Problem 1

a) The error terms e_i are normally distributed with mean 0 and variance σ^2.

$$X^T = [1 \dots 1]_{1*n}, \qquad C = [1], \qquad \beta = \mu$$

$$\hat{V} = \hat{\sigma}^2 (X^T X)^{-1} = s^2 * \frac{1}{n} = \frac{s^2}{n}$$

Therefore:

$$\frac{C^T \hat{\beta} - C^T \beta}{\sqrt{C^T \hat{V} C}} = \frac{\bar{y} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

And the $(100 - \alpha)\%$ confidence interval would be:

$$\bar{y} \pm t_{\frac{\alpha}{2}, n-1} * \frac{s}{\sqrt{n}}$$

b) $X_{new} = [1 \dots 1]$

$$X_{new}^T \hat{\beta} \pm t_{\frac{\alpha}{2}, n-1} \sqrt{X_{new}^T \hat{V} X_{new} + \hat{\sigma}^2} \rightarrow \bar{y} \pm t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{s^2}{n} + s^2} \rightarrow \bar{y} \pm t_{\frac{\alpha}{2}, n-1} \sqrt{s^2 \left(1 + \frac{1}{n}\right)}$$

c) $H = X(X^T X)^{-1} X^T = X * \frac{1}{n} * X^T = \frac{1}{n} * XX^T = \frac{1}{n} * \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{n*n}$

This implies that every value of H is 1/n. Therefore, every $h_{ii} = \frac{1}{n}$. Every observation has the same leverage of 1/n.

Problem 2

Constant Variance

Based on Figure 1 below, there does not appear to be any visible distortion. In Figure 1(a) there is one value with a much lower residual than the others, but there is no overall pattern, and the residual falls into normal pattern in Figure 1(c).

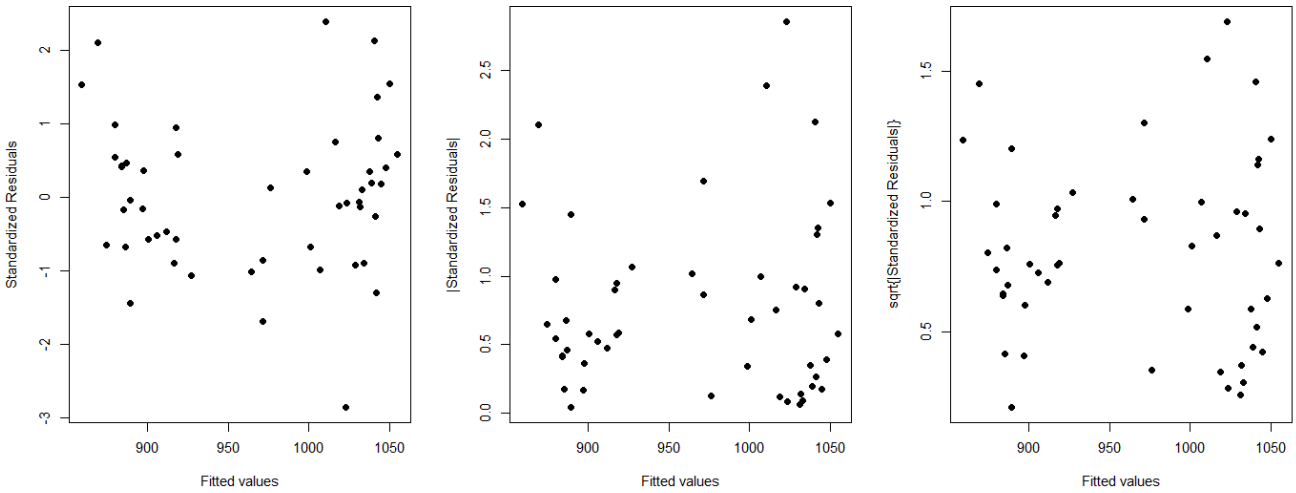

Figure 1: Fitted Values vs. Standardized Residuals

We then plotted the residuals vs the individual variables as seen in Figure 2 below. The same residual that showed in Figure 1(a) appears to be showing constantly in the four graphs in Figure 2. The first and third graph show no sign of visual distortion or pattern. However, graph number 2 (residual vs. ratio) has two points that appear to be outliers in the x-axis. These could be high leverage points, but otherwise the graph shows no obvious pattern. Graph number 4, shows a slight pattern of larger residuals for lower values and higher values of "takers" with residuals with less magnitude for moderate values of "takers". This might be something that needs to be investigated further.

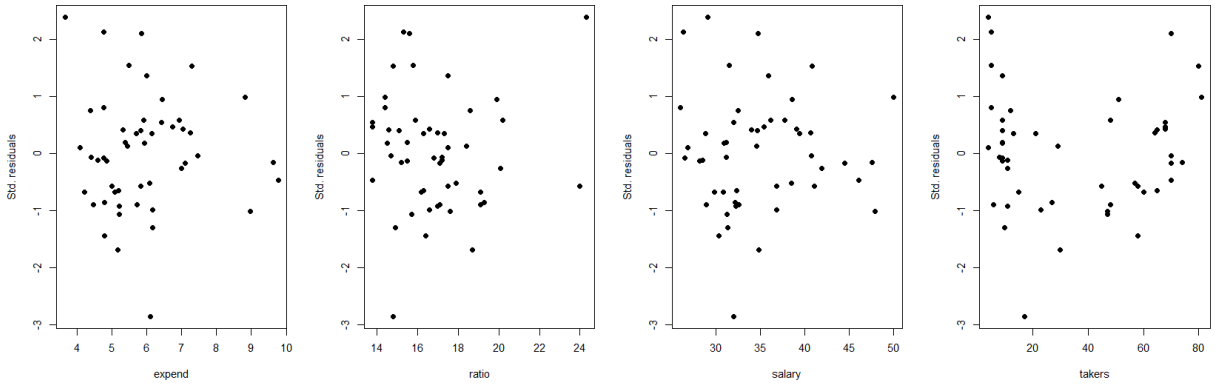

Figure 2: Residuals vs Individual Variables

Normality of the Residuals

We found that the residuals were approximately Normal. The Shapiro-Wilk normality test returned a p-value of 0.5607 indicating that there is not significant evidence to claim that the residuals are not Normal. The histogram in Figure 3 below shows that the residuals fit well to a normal curve, and the Normal-QQ Plot in Figure 3 below shows that residuals do follow closely to a Normal distribution.

```
        Shapiro-Wilk normality test

data:  rstandard(out)
W = 0.98021, p-value = 0.5607
```

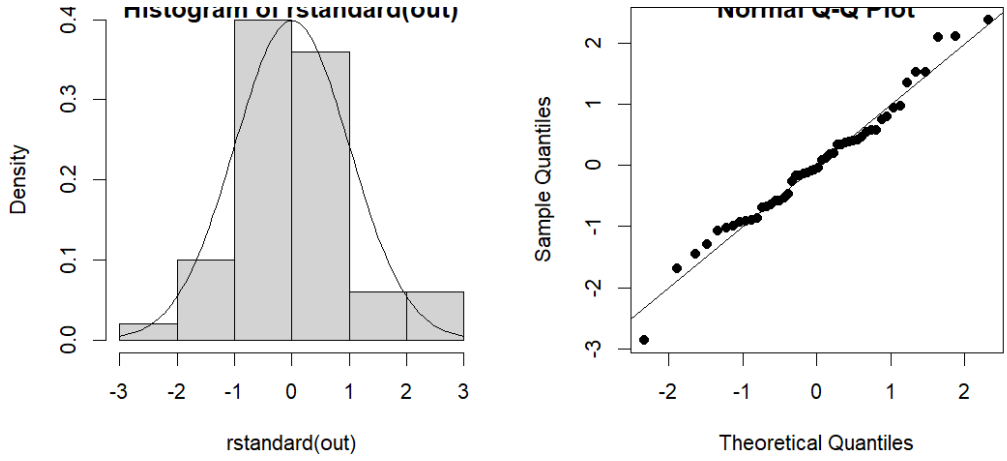

Figure 3: Checking Normality of Residuals

Leverage

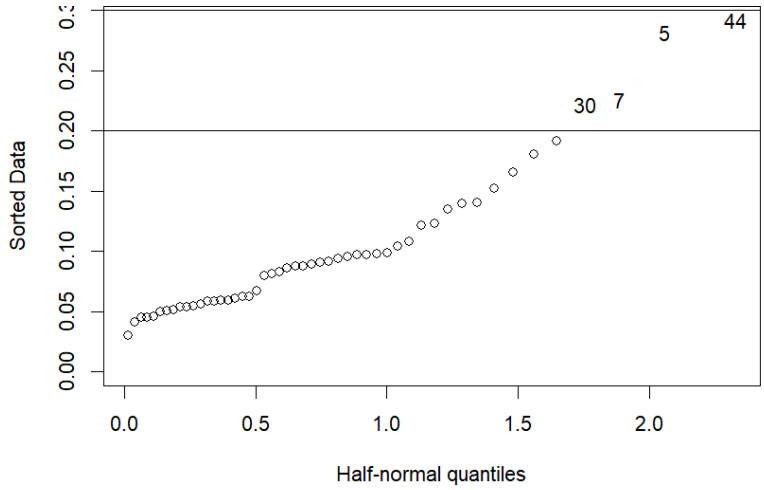

Figure 4:Checking for High Leverage Values

We found that there were no values that were considered high leverage values (lv > 3*mean(lv)), but there were four values that were considered a moderately high leverage value (lv > 2*mean(lv)).

```
> which(lv >= 2*mean(lv))
 California Connecticut  New Jersey        Utah
          5           7          30          44
```

Serial Correlation

There does not appear to be any correlation between the index and the residuals as seen in the figure below. When running the Durbin-Watson test, we recorded a p-value of 0.9459, meaning that there is no significant evidence that there is correlation between the index and the residuals.

```
        Durbin-Watson test

data:  total ~ . - verbal - math
DW = 2.4525, p-value = 0.9459
alternative hypothesis: true autocorrelation is greater than 0
```

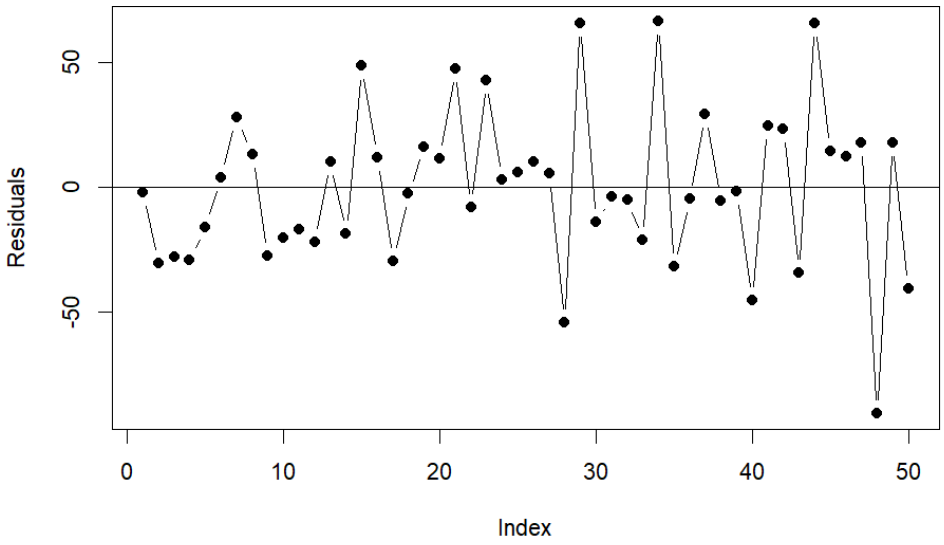

Figure 5: Residuals vs. Index

Problem 3

a) The prediction interval is wider than the confidence interval because of an extra σ^2 term in the "margin of error".

Confidence Interval: $C^T\beta \pm t_{\frac{\alpha}{2},n-rank(X)}\sqrt{C^T\hat{V}C}$

Prediction Interval: $X_{new}^T\beta \pm t_{\frac{\alpha}{2},n-rank(X)}\sqrt{X_{new}^T\hat{V}X_{new} + \sigma^2}$

This extra σ^2 is extra variability that we do not have control over. We cannot control the variability in the errors, making the prediction interval wider than the confidence interval.

b) The misspecification of the systematic component may lead to nonlinear patterns in our residual plot when our data does not follow a linear pattern. When trying to fit a linear model to data that does not follow a linear pattern, then the residual plot will show some kind of pattern as seen below. We would need to perform some kind of transformation on the data in order to achieve linearity, and therefore not see a pattern in the residual plot.

```r
# Simulate the data
n <- 1000
X <- runif(n, 0, 10)
e <- rnorm(n, mean = 0, sd = 1)
Y_true <- X^2 + e
```

Figure 6: Simulating Data

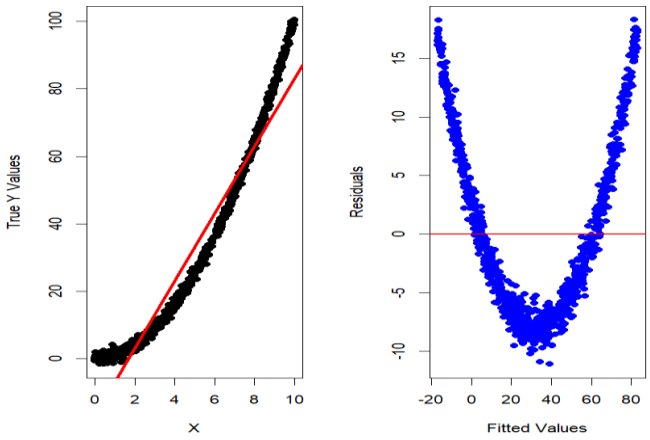

Figure 7: Plot of Data and Residual Plot

```r
#Achieving Linearity
out_linearized <- lm(sqrt(abs(Y_true)) ~ X)
plot(X, out_linearized$fitted.values,
     ylab = "Transformed Y Value")
abline(out_linearized, col = "red", lwd = 2)
```

Figure 8: Transforming Data to Achieve Linearity

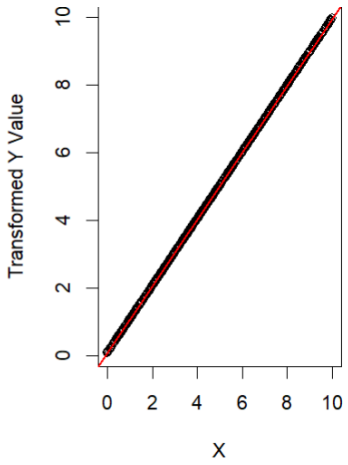

Figure 9: Plot of Transformed Data

Appendix: R Code

```
##############################
# Author: Evan Whitfield
# Date: 2-15-25
# Purpose: To answer HW4
##############################

##############################
#
# Problem 2
#
##############################

library(faraway)

head(sat)

#fitting the data
out <- lm(total ~ . - verbal - math, data = sat)
res <- resid(out)
fitted <- fitted(out)
res_std <- rstandard(out)

#plotting the residuals
par(mfrow = c(1,3), mar = c(4,4,0,2))
plot(fitted, res_std,
    xlab = "Fitted values",
    ylab = "Standardized Residuals", pch =19)
```

```
plot(fitted, abs(res_std),
    xlab = "Fitted values",
    ylab = "|Standardized Residuals|",
    pch =19)
plot(fitted, sqrt(abs(res_std)),
    xlab = "Fitted values",
    ylab = "sqrt{|Standardized Residuals|}",
    pch =19)

#plotting the residuals vs variables
par(mfrow=c(1,4), mar = c(4,4,0,2))
names <- c("expend", "ratio", "salary", "takers")
for(nm in names){
  plot(sat[[nm]], res_std, xlab = nm, ylab = "Std. residuals", pch = 19)
}



#checking normality of residuals
par(mfrow = c(1,2))
hist(rstandard(out), probability = TRUE)
curve(dnorm(x), to = 3, from = -3, add = TRUE)
qqnorm(rstandard(out), pch = 19)
abline(a = 0, b = 1)
shapiro.test(rstandard(out))

par(mfrow = c(1,1))
# leverage
lv <- hatvalues(out)
# with highest leverage
```

```
halfnorm(lv, nlab = 4)
# 3 times average of leverages
abline(h = 3*mean(lv))
# 2 times average of leverages
abline(h = 2*mean(lv))
# Identify high leverage points
which(lv >= 3*mean(lv))
# Identify moderately high leverage points
which(lv >= 2*mean(lv))



#Serial Correlation
library(lmtest)
dwtest(total ~ . - verbal - math, data = sat)

plot(resid(out), ylab = "Residuals", pch=19, type="b")
abline(h=0)




#####################################
#
# Problem 3
#
#####################################


# Simulate the data
n <- 1000
```

```
X <- runif(n, 0, 10)
e <- rnorm(n, mean = 0, sd = 1)
Y_true <- X^2 + e

# Fit a linear model
out_linear <- lm(Y_true ~ X)

# Get the fitted values and residuals
Y_predicted <- out_linear$fitted.values
residuals <- out_linear$residuals

# Plot data
plot(X, Y_true, pch = 19,
    xlab = "X", ylab = "True Y Values")
abline(out_linear, col = "red", lwd = 3)

#Plot residual plot
plot(Y_predicted, residuals,
    xlab = "Fitted Values", ylab = "Residuals",
    pch = 19, col = "blue")
abline(h = 0, col = "red", lwd = 2)

#Achieving Linearity
out_linearized <- lm(sqrt(abs(Y_true)) ~ X)
plot(X, out_linearized$fitted.values,
    ylab = "Transformed Y Value")
abline(out_linearized, col = "red", lwd = 2)
```