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.

Response

I have worked in the renewable energy industry as a data scientist. One project that I have worked on which utilized regression involved creating a model to predict which homes are likely candidates for photovoltaic (solar) panel installation. Some predictors that determine the likelihood of PV installation include: household income, type of home (detached, townhouse, etc.), age of occupants, geographic location, orientation (cardinal direction) of the roof of the home, proximity to other PV installations, and average sunshine duration (how sunny the area is where is the home is located). Based upon these factors, a regression model can fairly accurately predict the likelihood of a future PV installation for a given home.

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:

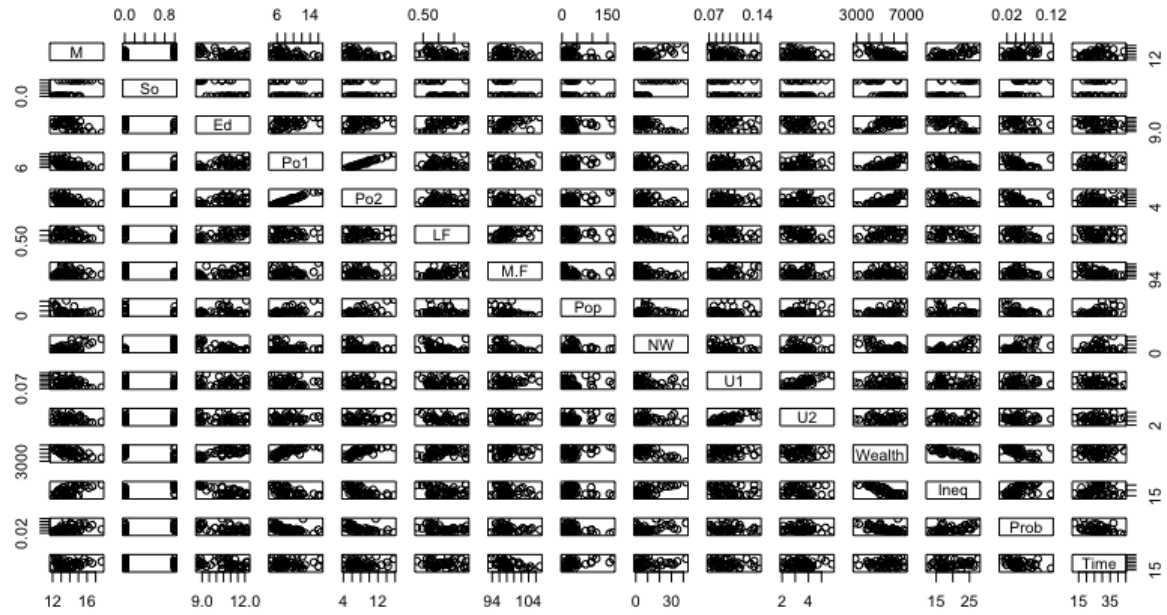
M = 14.0 So = 0 Ed = 10.0 Po1 = 12.0 Po2 = 15.5 LF = 0.640 M.F = 94.0 Pop = 150 NW = 1.1 U1 = 0.120 U2 = 3.6 Wealth = 3200 Ineq = 20.1 Prob = 0.04 Time = 39.0

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

(Please note that my R code for this question is contained in the file 'homework_5_Q8.2.R', and if you wish to run the code, you will need to change line 5 of my R code to your local directory which contains the 'temps.txt' data)

Response

To solve this problem, I implemented linear regression via R's built-in *lm* function. To begin, I read in the us crime data and saved it a as a variable called *crime_data*. *crime_data* contains 15 predictor variables and 1 response variable called *Crime*. *crime_data* contains *47* data points. To start analyzing this problem, I began by visualizing the data. I first created a figure in which I plotted each of the 15 predictor variables against every other predictor variable. This was done to visually check for highly correlated variables, which may lead to the issue of multicollinearity. This may be observed in the figure below:

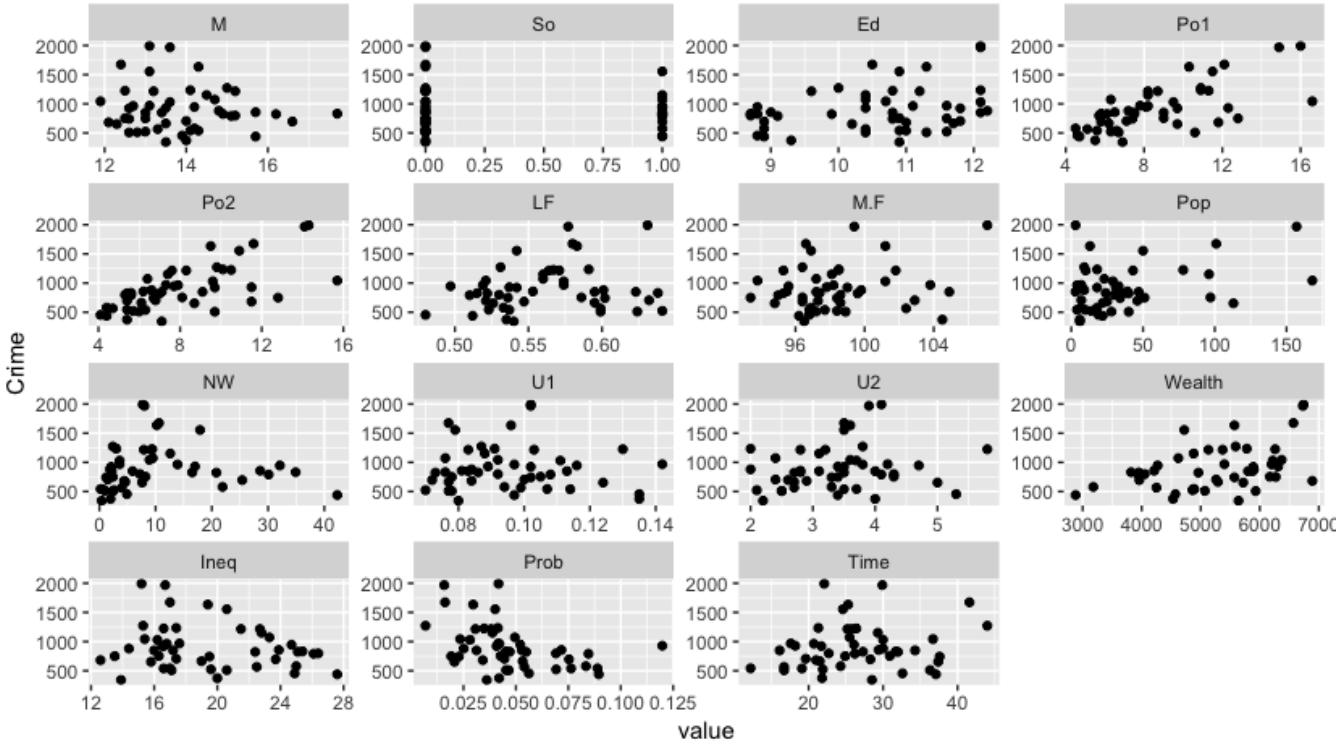


As may be observed in the figure above, it appears as though there may be a strong positive correlation between *Po1* and *Po2*. It also appears as though there may be a fairly strong negative correlation between *Wealth* and *Ineq*. This indicates that we will need to watch out for multicollinearity when we create our linear model later on. To more formally evaluate the relationships between the variables in the *crime_data* dataset, I next found the pearson correlations between each pair of variables. These correlations may be seen in the table below:

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	Wealth	Ineq	Prob	Time	Crime
M	1	0.58435534	-0.5302396	-0.5057369	-0.5131734	-0.1609488	-0.0286799	-0.2806376	0.59319826	-0.2243806	-0.2448434	-0.6700551	0.63921138	0.36111641	0.11451072	-0.0894724
So	0.58435534	1	-0.7027413	-0.3726363	-0.3761675	-0.5054695	-0.3147329	-0.0499183	0.76710262	-0.1724193	0.07169289	-0.6369454	0.73718106	0.53086199	0.06681283	-0.090637
Ed	-0.5302396	-0.7027413	1	0.48295213	0.49940958	0.56117795	0.43691492	-0.0172274	-0.6648819	0.01810345	-0.2156816	0.73599704	-0.7686579	-0.3899229	-0.2539735	0.32283487
Po1	-0.5057369	-0.3726363	0.48295213	1	0.99358648	0.1214932	0.03376027	0.52628358	-0.2137088	-0.0436976	0.18509304	0.78722528	-0.6305003	-0.473247	0.10335774	0.68760446
Po2	-0.5131734	-0.3761675	0.49940958	0.99358648	1	0.1063496	0.0228425	0.5137894	-0.2187682	-0.051712	0.16922422	0.79426205	-0.6481518	-0.4730273	0.07562665	0.66671414
LF	-0.1609488	-0.5054695	0.56117795	0.1214932	0.1063496	1	0.51355879	-0.1236722	-0.3412144	-0.2293997	-0.4207625	0.29463231	-0.2698865	-0.2500861	-0.1236404	0.18886635
M.F	-0.0286799	-0.3147329	0.43691492	0.03376027	0.0228425	0.51355879	1	-0.4106275	-0.3273045	0.3518919	-0.0186917	0.17960864	-0.1670887	-0.0508583	-0.4276974	0.21391426
Pop	-0.2806376	-0.0499183	-0.0172274	0.52628358	0.5137894	-0.1236722	-0.4106275	1	0.09515301	-0.0381199	0.27042159	0.30826271	-0.1262936	-0.3472891	0.46421046	0.33747406
NW	0.59319826	0.76710262	-0.6648819	-0.2137088	-0.2187682	-0.3412144	-0.3273045	0.09515301	1	-0.15645	0.08090829	-0.5901071	0.67731286	0.42805915	0.23039841	0.03259884
U1	-0.2243806	-0.1724193	0.01810345	-0.0436976	-0.051712	-0.2293997	0.3518919	-0.0381199	-0.15645	1	0.74592482	0.0448572	-0.0638322	-0.007469	-0.1698528	-0.0504779
U2	-0.2448434	0.07169289	-0.2156816	0.18509304	0.16922422	-0.4207625	-0.0186917	0.27042159	0.08090829	0.74592482	1	0.09207166	0.01567818	-0.0615925	0.10135833	0.17732065
Wealth	-0.6700551	-0.6369454	0.73599704	0.78722528	0.79426205	0.29463231	0.17960864	0.30826271	-0.5901071	0.0448572	0.09207166	1	-0.8839973	-0.5553347	0.00064856	0.44131995
Ineq	0.63921138	0.73718106	-0.7686579	-0.6305003	-0.6481518	-0.2698865	-0.1670887	-0.1262936	0.67731286	-0.0638322	0.01567818	-0.8839973	1	0.46532192	0.10182282	-0.1790237
Prob	0.36111641	0.53086199	-0.3899229	-0.473247	-0.4730273	-0.2500861	-0.0508583	-0.3472891	0.42805915	-0.007469	-0.0615925	-0.5553347	0.46532192	1	-0.4362463	-0.4274222
Time	0.11451072	0.06681283	-0.2539735	0.10335774	0.07562665	-0.1236404	-0.4276974	0.46421046	0.23039841	-0.1698528	0.10135833	0.00064856	0.10182282	-0.4362463	1	0.14986606
Crime	-0.0894724	-0.090637	0.32283487	0.68760446	0.66671414	0.18886635	0.21391426	0.33747406	0.03259884	-0.0504779	0.17732065	0.44131995	-0.1790237	-0.4274222	0.14986606	1

By looking at the correlations in the figures above, we can confirm that there is indeed a very strong positive correlation between *Po1* and *Po2* with a value of *0.99359* which indicates a nearly perfect linear relationship between these two predictor variables. We can also see that there is correlation of *-0.8839* between *Wealth* and *Ineq*, which indicates a very strong negative linear relationship between these two variables. This confirms that we will indeed need to consider multicollinearity when including these variables in our model (this will be further addressed below).

Next, since our primary objective is to predict a response for *Crime* in a new city, based upon values we are given for the 15 predictor variables, I plotted each of the 15 predictor variables variables in the *crime_data* dataset against the response variable *Crime.* The results of these plots may be observed in the figure below:



By observing the above figures, we can see that some plots are indicative of a strong relationship, while other plots do not appear to show a relationship between the predictor and response variable. However, to analyze this more rigorously, I first created a linear model in R using the *lm* function and included every predictor variable in the model. The results of this linear model can be seen below:

term	estimate	std.error	t-statistic	p-value
(Intercept)	-5984.2876	1628.31837	-3.6751336	0.00089299
M	87.8301732	41.7138664	2.10553902	0.04344339
So	-3.8034503	148.75514	-0.0255685	0.97976537
Ed	188.324315	62.0883761	3.03316541	0.00486143
Po1	192.804338	106.109676	1.81702882	0.07889198
Po2	-109.42193	117.477536	-0.9314285	0.35882957
LF	-663.82615	1469.72882	-0.4516657	0.65465409
M.F	17.4068555	20.3538427	0.85521225	0.39899533
Pop	-0.7330081	1.28955539	-0.5684193	0.57384523
NW	4.204461	6.48089218	0.64874725	0.52127912
U1	-5827.1027	4210.28904	-1.3840149	0.17623803
U2	167.799672	82.3359552	2.0379878	0.05016128
Wealth	0.09616624	0.10366605	0.92765416	0.36075378
Ineq	70.6720995	22.7165213	3.1110441	0.00398314
Prob	-4855.2658	2272.37462	-2.1366485	0.04062693
Time	-3.4790178	7.16527516	-0.4855386	0.63070844

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

When looking at the chart above, the column titled *p-value* is of particular interest to us. The *p*-value associated with a linear model is defined as the probability, that the null hypothesis, which states that the coefficient (called *estimate* in the table above) is equal to zero (meaning it has no effect). Therefore, coefficients that have a very small *p-value* associated with them are very likely to be meaningful predictor variables. For the purposes of this particular question we will consider any predictor that has a $p\text{-value} \leq 0.1$ associated with it to be meaningful.

The linear model defined above has an R^2 value equal to *0.8031* and an *adjusted* R^2 value equal to *0.708* associated with it. *Adjusted* R^2 values are always less than or equal to R^2 values, but a difference this substantial indicates that overfitting is present in our model, meaning that we have included too many predictor variables. This is not surprising given the number of predictor variables (15) relative to the number of data points in the *crime_data* data set (47). A ratio of 15 predictor variables to 47 data points is very likely to result in overfitting. Additionally, given the large *p-values* associated with some of the predictor variables, as well as the strong correlations between several of the variables, this result is not surprising. Finally, when we apply the coefficients of the model that is generated (defined by the estimate column in the table above), to a new city for which we are trying to predict *Crime,* and which is defined by the following predictor variables: $M = 14.0 \quad So = 0 \quad Ed = 10.0 \quad Po1 = 12.0 \; Po2 = 15.5 \quad LF = 0.640 \quad M.F = 94.0 \; Pop = 150 \; NW = 1.1 \quad U1 = 0.120 \quad U2 = 3.6 \; Wealth = 3200 \; Ineq = 20.1 \; Prob = 0.04 \; Time = 39.0$, then our model predicts that *Crime* will equal 155. This seems to be an unreasonably low estimate.

All of the above factors indicate that a linear model which includes all predictor variables is not capable of accurately forecasting *Crime*, and thus must be adjusted. To correct the model, I chose to create a linear regression model which includes only the terms that had a $p\text{-value} \leq 0.1$ in the above model. As can be seen from the chart above, this means that the model will only include the following predictors: *M, Ed, Po1, U2, Ineq,* and *Prob*.

When I created the new model, which includes only the terms *M, Ed, Po1, U2, Ineq,* and *Prob*, using R's *lm* function I found the following results:

term	estimate	std.error	statistic	p.value
(Intercept)	-5040.505	899.843385	-5.6015358	1.72E-06
M	105.019568	33.2992509	3.15381172	0.00305444
Ed	196.471201	44.7543733	4.38998887	8.07E-05
Po1	115.024191	13.7535885	8.36321305	2.56E-10
U2	89.3660431	40.905669	2.18468602	0.0348313
Ineq	67.6532159	13.935746	4.85465336	1.88E-05
Prob	-3801.8363	1528.09733	-2.4879543	0.01711387

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

When looking at the chart above, we can first notice that every predictor variable has an extremely small *p-value* associated with it. This indicates that every variable included in the model is very likely to be relevant. Additionally, we find an R^2 value of *0.7659*

associated with the model, and an *adjusted* R^2 value of *0.7307* associated with the model. This much smaller difference between the R^2 and *adjusted* R^2 values indicates that a minimal amount of overfitting is present in the model. Additionally, the ratio of 5 predictor variables to 47 data points is more reasonable than the ratio in the first model model. Also, when we apply the model, which is defined by the equation

$$Crime = -5040.5 + 105*M + 196.5*Ed + 115*Po1 + 89.4*U2 + 67.7*Ineq - 3801.8*Prob$$

to the relevant variables contained in the new city defined by the following predictor variables, $M = 14.0 \quad So = 0 \quad Ed = 10.0 \quad Po1 = 12.0 \; Po2 = 15.5 \quad LF = 0.640 \quad M.F = 94.0 \; Pop = 150 \quad NW = 1.1 \quad U1 = 0.120 \quad U2 = 3.6 \; Wealth = 3200 \; Ineq = 20.1 \; Prob = 0.04 \; Time = 39.0$, our model predicts that Crime will be equal to *1304*, which is a seemingly much more reasonable reslut. Finally, it should be pointed that the issue of multicollinearity is not an issue in this model. Since neither of the complete pairs of *Po1* and *Po2* nor *wealth* and *Ineq* are included in the model, nor any other pairs of variables which contain a strong very correlation, we do not need to consider multicollinearity when assessing the forecasting ability this model. Thus, given the above factors, which all indicate that this is a sound model, I feel confident in the prediction that the observed crime in the new city will be equal to *1304*.