JMP Academic Case Study 019

Contributions

Simple Linear Regression and Time Series

Produced by

Marlene Smith, University of Colorado Denver Business School

jmp STATISTICAL DISCOVERY

Contributions[1]

Simple Linear Regression and Time Series

Background

The Colorado Combined Campaign solicits Colorado government employees' participation in a fund-raising drive. Funds raised by the campaign go to over 700 Colorado charities in all, including the Humane Society of Boulder Valley and the Denver Children's Advocacy Center. Prominent state employees, such as university presidents, chancellors and lieutenant governors, head the annual campaigns. An advisory committee determines whether the charities receiving contributions provide the services claimed in a fiscally responsible manner.

All Colorado state employees may contribute to the fund. However, certain state institutions are targeted to receive promotional brochures and campaign literature. Employees in these targeted groups are referred to as "eligible" employees. Each year, the number of eligible employees is known in June. Fund-raising activities are then conducted throughout the fall. By year's end, total contributions raised that year are tabulated.

The Task

It is now June 2010. The number of eligible employees for 2010 has been determined to be 53,455. Does knowing the number of eligible employees help predict 2010 year-end contributions?

The Data **contributions.jmp**

This is an annual time-series from 1988 – 2009. The variables are contribution Year and:

Actual	Total contributions to the campaign for the year in dollars
Employees	Number of eligible employees that year

Analysis

The average level of contributions during this time period was $1,143,769, with a typical fluctuation of $339,788 around the average. The average number of eligible employees was 45,419, with a typical fluctuation of 9,791.

Exhibit 1 Summary Statistics for Actual and Employees

Tabulate

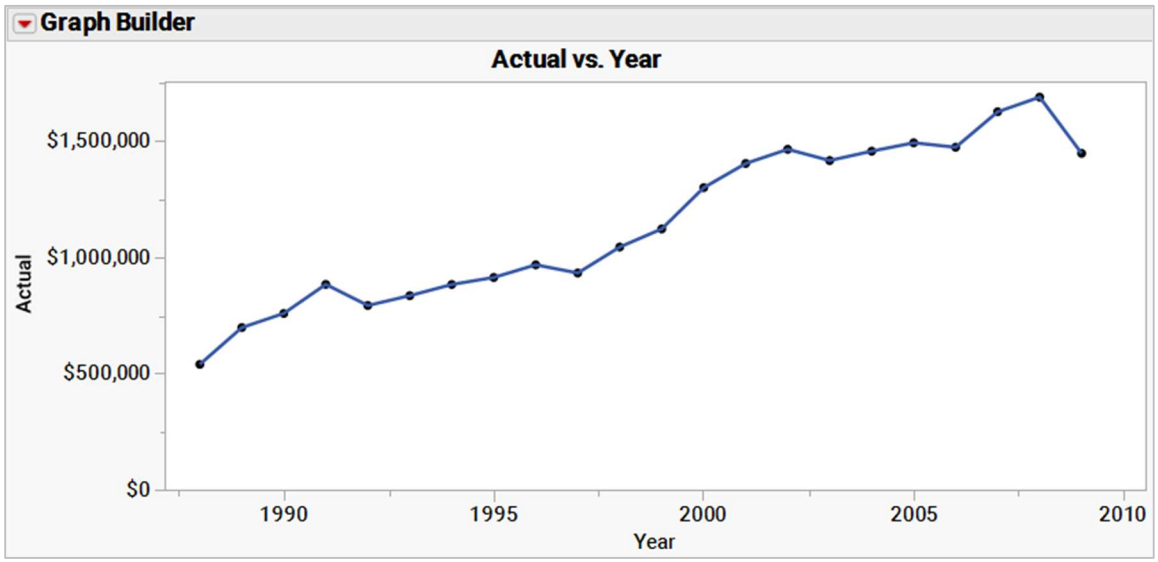
	Mean	Std Dev
Actual	1143769.45	339787.60
Employees	45768.35	9711.69

(Analyze > Tabulate; drag Mean and Std Dev from the middle panel to drop zone for columns, drag Actual and Employees in drop zone for rows as analysis columns.)

As we can see in Exhibit 2, contributions are growing over time:

[1]Mel Rael, Executive Director of the Colorado Combined Campaign, graciously provided these data.

Exhibit 2 Time Series Plot of Actual by Year

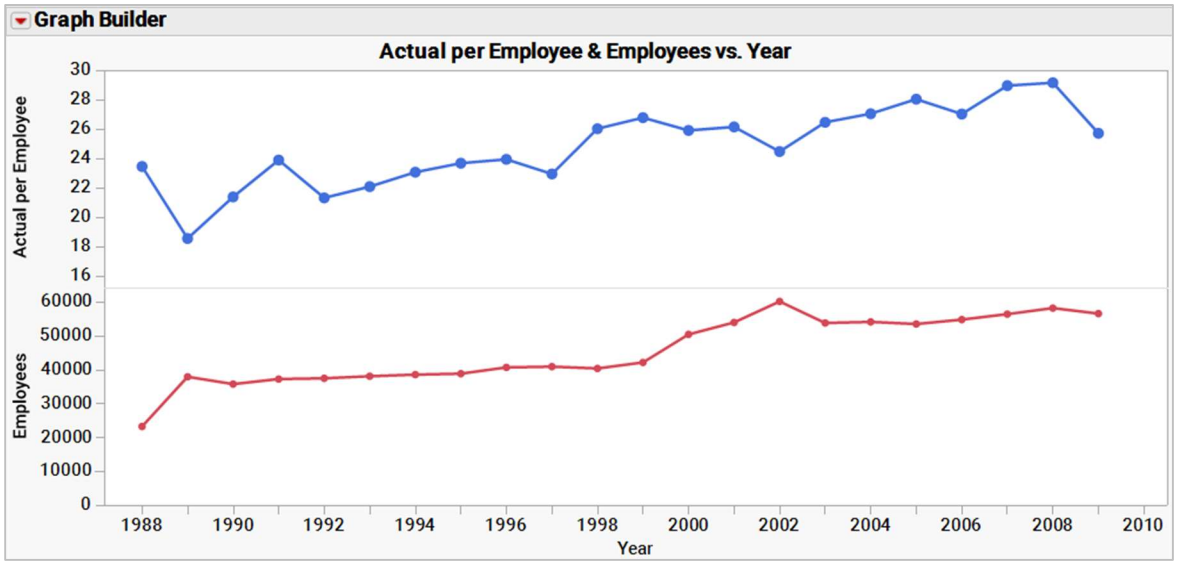


(Graph > Graph Builder; drag and drop Actual in Y and Year in X. Click on the smoother icon at the top to remove the smoother. Hold the shift key and click the line icon to add a line. Or, right click in the graph to select these options. Then, click Done.)

The long-term growth in contributions is attributable to two phenomena:

- The amount contributed *per eligible employee* is mostly upward (Exhibit 3, top).
- The number of eligible employees is on the rise, particularly in the 1999 to 2002 campaign years (Exhibit 3, bottom)

Exhibit 3 Time Series Plots of Actual per Employee and Employees

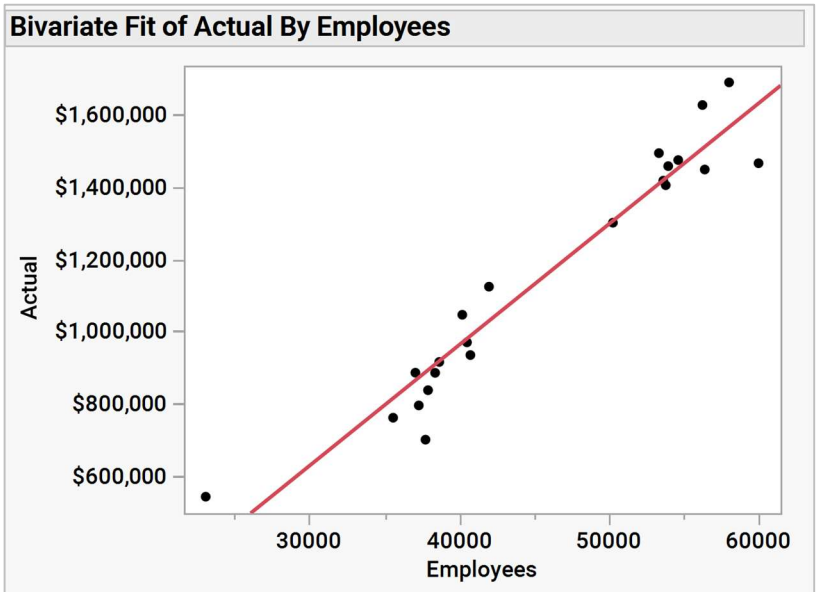


(Create a new column and rename it Actual per Employee, then use the Formula Editor to create the formula – Actual divided by Employees. Follow the instructions for Exhibit 2 to create the graph for Employees. Then, click and drag Actual per Employee above Employees in Y, and release. Right click on the X-axis and change the axis setting by making the bon increment to 1 and # minor ticks to 0. To change the markers for points, use the lasso from the toolbar to select the points (draw a circle around them). Then go to Rows > Markers and select a marker.)

The scatterplot and least squares regression line using Actual as the response variable and Employees as the predictor variable is shown in Exhibit 4. The formula for the regression line is found below the plot under Linear Fit. The slope of the fitted line, 33.555, estimates the contribution for each eligible employee over this time period. Hence, the model estimates an additional $33.56 in contributions for each eligible employee. Under Parameter Estimates, we see that the number of employees is a statistically significant predictor of year-end contributions; the p-value, listed as Prob > |t|, is < 0.0001.

The number of employees doesn't perfectly predict contributions. Just over 93% of the variability in contributions is associated with variability in number of eligible employees (RSquare = 0.934907). Comparing the standard deviation of Actual ($339,788) to the root mean square of the regression equation ((RMSE = $88,832) suggests that a substantial reduction in the variation in contributions occurs by using the regression model to explain variation in year-end contributions.

Exhibit 4 Regression with Actual (Y) and Employees (X)



Linear Fit

Actual = -380265.5 + 33.555042*Employees

Summary of Fit

RSquare	0.934907
RSquare Adj	0.931653
Root Mean Square Error	88831.66
Mean of Response	1143769
Observations (or Sum Wgts)	22

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-380265.5	91893.82	-4.14	0.0005*
Employees	33.555042	1.979813	16.95	<.0001*

(Analyze > Fit Y by X. Use Actual as Y, Response and Employees as X, Factor. Under the red triangle select Fit Line. Note: To remove the markers in Exhibit 3, go to the Rows menu and select Clear Row States.)

We've been informed that the number of eligible employees in 2010 is 53,455. To use the regression equation to forecast 2010 year-end contributions, we can plug this number into the regression equation.

If the number of Employees is 53,455, the predicted Actual contributions is:

Actual = -380265.5 + (33.555042) x Employees
= -380265.5 + (33.555042) x (53,455)
= 1413419.3 (or, $1,413,419)

In words, given that the number of eligible employees is 53,455, our model estimates that 2010 year-end contributions will be approximately $1.413 million.

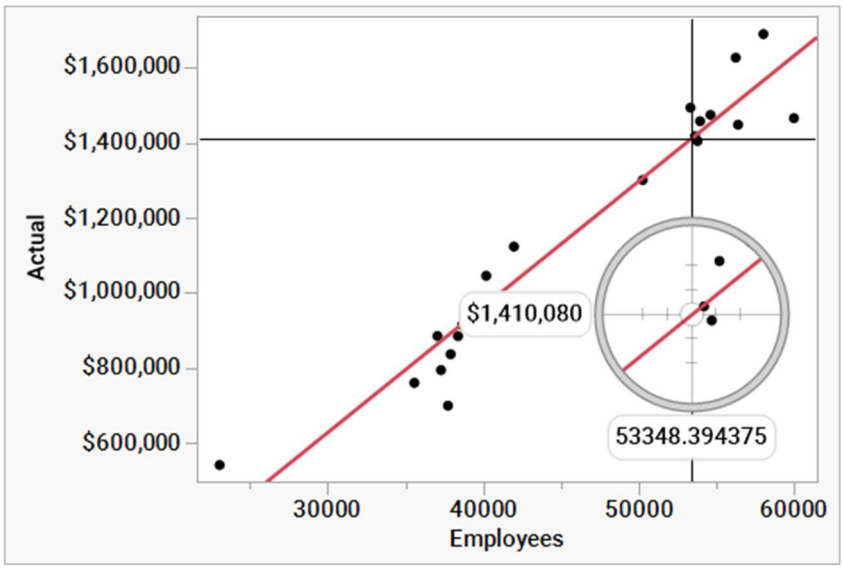
Easier still, we can skip the math exercise, save the regression formula and prediction intervals and ask JMP to calculate the estimated contributions for 2010 (Exhibit 5). Prediction intervals are useful, since the number of employees isn't a perfect predictor of contributions. The prediction interval gives us an estimate of the interval in which the 2010 year-end contributions will fall (with 95% confidence).

Exhibit 5 Predicted Value and Prediction Interval for 2010 Contribution

	Year	Actual	Employees	Actual per Employee	Predicted Actual	Lower 95% Indiv Actual	Upper 95% Indiv Actual
19	2006	$1,474,452	54605	27.002142661	1452007.5937	1258782.705	1645232.4824
20	2007	$1,627,071	56224	28.939082954	1506333.2063	1311685.1521	1700981.2606
21	2008	$1,689,947	58000	29.137017241	1565926.9606	1369467.7071	1762386.214
22	2009	$1,448,248	56377	25.688631889	1511467.1277	1316673.2489	1706261.0066
23	.	.	53455	.	1413419.2956	1221070.475	1605768.1163

(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 Actual. Also select Indiv Confidence Limit Formula to create upper and lower 95% for each individual prediction)

Exhibit 6 Using Cross-hair Tool to Explore Predicted Contribution

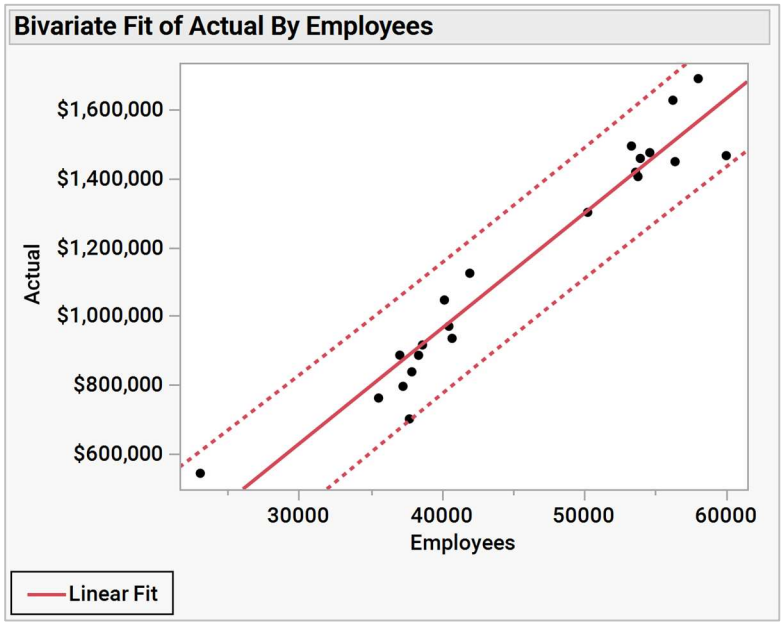


(Select the cross-hair tool on the toolbar. Click on the regression line at the value of the predictor to see the predicted response value.)

Predicted values can also be explored dynamically using the cross-hair tool. In Exhibit 6, we see that the predicted value for Actual, if Employees is 53,348, is around $1.410 million.

We can also graphically explore prediction intervals (Exhibit 7).

Exhibit 7 Prediction Intervals for Actual



(In the Bivariate Fit window, select Confid Curves Indiv under the red Triangle next to Linear Fit. Use the cross-hairs to find the upper and lower bounds for the prediction interval.)

Summary

Statistical Insights

Forecasting using regression involves substituting known or hypothetical values for X into the regression equation and solving for Y. In this case, values for the predictor variable in the forecasting horizon are known in advance; i.e., we know that the 2010 value for Employees is 53,455, so we plugged this value into the regression equation to forecast year-end contributions. In another setting, in which the same-year value for X is unknown, how would we proceed? One possibility is to forecast the value of the predictor variable. Another possibility, when theoretically and statistically justified, is to use lagged values of the original predictor variables in the regression model.

When building any regression model, residuals should be checked to ensure that the linear fit makes sense.

Managerial Implications

Regression has provided a prediction for year-end 2010 Colorado Combined Campaign contributions of $1.4M. In managerial settings such as this, where the response variable represents a business goal, managers often set *higher* expectations than the predicated value to motivate improved performance. One such choice here might be the upper 95% prediction limit of $1.6M.

This forecasting methodology can be repeated year after year. Once the final contributions to 2010 are known, they can be added to the data set and the regression line can be recalculated. By midyear of 2011, the number of eligible employees will be known.

Note that, in this case, we focused on trend analysis using only Year as the predictor. We could also fit a model with both Employee and Year. We will consider regression models with more than one predictor in a future case.

JMP Features and Hints

In this case we used Fit Y by X to develop a regression model. We used cross-hairs tool to explore the predicted value of the response at a given value of the predictor. Several options, such as saving predicted values and showing prediction intervals, are available under the red triangle for the fitted line. When the prediction formula is saved, a new column with the regression formula is created. Enter the value of X in a new row in the JMP data table, and the predicted value will display. To save prediction intervals to the data table for the value of X, use Fit Model.

Note that other intervals and model diagnostics are also available from both Fit Y by X and Fit Model. To generate residual plots from within Fit Y by X, select the option under the red triangle next to Linear Fit.

Exercises

A *regression trend analysis* uses only the information contained in the passage of time to predict a response variable.

1. Perform a trend analysis with the Colorado Combined Campaign data, using Actual as the response variable and Year as the predictor.

2. Forecast the 2010 - 2013 Colorado Combined Campaign contributions.

3. Compare your forecast for 2010 with that obtained from the simple linear regression model in which number of eligible employees is the predictor variable. Hint: Compare RMSE, RSquare, and the estimated contributions for 2010. Which model does a better job of explaining variation in contributions?
4. We've limited our analyses to one predictor variable at a time. Guestimate what would happen, in terms of RMSE, RSquare and model predictions if we were to build a model with both Year and Employees.