Math 183
Statistical Methods

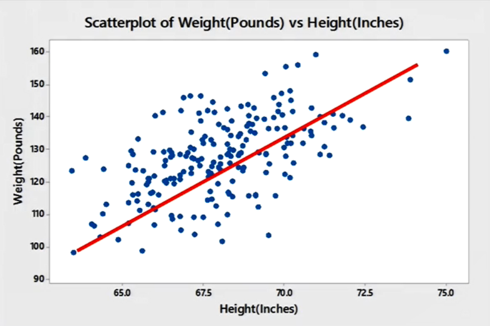
Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Today: Chapter 7 (end)

- Inference for linear regression
- Sampling distribution of the slope of the regression model
- Make C.I.'s for this slope
- Testing association

Inference About Regression



Recall our setup:

Take a data set where each data point has two values (here, height and weight)

Plot these and have the computer determine a line of best fit (or linear regression)

This line has the form

$$\hat{y} = b_0 + b_1 \cdot x$$

Here,

$$\widehat{\text{Weight}} = -111 + 3.51 \cdot \text{Height}$$

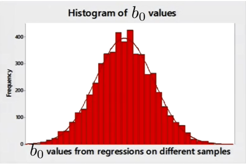
When There's A Sample, There's A Population

But... Those People are Just a Sample From the Population

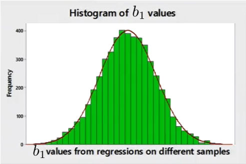
Population: Everyone in the U.S.
Parameter Model: $\hat{y} = \beta_0 + \beta_1 \cdot x$

From this population, we could virtually get many samples:

Sample: 250 people in the U.S. Statistic-Based Model: $\hat{y} = b_0 + b_1 \cdot x$	Sample: 250 people in the U.S. Statistic-Based Model: $\hat{y} = b_0 + b_1 \cdot x$	. . .

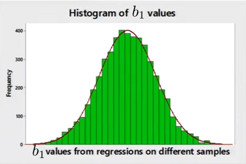


For both of these, we wonder about:
– Center, The SE
– Curve best fitting the histogram
– What conditions for this curve to actually fit the histogram



Exploring the Regression Slope b_1

We're not interested here in the intercept b_0. The important idea to explore is almost always the slope b_1 (which encodes variations!).



Where is the histogram of all the possible b_1's centered?

At the true population value β_1.

What about the spread?

$$SE_{b_1} = \frac{s_e}{s_x\sqrt{n-1}},$$

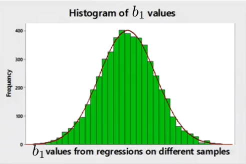
where

- s_e: Standard deviation of the residuals
- s_x: Standard deviation of the x values

Exploring the Regression Slope b_1

What curve best approximates the histogram?

Under the conditions below, the histogram is approximately t_{n-2}.



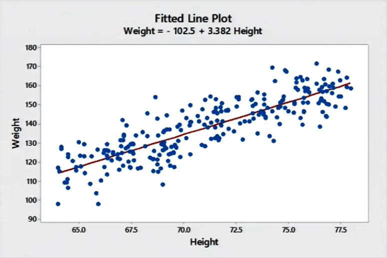
What conditions do we need to check to ensure the curve is t_{n-2}?

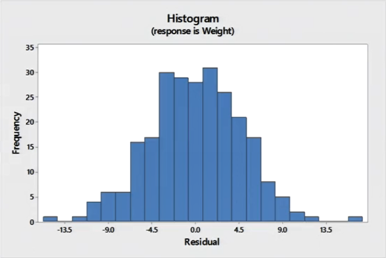
Those four conditions for creating a regression model:

- Roughly linear data
- Independence of observations
- Nearly normal residual histogram
- Constant variability around the regressio line

Example

You are curious how much an “inch of human being” weighs. To determine this, you plan to collect the data of 250 randomly picked Americans and build a regression model that predicts weight based on height.
You do so and get the below scatterplot and residuals plot.





Discuss if we meet the four conditions for linear regression.

Example

The scatterplot shows a linear trend, the residuals look roughly normal, we get independence from Randomization and the <10% rule, and the variability looks roughly constant at each value of x.

We get the regression line

$$\widehat{\text{Weight}} = -102.5 + 3.382 \cdot \text{Height}.$$

Why is it inappropriate to conclude that, on average, every inch of height adds 3.382 lbs?

The value $b_1 = 3.382$ is a statistic built on a sample of 250 Americans. A different sample would give rise to a different regression line and a different value for b_1.

Parameter VS Point Estimate... Again!

We wish to use statistical inference to estimate β_1, which is the weight, on average, for "an inch of American" (if we were to make a regression model based on **everyone** in the U.S.)

b_1 gives an estimate for β_1, and we saw earlier that b_1 is modelled by t_{n-2} centered at β_1 with $SE_{b_1} = \dfrac{s_e}{s_x\sqrt{n-1}}$.

If the conditions for inference are satisfied, we can build a confidence interval as we usually do:

$$\text{point estimate} \pm t^*_{n-2} \cdot SE.$$

Here, we would set our Confidence Interval as

$$C.I. = b_1 \pm t^*_{n-2} \cdot \frac{s_e}{s_x\sqrt{n-1}}.$$

Note: To find s_e, you'll need technology.
(Or a lot of time to lose doing it by hand!)

Reading These Values With Technology

You fit the line and notice this output:

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
8.19576  72.22%     72.11%      71.78%


Coefficients

Term          Coef  SE Coef  T-Value  P-Value
Constant   -102.50     9.48   -10.81    0.000
Height       3.382    0.133    25.39    0.000
```

From this we get:

- The estimated values $b_0 = -102.50$ and $b_1 = 3.382$
- The SE's for b_0 (9.48) and b_1 (0.133). This means that:

$$SE_{b_1} = \frac{s_e}{s_x\sqrt{n-1}} = 0.133.$$

Note: This output comes from the software Minitab. There are many software packages that focus on statistics/data science (see future slide).

Building Our Confidence Interval

From previous slides, $b_1 = 3.382$ and $SE_{b_1} = 0.133$.

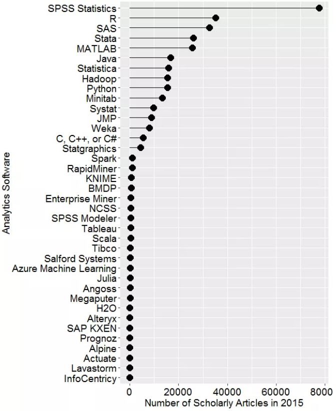
Here, $n = 250$, so for a 95% confidence level, a table gives $t^*_{248} \simeq 1.969$.

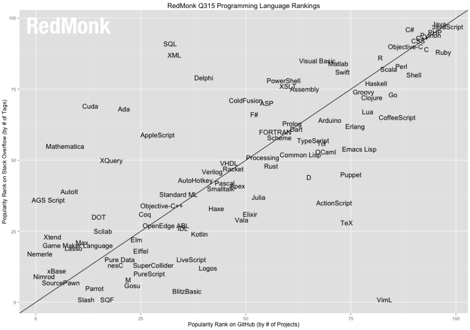
Our 95% confidence interval is

$$\begin{aligned} C.I. &= 3.382 \pm 1.969 \cdot 0.133 \\ &= (3.12\ lb/inch, 3.64\ lb/inch). \end{aligned}$$

We are 95% confident that the weight of an inch of American is between 3.12 lbs and 3.64 lbs.

The Data on Statistical Software

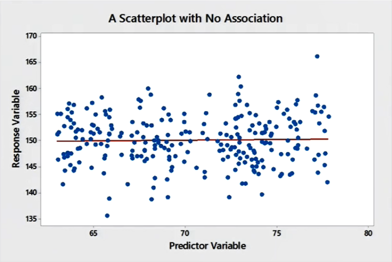




(Source 1, Source 2)

Hypothesis Testing on Slopes of Regression Lines

Typically, a hypothesis test on a slope sets H_0: $\beta_1 = 0$.



When two variables have no association, the slope of the regression line is 0 and the scatterplot looks like noise.

Here, x doesn't help predict y at all!

We tend to use a two-sided alternative H_A: $\beta_1 \neq 0$.

If the slope isn't 0, we have an association (which may be weak or strong, positive or negative).

As usual, we calculate a test-statistic by finding

$$\frac{\text{estimate} - \text{null value}}{SE}$$

In this case we find

$$T_{n-2} = \frac{b_1 - 0}{SE_{b_1}} = 25.39$$

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
8.19576	72.22%	72.11%	71.78%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	-102.50	9.48	-10.81	0.000
Height	3.382	0.133	25.39	0.000

We also a p-value $p = 0.000$. (from line "Height")

Since $p < 0.05$, we'd reject the null: there is an association between Teight and weight.

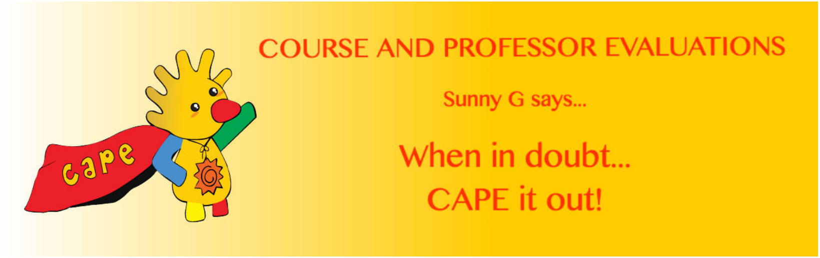
Indeed, our 95% C.I. for β_1 was $(3.12, 3.64)$ (which does not contain the value 0).

Remark: This p-value is always computed for a two-sided alternative hypothesis.

Course and Professor Evaluation (CAPE)

Don't forget to give (official) feedback on the course on

http://www.cape.ucsd.edu



Teaching and Beauty

Research were curious if the attractiveness of a professor would affect his/her teaching evaluations. (Source)
To test this, researchers collected data of 463 randomly picked professors:

- Average teaching evaluation:

  1 (worst) – 5 (best)

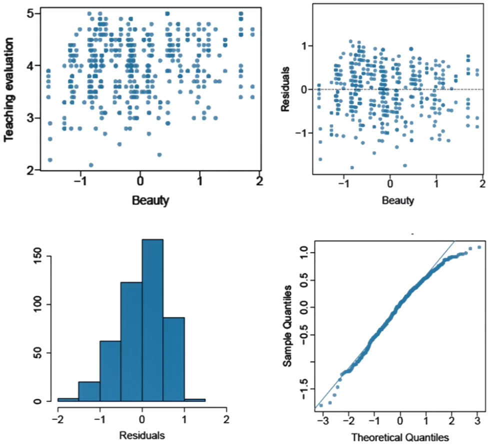
- Standardized attractiveness score:

  0 (average), - (< average), + (> average)

What are the null and alternative hypotheses for this study?

H_0:	Beauty and teaching have no association	$\beta_1 = 0$
H_A:	Beauty and evaluations have some associations	$\beta_1 \neq 0$

Given These 4 Plots, Should We Conduct the Study?



- The scatterplot almost looks like noise. Hard to say if it's linear. Note that weak associations will look a little like noise.
- Independence: Okay from randomization and the <10% rule.
- Normal residuals: Okay from the two bottom plots. Some worry about profs near the extremes of the beauty scale though.
- Constant variance: The residuals plot suggests this is true. Some concerns for the upper end of the beauty scale.

You get the below incomplete printout. Try and complete it.

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	4.010	0.0255	157.21	0.0000
Beauty	0.133	0.0322	4.13	0.0000

Under the null, the β_1 sampling distribution is modeled by t_{n-2}. Also, the test statistic is

$$T_{n-2} = \frac{\text{estimate} - 0}{SE}.$$

The output gives us

$$4.13 = \frac{\text{estimate} - 0}{0.0322},$$

thus we get

$$\text{estimate} = 4.13 \times 0.0322 \simeq 0.133.$$

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	4.010	0.0255	157.21	0.0000
Beauty	0.133	0.0322	4.13	0.0000

What is the regression for our particular sample?

$$\widehat{\text{Teach Score}} = 4.01 + 0.133 \cdot (\text{Beauty Score})$$

What does the value 4.010 mean?

It is the y-intercept of the regression line. So, it is the Teach Score we expect for professors with Beauty 0 (average).

What conclusion should the researcher draw about this test?

Given that the p-value is about 0, they should reject the null:
There does appear to be an association between teaching evaluations and beauty.

Back to Old Faithful

From our study that predicts (time until eruption) of Old Faithful based on (Time of last eruption) using 270 observations, we get this R printout.

Build a 90% C.I. for how much each second of eruption creates in waiting time for the next eruption. Is there really an association between these two ideas?

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 33.987808   1.181217   28.77   <2e-16 ***
Duration     0.176863   0.005352   33.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For inference on the slope of a regression,

$$C.I. = b_1 \pm t^*_{n-2} \cdot SE_{b_1}.$$

Based on the printout, we have

$$C.I. = 0.176 \pm t^*_{268} \cdot 0.00535.$$

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	31	1.31	1.70	2.04	2.45	2.74
	32	1.31	1.69	2.04	2.45	2.74
	33	1.31	1.69	2.03	2.44	2.73
	34	1.31	1.69	2.03	2.44	2.73
	35	1.31	1.69	2.03	2.44	2.72
	36	1.31	1.69	2.03	2.43	2.72
	37	1.30	1.69	2.03	2.43	2.72
	38	1.30	1.69	2.02	2.43	2.71
	39	1.30	1.68	2.02	2.43	2.71
	40	1.30	1.68	2.02	2.42	2.70
	41	1.30	1.68	2.02	2.42	2.70
	42	1.30	1.68	2.02	2.42	2.70
	43	1.30	1.68	2.02	2.42	2.70
	44	1.30	1.68	2.02	2.41	2.69
	45	1.30	1.68	2.01	2.41	2.69
	46	1.30	1.68	2.01	2.41	2.69
	47	1.30	1.68	2.01	2.41	2.68
	48	1.30	1.68	2.01	2.41	2.68
	49	1.30	1.68	2.01	2.40	2.68
	50	1.30	1.68	2.01	2.40	2.68
	60	1.30	1.67	2.00	2.39	2.66
	70	1.29	1.67	1.99	2.38	2.65
	80	1.29	1.66	1.99	2.37	2.64
	90	1.29	1.66	1.99	2.37	2.63
	100	1.29	1.66	1.98	2.36	2.63
	150	1.29	1.66	1.98	2.35	2.61
	200	1.29	1.65	1.97	2.35	2.60
	300	1.28	1.65	1.97	2.34	2.59
	400	1.28	1.65	1.97	2.34	2.59

Based on the table, $t^*_{268} \simeq 1.65$.

We get

$$C.I. = 0.176 \pm 1.65 \cdot 0.00535$$
$$= (0.167, 0.184).$$

We are 90% confident that each second of current eruption leads to between 0.167 to 0.184 second of waiting for the next eruption.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 33.987808   1.181217   28.77   <2e-16 ***
Duration     0.176863   0.005352   33.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Given the p-value $p < 210^{-16}$, we also believe that there is an association between the two variables we are studying.

The confidence interval gives a very good sense of how these variables are related.