Stat230 - HW 3

Owen Forman

```r
#loads necessary libraries
library(ggplot2)
library(patchwork)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

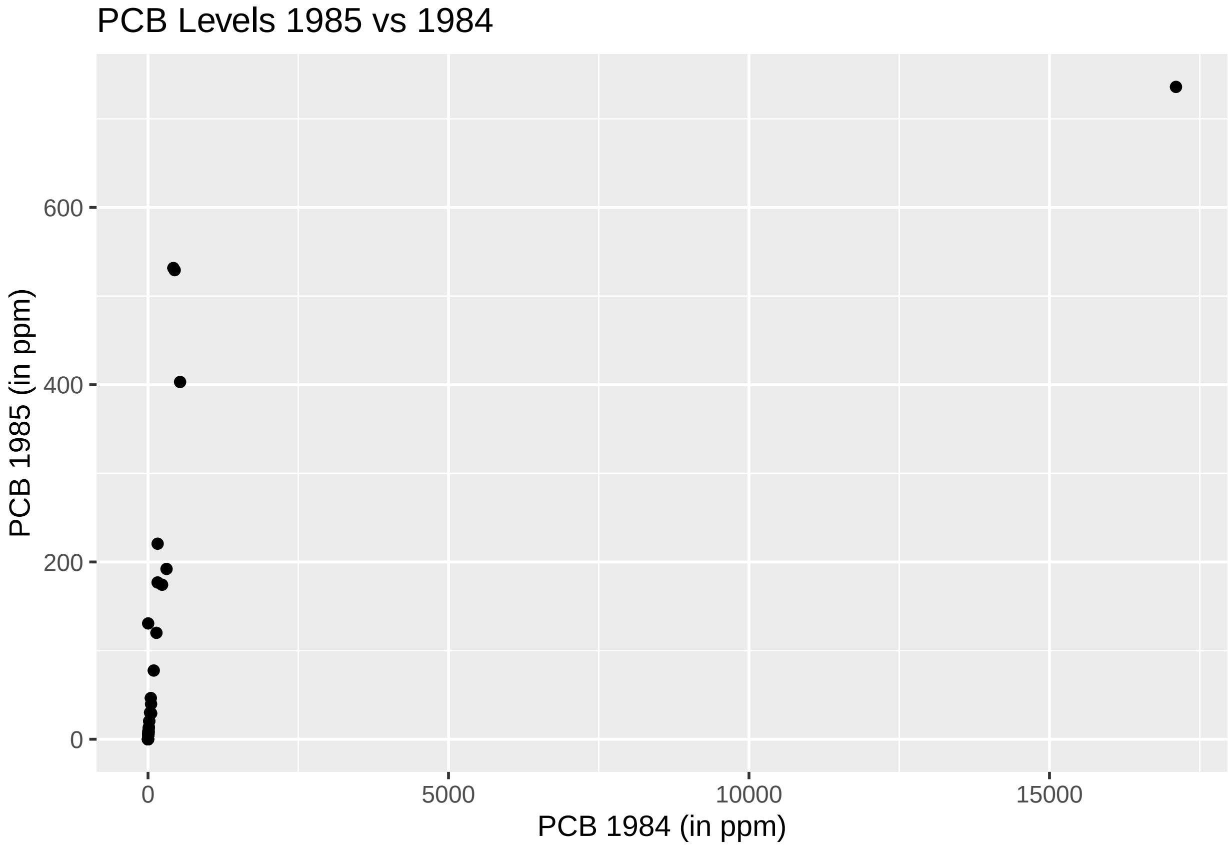
```r
library(broom)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

1b) **Answer:** Both models do not have constant variance. This is because they are both not following a linear trend, so when the SLR line is fitted, there will be non-constant variance as some fitted values are closer to the points than others. The difference between the two models, is that the simulated points in model 1 begin to diverge as x increases (aka residuals increase/decrease seemingly exponentially), whereas the simulated points in model 2 maintain their (still non-linear) original trend, which would cause the residuals to follow the same curved line as the model in the residual plot. If we tried to use model 2 to make a prediction about model 1 it is likely that we would get an incorrect answer, though we could use model 2 to potential predict some kind of mean value (or close to it) for model 1, as it appears like model 2 follows the same non-linear trend, just without divergence of points.

2a) **Answer:** The extreme outlier case was Boston Harbor, which had extremely high pcb levels in 1984. Upon removing the outlier though, it still turns out that the data is not yet suited for a SLR model. We can see this visually in the scatter plot “PCB Levels 1985 vs 1984 (No Outlier)” since the trend does not appear to be linear. Additionally, after attempting to fit a SLR model, it becomes even clearer that the model is not appropriate, as the residual plot”Pcb Resid plot (No outliers)” shows clear grouping around x = 0 and residuals become more widespread at larger x’s which is a clear patter. Thus an SLR model cannot be used despite removing the outlier.

```r
#reads in data
pcb <- read.csv("https://www.math.carleton.edu/ckelling/data/Pcb.csv")

#creates scatterplot of pcb85 vs pcb84.
ggplot(data = pcb, aes(x = pcb84, y = pcb85)) +
  geom_point() +
  labs(title = "PCB Levels 1985 vs 1984", x = "PCB 1984 (in ppm)", y ="PCB 1985 (in ppm)")
```



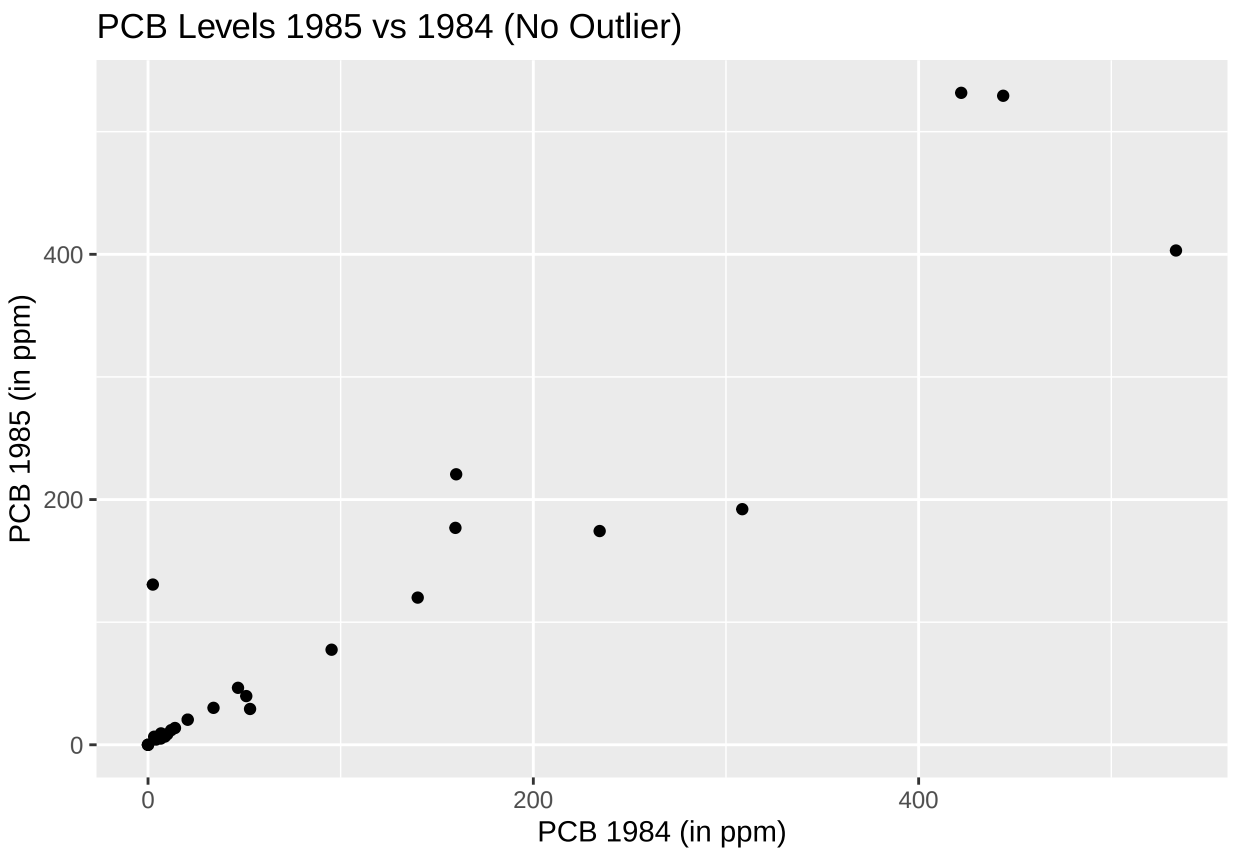
```
#identifies extreme outlier value
which(pcb$pcb84 > 10000)
```

```
## [1] 4
```

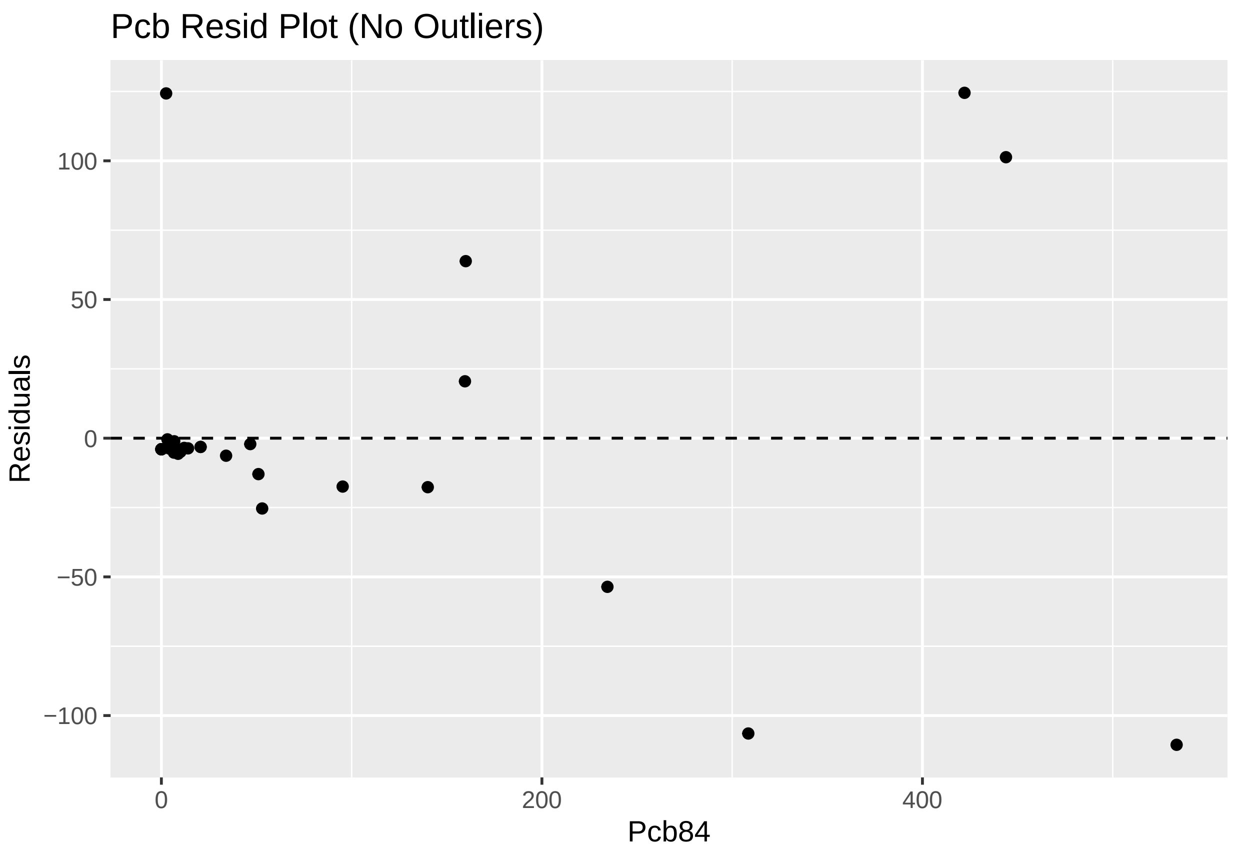
```
#prints row with outlier value
pcb[4,]
```

```
##              Site    pcb84 pcb85
## 4 Boston Harbor 17104.86   736
```

```
#removes outlier value (note that this data point was Boston Harbor)
pcb_new <- pcb[-4,]

#creates new scatterplot of pcb85 vs pcb84 - no outlier included
ggplot(data=pcb_new, aes(x = pcb84, y = pcb85)) +
  geom_point() +
    labs(title = "PCB Levels 1985 vs 1984 (No Outlier)", x = "PCB 1984 (in ppm)", y ="PCB 1985 (in ppm)
```



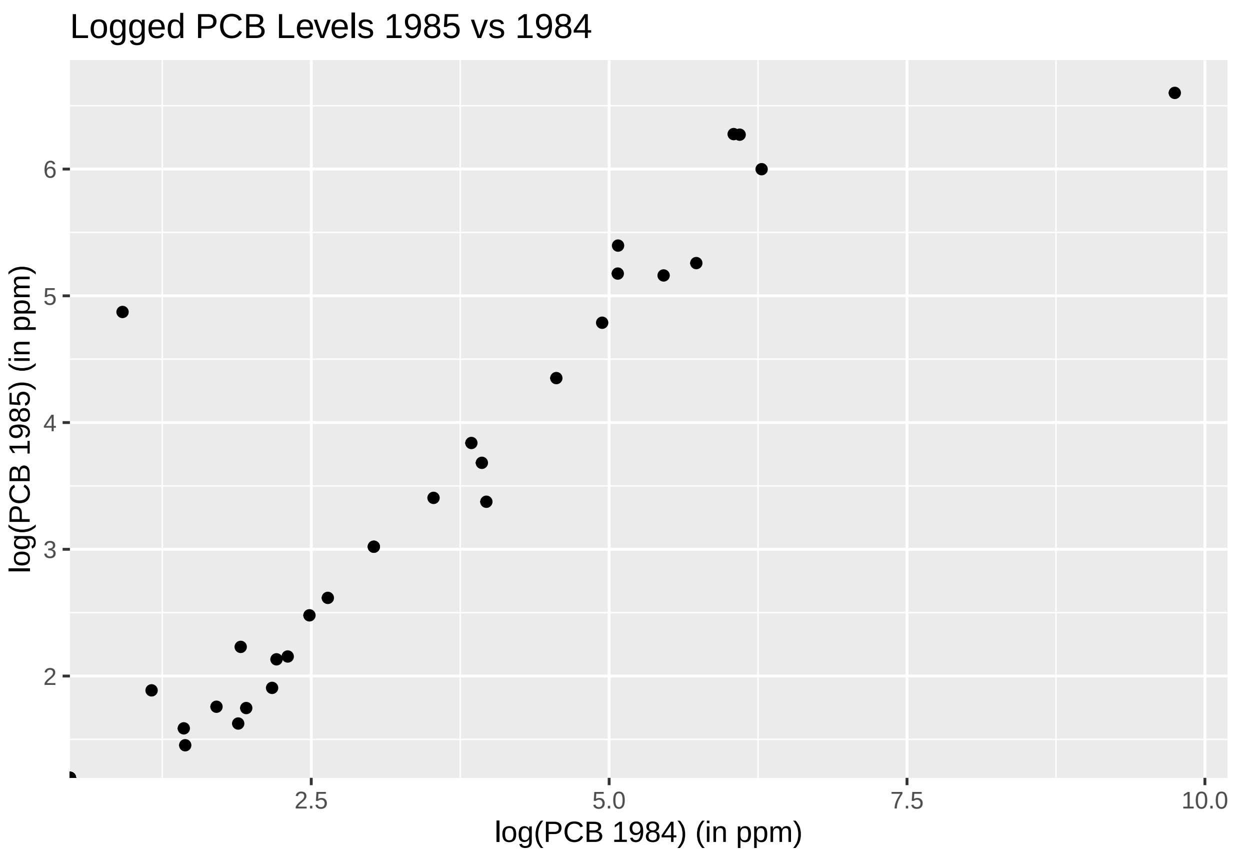
```
pcb_new %>%
  #creates SLR model on the new no outlier data
  lm(data = ., pcb85 ~ pcb84) %>%

  #augments model
  augment(.) %>%

  #creates and prints resid plot for model
  ggplot(., aes(x = pcb84, y = .resid)) +
    geom_point() +
    geom_hline(yintercept = 0, linetype = "dashed") +
    labs(title = "Pcb Resid Plot (No Outliers)", y = "Residuals", x = "Pcb84")
```



2b)

Answer: While we could not justify fitting a SLR model onto the previous graph in 2a), we can justify this model being used in the "Logged PCB Levels 1985 vs 1984" seen below. This is because the data has been transformed using the log function, which has uncovered the linear relationship between pcb84 and pcb85. This can clearly be seen visually, as the trend went from following an exponential line of some kind, to a more controlled linear line. While there are still outliers (which will be assessed in the coming problems), there are only really two or three of note, which may be the result of key information left out of this analysis. Though the reason these points are so extreme cannot be determined, it is safe to say given the low number of outliers, and the extremely strong linear relationship the rest of the data appears to follow, that we can still reasonably use an SLR model despite these outlier points.

```
#creates scatterplot of log(pcb85) vs log(pcb84).
ggplot(data=pcb, aes(x = log(pcb84), y = log(pcb85))) +
  geom_point() +
  labs(title = "Logged PCB Levels 1985 vs 1984", x = "log(PCB 1984) (in ppm)", y ="log(PCB 1985) (in ppm
```



2c)

Answer: Some of the values = -Inf because the test sites had 0 listed for pcb levels in 1984 and 1985. Since the lim x -> 0 lnx = -Inf it is clear to see that when we too the natural log of these 0 values R spits back -Inf as the answer, since ln(0) is technically undefined.

```r
#creates mutated dataset with log(pcb85) and log(pcb84) values
pcb_log <- pcb %>%
  mutate(., pcb84 = log(pcb84), pcb85 = log(pcb85))

#finds which of these values are less than 0, aka -Inf
which(pcb_log$pcb84 < 0)
```

```
## [1] 12 14 16 18 19 21 22 23
```

```r
which(pcb_log$pcb85 < 0)
```

```
## [1] 12 14 16 18 19 21 22 23
```

```r
#finds which of the original (non-logged) values are = 0
which(pcb$pcb84 == 0)
```

```
## [1] 12 14 16 18 19 21 22 23
```

```r
which(pcb$pcb85 == 0)
```

```
## [1] 12 14 16 18 19 21 22 23
```

```r
#locates which sites had -Inf values and prints them
pcb_log[c(12,14,16,18,19,21,22,23),]
```

```
##                    Site pcb84 pcb85
## 12     Pamilico Sound  -Inf  -Inf
## 14       Sapelo Sound  -Inf  -Inf
## 16          Tampa Bay  -Inf  -Inf
## 18         Mobile Bay  -Inf  -Inf
## 19       Round Island  -Inf  -Inf
## 21      Barataria Bay  -Inf  -Inf
## 22    San Antonio Bay  -Inf  -Inf
## 23 Corpus Christi Bay  -Inf  -Inf
```

```r
#filters -Inf values out
pcb_log_filtered <- pcb_log %>%
  filter(., pcb84 > 0, pcb85 >0)
```

2d)

Answer: Model Assumptions: As currently stands, our model assumptions are not met. Firstly, there appears to be a clear linear trend in the Residual plot, which indicates that our assumption of linearity and constant variance are not reasonable for this model.

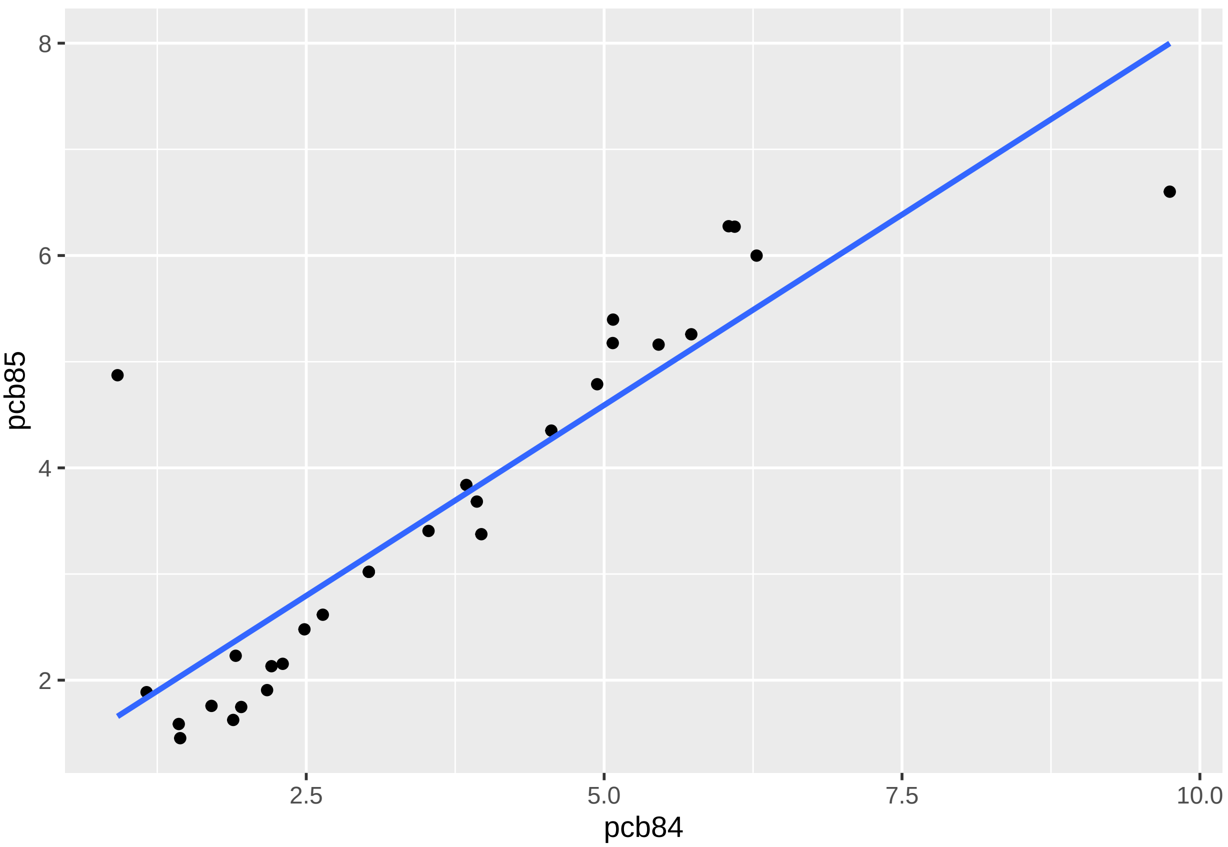
Next, for the assumption of Normality, we can observe the QQ plot. The QQ-plot is not perfect either, but does appear to generally track the expected normal quantiles. Other than the two outlier points which fall drastically off of the line, and a bit of bowing at both ends of the plot, it appears as if Normality may be met.

Lastly, independence, which is especially hard to check in this example since almost no information about how the data was collected was given. However, one possible cause for concern is that its possible some of these rivers lead into each other (there is no evidence for or against this given in the data description so while it may be unlikely it can't be completely disregarded). If two or more of these rivers attach it wouldn't be absurd to see the pcb levels of one river affect that of another, causing them to become dependent.

Outliers: The two obvious outliers are Boston Harbor and Delaware Bay. which correspond to rows 4 and 10 respectively (in the filtered log dataset)

```r
#creates scatterplot of log(pcb85) vs log(pcb84) filtered to have no -Inf values.
ggplot(data=pcb_log_filtered, aes(x = pcb84, y = pcb85)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

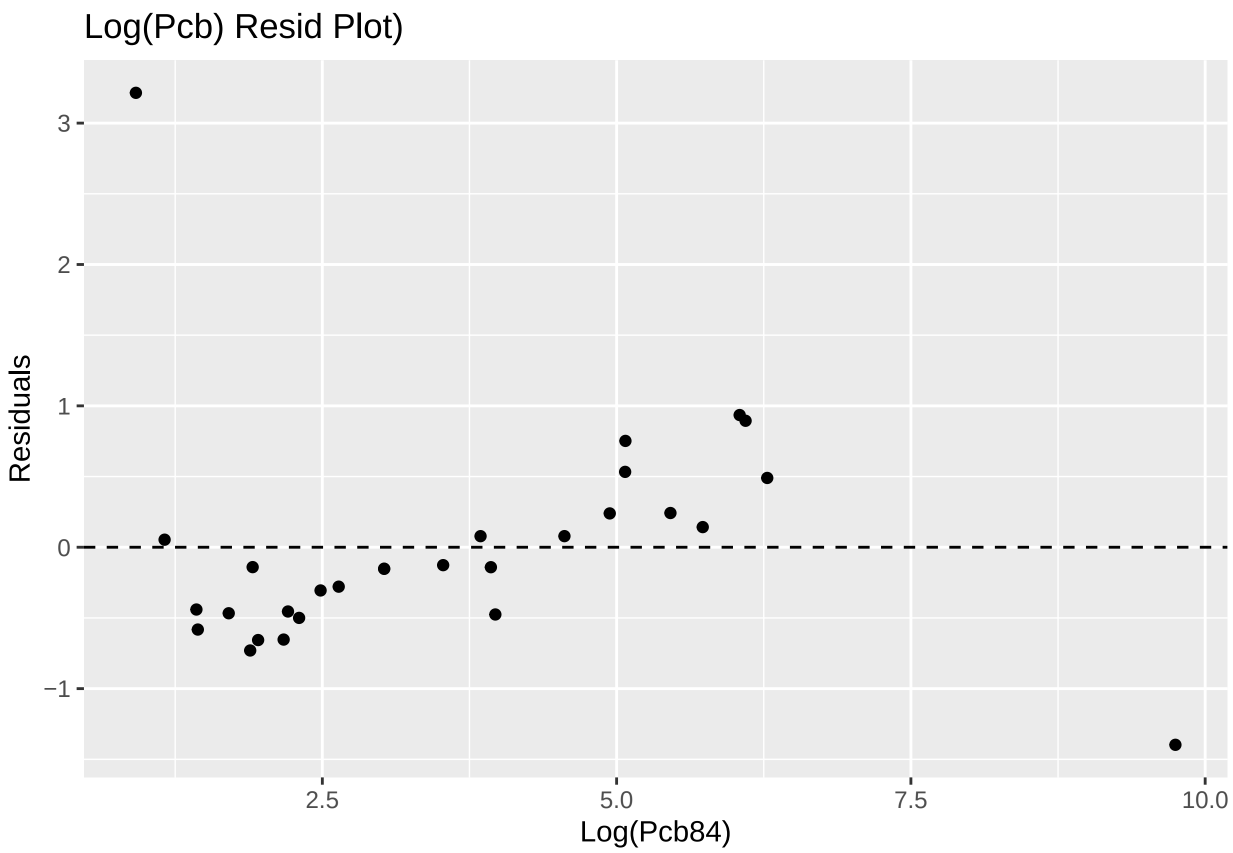
```
## `geom_smooth()` using formula = 'y ~ x'
```



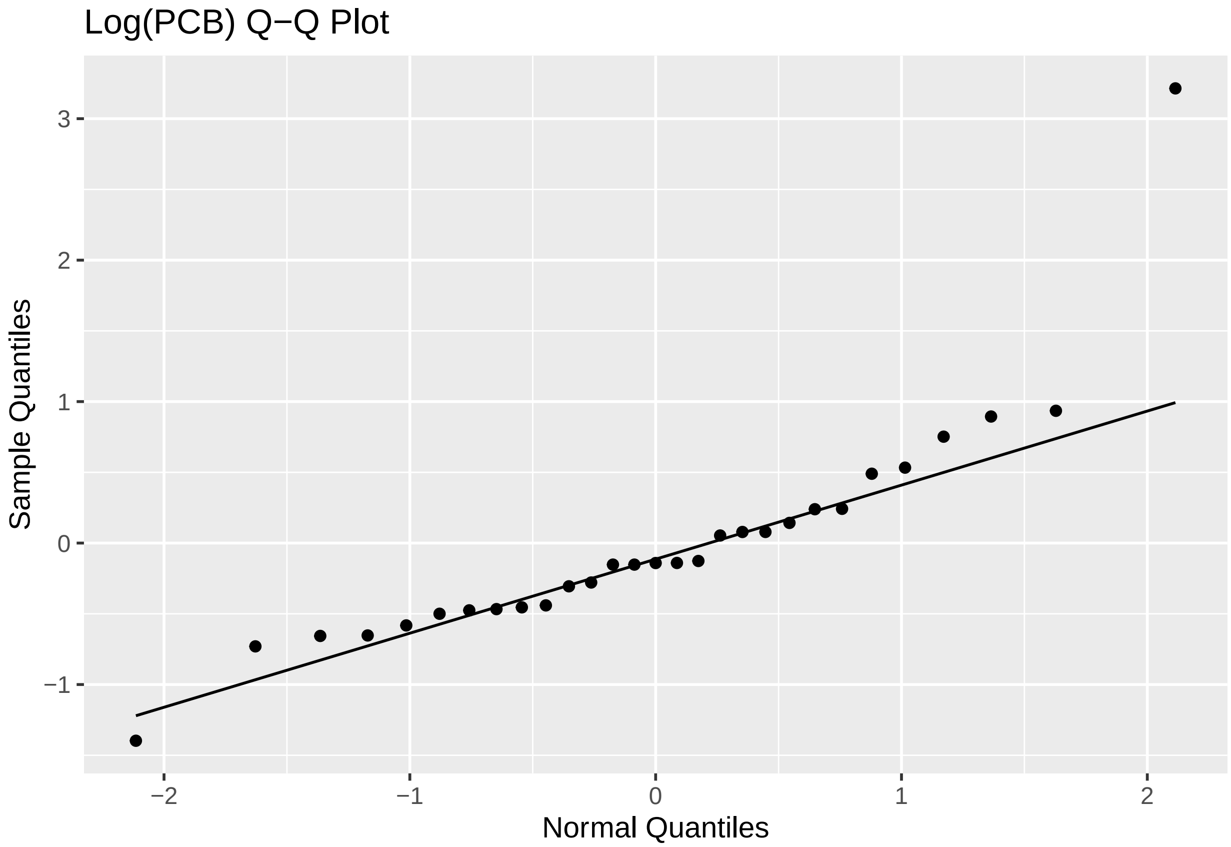
```
  labs(title = "Filtered logged PCB Levels 1985 vs 1984", x = "log(PCB 1984) (in ppm)", y ="log(PCB 198
```

```
## $x
## [1] "log(PCB 1984) (in ppm)"
##
## $y
## [1] "log(PCB 1985) (in ppm)"
##
## $title
## [1] "Filtered logged PCB Levels 1985 vs 1984"
##
## attr(,"class")
## [1] "labels"
```

```
  pcb_log_filtered_fit <- lm(data = pcb_log_filtered, pcb85 ~ pcb84)

  pcb_log_filtered_aug <- augment(pcb_log_filtered_fit)

  ggplot(pcb_log_filtered_aug, aes(x = pcb84, y = .resid)) +
    geom_point() +
    geom_hline(yintercept = 0, linetype = "dashed") +
    labs(title = "Log(Pcb) Resid Plot)", y = "Residuals", x = "Log(Pcb84)")
```



```
ggplot(pcb_log_filtered_aug, aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line() +
  labs(title = "Log(PCB) Q-Q Plot", y = "Sample Quantiles", x = "Normal Quantiles")
```



```
#finds which entries have resid > 2 (upper outlier) and resid < -1 (lower outlier)
which(pcb_log_filtered_aug$.resid > 2 | pcb_log_filtered_aug$.resid < -1)
```

```
## [1]  4 10
```

```
#prints out the entries (note we found which ones they were in the augmented dataset, and then locate t
pcb_log_filtered[c(4,10),]
```

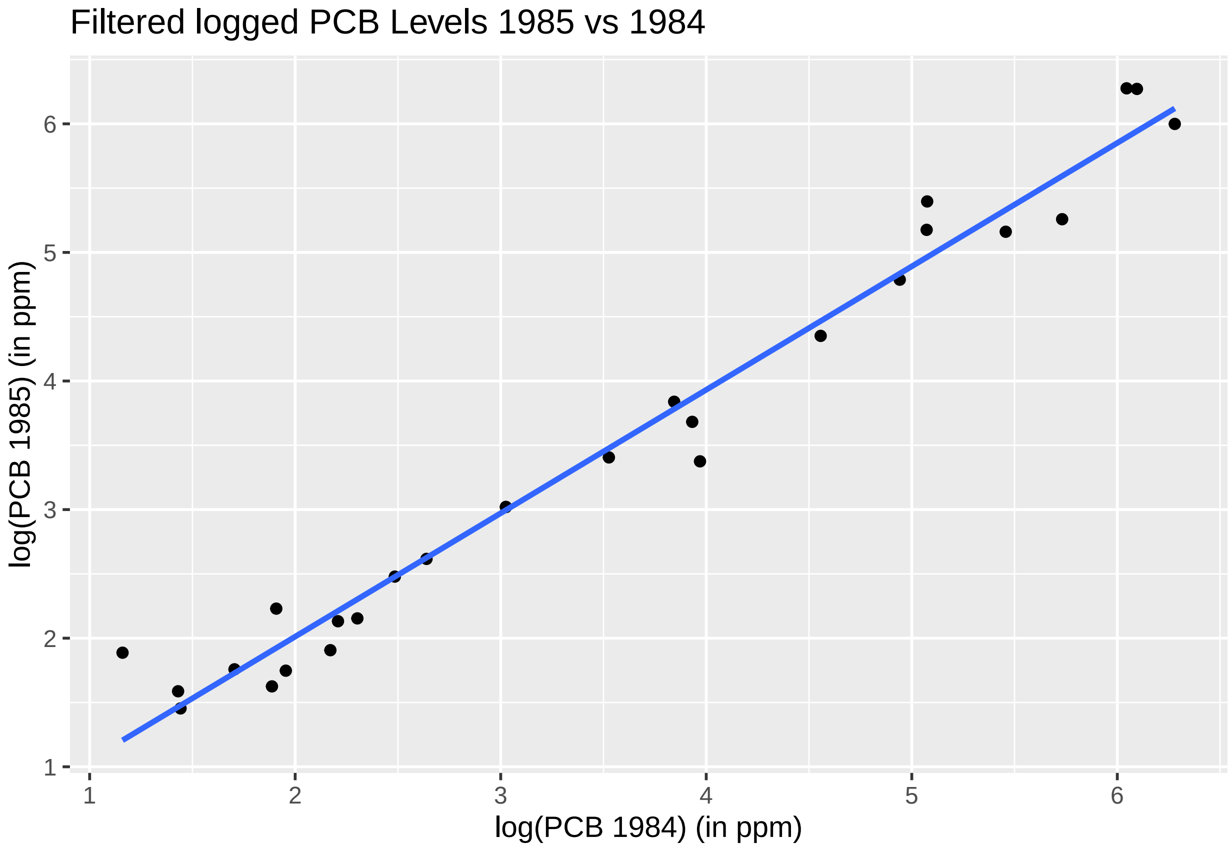
```
##               Site     pcb84    pcb85
## 4  Boston Harbor 9.7471179 6.601230
## 10  Delaware Bay 0.9162907 4.872675
```

2e)

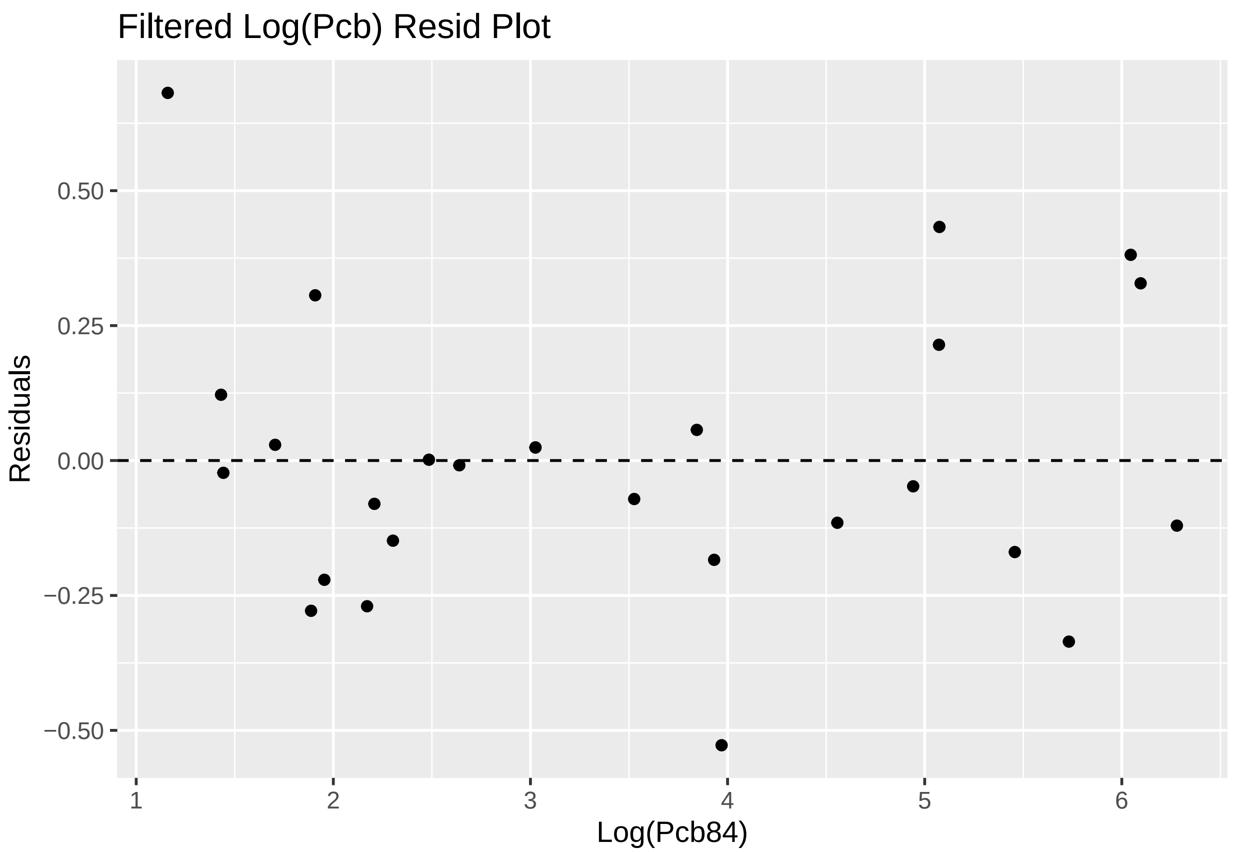
Answer: After removing the outliers the model now appears to better represent the mass of the data. Now, when looking at the residual plot we see a “good” plot, that appears to have a random collection of points around the y = 0 line. This means our assumptions for linearity and constant variance are most likely met. Having the outliers decreased the R^2 value and model slope.

```
#filters out the outlier values found in 2d
pcb_log_filtered_2 <- pcb_log_filtered[-c(4,10),]

#creates scatterplot of log(pcb85) vs log(pcb84) filtered to have no -Inf values and no outleirs
ggplot(data=pcb_log_filtered_2, aes(x = pcb84, y = pcb85)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Filtered logged PCB Levels 1985 vs 1984", x = "log(PCB 1984) (in ppm)", y ="log(PCB 198
```

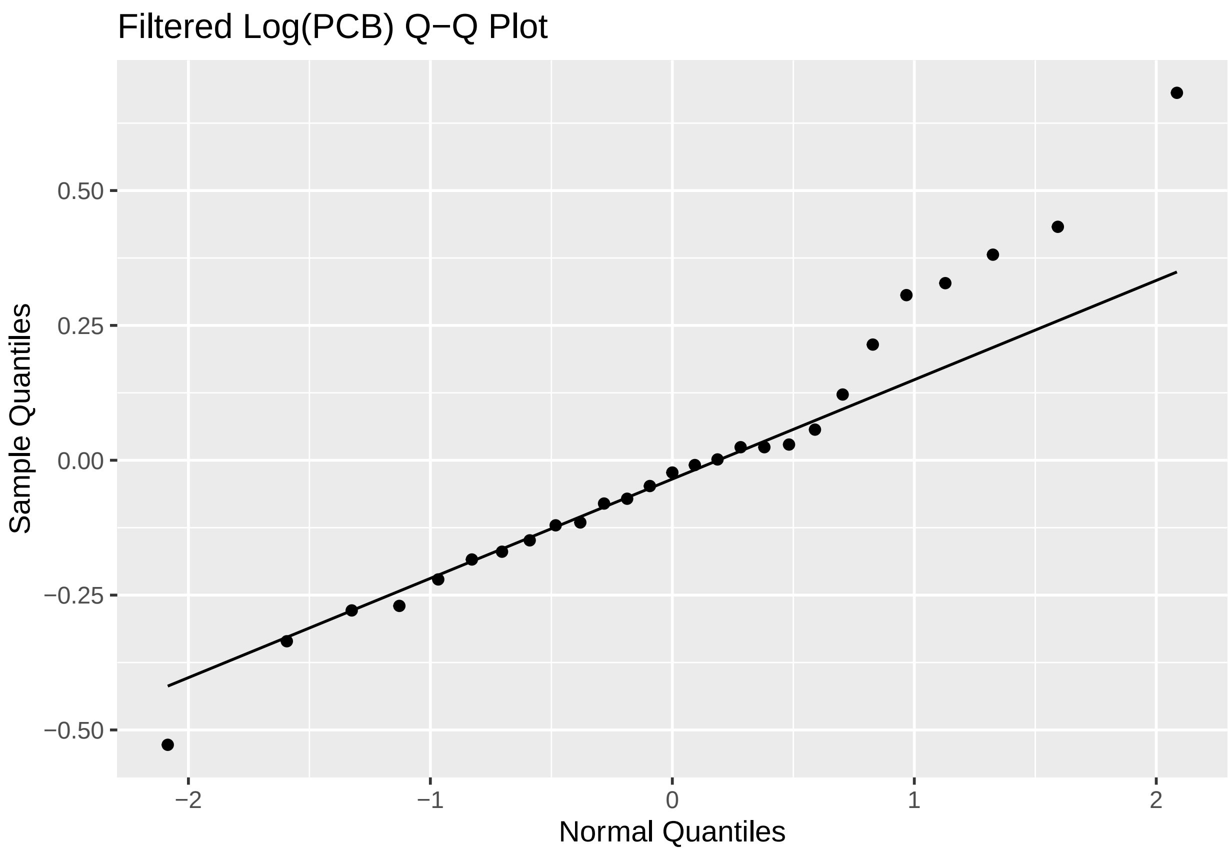
```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```
pcb_log_filtered_fit_2 <- lm(data = pcb_log_filtered_2, pcb85 ~ pcb84)

 pcb_log_filtered_aug_2 <- augment(pcb_log_filtered_fit_2)

 ggplot(pcb_log_filtered_aug_2, aes(x = pcb84, y = .resid)) +
   geom_point() +
   geom_hline(yintercept = 0, linetype = "dashed") +
   labs(title = "Filtered Log(Pcb) Resid Plot", y = "Residuals", x = "Log(Pcb84)")
```



```
ggplot(pcb_log_filtered_aug_2, aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line() +
  labs(title = "Filtered Log(PCB) Q-Q Plot", y = "Sample Quantiles", x = "Normal Quantiles")
```



```
summary(pcb_log_filtered_fit)$r.squared
```

```
## [1] 0.7651357
```

```
summary(pcb_log_filtered_fit_2)$r.squared
```

```
## [1] 0.9732116
```

```
pcb_log_filtered_fit
```

```
##
## Call:
## lm(formula = pcb85 ~ pcb84, data = pcb_log_filtered)
##
## Coefficients:
## (Intercept)        pcb84
##      1.0008       0.7179
```

```
pcb_log_filtered_fit_2
```

```
##
## Call:
## lm(formula = pcb85 ~ pcb84, data = pcb_log_filtered_2)
##
## Coefficients:
## (Intercept)        pcb84
##     0.09249      0.95983
```

2f)

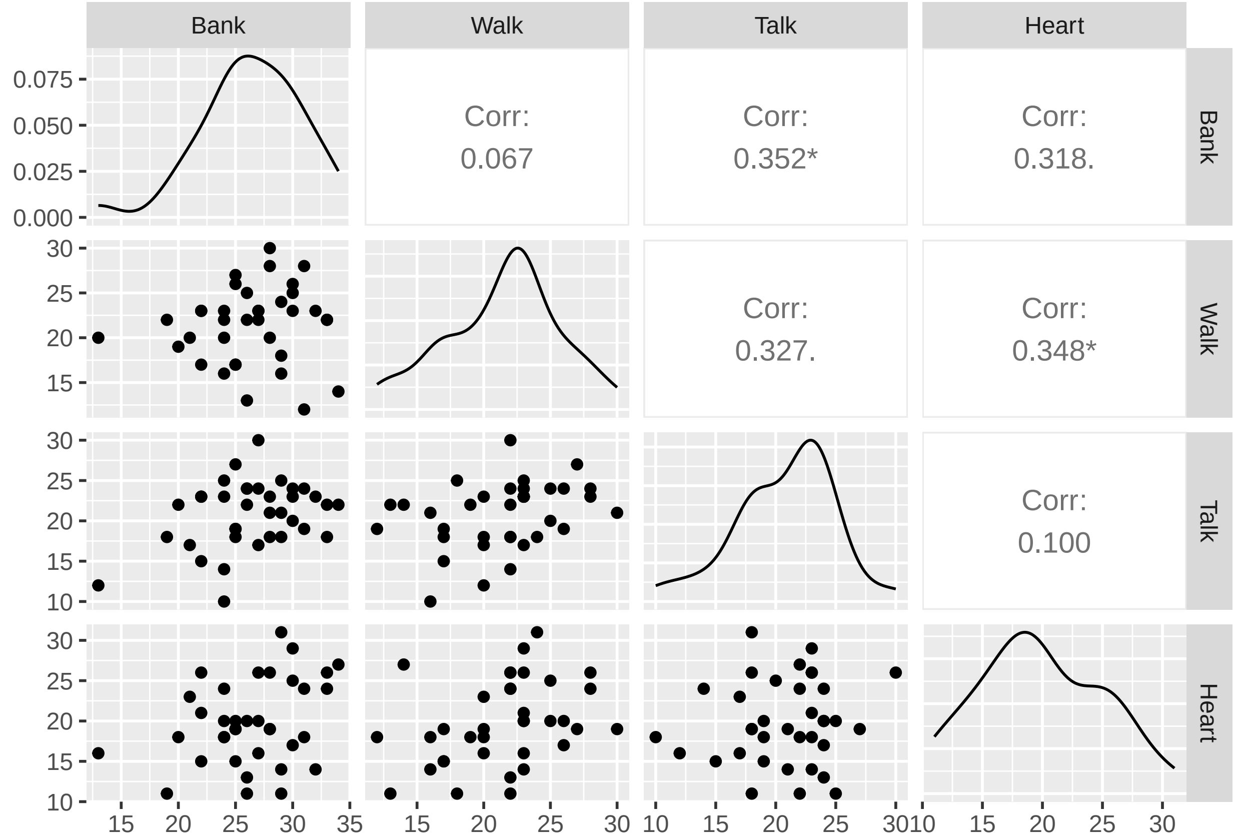
Answer: Interpretation of R^2: The R^2 value is .9732 The interpretation of this that approx 97.32% of the variation in PCB levels in 1985 can be explained by the PCB levels in 1984.

Interpretation of Slope: Doubling the PCB levels in 1984 is associated with an approximately 1.95-fold increase in the PCB levels in 1985 (Math attached as seperate pdf)

3.

3a)

```
#loads data for problem
data(ex0914, package = "Sleuth3")

#creates matrix of scatterplots
ggpairs(ex0914)
```



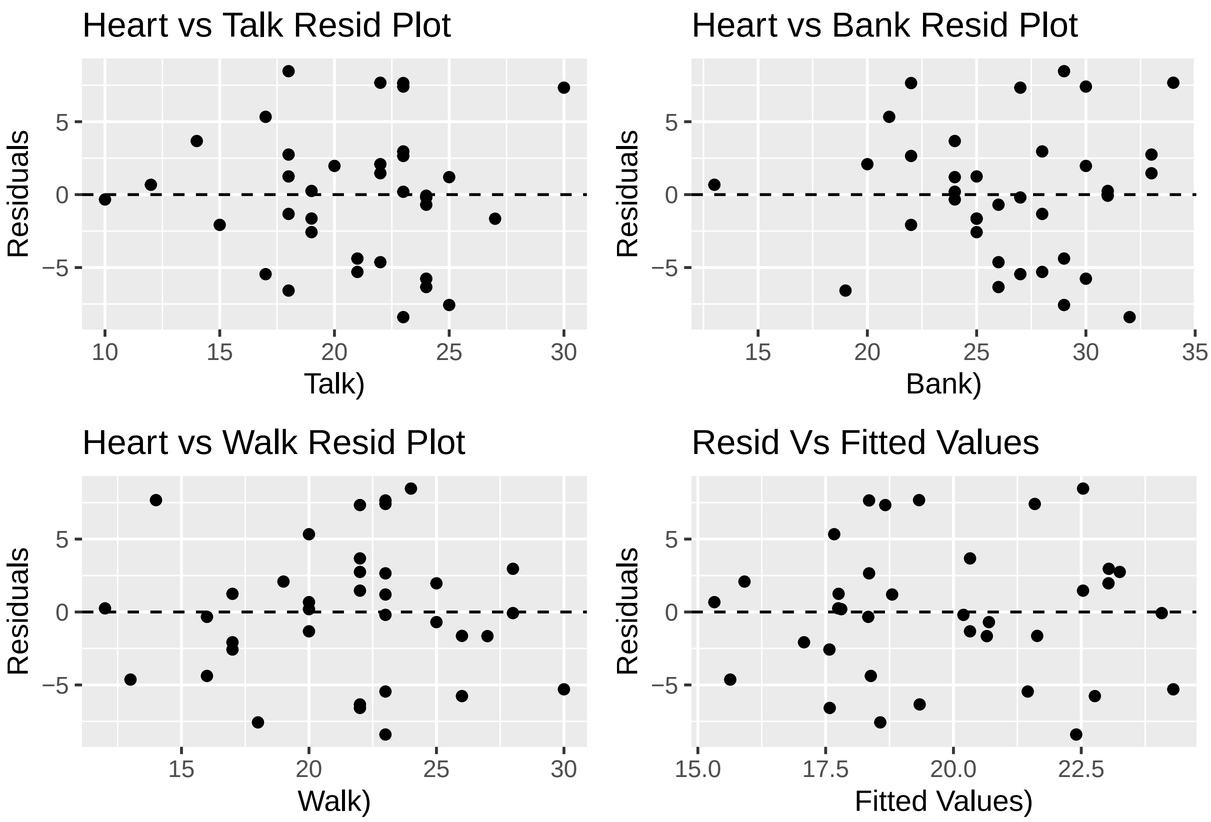
3b)

```
#fits linear model for Heart on Talk, bank, and walk
ex0914_fit <- lm(Heart ~ Talk + Bank + Walk, data = ex0914)
```

3c)

```
ex0914_aug <- augment(ex0914_fit)

#plots the residuals versus each of the predictors in addition to versus the fitted values
plot1 <- ggplot(ex0914_aug, aes(x = Talk, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Heart vs Talk Resid Plot", y = "Residuals", x = "Talk)")
```

```r
plot2 <- ggplot(ex0914_aug, aes(x = Bank, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Heart vs Bank Resid Plot", y = "Residuals", x = "Bank)")

plot3 <- ggplot(ex0914_aug, aes(x = Walk, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Heart vs Walk Resid Plot", y = "Residuals", x = "Walk)")

plot4 <- ggplot(ex0914_aug, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Resid Vs Fitted Values", y = "Residuals", x = "Fitted Values)")

#uses patchwork library to create 2x2 graph matrix
(plot1 + plot2)/(plot3 + plot4)
```


Heart vs Talk Resid Plot

Heart vs Bank Resid Plot

Heart vs Walk Resid Plot

Resid Vs Fitted Values

3d)

Answer: $\hat{Heart}$ = 3.178 - (0.1796)Talk + (0.4052)Bank + (0.4516)Walk SE: 6.3369, 0.2222, 0.1971, 0.2009

Interpretation: Fixing the variable Talk (“obtained by recording responses of postal clerks explain-ing the difference between regular, certified, and insured mail and by dividing the total number of syllables by the time of their response” - Sleuth) and the variable Walk (“walking speed of pedestrians over a distance of 60 feet during business hours on a clear summer day along a main downtown street” - Sleuth), a one unit increase (NOTE: units were not provided in the problem, could be minutes, seconds, etc) in the average time a bank clerk takes to make change for two $20 bills or to give $20 bills for change increases the age adjusted death rate from ischemic heart disease by .4052 (units not given but ostensibly) years.

```
ex0914_fit
```

```
##
## Call:
## lm(formula = Heart ~ Talk + Bank + Walk, data = ex0914)
##
## Coefficients:
## (Intercept)         Talk         Bank         Walk
##      3.1787      -0.1796       0.4052       0.4516
```

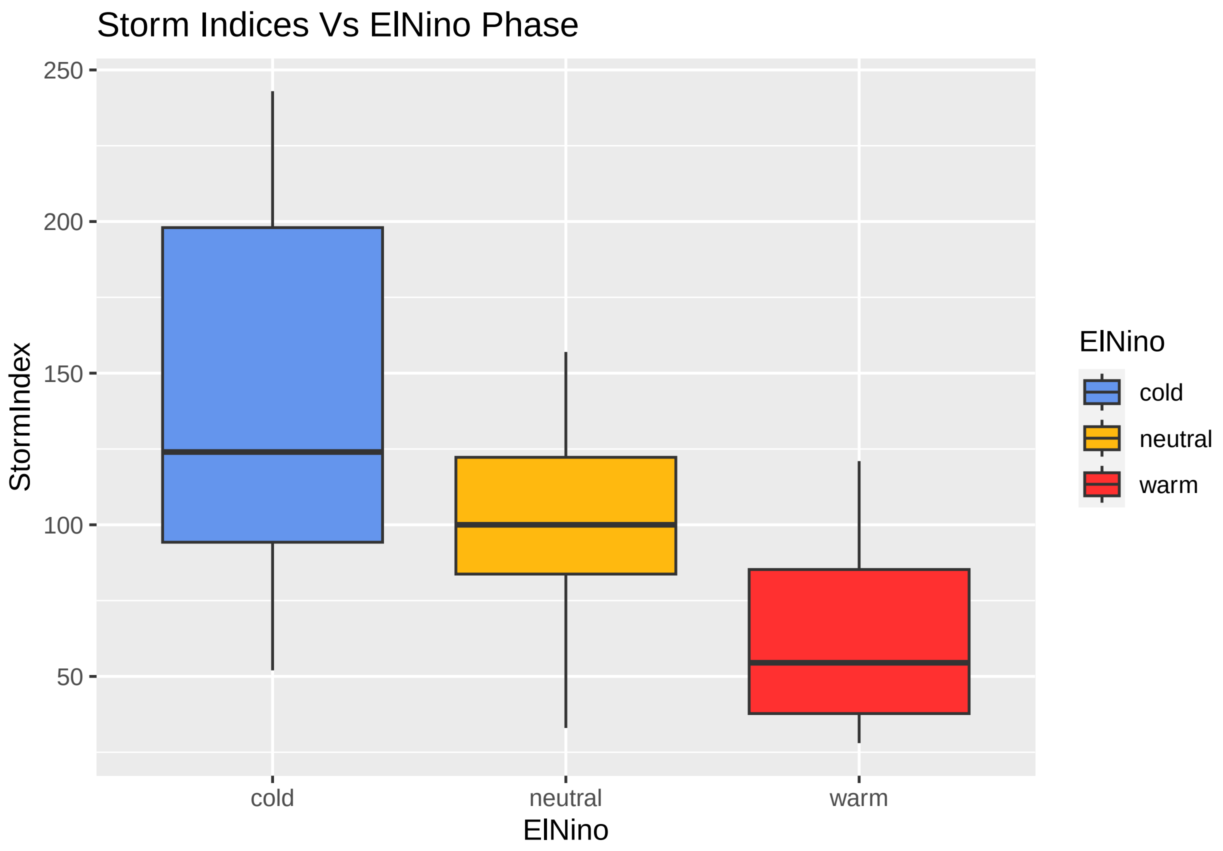
```
summary(ex0914_fit)
```

```
##
## Call:
## lm(formula = Heart ~ Talk + Bank + Walk, data = ex0914)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4014 -3.0263  0.0602  2.6748  8.4646
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.1787     6.3369   0.502   0.6194
## Talk         -0.1796     0.2222  -0.808   0.4249
## Bank          0.4052     0.1971   2.056   0.0480 *
## Walk          0.4516     0.2009   2.248   0.0316 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.805 on 32 degrees of freedom
## Multiple R-squared:  0.2236, Adjusted R-squared:  0.1509
## F-statistic: 3.073 on 3 and 32 DF,  p-value: 0.04162
```

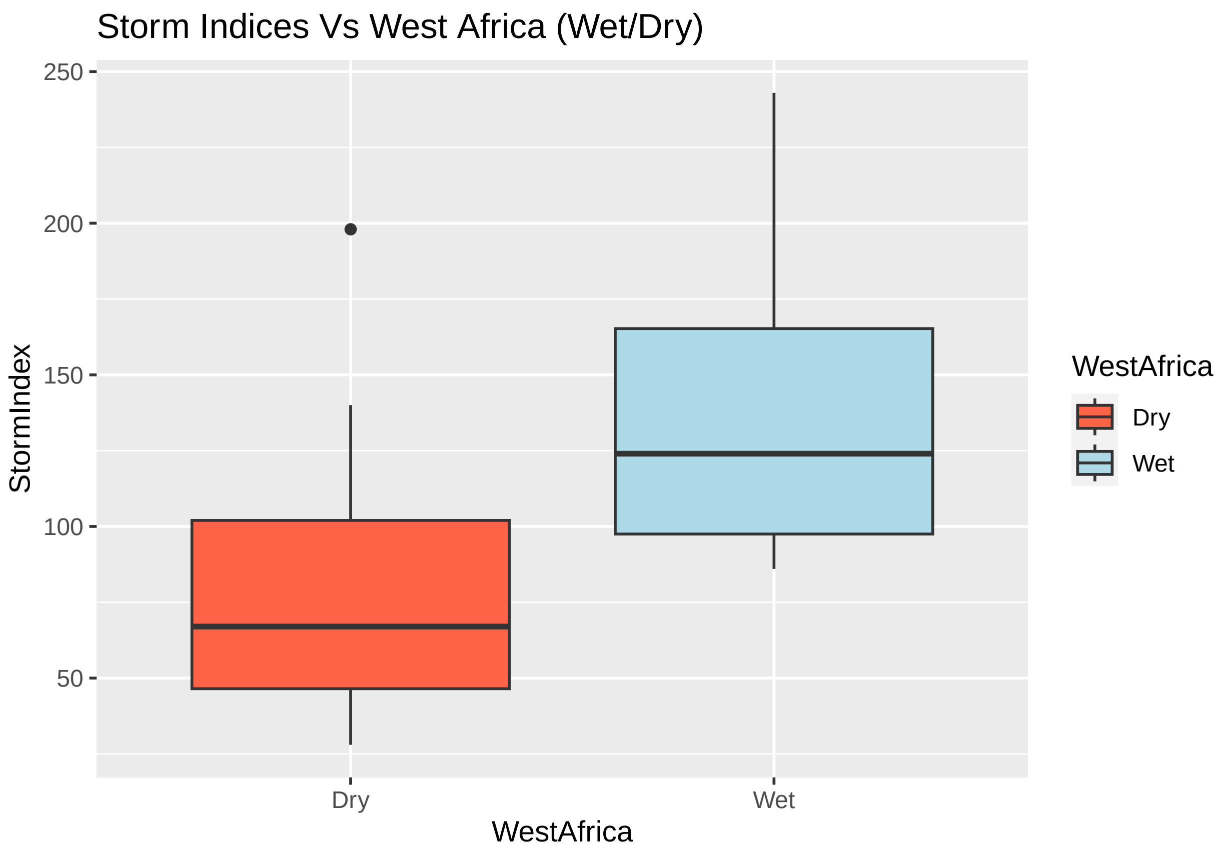
3e) Using the data we cannot provide evidence to the claim that walking quickly causes an increase in the risk of heart disease. This is because the study conducted was observational in nature and not controlled. Claiming that walking faster caused increase risk of heart disease would imply causation, which absolutely cannot be done for observational studies. Also, our data was only taken from US cities, so even if some kind of relationship could be drawn, it would be improper to expand this claim to all people, as our claim can only be about the population the sample was taken from.

4a) **Answer:** In the first graphic, Storm Indices Vs ElNino Phase, it appears that, as if the ElNino year is warmer, the average storm index level decreases, whereas if the ElNino year is colder than the average storm index level decreases. In the second graphic, Storm Indices Vs West Africa (Wet/Dry), it appears that the average Storm Index is higher if West Africa is wet than if it is dry.

```
#loads data for problem
data(ex1028, package = "Sleuth3")

ggplot(data = ex1028, aes(x = ElNino, y = StormIndex, fill = ElNino)) +
  geom_boxplot() +
  labs(title = "Storm Indices Vs ElNino Phase") +
  scale_fill_manual(values = c("cold" = "cornflowerblue",
                                "neutral" = "darkgoldenrod1",
                                "warm" = "firebrick1" ))
```

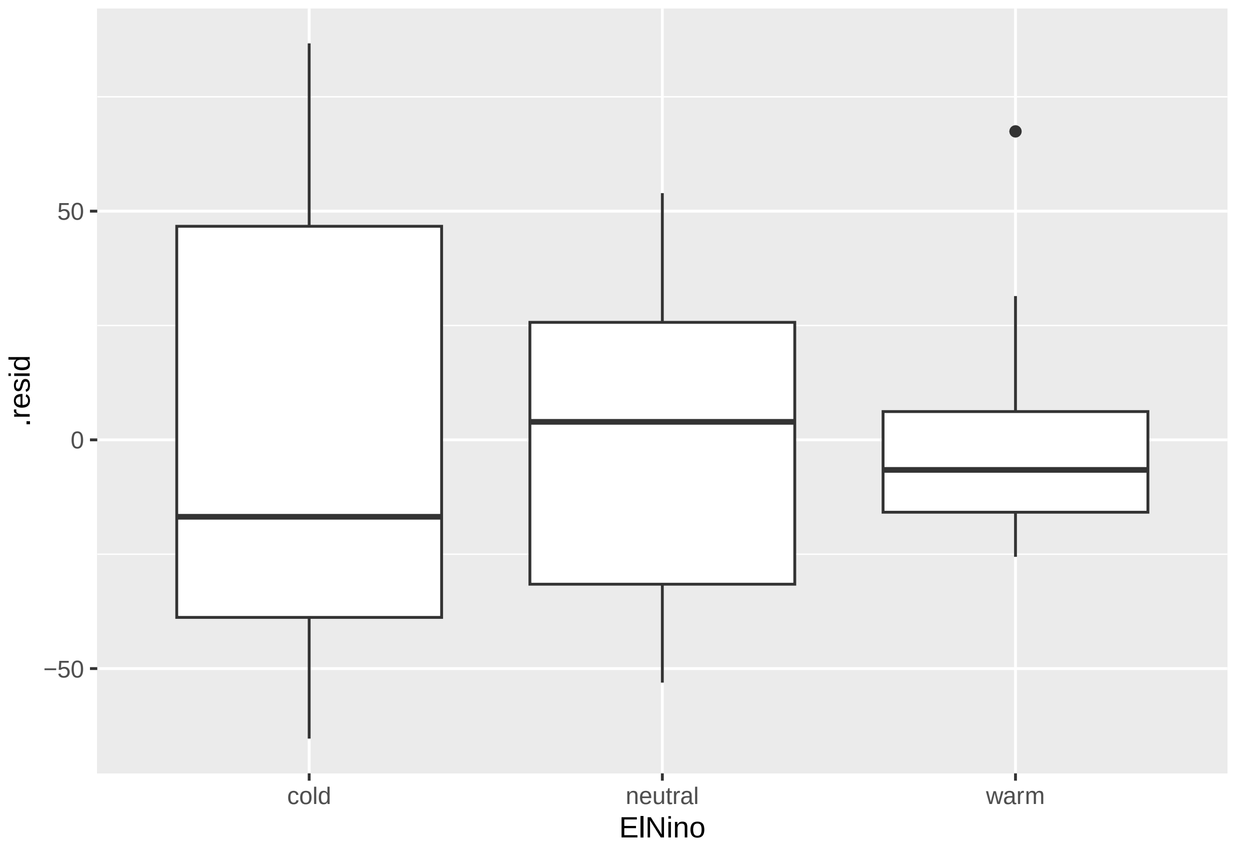


```
ex1028$WestAfrica <- factor(ex1028$WestAfrica, levels = c(0,1), labels = c("Dry", "Wet"))

ggplot(data = ex1028, aes(x = WestAfrica, y = StormIndex, fill = WestAfrica)) +
  geom_boxplot() +
  labs(title = "Storm Indices Vs West Africa (Wet/Dry)") +
  scale_fill_manual(values = c("Wet" = "lightblue",
                               "Dry" = "tomato1"))
```

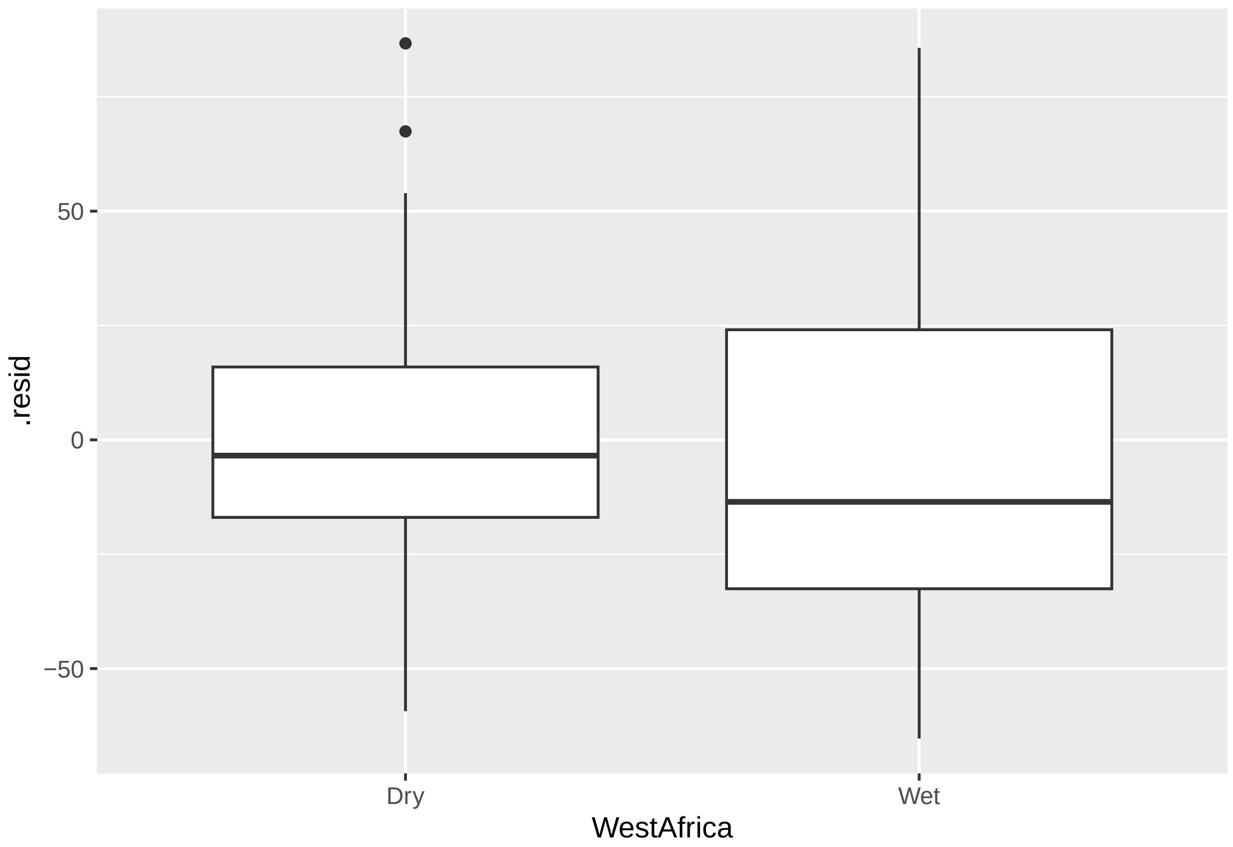


4b) **Answer:** While we haven't expressly look at what characteristics makes a "good"/"bad" residual boxplot, it is my understanding that if our assumptions for a SLR model are met we'd expect to see very similar boxplots across the board, and that the mean should fall on approx y=0. Additionally, we'd expect the sizes of the boxplots to fall relatively in line with eachother, to show that the residuals are the same for all factors and don't follow a trend based on specific factors. Clearly when we group by the ElNino variable these rulings are not met, as the mean residuals appear to differ drastically between the warm, cold, and neutral factors. When grouping by the WestAfrica variable it appears to be significantly more in line with expectations, although it does appear as if the mean residuals of the Wet factor are slightly lower, and the length of the box is slightly greater (meaning the IQR is greater for Wet vs Dry, when they should be pretty similar if our assumptions are met) which is not perfect, and again indicates that our assumptions are probably not met.

```
#fits linear model of StormIndex against ElNino and WestAfrica
ex1028_fit <- lm(StormIndex ~ ElNino + WestAfrica, data = ex1028)

#augments LM
ex1028_aug <- augment(ex1028_fit)

#creates resid boxplot for ElNino variable
ggplot(data = ex1028_aug, aes(x = ElNino, y = .resid)) +
  geom_boxplot()
```



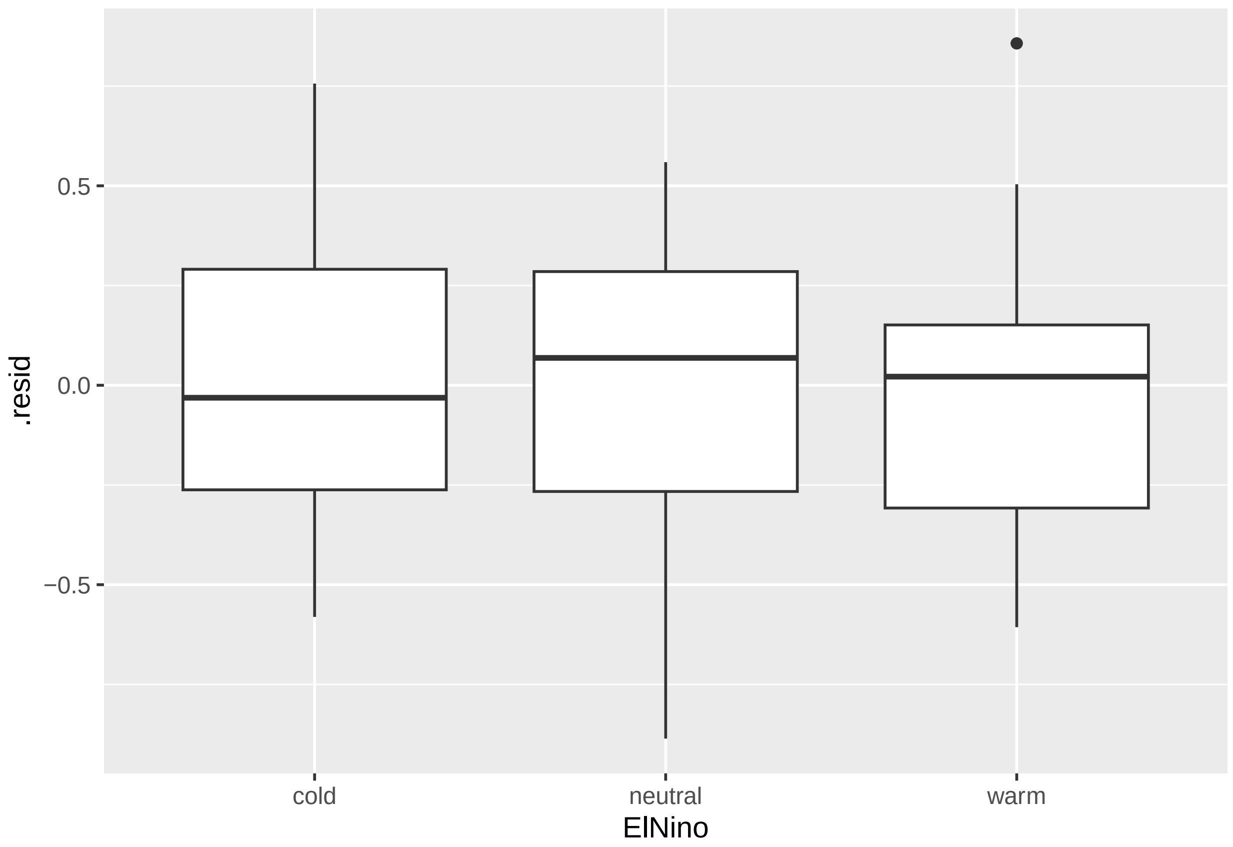
```
#creates resid boxplot for West Africa variable
ggplot(data = ex1028_aug, aes(x = WestAfrica, y = .resid)) +
  geom_boxplot()
```



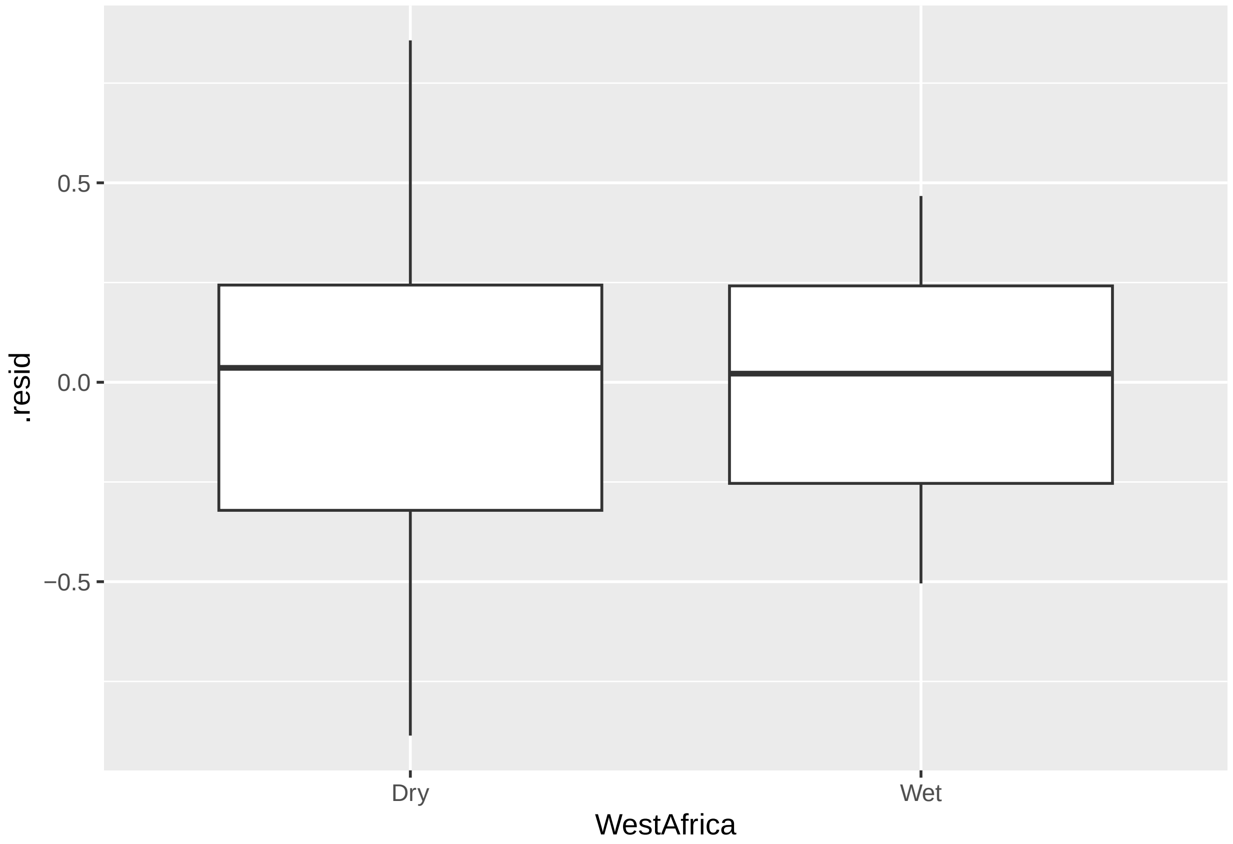
4c)

Answer: These residual boxplots are much more in line with what we'd expect to see given our assumptions are met. Namely, the IQR is relatively equal across factors (for both variables) and they all have mean residual values at approx 0, which is to be expected. While there is still slight variance in the means for each factor of the ElNino variable, these plots are significantly "better" than the previous un-logged ones, as they fall much more in line with whats to be expected for a SLR model.

```r
#fits linear model of log(StormIndex) against ElNino and WestAfrica
ex1028_fit_log <- lm(log(StormIndex) ~ ElNino + WestAfrica, data = ex1028)

#augments LM
ex1028_aug_log <- augment(ex1028_fit_log)

#creates resid boxplot for ElNino variable
ggplot(data = ex1028_aug_log, aes(x = ElNino, y = .resid)) +
  geom_boxplot()
```



```
#creates resid boxplot for West Africa variable
ggplot(data = ex1028_aug_log, aes(x = WestAfrica, y = .resid)) +
  geom_boxplot()
```



4d) **Answer:** The interpretation is: holding whether or not West Africa was Wet or Dry fixed, the estimated difference in mean Storm Index value between a warm and cold El-Nino is -57.76 (aka cold El Nino has 57.76 higher storm index holding Wet/Dry West Africa fixed)

```
ex1028_fit
```

```
##
## Call:
## lm(formula = StormIndex ~ ElNino + WestAfrica, data = ex1028)
##
## Coefficients:
##   (Intercept)  ElNinoneutral     ElNinowarm  WestAfricaWet
##        111.32         -25.25         -57.76          45.99
```