Categorical Variables, Part 1

Project Analysis for Today

First multiple regression

Add predictors to the initial model (with outliers held out) and interpret the coefficients in the multiple regression. Some of these new predictors (e.g., location) are categorical, and require the methods of today's class.

Review: Collinearity in Multiple Regression

What is collinearity? (Also known as multicollinearity.)

– Collinearity is correlation among the predictors in a regression.
– As such, collinearity does not "violate an assumption" in regression and is in fact a typical feature of most regression models.

What does collinearity do in regression? Consequences?

– Complicates interpretation, making it hard to separate the predictors.
– Inflates the SE's of the estimated coefficients.

How can I tell if collinearity is present?

– Graphically: Scatterplots help, but *leverage plots* are better.
  - Multiple "simple regression" views of one multiple regression.
  - Essential for identifying leverage points in multiple regression.
  -"Do I like the shown simple regression model?"
– Tests: Big F ratio, small t-ratio
– Diagnostic: Variance inflation factors (VIF)

$$\text{SE(slope estimate for } X_j) \approx \frac{\sigma}{\sqrt{n}} \frac{1}{\text{SD(Adjusted } X_j)}$$
$$= \frac{\sigma}{\sqrt{n}} \frac{\sqrt{\text{VIF}_j}}{\text{SD}(X_j)}$$
$$= \sqrt{\text{VIF}_j} * \text{(SE if no collinearity)}$$

What do I do about collinearity?

– Nothing. Collinearity complicates our ability to interpret, but in-sample prediction remains OK in the presence of collinearity.
– Reformulate predictors. Identify distinct concepts.
– Get rid of one of the offenders. Diagnostics (vif) help you decide which.
– Summary discussion on page **147** of the casebook.

Example of Multiple Regression

Automobile design Car89.jmp, page 109

"What is the predicted mileage for a 4000 lb. design, and what characteristics of the design are crucial?"

"How much does my 200 pound brother owe me for gas for carrying him 3,000 miles to California?" (Oops, it's urban mileage in example)

– Initial one-predictor model
- Transform response to gallons per 1000 mile scale.
- $\Uparrow$ 200 lbs for 3000 miles $\approx$ 8.2 gals
- RMSE = 4.23 (p 111)
- Skewness in residuals from regression with *Weight*. (p 112)
- *Predicted consumption @ 4000 lbs = 63.9 using JMP*

– Add variable for *Horsepower* (p 117)
- R^2 increases from 77% to 84% (added variable is significant, t=7.21)
- Predictors are correlated, higher SE for *Weight* (plot on p 120)
- $\Uparrow$ 200 lbs for 3000 miles $\approx$ 5.3 gals
- RMSE drops to 3.50
- Residuals evidently more normally distributed
- *Predicted consumption @ 4000 lbs, 200 HP = 65.0, [57.9, 72.1]*

Next steps for this model…

– What other factors are important for the design?
– How small can we make the RMSE?

Example with Extreme Collinearity in Multiple Regression

Stock prices and market indices Stocks.jmp, page 138

"What's beta for Walmart when regressed on returns of *two* indices?"

– Initial correlations and scatterplot matrix show outliers and high collinarity between the two market indices.
– Addition of sequence number to the plots also shows time series patterns.

– Fitted slope of stock returns on market estimate the beta for the stock.
– Initial beta estimate in regression of Walmart on S&P alone is 1.24 with SE .12 (is it significantly larger than one?)

– Add VW market returns to the regression.
– Huge collinearity (correlation between VW and S&P is 0.993), so almost no unique variation in either one given that other is in model.
– Either taken separately is a good predictor, but show weak effects (i.e., not significant) when used together.
– "Squished" leverage plots... little unique variation in either predictor available to explain the variation in the response. (p 144)

– More complete VW index is better predictor, as financial theory suggests, and we should use it alone to estimate the beta for Walmart.

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.020	0.006	3.22	0.0016
VW	1.239	0.118	10.49	<.0001

New Ideas and Terminology for Today

Categorical variable

– Represents group membership (e.g.: type of car, race, sex, religion).
– *JMP* denotes as "nominal" or "ordinal" in the column header.
– *JMP* does a lot of work in the background when these terms are added to a regression, building special variables to represent the groups and then adding these variables "in the background" – showing you the resulting regression coefficients.

Interaction

– Same concept as seen in anova:
Interaction implies that the effect of one predictor on the response depends on the value of other predictors.
– Measures how the slope of one predictor depends upon levels of others.
– Important in many models, crucial in models with categorical.
Interaction with categorical $\Rightarrow$ slope depends upon the group.

Important questions to answer when using categorical variables

– Are the fits in the different models parallel? (i.e., Is interaction present?)
– Are the error variances comparable? Heteroscedasticity can be a problem.

Messy part: Interpreting the output

– Take your time
– Write down the fit for each group, one at a time (until output is familiar).
– Be careful reading *JMP* output correctly.
– Term for one group will not be explicitly shown.

Analysis of covariance

A regression model that contains both categorical and continuous predictors, usually with a focus on the difference among the groups.

Categorical Predictors in Multiple Regression: Two Groups

Employee performance study Manager.jmp, page 161

Questions

"Do data support the claim that externally recruited managers do better?"

"Which of two prospective job candidate should we hire, the internal or the externally recruited manager?"

Data

150 managers, 88 of which are internal and 62 are external.

Analysis

• Initial comparison of the two groups
Average performance rating for internal managers is significantly lower than that for external managers,

difference = 0.72 with t = 2.98 (page 161)

(An aside: Regression with a two-group categorical variable alone is the same as a two sample t-test between the two groups.)

• Confounding issue
Salary is much higher for externally recruited managers... They occupy higher level positions within the company, and *Salary* is related to rating (p 164-165).

• Separate regressions of *Rating* on *Salary* for *In-House?* suggest reversed difference on means: at fixed salary, internal are more highly rated! (p166-67)

• Combined as one multiple regression
Slopes are parallel (i.e., no significant interaction), and model (page 168) implies that internal managers actually rate significantly higher

difference = -0.514 = 2 × -0.257 with t=-2.46

Conclude

After checking assumptions, conclude that ought to hire the internal candidate since at a given salary, we expect the internal manager to fare better.

JMP "tricks"

Point codes/labels using the values of a categorical variable.
Fitting several models in one *Fit Y by X* view.

Next Time

Review session

This Friday, reviewing material using multiple regression, with emphasis on how these ideas are relevant to the project.

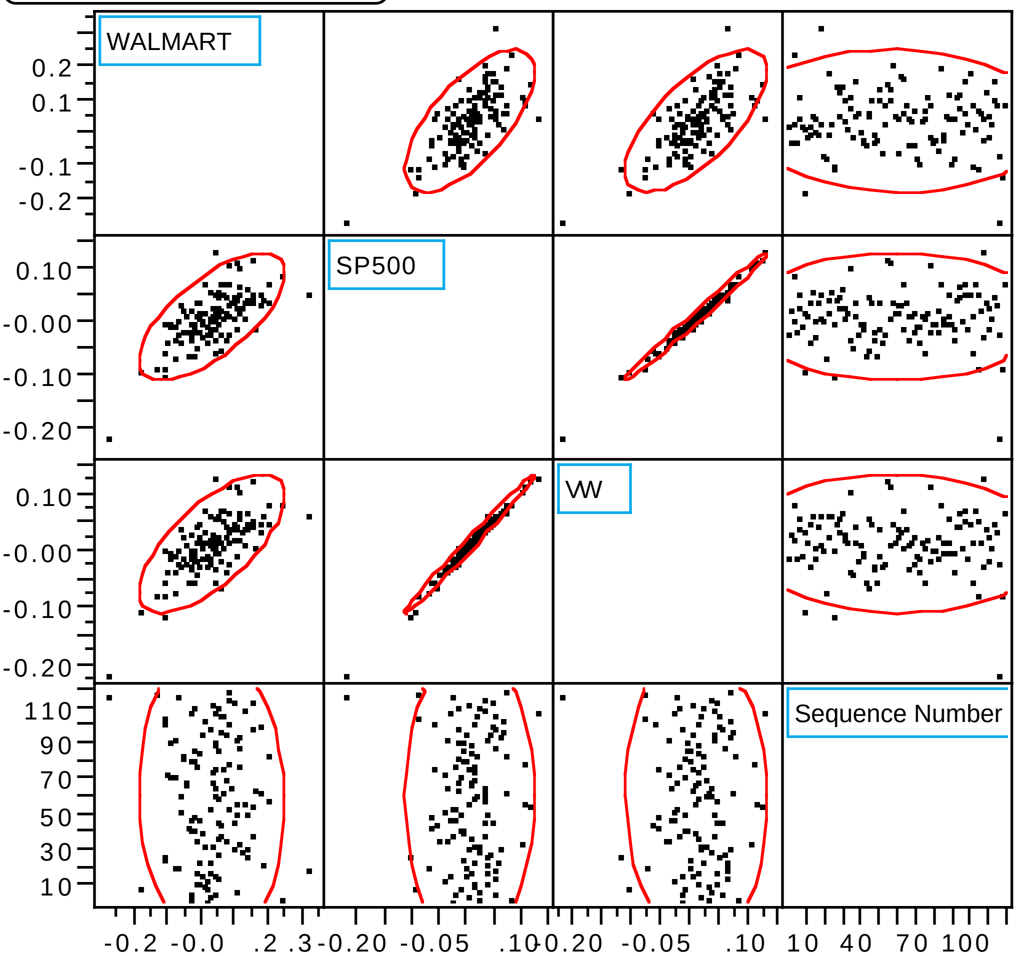
Categorical predictors, continued

Categorical predictors interact with the other predictors in the model. We'll do another example with these to help with the JMP notation.

Correlations

Variable	WALMART	SP500	VW	Sequence Number
WALMART	1.000	0.682	0.696	-0.055
SP500	0.682	1.000	0.993	0.002
VW	0.696	0.993	1.000	-0.036
Sequence Number	-0.055	0.002	-0.036	1.000

Scatterplot Matrix



Initial regression fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	0.024	0.006	4.02	0.0001	0
SP500	1.244	0.123	10.10	<.0001	1

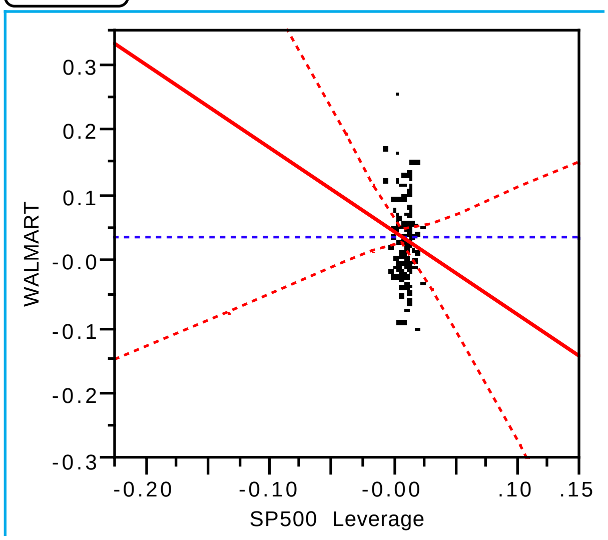
Fit using both market indices

Parameter Estimates

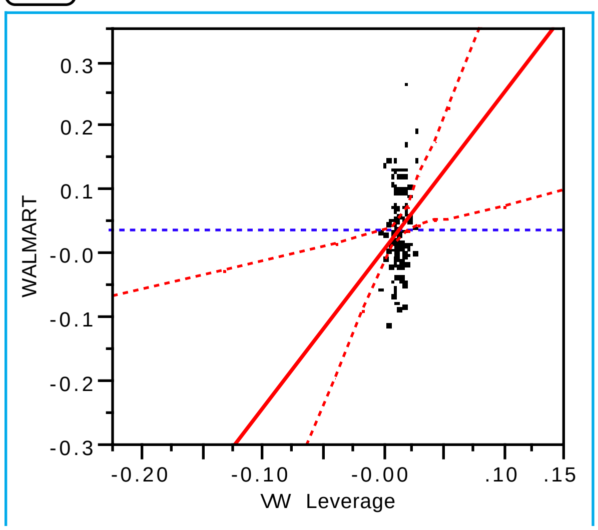
Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	0.015	0.007	2.13	0.0356	0.000
SP500	-1.258	1.041	-1.21	0.2294	74.297
VW	2.458	1.016	2.42	0.0171	74.297

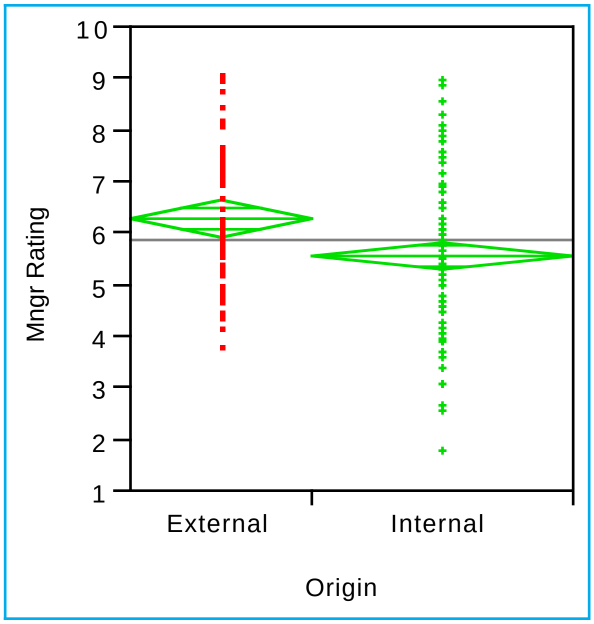
Leverage plots

SP500



VW





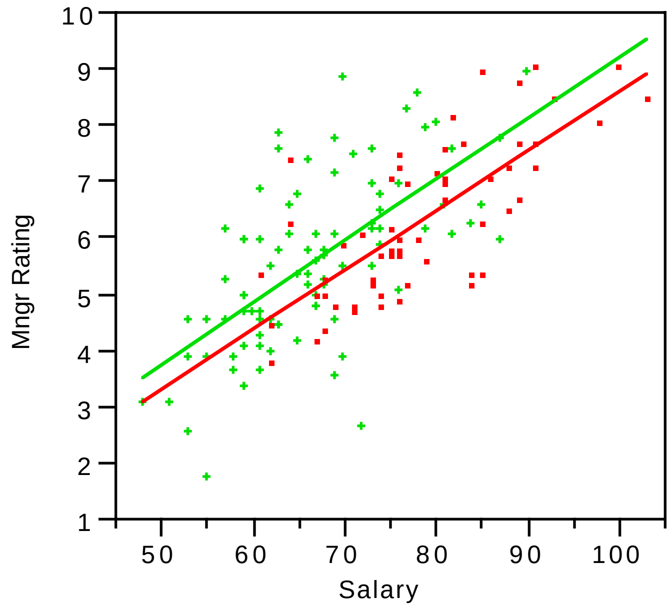
Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean
External	62	6.321	1.342	0.17044
Internal	88	5.605	1.518	0.16181

t-Test

	Difference	t-Test	DF	Prob>|t|
Estimate	0.716	2.984	148	0.0033
Std Error	0.240			
Lower 95%	0.242			
Upper 95%	1.191			

Assuming equal variances



Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.693524	0.949254	-1.78	0.0779
Salary	0.1090929	0.014066	7.76	<.0001

Internal

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.936941	0.986231	-1.96	0.0542
Salary	0.1053912	0.012499	8.43	<.0001

External

Fit as one multiple regression that combines these two

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.815	0.724	-2.51	0.0132
Salary	0.107	0.010	10.99	<.0001
Origin[Externa-Interna]	-0.122	0.724	-0.17	0.8667
Salary*Origin[Externa-Interna]	-0.002	0.010	-0.19	0.8499