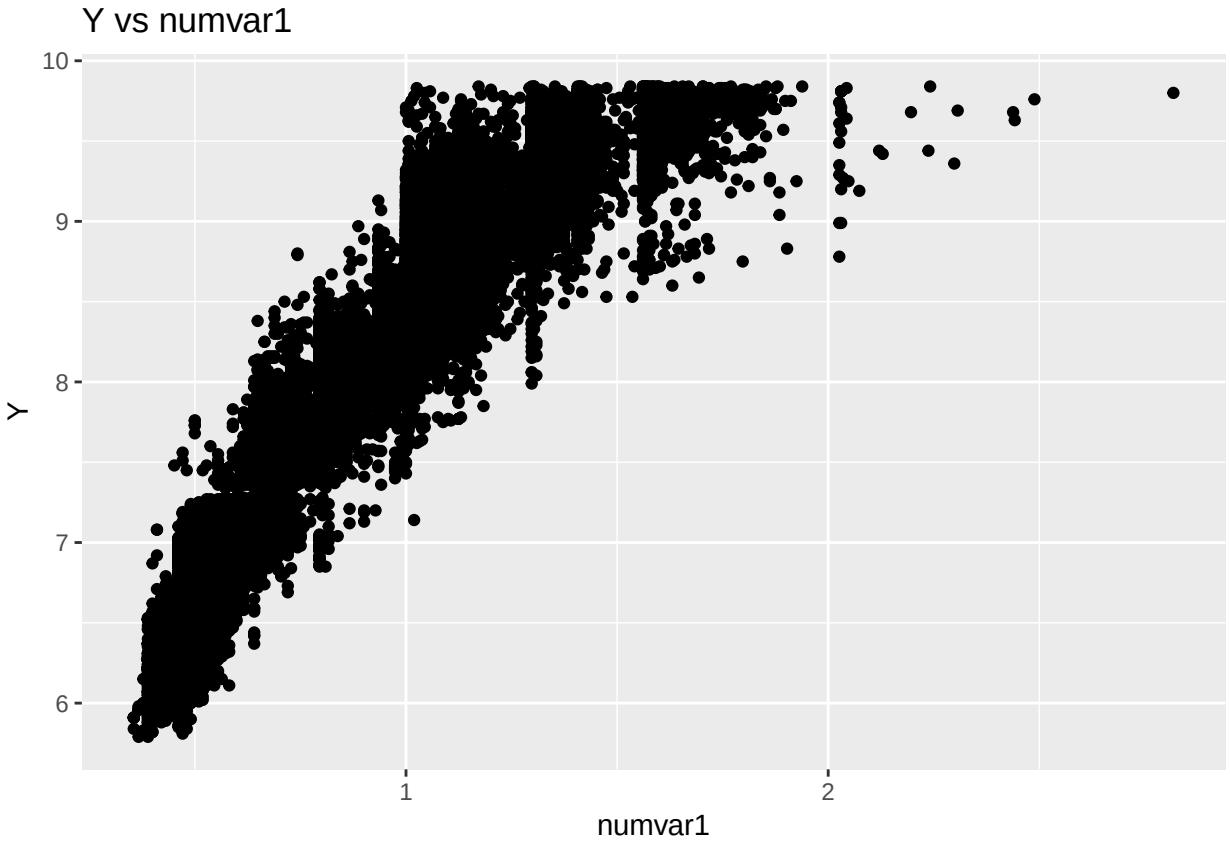
Linear Regression Assignment

Samuel Ivanecky

February 1, 2019

Part 1

Part 1.1



Based on the scatterplot, the variables $numvar1$ and Y appear to be associated; as $numvar1$ increases, Y also increases. The association between the two variables appears to be somewhat linear but there is noticeable curvature of the data which would not be well represented using a linear model. Using a $Y = \beta_0 + \beta_1 * numvar1 + \epsilon$ model could give a decent representation of the data but due to the curvature in shown in the plot, a polynomial model would be a more adequate choice to account for this curving.

Part 1.2

```
##
## Call:
## lm(formula = Y ~ numvar1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.0135 -0.1934  0.0149  0.2016  1.4574
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.259101   0.004573  1149.9   <2e-16 ***
## numvar1     3.035931   0.005130   591.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3249 on 39998 degrees of freedom
## Multiple R-squared:  0.8975, Adjusted R-squared:  0.8975
## F-statistic: 3.503e+05 on 1 and 39998 DF,  p-value: < 2.2e-16
```

Two parameters are estimated for the model. The first parameter is $\beta_0 = 5.259$ and the second is $\beta_1 = 3.0359$. The parameter β_0 represents the y-intercept for the linear model which indicates that the predicted value of Y is 5.259 if the value of $numvar1 = 0$. The second parameter, β_1 is the slope coefficient for the linear model. The value of 3.0359 indicates that increasing $numvar1$ by 1 will result in a 3.0359 increase in Y. The R^2 value of 0.8975 indicates the linear model accounts for approximately 90% of the variability in the response variable. In short, a high value of R^2 (close to 1) indicates the model is an appropriate fit for the dataset.

Part 2

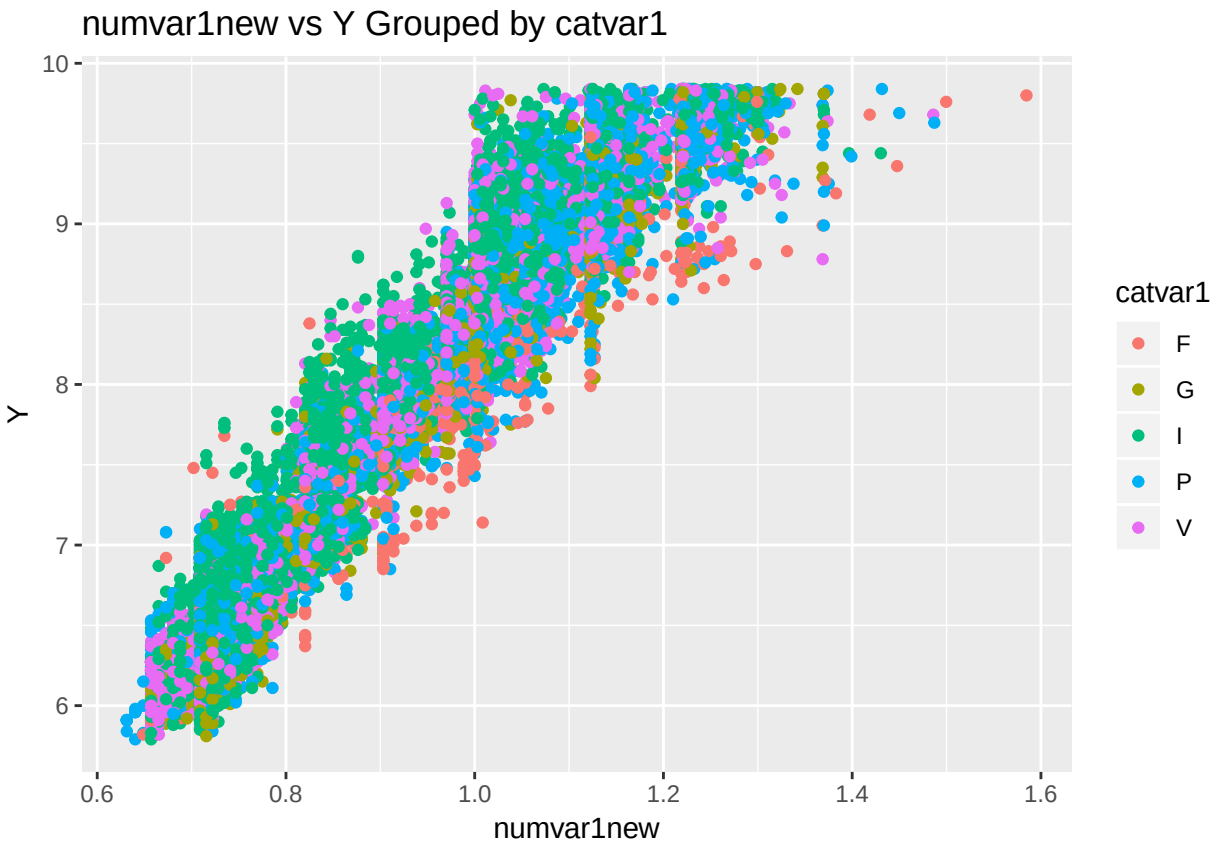
Part 2.1

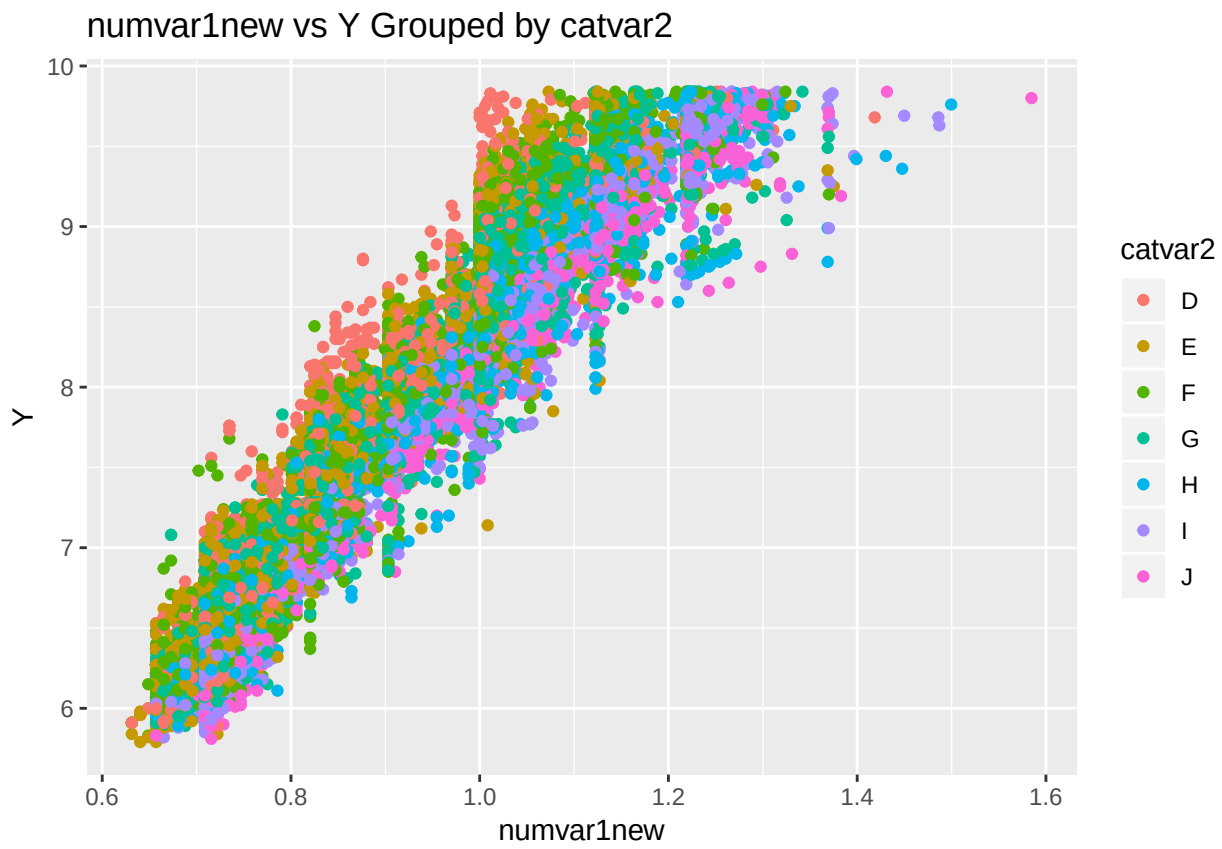
```
##
## Call:
## lm(formula = Y ~ numvar1new, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.33834 -0.16978  0.00022  0.17189  1.36788
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.979233   0.008311   238.1   <2e-16 ***
## numvar1new  6.410648   0.009043   708.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2755 on 39998 degrees of freedom
## Multiple R-squared:  0.9263, Adjusted R-squared:  0.9263
## F-statistic: 5.025e+05 on 1 and 39998 DF,  p-value: < 2.2e-16
```

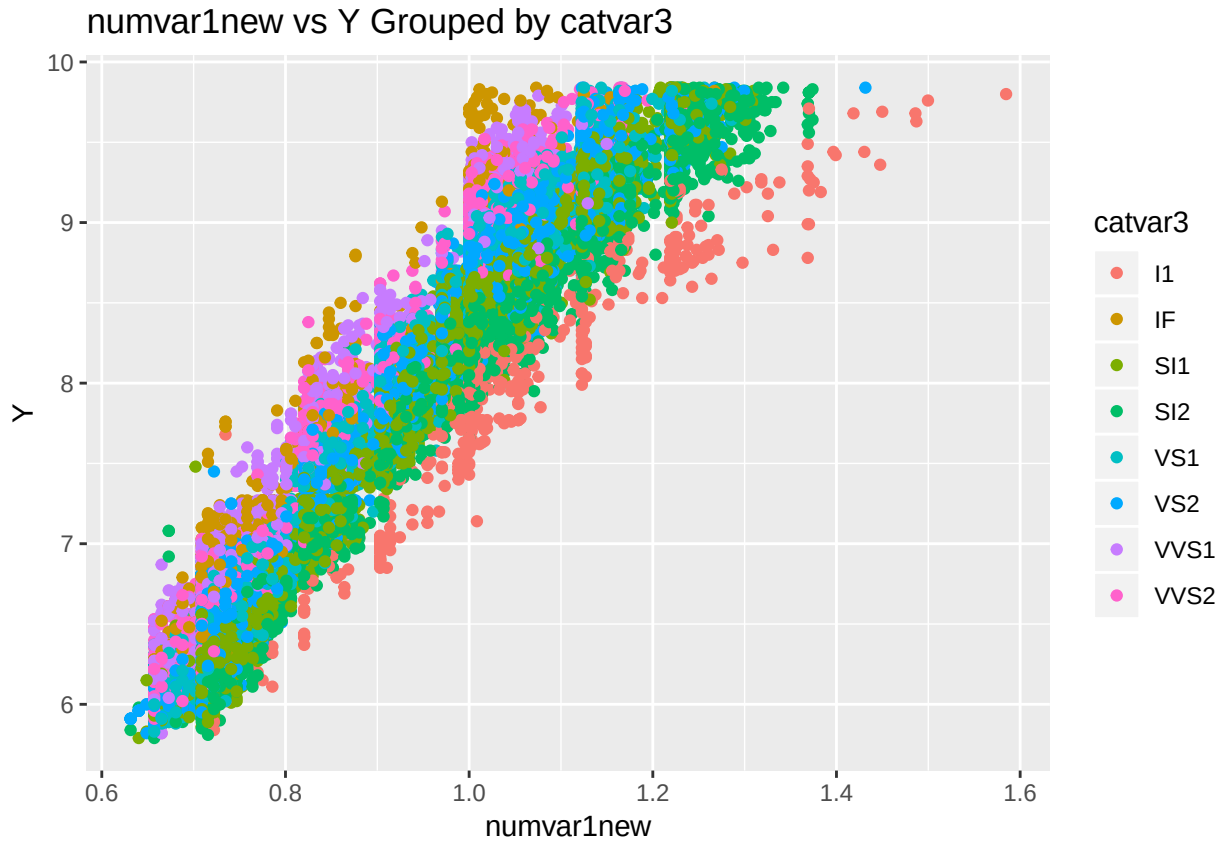
The second linear model, model2, also has two estimated parameters. For model2 the estimated parameters are $\beta_0 = 1.9792$ and $\beta_1 = 6.4106$. The value of β_0 indicates that if the value of $numvar1new = 0$, the predicted value of Y is 1.9792. The value of β_1 indicates that for a singular increase in $numvar1new$, the predicted value of Y would increase by 6.4106.

The R^2 value for model2 is 0.9263, indicating roughly 93% of the variability in the response variable is accounted for. Given that the R^2 for the model2 is greater than that for model1, model2 appears to be a better fit for the dataset.

Part 3


numvar1new vs Y Grouped by catvar1
Y
numvar1new
10
9
8
7
6
0.6
0.8
1.0
1.2
1.4
1.6
catvar1
F
G
I
P
V


numvar1new vs Y Grouped by catvar2
Y
numvar1new
10
9
8
7
6
0.6
0.8
1.0
1.2
1.4
1.6
catvar2
D
E
F
G
H
I
J

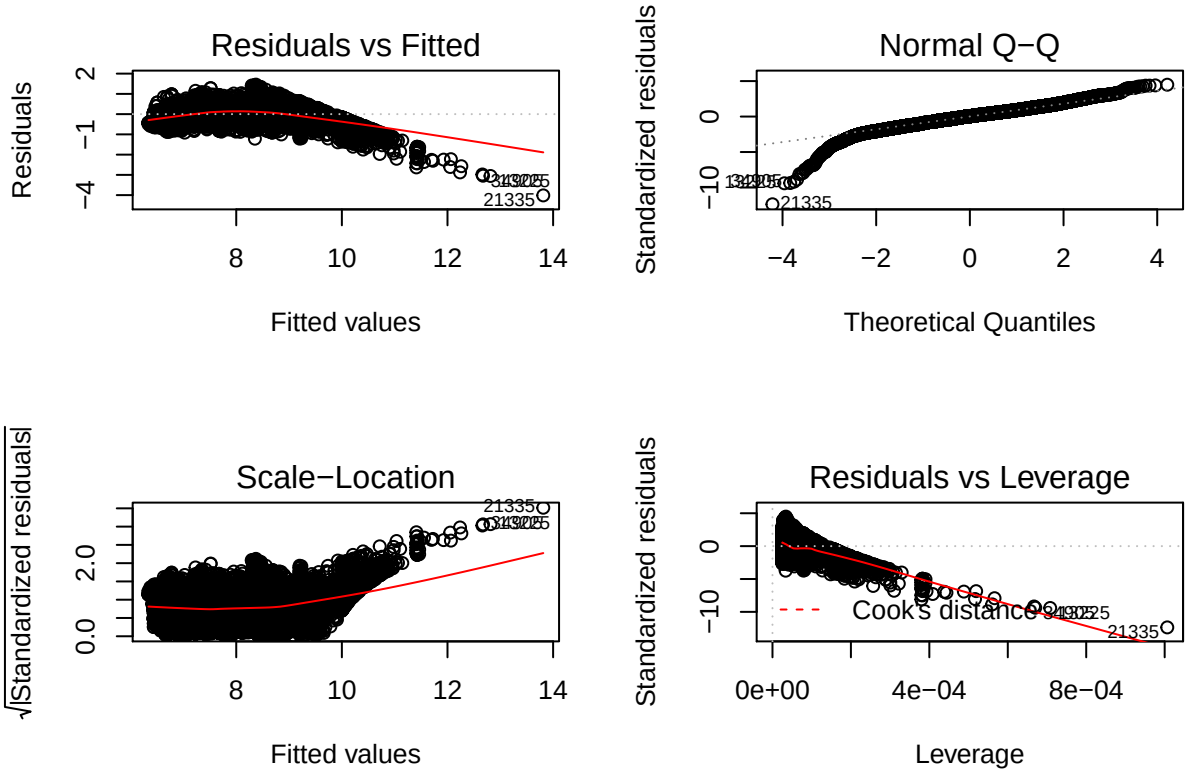


Based on the three scatterplots, including both *catvar*2 and *catvar*3 could be beneficial in producing a better linear model. The plot of *catvar*1 appears to have a random distribution among the different categories, showing no clear trends defined. For both *catvar*2 and *catvar*3, there appears to separation of the values based on which category they belong to, indicated by the layers of color in each respective plot. These categorical variables appear to have an association with the response variable Y and thus should be included in the model.

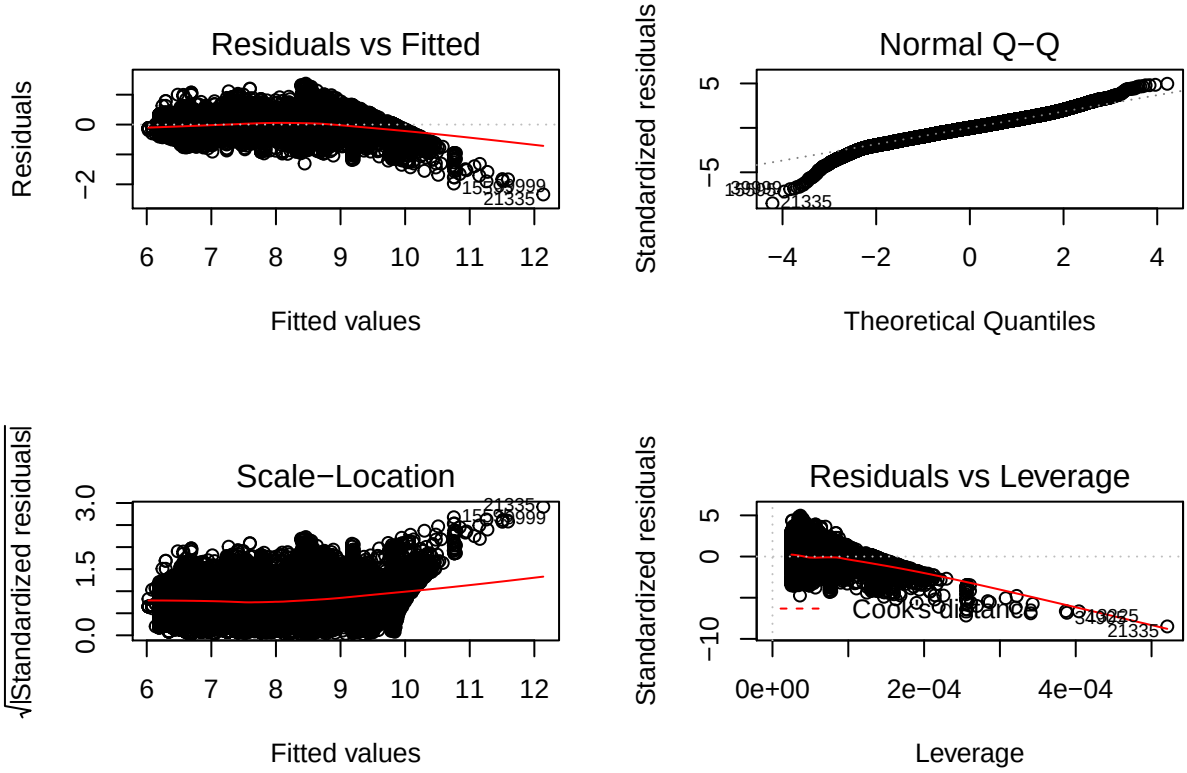
Part 4

```
##
## Call:
## lm(formula = Y ~ numvar1, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0135 -0.1934  0.0149  0.2016  1.4574
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.259101   0.004573  1149.9   <2e-16 ***
## numvar1     3.035931   0.005130   591.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3249 on 39998 degrees of freedom
```

```
## Multiple R-squared:  0.8975, Adjusted R-squared:  0.8975
## F-statistic: 3.503e+05 on 1 and 39998 DF,  p-value: < 2.2e-16
```

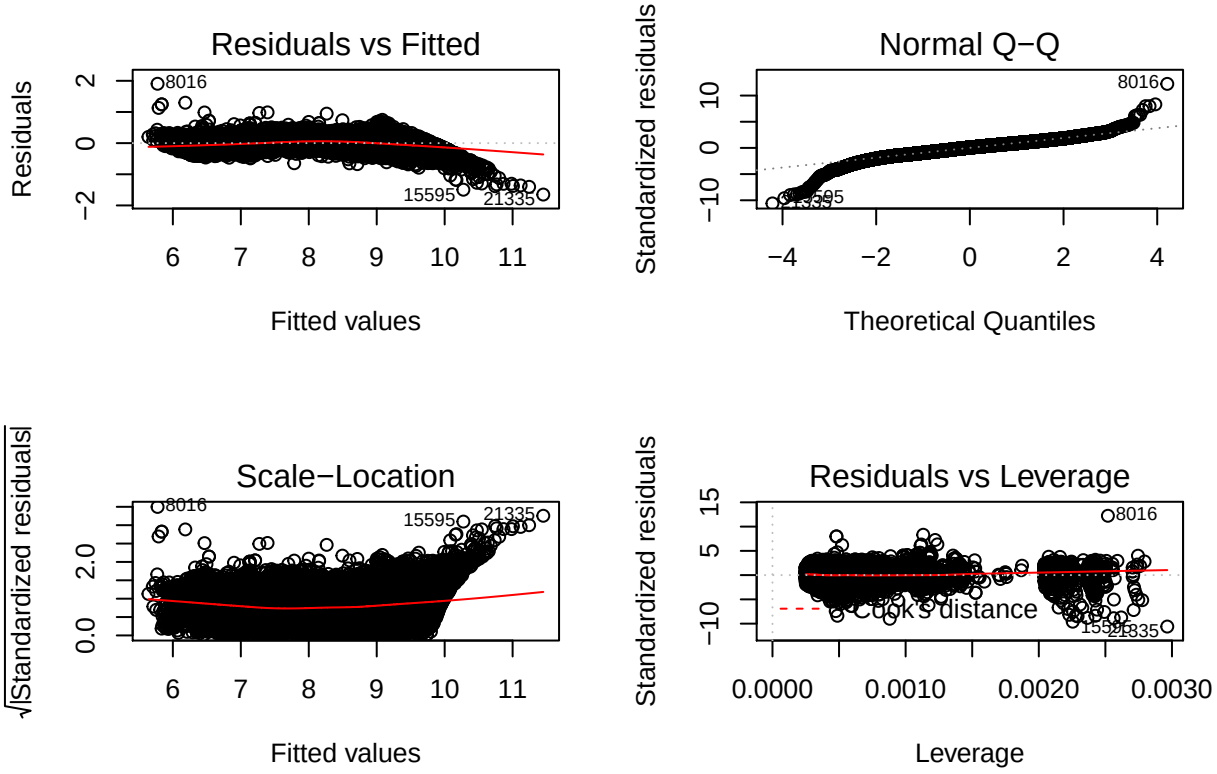



```
##
## Call:
## lm(formula = Y ~ numvar1new, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.33834 -0.16978  0.00022  0.17189  1.36788
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.979233   0.008311   238.1   <2e-16 ***
## numvar1new  6.410648   0.009043   708.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2755 on 39998 degrees of freedom
## Multiple R-squared:  0.9263, Adjusted R-squared:  0.9263
## F-statistic: 5.025e+05 on 1 and 39998 DF,  p-value: < 2.2e-16
```



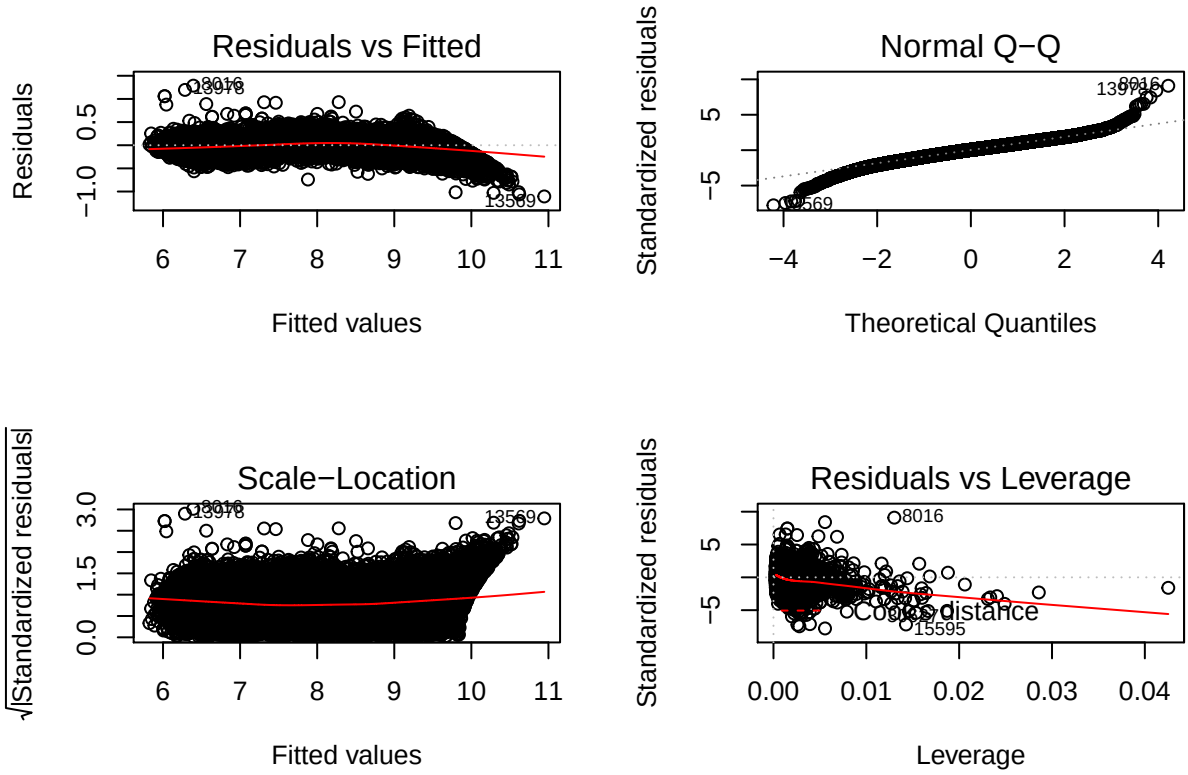

```
##
## Call:
## lm(formula = Y ~ numvar1new + catvar1 + catvar2 + catvar3, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65108 -0.09676  0.01071  0.10099  1.90498
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.565527   0.009772   57.87   <2e-16 ***
## numvar1new   7.216071   0.005876 1228.12   <2e-16 ***
## catvar1G     0.072865   0.005225   13.95   <2e-16 ***
## catvar1I     0.145861   0.004751   30.70   <2e-16 ***
## catvar1P     0.115007   0.004794   23.99   <2e-16 ***
## catvar1V     0.101861   0.004849   21.01   <2e-16 ***
## catvar2E    -0.056851   0.002859  -19.89   <2e-16 ***
## catvar2F    -0.092486   0.002890  -32.00   <2e-16 ***
## catvar2G    -0.161896   0.002827  -57.26   <2e-16 ***
## catvar2H    -0.268952   0.003007  -89.44   <2e-16 ***
## catvar2I    -0.406372   0.003374 -120.46   <2e-16 ***
## catvar2J    -0.549929   0.004150 -132.52   <2e-16 ***
## catvar3IF    1.117266   0.008167  136.81   <2e-16 ***
## catvar3SI1   0.643572   0.006970   92.34   <2e-16 ***
## catvar3SI2   0.461188   0.007006   65.83   <2e-16 ***
## catvar3VS1   0.851408   0.007114  119.69   <2e-16 ***
```

```
## catvar3VS2   0.782028   0.007007  111.60   <2e-16 ***
## catvar3VVS1  1.022110   0.007547  135.43   <2e-16 ***
## catvar3VVS2  0.966962   0.007335  131.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.156 on 39981 degrees of freedom
## Multiple R-squared:  0.9764, Adjusted R-squared:  0.9764
## F-statistic: 9.185e+04 on 18 and 39981 DF,  p-value: < 2.2e-16
```




```
##
## Call:
## lm(formula = Y ~ numvar1new * catvar1 + numvar1new * catvar2 +
##     numvar1new * catvar3, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.10592 -0.08659  0.00679  0.09189  1.28637
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.58687    0.05254  49.239  < 2e-16 ***
## numvar1new             5.30369    0.05102 103.960  < 2e-16 ***
## catvar1G              -0.48638    0.03536 -13.754  < 2e-16 ***
## catvar1I              -0.43293    0.03277 -13.211  < 2e-16 ***
## catvar1P              -0.13822    0.03291  -4.200 2.67e-05 ***
```

```
## catvar1V               -0.53776    0.03328 -16.159  < 2e-16 ***
## catvar2E               -0.04231    0.01764  -2.398  0.01651 *
## catvar2F               -0.05719    0.01770  -3.231  0.00123 **
## catvar2G               -0.02774    0.01692  -1.639  0.10113
## catvar2H                0.02593    0.01749   1.482  0.13824
## catvar2I                0.02692    0.01909   1.410  0.15841
## catvar2J                0.11210    0.02417   4.639 3.51e-06 ***
## catvar3IF              -0.82883    0.05461 -15.176  < 2e-16 ***
## catvar3SI1             -1.09504    0.04819 -22.724  < 2e-16 ***
## catvar3SI2             -0.84057    0.04861 -17.291  < 2e-16 ***
## catvar3VS1             -1.03305    0.04862 -21.246  < 2e-16 ***
## catvar3VS2             -0.97628    0.04817 -20.268  < 2e-16 ***
## catvar3VVS1            -0.96048    0.05119 -18.763  < 2e-16 ***
## catvar3VVS2            -1.02395    0.04965 -20.623  < 2e-16 ***
## numvar1new:catvar1G     0.56517    0.03596  15.715  < 2e-16 ***
## numvar1new:catvar1I     0.59822    0.03315  18.044  < 2e-16 ***
## numvar1new:catvar1P     0.24235    0.03318   7.303 2.86e-13 ***
## numvar1new:catvar1V     0.65931    0.03368  19.575  < 2e-16 ***
## numvar1new:catvar2E    -0.01249    0.02020  -0.618  0.53642
## numvar1new:catvar2F    -0.04479    0.02000  -2.239  0.02513 *
## numvar1new:catvar2G    -0.15813    0.01908  -8.288  < 2e-16 ***
## numvar1new:catvar2H    -0.32205    0.01937 -16.626  < 2e-16 ***
## numvar1new:catvar2I    -0.46370    0.02066 -22.446  < 2e-16 ***
## numvar1new:catvar2J    -0.68233    0.02503 -27.261  < 2e-16 ***
## numvar1new:catvar3IF    1.92007    0.05615  34.195  < 2e-16 ***
## numvar1new:catvar3SI1   1.64938    0.04594  35.901  < 2e-16 ***
## numvar1new:catvar3SI2   1.21107    0.04621  26.207  < 2e-16 ***
## numvar1new:catvar3VS1   1.81584    0.04661  38.954  < 2e-16 ***
## numvar1new:catvar3VS2   1.67574    0.04598  36.442  < 2e-16 ***
## numvar1new:catvar3VVS1  1.96719    0.05096  38.601  < 2e-16 ***
## numvar1new:catvar3VVS2  1.96242    0.04828  40.650  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1424 on 39964 degrees of freedom
## Multiple R-squared:  0.9803, Adjusted R-squared:  0.9803
## F-statistic: 5.686e+04 on 35 and 39964 DF,  p-value: < 2.2e-16
```




```
##
## Call:
## lm(formula = Y ~ numvar2new + numvar1new * catvar1 + numvar1new *
##     catvar2 + numvar1new * catvar3, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.09143 -0.08660  0.00656  0.09134  1.27418
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.42462    0.05382  45.050  < 2e-16 ***
## numvar2new             0.18074    0.01358  13.307  < 2e-16 ***
## numvar1new             4.84960    0.06128  79.133  < 2e-16 ***
## catvar1G              -0.46622    0.03532 -13.201  < 2e-16 ***
## catvar1I              -0.41251    0.03274 -12.601  < 2e-16 ***
## catvar1P              -0.11863    0.03287  -3.609 0.000307 ***
## catvar1V              -0.51787    0.03324 -15.580  < 2e-16 ***
## catvar2E              -0.04254    0.01761  -2.416 0.015698 *
## catvar2F              -0.05960    0.01766  -3.375 0.000740 ***
## catvar2G              -0.03132    0.01689  -1.855 0.063651 .
## catvar2H               0.02136    0.01746   1.224 0.221042
## catvar2I               0.02275    0.01905   1.194 0.232322
## catvar2J               0.10409    0.02412   4.316 1.60e-05 ***
## catvar3IF             -0.80893    0.05451 -14.839  < 2e-16 ***
## catvar3SI1            -1.08174    0.04809 -22.492  < 2e-16 ***
```

```
## catvar3SI2             -0.83368    0.04851 -17.186  < 2e-16 ***
## catvar3VS1             -1.01643    0.04853 -20.943  < 2e-16 ***
## catvar3VS2             -0.96179    0.04807 -20.006  < 2e-16 ***
## catvar3VVS1            -0.94401    0.05109 -18.476  < 2e-16 ***
## catvar3VVS2            -1.00524    0.04956 -20.282  < 2e-16 ***
## numvar1new:catvar1G     0.54123    0.03593  15.064  < 2e-16 ***
## numvar1new:catvar1I     0.57012    0.03315  17.199  < 2e-16 ***
## numvar1new:catvar1P     0.21407    0.03318   6.452 1.12e-10 ***
## numvar1new:catvar1V     0.63407    0.03366  18.836  < 2e-16 ***
## numvar1new:catvar2E    -0.01252    0.02016  -0.621 0.534376
## numvar1new:catvar2F    -0.04217    0.01996  -2.113 0.034594 *
## numvar1new:catvar2G    -0.15377    0.01904  -8.076 6.88e-16 ***
## numvar1new:catvar2H    -0.31638    0.01933 -16.366  < 2e-16 ***
## numvar1new:catvar2I    -0.45844    0.02062 -22.236  < 2e-16 ***
## numvar1new:catvar2J    -0.67295    0.02498 -26.935  < 2e-16 ***
## numvar1new:catvar3IF    1.89826    0.05605  33.866  < 2e-16 ***
## numvar1new:catvar3SI1   1.63599    0.04585  35.679  < 2e-16 ***
## numvar1new:catvar3SI2   1.20408    0.04611  26.111  < 2e-16 ***
## numvar1new:catvar3VS1   1.79851    0.04653  38.652  < 2e-16 ***
## numvar1new:catvar3VS2   1.66107    0.04590  36.192  < 2e-16 ***
## numvar1new:catvar3VVS1  1.95047    0.05087  38.346  < 2e-16 ***
## numvar1new:catvar3VVS2  1.94268    0.04819  40.311  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1421 on 39963 degrees of freedom
## Multiple R-squared:  0.9804, Adjusted R-squared:  0.9804
## F-statistic: 5.553e+04 on 36 and 39963 DF,  p-value: < 2.2e-16
```

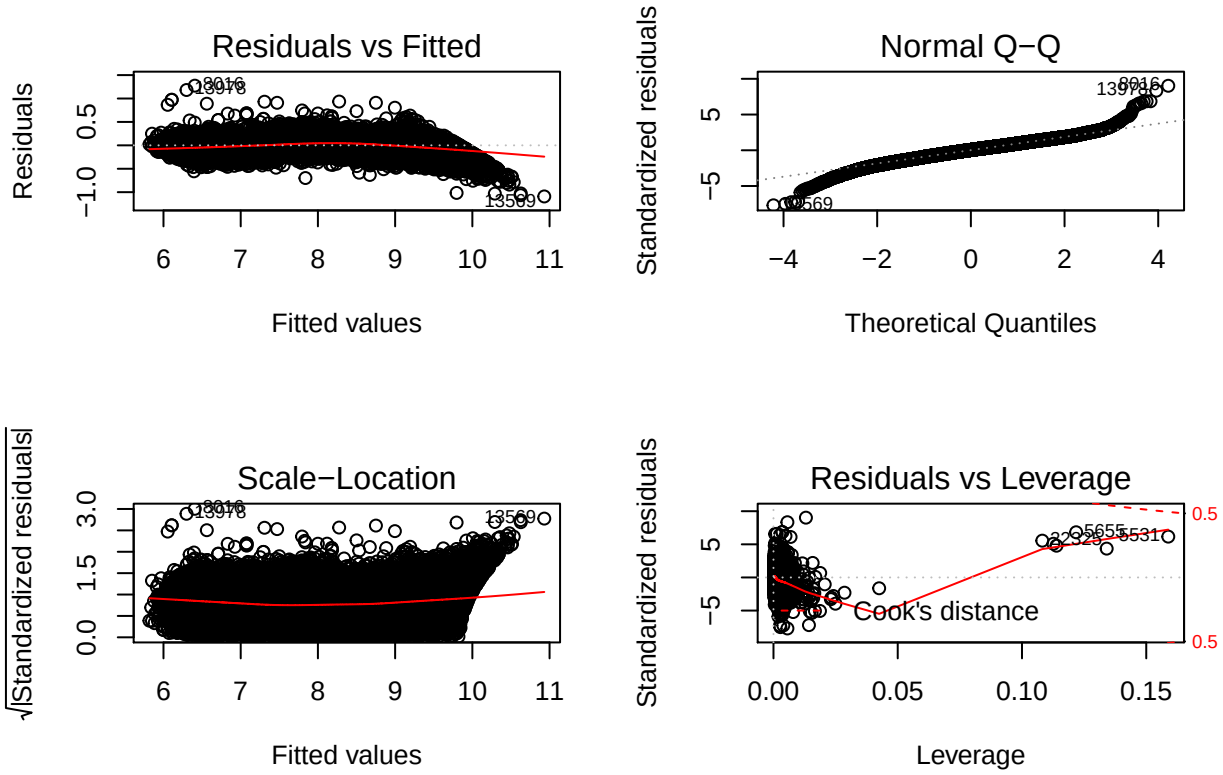
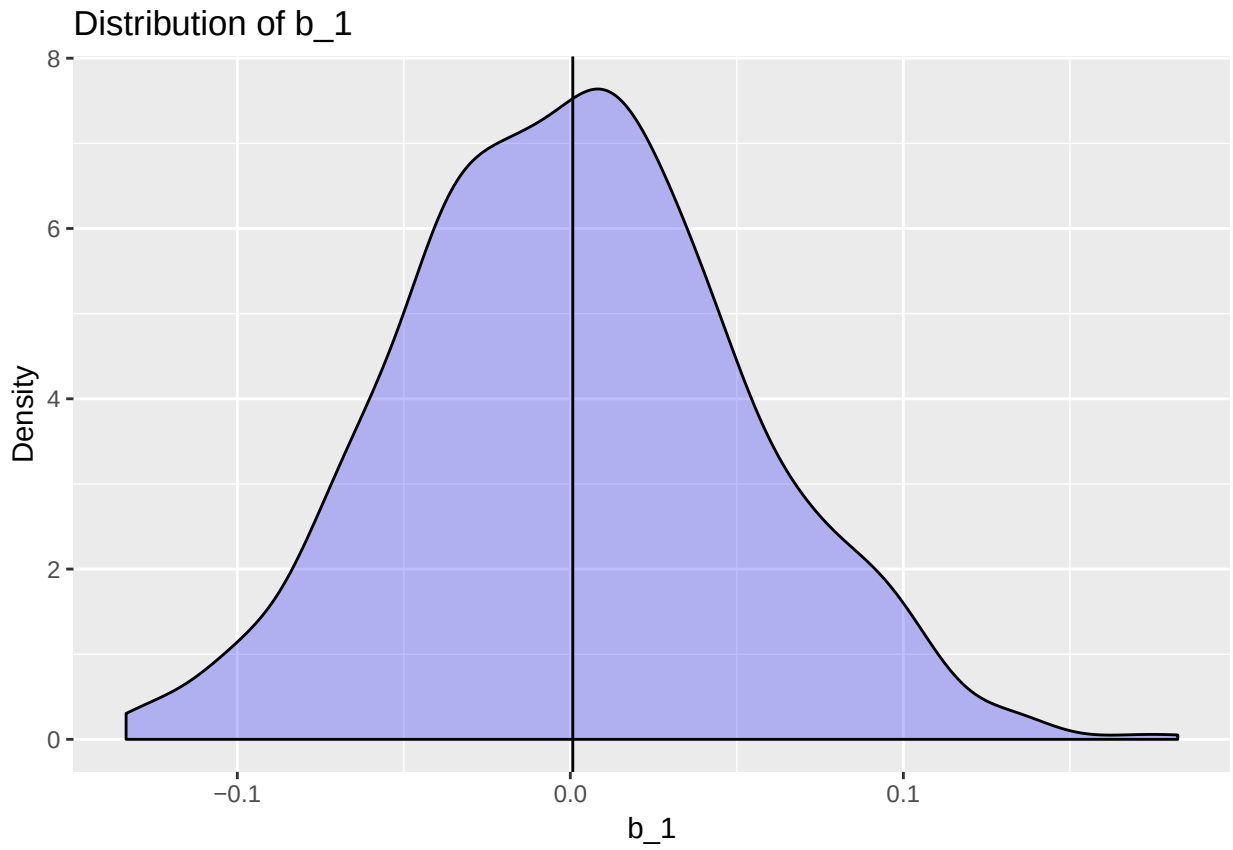


Table 1: R-Squared values for all models

Model1	Model2	Model3	Model4	Model5
0.8975	0.9263	0.9764	0.9803	0.9804

Comparing the five different models generated, model3 appears to be the most adequate fit for the data. Both model1 and model2 have substantially lower R^2 values than the other three models, indicating neither are the best fit. The final three, model1, model2 and model3, all have an R^2 value within 0.004 of one another making this statistic essentially the same between them. By analyzing the residual diagnostic plots of the last three models, model4 is the most appropriate for the given data. All three models show simliar plots comparing the residuals to fitted values but the Residuals vs Leverage plot for model5 shows eratic behavior in the Cook's Distance value, indicating this model is more effected by influential observations.

Part 5




```
## [1] 0.0007338344
```

Based on the density plot, the slope coefficients (b_1) for the 1000 simulated linear models appear to be roughly normally distributed with a mean of approximately zero. The mean of b_1 is 0.0007338344, which is approximately zero. Based on the mean value and the distribution, the slope coefficient of zero indicates that for randomly simulated values of X and Y there is no relationship, or a correlation coefficient of zero.

##Code Appendix:

```
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(ggplot2)
library(knitr)
library(captioner)
df <- read.csv("LinRegData.csv")
# Create scatterplot of Y vs numvar1
ggplot(df, aes(numvar1, Y)) +
  geom_point() +
  ggtitle("Y vs numvar1") +
  labs(x="numvar1", y="Y")
# Create linear model
model1 <- lm(Y~numvar1, data = df)

# Print model summary
summary(model1)
# New variable
df$numvar1new <- (df$numvar1)^(1/2.25)

# Create new linear model
model2 <- lm(Y ~ numvar1new, data = df)

# Print summary
summary(model2)

# Scatter plot for catvar1
ggplot(df, aes(numvar1new, Y, color = catvar1)) +
  geom_point() +
  ggtitle("numvar1new vs Y Grouped by catvar1") +
  labs(x="numvar1new", y="Y")

# Scatter plot for catvar2
ggplot(df, aes(numvar1new, Y, color = catvar2)) +
  geom_point() +
  ggtitle("numvar1new vs Y Grouped by catvar2") +
  labs(x="numvar1new", y="Y")

# Scatter plot for catvar3
ggplot(df, aes(numvar1new, Y, color = catvar3)) +
  geom_point() +
  ggtitle("numvar1new vs Y Grouped by catvar3") +
  labs(x="numvar1new", y="Y")
# Transform numvar2
df$numvar2new <- (df$numvar2)^(1/3.75)

# Create three new models
model3=lm(Y~numvar1new+catvar1+catvar2+catvar3, data=df)
model4=lm(Y~numvar1new*catvar1+numvar1new*catvar2+numvar1new*catvar3, data=df)
model5=lm(Y~numvar2new+numvar1new*catvar1+numvar1new*catvar2+numvar1new*catvar3, data=df)
# Summary
summary(model1)
# Create 4-1 plot for residuals
```

```r
par(mfrow=c(2,2), oma=c(0,0,0,0))
plot(model1)
# Summary
summary(model2)
# Create 4-1 plot for residuals
par(mfrow=c(2,2), oma=c(0,0,0,0))
plot(model2)
# Summary
summary(model3)
# Create 4-1 plot for residuals
par(mfrow=c(2,2), oma=c(0,0,0,0))
plot(model3)
# Summary
summary(model4)
# Create 4-1 plot for residuals
par(mfrow=c(2,2), oma=c(0,0,0,0))
plot(model4)
# Summary
summary(model5)
# Create 4-1 plot for residuals
par(mfrow=c(2,2), oma=c(0,0,0,0))
plot(model5)
# R-squared
rsq <- as.data.frame(cbind(0.8975, 0.9263, 0.9764, 0.9803, 0.9804))
names(rsq) <- c("Model1", "Model2", "Model3", "Model4", "Model5")
kable(rsq, caption = "R-Squared values for all models")

B=1000 ## number of simulation
b_1 =rep(NA, B)
for(i in 1:B)
{
  set.seed(i)
  X=rnorm(100,20,10)
  Y=rnorm(100,70,5)
  model <- lm(Y ~ X)
  b_1[i] = model$coefficients[2]
}
avg_b_1 <- mean(b_1)

# Plot density plots of coefficients
# b_1
ggplot(as.data.frame(b_1), aes(b_1)) +
  geom_density(fill = "blue", alpha = 0.25) +
  ggtitle("Distribution of b_1") +
  labs(x="b_1", y="Density") +
  geom_vline(xintercept = avg_b_1)

avg_b_1
```