HW3 - Josh Janda

Josh Janda

08 March, 2020

Problem 1.

Using the sat dataset, 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.

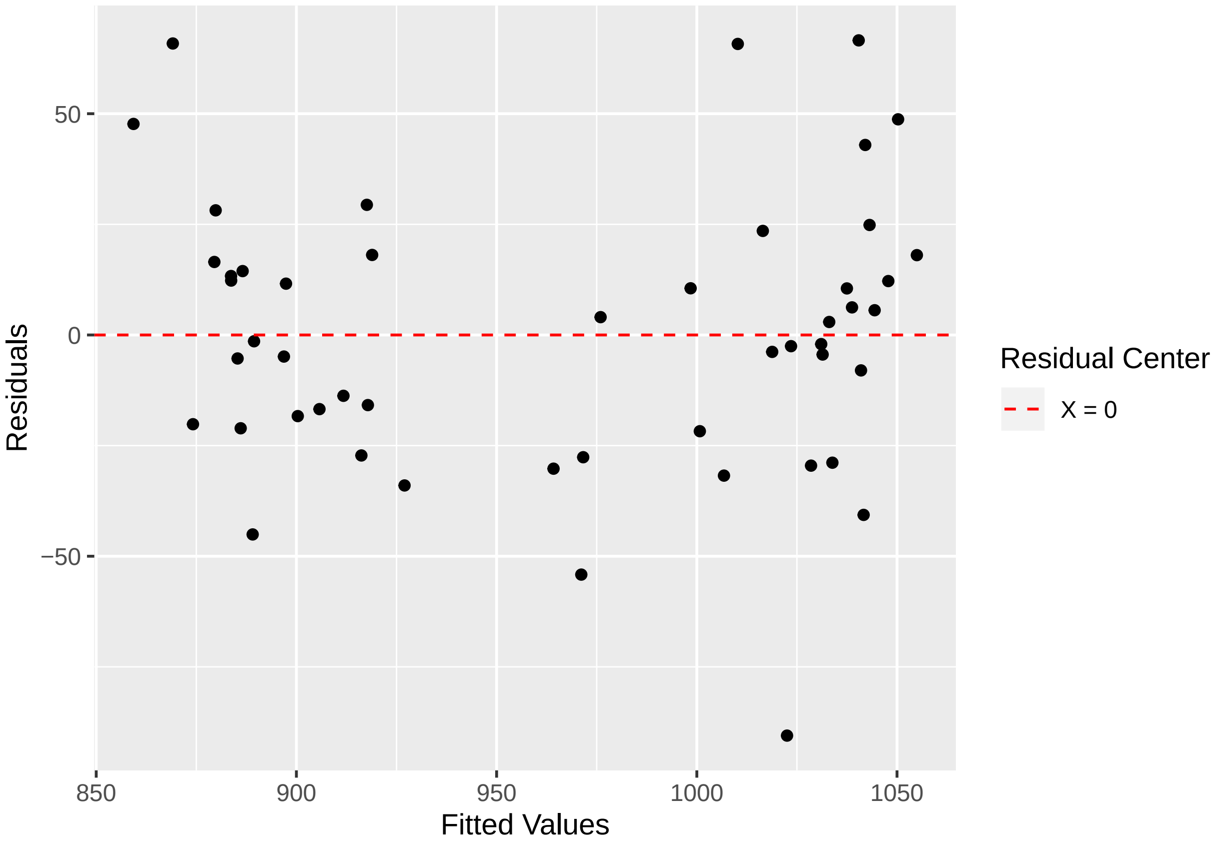
```r
library(faraway)
library(tidyverse)
library(lmtest)
library(knitr)
library(olsrr)
`%notin%` <- Negate(`%in%`)
```

```r
score_mod = lm(total ~ expend + salary + ratio + takers, data = sat)
```

A.

Check the constant variance assumption for the errors.

```r
ggplot() +
  geom_point(aes(x = score_mod$fitted.values,
                 y = score_mod$residuals)) +
  geom_hline(aes(yintercept = 0, linetype = "X = 0"), color = 'red') +
  labs(linetype = "Residual Center",
       x = "Fitted Values", y = "Residuals") +
  scale_linetype_manual(values = c(2))
```



Looking at the plot above, it does not appear that this model violates the constant variance assumptions for the residuals.

To further confirm homoskedasticity, I will utilize the Breusch-Pagan test. The null hypothesis for this test is that the model follows the constant variance assumption and is homoskedastic. The alternate hypothesis for this test is that the model violates the constant variance assumption and is heteroskedastic. The null hypothesis follows a χ^2_{p-1} distribution.

```r
score_mod_bpt = bptest(score_mod)
score_mod_bpt
```

```
##
##  studentized Breusch-Pagan test
##
## data:  score_mod
## BP = 2.1587, df = 4, p-value = 0.7066
```

Looking at the output above for the Breusch-Pagan test, we see:

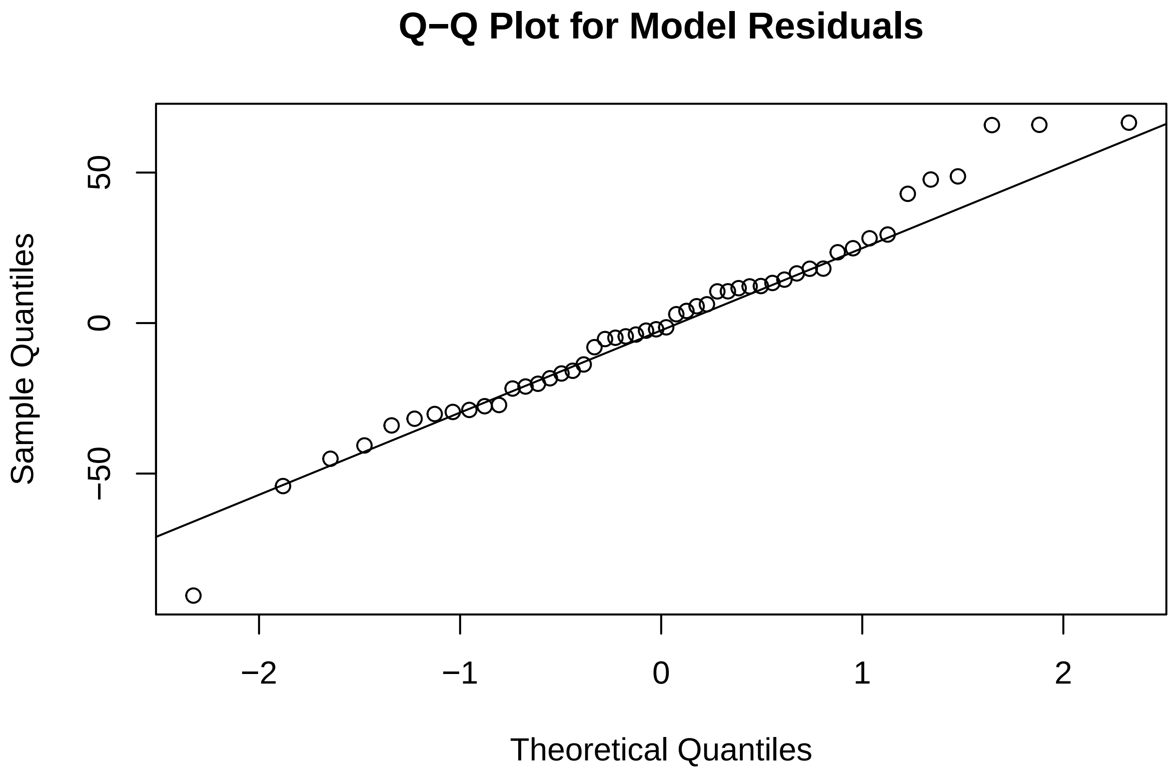
- BP Test Statistic of 2.1587006
- P-Value of 0.706597

Therefore, we fail to reject H_0 that the model follows the constant variance assumption and conclude that the model is homoskedastic.

B.

Check the normality assumption.

```
qqnorm(score_mod$residuals, main = 'Q-Q Plot for Model Residuals')
qqline(score_mod$residuals)
```



Looking at the Q-Q plot above, it appears that the residuals follow a normal distribution. The residuals follow a linear trend, which helps further demonstrate normality. It should be noted that samples in low/high sample quantiles do not fall on the linear trend line suggesting potential outliers.

To further confirm normality, I will utilize the Shapiro-Wilk test. The null hypothesis for this test is that the data is normally distributed. The alternate hypothesis is that the data is not normally distributed.

```
score_mod_swt = shapiro.test(score_mod$residuals)
score_mod_swt
```

```
##
##  Shapiro-Wilk normality test
##
## data:  score_mod$residuals
## W = 0.97691, p-value = 0.4304
```

Looking at the output above for the Shapiro-Wilk normality test, we see:

- W Test Statistic of 0.9769142

- P-Value of 0.4303922

Therefore, we fail to reject H_0 that the residuals follow a normal distribution and conclude they are normally distributed.

C.

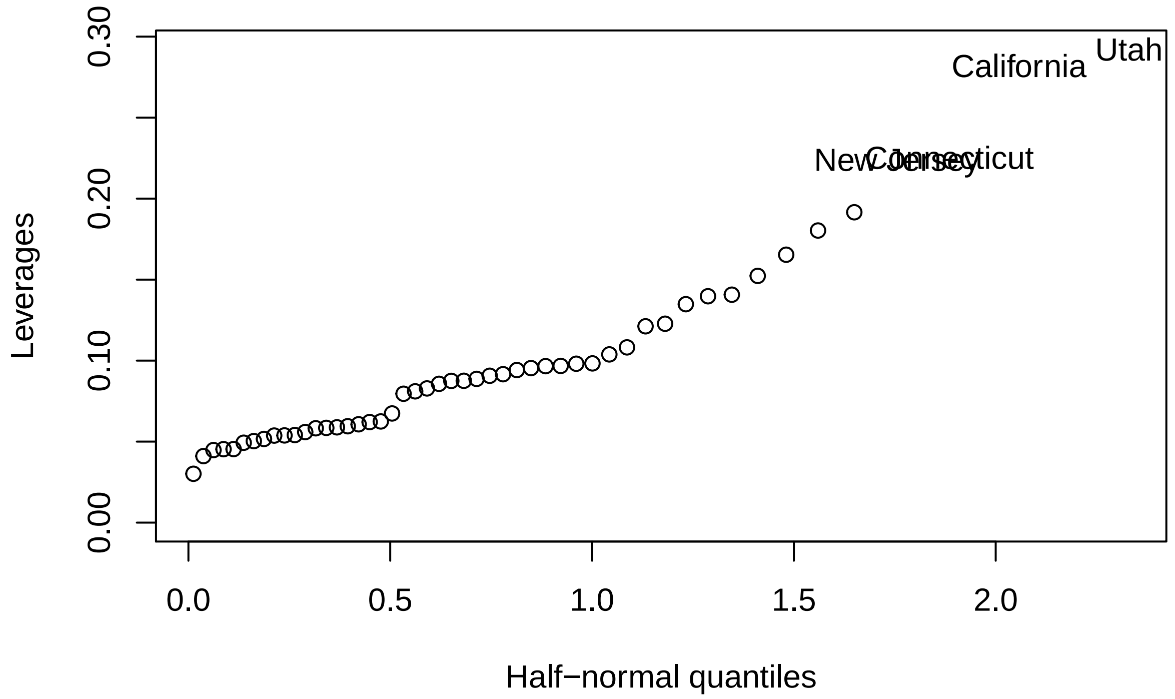
Check for large leverage points.

```r
p = length(coef(score_mod))
n = nrow(sat)
leverage = influence(score_mod)$hat
high_lev_pts = leverage[leverage > 2*(p/n)]
kable(high_lev_pts, col.names = "Leverage")
```

	Leverage
California	0.2821179
Connecticut	0.2254519
New Jersey	0.2220978
Utah	0.2921128

The output above displays high/large leverage points in our model. "Leverage" quantifies how far a data point is from the center of the whole sample. A data point with "high" leverage is considered any point with leverage larger than $2 * \frac{p}{n}$ where $p =$ number of predictors and $n =$ number of observations in the data. With that said, we see that states **California**, **Connecticut**, **New Jersey**, and **Utah** all have high leverage points.

```r
halfnorm(leverage, nlab = 4,
         labs = row.names(sat), ylab = 'Leverages')
```



Looking at the output above, we can further observe that these points have a much higher distance between data points compared to the rest of the data. These observations can be considered to have a larger influence on the model fitting.

D.

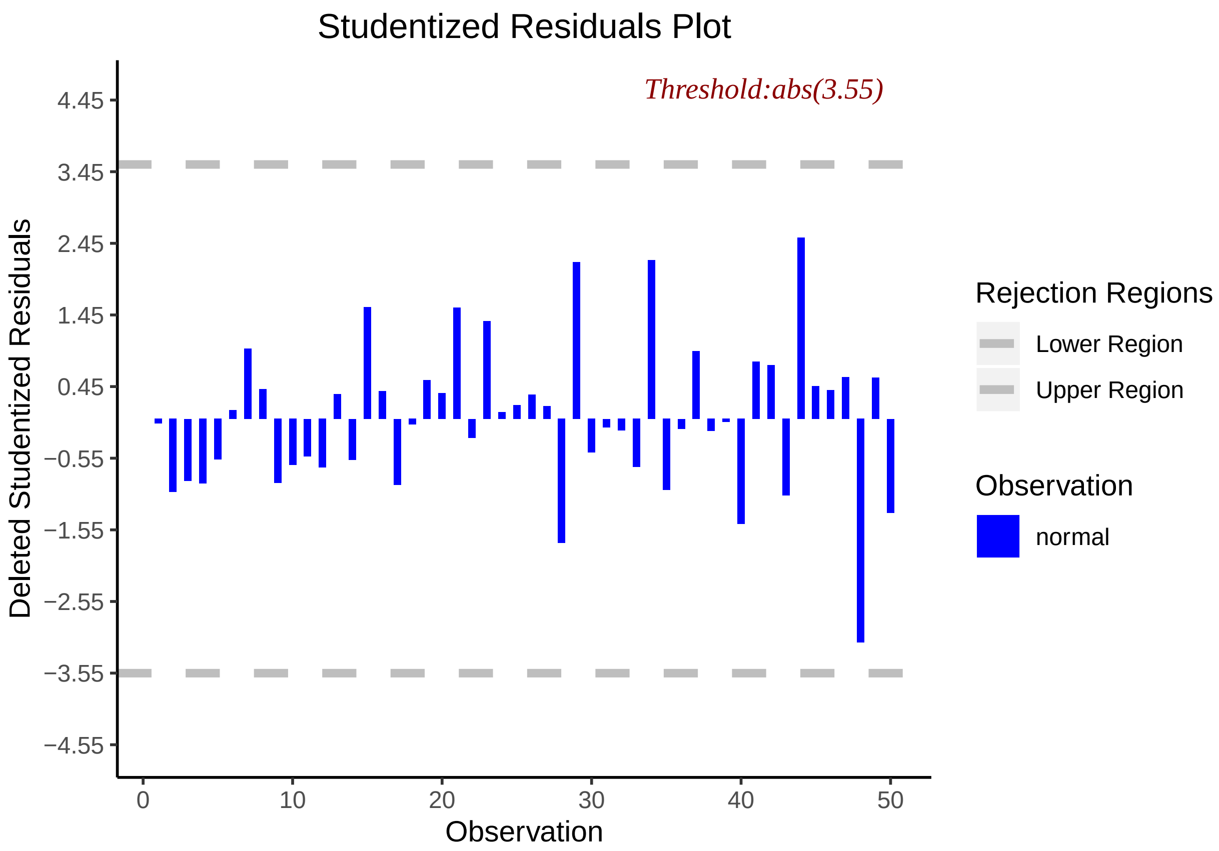
Check for outliers.

```
plot_outliers = function(model, dset) {

  rej_reg = qt(0.05 / (nrow(dset) * 2),
                     nrow(dset) - ncol(dset) - 3)
  d = data.frame(obs = 1:nrow(dset)) %>% mutate(dsr = rstudent(model),
                               color = ifelse(abs(dsr) < abs(rej_reg),
                                              'normal', 'outlier'),
                               fct_color = color,
                               txt = ifelse(color != 'normal',
                                            obs, NA))
  f = d[d$color == "outlier", c("obs", "dsr")]
  ymin = min(d$dsr) - 1
  ymax = max(d$dsr) + 1
  if (abs(ymin) < abs(ymax)) {
    ymin = -ymax
  }
  if (ymin < rej_reg) {
    ymin = rej_reg - 1
  }
  if (ymax < abs(rej_reg)) {
    ymax = abs(rej_reg) + 1
  }
  #generate outlier plot using above rules
  ggplot(d, aes(x = obs, y = dsr, label = txt)) +
  geom_bar(width = 0.5,
        stat = "identity", aes(fill = fct_color)) +
  scale_fill_manual(values = c("blue",
        "red")) + xlab("Observation") +
  ylab("Deleted Studentized Residuals") +
        labs(fill = "Observation") +
  ggtitle("Studentized Residuals Plot") +
  geom_hline(aes(yintercept = rej_reg, color = 'Lower Region'),
             linetype = 2, lwd = 1.5) +
  geom_hline(aes(yintercept = abs(rej_reg), color = 'Upper Region'),
             linetype = 2, lwd = 1.5) +
  geom_text(hjust = -0.2, nudge_x = 0.05,
        size = 2, na.rm = TRUE) +
  annotate("text", x = Inf, y = Inf,
        hjust = 1.2, vjust = 2, family = "serif", fontface = "italic",
        colour = "darkred", label = paste0("Threshold:",
                                           "abs(", round(abs(rej_reg), 2), ")")) +
  scale_y_continuous(limits = c(ymin, ymax),
                     breaks = round(seq(ymin, ymax, by = 1), 2)) +
  theme(plot.title = element_text(hjust = 0.5),
        axis.line = element_line(color = 'black'),
        panel.background =  element_blank()) +
  scale_color_manual(values = c('grey', 'grey')) +
  labs(color = 'Rejection Regions')

}

plot_outliers(score_mod, sat)
```



To check for outliers, we will be looking at studentized residuals. Under H_0, we can use the t-test to test whether the ith observation is an outlier or not. H_0 follows a t distribution with n-p-1 degrees of freedom.

Using the plot above, we see that no observations can be considered to be outliers. The red lines represent the rejection region, where any observations that have studentized residuals that go past either rejection region are considered outliers.

E.

Check for influential points.

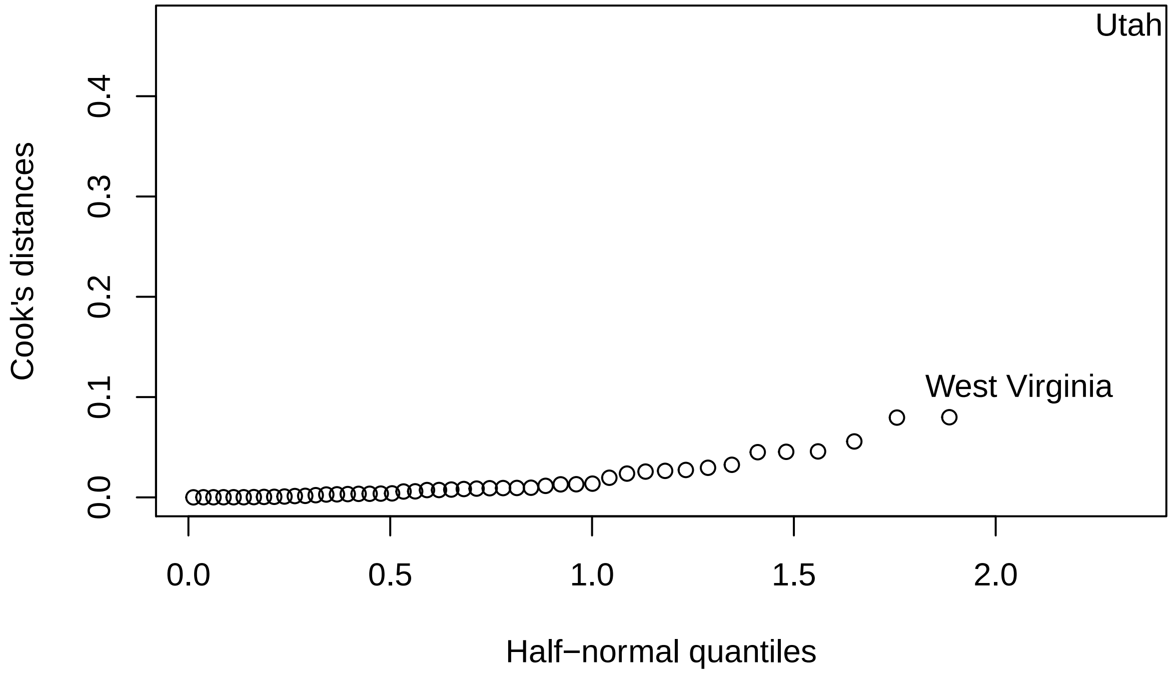
```
sat_cook = cooks.distance(score_mod)
high_influence = sat_cook[sat_cook >= 1]
length(high_influence)
```

```
## [1] 0
```

In order to check for influential points, I will utiize Cook's Distance. Cook's Distance measures how much specific observations influence a model's analysis. A highly influential point is considered any observation with a Cook's Distance greater than or equal to 1. High influential points indicate that the point is either an outlier, a high leverage point, or both.

Using the output above, we see that no observations have a Cook's Distance greater than or equal to one. This is understandable as before we have looked for outliers/high influence points, and found that we have no outliers and only four high leverage points where their leverages were actually not that large compared to other leverages.

```
halfnorm(sat_cook, labs=row.names(sat), ylab="Cook's distances")
```



Looking at the plot above, we see that the point with the largest Cook's distance is Utah. However, this observation still has a pretty low distance of ~.50 so cannot be considered highly influential. This plot confirms my statement before of having no high influence points.

F.

Check the structure of the relationship between the predictors and the response.

```
res_score_mod_m_expend = update(score_mod, ~. - expend)$res
res_expend = lm(expend ~ salary + ratio + takers, data = sat)$res
expend_lm = lm(res_score_mod_m_expend ~ res_expend)

res_score_mod_m_salary = update(score_mod, ~. - salary)$res
res_salary = lm(salary ~ expend + ratio + takers, data = sat)$res
salary_lm = lm(res_score_mod_m_salary ~ res_salary)

res_score_mod_m_ratio = update(score_mod, ~. - ratio)$res
res_ratio = lm(ratio ~ salary + expend + takers, data = sat)$res
ratio_lm = lm(res_score_mod_m_ratio ~ res_ratio)

res_score_mod_m_takers = update(score_mod, ~. - takers)$res
res_takers = lm(takers ~ salary + ratio + expend, data = sat)$res
takers_lm = lm(res_score_mod_m_takers ~ res_takers)
```

```r
compare_coefs = data.frame(Res.Coef = c(coef(expend_lm)[2],
                                        coef(salary_lm)[2],
                                        coef(ratio_lm)[2],
                                        coef(takers_lm)[2]),
                           Orig.Coef = coef(score_mod)[-1])
rownames(compare_coefs) = c('Expend', 'Salary', 'Ratio', 'Takers')
kable(compare_coefs)
```

	Res.Coef	Orig.Coef
Expend	4.462594	4.462594
Salary	1.637917	1.637917
Ratio	-3.624232	-3.624232
Takers	-2.904481	-2.904481

For checking the structure of the relationship between the predictors and the response, we want to know the relationship between the response (Y) and the predictor X_k after the effect of the other predictors has been removed. To remove these effects of other predictors, I must do the following procedure for each predictor:

1. Create model 1 $Y \sim X_1 + \cdots + X_{i-1} + \cdots + X_{i+1} + \cdots$
2. Create model 2 $X_i \sim X_1 + \cdots + X_{i-1} + \cdots + X_{i+1} + \cdots$
3. Obtain residuals from model 1, r_y
4. Obtain residuals from model 2, r_k^X
5. Create model 3 $r_y \sim r_k^X$

If there is a valid relationship between the predictor and the response, we should see the same slope in model 3 compared to the slope in the original model of the given predictor.

Looking at the table output above, we see that is the case and that all predictors have identical coefficients in the residuals model compared to the original model. This confirms that our model structure is valid.

Furthermore, if we plot r_y vs. r_k^X we should see points randomly scattered around a line through the origin with slope $\hat{\beta}_k$ if the model is valid.

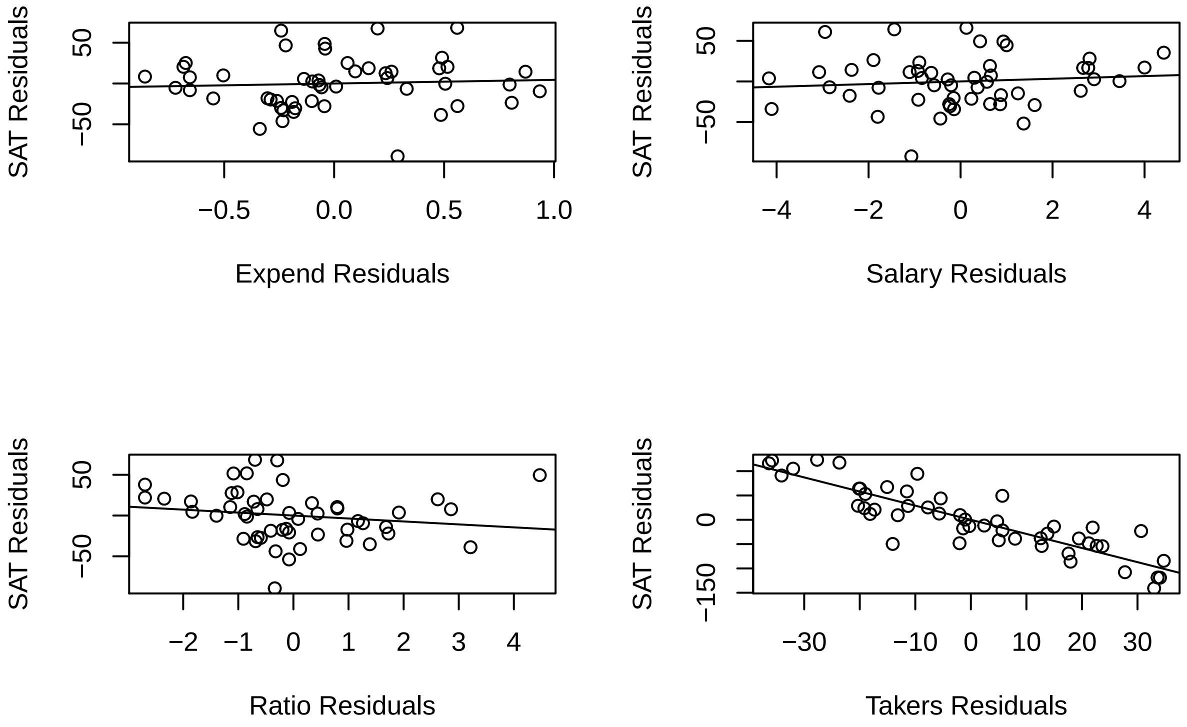
```r
par(mfrow = c(2, 2))

plot(res_expend, res_score_mod_m_expend,
     xlab = 'Expend Residuals', ylab = 'SAT Residuals')
abline(expend_lm)

plot(res_salary, res_score_mod_m_salary,
     xlab = 'Salary Residuals', ylab = 'SAT Residuals')
abline(salary_lm)

plot(res_ratio, res_score_mod_m_ratio,
     xlab = 'Ratio Residuals', ylab = 'SAT Residuals')
abline(ratio_lm)

plot(res_takers, res_score_mod_m_takers,
     xlab = 'Takers Residuals', ylab = 'SAT Residuals')
abline(takers_lm)
```



Looking at the plot above, we see that for all four predictors we have points randomly scattered around a line through the origin with their associated slope coefficient. Therefore, we can further confirm that our model is valid.

It should be noted that there seems to be some slightly strong negative correlation in the **Takers** residuals. This tells me that there is a strong relationship between the percentage of all eligible students taking the SAT and the average total score on the SAT. This could possibly be due to a higher percentage of eligible students being due to more smarter students, which will raise the average total score.

Overall, for this model there does not seem to be any obvious assumption violations and no real observations that can be considered worth removal consideration. I do not believe there are any real possible improvements or corrections to this model.

Problem 2.

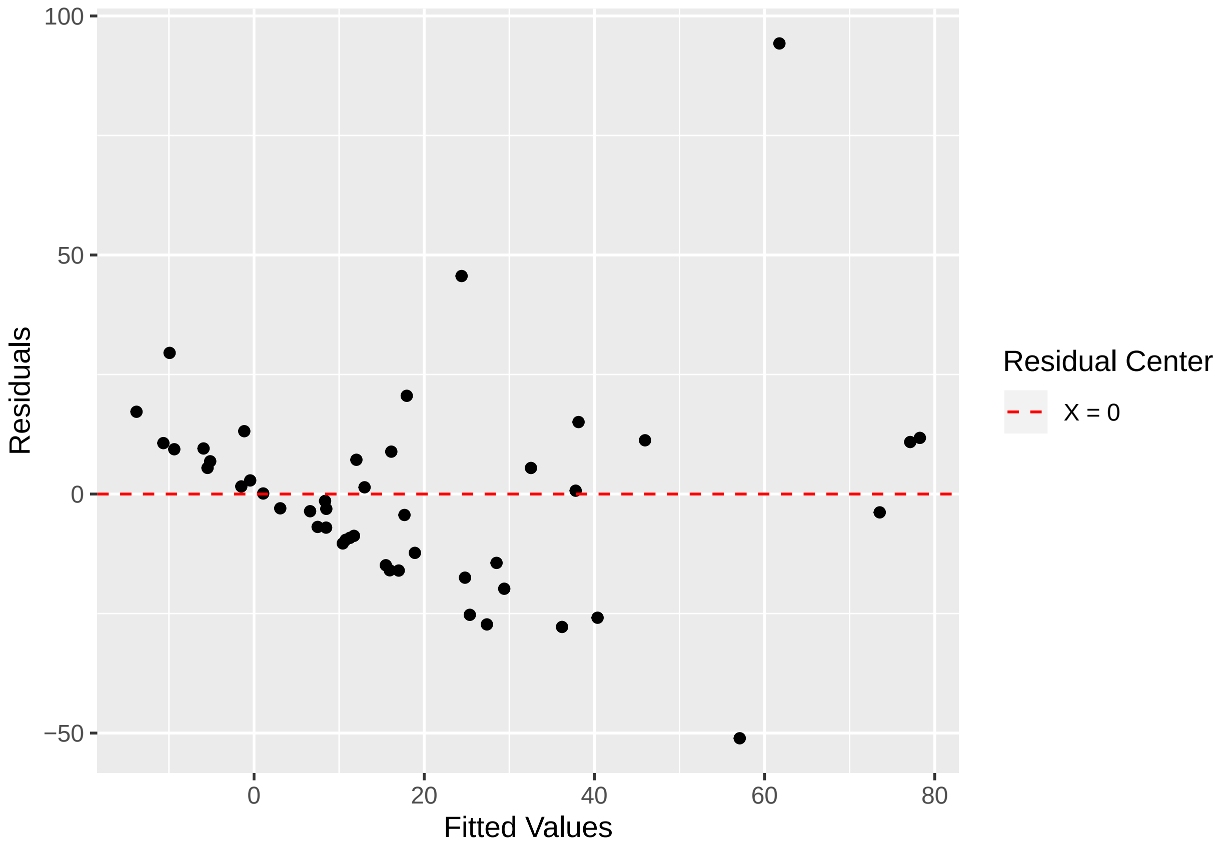
Using the teengamb dataset, fit a model with gamble as a response and the other variables as predictors. Answer the questions posed in the previous question.

```r
teengamb_mod = lm(gamble ~ ., data = teengamb)
```

A.

Check the constant variance assumption for the errors.

```r
ggplot() +
  geom_point(aes(x = teengamb_mod$fitted.values,
                 y = teengamb_mod$residuals)) +
  geom_hline(aes(yintercept = 0, linetype = "X = 0"), color = 'red') +
  labs(linetype = "Residual Center",
       x = "Fitted Values", y = "Residuals") +
  scale_linetype_manual(values = c(2))
```



Looking at the plot above, it does appear that this model violates the constant variance assumptions for the residuals.

To further confirm homoskedasticity, I will utilize the Breusch-Pagan test. The null hypothesis for this test is that the model follows the constant variance assumption and is homoskedastic. The alternate hypothesis for this test is that the model violates the constant variance assumption and is heteroskedastic. The null hypothesis follows a χ^2_{p-1} distribution.

```r
teengamb_mod_bpt = bptest(teengamb_mod)
teengamb_mod_bpt
```

```
##
##  studentized Breusch-Pagan test
##
## data:  teengamb_mod
## BP = 6.4288, df = 4, p-value = 0.1693
```

Looking at the output above for the Breusch-Pagan test, we see:

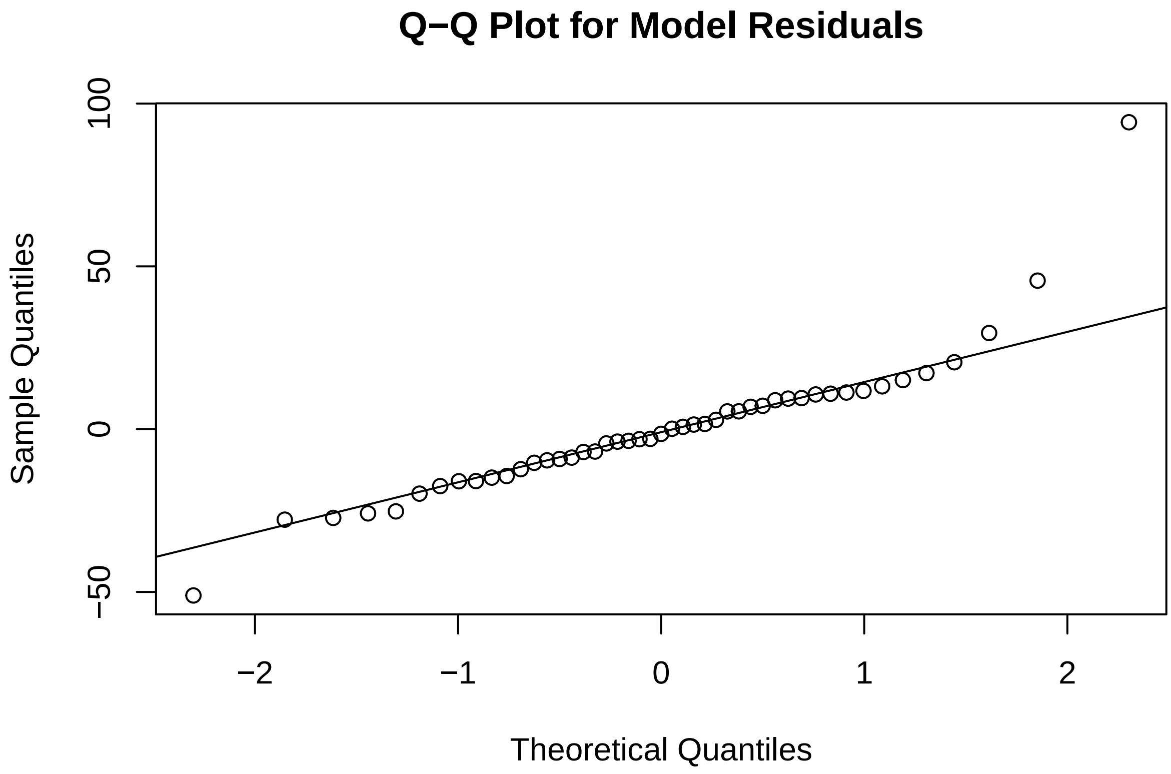
- BP Test Statistic of 6.4287646
- P-Value of 0.1693345

Therefore, we fail to reject H_0 that the model follows the constant variance assumption and conclude that the model is homoskedastic.

B.

Check the normality assumption.

```
qqnorm(teengamb_mod$residuals, main = 'Q-Q Plot for Model Residuals')
qqline(teengamb_mod$residuals)
```



Looking at the Q-Q plot above, it appears that the residuals follow a normal distribution. The residuals follow a linear trend, which helps further demonstrate normality. It should be noted that samples in low/high sample quantiles do not fall on the linear trend line suggesting potential outliers.

To further confirm normality, I will utilize the Shapiro-Wilk test. The null hypothesis for this test is that the data is normally distributed. The alternate hypothesis is that the data is not normally distributed.

```
teengamb_mod_swt = shapiro.test(teengamb_mod$residuals)
teengamb_mod_swt
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  teengamb_mod$residuals
## W = 0.86839, p-value = 8.16e-05
```

Looking at the output above for the Shapiro-Wilk normality test, we see:

- W Test Statistic of 0.8683906
- P-Value of 8.1604119×10^{-5}

Therefore, we reject H_0 that the residuals follow a normal distribution and conclude they are not normally distributed. I believe this is due to the extreme outlier that is seen in the Q-Q plot, where the sample quantile was seen to be 100 for an observation and all other observations were between sample quantiles of -50 to 50. This observation is highly influencing the models residuals causing non-normality.

C.

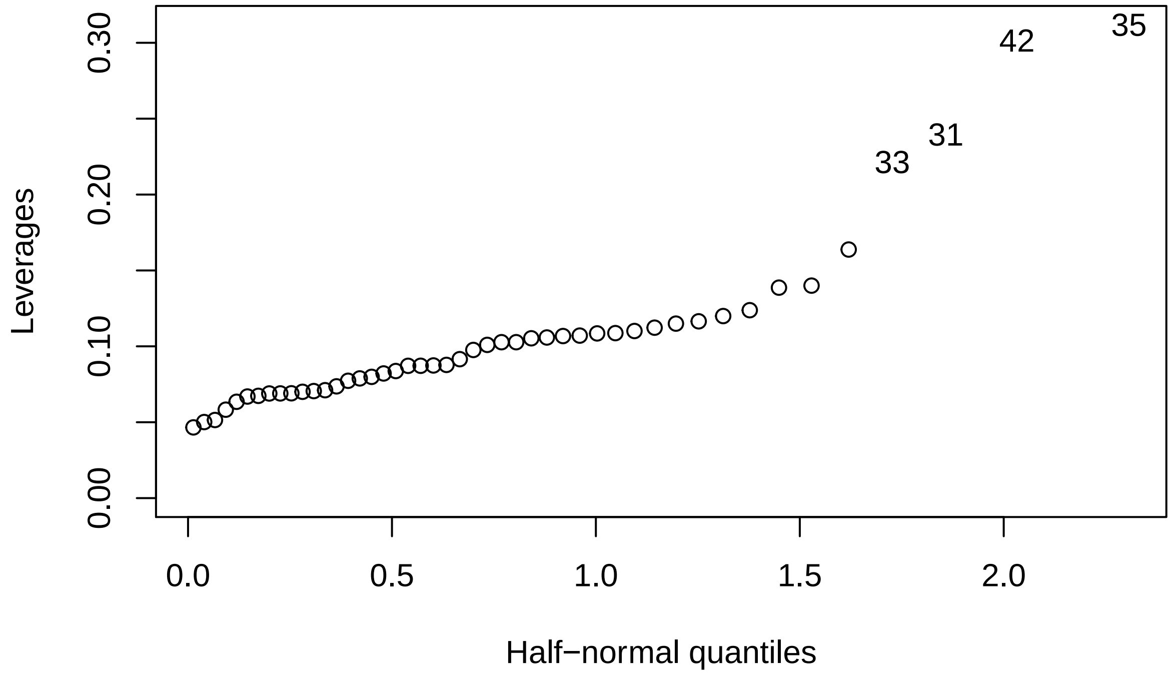
Check for large leverage points.

```
p = length(coef(teengamb_mod))
n = nrow(teengamb)
leverage = influence(teengamb_mod)$hat
high_lev_pts = leverage[leverage > 2*(p/n)]
kable(high_lev_pts, col.names = "Leverage")
```

	Leverage
31	0.2395031
33	0.2213439
35	0.3118029
42	0.3016088

Recall that a data point with "high" leverage is considered any point with leverage larger than $2 * \frac{p}{n}$ where p = number of predictors and n = number of observations in the data. With that said, using the table above, we can see that observations **31, 33, 35, and 42** are high leverage points.

```
halfnorm(leverage, nlab = 4,
         labs = row.names(teengamb), ylab = 'Leverages')
```

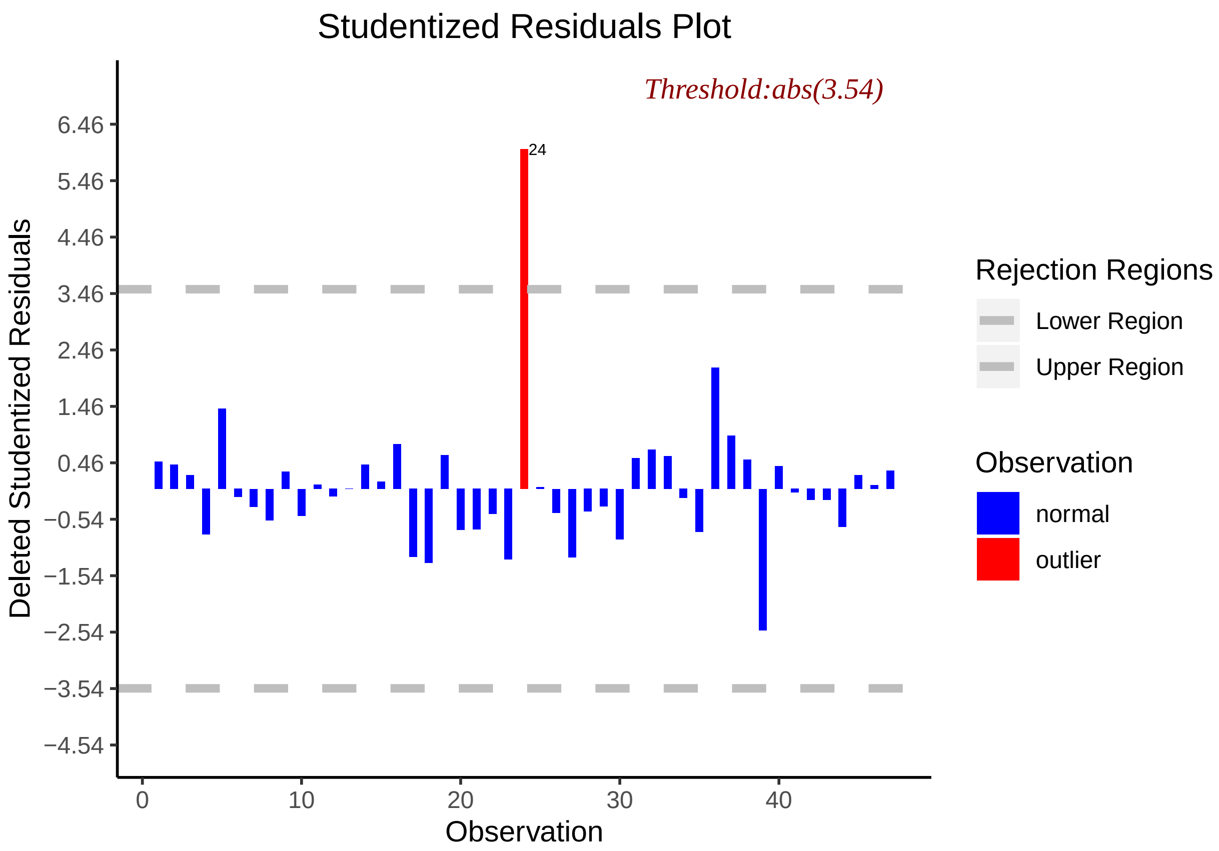


Looking at the output above, we can further observe that these points have a much higher distance between data points compared to the rest of the data. These observations can be considered to have a larger influence on the model fitting.

D.

Check for outliers.

```r
plot_outliers(teengamb_mod, teengamb)
```



To check for outliers, we will be looking at studentized residuals. Under H_0, we can use the t-test to test whether the ith observation is an outlier or not. H_0 follows a t distribution with n-p-1 degrees of freedom.

Looking at the plot above, we can see that there is 1 observation in this data that can be considered an outlier. This is observation **24**, in which the studentized residual is greater than the t-statistic for the rejection region. Therefore for this observation, we can reject H_0 and conclude that it is an outlier. Note that this outlier observation is not an observation with high leverage.

E.

Check for influential points.

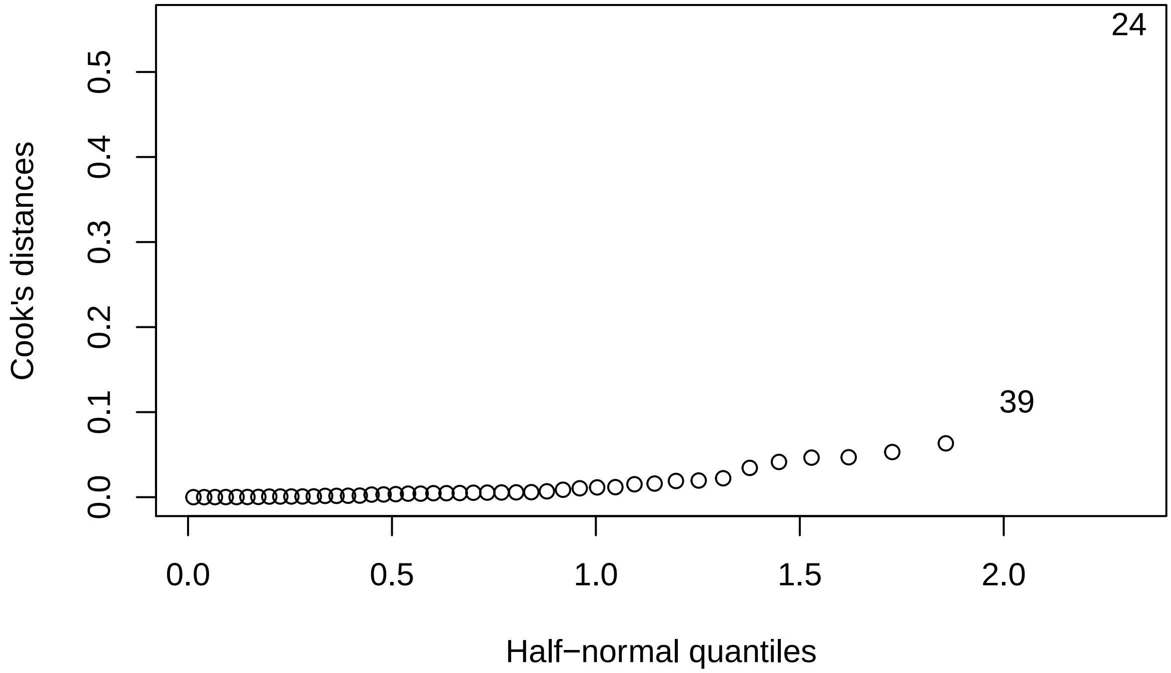
```
teengamb_cook = cooks.distance(teengamb_mod)
high_influence = teengamb_cook[teengamb_cook >= 1]
length(high_influence)
```

```
## [1] 0
```

Recall that for an observation to be considered influential it will have a Cook's distance greater than or equal to one. A Cook's distance greater than or equal to one can be clues that this observation is high leverage, an outlier, or both.

Using the output above, we see that no observations have a Cook's Distance greater than or equal to one. This is understandable as before we have looked for outliers/high influence points, and found that we have 1 outliers and only four high leverage points where their leverages were actually not that large compared to other leverages.

```r
halfnorm(teengamb_cook, labs=row.names(teengamb), ylab="Cook's distances")
```



Using the plot above, we see that points **39** and **24** can be considered to have a large Cook's distance compared to other observations. However, their Cook's distance is not greater than or equal to one. For observation **24**, we see a large Cook's distance compared to all other observations and this helps point towards a potential outlier as this observation was deemed an outlier in the previous question.

F.

Check the structure of the relationship between the predictors and the response.

```r
res_gamb_mod_m_sex = update(teengamb_mod, ~. - sex)$res
res_sex = lm(sex ~ status + income + verbal, data = teengamb)$res
sex_lm = lm(res_gamb_mod_m_sex ~ res_sex)

res_gamb_mod_m_status = update(teengamb_mod, ~. - status)$res
res_status = lm(status ~ sex + income + verbal, data = teengamb)$res
status_lm = lm(res_gamb_mod_m_status ~ res_status)

res_gamb_mod_m_income = update(teengamb_mod, ~. - income)$res
res_income = lm(income ~ sex + status + verbal, data = teengamb)$res
income_lm = lm(res_gamb_mod_m_income ~ res_income)

res_gamb_mod_m_verbal = update(teengamb_mod, ~. - verbal)$res
res_verbal = lm(verbal ~ sex + status + income, data = teengamb)$res
```

```r
verbal_lm = lm(res_gamb_mod_m_verbal ~ res_verbal)

compare_coefs = data.frame(Res.Coef = c(coef(sex_lm)[2],
                                        coef(status_lm)[2],
                                        coef(income_lm)[2],
                                        coef(verbal_lm)[2]),
                           Orig.Coef = coef(teengamb_mod)[-1])
rownames(compare_coefs) = c('Sex', 'Status', 'Income', 'Verbal')
kable(compare_coefs)
```

	Res.Coef	Orig.Coef
Sex	-22.1183301	-22.1183301
Status	0.0522338	0.0522338
Income	4.9619792	4.9619792
Verbal	-2.9594935	-2.9594935

For checking the structure of the relationship between the predictors and the response, we want to know the relationship between the response (Y) and the predictor X_k after the effect of the other predictors has been removed. To remove these effects of other predictors, I must do the following procedure for each predictor:

1. Create model 1 $Y \sim X_1 + \cdots + X_{i-1} + \cdots + X_{i+1} + \cdots$
2. Create model 2 $X_i \sim X_1 + \cdots + X_{i-1} + \cdots + X_{i+1} + \cdots$
3. Obtain residuals from model 1, r_y
4. Obtain residuals from model 2, r_k^X
5. Create model 3 $r_y \sim r_k^X$

If there is a valid relationship between the predictor and the response, we should see the same slope in model 3 compared to the slope in the original model of the given predictor.

Looking at the table output above, we see that is the case and that all predictors have identical coefficients in the residuals model compared to the original model. This confirms that our model structure is valid.

Furthermore, if we plot r_y vs. r_k^X we should see points randomly scattered around a line through the origin with slope $\hat{\beta}_k$ if the model is valid.

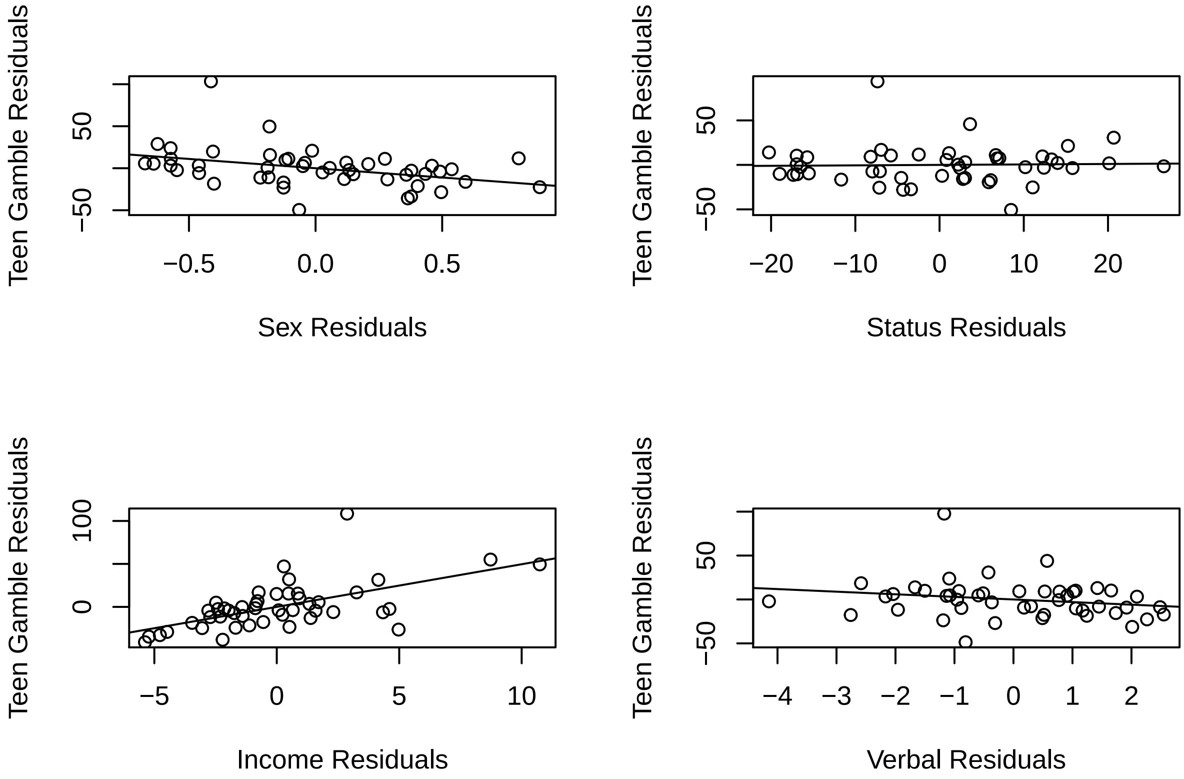
```r
par(mfrow = c(2, 2))

plot(res_sex, res_gamb_mod_m_sex,
     xlab = 'Sex Residuals', ylab = 'Teen Gamble Residuals')
abline(sex_lm)

plot(res_status, res_gamb_mod_m_status,
     xlab = 'Status Residuals', ylab = 'Teen Gamble Residuals')
abline(status_lm)

plot(res_income, res_gamb_mod_m_income,
     xlab = 'Income Residuals', ylab = 'Teen Gamble Residuals')
abline(income_lm)

plot(res_verbal, res_gamb_mod_m_verbal,
     xlab = 'Verbal Residuals', ylab = 'Teen Gamble Residuals')
abline(verbal_lm)
```



Looking at the outputs above, there seems to be slight correlation between the residuals for the **Sex** variable and the residuals from the model with **Sex** removed. There seems to also be slight correlation between the residuals for the **Income** variable and the residuals from the model with **Income** removed. However, the correlation does not appear very strong and the residuals appear to be somewhat randomly scattered around the lines. For **Status** and **Verbal**, there seems to be no correlation between residuals and points are randomly scattered around the lines. Our model structure appears to be valid.

Overall, for this dataset and model, we do not see any evidence of heteroskedasticity. We do see evidence of non-normal distribution for residuals, which can possibly be remedied using transformations suggested by the Box-Cox transformation. We do not see any high leverage or influential points, but we do see an outlier which is observation **24**. With further analysis, removing observation **24** may solve the non-normality problems and remove all outliers. The model structure is valid.