

Math233 Regression Project

Predicting Retail Sales

by

Christopher F J Wallis

Library Card No: 9709444

College: Lonsdale

Lecturer: Joe Whittaker

Tutor: Jan Currie

Aim of Analysis:

In this project, I am going to model data from past retail sales volumes, employment levels and national income wage & salary disbursements, with the aim of fitting a linear regression model both for each explanatory variable individually and also fitting a linear regression model for the Wages and Employment explanatory variables together. I will also use the Time regression model to predict future retail sales volumes and compare the values predicted to those given. I am unable to perform such a comparison for the either Wages or the Employment models, as I have no data for the explanatory variables to predict from, and using predicted data to then estimate other variables leads to notoriously erroneous results.

Review of Background Theory:

There are many different types of regression models available to statisticians, but here, I am concerned only with linear regression models - that is ones where the regression line is a straight one.

The most famous regression model is of the form:

$$y = \alpha + \beta \cdot x + z \quad \text{where } \alpha \text{ and } \beta \text{ are constants and } z \sim N(0, \sigma^2)$$

The simplest form of this model is where the x- and y-intercepts are at the origin:

$$y|x \sim N(\beta \cdot x, \sigma^2) \quad \text{i.e. } \alpha = 0$$

In the above models, the objective is to estimate β in such a way as to minimise σ^2 - that is to minimise the variability of β .

The more general linear model has equation:

$$y = X\theta + z \quad \text{with the following properties:}$$

- $E(z) = 0$ and $\text{var}(z) = \sigma^2 I$
- explanatory variables $X = [x_1 \quad x_2 \quad \dots \quad x_p]$ usually $x_1 = 1$
- unknown parameters $\theta = [\theta_1 \quad \theta_2 \quad \dots \quad \theta_p]^T$
- linearity $X\theta = [x_1 \quad \dots \quad x_p] \begin{bmatrix} \theta_1 \\ \theta_2 \\ \dots \\ \theta_p \end{bmatrix} = \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_p x_p$
- assumption $z = \begin{bmatrix} z_1 \\ \dots \\ z_n \end{bmatrix}$ has elements which come from a $N(0, \sigma^2)$ distribution

Leverage is an important concept to take into consideration, also. Leverage is the amount of “influence” that a particular point has on the regression line. Thus, if there is an out-lying point then it may be advantageous to eliminate it from the model, so that it does not have a detrimental effect on the regression model, and make the regression a less accurate reflection of the general trend of the data.

Once a model has been made, one can perform a residual analysis by way of a residual plot. Residuals are the difference between the values fitted by the model and the actual values used to construct the model. In the plot, standardised residuals are used, and all residuals should look to be mutually independent and look as if they come from a normal distribution. One can also construct a histogram of residuals, and, if it is a histogram of standardised residuals, most should lie between -2 and $+2$. If the histogram looks like a $N(0, \sigma^2)$ distribution, then the model is a good fitting one, if it is skewed, then it is not a good-fitting model.

The QQ plot is another useful tool for residual analysis. Here, the ordered residuals are plotted against the expected order statistics of a $N(0,1)$ distribution. If the model is a well-fitting one, then the plot should be fairly straight in the middle with possible slight variation at the tails. A poorly fitting model will not yield a very straight line. Thus it is possible to compare models by comparing QQ-plots of each - the one with the straightest QQ-plot is probably the better-fitting linear model of the two.

Log-likelihood is another useful tool for model analysis – the less negative the log-likelihood, the closer the estimated value for the parameters is to the actual values of the parameters.

In the case where the data does not suggest that there is a direct linear relationship, one can use transforms of the data, for example, by taking logarithms, square roots or other such, such that a linear model may be fitted to the transformed data.

[How to Replicate the Analysis:](#)

Please note that the regression analysis may be replicated by downloading and running a series of m-files, which can be found on my web site:

<http://www.lancs.ac.uk/ug/wallisc/courseworks/math233>

and then running them in the following order in Matlab:

salesssetup	prediction1
salesgraphs1	salesgraphs2
regress1	regress4
regress2	regressgraphs1
regress3	regressgraphs2

please note that all the files need to be downloaded for the analysis to work, and that a description of the function of each file is contained in each file.

[Set-up:](#)

To set up all the matrices required for the initial plots, I have written an m-file called "salesssetup.m". This file automatically loads the "salesdata.dat" file, providing that it is in the working directory, and then generates matrices for all of the variables for each of the four quarters of the year, as well as ones for the entire data set.

[Notation:](#)

Throughout this project, I have used certain abbreviations:

EMPL	≡	Employees on payrolls of non-agricultural establishments ('000s)
WASA	≡	National income and salary disbursements (\$ billions)
TIME	≡	Time quarter

Also, EMPL1≡ EMPL for first quarter of all years, etc., etc..

In the plots, different colours represent different quarters of the year:

blue	→	first quarter
green	→	second quarter
red	→	third quarter
black	→	fourth quarter
magenta	→	all quarters

Due to space limitations, it is not possible to show all graphs, QQ-plots, residual plots, etc., that I have used in this project. Therefore, I have included only the most important ones. The full set of graphs etc. may be viewed by replicating the project. (Details above)

Data:

The data for this project can be found at:

<http://lib.stat.cmu.edu/dasl> or in the file "salesdata.dat" on my website

The data is in a 44×4 matrix, and the columns represent: TIME; WASA; EMPL; GMER respectively.

Explanation of "Reganal.m"

"Reganal.m" is an m-function created for use in Matlab by the directors of the Math233 course. Below is the Matlab code contained in Reganal, and a brief explanation of what each line does:

```
Function [bhat,ci,yhat,yres,llik] = reganal(X,y)
[n,p] = size(X) ;
Bhat = (X'*X)^(-1) * X'*y ;
Yhat = X*bhat ;
Yres = y-yhat ;
sig2hat = yres'*yres/(n-p) ;
Varb = sig2hat * (X'*X)^(-1) ;
seb = (diag(varb)).^(0.5) ;
ci = [bhat - 1.96*seb, bhat + 1.96*seb] ;
Llik = -(log(yres'*yres/n) + 1)*n/2 ;
Return
```

% defines reganal as an m-function, with 2 input variables and 5 output variables
 % sets the values of n and p to be the size of the design matrix, X
 % estimates $\hat{\beta}$ using the formula: $\hat{\beta} = (XX^T)^{-1}(X^T y)$
 % computes \hat{y} by multiplying the design matrix, X, by the matrix of co-effs
 % computes the y-residuals
 % estimates the variance, σ^2
 % estimates the variance of $\hat{\beta}$ using the formula: $\text{var}(\hat{\beta}) = \hat{\sigma}^2 \times (X^T X)^{-1}$
 % estimates a 95% confidence interval for $\hat{\beta}$ using a variant on the formula:

$$ci = \hat{\beta}_j \pm 1.96 \sqrt{\text{var}(\hat{\beta}_j)}$$

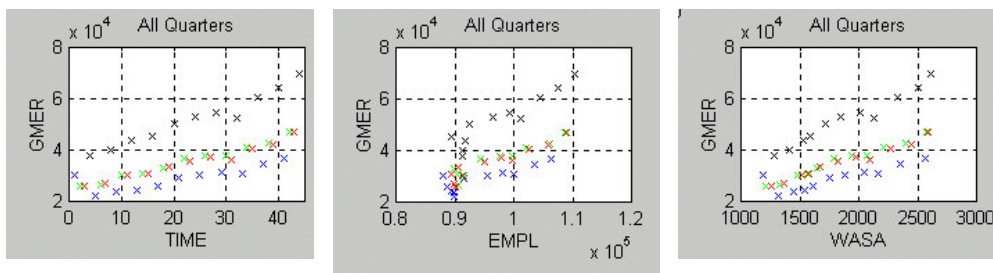
 % calculates the log-likelihood estimate, using the formula:

$$llik = -\frac{n}{2} \left(\log \left(\frac{(yres^T yres)}{n} \right) + 1 \right)$$

 % returns $\hat{\beta}$, ci, \hat{y} , $yres$, llik

Initial Plots:

The m-file "salesgraphs1.m" plots WASA, EMPL and TIME against GMER for each quarter and also for the entire data set. The graphs for the entire data set are below:



From the plots, it is clear that the time variable would lead to a good straight-line regression model.

It is also clear from the plots that GMER is dependant on what time of year it is, since there are clearly defined trends for each period, albeit the second and third quarters see similar GMER sales volumes.

From the TIME and WASA plots, it would appear that the very first value is an out-lier, so I will therefore exclude it from the rest of my analysis.

Simple Linear Model:

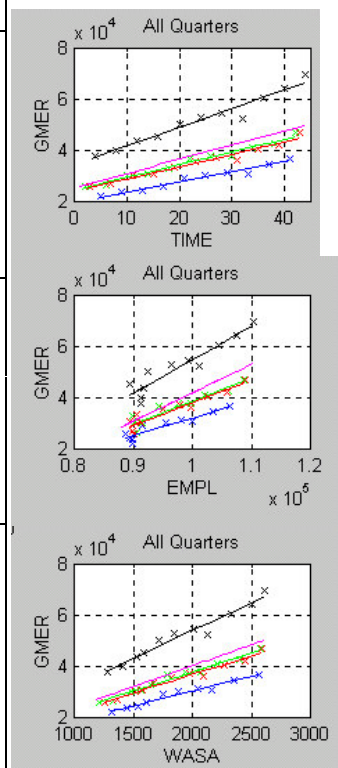
The first regression analysis I performed was by using each of the three explanatory variables in turn, using the "Reganal" m-file. Reganal estimates the coefficients of the model $\hat{y} = \hat{\alpha} + \hat{\beta}.x$ where $\hat{\alpha}$ and $\hat{\beta}$ are parameters, x is the explanatory variable, and \hat{y} is the estimated value of y at a given x . In all the regression analyses I will perform, I will assume that there is a y-axis intercept, such that $\alpha \neq 0$ in general, due to the suggestions from the initial plots.

I have created three m-files called "regress1.m", "regress2.m" and "regress3.m" which firstly alter the matrix for the first quarter so as to eliminate the very first piece of data, and then use the "reganal" math233 function to compute estimates for $\hat{\alpha}$ and $\hat{\beta}$ for each quarter and also for the entire data set.

The estimated coefficient values given by "reganal" are below:

Model	Quarter	$\hat{\alpha}$	$\hat{\beta}$	95% confidence interval		Log-likelihood
TIME	1	19,828	393	18,485	21,171	-72.6403
	2	24,470	500	23,213	25,726	-81.0243
	3	23,973	482	22,424	25,522	-82.9495
	4	34,401	735	31,570	37,231	-89.2182
	All Quarters	25,562	548	20,338	30,786	-420.0493
EMPL	1	-38,496	1	-56,444	-20,549	-79.0744
	2	-52,395	1	-74,439	-30,350	-90.5893
	3	-49,549	1	-67,993	-31,106	-88.8798
	4	-73,318	1	-102,820	-43,810	-94.3352
	All Quarters	-68,488	1	-103,800	-33,170	-418.0133
WASA	1	7,106	1.6	3,939	10,274	-73.3167
	2	7,858	5.0	4,972	10,744	-81.3241
	3	8,003	4.5	4,896	11,110	-82.0029
	4	10,495	2.0	4,771	16,219	-88.6440
	All Quarters	7,233	6.5	-4,263	18,728	-419.8832

The fitted regression lines may be seen in the plots below:



It is clear from both the plots and the estimates in the table that a fitted regression line for each quarter is better than a single one for all the data. This is backed up by smaller confidence intervals for the "quarter" model than for the model using all the data. The log-likelihood function is indicative towards this conclusion also. The QQ-plots confirm this assertion, as do the histograms of standard residuals and the plot of leverages against standard residuals. Unfortunately, there is not enough space to show this here!

Predicting GMER for 1990 using the TIME regression model:

To recap, the linear model for predicting GMER from TIME is:

$GMER = \hat{\alpha} + \hat{\beta}.x$ where $x = 45, 46, 47, 48$, depending on which quarter of the year we are predicting GMER for.

Below is the table of results of the predicted values:

	TIMEx Model	Time All Model	Actual Value	Closest Model
Quarter 1	37,528	50,220	41,832	TIMEx
Quarter 2	47,461	50,768	50,181	TIME All
Quarter 3	46,622	51,316	34,911	TIMEx
Quarter 4	69,663	51,864	36,543	TIME All

Clearly, neither model is particularly accurate, and the models need to be refined further to eliminate the high degree of inaccuracy.

Regression Model Using Two Explanatory Variables:

Next, I performed the regression analysis using two explanatory variables together - WASA and EMPL. It was not possible to include TIME in the multiple regression analysis, however, due to it being measured on a purely arbitrary scale. Thus the linear model becomes:

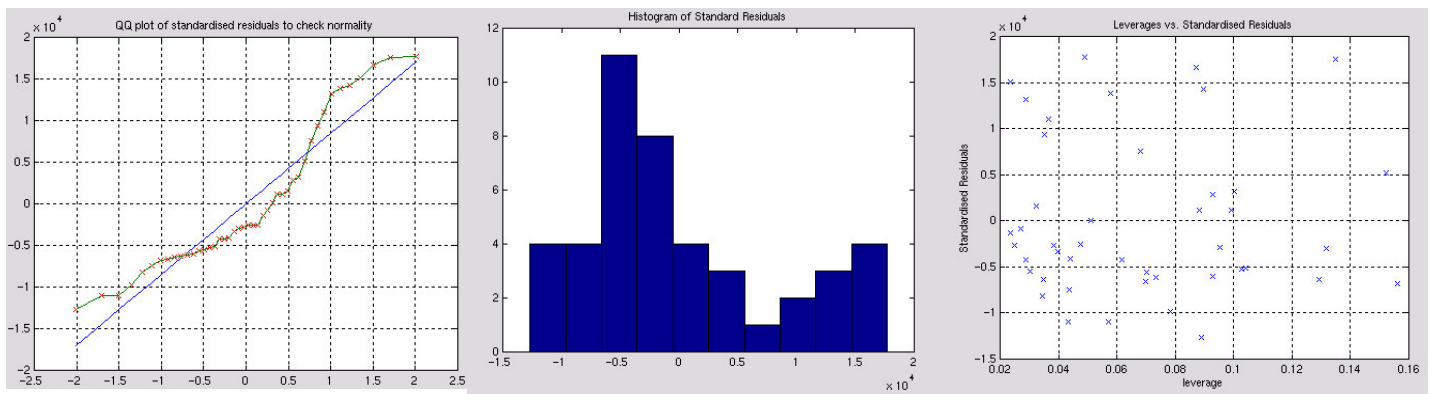
$$\underline{\hat{y}} = \begin{bmatrix} 1 & EMPL_1 & WASA_1 \\ M & M & M \\ 1 & EMPL_n & WASA_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

i.e. $\hat{y} = \hat{\alpha} + \hat{\beta}_1(WASA) + \hat{\beta}_2(EMPL)$

The estimated coefficient values given by "reganal" are below:

Quarter	$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	95% confidence interval	Log-likelihood
1	21,669	0	15	-7,810 51,148 -1 0 8 22	-72.6813
2	24,654	0	19	434 48,874 -1 0 13 24	-80.1679
3	14,778	0	16	-14,835 44,392 -1 0 9 23	-81.8647
4	26,504	0	26	-26,411 79,418 -1 1 13 38	-88.4045
All Quarters	-86,094	1	-4	-18,051 832 0 0 -3 2	-417.9299

Below, is the QQ-plot, histogram of standard residuals and plot of leverages vs. standard residuals:



The QQ-plot is not particularly straight, and is quite variable towards the extremities. This indicates a fairly poorly fitting model. The log-likelihood functions are similar to those in the simple regression model performed earlier. The histogram of standardised residuals is clearly skewed, which also implies

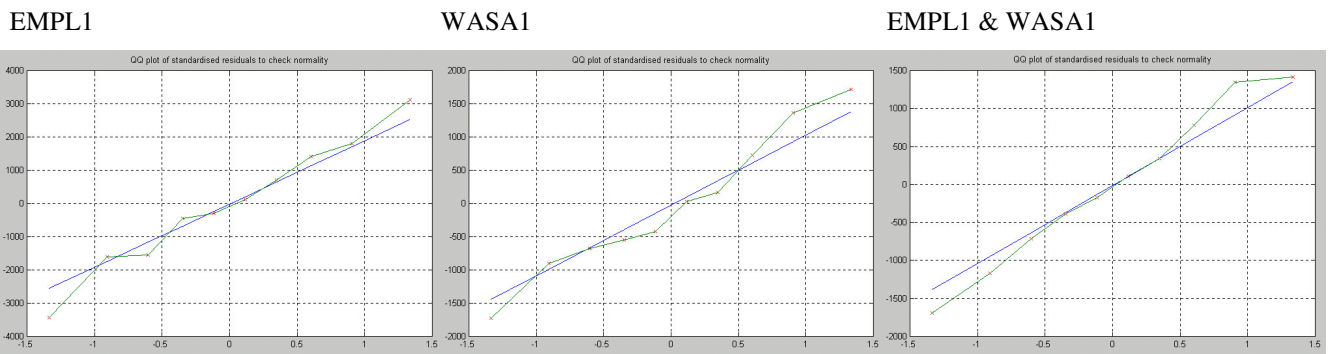
that this model does not fit the data very well. From the above table of estimated coefficients, it would appear that $\hat{\beta}_1$ (EMPL) has relatively little influence over the linear regression model compared to $\hat{\beta}_2$ (WASA). This begs the question of whether this model is superior to the first model that I constructed.

Certainly, the confidence intervals are proportionately much wider in this second model, which is indicative of a less-well fitting model than the first one. And, comparing the residuals, the residuals on the second model are greater than those from the first model, also indicating that the first model was the more superior of the two.

Comparison of the Two Types of Linear Models:

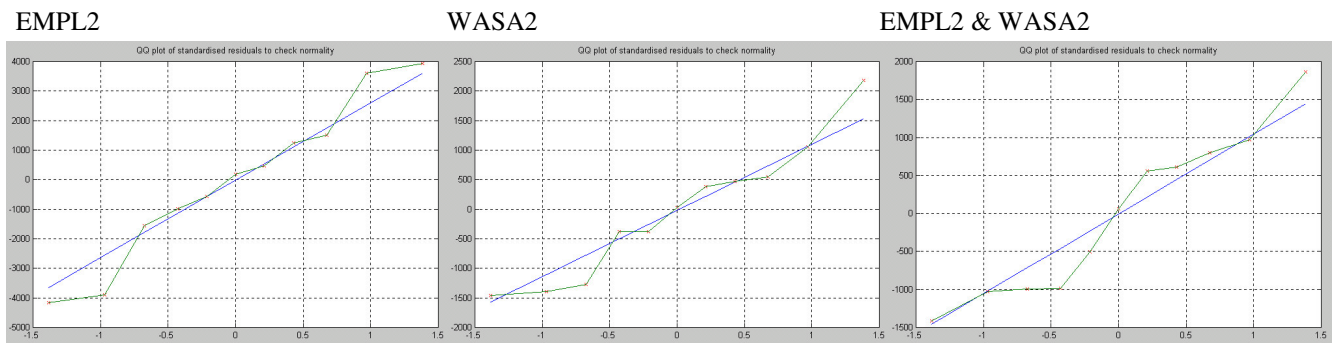
Due to space restrictions, it is not possible to demonstrate a comprehensive comparison of the two types of models, so I shall show a few selected examples of how the two types of linear models compare:

First, let me compare the QQ-plots of the models for the first quarter:



For the QQ-plots for the first quarter, it would appear that the second model is a better-fit, as the QQ-plot is almost straight in the central region, with small variances at either end, whereas for when modelling EMPL1 and WASA1 separately, there is clearly much more variance throughout.

Now, let's look at the QQ-plots of the models for the second quarter:



These seem to suggest that the first model, where EMPL and WASA are modelled separately yields a better-fitting model that the second model, where EMPL and WASA are used together to estimate GMER. Thus, it is easy to see that depending on what quarter of the year you want to predict values for, for some, the first model is a better estimate, and for others, the second model is more accurate.

As I have mentioned before, there is another slight complication to our analysis - that is whether we use the sub-model that calculates regression lines for each quarter separately, or whether we use the sub-model that takes into account the data for the entire year when calculating the regression lines. In the previous section, we saw in the "Simple Linear Model" section that sometimes the quarterly models were better, and other times the whole-dataset model gave a more reliable estimation.

Thus, I have drawn the following conclusions of which model is best for each quarter of the year:

	Best Model	Reasoning
Quarter 1	Model 2 - Quarter 1	Straightest QQ-plot Smallest standardised residuals
Quarter 2	Model 1 - Quarter 2	Straightest QQ-plot Normally-distributed residuals
Quarter 3	Model 2 - Quarter 3	Straightest QQ-plot Small standardised residuals Normally-distributed residuals
Quarter 4	Model 1 - Quarter 4	Normally-distributed residuals Small standard residuals

Postscript

Given more space, I would have like to have tried transforming some of the data. Whilst the TIME explanatory variable yields a fairly linear relationship with GMER, the linearity of WASA against GMER is perhaps slightly less clear (cf initial plots). However, I've run out of space, and can thus only suggest that this may provide a superior model to the ones that I have already investigated.