Regression Summary

Project Analysis for Today

First multiple regression

- Interpreting the location and wiring coefficient estimates
- Interpreting interaction terms
- Measuring significance

Second multiple regression

- Deciding how to extend a model
- Diagnostics (leverage plots, residual plots)

Review Questions: Categorical Variables

Where do those terms in a categorical regression come from?

- You cannot use a categorical term directly in regression (e.g. 2(“Yes”)=?).
- JMP converts each categorical variable into a collection of numerical variables that represent the information in the categorical variable, but are numerical and so *can be used* in regression.
- These special variables (a.k.a., dummy variables) use only the numbers +1, 0, and –1.
- A categorical variable with k categories requires (k-1) of these special numerical variables. Thus, adding a categorical variable with, for example, 5 categories adds 4 of these numerical variables to the model.

How do I use the various tests in regression?

What question are you trying to answer…

- Does this predictor add *significantly* to my model, improving the fit beyond that obtained with the other predictors? (t-ratio, CI, p-value)
- Does this collection of predictors add *significantly* to my model? Partial-F (Note: the only time you need partial F is when working with categorical variables that define 3 or more categories. In these cases, JMP shows you the partial-F as an “Effect Test”.)
- Does my full model explain “more than random variation”? Use the F-ratio from the Anova summary table.

How do I interpret JMP output with categorical variables?

```
Term                         Estimate  Std  Error  t  Ratio  Prob>|t|
Intercept                      179.59        5.62      32.0      0.00
Run Size                         0.23        0.02       9.5      0.00
Manager[a-c]                    22.94        7.76       3.0      0.00
Manager[b-c]                     6.90        8.73       0.8      0.43
Manager[a-c]*Run  Size           0.07        0.04       2.1      0.04
Manager[b-c]*Run  Size          -0.10        0.04      -2.6      0.01
```

- Brackets denote the JMP's version of dummy variables.
- Letters within the brackets tell you which values are used.

  Manager[a-c] Is 1 for manager a, –1 for c, 0 otherwise
  Manager[b-c] Is 1 for manager b, –1 for c, 0 otherwise

		Manager	
	A	B	C
Manager[a-c]	1	0	–1
Manager[b-c]	0	1	–1

- The fit for manager “a”, for example, uses the bracketed terms that include the letter “a”. The rest of the bracketed terms are zero.

  Fit for a is: Time = 180 + 0.23 Size + 23 + 0.07 Size
  = (180 + 23) + (0.23 + 0.07) Size
  Fit for b is: Time = 180 + 0.23 Size + 7 – 0.10 Size
  = (180 + 7) + (0.23 – 0.10) Size
- The fit for one category (the last, alphabetically) uses all of the bracketed terms, with the signs reversed (since the numerical variables that represent Manager in this example are all –1 for Manager c):

  Fit for c is: Time = 180 + 0.23 Size – (23 + 7) – (0.07 –0.10)
  = (180 – 30) + (0.23 + 0.03) Size
- Note that these 3 fits are those (up to rounding) found by fitting a simple regression to the data for each manager (page 191).

If the fits are the same, why bother with categorical variables?

- With the 3 fits together in one model, you can find differences among the slopes and intercepts without using heuristic methods.
- Tests based on the partial F test, as summarized in Effect Test summary in the JMP output.

Source	**Nparm**	**DF**	**Sum of Squares**	**F Ratio**	**Prob>F**
Run Size	1	1	22070.6	90.0	<.0001
Manager	2	2	4832.3	9.9	0.0002
Manager*Run Size	2	2	1778.7	3.6	0.0333

- In this example, the F-ratio for Manager (the categorical variable) implies significant differences among intercepts (p=.0002).
- Interaction term implies significant differences among the slopes (p=.033)

Should I take out Manager[b-c] since its not significant?

- Did you add this particular term? No.
  You added *Manager* as a whole, and are wise to leave it that way.
- The partial F-test (effect test) is much like the F-test that we used in the analysis of variance. It indicates whether some difference exists, but does not pinpoint its location.

What should I do about collinearity with interaction terms?

- Adding an interaction term often introduces collinearity, and collinearity can make the categorical term appear insignificant.

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

- Unless the interaction term is significant (partial F in the Effect Test), remove the interaction.
- If both the interaction and categorical term are insignificant, remove the interaction rather than the underlying categorical term. (page 173)

Why are the fitted lines parallel when I add a categorical term?

- With only one slope allowed for the continuous predictor, the regression model is forced to use the same slope for all categories.
- To allow different slopes, you *must* add an interaction term that combines the categorical variable with the continuous predictor.

I have several predictors. Do I need to use all of the interactions?

- Ideally, you ought to consider all of them.
- In practice, you ought to consider the ones that are most interesting from a substantive point of view.
- e.g., If you want to know whether advertisting impacts sales differently in different regions, you need an interaction of advertising with region.

Building a Regression Model

1. Before you gather and look at the data...

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

Prediction In/out of sample?
Allowable margin for error?
Interpreting slopes Expect 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 it representative? (sampling biases, measurement error)
Is there a hidden “clumping” factor? (dependence)

2. Assess the univariate and marginal relationships…

- Identify scales, ranges, distributions of the key factors.

Are data normal or skewed? Outliers present?
Distribution of Y command.

- 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)
Correlation of Y’s command, scatterplot matrix option.

- Check for special features in the data.

Discrete data, “hidden” categories?
Color code data for special categorical factors.
Use *Color by Col* command from *Rows* menu.

3. Fit an initial model…

- Fit the model suggested by your understanding of the problem.

Start with a regression that makes the most sense given your understanding of the context and data.

- Assess the coefficients of fitted model, using both substance and statistics.

Does your model explain much variation in the data? (F, RMSE, R^2)
Are the slopes significant? What is the size of confidence interval/SE’s?
Can you interpret the slopes, using appropriate units?
How do the partial slopes differ from the marginal slopes?
What is the impact of collinearity? Do you care, or should you ignore?

3. Fit initial model, continued.

- Evaluate model graphically.

  Do leverage plots look suitable for simple regression fits?
  How do leverage points or outliers affect the fit?
  Are residuals reasonable (i.e., constant variance, normal)?
  Do this initial check from the plot of residuals on the fitted values.

4. Revise the fitted model.

- Remove factors that are not needed.

  Are some predictors redundant? (VIFs, collapsing leverage plots)
  Small |t-ratio| < 2 (wide confidence interval, large p-value)

- Determine whether other, previously excluded factors are needed.

  Are variables appropriately transformed?
  What factors might explain the unexplained residual variation?
  Is an interaction called for?

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

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

5. Go back to step 2 until satisfied.

- Make sure that you can interpret the end result.
- Make sure that you can or cannot answer the question of interest.
- Run a careful check of residuals

  Does anything in the analysis suggest dependence?
  Do different groups have comparable variance?
  Do larger values on some scale have growing variance?
  Are the data normal (quantile plot from saved residuals)?

- Review regression handout from web page.

6. 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? (i.e., which predictors or groups are not known?)
  Are data really representative of new situations?
  Would more data help? Remember, more data without a better model does not improve in-sample prediction accuracy very much.

Concepts and Terminology

Stepwise regression

- Example of an automated method for finding a regression model.
- Other *data mining* methods include “neural networks”.
- Most useful when you know little about the problem, searching.
- Often combined with “response surface” option in JMP.
- Picks the predictors using a “greedy” incremental approach.

Overfitting

- Definition: Using more predictors in a regression than you should.
  That is, true slope is zero, but by chance your estimate differs from zero.
  As a result, the regression is more complex than it should be.
- Model claims high accuracy (low RMSE, high R^2), but in fact predicts poorly.
- Frequently occurs when automated method is asked to choose from too many.
- See casebook example (page 220) that illustrates how automated methods lead to a model for predicting stock returns that looks fine when fit to historical data ($R^2 > 75\%$) but predicts poorly.

Cross-validation

- Common sense idea for checking properties of regression model.
- Reserve some fraction of your data for testing the fit of the model.
- Try to predict this “hold back” sample using model fit to rest.
- Compare RMSE of model to SD of prediction errors.
  RMSE – estimates SD of errors from data used in modeling
  SD of prediction errors – should estimate same quantity

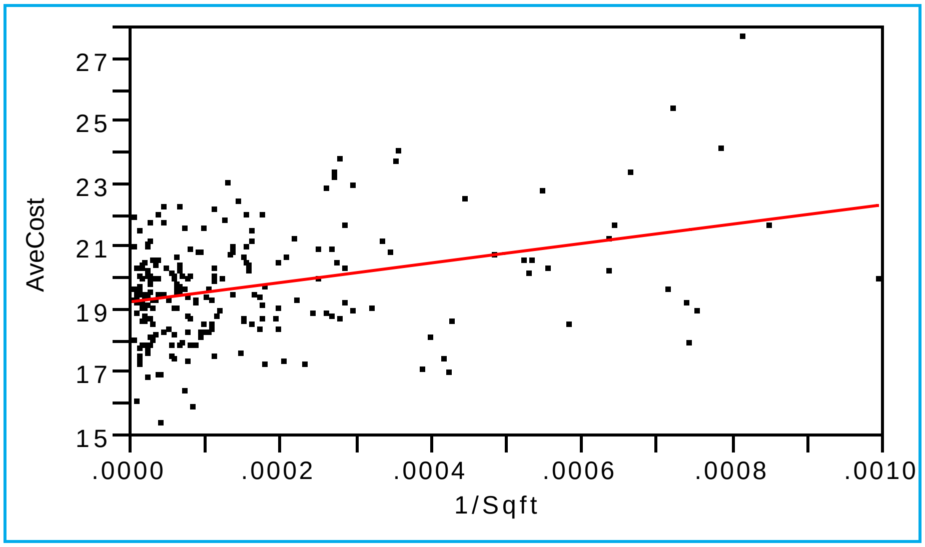
Bonferroni rule

- Simple method for validating regression models, avoid overfitting.
- Avoids need for a validation sample.
- Compare t-ratio to threshold which is larger than just 2 to make it harder for predictors to enter your model.
- Easily done by comparing p-values to 0.05/(# considered) rather than 0.05.
  e.g. If you have considered 20 predictors, then compare each t-ratio to
  $$0.05/20 = 0.0025$$
  rather than 0.05 to decide which predictors to keep.

Sample Project Analysis

Initial analysis

Not a whole lot of variation explained ($R^2 = 12\%$), with fixed costs of about $3100 and variable costs of $19.30/SqFt.

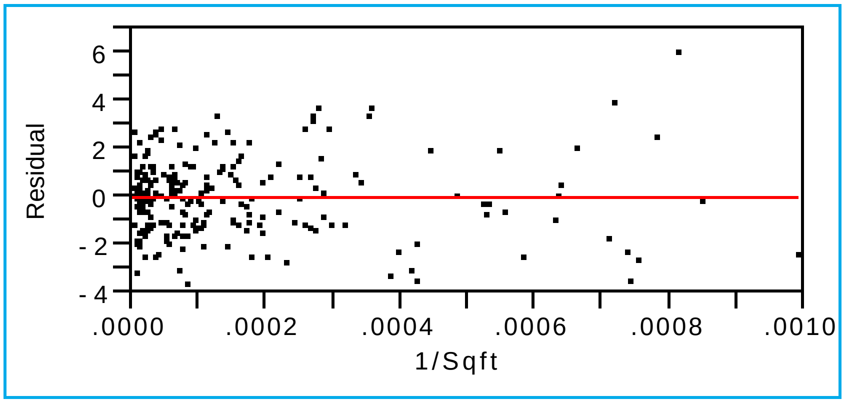


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	19.30	0.14	136.06	<.0001
1/Sqft	3072.29	569.78	5.39	<.0001

Summary of Fit

RSquare	0.116
RSquare Adj	0.112
Root Mean Square Error	1.637
Mean of Response	19.788
Observations (or Sum Wgts)	223.000



First multiple regression analysis

After skimming over data using scatterplot matrix, fit the indicated multiple regression as directed in project instructions. The goodness-of-fit improves considerably. (R^2 is higher, RMSE lower)

Summary of Fit

RSquare	0.601
RSquare Adj	0.583
Root Mean Square Error	1.122
Mean of Response	19.788
Observations (or Sum Wgts)	223.000

Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
1/Sqft	1	1	4.88	3.87	0.0504
Location	2	2	66.08	26.23	<.0001
Location*1/Sqft	2	2	0.03	0.01	0.9900
Parking/Sqft	1	1	31.50	25.01	<.0001
Location*Parking/Sqft	2	2	6.55	2.60	0.0766
Renovation	1	1	1.20	0.95	0.3296
Wiring	1	1	2.33	1.85	0.1752

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	19.144	0.158	121.54	<.0001
1/Sqft	1036.670	526.863	1.97	0.0504
Location[CITY-SUBOLD]	1.124	0.155	7.24	<.0001
Location[SUBNEW-SUBOLD]	-0.147	0.149	-0.99	0.3252
Location[CITY-SUBOLD]*1/Sqft	-7.222	730.178	-0.01	0.9921
Location[SUBNEW-SUBOLD]*1/Sqft	89.588	642.324	0.14	0.8892
Parking/Sqft	1216.024	243.169	5.00	<.0001
Location[CITY-SUBOLD]*Parking/	624.250	286.150	2.18	0.0302
Location[SUBNEW-SUBOLD]*Parking/	-109.429	291.808	-0.38	0.7080
Renovation	-0.005	0.005	-0.98	0.3296
Wiring[NO-YES]	-0.143	0.105	-1.36	0.1752

How do average lease costs depend upon location?

Significant variation in average variable costs (p-value < .0001). City costs about $20.30/Sqft(19.14+1.12), old suburbs about $18.17 (19.14–1.12+0.15), and new suburbs about $18.99 (19.14–0.15).

How much does parking cost in the city for one spot?

In the city, parking costs on average $1216+624=$1840 per spot.

How much does wiring add to costs cost, on average?

A bit less than $0.30, with ranging over the interval 2 [14.3±2(.105)]. Note that the interval includes zero.

Do fixed costs vary by location?

No, not significantly: interaction Location*(1/Sqft) is not significant. Any differences evidently already taken into account by handling of parking costs.

How much would 50,000 square feet of space with 20 executive parking spots, new wiring, and one year since renovation be expected to cost, in total?

Using a dummy observation with these factors (you might choose to remove some the insignificant interactions), JMP computes the predicted cost per square foot (for a city location – others will differ) as

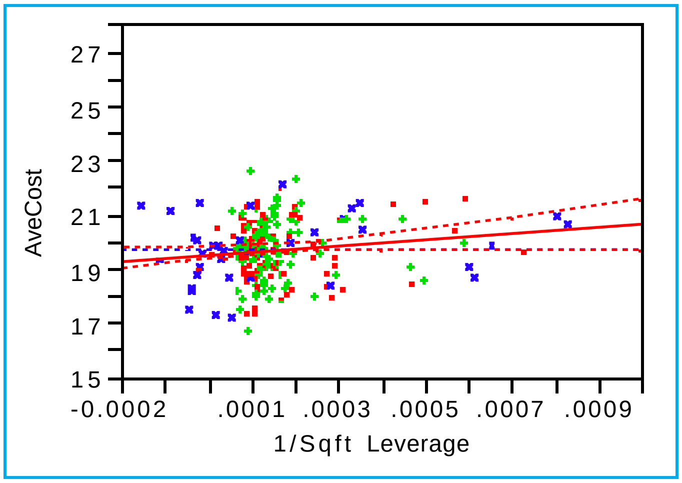
$21.162/SqFt times 50,000 $1.06 million

JMP will also do prediction intervals for you as well, which should roughly correspond to 50,000 [$21.162/SqFt ± 2 RMSE] or here about

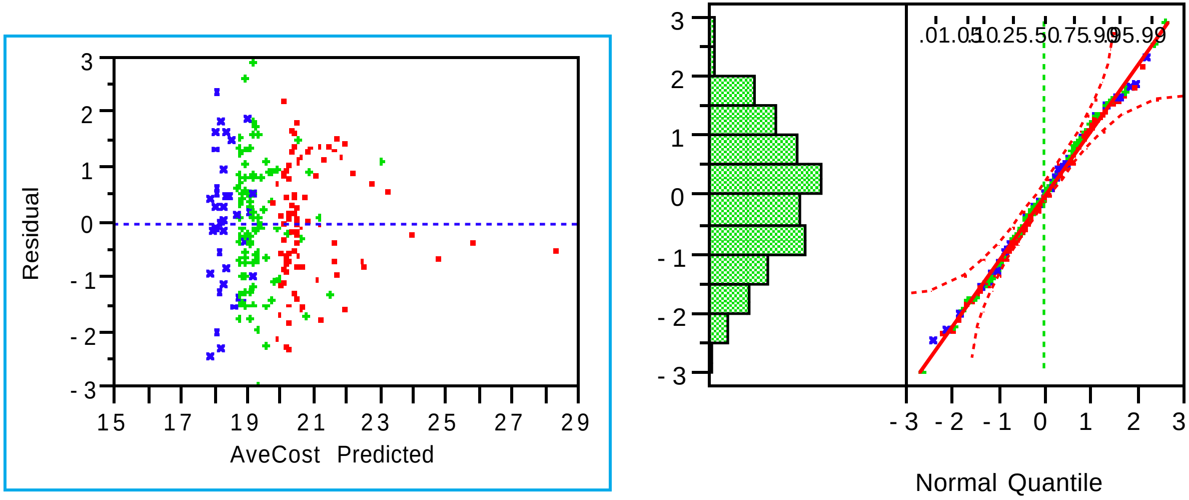
50,000[18.893, 23.431] ≈ [$945,000 , $1,170,000]

Assumptions?

Leverage plots



Residual plots



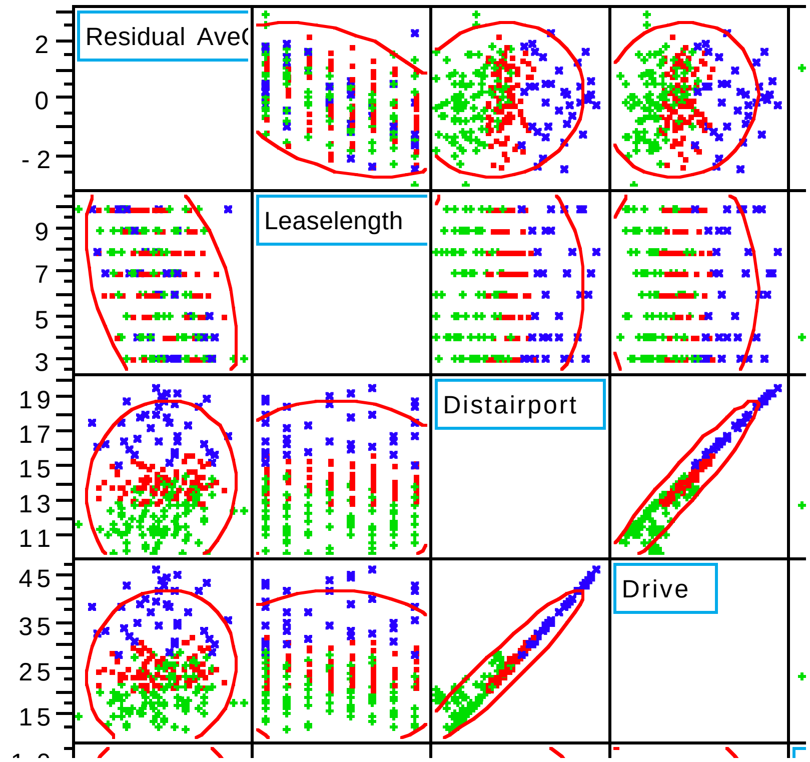
What other factors might improve the fit of this model?

One way is to look for factors related to the unexplained variation, namely the residuals…

Correlations

Variable	Residual	AveCost	Leaselength	Distairport	Drive	Occupancy
Residual AveCost		1.000	-0.411	0.086	0.080	-0.023
Leaselength		-0.411	1.000	0.003	-0.011	0.055
Distairport		0.086	0.003	1.000	0.953	-0.112
Drive		0.080	-0.011	0.953	1.000	-0.144
Occupancy		-0.023	0.055	-0.112	-0.144	1.000

Scatterplot Matrix



Another approach is to fit a really, really big regression (the “kitchen sink” approach) and try to sort out what you find, or you can let JMP’s stepwise regression tool sort through some of the factors for you. That can be a bit risky, though, and hard to interpret (see casebook for examples).