Statistics with Recitation: TA Session

Danny Po-Hsien Kang (康柏賢)

December 2, 2025

Today's agenda

1. Hypothesis Testing: ANOVA
   - ANOVA: `aov()`
2. Linear Regression
   - `lm()`
   - `summary()`
   - `geom_smooth()`

Final Week

- Final: 9:30 AM - 12:00 PM, Dec 10 (Wed)
- You can bring: One cheat sheet, Calculator.
- **No questions about R will appear in the final exam!**
- No TA Session in the final week.
  - TA office hour: 16:30-19:00 (Tue.)
- **No communication devices (smartwatches, cell phones, AI glasses,...) are allowed.**
- Remember to check FinalExam_2025h.pdf!

Today's Dataset

- Please download the two datasets from the OpenIntro website.
  - `mlb_players_18.csv`: Batter statistics for 2018 Major League Baseball season.
  - `possum.csv`: Data on possums in Australia, including site, population, sex, age, head length, and tail length.
- After that, import the data

```
mlb <- read.csv("data/mlb_players_18.csv") %>%
  filter(position != "P") %>% # Pitcher
  filter(position != "DH") # Designated Hitter

possum <- read.csv("possum.csv")
```

ANalysis Of VAriance: aov()

- **Syntax**

```
# perform ANOVA
aov_result <- aov(data$ValueVar ~ data$GroupVar)

# show result
summary(aov_result)
```

- **Example**

```
aov_result <- aov(mlb$OBP ~ mlb$position)
summary(aov_result)
```

ANalysis Of VAriance: aov()

- **Example**

```
               Df Sum Sq  Mean Sq F value  Pr(>F)
mlb$position    7 0.1074 0.015346   3.129 0.00297 **
Residuals     614 3.0118 0.004905
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fit Linear Model: `lm()`

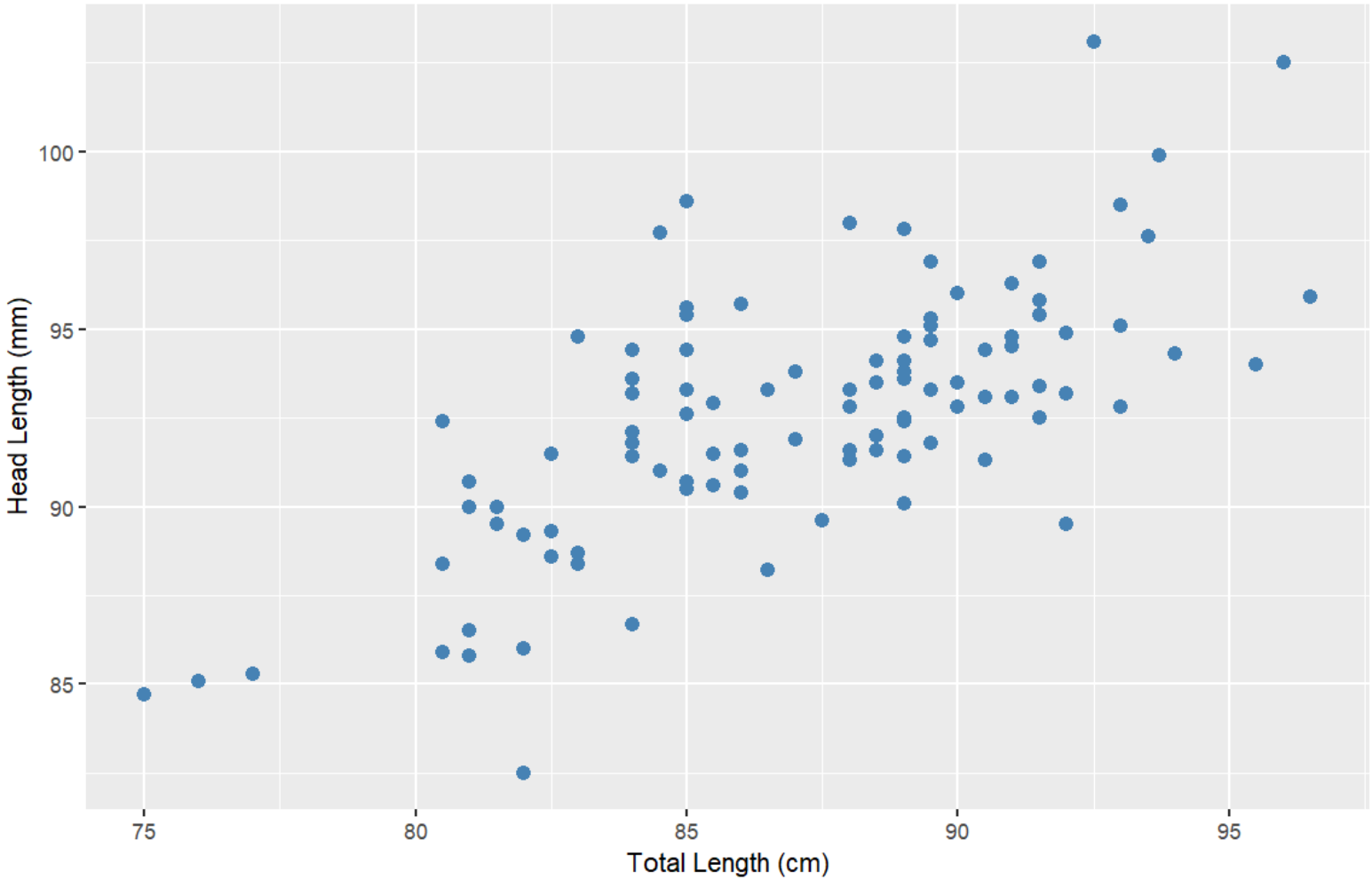

Figure: Head length and total length of the possums in Australia

Fit Linear Model: lm()

- **Syntax**

```
lm(myData$var1 ~ myData$var2)
lm(data = myData, var1 ~ var2 )
```

- **Example**

```
lm_model <- lm(possum$total_l ~ possum$head_l)
lm_model <- lm(data = possum, total_l ~ head_l)
print(lm_model)
```

Fit Linear Model: lm()

- **Output**

```
Call:
lm(formula = possum$total_l ~ possum$head_l)

Coefficients:
  (Intercept)  possum$head_l
       9.8882         0.8337
```

Obtain Regression Result: summary()

- **Syntax**

```
summary(lm_model)
```

- lm() fits the linear model and returns a model object with raw ingredients.
  - e.g. coefficients, residuals, fitted values ...
- summary() analyzes and reports that fit.
  - It computes the summary statistics most people want to read.
  - e.g. t-value, r-square ... these are not saved or computed in lm().
- Thus, we usually use summary() to report the results from lm().
  - Also for aov(), glm(), etc.
- Similar to t.test(), summary() and lm() save the results as lists.

Obtain Regression Result: summary()

- **Output**

```
Call:
lm(formula = possum$total_l ~ possum$head_l)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0881 -2.2935  0.2888  2.0801  7.4983

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.88823    8.00016   1.236    0.219
possum$head_l  0.83367    0.08633   9.657 4.68e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.131 on 102 degrees of freedom
Multiple R-squared:  0.4776,^^IAdjusted R-squared:  0.4725
F-statistic: 93.26 on 1 and 102 DF,  p-value: 4.681e-16
```

Adding Fitted Line: geom_smooth()

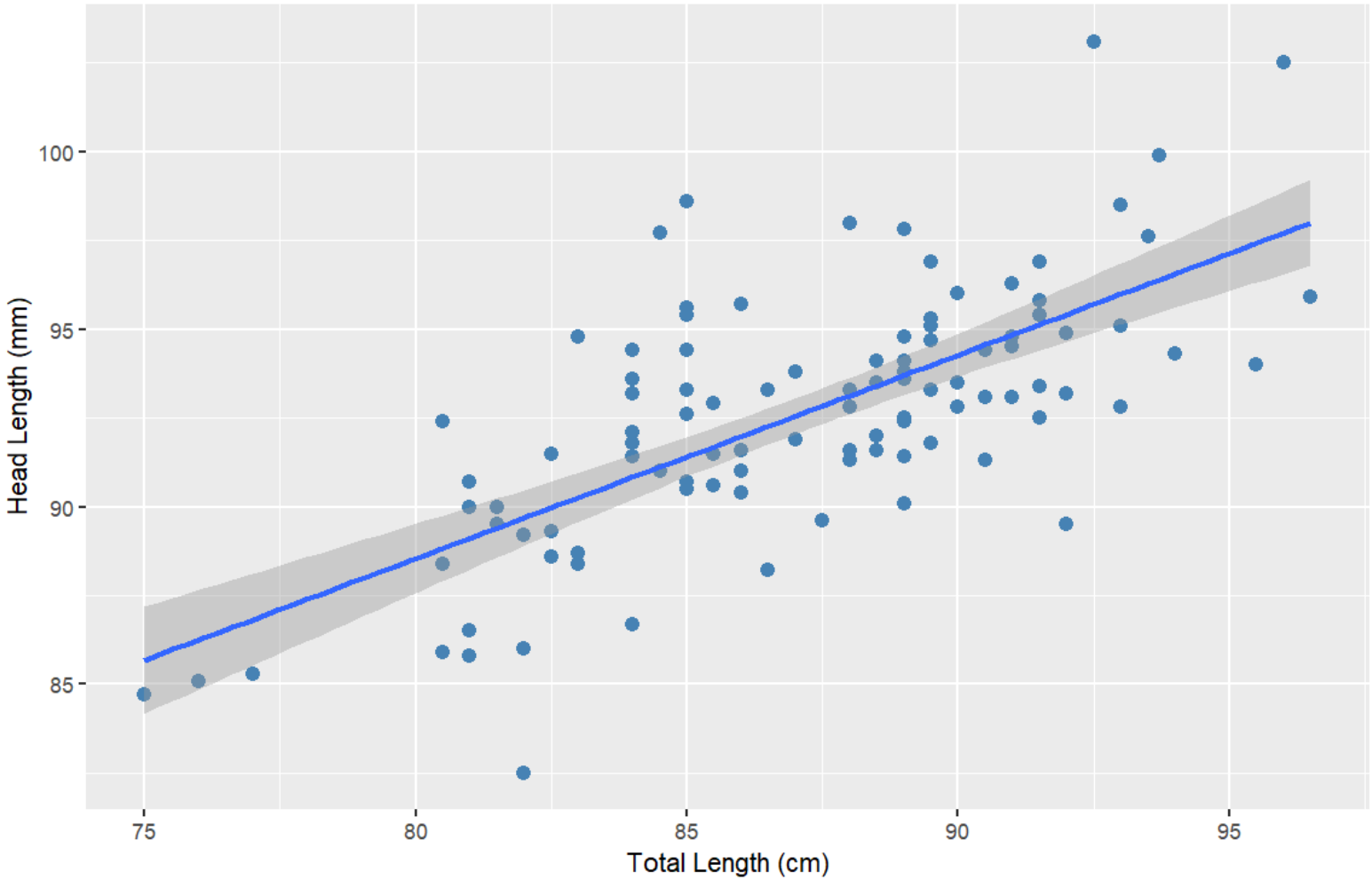

Figure: Head length and total length of the possums in Australia, with fitted line

Adding Fitted Line: geom_smooth()

- **Syntax**

```
Your_Plot +
  geom_smooth(method = "lm", se = TRUE)
```

- **Example**

```
ggplot(possum, aes(x = total_l, y = head_l)) +
  geom_point(color = "steelblue", size = 2.5) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(
    x = "Total Length (cm)",
    y = "Head Length (mm)"
  )
```