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 8.1

At my work, we analyzed the root cause of production failure based on number of parameters in order to improve the process and yield rate. Linear regression model could be applied to estimate the outcome of production (product dimension) by using input quantity, temperature, humidity, machine settings, etc. as predictors from previous production datasets.

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 8.2

```
# Clear global environment, load and preview dataset
> rm(list=ls())
> df1 <- read.table("/Users/…/uscrime.txt", header = T, stringsAsFactors = F)
> head(df1)
    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
```

Fit the linear regression model using crime as response and all other variables as predictors. P-value from the summary of the model suggests there are multiple predictors whose P-value fall below .1 significant level. Notice the R^2 value is 0.8031 with 15 factors, indicating possibilities of overfitting.

```
> model1 <- lm(Crime~., data = df1)
> summary(model1)

Call:
lm(formula = Crime ~ ., data = df1)

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
```

After plotting the test points against the previous model, result suggests the number of crime for these test points is around 155. However, as previous summary of the model suggests, some predictors are insignificant and may affect the estimated crime value.

```
> df <- 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)
> pred1 <- predict(model1, df)
> pred1
     1
155.4349
```

Model was refitted based on predictors whose p-value was below .1 threshold. Result from the new model suggests the number of crime for the same test points is around 1304, which seems more reasonable comparing to the result(155) from previous model. The R^2 value is 0.7659, which is not significantly different from the previous model as well. However, since P-value alone may not be the best measure for variable selection, cross validation will also be performed for variable selection.

```
> model2 <- lm(Crime~M+Ed+Po1+U2+Ineq+Prob, data = df1)
> summary(model2)

Call:
lm(formula = Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, data = df1)

Residuals:
   Min     1Q  Median     3Q    Max
-470.68  -78.41  -19.68  133.12  556.23

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

> pred2 <- predict(model2, df)
> pred2
     1
1304.245
```

Cross validation is performed using train command in the package caret by splitting the dataset into 5-fold, and test the model with 15 predictors. The R^2 value for model with 15 predictors is 0.4607 which is significantly lower than the R^2 value(0.8031) for linear regression model with identical predictors. Result suggests overfitting exists and model with 15 predictors may not be the best option for this dataset.

```
> install.packages("caret")
> library(caret)
> train1 <- trainControl(method = 'cv', number = 5)
> model1_cv <- train(Crime~., data = df1, method = "lm", trControl = train1)
> model1_cv
Linear Regression

47 samples
15 predictors

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 37, 38, 38, 37, 38
Resampling results:

  RMSE      Rsquared   MAE
  300.6023  0.4606907  234.7824

Tuning parameter 'intercept' was held constant at a value of TRUE
```

5-fold cross validation is also performed on model with 6 predictors. The R^2 value is 0.6031 which is still lower than the R^2 value(0.7659) for linear regression model with identical predictors. Result suggests overfitting still exists for model with 6 predictors for this dataset, but much less comparing to model with 15 predictors.

```
> model2_cv <- train(Crime~M+Ed+Po1+U2+Ineq+Prob, data = df1, method = "lm", trControl = train1)
> model2_cv
Linear Regression

47 samples
 6 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 38, 37, 39, 38, 36
Resampling results:

  RMSE      Rsquared   MAE
  219.9196  0.6030733  169.1969

Tuning parameter 'intercept' was held constant at a value of TRUE
```