Regression Analysis (I) Midterm solution

Po-Wei Chen

November 14, 2022

1. A simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad 3\%$$

where:

(1) Y_i: the value of the response variable in the ith trial.

(2) β_0 and β_1: parameters to be estimated.

(3) X_i: the value of the predictor variable in the ith trial.

(4) ϵ_i: a random error term with mean $E(\epsilon_i) = 0$ and variance $\sigma^2(\epsilon_i) = \sigma^2$. 1%

(5) ϵ_i and ϵ_j are uncorrelated so that their covariance is zero (i.e., $\sigma(\epsilon_i, \epsilon_j) = 0$ for all $i, j; i \neq j$) $i = 1, \cdots, n$. 1%

5% 2. To set up interval estimates and make tests, however, we need to make an assumption about the form of the distribution of the error terms ϵ_i: they are normally distributed.

10% 3. Diagnostic for the predictor variable to see if there are any outlying X values that could influence the appropriateness of the fitted regression function. We can draw the dot plot, the sequence plot, the stem-and-leaf plot, and the box plot to see if there are any outlying X values that could influence the appropriateness of the fitted regression function.

4. (1) The regression function is not linear. 2%

(2) The error terms do not have constant variance. 2%

(3) The error terms are not independent. 1%

(4) The model fits all but one or a few outlier observations. 2%

(5) The error terms are not normally distributed. 2%

(6) One or several important predictor variables have been omitted from the model. 1%

10% 5. Residual plot is a scatterplot of residuals against the fitted values or the predictor variable. It is used to examine (diagnose) six important types of departures from the simple linear regression model with normal errors.

6. Model misspecification due to either nonlinearity or the omission of important predictor variables tends to be serious, leading to biased estimates of the regression parameters and error variance. The presence of outliers can be serious for smaller data sets when their influence is large. The nonindependence of error terms results in estimators that are unbiased but whose variances are seriously biased. 5%

   Nonconstancy of error variance tends to be less serious, leading to less efficient estimates and invalid error variance estimates. 5%

7. Brown-Forsythe Test

   Assumption: the sample size needs to be large enough so that the dependencies among the residuals can be ignored.

5% (a) $\begin{cases} H_0: \text{ error variance is constant.} \\ H_a: \text{ error variance is not constant.} \end{cases}$

5% (b) Notations: use e_{i1} to denote the ith residual for group 1 and e_{i2} to denote the ith residual for group 2, the sample sizes of the two groups by n_1 and n_2, the medians of the residuals in the two groups by $\widetilde{e}_1$ and $\widetilde{e}_2$, and the absolute deviations of the residuals around their group median, to be denoted by d_{i1} and d_{i2}, where $d_{i1} = |e_{i1} - \widetilde{e}_1|$ and $d_{i2} = |e_{i2} - \widetilde{e}_2|$.

Test statistic: $t^*_{BF} = \dfrac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $\bar{d}_1$ and $\bar{d}_2$ are the sample means of the d_{i1} and d_{i2} respectively, and the pooled variance s^2 is $\dfrac{\sum(d_{i1} - \bar{d}_1)^2 + \sum(d_{i2} - \bar{d}_2)^2}{n-2}$.

(c) (1) 1% $\begin{cases} H_0: \text{ error variance is constant.} \\ H_a: \text{ error variance is not constant.} \end{cases}$

(2) Significance leval: $\alpha = 0.05$

1% (3) Test statistic: $t^*_{BF} = \dfrac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(1-\alpha/2; n-2)$

1% (4) Decision rule:

(a) If $|t^*_{BF}| > t(1-\alpha/2; n-2)$, then reject H_0.

(b) If p-value $< \alpha$, then reject H_0.

1% (5) Decision:

$\because$ P-value $= 0.8475129 > 0.05 = \alpha$

$\Rightarrow$ do not reject H_0 at $\alpha = 0.05$

1% (6) Conclusion:

Hence the error variance is constant.

8. (a) The lack of fit test assumes that the observations Y for given X are

2% (1) independent

2% (2) normally distributed

1% (3) the distributions of Y have the same variance σ^2

5% (b) $\begin{cases} \text{Full Model} : Y_{ij} = \mu_j + \epsilon_{ij},\ i = 1,2,3,4, j = 1,2,3,4, \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \\ \text{Reduced Model} : Y_{ij} = \beta_0 + \beta_1 X_j + \epsilon_{ij} \end{cases}$

(c) (1) $df_r = n - 2 = 16 - 2 = 14$

(2) $SSE = 3.234^2 \times 14 \approx 146.4226$

(3) $df_r - df_F = (n-2) - (n-c) = (16-2) - (16-4) = 2$

(4) $SSLF = SSE - SSPE = 146.4226 - 128.75 = 17.6726$

(5) $F^* = \dfrac{SSLF}{c-2} \div \dfrac{SSPE}{n-c} = \dfrac{17.6726}{4-2} \div \dfrac{128.75}{16-4} \approx 0.8236$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	14 2%	146.4226 2%				
2	12	128.75	2 2%	17.6726 2%	0.8236 2%	0.4622

(d) (1) 1%
$$\begin{cases} H_0 : E(Y) = \beta_0 + \beta_1 X \\ H_a : E(Y) \neq \beta_0 + \beta_1 X \end{cases}$$

(2) Significance leval: $\alpha = 0.01$

1% (3) Test statistic: $F^* = \dfrac{SSLF}{c-2} \div \dfrac{SSPE}{n-c} = \dfrac{MSLF}{MSPE}$

1% (4) Decision rule:

(a) If $F^* > F(1-\alpha; c-2, n-c)$, then reject H_0.

(b) If p-value $< \alpha$, then reject H_0.

1% (5) Decision:

$\because$ p-value $= 0.4622 > 0.01 = \alpha$

$\Rightarrow$ do not reject H_0 at $\alpha = 0.01$

1% (6) Conclusion:

Hence the regression function is linear.

9. (a) 5% A linear relation does not appear to be adequate here. The regression relation in the scatter plot appears to be curvilinear.

(b) 5% Clearly the regression relation appears to be curvilinear, so the simple linear regression model does not seem to be appropriate. Since the variability at the different X levels appears to be fairly constant, we shall consider a transformation on X. We consider initially the square root transformation $X' = \sqrt{X}$.

(c) We know that $E(e_i) = 0$ and $Var(e_i) = MSE$ for all $i = 1, ..., n$.

If $e \sim \mathcal{N}(0, MSE)$, then $Z = \dfrac{e-0}{\sqrt{MSE}} \sim \mathcal{N}(0, 1)$.

$\Rightarrow e = Z \times \sqrt{MSE}$

A good approximation of the expected value of the kth smallest observation in a random sample of n_i is

$$\sqrt{MSE} \times Z_{\left(\frac{k-0.375}{n+0.25}\right)}$$ 5%

where Z_a is the $a \times 100$ percentile of the standard normal distribution.

From Output (VI), we know that $MSE = 4.48$ and $n = 111$.

$\because n = 111$ and $k = 43$

$$\Rightarrow \frac{k-0.375}{n+0.25} = \frac{43-0.375}{111+0.25} \approx 0.3831$$

$\Rightarrow$ the expected value of e_1 under normality is $\sqrt{4.48} \times Z_{0.3831}$ 5%