Data605_Discussion605

Gabe Abreu

11/4/2020

Discussion 11

I'm going to use the built-in data set USArrests to demonstrate linear regression.

Is there a relationship between urban population and assaults?

```
data('USArrests')
summary(USArrests)
```

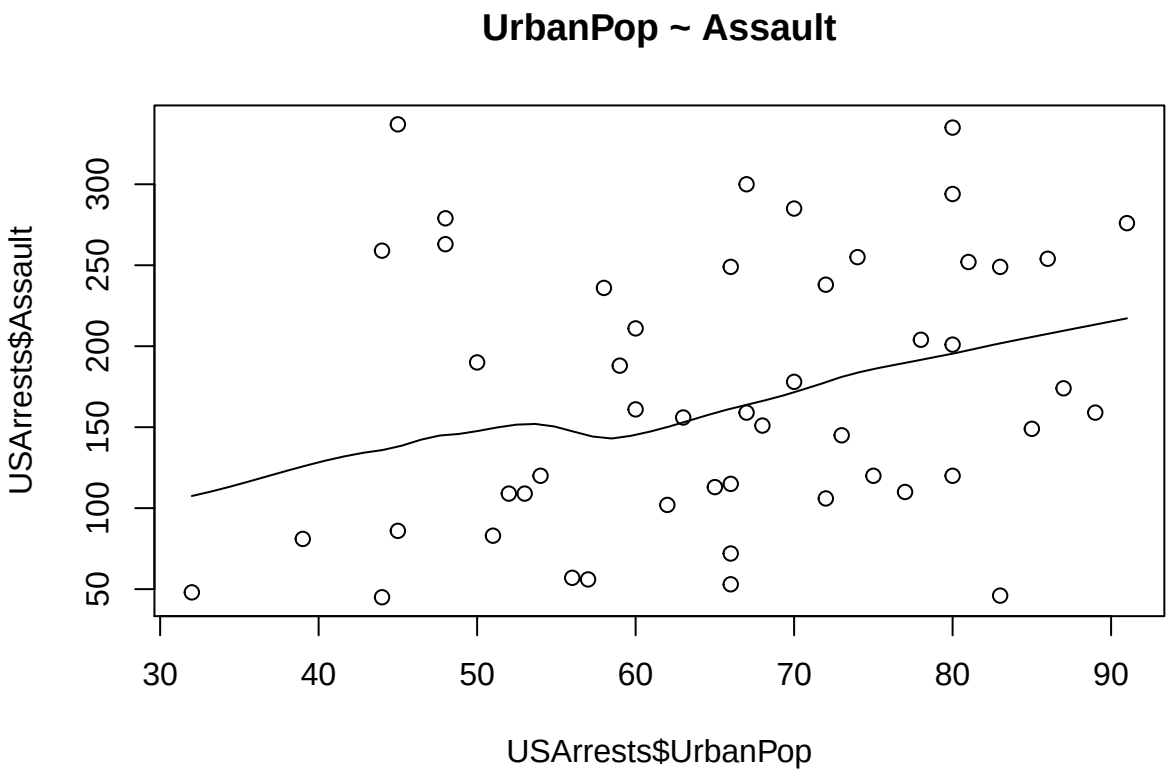
```
##      Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

The scatter plot can visualize linear relationships between the depdendent variable and indepdendent variable.

```
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```
scatter.smooth(x=USArrests$UrbanPop, y=USArrests$Assault, main="UrbanPop ~ Assault")
```



Let's build the linear regression and then print the summary.

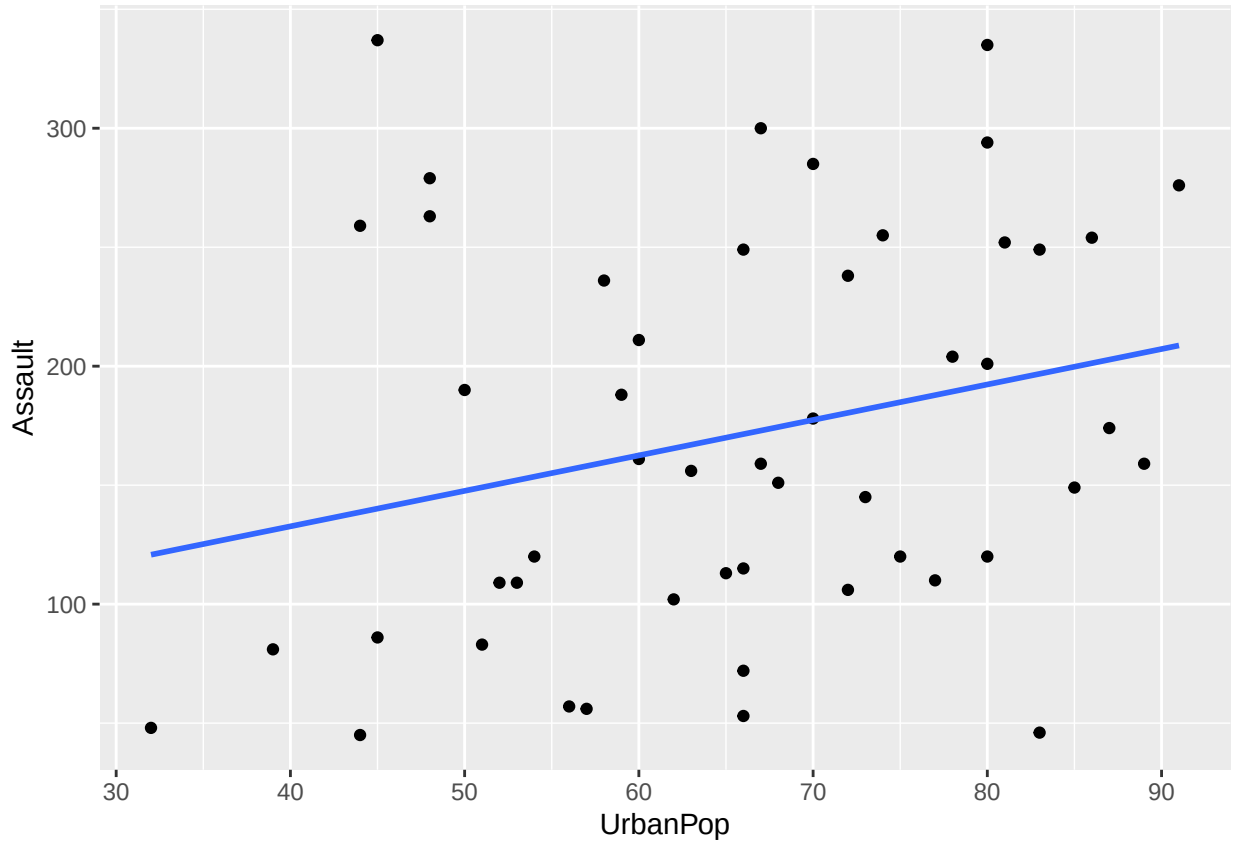
```
linear_reg <- lm(UrbanPop ~ Assault, data = USArrests)
summary(linear_reg)
```

```
##
## Call:
## lm(formula = UrbanPop ~ Assault, data = USArrests)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.020  -9.637   2.023  10.567  23.989
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 57.86213    4.59228  12.600   <2e-16 ***
## Assault      0.04496    0.02422   1.857   0.0695 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.13 on 48 degrees of freedom
## Multiple R-squared:  0.06701,    Adjusted R-squared:  0.04758
## F-statistic: 3.448 on 1 and 48 DF,  p-value: 0.06948
```

The p value is greater than 0.05, which means it's not statistically significant.

Plotting the linear regression with a scatter plot.

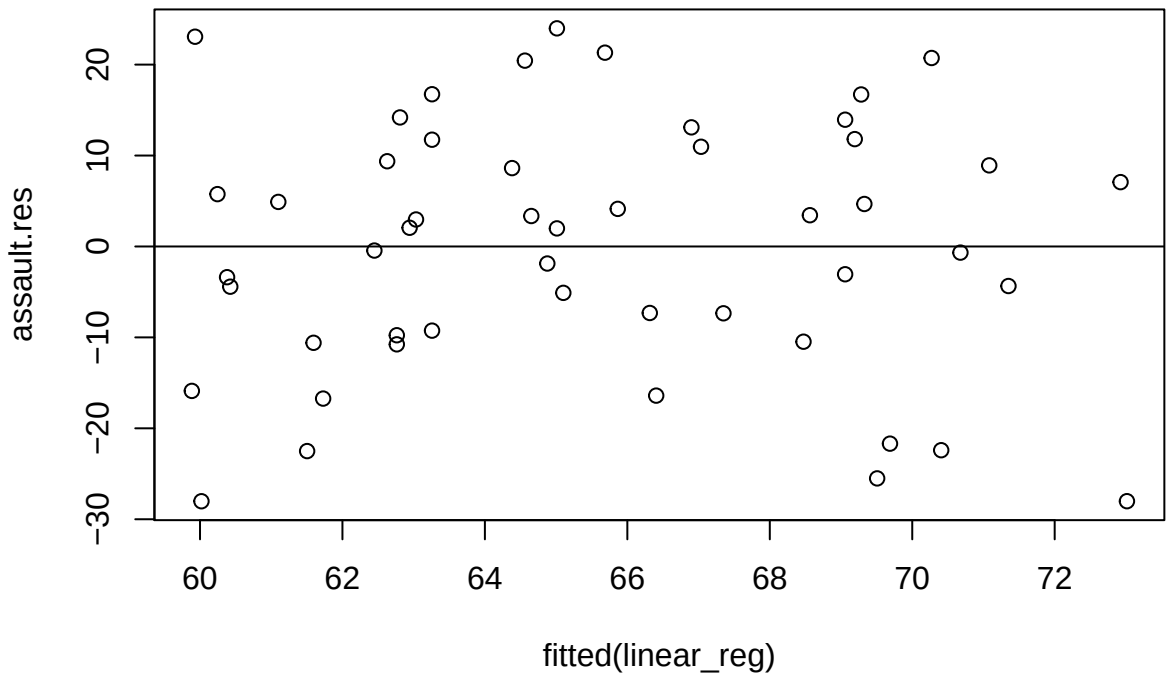
```
ggplot(data = USArrests, aes(x = UrbanPop, y = Assault)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE)
```



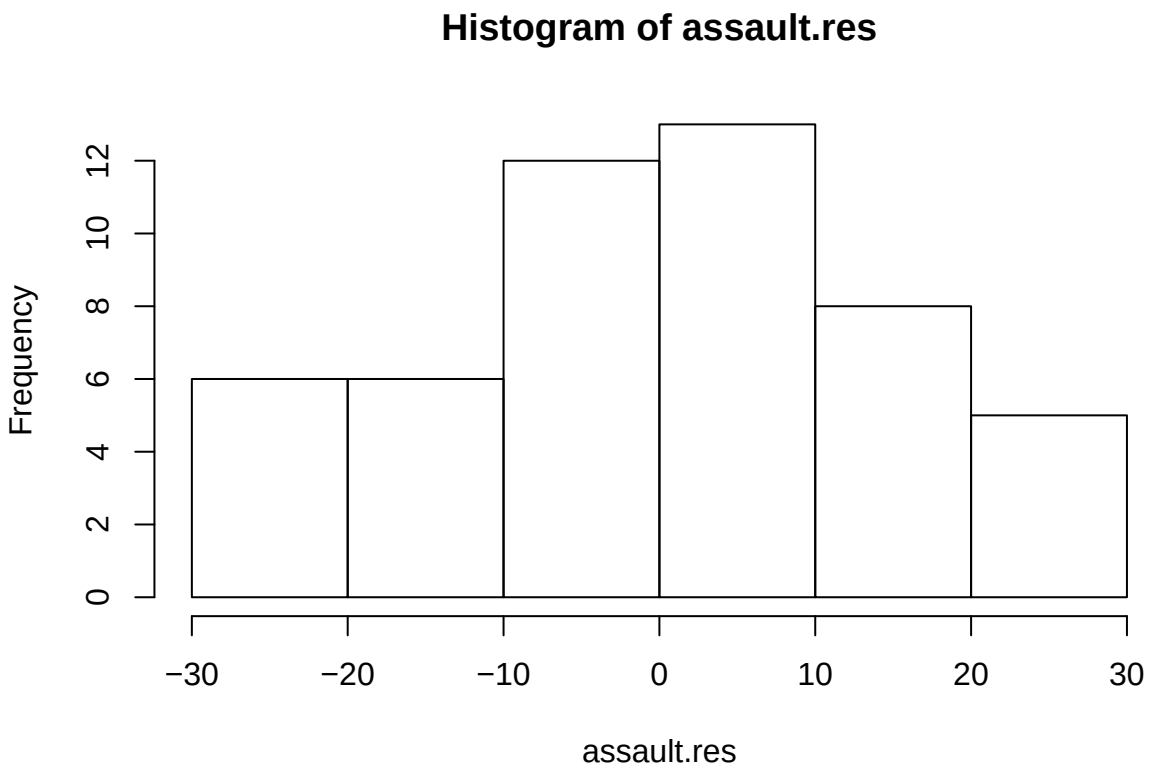
Residual Plot

```
assault.res = resid(linear_reg)

 plot(fitted(linear_reg), assault.res)
 abline(0,0)
```



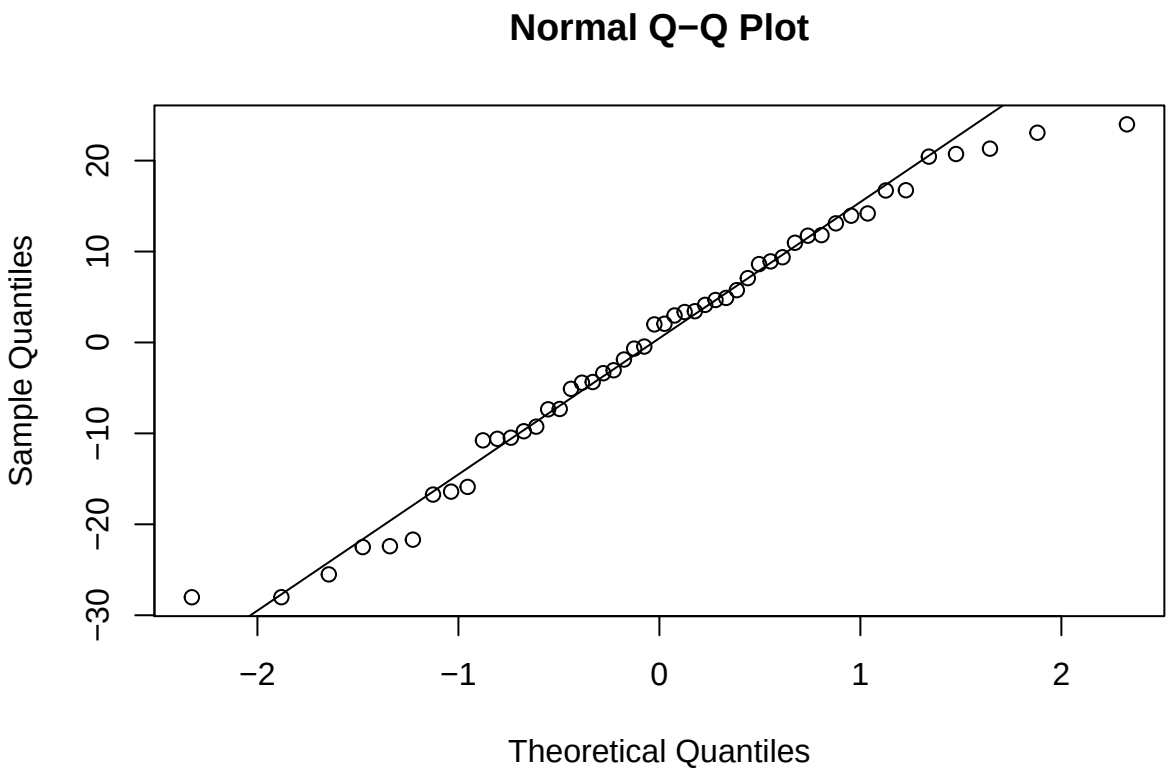

```
hist(assault.res)
```



Q-Q Plot

The plot below shows there are outliers at both ends of the dat, but more so towards upper quantiles.

```
qqnorm(resid(linear_reg))
qqline(resid(linear_reg))
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−30
−20
−10
0
10
20
−2
−1
0
1
2