Week 2: Linear Models

2.1 Simple Linear Regression (SLR) analysis

- **Introduction**

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** (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**.

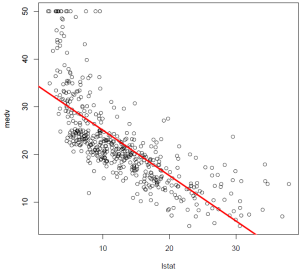


Figure: Scater plot of lstat versus medv in Boston data set.

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:

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

slop

intercept

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 $E(Y_i)$

error

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

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.

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

    - Solving the above equations gives the least squares estimates for the slope and intercept:

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

    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

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

R Code

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, called **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.

use t dist since we replace σ^2 with $\hat{\sigma}^2$, the distribution is t instead of normal

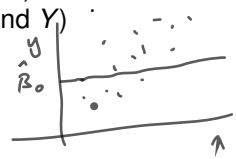
2.1 Simple Linear Regression (SLR) analysis

- **Cont. Step 4: Assessing the accuracy of the estimated coefficients**
  - **Hypothesis tests on the coefficients**
    We want to perform hypothesis tests on the coefficients

if $\beta_1 = 0$

$E(Y_i) = \beta_0$

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



- 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 $< \alpha$).

is the smallest α_0 which rejects the null hypothesis

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 to the data.
    - 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

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

variability that can be explained using regression

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**;

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

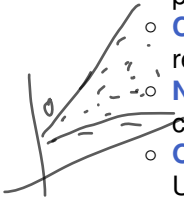
- In the **simple linear regression** setting, $R^2 = r^2$, where r is the **correlation coefficient**.

r is the estimation of $corr(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)\,var(Y)}}$

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.



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

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

$h_i > 2 \times 2/n$

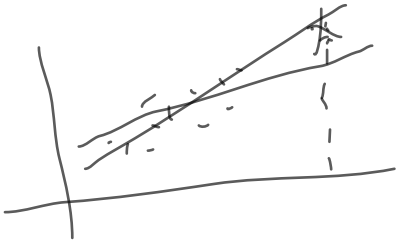
or $h_i > 3 \times 2/n$

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**

R Code.

2.2 Linear Models (LM)

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

$$\mathbb{E}(\mathrm{Y}_i) = \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta} \qquad \text{and} \qquad \mathrm{Y}_i \sim N(\mu_i, \sigma^2),$$

  There are three main models of this form:
  - **Multivariate regression**: association between a continuous response and several explanatory variables
  - **Analysis of variance (ANOVA)**: comparisons of more than two means
  - **Analysis of covariance (ANCOVA)**

- These **general linear** models are usually written as

error terms

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

  where $\mathbf{y}^\top = [\mathrm{Y}_1, \ldots, \mathrm{Y}_N]$, $\boldsymbol{\beta}^\top = [\beta_1, \beta_2, \ldots, \beta_p]$ and $\boldsymbol{\varepsilon}^\top = [\varepsilon_1, \ldots, \varepsilon_N]$ with $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ for $i = 1, \ldots, N$.

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

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

  - 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**.
- The model is **linear in the parameters**, for instance:

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

β_1^*

$\log \beta_1$

But NOT:

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

$\log Y_i = \quad + \beta_2 X_{i2}^*$

2.2 Linear Models (LM)

- **Estimation and accuracy of coefficient estimates in Linear Gaussian Models**
  There is 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 further 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)

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

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

while the information is of the form

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

p×p matrix

Proof.

2.2 Linear Models (LM)

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

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

vector $p \times 1$

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

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) can be written as

j-th element

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

$$\left(\frac{d\mu_i}{d\eta_i}\right)^2 \times \frac{d\eta_i}{d\mu_i}$$

  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

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

  And, therefore,

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

MLE

2.2 Linear Models (LM)

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

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

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

$$\mathcal{I}^{-1} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

    - **Normality**

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

2.2 Linear Models (LM)

$y = X\beta + \varepsilon$

$\varepsilon = y - X\beta$

- **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

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

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

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**, needs an assumption on the distribution of $\mathbf{y}$ to derive its distribution.

2.2 Linear Models (LM)

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

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

σ^2

    where $\sigma^2 = \mathrm{Var}(\varepsilon)$.
  - **Remark**: the $\mathrm{SE}(\hat{\beta}_1)$ 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:

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

    $k = 0, 1$, 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**

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

$H^T = H$

$H^2 = H$

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

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

    - $\boldsymbol{H}$ is **symmetric** and **idempotent** (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: $\hat{\boldsymbol{y}} := \boldsymbol{X}\hat{\beta}$ and the residuals are

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

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

Theorem

If the Gauss-Markov assumptions hold, then

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

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.
  - $\hat{\sigma}^2$ 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

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

    - 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.

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:

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

    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.

Code

2.2 Linear Models (LM)

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

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

  - 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

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

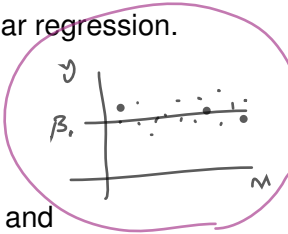
  - Fox (2002, p. 198) is among many authors who recommend $4/(n-2)$ as a rough cutoff for noteworthy values of D_i for simple linear regression.
  - In practice, it is important to look for gaps in the values of Cook's distance and not just whether values exceed the suggested cut-off.

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

$R^2 = 0.83$

83%

- **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.
- R^2 can be interpreted as the proportion of variance explained by the model.
- If there is a covariate, $R^2 = \mathbb{C}or(Y, X)^2$.
- 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).

Code

2.3 Hypothesis testing in Linear Models

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

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

with Y_i's independent, the deviance is:

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

!?

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

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

    - The scaled deviance can be used for model comparison.

M_1

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

Deviance M_0

    - $D_0 \sim \chi^2(N-q)$ 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

$$F = \frac{\overbrace{\frac{D_0 - D_1}{p-q}}^{\chi^2_{p-q}}}{\underbrace{\frac{D_1}{N-p}}_{\chi^2_{N-p}}} = \frac{\frac{\hat{\beta}_1^\top \mathbf{X}_1^\top \mathbf{y} - \hat{\beta}_0^\top \mathbf{X}_0^\top \mathbf{y}}{p-q}}{\frac{\mathbf{y}^\top \mathbf{y} - \hat{\beta}_1^\top \mathbf{X}_1^\top \mathbf{y}}{N-p}} \tag{2.3.3}$$

  - 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: $P(F_{(p-q,N-p)} > F)$.

2.3 Hypothesis testing in Linear Models

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

$$\begin{cases} H_0 : \underset{\sim}{\beta} = \beta_1 \text{ (intercept)} \\ H_1 : \underset{\sim}{\beta} = (\beta_1, \dots, \beta_p)^T \end{cases}$$

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

$M_0 \longrightarrow$ 1 parameter

$M_1 \longrightarrow$ p parameters

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, why?! $\mathbb{E}[(\mathrm{TSS} - \mathrm{RSS})/p] > \sigma^2$ 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.**

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

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

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 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

2.4 Confidence intervals and prediction intervals in Linear Models

- **Confidence and prediction intervals**
  - **Confidence vs Prediction Interval**
    Given a certain vector of predictors x^*, we want to find
    - confidence interval for the conditional mean $x^{*\top}\beta$
    - prediction interval for a future unobserved observation $Y^* = x^{*\top}\beta + \epsilon^*$ where ϵ^* is an error independent of ϵ_i, $i = 1, ..., n$, drawn from $N(0, \sigma^2)$.

values of predictors for 1 observation

$E(Y^*)$

error term

  - **Confidence interval** Y^* is Gaussian and $\boldsymbol{X}$ is the design matrix:

$$E(x^{*\top}\hat{\beta}) = x^{*\top}\beta \quad and \quad \mathrm{Var}(x^{*\top}\hat{\beta}) = \sigma^2 x^{*\top}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}x^*$$

where $\boldsymbol{X}$ is the design matrix for the fitted linear model. So

$$x^{*\top}\hat{\beta} \sim N(x^{*\top}\beta, \sigma^2 x^{*\top}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}x^*) \qquad \text{or} \qquad \frac{x^{*\top}\hat{\beta} - x^{*\top}\beta}{\sigma\sqrt{x^{*\top}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}x^*}} \sim N(0, 1).$$

unknown

2.4 Confidence intervals and prediction intervals in Linear Models

- **Cont. Confidence and prediction intervals**
  - $\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$ and $\boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$ are independent.
  - $x^{*\top}\hat{\boldsymbol{\beta}}$ and $(n-p)\hat{\sigma}^2/\sigma^2$ are independent.
  - Since $(n-p)\hat{\sigma}^2/\sigma^2$ has a χ^2_{n-p} distribution, the quotient of the two has a Student-t distribution with $(n-p)$ degrees of freedom:

$$\frac{x^{*\top}\hat{\beta} - x^{*\top}\beta}{\sigma\sqrt{x^{*\top}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}x^*}} \Big/ \sqrt{\frac{\frac{(n-p)\hat{\sigma}^2}{\sigma^2}}{n-p}} \sim t_{n-p} \quad \text{or} \quad \frac{x^{*\top}\hat{\beta} - x^{*\top}\beta}{\hat{\sigma}\sqrt{x^{*\top}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}x^*}} \sim t_{n-p}.$$

confidence interval

$x^{*T}\beta \in x^{*T}\hat{\beta} \pm t_{1-\frac{\alpha}{2},n-p}\,\hat{\sigma}\sqrt{x^{*T}(X^TX)^{-1}x^*}$

  - **Prediction intervals**

    Define $\hat{Y}^* = x^{*\top}\hat{\beta}$ and note that $E(Y^* - \hat{Y}^*) = 0$. Since $x^{*\top}\hat{\beta}$ and ϵ^* are independent,

$$\begin{aligned}\mathrm{Var}(Y^* - \hat{Y}^*) &= \mathrm{Var}(x^{*\top}\hat{\beta}) + \mathrm{Var}(\epsilon^*) \\ &= \sigma^2 x^{*\top}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}x^* + \sigma^2 = \sigma^2(1 + x^{*\top}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}x^*).\end{aligned}$$

$Var(x^{*T}\beta + \varepsilon^* - x^{*T}\hat{\beta}) = Var(-x^{*T}\hat{\beta} + \varepsilon^*)$

2.4 Confidence intervals and prediction intervals in Linear Models

- **Cont. Confidence and prediction intervals**
  - It can be shown that $Y^* - \hat{Y}^*$ and $(n-p)\hat{\sigma}^2/\sigma^2$ are independent.
  - Thus,

$$\frac{Y^* - \hat{Y}^*}{\sigma\sqrt{1 + x^{*\top}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}x^*}} \Big/ \sqrt{\frac{\frac{(n-p)\hat{\sigma}^2}{\sigma^2}}{n-p}} \sim t_{n-p}.$$

and upon simplifying

$$\frac{Y^* - \hat{Y}^*}{\hat{\sigma}\sqrt{1 + x^{*\top}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}x^*}} \sim t_{n-p}.$$

$Y^* \in \hat{Y}^* \pm t_{1-\alpha/2, n-p}\, \hat{\sigma}\sqrt{1 + \cdots}$

prediction interval

Code

2.5 ANOVA

- **Analysis of Variance (ANOVA)**
  - Analysis of variance is a method to compare **means of groups of continuous observations** where the groups are defined by the levels of the factors.
    - Y: continuous variable
    - **x**: categorical variable(s) $\longrightarrow$ Factors

    The element of **X** (design matrix) are dummy variables.

Code

2.5 ANOVA

Example: One-factor analysis

Genetically similar seeds are **randomly assigned** to be raised in either nutritionally enriched environment (treatment A or treatment B) or standard conditions (control group) using a completely randomised experimental design. After a predetermined time all plants are harvested, dried and weighted.

Y_i

X : $\underbrace{A, B, S}_{\text{levels of the factor } X}$

(treatment)

- This experiment is called a **completely randomised experiment**.
- The responses at level j, i.e. $Y_{j1}, \ldots, Y_{jn_j}$ are called replicates.
  - If $n_j = K$ for all j, it is called balanced. (**We will focus on this case**)
  - If $n_j = K_j$, the experiment is called unbalanced.

2.5 ANOVA

j is the number of treatments (levels of X)

- Let the response vector (of length $N = JK$) is given by:

$$\mathbf{y} = [\overbrace{Y_{11}, Y_{12}, \ldots, Y_{1K}}^{\text{level 1}}, \underbrace{Y_{21}, \ldots, Y_{2K}}_{\text{level 2}}, \ldots, \underbrace{Y_{J1}, \ldots, Y_{JK}}_{\text{level } J}]^\top$$

For $k = 1, \ldots, K$, we consider three specifications of the model:

Model 1. $\mathbb{E}(Y_{jk}) = \mu_j$

Model 2. $\mathbb{E}(Y_{jk}) = \mu + \alpha_j$

Model 3. $\mathbb{E}(Y_{jk}) = \mu + \alpha_j$, under constraint $\underbrace{\alpha_1 = 0}$.

$E(Y_{1k}) = \mu$

$E(Y_{2k}) = \mu + \alpha_2$

$E(Y_{Jk}) = \mu + \alpha_k$

2.5 ANOVA

- **Model 1.** $\mathbb{E}(\mathbf{Y}_{jk}) = \mu_j$ **for** $k = 1, \ldots, K$

  Model (1) can be re-written as $\mathbb{E}(Y_i) = \sum_{j=1}^{J} x_{ij}\mu_j$ for $i = 1, \ldots, N$ where x_{ij} represent an element of the design matrix through:

  - $x_{ij} = 1$ if response Y_i corresponds to level j
  - $x_{ij} = 0$ otherwise

  $Y_{ij} \longrightarrow x_{ij} = \begin{cases} 1 & Y_i \text{ belongs to level } j \\ 0 & \text{otherwise} \end{cases}$

  This gives $\mathbb{E}(\mathbf{y}) = \mathbf{X}\beta$ where

$$\beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_J \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{pmatrix}.$$

  The estimate $\hat{\beta}$ is the vector of sample means for each group

Code

2.5 ANOVA

- **Model 2 -** $\mathbb{E}(\mathbf{Y}_{jk}) = \mu + \alpha_j$ **for** $k = 1, \ldots, K$
  - μ is an average effect for all levels
  - α_j is an additional effect due to level j.

  In this case we have:

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

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_J \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{array}{c} \begin{array}{ccccc} \mu & \alpha_1 & & & \end{array} \\ \begin{pmatrix} \mathbf{1} & \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{pmatrix} \end{array} .$$

$N \times (J+1)$

sum the columns = first column

the columns of X are linearly dependent

X is not full rank

  - The design matrix as an additional column of elements equal to 1.
  - The first row (or column) of the $(J+1) \times (J+1)$ matrix $\mathbf{X}^\top \mathbf{X}$ is the sum of the remaining rows (or columns), therefore $\mathbf{X}^\top \mathbf{X}$ is **singular** and there is **no unique solution**.

2.5 ANOVA

- **Cont. Model 2 - $\mathbb{E}(\mathbf{Y}_{jk}) = \mu + \alpha_j$ for $k = 1, \ldots, K$**
  - The general solution can be written

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_J \end{bmatrix} = \frac{1}{K} \begin{bmatrix} 0 \\ \mathrm{Y}_{1.} \\ \vdots \\ \mathrm{Y}_{J.} \end{bmatrix} - \lambda \begin{bmatrix} -1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

why ?!

where λ is an arbitrary constant and $\mathrm{Y}_{j.} = \sum_{i=1}^{n_j} \mathrm{Y}_{ij}$. Usually a sum-to-one constraint is used, such that $\sum_{j=1}^{J} \alpha_j = 0$, i.e.

$$\frac{1}{K} \sum_{j=1}^{J} Y_{j.} - J\lambda = 0 \qquad \Longleftrightarrow \qquad \lambda = \frac{1}{JK} \sum_{j=1}^{J} Y_{j.} = \frac{Y_{..}}{N}$$

and therefore $\hat{\mu} = \frac{Y_{..}}{N}$ and $\hat{\alpha}_j = \frac{Y_{j.}}{K} - \frac{Y_{..}}{N}$ $\hat{\mu}$ $j = 1, \ldots, J$

Code

2.5 ANOVA

- **Model 3 - $\mathbb{E}(\mathbf{Y}_{jk}) = \mu + \alpha_j$ for $k = 1, \ldots, K$, under constraint $\alpha_1 = 0$**
  - μ represents the effect of the first level
  - α_j measures the difference between the first level and the j-th level of the factor. This is called a **corner point parametrisation**. We have

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_2 \\ \vdots \\ \alpha_J \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{array}{c} \begin{array}{ccccc} \text{level 1} & \text{level 2} & \cdots & & \text{level } J \end{array} \\ \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{pmatrix} \end{array} \underset{N \times J}{}.$$

columns of X are linearly indep

$\mathbf{X}^\top \mathbf{X}$ is **non-singular**, so there **a unique solution**:

$$\hat{\boldsymbol{\beta}} = \frac{1}{K} \begin{bmatrix} \mathrm{Y}_{1.} \\ \mathrm{Y}_{2.} - \mathrm{Y}_{1.} \\ \vdots \\ \mathrm{Y}_{J.} - \mathrm{Y}_{1.} \end{bmatrix} = (X^T X)^{-1} X^T Y$$

2.5 ANOVA

Based on Model 1

$H_0: \mu_1 = \mu_2 = \cdots = \mu_J = \mu$

H_1: they are not all equal

- **Cont. Model 3 - $\mathbb{E}(\mathbf{Y}_{jk}) = \mu + \alpha_j$ for $k = 1, \ldots, K$, under constraint $\alpha_1 = 0$**
  - In the analysis of the variance, it is important to compare the **alternative hypothesis** (means for each level differ) with the **null hypothesis** (means are all equal)
  - For the null model, $\mathbb{E}(\mathrm{Y}_{jk}) = \mu$ and the design matrix is a column vector of elements equal to 1, i.e. $\mathbf{X}^\top\mathbf{X} = N$ and $\mathbf{X}^\top\mathbf{y} = \mathrm{Y}_{..}$.

$$D_1 = \frac{1}{\sigma^2}(\mathbf{y}^\top\mathbf{y} - \hat{\beta}^\top\mathbf{X}^\top\mathbf{y}) \qquad \text{and} \qquad D_0 = \frac{1}{\sigma^2}\left[\sum_{j=1}^{J}\sum_{k=1}^{K}\mathrm{Y}_{jk}^2 - \frac{Y_{..}^2}{N}\right]$$

and the F-statistic

$$F = \frac{D_0 - D_1}{J-1} \Big/ \frac{D_1}{N-J}$$

$\sim F_{(J-1, N-J)}$

Code

2.6 Analysis of covariance (ANCOVA)

- some of the explanatory variables are **dummy** variables representing **factor levels** and others are **continuous** measurements called **covariates**.
- We compare means of subgroups defined by **factor levels**, but we consider that the covariates may also affect the response.
- $\Rightarrow$ we compare the means after adjustment for covariate effects.

Code

2.7 General linear models

- The term general linear models is used for **Gaussian models** with any combination of categorical and continuous explanatory variables.
- The factors can be
  - crossed: there are observations for each combination of levels of the factors (see two factors ANOVA)
  - nested: the combinations of factors are different

2.7 General linear models

ANOVA

Drug A_1	Drug A_2
B_1 B_2 B_3	B_1 B_2 B_3

- **Example on nested factors**
  - Two-factor nested design:

	Drug A_1			Drug A_2	
Hospitals	B_1	B_2	B_3	B_4	B_5
Responses	Y_{111}	Y_{121}	Y_{131}	Y_{241}	Y_{251}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	Y_{11n_1}	Y_{12n_2}	Y_{13n_3}	Y_{24n_4}	Y_{25n_5}

  - We want to compare the effects of the two drugs and possible differences among hospitals using the same drug.
  - The saturated model is

$$\mathbb{E}(Y_{jkl}) = \mu + \alpha_1 + \alpha_2 + (\alpha\beta)_{11} + (\alpha\beta)_{12} + (\alpha\beta)_{13} + (\alpha\beta)_{24} + (\alpha\beta)_{25}.$$

    under constraints $\alpha_1 = 0$, $(\alpha\beta)_{11} = 0$ and $(\alpha\beta)_{24} = 0$
  - Hospitals 1, 2 and 3 can be only compared within drug A_1 and hospitals 4 and 5 can be only compared within drug A_2.

2.7 General linear models

- This model is not different from other Gaussian models
  - Response variable are normally distributed
  - Response and explanatory variables are linearly related
  - The variance σ^2 is constant
  - The responses are independent
- These assumption must be checked by looking at the **residuals**.
- If the assumption of normality is not plausible, use the Box-Cox transformation:

$$y^* = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases} \tag{2.7.1}$$

  - if $\lambda = 1$, y is unchanged (except for a location shift)
  - if $\lambda = \frac{1}{2}$, the transformation is the square root
  - if $\lambda = -1$, the transformation is the reciprocal
  - if $\lambda = 0$, the transformation is the logarithm

Estimate λ which produces the "most normal" distribution by the method of maximum likelihood.

2.8 Extension

- **Non-additive associations**
  - The additive assumption means that the effect of changes in a predictor X_j on the response Y is **independent** of the values of **other predictors**.
  - In many situations, there is a synergy effect, i.e. increasing the level of one covariate may interact with the level of another. This is called **interaction** in statistics.

Example

Take

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

This means that

$$Y = \beta_0 + (\beta_1 + \beta_3 x_2) x_1 + \beta_2 x_2 + \varepsilon = \beta_0 + \tilde{\beta}_1 x_1 + \beta_2 x_2 + \varepsilon$$

where $\tilde{\beta}_1 = \beta_1 + \beta_3 x_2$, i.e. $\tilde{\beta}_1$ changes with x_2 and the effect of x_1 on Y is no longer constant.

2.8 Extension

$(\alpha\beta)_{11}$

$\alpha_1 \beta_1$

- **Cont. Non-additive associations**
  - Remark:
    - The hierarchical principle states that if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.
    - The concept of interactions applies to qualitative variables, to quantitative variables or to a combination of both.

Example and Code

2.8 Extension

- **Non-linear associations between X and Y**
  - A non-linear association can be suggested by looking at the **residuals**.
  - A popular model is a **U-shaped** association, that can be modelled by a **quadratic** association

$$\mathbb{E}(\mathrm{Y}_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

    - This is a **linear regression** since the equation is a linear combination of X and X^2.
  - In general, **centre** and **scale** the explanatory variables:

$$\tilde{x}_i = \frac{x_i - \bar{x}}{\mathrm{sd}(x)}$$

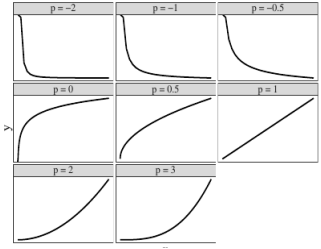
    - numerical accuracy of matrix manipulation is improved, in particular in presence of large values of the covariate
    - β_0 relates the average of y to the average of x, instead of the average of y with $x = 0$ (which is sometimes an impossible value)
    - the slope represents a one standard deviation change which is more meaningful than a one unit change (which can be very small or very large)

2.8 Extension

- **Fractional polynomials**
  - A range of functions can be investigated through fractional polynomials

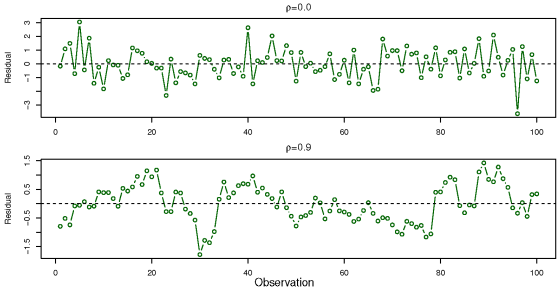
$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i^p \qquad p \neq 0$$

- Test several models ($p = 1$ is linear, $p = 2$ is quadratic, $p = -2$ is reciprocal quadratic) and investigate the best fit. If $p = 0$, use $\log(x_i)$.
- A large number of potential non-linear association can be investigated (Modify both the function and the slope parameter).



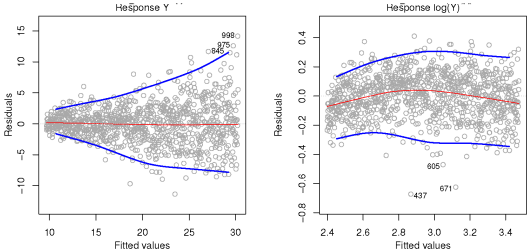
2.9 Potential problems

- **Correlation of the error terms**
  - An important assumption in linear model is that $\varepsilon_1, \ldots, \varepsilon_N$ are **uncorrelated**.
    - If there is correlation, then the estimated standard errors of the coefficients will tend to **underestimate** the true standard errors.
  - **When does it happen**? A classic situation is for time series.
  - Investigate the correlation of errors by plotting the **residuals w.r.t. time**:
    - If errors are uncorrelated, there should be no discernible pattern;
    - If errors terms positively correlated, we may see a trend for adjacent residuals.



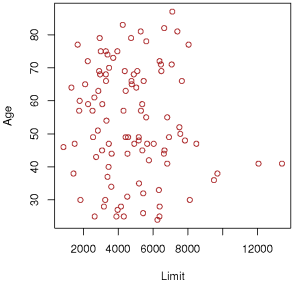
2.9 Potential problems

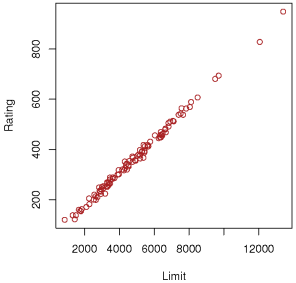
- **Non-constant variance**
  - Another important assumption of the linear model is $\mathbb{V}\text{ar}(\varepsilon_i) = \sigma^2$ for every i.
  - The case of **non-constant variance** is called **heteroscedasticity**.
  - A possible solution is to use a concave transformations, like $\log Y$ or $\sqrt{Y}$, to shrinkage the larger responses.
  - Sometimes we have an idea of the variance of each response: for example, each observation could be an average of n_i observations, there the average can have variance $\sigma_i^2 = \frac{\sigma^2}{n_i}$.
    Solution: Fit weighted least squares, with weights proportional to $w_i = n_i$.



2.9 Potential problems

- **Collinearity**
  - Collinearity occures when two or more predictors are **closely related**.
  - it will be difficult to separate out the individual effects of collinear variables on the response.
  - A small change in the data can cause the coefficient values to be estimated very differently. So, there is a great uncertainty in the estimates.
  - To detect collinearity take a look at the **correlation matrix of the predictors**.





2.9 Potential problems

- **Cont. Collinearity**
  - It is possible that collinearity exists among three or more variables even when no pair of variables has high correlation. This situation is called **multicollinearity**.
  - A way to inspect multicollinearity is the **variance inflation factor**, i.e. the ratio of the variance of $\hat{\beta}_j$ when fitting the full model divided by the variance of $\hat{\beta}_j$ if fit on its own.

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|x_{-j}}} \geq 1 \tag{2.9.1}$$

where $R^2_{x_j|x_{-j}}$ is the R^2 statistic from regression of X_j on all the other predictors.

- If $\text{VIF}(\hat{\beta}_j) = 1$ there is no collinearity
- If $\text{VIF}(\hat{\beta}_j) > 5$ there is a problem

Solutions:

- drop one of the problematic variables
- combine the collinear variables into a single predictor (e.g. taking the average of each pair of predictors)

Code