Stat 408

Stone Cai

2024-01-21

Inputting in data

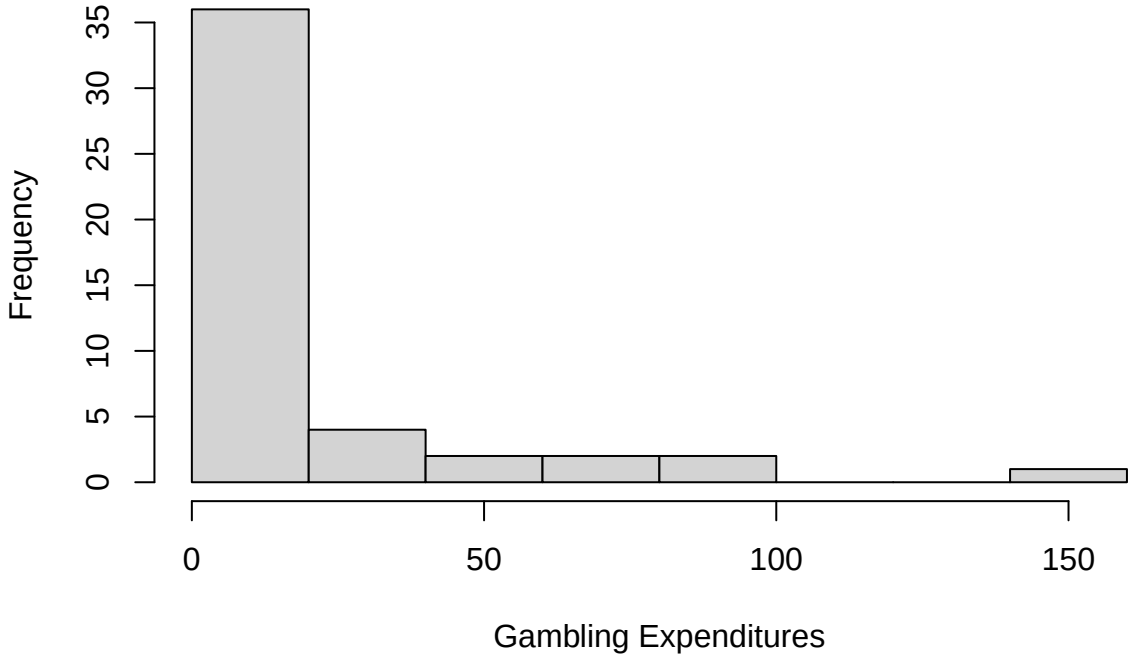
Problem 1a.

```
teengamb<-read.csv("C:/Users/stone/Documents/Stat 408/Homework Data Files/teengamb.csv")
teengamb$sex <- factor(teengamb$sex)
levels(teengamb$sex) <- c("male", "female")

summary(teengamb)
```

```
##      sex         status          income          verbal          gamble
##  male  :28   Min.   :18.00   Min.   : 0.600   Min.   : 1.00   Min.   :  0.0
##  female:19   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00   1st Qu.:  1.1
##              Median :43.00   Median : 3.250   Median : 7.00   Median :  6.0
##              Mean   :45.23   Mean   : 4.642   Mean   : 6.66   Mean   : 19.3
##              3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00   3rd Qu.: 19.4
##              Max.   :75.00   Max.   :15.000   Max.   :10.00   Max.   :156.0
```

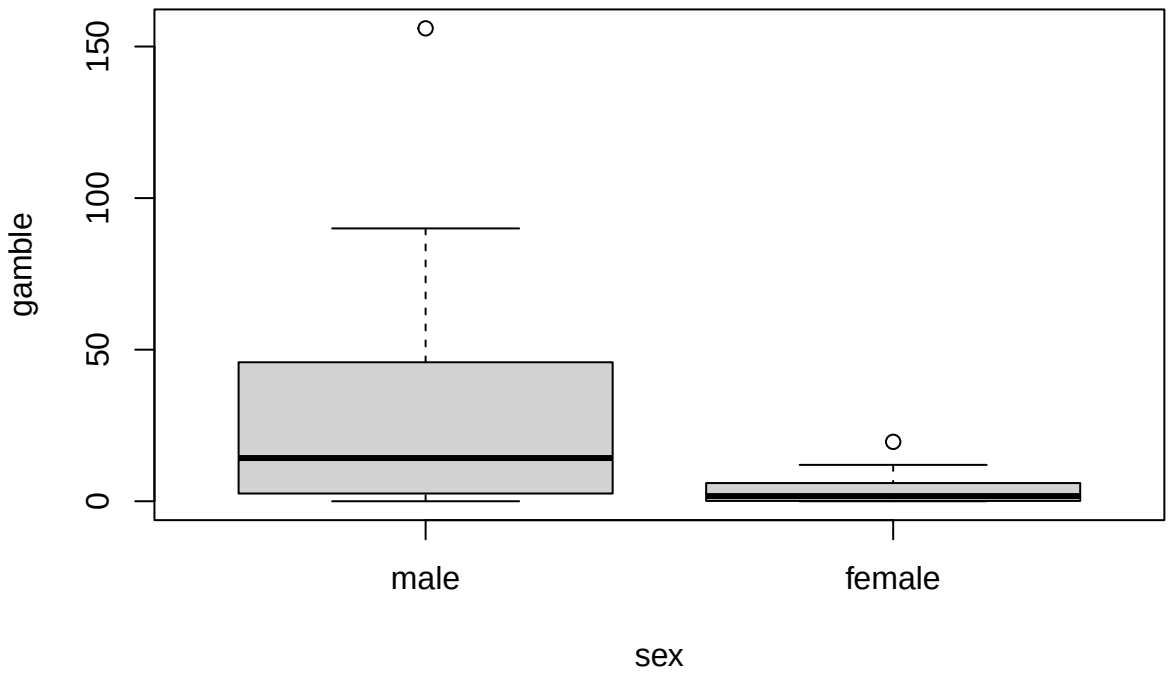
For contextual purposes it doesn't make a lot of sense to quantitatively summarize sex. Rather I am choosing to reclassify qualitatively.After doing, so we can see that more males than females were sampled for this study. Additionally, when looking at gambling expenditures, it seems that median gambling amounts is relatively low, but I notice that the mean is much higher than the median which may suggest a few large observations skewing the distribution. I'll plot a histogram of gambling expenditures to confirm. Nothing else quantitatively jumps out at first glance. It is important to note that both verbal and status variables are measured in discreet numbers while income and gamble are not.

```
hist(teengamb$gamble, xlab = "Gambling Expenditures", main = "")
```



The histogram that there is a heavy right skew in the distribution. A majority of teens seem to gamble very little or not at all. However, there seem to be a few large observations with a potential outlier notably maximum observation of 156 pounds.

```
plot(gamble~sex, data=teengamb)
```



At first glance, it is easy to question whether there is a difference in gambling expenditure based on sex. Taking a look at boxplot, we see that the median and 3rd quantile gambling expenditure observations for males are significantly larger than for the same female group. However, we also notice that both groups have 1 likely outlier observation.

Problem 2.

```
lmsex<-lm(gamble~sex,data=teengamb)
summary (lmsex)
```

```
##
## Call:
## lm(formula = gamble ~ sex, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.775 -18.325  -3.766   6.334 126.225
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.775      5.498   5.415 2.28e-06 ***
## sexfemale    -25.909      8.648  -2.996  0.00444 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.09 on 45 degrees of freedom
```

```
## Multiple R-squared:  0.1663, Adjusted R-squared:  0.1478
## F-statistic: 8.977 on 1 and 45 DF,  p-value: 0.004437
```

```
lmstatus<-lm(gamble~status,data=teengamb)
summary (lmstatus)
```

```
##
## Call:
## lm(formula = gamble ~ status, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.708 -17.903 -13.929   2.195 135.020
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.46486   13.14189   1.786   0.0809 .
## status      -0.09205    0.27180  -0.339   0.7364
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.82 on 45 degrees of freedom
## Multiple R-squared:  0.002542,   Adjusted R-squared:  -0.01962
## F-statistic: 0.1147 on 1 and 45 DF,  p-value: 0.7364
```

```
lmincome<-lm(gamble~income,data=teengamb)
summary (lmincome)
```

```
##
## Call:
## lm(formula = gamble ~ income, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.020 -11.874  -3.757  11.934 107.120
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.325      6.030  -1.049      0.3
## income         5.520      1.036   5.330 3.05e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.95 on 45 degrees of freedom
## Multiple R-squared:  0.387,  Adjusted R-squared:  0.3734
## F-statistic: 28.41 on 1 and 45 DF,  p-value: 3.045e-06
```

```
lmverbal<-lm(gamble~verbal,data=teengamb)
summary (lmverbal)
```

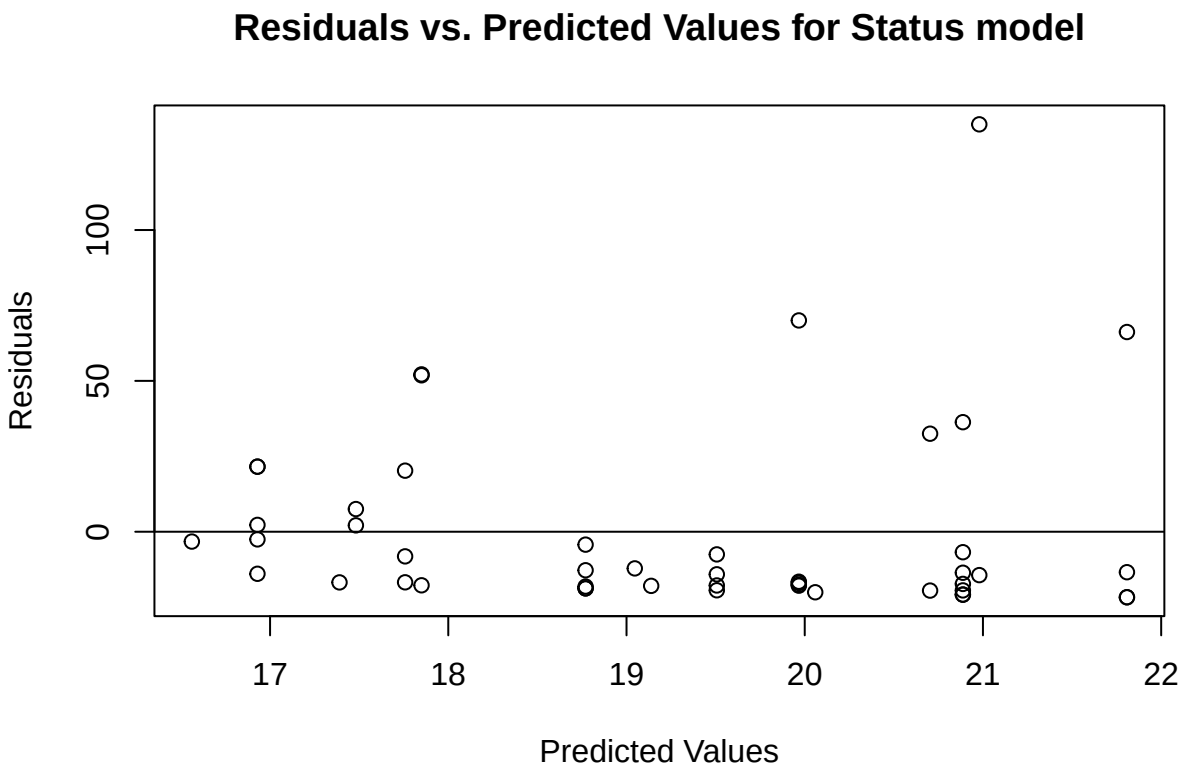
```
##
## Call:
```

```
## lm(formula = gamble ~ verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.036 -18.047 -13.294   4.271 126.764
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   44.178     17.053   2.591   0.0129 *
## verbal        -3.736      2.469  -1.513   0.1372
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.08 on 45 degrees of freedom
## Multiple R-squared:  0.04842,    Adjusted R-squared:  0.02728
## F-statistic:  2.29 on 1 and 45 DF,  p-value: 0.1372
```

The following outputs represent a simple linear regression model with the expenditure on gambling as the response and each of the following predictors individually: sex, status, income, and verbal.

Problem 3.

```
lmstatus<-lm(gamble~status,data=teengamb)
plot(x=predict(lmstatus), y= residuals(lmstatus),
     xlab='Predicted Values',
     ylab='Residuals',
     main='Residuals vs. Predicted Values for Status model')

abline(h=0)
```



Here is the plot of residuals vs predicted values. The plot suggests that most of the models residuals are negative, but observations with positive residuals have very high residuals.The range of the predicted values for status is much smaller than the range of status found in the original data.

Problem 4)

```
lmfull<-lm(gamble~sex+status+income+verbal,data=teengamb)
summary(lmfull)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sexfemale   -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

```
##Mean and residuals of full model
resids<- residuals(lmfull)
mean(resids)
```

```
## [1] -1.556914e-16
```

```
median(resids)
```

```
## [1] -1.451392
```

a) We look at R squared output. The 0.5267 value tells us that 52.67% of variation in the response is explained by the predictors.
b) Mean of residuals is nearly zero at -1.55 * 10^-16. Median is -1.451

```
cor( residuals(lmfull), fitted(lmfull) )
```

```
## [1] -6.215823e-17
```

```
cor(teengamb$income, residuals(lmfull))
```

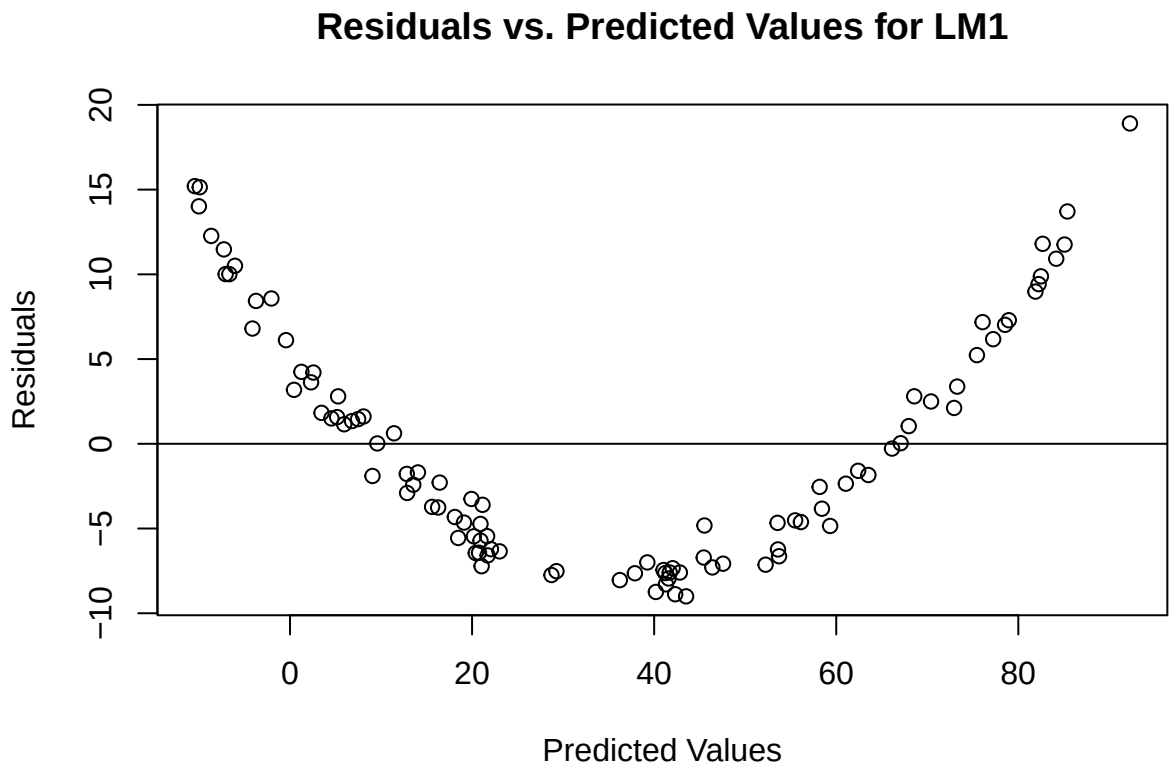
```
## [1] 3.247058e-17
```

c) correlation between residuals of the full model and fitted value is near zero at -6.21*10^-17
d) Correlation between Income and Residuals of full model is 3.24*10^-17
e) Since male are coded as zero, we consult the sex coefficient value. The coefficient of sex is -22.11833. Thus,when holding all other predictors constant, female teen gamblers spent 22.12 pounds less on gambling expenditures than male teen gamblers.

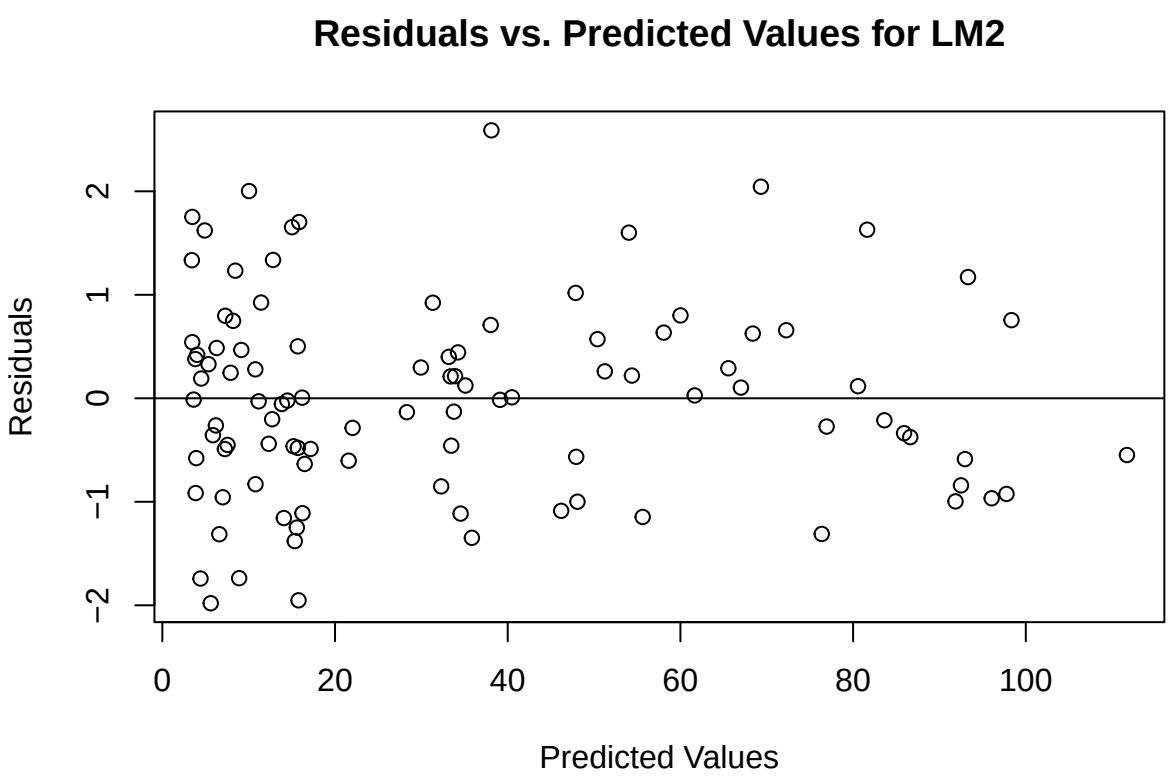
Problem 5

```
#Use the following code to simulate 100 observations
#from a linear model.
#setting a seed will give you the same random draw
#every time you run the code.
set.seed(1234)
#dont understand this? Try `help(runif)'
x <- runif(100,0,10)
#This is the true model. (In practice we don't know the truth)
y <- 3 + x + x^2 + rnorm(100,0,1)

lm1<-lm(y~x)
#Second Model (the true model)
lm2<-lm(y~x+I(x^2))

##summary(lm1)
```

```r
##summary(lm2)

plot(x=predict(lm1), y= residuals(lm1),
     xlab='Predicted Values',
     ylab='Residuals',
     main='Residuals vs. Predicted Values for LM1')
abline(h=0)
```




```r
plot(x=predict(lm2), y= residuals(lm2),
     xlab='Predicted Values',
     ylab='Residuals',
     main='Residuals vs. Predicted Values for LM2')
abline(h=0)
```



For the plot of LM1, I notice that the plot of residuals vs predicted values follows a parabolic arc. There seems to be an even distribution of positive and negative residuals. However, the shape of the graph suggests the model might not be linear as the residuals clearly are scattered in a trend. For plot 2, I notice the residuals are much more scattered. There seems to be a symmetric distribution of positive and negative residuals and they seem to be scattered horizontally with no specific pattern.There seems to be non-constant variance amongst the residuals.There seems to be no graphical indication from the residuals that the model is inherently wrong and needs to be improved upon.