STAC67H: Regression Analysis

Fall, 2014

Instructor: Jabed Tomal

Department of Computer and Mathematical Sciences
University of Toronto Scarborough
Toronto, ON
Canada

October 1, 2014

Diagnostics and Remedial Measures:

1. Graphic analysis of residual is inherently subjective.
2. There are occasions when one wishes to put specific questions to a test.

Diagnostics and Remedial Measures:

1. Tests for Randomness: Durbin-Watson test - will be covered in Chapter 12.
2. Tests for Outliers: Will be covered in Chapter 10.
3. Tests for Normality: Normal-probability plot has good reputation regarding test for normality.

Breusch-Pagan Test

1. A large sample test: this test assumes that the error terms are independent and normally distributed.
2. The variance of the error term ϵ_i, denoted by σ_i^2, is related to the level of X

$$\log_e \sigma_i^2 = \gamma_0 + \gamma_1 X_i.$$

3. The variance σ_i^2 either increases or decreases with the level of X, depending on the sign of γ_1.

Breusch-Pagan Test

1. The constancy of error variance holds when $\gamma_1 = 0$.
2. Fit a simple linear regression model regressing e_i^2 (response) against X_i (predictor) and obtain the *regression sum of squares* SSR^*.
3. Compute the test statistic as follows

$$\chi^2_{BP} = \frac{SSR^*}{2} \Big/ \left(\frac{SSE}{n}\right)^2$$

where SSE is the error sum of squares when regressing Y on X.

Breusch-Pagan Test

1. The hypotheses to be tested are

$$H_0 : \gamma_1 = 0 \quad \text{vs} \quad H_A : \gamma_1 \neq 0$$

2. Under H_0 and for large n, the approximate distribution of χ^2_{BP} is χ^2 with 1 degrees of freedom.
3. Reject H_0 at α level of significance if

$$\chi^2_{BP} \geq \chi^2(1 - \alpha, 1)$$

Breusch-Pagan Test: In our airfreight breakage problem ($n = 10$)

1. The error sum of squares is

$$SSE = 17.6$$

2. The regression sum of squares by regressing e_i^2 using X_i, we get

$$SSR^* = 6.4$$

3. Hence, the Breusch-Pagan test statistic is

$$\chi^2_{BP} = \frac{SSR^*}{2} / \left(\frac{SSE}{n}\right)^2 = 1.033058$$

Breusch-Pagan Test: In our airfreight breakage problem ($n = 10$)

1. At $\alpha = 0.05$ level of significance

$$\chi^2(0.95, 1) = 3.841459$$

2. Here,

$$\chi^2_{BP} \leq \chi^2(0.95, 1) = 3.841459$$

3. Decision: We don't have enough evidence to reject $H_0 : \gamma = 0$.

If the simple linear regression model ($Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$) is not appropriate for a data set, there are two basic choices:

1. Abandon the regression model and develop and use a more appropriate model.
2. Employ some transformation on the data so that the above regression model is appropriate for the transformed data.

Diagnostics and Remedial Measures:

Box-Cox Transformations

1. It is often difficult to determine from diagnostic plots which transformation of Y is appropriate for correcting skewness of the distributions of error terms, unequal error variances, and nonlinearity of the regression function.
2. The Box-Cox procedure automatically identifies a transformation from the family of power transformations on Y.

1. The family of power transformation is of the form

$$Y^* = Y^\lambda$$

A few examples of this power transformations

$$
\begin{array}{ll}
\lambda = 2 & Y^* = Y^2 \\
\lambda = 0.5 & Y^* = \sqrt{Y} \\
\lambda = 0 & Y^* = \log_e Y \quad \text{(by definition)} \\
\lambda = -0.5 & Y^* = \frac{1}{\sqrt{Y}} \\
\lambda = -1.0 & Y^* = \frac{1}{Y}
\end{array}
$$

With this power transformation, the normal error regression model becomes

$$Y_i^{\lambda} = \beta_0 + \beta_1 X_i + \epsilon_i$$

where the additional parameter λ needs to be estimated from the data using maximum likelihood method.

An Alternative to Maximum Likelihood

1. Conduct a numerical search in a grid of potential λ values; for example,

$$\lambda = -2, \lambda = -1.75, \cdots, \lambda = 1.75, \lambda = 2.$$

2. For each λ, standardize Y_i^λ using W_i so that the magnitude of the error sum of squares does not depend on the value of λ

$$W_i = \begin{cases} K_1(Y_i^\lambda - 1) & \lambda \neq 0 \\ K_2(\log_e Y_i) & \lambda = 0 \end{cases}$$

An Alternative to Maximum Likelihood

1. where
$$K_2 = \left(\prod_{i=1}^{n} Y_i \right)^{1/n}$$
the geometric mean of the Y_i observations,
2. and
$$K_1 = \frac{1}{\lambda K_2^{\lambda - 1}}$$
3. Fit regression model with response W_i and predictor variable X_i, and get SSE_λ.

An Alternative to Maximum Likelihood

1. The maximum likelihood estimate $\hat{\lambda}$ is that value of λ for which SSE_λ is a minimum.

Box-Cox Transformations

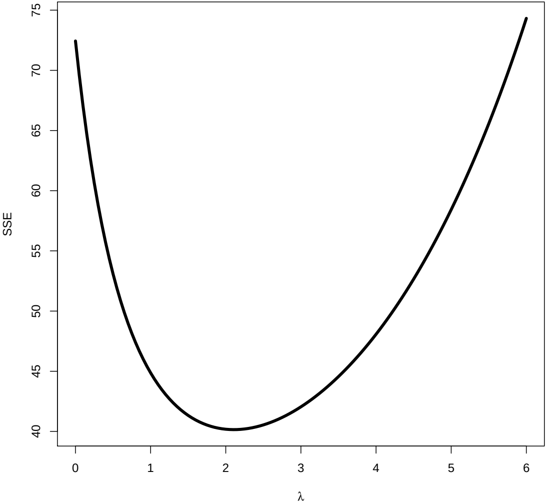

Figure: Plot of *SSE* against λ to determine appropriate transformation from the Box-Cox power transformations.

Comments

1. After a transformation has been selected, residual plots and other described analysis need to be employed to ascertain that the simple linear regression model is appropriate for the transformed data.
2. When transformed models are employed, the estimators b_0 and b_1 obtained by least squares have the least squares properties with respect to the transformed observations, not the original ones.
3. The error sum of squares SSE is often fairly stable in a neighborhood around the estimate. It is therefore often reasonable to use a nearby λ value for which the power transformation is easy to understand.
4. When the Box-Cox procedure leads to a λ value near 1, no transformation of Y may be needed.

Simultaneous Inferences:
Bonferroni Joint Confidence Intervals

1. The $100(1-\alpha)\%$ confidence interval of β_0 is

$$b_0 \pm t(1-\alpha/2; n-2)s\{b_0\}$$

2. The $100(1-\alpha)\%$ confidence interval of β_1 is

$$b_1 \pm t(1-\alpha/2; n-2)s\{b_1\}$$

3. What is the confidence coefficient of their joint intervals?

1. Let A_1 denote the event that the first confidence interval does not cover β_0. Then
$$P(A_1) = \alpha$$
2. Let A_2 denote the event that the second confidence interval does not cover β_1. Then
$$P(A_2) = \alpha$$
3. Here $A_1^C \cap A_2^C$ is the event which indicates that both of the confidence intervals cover β_0 and β_1.
$$P(A_1^C \cap A_2^C) = ?$$

1. Here,
$$A_1^C \cap A_2^C = (A_1 \cup A_2)^C$$
2. The probability theory of a complimentary event gives us
$$P(A_1^C \cap A_2^C) = 1 - P(A_1 \cup A_2) = 1 - P(A_1) - P(A_2) + P(A_1 \cap A_2)$$
3. After simplification, we write
$$P(A_1^C \cap A_2^C) = 1 - \alpha - \alpha + P(A_1 \cap A_2)$$

1. Since $P(A_1 \cap A_2) \geq 0$, we get

$$P(A_1^C \cap A_2^C) \geq 1 - 2\alpha.$$

This inequality is called *Bonferroni inequality*.

2. Hence, the confidence coefficient of containing both of the parameters in their respective intervals could be as low as

$$1 - 2\alpha.$$

3. Example: If both of the individual coefficients are 0.95, the joint confidence coefficient could be as low as 0.90.

1. We use Bonferroni inequality to obtain a family confidence coefficient of at least $1 - \alpha$ for estimating β_0 and β_1.
2. We do this by increasing the confidence coefficients for each β_0 and β_1 to
$$1 - \alpha/2.$$
3. This results to a Bonferroni bound of at least
$$1 - \alpha.$$

1. We use Bonferroni inequality to obtain a family confidence coefficient of at least $1 - \alpha$ for estimating β_0 and β_1.
2. We do this by increasing the confidence coefficients for each β_0 and β_1 to
$$1 - \alpha/2.$$
3. This results to a Bonferroni bound of at least
$$1 - \alpha.$$

1. Thus, the $1 - \alpha$ joint Bonferroni confidence limits for β_0 and β_1 for is

$$b_0 \pm Bs\{b_0\} \quad b_1 \pm Bs\{b_1\}$$

where $B = t(1 - \alpha/4; n - 2)$.

Airfreight breakage problem:

1. The estimates are

$$b_0 = 10.2, \quad b_1 = 4, \quad MSE = 2.2, \quad s^2\{b_0\} = 0.44, \quad s^2\{b_1\} = 0.22$$

2. For $\alpha = 0.05$

$$B = t(1 - 0.05/4; 10 - 2) = 2.751524.$$

3. Hence, the 95% simultaneous confidence intervals for β_0 and β_1 are

$$8.374846 \leq \beta_0 \leq 12.025154$$

and

$$2.709421 \leq \beta_1 \leq 5.290579$$

1. We consider the simultaneous predictions of g new observations on Y in g independent trials at g different levels of X.
2. With the Bonferroni procedure, the $1-\alpha$ simultaneous prediction intervals are:
$$\hat{Y}_h \pm Bs\{pred\}$$
where $B = t(1-\alpha/2g; n-2)$.

Simultaneous Inferences:

Bonferroni Joint Confidence Intervals

Airfreight breakage problem:

1. We want to get simultaneous prediction intervals for two new observations ($Y_{h1} = 18, X_{h1} = 2$) and ($Y_{h2} = 23, X_{h2} = 3$). The estimates are

$$\hat{Y}_{h1} = 18.2,\ \hat{Y}_{h2} = 22.2,\ MSE = 2.2,\ s^2\{\hat{Y}_{h1}\} = 2.64,\ s^2\{\hat{Y}_{h2}\} = 3.3$$

2. For $\alpha = 0.05$ and $g = 2$

$$B = t(1 - 0.05/4; 10 - 2) = 2.751524.$$

3. Hence, the 95% simultaneous confidence intervals for Y_{h1} and Y_{h2} are

$$13.7293 \leq Y_{h1} \leq 22.6707$$

$$17.20161 \leq Y_{h2} \leq 27.19839$$

1. We consider the simultaneous predictions of g new observations on Y in g independent trials at g different levels of X.
2. With the Scheffe procedure, the $1 - \alpha$ simultaneous prediction intervals are:
$$\hat{Y}_h \pm Ss\{pred\}$$
where $S^2 = gF(1 - \alpha; g, n - 2)$.

Airfreight breakage problem:

1. We want to get simultaneous prediction intervals for two new observations ($Y_{h1} = 18, X_{h1} = 2$) and ($Y_{h2} = 23, X_{h2} = 3$). The estimates are

$$\hat{Y}_{h1} = 18.2,\ \hat{Y}_{h2} = 22.2,\ MSE = 2.2,\ s^2\{\hat{Y}_{h1}\} = 2.64,\ s^2\{\hat{Y}_{h2}\} = 3.3$$

2. For $\alpha = 0.05$ and $g = 2$

$$S^2 = 2F(1 - 0.05; 2, 10 - 2) = 8.91794.$$

3. Hence, the 95% simultaneous confidence intervals for Y_{h1} and Y_{h2} are

$$13.34785 \leq Y_{h1} \leq 23.05215$$

$$16.77513 \leq Y_{h2} \leq 27.62487$$