Cochran's Theorem - (a quick tutorial)

Yang Feng

https://yangfengstat.github.io/

Importance of Cochran's Theorem

- Cochran's theorem tells us about the distributions of partitioned sums of squares of normally distributed random variables.
- Traditional linear regression analysis relies upon making statistical claims about the distribution of sums of squares of normally distributed random variables (and ratios between them)
  In the simple normal regression model:

$$\frac{SSE}{\sigma^2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi^2(n-2)$$

- Where does this come from?

Outline

- Establish the fact that the multivariate Gaussian sum of squares is $\chi^2(n)$ distributed
- Provide intuition for Cochran's theorem
- Prove a lemma in support of Cochran's theorem
- Prove Cochran's theorem
- Connect Cochran's theorem back to linear regression

χ^2 distribution

Theorem 1: Suppose Z_i are $i.i.d.$ $N(0,1)$, we have

$$\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$$

Proof:

- $Z_i^2 \sim \chi^2(1)$
- If $Y_1, \cdots, Y_n$ are i.i.d. random variables with moment generating functions (MGF) $m_{Y_1}(t), \cdots, m_{Y_n}(t)$. Then the moment generating function for $U = Y_1 + \cdots + Y_n$ is

$$m_U(t) = m_{Y_1}(t) \times m_{Y_2}(t) \cdots \times m_{Y_n}(t)$$

- MGF fully characterizes the distribution
- The MGF for $\chi^2(n)$ is $(1 - 2t)^{-n/2}$

Quadratic Forms and Cochran's Theorem

- Quadratic forms of normal random variables are of great importance in many branches of statistics
  - Least Squares
  - ANOVA
  - Regression Analysis
- General idea: Split the sum of the squares of observations into a number of quadratic forms where each corresponds to some cause of variation

Quadratic Forms and Cochrans Theorem

- The conclusion of Cochran's theorem is that, under the assumption of normality, the various quadratic forms are independent and χ^2 distributed.
- This fact is the foundation upon which many statistical tests rest.

Preliminaries: A Common Quadratic Form

- Let

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Lambda})$$

- Consider the quadratic form that appears in the exponent of the normal density

$$(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Lambda}^{-1}(\mathbf{X} - \boldsymbol{\mu})$$

- In the special case of $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Lambda} = \mathbf{I}$, this reduces to $\mathbf{X}'\mathbf{X}$ which by what we just proved we know is $\chi^2(n)$ distributed
- Let's prove it holds in the general case

Lemma 1

Let

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Lambda})$$

with $|\boldsymbol{\Lambda}| > 0$ and n is the dimension of $\mathbf{X}$, then

$$(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Lambda}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(n)$$

Proof

Let $\mathbf{Y} = \boldsymbol{\Lambda}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$, then we have $\mathbf{Y} \sim N(\mathbf{0}, \mathbf{I})$. Then,

$$(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Lambda}^{-1} (\mathbf{X} - \boldsymbol{\mu}) = \mathbf{Y}'\mathbf{Y} \sim \chi^2(n)$$

Cochran's Theorem

Let $X_1, X_2, \cdots, X_n$ be i.i.d. $N(0, \sigma^2)$- distributed random variables, and suppose that

$$\sum_{i=1}^{n} X_i^2 = Q_1 + Q_2 + \cdots + Q_k,$$

where $Q_1, Q_2, \cdots, Q_k$ are positive semi-definite quadratic forms in $X_1, X_2, \cdots, X_n$, i.e.,

$$Q_i = \mathbf{X}'\mathbf{A}_i\mathbf{X}, i = 1, 2, \cdots, k$$

Set $r_i = \text{rank}(\mathbf{A}_i)$. If $r_1 + r_2 + \cdots + r_k = n$, then

1. $Q_1, Q_2, \cdots, Q_k$ are independent.
2. $Q_i \sim \sigma^2 \chi^2(r_i)$

Several linear algebra results

- Let $\mathbf{X}$ be a normal random vector. The components of $\mathbf{X}$ are independent if and only if they are uncorrelated.
- Let $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, then $\mathbf{Y} = \mathbf{C}'\mathbf{X} \sim N(\mathbf{C}'\boldsymbol{\mu}, \mathbf{C}'\boldsymbol{\Lambda}\mathbf{C})$.
  - We can find an orthogonal matrix $\mathbf{C}$ such that $\mathbf{D} = \mathbf{C}'\boldsymbol{\Lambda}\mathbf{C}$ is a diagonal matrix. (Eigen Value Decomposition for Semi Positive Definite Matrix)
  - The components of $\mathbf{Y}$ will be independent and $\text{var}(Y_k) = \lambda_k$, where $\lambda_1, \cdots, \lambda_n$ are the eigenvalues of $\boldsymbol{\Lambda}$

Lemma 2

Let $X_1, X_2, \cdots, X_n$ be real numbers. Suppose that $\sum X_i^2$ can be split into a sum of positive semi-definite quadratic forms, that is

$$\sum X_i^2 = Q_1 + Q_2 + \cdots + Q_k$$

where $Q_i = \mathbf{X}'\mathbf{A}_i\mathbf{X}$ with $\text{rank}(A_i) = r_i$. If $\sum r_i = n$, then there exists an orthogonal matrix $\mathbf{C}$ such that, with $\mathbf{X} = \mathbf{CY}$, we have

$$\begin{aligned}
Q_1 &= Y_1^2 + Y_2^2 + \cdots + Y_{r_1}^2 \\
Q_2 &= Y_{r_1+1}^2 + Y_{r_1+2}^2 + \cdots + Y_{r_1+r_2}^2 \\
&\vdots \\
Q_k &= Y_{n-r_k+1}^2 + Y_{n-r_k+2}^2 + \cdots + Y_n^2
\end{aligned}$$

Remark

- Different quadratic forms contain different Y-variables and that the number of terms in each Q_i equals that rank, r_i, of Q_i
- The Y_i^2 end up in different sums, we'll use this to prove the independence of the different quadratic forms.
- Just prove for $k = 2$ case, the general case can be obtained by induction.

Proof

- For $k = 2$, we have $Q = \mathbf{X}'\mathbf{A}_1\mathbf{X} + \mathbf{X}'\mathbf{A}_2\mathbf{X}$
- There exists an orthogonal matrix $\mathbf{C}$ such that $\mathbf{C}'\mathbf{A}_1\mathbf{C} = \mathbf{D}$, where D is a diagonal matrix with eigenvalues of A_1.
- Since $rank(A_1) = r_1$, r_1 eigenvalues are positive and $n - r_1$ eigenvalues are 0.
- Suppose without loss of generality, the first r_1 eigenvalues are positive.
- Set $\mathbf{X} = \mathbf{C}\mathbf{Y}$, then we have $\mathbf{X}'\mathbf{X} = \mathbf{Y}'\mathbf{C}'\mathbf{C}\mathbf{Y} = \mathbf{Y}'\mathbf{Y}$.

Proof

- Therefore, $Q = \sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{r_1} \lambda_i Y_i^2 + \mathbf{Y'C'A_2CY}$
- Then, rearranging the terms we have

$$\sum_{i=1}^{r_1}(1-\lambda_i)Y_i^2 + \sum_{i=r_1+1}^{n} Y_i^2 = \mathbf{Y'C'A_2CY}$$

- Since $rank(\mathbf{A}_2) = r_2 = n - r_1$, we conclude that

$$\lambda_1 = \lambda_2 = \cdots = \lambda_{r_1} = 1$$

$$Q_1 = \sum_{i=1}^{r_1} Y_i^2, Q_2 = \sum_{i=r_1+1}^{n} Y_i^2$$

From this Lemma

- This lemma is about real numbers, not random variables
- It says that $\sum X_i^2$ can be split into a sum of positive semi-definite quadratic forms, then there is an orthogonal transformation $\mathbf{X} = \mathbf{CY}$ such that each of the quadratic forms has nice properties: Each Y_i appears in only one resulting sum of squares, which leads to the independence of the sum of squares.

Proof of Cochran's Theorem

1. Using the Lemma, $Q_1, \cdots, Q_k$ can be written using different Y_is, therefore, they are independent.
2. Furthermore, $Q_1 = \sum_{i=1}^{r_1} Y_i^2 \sim \sigma^2\chi^2(r_1)$. Other Q_is are the same.

Applications

- Sample variance is independent of the sample mean.
- Recall $SSTO = (n-1)s^2(Y)$,

$$SSTO = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$$

- Rearrange the term and express it in matrix format

$$\sum Y_i^2 = \sum (Y_i - \bar{Y})^2 + \frac{(\sum Y_i)^2}{n}$$

$$\mathbf{Y}'\mathbf{I}\mathbf{Y} = \mathbf{Y}'(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{Y} + \mathbf{Y}'(\frac{1}{n}\mathbf{J})\mathbf{Y}$$

- We know $\mathbf{Y}'\mathbf{I}\mathbf{Y} \sim \sigma^2\chi^2(n)$, $\text{rank}(\mathbf{I} - \frac{1}{n}\mathbf{J}) = n-1$ (next slide) and $\text{rank}(\frac{1}{n}\mathbf{J}) = 1$.
- As a result,

$$\sum (Y_i - \bar{Y})^2 \sim \sigma^2\chi^2(n-1)$$

$$\frac{(\sum Y_i)^2}{n} \sim \sigma^2\chi^2(1)$$

Rank of $\mathbf{I} - \frac{1}{n}\mathbf{J}$

Calculate $\text{rank}(\mathbf{I} - \frac{1}{n}\mathbf{J})$. First of all, we have

$$\text{rank}(\mathbf{I} - \frac{1}{n}\mathbf{J}) \geq \text{rank}(\mathbf{I}) - \text{rank}(\frac{1}{n}\mathbf{J}) = n - 1$$

On the other hand, since $(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{1} = \mathbf{0}$, we have

$$\text{rank}(\mathbf{I} - \frac{1}{n}\mathbf{J}) \leq n - 1$$

Therefore, we have

$$\text{rank}(\mathbf{I} - \frac{1}{n}\mathbf{J}) = n - 1$$

Another proof, noticing $\mathbf{I} - \frac{1}{n}\mathbf{J}$ is also idempotent and symmetric, therefore, $\text{rank}(\mathbf{I} - \frac{1}{n}\mathbf{J}) = \text{trace}(\mathbf{I}) - \text{trace}(\frac{1}{n}\mathbf{J}) = n - 1$

ANOVA

$$SSTO = \mathbf{Y}'[\mathbf{I} - \frac{1}{n}\mathbf{J}]\mathbf{Y}$$
$$SSE = \mathbf{Y}'[\mathbf{I} - \mathbf{H}]\mathbf{Y}$$
$$SSR = \mathbf{Y}'[\mathbf{H} - \frac{1}{n}\mathbf{J}]\mathbf{Y}$$

- Under the null hypothesis, when $\beta = 0$, we know $SSTO \sim \sigma^2\chi^2(n-1)$.
- From linear algebra: $\text{rank}(\mathbf{I} - \mathbf{H}) = n - p$ (next slide) and $\text{rank}(\mathbf{H} - \frac{1}{n}\mathbf{J}) = p - 1$.
- Then we have

$$SSE \sim \sigma^2\chi^2(n-p)$$
$$SSR \sim \sigma^2\chi^2(p-1)$$

- As a byproduct, $MSE = SSE/(n-p)$ is an unbiased estimator of σ^2, since the mean of $\chi^2(n-p)$ is $n-p$.

Rank of $\mathbf{I} - \mathbf{H}$

We have

$$\begin{aligned} \text{trace}(\mathbf{H}) &= \text{trace}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= \text{trace}((\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}) \\ &= \text{trace}(\mathbf{I}_p) \\ &= p \end{aligned}$$

Then,

$$\begin{aligned} \text{rank}(\mathbf{I} - \mathbf{H}) &= \text{trace}(\mathbf{I} - \mathbf{H}) \\ &= \text{trace}(\mathbf{I}) - \text{trace}(\mathbf{H}) \\ &= n - p \end{aligned}$$

Rank of $\mathbf{H} - \frac{1}{n}\mathbf{J}$

- First, since we have $\mathbf{H}\mathbf{1} = \mathbf{1}$ (This amounts to a multiple linear regression with the response always equal to 1, and therefore, the fitted value is still 1 because we can just use the constant to perfectly fit the model), then it is straightforward to check that $\mathbf{H} - \frac{1}{n}\mathbf{J}$ is an idempotent and symmetric matrix.
- Then, we have $\text{rank}(\mathbf{H} - \frac{1}{n}\mathbf{J}) = \text{trace}(\mathbf{H}) - \text{trace}(\frac{1}{n}\mathbf{J}) = p - 1$