



INSIDE THIS NEWSLETTER

Informative Missing Regression Coding in JMP® 111

Identifying Quality Issues and Misconduct Using Analyses of Digit Preference.....3

Fast Flexible Filling in Space Filling Designs.....5

When Responses Are Below the Limit of Detection.....8

Generalized Regression in JMP PRO 11.....12

Join the JMP User Community.....14

Designing Insightful Process Behavior Charts15

JMP Books From SAS Press19

What's New in JMP Training20

Don't Miss a Single Issue

Visit jmp.com/jmpercable to read current or back issues or to subscribe to the print version.



Informative Missing Regression Coding in JMP® 11

John Sall, Co-Founder and Executive Vice President, SAS



What if I told you that by adding a very simple feature, you could fit many models more accurately than before? At the same time, I could show you that your

previous answers were very biased, but the new ones were much less biased. Furthermore, rather than just fitting the nonmissing data, you could use all of the data, making predictions even when regressors are missing.

This big opportunity comes when the missingness of the data is predictive. Consider the JMP sample data table *Equity.jmp*, which contains home mortgage default data. The response is whether the customer defaulted on the loan. One of the regressors is **DEBTINC**, which is the ratio of debt to income. **DEBTINC** is missing for much of the data. Are the missing values predictive of whether the customer defaulted on the loan?

If I fit the usual model with all 12 of the other variables as regressors, the RSquare is .23, but this is on only 3,364 of the 5,960 rows. I can't predict the response when any of the regressors is missing.

Suppose I fit the model using the missingness of **DEBTINC** as an indicator variable. The RSquare in this

one-regressor model is better (.25), and I use all 5,960 rows to predict regardless of whether the data is missing. So missingness can be predictive, informative, and even outperform regular regressors.

The new idea is to create not only the missing value indicator variables but also make the original regressors useful without discarding rows of missing data.

In JMP Pro 11, the Fit Model launch window includes a new red triangle menu item called "Informative Missing." Although the feature is only in JMP Pro, you can, with some effort, achieve the same goal by adding formula columns to the data table, as described in this article.

A traditional way to do this is with imputation. You estimate the missing values. One standard imputation method is to predict values for a missing variable using all the other variables in the model that are nonmissing. When there are missing values in the data, the Multivariate platform supports imputation with the "Impute Missing Data" option. And there are other imputation approaches.

Informative Missing regression coding is much simpler than imputation and also more powerful. For every continuous variable in your model that has missing values, you substitute two variables. The first variable substitutes a mean

value for missing data in the column. The other variable is a missing value indicator – 1 if the original column is missing, 0 otherwise.

For example, in Equity.jmp, I create two new variables with formulas for the continuous regressor called **DEBTINC**. Here is the script for creating two new variables with formulas:

```
New Column( "DEBTINC Or Mean if Missing", Numeric, Continuous,
    Formula( If( Is Missing(
        :DEBTINC ), 33.78, :DEBTINC ) )
);

New Column( "DEBTINC Is Missing", Numeric, Continuous,
    Formula( If( Is Missing(
        :DEBTINC ), 1, 0 ) )
);
```

“33.78” in the formula is the mean of **DEBTINC**, though it can be any value you like.

Now instead of using **DEBTINC** in the model, you use the two new formula variables **DEBTINC Or Mean if Missing** and **DEBTINC Is Missing**.

The extra predictor – the indicator variable – can be strongly predictive of the response. In this example, the act of leaving the field missing is a strong clue about the risk of a loan to the applicant.

With these two new predictors, you have two variables that are never missing, so they can be used in all the data. The missing value indicator variable might be important in the fit, thus giving you a chance to improve the fit.

What about the other new regressor, **DEBTINC Or Mean if Missing**? It appears to be a primitive way to impute **DEBTINC** using just the mean. Replacing missing values with the mean does not affect the estimate of that variable. Rather, the parameter for the missing value indicator estimates the difference between the prediction and the mean value of that regressor.

You could substitute zero instead of the mean for missing **DEBTINC**. The parameter would then estimate the difference between the prediction and having a zero for that covariate. The plug-in value for

missing only affects the interpretation of the indicator parameter estimate.

This technique is not imputation. You do not substitute any predicted values for the missing values. Instead, you construct a coding system to recover information when the regressors are missing. Furthermore, you use missing values for their predictive value. It is assuming that having a missing value there is not some random event, but is for some cause that might have predictive value. That is why we call this coding technique “Informative Missing” regression coding.


Table 1 shows the dramatic differences between using imputation and Informative Missing for the Equity.jmp data.

Table 1. Comparing Imputation to Informative Missing

Method	Number of Observations Used	Rsquare
Old full model with row-wise missing exclusion	3,364	23.2%
One-variable model with DEBTINC Is Missing	5,960	25.6%
New full model with Informative Missing	5,960	45.6%

Figure 1 shows that missing value indicator variables are very significant, with **DEBTINC Is Missing** being more significant than any other regressor in the model.

As demonstrated here, you can create formula columns (or transforms) to regress with. With JMP Pro 11, you just have to select the “Informative Missing” red triangle menu option in Fit Model. Another option is downloading the “Informative Missing Coding” add-in from the [JMP File Exchange](#).

Missingness can be your friend if you treat it as being valuable. 

Parameter Estimates				
Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
DEBTINC Is Missing	-2.7829502	0.0976715	811.85	<.0001*
DELINQ Or Mean if Missing	-0.800191	0.0526308	231.16	<.0001*
Intercept	5.10298614	0.3720997	188.07	<.0001*
DEBTINC Or Mean if Missing	-0.0941618	0.0087834	114.93	<.0001*
VALUE Is Missing	-5.1082246	0.5357863	90.90	<.0001*
CLAGE Or Mean if Missing	0.00597567	0.0006806	77.08	<.0001*
DEROG Or Mean if Missing	-0.5219103	0.0625866	69.54	<.0001*
DEROG Is Missing	2.15892182	0.2988389	52.19	<.0001*
JOB[]	2.0441136	0.3603335	32.18	<.0001*
NINQ Or Mean if Missing	-0.145317	0.0261112	30.97	<.0001*
JOB[Sales]	-1.3772548	0.253729	29.46	<.0001*
CLNO Is Missing	-3.2591677	0.6339095	26.43	<.0001*
JOB[Office]	0.4906712	0.1405409	12.19	0.0005*
CLAGE Is Missing	-1.1434237	0.3437362	11.07	0.0009*
VALUE Or Mean if Missing	-3.9722e-6	1.2484e-6	10.12	0.0015*
YOJ Is Missing	0.58629045	0.1971033	8.85	0.0029*
JOB[Other]	-0.2945218	0.1043564	7.97	0.0048*
DELINQ Is Missing	1.16234203	0.4217969	7.59	0.0059*
MORTDUE Is Missing	-0.5600108	0.2054465	7.43	0.0064*
CLNO Or Mean if Missing	0.01305841	0.005329	6.00	0.0143*
YOJ Or Mean if Missing	0.01628617	0.0068979	5.57	0.0182*
MORTDUE Or Mean if Missing	3.60369e-6	1.7318e-6	4.33	0.0374*
LOAN	7.94464e-6	4.8327e-6	2.70	0.1002
JOB[Mgr]	-0.1984142	0.1340223	2.19	0.1388
JOB[ProfExe]	-0.0595748	0.1249524	0.23	0.6335
NINQ Is Missing	0.17622532	0.381414	0.21	0.6441
REASON[]	0.07006288	0.2078167	0.11	0.7360
REASON[DebtCon]	0.0254227	0.1144628	0.05	0.8242

Figure 1 Parameter estimates sorted by significance

Identifying Quality Issues and Misconduct Using Analyses of Digit Preference

Richard Zink, PhD, Principal Research Statistician Developer, JMP Life Sciences

The new Digit Preference option in JMP® Clinical 5.0 enables you to uncover quality issues and misconduct in clinical trials.

International guidelines recommend that clinical trial data should be actively reviewed or monitored to ensure data quality. The traditional interpretation of this guidance for pharmaceutical trials has led to extensive on-site monitoring¹. This on-site review is time-consuming and expensive (up to a third of the cost of a clinical trial). As is true for any manual effort, the on-site review is limited in scope and prone to error.

In contrast, risk-based monitoring (RBM) uses a central computerized review of clinical trial data and site metrics to determine whether sites should receive a more extensive quality review². Companies are interested in RBM for

The primary purpose of JMP Clinical is to simplify data discovery, analysis and reporting for clinical trials. With its straightforward user interface and main reliance on graphical summaries of results, everyone from the clinical trial team can explore data to identify safety or quality concerns.

reducing costs and improving operational efficiency. Regulatory agencies and sponsors use RBM to protect patient well-being and study integrity. For example, the FDA recently urged trial sponsors to proactively outline how data will be reviewed and to define thresholds for risk. The guidelines also recommend that trial sponsors document processes for preventing error and resolving issues³. In short, quality is built into the trial so that it becomes less of a reactionary brute-force process.

Most clinical trials enroll patients at numerous investigator sites. A study with multiple centers allows for greater subject diversity to support the general-

izability of findings to the larger patient population. Patients can be enrolled more quickly at multiple centers, which promotes a shorter clinical trial. Although there are numerous logistical challenges to managing multiple centers, analysts can use these independent data sources to help identify quality issues and misconduct within the clinical trial. One such example involves analyzing digit preference in the data collected during clinic visits (for example, in blood tests). This analysis is available from the **Clinical > Fraud Detection > Digit Preference** menu in JMP Clinical.

I can screen for unusual site-test combinations by treating each center as the “suspect site” in turn and comparing the distribution of the trailing digit (that is, the last digit) to that of all other sites combined. I use the Cochran-Mantel-Haenszel (CMH) row mean score statistic to take advantage of the ordinality of digit for greater power, and apply standardized midrank scores to account for the possibility that the observed digits might not be equally spaced from one another⁴. In general, i sites with j procedures performed will result in up to $i \times j$ comparisons. Initially, I can summarize all tests using a volcano plot to highlight the combinations of greater interest: tests with large numerical differences and/or tests that meet the criteria for statistical significance after applying a suitable multiplicity adjustment, such as the FDR method⁵.

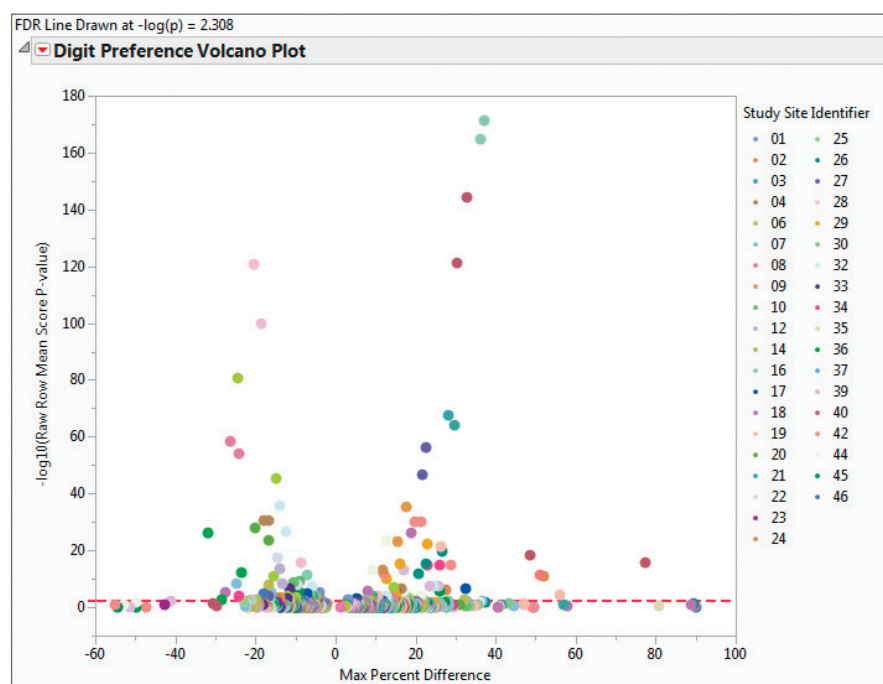


Figure 1 Digit preference volcano plot

In Figure 1, the x-axis represents the maximum difference between the suspect site versus the reference (all other sites) across all observed digits to represent the Max Percent Difference. The y-axis represents the raw p -value from the CMH row mean score test on the negative log10 scale, so the smaller the p -value, the larger the value vertically. In general, the interesting site-test combinations approach the upper corners and above the dashed-red reference line, which indicates markers that are significant accounting for multiple comparisons.

Further analysis of the top two blue markers provides the digit bar chart in Figure 2. Notice the investigator(s) at Site 16 were twice as likely to report a “0” in the trailing digit for diastolic blood pressure; a similar result (not shown) is available for systolic blood pressure. What could this result mean? This might be an example of a site that is not following the study protocol: The investigator might be collecting blood pressure manually and reading values from a gauge instead of using a machine to obtain the measurements. Alternatively, this investigator might have a tendency, compared to the other sites, to round measurements down. Further investigation would be required to understand the reason behind such a finding and whether the difference is substantial or important enough to intervene.

Digit preference analyses can also be used to detect instances where diagnostic equipment might be miscalibrated. Other important differences in subjective measurements can be identified, such as the investigator’s assessment of clinical signs using a Likert scale, which might suggest that additional training is needed.

Though I presented this analysis in the context of a clinical trial, it can be

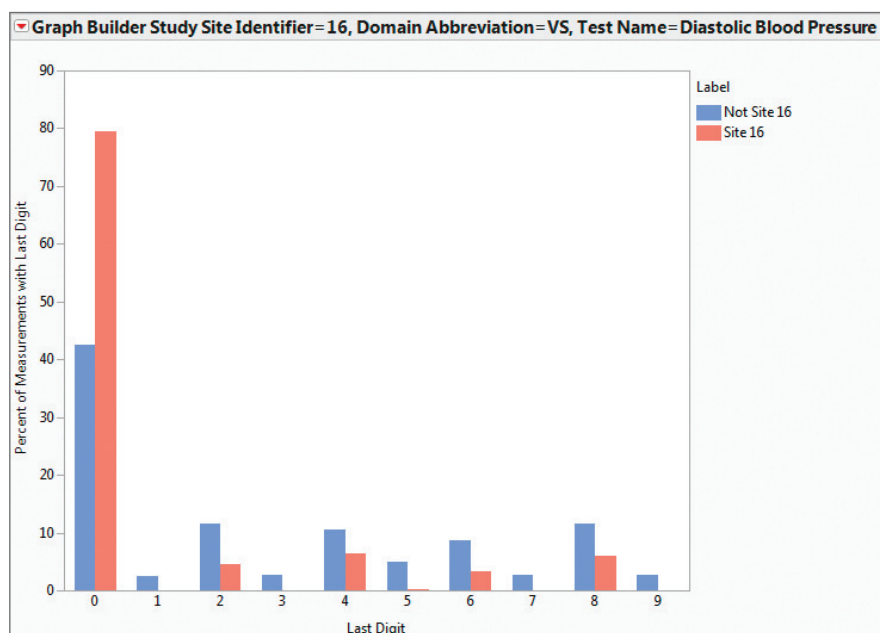



Figure 2 Digit bar chart for Diastolic Blood Pressure (site 16)

applied in any situation where multiple readers or machines are used to collect or report data. However, such analyses might identify natural differences between these independent sources. In the preceding example, the analyses could highlight differences in patient population or variations in technique between the sites that wouldn’t necessarily indicate a quality concern. These differences should still be considered in the subsequent analysis.

Although this analysis examined the trailing digit, similar analyses could be applied to the leading digit, either through the observed proportions or by applying the significant-digit or Benford’s law⁶. Further, similar analytical approaches (a volcano plot to screen and a drill-down for details) can identify issues in visit scheduling, noticeable differences in baseline characteristics, or the presence of a treatment effect prior to dosing. 

REFERENCES

1. International Conference of Harmonisation. (1996). E6: Guideline for Good Clinical Practice. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6_R1/Step4/E6_R1_Guideline.pdf
2. TransCelerate BioPharma Inc. (2013). Position paper: Risk-based monitoring methodology. Available at <http://transceleratebiopharmainc.com/>.
3. US Food and Drug Administration (2013). Guidance for industry: Oversight of clinical investigations – a risk-based approach to monitoring. Available at <http://www.fda.gov/downloads/Drugs/.../Guidances/UCM269919.pdf>.
4. Stokes, M.E., Davis, C.S., Koch, G.G. (2012). *Categorical Data Analysis Using SAS®, Third Edition*. Cary, NC: SAS Institute Inc.
5. Zink, R.C., Wolfinger, R.D., and Mann, G. (2013). Summarizing the incidence of adverse events using volcano plots and time windows. *Clinical Trials* 10: 398-406.
6. Hill, T.P. A statistical derivation of the significant-digit law, *Statistical Science*, 10, 354-363 (1996).

Fast Flexible Filling in Space Filling Designs

Bradley Jones, PhD, JMP Principal Research Fellow, SAS

Ryan Lekivetz, JMP Research Statistician Developer, SAS

A new tool in the Space-Filling Design platform supports placing design points in non-rectangular regions. Space-filling designs are very popular in experimentation with complex deterministic computer models. Such models give the same answer if you supply the same inputs, so controlling variability is not an issue.

Experimentation might take hours or even days to produce a single observation. The goal of experimentation with computer models is to find a much faster, faithful approximation to the computer model. You can then use this approximation to make predictions at untried points. Or, you can develop intuition about the relationships between the inputs and outputs of the computer model by making graphs using the simplified approximation model.

Building such graphs using the computer code itself takes too much time because of the huge number of points required to make such a plot. We call the new designs *Fast Flexible Filling* designs: they generate quickly, provide great flexibility in describing the region of experimentation, and do a good job of filling the space.

Example of a Space-Filling Design

Here is an example. Suppose that you were studying air quality over Georgia. You would likely be interested in covering the entire state in a uniform way. You have a limited budget of 100 air quality monitors. The design question is where to locate the monitors. Figure 1 shows the 100 point Fast Flexible Filling (FFF) design for this problem.

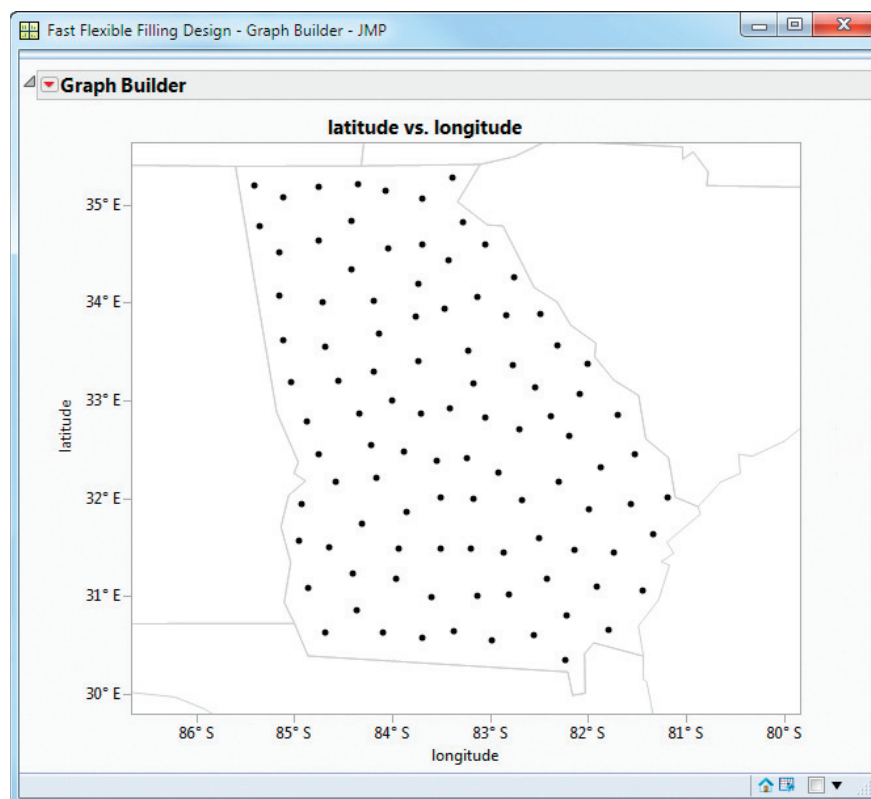


Figure 1 Uniform spacing of air quality monitors for Georgia

The points do a good job of covering the state in a uniform fashion.

Suppose that you want to run a script to repeat this experiment. Maps and fast flexible filling designs are scriptable in JMP. The following sections describe how to get the map coordinates for Georgia and create the space-filling design.

Get the Map Shape Coordinates

JMP installs a set of data tables that contain geographical map data. Each map consists of two JMP data tables named with a common prefix:

- The `-Name.jmp` data table contains the unique names for the different regions.

- The `-XY.jmp` data table contains the latitude and longitude coordinates for the regions.

Map files are installed in Maps directory for your version of JMP. On Windows, look in `C:/Program Files/SAS/JMP/11/Maps` (or in the `JMPPro/11/Maps` folder). On Macintosh, look in `/Library/Application Support/JMP/11/Maps`.

See [Create Maps](#) from the [Essential Graphing book](#) in the Help menu or [online documentation](#) for more details about the built-in map files.

1. Open `US-State-Name.jmp` from the Maps folder. Look for Georgia and notice the shape ID is 11.

- Open US-State-XY.jmp. Scroll down to row 903 and copy all rows for shape ID 11 into a new data table.
- Run the following script to get the coordinates in a format that the Space Filling Design platform can use.

```
dt = Current Data Table();
mymap = dt << Get as Matrix( {X,
Y} );
xx = mymap[0,1];
yy = mymap[0,2];
Show(min(xx), max(xx));
Show(min(yy), max(yy));
Show(xx, yy);
```

The minimum and maximum values of Georgia coordinates are written to the log, an abbreviated portion of which is shown in Figure 2.

Set Up the Fast Flexible Filling Design

Now that you have the map shape coordinates, you can create the space-filling design.

- In JMP, select **DOE > Space Filling Design**. Two factors are defined by default.
- Enter the minimum and maximum coordinates from the log into the Space Filling Designer.

Rename **X1** to **longitude**. Change the minimum and maximum values to **-86** and **-80**.

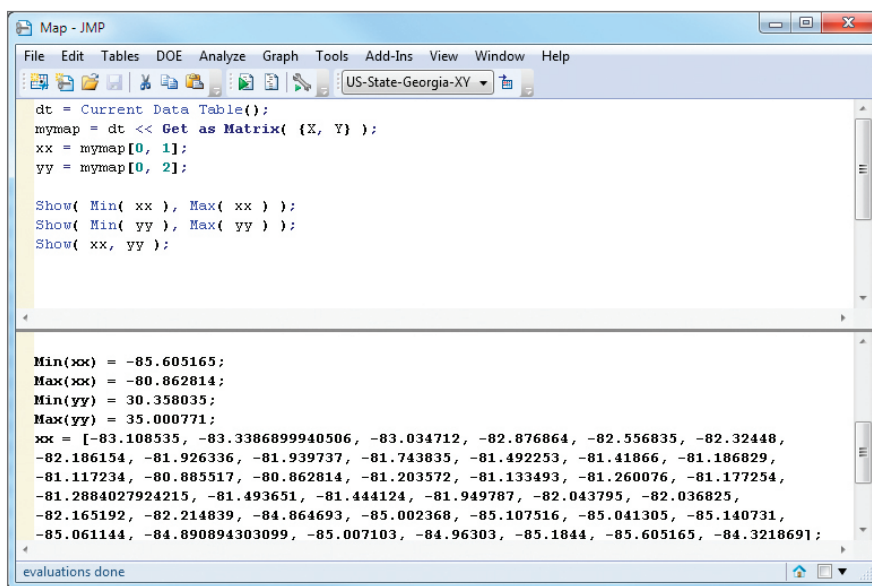


Figure 2 Coordinates for North Carolina shown in the log

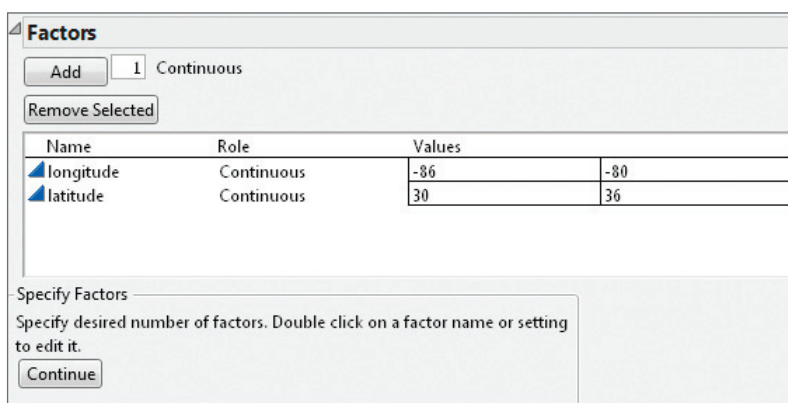


Figure 3 Factor table for the space filling design

Rename **X2** to **latitude**. Change the minimum and maximum values to **30** and **36**.

that your map shape only makes up a small fraction of it. The factor table appears in Figure 3.

NOTE: For other designs, make sure you don't define too big of a range so

Specify the Disallowed Combinations for Coordinates

The Space Filling Designer's Fast Flexible Filling algorithm uniformly distributes points within the map shape. To keep the points within the unconstrained design region, you define an expression that rules out infeasible input variable combinations. The expression evaluates whether a point in the space-filling design falls outside of the map shape.

This Disallowed Combinations expression must return a nonzero value for infeasible values. One way to do this is to create a Boolean expression that evaluates to true or false. Suppose that factors X1 and X2 range from -1 to 1 and want points to occur within the unit circle. You specify that disallowed combinations are $X1^2 + X2^2 > 1$. That is, disallow any points that fall outside of the unit circle.

1. Copy the xx and yy variables from the log window.
2. Select **Disallowed Combinations** from the Space Filling Design red triangle menu. The Disallowed Combinations window appears – the same window used in the Custom Designer.
3. Paste the xx and yy variables into the Disallowed Combinations window.
4. Enter the following code at the end of the window:

```
!In Polygon(longitude,
latitude,xx,yy)
```

This expression disallows any points outside of the map shape.

5. Click **OK**.
6. In the Space Filling Design window, click **Continue**.

Fast Flexible Filling is the only design available because of the disallowed combinations.

7. Change the number of runs to 100 to better visualize the results.
8. Click **Fast Flexible Filling** to create the design.
The design appears in a new window.
9. Click **Make Table** to create a data table that contains the Fast Flexible Filling design.

Viewing the Design in Graph Builder

Now you can view the design in Graph Builder to verify the design.

1. With the Fast Flexible Filling Design data table open, select **Graph > Graph Builder**.
2. Assign **longitude** to X and **latitude** to Y.
3. Right-click the graph, select **Graph > Background Map > US States**, and then click **OK**.

You should get something similar to the graph shown in Figure 1.

Final Thoughts

If you're dealing with maps that come in separate parts (such as states that have multiple Part ID numbers), we recommend creating FFF designs for each part ID individually.

Over a map, an FFF design is different from taking a simple random sample of points over the map shape. An FFF design is intended to make a more even spread of the design points, making for better coverage over the map. ✨

When Responses Are Below the Limit of Detection

Mark Bailey, Principal Analytical Training Consultant, Education and Training, SAS

To learn about a system or a process, there must be variation. If the characteristics or outcomes never change, then it is impossible to learn anything. We design experiments to provoke a large change in the response in the hope that the analysis will be more informative, both in kind (factor effects) and degree (precision). The determination of the response requires a measurement that is accurate (unbiased) and precise over a useful range. Many physical quantities are bounded by zero, and all measurements are limited by noise. When the response is absent or zero, the background signal can be translated in various ways into an upper bound on measurement, or *limit of detection* (LOD).¹ What value should you use in your analysis for a response reported to be below the limit of detection?

There are many intuitive practices for selecting the value for analysis when the response is below this threshold. Some analysts use zero. Other analysts use the LOD itself. Still others split the difference and use half of the LOD. Finally, some analysts regard such a case as indeterminate and leave the value missing.

These *ad hoc* approaches unfortunately do not address the central problem but instead introduce bias in any estimates, such as model parameters. A missing value reduces the sample size and, therefore, the power of the analysis, as if nothing is known about the response when, in fact, there is information available. Using zero biases your estimates downward; using the LOD biases your estimates upward. Using half the LOD might average out the bias, if you are

optimistic and tend to be lucky. Isn't there a better way? Isn't there a rigorous approach based on statistical theory that eliminates this bias and enables you to use all of the data?

The solution is found in an unrelated field of study that has nothing to do with chemistry or any other physical science. Investigators encountered the same problem at the start of *survival analysis*. In this analysis, the response is the *life-time* or the *time-to-event* where the event is death or the onset of disease. Subjects often survive or never incur the disease during the study period. What to do with their data? Ignoring it or using an arbitrary value would introduce bias as described above.

Analysts realized that two types of data existed in these studies: for one kind, the actual lifetime is known, and for the other kind, it is a lower bound on the actual lifetime. The second kind is called *censored data*. These lifetimes are *right-censored* because the actual lifetime is greater than the observation. In the same way, the responses that are below the LOD are called *left-censored* data.

Ordinary least squares regression is not able to analyze censored responses,

but *maximum likelihood estimation* accommodates censoring directly through the likelihood function.

JMP® provides a *parametric analysis* of survival models with multiple factors, which suits the case of our experiment. The normal distribution is not available for the likelihood function in the Parametric Survival platform, but the log-normal distribution is available. We merely transform the response by exponentiating the response as e^x for the analysis and then transform back using the natural logarithm after fitting the model for prediction. The following fictitious example illustrates the points above and shows how to use JMP for such an analysis.

Limit of Detection Example

This example illustrates censoring with responses below the LOD. The experiment includes four continuous factors, **X1-X4**. The response **Y** is simulated without censoring, and then an arbitrary LOD is applied. Perhaps Y is the level of a chemical impurity that you intend to minimize through judicious selection of factor levels. An arbitrary LOD (10) was selected to cause a few responses to become censored. A thorough study of this matter would involve many

X1	X2	X3	X4	Y	Left Censored Y	Right Censored Y
-1	1	1	1	23.186312846	11740530594	11740530594
-1	1	-1	0	9.3194701554	•	22026.465795
0	0	0	1	35.766932329	3.414926e+15	3.414926e+15
0	1	1	0	15.518144848	5488386.3693	5488386.3693
-1	-1	0	1	39.538689198	1.484002e+17	1.484002e+17

Figure 1 Original data (left) and new censored data columns (right)

¹ Detection Limit, http://en.wikipedia.org/wiki/Detection_limit.

simulated data sets. The single simulated set of responses here is presented only to illustrate the problem and how to deal with it.

The setup for this problem in JMP is simple. The original columns for the experiment are in the left red box in Figure 1. Three columns were added to facilitate the analysis as seen in the right box in the figure. We use *interval censoring* for this analysis. That is, we specify a lower and upper bound for each response in two new columns, here called **Left Censored Y** and **Right Censored Y**, respectively. These values are the same as the exponentiated actual response when it is above the LOD as seen in the first row. The left-censored value is missing and the right-censored value is the exponentiated LOD for censored responses as seen in the second row.

We will examine the bias caused by using one of the *ad hoc* corrections before examining the results of using the correct analysis.

The data were fit to a second order polynomial function with all two-factor interaction terms using ordinary least squares regression. The value in column **Y** was the simulated response levels before imposing the LOD. This regression represents the best we could do if there was no limit of detection. It is our benchmark for comparison with different ways of handling a LOD.

The OLS regression analysis is repeated with three different versions of the response **Y**.

1. The value in column **Y2** was set to 0 if the simulated response was below the LOD.
2. The value in column **Y3** was set to halfway between 0 and LOD (5 in this case) if the simulated response was below the LOD.
3. The value in column **Y4** was set to the LOD (10 in this case) if the simulated response was below the LOD.

The prediction formulas from all four OLS regressions are saved to the data table. The estimates for the parameters that are active in the simulated response are compiled from each of these four regressions along with the true value

from the simulation. Figure 2 shows the results in Graph Builder.

Notice that the estimates are close to the true value when there is no LOD (**Y**). On the other hand, the application of one of the three *ad hoc* methods when the response is below the LOD results in estimates that are not as close to the true value. (Note that this single example is not sufficient for proof. It is intended only to illustrate the problem.)

Now try the parametric survival model. In Fit Model, change the fitting Personality from Standard Least Squares to Parametric Survival, select the Lognormal for the Distribution, and

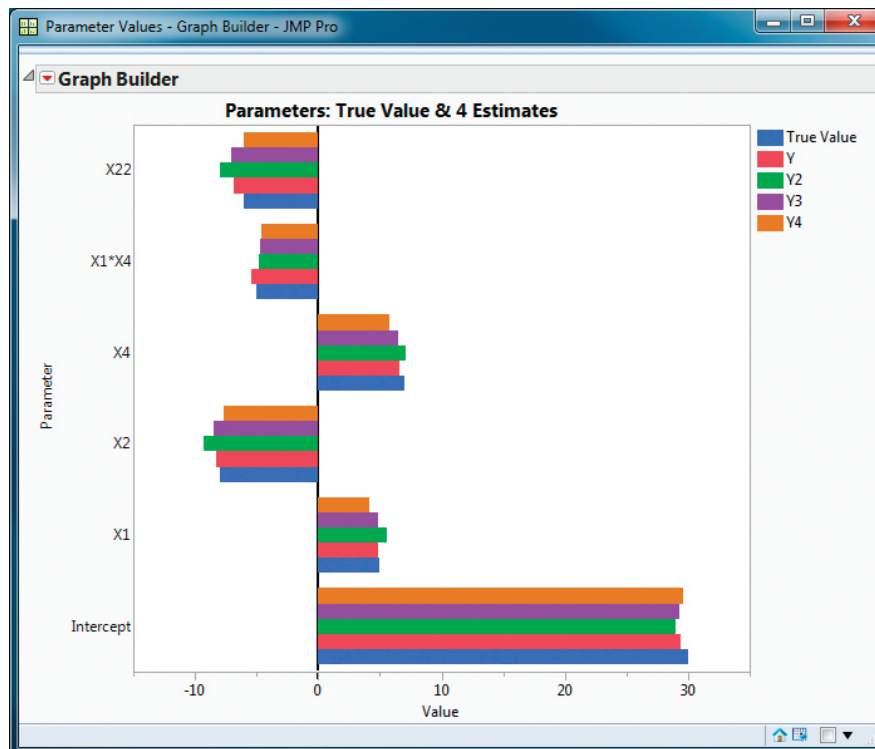


Figure 2 Actual versus predicted parameter estimates

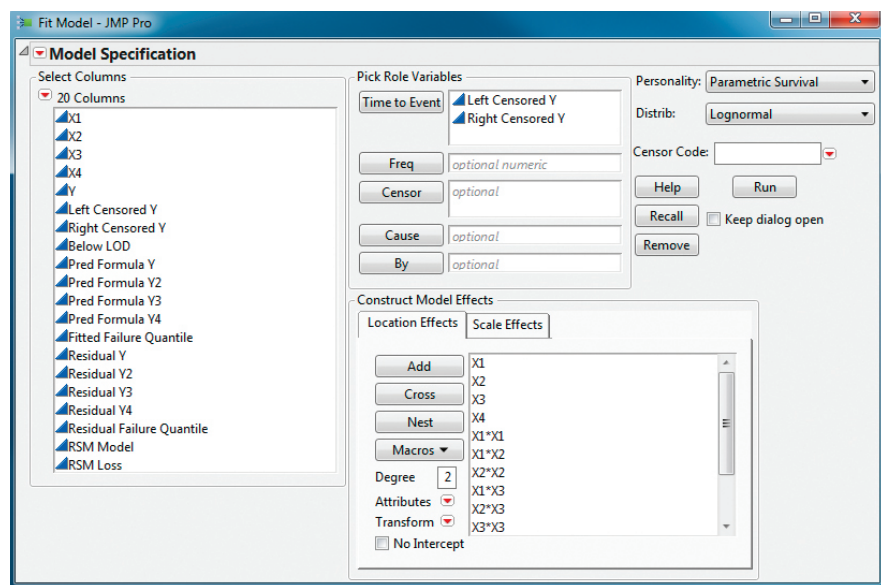


Figure 3 Fit Model specification for the Parametric Survival model

assign the **Left Censored Y** and **Right Censored Y** to the Time to Event role. The terms in this model (Location Effects) are the same as the terms used in the OLS regression above. There are no terms for the Scale Effects. This situation treats the variance as a constant, independent of factor levels.

The results must be translated as follows. The linear model determines the location parameter of the log-normal distribution and sigma estimates the standard deviation of the same distribution. Save the prediction formula for the *time quantile*. Click the red triangle at the top of the report, select **Save Quantile Function**, and then enter **0.5** for the probability. A new column called **Fitted Failure Quantile** is added to the data table. You can rename this column to reflect the original response.

The last step is to edit the prediction formula to transform back to the original

response. This step is easy. When you open the formula editor for the new column, the entire formula is already selected. Simply select the **Transcendental** group of functions, select **Log**, and then save the new formula.

Figure 4 shows the first five rows of the updated data table, which includes the original response **Y** and the new **Fitted Failure Quantile** formula column.

The four new **Y** columns were created by saving the prediction formula for the four OLS regressions and the parametric survival regression. Notice

that the **Fitted Failure Quantile** prediction is much closer to that from **Pred Formula Y**, the first OLS regression (our benchmark).

Examine the correspondence between the observed response and the predicted response from all of the models so far. In Figure 5, a graph of these predicted responses overlays a scatterplot. The identity line ($Y=X$) was added for reference.

The markers closest to the $Y=X$ line are those from the first OLS regression (no LOD imposed) and the parametric

Pred Formula Y	Pred Formula Y2	Pred Formula Y3	Pred Formula Y4	Fitted Failure Quantile	Y
23.120913319	23.684590797	23.628355238	23.57211968	23.139741208	23.1863128
9.7803716958	2.3401707614	7.2157265386	12.091282316	9.8467086644	9.31947015
36.490068108	37.676755566	37.111051607	36.545347647	36.483442238	35.7669323
14.717950223	13.689135167	14.052597165	14.416059163	14.707552447	15.5181448
40.119295847	40.883040338	40.828308183	40.773576027	40.131839786	39.5386891

Figure 4 Data table with new prediction columns

survival regression. The discrepancy becomes worse as the response approaches the LOD.

Perhaps a better way to see the distinctions between these different approaches is with a residual plot (Figure 6).

The residuals closest to 0 are those from the first OLS regression when no LOD was imposed (Residual Y) and the parametric survival regression (Residual Failure Quantile). The large residuals from the models using *ad hoc* adjustments to the response (increased bias) indicate that these models are less valuable if you need predictions near the LOD, as in the case of optimizing factor levels to reduce an impurity.

In conclusion, I have demonstrated that *ad hoc* methods for using a response below the detection limit result in biased parameter estimates and model predictions. On the other hand, using interval censoring with a parametric survival model avoids these problems. The survival model is easy to fit. The transformations allow the log-normal distribution to easily model the errors. You can use the prediction formula with tools such as the Prediction Profiler to find optimal settings. Further note that you can use the same approach when the limiting response is an upper bound by substituting a missing value for the Right Censored Y value. 🌟

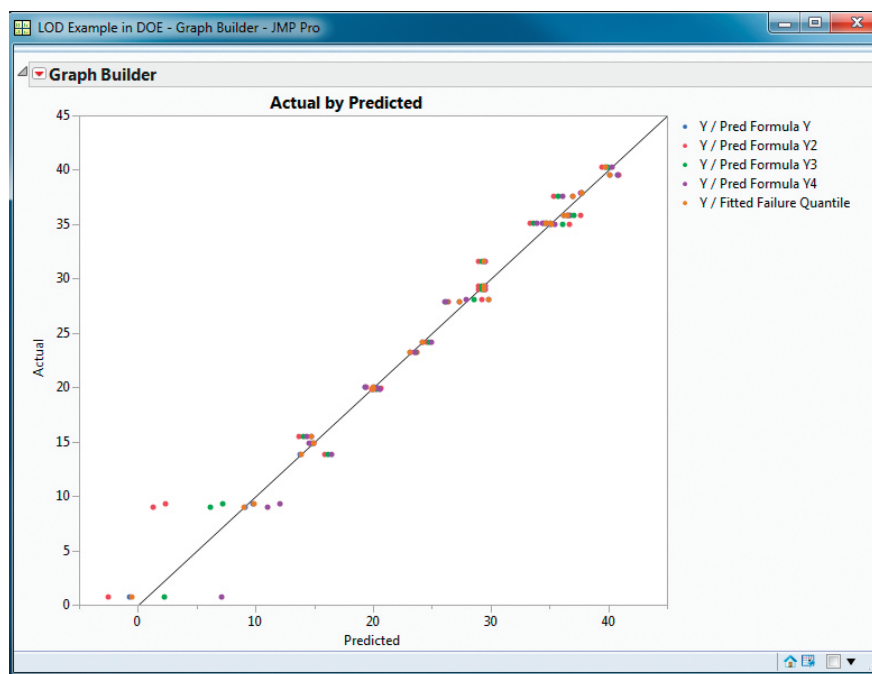


Figure 5 Graph of actual by predicted responses

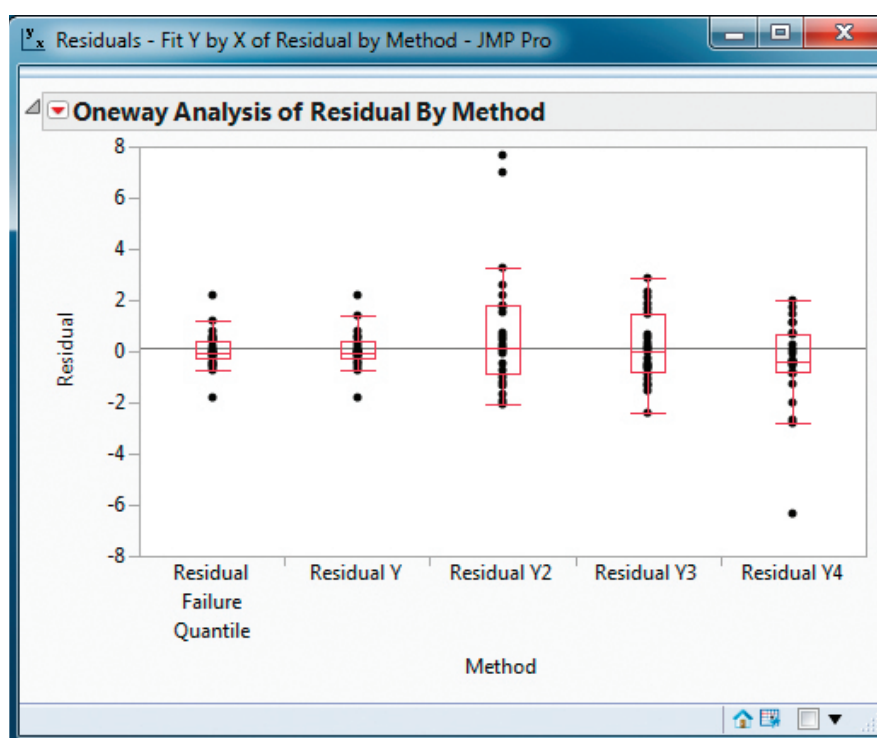


Figure 6 Residual plot of responses by method

Generalized Regression in JMP® Pro 11

Clay Barker, JMP Senior Research Statistician Developer, SAS

The Generalized Regression platform (new in JMP Pro 11) is designed for fitting penalized generalized linear models. That means that we can build models where the response is not normally distributed (for example, in logistic and Poisson regression models). It also means that we have the option to use the lasso, ridge regression, and the elastic net – three popular penalized regression techniques. Ignoring the generalized linear model piece of this platform for now, this article introduces some ideas about penalized regression. An example of using the lasso for variable selection is also included.

Why penalize?

Maximum likelihood is one of the workhorses of statistics and is a popular way to estimate the parameters in a regression model. The likelihood function tells us the probability of the observed responses for a given set of parameters. So the maximum likelihood estimator gives us the regression parameters that maximize the probability of the observed data. Likelihood theory also provides us theory for making inferences about our regression parameters. These are excellent qualities, but there is a catch: Maximum likelihood is most appropriate when you know in advance which predictors belong in your model. If you have an abundance of predictors and don't know which ones to include in your model, maximum likelihood might not be the best choice for you.

If we use a penalized likelihood instead, we can simultaneously do variable selection and parameter estimation. This article focuses on one such penalized

regression technique: the least absolute shrinkage and selection operator, better known as the lasso. Lasso estimates are obtained by minimizing

$$-likelihood(\beta) + \lambda \sum_j |\beta_j|$$

where β is a vector of regression coefficients and λ is a positive valued tuning parameter.

Minimizing this penalized likelihood makes sense intuitively: We want a model that fits our data well, but we don't want our model to get too complex. When the tuning parameter is zero, we get the usual maximum likelihood estimate. As we increase the tuning parameter, more and more coefficients are shrunk to zero. Because the lasso solution depends on the tuning parameter, we try a range of values of λ and keep the best (based on cross-validation, for example).

How does the lasso penalty force some of the coefficients to zero? To understand how the penalty works, it is helpful to think of the most basic case: simple linear regression. If we center the response (to eliminate the need for an intercept), the lasso problem looks like

$$\hat{\beta}_{Lasso} = \arg \min_{\beta} \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda |\beta|$$

which is equivalent to

$$\hat{\beta}_{Lasso} = \arg \min_{\beta} (\beta - \hat{\beta}_{ML})^2 + \lambda |\beta|$$

where $\hat{\beta}_{ML}$ is the maximum likelihood estimate. Writing the minimization problem in this way allows us to solve for the lasso

estimator as a function of the maximum likelihood estimate

$$\hat{\beta}_{Lasso} = \begin{cases} \hat{\beta}_{ML} + \lambda & \hat{\beta}_{ML} < -\lambda \\ 0 & -\lambda \leq \hat{\beta}_{ML} \leq \lambda \\ \hat{\beta}_{ML} - \lambda & \hat{\beta}_{ML} > \lambda \end{cases}$$

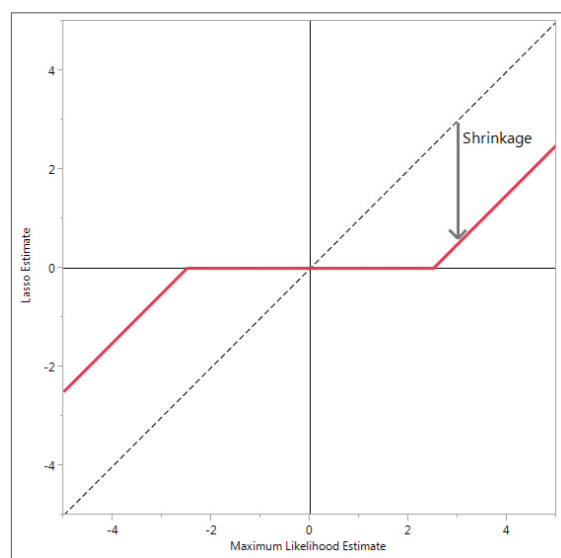


Figure 1 Comparing the maximum likelihood and lasso estimates

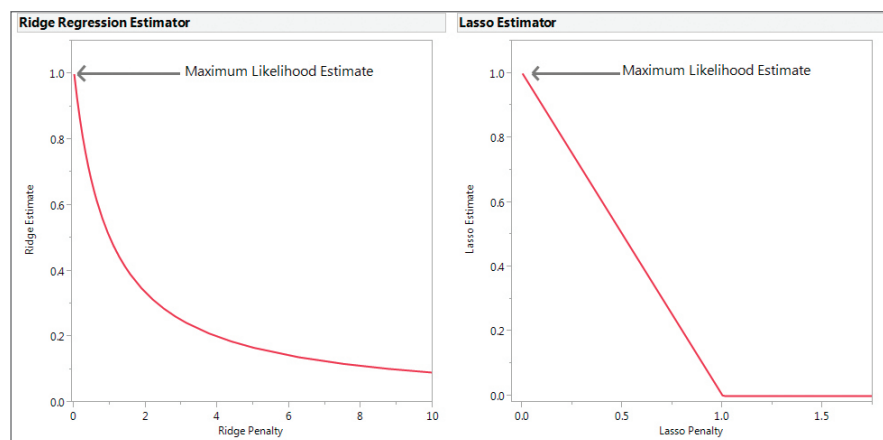


Figure 2 Comparing the ridge regression and lasso estimates

Now we can see that the lasso estimator is shrunk toward zero and can even take the value zero. Figure 1 shows an example of the relationship between the maximum likelihood estimate (dashed diagonal line) and the lasso estimate (red line).

It is also useful to look at how the lasso estimate changes as a function of the

tuning parameter. Figure 2 shows an example of the lasso estimate in contrast to the ridge regression estimate. Ridge regression is another penalized regression technique. However, it cannot do variable selection because the estimate never reaches zero. By shrinking some estimates to zero, the lasso performs variable selection and helps us avoid overfitting our data.

An Example

In this example, we use the Generalized Regression platform with the Hollywood Movies data from the JMP sample data folder. These data show the profitability of a sample of movies from 2011. Our goal is to learn about how profitability is associated with features such as genre, production budget and audience rating. The model specification (available from the Fit Model window) is in Figure 3.

Movie Name is an effect so that we can identify which movies are substantially more profitable than others. If a movie appears in our final model, we know that particular movie is substantially more (or less) profitable than otherwise expected.

Adding **Movie Name** as an effect allows us to screen and adjust for outliers. However, this produces more than 170 predictors, which is more than the number of rows in our data table. Luckily, the lasso performs variable selection, so we no longer have to worry about having more predictors than observations.

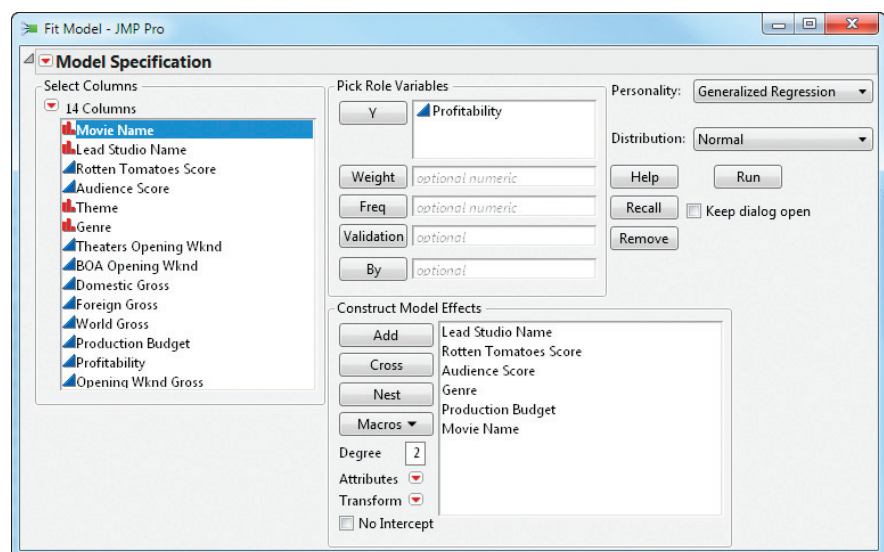


Figure 3 Fit Model specification

Figure 4 shows a portion of the report. The solution path plot in the left portion of the figure is crucial for understanding how variables enter the model. As we decrease the lasso penalty (moving from left to right in the plot), more and more parameters become nonzero. The red vertical line denotes the best fit based on the Bayesian information criterion (BIC). The line corresponds to the parameter estimates found in the right portion of Figure 4. Some of the results are not too surprising. Positive reviews

(from both audiences and movie critics) tend to be associated with better profitability. It might be surprising to see that dramas tend to be less profitable. A handful of movies exceeded expectations. For example, *Insidious* and *Bad Teacher* were far more profitable than expected. Notably absent from our final model is Production Budget. This suggests that you don't necessarily have to invest a fortune to make a profitable movie.

Conclusion

The Generalized Regression platform is a powerful tool for doing variable selection for generalized linear models. This article has hopefully shed some light on why we do penalized regression and how the lasso penalty works. By shrinking coefficients all the way to zero, the lasso performs variable selection, leaving us with models that are simpler and that should predict well for new observations. 🌟

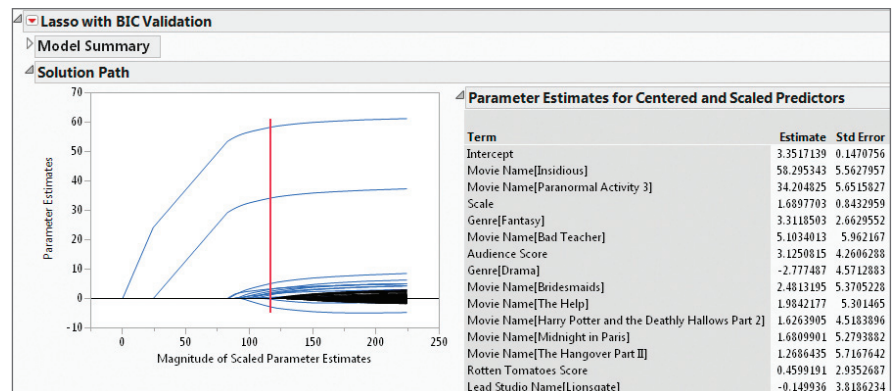


Figure 4 Lasso report

Join the JMP User Community



The JMP User Community is the largest online community of JMP users. It's the home of the File Exchange, where you can share and download JMP add-ins, scripts and sample data files. Contributed files are searchable and tagged with keywords that tell you how you can use them.

Submit a file today, and your contribution may become the featured file on the site.

The user community also includes the Discussion Forum, where you can get helpful answers to JMP questions and offer your own expertise. Earn points for taking part and be recognized in the community for your knowledge. The user community also keeps you up to date on the JMP Blog, links you to instructional webcasts and training, and connects you with JMP experts and users worldwide, all from one place. Join – or check back in – today!

community.jmp.com

About JMPer Cable

A Technical Publication for JMP® Users

Issue 29 Summer 2014

Editor: Sheila Loring

Designer: Diana Witt

Production: Melody Fountain

Contributors: Mark Bailey, Clay Barker, Bradley Jones, Ryan Lekivetz, Jose Ramirez, John Sall, Annie D. Zangi, Richard Zink

Don't Miss a Single Issue

Visit jmp.com/jmpcable to read current and back issues or to subscribe to the print version.

For additional user resources, go to community.jmp.com.

To order additional licenses of JMP, to learn about the most recent release, or to inquire about content of this JMPer Cable issue, call 1-877-594-6567.

Designing Insightful Process Behavior Charts

José G. Ramírez, PhD

Annie D. Zangi, JMP Research Statistician Developer, SAS

The Control Chart Builder platform was introduced in JMP® 10 as an easy and fun way to design process behavior charts. This becomes really useful in situations where different subgroups are possible, depending on how the data is organized. In this article we show how Control Chart Builder can be used to investigate different chart designs, helping select the one that reveals hidden differences in the data.

Injection Molding Example

Four hundred observations were collected in an injection molding operation that manufactures ball joint sockets, as shown in Figure 1 below.

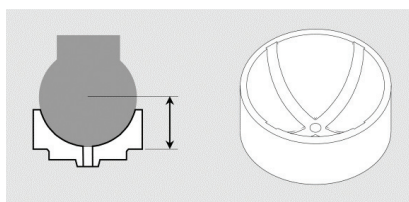


Figure 1 Scatterplot with exploratory fits

Figure 1 was adapted from Figure 3.9a of Ramírez and Ramírez (2010)¹.

The response of interest is the effective thickness of the ball joint socket (in 100th mm) in excess of 12 mm. Because this is a new process, management wants to know whether the process is ready for production. In other words, is the process producing ball joint sockets in a predictable way? We can use a process behavior chart to determine whether the process is in control or predictable.

It is easy to generate a process behavior chart. For example, in order to generate an individual and moving range chart we

select **Analyze > Quality and Process > Control Chart > IR** and assign **Thickness** to **Process**. Figure 2 does not show any points outside the control limits, but it does not look quite in “control.” Is the information displayed in this chart enough for us to qualify the process?

In this case, the individual and moving range charts do not reflect the structure of data, that is, the different sources of variation present in the data. So, no, this chart does not give us enough information to qualify the process.

Process Behavior Chart Design

A process behavior chart needs to be designed so that it can answer the questions of interest. What are these questions? They depend on which type of chart we are using. Observations in a process behavior chart represent “rational” subgroups of like things.

For example, the XBar and Range (or R)

charts plot the range of subgroup values on the R chart to monitor variation. They plot the average of the subgroup values on the XBar chart to monitor location. In their book, *Understanding Statistical Process Control*, Wheeler and Chambers clearly present the questions that the XBar and R process behavior charts pose.

- The R chart asks the question: “Making allowance for the average amount of variation within the subgroups, are the within subgroups differences **consistent**?”
- The XBar chart asks the question: “Making allowance for the amount of variation within the subgroups, are there **detectable differences** between the subgroup averages?”

As you can see, each chart answers a different question in terms of consistency within subgroups, and the ability to detect changes between

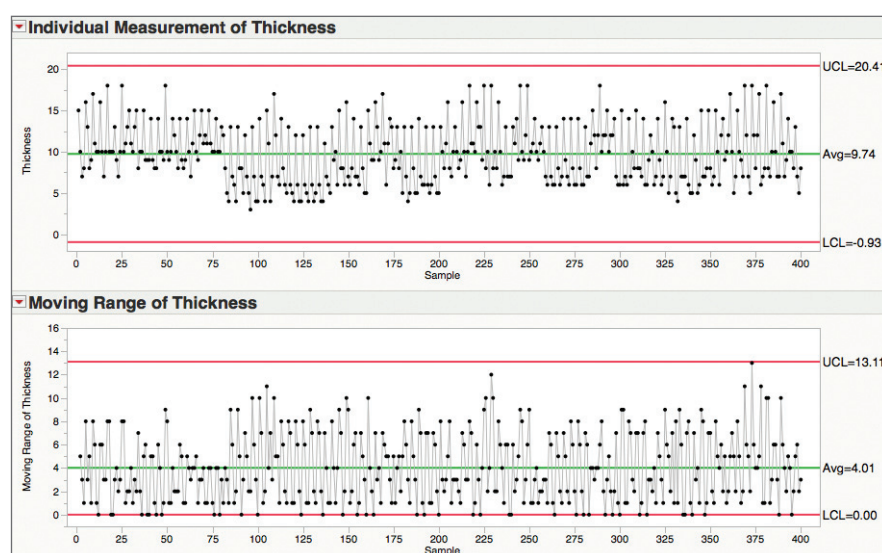


Figure 2 Individual Measurement and Moving Range chart of Thickness

subgroups. As we show below, the answers to these questions depend on how the data is organized into rational subgroups.

Rational Subgroups

Rational subgroups define the process of organizing the data into groups of like things that reflect the context and sources of variation present in the data. Let us revisit the data in Figure 2. What is the context of the ball joint socket data? The injection molding process produces sockets four at a time that come from a mold that has four cavities. To qualify the process, data was collected four times a day for five days. At each of those four periods during a day, five cycles of the press were performed, giving 20 parts per hour. This gave a total of 400 readings, as shown in Figure 2.

There are sources of variation in this data, as shown in Table 1.

Table 1. Sources of variation for ball joint socket data

Source of Variation
Hour-to-Hour
Cycle-to-Cycle
Cavity-to-Cavity

The design of the process behavior chart requires us to think about the allocation of these sources of variation to the rational subgroups, and the questions that the charts answer. The different allocations can be easily explored by means of Control Chart Builder.

Control Chart Builder

Control Chart Builder works much like Graph Builder with a drag-and-drop interface. It displays Individual and Moving Range charts the instant that you drag a variable onto the y-axis. When you drag a subgroup column

onto the x-axis, it switches to an XBar and R chart. Let us look at three different organizations of the ball joint socket data, according to how the sources of variation in Table 1 are allocated to the subgroups.

First Organization of the Data

The first organization of the thickness data allocates data to the different subgroups as shown in Table 2.

Table 2. First organization of the data

Source of Variation	Subgroup Allocation
Hour-to-Hour	Between subgroup
Cycle-to-Cycle	Between subgroup
Cavity-to-Cavity	Within subgroup

Note that due to this allocation the XBar chart confounds the Hour-to-Hour and the Cycle-to-Cycle information.

To generate the Individual and Moving Range charts shown in Figure 2 in Control Chart Builder, follow these steps:

1. Open Socket Thickness.jmp from the [JMP File Exchange](#). The data table has columns for **Day**, **Time**, **Hour**, **Cycle**, **Cavity** and **Thickness**.
2. Select **Analyze > Quality and Process > Control Chart Builder**.
3. Drag the **Thickness** column onto the graph.

Now follow these steps to generate the XBar and R charts shown in Figure 3.

4. Drag the **Hour** column to the x-axis area of the chart.
5. Drag the **Cycle** column to the drop zone just above the x-axis to nest **Cycle** within **Hour**.

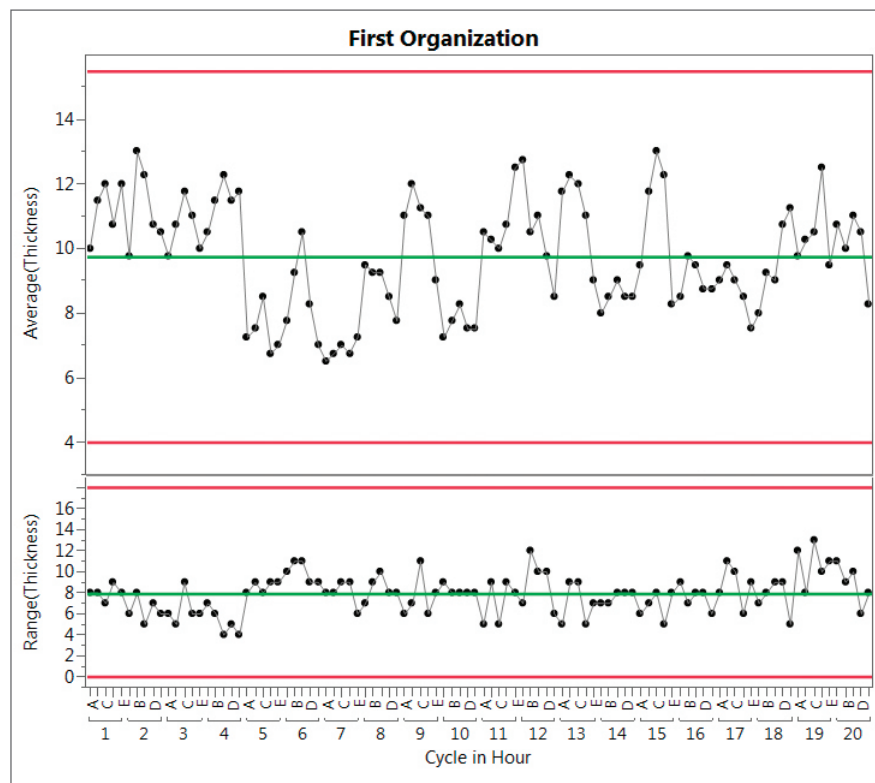


Figure 3 XBar and R chart showing the first organization of the thickness data

The charts in Figure 3 answer the following questions:

- The R chart asks the question: *“Are the Cavity-to-Cavity differences consistent?”*
- The XBar chart asks the question: *“Are there detectable differences from Hour-to-Hour and Cycle-to-Cycle?”*

The R chart shows that the Cavity-to-Cavity differences are consistent and centered around 0.08 mm. Although no points are outside the control limits in the XBar chart, the pattern does not seem random. We are not sure, then, if there are detectable differences Cycle-to-Cycle and Hour-to-Hour.

Second Organization of the Data

The second organization of the thickness data allocates data to the different subgroups as shown in Table 3.

Table 3. Second organization of the data

Source of Variation	Subgroup Allocation
Hour-to-Hour	Between subgroup
Cavity-to-Cavity	Between subgroup
Cycle-to-Cycle	Within subgroup

Note that, due to this allocation, the XBar chart confounds the Hour-to-Hour and the Cavity-to-Cavity information.

Rearrange the variables to update the XBar and R charts as shown in Figure 4.

6. Drag **Cycle** off of the x-axis.
7. Drag the **Cavity** column to the drop zone just above the x-axis to nest **Cavity** within **Hour**.

The charts in Figure 4 answer the following questions:

- The R chart asks the question: *“Are the Cycle-to-Cycle differences consistent?”*

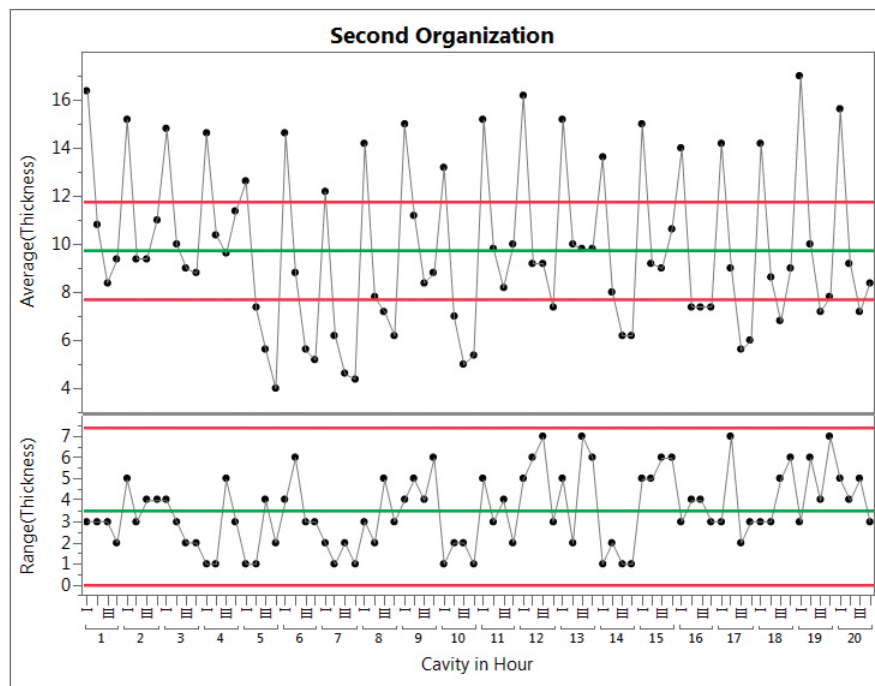


Figure 4 XBar and R chart showing the second organization of the thickness data

- The XBar chart asks the question: *“Are there detectable differences from Hour-to-Hour and Cavity-to-Cavity?”*

From the R chart in Figure 4, we can see that the cycle-to-cycle differences are consistent. This time, however, we see many points out of control in the XBar chart, indicating the presence of detectable differences Cavity-to-Cavity and Hour-to-Hour. Why didn't we see these points in Figure 3? The XBar chart in the first organization did not ask whether there were detectable differences between the cavities, just between Cycle-to-Cycle and Hour-to-Hour.

Selecting the high points in the XBar chart and looking back at the data table reveals that all the high-thickness readings come from Cavity I. This is useful information because it shows that even though the Cavity-to-Cavity differences are consistent, Cavity I in general produces sockets with a higher thickness. This is an indication that the cavities behave differently and their data should not be combined in a single process behavior chart.

Third Organization

To further investigate the differences between cavities discovered in Figure 4, we generate separate process behavior charts per cavity. Table 4 shows the reallocated data.

Table 4. Third organization of the data

Source of Variation	Subgroup Allocation
Hour-to-Hour	Between subgroup
Cycle-to-Cycle	Within subgroup

Rearrange the variables to update the XBar and R charts as shown in Figure 5.

8. Drag **Cavity** from the x-axis to the Phase drop zone above the graph.

Note: This example essentially makes separate control charts for each cavity.

Voilà! Figure 5 clearly shows how the thickness readings coming from Cavity I are higher than the other three cavities. Because there is only one source of variation allocated to each chart, we can confidently answer these questions:

- “For any given cavity, are the Cycle-to-Cycle differences consistent?”
Yes, and they are consistent for each of the cavities. The average ranges, green lines, are similar.
- “For any given cavity, are there detectable differences from Hour-to-Hour?”
Yes. Sockets are thicker one hour, thinner the next.

What Did We Learn?

The careful design of the process behavior chart can reveal patterns that a “default” software chart might mask. However, even when the charts are designed carefully, it is important to rationally think about the allocation of the different sources of variation to subgroups. For our three allocations, the first process behavior chart did not signal the Cavity I difference because it was not designed to detect differences between cavities. The second process behavior chart did signal the Cavity I difference because it was specifically designed to answer the question: Are there detectable differences from Cavity-to-Cavity? But it was the third allocation that clearly revealed not only the Cavity-to-Cavity differences, but also the Hour-to-Hour differences that the other allocations missed. Control Chart Builder made it very easy to design the charts, and helped reveal the hidden features in our data.

The Rest of the Story

The engineer in charge of the molding process sent the mold to the tool shop

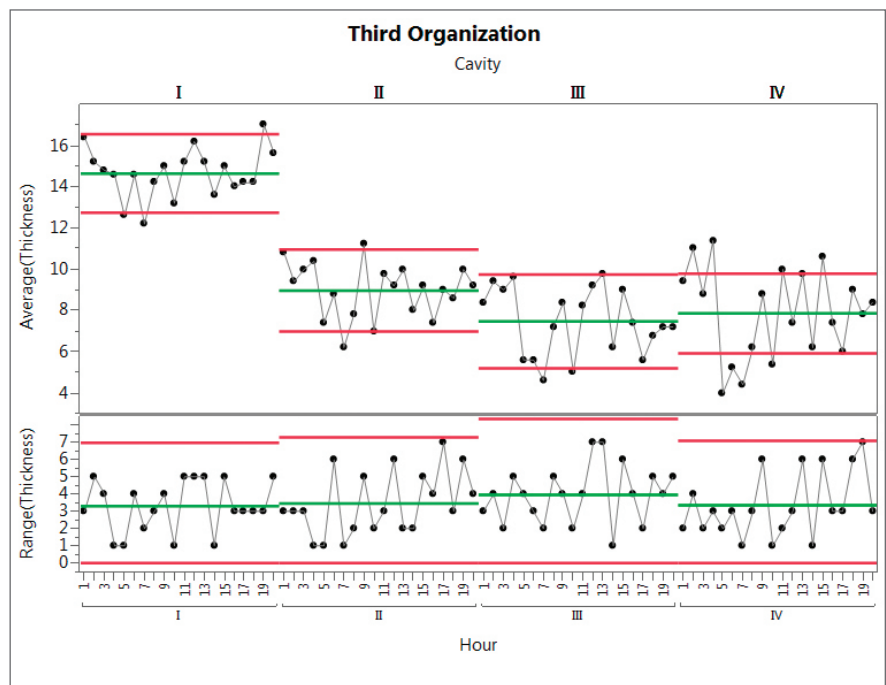


Figure 5 XBar and R chart showing the third organization of the thickness data

to have a 3-1,000th inch shim put behind Cavity I to solve the thickness problem shown in Figure 4. After he got the mold back, he asked the toolmaker if he had done anything else to the mold. The toolmaker said, “I did clean it up real good – there was a wax build-up on the face of the mold. I cleaned that off for you.” It was then that the process engineer realized that the process was out of control because the operators were not cleaning off the wax build-up often enough. ✨

REFERENCES

Wheeler, D.J., and D.S. Chambers. (1992) *Understanding Statistical Process Control*. Second Edition. Knoxville, TN: SPC Press. The data and example come from Section 5.6.

Ramírez, José G., and Ramírez, Brenda S. (2009) *Analyzing and Interpreting Continuous Data Using JMP: A Step-by-Step Guide*. Cary, NC: SAS Institute Inc.

JMP Quality Explorers Series

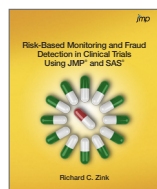
JMP® Books From SAS® Press

SAS Press, a very busy department within the Publications Division at SAS, continues to publish up-to-date books by outside authors. Topics by experts in a variety of fields are selected, carefully screened, edited and produced to offer SAS and JMP users the latest in theory, techniques, examples and case studies. In keeping with the latest in publishing technology, many books are now available as e-books.

Check out these upcoming JMP titles from SAS Press!

Risk-Based Monitoring and Fraud Detection in Clinical Trials Using JMP and SAS

(Anticipated publish date July 2014)
By Richard C. Zink



Improve efficiency while reducing costs in clinical trials with centralized monitoring techniques using JMP and SAS.

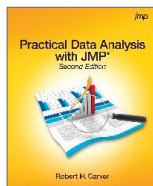
The well-being of trial participants and the validity and integrity of the final analysis results are at stake. Risk-based monitoring (RBM) makes use of central computerized review of clinical trial data and site metrics to determine if and when clinical sites should receive more extensive quality review or intervention.

Risk-Based Monitoring and Fraud Detection in Clinical Trials Using JMP and SAS describes analyses for RBM. Methods to detect potential misconduct, snapshot comparisons to identify new or modified data, and other novel techniques to enhance safety and

quality reviews are covered. The methods described in this book enable the clinical trial team to take a proactive approach to data quality and safety to streamline clinical development activities and address shortcomings while the study is ongoing.

Practical Data Analysis With JMP, Second Edition

(Anticipated publish date July 2014)
By Robert Carver



Understand the concepts and techniques of analysis while learning to reason statistically.

Practical Data Analysis With JMP, Second Edition by Robert H. Carver uses the powerful interactive and visual approach of JMP to introduce readers to the logic and methods of statistical thinking and data analysis. Three new review chapters help readers to integrate ideas and technique. In addition, the scope and sequence of the chapters has been updated with more coverage of data management and analysis of data.

Reflective of the broad applicability of statistical reasoning, the problems and examples come from a wide variety of disciplines, including engineering, life sciences, business and economics, among others, and include a number of international and historical real-world examples. This book introduces you to the major platforms and essential features of JMP and will leave you with a sufficient background and the confidence to continue your exploration independently.

Building Better Models With JMP Pro

(Anticipated publish date November 2014)
By Jim Grayson, Sam Gardner and Mia Stephens



Take a peek inside the black box of business analytics plus the methodology for managing and executing analytics projects.

Building Better Models With JMP Pro by Jim Grayson, Sam Gardner and Mia Stephens provides an applications-oriented introduction to data mining for the business student or entry-level business analyst with a proven process, which guides you in the application of data-mining concepts and tools. It will tell the what, why and how of using JMP Pro for data mining within the context of the business problem. Topics include: regression, logistic regression, classification and regression trees, neural networks, model cross-validation, model comparison and selection, and data reduction techniques.

Faculty members who teach predictive modeling, data mining and/or business analytics at the lower to upper graduate level and their students, or working professionals looking for an introduction to analytics with JMP, would benefit greatly from this book. It's also a great resource for business statistics, business analytics and predictive modeling. No prior experience with JMP is needed.

For more information about SAS Press and a complete list of available JMP and SAS books, go to support.sas.com/publishing.



100 SAS CAMPUS DRIVE, CARY, NC 27513

2014

DISCOVERY SUMMIT

EXPLORING DATA
INSPIRING INNOVATION

September 15 - 18
SAS World Headquarters
Cary, North Carolina

Find more information at
jmp.com/summit.



What's New in JMP® Training

New courses debut at JMP® Discovery Summit 2014

Don't miss the JMP Discovery Summit, Sept. 15 -18, in Cary, NC. SAS Education offers discounted JMP training to Discovery Summit attendees before and after the conference. When you register for the Summit, sign up for one of the new courses on Modern Screening Designs, Modeling Multidimensional Data or Consumer Choice Research. You can also sign up for the JMP Scripting Forum.

Visit jmp.com/about/events/summit2014/training.shtml for more information about the new courses and to register.

Contact Deborah Upchurch for more information or to schedule your course.

training@jmp.com

800-333-7660