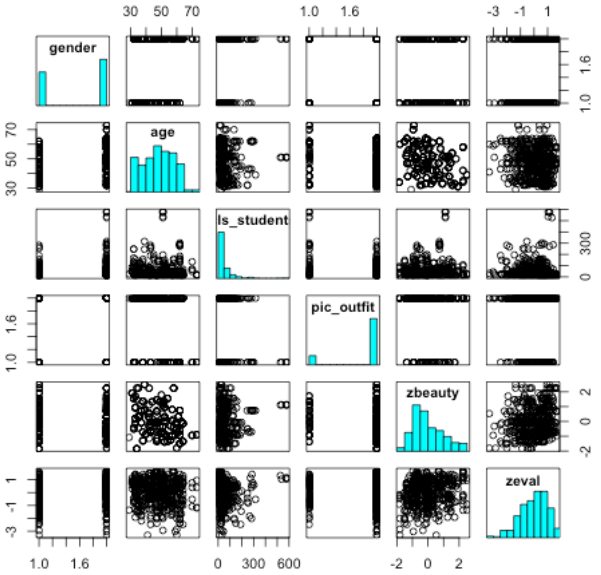
Assignment #11: Multiple Linear Regression

At the end of each semester, you complete course evaluation surveys to provide feedback on your courses and instructors. Results from those surveys are used to assess our teaching effectiveness.

The use of these surveys as indicators of teaching effectiveness is often criticized[1] since these evaluations may be biased by non-teaching-related factors, such as the gender, age, or race of the instructor[2].

In 2004, the University of Texas investigated the effect of physical appearance on course evaluations[3]. Do more physically attractive instructors receive higher course evaluation scores? If my evaluation scores are high, the answer is probably "yes," but we'll go ahead and analyze the dataset from this study.



Researchers sampled 94 instructors and collected data on the courses they taught between the years 2000-02. This led to a total of 463 course evaluations (with each instructor teaching between 1-3 courses).

For each of the 463 course evaluations, the researchers collected data on 24 variables[4]:

prof_id = professor ID : int 1 1 1 1 2 2 2 3 3 4 ...
course_eval = average course evaluation (1 = very unsatisfactory – 5 = excellent) : num 4.3 3.7 3.6 4.4 4.5 4 2.1 3.7 3.2 4.3 ...
prof_eval = average professor evaluation (1 = very unsatisfactory – 5 = excellent) : num 4.7 4.1 3.9 4.8 4.6 4.3 2.8 4.1 3.4 4.5 ...
rank = rank of professor (teaching, tenure-track, tenured) : Factor w/ 3 levels "teaching","tenure track",..: 2 2 2 2 3 3 3 3 3 3 ...
ethnicity = ethnicity of professor (minority, not minority) : Factor w/ 2 levels "minority","not minority": 1 1 1 1 2 2 2 2 2 2 ...
gender = gender of professor (female, male) : Factor w/ 2 levels "female","male": 1 1 1 1 2 2 2 2 2 1 ...
language = language of school where professor received education (english, non-english) : Factor w/ 2 levels : 1 1 1 1 1 1 1 1 ...
age = age of professor : int 36 36 36 36 59 59 59 51 51 40 ...
cls_perc_eval = percent of students in class completing survey: num 55.8 68.8 60.8 62.6 85 ...
cls_did_eval = number of students in class completing survey : int 24 86 76 77 17 35 39 55 111 40 ...
cls_students = total number of students in class : int 43 125 125 123 20 40 44 55 195 46 ...
cls_level = class level (lower, upper) : Factor w/ 2 levels "lower","upper": 2 2 2 2 2 2 2 2 2 2 ...
cls_profs = # of professors teaching sections in course : Factor w/ 2 levels "multiple","single": 2 2 2 2 1 1 1 2 2 2 ...
cls_credits = # of credits of class (one credit, multi-credit) : Factor w/ 2 levels "multi credit",..: 1 1 1 1 1 1 1 1 1 1 ...
bty_f1lower = beauty score of professor from lower-level female (1 = lowest, 10 = highest) : int 5 5 5 5 4 4 4 5 5 2 ...
bty_f1upper = beauty score of professor from upper-level female (1 = lowest, 10 = highest) : int 7 7 7 7 4 4 4 2 2 5 ...
bty_f2upper = beauty score of professor from 2nd upper-level female (1 = lowest, 10 = highest) : int 6 6 6 6 2 2 2 5 5 4 ...
bty_m1lower = beauty score of professor from lower-level male (1 = lowest, 10 = highest) : int 2 2 2 2 2 2 2 2 2 3 ...
bty_m1upper = beauty score of professor from upper-level male (1 = lowest, 10 = highest) : int 4 4 4 4 3 3 3 3 3 3 ...
bty_m2upper = beauty score of professor from 2nd upper-level male (1 = lowest, 10 = highest) : int 6 6 6 6 3 3 3 3 3 2 ...
bty_avg = average beauty score of professor : num 5 5 5 5 3 ...
pic_outfit = outfit of professor in picture (not formal, formal) : Factor w/ 2 levels "formal","not formal": 2 2 2 2 2 2 2 2 2 2 ...
pic_color = color of professor's picture (color, black&white) : Factor w/ 2 levels "black&white",..: 2 2 2 2 2 2 2 2 2 2 ...
pic_full_dept = whether or not all members of professor's department have photos available: Factor w/ 2 levels "no","yes": 2 2 2 …
class1 – class 30 = indicator for which of the classes with multiple professors the professor is teaching

I standardized the beauty ratings (the variables in red) by converting them to z-scores and then added them together to form a single beauty variable called **zbeauty**. I also standardized the course and professor evaluations (the variables in blue), added them together, and then standardized the composite evaluation score as **zeval.**

1. https://theconversation.com/students-dont-know-whats-best-for-their-own-learning-33835
2. http://sun.skidmore.union.edu/sunNET/ResourceFiles/Huston_Race_Gender_TeachingEvals.pdf
3. https://stat.duke.edu/courses/Fall12/sta101.002/CourseRatings.pdf
4. There were 54 variables in the full dataset, which you can download at: http://www.bradthiessen.com/html5/data/eval.csv

1. One condition of a regression analysis is that our data (or errors) are independent. Think about the data in this study (463 evaluations from courses taught by 94 instructors). We'll pretend as though all 463 observations are independent, but explain why these observations are <u>not</u> independent.

2. We want to investigate the extent to which an instructor's physical appearance (zbeauty) predicts course evaluation scores (zeval). To do this, let's compare the following models:

Reduced Model: Zeval = b_0

Least-squares line: y = 0
R^2 = 0; adjusted R^2 = 0
AIC = 1316.936
RMSE = 1

Full Model: Zeval = $b_0 + b_1$(Zbeauty)

Least-squares line: y = 0 + 0.19x
R^2 = 0.036; adjusted R^2 = 0.034
AIC = 1301.923
RMSE = 0.9829
F = 17.26 (p = 0.000039)

Interpret the slope of our full model:

0.19 represents: _____

Interpret the R-squared value:

0.036 represents: _____

Interpret the RMSE value for the full model:

0.9829 represents: _____

Use the R-squared value and sample size of n=463 to calculate the omnibus F-statistic to compare these models:

F = _____

3. Complete an ANOVA summary table to compare the two models. Remember the standard deviation of our Y variable is 1.0 (since we standardized the evaluation scores). Your MSR (F-statistic) should be the same as the omnibus F-statistic you just calculated (and the same as what is reported under the full model on the previous page).

Source	SS	df	MS	MSR (F)
Regression ($b_1 \mid b_0$)	_____	_____	_____	_____
Error	_____	_____	_____	p = _____
Total	_____	_____	MS_{total}	R^2 = _____

4. What conclusions can we make from this comparison of our full and reduced models? Does physical appearance predict course evaluation scores?

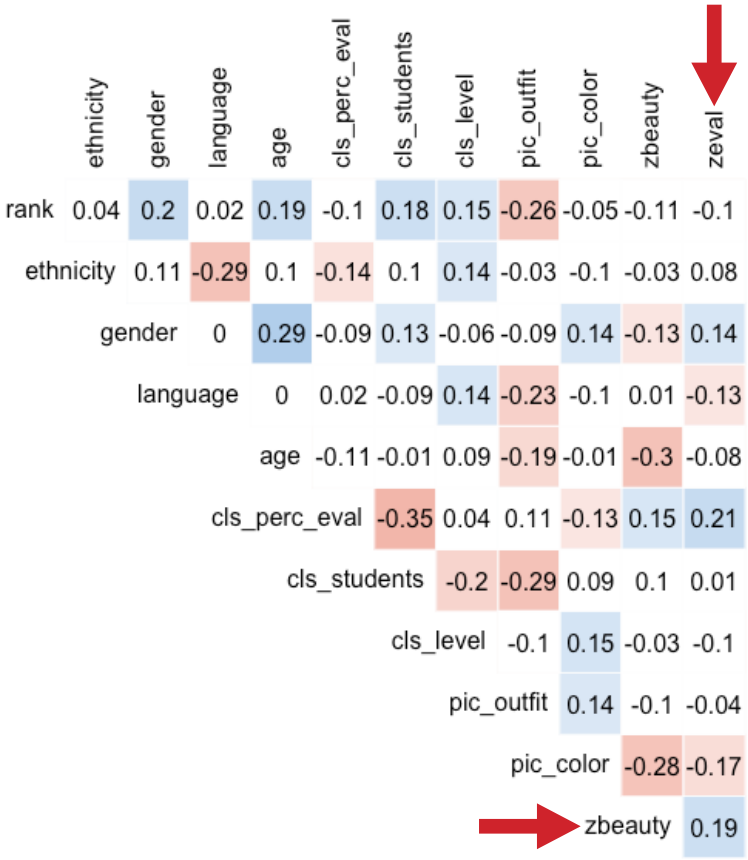
---

5. Let's investigate other models related to beauty and course evaluation scores. To the right, I've plotted a correlogram showing the correlations among pairs of variables in our dataset.

   Blue boxes represent larger positive correlations, while red boxes represent larger negative correlations I've used red arrows to highlight the evaluation and beauty variables we used in our previous full model. You can see these variables have a correlation of 0.19.

   The correlations can give us ideas for models to explore more in-depth. For example, evaluation scores are correlated to gender (higher evaluations for males), cls_perc_eval (higher evaluations from classes with a higher percentage of students completing evaluations), beauty, language (lower evaluations for non-native-English speaking instructors). Furthermore, it looks as though age and beauty have a negative correlation.

   Remember these correlations do not imply causation!



6. The model we constructed indicates physical appearance predicts course evaluation scores slightly. Would beauty still matter if we control for other variables, such as age, gender, and class size?

   Let's build some nested models and compare them. First, we'll start by determining if gender predicts course evaluation scores. To do this, we'll compare a reduced model (with no predictors) to a gender model (with a predictor of gender). Note that our gender variable has been coded so that 0 = female and 1 = male. We'll learn in the next activity that the method we're going to use is exactly the same as the independent samples t-test you learned in a previous statistics course.

   Here's some output from the gender model:

Gender Model:	Term	Coefficient	95% Confidence interval	p-value
R^2 = 0.02001	(intercept)	-0.1656438	-0.3051053 -0.02618226	0.0200
adj. R^2 = 0.018	gendermale	0.2861682	0.1028618 0.46947454	0.0023
AIC = 1309.579				
RMSE = 0.991				
F = 9.412				
p = 0.0023				

   From this output, we can conclude gender **IS** **IS NOT** a significant predictor of course evaluations.

7. Let's add age as a predictor to create a *gender-age* model. Here's the output from that model:

Gender-Age Model:	Term	Coefficient	95% Confidence interval	p-value
R^2 = 0.03595	(intercept)	0.44023637	-0.01317048, 0.893643219	0.05701
adj. R^2 = 0.03175	gendermale	0.36213731	0.17224907, 0.552025547	0.00022
RMSE = 0.984	age	−0.01343644	-0.02301111, -0.003861774	0.00605
AIC = 1303.987				
F = 8.576				
p = 0.00022				

Interpret the coefficient for the gender term:

0.362 represents: _____

Interpret the R-squared value:

0.03595 represents: _____

Why can't we simply compare the magnitude of the coefficients and conclude gender (coefficient = 0.36) is a more potent predictor of evaluations than age (coefficient of -0.013):

_____

Use the R-squared values of 0.03595 and 0.02001 to calculate the omnibus F-statistic to compare the gender and gender-age models:

F = _____

8. I compared these models in R and obtained the following output. Verify the F-statistic matches what you just calculated in the previous question. What can we conclude from this? Does age significantly predict course evaluation scores beyond gender?

```
Model 1: zeval ~ gender
Model 2: zeval ~ gender + age
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    461 452.76
2    460 445.39  1    7.3636 7.6051 0.006052 **
```

9. Calculate and interpret $R^2_{\text{age | gender}}$ = _____

10. Let's see if class size adds anything to our prediction. I constructed the following *gender-age-size* model:

Gender-Age-size:	Term	Coefficient	95% Confidence interval		p-value
R^2 = 0.03608	(intercept)	0.4500165509	-0.010087143	0.910120244	0.055217
adj. R^2 = 0.02978	gendermale	0.3654872346	0.173654674	0.557319796	0.000204
RMSE = 0.985	age	−0.0134997669	-0.023096688	-0.003902846	0.005934
AIC = 1305.921	cls_students	0.0001568858	-0.001367596	0.001053825	0.799110
F = 5.727					
p = 0.0007459					

Use the R-squared values of 0.03608 and 0.03595 to calculate the omnibus F-statistic to compare the gender-age and gender-age-size models:

F = _____

What can we conclude from this test? Does class size predict course evaluation scores beyond gender and age?

11. In question #2, we concluded that beauty does predict course evaluations. Let's see if beauty predicts evaluation scores after we control for the age and gender of the instructor.

Take a look at the output for these two competing models:

Gender-Age Model:	Term	Coefficient	95% Confidence interval		p-value
R^2 = 0.03595	(intercept)	0.44023637	-0.01317048,	0.893643219	0.05701
adj. R^2 = 0.03175	gendermale	0.36213731	0.17224907,	0.552025547	0.00022
RMSE = 0.984	age	−0.01343644	-0.02301111,	-0.003861774	0.00605
F = 8.576					
p = 0.00022					
AIC = 1303.987					

Gender-Age-beauty:	Term	Coefficient	95% Confidence interval		p-value
R^2 = 0.06903	(intercept)	0.161653055	-0.30453852	0.627844631	0.496
adj. R^2 = 0.06294	gendermale	0.379821404	0.19281717	0.566825638	.00008
RMSE = 0.968	age	−0.007888068	-0.01768662	0.001910483	0.114
F = 11.34	zbeauty	0.190751112	0.09793158	0.283570642	.00006
p = 0.00000034					
AIC = 1289.82					

Compare the AIC values and identify which model you would choose. Based on the following output, does beauty add to our prediction after controlling for age and gender?

```
Model 1: zeval ~ gender + age
Model 2: zeval ~ gender + age + zbeauty
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1    460 445.39
2    459 430.11  1    15.283 16.31 6.304e-05 ***
```

12. Take a look at the p-value for the age term in the gender-age-beauty model (p = 0.114, highlighted in a red font on the previous page). Based on this, what can we conclude about the age predictor?

13. Let's see if removing the age predictor significantly hurts our prediction of course evaluation scores.

    When I fit both of these models, I find: $R^2_{\text{gender, age, beauty}} = 0.06903$

    $$R^2_{\text{gender, beauty}} = 0.06395$$

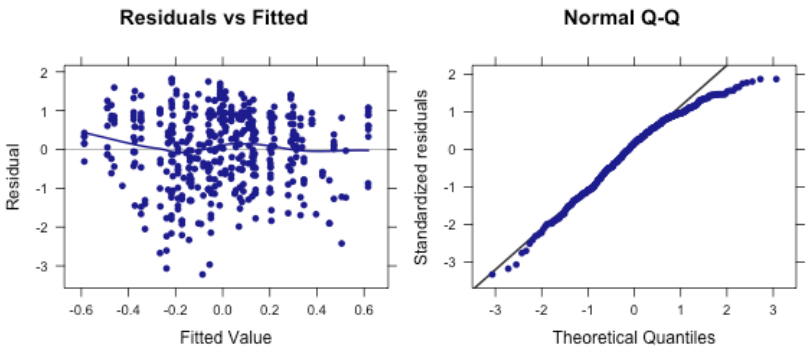
    Conduct an omnibus F-test to determine if removing age as a predictor significantly hurt our prediction of course evaluation scores.

    F = _____

    You can check your answer with the following output. From this, what conclusion can you make?

```
Model 1: zeval ~ gender + zbeauty
Model 2: zeval ~ gender + age + zbeauty
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    460 432.46
2    459 430.11  1    2.3452 2.5027 0.1143
```

14. To the right, I've pasted plots of the residuals for a model that includes gender and beauty as predictors. Based on these plots, the VIF, and the Durbin-Watson statistics, evaluate the assumptions necessary for multiple linear regression. Do any assumptions (other than independence) worry you?




```
> vif(noage)
  gender  zbeauty
1.016423 1.016423

> mean(vif(noage))
[1] 1.016423

> durbinWatsonTest(noage)
 lag Autocorrelation D-W Statistic p-value
   1        0.365388      1.267384       0
 Alternative hypothesis: rho != 0
```

15. Do you think there would be a significant interaction between beauty and gender? To check, we could compare:

$$R^2_{\text{gender, beauty}} = 0.06395$$

$$R^2_{\text{gender, beauty, gender x beauty}} = 0.07337$$

Gender-beauty-int:	Term	Coefficient	95% Confidence interval		p-value
R^2 = 0.07337	(intercept)	-0.1812859	-0.31859402	-0.04397773	0.009774
adj. R^2 = 0.06732	gendermale	0.3344626	0.15428209	0.51464315	0.000295
RMSE = 0.9658	zbeauty	0.1050824	-0.02631720	0.23648206	0.116742
F = 12.11	gndr x zbeauty	0.1963916	0.01775254	0.37503074	0.031256
p = 0.00000012					
AIC = 1287.654					

Interpret the coefficient for the interaction term. To do this, let's first rewrite the regression model separately for males and females. Substitute gender = 1 in the male formula and gender = 0 in the female formula. Then simplify.

Fill-in-the-blanks

Formula for males: -0.181 + 0.334(____) + 0.105(zbeauty) + 0.196(____ x zbeauty)

Formula for females: -0.181 + 0.334(____) + 0.105(zbeauty) + 0.196(____ x zbeauty)

Rewrite the formulas (simplify the above formulas, combining like terms):

Formula for males: _____

Formula for females: _____

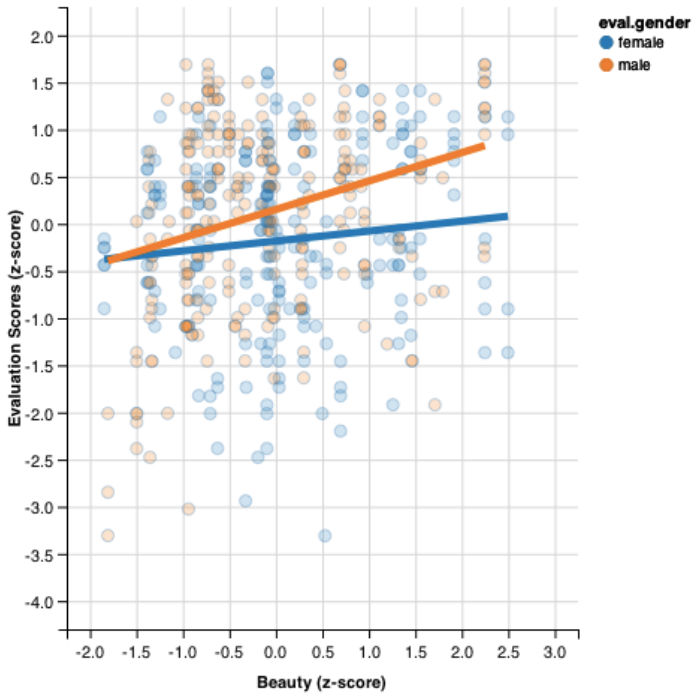
Using those formulas (and the graph to the right), explain what the interaction term means for this dataset:

_____

_____

_____

_____



16. Remember we've been ignoring the fact that the 463 observations in our dataset are not independent. We have instructors evaluated multiple times (and courses evaluated multiple times).

    A good way to deal with this dependence is to run a multi-level (nested) model in which instructors are nested within courses (or vice-versa). I'm hoping a student or two will investigate multilevel models for their course projects.

    For now, let's use a different (and much worse) method for dealing with the dependence. Let's just average the course evaluations for each instructor. When we do this, we end up with 94 average instructor evaluation scores.

    Which predictors should we use to predict the average instructor evaluations? Should we include beauty, age, and/or gender? To investigate this, let's use a cross-validation approach.

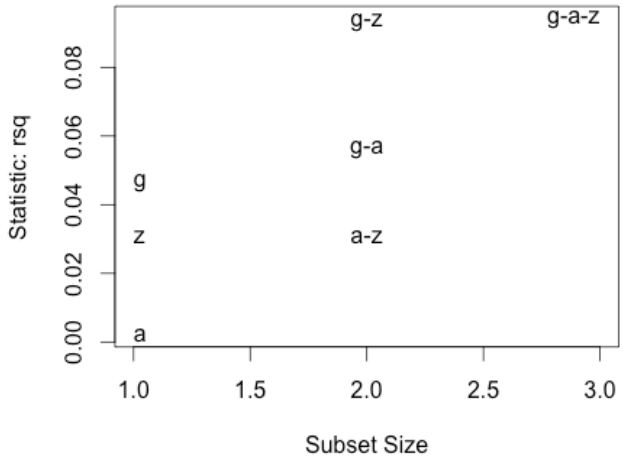
    Below, you can see the models I fit along with the average cross-validated mean square error for each:

Average C-V MSE	**Model**
0.760	Evaluations = f(age)
0.735	Evaluations = f(gender)
0.735	Evaluations = f(beauty)
0.740	Evaluations = f(age, beauty)
0.733	Evaluations = f(gender, age)
0.698	Evaluations = f(gender, beauty)
0.701	Evaluations = f(gender, age, beauty)
0.754	Evaluations = f(gender, age, gender x age)
0.740	Evaluations = f(age, beauty, age x beauty)
0.705	Evaluations = f(gender, beauty, gender x beauty)
0.701	Evaluations = f(gender, age, beauty, gender x age, gender x beauty, age x beauty)

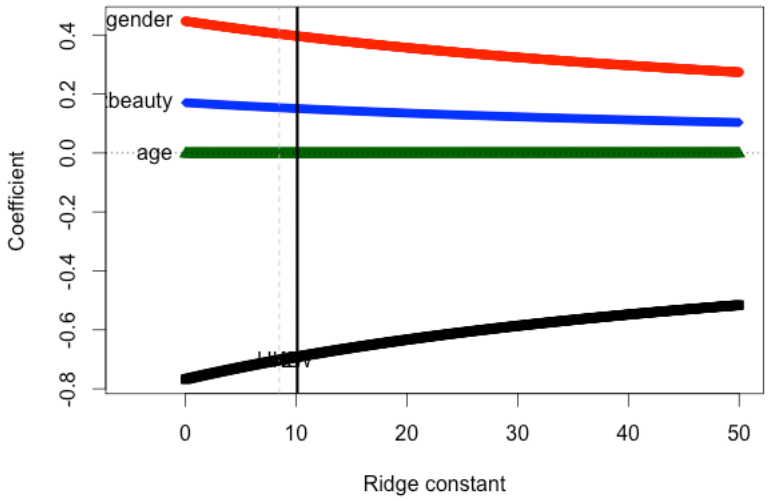
    The first group of models included a single predictor. The second group included two predictors. The third "group" was the model with all three predictors. The final group included interaction terms.

    Based on these results, which model would you conclude is the "best" model to predict course evaluations?

17. To the right, I've displayed results from a best subsets regression (a = age, g = gender, z = zbeauty). Based on this, which model would you choose to predict instructor evaluations?



18. To the right, I've displayed output using ridge regression (with a model that includes gender, age, and zbeauty). Based on this plot, what can we conclude about the stability of our coefficient estimates?



Assuming we have time, we'll try a little team competition during our next class. I'll provide a dataset and have teams of students work to produce their best possible predictions. Then, we'll test our predictions on a new dataset to see whose model was actually best.

We might also discuss this research report: http://bradthiessen.com/html5/stats/m301/referee.pdf