Categorical Predictors

Administrative Items

Deliverables

– Postpone Assignment #2 until Friday at **3** p.m.

– Further questions regarding the assignment?
We'll do the partial F ratio in an example.

Review from Last Time

Collinearity

Recognize from
Visually from scatterplot matrix and from the *leverage* plots
(diagonal ellipses) (narrow column)
Substantively from unstable, uninterpretable coefficient estimates
(negative beta)
Statistically from inflated standard errors (VIF), wide CIs.

Tests in multiple regression

– t-ratio for the impact of *one* predictor, taken incrementally

– F-ratio for the overall model, the effect of *all* of the predictors.

Diagnostics in multiple regression

– Whole model plot and overall residual plot

– Leverage plots for each coefficient in the model

Review Example

Questions

– Leverage plots, variance inflation factors, different tests.

– Is any of this relevant?

Improved parcel handling **Parcel.jmp, p. 148**

> “How can a shipping company improve performance at sorting centers?”

– Resource allocation: labor vs. capital.

– Predictors of # sorted packages are:
# sorting lines, # sorters, and # of truckers.

– All transformed to a log scale, so slopes are interpreted as an elasticity,

$$\text{slope} = \frac{\text{change in log Y}}{\text{change in log X}} = \frac{\text{\% change Y}}{\text{\% change X}}$$

The resulting model is *multiplicative* (so the fit is zero if any one of the predictors is zero, a Cobb-Douglas production function).

– Leverage plots for the predictors
These identify some leverage points (#11, #22), but indicate no major problems with the predictors in the multiple regression.

– Variance inflation factors.
Largest is about 3 – so some collinearity, but not extreme. It is enough, however, to permit slopes to change as other predictors join the model.

– Interpretation: marginal vs. partial elasticity for the # of sorting lines
• What is the source of the difference?
• Do the marginal and partial elasticities mean the same thing?
marginal elasticity for # lines is 0.70 (p 150)
partial elasticity for # lines is 0.37 (p 151)

– Management issue:
Can you just increase the number of sorting lines (the factor with the largest partial elasticity) without also increasing the number of workers?

– Illustration of partial-F test
Does adding *both* log(sorters) and log(lines) improve the fit?

Key Applications

Fitting models that capture differences between 2 groups

– Did cars from different regions differ in gasoline consumption?

– Are women systematically underpaid at this firm?

– Does this company discriminate in hiring of minorities?

Concepts and Terminology

Categorical variable

– Represents group membership (e.g. type of car, race, sex, religion)

– *JMP-IN* denotes such columns as "nominal" or "ordinal".

– Once handled by forming "dummy variables".

Fitting models with data from different groups

– Combine data from the groups when you have reason to believe that the groups are essentially similar, with some differences.

– Otherwise, fit a model to each group separately (*JMP* makes this easy).

– If the separate models are similar, *combine them to make comparisons*.

Interaction

– Measures how the slope of one predictor depends upon levels of others.

– Important in many models, crucial in models with categorical. E.g.
Does the effect of television advertising on sales (i.e., its slope) depend the region of the country (say, urban vs. rural)?

– Presence of interaction implies that the effect depends upon the group.

Questions to answer when using categorical variables

– Are the groups really so similar as to make sense to combine?

– Do slopes in the different models differ? (i.e., Is interaction present?)

– Are the error variances comparable? Heteroscedasticity can be a problem.

Interpreting the output

– Initially, take a careful approach
Write down the fit for each group, one at a time until you become familiar with the format of the output.

– Coefficients for all groups revealed only in the “expanded” estimates

Analysis of covariance

Fancy name for regression model that contains both categorical and continuous predictors, usually with a focus or emphasis on the difference in the intercepts among the groups.

JMP-IN Methods for Categorical

Column types

– Make sure that the columns in the spreadsheet you want JMP-IN to treat as categorical are marked as either *ordinal* or *nominal*.

– Check the status of each column by examining the left panel of the spreadsheet window: each column name is preceded by a symbol denoting how JMP-IN is treating that column:
c for “continuous”, *n* for “nominal” and *o* for “ordinal”.

– Change the “modeling type” by clicking on the status symbol itself.

Point codes and coloring the points

– Color/Mark by column
Plots of data that mix observations from different groups are usually more informative if you color/mark the points by their group membership.

– In the Fit Y by X view, you can even fit separate lines to each group by defining a “grouping variable”. (Use the “Group by…” option offered in the same menu that fits the line; it’s the one associated with the red triangle at the top of the scatterplot.)

Examples for Today

Employee performance study Manager.jmp, page 161

> "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

Average performance rating for external managers is significantly higher than that for internal managers

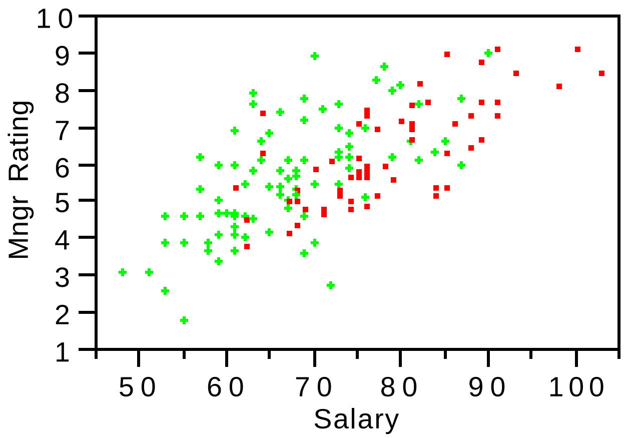
(avg ext.– avg int) = 0.72 with t = 2.98 (page 161)

t-Test

	Difference	t-Test	DF	Prob > |t|
Estimate	0.72	2.98	148	0.0033
Std Error	0.24			
Lower 95%	0.24			
Upper 95%	1.19			

Confounding issue…

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

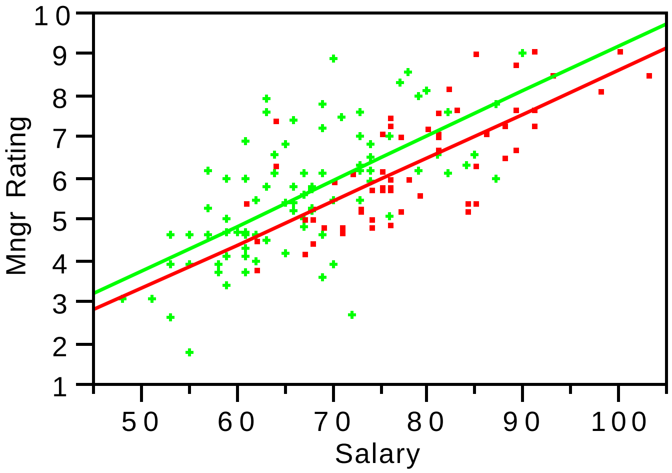


Collinearity in a new form…

Salary is related to the one predictor used to explain the rating of a manager, namely the internal/external status of a manager.

Separate fits…

Separate regressions of *Rating* on *Salary* for *In-House?* Reverse the difference: at a fixed salary, internal are more highly rated! (p166-67) The green line in the figure below (for internal managers) is consistently higher than the red line giving the expected rating for external managers.



Statistical test of difference

If we assume slopes are parallel (i.e., no evident interaction), a model using a categorical predictor (expanded estimates version of model on page 168) implies that internal managers actually rate significantly higher

Expanded Estimates

Term	**Estimate**	**Std Error**	**t Ratio**	**Prob>|t|**
Intercept	-1.84	0.71	-2.61	0.0100
Salary	0.107	0.01	11.14	<.0001
Origin[External]	-0.257	0.10	-2.46	0.0149
Origin[Internal]	0.257	0.10	2.46	0.0149

Fits for two groups

Internal	Predicted rating = –1.84 + 0.107 Salary + 0.257
External	Predicted rating = –1.84 + 0.107 Salary – 0.257

difference in intercepts = -0.514 = 0.257 – (-0.257) with t=-2.46

How can we be sure the slopes are parallel?

Rather than only look at the plot, we can fit a model that allows the slopes to differ and estimate the difference between the slopes directly. An interaction term captures the difference between the slopes. It is added to the model using the "cross" button. In this example, the interaction term is not significant (page 173)

Expanded 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[External]	-0.254	0.106	-2.39	0.0179
Origin[Internal]	0.254	0.106	2.39	0.0179
(Salary-71.63)*Origin[External]	-0.0018	0.010	-0.19	0.8499
(Salary-71.63)*Origin[Internal]	0.0018	0.010	0.19	0.8499

Fit for internal group

Internal Predicted rating = –1.815 + 0.107 Salary
+ 0.254 + 0.0018 (Salary–71.63)
= –1.815 + 0.254 – 0.0018(71.63) + (0.107+ 0.0018) Salary
= –1.690 + 0.109 Salary

As found in the initial simple regression for this group – but now we can compare this fit directly to the fit to the external group.

Note on the output

JMP-IN forms an interaction by subtracting the mean from the continuous predictor. Doing so reduces collinearity in this type of regression model.

Conclude

After checking assumptions (particularly, the assumption of equal error variance p 174), we conclude that ought to hire the internal candidate since at a given salary, we expect the internal manager to fare better.

Comment on presentation

This type of analysis would usually begin by checking for an interaction before looking at the difference of intercepts. Starting with the full model that includes interactions is a bit hard from a pedagogical point of view, so I built up the full model rather than start there.

Supplemental (Two groups, with a complex interaction)

Wage discrimination Salary.jmp, page 180

"Are men paid more than women in these management positions?"

Data

220 managers, 145 are men and 75 are women.

Confounding

- Marginally, women are paid less (t= –2.06) than men (p 181)
- Confounding effects: men have more experience and work in positions of higher responsibility (or position). (p. 182-183)

Regression with categorical predictor

Adjusting for confounding effects in *Position* and *Years of Experience* produces a model (p 184) with the reversed assessment of the difference between salaries: when adjusted for differences in experience and responsibility, men are paid about $2,200 *less* than women.

Interactions

The case includes discussion of various types of interaction, such as that between two continuous variables.

Practical comment

Watch out for collinearity when adding interaction terms. If the interactions are not significant, remove them from the fitted model.

Conclude

Comparison of groups with statistically adjusted backgrounds suggest that men are paid less. However, why do the women have less background?

Key Take-Away Points

Categorical predictors in regression

– Allow regression to estimate and test for differences between two groups.

– One approach to correcting for *confounding* caused by careless or convenient sampling that does not randomize the assignment of observations to the two groups.

– Categorical terms alter intercept, interactions alter associated slope. When the interaction is included, the resulting fitted model implied for each group is the same as if you had fit a separate model to each. However, the multiple regression gives t-ratios that allow you to compare those fits – a comparison that you can only do informally otherwise.

– Extend idea of t-test to problems where groups are not directly comparable.

Next Time

Allowing more categories...

– Analysis with more than two categories and interaction.

– Finally, the partial F test has a reason for existence.