Multivariate regression

Load the `evals.csv`. The data were gathered from end of semester student evaluations for a large sample of professors from the University of Texas at Austin. In addition, six students rated the professors' physical appearance. The result is a data frame where each row contains a different course and columns represent variables about the courses and professors.

- score a verage professor evaluation score: (1) very unsatisfactory - (5) excellent.
- rank r ank of professor: teaching, tenure track, tenured.
- ethnicity ethnicity of professor: not minority, minority.
- gender gender of professor: female, male.
- language l anguage of school where professor received education: english or non-english.
- age a ge of professor.
- cls_perc eval p ercent of students in class who completed evaluation.
- cls_did_eval n umber of students in class who completed evaluation.
- cls_students t otal number of students in class.
- cls_level c lass level: lower, upper.
- cls_profs n umber of professors teaching sections in
- course in sample: single, multiple. cls_credits n umber of credits of class: one credit (lab, PE, etc.), multi credit.
- bty_f1lower b eauty rating of professor from lower
- level female: (1) lowest - (10) highest.
- bty_f1upper b eauty rating of professor from upper level female: (1) lowest - (10) highest.
- bty_f2upper b eauty rating of professor from second level female: (1) lowest - (10) highest.
- bty_m1lower b eauty rating of professor from lower level male: (1) lowest - (10) highest.
- bty_m1upper b eauty rating of professor from upper level male: (1) lowest - (10) highest.
- bty_m2upper b eauty rating of professor from second upper level male: (1) lowest - (10) highest. bty_avg a verage beauty rating of professor.
- pic_outfit o utfit of professor in picture: not formal, formal.
- pic_color c olor of professor's picture: color, black & white.

```
evals=read.csv('evals.csv')
```

Explore Data

1. Describe the distribution of `score`. Is the distribution skewed? What does that tell you about how students rate courses? Is this what you expected to see? Why, or why not?
2. Excluding `score`, select two other variables and describe their relationship using an appropriate visualization (scatterplot or side-by-side boxplots).

Multiple Linear Regression

The data set contains several variables on the beauty score of the professor: individual ratings from each of the six students who were asked to score the physical appearance of the professors and the average of these six scores. Let's take a quick look at the relationship between one of these scores and the average beauty score.

```
plot ( evals $ bty_avg  ~  evals $ bty_f1lower )
cor ( evals $ bty_avg ,  evals $ bty_f1lower )
```

As expected the relationship is quite strong – after all, the average score is calculated using the individual scores. We can actually take a look at the relationships between all beauty variables (columns 13 through 19) using the following command:

```
plot ( evals [,  13 : 19 ])
```

These variables are collinear (correlated), and adding more than one of these variables to the model would not add much value to the model. In this application and with these highly-correlated predictors, it is reasonable to use the average beauty score as a single representative of these variables.

In order to see if beauty is still a significant predictor of professor score after we've accounted for the gender of theprofessor, we can add the gender term into the model.

```
m_bty_gen  <-  lm ( score  ~  bty_avg  +  gender ,  data =  evals )
summary ( m_bty_gen )
```

3. Is `bty_avg` still a significant predictor of `score`? Has the addition of `gender` to the model changed the parameter estimate for `bty_avg`?
4. What is the equation of the line corresponding to males? (Hint: For males, the parameter estimate is multiplied by 1.) For two professors who received the same beauty rating, which gender tends to have the higher course evaluation score?
5. Create a new model called `m_bty_rank` with `gender` removed and `rank` added in. How does R appear to handle categorical variables that have more than two levels? Note that the rank variable has three levels: teaching, tenure track, tenured.
6. The interpretation of the coefficients in multiple regression is slightly different from that of simple regression. The estimate for `bty_avg` reflects how much higher a group of professors is expected to score if they have a beauty rating that is one point higher w hile holding all other variables constant. In this case, that translates into considering only professors of the same rank with `bty_avg` scores that are one point apart.

The search for the best model

We will start with a full model that predicts professor score based on rank, ethnicity, gender, language of the university where they got their degree, age, proportion of students that filled out evaluations, class size, course level, number of professors, number of credits, average beauty rating, outfit, and picture color.

7. Which variable would you expect to have the highest p-value in this model? Why? Hint: Think about which variable would you expect to not have any association with the professor score.

Let's run the model...

```
m_full  <- lm ( score  ~  rank  +  ethnicity  +
gender  +  language  +  age  +  cls_perc_eval +  cls_students  +  cls_level  +  cls_profs
+ cls_credits  +  bty_avg +  pic_outfit  +  pic_color ,  data =  evals )
summary ( m_full )
```

8. Check your suspicions from the previous exercise. Include the model output in your response.
9. Interpret the coefficient associated with the `ethnicity` variable.
10. Drop the variable with the highest p-value and refit the model. Did the coefficients and significance of the other explanatory variables change? (One of the things that makes multiple regression interesting is that coefficient estimates depend on the other variables that are included in the model.) If not, what does this say about whether or not the dropped variable was collinear with the other explanatory variables?
11. Using backward-selection and p-value as the selection criterion, determine the best model. You do not need to show all steps in your answer, just the output for the final model. Also, write out the linear model for predicting score based on the final model you settle on.
12. Verify that the conditions for this model are reasonable using diagnostic plots.

*Adopted from openintro labs. http://openintro.org