Categorical Variables, Part 2

Project Analysis for Today

First multiple regression

Interpreting categorical predictors and their interactions in the first multiple regression model fit in the project.

Review sessions

– This Friday, for project-related questions.
– Prior to the final exam.

Final class

Issues in extending a multiple regression to use more predictors.

Questions???

Ideas and Terminology for Today

Categorical variable in regression

– Group membership (e.g.: type of car, race, sex, religion).
– Used in regression to compare regressions across two or more groups.

Questions when using categorical variables in regression

– If they are parallel, are the intercepts different?
– Are the fits in the different models parallel? (i.e., Is interaction present?)
– Are the error variances comparable? (i.e., Is heteroscedasticity present?)

Interaction

– Important in many models, crucial in models with categorical.
Interaction with categorical ⇒ slope depends upon the group.
– *Question:* Does a predictor affect the response in the same way for each group?
– *Rephrased:* Is there an interaction between the predictor and the categorical variable that identifies the groups.
– Same concept as in anova:
Interaction implies that the effect (slope) of one predictor on the response depends on the value of other predictors. Recall profile plot?

Reading the output

– Write down the fit that is implied for each group, one at a time.
– Substitute +1, –1, or zero for the categorical terms.
– Collect the common terms and compare the fitted models.
– Roles: categorical terms affect the intercept
interaction terms affect the slopes

Effect tests

– Categorical variables for 3 groups add 2 predictors simultaneously.
– A categorical variable with k levels introduces k-1 terms into the fit.
– Has the addition of *all* of these new terms improved the model?
– JMP reports an F-test for each variable in the “Effect Test”
section of the regression output. (initial table shown)

JMP Commands

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

Often you can find the answer to your question by looking for the item of interest in the index of the JMP manual (the book that came with the software). In the back, it has a summary of all of the menu commands

Keeping up with the class is easier if you not only read the casebook examples prior to class, but also try to reproduce the output as well. Then you'll know what to look for in the class demonstration.

Three places to look for JMP commands:

(1) In the top menu.
(2) At the bottom of the analysis window.
(3) Specialized buttons near a plot or table.

How do you add a categorical variable to a regression?

Just add it as a predictor, even though the underlying column is not numerical. JMP will convert it to numerical values prior to using it in the regression.

How do you add an interaction to a regression?

This is a little tricky the first time. The key is to remember that an interaction involves a pair of predictors. First, put all of the relevant predictors into the model (like Origin and Manager). To form an interaction of a pair of predictors, highlight (select) one of the predictors which is included in the model and find the other in the overall list of columns. When both are selected, the "Cross" buttons comes to life. Click on it and it will form the interaction term as an added predictors (as in Origin*Manager). It does not matter which name comes first or second. (See page 172)

How did you color code the points?

Use the Color/marker by Col command from the Rows menu.

How did you get the separate fits in the Fit Y by X view?

Use the "Grouping variable" option at the bottom of the fitting menu (where you tell JMP to fit a line).

Review Example
Categorical Predictors with 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.

Categorical Predictors with Three Groups

Timing production runs ProdTime.jmp, page 189

Special features: 3 groups, effect tests (partial F tests), significant interaction.

Question

"Are three line managers doing equally well supervising production?"

Data

20 runs for each of three managers, relating time to number of units.

Analysis

– A marginal comparison of the run times of the three managers would not be appropriate since initially you do not know if the run sizes are comparable.

– Three fitted models for the separate managers (page 190-191) show differences in setup costs (intercepts) and slopes (added time for each additional item produced). The difference in slopes is an *interaction*. (Recall the profile plot from anova.)

– Are the differences significant? Fit the full model (page 194). Observe that the full model reproduces all three of the original fits (6 terms in the full model = 3 slopes and 3 intercepts).

– Find significant interactions. Slopes differ more than can be attributed to random variation in data. Slope for manager "B" is particularly distinct.

– Check assumptions, particularly for equal variation in the residuals across the three managers. (page 196).

Conclude

Substantial differences exist among managers, both in setup times and in how they handle the impact of increased production run size. Manager "C" gets the job started the quickest, but manager "B" does well for larger jobs.

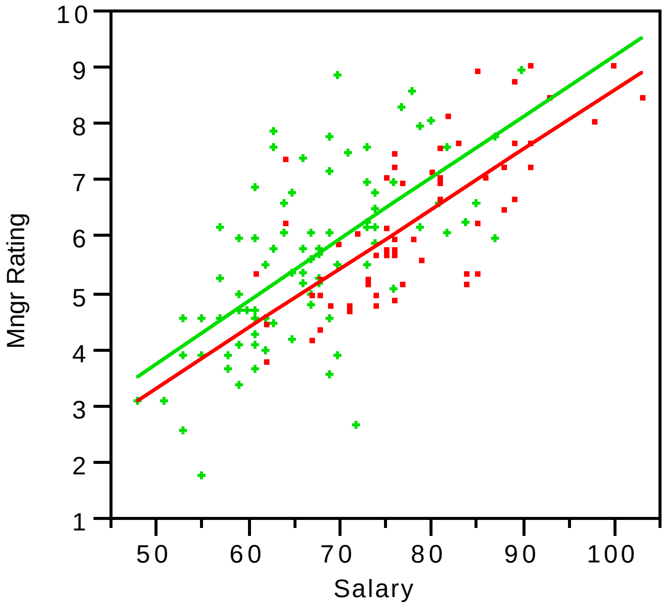
Further questions for management…

What does "B" do that helps make the large jobs run more quickly?
How does "C" get the process started quickly?
How can we help "A"?

Initial t-test shows a significant different: external do better

```
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
```

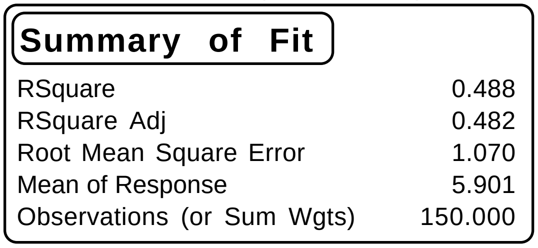
Separate regressions show the opposite: now appears that internal do better. Are the differences significant?



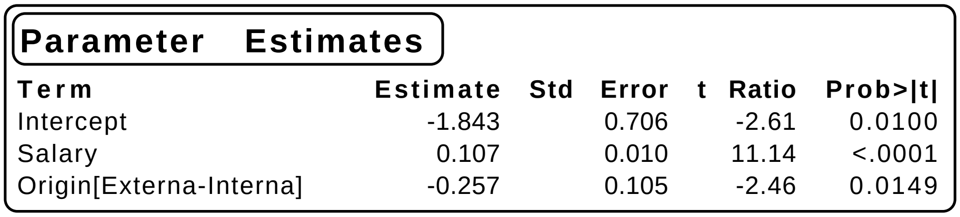
```
Term          Estimate    Std   Error   t  Ratio   Prob>|t|
Intercept    -1.693524     0.949254       -1.78     0.0779
Salary       0.1090929     0.014066        7.76     <.0001  Internal

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
```

First add a categorical variable (Origin) to a model with Salary. This model forces the fits in the two groups to be parallel. Is this reasonable?

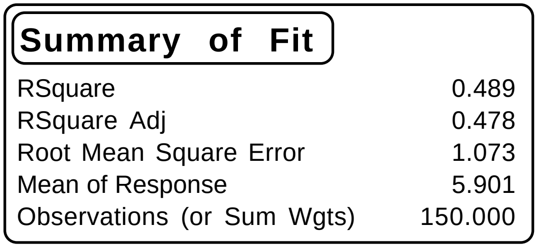


```
Summary of Fit
RSquare                          0.488
RSquare Adj                      0.482
Root Mean Square Error           1.070
Mean of Response                 5.901
Observations (or Sum Wgts)     150.000
```

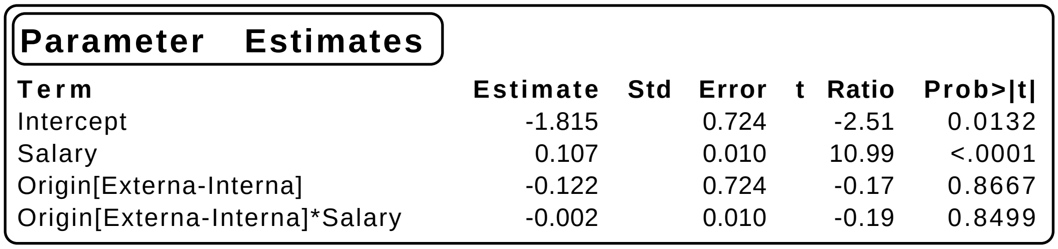


```
Parameter Estimates
Term                     Estimate  Std Error  t Ratio  Prob>|t|
Intercept                  -1.843      0.706    -2.61    0.0100
Salary                      0.107      0.010    11.14    <.0001
Origin[Externa-Interna]    -0.257      0.105    -2.46    0.0149
```

Fit as one multiple regression that combines these two simple regressions. This one multiple regression reproduces the fits of the two previous simple regression models.

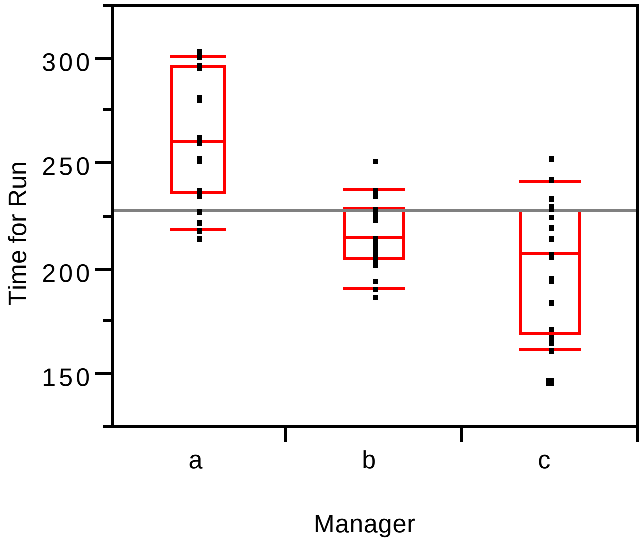


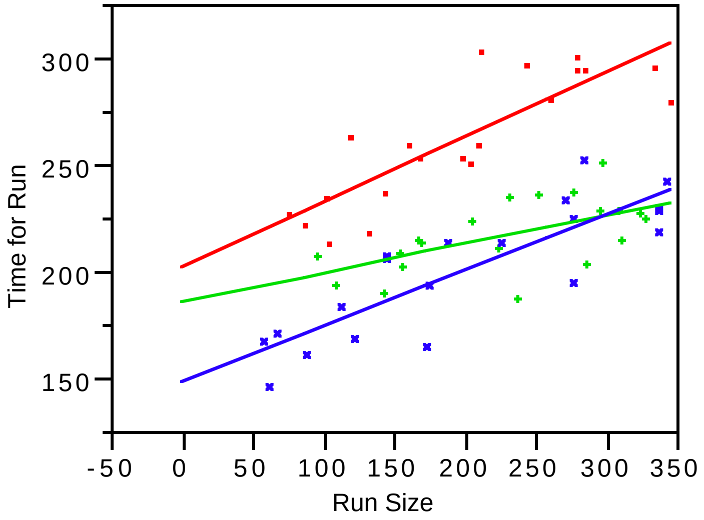
```
Summary of Fit
RSquare                          0.489
RSquare Adj                      0.478
Root Mean Square Error           1.073
Mean of Response                 5.901
Observations (or Sum Wgts)     150.000
```



```
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
Origin[Externa-Interna]*Salary    -0.002      0.010    -0.19    0.8499
```

Production Timing Example





Linear Fit Manager=a

Time for Run = 202.534 + 0.30727 Run Size (A is Red)

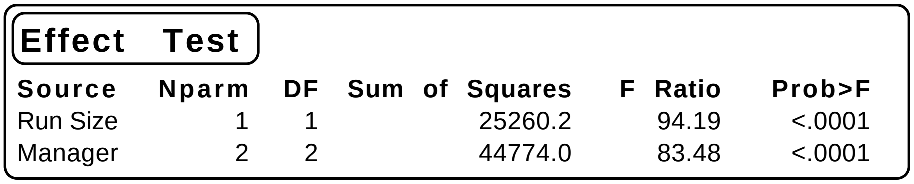
Linear Fit Manager=b

Time for Run = 186.494 + 0.13678 Run Size (B is Green)

Linear Fit Manager=c

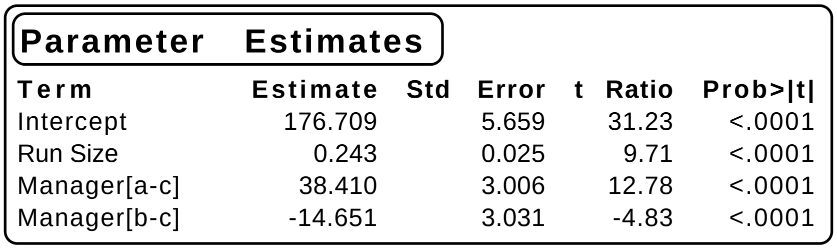
Time for Run = 149.748 + 0.25924 Run Size (C is Blue)

Fitting a multiple regression with Run Size and Manager as predictors forces a common slope. Not very reasonable to make that assumption here.



Effect Test

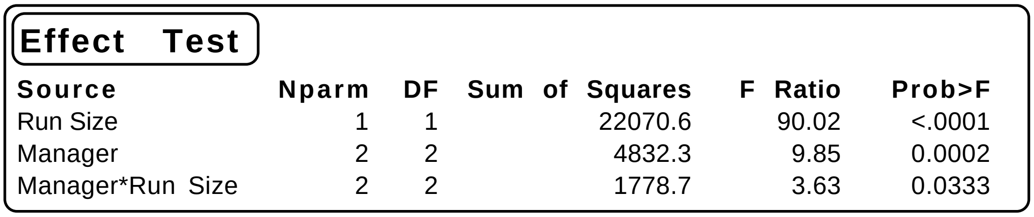
Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
Run Size	1	1	25260.2	94.19	<.0001
Manager	2	2	44774.0	83.48	<.0001



Parameter Estimates

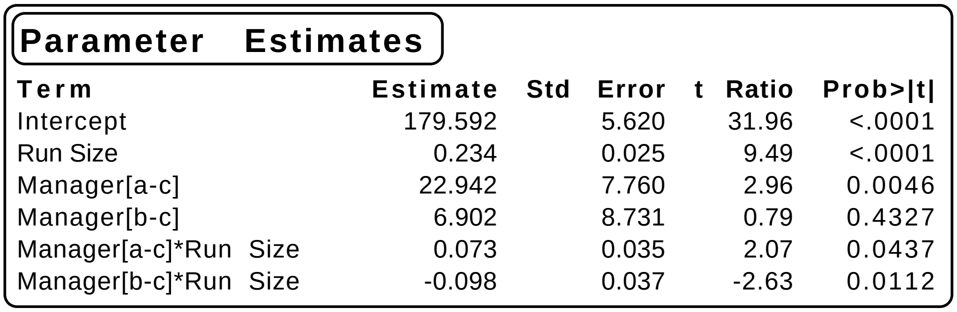
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	176.709	5.659	31.23	<.0001
Run Size	0.243	0.025	9.71	<.0001
Manager[a-c]	38.410	3.006	12.78	<.0001
Manager[b-c]	-14.651	3.031	-4.83	<.0001

Fitting a multiple regression with the interaction of Run Size and Manager added allows for differences among the slopes, and significantly improves the model.



Effect Test

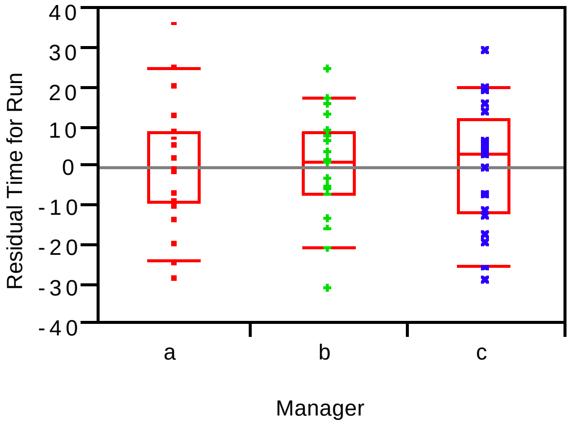
Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
Run Size	1	1	22070.6	90.02	<.0001
Manager	2	2	4832.3	9.85	0.0002
Manager*Run Size	2	2	1778.7	3.63	0.0333



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	179.592	5.620	31.96	<.0001
Run Size	0.234	0.025	9.49	<.0001
Manager[a-c]	22.942	7.760	2.96	0.0046
Manager[b-c]	6.902	8.731	0.79	0.4327
Manager[a-c]*Run Size	0.073	0.035	2.07	0.0437
Manager[b-c]*Run Size	-0.098	0.037	-2.63	0.0112

Need to check for constant residual variance.


Residual Time for Run
a
b
c
Manager