MA362: Linear Regression & Time Series
Project 1 | Indiana Demographics & Income
Nathan Gonyo, Aaron Neiger, Nick Donnelly
Dr. Rasitha Jayasekare

Table of Contents

Introduction & Background Information on Dataset

The dataset to be used in this project is entitled "indiana2016". The dataset consists of 28,304 randomly selected Indiana citizens who are between the ages of 18-75 and are employed. The data is from the 2016 American Community Survey of the Census Bureau. For the purposes of this course project, we are to include at least 2 quantitative predictor variables and at least 1 qualitative predictor variable. We will be using 3 quantitative predictor variables and 2 qualitative predictor variables. Here is a list of ALL the variables contained in the dataset. NOTE: we will NOT be using all variables in the dataset in our project.

AGE: continuous variable denoting age of individuals

SEX: 1= female; 0 = male

MARST: 0=never married/widowed/divorced/separated; 1=married

HEALTHINS: 0= no health insurance; 1= have health insurance

EDUC: number of years of education

HRSWORK: number of hours worked in week prior to survey

INCWAGE: wage income in dollars

INCTOTAL: total income in dollars

USCIT: 0 = non-US citizen: 1= US citizen

COW (class of worker): 0=private, for-profit; 1= private, non-profit; 2=government;

3= self-employed

TYPEHEALTHINS: 0=no insurance; 1= private insurance; 2=public insurance

For our project, the continuous response variable will be INCTOTAL, the total income in dollars of each individual. The three quantitative predictor (x) variables we will be using to predict income (y) are AGE, EDUC, HRSWORK, The two qualitative predictor (x) variables we will be using to predict income (y) are SEX & MARST.

Preprocessing & Outlier Removal: Summary of Variables

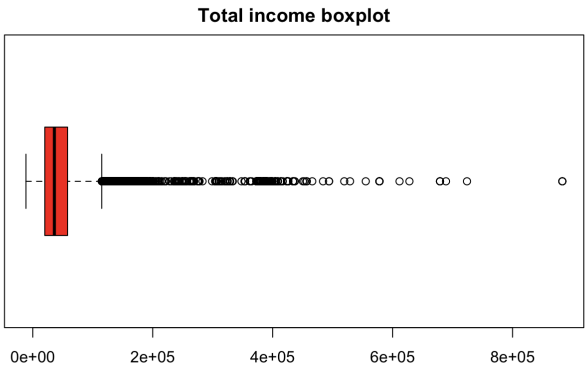
To start preprocessing, we gathered some basic summary statistics of each variable to be used in our model, to get a better understanding of each variable. Recall the variables which to be used in our model are INCTOTAL (predicted variable), AGE, EDUC, HRSWORK, SEX, MARST (predictor variables).

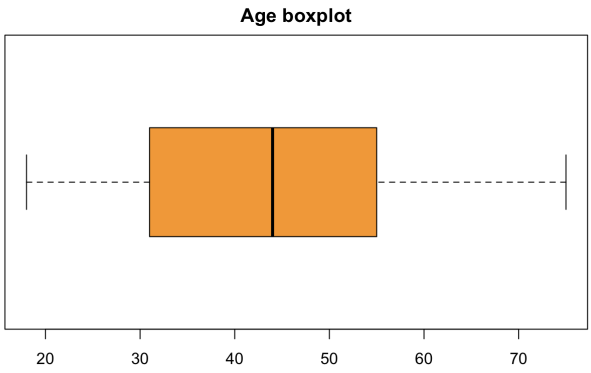
	Numerical Descriptions of Variables			
	Min	**Median**	**Mean**	**Max**
INCTOTAL	-11400	36,000	47242	883000
AGE	18	44	43.28	75
EDUC	6	12	12.37	18
HRSWORK	1	40	39.45	99
SEX	0	0	.4764	1
MARST	0	0	.5945	1

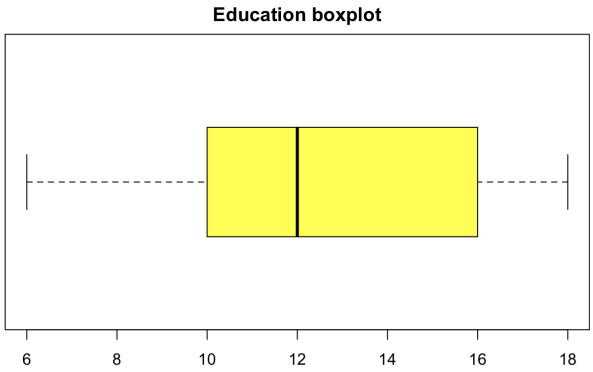
Note: SEX & MARST are binary variables taking on only values of 0 or 1 so summary statistics are essentially useless.

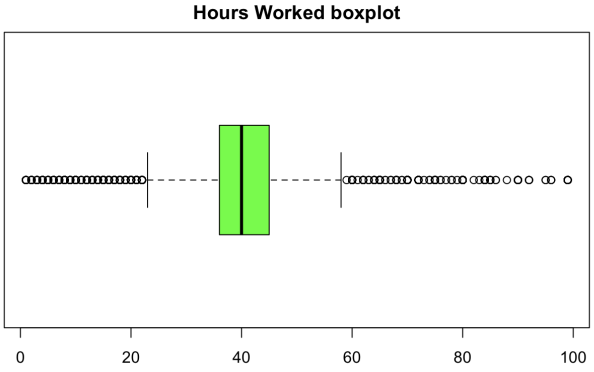
Preprocessing & Outlier Removal: Boxplots of Variables

The next step in preprocessing was to examine the presence of outliers in the dataset. We did this by generating boxplots for each variable to see whether there are outliers. A variable with outliers will contain dots on the extreme left or right side of the boxplot. Note that a box plot with equal lines to each. side is uniformly distributed. A boxplot with a longer line on the right signifies a right skew. Opposite is true for a left skew. **NOTE: SEX & MARST are binary variables taking on only values of 0 or 1 so they essentially cannot contain outliers and boxplots will not be provided.**

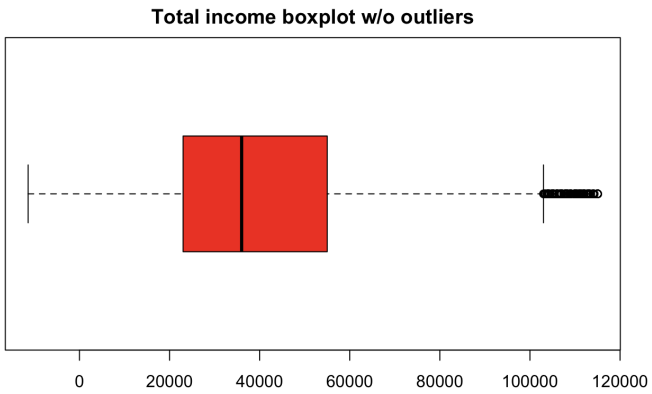








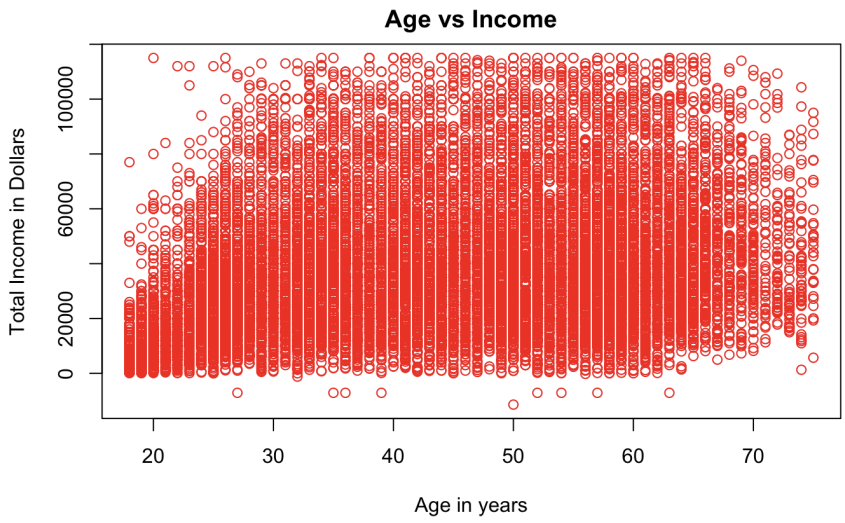
We can see from our boxplots that the variables total income and hours worked have a significant number of outliers, which makes sense for each variable given what they represent. For age and education, there are no outliers present which also makes sense as unless there was an error in data collection, all age and education values should be relatively close.



NOTE: After the first round of outlier removal, there are still some outliers present in the data. We cannot justify going through another round of outlier removal as we know these outliers are realistic numbers that actually occurred, not errors in data collection or dataset creation. It makes perfect sense that there would be people who work less than 10 or more than 80 hours per week. It also makes sense that there would be people that make more than 100,000 dollars per year, even if these are considered outliers from a purely statistical perspective. That being said, we can see that the outliers are less extreme after outlier removal, especially in the case of total income.

Scatterplots & Correlation Coefficients of Variables

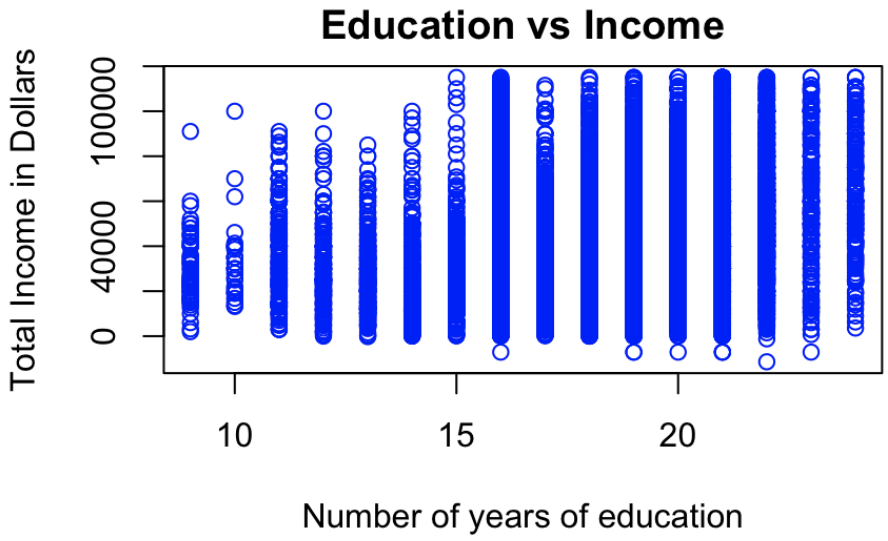
We will now create scatterplots for each pair of non-binary variables (excluding SEX & MARST). In each scatter plot, total income (INCTOTAL), will be the dependent (y) variable. We will also list the linear correlation coefficient between each variable and inctotal.



Linear Correlation Coefficient of (age, inctotal): 0.2671794

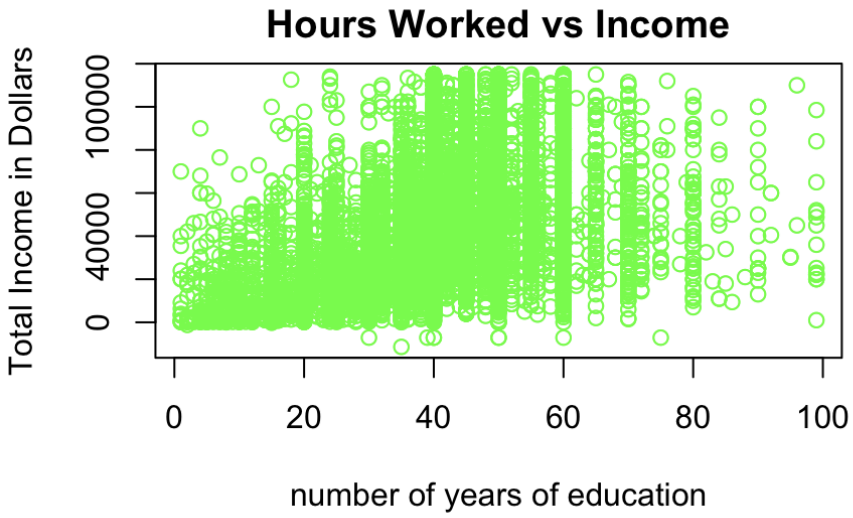
Here is the scatter plot of income vs age. The linear correlation coefficient shows a weak positive linear relationship between total income and age. Generally the older someone is the more likely they are to have promotions and experience in their field. This results in getting paid more for the work they do.

Although, this is not always the case. This is shown in the weaker positive relationship between the two variables.



Linear Correlation Coefficient of (educ, inctotal): 0.2628786

Here is the scatter play of total income vs number of years of education. The linear correlation coefficient shows a weak positive relationship between total income and number of years of education. This makes sense due to the fact that generally people make more money for their amount of higher education. Again, this is not always the case resulting in a weaker positive relationship between income and years of education.



Linear Correlation Coefficient of (hrswork, inctotal): 0.416332

Here is the scatter plot of total income vs hours work. The linear correlation coefficient shows a moderate positive relationship between hours worked per week and total income. The more someone works on a weekly basis the more they get paid makes sense. Whether they are working and getting paid for overtime or just for the more hours they work.

Creation of Training & Testing Subsets

Before we create our model, we will divide the dataset into training (80%) and testing (20%) subsets. 80% of Rows will be randomly selected to create the train dataset and moved into a new dataset. The remaining 20% will be used as the testing dataset. Since there are 22805 rows in our dataset, we need to take out (.8*22805= approx 18240) rows. We will randomly generate 18240 row numbers and extract those rows from Indiana2016.

Creation of Multiple Linear Regression Model

Regression Equation: inctotal = educ + age + hrswork + sex + marst

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-53668.39	1239.06	-43.31	<2e-16	***
TrainIndiana$educ	2542.98	56.19	45.25	<2e-16	***
TrainIndiana$age	405.63	10.93	37.10	<2e-16	***
TrainIndiana$hrswork	774.62	13.97	55.45	<2e-16	***
TrainIndiana$sex	-9543.18	295.23	-32.33	<2e-16	***
TrainIndiana$marst	5849.39	309.25	18.91	<2e-16	***

Residual standard error: 19220 on 18234 degrees of freedom

Multiple R-squared: 0.3537, Adjusted R-squared: 0.3536

F-statistic: 1996 on 5 and 18234 DF, p-value: < 2.2e-16

Interpretation of Parameters

BetaHat0 = -53668.39. There cannot be a worker with zero education, age, prior hours worked. Therefore, the interpretation of BetaHat0 is not meaningful.

BetaHat1 = 2542.98. For every one year increase in education, the average income total will go up by $2,542.98.

BetaHat2 = 405.63. For every one year increase in age, the average income total will go up by $405.63.

BetaHat3 = 774.62. For every one increase in the number of hours worked in the week prior to the survey, the average income total will go up by $774.62.

BetaHat4 = -9543.18. When the sex variable changes from male to female, the average income total will decrease by $9,543.18

BetaHat5 = 5849.39. When the marriage variable changes from not married to married , the average income will increase by $5,849.39

Standard Error of Regression:

Our model had a standard error of 19220. The average difference between observed total income and estimated total income is $19,220.

Coefficient of Determination

Our model has an R-squared value of 0.3537. Our coefficient of determination is 35.37%. That is, only 35.03% of the sample variation of total income can be explained by our model. Based on the coefficient of determination, our model is not favorable. However, since we have no model to compare this to, this number does not necessarily mean our model cannot or should not be used.

Coefficient of Variation (CV)

CV = S / Ybar = 19360 / 47242 = 0.4098

The estimated standard error of regression is about 41% of the sample mean total income. Since our model's CV is above 10%, the regression model is not favorable. However, since we have no model to compare this to, this number does not necessarily mean our model cannot or should not be used.

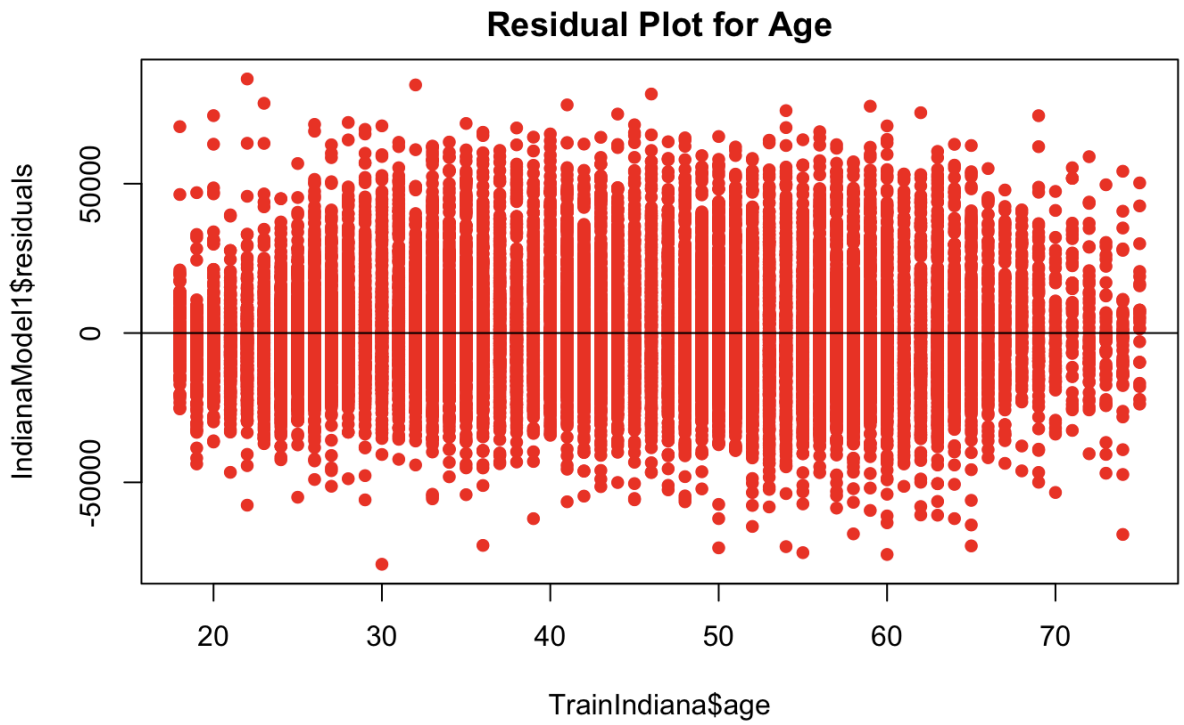
Residual Analysis (checking regression assumptions) & Interpretation

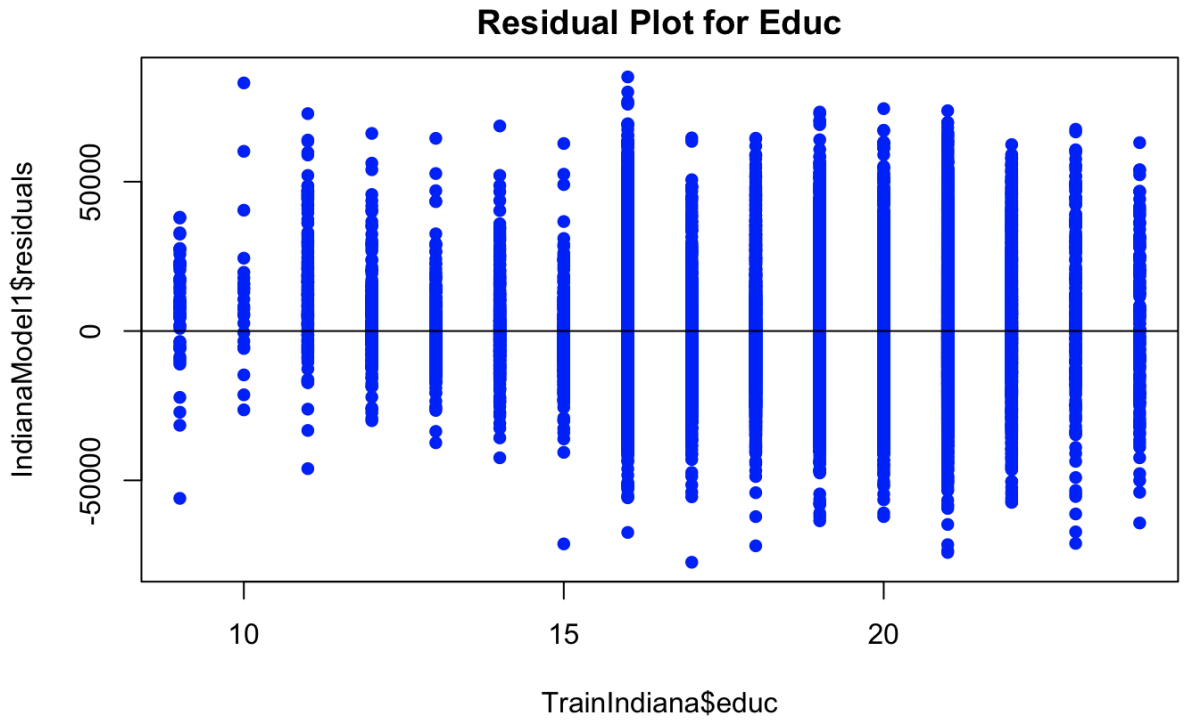
Recall the five assumptions of the MLRM model:

1. Error term has mean of, and therefore expected value of, 0
2. Error term exhibits constant variance, ie heteroscedasticity is not present
3. Error term normally distributed
4. Observations of response variable are independent from each other
5. Absence of multicollinearity, ie predictor variables do not contain strong linear relationships

Analysis of Residual Scatter Plots (Checking Assumptions 1 & 2)

The first two assumptions of the multiple linear regression model will be checked through the use of residual plots, which are below.

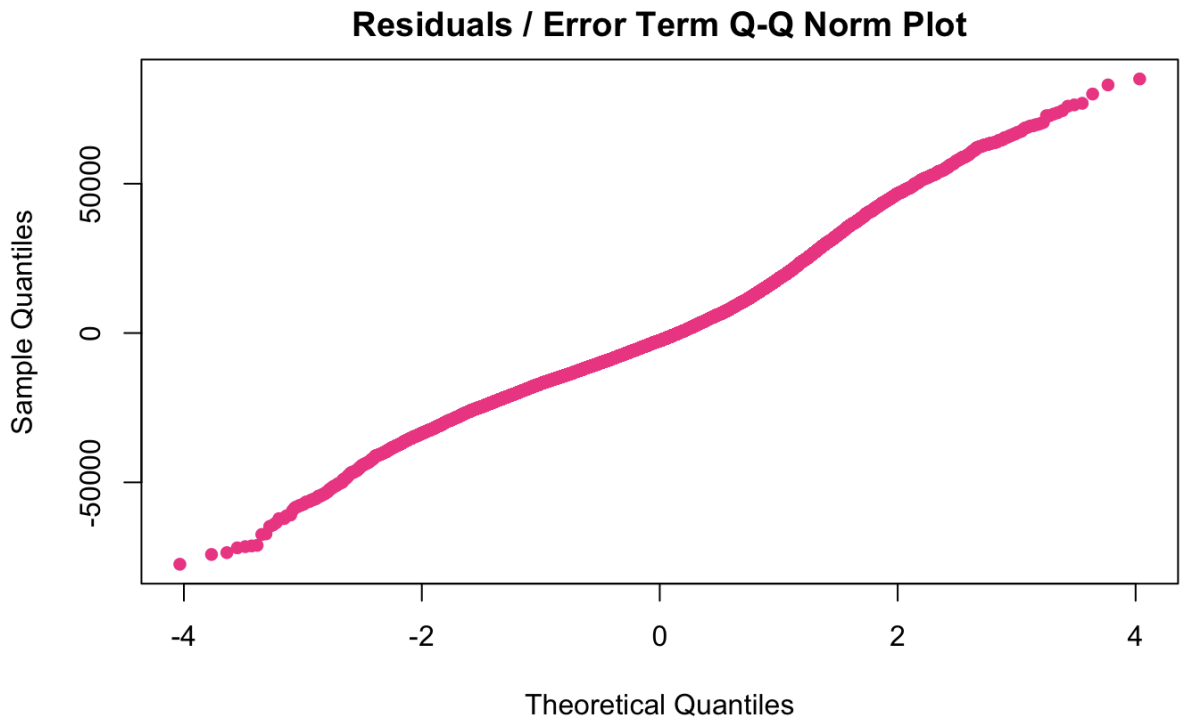

Residual Plot for Age
IndianaModel1$residuals
50000
0
-50000
20
30
40
50
60
70
TrainIndiana$age


Residual Plot for Educ
IndianaModel1$residuals
50000
0
-50000
10
15
20
TrainIndiana$educ

The first two assumptions state that the mean of the errors is equal to zero and the variance of errors is constant. For assumption 1, we refer to the above plots of the residuals versus our predictor variables. Because there are no trends, dramatic changes in variance, and less than 5% of the errors are outside the "2s" is 0, we can say that assumption 1 is satisfied. The second assumption is that the variance of errors is constant. We refer to the above plots to check for variance. Because there is a constant band around the x-axis, the data is homoscedastic. Therefore, we can say that assumption 2 is satisfied.

Analysis of Residual Q-Q Norm Plot (Checking Assumption 3)

The next assumption, 3, will be verified through use of a qq norm plot of the residuals, which is below.



The third assumption is that the error term is normally distributed. The Q-Q Norm Plot is mostly linear. Therefore, this chart shows the error term being normally distributed.

Performance of D-W Test (Checking Assumption 4)

The fourth assumption is tested through use of the Durbin Watson Test, which is done using the lmtest package in R. The output of this statistical test is below.

data: IndianaModel1

DW = 2.0025, p-value = 0.8632

The fourth assumption is that observations of response variable are independent from each other. We refer to the DW value from the D-W test. Because our DW value is very close to 2, we can say that assumption 4 is satisfied.

Checking for Multicollinearity (MLRM Assumption 5)

In order to adequately check for the presence of multicollinearity, there are a few steps we will take. First, we will check to see if any of the predictor variables have a strong linear correlation coefficient. The complete correlation matrix of all variables is below

	AGE	EDUC	HRSWRK	SEX	MARST
AGE	1	-0.0464	0.0366	0.0268	0.3306
EDUC	-0.0464	1	0.0534	0.1283	0.0621
HRSWRK	0.0366	0.0534	1	-0.2177	0.0828
SEX	0.0268	0.1283	-0.2177	1	-0.0200
MARST	0.3306	0.0621	0.0828	-0.0200	1

To check for multicollinearity, we check if there is a strong linear relationship between predictor variables. There are no values greater than 0.5 or less than -0.5, indicating weak, if any, linear relationship. Because none of the pairs of predictor variables have a strong linear relationships, we can say there is no multicollinearity so far in the model.

The next step in checking for multicollinearity is checking whether the beta coefficients of the model have opposite signs than expected. The second way to check for multicollinearity is if the beta coefficients have opposite signs than expected. In the above correlation matrix, all beta coefficients have signs that we would expect. Therefore, we can say there is no multicollinearity so far in the model.

The third step in checking for the presence of multicollinearity is seeing whether any of the predictor variables have a variance inflation factor (VIF) greater than 10. The VIF of each X variable is below.

TrainIndiana$educ	TrainIndiana$age	TrainIndiana$hrswork	TrainIndiana$sex	TrainIndiana$marst
1.033860	1.131406	1.063947	1.074664	1.136849

The third way to check for multicollinearity is to check if the variance inflation factor (VIF) values are greater than 10. All the above VIF values are less than 10. Therefore, we can say there is no multicollinearity so far in the model.

The final step in checking for multicollinearity is to look at the individual T-Tests of each predictor variable in our model, and ensure these results are consistent with the global F-Test. In the output for the regression model on page 10, we saw that all of the predictor variable T-Tests returned p values less than .05, meaning they are statistically significant in our model. The F-Test also returned a p value less than .05, meaning AT LEAST one of the predictor variables is statistically significant. Therefore, the global F-Test is consistent with the individual T-Tests, suggesting multicollinearity is NOT present.

The model has passed all four ways of checking for multicollinearity in the model. In the above pages, we have determined that the five assumptions have been satisfied. Therefore, we can conclude that all five assumptions of the Multiple Linear Regression Model are satisfied.

Checking Utility of Model: The Analysis of Variance F-Test

Ho: Beta1 = Beta2 = Beta3 = Beta4 = Beta5 = 0; None of the predictor variables are useful in the model

Ha: Beta(i) does not = 0; At least one of the predictor variables are useful in the model.

P-value = 2.2e-16 < 0.05 ; At a 5% significance level data supports the alternative hypothesis that at least one of the predictor variables is useful in the model.

Removal of Non-Useful Predictor Variables

Ho: Beta1 = 0

Ha: Beta 1 does not equal 0.

For a predictor variable to be useful in our model, we check for a p-value < 0.05. In the output on page 8, all predictor variables have a p-value of 2e-16, which is less than our alpha value. Therefore, the data provides evidence to support the alternative hypothesis that the predictor variable is useful in the model.

Prediction & Confidence Intervals

We will now create 95% prediction and confidence intervals for the dependent variable, inctotal based on our predictor variables used in the multivariate regression model. The intervals with the first 5 x values are provided below. We can see the confidence intervals are more precise / narrower than the prediction intervals, which is as we would expect.

Prediction Intervals:

	fit	lwr	upr
1	43624.87126	5947.341146	81302.40
2	45030.95086	7351.933291	82709.97
3	29179.33465	-8497.698167	66856.37
4	29724.31950	-7952.781040	67401.42
5	65580.62474	27893.139116	103268.11

Confidence Intervals:

	fit	lwr	upr
1	43624.87126	42988.89117	44260.851354
2	45030.95086	44312.22964	45749.672091
3	29179.33465	28573.53187	29785.137433
4	29724.31950	29114.31916	30334.319852
5	65580.62474	64506.02364	66655.225839

Prediction Accuracy

One way of checking the accuracy of our model is using mean absolute percentage error (MAPE). The formula for MAPE is the sum of differences between actual and predicted values, on a percent basis, divided by the number of observations.

MAPE: 0.7099669 (about 71%)