Identifiability and Inference for Linear Model

Lin Wang

Matrix representation of linear model

- We can write a linear model in matrix form

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

where $y = (y_1, \dots, y_n)^T, \beta = (\beta_1, \dots, \beta_{p-1})^T, \varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T,$

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix}$$

- The least-squares estimate of β is

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

Identifiability

- The least square estimate $\hat{\beta} = (X^T X)^{-1} X^T y$ relies on the the successful inverse of $X^T X$.
- Actually, $X^T X$ is not only related to identifiability. Later we will see that this matrix is the key in model estimation.
- If $X^T X$ is singular (not full rank), then there will be infinitely many solutions to the normal equations

$$X^T X \hat{\beta} = X^T y$$

- In this case, the model is unidentifiable
- Unidentifiability will occur if X's columns are linearly dependent (collinearity)
  - A person's weight is measured both in pounds and kilos
  - In a clinical trial, females are all assigned to treated group and males are all assigned to control group (gender is confounded with the treatment)

- We want to avoid collinearity between predictors in the data.

- Suppose we create a new variable for the Galápagos dataset

```
> gala$Adiff=gala$Area-gala$Adjacent
> lmod=lm(Species~Area+Elevation+Nearest+Scruz+Adjacent+Adiff, gala)
> summary(lmod)

Call:
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent +
    Adiff, data = gala)

Residuals:
     Min       1Q   Median       3Q      Max
-111.679  -34.898   -7.862   33.460  182.584

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
Adiff              NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,    Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

- More severe issue happens if we are close to unidentifiability
- Suppose we add a small random perturbation to new variable Adiff by adding a random variate from a uniform distribution U $[-0.0005, 0.0005]$
- This will break the exactly linear relationship, but it is still close to perfect

```
> set.seed(123)
> Adiffe <- gala$Adiff+0.001*(runif(30)-0.5)
```

```
> lmod <- lm(Species ~ Area+Elevation+Nearest+Scruz +Adjacent+Adiffe,
    gala)
> sumary(lmod)
                Estimate  Std. Error t value  Pr(>|t|)
(Intercept)       3.2964     19.4341    0.17      0.87
Area         -45122.9865  42583.3393   -1.06      0.30
Elevation         0.3130      0.0539    5.81 0.0000064
Nearest           0.3827      1.1090    0.35      0.73
Scruz            -0.2620      0.2158   -1.21      0.24
Adjacent      45122.8891  42583.3406    1.06      0.30
Adiffe        45122.9613  42583.3381    1.06      0.30

n = 30, p = 7, Residual SE = 60.820, R-Squared = 0.78
```

- All parameters are estimated, but the <u>standard errors</u> are very large
- We cannot estimate them in a stable way
  - For any "new" data from the same population, the corresponding "new" $\hat{\beta}$ will be very different

Inference

- What we have done now is to just estimate $\hat{\beta}$ from a **random sample** of observations
- The value of $\hat{\beta}$ is may change if a different sample is observed
- Therefore, $\hat{\beta}$ is a random variable
- Therefore, we hope to know how $\hat{\beta}$ varies when a different sample is observed
- We may want to test if the true β is in fact zero (why) or provide a confidence interval for the true β

- To do that, we need **assumptions** about the distribution of ε

Distribution Assumption in Linear Model

Assumptions

- The error ε follows a normal distribution

$$\varepsilon_i \sim N\left(0, \sigma^2\right), i = 1, 2, \ldots, n$$

- All errors of different observations follow a same normal distribution
- All errors are independent

$$\varepsilon_i \perp \varepsilon_j \text{ for } i \neq j$$

Note: These are assumptions, not facts, so they may not hold in reality. Therefore, we have to assess the assumptions in each application.

- With these assumptions, we have

$$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim N(0, \sigma^2 I_n)$$

Distribution of $\hat{\beta}$

- $y = X\beta + \varepsilon \sim N\left(X\beta, \sigma^2 I_n\right)$

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

- The variance term in the distribution explains why we have big standard error for coefficient estimation when two columns of X are close to linearly dependent
- Then each $\hat{\beta}_i$ follows a univariate normal distribution

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

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

Estimation of σ^2

The estimation of σ^2:

$$\hat{\sigma}^2 = \frac{SS_{Error}}{n-p} = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-p}$$

Why?

Residual-based estimator

- Residuals:

$$e_i = y_i - \hat{y}_i$$

- Residual sum of squares:

$$SS_{\text{Error}} = \sum_{i=1}^{n} e_i^2 \qquad \frac{SS_{\text{Error}}}{\sigma^2} \sim \chi^2_{n-p}$$

- Under the model assumptions:

$$\mathbb{E}(SS_{\text{Error}}) = (n-p)\sigma^2$$

Unbiased estimator

$$\hat{\sigma}^2 = \frac{SS_{\text{Error}}}{n-p}$$

Hypothesis Test: Each predictor

- One related question we want to ask is “Can one particular predictor be dropped from the model?”
- To answer this question, we need to test one predictor

$$H_0: \beta_i = 0$$
$$H_1: \beta_i \neq 0$$

- t-test: under H_0

$$t = \frac{\hat{\beta}_i - \beta_i}{\widehat{se(\hat{\beta}_i)}}$$

follows a t-distribution with *n-p* df.

Example: Testing one Predictor

```
> sumary(lmod)
              Estimate  Std. Error t value   Pr(>|t|)
(Intercept)    7.06822    19.15420    0.37     0.7154
Area          -0.02394     0.02242   -1.07     0.2963
Elevation      0.31946     0.05366    5.95  0.0000038
Nearest        0.00914     1.05414    0.01     0.9932
Scruz         -0.24052     0.21540   -1.12     0.2752
Adjacent      -0.07480     0.01770   -4.23     0.0003

n = 30, p = 6, Residual SE = 60.975, R-Squared = 0.77
```

Confidence Interval for β_i

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

$$\hat{\beta}_i \pm t_{n-p}^{0.025} * \widehat{se(\hat{\beta}_i)}$$

where i = 0, 1, 2,…, p – 1, and $t_{n-p}^{0.025}$ is the critical value of t-dist of n-p df with $\alpha = .05$

- In general, the $1 - \alpha$ confidence interval for true parameter β_i is

$$\hat{\beta}_i \pm t_{n-p}^{\alpha/2} * se(\hat{\beta}_i)$$

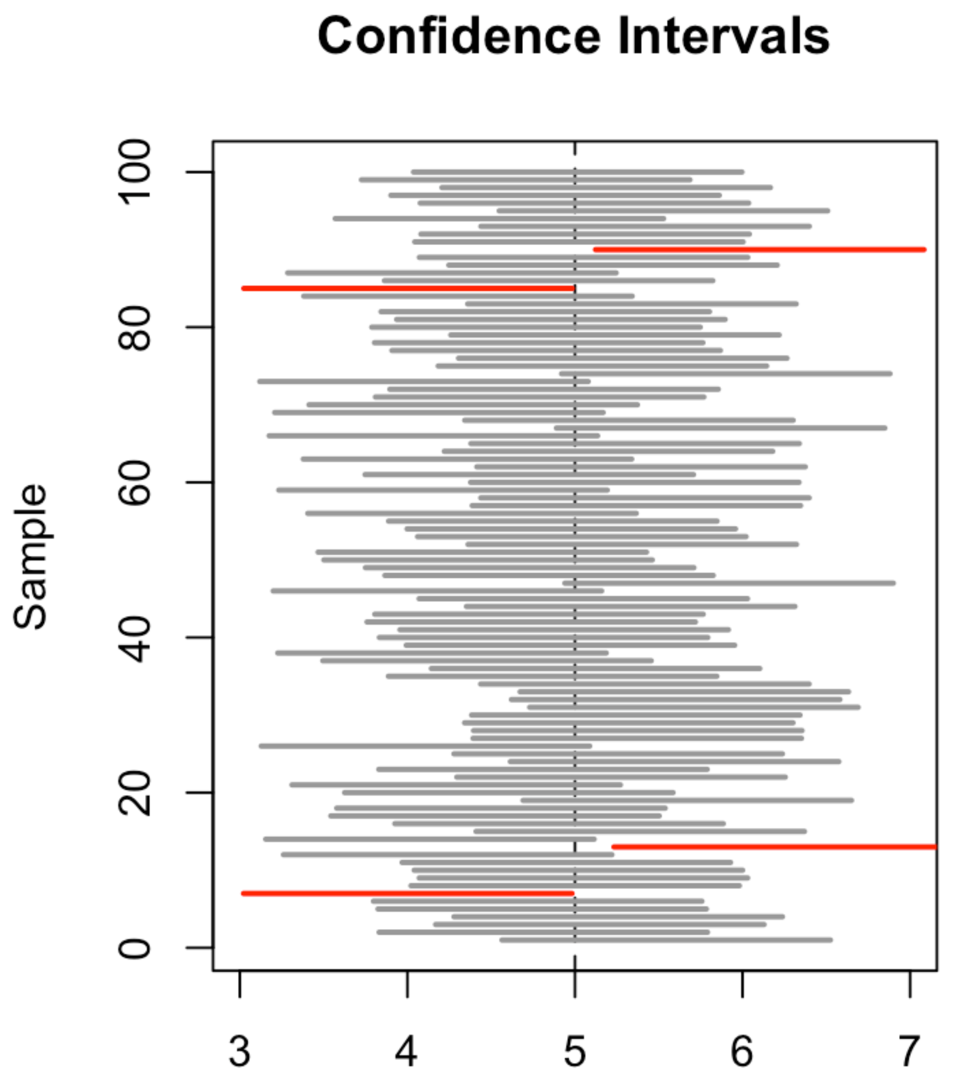
Confidence Interval for β_i

<u>Interpretation of CI:</u>

The construction of confidence interval relies on the data we have

Each dataset will give us one CI

Among all CIs constructed by different samples, (roughly) 95% of them cover the true β



Example

```
> lmod <- lm(Species ~ Area + Elevation + Nearest + Scruz  + Adjacent,
      gala)
> sumary(lmod)
            Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  7.06822   19.15420     0.37    0.7154
Area        -0.02394    0.02242    -1.07    0.2963
Elevation    0.31946    0.05366     5.95 0.0000038
Nearest      0.00914    1.05414     0.01    0.9932
Scruz       -0.24052    0.21540    -1.12    0.2752
Adjacent    -0.07480    0.01770    -4.23    0.0003

n = 30, p = 6, Residual SE = 60.975, R-Squared = 0.77
```

- We want to construct a 95% CIs for β_{Area}
- We need 97.5% percentiles of the t distribution with df=30-6=24

Hypothesis Test: Overall Significance of the Regression

Hypotheses

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \quad \text{(no linear relationship)}$$

$$H_A : \text{At least one } \beta_j \neq 0$$

(Intercept excluded from the test)

Test Statistic (F-test)

$$F = \frac{SS_{\text{Model}}/(p-1)}{SS_{\text{Error}}/(n-p)}$$

where

- $SS_{\text{Model}} = \sum(\hat{y}_i - \bar{y})^2$
- $SS_{\text{Error}} = \sum(y_i - \hat{y}_i)^2$

Connection to R^2

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

- Larger R^2 implies larger F
- The F-test provides a **formal inferential justification** for R^2

Sampling Distribution

Under H_0:

$$F \sim F_{p-1,\, n-p}$$

Decision Rule

- Compute the observed F statistic
- Reject H_0 if:

$$F > F_{p-1,n-p,\alpha}$$

or equivalently if the p-value $< \alpha$