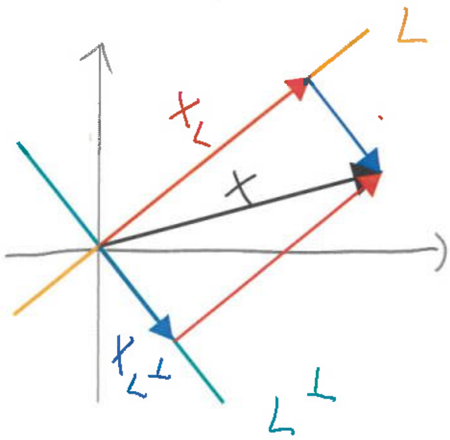
Recall that any $\mathbf{x} \in \mathbb{R}^n$ can be uniquely written as $\mathbf{x} = \mathbf{x}_L + \mathbf{x}_{L^\perp}$, where $\mathbf{x}_L \in L$ and $\mathbf{x}_{L^\perp} \in L^\perp$.



$\left.\begin{array}{l} Px_L = x_L \\ Px_{L^\perp} = 0 \end{array}\right] \Rightarrow Px = P(x_L + x_{L^\perp}) = Px_L = x_L$

Let $\mathbf{x} \in \mathbb{R}^n$. Then $P^2\mathbf{x} = P(\underbrace{P\mathbf{x}}_{x_L}) = P\mathbf{x}_L = P\mathbf{x}$. Hence, $P^2 = P$.

- $y = y_L + y_{L^\perp}$
- IF WE MULTIPLY AN ELEMENT (VECTOR) OF L WITH AN ELEMENT OF $L^\perp$ WE GET 0

For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

$$\mathbf{x}^T P^T \mathbf{y} = (\underbrace{P\mathbf{x}}_{\in L})^T \mathbf{y} = (\underbrace{P\mathbf{x}}_{\mathbf{x}_L})^T \mathbf{y}_L = \mathbf{x}_L^T P\mathbf{y} = \mathbf{x}^T P\mathbf{y}$$

$(Px)^T y_{L^\perp} = 0$

$Py = y_L$

$x^T P y = (x_L + x_{L^\perp})^T P y = x_L^T P y$ BECAUSE $x_{L^\perp}^T P y = 0$

Hence, $P^T = P$

$\boxed{\Leftarrow:}$

Let L be the space spanned by the columns of P. $L = \text{Span}(P) = \{Pz : z \in \mathbb{R}^m\}$

- Let $\mathbf{x} \in L$. Then $\exists \mathbf{z} \in \mathbb{R}^n : \mathbf{x} = P\mathbf{z}$. Hence, $P\mathbf{x} = P^2\mathbf{z} \overset{\text{idempot}}{=} P\mathbf{z} = \mathbf{x}$.
- Let $\mathbf{x} \in L^\perp$. Then for all $\mathbf{y} \in \mathbb{R}^n$: $(P\mathbf{x})^T\mathbf{y} = \mathbf{x}^T P^T \mathbf{y} \overset{\text{symm}}{=} \mathbf{x}^T \underbrace{P\mathbf{y}}_{\in L} = \mathbf{0}$. Hence $P\mathbf{x} = \mathbf{0}$.

$e_i = (0, \dots, 0, \underset{i\text{-th}}{1}, 0, \dots 0)$

- The projection matrix is unique. Indeed, for each i, the vector $\mathbf{e}_i$ can be uniquely written as $\mathbf{e}_i = \mathbf{x} + \mathbf{y}$, where $\mathbf{x} \in L$ and $\mathbf{y} \in L^\perp$. Then the ith column of P is $P\mathbf{e}_i = \mathbf{x}$.
- If $\mathbf{x}_1, \dots, \mathbf{x}_r$ are a basis of L then the projection onto L is given by

$$P = X(X^TX)^{-1}X^T,$$

where $X = (\mathbf{x}_1, \dots, \mathbf{x}_r)$. [prove this directly via the definition of the projection matrix or check $P^2 = P$, $P^T = P$, $\underbrace{\text{span}(P)}_{\text{space spanned by the columns of } P} = L$ or .]

$y \in \mathbb{R}^m$

$y = y_L + y_{L^\perp}$

$\rightarrow y_L = Xz$, FOR SOME $z \in \mathbb{R}^m$

$Py_L = X(X^TX)^{-1}X^Ty_L =$

$= X(X^TX)^{-1}X^TXz =$

$= Xz = y_L$

$Py_{L^\perp} = X(X^TX)^{-1}\underbrace{X^Ty_{L^\perp}}_{=0} = 0$

- If $\mathbf{x}_1, \dots, \mathbf{x}_r$ are an *orthonormal* basis then $P = XX^T$. $X^TX = I$
- $I_n - P$ is the projection matrix onto $L^\perp$

$y = y_L + y_{L^\perp}$

$y = Iy = (I-P)y + Py = (I-P)y + y_L$

$\rightarrow (I-P)y = y_{L^\perp}$

(can be checked using original definition).

Example 54
$n = 3$

If $L = \text{span}(\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix})$ then $P = \begin{pmatrix}1&0&0\\0&1&0\\0&0&0\end{pmatrix}$.

If $L = \text{span}(\begin{pmatrix}1\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix})$ then $P = \begin{pmatrix}1/2&1/2&0\\1/2&1/2&0\\0&0&1\end{pmatrix}$.

If $L = \text{span}(\mathbf{x})$ for some $\mathbf{x} \in \mathbb{R}^n$ then $P = \frac{\mathbf{x}\mathbf{x}^T}{\mathbf{x}^T\mathbf{x}}$.

$Py = \frac{x x^T y}{x^T x} = \frac{x^T y\, x}{\|x\|^2} = \frac{\langle x, y\rangle}{\|x\|^2} x$

Lemma 12
If A is an $n \times n$ projection matrix (i.e. $A = A^T$, $A^2 = A$) of rank r then

1. r of the eigenvalues of A are 1 and $n - r$ are 0,
2. $\text{rank}\, A = \text{trace}\, A$,

IDEMPOTENT

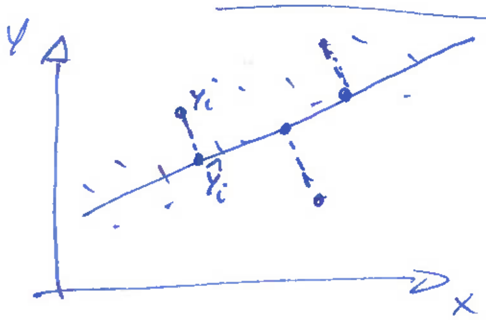
Proof Let $\mathbf{x}$ be an eigenvector of A, with eigenvalue λ. Then $\lambda\mathbf{x} = A\mathbf{x} = A^2\mathbf{x} = A\lambda x = \lambda A\mathbf{x} = \lambda^2\mathbf{x}$. $\overset{\mathbf{x}\neq 0}{\Longrightarrow} \lambda = \lambda^2 \implies \lambda \in \{0, 1\}$.

1. A symmetric $\implies \exists P$ (orthogonal) s.t. $P^{-1}AP = D$, where D is diagonal with 0s and 1s on the diagonal. Since P is non-singular, $\text{rank}\, A = \text{rank}\, D$. Hence D has r ones down the diagonal.
2. $\text{trace}(A) = \text{trace}(APP^{-1}) \overset{\text{Le6}}{=} \text{trace}(P^{-1}AP) = \text{trace}\, D = \text{rank}\, A$?

trace(AB) = trace(BA)

9.9 Residuals, Estimation of the variance

Definition 20
$\hat{\mathbf{Y}} = X\hat{\beta}$, where $\hat{\beta}$ is a least squares estimator, is called the *vector of fitted values*.



$\hat{Y} = X\hat{\beta}$ BY DEF OF $\hat{\beta}$

In the full rank case, $\hat{\mathbf{Y}} = \underbrace{X(X^TX)^{-1}X^T}_{P}\mathbf{Y}$.

Lemma 13
$\hat{\mathbf{Y}}$ is unique and

$$\hat{\mathbf{Y}} = P\mathbf{Y},$$

where P is the projection matrix onto the column space of X.

$\text{span}(X) = \{Xz : z \in \mathbb{R}^n\}$

Because of this lemma, P is sometimes called the *hat matrix* (it puts the hat on $\mathbf{Y}$, i.e. $\hat{\mathbf{Y}} = P\mathbf{Y}$).

Proof Suppose $\hat{\beta}$ is a LSE of β. We already know P is unique, hence $P\mathbf{Y}$ is unique. Thus it suffices to show $\hat{\mathbf{Y}} = P\mathbf{Y}$. Since $P\mathbf{Y} \in \text{span}(X)$ there exists γ s.t. $X\gamma = P\mathbf{Y}$. Then

$\gamma \in \mathbb{R}^m$

$S(\hat{\beta}) = \|Y - X\hat{\beta}\|^2$

$S(\gamma) = \|Y - X\gamma\|^2$

$$\begin{aligned}
S(\hat{\beta}) &= \|\mathbf{Y} - P\mathbf{Y} + P\mathbf{Y} - X\hat{\beta}\|^2 \\
&= \underbrace{\|\mathbf{Y} - P\mathbf{Y}\|^2}_{=S(\gamma)} + \|P\mathbf{Y} - X\hat{\beta}\|^2 + 2\underbrace{\underbrace{(\mathbf{Y} - P\mathbf{Y})^T}_{=\mathbf{Y}^T(I-P)}\underbrace{(P\mathbf{Y} - X\hat{\beta})}_{\in \text{span}(X)}}_{=0} \\
&\geq S(\hat{\beta}) + \|P\mathbf{Y} - X\hat{\beta}\|^2,
\end{aligned}$$

since $\hat{\beta}$ minimises S. Thus $\|P\mathbf{Y} - X\hat{\beta}\| = 0$. Therefore, $P\mathbf{Y} = X\hat{\beta}. = \hat{Y}$

BY DEF OF $\hat{Y}$

Definition 21
$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ is called the *vector of residuals*.

Remark Equivalent form: Using Lemma 13,

$e = (I - P)Y = QY$

$$\mathbf{e} = \mathbf{Y} - P\mathbf{Y} = Q\mathbf{Y},$$

where $Q = I - P$ is the projection matrix onto $\text{span}(X)^\perp$.

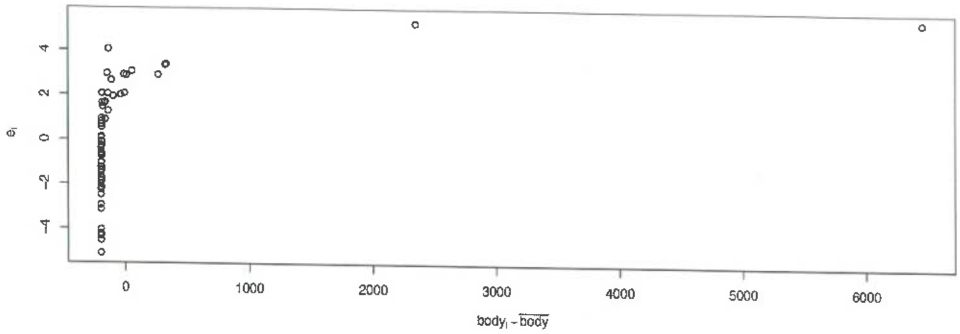
Remark

$E[Y] = X\beta$

$$\text{E}(\mathbf{e}) = \text{E}[Q\mathbf{Y}] = Q\,\text{E}\,\mathbf{Y} = \underbrace{QX}_{=0}\beta = \mathbf{0}$$

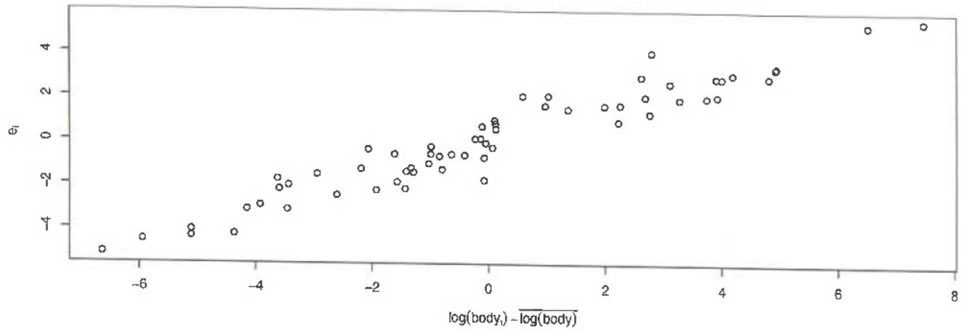
$\mathbf{e}$ can be used to see how well the model and the data agree and to see if certain observations are larger or smaller than predicted by the model.

Suppose we suspect $\mathbf{Z} = \mathbf{o}$ might be important. To investigate this, we plot $(Q\mathbf{o})_i = o_i - \bar{o}$ vs e_i. If the model (1) is true then the plot below should roughly look like the previous plot.



The fit of (1) does not seem to be good; however simply including $\mathbf{o}$ in the model does not seem to be reasonable because the above plot does not look like a linear relationship.

Let $z_j = \log(o_j)$. A plot of $(Q\mathbf{z})_i$ vs e_i:



This looks like a linear relationship with slope $\neq 0$ $\to$ could include $z_i\beta_2$ in model (1).

Residual Sum of Squares

Definition 22
$\mathrm{RSS} = \mathbf{e}^T\mathbf{e}$ is called the *residual sum of squares.*

RSS quantifies the departure of the data from the model. It is the minimum of $S(\beta)$.

Remark Other forms:

PY

- RSS $= \sum_{i=1}^n e_i^2$
- RSS $= S(\hat\beta) = \|Y - X\hat\beta\|^2$ $= \|Y - \hat{Y}\|^2 = \|e\|^2$
- RSS $= \underbrace{(Q\mathbf{Y})}_{e}^T Q\mathbf{Y} = \mathbf{Y}^T Q^T Q\mathbf{Y} = \mathbf{Y}^T Q\mathbf{Y}$

  ↳ Q IS SYMMETRIC AND IDEMPOTENT
- RSS $= \mathbf{Y}^T\mathbf{Y} - \hat{\mathbf{Y}}^T\hat{\mathbf{Y}}$.

  Indeed, RSS $= (\mathbf{Y}-\hat{\mathbf{Y}})^T(\mathbf{Y}-\hat{\mathbf{Y}}) = \mathbf{Y}^T\mathbf{Y} - 2\hat{\mathbf{Y}}^T\mathbf{Y} + \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} = \mathbf{Y}^T\mathbf{Y} - \hat{\mathbf{Y}}^T\hat{\mathbf{Y}}$. The last equality holds because $\hat{\mathbf{Y}}^T\mathbf{Y} = (P\mathbf{Y})^T\mathbf{Y} = (PP\mathbf{Y})^T\mathbf{Y} = \mathbf{Y}^T P^T P^T \mathbf{Y} = (P\mathbf{Y})^T P\mathbf{Y} = \hat{\mathbf{Y}}^T\hat{\mathbf{Y}}$.

  BY DEF $\hat{Y}$ ↳ $P^2 = P$

Theorem 8
$\hat\sigma^2 := \frac{\text{RSS}}{n-r}$ is an unbiased estimator of σ^2.

$I\sigma^2 = \text{cov}(\varepsilon)$
BY SOA

Recall: $r = \text{rank}(X)$ $= \text{rank}(P)$

Proof Let $Q = I - P$. Since P is a projection matrix, Q is a projection matrix as well. Hence, RSS $= Y^T QY$.

BECAUSE RSS IS A ONE-DIMENSIONAL OBJECT

trace(AB) = trace(BA)   BY LINEARITY OF TRACE

$$\begin{aligned}
\text{E(RSS)} &= \text{E trace RSS} = \text{E trace}(\mathbf{Y}^T Q\mathbf{Y}) \overset{\text{Le 6}}{=} \text{E trace}(Q\mathbf{Y}\mathbf{Y}^T) = \text{trace}(Q\,\text{E}(\mathbf{Y}\mathbf{Y}^T)) \\
&= \text{trace}(Q[\text{cov}\,\mathbf{Y} + \text{E}(\mathbf{Y})\,\text{E}(\mathbf{Y})^T]) = \text{trace}(Q\sigma^2) + \text{trace}(QX\beta(X\beta)^T) \\
&= \sigma^2\,\text{trace}(I-P) + 0 = \sigma^2(n - \text{trace}(P)) \\
&\overset{\text{Le 12}}{=} \sigma^2(n - \text{rank}(P)) = \sigma^2(n-r).
\end{aligned}$$

$\in \text{SPAN}(X)$ ⇝ $QX\beta = 0$

$E[YY^T] = \text{cov}(Y) + E[Y]E[Y]^T$
$\text{cov}(Y) = \sigma^2 I$

Remark This is a generalisation of the result that the sample variance s^2 is an unbiased estimator for σ^2 when $Y_1, \dots, Y_n$ are i.i.d. with unknown mean μ and unknown variance σ^2.

Indeed, we can write this iid setup as the linear model $\mathbf{Y} = \underbrace{\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}}_{X}\mu + \epsilon$ with $\text{E}\,\epsilon = \mathbf{0}$ and $\text{cov}\,\epsilon = \sigma^2 I$. Then $P = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \frac{1}{n}\mathbf{X}\mathbf{X}^T$ and thus $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - P\mathbf{Y} =$

$X^TX = n$

$= Y - \frac{1}{n}\begin{pmatrix}1 & \cdots & 1\\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1\end{pmatrix}Y$

$= Y - \begin{pmatrix}\bar{Y}\\ \vdots \\ \bar{Y}\end{pmatrix}$

$e = \mathbf{Y} - \begin{pmatrix} \bar{Y} \\ \vdots \\ \bar{Y} \end{pmatrix}$. Hence,

$$e^T e = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$\frac{\text{RSS}}{n-r} = \underbrace{\frac{\sum(Y_i - \bar{Y})^2}{n-1}}_{=s^2=\text{sample variance}} = s^2$$

which we already know is unbiased for σ^2.

$$E[s^2] = \sigma^2$$

Coefficient of Determination (R^2)

In the simplest model with only an intercept term, i.e. in

$$\mathbf{Y} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \beta_1 + \epsilon, \quad \mathrm{E}\,\epsilon = \mathbf{0}$$

we have RSS $= \sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Larger models, i.e. models with more columns in X will only lead to smaller RSS.

For models containing an *intercept term*, (i.e. X contains a column consisting of 1s (or any other constant)), a popular measure of the quality of a model is

$$R^2 = 1 - \frac{\text{RSS}}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2},$$

called the *coefficient of determination* or simply R^2. A smaller RSS is "better", thus we want a large R^2. Note: $0 \leq R^2 \leq 1$ and $R^2 = 1$ for a "perfect" model.

Remark (Intuitive interpretation) RSS $/n$ is an estimator of σ^2. $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is an estimator of σ^2 in the model with only the intercept term (let us call this the "total variance").

Thus $\frac{\text{RSS}/n}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \approx \frac{\text{Variance in the model}}{\text{total variance}}$ and hence

$$R^2 \approx \frac{\text{total variance} - \text{variance in model}}{\text{total variance}}$$

Hence, $R^2 \approx$ fraction of the total variance of the data that "is explained" by the model.