Homework4(LinearReg)

N/A

2023-09-22

```
library(ggplot2)
library(broom)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.
0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ lubridate 1.9.2     ✔ tibble    3.2.1
## ✔ purrr     1.0.2     ✔ tidyr     1.3.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts
() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

QUESTION 8.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a linear regression model would be appropriate. List some (up to 5) predictors that you might use.

ANS: In the real estate market, accurately determining house prices is paramount for buyers, sellers, and real estate agents alike. To address this challenge, a linear regression model can be a valuable tool, estimating house prices based on a range of predictors, providing valuable insights to assist stakeholders in making informed decisions. One of the fundamental predictors for estimating house prices is the square footage of the property; larger houses tend to command higher prices. Additionally, the number of bedrooms is a key factor influencing price, as larger families or those seeking extra space typically require more bedrooms. The location of the house plays a pivotal role in its valuation; desirable neighborhoods or cities with attributes such as good schools, low crime rates, and proximity to amenities often have higher property values. The year the house was built is another predictor, as newer houses may be more appealing due to modern features and reduced maintenance requirements. Finally, the presence of recent renovations or updates can significantly impact a house's price; houses that have undergone improvements are often more attractive to buyers, as they may require less immediate investment. By analyzing historical data on houses and incorporating these predictors, a linear regression model can establish relationships between these factors and house prices, facilitating accurate price predictions for both new and existing properties.

QUESTION 8.2

Using crime data from http://www.statsci.org/data/general/uscrime.txt (file uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html), use regression (a useful R function is lm or glm) to predict the observed crime rate in a city with the following data:

M = 14.0 So = 0 Ed = 10.0 Po1 = 12.0 Po2 = 15.5 LF = 0.640 M.F = 94.0 Pop = 150 NW = 1.1 U1 = 0.120 U2 = 3.6 Wealth = 3200 Ineq = 20.1 Prob = 0.04 Time = 39.0

Show your model (factors used and their coefficients), the software output, and the quality of fit. Note that because there are only 47 data points and 15 predictors, you'll probably notice some overfitting. We'll see ways of dealing with this sort of problem later in the course.

Let's first read in the data:

```
crime_data = read.table("uscrime.csv", header = TRUE)
```

Here is a brief description of each variable, taken from the website :

```
Variable   Description
M       percentage of males aged 14–24 in total state population
So      indicator variable for a southern state
Ed      mean years of schooling of the population aged 25 years or over
Po1     per capita expenditure on police protection in 1960
Po2     per capita expenditure on police protection in 1959
LF      labour force participation rate of civilian urban males in the age-gro
up 14-24
M.F     number of males per 100 females
Pop     state population in 1960 in hundred thousands
NW      percentage of nonwhites in the population
U1      unemployment rate of urban males 14–24
U2      unemployment rate of urban males 35–39
Wealth  wealth: median value of transferable assets or family income
Ineq    income inequality: percentage of families earning below half the media
n income
Prob    probability of imprisonment: ratio of number of commitments to number
of offenses
Time    average time in months served by offenders in state prisons before the
ir first release
Crime       crime rate: number of offenses per 100,000 population in 1960
```

Based on the requirements in the problem, I created a dataframe with the predictors. Before we use these however, we will run some exploratory statistics and analysis on the dataset.

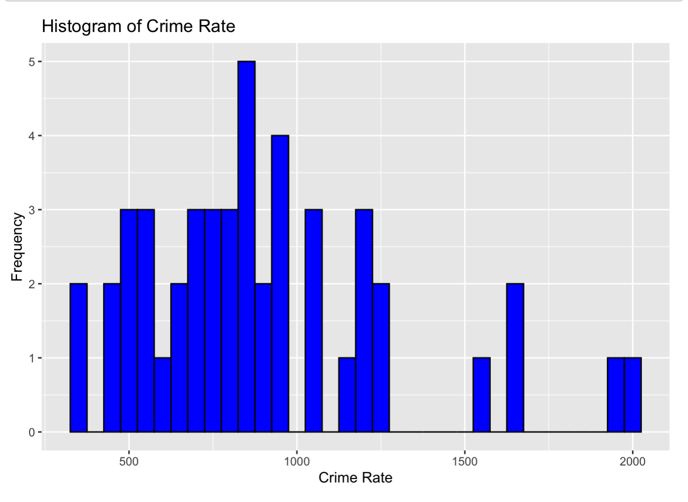
```
new_data <- data.frame(
  M = 14.0,
  So = 0,
  Ed = 10.0,
  Po1 = 12.0,
  Po2 = 15.5,
  LF = 0.640,
  M.F = 94.0,
  Pop = 150,
  NW = 1.1,
  U1 = 0.120,
  U2 = 3.6,
  Wealth = 3200,
  Ineq = 20.1,
  Prob = 0.04,
  Time = 39.0
)
```

EDA & Summary statistics

```
summary(crime_data)
```

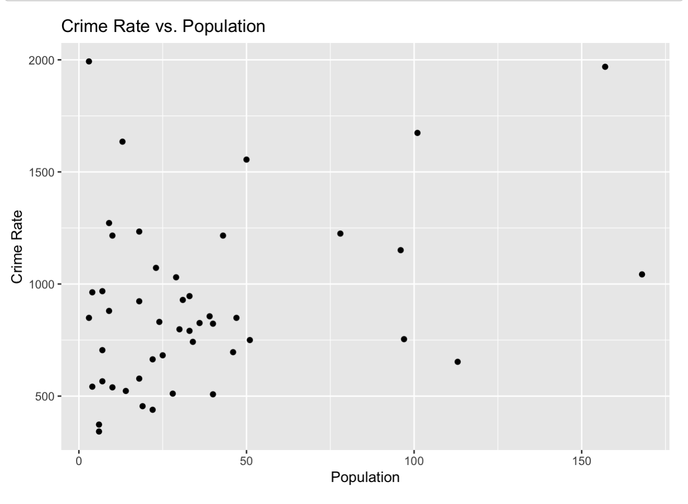
```
##        M               So               Ed              Po1
##  Min.   :11.90   Min.   :0.0000   Min.   : 8.70   Min.   : 4.50
##  1st Qu.:13.00   1st Qu.:0.0000   1st Qu.: 9.75   1st Qu.: 6.25
##  Median :13.60   Median :0.0000   Median :10.80   Median : 7.80
##  Mean   :13.86   Mean   :0.3404   Mean   :10.56   Mean   : 8.50
##  3rd Qu.:14.60   3rd Qu.:1.0000   3rd Qu.:11.45   3rd Qu.:10.45
##  Max.   :17.70   Max.   :1.0000   Max.   :12.20   Max.   :16.60
##       Po2              LF              M.F              Pop
##  Min.   : 4.100   Min.   :0.4800   Min.   : 93.40   Min.   :  3.00
##  1st Qu.: 5.850   1st Qu.:0.5305   1st Qu.: 96.45   1st Qu.: 10.00
##  Median : 7.300   Median :0.5600   Median : 97.70   Median : 25.00
##  Mean   : 8.023   Mean   :0.5612   Mean   : 98.30   Mean   : 36.62
##  3rd Qu.: 9.700   3rd Qu.:0.5930   3rd Qu.: 99.20   3rd Qu.: 41.50
##  Max.   :15.700   Max.   :0.6410   Max.   :107.10   Max.   :168.00
##        NW              U1                U2             Wealth
##  Min.   : 0.20   Min.   :0.07000   Min.   :2.000   Min.   :2880
##  1st Qu.: 2.40   1st Qu.:0.08050   1st Qu.:2.750   1st Qu.:4595
##  Median : 7.60   Median :0.09200   Median :3.400   Median :5370
##  Mean   :10.11   Mean   :0.09547   Mean   :3.398   Mean   :5254
##  3rd Qu.:13.25   3rd Qu.:0.10400   3rd Qu.:3.850   3rd Qu.:5915
##  Max.   :42.30   Max.   :0.14200   Max.   :5.800   Max.   :6890
##       Ineq             Prob              Time           Crime
##  Min.   :12.60   Min.   :0.00690   Min.   :12.20   Min.   : 342.0
##  1st Qu.:16.55   1st Qu.:0.03270   1st Qu.:21.60   1st Qu.: 658.5
##  Median :17.60   Median :0.04210   Median :25.80   Median : 831.0
##  Mean   :19.40   Mean   :0.04709   Mean   :26.60   Mean   : 905.1
##  3rd Qu.:22.75   3rd Qu.:0.05445   3rd Qu.:30.45   3rd Qu.:1057.5
##  Max.   :27.60   Max.   :0.11980   Max.   :44.00   Max.   :1993.0
```

```
ggplot(crime_data, aes(x = Crime)) +
  geom_histogram(binwidth = 50, fill = "blue", color = "black") +
  labs(title = "Histogram of Crime Rate", x = "Crime Rate", y = "Frequency")
```



Scatterplot of crime rate vs. population

```
ggplot(crime_data, aes(x = Pop, y = Crime)) +
  geom_point(fill = "blue", color = "black") +
  labs(title = "Crime Rate vs. Population", x = "Population", y = "Crime Rat
e")
```



Fit a linear regression model

```
model <- lm(Crime ~ ., data = crime_data)
summary(model)
```

```
##
## Call:
## lm(formula = Crime ~ ., data = crime_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -395.74  -98.09   -6.69  112.99  512.67
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.984e+03  1.628e+03  -3.675 0.000893 ***
## M            8.783e+01  4.171e+01   2.106 0.043443 *
## So          -3.803e+00  1.488e+02  -0.026 0.979765
## Ed           1.883e+02  6.209e+01   3.033 0.004861 **
## Po1          1.928e+02  1.061e+02   1.817 0.078892 .
## Po2         -1.094e+02  1.175e+02  -0.931 0.358830
## LF          -6.638e+02  1.470e+03  -0.452 0.654654
## M.F          1.741e+01  2.035e+01   0.855 0.398995
## Pop         -7.330e-01  1.290e+00  -0.568 0.573845
## NW           4.204e+00  6.481e+00   0.649 0.521279
## U1          -5.827e+03  4.210e+03  -1.384 0.176238
## U2           1.678e+02  8.234e+01   2.038 0.050161 .
## Wealth       9.617e-02  1.037e-01   0.928 0.360754
## Ineq         7.067e+01  2.272e+01   3.111 0.003983 **
## Prob        -4.855e+03  2.272e+03  -2.137 0.040627 *
## Time        -3.479e+00  7.165e+00  -0.486 0.630708
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 209.1 on 31 degrees of freedom
## Multiple R-squared:  0.8031, Adjusted R-squared:  0.7078
## F-statistic: 8.429 on 15 and 31 DF,  p-value: 3.539e-07
```

Predict the crime rate for the new data

```
predicted_crime_rate <- predict(model, newdata = new_data)
cat("Predicted Crime Rate:", predicted_crime_rate, "\n")
```

```
## Predicted Crime Rate: 155.4349
```

Now that we have the predicted crime rate, lets check the range of the crime datapoints

```
range(crime_data$Crime)
```

```
## [1]  342 1993
```

As stated in the homework question, since this dataset doesn't have very many data points, it is prone to overfitting. This can be one reason why our predicted number is way below the range of the data.To improve our model we can filter out some of the coefficients that are not significant. I will use the p-value to determine this, and I will remove and variables with coefficients that have a p-value of 0.05 or more.

Get the coefficients and their p-values from the summary, filter coefficients with p-values less than or equal to 0.05, create a data frame from the significant coefficients

```
coef_summary <- summary(model)$coef
significant_coeffs <- coef_summary[coef_summary[, 4] <= 0.05, 4]
significant_coeffs_df <- data.frame(P_Value = significant_coeffs)
print(significant_coeffs_df)
```

```
##                  P_Value
## (Intercept) 0.0008929887
## M           0.0434433942
## Ed          0.0048614327
## Ineq        0.0039831365
## Prob        0.0406269260
```

Here we can see that these are the only coefficients that have a p-value less than 0.5. Now I will create another model with them.

```
model_2<-lm(Crime~M+Ed+Ineq+Prob,data=crime_data,x=TRUE,y=TRUE)
summary(model_2)
```

```
##
## Call:
## lm(formula = Crime ~ M + Ed + Ineq + Prob, data = crime_data,
##     x = TRUE, y = TRUE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -532.97 -254.03  -55.72  137.80  960.21
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1339.35    1247.01  -1.074  0.28893
## M              35.97      53.39   0.674  0.50417
## Ed            148.61      71.92   2.066  0.04499 *
## Ineq           26.87      22.77   1.180  0.24458
## Prob        -7331.92    2560.27  -2.864  0.00651 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 347.5 on 42 degrees of freedom
## Multiple R-squared:  0.2629, Adjusted R-squared:  0.1927
## F-statistic: 3.745 on 4 and 42 DF,  p-value: 0.01077
```

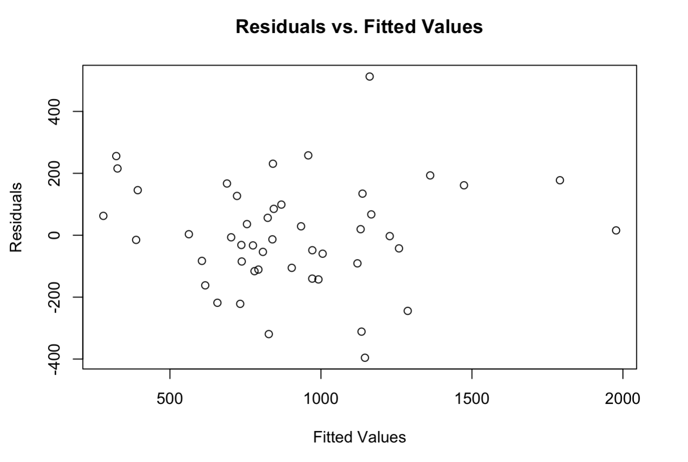
```
predicted_crime_rate <- predict(model_2,new_data)
cat("Predicted Crime Rate:", predicted_crime_rate, "\n")
```

```
## Predicted Crime Rate: 897.2307
```

Our value is now different makes more sense AND fits inside our range of crime values. When examining the Adjusted R-squared, a noticeable difference arises in this model. It's important to recognize that a higher Adjusted R-squared in the initial model doesn't necessarily indicate its superiority. In fact, a higher Adjusted R-squared can result from including numerous variables, some of which may not significantly contribute to prediction, leading to an overfit model. Removing variables from the model may reduce the Adjusted R-squared but doesn't provide insights into the model's overall quality. Moreover, our selection of variables with small p-values doesn't imply that other variables lack significance. Occasionally, when two variables exhibit a strong correlation, the model may designate one as a vital predictor, while the other might receive a higher p-value. In our analysis, we observed that the first model had a higher R-squared (R2) value compared to our second model. However, it's crucial to understand that a higher R2 does not necessarily equate to a better model. R-squared measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model.

We can also take a look at the residuals for both our models:

```
residuals <- resid(model)
plot(fitted(model), residuals, main = "Residuals vs. Fitted Values",
     xlab = "Fitted Values", ylab = "Residuals")
```

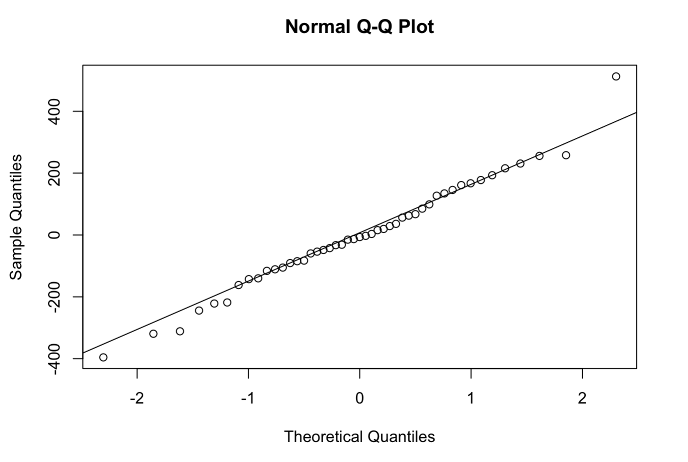


Calculate SSE & Visualization with QQ plot

```
SSE <- sum(residuals^2)
SSE
```

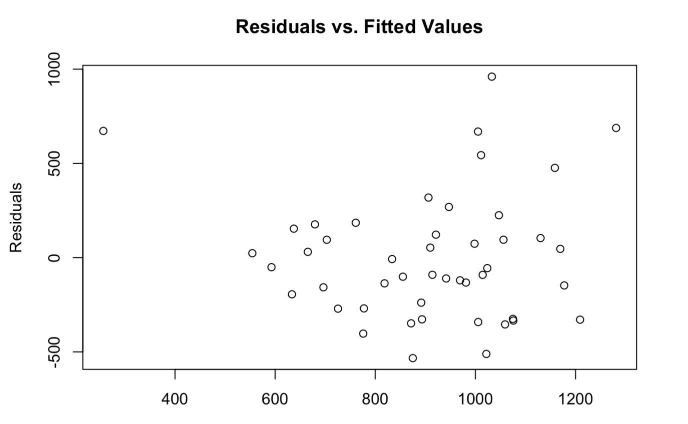
```
## [1] 1354946
```

```
qqnorm(residuals)
qqline(residuals)
```



Model 2

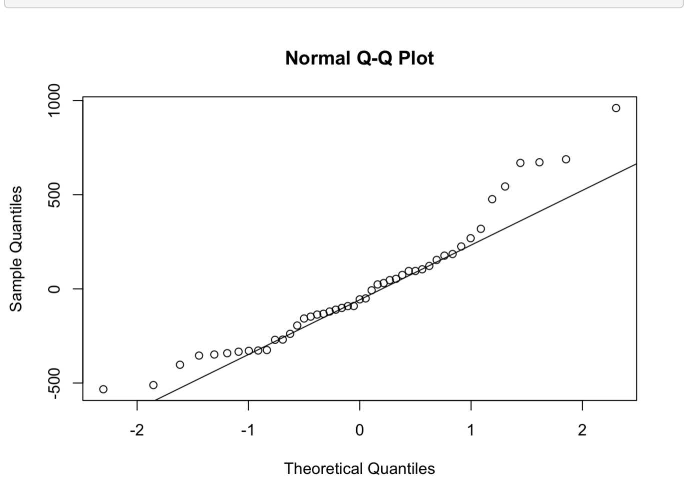
```
residuals <- resid(model_2)
plot(fitted(model_2), residuals, main = "Residuals vs. Fitted Values",
     xlab = "Fitted Values", ylab = "Residuals")
```




```
SSE <- sum(residuals^2)
SSE
```

```
## [1] 5071868
```

```
qqnorm(residuals)
qqline(residuals)
```



Based of residuals being scattered all over the place, it indicates that there may be a lack of linearity in the linear regression model. In a well-fitted linear regression model, you would expect the residuals to be randomly scattered around zero, with no discernible pattern. However, when the residuals are scattered in a non-random or unpredictable manner, it suggests that the relationship between the dependent variable and the independent variables may not be adequately captured by a linear model.

The SSE metric quantifies the overall goodness of fit of the model to the data, with lower values indicating a better fit. It is generally true for many regression problems, especially when the underlying assumptions of linear regression are met. In such cases, a lower Sum of Squared Errors (SSE) indeed indicates a better fit because it signifies that the model's predictions are closer to the observed data points, suggesting a higher level of explained variability. However, there are situations where this may not hold true, and higher SSE values might still be associated with a better model or a more appropriate fit. This could be due to the model complexity, outliers, heteroscedasticity, or other model assumptions.