homework5_jn2761

jiaying Ning

11/19/2020

```
rm(list=ls())
library(readxl)
library(tidyverse)
```

```
## -- Attaching packages -------------------------------------------------------------------------

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ----------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(broom)
library(dplyr)
```

Problem 1

```
#Import Data
antibodies_df=
   read.csv("./data/Antibodies.csv") %>%
  na.omit()
```

```
Normal = antibodies_df %>%
  filter(Smell == "Normal")

ALtered = antibodies_df %>%
  filter(Smell == "Altered")
```

non-parametric test to assess and comment on the difference in Ig-M levels between the two groups

The non parametric test we will be using is the Non-parametric Wilcoxon-Rank Sum test

Hypothesis

- H0:the median Ig_M levels are equal for both altered smell group and Normal smell group
- H1:the median Ig_M levels are not equal for both altered smell group and Normal smell group

calculations

```
wilcox.test(Normal$Antibody_IgM, ALtered$Antibody_IgM, mu=0)
```

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

```
test_stats_df=tidy(wilcox.test(Normal$Antibody_IgM, ALtered$Antibody_IgM, mu=0))
```

Test Stats

- W=5836
- p-value=0.0140605

Conclusion

- Using a 0.05 significance level, since we have p-value less than 0.05,, we reject H0 and conclude that the Normal smell and Altered Smell have significantly different Ig-M Level.

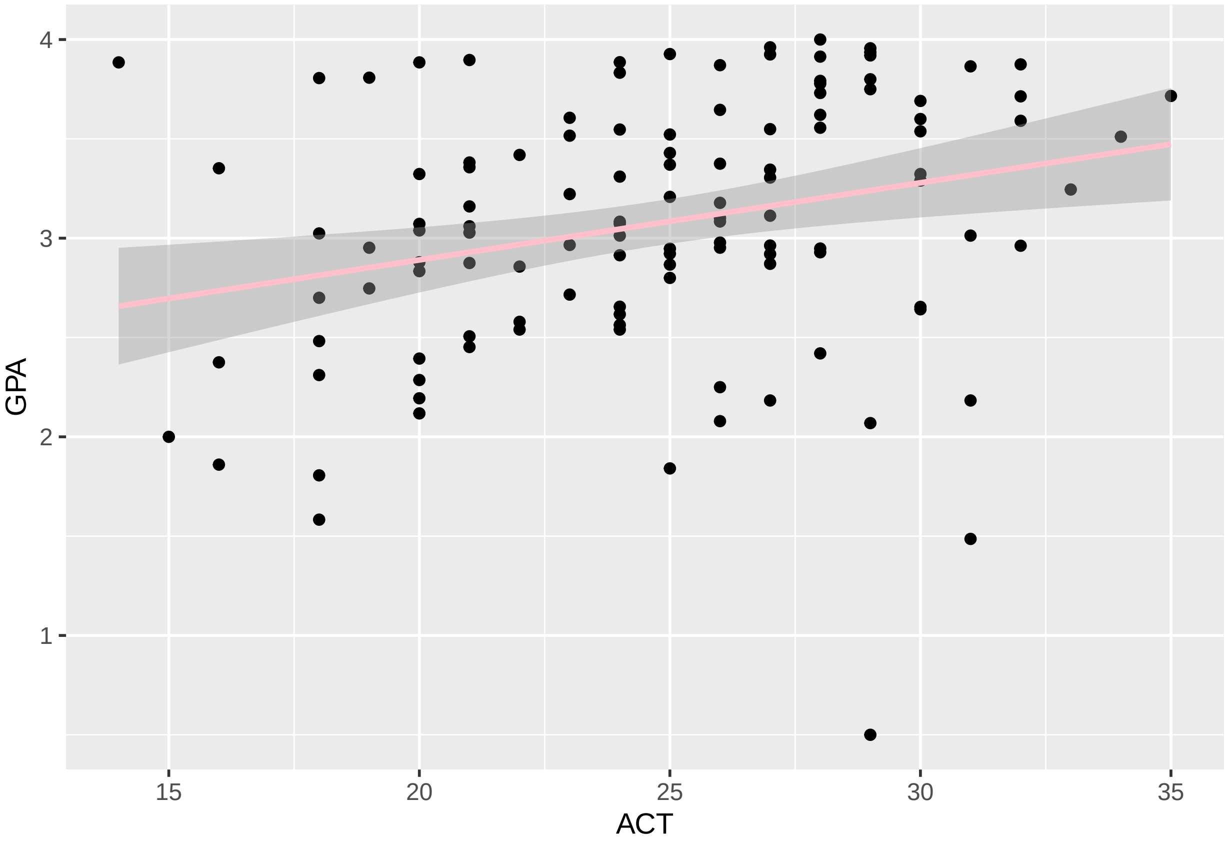
Problem 3

part 1

```
GPA_df=
  read.csv("./data/GPA.csv")
```

```
GPA_df %>%
  ggplot(aes(ACT, GPA)) + geom_point(color='black')  +
  geom_smooth(method='lm', se=TRUE,color="pink")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Hypothesis

- H0:There is no linear association exists between student's ACT score (X) and GPA at the end of the freshman year (Y).beta1=0
- H1:There is linear association exists between student's ACT score (X) and GPA at the end of the freshman year (Y).beta1!=0

Calculations

For the current test, we use the t-test to test whether or not the slope between ACT and GPA is significantly different from 0, if it is, we conclude that there is a linear association exists.

$$(beta1 - 0)/se(beta1)$$

```r
GPAlm=lm(GPA~ACT,data=GPA_df)
```

```r
summary(GPAlm)
```

```
##
## Call:
## lm(formula = GPA ~ 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 ***
## 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(GPAlm)
```

```
## # 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 ACT           0.0388    0.0128      3.04 0.00292
```

```
glance(GPAlm)
```

```
## # 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>
```

Decision

```
qt(0.975,118)
```

```
## [1] 1.980272
```

- **Critical Value**: t~(118,0.975) = 1.980272
- **Decision Rule**:
  - Reject H0: if |t| > t(118,0.975)
  - Fail to reject H0:|t| < t(118,0.975)

Conclusion

- For the current data, we have test stats of 3.040 and p-value of 2.916604e-03 for slope. Since t_stats:3.040 > critical value1.980272, we reject the null and conlcude that there is significant linear association exists between student's ACT score (X) and GPA at the end of the freshman year (Y).

part 2

$$GPA(estimated) = 2.11405 + 0.03883ACT$$

part 3

95% Confidence Interval for slope

$$slope\acute{s}t(118, 0.975) * se(beta1)$$

Calculation by r

```
confint(GPAlm,level=0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept) 1.47859015 2.74950842
## ACT         0.01353307 0.06412118
```

0.03883 ± 1.980272*0.01277302 = (0.01353595,0.06412405)

Conclusion

- With 95% confidence, we estimate that the mean GPA increases by somewhere between 0.01353595 and 0.06412405 for each additional point in ACT.
- the interval does not conclude 0. The director of admissions might be interested in whether the confidence interval includes zero because they are interested in learning about whether a higher ACT scores can be a potential predictor for higher gpa, in this case, it can be a potential predictor.

part 4

```
new_data <- data.frame(ACT=c(28))
predict(GPAlm, newdata=new_data, interval="confidence", level=0.95)
```

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

- **Interpretation**: For people who have ACT score = 28, the expected mean GPA score can vary between 3.061384 and 3.341033.

part 5

```
predict(GPAlm, newdata=new_data, interval="prediction", level=0.95)
```

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

- **Interpretation**: For a new person with ACT score of 28, her or his mean GPA score can vary between 1.959355 to 4.443063

part 6

- prediction interval is wider than the confidence interval because the prediction interval need to account for an additional error term whereas confidence interval don't.
- For confidence interval, we are using expected mean GPA for all ACT that equal to 28, but in prediciton, we are looking at the range for new people with ACT of 28, so prediciton have more error involved.