Statistics 712, Spring 1999

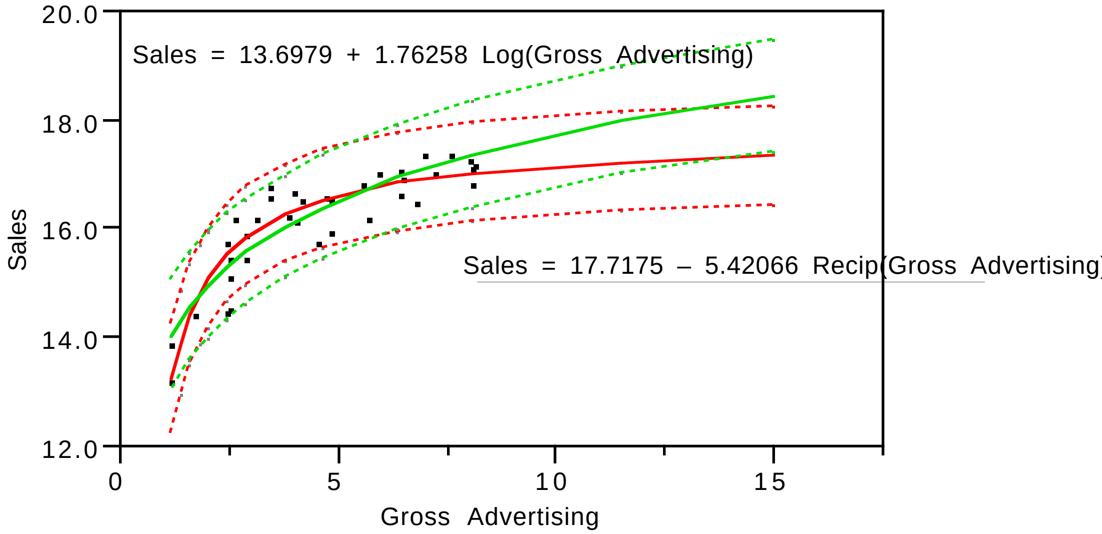
Final Exam Solutions Outline

This exam was intended to get you to review all of the material we have covered in the course and apply it within the context of one common problem. In that sense, it's a bit of a case study that pulls some of these ideas together.

The context for these questions revolves around decisions at a privately held company that sells several established computer software products. Its products are mature and incur very small production and distribution costs, about $0.10 per dollar in sales, not including promotional expenses. Promotional expenses are the main expenditure of this firm and have been increasing. In a battle to maintain market share, this company has steadily increased its advertising budget and is now considering plans to increase its level of promotion over the next year. Two managers have different opinions about how this planned level of expenditures for new advertising will affect sales. Manager A believes that sales will rise as expenditures increase, and offers a set of projections based on a regression model. Manager B is less optimistic and feels that the response of sales to more advertising will be flatter over this range of expenditures; he too has a model to support his conjecture. Both have submitted projections for the level of sales. Underlying data and these projections are in the file **Final.jmp** that you can get from the class web page. The data are quarterly with 36 quarters prior to the forecast period. The sales and advertising values in the data file are in millions of dollars ($US).

The ownership of the firm has changed recently; the firm was inherited by one of the nine children of the previous owner (quarter 34 in this data). This new owner has a current net worth of about $20M, and is considering taking the firm public.

(1) The models used by the two managers appear in the plot below (with 95% *individual*



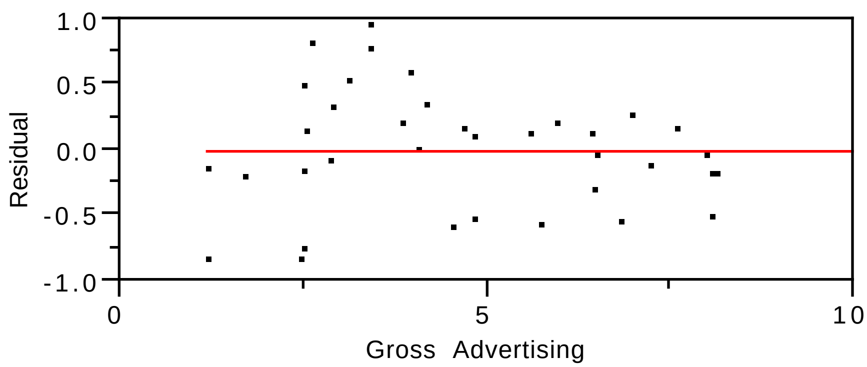
prediction limits). Manager A (shown in green, growing) used a model with log of advertising as the predictor, whereas Manager B (red, flattening out) used a model with the reciprocal of sales as the predictor. These are the same models that we considered previously in the example that considered the dependence of sales of a liquor store and amount of shelf space devoted to a product.

Both of these simple regression models show decreasing returns to scale in that we expect to see less additional sales for each additional dollar of advertising as the amounts increase. Said differently, the slopes of the models become flatter as the levels of advertising increase. The effects of increases in advertising diminish with more spending.

(2) You should consider both the numerical as well as graphical summaries of these models. The model of Manager A using the log of advertising produces an R2 of about 80% and RMSE = 0.463, whereas that of Manager B using the reciprocal has a slightly larger R2 of about 83% with a correspondingly smaller RMSE = 0.434. With only 36 observations,

RSquare	0.804	RSquare	0.828
RSquare Adj	0.798	RSquare Adj	0.823
Root Mean Square Error	0.463	Root Mean Square Error	0.434
Mean of Response	16.206	Mean of Response	16.206
Observations (or Sum Wgts)	36.000	Observations (or Sum Wgts)	36.000

these differences are quite small and not significant, though we have not considered a specific statistical test for this question.



Turning to the graphical evidence, the residual plots from these models show problems. Both sets of residuals show some evidence of systematic pattern or perhaps a lack of constant variance during the period of rapid sales growth (when advertising was less than 5). These simple, one-predictor models cannot accommodate the uneven growth rates in the early data. Perhaps other predictors are needed, but it is clear that neither model captures all of the trends here. You might also want to check for normality, but the samples are rather small.

(3) The p-values of the model obtained by the data miners are not, of course, to be taken at face value. These predictors were not chosen in advance, but rather were determined by using a search method (stepwise regression), as you can check by using JMP's multiple regression/stepwise tool with the response surface option. The issue, then, is how many predictors were considered in this process.

The output given with the question shows 7 additional predictors, and one of which, *Season*, is categorical. This one categorical term representing 4 categories requires (think back) 3 "dummy" variables to represent itself in a regression model. That gives a total of 10 (1 + 7 + 2) basic predictors. The data miners also used combinations of predictors to search for nonlinear terms. This gives a total of about 10 + 90/2 + 6 = 61 possible predictors (10 for the predictors themselves, plus all possible pairs, and six quadratic or squared terms as well). At this rate, there are more possible predictors than observations (61 vs. 36) so they have clearly had to use some form of sequential search, starting from a simple model and building up to the form shown in the exam. If they try to use them all at once, the fitting process is overwhelmed and has many ways to find a perfect fit to the data. The sequential searching process, as we found in class, can lead to an over-specified model that fits the data perfectly (R^2 = 1.00) and a clearly biased estimate of the noise variance (i.e., RMSE is way too small).

This combination of many, many predictors with a biased estimate of RMSE leads to grossly inflated t-statistics very small p-values that can "fool" the Bonferroni procedure. The simple Bonferroni bound is tricked by the biased estimates of noise variance. If we continue onward and use the Bonferroni rule, we get a bound of 0.05/61 ≈ .00082. Four of the predictors pass this test: Computer sales, Competitor Adv, Comp Sales*TV Share, and Comp Adv*TV Share. This use of Bonferroni does not go far enough since it gives a threshold that is based on the flawed assumption that we tried all 61 predictors in *one* model. Instead, the model was found via a sequential process that leads to bias in the estimated RMSE (which in turns makes the t-statistics very large since the RMSE is in the denominator for each). Another way to think about this is to realize that we get a new set of regression coefficients as the sequential process goes along. In a sense, we tried 61 coefficients in the first model, then 60 more in the second, and so forth. In effect, we have looked at many more than the 61 that Bonferroni would suggest. (As to grading, getting to the 61 or a similar value earned you full credit. I was curious if any would take it further. One or two did.)

As to prediction, the R^2 for this model is about 91%, about 10% higher than those of the two manager-chosen models with a smaller RMSE. Since the future values of many of these predictors are not known, this is not a big advantage since we'll still have to fill in the predictors with some estimates. It's questionable that it will really predict better given this additional uncertainty.

Finally, the data-driven model is very hard to interpret. What do these coefficients mean, substantively? Do you trust a model such as this that you cannot explain? Thinking further along these lines, this model does not include a transformation of Gross Advertising itself, and we cannot directly compare it to the manager's models using an F test. But think about that: it does not include the transformation to log or reciprocal that the simpler models of the managers use to capture dimishing returns to scale. The complex model has missed an important aspect of the simpler manager models.

(4) The analysis continues using the log model of Manager A.

(a) Using the multiple regression tool (with log of advertising as the one predictor), we can find the 95% individual interval precisely to be [17.44, 19.50]. If you read the values from the plot, you should come close to this interval. Since we are extrapolating in this case outside the range of data, we get an interval that is wider than the (prediction)±2 RMSE formula that works inside the range of the data. We have to allow for the statistical extrapolation penalty. It is not very large in this problem, but as we have seen it can be quite large and dominate the RMSE contribution.

(b) The interval for profits is a simple transformation of the interval for sales. Profits are defined to be 0.9(Sales) – Gross Adv, and we get the interval for profits by substituting our range for predicted sales into this equation,

$$0.9[17.44, 19.50] - 15 = [15.70, 17.55] - 15 = [0.70, 2.55]$$

(5) The owner has k=20, so if we use the approximate formula from class which applies when there is absolute risk aversion, we get a CEV of

$$\text{CEV} \approx \text{mean} - \text{variance}/2k \rightarrow \text{CEV} \approx 1.625 - .4625^2/40 = 1.625 - .0053 = 1.619$$

Should the owner proceed? Sure, the CEV is positive so it's worth it to continue. The range of uncertainty is small relative to the owner's wealth, so the positive mean value for profits implies a worthwhile investment. This answer, as noted in the question, assumes the truth of this regression model and the stability of this market. It also ignores the practical issue of whether the money might be better spent elsewhere than in this increasingly competitive market. The owner might be able to get more bang for the buck, so to speak, by spending a bit less on advertising. But then, that's a bit speculative – this campaign might be necessary to react to the moves of a competitor.

You could also do this problem with a simulation, by finding the average utility of the profits, and converting this average utility back to the profits scale. With a sample of 20,000 to represent the uncertainty in the interval from Q4, the average utility is 0.0778. Inverting the absolute risk utility function U(x) = 1 – e(-x/k) gives

$$x = k \log(1/(1-u)) = 20 \log 1/(1-.0778) = 20(.081) = 1.62$$

as found in the initial approximation.

(6) The interval found in Q4 above for Manager A is [0.7, 2.55]. The analogous interval for Manager B is

$$(0.9)[16.44, 18.27] - 15 = [-0.204, 1.443]$$

As Cauchy sources, these overlap and give the pooled intervals based on samples defined by

Manager A	1.625 + 0.925 (?cauchy)
Manager B	0.6195 + 0.8235 (?cauchy)

The pooled simulated 50% interval "pulls up" the interval of Manager B and is positive:

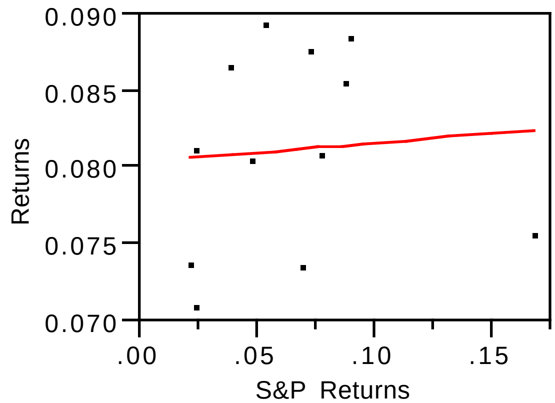
[0.58, 1.59] .

(7) These calculations can be done directly from the matched samples used to find the pooled interval in Q6. You don't need to simulate further since you already have a sample from this pooled source. Since the Cauchy is not absolute risk aversion (it is similar for positive values), you need to compute the utility as current wealth + profits. The utility is affected by the wealth of the individual. In this problem, the owner might not view the project so favorably were his wealth smaller.

(a) The expected utility of net wealth (20 + profits) using the logistic form is 0.739. If I start "from scratch" and use the interval found in Q6 directly (as 21.085+0.505(?cauchy) with the 20 added for current wealth), I get a mean utility of 0.740 from a sample of 20,000 – almost identical with that from the subset found in Q6.

(b) The CEV is the value *p* of profits that matches the average utility 0.486 just found. With the logistic utility function, this means that 0.740 = U(p) = 1/(1+e(-p/20)), or solving for the inverse function, p = k log(u/(1-u)) = 20 log(0.74/.26) = 20.92. Thus the CEV for the profits of this promotion is about $0.92 million.

(8) The correlation between the returns for this firm and the given S&P returns is quite



small, only 8% and is not significant (t=0.26). The plot shows the relationship. With so little data (only 12 observations), we have little chance to find any pattern, and the small correlation is not significant. Thus, as far as our investing procedure is concerned, these two investments are uncorrelated and can thus be treated as separate decisions. Assuming the usual form of risk aversion with k=wealth, the multiplier for each is given by the mean return over its variance. For the 12 quarters, the relevant summary statistics of the quarterly returns are

	This Firm	S&P
mean	0.081	0.065
variance	0.00004	0.0017
ratio	2025	38

Both series have positive returns, and the returns of this firm are higher. The difference is not quite significant using a paired t-test. The CI for the difference in means just barely includes zero. The risk-adjusted returns for this firm as given are huge. They should be reduced by the risk-free rate, but I forgot to include that in the file. Doing so will give a more reasonable number. You could have found something in the right ballpark from the data given out with the assignments, or just subtracted off a constant estimate such as two or three percent. Ignoring that aside, these ratios imply that you should buy as much as you can. But then, the risk-adjusted returns for the S&P are also quite large for these data. Much larger than what we found using a much more complete data set. Seems that we are looking at tool little data for this procedure, choosing a multiple of wealth as a ratio of mean to variance, to be very reliable.

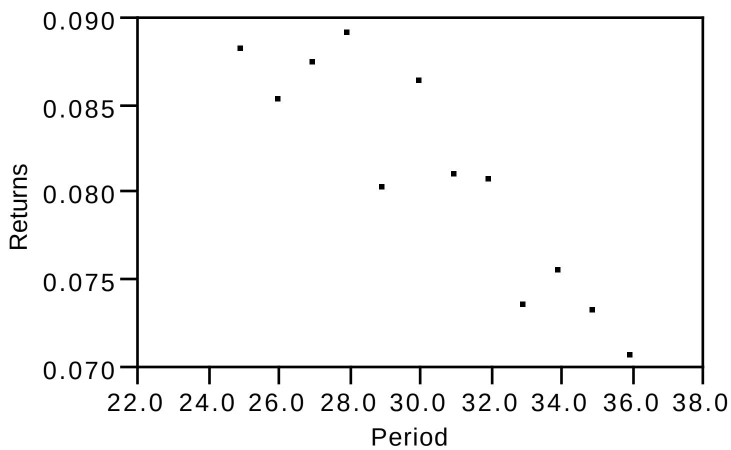
Given this analysis and the lack of data, it is important to notice the clear downward trend in the returns for this company. Given its competitive environment, this bodes poorly for the future. A plot appears in Q10 below.

(9) If you had to pool the intervals for quarters 37-40 into one…

(a) The idea is that you have the intervals, and now need to combine them. To do so, you would need to know the dependence among the prediction errors to find an interval for the next year. Since all of the predictions come from the same model, the predictions are likely to share common errors, such as missing a trend. It is the dependence among the prediction errors, not the source data, that is relevant. Even if the data are time series with autocorrelation, this is not a problem so long as the autocorrelation has been properly modeled. Rather, it's some error in our model that is of more concern. For example, suppose our model underpredicts future sales. It's likely to do that in all 4 quarters, so that the prediction errors are systematically related. Even if the raw data is itself not correlated, the prediction errors of the model are. Of course, you'd like to know about such systemic errors, but if you did, I'd assume that you would fix the problem. The hard part is to allow for the chance of such errors, even after you have fixed all the problems of which you are aware.

(b) This dependence among the forecast errors would be expected to make the pooled interval longer. Dependent sources are not so informative as independent sources, so the pooled interval would not be so narrow as one implied by assuming, incorrectly, that the forecast intervals were independent. You have to allow for a systematic error that would expand the range of possible outcomes.

(10) Certainly the trend in the profits shown below is disturbing. This firm is in a very competitive market and must invest more to "keep afloat" and maintain its share. One has to have qualms about its future. Nonetheless, profits have been very consistent and, though trending downward, suggest good returns for the near future. (The slope in the plot is small, though very steady. When compared to the variation in the S&P, this series is quite steady.)



The comment to "Keep in mind that there are a great many privately held firms in the US, and relatively few of these are taken public each year." was intended to make you think of selection bias. Companies that are taken public are not representative of the large collection of privately held companies. Most likely, they "look good" on paper in the sense of appealing to investors, and this appearance may be fleeting. Though a bit cynical, I am suggesting that the timing of an IPO is chosen to match a particularly good period for a company, and that the performance prior to the IPO may not be indicative of future returns. The hype of some IPO's also starts to lead to some of the things we considered in auctions, with excessive bidding and winner's curse. It is unlikely that a firm such as this would inspire frenzied bidding without some growth potential.

Finally, some of you pointed out that the company could be better served by reducing its planned advertising to a less expensive campaign. Since it is getting less reward (sales) for spending the additional dollars once the sales are large, it gets more profit by spending less in advertising. Again, this assumes that the large campaign is not, in some sense, necessary in the current market.