Predictive Modeling   Homework 3. Multiple Linear Regression   Due: Feb 8, 2024

1. **Likelihood Function for mean of normal distribution** Three automatic productions of bottled liquid are considered to be stable. A sample bottle was selected randomly from each productions and the volume of the content was measured. The deviation from the nominal volumes of $700.0, 750.0, 850.0$ ml was recorded respectively. The deviations (in ml) were $4.0, 6.5, 5.0$.

   The *model* is formulated

   1. Data: $Y_i = \mu_i + \epsilon_i$
   2. Assumptions:
      - Y_1, Y_2, Y_3 are independent
      - $Y_i \sim N(\mu_i, \sigma^2)$

   (a) What is the joint probability density function for Y_1, Y_2, Y_3?

   (b) What is the likelihood and log-likelihood functions?

   (c) What is the score function, observed and expected information respectively? and What are their corresponding expression for the observations of, $y_1 = 4.0, y_2 = 6.5, y_3 = 5.0$?

2. Let Y_1 and Y_2 be independent random variables with $Y_1 \sim \mathcal{N}(0,1)$ and $Y_2 \sim \mathcal{N}(3,4)$.

   (a) What is the distribution of Y_1^2?

   (b) If $\mathbf{y} = \begin{bmatrix} Y_1 \\ (Y_2-3)2 \end{bmatrix}$, obtain an expression for $\mathbf{y}^T\mathbf{y}$. What is its distribution?

   (c) If $\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$ and its distribution $\mathbf{y} \sim \text{MVN}(\mu, \mathbf{V})$, obtain a expression for $\mathbf{y}^T\mathbf{V}^{-1}\mathbf{y}$.

   hint: for some sample distributions, such as normal, χ^s, t-dist and f-dist please check out slides 13-14 from lecture note 02_2BasicLinearRegression.

3. Let $Y_1, \ldots, Y_n$ be independent random variables each with the distribution $\mathcal{N}(\mu, \sigma^2)$. Let

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

   (a) What is the distribution of $\bar{Y}$?

   (b) Show that $S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}(Y_i-\mu)^2 - n(\bar{Y}-\mu)^2\right]$.

   (c) From (b) it follows that $\sum_{i=1}^{n}(Y_i-\mu)^2/\sigma^2 = (n-1)S^2/\sigma^2 + \left[n(\bar{Y}-\mu)^2/\sigma^2\right]$.

   (d) What is the distribution of $(n-1)S^2/\sigma^2$?

   (e) What is the distribution of $\frac{\bar{Y}-\mu}{S/\sqrt{n}}$?

   hint: for relationships between sum of quares and χ^2, please check out lecture note on slides 31-33 from 03_1MultipleLinearRegression.

4. **Linear Regression** Data File Table 6.3 Carbohydrate diet.xls Please refer to Dobson Exercise 6.3 on page 119 and answer the following.
Consider the following two models

$$\begin{array}{lll} \text{Model A} & \mathrm{E}[Y_i] = \beta_0 + \beta_3 x_{i3} & x_3 = \text{Protein} \\ \text{Model B} & \mathrm{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} & x_1 = \text{Age}; x_3 = \text{Protein} \end{array}$$

   (a) Fit Model B and perform the following:
      1. Calculate the 95% confidence interval for β_1, the coefficient for age.
      2. Test the hypothesis (at 5%) that the response does not depend on age.

   (b) Fit Model A and calculate the 95% prediction interval when protein = 21.

   (c) Compare Model A with Model B, use deviance to test the hypothesis (at 5%) that the response does not depend on age.

5. **National Life Expentancies** data for the following questions. this exercise involves data filename "UNLifeExpectancy" from Frees page 137.

   This is considering Health system from $N = 185$ countries. Use LIFEEXP as response variable. You are asked to analyze the following:

   (a) You are asked to fit a regression model on LIFEEXP using three explanatory variables, FERTILITY, PUBLICEDUCATION, and lnHEALTH (the natural logarithm transform of PRIVATEHEALTH).
      i. Interpret the regression coefficient associated with public education.
      ii. Based on the model fit, is PUBLICEDUCATION a statistically significant variable? To respond to this question, use a formal test of hypothesis, State your null hypotheses, decision-making criterion, and decision-making rule.
      iii. Test whether PUBLICEDUCATION and LNHEALTH are jointly statistically significant (it means that the coefficients of the two variables are equal to zero versus the alternative that not both the coefficients are zero), using the F-test. State your null and alternative hypotheses, decision-making criterion, and decision making rules. Provide an approximate $p-$value for the test.

   (b) We now introduce the REGION variable, summarized in Table 4.11 (Frees, page 137), A box of plot of life expectancies versus REGION is given in Figure 4.8 (Frees, Page 138). Describe what we learn from the table and box plot about the effect of REGION on LIFEEXP.

   (c) Fit a regression model using three explanatory variables, FERTILITY, PUBLICEDUCATION, and LNHEALTH (the natural logarithmic transform of PRIVATEHEALTH), as well as the categorical variable REGION.
      i. You are examining a country that is not in the sample with values FERTILITY = 2.0, PUBLICEDUCATION = 5.0, and LNHEALTH = 1.0, produce two predicted life expectancy values by assuming that the country is from (1) an Arab state and (2) sub-Saharan Africa.

ii. Provide a 95% confidence interval for the difference in the life expectancies between an Arab state and sub-Saharan Africa.

iii. provide the (usual ordinary least squares) point estimate for the difference in the life expectancies between a country from sub-Saharan Africa and a high-income OECD country.

Region in National Life Expectancies

Region	Region Description	Number	Mean
1	Arab states	13	71.9
2	East Asia and the Pacific	17	69.1
3	Latin American and the Carribean	25	72.8
4	South Asia	7	65.1
5	Southern Europe	3	67.4
6	Sub-Saharan Africa	38	52.2
7	Central and Eastern Europe	24	71.6
8	High-income OECE	23	79.6
	All	150	67.4

Table 1: Average Life Expectancy by Region

Note: OECD stands Organization for Economic Co-operation and Development