MAT 3375
Regression Analysis

Chapter 3
Multiple Linear Regression

P. Boily (uOttawa)

Summer – 2023

Outline

Outline (cont.)

3 – Régression linéaire multiple

In practice, the situation is usually more complicated; there could be p **predictors** X_k, $k = 0, \ldots, p-1$.

Examples:

- X_1: age, X_2: sex; Y: height $(p = 3)$
- X_1: age; X_2: years of education, Y: salary $(p = 3)$
- X_1: income; X_2: infant mortality; X_3: fertility rate, Y: life expectancy $(p = 4)$
- etc.

In theory, we hope that there might be a **functional relationship** $Y = f(X_0, \ldots, X_{p-1})$ between $X_0(=1), X_1, \ldots, X_{p-1}$ and Y.

In practice (assuming that a relationship even exists), the best that we may be able to hope for is a **statistical relationship**

$$Y = f(X_0, X_1, \ldots, X_{p-1}) + \varepsilon,$$

where, as before, $f(X_0, X_1, \ldots, X_{p-1})$ is the **response function**, and ε is the **random error** (or noise).

In **general linear regression**, we assume that the response function is

$$f(X_0, X_1, \ldots, X_p) = \beta_0 X_0(=1) + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}.$$

The building blocks of regression analysis are the **observations**:

$$(X_{i,0}(=1), X_{i,1}, \ldots, X_{i,p-1}, Y_i), \quad i = 1, \ldots, n.$$

In an ideal setting, these observations are **(jointly) randomly sampled**, according to some appropriate design (which is a topic for other courses).

The **general linear regression model** (GLRM) is

$$Y_i = \beta_0 X_{i,0}(=1) + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where β_k, $k = 0, \ldots, p-1$ are **unknown parameters** and ε_i is the **random error on the ith observation** (or case).

Note that a predictor X_k can be a function of other predictors.

For instance, the following model is a GLR model:

$$\mathrm{E}\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2.$$

A GLR model need not necessarily be linear in X, but the mean response $\mathrm{E}\{Y\}$ must be **linear in the parameters** β_k, $k = 0, \ldots, p-1$.

In what follows, we write

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p-1} \end{pmatrix},$$

for the **response vector**, the **parameter vector**, and the **design matrix**, respectively.

In the design matrix $\mathbf{X}$, $\mathbf{X}_i$ represents the ith case (the ith row of $\mathbf{X}$), a single **multiple predictor level**.

The columns of the design matrix represent the values taken by the various predictor variables for all cases.

The **multiple linear regression model** is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{where} \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n).$$

Note that the SLR model fits into this framework, if we use $p = 2$ with

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & X_{1,1} \\ \vdots & \vdots \\ 1 & X_{n,1} \end{pmatrix}.$$

3.1 – Least Squares Estimation

We treat the predictor values $X_{i,k}$ as constant, for $i = 1, \ldots, n$, $k = 0, \ldots, p-1$ (i.e., we assume that there is **no measurement error**).

Since $\mathrm{E}\{\varepsilon_i\} = 0$, the **expected** (or mean) **response given** X_i is thus

$$\mathrm{E}\{Y_i \mid \mathbf{X}_i\} = \mathrm{E}\{\mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i \mid \mathbf{X}_i\} = \mathbf{X}_i\boldsymbol{\beta} + \mathrm{E}\{\varepsilon_i\} = \mathbf{X}_i\boldsymbol{\beta}.$$

The **deviation at** $\mathbf{X}_i$ is the difference between the observed response Y_i and the expected response $\mathrm{E}\{Y_i \mid \mathbf{X}_i\}$:

$$e_i = Y_i - \mathrm{E}\{Y_i \mid \mathbf{X}_i\};$$

the deviation can be **positive** (if the point lies **above** the hyperplane $Y = \mathbf{X}\boldsymbol{\beta}$) or **negative** (if it lies **below**).

How do we find **estimators** for $\boldsymbol{\beta}$? Incidentally, how do we determine if the fitted hyperplane is a **good model for the data**?

Consider the function

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \mathrm{E}\left\{Y_i \mid \mathbf{X}_i\right\}\right)^2 = \sum_{i=1}^{n} (Y_i - \mathbf{X}_i\boldsymbol{\beta})^2.$$

If $Q(\boldsymbol{\beta})$ is "small", then the sum of the **squared residuals** is "small", and so we would expect the hyperplane $Y = \mathbf{X}\boldsymbol{\beta}$ to be a good fit for the data.

The **least-square estimators** of the GLR problem is the vector $\mathbf{b} \in \mathbb{R}^p$ which minimizes the function Q with respect to $\boldsymbol{\beta} \in \mathbb{R}^p$.

We must then find critical points of $Q(\boldsymbol{\beta})$, i.e., solve $\nabla_{\boldsymbol{\beta}} Q(\mathbf{b}) = \mathbf{0}$.

3.1.1 – Matrix Notation

The LS regression function is $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$, where $\mathbf{b}$ minimizes

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - \mathbf{X}_i\boldsymbol{\beta})^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$
$$= (\mathbf{Y}^\top - \boldsymbol{\beta}^\top\mathbf{X}^\top)(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}^\top\mathbf{Y} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y} - \mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}.$$

Since $\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y}$ is a scalar, it is equal to its transpose $\mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta}$, so

$$Q(\boldsymbol{\beta}) = \mathbf{Y}^\top\mathbf{Y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{Y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}.$$

But $\mathbf{X}^\top\mathbf{X}$ is positive definite, so $Q(\boldsymbol{\beta})$ is minimized at $\nabla_{\boldsymbol{\beta}} Q(\mathbf{b}) = \mathbf{0}$.

3.1.2 – Normal Equations

The gradient vector of $Q(\boldsymbol{\beta})$ is

$$\nabla_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = -2\mathbf{X}^\top \mathbf{Y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta},$$

so the critical point $\mathbf{b}$ solves the **normal equations**

$$(\mathbf{X}^\top \mathbf{X})\mathbf{b} = \mathbf{X}^\top \mathbf{Y}.$$

The matrix $\mathbf{X}^\top \mathbf{X}$ is called the **sum of squares and cross products** (SSCP) matrix; when it is invertible, the **unique** solution of the normal equations is

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y},$$

also known as the **LS estimates** of the GLR problem.

The SSCP matrix is $p \times p$, and so is not usually too costly to invert, no matter the number of observations n.

For instance, say we have two predictors X_1, X_2 and three regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$. If we write $\mathbf{x} = (1, X_1, X_2)$, the **regression function** is

$$\mathrm{E}\{Y\} = \mathbf{x}\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

If the LS estimates are

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \big(0.5, -0.1, 2\big)^\top,$$

say, then the **estimated regression function** is

$$\hat{Y} = \mathbf{x}\mathbf{b} = 0.5 - 0.1 X_1 + 2 X_2.$$

3.1.3 – Residuals and Sums of Squares

The **fitted values** for the GLR problem are

$$\begin{pmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_n \end{pmatrix} = \hat{\mathbf{Y}} = \mathbf{Xb} = \underbrace{\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top}_{=\mathbf{H}}\mathbf{Y} = \mathbf{HY},$$

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

Theorem: $\mathbf{H}$, $\mathbf{I}_n - \mathbf{H}$ are idempotent and symmetric, and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{0}$.

Proof: we use the notation $\mathbf{M} = \mathbf{I}_n - \mathbf{H}$. We will first need to show that $\mathbf{H}^2 = \mathbf{H}$, $\mathbf{H}^\top = \mathbf{H}$, $\mathbf{M}^2 = \mathbf{M}$, and $\mathbf{M}^\top = \mathbf{M}$.

That this is the case is obvious:

$$\mathbf{H}^2 = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top = \mathbf{X}\mathbf{I}_n\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top = \mathbf{H}$$

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

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

$$\mathbf{M}^2 = (\mathbf{I}_n - \mathbf{H})^2 = \mathbf{I}_n^2 - \mathbf{I}_n\mathbf{H} - \mathbf{H}\mathbf{I}_n + \mathbf{H}^2 = \mathbf{I}_n - 2\mathbf{H} + \mathbf{H} = \mathbf{I}_n - \mathbf{H} = \mathbf{M}$$

$$\mathbf{M}^\top = (\mathbf{I}_n - \mathbf{H})^\top = \mathbf{I}_n^\top - \mathbf{H}^\top = \mathbf{I}_n - \mathbf{H} = \mathbf{M}.$$

Furthermore,

$$\mathbf{M}\mathbf{X} = (\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{X} - \mathbf{H}\mathbf{X} = \mathbf{X} - \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X} = \mathbf{X} - \mathbf{X}\mathbf{I}_n = \mathbf{0},$$

which completes the proof. ■

The i**th residual** is $e_i = Y_i - \hat{Y}_i$. Since $\mathbf{MX} = \mathbf{0}$, the **residual vector** is

$$\begin{aligned}\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} &= \mathbf{Y} - \mathbf{HY} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y} = \mathbf{MY} \\ &= \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{M}\boldsymbol{\varepsilon}.\end{aligned}$$

In other words, the residual vector is both a linear transformation of the response vector $\mathbf{Y}$ and of the random error vector $\boldsymbol{\varepsilon}$.

Just as in the SLR case (which is a special case of GLR), the residuals have a set of nice properties.

Theorem: the design matrix is orthogonal to the residual vector, i.e., $\mathbf{X}^\top \mathbf{e} = \mathbf{0}$ (the columns of $\mathbf{X}$ are orthogonal to $\mathbf{e}$).

Proof: from the normal equations, we get

$$\mathbf{X}^\top \mathbf{X}\mathbf{b} = \mathbf{X}^\top \mathbf{Y} \implies \mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\mathbf{b}) = \mathbf{0} \implies \mathbf{X}^\top(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0}.$$

But $\mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{e}$, so that $\mathbf{X}^\top \mathbf{e} = \mathbf{0}$. ■

Theorem: if the model has an intercept term β_0, we also have $\mathbf{1}_n^\top \mathbf{e} = 0$, $\overline{\mathbf{e}} = \overline{\mathbf{Y}} - \overline{\hat{\mathbf{Y}}} = 0$, and $\hat{\mathbf{Y}}^\top \mathbf{e} = 0$.

Proof: if there is an intercept term, the first column of the design matrix $\mathbf{X}$ is $\mathbf{1}_n$. Thus $\mathbf{1}_n^\top \mathbf{e}$ corresponds to the first entry of $\mathbf{X}^\top \mathbf{e} = \mathbf{0}$, which is to say, 0.

This also implies that $\overline{\mathbf{e}} = 0$. For the last part, recall that $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$. Thus, $\hat{\mathbf{Y}}^\top = \mathbf{b}^\top \mathbf{X}^\top$ and $\hat{\mathbf{Y}}^\top \mathbf{e} = \mathbf{b}^\top \mathbf{X}^\top \mathbf{e} = \mathbf{b}^\top \mathbf{0} = 0$. ■

We have already seen that SST is a quadratic form in $\mathbf{Y}$:

$$\mathrm{SST} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \mathbf{Y}^\top \left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{Y};$$

from the definition of the residuals, we see that this is also the case for SSE:

$$\begin{aligned}\mathrm{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}^\top\mathbf{e} = (\mathbf{MY})^\top\mathbf{MY} = \mathbf{Y}^\top\mathbf{M}^\top\mathbf{MY} \\ = \mathbf{Y}^\top\mathbf{M}^2\mathbf{Y} = \mathbf{Y}^\top\mathbf{MY} = \mathbf{Y}^\top\left(\mathbf{I}_n - \mathbf{H}\right)\mathbf{Y}.\end{aligned}$$

The sum of squares decomposition can then be re-written as:

$$\mathrm{SSR} = \mathrm{SST} - \mathrm{SSE}\,.$$

Thus, SSR is also a quadratic form in $\mathbf{Y}$:

$$\mathrm{SSR} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \mathbf{Y}^\top \left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n\right)\mathbf{Y} - \mathbf{Y}^\top(\mathbf{I}_n - \mathbf{H})\,\mathbf{Y}$$

$$= \mathbf{Y}^\top \left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n - \mathbf{I}_n + \mathbf{H}\right)\mathbf{Y} = \mathbf{Y}^\top \left(\mathbf{H} - \frac{1}{n}\mathbf{J}_n\right)\mathbf{Y}.$$

Theorem: $\mathrm{E}\{\mathrm{SSE}\} = (n-p)\sigma^2$, and thus $\mathrm{rank}(\mathbf{M}) = \mathrm{tr}(\mathbf{M}) = n - p$. Thus, SSE has $n - p$ degrees of freedom.

Proof: we have

$$\mathrm{SSE} = \mathbf{e}^\top\mathbf{e} = (\mathbf{M}\boldsymbol{\varepsilon})^\top\mathbf{M}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\top\mathbf{M}\boldsymbol{\varepsilon} = \sum_{i,j=1}^{n} m_{ij}\varepsilon_i\varepsilon_j = \sum_{i=1}^{n} m_{ii}\varepsilon_i^2 + \sum_{i\neq j} m_{ij}\varepsilon_i\varepsilon_j.$$

Since $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$,

$$\mathrm{E}\left\{\varepsilon_i^2\right\} = \sigma^2\left\{\varepsilon_i\right\} + \left(\mathrm{E}\left\{\varepsilon_i\right\}\right)^2 = \sigma^2 + 0 = \sigma^2, \quad i = 1, \ldots, n, \quad \text{and}$$
$$\mathrm{E}\left\{\varepsilon_i \varepsilon_j\right\} = \sigma\left\{\varepsilon_i, \varepsilon_j\right\} - \mathrm{E}\left\{\varepsilon_i\right\}\mathrm{E}\left\{\varepsilon_j\right\} = 0 - 0 = 0, \quad i \neq j.$$

Consequently,

$$\begin{aligned}
\mathrm{E}\left\{\mathrm{SSE}\right\} &= \mathrm{E}\left\{\sum_{i=1}^{n} m_{ii}\varepsilon_i^2 + \sum_{i \neq j} m_{ij}\varepsilon_i\varepsilon_j\right\} = \mathrm{E}\left\{\sum_{i=1}^{n} m_{ii}\varepsilon_i^2\right\} + \mathrm{E}\left\{\sum_{i \neq j} m_{ij}\varepsilon_i\varepsilon_j\right\} \\
&= \sum_{i=1}^{n} m_{ii}\mathrm{E}\left\{\varepsilon_i^2\right\} + \sum_{i \neq j} m_{ij}\mathrm{E}\left\{\varepsilon_i\varepsilon_j\right\} = \sigma^2 \sum_{i=1}^{n} m_{ii} = \sigma^2 \operatorname{tr}(\mathbf{M}) \\
&= \sigma^2 \operatorname{tr}(\mathbf{I}_n - \mathbf{H}) = \sigma^2[\operatorname{tr}(\mathbf{I}_n) - \operatorname{tr}(\mathbf{H})] = \sigma^2[n - \operatorname{tr}(\mathbf{H})].
\end{aligned}$$

But

$$\operatorname{tr}(\mathbf{H}) = \operatorname{tr}\Big(\underbrace{\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}}_{=A_{n\times p}} \underbrace{\mathbf{X}^\top}_{=B_{p\times n}} \Big) = \operatorname{tr}\Big(\underbrace{\mathbf{X}^\top}_{=B_{p\times n}} \underbrace{\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}}_{=A_{n\times p}} \Big) = \operatorname{tr}(\mathbf{I}_p) = p,$$

whence $\mathrm{E}\{\mathrm{SSE}\} = (n-p)\sigma^2$. ■

The **mean square error** MSE in the GLR model is

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-p},$$

which is not surprising as we have to estimate the p parameters β_k, $k = 0, \ldots, p-1$, in order to compute SSE. According to the previous theorem, MSE is an **unbiased estimator of the error variance** σ^2.

3.2 – Inference, Estimation, and Prediction

Assuming **normality** and **independence** of the random errors, the estimators $b_0, \ldots, b_{p-1}$ are then independent of SSE and

$$\frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-p).$$

This information allows us to test for the **significance of regression** using the **overall $F-$test**:

$H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$ **against** $H_1 : \beta_k \neq 0$ **for some** $k = 1, \ldots, p-1$

assuming that the GLR model holds.

Analysis of Variance

In particular, we have

$$Y_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), \quad i = 1, \ldots, n.$$

Whether H_0 holds or not, the unbiased estimator for the error variance is

$$\widehat{\sigma^2} = \text{MSE} = \frac{\text{SSE}}{n-p} \quad \left(\implies \frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-p) \right).$$

If H_0 holds, then $Y_1, \ldots, Y_n$ is an independent random sample drawn from $\mathcal{N}(\beta_0, \sigma^2)$. Our best estimate for σ^2 is thus

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{\text{SST}}{n-1} \quad \left(\implies \frac{\text{SST}}{\sigma^2} \sim \chi^2(n-1) \right).$$

Since $\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$, **Cochran's Theorem** implies that $\mathrm{SSE}, \mathrm{SSR}$ are **independent**, and that

$$\frac{\mathrm{SSR}}{\sigma^2} \sim \chi^2\big((n-1)-(n-p)\big) = \chi^2(p-1).$$

Thus, if H_0 holds, the quotient

$$F^* = \frac{\left(\frac{\mathrm{SSR}}{\sigma^2}\right) \Big/ (p-1)}{\left(\frac{\mathrm{SSE}}{\sigma^2}\right) \Big/ (n-p)} = \frac{\mathrm{SSR}\,/(p-1)}{\mathrm{SSE}\,/(n-p)} = \frac{\mathrm{MSR}}{\mathrm{MSE}} \sim F(p-1, n-p)$$

follows a Fisher F distribution with $p-1, n-p$ degrees of freedom.

The corresponding **ANOVA** table is

Source	**SS**	**df**	**MS**	$\mathbf{F}^*$
Regression	SSR	$p-1$	$\text{MSR} = \text{SSR}/(p-1)$	MSR/MSE
Error	SSE	$n-p$	$\text{MSE} = \text{SSE}/(n-p)$	
Total	SST	$n-1$		

The overall $F-$test's $\mathbf{p}-$**value** is

$$P(F(p-1, n-p) > F^*).$$

Decision Rule: at confidence level $1-\alpha$, we reject H_0 if

$$F^* > F(1-\alpha; p-1, n-p);$$

equivalently, we reject H_0 if $P(F(p-1, n-p) > F^*) < \alpha$.

Example: consider a dataset with $n = 12$ observations, a response variable Y and $p - 1 = 4$ predictors X_1, X_2, X_3, X_4. We build a GLR model

$$\begin{aligned} Y_i &= \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, 12 \\ &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \beta_4 X_{i,4} + \varepsilon_i, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{12}) \end{aligned}$$

The corresponding **ANOVA** table is

Source	**SS**	**df**	**MS**	$\mathbf{F}^*$
Regression	4957.2	4	1239.3	5.1
Error	1699.0	7	242.7	
Total	6656.2	11		

With a $p-$value $= P(F(4, 7) > 5.1) = 0.0303$, we **reject** H_0 at $\alpha = 0.05$ and conclude that the regression is **significant**.

Geometrical Interpretation

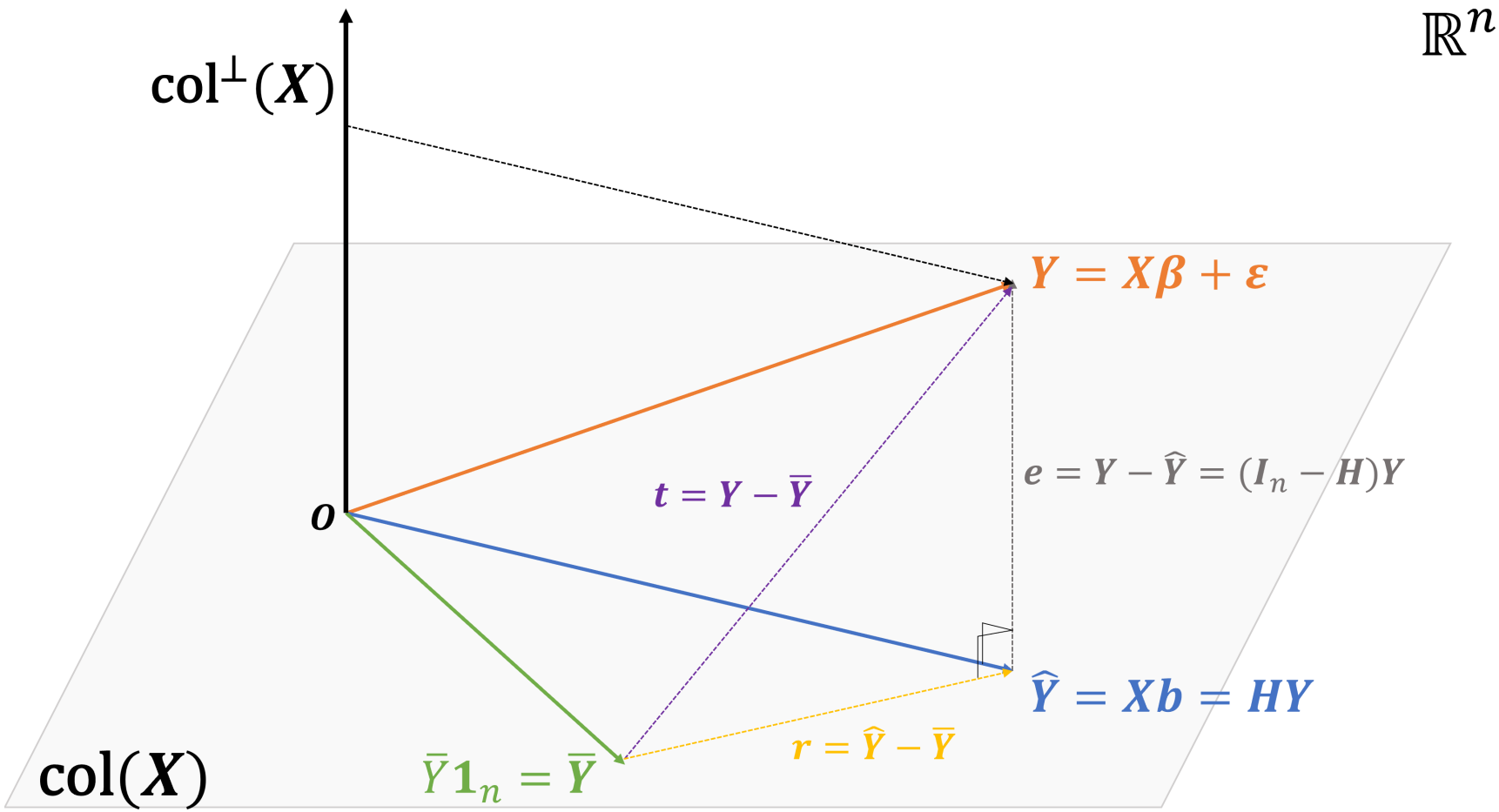
A number of GLR concepts become easier to understand when viewed through the prism of **geometry** and **vector algebra**. Let

$$\mathcal{M}(\mathbf{X}) = \text{col}(\mathbf{X}) = \{\mathbf{X}\boldsymbol{\gamma} \mid \boldsymbol{\gamma} \in \mathbb{R}^p\} < \mathbb{R}^n$$
$$\mathcal{M}^{\perp}(\mathbf{X}) = (\text{col}(\mathbf{X}))^{\perp} = \{\mathbf{v} \in \mathbb{R}^n \mid \mathbf{v} \cdot \mathbf{w} = 0, \ \forall \mathbf{w} \in \mathcal{M}(\mathbf{X})\}$$

The **vector of observations** $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ lies in $\mathbb{R}^n$, while the **fitted vector** $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{H}\mathbf{Y}$ lies in $\mathcal{M}(\mathbf{X})$ and

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$$

lies in $\mathcal{M}^{\perp}(\mathbf{X})$. The hat matrix $\mathbf{H}$ and $\mathbf{I}_n - \mathbf{H}$ are idempotent (they are the projection matrices on $\mathcal{M}(\mathbf{X})$ and $\mathcal{M}^{\perp}(\mathbf{X})$) and symmetric.


$\mathbb{R}^n$
$\mathbf{col}^{\perp}(\boldsymbol{X})$
$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$
$\boldsymbol{e} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}$
$\boldsymbol{t} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$
$\boldsymbol{o}$
$\widehat{\boldsymbol{Y}} = \boldsymbol{Xb} = \boldsymbol{HY}$
$\boldsymbol{r} = \widehat{\boldsymbol{Y}} - \overline{\boldsymbol{Y}}$
$\mathbf{col}(\boldsymbol{X})$
$\overline{Y}\mathbf{1}_n = \overline{\boldsymbol{Y}}$

The LS estimator $\mathbf{b}$ is such that $\mathbf{Xb}$ is the closest vector to $\mathbf{Y}$ in $\mathcal{M}(\mathbf{X})$:

$$\mathbf{b} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^p} \left\{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\gamma}\|_2^2\right\} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^p} \left\{\|\mathbf{e}\|_2^2\right\} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^p} \{\mathrm{SSE}\}.$$

If the GLR model has a constant term β_0, the mean vector $\overline{\mathbf{Y}} = \overline{Y}\mathbf{1}_n$ lies in $\mathcal{M}(\mathbf{X})$; indeed, for $\boldsymbol{\gamma}^* = (\overline{Y}, 0, \ldots, 0)^\top$, we have $\overline{\mathbf{Y}} = \mathbf{X}\boldsymbol{\gamma}^*$.

The triangle $\Delta\mathbf{Y}\hat{\mathbf{Y}}\overline{\mathbf{Y}}$ is thus a **right angle triangle**, with

$$\mathbf{t} = \mathbf{Y} - \overline{\mathbf{Y}} = (\mathbf{Y} - \hat{\mathbf{Y}}) + (\hat{\mathbf{Y}} - \overline{\mathbf{Y}}) = \mathbf{e} + \mathbf{r};$$

Pythagoras' Theorem then gives us

$$\|\mathbf{t}\|_2^2 = \mathrm{SST} = \mathrm{SSE} + \mathrm{SSR} = \|\mathbf{e}\|_2^2 + \|\mathbf{r}\|_2^2.$$

3.2.1 – Inference on Model Parameters

As was the case with the SLR model parameters, if $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, then

$$\mathbf{Y} \sim \mathcal{N}(\mathrm{E}\{\mathbf{Y}\}, \sigma^2\{\mathbf{Y}\}) = \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n).$$

If A is any compatible matrix, then

$$A\mathbf{Y} \sim \mathcal{N}(A\mathrm{E}\{\mathbf{Y}\}, A\sigma^2\{\mathbf{Y}\}A^\top) = \mathcal{N}(A\mathbf{X}\boldsymbol{\beta}, \sigma^2 AA^\top).$$

From the normal equations, the LS estimates for the GLR model are given by a **linear transformation** of the response vector $\mathbf{Y}$:

$$\mathbf{b} = \underbrace{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}_{p \times n}\mathbf{Y} = A\mathbf{Y}.$$

In particular,

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

so that $\mathbf{b}$ provides **unbiased estimators** of $\boldsymbol{\beta}$.

Furthemore,

$$\begin{aligned}\sigma^2\{\mathbf{b}\} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2\{\mathbf{Y}\}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]^\top \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\sigma^2\mathbf{I}_n\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right]^\top = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.\end{aligned}$$

Thus,

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

The **estimated variance-covariance matrix** for the estimators $\mathbf{b}$ is thus

$$\mathrm{s}^2\{\mathbf{b}\} = \mathrm{MSE} \cdot (\mathbf{X}^\top \mathbf{X})^{-1}, \quad \text{and} \quad \mathrm{s}\{\mathbf{b}\} = \sqrt{\mathrm{MSE}} \sqrt{\mathrm{diag}\left[(\mathbf{X}^\top \mathbf{X})^{-1}\right]}.$$

For each $k = 0, \ldots, p-1$, the **studentization** of b_k is

$$T_k = \frac{b_k - \beta_k}{\sqrt{\mathrm{MSE}}\sqrt{(\mathbf{X}^\top \mathbf{X})^{-1}_{k,k}}} = \underbrace{\frac{b_k - \beta_k}{\sigma\sqrt{(\mathbf{X}^\top \mathbf{X})^{-1}_{k,k}}}}_{=Z} \Bigg/ \sqrt{\underbrace{\frac{\mathrm{SSE}}{\sigma^2}}_{=U} \Bigg/ \underbrace{(n-p)}_{=\nu}} \sim t(n-p),$$

where $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}_{k,k}$ represents the $k+1$ entry in $\mathrm{diag}\left[(\mathbf{X}^\top \mathbf{X})^{-1}\right]$.

For a specific $k \in \{0, \ldots, p-1\}$, the $100(1-\alpha)\%$ C.I. for β_k is

$$\text{C.I.}(\beta_k; 0.95) \equiv b_k \pm t\left(1 - \tfrac{\alpha}{2}; n-p\right) \cdot \text{s}\{b_k\}.$$

The corresponding hypothesis tests for

$$H_0 : \beta_k = \beta_k^* \qquad \text{against} \quad H_1 : \begin{cases} \beta_k < \beta_k^* & \text{left-tailed test} \\ \beta_k > \beta_k^* & \text{right-tailed test} \\ \beta_k \neq \beta_k^* & \text{two-tailed test} \end{cases}$$

Under H_0, the computed test statistic

$$T_k = \frac{b_k - \beta_k^*}{\text{s}\{b_k\}} \sim t(n-p).$$

The **critical region** for the test depends on the **confidence level** $1-\alpha$ and on the **type** of the alternative hypothesis H_1.

Let t^* be the observed value of T_k. **We reject H_0 if t^* in the critical region**.

Alternative Hypothesis	Rejection Region
$H_1 : \beta_k < \beta_k^*$	$t^* < -t(1-\alpha; n-p)$
$H_1 : \beta_k > \beta_k^*$	$t^* > t(1-\alpha; n-p)$
$H_1 : \beta_k \neq \beta_k^*$	$\lvert t^* \rvert > t(1-\alpha/2; n-p)$

Example: consider the situation with $n = 12$ observations and $p - 1 = 4$ predictors as described previously.

We build the GLR model $\hat{\mathbf{Y}} = \mathbf{Xb}$ and obtain the following results:

Predictor	**Estimate**	**SE**	t
Intercept	-102.71	207.86	-0.49
X_1	0.61	0.37	1.64
X_2	8.92	5.3	1.68
X_3	1.44	2.39	0.60
X_4	0.01	0.77	0.02

Recall that $n - p = 7$; the 95% C.I. for β_2 is thus

$$\text{C.I.}(\beta_2; 0.95) \equiv 8.92 \pm t(0.975; 7) \cdot 5.3 = 8.92 \pm 2.365 \cdot 5.3 = [-3.6, 21.5].$$

We could also test for $H_0 : \beta_3 = 2$ against $H_1 : \beta_3 \neq 2$, say: under H_0,

$$T_3^* = \frac{b_3 - 2}{\text{s}\{b_3\}} \sim t(7).$$

The observed statistic is

$$t^* = \frac{1.44 - 2}{2.39} = -0.23;$$

we would reject H_0 at confidence level $1 - \alpha = 0.95$ if

$$|t^*| > t(0.975; 7) = 2.365;$$

as $-0.23 \not> 2.365$, the evidence is not strong enough to conclude that $\beta_3 \neq 2$.

While we can build a C.I. for β_2 and test a hypothesis about β_3, each at the $1 - \alpha = 0.95$ confidence level, we cannot do so **jointly**.

3.2.2 – Inference on the Mean Response

We can also conduct inferential analysis for the **expected response** at

$$\mathbf{X}^* = (1, X_1^*, \ldots, X_{p-1}^*) \quad \text{in the model's } \textbf{scope}.$$

In the GLR model, we assume that

$$\mathrm{E}\{Y^*\} = \mathbf{X}^*\boldsymbol{\beta} = \beta_0 + \beta_1 X_1^* + \cdots + \beta_{p-1} X_{p-1}^*.$$

The **estimated mean response** at $\mathbf{X}^*$ is

$$\hat{Y}^* = \mathbf{X}^*\boldsymbol{b} \ = b_0 + b_1 X_1^* + \cdots + b_{p-1} X_{p-1}^*.$$

The predictor values are **fixed**, thus $\hat{Y}^*$ is normally distributed with

$$\mathrm{E}\{\hat{Y}^*\} = \mathrm{E}\left\{\mathbf{X}^*\mathbf{b}\right\} = \mathbf{X}^*\mathrm{E}\left\{\mathbf{b}\right\} = \mathbf{X}^*\boldsymbol{\beta},$$

so that $\hat{Y}^*$ is an **unbiased estimator** of Y^*.

Furthermore,

$$\sigma^2\{\hat{Y}^*\} = \mathbf{X}^*\sigma^2\left\{\mathbf{b}\right\}\left(\mathbf{X}^*\right)^\top = \sigma^2\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}\left(\mathbf{X}^*\right)^\top,$$

so that

$$\mathrm{s}^2\{\hat{Y}^*\} = \mathrm{MSE}\cdot\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}\left(\mathbf{X}^*\right)^\top = \mathbf{X}^*\mathrm{s}^2\left\{\mathbf{b}\right\}\left(\mathbf{X}^*\right)^\top.$$

The **estimated standard error** is thus

$$\mathrm{s}\{\hat{Y}^*\} = \sqrt{\mathbf{X}^* \mathrm{s}^2\{\mathbf{b}\} (\mathbf{X}^*)^\top}.$$

Since

$$\hat{Y}^* = \mathbf{X}^*\mathbf{b} = \mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$$

is a **linear transformation** of $\mathbf{Y}$, and since

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

then

$$\hat{Y}^* \sim \mathcal{N}\left(\mathrm{E}\{\hat{Y}^*\}, \sigma^2\{\hat{Y}^*\}\right) = \mathcal{N}\left(\mathbf{X}^*\boldsymbol{\beta}, \sigma^2\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top\right).$$

Thus

$$Z = \frac{\hat{Y}^* - \mathrm{E}\{\hat{Y}^*\}}{\sigma\{\hat{Y}^*\}} = \frac{\hat{Y}^* - \mathbf{X}^*\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{X}^*\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{X}^*\right)^\top}} \sim \mathcal{N}(0,1).$$

The **studentization** of $\hat{Y}^*$ is then

$$T = \underbrace{\frac{\hat{Y}^* - \mathbf{X}^*\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{X}^*\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{X}^*\right)^\top}}}_{=Z} \Bigg/ \sqrt{\underbrace{\frac{\mathrm{SSE}}{\sigma^2}}_{=U} \Bigg/ \underbrace{(n-p)}_{=\nu}}$$

$$= \frac{\hat{Y}^* - \mathbf{X}^*\boldsymbol{\beta}}{\sqrt{\mathrm{MSE}}\sqrt{\mathbf{X}^*\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\left(\mathbf{X}^*\right)^\top}} \sim t(n-p).$$

For a specific predictor level $\mathbf{X}^*$, the $100(1-\alpha)\%$ C.I. for $\mathrm{E}\{Y^*\}$ is

$$\text{C.I.}(\mathrm{E}\{Y^*\}; 0.95) \equiv \hat{Y}^* \pm t\left(1-\tfrac{\alpha}{2}; n-p\right) \cdot \mathrm{s}\{\hat{Y}^*\}.$$

The corresponding hypothesis tests for

$$H_0 : \mathrm{E}\{Y^*\} = \gamma \qquad \text{against} \quad H_1 : \begin{cases} \mathrm{E}\{Y^*\} < \gamma & \text{left-tailed test} \\ \mathrm{E}\{Y^*\} > \gamma & \text{right-tailed test} \\ \mathrm{E}\{Y^*\} \neq \gamma & \text{two-tailed test} \end{cases}$$

Under H_0, the computed test statistic

$$T = \frac{\hat{Y}^* - \gamma}{\mathrm{s}\{\hat{Y}^*\}} \sim t(n-p).$$

The **critical region** for the test depends on the **confidence level** $1-\alpha$ and on the **type** of the alternative hypothesis H_1.

Let t^* be the observed value of T. **We reject H_0 if t^* in the critical region**.

Alternative Hypothesis	Rejection Region
$H_1 : \mathrm{E}\{Y^*\} < \gamma$	$t^* < -t(1-\alpha; n-p)$
$H_1 : \mathrm{E}\{Y^*\} > \gamma$	$t^* > t(1-\alpha; n-p)$
$H_1 : \mathrm{E}\{Y^*\} \neq \gamma$	$\lvert t^* \rvert > t(1-\alpha/2; n-p)$

Example: consider the situation with $n = 12$ observations and $p - 1 = 4$ predictors as described previously. We would like to predict the expected response at

$$\mathbf{X}^* = (1, 11.10, 20.74, 6.61, 182.38), \quad \text{in the model's scope.}$$

Thus

$$\begin{aligned}\hat{Y}^* &= \mathbf{X}^*\mathbf{b} \\ &= -102.71 + 0.61(11.10) + 8.92(20.74) + 1.44(6.61) + 0.01(182.38) \\ &= 100.40.\end{aligned}$$

Recall that $\mathrm{MSE} = 242.71$. Using the data, we computed

$$\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}\left(\mathbf{X}^*\right)^\top = 1.42,$$

so that

$$\mathrm{s}\{\hat{Y}^*\} = \sqrt{242.71}\sqrt{1.42} = 22.12.$$

Since $n - p = 7$; the 95% C.I. for $\mathrm{E}\{Y^*\}$ is

$$\begin{aligned}\mathrm{C.I.}(\mathrm{E}\{Y^*\}; 0.95) &\equiv 100.40 \pm t(0.975; 7) \cdot 22.12 \\ &= 100.40 \pm 2.365 \cdot 22.12 = [48.09, 152.71].\end{aligned}$$

We could also test for $H_0 : \mathrm{E}\{Y^*\} = 150$ against $H_1 : \mathrm{E}\{Y^*\} < 150$, say: under H_0,

$$T^* = \frac{\hat{Y}^* - 150}{\mathrm{s}\{\hat{Y}^*\}} \sim t(7).$$

The observed statistic is

$$t^* = \frac{100.40 - 150}{22.12} = -2.24.$$

We would reject H_0 at confidence level $1 - \alpha = 0.95$ if

$$t^* < -t(0.95; 7) = -1.89;$$

as $-2.24 < -1.89$, the evidence is strong enough to **reject**

$$H_0 : \mathrm{E}\{Y^*\} = 150 \quad \text{in favour of} \quad H_1 : \mathrm{E}\{Y^*\} < 150.$$

Note, however, that the two-sided 95% C.I. for $\mathrm{E}\{Y^*\}$ contains 150, so we **cannot reject**

$$H_0 : \mathrm{E}\{Y^*\} = 150 \quad \text{in favour of} \quad H_1 : \mathrm{E}\{Y^*\} \neq 150$$

at confidence level $1 - \alpha = 95\%$. As before, we cannot conduct **joint inferences** about various predictor levels $\mathbf{X}^*$ without modifications.

3.2.3 – Intervalles de prédiction

Let Y_p^* represent a **(future) response** at $\mathbf{X}^*$, so that

$$Y_p^* = \mathbf{X}^*\boldsymbol{\beta} + \varepsilon_p \quad \text{for some } \varepsilon_p.$$

If the average error is 0, the best prediction for Y_p^* is still the **fitted response at $\mathbf{X}^*$**:

$$\hat{Y}_p^* = \mathbf{X}^*\mathbf{b}.$$

The **prediction error** at $\mathbf{X}^*$ is thus

$$\text{pred}^* = Y_p^* - \hat{Y}_p^* = \mathbf{X}^*\boldsymbol{\beta} + \varepsilon_p - \mathbf{X}^*\mathbf{b}.$$

In the GLR model, the error ε_p and the estimators $\mathbf{b}$ are **normally distributed**. Consequently, so is the prediction error pred^*. Note that

$$\text{E}\left\{\text{pred}^*\right\} = \underbrace{\text{E}\left\{\mathbf{X}^*\boldsymbol{\beta} + \varepsilon_p^*\right\}}_{=\mathbf{X}^*\boldsymbol{\beta}} - \underbrace{\text{E}\left\{\mathbf{X}^*\mathbf{b}\right\}}_{=\mathbf{X}^*\boldsymbol{\beta}} = 0.$$

Because the residuals are uncorrelated, we also have

$$\begin{aligned}\sigma^2\left\{\text{pred}^*\right\} &= \sigma^2\left\{Y_p^*\right\} + \sigma^2\left\{\hat{Y}_p^*\right\} \\ &= \sigma^2 + \sigma^2\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top = \sigma^2\left[1 + \mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top\right].\end{aligned}$$

Thus

$$\text{pred}^* \sim \mathcal{N}\left(0, \sigma^2\left[1 + \mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top\right]\right).$$

The estimated standard error is thus

$$\mathrm{s}\left\{\mathrm{pred}^*\right\} = \sqrt{\mathrm{MSE}}\sqrt{1 + \mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top}.$$

As before, we can show that

$$T_p^* = \frac{\mathrm{pred}^* - 0}{\mathrm{s}\{\mathrm{pred}^*\}} \sim t(n-p), \quad \text{and so}$$

$$\mathrm{P.I.}(Y_p^*; 1-\alpha) \equiv \mathbf{X}^*\mathbf{b} \pm t(1-\tfrac{\alpha}{2}; n-p) \cdot \mathrm{s}\{\mathrm{pred}^*\}.$$

Note that $\mathrm{s}\{\hat{Y}^*\} < \mathrm{s}\{\mathrm{pred}^*\}$ so that the $\mathrm{C.I.}$ for the mean response is always **contained** in the $\mathrm{P.I.}$ for new responses.

Example: consider the situation with $n = 12$ observations and $p - 1 = 4$ predictors as described previously. We would like to predict the new responses at

$$\mathbf{X}^* = (1, 11.10, 20.74, 6.61, 182.38), \quad \text{in the model's } \textbf{scope}.$$

We have already seen that $\hat{Y}^* = \mathbf{X}^*\mathbf{b} = 100.40$. Recall that $\text{MSE} = 242.71$ and

$$\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top = 1.42,$$

so that

$$\text{s}\{\text{pred}^*\} = \sqrt{242.71}\sqrt{1 + 1.42} = 37.70.$$

Since $n - p = 7$, the 95% P.I. for Y^* is

$$\begin{aligned}\text{P.I.}(Y^*; 0.95) &\equiv 100.40 \pm t(0.975; 7) \cdot 37.70 \\ &= 100.40 \pm 2.365 \cdot 37.70 = [11.24, 189.56].\end{aligned}$$

3.2.4 – Joint Estimation and Prediction

At a **family confidence level** of $1 - \alpha$:

- the **Bonferroni** procedure can be used to jointly estimate g model parameters $\beta_{k\ell}$, g mean responses $\mathrm{E}\{Y_\ell^*\}$, or g new responses Y_ℓ^*, for $\ell = 1, \ldots, g$;
- the **Working-Hostelling** procedure can be used to jointly estimate g mean responses $\mathrm{E}\{Y_\ell^*\}$, for $\ell = 1, \ldots, g$;
- the **Scheffé** procedure can be used to jointly predict g new responses Y_ℓ^*, for $\ell = 1, \ldots, g$.

The process is identical to the SLR approach; depending on the task at hand, we pick the appropriate procedure that yields the **smallest interval**.

The sole difference lies in the composition of the **factors** that accompany the estimated standard errors in the construction of the **joint** confidence/prediction intervals at **family confidence level** $1-\alpha$:

- $t(1-\frac{\alpha/g}{2}; n-p)$ for the Bonferroni procedure;
- $\sqrt{pF(1-\alpha; p, n-p)}$ for the Working-Hotelling procedure, and
- $\sqrt{gF(1-\alpha; g, n-p)}$ for the Scheffé procedure.

Example: we can provide joint confidence intervals for the **model parameters** in the preceding example at family confidence level $1-\alpha = 0.95$, using $n - p = 7$ and $g = 5$.

The **Bonferroni** factor is $t\left(1 - \frac{0.05/5}{2}; 7\right) = t(0.995; 7) = 3.50$; the joint confidence intervals are

$$\text{C.I.}_{\text{B}}(\beta_k; 0.95) \equiv b_k \pm 3.50 \cdot \text{s}\{b_k\}.$$

Parameter	b_k	$\text{C.I.}_{\text{B}}(\beta_k; 0.95)$
β_0	-102.71	$[-830.22, 624.80]$
β_1	0.61	$[-0.685, 1.905]$
β_2	8.92	$[-9.63, 27.47]$
β_3	1.44	$[-6.925, 9.805]$
β_4	0.01	$[-2.685, 2.705]$

Individually, **none of the parameters** are significant at the family confidence level $1-\alpha = 0.95$ (all the confidence intervals contain 0), but the regression **as a whole** is significant (see overall F-test example).

Similarly, the **Working-Hotelling** joint confidence intervals for the estimated mean $\mathrm{E}\{Y_\ell^*\}$ at a variety of predictor levels $\mathbf{X}_\ell^*$, $\ell = 1, \ldots, g$ (family confidence level $1-\alpha = 0.95$) are

$$\begin{aligned}\mathrm{C.I.}_{\mathsf{WH}}(\mathrm{E}\{Y_\ell^*\}; 0.95) &\equiv \hat{Y}_\ell^* \pm \sqrt{5F(0.95;5,7)} \cdot \mathrm{s}\{\hat{Y}_\ell^*\} \\ &= \mathbf{X}_\ell^*\mathbf{b} \pm 4.46 \sqrt{\underbrace{242.71}_{=\mathrm{MSE}}} \sqrt{\mathbf{X}_\ell^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}_\ell^*)^\top}\end{aligned}$$

3.3 – Power of a Test

When we do hypothesis testing, we can make two types of errors.

Type I Error: reject a valid H_0

Type II Error: fail to reject H_0 when H_1 is valid

In fact there are 4 types of errors, but that is not important here.

The **level of significance** α is used to control the risk of making an error of type **I**; type **II** errors are harder to control, in general.

Suppose we are testing ($2-$sided test) for

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

Let α be the probability of making an error of type I.

The **power function**

$$K(\theta') = P(\text{reject } H_0 \text{ if } \theta = \theta'$$

is such that $K(\theta_0) = \alpha$.

If $\theta \neq \theta_0$, $t^* = \frac{\hat{\theta}-\theta_0}{\mathrm{s}\{\hat{\theta}\}} \sim t(\nu)$ with **non-centrality parameter**

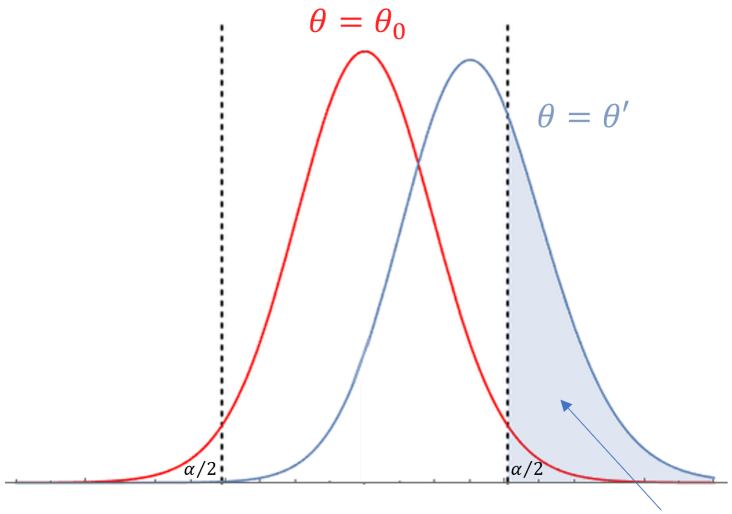
$$\delta = \frac{|\theta - \theta_0|}{\sigma\{\hat{\theta}\}} \approx \frac{|\theta - \theta_0|}{\mathrm{s}\{\hat{\theta}\}},$$

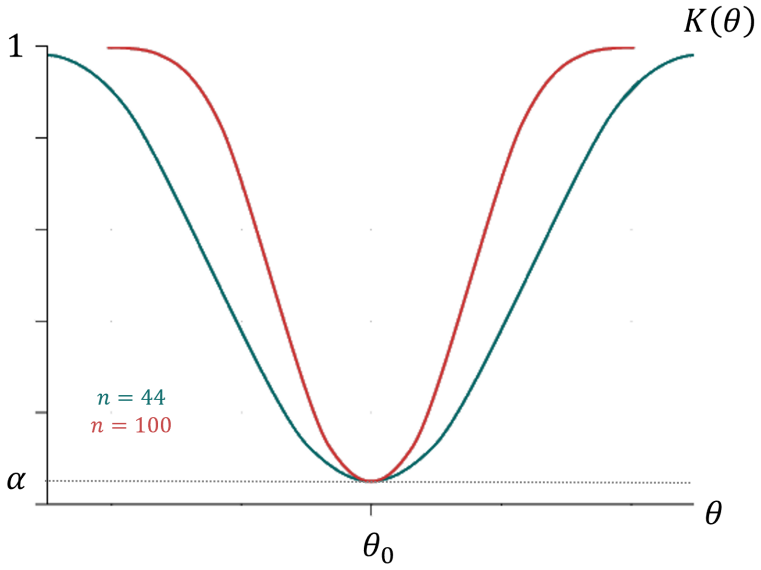
where θ is the true value and θ_0 is the value under H_0.

The **power of the test** is the probability of rejecting H_0 if $\theta = \theta'$):

$$K(\theta') = P(|t^*| > t(1 - \alpha/2; \nu); \delta).$$

To control the power, we can either increase n or decrease S_{xx}.





$$K(\theta') = P(\text{reject } H_0 \text{ if } \theta = \theta')$$

Example: we collect four responses for each of the following 5 predictor levels $(X = 5, 10, 15, 20, 25)$. In a preliminary sample, we collected data to approximate σ^2 *via* $\text{MSE} = 1532.1$. We build the model $\text{E}\{Y\} = b_0 + b_1 X$. If $\beta_1 = 6$, is it likely that we would conclude that the regression is significant?

Solution: we have

$$\sum_{i=1}^{20} X_i = 300 \quad \text{and} \quad \sum_{i=1}^{20} X_i^2 = 5500 \implies S_{xx} = \sum_{i=1}^{n} X_i^2 - 20\overline{X}^2 = 1000.$$

Thus

$$\delta \approx \frac{|\beta_1 - 0|}{\text{s}\{b_1\}} = \frac{|6 - 0|}{\sqrt{1532.1}/\sqrt{1000}} = 4.85.$$

From Table B.5 with $\nu = n - 2 = 18$ degrees of freedom and $\alpha = 0.05$, we see that the power of the significance test lies in $(0.97, 1.00)$ if $\beta_1 = 6$.

3.4 – Coefficients of Determination

The **coefficient of multiple determination** of a GLR model is

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}},$$

the proportion of the variation in Y which is explained by the regression.

If the GLR model incorporates an intercept term ($\beta_0 \neq 0$), then

$$R^2 = r^2_{Y\hat{Y}} = \frac{(s_{Y\hat{Y}})^2}{s_Y s_{\hat{Y}}};$$

this is not the case without an intercept term.

When the number of parameters p increases, so does R^2; however, the degrees of freedom, $n-p$ decrease. This typically means that the estimates are less precise. We can adjust R^2 to take into account this loss.

The **adjusted coefficient of multiple determination** of a GLR model is

$$R_a^2 = 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)} = 1 - \frac{n-1}{n-p} \cdot \frac{\text{SSE}}{\text{SST}} \quad \text{(which could be } < 0\text{)}.$$

Example: in the case we have been carrying around for a while, we had

$$\text{SST} = 6656.2, \quad \text{SSE} = 1699.0, \quad n-p = 7, \quad n-1 = 11,$$

so that

$$R^2 = 1 - \frac{1699.0}{6656.2} = 0.745 \quad \text{and} \quad R_a^2 = 1 - \frac{11}{7} \cdot \frac{1699.0}{6656.2} = 0.599.$$

3.5 – Diagnostics and Remedial Measures

We have seen that there are **four** GLR assumptions:

- **linearity** – $\mathrm{E}\{Y \mid \mathbf{X} = \mathbf{x}\} = \mathbf{x}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$
- **variance constancy (homoscedasticity)** – $\sigma^2\{\varepsilon_i\} = \sigma^2$, $i = 1, \ldots, n$
- **independence** – $\varepsilon_1, \ldots, \varepsilon_n$ are **independent** (**uncorrelated** is sufficient)
- **normality** – $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$

We have combined these assumptions in the simpler vector form

$$Y \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n).$$

In theory, these assumptions must be met before we can trust the GLR model (the model may prove useful even if they are not met, but that must be established **on a case-by-case basis**).

Recall that we have the following results on the **residuals**:

1. $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$, or $e_i = Y_i - \hat{Y}_i$, for $i = 1, \ldots, n$

2. if $\beta_0 \neq 0$, $\overline{\mathbf{e}} = 0$

3. $\sigma^2\{\mathbf{e}\} = \sigma^2(\mathbf{I}_n - \mathbf{H})$, so that $\sigma^2\{e_i\} = \sigma^2(1 - h_{ii})$, for $i = 1, \ldots, n$, and $\sigma\{e_i, e_j\} = \sigma\{e_j, e_i\} = -h_{ij}\sigma^2$ for $i \neq j = 1, \ldots, n$.

The **standard error** is $\mathrm{s}^2\{e_i\} = \mathrm{MSE}(1 - h_{ii})$ and the **internal studentization** is $r_i = \frac{e_i - \overline{e}}{\mathrm{s}\{e_i\}} \sim t(n-p)$, for $i = 1, \ldots, n$.

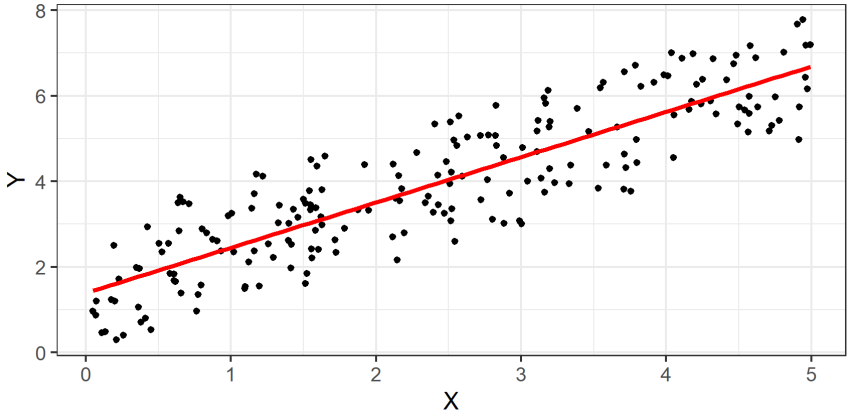
3.5.1 – Linearity

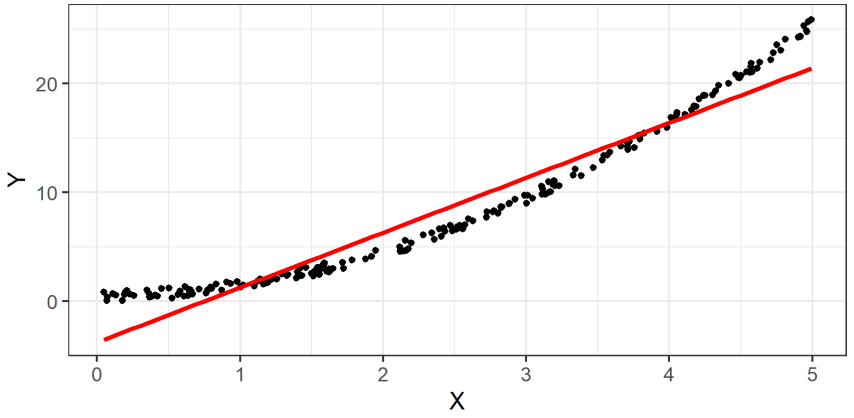
We plot the residuals e_i against the prediction $\hat{Y}_i$: if the linearity assumption is warranted, the points should appear **randomly scattered about** 0.

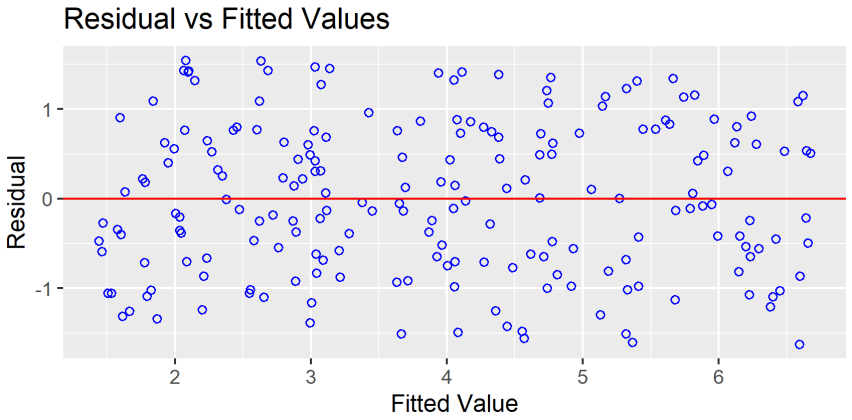
The **absence** of a trend suggests that the relationship between $X_1, \ldots, X_p$ and Y is indeed linear, the **presence** of a trend provides evidence against the linerity assumption (see next page).

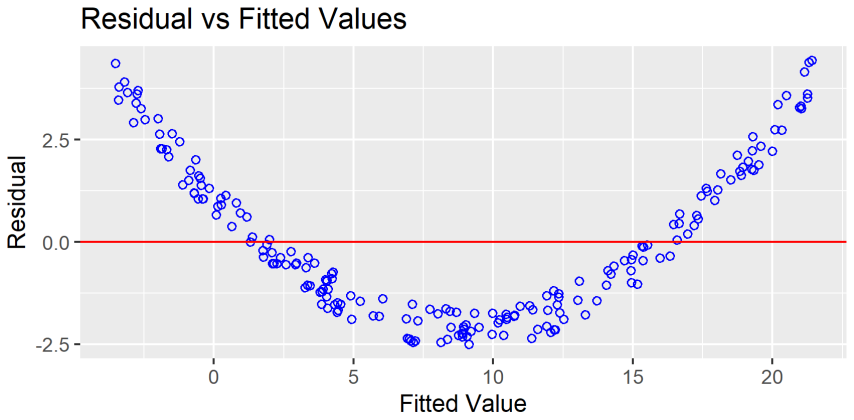
There are also formal tests, such as the test for **lack of fit**:

$$\begin{cases} H_0 : \mathrm{E}\{Y \mid \mathbf{X} = \mathbf{x}\} = \beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_{p-1}\mathbf{x}_{p-1} \\ H_1 : H_0 \text{ is false} \end{cases}$$


Y
X
0
1
2
3
4
5
6
8


Y
X
0
1
2
3
4
5
10
20


Residual vs Fitted Values
Residual
Fitted Value
-1
0
1
2
3
4
5
6


Residual vs Fitted Values
Residual
Fitted Value
-2.5
0.0
2.5
0
5
10
15
20

Let $\mathbf{W}^1 = (X_1^1, \ldots, X_{p-1}^1), \cdots, \mathbf{W}^c = (X_1^c, \ldots, X_{p-1}^c)$, be the c **distinct** predictor levels. The jth level has n_j observations $Y_{i,j}$. Assume that $\mathrm{E}\{Y\}$ has a **functional dependency on** $X_1, \ldots, X_{p-1}$, and that the residuals are **independent** and follow a **normal distribution** $\mathcal{N}(0, \sigma^2)$, and that **at least one** of the $p-1$ predictor levels X_k has **replicates**.

Let the **average observation** over the jth level be denoted by $\overline{Y}_j$, and write $\mathrm{SST}_j = \sum_i^{n_j} \left(Y_{ij} - \overline{Y}_j\right)^2$. The corresponding ANOVA table is

source	**SS**	**df**	**MS**	F^*
Regression	SSR	$p-1$	$\mathrm{SSR}/(p-1)$	$\mathrm{MSLF}/\mathrm{MSPE}$
Error	SSE	$n-p$	$\mathrm{SSE}/(n-p)$	
Lack of fit	SSLF	$c-p$	$\mathrm{SSLF}/(c-p)$	
Pure Error	SSPE	$n-c$	$\mathrm{SSPE}/(n-c)$	
Total	SST	$n-1$		

Recall that $\text{SST} = \text{SSE} + \text{SSR}$. We partition $\text{SSE} = \text{SSPE} + \text{SSLF}$, where $\text{SSPE} = \sum_{j=1}^{c} \text{SST}_j$ so that $\frac{\text{SSPE}}{\sigma^2} \sim \chi^2\left(\sum_{j=1}^{c}(n_j - 1)\right) = \chi^2(n-c)$.

Thus, according to **Cochran's Theorem**, when H_0 holds, $\frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-p)$, $\frac{\text{SSLF}}{\sigma^2} \sim \chi^2(c-p)$, and

$$F^* = \frac{\left(\frac{\text{SSLF}}{\sigma^2}\right) \Big/ (c-p)}{\left(\frac{\text{SSPE}}{\sigma^2}\right) \Big/ (n-c)} \sim F(c-p, n-c).$$

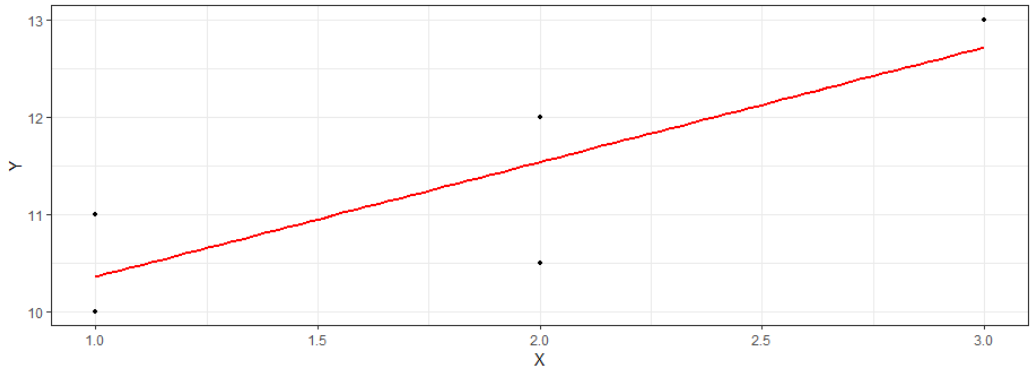
Decision Rule: If $F^* > F(1-\alpha; c-p, n-c)$, reject H_0 at a significance level of α.

Example: consider a dataset with the following (X, Y) observations

$$(1, 10), (1, 11), (2, 10.5), (2, 12), (3, 13).$$

Is the linear model $\mathrm{E}\{Y\} = \beta_0 + \beta_1 X$ warranted?

Solution: we have $n = 5$, $p = 2$, and $c = 3$. The OLS framework yields $Y = 9.18 + 1.18X$, and the scatterplot is shown below.



Visually, it does seem that the line would be a good model, but it is difficult to say with certainty since there are so few points in the chart. We use the formal test for lack of fitness: we have

$$\text{SST} = S_{yy} = 5.8, \quad \text{SSR} = b_1^2 S_{xx} = 3.8829, \quad \text{SSE} = \text{SST} - \text{SSR} = 1.91071,$$

$$\text{SSPE} = \text{SST}_1 + \text{SST}_2 + \text{SST}_3 = 0.5 + 1.125 + 0 = 1.625,$$

$$\text{SSLF} = \text{SSE} - \text{SSPE} = 1.91071 - 1.625 = 0.28751,$$

$$\text{MSLF} = \frac{\text{SSLF}}{c-p} = \frac{0.28571}{3-2} = 0.28571, \quad \text{MSPE} = \frac{\text{SSPE}}{n-c} = \frac{1.625}{5-3} = 0.8125,$$

so that

$$F^* = \frac{\text{MSLF}}{\text{MSPE}} = \frac{0.28571}{0.8125} = 0.3516.$$

Since the critical value of the $F(3-2, 5-3) = F(1,2)$ distribution at $\alpha = 0.05$ is 18.5, we **do not reject** the hypothesis of linearity.

3.5.2 – Constant Variance

We can use residual plots to determine whether the condition of homoscedasticity is met or not. But there are **formal tests** as well, such as the **Breusch-Pagan test** (requires normality of the residuals) or the **Brown-Forsythe** test (robust against departures from normality).

Let us take a look at the latter. Select a threshold $a \in \mathbb{R}$ and **partition** the residuals into 2 groups:

$$\text{Group 0: } \hat{Y} \leq a \text{ (the } e_{i,0}\text{'s)} \quad \text{vs.} \quad \text{Group 1: } \hat{Y} > a \text{ (the } e_{i,1}\text{'s).}$$

We pick a so that $|\text{Group 0}| = n_0 \approx n_1 = |\text{Group 1}|$. Let $\tilde{e}_j$ be the **median residual of group** j and let $d_{ij} = |e_{ij} - \tilde{e}_j|$ be the **absolute deviation of the** i**th residual in group** j **from** $\tilde{e}_j$, for $j = 0, 1$.

We use this framework rather than using the **mean** and the **square deviation** because of sensitivity to outliers (it is this choice that makes the test robust against departures from the normality assumption).

Set $\overline{d}_j = \dfrac{1}{n_j}\displaystyle\sum_i^{n_j} d_{ij}$, $j = 0, 1$. In order to test for

$$\begin{cases} H_0 : \overline{d}_0 = \overline{d}_1 & \text{(the variance is constant)} \\ H_1 : \overline{d}_0 \neq \overline{d}_1 & \text{(the variance is \textbf{not} constant)} \end{cases}$$

we compute the test statistic

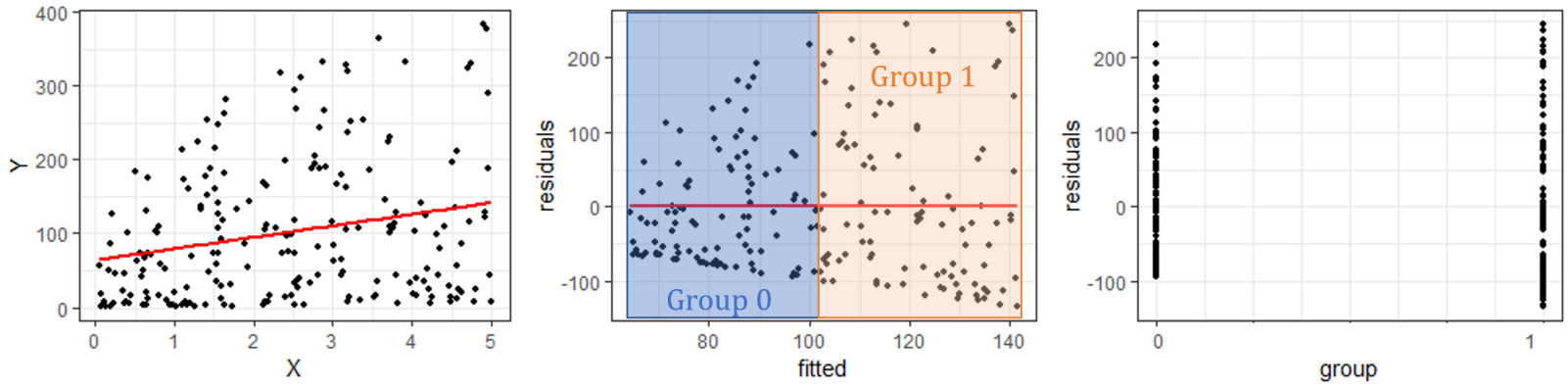
$$t^*_{\text{BF}} = \frac{\overline{d}_0 - \overline{d}_1}{s_p\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}},$$

where

$$s_p^2 = \frac{1}{n-2}\left[\sum_{i=1}^{n_0}(d_{i,0}-\overline{d}_0)^2 + \sum_{i=1}^{n_1}(d_{i,1}-\overline{d}_1)^2\right] = \frac{(n_0-1)s_0^2+(n_1-1)s_1^2}{n_0+n_1-2}$$

is the **pooled variance**. When H_0 holds, $t^*_{\text{BF}} \sim t(n_0+n_1-2) = t(n-2)$.

Decision Rule: If $|t^*_{\text{BF}}| > t(1-\alpha/2; n-2)$, reject H_0 at significance level α.



Example: in the charts on the previous slide, the median fitted value is $a = 101.5096$. Visually, the constant variance assumption does not seem to be met.

We divide the datasets into two groups, based on whether the fitted value falls below a (Group 0, in blue) or not (Group 1, in orange); there are $n_0 = n_1 = 100$ observations in each group.

The group median residuals are $\tilde{e}_0 = -15.6$, $\tilde{e}_1 = -22.9$. The mean and variance of the absolute deviations of the residuals to the median in each group are $\bar{d}_0 = 59.1$, ${s^2}_0 = 2197.745$, and $\bar{d}_1 = 86.3$, $s_0^1 = 4783.501$, respectively, which yield the pooled variance $s_p^2 = 3490.623$.

The BF test statistic is $t^*_{\text{BF}} = -3.21$; since $|t^*_{\text{BF}}| = 3.21 > t(0.975; 198) = 1.97$, we **reject** H_0 (equal variance) at significance level $\alpha = 0.05$.

3.5.3 – Independence

Independence of the error terms can be gauged visually by plotting the **residuals** e_i against the **fitted values** $\hat{Y}_i$.

If the errors are **independent**, the correlation between these should be small ($|\rho| \approx 0$); if a pattern or a trend emerges, then they are likely **dependent**. The residuals vs. fitted values chart of the previous example shows a **slight** pattern, for instance, but the correlation is so **small** ($\rho = -6 \times 10^{-18}$) that we can reasonably treat them as **independent**.

The GLR assumption is that the **errors** are independent, but we only ever work with the **residuals**, which are definitely **not independent** ($\overline{e} = 0$).

Other test: **Durbin-Watson** for auto-correlation in the residuals.

3.5.4 – Normality

If the error terms are $\mathcal{N}(0, \sigma^2)$, we expect the residuals to also be $\mathcal{N}(0, \sigma^2)$.

1. Thus, if the histogram of the **studentized residuals**

$$r_i = \frac{e_i}{\mathrm{s}\{e_i\}} = \frac{Y_i - \hat{Y}_i}{\sqrt{\mathrm{MSE}}\sqrt{1 - h_{ii}}}$$

   is not symmetrical, then they do not follow a standard normal distribution $\mathcal{N}(0, 1)$ and the error terms are unlikely to be normal.

2. If the histogram is symmetrical, we build the **normal probability** plot (also known as **quantile-quantile** plot, or **qq**-plot) from the **studentized residuals**.

For each $i = 1, \ldots, n$, we build the following table:

i	**studentized residual**	**rank**	**percentile**	$z-$**quantile**
1	r_1	k_1	p_1	z_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	r_i	k_i	p_i	z_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	r_n	k_n	p_n	z_n

The **rank** k_i is given in **increasing** order (ties use the average rank); the **approximate percentile** is

$$p_i = \frac{k_i - 0.375}{n + 0.25}, \quad \text{(blom plotting position)};$$

the **quantile** is $z_i = \Phi^{-1}(p_i)$, where $\Phi(z) = P(Z \leq z)$, $Z \sim \mathcal{N}(0, 1)$.

3. Plot the studentized residuals r_i against the quantiles z_i – the points should fall randomly about the "**normal**" line, with no systematic trend away from it. If not, the errors are unlikely to be normal.

4. Compute the **correlation** ρ between r_i and z_i, $i = 1, \ldots, n$. In order to test for

$$\begin{cases} H_0: & \text{error terms are normally distributed} \\ H_1: & H_0 \text{ is false} \end{cases}$$

we find the critical value ρ_α of the normal **probability plot correlation coefficient** (PPCC) for sample size n at a significance level α.

Decision Rule: If $\rho < \rho_\alpha$, **reject** H_0 at significance level α.

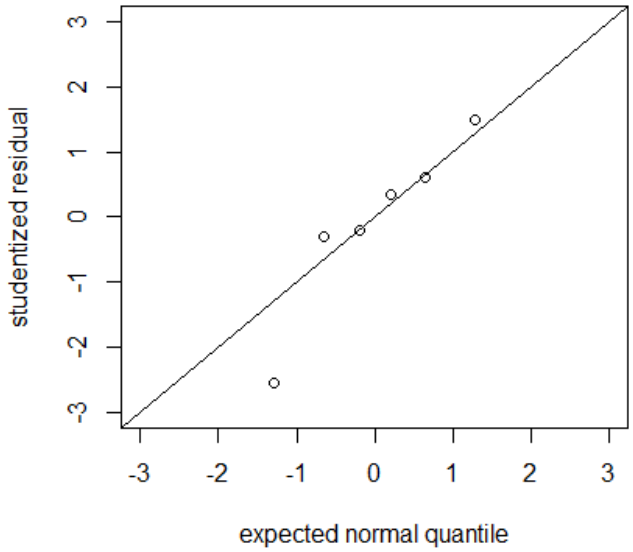
Example: consider a dataset with the following (X, Y) observations

$$(1, 7.4), (1, 8.0), (2, 7.0), (2, 10.4), (3, 19.1), (4, 20.3).$$

Assume a linear model $\mathrm{E}\{Y\} = \beta_0 + \beta_1 X$. Is the normality assumption of the error terms warranted?

Solution: the linear model is $\mathrm{E}\{Y\} = 1.802 + 4.722X$; the table is

x	y	**studentized residual**	**rank**	p	$z-$**quantile**
1	7.4	0.35	4	0.58	0.20
1	8.0	0.60	5	0.74	0.64
2	7.0	−2.57	1	0.10	−1.28
2	10.4	−0.29	2	0.26	−0.64
3	19.1	1.48	6	0.90	1.28
4	20.3	−0.21	3	0.42	−0.20



The correlation between the studentized residuals and the $z-$quantile is $\rho = 0.939$. At a significance level $\alpha = 0.05$, the critical value of the correlation in the PPCC table with $n = 6$ is 0.888, so we do not reject the normality assumption (not the same as accepting H_0).

3.5.5 – Remedial Measures

Transformations on X are used when the data exhibits a **monotone non-linear trend** with **variance constancy**; if the trend is ◜, we try $X' = \ln X$ or $X' = \sqrt{X}$; if the trend is ◞, we try $X' = e^X$ or $X' = X^2$; if it is ◟, we try $X' = \frac{1}{X}$ or $X' = e^{-X}$; if it is ◝, we try $X' = e^{-X^2}$.

Transformations on Y are used when the data exhibits **monotone non-linear trend** with **NO variance constancy**, but it is often hard to determine from the scatter plots which transformation on Y is best. The **Box-Cox** transformation helps us find a power λ which will be appropriate for the regression model

$$Y_i^{(\lambda)} = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon,$$

where $\mathbf{X}_i$ is the ith row of $\mathbf{X}$. Set

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln Y, & \lambda = 0 \end{cases}$$

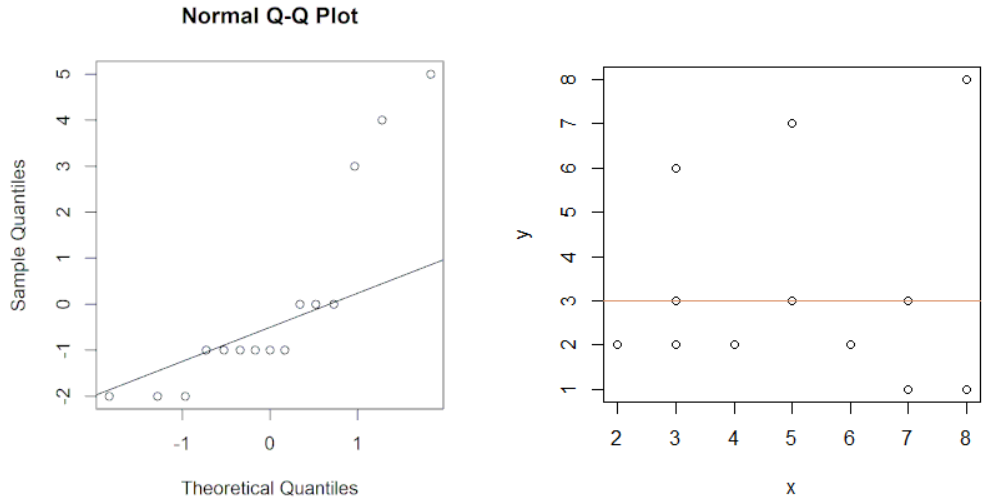
We pick the λ that minimizes the $\mathrm{SSE}(\lambda)$ resulting from the regressions.

Weighted Least-Squares are used if the data exhibits a **linear trend** with **no variance constancy**. An alternative would be to first use a transformation on Y to control the **variance**, and then a transformation on X to control the **linearity** that may have been destroyed by the first transformation.

Example: consider the following dataset

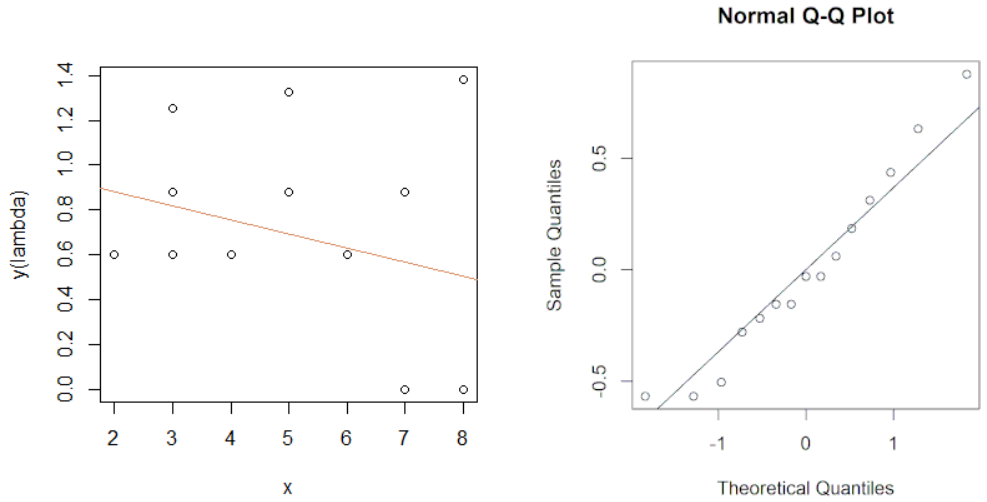
$$(7,1),\ (7,1),\ (8,1),\ (3,2),\ (2,2),\ (4,2),\ (4,2),\ (6,2),$$
$$(6,2),\ (7,3),\ (5,3),\ (3,3),\ (3,6),\ (5,7),\ (8,8).$$

The scatterplot, regression line, and normal qqplot are shown below.



The QQ plot shows that the error terms are unlikely to be normal, and so the regression model is not valid. The variance is not constant, so we use the Box-Cox transformation on Y: the optimal λ is -0.42.

The scatterplot, regression line, and normal qqplot on the transformed data are shown below.



IMPORTANT: the linear model on the original data is $\mathrm{E}\{Y\} = 3 + 0 \cdot X$. The linear model on the transformed data is

$$\mathrm{E}\left\{Y^{(-0.42)}\right\} = 1.00564 - 0.06264X$$

$\Longrightarrow$

$$\begin{aligned}
\mathrm{E}\{Y\} &= \left([\lambda\beta_0 + 1] + \lambda\beta_1 X\right)^{1/\lambda} \\
&= \left([-0.42(1.00564) + 1] + 0.42 \cdot 0.06264X\right)^{1/(-0.42)} \\
&= \frac{1}{(0.5776 + 0.0263X)^{2.380}}
\end{aligned}$$

which is **NOT** a straight line in the $xy-$plane.