**Where is the Next Croq' Pain?**



Group 4 Lei Duan | Nehal Jain | Qingqing Long | Xiaochen Li

Prof. Terrence August

Nov 07, 2017

## **Part 1. Case Overview**

### **1.1 Background**

Croq'Pain is a chain of French-style fast food restaurants that has wide branches in Paris. According to the store revenue report, seven out of ten new stores opened in last 10 years did not perform well. Thus, an improving system on location selection of new stores need to be constructed to help stakeholders in Croq'Pain identify the right locations for future store openings.

### **1.2 Objectives**

* Identify the most significant explanatory variables that affect operating earnings;
* Develop a regression model on operating earnings based on the most impactful variables from historical data;
* Select the optimal locations for new stores based on prediction.

## **Part 2. Data Processing**

### **2.1 Data Cleaning**

Firstly, we explored data by reading the description and viewing its histogram, scatterplot and correlation plots to check if abnormality exists. Two problems were found as shown below:

* There are five duplicate values in dataset;
* There is an extreme value in earning in store 1.

After cross checking all data, we believe that the outlier (see in *Exhibit 1*) is caused by a mistake of data entry. Since the unit of earning is $1000 and all other store earnings are ranged from -40 to 399 thousand dollars, it's highly possible that the original earning of store 1 was inputted in a raw format instead of being divided by 1000.

Thus, we removed all duplicates and divided the earning value in store 1 by 1000 for further processing.

### **2.2 Data Exploration**

#### 2.2.1 Histogram plot

After cleansing, we plotted the most variables' distribution, all of which look acceptable.

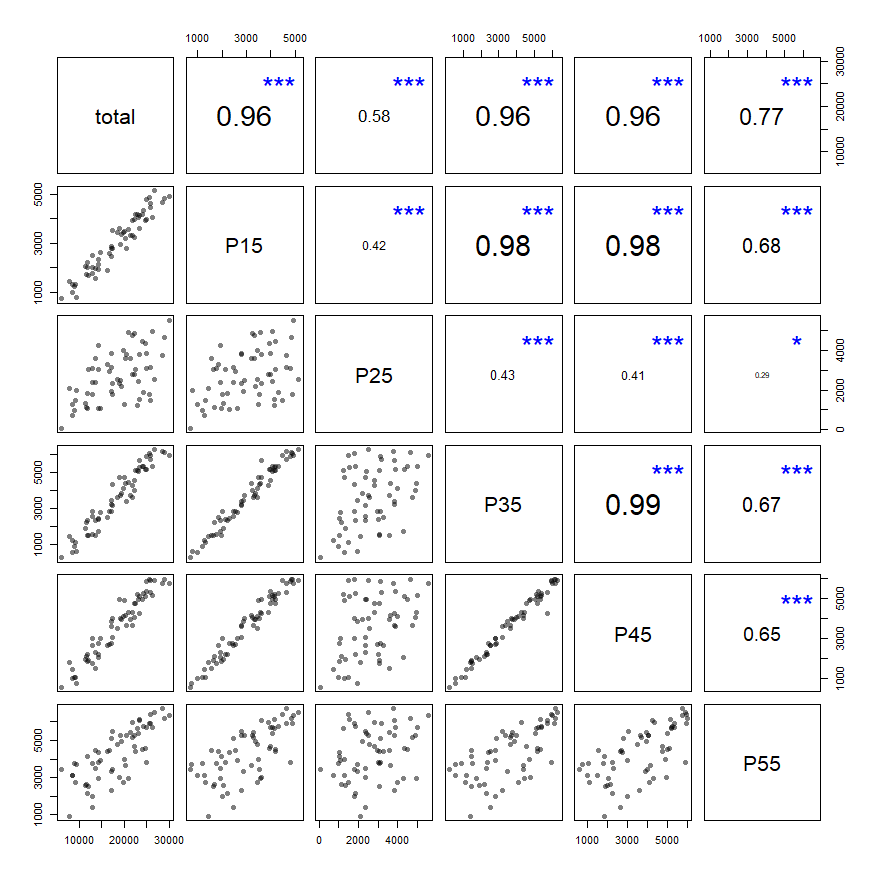
(See in *Exhibit 2*)

#### 2.2.2 Scatterplots and correlation plots on total and P15 - P55 variables

By drawing scatterplots and correlation plots, high correlation is found among the population in each age group with ‘total’ population within 3 km of the restaurant, which indicates the high potential of multi-collinearity. (see *Table* 1 and *Figure 1* and more details in *Exhibit 3*)

Correlation  
Data : CroqPainFix   
Method : pearson   
Variables: total, P15, P25, P35, P45, P55   
Null hyp.: variables x and y are not correlated  
Alt. hyp.: variables x and y are correlated  
  
Correlation matrix:  
 total P15 P25 P35 P45   
P15 0.96   
P25 0.58 0.42   
P35 0.96 0.98 0.43   
P45 0.96 0.98 0.41 0.99   
P55 0.77 0.68 0.29 0.67 0.65  
  
p.values:  
 total P15 P25 P35 P45   
P15 0.00   
P25 0.00 0.00   
P35 0.00 0.00 0.00   
P45 0.00 0.00 0.00 0.00   
P55 0.00 0.00 0.03 0.00 0.00

*Table 1*



*Figure 1*

## **Part 3. Model Building**

Firstly, we normalized variables ‘EARN’, ‘COMP’, ‘NCOMP’, ‘NREST’ and all age group variables (i.e. P15-P55) by total amount of residents. By comparing the dashboard of regression using un-normalized and normalized variables, we prefer to normalize data for further analysis based on more accountable variables (i.e., earnings per person and population proportion in each age group) to eliminate the impact of total population on other variables (see *Exhibit 4-5*).

Moreover, we used “stepwise” approach to select variables in regression model based on significance (see *Exhibit 6-7*) and made further adjustment by removing P15 and P25 (i.e. age groups from 15 to 34) and adding back P35 (i.e. age group from 35-44), as people aged from 35 to 44 are the target customers of Croq'Pain. (see following *Table* 2)

Thus, our final model contains the following variables as predictors:

* K: capital investment;
* INC: regional income level;
* Size: the size of the restaurant;
* P\_35\_total: the proportion of residents aged from 35-44 in the region;
* NREST\_total: the amount of non-restaurant business per person in the region;

Linear regression (OLS)  
Data : CroqPainFix   
Response variable : EARN\_total   
Explanatory variables: K, INC, SIZE, 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 0.037 0.000 1.000   
 K -0.449 0.130 -3.450 0.001 \*\*   
 INC 0.440 0.077 5.715 < .001 \*\*\*  
 SIZE 0.861 0.127 6.761 < .001 \*\*\*  
 P35\_total 0.307 0.083 3.707 < .001 \*\*\*  
 NREST\_total 0.468 0.082 5.712 < .001 \*\*\*  
  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
R-squared: 0.706, Adjusted R-squared: 0.678   
F-statistic: 25.886 df(5,54), p.value < .001  
Nr obs: 60   
  
Variance Inflation Factors  
 K SIZE P35\_total NREST\_total INC  
VIF 3.108 2.972 1.260 1.231 1.087  
Rsq 0.678 0.664 0.206 0.188 0.080

*Table 2*

## **Part 4. Model Validation**

To validate the model, we split the dataset into two parts: ‘50 historical restaurants before 1994’ and ‘10 restaurants after 1994’. Then we applied data of first 50 restaurants to amend regression model (result seen in *Table* 3) and to predict which restaurant opened after 1994 would reach the performance ratio target of Croq'Pain (i.e., 26%).

Linear regression (OLS)  
Data : CroqPainFix\_<= 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

*Table 3*

It's shown that only store 57 and store 60 (see in *Table* 4) could reach the 26% performance ratio according to our prediction model.

STOR pred\_Ratio  
1 57 0.3670888  
2 60 0.4060865

*Table 4*

In fact, three restaurants, which are store 51, store 57, store 60 (see in *Table* 5), achieved the goal and store 51 reached approximately 27% which is slightly over 26%, so we believe that our model accuracy is acceptable.

STOR real\_Ratio  
1 51 0.2787193  
2 57 0.3168194  
3 60 0.4033956

*Table 5*

## **Part 5. Model Prediction**

Based on the model, we predicted the potential earning of 10 restaurants opened after 1994 and calculated their performance ratios accordingly. It turned out that only "Toulouse" and "Montpellier" (see in *Table* 6) exceeded 26% level of performance ratio.

STOR pred\_Ratio  
1 Toulouse 0.3370087  
2 Montpellier 0.3620923

*Table 6*

## **Part 6. Conclusion and Advice in Location Choice**

According to the model and prediction result, we highly recommend that Craq'pain should take the following factors into consideration when selecting location for new stores: **capital investment, regional income level, restaurant size, the proportion of residents aging from 35-44 in the region and the number of non-restaurant business per person in the region**.

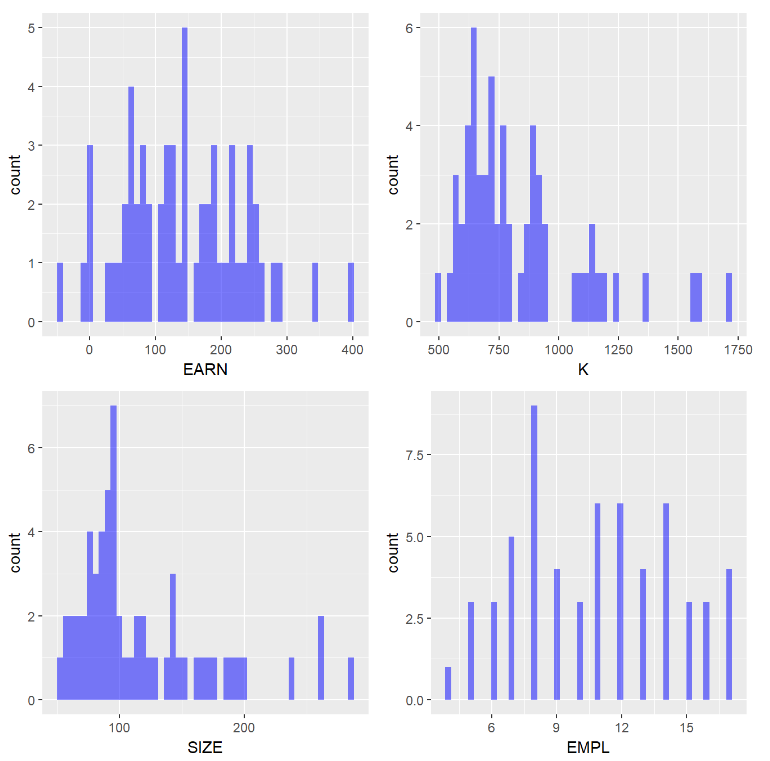
To reach performance ratio goal and achieve future success for new restaurants, we recommend choosing **‘Toulouse’** and **‘Montpellier’** to invest.

**Part 7. Appendix**

**Exhibit 1. Distribution plot of earning.**

#### 

#### Exhibit 2. Histogram of variables



#### C:\Users\Pauline\AppData\Local\Temp\1510090240(1).png

#### C:\Users\Pauline\AppData\Local\Temp\1510090278(1).png

#### C:\Users\Pauline\AppData\Local\Temp\1510090317(1).png

#### C:\Users\Pauline\AppData\Local\Temp\1510090348(1).png

#### Exhibit 3. Scatterplots of age group with total population

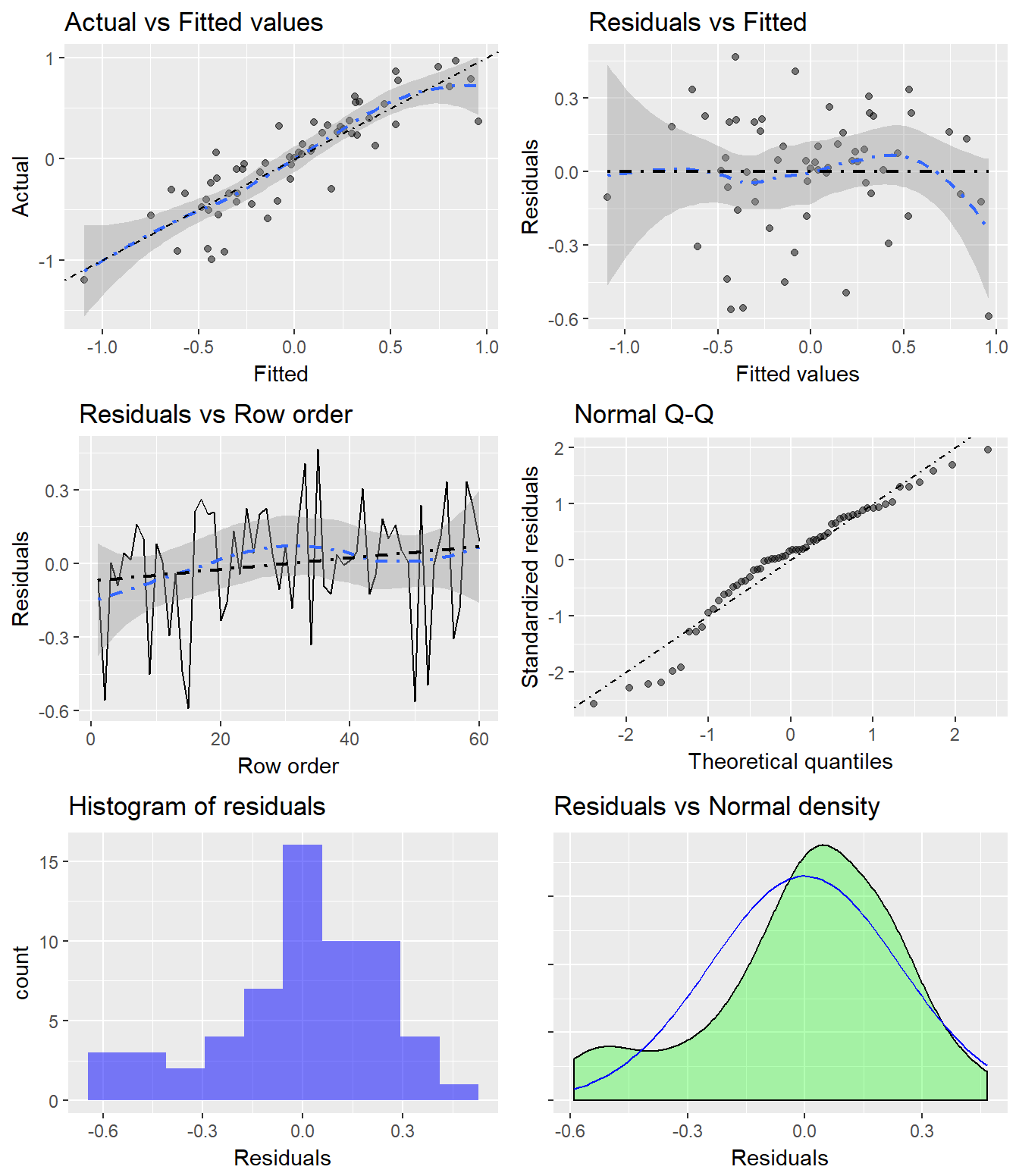
#### Exhibit 4. Regression with untransformed variables

Linear regression (OLS)  
Data : CroqPainFix   
Response variable : EARN   
Explanatory variables: K, SIZE, EMPL, total, P15, P25, P35, P45, P55, INC, COMP, NCOMP, NREST, PRICE, CLI   
Null hyp.: the effect of x on EARN is zero  
Alt. hyp.: the effect of x on EARN is not zero  
\*\*Standardized coefficients shown (2 X SD)\*\*  
  
 coefficient std.error t.value p.value   
 (Intercept) 0.000 0.026 0.000 1.000   
 K -0.516 0.319 -1.619 0.113   
 SIZE 0.851 0.251 3.389 0.001 \*\*   
 EMPL -0.066 0.060 -1.100 0.277   
 total -0.346 0.854 -0.405 0.687   
 P15 0.761 0.356 2.141 0.038 \*   
 P25 0.092 0.191 0.480 0.633   
 P35 0.109 0.417 0.260 0.796   
 P45 -0.089 0.556 -0.160 0.873   
 P55 0.079 0.216 0.365 0.717   
 INC 0.352 0.063 5.636 < .001 \*\*\*  
 COMP -0.059 0.061 -0.969 0.338   
 NCOMP -0.007 0.058 -0.113 0.910   
 NREST 0.365 0.060 6.116 < .001 \*\*\*  
 PRICE 0.084 0.190 0.445 0.659   
 CLI 0.040 0.063 0.642 0.524   
  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
R-squared: 0.875, Adjusted R-squared: 0.832   
F-statistic: 20.453 df(15,44), p.value < .001  
Nr obs: 60   
  
Variance Inflation Factors  
 VIF Rsq  
total 255.640 0.996  
P45 108.367 0.991  
P35 61.141 0.984  
P15 44.350 0.977  
K 35.614 0.972  
SIZE 22.127 0.955  
P55 16.407 0.939  
P25 12.739 0.922  
PRICE 12.656 0.921  
INC 1.372 0.271  
CLI 1.371 0.271  
COMP 1.291 0.225  
NREST 1.252 0.201  
EMPL 1.245 0.197  
NCOMP 1.177 0.150

#### 

#### Exhibit 5. Regression with normalized variables

Linear regression (OLS)  
Data : CroqPainFix   
Response variable : EARN\_total   
Explanatory variables: K, SIZE, EMPL, INC, PRICE, CLI, P15\_total, P25\_total, P35\_total, P45\_total, P55\_total, COMP\_total, NCOMP\_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 0.036 0.000 1.000   
 K -0.765 0.428 -1.788 0.081 .   
 SIZE 1.134 0.334 3.396 0.001 \*\*   
 EMPL -0.100 0.081 -1.223 0.228   
 INC 0.409 0.085 4.797 < .001 \*\*\*  
 PRICE 0.187 0.249 0.750 0.457   
 CLI 0.096 0.084 1.146 0.258   
 P15\_total 0.308 0.154 1.999 0.052 .   
 P25\_total 0.081 0.169 0.482 0.632   
 P35\_total 0.236 0.243 0.973 0.336   
 P45\_total -0.229 0.269 -0.853 0.398   
 P55\_total -0.038 0.225 -0.167 0.868   
 COMP\_total -0.189 0.107 -1.762 0.085 .   
 NCOMP\_total 0.135 0.092 1.474 0.147   
 NREST\_total 0.583 0.101 5.760 < .001 \*\*\*  
  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
R-squared: 0.766, Adjusted R-squared: 0.693   
F-statistic: 10.506 df(14,45), p.value < .001  
Nr obs: 60   
  
Variance Inflation Factors  
 VIF Rsq  
K 35.143 0.972  
SIZE 21.438 0.953  
P45\_total 13.888 0.928  
PRICE 11.895 0.916  
P35\_total 11.315 0.912  
P55\_total 9.701 0.897  
P25\_total 5.486 0.818  
P15\_total 4.563 0.781  
COMP\_total 2.217 0.549  
NREST\_total 1.967 0.492  
NCOMP\_total 1.609 0.379  
INC 1.395 0.283  
CLI 1.358 0.264  
EMPL 1.272 0.214



#### Exhibit 6. Regression model we initially built

Linear regression (OLS)  
Data : CroqPainFix   
Response variable : EARN\_total   
Explanatory variables: K, INC, PRICE, P15\_total, P25\_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  
\*\*Standardized coefficients shown (2 X SD)\*\*  
  
 coefficient std.error t.value p.value   
 (Intercept) 0.000 0.039 0.000 1.000   
 K 0.655 0.107 6.109 < .001 \*\*\*  
 INC 0.322 0.085 3.789 < .001 \*\*\*  
 PRICE -0.589 0.101 -5.822 < .001 \*\*\*  
 P15\_total 0.384 0.099 3.890 < .001 \*\*\*  
 P25\_total 0.203 0.093 2.193 0.033 \*   
 COMP\_total -0.200 0.096 -2.084 0.042 \*   
 NREST\_total 0.554 0.100 5.534 < .001 \*\*\*  
  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
R-squared: 0.679, Adjusted R-squared: 0.636   
F-statistic: 15.714 df(7,52), p.value < .001  
Nr obs: 60   
  
Variance Inflation Factors  
 K PRICE NREST\_total P15\_total COMP\_total P25\_total INC  
VIF 1.864 1.655 1.623 1.583 1.499 1.388 1.172  
Rsq 0.463 0.396 0.384 0.368 0.333 0.280 0.147

#### Exhibit 7. Regression with transformed variables with stepwise function

Start: AIC=-141.26  
EARN\_total ~ K + SIZE + EMPL + INC + PRICE + CLI + P15\_total +   
 P25\_total + P35\_total + P45\_total + P55\_total + COMP\_total +   
 NCOMP\_total + NREST\_total  
  
 Df Sum of Sq RSS AIC  
- P55\_total 1 0.00215 3.4576 -143.23  
- P25\_total 1 0.01785 3.4733 -142.96  
- PRICE 1 0.04318 3.4986 -142.52  
- P45\_total 1 0.05582 3.5112 -142.30  
- P35\_total 1 0.07272 3.5281 -142.01  
- CLI 1 0.10077 3.5562 -141.54  
- EMPL 1 0.11485 3.5703 -141.30  
<none> 3.4554 -141.26  
- NCOMP\_total 1 0.16694 3.6224 -140.43  
- COMP\_total 1 0.23828 3.6937 -139.26  
- K 1 0.24541 3.7008 -139.15  
- P15\_total 1 0.30691 3.7623 -138.16  
- SIZE 1 0.88547 4.3409 -129.58  
- INC 1 1.76719 5.2226 -118.48  
- NREST\_total 1 2.54757 6.0030 -110.12  
  
Step: AIC=-143.23  
EARN\_total ~ K + SIZE + EMPL + INC + PRICE + CLI + P15\_total +   
 P25\_total + P35\_total + P45\_total + COMP\_total + NCOMP\_total +   
 NREST\_total  
  
 Df Sum of Sq RSS AIC  
- PRICE 1 0.04106 3.4986 -144.52  
- P45\_total 1 0.06320 3.5208 -144.14  
- P35\_total 1 0.07898 3.5366 -143.87  
- CLI 1 0.09885 3.5564 -143.53  
- P25\_total 1 0.10248 3.5601 -143.47  
- EMPL 1 0.11279 3.5704 -143.30  
<none> 3.4576 -143.23  
- NCOMP\_total 1 0.16482 3.6224 -142.43  
- COMP\_total 1 0.23877 3.6964 -141.22  
- K 1 0.24564 3.7032 -141.11  
- P15\_total 1 0.33677 3.7943 -139.65  
- SIZE 1 0.90085 4.3584 -131.33  
- INC 1 1.76743 5.2250 -120.45  
- NREST\_total 1 2.56211 6.0197 -111.96  
  
Step: AIC=-144.52  
EARN\_total ~ K + SIZE + EMPL + INC + CLI + P15\_total + P25\_total +   
 P35\_total + P45\_total + COMP\_total + NCOMP\_total + NREST\_total  
  
 Df Sum of Sq RSS AIC  
- P45\_total 1 0.0703 3.5690 -145.32  
- EMPL 1 0.0811 3.5797 -145.14  
- P35\_total 1 0.0989 3.5975 -144.85  
- CLI 1 0.1097 3.6083 -144.67  
<none> 3.4986 -144.52  
- P25\_total 1 0.1395 3.6382 -144.17  
- NCOMP\_total 1 0.1580 3.6567 -143.87  
- COMP\_total 1 0.2719 3.7705 -142.03  
- P15\_total 1 0.3306 3.8292 -141.10  
- K 1 0.8302 4.3289 -133.74  
- INC 1 1.8326 5.3313 -121.25  
- NREST\_total 1 2.6964 6.1951 -112.24  
- SIZE 1 3.5636 7.0623 -104.38  
  
Step: AIC=-145.32  
EARN\_total ~ K + SIZE + EMPL + INC + CLI + P15\_total + P25\_total +   
 P35\_total + COMP\_total + NCOMP\_total + NREST\_total  
  
 Df Sum of Sq RSS AIC  
- P35\_total 1 0.0294 3.5984 -146.83  
- CLI 1 0.0777 3.6467 -146.03  
- EMPL 1 0.0871 3.6561 -145.88  
- NCOMP\_total 1 0.1200 3.6890 -145.34  
<none> 3.5690 -145.32  
- P25\_total 1 0.1530 3.7220 -144.81  
- P15\_total 1 0.2702 3.8392 -142.94  
- COMP\_total 1 0.2905 3.8595 -142.63  
- K 1 0.8365 4.4055 -134.69  
- INC 1 1.8656 5.4346 -122.09  
- NREST\_total 1 2.6302 6.1992 -114.20  
- SIZE 1 3.5138 7.0828 -106.20  
  
Step: AIC=-146.83  
EARN\_total ~ K + SIZE + EMPL + INC + CLI + P15\_total + P25\_total +   
 COMP\_total + NCOMP\_total + NREST\_total  
  
 Df Sum of Sq RSS AIC  
- CLI 1 0.0917 3.6901 -147.32  
- EMPL 1 0.1014 3.6998 -147.16  
- NCOMP\_total 1 0.1025 3.7009 -147.15  
<none> 3.5984 -146.83  
- P25\_total 1 0.1704 3.7688 -146.06  
- COMP\_total 1 0.3838 3.9822 -142.75  
- K 1 0.8998 4.4982 -135.44  
- P15\_total 1 1.0401 4.6386 -133.60  
- INC 1 1.8470 5.4454 -123.97  
- NREST\_total 1 2.6020 6.2004 -116.18  
- SIZE 1 3.7359 7.3344 -106.11  
  
Step: AIC=-147.32  
EARN\_total ~ K + SIZE + EMPL + INC + P15\_total + P25\_total +   
 COMP\_total + NCOMP\_total + NREST\_total  
  
 Df Sum of Sq RSS AIC  
- EMPL 1 0.0865 3.7765 -147.93  
- NCOMP\_total 1 0.1003 3.7904 -147.71  
<none> 3.6901 -147.32  
- P25\_total 1 0.1572 3.8472 -146.82  
- COMP\_total 1 0.3296 4.0197 -144.19  
- K 1 0.8096 4.4996 -137.42  
- P15\_total 1 1.1243 4.8144 -133.37  
- INC 1 1.8656 5.5557 -124.77  
- NREST\_total 1 2.5430 6.2331 -117.87  
- SIZE 1 3.6599 7.3500 -107.98  
  
Step: AIC=-147.93  
EARN\_total ~ K + SIZE + INC + P15\_total + P25\_total + COMP\_total +   
 NCOMP\_total + NREST\_total  
  
 Df Sum of Sq RSS AIC  
- NCOMP\_total 1 0.0903 3.8669 -148.51  
<none> 3.7765 -147.93  
- P25\_total 1 0.1635 3.9400 -147.39  
- COMP\_total 1 0.4331 4.2097 -143.42  
- K 1 0.8358 4.6124 -137.94  
- P15\_total 1 1.1119 4.8884 -134.45  
- INC 1 1.8112 5.5877 -126.43  
- NREST\_total 1 2.7495 6.5261 -117.11  
- SIZE 1 3.6805 7.4570 -109.11  
  
Step: AIC=-148.51  
EARN\_total ~ K + SIZE + INC + P15\_total + P25\_total + COMP\_total +   
 NREST\_total  
  
 Df Sum of Sq RSS AIC  
<none> 3.8669 -148.51  
- P25\_total 1 0.2518 4.1187 -146.73  
- COMP\_total 1 0.3453 4.2122 -145.38  
- K 1 0.8945 4.7614 -138.03  
- P15\_total 1 1.1331 5.0000 -135.09  
- INC 1 2.0322 5.8990 -125.17  
- NREST\_total 1 2.7938 6.6607 -117.89  
- SIZE 1 3.9542 7.8211 -108.25

----------------------------------------------------  
Backward stepwise selection of variables  
----------------------------------------------------  
Linear regression (OLS)  
Data : CroqPainFix   
Response variable : EARN\_total   
Explanatory variables: K, SIZE, EMPL, INC, PRICE, CLI, P15\_total, P25\_total, P35\_total, P45\_total, P55\_total, COMP\_total, NCOMP\_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 0.035 0.000 1.000   
 K -0.453 0.131 -3.468 0.001 \*\*   
 SIZE 0.903 0.124 7.292 < .001 \*\*\*  
 INC 0.396 0.076 5.228 < .001 \*\*\*  
 P15\_total 0.349 0.089 3.904 < .001 \*\*\*  
 P25\_total 0.154 0.084 1.840 0.071 .   
 COMP\_total -0.186 0.086 -2.155 0.036 \*   
 NREST\_total 0.554 0.090 6.129 < .001 \*\*\*  
  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
R-squared: 0.738, Adjusted R-squared: 0.703   
F-statistic: 20.907 df(7,52), p.value < .001  
Nr obs: 60   
  
Variance Inflation Factors  
 K SIZE NREST\_total P15\_total COMP\_total P25\_total INC  
VIF 3.381 3.039 1.621 1.587 1.476 1.385 1.140  
Rsq 0.704 0.671 0.383 0.370 0.323 0.278 0.123