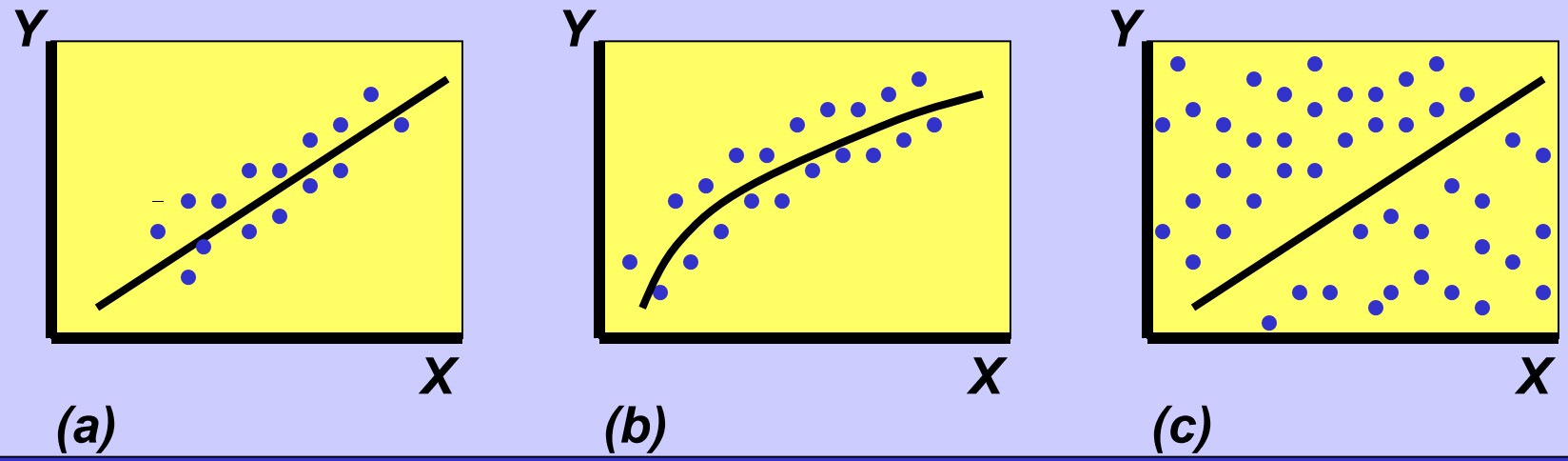
Regression Analysis

Goal: Develop a precise formula that relates Y and X

Try to predict Y given X

Example: Y_i = sales for months $i = 1, \ldots, n$

X_i = advertising expenditures for months $i = 1, \ldots, n$



- Regression is a curve fitting tool
- It also tells us how good the fit is

(a) advertise more?
(b) advertise less?
(c) stop advertising?

Basic Model

$(X_i, Y_i) i = 1, \ldots, n$ are data

X_i = deterministic and controllable

Y_i = outcome conditioned on X_i

$$Y_i = B_0 + B_1 X_i + \varepsilon_i \qquad i = 1, \ldots, n$$

$\varepsilon_i, \ldots, \varepsilon_n$ are *iid* $N(0, \sigma)$ random variables

Goal: estimate B_0, B_1 and σ from (x_i, y_i) $i = 1, \ldots, n$

Note

1) See statistics book for nonlinear transformations of X and/or *Y*

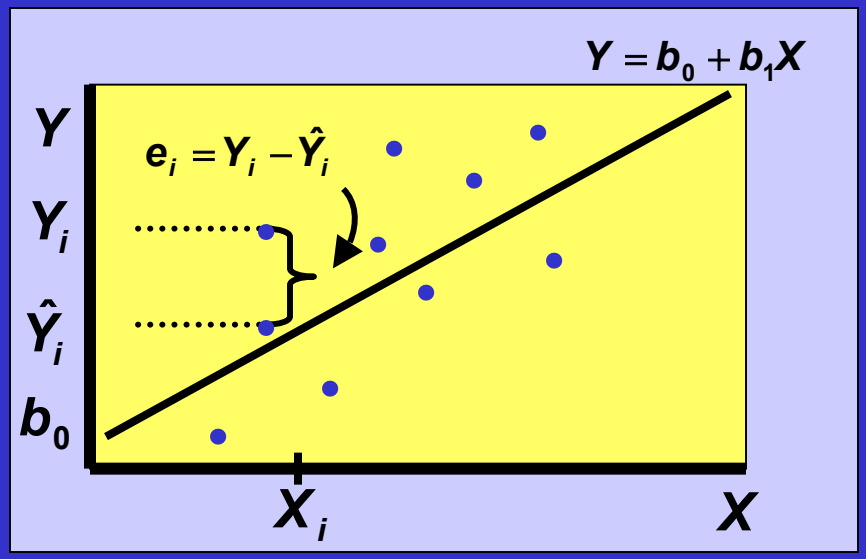
2) B_0 = baseline value of *Y*

3) B_1 = slope of line

4) $E[Y_i \mid X = x_i] = B_0 + B_1 x_i$

$VAR[Y_i \mid X = x_i] = \sigma^2$

Best Fit



How do we choose the line that best fits the data?

Suppose we choose the line $Y = b_0 + b_1X$

What is our best estimate for y_i?

$$\hat{Y}_i = b_0 + b_1X_i \qquad i = 1,\ldots,n$$

Our observed error, or residual, e_i, is

$$e_i = Y_i - \hat{Y}_i \quad (= \text{observed} - \text{predicted})$$

Regression chooses (the line) b_0 and b_1 that minimizes

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1X_i)^2$$

Multiple Regression

In Harmon Foods, many factors affect sales

$$Y = B_0 + B_1X_1 + ... + B_kX_k + \varepsilon$$

Interpreting the Results

Computer input

- Dependent variable Y
- Independent variables $X_1, ..., X_k$
- Row data $\left(Y^{(i)}, X_1^{(i)}, ..., X_K^{(i)}\right)$ $i = 1, ..., n$

Computer output

- Regression coefficients $b_0, b_1, ..., b_k$
- s_{b_0}, $s_{b_1}, ..., s_{b_k}$ = standard errors for $b_0, b_1, ..., b_k$
- R^2

Recall $\dfrac{m-\mu}{\frac{\sigma}{\sqrt{n}}} = Z$ if m = estimate, σ known

Now $\dfrac{m-\mu}{\frac{s}{\sqrt{n}}} = t_{n-k}$ m and s are estimates, σ unknown

i.e. $\dfrac{b_i - B_i}{s_{b_i}} = t_{n-k-1}$ for $i = 0,\ldots,k$

Now you can find (i) 95% confidence interval for B_i

$(ii) P(B_i > 0)$

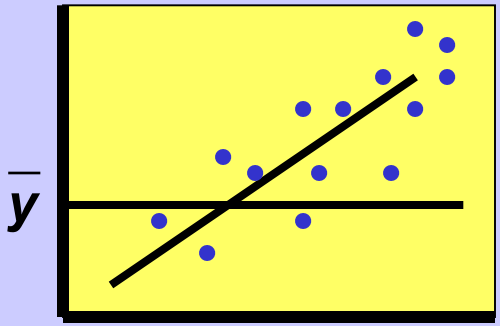
$R^2 =$ proportion of total variation that is explained by our regression line

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

if $k = 1$ (Y vs. X), then $R^2 = [CORR(X,Y)]^2$

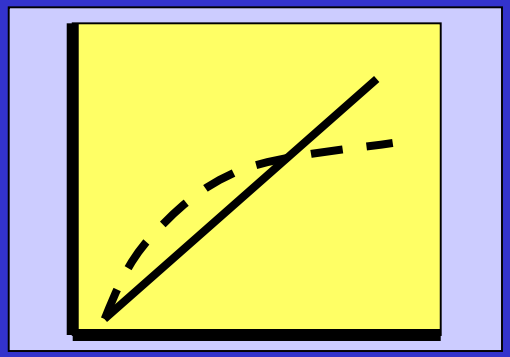
$R^2 = 1$ perfect fit

$R^2 = 0$ no relationship

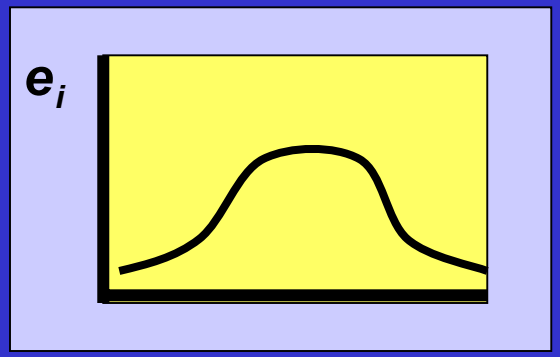


Model Validation

1) Check for nonlinearity

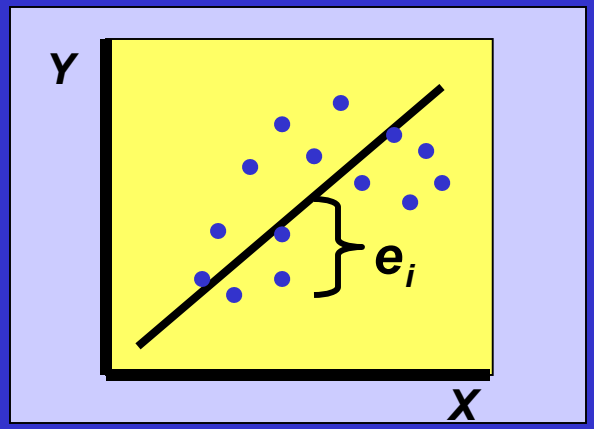


2) Are ε_i's normal? Plot $e_i = Y_i - \hat{Y}_i$



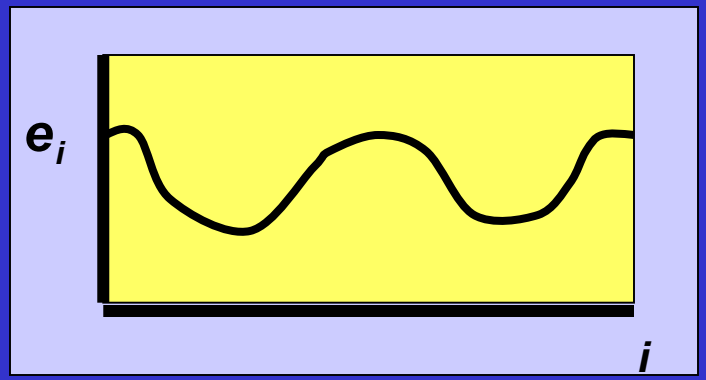
3) Heteroscedasticity:

Is $Var(\varepsilon_i)$ a function of X_i?



4) Are ε_i's independent?

Plot $e_1, e_2, ..., e_n$



Warnings

1) Too many independent variables

$$\frac{n}{k+2} \geq 10$$

2) Which independent variables to use?

- Stepwise Regression (guided by R^2)

3) Multicollinearity

- 2 or more X_i's are highly correlated
- ok if you estimate Y
- can give incorrect b_i's

4) Association vs. Causation

5) Overstating importance of b_i