HW3

Jinhong Du

目录

1 Problem

Find the relationship between the chi-square statistic X^2_{trend} and the regression coefficients from regressing $\hat{p}_i$ on S_i.

2 Chi-Square Test for Trend in Binomial Proportions

Noticing an trend in the proportions with respect to group numbers, we would like to employ a specific test to detect such trends. For this purpose a score variable S_i is introduced to correspond to the ith group. The score variable can represent some particular numeric attribute of the group.

Suppose there are k groups and we want to test whether there is an increasing (or decreasing) trend in the proportion of "successes" p_i (the proportion of units in the first row of the ith group) as i increases.

Denote the number of successes in the ith group by x_i , the total number of units in the ith group by n_i, and the proportion of successes in the ith group by $\hat{p}_i = \frac{x_i}{n_i}$. Denote the total number of successes over all groups by $x = \sum_{i=1}^{k} x_i$, the total number of units over all groups by $n = \sum_{i=1}^{k} n_i$, the overall proportion of successes by $\overline{p} = \frac{x}{n}$, and the overall proportion of failures by $\overline{q} = 1 - \overline{p}$.

Category	1	2	$\cdots$	k
Frequency	x_1	x_2	$\cdots$	x_k
Total	n_1	n_2	$\cdots$	n_k
Score	S_1	S_2	$\cdots$	S_k

We wish to test the hypothesis

$$
\begin{aligned}
H_0 &: \text{There is no trend among the } p_i\text{'s.} \\
H_1 &: \text{The } p_i \text{ are an increasing or decreasing function of the } S_i, \\
&\quad \text{expressed in the form } p_i = \alpha + \beta S_i \text{ for some constants } \alpha, \beta.
\end{aligned}
$$

The statistic is given by

$$
\begin{aligned}
X^2_{trend} &= \frac{\left[\sum_{i=1}^{k} n_i(\hat{p}_i - \overline{p})(S_i - \overline{S})\right]^2}{\overline{p}(1-\overline{p})\left[\sum_{i=1}^{k} n_i S_i^2 - n\overline{S}^2\right]} \\
&= \frac{\left(\sum_{i=1}^{k} x_i S_i - x\overline{S}\right)^2}{\overline{pq}\left[\sum_{i=1}^{k} n_i S_i^2 - n\overline{S}^2\right]}
\end{aligned}
$$

where

$$
\overline{S} = \sum_{i=1}^{k} \frac{n_i}{n} S_i \qquad \overline{p} = \sum_{i=1}^{k} \frac{n_i}{n} p_i
$$

3 Relationship

It is easy to see that

$$X^2_{trend} = \frac{(\mathbf{S}^\top \mathbf{N}\mathbf{p})^2}{\frac{1}{n}\overline{pq}\mathbf{S}^\top \mathbf{N}\mathbf{S}}$$

where

$$\mathbf{S} = \begin{bmatrix} S_1 - \overline{S} \\ S_2 - \overline{S} \\ \vdots \\ S_k - \overline{S} \end{bmatrix} \qquad \mathbf{N} = \begin{bmatrix} \frac{n_1}{n} & 0 & \cdots & 0 \\ 0 & \frac{n_2}{n} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{n_k}{n} \end{bmatrix} \qquad \mathbf{p} = \begin{bmatrix} \hat{p}_1 - \overline{p} \\ \hat{p}_2 - \overline{p} \\ \vdots \\ \hat{p}_k - \overline{p} \end{bmatrix}$$

Assumption:

1. Suppose that $\overline{p}$ is the ground truth value of the proportion of successes in each Bernoulli trial with respect to the score $\overline{S}$.
2. Suppose that $X_i \sim Binomial(n_i, \overline{p})$ $(i = 1, 2, \cdots, k)$ and $X \sim Binomial(n, \overline{p})$, then

$$VarX = n\overline{pq}$$

For random varibale p, the proportion of successes in one group,

$$X = np$$

and therefore

$$\begin{aligned} Varp &= \frac{1}{n}\overline{pq} \\ &\triangleq \sigma^2 \end{aligned}$$

We can describe the above as a regression problem,

$$\mathbf{p}_{k\times 1} = \mathbf{S}_{k\times 1}\beta_{1\times 1} + \boldsymbol{\varepsilon}_{k\times 1} \tag{1}$$

where

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 & \cdots & \varepsilon_k \end{bmatrix}^\top \sim N(0, \mathbf{N}^{-1}\sigma^2)$$

$$\sigma^2 = \frac{1}{n}\overline{pq}$$

The constant term can be ignore since by assumption 1, the regression line goes through $(\overline{S} - \overline{S}, \overline{p} - \overline{p}) = (0, 0)$.

To obtain the weighted least squares estimator of (1), first we let $\mathbf{p}' = \mathbf{N}^{\frac{1}{2}}\mathbf{p}$, $\mathbf{S}' = \mathbf{N}^{\frac{1}{2}}\mathbf{S}$, $\boldsymbol{\varepsilon}' = \mathbf{N}^{\frac{1}{2}}\boldsymbol{\varepsilon}$, then it becomes a ordinary least square problem,

$$\mathbf{p}' = \mathbf{S}'\beta + \boldsymbol{\varepsilon}' \tag{2}$$

where

$$\boldsymbol{\varepsilon}' \sim N(0, \sigma^2)$$

This weighted least squares estimator of (1) is given by the ordinary least squares estimator of (2),

$$\begin{aligned}
\hat{\beta}_{WLS} &= \hat{\beta}'_{OLS} \\
&= (\mathbf{S}'^{\top}\mathbf{S}')^{-1}\mathbf{S}'^{\top}\mathbf{p}' \\
&= (\mathbf{S}^{\top}\mathbf{N}\mathbf{S})^{-1}\mathbf{S}^{\top}\mathbf{N}\mathbf{p} \\
&\sim N(\beta, (\mathbf{S}^{\top}\mathbf{N}\mathbf{S})^{-1}\mathbf{S}^{\top}\mathbf{N}(\mathbf{N}^{-1}\sigma^2)\mathbf{N}\mathbf{S}(\mathbf{S}^{\top}\mathbf{N}\mathbf{S})^{-1}) \\
&\overset{H_0}{\sim} N(0, (\mathbf{S}^{\top}\mathbf{N}\mathbf{S})^{-1}\sigma^2)
\end{aligned}$$

Therefore, under H_0 we have

$$\begin{aligned}
\frac{1}{\sigma}(\mathbf{S}^{\top}\mathbf{N}\mathbf{S})^{\frac{1}{2}}\hat{\beta} &= \frac{1}{\sigma}(\mathbf{S}^{\top}\mathbf{N}\mathbf{S})^{\frac{1}{2}}(\mathbf{S}^{\top}\mathbf{N}\mathbf{S})^{-1}\mathbf{S}^{\top}\mathbf{N}\mathbf{p} \\
&\overset{H_0}{\sim} N(0,1) \\
\frac{1}{\sigma^2}(\mathbf{S}^{\top}\mathbf{N}\mathbf{S})\hat{\beta}^2 &= \frac{1}{\sigma^2}\frac{(\mathbf{S}^{\top}\mathbf{N}\mathbf{p})^2}{\mathbf{S}^{\top}\mathbf{N}\mathbf{S}} \\
&\overset{H_0}{\sim} \chi^2(1)
\end{aligned}$$

which gives the form of X^2_{trend}.

4 Rcode

The following results comfirm the relationship.

```
x <- c(320, 1206, 1011, 463, 220)
n_list <- c(1742, 5638, 3904, 1555, 626)
n <- sum(n_list)
S <- c(1, 2, 3, 4, 5)
p <- x / n_list

p_all <- sum(p*n_list) / n
p_overall <- p - p_all
S_overall <- S - sum(S*n_list) / n
N <- diag(n_list / n)
Xtrend <- crossprod(S_overall, N %*% p_overall)^2 / (1/n * p_all * (1 - p_all) *
            crossprod(S_overall, N  %*% S_overall))

# X2trend calculated by weighted least squares
cat('X2trend is ',Xtrend)
```

```
## X2trend is  129.012
```

```
# X2trend calculated by built-in function
prop.trend.test(x, n_list, score = S)
```

```
##
##  Chi-squared Test for Trend in Proportions
##
## data:  x out of n_list ,
##  using scores: 1 2 3 4 5
## X-squared = 129.01, df = 1, p-value < 2.2e-16
```