Categorical Predictors and Interaction

Preliminaries

Supplemental practice problems

– Multiple regression practice questions

Feedback and questions

– Please continue to use the in-class forms

– Academic reps, members of cohort QC team

– E-mail and office hours

Project

Get your data downloaded this week

– Download as text file

– Open text file and import data into JMP-IN spreadsheet

– Initial graphical analysis.
Look back at analysis of first problem on Assignment #2

Review of Key Points from Prior Class

Confounding effects

– Flaws in a marginal comparison
When the observations (in that example, the managers) have not been randomly assigned to the groups, *confounding* is possible.

– Confounding, in general
The differences that we detect in a marginal comparison are due to other factors, *not* the nominal labeling that defines the groups. Confounding is collinearity in another guise (e.g., it’s age not shoe size in IQ example).

– Confounding, in the hiring application
Internal managers look worse marginally, but better when adjusted to be doing comparable jobs (i.e., in the sense of jobs that pay equally).

– Why does it matter that the factors are confounded?
Consider the effects of actions taken based on the marginal comparison. In this application, you’d hire the less desirable employee (on average).

Adjusting for confounding

– Compare predictions of two regression models
Which group has the higher expected performance rating when filling a specific position in the company. The fit for the *internal* managers is higher.

– Question
Are the fitted regressions significantly different, or have we found a random difference rather than a systematic difference? Would the same effect show up in another sample, or is it just a property of this sample?

– Parallel
The comparison of the two regressions is "easy" if the fits are parallel, for then the difference when hiring a manager at $75,000 is the same as hiring one at $85,000 – or at zero for that matter. Thus we can focus on the difference of the intercepts, *if the fits are parallel.*

Common Question

Why not compare the intercepts using their confidence intervals?

– Recall the comparison of two averages using confidence intervals rather than the two-sample t-test in Stat 603.

– Such a procedure is OK when the differences are large, but…
(1) This procedure does not generalize naturally to problems having more than two groups (gets pretty tedious then).
(2) This procedure will miss some differences that are significant.

Key Applications for Today

Modeling with interaction terms

– Are the fitted slopes for internal and external managers different?

– Do spending patterns differ among consumers from different nations?

Concepts and Terminology

Combining several regressions into one

– How do we test for differences between the regressions fit to separate groups?

– It would be easy if we had
(1) a t-ratio for the difference in slopes and
(2) another for the difference in intercepts.

– “Glue” the separate simple regressions into *one* multiple regression.

Interaction

– Question: Does each predictor affect the response in the same way (i.e., have the same slope) for each group?

– Rephrased: Is there an *interaction* between the predictor and the categorical variable that identifies the groups.

– Consequence: Interaction complicates interpretation since the fits are no longer parallel, and thus the difference between fits is not constant.

Important questions to answer when using categorical variables

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

– Are the fits in the different models parallel? (i.e., Is interaction present?)

– If they are parallel, are the intercepts different?

– Are the error variances comparable? (i.e., Is heteroscedasticity present?) The multiple regression that joins the fits to the separate groups has one RMSE. Each fit to separate groups has its own. Is one good enough?

JMP-IN

Reading the output

– Each regression with categorical predictors has a "baseline" model common to all groups.

– Terms associated with a categorical predictor shift the equation up and down by changing the intercept.

– Interactions that combine categorical and continuous predictors shift the slope from the baseline model.

It looks easy in class, but when I try to do it, I can't figure out where to click next. What should I do?

Take a look in the index of the JMP-IN manual. It has a summary of all the menu commands. Try the on-line help too.

Also, keeping up with the class is easier if you not only skim the cases, but also try to reproduce the output as well. Then you'll know what to look for during the class demonstration.

How do you fit separate lines to each group in a scatterplot?

Use the red triangle pop-up menu at the upper left of the window. Select the "Group by" option and choose the categorical factor of interest. When you next fit a line in that scatterplot, JMP-IN will fit a separate line for each group.

How do you see results for both categories?

Be sure to ask JMP-IN for the *expanded* estimates (using the red triangle to get to the pop-up dialog in the multiple regression output).

How do you color-code the points?

Use the "Color/mark by Column" command from the *Rows* menu.

How do you add an interaction to a regression?

An interaction involves a pair of predictors. Select the two that you want to combine and click the "cross" button. JMP-IN will form the interaction term as an added predictor (labeled, e.g., as Origin*Manager). It does not matter which name comes first.

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.

Variables

• *Salary* is the starting salary of the employee when hired and indicates what sort of job the person was hired to do.

• *MngrRating* is an evaluation score of the employee in the job they do, indicating the “value” of the employee to the firm. Higher paying jobs typically earn higher ratings since such jobs provide the opportunity to offer more important contributions to the firm.

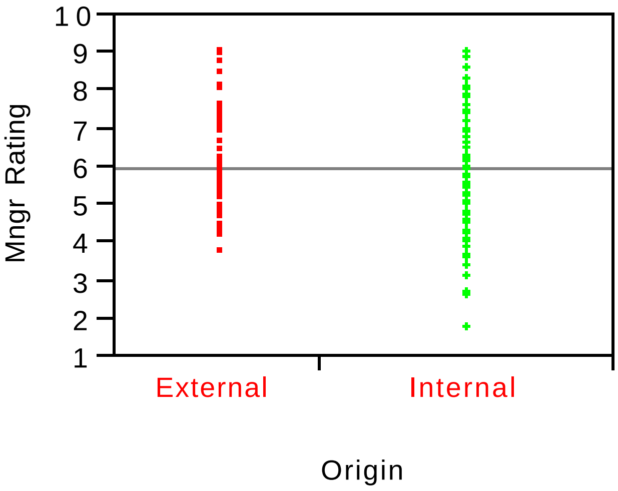
Analysis

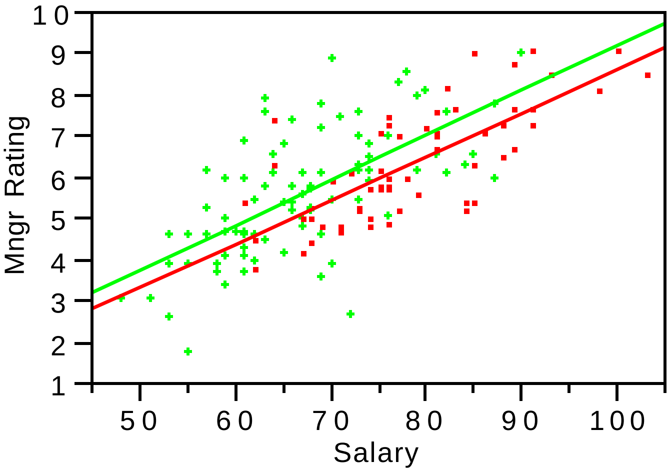
In a preliminary marginal analysis, the average performance rating for external managers is significantly higher than that for internal managers

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

Confounding issue…

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





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 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, if we assume that the slopes are the same.

***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, twice the coefficient in the regression summary.

Are the fitted models are parallel?

Rather than look at the picture, fit a model that allows the slopes to differ. Use it to estimate the difference between the slopes. The key is to use an *interaction.* Add it to the model using the "cross" button. (see the compact output on p. 173). In this example, the interaction is *not* significant

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

$$\begin{aligned}\text{Predicted rating} &= -1.815 + 0.107\,\textit{Salary}\\ &\qquad + 0.254 + 0.0018\,(\textit{Salary}-71.63)\\ &= -1.815 + 0.254 - 0.0018(71.63) + (0.107 + 0.0018)\,\textit{Salary}\\ &= -1.690 + 0.109\,\textit{Salary}\end{aligned}$$

This fit *is* the initial simple regression for this group – but now we can compare it to the fit for the external group using the t-ratios in the output.

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 on p. 174), conclude that the company ought to hire the internal candidate since at any given salary, we expect the internal manager to rate better by a statistically significant margin.

Comment on the order of our analysis

This type of analysis would usually begin by first checking to see if the slopes are parallel before looking at the difference of intercepts. Starting with the full model that includes interactions is a bit hard from a teaching point of view, so I have built up to the full model rather than start there.

Supplemental

Wage discrimination **Salary.jmp, page 180**

“Are men paid more than women in mid-level management positions?”

Data

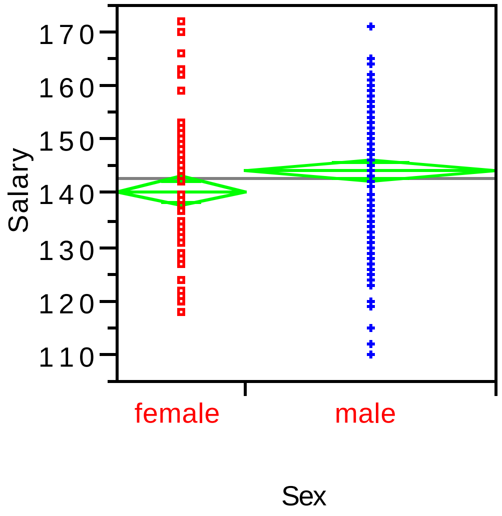
220 managers, 145 are men and 75 are women.

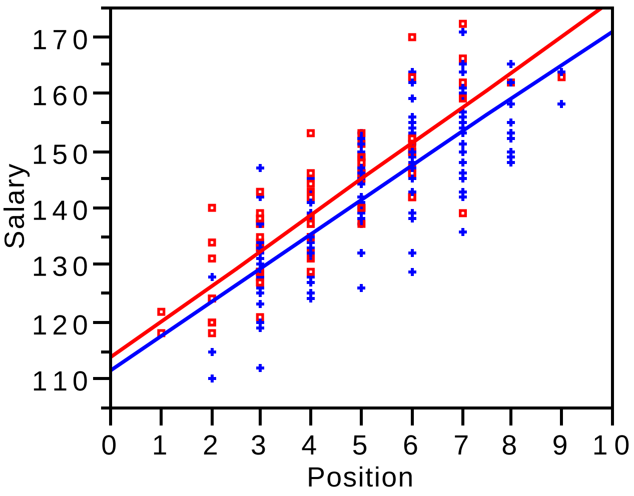
Marginal analysis

Men are paid more (p 181), to the tune of about $3600 per year. This difference is statistically significant.

Confounding

More “careful” analysis shows that men occupy higher-level positions than women in this sample. If the analysis is restricted to a subset that conditions on the position of the manager, the revised analysis indicates that men are paid less, not more. The loss of data leads to insignificant effect. (p 183)





Regression with categorical predictor

Uses of all of the data rather than a subset and finds that the salary difference is significant (table below shows *expanded* estimates)

Expanded Estimates

Term	**Estimate**	**Std Error**	**t Ratio**	**Prob>|t|**
Intercept	112.77	1.45	77.52	<.0001
Position	6.06	0.28	21.60	<.0001
Sex[female]	1.86	0.53	3.53	0.0005
Sex[male]	-1.86	0.53	-3.53	0.0005

Fits for the two groups

Women: predicted salary = 112.77 + 6.06 Position + 1.86

Men: predicted salary = 112.77 + 6.06 Position – 1.86

Thus we find that a man on average *at a given position* (rather than marginally) is being paid about 2(1.86)1000 = $3,720 *less* than a woman in the same position.

Check for interaction

This fit assumes that the two population lines are parallel, thus it forces the fitted lines to be parallel as well. To check this assumption, add an interaction (though it again looks clear there is no interaction).

Expanded Estimates

Term	**Estimate**	**Std Error**	**t Ratio**	**Prob>|t|**
Intercept	112.63	1.48	75.85	<.0001
Position	6.10	0.30	20.58	<.0001
Sex[female]	1.92	0.54	3.54	0.0005
Sex[male]	-1.92	0.54	-3.54	0.0005
Sex[female]*(Position-5.068)	0.15	0.30	0.49	0.6242
Sex[male]*(Position–5.068)	-0.15	0.30	-0.49	0.6242

Fits for the two groups

The model that includes both the categorical predictor and the interaction allows both the slopes and intercepts to differ (reproducing the original separate fits, but within one regression model)

Women: predicted salary = 112.63 + 6.10 Position + 1.92
+ 0.15 (Position–avg Position)

Men: predicted salary = 112.63 + 6.10 Position – 1.92
– 0.15 (Position–avg Position)

Interaction is not signficant

The t-ratio, confidence interval, or p-value all imply that the addition of the interaction term has not improved the fit of the model, so in the interest of *parsimony*, we ought to remove it and work with the simpler model.

Conclude

After checking assumptions (are variances comparable in these two groups) we find that a comparison of men with women having statistically adjusted backgrounds suggest that men are paid less. However, why do the women occupy lower level positions?

Key Take-Away Points

Categorical predictors in regression

– Test for differences between regression models fit to two groups.
By combining the two groups and fitting one multiple regression rather than two simple regressions, we can test for differences between

the slopes (i.e., test for interaction)

and between

the intercepts (test for shifts in the fit)

– Categorical terms alter intercept, interactions alter the associated slope.

– Relevant error assumption to check
Is the variation about the fit the same in all groups.

Next Time

Regression with more than two categories.

Building a regression model.