COVER PAGE

STAT 608 Homework 02, Summer 2017

Please write your name and email address below and attach as the first page of your homework upload.

NAME: RAJAN KAPOOR

EMAIL: r.kapoor@tamu.edu

18/22

Solution 1

$$S_y^2 = \frac{\sum_{j=1}^{n}(y_j - \bar{y})^2}{n-1} = 100 = \frac{SST}{n-1} \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = 10 = \frac{RSS}{n-2}$$

$SST = 100 \times 49 = 4900$ $\Rightarrow RSS = 480$

Since

$SST = SSreg + RSS$ and all terms are positive

So S_y^2 is related to total sample variability while $\hat{\sigma}^2$ is related to the variability unexplained by the model

Clearly, S_y^2 is not an estimator of σ^2. (C)

Solution 2

(a) From the standardized residual vs Distance plot, one value lies outside the -2 to 2 interval. Cook's distance for this point is found to be $> 4/15$ (identified as City No. 17). Clearly this is an outlier unexplained by the model described. Since the model is not valid, any claim based on the model are not useful.

As far as the claim of 99.4% variability explained by the model is concerned one of the points (City 13) lies far enough in the x direction to be classified as leverage point. This point has standardized residual inside -2 to 2 interval and is therefore a good leverage point. Good Leverage points can increase the value of R^2 and decrease the estimated standard errors of estimated regression coefficient.

(b) No. Clearly with current model there is a bad leverage point. Also there is a non-linear (non-random) pattern in Standard residual plot which shows that the assumption of constant variance of residual errors is not satisfied.

To improve the model, first check using box-plot / normal Q-Q plot for Fare and distance variables if the samples resemble normal distribution. Accordingly, consider transforming any one or both followed by regression. Box-Cox Method could be used for Distance variable while Inverse Response plots could be used for Fare variable.

Solution 3

From Taylor series results on Pg 77

$$Var(f(Y)) = [f'(E(Y))]^2 \, Var(Y)$$

$$= [f'(\mu)]^2 \mu^2$$

Given $\mathrm{Var}(f(y)) = 1$

$$\Rightarrow [f'(y)]^2_{y=\mu} = \frac{1}{\mu^2}$$

$$\Rightarrow f'(Y)\Big|_{Y=\mu} = \frac{1}{\mu} \Rightarrow f'(Y) = \frac{1}{Y}$$

$$\Rightarrow f(Y) = \log(Y)$$

Solution 4

$$Y_i = \beta x_i + e_i$$

$$\mathrm{Var}(e_i|x_i) = x_i^2 \sigma^2$$

Comparing with residual definition, $w_i = \frac{1}{x_i^2}$

$$\hat{e}_{wi} = \sqrt{w_i}\,(y_i - \hat{y}_i)$$

$$= x_i (y_i - \hat{y}_i)$$

$$WRSS = \sum_{i=1}^{n} \hat{e}_{wi}^2$$

$$= \sum_{i=1}^{n} x_i^2 (y_i - \hat{y}_{wi})^2$$

For least square estimate,

$$\frac{\partial WRSS}{\partial \beta} = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i^3 y_i = \sum_{i=1}^{n} x_i^4 \hat{\beta}$$

$$\Rightarrow \boxed{\hat{\beta} = \frac{\sum_{i=1}^{n} x_i^3 y_i}{\sum_{i=1}^{n} x_i^4}}$$

4

Solution 5

(a) Since Y_i is the median of n_i observations so $Var(Y_i) \propto \frac{1}{n_i}$
So weights $w_i = n_i$

(b) From Y_i vs $X1i$ and Yi vs $X2i$ plots the relationship doesnot seem like linear.
From standardized residual plot variance is not constant.
Significant no. of outliers from standardized residual plot.

(c)(i) Find transformations $\Psi_{S1}(x1i, \lambda_{x1})$ and $\Psi_{S2}(x2i, \lambda_{x2})$ that make X1 X2 close to normal using Box-cox. i.e. find $\lambda_{x1}, \lambda_{x2}$

(ii) Consider regression model

$$Y = g(\beta_0 + \beta_1 \Psi_{S1} + \beta_2 \Psi_{S2} + e)$$

Using inverse response plot, find g^{-1} and corresponding regression coefficients.