Coursework 2

MATH95012 – Statistical Modelling 1
Winter 2020

Instructions

To be handed in no later than Tuesday, 10th March 2020, 2pm.
Please hand in to the Mathematics Undergraduate Office.

- Some parts of this coursework will involve numerical computations. For this you can use any computer program of your choice (e.g. R, Matlab, Python). Do NOT use special functions for linear models (like `lm` in R). You may, however, use built-in functions for computing matrix inverses (like `solve` in R).
- **Regardless of which program you use, you must include a printout of your code** (ideally with the outputs between the commands).
- Your solution should not be longer than 5 sheets of paper (one sheet includes front and back).
- There are 20 marks possible on this assignment. The points possible for each sub-question are included in brackets (e.g., [x marks]) at the end of that sub-question.

Coursework 2

MATH95012 – Statistical Modelling 1
Winter 2020

Questions

In the first question of this coursework you will consider least squares estimation for the linear regression models

$$E(Y_i) = \beta_0 + x_i\beta_X \tag{1}$$
$$E(Y_i) = \gamma_0 + w_i\gamma_W + x_i\gamma_X \tag{2}$$

where $i = 1, \ldots, n$. For this question, you may assume the full rank (FR) and second order assumptions (SOA).

- For any $\mathbf{a} = (a_1, \ldots, a_n)^T \in \mathbb{R}^n$, define

$$\bar{a} := \frac{1}{n}\sum_{i=1}^{n} a_i \qquad \text{and} \qquad \bar{\mathbf{a}} := \bar{a}\mathbf{1}$$

  where $\mathbf{1} \in \mathbb{R}^n$ is a vector of ones.

- For any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, define $S_{ab} := (\mathbf{a} - \bar{\mathbf{a}})^T(\mathbf{b} - \bar{\mathbf{b}}) = \sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})$.

1. This question pertains to models (1) and (2) above.

(a) Show that the least squares estimator for model (2) is $\hat{\gamma} = (\hat{\gamma}_0, \hat{\gamma}_W, \hat{\gamma}_X)^T$ with

$$\begin{aligned}
\hat{\gamma}_0 &= \bar{Y} - \bar{w}\hat{\gamma}_W - \bar{x}\hat{\gamma}_X \\
\hat{\gamma}_W &= \frac{S_{wy}S_{xx} - S_{wx}S_{xy}}{S_{xx}S_{ww} - S_{xw}^2} \\
\hat{\gamma}_X &= \frac{S_{xy}S_{ww} - S_{wx}S_{wy}}{S_{xx}S_{ww} - S_{xw}^2}\,.
\end{aligned}$$

[6 marks]

(b) Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_X)^T$ be the least squares estimators from model (1). Use the results of 1(a) to establish a set of conditions under which

$$\hat{\gamma}_X = \hat{\beta}_X.$$

[2 marks]

(c) Compare the variances of $\hat{\gamma}_X$ in model (2) and $\hat{\beta}_X$ in model (1), under the condition you found in 1(b). Under what conditions is $\text{var}(\hat{\gamma}_X) < \text{var}(\hat{\beta}_X)$? Similarly, can we ever have $\text{var}(\hat{\gamma}_X) > \text{var}(\hat{\beta}_X)$? [2 marks]

2. This question pertains to the dataset contained in *fev.csv* and documented in *fev.doc*. Both files are available on the Blackboard page. The following is an excerpt of the documention:

> It is now widely believed that smoking tends to impair lung function. Much of the data to support this claim arises from studies of pulmonary function in adults who are long-time smokers. A question then arises whether such deleterious effects of smoking can be detected in children who smoke. To address this question, measures of lung function were made in 654 children seen for a routine check up in a particular pediatric clinic. The children participating in this study were asked whether they were current smokers.
>
> A common measurement of lung function is the forced expiratory volume (FEV), which measures how much air you can blow out of your lungs in a short period of time. A higher FEV is usually associated with better respiratory function.

The variables of interest are contained in the columns:

- `fev` – measured FEV (liters per second)
- `smoke` – smoking habits (1 = yes, 2 = no)
- `age` – subject age at time of measurement (years)

(a) Compute the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ based on the model

$$E(\texttt{fev}_i) = \beta_0 + \texttt{smoke}_i\beta_1 \,.$$

Use your results to estimate the difference in means between smokers and non-smokers

$$E(\texttt{fev} \mid \texttt{smoke==1}) - E(\texttt{fev} \mid \texttt{smoke==2}).$$

Based on these estimates, does smoking appear to impair lung function in children? [3 marks]

(b) Compute the least squares estimates $\hat{\gamma}_0$ and $\hat{\gamma}_1$ based on the model

$$E(\texttt{fev}_i) = \gamma_0 + \texttt{smoke}_i\gamma_1 + \texttt{age}_i\gamma_2 \,.$$

Use your results to estimate the difference in means between smokers and non-smokers, keeping age fixed

$$E(\texttt{fev} \mid \texttt{smoke==1}, \texttt{age}) - E(\texttt{fev} \mid \texttt{smoke==2}, \texttt{age}).$$

Based on these estimates, does smoking appear to impair lung function in children? Explain in one sentence how your condition from 1(c) applies to this problem. [3 marks]

(c) Assuming the model in 2(a) holds with the normal theory assumptions (NTA), report a 95% confidence interval for β_1 and conduct a hypothesis test of $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. Based on these estimates, does smoking appear to impair lung function in children? [2 marks]

(d) Assuming the model in 2(b) holds with NTA, report a 95% confidence interval for γ_1 and conduct a hypothesis test of $H_0 : \gamma_1 = 0$ against $H_1 : \gamma_1 \neq 0$. Based on these estimates, does smoking appear to impair lung function in children? [2 marks]