hw5_glm

HG

2022-09-27

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.

Answer: An example of when to use linear regression could be monitoring sales vs money spent inventory for a restaurant. To an extent, it makes sense that as inventory, so might sales and vice versa. Predictors I would use would be amount of money spent on inventory, number of employees, and even gross domestic product.

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

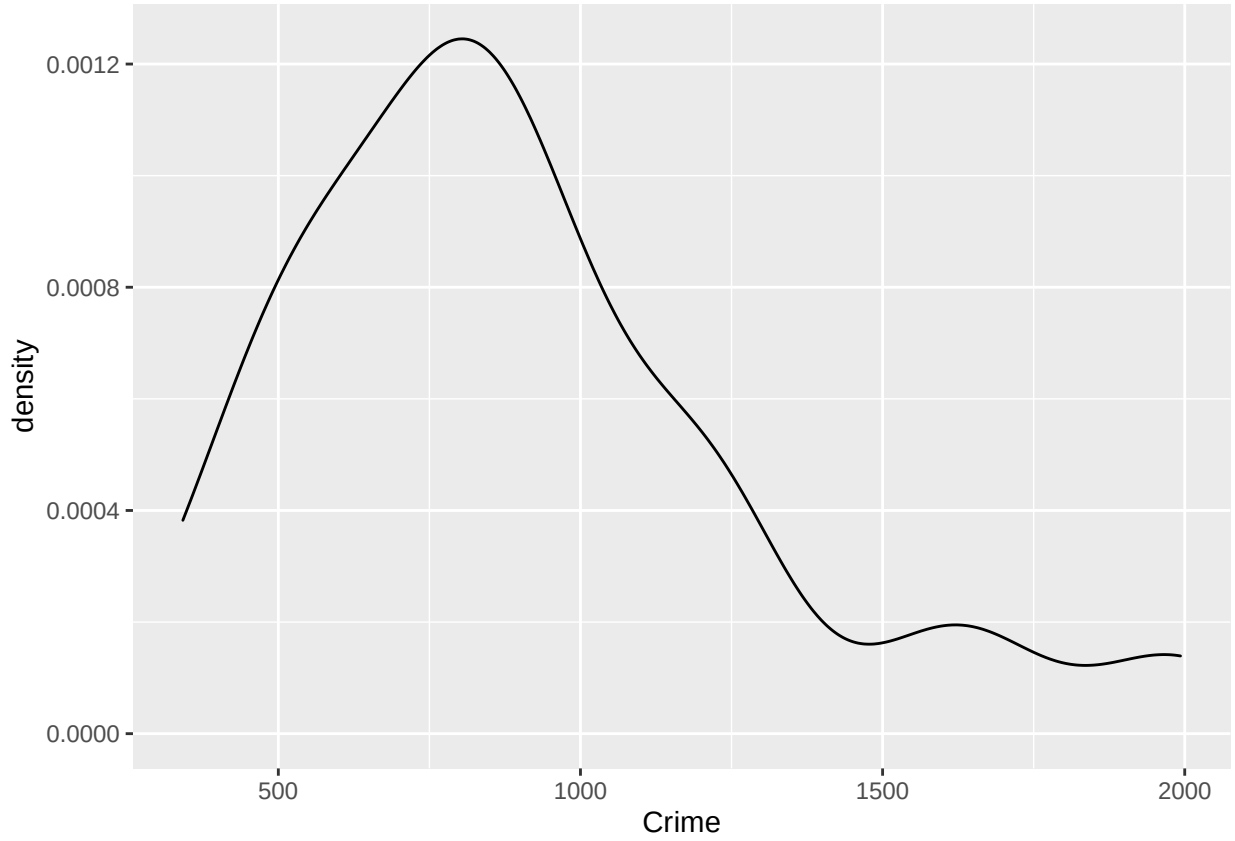
Show your model (factors used and their coefficients), the software output, and the quality of fit.

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.

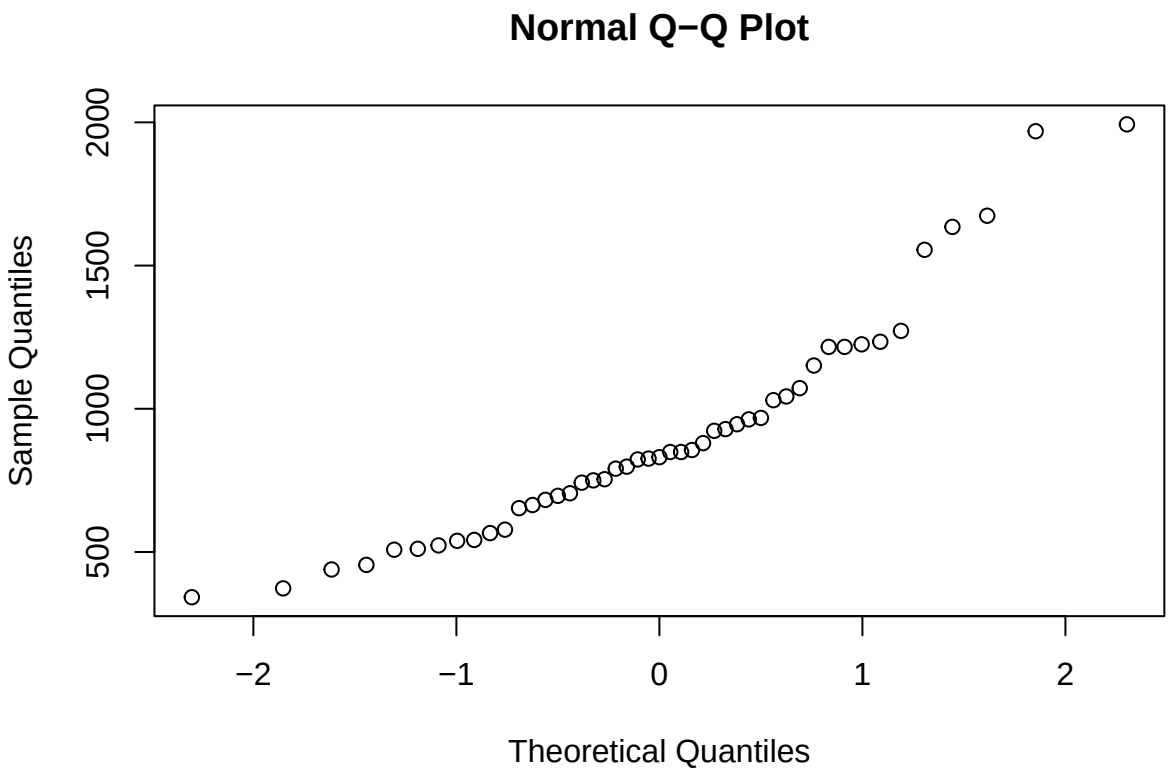
```
rm(list =ls())

library(ggplot2)

setwd("C:/Users/hkgha/OneDrive/Documents2/GATech/ISYE6501/hw5-SP22")
data <- read.table("data 8.2/uscrime.txt",stringsAsFactors = FALSE, header=TRUE)
head(data)
```

```
##      M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq     Prob
## 1 15.1  1  9.1  5.8  5.6 0.510  95.0  33 30.1 0.108 4.1   3940 26.1 0.084602
## 2 14.3  0 11.3 10.3  9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4 0.029599
## 3 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180 25.0 0.083401
## 4 13.6  0 12.1 14.9 14.1 0.577  99.4 157  8.0 0.102 3.9   6730 16.7 0.015801
## 5 14.1  0 12.1 10.9 10.1 0.591  98.5  18  3.0 0.091 2.0   5780 17.4 0.041399
## 6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6 0.034201
##      Time Crime
## 1 26.2011   791
## 2 25.2999  1635
## 3 24.3006   578
## 4 29.9012  1969
## 5 21.2998  1234
## 6 20.9995   682
```

```
ggplot(data=data,aes(x=Crime)) + geom_density()
```

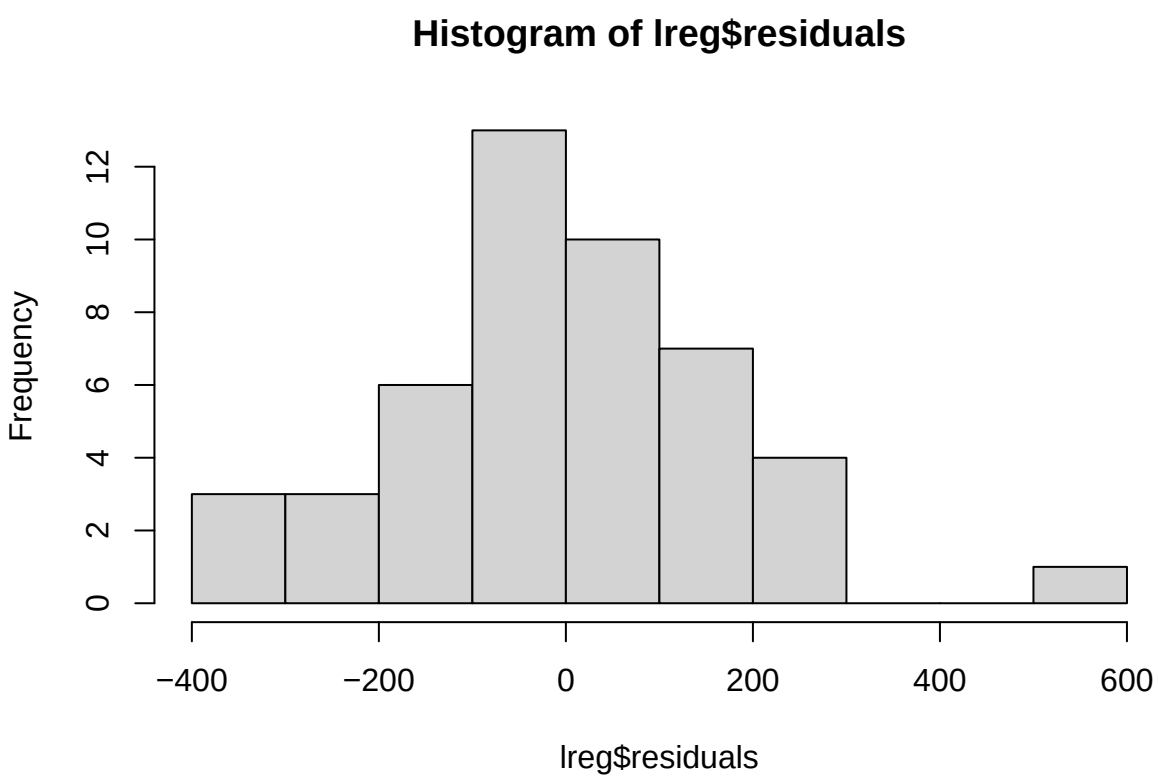



```
qqnorm(data$Crime)
```



I used glm funciton to run the linear regression model on crime data.I used all the predictors with residual value of 1355000 and an intercept value of -5.984e+03

```
lreg <- glm(Crime ~.,data=data,family="gaussian")
hist(lreg$residuals)
```




```
lreg
```

```
##
## Call:  glm(formula = Crime ~ ., family = "gaussian", data = data)
##
## Coefficients:
## (Intercept)            M           So           Ed          Po1          Po2
##  -5.984e+03    8.783e+01   -3.803e+00    1.883e+02    1.928e+02   -1.094e+02
##          LF          M.F          Pop           NW           U1           U2
##  -6.638e+02    1.741e+01   -7.330e-01    4.204e+00   -5.827e+03    1.678e+02
##      Wealth         Ineq         Prob         Time
##   9.617e-02    7.067e+01   -4.855e+03   -3.479e+00
##
## Degrees of Freedom: 46 Total (i.e. Null);  31 Residual
## Null Deviance:       6881000
## Residual Deviance: 1355000   AIC: 650
```

I modeled with a restricted predictors and resulted with Residual Deviance 2488000 and an intercept -5.957e+03 1

```
lreg2 <- glm(Crime ~ M + So + Ed + Pop + Po2+LF + M.F + Wealth +Time ,data = data,family ="gaussian")
lreg2
```

```
##
## Call:  glm(formula = Crime ~ M + So + Ed + Pop + Po2 + LF + M.F + Wealth +
##     Time, family = "gaussian", data = data)
##
## Coefficients:
```

```
## (Intercept)            M           So           Ed          Pop          Po2
##  -5.957e+03    5.916e+01    2.182e+02    6.427e+01    7.254e-01    1.148e+02
##          LF          M.F       Wealth         Time
##   8.445e+02    3.972e+01   -6.272e-02    1.100e+01
##
## Degrees of Freedom: 46 Total (i.e. Null);  37 Residual
## Null Deviance:       6881000
## Residual Deviance: 2488000   AIC: 666.6
```

```
summary(lreg2)
```

```
##
## Call:
## glm(formula = Crime ~ M + So + Ed + Pop + Po2 + LF + M.F + Wealth +
##     Time, family = "gaussian", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -485.98  -202.70    19.95   147.95   574.12
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.957e+03  1.706e+03  -3.491 0.001261 **
## M            5.916e+01  4.637e+01   1.276 0.210016
## So           2.182e+02  1.300e+02   1.679 0.101661
## Ed           6.427e+01  6.797e+01   0.946 0.350499
## Pop          7.254e-01  1.491e+00   0.486 0.629537
## Po2          1.148e+02  2.648e+01   4.338 0.000107 ***
## LF           8.445e+02  1.314e+03   0.643 0.524467
## M.F          3.972e+01  1.802e+01   2.204 0.033813 *
## Wealth      -6.272e-02  9.637e-02  -0.651 0.519139
## Time         1.100e+01  6.976e+00   1.576 0.123460
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 67230.49)
##
##     Null deviance: 6880928  on 46  degrees of freedom
## Residual deviance: 2487528  on 37  degrees of freedom
## AIC: 666.58
##
## Number of Fisher Scoring iterations: 2
```