Homework 5

9/2/2022

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

I am currently working with the marketing department of a certain retailer, and one interesting thing to understand would be to see if there is a relation between the revenue and the spending we do on different channels. So in this case, each datapoint would be for a specific point in time (months for example), the response would be the revenue and the features would be:

1. Spending in Printed Media
2. Spending in TV
3. Spending in Radio
4. Spending in Digital Channels (Social Media)

With this we could try and identify if any of those channels have a bigger impact than the other, and try to see if we could predict a bigger revenue by investing more in a specific one.

Question 8.2

Using crime data from http://www.statsci.org/data/general/uscrime.txt (http://www.statsci.org/data/general/uscrime.txt) (file uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html (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. 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.

Answer

We want to try and see if we can predict a specific response variable using several features, and for this we will use a linear regression model. First of all, let's inspect the dataset:

```
## Warning: package 'knitr' was built under R version 4.0.5
```

```
## Warning: package 'jtools' was built under R version 4.0.5
```

```
## Registered S3 methods overwritten by 'tibble':
##   method     from
##   format.tbl pillar
##   print.tbl  pillar
```

```
## Warning: replacing previous import 'vctrs::data_frame' by 'tibble::data_frame'
## when loading 'dplyr'
```

```
## Warning: package 'AICcmodavg' was built under R version 4.0.5
```

```
#loading up the data:
uscrime <- read.delim("uscrime.txt")

#understanding the data:
dim(uscrime)
```

```
## [1] 47 16
```

```
head(uscrime)
```

```
##      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
```

So we have a small dataset, with just 47 observations for 16 variables, all of them numeric. We want to see if we can predict the observed crime rate.That means we should try and build a linear regression model, using Crime as the response variable. Do note that most features are on a scale of 10, except Wealth. To verify if this could affect, we are going to do two models, one with scaled data and one with unscaled data and compare them.

```
uscrime_scaled <- data.frame(scale(uscrime))

#creating simple linear models
no_scale_model <- lm(Crime~.,data=uscrime)
scaled_model <- lm(Crime~.,data=uscrime_scaled)

#analyzing the models
export_summs(no_scale_model, scaled_model, scale = TRUE, error_format = "(p = {p.value})")
```

```
## Registered S3 methods overwritten by 'broom':
##   method            from
##   tidy.glht         jtools
##   tidy.summary.glht jtools
```

	Model 1	Model 2
(Intercept)	906.38 ***	0.00
	(p = 0.00)	(p = 0.98)
M	110.38 *	0.29 *
	(p = 0.04)	(p = 0.04)
So	-3.80	-0.01
	(p = 0.98)	(p = 0.98)
Ed	210.68 **	0.54 **
	(p = 0.00)	(p = 0.00)
Po1	572.99	1.48
	(p = 0.08)	(p = 0.08)
Po2	-305.96	-0.79
	(p = 0.36)	(p = 0.36)
LF	-26.83	-0.07
	(p = 0.65)	(p = 0.65)
M.F	51.29	0.13
	(p = 0.40)	(p = 0.40)
Pop	-27.91	-0.07

	(p = 0.57)	(p = 0.57)
NW	43.23	0.11
	(p = 0.52)	(p = 0.52)
U1	-105.06	-0.27
	(p = 0.18)	(p = 0.18)
U2	141.71	0.37
	(p = 0.05)	(p = 0.05)
Wealth	92.79	0.24
	(p = 0.36)	(p = 0.36)
Ineq	281.95 **	0.73 **
	(p = 0.00)	(p = 0.00)
Prob	-110.39 *	-0.29 *
	(p = 0.04)	(p = 0.04)
Time	-24.66	-0.06
	(p = 0.63)	(p = 0.63)
N	47	47
R2	0.80	0.80

All continuous predictors are mean-centered and scaled by 1 standard deviation. *** p < 0.001; ** p < 0.01; * p < 0.05.

Some key highlights when comparing the models:

1. Their coefficients are different -as expected- because of the scaling.
2. Do note that even though that is the case, their P-values for each feature is exactly the same. This means that the scaling doesn't affect the hypothesis used to calculate this value.
3. Both have the same r-squared and adjusted r-squared.

The R square is very high - above 70% -, which could indicate some overfitting.This is further backed up by many p-values being above 5% (all those features that don't have any asterisks near them). If we were to do some feature selection, we would only remain with M,Ed,Ineq and Prob, as they are the only ones with a p-value below 5%.

For now, as we are still not dealing with overfitting, we will continue with these two models. Let's do some predictions using them:

```
test_data <-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)

predict(no_scale_model,test_data)
```

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

```
predict(scaled_model,test_data)
```

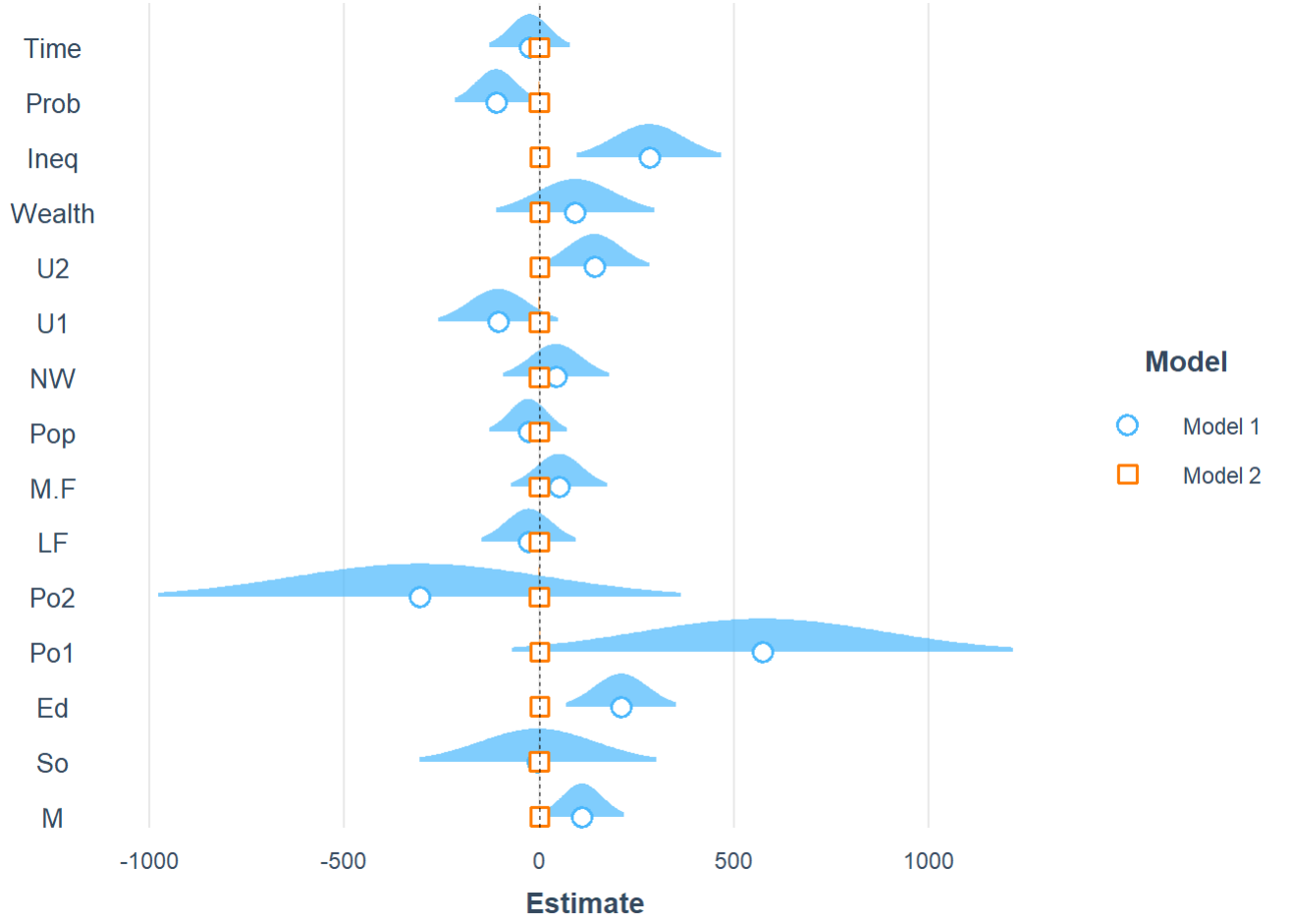
```
##        1
## 797.8641
```

```
range(uscrime$Crime)
```

```
## [1]  342 1993
```

We can see that the unscaled model gives a prediction of 155 for the Crime response, which is way below the minimum contained in the original dataset (342,1993).Now, this doesn't mean this datapoint is not possible, but it does beg some questions. This is specially true when we check out the scaled model and find that the 797 value is well within the range of values contained in the original dataset. This result can be explained by observing the estimation of the coefficient for each feature when done in unscaled data versus scaled data, in this particular case we have:

```
## Loading required namespace: broom.mixed
## Loading required namespace: broom.mixed
```



For example, observe the estimate for Po1 in the Model 1 (unscaled model) which is 572.99 and it's original distribution, it is heavily spread out. Comparing that to the scaled version, there is a big difference. This also applies to the rest of the variables in a bigger or lesser degree, which all compounded is giving us the difference in results.

To further verify if the scaled model is indeed the best one, let's compare them with the Akaike Information Criterion:

```
models <- list(no_scale_model, scaled_model)
model_names <- c("No scale Model", "Scaled Model")
aictab(cand.set = models, modnames = model_names)
```

```
##
## Model selection based on AICc:
##
##                  K   AICc Delta_AICc AICcWt Cum.Wt      LL
## Scaled Model    17 111.10       0.00      1      1  -28.00
## No scale Model 17 671.13     560.03      0      1 -308.01
```

As we can see, the Scaled Model has a lower AICc, which in this case means it has the higher likelihood, as can also be seen in the table. Therefore we would go with the Scaled Model as it is not only giving a more believable prediction, but also has the highest likelihood and therefore has the better quality fit.