ISYE Homework - Week 3 Markdown 2

2024-06-05

Question 8.2

We'll start off by loading in our data.

```
crime <- read.table("~/Documents/ISYE 6501/week_3_Homework-summer/week 3 data-summer/uscrime.txt", stri
print(head(crime))
```

```
##      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
```

Now, we will run a linear modeling function on our data, regressing all variables on Crime.

We'll print a summary as well, so we can more deeply assess the model itself.

```
full_model <- lm(Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 + Wealth + Ineq + Prob
summary(full_model)
```

```
##
## Call:
## lm(formula = 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
```

Upon first glance, there are many variables that appear not to be statistically significant. We're going to eliminate one of the Po variables simply because I have a hunch they are probably correlated with each other – based on the description of the data, this is the amount that was spent on police per capita in 1959 and 1960, so it would be reasonable to think that together, they might unduly influence our model.

After removing Po2 and a couple other variables with high p-values (Pop, Time), we are left with the following function:

lm(Crime ~ M + So + Ed + Po1 + LF + M.F + NW + U1 + U2 + Wealth + Ineq + Prob, data = crime)

It may be worth commenting here that this process likely seems a bit arbitrary. It feels that way to me, as well. But absent a more sophisticated manner of finding statistically significant variables (as we are presumably about to learn in the coming weeks), this is the process we'll go with for now.

I stopped removing variables here because removing more than that made the p values pretty volatile. Additionally, this is where I ended up with an adjusted R squared value of 0.7221 - the highest I've achieved with my various combinations of variables.

Below is a summary:

```
model1 <- lm(Crime ~ M + So + Ed + Po1 + LF + M.F + NW + U1 + U2 + Wealth + Ineq + Prob, data=crime)
summary(model1)
```

```
##
## Call:
## lm(formula = Crime ~ M + So + Ed + Po1 + LF + M.F + NW + U1 +
##     U2 + Wealth + Ineq + Prob, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -411.97 -103.51    5.39  116.00  450.73
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.834e+03  1.353e+03  -5.052 1.47e-05 ***
```

```
## M            8.967e+01  3.986e+01   2.249 0.031081 *
## So           1.130e+01  1.437e+02   0.079 0.937792
## Ed           1.798e+02  5.905e+01   3.046 0.004464 **
## Po1          9.115e+01  2.109e+01   4.323 0.000127 ***
## LF          -4.261e+02  1.360e+03  -0.313 0.756025
## M.F          2.482e+01  1.745e+01   1.422 0.164076
## NW           2.103e+00  5.902e+00   0.356 0.723830
## U1          -5.772e+03  4.012e+03  -1.439 0.159428
## U2           1.682e+02  7.971e+01   2.111 0.042229 *
## Wealth       8.083e-02  9.973e-02   0.811 0.423255
## Ineq         6.876e+01  2.131e+01   3.227 0.002764 **
## Prob        -3.890e+03  1.797e+03  -2.165 0.037528 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 203.9 on 34 degrees of freedom
## Multiple R-squared:  0.7946, Adjusted R-squared:  0.7221
## F-statistic: 10.96 on 12 and 34 DF,  p-value: 1.856e-08
```

Next, we will observe a new data point and the predicted crime rate for that observation.

We'll run the new observation through both our models - the full model that includes all the variables, and the model with some of the variables removed.

The equation for the full model is: -5984.0 + 87.83(14) + 188.3(10) + 192.8(12.5) -109.4(15.5) - 663.8(0.640) + 17.41(94.0) - 73.3(150) + 4.2(1.1) - 5827(0.12) + 167.8(3.6) + 0.09617(3200) + 70.67(20.1) - 4855(0.04) - 3.479(39.0)

And the equation for the updated model is: -6834 + 89.67(14) + 179.8(10) + 91.15(12.5) - 426.1(0.640) + 24.82(94.0) + 2.103(1.1) - -5772(0.12) + 168.2(3.6) + 0.08083(3200) + 68.76(20.1) - 3890(0.04)

First, let's create our new observation.

```
new_observation <- 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)
```

Now let's make our predictions:

```
predict(model1, new_observation)
```

```
##        1
## 774.4518
```

```r
predict(full_model, new_observation)
```

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

```r
summary(full_model)
```

```
##
## Call:
## lm(formula = 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
```

```r
summary(model1)
```

```
##
## Call:
## lm(formula = Crime ~ M + So + Ed + Po1 + LF + M.F + NW + U1 +
##     U2 + Wealth + Ineq + Prob, data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -411.97 -103.51    5.39  116.00  450.73
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.834e+03  1.353e+03  -5.052 1.47e-05 ***
## M            8.967e+01  3.986e+01   2.249 0.031081 *
## So           1.130e+01  1.437e+02   0.079 0.937792
## Ed           1.798e+02  5.905e+01   3.046 0.004464 **
## Po1          9.115e+01  2.109e+01   4.323 0.000127 ***
## LF          -4.261e+02  1.360e+03  -0.313 0.756025
## M.F          2.482e+01  1.745e+01   1.422 0.164076
## NW           2.103e+00  5.902e+00   0.356 0.723830
## U1          -5.772e+03  4.012e+03  -1.439 0.159428
## U2           1.682e+02  7.971e+01   2.111 0.042229 *
## Wealth       8.083e-02  9.973e-02   0.811 0.423255
## Ineq         6.876e+01  2.131e+01   3.227 0.002764 **
## Prob        -3.890e+03  1.797e+03  -2.165 0.037528 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 203.9 on 34 degrees of freedom
## Multiple R-squared:  0.7946, Adjusted R-squared:  0.7221
## F-statistic: 10.96 on 12 and 34 DF,  p-value: 1.856e-08
```

I want to call attention to the fact that our adjusted R-squared values are very similar, with the full model giving us an adjusted R-squared of 0.707, and the altered model giving us an adjusted R-squared of 0.722.

```
summary(full_model)$adj.r.square
```

```
## [1] 0.7078062
```

```
summary(model1)$adj.r.square
```

```
## [1] 0.7220505
```

Neither of those are the highest adjusted R-squared we've dealt with in this class, but aren't the lowest, either.

However, even though both models are in the same ballpark in terms of their R-squared values, the two models give us vastly different results. The full model predicts a crime rate of 155.43 per 100,000 residents, where the altered model predicts a crime rate of 774.45 per 100,000 residents.

Clearly, in order to achieve a more accurate model, we would want to pursue other methods of variable selection other than simply looking at p-values generated by our lm function in R. As the TA's have mentioned, in the coming modules we will be learning how to take a more methodical approach to variable selection, and hopefully that process will result in more reliable models... and therefore more accurate predictions.

```
predict(full_model, new_observation)
```

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

```r
predict(model1, new_observation)
```

```
##        1
## 774.4518
```