Assignment sheet 10

November 29, 2017

Assignment 1 (Extreme cases $p \in \{0, n\}$)**.** Consider a linea model $y = X\beta + \varepsilon$ where $X_{n\times p}$ is full rank and $\varepsilon \sim N_n(0, \sigma^2 I)$. Assume $n = p$. What are the estimators for β et σ^2, the errors and the fitted values ? Comment on it. Do the same for the cases $p = 0$ and $p = 1$.

Assignment 2 (R^2 and around)**.** Remember the definitions

$$R_0^2 = \frac{\|\widehat{y}\|^2}{\|y\|^2} \quad R^2 = \frac{\|\widehat{y}\|^2 - \|\overline{y}\mathbf{1}\|^2}{\|y\|^2 - \|\overline{y}\mathbf{1}\|^2} \quad R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p} \quad R_{0a}^2 = 1 - (1 - R_0^2)\frac{n}{n-p}.$$

(a) Show that R^2 is linked to the empirical correlaton

$$R^2 = \operatorname{corr}^2\left[(\widehat{y}_i)_{i=1}^n, (y_i)_{i=1}^n\right],$$

where

$$\operatorname{corr}\left[(y_i)_{i=1}^n, (z_i)_{i=1}^n\right] = \frac{\sum_{i=1}^n (y_i - \overline{y})(z_i - \overline{z})}{\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}\sqrt{\sum_{i=1}^n (z_i - \overline{z})^2}}$$

(0 is the denominator is 0).

(b) Show that $R^2 \le R_0^2$. When does it become an inequality ?

(c) She that $R_{0a}^2 \le R_0^2$. When does it become an equality ?

Remark. It can be shown as well that

$$R_{0a}^2 = 1 - \frac{\|e\|^2/(n-p)}{\|y\|^2/n}.$$

A similar result is true for R_a^2 and R^2, but the computations are more complicated.

Assignment 3 (On the Gauss-Markov theorem)**.** Let $Y = X\beta + \varepsilon$ with $\mathbb{E}(\varepsilon) = 0, \operatorname{Var}\varepsilon = \sigma^2 I$. Let $\hat{\beta}$ l'estimateur des moindres carrés de β, et $\tilde{\beta}$ another *linear and unbiased* estimator for β.
Show that

$$\operatorname{MSE}(c^t\tilde{\beta}) \ge \operatorname{MSE}(c^t\hat{\beta}), \quad \forall c \in \mathbb{R}^p,$$

is equivalent to the conclusion of Gauss-Markov theorem. Here $\operatorname{MSE}(\hat{\theta}) = \mathbb{E}\{(\hat{\theta} - \theta)^2\}$ is the mean square error for $\hat{\theta}$ ($\operatorname{MSE}(\hat{\theta}) = \operatorname{BIAS}(\hat{\theta})^2 + \operatorname{Var}\hat{\theta}$).

Assignment 4 (Diagnostic)**.** a) Figure 1 represents the the standardized residuals plotted against the fitted values for 4 different set of x_i's. For every case, discuss the fit of the model and explain briefly how you could fix mis-fits, if any are present.

b) Figure 2 represents 4 gaussian Q-Q plots. In every case, the covariates do not come from a Gaussian distribution. Actually, they are generated from distributions with

i) tails heavier than gaussian ;

ii) tails lighter than gaussian ;

iii) positive *skewness* ;

iv) negative skewness.

Match each case i)–iv) with a Q-Q plot from Figure 2 and comment.

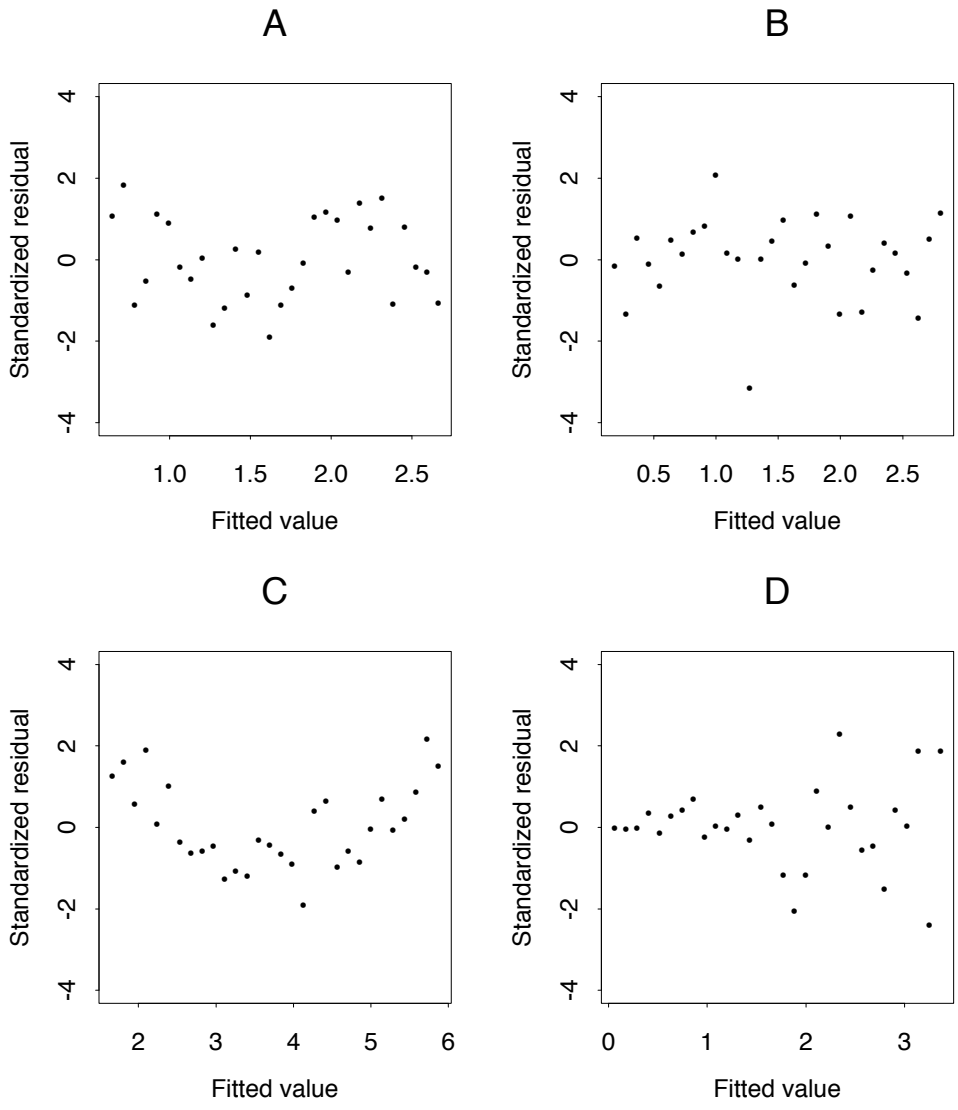

Figure 1 – Standardised residuals vs fitted values, Gaussian models

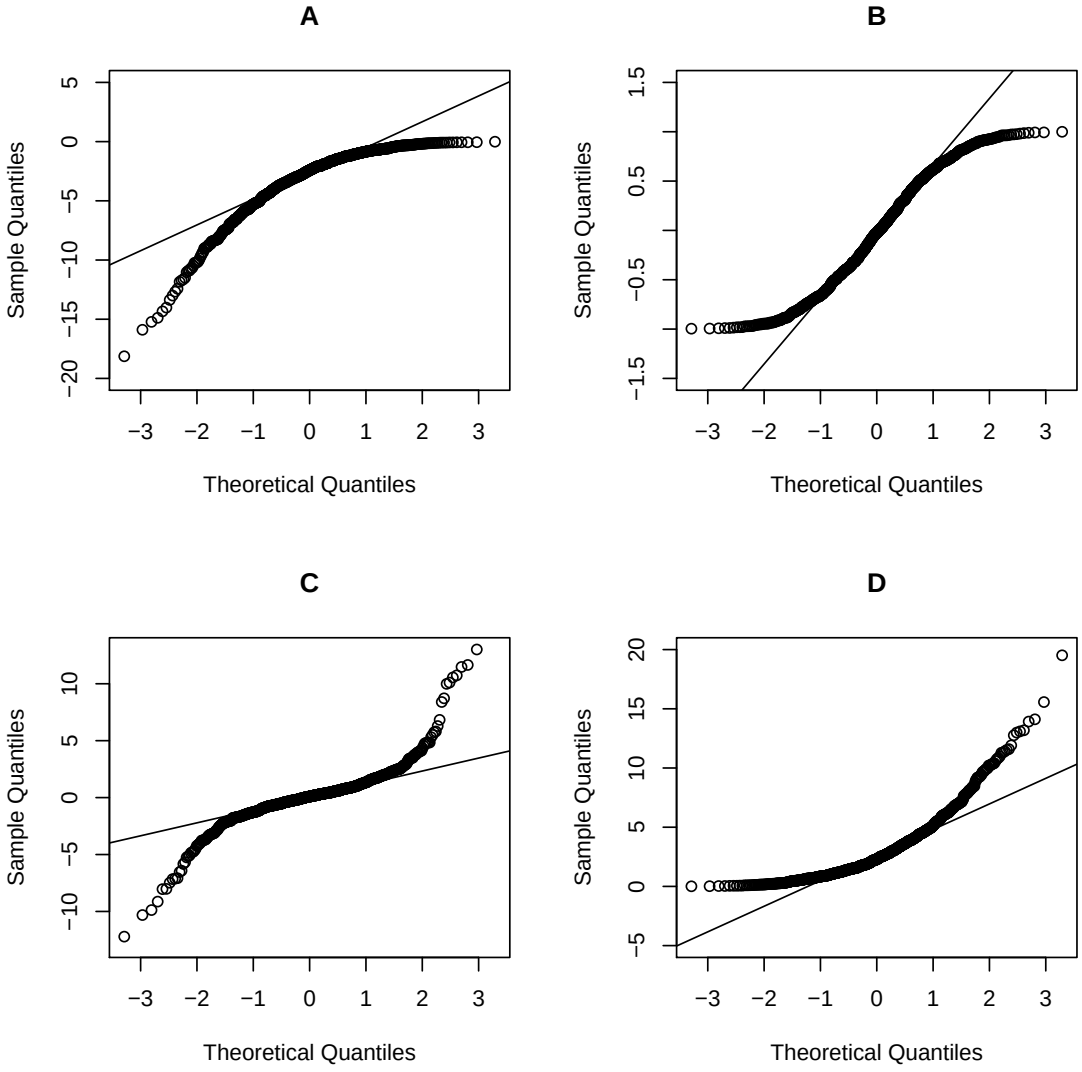

Figure 2 – Normal Q-Q plots for non Normal covariates.

Assignment 5 (confidence and prediction intervals)**.** The following table gives the estimators, the standard errors and the correlations for the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ fitted to $n = 13$ cement values from the example given in the course.

```
             Estimate   SE           Correlations of Estimates
(Intercept)  48.19      3.913            (Intercept)    x1    x2
         x1   1.70      0.205        x1     -0.736
         x2   0.66      0.044        x2     -0.416   -0.203
         x3   0.25      0.185        x3     -0.828    0.822  -0.089
```

a) Explain how R calculates the standard errors and the correlations that appear in the above table.

b) What is the prediction, under this model, of y when $x_1 = x_2 = x_3 = 1$? By how much it will change if instead $x_1 = 5$? And if $x_1 = x_2 = 5$?

c) Using only the above information and $t_9(0.975) = 2.262$, $t_9(0.95) = 1.833$, calculate under this model the confidence intervals for β_0, β_1, β_2 et β_3 at significance level $\alpha = 0.95$. Calculate a 0.9 confidence interval for $\beta_2 - \beta_3$.

Assignment 6. We fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$ to the cement dataset from the course ($n = 13$). R gives the following table :

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 48.19363    3.91330  12.315 6.17e-07 ***
x1           1.69589    0.20458   8.290 1.66e-05 ***
x2           0.65691    0.04423  14.851 1.23e-07 ***
x3           0.25002    0.18471   1.354    0.209
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

a) Explain in detail how R calculates the values in the columns "`t value`" and "`Pr(>|t|)`". What do these values mean ? Comment the observed numbers in the table.

b) Knowing that $\widehat{\text{corr}}(\hat{\beta}_2, \hat{\beta}_3) = -0.08911$, what is the p-value for the null hypothesis $\beta_2 - \beta_3 = 0$? For a 0.05 test, can we reject the null hypothesis ?

Assignment 7 (efficient computation of Cook's distance)**.** We have seen a measure of the influence of the k-th observation over the regression coefficient. This measure, *Cook's distance*, is defined as

$$C_k = \frac{1}{ps^2} \|\hat{y} - \hat{y}_{-k}\|^2,$$

where $\hat{y}_{-k} = X\hat{\beta}_{-k}$ and $\hat{\beta}_{-k}$ is the estimator of β without the k-th observation. It seems like one would need $n + 1$ regressions in order to calculate $C_1, \ldots, C_n$. We shall see that one can get the C_k's *using only the complete regression on* (y, X) by means of the formula

$$C_k = \frac{r_k^2 h_{kk}}{p(1 - h_{kk})}, \tag{1}$$

where r_k is the k-standardised residual and h_{kk} is the k-th diagonal element of the hat matrix $H = X(X^T X)^{-1} X^T$.

Let x_k^T be the k-th row of X, so that $x_k \in \mathbb{R}^p$ and

$$X^T = (x_1, \ldots, x_n)_{p \times n}.$$

Denote X_{-k} the $n \times p$ matrix whose l-th row is x_l^T if $l \neq k$ and whose kth row is $0 \in \mathbb{R}^p$. In symbols

$$X_{-k}^T = (x_1, \ldots, x_{k-1}, 0, x_{k+1}, \ldots, x_n).$$

In this exercise, you can use the identity

$$(x_1, \ldots, x_n) \begin{pmatrix} z_1^T \\ \vdots \\ z_n^T \end{pmatrix} = \sum_{i=1}^{n} x_i z_i^T \in \mathbb{R}^{p \times q},$$

where $x_i \in \mathbb{R}^p, z_i \in \mathbb{R}^q, i = 1, \ldots, n$.
Moreover, for compatible matrices A, B and C,

$$\mathrm{row}_j(AB) = \mathrm{row}_j(A) \cdot B,$$

$$\mathrm{col}_k(AB) = A \cdot \ \mathrm{col}_k(B)$$

$$(ACB)_{j,k} = \mathrm{row}_j(A) \cdot C \cdot \mathrm{col}_k(B),$$

where $\mathrm{row}_j(A)$ represents the j-th row of A, *as a row (rather than column) vector*, $\mathrm{col}_k(B)$ represnts the k-th column of B, as a column vector, and "$\cdot$" is the usual matrix product.

(i). Show that $X_{-k}^T X_{-k} = X^T X - x_k x_k^T$.

(ii). (a) Show the Sherman–Morrison formula

$$\left(A + uv^T\right)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u},$$

where $A_{n \times n}$ is invertible and $u, v \in \mathbb{R}^n$ satisfy $v^T A^{-1} u \neq -1$.

(b) Deduce that

$$(X_{-k}^T X_{-k})^{-1} = \left(I + \frac{1}{1 - h_{kk}} \left(X^T X\right)^{-1} x_k x_k^T \right) \left(X^T X\right)^{-1}.$$

(iii). Show that

(a) $X_{-k}^T y = X^T y - y_k x_k$,

(b) $x_k^T (X^T X)^{-1} X_{-k}^T y = (1 - h_{kk}) y_k - e_k$,

and conclude that

$$\hat{\beta}_{-k} = \hat{\beta} - \frac{e_k \left(X^T X\right)^{-1} x_k}{1 - h_{kk}}.$$

(iv). Lastly, show that $\|\hat{y} - \hat{y}_{-k}\|^2 = h_{kk} e_k^2 / (1 - h_{kk})^2$, and conclude (1).

Assignment 8 (best design)**.** Consider the linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \dots, n,$$

where $\beta_0, \beta_1 \in \mathbb{R}$, $\mathbb{E}[\epsilon] = 0$ and $\operatorname{Var} \epsilon = \sigma^2 I_n$ (and $n \geq 2$). This is called *simple linear regrssion.*

(i). Write down the design matrix for this model and give a necessary and sufficient condition for it to be of full rank.

(ii). Find the covariance matrix of the least squares estimator $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1)^T$.

(iii). Suppose that you can choose all the $x_1, \dots, x_n$ as you wish, but constrained to be in $[-1, 1]$. How would you choose them in order to minimise the variance of $\widehat{\beta}_1$?