Regression Analysis V ... More Model Building: Including Qualitative Predictors, Model Searching, Model "Checking"/Diagnostics

The session is a continuation of a version of Section 11.3 of MMD&S. It concerns the practical use of regression analysis in complicated business problems, where there are many many ways that one might think of modeling some important response. In Session 9 we began this discussion with consideration of building "new" predictors and responses from "old" ones. Here we consider

1. how to incorporate qualitative information in a MLR model

2. model "searching" among a large number of possible models for a few that look most promising in terms of effective prediction and simplicity **...** we consider

   - model comparison criteria

   - search "algorithms"

3. model "checking" (diagnostic) tools

   - plots

   - statistics/summary measures

Incorporating Qualitative Information in MLR:"Dummy"/ "Indicator" Predictors

Often there is information the might be used to predict y that is not really quantitative, but rather is qualitative. For example, a response might be different for females than for males. Or ...

Example Consider a hypothetical extension of the Real Estate Example, where one has not only $x_1 = size$ and $x_2 = condition$ to use in the prediction of $y = selling\ price$, but also knows in which quadrant of a city a given home is located. That is, one has available the information as to whether the home is in the NE, NW, SW or SE quadrant of a city, and expects that home prices are different in the 4 quadrants. How might this kind of information be incorporated into a MLR model?

A *naive* (**and ultimately completely unsatisfactory**) way to try to do this would be to invent a predictor variable

$$x_3 = \text{region number}$$

using a coding like

NW $x_3 = 2$	NE $x_3 = 1$
SW $x_3 = 3$	SE $x_3 = 4$

and try to model price as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

But this kind of model is silly. It forces as a model assumption that one gets the same (β_3) change in mean price for homes of a given size and condition as

one moves

$$\begin{aligned} \text{region } \#1 &\rightarrow \text{region } \#2 \\ \text{or region } \#2 &\rightarrow \text{region } \#3 \\ \text{or region } \#3 &\rightarrow \text{region } \#4 \end{aligned}$$

and there is absolutely no reason to *a priori* expect this to be a sensible model assumption. We need to be more clever than this in entering qualitative information into a MLR model. (For what it is worth, this failed attempt is sometimes called the method of "allocated codes.")

A "more clever" way of operating is to use "dummy" (or "indicator") variables as predictors. That is, suppose a qualitative factor A has I possible "levels" or settings. (A could, for example, be something like employee gender with $I = 2$ or city quadrant with $I = 4$.) It is then possible to represent A in MLR

through the creation of $I-1$ dummy variables. That is, one defines

$$x_{\mathsf{A1}} = \begin{cases} 1 & \text{if the observation is from level 1 of A} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{\mathsf{A2}} = \begin{cases} 1 & \text{if the observation is from level 2 of A} \\ 0 & \text{otherwise} \end{cases}$$

$$\vdots$$

$$x_{\mathsf{A},I-1} = \begin{cases} 1 & \text{if the observation is from level } I-1 \text{ of A} \\ 0 & \text{otherwise} \end{cases}$$

Then (for example initially considering a model that includes *only* the factor A

qualitative information), the model

$$y = \beta_0 + \beta_1 x_{\mathsf{A}1} + \beta_2 x_{\mathsf{A}2} + \cdots + \beta_{I-1} x_{\mathsf{A},I-1} + \epsilon$$

says that observations have means

$\beta_0 + \beta_1$	if observation is from level 1 of A
$\beta_0 + \beta_2$	if observation is from level 2 of A
$\vdots$	$\vdots$
$\beta_0 + \beta_{I-1}$	if observation is from level $I-1$ of A
β_0	if observation is from level I of A

Of course all of the MLR machinery is available to do inference for the β's and sums and differences thereof (that amount to means and differences in mean responses under various levels of A).

Example To return to the hypothetical extension of the Real Estate Example, what we are talking about here is the creation of an data table that looks like

y	x_1	x_2	x_{A1}	x_{A2}	x_{A3}	
-	-	-	1	0	0	
⋮	⋮	⋮	⋮	⋮	⋮	NE homes
-	-	-	1	0	0	
-	-	-	0	1	0	
⋮	⋮	⋮	⋮	⋮	⋮	NW homes
-	-	-	0	1	0	
-	-	-	0	0	1	
⋮	⋮	⋮	⋮	⋮	⋮	SW homes
-	-	-	0	0	1	
-	-	-	0	0	0	
⋮	⋮	⋮	⋮	⋮	⋮	SE homes
-	-	-	0	0	0	

What does this do for us? Consider the MLR model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{\text{A1}} + \beta_4 x_{\text{A2}} + \beta_5 x_{\text{A3}} + \epsilon$$

This model says that

- NE home selling prices are modeled as

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 + \epsilon \\ &= (\beta_0 + \beta_3) + \beta_1 x_1 + \beta_2 x_2 + \epsilon \end{aligned}$$

- NW home selling prices are modeled as

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 + \epsilon \\ &= (\beta_0 + \beta_4) + \beta_1 x_1 + \beta_2 x_2 + \epsilon \end{aligned}$$

- SW home selling prices are modeled as

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_5 + \epsilon \\ &= (\beta_0 + \beta_5) + \beta_1 x_1 + \beta_2 x_2 + \epsilon \end{aligned}$$

- SE home selling prices are modeled as

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \end{aligned}$$

That is, the SE quadrant (the Ith level of the qualitative factor) is taken as a baseline. The coefficients $\beta_3, \beta_4, \beta_5$ on the 3 (the $I - 1$) dummy variables shift the mean response up or down as the level of of the qualitative factor changes.

As we've laid it out so far, the basic relationship between mean price and the predictors size and condition is assumed to be the same in the 4 different quadrants. If we wanted to allow, for example, the rate of change of price with respect to size (the per square foot price change) to vary quadrant to quadrant, we could have this by making up **interaction** terms between $size$

and *quadrant dummies*, $x_1 x_{\mathsf{A1}}, x_1 x_{\mathsf{A2}}, x_1 x_{\mathsf{A3}}$. That is, the model could be extended to something like

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_{\mathsf{A1}} + \beta_4 x_{\mathsf{A2}} + \beta_5 x_{\mathsf{A3}} \\ &\quad + \beta_6 x_1 x_{\mathsf{A1}} + \beta_7 x_1 x_{\mathsf{A2}} + \beta_8 x_1 x_{\mathsf{A3}} + \epsilon \end{aligned}$$

The safest way to use this dummy variable idea is to make up the dummies "by hand." (Then one knows exactly what the software is doing.) However, it is possible to enter a single variable like

$$x = \text{quadrant number}$$

taking values $1, 2, \ldots, I$ and essentially get JMP to automatically make up a version of its own dummies (to replace the single variable x). By designating a predictor giving I levels of a factor A as a "nominal" variable (instead of

a "continuous" variable) and entering it as a predictor in a JMP `Fit Model` call, one gets JMP to create and use $I-1$ predictor variables

$$x'_{\mathsf{A1}} = \begin{cases} 1 & \text{if the observation is from level 1 of A} \\ -1 & \text{if the observation is from level } I \text{ of A} \\ 0 & \text{otherwise} \end{cases}$$

$$x'_{\mathsf{A2}} = \begin{cases} 1 & \text{if the observation is from level 2 of A} \\ -1 & \text{if the observation is from level } I \text{ of A} \\ 0 & \text{otherwise} \end{cases}$$

$$\vdots$$

$$x'_{\mathsf{A},I-1} = \begin{cases} 1 & \text{if the observation is from level } I-1 \text{ of A} \\ -1 & \text{if the observation is from level } I \text{ of A} \\ 0 & \text{otherwise} \end{cases}$$

This is a different coding (a $0/1/-1$ coding instead of a $0/1$ coding) of the qualitative information. The model

$$y = \beta_0 + \beta_1 x'_{\mathsf{A}1} + \beta_2 x'_{\mathsf{A}2} + \cdots + \beta_{I-1} x'_{\mathsf{A},I-1} + \epsilon$$

for example says that observations have means

$$\begin{array}{ll}
\beta_0 + \beta_1 & \text{if observation is from level 1 of A} \\
\beta_0 + \beta_2 & \text{if observation is from level 2 of A} \\
\quad\vdots & \qquad\qquad\vdots \\
\beta_0 + \beta_{I-1} & \text{if observation is from level } I-1 \text{ of A} \\
\beta_0 - \left(\sum_{i=1}^{I-1} \beta_i\right) & \text{if observation is from level } I \text{ of A}
\end{array}$$

With this coding, the sum of the means is $I\beta_0$ and thus β_0 is the arithmetic average of the I means. The other β's are then deviations of the "first" $I-1$ means from this arithmetic average of the I means. This coding of the qualitative information produces different interpretations for the β's, but exactly the same fitted values and subject matter inferences as the first one.

Model Searching

This is "enlightened mucking around" in the (typically very large) set of possible MLR models for a given problem, looking for some that are potentially useful (to subject to further scrutiny and then possible application). Qualitatively, one wants a MLR model

- that faithfully reproduces y (has $\hat{y}$ values that look like y's)

- is "small"/simple/uncomplicated

In this second regard, there are at least two motivations. One wants simplicity for purposes of ease of interpretation and use. But in addition, one wants to

avoid "overfitting." It is possible to use very complicated models that "do a lot of wiggling" and can therefore essentially "hit every data point in a data set," but because they are so focused on the data set in hand, really track a lot variation that is "noise" instead of "signal" and thus produce lousy predictions, especially for even slight interpolations or extrapolations.

Example Below is a plot of some data simulated from a SLR model, together with both a straight line and a 6th degree polynomial fit to the data. The 6th degree polynomial is an "overfit" for these data. It has a better R^2 than the fitted line and comes closer to hitting the data points in the data set. But it would be a terrible predictor of y's for any x's except those in the data set.

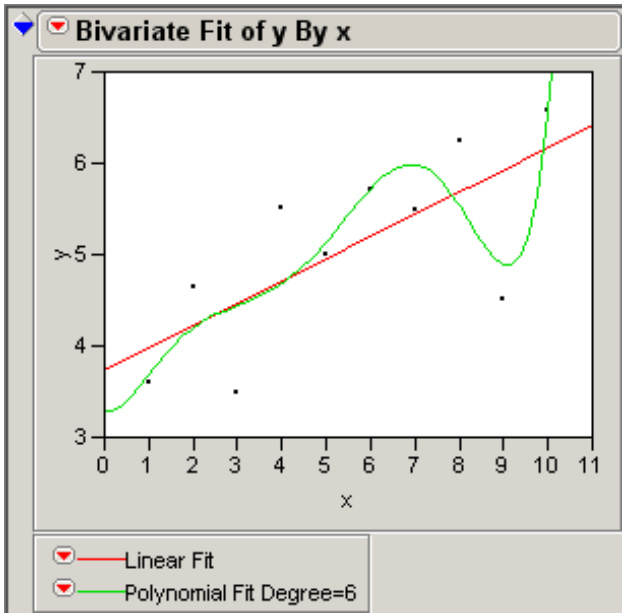

Figure 1: A Line and a 6th Degree Polynomial Fit to Simulated SLR Data

So then, what (numerical) *criteria* might one apply to compare the fits of MLR models? The most obvious ones are

- R^2 (generally, we'd like this to be big)

- s (generally we'd like this to be small)

As it turns out, these criteria can sometimes be in conflict. That is, it is possible to start with a large set of predictor variables for y, and find models (say, #1 and #2) each using some of the predictors, for which

$$R_1^2 > R_2^2 \text{ while } s_1 > s_2$$

AND, there are other popular comparison criteria, including Mallow's C_p and the Akaike Information Criterion AIC.

If one has k possible predictors and is contemplating the use of $p-1$ of those in a MLR model (so that including β_0 there are p regression coefficients β in the contemplated model), Mallows argued that

$$C_p = \frac{SSE_{\text{reduced}}}{MSE_{\text{full}}} + 2p - n$$

should be about p. (Some detail on the rationale for why this should be is given on the "Numerical Model Comparison Criteria" document on the course web page.) If this is substantially larger than p, there is indication that the corresponding model may not be such a good one.

The value of the Akaike Information Criterion for a model with p predictor variables is

$$AIC = n \ln\left(\frac{SSE}{n}\right) + 2\left(p+1\right)$$

and Akaike argued that one would like this to be small.

So, in mucking about in possible MLR models, it is desirable to find a **small model** (with p predictors, say, out of a possible k predictors) that has

big R^2

small s

$\frac{SSE_{\text{reduced}}}{MSE_{\text{full}}} + 2(p+1) - n$ not much (if any) larger than $p+1$

small AIC

It was formerly fashionable to use "forward selection," "backwards elimination," and "stepwise selection" algorithms to wander about in the set of all possible models, comparing R^2 values. JMP allows a "manual" version of this wandering using the `Stepwise` "personality" for `Fit Model`. One may add in and drop out predictors and immediately see the effects of doing so on $R^2, s, C_p,$

and AIC. (Formerly, more or less automatic versions of this were popular, based on F-test criteria of whether or not to add in or drop out predictors.)

For a given number of predictor variables, p, the above criteria all point to the same model as being most attractive. So the most important thing one could have for model searching purposes is a list of the best (biggest R^2) models of each "size." JMP will produce that automatically after running a `Stepwise` fit if one clicks on the red triangle and highlights `All Possible Models` . A list of R^2 values for *all* possible models is produced. If there are not too many predictors, one may simply sort through those directly. For large numbers of predictors, some way of paring down the list is needed. There is a text file (from SAS) on the course web page describing how to quickly list the single best model of each possible size. Basically, one right clicks on the `All Possible Models` output and saves it to a table. A new column is added to that table pasting in the following formula for the new column

$$\texttt{If}(\texttt{Row}() == 1, 1, \texttt{If}(: \texttt{Number}[\texttt{Row}()]! =: \texttt{Number}[\texttt{Row}() - 1], 1, 0))$$

That produces a column with 1's identifying the best model of each size. One can then select rows of the table with best models and make a new smaller table out of them only.

Model Checking/Diagnostics

Model searching methods give tools for screening a huge number of possible models down to a few for very careful examination, to see whether the normal MLR model is sensible. This careful examination is important, because all of the inference formulas we've studied depend for their reliability on the appropriateness of that model. Careful examination of a candidate model goes under the name "model checking" and/or "regression diagnostics."

The basic tool of diagnostics is the notion of residuals. Recall that we have used the notation

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= \text{the residual for the } i\text{th case} \end{aligned}$$

These are in some sense "what is left over or residual or unexplained" after one has fit a model. They are empirical approximations for the "errors" in the MLR model, the

$$\begin{aligned} \epsilon_i &= y_i - \mu_{y|x_{1i},x_{2i},\ldots,x_{ki}} \\ &= y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}) \end{aligned}$$

that the MLR model says are random draws from a normal distribution with mean 0 and standard deviation σ. So one hopes that residuals look like "noise," i.e. like (normal) random variation. One hopes (that a MLR model is sensible and therefore) that residuals

1. have a reasonably bell-shaped distribution (a bell-shaped histogram)

2. are essentially "patternless" when plotted against essentially any meaningful variable

JMP allows the saving of residuals from both `Fit Y by X` and `Fit Model` that then allows the examination of 1. and 2. Where 1. or 2. fails, there is indication that the modeling one has done is not sufficient and/or there are some aberrant cases in the data set (that suggest that while the model might be mostly sensible, its utility may not extend to the entire spectrum of situations represented in the data).

Below are some examples of kinds of plots typically made and patterns one might see.

Example Here are first a plot of some (x, y) data (simulated from a quadratic regression model) together with a fitted line, then a plot of e_i versus x_i. The positive/negative/positive pattern on the plots of residuals is indicative that y does not change linearly with x ... some curvature in x is needed to describe the relationship.

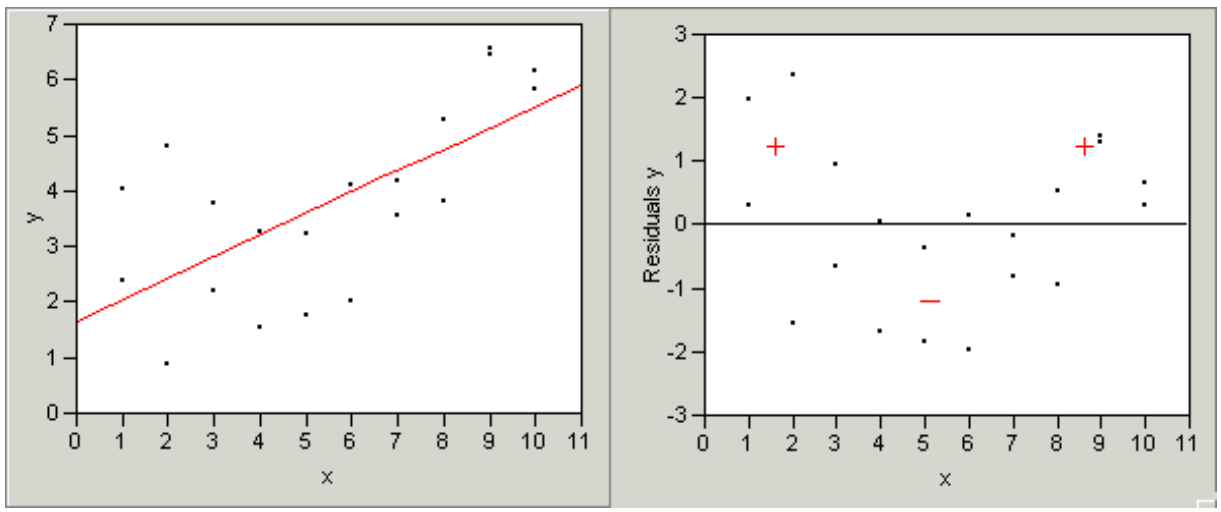

Figure 2: A Fitted Line and a Corresponding Plot of Residuals Against x

Common "residual plots" (for a MLR fit involving $x_1, x_2, \ldots, x_k$) are made for e_i versus each of

$$x_{1i}, x_{2i}, \ldots, x_{ki}, \text{ and } \hat{y}_i$$

Patterns on any of these plots indicates that something has been missed in the modeling.

Example JMP's standard output for `Fit Model` automatically includes a plot of e_i versus $\hat{y}_i$ (as a `Residual by Predicted Plot`). Lab #5 includes a problem where there is clear curvature on that plot. A residual plot like the one below indicate that the MLR model under-predicts both small and large y and over-predicts moderate values of y.

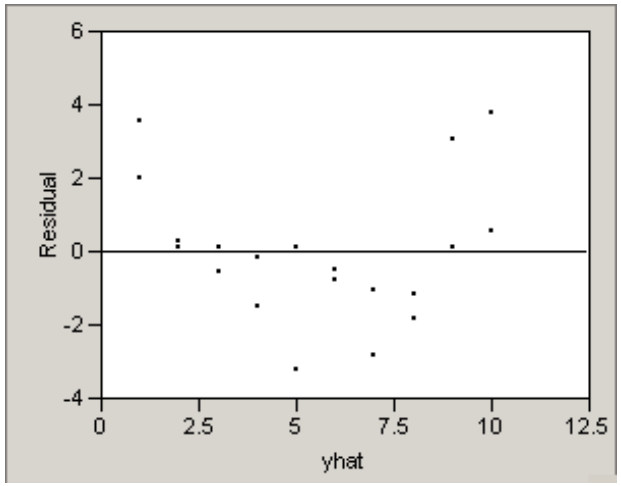

Figure 3: Hypothetical Plot of Residual Against Predicted

Example Below is a residual plot that indicates that variability in response increases with mean response ... something not allowed for in the usual (constant σ) MLR model.

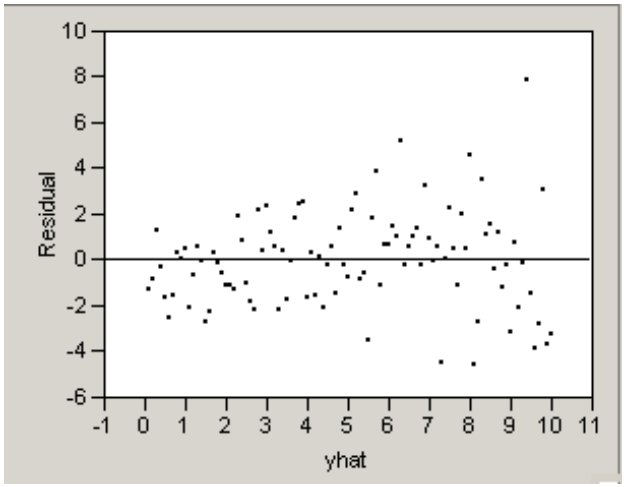

Figure 4: Residual Plot Indicating Problems with the "Constant σ" Part of the MLR Model Assumptions

Plotting residuals against "another variable" (one not already included in a regression model) can indicate (via an obvious pattern on the plot) that the variable should be included in the modeling in some fashion.

Example Consider a hypothetical extension of our basic Real Estate Example, where it turns out that homes $2, 3, 5$ and 10 are all located on one side (say East) of a town and the others are on the other side of town (West). The residuals for the 2-variable regression model

$$\begin{aligned} \hat{y} &= 9.78 + 1.87x_1 + 1.28x_2 \\ &= 9.78 + 1.87size + 1.28condition \end{aligned}$$

are negative for cases $2, 3, 5$ and 10 and positive for the other 6 cases. (The fitted model over-predicts on the East side and under-predicts on the West side.) In fact if we call the East side of town "Side #1" and the West "Side #2," a plot of residual against a "side of town" variable looks like the figure below, and clearly indicates the need to include the side of town information in modeling.

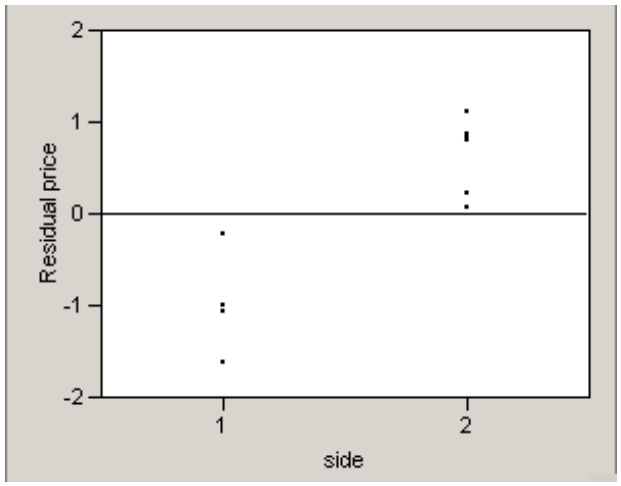

Figure 5: Residual versus "Side of Town"

The residuals $e_i = y_i - \hat{y}_i$ are "ordinary" residuals, and have the units of the response variable, y. There are other important types of residuals. For one, it is common to compute and plot **standardized (or "Studentized") residuals**

$$e_i^* = \frac{e_i}{\mathrm{SE}_{e_i}}$$

(where there is no "by hand" formula for the standard error of e_i). These are unitless and are typically between -2 and 2 (or at least -3 and 3). Big standardized residuals can flag potential recording errors, "odd" cases, and potentially "highly influential" data points. JMP `Fit Model` will allow a user to save both standard errors for the ordinary residuals (that vary case to case) and the e_i^*.

Other important residuals are the **deleted residuals**. Define

$$\hat{y}_{(i)} = \text{the value of the response predicted for case } i \text{ when the MLR model is fit using only the other } n-1 \text{ cases}$$

Then the ith deleted residual is

$$e_{(i)} = y_i - \hat{y}_{(i)}$$

One hopes that the deleted residuals $e_{(i)}$ (computed where the ith case is not in the data set used to predict its response) are "not too much bigger than" the ordinary residuals e_i. A means of summarizing the size of these deleted residuals is the so called "prediction error sum of squares" or "PRESS statistic"

$$PRESS = \sum e_{(i)}^2 = \sum \left(y_i - \hat{y}_{(i)}\right)^2$$

It is the case that the prediction error sum of squares is always at least as big as the error sum of squares

$$PRESS \geq SSE = \sum (y_i - \hat{y}_i)^2$$

and informally seeing how $PRESS$ compares to SSE is a way of judging how sensitive the model fit is to the exact values of the individual responses in

the data set. JMP will add the value of $PRESS$ to the `Fit Model` report through use of the `Row Diagnostics` item under the top red triangle on the report.

There are also **partial residuals**. These are the basis of JMP "effect leverage plots" that are part of the standard `Fit Model` report and that can be very useful for detecting mulitcollinearity. There is a discussion of these plots in the "Regression Diagnostics" document on the course web page. Because these are more subtle than the other diagnostics we're discussing here (are a somewhat more advanced topic) we'll not discuss them further here.

The final two diagnostic tools that we will consider are useful for identifying important or influential cases in a MLR. This is important, as one may not want to use a fitted model whose character is strongly dependent upon a few data cases ... or at least if such a model is to be used, it's important to know

that this is what is going on. These two new tools are based on what JMP calls **hats**. (In the non-SAS world, these are more often called "leverage values.")

The origin of the concept of hats is the mathematical fact:

> In a given MLR there are $n \times n$ constants h_{ij} so that the fitted values can be written as
>
> $$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{ii}y_i + \cdots + h_{in}y_n$$

So h_{ii} in some sense measures how important y_i (or the ith case) is to its own prediction. We'll call

$$h_{ii} = \text{the } i\text{th hat value}$$

It turns out that each

$$0 \leq h_{ii} \leq 1$$

and that

$$\sum h_{ii} = k + 1$$

This last fact implies that the hats average to $\frac{k+1}{n}$, and a corresponding common rule of thumb is that any case with

$$h_{ii} > 2\frac{k+1}{n}$$

is flagged as a "high leverage"/potentially highly influential case.

Large h_{ii}'s flag data points that are "near the edges" of the data set **as regards the values of the predictors** $x_1, x_2, \ldots, x_k$. (Hats have nothing to do with the observed responses.) We might want to consider some kind of measure

that takes into account the the observed responses. **Cook's Distance** is such a measure. Cook's distance for case i in a MLR is

$$\begin{aligned} D_i &= \frac{h_{ii}}{(k+1)\,MSE}\left(\frac{e_i}{1-h_{ii}}\right)^2 \\ &= \left(\frac{h_{ii}}{k+1}\right)\left(\frac{e_{(i)}}{s}\right)^2 \end{aligned}$$

To have a large value of Cook's Distance, a case must have *both* a large hat (must be near the "edge" of the data set) and have a big deleted residual (must be poorly predicted by a model fitted with using it in the fitting). JMP's `Fit Model` routine will allow one to save both hats and Cook's Distance values to the data table through use of the `Save Columns` item under the top red triangle on the report.

Example Below are an augmented data table for the Real Estate Example (including hats and Cook's Distances) and a plot of $(size, condition)$ data

pairs. Notice that Cases 2 and 10 have by far the largest hat values ... they are at the "edge" of the data set as regards $size$ and $condition$. But only case 10 has a large Cook's Distance. Only case 10 has a large (deleted) residual, so only case 10 is both on the edge of the data set and is poorly predicted by a model fit without including it in the fitting. (The practical upshot of this analysis is that one might want to be very cautious about believing predictions of prices for relatively large homes in relatively poor condition. One might even want to refit the model without using case 10 and see how much predictions for the other 9 cases change!)

	size	condition	price	Predicted price	Residual price	Studentized Resid price	h price	Cook's D Influence price
1	23	5	60	59.2044859	0.79551406	0.81816775	0.19029818	0.05244124
2	11	2	32.7	32.9188402	-0.2188402	-0.3109762	0.57585529	0.04376556
3	20	9	57.7	58.7042432	-1.0042432	-1.1510125	0.34802367	0.23573051
4	17	3	45.5	45.4225927	0.0774073	0.08045577	0.20720018	0.00056392
5	15	8	47	48.071426	-1.071426	-1.1939696	0.31031306	0.21380276
6	21	4	55.3	54.1844746	1.1155254	1.12411493	0.15656503	0.07818858
7	24	7	64.5	63.6317028	0.86829722	0.94622255	0.27878568	0.1153643
8	13	6	42.6	41.7732739	0.82672608	0.90120292	0.27923746	0.10488307
9	19	7	54.5	54.2770264	0.22297362	0.22402329	0.15153341	0.00298771
10	25	2	57.5	59.1119342	-1.6119342	-2.1143248	0.50218805	1.50322223

Figure 6: Augmented Real Estate Example Data Table for MLR

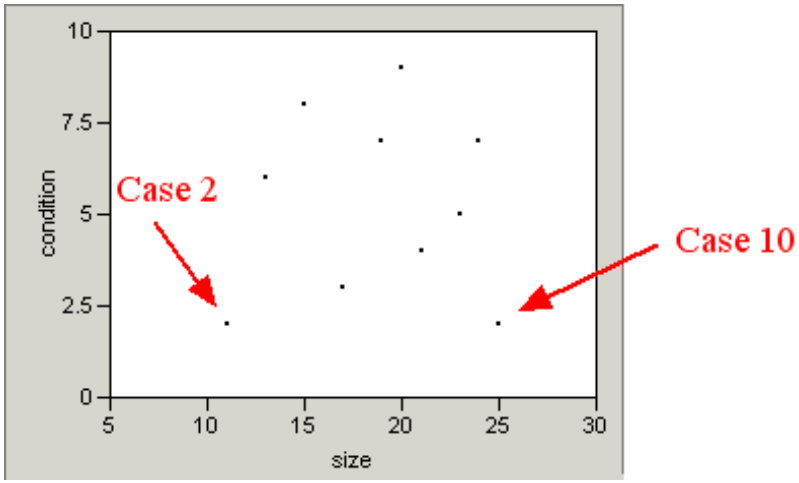

Figure 7: Scatterplot of $condition$ versus $size$ for the Real Estate Example