$z_k = \mathbb{1}(x_{k1} = 20)$

$\downarrow$

Homework Assignment 4
Due date: Nov. 2 請印出來以紙本繳交

pp 172-173

$y_k = \beta_0 + \beta_1 x_{k1} + \varepsilon_k$ vs. $y_k = \beta_0 + \beta_1 x_{k1} + \beta_2 z_k + \varepsilon_k$

1. Fit a line by least squares to the following points: $(4, 0.9), (3, 2.1), (2, 2.9), (1, 4.1)$ and $(20, 20)$. Obtain Studentized residuals and also plot the points and the estimated line. Does the point $(20, 20)$ appear as an outlier? Using a suitable indicator variable, numerically demonstrate the indicator variable interpretation of RSTUDENT's. Also demonstrate that DFFIT and the DFBETA's do indeed measure what has been claimed for them. $\Rightarrow \hat{\beta}_2 = \frac{\hat{e}_k}{1-h_{kk}}$, $\frac{\hat{\beta}_2}{\sqrt{Var(\hat{\beta}_2)}} = \frac{e_k}{s_k(1-h_{kk})} \sim t$

2. Consider the data in Exhibit 8.12 on male deaths per million in 1950 for lung cancer (y) and per capita cigarette consumption in 1930 (x).

1. Estimate a model expressing y as a linear function of x. Do any of the points look particularly influential? Delete the United States, rerun the model and check if the influences of Great Britain and Finland have been substantially altered. Now, put the U.S. back and delete Great Britain and examine the influence of points for the resultant model.

2. Try an appropriate broken line regression and examine the residuals. If you notice heteroscedasticity, run an appropriately weighted regression. Examine the data points for outliers or undue influence. $y_k = \beta_0 + \beta_1 x_{k1} + \beta_2 (x_{k1} - \underset{\substack{\uparrow \\ \text{change point}}}{c})\delta + \varepsilon_k$ $\delta = \mathbb{1}(x_{k1} - c)$

3. Do you think a plausible reason for using broken line regression is that the number of women who smoke might be much higher in countries with high per capita cigarette consumption?

4. Write a report discussing your various efforts and your final conclusion.

(A discussion of part 1 of this example is contained in Thfte, 1974, p. 78 et seq. The data was used in some of the earlier reports on Smoking and Health by the Advisory Committee to the U.S. Surgeon General. See reference to Doll, 1955.)

3. Consider the model $y_i = \beta x_i + \varepsilon_i$, where ε_i's are independent and distributed as $N(0, \sigma^2 x_i^2)$. Find the weighted least squares estimator for β and its variance. Give reasons why you would not wish to use ordinary least squares in this case.

4. Consider the model $y_i = \beta x_i + \varepsilon_i$, where $i = 1, ..., n$, x_i's are $k+1$-vectors, the ε_i's are independently distributed with means zero and variances $w_i^{-1}\sigma^2$ and the w_i's are known positive integers. Show that the weighted least squares estimate of β can be obtained using ordinary least squares in the following way: Construct a data set in which each of the cases (y_i, x_i) is repeated w_i times. Show that the ordinary least squares estimate of β obtained from this data set is $(X'WX)^{-1}X'Wy$, and an unbiased estimate of σ^2 is $\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2/(n-k-1)$, where $X' = (x_1, \ldots, x_n)$ and $W = diag(w_1, \ldots, w_n)$. (These estimates are therefore the same as the corresponding weighted least squares estimates using the w_i's as weights.)
[Hint: Let $1_i = (1, \ldots, 1)'$, be a vector of 1's of dimension w_i. To obtain the OLS estimator, we are using the model $D_y = DX\beta + D\varepsilon$, where

$$\begin{pmatrix} 1_1 & 0 & \ldots & 0 \\ 0 & 1_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1_n \end{pmatrix} : \sum_{i=1}^{n} w_i \times n.]$$

5. This example compares in-field ultrasonic measurements of the depths of defects in the Alaska oil pipeline to measurements of the same defects in a laboratory. The lab measurements were done in six different batches. The goal is to decide if the field measurement can be used to predict the more accurate lab measurement. In this analysis, the field measurement is the response variable and the laboratory measurement is the predictor variable. The data, in the file *W_pipeline.csv*, were given at `www.itl.nist.gov/div898/handbook/pmd/section6/pmd621.htm`. The three variables are called *Field*, the in-field measurement, *Lab*, the more accurate in-lab measurement, and *Batch*, the batch number.

1. Draw the scatterplot of Lab versus Field, and comment on the applicability of the simple linear regression model.

2. Perform a two-stage least squares approach that models the conditional variance of *Lab* as a function of *Field*. State your findings and conclusions.

3. In addition to the variable *Field*, consider the variable *Batch*, treating *Batch* as a class (or categorical) variable, in the regression model. Does the conditional variance vary with *Batch*? State your findings and conclusions.