STAT 408
Applied Regression Analysis

Miles Xi

Department of Mathematics and Statistics

Loyola University Chicago

Fall 2022

Statistical Inference in Multiple Linear Regression

Motivation

- Similar to simple linear model, the least square estimation

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

is still a random variable due to the randomness of sample data

- The statistical inference in multiple linear regression can
  1. Examine the distribution of $\hat{\beta}$
  2. Test the significance of single parameter β_j
  3. Jointly test the significance of multiple parameters βs
  4. Test the relationship among parameters, e.g., $\beta_j = \beta_k$

Distribution of Error

- We still start from our classical assumption "error term ϵ follows a normal distribution and different ϵ_i's are independent"

$$\epsilon_i \sim N\left(0, \sigma^2\right)$$

$$\epsilon_i \perp \epsilon_j \text{ for } i \neq j$$

- Using matrix notation

$$\varepsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

the error vector ε is a multivariate normal distribution with zero covariance

$$\varepsilon \sim N\left(\mathbf{0}, \sigma^2 I\right)$$

where **0** is a zero vector, σ^2 is common variance, I is identity matrix

Distribution of Response Variable

- In our linear model

$$y = X\beta + \varepsilon$$

both X and β are fixed, only ε is a random variable

- y is the sum of a "constant" and a multivariable normal random variable
- y is also a random variable, and follows a multivariable normal distribution as ε

- To show the complete distribution of y, we need to know its expectation and variance

Distribution of response variable

- Recall the expectation and variable operation in scaler case

$$E(a + X) = a + E(X)$$

$$V(a + X) = V(X)$$

where a is a constant and X is a random variable

- The sample rule applies to random variables in matrix form

$$E(y) = E(X\beta + \ \varepsilon) = E(X\beta) + E(\varepsilon) = X\beta$$

$$V(y) = V(X\beta + \varepsilon) = V(\varepsilon) = \sigma^2 I$$

Distribution of response variable

- Therefore, response variable y follows a multivariate normal distribution

$$y \sim N\left(X\beta, \sigma^2 I\right)$$

- Let's take a look at our main focus

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- Since $\hat{\beta}$ is a random variable, what distribution do you think it follows?

Distribution of least square estimator

- Since $\hat{\beta}$ is still a "constant" multiplied by a normal random variable, $\hat{\beta}$ also follows a multivariate normal distribution

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

- Recall if x is a scaler and a is a constant

$$V(ax) = a^2 V(x)$$

- Similar, in matrix form

$$V(MX) = MV(X)M^T$$

where M is a matrix and X is a multivariate random variable (vector of rv)

Distribution of least square estimator

- With this rule, the variance of $\hat{\beta}$ is

$$V(\hat{\beta}) = V\left(\left(X^T X\right)^{-1} X^T y\right) = \left(X^T X\right)^{-1} X^T V(y) [\left(X^T X\right)^{-1} X^T]^T$$

$$= \left(X^T X\right)^{-1} X^T \sigma^2 I [\left(X^T X\right)^{-1} X^T]^T = \sigma^2 \left(X^T X\right)^{-1} X^T X \left(X^T X\right)^{-1} = \left(X^T X\right)^{-1} \sigma^2$$

in which we use

$$V(y) = \sigma^2 I$$

$$(AB)^T = B^T A^T$$

$$(A^{-1})^T = \left(A^T\right)^{-1}$$

$$AA^{-1} = I \quad AI = A$$

Distribution of least square estimator

- To summarize, the least square estimation $\hat{\beta}$ follows a multivariate normal distribution as

$$\hat{\beta} \sim N(\beta, \left(X^T X\right)^{-1} \sigma^2)$$

1. $E\left(\hat{\beta}\right) = \beta$ indicates that the least square estimate is an <u>unbiased</u> estimator of model parameter β
2. Each individual $\hat{\beta}_j$ follows a normal distribution
3. $E\left(\hat{\beta}_j\right) = \beta_j$
4. $V\left(\hat{\beta}_j\right)$ is the jth diagonal element in the covariance matrix $\left(X^T X\right)^{-1} \sigma^2$
5. $Cov(\hat{\beta}_j, \hat{\beta}_k)$ is the jkth and kjth off-diagonal element in the covariance matrix $\left(X^T X\right)^{-1} \sigma^2$

Distribution of least square estimator

- In practice, we don't know σ^2 and has to estimate it

$$\hat{\sigma}^2 = \frac{RSS}{n-p}$$

- We can understand $\hat{\sigma}^2$ as "the average variation not explained by model"

Hypothesis Tests to Compare Models

- Given several predictors in the data, we might wonder if all are needed
- Consider a larger model, Ω, and a smaller model, ω, which consists of a subset of the predictors that are in Ω
  - We prefer ω if two model fits are "not very different" (for simplicity)
  - We prefer Ω if the large model fit is "improved" over small model

- Statistically, the previous judgement is a <u>hypothesis test</u>

$$H_0: \omega \text{ is better}$$

$$H_a: \Omega \text{ is better}$$

- How can we design a test statistic for this hypothesis test?

Hypothesis Tests to Compare Models

- The RSS is still a good choice, but like before, we need to consider the model complexity
- We use the follow F statistic

$$F = \frac{(\mathrm{RSS}_\omega - \mathrm{RSS}_\Omega)/(p-q)}{\mathrm{RSS}_\Omega/(n-p)}$$

where p = number of parameters in Ω, q = number of parameters in ω

- With the assumption of normal errors, under H_0 , $F \sim F_{p-q,n-p}$
- We reject H_0 if $F > F_{p-q,n-p}^{(\alpha)}$, where α is significant level

Hypothesis Tests to Compare Models

$$F = \frac{(\mathrm{RSS}_\omega - \mathrm{RSS}_\Omega)/(p-q)}{\mathrm{RSS}_\Omega/(n-p)}$$

- F statistics can be understood as the ratio of "average" residuals per predictor
- Remember that

$$df_\Omega = n - p$$

$$df_\omega = n - q$$

- Then the F statistics can be rewritten as

$$F = \frac{(\mathrm{RSS}_\omega - \mathrm{RSS}_\Omega)/(df_\omega - df_\Omega)}{\mathrm{RSS}_\Omega/df_\Omega}$$

Example: Test of All Predictors

- Let the full model Ω be $y = X\beta + \varepsilon$

- Let the small model ω be $y = \beta + \varepsilon$

- We call $y = \beta + \varepsilon$ "null model" and estimate β by $\bar{y}$ (least square estimation)

- If we want to test if the full model is better than the null model, we can use the following hypothesis test:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

H_a: At lease some predictors $\beta \neq 0$

Example: Test of All Predictors

```
# F test for model comparison
lm.model <- lm(insulin~., data = pima)
null.model <- lm(insulin~1, data=pima)
anova(null.model, lm.model)

# check coding 3.r for manually conducting F test
```

Example: Testing a Pair of Predictors

- Suppose we want to know whether the glucose or bmi had any relation to the response

- In other words

$$H_0: \beta_{glucose} = \beta_{bmi} = 0$$
$$H_a: \beta_{glucose} \neq 0 \text{ or } \beta_{bmi} \neq 0$$

Example: Testing a Pair of Predictors

```
# Test a Pair of Predictors
pima <- read.csv('pima.csv')
pima <- pima[complete.cases(pima), ]
lm.model <- lm(insulin~., data = pima)
small.model <- lm(insulin~pregnant+diastolic+triceps+diabetes+age+test,
data=pima)
anova(small.model, lm.model)
```

Example: Testing a Relationship

- We want to test whether the glucose and bmi have the same effect on insulin

$$H_0: \beta_{glucose} = \beta_{bmi}$$
$$H_a: \beta_{glucose} \neq \beta_{bmi}$$

- It is equivalent to say that we can merge glucose and bmi in linear model
  - Merging generates a small model

Example: Testing a Relationship

```
# Test relationship of two predictors
pima <- read.csv('pima.csv')
pima <- pima[complete.cases(pima), ]
lm.model <- lm(insulin~., data = pima)
small.model <- lm(insulin~I(glucose+bmi)+pregnant+diastolic+triceps+diabetes+age+test, data=pima)
anova(small.model, lm.model)
```

Example: Testing a Subspace

- Another example is to test whether a parameter can be set to a particular value

$$H_0: \beta_{glucose} = 2$$
$$H_a: \beta_{glucose} \neq 2$$

```
# Test a subspace
pima <- read.csv('pima.csv')
pima <- pima[complete.cases(pima), ]
lm.model <- lm(insulin~., data = pima)
small.model <- lm(insulin~offset(2*glucose)+bmi+pregnant+diastolic+triceps+diabetes+age+test, data=pima)
anova(small.model, lm.model)
```

When the F Test not Working

1. We cannot test a non-linear hypothesis, for example

$$H_0: \beta_j \beta_k = 1$$

2. We cannot compare models that are <u>not nested</u> using an F-test

Model one: glucose + bmi

Model two: glucose + pregnant + age

3. The models we compare use different datasets

Permutation Test

- The previous F test and t test all rely on the assumption of normal errors

$$\varepsilon \sim N\left(\mathbf{0}, \sigma^2 I\right)$$

- How can we perform hypothesis test <u>if this assumption is violated</u>?
- Recall our F statistics

$$F = \frac{(\mathrm{RSS}_\omega - \mathrm{RSS}_\Omega)/(df_\omega - df_\Omega)}{\mathrm{RSS}_\Omega/df_\Omega}$$

- Intuitively, if the response truly is related to predictors (full model is preferable), then F statistics should be "large"; otherwise it is "small"
- <u>This result is correct without the normal error assumption</u>

Permutation Test

- Our logic is
    1. Suppose that the H_a is preferable (full model is right), then the F statistic should be "large"
    2. If we randomly shuffle the response Y, then we break the relationship between response and predictors in each shuffled dataset
    3. The F statistics calculated under those shuffled datasets should be "small", because the model of <u>shuffled</u> data is wrong

- Let $\{F_j\}$, j=1,2, …, M be the set of those F statistics calculated based on each shuffled dataset (M is the number of random shuffle)
- If H_a is preferable, most F_js should be less than the original F, then we reject H_0
- If H_0 is preferable, then F is not different from other F_js , we cannot reject H_0

Permutation Test: Example

- We use the following method to conduct <u>permutation test</u>

$$\text{Permutation p-value} = \frac{Number\ of\ F_j s\ greater\ than\ F}{Number\ of\ shuffling\ M}$$

- We can use the value of this ratio as the p-value

- Small permutation p-value indicates most (shuffled) F_js are less than the (unshuffled) F statistic, thus we prefer full model

Permutation Test: Example

- Let's see one example. We first fit the full model and calculate the p-value of regular F test

- Then we conduct a permutation test to compare the permutation p-value and regular p-value

- See coding 3.r for example code

Permutation Test: One Predictor

- We can also use permutation test to <u>test one predictor</u>
- The idea is to break this predictor's relation with the response
- Instead of F statistic, we use t statistic $\hat{\beta}/se(\hat{\beta})$

- The method is straightforward:
  1. Randomly shuffle that predictor M times
  2. Each time, calculate shuffled t statistics t_js (absolute value)
  3. Permutation p-value = $\dfrac{Number\ of\ t_j s\ greater\ than\ original\ t\ (abs)}{Number\ of\ shuffling\ M}$

- See coding 3.r for example code

Confidence Interval for β

- Confidence intervals (CIs) provide another way to measure the uncertainty in the estimates of β
- Recall that $\hat{\beta}$ follows a multivariate normal distribution

$$\hat{\beta} \sim N(\beta, (X^T X)^{-1} \sigma^2)$$

- And we estimate σ^2 by

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n-p} = \frac{\text{RSS}}{n-p}$$

- Then any $\hat{\beta}_i$ also follows a univariate normal distribution

$$\hat{\beta}_i \sim N(\beta_i, se(\hat{\beta}_i))$$

where $se(\hat{\beta}_i)$ is the square root of the *i*th diagonal element in covariance matrix $\left(X^T X\right)^{-1} \sigma^2$

Confidence Interval for β

- Recall that in normal distribution, the critical value and standard deviation σ determines the probability

$$\hat{\beta}_i \sim N\left(\beta_i, se\left(\hat{\beta}_i\right)\right) \rightarrow \frac{\hat{\beta}_i - \beta}{se\left(\hat{\beta}_i\right)} \sim N(0,1)$$

- Therefore, we have

$$P\left(-z^{0.025} < \frac{\hat{\beta}_i - \beta}{se\left(\hat{\beta}_i\right)} < z^{0.025}\right) = 0.95$$

$$P\left(\hat{\beta}_i - z^{0.025} * se\left(\hat{\beta}_i\right) < \beta_i < \hat{\beta}_i + z^{0.025} * se\left(\hat{\beta}_i\right)\right) = 0.95$$

Confidence Interval for β

- Therefore, the 95% confidence interval for true parameter β_i is

$$\hat{\beta}_i \pm z^{0.025} * se(\hat{\beta}_i)$$

where i = 0, 1, 2,…, p – 1, and $z^{0.025} = 1.96$

- We can switch $z^{0.025}$ to other critical values for different confidence levels

- See coding 3.r for example code

Bootstrap Confidence Interval

- The previous construction of CI also relies on the normality assumption

$$\varepsilon \sim N\left(\mathbf{0}, \sigma^2 I\right)$$

- We can construct CIs without such assumptions
- Remember the uncertainty of $\hat{\beta}$ comes from the fact that we only have sample instead of the population
- If we can draw multiple samples from the population, and obtain one $\hat{\beta}$ for each sample, then we will have the <u>empirical distribution and CI of $\hat{\beta}$</u>
- One way to implement this is to <u>sample from our sample</u>, but with replacement to keep sample size same, which is called <u>bootstrap</u>

Bootstrap Confidence Interval

1. Randomly draw a sample (X, Y)* of size n with replacement from current data (X, Y)
2. Fit a linear model on (X, Y)* and obtain the estimated parameter $\hat{\beta}^*$
3. Repeat the process by multiple times and save all the $\hat{\beta}^*$s
4. Construct empirical CIs and standard deviation based on all the $\hat{\beta}^*$s

Gauss – Markov Theorem

- Recall that the least square estimate is an <u>unbiased</u> estimator of model parameter β

$$E(\hat{\beta}) = \beta$$

- Also, $\hat{\beta}$ is a <u>linear estimator</u> because it is essentially a linear transformation of response *y*

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- With the normality assumption for random error $\varepsilon \sim N(\mathbf{0}, \sigma^2 I)$, we have the Gauss–Markov Theorem:

<u>The least squares $\hat{\beta}$ estimator has the lowest variance within the class of linear unbiased estimators or “Best unbiased linear estimator (BLUE)”</u>