AP Statistics: Chapter 3 Exam Review

Alison Rector

2022-09-26

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

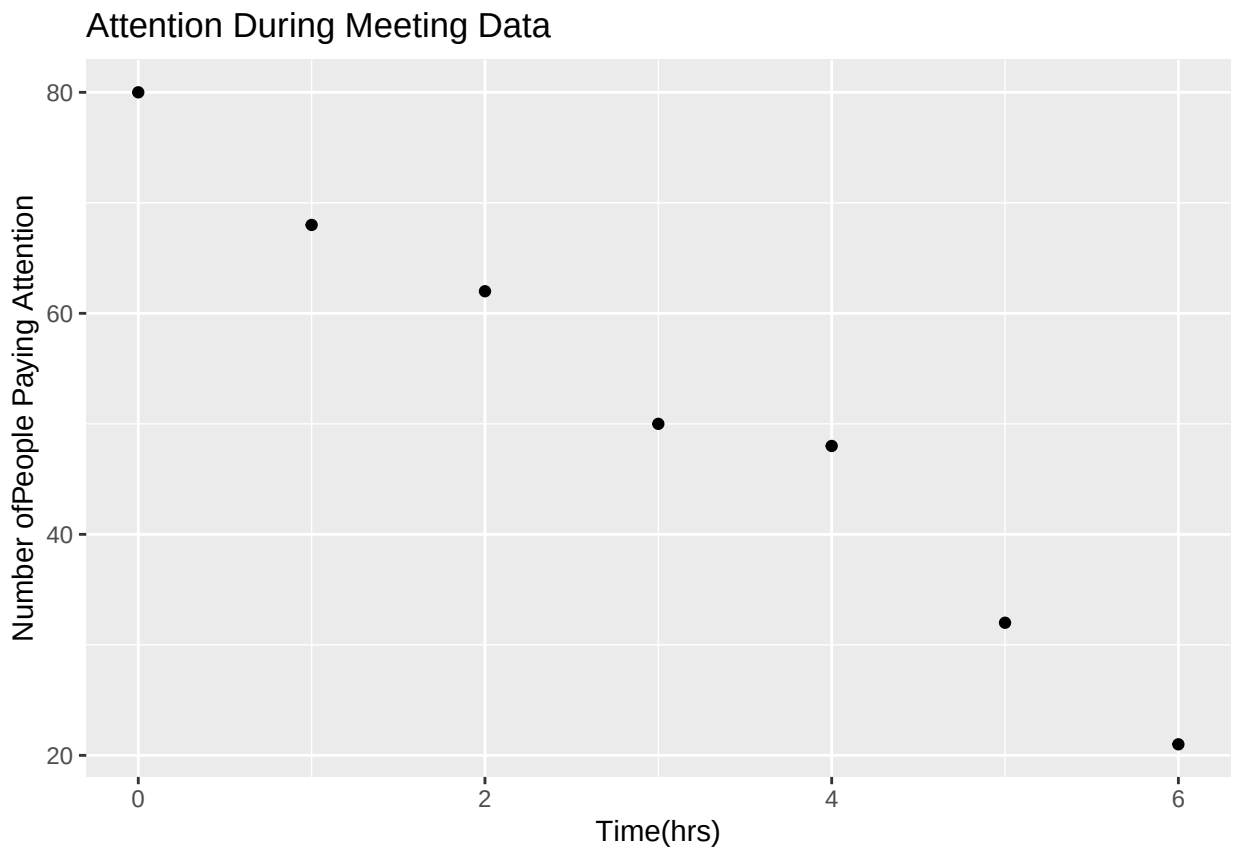
In my meeting yesterday, I noticed that the number of people paying attention to the training was declining throughout the day. I collected the following data: Time (hours): 0 1 2 3 4 5 6 Number of people paying attention: 80 68 62 50 48 32 21

1. Make a scatterplot of the data on your calculator and sketch it below.

```
meeting_time<-c(0,1,2,3,4,5,6)
n_people<-c(80, 68, 62, 50, 48, 32, 21)

meeting_df<-data.frame(time=meeting_time, attention_n=n_people)
```

```
library(ggplot2)
ggplot(data=meeting_df, aes(x=time, y=attention_n))+
  geom_point()+
  xlab("Time(hrs)")+
  ylab("Number ofPeople Paying Attention")+
  ggtitle("Attention During Meeting Data")
```



2. Describe the data (Direction, unusual features, form, strength)
The data shows a strong negative linear relationship between time and the number of people paying attention during the training. There are no unusual features nor obvious outliers.

3. Find the line of best fit using the equations for a and b. Confirm that it matched with the calculator's line of best fit.

```
mymod<-lm(attention_n~time, data=meeting_df)
```

$N = 79.750 - 9.393 * t$

4. Make a prediction for the # of people paying attention at t = 2.5 hours. $N(2.5) = 79.750 - 9.393 * (2.5) = 56.2675 \approx 56$ *people 5. What is r? Interpret.*

```
cor(meeting_df$time, meeting_df$attention_n)
```

```
## [1] -0.9901502
```

6. What is r2? Interpret.

```
summary(mymod)
```

```
##
## Call:
## lm(formula = attention_n ~ time, data = meeting_df)
##
## Residuals:
##       1       2       3       4       5       6       7
##  0.2500 -2.3571  1.0357 -1.5714  5.8214 -0.7857 -2.3929
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   79.750      2.142   37.24 2.63e-07 ***
## time          -9.393      0.594  -15.81 1.84e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.143 on 5 degrees of freedom
## Multiple R-squared:  0.9804, Adjusted R-squared:  0.9765
## F-statistic: 250.1 on 1 and 5 DF,  p-value: 1.84e-05
```

$r^2 \approx 0.9804$
This means that approximately 98% of the variation in y can be explained by x.

7. What is the slope of the LSRL? Interpret in the context of the problem. Every hour that passes there is a reduction of 9.39 people paying attention.

8. What is the y-intercept of the LSRL? Interpret in the context of the problem. 79.750
Approximately 80 people are paying attention at the start of the lecture.

9. Identify the values a and b in the Minitab output: Predictor Coef SE Coeff T P Constant 79.75 5.34 3.41 0 Time -9.393 4.23 -15.81 0 S = 3.143 R-Sq = 98.0% R-Sq (adj) = 97.2%

10. Interpret $s = 3.143$ in the context of the problem.
This tells us that the regression model predicts the number of people who are paying attention during the training with an average error of about 3.143.

11. Find the residual for a time of 4 hours.
$\hat{y_4} = 79.750 - 9.393 * (4) = 42.178$ $y_4 = 48$

$r_4 = (y_i - \hat{y}) = 48 - 42.178 = 5.822$

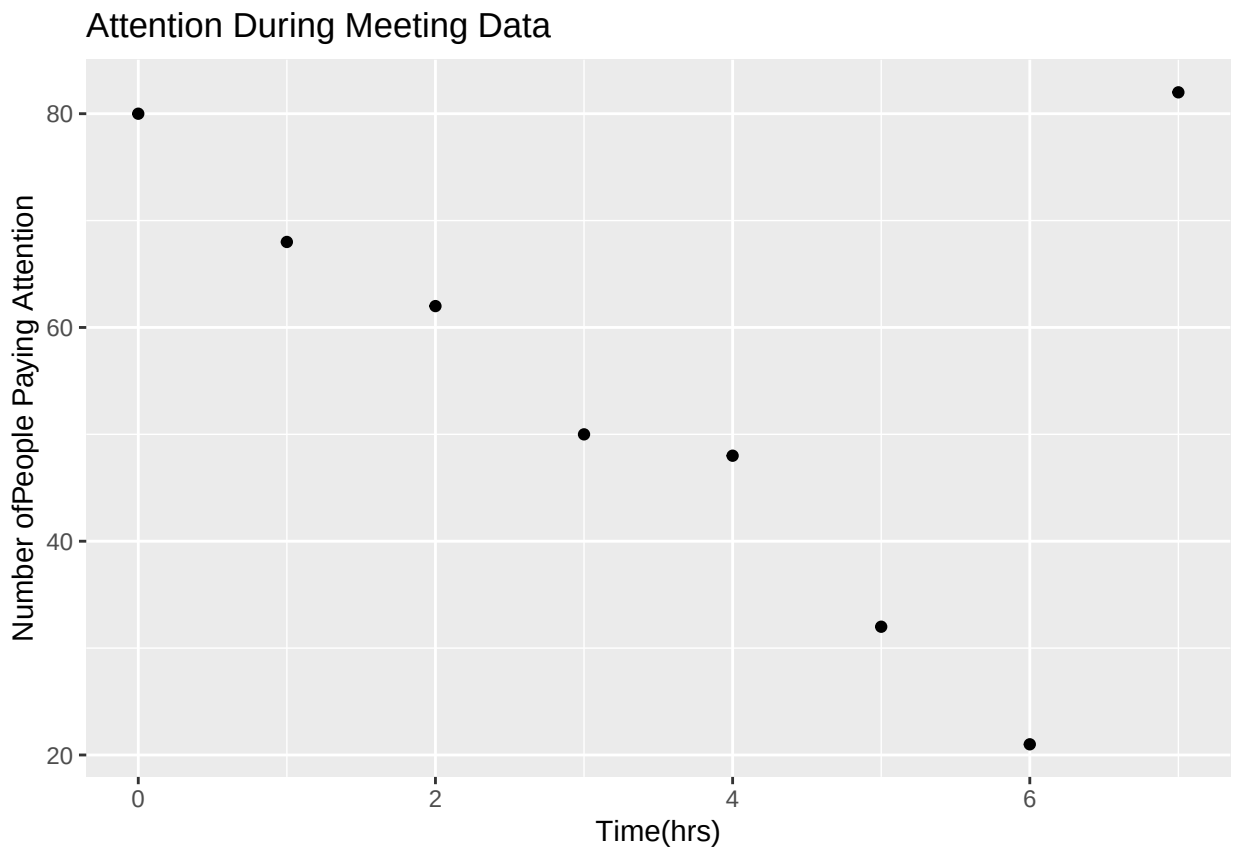
12. Make a residual plot on your calculator. Sketch it below.

13. Mr. Bakri walked into the meeting at t = 7 hours and all of a sudden 82 people were paying attention. Would this point be influential? Prove it.

```
meeting_time_new<-c(0,1,2,3,4,5,6, 7)
n_people_new<-c(80, 68, 62, 50, 48, 32, 21, 82)

meeting_df_new<-data.frame(time=meeting_time_new, attention_n=n_people_new)
```

```
library(ggplot2)
ggplot(data=meeting_df_new, aes(x=time, y=attention_n))+
  geom_point()+
  xlab("Time(hrs)")+
  ylab("Number ofPeople Paying Attention")+
  ggtitle("Attention During Meeting Data")
```




```
mymod<-lm(attention_n~time, data=meeting_df_new)
summary(mymod)
```

```
##
## Call:
## lm(formula = attention_n ~ time, data = meeting_df_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.060  -9.875  -2.238   5.378  39.667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   68.417     13.811   4.954  0.00257 **
## time          -3.726      3.301  -1.129  0.30214
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.4 on 6 degrees of freedom
## Multiple R-squared:  0.1751, Adjusted R-squared:  0.03765
## F-statistic: 1.274 on 1 and 6 DF,  p-value: 0.3021
```

$N = 68.417 - 3.726t$

14. What would happen to r when we add this point.

```
cor(meeting_df_new$time, meeting_df_new$attention_n)
```

```
## [1] -0.418481
```