Causal Forecasting Models

MIT Center for Transportation & Logistics

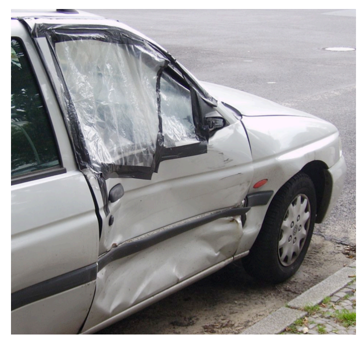
Causal Models

- Used when demand is correlated with some known and measurable environmental factor.
- Demand (y) is a function of some variables $(x_1, x_2, \ldots x_k)$

Dependent Variable

Independent Variables

Disposable Diapers
~ f(births, household income)



Car Repair Parts
~ f(weather/snow)

Promoted Item
~f(discount, placement, advertisements)

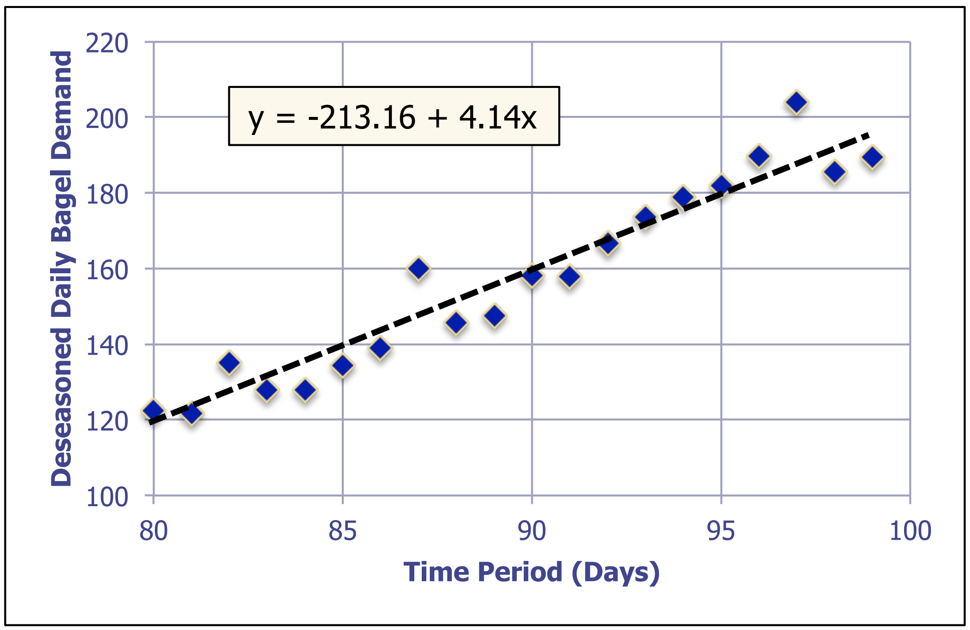
Agenda

- Simple Linear Regression
- Regression in Spreadsheets
- Multiple Linear Regression
- Model Transformations
- Model Fit and Validity

Simple Linear Regression

Example: Simple Linear Regression

- Recall from earlier lecture on exponential smoothing
- Estimating initial parameters for Holt-Winter (level, trend, seasonality)
- Removed seasonality in order to estimate initial level and trend



$$y_i = \beta_0 + \beta_1 x_i$$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad for\ i = 1,2,\ldots n$$

Observed

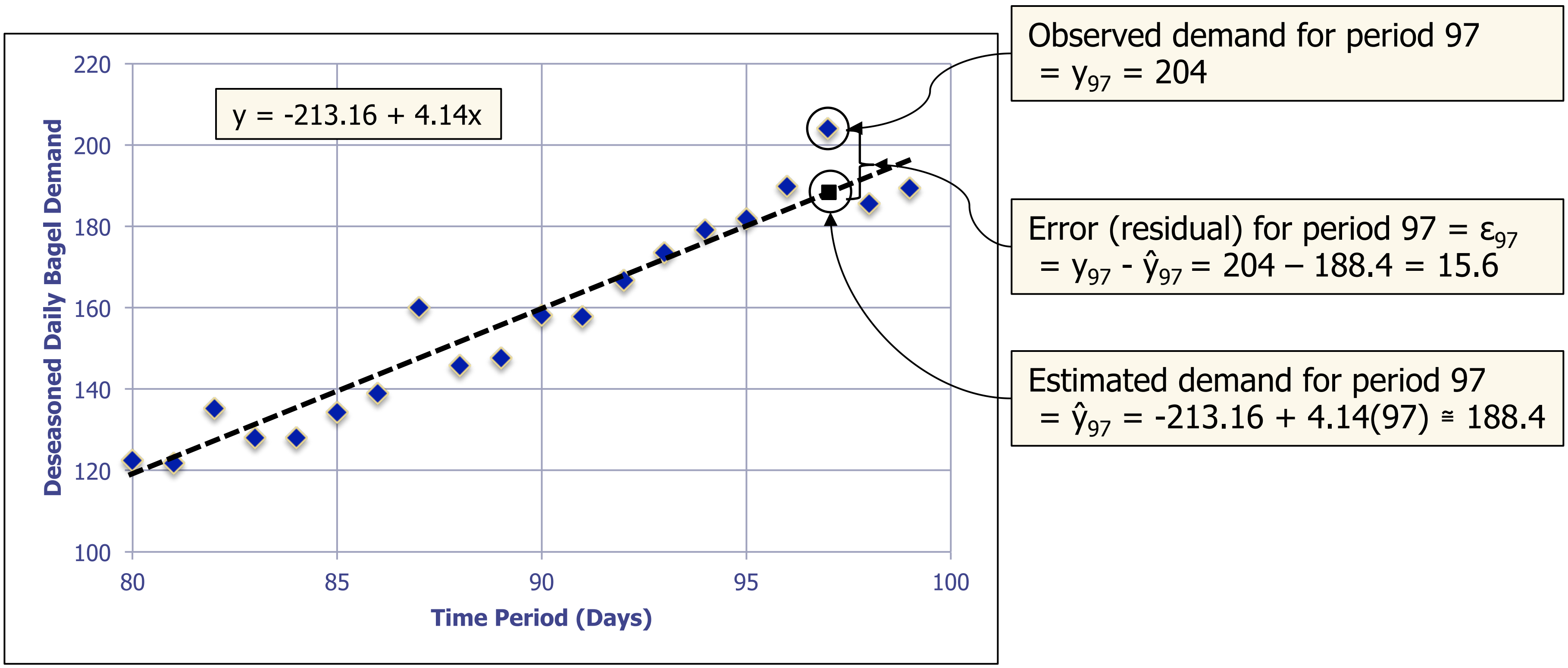
Unknown

$$E(Y \mid x) = \beta_0 + \beta_1 x$$

$$StdDev(Y \mid x) = \sigma$$

Simple Linear Regression

- The relationship is described in terms of a linear model
- The data (x_i, y_i) are the observed pairs from which we try to estimate the Beta coefficients to find the 'best fit'
- The error term, ε, is the 'unaccounted' or 'unexplained' portion
- The error terms are assumed to be iid ~N(0,σ)



Simple Linear Regression

- Residuals or Error Terms
  - Residuals, e_i, are the difference of actual minus predicted values
  - Find the b's that "minimize the residuals"

$$\hat{y}_i = b_0 + b_1 x_i \quad for\ i = 1,2,\ldots n$$

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i \quad for\ i = 1,2,\ldots n$$

- How should we "minimize" the residuals?
  - Min sum of errors - shows bias, but not accurate
  - Min sum of absolute error - accurate & shows bias, but intractable
  - Min sum of squares of error – shows bias & is accurate

$$\sum_{i=1}^{n}\left(e_i^2\right) = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right)^2$$

Simple Linear Regression

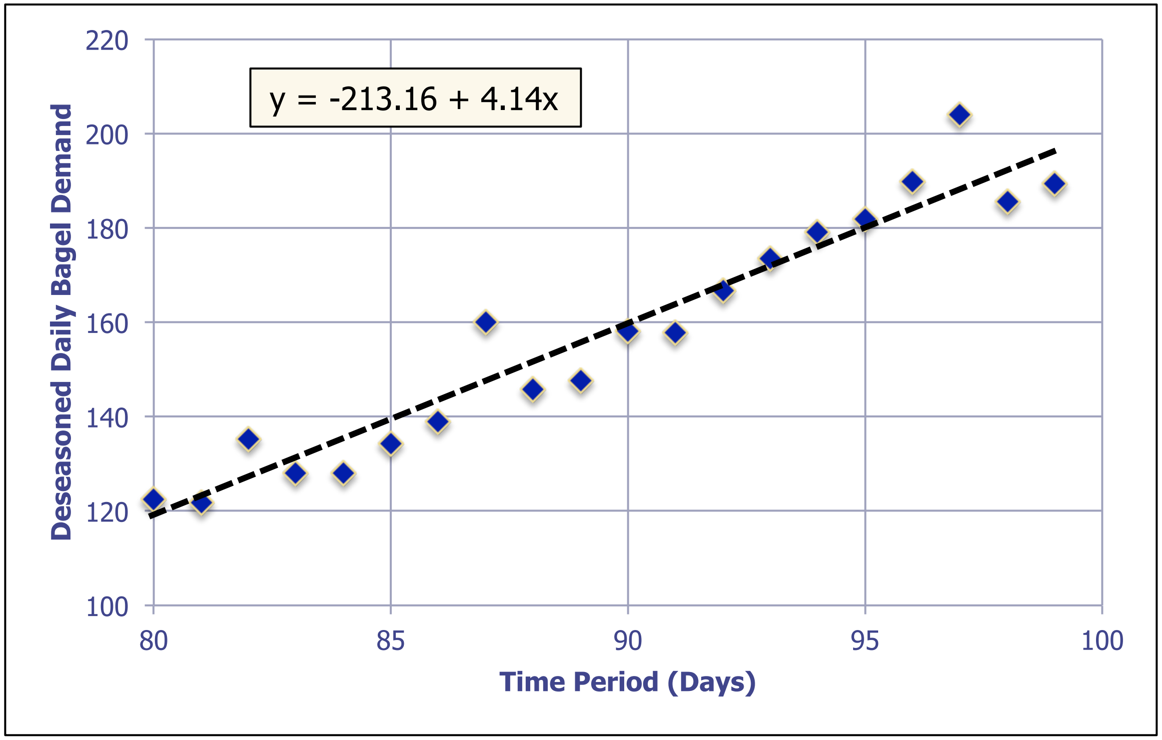
- Ordinary Least Squares (OLS) Regression
  - Finds the coefficients (b_0 and b_1) that minimize the sum of the squared error terms.
  - We can use partial derivatives to find the first order optimality condition with respect to each variable.

$$\sum_{i=1}^{n}\left(e_i^2\right) = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right)^2$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

We know from the data:
$\bar{x} = 89.5$ $\bar{y} = 157.4$



OLS Regression in Spreadsheet

Regression – By Hand

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Original data (y in column A, x in column B)

	A	B	C	D	E	F
1	Deseasoned Demand (y)	Time Period (x)	(x-avgx)	(y-avgy)	(x-avgx)^2	(x-avgx)* (y-avgy)
2	122.5	1	-9.50	-34.90	90.25	331.50
3	121.7	2	-8.50	-35.70	72.25	303.41
4	135.2	3	-7.50	-22.20	56.25	166.46
5	128.0	4	-6.50	-29.40	42.25	191.07
6	128.0	5	-5.50	-29.40	30.25	161.67
7	134.4	6	-4.50	-23.00	20.25	103.48
8	139.0	7	-3.50	-18.40	12.25	64.38
9	160.0	8	-2.50	2.60	6.25	-6.51
10	145.8	9	-1.50	-11.60	2.25	17.39
11	147.6	10	-0.50	-9.80	0.25	4.90
12	158.1	11	0.50	0.70	0.25	0.35
13	157.8	12	1.50	0.41	2.25	0.61
14	166.7	13	2.50	9.30	6.25	23.26
15	173.6	14	3.50	16.21	12.25	56.72
16	179.0	15	4.50	21.61	20.25	97.22
17	181.8	16	5.50	24.41	30.25	134.23
18	189.8	17	6.50	32.41	42.25	210.63
19	204.0	18	7.50	46.61	56.25	349.54
20	185.5	19	8.50	28.11	72.25	238.89
21	189.4	20	9.50	32.01	90.25	304.05
22	**157.4**	**10.5**	**0.0**	**0.0**	**665.0**	**2753.25**
23	Average		Sum			
24	b1 (trend) =	4.14				
25	b0 (intercept) =	113.92				

=B7−B22

=A7−A22

=C7*D7

=C7^2

=SUM(C2:C21) =SUM(D2:D21)
=SUM(E2:E21) =SUM(F2:F21)

=F22/E22

=A22-B24*B22

=AVERAGE(A2:A21) =AVERAGE(B2:B21)

Regression Equation
$y = b_0 + b_1 x$
$y = 113.92 + 4.14x$

Regression – Using LINEST function

Original data (y in column A, x in column B)

	A	B	C	D	E
1	Deseasoned Demand (y)	Time Period (x)			
2	122.5	1		4.14	113.92
3	121.7	2		0.27	3.22
4	135.2	3		0.93	6.94
5	128.0	4		236.84	18.00
6	128.0	5		11399.08	866.33
7	134.4	6			
8	139.0	7			
9	160.0	8			
10	145.8	9			
11	147.6	10			
12	158.1	11			
13	157.8	12			
14	166.7	13			
15	173.6	14			
16	179.0	15			
17	181.8	16			
18	189.8	17			
19	204.0	18			
20	185.5	19			
21	189.4	20			

=LINEST(A2:A21,B2:B21,1,1)

LINEST(known_y's, known_x's, constant, statistics)

b_1	b_0
s_{b1}	s_{b0}
R^2	s_e
F	d_f
SSR	SSE

- The LINEST is an array function
  - Receives and returns data to multiple cells
  - The equation will be bookended by {} brackets when active
  - While the function is the same in both LibreOffice and Excel, activating it differs slightly.
- LibreOffice
  - Type the formula into cell D2 and press the keyboard combination **Ctrl+Shift+Enter** (for Windows & Linux) or **command+shift+return** (for Mac OS X).
- Excel
  - Select a range of 2 columns by 5 rows, in this case (D2:E6).
  - Then, in the 'Insert Function' area, type the formula and press the keyboard combination **Ctrl+Shift+Enter** (for Windows & Linux) or **command+shift+return** (for Mac OS X).

Regression – Using LINEST

4.14	113.92
0.27	3.22
0.93	6.94
236.84	18
11399	866

b_1	b_0
s_{b1}	s_{b0}
R^2	s_e
F	d_f
SSR	SSE

n = number of observations
k = number of explanatory variables (NOT intercept)
d_f = degrees of freedom (n-k-1)
b_0 = estimate of the intercept $b_0 = \bar{y} - b_1\bar{x}$

b_1 = estimate of the slope (explanatory variable 1)

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Goodness of fit of the model – proportion of the variation in Y which is explained by X

Total Sum of Squares (SST)

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

"Explained" Portion Sum of Squares of Regression (SSR)

"Unexplained" Portion Sum of Squares of the Error (SSE)

R^2 = Coefficient of Determination: the ratio of "explained" to total sum of squares where $0 \le R^2 \le 1$

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

Regression – Using LINEST

4.14	113.92
0.27	3.22
0.93	6.94
236.84	18
11399	866

b_1	b_0
s_{b1}	s_{b0}
R^2	s_e
F	d_f
SSR	SSE

s_e = standard error of estimate: an estimate of variance of the error term around the regression line.

$$s_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-k-1}} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-k-1}}$$

s_{b0} = standard error of intercept

$$s_{b0} = s_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$$

s_{b1} = standard error of slope

$$s_{b1} = s_e\sqrt{\frac{1}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$$

How significant is the explanatory variable? Is it different from zero?

- Test the null hypothesis H_0: $b_1=0$ with alternate hypothesis H_A: $b_1 \neq 0$
- Use two-tailed t-test =TDIST(t_statistic, d_f, number tails) - always use 2 tail test
- Accepted thresholds for p-value ≤ 0.01, 0.05, or 0.10 (meaning we can reject the H_0 with 99%, 95%, and 90% probability respectively)

$$t_{b1} = \frac{b_1}{s_{b1}} = \frac{4.14}{0.27} = 15.33$$

p-value =TDIST(15.33, 18, 2) = $8.92 \times 10^{-12} < 0.01$

Regression – Using LINEST

b_1	b_0
s_{b1}	s_{b0}
R^2	s_e
F	d_f
SSR	SSE

Original data (y in column A, x in column B)

	A	B	C	D	E
1	Deseasoned Demand (y)	Time Period (x)			
2	122.5	1		4.14	113.92
3	121.7	2		0.27	3.22
4	135.2	3		0.93	6.94
5	128.0	4		236.84	18.00
6	128.0	5		11399.08	866.33
7	134.4	6			
8	139.0	7		b1	b0
9	160.0	8	T-Stat	15.39	35.35
10	145.8	9	P-value	0.0000%	0.0000%
11	147.6	10			
12	158.1	11			
13	157.8	12			
14	166.7	13			
15	173.6	14			
16	179.0	15			
17	181.8	16			
18	189.8	17			
19	204.0	18			
20	185.5	19			
21	189.4	20			

=LINEST(A2:A21,B2:B21,1,1)

LINEST(known_y's, known_x's, constant, statistics)

=D2/D3

=TDIST(D9,E5,2)

1. How is the overall fit of the model?
   - Look at Coefficient of Determination R^2
   - No hard rules, but ≥0.70 is preferred
2. Are the individual variables statistically significant?
   - Use t-test for each explanatory variable
   - Lower p-value is better
   - Generally used threshold values include 0.10, 0.05, 0.01

Multiple Linear Regression

Example: Monthly Iced Coffee Sales

	A	B	C	D	E	F
1	Demand	Time Period	Forecast Average Temp	Year	Month Name	School in Session
2	3025	1	37	1	Jan	No
3	3136	2	39	1	Feb	Yes
4	3414	3	46	1	Mar	Yes
5	3502	4	56	1	Apr	Yes
6	3736	5	67	1	May	Yes
7	3661	6	77	1	Jun	No
8	3553	7	82	1	Jul	No
9	3691	8	80	1	Aug	No
10	3474	9	73	1	Sep	Yes
11	3876	10	62	1	Oct	Yes
12	3865	11	52	1	Nov	Yes
13	3967	12	42	1	Dec	Yes
14	3596	13	37	2	Jan	No
15	4345	14	39	2	Feb	Yes
16	4413	15	46	2	Mar	Yes
17	4086	16	56	2	Apr	Yes
18	4377	17	67	2	May	Yes
19	4220	18	77	2	Jun	No
20	4238	19	82	2	Jul	No
21	4007	20	80	2	Aug	No
22	4086	21	73	2	Sept	Yes
23	4536	22	62	2	Oct	Yes
24	4291	23	52	2	Nov	Yes
25	4427	24	42	2	Dec	Yes

- Develop Forecasting Model #1
  - Level, trend, & avg. historical temperature
  - Develop OLS regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

DEMAND = LEVEL + TREND(period) + TEMP_EFFECT(temp)

- Using LINEST function
  - Follow earlier directions
  - {=LINEST(A2:A25,B2:C25,1,1)}
  - When activating, expand area to five (5) rows by k+1 columns
  - Output shifts for new variables
    - Top right is always b_0
    - Bottom left six cells don't change

Output

(0.27)	52.65	3,254.81
2.75	6.24	174.85
0.78	208.97	#N/A
36.36	21	#N/A
3,175,996	917,074	#N/A

b_2	b_1	b_0
s_{b2}	s_{b1}	s_{b0}
R^2	s_e	
F	d_f	
SSR	SSE	

Example: Monthly Iced Coffee Sales

(0.27)	52.65	3,254.81
2.75	6.24	174.85
0.78	208.97	#N/A
36.36	21	#N/A
3,175,996	917,074	#N/A

b_2	b_1	b_0
s_{b2}	s_{b1}	s_{b0}
R^2	s_e	
F	d_f	
SSR	SSE	

n= 24 observations
k = 2 variables
d_f = n-k-1 = 24-2-1= 21

- How is the overall fit of the model?
  - R^2 = 0.78 or 78%
- Are the individual variables statistically significant?
  - Run t-tests for each variable and the intercept

intercept

$$t_{b0} = \frac{b_0}{s_{b0}} = \frac{3255}{175} = 18.60$$

P-value =TDIST(18.6, 21, 2) < 0.0001

trend

$$t_{b1} = \frac{b_1}{s_{b1}} = \frac{52.65}{6.24} = 8.44$$

P-value =TDIST(8.44, 21, 2) < 0.0001

temperature effect

$$t_{b2} = \frac{b_2}{s_{b2}} = \frac{-0.27}{2.75} = -0.098$$

P-value =TDIST(0.27, 21, 2) = 0.9223

  - Both the intercept and trend coefficients are significant
  - Temperature effect is not, we cannot reject the H_0
- What next?
  - Try the model without the temperature effect

Example: Monthly Iced Coffee Sales

	A	B	C	D	E	F
1	Demand	Time Period	Forecast Average Temp	Year	Month Name	School in Session
2	3025	1	37	1	Jan	No
3	3136	2	39	1	Feb	Yes
4	3414	3	46	1	Mar	Yes
5	3502	4	56	1	Apr	Yes
6	3736	5	67	1	May	Yes
7	3661	6	77	1	Jun	No
8	3553	7	82	1	Jul	No
9	3691	8	80	1	Aug	No
10	3474	9	73	1	Sep	Yes
11	3876	10	62	1	Oct	Yes
12	3865	11	52	1	Nov	Yes
13	3967	12	42	1	Dec	Yes
14	3596	13	37	2	Jan	No
15	4345	14	39	2	Feb	Yes
16	4413	15	46	2	Mar	Yes
17	4086	16	56	2	Apr	Yes
18	4377	17	67	2	May	Yes
19	4220	18	77	2	Jun	No
20	4238	19	82	2	Jul	No
21	4007	20	80	2	Aug	No
22	4086	21	73	2	Sept	Yes
23	4536	22	62	2	Oct	Yes
24	4291	23	52	2	Nov	Yes
25	4427	24	42	2	Dec	Yes

- Develop Forecasting Model #2
  - Level and trend
  - Develop OLS regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$

DEMAND = LEVEL + TREND(period)

- Using LINEST function
  - Follow earlier directions
  - {=LINEST(A2:A25,B2:B25,1,1)}

52.55	3239.89
6.02	86.05
0.78	204.22
76.14	22
3175570	917500

b_1	b_0
s_{b1}	s_{b0}
R^2	s_e
F	d_f
SSR	SSE

- Model fit? R^2=0.78
- Variables?
  - p-value for b_0 and b_1 are both < 0.0001

Example: Monthly Iced Coffee Sales

- Compare the goodness of fit between models
  - Model 1:
    - DEMAND = LEVEL + TREND(period) + TEMP_EFFECT(temp)
    - $R^2 = 0.77594$
  - Model 2:
    - DEMAND = LEVEL + TREND(period)
    - $R^2 = 0.77584$
- If Model #2 is "better", why is the R^2 lower?
  - R^2 will never get worse (and will usually improve) by adding more variables – even bad ones!
  - Need to modify the metric – adjusted R^2
    - Model 1: adj $R^2 = 1 - (1-0.77594)(23/21) = 0.754600$
    - Model 2: adj $R^2 = 1 - (1-0.77584)(23/22) = 0.765651$

$$adj\ R^2 = 1 - \left(1 - R^2\right)\left(\frac{n-1}{n-k-1}\right)$$

Transforming Variables

Example: Monthly Iced Coffee Sales

	A	B	C
1	Demand	Time Period	School in Session
2	3025	1	0
3	3136	2	1
4	3414	3	1
5	3502	4	1
6	3736	5	1
7	3661	6	0
8	3553	7	0
9	3691	8	0
10	3474	9	1
11	3876	10	1
12	3865	11	1
13	3967	12	1
14	3596	13	0
15	4345	14	1
16	4413	15	1
17	4086	16	1
18	4377	17	1
19	4220	18	0
20	4238	19	0
21	4007	20	0
22	4086	21	1
23	4536	22	1
24	4291	23	1
25	4427	24	1

- Develop Forecasting Model #3
  - Level, trend, & school being open
  - Develop OLS regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_3 x_{3i} + \varepsilon_i$$

DEMAND = LEVEL + TREND(period) + OPEN_EFFECT(open)

- Need to create Dummy Variable
  - x_{3i} = 1 if School is in Session, =0 otherwise
  - Interpret β_3 as increase (decrease) in demand when school is in session
- Using LINEST function
  - {=LINEST(A2:A25,B2:C25,1,1)}

144.50	51.54	3156.12
85.35	5.81	96.30
0.80276	196.07	#N/A
42.74	21	#N/A
3285765	807305	#N/A

b_3	b_1	b_0
s_{b3}	s_{b1}	s_{b0}
R^2	s_e	
F	d_f	
SSR	SSE	

Example: Monthly Iced Coffee Sales (#3)

144.50	51.54	3156.12
85.35	5.81	96.30
0.80276	196.07	#N/A
42.74	21	#N/A
3285765	807305	#N/A

b_3	b_1	b_0
s_{b3}	s_{b1}	s_{b0}
R^2	s_e	
F	d_f	
SSR	SSE	

n= 24 observations
k = 2 variables
d_f = n-k-1 = 24-2-1= 21

- How is the overall fit of the model?
  - R^2 = 0.80276 with adj $R^2 \cong$ 0.78398 (better than #1 or #2)
- Are the individual variables statistically significant?
  - Run t-tests for each variable and the intercept
  - Intercept and trend coefficients are strongly significant, school flag is borderline

intercept

$$t_{b0} = \frac{b_0}{s_{b0}} = \frac{3156}{96.3} = 32.77$$

P-value =TDIST(32.77, 21, 2) < 0.0001

trend

$$t_{b1} = \frac{b_1}{s_{b1}} = \frac{51.54}{5.81} = 8.87$$

P-value =TDIST(8.87, 21, 2) < 0.0001

school in session

$$t_{b3} = \frac{b_3}{s_{b3}} = \frac{144.50}{85.35} = 1.69$$

P-value =TDIST(1.69, 21, 2) = 0.105

- Let's interpret this:

Demand = 3156 + 52(t) + 145(if in session)

  - We are forecasting a monthly demand level of 3,156 iced coffees with a monthly trend of ~52 additional cups each month and an increase of ~145 cups whenever school is in session.
  - My forecast for sales:
    - January year 3 = 3156 + 25(51.5) + 144.5(0) = 4444
    - February year 3 = 3156 + 26(51.5) + 144.5(1) = 4640

Model & Variable Transformations

- We are using linear regression, so how can we use dummy variables?
  - The model just needs to be linear in the parameters
  - For model #3: $y = \beta_0+\beta_1(\text{period})+\beta_3(\text{open\_flag})$
  - Many transformations can be used:

$$y = \beta_0 + \beta_1 x_1$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

$$y = \beta_0 + \beta_1 \ln(x_1)$$

$$y = ax^b \Rightarrow \ln(y) = \ln(a) + b\ln(x)$$

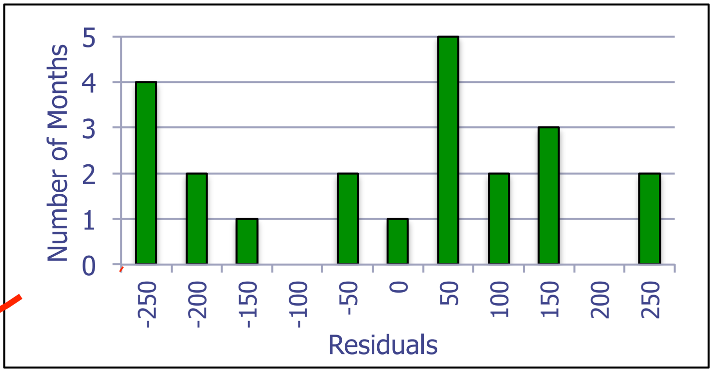
$$y = a x_1^{b1} x_2^{b2} \Rightarrow \ln(y) = \ln(a) + b_1 \ln(x_1) + b_2 \ln(x_2)$$

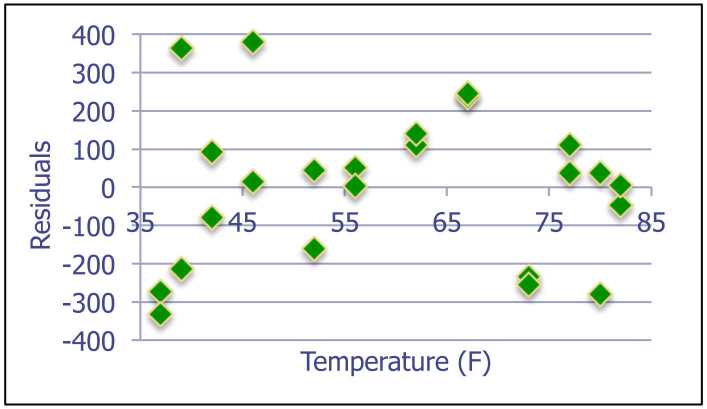
- Transformations and dummy variables allow for many models
  - For example:
    - x_{4i} = (x_{3i})*(temperature) if sales increase with temperature when school is in session
    - x_{5i} = 1 if competing store runs a sale, =0 otherwise
    - $x_{6i} = x_{1i}^2$, so that we can capture a tapering effect to the linear trend
  - But, be careful on interpretation of results

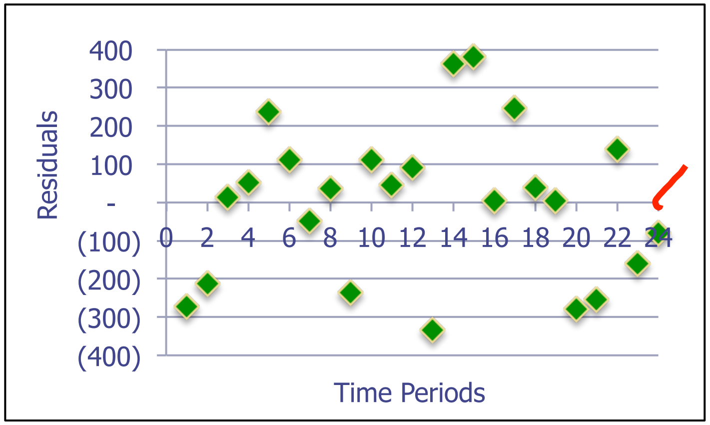
Model Fit & Validation

Model Validation

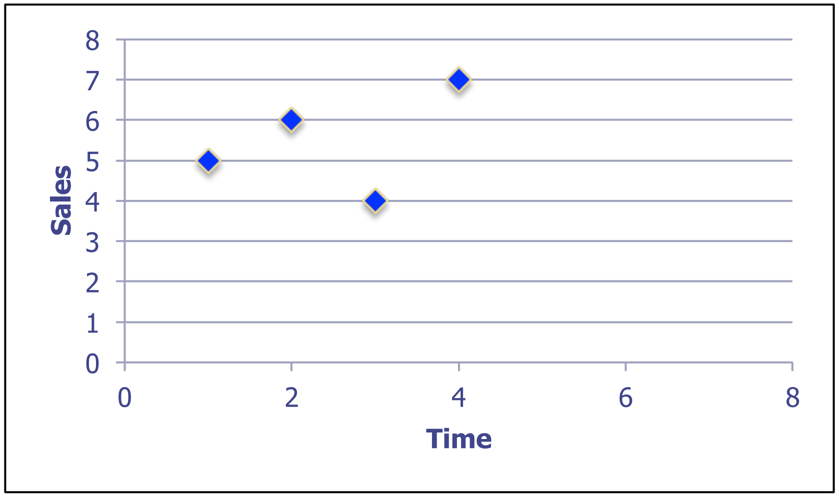
- Basic Checks
  - Goodness of Fit – look at the R^2 values
  - Individual coefficients – t-tests for p-value
- Additional Assumption Checks
  - Normality of residuals – look at histogram
  - Heteroscedasticity – look at scatter plot of residuals
    - Does the standard deviation of the error terms differ for different values of the independent variables?
  - Autocorrelation – is there a pattern over time
    - Are the residuals not independent?
  - Multi-Collinearity – look at correlations
    - Are the independent variables correlated?
    - Make sure dummy variables were not over specified
- Statistics Software
  - Most packages check for all of these
  - More sophisticated tests and remedies

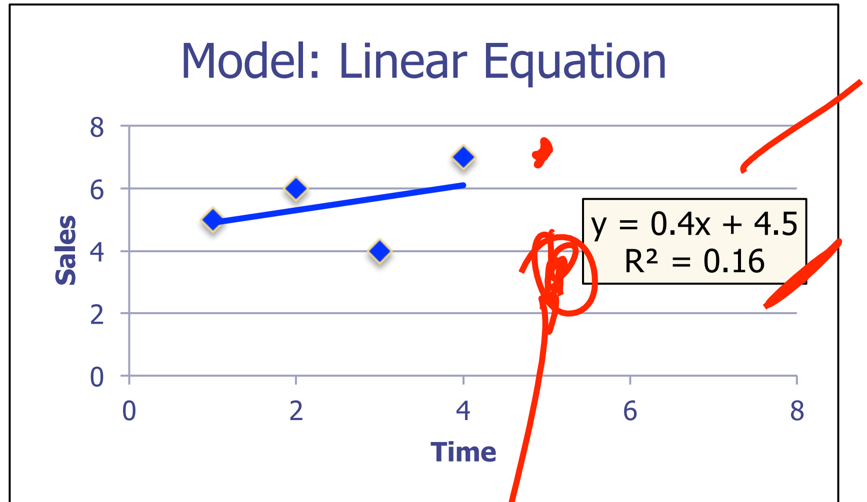


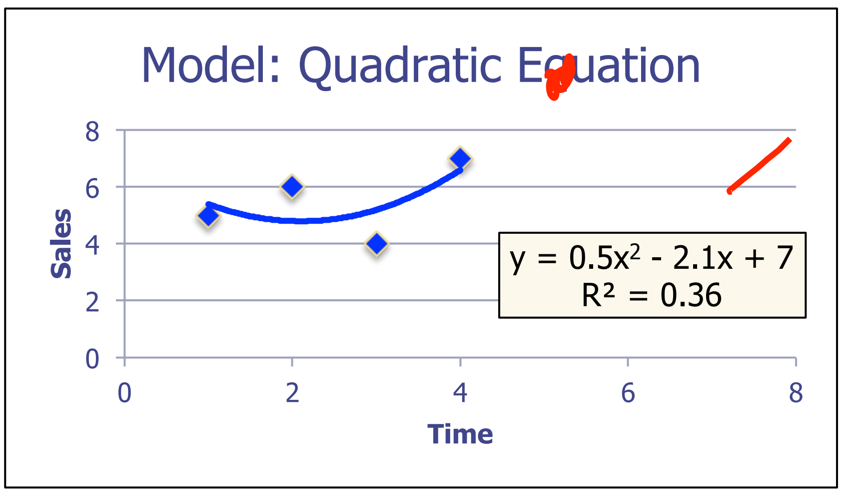


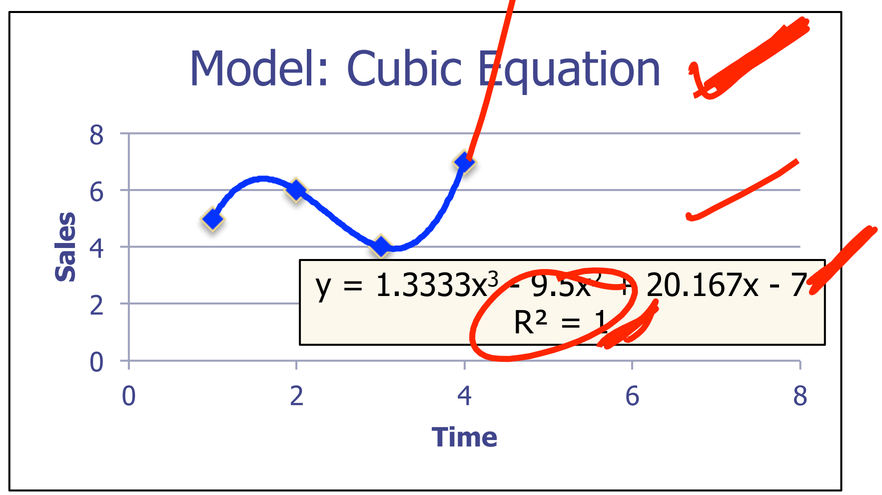


Modeling Results – which is best?









Avoid over-fitting. Objective is to forecast demand for planning purposes.

Key Points from Lesson

Key Points

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

- Regression finds correlations between
  - A single dependent variable (y)
  - One or more independent variables (x_1, x_2, …)
- Coefficients are estimates by minimizing the sum of the squares of the errors
- Always test your model:
  - Goodness of fit (R^2)
  - Statistical significance of coefficients (p-value)
- Some Warnings:
  - Correlation is not causation
  - Avoid over-fitting of data
- Why not use this instead of exponential smoothing?
  - All data treated the same
  - Amount of data required to store

CTL.SC1x -Supply Chain & Logistics Fundamentals

Questions, Comments, Suggestions? Use the Discussion!

"Casey"
Photo courtesy Yankee Golden Retriever Rescue (www.ygrr.org)

MIT Center for Transportation & Logistics

caplice@mit.edu

Image Sources

- "Pampers packages (2718297564)" by Elizabeth from Lansing, MI, USA - Pampers packagesUploaded by Dolovis. Licensed under Creative Commons Attribution 2.0 via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:Pampers_packages_(2718297564).jpg#mediaviewer/File:Pampers_packages_(2718297564).jpg
- "Kraft Coupon" by Julie & Heidi from West Linn & Gillette, USA - Grocery Coupons - Tearpad shelf display of coupons for Kraft Macaroni & Cheese. Licensed under Creative Commons Attribution-Share Alike 2.0 via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:Kraft_Coupon.jpg#mediaviewer/File:Kraft_Coupon.jpg
- "Damaged car door" by Garitzko - Own work. Licensed under Public domain via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:Damaged_car_door.jpg#mediaviewer/File:Damaged_car_door.jpg