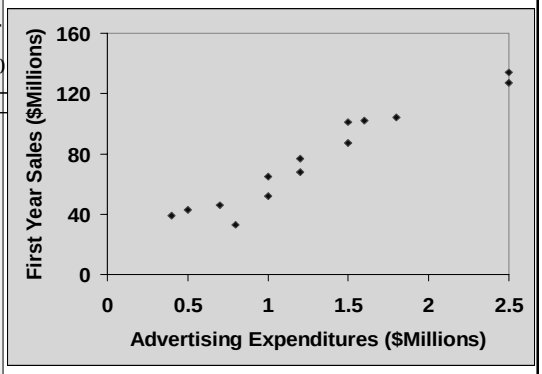
Multiple Linear Regression Review

Outline

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

Linear Regression: An Example

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:
a) How to relate advertising expenditure to sales?
b) What is expected first-year sales if advertising expenditure is $2.2 million?
c) How confident is your estimate? How good is the "fit?"

The Basic Model: Simple Linear Regression

Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

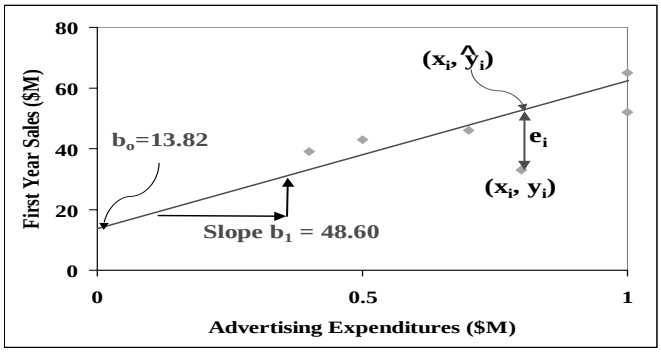
Model of the population: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are i.i.d. random variables, $N(0, \sigma)$
This is the true relation between Y and x, but we do not know β_0 and β_1 and have to estimate them based on the data.

Comments:

- $E(Y_i \mid x_i) = \beta_0 + \beta_1 x_i$
- $SD(Y_i \mid x_i) = \sigma$
- Relationship is linear – described by a "line"
- β_0 = "baseline" value of Y (i.e., value of Y if x is 0)
- β_1 = "slope" of line (average change in Y per unit change in 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)

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 errors (residual sum of squares):

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

Example: Sales of Nature-Bar ($ million)

region	sales	advertising	promotions	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 x variables.

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

Data: $(y_1, x_{11}, x_{21}, \ldots, x_{k1}), \ldots, (y_n, x_{1n}, x_{2n}, \ldots, x_{kn})$,

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

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are iid random variables, ~ N(0, σ)

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

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

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

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: $E[b_j] = \beta_j$.

Example:

b_0 = 65.705 (its interpretation is context dependent .

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 errors: 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

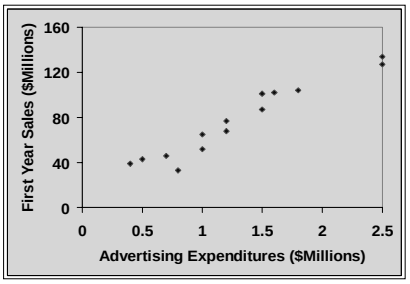
3) Degrees of freedom: #cases - #parameters, relates to over-fitting phenomenon

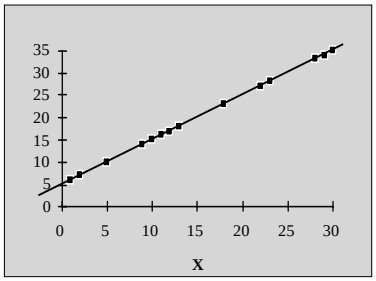
4) Standard errors of the coefficients: $s_{b0}, s_{b1}, \ldots, s_{bk}$
They are just the standard deviations of the estimates $b_0, b_1, \ldots, b_k$.

They are useful in assessing the quality of the coefficient estimates and validating the model.

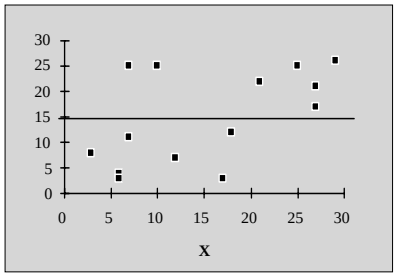
R^2 takes values between 0 and 1 (it is a percentage).

R^2 = 0.833 in our Appleglo Example





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



R^2 = 0; x values account for none variation in the Y values

Understanding Regression Output, Continued

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

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

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

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

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 simple regression, the R^2 is higher if the data points are better aligned along a line. But outliers – Anscombe example.
- 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."

Coefficient of Determination: R^2

- One should not include x variables unrelated to Y in the model, just to make the R^2 fictitiously high. (With more x variables there will be more freedom in choosing the b_i's to make the residual variation closer to 0).
- Multiple R is just the square root of R^2.

Validating the Regression Model

Assumptions about the population:

$Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i \quad (i = 1, \ldots, n)$

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are iid random variables, ~ N(0, σ)

1) Linearity

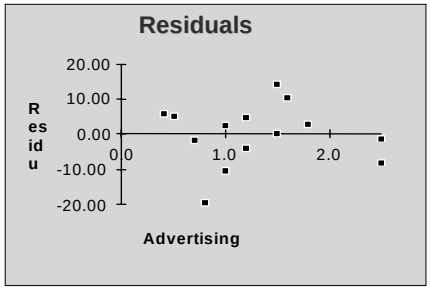
- If k = 1 (simple regression), one can check visually from scatter plot.
- "Sanity check:" the sign of the coefficients, reason for non-linearity?

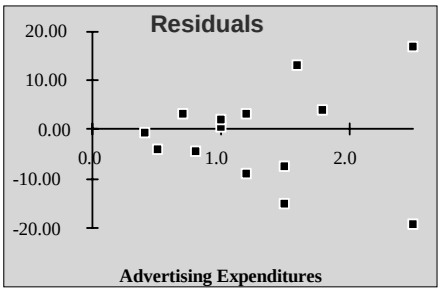
2) Normality of ε_i

- Plot a histogram of the residuals ($e_i = y_i - \hat{y}_i$).
- Usually, results are fairly robust with respect to this assumption.

3) Heteroscedasticity

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





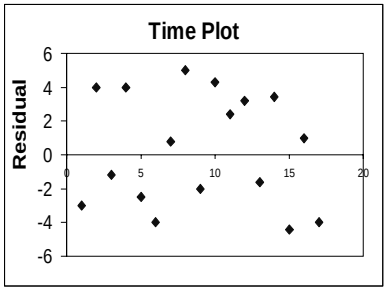
No evidence of heteroscedasticity

Evidence of heteroscedasticity

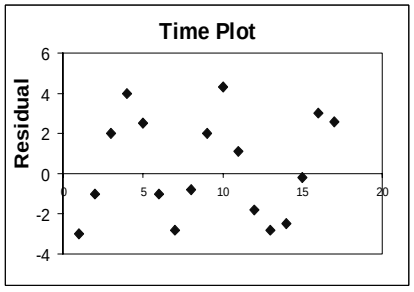
- May be fixed by introducing a transformation
- May be fixed by introducing or eliminating some independent variables

4) **Autocorrelation : Are error terms independent?**

Plot residuals in order and check for patterns



No evidence of autocorrelation



Evidence of autocorrelation

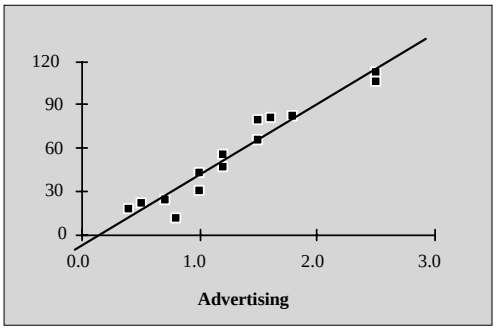
- **Autocorrelation may be present if observations have a natural sequential order (for example, time).**
- **May be fixed by introducing a variable or transforming a variable.**

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**



Validating the Regression Model

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

Example

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

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

Regression Models

- **In linear regression, we choose the "best" coefficients $b_0, b_1, \ldots, b_k$ as the estimates for $\beta_0, \beta_1, \ldots, \beta_k$.**
- **We know on average each b_j hits the right target β_j.**
- **However, we also want to know how confident we are about our estimates**

Back to Regression Output

Regression Statistics

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

Analysis of Variance

	df	Sum of Squares	Mean Square
Regression		16997.537	5665.85
Residual	11	3407.473	309.77
Total		20405.009	

	Coefficients	Standard Error	t Statistic	P-value	Lower 95%	Upper 95%
Intercept	65.71	27.73	2.37		4.67	126.74
Advertising	48.98	10.66	4.60		25.52	72.44
Promotions	59.65	23.63	2.53		7.66	111.65
Compet. Sales	-1.84	0.81	-2.26		-3.63	-0.047

Regression Output Analysis

1) **Degrees of freedom (dof)**
   - **Residual dof = n - (k+1)** **(We used up (k + 1) degrees of freedom in forming (k+1) sample estimates $b_0, b_1, \ldots, b_k$.)**

2) **Standard errors of the coefficients:** $s_{b0}, s_{b1}, \ldots, s_{bk}$
   - **They are just the SDs of estimates $b_0, b_1, \ldots, b_k$.**
   - **Fact: Before we observe b_j and s_{bj}, $\frac{b_j - \beta_j}{s_{b_j}}$ obeys a t-distribution with dof = (n - k - 1), the same dof as the residual.**
   - **We will use this fact to assess the quality of our estimates b_j.**
     - **What is a 95% confidence interval for β_j?**
     - **Does the interval contain 0? Why do we care about this?**

3) t-Statistic: $t_j = \frac{b_j}{s_{b_j}}$

- A measure of the statistical significance of each individual x_j in accounting for the variability in Y.
- Let c be that number for which

  $P(-c < T < c) = \alpha\ \%$,

  where T obeys a t-distribution with dof = (n - k - 1).
- If $|t_j| > c$, then the α % C.I. for β_j does not contain zero
- In this case, we are α% confident that β_j different from zero.

Example: Executive Compensation

Number	Pay ($1,000)	Years in position	Change in Stock Price (%)	Change in Sales (%)	MBA?
1	1,530	7	48	89	YES
2	1,117	6	35	19	YES
3	602	3	9	24	NO
4	1,170	6	37	8	YES
5	1,086	6	34	28	NO
6	2,536	9	81	-16	YES
7	300	2	-17	-17	NO
8	670	2	-15	-67	YES
9	250	0	-52	49	NO
10	2,413	10	109	-27	YES
11	2,707	7	44	26	YES
12	341	1	28	-7	NO
13	734	4	10	-7	NO
14	2,368	8	16	-4	NO

. . . . . .

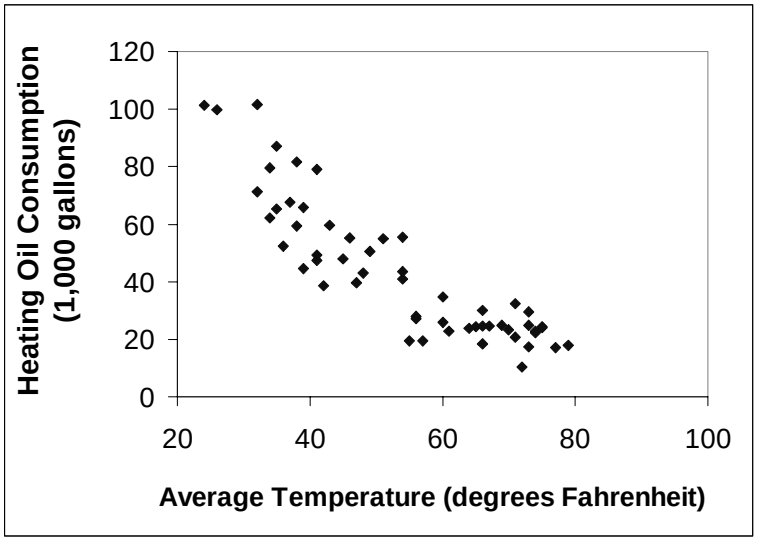
Dummy variables:

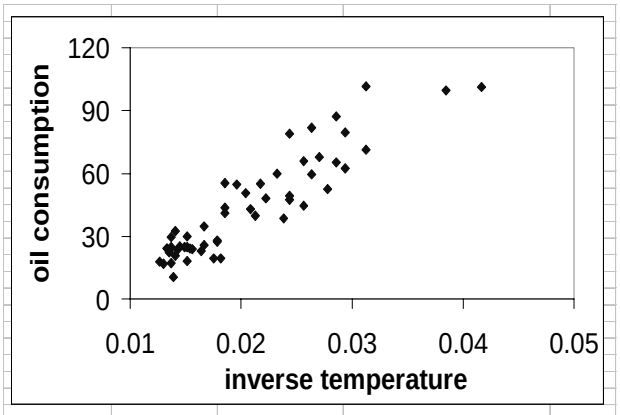
- **Often, some of the explanatory variables in a regression are *categorical* rather than *numeric*.**
- **If we think whether an executive has an MBA or not affects his/her pay, We create a *dummy* variable and let it be 1 if the executive has an MBA and 0 otherwise.**
- **If we think season of the year is an important factor to determine sales, how do we create dummy variables? How many?**
- **What is the problem with creating 4 dummy variables?**
- **In general, if there are m categories an x variable can belong to, then we need to create m-1 dummy variables for it.**

OILPLUS data

	Month	heating oil	temperature
1	August, 1989	24.83	73
2	September, 1989	24.69	67
3	October, 1989	19.31	57
4	November, 1989	59.71	43
5	December, 1989	99.67	26
6	January, 1990	49.33	41
7	February, 1990	59.38	38
8	March, 1990	55.17	46
9	April, 1990	55.52	54
10	May, 1990	25.94	60
11	June, 1990	20.69	71
12	July, 1990	24.33	75
13	August, 1990	22.76	74
14	September, 1990	24.69	66
15	October, 1990	22.76	61
16	November, 1990	50.59	49
17	December, 1990	79.00	41

. . . . . .





heating oil	temperature	inverse temperature
24.83	73	0.0137
24.69	67	0.0149
19.31	57	0.0175
59.71	43	0.0233
99.67	26	0.0385
49.33	41	0.0244

The Practice of Regression

- **Choose which independent variables to include in the model, based on common sense and context specific knowledge.**
- **Collect data (create dummy variables in necessary).**
- **Run regression –– the easy part.**
- **Analyze the output and make changes in the model –– this is where the action is.**
- **Test the regression result on "out-of-sample" data**

The Post-Regression Checklist

1) Statistics checklist:

Calculate the correlation between pairs of x variables —– watch for evidence of multicollinearity

Check signs of coefficients – do they make sense?

Check 95% C.I. (use t-statistics as quick scan) – are coefficients significantly different from zero?

R^2 :overall quality of the regression, but not the only measure

2) Residual checklist:

Normality – look at histogram of residuals

Heteroscedasticity – plot residuals with each x variable

Autocorrelation – if data has a natural order, plot residuals in order and check for a pattern

The Grand Checklist

- **Linearity:** scatter plot, common sense, and knowing your problem, transform including interactions if useful
- **t-statistics:** are the coefficients significantly different from zero? Look at width of confidence intervals
- **F-tests for subsets, equality of coefficients**
- **$R^{2:}$** is it reasonably high in the context?
- **Influential observations, outliers in predictor space, dependent variable space**

- **Normality:** plot histogram of the residuals
- **Studentized residuals**

- **Heteroscedasticity**: plot residuals with each x variable, transform if necessary, Box-Cox transformations
- **Autocorrelation:** "time series plot"
- **Multicollinearity**: compute correlations of the x variables, do signs of coefficients agree with intuition?
  - **Principal Components**
- **Missing Values**