Statistical Analysis + Simple Linear Regression.

What makes an analysis "statistical"?

<u>Statistics</u> : A framework for reasoning about quantitative evidence

Rough outline of a statistical analysis :

1. formulate a question.
2. Design a study and collect data
3. Choose a statistical model for the data.
4. Use the data to estimate the model
5. Make a judgment about the answer to the question.

This course is about linear models, which are designed to help answer questions like, "is response variable y related to independent variable x?" or "is y related to x after controlling for z?"

<u>Statistical Model</u> - A family of probability distributions encoding assumptions about how data were generated.

Statistical modeling is all about deciding what you want to assume to be true, and what you want to learn from data.

Simple linear model

responses : $y_1, y_2, \ldots, y_n$

covariate : $x_1, x_2, \ldots, x_n$

Question: is y related to x?

Model for y_i: $Y_i = b_0 + b_1 x_i + \varepsilon_i$

b_0, b_1 : unknown numbers (parameters)

$\varepsilon_i \overset{\text{ind}}{\sim} N(0, \sigma^2)$, σ^2: unknown parameter

ind = independent

$\Rightarrow$ Y_i is a random variable (RV)

$E(Y_i) = b_0 + b_1 x_i$ $\leftarrow$ linear model

Does this fit the definition of a statistical model?

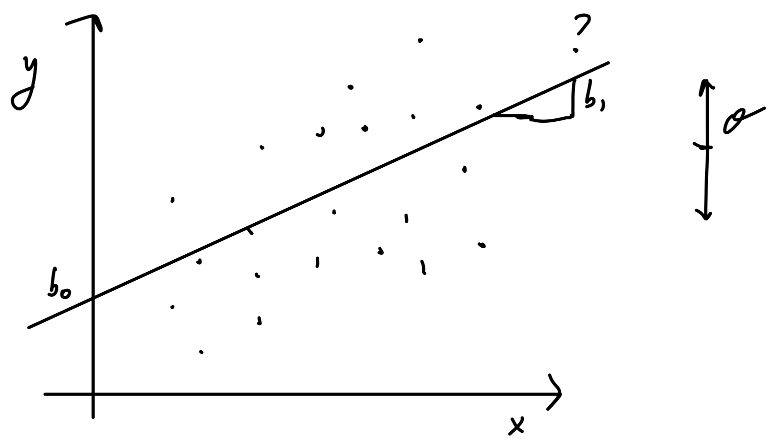
Are the model parameters relevant to our question?

Which ones?

What assumptions are encoded in this model, and how might those assumptions be violated?

Estimation: learning model parameters from data.

use $y_1, \ldots, y_n$ and $x_1, \ldots, x_n$ to get estimates of b_0, b_1, and σ^2



<u>estimate</u>: a function of the data $y_1, \ldots, y_n$ ← a number

<u>estimator</u>: random variable version of the estimate

a function of $Y_1, \ldots, Y_n$ ← a random variable

<u>example</u>: data $y_1, \ldots, y_n$

model $Y_1, \ldots, Y_n \overset{ind}{\sim} N(b_0, \sigma^2)$

estimate: $\hat{b}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i = 6.37$ (for example)

estimator: $\hat{B}_0 = \frac{1}{n} \sum_{i=1}^{n} Y_i \sim N\left(b_0, \frac{\sigma^2}{n}\right)$

The statistical model is useful because it allows us to something about how close to the true values of the parameters we expect our estimates to be.

Notation recap: b_0 = true value

$\hat{b}_0$ = estimate (number)

$\hat{B}_0$ = estimator (random variable)

B_0 = ?? (we'll come back to that)

<u>Desirable properties of estimators</u>

Bias: $E(\hat{B}) - b$ EV of estimator minus its <u>estimand</u>

we would like the bias to be small or 0

$E(\hat{B}) - b = 0$ → unbiased

$E(\hat{B}) - b = \frac{\text{something}}{n}$ → asymptotically unbiased

Variance: $Var(\hat{\beta})$

we would like the variance to be small and decrease with the sample size.

For many common estimators

$Var(\hat{\beta}) = \frac{\text{something}}{n}$ Usually not possible to beat a $1/n$ rate of convergence

$SE(\hat{\beta}) = \sqrt{Var(\hat{\beta})}$ standard error usually something$/\sqrt{n}$

For regression, we can usually find estimators that are unbiased, so we want to minimize variance.

Turns out that the <u>least squares</u> estimators do just that

<u>Sum of squares criterion</u>

$$rss(b_0^*, b_1^*) = \sum_{i=1}^{n} (y_i - b_0^* - b_1^* x_i)^2$$ residual sum of squares

b_0^*, b_1^*: arguments to rss function (like x is arg. to $f(x)$)

derivatives: $$\frac{\partial rss(b_0^*, b_1^*)}{\partial b_0^*} = -2 \sum_{i=1}^{n} (y_i - b_0^* - b_1^* x_i)$$

$$\frac{\partial rss(b_0^*, b_1^*)}{\partial b_1^*} = -2 \sum_{i=1}^{n} (y_i - b_0^* - b_1^* x_i) x_i$$

Show that setting derivatives $= 0$ yields

$$\hat{b}_1 = \sum (y_i - \bar{y})(x_i - \bar{x}) \Big/ \sum (x_i - \bar{x})^2 = sxy/sxx$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

$\boxed{\text{exercise: Verify this}}$

fitted values: $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$

residuals: $\hat{e}_i = y_i - \hat{y}_i$ $\qquad\qquad \hat{\varepsilon}_i = Y_i - \hat{B}_0 - \hat{B}_1 x_i$

$= y_i - \hat{b}_0 - \hat{b}_1 x_i \qquad \neq \varepsilon_i = Y_i - b_0 - b_1 x_i$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{b}_0 - \hat{b}_1 x_i\right)^2$$

why $1/n-2$? we'll study that later.

Properties of estimators

$$\hat{B}_1 = \frac{1}{s_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})Y_i = \sum_{i=1}^{n} c_i Y_i$$

linear function of $Y_1, \ldots, Y_n$

if $Y_1, \ldots, Y_n \overset{ind}{\sim}$ Normals $\Longrightarrow \sum c_i Y_i \sim$ Normal

$\hat{B}_0 = \bar{Y} - \hat{B}_1 \bar{x}$ $\leftarrow$ also linear

Since the estimators are normal, we just need to work out their expectations and variances.

Recall properties of linear functions of Random Variables:

RV's $A_1, \ldots, A_n$. numbers $c_1, \ldots, c_n$

$$E\left(\sum_{i=1}^{n} c_i A_i\right) = \sum_{i=1}^{n} c_i E(A_i)$$

$$Var\left(\sum_{i=1}^{n} c_i A_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j Cov(A_i, A_j)$$

$*$ $Cov(A_1, A_2) = Var(A_1) + 2Cov(A_1, A_2) + Var(A_2)$

$$E(\hat{B}_1) = E\left(\sum_{i=1}^{n} \frac{(x_i-\bar{x})}{Sxx} Y_i\right) = \sum_{i=1}^{n} \frac{(x_i-\bar{x})}{Sxx} E(Y_i)$$

$$= \frac{1}{Sxx} \sum_{i=1}^{n} (x_i-\bar{x})(b_0+b_1x_i) = \frac{1}{Sxx}\left[\sum_{i=1}^{n}(x_i-\bar{x})b_0 + \sum_{i=1}^{n}(x_i-\bar{x})x_ib_1\right]$$

$$= \frac{b_1}{Sxx}\left[\sum_{i=1}^{n}(x_i-\bar{x})x_i + \sum_{i=1}^{n}(x_i-\bar{x})\bar{x}\right] \quad 0$$

$$= \frac{b_1}{Sxx} \sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x}) = b_1 \frac{Sxx}{Sxx} = b_1$$

$$Var(\hat{B}_1) = Var\left(\sum_{i=1}^{n} \frac{(x_i-\bar{x})}{Sxx} Y_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{(x_i-\bar{x})(x_j-\bar{x})}{Sxx\quad Sxx} Cov(Y_i,Y_j)$$

$$= \frac{1}{(Sxx)^2} \sum_{i=1}^{n}(x_i-\bar{x})^2 Cov(Y_i,Y_i) = \frac{1}{(Sxx)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 Var(Y_i)$$

$$= \frac{1}{(Sxx)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\sigma^2 = \frac{\sigma^2}{Sxx}$$

$$E(\hat{B}_0) = E(\bar{Y} - \hat{B}_1\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) - \bar{x}E(\hat{B}_1)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(b_0+b_1x_i) - \bar{x}b_1$$

$$= \frac{1}{n}\sum_{i=1} b_0 + \frac{1}{n}b_1\sum_{i=1}^{n}x_i - b_1\bar{x} = \frac{nb_0}{n} = b_0 \quad \checkmark$$

$$Var(\hat{B}_0) = Var(\bar{Y} - \hat{B}_1\bar{x}) = Var(\bar{Y}) - 2\bar{x}Cov(\bar{Y},\hat{B}_1) + (\bar{x})^2 Var(\hat{B}_1)$$

$$Cov(\bar{Y},\hat{B}_1) = Cov\left(\frac{1}{n}\sum_{i=1}^{n}Y_i, \frac{1}{Sxx}\sum_{i=1}^{n}(x_i-\bar{x})Y_i\right)$$

$$= \frac{1}{n\cdot Sxx}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-\bar{x})Cov(Y_i,Y_j)$$

$$= \frac{1}{n\cdot Sxx}\sum_{i=1}^{n}(x_i-\bar{x})Var(Y_i) = \frac{\sigma^2}{n\cdot Sxx}\sum_{i=1}^{n}(x_i-\bar{x}) = 0$$

$$Var(\hat{B}_0) = \frac{\sigma^2}{n} + \frac{(\bar{x})^2\sigma^2}{Sxx} = \sigma^2\left[\frac{1}{n} + \frac{(\bar{x})^2}{Sxx}\right]$$

Putting this all together

$$\hat{B}_0 \sim N\left(b_0,\ \sigma^2\left(\frac{1}{n} + \frac{(\bar{x})^2}{Sxx}\right)\right)$$

$$\hat{B}_1 \sim N\left(b_1,\ \frac{\sigma^2}{Sxx}\right)$$

to get the full picture, also need to work out $Cov(\hat{B}_0, \hat{B}_1)$, which is not necessarily 0.

Testing:

Hypothesis about a parameter: $H_0: b_1 = b_1^*$

(often $b_1^* = 0$ is relevant)

Testing logic:

- Pick a statistic T
- decide which values of statistic constitute evidence against the null (e.g. $|T|$ large)
- decide form of decision rule (e.g. reject if $|T| \geq c$)
- Use sampling distribution of statistic to create decision rule with type 1 error probability equal to α.

Regression: $T = (\hat{B}_1 - b_1^*) / \sqrt{\hat{Var}(\hat{B}_1)} \sim t_{n-2}$

reject if $|T| \geq c$

$c = t_{n-2,\,0.975} \approx 2$

Confidence intervals: a set of plausible values for the parameter, given the data we have observed.

formally: a $1-\alpha$ confidence interval is a realization of a random interval that had probability $1-\alpha$ of containing the true parameter, regardless of its value

How to get one?

let $A = (a_1, a_2)$ be our conf. int.

define: $b_1^* \in A$ if we fail to reject $H_0: b_1 = b_1^*$ at level α

$b_1^* \notin A$ if we reject $H_0: b_1 = b_1^*$ at level α.

$$P(b_1 \in A) = P(\text{fail to reject } H_0: b_1 = b_1)$$
$$= 1 - P(\text{reject } H_0: b_1 = b_1)$$
$$= 1 - \alpha$$

This definition comports with our informal definition of a confidence interval as a set of plausible values: it contains those values that we couldn't rule out.