Review – chapters 3-6

MS 204 exam 2

Summary

The test covers the following units in the book:

- 3.1, 3.2
- 4.1, 4.2, 4.3
- Chapter 5
- 6.1, 6.2, 6.3

Format: HTML file over email

Review

```
library(oilabs)
library(mosaic)
data(evals)
head(evals)
```

```
## # A tibble: 6 x 21
##   score         rank    ethnicity gender language   age cls_perc_eval
##   <dbl>       <fctr>       <fctr> <fctr>   <fctr> <int>         <dbl>
## 1   4.7 tenure track     minority female  english    36      55.81395
## 2   4.1 tenure track     minority female  english    36      68.80000
## 3   3.9 tenure track     minority female  english    36      60.80000
## 4   4.8 tenure track     minority female  english    36      62.60163
## 5   4.6      tenured not minority   male  english    59      85.00000
## 6   4.3      tenured not minority   male  english    59      87.50000
## # ... with 14 more variables: cls_did_eval <int>, cls_students <int>,
## #   cls_level <fctr>, cls_profs <fctr>, cls_credits <fctr>,
## #   bty_f1lower <int>, bty_f1upper <int>, bty_f2upper <int>,
## #   bty_m1lower <int>, bty_m1upper <int>, bty_m2upper <int>,
## #   bty_avg <dbl>, pic_outfit <fctr>, pic_color <fctr>
```

We'll be working with the `evals` data set, which, as a reminder, contains evaluation information from the University of Texas in Austin.

For each of the following questions, assume a significance level of 95 percent.

- identify the appropriate hypothesis, using words and/or symbols
- state any assumptions required to do inference, and identify if those assumptions are met

- perform the test using the appropriate commands, either using R or by hand. For procedures involving the t-distribution, make sure to note the degrees of freedom.
- state appropriate conclusions, both techinical (reject/fail to reject, identification of p-value) and non-techical

For the test, you should be comfortable handling each of these questions **either** by hand or using R.

Hypothesis tests

1. Education codes require that 55 percent or fewer classes are taught by male professors, and Texas is worried that their male professor rates are too high. Test the hypothesis that 55 percent of courses are taught by male professors.

```
tally(~gender, data = evals)
```

```
## gender
## female   male
##    195    268
```

2. Having not taken the regression unit of statistics class, a student is interested in testing the overall average score by each gender. Create a 95 percent confidence interval for the average difference in evaluation score (`score`) by professor gender.

```
evals %>%
  group_by(gender) %>%
  summarise(ave.score = mean(score), sd.score = sd(score), n = n())
```

```
## # A tibble: 2 x 4
##   gender ave.score  sd.score     n
##   <fctr>     <dbl>     <dbl> <int>
## 1 female  4.092821 0.5638141   195
## 2   male  4.234328 0.5218958   268
```

3. School officials are curious as to the fraction of Texas students in each course that fill out evaluations. Using data from past years, officials hypothesize that the average rate of filling out evaluations should be around 40 percent. Using this data, test the claim that the average rate of students that fill out professor evaluations is 40 percent.

```
evals %>%
  summarise(ave.eval.rate = mean(cls_did_eval), sd.eval.rate = sd(cls_did_eval), n = n()
```

```
## # A tibble: 1 x 3
##   ave.eval.rate sd.eval.rate     n
##           <dbl>        <dbl> <int>
## 1      36.62419     45.01848   463
```

4. The language department at Texas is curious if there are any differences in the rates of native-english speakers (`langauge`) given the class level (`cls_level`), which is defined as either upper of lower. Test whether or not there is a difference in the percentage of english speakers by class level.

```
evals %>%
  group_by(cls_level) %>%
  summarise(rate.english = mean(language == "english"), n.classes = n())
```

```
## # A tibble: 2 x 3
##   cls_level rate.english n.classes
##      <fctr>        <dbl>     <int>
## 1     lower    0.9872611       157
## 2     upper    0.9150327       306
```

```
tally(language ~ cls_level, data = evals)
```

```
##              cls_level
## language      lower upper
##   english       155   280
##   non-english     2    26
```

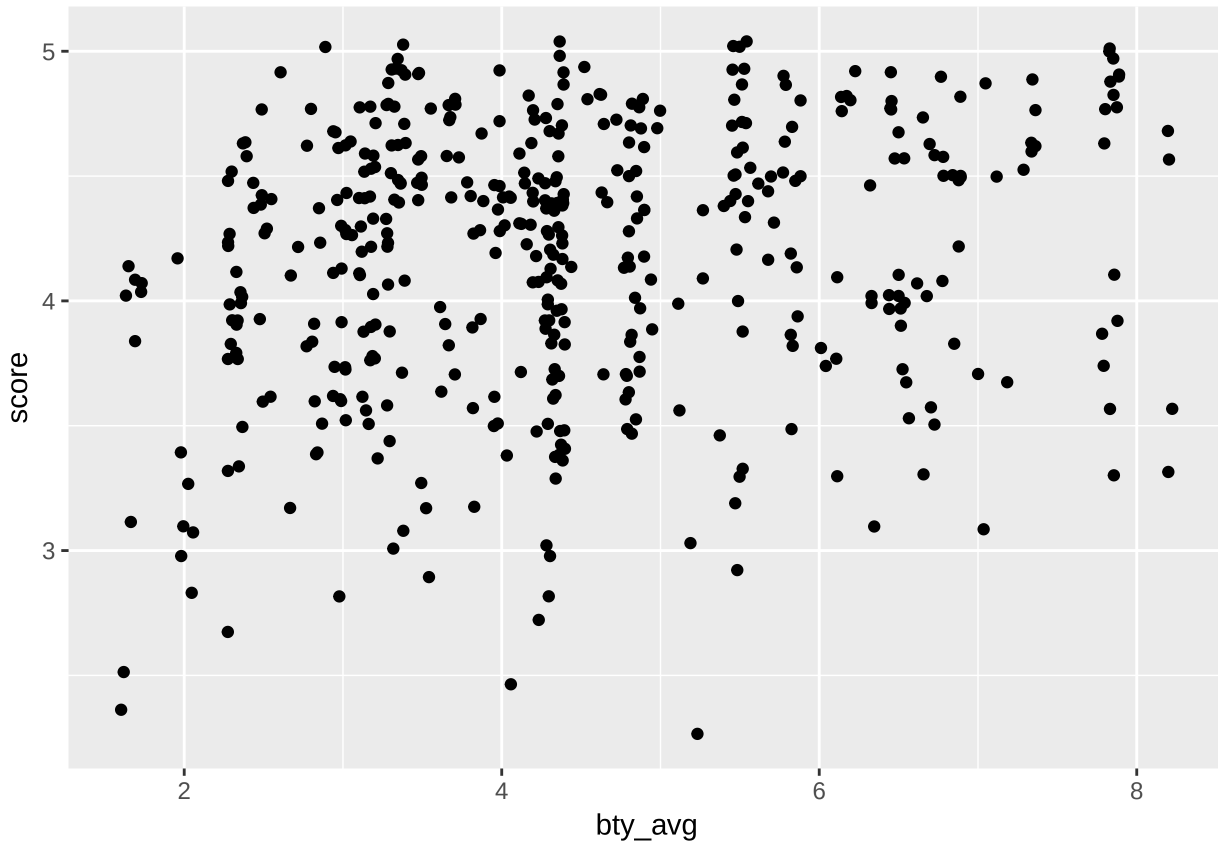
Additional questions

1. Using the teacher evaluation data set, design a question that **could** have been asked that would have required using a paired t-test, as in Chapter 4.2 of our book. You do not need to carry out this test.

2. A professor uses the following code. Run it, and answer the following questions.

```
evals <- evals %>%
  mutate(bty_avg_sq = bty_avg^2)

qplot(x = bty_avg, y = score, data = evals, geom = "jitter")
```




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

```
##
## Call:
## lm(formula = score ~ bty_avg + gender, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8305 -0.3625  0.1055  0.4213  0.9314
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.74734    0.08466  44.266  < 2e-16 ***
## bty_avg      0.07416    0.01625   4.563 6.48e-06 ***
## gendermale   0.17239    0.05022   3.433 0.000652 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5287 on 460 degrees of freedom
## Multiple R-squared:  0.05912,    Adjusted R-squared:  0.05503
## F-statistic: 14.45 on 2 and 460 DF,  p-value: 8.177e-07
```

- Identify the new variable, and use the scatter plot to propose why the professor created this new variable
- Write the estimated regression line
- Interpret the coefficient for `gender` in the regression model.
- Interpret the R-squared in the regression model.
- Professor Lopez is a male who is 35 years old, with a beauty average of 5 and an evaluation score of 4.3. Estimate his residual
- Is it worth keeping the quadratic term in the model?
- Check assumptions for the fit above, and compare to a fit that does not include a quadratic term for age. Which seems to more likely meet the assumptions of a linear model?