Solutions to selected exercises from Chapter 13 of *Wasserman — All of Statistics*

(1) Write $\mathbf{X}$ for the vector of X_i's, $\mathbf{1}$ for the vector of 1's of length n, $\hat{\mathbf{X}}$ for the matrix $\begin{pmatrix}\mathbf{1} & \mathbf{X}\end{pmatrix}$, $\mathbf{Y}$ for the vector of Y_i's, and $\beta = (\beta_0, \beta_1)^T$. We are trying to minimize the quantity

$$f(\beta_0,\beta_1) = \|\mathbf{Y}-\hat{\mathbf{X}}\beta\|^2 = (\mathbf{Y}-\hat{\mathbf{X}}\beta)^T(\mathbf{Y}-\hat{\mathbf{X}}\beta) = \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\hat{\mathbf{X}}\beta - \beta^T\hat{\mathbf{X}}^T\mathbf{Y} + \beta^T\hat{\mathbf{X}}^T\hat{\mathbf{X}}\beta.$$

The function f of β_0 and β_1 is convex and smooth with a single local minimum, which is the global minimum. The term $\mathbf{Y}^T\hat{\mathbf{X}}\beta$ is equal to

$$Y_1(\beta_0+\beta_1X_1)+\ldots+Y_n(\beta_0+\beta_1X_n).$$

Hence its derivative with respect to β_0 is $\mathbf{Y}\cdot\mathbf{1}$ and its derivative with respect to β_1 is $\mathbf{Y}\cdot\mathbf{X}$. These turn out to be the same as the derivatives for the term $\beta^T\hat{\mathbf{X}}^T\mathbf{Y}$. Finally, the term $\beta^T\hat{\mathbf{X}}^T\hat{\mathbf{X}}\beta$ is equal to

$$n\beta_0^2 + 2(\mathbf{1}\cdot\mathbf{X})\beta_0\beta_1 + (\mathbf{X}\cdot\mathbf{X})\beta_1^2.$$

The derivatives of this term with respect to β_0 and β_1 are $2n\beta_0 + 2(\mathbf{1}\cdot\mathbf{X})\beta_1$ and $2(\mathbf{1}\cdot\mathbf{X})\beta_0 + 2(\mathbf{X}\cdot\mathbf{X})\beta_1$.

Finally, setting the derivatives with respect to β_0 and β_1 equal to zero yields

$$n\beta_0 + (\mathbf{X}\cdot\mathbf{1})\beta_1 = \mathbf{Y}\cdot\mathbf{1} \text{ and } (\mathbf{1}\cdot\mathbf{X})\beta_0 + (\mathbf{X}\cdot\mathbf{X})\beta_1 = \mathbf{Y}\cdot\mathbf{X}.$$

In matrix form this is

$$\begin{pmatrix}\mathbf{1}\cdot\mathbf{1} & \mathbf{X}\cdot\mathbf{1}\\ \mathbf{1}\cdot\mathbf{X} & \mathbf{X}\cdot\mathbf{X}\end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} = \begin{pmatrix}\mathbf{Y}\cdot\mathbf{1}\\ \mathbf{Y}\cdot\mathbf{X}\end{pmatrix}.$$

Set $D = n(\mathbf{X}\cdot\mathbf{X}) - (\mathbf{X}\cdot\mathbf{1})^2$ as the determinant of the matrix on the left. Verify that

$$D = n\sum X_i^2 - n^2\left(\sum X_i\right)^2 = n\sum(X_i-\bar{X}_n)^2.$$

Then we have

$$\beta_1 = \frac{1}{D}((-\mathbf{X}\cdot 1)(\mathbf{Y}\cdot\mathbf{1}) + (\mathbf{1}\cdot\mathbf{1})(\mathbf{Y}\cdot\mathbf{X})) = \frac{1}{D}\left(-n^2\bar{X}_n\bar{Y}_n + n\sum X_iY_i\right)$$

which is equal to $\frac{n}{D}\sum(X_i-\bar{X}_n)(Y_i-\bar{Y}_n)$. Canceling a factor of n in the numerator and denominator then yields the desired expression for $\hat{\beta}_1$.

The first equation above, $n\beta_0 + (\mathbf{X}\cdot\mathbf{1})\beta_1 = \mathbf{Y}\cdot\mathbf{1}$ yields $n\beta_0 + n\bar{\mathbf{X}}\beta_1 = n\bar{\mathbf{Y}}$, as desired.

(2) We have

$$\hat{\beta}_1 = \frac{(\mathbf{X}-\overline{X_n}\mathbf{1})^T(\mathbf{Y}-\overline{Y_n}\mathbf{1})}{(\mathbf{X}-\overline{X_n}\mathbf{1})^T(\mathbf{X}-\overline{X_n}\mathbf{1})}$$

where $\mathbf{1}$ is the vector of all 1's, $\mathbf{X} = (X_1,\ldots,X_n)^T$, and $\mathbf{Y} = (Y_1,\ldots,Y_n)^T$. Hence, $\mathbb{V}(\hat{\beta}_1)$ is equal to

$$\frac{1}{(\mathbf{X}-\overline{X_n}\mathbf{1})^T(\mathbf{X}-\overline{X_n}\mathbf{1})}(\mathbf{X}-\overline{X_n}\mathbf{1})^T\Sigma(\mathbf{X}-\overline{X_n}\mathbf{1})$$

where Σ is the covariance matrix of $\mathbf{Y} - \overline{Y_n}\mathbf{1}$. Let's compute Σ.
We have $\mathbb{V}(Y_i - \overline{Y_n}) = \mathbb{V}(Y_i) + \mathbb{V}(\overline{Y_n}) - 2\operatorname{Cov}(Y_i, \overline{Y_n})$. We have $\mathbb{V}(Y_i) = \mathbb{V}(\beta_0 + \beta_1 X_i + \epsilon_i) = \mathbb{V}(\epsilon_i) = \sigma^2$. Similarly, $\mathbb{V}(\overline{Y_n}) = \frac{\sigma^2}{n}$. Finally, $\operatorname{Cov}(Y_i, Y_j) = \operatorname{Cov}(\epsilon_i, \epsilon_j) = \sigma^2 \delta_{ij}$ where δ_{ij} is the Dirac delta: $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. By bilinearity of Cov,

$$\operatorname{Cov}(Y_i, \overline{Y_n}) = \frac{1}{n} \sum_{j=1}^{n} \operatorname{Cov}(Y_i, Y_j) = \frac{\sigma^2}{n}.$$

Hence

$$\mathbb{V}(Y_i - \overline{Y_n}) = \sigma^2 \left(1 + \frac{1}{n} - 2\frac{1}{n}\right) = \frac{(n-1)\sigma^2}{n}.$$

For $i \neq j$ we have

$$\operatorname{Cov}(Y_i - \overline{Y_n}, Y_j - \overline{Y_n}) = \operatorname{Cov}(Y_i, Y_j) - \operatorname{Cov}(Y_i, \overline{Y_n}) - \operatorname{Cov}(Y_j, \overline{Y_n}) + \mathbb{V}(\overline{Y_n}).$$

By our previous calculations this is equal to $-\sigma^2/n$.
Finally then,

$$\Sigma = \frac{\sigma^2}{n} \begin{pmatrix} n-1 & -1 & -1 & -1 \\ -1 & n-1 & -1 & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \dots & n-1 \end{pmatrix}$$

and this yields

$$\mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{n} \cdot \frac{1}{\left(\sum (X_i - \overline{X_n})^2\right)^2} \cdot \left((n-1) \sum_{i=1}^{n} (X_i - \overline{X_n})^2 - \sum_{i \neq j} (X_i - \overline{X_n})(X_j - \overline{X_n}) \right).$$

We have that

$$-\sum_{i=1}^{n} (X_i - \overline{X_n})^2 - \sum_{i \neq j} (X_i - \overline{X_n})(X_j - \overline{X_n}) = -\left(\sum_{i=1}^{n} (X_i - \overline{X_n}) \right)^2 = 0.$$

So $\mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{\sum (X_i - \overline{X_n})^2} = \frac{\sigma^2}{n s_X^2}$, as claimed.
From $\hat{\beta}_0 = \overline{Y_n} - \hat{\beta}_1 \overline{X_n}$, we see

$$\operatorname{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \operatorname{Cov}(\overline{Y_n}, \hat{\beta}_1) - \overline{X_n} \operatorname{Cov}(\hat{\beta}_1, \hat{\beta}_1).$$

By our previous calculations,

$$\operatorname{Cov}(\overline{Y_n}, \hat{\beta}_1) = \frac{1}{n s_X^2} \sum (X_i - \overline{X_n}) \operatorname{Cov}(\overline{Y_n}, Y_i - \overline{Y_n}) = \frac{1}{n s_X^2} \sum (X_i - \overline{X_n}) \left(\frac{\sigma^2}{n} - \frac{\sigma^2}{n} \right) = 0$$

and this leaves $\operatorname{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\overline{X_n}\mathbb{V}(\hat{\beta}_1)$, as claimed.
Finally,

$$\mathbb{V}(\hat{\beta}_0) = \mathbb{V}(\overline{Y_n}) - 2\overline{X_n} \operatorname{Cov}(\overline{Y_n}, \hat{\beta}_1) + \overline{X_n}^2 \mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{n} + \frac{\sigma^2 \overline{X_n}^2}{n s_X^2} = \frac{\sigma^2 (s_X^2 + \overline{X_n}^2)}{n s_X^2}$$

and we calculate that $s_X^2 + \overline{X_n}^2 = \frac{1}{n} \sum X_i^2$. This gives the $(1,1)$ entry of the covariance matrix.

(3) We assume $Y_i = \beta X_i + \epsilon_i$ where the ϵ_i are iid with variance σ^2. Our model is $\hat{r}(X) = \hat{\beta} X$. For data points $(Y_1, X_1), \ldots, (Y_n, X_n)$, the sum of squared residuals is $\sum \hat{\epsilon}_i^2 = \sum (Y_i - \hat{\beta} X_i)^2$. This function of $\hat{\beta}$ is convex with a single local minimum. We have

$$\frac{\partial}{\partial \hat{\beta}} \left(\sum \hat{\epsilon}_i^2 \right) = 2 \sum \left(-X_i Y_i + \hat{\beta} X_i^2 \right).$$

Setting this equal to zero and solving for $\hat{\beta}$ yields

$$\hat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{\mathbf{X} \cdot \mathbf{Y}}{\mathbf{X} \cdot \mathbf{X}}.$$

where $\mathbf{X} = (X_1, \ldots, X_n)^T$ and similarly for $\mathbf{Y}$. Considering $\mathbf{X}$ to be constant, $\mathbb{V}(\hat{\beta}) = \frac{1}{(\mathbf{X} \cdot \mathbf{X})^2} (\mathbf{X})^T \Sigma (\mathbf{X})$ where Σ is the covariance matrix of $\mathbf{Y}$. As in the previous exercise, $\mathrm{Cov}(Y_i, Y_j) = \sigma^2 \delta_{ij}$ where δ_{ij} is the Dirac delta. So $\Sigma = \sigma^2 I$ where I is the $n \times n$ identity matrix and

$$\mathbb{V}(\hat{\beta}) = \frac{\sigma^2 \mathbf{X} \cdot \mathbf{X}}{(\mathbf{X} \cdot \mathbf{X})^2} = \frac{\sigma^2}{\sum X_i^2}.$$

As long as $\sum X_i^2 \to \infty$ we have $\mathbb{E}(\hat{\beta}) \to \beta$ and $\mathbb{V}(\hat{\beta}) \to 0$. Thus $\hat{\beta}$ converges to β in quadratic mean and therefore also in probability (see e.g. exercise 5.2).

(4) .....

(6) See the Jupyter Notebook 6.ipynb.

(7) See the Jupyter Notebook 7.ipynb.

(8) In this case, up to addition of a constant not depending on $\hat{\beta}$, the AIC is

$$\mathrm{AIC} = \ell_S - |S| = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - (X\hat{\beta})_i)^2 - |S|.$$

Mallow's C_p statistic is

$$\hat{R}_{\mathrm{tr}}(S) + 2|S|\sigma^2 = \sum_{i=1}^{n} (Y_i - (X\hat{\beta})_i)^2 + 2|S|\sigma^2.$$

Thus, up to adding a constant to AIC, which doesn't affect where the maximum AIC is achieved, we have $C_p = -2\sigma^2 \mathrm{AIC}$. So the $\hat{\beta}$ which maximizes AIC is the same as the $\hat{\beta}$ which minimizes C_p.

(11) See the Jupyter Notebook 11.ipynb.