# Homework 7

*Alyssa Ogle*

*11/16/2019*

**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 failing to reject the null hypothesis really mean instead? (1 point) We fail to reject the null hypothesis, because we do not have evidence to support the conclusion that there is an effect. A lack of evidence for an effect does not mean that the effect does not exist

(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) It was ridiculous to make this assertion, because the confidence interval from the study that had the conclusion of "no association" included values that indicated that there was a serious risk increase while using the anti-inflammatory drugs. Confidence intervals contain values that can be the population mean, so it does not make sense to assert that there is no association when this is the case.

(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) It was ridiculous, because the two studies were both examining the unintended effects of anti-inflammatory drugs and both studies arrived at similar conclusions, as they found a risk ratio of 1.2. The key difference between the two studies was the size of the intervals, as the statistically significant study had a smaller confidence interval that spanned from 9% to 33% rather than 3% to 48%.

(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) There is a bias regarding the "realness" of data, because a study that is statistically significant can be deemed as having results that are mrore "real" than results from a study that is not statistically significant, even though that is not necessarily the case.

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

Response from Homework 6, part e: The terms that should be included in the final regression model are displacement, transmission type, and an interaction variable of transmission type and mpg. These terms should be included because Part B had these terms and also the higest adjusted R squared value.

I would not change my final model suggestion, because the statistically signficance of the predictor variables was not my determining factor in what values to include. I used the r-squared value to decide which model to use and because the model from Part B had the largest r-squared value, I would still use this model.

**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){
  func_data <- data.frame(vec1, vec2)
```

```r
  slr_data <- lm(vec1~vec2, data = func_data)
  slr_data
  r_value <- summary(slr_data)$r.squared
  if(r_value >= .6)
  {
    return("Good fit")
  }
  else
  {
    return("Bad fit")
  }
}


#Should return "Bad fit"
x <- rnorm(10,2,1.3)
y <- rnorm(10,0,2)
my.SLR.fun(y,x)
```

```
## [1] "Bad fit"
```

```r
#Should return "Good fit"
x <- rnorm(10,2,1.3)
y <- x + rnorm(10,0,0.8)
my.SLR.fun(y,x)
```

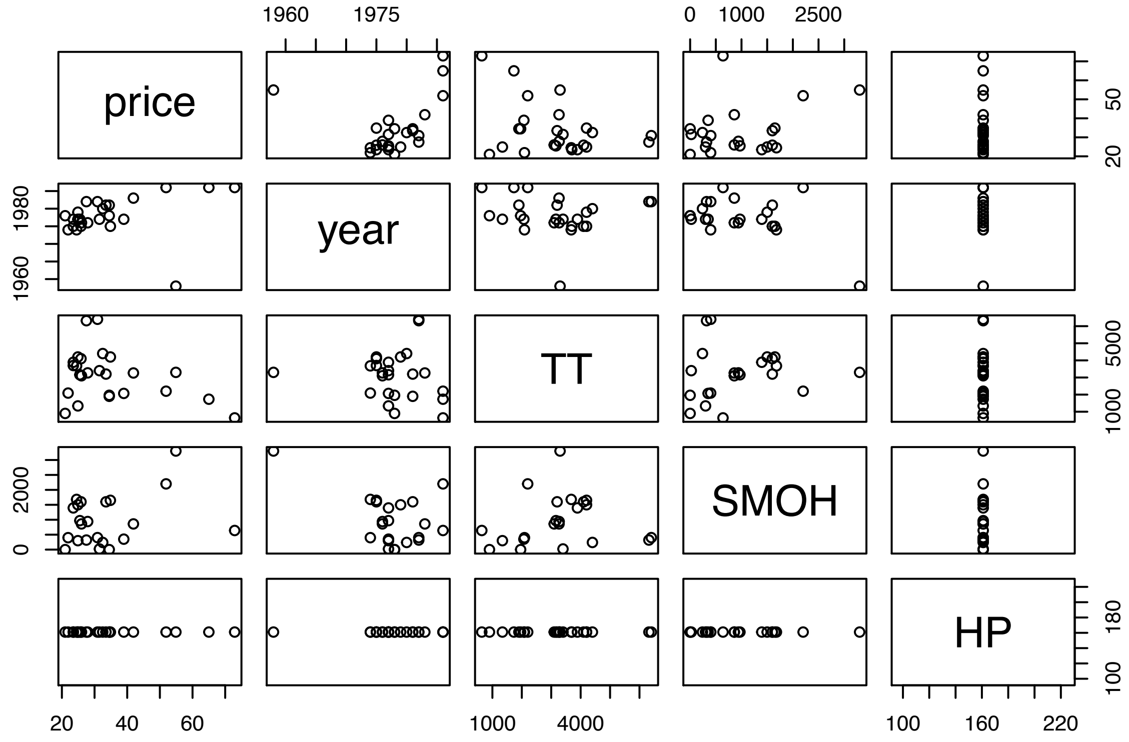
```
## [1] "Good fit"
```

**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)

```r
airplane <- read.csv("airplanes.csv")
airplane <- airplane[c(-26),]

airplane$LORAN <- as.factor(airplane$LORAN)

airplane$DME <- as.factor(airplane$DME)
airplane$paint <- as.factor(airplane$paint)

airplane %>% select(-c(IFR, DME, LORAN, ModeC, paint)) %>% pairs(pch=1)
```

```
## Warning: `lang()` is deprecated as of rlang 0.2.0.
## Please use `call2()` instead.
## This warning is displayed once per session.
## Warning: `new_overscope()` is deprecated as of rlang 0.2.0.
## Please use `new_data_mask()` instead.
## This warning is displayed once per session.
```

```
## Warning: `overscope_eval_next()` is deprecated as of rlang 0.2.0.
## Please use `eval_tidy()` with a data mask instead.
## This warning is displayed once per session.
```



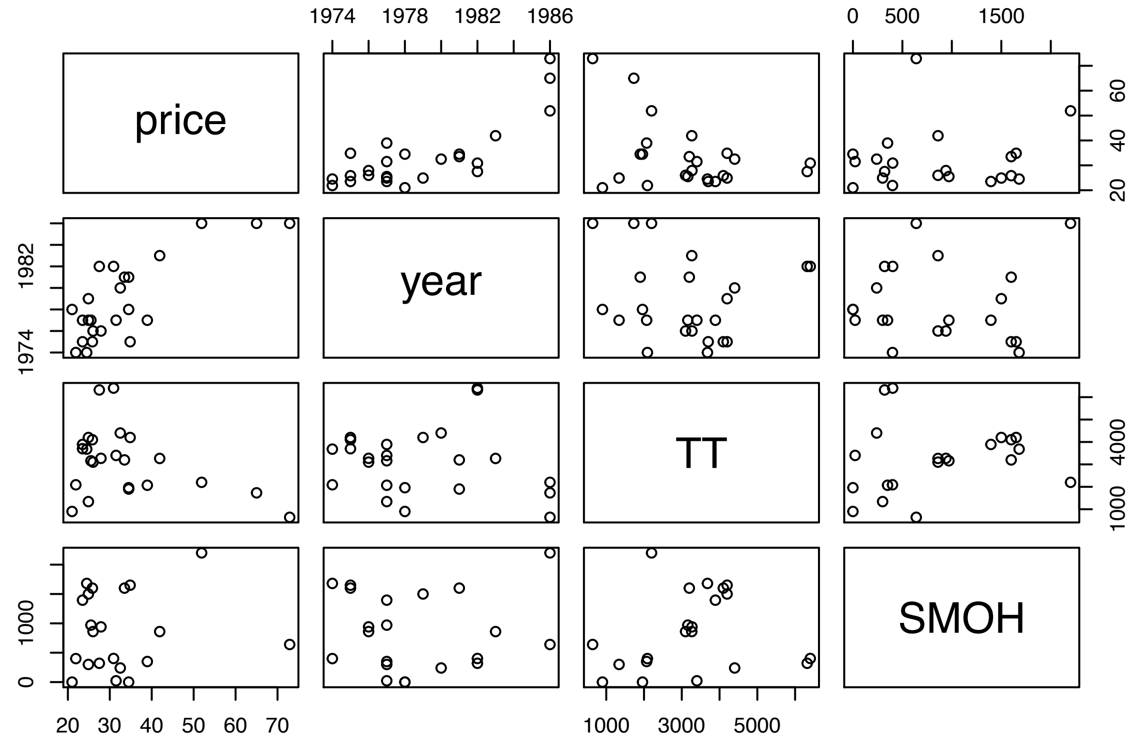
Comment on any notable features or patterns (or lack of thereof). What is notable about the variable HP? What do you think you should do with this variable? (2 points)

The price and year graph has a positive linear relationship. The price and TT grap had a somewhat positive linear relationship, but as the units increase a slight curve downwards starts to appear in the plot The SMOH and price scatterplot has a somewhat positive linear relationship, but has a funneled shape as well. The TT and SMOH graph shows a weak positive, linear relationship as it is funnelled shape. The year and TT graph does not demonstrate a strong linear relationship, as the points looked randomly plotted, though a weak positive, linear relationship can be discerned. All the graphs that include the variable HP are straight lines which indicates that this variable should probably not be included in the scatterplot matrix, because its lack of variation is uninformative.

(b) Which data point appears 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)

The data point from year 1954 appears to be an outlier. It may be an outlier because it is remarkably older than the older values, which would influence other variables such as TT, SMOH, and price.

```
airplane <- airplane[c(-13),]


airplane %>% select(-c(IFR, DME, LORAN, ModeC, paint, HP)) %>% pairs(pch=1)
```



(c) Fit a model with all of the predictor variables except SMOH.

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

```
##
## Call:
## lm(formula = price ~ year + TT + DME + LORAN + paint, data = airplane)
##
## 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) -5.815e+03  5.738e+02 -10.134 7.27e-09 ***
## year         2.955e+00  2.898e-01  10.199 6.58e-09 ***
## TT          -1.910e-03  6.415e-04  -2.977  0.00808 **
## DME1         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
```

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)

The value of R-squared is .9153. The estimated error variance is 18.6624. The estimated error variance represents the distribution of the random noise in the data. All of the variables appear to be significant at an alpha level of .05

(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)

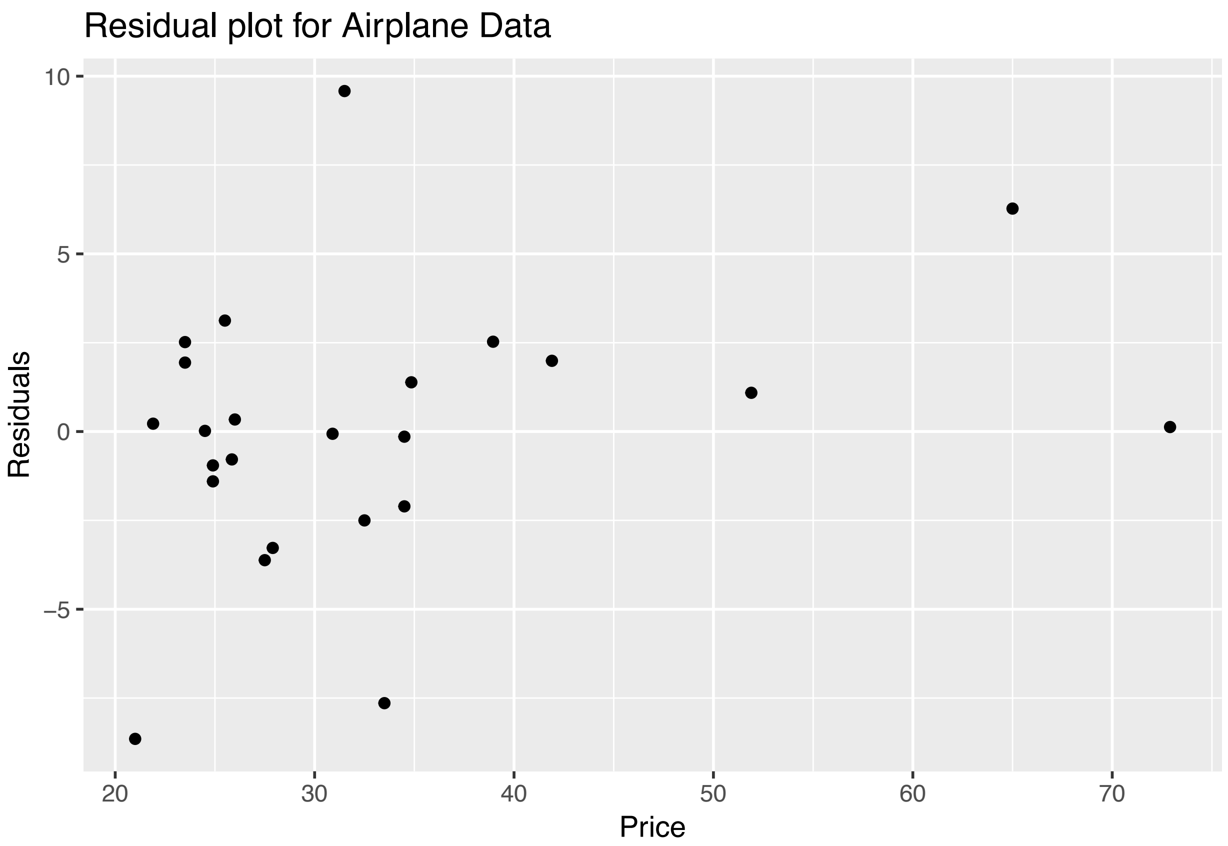
```
cor(airplane$TT, airplane$year)
```

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

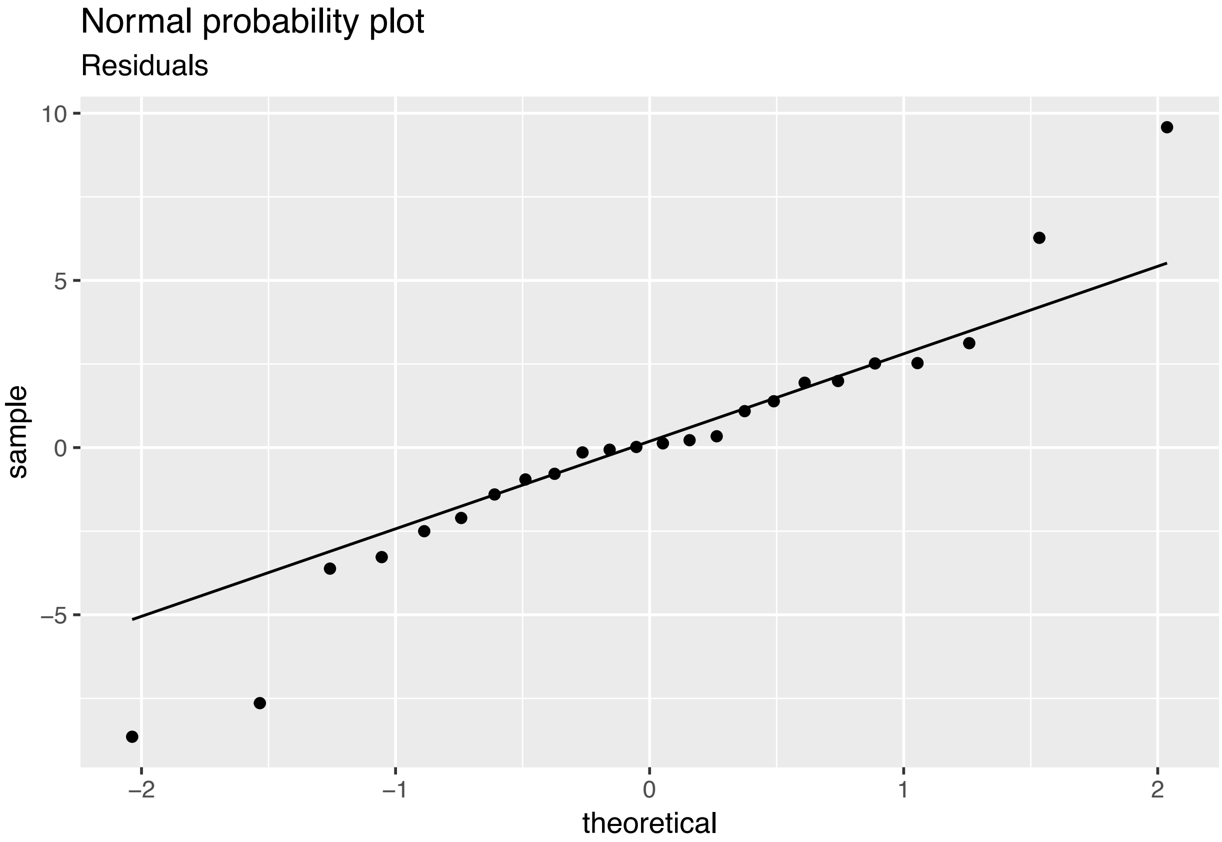
Because of the low correlation, it does not seem that TT and year are collinear, even though the scatterplot does show a somewhat strong, positive linear relationship (it is only somewhat strong because it has a slight curve downwards)

(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)

```
airplane$residuals <- residuals(LM_airplane)


res_plane <- ggplot(airplane) +
  geom_point(aes(x=price, y=residuals)) +
  labs(title="Residual plot for Airplane Data", x= "Price", y= "Residuals")
res_plane
```



```
plot1 <- ggplot(airplane, aes(sample = residuals))
plot1 + stat_qq() + stat_qq_line() + labs(title = "Normal probability plot", subtitle = "Residuals")
```



```
#looking at constant variance and linearity
```

The assumptions for the linear model are somewhat met. The normal probability plot indicates that there is a linear relationship with some outliers at the end. It does not seem like there is constant variance in the data set, the residual plot clearly has heteroskedasticity instead of having the random pattern expected from data sets with constant variance.

(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)

$\beta_0$ is -5.815e+03 and represents the average expected change in the response variable, price, with the reference categorical variable or when the x values of all the catergorical variables = 0.. $\beta_1$ is 2.955e+00 and represents the average expected change in price when the predictor variable, year, increases by one unit. $\beta_2$ is -1.910e-03 and represents the average expected change in price when the predictor variable, TT, increases by one unit. $\beta_3$ is 7.020e+00 and represents the expected change in y when the variable DME is changed from its ref category to category 1. $\beta$43 is 7.020e+00 and represents the expected change in y when the variable DME is changed from its ref category to category 1. $\beta_4$ is 1.196e+01 and representes the expected change in y when the variable LORAN is changed from its ref category to category 1. $\beta_5$ is 1.196e+01 and representes the expected change in y when the variable paint is changed from its ref category to category 1. The value of R-squared is .9153. The estimated error variance is 18.6624. The estimated error variance represents the distribution of the random noise in the data. All of the variables appear to be significant at an alpha level of .05