Simple Regression Model

INSR 260, Spring 2009

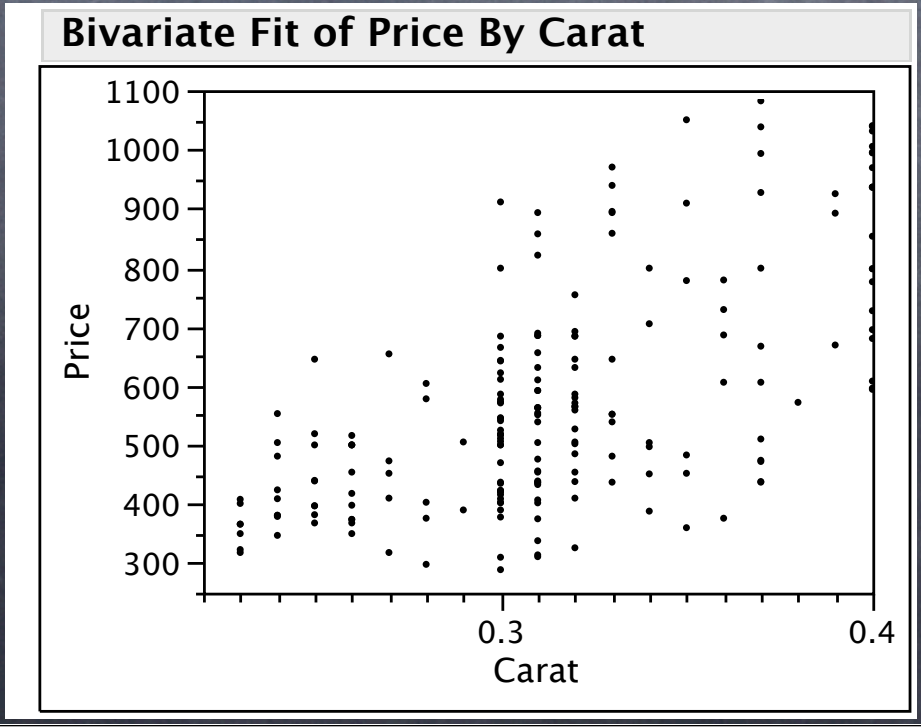
Bob Stine

Overview

- Simple Regression Model (SRM)
- Estimators, terminology
- Assumptions
- Inference
- Prediction
- Examples (both from Stat 102)
  - Promotion response
  - Diamond values

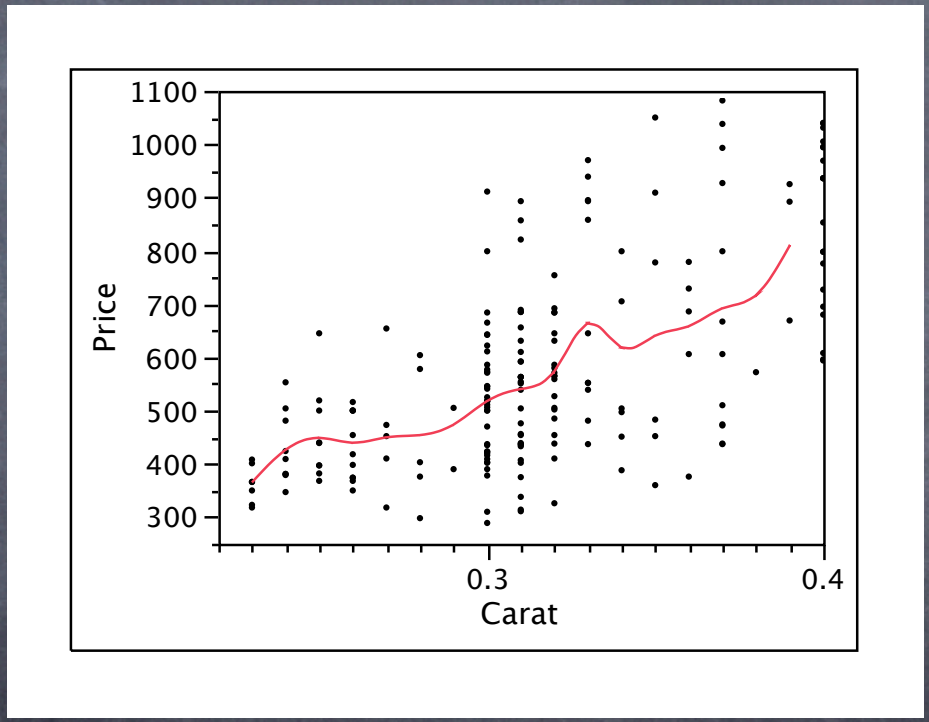
Example: Pricing Diamonds

- Questions
  - How much should I expect to pay for a diamond?
  - How accurate is such an estimate?
  - How do prices depend on the weight of the diamond?
- Data
  - n = 180 diamonds, relatively small (less than 0.4 carats)



Direct Approach

- Average the price of diamonds of each weight, then connect the dots with lines.



- Produces a "smooth" curve curve and an estimate of accuracy, but does not answer the question of how these are related?

Simple Regression Model

- Association versus causation
- Equation relates observed x and y (**n** cases)
  Conditional means: $E\ Y|X = \beta_0 + \beta_1 X = \mu_{y|x}$
  Observations: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Assumptions
  - Independent observations
  - Equal variance σ^2
  - Normal distribution around line

    $y_i \sim N(\mu_{y|x}, \sigma^2)$ $\varepsilon_i \sim N(0, \sigma^2)$
- Three parameters: β_0 , β_1 , σ^2

Least Squares

- Criterion
  - Find estimates that minimize sum of squared deviations

    $\min_{a0,a1} \Sigma(y_i - a_0 - a_1 x_i)^2$
  - Solution

    $b_1 = \text{cov}(x,y)/\text{var}(x)$ $b_0 = y - b_1 x$

$$\text{cov}(x,y) = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

$$\text{var}(x) = \frac{\sum_i (x_i - \overline{x})^2}{n-1}$$

- Fitted values, residuals
  - Fitted values (on the line) $\hat{y} = b_0 + b_1 x$
  - Residual deviations $e = y - \hat{y}$
- Standard error of regression (estimates σ^2)
  - $s^2 = \Sigma e_i^2/(n-2)$
  - degrees of freedom
  - aka: root mean squared error (RMSE), SD of residuals

Goodness of Fit

- R-squared statistic
  - Square of correlation between Y and X
  - Percentage of "explained" variation

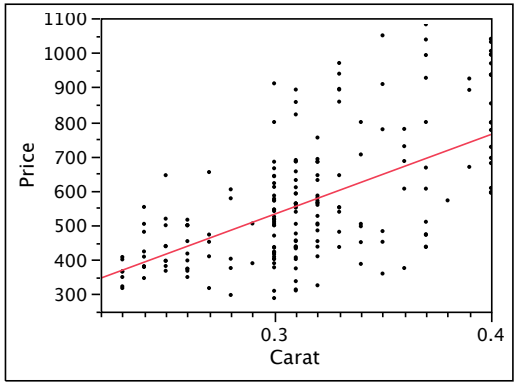
$$R^2 = \frac{\text{Explained SS}}{\text{Total SS}} = \frac{\sum_i (\hat{y}_i - \overline{y})^2}{\sum_i (y_i - \overline{y})^2} = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \overline{y})^2}$$

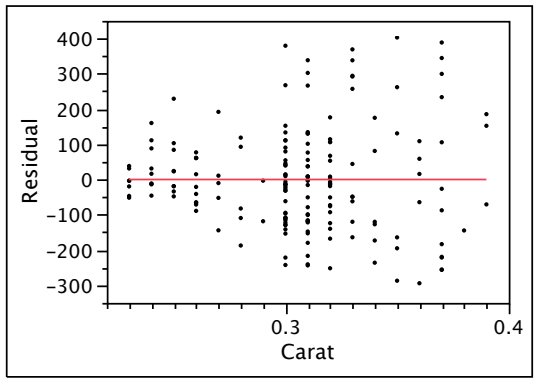
- Adjusted R-squared
  - Takes account of number of observations

$$\overline{R^2} = 1 - \frac{s^2}{\text{var}(y)}$$

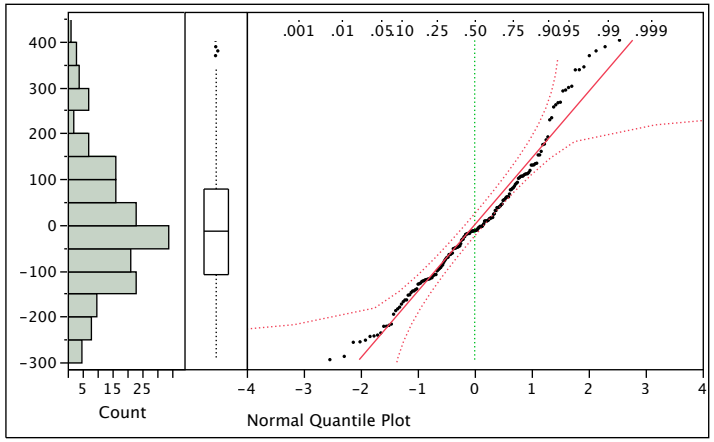
Checking Assumptions

- Scatterplots of Y on X, e on X





- Normal quantile plot of residuals



Inference

- Derived from the standard error of the intercept and slope

$$\mathrm{Var}(b_1) = \frac{\sigma^2}{\sum_i (x_i - \overline{x})^2}$$

- Three equivalent methods
  - Confidence interval
  - t-statistic
  - p-value
- Interpretation of estimates, uncertainty?

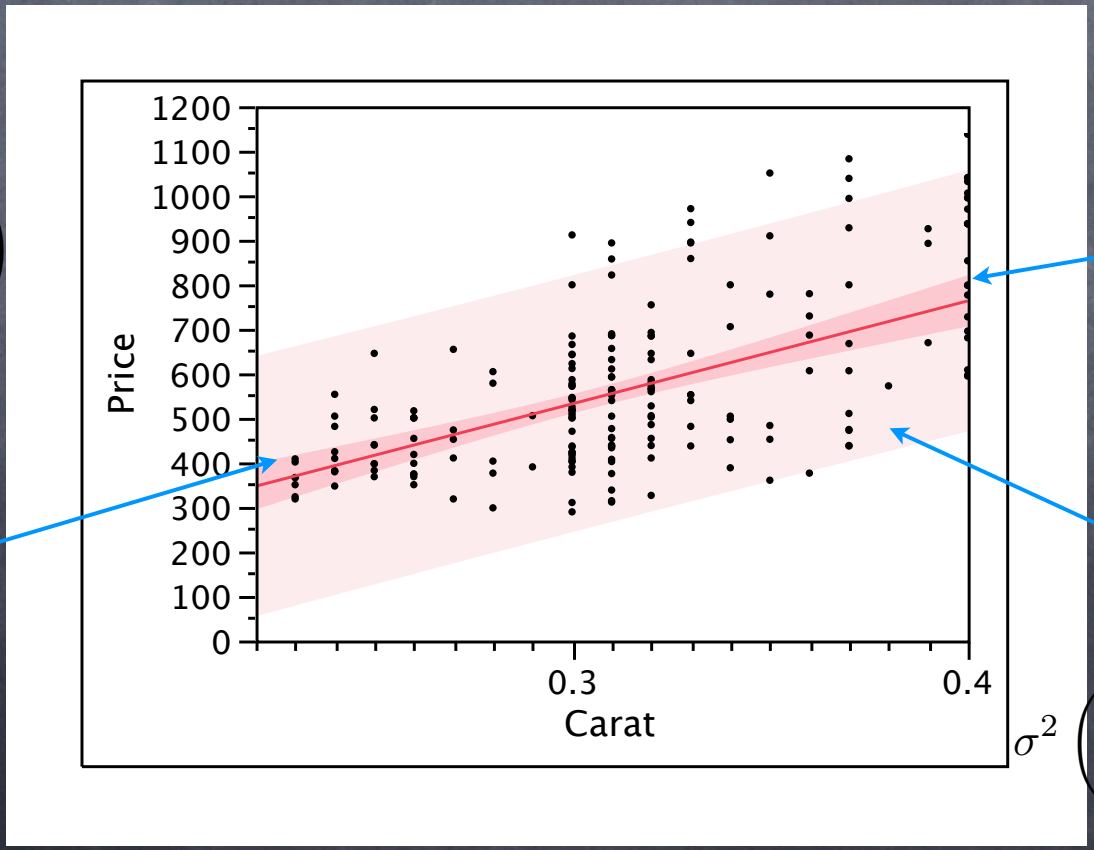
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-162.1909	87.42039	-1.86	0.0652
Carat	2312.3973	284.5074	8.13	<.0001*

Prediction

- Interpretation
  - Where is the population regression line? ($\mu_{Y|X}$)
    - What is the average price of all diamonds that weigh 0.3 carats?
  - What can be anticipated about a future observation?
    - How much for a specific 0.3 carat diamond?

$$E(\mu_{y|x} - \widehat{y}_x)^2 = \sigma^2\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_i (x_i - \overline{x})^2}\right)$$



Effect of extrapolation

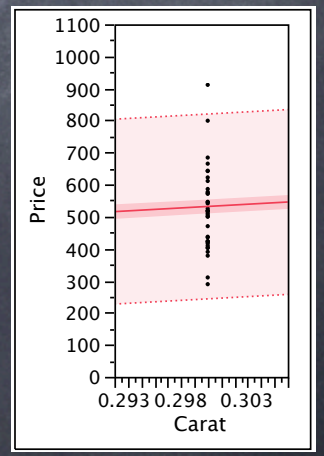
Conf Int for $\mu_{Y|X}$

Pred Int for Y

$$\sigma^2\left(1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_i (x_i - \overline{x})^2}\right)$$

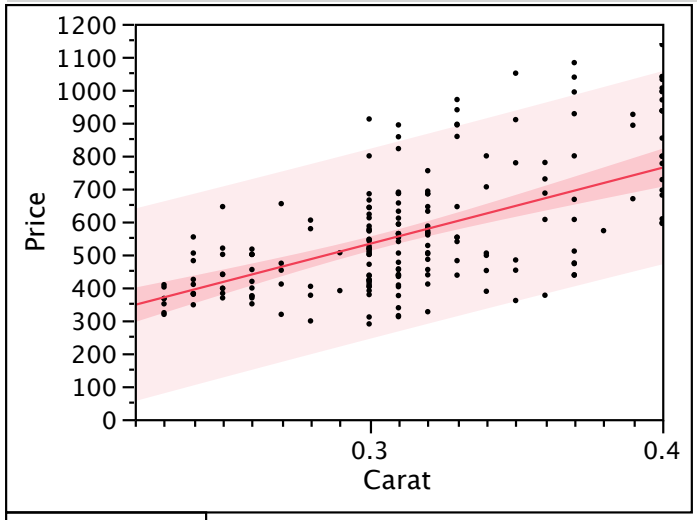
Prediction/Conf. Interval

- Estimates are same for both mean and specific diamond
  - $\hat{y}$ = -162.2 + 2312.4 Carats
    = -162.2 + 2312.4 (0.3)
    = $531.52
- Intervals differ in anticipated uncertainty
  - For the average, JMP (use Fit Model to save interval)
    $509.92 to $553.14
  - For a single diamond, prediction interval is
    $243.18 to $819.88
  - 95% Prediction interval is approximately
    $\hat{y} \pm 2$ RMSE



Diamond Example

Bivariate Fit of Price By Carat



Linear Fit

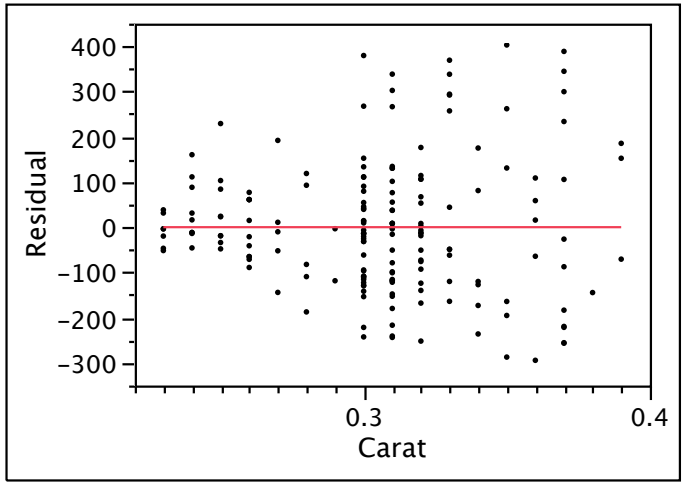
Price = -162.1909 + 2312.3973*Carat

Summary of Fit

RSquare	0.270671
RSquare Adj	0.266573
Root Mean Square Error	145.7105
Mean of Response	542.8333
Observations (or Sum Wgts)	180

Parameter Estimates

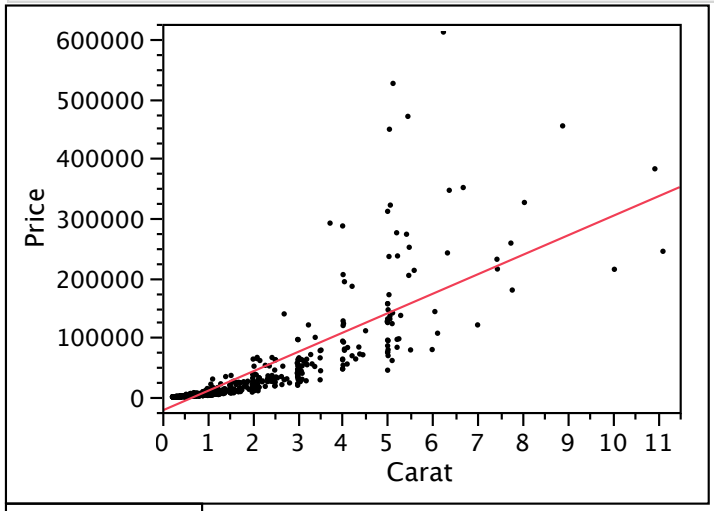
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-162.1909	87.42039	-1.86	0.0652
Carat	2312.3973	284.5074	8.13	<.0001*



Interpretation?
Comments?
Problems?

More Diamonds

Bivariate Fit of Price By Carat



Linear Fit

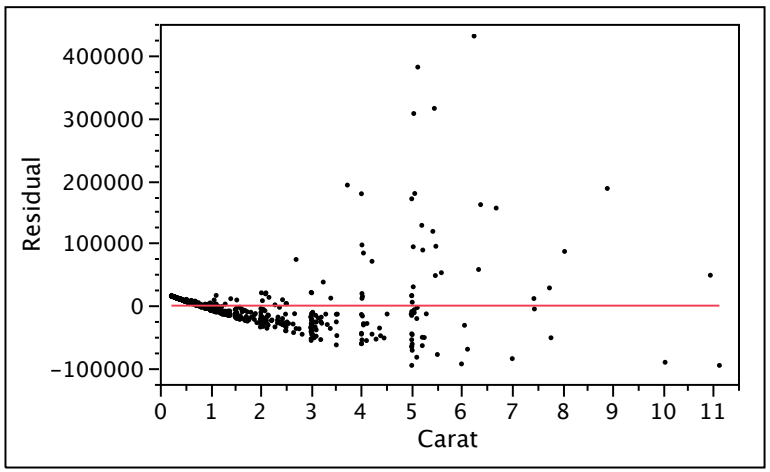
Price = -22655.47 + 32538.514*Carat

Summary of Fit

RSquare	0.652987
RSquare Adj	0.652639
Root Mean Square Error	34109.01
Mean of Response	21957.11
Observations (or Sum Wgts)	1000

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-22655.47	1491.049	-15.19	<.0001*
Carat	32538.514	750.8501	43.34	<.0001*



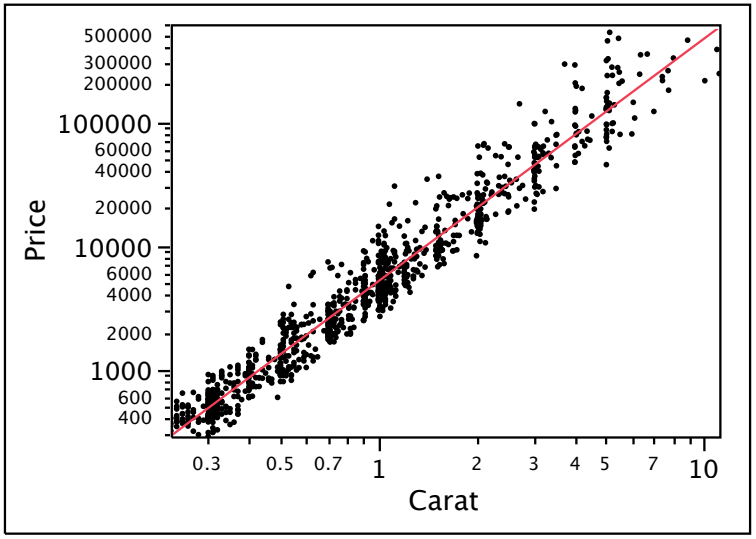
Estimate of cost/carat?

Negative intercept?

Possible to fix?

Log-Log Scale

- Same expanded sample, but on a log-log scale



Transformed Fit Log to Log

Log(Price) = 8.5485138 + 1.9571815*Log(Carat)

Summary of Fit

RSquare	0.950228
RSquare Adj	0.950178
Root Mean Square Error	0.377488
Mean of Response	8.415271
Observations (or Sum Wgts)	1000

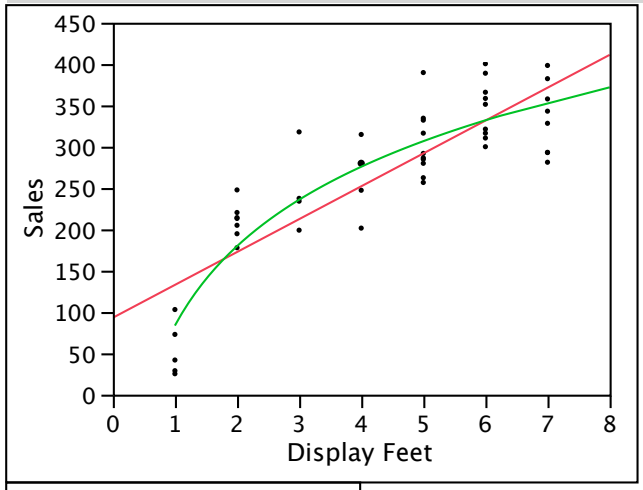
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	8.5485138	0.011976	713.79	0.0000*
Log(Carat)	1.9571815	0.014179	138.03	0.0000*

- Clearly a better fit, but what does it mean?

Promotion Response

Bivariate Fit of Sales By Display Feet



Linear Fit

Sales = 93.032311 + 39.75648*Display Feet

Summary of Fit

RSquare	0.711975
RSquare Adj	0.705574
Root Mean Square Error	51.59123
Mean of Response	268.13
Observations (or Sum Wgts)	47

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	93.032311	18.22782	5.10	<.0001*
Display Feet	39.75648	3.769509	10.55	<.0001*

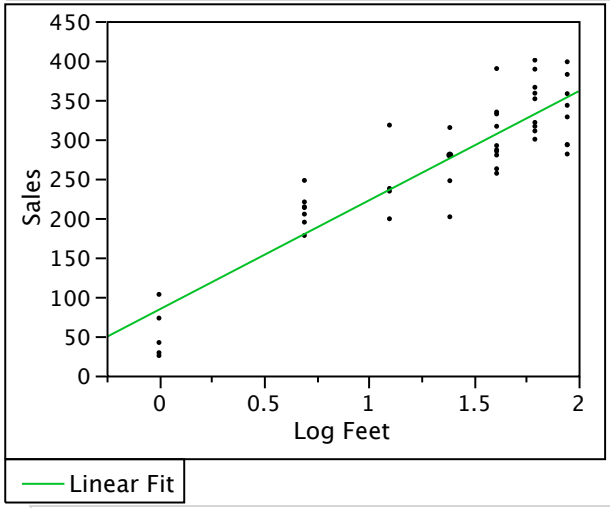
Transformed Fit to Log

Sales = 83.560256 + 138.62089*Log(Display Feet)

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	83.560256	14.41344	5.80	<.0001*
Log(Display Feet)	138.62089	9.833914	14.10	<.0001*

Bivariate Fit of Sales By Log Feet



Linear Fit

Sales = 83.560256 + 138.62089*Log Feet

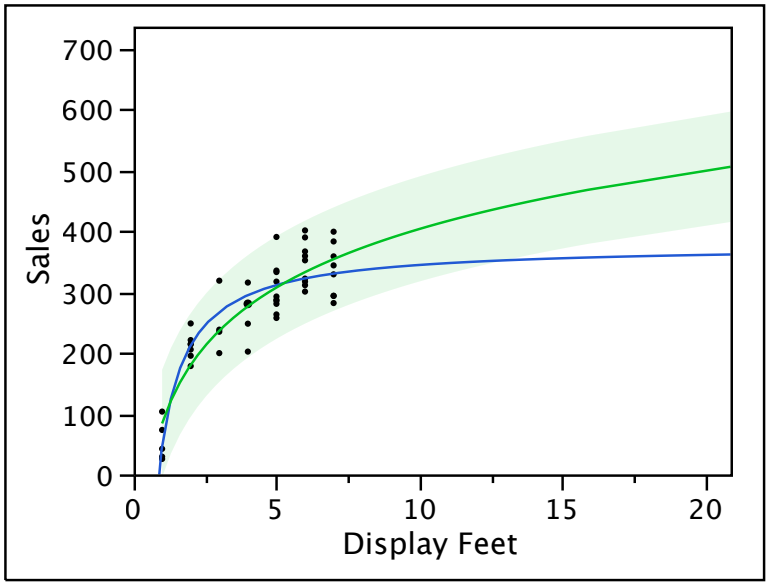
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	83.560256	14.41344	5.80	<.0001*
Log Feet	138.62089	9.833914	14.10	<.0001*

Same model, different scales on the axes

What is the interpretation of the coefficient of the log in this model?

Extrapolation



Comparable R^2, RMSE but very different extrapolations and implications for sales.

r s or t to pro

Sales = 376.69522 - 329.70421*Recip(Display Feet)

u ry o t

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

r t r st t s

r	st t	t rror	t t o	ro |t|
Intercept	376.69522	9.439455	39.91	<.0001*
Recip(Display Feet)	-329.7042	22.51988	-14.64	<.0001*

r s or t to o

Sales = 83.560256 + 138.62089*Log(Display Feet)

u ry o t

RSquare	0.815349
RSquare Adj	0.811246
Root Mean Square Error	41.3082
Mean of Response	268.13
Observations (or Sum Wgts)	47

r t r st t s

r	st t	t rror	t t o	ro |t|
Intercept	83.560256	14.41344	5.80	<.0001*
Log(Display Feet)	138.62089	9.833914	14.10	<.0001*

Summary

- Simple Regression Model (SRM)
- Least squares estimators
- Role of assumptions
- Methods of inference: tests and intervals
- Importance of transformations, particularly logs
- Prediction, role of extrapolation