Linear Regression 1

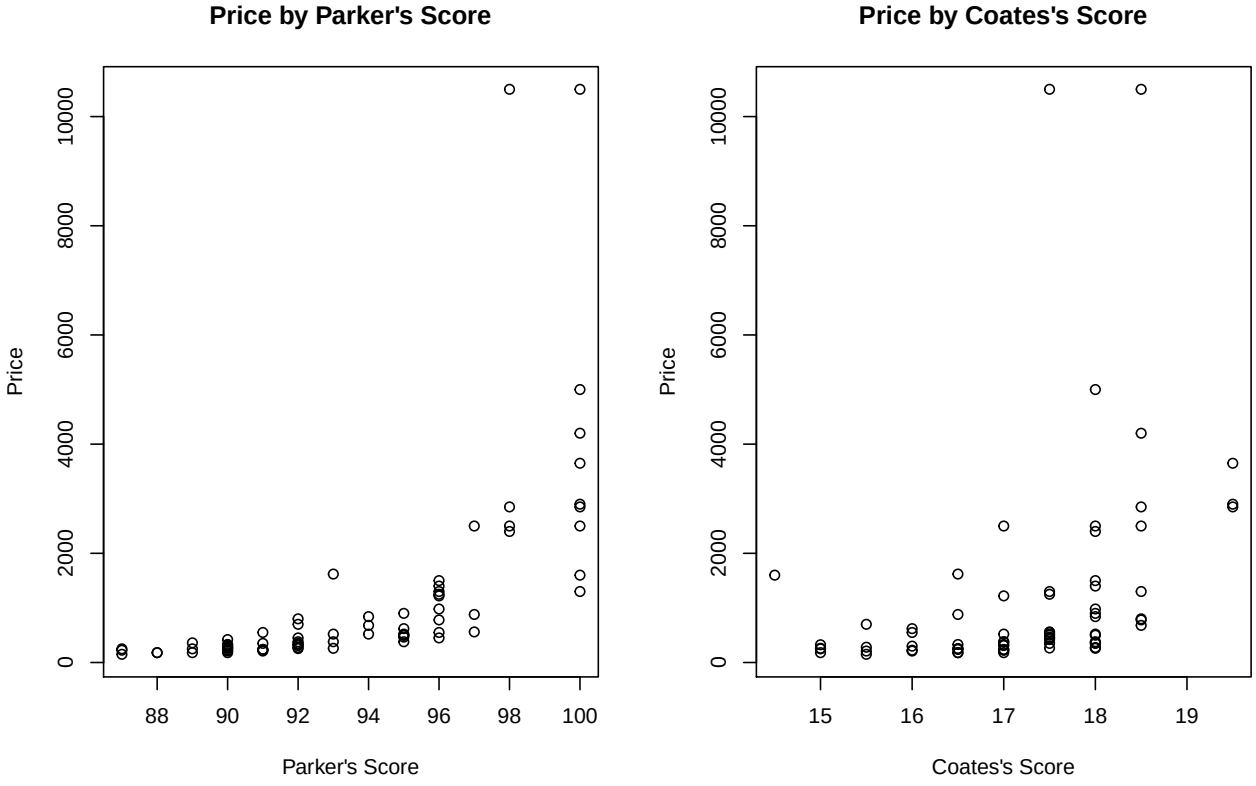
Ted Tinker, 3223468

April 6, 2017

Question 1, Part A

For the first part of this project, I loaded the CSV file *Bordeaux* into R Studio. The following code generates two simple scatterplots:

```
par(mfrow=c(2,2)) # Signifies that we wish to make two scatterplots side-by-side
plot(Bordeaux$ParkerPoints,Bordeaux$Price, xlab="Parker's Score",ylab="Price",
    main="Price by Parker's Score") # First plot
plot(Bordeaux$CoatesPoints,Bordeaux$Price, xlab="Coates's Score",ylab="Price",
    main="Price by Coates's Score") # Second plot
```



In these plots, we see a positive correlation between the score given by two famous wine critics and the price of the wine reviewed. However, the correlation does not appear to be perfectly linear. The distinct curve suggests an exponential trend, such that highly reviewed wines have exponentially higher prices while poorly reviewed wines have prices tending to zero.

Part B

This code finds the lines of best fit for the two plots and stores the information in variables:

```
ParkWine.fit <- lm(Price~ParkerPoints, data=Bordeaux)
CoatWine.fit <- lm(Price~CoatesPoints, data=Bordeaux)
```

Displaying Parker's data gives the following output (Coates' data is in similar form):

```
summary(ParkWine.fit) # Ask for the data to be displayed
Call:
lm(formula = Price ~ ParkerPoints, data = Bordeaux)

Residuals:
    Min      1Q  Median      3Q     Max
-1819.0  -652.8  -187.3   266.5  7992.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -27473.22    4469.16  -6.147 4.30e-08 ***
ParkerPoints    305.92      47.68   6.417 1.42e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1506 on 70 degrees of freedom
Multiple R-squared:  0.3704,Adjusted R-squared:  0.3614
F-statistic: 41.17 on 1 and 70 DF,  p-value: 1.419e-08
```

Using (Intercept) and ParkerPoints, and the corresponding values in Coates' data, we find best fit lines

$$Price = -27473.22 + 305.92(Parker's\ Score) \qquad Price = -9472.6 + 618.9(Coates'\ Score)$$

These indicate that, given our linear interpretation of the dataset, we expect a bottle of wine to cost about -$27,473 plus $306 for each point Parker awarded it, or -$9,472 plus $619 for each point Coates awarded it.

Here, -27473.22 and -9472.6 are β_0, the intercepts (the expected price of the wine if the reviewers awarded them no points). The fact that these values are negative reinforces the idea that a linear model might not be the best fit for this data, as wine prices should not be negative, no matter how poorly reviewed.

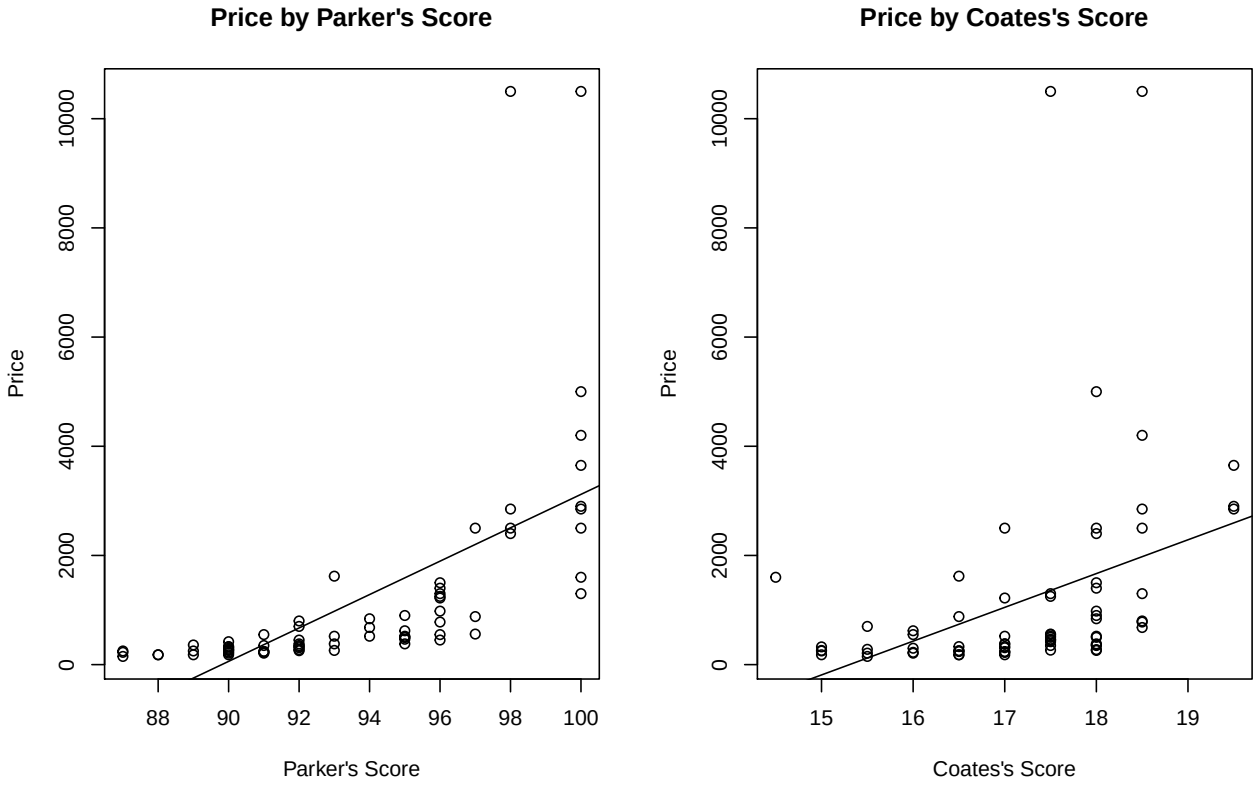
Similarly, 305.92 and 618.9 are β_1, the slopes of the Least-Squares-Regression lines, indicating the amount we expect the price of a bottle of wine to change given one point from either of our reviewers. Parker rates wines on a scale from 1-100, while Coates, Frenchman through and through, rates wines from 1-20. As one of Coates' points represents $\frac{1}{20}$ of the range while Parker's represents $\frac{1}{100}$, it makes sense for wine prices to change more for one of Coates' points than one of Parker's.

Part C

With some addition to our code, we produce two plots with regression lines:

```
ParkWine.fit <- lm(Price~ParkerPoints, data=Bordeaux)
CoatWine.fit <- lm(Price~CoatesPoints, data=Bordeaux) # Make lines of best fit, as in Part B

par(mfrow=c(2,2)) # We want two scatterplots, as in Part A
plot(Bordeaux$ParkerPoints,Bordeaux$Price, xlab="Parker's Score",ylab="Price",
    main="Price by Parker's Score") # First Plot
abline(ParkWine.fit$coefficients[1],ParkWine.fit$coefficients[2]) # Makes line on first plot
plot(Bordeaux$CoatesPoints,Bordeaux$Price, xlab="Coates's Score",ylab="Price",
    main="Price by Coates's Score") # Second Plot
abline(CoatWine.fit$coefficients[1],CoatWine.fit$coefficients[2]) # Makes line on second plot
```



On the left and right of both graphs, most of the points are above the line, while in the centers, most of the points are below the line. To me, this suggests that a constant σ^2 variance around the line of best fit does not properly represent the data. Wine prices also seem to be more widely varied for more highly-reviewed wines.

Part D

First, we address Part B by adjusting the datasets and making new lines of best fit:

```
newPrice <- log(Price)
newPPoints <- log(ParkerPoints)
newCPoints <- log(CoatesPoints)
newParkWine.fit <- lm(newPrice~newPPoints) # For each dataset, make a new dataset
newCoatWine.fit <- lm(newPrice~newCPoints) # whose entries are the log value of the original
```

```
summary(newParkWine.fit) # Ask for the data to be displayed
Call:
lm(formula = newPrice ~ newPPoints)

Residuals:
    Min       1Q   Median       3Q      Max
-0.90463 -0.34757  0.01913  0.25471  1.79930

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -94.132      6.860  -13.72   <2e-16 ***
newPPoints    22.158      1.511   14.66   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5086 on 70 degrees of freedom
Multiple R-squared:  0.7543,Adjusted R-squared:  0.7508
F-statistic: 214.9 on 1 and 70 DF,  p-value: < 2.2e-16
```

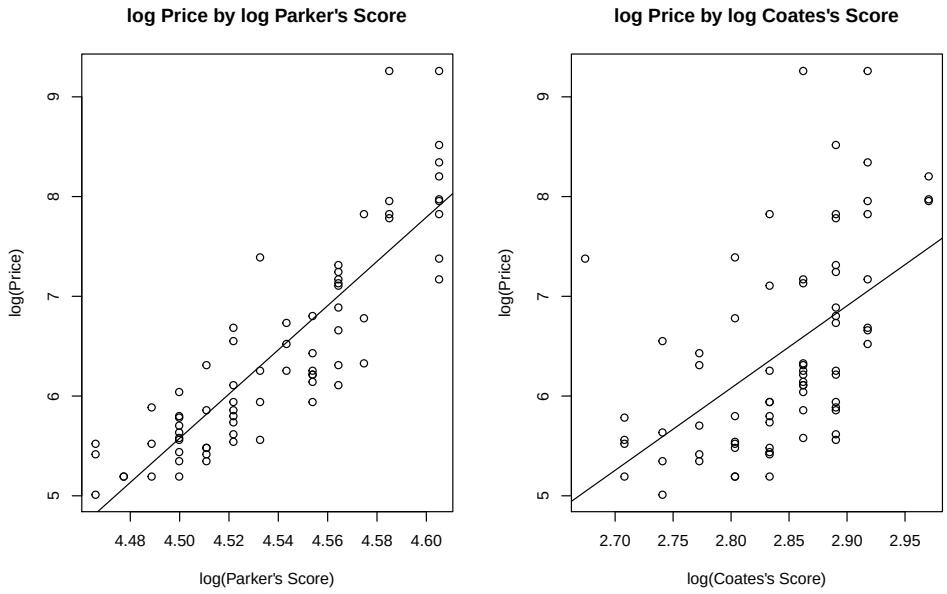
Using (Intercept) and newPPoints, and corresponding values in Coates' adjusted data, we find best fit lines

$log(Price) = -94.132 + 22.158(log(Parker's\ Score))$ $\quad$ $log(Price) = -17.014 + 8.248(log(Coates'\ Score))$

And, using the code

```
par(mfrow=c(2,2)) # We want two scatterplots, as in Part A
plot(newPPoints,newPrice, xlab="log(Parker's Score)",ylab="log(Price)",
 main="log Price by log Parker's Score") # First Plot
abline(newParkWine.fit$coefficients[1],newParkWine.fit$coefficients[2]) # Makes line on first plot
plot(newCPoints,newPrice, xlab="log(Coates's Score)",ylab="log(Price)",
 main="log Price by log Coates's Score") # Second Plot
abline(newCoatWine.fit$coefficients[1],newCoatWine.fit$coefficients[2]) # Makes line on second plot
```

we generate the following:



These new plots have much better lines of fit. Data is well distributed above and below the line, with no obvious patterns other than a clear upward linear trend in each. As before, there seems to be more variance as the predictive variable increases, but the effect is less drastic than before.

Part E

When we ask "who is the most influential wine critic," there are a few interpretations to consider. We might give the title to Parker, because in the log-log plots of Part D, his points seem to cluster around their line of best fit more regularly than the points of Coates' plot cluster around theirs. We might interpret this to mean that the wine industry considers Parker's opinion more valuable than Coates', and adheres to it more closely when determining their prices.

Alternatively, we could consider the opposite. What if Parker's scores are influenced (even subconsciously) by the wine's price? Then there would be a close relationship between his score and the price for a bottle of wine, but this relationship would not be indicative of Parker's influence on the French wine community. Parker's scores certainly seem like a more accurate reflection of the price, but without more information about how data was collected, it might be dangerous to conjecture which of the wine critics is truly more 'influential.'

Question 2, Part A

The least squares estimate of β is, by definition, the value for which minimizes the sum of the squares of the errors, $SSE = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \beta x_i)^2$. Calling this estimate $\hat{\beta}$, we understand

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

We assume in the problem that the regression goes through the origin, so $\overline{x}, \overline{y} = 0$, and therefore

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Part B

Knowing $\mathbb{E}[\hat{\beta}|X] = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} \times \mathbb{E}[y_i|X = x_i]$, and $\mathbb{E}[y_i|X = x_i] = \beta x_i$, we understand that $\mathbb{E}[\hat{\beta}|X] = \sum_{i=1}^{n} \frac{x_i}{x_i^2} \times \beta x_i = \beta \sum_{i=1}^{n} \frac{x_i^2}{x_i^2} = \beta$.

For variance, $\hat{\beta}[X] = \sum_{i=1}^{n} \left(\frac{x_i}{x_i^2}\right)^2 \times Var[y_i|X = x_1]$. We defined $\sigma^2 = Var[y_i|X = x_i]$, so this may be rewritten as

$$\sigma^2 \sum_{i=1}^{n} \frac{x_i^2}{x_i^4} = \frac{\sigma^2}{S_{XX}}$$