Multiple Regression

Preliminaries

Grading on this and other assignments

– Assignment will get placed in folder of first member of Learning Team.
– Will post solution set this week.
– Please see me if you have "non-routine" questions about grading.

Feedback

– To me: Class feedback form, e-mail, or cohort academic reps.
– To you: Continuous assessment questions on class web page
Practice questions on my web page.

Review of Regression with One Predictor

Regression model

0. *Equation* (Note that "X" and "Y" include needed transformation)

$$ave("Y_i"|"X_i") = \beta_0 + \beta_1 "X_i"$$

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:

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

Three unknown parameters identify the model:
slope, intercept and SD of the errors

Checking the assumptions (order IS important)

Linearity	Scatterplots(data/residuals)
Independence	Plot residuals sequentially, highlighting trend
Constant variance	Plot residuals on predictor values
Normality	Quantile plot of residuals (outliers, skewness)

Confidence intervals vs. "prediction intervals"

– To what degree does advertising affect sales? Interpreting when holds zero
Fitted slope ± 2 SE(fitted slope)
– What can I conclude about next month's sales? (an individual value)
Predicted value ± 2 RMSE (a.k.a. prediction interval)

for *in-sample* predictions. Extrapolation beyond experience requires more care. (See cottage profits example below)

Review Questions

How does the precision of the estimated slope (standard error) change when an removing an outlier from a regression?

It depends on where the outlier is located. If the point is a leveraged outlier (at the extremes of the predictor as with Center City in the Philadelphia housing data), then often the SE of the slope will increase. The explanation lies in the expression for the SE of a slope estimate: the more spread out the predictor, the smaller the SE of the slope.

What does R^2 tell me about a regression model?

– R^2 gives the proportion of "explained" variation, i.e., the proportion of the differences among the response values that can be interpreted in terms of differences among the corresponding values of the predictor.

– "Venn diagram" view of the regression process.

– Squared correlation between the predictor and the response.

How does R^2 compare to RMSE? Why have both?

– Both measure the "goodness-of-fit" of the model.

– RMSE is the *estimated* SD of the errors. If the RMSE is small relative to the variation in the data, then the R^2 is close to 1 and the data are concentrated near the fitted model.

$$RMSE^2 \approx (1 - R^2)\, Var(Y)$$

– RMSE has the units of the response whereas R2 is an index measuring the proportion of variation "explained" by the model.

– Different views of how well the predictor fits/explains the response.

– R^2 is a relative measure of what the model has accomplished.

– RMSE is an absolute measure of predictive accuracy of the model.

When can you compare the R^2's of different models?

– Since R^2 is relative (it's the proportion of the variation), you must have the same response.

What does it mean to say that a slope is "not significant"?

An estimated slope is "statistically significant" when any/all of the following occur:

- Zero is *not* in the confidence interval for the slope.
- The t-ratio is *larger* than two in absolute size.
- The p-value is *smaller* than 0.05.

When the slope *is* significant, we know

(a) The predictor has a non-random effect on the response
(i.e., unlikely this effect produced by sampling variation).

(b) The "direction" of the effect of this predictor on the response
(i.e., positive or negative effect).

When zero lies in the CI, the "true slope" may be zero – no effect. In any case, we cannot tell how if at all this predictor affects the response. For example, if the slope of sales on advertising is not significant, then you cannot expect a gain in sales when you increase advertising by $100,000, say.

What should you do about outliers?

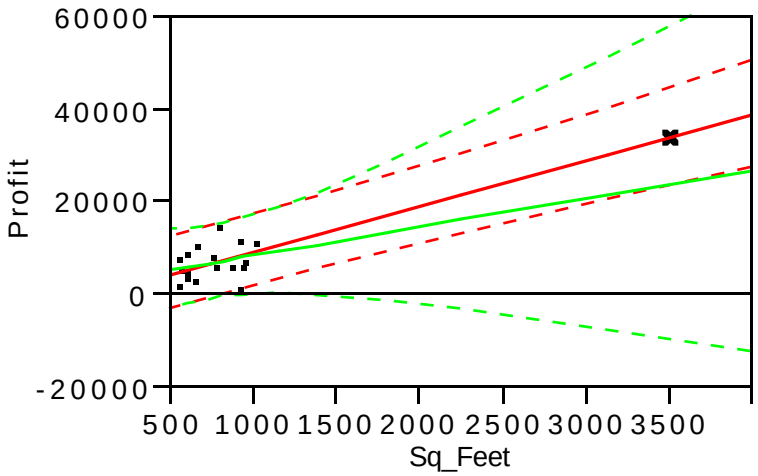
See the next example!

Review Example: Outliers and Confidence Intervals

Housing construction **Cottages.jmp, page 89**

"How much can a builder expect to profit from building a large home with 3,500 square feet?"

– Builder has previously constructed one large home, making this observation highly leveraged (extreme in terms of the predictor)



– With prior large house, $R^2 = 0.78$, RMSE ≈ 3500, and large constructions appear quite profitable, about $10 per square foot.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-416.9	1437.0	-0.29	0.7755
Sq_Feet	9.8	1.3	7.52	<.0001

– Without prior large house, $R^2 = 0.08$, RMSE ≈ 3600, and we cannot tell if profit grows with construction size (slope could be zero or negative).

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2245.4	4237.2	0.53	0.6039
Sq_Feet	6.1	5.6	1.10	0.2868

– Should we keep the large cottage, or do we exclude the large cottage?

Outliers in Regression

Terminology

– *Leveraged* observations = extremes of the predictor (e.g. CC Phila).

– *Influential* observations = fit changes when the point is removed. Use this term to refer to observations whose presence affects the slope or other feature (like R^2 or RMSE) of the fitted model.

Three canonical situations (see casebook, page 46)

– Not leveraged, large residual (direct mail example)

– Leveraged, far from fitted line (Philadelphia crime rate example)

– Leveraged, close to fitted line (construction profits in cottage example)

Outliers and assumptions

Outliers can suggest many ways in which your data do not conform to the assumed form of the fitted model:

- Wrong equation (in leveraged case)
- Variance is quite different under some conditions
- Errors are simply not normally distributed (think back to 603).

Impact of outlying values

– Fit can "improve" or "weaken" when the outlier is removed.
  - Fit got "better" without outlier in Philadelphia crime example.
  - Fit got "worse" in the housing construction example.

– Keep or set aside? Decision *must* be based on substantive grounds.

– Outliers often show where your model is deficient.
  (See discussion of Philadelphia housing example, p 68-70).

New Application for Today

Separating the factors that drive sales

– Which factor is the most important determinant of business growth? Advertising? Product loyalty? Price?

– Complicated because of the relationships among the predictors.

Concepts and Terminology

Multiple regression adds predictors to the equation

0. *Equation* adds more factors

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

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:

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

Interpretation: Implications of model equation

– Interpretation of slopes as effect of each predictor "holding other predictors fixed" (as in a partial derivative)

– Same slope for each predictor X_j regardless of values of other factors

– Factors add together (e.g., why not multiply as in production function?)

Interpretation: Marginal slope versus partial slope

– Marginal: "simple" regression slope

– Partial: multiple regression slope, adjusted for levels of other factors.

– Complications arise from collinearity, the common (almost universal) presence of correlation among the predictors in the model.

– Draw the "model graph" with variables as "nodes"

Inference: Determinants of SE for a slope

– Question about one predictor:
"Does this predictor improve a model containing *all* of the others?"

– Answer: Use the same procedure as with one predictor,
use the t-ratio or CI for the slope

– Relationships/correlation among predictors increase the SE of slopes.

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

Inference: Goodness-of-fit and R^2

– Question about the overall model:
"Does the model (taken collectively) explain significant variation?"

– Answer: Requires a procedure that was not needed with one predictor:
use the overall F-ratio (related to RMSE and R^2)

– R^2 = Proportion of variation in response captured by the fitted model.

$$R^2 = \frac{Explained\ Sums\ of\ Squares}{Total\ Sums\ of\ Squares}$$

– R^2 = Squared correlation of Y and predicted values from fitted model.

– Judge changes in R^2 by looking at what is left over...
Easy... 0.50 –> 0.51 Hard... 0.98 –> 0.99

Diagnostic: Leverage plots

– Graphical diagnostic for multiple regression.
"Do I like the shown simple regression model?"

– Reduces multiple regression to sequence of simple regressions.

JMP-IN Methods

Fit Model command

– Powerful generalization of the "Fit Y by X" command

– Allows you to save features of the model (predictions, residuals)

– Ability to save prediction formula is very useful
Add a dummy row(s) so that you can fill in values for the predictors and get the model prediction under these conditions.

Example for Today

Automobile design **Car89.jmp, page 109**

"What is the predicted mileage for a 4000 lb. design, and which 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)

– *Transform* response to gallons per 1000 mile scale

– *Simple regression* using only the weight of the car gives
- R^2 = 77%, RMSE = 4.23 (p 111)
- Prediction @ 4000 lbs. = 63.9,
- Cost for 200 lbs. for 3000 miles ≈ 8.2 gals

– Skewness in residuals from regression with Weight. (p 112)

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

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

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

– *Related predictors*: both typically increase together
- Implication is higher SE for Weight slope that in prior fit.
- Picture explains the increase in SE due to restricted range (p 120).

– Which predictor is "more important"?
Statistical significance (which offers more incremental improvement)
Substantive role of the coefficient in the model, $ impact on model.

– *Leverage plots* (p 125) show outlying, leveraged cars

– *Next steps*

What other factors are important for the design?

- How small can we make the RMSE?
- How do we avoid "false positives" by searching many predictors?

Key Take-Away Points

Role of outliers

– Leverage and influence terminology

– Outliers can nominally "improve" or "weaken" a fitted model.

– Require substantive insight to choose course of action (keep, delete)

Multiple regression

– Model simultaneously combines predictors

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

– New role for t-ratio as measure of incremental value.

"Collinearity"

– Predictors are correlated, making it hard to separate effects.

– Partial effect (multiple) not the same as marginal (simple)

Graphical diagnostics

– Plot of residuals on fitted (or predicted) values

– Leverage plots

Next Time

More multiple regression and more collinearity

– Growing the model with the addition of more predictors

– Emphasizing impact of collinearity on fitted model