# CroqPain_Case

*Team7: Shihan Wang, Zhengyu Jiang, Karina Mark, Yamini Nekkanti*

*10/27/2019*

## *CroqPain Case*

### Problem

Croq'Pain is a French fast food chain aimed at offering a good mix of quality food and price in central business districts. Currently, the customer base is made up of professionals and baby boomers, causing the new stores to move away from the college neighborhoods and escape from the competition with other fast food chains. The management team considered developing a regression model using appropriate parameters to help select new locations for stores.

### Data Summary

The data consists of earnings, population demographics, competitors, and capital invested per operating store. The stores considered have been in operation at least one full year up to and including the first half of 1994. For our analysis, we chose the first 50 stores that opened before 1994 as the training set and 10 stores that opened after 1994 as the validation set. Also, the 10 rows of the data set are new locations shortlisted by the experts for predicting the potential earnings.

### Final Model

Our optimal linear regression model predicting EARN_total on the K, SIZE, INC, P35_total, and NREST_total parameters:

EARN_total = - 22.334 - 0.006 * K + 0.055 * SIZE + 0.644 * INC + 31.541 * P35_total + 1.292 * NREST_total

EARN_total: operating earnings normalized by total population

K: capital invested in the store

SIZE: size of store

INC: average income in town or neighborhood around site

P35_total: Number of 35-44 year-olds normalized by total population in a 3km radius around site

NREST_total: number of non-restaurant businesses normalized by total population in 1km radius around site

The standardized coefficients: SIZE(0.676), NREST_total(0.576), INC(0.524), P35_total(0.377), K(-0.377)

After standardizing coefficients, the standardized coefficient of SIZE is the highest, indicating that the size of the store impacts operating earnings the most. It makes sense because the larger the size of store, the more customers the store can host. In general, the capacity of the store directly determines the profitability of the store. But intuitively, we know largest is not always the best and should consider other parameters as well in predicting the operating revenues.

## Conclusion

Size of the store impacts the earnings more than any other factor. Therefore, Croq'Pain should focus primarily more on the size of the store. In terms of deciding the location for growth expansion, the other factors Croq'Pain should consider are concentration of middle age( 35 - 44 years) working people with higher average income and non-restaurant business in 1km radius. On the other hand capital investment negatively impacts the earnings. So Croq'Pain should try to be more cost effective than spending too much capital to decorate the store. Last, the number of nearby competitors is not a very important factor while deciding the location of stores because its impact of earning is not significant enough.

Based on our analysis, out of the ten shortlisted stores two stores based out of Montpellier and Toulouse should be prioritized for opening up new stores ,as their performance ratio is exceeding the target performance ratio (26%)

## Concerns

However, the final model is still not perfect. There are assumptions and limitations our analysis couldn't confirm. First, the dataset is restricted to 50 stores for the training set and 10 stores for the validation set. The data size is significantly small to train and test. Second, we assume that the total groups of people equal the total population and that data is accurate. We didn't check for accountability. Third, casual effects can occur such as the increase in store size can cause an increase in operation earnings or vice versa. Fourth, we have limited variables to consider for the model because certain variables are highly correlated with others. The total population and groups of people are highly correlated with each other. We want data of high diversity.

## Appendix

Please refer to RMD file, Croq'Pain Case, for final regression model code. Please refer to RMD file, Case Analysis, for model checks and steps that led us to our conclusion.

### Steps to Achieve Our Final Model

1. Check distribution per X variable -> skewed
2. Check scatter plots Y vs X variables -> linear except for Nrest, Price, and CLI
3. Check correlation and VIF -> high correlations and VIF > 10
4. Normalized Earn, P15, P25, P35, P45, P55, Comp, NComp, and Nrest by total
5. Filter first 50 stores -> use stepwise -> pick K, Size, P35_total, Nrest_total
6. Examine correlation table -> keep only one population parameter -> pick P35 because it is the target group
7. Check residuals, normality, VIF, and p-value and predict earnings with the model
8. Follow steps 5 - 7 and regress with P25

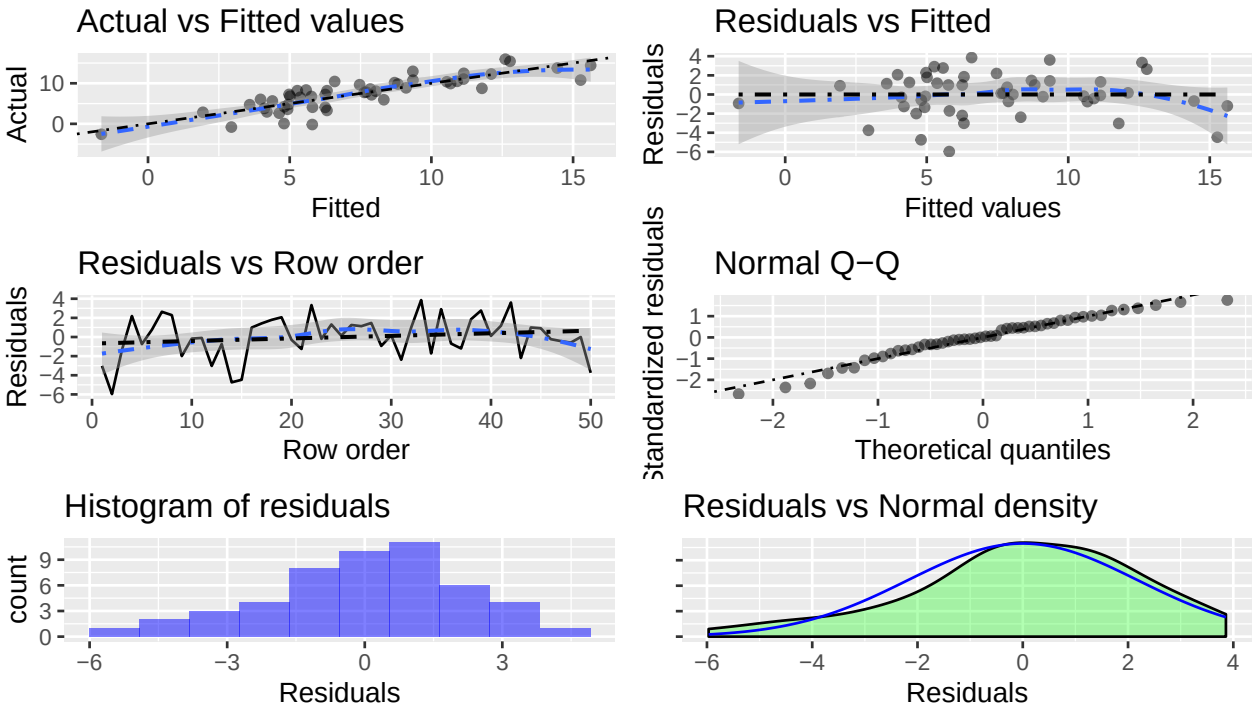
### Regression Model Check

1. Linearity (scatter plots show linearity)
2. Analyze and adjust outliers
3. Multicollinearity (VIF < 10)

4. Check correlation between variables when VIF $> 10$
5. Heteroskedasticty (residual vs fitted values are randomly scattered)
6. Normality of Residuals
7. Autocorrelation (residuals vs row order)

**Final Regression Model**

```
Linear regression (OLS)
Data     : CroqPain
Filter   : STOR<=50
Response variable    : EARN_total
Explanatory variables: K, SIZE, INC, P35_total, NREST_total
Null hyp.: the effect of x on EARN_total is zero
Alt. hyp.: the effect of x on EARN_total is not zero

             coefficient std.error t.value p.value
 (Intercept)     -22.334     3.554  -6.284  < .001 ***
 K                -0.006     0.002  -2.812   0.007 **
 SIZE              0.055     0.011   5.153  < .001 ***
 INC               0.644     0.106   6.102  < .001 ***
 P35_total        31.541     7.890   3.998  < .001 ***
 NREST_total       1.292     0.209   6.177  < .001 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.724,  Adjusted R-squared: 0.692
F-statistic: 23.067 df(5,44), p.value < .001
Nr obs: 50

Variance Inflation Factors
        K  SIZE P35_total NREST_total   INC
VIF 2.870 2.743     1.414       1.386 1.175
Rsq 0.652 0.635     0.293       0.278 0.149
```

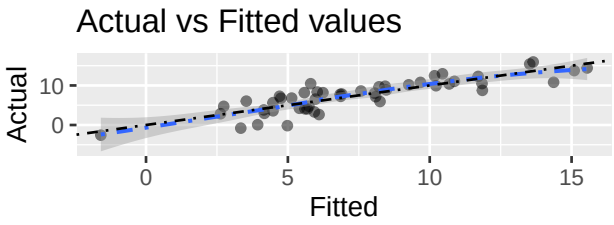


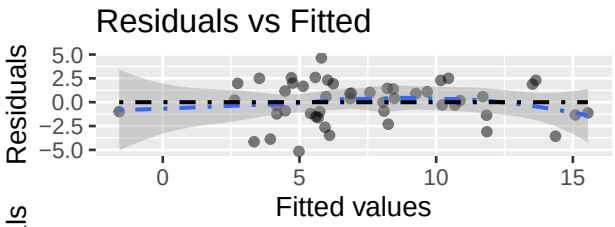

**Standardized Final Regression Model**

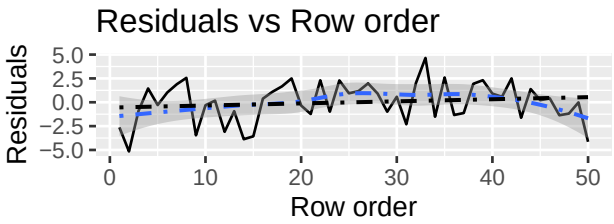
```
Linear regression (OLS)
Data     : CroqPain
Filter   : STOR<=50
Response variable    : EARN_total
Explanatory variables: K, SIZE, INC, P35_total, NREST_total
Null hyp.: the effect of x on EARN_total is zero
Alt. hyp.: the effect of x on EARN_total is not zero
**Standardized coefficients shown (2 X SD)**

             coefficient std.error t.value p.value
 (Intercept)      -0.000     0.039  -0.000   1.000
 K                -0.377     0.134  -2.812   0.007 **
 SIZE              0.676     0.131   5.153  < .001 ***
 INC               0.524     0.086   6.102  < .001 ***
 P35_total         0.377     0.094   3.998  < .001 ***
 NREST_total       0.576     0.093   6.177  < .001 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.724,  Adjusted R-squared: 0.692
F-statistic: 23.067 df(5,44), p.value < .001
Nr obs: 50

Variance Inflation Factors
        K  SIZE P35_total NREST_total   INC
VIF 2.870 2.743     1.414       1.386 1.175
Rsq 0.652 0.635     0.293       0.278 0.149
```

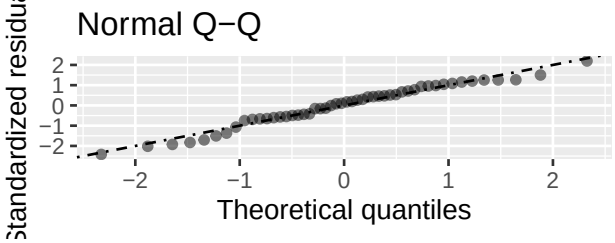
**Alternative Regression Model**

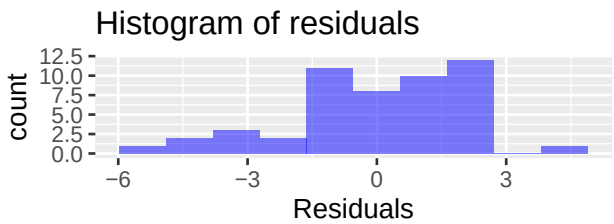
```
Linear regression (OLS)
Data     : CroqPain
Filter   : STOR<=50
Response variable    : EARN_total
Explanatory variables: K, SIZE, INC, P25_total, P35_total, COMP_total, NREST_total
Null hyp.: the effect of x on EARN_total is zero
Alt. hyp.: the effect of x on EARN_total is not zero

             coefficient std.error t.value p.value
 (Intercept)     -25.766     4.226  -6.097  < .001 ***
 K                -0.005     0.002  -2.474   0.017 *
 SIZE              0.053     0.011   4.887  < .001 ***
 INC               0.652     0.104   6.265  < .001 ***
 P25_total        11.226     5.855   1.917   0.062 .
 P35_total        37.288     9.006   4.140  < .001 ***
 COMP_total       -0.373     1.924  -0.194   0.847
 NREST_total       1.478     0.233   6.345  < .001 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-squared: 0.746,  Adjusted R-squared: 0.704
F-statistic: 17.64 df(7,42), p.value < .001
Nr obs: 50

Variance Inflation Factors
        K  SIZE P35_total NREST_total COMP_total P25_total   INC
VIF 3.289 2.976     1.913       1.784      1.674     1.277 1.186
Rsq 0.696 0.664     0.477       0.440      0.403     0.217 0.157
```

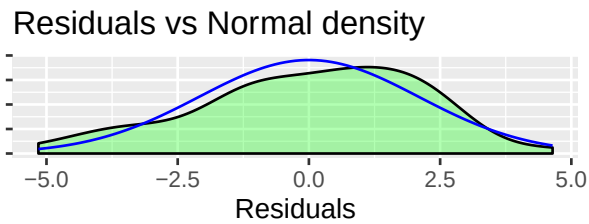


### Compare the performance of two models

```
Evaluate predictions for regression models
Data        : CroqPain
Filter      : STOR<=50
Results for : Both
Predictors  : pred_1, pred_2
Response    : EARN_total

     Type Predictor  n   Rsq  RMSE   MAE
 Training    pred_1 50 0.724 2.160 1.685
 Training    pred_2 50 0.746 2.071 1.695
     Test    pred_1 10 0.527 3.234 2.558
     Test    pred_2 10 0.461 3.523 2.885
```

### Prediction

Show 10 entries

Search:

| STOR | CITY | Performance_Ratio |
|---|---|---|
| All | All | All |
| 69 | Montpellier | 0.3333 |
| 64 | Toulouse | 0.3290 |
| 70 | Dijon | 0.2302 |
| 62 | Montchanin | 0.1296 |
| 66 | Marseilles-1 | 0.1064 |
| 68 | Clermont | 0.0824 |
| 65 | Torcy | 0.0724 |
| 61 | Calais | 0.0711 |
| 63 | Aubusson | 0.0588 |
| 67 | Marseilles-2 | -0.0251 |

Showing 1 to 10 of 10 entries

Previous 1 Next