Inference 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.*

Midterm Exam #2

Next Tuesday April 4, from 6-8 p.m. in Annenberg 110.

Preparing for the exam

Solutions to the old exams (no prediction intervals, "calculation" formulas)
Review assignments, material discussed in class.
Sample regression problems from text are 44, 46, 47a:
Interpret slope and intercept, use the equation to predict,
Know the relationship to correlation, answer if slope is $\neq 0$.

Assignment #4

Due today.

Regression Models

Inference questions

Is there a relationship between the predictor and response?
Can you use this measurement usefully to predict the response?
How accurately can you predict the response?
How accurately can you determine the slope?
What's a confidence interval for the slope? For the intercept?

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

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

Utopian model

What is the data generating process that produces data, according to this model?
See the utopian data example in the casebook (page 47).

Estimating the Regression Line from Data

Least squares

Minimize the sum of the squared vertical deviations from the fitted line. These deviations from the fitted line are called "residuals" and are in effect estimates of the unobserved error terms in our utopian model.

JMP will do all of the calculations for us, and print the least squares estimates of the slope and intercept in the fitted model.

Why squared, vertical deviations?

Squared: It makes calculus easy (e.g, what's the derivative of |x|?), and it works optimally if the data are normally distributed.

Vertical: Think of predicting Y from X; we want to chose the line that minimizes this type of error.

Inference in Regression

Accuracy of the estimators

The standard errors of the estimators can be used to form confidence intervals of the usual form (for example)

(estimated slope) ± 2 (SEs of estimated slope)

or test hypotheses, such as whether the slope is zero.

Note well: Slope ≠ 0 ←→ X & Y are related

One more estimator

How variable are the observations around the regression line? That is, what is the best estimator for the error variance, σ^2 (or the mean squared error).

As in anova, you simply sum the squared deviations from the fitted model (now a line) and divide by the degrees of freedom (here n-2, since we estimate both a slope and an intercept in computing the residuals).

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

A popular additional summary

R^2 in regression is the square of the usual correlation between the predictor X and the response Y, so $0 \le R^2 \le 1$. It has the important interpretation in regression as

$100\ R^2$ = percentage of variation in response *explained* by fitted model

JMP for "Simple" Regression (One Predictor)

All done via the fitting button appearing in the *Fit Y by X* view.

Examples of Inference in Regression

Some "interesting" questions...

(1) How much should you expect to pay for a diamond that weighs .25 carats? That weighs 1 carat?

(2) What level of promotion maximizes profit?
How close do you need to be to that value?

(3) Where are levels of cellular telephone use headed?
Where do you think subscription rates will be next year?

Linear Model: Confidence Intervals and Prediction Accuracy

Diamond data from web page

Prices (in Singapore $'s) for diamonds sold retail in 1990. Linear relationship of price and size of stone in carats.

Summary of fitted model

Linear Fit

Price (Singapore dollars) = -259.63 + 3721.02 Weight (carats)

Summary of Fit

RSquare	0.978
Root Mean Square Error	31.841
Mean of Response	500.083
Observations (or Sum Wgts)	48

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	2098596.0	2098596	2069.991
Error	46	46635.7	1014	Prob>F
C Total	47	2145231.7		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-259.6	17.3	-14.99	<.0001
Weight (carats)	3721.0	81.8	45.50	<.0001

Interpreting the summary

R^2: the correlation is quite high, with little variation in the response not captured (or explained) by the fitted model.

Standard error of intercept: The fitted intercept is 15 SEs below zero, and the confidence interval for the population intercept does not include zero.

Standard error of the slope: The slope is far from zero as well, indicating a significant relationship.

Confidence interval for slope: Based on the fitted model, the cost per carat for this type of diamond is \$3720, with a range of [3720 ± 2(82)]\$/carat.

Accuracy of prediction

The estimated error variance lies in the anova summary (as in anova) and the shown RMSE (for root of the mean squared error) indicates that the residual SD around the fitted line is about 32.

JMP will show these bands around the fitted line. (Use the Conf curve: individual button associated with the fitted curve in the Fit Y by X view.)

Prediction of the cost of a stone

For a stone weighing 0.25 carats, the fitted model implies a prediction of the expected cost (in Singapore dollars) as

Price (Singapore dollars) = -259.63 + 3721.02 (0.25) = \$670

with a 95% range given by ±2 SD of the residuals or,

\$670 ± 2(32)

Extrapolation to one carat

Like the intercept, trying to predict the cost of a one-carat stone from this data is foolish. Such a prediction requires that we believe that the linear relationship extends beyond the range of our data, a foolish proposition with this data. (Such a prediction would likely to grossly under-predict – imagine getting a 2 carat rock for the cost predicted by this model of slighly less than \$7,200.).

Non-Linear Model: CIs and Prediction Accuracy

Two questions

What is the optimal amount of display space?
How accurately is it determined from this data?

Fitted model

Here's the output for the fitted model using JMP. Note that the fit is clearly "significant", capturing % of the variation in sales with a slope that is clearly different from zero.

Transformed Fit to Log

Sales = 83.5603 + 138.621 Log(Display Feet)

Summary of Fit

RSquare	0.815
Root Mean Square Error	41.308
Mean of Response	268.130
Observations (or Sum Wgts)	47

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	339060.4	339060	198.7031
Error	45	76786.5	1706	**Prob>F**
C Total	46	415846.9		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	83.6	14.4	5.80	<.0001
Log(Display Feet)	138.6	9.8	14.10	<.0001

Interpreting the slope

In models with transformed variables, you have to think pretty hard about what the intercept and slope *mean*. In this example, the

intercept = predicted sales ($83.6) when one foot is used for the display

slope/100 = expected growth in sales ($1.4) per 1% change in display space

Optimal display space

If we accept the model using the log of space as a predictor (other transformations are close), then the fitted least squares equation is

Est Sales = 83.5603 + 138.621 Log(Display Feet)

If we assume that the cost for using the space is $50 per foot, then the profit as a function of the number of feet of display space used is

Profit(Feet) = 83.5603 + 138.621 Log(Feet) - 50(Feet)

which has its maximum (set the derivative to zero) at

139/50 = 2.78 feet

Confidence intervals are your friends

How can we get an interval for the location of the optimal display space? Simply plug the 95% interval for the fitted slope into the above equation and get

$$[139 \pm 2(9.8)]\ /50 = [119.4, 158.2]/50 = [2.38, 3.16] \text{ feet}$$

or about 2.5 to 3 feet of shelf space.

Accuracy of prediction

The fitted model explains much of the variation in the response, and the predictions (avoiding extrapolation) are accurate (with 95% confidence) to within about ±2 RMSE = ±2(41) = ± $82 .

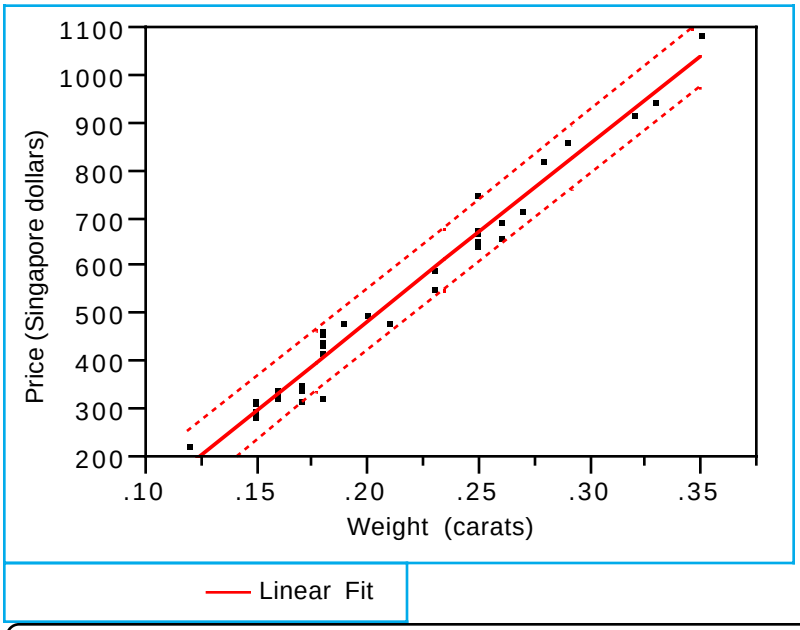
For example, if we predict the level of sales at 4 display feet, the we expect sales to be

Est Sales = 83.56 + 138.6 Log(4) = $275.7

which we can expect to be accurate to within about ±$82 of the actual sales.

So is this a good model?

What are you going to use the model to do? Are these prediction sufficiently accurate for your purpose. For many, a potential error of $82 on a prediction of $275 is too large, regardless of whether the fit is "statistically significant".



Linear Fit

Price (Singapore dollars) = -259.63 + 3721.02 Weight (carats)

Summary of Fit

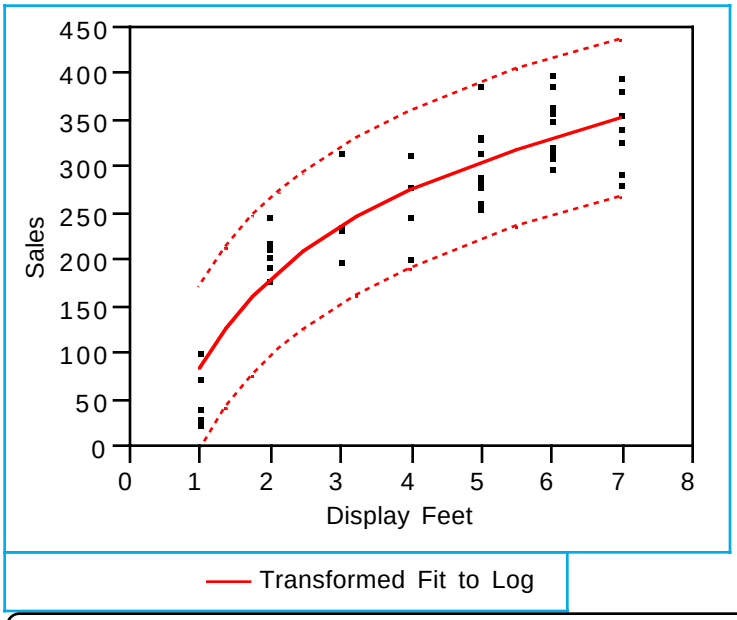
RSquare	0.978
RSquare Adj	0.978
Root Mean Square Error	31.841
Mean of Response	500.083
Observations (or Sum Wgts)	48.000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	2098596.0	2098596	2069.991
Error	46	46635.7	1014	**Prob>F**
C Total	47	2145231.7		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-259.63	17.32	-14.99	<.0001
Weight (carats)	3721.02	81.79	45.50	<.0001



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

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	339060.41	339060	198.7031
Error	45	76786.53	1706	**Prob>F**
C Total	46	415846.94		<.0001

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