Stat 021 Homework 7

Suzanne Thornton

Due: Saturday, Nov. 16, 12:00pm

Instructions: A **pdf** version of your homework must be submitted to Gradescope by **noon** on the due date. The course passcode is **MPKJ4Z**. If you are having trouble getting your *.Rmd* file to compile, you need to get help with this **before** the due date.

You are allowed to hand in **only one** late homework assignment throughout the semester. If you need to hand in this particular assignment late, you must let me know via email by noon on the due date.

You are encouraged to study with your peers to help complete the homework assignments but no copying is allowed. If I see that two or more homework assignments are copied, all students involved will receive a grade of 0 on that assignment and will forfeit (perhaps retroactively) the opportunity to hand in a late homework.

Q1) Read the article, "Scientists rise up against statistical significance" at https://www.nature.com/articles/d41586-019-00857-9.

(a) The article claims, "...researchers have been warned that a statistically non-significant result does not 'prove' the null hypothesis (the hypothesis that there is no difference between groups or no effect of a treatment on some measured outcome)." Explain why failing to reject the null hypothesis does not prove that there is no effect. (What does it mean instead?) (1 point)

Solution

Logically, it's more difficult to prove that a statement is true than that it is false, this is because the latter requires only one single counterexample but the former requires establishing a statement in general, wherever it may apply.

Failing to reject the null hypothesis simply means that we haven't yet discovered evidence towards disproving the null hypothesis.

(b) In the graphic "Beware false conclusions", results are shown from two studies: one that found "significant" results, and another that found "non-significant" results. The article claims that it is "ludicrous" to say that the second study found "no association." Briefly explain why this is the case. (1 point)

Solution

Since both confidence intervals are centered around the same point (1.2), the only way they differ can be in whatever the critical value is or in the standard error since the formulas for CIs are always:

$$\text{point estimate} \ \pm \ \text{critical value} \ \times \text{SE(point estimate)}$$

Most likely, both intervals were formed using the same critical value and so they probably only differ in SE, but we know that SE decreases as the sample size increases! This means that the blue CI on the plot probably uses *more* data than the red CI and is therefore *more* informative. In short, it looks like the study that produced the red CI may have cherry-picked the data they used to form the CI (whether or not this was intentional is less clear).

(c) Regarding the same two studies in part (b), the article claims that it is "absurd" to say that the two studies are in conflict, even though one was "significant" and the other was "not significant". Briefly explain why this is the case. (1 point)

Solution

Again, the studies most likely differ based on the size of the sample. Hence it's not that the resulting intervals are in conflict, it's just that we have less information (i.e. data) in the results with a wider interval. As a

general principle, we *prefer* more narrow CIs because they help us concentrate around a more precise region of the parameter space.

(d) In the section titled "Quit categorizing", the article claims that, "Statistically significant estimates are biased... Consequently, any discussion that focuses on estimates chosen for their significance will be biased." Briefly explain why this is the case. (1 point)

Solution

If the only scientific results being shared are the "significant" ones, then we're missing out on a bigger part of the picture of the scientific process. Scientific journals have thus made it seem as if all important discoveries coincide with small p-values (when in fact we know that we are bound to find false discoveries by chance, we just try to control this chance with our significance level, α). Furthermore, this makes it seems as if anything that doesn't produce a small p-value is unimportant. For example, pretend we have many (say 20) out-of-touch researchers studying a relationship between two totally unrelated events. Even if we know that there is no practically significant relationship, most likely there's going to be at least one of these researchers who finds a "statistically significant" relationship (because $1/20 = 0.05$) and that researcher has a better chance of publishing their results than the other 19 researchers who found no significant relationship.

(e) Now that you've read this article, going back to Q1 of HW 6, consider your answer for part (e) and explain if you would like to change your final model suggestion or not and why. (1 point)

Solution

Obviously, your answer here is going to depend on how you answered Q1 of HW 6. Basically, what I'm trying to get at here is that if you ultimately chose the largest model because it had the highest R squared value, you may have noticed that the individual t-tests for the individual significance of different predictors changes quite a lot compared to the smaller models.

Q 2) Create an R function that takes two numerical vectors as it's input and fits a SLR model using the second vector to predict the first one. The output of this function will be a phrase that either says "Good fit" or "Bad fit" depending on the R-squared value of the model. Let's say, for simplicity, that a model is a good fit if the R-squared value is 0.60 or higher. To get you started, you can use the following code as a template:

```
my.SLR.fun <- function(vec1, vec2){

  ## add you own code here

  if(){return("Good fit")}
  ##figure out what logical test needs to go into the if statment

  else{return("Bad fit")}
}

#test your function
y <- rnorm(10,0,2)
x <- rnorm(10,2,1.3)
my.SLR.fun(y,x)
```

Solution

```
my.SLR.fun <- function(vec1, vec2){
  R2 <- summary(lm(vec1 ~ vec2))$adj.r.squared
  if(R2 >= 0.6){return("Good fit")}
  else{return("Bad fit")}
}

#test your function
x1 <- rnorm(10,2,1.3)
```

```
y1 <- rnorm(10,0,2)
## note that neither x1 nor y1 are inherently related to each other,
## they both just happen to come from Gaussian distribution
my.SLR.fun(y1,x1)

x2 <- rnorm(10,2,1.3)
y2 <- x2 + rnorm(10,0,0.8)
##note that the way y2 was created was so that it IS linearly dependent on x2
my.SLR.fun(y2,x2)
```

Q 3) The dataset uploaded to Moodle called "airplanes.csv" was collected from national publication advertising the sale of used aircraft in the early 1990s. The variables included in this dataset include the year of the aircraft, TT (total flight time in hours), SMOH (hours since major overhaul), DME (distance measuring equipment), LORAN (long-range navigation based on satellite communication), HP (engine horsepower), paint (new or recent paint job), and price. The variables DME, LORAN, and paint are each binary categorical variables that indicate whether the corresponding item was mentioned as being present in the ad. The price is given in thousands of dollars.

(a) Read in the dataset from Moodle and be sure to double check that each of the variable types are what you want them to be. (Note: Please do **not** print out the entire data set in your final pdf document, instead just double check the data import process on the side on your own. Also, double check the data after reading it into R, get rid of any rows of only NA values and make sure each variable is the correct variable type.) Make a scatterplot matrix using only the numerical variables. Comment on any notable features or patterns (or lack of thereof). What is notable about the variable HP? What do you think you should do with this variable? (2 points)

Solution Looking at the matrix scatterplot, we immediately see that the variable HP takes on the same value for every data point. Because if this, there's not much discerning information contained in this variable and we will omit it in our model building.

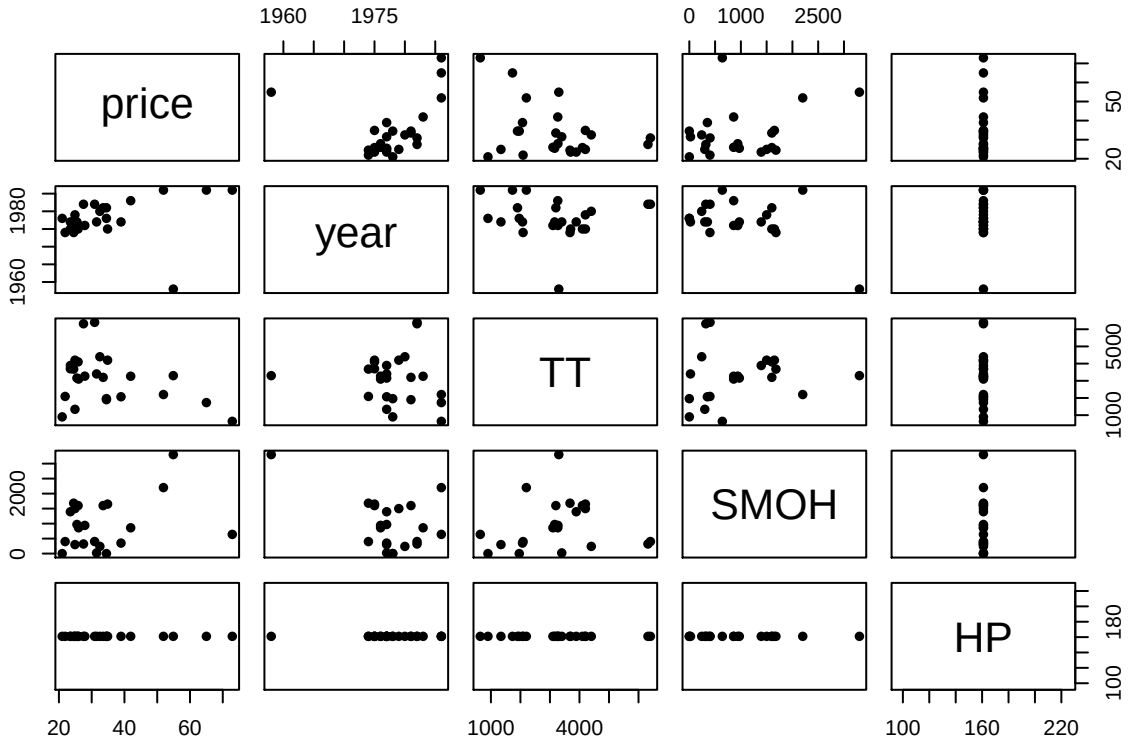
```
#After looking at the csv file and reading the description of the data, we make
#sure that R reads the variables DME, LORAN, and paint as factors and the other
#numerical variables are of type double
plane_data <- read_csv("~/Google Drive Swat/Swat docs/Stat 21/Data/airplanes.csv",
                       col_names=TRUE, cols(DME = col_factor(),
                                            LORAN = col_factor(),
                                            paint = col_factor()) )

#after we import the data into R, we look at the entire data object (only the last
#few rows are shown here for illustrative purposes) and notice that the last row
#correspons to all NA entries so we get rid of it
tail(plane_data)
```

```
## # A tibble: 6 x 10
##   price  year    TT  SMOH   IFR DME   LORAN ModeC    HP paint
##   <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <dbl> <dbl> <fct>
## 1  34.8  1975  4200  1650     1 1     1         1   161 0
## 2  25.8  1975  4100  1600     0 0     1         1   161 0
## 3  26    1976  3100   860     1 0     0         0   161 1
## 4  24.9  1979  4200  1500     1 0     0         1   161 0
## 5  21    1978   900     0     0 0     0         0   161 0
## 6  NA      NA    NA    NA    NA <NA>  <NA>     NA    NA <NA>
```

```
plane_data <- plane_data[-26,]
plane_data %>% select(price, year, TT, SMOH, HP) %>%
               pairs(labels=c("price", "year", "TT",
```

```
                    "SMOH", "HP"), pch = 16)
```

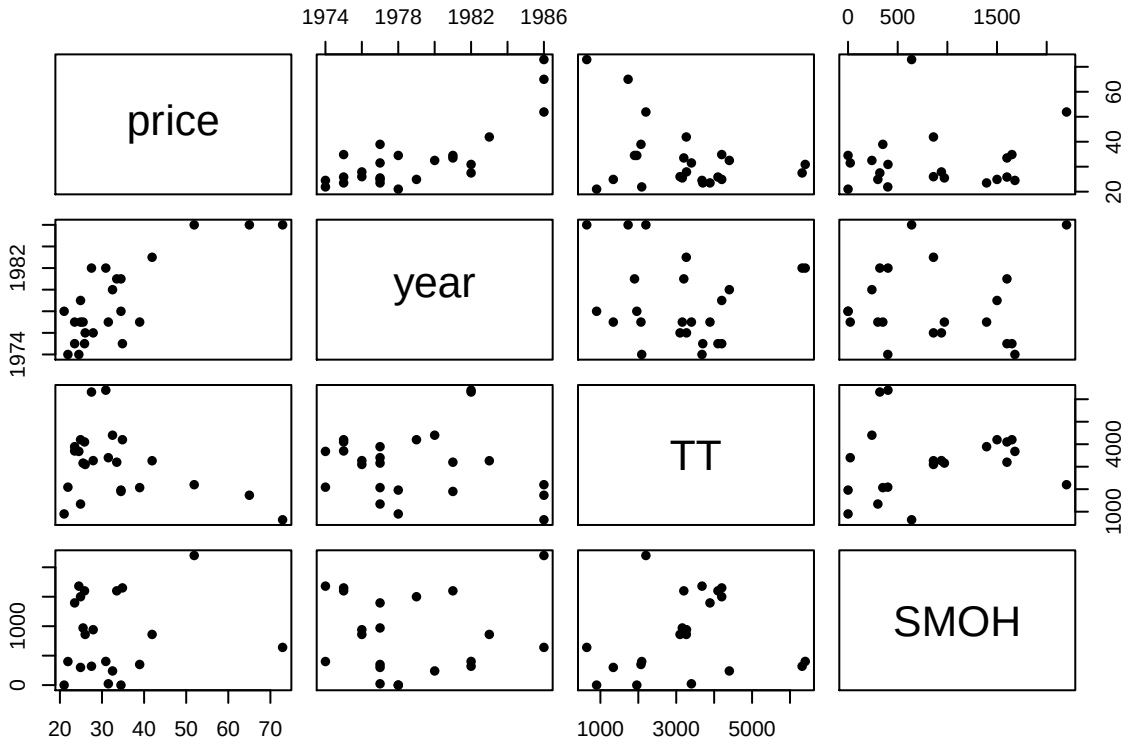


(b) Which data point(s) appear(s) to be an outlier? Can you guess why it might be an outlier? For the purposes of this assignment, let's delete this point from any further analyses since we don't know its correct value. Delete this outlier observational unit from the data and make a new scatterplot matrix of the quantitative variables. (1 point)

Solution

The point corresponding to the year 1958 is clearly far away from the rest of the data with respect to the year variable. It's possible that this is a typo since the rest of the data is from the 70s and 80s but we don't know this for sure.

```
plane_data2 <- plane_data %>% filter(year > 1960)
plane_data2 %>% select(price, year, TT, SMOH) %>%
                        pairs(labels=c("price", "year",
                                       "TT", "SMOH"), pch = 16)
```



(c) Fit a model with all of the predictor variables except SMOH. What is the value of R-squared? What is the estimated error variance? What is the interpretation of the latter? Which variables appear significant, and which do not? (2 points)

Solution

```
MLR_plane <- lm(price ~ year + TT + DME + LORAN + paint,
data = plane_data2)
summary(MLR_plane)
```

```
##
## Call:
## lm(formula = price ~ year + TT + DME + LORAN + paint, data = plane_data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6482 -1.5775  0.0729  1.9529  9.5815
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.808e+03  5.746e+02 -10.108 7.57e-09 ***
## year         2.955e+00  2.898e-01  10.199 6.58e-09 ***
## TT          -1.910e-03  6.415e-04  -2.977  0.00808 **
## DME0        -7.020e+00  2.475e+00  -2.836  0.01096 *
## LORAN1       1.196e+01  2.351e+00   5.089 7.66e-05 ***
## paint1       6.125e+00  2.685e+00   2.282  0.03491 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.32 on 18 degrees of freedom
## Multiple R-squared:  0.9153, Adjusted R-squared:  0.8917
## F-statistic: 38.89 on 5 and 18 DF,  p-value: 4.932e-09
```

The adjusted R-squared value is 0.8917286 and the estimate of the variance of the random errors is $\hat{\sigma}^2 = 4.3198483^2 = 18.6610891$. The estimate of the error variance represents how spread out we expect the random noise affiliated with the plane price to be. Based on the results of the individual t-tests for the predictor variables, it looks like every variable is "statistically significant".

(d) It seems plausible that TT and year might be collinear: older planes may have been flown more. Based on the scatterplot matrix and the estimated correlation between these variables, does this appear to be the case? (1 point)

Solution

```
cor(plane_data2$TT, plane_data2$year)
```

```
## [1] -0.12746
```

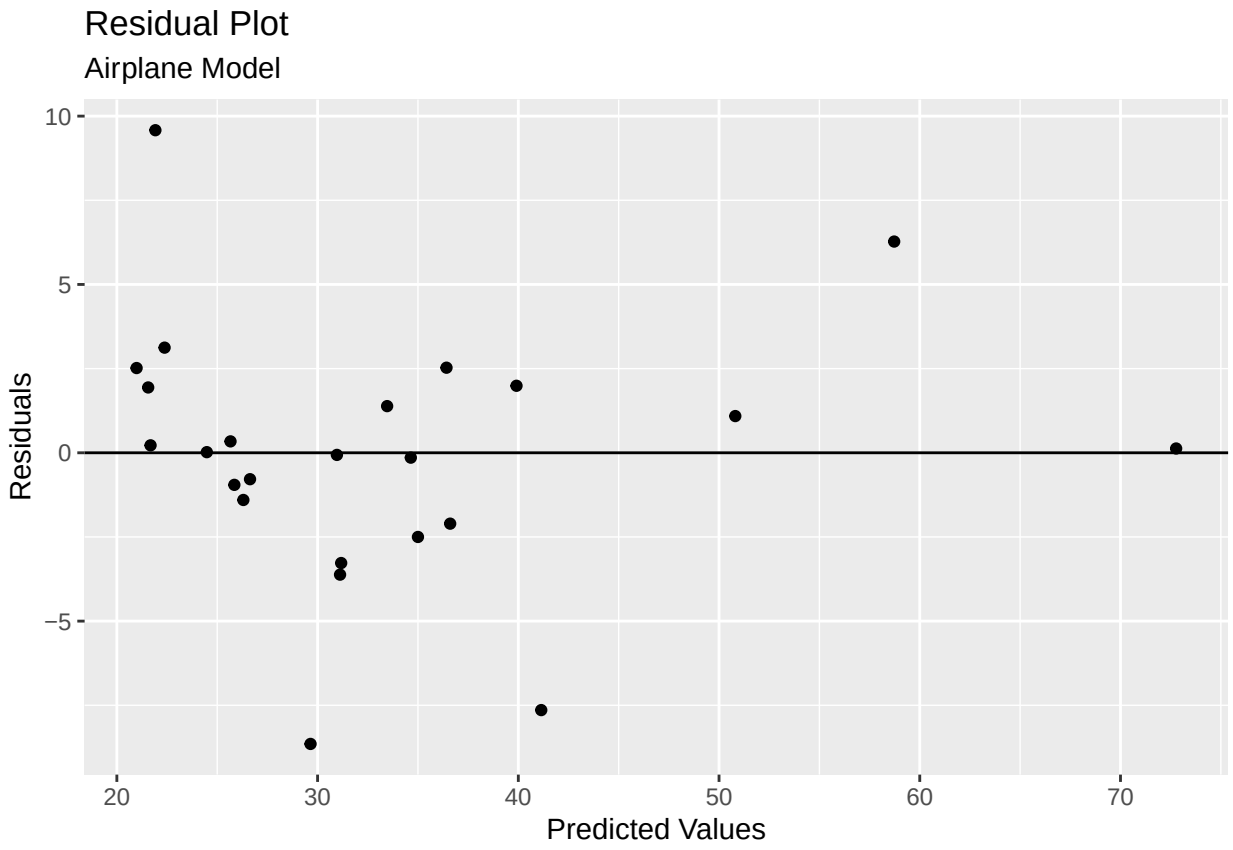
With a correlation of about -0.13, (and based on the lack of apparent linear trend between these two predictor variables), despite the non-statistical reason for thinking there may be collinearity present, the data does not seem to indicate that this is actually a problem.

(e) For our current model, do the regression assumptions appear to be satisfied? Make a residual plot (residuals vs predicted) and an Normal probability plot and comment on whether you think the assumptions are satisfied or whether there may be cause for concern. (2 points)

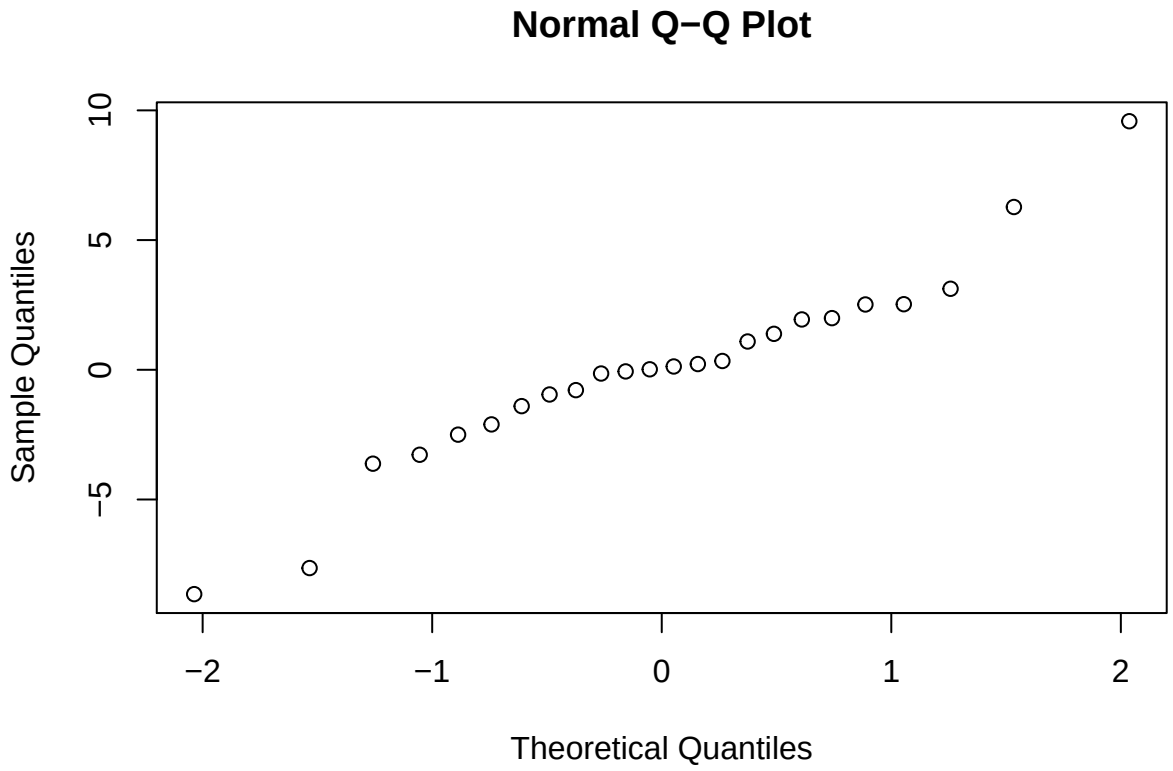
Solution

```
MLR_plane_sum <- summary(MLR_plane)
resid_plot_data <- plane_data2 %>%
                   mutate(residuals = MLR_plane_sum$residuals,
                          fitted_vals = MLR_plane$fitted.values)

ggplot(resid_plot_data) +
  geom_point(aes(x=fitted_vals, y=residuals)) +
  labs(title="Residual Plot", subtitle = "Airplane Model",
       x="Predicted Values", y="Residuals") +
  geom_hline(yintercept=0)
```




```
qqnorm(y=resid_plot_data$residuals)
```



In the residual plot we see some indication of non-constant variance and the data isn't evenly scattered about

the horizontal axis. Based on this residual plot, it doesn't look like our linear model is very reliable as is. Something we could try however that may fix both the non-linearity and possible the non-constant spread is to try some variable transformations (on both the predictors and the response). Furthermore, the tails of the QQ-plot are a bit too heavy to be normally distributed so the results of (say) confidence intervals for the mean response, will probably be overly optimistic (i.e. more narrow than they really should be) and thus could bias our conclusions toward false positives.

(f) Summarize your findings from this model. In particular, what do the regression coefficients mean? What is the estimated error variance and what does it represent? What is the R-squared? (2 points)

Solution (The last two questions are redundant and the solutions are the same as those in part (c), my bad!) The meaning of the regression coefficients based on this model are:

- From one year to the next, the **average** price of these planes increases by about 2960 dollars;
- For every additional hour the plan has spent traveling, we'd **expect** the price to decrease by about 1.9 dollars;
- The difference in the price of a plane with DME versus without DME is **on average** 7020 dollars;
- The difference in the price of a plane with LORAN versus without LORAN is **on average**, 12K dollars; and
- The **average difference** in the price of a plane with a paint job versus without a paint job is 6.13K dollars.

You have to be super careful here and realize that different categorical variables may have different reference levels. For example, since I did not specify how R should factor my categorical variables (for reasons unbeknownst to me) R automatically choose level 1 as the reference category for *DME*, level 0 as the reference category for *LORAN*, and level 0 as the reference category for *paint.* Also, the words that are bolded are of critical importance, non of these estimations is exact, rather its about central tendencies.