# Statistics: Regression

**Author:** Wenxiao Yang

**Institute:** Department of Mathematics, University of Illinois at Urbana-Champaign

**Date:** 2022

*All models are wrong, but some are useful.*

# Contents

# Chapter 1 Review of statistics

## 1.1 Random Vectors

Suppose a random vector $\vec{Z} \in \mathbb{R}^{m\times 1}$ has mean $\vec{\mu} = \mathbb{E}(\mathbf{Z}) = [\mathbb{E}(Z_1), \mathbb{E}(Z_2), \cdots, \mathbb{E}(Z_m)]^T$ and variance-covariance matrix: $\Sigma_{m\times m} = Cov(\mathbf{Z}) = \mathbb{E}((\mathbf{Z}-\mu)(\mathbf{Z}-\mu)^T) = \begin{bmatrix} Var(Z_1) & \cdots & Cov(Z_1, Z_m) \\ \cdots & \cdots & \cdots \\ Cov(Z_m, Z_1) & \cdots & Var(Z_m) \end{bmatrix}$.

## 1.2 Affine Transformation

(1) The affine transformation $\mathbf{W} = \mathbf{a}_{n\times 1} + \mathbf{B}_{n\times m}\mathbf{Z}_{m\times 1}$ should have following properties:

$$\mathbb{E}(\mathbf{W}) = \mathbf{a} + \mathbf{B}\mu,\ Cov(\mathbf{W}) = \mathbf{B}\Sigma\mathbf{B}^T$$

(2) The affine transformation $\mathbf{W} = \mathbf{v}^T\mathbf{Z} = v_1Z_1 + ... + v_mZ_m$ should have following properties:

$$\mathbb{E}(\mathbf{W}) = \mathbf{v}^T\mu = \sum_{i=1}^{m} v_i\mu_i$$

$$Var(\mathbf{W}) = \mathbf{v}^T\Sigma\mathbf{v} = \sum_{i=1}^{m} v_i^2 Var(Z_i) + 2\sum_{i<j} v_iv_jCov(Z_i, Z_j)$$

$$\text{i.e. } \mathbb{E}(\mathbf{AZ}) = \mathbf{A}\mathbb{E}(Z);\ Var(\mathbf{AZ}) = \mathbf{A}Var(\mathbf{Z})\mathbf{A}^T$$

(3)

$$\begin{aligned} Cov(\mathbf{AX}, \mathbf{BY}) &= \mathbb{E}[(\mathbf{AX} - \mathbf{A}\mathbb{E}(X))(\mathbf{BY} - \mathbf{B}\mathbb{E}(Y))^T] \\ &= \mathbf{A}\mathbb{E}[(\mathbf{X} - \mathbb{E}(X))(\mathbf{Y} - \mathbb{E}(Y))^T]\mathbf{B}^T \\ &= \mathbf{A}Cov(\mathbf{X}, \mathbf{Y})\mathbf{B}^T \end{aligned}$$

# Chapter 2 Regression Analysis (SLR)

## 2.1 Simple Linear Regression

It is a "tool" used to examine the relationship between **one Dependent Variable** or **Response** $y$, and one (or more) **Independent Variables** or **Regressors** or **Predictors** $X$.

$$y = \beta_0 + \beta_1 X$$

where $\beta_0$ is the *intercept*; $\beta_1$ is the *slope*. One *response* $y$; One *predictor* $X$.

$y$ is a random variable that has a distribution for every level of the independent variable. The data come in pairs: $(X, y) = \{(X_i, y_i)\}_{i=1,\dots,n}$.

**Definition 2.1 (Simple Linear Regression Model)**

*Suppose evert data observation can be interpreted as*

$$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

*where the intercept $\beta_0$, the slope $\beta_1$, and the error term $\varepsilon_i$ of observation $(X_i, y_i)$. (intercept $\beta_0$, the slope $\beta_1$, and error variance $\sigma^2$ are the model parameters.)*

♣

### 2.1.1 Assumptions

**Definition 2.2 (Assumptions (errors $\varepsilon$ form))**

*The errors $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ are assumed to*

- *have mean zero: $E(\varepsilon_i) = 0$.*
- *be uncorrelated: $Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$.*
- *be homoscedastic: $Var(\varepsilon_i) = \sigma^2$ does not depend on $i$.*

♣

The last two could be combined and written as: $Cov(\varepsilon_1, \varepsilon_j) = \sigma^2 \delta_{ij}$ where $\delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$

**Definition 2.3 (Assumptions ($Y|X$ form))**

*Based on the SLR model moment assumptions on the error terms, we have the following **equivalent** assumptions for the moments of $Y$ conditioning on $X$:*

- $E(y_i|X_i) = \beta_0 + \beta_1 x_i$
- $Var(y_i|X_i) = \sigma^2$

- $Cov(y_i, y_j|X_i, X_j) = 0,\ i \neq j$

♣

**Claim 2.1 (Interpretation of $\beta_1$, $\beta_0$)**

*$\beta_1$ is the **change in the mean** of the probability distribution function of $y$ per unit change in $x$.*

*$\beta_0$ is the **mean** of the probability distribution function of $y$(at $x = 0$) when $X = \vec{0}$, otherwise $\beta_0$ **has no particular meaning**.*

♠

### 2.1.2 Least Squares Estimations

Suppose the **Residual Sum of Squares (RSS)** is $RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$. We want to find estimates of $\beta_0$, $\beta_1$ to minimize the RSS:

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0,\beta_1)}{\operatorname{argmin}} RSS$$

**Proposition 2.1 (LS Estimators)**

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

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

*Alternative Representation of $\hat{\beta}_1$ is*

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} = r_{xy}\sqrt{\frac{S_{yy}}{S_{xx}}}$$

*where* $S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$, $S_{xx} = \sum_i (x_i - \bar{x})^2$, $S_{yy} = \sum_i (y_i - \bar{y})^2$, *and* $r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$.

♠

**Proof 2.1**

$\frac{\partial RSS}{\partial \beta_0} = 0 \Leftrightarrow -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \Leftrightarrow \beta_0 n + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$

$\frac{\partial RSS}{\partial \beta_1} = 0 \Leftrightarrow -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0 \Leftrightarrow \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$

**Lemma 2.1 (Unbiased Estimators)**

$\mathbb{E}\hat{\beta}_0 = \beta_0$, $\mathbb{E}\hat{\beta}_1 = \beta_1$.

♡

**Proof 2.2**

$\mathbb{E}\hat{\beta}_1 = \mathbb{E}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] = \mathbb{E}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})\beta_1 x_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] = \beta_1$. $\mathbb{E}\hat{\beta}_0 = \mathbb{E}\left[\bar{y} - \hat{\beta}_1 \bar{x}\right] = \beta_0$.

### 2.1.3 Fitted Values & Residuals

**Definition 2.4 (Fitted Values)**

*The Prediction of $y_i$ or the fitted value at $x_i$ is*

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

♣

$$\begin{aligned}\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i \\ &= \bar{y} + \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}(x_i - \bar{x})\end{aligned}$$

**Definition 2.5**

*The $i^{th}$ residual is $r_i = y_i - \hat{y}_i$.*

♣

**Proposition 2.2 (Properties of residuals)**

1. $\sum_i r_i = 0$.
2. $RSS = \sum_i r_i^2$ *is minimized.*
3. $\sum_{i=1} y_i = \sum_{i=1} \hat{y}_i$.
4. $\sum_i x_i r_i = 0$ *prove by first-order condition (proof:* $\sum_i x_i r_i = \sum_i x_i(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})) = \sum_i x_i y_i - n\bar{x}\bar{y} - \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}(\sum_i x_i^2 - n\bar{x}^2) = 0$*) (We can interpret Least Square elimates $Y$'s values in $X$'s direction, the residual should be perpendicular to $X$.)*
5. $\sum_i \hat{y}_i r_i = 0$ *(inferred from 4).*
6. *The regression line always goes through the point* $(\bar{x}, \bar{y})$.

♠

### 2.1.4 Degree of freedom

The **degree of freedom(df)** of the residuals is

$$df = \textit{(Sample size)} - \textit{(\# of parameters)}$$

$df = n - 2$ in this case.

### 2.1.5 (Sample) Error variance

The error variance is estimated by

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_i r_i^2$$

## 2.2 Goodness of Fit: $R-$square

### 2.2.1 TSS, RSS, FSS

$TSS : \sum_i (y_i - \bar{y})^2$

$RSS : \sum_i r_i^2$

$FSS : \sum_i (\hat{y}_i - \bar{y})^2$

$$\begin{aligned}\sum_i (y_i - \bar{y})^2 &= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_i (r_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_i r_i^2 + \sum_i (\hat{y}_i - \bar{y})^2 \\ TSS &= RSS + FSS\end{aligned}$$

### 2.2.2 Cofficient of Determination($R^2$)

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{FSS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$0 \leq R^2 \leq 1$

It measures the effect of $X$ in reducing the variation in $Y$.

The larger $R^2$ is, the more the total variation of $y$ is reduced by reducing the independent variable $x$.

$R^2$ can also represent the degree of linear association between $X$ and $Y$.

$r_{xy} = \pm\sqrt{R^2}$, where the sign is the sign of the slope.

$$\begin{aligned}r_{xy}^2 &= \frac{\left(\sum_i (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2} = \frac{\left(\sum_i (x_i - \bar{x})(r_i + \hat{y}_i - \bar{y})\right)^2}{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2} \\ &= \frac{\left(\sum_i (x_i - \bar{x})(\hat{y}_i - \bar{y})\right)^2}{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2} = \frac{\left(\sum_i \left(\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x}\right)(\hat{y}_i - \bar{y})\right)^2}{\sum_i \left(\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x}\right)^2 \sum_i (y_i - \bar{y})^2} \\ &= \frac{\left(\sum_i (\hat{y}_i - \bar{y})^2\right)^2}{\sum_i (\hat{y}_i - \bar{y})^2 \sum_i (y_i - \bar{y})^2} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = R^2\end{aligned}$$

## 2.3 Affine Transformations

Suppose we have a SLR model of $Y$ on $X$, i.e. $y_i = \beta_0 + \beta_1 x_i$

### 2.3.1 $\tilde{y}_i = ay_i + b$

1. Rescale $y_i$ by $\tilde{y}_i = ay_i + b$ and then regress $\tilde{y}_i$ on $x_i$. How would the $LS$ estimates and $R^2$ be affected?

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(ay_i + b - a\bar{y} - b)}{\sum_{i=1}^n (x_i - \bar{x})^2} = a\hat{\beta}_1$$

$$\tilde{\beta}_0 = a\bar{y} + b - \tilde{\beta}_1\bar{x} = a\hat{\beta}_0 + b$$

$$\tilde{R}^2 = \frac{\sum_i (a\hat{y}_i + b - a\bar{y} - b)^2}{\sum_i (ay_i + b - a\bar{y} - b)^2} = R^2$$

### 2.3.2 $\tilde{x}_i = ax_i + b$

2. Rescale $y_i$ by $\tilde{x}_i = ax_i + b$ and then regress $y_i$ on $\tilde{x}_i$. How would the $LS$ estimates and $R^2$ be affected?

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n (ax_i + b - a\bar{x} - b)(y_i - \bar{y})}{\sum_{i=1}^n (ax_i + b - a\bar{x} - b)^2} = \frac{\hat{\beta}_1}{a}$$

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1(a\bar{x} + b) = \hat{\beta}_0 - \frac{b}{a}\hat{\beta}_1$$

$$\tilde{R}^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = R^2$$

### 2.3.3 Regress $x$ on $y$ instead

3. Regress $x$ on $y$ instead

$$x = \tilde{\beta}_0 + \tilde{\beta}_1 y$$

$$\tilde{\beta}_1 = \frac{S_{xy}}{S_{yy}};\ \tilde{\beta}_0 = \bar{x} - \tilde{\beta}_1\bar{y};\ \tilde{R}^2 = r_{xy}^2 = R^2$$

## 2.4 Regression Through the Origin

$$y_i \approx \beta_1 x_i$$

(1) $\hat{\beta}_1$:

By LS: $\min_{\hat{\beta}_1} RSS = \sum_i (\hat{\beta}_1 x_i - y_i)^2$

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = \sum_i 2x_i(\hat{\beta}_1 x_i - y_i) = 0 \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

(2) $R^2$:

negative $R^2$ is possible since $R^2 = 1 - \frac{RSS}{TSS}$ and $RSS$ may be larger than $TSS$.

We use a modified R-square

$$\sum_i y_i^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i)^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i \hat{y}_i^2$$

$$\tilde{R}^2 = \frac{\sum_i \hat{y}_i^2}{\sum_i y_i^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i y_i^2} = 1 - \frac{RSS}{\sum_i y_i^2}$$

## 2.5 LS Estimators Properties

### 2.5.1 Unbiasedness of LS Estimators $E(\hat{\beta}_1) = \beta_1, E(\hat{\beta}_0) = \beta_0$

$x_i$'s ($\mathcal{X}$) are already known.

$$\begin{aligned}
\mathbb{E}_{\mathcal{Y}}\left(\hat{\beta}_1\right) &= \mathbb{E}\left[\frac{\sum_i (x_i - \bar{x})\, y_i}{\sum_i (x_i - \bar{x})^2}\right] = \frac{\sum_i (x_i - \bar{x}) \cdot \mathbb{E}(y_i)}{\sum_i (x_i - \bar{x})^2} \\
&= \frac{\sum_i (x_i - \bar{x}) \cdot \mathbb{E}(\beta_0 + \beta_1 x_i)}{\sum_i (x_i - \bar{x})^2} = \sum_i c_i (\beta_0 + \beta_1 x_i) \\
&= \beta_0 \sum c_i + \beta_1 \sum c_i x_i = \beta_1 \text{ , where } c_i = \frac{(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left(\hat{\beta}_0\right) &= \mathbb{E}\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) \\
&= \mathbb{E}(\bar{y}) - \bar{x} \cdot \mathbb{E}\left(\hat{\beta}_1\right) = \frac{1}{n}\sum_i \mathbb{E}(y_i) - \bar{x} \cdot \beta_1 \\
&= \frac{1}{n}\sum_i \mathbb{E}(\beta_0 + \beta_1 x_i) - \bar{x} \cdot \beta_1 \\
&= \beta_0 + \bar{x} \cdot \beta_1 - \bar{x} \cdot \beta_1 = \beta_0
\end{aligned}$$

### 2.5.2 Mean squared error(MSE) of LS Estimators $= Var(\hat{\beta}_1) = \sigma^2 \frac{1}{S_{xx}}$, $Var(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$

Mean squared error(MSE)$= \frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2$

Note that since both estimators are unbiased $\Rightarrow$ MSE = Variance.

1. MSE for slope

$$\begin{aligned}
\text{Var}\left(\hat{\beta}_1\right) &= \text{Var}\left[\frac{\sum_i (x_i - \bar{x})\, y_i}{\sum_i (x_i - \bar{x})^2}\right] = \text{Var}\left(\sum_i c_i y_i\right) (c_i \text{ as before }) \\
&= \sum_i c_i^2 \cdot \text{Var}(y_i) = \sum_i c_i^2 \sigma^2 (\text{ from model assumption }) \\
&= \sigma^2 \cdot \left(\frac{\sum_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}\right)^2 = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} = \sigma^2 \frac{1}{S_{xx}}
\end{aligned}$$

2. MSE for intercept

$$\text{Var}\left(\hat{\beta}_0\right) = \text{Var}\left(\bar{y} - \hat{\beta}_1\bar{x}\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

## 2.6 Normality Assumption

Additionally, we assume that

$$\varepsilon_i \sim^{iid} \mathcal{N}(0, \sigma^2)$$

Equivalently,

$$y_i \sim^{iid} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

($y_i$ are a linear shift of the $\varepsilon_i$, so it is also normally distributed)

(The $y_i$'s are jointly normal, and so are linear combinations of the $y_i$'s, since the errors are normally distributed and uncorrelated/independent.)

## 2.7 Distribution of LS Estimators

$\hat{\beta}_1$ and $\hat{\beta}_0$ are jointly normally distributed with

$$\mathbb{E}\left(\hat{\beta}_1\right) = \beta_1 \quad \text{Var}\left(\hat{\beta}_1\right) = \sigma^2\frac{1}{\text{S}_{xx}}$$
$$\mathbb{E}\left(\hat{\beta}_0\right) = \beta_0 \quad \text{Var}\left(\hat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$
$$\text{Cov}\left(\hat{\beta}_1, \hat{\beta}_0\right) = -\sigma^2\frac{\bar{x}}{S_{xx}}$$

$RSS = \sum_i \left(y_i - \hat{y}_i\right)^2 \sim \sigma^2\chi^2_{n-2}$ which implies that

$$\mathbb{E}\left(\hat{\sigma}^2\right) = \mathbb{E}\left(\frac{RSS}{n-2}\right) = \frac{\sigma^2(n-2)}{n-2} = \sigma^2$$

$\left(\hat{\beta}_0, \hat{\beta}_1\right)$ and RSS are independent.

## 2.8 Hypothesis Testing (T-test)

### 2.8.1 Testing for the Slope

$$\begin{cases} H_0 : \beta_1 = c(null) \\ H_\alpha : \beta_1 \neq c \text{ (alternative)} \end{cases}$$

where $c$ is an known constant. The test statistics is

$$t = \frac{\hat{\beta}_1 - c}{\sqrt{\text{Var}\left(\hat{\beta}_1\right)}} = \frac{\hat{\beta}_1 - c}{\hat{\sigma}/\sqrt{S_{xx}}}$$

The distribution of $t$ under the null is $T_{n-2}$. The $p$-value is twice the area under the $T_{n-2}$ distribution more extreme than the observed statistic $t$.

### 2.8.2 Testing for the Intercept

$$\begin{cases} H_0 : \beta_0 = c(null) \\ H_\alpha : \beta_0 \neq c \text{ (alternative)} \end{cases}$$

The test statistics is

$$t = \frac{\hat{\beta}_0 - c}{\sqrt{\text{Var}\left(\hat{\beta}_0\right)}}$$

The distribution of $t$ under the null is $T_{n-2}$. The $p$-value is twice the area under the $T_{n-2}$ distribution more extreme than the observed statistic $t$.

## 2.9 ANOVA Table & F-Test

### 2.9.1 Degrees of Freedom

$df_{TSS} = n - 1$ : one $df$ is lost, because the sample mean is used to estimate the population mean.

$df_{RSS} = n - 2$ : two $df$ are lost, because the two parameters are estimated in obtaining the fitted values $\hat{y}$

$df_{FSS} = 1$ : there are $n$ deviations $\hat{y}_i - \bar{y}$, but all the fitted values are associated with the same regression line.

$$df_{TSS} = df_{RSS} + df_{FSS}$$

| Sum of Squares | Expression | $df$ |
|---|---|---|
| TSS | $\sum_i (y_i - \bar{y})^2$ | $n - 1$ |
| FSS | $\sum_i (\hat{y}_i - \bar{y})^2$ | 1 |
| RSS | $\sum_i (y_i - \hat{y}_i)^2$ | $n - 2$ |

### 2.9.2 ANOVA Table

| Source | SS | df | MS | $F$ |
|---|---|---|---|---|
| Regression (model) | FSS | 1 | MSReg $= \frac{FSS}{1}$ | $F = \frac{MSReg}{MSE}$ |
| Error | RSS | $n-2$ | $MSE = \frac{RSS}{n-2}$ | |
| Total | TSS | $n-1$ | | |

**Figure 2.1**

### 2.9.3 F-Test (equivalent to t-test)

An alternative way to test for the model parameters is using the $F$ test:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_\alpha : \beta_1 \neq 0 \end{cases}$$

- Under $H_0$, the $F$-test statistic is

$$F = \frac{\text{MSReg}}{MSE} = \frac{FSS}{RSS/(n-2)} \sim F_{1,n-2}$$

- It can be shown that the $F$-test statistic is equal to the square of the $t$-test statistic and their $p$-values are the same. So, **this test is equivalent to the $t$-test before**.

## 2.10 Estimation and Prediction

### 2.10.1 Estimation (always reported for a parameter: $\beta_0 + \beta_1 x^* = \mathbb{E}(Y|x^*)$)

1. Estimation: We want to estimate the mean response at $x^*$. This is equivalent to estimate: $\beta_0 + \beta_1 x^*$

2. Accuracy of the estimation: is measured by the expected value of the squared difference between the point estimate and the target.

- For estimation the target is $\beta_0 + \beta_1 x^\star$ :

$$\begin{aligned}
&\mathbb{E}\left(\hat{\beta}_0 + \hat{\beta}_1 x^\star - \beta_0 - \beta_1 x^\star\right)^2 \\
=&\operatorname{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x^\star\right) \\
=&\operatorname{Var}\left(\hat{\beta}_0\right) + (x^\star)^2 \operatorname{Var}\left(\hat{\beta}_1\right) + 2x^\star \operatorname{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) \\
=&\sigma^2\left(\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right) \\
=&\sigma^2\left(\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{xx}}\right)
\end{aligned}$$

3. Confidence interval: An $(1-\alpha)100\%$ Confidence Interval for the Mean Response when $x = x^\star$ is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x^\star \pm T_{n-2}(\alpha/2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{xx}}}$$

### 2.10.2 Prediction (is reported for the value of a random variable $Y^*$)

1. Prediction: of an outcome of random variable $Y^*$ at a given value $x^*$ , where $Y^* \sim N(\beta_0 + \beta_1 x^*, \sigma^2)$

2. For prediction the target is $Y^\star = \beta_0 + \beta_1 x^\star + e^\star$, where $e^\star \sim N\left(0, \sigma^2\right)$ This new error $e^\star$ is independent of the previous $n$ data points, i.e. is independent of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{\beta}_0 + \hat{\beta}_1 x^\star - Y^\star\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\beta}_0 + \hat{\beta}_1 x^\star - \beta_0 - \beta_1 x^\star - e^\star\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\beta}_0 + \hat{\beta}_1 x^\star - \beta_0 - \beta_1 x^\star\right)^2\right] + \mathbb{E}\left[(e^\star)^2\right] \\
&= \sigma^2\left(1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{xx}}\right)
\end{aligned}$$

3. Prediction interval: An $(1-\alpha)100\%$ Prediction Interval for $\hat{Y}^\star$ when $x = x^\star$ is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x^\star \pm T_{n-2}(\alpha/2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{S_{xx}}}$$

### 2.10.3 Simultaneous Confidence

$\mu^* = \mathbb{E}[y|x^*] = \beta_0 + \beta_1 x^*$'s Confidence Interval:

$$I(x^*) = (\hat{\mu}^* \pm T_{n-2}(\frac{\alpha}{2}) se(\hat{\mu}^*))$$

Where

$$\hat{\mu}^* = \hat{\beta}_0 + \hat{\beta}_1 x^* \text{ and } se(\hat{\mu}^*) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

If we want confidence intervals at multiple points $(x_1^*, x_2^*, \cdots, x_m^*)$, we can use formula $(1)$ to have confidence intervals at the $m$ points: $I(x_1^*), I(x_2^*), \cdots, I(x_m^*)$.

We know that:

$$\mathbb{P}\left(\mu_i^* \in I\left(x_i^*\right)\right) = (1-\alpha)$$

This is the point-wise coverage probability for $\mu_i^*$ and formula (1) gives the point-wise CI.

What about the simultaneous coverage probability? i.e.:

$$\mathbb{P}\left(\mu_i^* \in I\left(x_i^*\right), \text{ for } i = 1, \ldots, m\right) = ?$$

To make sure that (for example):

$$\mathbb{P}\left(\mu_i^* \in I\left(x_i^*\right), \text{ for } i = 1, \ldots, m\right) = .95$$

we need to set $\alpha = 5\%/m$, which is known as the **Bonferroni correction**.

Let $\mathrm{A}_k$ denotes the event that the $k$ th confidence interval covers $\mu_k^*$ with:

$$\mathbb{P}\left(\mathrm{A}_k\right) = (1 - \alpha)$$

Then we can show:

$$\begin{aligned}
&\mathbb{P}\left(\text{ All CIs cover the corresponding } \mu_k^* \text{ values }\right) \\
&= \mathbb{P}\left(\mathrm{A}_1 \cap \mathrm{A}_2 \ldots \cap \mathrm{A}_m\right) \\
&= 1 - \mathbb{P}\left(\mathrm{A}_1^c \cup \mathrm{A}_2^c \ldots \cup \mathrm{A}_m^c\right) \\
&\geq 1 - \mathbb{P}\left(\mathrm{A}_1^c\right) - \ldots - \mathbb{P}\left(\mathrm{A}_m^c\right) \\
&= 1 - m\alpha
\end{aligned}$$

If we choose $\alpha/m$ instead of $\alpha$, the simultaneous coverage probability will be $(1 - \alpha)$

### 2.10.4 Confidence Band (Larger than CI)

Ideally we would like to construct a simultaneous confidence band (i.e., $m = \infty$ ) across all $x^*$ 's. (Scheffé's Theorem - 2959 ). Let

$$I(x) = (\hat{r}(x) - c\hat{\sigma}, \hat{r}(x) + c\hat{\sigma})$$

where

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x, c\hat{\sigma} = \sqrt{2F(\alpha, 2, n-2)}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Then,

$$\mathbb{P}(r(x) \in I(x) \text{ for all } x) \geq 1 - \alpha$$

Can we construct a simultaneous prediction band? No!

Are confidence bands always wider than point-wise confidence intervals? Yes! For SLR, at a location $x^*$, we have

$$\begin{aligned}
\text{band} &: \hat{\mu}^* \pm \sqrt{2F(\alpha, 2, n-2)}\,\mathrm{se}\left(\hat{\mu}^*\right) \\
\text{interval} &: \hat{\mu}^* \pm T_{n-2}(\alpha/2)\,\mathrm{se}\left(\hat{\mu}^*\right)
\end{aligned}$$

$$\sqrt{2F(\alpha, 2, n-2)} > T_{n-2}(\alpha/2) = \sqrt{2F(\alpha, 1, n-2)}$$

In fact, for any $\alpha$, we have

$$T_m(\alpha/2) = \sqrt{2F(\alpha, 1, m)} < \sqrt{kF(\alpha, k, m)}$$

## 2.11 Maximum likelihood estimators with normal error terms

We start with the statistical model, which is the Gaussian-noise simple linear regression model, defined as follows:

1. The distribution of $X$ is arbitrary (and perhaps $X$ is even non-random).

2. If $X = x$, then $Y = \beta_0 + \beta_1 x + \epsilon$, for some constants ("coefficients", "parameters") $\beta_0$ and $\beta_1$, and some random noise variable $\epsilon$.

3. $\epsilon \sim N\left(0, \sigma^2\right)$, and is independent of $X$.

4. $\epsilon$ is independent across observations.

$$p(y_i|x_i; \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma})^2}$$

Given any data set $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$, we can now write down the probability density, under the model, of seeing that data:

$$\prod_{i=1}^{n} p(y_i|x_i; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma})^2} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}}$$

Take the **log-likelihood**

$$L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

Then we can compute the **Maximum likelihood estimators** $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)$:

(1) $(\hat{\beta}_0, \hat{\beta}_1)$,

Obviously, maximizing $L(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)$ is as same as minimizing $\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$, then the **Maximum likelihood estimators** is exactly the **LS estimators**:

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

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

(2) $\hat{\sigma}^2$,

And the $\hat{\sigma}^2$ is exactly the in-sample mean squared error:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

# Chapter 3 Multiple Linear Regression (MLR): Basic

## 3.1 Basic

$x_1, x_2, \ldots, x_p$ be $p$ predictors of a response $y$.

The data will be of the form:

$$\begin{matrix} y_1 & x_{11} & x_{12} & \cdots & x_{1p} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_n & x_{n1} & x_{n2} & \cdots & x_{np} \end{matrix}$$

Model Equation:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n$$

where we denote $\mathbf{x_i} = (x_{i1,\ldots,x_ip})^T$, with $x_{i1} = 1$

$\left(\beta_1, \beta_2, \ldots, \beta_p; \sigma^2\right)$ are unknown true parameters.

$\beta_1$ is the intercept.

$\beta_2, \beta_3, \ldots, \beta_p$ are partial slopes.

$\sigma^2$ is the error variance

### 3.1.1 Assumptions of errors

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are the random errors. They usually assumed to satisfy the same conditions as in simple linear regression:

- zero mean: $\mathbb{E}\left(\varepsilon_i\right) = 0$

- uncorrelated: $\mathrm{Cov}\left(\varepsilon_\mathrm{i}, \varepsilon_\mathrm{j}\right) = 0, \mathrm{i} \neq \mathrm{j})$, and

- homoscedastic: $\mathrm{Var}\left(\varepsilon_\mathrm{i}\right) = \sigma^2$ does not depend on $i$ ).

### 3.1.2 Matrix Representation

Matrix Representation of the MLR Model:

$$\begin{matrix} \mathbf{y}_{n\times 1} & = & \mathbf{X}_{n\times p} & \beta_{p\times 1} & + & \varepsilon_{n\times 1} \\ \uparrow & & \uparrow & \uparrow & & \uparrow \\ \text{Response} & & \text{Design Matrix} & \text{Coefficients} & & \text{Error Term} \end{matrix}$$

- $n$ : sample size

- $p$ : number of predictors or columns of $X$

- By default the intercept is included in the model in which case the first column of $X$ is a vector of 1 's.

We set $\mathbb{E}(\varepsilon) = 0$ and $Cov(\varepsilon) = \sigma^2\mathbf{I}_n$, then we can infer that

$$\mathbb{E}(y) = \mathbf{X}\beta,\ Cov(y) = \sigma^2\mathbf{I}_n$$

## 3.2 Parameter Estimation $\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y$

- We want to estimate $\beta$, i.e. obtain:

$$\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p\right)^T$$

- The LS estimator of $\beta$ minimizes the sum of squared residuals:

$$RSS = \|y - \mathbf{X}\beta\|^2 = (y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta)$$

In order to minimize $RSS = (y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta)$, we take derivatives with respect to $\beta$ 's and set to zero (as before).

$$\frac{\partial RSS}{\partial \beta} = -2\mathbf{X}^T_{p\times n}(y - \mathbf{X}\beta)_{n\times 1} = \mathbf{0}_{p\times 1}$$

$\mathbf{X}^T(y - \mathbf{X}\beta) = \mathbf{0} \longrightarrow$ Normal Equations

$\left(\mathbf{X}^T\mathbf{X}\right)\beta = \mathbf{X}^T y$

$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y \rightarrow$ LS Estimators

**Remarks**

1. We assume that the rank of $X$ is $p$, i.e. no columns of $X$ is a linear combinations of the other columns of $X$.
2. Since $\mathbf{X}$ has rank $p$, the inverse of $\left(\mathbf{X}^T\mathbf{X}\right)$ exists.
3. if $\left(\mathbf{X}^T\mathbf{X}\right)$ is singular the $LS$ solutions is not unique (identifiability problem)

### 3.2.1

Fitted Values $\hat{y}_{n\times 1} = \mathbf{H}_{n\times n}y_{n\times 1}$

$$\begin{aligned}\hat{y}_{n\times 1} &= \mathbf{X}\hat{\beta} \\ &= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y \\ &= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^\top y \\ &= \mathbf{H}_{n\times n}y_{n\times 1}\end{aligned}$$

$\mathbf{H}_{n\times n} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^\top$ is called the hat matrix, since it returns the "y-hat" values.

## 3.3

Residuals $\mathbf{r} = (\mathbf{I} - \mathbf{H})y$, (Sample) error variance $\hat{\sigma}^2 = \frac{\mathbf{r}^T\mathbf{r}}{n-p}$

$$\mathbf{r}_{n\times 1} = y - \hat{y} = y - \mathbf{H}y = (\mathbf{I} - \mathbf{H})y$$

The residuals $\mathbf{r}$ are used to estimate the error variance:

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_i r_i^2 = \frac{\mathbf{r}^T\mathbf{r}}{n-p} = \frac{RSS}{n-p}$$

## 3.4 Properties of residuals

The LS estimator is the $\beta$ that satisfies the normal equations, that is

$$\mathbf{X}^T(y - \hat{y}) = \mathbf{X}^T(y - \mathbf{X}\hat{\beta}) = \mathbf{0}$$

This implies the following properties for the residuals, $r_{n\times 1} = y - \mathbf{X}\hat{\beta}$ :

### 3.4.1

$\mathbf{X}^T\boldsymbol{r} = 0$ 1. The cross-products between the residual vector $r$ and each column of $\mathbf{X}$ are zero, i.e.

$$\begin{aligned}\mathbf{X}^T\boldsymbol{r} &= \mathbf{X}^T(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}^T\boldsymbol{y} - \mathbf{X}^T\mathbf{X}\hat{\beta} \\ &= \mathbf{X}^T\boldsymbol{y} - \left(\mathbf{X}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{y} = 0\end{aligned}$$

### 3.4.2

$\hat{y}^T\boldsymbol{r} = 0$ 2. The cross-product between the fitted value $\hat{y}$ and the residual vector $r$ is zero, i.e.

$$\hat{y}^T r = \hat{\beta}^T X^T r = 0$$

This implies that the residual vector $r$ is **orthogonal** to each column of $X$ and $\hat{y}$.

## 3.5 Properties of H

### 3.5.1 HX = X

Let $c$ be any linear combination of the columns of $\mathbb{X}$, then

$$\mathbf{H}c = c$$

since $\mathbf{HX} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{X}$

### 3.5.2 Symmetric: $\mathbf{H}^T = \mathbf{H}$

**Symmetric**, since $\mathbf{H}^T = \left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{H}$

### 3.5.3 Idempotent: $\mathbf{HH} = \mathbf{HH}^T = \mathbf{H}^\top\mathbf{H} = \mathbf{H}$

**Idempotent**, i.e. $\mathbf{HH} = \mathbf{HH}^T = \mathbf{H}^\top\mathbf{H} = \mathbf{H}$. Indeed

$$\mathbf{HH} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{H}$$

### 3.5.4 $\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{0}$

This also implies that $\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{0}_{n\times n}$

### 3.5.5 $(\mathbf{I} - \mathbf{H})$ is also symmetric and idempotent

$$(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^T = \mathbf{I} - \mathbf{H}$$

### 3.5.6 $\mathrm{trace}(\mathbf{H}) = \mathrm{p}$

$\mathrm{trace}(\mathbf{H}) = \mathrm{p}$, the number of LS coefficients we estimated.

## 3.6 Geometric Representation of LS

### 3.6.1 Estimation Space

- The columns of $\mathbf{X}$ span a $p$-dimensional subspace in $\mathbb{R}^n$. This is a subspace that consists of vectors that can be written as linear combinations of the columns of $X$.

- The LS squares estimator $\hat{\beta}$ is obtained by minimizing the Euclidean distance between the vectors $\mathbf{y}$ and $\hat{\mathbf{y}}$, i.e. $\|y - \hat{y}\|^2$. $\hat{y}$ is the projection of $y$ onto the estimation space.

- $\mathbf{H}_{n\times n}$, projection/hat matrix is symmetric, unique, and idempotent.

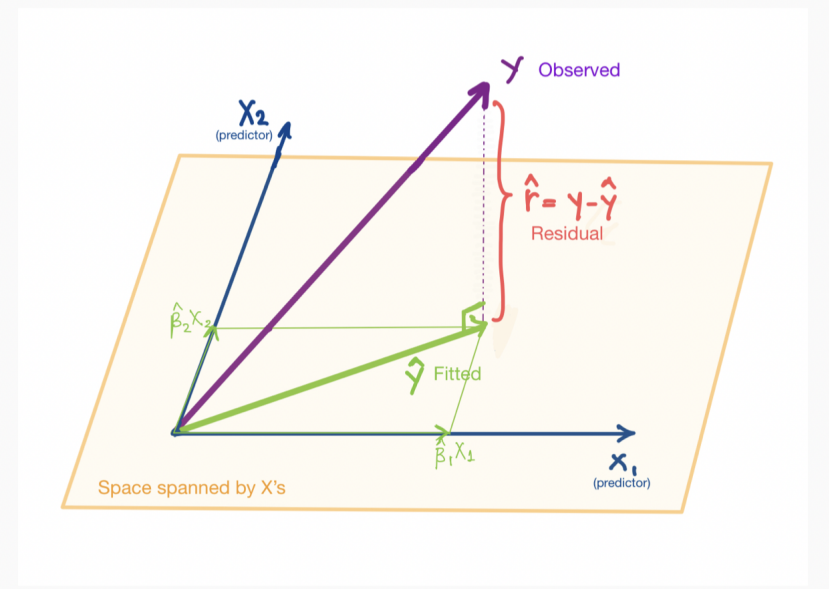


**Figure 3.1**

### 3.6.2 Error Space

- The error space is an $(n-p)$-dimensional space that is orthogonal to the estimation space. The projection matrix of the error space is $(\mathbf{I}-\mathbf{H})$.

- The residual $\mathbf{r}$ is the projection of $\boldsymbol{y}$ onto the error space, orthogonal to the estimation space. So, $\mathbf{r}$ is orthogonal to any vector in the estimation space, including each column of $X$.

- When the intercept is included in the model, then

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

In general, $\sum_{i=1}^{n} r_i X_{ij} = 0, j = 1, \ldots, p$ due to the normal equations.

## 3.7 Coefficient of determination, R-Square

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$$

An equivalent definition is

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

$0 \leq R^2 \leq 1$

## 3.8 Properties of LS Estimators

### 3.8.1 Unbiased:

$\mathbb{E}(\hat{\beta}) = \beta$

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbb{E}(y)$$
$$= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\beta = \beta$$

### 3.8.2

Variance-Covariance Matrix of $\hat{\beta}$: $\mathrm{Cov}(\hat{\beta}) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$

$$\mathrm{Cov}(\hat{\beta}) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\,\mathrm{Cov}(\mathbf{y})\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right)^\top$$
$$= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\sigma^2\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$$
$$= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$
$$= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$
$$se(\hat{\beta}_i) = \hat{\sigma}\sqrt{((\mathbf{X}^T\mathbf{X})^{-1})_{ii}}$$

### 3.8.3

$\hat{y}$: $\mathbb{E}(\hat{y}) = \mathbf{X}\beta$, $Cov(\hat{y}) = \sigma^2\mathbf{H}$

$$\mathbb{E}(\hat{y}) = \mathbb{E}(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^\top y) = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\beta = \mathbf{X}\beta$$

$$Cov(\hat{y}) = Cov(\mathbf{H}y) = \mathbf{H}Cov(y)\mathbf{H}^T = \sigma^2\mathbf{H}\mathbf{H}^T = \sigma^2\mathbf{H}$$

### 3.8.4

$\mathbf{r}$: $\mathbb{E}(\mathbf{r}) = 0$, $Cov(\mathbf{r}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$

$$\mathbb{E}(\mathbf{r}) = \mathbb{E}(y - \hat{y}) = 0$$

$$Cov(\mathbf{r}) = Cov(y - \hat{y}) = Cov((\mathbf{I} - \mathbf{H})y) = \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^T = \sigma^2(\mathbf{I} - \mathbf{H})$$

### 3.8.5

$\mathbb{E}(\hat{\sigma}^2) = \sigma^2$

$$\mathbb{E}(\hat{\sigma}^2) = \frac{1}{n-p}\mathbb{E}(\mathbf{r}^T\mathbf{r}) = \frac{1}{n-p}\sigma^2(n-p) = \sigma^2$$

*$n-p$: there are $p$ elements in the diagonal of $H$ are $1$ and others are $0$, so the diagonal of $\mathbf{I}_n - \mathbf{H}$ has $n-p$ 1. Then there are only $n-p$ $r_i^2 = \sigma^2$.*

$$\frac{\mathbf{r}^T\mathbf{r}}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$$

## 3.9 The Gauss-Markov Theorem: LS estimator is the BLUE(best linear unbiased estimator)

If the errors are 1. Mean zero, 2. umcorrelated, 3. homoscedastic, the LS estimators have **the lowest variance within the class of linear estimators.**

Suppose we are interested in estimating a linear combination of $\beta$ of the form:

$$\theta = \mathbf{c}^T\beta = \sum_{j=1}^{p} c_j\beta_j$$

LS estimators:

$$\hat{\theta}_{LS} = \mathbf{c}^T\hat{\beta} = \mathbf{c}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y$$

This is a **linear** (linear combination of $y_1, y_2, ..., y_n$) and **unbiased estimator** of $\theta$. Its mean square error can be calculated as:

$$MSE(\hat{\theta}_{LS}) = \mathbb{E}(\hat{\theta}_{LS} - \theta)^2 = Var(\hat{\theta}_{LS})$$

**Theorem 3.1 (Gauss-Markov Theorem)**

*$\hat{\theta}_{LS} = \mathbf{c}^T\hat{\beta}$ is the **BLUE**(best linear unbiased estimator) of the parameter $\mathbf{c}^T\beta$ for any vector $\mathbf{c} \in \mathbb{R}^p$* ♡

## 3.10 Maximum Likelihood Estimation, Distribution of LS estimates

### 3.10.1

$\mathbf{y} \sim \mathbf{N}_n\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right)$

Recall the normality assumption for the regression model:

$$y_i = \mathbf{x}_i^T\beta + \varepsilon_i, \quad i = 1, \ldots, n, \text{ with } \varepsilon_i \sim N\left(0, \sigma^2\right)$$

This implies that $\mathbf{y} \sim \mathbf{N}_n\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right)$

### 3.10.2 LS estimator is the Maximum Likelihood Estimator (MLE)

We can show that the likelihood function can be written as:

$$L\left(\beta, \sigma^2 \mid \mathbf{y}\right) \propto \frac{RSS}{n}^{-\frac{n}{2}}$$

The value of $\beta$ that maximizes the Likelihood function is *the Maximum Likelihood Estimator (MLE)* of $\beta$. This estimator is equal to the LS estimate of $\beta$.

### 3.10.3

$\hat{\beta} \sim \mathbf{N_p}\left(\beta, \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right)$, $\hat{\mathbf{y}} \sim \mathbf{N_n}\left(\mathbf{X}\beta, \sigma^2\mathbf{H}\right)$, $\hat{\mathbf{r}} \sim \mathbf{N_n}\left(\mathbf{0}, \sigma^2\left(\mathbf{I_n} - \mathbf{H}\right)\right)$

Recall the assumption for the linear regression model:

$$\mathbf{y} \sim \mathbf{N_n}\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right)$$

Any affine transformation of $\mathbf{y}$ will also have a Normal distribution [2]. We can show that:

$$\hat{\beta} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} \sim \mathbf{N_p}\left(\beta, \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right)$$

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \sim \mathbf{N_n}\left(\mathbf{X}\beta, \sigma^2\mathbf{H}\right)$$

$$\hat{\mathbf{r}} = \left(\mathbf{I_n} - \mathbf{H}\right)\mathbf{y} \sim \mathbf{N_n}\left(\mathbf{0}, \sigma^2\left(\mathbf{I_n} - \mathbf{H}\right)\right)$$

## 3.11 r Residuals' Properties

### 3.11.1

$\mathbf{r} \in \mathbb{R}^{n-p}$ Although $\mathbf{r}$ is a vector of dimension $n$, it always lies in a subspace of dimension $(n-p)$.

### 3.11.2

$\hat{\sigma}^2 \sim \sigma^2\frac{\chi^2_{n-p}}{n-p}$ r behaves like a random vector with a distribution $\mathbf{N}_{n-p}\left(\mathbf{0}, \sigma^2\mathbf{I}_{n-p}\right)$, so we have:

$$\hat{\sigma}^2 = \frac{\|\mathbf{r}\|^2}{n-p} \sim \sigma^2\frac{\chi^2_{n-p}}{n-p}$$

### 3.11.3

$\hat{\mathbf{y}}$ and $\boldsymbol{r}$ are independent It can be show that $\hat{\mathbf{y}}$ and $\boldsymbol{r}$ are uncorrelated since they are in orthogonal spaces. Since they also have a joint normal distribution, they are independent.

## 3.12 Testing Predictors (Coefficients)

### 3.12.1 Testing a Single Predictor $H_0 : \beta_j = 0$: $t-$test

Suppose you have a $p$ predictors in your regression model and you want to test the hypothesis:

$$H_0 : \beta_j = c \text{ vs. } H_\alpha : \beta_j \neq c$$

- The t-test statistic we use is:

$$t = \frac{\hat{\beta}_j - c}{se\left(\hat{\beta}_j\right)} = \frac{\hat{\beta}_j - c}{\hat{\sigma}\sqrt{\left[(\mathbf{X}^\top \mathbf{X})^{-1}\right]_{jj}}} \sim T_{n-p}$$

under the null hypothesis $H_0$.

- p-value $= 2\times$ the area under the curve of a $T_{n-p}$ distribution more extreme than the observed statistic.

- The $p$-value returned by the $Im$ function command is for $c = 0$.

### 3.12.2 Review the degree of freedom

The degrees of freedom of a $t$-test are determined by the denominator of the estimated variance $\hat{\sigma}^2$. Consider the following situations:

- In STAT 400: Test for $\theta = \alpha$, where $Z_1, \ldots, Z_n \sim \mathcal{N}\left(\theta, \sigma^2\right)$

$$\frac{\hat{\theta} - \alpha}{\text{se}(\hat{\theta})} = \frac{\overline{\mathbf{Z}} - \alpha}{\sqrt{\hat{\sigma}^2/n}} \sim T_{n-1}, \quad \hat{\sigma}^2 = \frac{\sum_i \left(Z_i - \bar{Z}\right)^2}{n-1}$$

- In SLR: Test for $\beta_1 = c$, we have

$$\frac{\hat{\beta}_1 - c}{\text{se}\left(\hat{\beta}_1\right)} = \frac{\hat{\beta}_1 - c}{\hat{\sigma}/\sqrt{S_{XX}}} \sim T_{n-2}, \quad \hat{\sigma}^2 = \frac{RSS}{n-2}$$

- In $MLR$ with $p$ predictors (including the intercept): Test for $\beta_j = c$,

$$\frac{\hat{\beta}_j - c}{\text{se}\left(\hat{\beta}_j\right)} = \frac{\hat{\beta}_j - c}{\hat{\sigma}\sqrt{\left[(\mathbf{X}^T \mathbf{X})^{-1}\right]_{jj}}} \sim T_{n-p}, \quad \hat{\sigma}^2 = \frac{RSS}{n-p}$$

### 3.12.3 Testing all predictors: $F-$test

Testing all predictors

$$\begin{cases} H_0 : \beta_2 = \beta_3 = \ldots = \beta_p = 0 \\ H_a : \beta_j \neq 0, \quad \text{for some } j, j = 2, \ldots, p \end{cases}$$

- Under the Null hypothesis, the test statistic:

$$F = \frac{FSS\left(X_2, \ldots, X_p\right)}{p-1} \div \frac{RSS\left(X_2, \ldots, X_p\right)}{n-p}$$
$$= \frac{MS(Reg)}{MS(\text{ Error })} \sim F_{p-1, n-p}$$

Large values of $F$ lead to conclusion $H_\alpha$.

- This is the overall $F$ test of whether or not there is a regression relation between the response variable $Y$ and the set of $X$ variables.

| Source | df | SS | MS | F-test |
|---|---|---|---|---|
| *Regression* | $p-1$ | FSS | FSS/$(p-1)$ | MS(Reg)/MSE |
| *Error* | $n-p$ | RSS | RSS/$(n-p)$ | |
| *Total* | $n-1$ | TSS | | |

**Figure 3.2**

### 3.12.4 Partial $F-$test

In general, consider the following partition of the design matrix into two sub-matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, that is

$$\mathbf{X}_{n\times p} = \left(\mathbf{X}_{1n\times(p-q)}, \mathbf{X}_{2n\times q}\right)$$

The corresponding partition of the regression parameter is:

$$\beta^T = \left(\beta_1^T, \beta_2^T\right)$$

where $\beta_1$ is $(p-q)\times 1$ and $\beta_2$ is $q\times 1$

This partition is used to test the hypothesis:

$$\begin{cases} H_0 : \beta_2 = \mathbf{0}, \text{ i.e.,} & \mathbf{y} = \mathbf{X}_1\beta_1 + \text{ error} \\ H_\alpha : \beta_2 \neq \mathbf{0}, \text{ i.e.,} & \mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \text{ error} \end{cases}$$

To test this hypothesis, the test statistic is:

$$F = \frac{\left(RSS_0 - RSS_\alpha\right)/q}{RSS_\alpha/(n-p)} \sim F_{q,n-p}$$

where $RRS_0$ = **Residual sum of squares for the model under** $H_0$ **;** $RRS_a$ = **Residual sum of squares for the model under** $H_\alpha$

**Numerator**: variation in the data not explained by the reduced model, but explained by the full model.

**Denominator**: variation in the data not explained by the full model (i.e., not explained by either model), which is used to estimate the error variance.

*Reject* $H_0$*, if* $F$ *test statistic is large*, that is, the variation missed by the reduced model, when being compared with the error variance, is significantly large.

F-test statistic can be rewirtten:

$$F = \frac{(RSS_0 - RSS_\alpha)/q}{RSS_\alpha/(n-p)} = \frac{((1-R_R^2)-(1-R_F^2))/(df_F - df_R)}{(1-R_F^2)/df_F} = \frac{(R_F^2 - R_R^2)/(df_R - df_F)}{(1-R_F^2)/df_F}$$

Note that this test statistic is not appropriate when the full and reduced regression models do not contain the intercept term $\beta_0$ .

## 3.13 Permutation Tests (When the normal distribution hypothesis doesn't hold)

The distribution of the data is *unknown*. - A test statistic is a function of the data; denote it $g($ data $)$.

- The test statistic tends to take extreme values under the alternative hypothesis $H_\alpha$.

### 3.13.1 Procedure

Procedure to conduct a permutation test

1. Form the test statistic $g($ data $)$ which tends to take extreme values under the alternative hypothesis.
2. Evaluate the test statistic on the observed data, denoted by $g_0$.
3. Find the distribution of $g($ data $)$, when data are generated from $H_0$.
4. Calculate the $p$-value, that is the following probability:

$$\mathbb{P}\left(g(\text{ data }) \text{ is more extreme than the observed } g_0 \mid \text{data} \sim H_0\right)$$

### 3.13.2 Calculation of the p-value: Monte Carlo method

We can obtain an approximation of $\mathbb{E}(Y)$ as follows:

1. Generate $N = 1000$ samples from this distribution, $Y_1, \ldots, Y_N$
2. Approximate the mean by

$$\mathbb{E}(Y) \approx \frac{1}{N}\sum_{i=1}^{N} Y_i$$

That is, population mean $\approx$ sample mean (when $N$ is large).

This method also works if we want to approximate the expected value of a *function* of a random variable:

$$\mathbb{E}(f(Y)) \approx \frac{1}{N}\sum_{i=1}^{N} f(Y_i)$$

## 3.14 Confidence Intervals for $\beta_j$, Confidence Region for $\beta$

### 3.14.1 Confidence Intervals for $\beta_j$

Recall that

$$\hat{\beta} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} \sim \mathbf{N_p}\left(\beta, \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right)$$

An $(1-\alpha)100\%$ CI(Confidence Interval) for $\beta_j$ can be written as

$$(\hat{\beta}_j \pm T_{n-p}(\frac{\alpha}{2})se(\hat{\beta}_j)) = (\hat{\beta}_j \pm T_{n-p}(\frac{\alpha}{2})\hat{\sigma}\sqrt{[(\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}})$$

Justification: The vector $\beta$'s confidence interval is a family/joint interval for all the betas, so it will be wider than the individual $\beta_j$'s intervals.

### 3.14.2 Confidence Region for $\beta$

$\beta$ is the entire vector,

$$\beta - \hat{\beta} \sim \mathbf{N_p}\left(0, \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right)$$

Thus the quadratic form:

$$\frac{(\beta-\hat{\beta})^T\mathbf{X}^T\mathbf{X}(\beta-\hat{\beta})}{p\hat{\sigma}^2} \sim F_{p,n-p}$$

We can construct a $(1-\alpha)100\%$ confidence region for $\beta$ to be all the points in the following ellipsoid,

$$\frac{(\beta-\hat{\beta})^T\mathbf{X}^T\mathbf{X}(\beta-\hat{\beta})}{p\hat{\sigma}^2} < F(\alpha; p, n-p)$$

Where $F(\alpha; p, n-p)$ is defined to be the point such that

$$\mathbb{P}(F_{p,n-p} > F(\alpha; p, n-p)) = \alpha$$

## 3.15 Confidence/Prediction Intervals for New Observations

Set $\mathbb{E}(Y|x^*) = \mu^* = (x^*)^T\beta$

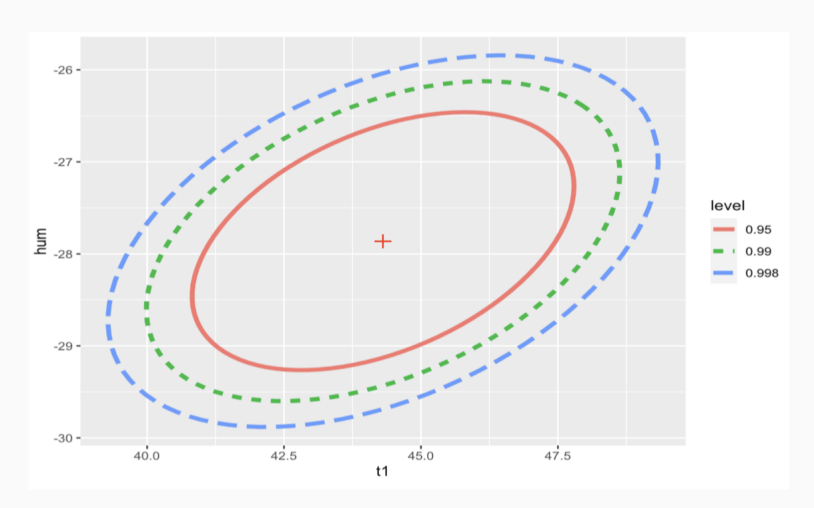


**Figure 3.3**

### 3.15.1 Confidence Interval for $\mu^* = (x^*)^T\beta$

The Gauss-Markov theorem tells us that the BLUE (Best Linear Unbiased Estimate) of $\mu^*$ is:

$$\hat{\mu}^* = (x^*)^T\hat{\beta} = (x^*)^T(X^TX)^{-1}X^Ty$$

Then,

$$\begin{aligned}\mathbb{E}[(\hat{\mu}^* - \mu^*)^2] &= Var(\hat{\mu}^*) \\ &= \sigma^2(x^*)^T(X^TX)^{-1}x^* \\ se(\hat{\mu}^*) &= \hat{\sigma}\sqrt{(x^*)^T(X^TX)^{-1}x^*}\end{aligned}$$

Where

$$\frac{\hat{\mu}^* - \mu}{se(\hat{\mu}^*)} \sim t(n-p)$$

A $(1-\alpha)100\%$ CI (confidence interval) for $\mu^*$ is:

$$\hat{\mu}^* \pm T_{n-p}(\frac{\alpha}{2})se(\hat{\mu}^*)$$

### 3.15.2 Prediction Interval for $y^* = (x^*)^T\beta + e^*$

The best estimate for $y^*$ at a future observation $x^*$ is also

$$\hat{y}^* = (x^*)^T\hat{\beta}$$

Then,

$$\begin{aligned} Var(\hat{y}^*) &= \mathbb{E}[((x^*)^T\hat{\beta} - y^*)^2] \\ &= \mathbb{E}[((x^*)^T\hat{\beta} - ((x^*)^T\beta + e^*))^2] \\ &= \mathbb{E}[((x^*)^T\hat{\beta} - ((x^*)^T\beta)^2] + \mathbb{E}[(e^*)^2] \\ &= Var(\hat{\mu}^*) + Var(e) \\ &= \sigma^2[1 + (x^*)^T(X^TX)^{-1}x^*] \\ se(\hat{y}^*) &= \hat{\sigma}\sqrt{1 + (x^*)^T(X^TX)^{-1}x^*} \end{aligned}$$

Where

$$\frac{\hat{y}^* - y^*}{se(\hat{y}^*)} \sim t(n-p)$$

A $(1-\alpha)100\%$ PI (prediction interval) for $y^*$ is:

$$\hat{y}^* \pm T_{n-p}(\frac{\alpha}{2})se(\hat{y}^*)$$

### 3.15.3 Mahalanobis distance

$$\mathbf{X}_{n\times p} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{bmatrix}^T$$

For any observation vector $\mathbf{x}_{p\times 1} = \begin{pmatrix} 1 \\ \mathbf{z} \end{pmatrix}$, where $\mathbf{z}$ denotes the value of predictors without the intercept.

We would like to use **Mahalanobis distance** to quantify the distance between observation vector $\mathbf{x}_{p\times 1}$ and its sample meaning $\bar{\mathbf{x}}$.

The sample convariance matrix of $\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \cdots & \mathbf{z}_n \end{bmatrix}$ is:

$$\hat{\Sigma}_{(p-1)\times(p-1)} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T$$

Then the $(x^*)^T(X^TX)^{-1}x^*$ can be written as

$$(x^*)^T(X^TX)^{-1}x^* = \frac{1}{n} + \underline{\frac{1}{n-1}(\mathbf{z}^* - \bar{\mathbf{z}})^T\hat{\Sigma}^{-1}(\mathbf{z}^* - \bar{\mathbf{z}})}$$

The second term in the right hand side ($\frac{1}{n-1}(\mathbf{z}^* - \bar{\mathbf{z}})^T\hat{\Sigma}^{-1}(\mathbf{z}^* - \bar{\mathbf{z}})$) is the so-called Mahalanobis distance from $\mathbf{z}^*$ to the center of the data $\bar{\mathbf{z}}$ (the sample mean).

Then we can write,

$$\begin{aligned}
\operatorname{se}\left(\hat{\mu}^*\right) &= \hat{\sigma}\sqrt{\mathbf{x}^{*\top}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{x}^*} \\
&= \hat{\sigma}\sqrt{\frac{1}{n}+\frac{1}{n-1}\left(\mathbf{z}^*-\overline{\mathbf{z}}\right)^T\hat{\Sigma}^{-1}\left(\mathbf{z}^*-\overline{\mathbf{z}}\right)} \\
se\left(\hat{y}^*\right) &= \hat{\sigma}\sqrt{1+\left(\mathbf{x}^*\right)^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{x}^*} \\
&= \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{1}{n-1}\left(\mathbf{z}^*-\overline{\mathbf{z}}\right)^T\hat{\Sigma}^{-1}\left(\mathbf{z}^*-\overline{\mathbf{z}}\right)}
\end{aligned}$$

Since $\operatorname{se}\left(\hat{y}^*\right)$ has an extra 1, when the sample size $n$ goes to infinity,

$$\operatorname{se}\left(\hat{\mu}^*\right) \to 0$$

$$\operatorname{se}\left(\hat{y}^*\right) \to \sigma$$

# Chapter 4 MLR: unusual observations

Recall, that we can write the MLR model as:

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{X}\beta, \sigma^2\mathbf{I}_n\right)$$

– Error: assumed to be iid, $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$

– Model: assumed to be linear in the parameters, i.e., $\mathbb{E}(\mathbf{y}) = \mathbf{X}\beta$

We might have unusual observations.

## 4.1 1. High leverage points: $h_i \geq \frac{2p}{n}$

### 4.1.1 Leverage Points

The diagonal elements of $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^\top$,

$$h_i = H_{ii} = \mathbf{x}_i\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_i^\top = \frac{Var(x_i^T\hat{\beta})}{\sigma^2}$$

are called **leverages** and are very useful diagnostics. $h_i$ gives a measure (invariant under any affine transformation of $\mathbf{X}$ ) of how far the $i$-th observation is from the center of the data (in the $X$-space).

For simple linear regression:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}$$

In general:

$$\begin{aligned} h_i &= \mathbf{x_i^T}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{x_i} \\ &= \frac{1}{n} + \frac{1}{n-1}\left(\mathbf{z}_i - \overline{\mathbf{z}}\right)^T\hat{\mathbf{\Sigma}}^{-1}\left(\mathbf{z}_i - \overline{\mathbf{z}}\right) \end{aligned}$$

where

$$\hat{\Sigma}^{-1}_{(p-1)\times(p-1)} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{z} - \overline{\mathbf{z}})(\mathbf{z} - \overline{\mathbf{z}})^T$$

is the sample covariance of the $(p-1)$ predictor variables. The second term in the right hand side ($\frac{1}{n-1}\left(\mathbf{z}_i - \overline{\mathbf{z}}\right)^T\hat{\mathbf{\Sigma}}^{-1}\left(\mathbf{z}_i - \overline{\mathbf{z}}\right)$) is the so-called **Mahalanobis distance** from $\mathbf{z}_i$ to the data center $\overline{\mathbf{z}}$

### 4.1.2 Properties of the Leverage: $0 < h_i < 1, \quad \sum_i h_i = p$

Recall that the hat matrix is idempotent $\mathbf{H} = \mathbf{H}\mathbf{H}^\top$ and has $\mathrm{tr}(\mathbf{H}) = p$

These imply that

$$\sum_i h_i = p \text{ and } \sum_j H_{ij}^2 = h_i$$

For a given $i$ we can decompose the last sum as follows:

$$\sum_j H_{ij}^2 = H_{ii}^2 + \sum_{i \neq j} H_{ij}^2 = h_i$$

$$\Rightarrow \sum_{i \neq j} H_{ij}^2 = h_i \left(1 - h_i\right) \Rightarrow h_i \left(1 - h_i\right) > 0$$

From this we can conclude the following properties of $h_i$ :

$$0 < h_i < 1, \quad \sum_i h_i = p$$

### 4.1.3 Fitted Values and Leverage: $Var(\hat{y}_i) = \sigma^2 h_i,\ Var(r_i) = \sigma^2(1 - h_i)$

Recall the equation $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$.

$$\begin{aligned} \hat{y}_i &= \sum_{t=1}^{n} H_{it} y_t \\ &= h_i y_i + \sum_{t \neq i}^{n} H_{it} y_t \end{aligned}$$

This means that $h_i = \frac{d\hat{y}_i}{dy_i}$

When $h_i$ is **large (close to 1)**, $\hat{y}_i$ relies heavily on $y_i$ (instead of using the information from other data points), therefore $\hat{y}_i$ will be "forced" to be **close** to the observed $y_i$.

Consequently, the variance for the residual $r_i$ will be small, and the variance for the fit $\hat{y}_i$ will be large (since the fit from another data set would be quite different).

$$Var(\hat{y}_i) = \sigma^2 h_i,\ Var(r_i) = \sigma^2(1 - h_i)$$

### 4.1.4 High-leverage Points: $h_i \geq \frac{2p}{n}$

*Good high-leverage points*: its $y$ point follows the pattern of the rest of the data, but with an $x_i$ value that is far away from the sample mean.

*Bad high-leverage points*: its $y$ value does not follow the pattern suggested by the rest of the data; the LS fitting might change a lot if we remove this point.

## 4.2 Residuals: Standardized Residuals vs. Studentized residuals

The residuals $r_i = y_i - \hat{y}_i$ do not have a constant variance. So they need to be standardized.

### 4.2.1 Difference between $\varepsilon$ and r

$\varepsilon$ : true residuals (our theoretical quantities)

**r** : estimated residuals - Both residuals are normally distributed, but:

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

where $\mathbf{H}$ is the projection/hat matrix.

- The errors $\varepsilon_i$ 's have equal variance and are independent, while the residuals $r_i$ 's have unequal variance and are correlated.

$-\mathbb{E}(\varepsilon) = \mathbb{E}(\mathbf{r}) = \mathbf{0}$. But

$$\sum_i \varepsilon_i \neq 0, \quad \sum_i r_i = 0$$

(by default we assume an intercept is included in the model)

### 4.2.2 Standardized Residuals: $r_i^* = \frac{r_i}{\hat{\sigma}\sqrt{1-h_i}}$

Since $r_i \sim \mathcal{N}\left(0, \sigma^2\left(1 - h_i\right)\right)$, it is reasonable to consider a standardization of the residuals in this form:

$$r_i^* = \frac{r_i}{\hat{\sigma}\sqrt{1-h_i}}, \quad i = 1, \ldots, n$$

- $\sum_i r_i^*$ is no longer zero.

- Since the $r_i$ is not independent of $\hat{\sigma}$, each $r_i^*$ is **not distributed as a $T$ distribution**.

- As an approximation, we can view the $r_i^*$ 's as iid $\mathcal{N}(0,1)$ random variables, although they are not Normally distributed and they are slightly correlated.

### 4.2.3 Studentized Residuals: $t_i = r_i^*\left(\frac{n-p-1}{n-p-r_i^{*2}}\right)^{1/2}$

- The studentized residuals are based on the idea of leave-one-out (also know as jackknife residuals).

- Here is the leave-one-out idea:

1. Run a regression model on the $(n-1)$ samples with the $i$-th sample $(x_i, y_i)$ removed.

2. Denote the leave-one-out estimates of the regression coefficient and error variance by $\hat{\beta}_{(i)}$ and $\hat{\sigma}_{(i)}$, where the notation (i) means "excluding the $i$-th observation."

3. Then, check the discrepancy between observations $y_i$ and the fitted value $\hat{y}_{(i)} = \mathbf{x}^T \hat{\beta}_{(i)}$

- Define the Studentized Residuals as:

$$t_i = \frac{y_i - \hat{y}(i)}{\hat{\sigma}_{(i)} \left(1 + x_i^T \left(\mathbf{X}_{(\mathbf{i})}^\top \mathbf{X}_{(i)}\right)^{-1} x_i\right)^{1/2}} = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1-h_i}}$$

which follows a $T_{n-p-1}$ distribution if $y_i \sim \mathcal{N}\left(\mathbf{x}_i^T \beta, \sigma^2\right)$

- One can also show that $r_i^*$ and $t_i$ are a monotone transformation of each other.

- We do not need to run the model $n$ times to get the estimates $\hat{\beta}_{(i)}$ and $\hat{\sigma}_{(i)}$ since it can be shown that:

$$t_i = r_i^* \left(\frac{n-p-1}{n-p-r_i^{*2}}\right)^{1/2}$$

## 4.3 2. Outlier (Large $|r_i^*|$)

### 4.3.1 Outlier test

Outliers are observations that do not fit the model, but Outliers are not necessarily observations with large residuals.

We need to used the studentized residuals for the outlier test.

Under the Null hypothesis $H_0$,

$$t_i \sim T_{n-p-1}$$

We can use t-test to test the $i^{th}$ observation: form a PI at $x_i$.

(an example of data snooping)

Large $|r_i^*| \Rightarrow$ Large $|t_i| \Rightarrow$ Reject Null hypothesis $\Rightarrow$ Outlier

### 4.3.2 Bonferroni Correction

Suppose we are testing $m$ hypothesis sinultaneously.

For each test, we use a significant level $\alpha$. That is, the chance to make a **overall** Type I error is $\alpha$.

Suppose we want to control the overall type I error rate (for all $m$ tests) to be $95\%$.

We should set the individual significance levels to be $\alpha = 5\%/m$

When we test outliers, since the $T$ distribution is bipartite, we need the confidence level: $\alpha/(2 * n)$

### 4.3.3 What we should do with outliers?

Points should not be routinely deleted simply because they do not fit the model. No data snooping!

Outliers, as well as other unusual observations discussed here, often flag potential problems of the current

model. **Instead of dropping them, maybe, try a new alternative model.**

## 4.4 3. Highly Influential Points: Large $D_i = \frac{r_i^{*2}}{p}(\frac{h_i}{1-h_i})$ $(D_i \geq 1)$

### 4.4.1 Influential observations

Observations whose removal greatly affects the regression analysis are called **influential observations**.

An **influential observations** may be (or may not) an **outlier** or a **high-leverage observation**; or may be both: an outlier and a high-leverage observation.

We will use the **Cook's distance** to detect influential observations.

$$D_i = \frac{\left\|\mathbf{X}\hat{\beta} - \mathbf{X}\hat{\beta}_{(i)}\right\|^2}{p\hat{\sigma}^2} = \frac{\left\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}\right\|^2}{p\hat{\sigma}^2} = \frac{r_i^{*2}}{p}\left(\frac{h_i}{1-h_i}\right)$$

which indicates that highly influential points are either outliers (large $|r_i^*|$) or high-leverage points (large $h_i$) or both.

A rule-of-thumb: observations with $D_i \geq 1$ are highly influential.

# Chapter 5 MLR Diagnostics: Checking Assumptions

## 5.1 Classical Linear Model (CLM) assumption: Gauss-Markov Assimption + $\varepsilon \sim^{IID} \mathcal{N}(0, \sigma^2)$

1. Constant Variance (Homoscedasticity)
2. Normality
3. Uncorrelated errors (No-Autocorrelation)
4. Linearity: $\mathbb{E}(y) = \mathbf{X}\beta$
5. Random Sampling

(1-5 calls Gauss-Markov Assimption)

6. $\mathbf{Y} = \beta\mathbf{X} + \varepsilon,$ where $\varepsilon \sim^{IID} \mathcal{N}(0, \sigma^2)$

(1-6 calls Classical Linear Model (CLM) assumption)

## 5.2 Check 1. Constancy of Variance(Homoscedasticity)

### 5.2.1 Method 1: graph *residuals* against *Fitted Values* $\hat{y}$

**SLR**:

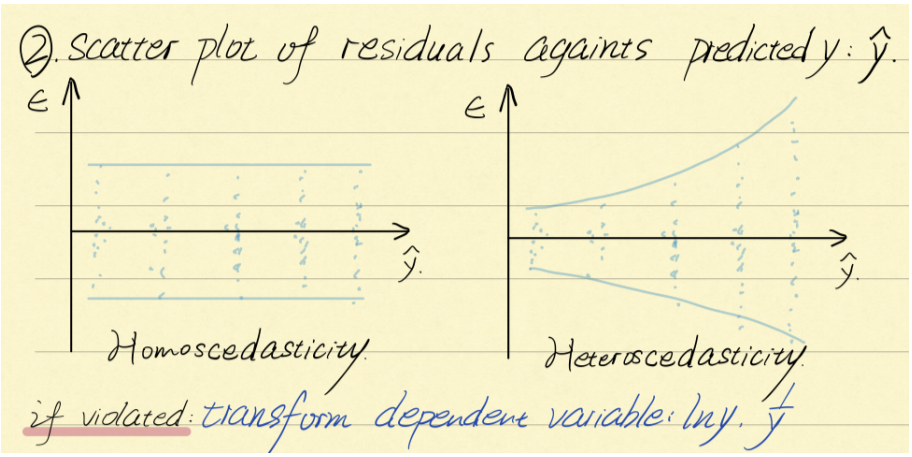


**Figure 5.1**

**MLR**:

If the variance is constant, the residuals will look like a football-shaped cloud. Check residual plots and look

for a “fan” type shape or trends.

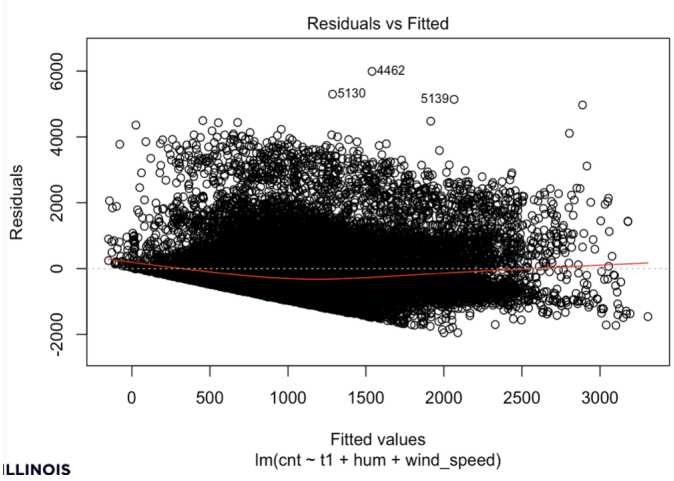


**Figure 5.2**

### 5.2.2 Method 2: Breusch-Pagan Test

$$\text{BP} = nR^2$$

where $R^2$ is the *coefficient of Determination* between the **squared residuals** $r_i^2$ of LS regression and the **covariates** (or a sub-set) $X_1, X_2, ..., X_p$.

$$H_0 : \text{BP} \sim \chi^2_{p-1} \text{ (asymptotically)}$$

### 5.2.3 What happen if Heteroscedasticity

**Unchanges:**

1. The estimaor will still be *unbiased* ($\mathbb{E}(\hat{\beta}) = \beta$), *consistent* ($\lim_{n\to\infty} \hat{\beta} = \beta$), but *inefficient*.
2. Interpretation of $R^2$ is not changed.

**Changes:**

1. $\hat{\beta}$ will be *inefficient*.
2. invalidates variance formulas for OLS estimators ($Var(\hat{\beta}_i)$).
3. usual $F-$test, $t-$test are not valid.
4. *Gauss-Markov Theorem* doesn't hold: OLS is no longer the BLUE (best linear unbiased estimator).

### 5.2.4 What can we do if Heteroscedasticity

(i) Take logarithms of each of the variables

(ii) Use suitably modified standard errors

(iii) Use a generalised least squares (GLS) procedure

### 5.2.5 Remedial measure: Variance Stabilizing Transformations $\sqrt{Y}$, $\log Y$, $\frac{1}{Y}$ or $\frac{1}{Y+1}$

**SLR**:

If violates Homoscedasticity: Transform *dependent variable*: $\ln y, \frac{1}{y}$

**MLR**:

Find a transformation of the response, $h(Y)$, to achieve constant variance.

How does it work?

- Suppose $h$ is a smooth function.

- Using Taylor's theorem, the expansion of $h(Y)$ around $\mathrm{E}(Y)$ is:

$$h(Y) = h(\mathbf{E}(Y)) + h'(\mathbf{E}(Y))(Y - \mathbf{E}(Y)) + \text{ Remainder}$$

- The remainder is assumed small with high probability and we can ignore it:

$$\mathrm{Var(h(Y))} \approx \left(\mathrm{h'(E(Y))}\right)^2 \mathrm{Var(Y)}$$

- We want to choose a transformation $h$ such that $\mathrm{Var(h}(Y))$ is approximately constant.

**Example 1**:

- Suppose that the variance of $Y$ is proportional to the mean of $Y$, i.e., $\mathrm{Var}(Y) \propto E(Y)$

- Select $h$ such that:

$$h'(z) = \frac{1}{\sqrt{z}} \Rightarrow h(z) \propto \sqrt{z}$$

- When plugging-in the value of $h'(z)$ evaluated at $E(Y)$ in the variance of $h(Y)$ equation, the variance of $h(Y)$ will be approximately constant. Indeed,

$$\mathrm{Var}(\sqrt{Y}) \approx \left(\frac{1}{\sqrt{E(Y)}}\right)^2 \mathrm{Var}(Y) = \frac{\mathrm{Var}(Y)}{E(Y)} \approx \text{ const}$$

**Example 2**:

- Suppose that the variance of $Y$ is proportional to the squared mean of $Y$, i.e., $\mathrm{Var}(Y) \propto (E(Y))^2$.

- Select $h$ such that:

$$h'(z) = \frac{1}{z} \Rightarrow h(z) = \log(z)$$

- Then,

$$\text{Var}(\log Y) \approx \frac{1}{(\text{E(Y)})^2} \text{Var(Y)} \approx \text{ const}$$

**Example 3:**

$\text{Var}(Y) \propto (E(Y))^4$.

$$h(Y) = \frac{1}{Y} \text{ or } \frac{1}{Y+1}$$

## 5.3 Check 2. Normality

### 5.3.1 Method 1: Histogram, graph *residuals* against its frequency

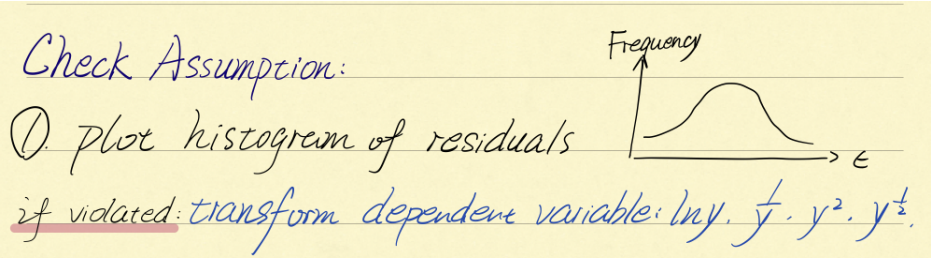


**Figure 5.3**

If violates Normality: Transform *dependent variable*: $\ln y, \frac{1}{y}, y^2, \sqrt{y}$

### 5.3.2 Method 2: QQ-Plot, graph *residuals* against its frequency

- Suppose that we have a sample $z_1, z_2, \ldots, z_n$.

- We wish to examine the hypothesis that the z's are a sample from a normal distribution with mean $\mu$ and variance $\sigma^2$.

**QQ-Plot**:

1. Order the z's: $z_{(1)}, z_{(2)}, \ldots, z_{(n)}$.
2. Compute $u_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$, where $\Phi$ is the cdf of the $N(0,1)$ and $i$ is the order if the data $(i = 1, 2, \ldots, n)$.

3. Plot $z_{(i)}$ against $u_i$.

$\Rightarrow$ If the z's are normal, the plot should be approximately a straight line.

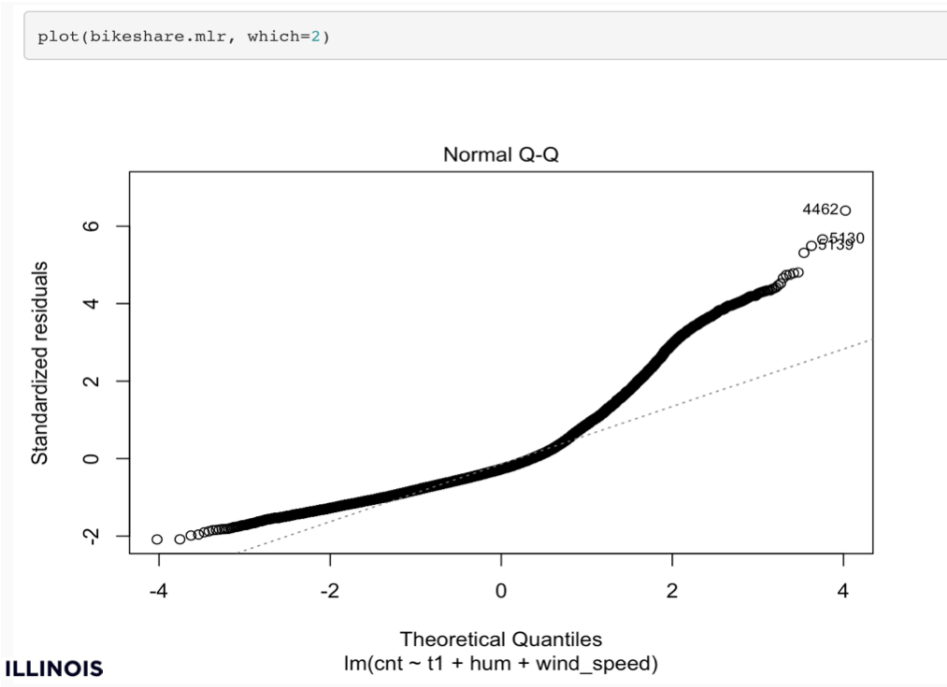


**Figure 5.4**

### 5.3.3 Method 3: Shapiro-Wilk Test (good for $n \leq 50$)

$$W = \frac{(\sum_{i=1}^{n} a_i r_{(i)})^2}{\sum_{i=1}^{n}(r_i - \bar{r})^2}$$

where $r_{(i)}$ is the $i$th largest value of the $r_i$'s and the $a_i$ terms are calculated using the means, variances, and covariances of the $r_i$s.

**Small values of** $W$ will lead to rejection of the null hypothesis.

### 5.3.4 Method 4: Kolmogorov-Smirnov Test (good for $n > 50$)

$$D_n = \sup_x |F_n(x) - \Phi(x)|$$

where $\Phi(x)$ is the cdf of the Normal and $F_n$ the empirical distribution function $F_n$ for $n$ i.i.d. ordered observations $X_i$ is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{[-\infty,x]}(X_i)$$

**Small values of** $D$ will lead to rejection of the null hypothesis.

### 5.3.5 Remedial measure: Box-Cox Transformations of $Y$

Suppose each $y_i > 0$, and consider the following transformation:

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

Choose $\lambda$ that maximizes the likelihood of the data, under the assumption that the transformed data $g_\lambda(\mathbf{y})$ has a normal distribution:

$$g_\lambda(\mathbf{y}) = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim N_n\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$$

- The log-likelihood function for $\lambda \neq 0$ is:

$$L(\lambda) = -\frac{n}{2}\log\left(RSS_\lambda/n\right) + (\lambda - 1)\sum_{i=1}^{n}\log\left(y_i\right)$$

where $RSS_\lambda$ is the $RSS$ when $g_\lambda(\mathbf{y})$ is the response, and for $\lambda = 0$ is:

$$L(0) = -\frac{n}{2}\log\left(RSS_0/n\right) - \sum_{i=1}^{n}\log\left(y_i\right)$$

The second term in these log-likelihood function corresponds to the Jacobian of the transformation.

In **R**, we can graph the log-likelihood as a function of $\lambda(L(\lambda))$ versus $\lambda \in (-2, 2)$ and then pick the maximizer $\hat{\lambda}$.

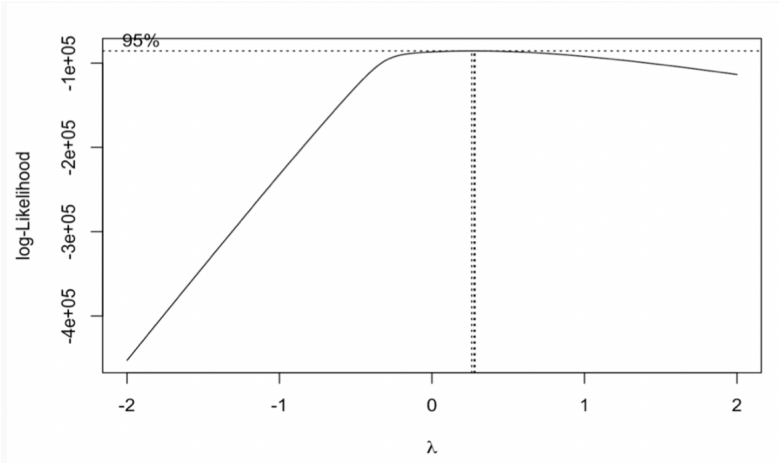


**Figure 5.5**

It is common to round $\hat{\lambda}$ to a nearby value like:

$$-1, -0.5, 0, 0.5, \text{ or } 1$$

then the transformation defined by $\hat{\lambda}$ is easier to interpret.

To answer the question whether we really need the transformation $g_\lambda$, we can do hypothesis testing $(H_0 : \lambda = 1)$, or equivalently construct a Confidence Interval for $\lambda$ as follows:

$$\{\lambda : L(\lambda) > L(\hat{\lambda}) - \frac{1}{2}\chi_1^2(1-\alpha)\}$$

## 5.4 Checking 3. Serial Dependence

### 5.4.1 Method 1: graph *residuals* against index variable(time or case number)

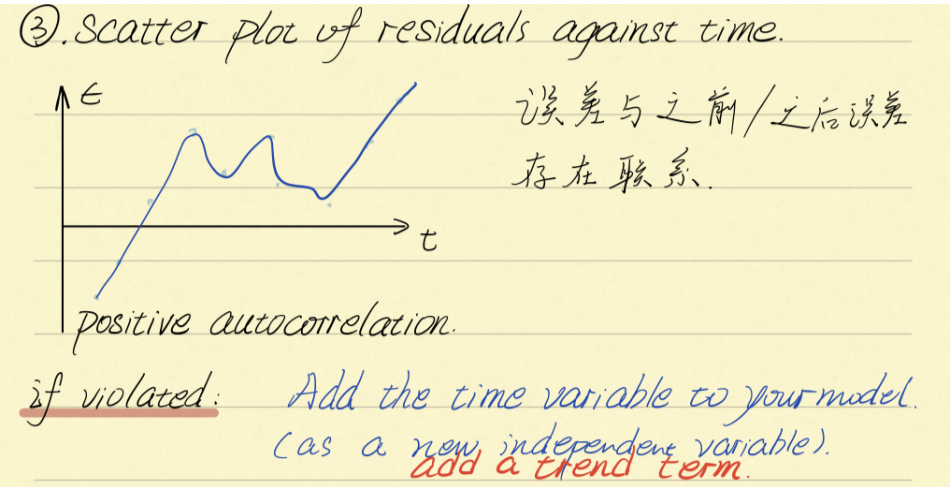


**Figure 5.6**

If violates No-Autocorrelation: add a new independent vaiable $(t)$.

### 5.4.2 Method 2: Durbin Watson test

**SLR:**

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Compare $d$ with $d_L$ and $d_U$.

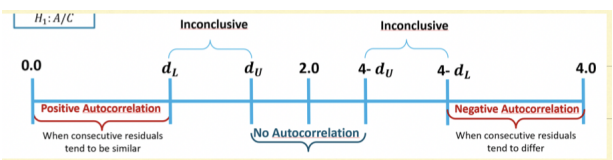


**Figure 5.7**

**MLR:**

$$DW = \frac{\sum_{k=1}^{n-1}(r_k - r_{k+1})^2}{\sum_{k=1}^{n} r_k^2}$$

if $DW < 2$, then there is evidence for positive serial dependence.

## 5.5 Checking 4. Non-Linearity

### 5.5.1 Method 1: Partial Regression Plots

We want to know **the relationship between the response** $Y$ **and a predictor** $X_k$ after the effect of the other predictors has been removed.

To remove the effect of the other predictors, run the following two regression models:

$$Y \sim X_1 + \ldots + X_{i-1} + X_{i+1} + \ldots \quad (1)$$

$$X_i \sim X_1 + \ldots + X_{i-1} + X_{i+1} + \ldots \quad (2)$$

Get the following residuals:

$$\mathbf{r}_y = \text{ residuals from } (1)$$

$$\mathbf{r}_k^X = \text{ residuals from } (2)$$

Plot $\mathbf{r}_y$ vs. $\mathbf{r}_k^X$ : For a valid model, the added-variable plot should produce points randomly scattered around a line through the origin with slope $\hat{\beta}_k$. This is also a useful plot to detect *high influential* data points.

### 5.5.2 Remedial measure: Linearizing Transformations

1. $\log(Y)$ vs. $\log(X)$, suitable when $\mathbb{E}(Y) = \alpha X_1^{\beta_1} ... X_p^{\beta_p}$
2. $\log(Y)$ vs. $X$, suitable when $\mathbb{E}(Y) = \alpha \exp \sum_j X_j \beta_j$
3. $\frac{1}{Y}$ vs. $X$, suitable when $\mathbb{E}(Y) = \frac{1}{\alpha + \sum_j X_j \beta_j}$

### 5.5.3 Remedial measure: Box-Cox Transformations of $Y$ also works

# Chapter 6 MLR Diagnostics: Collinearity

## 6.1 Exact Collinearity/ linearly dependent

There exists a set of constants $c_1, c_2, ..., c_p$ (at least one of them is non-zero) s.t.

$$\sum_{j=1}^{p} c_j \mathbf{X}_{.j} = 0$$

then the columns of X are called *linearly dependent* and there is *exact collinearity*.

## 6.2 What happens if exact collinearity

1. $(\mathbf{X}^T\mathbf{X})^{-1}$ does not exists.
2. The LS estimate $\hat{\beta}$ is not unique.
3. The corresponding linear model is not identifiable.

## 6.3 Approximate Collinearity

We generally do not need to worry about exact collinearity ($\mathbf{R}$ can detect it and fix it automatically), but *approximate collinearity*.

$$\sum_{j=1}^{p} c_j \mathbf{X}_{.j} \approx 0$$

$$\mathbf{X}_{.k} \approx -\sum_{j \neq k} c_j \mathbf{X}_{.j} / c_k$$

A simple diagnostic for this is to obtain the regression of $\mathbf{X}_{.k}$ on the remaining predictors, and if the corresponding $R_k^2$ is close to 1, we would diagnose approximate collinearity.

$$\mathbf{X}_{.k} \sim \mathbf{X}_{.1} + \ldots + \mathbf{X}_{.k-1} + \mathbf{X}_{.k+1} + \ldots \Rightarrow R_k^2$$

## 6.4 What happens if approximate collinearity: based on $\left(\frac{1}{1-R_k^2}\right)$ ($k-$th variance inflation factor (VIF), $VIF > 10$)

In a multiple regression $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + e$, the LS estimate $\hat{\beta}_k$ is unbiased with variance:

$$\text{Var}\left(\hat{\beta}_\text{k}\right) = \sigma^2 \left(\frac{1}{1 - \text{R}_\text{k}^2}\right) \left(\frac{1}{\sum_{\text{i}=1}^{n} \left(\text{x}_\text{ik} - \bar{\text{x}}_{\cdot\text{k}}\right)^2}\right)$$

where $R_k^2$ is the R-square from the regression of $\mathbf{X}_{\cdot k}$ on the remaining predictors. When $R_k^2$ is close to 1 , the variance of $\hat{\beta}_k$ is large.

Consequently we will have:

1. large Mean Square Error
2. large (inflated) $p$-value to the corresponding $t$-test, i.e, we could miss a significant predictor.

The quantity $\left(\frac{1}{1-R_k^2}\right)$ is called the $k$-th variance inflation factor (VIF).

$VIF > 10$ infers conllinearity

## 6.5 Possible symptoms of collinearity

1. high pair-wise (sample) correlation between predictors
2. high VIF
3. high condition number
4. $R^2$ is relatively large but none of the predictor is significant.

## 6.6 Global Measure of Collinearity: *condition number* of $\mathbf{X}^T\mathbf{X}$

*Condition number* of $\mathbf{X}^T\mathbf{X}$:

$$\kappa = (\text{largest eigenvalue/smallest eigenvalue})^{1/2}$$

An empirical rule for declaring collinearity is $\kappa \geq 30$

Note that $\kappa$ is *not scale-invariant*, so we should *standardized* each column of $X$ (i.e. each column should have *zero mean* and *sample variance* equal to 1, before calculating the condition number).

## 6.7 What to do with collinearity

1. Remove some predictors from highly correlated groups of predictors.
2. Regularize the model using penalized Least Squares estimation.

# Chapter 7 Generalized Least Squares (GLS)

What do we do if the errors are **correlated** or **heteroscedastic**?

Suppose $\varepsilon \sim \mathcal{N}_n(0, \Sigma)$, where $\Sigma$ is the variance-covariance matrix.

## 7.1 GLS, $\Sigma$ known ($\hat{\beta} = \left(\mathbf{X}^\top \Sigma^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\top \Sigma^{-1} \mathbf{y}$, $RSS = (\mathbf{y} - \mathbf{X}\beta)^\top \Sigma^{-1} (\mathbf{y} - \mathbf{X}\beta)$)

### 7.1.1 Method 1: Transform back to OLS

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

where $\varepsilon \sim \mathcal{N}_n(0, \Sigma)$ and $\Sigma$ is a known, symmetric, positive definite covariance matrix.

When the errors are heteroscedastic or correlated:

Transform this problem back to Ordinary Least-Squares (OLS):

1. Assume $S^{-1}$ exists and write

$$\Sigma = SS^\top$$

(We could use, for example, the Cholesky decomposition from linear algebra to obtain $S$.)

2. Multiply the model equation by $S^{-1}$ on both sides:

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

$$S^{-1}\mathbf{y} = S^{-1}(\mathbf{X}\beta + \varepsilon)$$

$$\underbrace{S^{-1}\mathbf{y}}_{:=\mathbf{y}^*} = \underbrace{S^{-1}\mathbf{X}}_{:=\mathbf{x}^*}\beta + \underbrace{S^{-1}\varepsilon}_{:=\varepsilon^*}$$

$$\mathbf{y}^* = \mathbf{X}^*\beta + \varepsilon^*$$

This implies that

$$\varepsilon^* \sim \mathcal{N}(S^{-1}\mathbf{0}, \underbrace{S^{-1}\Sigma\left(S^{-1}\right)^\top}_{=\text{Identity}}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$

since $S^{-1}\Sigma\left(S^{-1}\right)^\top = S^{-1}SS^\top\left(S^{-1}\right)^\top = I$

3. For the transformed model, we can solve for $\beta$ using OLS:

$$\mathbf{y}^* = \mathbf{X}^* \beta + \varepsilon^*$$

where $\mathbf{y}^* = S^{-1}\mathbf{y}, \mathbf{X}^* = S^{-1}\mathbf{X}$

So, the estimator for $\beta$ computes as

$$\begin{aligned}
\hat{\beta} &= \left(\mathbf{X}^{*\top}\mathbf{X}^*\right)^{-1} \mathbf{X}^{*\top}\mathbf{y}^* \\
&= (\mathbf{X}^\top \underbrace{\left(S^{-1}\right)^\top S^{-1}}_{=\Sigma^{-1}} \mathbf{X})^{-1} \mathbf{X}^\top \underbrace{\left(S^{-1}\right)^\top S^{-1}}_{=\Sigma^{-1}} \mathbf{y} \\
&= \left(\mathbf{X}^\top \Sigma^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\top \Sigma^{-1} \mathbf{y}
\end{aligned}$$

Note that the solution we obtained minimizes:

$$\|\mathbf{y}^* - \mathbf{X}^*\beta\|^2 = (\mathbf{y} - \mathbf{X}\beta)^\top \Sigma^{-1} (\mathbf{y} - \mathbf{X}\beta)$$

### 7.1.2 Weighted Least Squares (WLS)

Suppose that $\Sigma$ is a diagonal matrix of unequal error variances:

$$\Sigma = \operatorname{diag}\left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\right)$$

The GLS estimate of $\beta$ minimizes:

$$(\mathbf{y} - \mathbf{X}\beta)^\top \Sigma^{-1} (\mathbf{y} - \mathbf{X}\beta) = \sum_{i=1}^{n} \frac{\left(y_i - \mathbf{x}_i^\top \beta\right)^2}{\sigma_i^2}$$

This problem is known as the Weighted Least-Squares (WLS).

Note that the errors are weighted by

$$w_i = \frac{1}{\sigma_i^2}$$

smaller weights for samples with larger variances.

### 7.1.3 WLS special example : Replicated Observations

Suppose we collected multiple observations for each $\mathbf{x}_i$. We use double subscripts to indicate the replicate observations:

$$(\mathbf{x}_i, y_{i1}, y_{i2}, \ldots, y_{in_i})$$

Let $y_i$ denote the average of the $n_i$ observations sharing $\mathbf{x}_i$. Then the residual sum of squares for $\beta$ equals

$$\sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(y_{ij} - \mathbf{x}_i^\top \beta\right)^2 = \sum_{i=1}^{n} n_i \left(y_i - \mathbf{x}_i^\top \beta\right)^2 + \sum_{i=1}^{n} \sum_{j=1}^{n_i} (y_{ij} - y_i)^2$$

Minimizing the $RSS$ to solve for $\beta$ is the same as minimizing the first term on the right only.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} n_i (y_i - \mathbf{x}_i^T \beta)^2$$

### 7.1.4 Method 2: Likelihood Estimation

Model: $\mathbf{y} \sim N_n(\mathbf{X}\beta, \Sigma)$

Log-likelihood:

$$\begin{aligned}
&\log(p(\mathbf{y} \mid \beta, \Sigma)) \\
&= \log\left\{ \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[ -\frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^\top \Sigma^{-1} (\mathbf{y} - \mathbf{X}\beta) \right] \right\} \\
&= -\frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\beta) + \text{ Constant.}
\end{aligned}$$

Therefore the MLE is given by

$$\hat{\beta}_{mle} = \arg\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^\top \Sigma^{-1} (\mathbf{y} - \mathbf{X}\beta)$$

## 7.2 GLS, $\Sigma$ unknown

### 7.2.1 Method 1: Estimation of Variance ($r_i^2$)/Standard Deviation Function ($|r_i|$)

$$\sigma_i^2 = \mathbb{E}(\varepsilon_i^2) - (\mathbb{E}(\varepsilon_i))^2$$

Since we assume $E(\varepsilon_i) = 0$, we have

$$\sigma_i^2 = \mathbb{E}(\varepsilon_i^2)$$

Which implies $r_i^2$ is an estimator of $\sigma_i^2$; $|r_i|$ is an estimator of the standard deviation $\sigma_i$.

**Estimate Variance Function** $\hat{v}_i(x)$

1. Fit a regression model using OLS, and obtain the residuals $r_i$.
2. Regress the squared residuals $r_i^2$ against the appropriate predictor variables.

Denote $\hat{v}_i$ be the fitted value from variance function

$$w_i = \frac{1}{\hat{v}_i}$$

**Estimate Standard Deviation Function** $\hat{s}_i(x)$

1. Fit a regression model using OLS, and obtain the residuals $r_i$.
2. Regress the absolute residuals $|r_i|$ against the appropriate predictor variables.

Denote $\hat{s_i}$ be the fitted value from standard deviation function

$$w_i = \frac{1}{(\hat{s_i})^2}$$

The estimated variances are then placed in the variance-covariance matrix $\Sigma$ and the regression coefficients are estimated using the WLS (Weighted Least Squares method).

### 7.2.2 Method 2: iterative approach

1. Start with some initial guess of $\Sigma$
2. Use $\Sigma$ to estimate $\beta$
3. Use residuals (since we have known $\beta$) to estimate $\Sigma$
4. Iterate until convergence.

It looks like a good idea; however the methods will not work if we do not assume some structure about $\Sigma$ (too many parameters to be estimated).

Based on the application, we can assume a particular structure for $\Sigma$ that does not involve too many parameters. Then, we can model $\beta$ and $\Sigma$ simultaneously.

For example , for AR(1) times series (auto-regressive model of order 1), the structure of $\Sigma$ would be:

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots \\ \rho & 1 & \rho & \rho^2 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & \cdots & 1 \end{pmatrix}$$

$\Sigma$ as a function of $\rho$ and $\sigma^2$.

# Chapter 8 GLS Diagnostics: Lack of Fit Tests

## 8.1 Gaussian Assumption

Gaussian Assumption, which can be summarized concisely as:

$$y \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2\mathbf{I})$$

Under these assumptions:

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y \sim \mathcal{N}_p(\beta, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}),$$

$$\hat{y} = \mathbf{X}\hat{\beta} \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2\mathbf{H})$$

independently,

$$\hat{\sigma}^2 = \frac{RSS}{n-p} = \frac{||y-\hat{y}||^2}{n-p} \sim \sigma^2\frac{\chi^2_{n-p}}{n-p}$$

## 8.2 When $\sigma^2$ is known

Intuition:

If the model is correct, then $\hat{\sigma}^2$ is an unbiased estimate of $\sigma^2$.

If we know $\sigma^2$, we could construct a test based on the ratio $\frac{\hat{\sigma}^2}{\sigma^2}$, a measure of *lack-of-fit*.

In this case we want to test the hypothesis:

$$\begin{cases} H_0: \text{ There is no lack of fit.} \\ H_\alpha: \text{ There is lack of fit.} \end{cases}$$

We use the test statistic:

$$\frac{\hat{\sigma}^2}{\sigma^2} = \frac{RSS/(n-p)}{\sigma^2} \sim \frac{\chi^2_{n-p}}{n-p}$$

Lack of fit means the error variance is large related to the value of $\sigma^2$, i.e., the test statistic is large.

Conclude that there is lack of fit (i.e. Reject $H_0$), if:

$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \geq \chi^2_{n-p}(1-\alpha)$$

## 8.3 When $\sigma^2$ is unknown

### 8.3.1 Hypothesis

If $\sigma^2$ is unknown, a general approach is to compare an estimate of $\sigma^2$ based on a much bigger/general model. If we can derive the distribution (under $\mathrm{H}_0$ ) of $\hat{\sigma}^2_{\text{LinearModel}} / \hat{\sigma}^2_{\text{BigModel}}$, then we reduce this problem to a two model comparison test problem.

The null hypothesis is the current model:

$H_0 : \mathbb{E}(y_i) = \mathbf{x}_i^\top \beta, \quad i = 1, 2, \ldots, n, \quad$ for some vector $\beta$

The more general model is assumed under the alternative hypothesis:

$H_\alpha : \mathbb{E}(y_i) = f(\mathbf{x}_i), \quad i = 1, 2, \ldots, n, \quad$ for some function $f$

### 8.3.2 Under the null hypothesis $H_0$

$y_{ij} = \mathbf{x}_i^\top \beta + \varepsilon_{ij}$, some $\beta, \varepsilon_{ij} \sim$ iid $\mathcal{N}\left(0, \sigma^2\right)$

$RSS_0$ with $df = n - p$

### 8.3.3 Under the alternative big-model hypothesis $H_\alpha$ :

$y_{ij} = f(\mathbf{x}_i) + \varepsilon_{ij}$, some function $f, \varepsilon_{ij} \sim$ iid $\mathcal{N}\left(0, \sigma^2\right)$

**Can we estimate $\sigma^2$ for the big model in $H_\alpha$ ?**

- The answer is yes, if there is some replication in the data, i.e., there are multiple observations (replicates) for some (at least) of the same $\mathbf{x}_i$ values.

- Schematically we can represent these replicates as:

$$(\mathbf{x}_i, y_{i1}, y_{i2}, \ldots, y_{in_i}), \quad i = 1 : m, \quad n = \sum_i n_i$$

$RSS_a$ with $df = n - m = \sum_i (n_i - 1)$, where

$$RSS_a = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

### 8.3.4 $F-$Test

All of the degrees of freedom for $RSS_a$ come from the replications. Therefore, with replication we can do an F test for lack of fit:

$$F = \frac{(RSS_0 - RSS_a)/(m-p)}{RSS_a/(n-m)} \sim F_{m-p, n-m}$$

# Chapter 9 Polynomials Regression

## 9.1 Basic Function

$$y_i = f(x_i) + \varepsilon_i$$
$$y_i = \beta_0 + \sum_{j=1}^{d} b_j(x_i)\beta_j + \varepsilon_i$$
$$y_i = \beta_0 + \sum_{j=1}^{d} \beta_j x_i^j + \varepsilon_i$$

$d$ is the **degree of the polynomial component**.

## 9.2 Choose Order $d$

1.**Forward Approach**: Keep adding terms until the last added term is not significant.

2.**Backward Approach**: Start with a large $d$, and keep eliminating the terms that are not statistically significant, starting with the highest order term.

Once we pick up a $d$, we do not test the significance of the lower-order terms. (include all the lower-order terms in our model by default)

**Reasoning**: we do not want our results to be affected by a change of location/scale of the data. $(z_i - 2)^2 = z_i^2 - 4z_i + 4$.

**Exception**: particular polynomial function (physics law).

## 9.3 Orthogonal Polynomials

Successive predictors $x^j$ are **highly correlated** introducing multicollinearity problems.

$$y_i = \beta_0 + \beta_1 \mathbf{z}_i + ... + \beta_d \mathbf{z}_d + \varepsilon_i$$

where $\mathbf{z}_j = a_1 + b_2 x + ... + \kappa_j x^j$ is a polynomial of order $j$ with coefficients chosen such that $\mathbf{z}_j^T \mathbf{z}_j = 0$

## 9.4 Piece-wise Polynomials

If the true mean of $E(Y|X = x) = f(x)$ is too wiggly, we might need to fit a higher order polynomial, which is not always a good idea.

Instead we will consider **piece-wise polynomials**:

1.we divide the range of $x$ into several intervals, and

2.within each interval $f(x)$ is a low-order polynomial, e.g., cubic or quadratic, but the polynomial coefficients change from interval to interval;

3.in addition we require the overall $f(x)$ to be continuous up to certain derivatives.

## 9.5 Cubic Splines

### 9.5.1 Why Spline

**Polynomials**: smooth, but each point affects the fit globally.

**Piece-wise Polynomials**: localizes the influence of each data point, but are not smooth enough.

**Splines**: combines the beneficail aspects of both approaches.

### 9.5.2 Settings

A **Cubic Spline** is a curve constructed from sections of cubic polynomials, joined together so that the curve is *continuous up to second derivative*.

The points at which the sections join are called the **knots** of the spline.

We want to define a cubic spline function in the interval $[a, b]$

- Define $m$ knots such that: $a < \xi_1 < \xi_2 < \ldots < \xi_m < b$

- A function $g$ defined on $[a, b]$ is a **cubic spline with respect to knots** $\{\xi_i\}_{i=1}^m$ if:

1. $g$ is a cubic polynomial in each of the $m + 1$ intervals,

$$g(x) = d_i x^3 + c_i x^2 + b_i x + a_i, \quad x \in [\xi_i, \xi_{i+1}]$$

where $i = 0, \ldots, m, \xi_0 = a$ and $\xi_{m+1} = b$

2. $g$ is continuous up to the 2 nd derivative: since $g$ is continuous up to the 2nd derivative for any point inside

an interval, it suffices to check the following conditions:

$$g^{(0,1,2)}\left(\xi_i^+\right) = g^{(0,1,2)}\left(\xi_i^-\right), \quad i = 1:m$$

This expression indicates that the function and the first and second order derivatives are continuous at the knots.

### 9.5.3 Number of free parameters: $m + 4$

How many free parameters do we need to represent a cubic spline?

(i) 4 parameters $(d_i, c_i, b_i, a_i)$ for each of the $(m + 1)$ intervals.

(ii) 3 constraints at each of the $m$ knots (continuity constraints).

The **total number of free parameters** (similar to the number of degrees of freedom) is:

$$4(m + 1) - 3m = m + 4$$

### 9.5.4 Properties: linear combination also cubic spline

Given knots $\{\xi_i\}_{i=1}^m$, the linear combinations of cubic splines are also cubic splines.

That is, for a set of given knots, the corresponding cubic splines form a linear space (of functions) with $dim(m + 4)$.

### 9.5.5 Cubic Splines Basis

A set of basis functions for cubic splines (w.r.t knots $\{\xi_i\}_{i=1}^m$ ) is given by:

$$h_0(x) = 1$$

$$h_1(x) = x$$

$$h_2(x) = x^2$$

$$h_3(x) = x^3$$

$$h_{i+3}(x) = (x - \xi_i)_+^3, \quad i = 1, 2, \ldots, m$$

That is, any cubic spline can be uniquely expressed as:

$$\beta_0 + \sum_{j=1}^{m+3} \beta_j h_j(x)$$

Given knot locations, there are many alternative, but equivalent ways of writing down a basis for cubic splines.

**Example 9.1Other Basis** For example, another basis for cubic splines can be the following:

$$h_0(x) = 1$$

$$h_1(x) = x$$

$$h_{i+1}(x) = R\left(x, \xi_i^*\right), i = 1, \ldots, q-1$$

where

$$\begin{aligned} R(x,z) = &\left[(z-1/2)^2 - 1/12\right]\left[(x-1/2)^2 - 1/12\right]/4 \\ &- \left[(|x-z|-1/2)^4 - 1/2(|x-z|-1/2)^2 + 7/240\right]/24 \end{aligned}$$

## 9.6 Natural Cubic Splines (NCS)

A cubic spline on $[a, b]$ is a **Natural Cubic Spline** if its *second and third derivatives are zero at a and b.*

### 9.6.1 Degree of Freedom (Number of free parameters): $m$

This condition implies that NCS is a linear function in the two extreme intervals $[a, \xi_1]$ and $[\xi_m, b]$. The linear functions in the two extreme intervals are completely determined by their neighboring intervals.

The degree of freedom of NCS's with $m$ knots is:

$$4(m+1) - 3m - 4 = m$$

(We have 4 additional constraints.)

### 9.6.2 NCS Basis

A Natural Cubic Spline with $m$ knots is represented by $m$ basis functions, for example, one such basis is given by

$$N_1(x) = 1$$

$$N_2(x) = x$$

$$N_{k+2}(x) = d_k(x) - d_{k-1}(x)$$

where

$$d_k(x) = \frac{(x-\xi_k)_+^3 - (x-\xi_m)_+^3}{\xi_m - \xi_k}$$

Each of these derivatives can be seen to have zero second and third derivative for $x \geq \xi_m$.

### 9.6.3 Note: Waste of Data points

Recall that the **linear functions** in the two extreme intervals are completely determined by the other cubic splines. So data points in the two extreme intervals (i.e., outside the two boundary knots) are wasted since they do not affect the fitting.

## 9.7 Regression Splines

We can represent the model on the observed $n$ data points using matrix notation:

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}_{n\times 1} = \begin{pmatrix} h_0(x_1) & h_1(x_1) & \dots & h_{p-1}(x_1) \\ h_0(x_2) & h_1(x_2) & \dots & h_{p-1}(x_2) \\ & & \dots & \\ h_0(x_n) & h_1(x_n) & \dots & h_{p-1}(x_n) \end{pmatrix}_{n\times p} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_p \end{pmatrix}_{p\times 1}$$

where our design matrix is the matrix $\mathbf{F}$ of basis functions.

We can find $\beta$ by solving the problem:

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{F}\beta\|^2$$

## 9.8 $K-$Fold Cross-Validation

How to select the optimal number of knots (or df)?

**K-Fold Cross-Validation** Steps:

1. Set a fixed number of knots (or df).
2. Divide the set of observations into k groups (or folds).
3. Leave the first fold as a validation set (not used to fit the model). Fit the Regression Spline with a fixed number of knots using the remaining $k-1$ folds.
4. Calculate the *Mean Square Error* for fold 1: $MSE_1$.
5. Repeat the previous steps k times. Each time a new validation set is used to calculate $MSE_i$.
6. Calculate the average k-fold *Cross-Validation error*:

$$CV(k) = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

7. Repeat 2 to 6 with a new number of knots (or df).
8. Select the number of knots that **minimizes** the k-fold $CV$ error or $CV(k)$.

# Chapter 10 ANalysis of COVAriance (ANOVA): Basic

These are regression problems where some predictors are quantitative (i.e. numerical) and some are qualitative (i.e. categorical).

## 10.1 Two level example

For simplicity, we will focus on examples with just two predictors: $X$ (numerical) and $D$ (categorical).

$D$ has two levels: 0 or 1.

**General Model**

$$y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 (x \cdot d) + \varepsilon$$

**Model 1: Coincident regression lines**

$$y = \beta_0 + \beta_1 x + \varepsilon$$

**Model 1'**

$$y = \beta_0 + \beta_2 d + \varepsilon$$

**Model 2: Parallel regression lines**

$$y = \beta_0 + \beta_1 x + \beta_2 d + \varepsilon$$

**Model 3: Regression lines with equal intercepts but different slopes**

$$y = \beta_0 + \beta_1 x + \beta_3 (x \cdot d) + \varepsilon$$

**Model 4: Unrelated regression lines**

$$y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 (x \cdot d) + \varepsilon$$

**Hierarchical Rule** for interactions: *an interaction term will be included in a model only if all its main effects have been included.*

Due to this rule, we would include both $\beta_1$ and $\beta_2$, once $\beta_3$ is significant. So, we don't consider model 3 this place.

## 10.2 Two-level Test: t-test

**Which model to pick?**

1. test whether the interaction term is significant:

$$H_0 : \text{model 2} \quad H_\alpha : \text{model 4}$$

2. if don't reject $H_0$, test whether you can further reduce model 2 to model 1

$$H_0 : \text{model 1} \quad H_\alpha : \text{model 2}$$

## 10.3 Multi-level example

Model the response $Y$ by two predictors $X$ and $D$, where $X$ is a numerical variable and $D$ is categorical with $k$ levels.

We need to generate $k-1$ dummy variables: $D_2, \ldots, D_k$ where:

$$D_i = \begin{cases} 0, & \text{if not level i} \\ 1, & \text{if level i} \end{cases}$$

Level 1 is the reference level.

**Model 0**: $Y \sim 1$

**Model 1**: $Y \sim X$

**Model 1'**: $Y \sim D$

**Model 2**: $Y \sim D + X$

**Model 4**: $Y \sim D + X + D : X$

## 10.4 Multi-level Test: F-test

Note that when $D$ has more than two levels, the difference between model parameter number may not be one, so t-test is no longer appropriate.

1) Compare models:

$$H_0 : Y \sim X + D \quad \text{vs.} \quad H_\alpha : Y \sim D + X + D : X$$

If the interaction $D : X$ is significant, stop.

2) If $X$ is significant, keep $X$.

2′) If $D$ is significant, keep $D$.

3) If neither $X$ nor $D$ are significant, report the intercept model $Y \sim 1$

**Sequential ANOVA** We can use the anova function to get sequential F-tests. The sequence of $F$ -tests given by anova $(\text{lm}(Y \sim X + D + X : D))$

| $H_0$ | $H_\alpha$ |
|---|---|
| $Y \sim 1$ | $Y \sim X$ |
| $Y \sim X$ | $Y \sim X + D$ |
| $Y \sim X + D$ | $Y \sim X + D + X : D$ |

The sequence of $F$ -tests given by anova $(\text{lm}(Y \sim D + X + X : D))$

| $H_0$ | $H_\alpha$ |
|---|---|
| $Y \sim 1$ | $Y \sim D$ |
| $Y \sim D$ | $Y \sim D + X$ |
| $Y \sim D + X$ | $Y \sim D + X + X : D$ |

**Note**: Some of the F-stats and p-values from the sequential ANOVA table are **different** from the ones we calculated based on usual F-test (we learned) for comparing two nested models.

Suppose we want to compare:

$$H_0 : Y \sim X \quad \text{vs} \quad H_\alpha : Y \sim X + D$$

The usual $F$-stat is given by:

$$\frac{(RSS_0 - RSS_a)/(k-1)}{RSS_a/(n-p_a)} = \frac{(RSS_0 - RSS_a)/(k-1)}{\hat{\sigma}_a^2}$$

which follows $F_{k-1,n-p_a}$ under the null hypothesis. $k$ is the total number of categories of variable $D$

The $F$-stat from the sequential ANOVA table:

$$\frac{(RSS_0 - RSS_a)/(k-1)}{RSS_A/(n-p_A)} = \frac{(RSS_0 - RSS_a)/(k-1)}{\hat{\sigma}_A^2}$$

which follows $F_{k-1,n-p_A}$ under the null hypothesis, where $RSS_A$ denotes the RSS from the biggest model $Y \sim X + D + X : D$ and $p_A = 2k$

# Chapter 11 Model Adjustment: Variable Selection

## 11.1 Training and Test Errors

Training data: $(\mathbf{x}_i, y_i)_{i=1}^n$

Test data: $(\mathbf{x}_i, y_i^*)_{i=1}^n$ is an independent (imaginary) data set collected at the same location $\mathbf{x}_i$ 's (also known as in-sample prediction)

Assume the data comes from a linear model:

$\mathbf{y}_{n\times 1}, \mathbf{y}^*_{n\times 1}$ are i.i.d $\sim N_n\left(\mu, \sigma^2 \mathbf{I}_n\right)$ and $\mu = \mathbf{X}\beta$

We can also write:

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

$$\mathbf{y}^* = \mathbf{X}\beta + \varepsilon^*$$

with $\varepsilon_{n\times 1}, \varepsilon^*_{n\times 1}$ i.i.d $\sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right)$ are independent.

$$\begin{aligned}
\mathbb{E}(\text{ Test Error })^2 &= \mathbb{E}\left\|\mathbf{y}^* - \mathbf{X}\hat{\beta}\right\|^2 \\
&= \mathbb{E}\left\|(\mathbf{y}^* - \mathbf{X}\beta) + (\mathbf{X}\beta - \mathbf{X}\hat{\beta})\right\|^2 \\
&= \mathbb{E}\left\|\mathbf{y}^* - \mu\right\|^2 + \mathbb{E}\|\mathbf{X}\beta - \mathbf{X}\hat{\beta}\|^2 \\
&= \mathbb{E}\left\|\varepsilon^*\right\|^2 + \operatorname{Tr}\left(\mathbf{X}\operatorname{Cov}(\hat{\beta})\mathbf{X}^\top\right) \\
&= n\cdot\sigma^2 + \sigma^2 \operatorname{Tr}\mathbf{H} = n\cdot\sigma^2 + p.\sigma^2 \\
\mathbb{E}(\text{ Train Error })^2 &= \mathbb{E}||\mathbf{y} - \hat{\mathbf{y}}\,||^2 = \mathbb{E}||(\mathbf{I}-\mathbf{H})\mathbf{y}\|^2 \\
&= \operatorname{Tr}\left((\mathbf{I}-\mathbf{H})\operatorname{Cov}(\mathbf{y})(\mathbf{I}-\mathbf{H})^\top\right) \\
&= \sigma^2 \operatorname{Tr}((\mathbf{I}-\mathbf{H})) = (n-p)\cdot\sigma^2
\end{aligned}$$

Index each model (i.e., each subset of the p variables) by a $p$-dimensional binary vector $\gamma$ :

$$\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_p), \quad \gamma_j = 0/1$$

where $\gamma_j = 1$ indicates that $X_j$ is included in the model, and $\gamma_j = 0$ otherwise.

So there are a total of $2^p$ possible subsets or sub-models. In particular

$$\gamma = (1, 1, \ldots, 1)$$

refers to the full model including all $p$ variables (largest dim), and

$$\gamma = (0, 0, \ldots, 0)$$

refers to the intercept-only model (smallest dim).

Suppose that $\mu = \mathbf{X}\beta$ where $\mu$ is the mean of $\mathbf{y}$. If we fit the data $\mathbf{y}$ with respect to model $\gamma$, i.e., we fit a linear model with a sub-design matrix $X_\gamma$ where $X_\gamma$ contains only columns from $X$ such that $\gamma_j = 1$

We can show that the Testing Error and the Training error for model $\gamma$ are:

$$\mathbb{E}(\text{ Test Error }) = n\sigma^2 + p\sigma^2 + \text{Bias}_\gamma$$

$$\mathbb{E}(\text{ Training Error }) = n\sigma^2 - p\sigma^2 + \text{ Bias}_\gamma$$

**Bigger model (i.e., $p$ large) $\rightarrow$ small Bias, but large variance $\left(p\sigma^2\right)$;**

**Smaller model (i.e., $p$ small) $\rightarrow$ large Bias, but small variance $\left(p\sigma^2\right)$.**

So to reduce the test error (i.e., prediction error), the key is to find the best trade-off between Bias and Variance.

## 11.2 Model selection procedures

### 11.2.1 Testing-based procedures

Testing-based procedures: Select best model based on statistical tests for model comparison.

**Backward elimination**

- Start with all the predictors in the model.

1. Remove the predictor with highest $p-$value $> \alpha_0$ (most insignificant).

2. Refit the model, and repeat the above process.

3. Stop when all $p-$values $\leq \alpha_0$.

( $\alpha_0$ is often set to $15\%$ or $20\%$ which is higher than usual)

**Forward elimination**

1. Start with the intercept-only model.

2. For all predictors not in the model, check their $p$-value if being added to the model. Add the one with the lowest $p-$ value $\leq \alpha_0$ (most significant).

3. Refit the model, and repeat the above process.

4. Stop when no more predictors can be added.

**Pros and Cons of Testing-based procedures**

- Main advantage: Computation cost is low.

- Due to the "one-at-a-time" nature of adding/dropping variables, this type of procedures does not compare all possible models. So it's possible to miss the "optimal" model.

- It's not clear how to choose $\alpha_0$, the cut-off for $p$-values.

### 11.2.2 Criterion-based procedures

Criterion-based procedures: Select best model based on an information criteria (combining model fit and model complexity) for model comparison.

1. Score each model according to an information criteria
2. Use a searching algorithm to find the optimal model

Model selection criteria/scores often takes the following form:

Training error + Complexity-penalty

**Model Selection Criteria**:

**AIC/BIC**

$$AIC : -2 \times \log lik_\gamma + 2p_\gamma$$

$$BIC : -2 \times \log lik_\gamma + \log(n)p_\gamma$$

where $p_\gamma$ is the number of predictors included in model $\gamma$

For the linear regression model:

$$-2 \times \log lik_\gamma = n \log \frac{RSS_\gamma}{n}$$

The lower the AIC/BIC the better. Note that when $n$ is large, adding an additional predictor costs a lot more in BIC than AIC. So AIC tends to pick a bigger model than BIC.

**Adjusted $-R^2$ for model $\gamma$**

$$\begin{aligned} R_a^2 &= 1 - \frac{RSS/\left(n - p_\gamma - 1\right)}{TSS/(n-1)} \\ &= 1 - \left(1 - R^2\right)\left(\frac{n-1}{n - p_\gamma - 1}\right) \\ &= 1 - \frac{\hat{\sigma}_\gamma^2}{\hat{\sigma}_0^2} \end{aligned}$$

The higher the $R_a^2$ the better.

**Mallow's $C_p$**

$$C_p = \frac{RSS_\gamma}{\hat{\sigma}^2} + 2p_\gamma - n$$

where $\hat{\sigma}^2$ is the estimate of the error variance from the full model. Mallow's $C_p$ behaves very similar to AIC.

**Searching Algorithms:**

**Leap and Bounds:**

return *the global optimal solution* among all possible models, but *only feasible for less than* $50$ *variables*.

– Find the $p$ models with the smallest RSS amongst all models of the same size.

– Then evaluate the score on the $p$ models and report the optimal one.

**Greedy algorithms:**

fast, but *only return a local optimal solution* (which might be good enough in practice).

- Backward: start with the full model and sequentially delete predictors until the score does not improve.

- Forward: start with the null model and sequentially add predictors until the score does not improve.

- Stepwise: consider both deleting and adding one predictor at each stage.

# Chapter 12 Model Adjustment: Shrinkage Methods

Find a *trade-off* between *model bias* and *prediction error*.

## 12.1 Principal Components Regression (PCR)

When we have too many predictors, we need dimensionality reduction in the predictors space. Predictors might be highly correlated.

1. Take matrix X of predictors and center the columns of X to have zero mean. Consider X with no intercept column. (In order to focus on the vaiation).

2. Find directions of greater variation in the data. (Find the vector that can represnet $X$ best.)

### 12.1.1 Principal Component Analysis (PCA)

The steps to find directions of greater variation in matrix $\mathbf{X}$ :

- Find $\mathbf{u}_1$ to maximize variance of $\mathbf{u}_1^\top \mathbf{X}$ subject to $\mathbf{u}_1^\top \mathbf{u}_1 = 1$.

- Find $\mathbf{u}_2$ to maximize variance of $\mathbf{u}_2^\top \mathbf{X}$ subject to $\mathbf{u}_1^\top \mathbf{u}_2 = 0$ and $\mathbf{u}_2^\top \mathbf{u}_2 = 1$

- Continue looking for directions of greatest variation in the data which are orthogonal to the previous ones.

- Continue until the total number of dimensions is exhausted. The principal components are given by the columns of matrix $\mathbf{Z}$, where

$$\mathbf{Z} = \mathbf{XU}$$

$$[\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_m] = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m][\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_m]$$

$\mathbf{z}_i$ and $\mathbf{u}_i$ are the columns of $\mathbf{Z}$ and $\mathbf{U}$ respectively. $\mathbf{U}$ is called the **rotation matrix**. $\mathbf{Z}$ is a version of the data rotated in such a way that the resulting principal components are orthogonal.

- Each Principal Component is a linear combination of the original variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ with weights given by each column $\mathbf{u}_i$ of matrix $\mathbf{U}$ :

$$\mathbf{z}_i = u_{1i}\mathbf{x}_1 + u_{2i}\mathbf{x}_2 + \ldots + u_{mi}\mathbf{x}_m$$

- Principal components are very sensitive to outliers.

- The **Mahalanobis distance** can be used to measure the distance of a point to the data mean, after adjusting for correlation in the data.

- Under the multivariate normality assumption in $m$ dimensions, the **Mahalanobis distance**

$$d_i = \sqrt{(\mathbf{x}_i - \mu)^\top \Sigma^{-1} (\mathbf{x}_i - \mu)}$$

can be estimated using the sample estimators for $\mu$ and $\Sigma$ and the quantity $d_i^2$ follows a $\chi_m^2$ distribution. This can be used to detect outliers in higher dimensions.

### 12.1.2 After PCA

- Replace model $\mathbf{Y} \sim \mathbf{X}$ by the model $\mathbf{Y} \sim \mathbf{Z}$

- Only need to use the first few columns of $\mathbf{Z}$ as predictors

- Interpretation of the PCAs as predictors might be challenging. We need to use the values of $\mathbf{u}_i$ in the rotation matrix (also called the loadings) for interpretation.

- Sometimes we can make better predictions with a small number of PCs in $\mathbf{Z}$ than with a large number of predictors in $\mathbf{X}$

### 12.1.3 Use How many Principal Components?

- The trace of the sample variance-covariance $S$ of $Z$ (total sample variance) is equal to the sum of its eigenvalues:

$$\text{trace}(S) = s_1^2 + s_2^2 + \ldots + s_m^2 = \lambda_1 + \lambda_2 + \ldots + \lambda_m$$

Since, sample variance-covariance matrix is symmetric, the equation must hold.

- Most of the total variance of a data set is concentrated in the first principal components.

PCs have decreasing explanatory power, so we just need to use the first few:

1. Make a plot of the PCs standard deviations $\left(\sqrt{\lambda_i}\right)$ vs. the PC index $i$. This is called the scree plot.

- Look for the $PC$ index $i$ where there is a big change in slope (the elbow) in the scree plot.

**Example 12.1**

The first comp makes the first point to the second point,... , the fourth comp makes the fourth point to the fifth point, so here we only need the first four comp.

2. Another way is to calculate the cumulative variance explained by the first PCs, and retain the number of $PC$ s explaining between $70\%$ to $90\%$ of the total variation.

3. An alternative way is to discard PCs such that $\lambda_i < \bar{\lambda}$

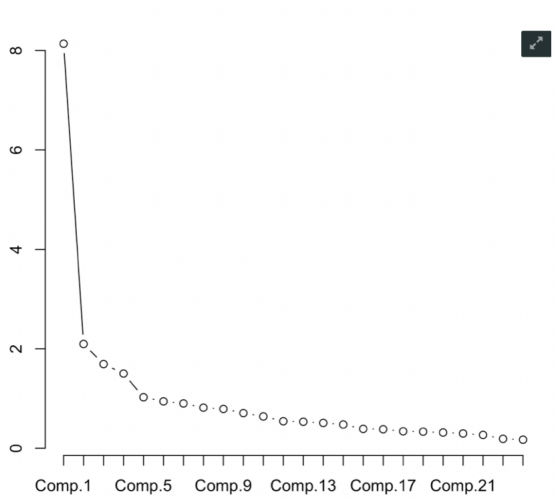


**Figure 12.1**

## 12.2 Ridge Regression

- Although the aim of PCR is to reduce dimensionality in the number of predictors, you still have to measure all the predictors since each $PC$ is a linear combination of all predictors.

- Ridge regression assumes that after normalization, some of the regression coefficients should not be very large.

- Ridge regression is very useful when you have collinearity and the LS regression coefficients are unstable.

- The method uses a **penalized regression** since the LS minimization problem has a penalty term:

$$\text{minimize}(y - X\boldsymbol{\beta})^\top (y - X\boldsymbol{\beta}) + \lambda \sum_j \boldsymbol{\beta}_j^2$$

for some $\lambda \geq 0$. The penalty term is $\sum_j \beta_j^2$

– Usually predictors are standardized first (centered by their means and scaled by their standard deviations) and the response $y$ is centered.

– The ridge regression estimates are:

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T y$$

Or, the $\beta$ minimize

$$(y - X\beta)^T (y - X\beta) \text{ subject to } \sum_j \boldsymbol{\beta}_j^2 \leq t^2$$

- The parameter $\lambda$ (or $t$) should be chosen to have stable estimates of $\beta$.

- Note that when $\lambda = 0$ the ridge regression estimation problem reduces to the standard LS problem, while when $\lambda \to \infty, \hat{\boldsymbol{\beta}} \to 0$

- It is useful to plot the values of $\hat{\beta}_j$ as a function of $\lambda$.

- The value of $\lambda$ can be also chosen using automated methods as **Generalized Cross-Validation (GCV)** (similar to Cross-Validation).

- Ridge regression coefficient estimates are **biased**.

## 12.3 Lasso Regression

- In this case the estimated $\hat{\boldsymbol{\beta}}$ minimizes:

$$\text{minimize}(y - X\boldsymbol{\beta})^{\top}(y - X\boldsymbol{\beta}) + \lambda \sum_{j} |\beta_j|$$

for some $\lambda \geq 0$. The penalty term is $\sum_j |\beta_j|$ ($L_1$ constraint)

Or, the $\beta$ minimize

$$(y - X\beta)^T(y - X\beta) \text{ subject to } \sum_{j} |\beta_j| \leq t$$

- In two-dimensions the constraint defines a square. In higher dimensions it defines a polytope.

- Lasso is useful when the response can be explained by few predictors with zero effect on the remaining predictors (Lasso is similar to a variable selection method).

- When $\beta_j = 0$ the corresponding predictor is eliminated. This is not the case for ridge regression.

- Use Lasso when the effect of predictors is **sparse**. This means that only few predictors will have an effect on the response (e.g. gene expression data) or when number of predictors is large ($p > n$)

- Use the lars $R$ package for Lasso

- Select $t$ in the constraint $\sum_{j=1}^{p} |\beta|_j \leq t$ by using **Cross-Validation (CV)**

- As $t$ increases, the number of predictors increases.

## 12.4 Comparing Ridge Regression and Lasso

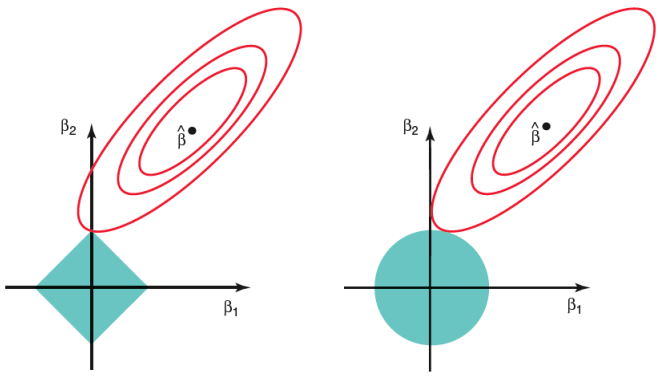


**Figure 12.2**

In 2 dimensions, Lasso is a square and Ridge is a circle.

- Lasso selects a sub-set of predictors (some coefficients equal to zero).

- Ridge regression performs better when the response is a function of many predictors with coefficients around the same size.

- Lasso will perform better when a relatively small number of predictors have large coefficients and the rest are very small or equal to zero.

- Since the number of predictors is never known a priori, cross-validation can be used to decide which approach is better for a particular data set.

# Chapter 13 ANOVA: Comparative Experiments

## 13.1 Terminology

- **Factor**: an Independent variable. They can be experimental or observational. In our example: Diet

- **Level**: A particular form of the factor. In our example: Levels of the Diet: $A, B, C, D$

- **Treatments**: Factor levels or factor level combinations (if the study contains more than one factors). They provide insights into mechanisms causing the variation being studied. Control treatments?

- **Complete Randomized Design**: Experimental units are randomly split into $r$ groups, and $r$ treatments are assigned, one per group.

## 13.2 Data

$$\begin{array}{lccccc} \text{group } 1 & y_{11}, & y_{12} & \dots & y_{1n_1} \\ \text{group } 2 & y_{21}, & y_{22} & \dots & y_{2n_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{group } r & y_{r1}, & y_{r2} & \dots & y_{rn_r} \end{array}$$

- $r$ is the number of groups

- $n_i$ denotes the number of obs in the ith group

- $n = \sum_{i=1}^{r} n_i$ is the total sample size

- $y_{ij} =$ observation $j$ for the ith factor.

## 13.3 ANOVA Model

### 13.3.1 ANOVA Means Model (Cell Means Model)

$$y_{ij} = \mu_i + \varepsilon_{ij}, i = 1, \dots, r; \quad j = 1, \dots, n_i$$

- $y_{ij}$ : the value of the response in the $j$ th trial for the ith factor.

- $\mu_i$ : the population mean for the ith factor level (treatment).

- $\varepsilon_{ij} \sim^{iid} \mathcal{N}\left(0, \sigma^2\right)$

### 13.3.2 Factor Effects Model

Define the effect of factor level $i$ on the response, i.e. the treatment effect as

$$\alpha_i = \mu_i - \mu$$

where $\mu$ is the overall mean.

Factor Effects Model:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, i = 1, \ldots, r; j = 1, \ldots, n_i$$

$$\varepsilon_{ij} \sim^{iid} \mathcal{N}\left(0, \sigma^2\right)$$

- The factor effects model has $r + 1$ model parameters, i.e.

$$(\mu, \alpha_1, \ldots, \alpha_r)$$

- In order for the $\alpha$ 's to be (uniquely) estimated, we need to impose restrictions.

- The restrictions on the $\alpha$ 's depend on how $\mu$ is defined.

| Model | $\mu$ Definition | $\alpha$ 's Restriction |
|---|---|---|
| Reference Cell | $\mu = \mu_1$ | $\alpha_1 = 0$ |
| Sum-to-Zero | $\mu = \frac{1}{r}\sum_i \mu_i$ | $\sum_i \alpha_i = 0$ |
| Weighted Sum-to-Zero | $\mu = \frac{1}{n}\sum_i n_i\mu_i$ | $\sum_i n_i\alpha_i = 0$ |

- The default in R is the Reference Cell model.

## 13.4 Model Properties

$- E\left(y_{ij}\right) = \mu_i$

$- \text{Var}\left(y_{ij}\right) = \text{Var}\left(\varepsilon_{ij}\right) = \sigma^2$

Thus, all observations have the same variance, regardless of factor level.

- $\varepsilon_{ij} \sim \mathcal{N}\left(0, \sigma^2\right)$ and independent

- $y_{ij} \sim \mathcal{N}\left(\mu_i, \sigma^2\right)$ and independent.

We can re-state the model as

$$y_{ij} \text{ are independent } \mathcal{N}\left(\mu_i, \sigma^2\right)$$

## 13.5 Model Estimation

Minimize the sum of squared deviations of the observations around their expected values with respect to the parameters:

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(y_{ij} - \mathbb{E}\left(y_{ij}\right)\right)^2$$

If we re-write $Q$ we have

$$Q = \sum_{j} \left(y_{1j} - \mu_1\right)^2 + \sum_{j} \left(y_{2j} - \mu_2\right)^2 + \ldots + \sum_{j} \left(y_{rj} - \mu_r\right)^2$$

So the **least squares estimator** of $\mu_i$, denoted by $\hat{\mu}_i$ is

$$\hat{\mu}_i = \bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

Using the appropriate constraints, we can easily extract the estimators for $\mu$ and $\alpha_i$.

- The $LS$ fit for $y_{ij}$ is the corresponding group mean

$$\hat{y}_{ij} = \bar{y}_i$$

- Residuals

$$r_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_i$$

- RSS

$$\sum_{i=1}^{r} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i\cdot}\right)^2$$

i.e. the within-group variation.

| Source of Variation | SS | df | MS |
|---|---|---|---|
| Between Groups | $FSS = \sum n_i \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2$ | $r-1$ | $\frac{FSS}{r-1}$ |
| Error (within Groups) | $RSS = \sum \sum \left(y_{ij} - \bar{y}_{i\cdot}\right)^2$ | $n-r$ | $\frac{RSS}{n-r}$ |
| Total | $TSS = \sum \sum \left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2$ | $n-1$ | |

## 13.6 F-test

- We want to test whether the means of the groups are really different. We can express this as

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_r \\ H_\alpha : \text{ not all } \mu_i, i = 1, \ldots, r \text{ are equal} \end{cases}$$

- or in terms of models

$$\begin{cases} H_0 : y_{ij} = \mu + \varepsilon_{ij} \\ H_\alpha : y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \end{cases}$$

- They are two nested models, so we can use the $F$-test

$$\frac{\left(RSS_0 - RSS_\alpha\right)/(r-1)}{RSS_\alpha/(n-r)} \sim F_{r-1,n-r}$$

under $H_0$.

- The test statistic can also be written as

$$\frac{FSS/(r-1)}{RSS/(n-r)} = \frac{\text{Between-group Variation }/(r-1)}{\text{Within-group Variation }/(n-r)}$$

where $FSS, RSS$ are defined in the ANOVA table.

## 13.7 Diagnostics for ANOVA Models

– Check for outliers/ unusual observations.

– Check the residuals vs. fitted values plot for departures from the constant variance assumption.

– Check the Q-Q plot for departures from the normality assumption.

**Levene's Test** for Equality of Variances:

– Run Regression abs(residuals)$\sim$ X, i.e. use abs(residuals) as the response in a new one-way ANOVA.

– If the p-value for the F-test is **greater** than 1% level, then we conclude that there is no evidence of a non-constant variance.

## 13.8 Inference for Factor Level Means (function about the $\mu_i$s)

### 13.8.1 A single factor level mean

- Estimation of

$$\mu_i : \hat{\mu}_i = \bar{y}_i$$

- Distribution of

$$\hat{\mu}_i : E\left(\hat{\mu}_i\right) = \mu_i, \quad \text{Var}\left(\hat{\mu}_i\right) = \frac{\sigma^2}{n_i}$$

- The estimated variance of $\bar{y}_i$. is

$$s^2_{\bar{y}_i} = \frac{1}{n_i} \cdot \frac{RSS}{n-r}$$

- Under the ANOVA model assumptions $\frac{\bar{y}_{i\cdot}-\mu_i}{s_{\bar{y}_i}}$ is distributed as $T_{n-r}$

- Confidence Interval for $\mu_i$ :

$$\mu_i \in \bar{y}_{i\cdot} \pm T_{n-r}(\alpha/2)s_{\bar{y}_i}$$

### 13.8.2 A difference between two factor level means

The difference between two factor level means (pairwise comparison) is defined as

$$D = \mu_i - \mu_{i'}$$

- Estimation of $D$:

$$\hat{D} = \bar{y}_i - \bar{y}_{i'}$$

- Distribution of $\hat{D}$:

$$E(\hat{D}) = \mu_i - \mu_{i'}, \mathrm{Var}(\hat{D}) = \sigma^2\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)$$

The estimated variance of $\hat{D}$ is

$$s^2_{\hat{D}} = \frac{RSS}{n-r} \cdot \left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)$$

- Under the ANOVA model assumptions

$$\frac{\hat{D}-D}{s_{\hat{D}}} \text{ is distributed as } T_{n-r}$$

- Confidence Interval for $D$:

$$D \in \hat{D} \pm T_{n-r}(\alpha/2)s_{\hat{D}}$$

- Hypothesis Test for $D$ :

$$\begin{cases} H_0 : \mu_i = \mu_{i'} \\ H_\alpha : \mu_i \neq \mu_{i'} \end{cases} \Leftrightarrow \begin{cases} H_0 : \mu_i - \mu_{i'} = D = 0 \\ H_\alpha : \mu_i - \mu_{i'} \neq 0 \end{cases}$$

The test statistic is

$$t = \frac{\hat{D}}{s_{\hat{D}}} \sim T_{n-r}$$

### 13.8.3 A contrast among factor level means

A contrast is a comparison involving two or more level means:

$$L = \sum_{i=1}^{r} c_i\mu_i, \quad \text{where } \sum_{i=1}^{r} c_i = 0$$

- Estimation of $L$:

$$\hat{L} = \sum_{i=1}^{r} c_i\bar{y}_i$$

- Distribution of $\hat{L}$:

$$E(\hat{L}) = \sum_{i=1}^{r} c_i \mu_i, \mathrm{Var}(\hat{L}) = \sigma^2 \sum_{i=1}^{r} \frac{c_i^2}{n_i}$$

The estimated variance of $\hat{L}$ is

$$s_{\hat{L}}^2 = \frac{RSS}{n-r} \cdot \sum_{i=1}^{r} \frac{c_i^2}{n_i}$$

- Under the ANOVA model assumptions $\frac{\hat{L}-L}{s_{\hat{L}}}$ is distributed as $T_{n-r}$

- Confidence Interval for $L$:

$$L \in \hat{L} \pm T_{n-r}(\alpha/2) s_{\hat{L}}$$

- Hypothesis Testing for $L$ :

$$\begin{cases} H_0 : L = 0 \\ H_\alpha : L \neq 0 \end{cases}$$

The test statistic is

$$t = \frac{\hat{L}}{s_{\hat{L}}} \sim T_{n-r}$$

### 13.8.4 A linear combination of factor level means

$$L = \sum_{i=1}^{r} c_i \mu_i, \quad \text{no restrictions on } c_i' s$$

- Point estimator and estimated variance same as before.

- Single Degree of Freedom Tests

$$\begin{cases} H_0 : L = c \\ H_\alpha : L \neq c \end{cases}$$

The test statistic here is

$$F = t^2 = \left( \frac{\hat{L} - c}{s_{\hat{L}}} \right)^2 \sim F_{1,n-r}$$

## 13.9 Limitations of Inference Procedures

The confidence coefficient $1 - \alpha$ for the estimation procedures described is a statement confidence coefficient and applies **only to a particular estimate**, **not to a series of estimates**.

Similarly the specified Type I error rate $\alpha$ applies only to a particular test and not to a series of tests.

## 13.10 Bonferroni Correction $\frac{\alpha}{m}$

**Example 13.1** If the confidence coefficents are $95\%$ for all individual $\mu_i$, the confidence coefficient will be $(95\%)^n < 95\%$ for family $f(\mu_1, ..., \mu_n)$.

So the $95\%$ confidence interval of family will be **wider** than individual.

When? The family of interest is a particular set of pairwise comparisons, contrasts, or linear combinations that is specified by the user.

- Suppose $m$ is the number of statements in the family.

- In order to control the family wise error rate to be $\alpha$, we need to reduce the error rate for each individual comparison to be $\alpha/m$.

- That is we need to increase the significance level from $(1-\alpha)$ to $(1-\alpha/m)$

- Not applicable when $m$ is large, since the Cls would be too wide due to the increase of the significant level.

i.e. We use $1 - \frac{5\%}{n}$ for all $\mu_i$ to form $95\%$ $f(\mu_1, ..., \mu_n)$

## 13.11 Tukey's Paired Comparison Procedures

When? the family of interest is a set of all pairwise comparisons of factor level means, i.e. it consists of estimates of all pairs $D = \mu_i - \mu_{i'}$

A confidence interval is given by

$$D \in \hat{D} + \frac{q(\alpha/2; r, n-r)}{\sqrt{2}} s(\hat{D})$$

where $q(\alpha/2; r, n-r)$ refers to the $\alpha/2$ upper quantile of the studentized range for $r$ means and $n-r$ degrees of freedom.

The coverage probability is exact when the sample sizes in each group are identical and is approximate otherwise.

Remark: The studentized range refers to the distribution of

$$\max_{i \neq j} \sqrt{n} \left(\bar{y}_i - \bar{y}_j\right) / \hat{\sigma}$$

where $\bar{y}_i$ and $\bar{y}_j$ are sample means from independent samples of size $n$ from normal distributions with common means and variance $\sigma^2$.

**Note**: Tukey is always better than Bonferroni and Scheffe in pairwise comparisons.

## 13.12 Scheffe's Method for Contrasts

When? The family of interest is the set of contrasts among the factor level means:

$$L = \sum c_i \mu_i, \text{ where } \sum c_i = 0$$

An confidence interval is given by

$$L \in \hat{L} + (r-1) F_{r-1,n-r}(\alpha) s_{\hat{L}}$$

# Chapter 14 Two Way ANOVA

Single-factor:

1. Do not explore the entire space of treatment combinations.
2. Interactions cannot be estimated.
3. Full randomization is not possible.
4. Multiple stages increase complexity of the analysis.

MultiFactor:

1. Efficient replication.
2. Assessment of Interactions.
3. Validity of Findings.

## 14.1 Factor Effects Model for Two Factors

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

| | Sum | Average |
|---|---|---|
| Cell $(i,j)$ | $y_{ij.} = \sum_{k=1}^{n} y_{ijk}$ | $\bar{y}_{ij.} = \frac{y_{ij}}{n}$ |
| Row $i$ | $y_{i..} = \sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}$ | $\bar{y}_{i..} = \frac{y_{i..}}{bn}$ |
| Column $j$ | $y_{j.j} = \sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk}$ | $\bar{y}_{j.} = \frac{y_{j.j}}{an}$ |
| Overall | $y_{...} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}$ | $\bar{y}_{...} = \frac{y_{...}}{nab}$ |

- Using least squares method, the estimated treatment means are:

$$\hat{\mu}_{ij} = \bar{y}_{ij}$$

- The factor effects estimators depend on the constraints that we impose. For example, under the sum-constraints we have

$$\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}\ldots, \quad \hat{\beta}_j = \bar{y}_{.j.} - \bar{y}\ldots$$

$$\hat{(\alpha\beta)}_{ij} = \bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}..$$

- The fitted values and residuals compute as usual as

$$\hat{y}_{ijk} = \bar{y}_{ij}, \quad r_{ij} = y_{ijk} - \hat{y}_{ijk}$$

## 14.2 Interaction Plots

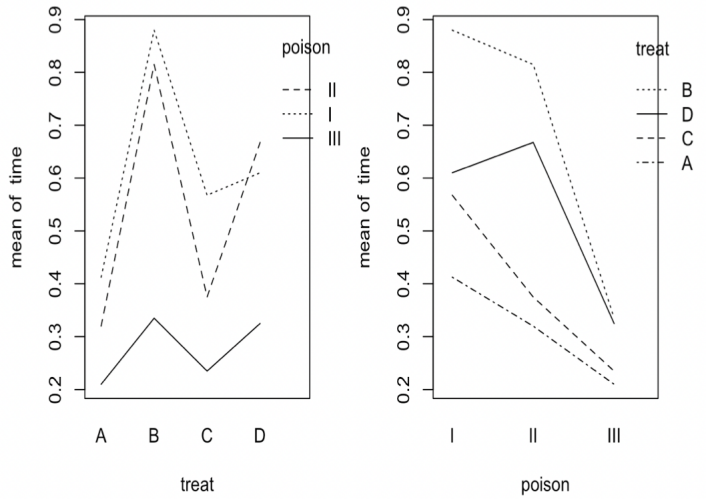


**Figure 14.1**

If the lines are not parallel, interaction is presented.

## 14.3 Partitioning of Total Sum of Squares

$$\underbrace{y_{ijk} - \bar{y}_{...}}_{\text{Total Deviation}} = \underbrace{\bar{y}_{ij} - \bar{y}_{...}}_{\substack{\text{Deviation of estimated} \\ \text{treatment mean around} \\ \text{overall mean}}} + \underbrace{y_{ijk} - \bar{y}_{ij.}}_{\substack{\text{Deviation} \\ \text{around estimated} \\ \text{treatment mean}}}$$

$$TSS = FSS + RSS$$

where

$$TSS = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2$$

$$FSS = n \sum_i \sum_j (\bar{y}_{ij\cdot} - \bar{y}_{...})^2$$

$$RSS = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij.})^2 = \sum_i \sum_j \sum_k e_{ijk}^2$$

## 14.4 Partitioning of Treatment Sum of Squares

$$\underbrace{\bar{y}_{ij.} - \bar{y}_{...}}_{\text{Deviation of estimated treatment mean around overall mean}} = \underbrace{\bar{y}_{i..} - \bar{y}_{...}}_{\text{A main effect}} + \underbrace{\bar{y}_{.j.} - \bar{y}_{...}}_{\text{B main effect}} + \underbrace{\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}}_{\text{A B interaction effect}}$$

$$FSS = SSA + SSB + SSAB \text{ (Orthogonal Decomposition)}$$

where

$$SSA = nb \sum_i (\bar{y}_{i..} - \bar{y} \ldots)^2$$

$$SSB = na \sum_j (\bar{y}_{.j.} - \bar{y} \ldots)^2$$

$$SSAB = n \sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y} \ldots)^2$$

## 14.5 ANOVA Table

| Source of Variation | SS | df | MS |
|---|---|---|---|
| Factor A | SSA | $a-1$ | MSA $= \frac{SSA}{a-1}$ |
| Factor B | SSB | $b-1$ | MSB $= \frac{SSB}{b-1}$ |
| AB Interactions | SSAB | $(a-1)(b-1)$ | MSAB $= \frac{SSAB}{(a-1)(b-1)}$ |
| Error | RSS | $ab(n-1)$ | MSE $= \frac{RSS}{ab(n-1)}$ |
| Total | TSS | nab $-1$ | |

## 14.6 F-test

- In order to test for the statistical significance of the interaction terms, we use partial $F$-tests. So, we fit a main effects model (i.e. no interactions)

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

- Then, we compare the two nested models:

$$\begin{cases} H_0: \text{ smaller model with } p_0 \text{ coefficients} \\ H_\alpha: \text{ larger model with } p_\alpha \text{ coefficients} \end{cases}$$

The $F$-test is formulated as

$$F = \frac{(RSS_0 - RSS_\alpha)\,/\,(p_\alpha - p_0)}{MSE_\alpha} \sim F_{p_\alpha - p_0, n - p_\alpha} \text{ under the } H_0$$

We can also perform F-tests directly using the ANOVA table, where for the interaction term we have:

$$F_{AB} = \frac{MSAB}{MSE} \sim F_{(a-1)(b-1),nab-1}$$

Hierarchy principle: we test for main effects only if the interaction term is not statistically significant.

## 14.7 Estimation of Factor Level Means

When interactions are not statistically significant, we analyze the factor level means:

- Factor Level Means:

$$\hat{\mu}_{i\cdot} = \bar{y}_{i\cdot\cdot}, s^2_{\hat{\mu}_{i\cdot}} = \frac{MSE}{bn}$$

- Differences of Factor Level Means:

$$\hat{\mu}_{i\cdot} - \hat{\mu}_{i'\cdot} = \bar{y}_{i\cdot\cdot} - \bar{y}_{i'\cdot\cdot}, s^2_{\hat{D}} = \frac{2MSE}{bn}$$

- Contrasts of Factor Level Means:

$$\hat{L} = \sum c_i \hat{\mu}_{i\cdot} = \sum c_i \bar{y}_{i\cdots}, s^2_{\hat{L}_{(i)}} = \frac{MSE}{bn} \sum c_i^2$$

where $\sum c_i = 0$

For individual hypothesis test and Cls, the multiplier is $T_{(n-1)ab}(\alpha/2)$.

For family hypothesis tests/intervals, we select the desired family multiplier:

- Tukey Multiplier:

$$\frac{1}{\sqrt{2}} q_{a,(n-1)ab}(1-\alpha)$$

- Bonferroni Multiplier: $B = T_{(n-1)ab}(1 - \alpha/2m)$, where $m$ refers to the number of multiple comparisons.

- Scheffé Multiplier:

- $S^2 = (b-1)F_{b-1,(n-1)ab}(1-\alpha)$, if the contrasts involve the $\mu_{i\cdot}$ and

- $S^2 = (a-1)F_{a-1,(n-1)ab}(1-\alpha)$, if the contrasts involve the $\mu_{\cdot j}$.

## 14.8 Estimation of Treatment Means

When interactions are statistically significant, we analyze the treatment means:

- Treatment Means:

$$\hat{\mu}_{ij} = \bar{y}_{ij.}, \quad s^2_{\hat{\mu}_{ij}} = \frac{MSE}{n}$$

- Differences of Treatment Means:

$$\hat{D} = \hat{\mu}_{ij} - \hat{\mu}_{i'j'} = \bar{y}_{ij.} - \bar{y}_{i'j'.}, \quad i, j \neq i', j' \text{ and } s^2_{\hat{D}} = \frac{2MSE}{n}$$

- Contrasts of Treatment Means:

$$\hat{L} = \sum\sum c_{ij}\hat{\mu}_{ij} = \sum\sum c_{ij}\bar{y}_{ij}, \text{ where } \sum\sum c_{ij} = 0$$

with variance $s^2_{\hat{L}} = \frac{MSE}{n}\sum c^2_{ij}$

For individual hypothesis test and CIs, the multiplier is $T_{(n-1)ab}(\alpha/2)$.

For family hypothesis tests/intervals, we choose the desired family multiplier:

- Tukey Multiplier: $\frac{1}{\sqrt{2}}q_{ab,(n-1)ab}(1-\alpha)$

- Bonferroni Multiplier: $B = T_{(n-1)ab}(1-\alpha/2m)$, where $m$ refers to the number of multiple comparisons.

- Scheffé Multiplier: $S^2 = (ab-1)F_{ab-1,(n-1)ab}(1-\alpha)$, if the contrasts involve the $\mu_{i..}$

# Chapter 15 Two Way ANOVA: Special Cases

## 15.1 Unbalanced ANOVA (Use Partial $F-$Test or ANOVA type III in R)

– **When the treatment sample sizes are unequal**, the analysis of variance for two-factor studies becomes more complex.

– The least-squares equations are no longer of a simple structure and the regular analysis of variance formulas are now inappropriate.

– Furthermore, the factor effect component sum of squares are no longer orthogonal; that is, they **do not sum up to TSS**.

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

– Due to the lack of orthogonality, the **ANOVA F-tests are not applicable**.

### 15.1.1 Use Partial $F-$Test

– We will express the ANOVA model as a regression model with **indicator (dummy) variables**.

– We need $a-1$ indicator variables for factor $A$ main effects and $b-1$ indicator variables for factor $B$ main effects. The interactions correspond to the cross products of the indicator variables for A and B.

Use **Partial $F-$Test**.

### 15.1.2 ANOVA type III in R

This type tests for the presence of an effect given that both the other effects are in the model.

**Example 15.1** Compute SSTotal

$$\begin{aligned} RSS_0 - RSS_\alpha =& (SSA + SSR) - SSR \quad && = SSA = 2.1340 \\ & (SSB + SSR) - SSR && = SSB = 11.7375 \\ & (SSAB + SSR) - SSR && = SSAB = 1.9800 \end{aligned}$$

$$SST = 2.1340 + 11.7375 + 1.9800 + 7.9151 = 23.7666$$

```
Anova(lm(1/time ~ treat*poison, data=newrats), type="III")
```

```
## Anova Table (Type III tests)
##
## Response: 1/time
##               Sum Sq Df F value    Pr(>F)
## (Intercept)  15.0605  1 66.5967 1.298e-09 ***
## treat         2.1340  3  3.1455   0.03723 *
## poison       11.7375  2 25.9514 1.225e-07 ***
## treat:poison  1.9800  6  1.4592   0.22073
## Residuals     7.9151 35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Figure 15.1**

## 15.2 Balanced ANOVA with $n = 1$ (Tukey's Test)

- Only one observation in each cell, so we cannot fit the interaction model.

- There are no degrees of freedom left for estimating the error.

- RSS = 0 when the model includes main effects and interaction term. (error is 0)

- All F-tests are valid, but the interaction model is not a candidate model.

### 15.2.1 Tukey's Test for Additivity

Consider the following model that includes interactions:

$$y_{ij} = \mu + \alpha_i + \beta_j + \theta\alpha_i\beta_j + \varepsilon_{ij}$$

Here, we assume that the interactions are of *multiplicative* nature, i.e.

$$(\alpha\beta)_{ij} = \theta\alpha_i\beta_j$$

- Consider the SSA, SSB as before and:

$$SSAB^* = \frac{\left(\sum_i \sum_j (\bar{y}_i - \bar{y}_{..})(\bar{y}_{.j} - \bar{y}_{..})\, y_{ij}\right)^2}{\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2}$$

- The TSS is computed as usual and is decomposed as

$$TSS = SSA + SSB + SSAB^* + SSRem^*$$

where the remainder is

$$SSRem^* = TSS - SSA - SSB - SSAB^*$$

- We want to test the following hypothesis

$$\begin{cases} H_0 : \theta = 0 & \text{(no interactions)} \\ H_\alpha : \theta \neq 0 & \text{(interactions)} \end{cases}$$

which is essentially a test for model additivity. - The test statistic computes as

$$F^* = \frac{SSAB^*/1}{SSRem^*/(ab - a - b)}$$

# Chapter 16 Experimental Designs: Introduction

## 16.1 Experimental vs. Observational Study

*Experimental Study* is a scientific procedure undertaken to make a discovery, test a hypothesis or verify a claim. *Observational Study* is one in which the experimenter observes the effect of a factor on the response, or measures an outcome without an attempt to affect the outcome by intervention.

## 16.2 Principles of Experimental Design

Randomization

- Random allocation of treatment and order
- Ensures that collected data are IID random variables
- Averages-out the effects of exogenous factors.

Replication

- Estimate of experimental error
- Higher Precision

Blocking

- Higher precision when comparisons of factors are made
- Reduced variability transmitted from nuisance factors.

## 16.3 Randomization Test

- A test based directly on re-randomizing - with the same kind of randomization originally used to assign the treatments - is called a **randomization test**.

- *Advantage*: No need for any distributional assumptions (independence, normality, etc.) - just need to assume that the treatment randomization was performed properly.
- *Disadvantage*: Requires more computation, and you must implement for yourself or use specialized software.

# Chapter 17 Experimental Designs: Blocking

## 17.1 Randomized Complete Block Design (RCBD) Model

Example:

- Treatment Factor ‘variety‘: 8 levels
- Block Factor ‘block‘: 5 levels
- Observe that we have *only one* observation per treatment-block combination.

```
xtabs(yield~variety + block, oatvar)
```

```
##         block
## variety   I  II III  IV   V
##       1 296 357 340 331 348
##       2 402 390 431 340 320
##       3 437 334 426 320 296
##       4 303 319 310 260 242
##       5 469 405 442 487 394
##       6 345 342 358 300 308
##       7 324 339 357 352 220
##       8 488 374 401 338 320
```

**Figure 17.1**

Suppose that there are $r$ treatments (factor levels) and $n_b$ blocks.

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

where

- $\mu$. is a constant

- $\tau_i$ are the treatment effects

- $\beta_j$ are the block effects

- $\varepsilon_{ij}$ are independent $\mathcal{N}\left(0, \sigma^2\right)$

- $i = 1, \ldots, n_b$ (total number of blocks), $j = 1, \ldots, r$ (total number of treatments)

Remarks

- $y_{ij}$ is the response for the $j$ th treatment in the ith block.

- There is a single observation per block. This implies that we have a limited ability to detect an interaction between treatment and block. So, we are working with the additive model.

- We can check for treatment and block main effects, but blocking is a feature of the design which means that if insignificant, we cannot gain the degrees of freedom.

ANOVA Display for the RCBD (Two factors: Treatments and Blocks)

| Source of Variation | Sum of Squares | Degrees of Freedom | Mean Square | $F_0$ |
|---|---|---|---|---|
| Treatments | $SS_{\text{Treatments}}$ | $r-1$ | $\frac{SS_{\text{Treatments}}}{r-1}$ | $\frac{MS_{\text{Treatments}}}{MS_g}$ |
| Blocks | $SS_{\text{Blocks}}$ | $n_b-1$ | $\frac{SS_{\text{Blocks}}}{n_b-1}$ | |
| Error | $SS_E$ | $(r-1)(n_b-1)$ | $\frac{SS_g}{(r-1)(n_b-1)}$ | |
| Total | $SS_T$ | $rn_b-1$ | | |

## 17.2 Latin Squares

### 17.2.1 Example

- A, B, C: 3 treatments

- Day of week (Monday, Wednesday, Friday): Blocking Variable

- Operator ID: 1, 2, 3: Blocking Variable

| | Operator | | |
|---|---|---|---|
| Day | 1 | 2 | 3 |
| Monday | $B$ | $A$ | $C$ |
| Wednesday | $A$ | $C$ | $B$ |
| Friday | $C$ | $B$ | $A$ |

* Each operator runs each treatment, and all treatments are run on each day

### 17.2.2 Features of a Latin Square Design

- There are $r$ treatments.

- There are 2 blocking variables, each containing $r$ classes.

- Each row and each column in the design square contains all treatments.

- Each treatment is assigned to each block only once.

Advantages

- Reduces more experimental error than with 1 blocking factor.

- Small scale studies can isolate important treatment features.

- Repeated measures designs can remove order effects.

Disadvantages

- Each blocking factor must have $r$ levels.

- No interactions among factors.

- With small $r$, we have very few error degrees of freedom.

- Complex Randomization.

### 17.2.3 Randomization in Latin Square Designs

- Determine $r$, the number of treatments, row blocks, and column blocks.

- Select a Standard Latin Square (from tables or with software).

- Use Capital Letters to represent treatments $(A, B, C, \ldots)$ and randomly assign treatments to labels.

- Randomly assign Row Block levels to Square Rows.

- Randomly assign Column Block levels to Square Columns.

### 17.2.4 Latin Square Model

$$y_{ijk} = \mu + \tau_i + \beta_j + \gamma_k + e_{ijk}$$

where

- $\mu$ is a constant

- $\tau_i$ treatment effect (latin letter)

- $\beta_j$ (column) blocking effect

- $\gamma_k$ (row) blocking effect

- $e_{ijk}$ are independent $\mathcal{N}\left(0, \sigma^2\right)$

- $i, k, j = 1, \ldots, r$

NOVA Display for the Latin Square Model (Three factors: Treatments, Rows and Cols)

| AOV | $df$ |
|---|---|
| Rows (blocks) | $r-1$ |
| Cols (blocks) | $r-1$ |
| Treatments | $r-1$ |
| Error | $(r-1)(r-2)$ |
| Total | $\left(r^2-1\right)$ |

## 17.3 Balanced Incomplete Block Design (BIBD)

- [$\rightarrow$] Why Balanced?

Each pair of treatments occur together $\lambda$ times.

- [$\rightarrow$] Why Incomplete?

Cannot fit all treatments in each block.

Notation

- $t$ treatments

- $b$ blocks

- $k$ treatments per block (block size)

- $r$ times each treatment occurs

- $N = t \cdot r = b \cdot k$ observations in total

- Treatment $i$ occurs in $r$ blocks.

- To have balance, each other treatment is equally likely to be treatment $i$ in a block.

- Since there are $k-1$ other units in a block and $t-1$ other treatments, the number of times each pair of treatments appears in the same block is

$\lambda =$ the probability that there existst treatment j in remaining units in one block $\times$ the number of blocks that have treatment

$$= \frac{k-1}{t-1} r = \frac{r(k-1)}{t-1}$$

where $\lambda$ is an integer.

Examples

- $t = 3, b = 3, k = 2 \rightarrow r = 2, \lambda = 1$.

$t$ =# treatments $\{A, B, C\} = 3$

The form of square: $k \times b = 2 \times 3$.

$r$ =# A occurs=# B occurs=# C occurs= 2

$\lambda$ =# A and B in one block=# A and C in one block=# B and C in one block= 1

| Block | | |
|---|---|---|
| 1 | 2 | 3 |
| A | B | A |
| B | C | C |

- $t = 4, k = 2, b = 6 \rightarrow r = 3, \lambda = 1$

| Block | | | | | |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 |
| A | A | A | B | B | C |
| B | C | D | C | D | D |

## 17.4 BIBD Remarks

Advantages:

- A BIBD enables us to run an experiment when the size of the available blocks of experimental units is smaller than the number of treatments.

- Estimates of treatment effects have equal precision and expressions for the variances of the estimated cell means and of contrasts of treatment means or effects are relatively simple.

- The presence of balance permits the use of Scheffé and Tukey procedures for the analysis of treatment effects.

Disadvantages:

- BIBD exist only for certain combinations of numbers of treatments, block sizes, and numbers of blocks.

- The assumption that there are no interactions between the blocking variable and the treatments is restrictive.

- The analysis of a BIBD is more complex than that of a RCBD.

$$y_{ij} = \mu + \tau_i + \beta_j + e_{ij}$$

- $\mu$ constant

- $\tau_i$ treatment effects

- $\beta_j$ the block effects

- $e_{ij}$ independent $N\left(0, \sigma^2\right)$

Remarks

- Not all $y_{ij}$ exist, because of incompleteness.

- Non-orthogonality of treatments and blocks.

# Chapter 18 Linear Models with Random Effects

## 18.1 Random Effects

One-way ANOVA model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Previously, we assumed the $\alpha_i$ s were parameters: fixed, unknown values. (We also gave a restriction, to make them identifiable.) These are fixed effects, corresponding to a fixed factor.

Now suppose that the $\alpha_i$ s are unobserved random variables:

$$\alpha_i \sim N\left(0, \sigma_\alpha^2\right)$$

and assume they are independent of each other and of the $\varepsilon_{ij}$ s.

These $\alpha_i$ s are called random effects, and the corresponding factor variable is a random factor.

- **Fixed effects** are appropriate when the levels of the factor are individually important or meaningful (e.g. treatments in a designed experiment, levels of education).

- **Random effects** are appropriate when the levels of the factor are meaningful only as representatives of a more general collection (e.g. as if sampled from a population, or representative of some hypothetical population).

- For the one-way ANOVA model with a random factor, the random effects satisfy

$$\mathrm{E}\left(\alpha_i\right) = 0 \quad \mathrm{Var}\left(\alpha_i\right) = \sigma_\alpha^2$$

so random effects contribute only to the variance structure of the model, not to the mean structure.

- The parameter $\sigma_\alpha^2$ is generally unknown, and we usually seek to estimate it ( or $\sigma_\alpha$) and test the null hypothesis $\sigma_\alpha^2 = 0$.

Parameters like $\sigma_\alpha^2$ (and $\sigma^2$) are called **variance components**.

## 18.2 Intraclass Correlation

- Under this model, the responses can be correlated:

$$\mathrm{Cov}\left(y_{ij}, y_{ij'}\right) = \sigma_\alpha^2 \quad \text{for } j \neq j'$$

So different observations from the same "class" (same level $i$ of the random factor) may have a nonzero correlation.

- The **intraclass correlation coefficient (ICC)** is the correlation between $y_{ij}$ and $y_{ij'}$ (for any $i$ and $j \neq j'$ ):

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma^2}$$

## 18.3 Mixed Models

- A **fixed effects model** has only fixed factors.

- A **random effects model** has only random factors.

- A **mixed (effects) model** has both fixed and random factors.

The general form (matrix-vector):

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{Z}_\gamma + \varepsilon$$

where **X** and **Z** are known design matrices, $\beta$ contains the fixed effect (mean-related) parameters, and

$$\gamma \sim N\left(0, \sigma^2 D\right) \quad \text{independent of} \quad \varepsilon \sim N\left(0, \sigma^2 I\right)$$

are the random effects and the errors, with D containing the (unknown) random effect parameters.

The variance components (in D) are typically estimated via one of three different methods:

- **ANOVA estimation**, based on quantities in an ANOVA table; complicated for general models

- **maximum likelihood**

- **restricted maximum likelihood (REML)**, generally less biased than maximum likelihood

For balanced data, REML and ANOVA estimation tend to coincide.

## 18.4 REML: restricted maximum likelihood

R package lme4 with mixed model function lmer

```
library(lme4)

## Loading required package:  Matrix
```

REML estimation

```
milk.reml = lmer(bac ~ (1|shipment), data=milk)
```

**Figure 18.1**

```
summary(milk.reml)

## Linear mixed model fit by REML ['lmerMod']
## Formula: bac ~ (1 | shipment)
##    Data: milk
##
## REML criterion at convergence: 184.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.66731 -0.60005  0.06269  0.50602  2.14554
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  shipment (Intercept) 29.74    5.454
##  Residual             22.29    4.721
## Number of obs: 30, groups:  shipment, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   15.167      2.587   5.863
```

**Figure 18.2**

- The formula bac (1—shipment) specifies the one-way random effects ANOVA model. As usual, there is an automatically-added intercept (representing $\mu$ ), and the term (1—shipment) represents the random effect term $\alpha_i$.

- Function Imer uses the REML method, by default. We see that the REML estimates of the variance components are

$$\widehat{\sigma}_\alpha^2 \approx 29.74 \quad \widehat{\sigma}^2 \approx 22.29$$

(The Std.Dev. column simply gives the square roots of these: $\widehat{\sigma}_\alpha$ and $\widehat{\sigma}$.)

- The only fixed effect is the intercept, $\mu$.

## 18.5 ML Estimation

```
# ML estimation

milk.ml = lmer(bac ~ (1|shipment), data=milk, REML=FALSE)
```

**Figure 18.3**

- So the MLE for $\sigma_\alpha^2$ is

$$\widehat{\sigma}_\alpha^2 \approx 23.05$$

```
summary(milk.ml)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: bac ~ (1 | shipment)
##    Data: milk
##
##      AIC      BIC   logLik deviance df.resid
##    194.1    198.3    -94.1    188.1       27
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.61636 -0.63533  0.08278  0.48918  2.19649
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  shipment (Intercept) 23.05    4.801
##  Residual             22.29    4.721
## Number of obs: 30, groups:  shipment, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   15.167      2.314   6.555
```

**Figure 18.4**

smaller than the REML estimate 29.74. MLEs for variance components are often biased low.

- The estimate listed for $\sigma^2$ is apparently not the MLE, but is still the REML estimate.

- In this case, the estimate of $\mu$ has remained unchanged, but its standard error has changed.

## 18.6 Testing and Confidence Intervals

- For the fixed effects (in $\beta$ ), likelihood ratio tests are available. (For this to work, the variance components should be estimated with MLE, not REML.)

(These LRTs are sometimes unreliable, so a parametric bootstrap approach can be used - later.)

- The methods of generalized least squares ( $F$-tests and $t$-tests) could alternatively be used (though this would ignore the additional uncertainty of replacing $\mathbf{D}$ with $\widehat{\mathbf{D}}$ ).

- There are also confidence intervals for fixed effect parameters based on the Wald approach or (perhaps more reliably) on profile likelihood.

## 18.7 Testing the Random Effect Variance

- For the random effects, the null hypothesis is usually that a variance component equals zero.

- For technical reasons, the usual chi-square approximation in the LRT often fails to be adequate (most often

leading to a test that is too conservative).

- An improvement is to use the parametric bootstrap to perform the LRT (see example later).

- Methods based on ANOVA are also available, and can be useful in single-factor or balanced cases.

- Profile likelihood confidence intervals for variance components can be computed (but may have problems, as the LRT does).

## 18.8 Parametric Bootstrap

The parametric bootstrap may be more accurate in small samples. Here are the steps:

1. Compute the LR statistics for the null and alternative models
2. Generate data under the null hypothesis model
3. Fit the null and alternative model for the generated data
4. Compute the LR statistic
5. Repeat steps 2 to 4 many times
6. Find the Bootstrap probability of exceeding the observed LR value