Interpreting Multiple Regression

Preliminaries

Project and assignments

– Hope to have some further information on project soon.

– Due date for Assignment #2.

Review of Key Points

Outliers

– *Leverage*: unusual in terms of the predictor
– *Influential*: the regression changes in an "important" way when the point is removed from the fit.
– Only leveraged points influence the *slope* of the model.
– Choice to retain or exclude an outlier driven by substance of problem.

Multiple regression adds predictors to the equation

0. *Equation* adds more factors

$$ave(Y_i \mid X_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

1. *Independent* observations (independent errors)

2. *Constant variance* observations (equal error variance)

3. *Normally distributed around* the regression line, written as an assumption about the unobserved errors:

$$\varepsilon_i \sim N(0,\sigma^2)$$

Interpreting regression coefficients

– *Marginal* coefficient in one-predictor regression "includes" effects of other correlated factors that are not in the regression model.

– *Partial* coefficient in multiple regression attempts to separate the effects of various predictors (i.e., "holding the other factors fixed" expression)

Prediction intervals

– In-sample: (prediction) +/– 2 RMSE.

– Extrapolation: Error is slope estimation is compounded as model is extrapolated farther out, leading to gradual widening of intervals.

Interpreting Regression Coefficients

Car fuel consumption example

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	9.4323	2.0545	4.59	<.0001
Weight(lb)	0.0136	0.0007	18.94	<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	11.6843	1.7270	6.77	<.0001
Weight(lb)	0.0089	0.0009	10.11	<.0001
Horsepower	0.0884	0.0123	7.21	<.0001

– Offer *several* points of view to help interpret these.

Geometric view

– "Table-top" geometry of points and planes in 3-D coordinates.

– Effect of correlation between two predictors on fitted slopes.

Graph view (a graph with "nodes" and "edges")

– Capture the effects of correlation among predictors

– Direct and indirect effects of predictors upon response

– Simple regression (marginal) slope combines direct and indirect effects.

Back-to-basics view: What does *any* slope mean?

– How does a slope in either model estimate what happens if we change the weight of a car, when in fact we *never* changed the weight of a car?

– Comparison of cars of different weight, not changing the weight.

– So how do you then get a multiple regression slope? (p. 121 for key plot)

Deciding which to use: marginal or partial?

– What question are you trying to answer?
  - Are you comparing two cars that have different weights.
  - Are you comparing two cars with same HP but different weights.

– Which is more appropriate for estimating cost of gas to CA?

Key Application

Separating the factors that influence sales

– Which factor is the most important determinant of business growth?

– How do we evaluate the whole model and the individual components?

Concepts and Terminology

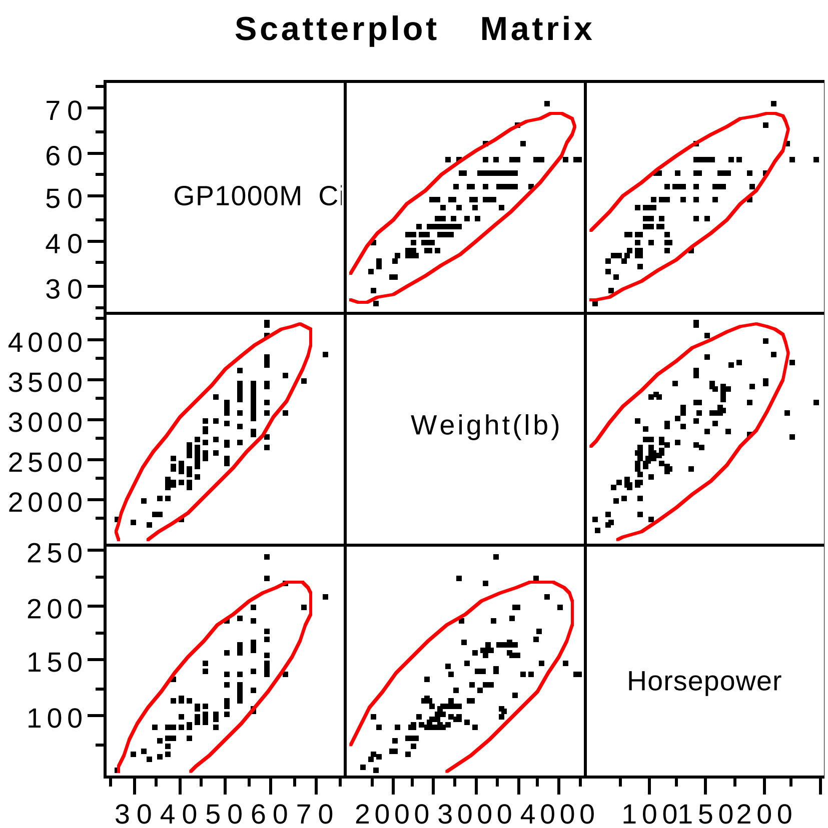
New plot: Scatterplot matrix

– Compact graphical summary of the pairwise associations among a collection of several variables. Visual correlation matrix – an array of plots rather than numbers.

– Put the response in the first row, the other predictors arranged below. This plot reveals collinearity among predictors as well as how each is related to the response.

– Generate this plot using JMP-IN's "Multivariate" command.

– Example (car data) on page 116.



Two types of inference for a multiple regression

– One coefficient t-ratio

“Is this slope different from zero?”

New interpretation in multiple regression, incremental improvement:

“Does this variable significantly improve a model containing rest?”

– All coefficients overall F-ratio

“Does this entire model explain significant amounts of variation?”

Rationale for different procedures

– Each addresses a specific aspect of the fitted model:

t-ratio considers one coefficient (intercept or slope)

F-ratio considers all *slopes*, simultaneously, without allocating.

– Why not just do a bunch of t-tests, one for each slope?

One has to watch out for problems due to the *multiplicity* of testing several conjectures. With 20 predictors, you expect one significant by chance alone! Too many things appear significant that are not meaningful.

New statistical summary: Analysis of variance (ANOVA)

– Summary of how much variation is being explained per predictor:

For car data with *Weight* and *Horsepower* as predictors…

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

– Model-fitting process converts data into regression coefficients.

– F-ratio answers the question “Which explains more variation... one observation or one slope?”

Two “new” plots for each type of test

– t-ratio

Leverage plot for each slope (p. 124, with more next time)

– F-ratio

Plot response on fitted values. The F-ratio tests the overall explanatory power of the model, as reflected in R^2, which happens to be the squared correlation between the fitted values (predicted values for observations, like the values on the simple regression line) and the actual values.

Examples for Today

Handouts

Two examples illustrate what can happen with collinearity in regression. Beginning from a simple regression with a significantly positive marginal slope, the multiple regression has

(a) Partial slope that is near zero, and
(b) Partial slope that is significantly negative.

Automobile design (cntd) Car89.jmp, p. 109

“What other factors are important for the design?”

– *Add* variable for Horsepower (p 117)
- Addition of HP is significant improvement since its t-ratio=7.21
- Cost for carrying additional 200 lbs. for 3000 miles ≈ 5.3 gals
- R^2 increases from 77% to 84% and RMSE drops to 3.50

– How small can we make the RMSE by adding other predictors?

– Add Cargo, Seating, Price (p 128)
- Leverage plots reveal other outliers that exert effects on fit. (p 129)
- Coefficients for “dubious” factors (e.g. Price) are sensitive to subsets

– How do we avoid “false positives” by searching over many predictors?
- Expanded data includes some “special” extra predictors
- R^2 always goes up, even if predictor does not really help.
- “Bonferroni method”: compare p-value to 0.05/(number considered)

– Revisit the problem of how to go about expanding (or in general, building) a regression model in lectures 9 and 10.

Key Take-Away Points

Multiple regression

– Partial vs. marginal slope: Which is the right one to use?

– Testing some/all of the coefficients using t-test or F-ratio.

– Adding predictors to a regression to improve its fit
  • Useful predictors lead to better predictions
  • Poor predictors claim to improve, but only add "noise"
    (Dangers of "data dreging" and Bonferroni method)

Collinearity

– Predictors are related, making it hard to separate effects:
Difference between marginal and partial coefficients

– Arises from correlation among the predictors

– Impact can be extreme
  - Change of sign in marketing handout example
  - Collapse of model's interpretation in advertising example.

Next Time

Collinearity and diagnostics for multiple regression

– How does one quantify the effects of collinearity?

– How does one check the residuals in a multiple regression, and learn how to add other factors?