Department of Statistics
The Wharton School
University of Pennsylvania

Statistics 621 Fall 2003

Module 5

Further Aspects of Multiple Regression

The ANOVA Table

All statistics programs including JMP provide an ANOVA (Analysis of Variance) table. This table includes the *F* ratio and p-value for testing the hypothesis

$$H_0: \beta_1 = 0, \ldots, \beta_K = 0$$

What does this hypothesis imply about the relationship between y and $x_1, \ldots, x_K$?

The *F* ratio is obtained as

$$F = \frac{Model\ SS / K}{Residual\ SS / (n-K-1)} = \frac{R^2/K}{(1-R^2)/(n-K-1)}$$

Large values of *F* and small *p*-values provide evidence against H_0. A useful rule of thumb[1] is to reject H_0 at the .05 level whenever $F > 4$.

However, it is easier and more accurate to use the familiar p-value strategy: If the p-value < .05, then H_0: $\beta_1 = 0, \ldots, \beta_K = 0$ can be rejected at the .05 level of significance.

[1] Use this rule if you do not have a p-value handy. This rule is "conservative": any time the F > 4, the p-value < 0.05. However, there are some cases in which the F < 4 but the p-value is less than 0.95 (p < 0.05).

Example (car89.jmp)

The multiple regression of GP1000M on Weight, Horsepower, Cargo and Seating yields

Response GP1000M

Summary of

RSquare	0.852239
RSquare Adj	0.846556
Root Mean Square Error	3.411697
Mean of Response	47.67511
Observations (or Sum Wgts)	109

Analysis of

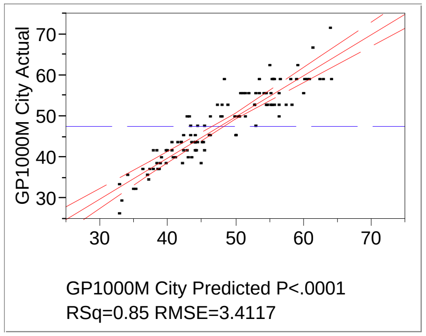
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	6981.9348	1745.48	149.9598
Error	104	1210.5264	11.64	Prob > F
C. Total	108	8192.4611		<.0001

Parameter

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	12.930547	2.020835	6.40	<.0001
Weight(lb)	0.0091318	0.001159	7.88	<.0001
Horsepower	0.0857712	0.01509	5.68	<.0001
Cargo	0.0346363	0.013277	2.61	0.0104
Seating	-0.476467	0.412437	-1.16	0.2506

What should we conclude from the ANOVA table?

The y vs $\hat{y}$ plot confirms this conclusion.



The F Test and Correlated Predictors

The ANOVA test comes in handy when, as usual, the predictors in a regression are correlated. The following example illustrates an extreme case.

Example: A Market Model [2]

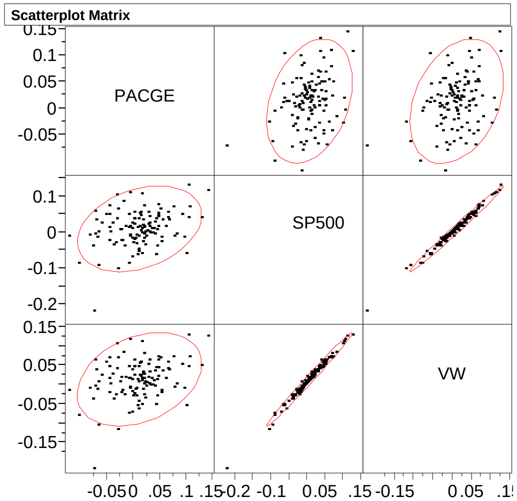
The file stocks.jmp contains monthly returns from 2/78 to 12/87 of VW, SP500, IBM, PACGE and Walmart. Let's focus on the relationship between PACGE, SP500 and VW.

Multivariate

Correlations

	PACGE	SP500	VW
PACGE	1.0000	0.3116	0.3249
SP500	0.3116	1.0000	0.9932
VW	0.3249	0.9932	1.0000

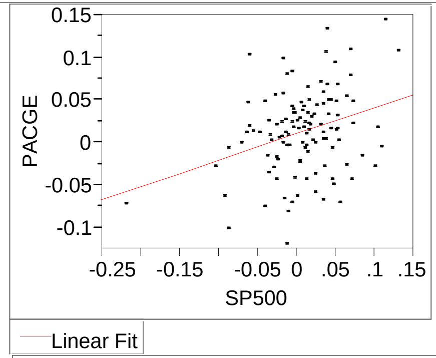
Scatterplot Matrix



[2] The BAR casebook example that uses this data (p 138) focuses instead on the relationship of these indices with the returns on Walmart stock. The results are similar and similar issues of collinearity arise there as well.

A simple regression of PACGE on SP500 yields

Bivariate Fit of PACGE By SP500



Linear Fit

PACGE = 0.0096823 + 0.3102945 SP500

Summary of Fit

RSquare	0.097073
RSquare Adj	0.089356
Root Mean Square Error	0.046161
Mean of Response	0.012718
Observations (or Sum Wgts)	119

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.02680342	0.026803	12.5786
Error	117	0.24931178	0.002131	Prob > F
C. Total	118	0.27611520		0.0006

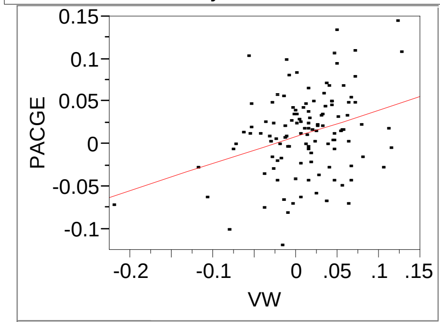
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0096823	0.004317	2.24	0.0268
SP500	0.3102945	0.08749	3.55	0.0006

What is the interpretation of $\hat{\beta}_1$ here?

A simple regression of PACGE on VW yields

Bivariate Fit of PACGE By VW



Linear Fit

Linear Fit

PACGE = 0.0083714 + 0.3156957 VW

Summary of Fit

RSquare	0.105532
RSquare Adj	0.097887
Root Mean Square Error	0.045945
Mean of Response	0.012718
Observations (or Sum Wgts)	119

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.02913890	0.029139	13.8040
Error	117	0.24697630	0.002111	Prob > F
C. Total	118	0.27611520		0.0003

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0083714	0.004371	1.92	0.0579
VW	0.3156957	0.08497	3.72	0.0003

What is the interpretation of $\hat{\beta}_1$ here?

Consider now what happens when *both* SP500 and VW are used together in a multiple regression

Response PACGE

Whole Model

Summary of Fit

RSquare	0.114681
RSquare Adj	0.099417
Root Mean Square Error	0.045906
Mean of Response	0.012718
Observations (or Sum Wgts)	119

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	0.03166507	0.015833	7.5131
Error	116	0.24445013	0.002107	Prob > F
C. Total	118	0.27611520		0.0009

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0054478	0.005119	1.06	0.2895
SP500	-0.821098	0.749946	-1.09	0.2758
VW	1.1114984	0.731784	1.52	0.1315

What has happened?[3]

[3] If returns on Walmart are the response, the regression shown in the casebook on page 143 finds a significant effect for the value-weighted index. Thus, in that case, VW significantly improves a regression with SP500 alone, but not vice versa. Adding SP500 to a model that already has VW does not improve the fit, agreeing with underlying finance.

Collinearity

In a multiple regression of y on $x_1, \ldots, x_K$, linear redundancy – or correlation – among $x_1, \ldots, x_K$, is called *collinearity*.

Effects of collinearity:

> Coefficient standard errors increase
> t-ratios decrease (and so p-values increase)
> Difficulty interpreting coefficients
> Coefficients change as others come and go.

These effects can be serious when collinearity is severe.

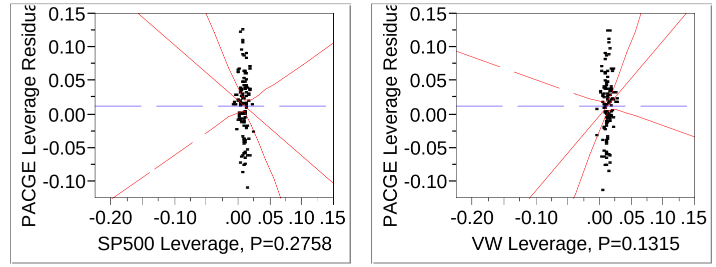
Why these effects happen:

> Key fact: In a multiple regression, $\hat{\beta}_k$ is the effect of adding x_k last. (As shown in the leverage plots)
>
> Variation of x_k with the other x's fixed is limited (p 121)
> This manifests itself as

$$\mathrm{SD}(\hat{\beta}_k) = \frac{\sigma_\varepsilon}{\sqrt{n}} \times \frac{1}{SD(adjusted\ x_k)}$$

> where adjusted x_k is the residual from a multiple regression of x_k on all the other x's
>
> The increase in SE($\hat{\beta}_k$) leads to smaller t-ratios.

The following leverage plots for the multiple regression of PACGE on SP500 and VW illustrate this phenomenon.



What to do if you have severe collinearity (p. 147)

- Suffer[4]

- Remove natural proxies

- Transform or combine some of your predictors

[4] Collinearity does not violate an assumption of the MRM. Rather, it causes problems in interpretation: the coefficients may not make much sense. If you only need to predict cases like the ones you have seen, it's not a problem. If you want to explain your predictions, it is.

Example: Market Segmentation

A marketing project identified a list of affluent customers for its new PDA. Should it focus on the younger or older members of this list?

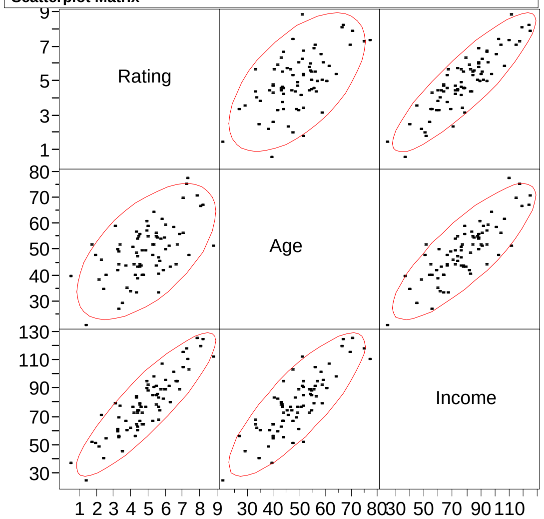
To answer this question, the marketing firm obtained a sample of 75 consumers and asked them to rate their "likelihood of purchase" on a scale of 1 to10. Age and income of consumers were also recorded.

Multivariate

Correlations

	Rating	Age	Income
Rating	1.0000	0.5867	0.8845
Age	0.5867	1.0000	0.8286
Income	0.8845	0.8286	1.0000

Scatterplot Matrix



The two simple regressions and multiple regression of *Rating* on *Age* and *Income* yields the following:

Regression of Rating on Age

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.067	0.487	4.24	<.0001
Age	0.059	0.009	6.19	<.0001

Regression of Rating on Income

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.596	0.352	-1.69	0.0951
Income	0.070	0.004	16.20	<.0001

Multiple Regression Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.736	0.295	-2.50	0.0149
Age	-0.047	0.008	-5.74	<.0001
Income	0.101	0.006	15.63	<.0001

What's going on?

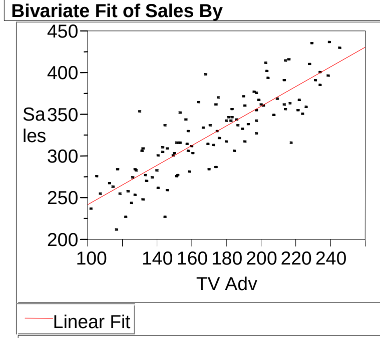
Based on these results, how should the marketing firm direct their marketing efforts?

Example: Advertising Allocation

A rapidly growing firm would like to improve its allocation of advertising dollars between television and print media. Television now gets the largest share. Should this continue?

An initial analysis quantifies the effect of television advertising.

Bivariate Fit of Sales By



Linear

Sales = 124.08525 + 1.1813206 TV Adv

Parameter

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	124.08525	12.71322	9.76	<.0001
TV Adv	1.1813206	0.072286	16.34	<.0001

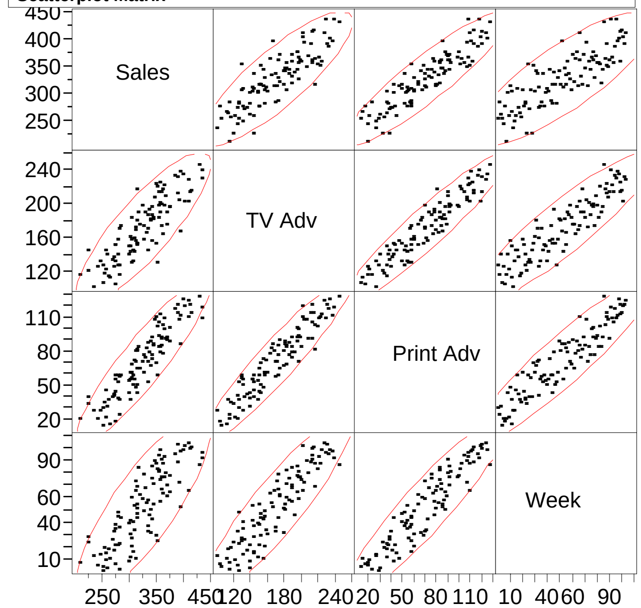
The scatterplot matrix (with "time" in the last column to show time trends) indicates that both sales and TV spending have grown over the two years, but so has print advertising.

Multivariate

Correlations

	Sales	TV Adv	Print Adv	Week
Sales	1.0000	0.8507	0.9135	0.8272
TV Adv	0.8507	1.0000	0.9428	0.9065
Print Adv	0.9135	0.9428	1.0000	0.9294
Week	0.8272	0.9065	0.9294	1.0000

Scatterplot Matrix



A multiple regression suggests a different impression for the effect of television advertising on sales.

Response					
Summary of					
RSquare	0.835424				
RSquare Adj	0.832165				
Root Mean Square Error	20.70725				
Mean of Response	327.3931				
Observations (or Sum Wgts)	104				

Analysis of				
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	219840.02	109920	256.3492
Error	101	43307.80	429	Prob > F
C. Total	103	263147.82		<.0001

Parameter				
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	228.91176	16.04249	14.27	<.0001
TV Adv	-0.13189	0.16816	-0.78	0.4347
Print Adv	1.6964939	0.204815	8.28	<.0001

Conclusions

Increased TV advertising – holding constant levels of print advertising – has no significant impact on sales. Why?

Increased print advertising would have a strong effect even when TV advertising was left unchanged. Why?

Might there be other collinear factors hidden from our analysis?

Finally, don't forget to *check assumptions*, in particular for trends in the residuals that might suggest important omitted factors.

Another Example

Just because we are doing multiple regression does not mean we should ignore transformations. Logs, in particular, can be very important in economic models.

Models with logs of Y and X lead to slope interpretations as *elasticities*. The BAR casebook gives an example (p 148).

Take-Away Review

The **F-test** allows for you to look at the importance of several factors simultaneously. When predictors are *collinear*, the F-test reveals their net effect rather than trying to separate their effects as a t-ratio does.

A **leverage** plot shows the contribution of each predictor to the regression, giving you a picture of what that variable adds to a model that contains *all* of the others.

Collinearity does not violate any assumption of the MRM, but it does make regression harder to interpret. In the presence of collinearity, slopes become less precise and the effect of one predictor depends on the others that happen to be in the model.

Next Module

Not all predictors are numerical. Some of the most important predictors of a response label an attribute of the observation, such as the sex or specialty of a doctor.

JMP allows you to easily include such categorical predictors in a regression, but leaves you with the burden of figuring out how to interpret the results. We'll start with that next time.