House Analysis

Surbhi Bector

house <- read.csv("C:/Users/Surbhi/Desktop/Stat 525 R/house.csv")  
regress = lm(CRIME~POP+POPCH+CHILD+LUNCH+INCOME+CRIMECH, data = house)  
sum <- summary(regress)  
sum

##   
## Call:  
## lm(formula = CRIME ~ POP + POPCH + CHILD + LUNCH + INCOME + CRIMECH,   
## data = house)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -117.010 -37.285 -2.292 28.046 241.633   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 67.5618 58.8035 1.149 0.2580   
## POP -5.2679 3.1061 -1.696 0.0983 .   
## POPCH 1.2279 0.9558 1.285 0.2069   
## CHILD -7.0990 1.6171 -4.390 9.11e-05 \*\*\*  
## LUNCH 2.8437 0.5304 5.361 4.60e-06 \*\*\*  
## INCOME 4.5877 1.6586 2.766 0.0088 \*\*   
## CRIMECH 0.5862 0.7550 0.776 0.4424   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 66.18 on 37 degrees of freedom  
## Multiple R-squared: 0.658, Adjusted R-squared: 0.6025   
## F-statistic: 11.86 on 6 and 37 DF, p-value: 2.187e-07

anova(regress)

## Analysis of Variance Table  
##   
## Response: CRIME  
## Df Sum Sq Mean Sq F value Pr(>F)   
## POP 1 61486 61486 14.0394 0.0006095 \*\*\*  
## POPCH 1 19407 19407 4.4313 0.0421361 \*   
## CHILD 1 30561 30561 6.9780 0.0120193 \*   
## LUNCH 1 165490 165490 37.7870 3.97e-07 \*\*\*  
## INCOME 1 32178 32178 7.3474 0.0101210 \*   
## CRIMECH 1 2641 2641 0.6029 0.4423983   
## Residuals 37 162043 4380   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The degrees of freedom in all of the variables is 1. For total population (POP), the sum of squares and MSE is 61486, meaning the data has a variablity of 61486. The F value of total population is 14.0394, also indicating a high variance in the data. The p value is .0006095 and hence small, we reject the null hypothesis that .

For percentage change in population over past several years (POPCH), the sum of squares and MSE is 19407, meaning the data has a variablity of 19407. The F value of total population is 4.4313, which is smaller but still indicates a high variance in the data. The p value is .0421361 and hence smaller than .05, we reject the null hypothesis that .

For percentage of children (under 18) in population(CHILD), the sum of squares and MSE is 30561, meaning the data has a variablity of 30561. The F value of total population is 6.9780, also indicating a high variance in the data. The p value is 0.0120193 and hence small, we reject the null hypothesis that .

For percentage free school lunch participation (LUNCH), the sum of squares and MSE is 165490, meaning the data has a high variablity of 165490. The F value of total population is 37.7870, also indicating a very high variance in the data. The p value is 3.97e-07 and hence very small, we reject the null hypothesis that .

For percentage change in household income over past several years(INCOME), the sum of squares and MSE is 32178, meaning the data has a variablity of 32178. The F value of total population is 7.3474, also indicating a high variance in the data. The p value is 0.0101210 and hence small, we reject the null hypothesis that .

For percentage change in crime rate over past several years.(CRIMECH), the sum of squares and MSE is 2641, meaning the data has a variablity of 2641. The F value of total population is 0.6029, also indicating a high variance in the data. The p value is 0.4423983 and hence still smaller than .05 , we reject the null hypothesis that .

The residual standard error is 162043, indicating that the average of the squares of errors is high, and the sum squared error is 4380 gives us a high deviation between the observations. The residual degrees of freedom is n-k-1 = 37.

3.1. : all = 0 and : all != 0

3.2. The value of the test statistic is 11.86.

3.3. The degrees of freedom of this test is n-k-1 = 37.

3.4. The p-value is 2.187e-07.

3.5. Because the p-value is less than alpha, .05, we should reject our null hypothesis.

3.6. Because we rejected our null hypothesis, the data shows that there is a regression relation between CRIME Y and the set of variables in the model. With the low p-value, our model can be assumed to be a “good” study of this relationship.

house <- read.csv("C:/Users/Surbhi/Desktop/Stat 525 R/house.csv")  
regress2 = lm(CRIME~POP+CHILD+LUNCH+INCOME, data = house)  
anova(regress2)

## Analysis of Variance Table  
##   
## Response: CRIME  
## Df Sum Sq Mean Sq F value Pr(>F)   
## POP 1 61486 61486 13.7787 0.0006411 \*\*\*  
## CHILD 1 24128 24128 5.4068 0.0253528 \*   
## LUNCH 1 154885 154885 34.7086 7.33e-07 \*\*\*  
## INCOME 1 59272 59272 13.2826 0.0007797 \*\*\*  
## Residuals 39 174035 4462   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(regress2,regress)

## Analysis of Variance Table  
##   
## Model 1: CRIME ~ POP + CHILD + LUNCH + INCOME  
## Model 2: CRIME ~ POP + POPCH + CHILD + LUNCH + INCOME + CRIMECH  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 39 174035   
## 2 37 162043 2 11992 1.3691 0.2669

4.1. : some = 0 and : some !=0

4.2. CRIME = +POP+POPCH+CHILD+LUNCH+INCOME+CRIMECH+

4.3. CRIME = +POP+CHILD+LUNCH+INCOME+

4.4. The value of the test statistic is 1.3691.

4.5. The degrees of freedom of this test is n-k-1 = 2.

4.6. The p-value is 0.2669.

4.7. Because the p-value is very large and less than alpha, .05, we should reject our null hypothesis.

4.8. Because we failed to reject our null hypothesis, the data shows that there is not enough evidence to say that there is a regression relation between CRIME Y and the set of updated variables in the model. With the two variables removed, the p-value is larger and therefore, there is less evidence in favor of the alternative hypothesis. This means that the new model does not conduct a better study of the relationship.

1. Looking at (POP):

t-test: t= = -5.2679/3.1061 = -1.696 F-test for = 14.0394

1. coefficients of partial determination:

SSEreduced = 4462\*39 = 174018

SSEfull = 4380\*37 = 162060

= (SSEreduced - SSEfull) / SSEreduced = .0687

library(usdm)

## Loading required package: sp

## Loading required package: raster

df = house  
vif(df)

## Variables VIF  
## 1 POP 1.289981  
## 2 POPCH 1.395811  
## 3 CHILD 3.065148  
## 4 LUNCH 3.288588  
## 5 INCOME 1.495688  
## 6 CRIME 2.923949  
## 7 CRIMECH 1.839426

Therefore, we see CHILD and LUNCH are heavily correlated. Our new model excludes percentage of children (under 18) in population:

regress3 = lm(CRIME~POP+POPCH+LUNCH+INCOME+CRIMECH, data = house)  
sum3 <- summary(regress)  
sum3

##   
## Call:  
## lm(formula = CRIME ~ POP + POPCH + CHILD + LUNCH + INCOME + CRIMECH,   
## data = house)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -117.010 -37.285 -2.292 28.046 241.633   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 67.5618 58.8035 1.149 0.2580   
## POP -5.2679 3.1061 -1.696 0.0983 .   
## POPCH 1.2279 0.9558 1.285 0.2069   
## CHILD -7.0990 1.6171 -4.390 9.11e-05 \*\*\*  
## LUNCH 2.8437 0.5304 5.361 4.60e-06 \*\*\*  
## INCOME 4.5877 1.6586 2.766 0.0088 \*\*   
## CRIMECH 0.5862 0.7550 0.776 0.4424   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 66.18 on 37 degrees of freedom  
## Multiple R-squared: 0.658, Adjusted R-squared: 0.6025   
## F-statistic: 11.86 on 6 and 37 DF, p-value: 2.187e-07