Collinearity and Multiple Regression

Preliminaries

Deliverables

– Postpone Assignment #2 until Friday at **3** p.m.

Academic reps / quality circle

– Need to hear from you so we can meet

Review of Key Points

Marginal and partial slopes in regression

– Which is the right one to use? What question is being asked?

– Different views of regression and associated slopes
(1) geometry
(2) path analysis
(3) comparison of averages

– Causation versus association
Regression finds association; we often interpret it as finding causation.

Collinearity

– Defined as simply correlation among the predictors in a multiple regression. Because of this “redundancy”, collinearity entangles the effects of the predictors, complicating the interpretation.

– Special case: Marginal slope = partial slope if no collinearity

Inference and testing

– New interpretation of a t-test as measuring the improvement offered by adding a single predictor to a model that includes all of the others.

– F-ratio as a measure of the overall explanatory power of the model. Has the model explained more than just random variation?

Plots

– Scatterplot matrix, a visual correlation matrix.

Review Questions

What's in the anova table?

– Table shows how much variability is being explained per predictor

```
Source      DF   Sum of Squares   Mean Square    F Ratio
Model        2        7062.5945       3531.30   288.3143
Error      109        1335.0408         12.25     Prob>F
C Total    111        8397.6353                   <.0001
```

– F-ratio measures how much variation is captured or represented by each slope (mean square for model) relative to the variation in each residual (mean square for error, MSE).

– F-ratio tests the size of the R^2 statistic

$$F = \frac{R^2}{1-R^2} \times \frac{DF\ for\ Error}{DF\ for\ Model} \approx \frac{R^2}{1-R^2} \times \frac{\# obs}{\# slopes}$$

Tests the null hypothesis H_0: *all* slopes are zero.

– Degrees of freedom in regression

– Related measures "hidden" in the anova table
- R^2: ratio of (model sum of squares) to the (total sum of squares).
- RMSE: square root of mean square for error.

What good is the F-test?

– Large overall F-ratio, but very "small" individual t-ratios

– The whole model "explains" significant variation, but you cannot attribute this ability to individual predictors.

– Bad date heuristic for the overall F-ratio: fit is significant, but not why.

Is the goal of regression a large F or a large R^2?

Neither! The goal is to understand how the various predictors are related to the response, with an eye toward taking some form of action. Consider the examples in the handout with Class 5 (PDA targeting marketing or allocating advertising expenses).

JMP-IN

Predictions using “Fit Model” tool

– Use the save dialog to compute predictions from a multiple regression model.

– Save the “prediction formula” to compute for other cases.

Key Topics for Today: Elaboration and Loose Ends

Loose ends: plots

– One slope
- use a t-test to measure its contribution to the model
- look for problems using a *leverage plot*. (more later)

– Whole model
- use the overall F-test to see if the model explains significant variation
- look for problems with the overall Y-by-Fit plot and residual plot.

Leverage plot (example follows)

– Analogous to the scatterplot view of a marginal slope, only that a leverage plot shows you a partial slope instead.

– Do you like the simple regression as shown in the leverage plot? If so, you are comfortable with the estimation of the multiple regression (partial) slope. (There may still be problems with interpretation.)

– Useful for
(a) Visually showing the impact of collinearity and
(b) Locating leveraged outliers.

Residual plot (example follows)

– A common overall diagnostic is the plot of the residuals on the fitted values (e.g., page 119)

– This plot is analogous to the plot of residuals on the one predictor in a simple regression. The fitted values are, after all, just a combination of the all of the predictors in the multiple regression.

– Useful for
(a) identifying large outliers in the “Y” direction and
(b) recognizing a lack of constant variance.
(The variance of the residuals often increases with the the predicted values).

Elaboration: Collinearity (a.k.a. multicollinearity.)

– What is collinearity?
Collinearity is correlation among the predictors in a regression.

– What does collinearity do in regression? What are the consequences?
Complicates interpretation, increases SE's, makes it hard to separate the roles of the predictors. Predictors are redundant, but we are asking the regression model to separate them.

– How can I tell if collinearity is present?

• Graphically:	Scatterplots help, leverage plots are better
• Tests:	Big F ratio, small t-ratio
• Diagnostic:	Variance inflation factors (VIF, see below)

– What do I do about collinearity?
- Nothing. Collinearity weakens ability to interpret, but in sample prediction works well (or at least is not injured!).
- Reformulate predictors. Identify distinct concepts.
- Get rid of one of the offenders.

– Summary discussion on page 147.

Variance inflation factor

– A variance inflation factor VIF_j tells you how much larger the SE of a slope has grown because of the presence of collinearity.

$$SE(slope\ \ for\ \ X_j) = \frac{RMSE}{\sqrt{n}} \times \frac{1}{SD(Adjusted\ \ X_j)}$$

$$= \left(\frac{RMSE}{\sqrt{n}} \times \frac{1}{SD(X_j)} \right) \times \sqrt{VIF_j}$$

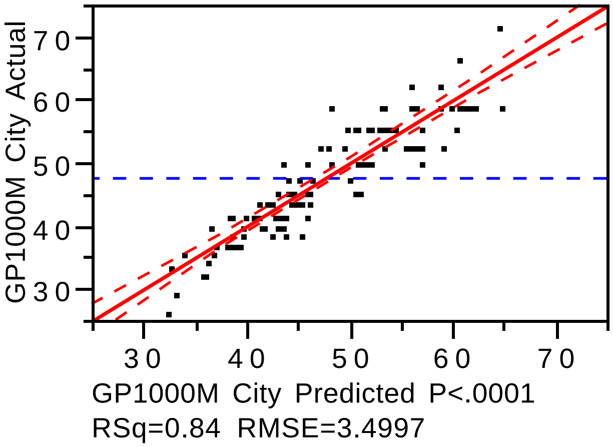
$$= \left(SE\ if\ no\, collinearity \right) \times \sqrt{VIF_j}$$

Examples of Plots in Multiple Regression

Whole model summary

– Closer to diagonal, the better the fit. Related to “calibration”
(i.e., When I predict a large score, is the score in fact large?)

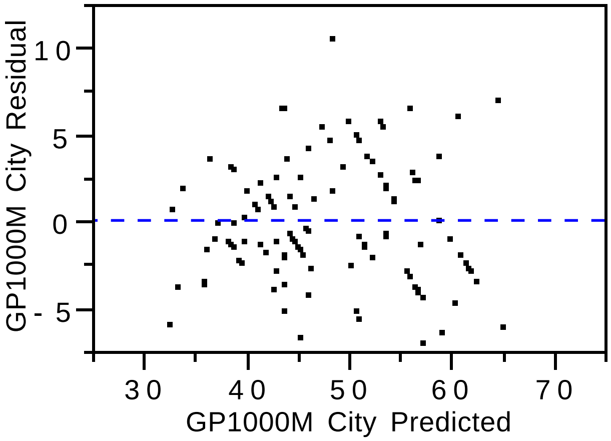
– Visual expression of the R^2 summary (correlation of predicted and actual)



Residual plot

– Plot residuals on at least the predicted values
(predicted values are weighted sum of predictors)

– Looking for outliers (who’s the large positive residual) and for a systematic change in the variance.



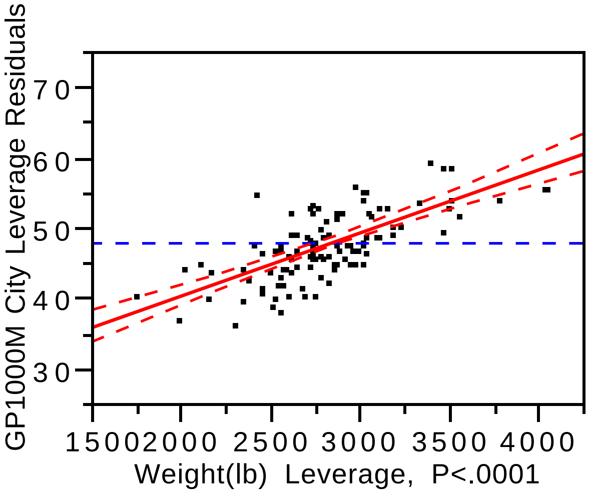
Leverage plots

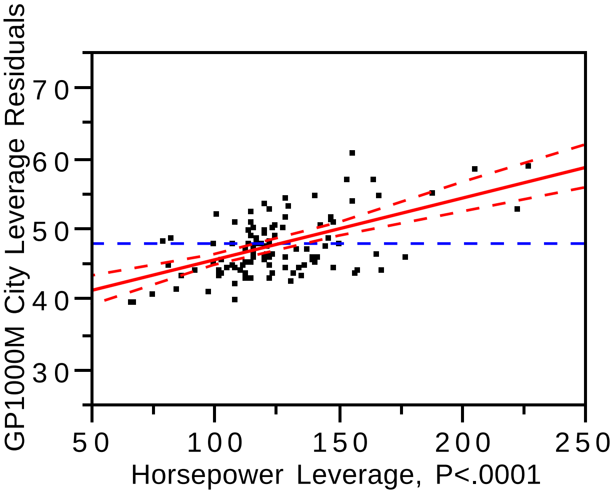
– Shows the partial slope (rather than the marginal slope)

– Useful to identify leverage points *and* see effect of collinearity

– One for each partial slope in the multiple regression

– Question to ask yourself:
Would you be satisfied with the fitted line if this were the plot of a predictor and response showing a marginal slope in simple regression?





New Casebook Examples

Stock prices and market indices **Stocks.jmp, p. 138**

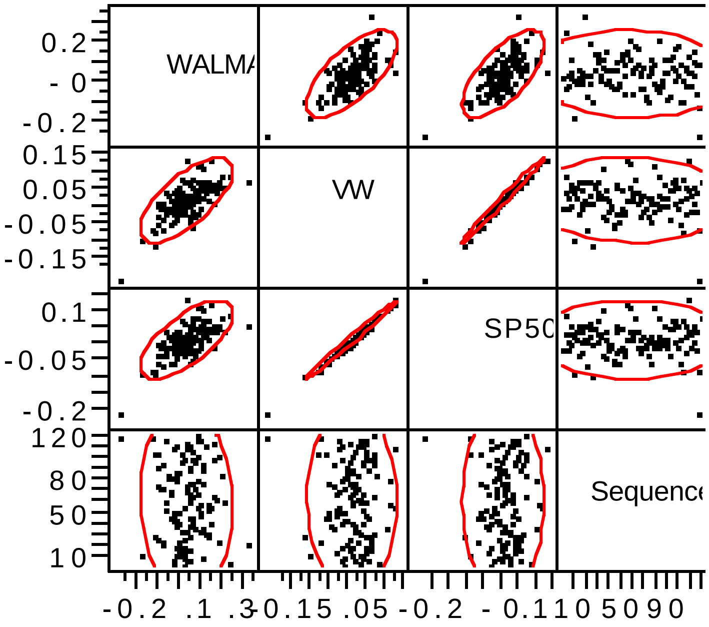
> "How is the return on Walmart related to the returns on two indices?"

– Fitted slope in regression of returns on stock on returns on market estimates the so-called *beta* for the stock, as considered in Assignment #1.
(Note: this example uses real returns, not excess returns. Why is this OK?)

– Huge collinearity (correlation between VW and S&P is 0.993), so almost no unique variation in either predictor given that other is in the model.

• Either taken separately is a good predictor

• Huge collinearity makes it impossible to separate their effects when used together in one model



– Multiple regression of *Walmart* on both indices is fine overall (significant F) but both predictors are "weak" (|t-ratio| around 2 or smaller) (p143)

Summary of Fit

RSquare	0.491
Root Mean Square Error	0.064

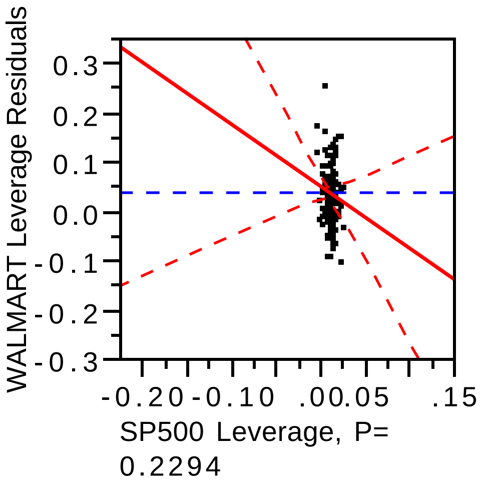
Analysis of Variance

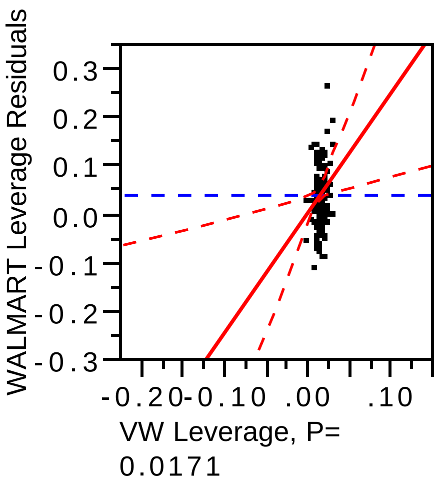
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Ratio**
Model	2	0.455	0.227	56.0082
Error	116	0.471	0.004	**Prob > F**
C. Total	118	0.926		<.0001

Parameter Estimates

Term	**Estimate**	**Std Error**	**t Ratio**	**Prob>|t|**	**VIF**
Intercept	0.02	0.01	2.13	0.0356	.
SP500	-1.26	1.04	-1.21	0.2294	74.3
VW	2.46	1.02	2.42	0.0171	74.3

– "Squished" leverage plots... little unique variation in either predictor available to explain the variation in the response. (p 144)





Conclusion

Remove SP500 index from model and use VW alone. The more inclusive VW index is a better predictor than SP500, as financial theory suggests.

Improved parcel handling Parcel.jmp, p. 148

"How can a shipping company improve performance at sorting centers?"

– Predictors of # sorted packages are:

# sorting lines, # sorters, and # of truckers.

– All predictors and response are transformed to a log scale, so slopes may be interpreted as

$$\text{slope} = \frac{\text{change in log Y}}{\text{change in log X}} = \frac{\text{\% change Y}}{\text{\% change X}}$$

– Multiplicative model, with each slope interpreted as an *elasticity*. For example, a Cobb-Douglas production function is a multiplicative model. (Note: multiplicative implies that zero for any predictor forces the prediction of the model to zero.)

– Differences between marginal and partial slope (elasticities in this example) for the # of sorting lines predictor

• What is the source of these differences?

• Do the marginal and partial elasticities mean the same thing?

marginal elasticity for # lines is 0.70 (p 150)
partial elasticity for # lines is 0.37 (p 151)

– Management issue:

Can you just increase the number of sorting lines (the factor with the largest partial elasticity) without also increasing the number of workers?

– Useful test:

Does adding *both* log(sorters) and log(lines) improve the fit? See the partial F-test example in the casebook.

Review of Take-Away Points

Partial and marginal slopes

– Which to use when.

Collinearity

Correlation among the predictors as seen in the
- scatterplot matrix and *leverage* plots
- changing/unstable coefficient estimates
- inflated standard errors (VIF)
- good overall fit but mediocre t-stats for separate coefficients.

Tests in multiple regression

– t-ratio for the impact of *one* predictor, taken incrementally

– partial F-ratio for the incremental impact of a *subset* of predictors.

– F-ratio for the overall model, the effect of *all* of the predictors. Located in the anova summary

Diagnostics in multiple regression

– Whole model plot and overall residual plot

– Leverage plots for each coefficient in the model

Next Time

Categorical information

– How can one use predictors that define categories of observations in a regression model?