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.

I am currently training for a half marathon in March, but not doing a very good job of it. I would think that regression would be good to help me predict how much I need to train to finish the half marathon under my goal. If I used past race finish times and their associated training runs, I could create a regression model to determine my probable finish time. Predictors would include total number of training runs as well as the distance, pace, and elevation change of each of those runs. I could also tell by the coefficients what predictors (distance, pace, or elevation change) seem to have the biggest impact on my finishing time.

Question 8.2

Using crime data from http://www.statsci.org/data/general/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:

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.

I first plotted the 15 predictors vs the crime value to see if I could get a better sense of the data and correlations. Some assumptions can be made, such as So and Time not showing a good correlation, but all together there isn't anything that is super obvious to me. These plots are shown in Figure 1.

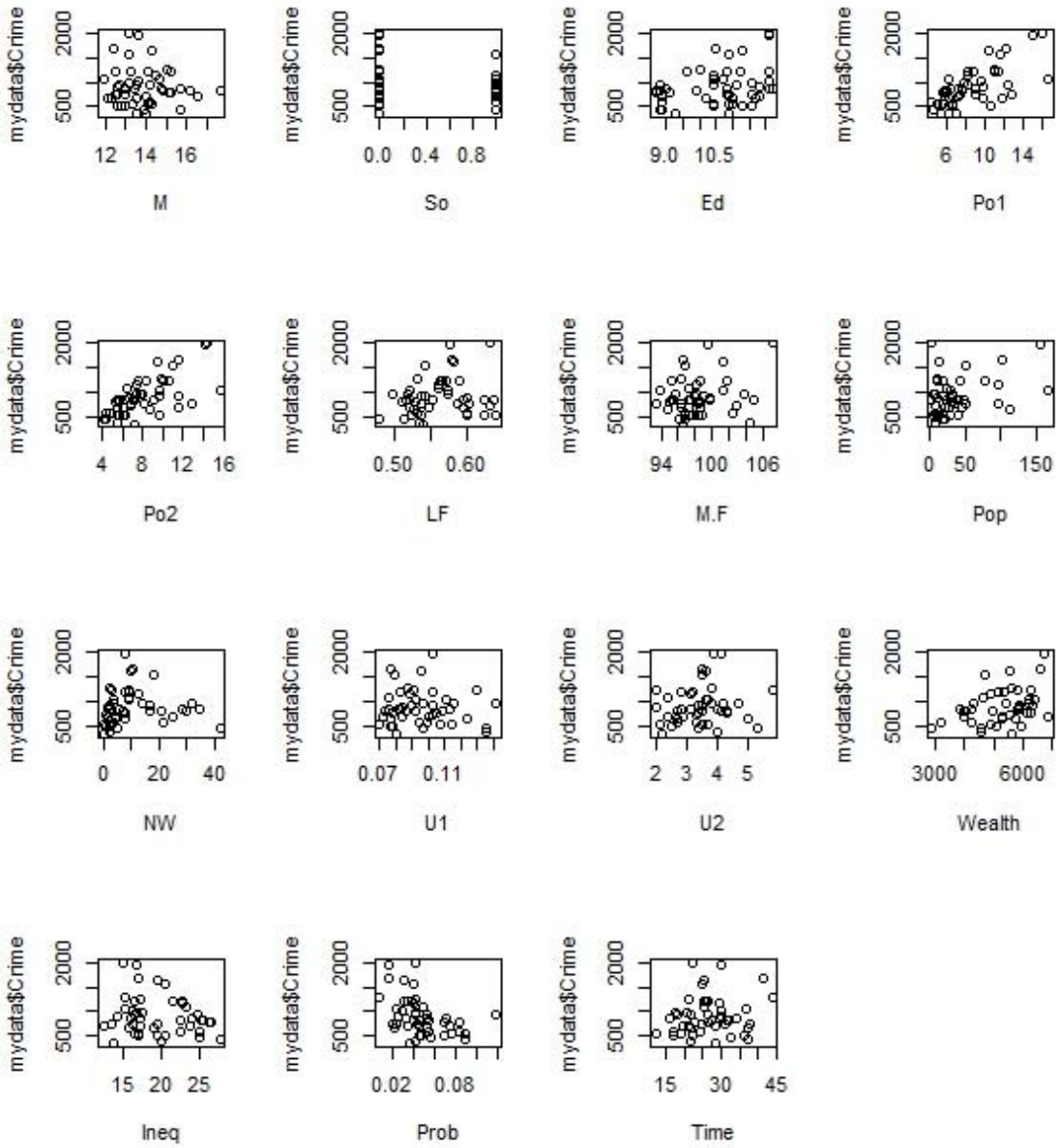

Figure 1. Predictors vs. Crime Data

Using the linear regression function lm() from the R stats package with all of the provided data gives the results shown in Figure 2, where the "Estimate" column values are the coefficients of the regression model for each predictor and the "Pr(>|t|)" column is the p-value, showing the predictors importance to the model. Using the model on the given point data, we calculate a crime rate of 155, which is extremely low.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.98e+03   1.63e+03   -3.68  0.00089 ***
M            8.78e+01   4.17e+01    2.11  0.04344 *
So          -3.80e+00   1.49e+02   -0.03  0.97977
Ed           1.88e+02   6.21e+01    3.03  0.00486 **
Po1          1.93e+02   1.06e+02    1.82  0.07889 .
Po2         -1.09e+02   1.17e+02   -0.93  0.35883
LF          -6.64e+02   1.47e+03   -0.45  0.65465
M.F          1.74e+01   2.04e+01    0.86  0.39900
Pop         -7.33e-01   1.29e+00   -0.57  0.57385
NW           4.20e+00   6.48e+00    0.65  0.52128
U1          -5.83e+03   4.21e+03   -1.38  0.17624
U2           1.68e+02   8.23e+01    2.04  0.05016 .
Wealth       9.62e-02   1.04e-01    0.93  0.36075
Ineq         7.07e+01   2.27e+01    3.11  0.00398 **
Prob        -4.86e+03   2.27e+03   -2.14  0.04063 *
Time        -3.48e+00   7.17e+00   -0.49  0.63071
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 209 on 31 degrees of freedom
Multiple R-squared:  0.803,     Adjusted R-squared:  0.708
F-statistic: 8.43 on 15 and 31 DF,  p-value: 3.54e-07
```

Figure 2. Model 1 Output Summary

From the summary we can see a sizable difference between the R-squared value and the adjust R-squared value, showing that the model is overfit as the adjusted R-squared penalizes for having too many predictors. We can also see that there are many predictors with large p-values, which shows that they have little correlation with the crime rate. Removing any predictors with a p-value greater than .8, which was chosen since Po1 was listed by the software to be a significant predictor, yields the below function:

Crime ~ M + Ed + Po1 + U2 + Ineq + Prob

Re-running the lm() with the new reduced function yields the summary shown in Figure 3 and running with the data point yields a predicted crime rate of 1304, which is much more realistic. We can see that all of the p-values are low and that the difference between the R-squared value and the adjust R-squared value has reduced significantly.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5040.5      899.8   -5.60  1.7e-06 ***
M              105.0       33.3    3.15   0.0031 **
Ed             196.5       44.8    4.39  8.1e-05 ***
Po1            115.0       13.8    8.36  2.6e-10 ***
U2              89.4       40.9    2.18   0.0348 *
Ineq            67.7       13.9    4.85  1.9e-05 ***
Prob         -3801.8     1528.1   -2.49   0.0171 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 201 on 40 degrees of freedom
Multiple R-squared:  0.766,     Adjusted R-squared:  0.731
F-statistic: 21.8 on 6 and 40 DF,  p-value: 3.42e-11
```

Figure 3. Model 2 Output Summary

To check the quality of the fit of the two models, we can compare the R^2 values, 0.803 vs 0.766. At first glance, model 1 looks superior, however, if we use the cross-validation regression function cv.lm() with 5 folds and calculate the R^2 values, we get 0.413 for model 1 and 0.638 for model 2. These values show that model two is the better overall model and is less overfit than model 1, but the large difference between model 2's R^2 values show that model 2 is still overfit and the original R^2 value of 0.766 was very optimistic.