

Data Analysis and Visualisation

Lecture 4

Linear Regression

Introduction

- ▶ Today we'll look at
 - ▶ Regression Equations
 - ▶ Standard Error of the Estimate
- ▶ To this point we have looked at the properties of data - describing its size and shape, and whether variables are related to each other
- ▶ Now we can go on to start using that data to make predictions

Introduction

- ▶ A prediction- a best guess as to the value of something.
- ▶ Good prediction
Minimal Difference between the predicted value and the actual value.
- ▶ Predicted value= Mean value If no further information about a value is available.

Example 1:

If Bob's Test Score is to be predicted and mean score for students like Bob is given as 50,

Prediction = Bob's score is 50

- ▶ A better prediction possible when more information available

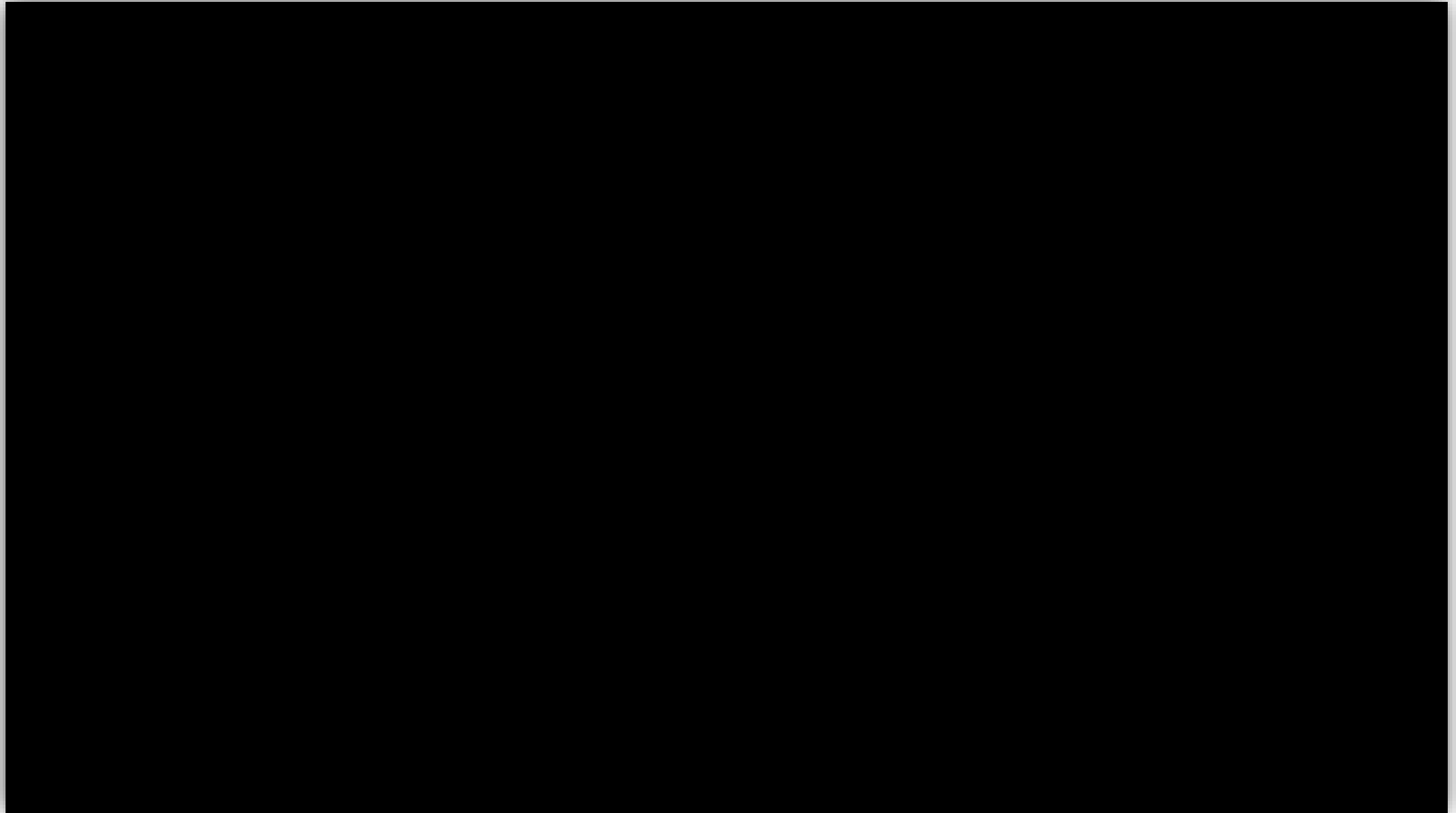
Introduction

▶ ***Example 2:***

▶ Given:

1. positive correlation between shoe size and weight.
2. Jim's shoe size is 7
3. John's shoe size is 12

What can you Predict for john's weight ??



Regression Equations

- ▶ Accuracy of predictions ↑ when *Regression equation* is used.

Simple Linear Regression

$$Y = b_0 + b_1X$$

Multiple Linear Regression

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_mX_m$$

$$Y = 0.9 + 1.2X_1 + 2X_2 + 4X_3 + 1X_4$$

- ▶ The higher the correlation, the better prediction
- ▶ r_{xy} is +1.00 or -1.00, then prediction is perfect

Regression Equations

$$\hat{Y} = \bar{Y} + b(X - \bar{X})$$

Where,

- ▶ $\hat{Y}$ - Predicted value of Y
- ▶ $\bar{Y}$ - Mean of the Y distribution
- ▶ $\bar{X}$ - Mean of the X distribution
- ▶ b is a weighting factor that represents the strength of the correlation between X and Y

$$b = r_{xy} \left(\frac{SD_y}{SD_x} \right)$$

where r_{xy} is Pearson's r_{xy} and SD_x , SD_y are the standard deviations of X and Y

▶ Example:

A researcher has been collecting information about teenage school achievement and family life. She discovered that the average high schooler in her community spends an average of 5.3 hours per week on homework, with $SD = 1.4$; she also learned that the parents of these students pay an average of \$1.32 for each grade of “A” their kids earn, with $SD = .35$. Finally, she calculated that there is a $r_{xy} = +.43$ between money paid for grades of “A” and time spent on homework. Putting all that in table form, we have the following :

Payment per grade of “A” (X): $\bar{X} = 1.32$

$$SD_x = 0.35$$

Hours of homework per week (Y): $\bar{Y} = 5.30$

$$SD_y = 1.40$$

$$r_{xy} = +.43$$

This researcher happens to have a teenage son, Marty, whom she pays a dollar and a half for every “A” he brings home. Can you predict how many hours per week Marty will study next term?

To find the answer to this question, first find the value of b:

- ▶ Two attributes, X and Y are found to have a correlation of 0.43

$$\bar{X} = 1.32, \quad SD_x = 0.35, \quad \bar{Y} = 5.30, \quad SD_y = 1.40$$

- ▶ Suppose we want to predict the value of Y for X=1.5

$$b = r_{xy} \left(\frac{SD_y}{SD_x} \right) = 0.43 \left(\frac{1.40}{0.35} \right) = 0.43(4.00) = 1.72$$

- ▶ The basic regression equation

$$\begin{aligned}\hat{Y} &= \bar{Y} + b(X - \bar{X}) \\ \hat{Y} &= 5.3 + 1.72(1.5 - 1.32) \\ \hat{Y} &= 5.3 + 1.72(0.18) \\ \hat{Y} &= 5.3 + 0.31 \\ \hat{Y} &= 5.62\end{aligned}$$

- ▶ The predicted value is slightly above the mean value of 5.30

Marty will spend 31 minutes more studying than the 5.30 hours the average kid and his mom was paying him a little more than the average kid was paid.

Example 2

How about Kindra, who doesn't get any money at all (i.e., $X = 0$) for the "A's" she earns?

Note: the rest of the values doesn't change

Example 2

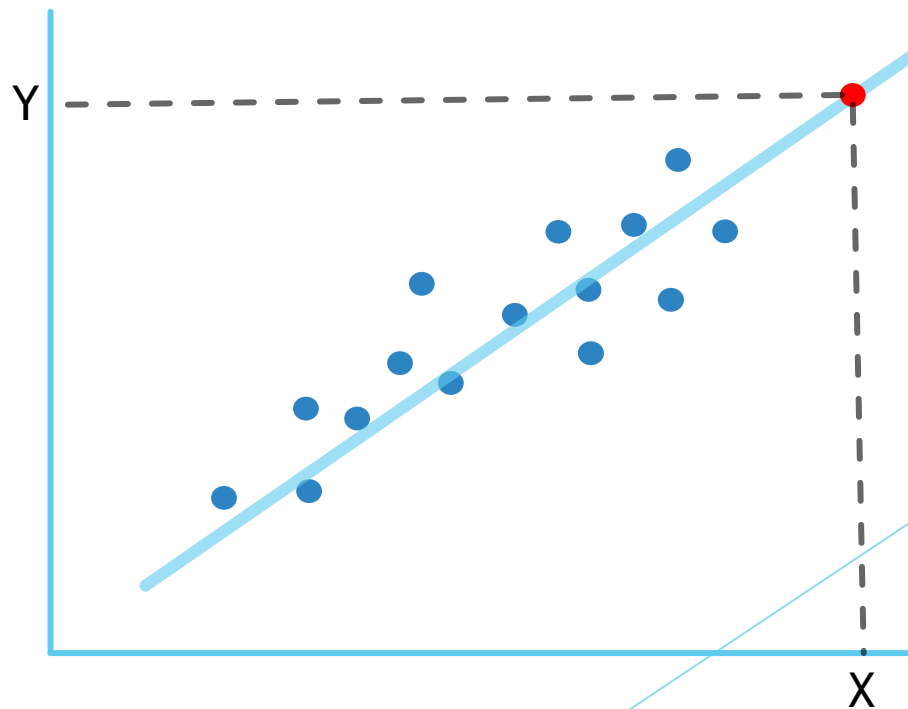
How about Kindra, who doesn't get any money at all (i.e., $X = 0$) for the "A's" she earns?

Note: the rest of the values doesn't change

$$\begin{aligned}\hat{Y} &= \bar{Y} + b(X - \bar{X}) \\ &= 5.30 + 1.72(0 - 1.32) \\ &= 5.30 + 1.72(-1.32) \\ &= 5.30 + (-2.27) \\ &= 3.03\end{aligned}$$

Graphical Interpretation

- ▶ Note that although we've made a prediction, we're not likely to be exactly right
 - ▶ We just want to be closer than chance
- ▶ We are effectively fitting a line through our correlated data and making a prediction along that line



Standard Error

- ▶ Informs how far our sample statistic deviates from the actual population mean
- ▶ Another example, ABC University uses an admissions test to predict the expected grade-point average (GPA) of incoming students. Over the years, it has been established that the mean score on the Test of Admissions (X) is 50.0 (SD = 10) and the mean Frosh GPA (Y) is 2.40 (SD = 0.60). The correlation between these two variables is +.70
- ▶ Linear regression analysis

$$\hat{Y} = \bar{Y} + b(X - \bar{X})$$

Denny, a twelfth grader, took the admissions test and got a score of 34. To predict his GPA using linear regression, we calculate prediction of GPA

$$\hat{Y} = ?$$

Standard Error

$$\hat{X}=50$$

$$SD_x=10$$

$$\hat{Y}=2.40$$

$$SD_y=0.60$$

$$R_{xy}=+0.70 \quad X=34$$

$$b = 0.70 \left(\frac{0.60}{10} \right)$$

$$= 0.042$$

$$\hat{Y} = 2.40 + 0.042(34 - 50)$$

$$= 2.40 - 0.672$$

$$\hat{Y} = 1.73$$

Standard Error

- ▶ The standard error is defined as

$$\sigma_{est} = \left(SD_y \sqrt{1 - r_{xy}^2} \right) \sqrt{\frac{N - 1}{N - 2}}$$

- ▶ Where ,

σ_{est} is the standard error of the estimate,

SD_y is the standard deviation of Y,

r_{xy}^2 is the square of the correlation coefficient,

N is the number of pairs of observations

- ▶ If N is large then the term in N is reduced to 1 and drops out of the equation
 - ▶ “Large” is normally defined as greater than or equal to 30 😊

Standard Error

- ▶ predicted value of $Y = 1.73$, what is the standard error?
- ▶ Assuming that we have a large sample

$$\sigma_{est} = \left(SD_y \sqrt{1 - r_{xy}^2} \right) \sqrt{\frac{N - 1}{N - 2}}$$

$$\sigma_{est} = \left(SD_y \sqrt{1 - r_{xy}^2} \right)$$

$$\sigma_{est} = \left(0.60 \sqrt{1 - 0.70^2} \right)$$

$$\sigma_{est} = \left(0.60 \sqrt{1 - 0.49} \right)$$

$$\sigma_{est} = 0.60(0.71)$$

$$\sigma_{est} = 0.43$$

- ▶ Note: $\sqrt{\frac{N-1}{N-2}}$ is reduced to 1 and drops out of the equation because the sample is so large
- ▶ The predicted number of GPA's is therefore written as 1.73 ± 0.43

Problem

- ▶ Suppose the average number of rats in a set of dockside warehouses is 57 per warehouse with an SD of 12. The average number of packages of rat poison purchased per warehouse per year is 4.2. with an SD of 1.1. The correlation between rats in warehouses and packages of rat poison purchased is 0.5
 - ▶ Kleensite bought 12 packages of rat poison last year. How many rats would you predict are in their warehouse?
 - ▶ Remember to calculate the standard error as well as the predicted value
- ▶ Hint: First work out what is Y , the thing you want to calculate. Then identify $\bar{X}$, SD_x , $\bar{Y}$, SD_y from the wording of the question

Problem

- ▶ Suppose the average number of rats in a set of dockside warehouses is 57 per warehouse with an SD of 12. The average number of packages of rat poison purchased per warehouse per year is 4.2, with an SD of 1.1. The correlation between rats in warehouses and packages of rat poison purchased is 0.5
 - ▶ Kleensite bought 12 packages of rat poison last year. How many rats would you predict are in their warehouse?
- ▶ Y is the number of rats in the warehouse
- ▶ X is the number of packages of rat poison bought = 12
- ▶ The correlation coefficient r_{xy} is 0.5
- ▶ $\bar{X} = 4.2$, $SD_x = 1.1$
- ▶ $\bar{Y} = 57$, $SD_y = 12$

Problem

- ▶ X and Y have a correlation of 0.50

$$\bar{X} = 4.2, \quad SD_x = 1.1, \quad \bar{Y} = 57 \quad SD_y = 12$$

- ▶ We want to predict the value of Y for X=12

$$b = r_{xy} \left(\frac{SD_y}{SD_x} \right) = 0.50 \left(\frac{12}{1.1} \right) = 0.50(10.9) = 5.45$$

- ▶ We can now plug this into the basic regression equation

$$\begin{aligned}\hat{Y} &= \bar{Y} + b(X - \bar{X}) \\ \hat{Y} &= 57 + 5.45(12 - 4.2) \\ \hat{Y} &= 57 + 5.45(7.80) \\ \hat{Y} &= 57 + 42.51 \\ \hat{Y} &= 99.51\end{aligned}$$

Problem

- ▶ X and Y have a correlation of 0.50

$$\bar{X} = 4.2, \quad SD_x = 1.1, \quad \bar{Y} = 57 \quad SD_y = 12$$

- ▶ We have predicted the value of Y for X=12 is 99.51 so what is the standard error for this?
- ▶ Assuming that we have a large sample

$$\sigma_{est} = \left(SD_y \sqrt{1 - r_{xy}^2} \right) \sqrt{\frac{N - 1}{N - 2}} = \sigma_{est} = \left(SD_y \sqrt{1 - r_{xy}^2} \right)$$

$$\sigma_{est} = \left(12 \sqrt{1 - 0.50^2} \right)$$

$$\sigma_{est} = \left(12 \sqrt{1 - 0.25} \right)$$

$$\sigma_{est} = 12(0.87)$$

$$\sigma_{est} = 10.44$$

- ▶ The predicted number of rats is therefore written as 99.51 ± 10.44

Summary

- ▶ Topics covered
 - ▶ Regression Equations
 - ▶ Standard Error of the Estimate
- ▶ Next week
 - ▶ The Sampling Experiment
 - ▶ Sample Values (statistics) and Population Values (Parameters)
 - ▶ The Null Hypothesis
 - ▶ Type I and Type II Errors
 - ▶ Statistical Significance and Type I Errors
- ▶ Inferential statistics allow us to make predictions based upon the sampling of a large population

References

- ▶ chapter 9 of *Statistics for the Terrified*.