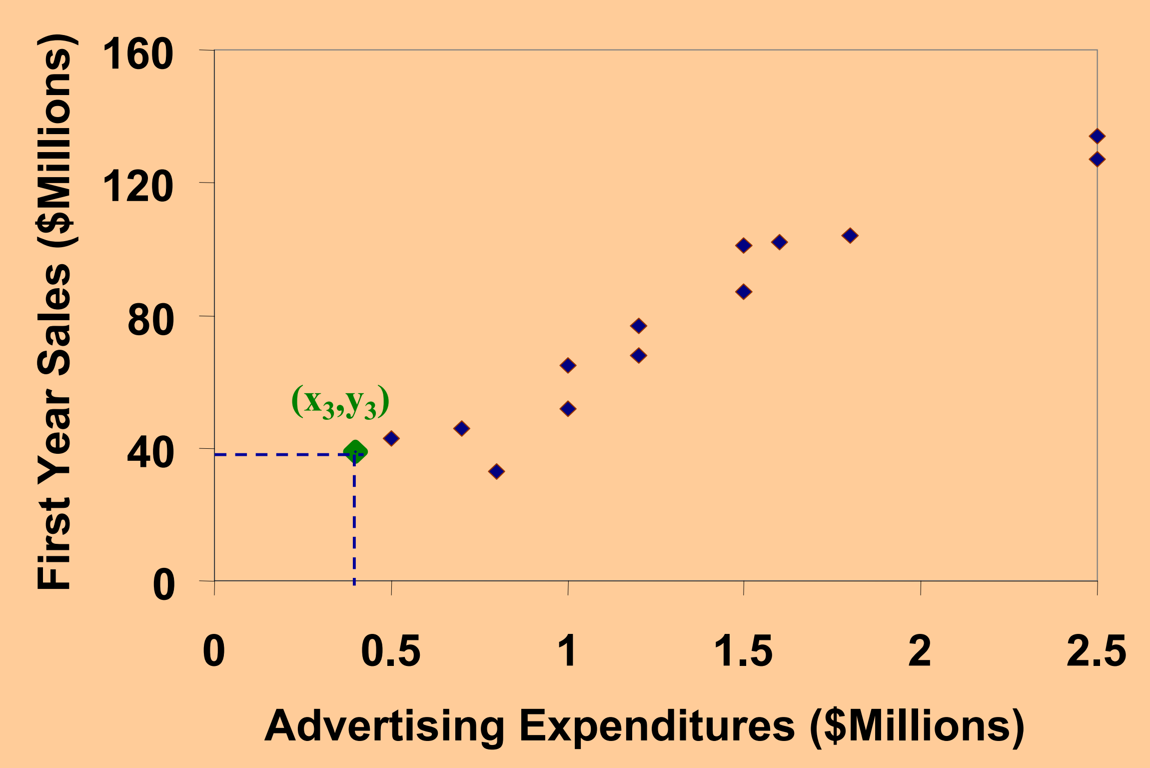
Regression Models

Summer 2003

Does Advertising Increase Sales?

Appleglo	First-Year Advertising Expenditures ($ millions)	First-Year Sales ($ millions)
Region	x	y
Maine	1.8	104
New Hampshire	1.2	68
Vermont	0.4	39
Massachusetts	0.5	43
Connecticut	2.5	127
Rhode Island	2.5	134
New York	1.5	87
New Jersey	1.2	77
Pennsylvania	1.6	102
Delaware	1.0	65
Maryland	1.5	101
West Virginia	0.7	46
Virginia	1.0	52
Ohio	0.8	33



Questions: i) How to relate advertising expenditure to sales?

ii) What is expected first-year sales if advertising expenditure is $2.2 million?

iii) How confident are you in your estimate?

Regression Analysis

GOAL: Develop a formula that relates two quantities

x: "independent" (also called "explanatory") variable
quantity typically under managerial control

Y: "dependent" variable
magnitude is determined (to some degree) by value of x
quantity to be predicted

Examples:

Y (dependent variable)	X (independent variable)
College GPA	SAT score
Lung cancer rate	Amount of cigarette smoking
Stock return	Spending in R&D
First-year sales	Advertising expenditures

Outline

- Simple Linear Regression
- Multiple Regression
- Understanding Regression Output
- Coefficient of Determination R^2
- Validating the Regression Model

The Basic Model: Simple Linear Regression

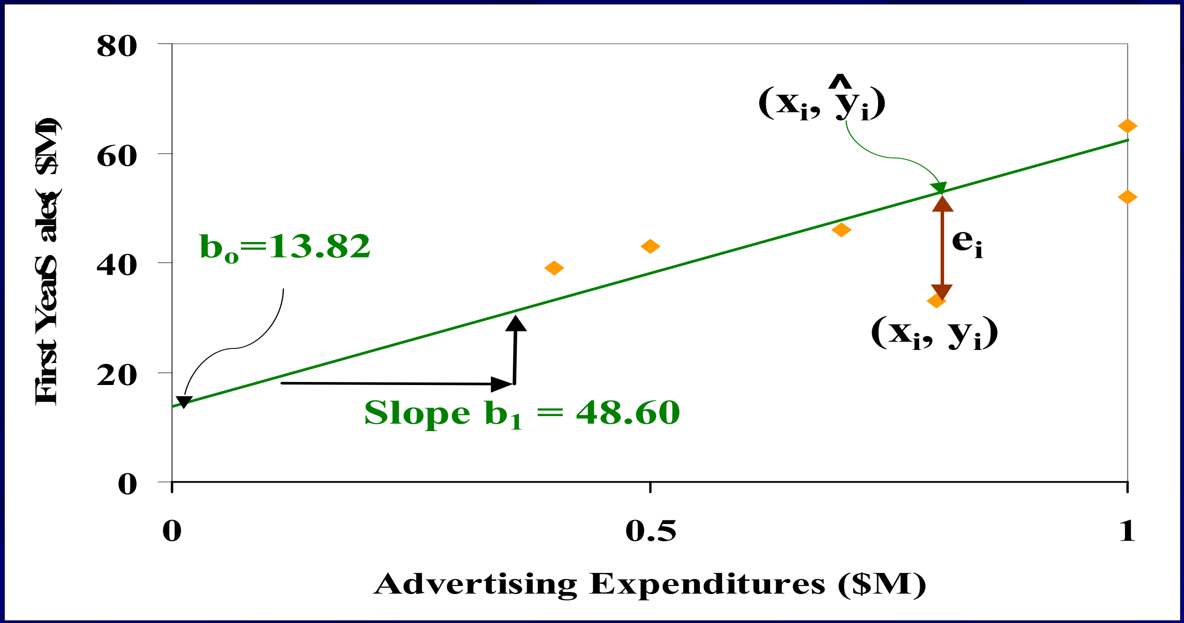
Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ (a sample of size n taken from the population of all (X,Y) values)

Model of the population*: $$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

<u>Comments:</u>

- The model assumes a *linear* relationship between x and Y, with *y intercept* β_0 and *slope* β_1
- β_0 and β_1 are the *parameters for the whole population*. We do not know them and will *estimate* them using b_0 and b_1 to be calculated from the data (i.e. from the sample of size n)
- ε_i is the called the *error term*. Since the Y's do not fall precisely on the line (i.e. they are r.v.'s) we need to add an error term to obtain an equality.
- ε_i is N(0, σ). Thus, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are i.i.d. Normally distributed r.v.'s.
- $E(Y_i \mid x_i) = \beta_0 + \beta_1 x_i$ Is the *expected value of Y for a given x value*. It is just the value on the line as that is where on average the Y_i value would fall for a given x_i value.
- $SD(Y_i \mid x_i) = \sigma$ Notice that The SD of Y_i is equal to the SD of ε_i and is a constant independent of the value of x.

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



Best choices:
b_o = 13.82
b_1 = 48.60

Regression coefficients: b_0 and b_1 are *estimates* of β_0 and β_1

Regression estimate for Y at x_i : $\hat{y}_i = b_0 + b_1 x_i$ (prediction)

Value of Y at x_i : $y_i = b_0 + b_1 x_i + e_i$ (use the error to obtain equality)

Residual (error): $e_i = y_i - \hat{y}_i$

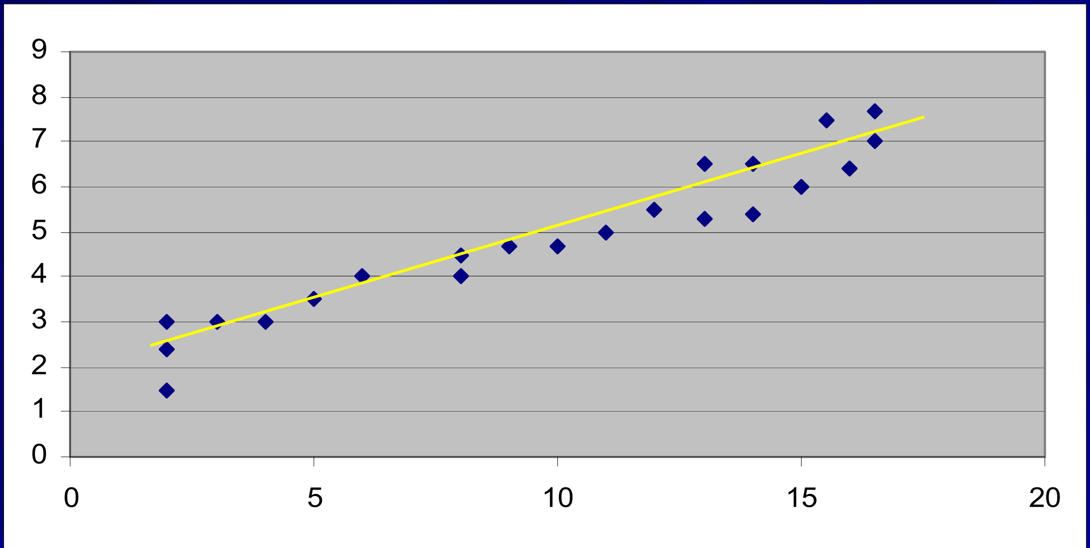
The "best" regression line is the one that chooses b_0 and b_1 to *minimize the total squared errors*:

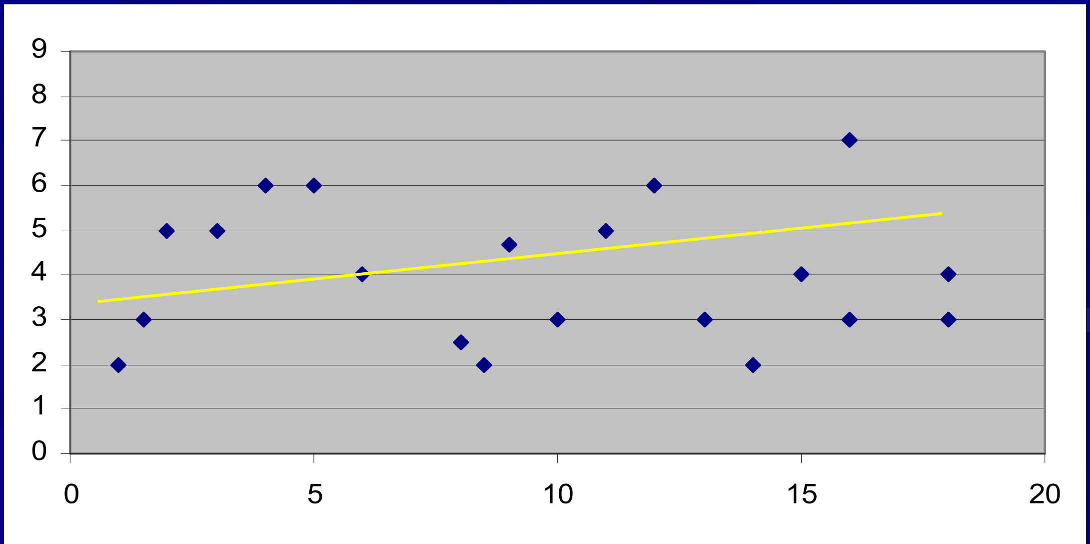
$$SSR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

SSR is the residual sum of squares, analogous to a variance calculation

How Good a Fit to the Line?

- std error s estimates σ, the std deviation of error ε_i
- lower figure has 10 times the error





Coefficient of Determination: R^2

- It is a measure of the *overall quality* of the regression.
  Specifically, it is the percentage of total variation exhibited in the y_i data that is accounted for or predicted by the sample regression line.

- The sample mean of Y: $\bar{y} = (y_1 + y_2 + \ldots + y_n)/\, n$

- Total variation in Y $= \sum_{i=1}^{n} (y_i - \bar{y})^2$

- Residual (unaccounted) variation in Y $= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

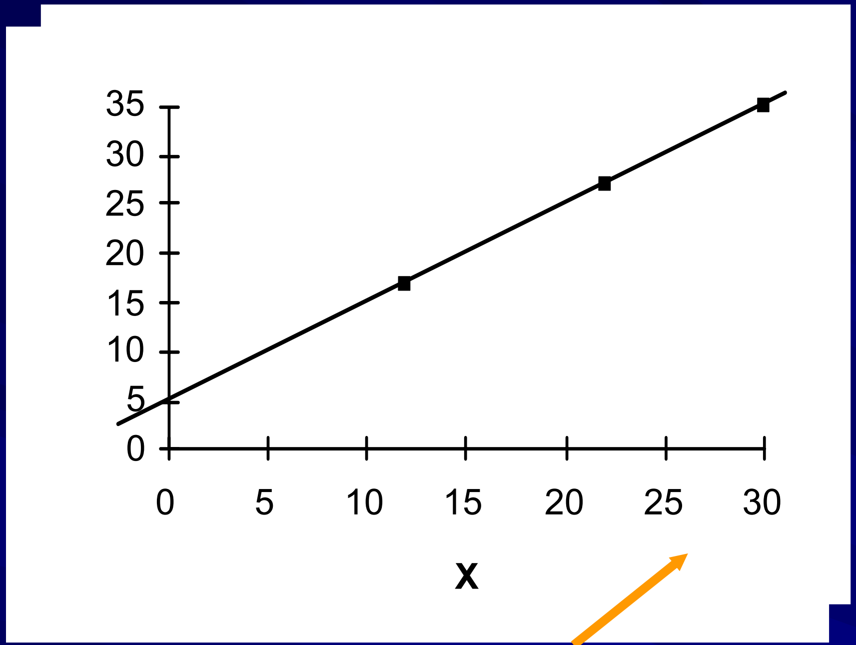
(even the linear model, $\hat{y}_i$, does not explain all the the variability in y_i)

$$R^2 = \frac{\text{variation accounted for by x variables}}{\text{total variation}}$$

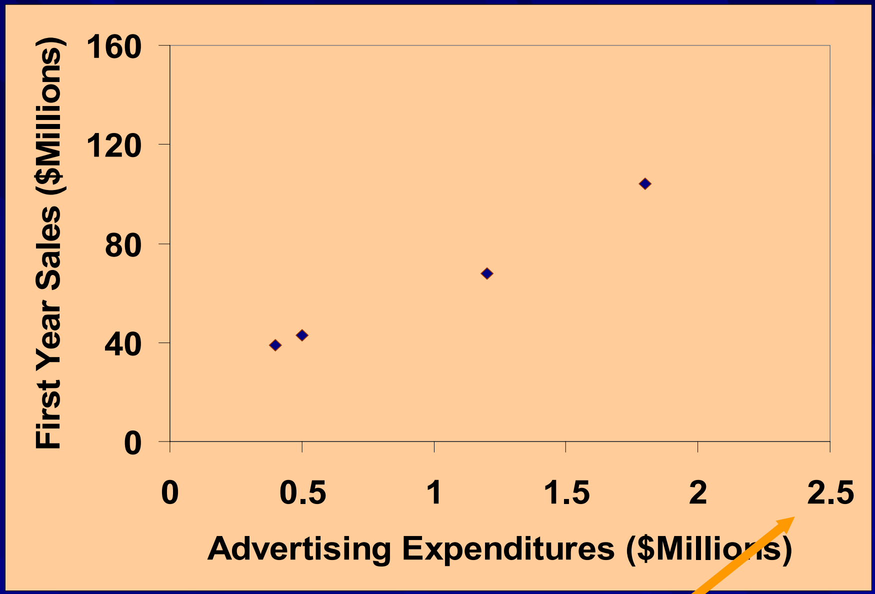
$$= 1 - \frac{\text{variation not accounted for by x variables}}{\text{total variation}}$$

$$\boxed{= 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

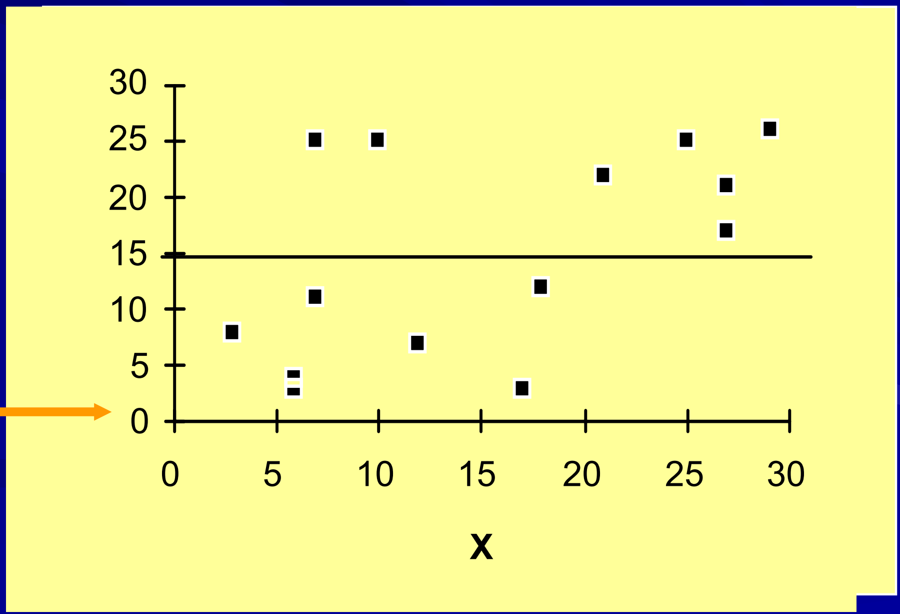
R^2 takes values between 0 and 1 (it is a percentage).



R^2 = 1; x values account for all variation in the Y values



R^2 = 0.833 in our Appleglo Example



R^2 = 0; x values account for no variation in the Y values

Correlation and Regression

- Simple regression is correlation in disguise
- Coefficient of Determination = squared correlation coefficient
- Regression coefficient: b_1 = correlation * s_y/s_x
- Appleglo: Sales = 13.82 + 48.60 * Advertising
- The coefficients are in units of sales and advertising. If advertising is $2.2 Million, then sales will be 13.82 + 48.60 * 2.2 = $120.74 M
- What if there are >1 predictor variable?

Sales of Nature-Bar ($ million)

region	$\underline{Y}$ sales	$\underline{x_1}$ advertising	$\underline{x_2}$ promotions	$\underline{x_3}$ competitor's sales
Selkirk	101.8	1.3	0.2	20.40
Susquehanna	44.4	0.7	0.2	30.50
Kittery	108.3	1.4	0.3	24.60
Acton	85.1	0.5	0.4	19.60
Finger Lakes	77.1	0.5	0.6	25.50
Berkshire	158.7	1.9	0.4	21.70
Central	180.4	1.2	1.0	6.80
Providence	64.2	0.4	0.4	12.60
Nashua	74.6	0.6	0.5	31.30
Dunster	143.4	1.3	0.6	18.60
Endicott	120.6	1.6	0.8	19.90
Five-Towns	69.7	1.0	0.3	25.60
Waldeboro	67.8	0.8	0.2	27.40
Jackson	106.7	0.6	0.5	24.30
Stowe	119.6	1.1	0.3	13.70

Multiple Regression

- In general, there are many factors in addition to advertising expenditures that affect sales
- Multiple regression allows more than one independent variable

Independent variables: $x_1, x_2, \ . \ . \ . \ , \ x_k$ (k of them)

Data: $(y_1, x_{11}, x_{21}, \ . \ . \ . \ , \ x_{k1}), \ . \ . \ . \ , (y_{n1}, x_{n1}, x_{n2}, \ . \ . \ . \ , \ x_{kn}),$

Population Model: $Y_i = \beta_0 + \beta_1 x_{1i} + \ . \ . \ . \ + \beta_k x_{ki} + \varepsilon_i$

$\varepsilon_1, \varepsilon_2, \ . \ . \ . \ , \varepsilon_n$ are i.i.d random variables, $\sim N(0, \sigma)$

Regression coefficients: $b_0, b_1, \ldots, b_k$ are estimates of $\beta_0, \beta_1, \ldots, \beta_k$.

Regression Estimate of y_i : $\hat{y}_i = b_0 + b_1 x_{1i} + \ . \ . \ . \ + b_k x_{ki}$

Goal: Choose $b_0, b_1, \ldots , b_k$ to minimize the residual sum of squares. i.e., minimize:

$$SSR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Regression Output (from Excel)

Regression Statistics	
Multiple R	0.913
R Square	0.833
Adjusted R Square	0.787
Standard Error	17.600
Observations	15

Standard error s: an estimate of σ

s^2 estimate of variance

Analysis of Variance					
	df	*Sum of Squares*	*Mean Square*	*F*	*Significance F*
Regression	3	16997.537	5665.85	18.290	0.000
Residual	11	3407.473	309.77		
Total	14	20405.009			

	Coefficients	*Standard Error*	*t Statistic*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	65.71	27.73	2.37	0.033	4.67	126.74
Advertising	48.98	10.66	4.60	0.000	25.52	72.44
Promotions	59.65	23.63	2.53	0.024	7.66	111.65
Competitor’s Sales	-1.84	0.81	-2.26	0.040	-3.63	-0.047

Understanding Regression Output

1) Regression coefficients: $b_0, b_1, \ldots, b_k$ are *estimates* of $\beta_0, \beta_1, \ldots, \beta_k$ based on sample data. Fact: $E[b_j] = \beta_j$.

(i.e., if we run the multiple regression many many times, the average value of the b_j's we get is β_j)

Example:

b_0 = 65.705 (its interpretation is context dependent, in this case, sales if no advertising, no promotions, and no competition)

b_1 = 48.979 (an additional $1 million in advertising is expected to result in an additional $49 million in sales)

b_2 = 59.654 (an additional $1 million in promotions is expected to result in an additional $60 million in sales)

b_3 = -1.838 (an increase of $1 million in competitor sales is expected to decrease sales by $1.8 million)

Understanding Regression Output, Continued

2) Standard error s: an estimate of σ, the SD of each ε_i.
   It is a measure of the amount of "noise" in the model.
   Example: $s = 17.60$

3) Degrees of freedom: to be explained later.

4) Standard errors of the coefficients: s_{b0} , s_{b1} , . . . , s_{bk}
   They are just the standard deviations of the estimates b_0 , b_1, . . . , b_k.

   They are useful in assessing the quality of the coefficient estimates and validating the model. (Explained later).

Coefficient of Determination: R^2

- A high R^2 means that most of the variation we observe in the y_i data can be attributed to their corresponding x values —– a desired property.
- In multiple regression, R is called "Multiple R"
- In simple regression, the R^2 is higher if the data points are better aligned along a line. The corresponding picture in multiple regression is a plot of predicted y_i vs. the actual y_i data.
- How high a R^2 is "good" enough depends on the situation (for example, the intended use of the regression, and complexity of the problem).
- Users of regression tend to be fixated on R^2, but it's not the whole story. It is important that the regression model is "valid."

Caution about R^2

- One should not include x variables unrelated to Y in the model, just to make the R^2 fictitiously high. New x variables will account for some additional variance by chance alone ("fishing"), but these would not be validated in new samples.
- Adjusted R^2 modifies R^2 to account for the number of variables and the sample size, therefore counteracting "fishing":

$$\text{Adjusted } R^2 = 1 - \frac{(n-1)}{[n-(k+1)]}(1-R^2)$$

Rule of thumb: $n >= 5(k+2)$ where n = sample size and k = number of predictor variables

Validating the Regression Model

Assumptions about the population:

$Y_i = b_0 + b_1 x_{1i} + \ldots + b_k x_{ki} + \varepsilon_i \quad (i = 1, \ldots, n)$

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are i.i.d random variables, ~ $N(0, \sigma)$

1) *Linearity*

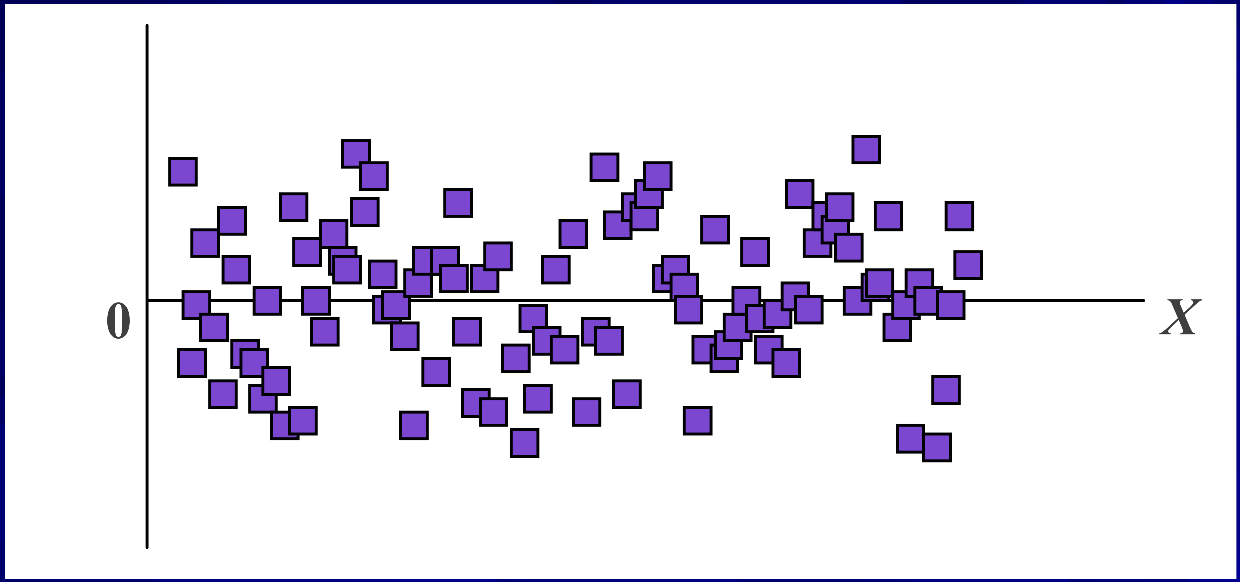
- If k = 1 (simple regression), one can check visually from scatter plot.
- "Sanity check": the sign of the coefficients, reason for non-linearity?

2) *Normality of ε_i*

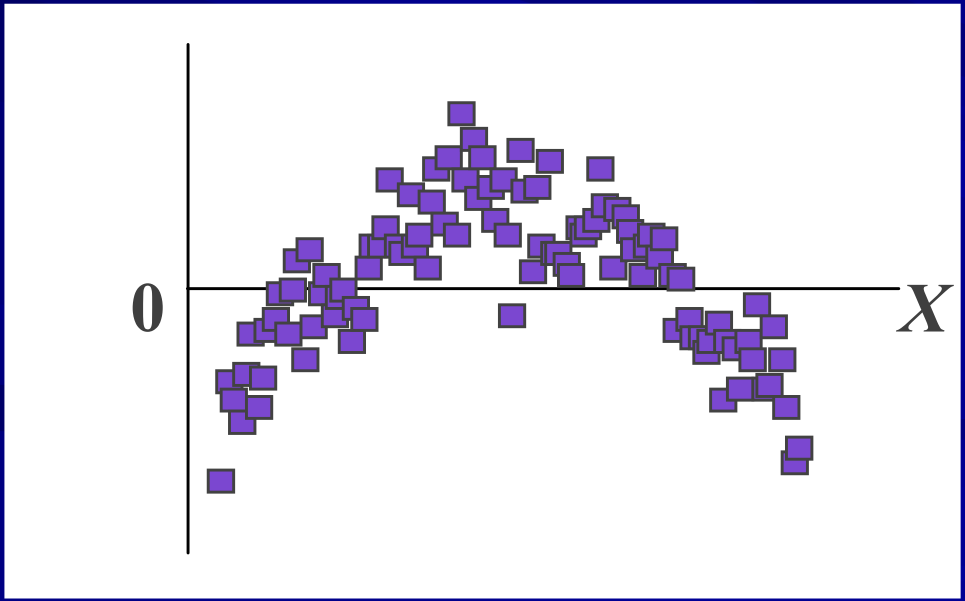
- Plot the residuals ($e_i = y_i - \hat{y}_i$).
- They should look evenly random – i.e. scattered.
- Then plot a histogram of the residuals. The resulting distribution should be approximately normal.

Usually, results are fairly robust with respect to this assumption.

Residual Plots



Healthy

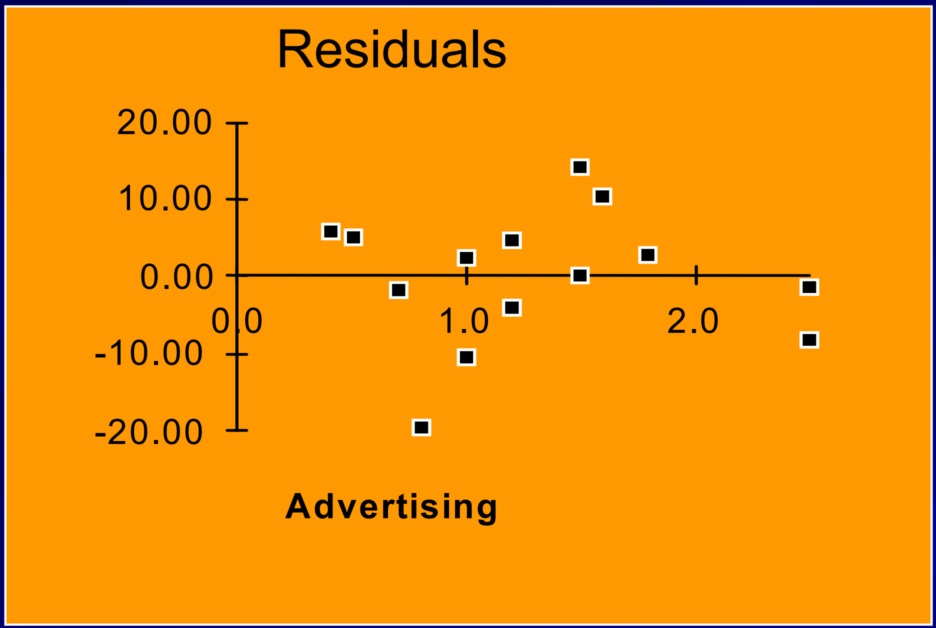


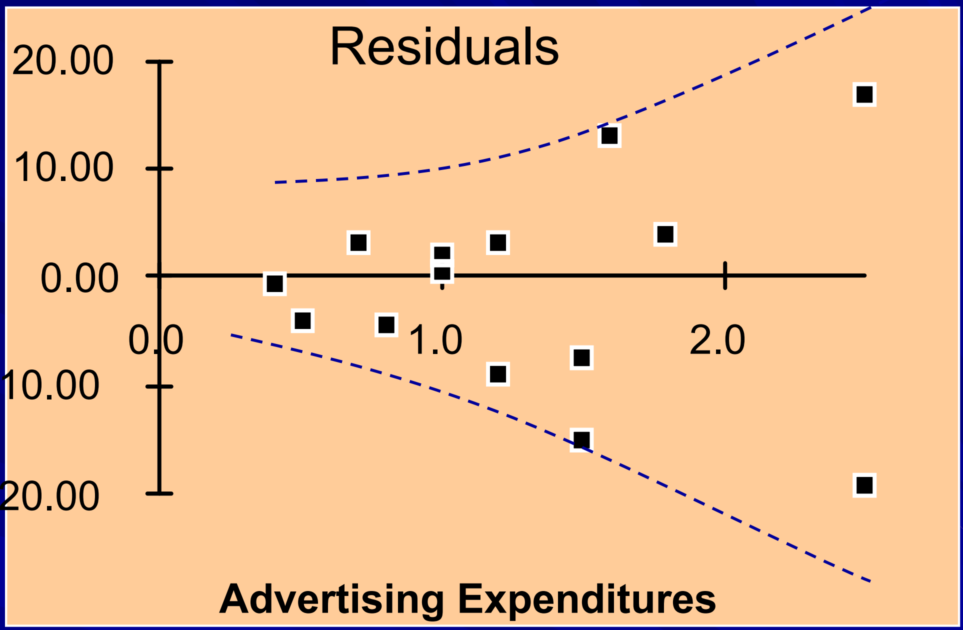
Nonlinear

Can sometimes be fixed, e.g., Insert x^2 as a variable.

3) *Heteroscedasticity*

- Do error terms have constant Std. Dev.? (i.e., $SD(\varepsilon_i) = \sigma$ for all i?)
- Check scatter plot of residuals vs. Y and x variables.





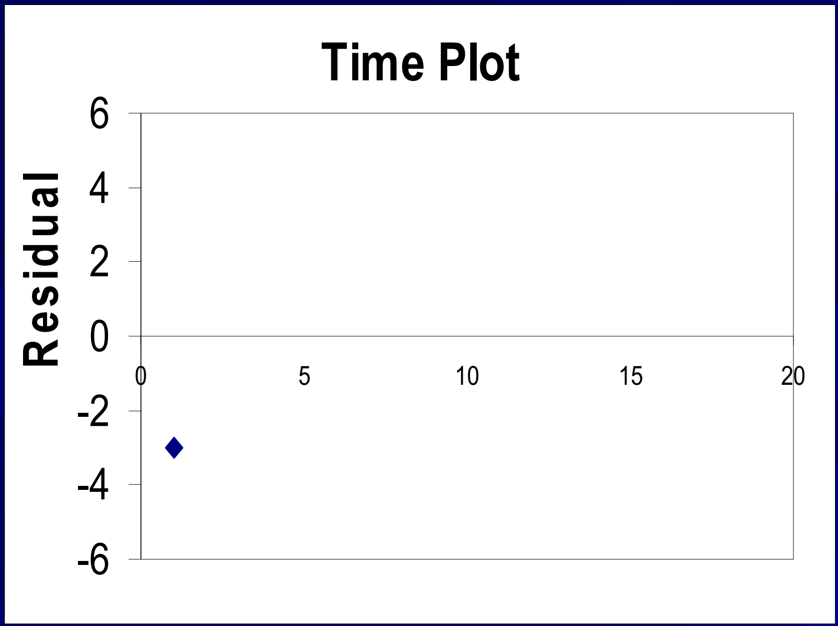
No evidence of heteroscedasticity

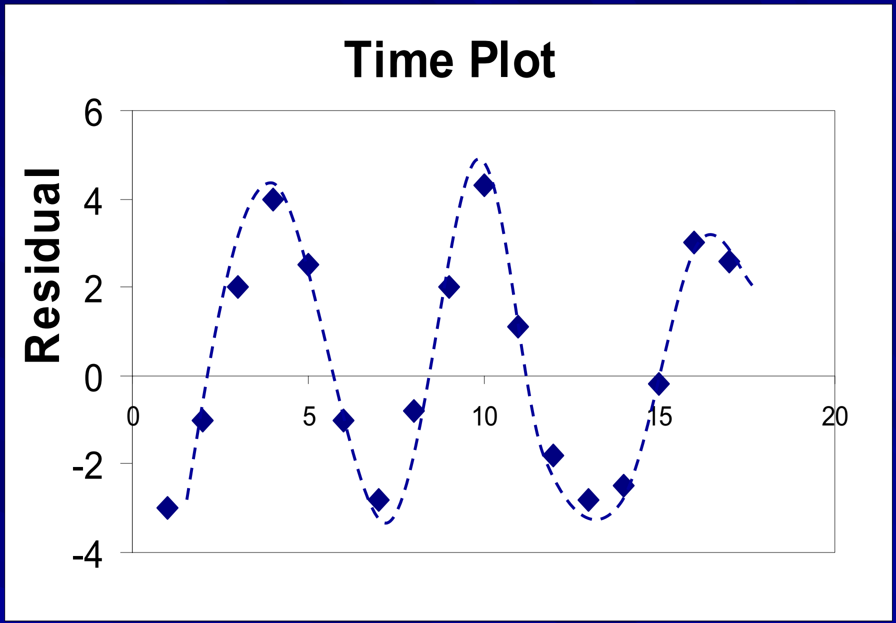
Evidence of heteroscedasticity

- May be fixed by introducing a transformation (e.g. use x^2 instead of x)
- May be fixed by introducing or eliminating some independent variables

4) *Autocorrelation: Are error terms independent?*

- Plot residuals in order and check for patterns





No evidence of autocorrelation

Evidence of autocorrelation

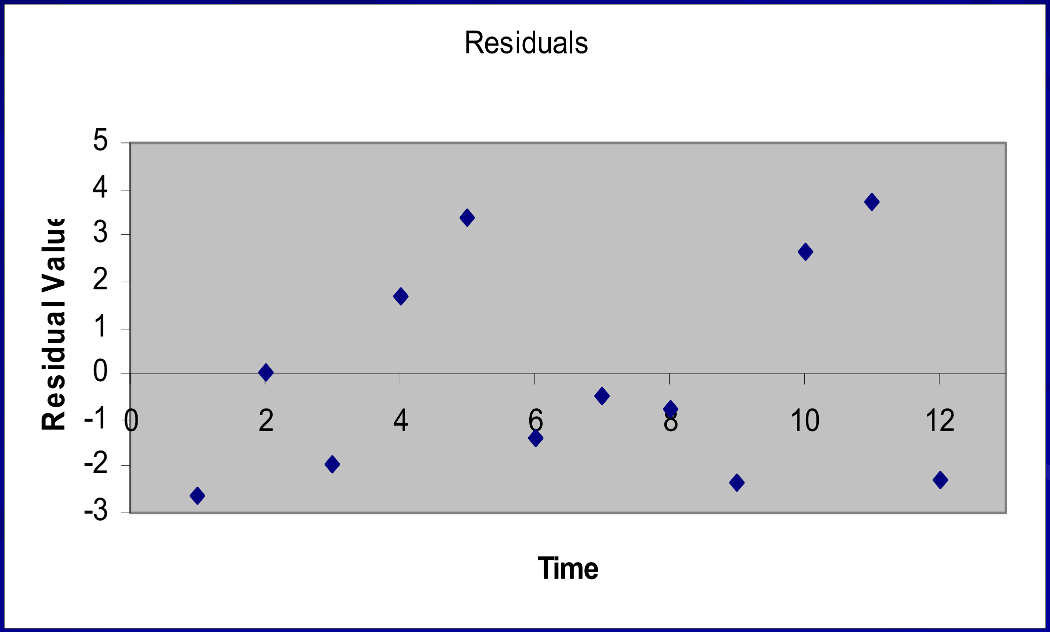
- Autocorrelation may be present if observations have a natural sequential order (for example, time).

- May be fixed by introducing a variable (frequently time) or transforming a variable.

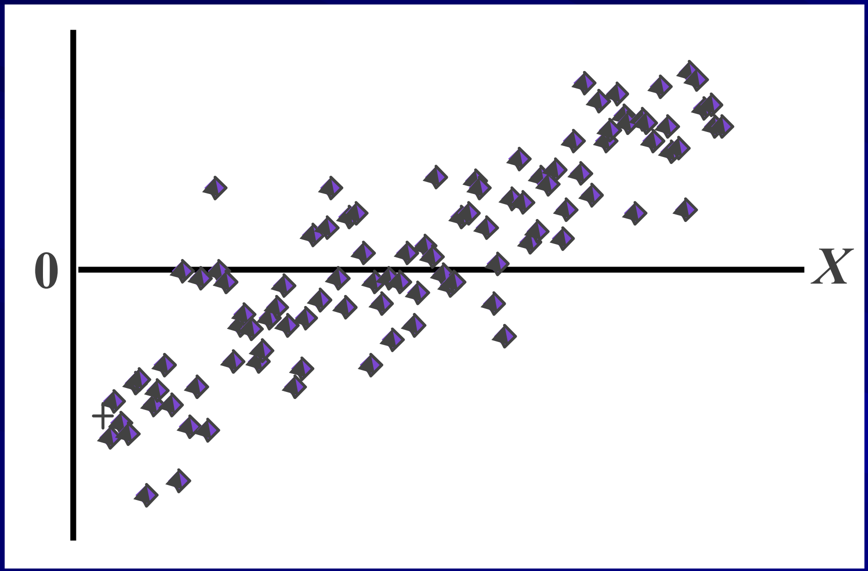
Validating the Regression Model: Autocorrelation

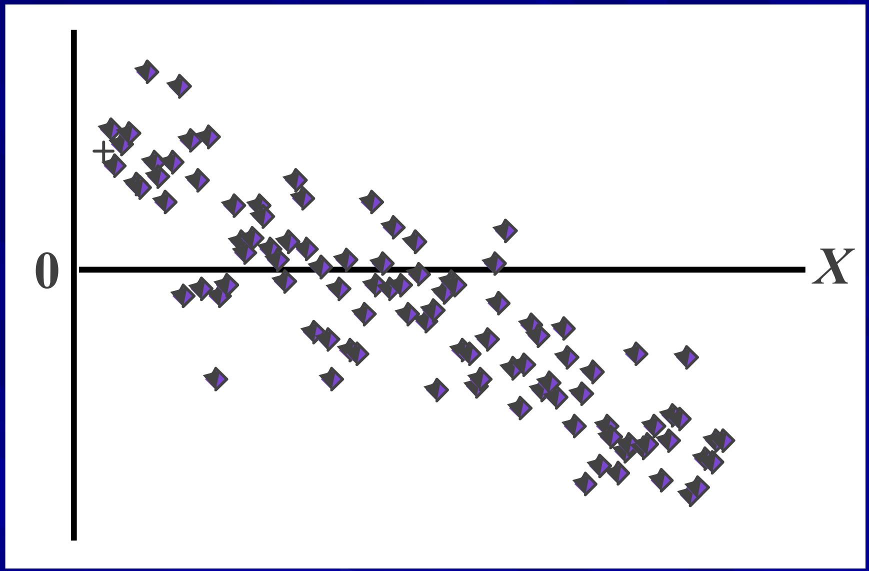
Sales ($ Thousands)	Promotions ($ Thousands)	Month
63.00	26	January
65.25	25	February
69.18	38.5	March
74.34	42	April
68.62	25.1	May
63.71	24.7	June
64.41	24.3	July
64.06	24.1	August
70.36	42.1	September
75.71	43	October
67.61	22	November
62.93	25	December

- Evidence of Autocorrelation in Simple Regression in Toothpaste monthly sales and promotions



Graphs of Non-independent Error Terms (Autocorrelation)





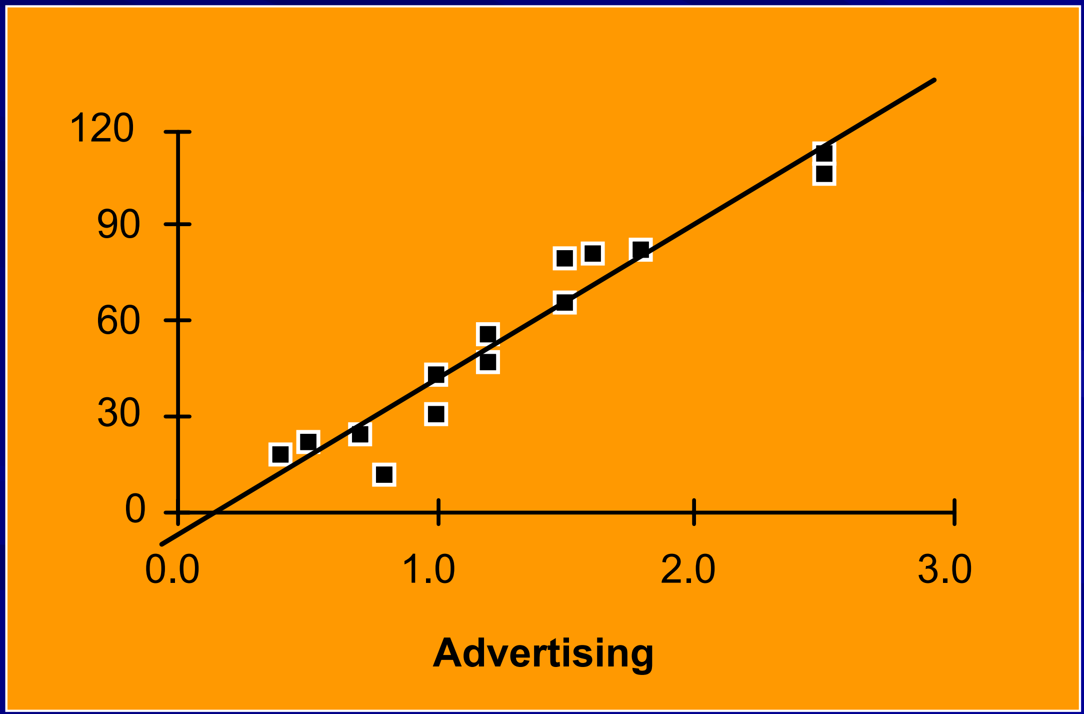
Possible solution: Insert time (sequence) of observation as a variable.

Pitfalls and Issues

1) *Overspecification*

- Including too many x variables to make R^2 fictitiously high.
- Rule of thumb: we should maintain that $n >= 5(k+2)$

2) *Extrapolating beyond the range of data (Carter Racing!!)*



Pitfalls and Issues

3) Multicollinearity

- Occurs when two of the x variable are strongly correlated.
- Can give very wrong estimates for β_i's.
- Tell-tale signs:
  - Regression coefficients (b_i's) have the "wrong" sign.
  - Addition/deletion of an independent variable results in large changes of regression coefficients
  - Regression coefficients (b_i's) not significantly different from 0
- May be fixed by deleting one or more independent variables

Can We Predict Graduate GPA from College GPA and GMAT?

Student Number	Graduate GPA	College GPA	GMAT
1	4.0	3.9	640
2	4.0	3.9	644
3	3.1	3.1	557
4	3.1	3.2	550
5	3.0	3.0	547
6	3.5	3.5	589
7	3.1	3.0	533
8	3.5	3.5	600
9	3.1	3.2	630
10	3.2	3.2	548
11	3.8	3.7	600
12	4.1	3.9	633
13	2.9	3.0	546
14	3.7	3.7	602
15	3.8	3.8	614
16	3.9	3.9	644
17	3.6	3.7	634
18	3.1	3.0	572
19	3.3	3.2	570
20	4.0	3.9	656
21	3.1	3.1	574
22	3.7	3.7	636
23	3.7	3.7	635
24	3.9	4.0	654
25	3.8	3.8	633

Regression Output

R Square	**0.96**	
Standard Error	**0.08**	
Observations	**25**	
	Coefficients	**Standard Error**
Intercept	**0.09540**	**0.28451**
College GPA	**1.12870**	**0.10233**
GMAT	**-0.00088**	**0.00092**

What happened?

College GPA and GMAT are highly correlated!

	Graduate	College	GMAT
Graduate	1		
College	0.98	1	
GMAT	0.86	0.90	1

Eliminate GMAT(HBS?)

R Square	0.958	
Standard Error	0.08	
Observations	25	
	Coefficients	Standard Error
Intercept	-0.1287	0.1604
College GPA	1.0413	0.0455

Checklist for Evaluating a Linear Regression Model

- Linearity: scatter plot, common sense, and knowing your problem.
- Signs of Regression Coefficients: do they agree with intuition?
- Normality: plot residual histogram
- R^2: is it reasonably high in the context?
- Heteroscedasticity: plot residuals against each x variable
- Autocorrelation: time series plot
- Multicollinearity: compute correlations between x variables
- Statistical test: are the coefficients significantly different from zero? (next time)

Summary and Look Ahead

- Regression is a way to make predictions from one or more predictor variables
- There are a lot of assumptions that must be checked to make sure the regression model is valid
- We may not get to Croq'Pain