, 32, 33, 34, 35, 36

Stat 231 - Statistics **Spring 2017**

Lecture 31, 32, 33, 34, 35, 36: July 12 - July 24, 2017

Lecturer: Suryapratim Banerjee *Notes By: Harsh Mistry*

31.1 Estimated Residuals

$\hat{r_i}$ = actual - predicted.

Standardized Estimated Residual

$$\hat{r_i} = \frac{\hat{r_i}}{s} \sim Z$$

where s is the standard error

31.2 Tests For The Regression Assumtion

The tests are graphical (and subjective) comes with experience

- Scatter Plot : We draw a scatter plot and check whether a linear relationship is appropriate
- Residual Plots
  - $\hat{r_i}$'s should be in a "small" band around zero
  - Variability in the $\hat{r_i}$'s should be more or less constant
  - Absence of any obvious patterns
- The QQ-plot. : If the assumptions are right, the q-q plot should be a 45 degree line

31.3 Two Population Problems

Equality of means

- Matched pair populations
- Unmatched data : equal variance populations
- Unmatched data : unequal variances and large sample sizes

Matched Pair

Consider the following :

$$B_1, \ldots, B_n \sim N(\gamma_1, \sigma_1^2)$$
$$A_1, \ldots, A_n \sim N(\gamma_2, \sigma_2^2)$$

Definition 31.1 $(B_i, A_i) \rightarrow$ *is a matched pair in the population, Given this this the units are the same or there is a natural match between the two populations*

In this case the NULL hypothesis and Challenging view are

$$H_0 : \gamma_1 = \gamma_2$$

$$H_1 : \gamma_1 \neq \gamma_2$$

Define $Y_i = A_i - B_i$, so

$$Y_i \sim N(\gamma_2 - \gamma_1, \sigma_1^2 + \sigma_2^2)$$

Using this we find

$$D = \mid \frac{\bar{Y}}{S/\sqrt{n}} \mid$$
$$d = \mid \frac{\bar{y}}{S/\sqrt{n}} \mid$$
$$S = \frac{1}{n-1} \sum (Y_i - \bar{y})$$
$$p - value = P(D \geq d) = P(\mid T_{n-1} \mid \geq d)$$

Unmatched Data

There is no natural pairing between the two populations.

Model

$$Y_{1i} \sim N(\gamma_1, \sigma^2)$$
$$Y_{2j} \sim N(\gamma_2, \sigma^2)$$

First Method for Equal Variance

In this case, we're assuming two populations have the same variability.

$$\text{From 1 } \hat{Y_1} \sim N(\gamma_1, \sigma^2/n_1)$$

$$\text{From 2 } \hat{Y_1} \sim N(\gamma_2, \sigma^2/n_2)$$

Thus,

$$\hat{Y_1} - \hat{Y_2} \sim N(\gamma_1 - \gamma_2, \sigma^2(\frac{1}{n_1} + \frac{1}{n_1}))$$

So,

$$D = \mid \frac{\hat{Y_1} - \hat{Y_2} - 0}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \mid \sim T_{n_1+n_2-2}$$
$$S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$
$$p - value = P(D \geq d) = P(\mid T_{n_1+n_2-2} \mid \geq d)$$

Second Method for Equal Variance

Define $X = \begin{cases} 0 \text{ if population} = 1 \\ 1 \text{ if population} = 2 \end{cases}$

Using this you can draw a linear graph where $\alpha = E(Y)$ and $B =$ change in 1 unit, As a result the p-value can be found easily.

$$D = \mid \frac{\tilde{\beta} - 0}{S/\sqrt{S_{XX}}} \mid$$
$$d = \mid \frac{\hat{\beta} - 0}{S/\sqrt{S_{XX}}} \mid$$
$$p - value = P(D \geq d) = P(\mid T_{n_1+n_2-2} \mid \geq d)$$

Unequal variance

$$Y_{1i} \sim G(\gamma_1, \sigma^2)$$
$$Y_{2j} \sim G(\gamma_2, \sigma^2)$$

Assume n_1, n_2 are large

$$D = \frac{(\bar{y_1} - y \bar{-} 2) - (y_1 - y_2)}{\sqrt{(S_1^2/n_1) + (S_2^2/n_2)}}$$

31.4 Test For Goodness Of Fit

In so,e situations, the unknown parameter of interest is a vector

$$\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_m)$$
$$H_0 = \underset{\sim}{\theta} = \underset{\sim}{\theta}(\alpha)$$

Recall :

$$\triangle(\theta) = -2\log\frac{L(\theta)}{L(\hat{\theta})} \sim X_1^2$$

Theorem 31.2 *If $\underset{\sim}{\theta}$ is a vector then*

$$\triangle(\theta) = -2\log\frac{L(\theta)}{L(\hat{\theta})} \sim X_n^2$$

Where $n =$ the number of independent unrestricted parameters of θ + the number of parameters estimated under H_0

For a Multinomial Problem

$$\triangle = 2\sum Y_i \cdot \ln \frac{Y_i}{E_i} \sim X^2 + n - 1 - 1$$

where

- Y_i = observation frequency of category i
- E_i = expected frequency of category i if H_0 is true

Given this we can determine that

- $E_i = n \times p_i$
- p_i = expected probabilities of category i under H
- n = sample size

Poisson Problem

Divide the sata into categories and compute the observed frequency of each category

$$\hat{p_i} = \frac{e^{\bar{x}}\bar{x}^i}{i!}$$

$$E_i = nxp_i$$

Exponential Problem

Produce a table consisting of the frequency associated with the interval, then use the following to determine the expected values

$$\hat{\theta} = \bar{x}$$

$$\hat{p_i} = \int_{interval_{min}}^{interval_{max}} \frac{1}{\bar{x}} e^{-\frac{2}{\bar{x}}} dx$$

Restrictions :

- n needs to be large
- $n_i \geq 5 \forall i$

Two Categorical Variables

$$e_{ij} = \frac{\text{ith Row Total} \times \text{jth Column Total}}{\text{Total sample size}}$$

$$\lambda = 2\sum_i \sum_j y_{ij} \ln \frac{y_{ij}}{e_{ij}}$$

Normal Problem

$$H_0 : X_i \sim G(\gamma, \sigma^2)$$

1. Divide the data into mutually exclusive and exhaustive categories and calculate the frequencies of each category.
   We need atleast 3 categories
2. Assume the null hypothesis is true
   Estimate $\hat{p}_i$ = estimate probability of each category.
$$\hat{\gamma} = \bar{x}$$
$$p_i = P(\frac{interval_{min} - \bar{x}}{\text{number of elements}} \leq x \leq \frac{interval_{max} - \bar{x}}{\text{number of elements}}$$
3. Calculate e_i
$$e_i = n \times \hat{p}_i$$
4. Compute λ
$$\lambda = 2 \sum_i y_i \ln \frac{y_i}{e_i}$$
5. Compute the p-value
$$p - value = P(X^2_{\text{number of categories}} \geq \lambda)$$

General Problem

$$f(y_i; \theta) = \frac{2y}{\theta} e^{-y^2/\theta} y \geq 0$$
$$H_0 : Y_i \sim f(y_i; \theta);$$

- Compute $\theta = MLE$ for θ from your data and use that to compute $\hat{p}_i$ and thus e_i

Pearson's Chi-Squared Statistic :

$$k = \sum_{i=1}^{n} \frac{(Y_i - E_i)^2)}{E_i}$$

This also follows X^2 with the same degree of freedom, but is less powerful than the LRTS.

31.5 Design of Experiments

We have to design the experiments in such as way that confounding variables are taken into account.

- Blocking : we collect data holding the value if confounding variable constant, the problem is identifying all variables.
- Randomization : we divide the data into two groups with the expectation that confounding factors cancel each other.