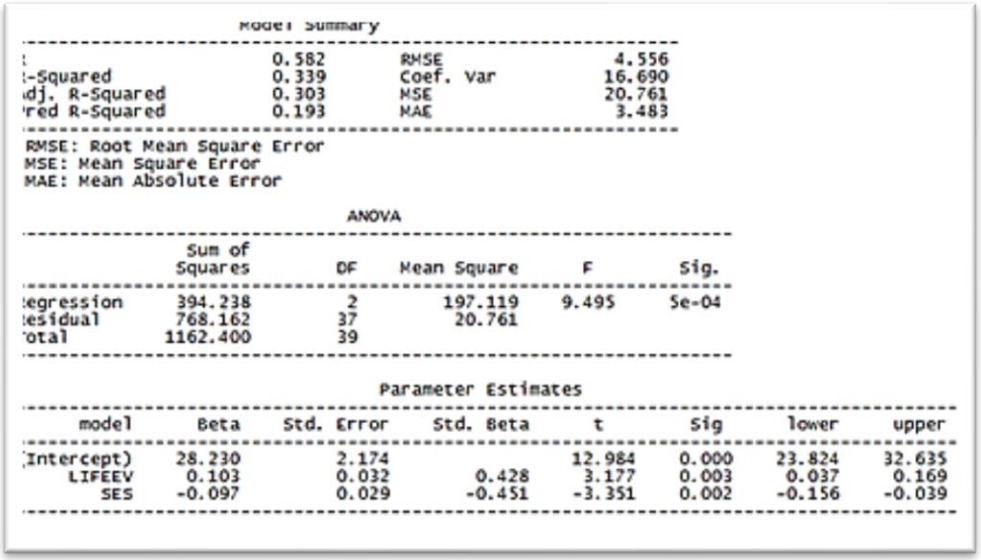
Stats Midterm Review

Review of Main Concepts:

Example Data: MHDATA
This data is from a study in Alachua County, FL that examines mental health and several predictor variables

- **DV:** Mental Impairment (MI): includes many dimensions of psychiatric symptoms (e.g., depression, anxiety). Higher scores indicate more impairment.
- **IV:** Life Events Score (LIFEEV): composite measure of the number and severity of major life events (e.g. death, affair, new job) experienced during the past 3 years. Higher scores indicate more severe life events.
- **IV:** Socioeconomic Status (SES): composite index of occupation, income, and education. Higher scores indicate more affluent status.



Model Summary

	0.582	RMSE	4.556
-Squared	0.339	Coef. Var	16.690
dj. R-Squared	0.303	MSE	20.761
red R-Squared	0.193	MAE	3.483

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
egression	394.238	2	197.119	9.495	5e-04
esidual	768.162	37	20.761		
otal	1162.400	39			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	28.230	2.174		12.984	0.000	23.824	32.635
LIFEEV	0.103	0.032	0.428	3.177	0.003	0.037	0.169
SES	-0.097	0.029	-0.451	-3.351	0.002	-0.156	-0.039

R
Correlation between predicted (y hat) and observed (y i) scores

---

R^2
$R^2 = SS_{regression} / SS_{total}$
Example for our data: Roughly 44% of the variability in MI can be predicted by both SES and LifeEvents (though individual contribution varies)

Adjusted R^2

Corrects for a problem with regular R2= that is R2 the denominator (SST) is fixed (unchanging) and the numerator (SSR) can ONLY increase. Therefore, each additional variable used in the equation will, at least, not decrease the numerator and will probably increase the numerator (at least to a small extent), resulting in a higher R2 even when/if the added variable(s) add nothing to the model. With adjusted R2, Adj R2 can decline in value of the contribution of the explained variance by the additional variables is less than the impact on the degrees of freedom. In other words, adjusted R2 rewards parsimony.

$$Adj\_R^2 = 1 - \left((1 - R^2) \frac{(n-1)}{(n-k-1)} \right)$$

Where n is the sample size and k is the number of predictors.

Predicted R^2

Predicted R_2 also attempts to prevent overfitting. It indicates how well a regression model predicts responses for new observations. It is calculated by systematically removing each observation from the data set, estimating the regression equation, and determining how well the model predicts the removed observation. Both adjusted and predicted R_2 can be negative and are always smaller than R_2.

A key benefit of adjusted and predicted R-squared is that it can prevent you from overfitting a model.

Sum of Squares values (in ANOVA table)

SS_{total}
Represents the total variability present in the model
$SS_{total} = SS_{regression} + SS_{residual}$

$SS_{regression}$
Represents the variability in Y that can be accounted for by the model
$SS_{regression} = SS_{total} - SS_{residual}$

$SS_{residual}$
Represents the variability in Y that is attributed to error
$SS_{residual} = SS_{total} - SS_{regression}$

Model Summary

t	0.582	RMSE	4.556
t-Squared	0.339	Coef. Var	16.690
idj. R-Squared	0.303	MSE	20.761
'red R-Squared	0.193	MAE	3.483

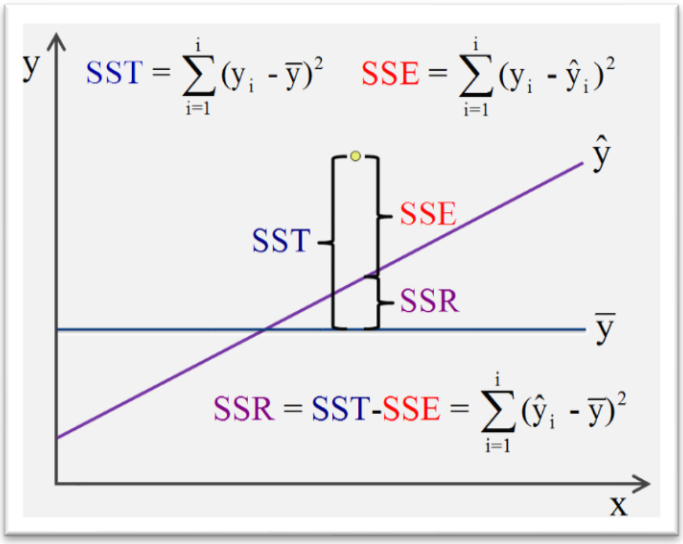
RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
tegression	394.238	2	197.119	9.495	5e-04
tesidual	768.162	37	20.761		
rotal	1162.400	39			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	28.230	2.174		12.984	0.000	23.824	32.635
LIFEEV	0.103	0.032	0.428	3.177	0.003	0.037	0.169
SES	-0.097	0.029	-0.451	-3.351	0.002	-0.156	-0.039



Mean Squares (ANOVA table):

MSregression

Represents the amount of variability in Y explained by the model adjusted for degrees of freedom.

MSregression = SSregression / df

MSresidual

Represents the amount of variability in Y attributed to error, adjusted for degrees of freedom.

Msresidual = SSresidual / df

Degrees of Freedom
Regression: *k* (number of predictors)
Residual: *n – k – 1*
Total: *n – 1*

F statistic
The ratio of explained variance to unexplained variance
$F = MS_{regression} / MS_{residual}$
If F is significant, then we reject the null hypothesis that all the regression coefficients are equal to zero.
F is an omnibus test. It doesn't tell us anything about individual predictors, only the model as a whole.

Beta

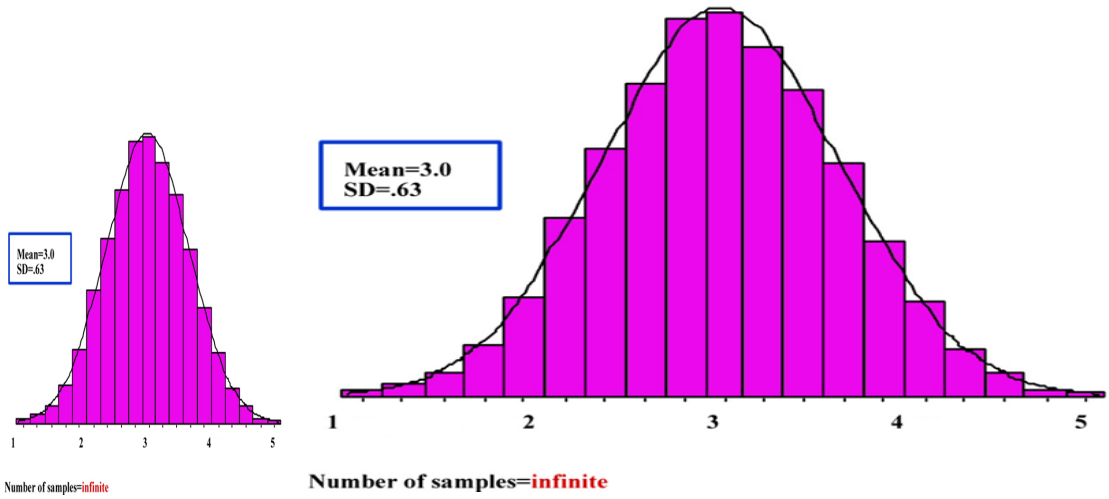
Intercept= Y-intercept of the linear model. The value of Y when all predictors equal zero.
Example for our data: predicted MI score when life events and SES are average (OR 0 if the mean is centered at zero)

Predictors= regression coefficients
SLR: The expected change in Y for every one unit increase in x.
Examlple of our data: If we were only looking at SES and MI, we would interpret that for every one unit increase in SES, we would predict a change in MI score of -.097
MLR: The expected change in Y for every one unit increase in x *holding all other predictors constant.*
Beta is used in the calculation of *t** and *p*

Example of Beta Interpretation for MLR: The estimate for the slope of LIFEEV indicates that, <u>while holding SES constant</u>, every one unit increase in LIFEEV results in a 0.103 unit increase in predicted MI score. The p value is less than alpha; therefore this estimate is statistically significant. In other words, there is less than a 5% chance that the estimated slope (or a more extreme slope) would be obtained if the null hypothesis was true. Therefore, we reject the null hypothesis that the slope is equal to zero.

Standard Error:
Sampling Distribution
Fit the same model in a large number of random samples and plot a histogram for the estimates for beta.
Standard Error
The standard deviation of the frequency distribution for beta

Standard error is used in the calculation of t* and p



Standard Beta
Estimates for Beta with all variables express as Z-scores.

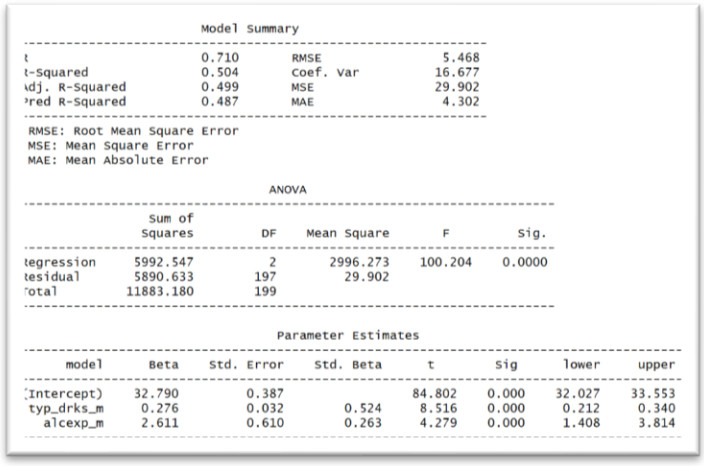
For MLR: the expected change in standard deviations for Y, given a one standard deviation increase in X *while holding all other predictors constant.*
For SLR: the expected change in standard deviations for Y, given a one standard deviation increase in X

For SLR, Std. Beta = Pearson's R
|Std. Beta| = R from model summary

T-statistic
This is *t**, the test statistic for each parameter estimate. Essentially, a ratio of systematic variance to unsystematic variance
T* is the number of standard errors away from the null.
*t** = beta / standard error
t* is compared to t_{crit}
If t* is larger, then we reject the null hypothesis that beta = 0

t_{crit} represents the largest t value we would expect to find 95% of the time <u>if the null hypothesis were true</u>.
t_{crit} changes depending on the degrees of freedom. Chart is in the appendix of the book.

Sig. = P-value: the probability of obtaining the observed t* (or one that is more extreme) if the null hypothesis were true.



Model Summary

t	0.710	RMSE	5.468
t-Squared	0.504	Coef. Var	16.677
djj. R-Squared	0.499	MSE	29.902
red R-Squared	0.487	MAE	4.302

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
tegression	5992.547	2	2996.273	100.204	0.0000
tesidual	5890.633	197	29.902		
otal	11883.180	199			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	32.790	0.387		84.802	0.000	32.027	33.553
typ_drks_m	0.276	0.032	0.524	8.516	0.000	0.212	0.340
alcexp_m	2.611	0.610	0.263	4.279	0.000	1.408	3.814

Example from Kim's Notes: The null hypothesis for the intercept in this example isn't overly interesting so we will ignore it. The t* for each slope is statistically significant, indicating that the independent effect of each, adjusting or controlling for the other, is unlikely to be 0 in the population. Therefore, we reject the null hypothesis for each slope. Just as was the case in SLR, the 95% CI gives use a range of plausible values for each parameter estimate and demonstrates the precision of these estimates.

Confidence Intervals

An interval that contains 95% of the sampling distribution of the parameter estimate.
If we estimate the same model in many random samples, the parameter estimate in 95% of the models will fall within the confidence interval
The confidence interval gives an indication of the range of feasible values for the parameter estimate.

In order to calculate the confidence interval for an estimate (for example a regression slope), you should do the following:

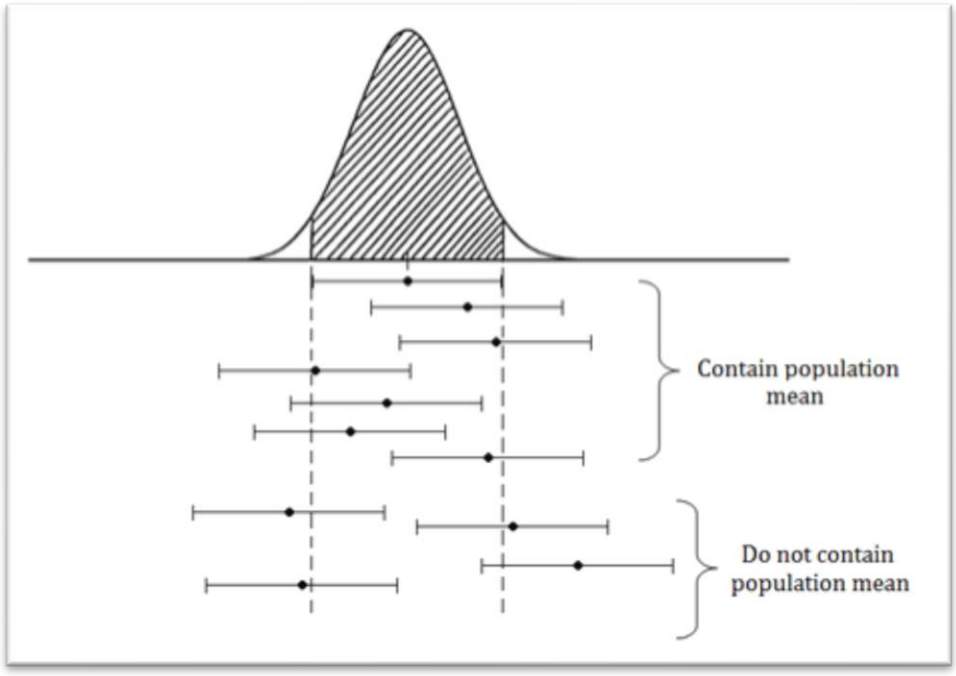
You start by finding the critical t, this is the same critical t that you use to compare each t* in the parameter estimates table (to determine statistical significance) and is obtained by choosing your desired alpha (.05, for example), the number of tails for the test (2 for a non-directional hypothesis – which is what we focus on in class), and the degrees of freedom for the model (n – 1 - # of predictors). Because it is a two-sided test, we divide alpha by 2, and put .025 in each tail (so the critical t is t-value corresponding to the 2.5 and 97.5 percentile of the student's t distribution). The critical t is calculated using the qt function in R (or you can find it in the back of a heavy statistics book).

Next, to get the corresponding confidence interval you take the estimate plus/minus the critical t times the standard error.

As an example, let's imagine that you fit a simple linear regression in a sample of size 200, you set alpha to .05, and want to test a non-directional (2-sided) test, then the df are 198. You can type qt(c(.025, .975, df = 198)) into R and it will return a critical t of 1.97. Imaging the slope is 1.935, and a standard error of .148. The 95% CI is then calculated as:

Lower: 1.935 + 1.97*.148 = 1.64

Upper: 1.935 – 1.97*.148 = 2.23



CI = Beta +/- t_{crit} * s.e.

Standard error describes the sampling distribution.

Calculating a confidence interval centers the sampling distribution at the estimate for beta, then finds the range of values that contain 95% of that distribution

If the C.I. does not contain zero, then the parameter estimate is statistically significant.

Partial F Test

Adding additional predictors to a regression model will always reduce $SS_{residual}$ (i.e. will explain more variance)
However, we must perform a test to determine if the improvement is statistically significant.
The Partial F Test compares two nested models and tests the null hypothesis that the reduction in $SS_{residual}$ resulting from the additional predictors in the more complex model is equal to zero.

$$Partial\ F^* = \frac{\frac{SSE_{\text{Reduced}} - SSE_{Full}}{df_{\text{Reduced}} - df_{Full}}}{\frac{SSE_{Full}}{df_{Full}}}$$

Example: For our research question, we are interested in predicting job satisfaction, which leads to a variety of positive work-related outcomes (such as higher commitment to an organization, lower counter-productive behaviors, and lower turnover). With so much data available, we need to decide which variables make sense to include.

We specify a hierarchical regression, a set of sequential models, to determine if the full model significantly adds to our ability to predict the outcome over our reduced model. The null hypothesis is that the full model (e.g. model 2 doesn't explain any additional variability over our reduced model (Model 1). The alternative hypothesis is that Model 2 explains significantly more variability than Model 1.

All Variables:
HLPEQUIP: I receive enough help and equipment to get the job done
HAVEINFO: I have enough information to get the job done
CONDEMND: I am free from the conflicting demands that other people make of me
OPDEVEL: I have an opportunity to develop my own special abilities
FRINGEOK: My fringe benefits are good
SUPHELP: My supervisor is helpful in getting the job done
COWRKHLP: The people I work with can be relied on when I need help

Building the Model
We want to control for job constraints: HLPEQUIP, HAVEINFO and CONDEMND (Model 1)
THEN, look at the effects of positive job characteristics: OPDEVEL, FRINGEOK, SUPHELP, COWRKHLP in addition to model 1 (model 2)

To test the hypothesis, we calculate a partial F-test. The formula for the partial F-test is above.
First, we need to calculate the critical value of F for the partial F-test. The df for the numerator of the critical F is equal to the difference in the number of predictors between the full and reduced models (in this case, 4) , and the df for the denominator of our critical F is equal to the SSE for the full model (n - 1 - # of predictors).

```
qf(.95, df1 = 3, df2 = 194)
[1] 2.651153
```

Partial F Test

```
> anova(lm.GenSocSurvey, lm.GenSocSurvey.full, test="F")
Analysis of Variance Table

Model 1: SATJOB1 ~ HLPEQUIP + HAVEINFO + CONDEMND
Model 2: SATJOB1 ~ HLPEQUIP + HAVEINFO + CONDEMND + OPDEVEL + FRINGEOK +
    SUPHELP + COWRKHLP
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1183 759.17
2   1179 627.58  4    131.59 61.804 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation: The partial F test was statistically significant (61.804), indicating that the reduced model has significantly worse fit than the full model. We can reject the null hypothesis that the reduction in $SS_{residual}$ resulting from the additional predictors is equal to zero.
If the F value was non-significant, we would not have been justified in adding the other variables because we want the simplest model that explains the most amount of variance.

Other notes:
For each slope, our null hypothesis is that the partial regression coefficient is 0, and our alternative hypothesis is that the partial regression coefficient is not 0. For the intercept, our null hypothesis is that the intercept is 0 (i.e., the predicted value of y when all x variables are 0 is 0), and the alternative is that the intercept is not 0. The null hypothesis for the F* test considers the overall model and asserts that the $R_2 = 0$. The alternative is that R_2 is greater than 0.
Just as we have for all other inferential tests, we first need to set alpha and obtain the critical value of t (for testing the parameter estimates (e.g., the slopes)) and the critical value of F (for determining whether a significant amount of variance in the outcome is explained by the model).
In a MLR, the degrees of freedom for the critical t is calculated as *n* - 1 - # of predictors. In our example that is 200 - 1 - 2 = 197. Thus, critical t (for alpha of .05, 2-sided) is 1.97. This is what we compare our t* (estimate/se) to for each parameter estimate in the model.

The critical value of F has degrees of freedom equal to the number of predictors for the numerator (2 in this case) and N – 1 – number of predictors for the denominator (197 in this case). This equates to 3.04. This is what we will compare our F* to in order to determine if our set of predictors explains a significant portion of the variability in the outcome.

qt(c(.025, .975), df = 197)
qf(.95, df1 = 2, df2 = 197)

The critical value of F has degrees of freedom equal to the number of predictors for the numerator (2 in this case) and N – _1 – _number of predictors for the denominator (197 in this case). This equates to 3.04. This is what we will compare our F* to in order to determine if our set of predictors explains a significant portion of the variability in the outcome.

Critical values of the F-distribution are found on pg. 26156 in book (appendix)

Partial F test for Model Trimming

We can also use a Partial F-test for modeling trimming, that is, to determine if removing variables is warranted. This is a desir-able approach if one seeks to arrive at the most parsimonious, yet most predictive model possible. The models still must be nested. With this approach, a full model is compared to a subset model in which some of the predictors in the full model have been removed. Let's consider an example. In model 2 that we just estimated, two of the variables are significant _(alc_gm and weight), but the remaining predictors (alcexp, typ_drks, and absorb) are not. We can determine if removing these predictors makes sense. In this example the critical value for the Partial F-test just happens to be the same as the critical value of F calculated to compare Model 1 to Model 2. The numerator df is the difference in the number of predictors between the two models (3), and the denominator df is the df for the SSE for the full model (194). Therefore, the critical value of F re-mains 2.65.

The R Code looks like this: (we create a mod3) and then compare with an "anova" function

```
mod3 <- lm(data = obs, bac100 ~ weight_m + alc_gm_m)
ols_regress(mod3)

anova(mod3, mod2, test = "F")
```

```
Analysis of variance Table

Model 1: bac100 ~ weight_m + alc_gm_m
Model 2: bac100 ~ alcexp_m + typ_drks_m + absorb_m + weight_m + alc_gm_m
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    197 71.759
2    194 71.452  3   0.30674 0.2776 0.8415
```

$F^* = .2776$, this does not exceed critical value of F, therefore we do not reject the null hypothesis. Removing alcexp, typ_drks and absorb doesn't significantly erode our ability to predict bac.

Semi Partial and Partial Correlations

Simple Correlation: The relationship between X & Y

Below: Numerator is the variance explained by all predictors. Denominator is the total variance of Y.

$$R^2_{YX_1} = \frac{a+c}{a+b+c+d}$$

Partial Correlation: Varaince expained in Y after controlling for all other predictors

Below: Numerator is the residual of X1 after accounting for X2. Denominator is the residual variance of Y after accounting for X2.

$$R^2_{YX_1|X_2} = \frac{a}{a+d}$$

Partial Correlation

For a predictor (SES) and outcome (MI) of interest, partial correlation first removes from both SES and MI all variance which may be ac-counted for by the other predictors (in this case, just one, LIFE EVENTS), then correlates the remaining variance of MI (the residual) with the remaining variance of MI (the residual).
Here, the partial correlation between SES and MI is _____ (interpreted from r with cor(obs1$resid_ry, obs2@resid_rx)). Notice that this is the correlation of the two residuals that we obtained in our previous activity.
(pg. 18 in notes)

Semi Partial Correlations: The proportion of the total variance of Y explained by X2 above and beyond other variables.

Below: Numerator is the variance of Y explained by X1. Denominator is the total variance of Y

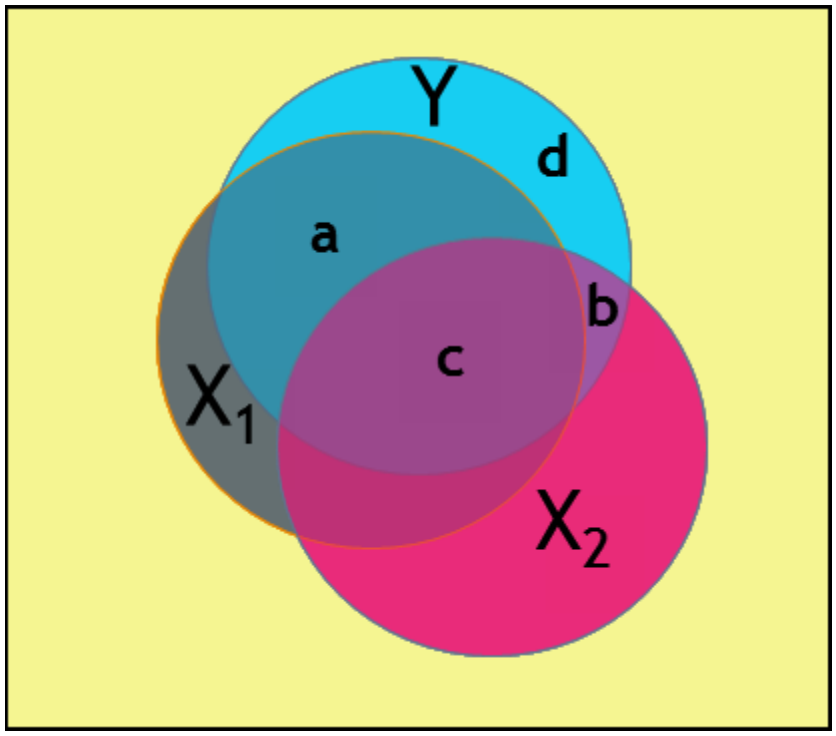
$$R^2_{YX_1|X_2} = \frac{a}{a+b+c+d}$$

If you have a reduced model and a full model with <u>just one</u> additional predictor
Semi-partial correlation is equal to:

R^2(full model) – R^2(reduced model)
OR
{$SS_{regression}$(full model) – $SS_{regression}$(reduced model)}/ SS_{total}(full model)

Semi-Partial (Part) Correlation

For a predictor (SES) and outcome (MI) of interest, semi-partial correlation first removes from the predictor (SES) all variance which may be accounted for by the other predictors (in this case LIFE EVENTS), then correlates the remaining variance of MI (the residual) with y.
Here the semi-partial (part) correlation between SES and MI is _____. This is concluded from the cor(obs1$MI, OBS1$resid_rx) function.


Y
d
a
b
c
X1
X2

Notes/Equations for Potential Reference

We use the least squares criterion to find the best fit line. This process starts by considering each case's residual. The residual is the difference between the observed value of y and the predicted value of y (the value that falls on the regression line — this is called y-hat).

Each case (e.g., individual—represented as id_i) in the dataset has an observed y, a y-hat, and a residual.

y_i Observed y (e.g., observed pounds lost)

$\hat{y}_i$ Predicted y (e.g., the value of y that we predict based on the individual's score on x) - the point on the best fit line

e_i Residual—calculated as observed y minus predicted y

Any line drawn through the data points, including the best fitting line as determined by the LSC, can be written in equation form.

The residual (e_i) is the difference between an individuals observed and predicted score on y.

$$y_i = b_0 + b_1 x_i + e_i$$

The intercept (b_0) is the predicted value of y when x=0.

The slope (b_1) is the predicted change in y for a 1 unit increase in x.

In order to determine the best fitting line, we rely on two equations called the normal equations.

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{770.30}{2613.56} = .29$$ ← Equation for the slope

$$b_0 = \bar{y} - b_1\bar{x} = 3.06 - .29(10.80) = -.12$$ ← Equation for the intercept

$$y_i = -.12 + .29x_i + e_i$$

This value (the residual) accounts for the fact that we do not perfectly predict each individual's score.

Among men who accumulate no caloric deficit (i.e., x=0), we predict .12 pounds of weight gain during the course of the program.

For each one unit (i.e., 1000 calories) increase in the accumulated caloric deficit, we predict .29 pounds of weight loss.