Final!!

Monique Sierra

12/8/2020

library (readr)  
Chicago.Final.Num <- read.csv("~/Downloads/Chicago.Final.Num.csv")  
  
chicago <- Chicago.Final.Num

## New Names

library(ggformula)

## Loading required package: ggplot2

## Loading required package: ggstance

##   
## Attaching package: 'ggstance'

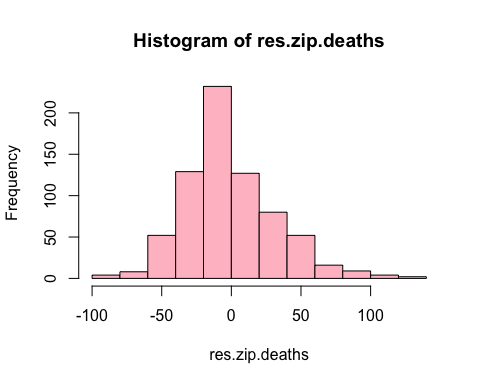
## The following objects are masked from 'package:ggplot2':  
##   
## geom\_errorbarh, GeomErrorbarh

##   
## New to ggformula? Try the tutorials:   
## learnr::run\_tutorial("introduction", package = "ggformula")  
## learnr::run\_tutorial("refining", package = "ggformula")

c.weekly <-chicago$Cases...Weekly  
week.num <-(chicago$Week.Number)  
p.d.fact <- chicago$pop.density.factor  
income <-chicago$income.factor  
t.r.weekly <- chicago$Test.Rate...Weekly  
d.r.cuml <-chicago$Death.Rate...Cumulative  
d.r.weekly <-chicago$Death.Rate...Weekly  
hh.avg.in <-chicago$HH.Avg.Income  
c.r.weekly <-chicago$Case.Rate...Weekly  
d.weekly <-chicago$Deaths...Weekly

### 1. In Chicago, were all neighborhoods equally affected by death from covid? In this context by ‘neighborhood’ I mean ZIP code. Present evidence for your conclusion.

zip.c.d <- lm(d.r.cuml ~ as.factor(chicago$ZIP), data = chicago)  
  
res.zip.deaths <-resid(zip.c.d)  
hist(res.zip.deaths, col = "pink")

 ##### First we plotted a line of the cumulative deaths across all zip codes. Then we took the residuals of our regression to see if our model was good at predicting any type of trend and took the residuals. Although many of our residuals fell around zero, a substantial amount were scattered outside the line. If the ZIP codes were similar and accurately predicted by our model, we would see a lower level of residuals present. My conclusion is that there is variation among the number of deaths across all zip codes.

### 2. Is there a statistically significant difference between the weekly deaths in zip=60618 and zip 60624? Show evidence for your conclusion.

PLEASE <- data.frame(chicago[chicago$ZIP== "60618" | chicago$ZIP== "60624", c("Death.Rate...Weekly", "Deaths...Weekly","Week.Number", "ZIP")])  
PLEASE$ZIP <- as.factor(PLEASE$ZIP)  
  
something.whack <- aov(Deaths...Weekly ~ ZIP, data = PLEASE)  
  
summary(something.whack)

## Df Sum Sq Mean Sq F value Pr(>F)  
## ZIP 1 6.5 6.500 1.027 0.321  
## Residuals 24 151.8 6.327

##### The P-value is high at .321, showing there is no evidence to support a statistically significant difference between the number of weekly deaths between the two zip codes.

### 3. Does it appear that the population density has an effect on Case.Rate…Weekly? Back up your arguments with evidence.

p.d.fact <- cut(chicago$people.sqmi, right = F,  
 breaks = c(0,14,27,40,60),  
 labels = c("Rural", "ExUrban", "Suburban", "Urban"))  
summary(p.d.fact)

## Rural ExUrban Suburban Urban NA's   
## 169 169 169 195 13

pd.crw <- aov(c.r.weekly ~ p.d.fact, data = chicago)  
summary(pd.crw)

## Df Sum Sq Mean Sq F value Pr(>F)   
## p.d.fact 3 103095 34365 2.636 0.049 \*  
## Residuals 570 7430609 13036   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 141 observations deleted due to missingness

##### P-value is 0.049 which is JUST under the significance level of .05, making the differences in case rate weekly according to location statistically significant.

### 4. This time create Chicago.Final$income.factor with levels Poor [0, 30000), M\_Class [30000, 51000), and Wealthy [51000, 105000) (again, closed on left, open on right.)

Does it appear that wealth has an effect on Case.Rate…Weekly? Back up your arguments with evidence.

income <- cut(hh.avg.in, right = F,  
 c(0,30000,51000,105000),  
 labels = c("Poor","M\_Class","Wealthy")  
 )  
summary(income)

## Poor M\_Class Wealthy   
## 169 364 182

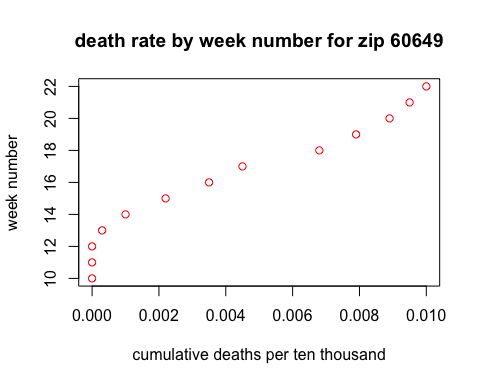
money.cases <-aov(c.r.weekly ~ income,data = chicago)  
summary(money.cases)

## Df Sum Sq Mean Sq F value Pr(>F)   
## income 2 769083 384541 32.56 3.95e-14 \*\*\*  
## Residuals 582 6873989 11811   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 130 observations deleted due to missingness

##### Yes, it does appear that wealth has an effect on case rate weekly. When we run ANOVA, we see that the number of cases between the three levels of income produced a P-value that was tiny, (3.95e-14), showing they are significant.

### 5. For zip code 60649 graph cumulative deaths per ten thousand vs week number. Comment on the results.

chicago$Deaths...Cumulative <- chicago$Deaths...Cumulative/10000  
  
newdata <- subset(chicago, ZIP==(60649), select= c(Deaths...Cumulative, Week.Number))  
  
plot(newdata, xlab = "cumulative deaths per ten thousand", ylab = "week number", main = "death rate by week number for zip 60649", col = "red")

 ##### After changing cumulative deaths to deaths per ten thousand, we see the number of deaths remain near constant for the first 12 weeks, but then around week 14 it steadily increased. Most recently it has shown signs of slowing, but the number of deaths are still ultimately increasing.

### 6. Use multiple linear regression to predict Cases…Weekly using Week.Number, income.factor, pop.density.factor, and Test.Rate…Weekly. Tell me about significance of coefficients and R squared adjusted.

c.weekly <-chicago$Cases...Weekly  
week.num <-chicago$Week.Number  
p.d.fact <- chicago$pop.density.factor  
income <-chicago$income.factor  
t.r.weekly <- chicago$Test.Rate...Weekly  
  
den <- lm(c.weekly ~ t.r.weekly + week.num + p.d.fact +income, data = chicago)  
summary(den)

##   
## Call:  
## lm(formula = c.weekly ~ t.r.weekly + week.num + p.d.fact + income,   
## data = chicago)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -374.64 -41.23 -11.52 27.44 414.27   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 104.7256 21.8099 4.802 2.03e-06 \*\*\*  
## t.r.weekly 0.1335 0.0119 11.217 < 2e-16 \*\*\*  
## week.num -5.2266 1.5061 -3.470 0.00056 \*\*\*  
## p.d.factRural -56.1921 10.1701 -5.525 5.07e-08 \*\*\*  
## p.d.factSuburban 1.7273 9.2584 0.187 0.85207   
## p.d.factUrban 21.0435 9.4658 2.223 0.02661 \*   
## incomePoor -23.4168 8.9696 -2.611 0.00928 \*\*   
## incomeWealthy -53.1700 8.0793 -6.581 1.08e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 77.48 on 555 degrees of freedom  
## (152 observations deleted due to missingness)  
## Multiple R-squared: 0.3431, Adjusted R-squared: 0.3348   
## F-statistic: 41.41 on 7 and 555 DF, p-value: < 2.2e-16

##### The P-value of week number is very low, making it a good indicator of the number of cases weekly. Test rate weekly is has an even lower p-value, which makes sense because more testing will clearly be linked to more reported cases. Between the three population density factors, the rural level showed to have the most significance. Among the two income factors, “wealthy” had the lower p-value, indicating their values are more significant than “poor.” The adjusted R-squared value of .3348 is not terrific, as the higher it is the better it follows the model.

### 7. Use the regression equation in #6 above and the “predict” function to predict “Cases…Weekly” for predictors with values Week.Number=20, income.factor=‘Poor’, pop.density.factor=‘Urban’, Test.Rate…Weekly=50. What is the prediction? Hint: this might help:

<http://www.sthda.com/english/articles/40-regression-analysis/166-predict-in-r-model-predictions-and-confidence-intervals/>

den <- lm(c.weekly ~ t.r.weekly + week.num + p.d.fact + income, data = chicago)  
  
  
hola<- data.frame(  
   
t.r.weekly = c(50),  
p.d.fact = c('Urban'),  
income = c('Poor'),  
week.num = c(20)  
)  
  
predict(den, newdata = hola)

## 1   
## 4.492695

##### The predicted number of cases with our variable values specifically set gives us 8.59, or around 9 cases. This means that given a test rate of 50, an urban population density factor, and an income categorized as “poor” during week 20, we would supposedly get a value close to our prediction of nine cases.

### 8. Run gf\_qq and gf\_qqline on the residuals from the above regression. What do you conclude?

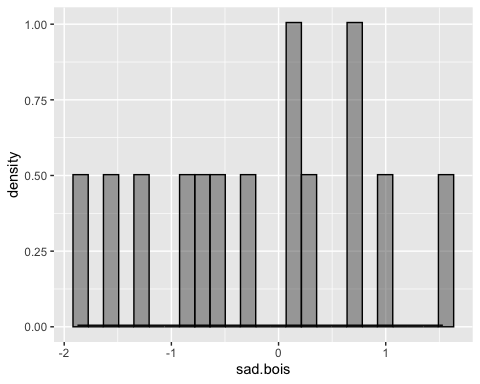
pupper <-resid(den)  
res.mean <- mean(pupper)  
sd(pupper)

## [1] 76.99165

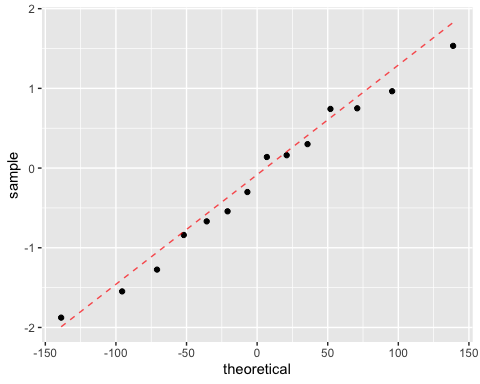
res.mean

## [1] 3.219268e-16

sad.bois <-rnorm(den)  
  
gf\_dhistogram(~sad.bois, color = "black") %>%   
 gf\_function(dnorm, args=c(3.219268e-16,76.99165),  
 xlim = 100:100)

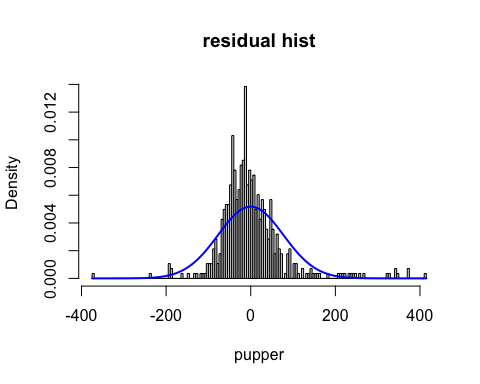


gf\_qq(~sad.bois, distribution = qnorm,  
 dparams = c(3.219268e-16,76.99165)) %>%  
 gf\_qqline(~sad.bois, distribution=qnorm,  
 dparams = c(3.219268e-16,76.99165),  
 color = "red")

 ##### We can see it’s pretty close to normal, but there appears to be a few outlines scattered here and there. Variation is shown but that is to be expected.

### 9. Do a histogram of residuals with the appropriate normal distribution layered over the top. Comment on the fit you see.

den <- lm(c.weekly ~ t.r.weekly + week.num + p.d.fact + income, data = chicago)  
  
pupper <-residuals(den)  
hist(pupper, breaks = 125, probability = T,  
 main = "residual hist")  
  
res.mean <- mean(pupper)  
sd.rm <-sqrt(var(pupper))  
   
  
curve(dnorm(x,mean = res.mean, sd=sd.rm),  
 col= "blue", lwd = 2, add = TRUE)

 ##### Our histogram of residuals from the linear model appears to have a normal distribution. Although it is not perfect, it fits the model well. The fact that the majority of the residuals fall around 0 shows our line is a relatively good predictor.

### 10. Using R code, find the zip code with the largest Death.Rate…Cumulative and look at the income.factor, pop.density.factor, Test.Rate…Weekly. Think about how these or other factors could influence the Death.Rate…Cumulative and discuss. What else might contribute to a large death rate?

highest.death <- sort(d.r.cuml, decreasing = TRUE, na.last = NA)  
head(highest.death)

## [1] 217.3 206.4 193.4 187.0 174.9 171.6

bofa <- subset(chicago,   
 d.r.cuml==217.3,   
 select= c(income.factor, Week.Number, Test.Rate...Weekly, pop.density.factor)  
 )  
print(bofa)

## income.factor Week.Number Test.Rate...Weekly pop.density.factor  
## 579 Poor 22 874 Suburban

##### First we found the highest cumulative death rate, then we created a subset where we set deaths equal to that value and found the correlating variables. We see that the highest death value had an income factor of “poor,” and is from a suburban setting where the test rate weekly was 874. The higher rate may be due to the fact that “poorer” communities are faced with jobs that can’t be fulfilled remotely, or it could also be the fact that they live in the suburbs and they’re closer to each other. Another factor may be their lower wealth grants them weaker healthcare, contributing to their higher death rates.