CSCI E-106:Assignment 6

Problem 1

Use the divusa data under library(faraway).(type following commands in R Console: library(faraway);data("divusa")) (25 points, 5 points each)

```
suppressPackageStartupMessages({
  library(ggplot2)
  library(faraway)
  library(lattice)
  library(caret)
  library(ggfortify)
  library(dplyr)
  library(car)
  library(gridExtra)
  library(lmtest)
  library(MASS)
  library(GGally)
  library(olsrr)
  library(caret)
})
divusa_data <- as.data.frame(divusa)
```

a-) Fit a regression model with divorce as the response and unemployed, femlab, marriage, birth and military as predictors.

```
divusa_lm <- lm(divorce ~ unemployed + femlab + marriage + birth + military, data=divusa_data)
summary(divusa_lm)
```

```
##
## Call:
## lm(formula = divorce ~ unemployed + femlab + marriage + birth +
##     military, data = divusa_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8611 -0.8916 -0.0496  0.8650  3.8300
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.48784    3.39378   0.733   0.4659
## unemployed  -0.11125    0.05592  -1.989   0.0505 .
## femlab       0.38365    0.03059  12.543  < 2e-16 ***
## marriage     0.11867    0.02441   4.861 6.77e-06 ***
## birth       -0.12996    0.01560  -8.333 4.03e-12 ***
## military    -0.02673    0.01425  -1.876   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

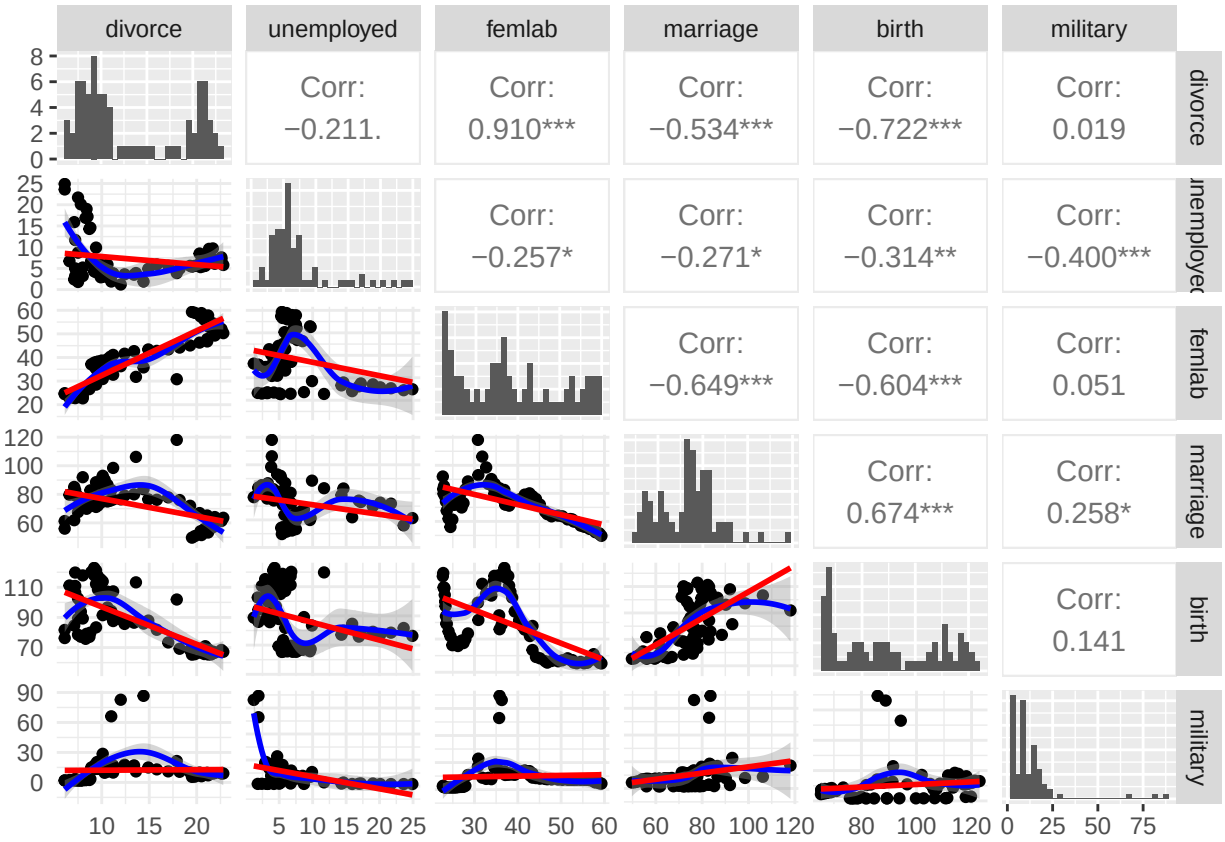
```
## Residual standard error: 1.65 on 71 degrees of freedom
## Multiple R-squared:  0.9208, Adjusted R-squared:  0.9152
## F-statistic: 165.1 on 5 and 71 DF,  p-value: < 2.2e-16
```

The residuals represent the differences between the observed and predicted divorce values. Most residuals are relatively small, ranging from -3.86 to 3.83, with the majority clustering around the lower quartiles (-0.89, 0.87), meaning that the model is reasonably accurate for many predictions. When examining the coefficients, we observe that the most influential and significant variables are femlab, birth and marriage. These have strong effects on divorce and highly significant p-values. The unemployment and military variables are marginally significant, suggesting a weak or less clear effect on divorce. The model explains a large portion of the variance, given the R^2 of 0.928, in divorce rates and is statistically significant, but some predictors have more influence than others.

b-) For the same model, compute the VIFs. Is there evidence that collinearity causes some predictors not to be significant? Explain.

First, I create a pairwise plot matrix to visualize the relationships between the variables in our dataset. The lower triangle of the matrix displays smoothed scatterplots with both the smooth lowess and the regression line, the diagonal elements display histograms (provide insight into the univariate distributions of the selected variables), and the upper triangle of the matrix displays the correlation coefficients.

Note: I suppress the code/messages as it outputs many lines of code about binwidth. However, it is all there in the .rmd



The significant correlation (***) between femlab, birth, and marriage with divorce and with each other suggests potential multicollinearity. This could inflate the standard errors for predictors like unemployed and military (which seem to be significantly collinear), causing them to appear less significant as predictors despite their relationship with the response variable.

To investigate collinearity empirically, we can use the Variance Inflation Factor (VIF), which quantifies

how much the variance of a coefficient is inflated due to collinearity. A VIF value above 5 or 10 indicates problematic multicollinearity.

```
ols_vif_tol(divusa_lm)
```

```
##    Variables Tolerance      VIF
## 1 unemployed 0.4438747 2.252888
## 2     femlab 0.2767572 3.613276
## 3   marriage 0.3490567 2.864864
## 4      birth 0.3867746 2.585485
## 5   military 0.8002586 1.249596
```

All the VIF values are below 10, indicating that multicollinearity is not a severe issue. However, `femlab` has the highest VIF at 3.61, suggesting some degree of collinearity, but within acceptable limits.

c-) Does the removal of insignificant predictors from the model reduce the collinearity? Investigate.

Given `military` and `unemployed` are insignificant predictors, we build a new model, `divusa_rmv_insig_lm`, without these independent variables and examine the updates to the model performance.

However, first, let us use the General F-test to confirm whether we can justify dropping these variables.

H_o: $\beta_{military} = \beta_{unemployed} = 0$
H_a: $\beta_{military} \neq 0$ or $\beta_{unemployed} \neq 0$, one of them are significant so we cannot drop them

```
divusa_rmv_insig_lm <- lm(divorce ~ femlab + marriage + birth, data=divusa_data)

anova(divusa_rmv_insig_lm,divusa_lm)
```

```
## Analysis of Variance Table
##
## Model 1: divorce ~ femlab + marriage + birth
## Model 2: divorce ~ unemployed + femlab + marriage + birth + military
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     73 209.84
## 2     71 193.40  2    16.444 3.0185 0.05519 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The `anova()` function performs a hypothesis test comparing the two models, and in this case, the p-value is 0.05 which is lower than or equal to our chosen significance $\alpha = 0.05$, so we cannot reject the null hypothesis and we conclude that these variables, military and unemployed, are not significant and we can drop them from our model.

```
summary(divusa_rmv_insig_lm)
```

```
##
## Call:
## lm(formula = divorce ~ femlab + marriage + birth, data = divusa_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6923 -1.1934 -0.0534  1.2265  3.6701
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.54545    2.21247  -0.699    0.487
```

```
## femlab       0.41337    0.02275  18.174  < 2e-16 ***
## marriage     0.12609    0.02199   5.735 2.07e-07 ***
## birth       -0.11627    0.01412  -8.235 5.10e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.695 on 73 degrees of freedom
## Multiple R-squared:  0.9141, Adjusted R-squared:  0.9106
## F-statistic: 258.9 on 3 and 73 DF,  p-value: < 2.2e-16
```

```
vif(divusa_rmv_insig_lm)
```

```
##   femlab marriage    birth
## 1.893390 2.201891 2.008469
```

Table 1.1: Examining Model VIF Values: Pre and Post Removal of Insignificant Predictors

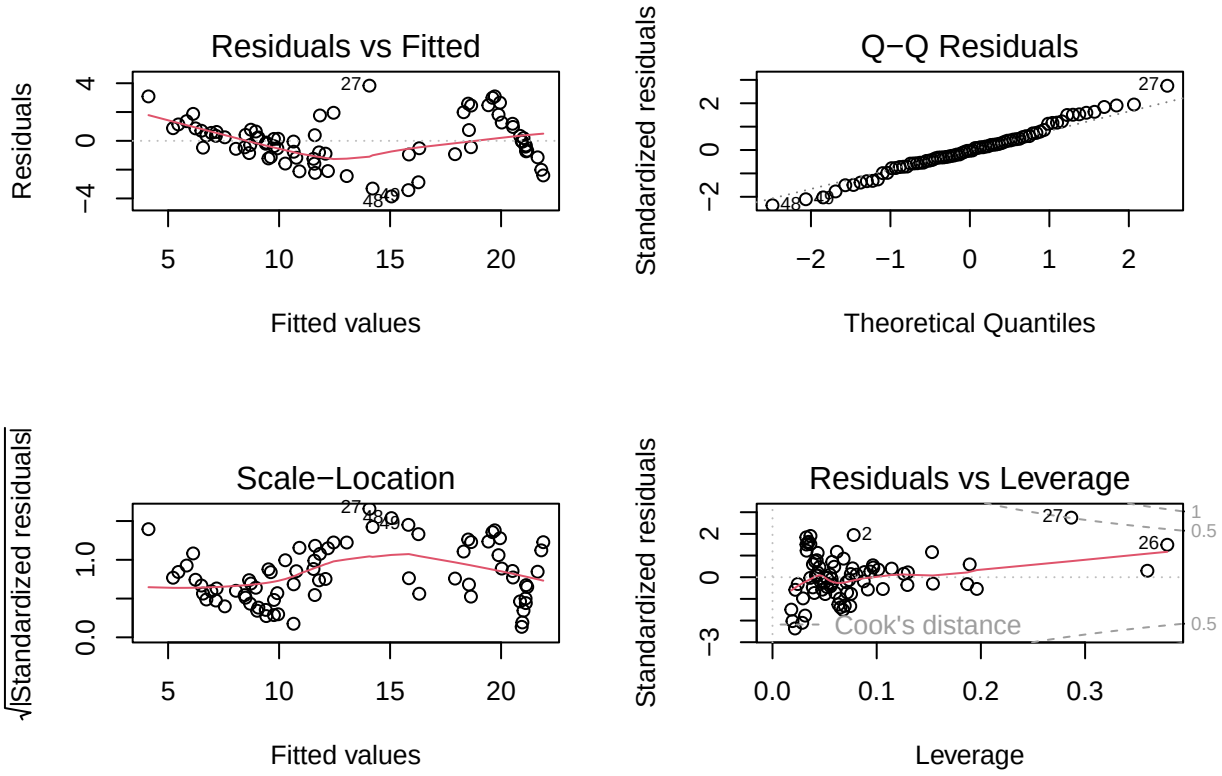
Variable	Full Model	Only Significant Model
femlab VIF	3.613276	1.893390
marriage VIF	2.864864	2.201891
birth VIF	2.585485	2.008469
R^2	0.9208	0.9141
Adj. R^2	0.9152	0.9106

After removing the insignificant predictors from the model and recalculating VIF, we observe the VIF values decreased for femlab, marriage and birth, indicating a reduction in collinearity among the remaining predictors; however, the R^2 also decreased, which is common when predictor variables are removed. Reduced VIFs indicate that the model is now more robust, with significant predictors better representing the underlying relationships without the confounding effects of the previously included insignificant predictors.

d-) Visually using the appropriate graphs, comment on the regression model assumptions.

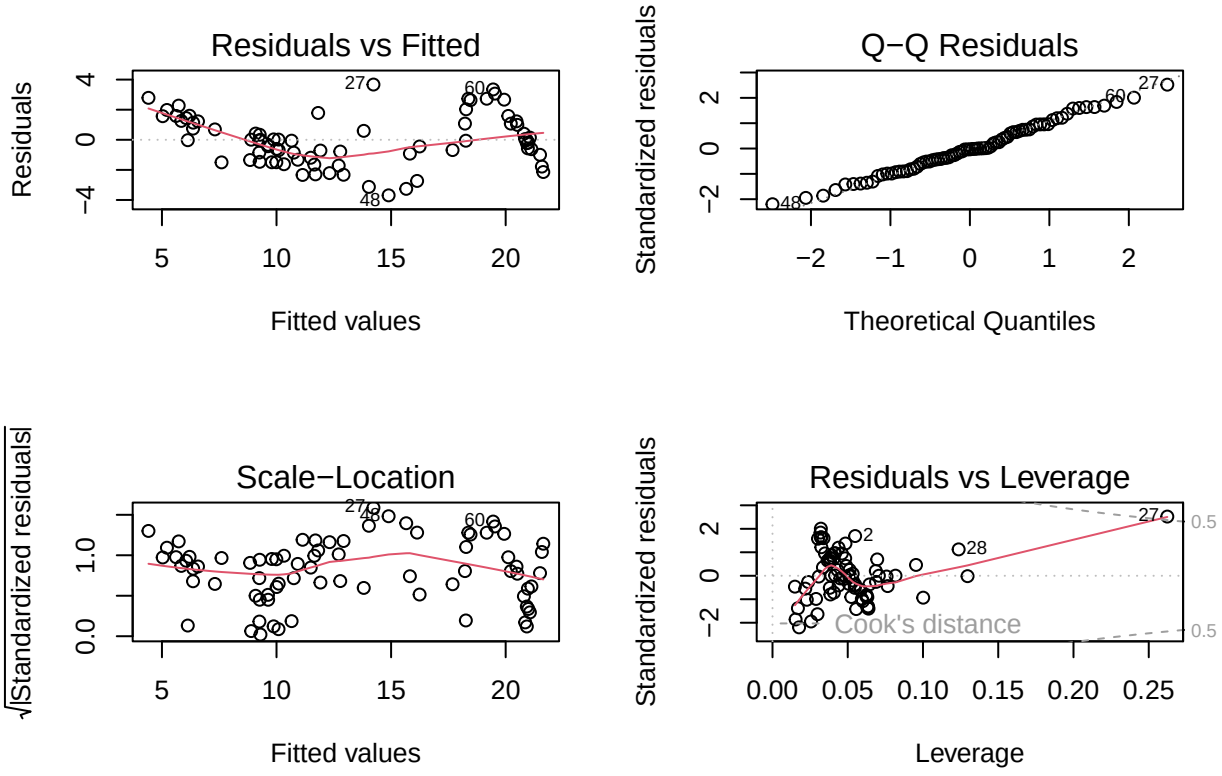
Full Model

```
par(mfrow=c(2,2))
plot(divusa_lm)
```



Insignificant Predictors Removed Model

```
par(mfrow=c(2,2))
plot(divusa_rmv_insig_lm)
```



Between the two models, the diagnostic plots (Residuals vs Fitted, Scale Location, and Residuals vs Leverage)

show similar patterns overall. However, in the reduced model, there is a more pronounced peak at a leverage of 0.05, compared to a broader peak in the full model. Additionally, data point 27 stands out as a potential outlier with a Cook's Distance of 0.5 in the reduced model, whereas it only marginally crosses the Cook's Distance threshold in the full model. Notably, the Q-Q plot in the reduced model shows residuals clustering more closely along the expected diagonal, suggesting an improvement in the normality of residuals after removing insignificant predictors.

e-) Conduct the Breusch Pagan Test test.

First, let us state the hypotheses:
H_o: $\sigma_i^2 = \sigma^2$ for all i: There is homoscedasticity (constant variance) in the residuals. In other words, the variance of the errors does not depend on the independent variables in the regression model.
H_a: $\sigma_i^2 \neq \sigma^2$: There is heteroscedasticity (non-constant variance) in the residuals. This means that the variance of the errors is related to the independent variables in the regression model.

Next, let us state the decision rule:
1. If p-value $\geq \alpha = 0.05$ or $BP < \chi^2_{1-\alpha,df}$, then we fail to reject H_0, concluding that the residuals have constant variance.
2. If p-value $< \alpha = 0.05$ or $BP > \chi^2_{1-\alpha,df}$, then we reject H_0, concluding that the residuals do not have constant variance.

```
bp_test_org <- bptest(divusa_lm, studentize = FALSE)

print(bp_test_org)
```

```
##
##  Breusch-Pagan test
##
## data:  divusa_lm
## BP = 16.581, df = 5, p-value = 0.005368
```

```
df_org <- length(coefficients(divusa_lm)) - 1
alpha <- 0.01
critical_value_org <- qchisq(1 - alpha, df = df_org)

cat("Critical value Breusch-Pagan test of Full Model at 99% confidence level:", critical_value_org, "\n
```

```
## Critical value Breusch-Pagan test of Full Model at 99% confidence level: 15.08627
```

```
bp_test_insig <- bptest(divusa_rmv_insig_lm, studentize = FALSE)

print(bp_test_insig)
```

```
##
##  Breusch-Pagan test
##
## data:  divusa_rmv_insig_lm
## BP = 11.037, df = 3, p-value = 0.01153
```

```
df_insig <- length(coefficients(divusa_rmv_insig_lm)) - 1
alpha <- 0.01
critical_value_insig <- qchisq(1 - alpha, df = df_insig)

cat("Critical value Breusch-Pagan test of Only Significant P.V. Model at 99% confidence level:", critica
```

```
## Critical value Breusch-Pagan test of Only Significant P.V. Model at 99% confidence level: 11.34487
```

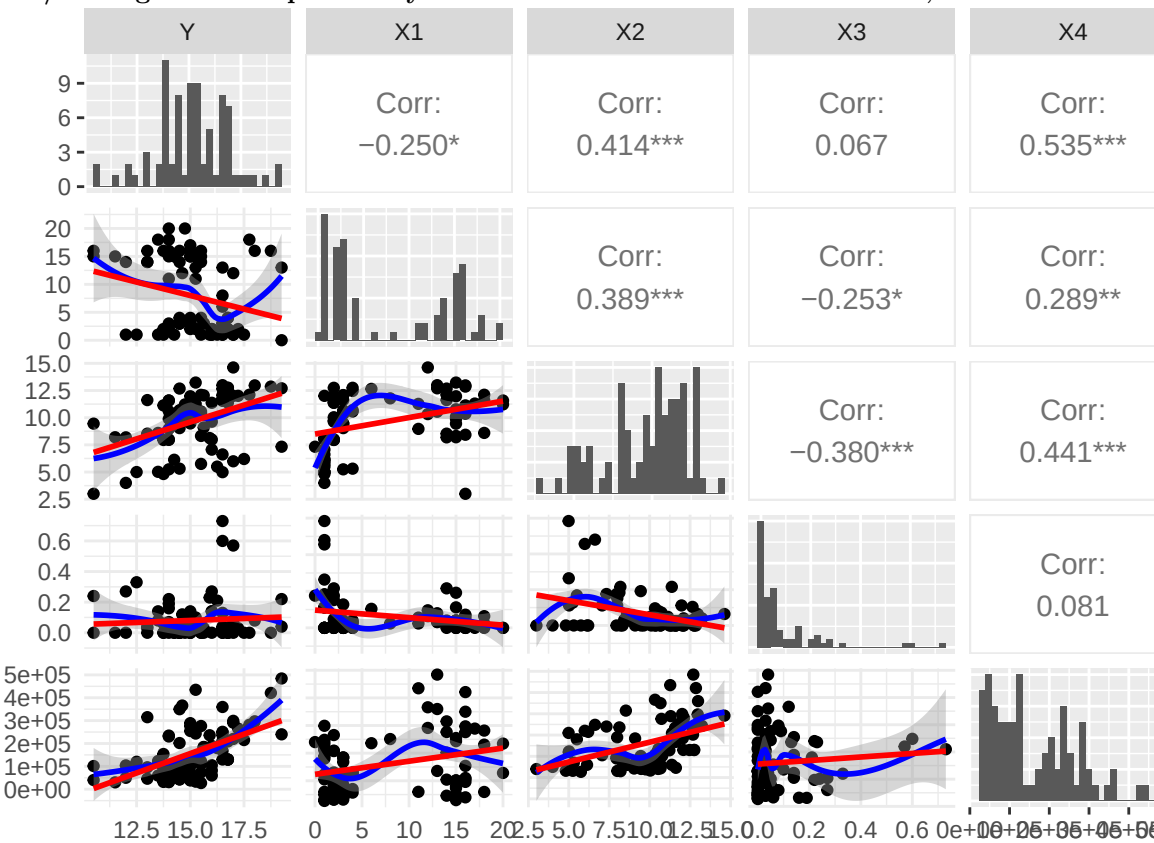
By removing the insignificant variables and comparing the results of the Breusch-Pagan tests at a significance level of $\alpha = 0.01$, we observe that the model without the insignificant predictors maintains constant variance ($BP = 11.037 < \chi^2_{\alpha-1,df} \approx 11.34487$ and p-value $= 0.012 \geq \alpha = 0.01$). In contrast, the full model violates this assumption ($BP = 16.581 > \chi^2_{\alpha-1,df} \approx 15.08627$ and p-value $= 0.005 < \alpha = 0.01$). Therefore, we conclude that the constancy of variance holds only for the only-significant-predictor-variables model, and not in the full model; however, increasing the value of α would ensure both the models support homoscedasticity.

Problem 2

Commercial properties data set. A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data are taken from 81 suburban commercial properties include the age (X1), operating expenses and taxes (X2), vacancy rates (X3), total square footage (X4), and rental rates (Y). (45 points)

```
com_prop_data <- read.csv('/Users/shreyabajpai/CSCI E-106 - Data Modeling/CSCI E-106 Assignment 6/Comme
Y <- com_prop_data$Y
X1 <- com_prop_data$X1
X2 <- com_prop_data$X2
X3 <- com_prop_data$X3
X4 <- com_prop_data$X4
```

a-) Obtain the scatter plot matrix and the correlation matrix. Interpret these and state your principal findings.(5 points)
Note: I suppress the code/messages as it outputs many lines of code about binwidth. However, it is all there in the .rmd



The histograms on the scatterplot matrix diagonal show that Y and X2 have concentrated value ranges, while X1, X3, and X4 are more broadly distributed. Notably, Y has strong positive correlations with X4 (0.535)

and X2 (0.414), while X2 correlates moderately with X3 (-0.380) and X4 (0.441). X3 shows weak correlations overall, suggesting a minimal linear relationship.

The scatterplots reveal complexities in the data, with non-linear patterns that the regression lines do not fully capture. For instance, Y and X1's relationship is curvilinear, and Y and X4 display a hyperbolic trend. These findings suggest that a linear regression may not sufficiently model the variable interactions, especially given the non-linear behaviors highlighted by the lowess smooths.

b-) Fit regression model for four predictor variables to the data. State the estimated regression function.(5 points)

```
com_prop_lm <- lm(Y ~ X1 + X2 + X3 + X4, data=com_prop_data)
summary(com_prop_lm)
```

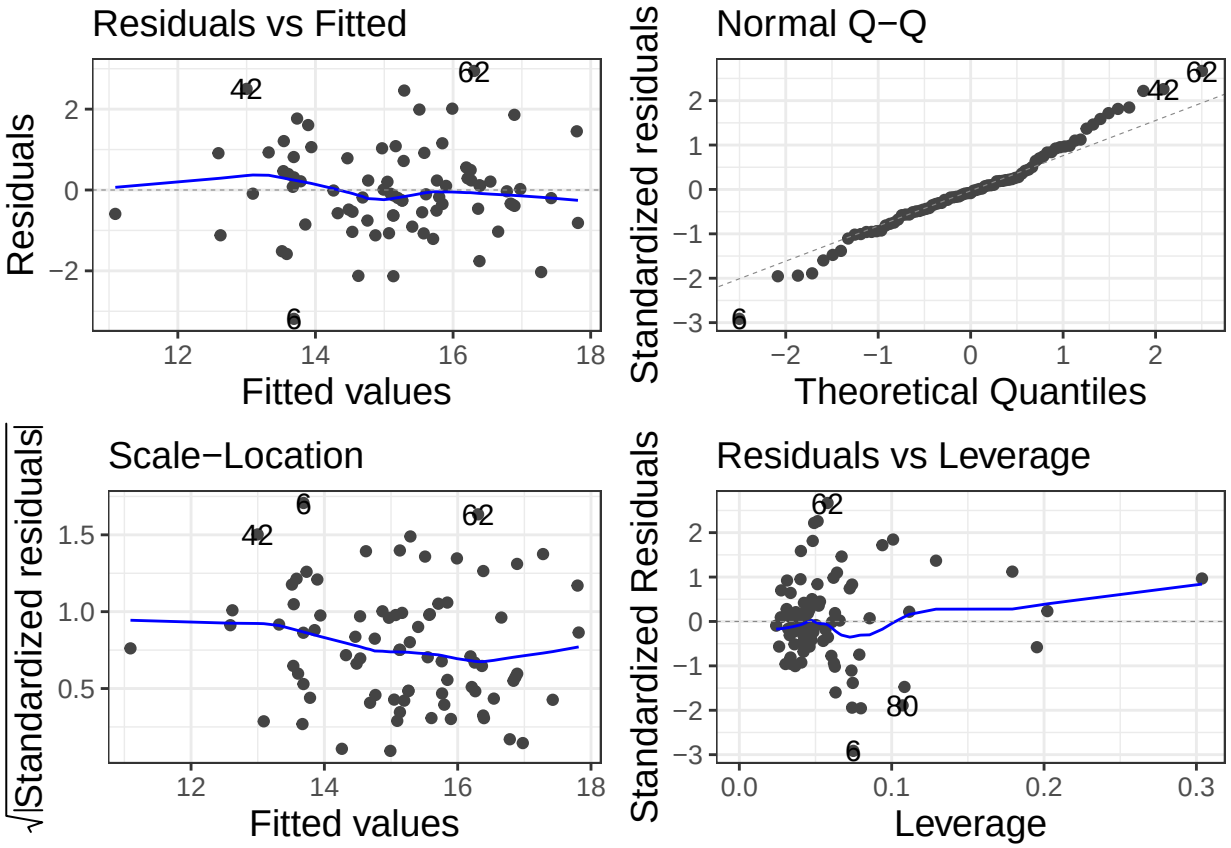
```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = com_prop_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1872 -0.5911 -0.0910  0.5579  2.9441
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.220e+01  5.780e-01  21.110  < 2e-16 ***
## X1          -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## X2           2.820e-01  6.317e-02   4.464 2.75e-05 ***
## X3           6.193e-01  1.087e+00   0.570     0.57
## X4           7.924e-06  1.385e-06   5.722 1.98e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.137 on 76 degrees of freedom
## Multiple R-squared:  0.5847, Adjusted R-squared:  0.5629
## F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

The estimated regression function is: $Y = 12.20 - 0.142 \times X1 + 0.282 \times X2 + 0.6193 \times X3 + 0.000007924 \times X4$.

c-) Visually using the appropriate graphs, comment on the regression model assumptions. (5 points)

Linear regression has the following assumptions:
(1) The residuals (errors) are identically and independently distributed.
(2) The residuals (errors) are normally distributed, with mean 0 and equal variance (homoscedasticity).

```
autoplot(com_prop_lm) + theme_bw() + theme(
    text = element_text(size = 12),
    axis.title = element_text(size = 14),
    legend.position = "bottom"
  )
```



The diagnostic plots above show residuals in four different ways:

1. **Residuals vs Fitted** This plot checks whether the assumption of a linear relationship holds in the model. Ideally, residuals should be scattered randomly around zero, showing a balanced fit. In this case, a "wave" pattern suggests that the relationship between the predictors and the response variable is not strictly linear, especially with the clustering of residuals in the low trough of the "wave" and the few highlighted points that may be pulling the "wave" to its peak (42, 62) and troughs (6).

2. **Normal QQ** This plot evaluates whether the residuals follow a normal distribution, with the ideal scenario being that all points align closely with the diagonal line. In this case, the residuals fall primarily on the dotted gray line with heavy tails and few extreme values highlighted which could indicate that these residuals deviate significantly from what's expected under normal distribution, suggesting the presence of outliers. This non-normality can skew the model's predictions and lead to less reliable parameter estimates, as it violates a core assumption of linear regression.

3. **Scale-Location** This plot helps assess whether the residuals have constant variance, a key assumption in linear regression. Ideally, the residuals should be scattered randomly around a horizontal line, which is what we see for the most part, but the slightly downward "wave" curve could violate the assumption of constant error variance, which could lead to inefficiencies in the model's predictions and affect its reliability for unknown data.

4. **Residuals vs Leverage** The plot is crucial for identifying influential data points that could disproportionately impact the model. Ideally, residuals should be randomly scattered which is what we observe here with slight curve that tapers upwards as leverage increase. This pattern suggests that certain data points in the mean of the data set could be exerting an influence on the model's fit. These high-leverage points (i.e., 6, 62, 80) could be pulling the model in their direction, making it less accurate overall.

At first glance, there does not seem to be too much deviation between the expected nature of linear regression and the residuals plotted above. The possibility of potential outliers (as evidence in the heavy tails of the QQ-plot) exists and may be the reason we see the slight deviations from normality across the graphs.

d-) Divide the 81 cases into two groups. placing the 40 cases with the smallest fitted values into

group 1 and the remaining cases into group 2. Conduct the Brown-Forsythe test.(10 points)

The hypotheses are:
H_o: Error variance is constant.
H_a: Error variance is not constant.
The decision rule is:
If $t^*_{BF} \leq t(1-\alpha/2; n-2)$ or $p-value \geq \alpha$, then we fail to reject H_0, concluding the error variances are constant.
2. $t^*_{BF} > t(1-\alpha/2; n-2)$ or $p-value < \alpha$, then we fail to reject H_0, concluding the error variances are constant.

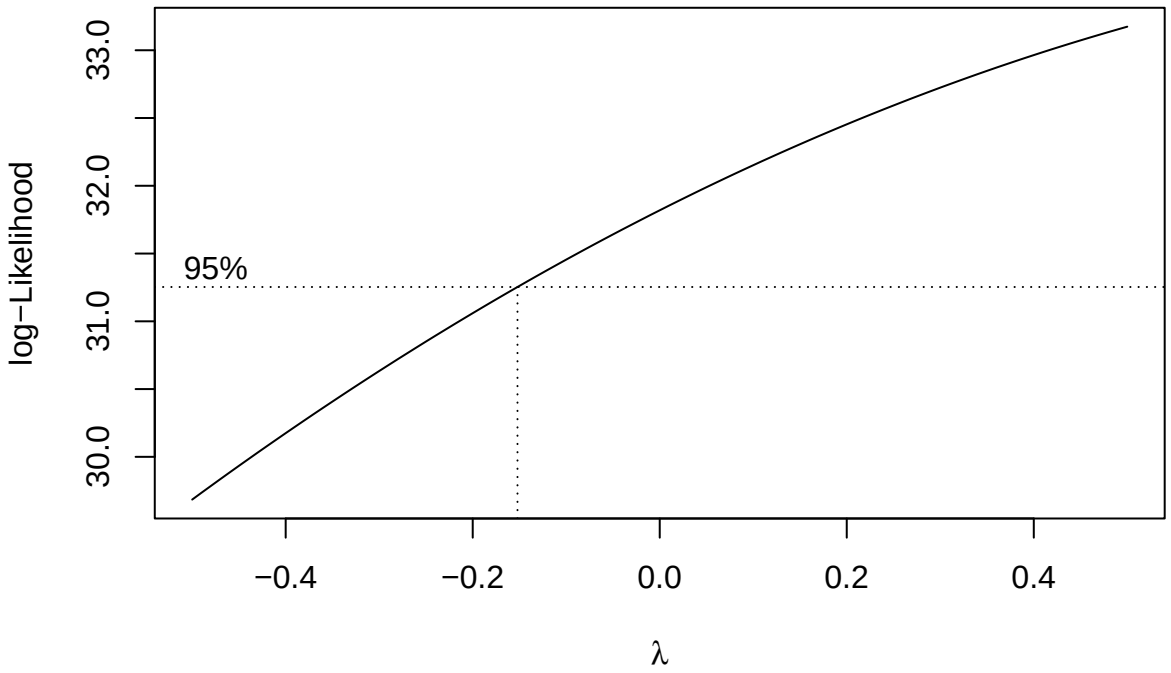
```
fitted_values <- fitted(com_prop_lm)

sorted_indices <- order(fitted_values)

group_1_indices <- sorted_indices[1:40]
group_2_indices <- sorted_indices[41:81]

com_prop_data$group <- NA

com_prop_data$group[group_1_indices] <- 1
com_prop_data$group[group_2_indices] <- 2

com_prop_data$group <- as.factor(com_prop_data$group)

residuals_model <- residuals(com_prop_lm)

leveneTest(residuals_model ~ com_prop_data$group, center = median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.3048 0.5824
##       79
```

The Brown-Forsythe test assesses whether different groups exhibit equal variances, a condition known as homogeneity of variances, while being robust to non-normality in the data. In our analysis, the F statistic is 0.3048, with a p-value of 0.5824—significantly above the 0.05 threshold, so we are led to accept H_0 concluding that the error variances are constant, thereby preserving homoscedasticity. In other words, the variability of the residuals does not change across levels of the fitted values.

e-) Perform a Box-Cox transformation analysis and decide if Y needs to be transformed.(5 points)

```
boxcox_result <- boxcox(com_prop_lm, lambda = seq(-0.5,0.5,0.01))
```




```
optimal_lambda <- boxcox_result$x[which.max(boxcox_result$y)]
print(paste("Optimal Lambda:", optimal_lambda))
```

```
## [1] "Optimal Lambda: 0.5"
```

The graph (*log-likelihood function for a Box-Cox transformation*) shows that the optimal Box-Cox transformation parameter lies near the peak of the curve, which appears to be around $\lambda = 0.5$. This suggests that a power transformation close to the optimal λ will best stabilize the variance and improve model fit. The optimal λ is 0.5, which suggests either a square root or a power transformation might be ideal. The Box-Cox transformation formula for the power transformation is $\lambda \neq 0$ is: $y' = \frac{y^{\lambda}-1}{\lambda}$, since $\lambda = 0.5$, $Y_{transformed} = \frac{Y^{0.5}-1}{0.5}$.

```
com_prop_data$Y_transformed <- (com_prop_data$Y^optimal_lambda - 1) / optimal_lambda
com_prop_data$Y_transformed_2 <- sqrt(com_prop_data$Y)

com_prop_trnsfrm_lm <-lm(Y_transformed ~ X1 + X2 + X3 + X4, data = com_prop_data)
com_prop_trnsfrm_lm_2 <-lm(Y_transformed_2 ~ X1 + X2 + X3 + X4, data = com_prop_data)

summary(com_prop_trnsfrm_lm)
```

```
##
## Call:
## lm(formula = Y_transformed ~ X1 + X2 + X3 + X4, data = com_prop_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90919 -0.15848 -0.02097  0.13126  0.70851
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.993e+00  1.501e-01  33.256  < 2e-16 ***
## X1          -3.752e-02  5.545e-03  -6.766 2.41e-09 ***
## X2           7.594e-02  1.641e-02   4.627 1.50e-05 ***
## X3           1.605e-01  2.823e-01   0.568    0.572
```

```
## X4           2.005e-06  3.597e-07   5.574 3.63e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2953 on 76 degrees of freedom
## Multiple R-squared:  0.5863, Adjusted R-squared:  0.5645
## F-statistic: 26.92 on 4 and 76 DF,  p-value: 6.339e-14
```

```
summary(com_prop_trnsfrm_lm_2)
```

```
##
## Call:
## lm(formula = Y_transformed_2 ~ X1 + X2 + X3 + X4, data = com_prop_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45460 -0.07924 -0.01049  0.06563  0.35426
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.497e+00  7.507e-02  46.577  < 2e-16 ***
## X1          -1.876e-02  2.772e-03  -6.766 2.41e-09 ***
## X2           3.797e-02  8.206e-03   4.627 1.50e-05 ***
## X3           8.023e-02  1.412e-01   0.568    0.572
## X4           1.003e-06  1.799e-07   5.574 3.63e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1477 on 76 degrees of freedom
## Multiple R-squared:  0.5863, Adjusted R-squared:  0.5645
## F-statistic: 26.92 on 4 and 76 DF,  p-value: 6.339e-14
```

Between the two possible transformations with the lambda we obtained, we observe that both the model's R^2 are equivalent; however, the residuals are tighter, and the RSE, coefficient estimates and standard errors are lower in the *square root transformation model* when compared to the power transformation model. We aim to stabilize variance and improve the normality of the response variable. We now examine the performance improvements we may have achieved as a result of the **square root** transformation on the dataset.

```
# Non-Transformed Model
Predicted <- predict(com_prop_lm, com_prop_data)

ModelData <- data.frame(obs = com_prop_data$Y, pred = Predicted)

ModelFull <- defaultSummary(ModelData)

# Transformed Model
PredictedTrnsfrm <- predict(com_prop_trnsfrm_lm_2, com_prop_data)

# Inverse transformation to bring predictions back to original scale
PredictedTrnsfrm_OriginalScale <- (PredictedTrnsfrm)^2

ModelTrnsfrmData <- data.frame(obs = com_prop_data$Y, pred = PredictedTrnsfrm_OriginalScale)

ModelTrnsfrm <- defaultSummary(ModelTrnsfrmData)

trns = rbind(ModelFull, ModelTrnsfrm)
```

```
print(trns)
```

```
##                   RMSE  Rsquared       MAE
## ModelFull    1.101237 0.5847496 0.8268344
## ModelTrnsfrm 1.102375 0.5840572 0.8265651
```

The transformed model does not show significant improvements over the non-transformed model. The Root Mean Squared Error (RMSE), which reflects the average magnitude of prediction error, remains nearly identical between the two models (1.101 for the non-transformed and 1.102 for the transformed). Similarly, the Mean Absolute Error (MAE), which captures the average absolute difference between predicted and actual values, shows minimal change (0.8268 for the non-transformed model versus 0.8266 for the transformed). This suggests that the transformed model does not offer closer predictions to the true values compared to the non-transformed model.

The R^2 value, which indicates the proportion of variance explained by the model, decreased slightly from 0.5847 in the non-transformed model to 0.5841 in the transformed model. This small drop further indicates that the transformed model does not provide better explanatory power for the variation in the response variable.

The Box-Cox transformation, despite being a useful tool for stabilizing variance and improving model accuracy in many cases, has not enhanced the predictive performance in this instance. The minimal changes in RMSE, MAE, and R^2 suggest that the original model was already fitting the data well, and the transformation did not offer any substantial benefits. This outcome highlights that while transformations can be effective, they are not always necessary if the data is already well-suited for linear modeling.

f-) Fit a second order model ($Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_1^2 + b_6X_2^2 + b_7X_3^2 + b_8X_4^2 + b_9X_1X_2 + b_{10}X_1X_3 + b_{11}X_1X_4 + b_{12}X_2X_3 + b_{13}X_2X_4 + b_{14}X_3X_4 + b_{15}X_1X_2X_3 + b_{16}X_1X_2X_4 + b_{17}X_2X_3X_4 + b_{18}X_1X_2X_3X_4 + b_{19}X_1X_3X_4$) Drop the insignificant variables from the model one at a time, refit the model and repeat this until all the variables you have in the model are significant. Compare this again the model in part b) (15 points)

```
full_mod <- lm(Y ~ X1 + X2 + X3 + X4 +
                 I(X1^2) + I(X2^2) + I(X3^2) + I(X4^2) +
                 X1:X2 + X1:X3 + X1:X4 + X2:X3 + X2:X4 + X3:X4 +
                 X1:X2:X3 + X1:X2:X4 + X2:X3:X4 + X1:X2:X3:X4 + X1:X3:X4, data=com_prop_data)
summary(full_mod)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + I(X1^2) + I(X2^2) + I(X3^2) +
##     I(X4^2) + X1:X2 + X1:X3 + X1:X4 + X2:X3 + X2:X4 + X3:X4 +
##     X1:X2:X3 + X1:X2:X4 + X2:X3:X4 + X1:X2:X3:X4 + X1:X3:X4,
##     data = com_prop_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01237 -0.60959 -0.02205  0.50532  1.96870
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.838e+00  1.933e+00   5.089  3.7e-06 ***
## X1          -4.979e-02  1.718e-01  -0.290   0.7729
## X2           8.365e-01  4.257e-01   1.965   0.0539 .
## X3          -3.720e+01  1.620e+01  -2.296   0.0251 *
```

```
## X4           2.003e-05  1.318e-05   1.519   0.1338
## I(X1^2)      2.502e-03  6.782e-03   0.369   0.7134
## I(X2^2)     -3.280e-02  2.658e-02  -1.234   0.2220
## I(X3^2)     -9.839e+00  1.119e+01  -0.879   0.3827
## I(X4^2)     -6.644e-12  1.388e-11  -0.479   0.6338
## X1:X2       -6.437e-03  1.521e-02  -0.423   0.6736
## X1:X3        7.327e+00  3.944e+00   1.858   0.0681 .
## X1:X4       -2.753e-06  1.289e-06  -2.136   0.0367 *
## X2:X3        4.353e+00  2.052e+00   2.121   0.0380 *
## X2:X4       -8.255e-07  1.391e-06  -0.593   0.5551
## X3:X4        2.695e-04  1.391e-04   1.938   0.0573 .
## X1:X2:X3    -8.431e-01  4.305e-01  -1.958   0.0547 .
## X1:X2:X4     2.283e-07  1.153e-07   1.980   0.0522 .
## X2:X3:X4    -3.062e-05  1.619e-05  -1.892   0.0632 .
## X1:X3:X4    -4.626e-05  2.421e-05  -1.911   0.0607 .
## X1:X2:X3:X4  4.881e-06  2.375e-06   2.055   0.0441 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9804 on 61 degrees of freedom
## Multiple R-squared:  0.7521, Adjusted R-squared:  0.6749
## F-statistic: 9.743 on 19 and 61 DF,  p-value: 3.775e-12
```

Given that only a handful of independent variables are significant (X3, X1:X4, X2:X3, X1:X2:X3:X4), we can build a function to scan the p-value of each of the independent variables and drop the ones which are non-significant ($\alpha \geq 0.05$).

```
drop_insignificant_vars <- function(model, data) {
  while (TRUE) {
    p_values <- summary(model)$coefficients[, 4]
    max_p_value <- max(p_values[-1])  # Ignore intercept (p-values[-1])
    if (max_p_value <= 0.05) break
    var_to_drop <- names(p_values)[which.max(p_values)]
    formula <- as.formula(update(model, paste(". ~ . -", var_to_drop)))
    model <- lm(formula, data = data)
    print(summary(model))
  }
  return(model)
}

redu_mod <- drop_insignificant_vars(full_mod, com_prop_data)
```

```
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.06806 -0.59713 -0.00695  0.50039  1.91829
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.534e+00  1.611e+00   5.919 1.52e-07 ***
## X2           8.795e-01  3.960e-01   2.221  0.03001 *
## X3          -3.675e+01  1.600e+01  -2.296  0.02507 *
```

```
## X4           2.149e-05  1.209e-05   1.777  0.08041 .
## I(X1^2)      1.331e-03  5.407e-03   0.246  0.80637
## I(X2^2)     -3.473e-02  2.553e-02  -1.360  0.17868
## I(X3^2)     -9.700e+00  1.110e+01  -0.874  0.38548
## I(X4^2)     -8.060e-12  1.289e-11  -0.625  0.53414
## X2:X1       -9.252e-03  1.162e-02  -0.796  0.42905
## X3:X1        7.192e+00  3.888e+00   1.850  0.06910 .
## X4:X1       -2.959e-06  1.067e-06  -2.773  0.00732 **
## X2:X3        4.315e+00  2.033e+00   2.123  0.03779 *
## X2:X4       -9.179e-07  1.344e-06  -0.683  0.49721
## X3:X4        2.662e-04  1.376e-04   1.935  0.05761 .
## X2:X3:X1    -8.324e-01  4.257e-01  -1.955  0.05506 .
## X2:X4:X1     2.483e-07  9.167e-08   2.709  0.00872 **
## X2:X3:X4    -3.035e-05  1.604e-05  -1.892  0.06311 .
## X3:X4:X1    -4.519e-05  2.374e-05  -1.903  0.06164 .
## X2:X3:X4:X1  4.786e-06  2.335e-06   2.050  0.04461 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9731 on 62 degrees of freedom
## Multiple R-squared:  0.7518, Adjusted R-squared:  0.6797
## F-statistic: 10.43 on 18 and 62 DF,  p-value: 1.184e-12
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05044 -0.60860 -0.02596  0.50998  1.91635
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.700e+00  1.452e+00   6.681 7.17e-09 ***
## X2           8.505e-01  3.752e-01   2.267  0.02684 *
## X3          -3.827e+01  1.465e+01  -2.613  0.01121 *
## X4           2.068e-05  1.154e-05   1.791  0.07806 .
## I(X2^2)     -3.443e-02  2.531e-02  -1.360  0.17866
## I(X3^2)     -1.036e+01  1.069e+01  -0.969  0.33643
## I(X4^2)     -8.810e-12  1.243e-11  -0.709  0.48112
## X2:X1       -6.657e-03  4.861e-03  -1.370  0.17569
## X3:X1        7.479e+00  3.681e+00   2.032  0.04641 *
## X4:X1       -2.836e-06  9.349e-07  -3.033  0.00351 **
## X2:X3        4.510e+00  1.858e+00   2.427  0.01809 *
## X2:X4       -8.058e-07  1.255e-06  -0.642  0.52320
## X3:X4        2.782e-04  1.276e-04   2.181  0.03294 *
## X2:X3:X1    -8.641e-01  4.028e-01  -2.145  0.03579 *
## X2:X4:X1     2.365e-07  7.746e-08   3.053  0.00332 **
## X2:X3:X4    -3.171e-05  1.494e-05  -2.123  0.03773 *
## X3:X4:X1    -4.672e-05  2.274e-05  -2.055  0.04405 *
## X2:X3:X4:X1  4.940e-06  2.232e-06   2.213  0.03052 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.9658 on 63 degrees of freedom
## Multiple R-squared:  0.7516, Adjusted R-squared:  0.6845
## F-statistic: 11.21 on 17 and 63 DF,  p-value: 3.546e-13
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.12015 -0.57464 -0.05508  0.53453  1.84410
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.711e+00  1.445e+00   6.720 5.75e-09 ***
## X2           9.225e-01  3.564e-01   2.588  0.01193 *
## X3          -4.060e+01  1.412e+01  -2.875  0.00549 **
## X4           1.394e-05  4.781e-06   2.916  0.00489 **
## I(X2^2)     -4.283e-02  2.157e-02  -1.986  0.05133 .
## I(X3^2)     -1.344e+01  9.509e+00  -1.413  0.16236
## I(X4^2)     -1.256e-11  1.093e-11  -1.149  0.25484
## X2:X1       -6.486e-03  4.831e-03  -1.342  0.18420
## X3:X1        7.474e+00  3.664e+00   2.040  0.04550 *
## X4:X1       -2.600e-06  8.558e-07  -3.038  0.00344 **
## X2:X3        4.881e+00  1.757e+00   2.778  0.00718 **
## X3:X4        3.164e-04  1.124e-04   2.816  0.00646 **
## X2:X3:X1    -8.724e-01  4.007e-01  -2.177  0.03316 *
## X2:X4:X1     2.131e-07  6.806e-08   3.131  0.00263 **
## X2:X3:X4    -3.598e-05  1.332e-05  -2.701  0.00884 **
## X3:X4:X1    -4.752e-05  2.260e-05  -2.103  0.03944 *
## X2:X3:X4:X1  5.074e-06  2.212e-06   2.293  0.02513 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9614 on 64 degrees of freedom
## Multiple R-squared:  0.7499, Adjusted R-squared:  0.6874
## F-statistic:    12 on 16 and 64 DF,  p-value: 1.206e-13
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.27936 -0.59779 -0.02366  0.56374  1.95736
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.006e+01  1.416e+00   7.104 1.14e-09 ***
## X2           8.540e-01  3.523e-01   2.424  0.01813 *
## X3          -3.797e+01  1.397e+01  -2.718  0.00842 **
## X4           9.780e-06  3.131e-06   3.123  0.00267 **
## I(X2^2)     -3.756e-02  2.112e-02  -1.778  0.08012 .
## I(X3^2)     -1.535e+01  9.385e+00  -1.636  0.10676
```

```
## X2:X1       -5.267e-03  4.725e-03  -1.115  0.26907
## X3:X1        6.279e+00  3.522e+00   1.783  0.07930 .
## X4:X1       -2.486e-06  8.521e-07  -2.917  0.00485 **
## X2:X3        4.617e+00  1.747e+00   2.643  0.01028 *
## X3:X4        3.201e-04  1.126e-04   2.843  0.00596 **
## X2:X3:X1    -7.321e-01  3.826e-01  -1.914  0.06009 .
## X2:X4:X1     1.937e-07  6.611e-08   2.931  0.00466 **
## X2:X3:X4    -3.619e-05  1.335e-05  -2.710  0.00858 **
## X3:X4:X1    -4.427e-05  2.248e-05  -1.970  0.05316 .
## X2:X3:X4:X1  4.674e-06  2.190e-06   2.134  0.03662 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9638 on 65 degrees of freedom
## Multiple R-squared:  0.7448, Adjusted R-squared:  0.6859
## F-statistic: 12.65 on 15 and 65 DF,  p-value: 6.058e-14
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54424 -0.51198 -0.01418  0.39313  1.93535
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.010e+01  1.418e+00   7.121 9.85e-10 ***
## X2           8.149e-01  3.512e-01   2.321 0.023411 *
## X3          -3.936e+01  1.394e+01  -2.823 0.006278 **
## X4           1.210e-05  2.339e-06   5.175 2.32e-06 ***
## I(X2^2)     -3.802e-02  2.116e-02  -1.797 0.076976 .
## I(X3^2)     -1.372e+01  9.288e+00  -1.478 0.144294
## X3:X1        6.480e+00  3.524e+00   1.839 0.070429 .
## X4:X1       -2.928e-06  7.556e-07  -3.874 0.000248 ***
## X2:X3        4.955e+00  1.723e+00   2.876 0.005425 **
## X3:X4        3.106e-04  1.125e-04   2.761 0.007445 **
## X2:X3:X1    -7.948e-01  3.791e-01  -2.096 0.039883 *
## X2:X4:X1     2.094e-07  6.472e-08   3.235 0.001901 **
## X2:X3:X4    -3.648e-05  1.337e-05  -2.727 0.008168 **
## X3:X4:X1    -4.073e-05  2.229e-05  -1.827 0.072202 .
## X2:X3:X4:X1  4.547e-06  2.191e-06   2.075 0.041887 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9655 on 66 degrees of freedom
## Multiple R-squared:  0.7399, Adjusted R-squared:  0.6847
## F-statistic: 13.41 on 14 and 66 DF,  p-value: 2.857e-14
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -2.64598 -0.63088  0.00175  0.63967  1.95026
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.743e+00  1.410e+00   6.911 2.19e-09 ***
## X2           9.531e-01  3.414e-01   2.791 0.006830 **
## X3          -3.105e+01  1.287e+01  -2.413 0.018575 *
## X4           1.285e-05  2.304e-06   5.578 4.72e-07 ***
## I(X2^2)     -4.810e-02  2.020e-02  -2.381 0.020129 *
## X3:X1        4.869e+00  3.380e+00   1.440 0.154445
## X4:X1       -3.035e-06  7.588e-07  -3.999 0.000161 ***
## X2:X3        3.707e+00  1.515e+00   2.447 0.017051 *
## X3:X4        1.835e-04  7.313e-05   2.510 0.014513 *
## X2:X3:X1    -6.143e-01  3.621e-01  -1.697 0.094416 .
## X2:X4:X1     2.172e-07  6.507e-08   3.337 0.001383 **
## X2:X3:X4    -2.248e-05  9.524e-06  -2.361 0.021168 *
## X3:X4:X1    -3.169e-05  2.162e-05  -1.465 0.147508
## X2:X3:X4:X1  3.493e-06  2.090e-06   1.671 0.099346 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.974 on 67 degrees of freedom
## Multiple R-squared:  0.7313, Adjusted R-squared:  0.6792
## F-statistic: 14.03 on 13 and 67 DF,  p-value: 2.027e-14
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.60284 -0.57735 -0.01582  0.68584  1.95512
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.880e+00  1.418e+00   6.968 1.62e-09 ***
## X2           9.345e-01  3.439e-01   2.718 0.008336 **
## X3          -1.991e+01  1.037e+01  -1.921 0.058969 .
## X4           1.290e-05  2.322e-06   5.556 4.98e-07 ***
## I(X2^2)     -4.784e-02  2.036e-02  -2.349 0.021724 *
## X4:X1       -3.405e-06  7.194e-07  -4.733 1.16e-05 ***
## X2:X3        2.470e+00  1.258e+00   1.963 0.053693 .
## X3:X4        1.225e-04  6.007e-05   2.039 0.045331 *
## X2:X3:X1    -9.877e-02  5.507e-02  -1.793 0.077352 .
## X2:X4:X1     2.499e-07  6.145e-08   4.067 0.000126 ***
## X2:X3:X4    -1.558e-05  8.296e-06  -1.878 0.064650 .
## X3:X4:X1    -5.015e-06  1.125e-05  -0.446 0.657280
## X2:X3:X4:X1  8.114e-07  9.573e-07   0.848 0.399602
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9817 on 68 degrees of freedom
## Multiple R-squared:  0.723,  Adjusted R-squared:  0.6741
```

```
## F-statistic: 14.79 on 12 and 68 DF,  p-value: 1.336e-14
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.59950 -0.56491 -0.00854  0.65982  2.03161
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.005e+01  1.358e+00   7.397 2.53e-10 ***
## X2           8.995e-01  3.329e-01   2.702  0.00866 **
## X3          -2.007e+01  1.030e+01  -1.948  0.05544 .
## X4           1.265e-05  2.238e-06   5.651 3.31e-07 ***
## I(X2^2)     -4.572e-02  1.969e-02  -2.322  0.02316 *
## X4:X1       -3.569e-06  6.147e-07  -5.806 1.78e-07 ***
## X2:X3        2.487e+00  1.250e+00   1.990  0.05059 .
## X3:X4        1.162e-04  5.807e-05   2.002  0.04925 *
## X2:X3:X1    -1.120e-01  4.611e-02  -2.429  0.01774 *
## X2:X4:X1     2.645e-07  5.167e-08   5.120 2.64e-06 ***
## X2:X3:X4    -1.482e-05  8.070e-06  -1.836  0.07064 .
## X2:X3:X4:X1  4.071e-07  3.036e-07   1.341  0.18435
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.976 on 69 degrees of freedom
## Multiple R-squared:  0.7222, Adjusted R-squared:  0.6779
## F-statistic:  16.3 on 11 and 69 DF,  p-value: 3.516e-15
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.57383 -0.64001  0.03223  0.52384  1.82291
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.023e+01  1.359e+00   7.524 1.37e-10 ***
## X2           8.417e-01  3.319e-01   2.536   0.0135 *
## X3          -1.518e+01  9.689e+00  -1.567   0.1217
## X4           1.215e-05  2.220e-06   5.474 6.45e-07 ***
## I(X2^2)     -4.196e-02  1.960e-02  -2.141   0.0357 *
## X4:X1       -3.474e-06  6.141e-07  -5.658 3.11e-07 ***
## X2:X3        1.678e+00  1.101e+00   1.524   0.1320
## X3:X4        7.598e-05  4.999e-05   1.520   0.1330
## X2:X3:X1    -6.295e-02  2.822e-02  -2.230   0.0289 *
## X2:X4:X1     2.611e-07  5.189e-08   5.031 3.62e-06 ***
## X2:X3:X4    -7.743e-06  6.141e-06  -1.261   0.2115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.9815 on 70 degrees of freedom
## Multiple R-squared:  0.7149, Adjusted R-squared:  0.6742
## F-statistic: 17.55 on 10 and 70 DF,  p-value: 1.885e-15
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62099 -0.54338 -0.03675  0.73553  1.82049
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.887e+00  1.338e+00   7.391 2.24e-10 ***
## X2           9.049e-01  3.295e-01   2.746  0.00763 **
## X3          -4.510e+00  4.739e+00  -0.952  0.34444
## X4           1.118e-05  2.089e-06   5.350 1.02e-06 ***
## I(X2^2)     -4.398e-02  1.961e-02  -2.243  0.02804 *
## X4:X1       -3.243e-06  5.885e-07  -5.511 5.41e-07 ***
## X2:X3        4.802e-01  5.590e-01   0.859  0.39324
## X3:X4        1.530e-05  1.358e-05   1.127  0.26368
## X2:X3:X1    -7.822e-02  2.560e-02  -3.055  0.00317 **
## X2:X4:X1     2.425e-07  4.997e-08   4.853 6.97e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9856 on 71 degrees of freedom
## Multiple R-squared:  0.7085, Adjusted R-squared:  0.6715
## F-statistic: 19.17 on 9 and 71 DF,  p-value: 8.784e-16
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6192 -0.5164  0.0395  0.6678  1.9397
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.671e+00  1.311e+00   7.374 2.24e-10 ***
## X2           9.426e-01  3.260e-01   2.892  0.00506 **
## X3          -1.248e+00  2.830e+00  -0.441  0.66040
## X4           1.141e-05  2.068e-06   5.516 5.14e-07 ***
## I(X2^2)     -4.527e-02  1.952e-02  -2.319  0.02323 *
## X4:X1       -3.309e-06  5.825e-07  -5.680 2.66e-07 ***
## X3:X4        1.324e-05  1.334e-05   0.992  0.32439
## X2:X3:X1    -6.534e-02  2.071e-02  -3.154  0.00235 **
## X2:X4:X1     2.464e-07  4.967e-08   4.962 4.52e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.9838 on 72 degrees of freedom
## Multiple R-squared:  0.7054, Adjusted R-squared:  0.6727
## F-statistic: 21.55 on 8 and 72 DF,  p-value: 2.584e-16
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.56739 -0.50602  0.03801  0.64997  1.97139
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.641e+00  1.302e+00   7.402 1.86e-10 ***
## X2           9.302e-01  3.230e-01   2.880 0.005214 **
## X4           1.159e-05  2.013e-06   5.759 1.87e-07 ***
## I(X2^2)     -4.424e-02  1.927e-02  -2.295 0.024585 *
## X4:X1       -3.275e-06  5.743e-07  -5.703 2.35e-07 ***
## X4:X3        7.842e-06  5.292e-06   1.482 0.142679
## X2:X1:X3    -6.857e-02  1.927e-02  -3.559 0.000659 ***
## X2:X4:X1     2.439e-07  4.905e-08   4.972 4.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9784 on 73 degrees of freedom
## Multiple R-squared:  0.7046, Adjusted R-squared:  0.6763
## F-statistic: 24.88 on 7 and 73 DF,  p-value: < 2.2e-16
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54242 -0.47139 -0.03843  0.66551  1.95317
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.763e+00  1.310e+00   7.451 1.40e-10 ***
## X2           9.327e-01  3.256e-01   2.865  0.00543 **
## X4           1.338e-05  1.624e-06   8.240 4.56e-12 ***
## I(X2^2)     -4.680e-02  1.935e-02  -2.419  0.01804 *
## X4:X1       -3.468e-06  5.638e-07  -6.152 3.58e-08 ***
## X2:X1:X3    -5.684e-02  1.771e-02  -3.210  0.00196 **
## X2:X4:X1     2.541e-07  4.895e-08   5.190 1.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9862 on 74 degrees of freedom
## Multiple R-squared:  0.6957, Adjusted R-squared:  0.6711
## F-statistic:  28.2 on 6 and 74 DF,  p-value: < 2.2e-16
```

```r
PredictedSig<-predict(redu_mod,com_prop_data)

ModelSigData<-data.frame(obs = com_prop_data$Y,pred=PredictedSig)

ModelSig=defaultSummary(ModelSigData)

trns = rbind(ModelFull, ModelTrnsfrm, ModelSig)
print(trns)
```

```
##                    RMSE  Rsquared       MAE
## ModelFull    1.1012372 0.5847496 0.8268344
## ModelTrnsfrm 1.1023752 0.5840572 0.8265651
## ModelSig     0.9426472 0.6957387 0.7492499
```

The significant model (`ModelSig`) performs better than the full model (`ModelFull`) and the earlier-transformed-via-box-cox transformed model (`ModelTrnsfrm`) across all three metrics. It has lower RMSE Original RMSE: 1.10, Transformed RMSE: 1.10, Significant RMSE: 0.94) and MAE (Original MAE: 0.83, Transformed MAE: 0.83, Significant MAE: 0.75) values, suggesting that it has more accurate predictions, and a higher R^2 value (Original: 0.5847, Transformed: 0.5863, Significant: 0.6957), indicating that it explains a greater proportion of the variance in the dependent variable. The significant model retains essential predictors while eliminating those that do not contribute significantly to the model's explanatory power, resulting in better overall performance.

Problem 3

Using the sat data set (type following commands in R Console: library(faraway);data("sat")). Fit a model with the total SAT score as the response and expend, salary, ratio and takers as predictors. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant. Do not provide any plots about which you have nothing to say. Suggest possible improvements or corrections to the model where appropriate. (30 points, 5 points each)

```r
sat_data <- as.data.frame(sat)
sat_lm <- lm(total ~ expend + salary + ratio + takers, data=sat_data)
summary(sat_lm)
```

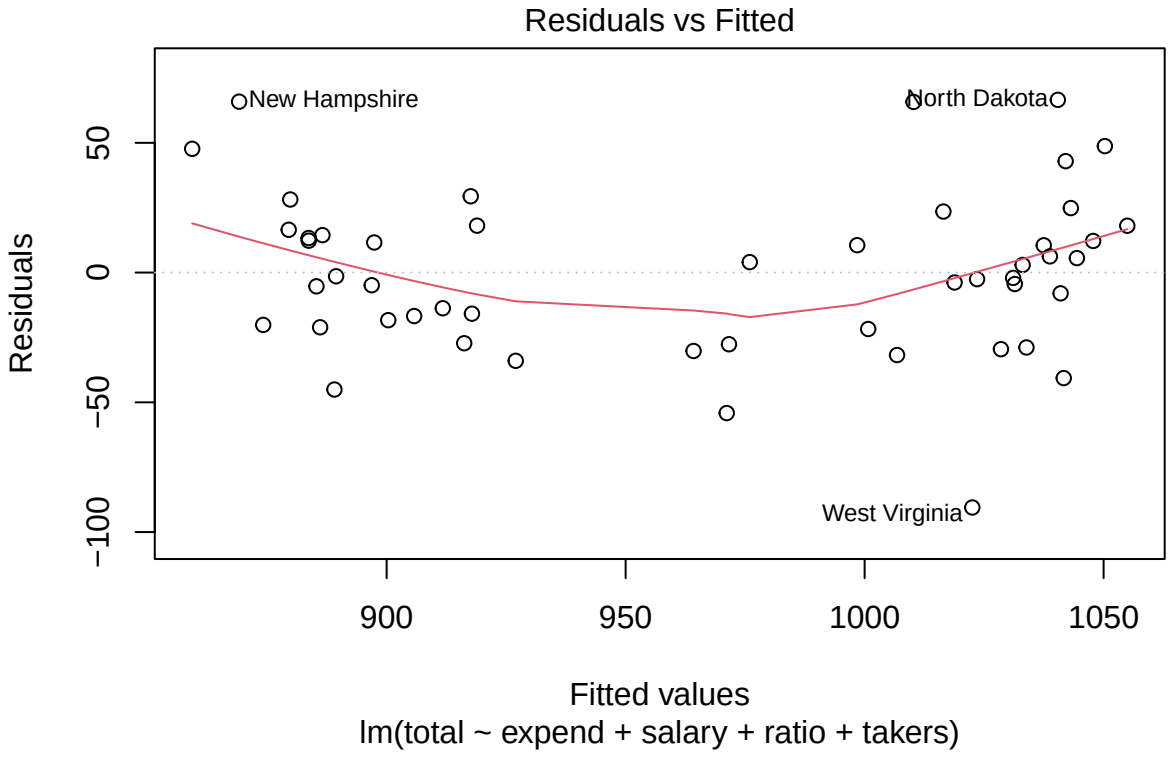
```
##
## Call:
## lm(formula = total ~ expend + salary + ratio + takers, data = sat_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.531 -20.855  -1.746  15.979  66.571
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
## expend         4.4626    10.5465   0.423    0.674
## salary         1.6379     2.3872   0.686    0.496
## ratio         -3.6242     3.2154  -1.127    0.266
## takers        -2.9045     0.2313 -12.559 2.61e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
```

```
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16
```

At first glance, it seems only `takers` is a significant predictor variable, with others not being significant. However, let us examine the underlying data to see what is going on and to see whether we can improve the performance of the model.

a-) Check the constant variance assumption for the errors.

```
plot(sat_lm, which=1)
```



To check the assumption of constant variance we plot fitted values against the residuals to look for any structure in the distribution of values about the theoretical mean value line $E[\epsilon] = 0$. For the most part, the variance shows random scatter, with a few potential outliers leading to the slight curvilinear nature of the line.

To verify the constancy of variance, we will conduct the Brown-Forsythe Test.
H_o : $\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$ The error variances are constant (homoscedasticity).
H_a : $\sigma_i^2 \neq \sigma_j^2$ The error variances are not constant (heteroscedasticity), where at least one group variance is not equal.
The decision rule is as follows:
1. If $p - value \geq \alpha$, then we fail to reject H_0, concluding the error variances are constant.
2. If $p - value < \alpha$, then we reject H_0 and conclude H_a, concluding the error variances are not constant.

```
ei <- residuals(sat_lm)
bf_data <- data.frame(expend = sat_data$expend, salary = sat_data$salary, ratio = sat_data$ratio, taker
median_takers <- median(sat_data$takers)
bf_data$takers_group <- as.factor(ifelse(bf_data$takers < median_takers, 1, 0))
leveneTest(ei ~ takers_group, data = bf_data, center = median)
```

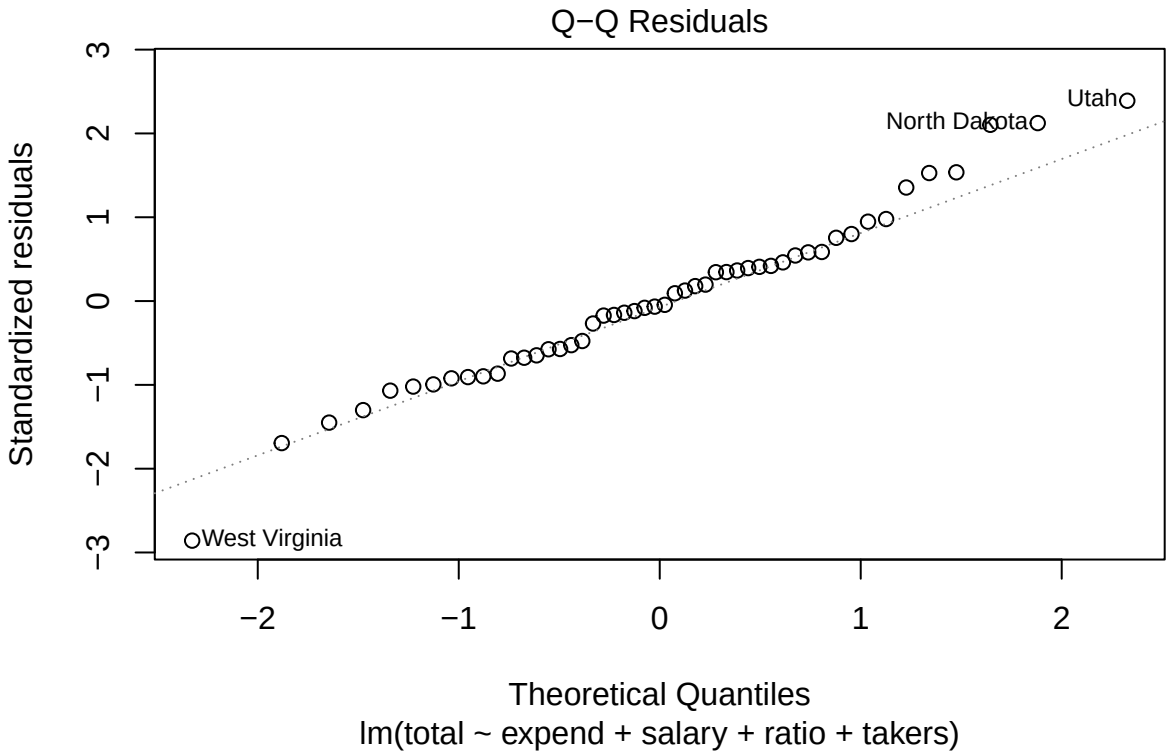
```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
```

```
## group  1  0.2056 0.6523
##       48
```

The Brown-Forsythe test assesses whether different groups exhibit equal variances, a condition known as homogeneity of variances. One advantage of this test is its robustness to non-normality in the data. In our analysis, the F statistic is 0.2056, with a p-value of 0.6523—significantly above the 0.05 threshold, so we are led to accept H_0 concluding that the error variances are constant, thereby indicating homoscedasticity. In other words, the variability of the residuals does not change across levels of the fitted values, and remains uniform.

b-) Check the normality assumption.

```
plot(sat_lm, which = 2)
```



The residuals appear normally distributed in the middle of the range; however, the distribution is slightly right skewed and there are a couple of points on the upper quantile (North Dakota, Utah) and one on the lower quantile (West Virgina) that deviates from the theoretical distribution which might be potential outliers.

To test this using a non-parametric metric, I will conduct the **Shapiro-Wilk** test.
First, let us state the hypotheses:
H_o: The residuals follow a normal distribution.
H_a: The residuals do not follow a normal distribution.

```
residuals <- residuals(sat_lm)
shapiro_test_result <- shapiro.test(residuals)
print(shapiro_test_result)
```

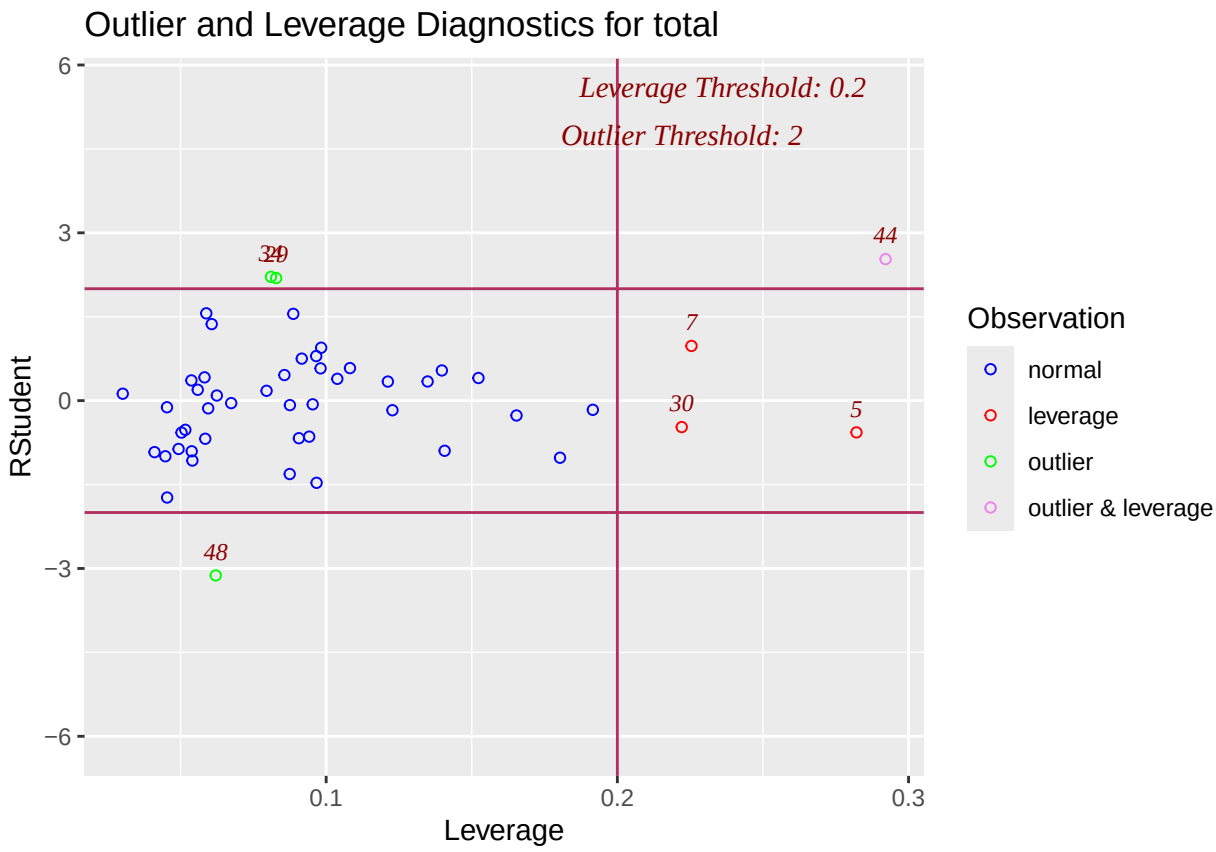
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
```

```
## W = 0.97691, p-value = 0.4304
```

The W statistic measures how closely the sample distribution matches a normal distribution. Values close to 1 indicate that the sample distribution is similar to a normal distribution. A p-value of 0.4304 is quite high and above the chosen significance level $\alpha = 0.05$, thereby we accept H_o and conclude there is not enough evidence to conclude that the residuals deviate significantly from normality. The residuals from our model do not show evidence of significant departure from normality based on the Shapiro-Wilk test. Thus, we can reasonably assume that the normality assumption for the residuals of our model is satisfied.

c-) Check for large leverage points.

```
ols_plot_resid_lev(sat_lm)
```




```
influence_data <- influence(sat_lm)

leverage <- hatvalues(sat_lm) # Leverage values
standardized_residuals <- rstandard(sat_lm) # Standardized residuals

# Define thresholds for outliers and high leverage
# Threshold for high leverage: 2*(p+1)/n where p is the number of predictors, n is the number of observ
n <- nrow(sat_data)
p <- length(coef(sat_lm)) - 1
leverage_threshold <- 2 * (p + 1) / n

# Threshold for outliers (standardized residuals > 2 or < -2)
outlier_threshold <- 2

# Find indices of high leverage points
high_leverage_indices <- which(leverage > leverage_threshold)
```

```r
# Find indices of outliers (standardized residuals)
outlier_indices <- which(abs(standardized_residuals) > outlier_threshold)

# Find indices of points that are both outliers and high leverage
both_indices <- intersect(high_leverage_indices, outlier_indices)

cat("Outliers: ", outlier_indices, "\n")
```

```
## Outliers:  29 34 44 48
```

```r
cat("High Leverage Points: ", high_leverage_indices, "\n")
```

```
## High Leverage Points:  5 7 30 44
```

```r
cat("Both Outliers and High Leverage: ", both_indices, "\n")
```

```
## Both Outliers and High Leverage:  44
```

```r
significant_indices <- which(leverage > leverage_threshold)
sat_data_filtered <- sat_data[-significant_indices, ]
sat_lm_wo_high_levg_pts <- lm(total ~ expend + salary + ratio + takers, data = sat_data_filtered)
summary(sat_lm_wo_high_levg_pts)
```

```
##
## Call:
## lm(formula = total ~ expend + salary + ratio + takers, data = sat_data_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -93.821 -17.364   1.386  13.667  66.319
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1088.4407    59.4084  18.321  < 2e-16 ***
## expend         2.7706    11.4915   0.241   0.8107
## salary         2.3269     2.6397   0.882   0.3832
## ratio         -7.0333     3.8944  -1.806   0.0783 .
## takers        -2.9282     0.2259 -12.963 4.32e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.75 on 41 degrees of freedom
## Multiple R-squared:  0.8347, Adjusted R-squared:  0.8186
## F-statistic: 51.75 on 4 and 41 DF,  p-value: 1.711e-15
```

Table 1.2: Examining Model Performance with and without High Leverage Points

Variable	Model with High Leverage Points	Model without High Leverage Points
RSE	32.7	31.75
df	45	41
R^2	0.8246	0.8347
Adj. R^2	0.809	0.8186
Residuals - Min.	-90.531	-93.821

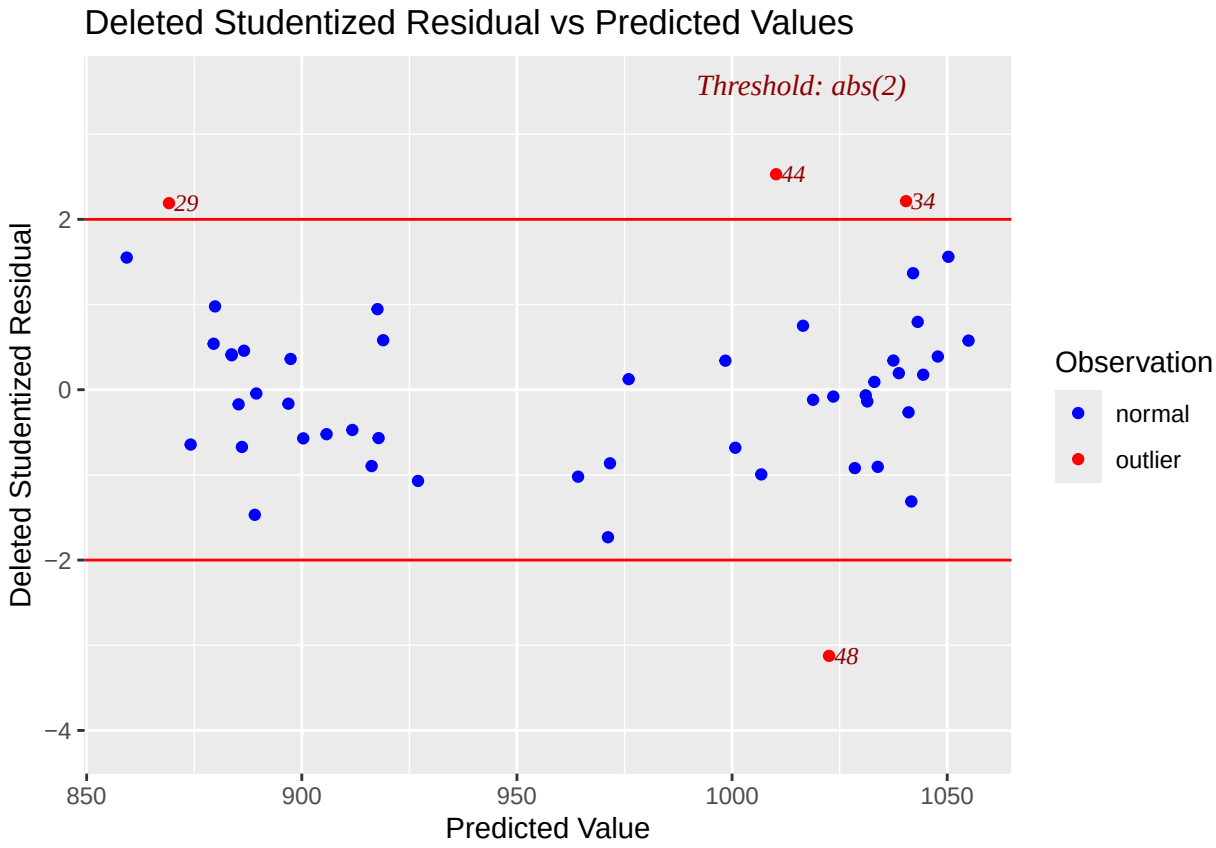
Variable	Model with High Leverage Points	Model without High Leverage Points
Residuals - Max.	66.571	66.319

Removing high leverage observations from the model leads to a negligible enhancement in its performance. Specifically, we observe slight increment in R^2 value, indicating that the model explains a slightly greater amount of the variance in the response variable. Additionally, the range of residuals widens, suggesting reduced accuracy in predictions. There is a slight reduction in the Residual Standard Error (RSE) further confirms this trend, indicating a slightly more precise fit to the data. Importantly, the significance of t-values of each of the non-significant independent variables reduces closer to the significance level, but still remains far. This transformation to address high leverage points does not make a a strong case as a sole transformation on this data model to improve performance.

d-) Check for outliers.

We venture to identify outliers by visualizing a diagnostic plot that displays studentized residuals versus fitted values in an ordinary least squares (OLS) regression model. Our graph identifies observation(s) 29, 44, 48 and 34 as potential outliers with a threshold of |2|, lower than the theoretical standardized residual threshold |3|.

```
ols_plot_resid_stud_fit(sat_lm)
```




```
std_residuals <- rstudent(sat_lm)
outlier_indices <- which(abs(std_residuals) > 2)
sat_data_wo_outliers <- sat_data[-outlier_indices, ]
sat_lm_wo_outliers <- lm(total ~ expend + salary + ratio + takers, data = sat_data_wo_outliers)
summary(sat_lm_wo_outliers)
```

```
##
## Call:
## lm(formula = total ~ expend + salary + ratio + takers, data = sat_data_wo_outliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.174 -12.506   1.451  13.995  46.244
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1075.7621    43.0392  24.995   <2e-16 ***
## expend         2.1782     8.1187   0.268   0.7898
## salary         3.0492     1.8528   1.646   0.1075
## ratio         -7.3212     2.7137  -2.698   0.0101 *
## takers        -3.0714     0.1782 -17.240   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.36 on 41 degrees of freedom
## Multiple R-squared:  0.8984, Adjusted R-squared:  0.8884
## F-statistic:  90.6 on 4 and 41 DF,  p-value: < 2.2e-16
```

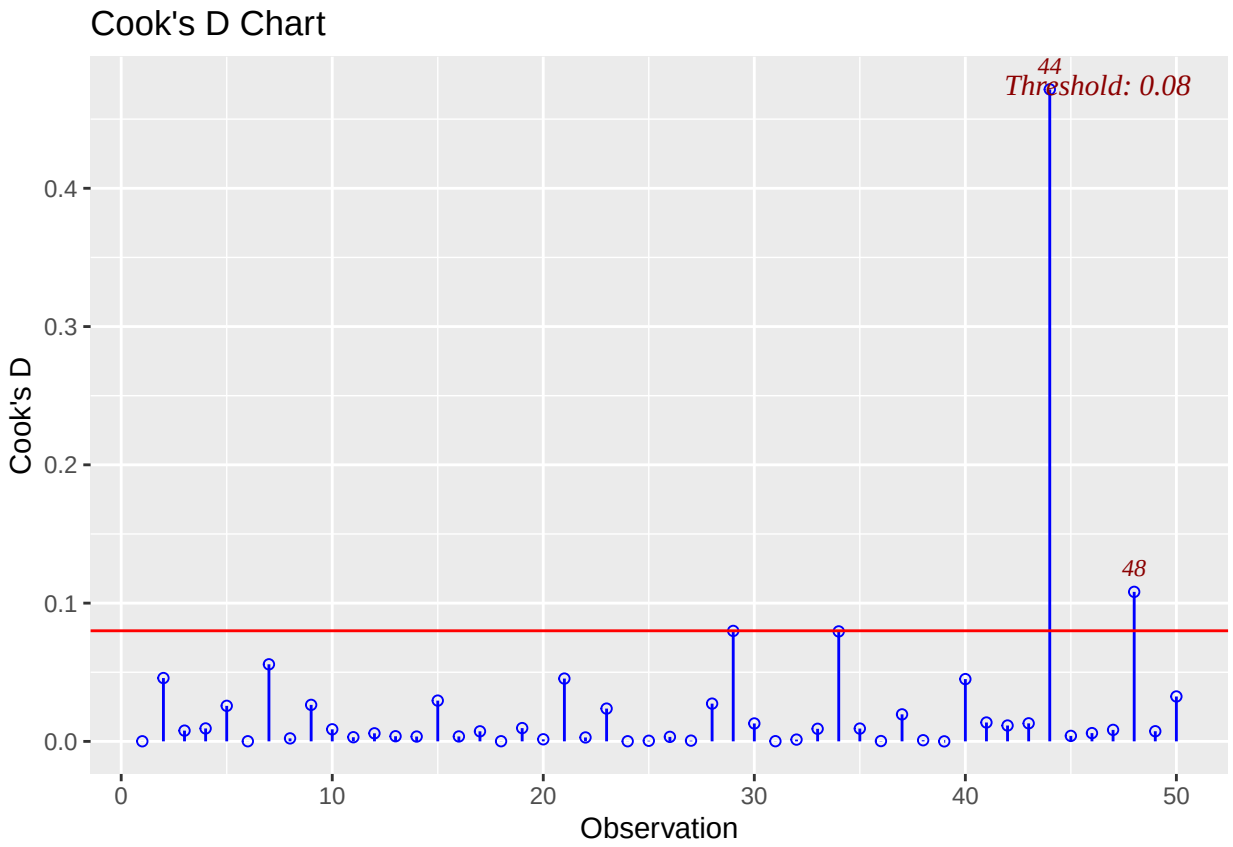
Table 1.3: Examining Model Performance with and without Potential Outliers

Variable	Model with Potential Outliers	Model without Potential Outliers
RSE	32.7	24.36
df	45	41
R^2	0.8246	0.8984
Adj. R^2	0.809	0.8884
Residuals - Min.	-90.531	-47.174
Residuals - Max.	66.571	46.244

Removing potential outlying observations from the model leads to a notable enhancement in its performance. Specifically, we observe a substantial increase in the R^2 value, indicating that the model now explains a greater proportion of the variance in the response variable. Additionally, the range of residuals tightens, suggesting improved accuracy in predictions. The reduction in the Residual Standard Error (RSE) further confirms this trend, indicating a more precise fit to the data. Importantly, the significance of the independent variable ratio emerges at the $\alpha = 0.01$ level, with a t_{ratio} of 0.01. This revelation underscores the model's refined ability to identify relevant predictors, highlighting the importance of addressing potential outlying points in regression analysis.

e-) Check for influential points.

```
ols_plot_cooksd_chart(sat_lm)
```




```
cooks_threshold <- 0.08  # As identified from graph above
cooks_distances <- cooks.distance(sat_lm)
high_influence_points <- which(cooks_distances > cooks_threshold)
cat("High influence points (Cook's Distance > ", cooks_threshold, "):\n")
```

```
## High influence points (Cook's Distance >  0.08 ):
```

```
print(high_influence_points)
```

```
##          Utah West Virginia
##            44            48
```

```
cooksd_df <- data.frame(Index = 1:length(cooks_distances), Cook = cooks_distances)
```

There are two influential points (West Virginia and Utah) identified by Cook's Distance. Let us examine the performance of the model if we remove these influential points.

```
sat_lm_wo_infl_pts <- lm(total ~ expend + salary + ratio + takers, data = sat_data, subset = (cooks_dis
summary(sat_lm_wo_infl_pts)
```

```
##
## Call:
## lm(formula = total ~ expend + salary + ratio + takers, data = sat_data,
##     subset = (cooks_distances < cooks_threshold))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -48.477 -16.066  -0.417  12.046  63.545
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1106.9838    48.0189  23.053   <2e-16 ***
## expend         1.8681     9.1687   0.204    0.840
## salary         2.5734     2.0917   1.230    0.225
## ratio         -8.0628     3.0739  -2.623    0.012 *
## takers        -3.0006     0.1971 -15.226   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.69 on 43 degrees of freedom
## Multiple R-squared:  0.8737, Adjusted R-squared:  0.8619
## F-statistic: 74.34 on 4 and 43 DF,  p-value: < 2.2e-16
```

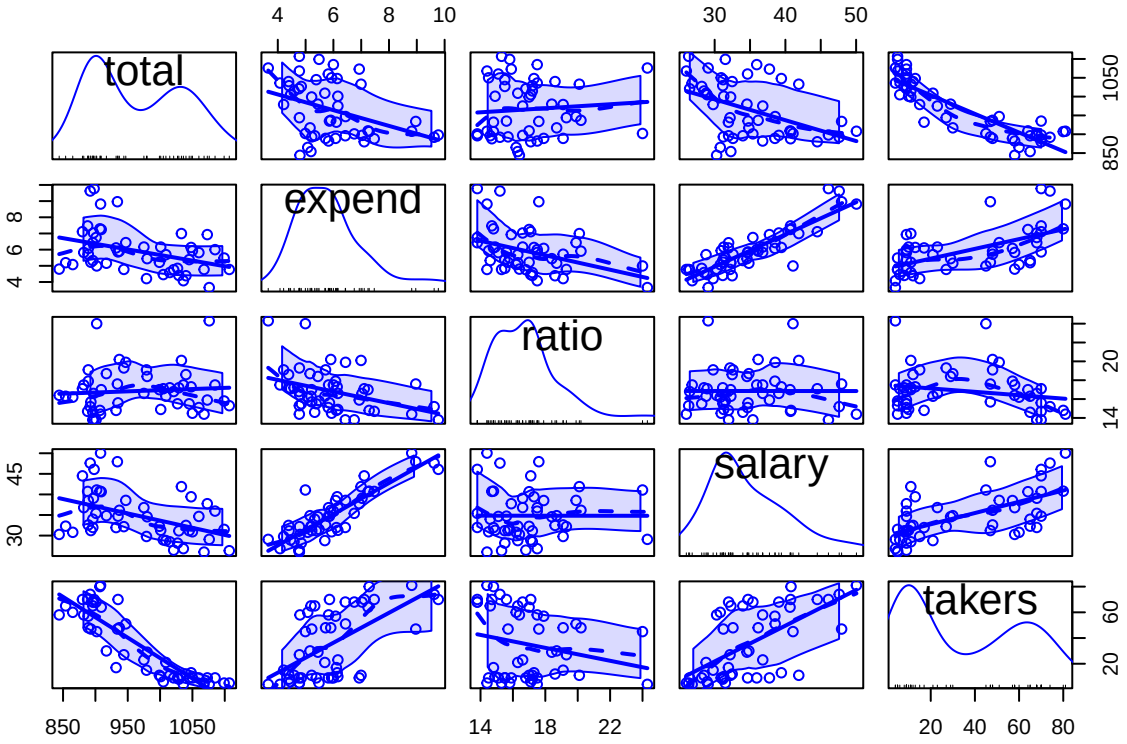
Table 1.4: Examining Model Performance with and without Influential Observations

Variable	Model with High Leverage Points	Model without High Leverage Points
RSE	32.7	27.69
df	45	43
R^2	0.8246	0.8737
Adj. R^2	0.809	0.8619
Residuals - Min.	-90.531	-48.477
Residuals - Max.	66.571	63.545

Removing influential observations from the model leads to a notable enhancement in its performance. Specifically, we observe a substantial increase in the R^2 value, indicating that the model now explains a greater proportion of the variance in the response variable. Additionally, the range of residuals tightens, suggesting improved accuracy in predictions. The reduction in the Residual Standard Error (RSE) further confirms this trend, indicating a more precise fit to the data. Importantly, the significance of the independent variable ratio emerges at the $\alpha = 0.01$ level, with a t_{ratio} of 0.01 (at a slightly lower significance than was observed when removing potential high leverage points $t_{ratio} = 0.0101$). This revelation underscores the model's refined ability to identify relevant predictors, highlighting the importance of addressing influential points in regression analysis.

f-) Check the structure of the relationship between the predictors and the response. Visualization

```
scatterplotMatrix(sat[c("total","expend","ratio","salary","takers")], smooth=TRUE, regLine = TRUE)
```



From our scatterplot analysis at the beginning of the analysis, we observed a deviation between the smoothed lowess and the regression line plotted. Let's dive deeper into the model to see what else we can gather about the non-linear nature of this dataset.

ANOVA

Let's examine the anova of all four of the models to understand whether the predictors collectively explain a significant amount of variance in the response.

```
combined_anova <- rbind(anova(sat_lm), anova(sat_lm_wo_infl_pts), anova(sat_lm_wo_outliers), anova(sat_l
print(combined_anova)
```

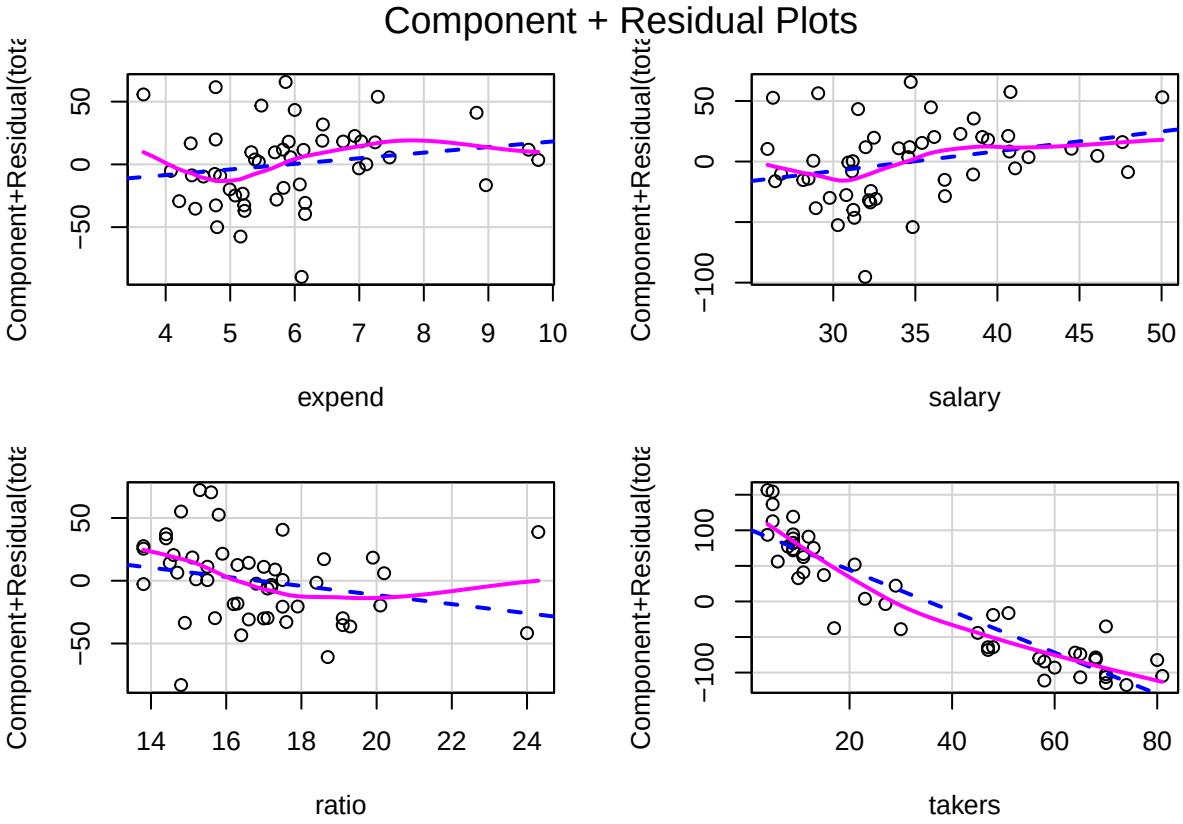
```
## Analysis of Variance Table
##
## Response: total
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## expend        1  39722   39722  37.1436 2.260e-07 ***
## salary        1  13361   13361  12.4933 0.0009585 ***
## ratio         1   4413    4413   4.1266 0.0481449 *
## takers        1 168688  168688 157.7379 2.607e-16 ***
## Residuals    45  48124    1069
## expend1       1  31479   31479  41.0640 9.314e-08 ***
## salary1       1  17812   17812  23.2357 1.819e-05 ***
## ratio1        1    943     943   1.2300 0.2735785
## takers1       1 177723  177723 231.8401 < 2.2e-16 ***
## Residuals1   43  32963     767
## expend2       1  25804   25804  43.4775 6.194e-08 ***
## salary2       1  12133   12133  20.4435 5.150e-05 ***
## ratio2        1    758     758   1.2771 0.2650000
## takers2       1 176395  176395 297.2067 < 2.2e-16 ***
## Residuals2   41  24334     594
## expend3       1  26458   26458  26.2485 7.536e-06 ***
## salary3       1  10358   10358  10.2754 0.0026124 **
```

```
## ratio3      1   2469    2469   2.4493 0.1252639
## takers3     1 169386  169386 168.0422 4.317e-16 ***
## Residuals3 41  41328    1008
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table summarizes the contribution of four predictor variables—expend, salary, ratio, and takers—in explaining the variability in the response variable total. I compared the models in order of creation: `sat_lm`, `sat_lm_wo_infl_pts`, `sat_lm_wo_outliers` and `sat_lm_wo_high_levg_pts`. Overall, the significant predictors—especially `takers`, followed by `expend` and `salary`—provide substantial explanatory power. However, `ratio`, while significant, does not contribute as strongly to explaining the variation in `total` in any of the models. The residual sum of squares (RSS) is reduced the most in the model without outliers, `sat_lm_wo_outliers`, at 24334 with the lowest mean squared error (MSE = 594). This suggests that the model, without a few of its challenging data points, fits the data reasonably well, but there is still some variability that is not captured by the predictors. This analysis implies that focusing on `takers` and `expend` could offer the most insight when predicting total, but further refinement of the model *(perhaps combining the removal of both high influence and outliers may generate a model that outperforms `sat_lm_wo_outliers`)* may help improve the explanation of residual variance.

Component + Residual Plots to Detect Non-Linear Relationships between the Response and each Predictor Variable

```
crPlots(sat_lm)
```



Interpretation of the plots:
1. Expend: The residuals seem randomly scattered around zero, but there is a slight curvature suggesting that the relationship between expend and the response may not be purely linear.
2. Salary: The residuals are also fairly scattered around zero, with some indication of curvature, hinting at a possible non-linear relationship between salary and the response.
3. Ratio: There's some variability in the residuals, though the overall pattern is more consistent around zero.

This suggests that ratio is a reasonably good linear predictor, but the slight curvature might indicate the need for further exploration.
4. Takers: The residuals show a clear pattern of non-linearity, with the fitted curve (solid line) deviating significantly from a straight line. This strong curvature suggests that the relationship between "takers" and the response is highly non-linear.

The lowess smooth, the magenta curve, suggests that the linear model might not fully capture the complexity in the data for variables `takers`, `expend`, `ratio` and `salary`. The blue dashed lines represent the linear regression fit, and while they capture some instances of the relationship, they fail to reflect the curvilinear nature that arises. This suggests that transformations of these predictors, or considering a non-linear model, may improve the fit.

Multicolinearty Exploration
Now, we examine whether there is multicolinearity between the predictors.

```
ols_vif_tol(sat_lm)
```

```
##   Variables Tolerance      VIF
## 1    expend 0.1056488 9.465320
## 2    salary 0.1084924 9.217237
## 3     ratio 0.4109808 2.433204
## 4    takers 0.5697714 1.755090
```

VIF values greater than 5 or 10 indicate multicollinearity issues, meaning that predictors may be highly correlated with each other, potentially distorting the model. We clearly see the presence of multicollinearity of expend and salary. High multicollinearity can inflate the standard errors of the coefficients, making it harder to assess their individual significance.

We venture now to see if removing one of the multicollinear predictors will improve the performance of the model. As per the initial summary(sat_lm), expend (p-value = 0.674) is less significant than salary (p-value = 0.496), so we attempt to build a simplified model without this predictor.

First, we perform the General F-test to test whether we can justify dropping the `expend` variable:
H_o: $\beta_{expend} = 0$
H_a: $\beta_{expend} \neq 0$ and it is significant so we cannot drop it.

```
sat_lm_simplified <- lm(total ~ salary + ratio + takers, data = sat_data)
anova(sat_lm_simplified,sat_lm)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ salary + ratio + takers
## Model 2: total ~ expend + salary + ratio + takers
##   Res.Df   RSS Df Sum of Sq     F Pr(>F)
## 1     46 48315
## 2     45 48124  1    191.47 0.179 0.6742
```

The `anova()` function performs a hypothesis test comparing the two models, and in this case, the p-value is 0.6742 which is higher than our chosen significance $\alpha = 0.05$, so we fail to reject the null hypothesis and we can drop the predictor variable, `expend`.

```
ols_vif_tol(sat_lm_simplified)
```

```
##   Variables Tolerance     VIF
## 1    salary 0.6018078 1.66166
## 2     ratio 0.9272221 1.07849
## 3    takers 0.5744914 1.74067
```

```
cat("Original Model: Adjusted R-squared =", summary(sat_lm)$adj.r.squared, "AIC =", AIC(sat_lm), "BIC =
```

```
## Original Model: Adjusted R-squared = 0.8089679 AIC = 497.3694 BIC = 508.8415
```

```
cat("Simplified Model: Adjusted R-squared =", summary(sat_lm_simplified)$adj.r.squared, "AIC =", AIC(sa
```

```
## Simplified Model: Adjusted R-squared = 0.8123772 AIC = 495.568 BIC = 505.1281
```

Upon removing one of the multicolinear predictor, `expend`, we see no presence of significant multicolinearity in the variance inflation factors (VIFs) of the simplified model. There is a slight increase in the adjusted R^2 when comparing the original model and the simplified model. The Akaike Information Criterion (AIC), which assesses the trade-off between model complexity and goodness of fit, is lower in the simplified model suggesting that it has a better balance between goodness of fit and model complexity. The BIC (Bayesian Information Criterion), which penalizes model complexity but is generally stricter (especially when the sample size is large–not the case for us as we only have 50 observations in `sat_data`), is lower in the simplified model (505.13 compared to 508.84), which reinforces the conclusion that the simplified model is optimal.

As an alternative, we can combine the two multicolinear variables in a weighted approach based on the correlation between each variable and the response variable (e.g., total) and assigning higher weight to the variable that has a stronger correlation.

```
cor_salary <- cor(sat_data$salary, sat_data$total)
cor_expend <- cor(sat_data$expend, sat_data$total)

# Normalize the correlations to sum to 1
total_cor <- cor_salary + cor_expend
weight_salary <- cor_salary / total_cor
weight_expend <- cor_expend / total_cor

# Create the index using the correlation-based weights
sat_data$salary_expend_index <- weight_salary * sat_data$salary + weight_expend * sat_data$expend

sat_lm_comb <- lm(total ~ salary_expend_index + ratio + takers, data = sat_data)
summary(sat_lm_comb)
```

```
##
## Call:
## lm(formula = total ~ salary_expend_index + ratio + takers, data = sat_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -89.797 -21.093  -1.388  17.061  67.275
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1052.4363    45.3243  23.220   <2e-16 ***
## salary_expend_index    4.0714     1.5853   2.568   0.0135 *
## ratio                 -4.2161     2.0968  -2.011   0.0502 .
## takers                -2.9121     0.2268 -12.840   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.37 on 46 degrees of freedom
## Multiple R-squared:  0.8243, Adjusted R-squared:  0.8129
## F-statistic: 71.95 on 3 and 46 DF,  p-value: < 2.2e-16
```

```r
anova(sat_lm_comb, sat_lm)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ salary_expend_index + ratio + takers
## Model 2: total ~ expend + salary + ratio + takers
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     46 48188
## 2     45 48124  1    64.027 0.0599 0.8078
```

```r
cat("Original Model: Adjusted R-squared =", summary(sat_lm)$adj.r.squared, "AIC =", AIC(sat_lm), "\n")
```

```
## Original Model: Adjusted R-squared = 0.8089679 AIC = 497.3694
```

```r
cat("Simplified Model: Adjusted R-squared =", summary(sat_lm_comb)$adj.r.squared, "AIC =", AIC(sat_lm_c
```

```
## Simplified Model: Adjusted R-squared = 0.8128721 AIC = 495.4359
```

When we perform the combined transformation to avoid the adverse effects of multicollinearity, we see a slight improvement in the Adjusted R^2 and a reduced AIC similar to what we observed when we dropped one of the insignificant multicollienar predictor variables.

Interaction Effects An interaction effect occurs when the effect of one or more independent variables on a dependent variable changes depending on the value of another independent variable. I will first build a second-order model of the centered variables in the sat_data, then using the observations from the scatter plot matrix, I will create other versions of the model.

```r
sat_data$expend_centered <- sat_data$expend - mean(sat_data$expend)
sat_data$salary_centered <- sat_data$salary - mean(sat_data$salary)
sat_data$ratio_centered <- sat_data$ratio - mean(sat_data$ratio)
sat_data$takers_centered <- sat_data$takers - mean(sat_data$takers)

# Full Model
model_full_interaction <- lm(total ~ expend_centered * salary_centered * ratio_centered * takers_center
summary(model_full_interaction)
```

```
##
## Call:
## lm(formula = total ~ expend_centered * salary_centered * ratio_centered *
##     takers_centered, data = sat_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.654 -14.430  -0.575  12.294  57.020
##
## Coefficients:
##                                                    Estimate
## (Intercept)                                      965.330599
## expend_centered                                   27.697558
## salary_centered                                    1.001970
## ratio_centered                                     0.081049
## takers_centered                                   -2.998041
## expend_centered:salary_centered                   -1.840085
## expend_centered:ratio_centered                     7.055487
## salary_centered:ratio_centered                     0.615823
## expend_centered:takers_centered                    0.631314
## salary_centered:takers_centered                    0.040446
```

```
## ratio_centered:takers_centered                                          -0.496914
## expend_centered:salary_centered:ratio_centered                          -1.995187
## expend_centered:salary_centered:takers_centered                         -0.059588
## expend_centered:ratio_centered:takers_centered                          0.601127
## salary_centered:ratio_centered:takers_centered                         -0.043469
## expend_centered:salary_centered:ratio_centered:takers_centered          0.003613
##                                                                       Std. Error
## (Intercept)                                                             6.412489
## expend_centered                                                        13.239482
## salary_centered                                                         2.564902
## ratio_centered                                                          4.786688
## takers_centered                                                         0.331144
## expend_centered:salary_centered                                         0.993399
## expend_centered:ratio_centered                                          4.469899
## salary_centered:ratio_centered                                          0.725275
## expend_centered:takers_centered                                         0.542846
## salary_centered:takers_centered                                         0.104511
## ratio_centered:takers_centered                                          0.257044
## expend_centered:salary_centered:ratio_centered                          0.866231
## expend_centered:salary_centered:takers_centered                         0.040481
## expend_centered:ratio_centered:takers_centered                          0.239397
## salary_centered:ratio_centered:takers_centered                          0.043661
## expend_centered:salary_centered:ratio_centered:takers_centered          0.020255
##                                                                       t value Pr(>|t|)
## (Intercept)                                                           150.539  < 2e-16
## expend_centered                                                         2.092   0.0440
## salary_centered                                                         0.391   0.6985
## ratio_centered                                                          0.017   0.9866
## takers_centered                                                        -9.054  1.4e-10
## expend_centered:salary_centered                                        -1.852   0.0727
## expend_centered:ratio_centered                                          1.578   0.1237
## salary_centered:ratio_centered                                          0.849   0.4018
## expend_centered:takers_centered                                         1.163   0.2529
## salary_centered:takers_centered                                         0.387   0.7012
## ratio_centered:takers_centered                                         -1.933   0.0616
## expend_centered:salary_centered:ratio_centered                         -2.303   0.0275
## expend_centered:salary_centered:takers_centered                        -1.472   0.1502
## expend_centered:ratio_centered:takers_centered                          2.511   0.0170
## salary_centered:ratio_centered:takers_centered                         -0.996   0.3265
## expend_centered:salary_centered:ratio_centered:takers_centered          0.178   0.8595
##
## (Intercept)                                                           ***
## expend_centered                                                       *
## salary_centered
## ratio_centered
## takers_centered                                                       ***
## expend_centered:salary_centered                                       .
## expend_centered:ratio_centered
## salary_centered:ratio_centered
## expend_centered:takers_centered
## salary_centered:takers_centered
## ratio_centered:takers_centered                                        .
## expend_centered:salary_centered:ratio_centered                        *
## expend_centered:salary_centered:takers_centered
```

```
## expend_centered:ratio_centered:takers_centered                 *
## salary_centered:ratio_centered:takers_centered
## expend_centered:salary_centered:ratio_centered:takers_centered
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.65 on 34 degrees of freedom
## Multiple R-squared:  0.891,  Adjusted R-squared:  0.8429
## F-statistic: 18.53 on 15 and 34 DF,  p-value: 4.035e-12
```

```
# Only Significant Model
model_significant <- lm(total ~ takers_centered + expend_centered + expend_centered:salary_centered:rati
summary(model_significant)
```

```
##
## Call:
## lm(formula = total ~ takers_centered + expend_centered + expend_centered:salary_centered:ratio_cente
##     expend_centered:ratio_centered:takers_centered, data = sat_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -89.111 -22.564   3.234  19.776  70.244
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                   965.93831    4.76159 202.860
## takers_centered                                -2.86169    0.22066 -12.969
## expend_centered                                13.44384    5.24010   2.566
## expend_centered:salary_centered:ratio_centered -0.20215    0.37535  -0.539
## takers_centered:expend_centered:ratio_centered  0.06669    0.10898   0.612
##                                                Pr(>|t|)
## (Intercept)                                      <2e-16 ***
## takers_centered                                  <2e-16 ***
## expend_centered                                  0.0137 *
## expend_centered:salary_centered:ratio_centered   0.5928
## takers_centered:expend_centered:ratio_centered   0.5437
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.04 on 45 degrees of freedom
## Multiple R-squared:  0.821,  Adjusted R-squared:  0.805
## F-statistic: 51.59 on 4 and 45 DF,  p-value: 3.024e-16
```

```
anova(model_significant, model_full_interaction)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ takers_centered + expend_centered + expend_centered:salary_centered:ratio_centered +
##     expend_centered:ratio_centered:takers_centered
## Model 2: total ~ expend_centered * salary_centered * ratio_centered *
##     takers_centered
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1     45 49111
## 2     34 29894 11     19217 1.9869 0.06187 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The second-order model reveals many significant terms due to the sheer volume of degrees of freedom, with the following significant at $\alpha = 0.05$: `expend_centered`, `takers_centered`, `expend_centered:salary_centered:ratio_centere` and `expend_centered:ratio_centered:takers_centered`. When we build a model with only significant predictors, it explains less variation than the additional interaction terms in the full model and it does not provide a statistically significant improvement in model fit at the 0.05 level ($p = 0.06187$). Neither of the models is successful at explaining the data significantly as we gathered when iterating through the second-order model earlier–you get different results with each variable you remove, so it needs to be an incremental process.

Now, we opt to go step-wise in a manual manner to identify the most significant model for the SAT dataset.

```r
drop_insignificant_vars <- function(model, data) {
  while (TRUE) {
    p_values <- summary(model)$coefficients[, 4]
    max_p_value <- max(p_values[-1])  # Ignore intercept (p-values[-1])
    if (max_p_value <= 0.05) break
    var_to_drop <- names(p_values)[which.max(p_values)]
    formula <- as.formula(update(model, paste(". ~ . -", var_to_drop)))
    model <- lm(formula, data = data)
    print(summary(model))
  }
  return(model)
}

sat_redu_lm <- drop_insignificant_vars(model_full_interaction, sat_data)
```

```
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -79.73 -14.38  -0.55  12.24  56.98
##
## Coefficients:
##                                                                 Estimate
## (Intercept)                                                  965.306888
## expend_centered                                               27.538554
## salary_centered                                                1.033410
## takers_centered                                               -2.996591
## expend_centered:salary_centered                               -1.836281
## expend_centered:ratio_centered                                 7.036223
## salary_centered:ratio_centered                                 0.618756
## expend_centered:takers_centered                                0.633235
## salary_centered:takers_centered                                0.039910
## takers_centered:ratio_centered                                -0.494419
## expend_centered:salary_centered:ratio_centered                 -1.988694
## expend_centered:salary_centered:takers_centered                -0.059739
## expend_centered:takers_centered:ratio_centered                 0.600301
## salary_centered:takers_centered:ratio_centered                -0.043425
## expend_centered:salary_centered:takers_centered:ratio_centered  0.003454
##                                                              Std. Error
## (Intercept)                                                    6.167697
## expend_centered                                                9.198505
```

```
## salary_centered                                                 1.744013
## takers_centered                                                 0.315271
## expend_centered:salary_centered                                 0.953743
## expend_centered:ratio_centered                                  4.260492
## salary_centered:ratio_centered                                  0.694157
## expend_centered:takers_centered                                 0.523221
## salary_centered:takers_centered                                 0.098182
## takers_centered:ratio_centered                                  0.207570
## expend_centered:salary_centered:ratio_centered                  0.765555
## expend_centered:salary_centered:takers_centered                 0.038915
## expend_centered:takers_centered:ratio_centered                  0.230996
## salary_centered:takers_centered:ratio_centered                  0.042954
## expend_centered:salary_centered:takers_centered:ratio_centered  0.017685
##                                                                t value Pr(>|t|)
## (Intercept)                                                    156.510  < 2e-16
## expend_centered                                                  2.994  0.00503
## salary_centered                                                  0.593  0.55729
## takers_centered                                                 -9.505 3.15e-11
## expend_centered:salary_centered                                 -1.925  0.06234
## expend_centered:ratio_centered                                   1.652  0.10758
## salary_centered:ratio_centered                                   0.891  0.37881
## expend_centered:takers_centered                                  1.210  0.23429
## salary_centered:takers_centered                                  0.406  0.68685
## takers_centered:ratio_centered                                  -2.382  0.02279
## expend_centered:salary_centered:ratio_centered                  -2.598  0.01364
## expend_centered:salary_centered:takers_centered                 -1.535  0.13375
## expend_centered:takers_centered:ratio_centered                   2.599  0.01360
## salary_centered:takers_centered:ratio_centered                  -1.011  0.31898
## expend_centered:salary_centered:takers_centered:ratio_centered   0.195  0.84628
##
## (Intercept)                                                    ***
## expend_centered                                                **
## salary_centered
## takers_centered                                                ***
## expend_centered:salary_centered                                .
## expend_centered:ratio_centered
## salary_centered:ratio_centered
## expend_centered:takers_centered
## salary_centered:takers_centered
## takers_centered:ratio_centered                                 *
## expend_centered:salary_centered:ratio_centered                 *
## expend_centered:salary_centered:takers_centered
## expend_centered:takers_centered:ratio_centered                 *
## salary_centered:takers_centered:ratio_centered
## expend_centered:salary_centered:takers_centered:ratio_centered
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.23 on 35 degrees of freedom
## Multiple R-squared:  0.891,  Adjusted R-squared:  0.8474
## F-statistic: 20.44 on 14 and 35 DF,  p-value: 8.665e-13
##
##
## Call:
```

```
## lm(formula = formula, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.834 -14.636  -0.005  12.236  56.660
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                     965.45701    6.03731 159.915
## expend_centered                                  27.77204    8.99782   3.087
## salary_centered                                   0.97000    1.69048   0.574
## takers_centered                                  -2.98049    0.30020  -9.928
## expend_centered:salary_centered                  -1.89578    0.89163  -2.126
## expend_centered:ratio_centered                    7.55969    3.26732   2.314
## salary_centered:ratio_centered                    0.62689    0.68359   0.917
## expend_centered:takers_centered                   0.65574    0.50351   1.302
## salary_centered:takers_centered                   0.03846    0.09658   0.398
## takers_centered:ratio_centered                   -0.47780    0.18677  -2.558
## expend_centered:salary_centered:ratio_centered   -1.95337    0.73388  -2.662
## expend_centered:salary_centered:takers_centered  -0.06199    0.03667  -1.690
## expend_centered:takers_centered:ratio_centered    0.59046    0.22241   2.655
## salary_centered:takers_centered:ratio_centered   -0.04037    0.03948  -1.023
##                                                 Pr(>|t|)
## (Intercept)                                      < 2e-16 ***
## expend_centered                                  0.00388 **
## salary_centered                                  0.56967
## takers_centered                                 7.53e-12 ***
## expend_centered:salary_centered                  0.04041 *
## expend_centered:ratio_centered                   0.02650 *
## salary_centered:ratio_centered                   0.36521
## expend_centered:takers_centered                  0.20106
## salary_centered:takers_centered                  0.69281
## takers_centered:ratio_centered                   0.01488 *
## expend_centered:salary_centered:ratio_centered   0.01154 *
## expend_centered:salary_centered:takers_centered  0.09963 .
## expend_centered:takers_centered:ratio_centered   0.01174 *
## salary_centered:takers_centered:ratio_centered   0.31325
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.83 on 36 degrees of freedom
## Multiple R-squared:  0.8909, Adjusted R-squared:  0.8515
## F-statistic: 22.61 on 13 and 36 DF,  p-value: 1.8e-13
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.382 -14.028  -0.076  12.273  57.697
##
## Coefficients:
##                                                  Estimate Std. Error t value
```

```
## (Intercept)                                       965.61443    5.95546 162.139
## expend_centered                                    27.44439    8.85766   3.098
## salary_centered                                     1.01709    1.66705   0.610
## takers_centered                                    -2.93394    0.27335 -10.733
## expend_centered:salary_centered                    -1.88973    0.88130  -2.144
## expend_centered:ratio_centered                      7.45584    3.21965   2.316
## salary_centered:ratio_centered                      0.70294    0.64886   1.083
## expend_centered:takers_centered                     0.82594    0.26317   3.138
## takers_centered:ratio_centered                     -0.44345    0.16377  -2.708
## expend_centered:salary_centered:ratio_centered     -1.99035    0.71965  -2.766
## expend_centered:salary_centered:takers_centered   -0.06330    0.03611  -1.753
## expend_centered:takers_centered:ratio_centered      0.63172    0.19456   3.247
## salary_centered:takers_centered:ratio_centered    -0.05014    0.03059  -1.639
##                                                  Pr(>|t|)
## (Intercept)                                       < 2e-16 ***
## expend_centered                                   0.00371 **
## salary_centered                                   0.54552
## takers_centered                                  6.44e-13 ***
## expend_centered:salary_centered                   0.03865 *
## expend_centered:ratio_centered                    0.02622 *
## salary_centered:ratio_centered                    0.28566
## expend_centered:takers_centered                   0.00333 **
## takers_centered:ratio_centered                    0.01019 *
## expend_centered:salary_centered:ratio_centered    0.00881 **
## expend_centered:salary_centered:takers_centered   0.08784 .
## expend_centered:takers_centered:ratio_centered    0.00248 **
## salary_centered:takers_centered:ratio_centered    0.10966
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.5 on 37 degrees of freedom
## Multiple R-squared:  0.8904, Adjusted R-squared:  0.8549
## F-statistic: 25.05 on 12 and 37 DF,  p-value: 3.76e-14
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.063 -12.914   0.391  10.177  56.722
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                       965.23930    5.87450 164.310
## expend_centered                                    30.32294    7.43435   4.079
## takers_centered                                   -2.92144    0.27032 -10.807
## expend_centered:salary_centered                   -1.91937    0.87266  -2.199
## expend_centered:ratio_centered                     7.15027    3.15407   2.267
## salary_centered:ratio_centered                     0.82679    0.61118   1.353
## expend_centered:takers_centered                    0.81987    0.26080   3.144
## takers_centered:ratio_centered                    -0.44699    0.16231  -2.754
## expend_centered:salary_centered:ratio_centered    -1.93711    0.70842  -2.734
## expend_centered:takers_centered:salary_centered   -0.05586    0.03370  -1.658
```

```
## expend_centered:takers_centered:ratio_centered    0.62996    0.19292   3.265
## takers_centered:salary_centered:ratio_centered   -0.04894    0.03027  -1.617
##                                                 Pr(>|t|)
## (Intercept)                                      < 2e-16 ***
## expend_centered                                 0.000223 ***
## takers_centered                                 3.78e-13 ***
## expend_centered:salary_centered                 0.034002 *
## expend_centered:ratio_centered                  0.029164 *
## salary_centered:ratio_centered                  0.184118
## expend_centered:takers_centered                 0.003232 **
## takers_centered:ratio_centered                  0.008983 **
## expend_centered:salary_centered:ratio_centered  0.009439 **
## expend_centered:takers_centered:salary_centered 0.105647
## expend_centered:takers_centered:ratio_centered  0.002319 **
## takers_centered:salary_centered:ratio_centered  0.114208
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.27 on 38 degrees of freedom
## Multiple R-squared:  0.8893, Adjusted R-squared:  0.8573
## F-statistic: 27.76 on 11 and 38 DF,  p-value: 8.26e-15
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.241 -12.543  -1.084  14.522  55.707
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                     968.01829    5.56187 174.045
## expend_centered                                  30.02502    7.50975   3.998
## takers_centered                                  -2.87976    0.27140 -10.611
## expend_centered:salary_centered                  -1.67459    0.86273  -1.941
## expend_centered:ratio_centered                    6.96299    3.18439   2.187
## expend_centered:takers_centered                   0.70525    0.24926   2.829
## takers_centered:ratio_centered                   -0.35077    0.14744  -2.379
## expend_centered:salary_centered:ratio_centered   -1.78435    0.70676  -2.525
## expend_centered:takers_centered:salary_centered  -0.06643    0.03313  -2.005
## expend_centered:takers_centered:ratio_centered    0.53199    0.18071   2.944
## takers_centered:salary_centered:ratio_centered   -0.03815    0.02951  -1.293
##                                                 Pr(>|t|)
## (Intercept)                                      < 2e-16 ***
## expend_centered                                 0.000275 ***
## takers_centered                                 4.65e-13 ***
## expend_centered:salary_centered                 0.059507 .
## expend_centered:ratio_centered                  0.034836 *
## expend_centered:takers_centered                 0.007331 **
## takers_centered:ratio_centered                  0.022345 *
## expend_centered:salary_centered:ratio_centered  0.015760 *
## expend_centered:takers_centered:salary_centered 0.051913 .
## expend_centered:takers_centered:ratio_centered  0.005436 **
```

```
## takers_centered:salary_centered:ratio_centered  0.203743
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.57 on 39 degrees of freedom
## Multiple R-squared:  0.884,  Adjusted R-squared:  0.8542
## F-statistic: 29.72 on 10 and 39 DF,  p-value: 3.485e-15
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -78.237 -12.643  -1.255  15.539  55.951
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                    968.76922    5.57765 173.688
## expend_centered                                 29.44404    7.55890   3.895
## takers_centered                                 -3.05826    0.23559 -12.981
## expend_centered:salary_centered                 -1.77637    0.86631  -2.051
## expend_centered:ratio_centered                   8.63636    2.93368   2.944
## expend_centered:takers_centered                  0.75741    0.24803   3.054
## takers_centered:ratio_centered                  -0.41507    0.13995  -2.966
## expend_centered:salary_centered:ratio_centered  -1.86910    0.70959  -2.634
## expend_centered:takers_centered:salary_centered -0.05113    0.03120  -1.639
## expend_centered:takers_centered:ratio_centered   0.45508    0.17205   2.645
##                                                Pr(>|t|)
## (Intercept)                                     < 2e-16 ***
## expend_centered                                0.000364 ***
## takers_centered                                6.32e-16 ***
## expend_centered:salary_centered                0.046907 *
## expend_centered:ratio_centered                 0.005376 **
## expend_centered:takers_centered                0.004008 **
## takers_centered:ratio_centered                 0.005072 **
## expend_centered:salary_centered:ratio_centered 0.011941 *
## expend_centered:takers_centered:salary_centered 0.109113
## expend_centered:takers_centered:ratio_centered 0.011618 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.8 on 40 degrees of freedom
## Multiple R-squared:  0.879,  Adjusted R-squared:  0.8518
## F-statistic: 32.29 on 9 and 40 DF,  p-value: 1.31e-15
##
##
## Call:
## lm(formula = formula, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.356 -11.366  -1.428  17.702  66.164
##
```

```
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                     969.4662     5.6746 170.844
## expend_centered                                  22.0319     6.1795   3.565
## takers_centered                                  -3.1683     0.2304 -13.750
## expend_centered:salary_centered                  -2.3124     0.8185  -2.825
## expend_centered:ratio_centered                    8.0392     2.9702   2.707
## expend_centered:takers_centered                   0.7588     0.2531   2.998
## takers_centered:ratio_centered                   -0.3654     0.1394  -2.621
## expend_centered:salary_centered:ratio_centered  -1.2129     0.5977  -2.029
## expend_centered:takers_centered:ratio_centered   0.3446     0.1615   2.134
##                                                 Pr(>|t|)
## (Intercept)                                      < 2e-16 ***
## expend_centered                                 0.000941 ***
## takers_centered                                  < 2e-16 ***
## expend_centered:salary_centered                 0.007266 **
## expend_centered:ratio_centered                  0.009863 **
## expend_centered:takers_centered                 0.004596 **
## takers_centered:ratio_centered                  0.012237 *
## expend_centered:salary_centered:ratio_centered 0.048967 *
## expend_centered:takers_centered:ratio_centered 0.038910 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.39 on 41 degrees of freedom
## Multiple R-squared:  0.8709, Adjusted R-squared:  0.8457
## F-statistic: 34.57 on 8 and 41 DF,  p-value: 7.602e-16
```

```
vif(sat_redu_lm, type = "predictor")
```

```
## GVIFs computed for predictors
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## expend_centered   1.00000  8        1.000000
## takers_centered  67.04746  6        1.419706
## salary_centered 168.48755  4        1.898108
## ratio_centered    1.00000  8        1.000000
##                                                  Interacts With
## expend_centered salary_centered, ratio_centered, takers_centered
## takers_centered                  expend_centered, ratio_centered
## salary_centered                  salary_centered, ratio_centered
## ratio_centered  ratio_centered, takers_centered, salary_centered
##                 Other Predictors
## expend_centered               --
## takers_centered  salary_centered
## salary_centered  takers_centered
## ratio_centered                --
```

Given the high collinearity values of `takers_centered` and `salary_centered`, we aim to remove the lowest significant interaction term with either of this model.

```
model <- lm(total ~ expend_centered + takers_centered +
            expend_centered:salary_centered +
            expend_centered:ratio_centered +
            expend_centered:takers_centered +
            takers_centered:ratio_centered +
```

```
            expend_centered:salary_centered:ratio_centered +
            expend_centered:takers_centered:ratio_centered,
            data = sat_data)
sat_redu_lm_L1 <- update(model, . ~ . -expend_centered:takers_centered:ratio_centered)
summary(sat_redu_lm_L1)
```

```
##
## Call:
## lm(formula = total ~ expend_centered + takers_centered + expend_centered:salary_centered +
##     expend_centered:ratio_centered + expend_centered:takers_centered +
##     takers_centered:ratio_centered + expend_centered:salary_centered:ratio_centered,
##     data = sat_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.593 -16.229  -2.669  19.338  67.305
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                   965.11216    5.51427 175.021
## expend_centered                                14.74776    5.36415   2.749
## takers_centered                                -3.02581    0.22967 -13.175
## expend_centered:salary_centered                -1.73018    0.80363  -2.153
## expend_centered:ratio_centered                  3.21244    2.00432   1.603
## expend_centered:takers_centered                 0.64194    0.25731   2.495
## takers_centered:ratio_centered                 -0.21259    0.12456  -1.707
## expend_centered:salary_centered:ratio_centered  -0.01351    0.21156  -0.064
##                                               Pr(>|t|)
## (Intercept)                                    < 2e-16 ***
## expend_centered                                0.00877 **
## takers_centered                                < 2e-16 ***
## expend_centered:salary_centered                0.03711 *
## expend_centered:ratio_centered                 0.11648
## expend_centered:takers_centered                0.01662 *
## takers_centered:ratio_centered                 0.09525 .
## expend_centered:salary_centered:ratio_centered 0.94937
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.61 on 42 degrees of freedom
## Multiple R-squared:  0.8566, Adjusted R-squared:  0.8327
## F-statistic: 35.83 on 7 and 42 DF,  p-value: 1.002e-15
```

```
vif(sat_redu_lm_L1, type = "predictor") # SALARY_CENTERED COL. STILL HIGH
```

```
## GVIFs computed for predictors
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## expend_centered   1.00000  7        1.000000
## takers_centered   7.70665  5        1.226554
## salary_centered  14.71790  4        1.399525
## ratio_centered    1.00000  7        1.000000
##                                                Interacts With
## expend_centered salary_centered, ratio_centered, takers_centered
## takers_centered                 expend_centered, ratio_centered
```

```
## salary_centered                  salary_centered, ratio_centered
## ratio_centered  ratio_centered, takers_centered, salary_centered
##                 Other Predictors
## expend_centered             --
## takers_centered  salary_centered
## salary_centered  takers_centered
## ratio_centered              --
```

```
sat_redu_lm_L2 <- update(model, . ~ . -expend_centered:salary_centered:ratio_centered)
summary(sat_redu_lm_L2)
```

```
##
## Call:
## lm(formula = total ~ expend_centered + takers_centered + expend_centered:salary_centered +
##     expend_centered:ratio_centered + expend_centered:takers_centered +
##     takers_centered:ratio_centered + expend_centered:takers_centered:ratio_centered,
##     data = sat_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -81.301 -14.192  -5.683  18.042  67.250
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                   965.24076    5.47109 176.426
## expend_centered                                16.31906    5.70142   2.862
## takers_centered                                -2.99584    0.22198 -13.496
## expend_centered:salary_centered                -1.76291    0.80055  -2.202
## expend_centered:ratio_centered                  3.48653    2.01730   1.728
## expend_centered:takers_centered                 0.70391    0.26079   2.699
## takers_centered:ratio_centered                 -0.19021    0.11344  -1.677
## expend_centered:takers_centered:ratio_centered  0.03637    0.05689   0.639
##                                               Pr(>|t|)
## (Intercept)                                    < 2e-16 ***
## expend_centered                                0.00653 **
## takers_centered                                < 2e-16 ***
## expend_centered:salary_centered                0.03320 *
## expend_centered:ratio_centered                 0.09128 .
## expend_centered:takers_centered                0.00997 **
## takers_centered:ratio_centered                 0.10102
## expend_centered:takers_centered:ratio_centered 0.52614
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.46 on 42 degrees of freedom
## Multiple R-squared:  0.8579, Adjusted R-squared:  0.8342
## F-statistic: 36.23 on 7 and 42 DF,  p-value: 8.222e-16
```

```
vif(sat_redu_lm_L2, type = "predictor") #COL. ALL IN CONTROL
```

```
## GVIFs computed for predictors
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## expend_centered 1.000000  7        1.000000
## takers_centered 4.744047  6        1.138533
## salary_centered 6.902742  2        1.620897
```

```
## ratio_centered  4.744047  6        1.138533
##                                                 Interacts With
## expend_centered salary_centered, ratio_centered, takers_centered
## takers_centered                  expend_centered, ratio_centered
## salary_centered                                 salary_centered
## ratio_centered                   ratio_centered, takers_centered
##                                Other Predictors
## expend_centered                                --
## takers_centered                 salary_centered
## salary_centered takers_centered, ratio_centered
## ratio_centered                  salary_centered
```

```
Predicted_Simpl <- predict(sat_redu_lm_L2, sat_data)
ModelDataSimpl <- data.frame(obs = sat_data$total, pred = Predicted_Simpl)

Predicted_Full <- predict(sat_lm, sat_data)
ModelDataFull <- data.frame(obs = sat_data$total, pred = Predicted_Full)

rbind(defaultSummary(ModelDataSimpl), defaultSummary(ModelDataFull))
```

```
##          RMSE  Rsquared      MAE
## [1,] 27.91839 0.8579266 21.60825
## [2,] 31.02383 0.8245623 23.99230
```

```
cat("[1,] Simplified Model: Adjusted R-squared =", summary(sat_redu_lm_L2)$adj.r.squared, "AIC =", AIC(s
```

```
## [1,] Simplified Model: Adjusted R-squared = 0.8342477 AIC = 492.8224
```

```
cat("[2,] Original Model: Adjusted R-squared =", summary(sat_lm)$adj.r.squared, "AIC =", AIC(sat_lm), "
```

```
## [2,] Original Model: Adjusted R-squared = 0.8089679 AIC = 497.3694
```

When utilizing the Simplified Model with further modification to reduce multicollinearity we see a reduction in the error metrics (RMSE, MAE), a reduction in multicollinearity (recall: the vif values were near 9 of expend and salary in the original model), an increase in Adjusted R^2 ($0.8342 > 0.8090$) and a decrease in AIC when compared to the Original Model ($492.82 < 497.37$). This concludes our analysis of the structure of the relationship between the predictor variable(s) and the response variable wherein we leveraged ANOVA to compare the variation in total explained by each predictor variable in all four of the linear models we created, deciphered the component and residual plots to expose the curvilinear nature of the variables, explored and then reduced the multicollinearity present among the predictor variables and divulged the interaction effects that could be playing a role in weakening other variables (i.e., salary shows it is significant in reducing variation, but it is also highly collinear with takers and expend and removing it improves the performance of the model when it is leveraged in an interaction term with expend). This extensive analysis shows how much is hidden when we build a linear model with all predictors, and it is not until we dive deep into the data can we truly identify what the dat and variables intend to tell us.