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 sense~~s~~ 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: $\widehat{Super} = 0.191 + 1.133(\text{Walmart})$

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 Supermarket (y) prices at the Supermarket 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 $S_y = 1.407$

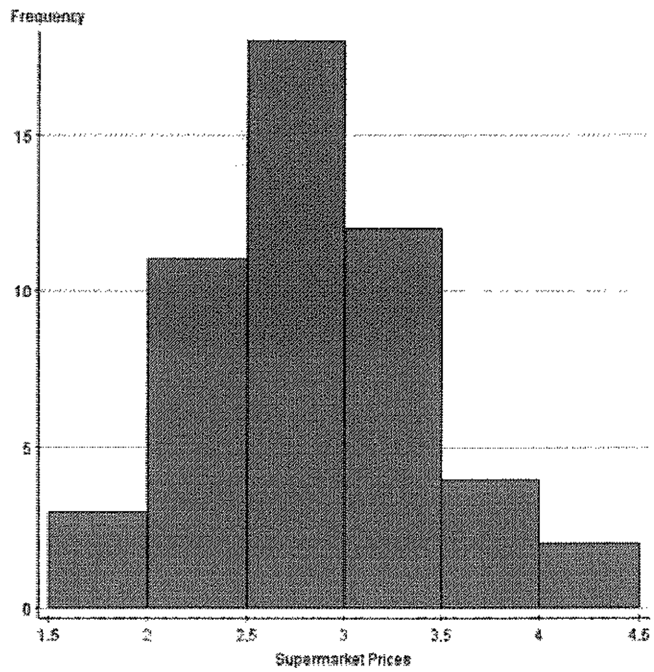
Stat II Time – Inference for Regression

Questions:

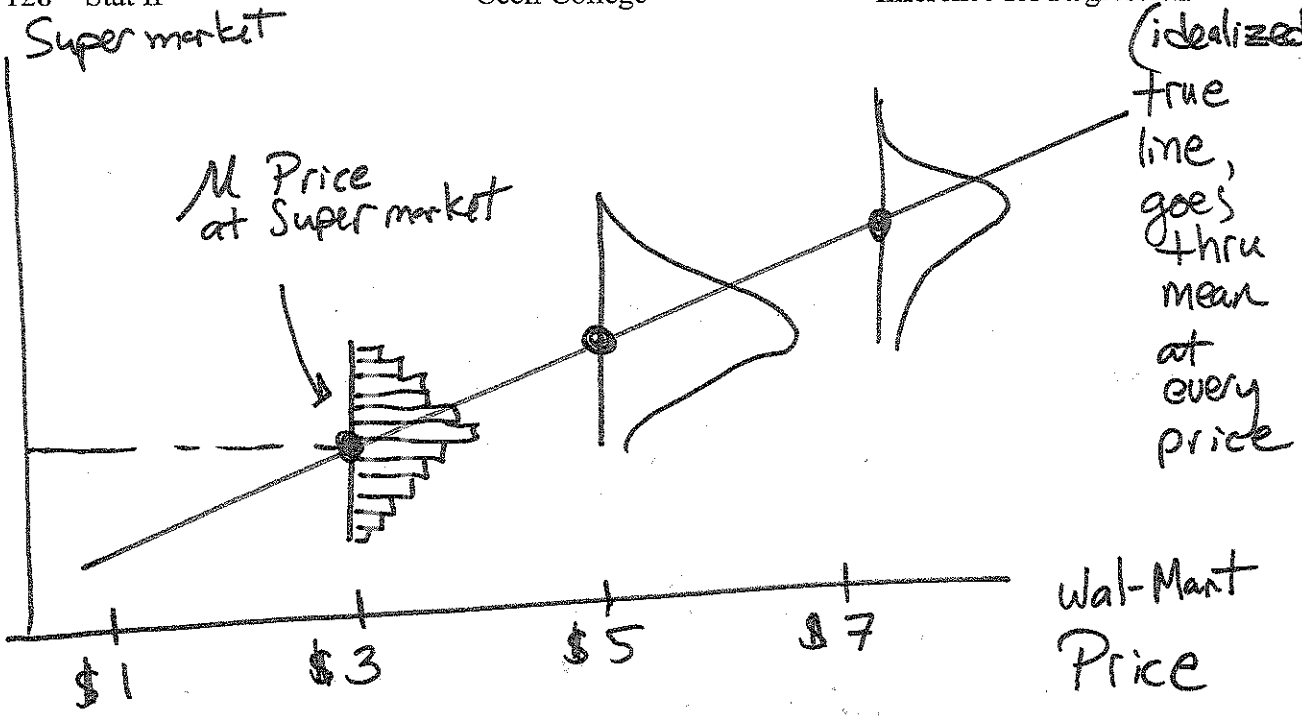
- Could the positive relationship we see between the Wal-Mart price and the supermarket price just be due to chance? Probably not.
- We can estimate the slope, but how reliable is our estimate? Probably good.

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 <u>model</u> the relationship between the two variables and we can imagine an <u>idealized</u> regression line.
- This idealized regression assumes the <u>mean</u> 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 poi~~nt, 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	$\neq 0$	93	1.2196608	0.2257
Slope	1.1329474	0.055957485	$\neq 0$	93	20.246574	<0.0001

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

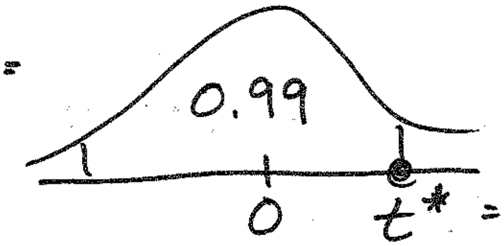
$$b_1 \pm t^* \times (SE(b_1))$$

$$df = n-2$$

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

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

$t^* =$



$t^* = 2.6297$

$SE(b_1) = 0.05596$

$1.133 \pm 2.6297(0.05596)$

$(0.986, \quad 1.280)$

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

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

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

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

$$0.00165 \pm 0.000289$$

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

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

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

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

---