Multiple Regression

Project Analysis for Today

First steps

Transforming the data into a form that lets you estimate the fixed and variable costs of a lease using a regression model that meets the three key assumptions.

Review of Multiple Regression from Last Week

Objective

Isolate the key factors that influence the response and separate their effects.

Model

"Y" $= \beta_0 + \beta_1$ "X_1" + ... + β_k "X_k" + Error

Sales $= \beta_0 + \beta_1$ Adv$ + β_2 Price + Error

with

- Independence
- Constant variance σ^2 about regression line
- Normally distributed errors about the regression line.

Discussion

– Model is additive
– Geometry of multiple regression
– Slopes measure effect of each predictor "holding others fixed"
"Simple" regression slope vs multiple regression slope

Relationship between R^2 and RMSE

– Both describe "goodness-of-fit"
– R^2 is relative whereas RMSE is absolute.
– They are related as follows:

$$RMSE^2 = \text{Var}(\textit{residuals}) \approx (1 - R^2)\ \text{Var}(\textit{response})$$

– Same interpretation in simple (one predictor) and multiple regression.

Inference in Multiple Regression

Inference in multiple regression

– One coefficient t-ratio (estimate/SE)
"Is this slope different from zero?"
"Does this variable significantly improve a model containing rest?"
– All coefficients overall F-ratio (anova table)
"Does this entire model explain significant amounts of variation?"

Analysis of variance (ANOVA) summary (page 141)

– Summary of how much variation is being explained *per* predictor.
– Example for the 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

Why do we need different tests?

– Each addresses a specific aspect of the fitted model:
t-ratio considers one coefficient (intercept or slope)
F-ratio considers all *slopes*, simultaneously
– Why not just do a bunch of t-tests, one for each slope?
With 20 predictors and 95% CI, you can expect one significant (not zero) by chance alone! Too many things will appear significant that really are not meaningful.
– Recall the use of multiple comparisons in anova.

Collinearity in Multiple Regression

What is collinearity? (Also known as multicollinearity.)
- Collinearity is correlation among the predictors in a regression.
- As such, collinearity does not "violate an assumption" in regression.

What does collinearity do in regression? Consequences?
- Complicates interpretation, making it hard to separate the predictors.
- Inflates the SE's of the estimated coefficients.

$$\text{SE(slope estimate for } X_j) \approx \frac{\sigma}{\sqrt{n}} \frac{1}{\text{SD(Adjusted } X_j)}$$
$$= \frac{\sigma}{\sqrt{n}} \frac{\sqrt{\text{VIF}_j}}{\text{SD}(X_j)}$$
$$= \sqrt{\text{VIF}_j} * \text{(SE if no collinearity)}$$

How can I tell if collinearity is present?
- Graphically: Scatterplots help, but *leverage plots* are better.
  - Multiple "simple regression" views of one multiple regression.
  - Essential for identifying leverage points in multiple regression.
  -"Do I like the shown simple regression model?"
- Tests: Big F ratio, small t-ratio
- Diagnostic: Variance inflation factors (VIF)

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. Stats help you decide which one.
- Summary discussion on page **147** of the casebook.

Example of Multiple Regression

Automobile design Car89.jmp, page 109

"What is the predicted mileage for a 4000 lb. design, and what characteristics of the design are crucial?"

"How much does my 200 pound brother owe me for gas for carrying him 3,000 miles to California?" (Oops, it's urban mileage in example)

– Initial one-predictor model
- Transform response to gallons per 1000 mile scale.
- Cannot compare R^2's since two model use different dependent variables (MPG and GPM)
- Effect of scaling from GPM to GP1000M.
- RMSE = 4.23 (p 111)
- Skewness in residuals from regression with *Weight*. (p 112)
- Prediction @ 4000 lbs = 63.9, ⇑ 200 lbs for 3000 miles ≈ 8.2 gals

– Add variable for *Horsepower* (p 117)
- R^2 increases from 77% to 84% (added variable is significant, t=7.21)
- RMSE drops to 3.50
- Predictors are related, both increase together, higher SE for *Weight*.
- Picture explains the increase in SE due to restricted range (p 120).
- ⇑ 200 lbs for 3000 miles ≈ 5.3 gals
- Prediction from multiple regression

– Add a predictor less correlated with Weight, use *HP/Pound* (p 123)
- *Weight* and *HP/Pound* less related, more distinct properties of these cars.
- Engineer can manipulate these separately, unlike HP and weight.

Residual plots
– Show residuals plotted on fitted values
– Inspect for deviations from assumptions (such as lack of constant variance)

Leverage plots (p 125)
– Diagnostic plot, designed especially for multiple regression
– Reveals leveraged observations in *multiple* regression.

Next steps for this model…
– What other factors are important for the design?
– How small can we make the RMSE?

Example with Extreme Collinearity in Multiple Regression

Stock prices and market indices Stocks.jmp, page 138

"What's the beta for Walmart when regressed on *two* indices?"

– Fitted slope of stock returns on market estimate the beta for the stock.
– Huge collinearity (correlation between VW and S&P is 0.993), so almost no unique variation in either one given that other is in model.
– Either taken separately is a good predictor, but show weak effects when used together.
– "Squished" leverage plots... little unique variation in either predictor available to explain the variation in the response. (p 144)

– More complete VW index is better predictor, as financial theory suggests.

Next Time

Categorical predictors...

Categorical predictors allow us to compare regression models for different groups, judging if the models for the different groups are comparable.

Response: GP1000M City

Summary of Fit

RSquare	0.765
RSquare Adj	0.763
Root Mean Square Error	4.233
Mean of Response	47.595
Observations (or Sum Wgts)	112.000

Lack of Fit

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

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	6426.44	6426.44	358.6195
Error	110	1971.19	17.92	Prob>F
C Total	111	8397.64		<.0001

Response: GP1000M City

Summary of Fit

RSquare	0.841
RSquare Adj	0.838
Root Mean Square Error	3.500
Mean of Response	47.595
Observations (or Sum Wgts)	112.000

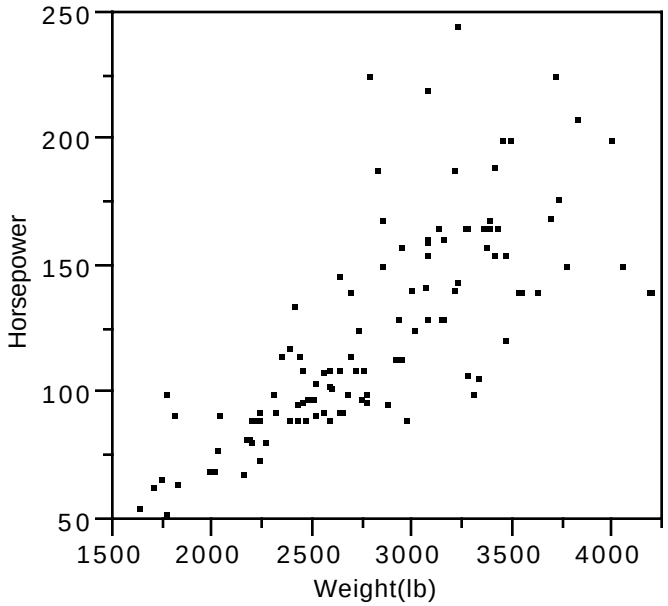
Lack of Fit

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

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	7062.59	3531.30	288.3143
Error	109	1335.04	12.25	Prob>F
C Total	111	8397.64		<.0001

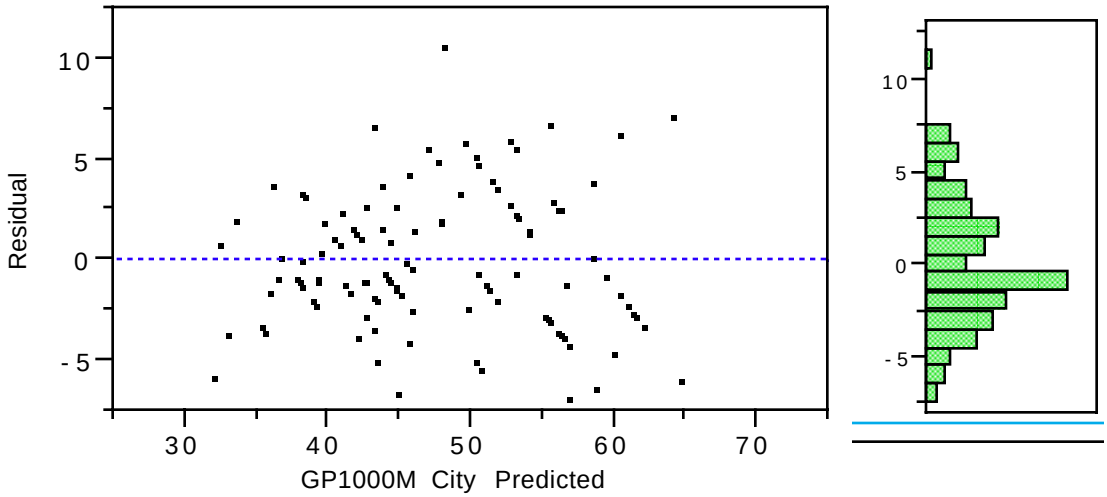


$$\text{SE(slope estimate for } X_j) \approx \frac{\sigma}{\sqrt{n}} \frac{1}{\text{SD(Adjusted } X_j)}$$

$$= \frac{\sigma}{\sqrt{n}} \frac{\sqrt{\text{VIF}_j}}{\text{SD}(X_j)}$$

$$= \sqrt{\text{VIF}_j} * \text{(SE if no collinearity)}$$

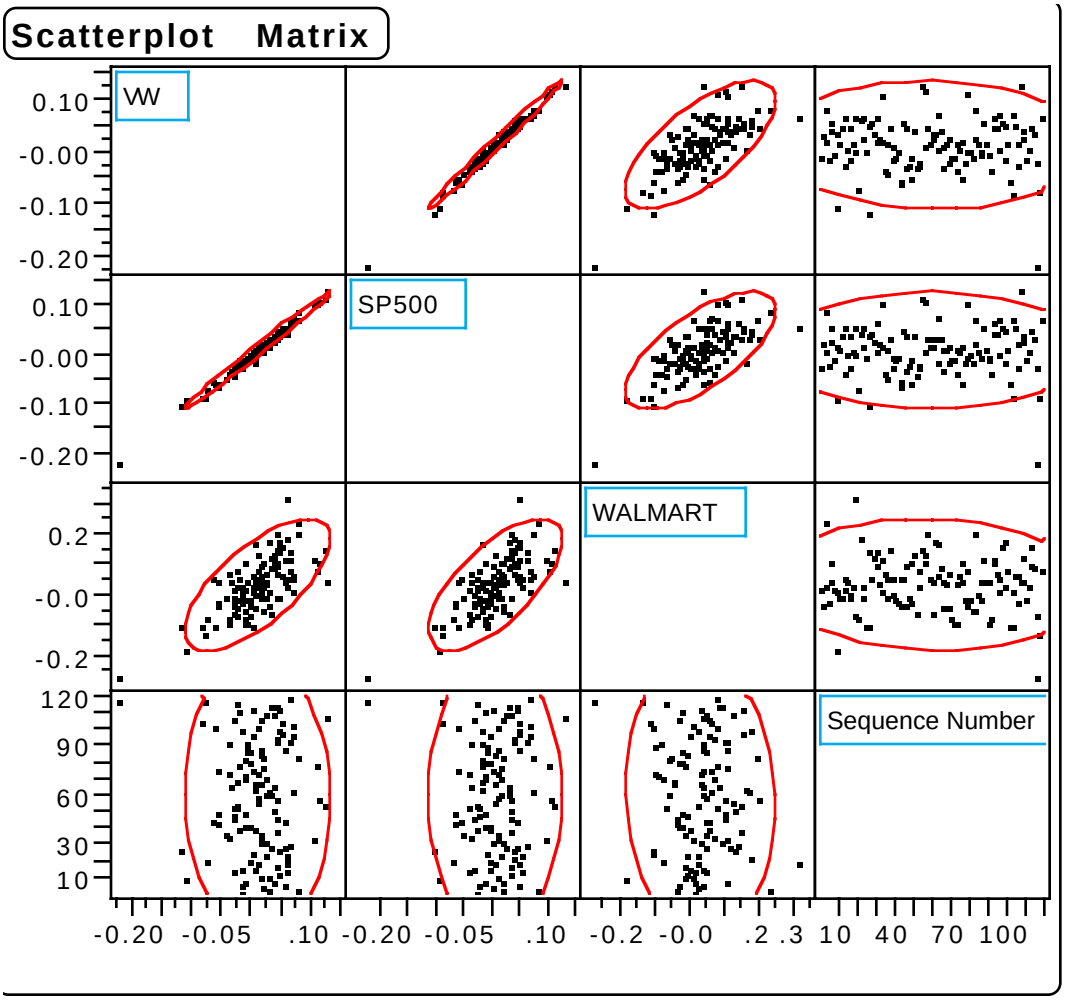
Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	11.6843	1.72704	6.77	<.0001	0.000
Weight(lb)	0.0089	0.00088	10.11	<.0001	2.202
Horsepower	0.0884	0.01226	7.21	<.0001	2.202



Correlations

Variable	VW	SP500	WALMART	Sequence Number
VW	1.000	0.993	0.696	-0.036
SP500	0.993	1.000	0.682	0.002
WALMART	0.696	0.682	1.000	-0.055
Sequence Number	-0.036	0.002	-0.055	1.000

Scatterplot Matrix

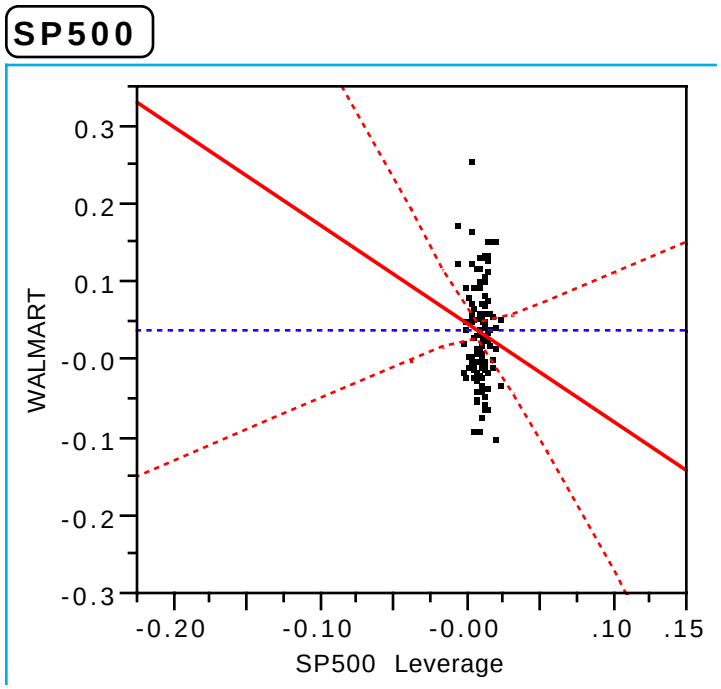


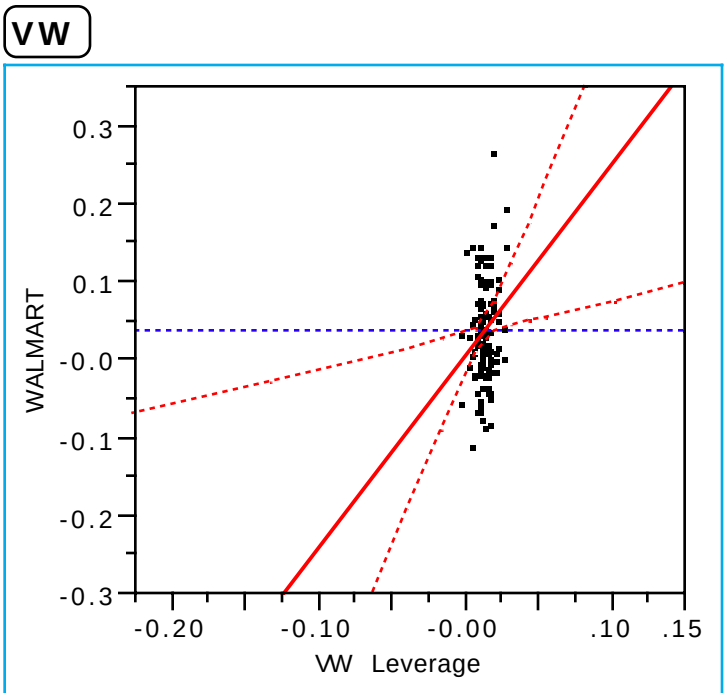
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	0.024	0.006	4.02	0.0001	0
SP500	1.244	0.123	10.10	<.0001	1

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	0.015	0.007	2.13	0.0356	0.000
SP500	-1.258	1.041	-1.21	0.2294	74.297
VW	2.458	1.016	2.42	0.0171	74.297


SP500
WALMART
0.3
0.2
0.1
-0.0
-0.1
-0.2
-0.3
-0.20
-0.10
-0.00
.10
.15
SP500 Leverage


VW
WALMART
0.3
0.2
0.1
-0.0
-0.1
-0.2
-0.3
-0.20
-0.10
-0.00
.10
.15
VW Leverage