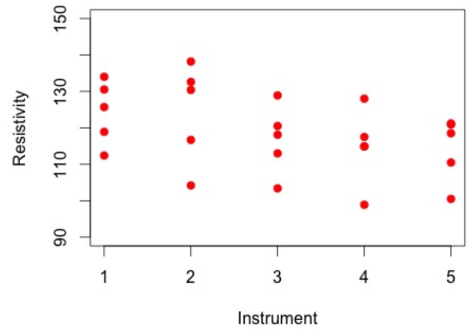
Lecture 13. Linear models with normal assumptions

One way analysis of variance

Example 13.1

Resistivity of silicon wafers was measured by five instruments.
Five wafers were measured by each instrument (25 wafers in all).




```
y=c(130.5,112.4,118.9,125.7,134.0,
    130.4,138.2,116.7,132.6,104.2,
    113.0,120.5,128.9,103.4,118.1,
    128.0,117.5,114.9,114.9, 98.9,
    121.2,110.5,118.5,100.5,120.9)
```

Let $Y_{i,j}$ be the resistivity of the jth wafer measured by instrument i, where $i, j = 1, .., 5$.

A possible model is, for $i, j = 1, .., 5$.

$$Y_{i,j} = \mu_i + \varepsilon_{i,j},$$

where $\varepsilon_{i,j}$ are independent $N(0, \sigma^2)$ random variables, and the μ_i's are unknown constants.

This can be written in matrix form: Let

$$\underset{25\times 1}{\mathbf{Y}} = \begin{pmatrix} Y_{1,1} \\ . \\ . \\ Y_{1,5} \\ Y_{2,1} \\ . \\ . \\ Y_{2,5} \\ . \\ . \\ Y_{5,1} \\ . \\ . \\ Y_{5,5} \end{pmatrix}, \quad \underset{25\times 5}{X} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ . & . & \dots & . \\ . & . & \dots & . \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ . & . & \dots & . \\ . & . & \dots & . \\ 0 & 1 & \dots & 0 \\ . & . & \dots & . \\ . & . & \dots & . \\ 0 & 0 & \dots & 1 \\ . & . & \dots & . \\ . & . & \dots & . \\ 0 & 0 & \dots & 1 \end{pmatrix}, \quad \underset{5\times 1}{\boldsymbol{\beta}} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \end{pmatrix}, \quad \underset{25\times 1}{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_{1,1} \\ . \\ . \\ \varepsilon_{1,5} \\ \varepsilon_{2,1} \\ . \\ . \\ \varepsilon_{2,5} \\ . \\ . \\ \varepsilon_{5,1} \\ . \\ . \\ \varepsilon_{5,5} \end{pmatrix},$$

Then

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

$$X^T X = \begin{pmatrix} 5 & 0 & \dots & 0 \\ 0 & 5 & \dots & 0 \\ . & . & \dots & . \\ 0 & 0 & .. & 5 \end{pmatrix}.$$

Hence

$$(X^T X)^{-1} = \begin{pmatrix} \frac{1}{5} & 0 & \dots & 0 \\ 0 & \frac{1}{5} & \dots & 0 \\ . & . & \dots & . \\ 0 & 0 & .. & \frac{1}{5} \end{pmatrix},$$

so that

$$\hat{\boldsymbol{\mu}} = (X^T X)^{-1} X^T \mathbf{Y} = \begin{pmatrix} \overline{Y_{1.}} \\ .. \\ \overline{Y_{5.}} \end{pmatrix}$$

RSS $= \sum_{i=1}^5 \sum_{j=1}^5 (Y_{i,j} - \hat{\mu}_i)^2 = \sum_{i=1}^5 \sum_{j=1}^5 (Y_{i,j} - \overline{Y_{i.}})^2$ on $n - p = 25 - 5 = 20$ degrees of freedom.

For these data, $\tilde{\sigma} = \sqrt{\text{RSS}/(n-p)} = \sqrt{2170/20} = 10.4$.

Assuming normality

- We now make a Normal assumption

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \text{N}_n(\mathbf{0}, \sigma^2 I), \qquad \text{rank } (X) = p(< n).$$

- This is a special case of the linear model of Lecture 12, so all results hold.
- Since $\mathbf{Y} \sim \text{N}_n(X\boldsymbol{\beta}, \sigma^2 I)$, the log-likelihood is

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2} S(\boldsymbol{\beta}),$$

where $S(\boldsymbol{\beta}) = (\mathbf{Y} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\boldsymbol{\beta})$.

- Maximising ℓ wrt $\boldsymbol{\beta}$ is equivalent to minimising $S(\boldsymbol{\beta})$, so MLE is

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

the same as for least squares.

- For the MLE of σ^2, we require

$$\left.\frac{\partial \ell}{\partial \sigma^2}\right|_{\hat{\boldsymbol{\beta}}, \hat{\sigma}^2} = 0,$$

$$\text{i.e.} \quad -\frac{n}{2\hat{\sigma}^2} + \frac{S(\hat{\boldsymbol{\beta}})}{2\hat{\sigma}^4} = 0$$

- . So

$$\hat{\sigma}^2 = \frac{1}{n} S(\hat{\boldsymbol{\beta}}) = \frac{1}{n}(\mathbf{Y} - X\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - X\hat{\boldsymbol{\beta}}) = \frac{1}{n}\text{RSS},$$

where RSS is 'residual sum of squares' - see last lecture.

- See example sheet for $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ for simple linear regression and one-way analysis of variance.

Lemma 13.2

(i) *If $\mathbf{Z} \sim N_n(\mathbf{0}, \sigma^2 I)$, and A is $n \times n$, symmetric, idempotent with rank r, then $\mathbf{Z}^T A \mathbf{Z} \sim \sigma^2 \chi^2_r$.*

(ii) *For a symmetric idempotent matrix A, rank(A) = trace(A)*

Proof:

- (i) $A^2 = A$ since idempotent, and so eigenvalues of A are $\lambda_i \in \{0, 1\}, i = 1, .., n, \qquad [\lambda_i \mathbf{x} = A\mathbf{x} = A^2\mathbf{x} = \lambda_i^2 \mathbf{x}]$.
- A is also symmetric, and so there exists an orthogonal Q such that

$$Q^T A Q = \text{diag } (\lambda_1, .., \lambda_n) = \text{diag } (\underbrace{1, .., 1}_{r}, \underbrace{0, ..., 0}_{n-r}) = \Lambda \text{ (say)}.$$

- Let $\mathbf{W} = Q^T \mathbf{Z}$, and so $\mathbf{Z} = Q\mathbf{W}$. Then $\mathbf{W} \sim \text{N}_n(\mathbf{0}, \sigma^2 I)$ by Proposition 11.1(i). (since cov($\mathbf{W}$) $= Q^T \sigma^2 I Q = \sigma^2 I$).
- Then

$$\mathbf{Z}^T A \mathbf{Z} = \mathbf{W}^T Q^T A Q \mathbf{W} = \mathbf{W}^T \Lambda \mathbf{W} = \sum_{i=1}^r w_i^2 \sim \sigma^2 \chi^2_r,$$

from the definition of χ^2.

- (ii)

$$\begin{aligned} \text{rank }(A) &= \text{rank }(Q^T A Q) && \text{if } Q \text{ orthogonal} \\ &= \text{rank }(\Lambda) \\ &= \text{trace }(\Lambda) \\ &= \text{trace }(Q^T A Q) \\ &= \text{trace }(A Q^T Q) && \text{since } \operatorname{tr}(AB) = \operatorname{tr}(BA) \\ &= \text{trace }(A) \end{aligned}$$

□

Theorem 13.3

For the normal linear model $\mathbf{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 I)$,

(i) $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$.

(ii) *RSS* $\sim \sigma^2 \chi^2_{n-p}$, *and so* $\hat{\sigma}^2 \sim \frac{\sigma^2}{n} \chi^2_{n-p}$.

(iii) $\hat{\boldsymbol{\beta}}$ *and* $\hat{\sigma}^2$ *are independent.*

Proof:

- (i) $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y}$, say $C\mathbf{Y}$.
  Then from Proposition 11.1(i), $\hat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$.

(ii) We can apply Lemma 13.2(i) with $\mathbf{Z} = \mathbf{Y} - X\boldsymbol{\beta} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 I_n)$ and $A = (I_n - P)$, where $P = X(X^T X)^{-1} X^T$ is the projection matrix covered after Definition 12.3.

- (P is also known as the 'hat' matrix since it projects from the observation $\mathbf{Y}$ onto the fitted values $\hat{\mathbf{Y}}$.)
- P is symmetric and idempotent, so $I_n - P$ is also symmetric and idempotent (check).
- By Lemma 13.2(ii),

  $$\operatorname{rank}(P) = \operatorname{trace}(P) = \operatorname{trace}(X(X^T X)^{-1} X^T) = \operatorname{trace}((X^T X)^{-1} X^T X) = p,$$

  so $\operatorname{rank}(I_n - P) = \operatorname{trace}(I_n - P) = n - p$.
- Note that $(I_n - P)X = 0$ (check) so that

  $$\mathbf{Z}^T A \mathbf{Z} = (\mathbf{Y} - X\boldsymbol{\beta})^T (I_n - P)(\mathbf{Y} - X\boldsymbol{\beta}) = \mathbf{Y}^T (I_n - P)\mathbf{Y} \text{ since } (I_n - P)X = 0.$$

  We know $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = (I_n - P)\mathbf{Y}$ and $(I_n - P)$ is symmetric and idempotent, and so

  $$\mathrm{RSS} = \mathbf{R}^T \mathbf{R} = \mathbf{Y}^T (I_n - P)\mathbf{Y} \qquad (= \mathbf{Z}^T A \mathbf{Z}).$$

  Hence by Lemma 13.2(i), RSS $\sim \sigma^2 \chi^2_{n-p}$ and $\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n} \sim \frac{\sigma^2}{n} \chi^2_{n-p}$.

- (iii) Let $\underset{(p+n)\times 1}{V} = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \mathbf{R} \end{pmatrix} = D\mathbf{Y}$, where $D = \begin{pmatrix} C \\ I_n - P \end{pmatrix}$ is a $(p+n) \times n$ matrix.
- By Proposition 11.1(i), V is multivariate normal with

$$\begin{aligned} \operatorname{cov}(V) = \sigma^2 D D^T &= \sigma^2 \begin{pmatrix} CC^T & C(I_n - P)^T \\ (I_n - P)C^T & (I_n - P)(I_n - P)^T \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} CC^T & C(I_n - P) \\ (I_n - P)C^T & (I_n - P) \end{pmatrix}. \end{aligned}$$

- We have $C(I_n - P) = 0$ (check) $[(X^T X)^{-1} X^T (I_n - P) = 0$ because $(I_n - P)X = 0]$.
- Hence $\hat{\boldsymbol{\beta}}$ and $\mathbf{R}$ are independent by Proposition 11.2(ii).
- Hence $\hat{\boldsymbol{\beta}}$ and RSS=$\mathbf{R}^T \mathbf{R}$ are independent, and so $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent. □.

From (ii), $\mathbb{E}(\mathrm{RSS}) = \sigma^2 (n - p)$, and so $\tilde{\sigma}^2 = \frac{\mathrm{RSS}}{n-p}$ is an unbiased estimator of σ^2.

$\tilde{\sigma}$ is often known as the *residual standard error on $n - p$ degrees of freedom.*

Example 12.1 continued

The RSS = residual sum of squares is the sum of the squared vertical distances from the data-points to the fitted straight line.

$\text{RSS} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - \hat{a}' - \hat{b}(x_i - \bar{x})^2 = 67968.$

So the estimate of

$$\tilde{\sigma}^2 = \frac{\text{RSS}}{n-p} = \frac{67968}{(24-2)} = 3089.$$

Residual standard error is $\tilde{\sigma} = \sqrt{3089} = 55.6$ on 22 degrees of freedom.

The F distribution

- Suppose that U and V are independent with $U \sim \chi^2_m$ and $V \sim \chi^2_n$.
- Then $X = (U/m)/(V/n)$ is said to have an F **distribution** on m and n degrees of freedom.
- We write $X \sim F_{m,n}$.
- Note that, if $X \sim F_{m,n}$ then $1/X \sim F_{n,m}$.
- Let $F_{m,n}(\alpha)$ be the upper $100\alpha\%$ point for the $F_{m,n}$-distribution so that if $X \sim F_{m,n}$ then $\mathbb{P}(X > F_{m,n}(\alpha)) = \alpha$. These are tabulated.
- If we need, say, the lower 5% point of $F_{m,n}$, then find the upper 5% point x of $F_{n,m}$ and use $\mathbb{P}(F_{m,n} < 1/x) = \mathbb{P}(F_{n,m} > x)$.
- Note further that it is immediate from the definitions of t_n and $F_{1,n}$ that if $Y \sim t_n$ then $Y^2 \sim F_{1,n}$, since ratio of independent χ^2_1 and χ^2_n variables.