INTRO TO MLR

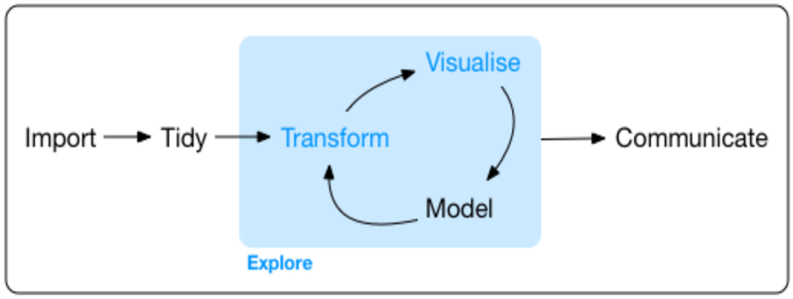
Research Methods in Psychology I & II ▪ Department of Psychology ▪ Colorado State University

BY THE END OF THIS UNIT YOU WILL:

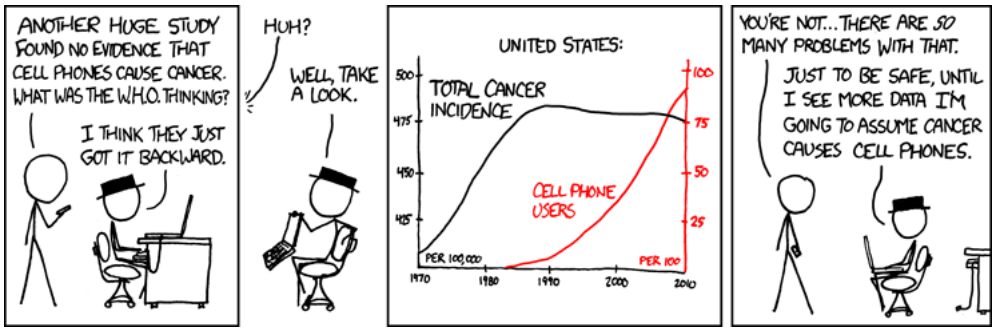
1. Understand the utility of including multiple predictor variables in a MLR model.
2. Know how to fit a Multiple Linear Regression model (MLR) in R.
3. Have an intuitive sense of what it means to control for variables when looking at the effect of others.
4. Be able to define partial and semi-partial correlation and understand the difference.
5. Know how to fit a series of nested models and be able to determine if one model fits significantly better than another.
6. Know how to plot the results of a MLR model.

What is Multiple Linear Regression (MLR)?

MLR extends SLR to consider more than one predictor. We'll continue working with one continuously distributed dependent variable (i.e., outcome, y), but now we'll consider multiple continuously distributed independent variables (i.e., predictors, x's).



Wickham & Grolemund—R for Data Science



© Kimberly L. Henry, Ph.D. ▪ kim.henry@colostate.edu

Dataset Description

A research team sought to examine factors associated with 21st birthday drinking among female students at a large University. Female students who were nearing age 21 and self-classified as regular drinkers were eligible for the study. In total, 200 students were recruited and agreed to take part in the study. Students were instructed to report to the lab two weeks prior to their 21st birthday. During this lab session, students completed a brief survey that measured alcohol use during the past month (using the Timeline Follow Back Method) and their weight was recorded. One week prior to their 21st birthday, participants were sent a link for an online survey to measure positive alcohol expectancies for drinking on their 21st birthday. Within three days prior to their 21st birthday, students reported to the lab and were given a diary-based data collection form to record several items on their 21st birthday. Students were instructed to record the food that they consumed during the day, the degree to which they were in a partying mood just prior to the celebration, and the quantity and type of drinks that they consumed during the first two hours of the celebration. The students were also given a small breathalyzer machine to measure BAC 2 hours after consumption of their first drink.

The dataset called bac_obs.csv contains the data:

- weight: weight in kilograms
- alcexp: positive alcohol expectancy for drinking on the impending 21st birthday, a multi-item scale that ranges from 1-7, where a higher score indicates more positive expectations about the role alcohol will play
- typ_drks: the number of standard alcohol drinks consumed in the past 30 days
- pmood: a rating on a scale from 1-9 on the respondent's mood to party on the 21st birthday, where 1 means never been less in the mood to party, and 9 means never been more in the mood to party
- absorb: a score calculated from the food diaries to determine how full the participant was when they began drinking, the score ranges from 1 to 8, where 1 means a completely full stomach, and 8 means a completely empty stomach
- alc_gm: a score calculated from the drinking diary to estimate the grams of alcohol consumed on the 21st birthday
- bac: the participant's blood alcohol content, measured as grams of alcohol per deciliter of blood on the 21st birthday

Set Up the Notebook

BAC_Notebook.Rmd*

Preview

```
---
title: "R Notebook for BAC data"
output: html_notebook
---


# Load libraries
```{r, message = FALSE}

library(huxtable)
library(olsrr)
library(GGally)
library(psych)
library(modelr)
library(tidyverse)

```

Import data
```{r, message = FALSE}

obs <- read_csv("bac_obs.csv")
exper <- read_csv("bac_exp.csv")

```
```

Please begin by setting up your notebook. Load the libraries (note that there is one new library—huxtable) and import the data. We will use the bac_exp dataset later, but we'll import it now to have it handy.

Descriptive Statistics

describe(obs)

```
         vars   n   mean    sd median trimmed   mad   min    max  range  skew kurtosis   se
id          1 200 100.50 57.88 100.50  100.50 74.13  1.00 200.00 199.00  0.00    -1.22 4.09
weight      2 200  68.46  9.92  68.90   68.52  9.56 37.60  91.40  53.80 -0.18     0.16 0.70
typ_drks    3 200  46.86 14.67  46.50   47.11 14.08  4.00  79.00  75.00 -0.15    -0.29 1.04
alcexp      4 200   4.09  0.78   4.13    4.11  0.77  2.01   6.06   4.05 -0.26    -0.07 0.06
pmood       5 200   5.12  1.40   5.00    5.09  1.48  1.00   9.00   8.00  0.18    -0.05 0.10
absorb      6 200   4.69  0.91   4.64    4.68  0.95  2.67   6.80   4.13  0.11    -0.60 0.06
alc_gm      7 200  32.79  7.73  33.00   32.76  8.90  8.00  58.00  50.00  0.04     0.43 0.55
bac         8 200   0.08  0.02   0.08    0.08  0.02  0.02   0.15   0.13  0.22     0.40 0.00
```

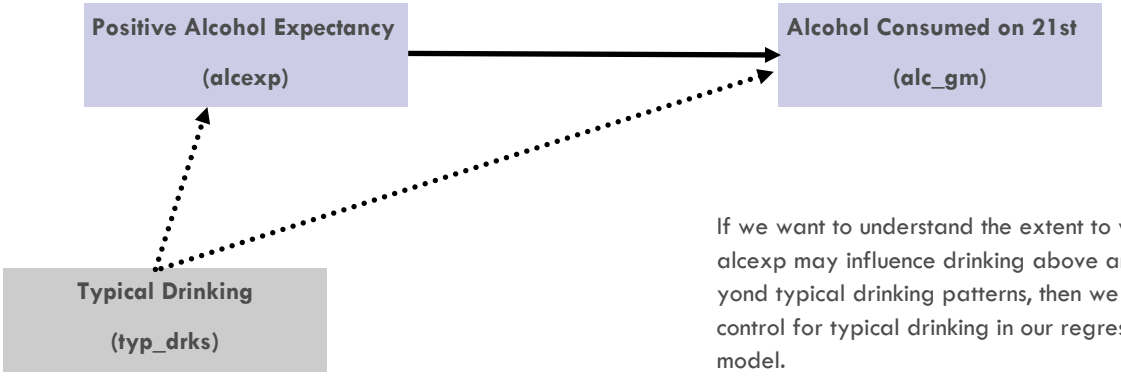
Initial Research Question

Let's begin with a simple research question:

Does a positive alcohol expectancy predict more drinking on a female's 21st birthday?

This is an important research question that considers the role positive alcohol expectancies may play in influencing drinking on the 21st birthday. If it were found that holding positive alcohol expectancies is causally associated with drinking, then an intervention designed to challenge these expectancies might be an effective strategy to reduce risky drinking in this setting. While only an experiment can prove cause and effect, we can attempt to better understand the potential causal effect of expectancies on consumption if we think carefully about common causes of these two constructs and control for them.

These common causes are potential confounders. A confounder is a pre-exposure variable that causes the exposure, and conditional on the exposure, the outcome. In our example, alcohol expectancy is the exposure and typical drinking is the potential confounder. If typ_drks is a confounder and is left out of the equation, the observed effect of alcexp on alc_gm is spurious. A spurious association is a relationship between two variables that does not result from any direct relation between them, but rather because they are both caused by another variable (or set of variables).

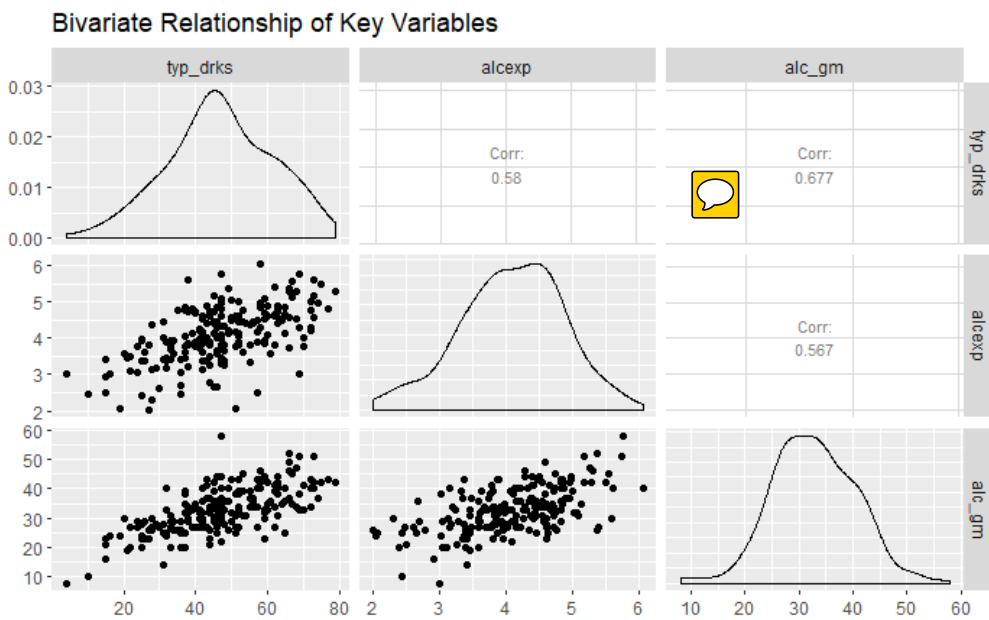


If we want to understand the extent to which alcexp may influence drinking above and beyond typical drinking patterns, then we need to control for typical drinking in our regression model.

Scatterplot Matrix

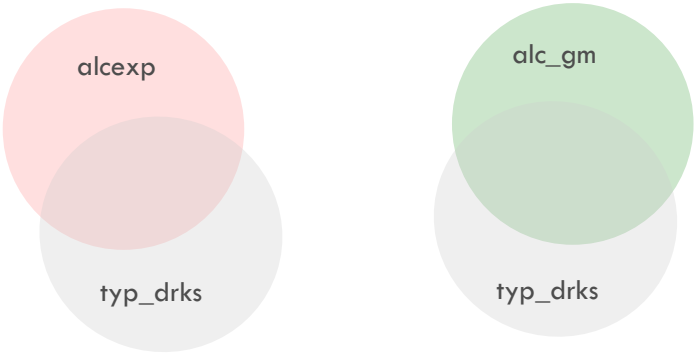
Let's take a look at a scatterplot matrix of the three variables of interest.

```
scatterplot <- ggpairs(obs, columns = c("typ_drks", "alcexp", "alc_gm"),
        upper = list(continuous = wrap("cor", size=3)),
        title = "Bivariate Relationship of Key Variables")
print(scatterplot, progress=FALSE )
```

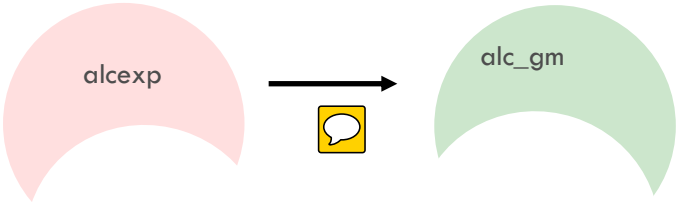


Partialling Out the Effect of a Potential Confounder

We desire to partial out the effect of typical drinking and estimate the independent effect of alcohol expectancies on alcohol consumption on the 21st birthday. From the correlation matrix, we see that 34% of the variability in alcexp is associated with typ_drks ($.580^2$) and 46% of the variability in alc_gm is associated with typ_drks ($.677^2$).



Imagine that we could pull out the variability in alc_gm and alcexp associated with typ_drks, then we'd be able to assess the association of alcexp on alcgms, independent of the effect of typ_drks. That is, the part of the relationship between alcexp and alc_gm that isn't associated with typ_drks.



Multiple linear regression (MLR) allows us to accomplish this task. The slope for each variable in the MLR will represent the unique effect of that variable holding constant/adjusting for/partialing out (these all mean the same thing) the effect of all others variables in the model.

Explore the Meaning of Partial Regression Coefficints in a MLR

The regression slopes in a MLR are called partial regression coefficients because the effect of the other variables has been partialled out. Let's explore what that means.

```
# step 1 - regress alc_gm (y) on variable(s) we want to partial out (typ_drks)
ry <- lm(data = obs, alc_gm ~ typ_drks)

# step 2 - regress alcexp (key predictor) on variable(s) we want to partial out (typ_drks)
rx <- lm(data = obs, alcexp ~ typ_drks)

# gather residuals from step 1 and 2, and add them to our dataset
obs1 <- obs %>%
  add_residuals(ry, var = "resid_ry") %>%
  add_residuals(rx, var = "resid_rx")
```

Name of the new variable that contains the residuals.

Name of the model used to calculate the residuals

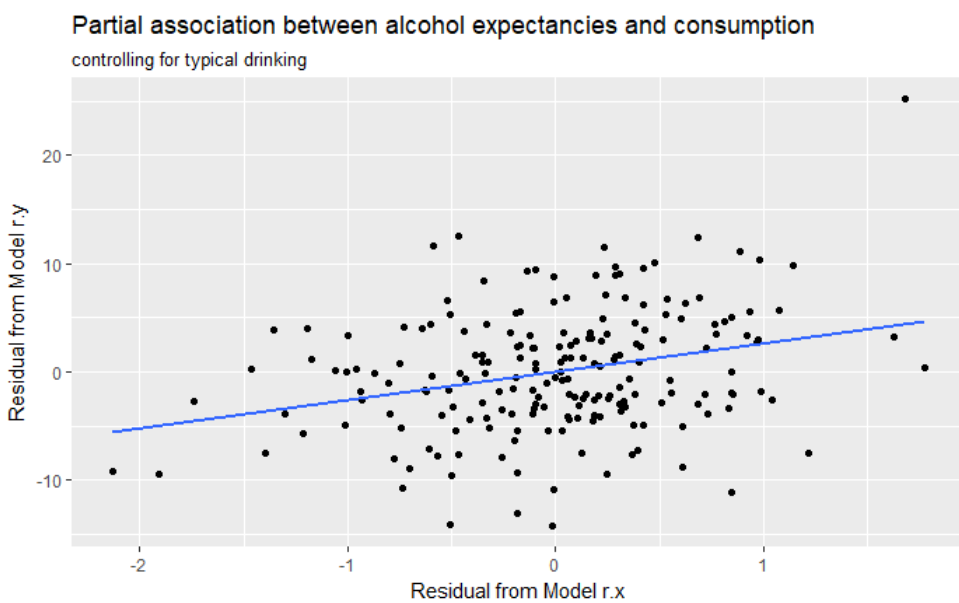
BAC_Notebook.Rmd* obs1

Filter

	id	weight	typ_drks	alcexp	pmood	absorb	alc_gm	bac	resid_ry	resid_rx
1	1	64.9	36	2.69	5	4.80	29	0.073	0.08214694	-1.060712219
2	2	62.6	42	3.61	4	5.05	32	0.085	0.94283924	-0.325512098
3	3	71.1	64	4.50	5	4.79	35	0.081	-3.90128900	-0.113111654
4	4	67.6	25	2.40	4	4.99	20	0.052	-4.99578894	-1.011912441
5	5	70.2	59	3.95	3	4.02	23	0.052	-14.11853258	-0.509111755
6	6	74.0	67	5.13	5	6.09	35	0.087	-4.97094285	0.424488406
7	7	57.2	15	2.92	4	4.63	16	0.041	-5.43027611	-0.183912643
8	8	69.8	69	4.30	3	4.77	33	0.082	-7.68404542	-0.467111553
9	9	69.8	52	5.01	6	5.90	39	0.095	4.37732640	0.766488104
10	10	59.2	42	3.13	5	5.17	30	0.081	-1.05716076	-0.805512098
11	11	71.6	63	4.41	7	5.84	41	0.085	2.45526228	-0.172311674
12	12	59.3	47	4.82	4	4.24	29	0.086	-3.83991718	0.730488003
13	13	74.4	66	5.30	4	6.09	46	0.097	6.38560843	0.625288386

Scatterplot of Residuals

```
# step 3 - plot residual from step 1 (on y) and residual from step 2 (on x) on a scattergram
ggplot(data=obs1, aes(x = resid_rx, y = resid_ry)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Partial association between alcohol expenctancies and consumption",
      subtitle = "controlling for typical drinking",
      x = "Residual from Model r.x", y = "Residual from Model r.y")
```



Even after partialling out the effect of typical drinking, we still see a positive relationship between alcohol expectancies and alcohol consumption.

Regression of alc_gm residual on alcexp residual

```
# step 4 - regress residual from step 1 on residual from step 2

ryx <- lm(data = obs1, resid_ry ~ resid_rx)

ols_regress(ryx)
```

```
                          Model Summary
---------------------------------------------------------------------
R                       0.292       RMSE                        5.454
R-Squared               0.085       Coef. Var          -3.301718e+16
Adj. R-Squared          0.080       MSE                        29.751
Pred R-Squared          0.063       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.
---------------------------------------------------------------------
Regression      547.609         1        547.609    18.407    0.0000
Residual       5890.633       198         29.751
Total          6438.243       199
---------------------------------------------------------------------

                                Parameter Estimates
-----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t       Sig      lower    upper
-----------------------------------------------------------------------------------------
(Intercept)     0.000         0.386                 0.000    1.000    -0.761    0.761
   resid_rx     2.611         0.609        0.292    4.290    0.000     1.411    3.811
-----------------------------------------------------------------------------------------
```

```
cor(obs1$resid_ry, obs1$resid_rx)
```

```
[1] 0.2916431
```

The regression coefficient for resid_rx is positive, indicating that partialling out typ_drks, a one-unit increase in alcexp is associated with a 2.611 unit increase in alcohol consumption. Note that the correlation is .29 (which of course is the same as the standardized beta) and the R^2 is .085, indicating that a little more than 8% of the remaining variability in alcohol consumption (after partialling out typical drinking) can be predicted by the remaining variability in alcohol expectancies (after partialling out typical drinking). We won't worry about the inference aspects of the model for now.

SLR: typ_drks → alc_gm

Before we fit the full MLR, let's take a look at the two SLR models, one for each predictor. Here's the first.

```
m1 <- lm(data = obs, alc_gm ~ typ_drks)

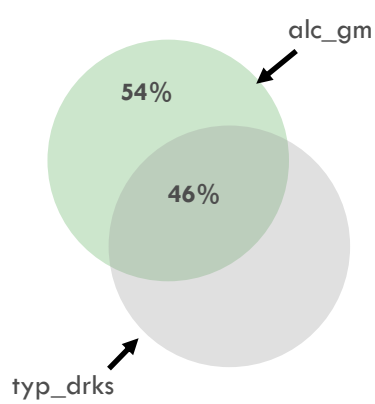
ols_regress(m1)
```

```
                        Model Summary
---------------------------------------------------------------
R                       0.677       RMSE                 5.702
R-Squared               0.458       Coef. Var           17.390
Adj. R-Squared          0.455       MSE                 32.516
Pred R-Squared          0.448       MAE                  4.441
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
-----------------------------------------------------------------------
                Sum of
               Squares         DF    Mean Square       F         Sig.
-----------------------------------------------------------------------
Regression     5444.937         1       5444.937    167.452    0.0000
Residual       6438.243       198         32.516
Total         11883.180       199
-----------------------------------------------------------------------

                                Parameter Estimates
-----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
-----------------------------------------------------------------------------------------
(Intercept)    16.082         1.353                 11.889    0.000    13.415    18.749
   typ_drks     0.357         0.028        0.677    12.940    0.000     0.302     0.411
-----------------------------------------------------------------------------------------
```

If a participant reports no drinking in the past 30 days (a typ_drks score of 0), we predict their alc_gm score to be 16.1. A standard drink is 14 grams of alcohol, so this equates to a little more than one drink. Now, recall that the sample represents drinkers, and the lowest typ_drks score is 4 — so a score of 0 is a bit outside of our data. Centering would be wise, and we will do this moving forward to the MLR. The slope for typ_drks equals .36; therefore, for each addition drink that a participant reported consuming during the prior 30 days, we predict that she will consume .36 additional grams of alcohol on her 21st birthday. This is a sizeable effect (the correlation is .68, and about 46% of the variability in alcohol consumed is predicted by drinking pattern during the past 30 days. The slope is significantly different from 0.



SLR: alcexp → alc_gm

Here's the second.

```
m2 <- lm(data = obs, alc_gm ~ alcexp)

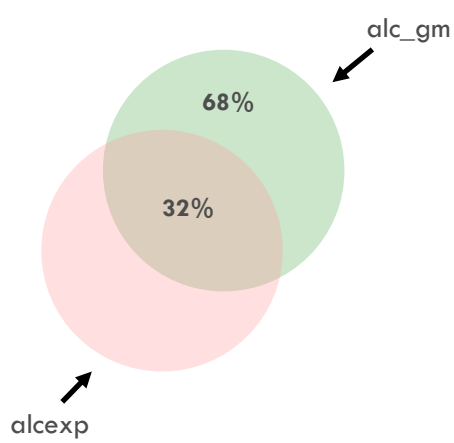
ols_regress(m2)
```

```
                        Model Summary
---------------------------------------------------------------
R                       0.567       RMSE                 6.380
R-Squared               0.322       Coef. Var           19.457
Adj. R-Squared          0.318       MSE                 40.703
Pred R-Squared          0.306       MAE                  5.105
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                ANOVA
-----------------------------------------------------------------------
                 Sum of
                Squares         DF    Mean Square      F         Sig.
-----------------------------------------------------------------------
Regression     3823.957          1       3823.957    93.947    0.0000
Residual       8059.223        198         40.703
Total         11883.180        199
-----------------------------------------------------------------------

                                Parameter Estimates
-------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta      t       Sig     lower     upper
-------------------------------------------------------------------------------------
(Intercept)     9.819         2.412                 4.070    0.000    5.062    14.577
     alcexp     5.623         0.580        0.567    9.693    0.000    4.479     6.767
-------------------------------------------------------------------------------------
```

How do you interpret this model?



Fit the MLR: alcexp + typ_drks → alc_gm

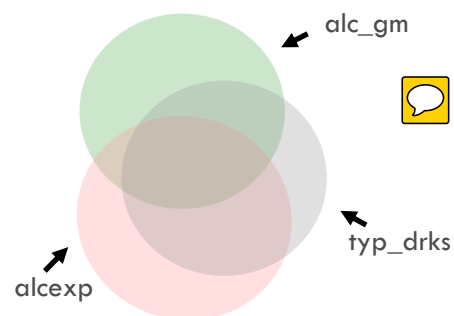
```
obs <- obs %>%
  mutate(alcexp_m = alcexp - mean(alcexp),
         typ_drks_m = typ_drks - mean(typ_drks))
m3 <- lm(data = obs, alc_gm ~ typ_drks_m + alcexp_m)
ols_regress(m3)
```

First, we will center our predictors at the mean, and then fit the MLR. To add multiple predictors, simply list each x variable on the right hand side of the tilde, separated by a + sign.

```
                        Model Summary
---------------------------------------------------------------
R                       0.710       RMSE                 5.468
R-Squared               0.504       Coef. Var           16.677
Adj. R-Squared          0.499       MSE                 29.902
Pred 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.
-----------------------------------------------------------------------
Regression     5992.547          2       2996.273    100.204    0.0000
Residual       5890.633        197         29.902
Total         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
-----------------------------------------------------------------------------------------
```



The venn diagram is for illustrative purposes and not drawn to perfect scale.

Let's begin digging into the parameter estimates. The intercept is 32.790, this is the predicted alcohol consumed for a female who reported average typical drinking and held average alcohol expectancies. Now consider our key predictor — alc_exp. In the SLR, the slope for alc_exp was 5.623, in the MLR it is 2.611. How are these effects interpreted across the two models?

In SLR: Each one unit increase in alc_exp is associated with a 5.6 unit increase in alc_gm.

In MLR: Holding constant/adjusting for/partialling out typ_drks, each one unit increase in alcexp is associated with a 2.6 unit increase in alc_gm.

In a MLR, the slope of each variable captures the unique effect of that particular variable, holding constant or partialling out the effect of the other predictors in the model. Notice that the regression coefficient for alcexp is the same as what we obtained using the residuals in our previous exercise.

Each regression slope is interpreted in the same manner, it's the expected change in y for a one-unit increase in x, holding constant all other predictors.

$$alc_gm_i = 32.79 + .28typ_drks_m_i + 2.61alcexp_m_i + e_i$$

Predicted Scores

$$alc_gm_i = 32.79 + .28typ_drks_m_i + 2.61alcexp_m_i + e_i$$

```
predgrid_demo <- data_grid(obs,
                    alcexp_m = c(-1, 0, 1),
                    typ_drks_m = c(1, 1, 1)) %>%
 add_predictions(m3)

predgrid_demo
```

Here we create a prediction matrix where alcexp_m varies, but typ_drks_m is held constant.

BAC_Notebook.Rmd* predgrid_demo

Filter

	alcexp_m	typ_drks_m	pred
1	-1	1	30.45496
2	0	1	33.06613
3	1	1	35.67729

33.06613 - 30.45496 = 2.61

35.67729 - 33.06613 = 2.61

Holding constant typ_drks, our predicted values of alc_gm demonstrate that each 1 unit increase in alcexp is associated with a 2.61 unit increase n alc_gm.

Make a Plot of the Fitted Model

Now we create a prediction matrix that spans the whole range of alcexp_m (that is what seq_range does) with 10 evenly spaced values of alcexp_m between the min and max values. We specify 3 prototypical values of weight. The data_grid function then produces all possible combos of our specification—so, the range of alcexp_m for each of the three prototypical weights.

```
predgrid_plot <- data_grid(obs,
                alcexp_m = seq_range(alcexp_m, 10),
                typ_drks_m = c(-14.67, 0, 14.67)) %>%
 mutate(typ_drks_m.f = factor(typ_drks_m,
                    levels = c(-14.67, 0, 14.67),
                    labels = c("1 SD Below Mean", "Mean", "1 SD Above Mean"))) %>%
 add_predictions(m3)


ggplot(data = predgrid_plot, aes(x = alcexp_m, y = pred, group = typ_drks_m.f, color = typ_drks_m.f)) +
 geom_line() +
 guides(color=guide_legend("Typical Drinking")) +
 labs(title = "Relationship between alcohol expectancy and alcohol consumed by typical use",
    x = "Positive Alcohol Expectancy (centered at the mean)", y = "Alcohol Consumed on 21st Birthday (grams)")
```

BAC_Notebook.Rmd | predgrid_plot

Filter

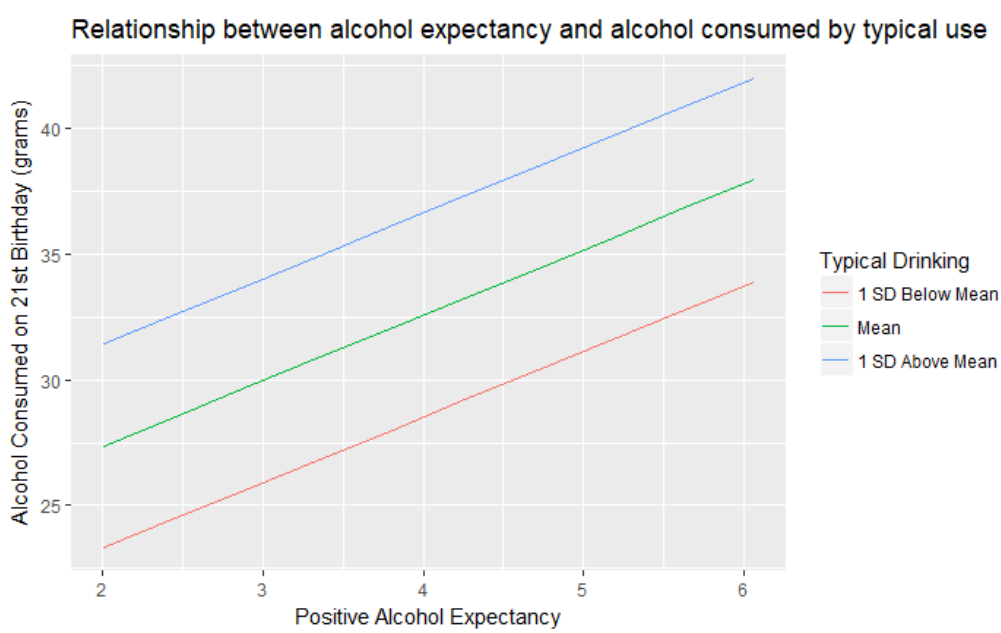
	alcexp_m	typ_drks_m	typ_drks_m.f	pred	alcexp
1	-2.0752	-14.67	1 SD Below Mean	23.32052	2.01
2	-2.0752	0.00	Mean	27.37131	2.01
3	-2.0752	14.67	1 SD Above Mean	31.42210	2.01
4	-1.6252	-14.67	1 SD Below Mean	24.49554	2.46
5	-1.6252	0.00	Mean	28.54633	2.46
6	-1.6252	14.67	1 SD Above Mean	32.59712	2.46
7	-1.1752	-14.67	1 SD Below Mean	25.67057	2.91
8	-1.1752	0.00	Mean	29.72136	2.91
9	-1.1752	14.67	1 SD Above Mean	33.77215	2.91
10	-0.7252	-14.67	1 SD Below Mean	26.84559	3.36
11	-0.7252	0.00	Mean	30.89638	3.36
12	-0.7252	14.67	1 SD Above Mean	34.94717	3.36

Modify Plot to Put x-axis on Original Scale

```
predgrid_plot <- data_grid(obs,
                    alcexp_m = seq_range(alcexp_m, 10),
                    typ_drks_m = c(-14.67, 0, 14.67)) %>%
  mutate(typ_drks_m.f = factor(typ_drks_m,
                         levels = c(-14.67, 0, 14.67),
                         labels = c("1 SD Below Mean", "Mean", "1 SD Above Mean"))) %>%
  add_predictions(m3) %>%
  mutate (alcexp = alcexp_m + mean(obs$alcexp))



ggplot(data = predgrid_plot, aes(x = alcexp, y = pred, group = typ_drks_m.f, color = typ_drks_m.f)) +
  geom_line() +
  guides(color=guide_legend("Typical Drinking")) +
  labs(title = "Relationship between alcohol expectancy and alcohol consumed by typical use",
      x = "Positive Alcohol Expectancy", y = "Alcohol Consumed on 21st Birthday (grams)")
```



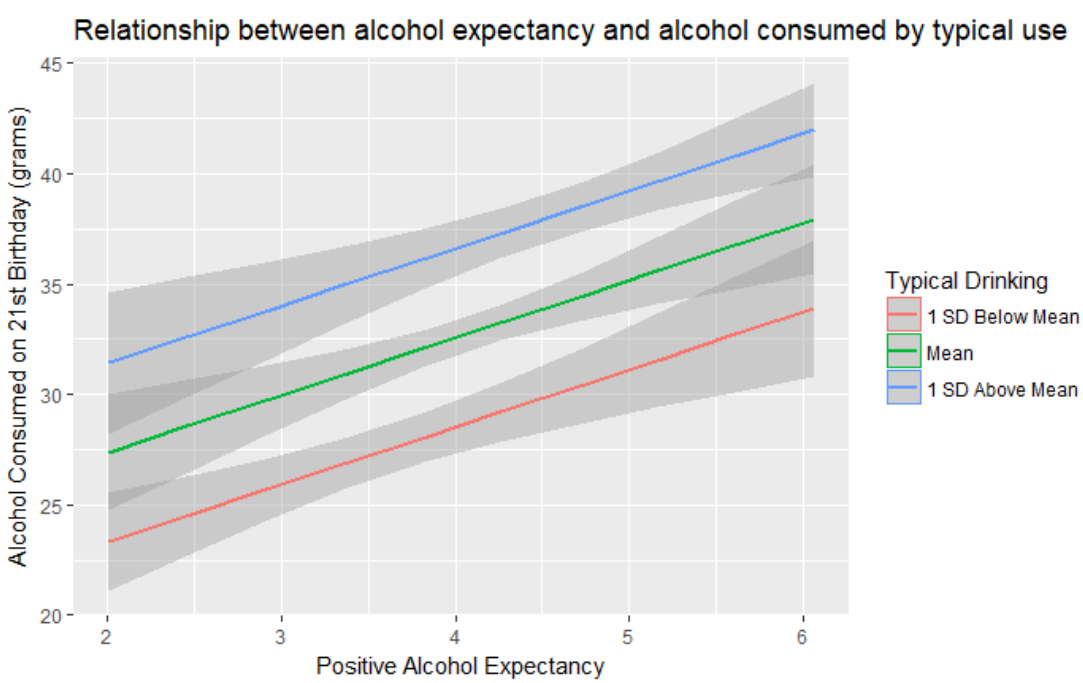
One Last Advanced Graph with Confidence Intervals

```
pred1 <- data_grid(obs,
                   alcexp_m = seq_range(alcexp_m, 10),
                   typ_drks_m = c(-14.67, 0, 14.67)) %>%
  mutate(typ_drks_m.f = factor(typ_drks_m,
                               levels = c(-14.67, 0, 14.67),
                               labels = c("1 SD Below Mean", "Mean", "1 SD Above Mean")))

pred2 <- predict(m3, pred1,  interval = "confidence") %>%
  as_data_frame()

pred3 <- cbind(pred1, pred2) %>%
  mutate (alcexp = alcexp_m + mean(obs$alcexp))


ggplot(data = pred3, aes(x = alcexp, y = fit, group = typ_drks_m.f, color =typ_drks_m.f)) +
  geom_ribbon(aes(ymin = lwr, ymax = upr, color = NULL), alpha = .4, fill = "grey60") +
  geom_line(size = 1) +
  guides(color=guide_legend("Typical Drinking")) +
  labs(title = "Relationship between alcohol expectancy and alcohol consumed by typical use",
       x = "Positive Alcohol Expectancy", y = "Alcohol Consumed on 21st Birthday (grams)")
```



R^2 for MLR Model

The R^2 for this MLR is 50.4% (obtained by SSR/SST) meaning that half of the variability in alc_gm can be predicted by alcexp and typ_drks. Notice that some of the explained variability is predicted just by typ_drks, some just by alcexp, and some can be predicted by both alcexp and typ_drks. MLR allows us to divvy up this variability and determine how much each variable uniquely contributes, as well as how the variables come together to predict the overall variability in the outcome.

Model Summary

R	0.710	RMSE	5.468
R-Squared	0.504	Coef. Var	16.677
Adj. R-Squared	0.499	MSE	29.902
Pred R-Squared	0.487	MAE	4.302

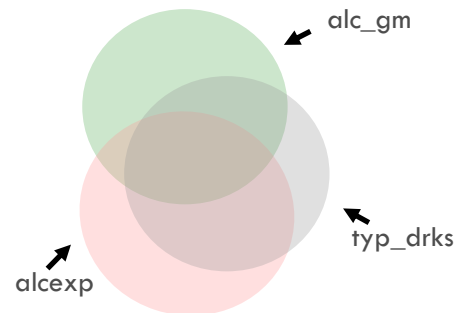
RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	5992.547	2	2996.273	100.204	0.0000
Residual	5890.633	197	29.902		
Total	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



The venn diagram is for illustrative purposes and not drawn to perfect scale.

The **adjusted R^2** corrects for a problem with regular R^2 – that is, in R^2 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 R^2, even when the added variable(s) adds nothing to the model. On the other hand, the adjusted R^2 can decline in value if the contribution to the explained variance by the additional variable(s) is less than the impact on the degrees of freedom. ADJUSTED R^2 REWARDS PARSIMONY.

While the R^2 can be interpreted as a proportion, the Adjusted R^2 CANNOT – so please don't interpret it in this way.

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

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

Predicted R^2 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.

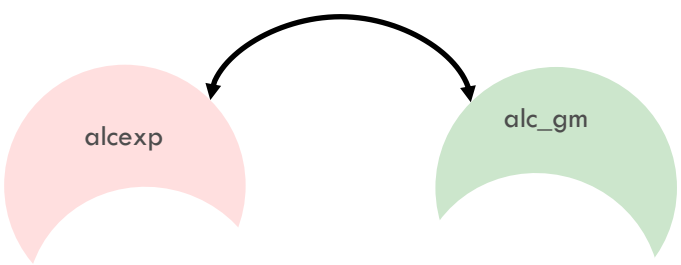
Partial and Semi-Partial (Part) Correlations

```
cor(obs1$resid_ry, obs1$resid_rx)
cor(obs1$alc_gm, obs1$resid_rx)
```

```
[1] 0.2916431
[1] 0.2146688
```

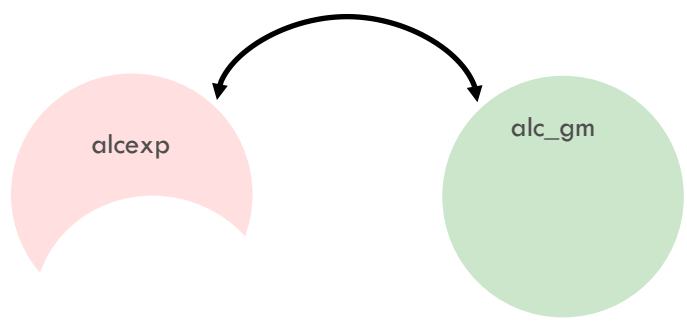
Partial Correlation

For a predictor (alcexp) and outcome (alc_gm) of interest, partial correlation first removes from both alcexp and alc_gm all variance which may be accounted for by the other predictors (in this case, just one, typ_drks), then correlates the remaining variance of alcexp (the residual) with the remaining variance of alc_gm (the residual). Here, the partial correlation between alcexp and alc_gm is .292. Notice that this is the correlation of the two residuals that we obtained in our previous activity.



Semi-Partial (Part) Correlation

For a predictor (alcexp) and outcome (alc_gm) of interest, semi-partial correlation first removes from the predictor (alcexp) all variance which may be accounted for by the other predictors (in this case typ_drks), then correlates the remaining variance of alcexp (the residual) with y. Here the semi-partial (part) correlation between alcexp and alc_gm is .215.




```
ols_correlations(m3)
```

```
                Correlations
-------------------------------------------
Variable       Zero Order    Partial    Part
-------------------------------------------
typ_drks_m          0.677      0.519   0.427
alcexp_m            0.567      0.292   0.215
-------------------------------------------
```

Squared Partial and Semi-Partial (Part) Correlations

ols_correlations(m3)^2

```
           Zero-order    Partial       Part
typ_drks_m  0.4582054 0.26908172 0.18249236
alcexp_m    0.3217958 0.08505568 0.04608271
```

Squared Partial Correlation

The square of the partial correlation indicates what proportion of the residualized y (residual of alc_gm) is predicted by the residualized x_1 (residual of alcexp). In our example, 8.5% of the residual variability in alc_gm is predicted by the residual variables of alcexp.

Squared Semi-Partial (Part) Correlation

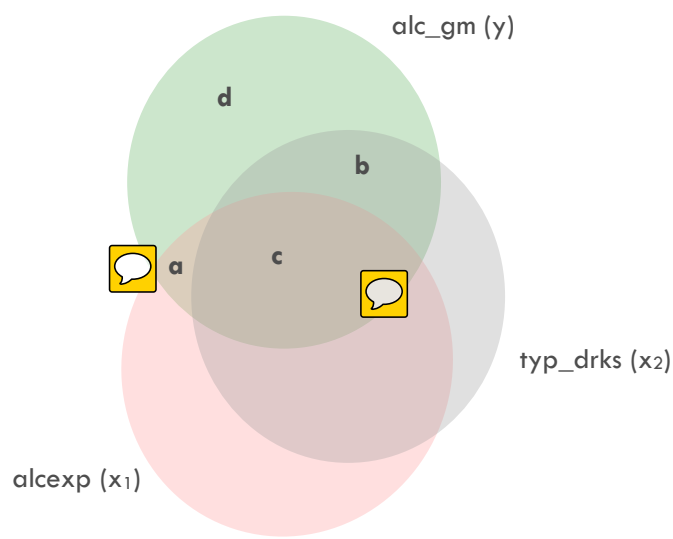
The square of the semi-partial correlation indicates the increment in R^2 that is achieved when we add our x of interest (x_1, alcexp) to a model that already includes all other predictors (e.g., x_2, typ_drks). In our example, 4.6% of the **total** variability of alc_gm is uniquely predicted by alcexp.

Squared Partial Correlation

$$R^2_{yx_1|x_2} = \frac{a}{a+d}$$

Squared Semi-Partial Correlation

$$R^2_{y(x_1|x_2)} = \frac{a}{a+b+c+d}$$



Squared Zero-Order (Simple) Correlation

$$R^2_{yx_1} = \frac{a+c}{a+b+c+d}$$

Inference for MLR

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)

```
[1] -1.972079  1.972079
[1] 3.041753
```

F* equals 100.204, this clearly exceeds the critical value of F, therefore we reject the null hypothesis. These variables explain variability in the outcome.

```
                        Model Summary
-------------------------------------------------------------
R                       0.710       RMSE                5.468
R-Squared               0.504       Coef. Var          16.677
Adj. R-Squared          0.499       MSE                29.902
Pred 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.
---------------------------------------------------------------------
Regression     5992.547          2       2996.273    100.204    0.0000
Residual       5890.633        197         29.902
Total         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
-----------------------------------------------------------------------------------------
```

The null hypothesis for the intercept in this example isn't overly interesting, so we'll 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.

Hierarchical Regression: Model Building

In a hierarchical regression, we build a series of nested regression models. Nested means that one smaller model is a subset of a larger model. To explore this technique, let's consider two models to predict blood alcohol content in the bac_obs dataset.

1. Model 1: Regress blood alcohol content (bac) on alcohol expectancies (alcexp) and typical drinking (typ_drks). We call this the reduced model in the hierarchical set.
2. Model 2: Build on Model 1, add three variables that are directly related to bac — the participants weight (weight), the absorption rate based on food consumed during the day (absorb), and alcohol consumed (alc_gm). We call this the full model in the hierarchical set.

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.

To test the hypothesis, we calculate a partial F-test. The formula for the partial F-test is below.

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, 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
```

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

Hierarchical Regression Syntax

Model 1

```
# bac is a very small number -- we will multiply it by 100 so it's easier to read the output,
# we will also center all of the predictors at the mean
obs <- obs %>%
  mutate(bac100 = bac*100,
         alc_gm_m = alc_gm - mean(alc_gm),
         weight_m = weight - mean(weight),
         absorb_m = absorb - mean(absorb),
         alcexp_m = alcexp - mean(alcexp),
         typ_drks_m = typ_drks - mean(typ_drks))

mod1 <- lm(data = obs, bac100 ~ alcexp_m + typ_drks_m)
ols_regress(mod1)
```

Model 2

```
mod2 <- lm(data = obs, bac100 ~ alcexp_m + typ_drks_m + absorb_m + weight_m + alc_gm_m)
ols_regress(mod2)
```

Compare Models 1 and 2

```
anova(mod1, mod2, test = "F")
```

List the reduced model first, then the full model second

Hierarchical Regression Results

Results of Model 1

```
                        Model Summary
-------------------------------------------------------------
R                       0.683       RMSE                1.587
R-Squared               0.467       Coef. Var          19.158
Adj. R-Squared          0.461       MSE                 2.520
Pred R-Squared          0.449       MAE                 1.277
-------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
-------------------------------------------------------------------
               Sum of
              Squares         DF    Mean Square      F         Sig.
-------------------------------------------------------------------
Regression    434.411          2        217.205    86.204    0.0000
Residual      496.377        197          2.520
Total         930.788        199
-------------------------------------------------------------------

                                Parameter Estimates
-----------------------------------------------------------------------------------------
      model     Beta    Std. Error    Std. Beta      t        Sig     lower    upper
-----------------------------------------------------------------------------------------
(Intercept)    8.286         0.112                 73.818    0.000    8.064    8.507
   alcexp_m    0.907         0.177        0.327     5.119    0.000    0.557    1.256
 typ_drks_m    0.065         0.009        0.440     6.886    0.000    0.046    0.083
-----------------------------------------------------------------------------------------
```

Results of Model 2

```
                        Model Summary
-------------------------------------------------------------
R                       0.961       RMSE                0.607
R-Squared               0.923       Coef. Var           7.325
Adj. R-Squared          0.921       MSE                 0.368
Pred R-Squared          0.918       MAE                 0.447
-------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
-------------------------------------------------------------------
               Sum of
              Squares         DF    Mean Square       F         Sig.
-------------------------------------------------------------------
Regression    859.335          5        171.867    466.635    0.0000
Residual       71.452        194          0.368
Total         930.788        199
-------------------------------------------------------------------

                                Parameter Estimates
-----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta       t        Sig      lower     upper
-----------------------------------------------------------------------------------------
(Intercept)     8.286         0.043                 193.075    0.000     8.201     8.370
   alcexp_m     0.023         0.073        0.008      0.320    0.749    -0.120     0.167
 typ_drks_m     0.003         0.004        0.019      0.675    0.501    -0.005     0.011
   absorb_m    -0.010         0.048       -0.004     -0.206    0.837    -0.104     0.084
   weight_m    -0.090         0.005       -0.414    -20.062    0.000    -0.099    -0.081
   alc_gm_m     0.252         0.008        0.899     31.057    0.000     0.236     0.268
-----------------------------------------------------------------------------------------
```

Comparison of Models. F* = 384.57, this greatly exceeds our critical value of F, therefore we reject the null hypothesis. Model 2 predicts significantly more variance in bac than Model 1.

```
Analysis of Variance Table

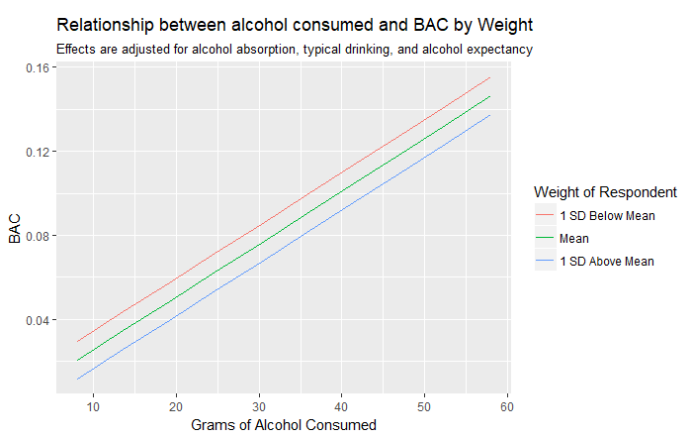
Model 1: bac100 ~ alcexp_m + typ_drks_m
Model 2: bac100 ~ alcexp_m + typ_drks_m + absorb_m + weight_m + alc_gm_m
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    197 496.38
2    194  71.45  3    424.92 384.57 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Plot Model 2

Let's plot Model 2 to see how we plot the fitted model relating alc_gm to BAC by three prototypical weights. We will hold all of the other variables constant at the mean.

```
predgrid <- data_grid(obs,
                      alc_gm_m = seq_range(alc_gm_m, 10),
                      weight_m = c(-9.92, 0, 9.92),
                      absorb_m = 0,
                      alcexp_m = 0,
                      typ_drks_m = 0) %>%
  mutate(weight_m.f = factor(weight_m, levels = c(-9.92, 0, 9.92),
           labels = c("1 SD Below Mean", "Mean", "1 SD Above Mean"))) %>%
  add_predictions(mod2) %>%
  mutate (bac = pred/100, alc_gm = alc_gm_m + mean(obs$alc_gm))


ggplot(data = predgrid, aes(x = alc_gm, y = bac, group = weight_m.f, color = weight_m.f)) +
  geom_line() +
  guides(color=guide_legend("Weight of Respondent")) +
  labs(title = "Relationship between alcohol consumed and BAC by Weight",
      subtitle = "Effects are adjusted for alcohol absorption, typical drinking, and alcohol expectancy",
      x = "Grams of Alcohol Consumed", y = "BAC")
```



Partial F 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 desirable 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 predictors (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 remains 2.65.

Model 3

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

Compare Models 2 and 3

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

```
                        Model Summary
---------------------------------------------------------------
R                       0.961       RMSE                0.604
R-Squared               0.923       Coef. Var           7.284
Adj. R-Squared          0.922       MSE                 0.364
Pred R-Squared          0.920       MAE                 0.449
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                 ANOVA
-----------------------------------------------------------------------
               Sum of
              Squares         DF    Mean Square       F          Sig.
-----------------------------------------------------------------------
Regression    859.029          2        429.514    1179.142    0.0000
Residual       71.759        197          0.364
Total         930.788        199
-----------------------------------------------------------------------

                                   Parameter Estimates
----------------------------------------------------------------------------------------
      model      Beta    Std. Error    Std. Beta       t        Sig      lower     upper
----------------------------------------------------------------------------------------
(Intercept)     8.286         0.043                 194.146    0.000     8.201     8.370
   weight_m    -0.091         0.004       -0.418    -20.953    0.000    -0.100    -0.082
   alc_gm_m     0.257         0.006        0.918     46.056    0.000     0.246     0.268
----------------------------------------------------------------------------------------
```

```
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.

Describe Results

Make a table of the results from all three models

```
huxreg("Model 1" = mod1, "Model 2" = mod2, "Model 3" = mod3, error_format = "({std.error})",
    statistics = c("N" = "nobs", "R-Squared" = "r.squared", "F statistic" = "statistic", "P value" = "p.value"))
```

	Model 1	Model 2	Model 3
(Intercept)	8.286 ***	8.286 ***	8.286 ***
	(0.112)	(0.043)	(0.043)
alcexp_m	0.907 ***	0.023	
	(0.177)	(0.073)	
typ_drks_m	0.065 ***	0.003	
	(0.009)	(0.004)	
absorb_m		-0.010	
		(0.048)	
weight_m		-0.090 ***	-0.091 ***
		(0.005)	(0.004)
alc_gm_m		0.252 ***	0.257 ***
		(0.008)	(0.006)
N	200	200	200
R-Squared	0.467	0.923	0.923
F statistic	86.204	466.635	1179.142
P value	0.000	0.000	0.000

*** p < 0.001; ** p < 0.01; * p < 0.05.

The purpose of our analysis was to predict blood alcohol content (BAC) measured 2 hours after the start of female student's 21st birthday celebration. We used a hierarchical regression to examine two models. The first considered two predictors measured prior to the student's 21st birthday: typical drinking (measured as the number of standard drinks consumed during the 30 prior days) and positive expectancies for the role alcohol would play in the 21st birthday celebration. The second model added three additional variables measured on the day of the 21st birthday and highly relevant to BAC: the predicted absorption rate of the alcohol based on the amount of food consumed during the day, the participant's weight, and the grams of alcohol consumed. The two models were compared using a Partial F-test. The results of each model are presented in the Table 1. Model 1 predicted 46.7% of the variance in BAC ($F(2,197)=86.20$, $p<.001$). Model 2 predicted 92.3% of the variance in BAC ($F(5,194)=466.64$ $p<.001$). A Partial F-test indicated substantial improvement in the prediction of BAC for Model 2 over Model 1 ($F(3,194) = 384.57$, $p<.001$).

Assessment of the parameter estimates in Model 2 revealed that only 2 of the variables independently predicted BAC (weight and alcohol consumed on the 21st birthday). Moreover, while typical drinking and alcohol expectancies predicted BAC in Model 1, once the additional variables in Model 2 were added, neither variable independently predicted BAC. We elected to examine a final model that trimmed the three non-significant predictors from the model. The results are presented in Table 1, Model 3. A Partial F-test was used to compare Model 2 and 3 ($F(3, 194)=.28$, *NS*), the non-significant F-test indicates that the removal of these three variables did not significantly erode the overall model fit. Therefore, we consider Model 3 to be our final model.

*We will build on this model in future Units to consider a potential interaction between alc_gm and weight (i.e., perhaps the effect of each additional gram of alcohol is different depending on how much the student weighs), and whether alcohol consumption may mediate (i.e., explain) the effect of typical drinking and alcohol expectancies on BAC (i.e., the indirect effect0.