The Stata Journal (2010)
10, Number 2, pp. 215–225

Model fit assessment via marginal model plots

Charles Lindsey
Texas A & M University
Department of Statistics
College Station, TX
lindseyc@stat.tamu.edu

Simon Sheather
Texas A & M University
Department of Statistics
College Station, TX
sheather@stat.tamu.edu

Abstract. We present a new Stata command, `mmp`, that generates marginal model plots (Cook and Weisberg, 1997, *Journal of the American Statistical Association* 92: 490–499) for a regression model. These plots allow for the comparison of the fitted model with a nonparametric or semiparametric model fit. The user may precisely specify how the alternative fit is computed. Demonstrations are given for logistic and linear regressions, using the `lowess` smoother to generate the alternate fit. Guidelines for the use of `mmp` under different models (through `glm` and other commands) and different smoothers (such as `lpoly`) are also presented.

Keywords: st0189, mmp, regress, glm, lpoly, logit, logistic, marginal model plots

1 Theory/motivation

Graphical assessment of a model's fit can be an intuitive (even essential) tool in regression analysis. Ordinary least-squares linear regression allows powerful graphical assessment diagnostics through the model's residuals. In other forms of generalized linear model regression, this may not be the case. For example, when we are performing binary logistic regression, all the residuals normally used (Pearson, deviance, and response) will display a nonrandom and uninterpretable pattern when plotted against the model's predictors, even if the correct model has been fit. Bypassing the use of the residuals, Cook and Weisberg (1997) provide an alternative graphical assessment for regression model fit: marginal model plots. The assessment is performed as follows.

Suppose that we have k predictors. If all of them are continuous, we will produce $k + 1$ plots, one for each predictor and one for the $\beta' x$ linear forms. If a predictor is not reasonably continuous (more than two values), then we omit its plot. Each of the generated plots is called a marginal model plot. In each of the plots, a scatterplot is drawn, with the response on the vertical axis and the predictor or linear form on the horizontal axis.

The regression model's estimate of the response's mean, conditional on the predictor (linear form), is then computed at each value of the predictor (linear form). A line is passed through the estimated points. This is the model line. Now a nonparametric estimate (smooth) of the response's mean, conditional on the predictor (linear form), is computed at each value of the predictor (linear form). A new line (usually of a different color or pattern) is then passed through these points. This is the alternative line.

When we pick a good nonparametric estimator for generating the alternative line, we can assess the fit of the model by judging how closely the lines overlap. If the lines match each other closely in each of the generated plots, we conclude that our regression model is a good fit. If there is significant disparity between the lines in any of the plots, our regression model is not a good fit. Figure 1 shows a set of marginal model plots that demonstrate the good fit of a linear regression model. We will discuss this figure futher in section 2.1. The dashed line is the model line.

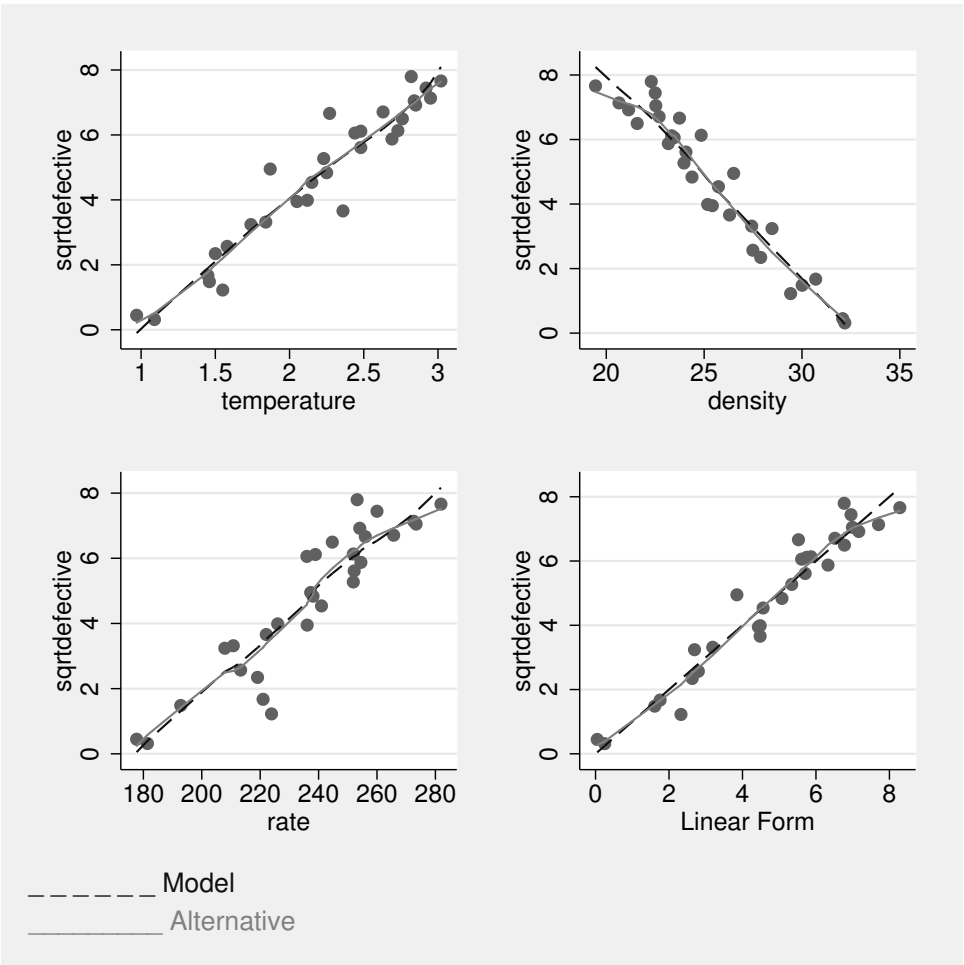

Figure 1. Marginal model plot example

Choosing a good nonparametric estimator is key to correctly use this method. There are many options. Cook and Weisberg (1997) use a lowess smoother to make their estimates. In this article, we will use the lowess smoother provided by Stata (see [R] **lowess**) with a tuning parameter $\alpha = 2/3$.

Estimation of the alternative line is completely dependent on the nonparametric estimator used, but there are interesting general results used in estimation of the model line. We will discuss these briefly.

Suppose that we observe the response y and the predictors $x = (x_1, \ldots, x_k)'$. We have as a regression model the generalized linear model $E_M(y \mid x) = g(\beta' x)$. To find the model line, we estimate $E_M(y \mid \alpha' x)$, where $\alpha' x$ is the linear form $\widehat{\beta}' x$ or a single predictor $x_1, \ldots, x_k$. Cook and Weisberg (1997) saw that for any α vector of the proper dimension,

$$E_M(y \mid \alpha' x) = E\left\{E_M(y \mid x) \mid \alpha' x\right\}$$

We estimate $E_M(y \mid x)$ by $g(\widehat{\beta}' x)$. So the estimator of $E_M(y \mid \alpha' x)$ is an estimator of $E\left\{g(\widehat{\beta}' x) \mid \alpha' x\right\}$. The closed form solution of this expectation for an arbitrary α is probably unknown or rather complicated. We can estimate it by using the nonparametric estimator that generates the alternative line.

So we will use the nonparametric estimator twice, once for the alternative line $(y \mid \alpha' x)$ and again for the model line $\left\{g(\widehat{\beta}' x) \mid \alpha' x\right\}$. Again note that $\alpha' x$ is $\widehat{\beta}' x$ or a single predictor $x_1, \ldots, x_k$.

Generally, the closer the estimated line is to the points in the plot, the more accurate the estimation method is. So one may extend the methodology and use a semiparametric estimate or even a parametric estimate to generate the alternative line. The marginal model plots can then be used to assess whether the alternative estimator is as good as the model estimator, or vice versa.

We present a new Stata command, `mmp`, that creates marginal model plots. With very mild restrictions, it allows users to construct a traditional marginal model plot using any nonparametric smoother and generalized linear model that they wish. In addition, a user may try some of the less orthodox strategies already presented if desired.

We will restrict ourselves to linear and logistic regressions, because the appropriateness of the lowess smoother with tuning parameter $\alpha = 2/3$ is well documented.

In this article, we are only concerned with estimating the mean of the response given the predictors. We note that Cook and Weisberg (1997) extended the marginal model plot method for variance estimation as well. We also emphasize that our method is only appropriate after running generalized linear model estimation commands and that the input sample should have independent observations.

(*Continued on next page*)

2 Use and examples

`mmp` is to be executed after an estimation command that performs a regression. The `mmp` command has the following syntax:

`mmp, mean(`*string*`) smoother(`*string*`)` [`smooptions(`*string*`) linear predictors varlist(`*varlist*`) generate indgoptions(`*string*`) goptions(`*string*`)`]

With the `mean()` option, the user informs `mmp` how it should use the `predict` command to generate the estimated response mean. For example, if the user specifies `mean(xb)`, then `mmp` will generate the estimated response mean by calling `predict, xb`.

The `smoother()` option tells `mmp` which nonparametric estimation (smoothing) command to use for generating the alternative and model lines. The only restriction on the smoothing command is that it must have a `generate()` option that takes a single variable (where the smoothed values are stored); for example, the `lowess` command has a `generate()` option and so is appropriate to specify in `smoother()`. Additional options for the smoothing command are passed into the `smooptions()` option.

When `linear` is specified, a marginal model plot will be generated for the $\widehat{\beta}'x$ linear forms. If `predictors` is specified, then a marginal model plot is generated for each predictor. Marginal model plots for single predictors (or even unrelated variables) can be generated through placement in the optional *varlist* in the `varlist()` option.

The `generate` option makes `mmp` save the lowess estimates for the model and alternative lines as variables for each of the produced plots. If a plot is produced for predictor `x`, then variables `x_model` and `x_alt` are added to the data. If the `linear` option is specified to draw a plot for the linear form, then variables `linform_model` and `linform_alt` are added to the data.

The last two options of `mmp` concern the visual presentation of the plots. Graphical options passed into `indgoptions()` affect the display of individual marginal model plots, while options passed into `goptions()` affect the display of the combined array of marginal model plots.

We will now demonstrate the use of these options with some examples.

2.1 Ordinary least-squares example: Defective parts

First, we investigate the regression that yielded figure 1. These data were taken from Sheather (2009). The manufacturing criteria `temperature`, `density`, and `rate` are used to predict the number of defective parts produced, `defective`. See figure 2.

```
. use defects
. regress defective temperature density rate
```

Source	SS	df	MS		
Model	9609.44261	3	3203.14754	Number of obs =	30
Residual	1314.43141	26	50.5550544	F(3, 26) =	63.36
				Prob > F =	0.0000
Total	10923.874	29	376.685311	R-squared =	0.8797
				Adj R-squared =	0.8658
				Root MSE =	7.1102

defective	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
temperature	16.07792	8.294106	1.94	0.063	-.9708617	33.12669
density	-1.827292	1.497068	-1.22	0.233	-4.90456	1.249976
rate	.1167321	.1306268	0.89	0.380	-.1517751	.3852394
_cons	10.32437	65.92648	0.16	0.877	-125.1894	145.8382

```
. mmp, mean(xb) smoother(lowess) smooptions(bwidth(.6666667)) predictors linear
> goptions(xsize(10) ysize(10)) generate
```

(*Continued on next page*)

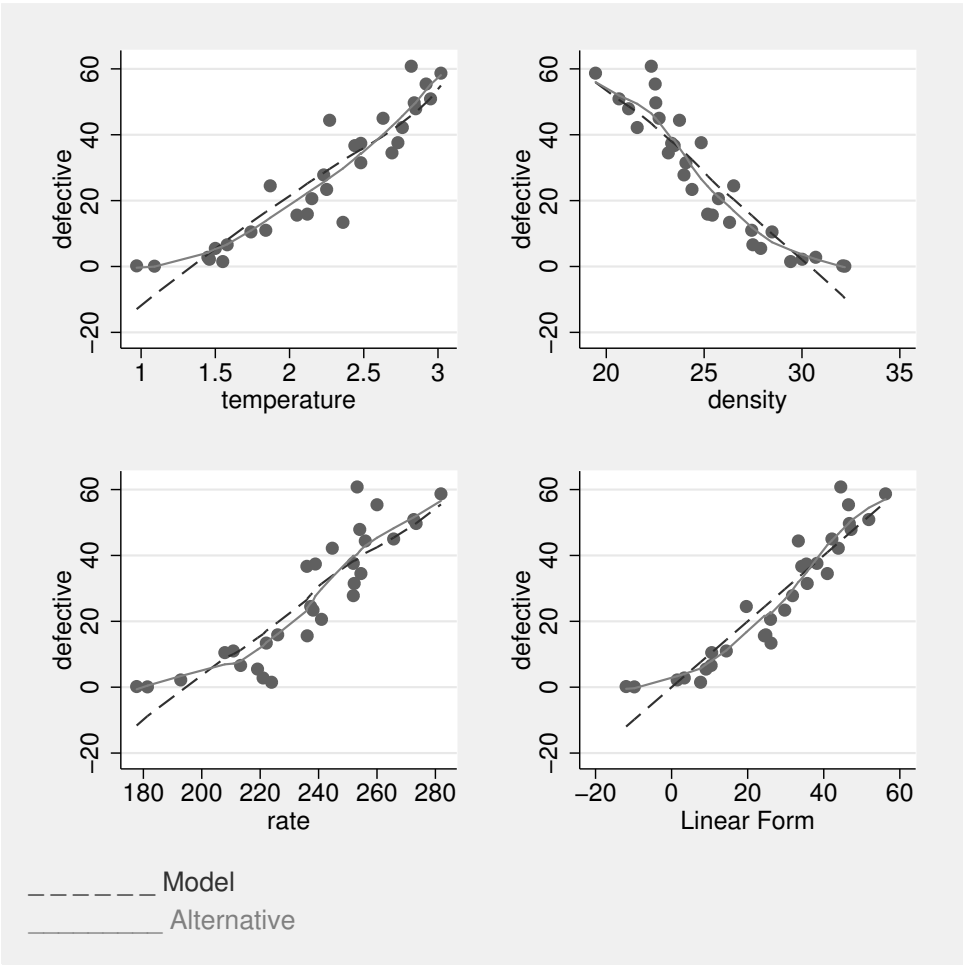

Figure 2. Defective marginal model plots

In the code, we specified that the mean of `defective` be estimated with `predict, xb`. The smoother would be `lowess` with a bandwidth of 2/3. We also specified that the resulting graphic would be square with a width of 10 centimeters.

The marginal model plots show us that the model is not perfect. When we examine the first plot, we see a quadratic trend to the fit. Further investigation in Sheather (2009) shows that it is appropriate to transform `defective`. The regression with $\texttt{defective}^{1/2}$ yielded the well-fitting marginal model plots in figure 1. We used the `generate` option in this example, so we will use the `summarize` command (see [R] **summarize**) to compare the `*_model` and `*_alt` variables. As suggested by the plots, the model estimates and the alternative estimates have notable differences.

```
. summarize temperature_model temperature_alt density_model density_alt
> rate_model rate_alt linform_model linform_alt
```

Variable	Obs	Mean	Std. Dev.	Min	Max
temperatur~l	30	27.18532	18.28548	-12.95191	54.9467
temperatur~t	30	27.62225	18.2362	-.3680072	58.10513
density_mo~l	30	27.0847	17.79125	-9.657054	55.92059
density_alt	30	27.34462	17.56772	-.214293	55.99818
rate_model	30	27.21515	17.05721	-11.64078	55.50286
rate_alt	30	27.44537	17.01681	-.6469474	56.58284
linform_mo~l	30	27.14333	18.2033	-11.95628	56.24563
linform_alt	30	27.49958	18.17563	-.4954669	57.07471

2.2 Logistic example: Michelin

Now we will try a logistic regression example. Another example in Sheather (2009) predicts the log odds of inclusion of a French restaurant in New York City in the Michelin restaurant guide, `inmichelin`, with the cost of a standard dinner at the restaurant, `cost`, and scores from the Zagat restaurant guide of the restaurant criteria, `decor`, `food`, and `service`. See figure 3.

```
. use michelin, clear
. logit inmichelin food decor service cost, nolog
Logistic regression                             Number of obs   =        164
                                                LR chi2(4)      =      77.39
                                                Prob > chi2     =     0.0000
Log likelihood = -74.198473                     Pseudo R2       =     0.3428
```

inmichelin	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
food	.4048471	.131458	3.08	0.002	.1471942	.6624999
decor	.0999727	.0891936	1.12	0.262	-.0748436	.274789
service	-.1924241	.1235695	-1.56	0.119	-.4346159	.0497677
cost	.0917195	.031753	2.89	0.004	.0294848	.1539542
_cons	-11.19745	2.308961	-4.85	0.000	-15.72293	-6.671971

```
. mmp, mean(pr) smoother(lowess) smooptions(bwidth(.6666667)) predictors linear
> goptions(xsize(9) ysize(10))
```

(*Continued on next page*)

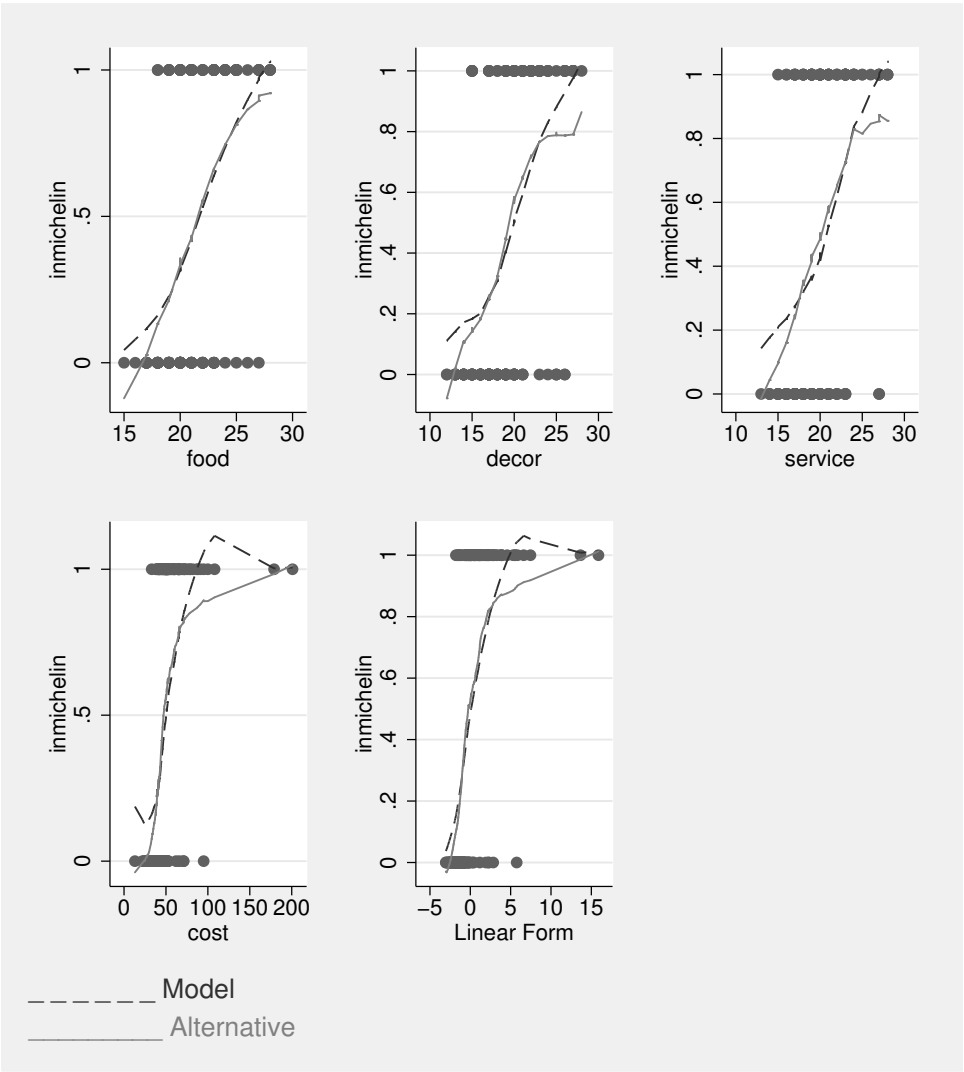

Figure 3. Michelin marginal model plots

Here we specified `mean(pr)`. If we had used `mean(xb)`, we would have estimated the log odds of Michelin inclusion. The marginal model plots for our model indicate that it does not fit particularly well. Note how the model line jumps above the value 1 for the predictor `cost` and the linear form. Our smoother does not truncate its output so that it must fall between 0 and 1. If the lowess estimate of the conditional mean exceeds 1, the excessive number is reported. Here the actual points used to calculate the estimate will of course not exceed 1, but the trend they suggest to the lowess calculation may lead to a surprisingly high value. As detailed in Sheather (2009), after some diagnostic

checks that examine the conditional distributions of the predictors under `inmichelin`, we find a superior model. This superior model is obtained by adding an interaction term for `service` and `decor` and the natural logarithm of `cost` to the model. These marginal model plots indicate a good fit for our model. See figure 4.

```
. generate srvdr = service*decor
. generate lncost = ln(cost)
. logit inmichelin food decor service cost lncost srvdr, nolog
Logistic regression                               Number of obs   =        164
                                                  LR chi2(6)      =      95.97
                                                  Prob > chi2     =     0.0000
Log likelihood = -64.910085                       Pseudo R2       =     0.4250
```

inmichelin	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
food	.6699594	.182764	3.67	0.000	.3117486	1.02817
decor	1.297883	.4929856	2.63	0.008	.3316491	2.264117
service	.9197059	.4882945	1.88	0.060	-.0373338	1.876746
cost	-.0745649	.0441641	-1.69	0.091	-.1611249	.0119951
lncost	10.96401	3.228443	3.40	0.001	4.63638	17.29164
srvdr	-.0655087	.0251228	-2.61	0.009	-.1147485	-.016269
_cons	-70.85311	15.45785	-4.58	0.000	-101.1499	-40.55628

```
. mmp, mean(pr) smoother(lowess) varlist(food decor service cost)
> smooptions(bwidth(.6666667)) linear goptions(xsize(9) ysize(10))
```

(*Continued on next page*)

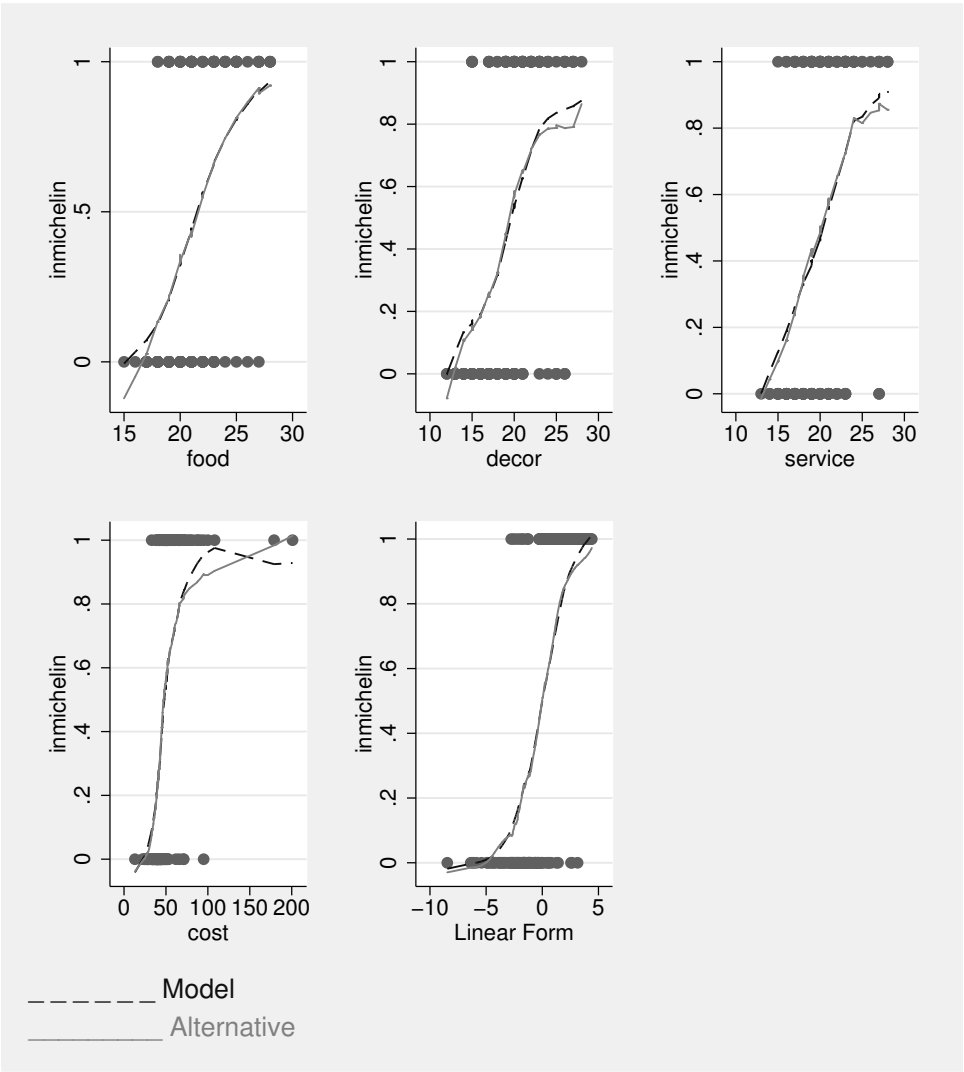

Figure 4. Michelin marginal model plot `srvdr`, log(`cost`)

3 Conclusion

Both the theory and the practice of marginal model plots have been explained in this article. We have demonstrated the use of marginal model plots in both linear and logistic regressions. The `mmp` command was fully defined as a method for using marginal model plots in Stata.

4 References

Cook, R. D., and S. Weisberg. 1997. Graphics for assessing the adequacy of regression models. *Journal of the American Statistical Association* 92: 490–499.

Sheather, S. J. 2009. *A Modern Approach to Regression with R*. New York: Springer.

About the authors

Charles Lindsey is a PhD candidate in statistics at Texas A & M University. His research is currently focused on nonparametric methods for regression and classification. He currently works as a graduate research assistant for the Institute of Science Technology and Public Policy within the Bush School of Government and Public Service. In the summer of 2007, he worked as an intern at StataCorp. Much of the groundwork for this article was formulated there.

Simon Sheather is a professor in and head of the Department of Statistics at Texas A & M University. Simon's research interests are in the fields of flexible regression methods, and nonparametric and robust statistics. In 2001, Simon was named an honorary fellow of the American Statistical Association. Simon is currently listed on http://www.ISIHighlyCited.com among the top one-half of one percent of all mathematical scientists, in terms of citations of his published work.