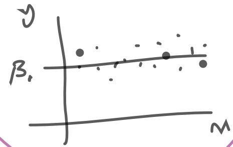
2.3 Hypothesis testing in Linear Models

- **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

$R^2 \approx 0.83$

83%.

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

$$\boxed{R^2} := \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} \qquad (2.3.1)$$

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

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

with Y_i's independent, the deviance is:

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

  - **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.

Deviance M_0

$$\Delta D = D_0 - \overset{M_1}{D_1} = \frac{1}{\sigma^2}\left[(\mathbf{y}^\top \mathbf{y} - \hat{\boldsymbol{\beta}}_0^\top \mathbf{X}_0^\top \mathbf{y}) - (\mathbf{y}^\top \mathbf{y} - \hat{\boldsymbol{\beta}}_1^\top \mathbf{X}_1^\top \mathbf{y})\right] = \frac{1}{\sigma^2}\left[\hat{\boldsymbol{\beta}}_1^\top \mathbf{X}_1^\top \mathbf{y} - \hat{\boldsymbol{\beta}}_0^\top \mathbf{X}_0^\top \mathbf{y}\right]$$

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

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

    - 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 ✓

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

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

M_0 → 1 parameter

M_1 → p parameters

and

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

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

- $\mathbb{E}(\mathrm{RSS}/(N-p)) = \sigma^2$
- under H_0, $\mathbb{E}[(\mathrm{TSS} - \mathrm{RSS})/p] = \sigma^2 0$
- 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

    values of predictors for 1 observation

    - confidence interval for the conditional mean $x^{*\top}\beta$

      $E(Y^*)$

    - prediction interval for a future unobserved observation $\boxed{Y^*} = \underbrace{x^{*\top}\beta}_{} + \boxed{\epsilon^*}$ where ϵ^* is an error independent of ϵ_i, $i = 1, ..., n$, drawn from $N(0, \sigma^2)$.

      error term

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

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

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

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

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.
  - $x^{*\top}\hat{\beta}$ 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,

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

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

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,

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

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

prediction interval

and upon simplifying

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

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) → 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 :

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

- 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 $\underline{\alpha_1 = 0}$.

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

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

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

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

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

  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:

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

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

→ 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 $\underline{\sum_{j=1}^{J} \alpha_j = 0}$, 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

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

columns of X are linearly indep

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

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

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, $\mathbb{E}(Y_{jk}) = \mu$ and the design matrix is a column vector of elements equal to 1, i.e. $\mathbf{X}^\top\mathbf{X} = N$ and $\mathbf{X}^\top\mathbf{y} = Y_{\cdot\cdot}$.

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

and the F-statistic

$$F = \frac{D_0 - D_1}{J-1} \Big/ \frac{D_1}{N-J} \quad \checkmark \sim F_{(J-1,\, N-J)}$$

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

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

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

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

    - 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:

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

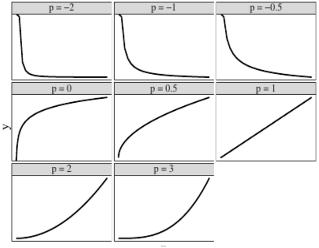
    - numerical accuracy of matrix manipulation is improved, in particular in presence of large values of the covariate
    - β_0 relates the average of y to the average of x, instead of the average of y with $x = 0$ (which is sometimes an impossible value)
    - the slope represents a one standard deviation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

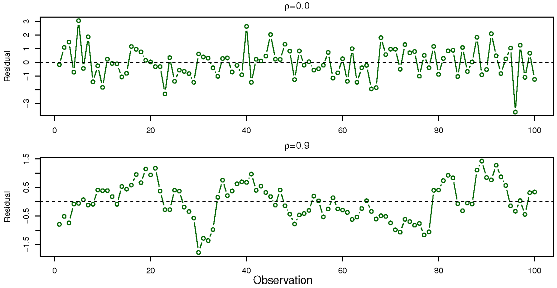
$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i^p \qquad p \neq 0$$

- 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

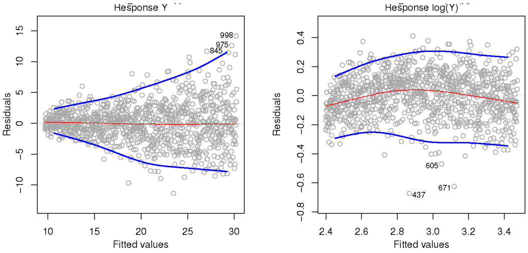
- **Correlation of the error terms**
  - An important assumption in linear model is that $\varepsilon_1, \ldots, \varepsilon_N$ are **uncorrelated**.
    - If there is correlation, then the estimated standard errors of the coefficients will tend to **underestimate** the true standard errors.
  - **When 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

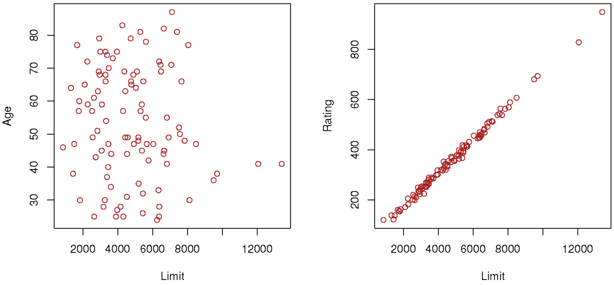
- **Non-constant variance**
  - Another important assumption of the linear model is $\mathbb{V}\text{ar}(\varepsilon_i) = \sigma^2$ for every i.
  - The case of **non-constant variance** is called **heteroscedasticity**.
  - A possible solution is to use a concave transformations, like $\log \text{Y}$ or $\sqrt{\text{Y}}$, 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.

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

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

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