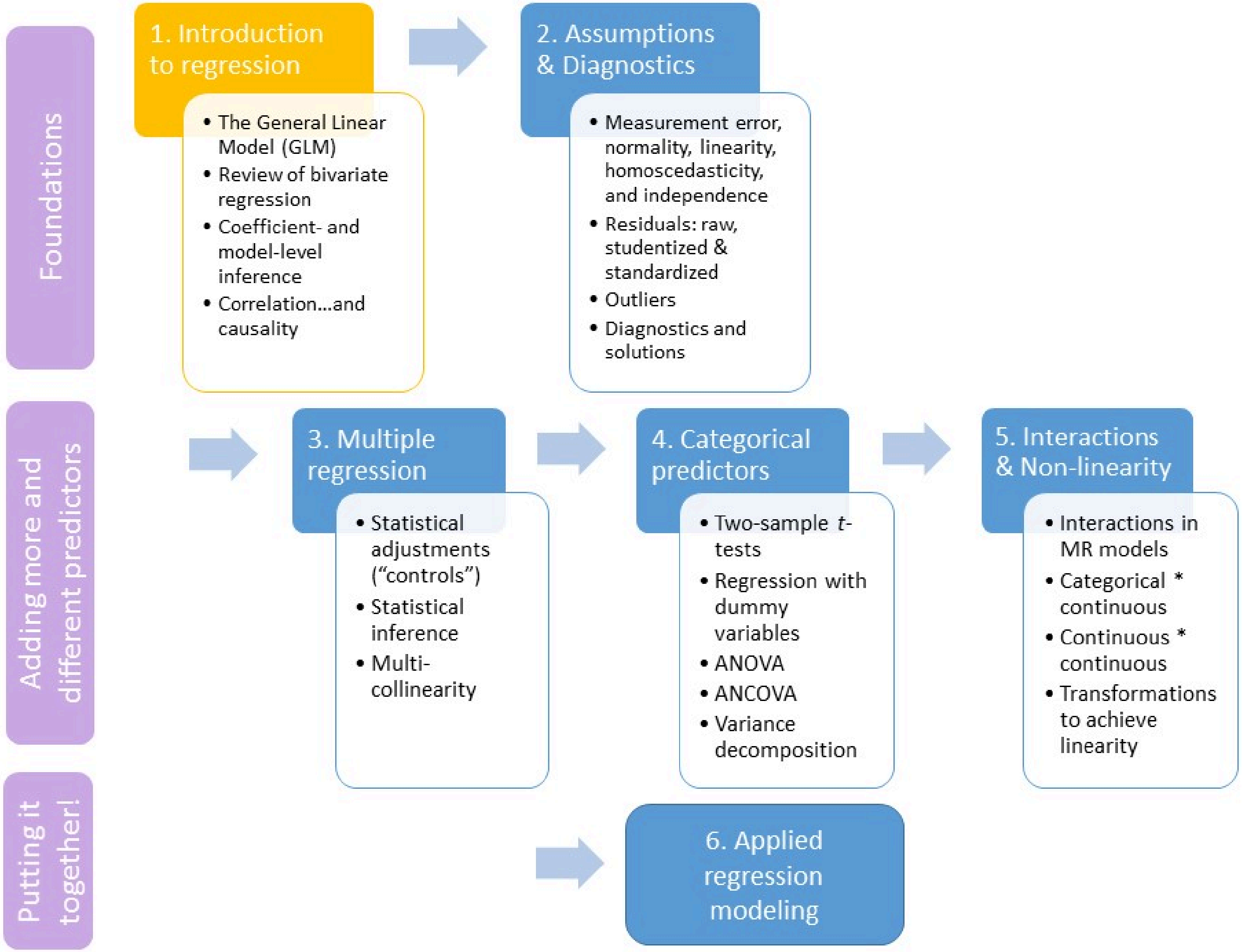
Regression inference

EDUC 643: Unit 1

David D. Liebowitz

UNIVERSITY OF OREGON

Roadmap



Foundations

1. Introduction to regression

- The General Linear Model (GLM)
- Review of bivariate regression
- Coefficient- and model-level inference
- Correlation...and causality

2. Assumptions & Diagnostics

- Measurement error, normality, linearity, homoscedasticity, and independence
- Residuals: raw, studentized & standardized
- Outliers
- Diagnostics and solutions

Adding more and different predictors

3. Multiple regression

- Statistical adjustments ("controls")
- Statistical inference
- Multi-collinearity

4. Categorical predictors

- Two-sample *t*-tests
- Regression with dummy variables
- ANOVA
- ANCOVA
- Variance decomposition

5. Interactions & Non-linearity

- Interactions in MR models
- Categorical * continuous
- Continuous * continuous
- Transformations to achieve linearity

Putting it together!

6. Applied regression modeling

Goals for the unit

- Characterize a bivariate relationship along five dimensions (direction, linearity, outliers, strength and magnitude)
- Describe how statistical models differ from deterministic models
- Mathematically represent the population model and interpret its deterministic and stochastic components
- Formulate a linear regression model to hypothesize a population relationship
- Describe residuals and how they can describe the degree of our OLS model fit

- Explain R^2, both in terms of what it tells us and what it does not
- Estimated a fitted regression line using Ordinary-Least Squares regression
- Conduct an inference test for a regression coefficient and our regression model

- Calculate a correlation coefficient (r) and describe its relationship to R^2
- Distinguish between research designs that permit correlational associations and those that permit causal inferences

A motivating question

Researchers (including two from the **University of Oregon**), Nichole Kelly, Elizabeth Cotter and Claire Guidinger (2018), set out to understand the extent to which young men who exhibit overeating behaviors have weight-related medical and psychological challenges.

Using real-world data (generously provided by Nichole Kelly) about the dietary habits, health, and self-appraisals of males 18-30, we are going to attempt to answer a similar question.

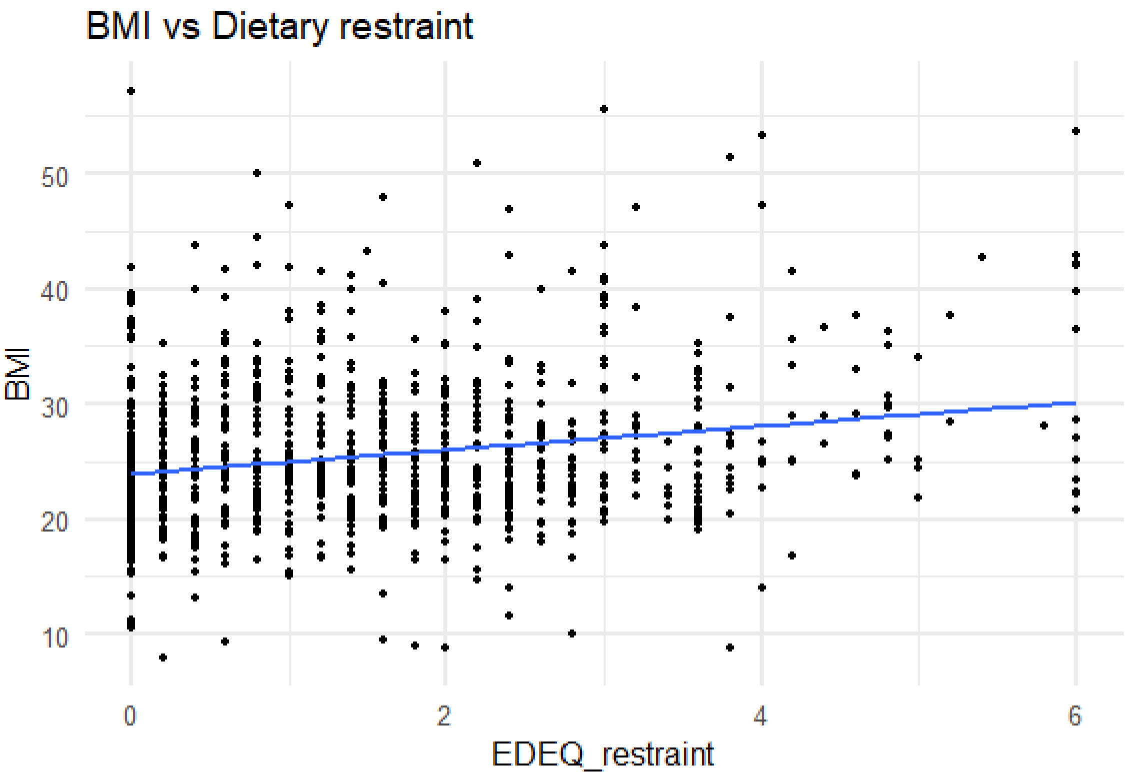
In particular, we are going to explore the **relationship** between **dietary restraint behaviors** (self-reports on the extent to which participants consciously restricted/controlled their food intake) and **over-eating frequency** (participants' self-reported frequency of over-eating episodes).

A preliminary analysis

Before we get to the core question of the Kelly et al. study--how are dietary restraint behaviors related to over-eating frequency?--we are going to explore another important relationship in the data that may also be related to our main research question: the **relationship** between dietary restraint behaviors (self-reports on the extent to which participants consciously restricted/controlled their food intake) and body-mass index (BMI). In particular, we are going to operationalize this by examining the relationship in our sample of young men between our predictor variable (***EDEQ_restraint***) and their body-mass index (***BMI***).

> We are examining this relationship so that we can better understand how all three of these variables (***OE_frequency***, ***EDEQ_restraint*** and ***BMI***) are related in Unit 3. Additionally, the properties of the variable ***BMI*** are pedagogically helpful in demonstrating the assumptions of OLS.
>
> However, we recognize that BMI has been shown to be relatively uninformative about individuals' overall health and categorizes individuals based on distributions initially derived exclusively from white Western European (French and Scottish) study participants. We use the measure for pedagogical purposes because the variable is one of the few continuous measures in one of the few datasets that our UO colleagues shared with us, while noting its problematic historical use.

Our bivariate relationship



Our linear model:

$$BMI_i = \beta_0 + \beta_1(EDEQ\_restraint_i) + \varepsilon_i$$

OLS regression

Fit a regression

Let's try fitting our regression in R.

```
fit <- lm(BMI ~ EDEQ_restraint, data=do)
summary(fit)
```

```
Call:
lm(formula = BMI ~ EDEQ_restraint, data = do)

Residuals:
    Min      1Q  Median      3Q     Max
-19.047  -3.955  -0.922   2.701  33.282

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     23.9223     0.2647  90.384  < 2e-16 ***
EDEQ_restraint   1.0367     0.1370   7.566 8.18e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.089 on 1083 degrees of freedom
Multiple R-squared:  0.05021,    Adjusted R-squared:  0.04933
F-statistic: 57.25 on 1 and 1083 DF,  p-value: 8.177e-14
```

Evaluating regressions: Coefficients

Here we can find our intercept and slope coefficients for our linear regression.

```
...
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     23.9223     0.2647  90.384  < 2e-16 ***
EDEQ_restraint   1.0367     0.1370   7.566 8.18e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.089 on 1083 degrees of freedom
Multiple R-squared:  0.05021,    Adjusted R-squared:  0.04933
F-statistic: 57.25 on 1 and 1083 DF,  p-value: 8.177e-14
...
```

The predicted BMI for a young male with a dietary restraint rating of 0 is 23.92 (β_0).

Evaluating regressions: Coefficients

Here we can find our intercept and slope coefficients for our linear regression.

```
...
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     23.9223     0.2647  90.384  < 2e-16 ***
EDEQ_restraint   1.0367     0.1370   7.566 8.18e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.089 on 1083 degrees of freedom
Multiple R-squared:  0.05021,    Adjusted R-squared:  0.04933
F-statistic: 57.25 on 1 and 1083 DF,  p-value: 8.177e-14
...
```

A one unit difference in dietary restraint is positively associated with a 1.04 difference in BMI.

Even better: **We reject the null hypothesis, and conclude that *on average, in the population* there is a relationship between dietary restraint and BMI. We estimate that young men who are one unit apart on dietary restraint index will have a BMI score 1.04 points different from each other.**

Why not just say "increase" or "decrease"? Be careful of causal language! More on this next class!

Evaluating regressions: Std errors

Standard errors represent how precisely we have estimated our regression coefficient, given our sample size, the quality of our model fit, and the variability in our predictor.

```
...
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     23.9223     0.2647  90.384  < 2e-16 ***
EDEQ_restraint   1.0367     0.1370   7.566 8.18e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.089 on 1083 degrees of freedom
Multiple R-squared:  0.05021,    Adjusted R-squared:  0.04933 
F-statistic: 57.25 on 1 and 1083 DF,  p-value: 8.177e-14
...
```

Not critical you understand the formula, but standard errors are important pieces of information that we'll examine in more detail momentarily:

$$SE_{\hat{\beta}_1} = \sqrt{\frac{1}{n-2} * \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(x_i - \bar{x})^2}}$$

Evaluating regressions: Model fit

```
...
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     23.9223     0.2647  90.384  < 2e-16 ***
EDEQ_restraint   1.0367     0.1370   7.566 8.18e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.089 on 1083 degrees of freedom
Multiple R-squared:  0.05021,    Adjusted R-squared:  0.04933
F-statistic: 57.25 on 1 and 1083 DF,  p-value: 8.177e-14
...
```

The residual standard error (RSE) is the standard deviation of the residuals. This summarizes the variability of observed values around the model-predicted values, in the original units of the outcome.

$$RSE = 6.089$$

This means observed values vary around our model-predicted BMI with a standard deviation of 6.089. In BMI, 6 units is quite large!

Degrees of freedom

Both RSE and the related measure of model fit, Root Mean Square Error (RMSE), depend on the number of degrees of freedom (df) in your regression. Though it's not critical that you learn how to calculate RSE or RMSE, it is important to understand that is is a function of the **degrees of freedom** (df) in your regression:

$$RSE = \sqrt{\frac{\text{sum of squared residuals}}{n - (\#\text{ of parameters estimated})_{SS}}}$$

Our degrees of freedom decrease each time we use another parameter (add a predictor to our regression) to calculate the sum of squares. In a bivariate regression, our degrees of freedom (aka, the denominator) will always be $n - 2$ because we are estimating two parameters β_0 and β_1. With smaller samples and many covariates, we may quickly use up our degrees of freedom.

What happens to our model's precision as our degrees of freedom decrease?

Use the stored model statistics to calculate RSE by hand:

```
sqrt(sum(fit$residuals^2) / fit$df)
```

```
## [1] 6.088998
```

Evaluating regression: R^2

Here is our summary of model performance.

```
...
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     23.9223     0.2647  90.384  < 2e-16 ***
EDEQ_restraint   1.0367     0.1370   7.566 8.18e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.089 on 1083 degrees of freedom
Multiple R-squared:  0.05021,    Adjusted R-squared:  0.04933
F-statistic: 57.25 on 1 and 1083 DF,  p-value: 8.177e-14
...
```

The R-squared value is .05. This means that our model accounts for 5% of the variance in BMI. Since our model has only one predictor, we can alternatively say Dietary Restraint accounts for 5% of the variance in BMI.

- The rest? Measurement error, random individual variation, other unobserved causes

What does R^2 mean?

R^2 describes what proportion of the variation in the outcome the full regression model has explained.

Whether or not your model has a high or low R^2 is:

- Disciplinary dependent
- Entirely independent from whether or not your model accurately characterizes the relationship

R^2 does **NOT**:

- Imply anything about causality
- Tell us anything about whether there exists a linear or non-linear relationship (more on this soon)
- Tell us anything about the magnitude (steepness/shallowness) of the slope

A model can have a low R^2, and yet your estimated coefficient of interest can meaningfully predict variation in your outcome.

Regression inference

A review of inference

So far, we've been evaluating the statistics generated by models fit on our sample, but remember our primary interest is in making **an inference from the sample to the population.**

Go back to EDUC 641 for a refresher on Null-Hypothesis Significance Testing (NHST), the Central Limit Theorem and t-distributions.

...but we're going to provide a quick review now before applying these concepts to understanding standard errors and using them to construct confidence intervals (CI).

Basic review of NHST

Start by imagining a hypothetical world in which there is **no relationship** between x and y in our population.

Last term, we formulated a set of null hypotheses about relationship between categorical data (e.g., no relationship between race of victim and death penalty sentence) or about whether a sample mean was different from a "known" population mean (e.g., life expectancy in Mediterranean countries is the same as in the full population of countries).

We can use the same framework to test elements of the General Linear Model and, in particular, regression coefficients. In particular, here we will test the null hypothesis that $\beta_1 = 0$ (there is no relationship between dietary restraint behaviors and BMI).

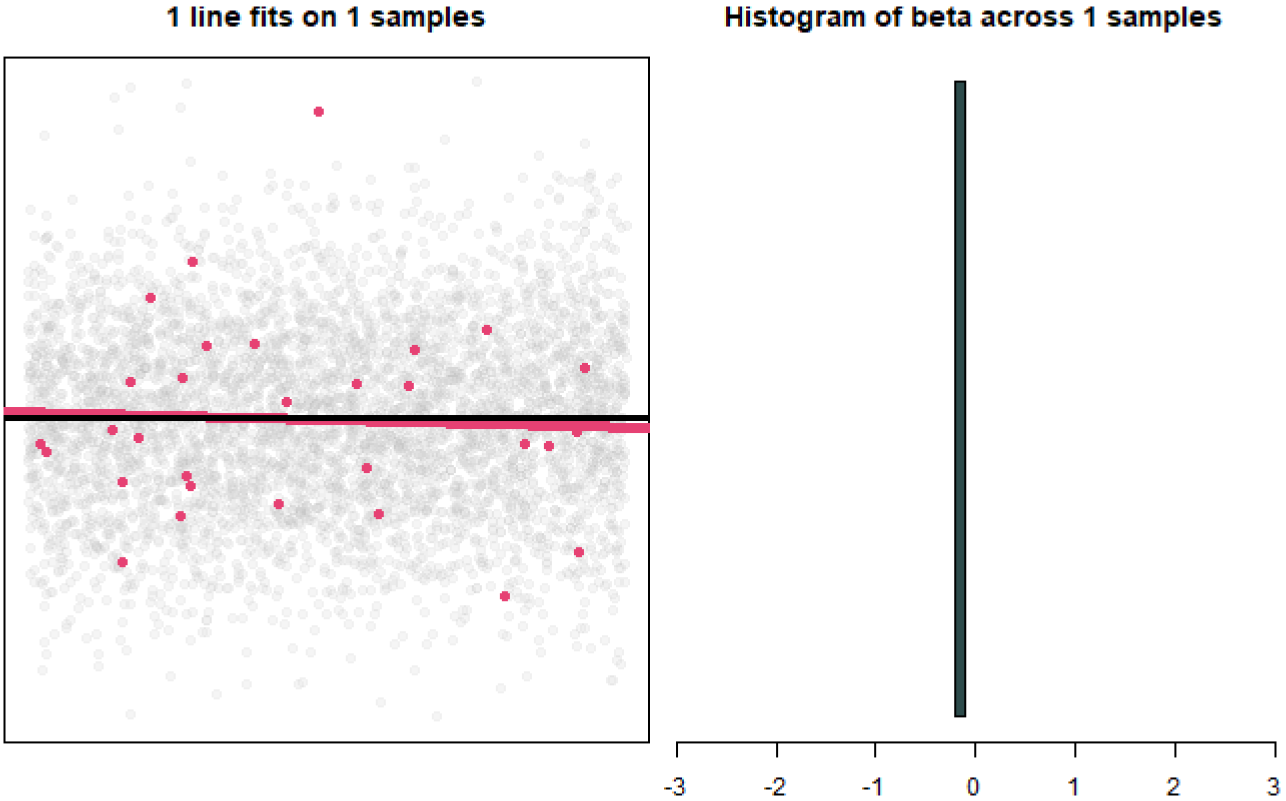
$$H_0 : \beta_1 = 0$$

Our alternative hypothesis is that there is a relationship:

$$H_1 \mathrm{or} H_A : \beta_1 \neq 0$$

The beauty of the CLT

We can imagine drawing samples over and over again (say...5,000 times) from this hypothetical null population. What values of $\hat{\beta}_1$ might we observe?



h/t Simon Hess (@simonhhess)

The beauty of the CLT

- Random sampling from the population **will return** sample means that will be asymptotically (approaching) normal in their distribution as the number of samples approaches infinity (i.e., they will take on a Z-distribution). This is (mathematically) what the Central Limit Theorem demonstrates.

- Because of that mathematical fact, we can conduct inference in the statistics.

- We know what the distribution of sample means from a null hypothesis will look like, so we can determine the probability of observing a sample statistic as extreme as we did

- REMEMBER: this requires no assumptions about the shape of the sample variable(s); they do **NOT** have to be normally distributed for the CLT to hold

- ...but, how do we know how likely it is to have observed such an extreme sample statistic in the presence of a null relationship in the population?

p–values

The statistic that captures the likelihood that one would observe a value of $\hat{\beta}_1$ of a given magnitude in a particular sample, in the presence of a null population, is called the *p*–value.

Prior to interacting with our data, we set an alpha threshold; a probability threshold, below which we will consider p to be so small that it is unlikely that we would have gotten this result if the null were true, and we will reject the null hypothesis. Above this value, we will fail to reject the null.

In social science research, it is customary to (arbitrarily) set that threshold at **5 percent** $(p < 0.05)$.[1] In other words, we say that if the difference between our observed data and our expected data would have happened in fewer than 1 out of 20 randomly drawn samples, that the difference reflects a true difference in the population.

[1] Reminder that in some disciplines, the convention is to set alpha thresholds at 0.01 or 0.1, highlighting that these are subjective and arbitrary.

What are we testing?

There are multiple inferential tests in a regression model:

- Tests of the coefficients
- Test of the model (omnibus test)

```
...
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     23.9223     0.2647  90.384  < 2e-16 ***
EDEQ_restraint   1.0367     0.1370   7.566 8.18e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.089 on 1083 degrees of freedom
Multiple R-squared:  0.05021,    Adjusted R-squared:  0.04933 
F-statistic: 57.25 on 1 and 1083 DF,  p-value: 8.177e-14
...
```

Here, the p-value refers to the probability of obtaining a slope equal to or more extreme than β assuming H_0 is true. But where do we get that p-value from?

Student's t-distribution

- William Sealy Gosset, then a Head Experimental Brewer at Guinness Beer, wrote a pseudonymously published article in 1908 showing that estimates of $(\hat{\beta}_1)$ divided by their standard error form a defined distribution
- This distribution is now known as Student's t-distribution
- Why *Student's* t-distribtion? Gosset's pseudonym was "Student"

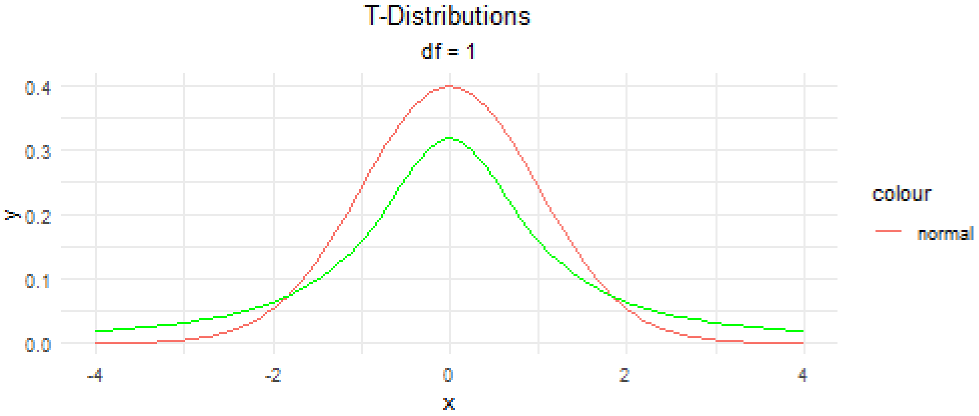
The t-statistic represents an estimate of how many standard errors $\hat{\beta}_1$ lies away from 0 in the t-distribution.

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$$

- When we posit a null hypothesis, we assume that $\beta_1 = 0$

t-distributions

- The degrees of freedom for our t-statistic is always *n*-1, where *n* is our sample size
- t-distributions with fewer degrees of freedom have "fatter" tails
- As the degrees of freedom get larger, the t-distribution approaches a standard normal distribution



How large is enough?

Generally, as we get to around 50 degrees of freedom, our t-distribution approaches a standard normal distribution, and our inferences are straightforward because the p-values for our t-test are the same as p-values are in the standard normal distribution.

Critical values of t_{observed}

df	**0.10**	**0.05**	**0.01**
10	1.81	2.23	3.17
20	1.72	2.09	2.85
30	1.70	2.04	2.75
50	1.68	2.01	2.68
100	1.66	1.98	2.63
$\to \infty$	**1.64**	**1.96**	**2.58**

Regression coefficients: $\hat{\beta}_0$

Our output shows us the significance of the intercept (typically, we are not interested in whether the intercept differs from 0).

```
...
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     23.9223     0.2647  90.384  < 2e-16 ***
EDEQ_restraint   1.0367     0.1370   7.566 8.18e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.089 on 1083 degrees of freedom
Multiple R-squared:  0.05021,	Adjusted R-squared:  0.04933
F-statistic: 57.25 on 1 and 1083 DF,  p-value: 8.177e-14
...
```

Regression coefficients: $\hat{\beta}_1$

```
...
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     23.9223     0.2647  90.384  < 2e-16 ***
EDEQ_restraint   1.0367     0.1370   7.566 8.18e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.089 on 1083 degrees of freedom
Multiple R-squared:  0.05021,    Adjusted R-squared:  0.04933 
F-statistic: 57.25 on 1 and 1083 DF,  p-value: 8.177e-14
...
```

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{1.0367}{0.137} = 7.57$$

$$Pr(t < -7.57 \text{ or } t > 7.57)|H_0 = 0.000000000000082 \text{ or } p < 0.001$$

Regression coefficients

```
...
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     23.9223     0.2647  90.384  < 2e-16 ***
EDEQ_restraint   1.0367     0.1370   7.566 8.18e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.089 on 1083 degrees of freedom
Multiple R-squared:  0.05021,    Adjusted R-squared:  0.04933 
F-statistic: 57.25 on 1 and 1083 DF,  p-value: 8.177e-14
...
```

Under the null hypothesis, it is extremely unlikely to obtain a Dietary Restraint slope of 1.04 ($p < .001$). Therefore, we can reject the null hypothesis and conclude that there is a positive relationship between Dietary Restraint rating and BMI, on average in the population.

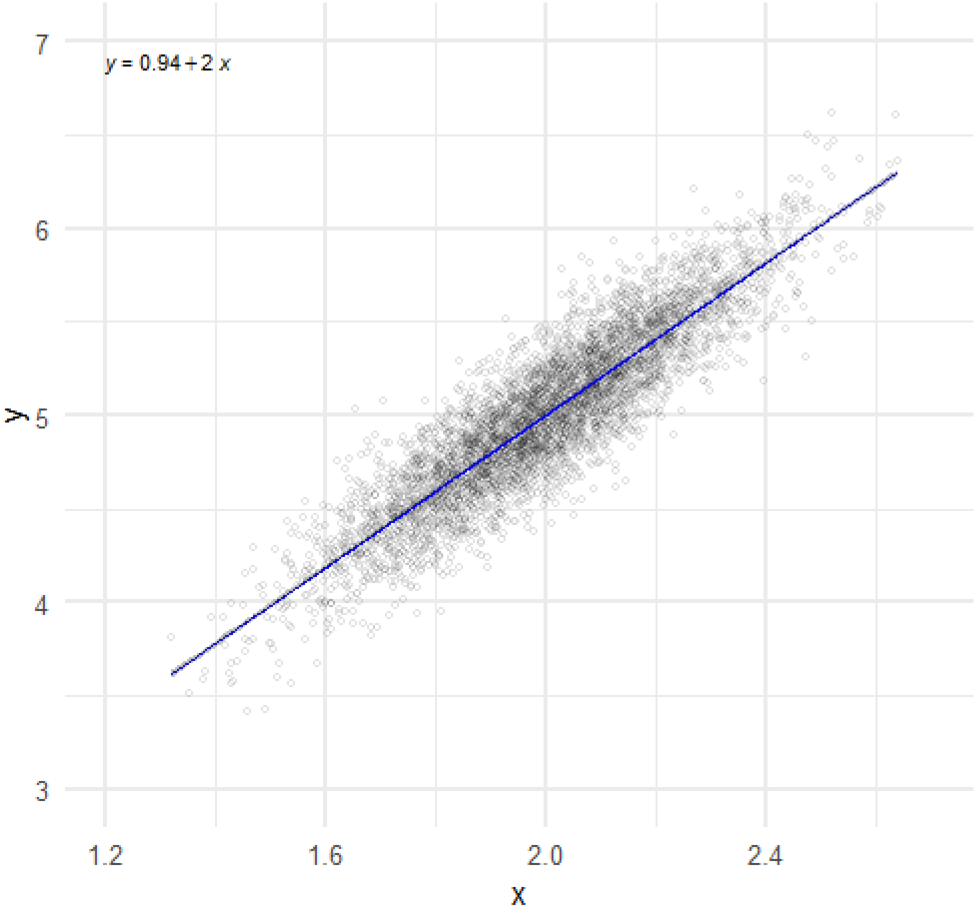
Confidence intervals (CIs)

It would be nice to say something more concrete about how accurately we have estimated this relationship.

Perhaps, we can identify a range of plausible values for $\hat{\beta}_1$. We may be able to use information about the quality of our model fit and the variability in our to construct intervals that offer a range of plausible values for the population parameter.

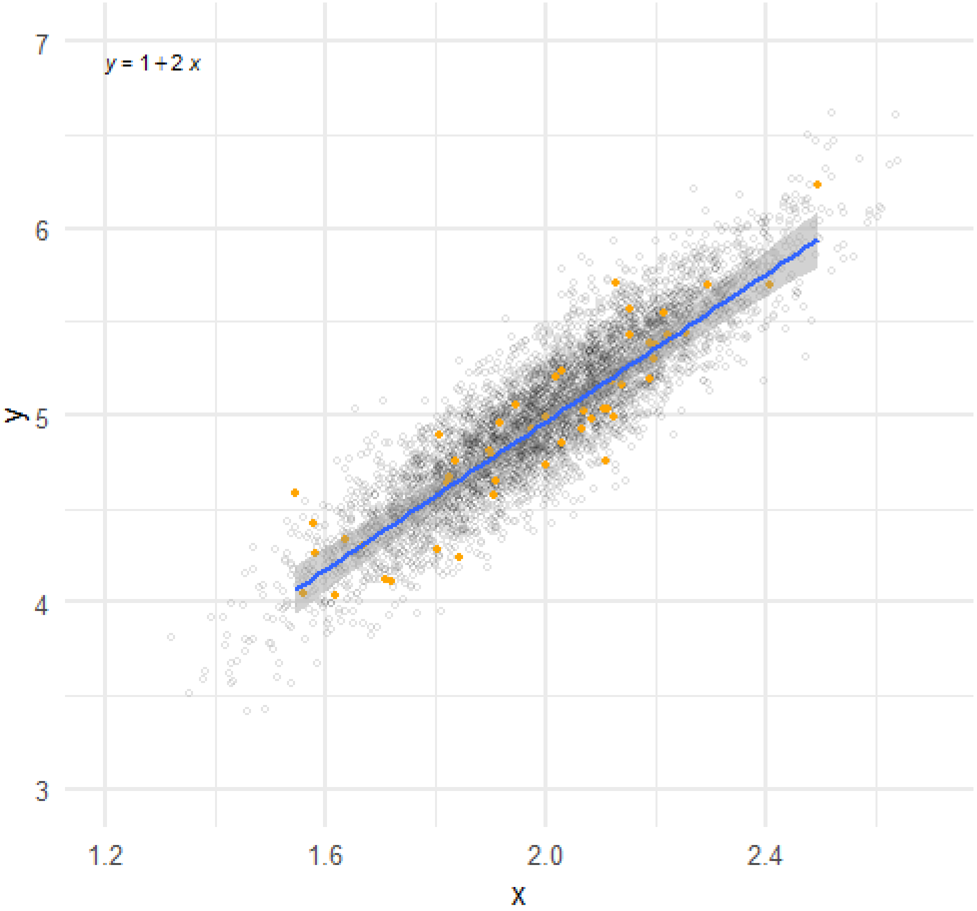
Let's develop some intuition around the idea of confidence intervals (CIs).

Confidence interval intuition



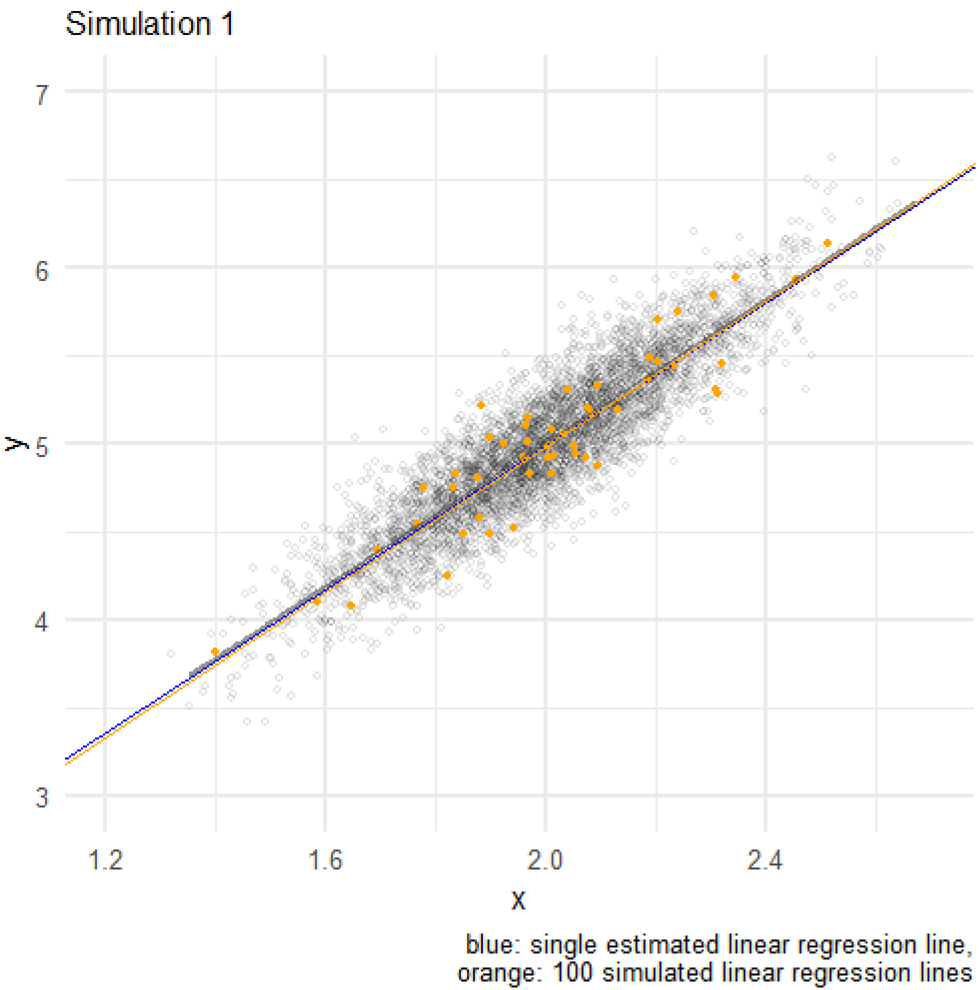
blue: hypothetical linear regression line on full population,
gray: 95% CI

Confidence interval intuition



blue: linear regression line on random sample of 50 observations,
gray: 95% CI on random sample estimate

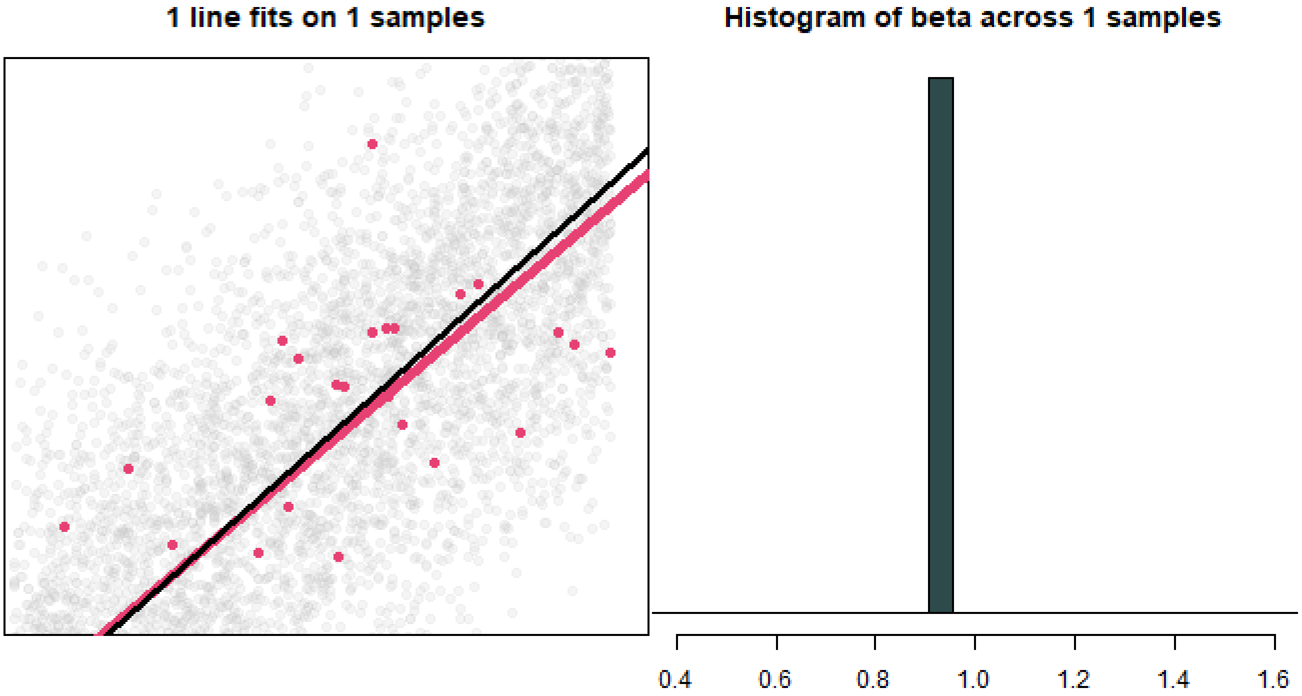
Confidence interval intuition



blue: single estimated linear regression line,
orange: 100 simulated linear regression lines

Slope ($\hat{\beta}_1$) sampling distribution

We can use this same approach to construct a distribution of estimated slopes $\hat{\beta}_1$. Here, we imagine the "true" population slope is 1.



h/t Simon Hess (@simonhhess)

Formal confidence intervals

The intuition to take from the preceding simulation is that the confidence interval corresponds to the frequency with which the underlying "true" population parameter will fall within the range of the confidence interval over repeated sampling from that population.

For a given α-threshold, we are positing that 90 or 95, 99 or 99.9 percent of future confidence intervals drawn from repeated samples will encompass the true value of the population parameter.

Colloquially: "If we drew a sample over and over again and computed a 95% confidence interval for each replication, then 95% of those intervals would contain the true mean."

NOT: "there is a 95% probability that the true mean lies within the confidence interval"

Confidence interval for $\hat{\beta}_1$:

$$\hat{\beta}_1 \pm t_{n-2}^{critical}[se(\hat{\beta}_1)]$$

Confidence intervals (CIs)

```
...
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     23.9223     0.2647  90.384  < 2e-16 ***
EDEQ_restraint   1.0367     0.1370   7.566 8.18e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.089 on 1083 degrees of freedom
Multiple R-squared:  0.05021,    Adjusted R-squared:  0.04933
F-statistic: 57.25 on 1 and 1083 DF,  p-value: 8.177e-14
...
```

95% CIs:

$$\hat{\beta}_1 \pm t_{n-2}[se(\hat{\beta}_1)]$$

$$1.037 \pm 1.96(0.137)$$

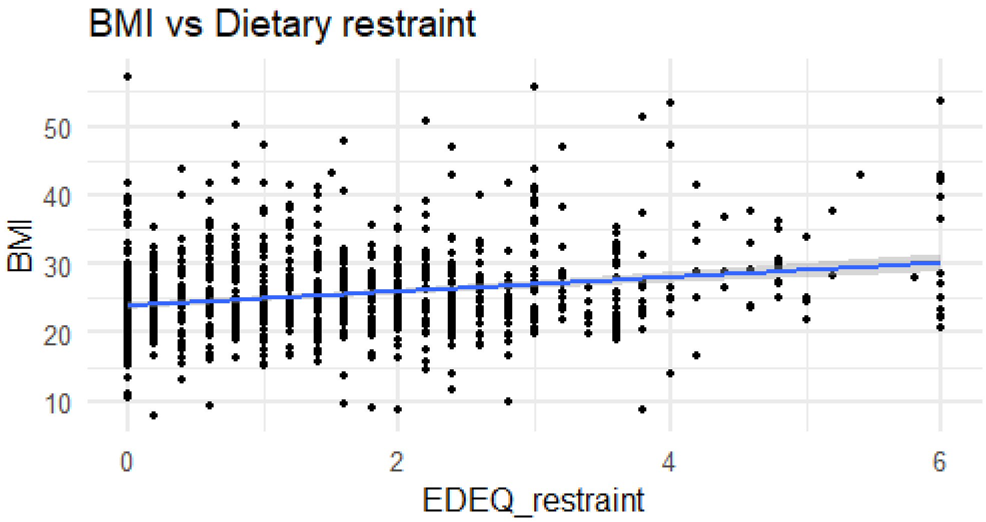
$$[0.768, 1.306]$$

Confidence intervals (CIs)

Let's do the same with R

```
tidy(fit, conf.int=T)
```

```
## # A tibble: 2 x 7
##   term           estimate std.error statistic  p.value conf.low conf.high
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)       23.9      0.265     90.4  0           23.4      24.4
## 2 EDEQ_restraint     1.04     0.137      7.57 8.18e-14    0.768      1.31
```



F-Distributions and omnibus tests

The omnibus test uses the F-distribution to test the ratio of two variances (i.e., explained vs unexplained variance). This is a different, but similar, distribution to the t-distribution. For now, it's not critical that you know how it differs.

Null hypothesis: The model does not account for any variance in Y.

If we reject the null, then the model accounts for more variance than we would expect by chance.

F-Distributions and omnibus tests

Just like tests with other probability distribution, we are testing the probability of obtaining a value (or more extreme value) of F under the F-Distribution.

$$F = \frac{MS_{model}}{MS_{residual}}$$

Mean Squares (MS) are the Sums of Squares divided by their respective degrees of freedom.

Interpreting the omnibus test

```
...
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     23.9223     0.2647  90.384  < 2e-16 ***
EDEQ_restraint   1.0367     0.1370   7.566 8.18e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.089 on 1083 degrees of freedom
Multiple R-squared:  0.05021,    Adjusted R-squared:  0.04933 
F-statistic: 57.25 on 1 and 1083 DF,  p-value: 8.177e-14
...
```

Our omnibus test is significant at an α-threshold of 0.05 $(p < 0.001)$, therefore we can reject the null hypothesis that the full model accounts for no variability in individuals' BMI.

Summarizing regression results

```r
modelsummary(fit, stars=T,
  gof_omit = "Adj.|AIC|BIC|Log",
  coef_rename = c("EDEQ_restraint" = "Dietary Restraint Index (0-6)"),
  escape=F) # <- necessary for html, but don't need this for Word
```

	(1)
(Intercept)	23.922***
	(0.265)
Dietary Restraint Index (0-6)	1.037***
	(0.137)
Num.Obs.	1085
R2	0.050
RMSE	6.08

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Summarizing regression results

	(1)
(Intercept)	23.922***
	(0.265)
Dietary Restraint Index (0-6)	1.037***
	(0.137)
Num.Obs.	1085
R2	0.050
RMSE	6.08
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001	

We postulated a linear model which we estimated via Ordinary-Least Squares regression to assess whether there is a relationship between BMI and Dietary Restraint, on average, in the population of young adult males. At an alpha threshold of 0.05, we found that Dietary Restraint was a significant predictor of BMI and accounted for approximately 5 percent of the variance in BMI. We estimate that young men who are one unit apart on a dietary restraint index will have a BMI score 1.04 (*p*<0.001, 95% CI: 0.77, 1.31) points different from each other.

Synthesis and wrap-up

Putting it all together…so far

- **Understand your data first**
  - Summarize and visualize each variable independently
  - Start with a visual representation of the relationship between your two variables
- **Your General Linear Model represents your hypothesis about the population**
  - When you fit a regression model, you are estimating *sample* values of *population* parameters that you will not directly observe
  - The goal of classical regression inference is to understand how likely the observed data in your sample are in the presence of no relationship in the unobserved population
  - State a null hypothesis
  - Establish an alpha threshold
- **Use Ordinary Least Squares regression to estimate the relationship**
  - Interpret coefficients, standard errors, R^2
  - *Assess assumptions*
  - Conduct an inference test
  - Reject (or fail to reject) the null
  - Substantively interpret

Class goals

- Formulate a linear regression model to hypothesize a population relationship
- Explain R^2, both in terms of what it tells us and what it does not
- Estimated a fitted regression line using Ordinary-Least Squares regression
- Conduct an inference test for a regression coefficient and our regression model

To-Dos

Reading:

- **If you have not already**: LSWR Chapter 15.1 - 15.2 and 15.4 - 15.7 and Hu (2021)
- **By January 21 class**: LSWR Chapter 5.7

Quiz:

- Opens Jan. 21st

Assignment 1:

- Now live
- Have all information you need to complete parts 1 & 2
- Due Feb. 3