Diagnostics for Multiple Regression

Preliminaries

Office hours

– Wednesday 3-5:30, Friday from 9-12 noon.

– E-mail

Review of Key Points from Lecture 10

Tukey-Kramer procedure

– Use this method to *compare* the average values of the response after adjusting for the other predictors in the regression.

– This procedure gives a confidence interval for each of the pairwise comparison, regardless of how many pairwise comparisons are being done.

– The results are fully adjusted for all terms in the regression model. The method takes into account, therefore, all of the problems of confounding as well as the small constants introduced by interactions.

– The results are obtained by the "red triangle" near the leverage plot for the categorical effect.

Bonferroni approach

– Handles the problem that occurs when one scans over many, many t-statistics: persistence rewards you with a statistically significant result even when none of the effects are meaningful.

– The implementation of the procedure is simple. When searching for a significant effect (as opposed to be guided by an a priori theory), compare the p-value for a coefficient to

0.05/(number of tested coefficients)

rather than to 0.05 itself.

– The effect of the procedure is to judge serendipitous by a tougher standard than the terms that are motivated by a clear theory suited to the problem at hand.

Review Questions

Which predictors should I consider in interactions?

In some sense, all of them, particularly when it is easy to see that they all might depend on the location. Such an approach, however, is not practical when you recognize that interactions can occur between virtually every pair of predictors. For the project, stick to the obvious interactions (as motivated by the question) that involve location.

How do I decide whether to retain or exclude outliers?

If you think a point is an outlier, ideally, one should go back and check the source of the data for coding errors. You cannot do that for the project. For the project, any lease with unexplained costs that are two or three times the cost of others is clearly unusual. Similarly, its hard to see why a lease of a 250 sq. ft. property should have much influence in determining the estimated fixed costs for 220 other leases of much larger properties.

What other fixed costs should I be worried about?

This is a question of semantics. Clearly, parking spots are a fixed cost since these add costs to the lease regardless of the size of the lease. The coefficient for *1/SqFt* captures other, unnamed fixed costs such as those due to legal fees related to the lease. We do not have specific columns that identify these effects.

Why keep all of the relevant predictors in the "first model"?

Because I want you to interpret each of these, whether significant or not. For the first model, your regression should include as predictors all of

1/SqFt	*Location*
Parking/SqFt	*Restaurant*
Renovation	*Wiring*
	Exercise

as well as any necessary interactions. Do not retain interactions that are not statistically significant in your model.

For the *second model* (6-10), use a parsimonious model that contains *only* the statistically meaningful predictors.

Writing up the Project Results

Executive summary

– One page. Only one double-spaced, 10/12 pt font page. Just one.

– Summarize key findings.

– No technical language (e.g., p-value, standard error, RMSE, F-ratio…)

– Make recommendations based on your analysis, such as where the firm should locate given their preference for the city and the associated costs.

– Round values appropriately here and in answers to the subsequent questions. It's *crazy* to report total lease costs as something like

$1,253,159.22 +/– 336,724.55.

Such a range should be presented rounded off to reflect the uncertainty of this range, as in

$1,250,000 +/– 350,000.

Answers to specific questions

– Answer the question that is asked, directly.

– Here is an illustrative this answer to question #4. The associated row excerpted from the summary of my regression model for parts 1–5 is:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Renovation	0.0023	0.0047	0.49	0.6253

How much can the company save by obtaining a property that was renovated five years ago, rather than one year ago?

The difference in costs for leasing a property one year after renovation versus 5 years after renovation is quite small. My analysis of this data suggests that properties that were renovated 5 years ago are slightly more expensive than those renovated one year ago. The difference in costs is about $0.01 /SqFt, or $500 more for leasing a 50,000 square foot property renovated 5 years ago. The effect of renovation is not only small, but also poorly determined from this data. My analysis suggests that a property renovated 5 years ago could cost as much as $0.047/SqFt more or little as $0.028/SqFt less than one renovated one year ago.

General guidelines

– *At most* 5 pages, double-spaced, 10/12 pt. font.

– Use the stated predictors for questions 1–5. Be selective for 6–10. If one of the questions 6-9 ask about a predictor that you did not use, simply indicate that it is not useful/not statistically significant in your model.

– No technical language for these answers either. Save that for the appendix.

– Answer the questions directly and concisely.

Appendix

– Show no graph that is not discussed. If you do not discuss the relevance of the plot, do not show it. Saying "Here are the leverage plots; I looked at them." is not adequate to discuss several leverage plots.

– Describe why you excluded any outliers. Why just these? How did you find them? Show the points using the plot that helped you find them. Identify them clearly in the plot.

– Complete summary of regression model used for 1–5.

– Diagnostic analysis of this model, including assumption of equal variance and the assumption of normality.

– How did you go from this model to the "final" model used for 6–10. Suggest the steps that you took, perhaps with one or two illustrations or a summary table of what was done.

– Complete summary of the regression model used for 6–10.

– Diagnostic analysis of the second model, including assumption of equal variance and the assumption of normality.

Key Application for Today

How do I look for other factors?

– Residual methods that are effective when there is not too much collinearity.

– Automated selection methods, like stepwise regression.
See the casebook, pages 220 – 228 for an example of this method

Project Analysis (Stage 3, Wrapping Up)

Model summary

After removing terms to just those that are significant or nearly so…

Summary of Fit

RSquare	0.772
Root Mean Square Error	0.854
Mean of Response	16.967
Observations (or Sum Wgts)	223.000

Expanded Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	15.062	0.466	32.29	<.0001
1/Sqft	1759.187	374.886	4.69	<.0001
Park/Sqft	1560.990	260.249	6.00	<.0001
Location[CITY]	0.972	0.138	7.06	<.0001
Location[SUBNEW]	-0.047	0.114	-0.41	0.6786
Location[SUBOLD]	-0.925	0.106	-8.69	<.0001
Wiring[NO]	-0.154	0.088	-1.76	0.0804
Wiring[YES]	0.154	0.088	1.76	0.0804
Location[CITY]*(Park/Sqft-.0002)	1483.630	269.231	5.51	<.0001
Location[SUBNEW]*(Park/Sqft-.0002)	-310.632	425.493	-0.73	0.4662
Location[SUBOLD]*(Park/Sqft-.0002)	-1172.998	280.483	-4.18	<.0001
Leaselength	-0.042	0.023	-1.81	0.0722
Distcity	-0.095	0.056	-1.70	0.0906
Occupancy	2.059	0.492	4.19	<.0001

Check the interpretation

– Signs in expected direction, effects make sense? Collinearity?

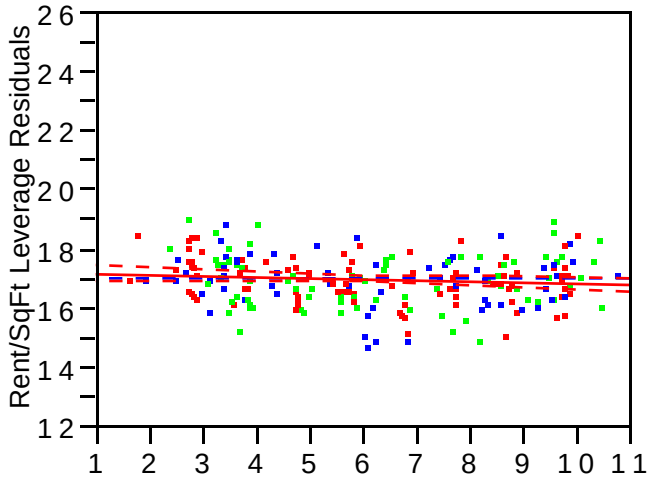
Diagnostic plots

– Leverage plots

– Residual plots

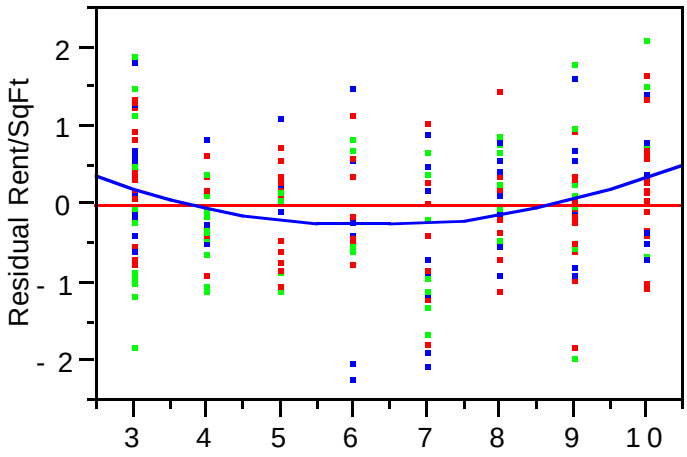
Lease-length weirdness

– In the leverage plot, you can see a bend (as in Assignment #2)



– This type of nonlinearity is usually more clear if you l plot the residuals on the original predictor itself. The correlation between the two is zero by construction, but a nonlinear pattern would remain. (To do this, save the residuals from the multiple regression and use Fit Y by X.).

– The next plot shows the result of fitting a line (horizontal) and a quadratic.



– Return to the model and add a quadratic term (using the cross or build it explicitly). If you "cross" lease length with itself, subtracting the mean as JMP-IN does reduces collinearity. Here's the new output (not expanded)

Summary of Fit

RSquare	0.787
Root Mean Square Error	0.829
Mean of Response	16.967
Observations (or Sum Wgts)	223.000

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	14.768	0.459	32.16	<.0001
1/Sqft	1793.872	363.801	4.93	<.0001
Park/Sqft	1643.869	253.425	6.49	<.0001
Location[CITY]	0.903	0.135	6.70	<.0001
Location[SUBNEW]	0.012	0.112	0.11	0.9151
Wiring[NO]	-0.146	0.085	-1.72	0.0870
Location[CITY]*(Park/Sqft-.0002)	1344.623	263.768	5.10	<.0001
Location[SUBNEW]*(Park/Sqft-.0002)	-104.482	416.374	-0.25	0.8021
Leaselength	-0.052	0.023	-2.30	0.0225
Distcity	-0.119	0.055	-2.18	0.0301
Occupancy	2.206	0.479	4.61	<.0001
(Leaselength-6.15)*(Leaselength-6.15)	0.045	0.012	3.78	0.0002

Going further

– Look at plots of the residuals from this model versus included *and* excluded predictors.

– Scatterplot matrix of residuals vs. other factors.

Done?

– Check the assumptions of
- Constant variance
- Normality

Automatic Methods

Stepwise regression and "data mining"

– Automated search over all of the predictors and their interactions.

– Runs fast, but can you sort out the results.

– Obtained via changing the personality of the Fit Model dialog and then using a "Response Surface" to add all of the possible factors.

– See the casebook example (page 220) for an illustration of how the method can go wrong when used carelessly.

Key Take-Away Points

Project

– Reaching closure

– Residual diagnostics

Next Time

Analysis of variance

– Experiments

– Conjoint analysis and marketing research.