Answer sheet 10

Assignment 1. The matrix X is of full rank, hence invertible. We obtain

$$\widehat{\beta} = (X^tX)^{-1}X^ty = X^{-1}y, \qquad \widehat{y} = X\widehat{\beta} = y, \qquad e = y - \widehat{y} = 0, \qquad \widehat{\sigma}^2 = n^{-1}\|e\|^2 = 0.$$

We see that the errors are 0 and the fit of the model is perfect ($R_0^2 = R^2 = 1$). The unbiased estimator of the variance $S^2 = \|e\|^2/(n-p) = 0/0$ is not defined and $H = X(X^tX)^{-1}X^t = I_n$. The column space of X is $\mathcal{M}(X) = \mathbb{R}^n$.
If $p = 0$, then we don't have parameters, not even the costant, and we project on the space $\text{span}(\emptyset) = \{0\}$. In particular $H = 0$ and $\widehat{y} = 0$, $e = y$ and $S^2 = \widehat{\sigma}^2 = \|y\|^2/n$. The model assumes that y_i are iid *and of zero mean.* This is why the unbiased estimator for the variance is $\|y\|^2 = \sum_{i=1}^n y_i^2$ and not $\sum_{i=1}^n (y_i - \overline{y})^2$ and we divide by n and not by $n-1$.
If $p = 1$ then X is just a vector, so $\widehat{\beta} = X^ty/\|X\|^2$. In the special case where X is a vector of ones, $\widehat{\beta} = \overline{y}$, $\widehat{y} = \overline{y}\mathbf{1}_n$, $e_i = y_i - \overline{y}$ and $S^2 = \|e\|^2/(n-1) = \sum(y_i - \overline{y})^2/(n-1)$. This is just estimating the mean of a distribution from a sample of size n.

Assignment 2.

a) We know that e and $\widehat{y}$ are independent and that the standardized residuals follow approximatively a standard normal (basically, the standardized residuals need to take values between -2 and 2 independently on the values of $\hat{y}_j$). The hyphotesis are respectively ($\epsilon \sim N(0, \sigma^2 I)$, ...).
   — Picture A : the fit seems good.
   — Picture B : there is an outlier $r_j < -2$.
   — Picture C : The fitted values and the standardized residuals are dependent. It seems that there is a quadratic relation between the two and that we might need to add a quadratic term to the model. It could be useful to plot r and each of the column of X to see if there are really quadratic relations.
   — Picture D : the hyphotesis of homoskedasticity is not satisfied because the variance of the residuals is not constant. A possibly way to deal with heteroskedasticity is to use weighted regression.

b) Is the distribution of the covariates has the left tail heavier than the Normal, the empirical quantiles on the left of the picture will be below the diagonal on the QQ plot. In this case indeed, $G(x) \gg F(x)$ whenever $x \to -\infty$, where G represents the c.d.f. of the covariates and F the normal law. For $alpha > 0$ small, $x = F^{-1}(\alpha)$, then $G(x) \gg \alpha$, hence $G^{-1}(\alpha) < x$, and this means that the quantiles are below the diagonal line. On the contrary, if the left tail is lighter, then the empirical quantiles will be above the diagonal line.

   In the same way, a heavy tail on the right implies $1 - G(x) \gg 1 - F(x)$ when $x \to \infty$, hence for α close to 1, $G^{-1}(\alpha) > F^{-1}(\alpha)$ and the empirical quantiles on the right of the picture as going to be above the line.
   — Pic A : lighter tail on the right and heavier on the left : negative skewness.
   — Pic B : tails less heavy than the ones of a gaussian.
   — Pic C : tails heavier than gaussian.
   — Pic D : lighter tail on the left and heavier on the right : positive skewness.

Assignment 3. a) The covariance matrix of $\hat{\beta}$ is $\sigma^2(X^TX)^{-1}$. Since we don't know σ^2, we estimate the latter by $\widehat{var}\hat{\beta} = S^2(X^TX)^{-1}$. Denote $v_{ij} = ((X^TX)^{-1})_{ij}$, $i = 0, 1, 2, 3$, $j =$

$0, 1, 2, 3$ (the indexing starts at 0). Then the i-th standard error is estimated by $\widehat{\mathrm{SE}}(\hat{\beta}_i) = \sqrt{\widehat{var}\hat{\beta}_{ii}} = \sqrt{S^2 v_{ii}}$. For the correlation, we have

$$\widehat{\mathrm{corr}}(\hat{\beta}_i, \hat{\beta}_j) = \frac{[\widehat{var}\hat{\beta}]_{ij}}{\sqrt{[\widehat{var}\hat{\beta}]_{ii}}\sqrt{\widehat{var}[\hat{\beta}]_{jj}}} = \frac{S^2 v_{ij}}{\sqrt{S^2 v_{ii}}\sqrt{S^2 v_{jj}}} = \frac{v_{ij}}{\sqrt{v_{ii} v_{jj}}}.$$

b) The prediction is

$$\hat{y}_+ = x_+^T \hat{\beta}$$

and so

$$\hat{y}_+ = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3.$$

In comparison with the case $x_1 = x_2 = x_3 = 1$, the expectation will increase by $4\hat{\beta}_1 = 4 \times 1.70 = 6.80$ if $x_1 = 5$, and by $4\hat{\beta}_2 = 4 \times 0.66 = 2.64$ if $x_2 = 5$. More explicitly,

$$\begin{aligned} x_1 = x_2 = x_3 = 1 &\implies \hat{y}_+ = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 = 48.19 + 1.70 + 0.66 + 0.25 = 50.80 \\ x_1 = x_2 = 5, x_3 = 1 &\implies \hat{y}_+ = 48.19 + 1.70 \times 5 + 0.66 \times 5 + 0.25 = 60.24 \end{aligned}$$

c) For the i-th coordinate of β, the confidence interval is

$$\hat{\beta}_i \pm \sqrt{S^2 v_{ii}} t_{n-p}(1 - \alpha/2) = \hat{\beta}_i \pm \widehat{\mathrm{SE}}(\hat{\beta}_i)\, t_{n-p}(1 - \alpha/2), \quad i = 0, 1, 2, 3.$$

Here $n = 13$, $p = 4$, $\alpha = 0.05$, $t_9(0.975) = 2.262$, so the intervals are

$$[39.34, 57.04], \quad [1.236, 2.164], \quad [0.5605, 0.7595], \quad [-0.1685, 0.6685].$$

More generally, if $c \in \mathbb{R}^p$, the confidence interval for $c^T\beta$ is

$$c^T \hat{\beta} \pm t_{n-p}(1 - \alpha/2) \sqrt{S^2 c^T (X^T X)^{-1} c}.$$

In the specific case of the question, the relevant choice for c is $c = (0, 0, 1, -1)^T$. This gives

$$\begin{aligned} S^2 c^T (X^T X)^{-1} c &= S^2 v_{22} + S^2 v_{33} - 2 \frac{v_{23}}{\sqrt{v_{22} v_{33}}} \sqrt{S^2 v_{22}} \sqrt{S^2 v_{33}} \\ &= \left(\widehat{\mathrm{SE}}(\hat{\beta}_2)\right)^2 + \left(\widehat{\mathrm{SE}}(\hat{\beta}_3)\right)^2 - 2\, \widehat{\mathrm{corr}}(\hat{\beta}_2, \hat{\beta}_3)\, \widehat{\mathrm{SE}}(\hat{\beta}_2)\, \widehat{\mathrm{SE}}(\hat{\beta}_3) \\ &= 0.044^2 + 0.185^2 - 2 \cdot (-0.089) \cdot 0.044 \cdot 0.185 \end{aligned}$$

We get the following confidence interval for $\beta_2 - \beta_3$ at level 0.90 :

$$0.66 - 0.25 \pm \left\{0.044^2 + 0.185^2 - 2 \cdot 0.044 \cdot 0.185 \cdot (-0.089)\right\}^{1/2} t_9(0.95) = [0.055, 0.765].$$

The R Commands

```
library(MASS)
fit<-lm(y~1+x1+x2+x3, data=cement)
confint(fit)
```

give the following as confidence intervals for each coordinate of β :

```
                2.5 %     97.5 %
(Intercept) 39.3411244 57.0461442
x1           1.2330935  2.1586869
x2           0.5568501  0.7569797
x3          -0.1678276  0.6678628
```

Assignment 4. a) The column "`t value`" gives the t-statistics for the null hypothesis $\beta_i = 0$. These are defined as

$$T_i = \frac{\hat{\beta}_i}{\sqrt{S^2 v_{ii}}} = \frac{\hat{\beta}_i}{\widehat{\mathrm{SE}}(\hat{\beta}_i)},$$

where v_{ii} is the i-th diagonal element of $(X^T X)^{-1}$. When the null hypothesis $\beta_i = 0$ is true, $T_i \sim t_{n-p}$ and we reject the hypothesis when the observed value of $|T_i|$ is large.

The column "`Pr(>|t|)`" gives the p-values for the bilateral t-tests of the preceding paragraph. Denote the observed value of T_i by $T_{i,\mathrm{obs}}$. The the p-value for the i-th test is

$$p_i = P(|T_i| > |T_{i,\mathrm{obs}}|) = 2(1 - F_{n-p}(|T_{i,\mathrm{obs}}|)) = 2F_{n-p}(-|T_{i,\mathrm{obs}}|)),$$

where F_{n-p} is the distribution function of a t_{n-p} random variable. If $p_i < 0.05$ we reject the i-th null hypothesis at significance level 5%. In this case, at 5%, we reject the null hypothesis $\beta_i = 0$ for $i = 0, 1, 2$, but not for $i = 3$.

b) In this case, the T statistic is

$$T = \frac{c^T \hat{\beta}}{\sqrt{S^2 c^T (X^T X)^{-1} c}}$$

for $c = [0, 0, 1, -1]^T$. Since

$$\begin{aligned} S^2 c^T (X^T X)^{-1} c &= \left(\widehat{\mathrm{SE}}(\hat{\beta}_2)\right)^2 + \left(\widehat{\mathrm{SE}}(\hat{\beta}_3)\right)^2 - 2\,\widehat{\mathrm{corr}}(\hat{\beta}_2, \hat{\beta}_3)\,\widehat{\mathrm{SE}}(\hat{\beta}_2)\,\widehat{\mathrm{SE}}(\hat{\beta}_3) \\ &= 0.04423^2 + 0.18471^2 - 2 \cdot (-0.08911) \cdot 0.04423 \cdot 0.18471 = 0.03753 \end{aligned}$$

we get that

$$T = \frac{0.65691 - 0.25002}{\sqrt{0.03753}} = 2.10033,$$

and find the corresponding p-value

$$p = 2 \cdot F_{13-4}(-2.10033) = 0.06508.$$

We cannot therefore reject the null hypothesis at 5%.

Assignment 5. (i). The design matrix X has dimension $n \times 2$, and $X^T = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix}$. The columns of X (rows of X^T) are dependent if and only if $x_1 = \cdots = x_n$ (which always happens when $n = 1$).

(ii). The covariance matrix we seek is $\sigma^2 A^{-1}$, with

$$A = X^T X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} = \begin{pmatrix} n & n\overline{x} \\ n\overline{x} & \sum x_i^2 \end{pmatrix}.$$

To find the inverse of A we look for a matrix $B = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ such that $BA = I_2$. This gives four linear equations in 4 variables that is easy to solve, and we get

$$\sigma^2(X^TX)^{-1} = \frac{\sigma^2}{nS_{xx}} \begin{pmatrix} \sum x_i^2 & -n\overline{x} \\ -n\overline{x} & n \end{pmatrix}, \qquad S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2.$$

Note that $S_{xx} = 0$ if and only if $x_1 = \cdots = x_n$ if and only if X^TX is not invertible.

(iii). The variance of $\widehat{\beta}_1$ is σ^2/S_{xx}, where σ^2 does not depend on x. Minimising this variance is equivalent to maximising S_{xx} over $[-1,1]^n$, and this can be done by elementary methods. For all j

$$\frac{\partial}{\partial x_j} S_{xx} = 2x_j - 2n\overline{x}\frac{\partial}{\partial x_j}\overline{x} = 2(x_j - \overline{x}) \quad \begin{cases} > 0 & x_j > \overline{x} \\ < 0 & x_j < \overline{x}. \end{cases}$$

The stationary point $x_j = \overline{x}$ gives a mimimum and not a maximum. It follows that the maximum is attained at the boundary ; $x_j = \pm 1$. This is true for all j. Thus we need to choose k of x_j's to be 1 and the others -1. For such choice $\sum x_i^2 = n$ (indpendently of k) and $\overline{x} = 2k/n - 1$. We need to minimise $\overline{x}^2$ with respect to k, which consequently must be as closed as possible to $n/2$. If n is odd, we choose $k = n/2$; otherwise $k = (n+1)/2$ or $(n-1)/2$.

Intuition. With this choice of x, we have very good estimators for $\mathbb{E}[y|x = 1]$ and $\mathbb{E}[y|x = -1]$, and we estimate $\mathbb{E}[Y|x = x]$ with the corresponding line. This gives the best estimator in terms of the variance of $\widehat{\beta}_1$. However, this choice leaves us with no information whatsoever on the distribution of y when $|x| < 1$. If linearity breaks down, we will not be able to detect it with the diagnostic tools !

Remark. One can argue alternatively that $S_{xx} : \mathbb{R}^n \to \mathbb{R}_+$ is convex. Indeed, it is a quadratic form with Hessian matrix

$$\begin{pmatrix} 2-2/n & -2/n & -2/n & \dots & -2/n \\ -2/n & 2-2/n & -2/n & \dots & -2/n \\ -2/n & -2/n & 2-2/n & \dots & -2/n \\ & & & \ddots & \\ -2/n & -2/n & \dots & -2/n & 2-2/n \end{pmatrix} = \left(2 - \frac{2}{n}\right) \begin{pmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & \rho & \dots & \rho \\ & & & \ddots & \\ \rho & \rho & \dots & \rho & 1 \end{pmatrix} = \left(2 - \frac{2}{n}\right) A,$$

where $\rho = -1/(n-1)$. Let us find the eigenvalues and eigenvectors of A. If $v = (v_1, \dots, v_n)$ then

$$Av = \begin{pmatrix} \rho \sum v_i + (1-\rho)v_1 \\ \vdots \\ \rho \sum v_i + (1-\rho)v_n \end{pmatrix} = \begin{cases} [1 + (n-1)\rho]v & v_1 = \cdots = v_n \\ (1-\rho)v & \sum v_i = 0. \end{cases}$$

We see that A has two eigenvalues : $1 + (n-1)\rho$ with multiplicity 1 and $1 - \rho$ with multiplicity $n-1$. This matrix is nonnegative definite if and only if $-1/(n-1) \le \rho \le 1$, which holds (just barely) in our setup $\rho = -1/(n-1)$. Thus S_{xx} is convex and the maximum is attained in the boundary.

Matrices of the form of A appear in statistics in *random effects models.*