Lecture 7: Statistical Inference II

Ailin Zhang

2023-05-24

Agenda

- t-Test of a Single β_j (Lec 6)
- Confidence Intervals for Coefficients (Lec 6)
- Confidence Intervals for Prediction
- Prediction Intervals for Prediction
- Sum of Squares
- Model Comparison (F-test)

Recap: Statistical Inference for β_j

- $\hat{\beta} \sim \mathrm{N}(\beta, \sigma^2 (X^T X)^{-1})$
- We can also denote it as $\hat{\beta} | X \sim \mathrm{N}(\beta, \sigma^2 (X^T X)^{-1})$
- Therefore, for $j = 0, 1, \ldots, p$:

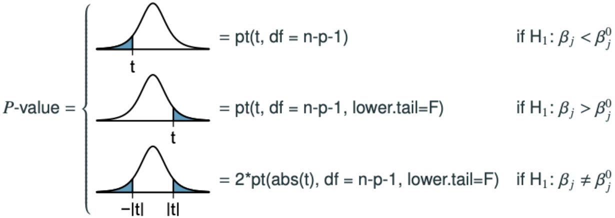
$$\hat{\beta}_j | X \sim \mathrm{N}(\beta_j, \mathrm{Var}(\hat{\beta}_j | X))$$

$\mathrm{Var}(\hat{\beta}_j | X)$: j+1 th diagonal entry of $\sigma^2 (X^T X)^{-1}$

- $\dfrac{\hat{\beta}_j - \beta_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j | X)}} \sim \mathrm{N}(0, 1)$
  - Issue: σ^2 is unknown
- $\dfrac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_j | X)}} = \dfrac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim t_{n-p-1}$

Recap: t-Test

- $H_0 : \beta_j = \beta_j^0$
  - The t-statistic for testing H_0 is $t = \dfrac{\hat{\beta}_j - \beta_j^0}{s.e.(\hat{\beta}_j)}$ which has a t-distribution with $df = n - p - 1$, where $s.e.(\hat{\beta}_j)$ is given on the previous slide.
  - The P-value can be calculated using `pt()` based on the alternative hypothesis H_1.



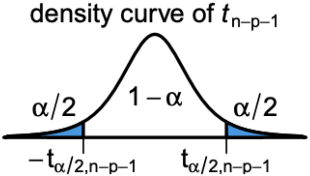
- Decision Rule:
  - Reject H_0 if P-value $< \alpha$

Recap: Confidence Interval

- $100(1-\alpha)\%$ confidence interval for β_j is:

$$\hat{\beta}_j \pm t_{n-p-1,\alpha/2} s.e.(\hat{\beta}_j)$$

where $t(n-p-1, \alpha/2)$ is the critical value for the t_{n-p-1} distribution at confidence level $1-\alpha$



which can be found using either of the following R commands:

```
qt(alpha/2, df=n-p-1, lower.tail=FALSE)
qt(1-alpha/2, df=n-p-1)
```

Estimation vs Prediction of Y

There are TWO kinds of predictions for the response Y given $X = x_0$ based on a SLR model $Y = \beta_0 + \beta_1 X + \epsilon$:

- given $X = x_0$, estimation of the mean response

$$E[Y|X = x_0] = \beta_0 + \beta_1 x0$$

- given $X = x_0$, prediction of the response for one specific observation

$$Y = \beta_0 + \beta_1 x0 + \epsilon$$

- The first one is an estimation problem it only involve fixed parameters β_0,β_1, and a known number x_0.
- The second one is a prediction problem as it involves an extra random number ϵ

Estimated Value and Predicted Value

Both

$$E[Y|X = x_0] = \beta_0 + \beta_1 x0 \quad \text{and} \quad Y = \beta_0 + \beta_1 x0 + \epsilon$$

are estimated/predicted by

$$\hat{\beta_0} + \hat{\beta_1} x0$$

This is because $E(\epsilon) = 0$

Variance of estimation and prediction

For SLR, $(X^TX)^{-1}$ equals

$$\begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} & -\frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ -\frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} & \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{bmatrix}$$

We get $\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right), \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$

$\mathrm{cov}(\hat{\beta}_1, \hat{\beta}_0) = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$

Variance of estimation and prediction

- $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$
- $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 + \epsilon) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) + \sigma^2$
- As n gets larger,
  - $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0)$ would go down to zero
  - $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 + \epsilon)$ just goes down to σ^2

Factors that control the variance

Variance for estimating $E(Y|X = x_0)$: $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$

Variance to predict Y when $X = x_0$: $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 + \epsilon) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) + \sigma^2$

Less variance comes from:

- small σ^2
- large sample size n
- large $\sum_{i=1}^{n}(x_i - \bar{x})^2$ (more spread in predictors)
- small $(x_0 - \bar{x})^2$

Confidence Intervals VS Prediction Intervals

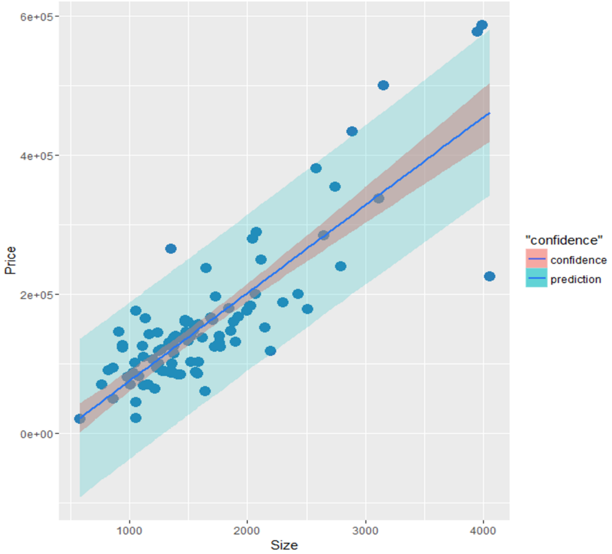
- $100(1-\alpha)\%$ confidence interval for $\beta_0 + \beta_1 x_0$ is:

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}}$$

- $100(1-\alpha)\%$ prediction interval for $Y = \beta_0 + \beta_1 x_0 + \epsilon$ is:

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum\limits_{i=1}^{n}(x_i-\bar{x})^2}}$$

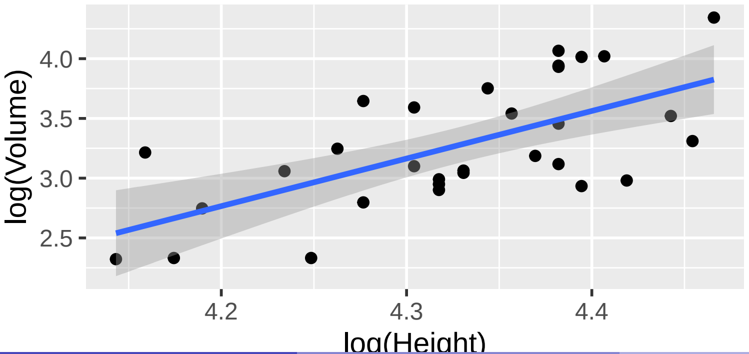
Confidence Intervals VS Prediction Intervals



Confidence Intervals in R

geom_smooth(method='lm') in ggplot() by default includes the 95% confidence intervals for estimating $E(y|X = x_0)$.

```r
library(ggplot2)
ggplot(trees, aes(x=log(Height), y=log(Volume))) +
  geom_point() + geom_smooth(method='lm', formula='y~x')
```



Example: Tree Data

```r
lm1 = lm(log(Volume) ~  log(Height), data=trees)
predict(lm1, data.frame(Height=71), interval="confidence")
```

```
##        fit      lwr      upr
## 1 3.015679 2.827221 3.204138
```

```r
predict(lm1, data.frame(Height=71), interval="prediction")
```

```
##        fit      lwr      upr
## 1 3.015679 2.161432 3.869927
```

- For trees with a height of 71 ft, the **average** of log(Volume) is estimated to be 3.0157 (measured in cubic ft) with a 95% confidence interval from 2.8272 to 3.2041.
- For a randomly selected tree with a height of 71 ft, the log(Volume) is between 2.1614 to 3.8699 with 95% confidence.

Example: Tree Data

- Both the confidence intervals and the prediction intervals are narrowest when $x_0 = \bar{X}$.
- Prediction interval is wider.

Accuracy of Predictions for MLR

An MLR model $Y = \beta_0 + \beta_1 X1 + \cdots + \beta_p X_p + \epsilon$ also has TWO kinds of predictions give the values of the predictors:

$$X_1 = x_{01}, \ldots, X_p = x_{0p}$$

- Estimation of the mean response:

$$E(Y|X_0) = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p}$$

- Prediction of the response for one specific observation at X_0

$$Y = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + \epsilon$$

Just like SLR, two problems have identical estimated/predicted values:

$$\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p}$$

but their standard errors are different

$$s.e.(E(\hat{Y}|X_0)) = \hat{\sigma}\sqrt{x_0^T (X^T X)^{-1} x_0}$$

$$s.e.(\hat{Y}|X_0) = \hat{\sigma}\sqrt{1 + x_0^T (X^T X)^{-1} x_0}$$

where $x_0^T = (1, x_{01}, \ldots, x_{0p})^T$

Confidence Intervals and Prediction Intervals

- $100(1-\alpha)\%$ confidence interval for $E(Y|X_0) = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p}$ is:

$$\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p} \pm t_{n-p-1,\alpha/2} s.e.(E(\hat{Y}|X_0))$$

- $100(1-\alpha)\%$ prediction interval for $Y = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + \epsilon$ is:

$$\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p} \pm t_{n-p-1,\alpha/2} s.e.(\hat{Y}|X_0)$$

Example: Tree Data

```r
lmtrees = lm(log(Volume) ~ log(Diameter) + log(Height),
             data=trees)
predict(lmtrees, data.frame(Diameter=10, Height = 70)
        ,interval = "confidence")
```

```
##        fit      lwr      upr
## 1 2.679696 2.633357 2.726035
```

```r
predict(lmtrees, data.frame(Diameter=10, Height = 70)
        ,interval = "prediction")
```

```
##        fit      lwr      upr
## 1 2.679696 2.506664 2.852728
```

- The **mean** log(Volume) for all 70-ft-tall, 10 ft in diameter, cherry trees is estimated to between 2.633 to 2.726, at 95% confidence level.
- The log(Volume) for a randomly selected 70-ft-tall cherry tree with a diameter of 10 ft is predicted to be between 2.507 to 2.853 with 95% confidence.

Example: Tree Data

```r
predict(lmtrees, data.frame(Diameter=10, Height = 70)
        ,interval = "confidence")
```

```
##        fit      lwr      upr
## 1 2.679696 2.633357 2.726035
```

```r
predict(lmtrees, data.frame(Diameter=10, Height = 70)
        ,interval = "prediction")
```

```
##        fit      lwr      upr
## 1 2.679696 2.506664 2.852728
```

One can exponentiate the intervals to get intervals for Volume rather than for log(Volume).

- The **mean** Volume for all 70-ft-tall, 10 ft in diameter, cherry trees is estimated to between $e^{2.633} \approx 13.92$ to $e^{2.726} \approx 15.27$ cubic ft, at 95% confidence level.
- The Volume for a randomly selected 70-ft-tall cherry tree with a diameter of 10 ft is predicted to be between $e^{2.507} \approx 12.26$ to $e^{2.853} \approx 17.34$ cubic ft with 95% confidence.

Sum of Squares

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SSE}$$

(Proof on board)

- SST = total sum of squares
  - total variability of Y
  - depends on the response Y only, not on the form of the model
- SSR = regression sum of squares
  - variability of Y explained by $X_1, \ldots, X_p$
- SSE = error (residual) sum of squares
  - variability of Y not explained by $X_1, \ldots, X_p$

Distribution of Sum of Squares

- For MLR model $Y = X\beta + \epsilon$, $\epsilon_i \sim N(0, \sigma^2)$
- $\dfrac{SSE}{\sigma^2} \sim \chi^2_{n-p-1}$
- If we further assume that $\beta_1 = \beta_2 = \cdots = \beta_p = 0$, then

$$\frac{SST}{\sigma^2} \sim \chi^2_{n-1}, \qquad \frac{SSR}{\sigma^2} \sim \chi^2_p$$

- The degrees of freedom also follows the summation rule:

$$df_{SST} = df_{SSR} + df_{SSE}$$

Multiple R^2 and Adjusted R^2

- Multiple R^2 also called the coefficient of determination, is defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Interpretation: proportion of variability in Y explained by $X_1, \ldots, X_p$
- $0 \leq R^2 \leq 1$
- For SLR, $R^2 = r_{xy}^2$ is the square of the correlation between X and Y. So multiple R^2 is a generalization of the correlation
- For MLR, R^2 is the square of the correlation between Y and $\hat{Y}$
- When more terms are added into a model, R^2 may increase or stay the same but never decrease.
    - Is large R^2 always preferable?

Adjusted R^2

Since R^2 always increases as we add terms to the model, some people prefer to use an adjusted R^2 defined as

$$R^2_{adj} = 1 - \frac{SSE/df_{SSE}}{SST/dt_{SST}} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$
$$= 1 - \frac{n-1}{n-p-1}(1-R^2)$$

- $-\frac{p}{n-p-1} \leq R^2_{adj} \leq R^2 \leq 1$
- R^2_{adj} can be negative
- R^2_{adj} does not always increase as more variables are added. In fact, if unnecessary terms are added, R^2_{adj} may decrease.

Example: Tree Data

```r
summary(lmtrees)
```

```
##
## Call:
## lm(formula = log(Volume) ~ log(Diameter) + log(Height), data = trees)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.168561 -0.048488  0.002431  0.063637  0.129223
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.63162    0.79979  -8.292 5.06e-09 ***
## log(Diameter)  1.98265    0.07501  26.432  < 2e-16 ***
## log(Height)    1.11712    0.20444   5.464 7.81e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08139 on 28 degrees of freedom
## Multiple R-squared:  0.9777, Adjusted R-squared:  0.9761
## F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

- The predictors log(Diameter) and log(Height) can explain 97.77% of the variation in log(Volume).
- 0.9761 is the modified R^2 to account for the number of variables and the sample size. .