

Regression Analysis IV ... More MLR and Model Building

This session finishes up presenting the formal methods of inference based on the MLR model and then begins discussion of "model building" (use of regression analysis in the practical description of complicated real world relationships ... a version of what is in Section 11.3 of MMD&S).

1

t Tests (for Individual Coefficients) in MLR

It is possible to test $H_0: \beta_j = \#$ in the MLR model. A t statistic for this is

$$t = \frac{b_j - \#}{SE_{b_j}}$$

(where both b_j and SE_{b_j} must come from a regression program). p -values come from the t distribution with $df = n - k - 1$. The most common version of this test is the $\# = 0$ version. The hypothesis $H_0: \beta_j = 0$ is that $\mu_{y|x_1, x_2, \dots, x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$ doesn't really depend upon x_j . That is, it says that the j th predictor variable can be dropped from the prediction equation without detectable loss in ability to predict/explain y . Regression analysis programs like JMP's Fit Model routine typically print out

$$t = \frac{b_j - 0}{SE_{b_j}} = \frac{b_j}{SE_{b_j}}$$

2

and two-sided p -values for testing these hypotheses.

Example In the Real Estate Example with

$$y = \text{price}, x_1 = \text{size}, \text{ and } x_2 = \text{condition}$$

for testing $H_0: \beta_1 = 0$, the observed value of the test statistic is

$$t = \frac{b_1}{SE_{b_1}} = \frac{1.87}{.076} = 24.56$$

and the p -value is

$$P(\text{a } t \text{ r.v. with } df = 7 \text{ is } > 24.56 \text{ or is } < -24.56)$$

This is tiny, smaller than 2(.0005) according to the t table in MMD&S (the upper .0005 point of that t distribution is 5.408). Similarly, for testing $H_0: \beta_2 = 0$, the observed value of the test statistic is

$$t = \frac{b_2}{SE_{b_2}} = \frac{1.28}{.144} = 8.85$$

3

and the p -value is

$$P(\text{a } t \text{ r.v. with } df = 7 \text{ is } > 8.85 \text{ or is } < -8.85)$$

which is again smaller than 2(.0005).

In this example, BOTH $H_0: \beta_1 = 0$ and $H_0: \beta_2 = 0$ are implausible on the basis of our data analysis. Both x_1 and x_2 are helpful in predicting/explaining y . (Notice, by the way, that this was already obvious from the confidence intervals for β_1 and β_2 made in the last session, as neither of those include 0.)

The location of these test statistics and p -values on a JMP report is shown below.

4

The screenshot shows the 'Parameter Estimates' section of a JMP MLR report. It includes a table with columns for Term, Estimate, Std Error, t Ratio, Prob > |t|, Lower 95%, and Upper 95%. Red arrows point to the 't Ratio' and 'Prob > |t|' columns, with a label 'p-values for two-sided alternatives' pointing to the probability values.

Term	Estimate	Std Error	t Ratio	Prob > t	Lower 95%	Upper 95%
Intercept	9.7822706	1.630481	6.00	0.0005	5.9267965	13.637745
size	1.8709353	0.076174	24.56	<.0001	1.6908134	2.0510572
condition	1.2781408	0.1444	8.85	<.0001	0.9366883	1.6195933

Figure 1: JMP MLR Report and t Tests of $H_0:\beta_1 = 0$ and $H_0:\beta_2 = 0$

Regarding the interpretation of tests for individual coefficients, be aware that testing $H_0:\beta_j = 0$ in a MLR is *not equivalent* (either numerically or conceptually) to testing $H_0:slope = 0$ in a SLR involving only x_j . Testing the first of these is a way of asking

"In the presence of the other predictors, can I do without x_j ?"

Testing the second of these is a way of asking

"If all I have to predict with is x_j , can I do without it?"

This distinction is an important one. If there are other predictors that are correlated with x_j , the answer to the first of these questions can be "YES!"

while simultaneously the answer to the second is "NO!" This point is strongly related to the concept of "multicollinearity" discussed more fully later in this session.

The "Overall F Test"/"Model Utility Test" in MLR

The normal MLR model supports the testing of the hypothesis $H_0:\beta_1 = \beta_2 = \dots = \beta_k = 0$. This is done using an F statistic and the calculations are typically organized in an ANOVA table for MLR. The MLR model says that

$$\mu_{y|x_1, x_2, \dots, x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

so that under this model, the hypothesis means that

$$\mu_{y|x_1, x_2, \dots, x_k} = \beta_0$$

and the mean response doesn't depend upon *any of the predictors* $x_1, x_2, x_3, \dots, x_k$. Testing this hypothesis is often treated as a way of asking

"Is there somewhere in the predictors $x_1, x_2, x_3, \dots, x_k$ a detectable amount of predictive power for y ?"

Making this test can be thought of as a way to attach a p -value to the coefficient of determination for the MLR, R^2 .

In Session 8 we learned what is meant by SSE and $SSR = SSTot - SSE$ in MLR (and were reminded that $SSTot$ has nothing to do with whether or not predictor variables are being employed). These can be placed into an ANOVA table that looks very much like the one for SLR. That is, the basic ANOVA table for MLR is as below.

Source	SS	df	MS	F
Regression	SSR	k	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$
Error	SSE	$n - k - 1$	$MSE = \frac{SSE}{n-k-1}$	
Total	$SSTot$	$n - 1$		

Just as in SLR, if a fitted MLR model is effective, SSE will be small, SSR will be large, and F will be large. That is, large observed F count as evidence against $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$. p -values will come from right tail areas for the F distribution with $df_{num} = k$ and $df_{denom} = n - k - 1$. Just as in SLR, note that the entries in this table can also produce $R^2 = SSR/SSTot$ and that for MLR, $s = \sqrt{SSE/(n - k - 1)}$ so that in the present notation

$$MSE = s^2$$

9

Example In the Real Estate Example with

$$y = price, x_1 = size, \text{ and } x_2 = condition$$

for testing $H_0: \beta_1 = \beta_2 = 0$, the basic ANOVA table is

Source	SS	df	MS	F
Regression (size and condition)	819.33	2	409.66	350.9
Error	8.17	7	1.17	
Total	827.50	9		

The observed value

$$F = 350.9$$

is to be compared to tabled percentage points for the F distribution with 2 and 7 degrees of freedom. The p -value here is tiny. (Remember that $R^2 = .99$

10

... this is huge.) The data provide definitive evidence that together x_1 and x_2 provide a detectable level of explanatory power for describing y .

Note too that (just as expected) $R^2 = .99 = 819.33/827.50$ and that $s = 1.08 = \sqrt{1.17}$.

The JMP version of the basic ANOVA table and the overall F test is shown below.

11

The screenshot shows the JMP software interface for a regression analysis. The 'Response price' window is open, displaying the following information:

- Whole Model**
 - Actual by Predicted Plot
 - Summary of Fit

RSquare	0.990123
RSquare Adj	0.987301
Root Mean Square Error	1.080545
Mean of Response	51.73
Observations (or Sum Wgts)	10
 - Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	819.32795	409.664	350.8665
Error	7	8.17305	1.168	Prob > F
C. Total	9	827.50100		<.0001

Figure 2: JMP Report Showing the Basic ANOVA Table and Overall F Test for the Real Estate Example

12

Exercise For the augmented version of the small fake data set, find t statistics for testing $H_0:\beta_1 = 0$ and then $H_0:\beta_2 = 0$ on a handout MLR printout. Then use earlier hand work (that produced $SSE = .4$ and $SSR = 15.6$ for MLR starting from the fitted equation) and make the ANOVA table and F statistic for testing $H_0:\beta_1 = \beta_2 = 0$.

Multicollinearity

It is an important matter of interpretation to realize that using an F test to test $H_0:\beta_1 = \beta_2 = 0$ in a MLR model with 2 predictors x_1 and x_2 is NOT at all the same as separately testing (e.g. using two t tests) $H_0:\beta_1 = 0$ and then $H_0:\beta_2 = 0$ in that MLR model! It is quite possible to *simultaneously* obtain

13

- a tiny p -value for testing $H_0:\beta_1 = \beta_2 = 0$
- large p -values for testing BOTH
 - $H_0:\beta_1 = 0$
 - $H_0:\beta_2 = 0$

This kind of thing can happen when *predictors* in MLR are correlated. The standard jargon for this kind of circumstance is **multicollinearity**. Multicollinearity is extremely common in practice and creates all kinds of headaches of interpretation for MLR.

Example Consider the small fake example of $n = 5$ cases of (x_1, x_2, y) in the table below.

14

x_1	x_2	y
.9	2.1	3.9
2.1	3.9	6.9
2.9	5.9	9.9
4.1	7.9	13.1
5.1	10.1	15.9

It's easy enough to see from the data table that

$$x_2 \approx 2x_1$$

and

$$y \approx 1 + 3x_1$$

$$\approx 1 + 1.5x_2$$

Actually doing regression analyses with these "data" produces

15

- For SLR on x_1

$$\hat{y} = 1.194 + 2.89x_1$$

$$R^2 = .9972$$

$$t = 33 \text{ for testing } H_0 : \beta_1 = 0 \text{ (in SLR) with a tiny } p\text{-value}$$
- For SLR on x_2

$$\hat{y} = .923 + 1.508x_2$$

$$R^2 = .9979$$

$$t = 38.4 \text{ for testing } H_0 : \beta_2 = 0 \text{ (in SLR) with a tiny } p\text{-value}$$
- For MLR on x_1 and x_2

$$\hat{y} = 1.03 + 1.239x_1 + .864x_2$$

16

$$R^2 = .9989$$

$$F = 878.3 \text{ (} p\text{-value } .0011 \text{ for testing } H_0 : \beta_1 = \beta_2 = 0 \text{)}$$

$$t = 1.26 \text{ for testing } H_0 : \beta_1 = 0 \text{ (in MLR) with } p\text{-value of } .3360$$

$$t = 1.68 \text{ for testing } H_0 : \beta_2 = 0 \text{ (in MLR) with } p\text{-value of } .2344$$

In contexts like this (where predictors are correlated) I can't really separate their effects on y and need to be cautious in my interpretation of b 's. Geometrically, I have something going on like what is pictured below.

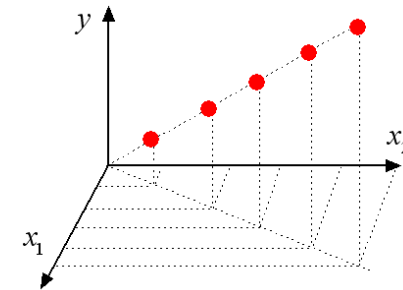


Figure 3: Cartoon Illustrating Multicollinearity

x_1 and x_2 are strongly linearly related to each other, y is strongly linearly related to either/both, so that R^2 is large and one can effectively predict y from x_1 , from x_2 , or from the pair (x_1, x_2) . BUT there are MANY planes that do about the same job of running through the points "plotted" in 3 dimensions ... since they come close to falling on a line. EXACTLY which plane (which set of β 's) is really appropriate is not clear from the data ... one can't really separate the effects of the two predictors. Further, when one does try to pick out a plane containing the points, since that plane is very poorly determined, one can predict effectively for data like the ones in hand, but extrapolation becomes extremely unreliable! Further the whole standard interpretation of coefficients as "rates of change of response with respect to one predictor with the others held fixed" breaks down, to the extent that the evidence in the sample is that one may not be able to change one predictor without a corresponding change in others!

Partial F Tests in MLR

There is another type of F test associated with MLR, called a "partial F test." MMD&S do something equivalent to this in terms of R^2 values rather than sums of squares. See their page 661.

The partial F tests (or comparison of R^2 values per MMD&S) is a way of comparing a **Full Model**

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k + \epsilon$$

to some **Reduced Model** involving $l < k$ predictors

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_lx_l + \epsilon$$

It is a way of asking

"Is the reduced model adequate to predict/explain y ? Or is there some additional predictor or predictors (in the full model) that is/are needed in order to effectively describe y ?"

Since it is always the case that $SSR_{full} \geq SSR_{reduced}$, it is similarly true that $R_{full}^2 \geq R_{reduced}^2$, and a partial F test can be thought of as a way of attaching a p -value to the increase in R^2 associated with moving from the reduced model to the larger full model.

Formally, the partial F test is a test of $H_0: \beta_{l+1} = \beta_{l+2} = \dots = \beta_k = 0$ in the MLR model. Calculations are usually organized in an expanded version of the basic MLR ANOVA table. This expanded table is illustrated below.

(Expanded) ANOVA Table (for MLR)

Source	SS	df	MS	F
Regression	SSR_{full}	k		
x_1, \dots, x_l	SSR_{red}	l		
$x_{l+1}, \dots, x_k x_1, \dots, x_l$	$SS = SSR_{full} - SSR_{red}$	$k - l$	$MS = \frac{SS}{k-l}$	$F = \frac{MS}{MSE_{full}}$
Error	SSE_{full}	$n - k - 1$	MSE_{full}	
Total	$SSTot$	$n - 1$		

Example As a completely hypothetical example, suppose that a data set containing $n = 50$ cases has $k = 5$ predictors and one is interested in the two models

$$\text{Full Model: } y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_5x_5 + \epsilon$$

and

$$\text{Reduced Model: } y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \epsilon$$

Here $l = 2$ and one is asking whether predictors x_1 and x_2 together are sufficient to predict y , or whether one or more of x_3, x_4, x_5 are additionally important to explaining/predicting y . The partial F test is a test of $H_0: \beta_3 = \beta_4 = \beta_5 = 0$ in the full model.

Suppose for sake of illustration that $SSTot = 100$, the full model has $SSR_{full} = 75$ and the reduced model has $SSR_{reduced} = 60$. This is a case where the reduced model has $R_{red}^2 = 60/100 = .6$ and the full model has $R_{full}^2 = .75$. A partial F test is a way of assigning a p -value to the increase in R^2 and thereby determining whether there is detectable additional predictive power in the last 3 predictors after that of the first 2 has been accounted for. The appropriate ANOVA table is then

(Expanded) ANOVA Table

Source	SS	df	MS	F
Regression	75	5		
x_1, x_2	60	2		
$x_3, x_4, x_5 x_1, x_2$	$15 = 75 - 60$	$3 = 5 - 2$	$5 = \frac{15}{3}$	$F = \frac{5}{25/44}$
Error	25	44	$\frac{25}{44}$	
Total	100	49		

and the observed value

$$F = \frac{5}{25/44} = 8.8$$

is to be compared to tabled F percentage points for degrees of freedom 3 and 44.

Example In the Real Estate Example with

$$y = \text{price}, x_1 = \text{size}, \text{ and } x_2 = \text{condition}$$

we have seen that $R_{full}^2 = .99$, while for SLR on size alone, we had $R_{reduced}^2 = .88$. I might wish to ask whether this increase in R^2 is statistically significant. This can be approached by testing $H_0: \beta_2 = 0$ in the full model ... which we have previously done using a t test. It is also possible to use (what turns out in this case to be an equivalent) partial F test. (When a full model has exactly one more x than a reduced model, the partial F test comparing full and reduced models is the same as a two-sided t test for that coefficient in the full model.)

Recall for this example that $SSTot = 827.5$, $SSR_{full} = 819.33$ from earlier this session, and $SSR_{red} = 727.85$ from Session 8. Then the expanded ANOVA table here is

(Expanded) ANOVA Table

Source	SS	df	MS	F
Regression	819.33	2		
x_1	727.85	1		
$x_2 x_1$	91.48	1	91.48	78.2
Error	8.17	7	1.17	
Total	827.50	9		

and $F = 78.2$ is to be compared to tabled F percentage points for 1 and 7 degrees of freedom.

Note too that earlier this session we had an observed value of $t = 8.85$ for testing this hypothesis. Except for some rounding error, $78.2 = (8.85)^2$ and the two ways of testing this hypothesis are equivalent.

Exercise The Summer 2003 Stat 328 regression exam posted on the course web site is based on data from the Ames City Assessor for the prices at which

$n = 88$ homes fitting a particular description sold in Ames over a 1 year period. Not 2, but rather $k = 14$ different explanatory variables were available for model building. A 14-predictor model had $R^2 = .7516$ while a smaller model with only $l = 8$ of those predictors had $R^2 = .7286$. The total sum of squares (for prices in thousands of dollars) was $SSTot = 121,386$. Make up an expanded ANOVA table and the partial F statistic for investigating whether the 8-predictor model is adequate as a description of price.

Model Building ... Creating "New" Predictors and Responses

At first glance, the multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

may look fairly specialized and of limited applicability. But because there are many ways that one might *reexpress a predictor x_j or the response y* , and many ways to take several existing predictors and create new ones, the MLR model is really quite flexible and widely useful.

Lab 4 provides an effective example of the first of these points. A SLR model for

$$x = \text{length of stay and } y = \text{reimbursed cost}$$

didn't seem appropriate. But the lab led you through the effective use of a SLR model for the transformed variables $x' = \sqrt{x}$ and $y' = \sqrt{y}$. Reexpressing x and y on square root scales makes the SLR model useful. One isn't limited to using the scales on which data x_j and/or y first come to one. Transformations are possible and multiply the applicability of the MLR model.

Another kind of application of this basic notion is "polynomial regression." That is, instead of just fitting a SLR model

$$y = \beta_0 + \beta_1x + \epsilon$$

to a set of (x, y) data (and beginning with a "straight line relationship" between x and mean y), a more complicated ("curved line") relationship like

$$y = \beta_0 + \beta_1x + \beta_2x^2 + \epsilon$$

or even

$$y = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + \epsilon$$

might be of interest. Notice that by calling

$$\begin{aligned} x_1 &= x \\ x_2 &= x^2 \\ x_3 &= x^3 \end{aligned}$$

29

we can think of the problem of polynomial regression as a special case of MLR. In order to fit a quadratic relationship between y and x in JMP, I simply make up a column of the squares of x and use Fit Model with predictors x and x^2 . Alternatively, I can make use of a special facility in Fit Y by X that fits and plots polynomials directly. These two possibilities are illustrated below for a small set of artificial data.

30

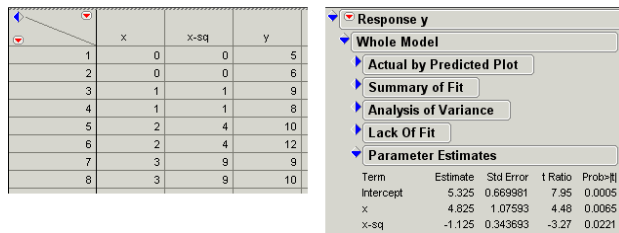


Figure 4: JMP Data Table and Fit Model Report for a Quadratic Regression of y on x

31

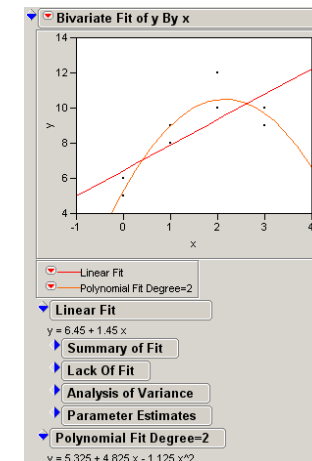


Figure 5: JMP Fit Y by X Report for the Quadratic Regression of y on x

32

Once one begins to consider making "new" variables from "old" ones, the possibility arises of "combining" more than one variable into a single additional predictor. This is a useful idea, but one needs to understand some qualitative implications of doing this. A most important one is the idea of "interaction." If I start with k independent/genuinely different predictors and make up some new ones from the original set, as long as in making up any new predictor, I use only one of the originals at a time, the model is said to be "additive" in the original predictors ... the model doesn't involve "interaction" effects of the original predictors. This means that the model hypothesizes "parallel response traces." But when two or more predictors are combined to make up new variables, the model allows for interactions and non-parallel response traces.

Example Consider two generalizations of the simple MLR model

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \epsilon$$

33

the first where I add a new predictor made up as the square of x_1 alone, i.e.

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1^2 + \epsilon$$

and the where I add a new predictor made up as the product of x_1 and x_2 , i.e.

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2 + \epsilon$$

Below is a graphic comparing what these 3 models say about how mean y varies with the predictors.

34

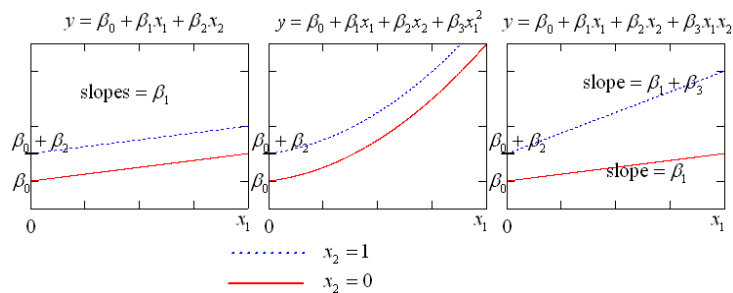


Figure 6: Three Models for Mean Response

35

Notice that in the first two of the panels, the two traces of mean response (for $x_2 = 0, 1$) are *parallel* (there are no "interaction" effects ... changing x_1 produces the same change in y no matter what be the value of x_2). The difference between the first and second panels is that the squared term for x_1 in the second form of mean response allows for curvature in the relationship between x_1 and y for any given x_2 . The third panel is an illustration of a situation involving interaction. The the two traces of mean response (for $x_2 = 0, 1$) are *not parallel* (there are "interaction" effects ... changing x_1 produces different changes in y for different values of x_2). This phenomenon can be traced to the fact that the third form of the relationship between y and x_1 and x_2 involves a cross-product term (x_1x_2) ... a predictor made up by "combining" x_1 and x_2 .

36