Prediction and Outliers in Regression

Administrative Items

Get help!

– See me Monday 3-5:30, Wednesday from 4-5:30, or make an appointment.
– Send an e-mail to stine@wharton.
– Visit the StatLab/TAs, *particularly for help using the computer*.

Review of Regression

Questions

Can you use this measurement to predict the response?
How accurately can you predict the response?
How do the various observations influence this prediction?

Utopian model for regression

If we let Y denote the response and X the predictor, then

$$\begin{aligned} \text{Ave}(Y \mid X) &= \text{Intercept} + \text{Slope}\ (X) \\ &= \beta_0 + \beta_1\ (X) \end{aligned}$$

where we assume that the underlying observations are

(a) Independent
(b) Have constant variance
(c) Are normally distributed around the "true" regression line

$$Y_i = \beta_0 + \beta_1\ X_i + \varepsilon_i\ , \qquad \varepsilon_i \sim N(0,\sigma^2)$$

Estimation

Choose the line that minimizes the sum of the squared *residuals*, the vertical deviations or fitting errors that separate the observed data points from the line.

Confidence intervals and tests

The standard error of the slope is

$$\text{SE(slope estimate)} \approx \frac{\sigma}{\sqrt{n}}\,\frac{1}{\text{SD}(X)}$$

It produces confidence intervals of the usual form

$$\text{(estimated slope)} \pm 2\ \text{(SEs of estimated slope)}$$

and leads to tests of hypotheses, such as whether the slope is zero, by counting the number of standard errors that separate the fitted slope from zero (t-ratio).

Regression and Prediction

Accuracy of prediction

Determined by the variability of points around the fitted regression line. In the utopian model, the variance of the errors is σ^2 (or the mean squared error).

Prediction and R^2

R^2 is the square of the usual correlation between the predictor X and the response Y, so $0 \le R^2 \le 1$. In regression it may also be computed as the ratio

$$R^2 = \frac{\text{Variation captured by fitted model}}{\text{Variation in Response}}$$

so that 100 R^2 is interpreted as the percentage of variation in the response which has been *explained* by the fitted model. For a given set of data, the larger the value of R^2, the smaller the MSE and thus the more accurate the prediction. Roughly,

$$MSE = (1- R^2)\ Var(Y)$$

Prediction interval

Once you have an estimate of $MSE = \sigma^2$, under the assumption of normality, roughly 95% of the observations are within ± 2 $(\sqrt{MSE} = RMSE)$ of the fitted line.

Extrapolation penalty

The previous interval is only accurate for predictions in the range of the observed data. Extrapolation beyond that range is not so accurate as this expression would suggest.

Importance of the normality assumption

In most problems, the Central Limit Theorem means that the estimator (like the sample average) is close to normally distributed, so confidence intervals are accurate even if the data are not normal. For prediction intervals, however, the assumption of normality is crucial.

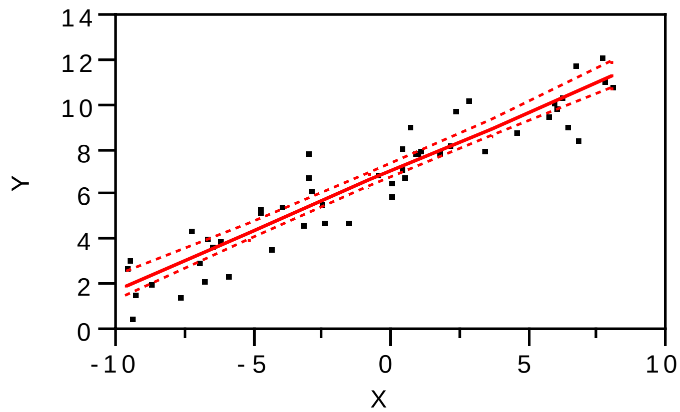
Outliers

Leverage and influence

Single values can have substantial effect on a fitted model.

Observations with unusual values of the predictor are said to be *leveraged*. Removing *influential* observations lead to changes in the fitted model.

Two Intervals for the Regression



Confidence intervals for the regression line

"Where do I think the population regression line lies?"

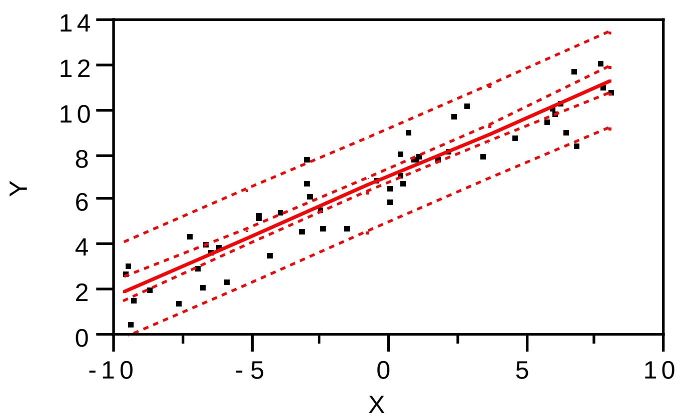
– Fitted line ± 2 SE(Fitted line)

– Regression line gives *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)?



Prediction intervals for individual observations

"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 CI 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.

Example: Prediction and Outliers in Regression

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 ≈ 3500 (p 90)
- Predictions suggest profitable

– Model without "special cottage"
- $R^2 \approx 0.08$, RMSE ≈ 3500 (p94-95)
- Predictions are useless

– *Should we keep the outlier, or should we exclude the outlier?*

Liquor sales and display space Display.jmp, page 99

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

– Fit of the two models is not distinguishable over the range of observed data.

– Predictions out to 20 feet are *very* sensitive to transformation
Prediction interval at 20 feet is far from range of data.
Very sensitive: Log pred. interval does not include reciprocal pred (p111)

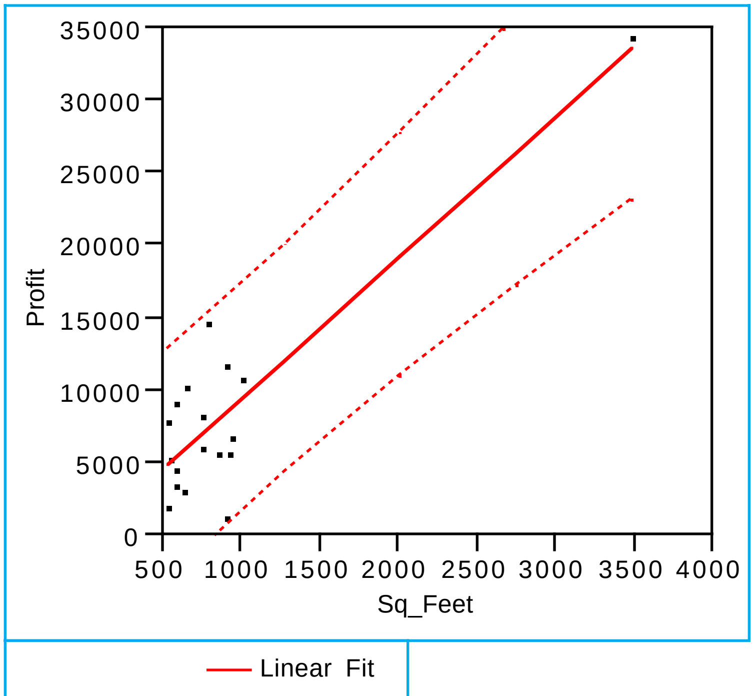
– *Have we captured the "true" uncertainty*

Philadelphia housing prices Phila.jmp, page 62
Further example of issues with outlying observations.

Next Time

Multiple regression
Using more than one predictor to reduce the unexplained variation and control for other sources of variation.

Profit By Sq_Feet



Linear Fit

Profit = -416.86 + 9.75055 Sq_Feet

Summary of Fit

RSquare	0.780
RSquare Adj	0.766
Root Mean Square Error	3570.379
Mean of Response	8347.799
Observations (or Sum Wgts)	18.000

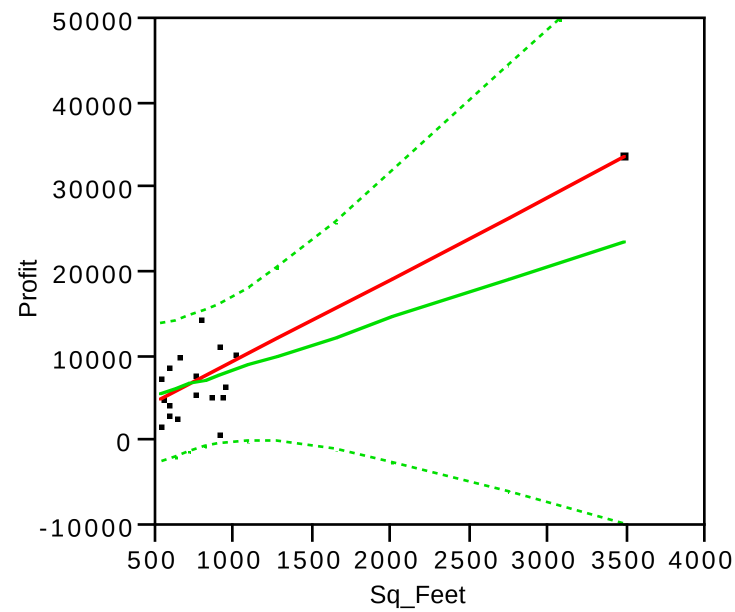
Analysis of Variance

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Ratio**
Model	1	721736309	7.2174e8	56.6174
Error	16	203961649	12747603	**Prob>F**
C Total	17	925697959		<.0001

Parameter Estimates

Term	**Estimate**	**Std Error**	**t Ratio**	**Prob>|t|**
Intercept	-416.86	1437.02	-0.29	0.7755
Sq_Feet	9.75	1.30	7.52	<.0001

Profit By Sq_Feet



Linear Fit
Linear Fit

Linear Fit

Linear Fit

Profit = 2245.4 + 6.13702 Sq_Feet

Summary of Fit

RSquare	0.075
RSquare Adj	0.014
Root Mean Square Error	3633.591
Mean of Response	6822.899
Observations (or Sum Wgts)	17.000

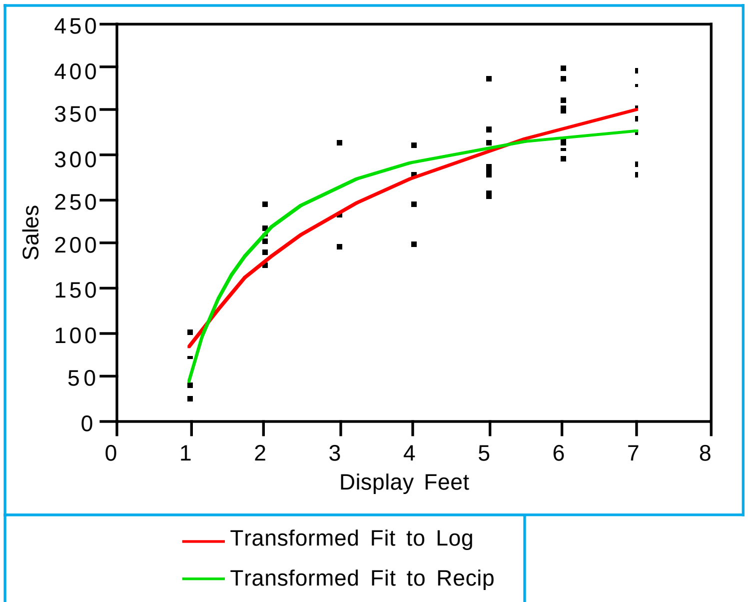
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	16105172	16105172	1.2198
Error	15	198044803	13202987	**Prob>F**
C Total	16	214149975		0.2868

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2245.40	4237.25	0.53	0.6039
Sq_Feet	6.14	5.56	1.10	0.2868

Sales By Display Feet



Transformed Fit to Log

Sales = 83.5603 + 138.621 Log(Display Feet)

Summary of Fit

RSquare	0.815
RSquare Adj	0.811
Root Mean Square Error	41.308
Mean of Response	268.130
Observations (or Sum Wgts)	47.000

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	83.56	14.41	5.80	<.0001
Log(Display Feet)	138.62	9.83	14.10	<.0001

Transformed Fit to Recip

Sales = 376.695 – 329.704 Recip(Display Feet)

Summary of Fit

RSquare	0.826487
RSquare Adj	0.822631
Root Mean Square Error	40.04298
Mean of Response	268.13
Observations (or Sum Wgts)	47

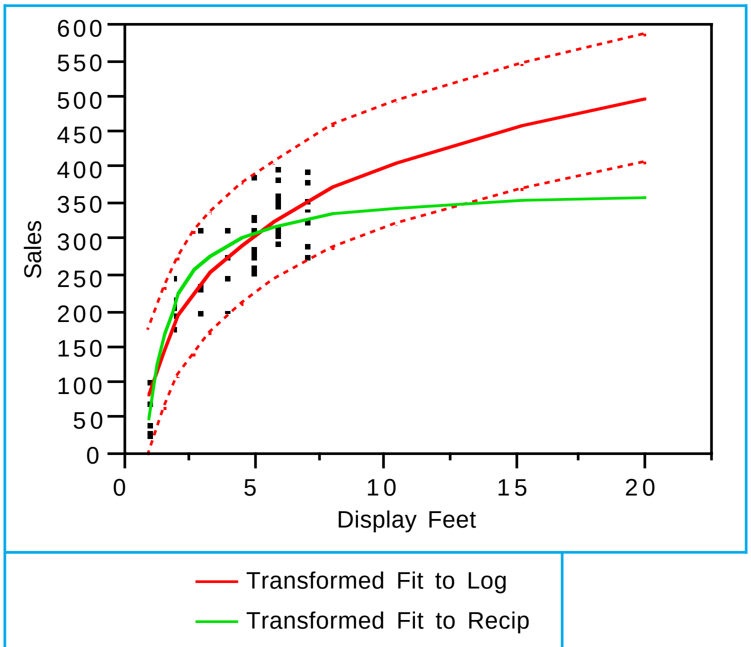
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	343692.15	343692	214.3468
Error	45	72154.79	1603	**Prob>F**
C Total	46	415846.94		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	376.70	9.44	39.91	<.0001

Sales By Display Feet



Transformed Fit to Log

Sales = 83.5603 + 138.621 Log(Display Feet)

Summary of Fit

RSquare	0.815
RSquare Adj	0.811
Root Mean Square Error	41.308
Mean of Response	268.130
Observations (or Sum Wgts)	47.000

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	83.56	14.41	5.80	<.0001
Log(Display Feet)	138.62	9.83	14.10	<.0001

Transformed Fit to Recip

Sales = 376.695 – 329.704 Recip(Display Feet)

Summary of Fit

RSquare	0.826
RSquare Adj	0.823
Root Mean Square Error	40.043
Mean of Response	268.130
Observations (or Sum Wgts)	47.000

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	376.70	9.44	39.91	<.0001
Recip(Display Feet)	-329.70	22.52	-14.64	<.0001