Regression Analysis

- 1. Simple Linear Regression
- 2. Inference in Regression Analysis
- 3. Diagnostics
- 4. Simultaneous Inference
- 5. Matrix Algebra
- 6. Multiple Linear Regression
- 7. Extra Sums of Squares
- 8.-10. Building the Regression Model
- 11 Qualitative Predictor Variables

1. Simple Linear Regression

Suppose that we are interested in the average height of male undergrads at UF. We put each guy's name (**population**) in a hat and randomly select 100 (**sample**). Here they are: $Y_1, Y_2, \ldots, Y_{100}$.

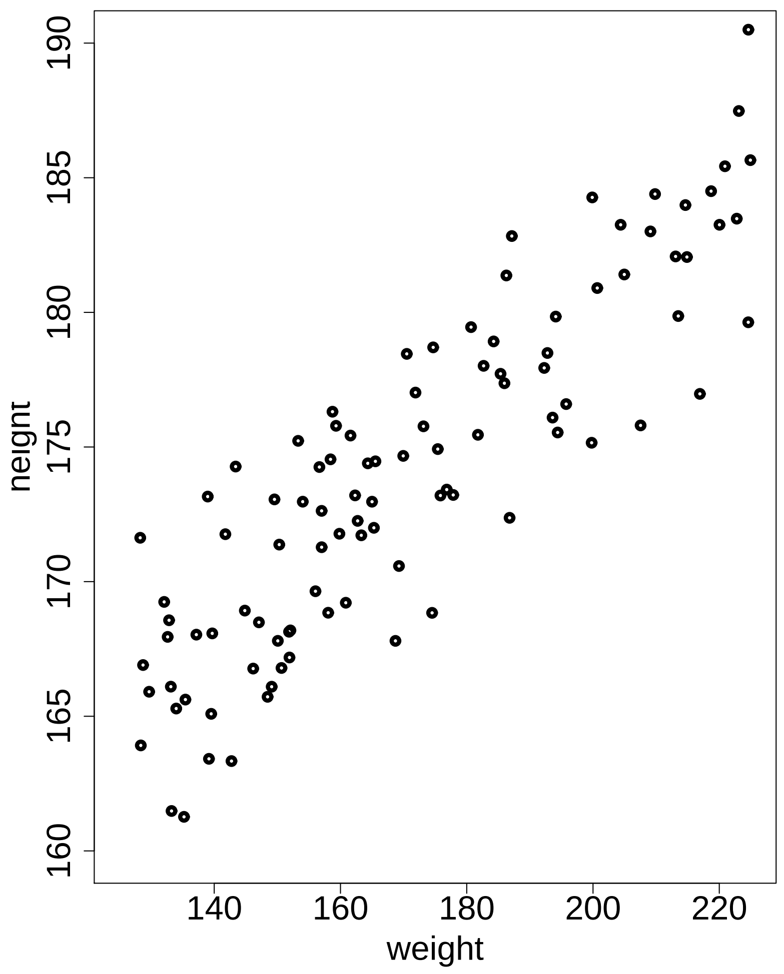
Suppose, in addition, we also measure their weights and the number of cats owned by their parents. Here they are: $W_1, W_2, \ldots, W_{100}$ and $C_1, C_2, \ldots, C_{100}$.

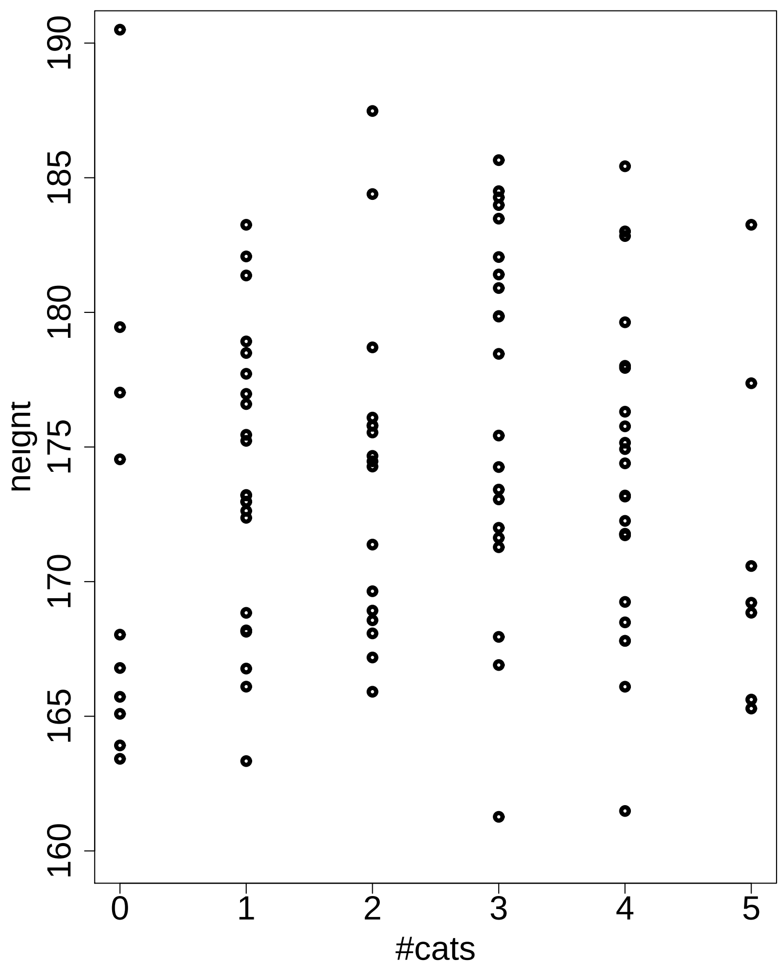
Questions:

1. How would you use this data to estimate the average height of a male undergrad?

2. male undergrads who weigh between 200-210?

3. male undergrads whose parents own 3 cats?


height
weight
160
165
170
175
180
185
190
140
160
180
200
220


height
#cats
160
165
170
175
180
185
190
0
1
2
3
4
5

Answers:

1. $\bar{Y} = \frac{1}{100}\sum_{i=1}^{100} Y_i$, the sample mean.

2. average the Y_i's for guys whose X_is are between 200-210.

3. average the Y_i's for guys whose C_is are 3? **No!**
   Same as in 1., because height certainly do not depend on the number of cats.

Intuitive description of regression:
(height) Y = variable of interest = response variable = dependent variable
(weight) X = explanatory variable = predictor variable = independent variable

Fundamental assumption of regression

1. For each particular value of the predictor variable X, the response variable Y is a random variable whose mean (expected value) depends on X.

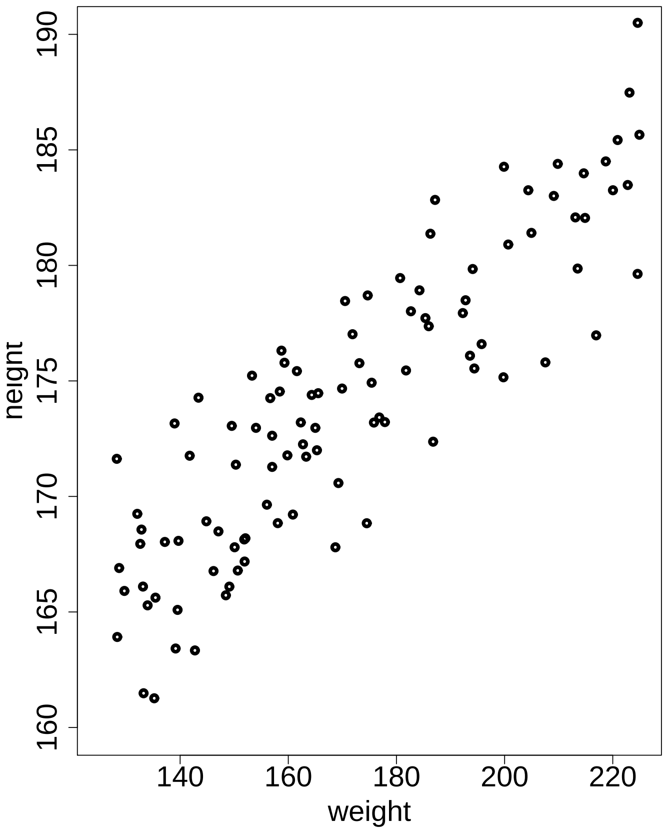
2. The mean value of Y, $\mathsf{E}(Y)$, can be written as a deterministic function of X.

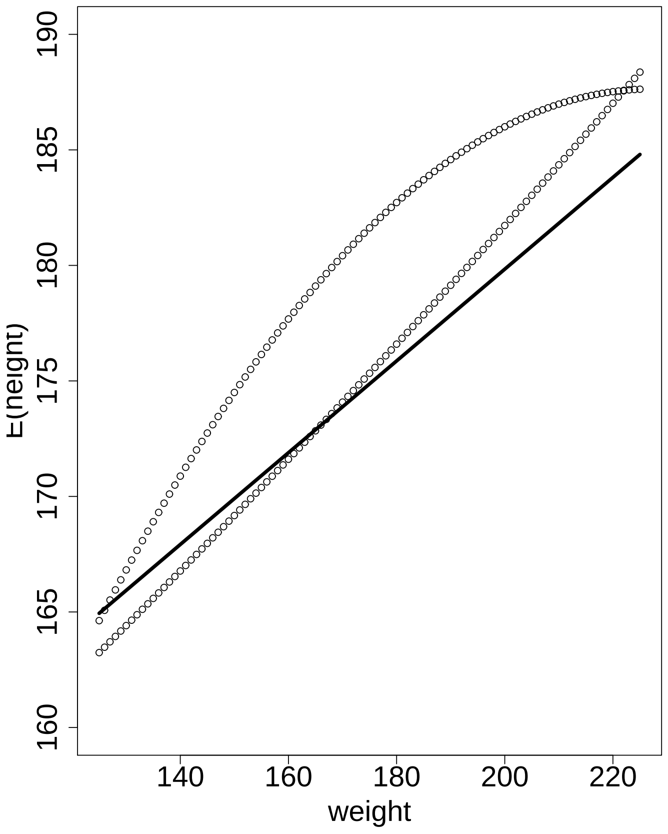
Example: $\mathsf{E}(height_i) = f(weight_i)$

$$\mathsf{E}(height_i) = \begin{cases} \beta_0 + \beta_1(weight_i) \\ \beta_0 + \beta_1(weight_i) + \beta_2(weight_i^2) \\ \beta_0 \exp[\beta_1(weight_i)], \end{cases}$$

where β_0, β_1, and β_2 are **unknown parameters!**

Scatterplot $weight$ versus $height$ and $weight$ versus $\mathrm{E}(height)$:

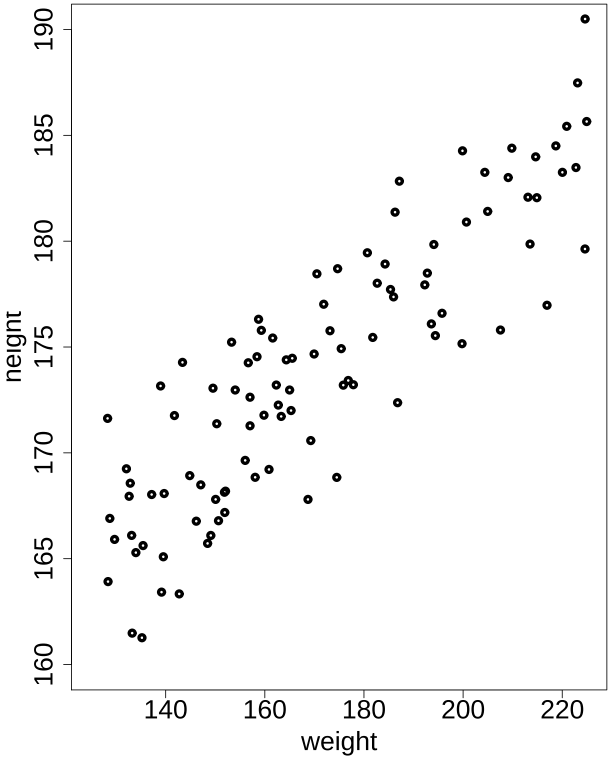


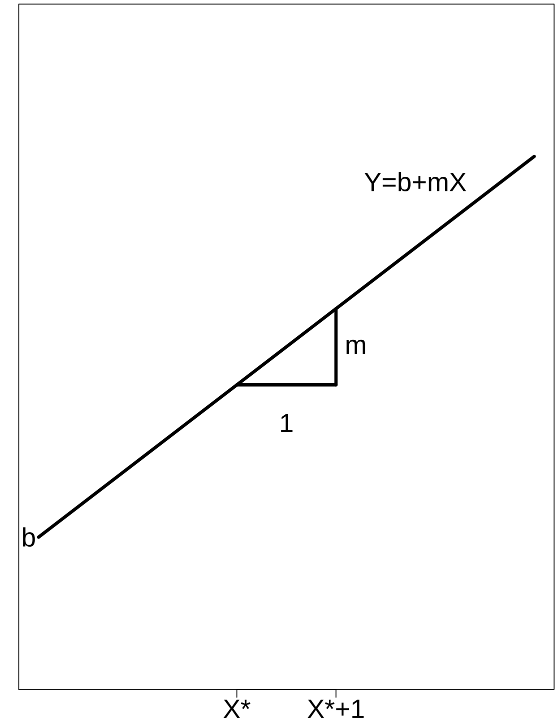


Simple Linear Regression (SLR)

A scatterplot of 100 (X_i, Y_i) pairs $(weight, height)$ shows that there is a **linear trend**.

Equation of a line: $Y = b + m \cdot X$ (**slope and intercept**)





At X^*: $Y = b + mX^*$

At $X^* + 1$: $Y = b + m(X^* + 1)$

Difference is: $(b + m(X^* + 1)) - (b + mX^*) = m$

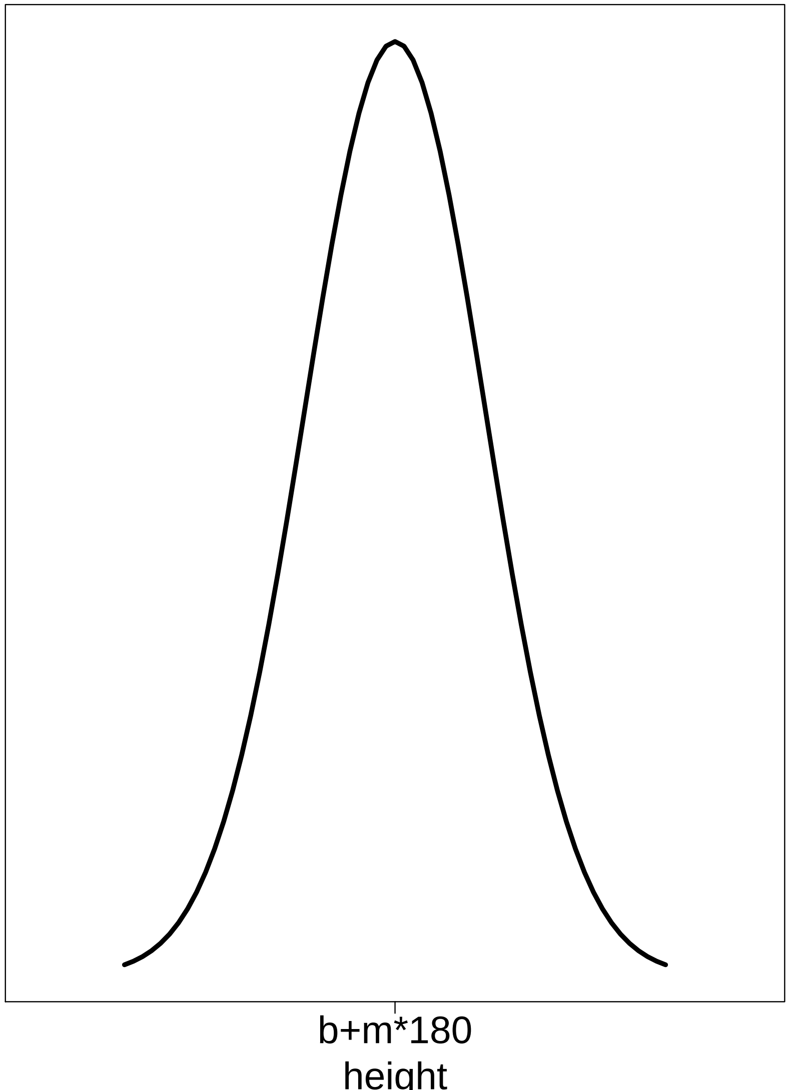
Is: $height = b + m \cdot weight$? (**functional relation**)

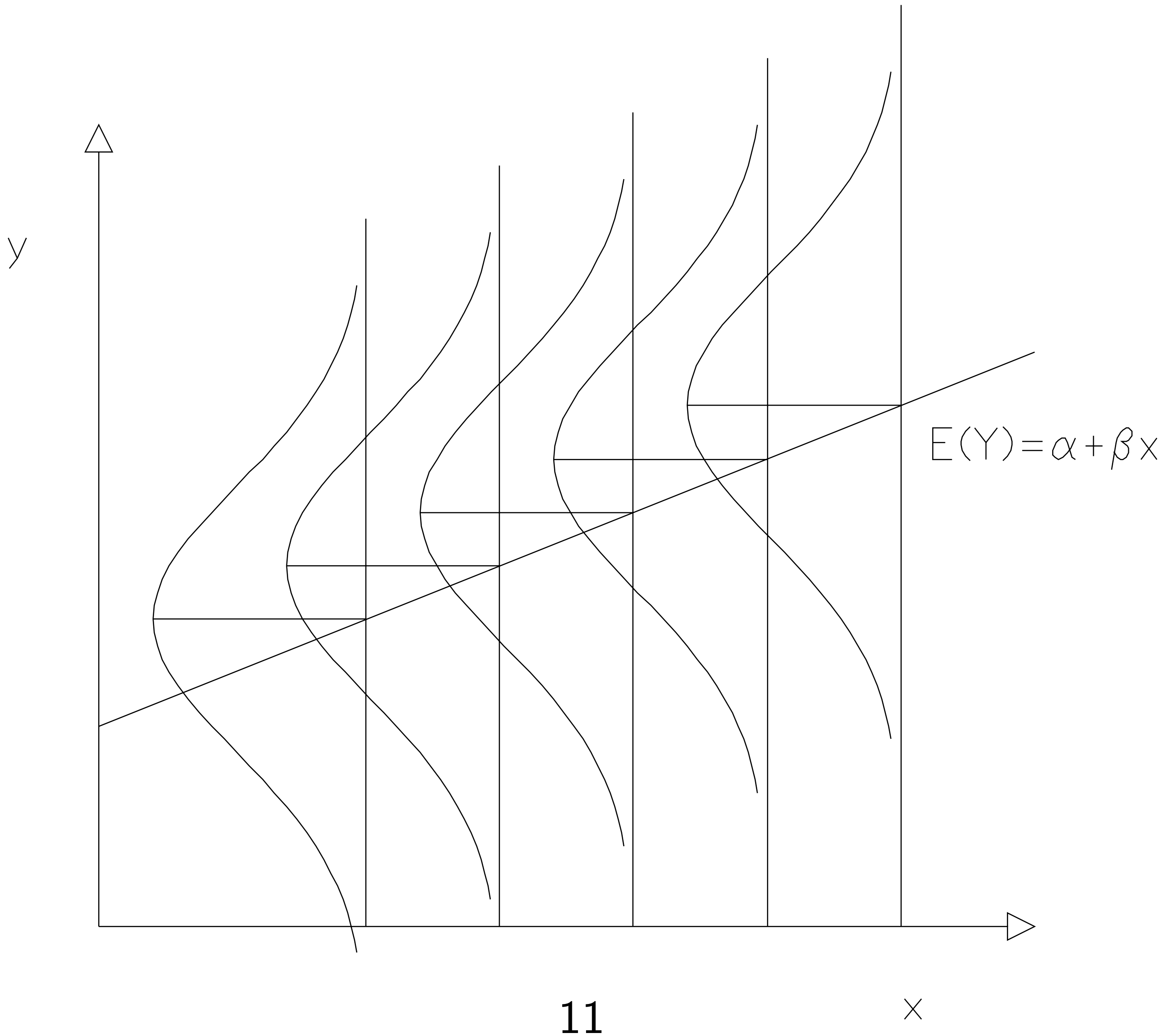
No! The relationship is far from perfect (**it's a statistical relation**)!

We can say that: $\mathsf{E}(height) = b + m \cdot weight$

That is, **height is a random variable, whose expected value is a linear function of weight**.

Distribution of height for a person who is 180lbs, i.e. Mean $\mathsf{E}(height) = b + m \cdot 180$.


b+m*180
height


y
E(Y)=α+βx
x

Formal Statement of the SLR Model

Data: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

Equation:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

Assumptions:

- Y_i is the value of the **response variable** in the ith trial
- X_i's are **fixed known constants**
- ϵ_i's are uncorrelated and identically distributed **random errors** with $\mathsf{E}(\epsilon_i) = 0$ and $\mathsf{var}(\epsilon_i) = \sigma^2$.
- β_0, β_1, and σ^2 are **unknown parameters** (constants).

Consequences of the SLR Model

- The response Y_i is the sum of the constant term $\beta_0 + \beta_1 X_i$ and the random term ϵ_i. Hence, Y_i is a random variable.

- The ϵ_i's are uncorrelated and since each Y_i involves only one ϵ_i, the Y_i's are uncorrelated as well.

- $\mathsf{E}(Y_i) = \mathsf{E}(\beta_0 + \beta_1 X_i + \epsilon_i) = \beta_0 + \beta_1 X_i$.
  Regression function (it relates the mean of Y to X) is

$$\mathsf{E}(Y) = \beta_0 + \beta_1 X.$$

- $\mathsf{var}(Y_i) = \mathsf{var}(\beta_0 + \beta_1 X_i + \epsilon_i) = \mathsf{var}(\epsilon_i) = \sigma^2$.
  Thus $\mathsf{var}(Y_i) = \sigma^2$ (same constant variance for all Y_i's).

Why is it called SLR?

$Simple$: only one predictor X_i

$Linear$: regression function, $\mathsf{E}(Y) = \beta_0 + \beta_1 X$, is linear in the parameters.

Why do we *care about* the regression model?

If the model is realistic and we have reasonable estimates of β_0 and β_1 we have:

1. The ability to predict new Y_i's given a new X_i

2. An understanding of how the mean of Y_i, $\mathsf{E}(Y_i)$, changes with X_i

Repetition – The Summation Operator:

Fact 1: If $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ then

$$\sum_{i=1}^{n}(X_i - \bar{X}) \quad = \quad 0$$

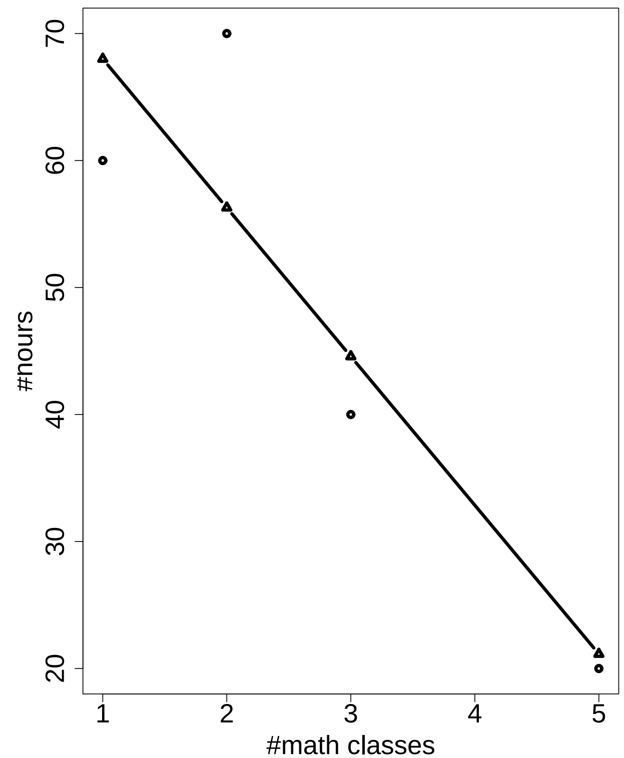
Fact 2:

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \bar{X})X_i = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

Least Squares Estimation of regression parameters β_0 and β_1

$X_i = \#$math classes taken by ith student in spring
$Y_i = \#$hours student i spends writting papers in spring

Randomly select 4 students
$(X_1, Y_1) = (1, 60)$, $(X_2, Y_2) = (2, 70)$,
$(X_3, Y_3) = (3, 40)$, $(X_4, Y_4) = (5, 20)$



If we assume a SLR model for these data, we are assuming that at each X, there is a distribution of $\#$hours and that the means (expected values) of these responses all lie on a line.

We need **estimates of the unknown parameters** β_0, β_1, and σ^2. Let's focus on β_0 and β_1 for now.

Every (β_0, β_1) pair defines a line $\beta_0 + \beta_1 X$. The **Least Squares Criterion** says choose the line that **minimizes** the sum of the squared vertical distances from the data points (X_i, Y_i) to the line $(X_i, \beta_0 + \beta_1 X_i)$.

Formally, the least squares estimators of β_0 and β_1, call them b_0 and b_1, minimize

$$Q = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

which is the sum of the squared vertical distances from the points to the line.

Instead of evaluating Q for every possible line $\beta_0 + \beta_1 X$, we can find the best β_0 and β_1 using calculus. We will minimize the function Q with respect to β_0 and β_1

$$
\begin{aligned}
\frac{\partial Q}{\partial \beta_0} &= \sum_{i=1}^{n} 2(Y_i - (\beta_0 + \beta_1 X_i))(-1) \\
\frac{\partial Q}{\partial \beta_1} &= \sum_{i=1}^{n} 2(Y_i - (\beta_0 + \beta_1 X_i))(-X_i)
\end{aligned}
$$

Set it to 0 (and change notation) yields the **normal equations (very important)!**

$$
\begin{aligned}
\sum_{i=1}^{n}(Y_i - (b_0 + b_1 X_i)) &= 0 \\
\sum_{i=1}^{n}(Y_i - (b_0 + b_1 X_i))X_i &= 0
\end{aligned}
$$

Solving these equations simultaneously yields

$$
\begin{aligned}
b_1 &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
b_0 &= \bar{Y} - b_1 \bar{X}
\end{aligned}
$$

This result is **even more important!** Use second derivative to show that a minimum is attained.

A more efficient formula for the calculation of b_1 is

$$
\begin{aligned}
b_1 &= \frac{\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}(\sum_{i=1}^{n} X_i)^2} \\
&= \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{S_{XX}}
\end{aligned}
$$

where $S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2$.

Example:

Let us calculate the estimates of slope and intercept of our example:

$\sum_i X_i Y_i = 60 + 140 + 120 + 100 = 420$
$\sum_i X_i = 11$, $\sum_i Y_i = 190$, $\sum_i X_i^2 = 39$

$$b_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{1}{n}(\sum_{i=1}^n X_i)(\sum_{i=1}^n Y_i)}{\sum_{i=1}^n X_i^2 - \frac{1}{n}(\sum_{i=1}^n X_i)^2}$$

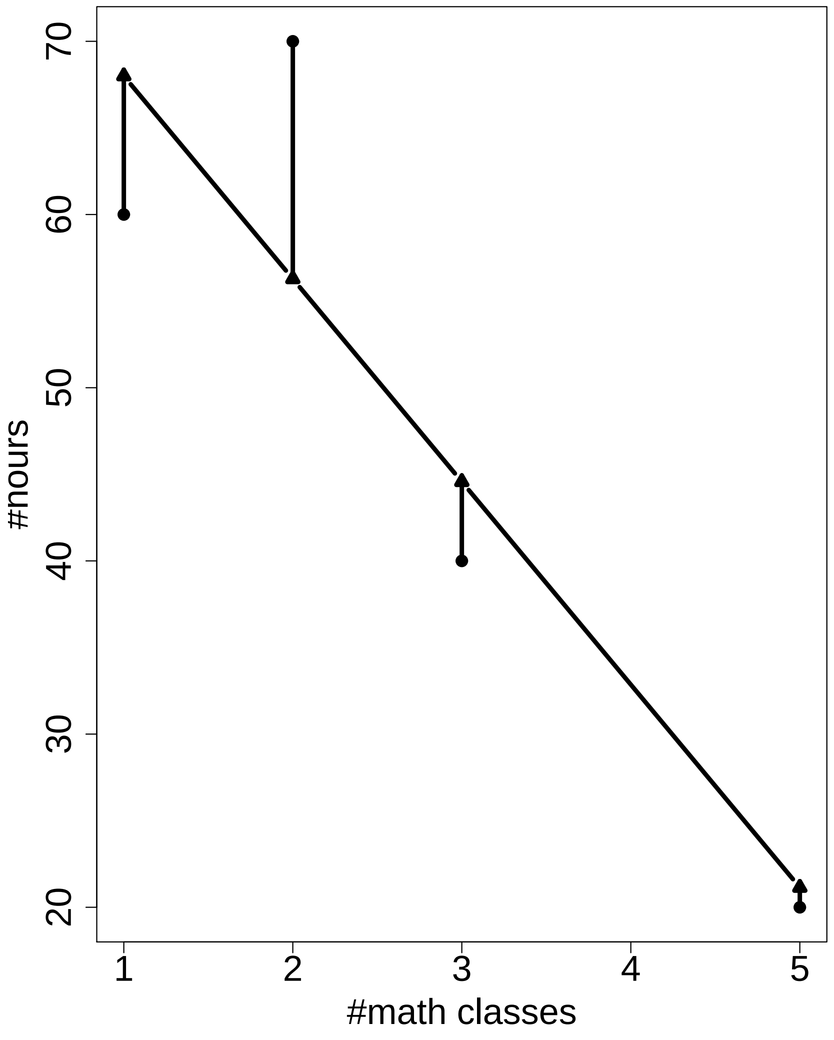
$$= \frac{420 - \frac{1}{4}(11)(190)}{39 - \frac{1}{4}(11)^2} = \frac{-102.5}{8.75} = -11.7$$

$$b_0 = \bar{Y} - b_1 \bar{X} = \frac{1}{4}190 - (-11.7)(\frac{1}{4}11) = 80.0$$

Estimated regression function

$$\widehat{\mathsf{E}(Y)} = 80 - 11.7X$$

At $X = 1$: $\widehat{\mathsf{E}(Y)} = 80 - 11.7(1) = 68.3$
At $X = 5$: $\widehat{\mathsf{E}(Y)} = 80 - 11.7(5) = 21.5$


70
60
50
40
30
20
#hours
1
2
3
4
5
#math classes

Properties of Least Squares Estimators

An important theorem, called the *Gauss Markov Theorem*, states that the Least Squares Estimators are **unbiased** and have **minimum variance** among all unbiased linear estimators.

Point Estimation of the Mean Response:
Under the SLR model, the regression function is

$$\mathsf{E}(Y) = \beta_0 + \beta_1 X.$$

We use our estimates of β_0 and β_1 to construct the **estimated regression function**

$$\widehat{\mathsf{E}(Y)} = b_0 + b_1 X$$

Fitted Values: Define

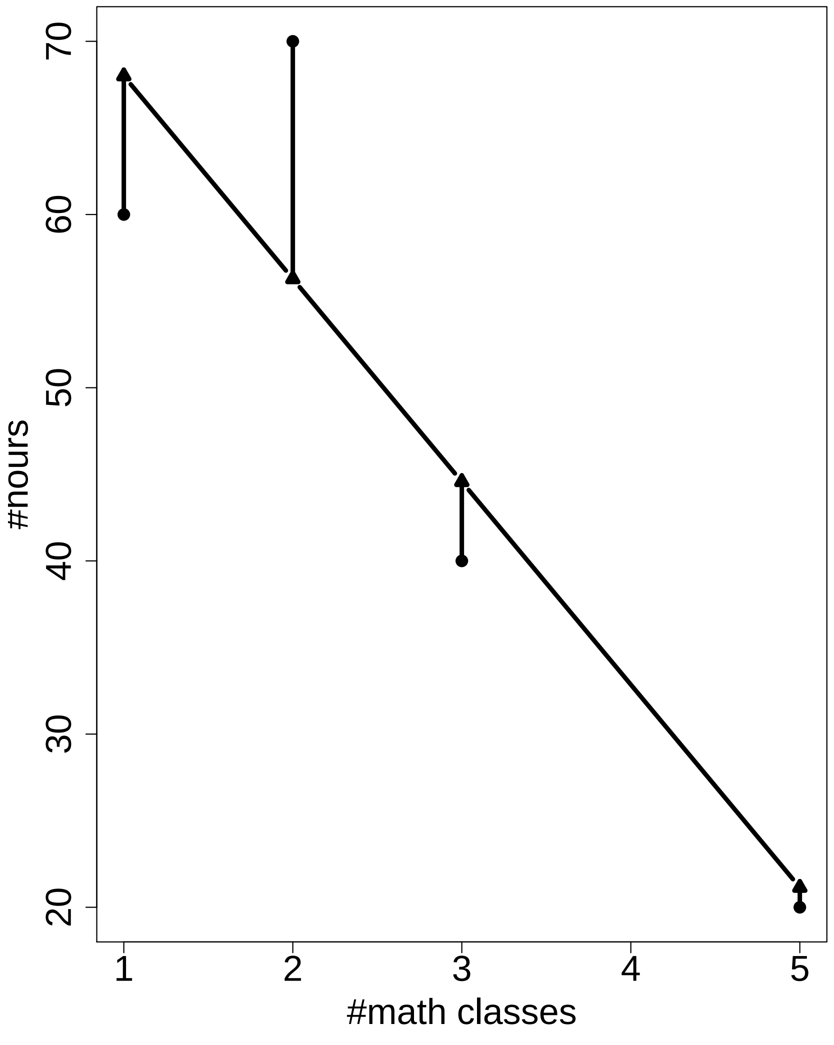
$$\hat{Y}_i = b_0 + b_1 X_i, \quad i = 1, 2, \ldots, n$$

$\hat{Y}_i$ is the fitted value at X_i.

Residuals: Define

$$e_i = Y_i - \hat{Y}_i, \quad i = 1, 2, \ldots, n$$

e_i is called ith residual. The vertical distance between the ith Y value and the line.



70
60
50
40
30
20
#hours
1
2
3
4
5
#math classes

Properties of Fitted Regression Line

- The sum of the residuals is zero:

$$\sum_{i=1}^{n} e_i = 0.$$

- The sum of the squared residuals, $\sum_{i=1}^{n} e_i^2$, is a minimum.

- The sum of the observed values equals the sum of the fitted values:

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i.$$

- The sum of the residuals weighted by X_i is zero:

$$\sum_{i=1}^{n} X_i e_i = 0.$$

- The sum of the residuals weighted by $\hat{Y}_i$ is zero:

$$\sum_{i=1}^{n} \hat{Y}_i e_i = 0.$$

- The regression line always goes through the point $(\bar{X}, \bar{Y})$.

Errors versus Residuals

$$
\begin{aligned}
e_i &= Y_i - \hat{Y}_i \\
&= Y_i - b_0 - b_1 X_i
\end{aligned}
$$

$$
\epsilon_i = Y_i - \beta_0 - \beta_1 X_i
$$

So e_i is like $\hat{\epsilon}_i$, but ϵ_i is **not** a parameter!

Estimation of σ^2 in SLR:

Motivation from iid (independent & identically distributed) case, where $Y_1, \ldots, Y_n$ iid with $\mathsf{E}(Y_i) = \mu$ and $\text{var}(Y_i) = \sigma^2$.

Sample variance (two steps)

1. find
$$\sum_{i=1}^{n} (Y_i - \widehat{\mathsf{E}(Y_i)})^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$
   Square the difference between each observation and the estimate of its mean.

2. divide by degrees of freedom
$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

   Lost 1 degree of freedom, because we estimated 1 parameter, μ.

SLR model with $\mathsf{E}(Y_i) = \beta_0 + \beta_1 X_i$ and $\mathsf{var}(Y_i) = \sigma^2$, independent but not identically distributed.

Let's do the same two steps.

1. find
$$\sum_{i=1}^{n}(Y_i - \widehat{\mathsf{E}(Y_i)})^2 = \sum_{i=1}^{n}(Y_i - (b_0 + b_1 X_i))^2 = \text{SSE}.$$
Square the difference between each observation and the estimate of its mean.

2. divide by degrees of freedom
$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - (b_0 + b_1 X_i))^2 = \text{MSE}.$$
Lost 2 degree of freedom, because we estimated 2 parameters, β_0 and β_1.

SSE: *error (residual) sum of squares*; MSE: *error (residual) mean square*

Properties of the point estimator of σ^2:

$$
\begin{aligned}
s^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - (b_0 + b_1 X_i))^2 \\
&= \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\
&= \frac{1}{n-2}\sum_{i=1}^{n} e_i^2
\end{aligned}
$$

MSE is an **unbiased estimate** of σ^2, that is

$$\mathsf{E}(\mathsf{MSE}) = \sigma^2.$$

Normal Error Regression Model

No matter what may be the form of the distribution of the error terms ϵ_i the **least squares** method provides **unbiased** point estimators of β_0 and β_1 that have **minimum variance** among all unbiased linear estimators.

To set up interval estimates and make tests, however, we need to make assumptions about the distribution of the ϵ_i.

The **normal error regression model** is as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

Assumptions:

- Y_i is the value of the **response variable** in the ith trial
- X_i's are **fixed known constants**
- ϵ_i's are independent $N(0, \sigma^2)$ **random errors**.
- β_0, β_1, and σ^2 are **unknown parameters** (constants).

This implies, that the responses are independent random variates with

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2).$$

Motivate Inference in SLR Models

Let X_i = #siblings and Y_i = #hours spent on papers. Data $(1, 20), (2, 50), (3, 30), (5, 30)$ gives

$$\widehat{\mathsf{E}(Y)} = 33 + 0.3X$$

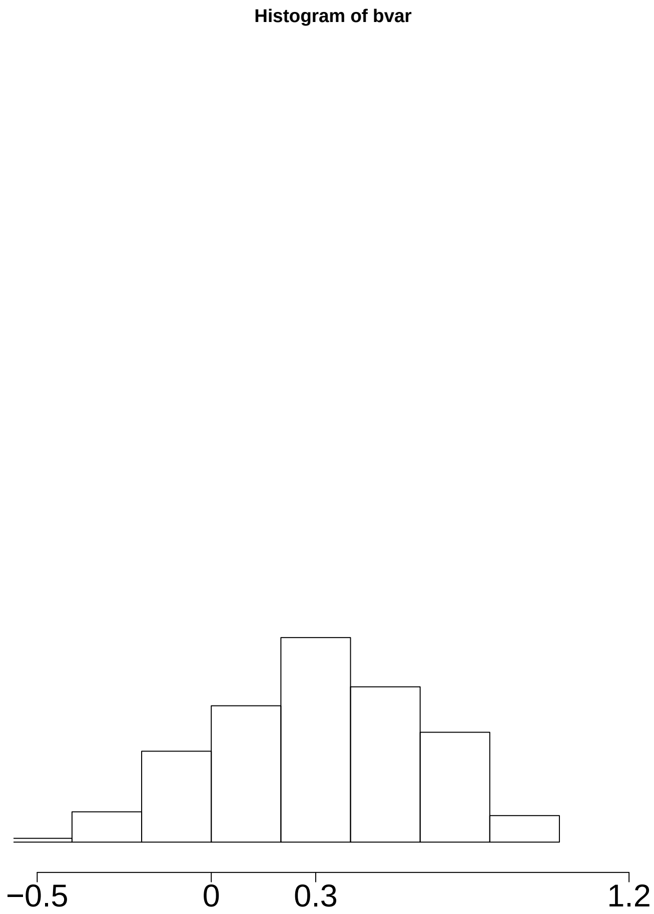
Conclusion: b_1 is not zero, so
#siblings is linearly related to #hours,right?

WRONG!

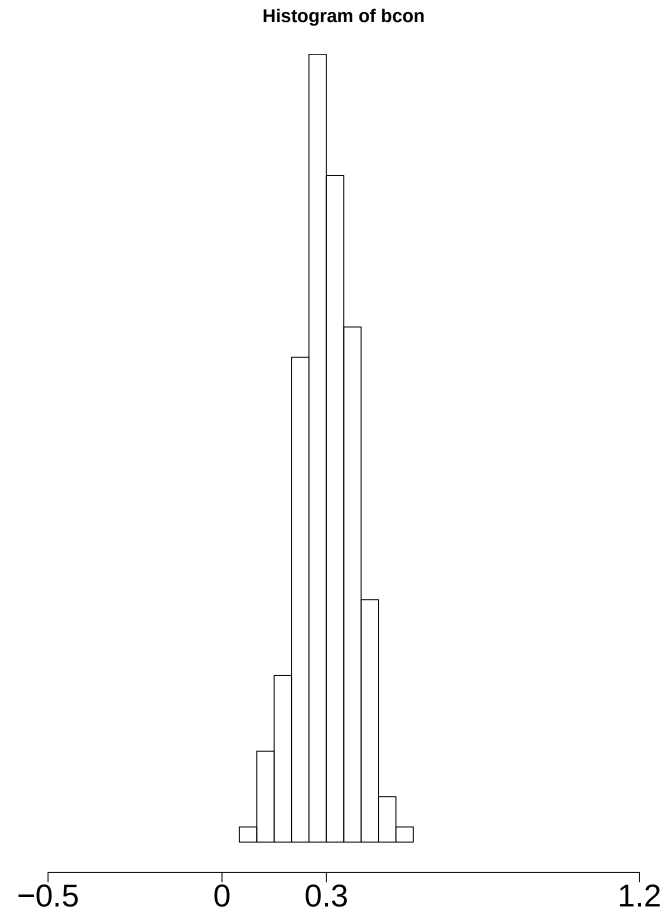
b_1 is a random variable because it depends on the Y_i's.

Think of consecutively collecting data and recalculating b_1 for each data. We draw the histogram of these b_1's

Scenario 1: Highly variable



Scenario 2: Highly concentrated



Think about $H_0 : \beta_1 = 0$
Is H_0 false? Scenario 1: not sure
Scenario 2: definitely

If we know the exact dist'n of b_1, we can formally decide if H_0 is true. We need formal statistical test of
$H_0 : \beta_1 = 0$ (not)
$H_A : \beta_1 \neq 0$ (there is a linear relationship between E(Y) and X)