Session 3

Simple Linear Regression (I): Linear Fit

Managerial Decisions

What is the relationship between price and demand for our product?

What is the relationship between commission and sales?

What is the relationship between return of this stock and the market?

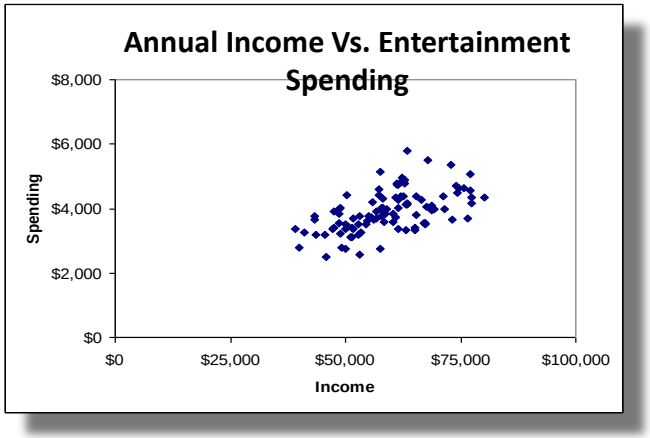
What is the relationship between advertising and sales?

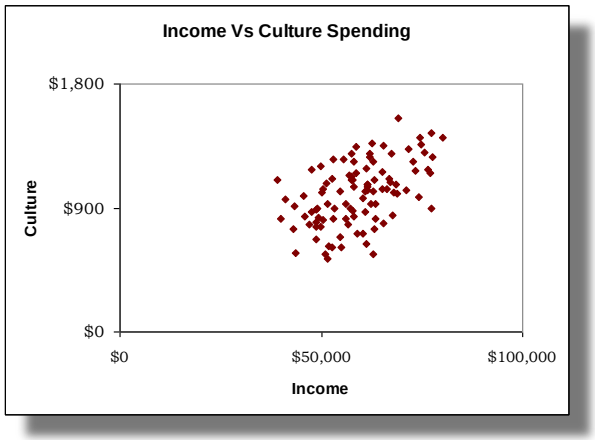
(I) The direction of the relationship? (II) The form of the relationship? (III) The strength of the relationship?

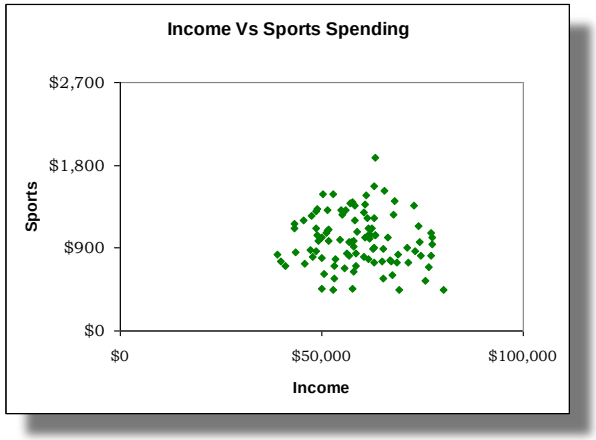
Learning Objectives

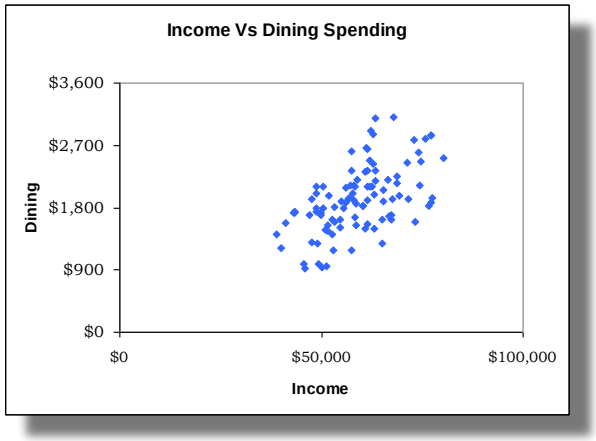
- How to provide a simple characterization of the relationship between variables?
- How to estimate a linear fit?
- How to interpret the slope and the intercept of the fitted line?
- How to quantify the goodness of fit?

Visual Method for Linear Association (Ex: Household Spending)









Quantitative Method for Linear Association

Covariance

	Income	Culture	Sports	Dining	Spending
Income	91130279	1105845	-221239	2590601	3475207
Culture	1105845	52315	-33947	18938	37306
Sports	-221239	-33947	81427	36830	84310
Dining	2590601	18938	36830	236188	291955
Spending	3475207	37306	84310	291955	413572

$$Cov(X,Y) = S_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

Correlation

	Income	Culture	Sports	Dining	Spending
Income	1.00	0.51	-0.08	0.56	0.57
Culture	0.51	1.00	-0.52	0.17	0.25
Sports	-0.08	-0.52	1.00	0.27	0.46
Dining	0.56	0.17	0.27	1.00	0.93
Spending	0.57	0.25	0.46	0.93	1.00

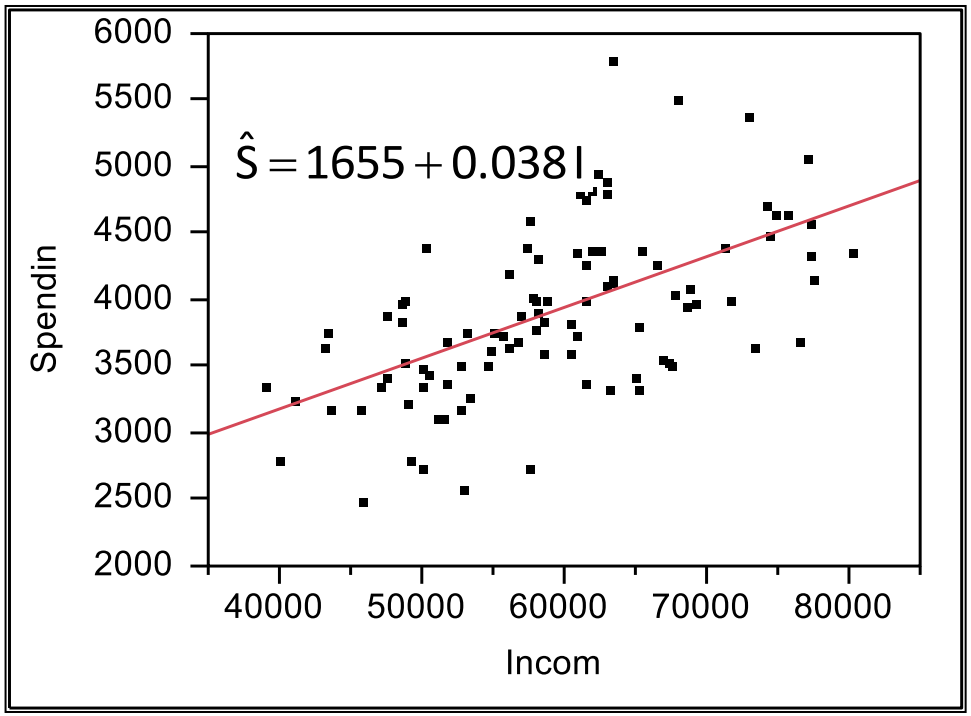
$$Correl(X,Y) = r_{XY} = \frac{Cov(X,Y)}{SD(X).SD(Y)} = \frac{S_{XY}}{S_X S_Y}$$

Covariance and Correlation

- Covariance can be used to decipher the direction of relationship between two variables but not the strength
- Random variables with zero covariance are called uncorrelated
  - All independent variables are uncorrelated
  - All uncorrelated variables are not independent
- It is is difficult to establish the strength of the relationship using covariance because it depends on the unit of measurement
- Construct $W = aX + bY$; a, b are any constants; X, Y are two random variables
  - Mean: $E[W] = aE[X] + bE[Y]$
  - Variance: $Var[W] = a^2Var[X] + b^2Var[Y] + 2abCov[X,Y]$
- The variance of the combination increases or decreases depending on the sign of the covariance term (+ or -)

- Correlation is dimensionless because it is standardized using standard deviations
  - It always takes a value between -1 (linear relationship with positive slope) and 1 (linear relationship with negative slope)
  - Close to 0 implies that there is no linear relationship
- Correlation describes the direction and strength of the relationship but cannot readily be used for "predictive" purposes
  - e.g., If the annual salary goes up by a $1000, how much do we expect the entertainment spending to change?
- Correlation is not not the same as causation; can result from "omitted/lurking" variables
  - e.g., Fire damage and number of fire fighters

Linear Fit: Beyond Correlation and Covariance



- The fitted line is denoted by

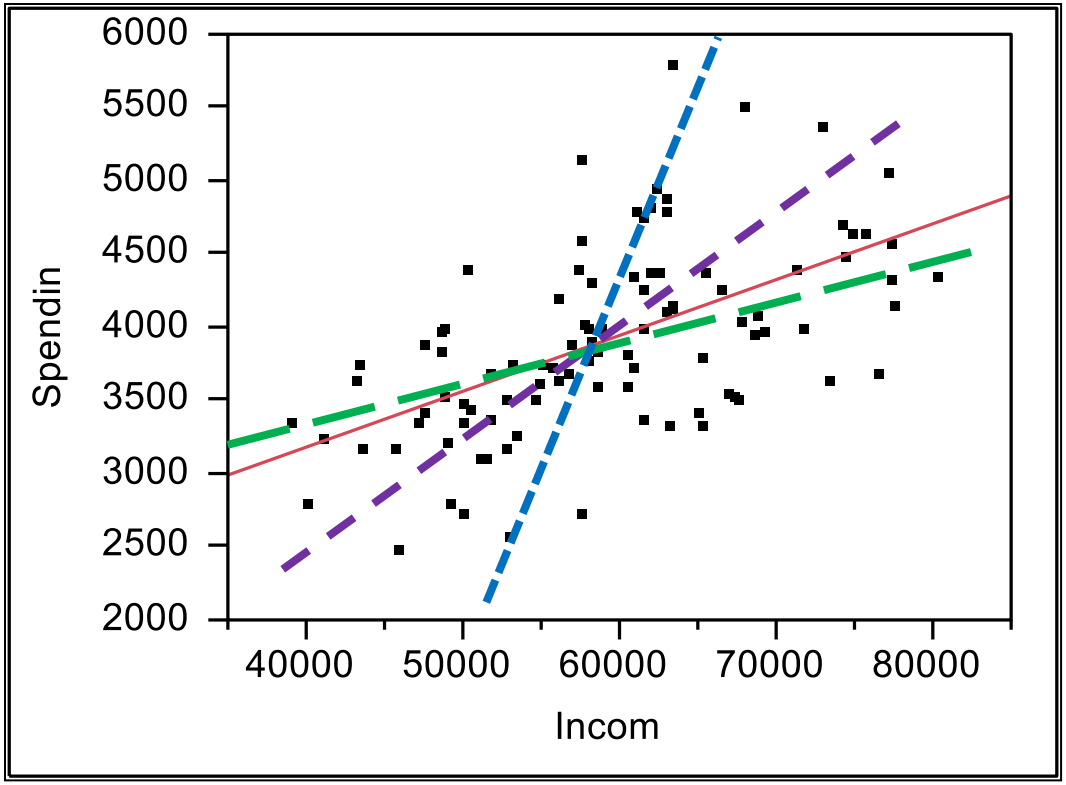
$$\hat{y} = b_0 + b_1 x$$

  - b_0 is the intercept
  - b_1 is the slope
  - $\hat{y}$ is an (point) estimate or fitted value of y for a given x value

- Interpretation
  - b_1 = 0.038
  - b_0 = \$1655

- \$1000 increase in income is associated with \$38 increase in average spending
- \$1665 is that part of the spending that is not associated with income

Which Line Should We Fit?



- The line that most closely matches the observed relationship between *Y* and *X*

Choosing the Right Line

- The error in estimation is given by:

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$

- The e_i are called the residuals

- Choose b_0 and b_1 such that they minimize the sum of squared residuals

$$\underset{b_0, b_1}{Min} \sum_i e_i^2$$

- Why should we square the residuals?

Deriving Normal Equations

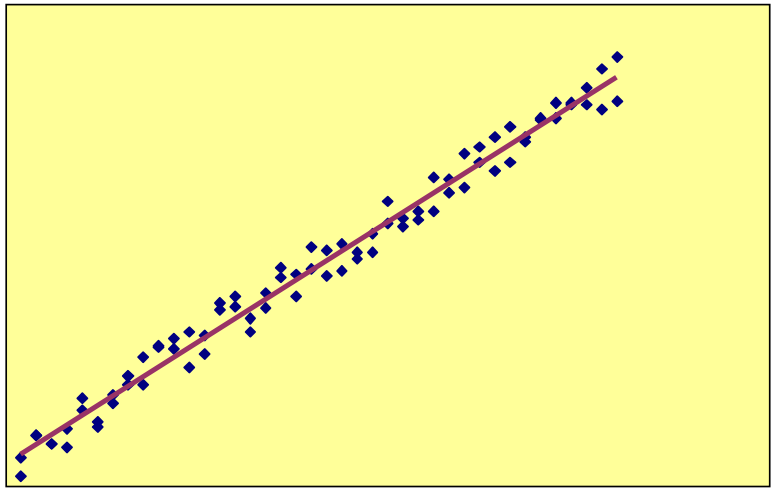
Ordinary Least Squares

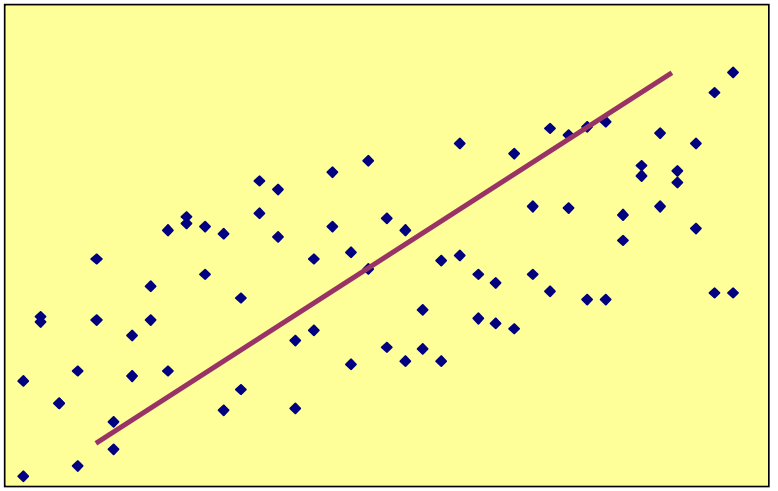
- The resulting estimators b_0 and b_1 are called the Ordinary Least Squares (OLS) estimates

$$b_1 = \frac{cov(X,Y)}{var(X)} \qquad b_0 = \bar{y} - b_1\bar{x}$$

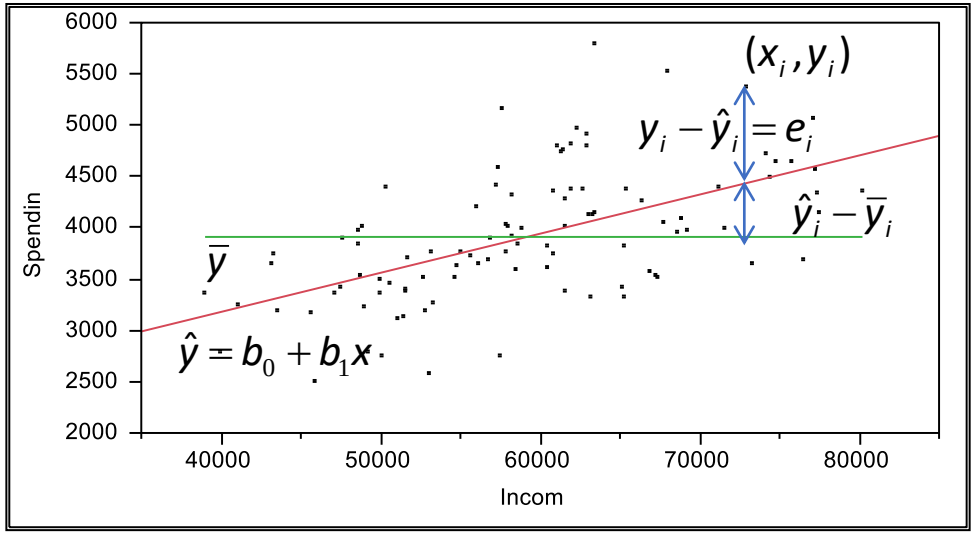
- Regression line passes through $(\bar{x}, \bar{y})$
- Cov(X,Y) determines the direction of the line
- Sum of the residuals around the best fitted line is zero $\sum_i e_i = 0$

How Good is the Model (Best Line) Fit?





How Good is the Model Fit?



$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

Sum of Squares Total (SST)

Sum of Squares Regression (SSR)

Sum of Squares Error (SSE)

$$R^2 = \frac{SSR}{SST}$$

Proportion of total variation in Y explained by X

$$RMSE = \sqrt{\frac{SSE}{(n-2)}} = \sqrt{\frac{(e_1^2 + e_2^2 + \ldots + e_n^2)}{(n-2)}}$$

Standard deviation of the residuals

Obtaining Model Fit

Summary of Fit

RSquare	0.320441
RSquare Adj	0.313507
Root Mean Square Error	532.8359
Mean of Response	3911.1
Observations (or Sum Wgts	100

$$RMSE = \sqrt{\frac{SSE}{(n-2)}}$$

$$532.83 = \sqrt{\frac{27823576}{98}}$$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	13120003	13120003	46.2112
Error	98	27823576	283914.05	Prob > F
C. Total	99	40943579		<.0001 *

$$R^2 = \frac{SSR}{SST}$$

$$0.32 = \frac{13120003}{40943579}$$

Approximately 32% of the variation in Spending is explained by variation in Income

Summary of Session

- What is regression?
  - Characterizing the relationship between two variables should start with scatter plots and calculation of correlation and covariance
- How to estimate a linear fit?
  - To estimate the linear relationship, fit a line that minimizes the sum of squared residuals (Least squares)
- How to interpret the slope and the intercept of the fitted line?
  - Slope of the estimated linear model can be interpreted as the change in the dependent variable associated with a unit change in the independent variable
  - Intercept can be interpreted as that part of the dependent variable that is not associated with change in the independent variable.
- How to quantify the goodness of fit?
  - "Goodness of Fit" of a linear model can be measured using R^2 or RMSE
    - RMSE has units of y variable
    - $R^2 = SSR/SST$ has no units and lies on a standardized scale between 0 and 1