Calibration

Overview

This lecture introduces calibration along with methods for checking whether a model is calibrated.

Calibration

- Calibration is a fundamental property of any predictive model. If a model is not calibrated its predictions may systematically lead to poor decisions.
- Test for calibration uses the plot of Y on the fitted values. It is similar to checking for a nonlinear pattern in the scatterplot of "Y versus X" with a transformation.

New tool

- Smoothing splines fit smooth trends that may not be linear.
- Calibrating a model using smoothing splines improves its predictions of the response, albeit without offering more in the way of explanation.

Introduction

Better models produce better predictions in several ways:

- The better the model, the more closely its predictions track the *average* of the response.
- The better the model, the more precisely its predictions match the response (*e.g.*, smaller prediction intervals).
- The better the model, the more likely it is that the distribution of the prediction errors are normally distributed (as assumed by the MRM).

Consequences of systematic errors

- The first aspect, to be right on average, is critical.
- Unless predictions from a model are right on average, the model cannot be used for economic decisions.

Calibration is about being right on average.

High $R^2 \neq$ calibrated. Calibration is neither a consequence nor precursor to a large R^2. A model with $R^2 = 0.05$ may be calibrated, and a model with $R^2 = 0.999$ need not be calibrated.

In simple regression, calibration is related to the choice of transformations that capture nonlinear transformations.

Calibration

Definition

A model is calibrated if its predictions are "right" on average:

ave(Response | Predicted value) = Predicted value.

$$E\left(Y \middle| \hat{Y}\right) = \hat{Y}$$

Calibration is a basic property of any predictor, whether the prediction results from the equation of a regression or a gypsy's forecast from a crystal ball.

A regression should be calibrated before we use its predictions.

Calibration in other situations

Physician's opinions on risk of an illness

Weather forecasts

Political commentary on the state of financial markets

Example

A bank has made a collection of loans. It expects to earn \$100 interest *on average* from these loans each month.

If this claim is calibrated, then the average interest earned over following months will be \$100.

If the bank earns \$0 half of the months and \$200 during the other months, then the estimate is calibrated. (Not very precise, but it would be calibrated.)

If the bank earns \$105 every month, then the estimate is not calibrated (even though it is arguably a better prediction).

Role in demand models (Module 6)

Getting the mean level of demand correct is the key to making a profitable decision. Calibration implies that predicted demand is indeed the mean demand μ under the conditions described by the explanatory variables.

Example: If you collect days on which we predict to sell 150 items, the average sales on these days is 150.

Calibration does not imply small variance or a good estimate of σ. It only means that we are right on average.

Another Reason for Calibration

A gift to competitors

If a model is not calibrated, a rival can improve its predictions without having to model the response.

Rivals have an easy way to beat your estimates.

Example

You spend a lot of effort to develop a model that predicts the credit risk of customers.

If these predictions are not calibrated, a rival with *no* domain knowledge can improve these predictions, obtaining a better estimate of the credit risk and identify the more profitable customers.

Do you really want to miss out?

Checking the Calibration of a Regression

Basic procedure is familiar

No new tools are absolutely necessary: you've been checking for calibration all along, but did not use this terminology.

Checking for calibration uses methods that you have already seen in other contexts.

Calibration in simple regression

A simple regression is calibrated if the fit of the equation tracks the average of the response.

Often, to get a calibrated model, you transform one or both of the variables (*e.g.*, using logs or reciprocals)

Calibration in multiple regression

The "calibration plot" is part of the standard output produced by JMP that describes the fit of a multiple regression.

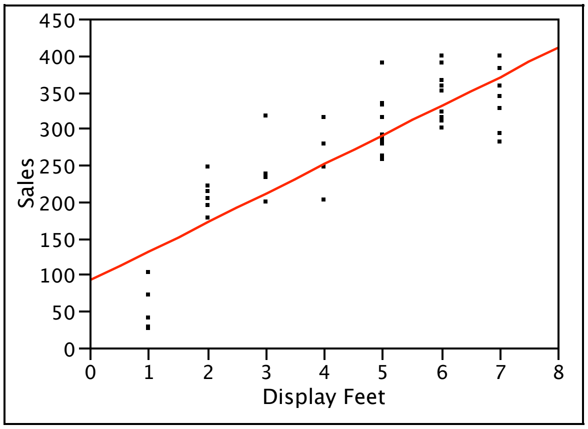
Calibration in Simple Regression (display.jmp)

Relationship between promotion and sales[1]

How much of a new beverage product should be put on display in a chain of liquor stores?

[1] This example begins in BAUR on page 12 and shows up several times in that casebook. This example illustrates the problem of resource allocation because to display more of one item requires showing less of items. The resource is the limited shelf space in stores.

Scatterplot and linear fit



Sales = 93 + 39.8 Display Feet

The linear fit misses the average amount sold for each amount of shelf space devoted to its display. The objective is to identify a smooth curve that captures the relationship between the display feet and the amount sold at the stores.

Smoothing spline

We can show the average sales for each number of display feet by adding a **smoothing spline** to the plot.[2]

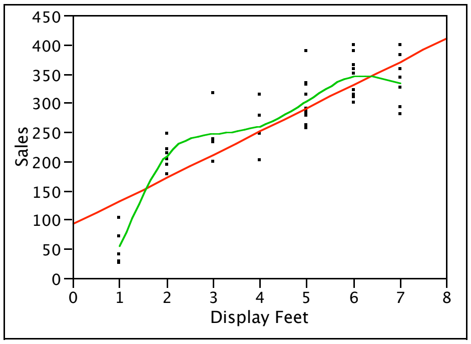
A spline is a *smooth* curve that connects the average value of *Y* over a subset of values of identified by a subset of adjacent values of *X*.

The smoothness of the spline depends on the data and can be controlled by a slider in the software. The slider controls the size of the averaged subset by controlling the range of values of *X*.

Example of spline

In this example, a "smoothing spline" joins the mean sales for each amount sold with a smooth curve.[3]

[2] In the Fit Y by X view of the data, choose the Fit Spline option from the tools revealed by the red triangle in the upper left corner of the window. Pick the "flexible" option.



With **replications** (several stores display the same amount; these data have several *y*'s for each *x*), the data form "columns" of dots in the plot.

If the range of adjacent values of *X* is small enough (less than 1) then the spline joins the means of these columns.[4]

Without replications, the choice of a spline is less obvious and can require a bit of practice (*i.e.*, guesswork).

Compare the spline to the linear fit

With one foot on display, the linear fit seems too large. The mean is about $50.

With two feet on display, the linear fit seems too small. The mean of sales is around $200.

The linear fit explains R^2 = 71% of the variation. A spline *always* does as least as well as the linear fit. For the spline, R^2 = 86%.

The situation resembles multiple regression: R^2 grows as we add predictors. In a sense, a spline "adds predictors" as well (but JMP does not tell us how many).

[3] The spline tool is easy to control. Use the slider to vary the smoothness of the curve shown in the plot. See BAUR casebook, p. 13, for further discussion of splines and these data.

[4] Replications (several cases with matching values of all of the X*s*) are a great asset if you have them. Usually, they only show up in simple regression because there are too many combinations of predictors in multiple regression. Because this dataset has replications, we can use JMP's "Lack of Fit" test to check for calibration. See the BAUR casebook, page 247.

Testing for a Lack of Calibration

Objective

Does the spline really fit the data better that the line (able to predict new observations more accurately)?

Does the spline capitalize on random, idiosyncratic features of the data that won't show up in new observations?

Approximate spline with polynomial

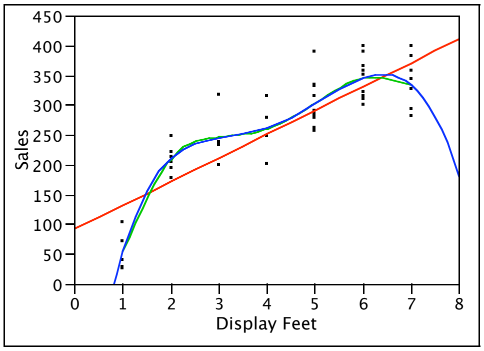
Approximate the spline by fitting a polynomial that roughly matches its fit to the data. We *can* test the polynomial fit to see if its better than the linear fit.

Fit a polynomial model[5]

This plot superimposes the fit of a 5th degree polynomial

$$\hat{y} = b_0 + b_1\ X + b_2\ X^2 + b_3\ X^3 + b_4\ X^4 + b_5\ X^5$$

The output centers the predictor (subtracting the mean 4.4 to reduce the collinearity). We only need R^2.



RSquare	**0.8559**
Root Mean Square Error	38.22968

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	109.52749	70.2328	1.56	0.1266
Display Feet	37.9317	15.51394	2.45	0.0189
(Display Feet-4.40426)^2	12.189222	9.094144	1.34	0.1875
(Display Feet-4.40426)^3	-2.36853	6.066507	-0.39	0.6982
(Display Feet-4.40426)^4	-2.070732	1.236305	-1.67	0.1016
(Display Feet-4.40426)^5	0.108778	0.560066	0.19	0.8470

[5] To add a polynomial to a scatterplot, click the red triangle in the upper left corner of the output window and pick Fit Polynomial. To match the spline often requires 5 or 6 powers.

Partial F-test

Do not use the t-statistics to assess the polynomial because of collinearity (even with centering). We're not interpreting the coefficients; all we want to learn is whether R^2 significantly improved.

The partial F-test does what we need. It finds a very significant increase for the **4** added predictors.

The degrees-of-freedom divisor in the numerator of the ratio is 4 (not 5) because the polynomial adds 4 more predictors. The d.f. for the denominator of F is $n - 2$ (for the original fit) – 4 (added by the spline)

$$\begin{aligned} F &= \frac{(0.8559 - 0.7120)/4}{(1 - 0.8559)/(47 - 6)} \\ &= \frac{41}{4} \times \frac{0.1439}{0.1441} \\ &= 10.2 \end{aligned}$$

Conclude: The *linear* regression is not calibrated.[6]

Calibrating simple regression

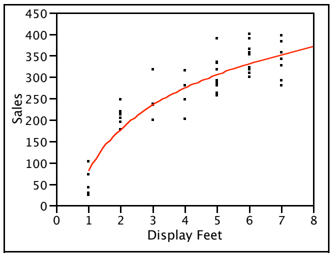
Transform one or both variables to capture the pattern in the means and obtain a more interpretable model that fits almost as well.

Substantively, model the *diminishing returns to scale* using a log of the number of display feet as a predictor.

The R^2 of the fit using log(display feet) is 81.5% and the partial F improvement offered by the polynomial is no longer significant (F = 2.9).[7]

[6] The improvement is statistically significant even though *none* of the t-stats for the added coefficients is significant. There's too much collinearity. When you add several terms at once, use the partial F to judge the increase.

[7] F= ((0.8559-0.815)/4)/((1-0.8559)/41) = 2.9



Sales = 83.56 + 138.62 Log(Display Feet)

Calibration Plot

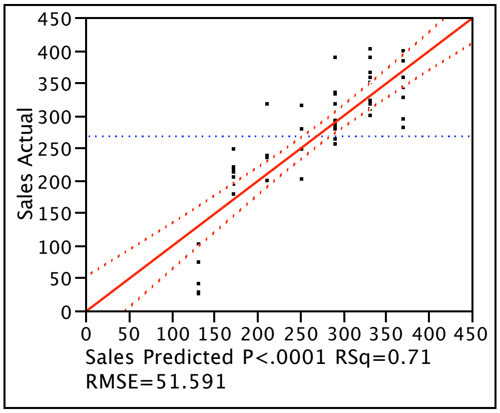
Generalize to multiple regression

A calibration plot offers an equivalent test of calibration, one that *generalizes to multiple regression*.

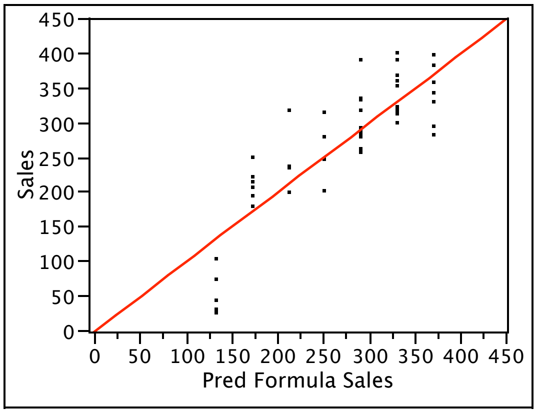
We need this generalization to test for lack of calibration more convenient in models with more than one predictor.

Definition

A **calibration plot** is a scatterplot of the actual values of the response on the y-axis and the predicted values on the x-axis. (This plot is part of the default output produced by "Fit Model".)



We need to make our own calibration plot in order to test for calibration. *Save the prediction formula* (or the predicted/fitted values) from the simple regression, then scatterplot the actual values on the predicted values.



The slope of this fit must be $b_0 = 1$ and the intercept must be $b_1 = 0$. The R^2 and RMSE match those of the previous regression. Why is this true?

```
                    Linear Fit[8]
             Sales = 0 + 1 Pred Formula Sales

                    Summary of Fit
         RSquare                            0.712
         Root Mean Square Error            51.591
         Observations (or Sum Wgts)            47

                 Parameter Estimates
Term                  Estimate   Std Error   t Ratio   Prob>|t|
Intercept                    0       26.51      0.00     1.0000
Pred Formula Sales           1        0.09     10.55     <.0001
```

Hint: Recall that R^2 in regression is the square of the correlation between the response and the predicted response.

[8] JMP sometimes uses "scientific notation" for numbers very close to zero, something like 1.0e-13, which translates to 1×10^{-13}. The difference from zero in this output is due to round-off errors in the underlying numerical calculations; computers use binary arithmetic that does not represent *exactly* every number that you write down.

Checking the Calibration

Using the calibration plot

We can check the calibration of *any regression* from the calibration plot.

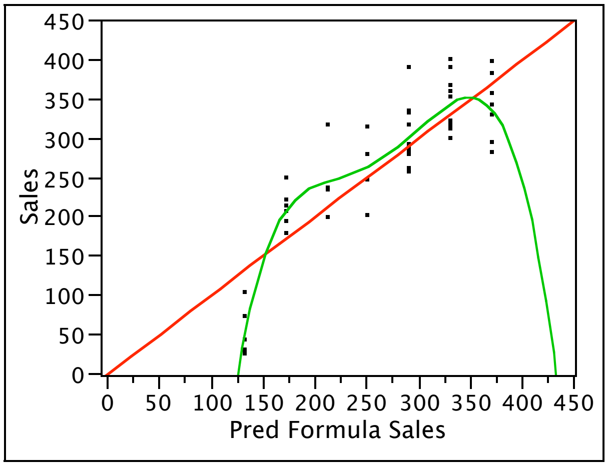
Example

This summary of what happens when we add the 5th order polynomial for the display example.[9]

The coefficients of the polynomial differ from those in the prior fit, but R^2 and RMSE of the fit are the same.

Caution: Be careful computing the partial F test. We want the degrees of freedom as though we added these terms to the original model as done previously, not started over.

Hence, we reach the same conclusion from the calibration plot as when working with the scatterplot of Y on X: the linear model not calibrated.



The summary of the fit of this model is on the next page.

[9] Why use a 5th order polynomial, you might ask? It's arbitrary, but seems to match the fit of a smoothing spline so long as we stay *away from the edges of the plot*.

```
                    Summary of Fit
            RSquare                        0.8559
            Root Mean Square Error        38.2297
            Mean of Response             268.1300
            Observations (or Sum Wgts)    47.0000

                    Parameter Estimates
Term                               Estimate   Std Error   t Ratio   Prob>|t|
Intercept                         20.765257    106.1972      0.20     0.8459
Pred Formula Sales                0.9541011    0.390224      2.45     0.0189
(Pred Formula Sales-268.13)^2     0.0077119    0.005754      1.34     0.1875
(Pred Formula Sales-268.13)^3     -0.000038    0.000097     -0.39     0.6982
(Pred Formula Sales-268.13)^4      -8.29e-7      4.9e-7     -1.67     0.1016
(Pred Formula Sales-268.13)^5        1.1e-9      5.6e-9      0.19     0.8470
```

Warning: Never extrapolate a high-order polynomial…

The fit of a polynomial of this degree “goes wild” when extrapolated beyond the data, even though it seems reasonable within the observations.

The spline is much more useful for this purpose.

When to check for calibration?

Calibration is a diagnostic, something done once you feel you have a reasonable model and are ready to start checking residuals.

Checking Calibration in Multiple Regression

Procedure

Use the calibration plot of y on the fitted values $\hat{y}$ (*a.k.a.*, predicted values) from the regression.

Test the calibration by seeing whether a polynomial fit in the plot of y on $\hat{y}$ significantly improves upon the linear fit that summarizes the R^2 and RMSE of the regression.

Tricky part

The degrees of freedom for the associated partial F test depend on the number of slopes in the original multiple regression.

Thankfully, this adjustment has little effect unless you have a small sample.

Example data

Model of housing prices, from the analysis of the effects of pollution on residential home prices in the 1970's in the Boston area (introduced in Module 3).

In the examples of interactions, we observed curvature in the residuals of the models. That's a lack of calibration.

Illustrative multiple regression

This regression with 5 predictors explains almost 70% of the variation in housing values. All of the predictors are statistically significant (if we accept the MRM) with tiny *p*-values.[10]

The model lacks interactions; every predictor is assumed to have a linear partial association with the prices, with a fixed slope regardless of other explanatory variables.

[10] This model is part of the JMP data file; run the "Uncalibrated regression" script.

Summary of Fit

RSquare	0.689
Root Mean Square Error	4.951
Mean of Response	22.248
Observations	500

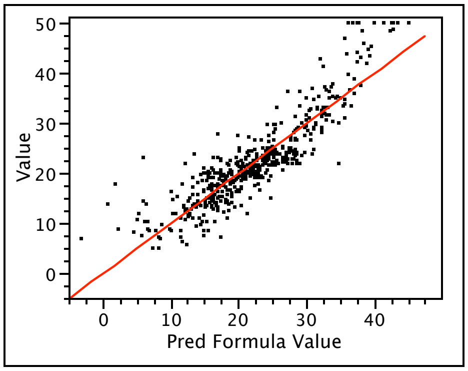
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	72.10	3.11	23.18	<.0001
NOx	-20.83	3.25	-6.41	<.0001
Distance	-1.55	0.19	-8.18	<.0001
Lower class	-0.74	0.04	-17.66	<.0001
Pupil/Teacher	-1.28	0.12	-10.88	<.0001
Zoning	0.05	0.01	3.48	0.0006

Calibration plot

We find curvature in the plot of the housing values on the fitted values from the regression.

As happens often, a lack of calibration is most evident at the extremes with the smallest and largest fitted values.



RSquare	0.689
Root Mean Square Error	4.93

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-7.5e-14	0.706	0.00	1.0000
Pred Formula Value	1	0.030	33.18	<.0001

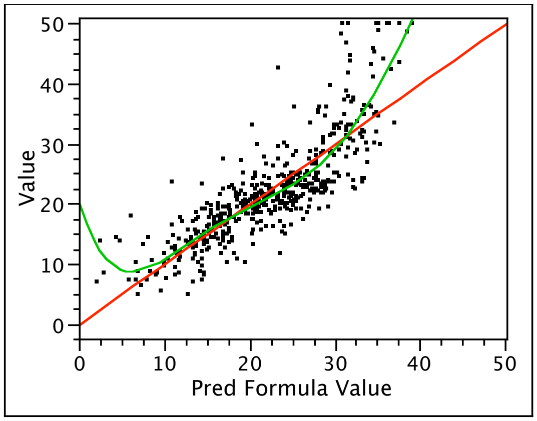
The fitted slope and intercept are 1 and 0, with R^2 and RMSE matching those in the multiple regression.

Testing for a lack of calibration

Use the procedure that we applied to simple regression, starting from a scatterplot of y on $\hat{y}$.

The fitted 5th order polynomial (or spline) shows that average housing values are larger than the fitted values at *both* sides of the plot. The fitted values are too small for *both* inexpensive and expensive housing.

The effect for inexpensive tracts (left) seems less noticeable than for the expensive tracts (right).



The partial F-test for calibration shows that there's room to improve the model.[11]

Summary of Fit

RSquare	0.742
Root Mean Square Error	4.506
Observations (or Sum Wgts)	500

$$F = \frac{(0.742 - 0.689)/4}{(1 - 0.742)/(500 - 6 - 4)}$$

$$= \frac{490}{4} \times \frac{0.053}{0.258} = 25.2$$

[11] The degrees of freedom in the F are 4 (for the 4 nonlinear terms in the polynomial) and $n - 6 - 4$. (6 for the intercept and slopes in the original fit plus 4 for the non-linear terms).

Calibrating a Model

What should be done when a model is not calibrated?

Simple regression. *Ideally*, find a substantively motivated transformation, such as a log, that captures the curvature. *Data mining is no substitute for knowing the substance of the problem*.

Multiple regression. Again, find the right transformation or a missing predictor. This can be hard to do, but some methods can find these (and indeed work for this data set and model).

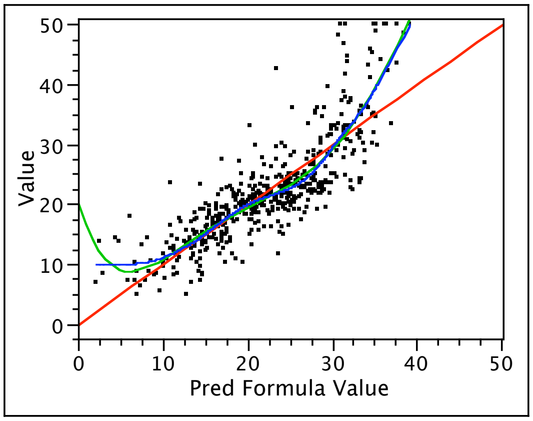
If you only need predictions, calibrate the fit.

(a) Use predictions from the polynomial used to test for calibration, or better yet…

(b) Use a spline that matches the polynomial fit in the test for calibration. The spline avoids the "edge effects" that make polynomials go wild when extrapolating.

Example

The matching spline (lambda=238.43) has similar R^2 to the fit of the polynomial model used in the test (74.2%), but the spline controls the behavior at the edges of the plot.



```
R-Square                    0.748328
Sum of Squares Error        9784.732
```

Predicting with a Smoothing Spline

Use this three-step process:

1. Save predictions $\hat{y}$ from the original, uncalibrated multiple regression (used to test the calibration).
2. Match a smoothing spline to the polynomial in the calibration plot (as done above).
3. Save the prediction *formula* from this smoothing spline as a column in the data table.
4. Insert values of $\hat{y}$ from the regression into the formula for the smoothing spline to get the final prediction.

Example

Predict the first row. This row is missing the response column and matches the predictors held in row #206, a case that the fitted regression under-predicts.

Predicted value from multiple regression = 39.0

Predict the response using the spline fit for cases with this predicted value. If you have saved the formula for the spline fit, this calculation happens automagically.

Predicted value from the spline = 49.5

The actual value for the matching case in row #206 is 50.

The spline gives a much better estimate!

The use of a spline after a regression fixes the calibration, but *avoids any attempt at explanation*.

That's a weakness. Knowing why the model under-predicts housing values at either extreme might be important.

Other Forms of Calibration

When a model predicts the response right, on average, we might ought to say that the model is *calibrated to first order*.

Higher-order calibration matters as well:

(a) *Second-order calibrated*: the variation around every prediction is a known or correctly estimated (including capturing changes in error variation).

(b) *Strongly calibrated*: the distribution of the prediction errors is known or correctly estimated (such as with a normal distribution).

We typically check these as part of the usual regression diagnostics, after we get the right equation for the mean value.

That's the message here as well:

> Get the predicted value right on average, then worry about the other attributes of your prediction.