Prediction and Confidence Intervals in Regression

Preliminaries

Teaching assistants

- See them in Room 3009 SH-DH.
- Hours are detailed in the syllabus.

Assignment #1 due Friday

– Substantial penalty if not turned in until Monday.

Feedback to me

- In-class feedback form
- e-mail from web page
- Cohort academic reps, quality circle

Feedback questions on class web page

Review of Key Points

Regression model adds assumptions to the equation

0. Equation, stated either in terms of the average of the response $ave(Y_i \mid X_i) = b_0 + b_1 X_i$

or by adding an error term to the average, obtaining for each of the individual values of the response

$$Y_i = ave(Y_i \mid X_i) + \varepsilon_i$$
$$= \beta_0 + \beta_1 X_i + \varepsilon_i$$

- 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:

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

- Least squares estimates of intercept, slope and errors

– Importance of the assumptions evident in the context of prediction.

Checking the assumptions (order IS important)

Linearity	Scatterplots(data/residuals) (nonlinearity)	
Independence	Plots highlighting trend	(autocorrelation)
Constant variance	Residual plots	(heteroscedasticity)
Normality	Quantile plots	(outliers, skewness) (summary on p. 46)

One outlier can exert much influence on a regression

- JMP-IN script illustrating effect of outlier on the fitted least squares regression line (available from my web page for the class).
- Another version of this script is packaged with the JMP software and installed on your machine (as part of the standard installation).

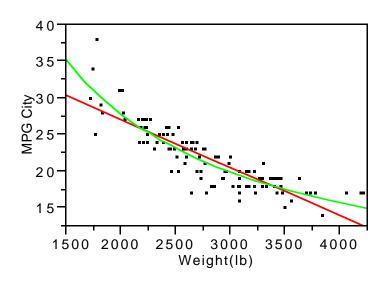
Review Questions from Feedback

Can I use the RMSE to compare models?

– Yes, but ...

You need to ensure that the comparison is "honest". As long as you have the same response in both models, the comparison of two RMSE values makes sense. If you change the response, such as through a log transformation of the response as in the second question of Assignment #1, you need to be careful.

- The following figure compares a linear fit to a log-log fit for the MPG versus the weight:



If you compare the nominal RMSEs shown for the two models, you will get the following

RMSE linear	2.2
RMSE log-log	0.085

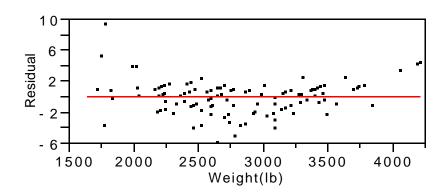
While the linear fit looks inferior, it's not that much worse!

- -The problem is that we have ignored the units; these two reported RMSE values are on different scales. The linear RMSE is 2.2 MPG, but the log-log RMSE is 0.085 log MPG, a very different scale.
- JMP-IN helps out if you ask it. The following summary describes the accuracy of the fitted log-log model, but in the original units. That summary gives an RMSE of 1.9 MPG, close to the linear value.

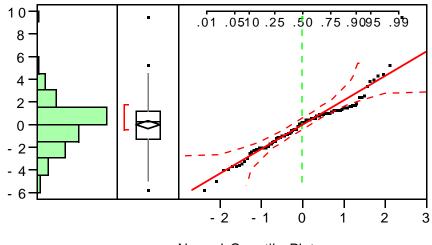
Fit Measured on Original	Scale
Sum of Squared Error	394.518
Root Mean Square Error	1.894
RSquare	0.803
Sum of Residuals	9.242

What does RMSE tell you?

- The RMSE gives the SD of the residuals. The RMSE thus estimates the concentration of the data around the fitted equation. Here's the plot of the residuals from the linear equation.



- Visual inspection of the normal quantile plot of the residuals suggests the RMSE is around 2-3. If the data are roughly normal, then most of the residuals lie within about ± 2 RMSE of their mean (at zero):



Normal Quantile Plot

Key Applications

Providing a margin for error...

For a feature of the "population"

- How does the spending on advertising affect the level of sales?

- Form *confidence* intervals as (estimate) +/- 2 SE(estimate)

For the prediction of another observation

- Will sales exceed \$10,000,000 next month?
- Form *prediction* intervals as (prediction) +/- 2 RMSE

Concepts and Terminology

Standard error

- Same concept as in Stat 603/604, but now in regression models
 Want a standard error for both the estimated *slope* and intercept.
- Emphasize slope since it measures expected "impact" of changes in the predictor upon values of the response.
- Similar to expression for SE(sample average), but with one important adjustment (page 91)

$$SE(estimated \ slope) = SE(\hat{\beta})$$

$$= \frac{SD(error)}{\sqrt{sample \ size}} \times \frac{1}{SD(predictor)}$$
$$= \frac{\sigma}{\sqrt{n}} \times \frac{1}{SD(X)}$$

Confidence intervals, t-ratios for parameters

 As in Stat 603/604, the empirical rule + CLT give intervals of the form Fitted intercept +/- 2 SE(Fitted intercept)
 Fitted slope +/- 2 SE(Fitted slope)

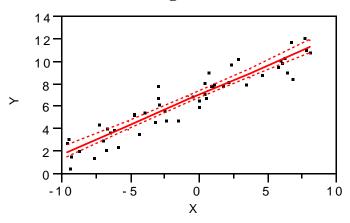
- t-ratio answers the question

"How many SEs is the estimate away from zero?"

In-sample prediction versus extrapolation

- In-sample prediction... able to check model properties.
- Out-of-sample prediction... hope form of model continues.
- Statistical *extrapolation penalty* assumes model form continues.
 See the example of the display space transformation (p. 101-103).
- Another approach to extrapolation penalty is to compare models
 e.g., Hurricane path uncertainty formed by comparing the predictions of different models.

Confidence intervals for the regression line

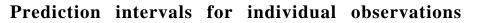


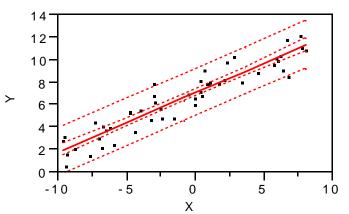
- Answers "Where do I think the population regression line lies?"
- Fitted line ± 2 SE(Fitted line)
- Regression line is *average* value of response for chosen values of X.
- "Statistical extrapolation penalty"

CI for regression line grows wider as get farther away from the mean of the predictor.

- Is this penalty reasonable or "optimistic" (i.e., too narrow)?

JMP-IN: "Confid Curves Fit" option from the fit pop-up.





- Answers "Where do I think a single new observation will fall?"

- Interval captures *single* new random observation rather than *average*.
- Must accommodate random variation about the fitted model.
- Holds about 95% of data surrounding the fitted regression line.
- Approximate *in sample* form: Fitted line ± 2 RMSE
- Typically more useful that CIs for the regression line:
 More often are trying to predict a new observation, than wondering where the average of a collection of future values lies.

JMP-IN: "Confid Curves Indiv" option from the fit pop-up.

Goodness-of-fit and R²

- Proportion of variation in response captured by the fitted model and is thus a relative measure of the goodness of fit – RMSE is an absolute measure of the accuracy in the units of "Y".
- Computed from the variability in response and residuals (p 92)
 ... Ratio of sums of squared deviations.
- Related to RMSE, RMSE² approximately equals $(1 R^2)$ Var(Y)
- Square of the correlation between the predictor X and response Y. $corr(X,Y)^2 = R^2$

Another way to answer "What's a big correlation?"

Examples for Today

The ideal regression model Utopia.jmp, page 47

"Which features of a regression change when the sample size grows?"

- Simulate data from a normal population.

- Add more observations. Use formula to generate more values.
- Some features stay "about the same": intercept, slope, RMSE, R²
 (Note: Each of these estimates a population feature.)
- Standard errors shrink and confidence intervals become more narrow.

"How do crime rates impact the average selling price of houses?"

- Initial plot shows that Center City is "leveraged" (unusual in X).
- Initial analysis with all data finds \$577 impact per crime (p 64).
- Residuals show lack of normality (p 65).
- Without CC, regression has much steeper decay, \$2289/crime (p 66).
- Residuals remain non-normal (p 67).
- Why is CC an outlier? What do we learn from this point?
- Alternative analysis with transformation suggests may be not so unusual. (see pages 68-70)

Housing construction

Cottages.jmp, page 89

"How much can a builder expect to profit from building larger homes?"

– Highly leveraged observation ("special cottage") (p 89)

- Contrast confidence intervals with prediction intervals.

- role of assumptions of constant variance and normality.
- Model with "special cottage"
 - $R^2 \approx 0.8$, RMSE $\approx 3500 (p 90)$
 - Predictions suggest profitable
- Model without "special cottage"
 - $R^2 \approx 0.08$, RMSE $\approx 3500 (p94-95)$
 - Predictions are useless
- Recall demonstration with JMP-IN "rubber-band" regression line.

– Do we keep the outlier, or do we exclude the outlier?

Liquor sales and display space Display.jmp, page 99

"How precise is our estimate of the number of display feet?"

"Can this model be used to predict sales for a promotion with 20 feet?"

- Lack of sensitivity of optimal display footage to transformation
 Log gives 2.77 feet (p 19-20), whereas reciprocal gives 2.57 feet (p 22)
- 95% confidence interval for the optimal footage (p 100).95% CI for optimal under log model is [2.38, 3.17]
- Predictions out to 20 feet are very sensitive to transformation Prediction interval at 20 feet is far from range of data.
 Very sensitive: Log interval does not include reciprocal pred (p111)
 Management has to decide which model to use.
- Have we captured the "true" uncertainty?

Key Take-Away Points

Standard error again used in inference

- Confidence interval for the slope is (fitted slope) ± 2 SE(slope)
- As the sample size grows, the confidence interval shrinks.
- If 0 is outside the interval, the predictor is "significant".
- Formula for standard error is different from that for an average
 - depends on the variation in the predictor
 - more variation in predictor, more precise slope

R² as a measure of goodness of fit

- Square of usual correlation between predictor and response
- Popular as "percentage of explained variation"

Prediction intervals used to measure accuracy of predictions

- Formed as predicted value (from equation) with margin for error set as plus or minus twice the SD of the residuals,
 - (predicted value) +/- 2 RMSE (in-sample only!)
- RMSE provides baseline measure of predictive accuracy

In-sample prediction vs. out-of-sample extrapolation

- Models are more reliable in-sample.
- Statistical "extrapolation penalty" is often too small because it assumes the model equation extends beyond the range of observation.

Role of outliers

- Single leveraged observation can "skew" fit to rest of data.
- Outliers may be very informative; not a simple choice to delete them.

Next Time

Multiple regression

Combining the information from several predictors to model more variation, the only way to more accurate predictions.