Statistical Linear Models

Recall

Goal: 1. Find the best possible estimates for β_0 and β_1. (Least squares)

↑ why is it "best"?

2. Assess "signal-to-noise".

• Hypothesis Test. → $H_0: \beta_1 = 0$ [line is flat → no linear relationship]

$H_a: \beta_1 \neq 0$ [line isn't flat → linear relationship]

Recall from last time:

The simple linear regression model relates two quantitative variables X and Y through the linear model

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$$

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

indep + homoskedastic

We investigated ways to fit a regression line by obtaining least squares estimates for the model parameters:

$$\boxed{\begin{aligned} \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 &= \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \end{aligned}}$$

"What makes the least squares estimates so good?"

1. They are unbiased. $E\left[\underset{\substack{\uparrow \\ \text{sample}}}{\text{estimate}}\right] = \underset{\substack{\uparrow \\ \text{population}}}{\text{parameter}}$

$$\underbrace{E[\hat{\beta}_0] = \beta_0} \quad + \quad \underbrace{E[\hat{\beta}_1] = \beta_1}$$

Try these at home.

2. Among all linear, unbiased estimators, they have

the smallest variance. (Gauss Markov Theorem pg 15-16)
They are called BLUEs.
Best Linear Unbiased Estimators

Fitting the model is EASY! The skill is in assessing its adequacy!

0). Does a linear model make sense? Scatterplot
1). Are the assumptions reasonable? Residual Plots
2). How well is Y's variability explained by X? R^2

$$R^2 = 1 - \frac{SSE}{SST} \quad \text{where} \quad SSE = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$
$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

3) Does a linear model make sense? Part II Test $H_0: \beta_1 = 0$

4) Can we make predictions from our model? $H_a: \beta_1 \neq 0$

If all we have is a sample statistic $(\bar{X}, \hat{\beta}_0 \text{ or } \hat{\beta}_1)$, how do we draw inference about the population parameters, (μ, β_0, β_1)? Hypothesis Testing

Defn: A hypothesis is a claim or assertion about the value of a parameter of a probability distn. (some characteristic about the population)

null hypothesis: "H_0" is the "prior belief" claim initially assumed to be true

(hypothesis of no consequence)

alternative hypothesis: "H_a" is an assertion that is somehow contradictory to H_0

(alternative explanation)

The null hypothesis is innocent until proven guilty

The null hypothesis is innocent until proven guilty beyond a reasonable doubt. We require a large body of evidence (in the form of our data) before we feel comfortable rejecting it.

Can either
- reject H_0 (if we have evidence beyond a reasonable doubt)
- fail to reject H_0

How do we test our hypothesis?

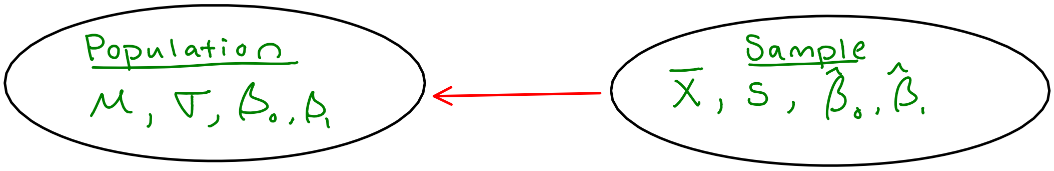
1. State null & alternative hypotheses.
2. Decide upon a test statistic calculated from data on which decision to reject H_0 (or not) will be base
3. Determine a rejection region such that if test statistic falls in that region, H_0 will be rejected.
4. Collect data, and Compute test statistic from data
5. If the value of the test statistic lies in rejection region, then reject H_0.

We decide upon a rejection region to avoid making errors:

There's a tradeoff here.

Type I error: rejecting H_0 when it's true [convicting a guilty person]

Type II error: failing to reject H_0 when it's false [setting a guilty person free]



Population
$\mu, \sigma, \beta_0, \beta_1$
Sample
$\bar{X}, S, \hat{\beta}_0, \hat{\beta}_1$

What do we Mean by Mean?

The word Mean has 2 meanings

The True Mean of X: also known as

- μ
- population mean
- Expected value or $E[x]$

→ Denotes the average value of a random variable if we could examine All points in our sample space.

- It is a real number (a constant) that is usually forever unknown to us.

Examples

1. The fraction of voters planning to vote independent in the next election.

2. The average lifetime of a component.

The Sample Mean of X also known as $\bar{X}$ or $\bar{X}_n$.

→ Denotes the average value of N independent random variables with the same distribution as X.

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

Examples

1. $\bar{X}$ = The fraction of 1000 voters independent.

2. $\bar{X}$ = The average lifetime of 25 randomly chosen components.

Subtle Point:

Before we observe the numbers $X_1, .., X_n$, $\bar{X}$ is itself a random variable.

We call $\bar{X}$ an estimator of μ.

After we observe the numbers $X_1, ..., X_n$, $\bar{X}$ is just a number.

Now we call $\bar{X}$ an estimate of μ.

When n is large, then we almost always have $\bar{X} \approx \mu$.

One of the goals of statistics is to determine how large n must be before we are "close enough."

Likewise, a random variable X has a

TRUE Variance, also Known as
- σ^2
- population variance
- $Var(X)$

→ Measures the underlying average squared difference from the true mean.

$$\sigma^2 = E[(X-\mu)^2]$$

Since we usually don't know σ^2 exactly, we estimate it using the

SAMPLE Variance also Known as
- S^2

→ Given n indep. observations (if we somehow Knew μ) then

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \text{"}\mu\text{"})^2}{n}$$

(average squared distance from μ)

If we don't know μ, then we estimate μ with $\bar{X}$.
Thus, it makes sense to use

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

However, it turns out to be better to define the Sample variance as:

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

and the sample standard deviation as:

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

Question? Why is it better to divide by $n-1$?

Short answer: S^2 is an unbiased estimator of σ^2

That is: $E[s^2] = \sigma^2$

Where as: $E\left(\frac{\Sigma (X_i - \bar{X})^2}{n}\right) = \frac{n-1}{n}\sigma^2$

For proof: see below

$$E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = E\left[\Sigma X_i^2 - 2\bar{X}\Sigma X_i + \Sigma \bar{X}^2\right]$$

$$= E\left[\Sigma X_i^2 - 2n\bar{X}^2 + n\bar{X}^2\right]$$

$$= E\left[\Sigma X_i^2 - n\bar{X}^2\right]$$

$$= \Sigma E(X_i^2) - nE(\bar{X}^2)$$

Since $E(Y^2) = Var(Y) + (E(Y))^2$

$$= \Sigma\left[Var(X_i) + (E(X_i))^2\right] - n\left[Var(\bar{X}) + (E(\bar{X}))^2\right]$$

$$= n\sigma^2 + n\mu^2 - n\left(\sigma^2/n + \mu^2\right)$$

$$= (n-1)\sigma^2$$

$$\therefore\ E(s^2) = E\left[\frac{\Sigma(X_i - \bar{X})^2}{n-1}\right] = \frac{(n-1)\sigma^2}{n-1} = \sigma^2$$

Why is this important?

The observations on X and Y taken during an experiment are random variables. If the experiment were repeated (with different subjects) then the observations would be slightly different. Consequently, parameter estimates would be slightly different also!!

In this sense, the parameter estimates obtained from the data are random variables. (why?)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \underset{\substack{\uparrow \\ \text{try at home} \\ \text{or see book for} \\ \text{details.}}}{\cdots} = \frac{\sum_{i=1}^{n}\boxed{(X_i - \bar{X})}\,Y_i}{\boxed{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \sum_{i=1}^{n} c_i Y_i$$

where $c_i = \frac{(X_i - \bar{X})}{S_{xx}}$

Assuming X (our indep. random variable) is fixed, $\hat{\beta}_1$ is just a linear combination of the (random) response variable

Assuming X (our indep. random variable) is fixed, $\hat{\beta}_1$ is just a linear combination of the (random) response variable Y!

So what is it's mean, variance + distribution?

Mean of $\hat{\beta}_1$: $E[\hat{\beta}_1] = E\left[\sum_{i=1}^{n} c_i Y_i\right] = \sum_{i=1}^{n} c_i \, E[Y_i]$

$$= \sum_{i=1}^{n} c_i \, E[\underbrace{\beta_0 + \beta_1 X_i}_{\text{const.}} + \underbrace{\varepsilon_i}_{\text{Random v. } (E[\varepsilon_i]=0)}]$$

$$= \sum_{i=1}^{n} c_i \beta_0 + \sum_{i=1}^{n} c_i \beta_1 x_i + 0$$

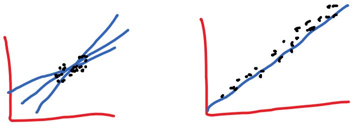
$$= \beta_0 \underbrace{\sum_{i=1}^{n} c_i}_{=0} + \beta_1 \underbrace{\sum_{i=1}^{n} c_i x_i}_{=1} = \beta_1$$

← prove at home!

Our least squares estimate is unbiased!!

Variance of $\hat{\beta}_1$: $Var(\hat{\beta}_1) = Var\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} c_i^2 \, Var(Y_i)$

$$= \sum_{i=1}^{n} c_i^2 \, Var(\underbrace{\beta_0 + \beta_1 X_i}_{\text{const.}} + \underbrace{\varepsilon_i}_{\text{R.V. } Var(\varepsilon_i)=\sigma^2})$$



$$= \sum_{i=1}^{n} c_i^2 \, \sigma^2$$

$$= \sigma^2 \underbrace{\sum_{i=1}^{n} c_i^2}_{= \frac{1}{\sum (X_i - \bar{X})^2} = \frac{1}{S_{xx}}} = \frac{\sigma^2}{S_{xx}}$$

← our slope estimate is more stable (smaller variance) when the X_i are spread out

Distribution of $\hat{\beta}_1$:

Since $\hat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i = \sum_{i=1}^{n} c_i (\beta_0 + \beta_1 X_i + \overset{\text{normal}}{\varepsilon_i})$

$$\boxed{\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)}$$

We will very often not know σ^2 directly. However, we can estimate σ^2 with s^2 where

$$s^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y})^2}{n-2} = \frac{SSE}{n-2} = \text{"Mean Square Error"} = MSE$$

Replacing σ^2 by its estimate s^2, the sample variance of $\hat{\beta}_1$ becomes:

$$\boxed{S^2_{\hat{\beta}_1} = \frac{s^2}{S_{xx}} = \frac{MSE}{S_{xx}}}$$

Confidence Interval: We can formulate a $100(1-\alpha)\%$ conf. interval for the slope parameter

$$\boxed{\hat{\beta}_1 \pm t_{\alpha/2, n-2} \frac{s}{\sqrt{S_{xx}}}}$$

Hypothesis Testing: We can also formulate hypotheses regarding the slope parameter and test them using our distribution assumptions above.

Null Hypothesis : $H_0: \beta_1 = 0$

Alternative Hypothesis : $H_a: \beta_1 \neq 0$

using the test statistic $T = \dfrac{\hat{\beta}_1 - 0}{s / \sqrt{S_{xx}}}$

where $T \sim t_{\alpha/2, n-2}$

We reject H_0 if $|T| > t_{\alpha/2, n-2}$

If we reject H_0, then we can conclude that the line has a nonzero slope.

If we fail to reject H_0, then we can't conclude that a linear relationship exists between Y and X.