**Introduction to regression**

We will briefly examine regression in this section. If you are interested in learning more about the topic, there are whole courses available on it!

Suppose we are interested in estimating the average GPA of all students at this university. How would we do this? Assume we do not have access to any student records.

* 1. Define the random variable: let Y denote student GPA
	2. Define the population: all students at this university
	3. Define the parameter that we are interested in: μ = population mean GPA
	4. Take a representative sample from the population: suppose a random sample of 100 students is selected
	5. Calculate the statistic that estimates the parameter:  = sample mean GPA
	6. Make an inference about the value of the parameter using the statistical science: construct confidence intervals or perform hypothesis tests using the sample mean and sample standard deviation

The diagram below demonstrates these steps. Note that not all GPAs could be shown in the diagram.



What factors may be related to GPA?

1. High school (HS) GPA
2. ACT score
3. Involvement in activities
4. And many more…

Suppose we are interested in the relationship between college and HS GPA, and we want to use HS GPA to predict college GPA. How could we do this? Assume we do not have access to any student records.

Use similar steps as on page 1, but now with regression models.



Data shown as: (HS GPA, College GPA)

Example: HS and College GPA (gpa.R, gpa.txt)

A random sample of 20 students from our university is taken producing the data set below (data is different from above).

> gpa <- read.csv(file = "gpa.csv")

> gpa

 HS.GPA College.GPA

1 3.04 3.10

2 2.35 2.30

3 2.70 3.00

4 2.55 2.45

5 2.83 2.50

6 4.32 3.70

7 3.39 3.40

8 2.32 2.60

9 2.69 2.80

10 2.83 3.60

11 2.39 2.00

12 3.65 2.90

13 2.85 3.30

14 3.83 3.20

15 2.22 2.80

16 1.98 2.40

17 2.88 2.60

18 4.00 3.80

19 2.28 2.20

20 2.88 2.60

Plot the observation pairs in a scatter plot:



Regression allows us to develop an equation, like  = 1.09 + 0.61×(HS GPA), to predict College GPA from HS GPA:



Notice that the regression model does not perfectly predict the college GPAs. There is some error in the prediction. This error can be quantified through the use of probability distributions!

Goal of this section:

Develop a model (equation) that numerically describes the relationship between two variables using **simple linear regression**.

### Algebra Review

|  |  |
| --- | --- |
| x | y |
| -1/2 | 0 |
| 0 | 1 |
| 1 | 3 |

y = dependent variable

x = independent variable

b = y-intercept

m= slope of line; measures how fast (or slow) that y changes as x changes by a one-unit increase

The simple linear regression model

Suppose you are interested in studying the relationship between two variables x and Y (x may be HS GPA and Y may be college GPA)



where

Y = dependent (response) random variable value

y = observed value of Y

x = independent (explanatory) variable value (this is

 assumed to be a fixed constant here)

Y = β0 + β1x + ε is the population regression model

* ε = random variable (random error term) that has a normal probability distribution mean 0 and variance 
* E(Y) = β0 + β1x is what Y is expected to be on average for a specific value of x because

 E(Y) = E(β0 + β1x + ε)

= E(β0 + β1x) + E(ε)

= β0 + β1x + 0

= β0 + β1x

* β0 = y-intercept for the population regression model
* β1 = slope for the population regression model
* β0 and β1 are parameters that need to be estimated

 is the sample regression model (estimated regression model, equation, or line and fitted line)

*  estimates E(Y) = β0 + β1x
* is the estimated value of β0; y-intercept for the sample regression model
*  is the estimated value of β1; slope for the sample regression model
*  and  are statistics
	+ They can be thought of as random variables and observed values calculated from the sample.
	+ Notationally, it is rare for someone to write these quantities in a capitalized form when they are random variables.
*  is the estimated or predicted value of E(Y); based on the sample, we can calculate .

Note that x is a constant value – not a random variable. In settings like the GPA example, it makes sense for HS GPA to be a random variable. Even if it is, the estimators derived and inferences made in this chapter would remain the same.

Below are two nice diagrams showing what is being done here.

Population:



Population and sample:

