Stat 230 - A review of Simple Linear Regression (SLR) in R - HANDOUT- SOLUTION

P.B. Matheson adapted from A.S. Wagaman

Feb 9, 2022

R Markdown notes: We install the readr package. The goal of 'readr' is to provide a fast and friendly way to read rectangular data (like 'csv', 'tsv', and 'fwf'). It is designed to flexibly parse many types of data found in the wild, while still clearly failing when data unexpectedly changes. We also install the broom package. The broom package takes the messy output of built-in functions in R, such as lm , nls , or t. test , and turns them into tidy tibbles.

SLR Review - CHOOSE-FIT-ASSESS-USE - How is price related to the quality of food in NYC?

The Zagat guide contains restaurant ratings and reviews for many major world cities. Assume that you are heading to New York City, and want to take your significant other out for a meal. Being a good analyst, you do some research prior to the trip. To understand variation in the average *Price* of a dinner in Italian restaurants in New York City, we want to know how customer ratings (measured on a scale of 0 to 30) of the *Food* is associated with the average *Price* of a meal. The data contains ratings and prices for 168 Italian restaurants in 2001.

```
NYC <- read_csv("https://pmatheson.people.amherst.edu/nyc.csv")
summary(NYC) #provides a basic summary -
```

```
##       ...1             Case         Restaurant           Price
##  Min.   :  1.00   Min.   :  1.00   Length:168         Min.   :19.0
##  1st Qu.: 42.75   1st Qu.: 42.75   Class :character   1st Qu.:36.0
##  Median : 84.50   Median : 84.50   Mode  :character   Median :43.0
##  Mean   : 84.50   Mean   : 84.50                      Mean   :42.7
##  3rd Qu.:126.25   3rd Qu.:126.25                      3rd Qu.:50.0
##  Max.   :168.00   Max.   :168.00                      Max.   :65.0
##       Food          Decor          Service          East
##  Min.   :16.0   Min.   : 6.00   Min.   :14.0   Min.   :0.000
##  1st Qu.:19.0   1st Qu.:16.00   1st Qu.:18.0   1st Qu.:0.000
##  Median :20.5   Median :18.00   Median :20.0   Median :1.000
##  Mean   :20.6   Mean   :17.69   Mean   :19.4   Mean   :0.631
##  3rd Qu.:22.0   3rd Qu.:19.00   3rd Qu.:21.0   3rd Qu.:1.000
##  Max.   :25.0   Max.   :25.00   Max.   :24.0   Max.   :1.000
```

```
   #can be lengthy and providing means/sds for non quantitative data (see EAST) is wrong
   #just because it's in the output doesn't mean it's correct.  YOU MUST BE THE JUDGE OF THAT.
glimpse(NYC) #quick glimpse
```

```
## Rows: 168
## Columns: 8
## $ ...1       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
## $ Case       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ~
```

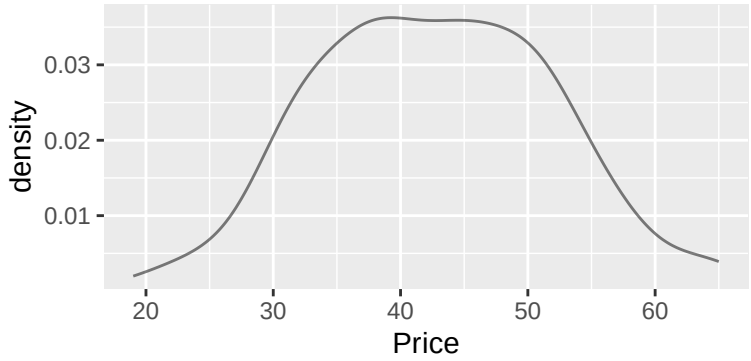
```
## $ Restaurant <chr> "Daniella Ristorante", "Tello's Ristorante", "Biricchino", ~
## $ Price      <dbl> 43, 32, 34, 41, 54, 52, 34, 34, 39, 44, 45, 47, 52, 35, 47,~
## $ Food       <dbl> 22, 20, 21, 20, 24, 22, 22, 20, 22, 21, 19, 21, 21, 19, 20,~
## $ Decor      <dbl> 18, 19, 13, 20, 19, 22, 16, 18, 19, 17, 17, 19, 19, 17, 18,~
## $ Service    <dbl> 20, 19, 18, 17, 21, 21, 21, 21, 22, 19, 20, 21, 20, 19, 21,~
## $ East       <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
```

PRELIMINARY ANALYSIS: We should check basic univariate descriptive statistics and the correlation between Food and Price before fiting the linear regression model.

```
favstats(~ Price, data = NYC)
```

```
##  min Q1 median Q3 max     mean       sd   n missing
##   19 36     43 50  65 42.69643 9.292814 168       0
```

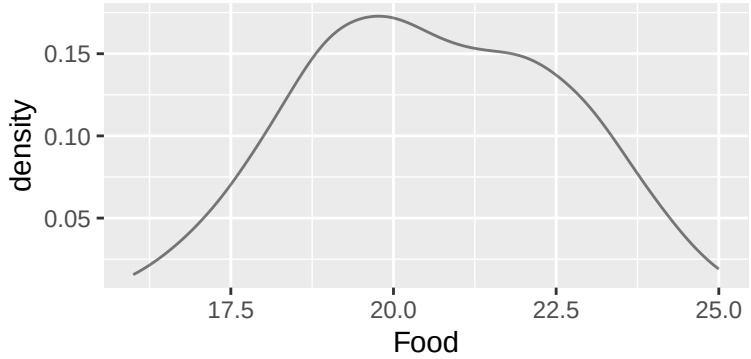
```
gf_dens (~Price, data=NYC)
```




```
favstats(~ Food, data = NYC)
```

```
##  min Q1 median Q3 max     mean       sd   n missing
##   16 19   20.5 22  25 20.59524 1.982674 168       0
```

```
gf_dens (~Food, data=NYC)
```




```
cor(Price ~ Food, data = NYC, use = "pairwise")
```

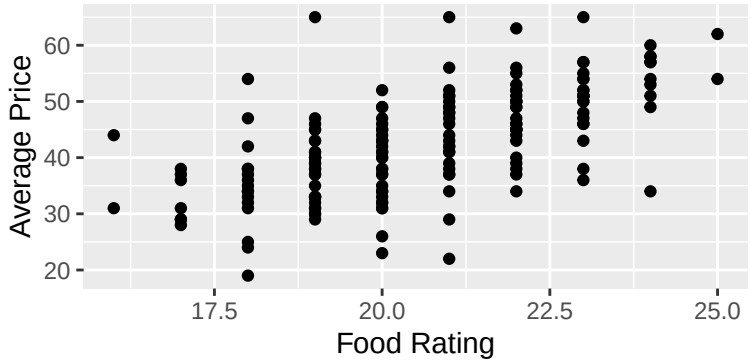
```
## [1] 0.6270435
```

1. Describe the data for price and food based on shape, center and spread > ANSWER Both price and food ratings appear to be relatively normally distributed so we'll use the mean and sd to describe them. Mean price = $42.7, sd=$9.3, n=168 Mean food rating =20.6, sd=1.98, n=168
2. Does the correlation agree with what you observed in the scatterplot?

ANSWER Yes there is a moderate positive linear relationship with no influential outliers or heteroscedasticity. There is one outlier with a high price (>$60) but a below average rating on food (<20) but it does not appear to exert too much leverage. The granularity in the scatterplot is caused by the method of ratings (only integers between 0-30 are used). With a correlation coefficient of .627, there appears to be agreement.

CHOOSE To determine if simple linear regression is appropriate, we return to the scatterplot. There are many ways to generate scatterplots in R, but we are using ggformula syntax here adding a regression line and labels.

```
#basic scatterplot with labeled axes
gf_point(Price ~ Food, data = NYC) %>%
  gf_labs(x = "Food Rating", y = "Average Price")
```



3. Does linear regression appear to be appropriate?

   ANSWER Yes, at first pass from a visual standpoint we can proceed. We will need to use the LINE pnuemonic to consider during the ASSESS phase.

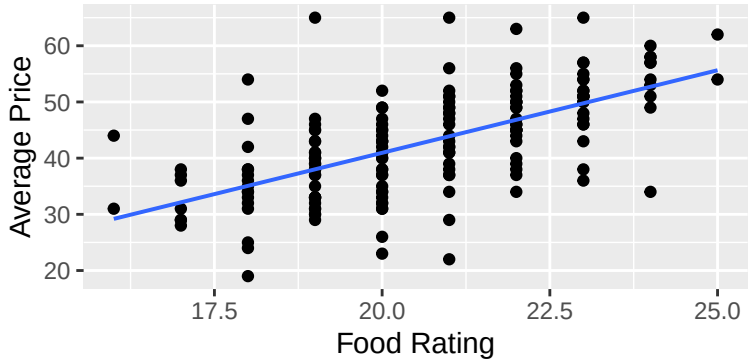
Linearity - Assess via a scatterplot of X and Y

Independence - Read the research design/methods to see if it observations were independent (randomly selected)

Normality - Assess the distribution of errors with histograms/qqplots of residuals errors should be centered at zero (ZERO MEAN), no skew or pattern (RANDOM)- necessary for inference

Equal (Constant) Variance - Assess with a residual vs. fitted plot. Looking for no pattern

```
#scatterplot with regression line
gf_point(Price ~ Food, data = NYC) %>%
  gf_labs(x = "Food Rating", y = "Average Price") %>%
  gf_lm()
```



Let's fit the linear model. We have to get information about the model coefficients, error and variance accounted for.

```
fm <- lm(Price ~ Food, data = NYC)
#fm is just a name we gave to the model
#you can change the name from fm (fitted model) but need to be consistent throughout code
msummary(fm) #reports the summary; msummary and summary are very similar
```

FIT

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.8321     5.8631  -3.041  0.00274 **
## Food          2.9390     0.2834  10.371  < 2e-16 ***
##
## Residual standard error: 7.261 on 166 degrees of freedom
## Multiple R-squared:  0.3932, Adjusted R-squared:  0.3895
## F-statistic: 107.6 on 1 and 166 DF,  p-value: < 2.2e-16
```

The typical output contains the slope and intercept estimates as well as their standard errors,associated t-statistics, and p-values assuming conditions for inference are met. We also see the R squared and residual standard error reported below the coefficients box.

We can get confidence intervals for the regression coefficients, those are obtained with:

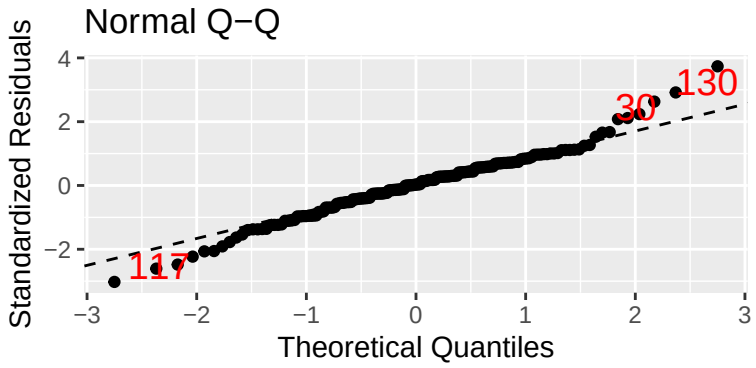
```
confint(fm, level = 0.95) #you can adjust the level
```

```
##                 2.5 %    97.5 %
## (Intercept) -29.408044 -6.256253
## Food          2.379464  3.498455
```

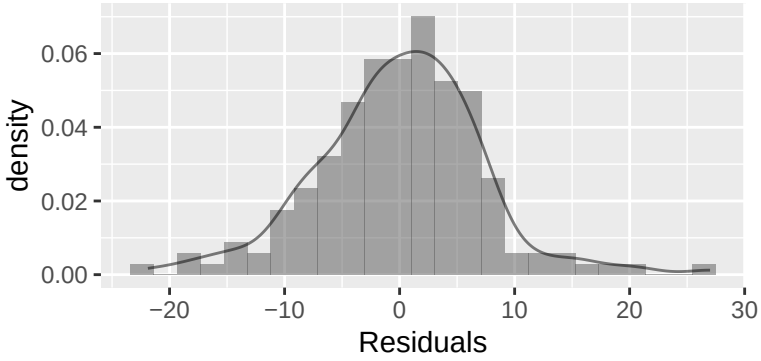
Let's see how well does our model fit the data? Time to turn to the ASSESS phase.

ASSESS Start with the normality condition of the errors.

```
#easiest way to get QQplot for residuals
mplot(fm, which = 2)
```




```
#generates a histogram with fitted density curve
gf_dhistogram(~ residuals(fm), xlab = "Residuals") %>%
  gf_dens()
```



4. What do these plots suggest to you about whether this condition is satisfied? Are the errors normally distributed? Centered at zero and show no pattern?

   ANSWER The histogram is relatively normal and centered at zero (the latter condition is always true for least squares technique). There is one high resid >20 that may need to be considered. All the other residuals fall between +/- 20.

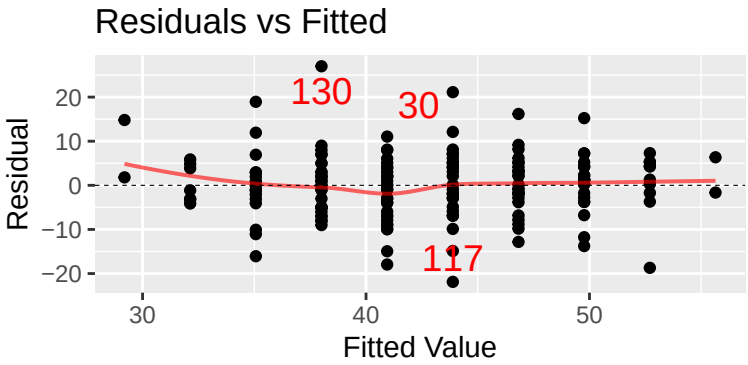
What you need to look at in QQ Plots is whether the points are on the straight line going from bottom left to top right. When deviations occur, they are often located at the lower or higher end of the line, whereas deviations in the middle are less likely. If you see any type of an S form, an exponential curve, or another shape than a straight line, you have a problem. To the extent most of the points are on the line, there is no real shape and only a few are off at the ends, you should be ok.

Next, we can check the condition of equal/constant variance of errors (as well as sometimes seeing issues with linearity) by checking the residual versus fitted plots.

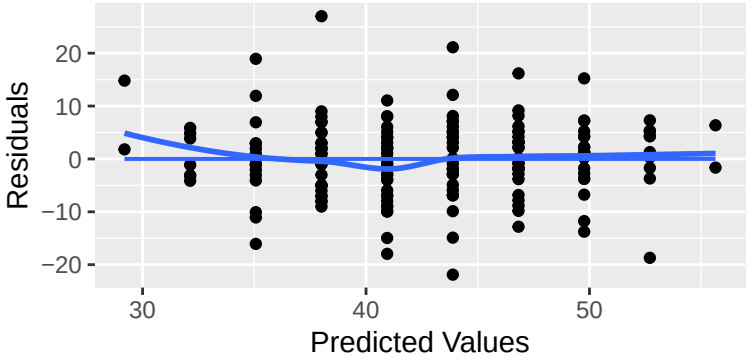
You can make the residual vs. fitted plots multiple ways (2 are shown below)

```
#easiest way
mplot(fm, which = 1)
```

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




```
#Plot using Y~X format
gf_point(residuals(fm) ~ fitted(fm)) %>%
  gf_labs(x = "Predicted Values", y = "Residuals") %>%
  gf_lm() %>% gf_smooth(method = "loess", se = FALSE)
```



5. What do these plots suggest to you about whether the constant variance and linearity conditions are satisfied?

   ANSWER A few outliers are noted in red but nothing exerting a great deal of leverage (unduly influencing the overall relationship). There is no pattern of errors (looks like a cloud) and is mostly linear. IF the variance in Y was not equal across X we could see a fan shape indicating heteroscedastcity which would need to be resolved with a transformation.

6. Which conditions are we unable to check with graphs?

   ANSWER Independence - Are the observations used independent? There is no reason to believe that there are any relationships between the restaurants in the sample. One restaurant own changing prices would probably not affect another restaurant owner. Random - Were the units/people in the study randomly selected? The study description does not tell us how the 168 restaurants were selected. If it was random we can make inferences going forward.

USE Let's interpret the regression coefficients. You should be comfortable interpreting the slope and intercept from Stat 101.

The slope here of 2.939 says that we expect the average Price of a meal at Italian restaurants to increase by 2.939 dollars for each one additional rating point.

7. Would you want to interpret the intercept of -17.832?

   ANSWER No, while it is an actual value (price predicted) in the regression equation, it is not meaningful. The interpretation would be that if a restaurant had a food rating (X) of 0, they would pay us $17.83 to eat there!

Let's try obtaining a predicted value from the model? There are multiple ways to do this, but the easiest way is to create a function that computes it.

```
#creates a function that computes predicted values
fit.price <- makeFun(fm)  #fit.price is a name we gave to the function, you can call it anything you wa
 #just stay consistent throughout the R code when referring to this model.
#obtain the predicted value of Price when Food is 23 - USE
fit.price(Food = 23)
```

```
##        1
## 49.76393
```

The *augment* function in the broom package is another way we can get predicted values easily IF the value of interest for a variable is already in the data set. *augment* saves fitted values and residuals for the model and the data set used (which removes observations with missing values for the variables used) to a new data set that you can look at.

```
augmentNYC <- augment(fm)
head(augmentNYC)
```

```
## # A tibble: 6 x 8
##   Price  Food .fitted  .resid    .hat .sigma     .cooksd .std.resid
##   <dbl> <dbl>   <dbl>   <dbl>   <dbl>  <dbl>       <dbl>      <dbl>
## 1    43    22    46.8 -3.82   0.00896   7.28 0.00127       -0.529
## 2    32    20    40.9 -8.95   0.00649   7.25 0.00499       -1.24
## 3    34    21    43.9 -9.89   0.00620   7.24 0.00582       -1.37
## 4    41    20    40.9  0.0530 0.00649   7.28 0.000000175    0.00732
## 5    54    24    52.7  1.30   0.0236    7.28 0.000395       0.181
## 6    52    22    46.8  5.18   0.00896   7.27 0.00232        0.716
```

Here, to find a prediction for when Food is 23, we just have to find an observation with Food of 23, which looks to happen for observation 21. (Found by just viewing the data set.)

```
augmentNYC[21, ] #says give row 21, but all columns
```

```
## # A tibble: 1 x 8
##   Price  Food .fitted .resid   .hat .sigma .cooksd .std.resid
##   <dbl> <dbl>   <dbl>  <dbl>  <dbl>  <dbl>   <dbl>      <dbl>
## 1    54    23    49.8   4.24 0.0148   7.28 0.00259      0.588
```

8. What is the predicted price for dinner when the food rating is 23?

   ANSWER $49.80

How is the rating of food related to price? Our ASSESS step showed no major problems with conditions for inference, so we can use the R output to perform the t-test for slope, or use a confidence interval.

```
msummary(fm)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.8321     5.8631  -3.041  0.00274 **
## Food          2.9390     0.2834  10.371  < 2e-16 ***
##
## Residual standard error: 7.261 on 166 degrees of freedom
## Multiple R-squared:  0.3932, Adjusted R-squared:  0.3895
## F-statistic: 107.6 on 1 and 166 DF,  p-value: < 2.2e-16
```

```
confint(fm)#default level is .95
```

```
##                  2.5 %    97.5 %
## (Intercept) -29.408044 -6.256253
## Food          2.379464  3.498455
```

9.Is *Food* is a significant predictor of *Price*.

   ANSWER (hint: look at both the p value for the t-test and the confidence interval)

Looking at the R output summary table we can see that for the coefficient of *Food* (our one predictor), the t-test statistic is 10.37 and resulting two-sided p-value is very small (p-value: $<$ 2.2e-16).

Similarly, the confidence interval for slope does not contain 0. We have evidence that *Food* is a significant predictor of *Price*. Moving away from the language of "significance", we can say that it appears there is a relationship between the *Food* rating and average *Price* of a meal based on our data.

10. How would you interpret the Confidence interval and Confidence level?

    ANSWER CI: 95 percent confident that the true slope for predicting average price using food rating lies in the interval (2.379, 3.498). Or perhaps you say: we have 95 percent confidence that the average price is expected to increase by between 2.379 and 3.498 dollars for each one unit increase in food rating. Confidence Level: If we repeated this study, and generated many 95

percent CIs for the true slope predicting average price using food rating, we would expect our intervals to contain the true slope about 95% of the time.