Final Project

Jake Summaria

5/7/2024

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.3.2
```

```
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 4.3.3
```

```
## Warning in check_dep_version(): ABI version mismatch:
## lme4 was built with Matrix ABI version 1
## Current Matrix ABI version is 0
## Please re-install lme4 from source or restore original 'Matrix' package
```

```
nba = read.csv("NBA.csv", header = TRUE, sep = ",")
```

```
model = lm(usg_pct ~ pts + ast + ts_pct + ast_pct + net_rating + conference, nba)
summary(model)
```

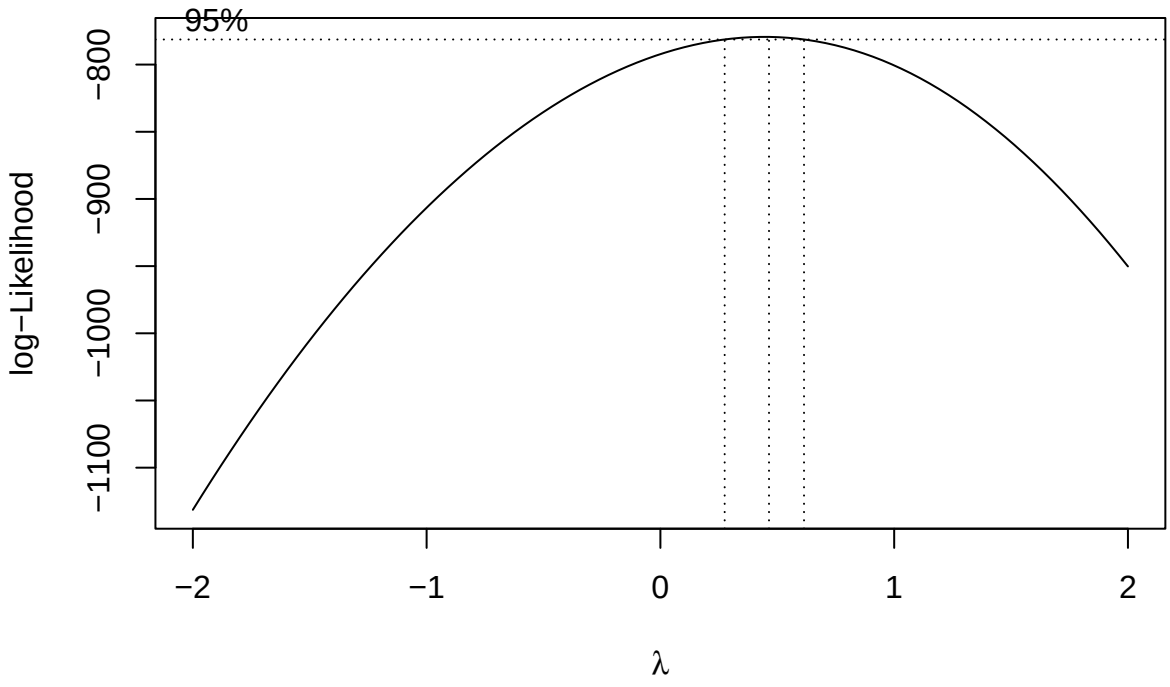
```
##
## Call:
## lm(formula = usg_pct ~ pts + ast + ts_pct + ast_pct + net_rating +
##     conference, data = nba)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09823 -0.01955 -0.00368  0.01459  0.31770
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1395530  0.0108668  12.842  < 2e-16 ***
## pts          0.0082326  0.0003456  23.818  < 2e-16 ***
## ast         -0.0208989  0.0019439 -10.751  < 2e-16 ***
## ts_pct      -0.0786704  0.0178397  -4.410 1.25e-05 ***
## ast_pct      0.3874925  0.0345301  11.222  < 2e-16 ***
```

```
## net_rating   0.0007285  0.0001331   5.473 6.85e-08 ***
## conference  -0.0010755  0.0029231  -0.368    0.713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03349 on 526 degrees of freedom
## Multiple R-squared:  0.6497, Adjusted R-squared:  0.6457
## F-statistic: 162.6 on 6 and 526 DF,  p-value: < 2.2e-16
```

---

a: For every 1 point increase in pts, I would estimate the average usg_pct to increase by 0.008, holding all other variables constant. b: For a player in the Eastern Conference, I would estimate the average usg_pct to be 0.0017 lower than a player in the Western Conference, holding all other variables constant. c: The conference variable has a p-value of 0.56, which is greater than the usual significance level of 0.05, which means that the conference variable is not statistically significant. d: The baseline level is when Conference = 0, which represents the Western Conference. ***

```
boxcox(model)
```




```
nba$new_usg_pct = sqrt(nba$usg_pct)
sqrt_model = lm(new_usg_pct ~ pts + ast + ts_pct + ast_pct + net_rating + conference, nba)
```

The Box-Cox suggests a square root transformation as lambda is roughly equal to 0.5. I will use this transformation for the remainder of my analysis to get a more normalized distribution and to stabilize variance. ***

```
model2 = lm(new_usg_pct ~ pts + ast + ts_pct + ast_pct + poly(net_rating, 2) + conference, nba)
summary(model2)
```

```
##
## Call:
## lm(formula = new_usg_pct ~ pts + ast + ts_pct + ast_pct + poly(net_rating,
##     2) + conference, data = nba)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12543 -0.02200 -0.00201  0.02077  0.14678
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.3500396  0.0113500  30.841  < 2e-16 ***
## pts                     0.0097005  0.0003597  26.971  < 2e-16 ***
## ast                    -0.0252552  0.0020197 -12.504  < 2e-16 ***
## ts_pct                 -0.0594423  0.0186225  -3.192   0.0015 **
## ast_pct                 0.4833386  0.0359390  13.449  < 2e-16 ***
## poly(net_rating, 2)1    0.1617455  0.0360128   4.491 8.71e-06 ***
## poly(net_rating, 2)2    0.4090085  0.0354114  11.550  < 2e-16 ***
## conference             -0.0025768  0.0030390  -0.848   0.3969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03479 on 525 degrees of freedom
## Multiple R-squared:  0.7117, Adjusted R-squared:  0.7078
## F-statistic: 185.1 on 7 and 525 DF,  p-value: < 2.2e-16
```

I created a polynomial term for the net_rating variable, and based on the summary, the polynomial terms are statistically significant, which means the polynomial transformation contributes to variation in the new_usg_pct variable. ***

```
model3 = step(model2, direction = 'backward')
```

```
## Start:  AIC=-3572.07
## new_usg_pct ~ pts + ast + ts_pct + ast_pct + poly(net_rating,
##     2) + conference
##
##                        Df Sum of Sq     RSS     AIC
## - conference            1   0.00087 0.63639 -3573.3
## <none>                              0.63552 -3572.1
## - ts_pct                1   0.01233 0.64785 -3563.8
## - poly(net_rating, 2)   2   0.18879 0.82431 -3437.4
## - ast                   1   0.18927 0.82478 -3435.1
## - ast_pct               1   0.21895 0.85446 -3416.3
```

```
## - pts                  1   0.88057 1.51609 -3110.7
##
## Step:  AIC=-3573.34
## new_usg_pct ~ pts + ast + ts_pct + ast_pct + poly(net_rating,
##     2)
##
##                        Df Sum of Sq     RSS     AIC
## <none>                              0.63639 -3573.3
## - ts_pct                1   0.01220 0.64859 -3565.2
## - poly(net_rating, 2)   2   0.18793 0.82431 -3439.4
## - ast                   1   0.19121 0.82759 -3435.3
## - ast_pct               1   0.22283 0.85921 -3415.3
## - pts                   1   0.88192 1.51831 -3111.9
```

For the model selection process, I chose to use a backward selection with AIC metric. The resulting model is new_usg_pct ~ pts + ast + ts_pct + ast_pct + poly(net_rating, 2) ***

```
summary(model2)$r.squared
```

```
## [1] 0.7116919
```

```
summary(sqrt_model)$r.squared
```

```
## [1] 0.6384305
```

The R^2 of the model with the polynomial term is 0.712, which means that 71.2% of the variance in new_usg_pct is explained by the predictor variables, and the R^2 of the sqrt model is 0.638, which means that 63.8% of the variance in new_usg_pct is explained by the predictor variables. ***

---

a: I have decided to go with the model that was selected from the model selection process because it only includes variables that are significant in the variance of usg_pct. ***

```
summary(model3)$coef
```

```
##                          Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)           0.348353497 0.0111714440  31.182495 1.168975e-121
## pts                   0.009706199 0.0003595023  26.998993 2.185318e-101
## ast                  -0.025347303 0.0020162727 -12.571367  7.073631e-32
## ts_pct               -0.059108947 0.0186133542  -3.175620  1.582810e-03
## ast_pct               0.485890009 0.0358032307  13.571122  3.484087e-36
## poly(net_rating, 2)1  0.159692189 0.0359216647   4.445568  1.069848e-05
## poly(net_rating, 2)2  0.407782277 0.0353724123  11.528257  1.429834e-27
```

---

b: fitted model: new_usg_pctˆ = 0.3484 + 0.0097 x pts - 0.0253 x ast - 0.0591 x ts_pct + 0.4859 x ast_pct + 0.1597 x net_rating + 0.4078 x (net_rating)ˆ2 ***

```
dim(nba)[1]
```

```
## [1] 533
```

---

c: n = 533, p = 7 ***

```
sd(nba$new_usg_pct)
```

```
## [1] 0.06436933
```

```
sd(residuals(model3))
```

```
## [1] 0.03458633
```

---

d: standard deviation of new_usg_pct = 0.0644 standard deviation of model3 = 0.0346 Because the estimated standard deviation of the model is much lower than the standard deviation of new_usg_pct, it suggests that the model is providing a good fit to the data and that the predictors that are included are useful and significant. ***

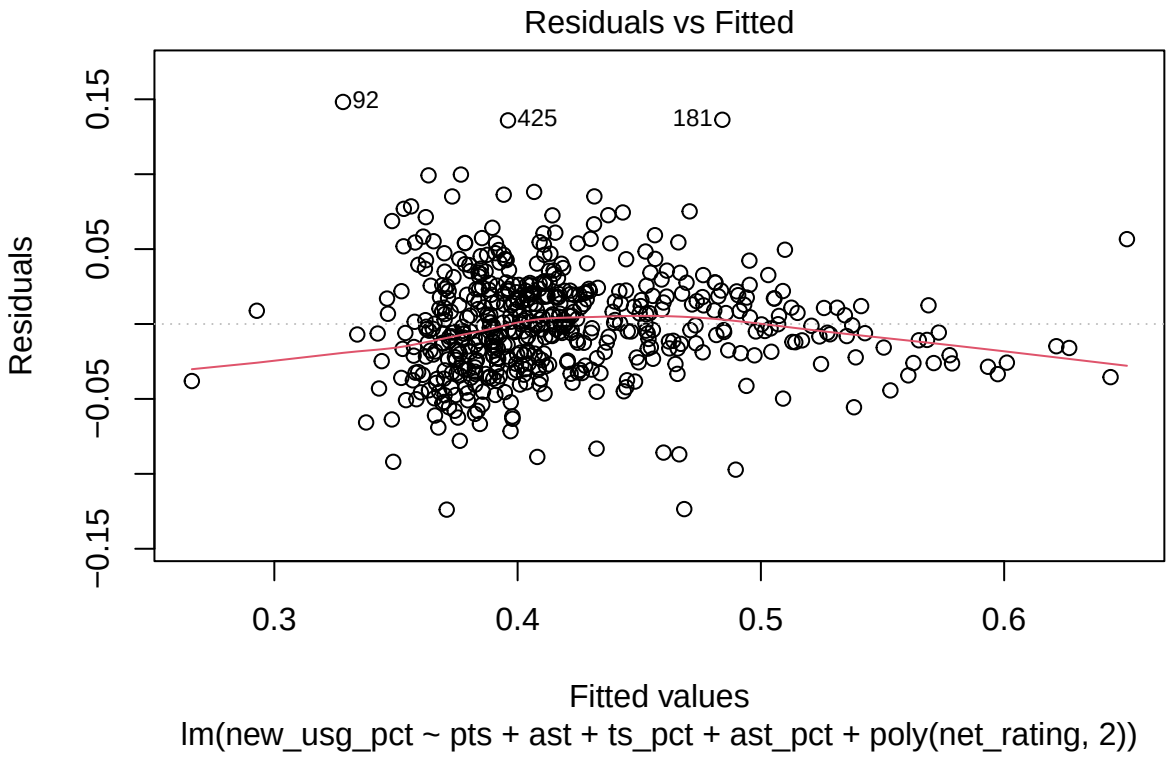
```
vif(model3)
```

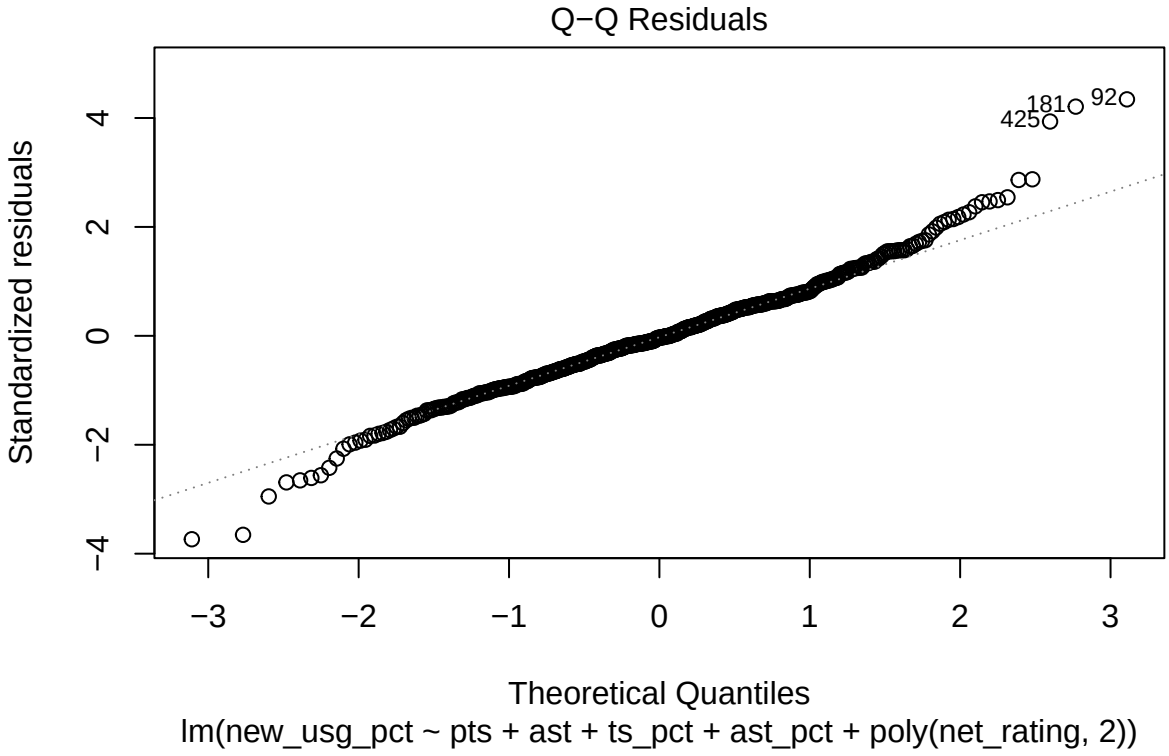
```
##                  pts                  ast               ts_pct
##             2.648750             6.692520             1.217779
##              ast_pct poly(net_rating, 2)1 poly(net_rating, 2)2
##             4.187188             1.066542             1.034176
```

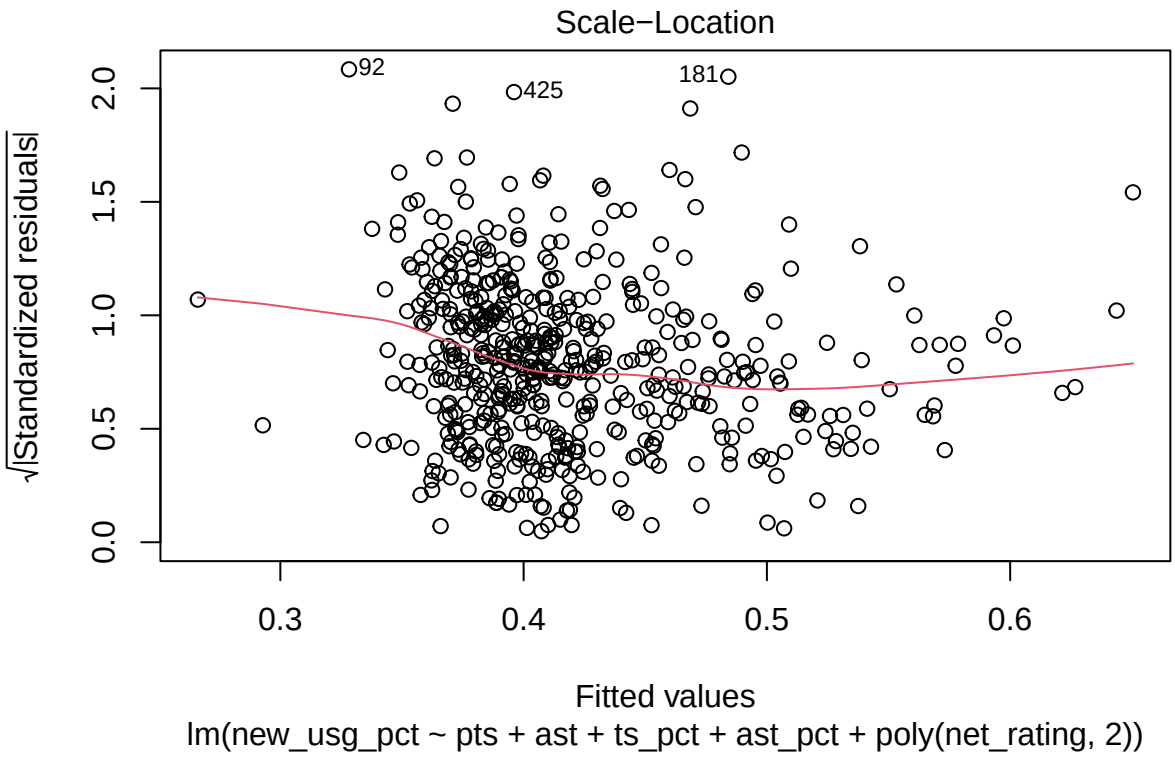
---

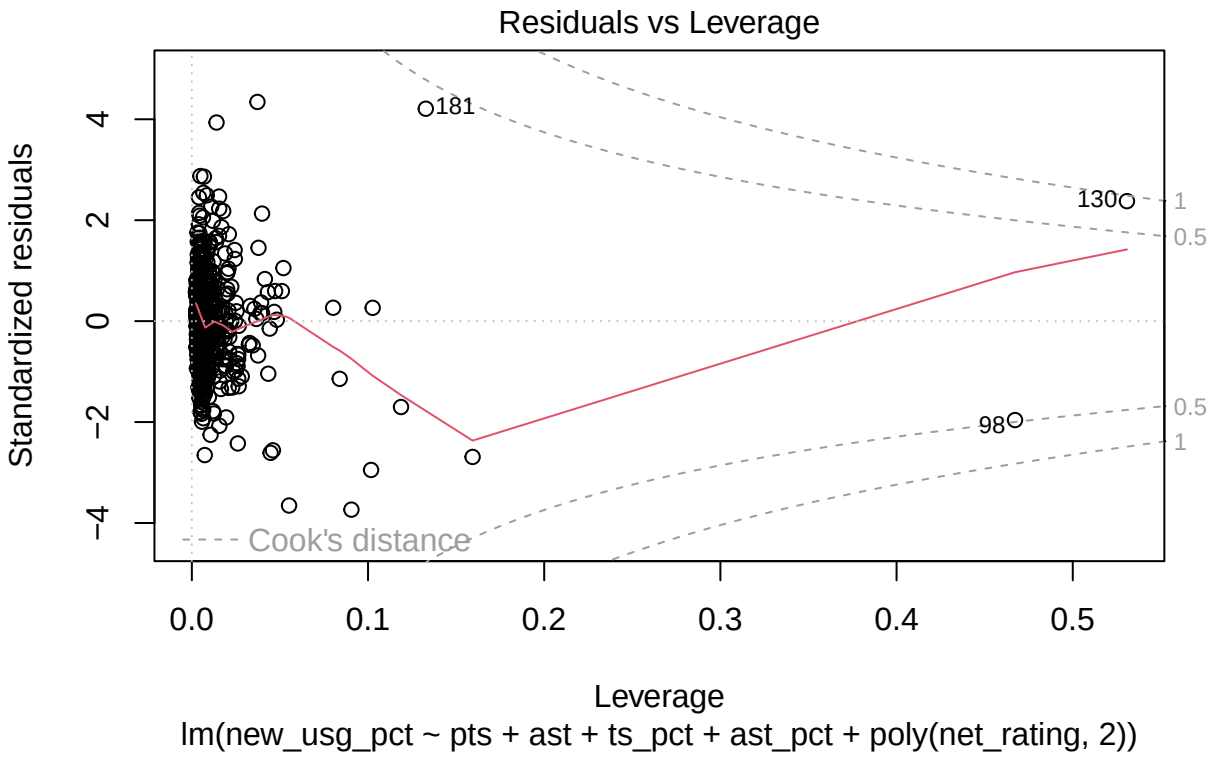
e: Based on the results, the only variable that raises some concern over collinearity is the ast variable, because it's VIF value 6.693 > 5, which is a problem. ***

```
plot(model3)
```


Residuals vs Fitted
Residuals
0.15
0.05
−0.05
−0.15
92
425
181
0.3
0.4
0.5
0.6
Fitted values
lm(new_usg_pct ~ pts + ast + ts_pct + ast_pct + poly(net_rating, 2))


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(new_usg_pct ~ pts + ast + ts_pct + ast_pct + poly(net_rating, 2))
181
92
425
−4
−2
0
2
4
−3
−2
−1
0
1
2
3


Scale−Location
92
425
181
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0.3
0.4
0.5
0.6
Fitted values
lm(new_usg_pct ~ pts + ast + ts_pct + ast_pct + poly(net_rating, 2))



f: The Residuals vs Fitted model has a red line that is not very flat, which means that there might not be a linear relationship between the new_usg_pct variable and all of the predictor variables. Also, this shows that the spread of residuals narrows as the fitted values increase, which suggests heteroscedasiticity, violating the assumption of equal variance. The QQ plot shows that the residuals follow the predicted line well, which suggests that the residuals follow a normal distribution. The scale location plot shows that the spread of residuals closely resembles a funnel, which means that the variability of the residuals is not constant. The residuals vs leverage plot has a line that is not flat, which also indicates that the linear assumption is not met.

---

g: unusual observations: based on the scale-location plot, there are three observations that stand out, labeled as 181, 130, and 98. Because a couple of them have a Cook's Distance of about 0.5, I would fit a new model that would exclude these observations due to them having a high influence. ***

```
sqrt(mean((resid(model3) / (1 - hatvalues(model3))) ^ 2))
```

```
## [1] 0.03586181
```

---

h: estimated errors: 0.0358 This means that the model's predictions have an average error of about 0.03586 units, and because it is a small value, it suggets that the model is making reasonably accurate predictions. ***

---

i: model complexity: n / p = 533 / 7 > 10 Using the rule of thumb of having at least 10 observations for every coefficient, there is no concern for the model complexity. ***

```
summary(model3)
```

```
##
## Call:
## lm(formula = new_usg_pct ~ pts + ast + ts_pct + ast_pct + poly(net_rating,
##     2), data = nba)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.123887 -0.021749 -0.001267  0.019955  0.148211
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.3483535  0.0111714  31.182  < 2e-16 ***
## pts                   0.0097062  0.0003595  26.999  < 2e-16 ***
## ast                  -0.0253473  0.0020163 -12.571  < 2e-16 ***
## ts_pct               -0.0591089  0.0186134  -3.176  0.00158 **
## ast_pct               0.4858900  0.0358032  13.571  < 2e-16 ***
## poly(net_rating, 2)1  0.1596922  0.0359217   4.446 1.07e-05 ***
## poly(net_rating, 2)2  0.4077823  0.0353724  11.528  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03478 on 526 degrees of freedom
## Multiple R-squared:  0.7113, Adjusted R-squared:  0.708
## F-statistic:   216 on 6 and 526 DF,  p-value: < 2.2e-16
```

Null Hypothesis: There is no significant relationship between the new_usg_pct variable and the pts variable. Alternative Hypothesis: There is a significant relationship between the new_usg_pct variable and the pts variable. test statistic = 26.999 p-value = < 2e-16 Based on the p-value being less than the threshold of 0.05, I would reject the null hypothesis and conclude that there is a significant relationship between the new_usg_pct variable and the pts variable. ***