Given,

$\tilde{Y} : b_0 + b_1 X$, $b_1 \neq 0$ with $(\bar{X}, \bar{Y})$

where $\sum (Y_i - \bar{Y})^2 = \sum (\tilde{Y}_i - \bar{Y})^2 + \sum (Y_i - \tilde{Y}_i)^2$

s.t. $\tilde{Y}_i - \bar{Y} = b_1 (X_i - \bar{X})$

Show:

$b_0 = \bar{Y} - b_1 \bar{X}$

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \quad \text{or} \quad \frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2}$$

proof:

Prove that b_0 & b_1 are related. substitute $(\bar{X}, \bar{Y})$ into the linear fit:

$$\bar{Y} = b_0 + b_1 \bar{X}$$

$$b_0 = \bar{Y} - b_1 \bar{X} \quad \text{or}$$

Now we know b_1 & b_0 are related. so we show b_0 is LS estimator. We start with the SS properties.

$$\sum (Y_i - \bar{Y})^2 = \sum (\tilde{Y}_i - \bar{Y})^2 + \sum (Y_i - \tilde{Y}_i)^2$$

$$\sum (Y_i - \bar{Y})^2 = \sum \left[b_1 (X_i - \bar{X}) \right]^2 + \sum \left[Y_i - \bar{Y} - b_1 (X_i - \bar{X}) \right]^2$$

$$\sum (Y_i - \bar{Y})^2 = b_1^2 \sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2 - 2b_1 \sum (X_i - \bar{X})(Y_i - \bar{Y}) + \sum b_1^2 (X_i - \bar{X})^2$$

$$\sum (Y_i - \bar{Y})^2 = 2b_1^2 \sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2 - 2b_1 \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

$$2b_1^2 \sum (X_i - \bar{X})^2 = 2b_1 \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

is the least squares estimator for b_1 and b_0 from the normal equations. ∎

* Essential Steps: ① Use 1st property $(\bar{X}, \bar{Y})$ on $\tilde{Y}$ to show b_0 relation to b_1

② Start w/ the sum of squares special property and substitute out all $\tilde{Y}$ estimators to obtain b_1.

HW2: SLR Inference and Diagnostics (ALRM Ch. 2, 3)

MATH/STAT 571A

DUE: 09/22/2023 11:59pm

Homework Guidelines

Please submit your answers on Gradescope as a PDF with pages matched to question answers.

One way to prepare your solutions to this homework is with R Markdown, which provides a way to include mathematical notation, text, code, and figures in a single document. A template `.Rmd` file is available through D2L.

Make sure all solutions are clearly labeled, and please utilize the question pairing tool on Gradescope. You are encouraged to work together, but your solutions, code, plots, and wording should always be your own. Come and see me and/or our TA during office hours or schedule an appointment when you get stuck and can't get unstuck.

I. Mathematical Foundations [12 pts]

(1) [5 pts] Let $Y_1, \ldots, Y_n$ be a sample of response values associated with predictor values $X_1, \ldots, X_n$. Suppose that $\tilde{Y} = b_0 + b_1 X$, $b_1 \neq 0$, is a line that passes through $(\bar{X}, \bar{Y})$ and satisfies the sums of squares relationship $\sum(Y_i - \bar{Y})^2 = \sum(\tilde{Y}_i - \bar{Y})^2 + \sum(Y_i - \tilde{Y}_i)^2$. Show that b_0 and b_1 must be the least-squares estimates of the simple linear regression of Y_i onto X_i (hint: lines that pass through $(\bar{X}, \bar{Y})$ must satisfy $\tilde{Y}_i - \bar{Y} = b_1(X_i - \bar{X})$).

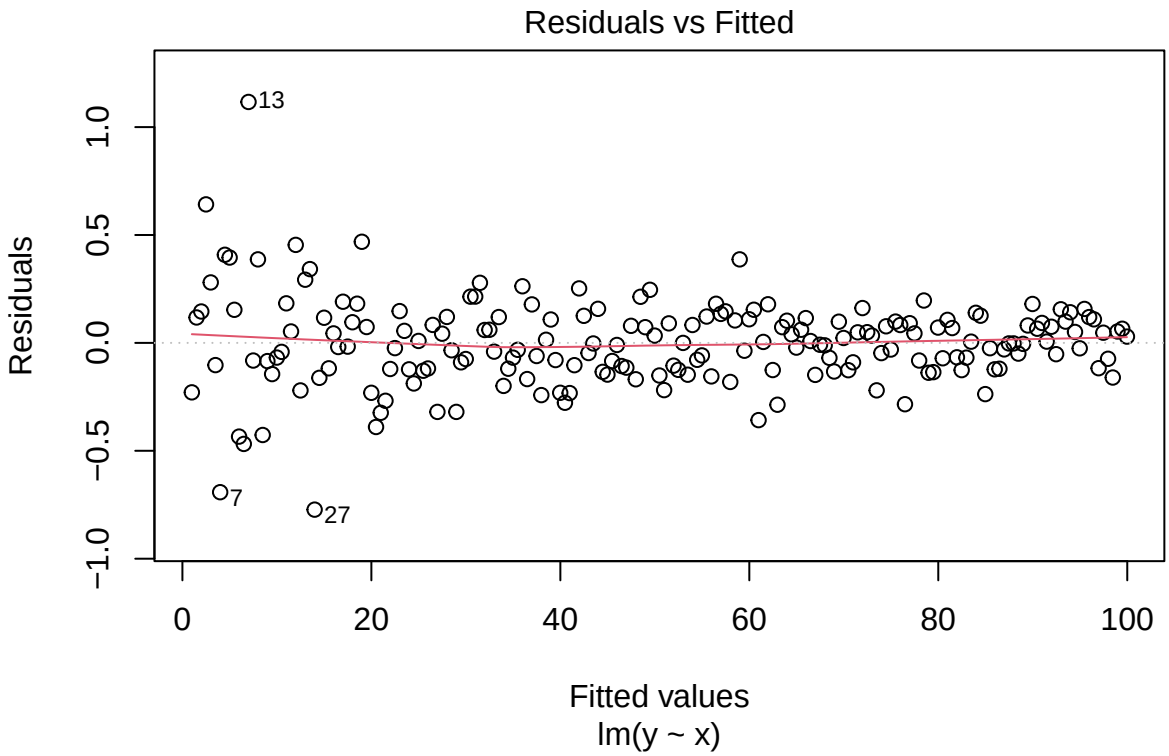
(2) [3 pts] ALARM 2.18: For conducting statistical tests concerning the parameter β_1, why is the t test more versatile than the F test?

The T-test is more versatile because it can be used for one-sided alternative hypotheses which the F-test cannot.

(3) [4 pts] ALRM 3.2: Prepare a prototype residual plot for each of the following cases. You can prepare your plots in R, or draw them neatly by hand.

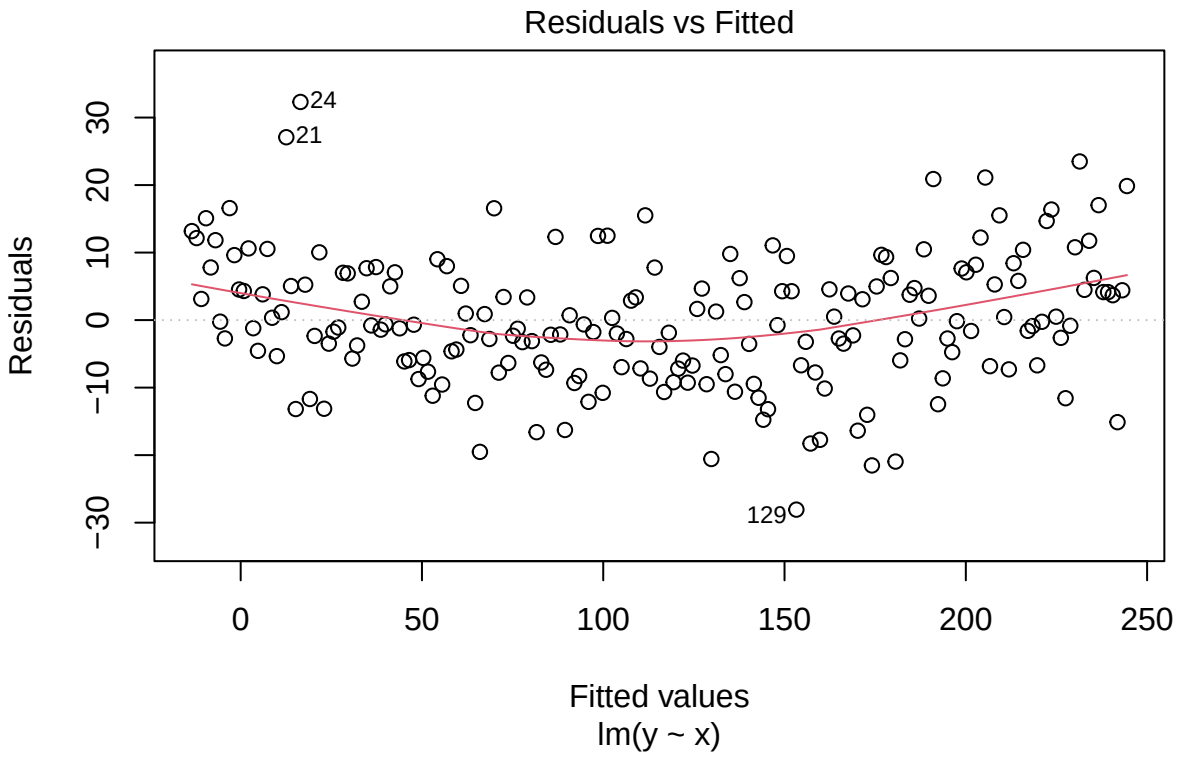
(i) error variance decreases with X

```
x = seq(1,100,by=.5)
b0 = 0
b1 = 1
sigma2 = x^-1
eps = rnorm(x, mean = 0, sd = sqrt(sigma2))
y = b0 + b1*x + eps
fit = lm(y ~ x)
plot(fit,1)
```



(ii) the true regression function is "U"-shaped, but a linear regression function is fitted.

```
x = seq(1,100,by=.5)
b0 = 0
b1 = 1
sigma = 10
eps = rnorm(x, mean = 0, sd = sigma)
y = b0 + b1*x^1.2 + eps
fit = lm(y ~ x)
plot(fit,1)
```



II. Tree Cover in Tucson [28 pts]

Download the Tree Equity Scores data set from Module 1 on D2L (`Tree_Equity_Scores_Tucson_noNA.csv`). Use the `read.csv()` function to import the comma spaced values as a dataframe in R. There are several variables recorded in this data set, but we will continue to focus primarily on two (read more here: https://gisdata.tucsonaz.gov/datasets/cotgis::tree-equity-scores-tucson-1/about). The first is `PCTTreeCover`, the percent of each corresponding census tract that is covered by tree canopy. The second is `PCTPoverty`, the percentage of households in each census tract with an income below the poverty line. We will also look at the percentage of residents in each tract that are children (under 17), `PCTChildren`, for one question.

Treat percent tree cover as the response variable and percent below poverty line as a predictor.

(4) [2 pts] Obtain a 99% confidence interval for β_1, the effect of percent below poverty line on tree cover. Does it include 0?

The 99% confidence interval for beta_1 is [-0.07013332, -0.006111215] and does not contain 0.

```
trees <- read.csv('./Datasets/Tree_Equity_Scores_Tucson_noNA.csv')
fit_trees <- lm(PCTTreeCover ~ PCTPoverty, data = trees)
confint(object = fit_trees, level = 0.99)
```

```
##                   0.5 %       99.5 %
## (Intercept)  9.86730128 12.732668061
## PCTPoverty  -0.07013332 -0.006111215
```

(5) [2 pts] Report the p-value for the two sided t-test for whether or not a linear relationship exists between the predictor and response. Do you reject the null hypothesis of no linear relationship at the $\alpha = 0.01$ level? How does your conclusion reconcile with your confidence interval from the previous question?

The p-value is 0.0022 which allows us to reject the null hypothesis at a 0.01 alpha level. This agrees with our 99% confidence which does not include 0.

```
summary(fit_trees)
```

```
##
## Call:
## lm(formula = PCTTreeCover ~ PCTPoverty, data = trees)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2780  -3.0233  -0.3271   2.8446  20.4342
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.29998    0.55357  20.413   <2e-16 ***
## PCTPoverty  -0.03812    0.01237  -3.082   0.0022 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.771 on 403 degrees of freedom
## Multiple R-squared:  0.02303,    Adjusted R-squared:  0.02061
## F-statistic:   9.5 on 1 and 403 DF,  p-value: 0.002196
```

(6) [3 pts] Compute and report the standard error, $s(\hat{Y}_h)$, for a prediction of the expected percent tree cover in census tracts where 20% of residents have incomes below the poverty line. Obtain a 90% confidence interval for the expected percent tree cover in census tracts with 20% of residents below the poverty line. What is the width of your interval?

The width is 1.142882.

```
pred_interval = predict(fit_trees, newdata = data.frame(PCTPoverty = 20),
                        interval = 'confidence',
                        level = .9)
pred_interval
```

```
##        fit      lwr      upr
## 1 10.53754 9.966098 11.10898
```

```
pred_interval[[3]]-pred_interval[[2]]
```

```
## [1] 1.142882
```

(7) [3 pts] Compute and report the standard error, $s(\text{pred})$, for a prediction of percent tree cover in a new census tract where 20% of residents have incomes below the poverty line. Obtain a 90% confidence interval for the prediction at a new tract. What is the width of your interval?

The width is 15.77207.

```
pred_interval = predict(fit_trees, newdata = data.frame(PCTPoverty = 20),
                        interval = 'predict',
                        level = .9)
pred_interval
```

```
##        fit      lwr      upr
## 1 10.53754 2.651505 18.42357
```

```
pred_interval[[3]]-pred_interval[[2]]
```

```
## [1] 15.77207
```

(8) [3 pts] Determine the boundary values of the 90% Working-Hotelling confidence band for the regression line when the percent of residents with an income below the poverty line is 20%. What is the width of your interval?

The width is 1.605366.

```
# Taken from Session 4
pred_WH <- function(object, newdata, level = 0.95){
  fit <- predict(object, newdata)
  MSE <- summary(object)$sigma^2
  n <- nrow(object$model)
  W <- qf(level, 2, n - 2)

  X_obs <- object$model[, attr(object$terms, "term.labels")]
  X <- newdata[, attr(object$terms, "term.labels")]
  ME <- W * sqrt(MSE * (1 / n + (X - mean(X_obs))^2 /
                          sum((X_obs - mean(X_obs))^2)))
  upr <- fit + ME
  lwr <- fit - ME
  return(cbind(fit, lwr, upr))
}
WH_pred <- pred_WH(fit_trees, newdata = data.frame(PCTPoverty = 20), level = 0.9)
WH_pred
```

```
##        fit      lwr      upr
## 1 10.53754 9.734856 11.34022
```

```
WH_pred[[3]] - WH_pred[[2]]
```

```
## [1] 1.605366
```

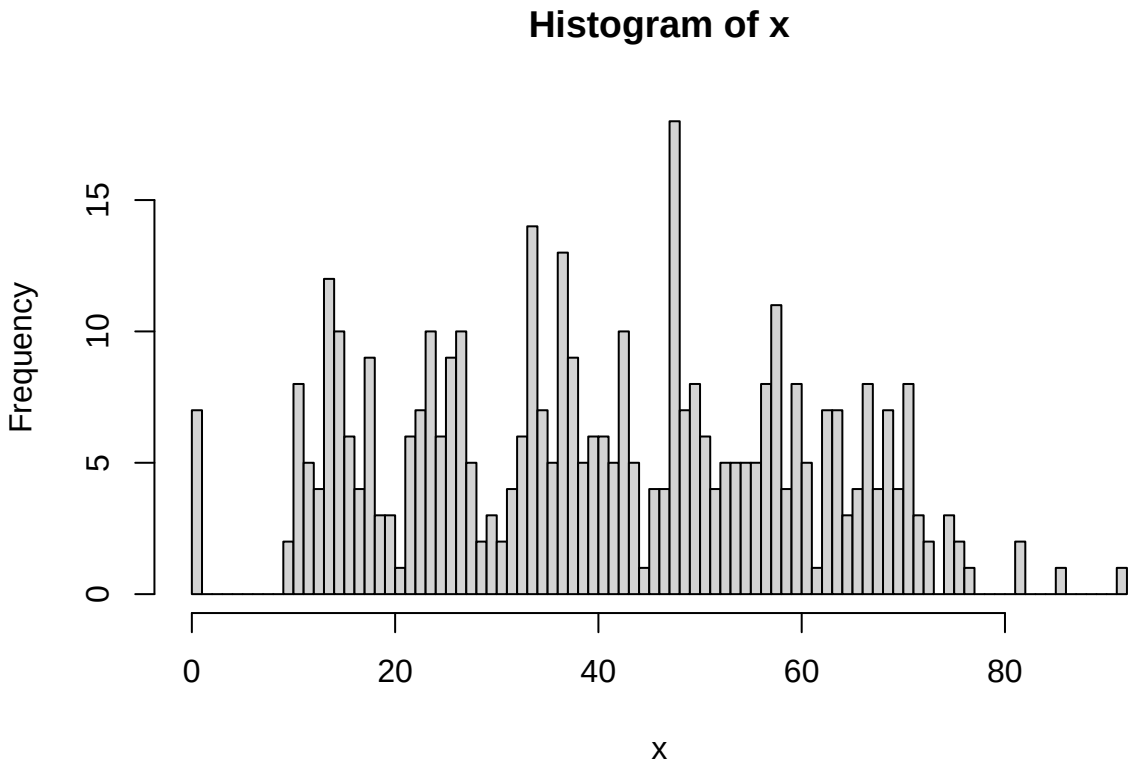
(9) [2 pts] Prepare a histogram of the predictor values. Do you notice any potential outliers?

There appear to be 7 outliers at the 0 value for PCTPoverty.

```
x = trees$PCTPoverty
head(sort(x), n = 10)
```

```
##  [1]  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
##  [8]  9.896907 10.000000 10.021946
```

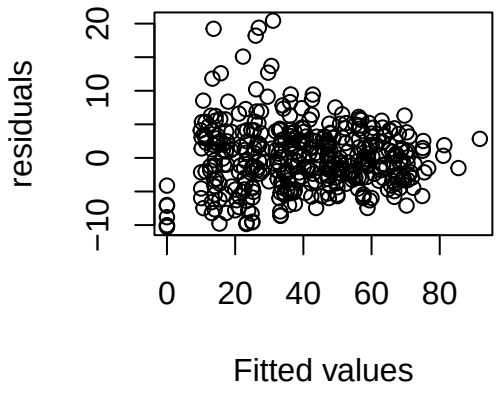
```
hist(x, breaks = 100)
```



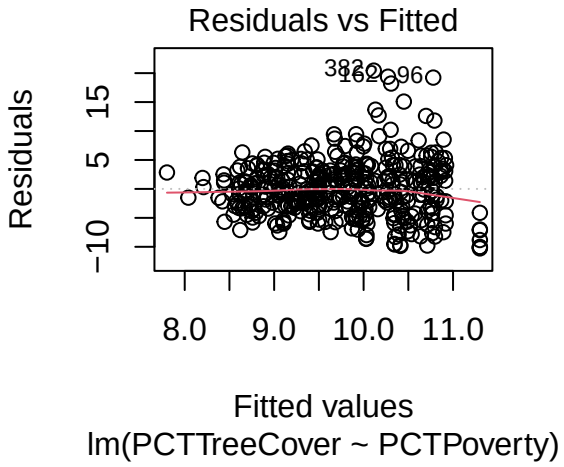
(@) [3 pts] Prepare plots of the residuals vs. both the fitted values and predictor values. What potential issue(s) with the normal SLR model does your two diagnostic plots raise? Describe what concerning feature(s) you see in the figures.

We see heteroskedasticity: -lower predictor values correspond to higher variance in predictions -higher fitted values have higher variance in predictions Overall, this indicates a negative relationship that has decreasing variation with larger predictor values.

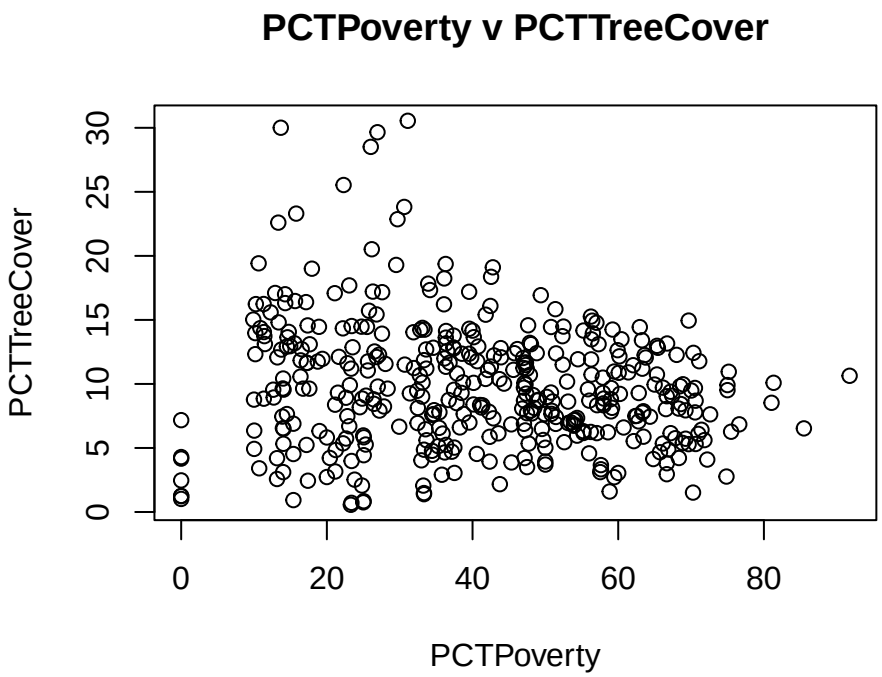
```
# Residuals v. predictor
plot(trees$PCTPoverty, fit_trees$residuals, ylab = "residuals", xlab = "Fitted values")
```




```
# Residuals v. fitted
plot(fit_trees,1)
```



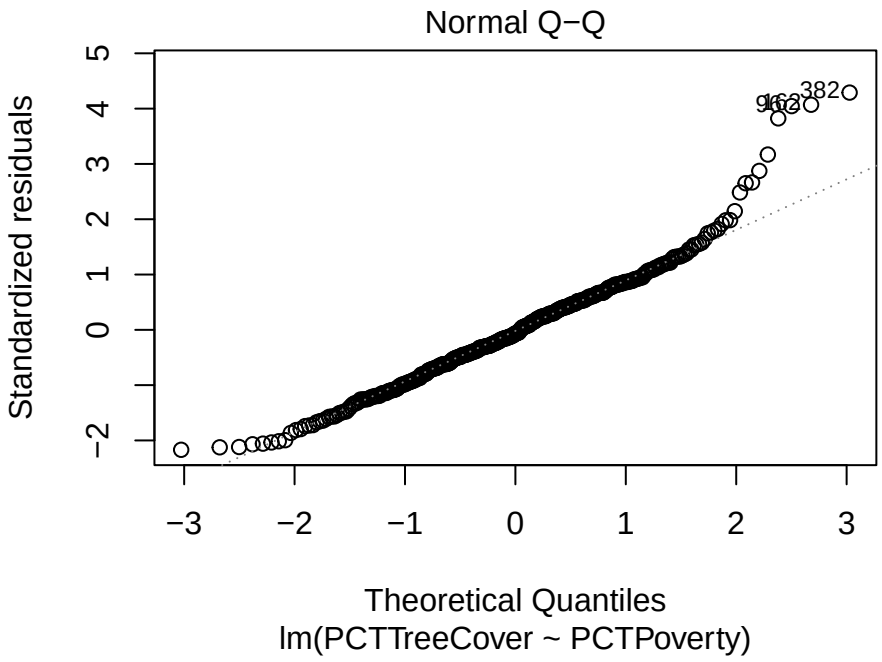

```
plot(trees$PCTPoverty, trees$PCTTreeCover, main = "PCTPoverty v PCTTreeCover",
     xlab = "PCTPoverty", ylab = "PCTTreeCover")
```



(@) [2 pts] Prepare a QQ-plot comparing the residuals to a normal reference distribution. What potential issue(s) with the normal SLR model does your two diagnostic plots raise? Describe what concerning feature(s) you see in the figure.

The QQ plot reveals that there is higher variation at the larger fitted values which can be inferred from the visibly fatter tail on the right.

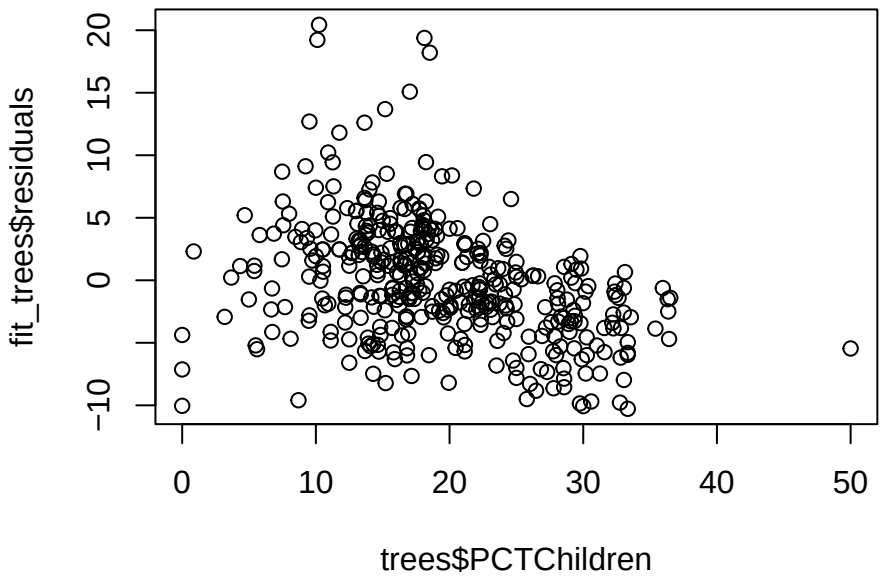
```
plot(fit_trees,2)
```



(10) [3 pts] Make a plot of the residuals on the vertical axis and the percentage of residents who are children (under 17). Does your plot suggest that this might be an important omitted variable for predicting percent tree cover? Why/why not?

It appears that PCTChildren is an important variable that is related to the model. The variation decreases with increasing values of PCTChildren.

```
plot(x = trees$PCTChildren, y = fit_trees$residuals)
```



(11) [2 pts] Suggest a transformation to either the predictor or response variable that might address one of the concerns you raised in (10). Be sure to indicate what issue the transformation is intended to address and why you think it might help.

I would like to use the boxcox transformation to address the high variance at lower-valued predictor values.

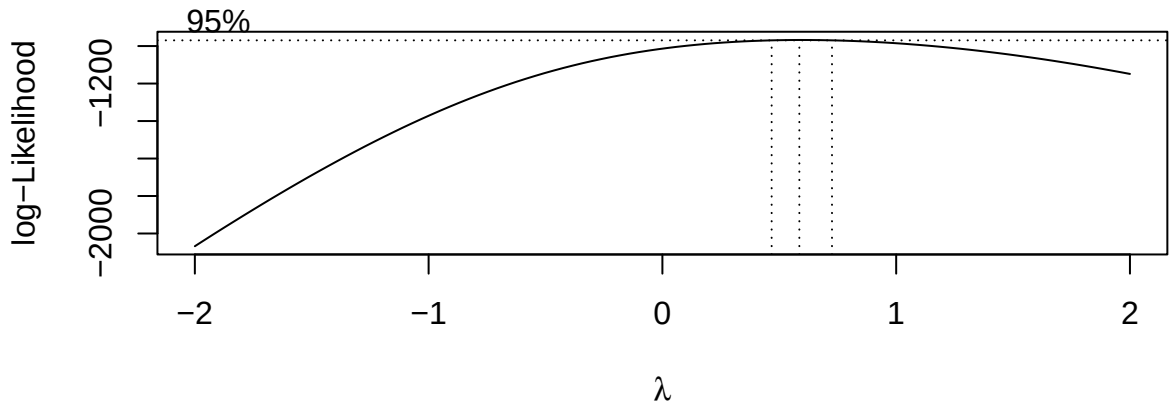
(12) [3 pts] Implement the transformation you suggested in (13) and comment on how much it helped.

It has helped reduce the changing variance drastically.

```
library(MASS)
```

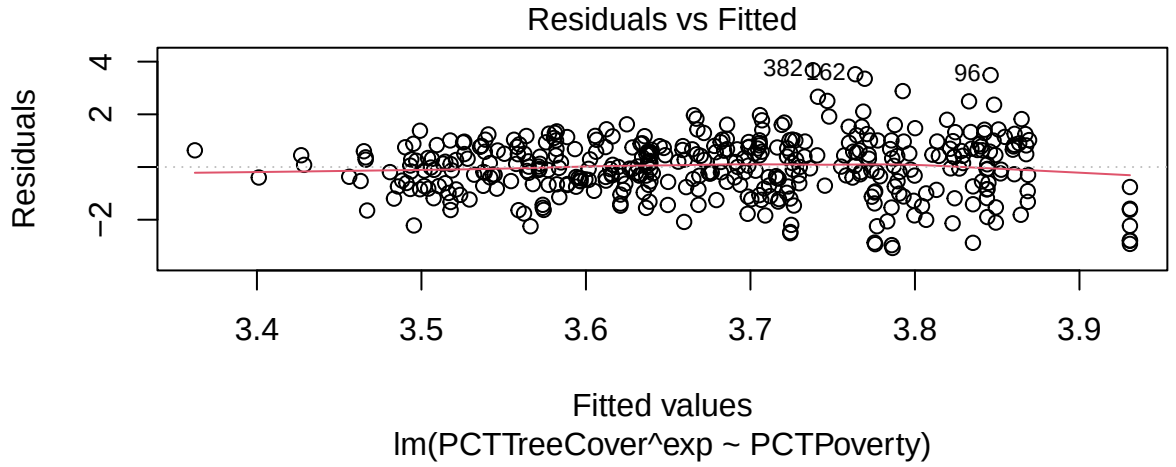
```
## Warning: package 'MASS' was built under R version 4.2.3
```

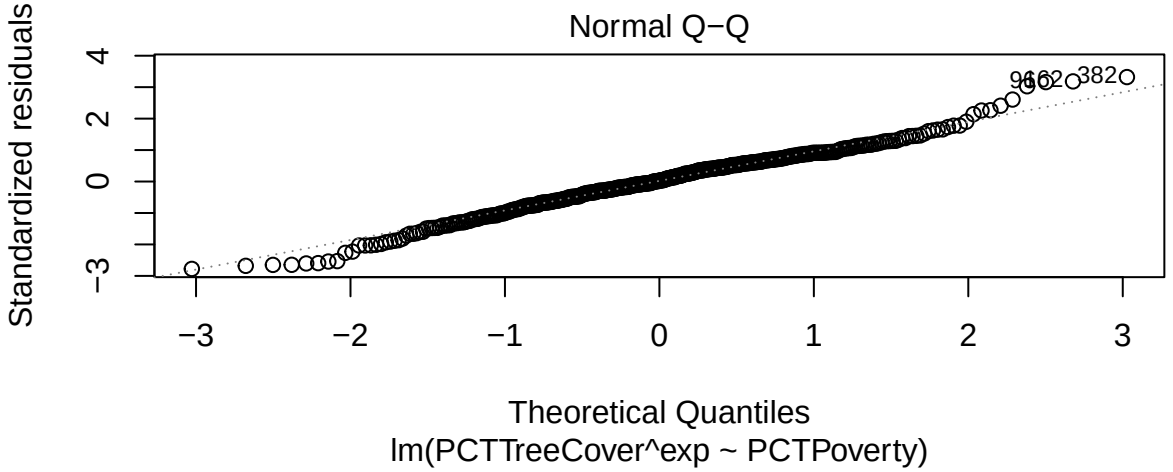
```
bc <- boxcox(fit_trees)
```




```
exp <- bc$x[which.max(bc$y)]

fit_bc <- lm(PCTTreeCover^exp ~ PCTPoverty, data = trees)
plot(fit_bc, c(1,2))
```





(*BONUS*) [3 pts] Conduct a test for lack of fit by binning predictor values into bins of width 5 and performing an appropriate general linear test. Is there sufficient evidence to reject the linear model at the $\alpha = 0.01$ level?