Simple Linear Regression

Inference & Prediction

Dr. Maria Tackett

09.11.19

[Click for PDF of slides]

Announcements

- HW 01 - **due Wednesday, 9/18 at 11:59p**
- Reading 01: ANOVA
- Use Piazza for questions instead of email
    - access it through Sakai
    - feel free to reply if you know the answer to question
    - let me know if you're not on Piazza

Check in

- Any questions from last class?

STA 210

Today's Agenda

- Inference for regression
- Prediction
- Cautions

Packages and Data

```
library(tidyverse)
library(broom)
library(knitr)
library(MASS) #cats dataset
```

Cats!

- When veterinarians prescribe heart medicine for cats, the dosage often needs to be calibrated to the weight of the heart.
- It is very difficult to measure the heart's weight, so veterinarians need a way to estimate it.
- One way to estimate it is using a cat's body weight which is more feasible to obtain (though still difficult depending on the cat!).
- **Goal**: Fit a regression model that describes the relationship between a cat's heart weight and body weight.

The Data

We will use the **cats** dataset from the MASS package. It contains the following characteristics for 144 cats:

- `Sex`: Male (M) or Female (F)
- `Bwt`: Body weight in kilograms (kg)
- `Hwt`: Heart weight in grams (g)

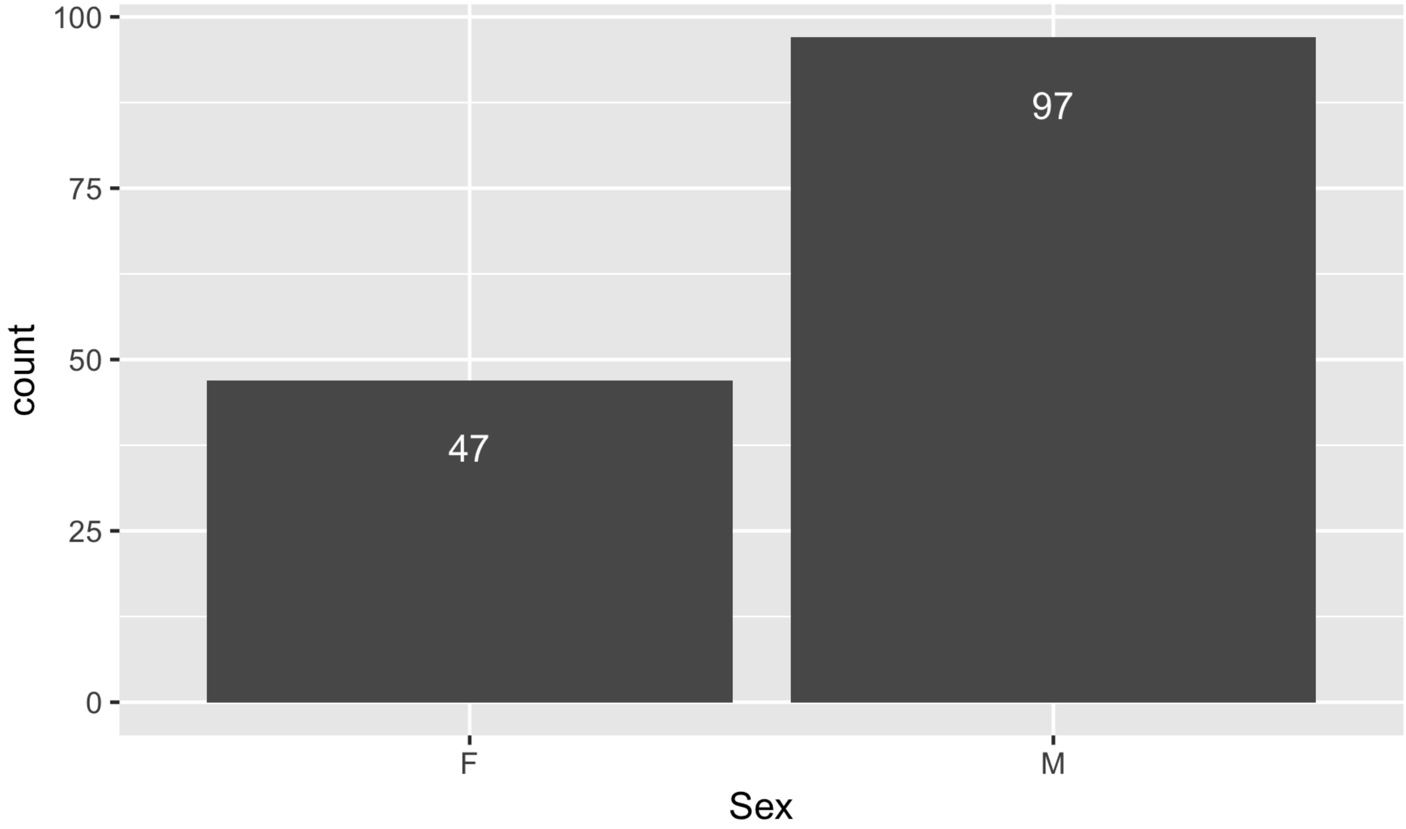
```
cats %>% slice(1:10)
```

```
##    Sex Bwt Hwt
## 1    F 2.0 7.0
## 2    F 2.0 7.4
## 3    F 2.0 9.5
## 4    F 2.1 7.2
## 5    F 2.1 7.3
## 6    F 2.1 7.6
## 7    F 2.1 8.1
## 8    F 2.1 8.2
## 9    F 2.1 8.3
## 10   F 2.1 8.5
```

Exploratory Data Analysis

STA 210

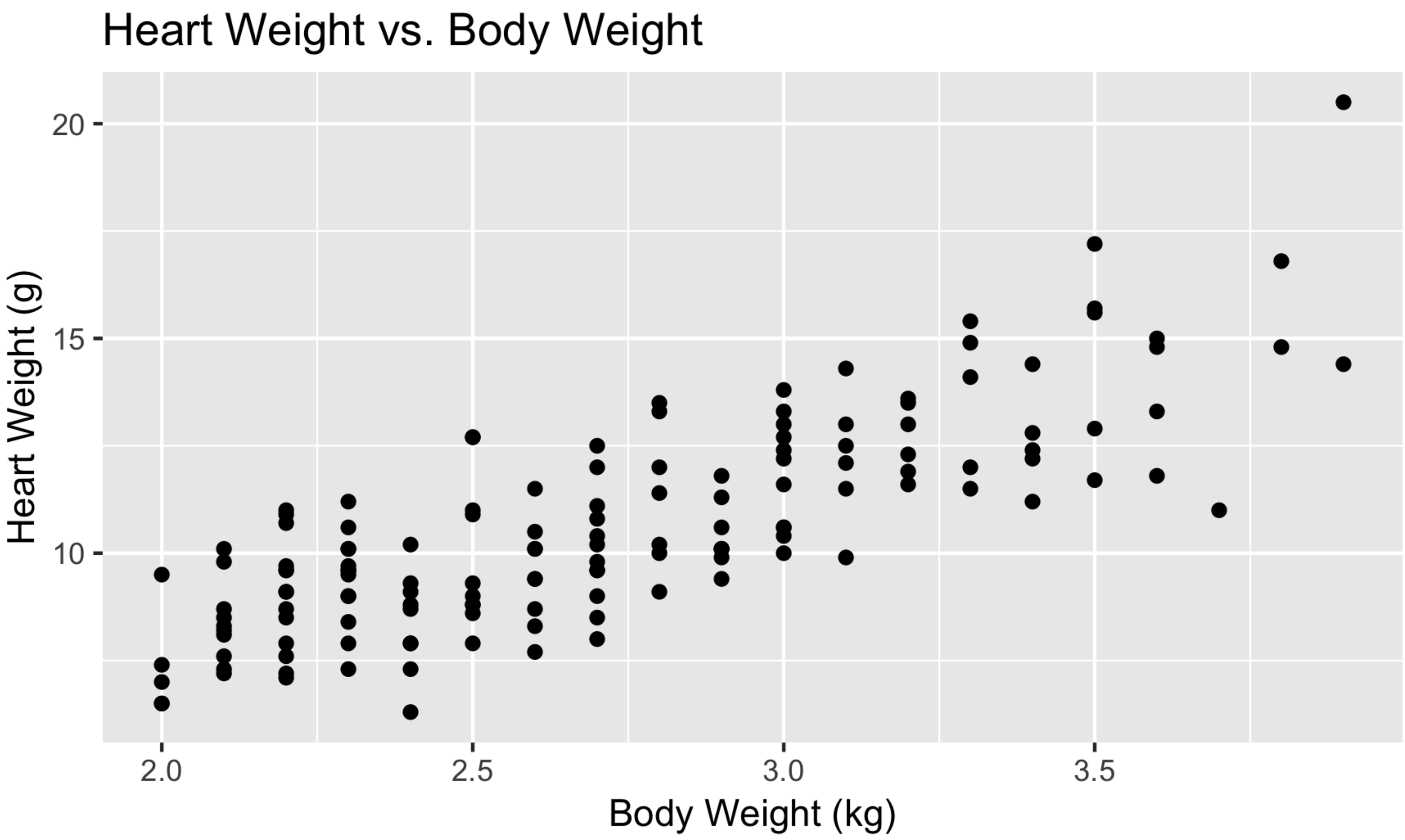
Exploratory Data Analysis



Exploratory Data Analysis

```
## Skim summary statistics
##  n obs: 144
##  n variables: 3
##
## ── Variable type:factor ─────────────────────────────────────────────
##  variable missing complete   n n_unique           top_counts ordered
##       Sex       0      144 144        2 M: 97, F: 47, NA: 0   FALSE
##
## ── Variable type:numeric ────────────────────────────────────────────
##  variable missing complete   n  mean   sd  p0  p25  p50   p75 p100
##       Bwt       0      144 144  2.72 0.49 2    2.3  2.7  3.02  3.9
##       Hwt       0      144 144 10.63 2.43 6.3 8.95 10.1 12.12 20.5
##      hist
##  ▇▆▇▂▃▃▁▁
##  ▂▇▇▅▂▁▁▁
```

Exploratory Data Analysis



Applicaiton Exercise

- Make a copy of the **cats** project on RStudio Cloud.
- Work with your lab groups to complete **Part 2: Fit the Model & Check Assumptions**
- Put your name at the top of the document. You can put everyone's name on the same document if you're working off of one computer.
- We'll look at one group's Rmd file and discuss as a class after about 10 minutes.

STA 210

Is there truly a linear relationship between the response and predictor variables?

Recall: Outline of Hypothesis Test

1. State the hypotheses
2. Calculate the test statistic
3. Calculate the p-value
4. State the conclusion in the context of the problem

STA 210

1. State the hypotheses

- We are often interested in testing whether there is a significant linear relationship between the predictor and response variables
- If there is truly no linear relationship between the two variables, the population regression slope, β_1, would equal 0
- We can test the hypotheses:

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_a &: \beta_1 \neq 0 \end{aligned}$$

- This is the test conducted by the `lm()` function in R

2. Calculate the test statistic

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

Test Statistic:

$$\text{test statistic} = \frac{\text{Estimate} - \text{Hypothesized}}{SE}$$

$$= \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

STA 210

3. Calculate the p-value

p-value is calculated from a t distribution with $n - 2$ degrees of freedom

$$\text{p-value} = P(t \geq |\text{test statistic}|)$$

STA 210

4. State the conclusion

Magnitude of p-value	Interpretation
p-value < 0.01	strong evidence against H_0
0.01 < p-value < 0.05	moderate evidence against H_0
0.05 < p-value < 0.1	weak evidence against H_0
p-value > 0.1	effectively no evidence against H_0

Note: These are general guidelines. The strength of evidence depends on the context of the problem.

Cats!: Hypothesis test for β_1

- Refer back to the **`cats`** application exercise to answer the following questions:

a. State the hypotheses in (1) words and (2) statistical notation.

b. What is the meaning of the test statistic in the context of the problem?

c. What is the meaning of the p-value in the context of the problem?

d. State the conclusion in context of the problem.

Predictions

Predictions for New Observations

- We can use the regression model to predict for a response at x_0

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

- Because the regression models produces the mean response for a given value of x_0, it will produce the same estimate whether we want to predict the mean response at x_0 or an individual response at x_0

STA 210

Predicting Mindy's heart weight

My cat Mindy weighs about 3.18 kg (7 lbs).

What is her predicted heart weight?

$$\begin{aligned} \hat{hwt} &= -0.3567 + 4.0341 \times \color{red}{3.18} \\ &= 12.472 \text{ grams} \end{aligned}$$

STA 210

Uncertainty in predictions

- There is uncertainty in our predictions, so we need to calculate an a standard error (SE) to capture the uncertainty
- The SE is different depending on whether you are predicting an average value or an individual value
- SE is larger when predicting for an individual value than for an average value

STA 210

Standard errors for predictions

Predicting the mean response

$$SE(\hat{\mu}) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}}$$

Predicting an individual response

$$SE(\hat{y}) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}}$$

CI for predicted heart weight

- Calculate a 95% prediction interval for Mindy's predicted heart weight.

```
x0 <- data.frame(Bwt = c(3.18))
predict.lm(bwt_hwt_model, x0, interval = "prediction",
           conf.level = 0.95)
```

```
##        fit      lwr      upr
## 1 12.47166 9.581804 15.36151
```

- Calculate a 95% confidence interval for the predicted mean heart weight for the subset of cats who weigh 3.18 kg.

```
x0 <- data.frame(Bwt = c(3.18))
predict.lm(bwt_hwt_model, x0, interval = "confidence",
           conf.level = 0.95)
```

```
##        fit      lwr      upr
## 1 12.47166 12.14269 12.80063
```

Cautions

Caution: Extrapolation

- The regression is only useful for predictions for the response variable y in the range of the predictor variable x that was used to fit the regression

- It is risky to predict far beyond that range of x, since you don't have data to tell whether or not the relationship continues to follow a straight line

STA 210

Caution: Extrapolation

My cat Andy weighs about 8.60 kg (10 lbs).

Should we use this regression model to predict his heart weight?

min	q1	median	q3	max
2	2.3	2.7	3.025	3.9

The heaviest cat in this dataset weighs 3.9 kg (8.6 lbs). We should **not** use this model to predict Andy's heart weight, since that would be a case of extrapolation.

Caution: Correlation $\neq$ Causation

- The regression model is <u>**not**</u> a statement of causality
    - The regression model provides a description of the averages of Y for different values of X
- The regression model alone <u>**cannot**</u> prove causality. You need either
    - Randomized experiment
    - Observational study in which all relevant confounding variables are controlled for adequately