Homework 3

Claudia Nixon and Catherine Allred

4/26/2021

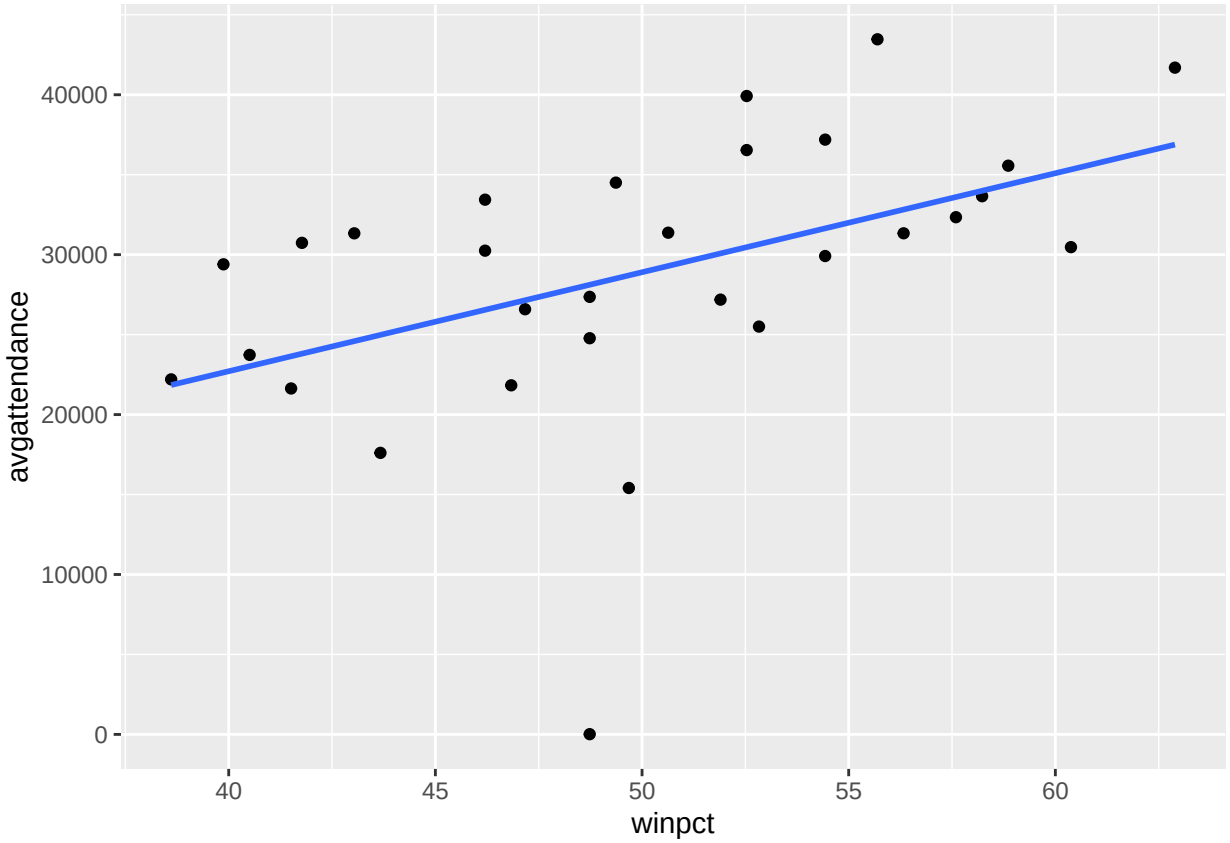
#1 Import the mlbattendance2015 data set from the class asulearn page into R.

```
mlbattendance2015_1_ <- read_excel("~/mlbattendance2015 (1).xlsx")
```

#2 Build a regression model to predict average attendance from Winning Percentage (Winpct). Also provide a graphical summary of the relationship.

```
# x <- winpct
# y <-  avgattendance
ggplot(data = mlbattendance2015_1_) + geom_point(aes(x=winpct, y=avgattendance)) + geom_smooth(aes(x=wi
```

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



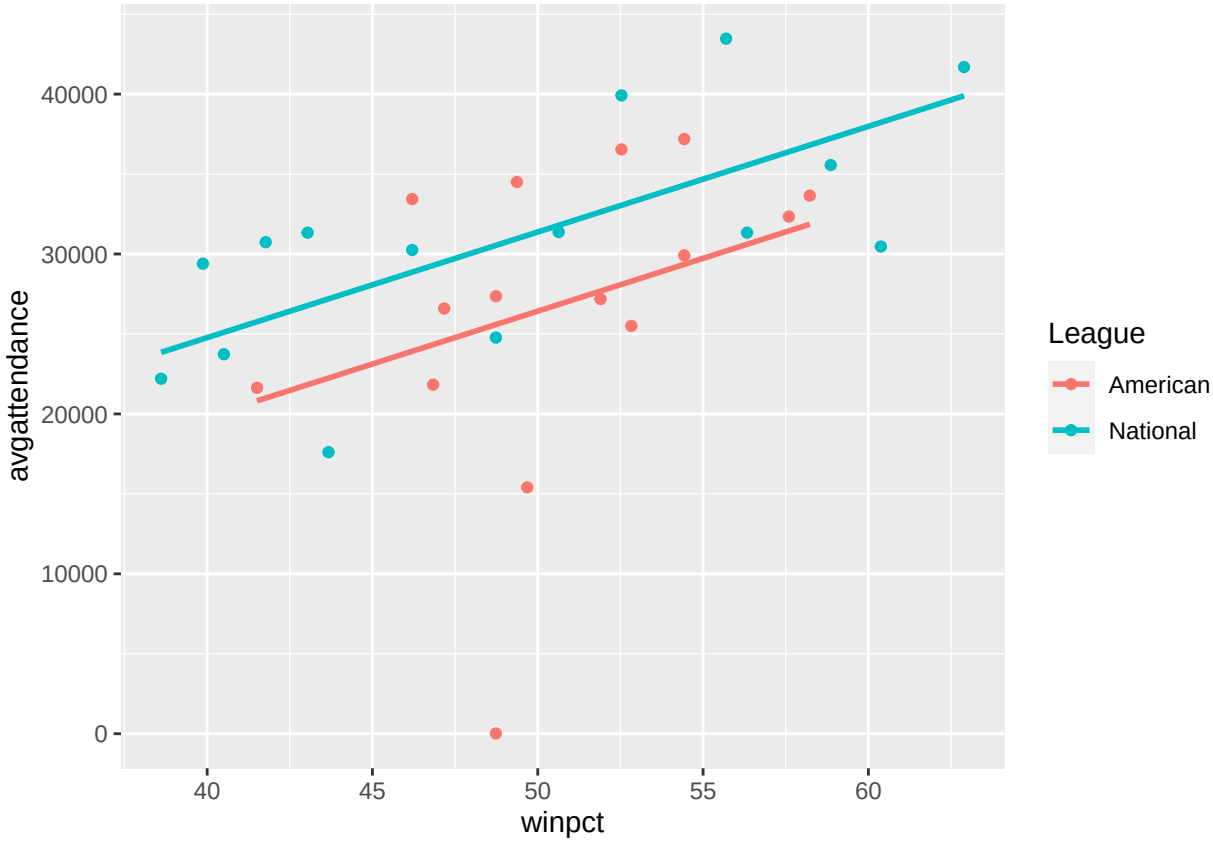

```
summary(lm(avgattendance ~ winpct, data = mlbattendance2015_1_))
```

```
##
## Call:
```

```
## lm(formula = avgattendance ~ winpct, data = mlbattendance2015_1_)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28101.7  -2670.1      6.8   5883.8  11039.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2054.6    11166.8  -0.184  0.85534
## winpct         619.1      221.6   2.794  0.00928 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7730 on 28 degrees of freedom
## Multiple R-squared:  0.2181, Adjusted R-squared:  0.1901
## F-statistic: 7.808 on 1 and 28 DF,  p-value: 0.009281
```

#3 Next, add League as a predictor to the previous model, in a parallel lines model.

```
# x <- winpct + league
# y <-  avgattendance
ggplot(data = mlbattendance2015_1_) + geom_point(aes(y=avgattendance, x=winpct, color = League)) + geom_
```




```
summary(lm(avgattendance ~ winpct + League, data = mlbattendance2015_1_))
```

```
##
## Call:
## lm(formula = avgattendance ~ winpct + League, data = mlbattendance2015_1_)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25573.8  -2296.0    857.8   4671.3   9521.4
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -6609.3    11026.2  -0.599  0.55389
## winpct            660.7      214.3   3.084  0.00467 **
## LeagueNational   4950.8     2729.6   1.814  0.08085 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7432 on 27 degrees of freedom
## Multiple R-squared:  0.303,  Adjusted R-squared:  0.2513
## F-statistic: 5.868 on 2 and 27 DF,  p-value: 0.007653
```

#4 Finally, add potential interaction between League and Winning Percentage as a predictor, creating a possible non-parallel lines model.

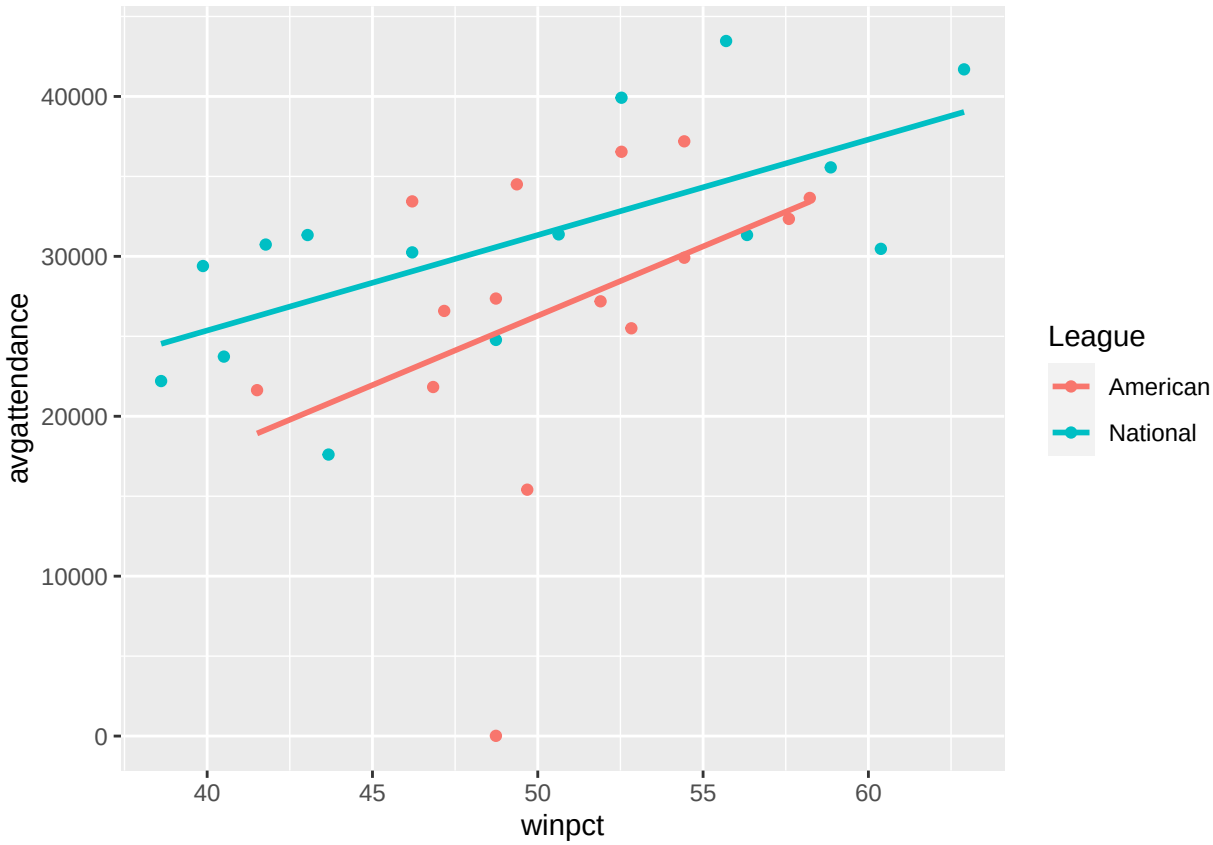
```
# x <- winpct * league
# y <-  avgattendance
summary(lm(avgattendance ~ winpct*League, data = mlbattendance2015_1_))
```

```
##
## Call:
## lm(formula = avgattendance ~ winpct * League, data = mlbattendance2015_1_)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25172.5  -2231.2     10.5   4144.8  10445.3
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -17069.6    22739.0  -0.751   0.4596
## winpct                   867.1      447.0   1.940   0.0633 .
## LeagueNational         18555.3    25903.2   0.716   0.4802
## winpct:LeagueNational   -270.2      511.4  -0.528   0.6018
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7534 on 26 degrees of freedom
## Multiple R-squared:  0.3104, Adjusted R-squared:  0.2308
## F-statistic: 3.901 on 3 and 26 DF,  p-value: 0.01998
```

#5 Provide a graph of the data that incorporates both winning percentage and league as predictors and avg attendance as the response.

```
ggplot(mlbattendance2015_1_, aes(x=winpct, y=avgattendance, color=League)) +
  geom_point()+
  geom_smooth(method = lm, se = FALSE)
```

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



#6 Which of the models in (2), (3), and (4) do you think is the best? Clearly justify your answer.

Model 2 in question 3 is the best model out of the three options. The R-squared value for model 2 was 0.25, higher than the other two models (0.19 and 0.23). The p-value for the second model is also lower and thus more significant than the other two models making it . Therefore, model 3 had the best predictions.

#7 Predict average attendance for an American League team with a 60% winning percentage, and for a National League Team with a 50% winning percentage based upon your best model from steps (2), (3), (4).

```
#American league with 60% winning percentage
x <- 60
sum(-6609.3+660.7*x)
```

```
## [1] 33032.7
```

```
#National league with 50% winning percentage
x <- 50
sum(-6609.3+660.7*x+4950)
```

```
## [1] 31375.7
```

Model 2 predicts that an American League team with a 60% winning percentage would have 33032.7 average attendance and a Nation League team with 50% winning percentage would have 31375.7 average attendance.