Quick Review of Stat I – Simple Linear Regression in Two Pages

- We will investigate the relationship between two quantitative variables.
- The x variable is the independent variable. In statistics, it's the explanatory variable.
- The y variable is the dependent variable. In statistics, it's the response variable.
- The correct graph to make is a Scatterplot.
- The number one requirement is that the relationship looks linear. If it doesn't, stop the analysis and do something else (details to come).
- We can measure the strength and direction of the linear relationship between two quantitative variables by calculating the correlation coefficient, r.
- Correlation goes from -1, which is a perfect negative linear relationship, to 0, which is no linear relationship, to $+1$ which is a perfect positive linear relationship.
- Correlations beyond the ± 0.8 are considered strong.
- If the data look linearly related and the correlation is worthwhile (context dependent), the next step is to create a linear model to predict y values based on x values.
- The equation of a straight line is:

$$y = mx + b \quad \text{(Math class)}$$
$$\hat{y} = b_0 + b_1 x \quad \text{(stat class)}$$

- The slope has a formula, though we don't use it much:

$$b_1 = r \cdot \frac{S_y}{S_x}$$

S_y, S_x are SD of y, x variables

- So does the y-intercept:

$$b_0 = \bar{y} - b_1 \bar{x}$$

- Technology will output the linear regression model, so don't waste lifetime doing it by hand. We need to be able to interpret the model's important features.

- The slope tells us how much y changes for a one-unit change in x. Put it in context.

- The y-intercept tells us the value of y if x = 0. Sometimes this value makes no sense because x = 0 makes no ~~senses~~ sense in context. Sometimes this value makes no senses because the y-intercept value makes no sense (our model may be off because our data wasn't the best sample). Use your head when interpreting this value with words.

- We can grade our linear regression model by looking at R^2. This number ranges from 0% to 100%, and higher is better. It is the percentage of the variation in the y-variable that is explained by knowing the value of the x-variable.

- The whole linear modeling idea is based on minimizing the residuals. A residual is the vertical distance between a data point and the regression line. In words and symbols: Residual = Observed y − Expected y

$$e = y - \hat{y}$$

- A key summary number is the standard deviation of the residuals, s_e, also known as the standard deviation of the errors. We want this to be small.

- We can compare the standard deviation of the residuals to the standard deviation of the y-variable to get a sense of how much variation the model accounted for.

- Also, we can look at standardized residuals. The 68-95-99.7% rule kicks in, and residuals larger than ±2 are unusual and should be looked at.

- Don't extrapolate. We build regression models to do prediction – in other words, plug in an x value and see what the model says for the value of y. We must only do this for the range of x-values we have data for.

- Finally, don't infer that the x-variable causes the y-variable to change, even if there is a strong linear relationship. There is always the chance for lurking variables.

Example: Open up the "***Stat II Wal-Mart Supermarket***" dataset on StatCrunch run an entire linear regression analysis, hitting all points. We use the Wal-Mart price to predict the supermarket price. Go.

x = Walmart price , y = Supermarket Price

Scatterplot: Linear, positive, strong, about 5 products much more expensive than the rest.

Correlation: $r = 0.903$ (Strong positive)

Model: $\widehat{Super} = 0.191 + 1.133(Walmart)$

e.g. for prediction →
If something costs \$2 at WalMART,
we predict it to cost
$\widehat{Super} = 0.191 + 1.133(2) = \$2.457$
at the Supermarket.

Slope: $b_1 = 1.133$

For every \$1.00 increase at WalMART,
we expect the supermarket
price to increase by ~ \$1.13

y-intercept: If something cost $x = \$0$ at Walmart, the model predicts $\hat{y} = \$0.191$ at the Supermarket.

Meaningless answer!

$R^2 = 81.5\%$

So 81.5% of the variation in prices at the Supermarket (y) is explained by the Wal-Mart price (x).

18.5% is explained by other x-variables

Standard deviation of the errors:

$S_e = 0.608$

Look at $S_y = 1.407$

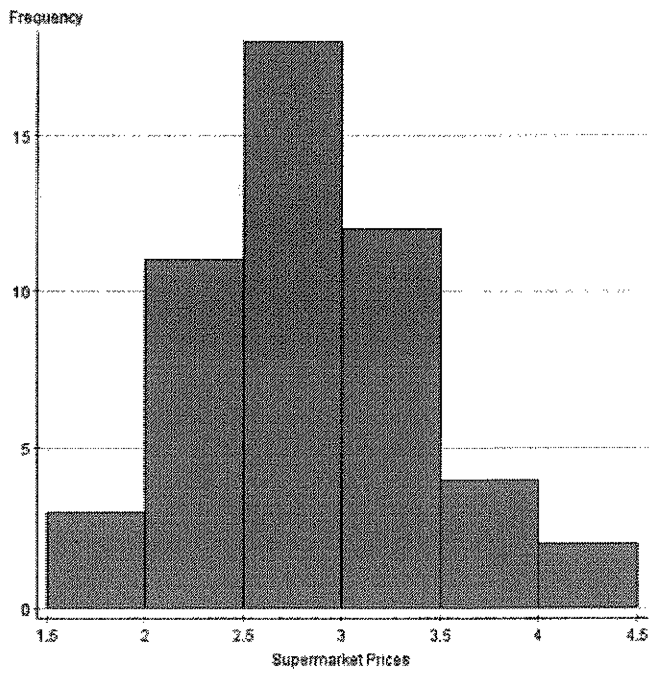
Stat II Time – Inference for Regression

Questions:

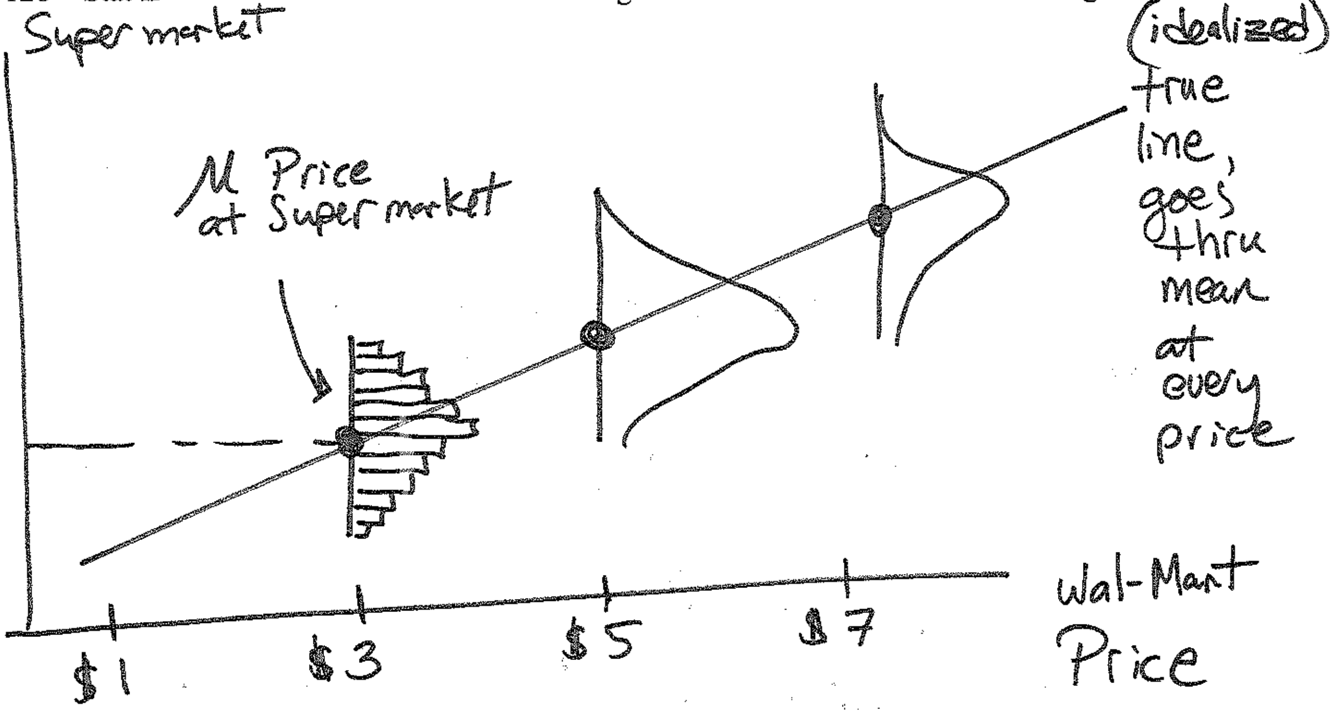
- Could the positive relationship we see between the Wal-Mart price and the supermarket price just be due to chance? Probably not.
- We can estimate the slope, but how reliable is our estimate? Probably good.

Wal-Mart Example: Our Sample versus the Entire Population

- Understand that even if we knew prices of every single product at Wal-Mart and the supermarket, our data would not line up perfectly on a straight line.
- Rather, we want to model the relationship between the two variables and we can imagine an idealized regression line.
- This idealized regression assumes the mean supermarket price falls exactly on the line for every price point at Wal-Mart.
- For a product at Wal-Mart (x) that costs \$2.49, there is a distribution of supermarket prices:



- This is true at every possible Wal-Mart price point. Sketch a graph of this on the next page:



- Our fitted line equation is: $\hat{y} = b_0 + b_1 x$
- The generalized model equation is: $\mu_y = \beta_0 + \beta_1 x$
- All models have errors. So does ours. F~~or each data poin~~t, we have:

$$y = \beta_0 + \beta_1 x + \varepsilon,$$ true for each data point.

Conditions and Assumptions for Linear Regression Inference

1. Make a scatterplot. Look for linearity.
2. If the data are straight enough, use a computer to find the regression equation, the residuals e and the predicted values $\hat{y}$.
3. Make a scatterplot of the residuals against the *x* variable – the plot should have no pattern. This ensures the residuals have equal spread at every *x*-value and that a linear model is appropriate.
4. If your *x*-variable is some measurement of time or your data values were collected in some meaningful numerical order, plot your residuals against time to check for evidence that the data values were not independent.
5. If the residual plots look OK, check the residuals for normality. We need this to do inference on our regression parameters.

Example: For the Wal-Mart and supermarket problem, check all conditions, using StatCrunch. Also note how to automatically calculate the residuals and the predicted values.

✓ Boxes for residuals, fitted values.

Scatterplot: x = WALMART
y = Residuals

No pattern ✓ Equal Spread ✓

Histo / QQ Residuals

Unimodal & Symmetric ✓
QQ off a bit ....
n = 95 so not too concerned
(n = 30 is the magic number for inference)

Regression Inference – A List of Things We Can Test

1. First, test if the slope of the regression equation is different from 0. If this test has statistically significant results, it means your regression is technically meaningful.

2. Calculate a confidence interval for the true population slope. We know our slope is just one example from the actual data we happened to collect. A confidence interval will give us a range of plausible values and deepen our understanding.

3. For a particular *x*-value of interest, we can find a confidence interval for the mean predicted *y*-value.

4. Also, we can find a prediction interval for an individual new observation at a particular *x*-value.

1. Testing if the Slope Differs From Zero

We will never calculate this by hand, but here is the formula for the standard error of the slope:

$$SE(b_1) = \frac{S_e}{\sqrt{n-1}\, S_x}$$

It clearly depends on three things →

1. The spread about the regression line, S_e. More spread about the regression line results in more variation in your slope, sample-to-sample.

2. The spread of your x values. The more spread out your x-values are, the less your slope will vary, sample-to-sample.

3. The sample size n. If you have a large sample size, your slope should vary less sample-to-sample.

Example: For fun, verify the standard error of the slope.

Summary statistics:

	Column	n	Mean	Std. Dev.	Median	Q1	Q3	Variance
X	Walmart Price	95	2.5623157	1.1208996	2.5	1.78	3.18	1.2564158
Y	Supermarket Price	95	3.0936842	1.4066174	2.99	2	3.99	1.9785725

Simple linear regression results:
Dependent Variable: Supermarket Price
Independent Variable: Walmart Price
Supermarket Price = 0.1907152 + 1.1329474 Walmart Price
Sample size: 95
R (correlation coefficient) = 0.9028
R-sq = 0.81508136
Estimate of error standard deviation: 0.6081193

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Value
Intercept	0.1907152	0.1563674	$\neq 0$	93	1.2196608	0.2257
Slope	1.1329474	0.055957485	$\neq 0$	93	20.246574	<0.0001

$$SE(b_1) = \frac{0.6081193}{\sqrt{95-1}\,(1.1208996)} \approx 0.05596$$

Question: Why test if the slope equals zero?

Answer: The regression model is: $\hat{y} = b_0 + b_1 x$

If the slope equal zero, the regression model becomes: $\hat{y} = b_0 = \bar{y}$

With no x in the model, y doesn't depend on x at all!

The intercept b_0 would turn out to be $\underline{\bar{y}}$.

In fact, for any x value, the predicted y value would be $\underline{\bar{y}}$.

Always guess $\bar{y}$ for any value of x.

Official Steps for Testing the Slope

1. The hypotheses are:

$H_0: b_1 = 0$ vs. $H_A: b_1 \neq 0$

Can do <, > tests

Can test against a non-zero number.

2. The test statistic is:

$$t = \frac{b_1 - \text{Usually Zero}}{SE(b_1)}, \quad df = n-2$$

3. The P-value would be found by shading in both directions (for our general $\neq$ test) under the Student's t model with $n-2$ degrees of freedom.

4. When running regressions using the computer, this test is automatically performed, even if you didn't want to:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Value
Intercept	0.1907152	0.1563674	≠ 0	93	1.2196608	0.2257
Slope	1.1329474	0.055957485	≠ 0	93	20.246574	<0.0001

5. We can calculate a confidence interval for the true slope of the population using:

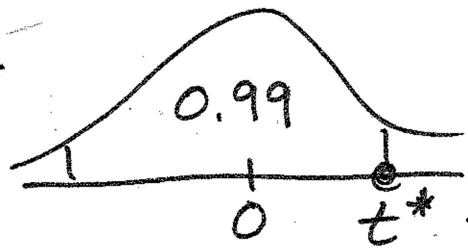
$$b_1 \pm t^* \times (SE(b_1))$$

$$df = n-2$$

Example: Back to Wal-Mart, calculate by hand a 99% confidence interval for the true population slope. Interpret. Then demonstrate on technology.

$b_1 = 1.133$ $df = 95-2 = 93$

$t^* =$



$t^* = 2.6297$

$SE(b_1) = 0.05596$

$1.133 \pm 2.6297(0.05596)$

$(0.986 , 1.280)$

We're 99% confident the true slope falls inside this interval.

A Quick Note on the *y*-Intercept

- We can test if the intercept is different from zero; it's just not that important.
- On computerized output, the intercept is also always tested – check for Wal-Mart.

Example: The Honshu Earthquake that hit Japan a few years ago caused massive destruction. Is there value in trying to predict the magnitude based on the depth of the earthquake? Run a preliminary linear regression analysis, checking the conditions and testing for the slope.

x = Depth, y = magnitude

Scatterplot: Not really linear, $r = -0.088$, very weak negative. A few outliers, namely the big one → depth = 24.4
mag = 8.9

EQUATION: $\widehat{MAG} = 5.32 - 0.0054\,(\text{Depth})$

$R^2 = 0.77\%$

* Residuals vs. x = depth, no pattern, pretty equal spread ✓

  See "big one" is almost 4 SD above predicted value.

* Data is in chronological order, so check residuals in time order (index)
  - Upward trend near "big one", suggest data values not independent.

* Residuals not normal, suggesting caution. Data set has $n = 446$ values so CLT takes over and normality not crucial.

* $H_0: b_1 = 0$ vs. $H_A: b_1 \neq 0$ (test slope)

  P-Value = 0.0638, fail to reject H_0. Since we're not convinced the slope differs from 0, this regression has limited value.

Example: Back in the Fall 2012 semester, Cecil College students randomly selected a book from the college library, and recorded the number of pages and the weight (lbs). Open the "*Library Data – Quiz 2*" dataset.

Check the conditions to predict a book's weight based on its number of pages. Test if the model is useful. Create a 95% confidence interval for the true slope. Interpret.

Scatterplot → Linear, positive, a few large books, $r = 0.729$

Residual Plot → No pattern, suggesting linearity OK. Large residuals for large books (funneling) so we must exercise caution.

Histogram of Residuals: Unimodal & Symmetric and $n = 115$ (large) so OK here.

Model: $\widehat{\text{Weight}} = 0.126 + 0.00165(\text{Pages})$

Test Slope: $H_0: b_1 = 0$

$H_A: b_1 \neq 0$

Test Stat: $t = 11.33$

P-Value < 0.0001

Reject H_0 and conclude slope is not 0.

This means that # of pages does have predictive value for weight of a book.

95% CI for true slope β_1 is

$$b_1 \pm t^*_{\underset{df}{n-2}} (\text{SE of slope})$$

$$0.00165 \pm 1.981\,(0.000146)$$

$$0.00165 \pm 0.000289$$

$$(0.001361 \text{ to } 0.001939)$$

95% confident that, on average, between 0.001361 and 0.001939 lbs. are added to a book's weight for each additional page.

STATCRUNCH:

Select CI in the regression menus. Default is HT.

Confidence Intervals for Prediction

- When we plug in an x value into our regression equation, the y value we solve for is at best an educated guess.
- There are two questions we can answer with predictions:
  - For a given x value, give a confidence interval for the mean y value.
    - **Example:** When predicting the weight of a book based on the number of pages, suppose you're interested in 600 page books. We can create a 95% confidence interval for the ***mean*** weight of 600 page books.
  - For a given x value, give a prediction interval for a particular y value.
    - **Example:** We pull a 600 page book off the shelves. We can create a 95% prediction interval for the weight of that ***particular*** book.
- Confidence intervals for mean y values at a particular x value are much fixed skinnier than the same prediction interval. It is much harder to predict the y value of the next observation.

Formulas

The symbol for a new x value is: x_v "x sub-new"

The symbol for a predicted y value is: $\hat{y}_v$ "y-hat sub-new"

Both intervals have this form, with $n-2$ degrees of freedom for the t critical value:

$$\hat{y}_v \pm t^*_{n-2} \times SE$$

The SE = standard error is different depending on which interval we are calculating.

Confidence Interval Formula for the Mean *y* value at a Particular *x* Value

$$\hat{y}_\nu \pm t^*_{n-2} (SE_{\hat{\mu}_\nu}) \quad \text{with}$$

$$SE(\hat{\mu}_\nu) = \sqrt{SE^2(b_1)\cdot(x_\nu - \bar{x})^2 + \frac{S_e^2}{n}}$$

Prediction Interval Formula for the *y* value at a Particular *x* Value

$$\hat{y}_\nu \pm t^*_{n-2} \cdot SE(\hat{y}_\nu) \quad \text{with}$$

$$SE(\hat{y}_\nu) = \sqrt{SE^2(b_1)\cdot(x_\nu - \bar{x})^2 + \frac{S_e^2}{n} + S_e^2}$$

- The standard error of a single predicted value has extra variability when compared to the standard error for the mean. This is because individuals vary around the predicted mean.
- In the formula, there is an extra term in our prediction intervals: S_e^2.
- This extra term makes prediction intervals wider, as expected.
- StatCrunch will output these values, but first, one example by hand.

Example: For the Library dataset, calculate a 95% confidence interval for the mean weight of a book with 600 pages. Interpret. Then calculate a 95% prediction interval for the same book. Compare.

$n = 115$, $df = 113$ and $t^* = 1.981$

$SE(b_1) = 0.000146$, $S_e = 0.355$

$x_\nu = 600$, $\bar{x} = 402.096$

$\hat{y}_\nu = 0.126 + 0.00165(600) = 1.116$

95% CI for mean weight at $x = 600$ pages:

$$1.116 \pm 1.981 \sqrt{(0.000146)^2(600-402.096)^2 + \frac{(0.355)^2}{115}}$$

$$1.116 \pm 0.087 \Rightarrow (1.029, 1.203)$$

We're 95% confident the interval contains the mean weight of a 600-page book.

95% PI for weight of the next 600 page book:

$$1.116 \pm 1.981 \sqrt{(0.000146)^2(600-402.096)^2 + \frac{0.355^2}{115} + 0.355^2}$$

$$1.116 \pm 0.709 \Rightarrow (0.407, 1.825)$$

We're 95% confident the next 600 page book with weigh between 0.407 lb. and 1.825 lbs.

Example: For the Wal-Mart dataset, use StatCrunch to compute a 99% confidence interval for the mean price at the supermarket if something costs \$3.49 at Wal-Mart. Then determine a 99% prediction interval for the same item. Interpret.

Under Regression menus, ☑ Predict y for x = 3.49 and 99%

We are 99% confident that the mean price at the Supermarket is between ~ ~~\$3.98~~ 3.93 and ~ ~~\$4.31~~ 4.36 if something costs \$3.49 at WalMart.

We are 99% conf. that the next \$3.49 item at WalMart will cost Between \$2.53 and \$5.76 at the SuperMarket

---

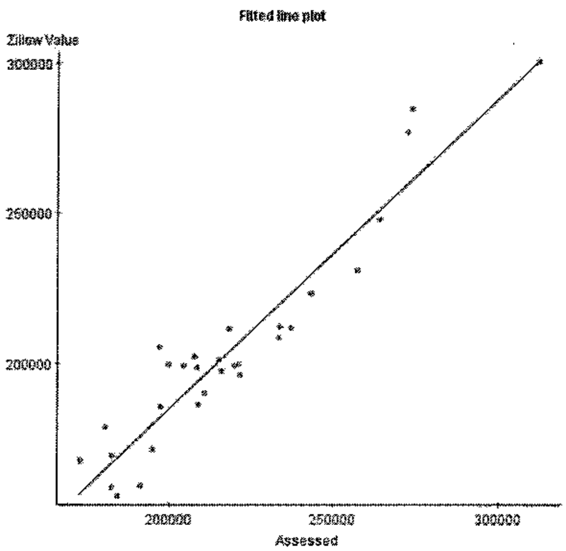
Regression Problems

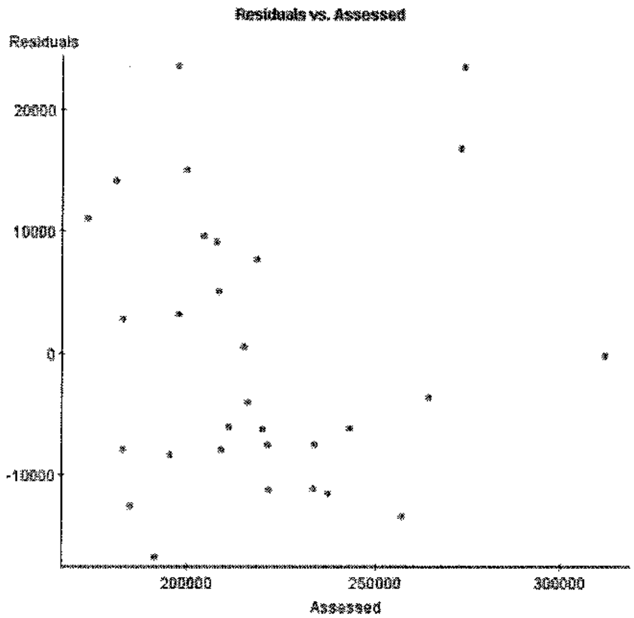
Example: How does the ***Zillow Value*** of a house depend on its ***Assessed Value***? Data from Professor Kupe's neighborhood produced the following regression output and graphs.

Simple linear regression results:
Dependent Variable: Zillow Value
Independent Variable: Assessed
Zillow Value = -22913.75 + 1.038321 Assessed
Sample size: 30
R (correlation coefficient) = 0.9478
R-sq = 0.8983507
Estimate of error standard deviation: 11460.4795

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Value
Intercept	-22913.75	14591.321	$\neq 0$	28	-1.5703685	0.1276
Slope	1.038321	0.06600575	$\neq 0$	28	15.730766	<0.0001





a. The intercept is negative. Discuss its value, taking note of its *P*-value.

b_0 = -$22,913.75

~~If we test~~ $H_0: \beta_0 = 0$, P-Value = 0.1276

is "fail to reject" H_0 and we don't have evidence it differs from 0.

(It shouldn't by the way)

b. Is the regression meaningful? Why?

We test $H_0: \beta_1 = 0$ and reject H_0 with P-Value < 0.0001.

Yes it is meaningful, we have value in Predicting Zillow Value based on Assessed value.

c. The output reports $s_e = 11,460.4795$. What does this number mean in this context?

Amounts by which Zillow Values differ from predictions made by this model vary, with a standard deviation of $11,460.48.

d. What's the value of the standard error of the slope of the regression line? What's it mean in context.

$SE(b_1) = 0.066$

If we constructed other models based on different samples of houses, we'd expect the slopes of the regression lines to vary, with standard deviation of about 0.066.

Example: In the Chicken Sandwiches dataset, we'd like to predict ***Calories*** based on ***Fat***.

a. To use a linear model, we must check many conditions. Check them all.

Scatterplot → Linear, positive ✓

Residuals vs. Fat → Random Scatter ✓

Equal spread ✓

Residuals → Unimodal, Symmetric ✓

b. Burger King has been toying with a new sandwich that would have 59 grams of fat. Predict its calories or explain why you would be hesitant to do so.

$x = 59$ g fat is extrapolation.

Biggest x-value we have is Sonic Club Toaster $x = 46$ g

Predictions, beware.

c. Give the value of the residual for the Subway Oven Roasted and explain what it means in context.

$e = 37.37$ calories

Sandwich has 37.37 more calories than predicted, based on having 5 grams of fat.

d. Predict, with a 95% interval, the number of calories the next newly-invented chicken sandwich would have if we knew it was going to have 34 grams of fat.

95% prediction interval if $x = 34$ g.

(503.4 to 759.5 calories)

e. Is there evidence of a relationship between fat grams and number of calories? Back up your answer.

We test if slope is zero

$H_0: \beta_1 = 0$ P-Value < 0.0001

$H_A: \beta_1 \neq 0$ Reject H_0

There is evidence that fat grams usefully can predict # of calories

f. Give a 95% confidence interval for the mean number of calories for sandwiches with 19 grams of fat.

We're 95% confident that 19 g fat sandwiches average between 428.0 and 463.7 calories.

g. What percentage of the variation in calories is explained by grams of fat?

$R^2 = 82.5\%$

Before ANOVA, Review Inference One Last Time

Example: Some college professors make bound lecture notes available to their students in an effort to improve teaching effectiveness. At one college, two groups of students were surveyed – 86 students enrolled in a promotional strategy class that required the purchase of lecture notes and 35 students enrolled in a sales / retailing elective that did not offer the notes. At the end of the semester, students were asked to respond: "Having a copy of the lecture notes was [would be] helpful in understanding the material". Responses were measured on a 9-point scale, where 1 = strongly disagree and 9 = strongly agree.

Summarized results are found in the table:

	Classes Buying Lecture Notes	Classes Not Buying Lecture Notes
Sample Size	**86**	**35**
Sample Mean	**8.48**	**7.80**
Sample Variance	**0.94**	**2.99**

a. Do the samples provide sufficient evidence to conclude there is a difference in the mean responses of the two groups of students? Test using $\alpha = 0.01$.

b. If we made an error, what kind of error? What would this mean in the context of the problem?

c. Interpret the *P*-value in context.

d. Construct a 99% confidence interval for the difference in sample means. Interpret the result.

a.) 2-Sample T Test for a difference in means.

$H_0: \mu_{NOTES} = \mu_{NO\ NOTES}$

$H_A: \mu_{NOTES} \neq \mu_{NO\ NOTES}$

TI: 2 Sample T Test

$t = 1.32$ P-Value = 0.195

Fail to Reject H_0, conclude there is not enough evidence to say there is a difference in mean responses.

b.) Since we failed to reject H_0, could have made a Type II error.
This means that the means do differ but our collected data did not show it.

c.) P-Value = 0.195. If the two population means were actually equal, we'd get sample means of $\bar{Y}_{NOTES} = 8.48$ and $\bar{Y}_{NO\ NOTES} = 7.80$ or ones even more different 19.5% of the time.

(19.5% is not unusual enough to convince us the $\bar{Y}$'s are different)

d.) 2-SampT Int
(-0.72 to 2.08)

99% confident this interval captures the true difference in means.
As 0 is in the interval, no difference is plausible.

Example: In the Journal of Genetic Psychology, 13 male college students were asked to complete the following survey about each of his parents: "My relationship with my father (mother) can best be described as (1) Awful! (2) Poor, (3) Average, (4) Good, or (5) Great!

The data obtained are in the table:

Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Attitude Toward Father	2	5	4	4	3	5	4	2	4	5	4	5	3
Attitude Toward Mother	3	5	3	5	4	4	5	4	5	4	5	4	3
	-1	0	1	-1	-1	1	-1	-2	-1	1	-1	1	0

a. Specify the appropriate hypotheses for testing whether male students' attitudes towards their fathers differs from their attitude towards their mothers, on average.

b. Conduct the test, using a 5% level of significance.

c. What assumptions about the sample and the population did you have to make in order to ensure the validity of the test?

a.) Dependent samples, so just run a T Test on the differences.

b.) $H_0: \mu_d = 0 \quad H_A: \mu_d \neq 0$

Put differences in TI or SC, run T Test.

$\bar{y} = -0.308$ (favoring mother)

$S_d = 1.032$

$t = -1.075 \quad$ P-Value $= 0.3033$

Fail to Reject Ho, not convinced students attitudes differ on father vs. mother.

c.) $n = 13$ so differences must be normal

Also, these guys are unbiased, represent all guys.

Example: Obstructive sleep apnea is a disorder that causes a person to stop breathing momentarily and then awaken briefly. These sleep interruptions, which can occur hundreds of times in a night, can drastically reduce the quality of rest and cause fatigue during the waking hours.

Researchers at Stanford University studied 159 commercial truck drivers and found that 124 of them suffered from obstructive sleep apnea (*Chest*, May 1995).

a. Use the study results, to estimate, with 90% confidence, the fraction of truck drivers who suffer from the sleep disorder.

b. Sleep researchers believe that about 25% of the general population suffer from obstructive sleep apnea. Comment on whether or not this value represents the true percentage of truck drivers who suffer from the sleep disorder.

a.) 1-Prop Z interval on TI

Assume unbiased sample, have 10+ S/F ✓✓

$\hat{p} = \frac{124}{159} = 0.7799$

We're 90% confident that between 72.6% and 83.4% of all truckers have sleep apnea.

b) We'd test $H_0: p = 75\%$ vs.

$H_A: p \neq 75\%$ and

clearly reject H_0.

I would question the apnea testing methods! Are truckers really <u>that</u> different?

Example: One study of gambling newsletters that purport to improve a bettor's odds of winning bets on NFL football games indicates that the newsletters' betting schemes were not profitable. Suppose a random sample of 50 games is selected to test one gambling newsletter. Following the newsletter's recommendations, 30 of the 50 games produced winning wagers.

Do we have evidence that the newsletter can be said to significantly increase the odds of winning over what one could expect by selecting the winner at random? Interpret the *P*-value in context.

$H_0: p = 0.5$ (random picks should win 50% of the time)

$H_A: p > 0.5$ 1-Prop Z-TEST

$\hat{p} = \frac{30}{50} = 60\%$ and P-Value = 0.0786

If we picked our teams at random (say flip a coin) we'd get 30 right or more 7.86% of the time.

Our data is not super-convincing that the newsletter's picks do better, but there is some evidence they do.

Proceed with caution.

(25,600 customers week 1)
Joe, notes

Example: Census data for NYC indicate that 29.2% of the under-18 population is white, 28.2% black, 31.5% Latino, 9.1% Asian, and 2% other ethnicities.

The New York Civil Liberties Union points out that of 26,181 police officers, 64.8% are white, 14.5% are black, 19.1% are Hispanic, and 1.4% are Asian. Do the police officers reflect the ethnic composition of the city's youth? Test an appropriate hypothesis and state your conclusion.

	White	Black	Hispanic	Asian	Other
Observed:	16965	3796	5001	367	52
Expected:	7644.852	7383.042	8247.015	2382.471	523.62

CHI-SQUARE, GOODNESS OF FIT

H_0: NYPD distribution of cops matches the city's youth.

H_A: It doesn't.

df = 4, run on TI

$\chi^2 = 16512.75$

P-Value ≈ 0

Reject H_0, conclude distribution of cops doesn't match distribution of NY youth ethnicities.

(Does it have to?)