Categorical Predictors, Building Regression Models

Preliminaries

Supplemental notes on main Stat 621 web page

– Steps in building a regression model: see class web page.

– Glossary of terms (or consult the index in the casebook)

Review of Key Points from Class 8

Categorical predictors with two groups in regression

– Test for differences between regression models fit to *two* groups.
Combine the two simple regressions into one multiple regression, obtaining t-ratios for testing the
(a) difference between the slopes (i.e., testing for interaction) and the
(b) difference between the intercepts (i.e., testing for a shift).

– Roles of the terms
Categorical terms alter intercepts.
Interactions alter the associated slopes.

Need for the adjustment

– Confounding implies that differences in the mean response could be due to other differences between the groups rather than just the nominal labeling.

– Problem of confounding goes away with a randomized experiment or with balanced sampling designs.

Review Example

Fedex packaging

– Two groups, with the question of interest being to identify the group that offers the most sales with the least effort.

Common Questions

What's the difference between interaction and collinearity?

– Collinearity is correlation among the predictors
The model is "right", it's just hard to separate the predictors.

– Interaction implies that changing the value of one predictor affects the slope of another predictor:
The model without the interaction is "wrong" and not linear.

– Example in car context for two continuous predictors:
Adding horsepower might have a larger effect for smaller cars than larger cars.

– Interaction is most easily recognized and handled in models having categorical predictors.

If the multiple regression with categorical terms and interactions has the same fit as simple regressions fit within each of the groups, why go to this trouble?

Two reasons:

(a) First, to compare the slopes and intercepts. Without the output offered by the multiple regression (namely the t-ratios associated with the categorical terms), we have to resort to vague methods for comparing the fits. For example, we can check for overlapping confidence intervals, but this procedure is inadequate in regression.

(b) Second, what are you going to do when there is more than one predictor? Fit different multiple regressions to the separate groups? In such cases, the comparisons become even more difficult and numerous.

Project Analysis (Stage 1): Fitting initial models

Initial predictors

– Questions 1-5 restrict attention to limited subset of the collection of predictors. We can only “fill in the blanks” for a few predictors at this stage, so why use more than these?

– Predictors to consider:

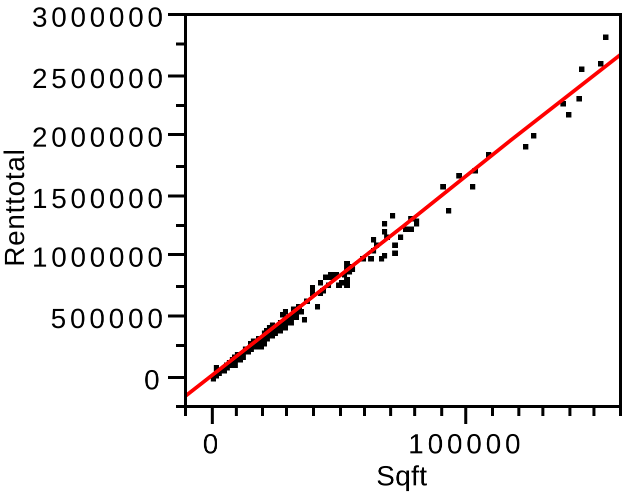
Location: city vs two types of suburban sites
Parking: how should this be represented
Amenities (wiring, exercise, restaurant)
Years since renovation
Fixed costs

– Scatterplot matrix of these: Note the general patterns (plot omitted)

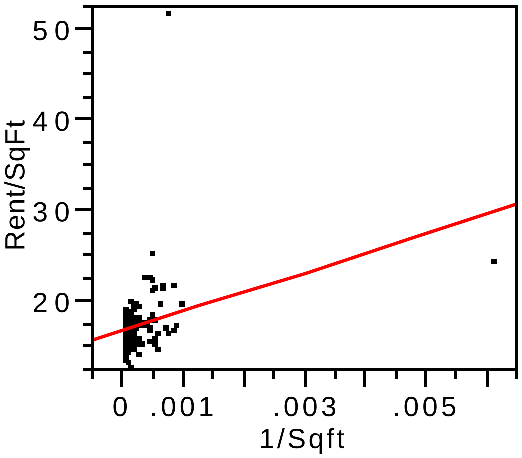
Initial approach from assignment #2: size matters

– May need to deal with outliers.

– Establish initial estimates of fixed and variable costs.



Renttotal = 287. + 16.6 Sqft



Rent/SqFt = 16.8 + 2133 1/Sqft

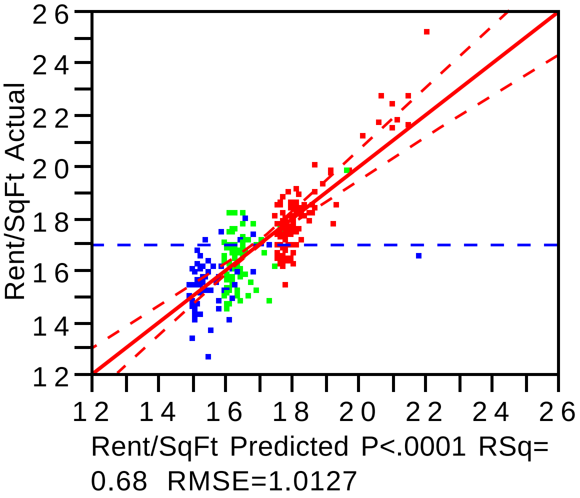
Color code by location

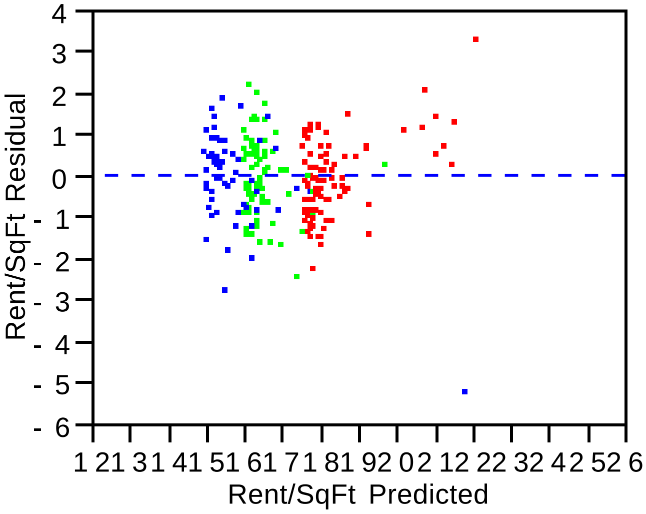
– Makes patterns associated with location apparent.

Build a multiple regression with the factors of interest

Summary of Fit

RSquare	0.677
Root Mean Square Error	1.013
Mean of Response	16.967
Observations (or Sum Wgts)	223.000



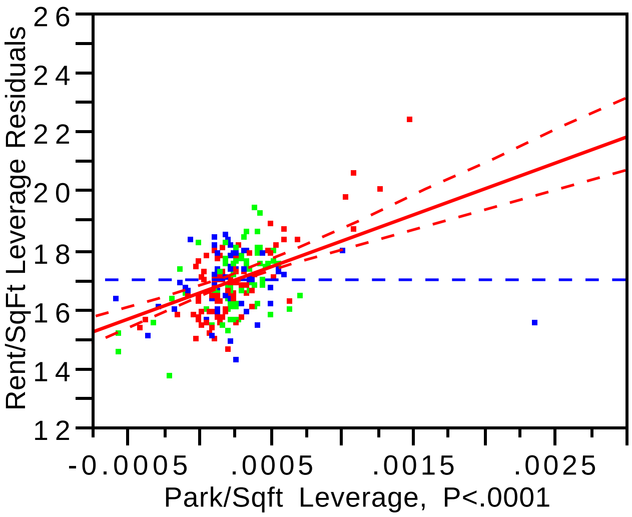


Expanded Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	16.135	0.133	120.96	<.0001
1/Sqft	2107.127	442.724	4.76	<.0001
Park/Sqft	1771.293	227.153	7.80	<.0001
Location[CITY]	1.309	0.111	11.76	<.0001
Location[SUBNEW]	-0.117	0.105	-1.12	0.2642
Location[SUBOLD]	-1.192	0.109	-10.91	<.0001
Renovation	0.004	0.005	0.76	0.4506
Restaurant[NO]	0.016	0.110	0.15	0.8810
Restaurant[YES]	-0.016	0.110	-0.15	0.8810
Wiring[NO]	-0.087	0.109	-0.80	0.4244
Wiring[YES]	0.087	0.109	0.80	0.4244
Exercise[NO]	-0.063	0.134	-0.47	0.6365
Exercise[YES]	0.063	0.134	0.47	0.6365

Check for problems

– Leverage plots for outliers, anomalous clusters



Some further questions:

– What assumptions are implicit in this model, and how can you check them?
Are the effects of the predictors the same in all of 3 of the locations?
Is the variability consistent over the different locations?

– BTW,
How do you work with categorical variables with more than two levels?

– Why/how has the estimate of fixed costs changed since the original simple regression?

– Why do the terms that are not significant not contribute to the model? Is it that they simply do not affect costs, or is it that they are redundant?
What is the mgmt. implication of whether or not they affect costs?

– What to do about the terms that are not significant? What, if anything do we learn about them from this output?

JMP-IN

– Save the model with the data table for later steps.

– Let JMP-IN do the predictions for you by adding an extra “dummy” row to the data table, filling in the columns as chosen by management.

Key Application for Today

Building models with more than two groups

– How does advertising affect sales in the different marketing regions?

– How can I compare the performance of three managers?

Concepts and Terminology

Effect tests

– Adding a two-level categorical variable (e.g., sex) …
- The *two* terms are redundant since they have opposite sign (i.e., they sum to zero)
- Two levels –> added one new term to model.
- Test either term with a t-test (same conclusion, opposite sign)

– Categorical variable for 3 groups (e.g., three managers)…
- The *three* terms are redundant since they have opposite sign (i.e., they sum to zero)
- Three levels –> added two new terms.
- How to test for the addition of both new terms at once?

– Question requires a *partial F-test* to answer.

– JMP-IN reports the relevant partial F-test in the “Effect Test” summary
- Don’t need to do this partial F by hand.

Examples for Today

Timing production runs **Prodtime.jmp, page 189**

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

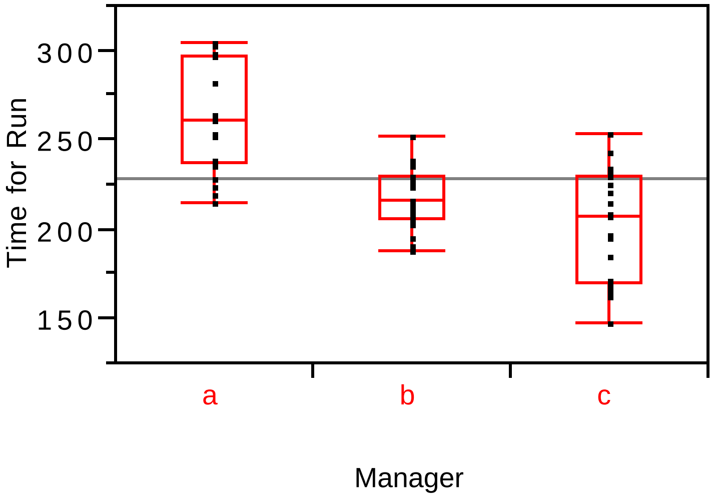
Data

20 production runs supervised by each of three managers, relating the time (in minutes) to complete the task and the number units produced in the job.

Analysis

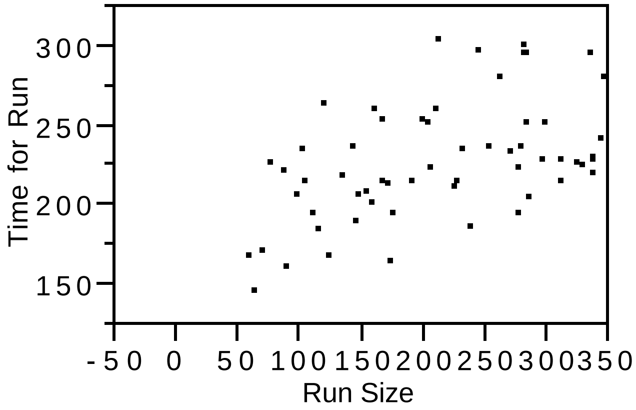
Marginal comparison

A marginal comparison of the run times of the three managers would not be appropriate since, at least initially, you do not know whether the run sizes are comparable.



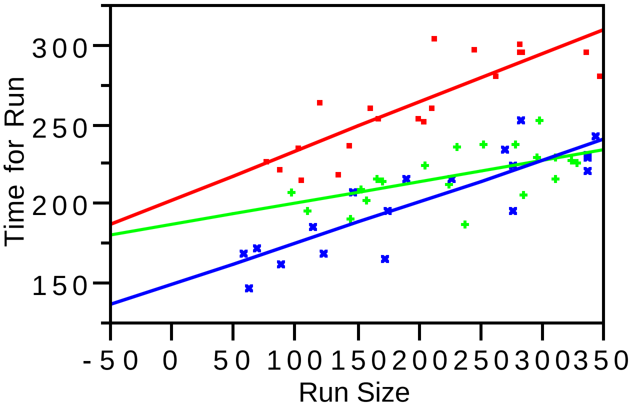
Possible confounding

Hard to identify what's happening without coloring the points.



With points color coded

An interaction (the slopes differ) and other features of the performances of the three managers becomes pretty clear. (p 190-191)



Immediate initial interpretation

Who do you want to use for a large job? For a small job? How do we tell if these apparent differences are meaningful (i.e., not just random noise)?

Results from model with categorical and interaction

It's clear we need an interaction term in this model, so let's include it right from the start. The full model (i.e., model with categorical and interaction) reproduces all three of the original simple fits.

Before looking at the slopes…

First check to see if there are significant effects (effect test)

Since there are more than 2 groups, we need to consider the results summarized in the Effect Test summary (partial F-test). The summary shows that the interaction is indeed significant, though just barely. Why is the interaction (so visible in the plot) not more significant?

Effect Tests

Source	**Nparm**	**DF**	**Sum of Squares**	**F Ratio**	**Prob > F**
Run Size	1	1	22070.61	90.02	<.0001
Manager	2	2	43981.45	89.69	<.0001
Manager*Run Size	2	2	1778.66	3.63	0.0333

After checking the effect tests, take a look at the rest.

Expanded Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	179.59	5.62	31.96	<.0001
Run Size	0.234	0.02	9.49	<.0001
Manager[a]	38.19	2.90	13.17	<.0001
Manager[b]	-13.54	2.94	-4.61	<.0001
Manager[c]	-24.65	2.89	-8.54	<.0001
Manager[a]*(Run Size-209.317)	0.073	0.04	2.07	0.0437
Manager[b]*(Run Size-209.317)	-0.098	0.04	-2.63	0.0112
Manager[c]*(Run Size-209.317)	0.025	0.03	0.77	0.4444

Practice: decode the fit for manager A

Predicted value = baseline model + effects for Manager A

= 179.59 + 0.234 Run Size + 38.19 + 0.073(Run Size–209.317)

= (179.59 + 38.19 – 0.07(209.317)) + (0.23+0.07) Run Size

= 202.50 + 0.307 Run Size

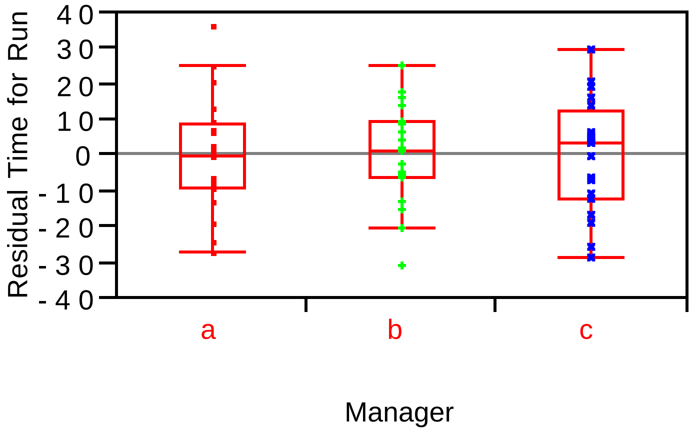
Notes

(1) This model for Manager A reproduces (up to rounding errors) the original simple regression for Manager A shown on page 191.

(2) The Manager terms add to zero as well as the interaction terms. Each set is redundant, so each adds only two new terms to the model.

Assumptions

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 (intercepts) and in how they handle the impact of increased production run size (slopes). Manager "C" gets the job started the quickest, but manager "B" does well for larger jobs.

Further questions for management include…
- What does "B" do that helps make the large jobs run more quickly?
- How does "C" get the process started so quickly?
- How can we help "A"?

Key Take-Away Points

Categorical predictors in regression

– Allow regression to estimate and test for differences among many groups, not just two groups.

– The baseline model serves as an overall point of reference, with the categorical terms representing how the fits for the separate groups differ from this common baseline.

Effect tests

– Test for the value of adding a categorical term (or the associated interaction) when the categorical term has more than two categories. When there are just two categories, the effect test is redundant with the t-test.

Next Time

The model-building process.

Overfitting when using automated methods to build regression models.