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:
You could use a linear regression model to predict a person's salary. A few of the predictors you could use include:

1. Education level - Higher education levels could be associated with higher salaries.
2. Years of experience - More years of experience could be associated with higher salaries.
3. Job title - Different job titles could have different salary ranges.
4. Geographic location - Salaries can vary significantly by location
5. Industry - Some industries may pay higher salaries than others.

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.

ANSWER:
First I loaded the data set into R and created a model using all of the variables as predictors.

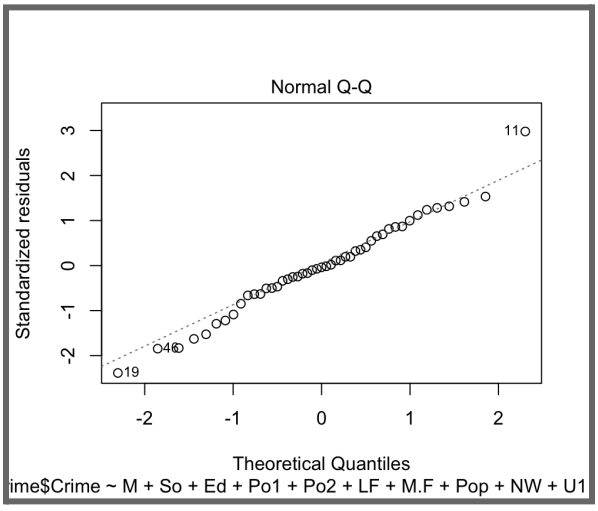
```
crime <- read.table("/Users/[redacted]/Desktop/GATech/Intro_To_Analytics_Modeling/hw5-SP22-1/uscrime.txt", header = TRUE)
head(crime)

model <- lm(formula = crime$Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 + Wealth + Ineq + Prob + Time, data = crime)
```

Next, I evaluated the summary of the model to determine which coefficients were statistically significant. I noticed that M, Ed, Ineq, and Prob were significant. I also noted that Po1 and U2 were very close to being significant. The r-squared value was .80 so approximately 80% of the data could be explained by this model. I did notice that the coefficients

```
Call:
lm(formula = crime$Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F +
    Pop + NW + U1 + U2 + Wealth + Ineq + Prob + Time, data = crime)

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
```

My immediate thought was that I could create a new model using only the predictors that were significant or close to significant and see if that improved the accuracy of the model. Before I did that, though, I wanted to evaluate the error plots to ensure that they were all showing that this model was appropriate.

Below is the normal QQ plot, which seems to show that the errors are normally distributed as they largely follow a straight line

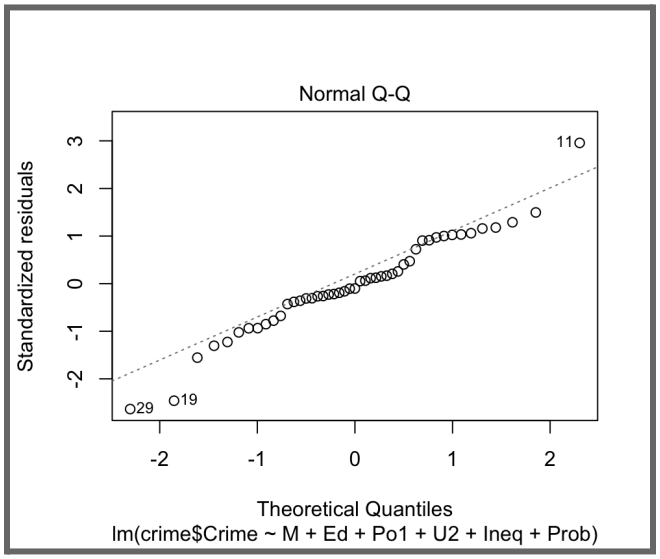


From this plot, I am fairly certain that my initial model will work, so I uploaded the test data point and ran the model against it to see the expected crime for that point. It was 155.4348, which is quite low compared to the crime values in the data set.

```
test_point <- data.frame(M=14.0, So =0, Ed = 10.0, Po1 = 12.0, Po2 = 15.5, LF = .640,
                         M.F = 94.0, Pop = 150, NW = 1.1, U1 = .120, U2 = 3.6,
                         Wealth = 3200, Ineq = 20.1, Prob = 0.040, Time = 39.0)

pred_model <- predict(model, test_point)
pred_model
```

For the last part of my analysis I wanted to recreate the model using only the variables that were significant or close to significant. I created a similar QQ plot for this new model and confirmed that the residuals largely fell along the line

```
model2 <- lm(formula = crime$Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data = crime)
plot(model2)
```



I then wanted to look at the summary of the model to see the coefficients, p values, and the r-squared. I noticed that this new model had very strong p-values for the predictors, a slightly worse r-squared, but a slightly better adjusted r-squared.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5040.50     899.84  -5.602 1.72e-06 ***
M             105.02      33.30   3.154  0.00305 **
Ed            196.47      44.75   4.390 8.07e-05 ***
Po1           115.02      13.75   8.363 2.56e-10 ***
U2             89.37      40.91   2.185  0.03483 *
Ineq           67.65      13.94   4.855 1.88e-05 ***
Prob        -3801.84    1528.10  -2.488  0.01711 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 200.7 on 40 degrees of freedom
Multiple R-squared:  0.7659,    Adjusted R-squared:  0.7307
F-statistic: 21.81 on 6 and 40 DF,  p-value: 3.418e-11
```

Also, notably, when i tested the test data point, my predicted crime rate was 1,304, which was more in line with the actual crime rates for other data points in the crime data set.