Modern Regression Methods MATH EPFL
A.C. Davison

Solutions 1

Exercise 1.

a) The likelihood of (μ, σ^2) is

$$L(\mu,\sigma^2) = \prod_{i=1}^n p\left(y_i;\mu,\frac{\sigma^2}{w_i}\right) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2/w_i}} \exp\left\{-\frac{w_i}{2\sigma^2}(y_i-\mu)^2\right\}.$$

So, the log-likelihood is

$$\begin{aligned}\ell(\mu,\sigma^2) = \log L(\mu,\sigma^2) &= \sum_{i=1}^n \left\{-\frac{1}{2}\log\left(2\pi\frac{\sigma^2}{w_i}\right) - \frac{w_i}{2\sigma^2}(y_i-\mu)^2\right\} \\ &= -\frac{1}{2}\left\{n\log(2\pi\sigma^2) - \sum_{i=1}^n \log(w_i) + \frac{1}{\sigma^2}\sum_{i=1}^n w_i(y_i-\mu)^2\right\}.\end{aligned}$$

Thus we get,

$$\begin{aligned}\frac{\partial \ell}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^n w_i(y_i-\mu), \\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n w_i(y_i-\mu)^2.\end{aligned}$$

Annulation of the first partial derivative gives

$$\hat\mu = \frac{1}{\sum_{i=1}^n w_i}\sum_{i=1}^n w_i y_i = \bar y_w.$$

Annulation of the second partial derivative gives

$$-n + \frac{1}{\sigma^2}\sum_{i=1}^n w_i(y_i-\mu)^2 = 0 \tag{1}$$

and so

$$\hat\sigma^2 = \frac{1}{n}\sum_{i=1}^n w_i(y_i-\hat\mu)^2 = \frac{1}{n}\sum_{i=1}^n w_i(y_i-\bar y_w)^2.$$

Straightforward calculations gives that the Hessian matrix for $(\hat\mu,\hat w)$ is

$$H\big|_{(\mu,w)=(\hat\mu,\hat w)} = \begin{bmatrix} -\frac{\sum_{i=1}^n w_i}{\hat\sigma^2} & 0 \\ 0 & -\frac{n}{2\hat\sigma^2}\end{bmatrix}.$$

This matrix is negative definite and so, $(\hat\mu,\hat w) = (\hat\mu,\hat\sigma^2)$ is a maximum.

b) The likelihood of (β,σ^2) is

$$L(\beta,\sigma^2) = \prod_{i=1}^n p\left(Y_i;X_i^T\beta,\frac{\sigma^2}{w_i}\right) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2/w_i}}\exp\left\{-\frac{w_i}{2\sigma^2}(Y_i - X_i^T\beta)^2\right\},$$

where X_i is the i-th row of X. So, the log-likelihood is

$$\begin{aligned}
\ell(\mu,\sigma^2) = \log L(\mu,\sigma^2) &= \sum_{i=1}^{n}\left\{-\frac{1}{2}\log\left(2\pi\frac{\sigma^2}{w_i}\right) - \frac{w_i}{2\sigma^2}(y_i - X_i^T\beta)^2\right\}\\
&= \sum_{i=1}^{n}\left\{-\frac{1}{2}\log\left(2\pi\frac{\sigma^2}{w_i}\right)\right\} - \frac{1}{2\sigma^2}\{Y' - X'\beta\}^T\{Y' - X'\beta\}\\
&= -\frac{1}{2}\left\{n\log(2\pi\sigma^2) - \sum_{i=1}^{n}\log(w_i)\right\} - \frac{1}{2\sigma^2}\{Y' - X'\beta\}^T\{Y' - X'\beta\},
\end{aligned}$$

where $Y' = \sqrt{W}Y$ and $X' = \sqrt{W}X$. Taking partial derivative with respect to β gives

$$\frac{\partial\ell}{\partial\beta} = \frac{(X')^T}{\sigma^2}\left[\{Y' - X'\beta\}\right]$$

Thus, annulation of the previous equation leads to

$$\widehat{\beta} = \left\{(X')^T X'\right\}^{-1}(X')^T Y' = \left(X^T W X\right)^{-1} X^T W Y.$$

Thus, with the modified variables Y' and X', we have a classical regression setting, so

$$\widehat{\beta} \sim \mathcal{N}\left[\beta, \sigma^2\{(X')^T X'\}^{-1}\right] = \mathcal{N}\left[\beta, \sigma^2\{(X)^T W X\}^{-1}\right],$$

which gives the variance and unbiasedness of the estimator.

c) Partial derivative with regards to σ^2 of the log-likelihood derived in question b) with modified variables Y' and X' gives

$$\frac{\partial\ell}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\{Y' - X'\beta\}^T\{Y' - X'\beta\}.$$

and its annulation leads to

$$\sigma^2 = \frac{1}{n}\{Y' - X'\beta\}^T\{Y' - X'\beta\}.$$

Then the estimator $\widehat{\sigma}^2$ is

$$\widehat{\sigma}^2 = \frac{1}{n}\{Y' - X'\hat{\beta}\}^T\{Y' - X'\widehat{\beta}\}.$$

However, $\widehat{\sigma}^2$ is a biased estimator of σ^2 because $\{Y' - X'\hat{\beta}\}^T\{Y' - X'\widehat{\beta}\}/\sigma^2 \sim \chi_{n-p}$, so an unbiased estimator of σ^2 is

$$\begin{aligned}
s^2 &= \frac{1}{n-p}\{Y' - X'\hat{\beta}\}^T\{Y' - X'\widehat{\beta}\}, \tag{2}\\
&= \frac{1}{n-p}\left[Y' - X'\left\{(X')^T X'\right\}^{-1}(X')^T Y'\right]^T\left[Y' - X'\left\{(X')^T X'\right\}^{-1}(X')^T Y'\right], \tag{3}\\
&= \frac{1}{n-p}(Y')^T\left[1 - X'\left\{(X')^T X'\right\}^{-1}(X')^T\right]^T\left[1 - X'\left\{(X')^T X'\right\}^{-1}(X')^T\right]Y', \tag{4}\\
&= \frac{1}{n-p}(Y')^T\{1 - X'\left\{(X')^T X'\right\}^{-1}(X')^T\}Y', \tag{5}\\
&= \frac{1}{n-p}Y^T\{W - WX\left\{X^T W X\right\}^{-1}X^T W\}Y. \tag{6}
\end{aligned}$$

The simplification between equation (4) and (5) is a consequence of the idempotence of $(X')^T\left\{(X')^T X'\right\}^{-1}(X')^T$, as a projection matrix. We thus get an unbiased estimator of σ^2.

Exercise 2.

a) The column "`t value`" gives the t-statistics for the null hypothesis $\beta_i = 0$, defined by

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

where v_{ii} is the i-th diagonal element of the matrix $V = (X^{\mathrm{T}}X)^{-1}$. When the hypothesis $\beta_i = 0$ is verified, T_i follows a Student-t distribution with $n - p$ degrees of freedom. The null hypothesis $\beta_i = 0$ is rejected for high values of $|T_i|$.

The column "`Pr(>|t|)`" gives the p-values for bilateral t tests. For an observed statistic $T_{i,\mathrm{obs}}$, the p-value 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 for the law t_{n-p}. If $p_i < 0.05$, we reject i-th null-hypothesis with a 5% significance level. For this example, with a 5% significance level, we can reject the null-hypothesis $\beta_i = 0$ for $i = 0, 1, 2$, but nor for $i = 3$.

b) The test statistic T is

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

for $c = [0, 0, 1, -1]^{\mathrm{T}}$. Following the reminder,

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

So

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

and the corresponding p-value is

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

With a 5% significance level, we cannot reject the null-hypothesis.

Exercise 3.

a) We use the F-test given in the exercice to compare the models. Its critical value for a 5% confidance level is 5.32.

Forward selection For each step, we consider adding the variable inducing the highest RSS reduction.

— Initial model : $y = \beta_0 + \epsilon$
— Step 1 : $y = \beta_0 + \beta_4 x_4 + \epsilon$, $F = \frac{2715.8-883.9}{47.9/(13-5)} = 305.95 > 5.32$.
— Step 2 : $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \epsilon$, $F = 135.13 > 5.32$.
— Step 3 : $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, $F = 4.47 < 5.32$.
Final model : $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \epsilon$.

Backward selection For each step, we consider removing the variable inducing the lowest RSS increase.

— Initial model : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$
— Step 1 : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \epsilon$, $F = \frac{48-47.9}{47.9/(13-5)} = 0.0167 < 5.32$.
— Step 2 : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, $F = 1.65 < 5.32$.
— Step 3 : $y = \beta_0 + \beta_2 x_2 + \epsilon$, $F = 141.70 > 5.32$.
Final model : $y = \beta_0 + \beta_2 x_2 + \beta_1 x_1 + \epsilon$.

b) i) Mallows C_p is similar to AIC : the model with lowest C_p is the preferred model. To compute the missing C_p, we need to know s^2, which can be found using any known C_p or we also have

$$s^2 = \frac{\|e_{\text{full}}\|^2}{n-p} = \frac{\text{RSS}_{\text{full}}}{13-5} = \frac{47.9}{8} = 5.99.$$

The completed table is then

Model	RSS	C_p	Model	RSS	C_p	Model	RSS	C_p
- - - -	2715.8	442.58	1 2 - -	57.9	2.67	1 2 3 -	48.1	3.03
			1 - 3 -	1227.1	197.94	1 2 - 4	48.0	3.02
1 - - -	1265.7	202.39	1 - - 4	74.8	5.49	1 - 3 4	50.8	3.48
- 2 - -	906.3	142.37	- 2 3 -	415.4	62.38	- 2 3 4	73.8	7.325
- - 3 -	1939.4	314.90	- 2 - 4	868.9	138.12			
- - - 4	883.9	138.62	- - 3 4	175.7	22.34	1 2 3 4	47.9	5

ii) Forward selection leads to $y = \beta_0 + \sum_{i\in\{1,2,4\}} \beta_i x_i$, whereas with backward selection, we get $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. The backward selection model has the lowest C_p, so it is the best overall model.

Exercise 4.

a) Following the hypothesis of the model, the residuals e and $\widehat{y}$ are independent and the standardized residuals follows a centered Gaussian distribution with unit variance (in practice, it means residuals values are between -2 and 2 and independent of $\hat{y}_j$).

— Plot A : the model seems reasonable.

— Plot B : there is an outlier with a $r_j < -2$. We should check if this is not the result of a systematic measurement error.

— Plot C : fitted values and standardized residuals seems dependent. The addition of a quadratic term would probably be an efficient solution to fix to this problem.

— Plot D : here, homoscedasticity is not verified, because the residuals present a varying variance. The solution here would be tu use a weighted least square (see Exercice 1).

b) If the data distribution has a lower tail (left side of the distribution) heavier than the normal law, the empirical quantiles on the lower left part of the quantile-quantile plot will be under the diagonal $y = x$. Indeed in this case, $G(x) \gg F(x)$ when $x \to -\infty$, where G is the empirical distribution function and F the distribution of the Gaussian law. For small $\alpha > 0$, if $x = F^{-1}(\alpha)$, then $G(x) \gg \alpha$, and so $G^{-1}(\alpha) < x$, which means that the points will be under the line $y = x$. On the contrary, if the lower tail is lighter than the normal distribution, then empirical quantiles will be above the diagonal. Similar considerations allows to conclude for the behaviour of the upper (right) tail.

— Plot A : heavy lower tail and light upper tail : there is a negative skewness parameter.

— Plot B : tails are lighter than the Gaussian law.

— Plot C : tails are heavier than the Gaussian law.

— Plot D : light lower tail and heavy upper tail : there is a positive skewness parameter.

Exercise 5.

a) I. $X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$, $\beta = (\beta_0, \alpha_1, \alpha_2)^{\mathrm{T}}$.

II. We note that $X(1,-1,-1)^{\mathrm{T}} = 0$ then X is not injective. The consequence is that we cannot inverse the matrix $X^{\mathrm{T}}X$; statistically, it means that the three parameters cannot be estimated jointly, i.e., the model is not *identifiable*.

III. $X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$.

Suppressing the column for α_1 is equivalent to set $\alpha_1 = 0$. β_0 is the mean of the observations in the group $a_j =$ "1" and α_2 is the difference between the means of the groups $a_j =$ "1" and $a_j =$ "2".

IV. For the model $y \sim a$, $X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$

For the model $y \sim a + b$, $X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}$

For the model $y \sim x + a - 1$, $X = \begin{pmatrix} 152 & 1 & 0 \\ 93 & 1 & 0 \\ 127 & 1 & 0 \\ 109 & 0 & 1 \\ 141 & 0 & 1 \\ 136 & 0 & 1 \end{pmatrix}$.

For the model $y \sim b + x - 1$, $X = \begin{pmatrix} 1 & 0 & 0 & 152 \\ 0 & 1 & 0 & 93 \\ 0 & 0 & 1 & 127 \\ 1 & 0 & 0 & 109 \\ 0 & 1 & 0 & 141 \\ 0 & 0 & 1 & 136 \end{pmatrix}$.

b) For the model $y \sim a : x$,

$X = \begin{pmatrix} 1 & 152 & 0 \\ 1 & 93 & 0 \\ 1 & 127 & 0 \\ 1 & 0 & 109 \\ 1 & 0 & 141 \\ 1 & 0 & 136 \end{pmatrix}$; the columns are linearly independent.

For the model $y \sim a : b$, $X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$; the columns are not linearly independent.

For the model $y \sim a{+}b : x$, $X = \begin{pmatrix} 1 & 0 & 152 & 0 & 0 \\ 1 & 0 & 0 & 93 & 0 \\ 1 & 0 & 0 & 0 & 127 \\ 1 & 1 & 109 & 0 & 0 \\ 1 & 1 & 0 & 141 & 0 \\ 1 & 1 & 0 & 0 & 136 \end{pmatrix}$; the columns are linearly independent.

Pour le modèle $y \sim a + a : b : x$, $X = \begin{pmatrix} 1 & 0 & 152 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 93 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 127 & 0 \\ 1 & 1 & 0 & 109 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 141 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 136 \end{pmatrix}$; the columns are not linearly independent.