Croq'Pain Case Study

Asad Adnan, Charley Conroy, Elisabeth Gangwer, Yun-Shiuan Hsu

Executive Summary

Objective

Determine the ideal locations for French restaurant chain Croq' Pain to open new stores in 1996 and accurately predict the store performance ratio using an improved regression model.

Problem

The regression model currently being used violates the principles of multiple regression: multicollinearity and heteroscedasticity. This means that the current regression model has too many insignificant explanatory variables clouding the recommendation and does not provide accurate suggestions to the company.

Findings

The original regression model created by Croq' Pain shows multiple signs of multicollinearity and heteroscedasticity. Our goal is to produce an optimized model that provides accurate results to Croq' Pain executives. Allowing them to make accurate and informed decisions about potential store locations. We began with visualizing multiple variables provided to us. This allows us to have a better understanding of the data that was provided (Figures A - E). Through correlation testing, we found that the variables P15, P35, and P45 have very high correlation coefficients, suggesting that there is multicollinearity within the original model. Furthermore, we created new variables for different population proportions and per capita variables to better compare and understand the model.

We then looked into the original model and discovered that there were too many explanatory variables with a high p-value and no significance (Figure F). This factored into our decision making for an improved regression model. To create an improved regression model, we began with trial and error testing to build up various linear regression models and compared them to decide on the best model for Croq' Pain.

Through our trial and error testing, we discovered that the explanatory variables of the total earnings should be the capital invested (*K*), average income of the area (*INC*), the natural log of the size of the store (*SIZE*), the natural log of the number of non-restaurants competitors (*NREST*), and the natural log of total population (*total*). These variables combined will create a more complete and valid regression model (Figure G). To validate the model we ran various

validation tests to solidify our new linear regression model (Figures H - K). Thus, allowing Croq'Pain to better maximize performance and evaluate the best potential store location for 1996.

By applying our new regression model, we accessed the first 50 stores opened before 1994 to check for multicollinearity and the new coefficients. We solidified its accuracy and applied it to the 10 new stores opened in 1994. The application of the prediction of the new model indicates that 7 out of the 10 new stores should not have been opened. We did this by calculating the performance ratio. The performance ratio is the potential earnings divided by the capital investments into account, which Croq' Pain set at 26%. The three stores that should be opened from 1994 are Store 51(*26.1%*), Store 57 (*31%*) and Store 60 (*37.3%*). These three stores meet and exceed the 26% performance ratio.

When going through the predictions of the new regression model, we realized that the top earning store Store 58 ($300,000) but was not predicted as a top performing store. Notably the predicted performance ratio for store 58(*12.6%*) has a significant difference from the actual performance ratio (*22%*). This difference is due to the extremely high capital investment in the store. The finding shows that we cannot fully rely on the linear regression model for prediction of the success of a business but as a reference. Due to other factors playing a role in whether a business is successful or not, which might be difficult to capture with mere data, and resulting in possible outliers. After our validation of the new regression model, we predicted the ideal store locations that Croq' Pain should open in 1996. Our recommendation is that Toulouse (*42.1%*), Montpellier (*31.9%*), and Dijon (*34.6%*) are the ideal locations for Croq' Pain to open.

Conclusion:

According to our model's prediction, we recommend that the Toulouse, Montpellier, and Dijon stores are the ideal locations where Croq'Pain should open stores. All three locations have low capital investment, big store sizes, with a high number of non-restaurant businesses in the area, and are around a middling income and high total population areas. This connects back to the origins of Croq'Pain as a better quality alternative to fast food popular as a quick lunch among businessmen. Big stores in populated areas around a busy middle class are the spots where Croq'Pain can thrive.

Appendix

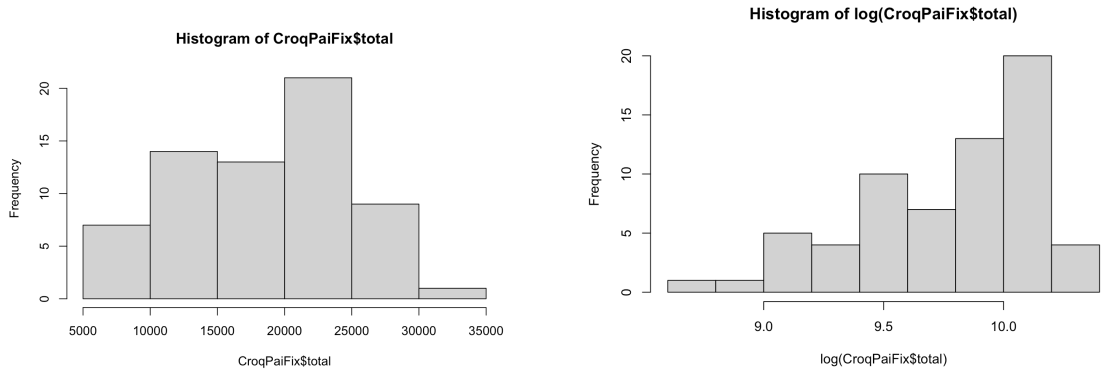

Figure A: Histogram of the total population and the natural log of total in the CroqPaiFix dataset. In our regression model, we used the natural log of total to reduce the large digits of total.

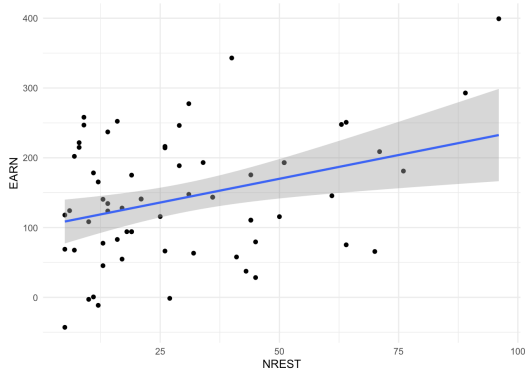

Figure B: Visualization of the linearity between NREST & EARN variables in dataset.

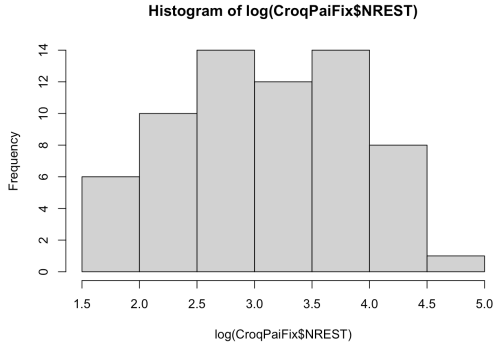

Figure C: Histogram of the natural log of NREST. We used the natural log of the NREST variable to reduce the skewness of the distribution of NREST. This helped us form a more complete and accurate model.

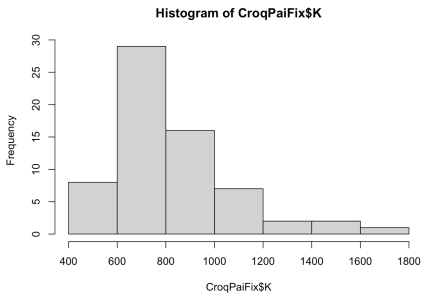

Figure D: Histogram of the capital investment (K) in the CroqPaiFix dataset.

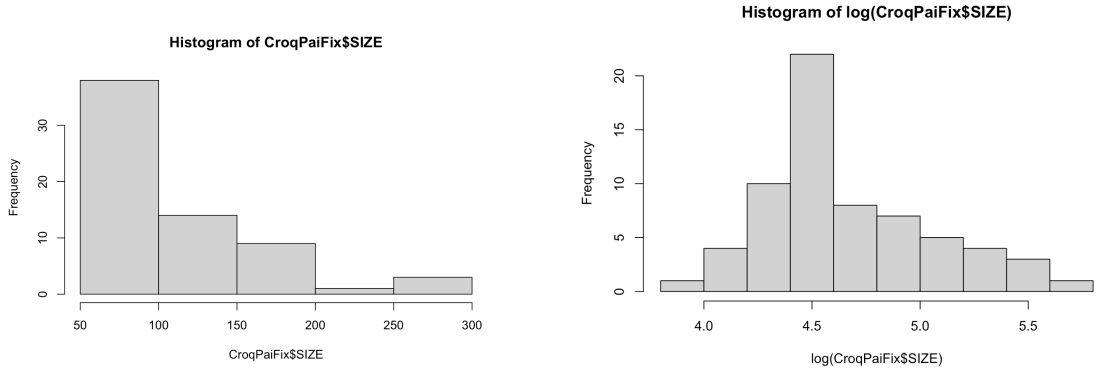

Figure E: Histogram of Size & the Natural Log of Size. We used the natural log of Size in our regression model to reduce the skew of the distribution of SIZE.

```
Call:
lm(formula = EARN ~ SIZE + EMPL + total + P15 + P25 + P35 + P45 +
    P55 + INC + COMP + NCOMP + NREST + PRICE + CLI, data = CroqPaiFix)

Residuals:
    Min      1Q  Median      3Q     Max
-71.042 -20.590   5.543  22.047  68.578

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.724e+02  9.041e+01  -4.119 0.000143 ***
SIZE         7.736e-01  9.443e-02   8.192 8.41e-11 ***
EMPL        -1.138e+00  1.397e+00  -0.814 0.419425
total       -1.083e-02  1.196e-02  -0.906 0.369378
P15          5.756e-02  2.647e-02   2.175 0.034403 *
P25          1.393e-02  1.260e-02   1.105 0.274296
P35          1.423e-02  2.105e-02   0.676 0.502038
P45          2.306e-03  3.019e-02   0.076 0.939402
P55          1.077e-02  1.333e-02   0.808 0.422787
INC          8.811e+00  1.584e+00   5.562 1.04e-06 ***
COMP        -3.090e+00  2.218e+00  -1.393 0.169787
NCOMP       -7.669e-01  1.468e+00  -0.523 0.603580
NREST        1.475e+00  2.313e-01   6.379 5.61e-08 ***
PRICE       -3.110e+00  8.931e-01  -3.483 0.001041 **
CLI          4.974e-01  6.348e-01   0.784 0.436936
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36.93 on 50 degrees of freedom
Multiple R-squared:  0.8631,    Adjusted R-squared:  0.8248
F-statistic: 22.52 on 14 and 50 DF,  p-value: < 2.2e-16
```

Figure F: Summary of the original linear model made by Michel. While looking at the original linear model, there are multiple insignificant explanatory values that could be hindering the results.

```
Call:
lm(formula = EARN ~ K + log(SIZE) + INC + log(NREST) + log(total),
    data = CroqPaiFix)

Residuals:
   Min     1Q Median     3Q    Max
-60.23 -33.42  -2.11  27.00 112.42

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.218e+03  1.669e+02 -13.292  < 2e-16 ***
K           -1.021e-01  3.127e-02  -3.263  0.00183 **
log(SIZE)    1.495e+02  2.032e+01   7.360 6.63e-10 ***
INC          1.144e+01  1.662e+00   6.883 4.27e-09 ***
log(NREST)   4.024e+01  6.638e+00   6.062 1.02e-07 ***
log(total)   1.293e+02  1.377e+01   9.393 2.54e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.93 on 59 degrees of freedom
Multiple R-squared:  0.8016,    Adjusted R-squared:  0.7848
F-statistic: 47.68 on 5 and 59 DF,  p-value: < 2.2e-16
```

Figure G: Summary of the improved regression model. This model has the explanatory variables as the capital investment (*K*), the natural log of the size of the store (*log(STORE)*), the average income of the area (*INC*), the natural log of Non-Restaurant competitors (*log(NREST)*), and the natural log of the population total (*log(total)*). The R-Squared Value is 0.8016 and Adjusted R-Squared is 0.7848.

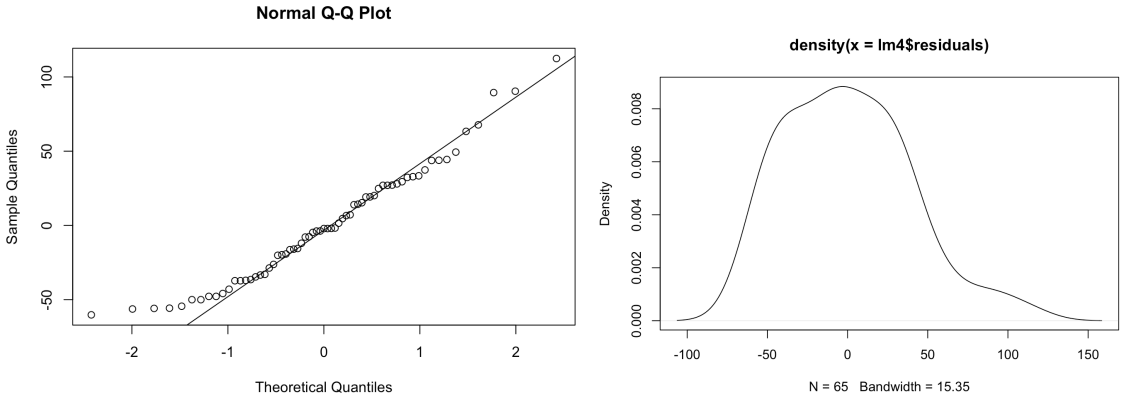

Figure H: Validation test of the improved regression model. The 'Normal Q-Q plot' and the 'density(x = lm4$residuals)' confirms that error follows normal distribution.

```
        K  log(SIZE)        INC log(NREST) log(total)
 2.488861   2.486259   1.060297   1.031791   1.017847
```

Figure I: Validation test of the improved regression model. Testing for multicollinearity within the improved regression model, using the variance inflation factor (VIF). The optimal values being under 5. There is no multicollinearity within the improved regression model.

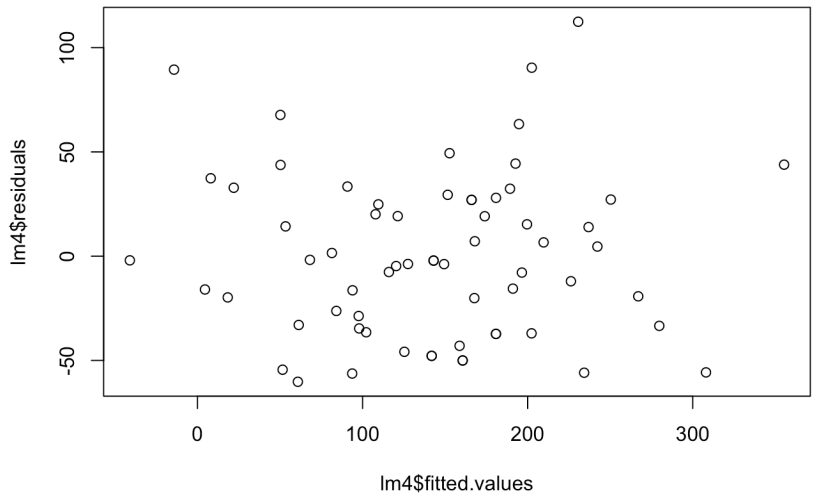

Figure J: Validation test of the improved regression model. Testing the fitted values and residuals of the regression model, confirming homoscedasticity.

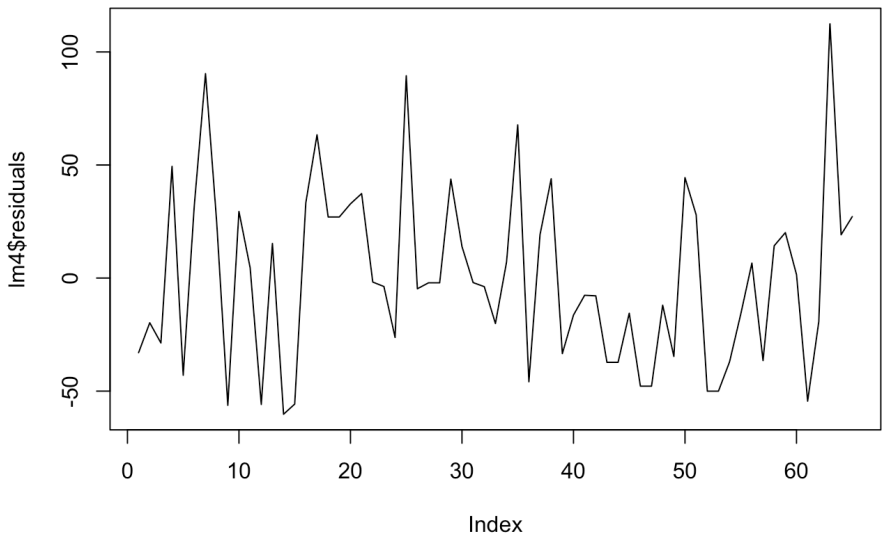

Figure K: Validation test of the improved regression model. Testing for autocorrelation within the regression model. There is no evidence of autocorrelation within the improved regression model.

```
# A tibble: 3 × 31
   STOR  EARN     K  SIZE  EMPL total   P15   P25   P35   P45   P55   INC  COMP
  <int> <dbl> <dbl> <int> <int> <int> <int> <int> <int> <int> <int> <dbl> <int>
1    51  216.  776.   146    11 17440  2800  2350  3180  3050  5490  32.1     2
2    57  248.  782.   119     7 23400  3620  3820  5680  4260  6060  33.4     2
3    60  278.  688.    92    12 25490  4890  1800  6070  5960  5890  36       1
```

Figure L: Using the newly created linear model, Stores 51, 57, and 60 all meet the performance ratio of 26% in 1994.

```
           STOR   K SIZE EMPL total  P15  P25  P35  P45  P55  INC COMP NCOMP
4     Toulouse 836  245 <NA> 11350 3400 3000 2570 1200 1350 37.0    5     8
9  Montpellier 584  149 <NA> 19020 2500 2680 4600 4567 3000 28.6    4     5
10       Dijon 681  150 <NA> 12650 1650 1320 1000 3400 2370 34.9    3    12
   NREST PRICE CLI P15_total P25_total  P35_total P45_total P55_total
4     62  12.5 136 0.2995595 0.2643172 0.22643172 0.1057269 0.1189427
9     26  13.4 128 0.1314406 0.1409043 0.24185068 0.2401157 0.1577287
10    54  15.4 128 0.1304348 0.1043478 0.07905138 0.2687747 0.1873518
     COMP_total  NCOMP_total NREST_total  predict performance
4  0.0004405286 0.0007048458 0.005462555 316.4269   0.3785010
9  0.0002103049 0.0002628812 0.001366982 203.4823   0.3484287
10 0.0002371542 0.0009486166 0.004268775 243.3160   0.3572922
```

Figure K: Using our regression model, Toulouse, Montpellier, and Dijon are the ideal locations for the new Croq' Pain stores in 1996. With predicted performance ratios of 37.9% (Toulouse), 34.8%(Montpellier), and 35.7% (Dijon).