FORMAL STATEMENT OF CLASSICAL SIMPLE LINEAR REGRESSION MODEL:

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

where Y_i is the value of the response variable Y in the ith trial

β_0 and β_1 are parameters

intercept parameter

slope parameter

X_i is a known constant, i.e., the value of the predictor in the ith trial

ε_i is a random error terms with

$\mathbb{E}[\varepsilon_i] = 0$ mean of errors is zero

$Var(\varepsilon_i) = \sigma^2$ homoscedasticity

$Corr(\varepsilon_i, \varepsilon_j) = 0$ no serial correlation no autocorrelation

It's $\boxed{\text{SIMPLE}}$ there is one independent variable.

It's $\boxed{\text{CLASSICAL}}$ because there are no weird bells or whistles → some folks take this word to mean that the error terms are assumed Gaussian.

It's $\boxed{\text{LINEAR}}$ in the parameters and linear in the predictor variables.

EXAMPLES: $Y_i = \beta_0/\beta_1 + \beta_1 X_i + \varepsilon_i$ ← non-linear; divide β_0 by β_1

$Y_i = \beta_0 + \beta_1 X_i^2 + \varepsilon_i$ ← no longer linear in X_i; however, it's linear in X_i^2

NOTE: It's useful to view a linear regression model as having a predictable and unpredictable component.

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\substack{\text{predictable part;} \\ \text{deterministic part}}} + \underbrace{\varepsilon_i}_{\substack{\text{random component;} \\ \text{stochastic part}}} \qquad (1)$$

DEF: The <u>regression function</u> is obtained by taking the expected value of (1):

$$\mathbb{E}[Y_i] = \mathbb{E}[\beta_0 + \beta_1 X_i + \varepsilon_i] = \mathbb{E}[\beta_0] + \mathbb{E}[\beta_1 X_i] + \overset{0}{\cancel{\mathbb{E}[\varepsilon_i]}}$$

$$= \beta_0 + \beta_1 X_i$$

this is the mean value of the response variable at X_i

Qn: What is the variance of Y_i? What is the variance of a response variable?

$$\Rightarrow \mathrm{Var}(Y_i) = \mathrm{Var}(\beta_0 + \beta_1 X_i + \varepsilon_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2$$

CONCLUSION: The model dictates that the Y_i have mean $\beta_0 + \beta_1 X_i$ and variance σ^2.

Qn: What is $\mathrm{CORR}(Y_i, Y_j) = \rho(Y_i, Y_j)$, $i \neq j$?

$$\mathrm{CORR}(\beta_0 + \beta_1 X_i + \varepsilon_i, \beta_0 + \beta_1 X_j + \varepsilon_j) = \mathrm{CORR}(\varepsilon_i, \varepsilon_j) = 0$$

QN: How do we interpret β_0 and β_1?

EXAMPLE: $Y_i = 10 + 0.5X_i + \varepsilon_i$

Y_i: height of plant in cm

X_i: cumulative water in week one, in liters

$\boxed{\beta_0 = 10}$ If a plant is given no water, the height of the plant will be 10 cm, on average, after one week.

$\boxed{\beta_1 = 0.5}$ For each additional liter of water, the plant will grow 0.5 cm, on average, after its first week.

NOTE: Another alternative representation of this model is:

this is the average over X_i's

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(X_i - \bar{X}) + \beta_1 \bar{X} + \varepsilon_i \\ &= \underbrace{(\beta_0 + \beta_1 \bar{X})}_{\text{"new" intercept coefficient}} + \beta_1 (X_i - \bar{X}) + \varepsilon_i \end{aligned}$$

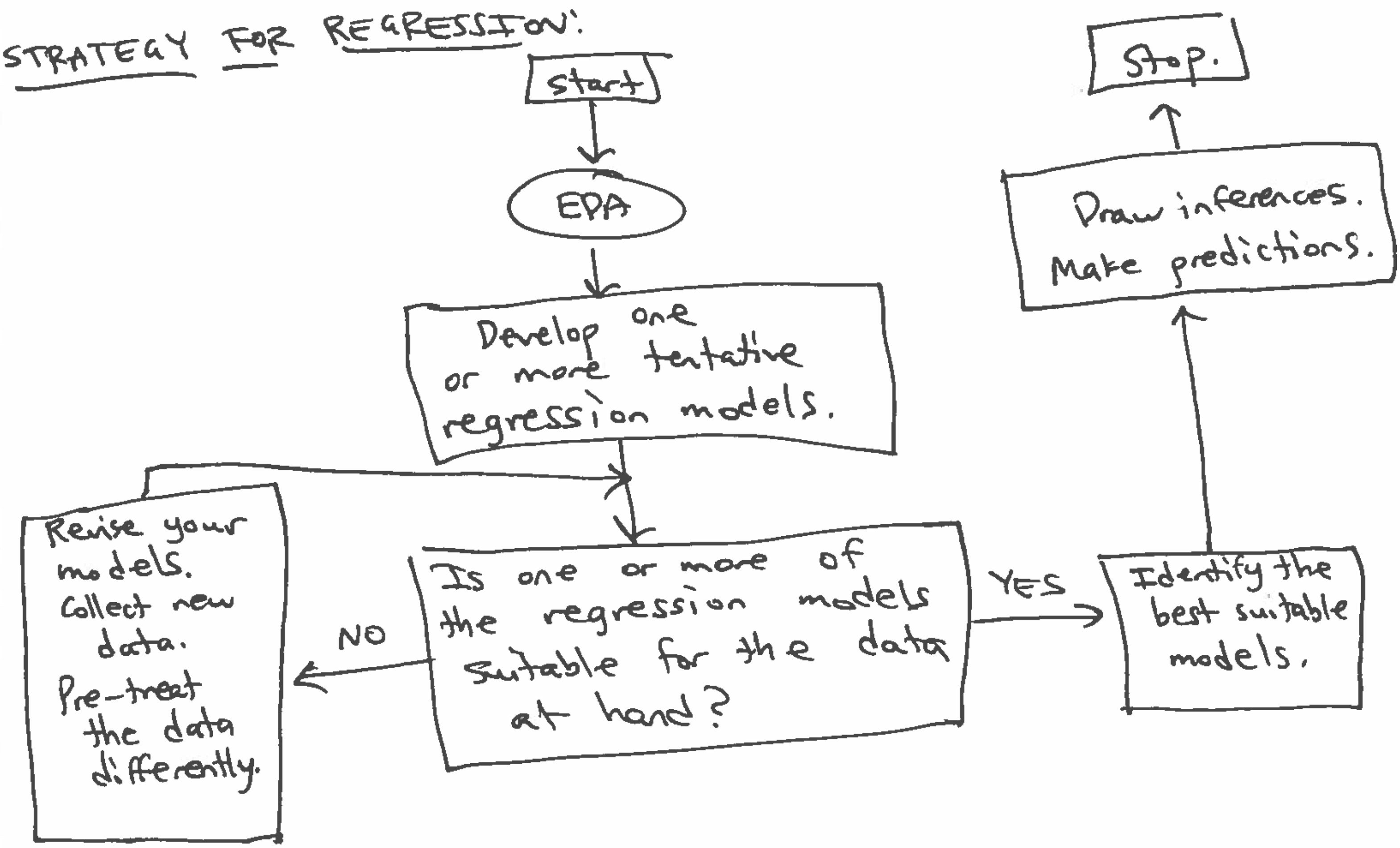
same slope coefficient

TWO MAJOR KINDS OF EXPERIMENTAL DESIGNS:

(1) **Observational Studies:** data sets collected after the fact, often as convenience samples, in which there is limited assignment of subjects to treatments. Because hidden variables can be related to both X_i and Y_i of interest, causal inference can be limited.

(2) **Controlled Experiment:** Various levels of treatment — for at some — if not all — variables are assigned at random. In general, you can infer a greater level of causality depending on the degree to which the assignment of all treatment levels was completely random.

STRATEGY FOR REGRESSION:



Obtain the least-squares estimators of β_0 and β_1.

<u>Step 1</u>: Construct the deviations between the response variable and the predictable component of the model:

$$\underbrace{Y_i - (\beta_0 + \beta_1 X_i)}_{\text{This measures the error terms we see}} = \varepsilon_i \quad \text{for } i = 1, \dots, n,$$

<u>Step 2</u>: Construct a loss function.

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

$$(b_0, b_1) = (\hat{\beta}_0, \hat{\beta}_1) = \underset{\substack{\beta_0 \in \mathbb{R} \\ \beta_1 \in \mathbb{R}}}{\arg\min} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\overline{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \Rightarrow n\overline{XY} = \sum_{i=1}^{n} X_i Y_i$$

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} Y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} X_i = 0 \\ \sum_{i=1}^{n} X_i Y_i - \beta_0 \sum_{i=1}^{n} X_i - \beta_1 \sum_{i=1}^{n} X_i^2 = 0 \end{cases}$$

NORMAL EQUATIONS

the first-order conditions

$$n\overline{X^2} = \sum_{i=1}^{n} X_i^2$$

$$\begin{cases} n\bar{Y} - n\beta_0 - \beta_1 n\bar{X} = 0 \\ n\overline{XY} - n\beta_0\bar{X} - n\beta_1\overline{X^2} = 0 \end{cases}$$

$$\begin{cases} \bar{Y} - \beta_0 - \beta_1\bar{X} = 0 & (1) \\ \overline{XY} - \beta_0\bar{X} - \beta_1\overline{X^2} = 0 & (2) \end{cases}$$

Solve (2) for β_1:

$$\beta_1 = \frac{\overline{XY} - \beta_0\bar{X}}{\overline{X^2}}$$

$$\boxed{\text{From (1), } \beta_0 = \bar{Y} - \beta_1\bar{X}}$$

$$= \frac{\overline{XY} - (\bar{Y} - \beta_1\bar{X})\bar{X}}{\overline{X^2}}$$

$$= \frac{\overline{XY} - \bar{X}\bar{Y} + \beta_1\bar{X}\bar{X}}{\overline{X^2}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X}$$

$$\Uparrow$$

$$\Rightarrow \beta_1 - \frac{\bar{X}\bar{X}}{\overline{X^2}}\beta_1 = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2}}$$

$$\Rightarrow \frac{\overline{X^2}}{\overline{X^2}}\beta_1 - \frac{\bar{X}\bar{X}}{\overline{X^2}}\beta_1 = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2}}$$

$$\hat{\beta}_1$$

$$\parallel$$

$$\Rightarrow \beta_1 = \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{XX} - \bar{X}\bar{X}} = \frac{\text{sample cov btw } X \text{ and } Y}{\text{sample var of the } X_i\text{'s}}$$

The critical value for (β_0, β_1) is obtained at

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i - \left(\sum_{i=1}^{n} X_i\right)^2}$$

and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

QUICK CHECK FOR CONCAVITY:

$$\frac{\partial^2 Q}{\partial \beta_0^2} = +2n \qquad \frac{\partial^2 Q}{\partial \beta_1^2} = +2\sum_{i=1}^{n} X_i^2$$

$$\frac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} = 2\sum_{i=1}^{n} X_i$$

$$H_Q(\beta_0, \beta_1) = \begin{pmatrix} 2n & 2\sum_{i=1}^{n} X_i \\ 2\sum_{i=1}^{n} X_i & 2\sum_{i=1}^{n} X_i^2 \end{pmatrix}$$

Qn: why is this positive definite?

$$\det H = 2\left(n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right)$$

CHECK: why is this quantity positive?

HINT: This is related to a shortcut formula for the sample variance ~~of~~ the X_i's

<u>DEF</u>: The ith (fitted) residual is the difference between the observed value Y_i and the corresponding fitted value $\hat{Y}_i$. Denote it by

$$\hat{\varepsilon}_i = e_i = Y_i - \hat{Y}_i$$

NOTE: Given sample estimators b_0 and b_1 for the regression function parameters β_0 and β_1 in the

$$\mathbb{E}[Y_i] = \beta_0 + \beta_1 X_i$$

we would estimate this function by

$$\hat{Y}_i = b_0 + b_1 X_i$$

↑ called a fitted value
"a value of the response variable"
"best prediction of Y given X
and our data"

<u>EX</u>: Suppose we collect $n=90$ data points for our plant and watering example. (We knew, from God, that $\beta_0 = 10$ and $\beta_1 = 0.5$.) From our data, we estimate $b_0 = 9$ and $b_1 = 0.6$. How would you predict the height of a plant given 1.5 liters of water?

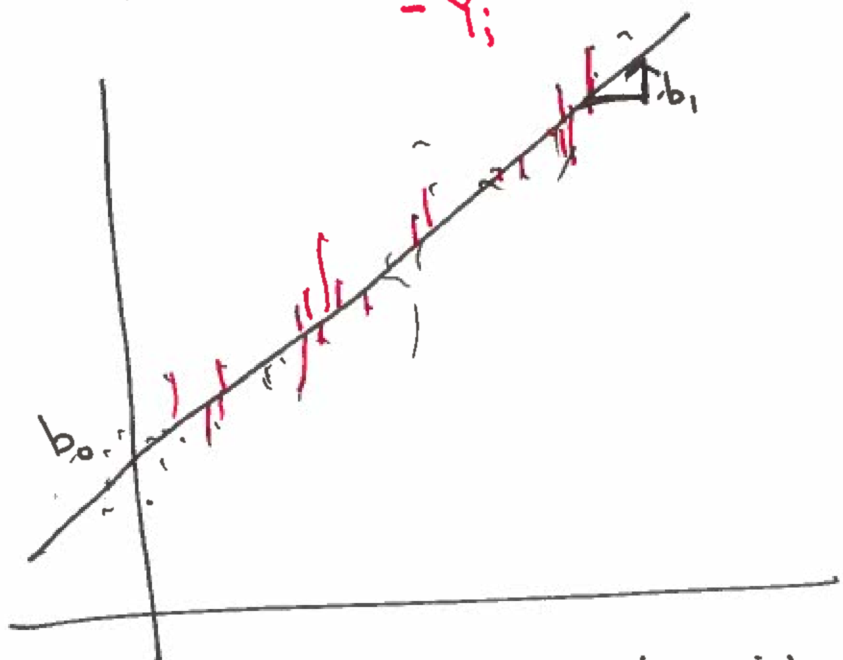
$$\hat{Y} = 0.6(1.5) + 9 = b_1 X + b_0 = 9.9$$

SIX PROPERTIES OF THE FITTED REGRESSION MODEL:

1) The fitted residuals must sum to zero: $\sum_{i=1}^{n} e_i = 0$

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \{Y_i \underbrace{- b_0 - b_1 X_i}_{-\hat{Y}_i}\} = \sum_{i=1}^{n} Y_i - n b_0 - b_1 \sum_{i=1}^{n} X_i = 0$$

because (b_0, b_1) solves/satisfies the normal equations



2) The sum of the squared fitted residuals is at a minimum; you cannot change b_0 and b_1 to make $\sum_{i=1}^{n} e_i^2$ any smaller.

3) The sum of the observed values of the response variable is equal to the sum of the fitted values of the response variable.

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

Check for yourself.

4) The ~~residuals~~ fitted residuals, when weighted by the levels of the predictor variable, sum to zero:

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

Check for yourself.

this is sometimes called the endogeneity property, and it related to the question "Are there missing independent variables from my model?"

5) The fitted residuals, when weighted by the levels of the fitted values $\hat{Y}_i$, sum to zero.

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

check for homework (related to homoscedasticity)

6) The point $(\bar{X}, \bar{Y})$ lies on the fitted regression line.

$$\begin{aligned}\bar{Y} &= b_0 + b_1 \bar{X} \\ &= \bar{Y} - b_1\bar{X} + b_1\bar{X} \\ &= \bar{Y}\end{aligned}$$

← obviously an identity

There is one additional parameter we haven't dealt with yet!

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

$\hookrightarrow \mathbb{E}[\varepsilon_i] = 0$

$\hookrightarrow Var(\varepsilon_i) = \sigma^2 > 0$

$\hookrightarrow corr(\varepsilon_i, \varepsilon_j) = 0$

THE ANSWER TO HOW TO ESTIMATE:

$$\hat{\sigma}^2 = s_e^2 = s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2}{n-2}$$

the fitted (or estimated) regression variance

NOTE: We will call $SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$, and it's called the "sum of the squared errors."

DEF: The MSE, or mean-squared error, of the regression model is $MSE = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$.

One can show, but I will not, that $\mathbb{E}[MSE] = \sigma^2$.

DEF: The normal simple linear regression model is

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

$$\varepsilon_i \sim N(0, \sigma^2)$$

and $\varepsilon_i \perp \varepsilon_j$ for $i \neq j$

–or– $\operatorname{corr}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.

You will agree that $\boxed{Y_i \mid X = X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)}$

It also means that

$$f_{Y_i|X_i}(y|x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(y - \beta_0 - \beta_1 x)^2}{2\sigma^2} \right\}.$$

Suppose that we collect a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$.

$$L((X_1,Y_1), \ldots, (X_n,Y_n); \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right\}$$

$$\ell(\beta_0, \beta_1, \sigma^2) = \sum_{i=1}^{n} -\log \sigma\sqrt{2\pi} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

$\frac{\partial \ell}{\partial \beta_0} = 0$ and $\frac{\partial \ell}{\partial \beta_1} = 0$ are normal equations (1) and (2)