Regression Analysis
Chapter 6: Testing and ANOVA

Shaobo Jin

Department of Mathematics

Residual Sum of Squares

Consider the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}.$$

Consider the conditions

1. $\mathrm{E}\left(Y \mid \boldsymbol{X} = \boldsymbol{x}\right) = \boldsymbol{x}^T\boldsymbol{\beta}$ with $\boldsymbol{\beta}_2 = \mathbf{0}$ is correctly specified,
2. $\boldsymbol{e} \mid \boldsymbol{X} \sim N\left(\mathbf{0}, \sigma^2\boldsymbol{I}\right)$.

The residual sum of squares is

$$\mathrm{RSS} = \boldsymbol{y}^T\left(\boldsymbol{I} - \boldsymbol{H}\right)\boldsymbol{y} = \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^T\left(\boldsymbol{I} - \boldsymbol{H}\right)\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right).$$

Intuitively speaking, RSS is a sum of square normal random variables and

$$\frac{\mathrm{RSS}}{\sigma^2} \sim \chi^2\left(n - p\right).$$

Restrictions

Suppose that we want to test

$$H_0: \boldsymbol{L}\boldsymbol{\beta} = \mathbf{0} \quad \text{vs} \quad H_1: \boldsymbol{L}\boldsymbol{\beta} \neq \mathbf{0}.$$

We can fit two models:

1. Model 1 ignores the restriction $\boldsymbol{L}\boldsymbol{\beta} = \mathbf{0}$. Its OLS estimator is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

   Denote its residual sum of squares by RSS_1.

2. Model 0 has the restriction $\boldsymbol{L}\boldsymbol{\beta} = \mathbf{0}$, that is, the estimator must satisfy $\boldsymbol{L}\hat{\boldsymbol{\beta}} = \mathbf{0}$. The estimator (without proof) is

$$\hat{\boldsymbol{\beta}}_L = \hat{\boldsymbol{\beta}} - \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{L}^T\left[\boldsymbol{L}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{L}^T\right]^{-1}\boldsymbol{L}\hat{\boldsymbol{\beta}}.$$

   Denote its residual sum-of-squares by RSS_0.

Residual Sums of Squares

Consider the assumptions

1. $\mathrm{E}\left(Y \mid \boldsymbol{X}=\boldsymbol{x}\right)=\boldsymbol{x}^T\boldsymbol{\beta}$ where $\boldsymbol{L}\boldsymbol{\beta}=\boldsymbol{0}$,
2. $\boldsymbol{e} \mid \boldsymbol{X}=\boldsymbol{x} \sim N\left(\boldsymbol{0}, \sigma^2\boldsymbol{I}\right)$.

Intuitively speaking,

$$\begin{aligned}\frac{\mathrm{RSS}_0}{\sigma^2} &\sim \chi^2\left(n-p_0\right),\\ \frac{\mathrm{RSS}_1}{\sigma^2} &\sim \chi^2\left(n-p\right),\end{aligned}$$

where

$$p_0 = \operatorname{rank}\left(\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{L}\end{bmatrix}\right)-\operatorname{rank}\left(\boldsymbol{L}\right).$$

Special Case: I

Suppose that we want to test $\beta_1 = 0$ in a model with intercept and $p-1$ covariates.

- It is the same as

$$\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \end{bmatrix}_{1\times p} \boldsymbol{\beta}_{p\times 1} = 0.$$

- For the model without restriction, $\mathrm{RSS}_1 \sim \chi^2(n-p)$.
- For the model with restriction,

$$\mathrm{RSS}_0 \sim \chi^2(n-p_0),$$

where

$$p_0 = \operatorname{rank}\left(\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{L} \end{bmatrix}\right) - \operatorname{rank}(\boldsymbol{L}) = p-1.$$

Special Case: II

Suppose that we want to test if the model only contains the intercept (null model) with intercept and $p-1$ covariates.

- It is the same as

$$\begin{bmatrix}\mathbf{0}_{p-1\times 1} & \boldsymbol{I}_{p-1\times p-1}\end{bmatrix}\boldsymbol{\beta}_{p\times 1} = \mathbf{0}_{p-1\times 1}.$$

- For the model without restriction, $\mathrm{RSS}_1 \sim \chi^2(n-p)$.
- For the model with restriction,

$$\mathrm{RSS}_0 \sim \chi^2(n-p_0),$$

where

$$p_0 = \mathrm{rank}\left(\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{L}\end{bmatrix}\right) - \mathrm{rank}(\boldsymbol{L}) = p-(p-1) = 1.$$

F-Test

Consider the assumptions

1. $\mathrm{E}\left(Y \mid \boldsymbol{X}=\boldsymbol{x}\right)=\boldsymbol{x}^T\boldsymbol{\beta}$ where $\boldsymbol{L}\boldsymbol{\beta}=\mathbf{0}$,
2. $\boldsymbol{e} \mid \boldsymbol{X}=\boldsymbol{x} \sim N\left(\mathbf{0}, \sigma^2\boldsymbol{I}\right)$.

Intuitively speaking,

$$\frac{\mathrm{RSS}_0}{\sigma^2} \sim \chi^2\left(n-p_0\right), \qquad \frac{\mathrm{RSS}_1}{\sigma^2} \sim \chi^2\left(n-p\right),$$

and

$$\frac{\mathrm{RSS}_0-\mathrm{RSS}_1}{\sigma^2} \sim \chi^2\left(p-p_0\right).$$

Then, we may have

$$F=\frac{\left(\mathrm{RSS}_0-\mathrm{RSS}_1\right)/\left(p-p_0\right)}{\mathrm{RSS}_1/\left(n-p\right)} \sim F\left(p-p_0, n-p\right).$$

Special Cases

1. We want to test $H_0 : \beta_1 = 0$ with $p_0 = p - 1$. Then,

$$F = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1 / (n-p)} \quad \sim \quad F\left(1, n-p\right),$$

   which will be the same as the squared t-value.

2. We want to test if the model only contains the intercept (null model), with $p_0 = 1$. Then,

$$F = \frac{\left(\mathrm{RSS}_0 - \mathrm{RSS}_1\right) / (p-1)}{\mathrm{RSS}_1 / (n-p)} \quad \sim \quad F\left(p-1, n-p\right),$$

   is the F-statistics reported in R.

F-Statistic and R^2

In the special case where we want to test all regression slopes are zero, i.e., the model is $Y = \beta_0 + e$, the F-statistic becomes

$$F = \frac{\left[\sum_{i=1}^{n} (y_i - \bar{y})^2 - \mathrm{RSS}_1\right] / (p-1)}{\mathrm{RSS}_1 / (n-p)} \sim F(p-1, n-p).$$

We can rewrite F as

$$F = \frac{n-p}{p-1}\left(\frac{1}{1-R^2} - 1\right)$$

or

$$R^2 = \left[1 + \frac{n-p}{(p-1)F}\right]^{-1}.$$

This means that if F is large, then R^2 is also large.

Compare Several Population Means

Sometimes we have several populations as

$$\begin{aligned} \text{Population } 1: &\quad y_{11}, y_{12}, ..., y_{1n_1} \\ \text{Population } 2: &\quad y_{21}, y_{22}, ..., y_{2n_2} \\ &\vdots \\ \text{Population } g: &\quad y_{g1}, y_{g2}, ..., y_{gn_g} \end{aligned}$$

Analysis of variance (ANOVA) can be used to investigate whether the population means are the same.

Assumptions

We need the following assumptions:

1. $Y_{\ell 1}$, $Y_{\ell 2}$, ..., $Y_{\ell n_\ell}$ is a random sample of size n_ℓ , from a population with mean μ_ℓ, $\ell = 1, 2, ..., g$.
2. Then random sample from different populations are independent.
3. Each population is multivariate normal.
4. All populations have a common varinace σ^2.

ANOVA Model

The ANOVA model for comparing g population means is

$$Y_{\ell j} = \underset{\text{overall mean}}{\mu} + \underset{\text{treatment effect}}{\tau_\ell} + \underset{\text{random eror}}{e_{\ell j}},$$

for $j = 1, 2, ..., n_\ell$ and $\ell = 1, 2, ..., g$, where $e_{\ell j} \sim N\left(0, \sigma^2\right)$.

- The ANOVA model is unidentified. We often impose the identification restriction $\sum_{\ell=1}^{g} n_\ell \tau_\ell = 0$, or equivalent.
- Our H_0 is$\tau_\ell = 0$ for all ℓ and H_1 is some τ_ℓ is not zero.
- The sample decomposition is

$$y_{\ell j} = \underset{\text{overall sample mean}}{\bar{y}} + \underset{\text{estimated treatment effect}}{\bar{y}_\ell - \bar{y}} \\ \underset{\text{residual}}{+ y_{\ell j} - \bar{y}_\ell}.$$

Sum of Squares Decomposition

The total sum of squares and cross products satisfy

$$\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}\left(y_{\ell j}-\bar{y}\right)^2 = B+W,$$

where

$$W = \sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}\left(y_{\ell j}-\bar{y}_\ell\right)^2$$

is the within sum of squares and cross products, and

$$B = \sum_{\ell=1}^{g} n_\ell\left(\bar{y}_\ell-\bar{y}\right)^2$$

is the between sum of squares and cross products.

ANOVA Table

Source of variation	Sum of Squares	Degrees of Freedom
Treatment	B	$g-1$
Residual	W	$\sum_{\ell=1}^{g}(n_\ell - 1) = n-g$
Total	$B+W$	$\sum_{\ell=1}^{g} n_\ell - 1 = n-1$

If $\tau_\ell = 0$ for all ℓ , then

$$\frac{B/(g-1)}{W/(n-g)} \sim F(g-1, n-g).$$

F Test in ANOVA and Linear Regression

We can write the ANOVA model

$$Y_{\ell j} = \mu + \tau_\ell + e_{\ell j},$$

as

$$Y_i = \mu + \sum_{\ell=1}^{g} \beta_\ell 1\left(\text{individual } i \text{ belongs to group } \ell\right) + e_i.$$

Hence, ANOVA is just the F-test in linear regression!

Multiple Testing

Suppose that we have performed an ANOVA analysis and show that some τ_ℓ's are not zero. Now, we need to decide which ones are not zero.

- One idea is to make pairwise comparison using t-test or other methods: G1 versus G2, G1 versus G3, etc.
- Suppose that $g = 10$ and we choose $\alpha = 0.05$.
- For simplicity, we assume that all tests are independent and $\tau_\ell = 0$ for all ℓ.
- Then, our type I error is out of control!

```
# k = 10
1 - dbinom(0, 10 * (10 - 1) / 2, 0.05)

## [1] 0.9005597
```

Multiple Testing

There are many methods that allow us to avoid such exploding type I error probability.

1. Family-wise error rate (FWER): some methods (e.g, Bonferroni, Holm and Tukeys HSD) focus on the probability of at least one false positive result (type I error).
    - It becomes conservative if we have many tests, i.e., low power and it is hard to discover true positive results.
    - It is often used if we want to avoid false positive results.
2. False discovery rate (FDR): some methods (e.g., Benjamini-Hochberg) focus on the probability of false positive among all positive results.
    - We have a better chance to discover some treatment effects.
    - It is often used if we want to discover effects but are ready to accept the risk of some false positive results.

Two-Way ANOVA

If we have more two factors, we can perform a two-way ANOVA:

$$y_i = \mu + \tau_\ell + \alpha_j + \epsilon_i$$

- τ_ℓ is the effect of group ℓ of one factor (e.g., effect of gender),
- α_j is the effect of group j of another factor (e.g., effect of different vaccines).

If we want to include an interaction between two factors, we can consider

$$y_i = \mu + \tau_\ell + \alpha_j + \gamma_{\ell j} + \epsilon_i,$$

where $\gamma_{\ell j}$ is the interaction.