Lecture 6: Statistical Inference

Hypothesis Testing and Confidence Interval

Ailin Zhang

2023-05-23

Recap: Gauss-Markov Model

Under the Gauss-Markov Model,

- $E(\hat{\beta}) = \beta$
- $\mathrm{cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$
- $E(\hat{\sigma}^2) = \sigma^2$, where $\hat{\sigma}^2 = \frac{RSS}{n-(p+1)} = \frac{\sum_{i=1}^{n} \hat{\epsilon_i}^2}{n-(p+1)}$

Recap: Gaussian/Normal linear model

- We would like to further study the distribution of the OLS estimator
  - Enable statistical Inference
- To derive the distribution, we need stronger assumptions
  - The assumption will focus on the Gaussian/Normal linear model:
    - $\epsilon \sim \mathrm{N}(0, \sigma^2 I_n)$
    - $Y \sim \mathrm{N}(X\beta, \sigma^2 I_n) \Longleftrightarrow y_i \overset{\text{ind}}{\sim} N(x_i^T \beta, \sigma^2), \qquad i = 1, \ldots, n$
  - where X is fixed such that $X^T X$ is non-degenerate, and (β, σ^2) are fixed but unknown parameters.
  - The modeling assumption is extremely strong, but it is canonical in statistics.
  - It allows us to derive elegant formulas, and also justifies the output of the linear regression function in many statistical packages.

Recap: Joint Distribution of $(\hat{\beta}, \hat{\sigma}^2)$

Theorem

Under the Gaussian linear model,

$$\begin{pmatrix} \hat{\beta} \\ \hat{\epsilon} \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} \beta \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} (X^T X)^{-1} & 0 \\ 0 & I_n - H \end{pmatrix} \right\}$$

so $\hat{\beta} \perp\!\!\!\perp \hat{\epsilon}$*;* $\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-(p+1)}$*, and* $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$

Recap: Joint Distribution: $(\hat{Y}, \hat{\epsilon})$

Theorem

Under the Gaussian linear model,

$$\begin{pmatrix} \hat{Y} \\ \hat{\epsilon} \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} H & 0 \\ 0 & I_n - H \end{pmatrix} \right\}$$

so $\hat{Y} \perp\!\!\!\perp \hat{\epsilon}$

- Orthogonal: a linear algebra fact without assumptions. $< x, y >= 0$
- Independent: a statistical property under the Gaussian linear model. $f(x, y) = f(x)f(y)$
- Uncorrelated: $\mathrm{cov}(x, y) = 0$

Agenda

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

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
- $\dfrac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_j|X)}} = \dfrac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim t_{n-p-1}$

Example: The Trees Data

The trees data are measurements of the diameter, height and volume of timber in 31 felled black cherry trees. The variables are - Girth: Tree diameter (rather than girth, actually) in inches measured at 4 ft 6 in above the ground

- Height: Height in ft
- Volume: Volume of timber in cubic ft

The trees data are build-in in R. One can load the the data by the command

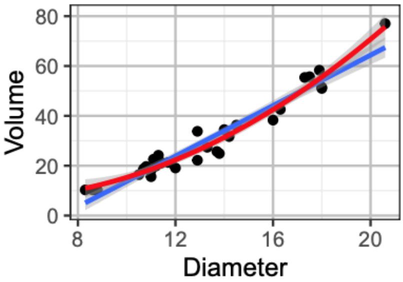
```r
data("trees")
```

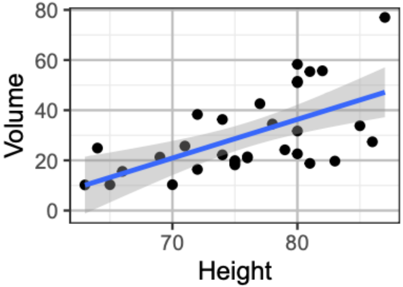
Let's rename the misleading Girth variable as Diameter

```r
trees$Diameter = trees$Girth
```

Data Exploration

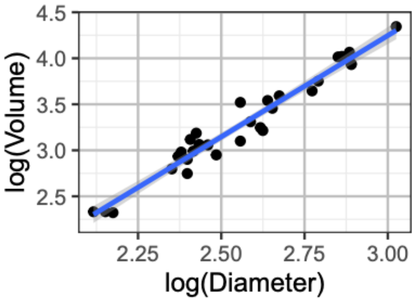
```
library(ggplot2)
ggplot(trees, aes(x=Diameter, y=Volume)) + geom_point() +
geom_smooth(method='lm', formula='y~x') +
geom_smooth(method='lm', formula='y~x+I(x^2)', col="red")
ggplot(trees, aes(x=Height, y=Volume)) + geom_point() +
geom_smooth(method='lm', formula='y~x')
```

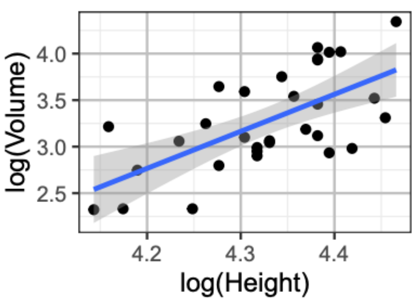




Log Transformation

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





t-Test

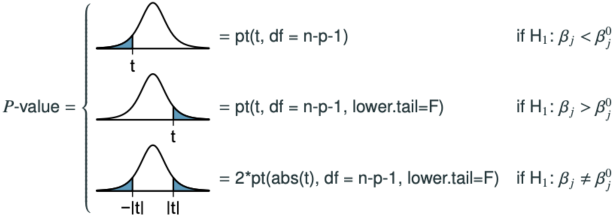
```r
lmtrees = lm(log(Volume) ~ log(Diameter) + log(Height), data=trees)
summary(lmtrees)
```

```
##
## 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 column “estimate” shows the LS estimates
- The column “std. error” gives the standard errors

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.



t-Test: Significance level α

Decision Rule:

- Reject H_0 if P-value $< \alpha$
- Reject H_0 if $t < -t_\alpha$ or $t > t_\alpha$ (for one tail)
- Reject H_0 if $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$ (for two tails)

To find the critical t-score in R, you can use `qt()`

```
qt(p, df, lower.tail=TRUE)
```

Hypothesis test with $\alpha = 0.05$

```r
summary(lmtrees)$coef
```

```
##                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   -6.631617 0.79978973 -8.291701 5.057138e-09
## log(Diameter)  1.982650 0.07501061 26.431592 2.422550e-21
## log(Height)    1.117123 0.20443706  5.464388 7.805278e-06
```

Suppose now we want to test

- $H_0 : \beta_1 = 2$, $H_1 : \beta_1 \neq 2$

$$t_1 = \frac{\hat{\beta}_1 - 2}{s.e.(\hat{\beta}_1)} = \frac{1.983 - 2}{0.075} \approx -0.227 \text{ with } df = 31 - 2 - 1 = 28$$

```r
# Compute p value
2*pt(0.227,df=28, lower.tail=F)
```

```
## [1] 0.822073
```

Conclusion: Since $0.82 > 0.05$, we can not reject H_0

- Exercise: $H_0 : \beta_2 = 1$, $H_1 : \beta_2 \neq 1$

```r
summary(lmtrees)$coef
```

```
##                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     -6.631617 0.79978973 -8.291701 5.057138e-09
## log(Diameter)    1.982650 0.07501061 26.431592 2.422550e-21
## log(Height)      1.117123 0.20443706  5.464388 7.805278e-06
```

1. The column "**t value**" shows the t-statistic for testing $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$,

$$t_0 = \frac{\hat{\beta}_0 - 0}{s.e.(\hat{\beta}_0)} = \frac{-6.632 - 0}{0.7998} = -8.292$$

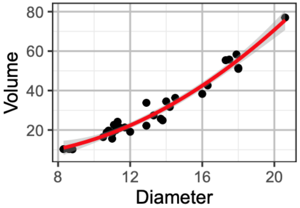
$$t_1 = \frac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)} = \frac{1.983 - 0}{0.07501} = 26.432$$

$$t_2 = \frac{\hat{\beta}_2 - 0}{s.e.(\hat{\beta}_2)} = \frac{1.117 - 0}{0.2044} = 5.464$$

2. The column "**Pr(>|t|)**" shows the 2-sided P-values for testing $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$

Potential Application: Check Non-Linearity

Recall we said earlier that the relation between Volume and Diameter is slightly nonlinear. We can check nonlinearity by fitting the polynomial model



$$\text{Volume} = \beta_0 + \beta 1\text{Diameter} + \beta_2\text{Diameter}^2 + \epsilon$$

Test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$

```
lm2 = lm(Volume ~ Diameter + I(Diameter^2), data=trees)
summary(lm2)$coef
```

```
##                   Estimate Std. Error    t value    Pr(>|t|)
## (Intercept)     10.7862655 11.2228199  0.9611012 0.344728171
## Diameter        -2.0921396  1.6473417 -1.2700095 0.214534409
## I(Diameter^2)    0.2545376  0.0581716  4.3756327 0.000152389
```

Digression: Vector Notation in Statistical Inference

Let's consider $c \in \mathbb{R}^{p+1}$. if $c = e_{j+1} = (0, \ldots, 1, \ldots, 0)^T$ with the only $j+1$th element being one. $c^T\beta = \beta_j$

- $c^T\hat{\beta} \sim \mathrm{N}\{c^T\beta, \sigma^2 c^T(X^TX)^{-1}c\}$
-
$$
\begin{aligned}
T_c = \frac{c^T\hat{\beta} - c^T\beta}{\sqrt{\hat{\sigma}^2 c^T(X^TX)^{-1}c}} &= \frac{c^T\hat{\beta} - c^T\beta}{\sqrt{\sigma^2 c^T(X^TX)^{-1}c}} \Big/ \sqrt{\frac{\hat{\sigma}^2}{\sigma^2}} \\
&\sim \frac{\mathrm{N}(0,1)}{\sqrt{\chi^2_{n-p-1}/(n-p-1)}} \\
&\sim t_{n-p-1}
\end{aligned}
$$

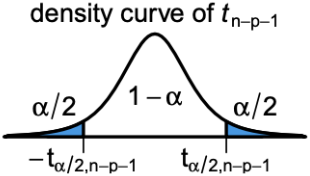
- Sometimes we may also be interested in $\beta_j - \beta_k$, the difference between the coefficients of two covariates, which corresponds to $c = (0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)^T = e_{j+1} - e_{k+1}$

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)
```

Example: CI for β_1

```r
summary(lmtrees)$coef
```

```
##                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)     -6.631617 0.79978973 -8.291701 5.057138e-09
## log(Diameter)    1.982650 0.07501061 26.431592 2.422550e-21
## log(Height)      1.117123 0.20443706  5.464388 7.805278e-06
```

95% confidence intervaal for β_1 is

$$\hat{\beta}_1 \pm t_{0.025,28} s.e.(\hat{\beta}_1)$$

where $t_{0.025,28}$ can be found using either R commands below

```r
qt(0.05/2, df=28, lower.tail=F)
```

```
## [1] 2.048407
```

```r
qt(0.975, df=28)
```

```
## [1] 2.048407
```

Finding CIs for Coefficients Using `confint()`

The `confint()` command in R can produce confidence intervals for the coefficients as well.

```
confint(lmtrees)
```

```
##                      2.5 %    97.5 %
## (Intercept)   -8.269912 -4.993322
## log(Diameter)  1.828998  2.136302
## log(Height)    0.698353  1.535894
```

```
confint(lmtrees, level = 0.9) # changing the confidence level to 90%
```

```
##                         5 %      95 %
## (Intercept)   -7.9921642 -5.271070
## log(Diameter)  1.8550470  2.110253
## log(Height)    0.7693491  1.464898
```

```
confint(lmtrees, level = 0.95, "log(Diameter)")
```

```
##                    2.5 %   97.5 %
## log(Diameter) 1.828998 2.136302
```

Interpretation: We have 95% confidence that every unit increase in X_j will increase/decrease Y by xx to xx on average while holding other variables as constant.

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

Confidence Intervals VS Prediction Intervals

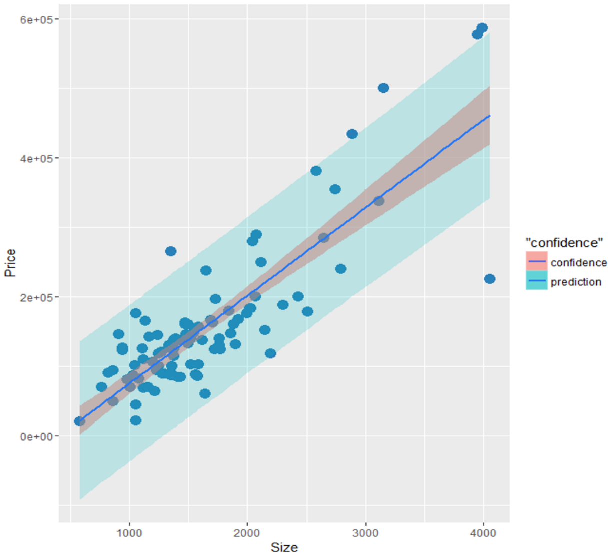
- $100(1-\alpha)\%$ confidence interval for $\beta_0 + \beta_1 x_0$ is:

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

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

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

Confidence Intervals VS Prediction Intervals



Confidence Intervals in R

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

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