

# Linear Regression

---

Max Turgeon

STAT 3150—Statistical Computing

# Lecture Objectives

- Fit linear regression models using R.
- Understand the output.
- Compare and contrast with t-tests and ANOVA.

# Motivation

- In the last module, we discussed how bootstrap can be used to approximate the sampling distribution.
  - The general idea is based on replacing the true CDF by the empirical CDF (i.e. sampling with replacement).
- If we make assumptions about the data-generating mechanism, we can sometimes improve on the general bootstrap.
- We will explore this idea using **linear regression**.
  - But first, we will review it!

## General notation

- Linear regression estimates the relationship between a single variable  $Y$ , called the *outcome variable*, and a series of variables  $X_1, \dots, X_p$ , called *covariates*.
  - Machine learning uses target and features, respectively.
- The outcome  $Y$  is typically a continuous variable.
  - Eg. Height, income, blood pressure, etc.
- The covariates  $X_1, \dots, X_p$  can be anything.
- We want to collect all variables  $Y, X_1, \dots, X_p$  on the same unit of observation (e.g. person, school, animal, olive oil).

# Simplest linear regression

- The simplest linear regression only has an outcome variable  $Y$ , no covariates.
  - It's equivalent to a one-sample t-test.
- The linear regression equation can be written as

$$E(Y) = \beta_0.$$

- In other words, we are saying the population mean of  $Y$  (i.e.  $E(Y)$ ) is equal to a parameter  $\beta_0$ .
  - This notation is a bit overkill, but it will make more sense soon...

## Example i

```
library(readr)
dataset <- read_csv("heart.csv")

# Use function lm
fit <- lm(age ~ 1, data = dataset)
fit
```

## Example ii

```
##  
## Call:  
## lm(formula = age ~ 1, data = dataset)  
##  
## Coefficients:  
## (Intercept)  
##          54.37
```

```
# To compute confidence interval  
# use confint  
confint(fit)
```

## Example iii

```
##                2.5 %    97.5 %  
## (Intercept) 53.3396 55.39307
```



# One binary covariate i

- The next simplest linear regression has a single covariate  $X$  which can take only two values: 0 or 1.
- The idea is that it encodes a binary variable.
  - Eg. Male: 1; Female 0. CS Major: 1; Non-CS Major: 0.
- The linear regression equation for this situation can be written as

$$E(Y|X) = \beta_0 + \beta_1 X.$$

- Let's unpack this:
  - When  $X = 0$ , the RHS simplifies to  $\beta_0$ , and we get

$$E(Y|X = 0) = \beta_0.$$

## One binary covariate ii

- When  $X = 1$ , we get

$$E(Y|X = 1) = \beta_0 + \beta_1.$$

- In other words,  $\beta_0$  represents the population mean of  $Y$  when  $X = 0$ , but  $\beta_1$  represents the **difference** in population means between the two subgroups.
  - If  $\beta_1$  is significantly different from 0, then we have evidence of a difference in means between the two groups!

## Example i

```
fit <- lm(age ~ sex, data = dataset)
```

```
fit
```

```
##
```

```
## Call:
```

```
## lm(formula = age ~ sex, data = dataset)
```

```
##
```

```
## Coefficients:
```

```
## (Intercept)          sex
```

```
##      55.677      -1.919
```

## Example ii

```
confint(fit)
```

```
##                2.5 %      97.5 %  
## (Intercept) 53.858832 57.4953350  
## sex         -4.118464  0.2812059
```

```
# What if we change the coding 0/1 to female/male?  
dataset <- transform(dataset, sex = ifelse(sex == 1,  
                                             "male",  
                                             "female"))
```

## Example iii

```
fit <- lm(age ~ sex, data = dataset)
```

```
fit
```

```
##
```

```
## Call:
```

```
## lm(formula = age ~ sex, data = dataset)
```

```
##
```

```
## Coefficients:
```

```
## (Intercept)      sexmale
```

```
##      55.677      -1.919
```

## Summary so far

- We saw how linear regression connects the average value of an outcome variable with covariates.
- **Important:** the regression coefficient  $\beta_1$  measures a *difference* in means.
- With a single binary covariate, we recover the two-sample t-test.

# Exercise

Using the heart dataset, determine whether average cholesterol levels are different between men and women.

## Solution i

```
fit <- lm(chol ~ sex, data = dataset)
confint(fit)
```

```
##                2.5 %    97.5 %
## (Intercept) 251.08108 271.52308
## sexmale     -34.37824  -9.64622
```



## One continuous covariate i

- Next we look at the case of a single *continuous* covariate  $X$
- The linear regression equation for this situation can also be written as

$$E(Y|X) = \beta_0 + \beta_1 X.$$

- Let's unpack this:
  - When  $X = 0$ , the RHS still simplifies to  $\beta_0$ , and we get

$$E(Y|X = 0) = \beta_0.$$

- Let's compare two values of  $X$  that differ by 1 unit, e.g.  $x$  and  $x + 1$ . We have

$$\begin{aligned}E(Y|X = x) &= \beta_0 + \beta_1 x \\E(Y|X = x + 1) &= \beta_0 + \beta_1(x + 1) \\&= (\beta_0 + \beta_1 x) + \beta_1 \\&= E(Y|X = x) + \beta_1.\end{aligned}$$

- Rearranging, we get

$$\beta_1 = E(Y|X = x + 1) - E(Y|X = x).$$

## One continuous covariate iii

- In other words,  $\beta_1$  represents the **difference** in population means between two subgroups that differ by 1 unit in their value of the covariate  $X$ .
- $\beta_0$  still represents the population mean of  $Y$  when  $X = 0$ .
  - But depending on what  $X$  represent (e.g. age, cholesterol),  $X = 0$  may not be possible!

## Example i

```
fit <- lm(age ~ chol, data = dataset)
```

```
fit
```

```
##
```

```
## Call:
```

```
## lm(formula = age ~ chol, data = dataset)
```

```
##
```

```
## Coefficients:
```

```
## (Intercept)          chol
```

```
##      45.14573      0.03744
```

## Example ii

```
confint(fit)
```

```
##                2.5 %        97.5 %  
## (Intercept) 40.25979148 50.03166732  
## chol        0.01802571  0.05685821
```

- *Interpretation:* the estimated value of  $\beta_1$  is 0.04, which means that two groups of people from the study who differ in their cholesterol levels by 1 unit on average differ in their age by 0.04 year (i.e. about two weeks), with a higher cholesterol level being associated with being older.

## Exercise

It is perhaps more natural to think of the difference in cholesterol levels for groups of different ages. Using the heart dataset, determine the average difference in cholesterol levels for people in the study whose age differ by one year.

## Solution i

```
fit <- lm(chol ~ age, data = dataset)
```

```
fit
```

```
##
```

```
## Call:
```

```
## lm(formula = chol ~ age, data = dataset)
```

```
##
```

```
## Coefficients:
```

```
## (Intercept)          age
```

```
##      179.967         1.219
```

## Solution ii

```
confint(fit)
```

```
##                2.5 %      97.5 %  
## (Intercept) 145.1132605 214.821681  
## age         0.5870764   1.851806
```

- *Interpretation:* the estimated value of  $\beta_1$  is 1.22, which means that two groups of people from the study who differ in their age by 1 year on average differ in their cholesterol levels by 1.22 mg/dl, with being older being associated with higher cholesterol level.



# Inspecting the model fit i

- We can use a scatter plot to investigate the model fit, i.e. whether the regression equation is a good description of the data.
  - But only really helpful when both  $Y$  and the single covariate  $X$  are continuous.

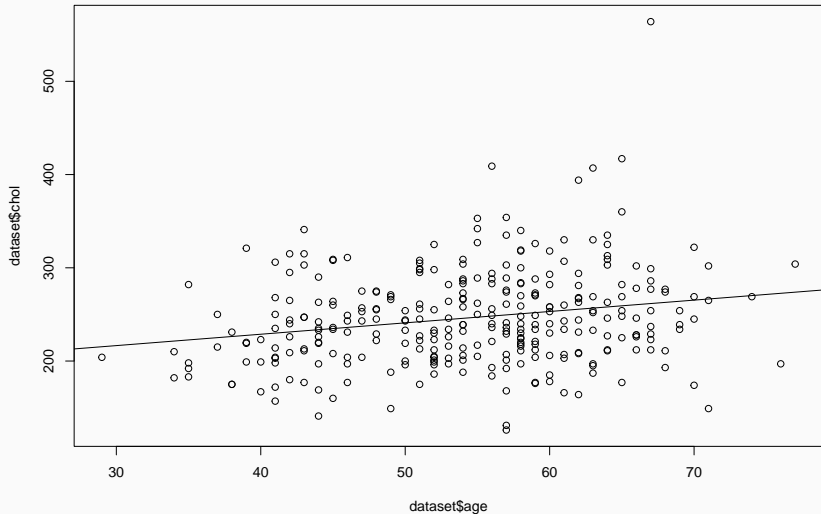
```
fit <- lm(chol ~ age, data = dataset)
# Extract coefficient estimates with coef
coef(fit)
```

## Inspecting the model fit ii

```
## (Intercept)          age  
## 179.967471      1.219441
```

```
plot(dataset$age, dataset$schol)  
abline(a = coef(fit)[1],  
       b = coef(fit)[2])
```

## Inspecting the model fit iii



## Categorical covariate $i$

- Now, let's assume we measured a continuous outcome variable  $Y$  across different subgroups.
  - For simplicity, we'll assume only three subgroups, but this can easily be generalized.
- Let  $Z$  keep track of which subgroup an observation is from.
  - Eg.  $Z = 1$  for CS major,  $Z = 2$  for Psych Major, and  $Z = 0$  for non-CS, non-Psych major.
- We want to compare the average value of  $Y$  between all subgroups.
  - How can we fit this into linear regression?

## Categorical covariate ii

- Solution: we introduce *dummy* variables  $X_1$  and  $X_2$ .
  - $X_1 = 1$  if  $Z = 1$ , and  $X_1 = 0$  otherwise.
  - $X_2 = 1$  if  $Z = 2$ , and  $X_2 = 0$  otherwise.

$Z$	$X_1$	$X_2$
0	0	0
1	1	0
2	0	1

## Categorical covariate iii

- The linear regression equation for this situation can then be written as

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

- Let's unpack this:
  - When  $Z = 0$ , both  $X_1 = 0$  and  $X_2 = 0$ , and so the RHS simplifies to  $\beta_0$ :

$$E(Y|Z = 0) = \beta_0.$$

- When  $Z = 1$ , we have  $X_1 = 1$  and  $X_2 = 0$ , and so the RHS simplifies to  $\beta_0 + \beta_1$ :

$$E(Y|Z = 1) = \beta_0 + \beta_1.$$

- When  $Z = 2$ , we have  $X_1 = 0$  and  $X_2 = 1$ , and so the RHS simplifies to  $\beta_0 + \beta_2$ :

$$E(Y|Z = 2) = \beta_0 + \beta_2.$$

- In other words,  $\beta_1$  represents the **difference** in population means between the two subgroups  $Z = 0$  and  $Z = 1$ , and  $\beta_2$  represents the **difference** between the two subgroups  $Z = 0$  and  $Z = 2$
- $\beta_0$  still represents the population mean of  $Y$  when  $Z = 0$ .

## Example i

```
library(dslabs)
```

```
table(olive$region)
```

```
##
```

```
## Northern Italy      Sardinia Southern Italy
```

```
##           151           98           323
```

```
fit <- lm(oleic ~ region, data = olive)
```

```
fit
```



## Example ii

```
##  
## Call:  
## lm(formula = oleic ~ region, data = olive)  
##  
## Coefficients:  
## (Intercept) regionSardinia regionSouthern Italy  
## 77.93 -5.25 -6.93  
  
confint(fit)
```

## Example iii

##		2.5 %	97.5 %
## (Intercept)		77.484107	78.37695
## regionSardinia		-5.961921	-4.53873
## regionSouthern Italy		-7.471234	-6.38964

## Example iv

### Interpretation

- The average level of oleic acid for olive oils in Northern Italy (i.e. the reference category) is  $\beta_0 = 77.9$ .
- The average level of oleic acid for olive oils in Sardinia is  $\beta_0 + \beta_1 = 72.7$ .
- The average level of oleic acid for olive oils in Southern Italy is  $\beta_0 + \beta_2 = 71$ .
- The average level of oleic acid is highest in Northern Italy, and it's significantly different from that of other regions.
  - But these confidence intervals can't tell us whether the average levels are different between Sardinia and Southern Italy.

# Summary

- We introduced linear regression as a general framework for comparing means between subgroups.
- We saw how one-sample and two-sample t-tests are special cases.
- By introducing dummy variables, we can also get ANOVA as a special case.
- Next lecture: we will discuss the assumptions behind linear regression.