HW 04

Group 10

2022-11-02

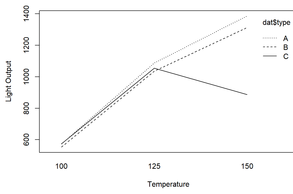
Question 3

Load in data

```
dat = data.frame(type=c(rep("A",9), rep("B",9), rep("C",9)),
temp=rep(c(100,125,150), 9),
y=c(580,1090,1392,568,1087,1380,570,1085,1386,550,1070,1328,530,1035,1312,
579,1000,1299,546,1045,867,575,1053,984,599,1066,809))
```

3(a)

```
interaction.plot(x.factor = dat$temp,
                 trace.factor = dat$type,
                 response = dat$y, xlab = "Temperature", ylab = "Light Output")
```



3(b)

```
fit_3b = lm(y ~ as.factor(temp)*type, data = dat)
summary(fit_3b)
```

```
##
## Call:
## lm(formula = y ~ as.factor(temp) * type, data = dat)
##
## Residuals:
##    Min     1Q  Median      3Q     Max
## -35.000  -5.333  -0.333   6.667  35.000
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 572.6667    11.0381  51.881  < 2e-16 ***
## as.factor(temp)125          514.6667    15.6102  32.970  < 2e-16 ***
## as.factor(temp)150          813.3333    15.6102  52.103  < 2e-16 ***
## typeB                       -19.6667    15.6102  -1.260   0.2238
## typeC                         0.6667    15.6102   0.043   0.9664
## as.factor(temp)125:typeB    -32.6667    22.0762  -1.480   0.1562
## as.factor(temp)150:typeB    -53.3333    22.0762  -2.416   0.0255 *
## as.factor(temp)125:typeC    -33.3333    22.0762  -1.510   0.1484
## as.factor(temp)150:typeC    -500.0000    22.0762 -22.649 1.11e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.12 on 18 degrees of freedom
## Multiple R-squared:  0.9973, Adjusted R-squared:  0.9961
## F-statistic: 824.8 on 8 and 18 DF,  p-value: < 2.2e-16
```

3(c)

```
summary_3c = summary(fit_3b)
summary_3c
```

```
##
## Call:
## lm(formula = y ~ as.factor(temp) * type, data = dat)
##
## Residuals:
##    Min     1Q  Median      3Q     Max
## -35.000  -5.333  -0.333   6.667  35.000
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 572.6667    11.0381  51.881  < 2e-16 ***
## as.factor(temp)125          514.6667    15.6102  32.970  < 2e-16 ***
## as.factor(temp)150          813.3333    15.6102  52.103  < 2e-16 ***
## typeB                       -19.6667    15.6102  -1.260   0.2238
## typeC                         0.6667    15.6102   0.043   0.9664
## as.factor(temp)125:typeB    -32.6667    22.0762  -1.480   0.1562
## as.factor(temp)150:typeB    -53.3333    22.0762  -2.416   0.0255 *
## as.factor(temp)125:typeC    -33.3333    22.0762  -1.510   0.1484
## as.factor(temp)150:typeC    -500.0000    22.0762 -22.649 1.11e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.12 on 18 degrees of freedom
## Multiple R-squared:  0.9973, Adjusted R-squared:  0.9961
## F-statistic: 824.8 on 8 and 18 DF,  p-value: < 2.2e-16
```

$$H_0 : \beta_{temperature*} = \beta_{type*} = \beta_{temperature*type*} = 0$$
$$H_1 : At\ least\ one\ of\ \beta\ s \beta_{temperature*} = \beta_{type*} = \beta_{temperature*type*} \neq 0$$

Since the overall p-value of the model is much less than 0.05, we can reject the null and conclude that at least one of the $\beta's \neq 0$.

3(d)

```
summary(fit_3b)
```

```
##
## Call:
## lm(formula = y ~ as.factor(temp) * type, data = dat)
##
## Residuals:
##    Min     1Q  Median      3Q     Max
## -35.000  -5.333  -0.333   6.667  35.000
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 572.6667    11.0381  51.881  < 2e-16 ***
## as.factor(temp)125          514.6667    15.6102  32.970  < 2e-16 ***
## as.factor(temp)150          813.3333    15.6102  52.103  < 2e-16 ***
## typeB                       -19.6667    15.6102  -1.260   0.2238
## typeC                         0.6667    15.6102   0.043   0.9664
## as.factor(temp)125:typeB    -32.6667    22.0762  -1.480   0.1562
## as.factor(temp)150:typeB    -53.3333    22.0762  -2.416   0.0255 *
## as.factor(temp)125:typeC    -33.3333    22.0762  -1.510   0.1484
## as.factor(temp)150:typeC    -500.0000    22.0762 -22.649 1.11e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.12 on 18 degrees of freedom
## Multiple R-squared:  0.9973, Adjusted R-squared:  0.9961
## F-statistic: 824.8 on 8 and 18 DF,  p-value: < 2.2e-16
```

```
print("-----------------------------------------------------")
```

```
## [1] "-----------------------------------------------------"
```

```
fit_3d = aov(y ~ as.factor(temp)*type, data = dat)
summary(fit_3d)
```

```
##                  Df  Sum Sq Mean Sq F value   Pr(>F)
## as.factor(temp)   2 1978335  989167  2695.3  < 2e-16 ***
## type              2  150865   75432   206.4 3.89e-13 ***
## as.factor(temp):type  4  290552   72638   198.7 1.25e-14 ***
## Residuals        18    6579     366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$H_0 : \beta_{temperature*type*} = 0$$
$$H_1 : \beta_{temperature*type*} \neq 0$$

Since the overall p-value of the model is much less than 0.05, we can reject the null and conclude that $\beta_{temperature*type*} \neq 0$.

3(e)

We can see from 3(d), the temp is greatly effected when typeC is present. When type B is present there is little to no effect on the light output. We can also see from the interaction plot that light output takes a sharp turn down when type C face plate being used. The line continues upwards when type A or B are being used. In conclusion, we can see face plate C is causing something significant to happen and the light output to be lessened at higher temperatures.

3(f)

Since we only have three different temperatures in this data set, it is not appropriate to have it as numerical. It would only be appropriate if we had a more continuous looking temperature column.

Question 4

Read data, add features

```
library(MASS)
library(car)
```

```
## Loading required package: carData
```

```
df = read.csv("bike.csv")
df = subset(df, select = -c(1, 46))

normalize <- function(x) {
return ((x - min(x)) / (max(x) - min(x)))
}
df$cbd_indicator = (normalize(df$Limited_Business_License) +
                      normalize(df$Retail_Food_Establishment) +
                      normalize(df$CTA_BUS_STATIONS) +
                      normalize(df$CTA_TRAIN_STATIONS) +
                      normalize(df$POPULATION_SQ_MILE)) / 5

df$effect_rides = (normalize(df$ARSON) + normalize(df$BURGLARY) + normalize(df$HOMICIDE)) / 3
```

Get model

Explainable model

```
hand_reduced_model = lm(trips ~ PARK_AREA_ACRES + CAPACITY + MINORITY +
                          cbd_indicator + effect_rides, data = df)

summary(hand_reduced_model)
```

```
##
## Call:
## lm(formula = trips ~ PARK_AREA_ACRES + CAPACITY + MINORITY +
##     cbd_indicator + effect_rides, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01927 -0.31648  0.05557  0.36211  1.67899
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     10.7109901  0.2631561  40.702  < 2e-16 ***
## PARK_AREA_ACRES -0.0008734  0.0001793  -4.870 1.76e-06 ***
## CAPACITY         0.0550498  0.0071447   7.705 2.02e-13 ***
## MINORITY        -1.7486482  0.1298993 -13.462  < 2e-16 ***
## cbd_indicator    1.2476848  0.1954596   6.383 6.74e-10 ***
## effect_rides    -1.0473499  0.2112240  -4.958 1.20e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5222 on 294 degrees of freedom
## Multiple R-squared:  0.7088, Adjusted R-squared:  0.7038
## F-statistic: 143.1 on 5 and 294 DF,  p-value: < 2.2e-16
```

```
print("---------------------------------VIF----------------------------------")
```

```
## [1] "---------------------------------VIF----------------------------------"
```

```
vif(hand_reduced_model)
```

```
## PARK_AREA_ACRES       CAPACITY       MINORITY   cbd_indicator    effect_rides
##        1.026444        1.619272        1.213232        1.688692        1.137146
```

Conclusion (Explanation of model above)

There are a few takeaways from the model above:

The first take away is that densely populated areas generally have more diverse modes of transportation than more rural areas. I used to live in a very rural area, the only mode of transportation really was using a car. Moving to Evanston I see many people using other modes of transportation because everything is closer together. This is why I believe the following are significant:

- Area of park in acres (PARK_AREA_ACRES): The more park areas a city has the less urban it is. Thinking about it, places with a lot of parks are generally not urban, and thus would not be practical for biking. For example, Chicago is a very urban city. It doesn't have very much park space other than near the bean. Since everything is closer together, it is a perfect city to rent a bike in.

- Capacity of bikes (CAPACITY): As you increase the number of bikes in a rack, this indicates that the company has data suggesting high demand in that area. Because they have this data, they have placed more bikes in that area. This is why larger capacity indicates more trips.

- Central business indicator (cbd_indicator): A central business district is the business center of a city. Often referred to as downtown. For example, Downtown Chicago. $\beta_{CBD\ Indicator}$ is intuitive. If you are in a central business district, buildings are closer together and thus, more accessible through bikes and other smaller forms of transportation. This is why you see more trips in central business districts.

The next variable that is significant is minority (MINORITY). Research shows that African Americans are more likely to commit crimes than non minority citizens. Since African Americans represent the largest minority group in the country, we can say the minority feature variable represents more African Americans than other races. This model shows that crime and minority are significant and negative. It makes sense that people wouldn't want to rent bikes where crime is happening as it is less safe because they are putting themselves at risk. This explains why the β is negative. This also explains why the β for minority is negative.

The crimes included in my composite crime variable are arson, burglary, and homicide. These generally go together in a home invasion. Thinking about it, most of the crimes can happen anywhere. These crimes are more area specific. Some areas will have more break-ins than other, more secure areas. This break in will include the crimes in the composite variable. Areas where these crimes happen a lot will also be areas where minorities live.

Crime and Minority are related. We aren't sure why this is. This could be an omitted variable. More research is needed to come to a solid conclusion.