Homework 5

2025-09-23

Question 8.1

Let's say someone is trying to lose weight and wants to predict when they will lose 20 pounds. They measure their weight over the span of two months. A linear regression model would be appropriate in this case because I want to see if our weight decreases over time, and try to predict when we will reach our desired weight. In our example the x-axis would be # of weeks since starting, and the y-axis would be weight (in lbs). Some predictors I might use are:

1. How many calories they ate in a given day.
2. How many minutes they exercised in a given day.
3. Time they weighed yourself (I would need to make sure that I weigh myself at a consistent time, before breakfast in the morning for instance.)
4. How much we hydrated (in the short term we can see changes in weight based on how much water we drink. Long term, not so much.)
5. Sleep quality? If we are getting low quality sleep on a regular basis, this can decrease our weight. But I don't think this is as important as the other predictors.

Question 8.2

Methodology

1. My goal is to predict the crime rate (our response variable) using several predictors from our crime dataset.
2. Let's start by fitting a linear regression model on our crime data. I want to use all predictors initially, and we can finetune our model later based on the results.
3. I call a summary function on the model that gives us residuals, and coefficients.
4. Let's now observe our coefficients from our model (each have an estimate, standard error, t value, and probability).
5. Now I want to try to predict the crime rate given certain values of predictors. So we create a test point as seen below with specified values.
6. We run a Q-Q plot on the crime column to check if the datapoints come from a normal distribution (a roughly straight line would be a result of this).
7. I run cross-validation on our lm model with the crime data for 4 folds. I want to know how well our model is performing on new data.
8. I note on my model's results based on what we have learned.

Analysis and Discussion

Before we start, I suspect that the wealth, inequality, and population will be valuable predictors for us. Overfitting is likely because we have few observations. Is the model trained enough to perform well on new observations? We'll have to see.

If we observe our residuals, we see that our median is -6.69, which means that our observed crime rate was on average 6.69 lower than our predicted crime rate. Our ideal situation would obviously be where the median is 0 (model predicted crime is equal to the actual crime) but roughly 0 is a good model. We can also see that the other quartiles show symmetrical data. Quartile 1 is just about -100, and Quartile 3 is just over 100. There is a slight skewness to the right because of the larger maximum and Quart. 3 but it is roughly symmetrical.

Let's move on to our coefficients. The standard error is how much our predicted values deviate from our observed values. The t-value is calculated as estimate/standard error and generally we want our coefficients to have a large t-statistic. The p-value estimates the probability that the coefficient is 0. We can say that if the $p < 0.05$, the coefficient is significant. The other coefficients can be removed from our model. It's important to note that this is a probability, not a certainty. We are guessing these predictors are important (with valid reason, of course.). If we look at the R-squared, we get 0.8031. But does this tell us much? We know that the r-squared increases as we have more predictors. The adjusted R-squared of 0.7078 might be better to use for this reason.

When we predict crime rate using our sample point 1, we get 155.4349. This is quite low compared to the rest of our values. One possible reason why it is due to our model overfitting new data, as the predictors were not so abnormal. How about we use the values for one of our observations to predict the crime rate. We predict a crime rate of 750.42 which is very close to the actual crime rate of 705.

For our q-q plot ideally we would like to see a straight line. We do see a couple of outliers on either end, the middle is normally distributed so overall looks okay. From our CV plot below, all folds perfomed consistently. There wasn't a fold that showed abnormal behavior and overall, our model did a good job of predicting the crime rate.

I think our model performed well given the size of the dataset. Of course, having more datapoints would improve it, but it was able to predict non-outliers quite well. For outliers, it was sensitive to small changes in predictors but this shouldn't come as a surprise to us.

```
rm(list = ls()) # clears your R environment, ensures a clean and reproducible environment.

uscrime <- read.table("Homework5_ISYE6501_Data/uscrime.txt", stringsAsFactors = FALSE, header = TRUE)

lm_uscrime <- lm(Crime ~ ., uscrime) # linear model

summary(lm_uscrime) # quartile distribution, and coeffients.
```

```
##
## Call:
## lm(formula = Crime ~ ., data = uscrime)
##
## 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
```

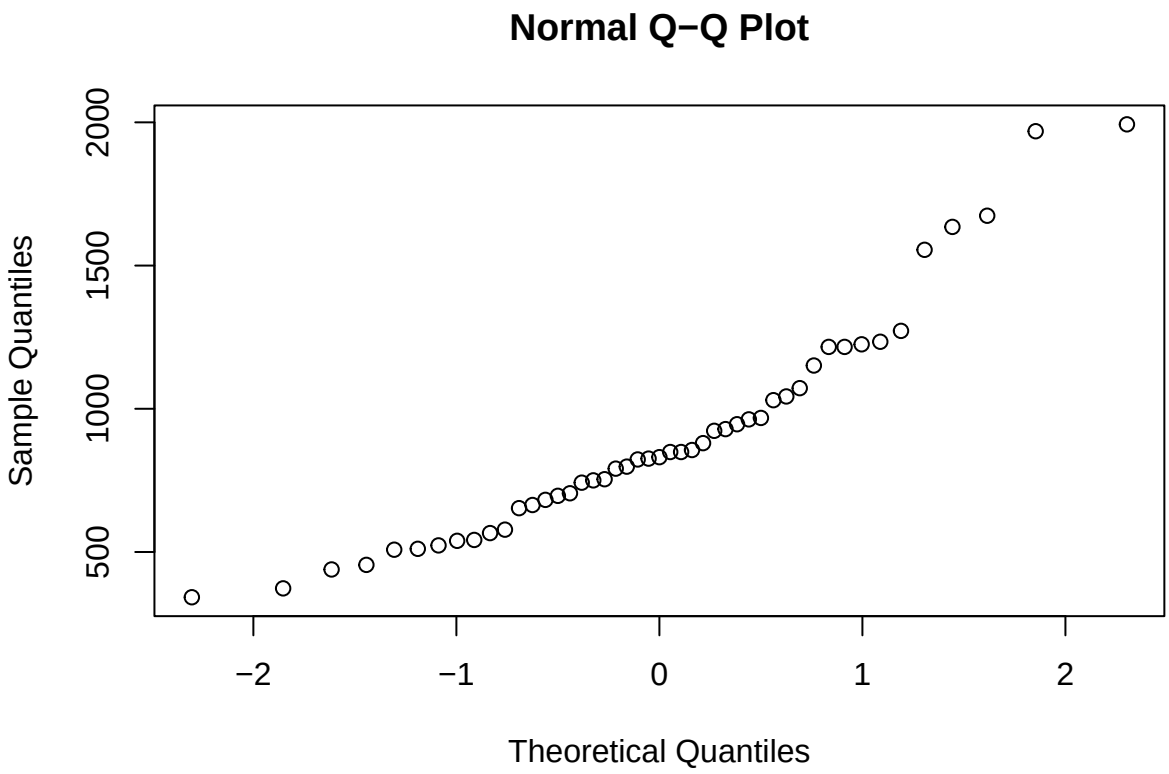
```
test_point <- data.frame(M = 14.0, So = 0, Ed = 10.0, Po1 = 12.0, Po2 = 15.5,
                         LF = 0.640, M.F = 94.0, Pop = 150, NW = 1.1,
                         U1 = 0.120, U2 = 3.6, Wealth = 3200, Ineq = 20.1,
                         Prob = 0.040, Time = 39.0)

test_point2 <- data.frame(M = 14.0, So = 0, Ed = 11.8, Po1 = 7.1, Po2 = 6.8,
                         LF = 0.632, M.F = 102.9, Pop = 7, NW = 1.5,
                         U1 = 0.100, U2 = 2.4, Wealth = 5260, Ineq = 17.4,
                         Prob = 0.044498, Time = 15.5994)

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

```
##        1
## 155.4349
```

```
pred_model2 <- predict(lm_uscrime, test_point2)
pred_model2
```

```
##        1
## 750.4241
```

```
qqnorm(uscrime$Crime)
```

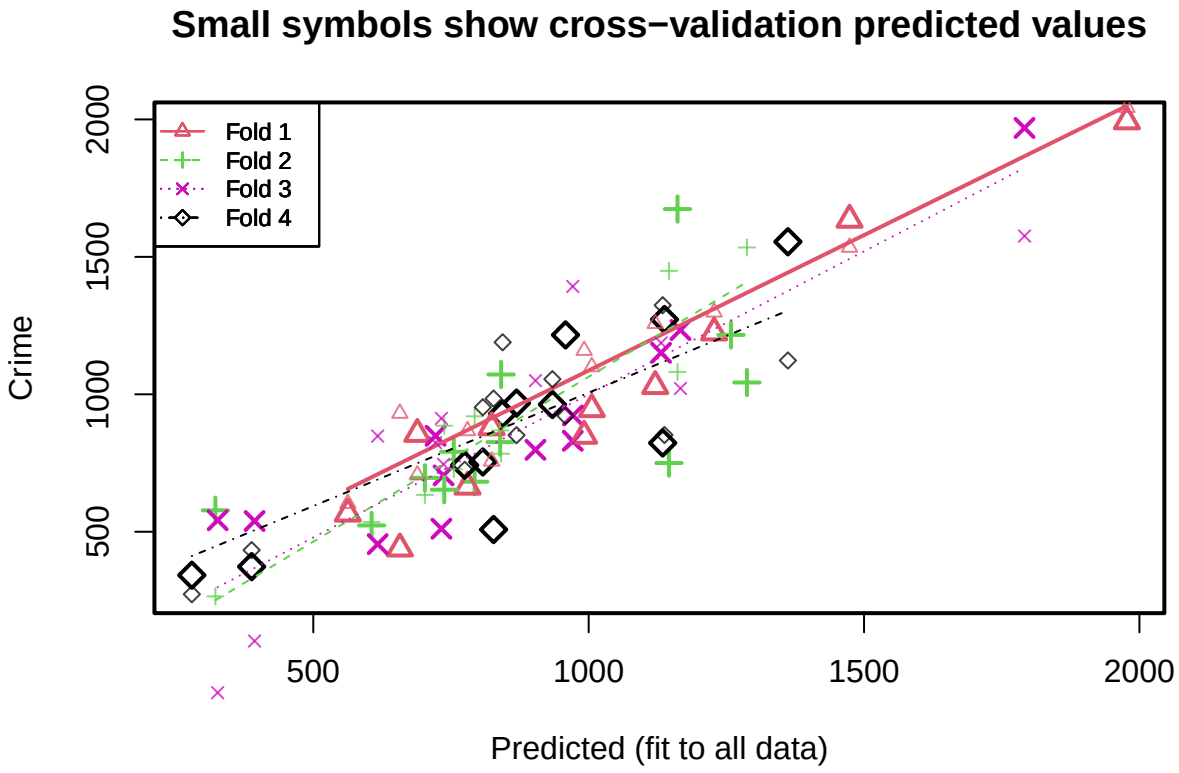



```
#the DAAG function has a cv.lm() function that we can use to get a more accurate measure of the quality
library(DAAG)
```

```
## Warning: package 'DAAG' was built under R version 4.3.3
```

```
set.seed(11)

lm_uscrime_cv <- cv.lm(uscrime, lm_uscrime, m=4)
```

```
## Warning in cv.lm(uscrime, lm_uscrime, m = 4):
##
##  As there is >1 explanatory variable, cross-validation
##  predicted values for a fold are not a linear function
##  of corresponding overall predicted values.  Lines that
##  are shown for the different folds are approximate
```




```
##
## fold 1
## Observations in test set: 11
##                      2        9        14        16         20        22
## Predicted   1473.6764 688.8682  780.0401 1005.6569 1227.83873  657.2092
## cvpred      1534.7127 706.3550  867.2990 1100.2138 1298.87335  930.7865
## Crime       1635.0000 856.0000  664.0000  946.0000 1225.00000  439.0000
## CV residual  100.2873 149.6450 -203.2990 -154.2138  -73.87335 -491.7865
##                     26        38       41        44        47
## Predicted   1977.37067 562.69340 823.7419 1120.8227  991.7629
## cvpred      2043.30084 602.84891 756.5861 1257.3099 1159.4722
## Crime       1993.00000 566.00000 880.0000 1030.0000  849.0000
## CV residual  -50.30084 -36.84891 123.4139 -227.3099 -310.4722
##
## Sum of squares = 512056.6    Mean square = 46550.6    n = 11
##
## fold 2
## Observations in test set: 12
##                      1        3         6        11       19        25
## Predicted   755.03222 322.2615  792.9301 1161.3291 1145.738 605.88241
## cvpred      727.72778 264.9473  919.9553 1081.5238 1449.050 535.14166
## Crime       791.00000 578.0000  682.0000 1674.0000  750.000 523.00000
## CV residual  63.27222 313.0527 -237.9553  592.4762 -699.050 -12.14166
##                     28        29       30        33        35       39
## Predicted   1258.484227 1287.3917 702.6945  840.9992  737.7888 839.2864
## cvpred      1219.776055 1534.1148 634.0349  783.9328  885.6665 868.7428
```

```
## Crime       1216.000000 1043.0000 696.0000 1072.0000  653.0000 826.0000
## CV residual   -3.776055 -491.1148  61.9651  288.0672 -232.6665 -42.7428
##
## Sum of squares = 1382466    Mean square = 115205.5    n = 12
##
## fold 3
## Observations in test set: 12
##                     4        5        10        12        13        15       17
## Predicted   1791.3619 1166.684 736.50802 722.04080  732.6412  903.3541 393.3633
## cvpred      1575.8534 1021.363 745.05052 823.97648  912.3609 1049.9306 102.5037
## Crime       1969.0000 1234.000 705.00000 849.00000  511.0000  798.0000 539.0000
## CV residual  393.1466  212.637 -40.05052  25.02352 -401.3609 -251.9306 436.4963
##                   34        37         40        42        45
## Predicted   971.4558  971.1513 1131.45326 326.33240  616.8983
## cvpred      823.3683 1392.2943 1186.78661 -85.46134  848.0354
## Crime       923.0000  831.0000 1151.00000 542.00000  455.0000
## CV residual  99.6317 -561.2943  -35.78661 627.46134 -393.0354
##
## Sum of squares = 1491541    Mean square = 124295.1    n = 12
##
## fold 4
## Observations in test set: 12
##                      7         8        18        21        23       24
## Predicted    934.16366 1361.7468  843.8072 774.85062  957.9918 868.9805
## cvpred      1055.08192 1123.0151 1189.2623 725.25486  921.6292 851.1620
## Crime        963.00000 1555.0000  929.0000 742.00000 1216.0000 968.0000
## CV residual  -92.08192  431.9849 -260.2623  16.74514  294.3708 116.8380
##                    27        31        32        36        43        46
## Predicted   279.47716 388.03339  807.8167 1137.6171 1134.4172  827.3543
## cvpred      272.65799 433.14096  952.7789  851.8925 1323.9585  984.4175
## Crime       342.00000 373.00000  754.0000 1272.0000  823.0000  508.0000
## CV residual  69.34201 -60.14096 -198.7789  420.1075 -500.9585 -476.4175
##
## Sum of squares = 1065774    Mean square = 88814.49    n = 12
##
## Overall (Sum over all 12 folds)
##       ms
## 94719.93
```