LINEAR MODELS

Zhaoxia Yu

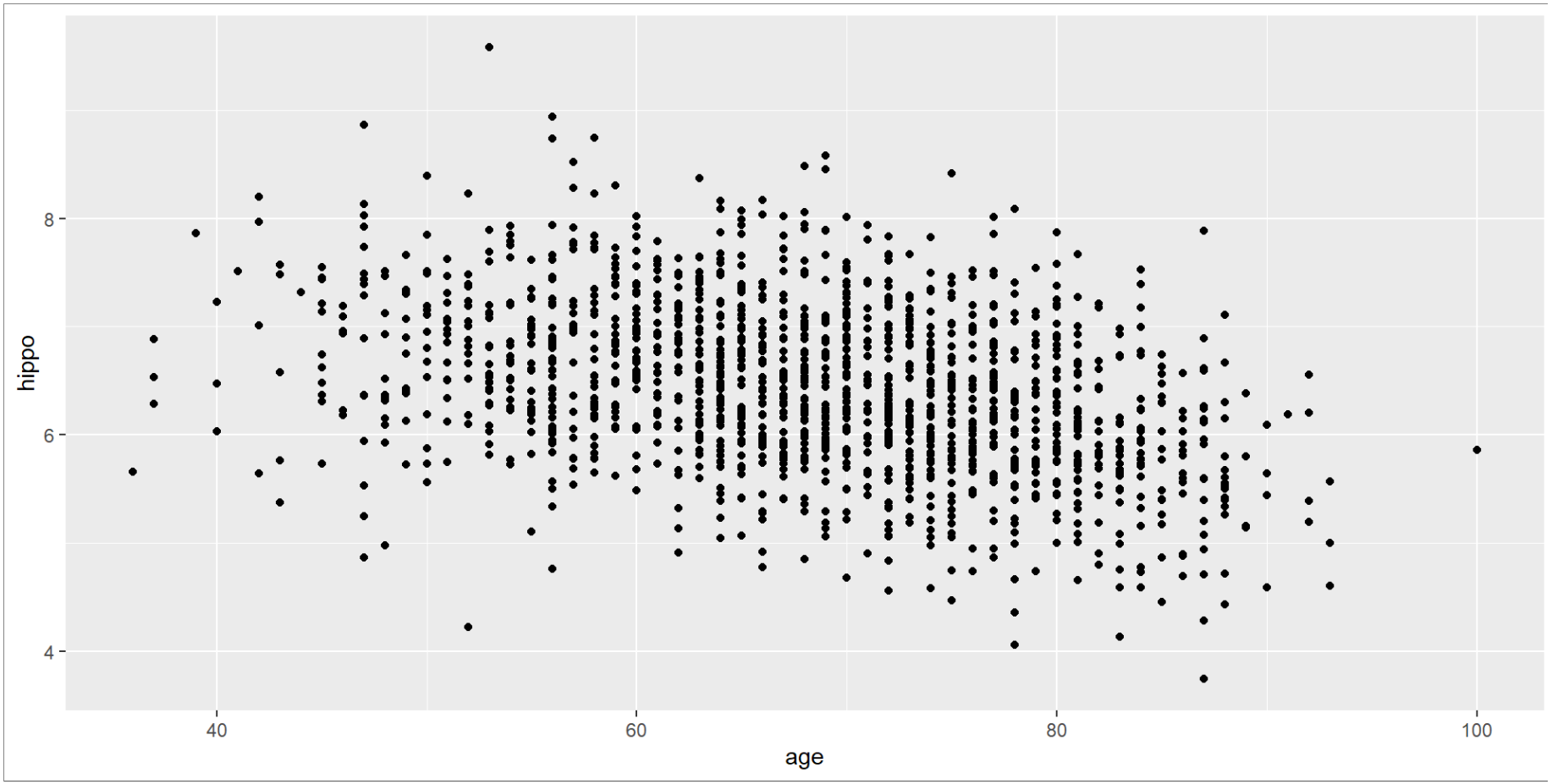
LEARNING OBJECTIVES

- Simple linear regression
  - Interpretation of (estimated) parameters
  - Outliers
- Multiple linear regression
- Design matrices
- Theory of Linear Models

SIMPLE LINEAR REGRESSION

SIMPLE LINEAR REGRESSION

- Simple linear regression is the regression model with only one explanatory variable.
- It can be used to assess the linear relationship between two variables. Example, age and hippocampus volume in healthy subjects


hippo
8
6
4
40
60
80
100
age

SIMPLE LINEAR REGRESSION

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \cdots, n$
  - β_0 is the intercept. It is the expected value when $x = 0$.
  - β_1 is the slope. It is the expected change of the response variable when x increases by 1 (unit).
  - ϵ_i is assumed to have mean 0.
- The least squares fit minimizes the sum of squared errors, i.e., $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained by minimizing

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

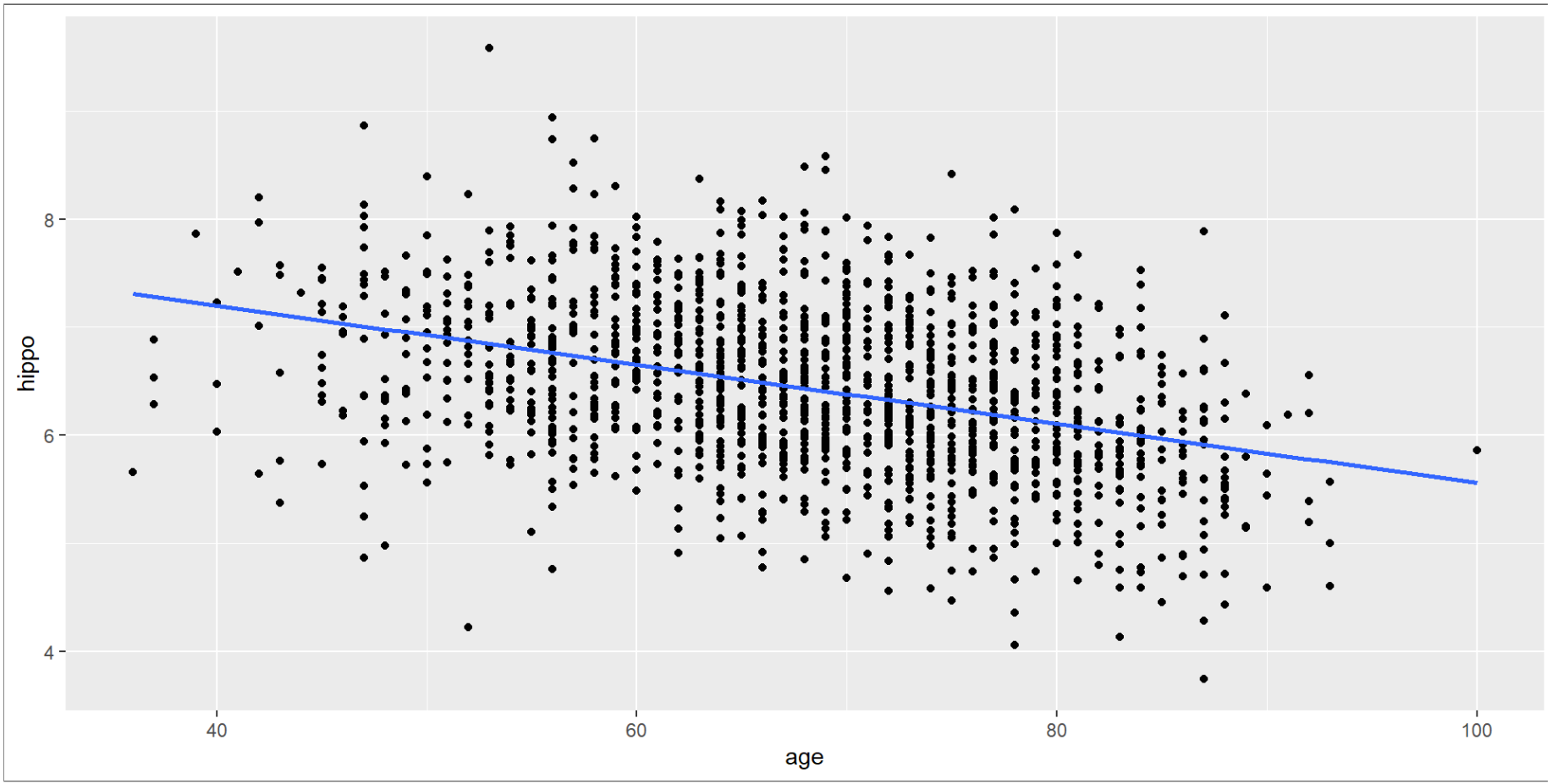
THE FITTED LINE USING LEAST SQUARES ESTIMATE

- The LSE of β_0, β_1 is:
  - $b_0 = \hat{\beta}_0 = \bar{y} - b_1 \bar{x}$
  - $b_1 = \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- There are alternative ways to express $\hat{\beta}_1$. Please verify the following
  - $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = r_{xy}\dfrac{s_y}{s_x}$ where
  - $r_{x,y}$ is the sample correlation coefficient between x and y , s_x and s_y are the sample standard deviation

EXAMPLE: LSE AND THE FITTED LINE

```r
obj=lm(hippo~age, data=alzheimer0)
coef(obj) #alternatively: obj$coefficients
```

```
(Intercept)         age
 8.29252018 -0.02732632
```



The fitted line is $\boldsymbol{hippo = 8.29 - 0.027 \times age}$
For each added year, the hippocampus size decreased by 0.027 (cc)

MODEL ASSUMPTIONS

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \cdots, n$
- $E(\epsilon_i) = 0$
- The n observations are independent
- Reasonally large n or normality assumption: $\epsilon_i \sim N(0, \sigma^2)$

RESIDUALS

- $e_i = y_i - \hat{y}_i$ where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Residual sum of squares (RSS): $RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- An unbiased estimate of σ^2 is s^2:
  $s^2 = \frac{RSS}{n-2}$

STANDARD ERROR AND CONFIDENCE INTERVAL

- The standard error of $\hat{\beta}_1$ is

  $$se(\hat{\beta}_1) = \frac{\sqrt{s^2}}{\sqrt{\sum(x_i - \bar{x})^2}}$$

- A $100(1-\alpha)\%$ confidence interval for β_1 is

  $$\hat{\beta}_1 \pm t_{\alpha/2,n-2} se(\hat{\beta}_1)$$

T AND P-VALUE

```
1 summary(obj)
```

```
Call:
lm(formula = hippo ~ age, data = alzheimer0)

Residuals:
     Min       1Q   Median       3Q      Max
-2.64965 -0.48857 -0.01641  0.48528  2.73487

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.292520   0.118318   70.09   <2e-16 ***
age         -0.027326   0.001712  -15.96   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7157 on 1507 degrees of freedom
```

```
1 confint(obj)
```

```
                  2.5 %      97.5 %
(Intercept)  8.06043571  8.52460464
age         -0.03068437 -0.02396827
```

ESTIMATION VS PREDICTION

- Estimate $E(y|x = x_h) = \beta_0 + \beta_1 x_h$
  - The estimate is $\hat{\beta}_0 + \hat{\beta}_1 x_h$
  - Confidence interval:

$$\hat{y}_h \pm t_{\alpha/2,n-2}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x}^2)}\right)}$$

- Predict the response value when $x = x_h$
  - The prediction is $\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$
  - Prediction interval:

$$\hat{y}_h \pm t_{\alpha/2,n-2}\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x}^2)}\right)}$$

EXAMPLE: ESTIMATION VS PREDICTION

```
1 predict(obj, newdata = data.frame(age=44), interval = 'confidence')
```

```
       fit      lwr      upr
1 7.090162 7.001009 7.179315
```

```
1 predict(obj, newdata = data.frame(age=44), interval = 'predict')
```

```
       fit      lwr      upr
1 7.090162 5.683368 8.496956
```

OUTLIERS

Outliers are points that do not follow the general pattern of the majority of the data.

- **Leverage points** are also called x-outliers.
- **Influential Outliers:** These are data points that, when removed, lead to a significant change in the regression model. They can drastically affect the slope and intercept of the regression line.
- Residuals ($e_i = y_i - \hat{y}_i, i = 1, \cdots, n$) and their standardized versions are often used to visually inspect assumptions.
- **Outliers** often have large residuals. Residual plots can be used to visually detect these outliers. An observation with a residual significantly larger or smaller than the other residuals may be an outlier.

OUTLIERS: EXAMPLES

MODEL DIAGNOSTICS

- **Linearity**: The relationship between predictors and response is linear. Violation seen when residuals display a systematic pattern.
- **Independence**: The residuals are independent. Violation seen when residuals display trends over time or space.
- **Homoscedasticity**: The variance of the residuals is constant. Violation seen when residuals show a “funnel” shape pattern.
- **Normality**: The residuals are normally distributed. Violation seen when residuals don’t follow a bell-shaped pattern in a histogram or a straight line on a Q-Q plot.

MULTIPLE LINEAR REGRESSION

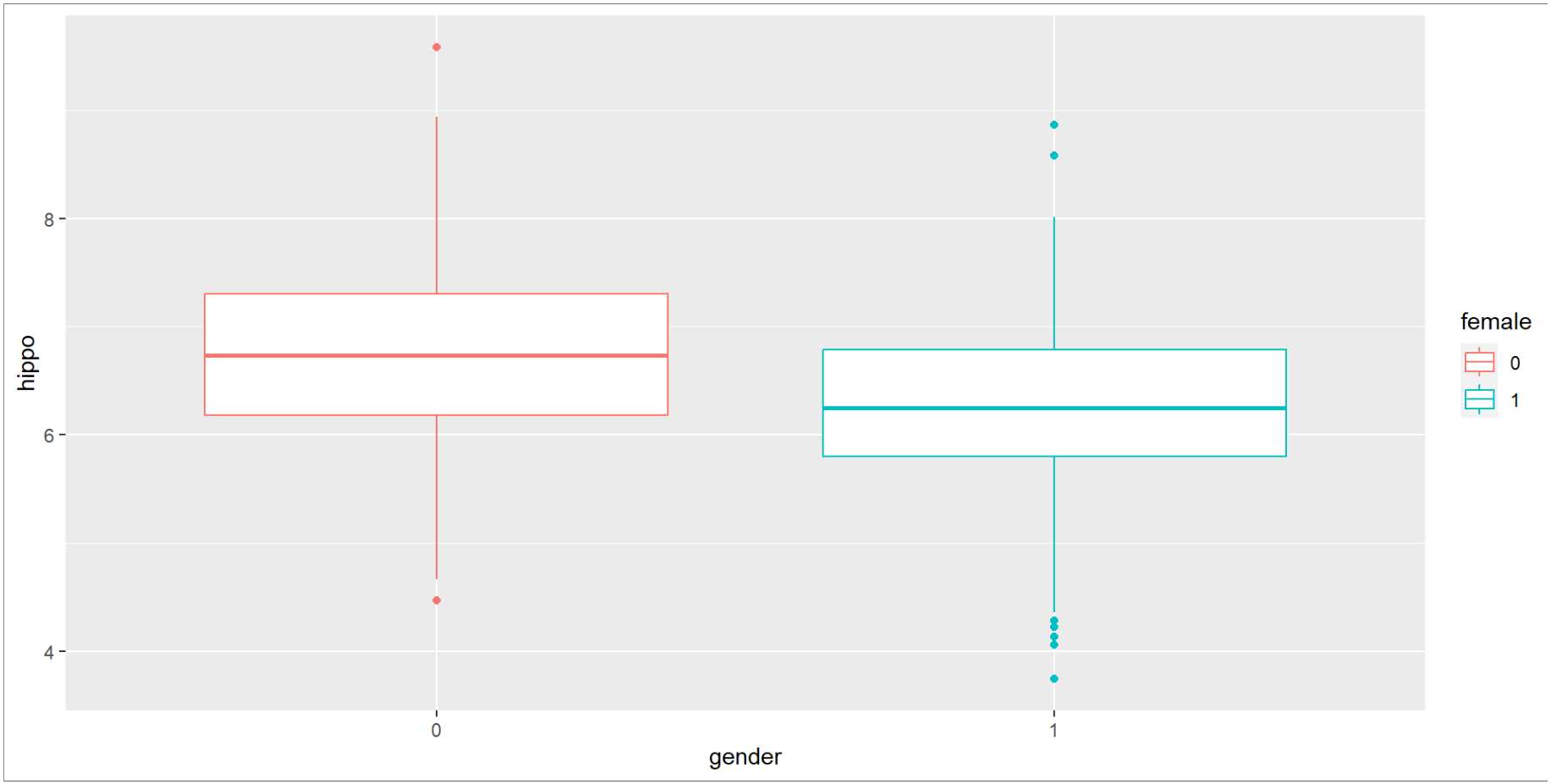
MOTIVATING EXAMPLE

- Are healthy men and women differ in hippocampus volume?

```
1  ggplot(alzheimer0, aes(x=female, y=hippo, color=female)) + geom_boxplot()
2    labs(x = "gender")
```

```
1  summary(lm(hippo~female, data=alzheimer0))
```

```
Call:
lm(formula = hippo ~ female, data = alzheimer0)

Residuals:
     Min       1Q   Median       3Q      Max
-2.51650 -0.49828 -0.01098  0.54050  2.83082

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.74828    0.03247  207.84   <2e-16 ***
female1     -0.48879    0.04005  -12.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7383 on 1507 degrees of freedom
```


hippo
8
6
4
0
1
gender
female
0
1

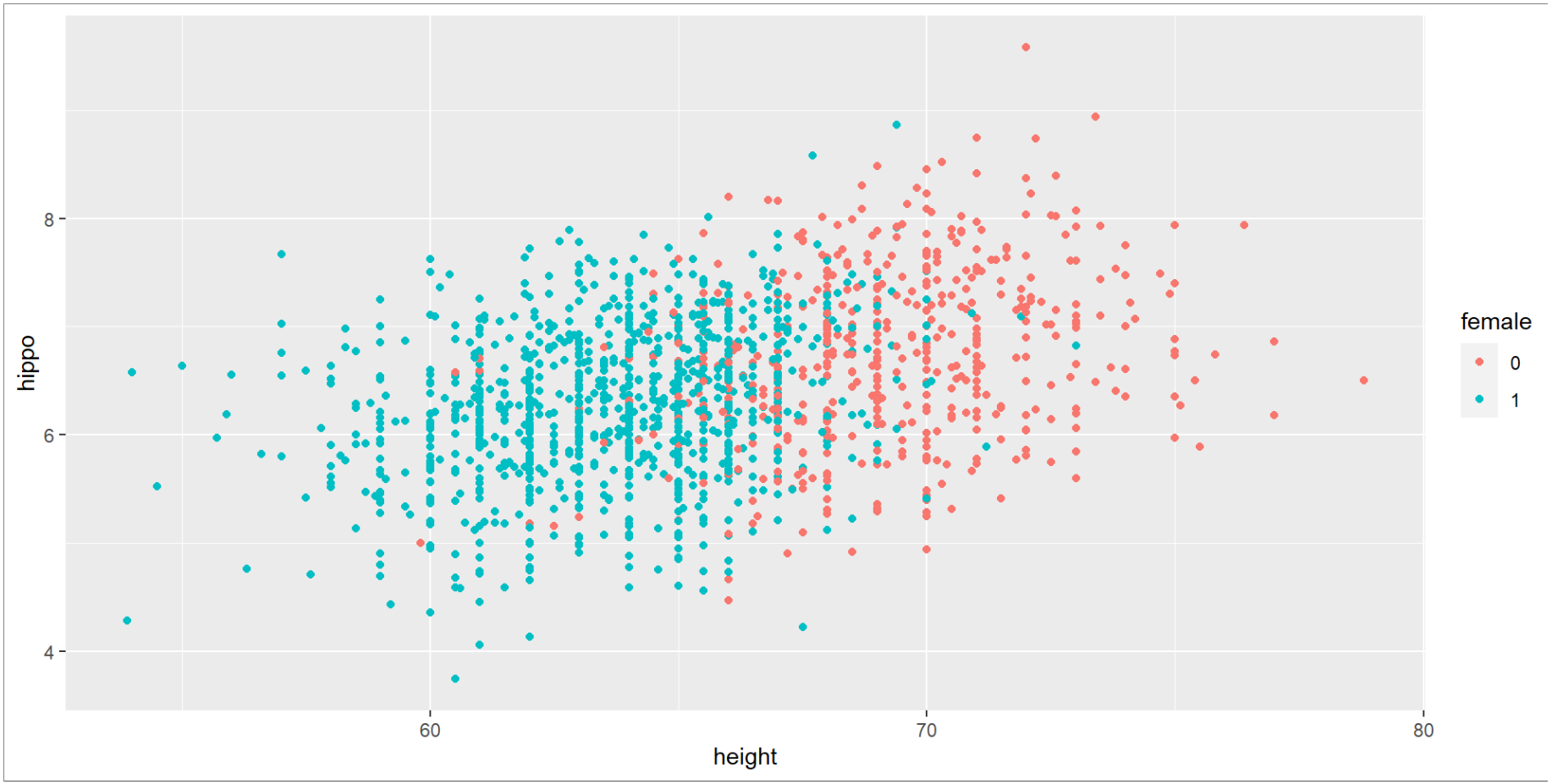
CAN THE DIFFERENCE BE EXPLAINED BY HEIGHT?

- Men have larger hippocampus volume

```
1  ggplot(alzheimer0, aes(x=height, y=hippo, color=female))+geom_point()
```

```
1  summary(lm(hippo ~ height + female, data=alzheimer0))
```

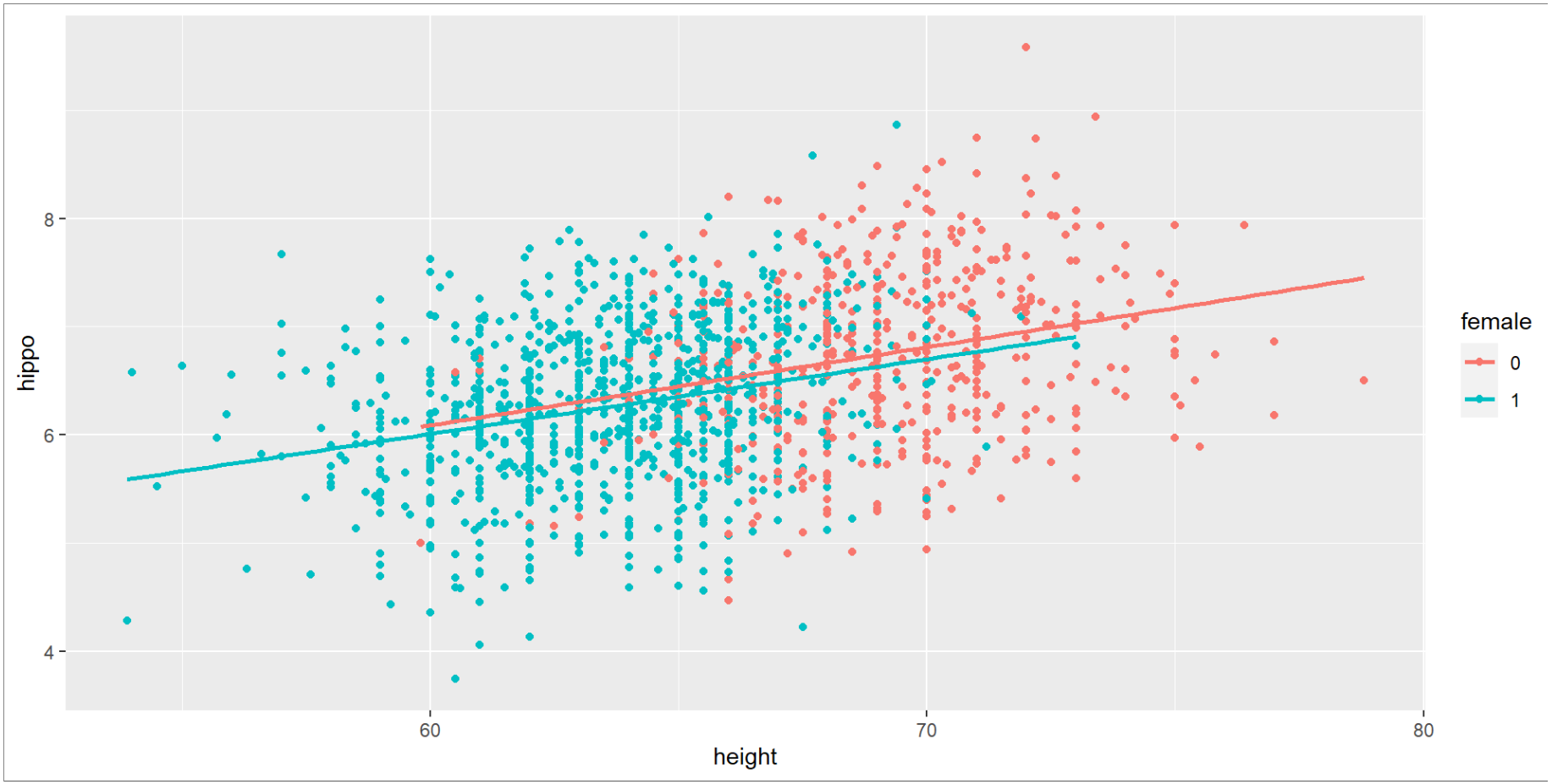
```
Call:
lm(formula = hippo ~ height + female, data = alzheimer0)

Residuals:
     Min       1Q   Median       3Q      Max
-2.30992 -0.46918  0.01402  0.50658  2.62898

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.896543   0.461747   4.107 4.22e-05 ***
height       0.070189   0.006665  10.532  < 2e-16 ***
female1     -0.102456   0.053294  -1.922   0.0547 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



- The **adjusted** difference between men and women is much smaller

DO MEN AND WOMEN HAVE THE SAME HIPPO~HEIGHT RELATIONSHIP?

```
1  ggplot(alzheimer0, aes(x=height, y=hippo, color=female))+
2    geom_point()+
3    geom_smooth(method=lm, se=F)
```


hippo
8
6
4
60
70
80
height
female
0
1

DO MEN AND WOMEN HAVE THE SAME HIPPO~HEIGHT RELATIONSHIP?

```
1  summary(lm(hippo ~ height* female, data=alzheimer0))
```

```
Call:
lm(formula = hippo ~ height * female, data = alzheimer0)

Residuals:
     Min       1Q   Median       3Q      Max
-2.30535 -0.46769  0.00966  0.50828  2.62261

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.743321   0.783010   2.226   0.0261 *
height         0.072405   0.011318   6.397 2.11e-10 ***
female1        0.125680   0.942902   0.133   0.8940
height:female1 -0.003394   0.014006  -0.242   0.8086
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

COMPARE TWO LINEAR MODELS

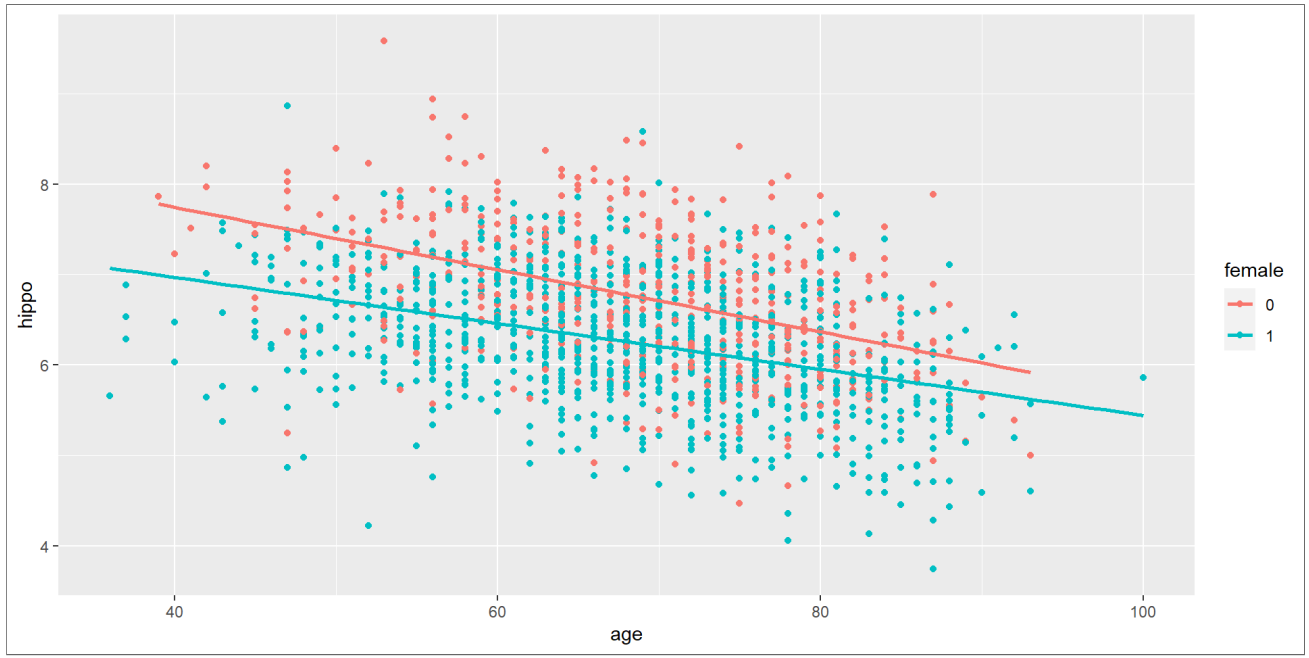
- Code

```
Analysis of Variance Table

Model 1: hippo ~ height
Model 2: hippo ~ height * female
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1507 766.89
2   1505 764.98  2    1.9072 1.8761 0.1535
```

INTERPRET INTERACTIONS

- Do men and women shrink brain in the same speed?

▶ Code



INTERPRET INTERACTIONS

▶ Code

```
Call:
lm(formula = lhippo ~ age * female, data = alzheimer0)

Residuals:
     Min       1Q   Median       3Q      Max
-1.38031 -0.23026 -0.00833  0.25027  1.24240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.532691   0.103765  43.682  < 2e-16 ***
age         -0.017258   0.001489 -11.590  < 2e-16 ***
female1     -0.587113   0.125421  -4.681 3.11e-06 ***
age:female1  0.004712   0.001807   2.608   0.0092 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

INTERPRET INTERACTIONS

- The estimated brain shrinking speed for men is 0.017cc per year
- The estimated brain shrinking speed for women is 0.013cc per year
- The result about the interaction item indicates that the difference in brain shrinking speed between men and women is 0.005cc per year and it is statistically significant

INTERPRET TRANSFORMED DATA

- Weight is a little bit skewed. Regress hippocampus volume on log(weight). How to interpret the results?

▶ Code

```
Call:
lm(formula = hippo ~ log(weight), data = alzheimer0)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.43292 -0.53018 -0.00561  0.51543  2.75324 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.25174    0.46350   2.701    0.007 ** 
log(weight)  1.01324    0.09067  11.175   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7437 on 1507 degrees of freedom
```

INTERPRET TRANSFORMED DATA

- Regress log(weight) on height. How to interpret the results?

▶ Code

```
Call:
lm(formula = log(weight) ~ height, data = alzheimer0)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.52837 -0.12566 -0.01422  0.11537  0.58745 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.184590   0.079885   39.87   <2e-16 ***
height      0.029356   0.001217   24.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1795 on 1507 degrees of freedom
```

LINEAR MODELS ARE A UNIFIED TOOL

CONDUCT T-TEST USING LM

- Code

```
    Welch Two Sample t-test

data:  hippo by female
t = 11.786, df = 950.08, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1
is not equal to 0
95 percent confidence interval:
 0.4074032 0.5701731
sample estimates:
mean in group 0 mean in group 1
       6.748284        6.259496
```

- Code

```
Call:
lm(formula = hippo ~ female, data = alzheimer0)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-2.51650 -0.49828 -0.01098  0.54050  2.83082

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.74828    0.03247  207.84   <2e-16 ***
female1     -0.48879    0.04005  -12.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CONDUCT ANOVA USING LM

▶ Code

```
             Df Sum Sq Mean Sq F value Pr(>F)
diagnosis     2  100.5   50.27   247.1 <2e-16 ***
Residuals  2697  548.6    0.20
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

▶ Code

```
Call:
lm(formula = lhippo ~ diagnosis, data = alzheimer)

Residuals:
     Min       1Q   Median       3Q      Max
-2.14029 -0.30148 -0.00139  0.30199  1.85171

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.18148    0.01152  276.29   <2e-16 ***
diagnosis1  -0.29680    0.02155  -13.77   <2e-16 ***
diagnosis2  -0.46009    0.02237  -20.57   <2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

▶ Code

DESIGN MATRIX

- Definition: The design matrix, often denoted as $\mathbf{X}$, is a matrix of observed data in which each row represents an observation, and each column corresponds to a predictor or independent variable in the linear model.
- Matrix format: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$:
  - $\mathbf{Y}$ is the $n \times 1$ response vector.
  - $\mathbf{X}$ is the $n \times p$ design matrix.
  - $\boldsymbol{\beta}$ is the $p \times 1$ vector of coefficients.
  - $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector of error terms.

DESIGN MATRIX: EXAMPLE 1

▶ Code

```
[1] 1509    2
```

▶ Code

```
  (Intercept) age
1           1  74
2           1  56
3           1  77
4           1  74
6           1  72
7           1  64
```

DESIGN MATRIX: EXAMPLE 1

- The model in Example 1 is
  $Y_i = \beta_0 + \beta_1 x_i + \varepsilon, i = 1, \ldots, n$
- Align the n equations vertically,
  $$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$
- The matrix format is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

DESIGN MATRIX: EXAMPLE 2

▶ Code

```
[1] 1509    2
```

▶ Code

```
  (Intercept) female1
1           1       0
2           1       1
3           1       1
4           1       1
6           1       1
7           1       1
```

DESIGN MATRIX: EXAMPLE 2

▶ Code

```
Call:
lm(formula = hippo ~ female, data = alzheimer0)

Residuals:
     Min       1Q   Median       3Q      Max
-2.51650 -0.49828 -0.01098  0.54050  2.83082

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.74828    0.03247  207.84   <2e-16 ***
female1     -0.48879    0.04005  -12.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7383 on 1507 degrees of freedom
```

- How to interpret the results?

DESIGN MATRIX: EXAMPLE 2

```
Call:
lm(formula = hippo ~ female - 1, data = alzheimer0)

Residuals:
     Min       1Q   Median       3Q      Max
-2.51650 -0.49828 -0.01098  0.54050  2.83082

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
female0  6.74828    0.03247   207.8   <2e-16 ***
female1  6.25950    0.02344   267.0   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7383 on 1507 degrees of freedom
```

```
  female0 female1
1       1       0
2       0       1
3       0       1
4       0       1
6       0       1
7       0       1
```

DESIGN MATRIX: EXAMPLE 3

▶ Code

```
[1] 2700    3
```

▶ Code

```
  (Intercept) diagnosis1 diagnosis2
1           1          0          0
2           1          0          0
3           1          0          0
4           1          0          0
5           1          1          0
6           1          0          0
```

- According to the design matrix, which diagnosis group is used as the reference group?

DESIGN MATRIX: EXAMPLE 3

▶ Code

```
Call:
lm(formula = lhippo ~ diagnosis, data = alzheimer)

Residuals:
     Min       1Q   Median       3Q      Max
-2.14029 -0.30148 -0.00139  0.30199  1.85171

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.18148    0.01152  276.29   <2e-16 ***
diagnosis1  -0.29680    0.02155  -13.77   <2e-16 ***
diagnosis2  -0.46009    0.02237  -20.57   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- How to interpret the results?

DESIGN MATRIX: EXAMPLE 3

- Let's remove the intercept

▶ Code

```
Call:
lm(formula = lhippo ~ diagnosis - 1, data = alzheimer)

Residuals:
     Min       1Q   Median       3Q      Max
-2.14029 -0.30148 -0.00139  0.30199  1.85171

Coefficients:
           Estimate Std. Error t value Pr(>|t|)
diagnosis0  3.18148    0.01152   276.3   <2e-16 ***
diagnosis1  2.88468    0.01822   158.4   <2e-16 ***
diagnosis2  2.72139    0.01918   141.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- How to interpret the results?

DESIGN MATRIX: EXAMPLE 3

- We can also change the reference group

▶ Code

```
Call:
lm(formula = lhippo ~ relevel(diagnosis, ref = "1"), data = alzheimer)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.14029 -0.30148 -0.00139  0.30199  1.85171 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      2.88468    0.01822 158.360  < 2e-16 ***
relevel(diagnosis, ref = "1")0   0.29680    0.02155  13.772  < 2e-16 ***
relevel(diagnosis, ref = "1")2 -0.16329    0.02645  -6.173 7.69e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- How to interpret the results?

INFERENCE OF A LINEAR COMBINATION

- The lm function in R produces not only estimates but also the variance-covariance matrix

▶ Code

```
diagnosis0 diagnosis1 diagnosis2
  3.181481   2.884682   2.721390
```

▶ Code

```
             diagnosis0   diagnosis1   diagnosis2
diagnosis0 0.0001325996 0.0000000000 0.0000000000
diagnosis1 0.0000000000 0.0003318234 0.0000000000
diagnosis2 0.0000000000 0.0000000000 0.0003678259
```

- What if we are interested in a specific linear combination of the coefficients, say $\frac{\beta_1+\beta_2}{2} - \beta_0$

INFERENCE OF A LINEAR COMBINATION

- Note that $\frac{\beta_1+\beta_2}{2} - \beta_0 = (-1, \frac{1}{2}, \frac{1}{2}) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$
- $cov(a^T\hat{\boldsymbol{\beta}}) = a^T cov(\hat{\boldsymbol{\beta}})a$

▶ Code

```
[1] "estimate"
```

▶ Code

```
            [,1]
[1,] -0.3784448
```

▶ Code

```
[1] "standard error"
```

▶ Code

```
            [,1]
[1,] 0.01753602
```

INFERENCE OF A LINEAR COMBINATION

▶ Code

```
        [,1]
[1,] -21.581
```

▶ Code

```
[1] "p-value"
```

▶ Code

```
     [,1]
[1,]    0
```

THEORY OF LINEAR MODELS

OUTLINE

- Warning: due to limited time, we cannot review everything needed to establish all the results. The goal of this one-hour lecture is to get the big picture. To fully understand everything, we need 10+ lectures.
- The normal equation and LSE
- The Gauss-Markov theorem
- Multivariate normal distributions
- Chi-squared distributions
- F-test and t-test

LEAST SQUARES ESTIMATE: AN OPTIMIZATION PROBLEM

- Suppose $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. The LSE is obtained by minimizing the sum of squared errors
  $Q(\boldsymbol{\beta}) = ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$
- Differentiation with respect to a vector of parameters. Let $f(\boldsymbol{\beta})$ be a scaluar-valued function of a vector $\boldsymbol{\beta}$. The differentiation with respect to (wrt) the vector $\boldsymbol{\beta}$ is defined as

$$\frac{d}{d\boldsymbol{\beta}} f(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial f}{\partial \beta_0} \\ \frac{\partial f}{\partial \beta_1} \\ \cdots \\ \frac{\partial f}{\partial \beta_{p-1}} \end{pmatrix}$$

- Note. This is known as the gradient of $f(\boldsymbol{\beta})$

LEAST SQUARES ESTIMATE: DERIVATION FOR SIMPLE LINEAR REGRESSION

- $Q(\boldsymbol{\beta}) = Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1)^2$

$$\frac{d}{d\boldsymbol{\beta}} f(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial f}{\partial \beta_0} \\ \frac{\partial f}{\partial \beta_1} \end{pmatrix} = \begin{pmatrix} -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\ 2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i \end{pmatrix}$$

- Set them to zero, we have

$$\beta_0 + \beta_1 \bar{x} = \bar{y}$$

$$\sum_{i=1}^{n} x_i^2 \beta_1 + \beta_0 n \bar{x} = \sum_{i=1}^{n} x_i y_i$$

- It follows immediately that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

LEAST SQUARES ESTIMATE: DERIVATION FOR MULTIPLE LINEAR REGRESSION

- The same principal applies but knowledge of matrix theory is needed to understand the derivation.
- Don't know or cannot remember results of matrix algebra? Not a big deal. Just search "**the matrix cookbook**"
  $Q(\boldsymbol{\beta}) = ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 = \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$

  $$\frac{\boldsymbol{d}}{\boldsymbol{d\beta}}\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}$$

  $$\frac{\boldsymbol{d}}{\boldsymbol{d\beta}}\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

- Thus, $0 = \frac{d}{d\boldsymbol{\beta}}Q(\boldsymbol{\beta})$ gives the so called **normal equation**
  $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}$

THE NORMAL EQUATION

- The normal equation is
  $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}$
- The normal equation includes a list of equations. A solution to it is an LSE of $\boldsymbol{\beta}$.
- LSE exists but might not be unique.
- When $\boldsymbol{rank}(\mathbf{X}_{n\times p}) = p$, $\mathbf{X}^T\mathbf{X}$ is invertible and the LSE is unique

  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$

THE NORMAL EQUATION: EXAMPLE

- Use the lm function in R

▶ Code

```
(Intercept)         height        female1 height:female1
1.743321211    0.072405185    0.125679857   -0.003394123
```

- Use the LSE formula in the previous slide:

▶ Code

```
                      [,1]
(Intercept)     1.743321211
height          0.072405185
female1         0.125679857
height:female1 -0.003394123
```

- Note, R and other software do not calculate the inverse of $\mathbf{X}^T\mathbf{X}$ directly; they use more efficient matrix decomposition methods

GAUSS-MARKOV THEOREM

- The LSE is optimal (minimum variance among all linear unbiased estimators) when
  - $\boldsymbol{\varepsilon} = \mathbf{0}$
  - $cov(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$
- This result is know as Gauss-Markov theorem: the LSE is the Best Linear Unbiased Estimator (BLUE) when the two above conditions hold.
- When the covariance assumption is violated, the LSE is still unbiased but it does not attain the minimum variance
  - Transform the response variable
  - Use Generalized Least Squares Estimate (GLSE).
  - Use GLM if it applies

GAUSS-MARKOV THEOREM

- Suppose that $cov(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$. Note $cov(\mathbf{Y}) = \boldsymbol{\Sigma}$.
- If $\boldsymbol{\Sigma}$ is known, we can model $\boldsymbol{\Sigma}^{-1/2}\mathbf{Y}$ on $\boldsymbol{\Sigma}^{-1/2}\mathbf{X}$ and find its LSE. This is because the transformed data $\boldsymbol{\Sigma}^{-1/2}\mathbf{Y}$ has covariance $\mathbf{I}$.
- A more practical situation is when $cov(\boldsymbol{\varepsilon}) \propto \boldsymbol{\Sigma}$. The above strategy still works.
- Example

$$cov(\boldsymbol{\varepsilon}) = \sigma^2 \begin{bmatrix} n_1 & & & \\ & n_2 & & \\ & & \ddots & \\ & & & n_K \end{bmatrix}$$

PDF OF NORMAL DISTRIBUTIONS

- Univariate normal distribution:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

- Bivariate normal distribution:

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(y_1-\mu_1)^2}{\sigma_1^2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2} - 2\rho\frac{(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2}\right)}$$

  The formula for a $p \geq 3$-dimensional multivariate normal distribution is much messier, so we use a compact way:

- Multivariate normal distribution:

$$f(\mathbf{Y}) = \frac{1}{\sqrt{(2\pi)^p |\mathbf{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{Y}-\boldsymbol{\mu})}$$

ONCE NORMAL, ALWAYS NORMAL

- Of course this cannot be true!
- What it means is that if we start from a random vector following a MVN, then any linear function of it follows a normal distribution
- Let $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then using a tool called moment generating function (MGF), one can show that
  - $\mathbf{Y} - \mathbf{M} \sim N(\boldsymbol{\mu} - \mathbf{M}, \boldsymbol{\Sigma})$
  - $a^T\mathbf{Y} \sim N(a^T\boldsymbol{\mu}, a^T\boldsymbol{\Sigma}a)$, where a is a column vector
  - $A\mathbf{Y} \sim N(A\boldsymbol{\mu}, A\boldsymbol{\Sigma}A^T)$, where A is a matrix
  - $A\mathbf{Y} \perp B\mathbf{Y} \iff A\boldsymbol{\Sigma}B^T = 0$

ONCE NORMAL, ALWAYS NORMAL: EXAMPLE 1 (THE LSE)

- Suppose that $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. $\hat{\boldsymbol{\beta}}$ is a linear function of $\mathbf{Y}$. Therefore,
  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$
  $\sim N((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1})$
  $\sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$
- $\frac{1}{\sigma}(\mathbf{X}^T\mathbf{X})^{1/2}\hat{\boldsymbol{\beta}}$ is a linear function of $\hat{\boldsymbol{\beta}}$. Therefore,

$$\frac{1}{\sigma}(\mathbf{X}^T\mathbf{X})^{1/2}\hat{\boldsymbol{\beta}} \sim N(\frac{1}{\sigma}(\mathbf{X}^T\mathbf{X})^{1/2}\boldsymbol{\beta}, \mathbf{I})$$

- We can also center it, leading to
  $\frac{1}{\sigma}(\mathbf{X}^T\mathbf{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N(\mathbf{0}, \mathbf{I})$

ONCE NORMAL, ALWAYS NORMAL: EXAMPLE 2 (THE RESIDUALS)

- The vector of residuals is
$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$
$$= [\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{Y}$$
- The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the so-called Hat matrix because $\mathbf{HY} = \hat{\mathbf{Y}}$.
- It is easy to verify that $\mathbf{H}$ satisfies two conditions
  - symmetric: $\mathbf{H} = \mathbf{H}^T$
  - idempotent: $\mathbf{H}^2 = \mathbf{H}$.
- A matrix that is both symmetric and idepotent is a projection matrix.
- You can verify that $\mathbf{I} - \mathbf{H}$ is also a projection matrix.

ONCE NORMAL, ALWAYS NORMAL: EXAMPLE 2 (THE RESIDUALS)

- $\mathbf{e}$ is linear in $\mathbf{Y}$ because $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$. It is not difficult to verify that $(\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$
- Therefore $\mathbf{e} = (\mathbf{I} - \mathbf{H})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$

PROJECTION MATRICES

- As stated earlier, a projection matrix is both symmetric and idempotent.
- An imporant and useful property is that if P is a projection matrix, then
  - $rank(P) = trace(P)$
  - Spectral decomposition:
    $P = \Gamma \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \Gamma^T$
    where Γ is an orthogonal matrix, i.e., $\Gamma\Gamma^T = \Gamma^T\Gamma = \mathbf{I}$.
- Example

▶ Code

```
[1] 4
```

▶ Code

CHI-SQUARED DISTRIBUTION: DEFINITION

- The simplest chi-squared distributed random variable can be obtained by squaring a $N(0,1)$ random variable. In other words, if $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$.
- Suppose $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I})$ and the length of $\mathbf{Z}$ is k. Then $\mathbf{Z}^T\mathbf{Z} = \sum_{i=1}^{k} Z_i^2 \sim \chi^2_k$ based on a definition of χ^2_{df}.

CONSTRUCT CHI-SQUARED RANDOM VARIABLES FROM QUADRATIC FORMS

- Let P be a projection matrix with rank df. Then

  $$\mathbf{Z}^T P \mathbf{Z} = \mathbf{Z}^T \Gamma \begin{pmatrix} \mathbf{I}_{df} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \Gamma^T \mathbf{Z}$$

  - "Once normal always normal" indicates that $\Gamma^T \mathbf{Z} \sim N(\mathbf{0}, \Gamma^T \Gamma) \sim N(\mathbf{0}, \mathbf{I})$ because Γ s.t. $\Gamma \Gamma^T = \mathbf{I}$.
  - If we let $\tilde{\mathbf{Z}} = \Gamma^T \mathbf{Z}$. We have $\tilde{\mathbf{Z}} \sim N(\mathbf{0}, \mathbf{I})$; in other words, $(\tilde{Z}_1, \cdots, \tilde{Z}_n)$ are iid $N(0, 1)$. Also

$$\mathbf{Z}^T P \mathbf{Z} = \tilde{\mathbf{Z}}^T \Gamma \begin{pmatrix} \mathbf{I}_{df} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tilde{\mathbf{Z}} = \sum_{i=1}^{df} \tilde{Z}_i^2 \sim \chi^2_{df}$$

CHI-SQUARED RANDOM VARIABLES IN LINEAR MODELS

- Recall that $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$; therefore,
  $\frac{1}{\sigma}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \sim N(\mathbf{0}, \mathbf{I})$
- Let $P = \mathbf{I} - \mathbf{H}$, the rank of which is $n - p$. (why? A property of trace needs to be used to find the rank of a projection matrix easily)
- We have $\frac{1}{\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{I} - \mathbf{H})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \sim \chi^2_{n-p}$.
- Recall that $\boldsymbol{e} = (\mathbf{I} - \mathbf{H})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ and $RSS = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}^T\boldsymbol{e}$. The result above indicates that
  $\frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$

CHI-SQUARED RANDOM VARIABLES IN LINEAR MODELS

- Recall that $\frac{1}{\sigma}(\mathbf{X}^T\mathbf{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N(\mathbf{0}, \mathbf{I}_p)$. We have
  $[\frac{1}{\sigma}(\mathbf{X}^T\mathbf{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]^T[\frac{1}{\sigma}(\mathbf{X}^T\mathbf{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \sim \chi^2_p$
- It can be simplified to
  $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\mathbf{X}^T\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2_p$
- In most situations, we are not interested in all the p elements of $\boldsymbol{\beta}$
- Let A be a $q \times p$ matrix with rank q. Then $A\boldsymbol{\beta}$ produces q linearly independent (i.e., non-redundant) linear functions of $\boldsymbol{\beta}$. How to construct a chi-squared random variable that involving $A\boldsymbol{\beta}$?

CHI-SQUARED RANDOM VARIABLES IN LINEAR MODELS

- Recall that $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N(0, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$.
- "Once normal always normal" implies that
  $A(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N(0, \sigma^2 A(\mathbf{X}^T\mathbf{X})^{-1} A^T)$
- Use "once normal always normal" again, we have
  $(\sigma^2 A(\mathbf{X}^T\mathbf{X})^{-1})A^T)^{-1/2} A(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N(0, \mathbf{I}_q)$
- The quadratic form
  $[(\sigma^2 A(\mathbf{X}^T\mathbf{X})^{-1})A^T)^{-1/2} A(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]^T [(\sigma^2 A(\mathbf{X}^T\mathbf{X})^{-1})A^T)^{-1/2} A(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \sim \chi^2_q$
  which can be simplified to
  $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T [A(\mathbf{X}^T\mathbf{X})^{-1} A^T]^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2_q$

F-STATISTIC

- Having constructed two chi-square distributed random variables, we only need one more ingredient: independence
- In RSS, the only random part is $(\mathbf{I} - \mathbf{H})\mathbf{Y}$; in the other one, the only random part is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.
- It can be verified that $cov((\mathbf{I} - \mathbf{H})\mathbf{Y}, \hat{\boldsymbol{\beta}}) = \mathbf{0}$, indicating independence, which further implies that the two quadratic forms are independent.
- We now have all the ingredients we need to bake and F-statistic
  - $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T[A(\mathbf{X}^T\mathbf{X})^{-1}A^T]^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2_q$
  - $RSS/\sigma^2 \sim \chi^2_{n-p}$
  - They are independent.

F-STATISTIC

- Because an F-distributed random variable can be constructed using two independent chi-square distributed random variables, we have

$$F = \frac{\frac{1}{\sigma^2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T [A(\mathbf{X}^T\mathbf{X})^{-1} A^T]^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/q}{\frac{1}{\sigma^2} RSS/(n-p)} \sim F_{q,n-p}$$

- F can be simplified to

$$F = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T [A(\mathbf{X}^T\mathbf{X})^{-1} A^T]^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/q}{RSS/(n-p)}$$

T-STATISTIC

- Consider $a^T\boldsymbol{\beta}$. We have shown that $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N(0, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$
- By "once normal always normal", we have
  $a^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N(0, \sigma^2 a^T(\mathbf{X}^T\mathbf{X})^{-1}a)$
- Scale the LHS, we have
  $$\frac{a^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sqrt{\sigma^2 a^T(\mathbf{X}^T\mathbf{X})^{-1}a}} \sim N(0,1)$$
- Finally,
  $$t = \frac{\frac{a^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sqrt{\sigma^2 a^T(\mathbf{X}^T\mathbf{X})^{-1}a}}}{\sqrt{\frac{RSS}{\sigma^2}/(n-p)}} \sim t_{n-p}$$
- The LHS can be simplified to
  $$t = \frac{a^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sqrt{s^2 a^T(\mathbf{X}^T\mathbf{X})^{-1}a}}$$
  where $s^2 = \frac{RSS}{n-p}$.

- Note, $\sqrt{s^2 a^T (\mathbf{X}^T \mathbf{X})^{-1} a} = se(a^T \boldsymbol{\beta})$.

A FINAL NOTE

- Most of questions of interested can be expressed into $H_0 : a^T \boldsymbol{\beta} = c$ or $H_0 : A\boldsymbol{\beta} = C$
- F or t-tests can be used.
- We discussed LSE of $\boldsymbol{\beta}$. With the assumption of MVN, we can show that the MLE and LSE of $\boldsymbol{\beta}$ are the same.