P8130 Biostatistical Methods Homework 5

Emil Hafeez (eh2928)

11/13/2020

Problem 1

```
#Read the CSV data into a dataframe
antibodies_df <- read.csv("./data/Antibodies.csv")

# Make AgeCategory, Smell, and Gender appopriate datatypes. Helps to ensure we know all unique values,
unique(antibodies_df$AgeCategory)
    antibodies_df <- antibodies_df %>%
        mutate(AgeCategory = factor(AgeCategory, labels = c("18-30", "31-50", "51+") ))
unique(antibodies_df$Smell)
    antibodies_df <- antibodies_df %>%
        mutate(Smell = factor(Smell, levels = c("Normal","Altered", "Unanswered/Others")))
    antibodies_df = antibodies_df %>% filter(Smell != "Unanswered/Others")
unique(antibodies_df$Gender)
    antibodies_df <- antibodies_df %>%
        mutate(Gender = factor(Gender, levels = c("Male","Female")))
```

In order to assess the difference in IgM levels between the two smell factor groups (Normal vs Altered) given non-normal distributions, we opt for a non-parametric test called the Wilcoxon Rank-Sum test. It's the nonparametric equivalent of the two-sample independent t-test, and examines if the medians of the two populations are equal versus not equal:

H_0 = the medians of the two groups' IgM levels are equal, and H_A = the medians of the two groups' IgM levels are not equal. The decision rule is that we reject H_0 = if $T > z_{1-(\alpha/2)}$.

The test statistic is computed, with a continuity correction, one of two ways. We first combine the data from the two groups, order the values from lowest to highest, assign ranks to the individual values (1 to n), and if ties, assign the average rank. Then, select a group and compute the sum of ranks T_1 for the first group, and then use the appropriate test statistic formula.

With no ties (referring to two equally ranked values once the values are listed), the test statistic is

$$T = \frac{\left|T_1 - \frac{n_1(n_1+n_2+1)}{2}\right| - \frac{1}{2}}{\sqrt{(n_1 n_2/12)(n_1+n_2+1)}}$$

and with ties, the test statistic is

$$T = \frac{\left|T_1 - \frac{n_1(n_1+n_2+1)}{2}\right| - \frac{1}{2}}{\sqrt{(n_1 n_2/12)\left[(n_1+n_2+1) - \sum_{i=1}^{g} t_i\left(t_i^2-1\right)/(n_1+n_2)(n_1+n_2-1)\right]}}$$

where t_i refers to the number of observations with the same absolute value in the i^{th} group and g is the number of tied groups.

In our case, the test statistic calculated by R is slightly different, since it does not by default add the $n_1(n_1+1)/2$ term (and is denoted by W).

The p-value under the normal approximation, with n_1 and $n_2 \geq 10$ is described by $2 * [1 - \Phi(T)]$.

```
antibodies_df2 =
  antibodies_df %>%
  pivot_wider(
        names_from = Smell,
        values_from = Antibody_IgM
    )

wilcox.test(antibodies_df2$Normal, antibodies_df2$Altered, mu = 0)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  antibodies_df2$Normal and antibodies_df2$Altered
## W = 5836, p-value = 0.01406
## alternative hypothesis: true location shift is not equal to 0
```

In context, and ignoring missing values and the unanswered smell category, we find evidence to reject the null hypothesis and conclude that the true location shift between the Normal and Altered smell categories is not equal to zero (in other words, the median IgM values are different for the two groups).

Problem 2

Problem 2 Part 1

$L\left(\mu, \sigma^2 \mid Y\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(Y_i-\mu)^2}{2\sigma^2}}$

$L\left(\beta_0, \beta_1, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(Y_i-\beta_0-\beta_1 X_i)^2}{2\sigma^2}} = \left(2\pi\sigma^2\right)^{-n/2} \cdot e^{\frac{-\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1 X_i)^2}{2\sigma^2}}$

$$\ln L\left(\beta_0, \beta_1, \sigma^2\right) = \log\left[\left(2\pi\sigma^2\right)^{-\frac{n}{2}} \cdot e^{\frac{-\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1 X_i)^2}{2\sigma^2}}\right]$$

$$\ln L\left(\beta_0, \beta_1, \sigma^2\right) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \sum_{i=1}^{n} \frac{(Y_i-\beta_0-\beta_1 X_i)^2}{2\sigma^2}$$

$\ln L\left(\beta_0, \beta_1, \sigma^2\right) =$

$$\frac{d}{d\beta_1} ln(\beta_0, \beta_1, \sigma) = -\frac{1}{2\sigma^2}\frac{d}{d\beta_1}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta 1\right)^2$$
$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta 1\right)$$

$0 = \frac{1}{\sigma^2}\Sigma_{i=1}^{n}(Y_i - \beta_1)$

$\beta_1 = \frac{1}{n}\Sigma_{i=1}^{n}(Y_i - \beta_1)$

$= \frac{cov(XY)}{var(X)}$

Segmented

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

/ Let's try that again.

$$\text{SSE} = \sum_{i}^{n}\left(y_i - \hat{y}_i\right)^2$$
$$= \sum_{i}^{n}\left[y_i - (\beta_1 x_i + \beta_0)\right]^2$$
$$= \sum_{i}^{n}\left(y_i - \beta_1 x_i - \beta_0\right)^2$$

$\frac{\partial}{\partial \beta_0}\text{SSE} = -2\sum_i^n (y_i - \beta_1 x_i - \beta_0)$
$= -2\sum_i^n y_i + 2\beta_1 \sum_i^n x_i + 2n\beta_0$ and then solving for the intercept when we set the derivative equal to 0
$= -2n\bar{y} + 2\beta_1 n\bar{x} + 2n\beta_0$
$= \beta_0 = \bar{y} - \beta_1 \bar{x}$

"We used the fact that the sum over the values of a variable is equal to the mean of those values multiplied by how many values we have."

Then, we can utilize this result to find the slope β_1 by using the intercept solution and isolating the slop term, such that

$\frac{\partial}{\partial \beta_1}\text{SSE} = -2\sum_i^n (y_i - \beta_1 x_i - \beta_0) x_i$
$= -2\sum_i^n x_i y_i + 2\beta_1 \sum_i^n x_i^2 + 2\beta_0 \sum_i^n x_i$
$= -2\sum_i^n x_i y_i + 2\beta_1 \sum_i^n x_i^2 + 2(\bar{y} - \beta_1 \bar{x}) \sum_i^n x_i$
$= -2\sum_i^n x_i y_i + 2\beta_1 \sum_i^n x_i^2 + 2\bar{y} \sum_i^n x_i - 2\beta_1 \bar{x} \sum_i^n x_i$
$= 2\beta_1 \sum_i^n x_i^2 - 2\beta_1 \bar{x} \sum_i^n x_i + 2\bar{y} \sum_i^n x_i - 2\sum_i^n x_i y_i$
$= 2\beta_1 \left(\sum_i^n x_i^2 - \bar{x} \sum_i^n x_i\right) + 2\bar{y} \sum_i^n x_i - 2\sum_i^n x_i y_i$

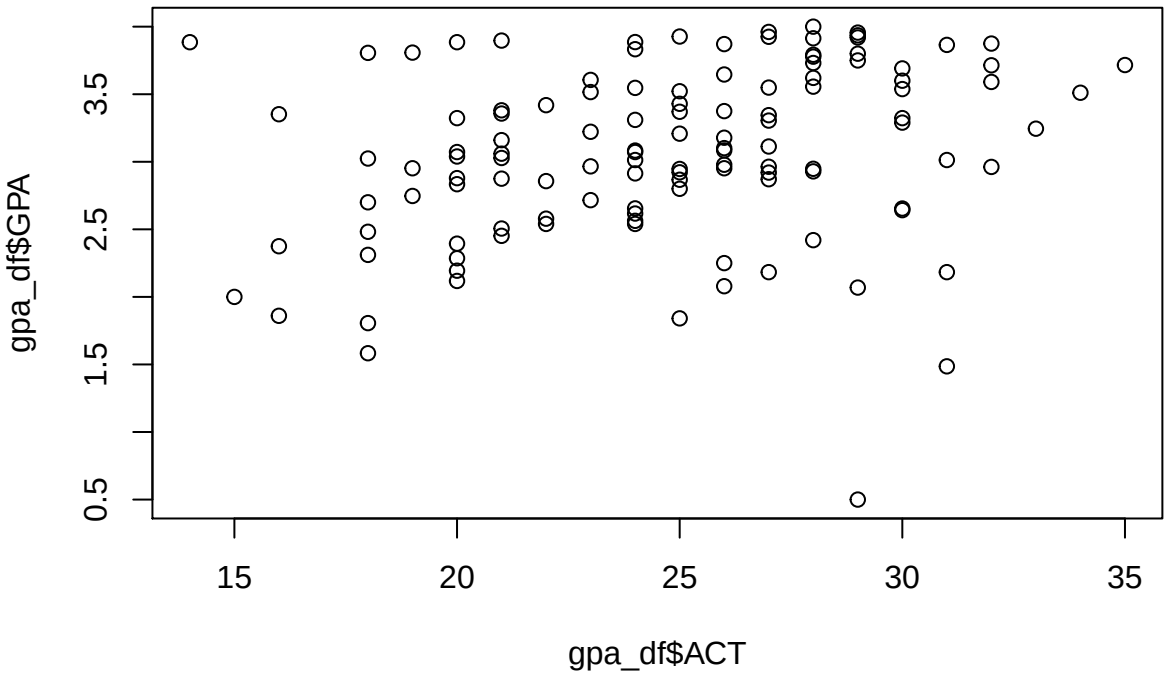
Problem 3

Problem 3 Part 1

Load the data and plot it.

```
gpa_df <- read.csv("./data/GPA.csv")

plot(gpa_df$ACT, gpa_df$GPA)
```



Test whether a linear association exists between ACT score (x) and GPA at the end of freshman year (Y).

Let $\alpha = 0.05$ and let β_1 represent the true slope to be estimated.

The null hypothesis is $H_0 : \beta_1 = \beta_{10}$ where $\beta_{10} = 0$. The alternative hypothesis is $H_A : \beta_1 \neq \beta_{10}$. In context, testing $H_0 : \beta_1 = 0$ examines whether a student's GPA at the end of freshman year can be predicted from the ACT test score.

The test statistic follows the t distribution with n-2 degrees of freedom, such that

$t = \frac{\widehat{\beta_1} - \beta_{10}}{se\left(\widehat{\beta_1}\right)} \sim t_{n-2}$, under $H_0 = \frac{0.03883 - 0}{0.01277} = 3.040$ using degrees of freedom $= n = 120, df = n - 2$

The corresponding critical value is fixed by $t_{n-2,1-(\alpha/2)}$ and the decision rule is that we reject H_0 if $|t| > t_{n-2,1-(\alpha/2)}$ and fail to reject H_0 if $|t| \leq t_{n-2,1-(\alpha/2)}$. As such, the critical value is $t_{118,0.975} = 1.980272$.

Therefore, $|t| > t_{118,0.975}$ using the 5% significance level, we find evidence to reject the null hypothesis and conclude that there is a significant linear association between students' ACT scores and GPA at the end of freshman year.

```
reg_admit<-lm(gpa_df$GPA ~ gpa_df$ACT, data = gpa_df)

# Summarize regression
summary(reg_admit)
```

```
##
## Call:
## lm(formula = gpa_df$GPA ~ gpa_df$ACT, data = gpa_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74004 -0.33827  0.04062  0.44064  1.22737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
## gpa_df$ACT   0.03883    0.01277   3.040  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6231 on 118 degrees of freedom
## Multiple R-squared:  0.07262,    Adjusted R-squared:  0.06476
## F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

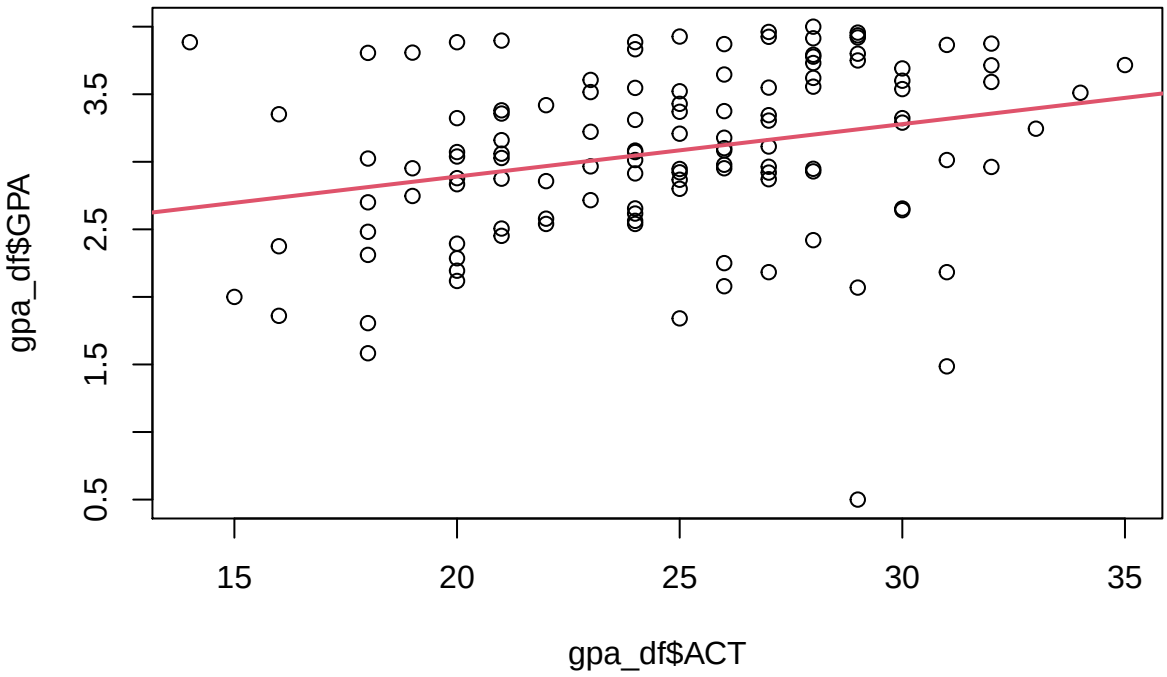
```
tidy(reg_admit)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic       p.value
##   <chr>          <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept)   2.11      0.321       6.59 0.00000000130
## 2 gpa_df$ACT    0.0388    0.0128      3.04 0.00292
```

```
glance(reg_admit)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1    0.0726        0.0648 0.623      9.24 0.00292     1  -113.  231.  239.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
# Scatterplot and regression line overlaid
plot(gpa_df$ACT, gpa_df$GPA)
abline(reg_admit,lwd=2,col=2)
```



Problem 3 Part 2

The basic regression model follows the form $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ and the estimated regression model is given by $\widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_i, i = 1, 2, 3...n$. In our context, this means that the estimated GPA value is equal to the sum of the intercept β_0 and the estimated beta one times the student's ACT score, such that $\widehat{GPA} = 2.11405 + 0.03883 \cdot ACT$.

Problem 3 Part 3

The 95% confidence interval for the true slope β_1 is given by $\widehat{\beta_1} \pm t_{n-2,1-(\alpha/2)} \cdot se(\widehat{\beta_1})$where $se(\widehat{\beta_1}) = \sqrt{MSE/\Sigma_{i=1}^{n}(X_i - \overline{X})^2}$.

In our context, the 95% confidence interval for the true slope is equal to $0.03883 \pm 1.980272 \cdot 0.01277 = (0.01354, 0.06412)$. This confidence interval does not include 0. The Director of Admissions may be interested in whether this confidence interval includes 0 because 0 is the null value. As it stands, with 95% confidence we estimate that the student GPA score increases by between (0.06412, 0.01354) points for each additional ACT point scored priorly. If the interval were to include 0, we may suspect that the true increase in student GPA score per additional ACT points may include 0, which is to say that the ACT may have no utility informing our admissions process. Perhaps there is other evidence we may consider that informs whether we retain ACT requirements, too, though.

Problem 3 Part 4

The 95% confidence interval for the mean freshman GPA for students whose ACT score is 28 is given by:

$$\widehat{\beta_0} + \widehat{\beta_1} X_h \pm t_{n-2,1-\alpha/2} \cdot \text{se}\left(\widehat{\beta_0} + \widehat{\beta_1} X_h\right)$$

$$\text{se}\left(\widehat{\beta_0} + \widehat{\beta_1} X_h\right) = \sqrt{MSE\left\{\frac{1}{n} + \left[\left(X_h - \bar{X}\right)^2 / \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right]\right\}}$$

```
reg_admit <- lm(GPA ~ ACT, data = gpa_df)

ACT_to_predict_from = data.frame(ACT = 28)
```

```r
predict(reg_admit, ACT_to_predict_from, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 3.201209 3.061384 3.341033
```

In context, this means that with 95% confidence we estimate the mean freshman GPA for students whose ACT scores are 28 to be between 3.061384 and 3.341033 (3.061384, 3.341033).

Problem 3 Part 5

The 95% prediction interval for Anne's freshman GPA is calculated as below.

$$\widehat{\beta_0} + \widehat{\beta_1} X_h \pm t_{n-2,1-\alpha/2} \cdot \text{se}\left(\widehat{\beta_0} + \widehat{\beta_1} X_h\right)$$

$$\text{se}\left(\widehat{\beta_0} + \widehat{\beta_1} X_h\right) = \sqrt{MSE\left\{\frac{1}{n} + \left[\left(X_h - \bar{X}\right)^2 / \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] + 1\right\}}$$

```r
reg_admit <- lm(GPA ~ ACT, data = gpa_df)

ACT_to_predict_from = data.frame(ACT = 28)

predict(reg_admit, ACT_to_predict_from, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 3.201209 1.959355 4.443063
```

We predict with 95% confidence that Anne's freshman year GPA will be between 1.959355 and 4.443063, as predicted from her ACT test score of 28.

Problem 3 Part 6

Prediction interval will always be wider than the CI, because it has an additional term to account for (the non-normally distributed error term). This is visible in the above formulas, where the confidence interval and prediction interval are calculated using similar formats, but the standard errors are quite different: the prediction interval focuses on one specific new value of Y_h, and since we do not calculate an expected mean the errors do not reduce to 0, and so the SE formula for prediction includes a +1 in the denominator, widening the interval overall.