JMP Academic Case Study 020

Direct Mail

Regression and Forecasting

Produced by

Marlene Smith, University of Colorado Denver Business School

jmp STATISTICAL DISCOVERY

Direct Mail

Regression and Forecasting

Background

An antique dealer recently opened a shop. The store is open to the public on Tuesdays through Sundays from 10am to 6pm. Direct mail (unsolicited flyers and informational brochures delivered via U.S. post) is the primary advertising outlet. Direct mailings are sent out every Wednesday.

The Task

Use regression to determine if sales are related to the direct mail campaign. If they are related, determine the nature of the relationship.

The Data **directmail.jmp**

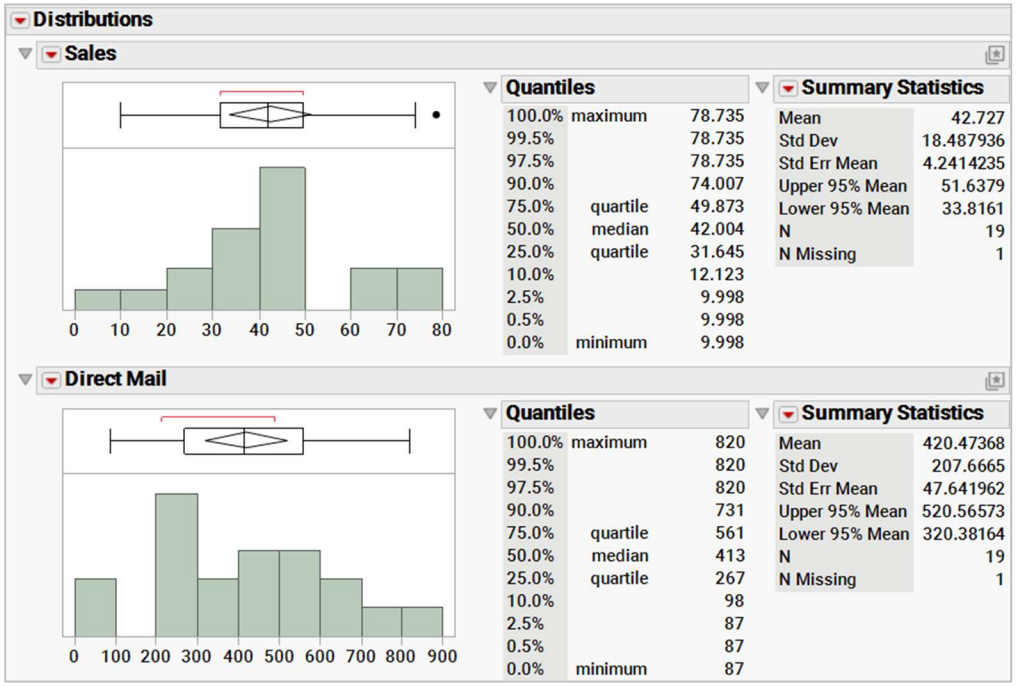
This is a weekly time-series using all available data since the opening of the dealership. The variables in the data set are Week and:

Sales	Weekly (Tuesday through Sunday) total revenues in $1,000
Direct Mail	Weekly direct mail costs, including supplies and postage, in dollars

Analysis

The average level of sales over the first 19 weeks was around $42,730 with a typical fluctuation of $18,500 (see Exhibit 1). Direct mail costs averaged roughly $420 per week, although there was a lot of variability in direct mail expenditures from week to week, with a typical fluctuation of $208.

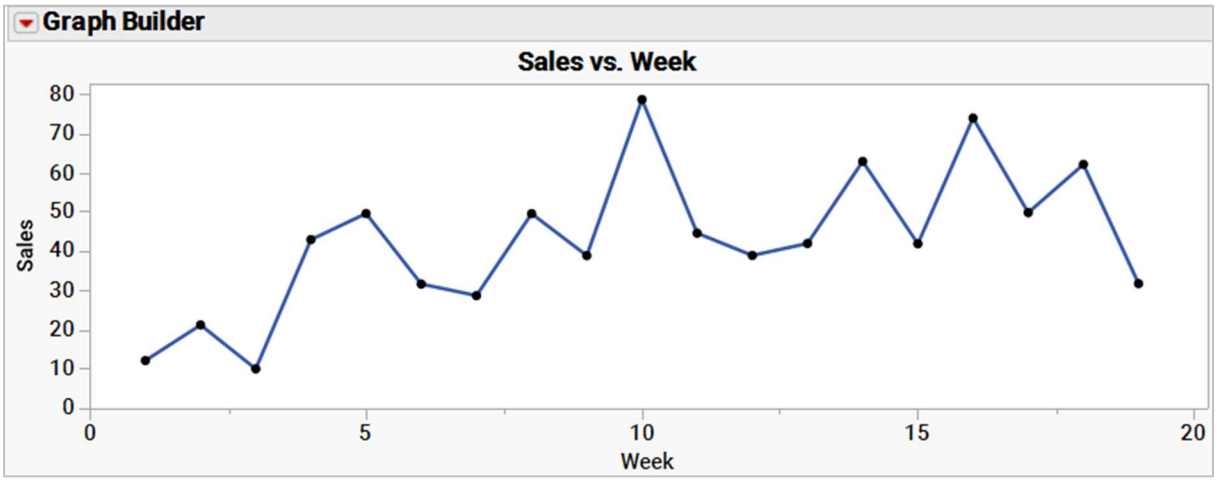
Exhibit 1 Distribution of Sales and Direct Mail



(Analyze > Distribution; Select Sales and Direct Mail as Y, Columns, and click OK. For a horizontal layout select Stack under the top red triangle.)

As shown in Exhibit 2, sales grew initially, but may be leveling off.

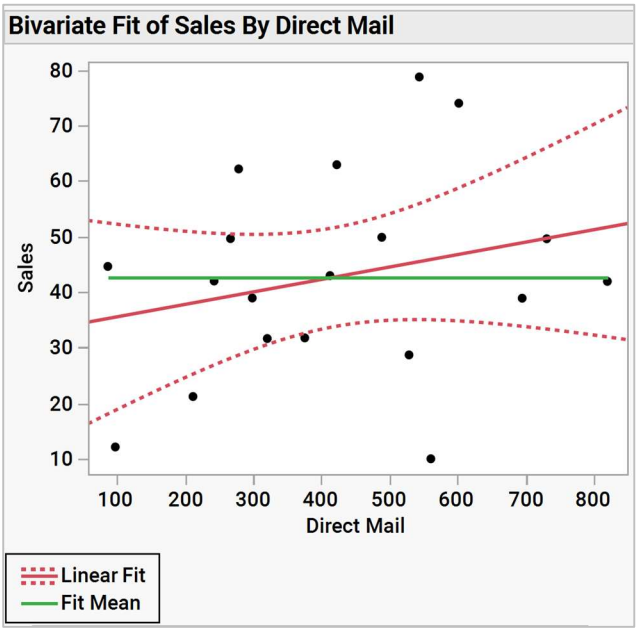
Exhibit 2 Time Series Plots of Sales



(Graph > Graph Builder; drag and drop Sales in Y, Week in X. Click the smoother icon to remove the smooth line. Then, hold the Shift key and click the line icon at the top. Or, right-click in the graph, and select Smoother > Change to > Line.)

To examine the relationship between direct mail expenditures and sales, we construct a regression model (Exhibit 3). We display the overall mean Sales (the horizontal line) and fit confidence bands for the predicted mean Sales (the curved dashed lines) to help us interpret the output.

Exhibit 3 Regression with Sales (Y) and Direct Mail (X)



Linear Fit

Sales = 33.300517 + 0.0224187*Direct Mail

Summary of Fit

RSquare	0.063413
RSquare Adj	0.00832
Root Mean Square Error	18.41087
Mean of Response	42.727
Observations (or Sum Wgts)	19

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	390.1460	390.146	1.1510
Error	17	5762.3223	338.960	Prob > F
C. Total	18	6152.4682		0.2983

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	33.300517	9.748879	3.42	0.0033*
Direct Mail	0.0224187	0.020896	1.07	0.2983

(Analyze >, Fit Y by X. Use Sales as Y, Response and Direct Mail as X, Factor. Under the red triangle select Fit Line. To explore the significance of the fitted line, select Fit Mean under the top red triangle, and select Confid Curves Fit under the red triangle next to Linear Fit to plot confidence bands.)

Exhibit 3 demonstrates that although 6.3% of the variability in sales is attributed to variation in direct mail costs (RSquare = 0.063), direct mail does *not* have a statistically significant influence on sales, since

- the p-value for Direct Mail (Prob > |t)| is 0.2983, and
- the mean line is contained within the 95% confidence bands, indicating that our model predicts no better than the overall mean Sales.

Our advice to this small business owner is to stop all direct mail activities immediately, since this advertising expense is not affecting the bottom line.

Wait!

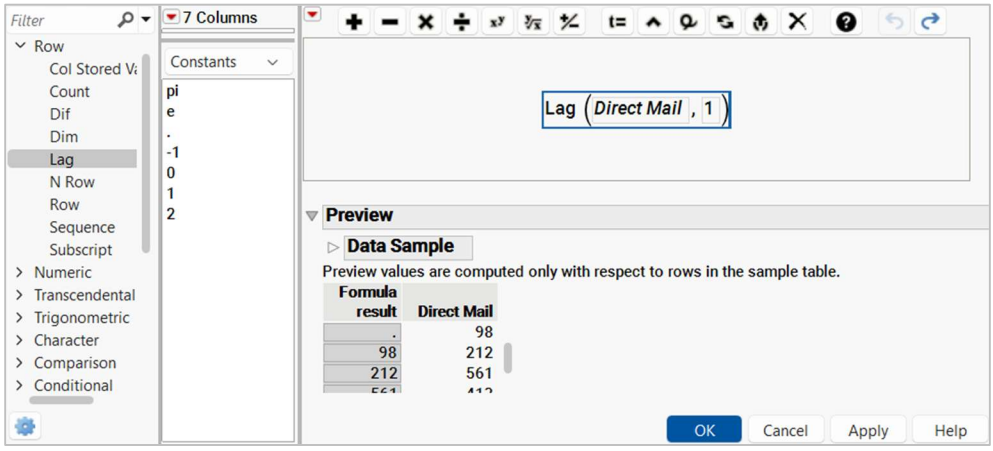
Look again, more closely, at the *time sequencing* of the relationship between sales and direct mail. According to the business owner, mailings are posted on Wednesdays of each week. With the lag in delivery time, most mail won't arrive until the weekend. Is it possible that customers don't respond to this week's advertisements until next week? If so, our current model doesn't capture this time lag. To see this, look at the partial listing of the data set shown in Exhibit 4.

Exhibit 4 Partial View of Data Set

	Week	Sales	Direct Mail
1	1	12.123	98
2	2	21.209	212
3	3	9.998	561
4	4	42.978	413

The data set as constructed shows us the relationship between direct mail in any given week and sales *in that same week*, since a row in the data set corresponds to each week's figures. Suppose, though, that the statistical relationship is one in which *this week's* sales is related to *last week's* direct mail activities. We would then need to rearrange the data set so that a row in the data set displays the current week's sales and the prior week's direct mail. Such a rearrangement involves *lagging a time-series variable*. (Exhibit 5).

Exhibit 5 Creating Direct Mail Lagged



(Create a new column, and rename the column Direct Mail Lagged. Right click on the column header, and select Formula to open the Formula Editor. To create the formula: 1. From Functions select Row > Lag , 2. Select Direct Mail from the columns list., 3. Type n = 1 > Click OK.)

The resulting variable is shown in Exhibit 6.

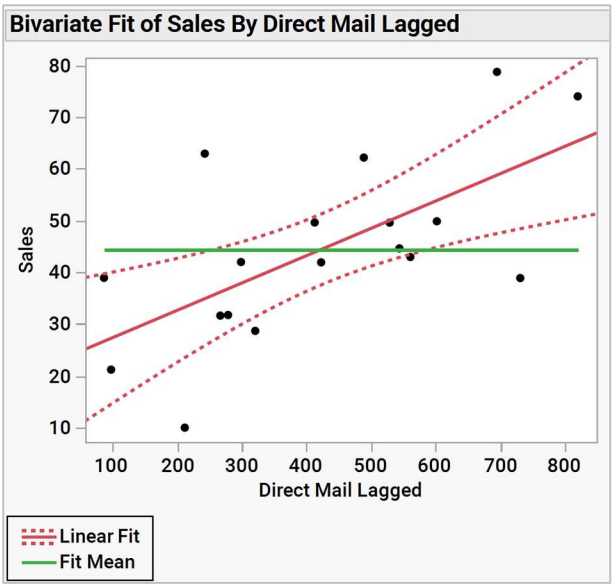
Exhibit 6 Data Table with Direct Mail Lagged

	Week	Sales	Direct Mail	Direct Mail Lagged
1	1	12.123	98	•
2	2	21.209	212	98
3	3	9.998	561	212
4	4	42.978	413	561

Lagging a time-series variable involves nothing more than moving it down one row. Note that the first observation (Week 1) for Direct Mail Lagged is missing, since there is no observation prior to the one at the beginning of the time-series.

In Exhibit 7 we see the regression results using last week's direct mail as the predictor variable.

Exhibit 7 Regression with Sales (Y) and Direct Mail Lagged (X)



Linear Fit

Sales = 22.089289 + 0.0528153*Direct Mail Lagged

Summary of Fit

RSquare	0.418199
RSquare Adj	0.381837
Root Mean Square Error	13.70292
Mean of Response	44.42722
Observations (or Sum Wgts)	18

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	2159.5093	2159.51	11.5008
Error	16	3004.3205	187.77	Prob > F
C. Total	17	5163.8298		0.0037*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	22.089289	7.336104	3.01	0.0083*
Direct Mail Lagged	0.0528153	0.015574	3.39	0.0037*

From Exhibit 7, we conclude that:

- There is a *statistically* significant relationship between last week's direct mail and this week's sales (the p-value is 0.0037, and the mean line has crossed the 95% confidence bands.)
- The model estimates that each additional dollar in direct mail costs is related to roughly a $53 increase in next week's sales. (Recall that direct mail is measured in dollars and sales in measured in $1,000.)
- However, only about 42% of the variability in sales is related to variability in prior week's direct mail activities (RSquare = 0.4182), and there is still a lot of unexplained variation in weekly sales figures (RMSE = 13.702, or $13,702).

Consider yourself lucky if you find that a response variable can be modeled with *lagged* predictors. In such instances, forecasting *future* values of the response variable is straightforward. For example, our regression equation here is:

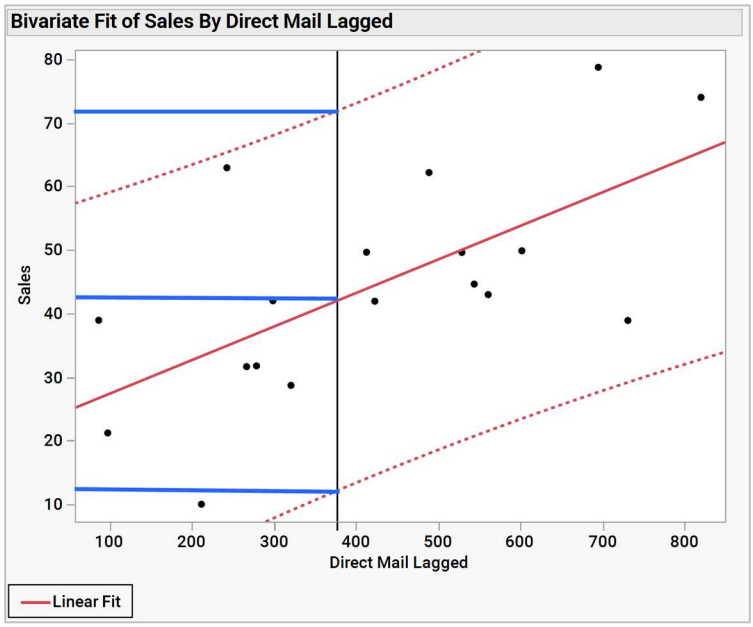
Sales = 22.09 + 0.053 Direct Mail Lagged

This equation says that to forecast sales in any one week, plug the value for direct mail expenditures from last week into the equation. Suppose that we are now in week 20 (recall that the data set ended in week 19). What is our forecast for sales in week 20? Since 376 is the value shown in the data set for direct mail in week 19:

Sales (in week 20) = 22.09 + 0.053 Direct Mail (in week 19)
= 22.09 + 0.053 (376)
= 42.018 (or $42,018)

We can use 95% prediction intervals to estimate the interval within which week 20 sales will fall. Prediction intervals are displayed as dashed lines in Exhibit 8. The prediction interval for Week 20 Sales is roughly 12 – 72 (or, $12K - $72K).

Exhibit 8 Prediction Intervals for Sales



(Select Confid Curves Indiv under the red triangle next to Linear Fit to plot prediction intervals. To add a vertical reference line double-click on the axis for Direct Mail Lagged, type the value for the reference line and click Add. Blue reference lines were drawn manually for illustration.)

Exhibit 9 Predicted Value and Prediction Interval for Sales

	Week	Sales	Direct Mail	Direct Mail Lagged	Predicted Sales	Lower 95% Indiv Sales	Upper 95% Indiv Sales
1	1	12.123	98	•	•	•	•
2	2	21.209	212	98	27.265187192	-4.449308272	58.979682656
3	3	9.998	561	212	33.286130239	2.6394316249	63.932828853
4	4	42.978	413	561	51.718666409	21.527729556	81.909603262
5	5	49.637	267	413	43.902003506	14.055296343	73.748710669
6	6	31.645	321	267	36.190971183	5.9052435009	66.476698864
7	7	28.698	529	321	39.042996837	9.0089150693	69.077078604
8	8	49.611	731	529	50.028577133	19.97898259	80.078171676
9	9	38.908	695	731	60.69726569	29.167018664	92.227512717
10	10	78.735	544	695	58.795915254	27.628735768	89.96309474
11	11	44.631	87	544	50.820806481	20.709490123	80.932122839
12	12	38.941	299	87	26.684219003	-5.154958511	58.523396517
13	13	42.004	243	299	37.881060459	7.7569364852	68.005184433
14	14	62.938	423	243	34.923404225	4.492957417	65.353851034
15	15	41.937	820	423	44.430156405	14.585254992	74.275057818
16	16	74.007	602	820	65.39782649	32.800912111	97.994740869
17	17	49.873	489	602	53.884093295	23.459362166	84.308824424
18	18	62.178	279	489	47.915965538	17.991491602	77.840439473
19	19	31.762	376	279	36.824754661	6.6038549057	67.045654417
20	•	•	•	376	41.94783778	12.062720393	71.832955168

(In the Bivariate Fit window, select Save Predicteds under the red Triangle for Linear Fit. JMP will create a new column with the prediction formula for Sales. Create a new row- since Direct Mail Lagged has a formula, the value for week 19 will auto-fill in the Direct Mail Lagged column and the predicted value for Sales will display. To save prediction intervals, select Indiv Confidence Limits Formula under the red triangle for linear fit. Or one cal also use Analyze > Fit Model; select Sales as Y and Direct Mail Lagged as a model effect, and hit Run. Under the red triangle select Save Columns > Indiv Confidence Intervals.)

In Exhibit 9 we see that our forecast of $42,000 for sales in week 20 is not very precise, since the prediction interval is $12,062 to $71,832. Although we have developed a statistically significant model, there is too much unexplained variation in sales and the current model's value as a forecasting tool is limited.

Can we build a better model with the existing data? We will explore this question in an exercise.

Summary

Statistical Insights

In addition to allowing you further practice in interpreting regression displays, this case reveals three important statistical issues:

1. Look for time-lagged relationships between variables in time-series regression models. Here, we discovered a *one-period* lagged relationship, although in other applications you might want to look for higher-order lags. For instance, a two-period lag would model the current response variable against the predictor variable lagged twice.
2. Knowing something about the organizational behavior (i.e., that direct mail, posted on Wednesdays, might have a lagged effect on sales) allowed us to discover an important statistical relationship in the data that might have otherwise been overlooked. Good statistical analysis rarely occurs in a vacuum.
3. Check your prediction intervals to assess how precisely your model predicts. Just because a model fits the data well doesn't mean that it forecasts well, or that the model is of much practical use. We might be able to do a better job had we been given additional predictor variables with which to forecast sales, a methodology called multiple regression (forthcoming). Alternatively, other univariate time-series methods, like ARIMA, might have done a better job forecasting these data.

Managerial Implications

The return on direct mail is very high: almost $53 to $1. Expect a time lag, though, between mailings and sales. Re-evaluate the effectiveness of direct mail in the next few months, since it's unlikely that such phenomenal returns on advertising will be experienced forever. The relationship between direct mail activities and sales will likely change over time as the store becomes an established business entity.

JMP Features and Hints

JMP will calculate and display both prediction intervals and confidence intervals. It's your job, as the analyst, to know which interval to apply.

1. Prediction Intervals (used for forecasts of individual values)
    - To graphically display the 95% prediction bands in Fit Y by X, use *Confid Curves Indiv.*
    - To save values for the 95% prediction intervals to the data table, use Fit Model. Save the columns using *Indiv Confidence Interval.*

2. Confidence Intervals (used for forecasts of average values of Y)
    - To graphically display the 95% confidence bands, use *Confid Curves Fit*.
    - To save values for 95% confidence intervals to the data table, use Fit Model. Save the columns using *Mean Confidence Interval.*

Remember: "Individual" for prediction intervals; "Mean" or "Fit" for confidence intervals.

Exercises

Determine whether a two-period lag for Direct Mail would model Sales better than the one-period lag model.

1. First, create a variable Direct Mail Lagged 2.
2. Then, fit a regression model for Sales and Direct Mail Lagged 2.
3. Compare the p-values, RSquare, RMSE, the predicted value, and the prediction interval.
4. Which, of the two, is the better model?
5. Why would it be unwise to consider higher-order lags, say, a 10-week lag for Direct Mail?