Regression Analysis Project

Joshua Marple

November 2014

Introduction

The following project attempts to find the most predictive elements of the Index of Consumer Satisfaction (ICS) through the methods learned in MATH 605. This data set was a survey of consumers and can be found at http://www.sca.isr.umich.edu/subset/subset.php.

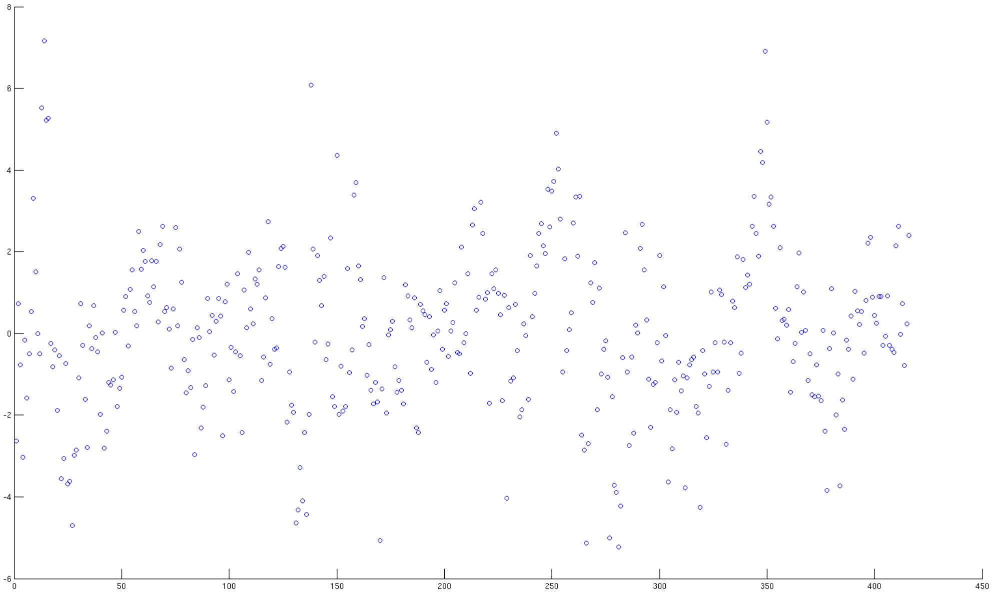
A number of factors were selected for analysis.

- PAGO - Current Financial Situation Compared with a Year Ago
- PEXP - Expected Change in Financial Situation in a Year
- PINC2 - Probability of Personal Income Increase During the Next Year
- PINC - Probability of Real Income Gains During the Next 5 Years
- PJOB - Probability of Losing a Job During the Next 5 Years
- PSSA - Probability of Adequate Retirement Income
- PSTK - Probability of Increase in Stock Market in Next Year
- NEWS - News Heard of Recent Changes in Business Conditions
- RATEX - Expected Change in Interest Rates During the Next Year
- PX1 - Expected Change in Prices During the Next Year
- GOVT - Opinions About the Government's Economic Policy
- DUR - Buying Conditions for Large Household Durables
- VEH - Buying Conditions for Vehicles
- VEHRN - Reasons for Opinions for Buying Conditions for Vehicles

Model Explanation

In this model, ICS is our response variable and the factors listed in the Introduction are our regressor variables. The model parameters are estimated according to the code attached in the appendix, and were estimated with the tools learned in this course. The errors of the model were then analyzed. Through inspection, we have a rough verification of the model. The error follows no consistent pattern, and appears to be mostly normally scattered. There are a few outliers, but for the most part the model error is consistent with what one might expect in a well fitted model. Additionally, the model error was $-5.5962e-09$, which while it does prove that a linear model is best, does not raise any flags.

Figure 1: Scatter Plot of the Errors



ANOVA Testing

The code for the ANOVA test is contained within the appendix.

Our ANOVA table is as follows:

	DF	SS	MS	F
Model	12	68769.111634	5730.759303	1439.002475
Error	416	1604.928442	3.982453	
Total	428	70374.039904		

We then use this F statistic to test the hypothesis that all of our regressors are equal to zero. This provides us with the P value of 3.2619e-20, so we can very safely reject the null hypothesis.

Regressor Testing

Now we pick one of the regressors, and test that it is not equal to zero. Again, the code that does this is contained within the appendix, so we will summarize the results here. The regressor picked was PAGO, or Financial Situation compared to a year ago. As it turns out, this regressor is an extremely important variable, with a P value of 2.8801e-19.

Prediction Interval

Now we create a fictitious data point and see how our model would predict in (within 95% accuracy). First, we create the data point by examining the mean of the data set and the standard deviation. The data point ([112; 117; 78; 55; 3.9; 40; 5; 100; 127; 130; 8], ordered as [pago_r_all, pexp_r_all, news_r_all, ratex_r_all, px1_mean_all, px1_var_all, px1_std_all, govt_r_all, dur_r_all, veh_r_all, vehrn_np_all]) is within one standard deviation of the means.

Now we evaluate our data point. We do so with an inverse t test, at a 95% probability. This allows us to create our prediction interval, which comes out to 82.936475 to 84.485553.

Interpretation/ Conclusions

It is now plain to see that at least some of our regressors predict the Index of Consumer Satisfaction adequately. Some regressors, such as PAGO, explain much of the variance while others may do less. Through our empirical analysis of the errors, we can clearly see that a linear model fits the data appropriately, and through our ANOVA table, we have a good picture of our variance in the data set.

Appendix: Code

All code was written for MATLAB 2014, and does not work with any previous versions. import_data_GEN is also not shown, as it is a file automatically generated by MATLAB.

```
import_data_GEN; %takes care of formatting, an autogenned file

%% Simple Calculations

y = ics_all;
x = [ones(length(y), 1), pago_r_all, pexp_r_all, news_r_all, ratex_r_all, px1_mean_all ...
    px1_var_all, px1_std_all, govt_r_all, dur_r_all, veh_r_all, vehrn_np_all];

n = length(y);
p = size(x);
p = p(2);

b = inv( transp(x) * x ) * transp(x) * y; %#ok<MINV>

u = x * b;

e = y - u;
scatter(1:length(e), e)
fprintf('STD DEV of e: %f \n', std(e));
fprintf('MEAN of e: %f \n', mean(e));

s2 = transp(e) * e / (n - p - 1);

var = inv(transp(x) * x);

%% ANOVA
%creating the anova table

SSR = transp(b) * transp(x) * y - n * mean(y)^2;
SSE = transp(y - x * b) * (y - x * b);
SST = sum((y - mean(y)).^2);
fprintf('SSE: %f\n', SSE);
fprintf('SSR: %f\n', SSR);
if (abs(SST - (SSR + SSE)) > 1)
    fprintf('SUMS DO NOT MATCH, ERROR\n');
    fprintf('%f != %f\n', SST, SSR + SSE);
else
    fprintf('SST = SSR + SSE, proceeding\n');
    fprintf('%f = %f\n', SST, SSR + SSE);
```

```
end

MSR = SSR/p;
MSE = SSE/(n-p-1);
fprintf(’MSR: %f, MSE: %f\n’, MSR, MSE);

F = MSR/MSE;
fprintf(’F: %f\n’, F);
pval = fpdf( F, n-p-1, p);
fprintf(’P value: %f\n’, pval);
if (pval > 0.05)
    fprintf(’Fail to reject null hypothesis, as our p value is so large\n’);
else
    fprintf(’Reject the null hypothesis, as our p value is so small\n’);
end

%% Testing Important Regressors
% testing if financial situation compared to a year ago is an important regressor

T = b(2) / sqrt(s2 * var(2,2));

pval = 2*tcdf( T, n - p -1, ’upper’);


fprintf(’P-value when testing financial situation compared to a year ago %f\n’, pval);


%% Prediction
% make a prediction given a relatively reasonable data set

test_val = [1; 112; 117; 78; 55; 3.9; 40; 5; 100; 127; 130; 8];
t_val = tinv(.95, n-p-1)* sqrt(s2) * sqrt(transp(test_val) * inv(1+transp(x) * x) * test_val);

pred_int = [transp(test_val) * b - t_val, transp(test_val) * b + t_val];
fprintf(’Prediction interval: %f - %f\n’, pred_int(1), pred_int(2));
```

The raw output of this script is as follows:

```
STD DEV of e: 1.966545
MEAN of e: -0.000000
SSE: 1604.928442
SSR: 68769.111634
SST = SSR + SSE, proceeding
70374.039904 = 70374.040075
MSR: 5730.759303, MSE: 3.982453
```

```
F: 1439.002475
P value: 0.000000
Reject the null hypothesis, as our p value is so small
P-value when testing financial situation compared to a year ago 0.000000
Prediction interval: 82.936475 - 84.485553
```