

# Lecture 4: Simple Linear Regression Assumptions

Module 1: part 3

Spring 2024

# Logistics

- Wrap up Module 1 today
- Module assessment due on Feb 11 11:59pm
- Module 2 will consider regression with multiple covariates
- Office hour locations: Daniel and Tathagata (Comstock 1187); Nayel in Surge B 159.

## Recap

The equation for a line can be put into the following form

$$Y = b_0 + b_1X \quad (1)$$

# Recap

The equation for a line can be put into the following form

$$Y = b_0 + b_1X \quad (1)$$

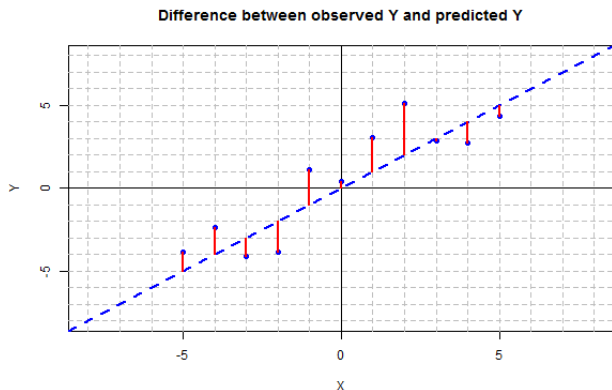
- $X$  and  $Y$  are variables
- $b_0$  is the **Y-intercept**. It is the value of the  $Y$  coordinate when  $X = 0$
- $b_1$  is the **slope**. It describes how  $Y$  changes as  $X$  changes.

# Recap

Suppose we observe  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ .

To select a “best line” we consider the difference between the predicted point and observed value of  $y_i$  and choose  $\hat{b}_0$  and  $\hat{b}_1$  to minimize the RSS:

$$RSS(\hat{b}_0, \hat{b}_1) = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (\hat{b}_0 + \hat{b}_1 x_i))^2 \quad (2)$$



# Recap

## Outliers:

- Points which have  $x$  values far from  $\bar{x}$  have high leverage
- Points which have high leverage may also have high influence; i.e., change the estimate when included/excluded
- When to include or exclude points with high influence?

# Linear Model

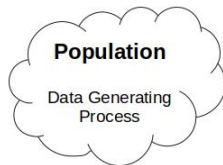
# Interpretation

Let's take a step back and consider what we have calculated

- Still have “hat's” on  $\hat{b}_0$  and  $\hat{b}_1$  because they are calculated from the sample data
- We want to use the sampled data to infer something about the population



# Sample data vs population distribution



**Data**

A screenshot of a data table with multiple columns and rows of numerical data.



**Statistic**



# Linear Models

Much of what we've talked about so far involves calculating coefficients which describe a specific set of data

- Given a sample of data  $(x_1, y_1), (x_2, y_2) \dots, (x_n, y_n)$ , calculate line which minimizes RSS
- Sample is all we have, but most often we are interested in quantities which describe a population
- Given a new sample (potentially repeating the experiment) will give different estimates of  $\hat{b}_0$  and  $\hat{b}_1$
- What can we say about  $\hat{b}_0, \hat{b}_1$  and the “true” population process?

# Linear Model Assumptions

Commonly used linear model where  $\varepsilon_i$  is an error term:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

# Linear Model Assumptions

Commonly used linear model where  $\varepsilon_i$  is an error term:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

## Assumptions of the model:

- Linear function:  $E(Y_i | X_i = x) = b_0 + b_1 x$
- Independence across observations:  $\varepsilon_i$  is independent of  $\varepsilon_k$  where  $i$  and  $k$  denote different observations
- Independence of errors:  $\varepsilon_i$  is independent of  $X_i$  with mean 0 and variance  $\sigma^2$

# Linear Model Assumptions

Commonly used linear model where  $\varepsilon_i$  is an error term:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

## Assumptions of the model:

- Linear function:  $E(Y_i | X_i = x) = b_0 + b_1 x$
- Independence across observations:  $\varepsilon_i$  is independent of  $\varepsilon_k$  where  $i$  and  $k$  denote different observations
- Independence of errors:  $\varepsilon_i$  is independent of  $X_i$  with mean 0 and variance  $\sigma^2$

## Less important assumption:

- Normality: sometimes, we assume that  $\varepsilon_i \sim N(0, \sigma^2)$

# Model Implications

Conditional expectation:  $E(Y_i | X_i = x) = b_0 + b_1x$

# Model Implications

Conditional expectation:  $E(Y_i | X_i = x) = b_0 + b_1x$

## Interpretation

- $b_0$  is the expected value of  $Y_i$  when conditioning on  $X_i = 0$
- $b_1$  is the difference of the expected value of  $Y_i$  when conditioning on values of  $X_i$  which differ by 1 unit.

$$b_1 = E(Y_i | X_i = x + 1) - E(Y_i | X_i = x)$$

# Conditional Expectation

In general, the conditional expectation is not the same as “intervening” on  $X$



# Conditional Expectation

In general, the conditional expectation is not the same as “intervening” on  $X$

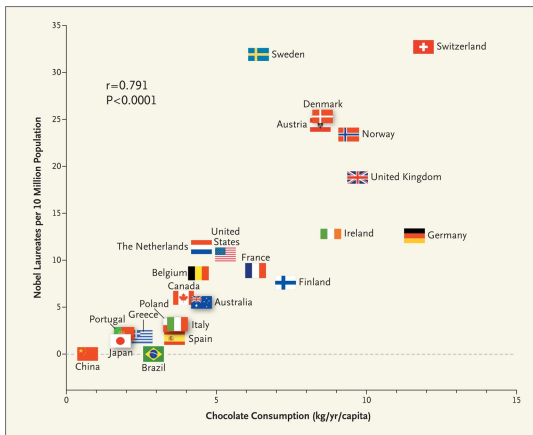


Figure: Messerli 2012, NEJM

# Interpretation

## Correct Interpretations

- Given two observations whose  $X$  values differ by 1 unit, we would **expect** the observation with the larger  $X$  value to have a  $Y$  value  $b_1$  units larger than the observation with the smaller  $X$  value
- Given two observations whose  $X$  values differ by 1 unit, **on average** the observation with the larger  $X$  value will have a  $Y$  value  $b_1$  units larger than the observation with the smaller  $X$  value

# Interpretation

## Correct Interpretations

- Given two observations whose  $X$  values differ by 1 unit, we would **expect** the observation with the larger  $X$  value to have a  $Y$  value  $b_1$  units larger than the observation with the smaller  $X$  value
- Given two observations whose  $X$  values differ by 1 unit, **on average** the observation with the larger  $X$  value will have a  $Y$  value  $b_1$  units larger than the observation with the smaller  $X$  value
- A 1 unit difference in  $X$  is associated with a  $b_1$  unit difference in  $Y$

# Interpretation

## Correct Interpretations

- Given two observations whose  $X$  values differ by 1 unit, we would **expect** the observation with the larger  $X$  value to have a  $Y$  value  $b_1$  units larger than the observation with the smaller  $X$  value
- Given two observations whose  $X$  values differ by 1 unit, **on average** the observation with the larger  $X$  value will have a  $Y$  value  $b_1$  units larger than the observation with the smaller  $X$  value
- A 1 unit difference in  $X$  is associated with a  $b_1$  unit difference in  $Y$

## Incorrect Interpretations

- Increasing  $X$  by 1 unit increases  $Y$  by  $b_1$  units
- A 1 unit increase in  $X$  causes  $Y$  to increase by  $b_1$  units

# Statistic is unbiased

Under the assumptions that  $\varepsilon_i$  is independent of  $X_i$ , we have:

$$E(\hat{b}_1) = b_1$$

$$E(\hat{b}_0) = b_0$$

so that the estimated values are “unbiased” estimators of the true values

- If you replicate the experiment many different times, you will get a different estimate, each time, but the average will be the “truth”

## Potentially helpful (but not necessary) math

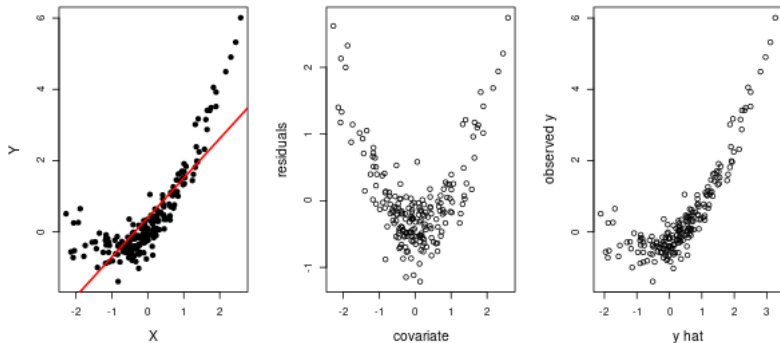
Under the assumptions, we have:

$$\bar{y} = \frac{1}{n} \sum_i (b_0 + b_1 x_i + \varepsilon_i) = b_0 + \frac{1}{n} \sum_i b_1 x_i + \frac{1}{n} \sum_i \varepsilon_i = b_0 + b_1 \bar{x} + \bar{\varepsilon}$$

$$\begin{aligned} E(\hat{b}_1 | X) &= E \left( \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \mid X \right) \\ &= E \left( \frac{\sum_i (b_0 + b_1 x_i + \varepsilon_i - b_0 - b_1 \bar{x} - \bar{\varepsilon})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \mid X \right) \\ &= E \left( \frac{b_1 \sum_i (x_i - \bar{x})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \mid X \right) + \underbrace{E \left( \frac{\sum_i (\varepsilon_i - \bar{\varepsilon})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \mid X \right)}_{\text{cov}(\varepsilon_i, X_i)=0} \\ &= b_1 + 0 \end{aligned}$$

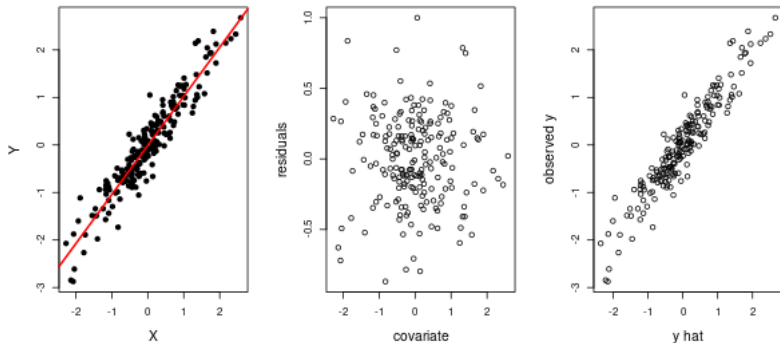
# Model Assumptions: Linearity

Look for patterns in residuals if the linearity assumption is violated



# Model Assumptions: Linearity

Look for patterns in residuals if the linearity assumption is violated





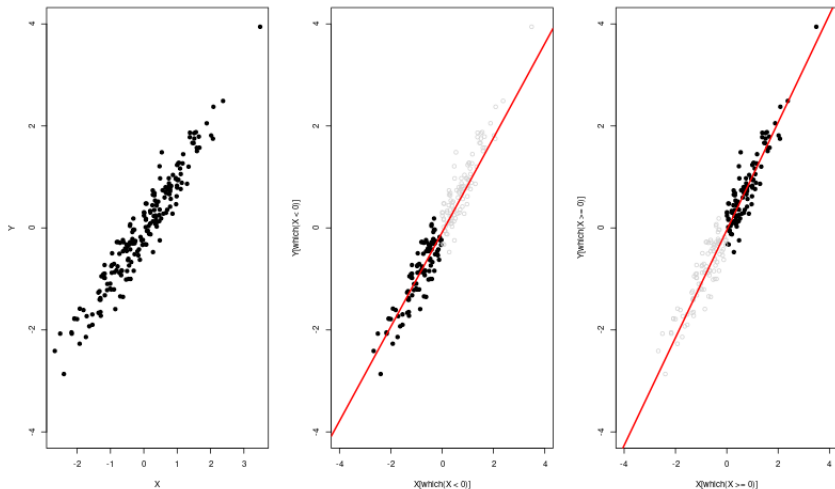
# Model Assumptions: Linearity

What happens if the linearity assumption is violated?

- Consider transforming your data with a non-linear transformation
- Adding other covariates can be “helpful”
- $b_1$  no longer corresponds to change in conditional expectation, but the sign of coefficient can still be useful for interpretability
- Parameters are the best “linear approximation”
- Best linear approximation depends on the range of the  $X$  values

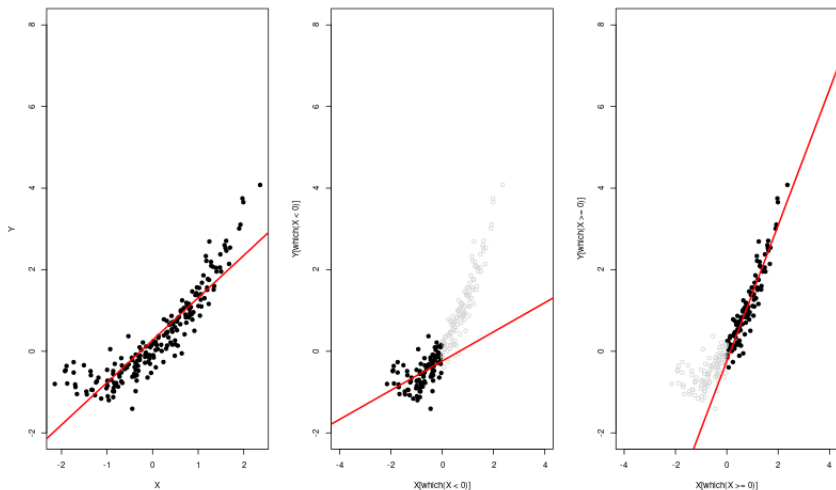
# Model Assumptions: Linearity

Best linear approximation depends on the range of the  $X$  values



# Model Assumptions: Linearity

Best linear approximation depends on the range of the  $X$  values

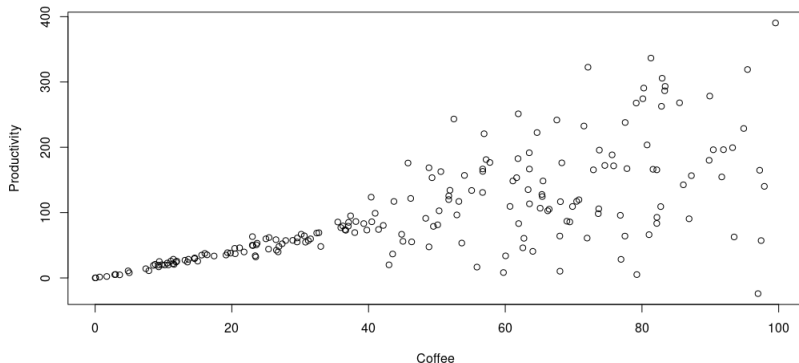


# Model Assumptions: independence across observations

# Model Assumptions: independence of error and covariate

We made a strong assumption that  $\varepsilon_i$  is mean 0 and independent of  $X_i$

- What if the variance of  $\varepsilon_i$  depends on  $X_i$ ? i.e., model is heteroscedastic
- As long as  $E(\varepsilon_i | X_i) = 0$ , estimates are still unbiased  $E(\hat{b}_1) = b_1$
- Will effect testing procedures!



# Discussion

- What is a scientific question that you are interested in?
- Are you trying to do prediction or modeling?
- Are the assumptions we discussed today reasonable for your setting?
  - Linearity
  - Independence across observations
  - Independence of errors and covariates

# Assessing explanatory power

# Components of the squared error

How can we assess how useful the explanatory variable is for predicting the response variable?

$$\begin{aligned}(y_i - \bar{y}) &= (y_i - \hat{y}_i + \hat{y}_i - \bar{y}) \\ &= (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \\ &= \text{residual} + \text{predicted deviation from mean}\end{aligned}\tag{3}$$



# Components of the squared error

How can we assess how useful the explanatory variable is for predicting the response variable?

$$\begin{aligned}(y_i - \bar{y}) &= (y_i - \hat{y}_i + \hat{y}_i - \bar{y}) \\ &= (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \\ &= \text{residual} + \text{predicted deviation from mean}\end{aligned}\tag{3}$$

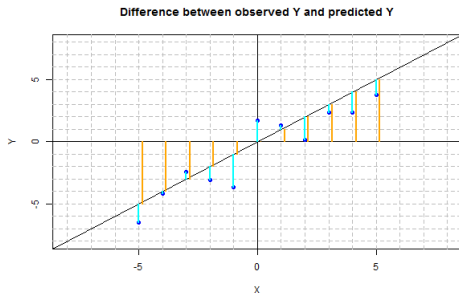
Using a bit of algebra, we can decompose the total sum of squares for  $Y$  into

$$SS_{total} = \sum_i (y_i - \bar{y})^2 = \underbrace{\sum_i (\hat{y}_i - \bar{y})^2}_{SS_{\text{regression}}} + \underbrace{\sum_i (y_i - \hat{y}_i)^2}_{SS_{\text{error}}}\tag{4}$$

# Components of the squared error

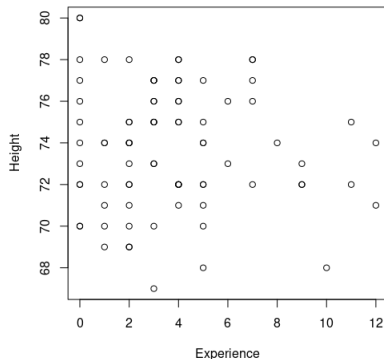
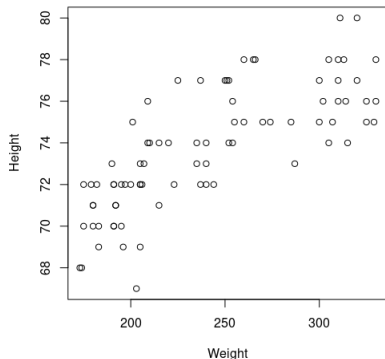
If  $SS_{\text{regression}}$  is large compared to  $SS_{\text{error}}$ , then the explanatory variable is a good predictor of the response variable

$$\underbrace{1 - \frac{SS_{\text{error}}}{SS_{\text{total}}} = \frac{SS_{\text{regression}}}{SS_{\text{total}}} = \frac{\sum_i (\hat{y}_i - \bar{y})}{\sum_i (y_i - \bar{y})} = r_{XY}^2}_{\text{Referred to as } R^2} \quad (5)$$



## Example: Components of the squared error

The  $R^2$  for height and weight is .59 while the  $R^2$  for height and experience is .01.



# Wrap-up

- If we assume the true population process is a linear model, we can describe properties of the estimated regression coefficients
- Estimated slope is estimated difference in conditional expectation associated with difference in  $X$
- If assumptions are violated, interpretation is not as straightforward
- Explanatory power of regression can be summarized by  $R^2$  value
- Next module will consider setting with more than 1 covariate