# Chapter 9: Outline Solutions

Solutions were generated using Minitab and SPSS unless otherwise stated.

### 9.1

Define *X*3 = (0 if female, 1 if male). Given the expression:



When *X3* = 0 the equation reduces to



When *X3* = 1 the equation becomes



### 9.2

40 observations were generated for the model:



The coefficients X1 – X4 represent the four quarters of the year, so the data correspond to 10 years of data. The error terms are random numbers generated from independent normal distributions with mean 0 and variance 1. A data file is available on the exercise solutions web site, Data\_Exercise\_9.1.xlsx. The results of analyzing the data will differ due to the use of a random number generator.

*Minitab Output* [The results for SAS, SPSS and other major statistical programs are identical]

**Regression Analysis: Y versus X1, X2, X3, X4**

\* X4 is highly correlated with other X variables

\* X4 has been removed from the equation.

The regression equation is

Y = 18.0 - 12.0 X1 - 10.4 X2 - 8.62 X3

Predictor Coef SE Coef T P

Constant 17.9559 0.3378 53.16 0.000

X1 -12.0403 0.4777 -25.20 0.000

X2 -10.4218 0.4777 -21.82 0.000

X3 -8.6195 0.4777 -18.04 0.000

S = 1.06818 R-Sq = 95.5% R-Sq(adj) = 95.1%

Analysis of Variance

Source DF SS MS F P

Regression 3 863.62 287.87 252.30 0.000

Residual Error 36 41.08 1.14

Total 39 904.69

At first sight this looks nothing like the generating equation. However, if we recall that:



We may substitute for X4in the original equation to obtain:



The estimated coefficients are all close to the original input values. Also, note that the program automatically deletes the fourth variable because of the linear relationship. If you delete one of the other Xs instead, the results look different until analyzed as above, when all are equivalent.

The S, R2 and ANOVA results are always identical.

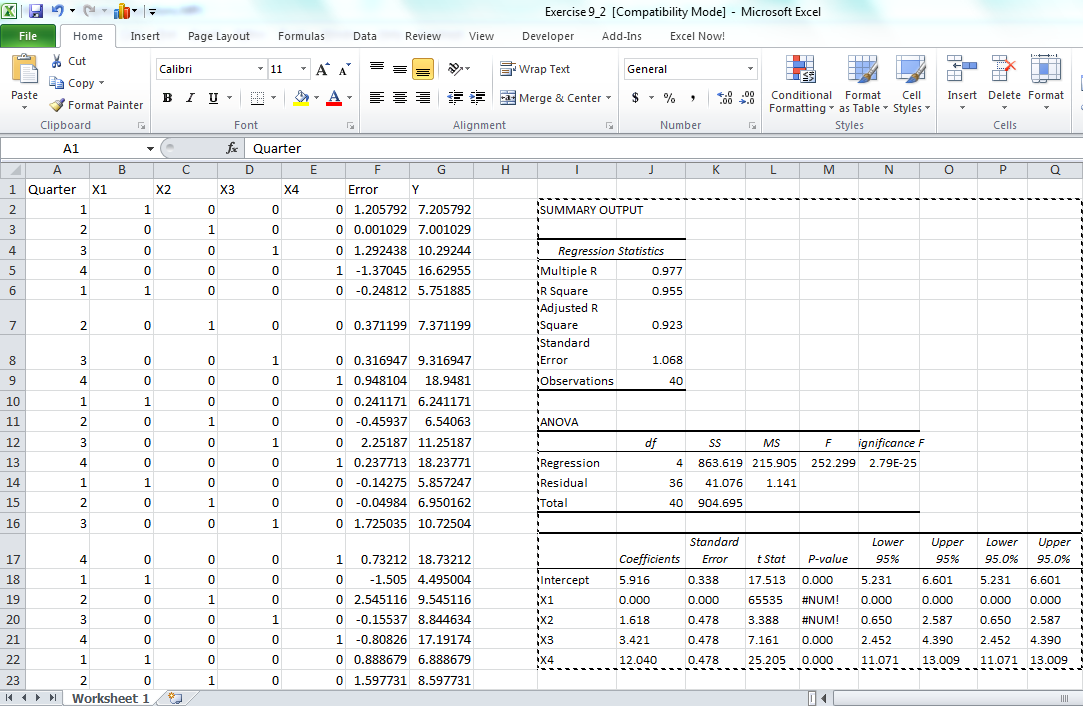
*Excel Output [Using Analysis ToolPak; edited to fit page]]*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.977 |  |  |  |  |  |  |  |
| R Square | 0.955 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.923 |  |  |  |  |  |  |  |
| Standard Error | 1.068 |  |  |  |  |  |  |  |
| Observations | 40 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 4 | 863.619 | 215.905 | 252.299 | 2.79E-25 |  |  |  |
| Residual | 36 | 41.076 | 1.141 |  |  |  |  |  |
| Total | 40 | 904.695 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 5.916 | 0.338 | 17.513 | 0.000 | 5.231 | 6.601 | 5.231 | 6.601 |
| X1 | 0.000 | 0.000 | 65535 | #NUM! | 0.000 | 0.000 | 0.000 | 0.000 |
| X2 | 1.618 | 0.478 | 3.388 | #NUM! | 0.650 | 2.587 | 0.650 | 2.587 |
| X3 | 3.421 | 0.478 | 7.161 | 0.000 | 2.452 | 4.390 | 2.452 | 4.390 |
| X4 | 12.040 | 0.478 | 25.205 | 0.000 | 11.071 | 13.009 | 11.071 | 13.009 |

The coefficients, standard errors and *t*-values are correct if we substitute for X1 in the original equation and ignore the entry for X1 in the output table. However, the P-value for X2 is undefined, instead of the correct value of 0.002.

The Excel ANOVA gives the correct results, apart from incorrect DF. The results remain incorrect even in the latest 2016 release of Excel.

The full Excel worksheet is shown below:



### 9.3

To check the consistency of different choices of indicator variables, it is sufficient to look at two distinct cases. We considered WFJ Sales on X1 – X4 and again on X2 – X5. The first 56 observations were used as the estimation sample and forecasts generated for the last six weeks.

**Regression Analysis: WFJ Sales versus X1, X2, X3, X4**

Predictor Coef SE Coef T P

Constant 34986 1561 22.41 0.000

X1 -8403 2031 -4.14 0.000

X2 -1246 2208 -0.56 0.575

X3 -497 3662 -0.14 0.893

X4 -206 1842 -0.11 0.911

S = 4683.95 R-Sq = 36.7% R-Sq(adj) = 31.7%

Analysis of Variance

Source DF SS MS F P

Regression 4 649002371 162250593 7.40 0.000

Residual Error 51 1118907940 21939371

Total 55 1767910310

Predicted Values for New Observations

New Obs Fit SE Fit 95% CI 95% PI

1 34780 977 (32819, 36741) (25174, 44385)

2 34986 1561 (31851, 38120) (25074, 44898)

3 34986 1561 (31851, 38120) (25074, 44898)

4 34780 977 (32819, 36741) (25174, 44385)

5 34780 977 (32819, 36741) (25174, 44385)

6 34780 977 (32819, 36741) (25174, 44385)

**Regression Analysis: WFJ Sales versus X2, X3, X4, X5**

Predictor Coef SE Coef T P

Constant 26583 1299 20.46 0.000

X2 7157 2031 3.52 0.001

X3 7905 3558 2.22 0.031

X4 8197 1625 5.04 0.000

X5 8403 2031 4.14 0.000

S = 4683.95 R-Sq = 36.7% R-Sq(adj) = 31.7%

Analysis of Variance

Source DF SS MS F P

Regression 4 649002371 162250593 7.40 0.000

Residual Error 51 1118907940 21939371

Total 55 1767910310

Predicted Values for New Observations

New Obs Fit SE Fit 95% CI 95% PI

1 34780 977 (32819, 36741) (25174, 44385)

2 34986 1561 (31851, 38120) (25074, 44898)

3 34986 1561 (31851, 38120) (25074, 44898)

4 34780 977 (32819, 36741) (25174, 44385)

5 34780 977 (32819, 36741) (25174, 44385)

6 34780 977 (32819, 36741) (25174, 44385)

### 9.4

The notation is Lj for Lag j.

The general characteristics of the fitted model are similar to the model in Section 9.3. However, note that a number of observation (in particular 2008 Q4) are flagged as “unusual” heralding the onset of the Great recession. As the figure shows, the forecast function recovers quite well.

**Regression Analysis: Walmart Sales versus L1\_GDP, L1\_Sales, L4\_Sales, and seasonal dummies**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | | Unstandardized Coefficients | | t | Sig. |
| B | Std. Error |
| 1 | (Constant) | 5.537 | 6.902 | 0.802 | 0.428 |
| Quarter\_1 | -20.552 | 2.938 | -6.995 | 0.000 |
| Quarter\_2 | -7.773 | 1.052 | -7.389 | 0.000 |
| Quarter\_3 | -12.677 | 1.622 | -7.815 | 0.000 |
| L1\_GDP | 0.002 | 0.001 | 1.593 | 0.121 |
| L1\_Sales | 0.713 | 0.136 | 5.258 | 0.000 |
| L4\_Sales | 0.126 | 0.097 | 1.296 | 0.204 |

Regression Equation

Sales\_2013 = 5.54 - 20.55 Quarter\_1 - 7.77 Quarter\_2 - 12.68 Quarter\_3 + 0.001530 L1\_GDP

+ 0.713 L1\_Sales + 0.1258 L4\_Sales

Fits and Diagnostics for Unusual Observations

Obs Sales\_2013 Fit Resid Std Resid

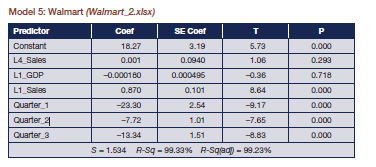
24 108.00 111.21 -3.21 -2.27 R

35 109.50 106.68 2.82 2.03 R

37 112.30 109.35 2.95 2.16 R

R Large residual



The results are similar to those shown for the full data set. 

|  |  |  |
| --- | --- | --- |
|  | Absolute Error | % Absolute Error |
| Average | 2.72 | 2.28 |
| Median | 2.51 | 2.09 |

The forecast error statistics are: The MAE is 2.72 and the MAPE is 2.28%.

But all forecasts were above the actuals though within the prediction Intervals.



### 9.5

The notation is Lj for Lag j. We now use the estimation sample to 2010.

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 3.02 8.96 0.34 0.740

Quarter\_1 -20.24 3.62 -5.59 0.000 37.58

Quarter\_2 -7.17 1.09 -6.55 0.000 3.43

Quarter\_3 -12.93 1.96 -6.58 0.000 11.05

L1\_GDP 0.00189 0.00116 1.63 0.118 19.21

L1\_Sales 0.679 0.173 3.92 0.001 78.62

L4\_Sales 0.130 0.111 1.17 0.256 42.52

Regression Equation

Sales\_2010 = 3.02 - 20.24 Quarter\_1 - 7.17 Quarter\_2 - 12.93 Quarter\_3 + 0.00189 L1\_GDP + 0.679 L1\_Sales + 0.130 L4\_Sales

Model Summary

S R-sq R-sq(adj) R-sq(pred)

1.35351 99.26% 99.05% 98.74%

Fits and Diagnostics for Unusual Observations

Obs Sales\_2010 Fit Resid Std Resid

24 108.000 111.113 -3.113 -2.52 R

R Large residual



There is substantial evidence of the effect of the Great recession

|  |  |  |
| --- | --- | --- |
|  | Absolute Error | %Absolute Error |
| Mean | 4.14 | 3.54 |
| Median | 4.56 | 3.57 |

### 9.6

The notation is Ln for logarithm, Lj for Lag j. The patterns are qualitatively similar, but the forecasts are consistently above the actual values.

**Regression Analysis: LN\_sales versus LN\_L1\_GDP, LN\_L4\_Sales, ...**



Error statistics are shown below, the forecasts being worse than the untransformed model:

|  |  |  |
| --- | --- | --- |
|  | Absolute error | APE |
| Mean | 4.67 | 3.84 |
| Median | 5.17 | 4.43 |

### 9.7

We will use the analysis of the log model as illustrative. The notation is D for difference, LN for logarithm, Lj for Lag j.

The calculation of one-step-ahead forecasts requires some care:

Step 1: Forecast for LN(Yt) = Actual for LN(Yt-1) + Forecast for D\_LN(Yt)

Step 2: Convert back to original units

The prediction intervals are calculated in the same way.

|  |  |  |
| --- | --- | --- |
|  | Absolute Error | % Error |
| Mean | 1.95 | 1.60 |
| Median | 1.67 | 1.44 |

**Regression Analysis: D\_LN\_Sales versus D\_LN\_L1\_GDP, D\_LN\_L4\_Sales, plus seasonal dummies... using data to 2013Q4.**

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 0.0719 0.0216 3.33 0.002

Quarter\_1 -0.1095 0.0378 -2.90 0.007 29.59

Quarter\_2 -0.0849 0.0284 -2.99 0.005 17.98

Quarter\_3 -0.0701 0.0226 -3.09 0.004 11.43

D\_Ln\_L1\_GDP 0.692 0.408 1.70 0.099 1.18

D\_Ln\_L1\_Sales -0.311 0.168 -1.85 0.073 32.30

D\_Ln\_L4\_Saes 0.383 0.156 2.46 0.020 28.92

Regression Equation

D\_ln\_Sales = 0.0719 - 0.1095 Quarter\_1 - 0.0849 Quarter\_2 - 0.0701 Quarter\_3

+ 0.692 D\_Ln\_L1\_GDP - 0.311 D\_Ln\_L1\_Sales + 0.383 D\_Ln\_L4\_Saes

Fits and Diagnostics for Unusual Observations

Obs D\_ln\_Sales Fit Resid Std Resid

24 0.10130 0.14575 -0.04445 -2.64 R

R Large residual





### 9.8

The timing of the Great Recession and its effects on retail sales is uncertain. Certainly by the end of 2008 we see in the Walmart data an outlier. Setting a recession indicator as =1 for Q3 and Q4, 2008, we get the following (estimated on data to 2013 – compare with exercise 9.6)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summaryb** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .996a | .992 | .990 | .01765 |
| a. Predictors: (Constant), Recession, ln\_L1\_Sales, Quarter\_3, Quarter\_2, Quarter\_1, ln\_L1\_GDP, ln\_L4\_Sales | | | | |
| b. Dependent Variable: ln\_Sales\_2013 | | | | |



While the variable is not significant in this model (one-tailed test) it does decrease the standard error. It also leads to a better specified model that includes GDP.

### 9.9

The log model is as follows:



With forecast error statistics which compare with Table 4.3.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Absolute Error | % AE | ME |
| Average | 31.40 | 12.13 | 31.40 |
| Median | 30.44 | 10.95 | 30.44 |

The four-variable log model performs better but the forecasts are always below the ‘actual’ because of the effect of SP500. The effects of the Great Recession lead to outliers, starting June 2009.

All variables are significant, however a log model with lagged Crude and lagged Unleaded gives much better one-step ahead forecast.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Absolute Error | % AE | ME |
| Average | 7.76 | 3.10 | 3.49 |
| Median | 6.88 | 2.61 | 3.09 |

### 9.10

Solution follows 9.9 above

### 9.11

We will use the two-variable model which includes L3 variables in crude and unleaded. In fact, the lagged crude price is insignificant and the preferred model is autoregressive with lag 3. The summary error measures for the 3-period autoregressive model, estimated to Dec2008 are:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Absolute Error | % AE | ME |
| Average | 23.99 | 10.30 | 11.66 |
| Median | 20.28 | 7.30 | 16.37 |

The forecast error for Jan 2009 was major, raising the question as to why – but of course the recession was developing rapidly and relationships in the economy were changing.

As seen compared to 9.9 the deterioration has been substantial. While the median errors are lower they remain substantial.

### 9.12

We use the simple model of example 9.4 where only *Crude\_price* has been included. We could additionally use the recession dummy, or specify the model in logs. These latter two extensions would be sensible from earlier analysis.

We use EViews (see the discussion in Section 10.5) where a variety of criteria are available to help specify an appropriate lag length. Alternatively, forward selection in regression could be used. We just present AIC and BIC here:



The selected minimum is for 9 lags using AIC and just 2 if using BIC. Simplicity argues for the latter. The regression model with autoregressive error (Section 9.4.4) is a constrained model of the one we propose here with two lags in the two variables. When estimated the results are:



The above table showing evidence of the need to include longer lags in the unleaded model.

### 9.13

There is a lack of clarity in this question as to which models should be considered. We suggest the models with the following explanatory variables:

1. Lagged: *Unleaded, Crude*, *PDI, Unemp, RR\_Sales, SP500, Demand* as well as a recession dummy: NB: *demand* is measured in gallons and therefore should not be deflated
2. The differenced model (see Table 9.7)
3. Optionally the log model in the above variables.

Using forward selection on model 1 only *Unleaded, Crude* and the Recession Indicator are included.

Model 1:

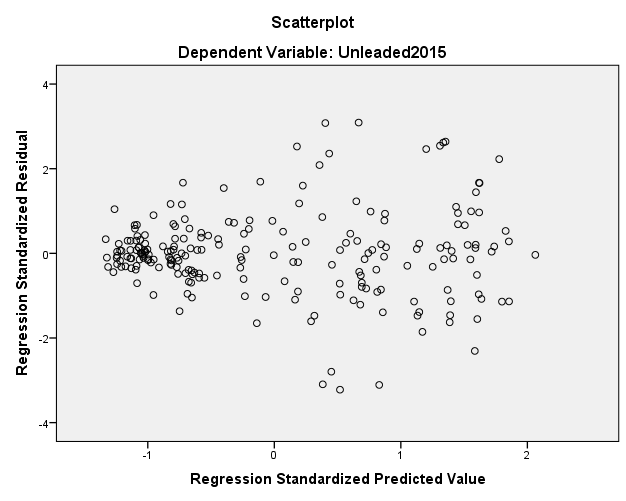
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summaryd** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .979a | .959 | .959 | 6.89063 |
| 2 | .983b | .967 | .966 | 6.24248 |
| 3 | .987c | .974 | .974 | 5.54149 |
| a. Predictors: (Constant), L1\_R\_Unleaded | | | | |
| b. Predictors: (Constant), L1\_R\_Unleaded, Recession | | | | |
| c. Predictors: (Constant), L1\_R\_Unleaded, Recession, L1\_R\_Crude | | | | |
| d. Dependent Variable: R\_Unleaded2015 | | | | |

Note the std. error is not directly compatible with the nominal (rather than real) data.

We may start from the full model and gradually remove redundant terms. The most striking feature is the extremely high VIF values for PDI and CPI (hardly surprising). Thus, we may drop one of these terms with minimal effect on the goodness of fit, as seen in the second analysis.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summaryb** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .992a | .984 | .983 | 11.94090 |
| a. Predictors: (Constant), L1\_CPI, Recession, L1\_Crude\_price, L1\_Unleaded, L1\_PDI | | | | |
| b. Dependent Variable: Unleaded2015   |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Model Summaryb** | | | | | | Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | | 1 | .992a | .983 | .983 | 12.11438 | | a. Predictors: (Constant), L1\_PDI, Recession, L1\_Crude\_price, L1\_Unleaded | | | | | | b. Dependent Variable: Unleaded2015 | | | | | | | | | |



The residual characteristics are similar for both models showing higher variance for larger predicted values. Also autocorrelation.

The Error summary statistics are:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Absolute Error | | % Absolute Error | |
|  | Real | Nominal | Real | Nominal |
| Average | 12.99 | 13.69 | 4.85 | 5.09 |
| Median | 9.38 | 8.84 | 3.63 | 3.78 |

Model 2:

The model for the real data is: 

While for the nominal data:



The residuals for both models are reasonably well-behaved apart from some extreme outliers, despite the inclusion of the Great Recession dummy.

To calculate the forecasts for the real model the following transformation of the forecast of the real differences needs to be carried out.



The error statistics shown below marginally favor the real model. From our earlier analysis, consider dropping the SP500.



### 9.14

The effect of a unit increase in *X*in period (t+1) adds *β1* to what *Y* would otherwise have been, all other terms in the equation remaining the same.

But at period (t+2), *Yt+1* is higher than it would have been (without the unit increase in *X)*. Thus, the forecast for period (t+2) is *β1+φ1 β1+ β2,* this last term arising from the increase from *Xt-1­.* [N.B. note that the ‘answer’ is incorrect in the text].

While it is possible to continue this calculation, the easy way is to recognize that there is a long-term increment of 1 for both Xt and Xt-1 while in equilibrium (as t gets large), *Yt, Yt-1 and Yt-2*all must converge to the same value, µ say.

So



Giving the result. Note that this is the incremental effect above that if X had remained at the same level. This leads to  being eliminated.

### 9.15

Model Summary

S R-sq R-sq(adj) R-sq(pred)

10.3737 93.89% 92.45% 87.69%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 755.8 30.5 24.74 0.000

advert 1.912 0.325 5.89 0.000 83.27

AdvertSq -0.003672 0.000832 -4.42 0.000 84.98

newdum1 19.86 5.14 3.87 0.001 1.35

newdum2 -13.11 8.18 -1.60 0.127 2.03

Regression Equation

sales = 755.8 + 1.912 advert - 0.003672 AdvertSq + 19.86 newdum1 - 13.11 newdum2

Fits and Diagnostics for Unusual Observations

Obs sales Fit Resid Std Resid

1 874.53 897.87 -23.34 -2.83 R

R Large residual



While the residual plots for the log-log model don’t look much different the forecasts (measured by MAE and MdAE – chosen since the absolute amount of sales is more relevant than the relative APE) are better for the levels model rather than the log model. In summary, neither model describes all the data.

The log-log model proved poorer at the end of the 22 periods although there is no evidence of autocorrelation.



### 9.16

Week 52 had the assignment 15-30s: 25 - 75 split. The consequence of including an indicator for the outlier is to adjust the effect of that advertising measure; the others remain unchanged (compare p.281). However, the fit of the model is improved, as we would expect.

As before only the “all 15s” strategy stands out as inferior to the others.

**Regression Analysis: WFJ Sales versus 100% 15s, 15-30s: 75-25 split, ...**

Predictor Coef SE Coef T P

Constant 34488 3332 10.35 0.000

100% 15s -7905 3579 -2.21 0.031

15-30s: 75-25 split -749 3683 -0.20 0.840

15-30s: 25 - 75 split -646 3622 -0.18 0.859

15-30s: 50-50 split -305 3453 -0.09 0.930

S = 4711.76 R-Sq = 31.2% R-Sq(adj) = 26.3%

Analysis of Variance

Source DF SS MS F P

Regression 4 573206576 143301644 6.45 0.000

Residual Error 57 1265437104 22200651

Total 61 1838643680

**Regression Analysis: WFJ Sales versus 100% 15s, 15-30s: 75-25 split, ...**

Predictor Coef SE Coef T P

Constant 34488 2844 12.12 0.000

100% 15s -7905 3055 -2.59 0.012

15-30s: 75-25 split -749 3145 -0.24 0.813

15-30s: 25 - 75 split -2453 3116 -0.79 0.434

15-30s: 50-50 split -305 2948 -0.10 0.918

Outlier 19879 4219 4.71 0.000

S = 4022.69 R-Sq = 50.7% R-Sq(adj) = 46.3%

Analysis of Variance

Source DF SS MS F P

Regression 5 932448569 186489714 11.52 0.000

Residual Error 56 906195111 16182056

Total 61 1838643680

9.17 The Excel file Lasso\_results includes a sheet named “Estimation data”. If you go to column AH in that sheet, it provides step-by-step details of how to compute the Lasso coefficients. The results are summarized in the sheet headed “Lasso results” and the hold-out data are available in the sheet “Hold-out Data”.

The Lasso method constrains all the coefficients, whereas the Group Lasso allows the two key variables *L1\_Unleaded* and *L1\_Crude\_price* to be weighted separately. As may be seen from the table, the Lasso method loses information whereas the Group approach downplays the effect of the less relevant variables and yields slightly more accurate results.

The detailed calculations are available in Lasso\_results\_Ex9.17 in the Solutions file.

|  |  |  |  |
| --- | --- | --- | --- |
|  | ME | MAE | RMSE |
| OLS | -0.050 | 0.147 | 0.194 |
| Lasso (1.0) | -0.324 | 0.358 | 0.445 |
| Group (0.5) | -0.061 | 0.140 | 0.189 |

**Minicase 9.1**

We will make comments on the data base to 2010 with the next two years as hold-out sample

1. Starting with the full regression model of all the contemporaneous variables, Housing may be dropped as unimportant.

2. Bringing in the VIF, shows the redundancy between CPI and PDI. Retaining CPI makes more sense as gas prices factor into that index (although that is hardly explanation!).

3. Removing PDI leaves S&P500 as not significant, so that might also be removed.

4. The small value of the DW statistics means we need to allow for autocorrelation: add L1\_Unleaded and possibly L1\_crude.

There are many ways to go and the reader should explore the options.

****

**Regression Analysis: Unleaded versus *Crude, Unemp*, etc but including lagged *Unleaded\_price* and *lagged crude* + recession dummy.** (We know for earlier analysis that these are crucial)

Stepwise gives a fairly simple model that includes *Production* with Std.Error of 9.15.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
| B | Std. Error | Beta | Tolerance | VIF |
| 5 | (Constant) | 53.088 | 14.437 |  | 3.677 | 0.000 |  |  |
| Crude\_price | 1.909 | 0.191 | 0.693 | 9.998 | 0.000 | 0.018 | 56.609 |
| L1\_Unleaded | 0.668 | 0.051 | 0.666 | 13.037 | 0.000 | 0.032 | 30.772 |
| L1\_Crude\_price | -1.034 | 0.259 | -0.373 | -3.995 | 0.000 | 0.010 | 102.943 |
| Recession | -13.861 | 5.543 | -0.034 | -2.501 | 0.013 | 0.470 | 2.126 |
| Production | -0.168 | 0.067 | -0.035 | -2.492 | 0.014 | 0.425 | 2.353 |
| a. Dependent Variable: Unleaded | | | | | | | | |

The sign is probably appropriate – higher production, lower prices (though the causality is an issue –why: lagged production?). And how does that compare with lags. Note that the relationship with *Crude* contains both current and lagged effects.

A backward selection model always includes many more variables and as discussed, *CPI* and *PDI* both have high VIF. (So do crude and lagged crude but that’s typical and ok). We would need to eliminate one – and from earlier analysis perhaps using real variables would help – check it. But the standard error is noticeably lower from this more complex model, Std.Error=8.23. Perhaps there is an explanation of the dramatic changes in the last two years of the data.

But we also know that differencing the data might well work well both as offering an explanation and a predictive model. So let’s difference the variables.

Now the Unleaded\_price is probably non-stationary. And the evidence suggests a differenced model as a possibility. Using forward selection gives:

**Regression Analysis: *D\_Unleaded* versus *D\_Crude, L1\_Unleaded* etc+ *Production.*..**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | | Unstandardized Coefficients | | t | Sig. | Collinearity Statistics | |
| B | Std. Error | Tolerance | VIF |
| 6 | (Constant) | 54.478 | 13.531 | 4.026 | 0.000 |  |  |
| D\_L1\_Crude | 0.872 | 0.171 | 5.105 | 0.000 | 0.615 | 1.627 |
| Recession | -0.714 | 5.795 | -0.123 | 0.902 | 0.378 | 2.647 |
| Crude\_price | 1.851 | 0.179 | 10.326 | 0.000 | 0.018 | 56.552 |
| L1\_Crude\_price | -1.215 | 0.245 | -4.956 | 0.000 | 0.010 | 104.667 |
| L1\_Unleaded | -0.267 | 0.050 | -5.365 | 0.000 | 0.031 | 32.770 |
| Production | -0.189 | 0.063 | -2.992 | 0.003 | 0.426 | 2.349 |

The various coefficients associated with *Crude\_price* support the view that the lag structure is complicated and that crude prices feed through into retail unleaded over a number of periods (the issue in exercise 9.12). A degree of simultaneity is to be expected in that the data is monthly and it would not (necessarily) take a month for prices at the pump to change.

Once we have established a number of plausible models and we might develop new compound variables (e.g. supply=Imports+production-demand), we would need to examine the residuals and of course forecasting performance. So, yes, there is a difference between an explanatory model and a forecasting model. The out-of-sample performance of the explanatory model is one important test of the validity of a model. But a forecasting model can use indicator or lagged variables which may in themselves not describe the economic processes under analysis. If we now use lagged variables from the above model we get:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | | Unstandardized Coefficients | | t | Sig. | Collinearity Statistics | | Std. Error |
| B | Std. Error | Tolerance | VIF |  |
| 4 | (Constant) | 1.403 | 0.908 | 1.546 | 0.124 |  |  |  |
| Recession | -25.107 | 5.472 | -4.588 | 0.000 | 0.791 | 1.264 |  |
| D\_L1\_Crude | 1.357 | 0.206 | 6.591 | 0.000 | 0.791 | 1.264 |  |
| D\_L1\_Demand | 0.867 | 0.546 | 1.589 | 0.114 | 0.996 | 1.004 | 11.72 |
| 5 | (Constant) | 1.421 | 0.912 | 1.559 | 0.121 |  |  |  |
| Recession | -25.346 | 5.494 | -4.614 | 0.000 | 0.792 | 1.263 |  |
| D\_L1\_Crude | 1.368 | 0.207 | 6.621 | 0.000 | 0.792 | 1.263 | 11.77 |

Check diagnostics, forecasting performance when compared to a simple ARIMA or exponential smoothing model. *D\_L1\_Demand*  has turned out to have somewhat robust characteristics (see Table 9.7).

### **Minicase 9.2**

An initial analysis just based upon indicators for the months clearly shows the need to model the intervention. When Inter is added, everything checks out well. Discussions on possible trends, autocorrelation, etc. can be explored.

**Regression Analysis: Injuries versus Feb, Mar, ...**

Predictor Coef SE Coef T P

Constant 1591.30 52.43 30.35 0.000

Feb -210.80 74.14 -2.84 0.005

Mar -111.40 74.14 -1.50 0.136

Apr -240.60 74.14 -3.25 0.002

May -147.30 74.14 -1.99 0.049

Jun -188.40 74.14 -2.54 0.012

Jul -134.80 74.14 -1.82 0.072

Aug -103.90 74.14 -1.40 0.164

Sep -13.30 74.14 -0.18 0.858

Oct 109.40 74.14 1.48 0.143

Nov 283.90 74.14 3.83 0.000

Dec 426.60 74.14 5.75 0.000

S = 165.792 R-Sq = 61.1% R-Sq(adj) = 57.2%

Analysis of Variance

Source DF SS MS F P

Regression 11 4665170 424106 15.43 0.000

Residual Error 108 2968586 27487

Total 119 7633757



**Regression Analysis: Injuries versus Feb, Mar, ...**

Predictor Coef SE Coef T P

Constant 1636.31 34.96 46.81 0.000

Feb -195.80 49.16 -3.98 0.000

Mar -96.40 49.16 -1.96 0.052

Apr -225.60 49.16 -4.59 0.000

May -132.30 49.16 -2.69 0.008

Jun -173.40 49.16 -3.53 0.001

Jul -119.80 49.16 -2.44 0.016

Aug -88.90 49.16 -1.81 0.073

Sep 1.70 49.16 0.03 0.972

Oct 124.40 49.16 2.53 0.013

Nov 298.90 49.16 6.08 0.000

Dec 441.60 49.16 8.98 0.000

Inter -300.05 25.46 -11.78 0.000

S = 109.878 R-Sq = 83.1% R-Sq(adj) = 81.2%

Analysis of Variance

Source DF SS MS F P

Regression 12 6341931 528494 43.77 0.000

Residual Error 107 1291826 12073

Total 119 7633757

Durbin-Watson statistic = 1.68752

