Linear Statistical Models

2014-04-25

Question 1

a) Under the assumption that $\beta_1 = \beta_2$ we can rewrite the model as:

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} \\ &= \alpha + \beta_1 x_{i1} + \beta_1 x_{i2} \\ &= \alpha + \beta_1 (x_{i1} + x_{i2}) \\ &= \alpha + \beta_1 z_i \end{aligned}$$

where $z_i = x_{i1} + x_{i2}$. Use the usual formula for the least squares in terms of (y_i, z_i)

b) Similarly to a) we can write

$$\begin{aligned} y_i &= \alpha + \beta_1 x_{i1} + \beta_2 x_{i1} \\ &= \alpha + 2\beta_2 x_{i1} + \beta_2 x_{i2} \\ &= a + \beta_2 (2x_{i1} + x_{i2}) \\ &= \alpha + \beta_2 z_i \end{aligned}$$

where $z_i = 2x_{i1} + x_{i2}$. Use again the usual formula for the least squares, in term of (y_i, z_i)

c) This is a minimization problem with inequality constraints. We introduce the constraint into the least squares functions using Lagrange multipliers as follows

$$Q(\alpha, \beta_1, \beta_2) = \sum_{i=1}^{n} [y_i - (\alpha + \beta_1 x_{i1} + \beta_2 x_{i2})]^2 + \lambda(\beta_1 - \beta_2)$$

The partial derivatives w.r.t. β_1 and β_2 are

$$\frac{\partial Q}{\partial \beta_1} = \sum_{i=1}^{n} -2x_{i1}[y_i - (\alpha + \beta_1 x_{i1} + \beta_2 x_{i2})] + \lambda$$

$$\frac{\partial Q}{\partial \beta_2} = \sum_{i=1}^{n} -2x_{i2}[y_i - (\alpha + \beta_1 x_{i1} + \beta_2 x_{i2})] - \lambda$$

We can't write down the solution explicitly, but we can write it in terms of λ

d) For a) and b) we can just test that $\beta_1 - \beta_2 = 0$ and $\beta_1 - 2\beta_2 = 0$, respectively, using a t-test, since the sum of t distributed random variables is t. For c) we can do a one-sided t-test for $\beta_1 - \beta_2 > 0$

Question 2

a) No, it's not correct since that model doesn't gurantee that θ_i is between 0 and 1.

b) A minimal requirement is that $0 \leq F \leq 1$. The logistic function in c) is one such function.

c) The odds are equal to $e^{\beta_1} = e^{0.1} = 1.1052$. This means that the odds of not surviving the procedure increase, in average, about 11% each year (regarding the age of the patient).

d) $L(\theta_i) = \prod_{i=1}^{n} \theta_i^{y_i} (1-\theta_i)^{y_i}$

The log-likelihood is

$$l(\theta_i) = \log(L(\theta_i)) = \sum_{i=1}^{n} y_i \log(\theta_i) + (1-y_i)\log(1-\theta_i)$$

$$= \sum_{i=1}^{n} y_i \log\left(\frac{\theta_i}{1-\theta_i}\right) + \log(1-\theta_i)$$

Question 3

a) Hard to make coments about overall fit, but the estimates for each parameter don't have particularly high significance.

b) It just affects β_0 (intercept), but doesn't change the magnitude or meaning of β_1 (age). It reduces the correlation between β_0 and β_1

c) It doesn't make much sense to add age^2 to the model from a modeling point of view. Again, hard to tell about overall fit but the significance of the estimates decreases in comparison to models 1 and 2.

d) Can use bootstrap. Compute the estimates of β_1 and β_2 for a number of bootstrap samples and then compute the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$.

Question 4

a) The z-scores are defined as $z_j = \dfrac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$, where v_j is the j-th diagonal element of $(X^TX)^{-1}$ and is distributed $t_{(n-p-1)}$. A z-score larger than 2 in absolute value is approximately significant (non-zero) at the 5% level. lcavol, lweight and svi are the most significant parameters

b) For a description look at the course notes. In general, different methods pick different models. Which one to choose depends on the goals of the modeling.

Question 5

a) $\tilde{\beta}$ is an unbiased estimator of β if and only if $E[\tilde{\beta}]=\beta$
We start by computing its expected value then.

$$E[\tilde{\beta}] = E\{[(X^TX)^{-1}X^T+D]y\} = E\{[(X^TX)^{-1}X^T+D]y\,(X\beta+\varepsilon)\}$$
$$= E\{[(X^TX)^{-1}X^T+D]X\beta\} + E\{[(X^TX)^{-1}X+D]\varepsilon\}$$
$$= [(X^TX)^{-1}X^T+D]X\beta + [(X^TX)^{-1}X+D]\underbrace{E[\varepsilon]}_{0}$$
$$= (X^TX)^{-1}X^TX\beta + DX\beta$$
$$= [I+DX]\beta \quad \text{which is equal to } \beta \iff DX=0$$

b) We compute the variance of $\tilde{\beta}$. Let $C=(X^TX)^{-1}X^T+D$, then
$$Var(\tilde{\beta}) = Var(Cy) = C\,Var(y)\,C^T = \sigma^2CC^T$$
$$= \sigma^2[(X^TX)^{-1}X^T+D][(X^TX)^{-1}X^T+D]^T$$
$$= \sigma^2[(X^TX)^{-1}X^T+D][X(X^TX)^{-1}+D^T]$$
$$= \sigma^2[(X^TX)^{-1}X^TX(X^TX)^{-1} + (X^TX)^{-1}X^TD^T$$ [redacted] $$+ DX(X^TX)^{-1} + DD^T$$
$$= \sigma^2[(X^TX)^{-1} + X^TX\underbrace{(DX)^T}_{0} + \underbrace{DX}_{0}(X^TX)^{-1} + DD^T]$$
$$= \sigma^2(X^TX)^{-1} + \sigma^2DD^T$$
$$= Var(\tilde{\beta}) + \sigma^2DD^T > Var(\tilde{\beta})$$

Question 6

a) Yes, continuous variables seem to require transformation

b) Inflation has the strongest effect. The effects by country vary.

c) some outliers need to be removed, the normality assumption may be violated.

Question 7

a) Not for the left-hand-side panel, but for the right-hand-side the regression line will divide the observations in two groups (it will be an horizontal line at about $y=1$)

b) Look at the regression line and classify observation above it as belonging to one group and observations below it as belonging to the other group. To check the classification rule, the regression line could be computed on a training set and tested on a test set (divide the data into train and test before performing the regression).

c) There is a number of classification methods, most of them better since regression itself may not work always (as for the case in the left-hand-side panel).