This Week

Monday

- Recap: Project Risk Analysis

Wednesday

- Lab Exercise: Trend Estimation for Existing Liquair-Pro Customers

Topics

- Simple Linear Regression
- Evaluating Regression Output
- Data Transformation

A Problem Solving Framework

1. *Define the Problem*
2. *Collect and Organize Data*
3. *Characterize Uncertainty and Data Relationships*
4. *Build an Evaluation Model*
5. *Formulate a Solution Approach*
6. *Evaluate Potential Solutions*
7. *Recommend a Course of Action*

Regression Models

- Regression models ***quantify relationships between variables***
- Used for: Description / Prediction / Control
- Applications in all areas of business and government:
  - Target customers for direct marketing / having low credit risk
  - Forecast sales / demand / market share / investment return
  - Set economic/monetary policy factors to control inflation

Sales = ***f*** (***Advertising***, ***Sales force***, ***Shop surface***, ***Population density***, ...)

Response variable

What we are trying to explain or predict

Explanatory variables

Factors that influence the response variable

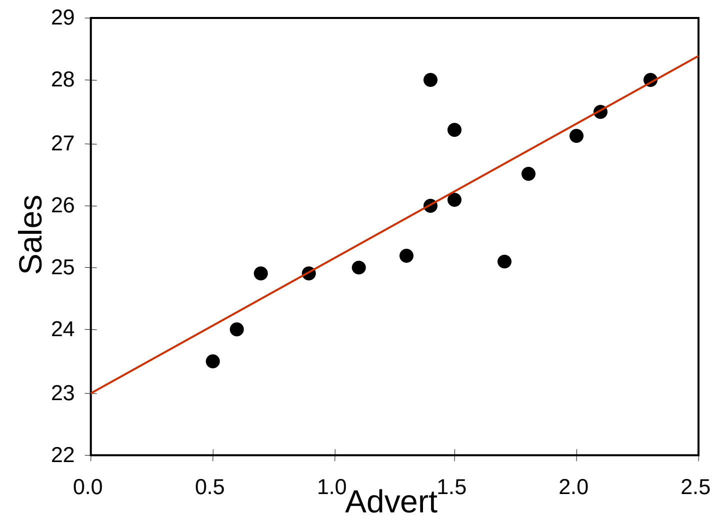
Simple Linear Regression

- A ***simple regression*** model assumes that a ***linear relationship exists between two variables***, plus a random error term:

$$\boxed{Y_i = a + b \cdot X_i} + e_i$$

response variable (Y_i), **intercept coefficient** (a), **slope coefficient** (b), **explanatory variable** (X_i), **residual error** (e_i)

$$\underbrace{\text{Observed Value}}_{} = \underbrace{\text{Fitted Value}}_{\hat{Y}_i} + \underbrace{\text{Residual}}_{}$$

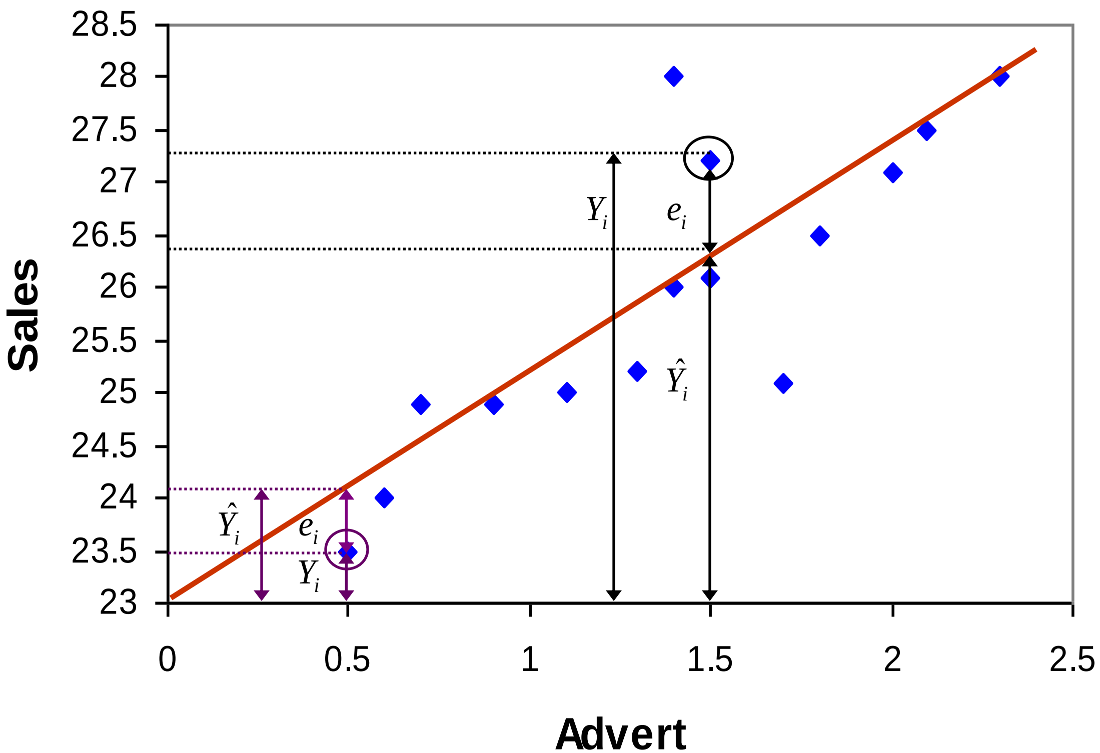


- The goal is to find estimates for ***a*** and ***b*** (denoted $\hat{a}$ and $\hat{b}$) that explain, or “fit”, the observed data best. Intuitively, we want to find slope and intercept coefficients that result in ***small residual errors***.

Residual Errors

Sales = ***a*** + ***b***·*Advert* + ***error***

- A ***residual error*** e_i is defined as the difference between the observed value of a response variable (Y_i) and its fitted value on the regression line ($\hat{Y}_i = \hat{a} + \hat{b} \cdot X_i$). Errors can be positive or negative.



Finding the "Best Fit" Line

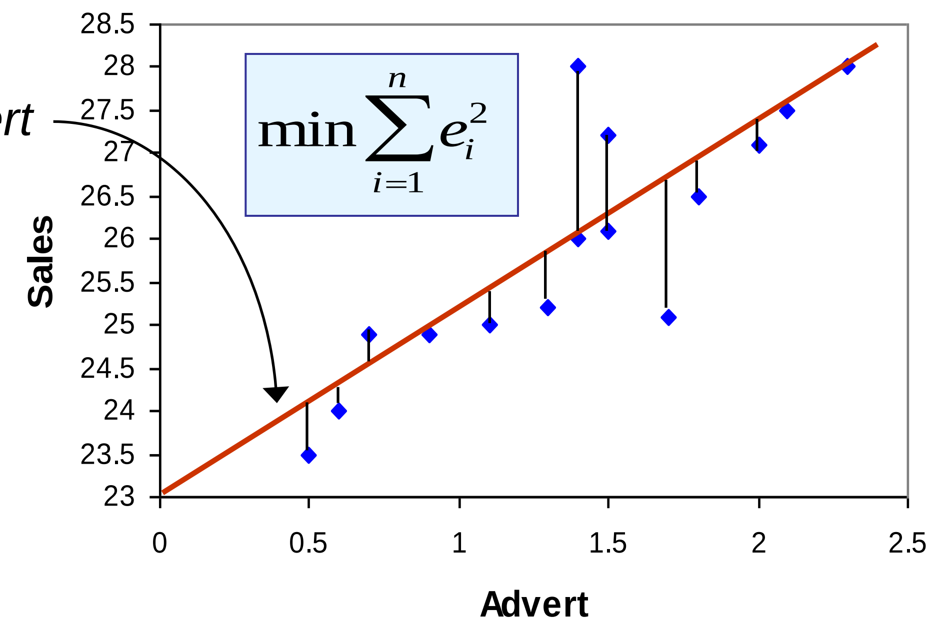
$$Sales = \mathbf{a} + \mathbf{b} \cdot Advert + \mathbf{error}$$

- To determine estimates for ***a*** and ***b*** that "fit" the data best, we find the ***least squares line*** – the line that minimizes the sum of the squared residual errors:

Least squares line is:

$$Sales = \underbrace{\overset{\hat{a}}{22.94}}_{} + \overset{\hat{b}}{2.16} \cdot Advert$$

Best guess ***estimates*** for the true ***a*** and ***b*** – these estimates have ***sampling error***.



Excel: =INTERCEPT(Sales, Advert) = 22.94 $(\hat{a})$

=SLOPE(Sales, Advert) = 2.16 $(\hat{b})$

How Good Is the Model?

$$Sales = \mathbf{22.94} + \mathbf{2.16} \cdot Advert + \textbf{\textit{error}}$$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.83
R Square	0.70
Adjusted R Square	0.67
Standard Error	0.81
Observations	15

	Coeffts	*Std Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	22.94	0.59	39.13	0.000000	21.67	24.21
Advert	2.16	0.39	5.47	0.000108	1.31	3.01

"Goodness" measures:

- Significance of explanatory variable (Advert)
- Standard Error of Estimate (s_e)
- R Square and Adjusted R Square statistics

Significance of Explanatory Variable

	Coeffts	*Std Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	22.94	0.59	39.13	0.000000	21.67	24.21
Advert	2.16	0.39	5.47	0.000108	1.31	3.01

$\hat{b}$

SE of $\hat{b}$

Number of standard errors that $\hat{b}$ **is away from zero**

p-value of $\hat{b}$ **with** H_0**:** $b = 0$

95% CI for b

Want a small p-value (typically, < 0.05)

- Are sales related to advertising spending?
- Should *Advert* be in the model?
- Is the *Advert* coefficient significantly different from 0?

Standard Error of Estimate

SUMMARY OUTPUT

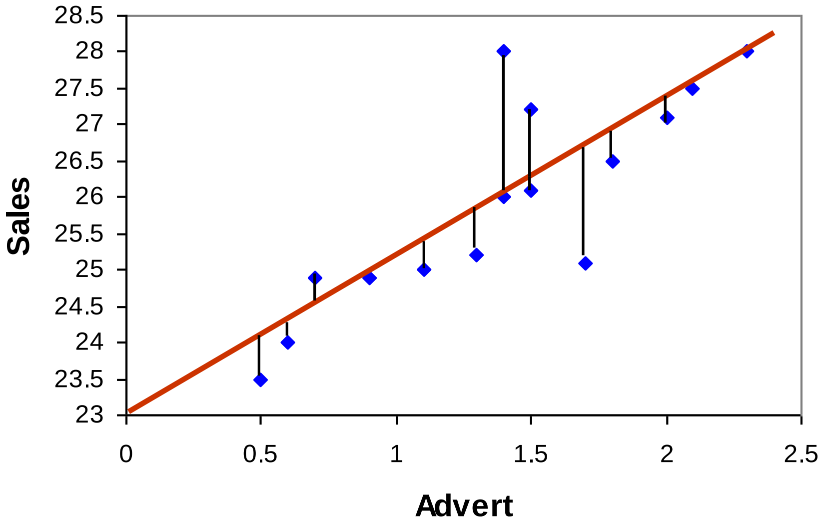
Regression Statistics	
Multiple R	0.83
R Square	0.70
Adjusted R Square	0.67
Standard Error	0.81
Observations	15

- ***Standard Error of Estimate*** or s_e: (Roughly) the standard deviation of the residual errors. The smaller the s_e, the better the model:

Excel: =STEYX(Sales, Advert) = 0.81

$$s_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

Number of coefficients being estimated



R Square Statistics

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.83
R Square	0.70
Adjusted R Square	0.67
Standard Error	0.81
Observations	15

- ***R Square***: R^2, or the ***coefficient of determination***, is the percentage of the variation in the response variable that is explained by the regression line: 0% < R^2 < 100%. The higher the R^2, the better the linear fit.

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \right) = 1 - \left(\frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \right)$$

Excel: =RSQ(Sales, Advert) = 0.70

- ***Adjusted R Square***: R^2 adjusted for the number of explanatory variables (important for multiple regression).

Using Regression Models to Predict

- Regression Model: *Sales* = **22.94** + **2.16** · *Advert* + ***error***

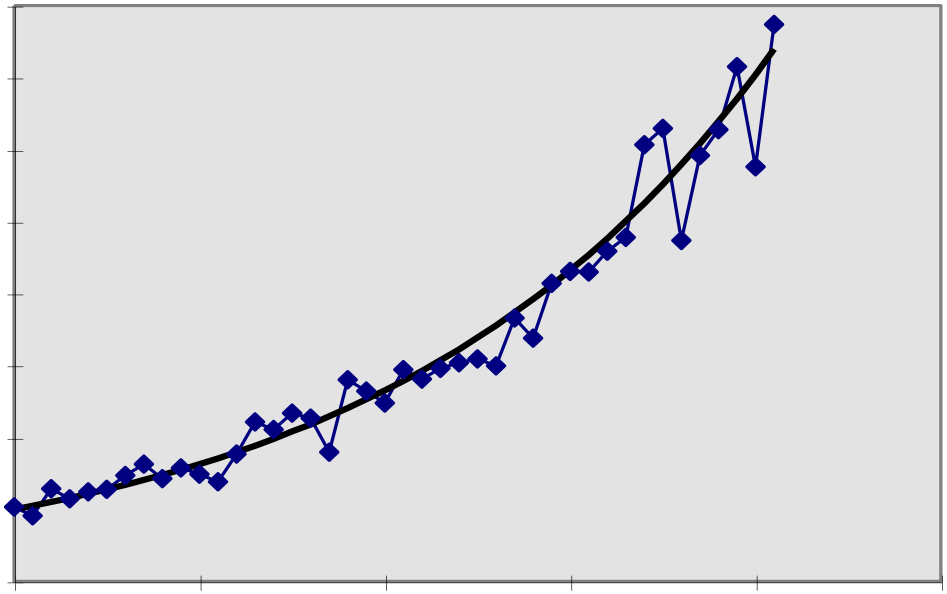
  $\hat{a}$ $\hat{b}$

- Sales ***prediction*** (point estimate) for a store spending $1.5M on advertising:

$$\hat{Y}_i = \hat{a} + \hat{b} \cdot X_i$$

Sales = **22.94** + **2.16** (1.5) = 26.18

Nonlinear Data Relationships



- What can we do if a data relationship appears to be ***nonlinear?***

➔ Variables which exhibit certain types of nonlinear relationships may be ***transformed into other variables which ARE linearly related***.

Transforming Data

Exponential Relationship:

$$Y = a \cdot e^{bX}$$

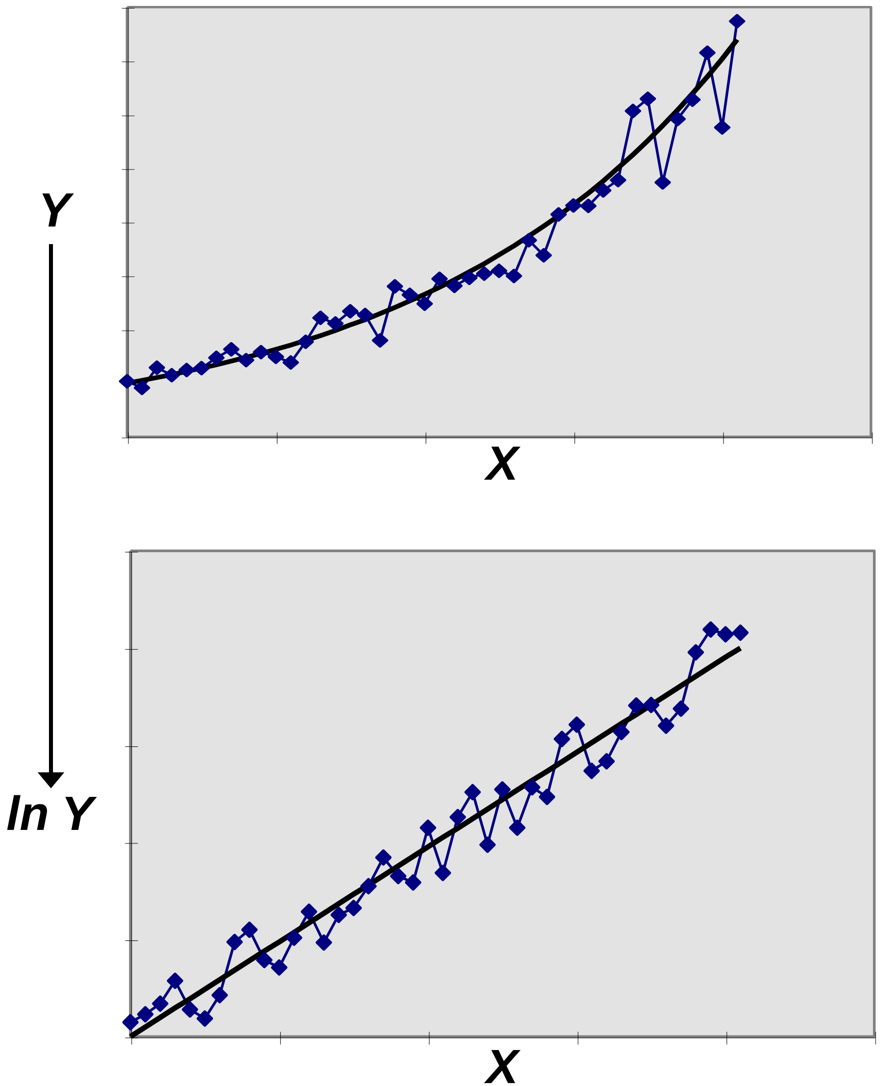
Make a <u>new response variable</u> ln(Y) which has a linear relationship with X.

$$\mathbf{\ln Y} = \ln a + b\mathbf{X}$$

In terms of parameters ***a*** and ***b***:

Slope *of new model =* ***b***

Intercept *of new model =* ***ln a***



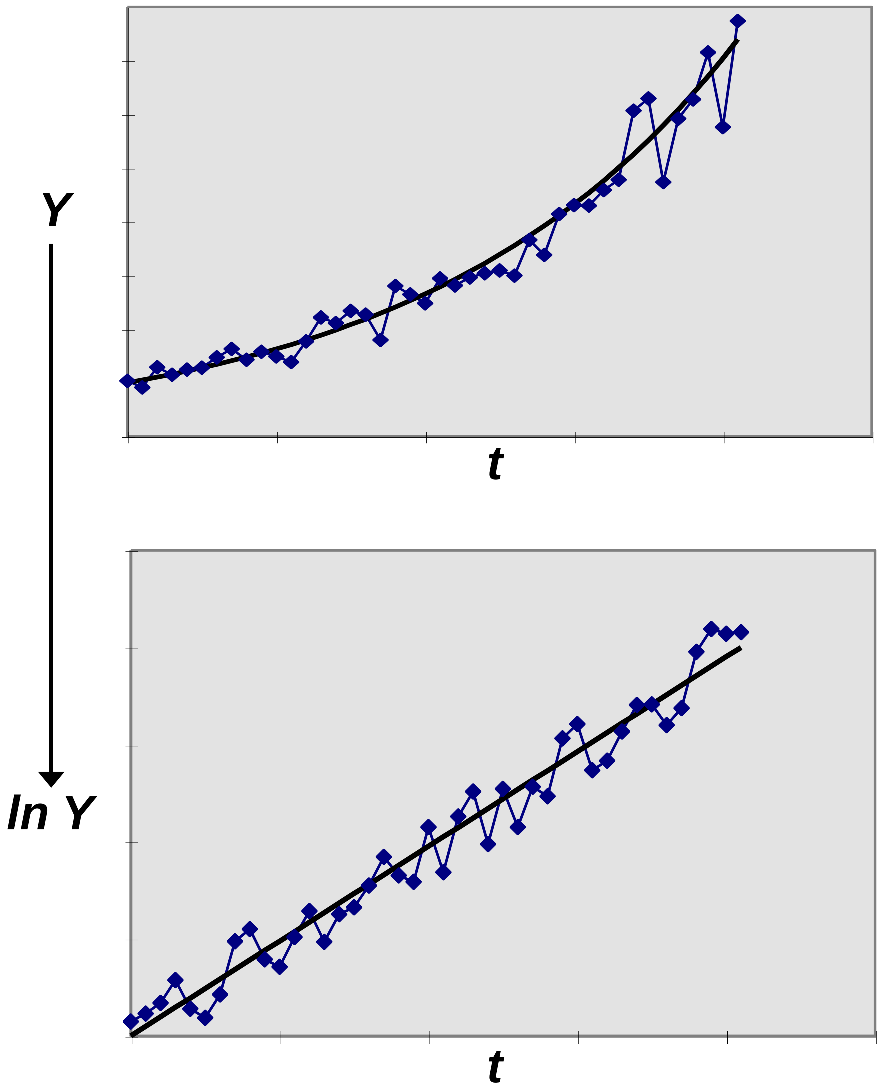
Liquair-Pro Data

- Some of Liquair-Pro's customers exhibit demand (***Y***) that grows exponentially in time (***t***):

$$Y_t = Y_0 \cdot (1 + r)^t$$

where ***r*** is the annual growth rate.

- To get point estimates for (future) Y_t and ***r***, as well as estimates for their standard errors, we can find the best fit exponential model $Y_t = a \cdot e^{bt}$ where $e^b = (1 + r)$.



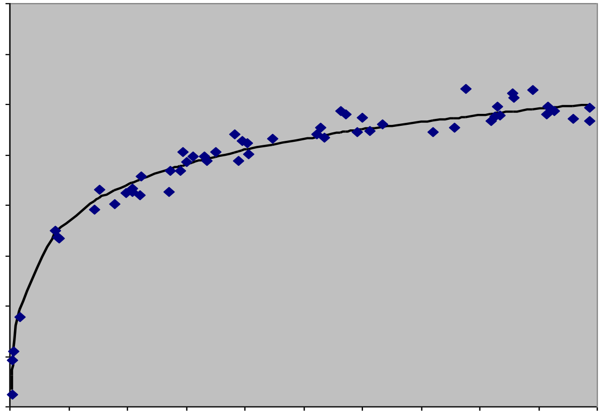
Other Common Transformations

Nonlinear Relationship:

Linearized form:

Logarithmic

Y = a + b ln X

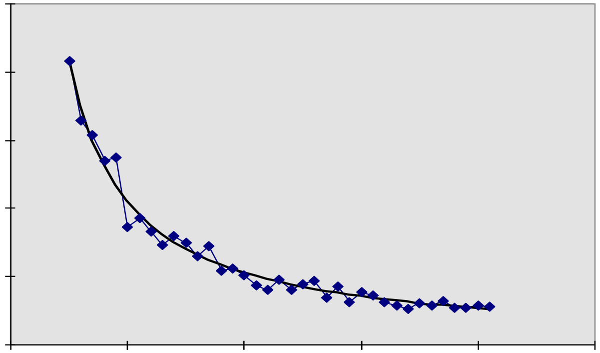


Fit Y against log X:

Y = *a* + *b*(***ln X***)

Reciprocal

Y = a + b/X

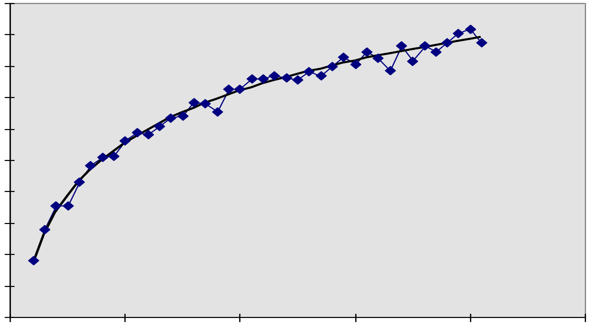


Fit Y against (1/X):

Y = *a* + *b* (***1/X***)

Multiplicative

Y = a·X^b



Fit log Y against log X:

ln Y = *ln a* + *b*(***ln X***)

Which Transformation?

- If there are multiple possible data transformations that "straighten out the data" equally well, then use the one that is ***easiest to interpret***:
  - ***Fewest*** explanatory variables (if multiple regression)
  - ***Logarithmic*** transformations typically have nice interpretations in terms of relating changes in X to changes in Y