Department of Statistics

STATS 330 Assignment 1 - Model Answers 23 August 2002

Evaluating Sales Performance

Arden Miller

Executive Summary

A model was identified that predicts the sales for a territory based on five explanatory variables: the length of time the salesperson has been with the company, industry sales in units for the territory, the dollar expenditures on advertising, the weighted average of market share for 4 previous years and the weighted average of market share for 4 previous years. For specified levels of these variables an interval can be generated that will contain the sales for a territory with a specified level of confidence. These intervals can be interpretted as containing the plausible range of values for sales given the values of the explanatory variables. Thus the sales of a territory can be considered as unusually low if they fall below the appropriate prediction interval for that territory.

Sales Data

The data investigated in this report represents a random sample of 25 sales territories for a company. The following measurements were recorded for each territory:

SALES	sales in units for the territory (response).
TIME	length of time the salesperson has been with the company.
POTENT	industry sales in units for the territory.
ADV	dollar expenditures on advertising.
SHARE	weighted average of market share for 4 previous years.
SHARECHG	change in market share over the 4 previous years.
ACCTS	total number of accounts assigned to the salesperson.
WORKLOAD	an index that measures the average workload per account.
RATING	an aggregate rating of performance by the field sales manager.

Box plots for these measurements are presented in Figure 1. The plot for SALES indicates that the values range from (approximately) 1600 to 6500 with 50% of the values being between 2400 and 3400. The box plots for the other variables indicate the range and distribution of these variables. As well these plots indicate there is one unusual value for each of the variables TIME, SHARECHG, ACCTS, and WORKLOAD.

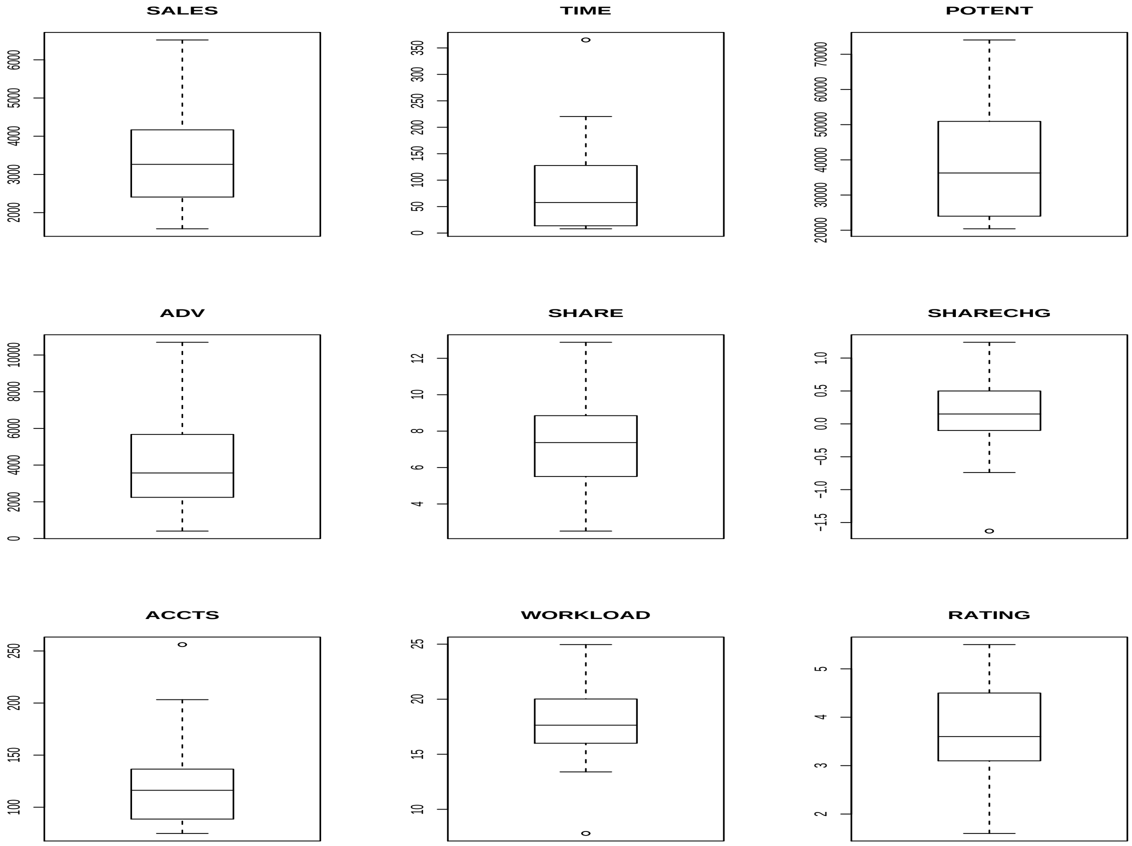


Figure 1: Box Plots of the Measurements.

Predicting Sales

The following model was selected to predict Sales:

$$\begin{aligned}\mathsf{SALES} &= \exp(6.80 + .00114 \times \mathsf{TIME} + .0000124 \times \mathsf{POTENT} \\ &\qquad + .0000395 \times \mathsf{ADV} + .0650 \times \mathsf{SHARE} + .0842 \times \mathsf{SHARECHG})\end{aligned}$$

This model only contains five out of the eight possible explanatory variables contained in the data. This does not mean that the three variables (ACCTS, WORKLOAD, RATING) that do not appear in the model are not related to the response. Rather it is a case of these variables not providing useful additional information to that provided by the five variables that were included in the model. It should be noted that this model should only be used to make predictions for territories that have values of TIME, POTENT, ADV, SHARE, and SHARECHG that fall in the ranges of values for these variables indicated by the plots in Figure 1.

The fitted model indicates that the predicted value of SALES increases as the value of each of the five explanatory variables increases. The impact that each explanatory has on the predicted sales is summarised in Figure 2. For each plot, the predicted SALES is plotted versus one of the explanatory variable that is varied over the range of values it exhibited in the data set while all the other variables are held constant at their mean values. These plots show that predicted

SALES is most affected by changes in POTENT (industry sales in units for the territory) and SHARE (weighted average of market share for 4 previous years).

When using the above model to predict SALES for individual territories it is necessary to take into account the estimation error. The standard statistical approach to this problem is to create prediction intervals. These are intervals that will contain the value of SALES with a specified level of confidence for a single (future) observation that has a specified set of values for the explanatory variables. For example if the values of the explanatory variables are set to their means (POTENT = 38858, TIME = 87.64, ADV = 4357.4, SHARE = 7.565, SHARECHG = 0.0956) then the predicted value of SALES is 3150.4 and the 95% prediction interval is (2454.2, 4044.2).

To evaluate, the precision of predictions made using the above model, 95% prediction intervals were calculated for each point in the data set. The width of these intervals varied from approximately $1000 to nearly $4000. This indicatcates that the width of intervals are severely affected by the values of the explanatory variables. Thus it will be necessary to generate an interval for each combination explanatory variables for which we want to make predictions.

To use this model to identify territories in which sales are unusually low it will be necessary to account for the uncertainty in the predicted value of SALES. One approach would be to generate a prediction interval with a specified confidence level (e.g. 99 %) level for each territory we wish to evaluate. This interval can be interpretted as the reasonable range of values for SALES for that territory. Thus if the observed SALES is below the lower boundary of the prediction interval, this territory has unusually low sales.

Statistical Appendix

First, the model that used log(SALES) as the response and contained all 8 explanatory variables was tried. For this model, 4 of the explanatory variables hand non-significant P-values. I removed ACCTS as it had the largest P-value and refitted the model. There were still unnecessary variables, so I repeated the process of dropping the explanatory variable with the largest P-value and refitting the model. Following this procedure I dropped WORKLOAD next and then RATING. At this point the output from the summary command in *R* looked like:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.801e+00  1.142e-01  59.557  < 2e-16 ***
TIME        1.147e-03  3.214e-04   3.571 0.002041 **
POTENT      1.242e-05  1.831e-06   6.782 1.78e-06 ***
ADV         3.949e-05  1.007e-05   3.921 0.000919 ***
SHARE       6.497e-02  1.064e-02   6.104 7.20e-06 ***
SHARECHG    8.417e-02  4.277e-02   1.968 0.063868 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The variables TIME, POTENT, ADV and SHARE are clearly needed in the model but SHARECHG is marginal. I decided to keep it in the model but it would be reasonable to drop it as well.

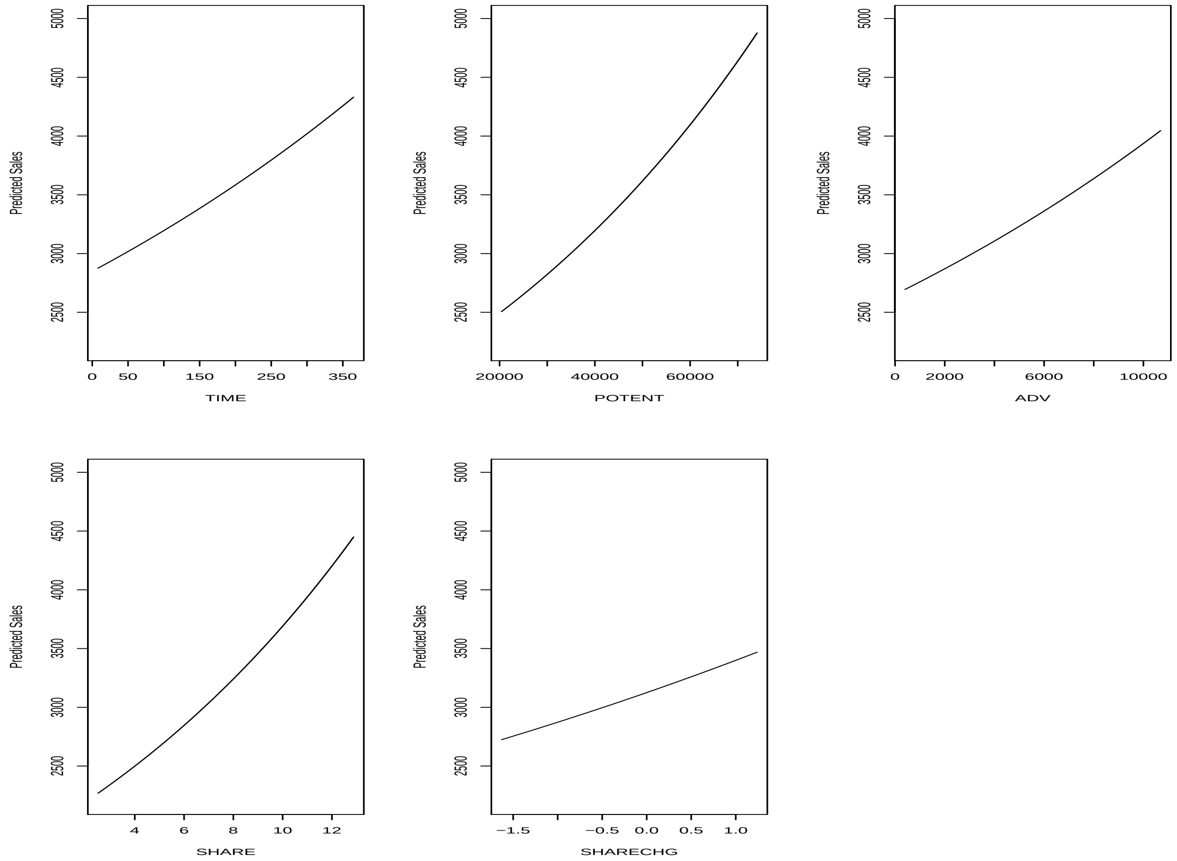
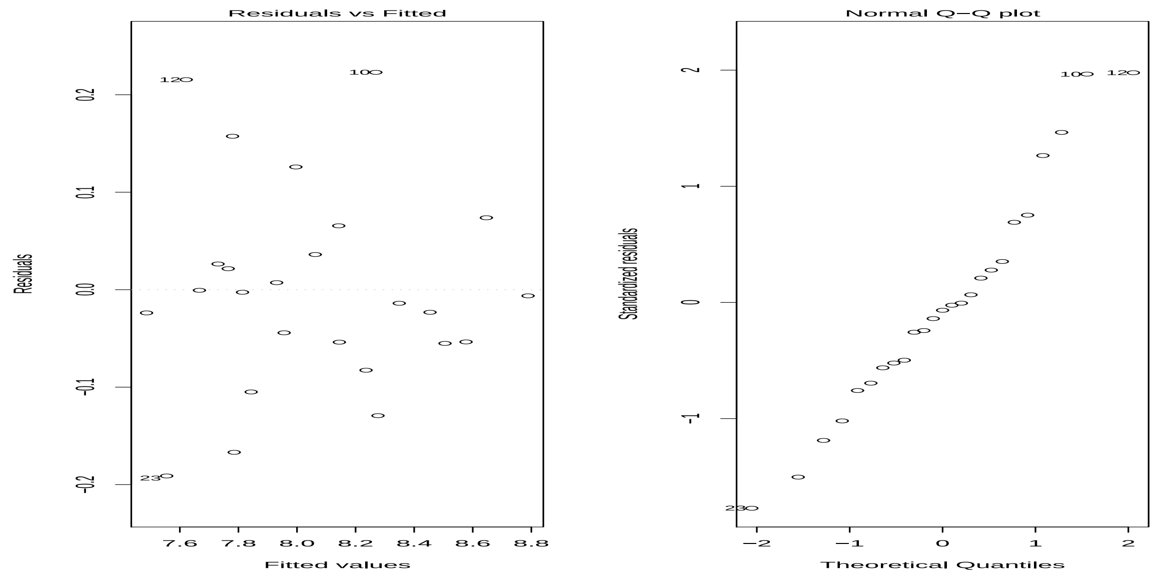


Figure 2: Predicted Sales as a Function of each Explanatory Variable

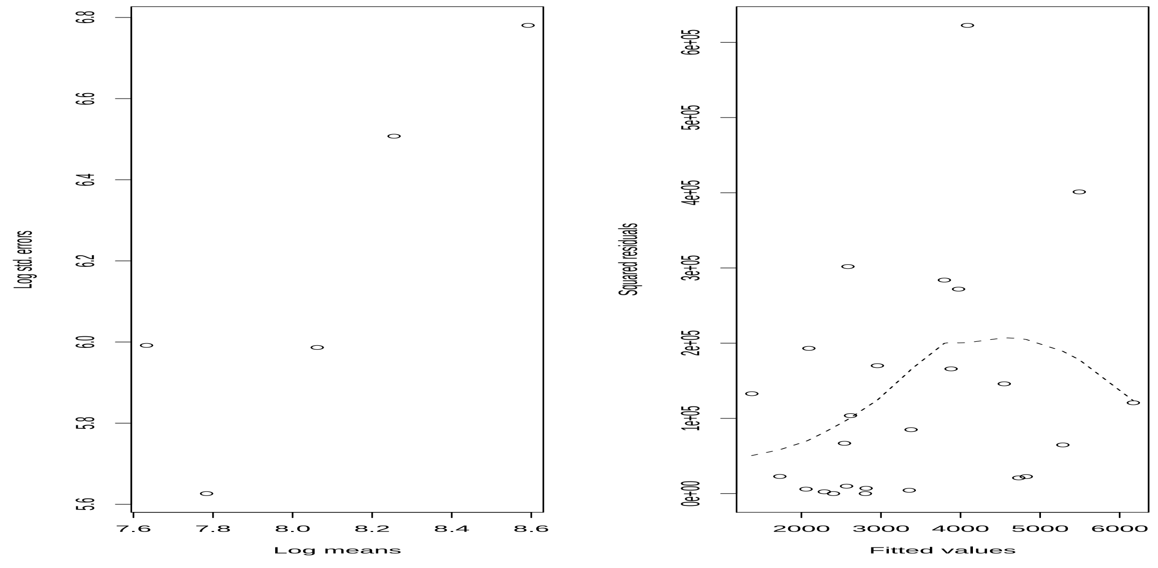
I evaluated the precision of the predictions made from the model by considering 95% prediction intervals. I chose to use prediction intervals rather than confidence intervals since the model was to be used to assess the value of SALES for individual territories. To get prediction intervals for SALES using my model: First I created 95% prediction intervals for log(SALES) and then I used the exponential function to transform these into intervals for SALES.

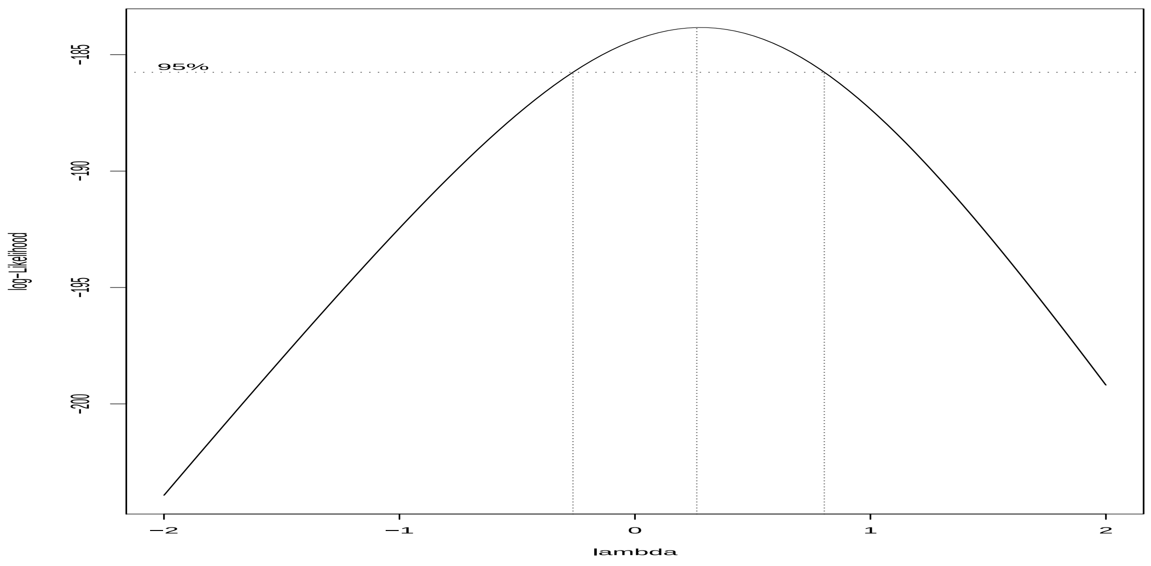
The use of log(SALES) as the response seems reasonable. The plot of the residuals versus the fitted values for the identified model does not show clear evidence of a trend or of a funnel effect and the normal probability plot of the residuals is reasonably linear.



I also checked the "funnel" plots and the partial residual plots. These did not indicate that either non-constant variance or non-linearity was a problem.

I also tried fitting the model using SALES as the response. The plots of the residuals versus fitted values and the probability plot of the residuals both seem reasonable for this model. However, the first plot produced by the funnel command (logged standard deviations versus logged means) indicates a positive trend with a slope of 1.06. This suggests that a log transformation of the response is sensible. I also did a Box-Cox plot of this data which also suggested that a log transformation of the response was reasonable.




log−Likelihood
-185
-190
-195
-200
95%
−2
−1
0
1
2
lambda