Building a Regression Model

Preliminaries

Practice questions on my web page

– Review questions covering regression with categorical predictors.

Office hours

– Monday, Wednesday 3-5:30, Tuesday from 12-3.

– E-mail in evenings

– Etiquette
If office hours become crowded, please try to limit your questions while others are waiting. Also, please do not expect TAs to use computer to analyze *your* project data.

Review of Key Points from Lecture 9

Categorical predictors in regression

– Allow model to incorporate differences among many groups:
(a) differences in intercepts (as captured by the categorical terms)
(b) differences in slopes (as captured by the intercepts)

– The baseline model is a point of reference, with the categorical terms representing how the fits for the separate groups differ from this common baseline model.

Effect tests

– Test for the value of adding a categorical term (or the associated interaction) when the categorical term has more than two categories.

– Like other F tests, the partial-F associated with a categorical factor having 3 or more levels (or its interaction) does not isolate where the difference among intercepts (or slopes) occurs. It only indicates if such a difference occurs, but does not pinpoint where.

Building a Regression Model: An Outline

Before you gather and look at the data...

– Identify the question of interest, the goals of the analysis.

Prediction	In/out of sample? Extrapolation? Allowable margin for error? What sort of RMSE is acceptable?
Interpretation	Does the estimate "make sense"? Is there collinearity? How much? Need marginal or partial slope?

– Anticipate important features of the model.
  Which variables do you expect to find important?
  Do you anticipate nonlinear patterns or interactions?
  What do you expect the coefficients to be? e.g., positive or negative?

– Evaluate the data.
  Is there enough? (role of preliminary or "pilot" study)
  Is the data representative? (sampling biases, measurement error)
  Is there a hidden "clumping" factor? (dependence)

Assess the univariate and marginal relationships

– Identify scales, ranges, distributions of the various factors.
  Are data normal or skewed? Outliers present?
  Distribution command (i.e., look at histograms, frequency tables)

– Look at bivariate scatterplots of factors, time series plots (if appropriate).
  Nonlinear (curvature)? Outliers, leverage points?
  Marginal associations with response?
  Correlation among predictors? (suggests collinearity)
  Multivariate command, scatterplot matrix.

– Check for special features in the data.
  Discrete data, "hidden" categories?
  Color code data for important categorical factors.
  Use *Color by Col* command from *Rows* menu.

Fit an initial model

– Modeling is an iterative process. No one gets it right the first time.

– Fit the model suggested by your understanding of the problem, in form that makes the most sense given the context. Often the needs of a "client" dictate the structure of the model.

– Assess the parameters (slopes, intercept) of the fitted model, focusing on a mixture of statistics and substance.

Does model explain much variation in data? (RMSE, F and R^2)
Are estimates significant? What is the length of CIs?
Can you interpret the slopes, using appropriate units?
How do the partial slopes differ from the marginal slopes?
What is the impact of collinearity? Can you ignore it?

- Evaluate your model graphically.

Do leverage plots indicate problems? Unusual subsets?
Do leverage points, outliers affect the fit?
Are residuals reasonable (i.e., constant variance, normal)?
(Don't dwell on these until get a decent model.)

Revise the fitted model.

– Procedure depends on the use of the model

– Key use is interpretation

Collinearity is an issue, since it obscures the effects of predictors.
VIF, collapsing leverage plots

– Key use is prediction

Parsimony is essential: don't use factors that are not contributing significantly to the model – they only add error to the prediction.
Check these with the t-statistics, effect tests.

– Identify other omitted factors

Are variables appropriately transformed?
What factors explain the unexplained residual variation?

– Use a cautious, one-at-a-time strategy.

Removing several is dangerous if collinearity is present.
Check for missed nonlinearity.

Go back to the prior step until satisfied. Then consider these:

– Make sure that you can interpret the end result.

– Make sure that you can answer the question of interest.

– Run a careful check of residuals
  Does anything in the analysis suggest dependence?
  Do different groups have comparable variance?
  Are the data normal (quantile plot from saved residuals)?

– Review regression handout!

Plan next steps.

– Determine how to communicate results to others. Know your audience.
  Do they know statistics?
  Do they appreciate subtleties of analysis, such as plots?
  What common beliefs does your analysis support? Contradict?

– Focus on things that would make analysis simpler, better.
  What data are missing? Which predictors are missing? (FedEx case)
  Would more data help? Remember, more data without a better model does not improve in-sample prediction accuracy by much.

Project Analysis (Stage 2)

Administrative

– Read the project description! Answer questions as presented.
  • Use one model for *all* of 6-10.
  • Heed page/layout instructions.
  Limit to 1, 5, and 10 **double-spaced** pages, respectively.

– Discussion with classmates allowed, responsibility for your report is yours.

– Save time to write your report… Due Friday, **October 13, 3 p.m.**.

Update from last model

– Checked some other interactions
  Parking/Sqft x *Location*
  1/Sqft x *Location*
  Only the parking effect varies by location. (Effect test).

– Why consider these interactions? Should we check others as well?

– Why does the apparent difference in fixed costs seen in the original plots "go away" in this multiple regression?

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
1/Sqft	1	1	16.58	20.32	<.0001
Park/Sqft	1	1	26.72	32.74	<.0001
Location	2	2	150.05	91.93	<.0001
Renovation	1	1	0.29	0.35	0.5538
Restaurant	1	1	0.04	0.05	0.8265
Wiring	1	1	1.19	1.46	0.2276
Exercise	1	1	0.18	0.22	0.6379
Location*1/Sqft	2	2	0.58	0.35	0.7020
Location*Park/Sqft	2	2	26.29	16.11	<.0001

Key Technique for Today

Assessing the differences among many groups

– When the effect test is significant, how should I decide which differences are important?

– The effect test (i.e., the partial F) indicates a difference among the associated intercepts or slopes, but does not indicate which differences are meaningful.

Concepts and Terminology ("Class 9" in casebook)

Multiple comparisons

– Easy to find a significant effect when many comparisons are made.

–Want to do 10 "honest" t-tests at one time?
The "Bonferroni method" suggests that you should use a p-value of 0.05/10 = 0.005 for each one rather than usual 0.05 cutoff. (p 229-32)

– Moral: Being "punished" for not having better, sharper theory.

– *Same issues arise in regression*, such as when you go exploring for other interaction terms in a model or consider many combinations of predictors.

Tukey-Kramer comparisons

– Alternative approach to Bonferroni, one better suited to pairwise comparisons: e.g. Which locations differ from which other locations?

– Goal
Locate important differences while avoiding false claims of significant differences.

JMP-IN Methods

Where are the Tukey-Kramer comparisons?

– See the output associated with the categorical term in the regression… near its leverage plot.

– Use the "magic" red triangle to get the Tukey-Kramer comparison table. (We did this back in Stat 603 in case this seems at all familiar.)

Casebook Example

Selecting the best vendor Repairs.jmp, page 233

"Is there a difference among the vendors in cost of service, or can the observed differences be attributed to chance?"

Data

– 10 service calls for each of 10 vendors
– Response is the price of a comparable type of computer repair.

Analysis

– Multiple regression using just one categorical variable
– t-ratios for the slopes (just differences from the overall average) suggest that costs for Vendor 1 are significantly less than overall costs
– Is this analysis appropriate? (p 234)

Expanded Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	150.39	1.12	134.07	<.0001
Vendor[1]	*-7.39*	*3.37*	*-2.20*	*0.0307*
Vendor[2]	3.11	3.37	0.92	0.3579
Vendor[3]	-4.69	3.37	-1.39	0.1668
Vendor[4]	0.01	3.37	0.00	0.9976
Vendor[5]	0.91	3.37	0.27	0.7875
Vendor[6]	-1.09	3.37	-0.32	0.7468
Vendor[7]	2.11	3.37	0.63	0.5322
Vendor[8]	4.21	3.37	1.25	0.2142
Vendor[9]	2.61	3.37	0.78	0.4400
Vendor[10	0.21	3.37	0.06	0.9504

Adjusting for multiplicity

Many tests increase the chances for "false positives", i.e., thinking we have a significant effect when none is present. That's what happens in this regression. To fix this problem, we can use

• *Bonferroni* Rather than compare 10 p-values to 0.05, compare them to the threshold 0.05/10 = 0.005. Nothing is significant.

• *F-ratio* Rather than look at each factor, test whether overall there is some difference among them. The overall F is not significant: no significant difference.

Little reason to go farther here…

Since there's no difference indicated by the F, we have little reason to continue. Sometimes you will subsequently find a significant effect, but not often. We'll continue here to illustrate the tools.

Tukey-Kramer comparison

Using the "LSMeans Tukey HSD" option, JMP produces a comparison table. The table highlights significantly different values. None are highlighted here – none of the vendors is statistically significantly different from another.

Conclude

Differences among the vendors are *not* significant. Get more data and have a more focused comparison next time.

The casebook illustrates the "deceptive" use of an inappropriate analysis to conclude a significant effect does exist (p 242).

Back to the Project

Multiple comparisons relevant in 2 ways

(1) Comparisons among the several locations

(2) Lots and lots of t-tests associated with all of the models some will try to fit in modeling their data.

– Statistical tests "reward" persistence with false positives.
That is, if you try enough predictors, you will find things to be significant by chance alone even though such a factor is actually irrelevant.

Location example from multiple regression

– Which locations have different costs per square foot when adjusted for the other factors in the prior multiple regression?

– Effect tests say a difference exists, but where?

– Tukey Kramer analysis says the differences are significant everywhere. E.g., city costs 2.36$/SqFt more than old suburbs, and this is significant. The 95% interval is $1.95 to $2.77 more per square foot.

LSMeans Differences Tukey HSD

Alpha=0.050 Q=2.36051 LSMean[i] By LSMean[j]

Mean[i]-Mean[j] Std Err Dif Lower CL Dif Upper CL Dif	CITY	SUBNEW	SUBOLD
CITY	0 0 0 0	1.37575 0.18649 0.93555 1.81596	*2.36148* 0.17457 1.94941 2.77356
SUBNEW	-1.3758 0.18649 -1.816 -0.9355	0 0 0 0	0.98573 0.18347 0.55265 1.41881
SUBOLD	-2.3615 0.17457 -2.7736 -1.9494	-0.9857 0.18347 -1.4188 -0.5527	0 0 0 0

Application to interactions?

– Sorry, JMP will not do these for you (its not a simple comparison of average effects – it's the difference in slopes).

– Stick to the effect tests and common sense for these.

Not needed with 2-level categorical comparisons.

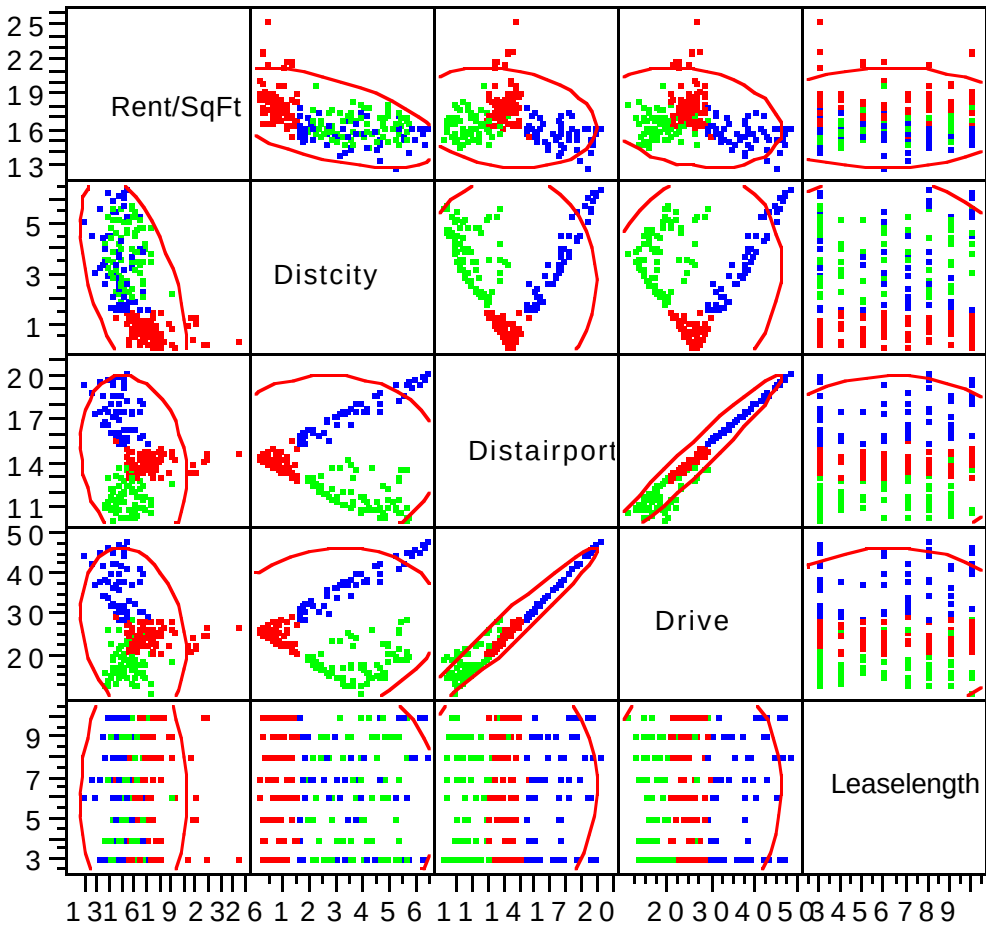
– Only comparing one to another, so T-K not needed here. Only doing one comparison, so no need to use a special method.

Project Analysis: Revising the Model and Adding Factors

More client issues

– Positioning: Distance to airport, time to airport
Lease: renewable leases, lease length
Building features: # elevators, % occupied

– Wise to explore these new factors graphically before add to regression.
Some pretty "interesting" patterns, especially with the distances.



– What happens when these (and the other factors of interest) are added to prior regression model?

Revised model

– RMSE drops to 0.87 $/SqFt with R^2 at 77% for this data set.

– Your results could be better or worse.

Slopes associated with new factors

– Only shows the slopes for the added terms.

– Want to keep all of these? Which are useful?

– Collinearity relevant?

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|
Intercept	15.6489	1.3234	11.83	<.000
...	.	.	.	
Distairport	-0.0220	0.1327	-0.17	0.868
Drive	-0.0154	0.0334	-0.46	0.645
Leaselength	-0.0452	0.0244	-1.85	0.065
Elevator	-0.0016	0.0286	-0.06	0.955
Occupancy	2.0434	0.5064	4.04	<.000
Renewable[NO]	-0.0990	0.1102	-0.90	0.370

Reaching closure

– Time to start giving residuals more attention, checking assumptions.

– Did we miss any features associated with any of these predictors (look back at Assignment 2)

– What other factors, interactions might be relevant?
Need a prediction for the final offering? Is it a good price?

Key Take-Away Points

Project

– Deciding which predictors to retain, which to exclude

– Importance of keeping the goals of the model in mind.

Multiple comparisons

– If you are persistent and use the p-value < 0.05 rule (or |t|>2), you will find significant effects, whether real or not, by doing enough tests.

– Bonferroni procedure offers simple protection (0.05/number of tests)
Tukey-Kramer helps when you want pairwise comparisons of averages.

Next Time

Project analysis continues

– Focus on how to answer questions, executive summary

– Diagnostics, other subtle effects.

Analysis of variance

– Experiments

– Conjoint analysis and marketing research.