ORIE 476
Applied Linear Statistical Models
http://www.stat.ufl.edu/~jbooth

- 1. Simple Linear Regression
- 2. Inference in Regression Analysis
- 3. Diagnostics
- 5. Matrix Approach to SLR Analysis
- 6. Multiple Regression
- 16. Single-Factor ANOVA
- 19. Two-factor ANOVA: Equal Sample Size
- 22. Two-factor ANOVA: Unequal Sample Size
- 25. Analysis of Covariance

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**).

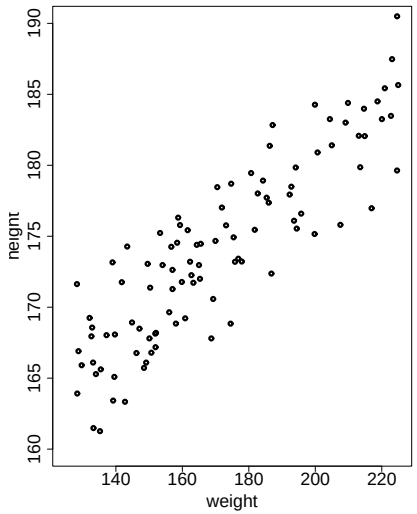
Here they are: $Y_1, Y_2, \ldots, Y_{100}$.

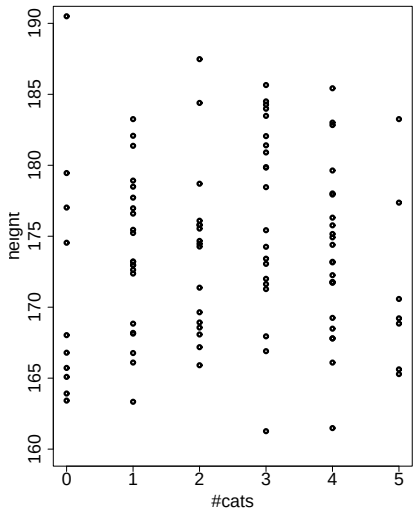
Suppose, in addition, we also measure their weights and the number of cats owned by their parents.

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?





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 does 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

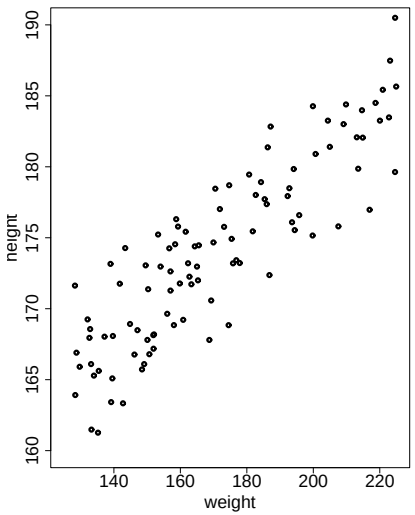
1. For each particular value of the predictor variable X, the response variable Y is a random variable whose mean (expected value) depends on X.
2. The expected value of Y, $E(Y)$, can be written as a deterministic function of X.

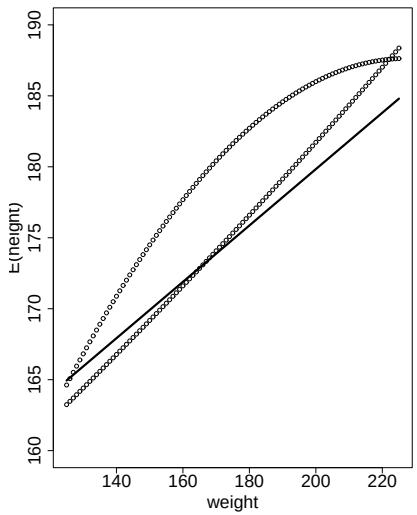
Example: $E(height_i) = f(weight_i)$

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

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

Scatterplot *weight* versus *height* and
weight versus $E(height)$:

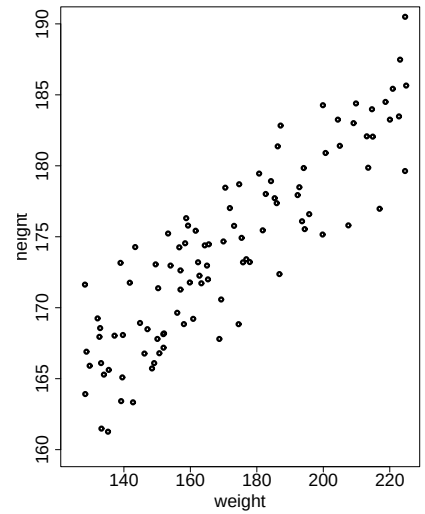


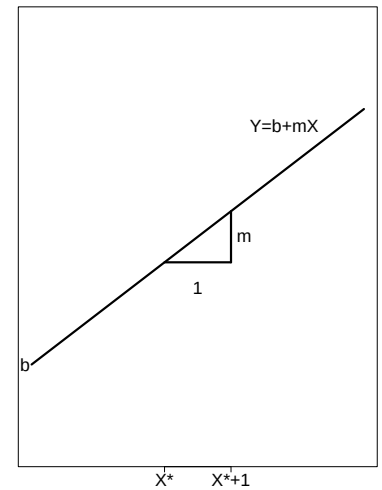


Simple Linear Regression (SLR)

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

Equation of a line: $Y = b + mX$
(b =intercept and m =slope)





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

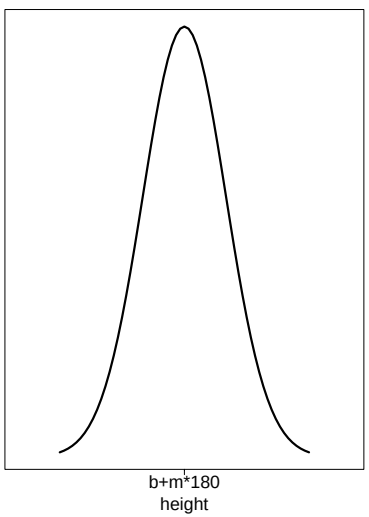
Is: $height = b + m * weight$? (**functional relation**)

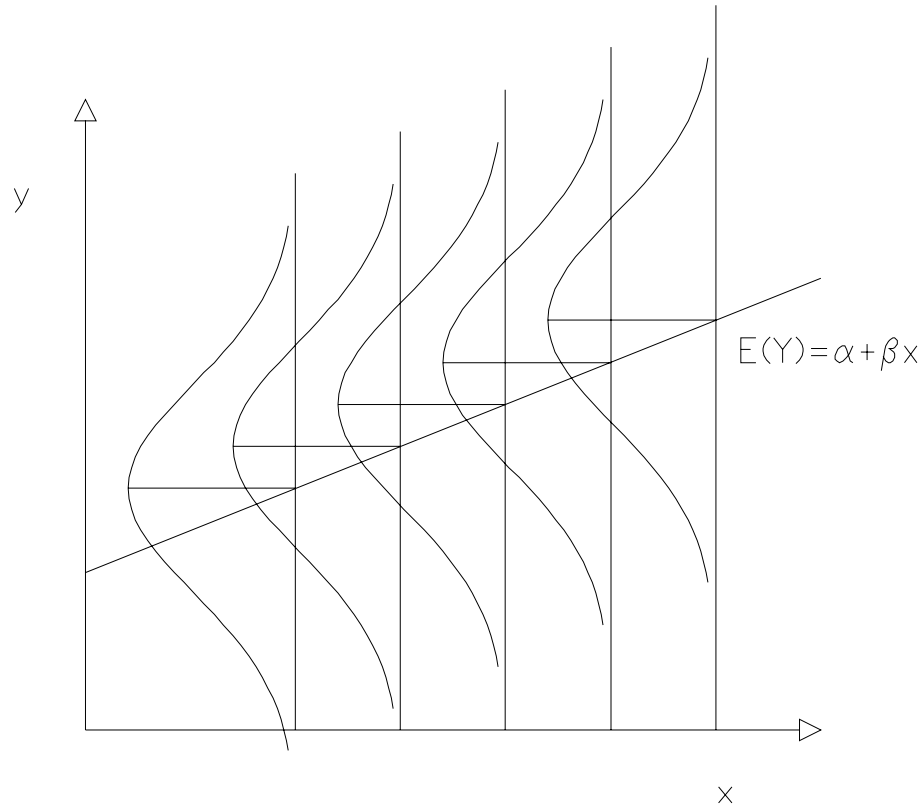
No! The relationship is far from perfect (**it's a statistical relation**)!

We can say that: $E(height) = b + m * weight$

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

Distribution of height for male students whose weight is 180lbs:





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 values of a **predictor variable** - may be fixed by the experimenter (**designed study**), or randomly sampled (**observational study**).
- ϵ_i's are uncorrelated and identically distributed **random errors** with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$.
- β_0, β_1, and σ^2 are fixed but **unknown parameters**

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

- Strictly speaking the model specifies the **conditional distribution** of Y given X.
- The **Regression function** (which relates the mean of Y to X) is

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

- Constant (homogeneous) variance across values of X.

$$Var(Y_i|X_i) = Var(\epsilon_i) = \sigma^2$$

- The ϵ_i's are uncorrelated and since each Y_i involves only one ϵ_i, the Y_i's are uncorrelated as well.
- The conditioning notation "$|X$" is suppressed in what follows.

Consequences of the SLR Model

Why is it called *SLR*?

Simple: only one predictor X_i

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

Regression Francis Galton 19th century scientist - tendency for regression of children's heights towards the mean.

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's given a new X
2. An understanding of how the mean of Y, $E(Y)$, changes with X

Review – The Summation Operator:
(see also Appendix A of the textbook)

Define

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \ldots + a_n$$

Fact 1:

$$\sum_{i=1}^{n}(a_i + b_i) \quad = \quad \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i$$

Fact 2: If $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ then

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

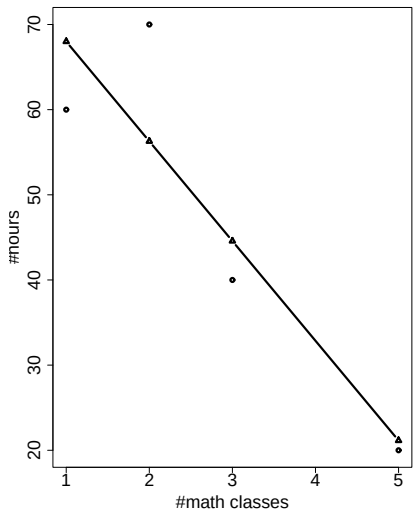
Fact 3:

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

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 writing 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

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

the **normal equations (very important)!**

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

Data: (1,60), (2,70), (3,40), (5,20)

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

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

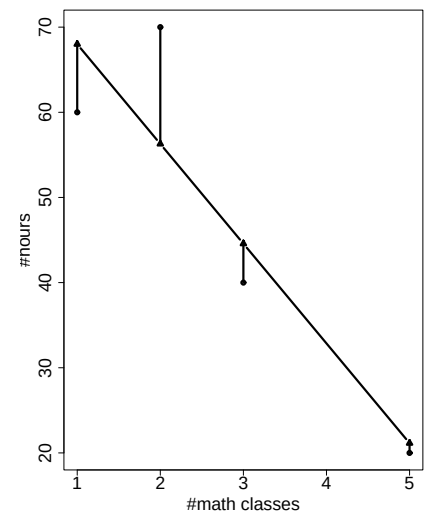
$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^n X_i Y_i - \frac{1}{n}(\sum_{i=1}^n X_i)(\sum_{i=1}^n Y_i)}{\sum_{i=1}^n X_i^2 - \frac{1}{n}(\sum_{i=1}^n X_i)^2} \\ &= \frac{420 - \frac{1}{4}(11)(190)}{39 - \frac{1}{4}(11)^2} = \frac{-102.5}{8.75} \\ &= -11.7 \end{aligned}$$

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

Estimated regression function

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

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



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

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

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

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

Fitted Values: Define

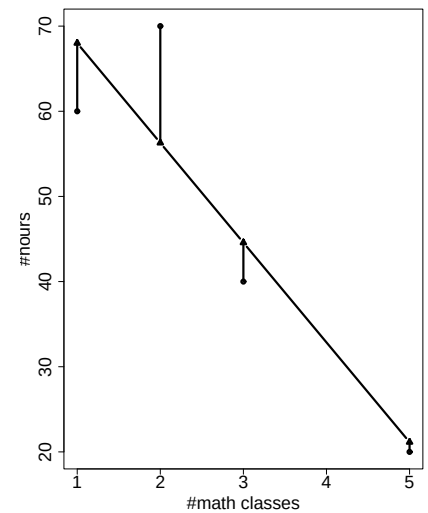
$$\hat{Y}_i = b_0 + b_1 X_i, \quad i = 1, 2, \ldots, n$$

$\hat{Y}_i$ 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 the ith residual. The vertical distance between the ith Y value and the line.



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 minimizes Q with respect to β_0 and β_1.

- 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!
Prediction versus estimation.

Estimation of σ^2 in SLR:

Motivation from iid (independent & identically distributed) case, where $Y_1, \ldots, Y_n$ iid with $E(Y_i) = \mu$ and

$$\sigma^2 = Var(Y_i) = E\left\{(Y_i - \mu)^2\right\} .$$

Sample variance (two steps)

1. Find

$$\sum_{i=1}^{n}(Y_i - \widehat{E(Y_i)})^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

   Square the difference between each observation and the estimate of its mean.

2. Take the average, except divide by **degrees of freedom** rather than n

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

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

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

Let's do the same two steps.

1. find

$$\sum_{i=1}^{n}(Y_i - \widehat{E(Y_i)})^2 = \sum_{i=1}^{n}(Y_i - (b_0 + b_1 X_i))^2 = SSE.$$

   Square the difference between each observation and the estimate of its mean.

2. divide by degrees of freedom

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - (b_0 + b_1 X_i))^2 = MSE.$$

   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

$$E(s^2) = E(MSE) = \sigma^2.$$

No matter what 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.

Normal Error Regression Model

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

Assumptions:

- Y_i is the value of the **response variable** in the ith trial
- X_i's are **predictor values**
- 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

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

The assumption of normal errors allows us to make exact inferences about the model parameters.

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

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

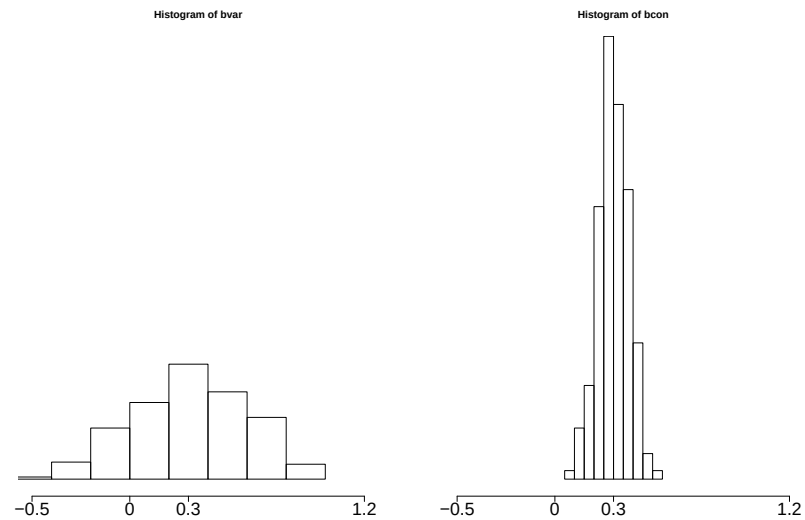
Conclusion: b_1 is not zero, so
#siblings is linearly related to #hours,right?

WRONG!

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

Think of repeatedly collecting data **(repeated sampling)** and recalculating b_1 for each data. Then 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
$H_0 : \beta_1 = 0$ (no association between Y and X)
$H_A : \beta_1 \neq 0$ (there is a linear relationship between $E(Y)$ and X)