\top = [\beta_1, \beta_2, \ldots, \beta_p]$ and $\varepsilon^\top = [\varepsilon_1, \ldots, \varepsilon_N]$ with $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ for $i = 1, \ldots, N$.

  - $\mathbf{X}$ is an $N \times p$ **design matrix**, and in a **multiple regression** is set to

$$\mathbf{X} = \begin{pmatrix} 1 & X_{12} & X_{13} & \ldots & X_{1p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{N2} & X_{N3} & \ldots & X_{Np} \end{pmatrix}.$$

$\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$

intercept

  - β_j is interpreted as the average effect on Y of a one unit increase in the covariate x_j, **holding all the other predictors fixed**.

2.2 Linear Models (LM)

- **Cont. Introduction to Linear Gaussian Models**
  - **Error** is all the terms we have missed with the model and is usually considered independent from **X**.
- The model is **linear in the parameters**, for instance:

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3}^2 \quad \text{or} \quad \mathbb{E}(Y_i) = \beta_1 + \gamma_1 \delta_1 X_{i2} + \exp(\beta_2) X_{i3}.$$

But NOT:

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 X_{i2}^{\beta_2} \quad \text{or} \quad \mathbb{E}(Y_i) = \beta_1 \exp(\beta_2 X_{i2}).$$

2.2 Linear Models (LM)

- **Estimation and accuracy of coefficient estimates in Linear Gaussian Models**
  There exists several methods to estimate the coefficients of a linear (Gaussian) model.
  - **Maximum likelihood estimation**: Use the distributional assumptions to derive the likelihood.
  - **Least squares estimation**: Don't make any assumptions about the distribution of Y.

  We will quantify the uncertainty that comes with the estimation by constructing confidence intervals.

2.2 Linear Models (LM)

in normal dist

$E(Y_i) = \mu_i = x_i^T \beta = \eta_i$

- **Maximum likelihood estimation**
  The score function is given by

for normal dist $\frac{\partial \mu_i}{\partial \eta_i} = 1$

$$U_j = \sum_{i=1}^{N} \left[\frac{(y_i - \mu_i)}{\mathbb{V}ar(Y_i)} x_{ij} \left(\frac{d\mu_i}{d\eta_i} \right) \right]$$

while the information is of the form

$$\mathcal{I} = \frac{1}{\sigma^2} \mathbf{X}^T \mathbf{X}$$

$= \frac{1}{\sigma^2} X^T I X$

$= X^T W X$ → $\text{diag}\left(\frac{1}{\sigma^2}\right)$

Proof.

2.2 Linear Models (LM)

newton-Raphson

- **Cont. Maximum likelihood estimation**
  Apply the method of scoring to approximate the MLE:

$$\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}^{(m-1)} + \left[\mathcal{I}^{(m-1)}\right]^{-1} \mathbf{u}^{(m-1)}$$

$$\left[\mathcal{I}^{(m-1)}\right] \hat{\boldsymbol{\beta}}^{(m)} = \left[\mathcal{I}^{(m-1)}\right] \hat{\boldsymbol{\beta}}^{(m-1)} + \mathbf{u}^{(m-1)} \tag{2.2.4}$$

From Equation (2.2.2), the information matrix can be written as

$$\mathcal{I} = \mathbf{X}^\top \mathbf{W} \mathbf{X} \tag{2.2.5}$$

where $w_{ii} = \frac{1}{\mathbb{V}\text{ar}(Y_i)} \left(\frac{d\mu_i}{d\eta_i}\right)^2 = \frac{1}{\sigma^2}$.

2.2 Linear Models (LM)

- **Cont. Maximum likelihood estimation**

  The expression on the right hand side of (2.2.4) is the vector with elements

$$\sum_{k=1}^{p}\sum_{i=1}^{N}\frac{X_{ij}X_{ik}}{\mathbb{V}\mathrm{ar}(Y_i)}\left(\frac{d\mu_i}{d\eta_i}\right)^2\hat{\beta}_k^{(m-1)}+\sum_{i=1}^{N}\frac{(Y_i-\mu_i)X_{ij}}{\mathbb{V}\mathrm{ar}(Y_i)}\left(\frac{d\mu_i}{d\eta_i}\right)$$

  which can be written in matrix terms as

$$\mathcal{I}^{(m-1)}\hat{\boldsymbol{\beta}}^{(m-1)}+\mathbf{u}^{(m-1)}=\mathbf{X}^\top\mathbf{W}\mathbf{z} \qquad (2.2.6)$$

  where $\mathbf{z}$ is the vector with elements

$$z_i=\sum_{k=1}^{p}X_{ik}\hat{\beta}_k^{(m-1)}+(Y_i-\mu_i)\left(\frac{d\eta_i}{d\mu_i}\right)=\sum_{k=1}^{p}X_{ik}\hat{\beta}_k^{(m-1)}+\left(Y_i-\sum_{k=1}^{p}X_{ik}\beta_k^{(m-1)}\right)=Y_i$$

  And, therefore,

$$\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}\hat{\beta}=\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{y}\quad\Longrightarrow\quad\hat{\boldsymbol{\beta}}=(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

2.2 Linear Models (LM)

- **Cont. Maximum likelihood estimation**
  - **Properties**:
    - **Unbiasedness**

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbb{E}(\mathbf{Y}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

    - The **variance-covariance matrix** is $\mathcal{I}^{-1}$, since

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= Var((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}) \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Var(\mathbf{Y})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = \mathcal{I}^{-1} \end{aligned}$$

    - **Normality**

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$$

2.2 Linear Models (LM)

- **Least squares estimation**
  - Derive an estimator without any further assumption on the distribution of $\mathbf{y}$.
  - Let $N > p$. Under **Gauss-Markov assumptions**, i.e. $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 \mathbb{I}_N$, the **least squares function** is

$$\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) = \varepsilon^\top \varepsilon \tag{2.2.7}$$

  - This is a multivariate function in β and (strictly) **convex**.
  - There is a unique minimiser $\hat{\beta}$, satisfying

$X^T X$ is symmetric

$$\frac{d}{d\beta}(\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) = \frac{d}{d\beta}\left[\mathbf{y}^\top \mathbf{y} - 2\beta^\top \mathbf{X}^\top \mathbf{y} + \beta^\top \mathbf{X}^\top \mathbf{X}\beta\right]$$

$$= -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\beta = 0 \quad \Rightarrow \quad \mathbf{X}^\top \mathbf{X}\beta = \mathbf{X}^\top \mathbf{y}$$

Assuming that $\mathbf{X}$ has rank $N \geq p$, we can write

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

$\hat{\beta}$ is **unbiased**, but needs an assumption on the distribution of $\mathbf{y}$ to derive its distribution.

2.2 Linear Models (LM)

$I^{-1} \longrightarrow SE(\hat{\beta}_i)$

- **Confidence Intervals for regression parameters**
  - The standard error, $SE(\hat{\boldsymbol{\beta}})$, is the estimate of the uncertainty about $\hat{\boldsymbol{\beta}}$:

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{N} (X_i - \bar{X})^2} \right] \quad \text{and} \quad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{N} (X_i - \bar{X})^2},$$

    where $\sigma^2 = \text{Var}(\varepsilon)$.
  - **Remark**: the $SE(\hat{\beta}_1)$ is smaller when the X_i are more spread out and we are more able to estimate the slope of the line.
  - $(1-\alpha)100\%$ confidence intervals are:

Simple Linear regression $p=2$ $\beta_0 + \beta_1 n$

$$\left[\hat{\beta}_k - t_{1-\alpha/2,n-2} SE\left(\hat{\beta}_k\right), \hat{\beta}_k + t_{1-\alpha/2,n-2} SE\left(\hat{\beta}_k\right) \right],$$

    $k = 0, 1$, where $t_{1-\alpha/2,n-2}$ represents the $1-\alpha/2$ quantile of the Student-t distribution with $n-2$ degrees of freedom.

Example

2.2 Linear Models (LM)

- **Distribution of residuals in Linear Gaussian Models**
  - **hat matrix**

$$\boldsymbol{H} := \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top.$$

    - The hat matrix puts the hat on $\boldsymbol{y}$:

$$\text{Since } \hat{\beta} = (X^T X)^{-1} X^T \mathbf{y} \text{ and } \hat{\mathbf{y}} = X\hat{\beta} \;\Rightarrow\; \boldsymbol{H}\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{X}\hat{\beta} = \hat{\boldsymbol{y}}.$$

    - $\boldsymbol{H}$ is **symmetric** and **idempotent** ($\boldsymbol{H}^2 = \boldsymbol{H}$) $\rightarrow$ a **projection matrix**.
    - transforms $\boldsymbol{y}$ in N-dimensional space to vector $\hat{\boldsymbol{y}}$ in a subspace such that $\hat{\boldsymbol{y}}$ is as close to $\boldsymbol{y}$ as possible.
  - $\hat{y}_i = \sum_{j=1}^{n} H_{ij} y_j$ is a weighted sum of the y_j's .
  - The effect that y_i has on its fitted value is H_{ii}, the ith diagonal entry of H, which gives the **leverage**, (used to diagnosing influential points in the regression).

For an indempotent matrix $\boldsymbol{A}$: $\text{rank}(\boldsymbol{A}) = \text{tr}(\boldsymbol{A})$

2.2 Linear Models (LM)

- **Cont. Distribution of residuals in Linear Gaussian Models**
  - Recall: $\hat{\boldsymbol{y}} := \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and the residuals are

$$\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}. \tag{2.2.8}$$

$Y \sim N(X\beta, \sigma^2 I)$

Theorem

If the Gauss-Markov assumptions hold, then

$$r \sim \mathcal{N}\left(0, \sigma^2(\boldsymbol{I} - \boldsymbol{H})\right)$$

and

$$\sigma^{-2}\boldsymbol{r}^\top\boldsymbol{r} = \sigma^{-2}\sum_{i=1}^{N} r_i^2 \sim \chi^2_{N-p}$$

$p = 2$ in SLR

2.2 Linear Models (LM)

- **Cont. Distribution of residuals in Linear Gaussian Models**
  - $\hat{\sigma}^2 := \sum_{i=1}^{N} r_i^2/(N-p)$ is an unbiased estimator of σ^2.
  - $\hat{\sigma}$ is called the **residual standard error** and is used to estimate the coefficient standard error.
- **Residual plots** are useful tools for identifying **non-linearity** in the data:
  - plot the residuals $(Y_i - \hat{Y}_i)$ versus the fitted values $\hat{Y}_i$.
  - Ideally the residual plot will not show any discernible pattern.
  - If the residual plot indicates that there are non-linear associations in the data, then a simple approach is to use non-linear transformations of the predictors, i.e. $\log x$, $\sqrt{x}$, x^2, etc.

Plots on Ed

2.2 Linear Models (LM)

- **Assessing model assumptions**
  We have made several assumptions for the model to be valid. It is therefore needed to check if these assumptions hold. In order to do so we look into:
  - The standardised residuals
  - The presence of high leverage points
  - The Cook's distance

2.2 Linear Models (LM)

- **Assessing model assumptions: Standardised residuals**
  - The residuals vs fitted values plot may reveal possible violations of linearity or homoscedasticity.
    - Standardising the residuals may lead to a better feeling for their magnitude.
    - The variance of the i-th residual is

$$\text{Var}[r_i] = \text{Cov}[\boldsymbol{e}_i^\top \boldsymbol{r}] = \boldsymbol{e}_i^\top \sigma^2(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{e}_i = \sigma^2(1 - h_{ii}) \tag{2.2.9}$$

    - The standardised residuals are:

$$r_{0i} = \frac{r_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}.$$

  - **Recommendations about outliers**:
    - Points should not be routinely deleted from an analysis just because they do not fit the model.
    - Outliers and bad leverage points are signals, flagging potential problems with the model.
    - Outliers often point out an important feature of the problem not considered before. They may point to an alternative model in which the points are not an outlier.

R Code

2.2 Linear Models (LM)

- **Assessing model assumptions: Leverage**
  - The i-th diagonal entry h_{ii} of $\boldsymbol{H}$ is called the **leverage** of the i-th observation.
  - Let $\hat{y}_i^{-i}$ denote the fitted value at x_i where (x_i, y_i) is removed. Then

$$\frac{\hat{y}_i - \hat{y}_i^{-i}}{r_i} = \frac{h_{ii}}{1 - h_{ii}}, \qquad r_i = y_i - \hat{y}_i.$$

  - Model fits are sensitive to data with high leverage.
  - Residuals at points with high leverage have small variance.
  - For SLR, $y_i = \beta_0 + \beta_1 x_1$ the leverage is largest at the most extreme x-values.
  - The sum of leverages equals the number of parameters:

$$\sum_{i=1}^{N} h_{ii} = \text{tr}(\boldsymbol{H}) = \text{tr}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \text{tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}) = \text{tr}(\boldsymbol{I_p}) = p$$

    Therefore, $\frac{1}{N}\sum_{i=1}^{N} h_{ii} = \frac{p}{N}$. As a rule of thumb, if h_{ii} is greater than two or three times p/N, it may be a concern.

R Code

2.2 Linear Models (LM)

- **Assessing model assumptions: Cook's distance**
  - The Cook's distance is defined by:

$$D_i = \frac{1}{p\hat{\sigma}^2}|\hat{\mathbf{y}} - \hat{\mathbf{y}}^{-i}|^2$$

  - Cook's distance measures the (rescaled) sum of squared differences between fitted values when the i-th datum is removed.
  - It is a measure for the influence of the i-th datum on the entire model fit.
  - It can be shown that

$$D_i = \frac{1}{p}\frac{h_{ii}}{1-h_{ii}}r_{0i}^2$$

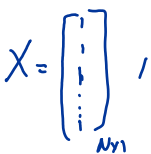
  - Fox (2002, p. 198) is among many authors who recommend $4/(n-2)$ as a rough cutoff for noteworthy values of D_i for simple linear regression.
  - In practice, it is important to look for gaps in the values of Cook's distance and not just whether values exceed the suggested cut-off.

R Code

2.3 Hypothesis testing in Linear Models

- **Coefficient of determination**
  The strength of a **linear relationship** is measured by the sample correlation coefficient, R.
  - An R close to 1 indicates a positive linear relationship
  - An R close to -1 indicates a negative linear relationship.
- Equivalently, R^2 close to one indicates the strength of the linear regression.
- $\mathrm{RSS} = \sum_{i=1}^{N} \varepsilon_i^2 = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ is minimised by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$, therefore

$$\widehat{\mathrm{RSS}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}^\top \mathbf{y} - \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{y}$$



- For the **minimal model**, $Y_i = \beta_0 + \varepsilon_i$, we know that $\mathbf{X}^\top \mathbf{X} = N$ and $\mathbf{X}^\top \mathbf{y} = \sum_{i=1}^{N} y_i$, then RSS is minimised by $\hat{\boldsymbol{\beta}} = \hat{\beta}_0 = \bar{y}$, and $\mathrm{RSS}_0 = \sum_{i=1}^{N} (y_i - \bar{y})^2$.
  - RSS_0 is the worst possible value for RSS, also known as **the total sum of squares** (TSS).

2.3 Hypothesis testing in Linear Models

- **Cont. Coefficient of determination**
  - If **parameters are added** to the model, then RSS must **decrease**. The relative amount of decrease

$$R^2 := \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} \tag{2.3.1}$$

is called the **coefficient of determination**. It is the **proportion of the total variation in the data which is explained by the model**.

- For the maximal model, $\mathrm{RSS} = 0$ and $R^2 = 1$.
- R^2 always increases when more variables are added to the model.
- If adding a variable leads to a small increase in R^2, the contribution of that variable is small.
- If there is a covariate, $R^2 = \mathbb{C}or(Y, X)^2$. in SLR
- In multiple regression, $R^2 = \mathbb{C}or(Y, \hat{Y})^2$ (the property of the least squares estimates is that they maximises the correlation among the responses and the fitted linear model among all the possible linear models).

R Code

2.3 Hypothesis testing in Linear Models

- **The F-statistic in Linear Models**
  - For the Linear Gaussian Model

$$\boldsymbol{E}[Y_i] = \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad Y_i \sim N(\mu_i, \sigma^2)$$

with Y_i's independent, the deviance is:

$$D = \frac{1}{\sigma^2}(\mathbf{y}^\top \mathbf{y} - \hat{\boldsymbol{\beta}}^\top \mathbf{X}^\top \mathbf{y}) \quad (2.3.2)$$

  - **Select between two competing models M_0 and M_1**
    - Consider a null hypothesis H_0 and an alternative hypothesis H_1.

$$H_0 = \boldsymbol{\beta} = \boldsymbol{\beta}_0 = \begin{bmatrix}\beta_1 & \cdots & \beta_q\end{bmatrix}^\top, \quad H_1 = \boldsymbol{\beta} = \boldsymbol{\beta}_1 = \begin{bmatrix}\beta_1 & \cdots & \beta_p\end{bmatrix}^\top, \quad (p > q).$$

    - The scaled deviance can be used for model comparison.

$$\Delta D = D_0 - D_1 = \frac{1}{\sigma^2}\left[(\mathbf{y}^\top \mathbf{y} - \hat{\boldsymbol{\beta}}_0^\top \mathbf{X}_0^\top \mathbf{y}) - (\mathbf{y}^\top \mathbf{y} - \hat{\boldsymbol{\beta}}_1^\top \mathbf{X}_1^\top \mathbf{y})\right] = \frac{1}{\sigma^2}\left[\hat{\boldsymbol{\beta}}_1^\top \mathbf{X}_1^\top \mathbf{y} - \hat{\boldsymbol{\beta}}_0^\top \mathbf{X}_0^\top \mathbf{y}\right]$$

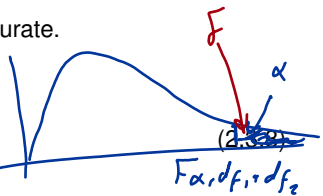
    - $D_0 \sim \chi^2(N-q)$ and $D_1 \sim \chi^2(N-p)$, and thus, for large N,

$$\Delta D \sim \chi^2(p-q).$$

2.3 Hypothesis testing in Linear Models

- **Cont. The F-statistic in Linear Models**
  - If the values of ΔD is in the critical region, reject H_0 in favour of H_1 (model M_1 provides a significantly better description of the data).
  - The standard deviation σ^2, however, is unknown;
    - replace it by its estimate $\hat{\sigma}^2$ results in (2.3.2) being inaccurate.
    - eliminate σ^2 by using the ratio

$$F = \frac{\frac{D_0 - D_1}{p-q}}{\frac{D_1}{N-p}} = \frac{\frac{\hat{\beta}_1^\top \mathbf{X}_1^\top \mathbf{y} - \hat{\beta}_0^\top \mathbf{X}_0^\top \mathbf{y}}{p-q}}{\frac{\mathbf{y}^\top \mathbf{y} - \hat{\beta}_1^\top \mathbf{X}_1^\top \mathbf{y}}{N-p}}$$



  - Under the null hypothesis H_0 (Model M_0), against the alternative hypothesis H_1 (Model M_1), $F \sim F(p-q, N-p)$.
  - Reject H_0 if $F > F_\alpha(p-q, N-p)$, where α is the size of the test (typically 0.05) and $F_\alpha(p-q, N-p)$ is the $(1-\alpha)$th quantile of the F(p-q, N-p) distribution.
  - Alternatively, we can compute the P-value: $P(F_{(p-q,N-p)} > F)$.

p-value $< \alpha$ $\Rightarrow$ reject H_0

2.3 Hypothesis testing in Linear Models

- **Cont. The F-statistic in Linear Models**
  - The F-statistic is usually used to test the hypothesis

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_p = 0 \quad \text{(Minimal Model)}$$
$$H_1 : \text{at least one } \beta_j \text{ is non-zero}$$

and

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/(p-1)}{\mathrm{RSS}/(N-p)} \tag{2.3.4}$$

where $\mathrm{TSS} = \sum_{i=1}^{N}(\mathrm{Y}_i - \bar{\mathrm{Y}})^2$ and $\mathrm{RSS} = \sum_{i=1}^{N}(\mathrm{Y}_i - \hat{\mathrm{Y}}_i)^2$.

- $\mathbb{E}(\mathrm{RSS}/(N-p)) = \sigma^2$
- under H_0, $\mathbb{E}[(\mathrm{TSS} - \mathrm{RSS})/p] = \sigma^2$
- Therefore, under H_0, the F-statistic is expected to be close to 1, while under H_1, $\mathbb{E}[(\mathrm{TSS} - \mathrm{RSS})/p] > \sigma^2$ and the F-statistic is larger than 1.

2.3 Hypothesis testing in Linear Models

- **Cont. The F-statistic in Linear Models**
  - **Relationship between the F-statistics and the R^2 coefficient.**

$$R^2 \times \mathrm{TSS} = (\mathrm{TSS} - \mathrm{RSS}) \quad \text{and} \quad \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - R^2.$$

and from (2.3.4) we have

$$F = \frac{\frac{\mathrm{TSS}-\mathrm{RSS}}{p-1}}{\frac{\mathrm{RSS}}{N-p}} = \frac{R^2 \times \mathrm{TSS}}{\mathrm{RSS}} \frac{N-p}{p-1} = \frac{R^2}{1-R^2} \frac{N-p}{p-1} \sim F_{p-1,N-p}.$$

  - **Remark 1**: If you test for the effect of any predictor separately and without any correction, about 5% of the p-values will be under α (e.g. 0.05) by chance. The F-statistic does not suffer from this problem because it adjusts for the number of predictors.
  - **Remark 2**: The approach using the F-statistic works when $p < N$; For $p > N$, multiple regression cannot be fitted and the F-statistic cannot be used.

Code