STAT2008 Assignment 1

Student ID: 7205329

Part a - Data Cleaning & Fitting a SLR Model

We are interested in the relationship between `Precent below poverty level V13` and `Per capita income V15`. After removing the rows with missing values, we attain the cleaned data set with the sample size of 420.

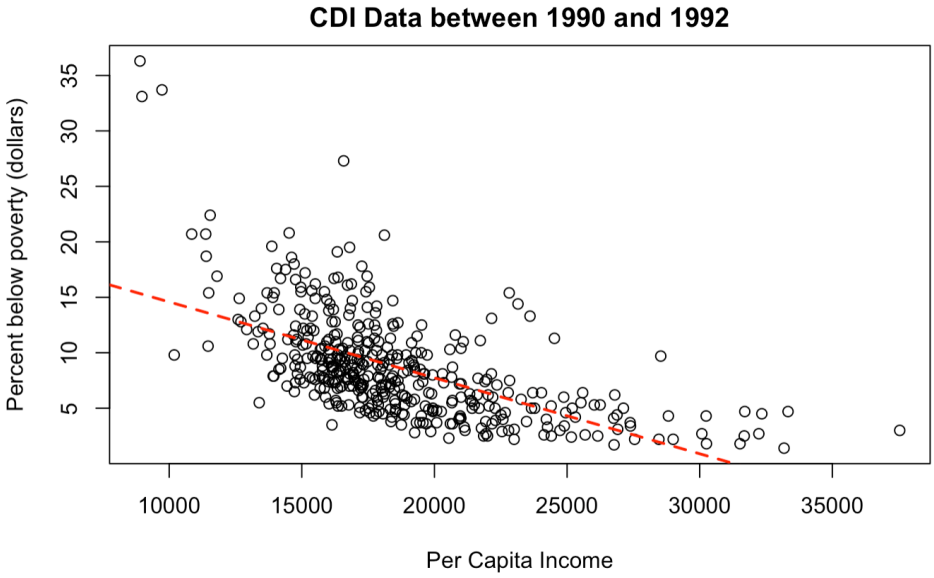
```
mymodel = lm(below_poverty ~ per_cap_income)
summary(mymodel)
```

```
Call:
lm(formula = below_poverty ~ per_cap_income)

Residuals:
   Min     1Q Median     3Q    Max
-6.882 -2.654 -0.613  1.618 20.964

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.143e+01  8.459e-01   25.33   <2e-16 ***
per_cap_income -6.844e-04  4.445e-05  -15.40   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.761 on 418 degrees of freedom
Multiple R-squared:  0.362,     Adjusted R-squared:  0.3604
F-statistic: 237.1 on 1 and 418 DF,  p-value: < 2.2e-16
```

We found the regression coefficients are $b_0 = 21.427$ and $b_1 = -0.000068$, hence the estimated regression function is

$$\hat{Y} = 21.427 - 0.000068X.$$

This estimated line is plotted by



It appears that the fitted function is not a good description of the relationship between `Per capita income` and `Percent below poverty level`.

As $b_0 = 21.4269$, it is estimated that when `Per capita income` equals to 0, the `Percent below poverty level` is approximately 21.43% with a standard deviation of 0.8459%. At the same time, we have the slope coefficient $b_1 = -0.000068$ which indicates that one dollar increase in the `Per capita income` corresponds to 0.00068 unit decrease in the mean of `Percent below poverty level`, with a standard deviation of 4.445×10^{-5} percent.

Part b - Construct the ANOVA Table

```
anova(mymodel)
```

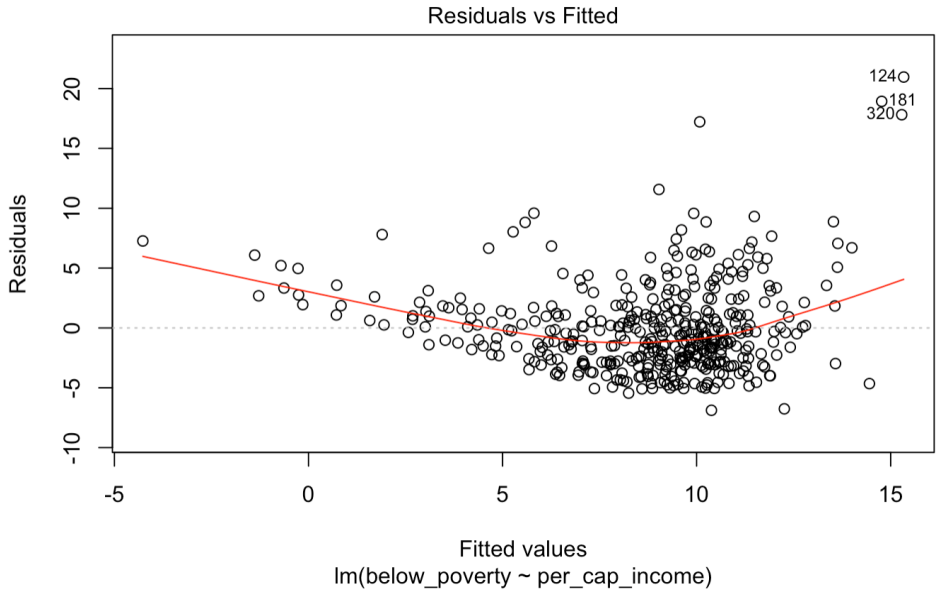
```
Analysis of Variance Table

Response: below_poverty
                Df Sum Sq Mean Sq F value    Pr(>F)
per_cap_income   1 3354.7  3354.7  237.13 < 2.2e-16 ***
Residuals      418 5913.5    14.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Given the following hypotheses:

$$H_0 : \frac{SSR}{SSE} = 1, \quad H_a : \frac{SSR}{SSE} > 1$$

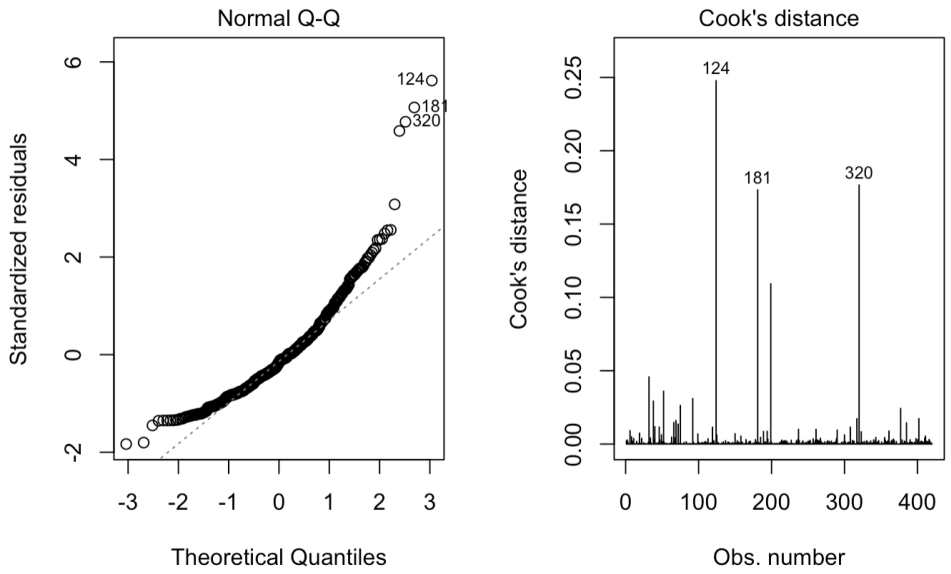
From the ANOVA table, we found our F-value is 273.13 and its corresponding p-value is much smaller than $\alpha = 0.05$. This indicates that there is enough evidence for us to reject the null hypothesis in favour of the alternative. Therefore, we conclude that SSR is larger than SSE which would further suggest that $\beta_1 \neq 0$. Hence, the model involving `Per capita income` is explaining a significant proportion of the variability in `Percent below poverty level`.

Part c - Diagnostic Checks



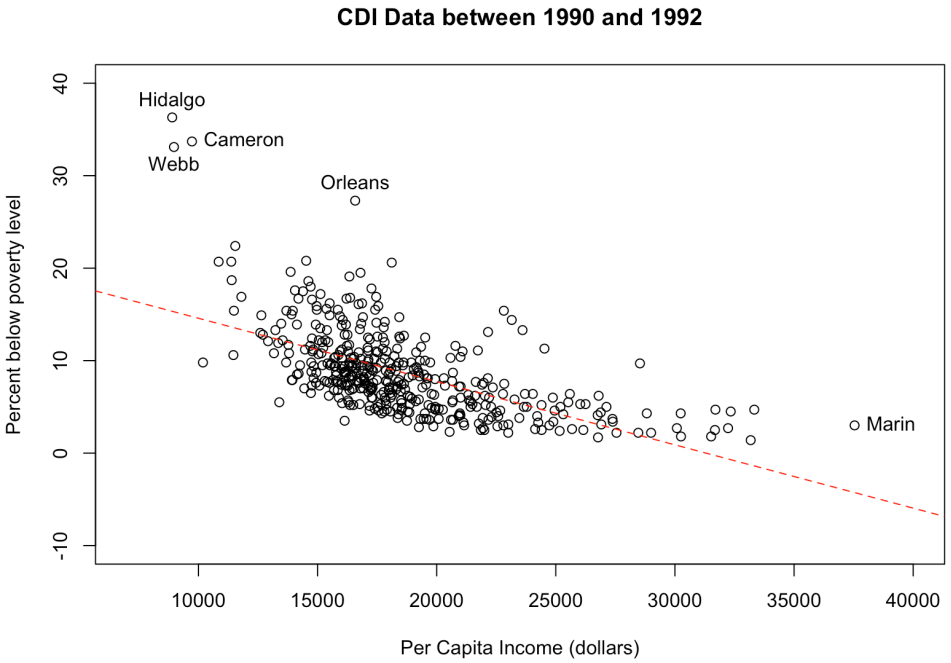
From the "Residuals vs Fitted" plot, it is clear that there are three major problems:

1. **Non-constant variance** (variation in the residuals grows as the fitted values increases from -5 to 10),
2. **A violation in the linearity assumption** (a curvature shape is apparent), and
3. **A number of outliers** (Observation 124, 181, and 320).



The "Normal Q-Q plot" suggests that the normality assumption on the error terms is not met. In fact, the residuals are drawn from a right-skewed distribution. Furthermore, the "Cook's Distance" plot confirms the three influential observations 124, 181, and 320.

Part d - Identify the outliers



We have identified the counties Hidago, Cameron, Webb, Orleans and Marin to be the outliers of our model. By further investigation into the data, we found that Hidago is a county in Texas with a population of more than 380,000 people in 1990.

Its per capita income was 8899 dollars with 36.3% of the population was below the poverty level. Another outlier is Marin, a county in California with a population of around 230,000 in 1990. In that year, Marin's per capita income was 37,541 dollars with 3% of its population was below the poverty level.

Part e - Transforming Variables

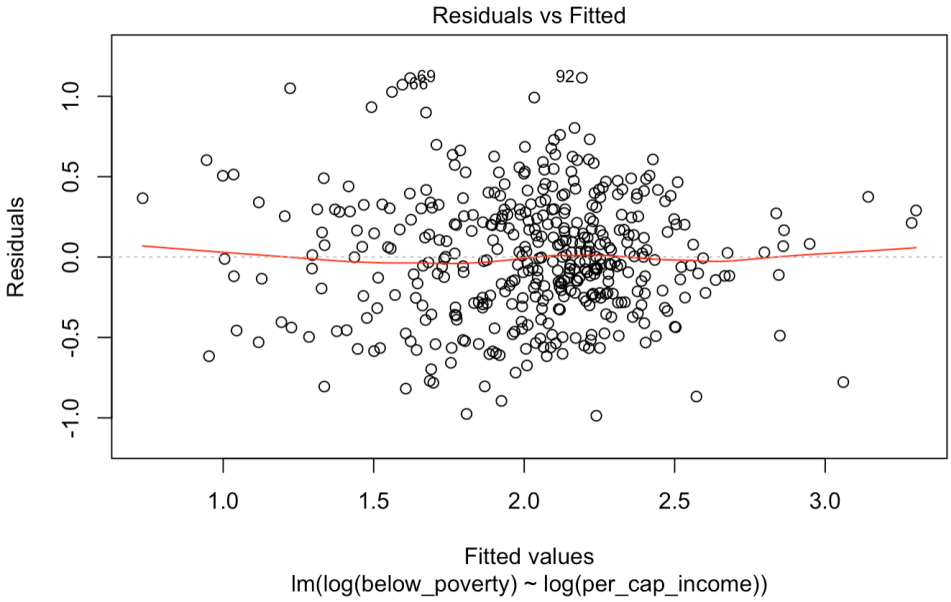
After experimenting, it was found that the natural log transformation applied to both variables of `Per capita income` and `percent_below_poverty_level` led to a satisfactory linear fit (based on scatter plots and correlation coefficients). Our transformed model is now

$$log(Y) = \beta_0 + \beta_1 \times log(X) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

where Y: `Percent below poverty level` and X: `Per capita income` (dollars).

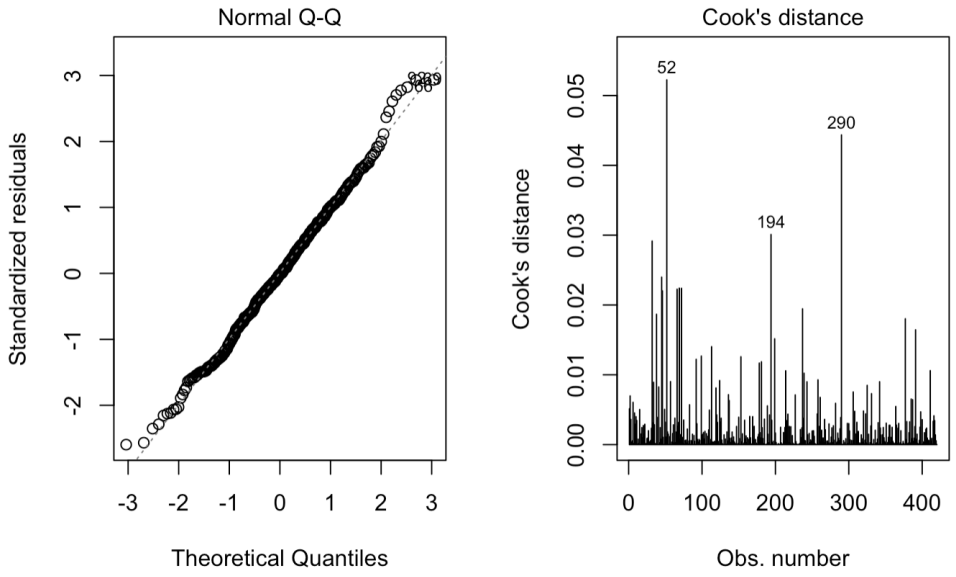
The correlation coefficient for this transformed model is -0.7026 (4 dp). The negative sign suggests that this relationship is negative, i.e. as the `log(Per capita income)` increases, the `Log(Percent below poverty level)` decreases.

Part f - Diagnostic Check of Transformed Model



In the "Residuals vs Fitted" plot for transformed variables, we do not see any clear pattern comparing to prior transformation. The residuals are randomly distributed

around the line h = 0 and cluster towards the middle of the plot. Outliers is also no longer a problem here.



While there are three points that stand out with higher Cook's distances (Observation 52, 194, and 290), they are not significantly larger than other data points considering the small scale of the y-axis. The "Normal Q-Q" also shows satisfaction in the normal assumption. Thus, we can conclude that this is a more appropriate model and continue our investigation.

Part g - Test whether the slope is smaller than -1

```
summary(mymodel_transformed)
```

```
Call:
lm(formula = log(below_poverty) ~ log(per_capita_income))

Residuals:
    Min      1Q  Median      3Q     Max
-0.9869 -0.2525 -0.0049  0.2602  1.1153

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           19.53439    0.86730   22.52   <2e-16 ***
log(per_capita_income) -1.78501    0.08842  -20.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3807 on 418 degrees of freedom
Multiple R-squared:  0.4937,    Adjusted R-squared:  0.4925
F-statistic: 407.6 on 1 and 418 DF,  p-value: < 2.2e-16
```

A one-tailed test was conducted on β_1 of the transformed model, assuming the hypotheses are $H_0 : \beta_1 \geq -1, \quad H_a : \beta_1 < -1.$ According to the summary table above, the slope coefficient b_1 is -1.785 and its corresponding standard error is 0.088. Hence, we have the test statistics

$$t^\star = \frac{b_1 - (-1)}{s(b_1)} = \frac{-1.785 - (-1)}{0.088} \approx -8.8786.$$

Using $\alpha = 0.05$, the critical value is $t(0.05, 418) = -1.65$. As the test statistics is within the rejection region $(-\infty, -1.65)$, we can reject the null hypothesis in favour of the alternative and conclude that the slope coefficient of the transformed model is indeed smaller than -1.

Part h - Estimated Error Variance and R^2

With the selected model, the estimated error variance is the MSE = 0.145 according to the ANOVA table of the transformed model. That is, the variation in the `log(percent below poverty level)` around the mean for different `log(per capita income)` is estimated to 0.145.

```
Analysis of Variance Table

Response: log(below_poverty)
                        Df Sum Sq Mean Sq F value
log(per_capita_income)   1 59.086  59.086  407.59
Residuals              418 60.595   0.145
                          Pr(>F)
log(per_capita_income) < 2.2e-16 ***
Residuals
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

To calculate R^2, we determine SSTO which can be derived from the same ANOVA table. Given that SSE = 60.595 and SSR = 59.086, we have

$$SSTO = SSE + SSR = 119.681,$$

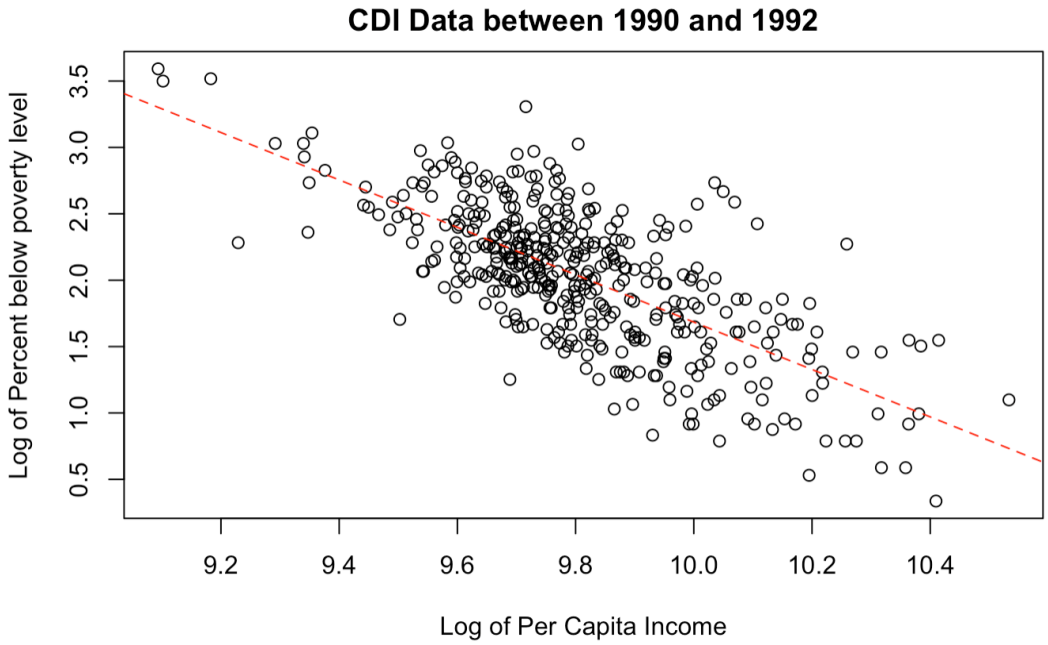
and the coefficient of determination

$$R^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{60.595}{119.6817} \approx 0.49.$$

This R-squared value is consistent with part e as

$$r^2 = (-0.70)^2 = 0.49 = R^2.$$

This number can be interpreted as approximately 50% of the variation in the `log(Per capita income)` can be explained by the `log(Percent below poverty)`. While the linear association is strong, the moderate R^2 and the scatter plot (as shown below) show that there is a lot of variation around the new fitted line (50% of the data points fall on the new regression line).



Part i

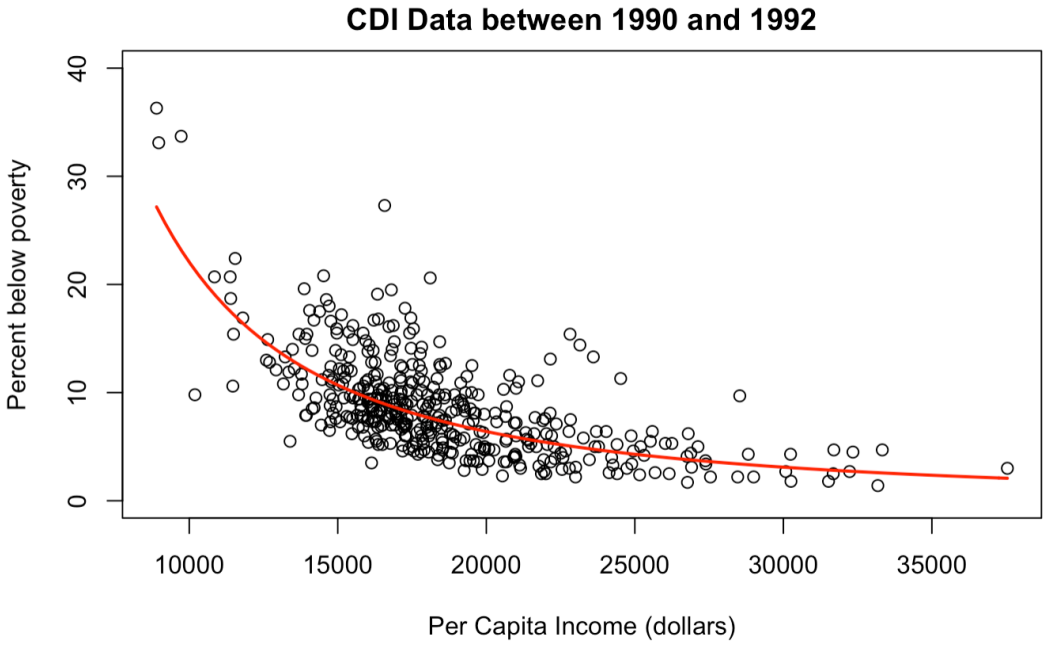
The new SLR model in terms of original untransformed variables is

$$\hat{Y} = e^{b_0} \times X^{b_1} = e^{19.534} \times X^{-1.785}$$

When the value of X is multiplied by a factor of k, the value of $\hat{Y}$ is multiplied by a factor of k^{b_1}. Indeed,

$$\hat{Y}' = e^{b_0} \times (kX)^{b_1} = k^{b_1} e^{b_0} \times (X)^{b_1} = k^{b_1} \hat{Y}$$

Below is the scatter plot of X and Y on the original scale, along with the new fitted model. We see that the new regression function is a much better fit for our data.



Part j

Let us find a 90% confidence interval for the mean of `Percent below poverty level` for a county whose `Per capita income` was 25,000 dollars in 1990. We found the point estimate

$$\hat{Y}_h = e^{19.53} + (25000)^{-1.785} = 4.30,$$

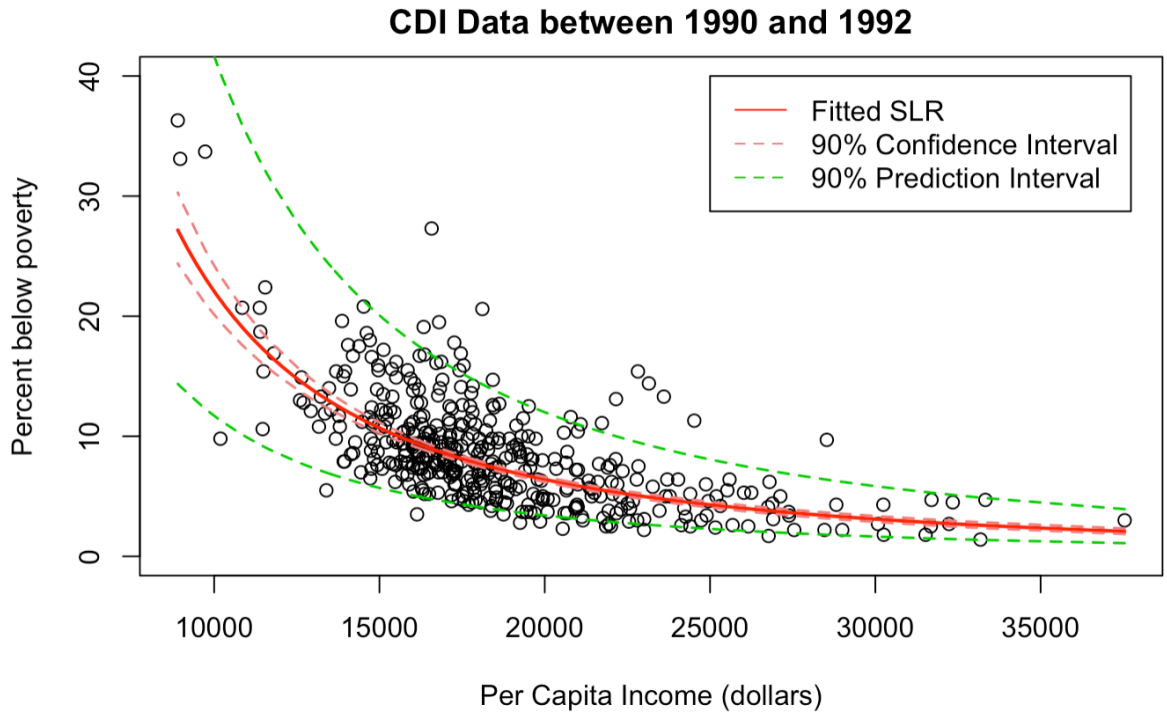
and its corresponding estimated standard deviation $s(\hat{Y}_h) = 0.034$. For a 90 percent confidence coefficient, we require $t(.95, 420) = 0.0557$. Hence our confidence coefficient .90 is given by

$$4.24 \leq E[Y_h] \leq 4.35$$

It can be concluded that with $\alpha = 0.10$, the mean number of `Percent below poverty level` when the `Per capita income` is 25,000 dollars is between 4.24 and 4.35 percent. This estimate is valid as 25,000 is within the range of our original data. It appears that the confidence interval is fairly narrow which indicates that there is a linear association between the transformed variables.

However, as we plot the 90% CI, we see that interval is very tight and many of the observations lie outside of the interval. That is, there is a lot of variability in this relationship (which we have seen in part h) and thus precise prediction cannot be

made of the `Percent below poverty` from `Per capita income`. Indeed, we find the 90% Prediction Intervals cover almost the entire range of the observations.



Finally, it can be concluded that while there is strong indication of a linear relationship between `log(Per capita income)` and `log(Percent below poverty level)`, there exists significant variability in this relationship. That is, there might be other variables contributing to changes in `Percent below poverty level` in conjunction with `Per capita income`. Furthermore, `Per capita income` should not be used solely to represent individual's household income as it is sensitive to outliers. Therefore, further investigation must be made before we can made any definite conclusions.