Paulette Rodriguez
Prof. Loh
Applied Statistics
24 September 2019

Homework 2

#1.106

N(150,35)
μ = 150
σ = 35
Scores = [150, 140, 100, 180, 230]

Zscores1 = $Z = \frac{(x-\mu)}{\sigma} = \frac{(150-150)}{35} = 0$
Zscores2 = $Z = \frac{(x-\mu)}{\sigma} = \frac{(140-150)}{35} = -0.2857143$
Zscores3 = $Z = \frac{(x-\mu)}{\sigma} = \frac{(100-150)}{35} = -1.428571$
Zscores4 = $Z = \frac{(x-\mu)}{\sigma} = \frac{180-150}{35} = 0.8571429$
Zscores5 = $Z = \frac{(x-\mu)}{\sigma} = \frac{(230-150)}{35} = 2.285714$

#1.108

N(288,32)

Percentile	Score
10%	246
25%	279
50%	290
75%	311
90%	328

By using the percentiles to assess whether or not the NAEP U.S. History scores for 12-th grade students are approximately Normal, we were able to find the following z-scores:

- 10%, z-score ~ -1.28
- 25%, z-score ~ -0.67
- 50%, z-score = 0.00
- 75%, z-score ~ 0.67
- 90%, z-score ~ 1.28

Now, for each of the z-scores that were found for the given percentiles, we can find the test-scores of the students to see if they have a normal distribution. We will find the test scores by using

$$x = \mu + z * \sigma$$

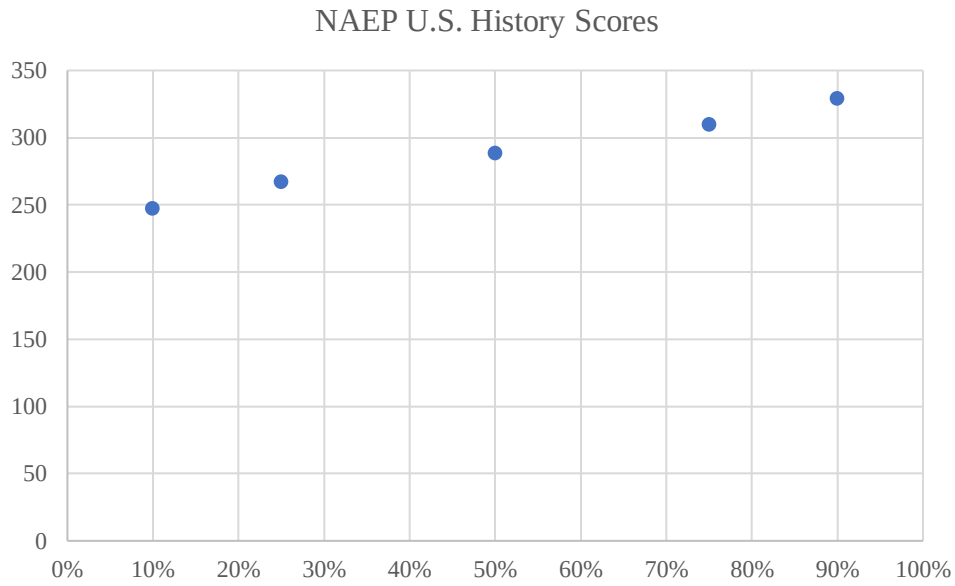
which is derived from $z = \frac{(x-\mu)}{\sigma}$.

The new test-scores are:

- x = 288 + (-1.28)*32 = 247.04 for percentile of 10%
- x = 288 + (-0.67)*32 = 266.56 for percentile of 25%
- x = 288 + (0.00)*32 = 288 for percentile of 50%
- x = 288 + (0.67)*32 = 309.44 for percentile of 75%
- x = 288 + (1.28)*32 = 328.96 for percentile of 90%

As we can see from the calculations, the difference between the actual test-scores and the test-scores that were obtained is very small. Therefore, we can see say that the history scores are approximately Normal.

We can also see that by making a scatterplot the scores have an approximately linear relationship. Hence, the data are normally distributed.



#1.110

(a). By using the 68-95-99.7 rule to describe this distribution we were able to find that:

- 68% of women spoke between 7,856 and 20,738 words per day.
- 95% of women spoke between 1,415 and 27,179 words per day.
- 99.7% of women spoke between -5,026 and 33,620 words per day.

(b). This distribution cannot truly be Normal since it is impossible to speak a negative number of words per day as how we saw that 99.7% of this distribution lies between -5,026 and 33,620. Therefore, using this rule in this situation might not be useful.

(c). By using the 68-95-99.7 rule to describe this distribution we were able to find that:

- 68% of men spoke between 5,004 and 23,116 words per day.
- 95% of men spoke between -4,052 and 32,172 words per day.
- 99.7% of men spoke between -13,108 and 41,228 words per day.

This distribution is not Normal since the lower end of 95% of this distribution, which ranges from -13,108 to -4,052, is less than zero. Again, it is not possible to speak a negative number of words in a day. Therefore, using this rule in this situation might not be useful.

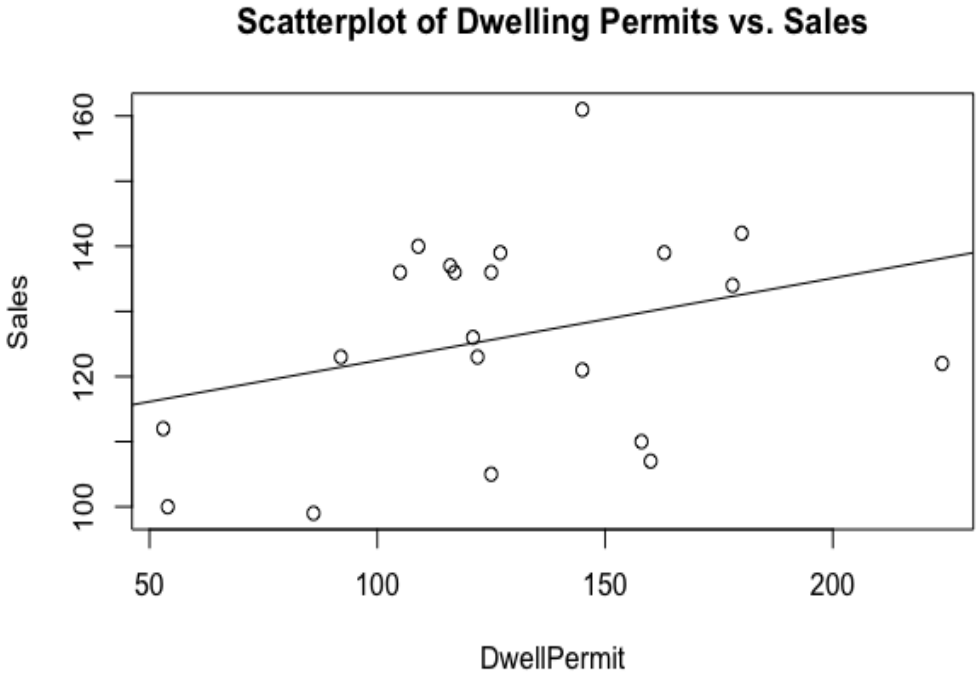
(d). At the beginning of this study conventional wisdom suggested that women are more talkative than men. Therefore, one study collected data on 42 women and 35 men and examined it based on the given stereotype. After computing these results and using the 68-95-99.7 rule we can see that men have a larger standard deviation, meaning that 68% of the women distribution lies in the 68% region of the men. We can conclude that it might be possible that men speak less than women but also that they might speak more words in a day than women.

#2.144

```
> MEIS <- read.csv(file.choose(), header=T,sep=,)
> attach(MEIS)
> plot(DwellPermit, Sales, main="Scatterplot of Dwelling Permits vs. Sales", xlab="DwellPermit", ylab = "Sales")
> mod <- lm(Sales ~ DwellPermit)
> abline(mod)
> coef(mod)
        (Intercept) DwellPermit
        109.8204261   0.1263479
> summary(mod)

        Call:
        lm(formula = Sales ~ DwellPermit)

        Residuals:
        Min      1Q      Median      3Q     Max
        -23.036 -16.122   1.556  11.397  32.859

        Coefficients:
        Estimate Std. Error t value
        (Intercept) 109.8204    11.5582   9.501
        DwellPermit   0.1263     0.0857   1.474
        Pr(>|t|)
        (Intercept) 1.19e-08 ***
        DwellPermit    0.157
        ---
        Signif. codes:
        0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        Residual standard error: 15.7 on 19 degrees of freedom
        Multiple R-squared:  0.1027,  Adjusted R-squared:  0.05543
        F-statistic: 2.174 on 1 and 19 DF, p-value: 0.1568
```

(a). and (b).

Making a scatterplot with Sales as the response variable and Dwelling Permits as the explanatory variable helps us display the relationship between the two variables. As we can see from the scatterplot attached below, we can see that there are indeed a few outliers that lie around 100-110 and 160 at Sales.



By using the lm function in R to find the least-squares regression line we find that this line between Sales and Dwell Permits is:

$$y = 0.1263*x + 109.8204$$

(c).

The slope is an indicator of how Sales is expected to change and by how much as Dwell Permits increases. In the case of this example, 0.1263 is a positive slope, but it is pretty low. Meaning, that as Dwell Permits increases Sales will still increase but a very low rate and slow pace.

(d).

The intercept in our example will help us see that if Dwelling Permits equal zero, then Sales will start at 109.8204, or 109. The intercept in this case is not very useful in explaining the relationship between the two variables Sales and Dwelling Permits simply because this information doesn't make a lot of sense for the regression line used in our plot.

(e). and (f).

By filtering out Canada with an index of 224 for Dwelling Permits and using the augment function in R we will be able to find the prediction value of Sales for Canada and the residuals for it.

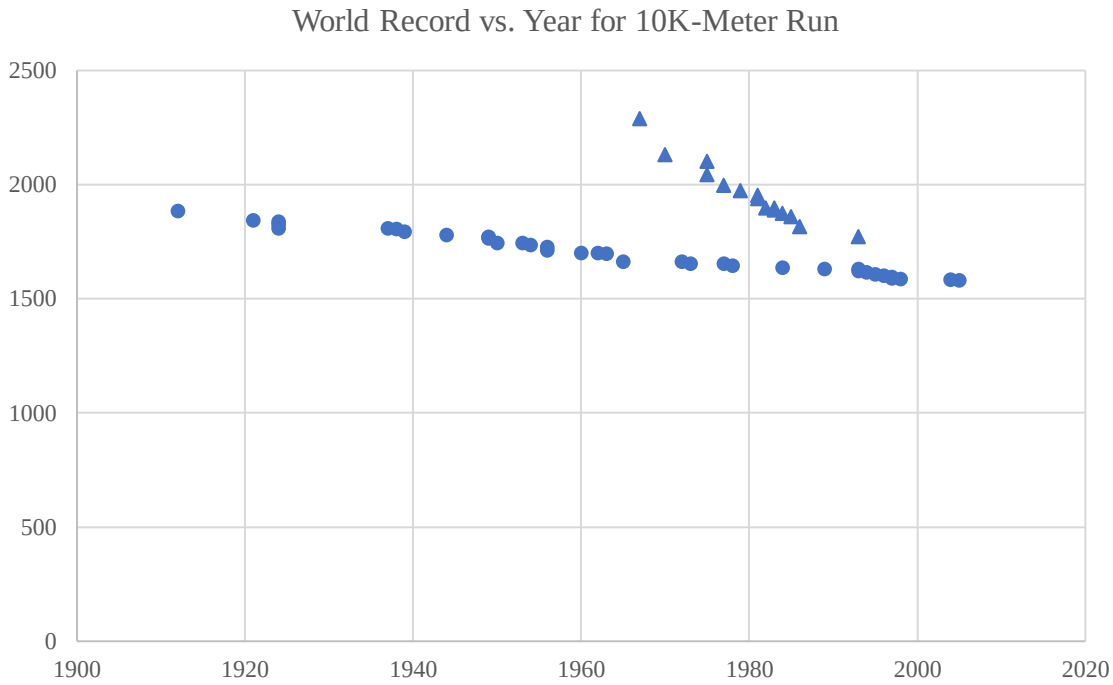
Sales	DwellPermit	.fitted	.se.fit	.resid
122	224	138.1224	8.847853	-16.12235

Canada, with an index of 224, has the predicted value of 138.12 and -16.12 as the residual.

(g).
By the above code we can see that the percent of the variation in sales that is explained by dwelling permits is only 10%. The remaining 90% is, or can be, due to factors that are either unexplained or random.

#2.150

(a).



In the above scatterplot, the solid circle symbol is used for men and the solid triangles are used for women.

As we can see from the plot, men started distance running years before women started doing so. The pattern we see for the men shows that throughout the years the amount of time, in seconds, they took to complete the distance running started decreasing slowly becoming slightly constant towards recent years. This shows that the men were able to complete the distance running in shorter amounts of time, making slow improvements.

As we can see from the plot, women started distance running years after men started doing so. The pattern we see for the women shows that in short amount of time the women were able to get faster at the distance running, decreasing the amount of time, in seconds, they took to complete the run; improving at a faster rate than men.

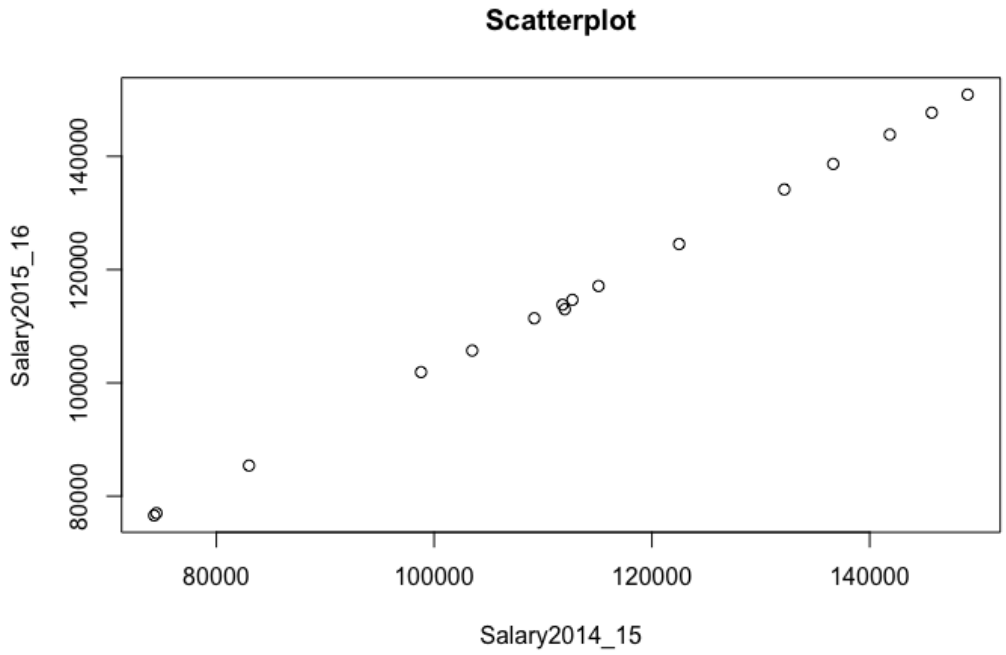
(b). Yes, the data seems to support the claim that even though women began running this long distance later than men, it might be expected that their improvements happen more rapidly. Moreover, it is often said that men have little advantage over women in distance running as opposed to sprints, where muscular strength plays a greater role. The data shows that during the

late 1990s and early 2000s women seem to be getting faster, but men are still a bit faster than the women.

#2.160

(a).

```
> FACULTY <- read.csv(file.choose(), header=T, sep=,)
> attach(FACULTY)
> plot(Salary2014_15, Salary2015_16, main = "Scatterplot", xlab="Salary 2014-15", ylab="Salary 2015-16")
```



(b). As we can see from the scatterplot provided above, the form, direction and relationship of the data is really strong and positive. The data has a strong linear form and relationship with a positive slope, which we can interpret to mean that the data is normally distributed. There are no outliers in the data, or any influential outliers that may cause some changes in the data or any computations.

(c).

```
 > cor(FACULTY)
                Salary2014_15 Salary2015_16
Salary2014_15     1.0000000     0.9998708
Salary2015_16     0.9998708     1.0000000
> mod <- lm(FACULTY)
> summary(mod)
        Call:
        lm(formula = FACULTY)

        Residuals:
```

```
    Min      1Q Median      3Q     Max
-895.19 -81.05 -49.78   91.03 1119.95
Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.984e+03 5.118e+02   -5.83 4.37e-05 ***
Salary2015_16 1.008e+00 4.329e-03 232.74 < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1

Residual standard error: 393.6 on 14 degrees of freedom
Multiple R-squared:  0.9997,  Adjusted R-squared:  0.9997
F-statistic: 5.417e+04 on 1 and 14 DF, p-value: < 2.2e-16
```

(c). 99.97% of the variation in 2015–2016 salaries is explained by 2014 – 2015 salaries.

#2.162

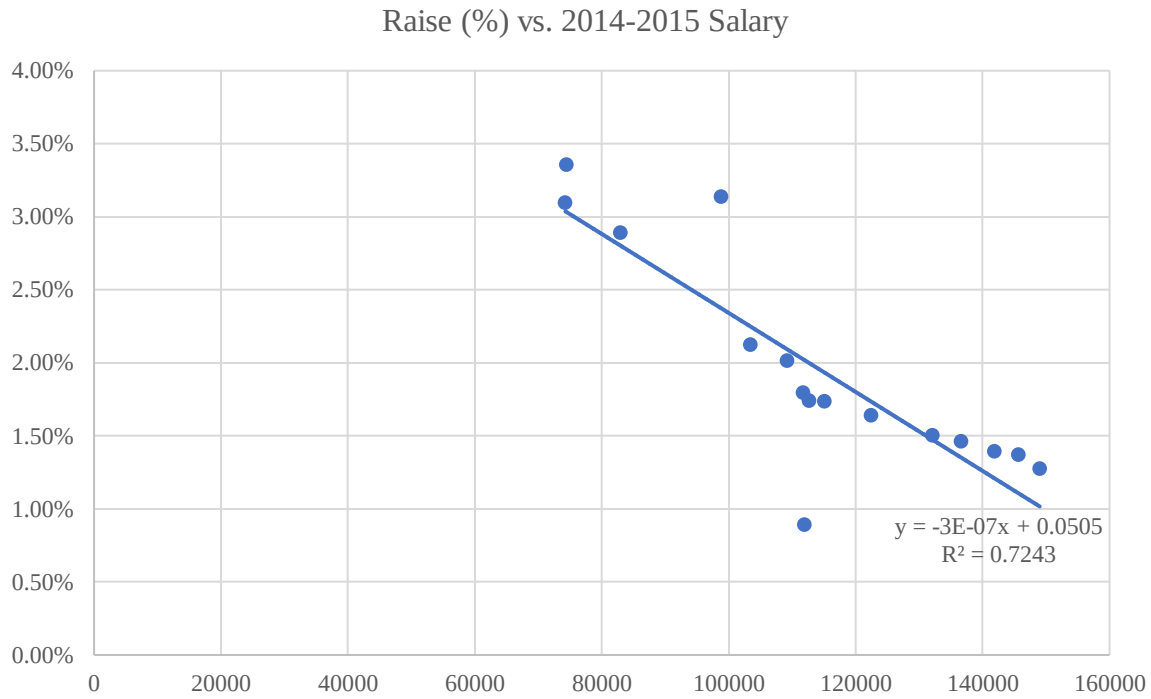
(a).
By taking the difference between the 2015–2016 salaries and the 2014–2015 salaries, diving by the 2014–2015 salaries, then multiplying by 100 we were able to compute the percent raise for each faculty member.

This table shows the percent raise that was computed for each faculty member.

Salary2014_15 ($)	Raise (%)	Salary2015_16 ($)
145700	1.37%	147700
112700	1.74%	114660
109200	2.01%	111400
98800	3.14%	101900
112000	0.89%	113000
111790	1.80%	113800
103500	2.13%	105700
149000	1.28%	150900
136650	1.46%	138650
132160	1.51%	134150
74290	3.10%	76590
74500	3.36%	77000
83000	2.89%	85400
141850	1.40%	143830
122500	1.64%	124510
115100	1.74%	117100

Now, we will make a scatterplot with raise as the response variable and the 2014–2015 salaries as the explanatory variable. As we can see from the scatterplot provided below, the relationship we see between the raise and the salaries is that of a negative linear relationship. As the salaries for these faculty members increase, the less percent raise they will receive. For those that make less,

whose salaries are lower, we can see from the scatterplot that they have the highest raise percentage.



(b). The least-squares regression line for this plot is y = -0.0000003*x + 0.0505.

(c). As we can see from the graph there are a couple of outliers. By removing the outlier at point (112000, 0.89%) it doesn't really change the graph or regression line that much.

(d). Yes, as we've mentioned before there is evidence in the data to support the idea that greater percent raises are given to those with lower salaries. For example, the highest percent raise is given to those that make less than 100,000 and 80,000. Those who make more than 100,000, as shown in the plot, the percentage in raise is seen to be decreasing. Also, if we were to sort the data with its corresponding percent raise, we will be able to see this as well.

Salary2014_15 ($)	Raise (%)
74290	3.10%
74500	3.36%
83000	2.89%
98800	3.14%
103500	2.13%
109200	2.01%
111790	1.80%
112000	0.89%
112700	1.74%
115100	1.74%
122500	1.64%
132160	1.51%
136650	1.46%
141850	1.40%
145700	1.37%
149000	1.28%

#2.170

(a).

	Smoker?	
	Yes	**No**
Dead	139	230
Alive	443	502
Total	582	732

The percent of the smokers that stayed alive for 20 years is:

$$\frac{443}{582} * 100 = 76\%$$

The percent of nonsmokers that survived is:

$$\frac{502}{732} * 100 = 68.6\% \sim 69\%$$

(b). The age of the women at the time of the study is a lurking variable. In the following computations we will show that within each of the three age groups in the data, a higher percent of nonsmokers remained alive 20 years later.

	Age 18 to 44	
	Yes	**No**
Dead	19	13
Alive	269	327
Total	288	340

Smoker remained alive: $\frac{269}{288} * 100 = 93\%$

Nonsmoker remained alive: $\frac{327}{340} * 100 = 96\%$

	Age 45 to 64	
	Yes	**No**
Dead	78	52
Alive	167	147
Total	245	199

Smoker remained alive: $\frac{167}{245} * 100 = 68\%$

Nonsmoker remained alive: $\frac{147}{199} * 100 = 73\%$

	Age 65+	
	Yes	**No**
Dead	42	165
Alive	7	28
Total	49	193

Smoker remained alive: $\frac{7}{49} * 100 = 14.3\%$

Nonsmoker remained alive:

$$\frac{28}{193} * 100 = 14.5\%$$

As we can see, this is an example of Simpson's paradox. In part (a) is seemed surprising that a higher percentage of smoker stayed alive while the nonsmokers had a lower percentage. In part (b) we see that this is not the case. Here, we have shown that within each of the three age groups in the data, a higher percent of nonsmokers remained alive 20 years later.

(c). Yes, by looking at the percent of smokers in the three age groups that we found in part (b) we can see that it verifies this explanation. The difference between the smokers and the nonsmokers who remained alive in group Age 65+, we can see that the difference is very small with a difference of only 0.2%. We can also take a look at the total of smokers from that same group. Compared to the total number of smokers from the other two age groups, the women who smoked in the group Age 65+ was smaller than that of the other groups. Therefore, the explanation that "few of the older women (over 65 at the original survey) were smokers, but many of them had died by the time of follow-up" holds up.