Building Predictive Models

Duffy Industries

- Robin Curtin, the Transportation vice president for Duffy Industries, a food manufacturer, is trying to understand their transportation rates. She needs to estimate what rates should be for full truckload (TL) service from a new facility. Duffy uses contract "over-the-road" trucking companies for TL shipments. These are moves directly from one point to another with no intermediate stops.
- Robin only has a snapshot of data showing the costs and some other characteristics of about 100 TL shipments.
- Some questions she would like to answer:
  - What characteristics are driving the rates for my TL services?
  - What rates should I expect if I open new lanes?

Duffy Industries

- Let's take a look at a snapshot of the data

ID	Cost Per Load	Distance (miles)	LeadTime (days)	Trailer Length (ft)	Weight (lbs)	Equipment
1	$3,692	1,579	1	53	20,559	DRY
2	$3,279	1,298	12	48	17,025	REF
3	$3,120	1,382	11	48	26,736	DRY
4	$3,205	1,033	1	53	26,176	DRY
5	$3,188	1,320	3	53	17,994	DRY
6	$2,835	1,103	9	53	32,207	DRY
7	$2,364	743	1	48	18,589	DRY

ID	unique identification number for the load
CPL	cost per load ($)
Dist	distance hauled for shipment (miles)
LdTime	lead time from offer to tender to carrier (days) 0 = same day
TrlLng	trailer length (feet)
Wgt	weight of goods in trailer (lbs)
Eqpt	equipment type (Dry Van or Refrigerated)

Duffy Industries – Quick Statistics

	CPL	Dist	LdTime	TrlLng	Wgt
Min	$1,660	502	0.00	48	15,100
25th Pct	$2,632	904	2.75	48	21,221
Mode	$3,730	#N/A	1.00	53	#N/A
Median	$3,166	1273	6.00	53	26,514
Mean	$3,132	1207	5.87	51.3	26,709
75th Pct	$3,701	1538	9.00	53	32,277
Max	$4,301	1793	13.00	53	39,932
Range	$2,641	1291	13.00	5	24,832
IQ	$1,070	634	6.25	5	11,056
StdDev.s	$652	385	3.94	2.38	7,034
CORR(CPL,X)	1.00	0.90	-0.09	0.14	0.08

Eqpt: 60 Ref and 40 Dry shipments

What to do next?

- Explore how CPL is influenced by other variables
- Develop a descriptive model where CPL=f(Dist, LdTime, . . . ?)
- Develop a predictive model for CPL

Setting up the Variables

Dependent vs. Independent Variables

- We want to measure movement of one (dependent) variable to a small set of relevant (independent) variables.
- Dependent variable Y is a function of independent variables X.
- Examples:
  - Property Values = f(area, location, # bathrooms, …)
  - Sales = f(last month's sales, advertising budget, price, seasonality, .. .)
  - Probability Taking Transit versus Driving = f(income, location, . . .)
  - Height = f(age, gender, height of parents, . . .)
  - GPA = f(GMAT, age, undergraduate Grades, …)
  - Number of Fliers = f(Economic activity, size of origin and destination cities, competitor's price, . . .)
  - Condominium fees = f(area, story . . .)

Variables Types

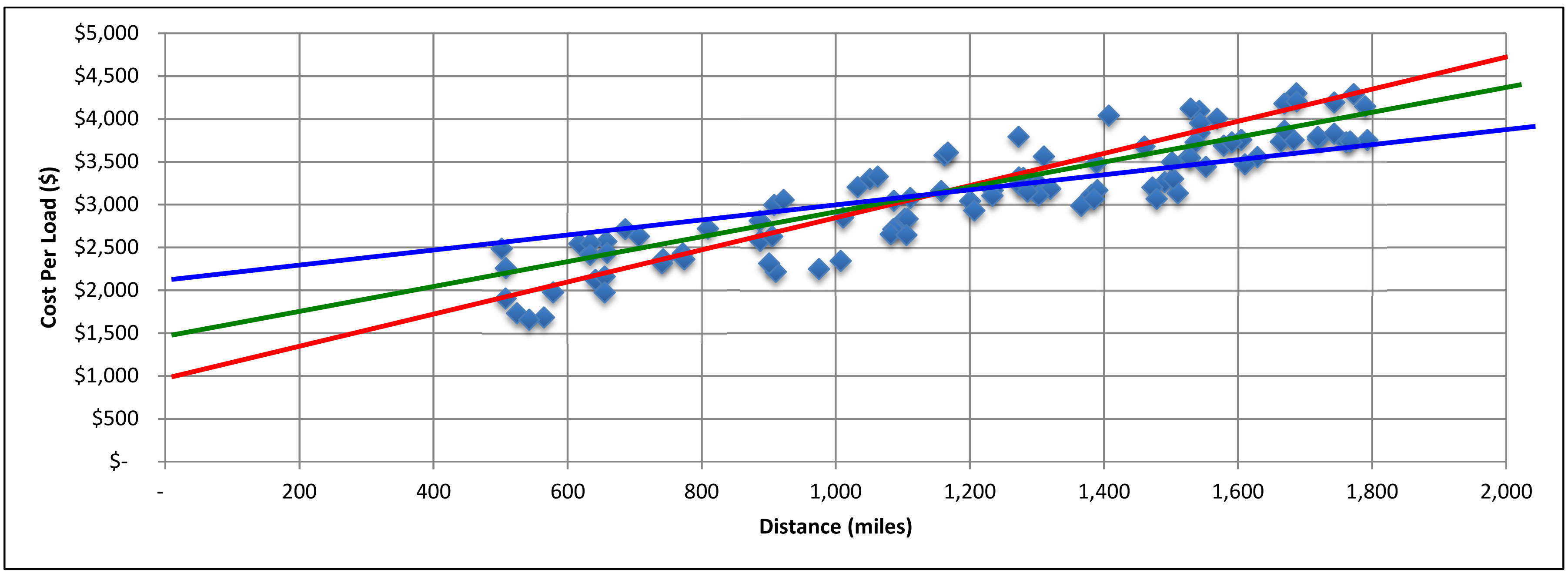
- Variables have different scales
  - Nominal Scale / Categorical – Groupings without related value
    - Income status (1=retired, 2=student, 3=two income, etc.)
    - Country of origin (1=US, 2=Canada, 3=China, etc.)
    - Organizational structure (1=centralized, 2=decentralized, etc.)
  - Ordinal Scale – indicates ranking but not proportionality between values
    - Job satisfaction scale 1 to 5 (a 2 is not twice as good as a 1)
    - Planning versus Response Profile (0 = Planner, 4 = Responder)
    - Education level (1=High School, 2=Undergraduate, 3=Masters, etc.)
  - Ratio Scale – value indicating ranking and relation
    - Examples; Age, Income, Cost, Distance, Weight, ...
- Form of the Dependent Variable dictates the method used
  - Continuous – takes any value
  - Discrete – takes only integer values
  - Binary – is equal to 0 or 1

We will focus on Linear Regression of continuous, ratio scaled dependent variables

Duffy Industries

- Dependent variable, Y: CPL or cost per load
- Potential independent variables, X_i:
  - Dist – distance
  - LdTime – lead time
  - TrlLng – length of trailer
  - Wgt – weight
  - Eqpt – equipment type

- Start with simple linear model
- Draw "best fit" line
- CPL = f(Dist) = $\beta_0 + \beta_1 X_1$



Linear Regression

Linear Regression Model

- Formally,
  - 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 Β 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) and catch all of the factors ignored or neglected in the model

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

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

Observed

Unknown

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

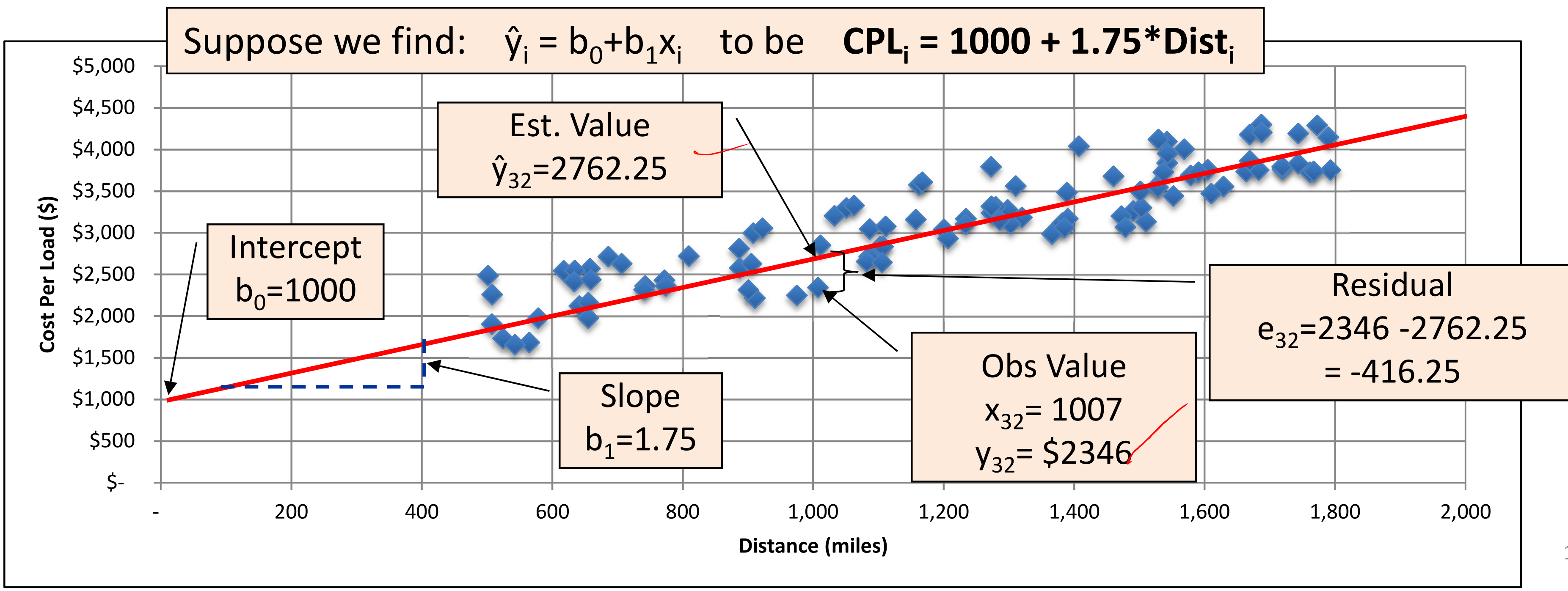
$$StdDev(Y \mid x) = \sigma$$

Linear Regression - Residuals

- Residuals
  - Predicted or estimated values found by using regression coefficients, **b**.
  - Residuals, e_i, are the difference of actual, y_i, minus predicted, $\hat{y}_i$, values
  - We want to select the "best" **b** values that "minimize the residuals"

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

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



Linear Regression – Best Fit

- How should I determine the "best fit" with the residuals?
  - Min sum of errors? $\min \Sigma(y_i - b_0 + b_1 x_i)$
  - Min sum of absolute error? $\min \Sigma|y_i - b_0 + b_1 x_i|$
  - Min sum of squares of error? $\min \Sigma(y_i - b_0 + b_1 x_i)^2$

We will select the model that minimizes the residual sum of squares . . . WHY?

- Ordinary Least Squares (OLS) Regression
  - Finds the optimal value of the coefficients (b_0 and b_1) that minimize the sum of the squares of the errors.

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

$$b_0 = \bar{y} - b_1 \bar{x} \qquad b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Formal Definitions

- Where did this come from?
  - Unconstrained optimization (think back to first week!)
  - Partial derivatives to find the first order optimality condition with respect to each variable.

$$\frac{\partial \sum e_i^2}{\partial b_0} = \frac{\partial \sum (y_i - b_0 - b_1 x_i)^2}{\partial b_0} = \sum_{i=1}^{n} -2(y_i - b_0 - b_1 x_i) = 0$$

$$= -2\sum_{i=1}^{n}(y_i) + 2\sum_{i=1}^{n}(b_0) + 2\sum_{i=1}^{n}(b_1 x_i) = 0$$

$$\sum_{i=1}^{n}(b_0) = \sum_{i=1}^{n}(y_i) - \sum_{i=1}^{n}(b_1 x_i)$$

$$nb_0 = \sum_{i=1}^{n}(y_i) - b_1 \sum_{i=1}^{n}(x_i)$$

$$b_0 = \frac{\sum_{i=1}^{n}(y_i)}{n} - b_1 \frac{\sum_{i=1}^{n}(x_i)}{n} = \bar{y} - b_1 \bar{x}$$

- We can expand to multiple variables
  - So, for k variables we need to find k regression coefficients

$$Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i \quad \text{for } i = 1, 2, \ldots n$$

$$E(Y \mid x_1, x_2, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

$$StdDev(Y \mid x_1, x_2, \ldots, x_k) = \sigma$$

$$\sum_{i=1}^{n}\left(e_i^{\,2}\right) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_{1i} - \ldots - b_k x_{ki})^2$$

Validating a Model – First Steps

Model Evaluation Metrics

- All packages will provide statistics for evaluation
  - Names and format will differ slightly package by package
- Model Output
  - Model Statistics (Regression Statistics or Summary of Fit)
    - Coefficient of Determination or Goodness of Fit (R^2)
    - Adjusted R^2
    - Standard Error (Root Mean Squared Error)
  - Analysis of Variance (ANOVA)
    - Sum of the Squares (Model, Residual/Error, and Total)
    - Degrees of Freedom
  - Parameter Statistics (Coefficient Statistics)
    - Coefficient (b value)
    - Standard error
    - t-Statistic
    - p-value
    - Upper and Lower Bounds

Model Validation 1 – Overall Fit

ID	CPL (Y)	Dist (X)
1	$3,692	1,579
2	$3,279	1,298
3	$3,120	1,382
4	$3,205	1,033
5	$3,188	1,320
6	$2,835	1,103
7	$2,364	743
8	$2,434	772
9	$3,486	1,389
10	$3,730	1,761
11	$3,735	1,664
12	$4,096	1,542
13	$2,123	641
14	$3,560	1,527
15	$4,041	1,407
16	$3,765	1,720
17	$3,565	1,310
18	$1,686	565
19	$3,045	1,200
20	$2,933	1,207
.		
.		
100	$4,208	1,687

- How much variation in dependent variable, y, can we explain?
  - If we only have the mean?
  - If we can make estimates?
- What is the total variation of CPL?
  - Find dispersion around the mean
  - Called the Total Sum of Squares

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

- What if we now make estimates of y for each x?
  - Find variation not accounted for by estimates
  - Called the Error or Residual Sum of Squares

$$e_i = y_i - \hat{y}_i$$

$$RSS = e_i^2 = \sum (y_i - \hat{y}_i)^2$$

- The regression model "explains" a certain percentage of the total variation of the dependent variable
  - Coefficient of Determination or R^2
  - Ranges between 0 and 1
  - The adj R^2 corrects for additional variables
    - n= # observations
    - k = # independent variables (not b_0)

$$R^2 = \boxed{1 - \frac{RSS}{TSS}} = 1 - \frac{(y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2}$$

$$\boxed{adjR^2} = 1 - \left(\frac{RSS}{TSS}\right)\left(\frac{n-1}{n-k-1}\right)$$

MIT Center for Transportation & Logistics

Model Validation 2 – Individual Coefficients

- Each Independent variable (and b_0) will have:
  - An estimate of coefficient (b_1),
  - A standard error (s_{bi})
    - s_e = Standard error of the model
    - s_x = Standard deviation of the independent variable
    - n = number of observations

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}}$$

  - The t-statistic ~Student-t, df=n-k-1
    - k = number of independent variables
    - b_i = estimate or coefficient of independent variable

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

  - Corresponding p-value – Testing the Slope
    - If no linear relationship exists between the two variables, we would expect the regression line to be horizontal, that is, to have a slope of zero.
    - We want to see if there is a linear relationship, i.e. we want to see if the slope (b_1) is something other than zero. So: H_0: $b_1 = 0$ and H_1 $b_1 \neq 0$
- Confidence Intervals
  - We can estimate an interval for the slope parameter, (b_1)

$$b_1 \pm t_{\alpha/2} s_{b_1} \qquad \nu = n-2$$

MIT Center for Transportation & Logistics

Model 1: CPL=f(Dist)

Model 1: $CPL = b_0 + b_1(Dist)$

R^2	0.818
adj R^2	0.816
s_e	281.3
RSS	7,754,694
TSS	42,519,984

Estimation Model

$CPL = 1282 + 1.532\ (Dist)$

	Coefficient	Std Error (s_{bi})	t-stat	p-value	Lower CI (95%)	Upper CI (95%)
Intercept (b_0)	1,282.47	92.596	13.85	<0.0001	1,099	1,466
Distance (b_1)	1.532	0.073	20.961	<0.0001	1.387	1.677

Interpretation:

- Model explains ~82% of total variation in CPL (very good!)
- Both the b_0 and b_1 terms make sense in terms of magnitude and sign and are statistically valid ($p<0.0001$)

How can we improve the model?

ID	Cost Per Load	Distance (miles)	LeadTime (days)	Trailer Length (ft)	Weight (lbs)	Equipment
1	$3,692	1,579	1	53	20,559	REF
2	$3,279	1,298	12	48	17,025	DRY
3	$3,120	1,382	11	48	26,736	REF
4	$3,205	1,033	1	53	26,176	REF
5	$3,188	1,320	3	53	17,994	REF
6	$2,835	1,103	9	53	32,207	REF
7	$2,364	743	1	48	18,589	REF

- What potential additions can we make?
  - Does the equipment type matter?
  - Does lead time have an impact?
  - Does the trailer length have an effect?
  - Does the weight influence rates?
  - Does the CPL have a non-linear relationship with distance? weight?
- Be logical in approach and exploration – always have a hypothesis going in!

Model 2: CPL = f(Dist, Wgt)

Model 2: $CPL = b_0 + b_1(Dist) + b_2(Wgt)$

R^2	0.819
adj R^2	0.815
s_e	281.951
RSS	7,711,126
TSS	42,519,984

Estimation Model

$$CPL = 1282 + 1.532\,(Dist) – 0.003\,(Wgt)$$

	Coefficient	Std Error (s_{bi})	t-stat	p-value	Lower CI (95%)	Upper CI (95%)
Intercept (b_0)	1,354	134.4	10.077	<0.0001	1,088	1,621
Distance (b_1)	1.538	0.074	20.852	<0.0001	1.392	1.685
~~Weight (b_2)~~	-0.003	0.004	-0.74	0.461	-0.011	0.005

Interpretation:

- Model explains ~82% of total variation in CPL (still very good!)
- Note that while R^2 improved from Model 1, the adj R^2 got worse!
- Both the b_0 and b_1 terms make sense in terms of magnitude and sign and are statistically valid (p<0.0001)
- b_2 does not make sense (more weight costs less?) and has poor p-value

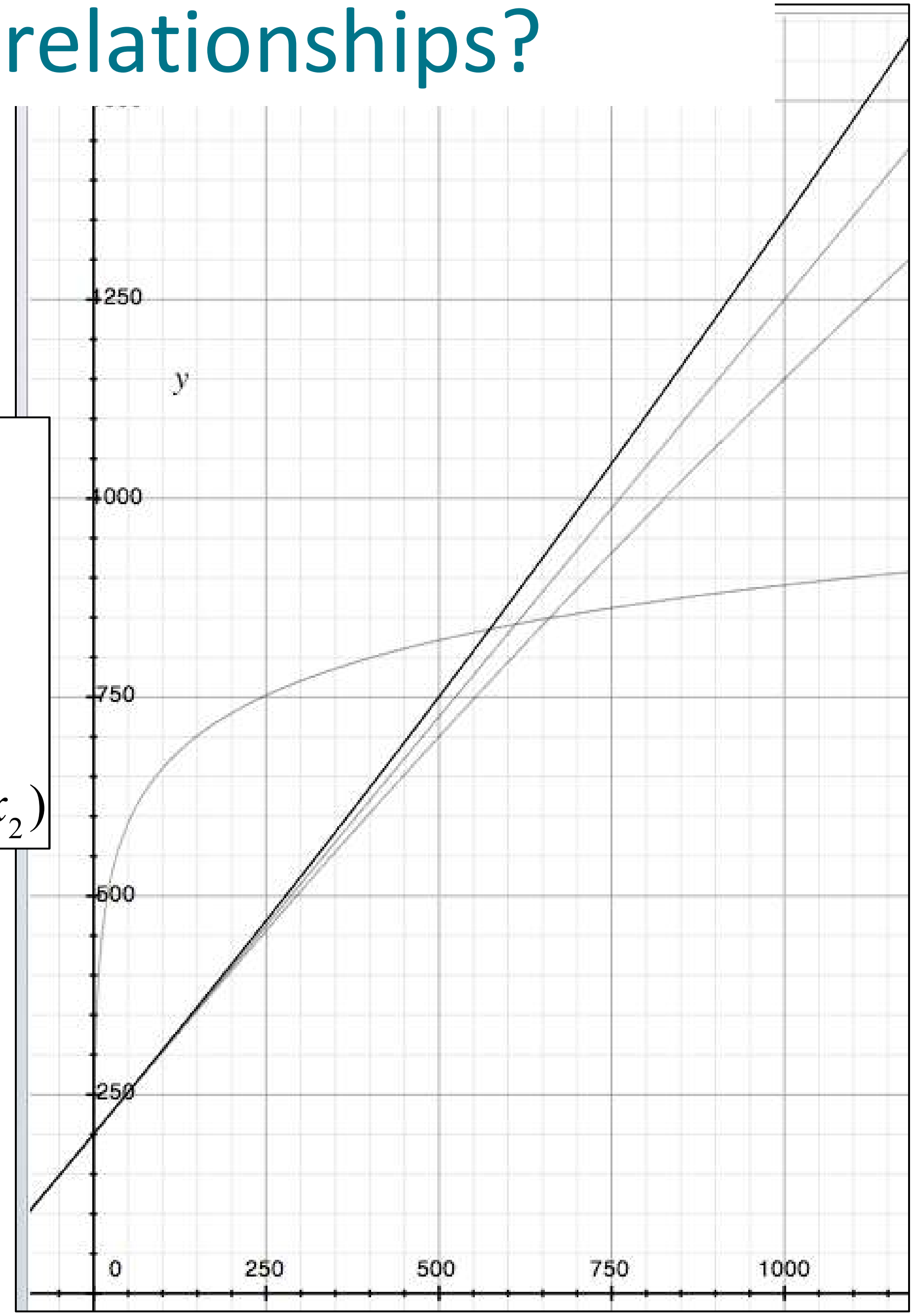
MIT Center for Transportation & Logistics

Linear Transformations

What about non-linear relationships?

Suppose we think that CPL has some non-linear relationship with some of our independent variables . . .

$$y = \beta_0 + \beta_1 x_1$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$
$$y = \beta_0 + \beta_1 \ln(x_1)$$
$$y = ax^b \Rightarrow \ln(y) = \ln(a) + b\ln(x)$$
$$y = a x_1^{b1} x_2^{b2} \Rightarrow \ln(y) = \ln(a) + b_1 \ln(x_1) + b_2 \ln(x_2)$$



Recall,

e=2.71828 . . .

- ln(1) = 0
- ln(e) = 1
- ln(xy) = ln(x) + ln(y)
- ln(x/y) = ln(x) – ln(y)
- ln(x^a) = aln(x)

Modeling Techniques 1

- Linear Transformations
  - We assume a linear model: $y_i = b_0 + b_1x_{i1} + b_2x_{i2} + \ldots + b_kx_{ik}$
  - What if we have a non-linear relationship with an independent variable?
  - OLS Regression is ok as long as the estimated equation **is linear in all of its independent variables**

- Let's Try! Model 3: CPL = f(Dist, $Dist^2$)
  - Testing whether the distance effect tapers off at longer distances
    - Create new variable: DistSq = DIST^2
    - Run Regression: CPL = f(Dist, DistSq)

Model 3: $CPL = b_0 + b_1(Dist) + b_2(DistSq)$

R^2	0.818
adj R^2	0.814
s_e	282.698
RSS	7,752,076
TSS	42,519,984.

Estimation Model

$$CPL = 1282 + 1.532\ (Dist) - 0.00004\ (DistSq)$$

	Coefficient	Std Error (s_{bi})	t-stat	p-value	Lower CI (95%)	Upper CI (95%)
Intercept (b_0)	1,236	271	4.557	<0.0001	698	1,775
Distance (b_1)	1.622	0.503	3.222	0.002	0.623	2.621
Distance^2 (b_2)	-0.00004	0.00022	-0.181	0.857	-0.0005	0.0004

Interpretation:

- Model explains ~82% of total variation in CPL (still very good!), but squared term did not improve the adj R^2 from Model 1.
- b_0 and b_1 still are good (note slight degradation of p-value for b_1)
- b_2 sign and magnitude make sense (at 1000 miles, reduces effect by $40) but has poor p-value

Modeling Categorical Variables

Modeling Techniques 2

- So far, we assumed that independent variables are continuous & ratio scalar.

- What if we have a nominal/categorical independent variable?

- Create a Dummy Variable
  - Suppose you think equipment type impacts CPL?
    - Create a binary dummy variable RefFlag =1 if Refrigerated, =0 o.w.
    - Run the Regression CPL = (Dist, RefFlag)
    - Coefficient of RefFlag captures the differential impact of refrigerated trailers versus Dry vans
- Notes:
  - You do not need to create two dummy variables (RefFlag and DryFlag) – in fact it will fail! This is over-specifying.
  - If we create a DryFlag variable and run CPL=f(Dist, DryFlag) we will get the same estimates for each observation!

Model 4: $CPL = b_0 + b_1(Dist) + b_2(RefFlag)$

R^2	0.822
adj R^2	0.818
s_e	279.634
RSS	7,584,957
TSS	42,519,984.

Estimation Model

$$CPL = 1320 + 1.529\ (Dist) + 84\ (RefFlag)$$

	Coefficient	Std Error (s_{bi})	t-stat	p-value	Lower CI (95%)	Upper CI (95%)
Intercept (b_0)	1,235.858	97.333	12.697	<0.0001	1,042.68	1,429.037
Distance (b_1)	1.529	0.073	21.032	<0.0001	1.384	1.673
RefFlag(b_2)	84.135	57.106	1.473	0.144	-29.204	197.474

Interpretation:

- Model explains ~82% of total variation in CPL (adj R^2 improved from Model 1)
- Both the b_0 and b_1 terms are still fine
- b_2 does not make sense in terms of sign (refrigeration costs more) but has a poor p-value. Perhaps it is more of a function of distance . . . .
- Let's test – new variable RefDist = RefFlag*Dist

Model 5: $CPL = b_0 + b_1(Dist) + b_2(RefDist)$

R^2	0.821
adj R^2	0.817
s_e	280.298
RSS	7,620,996
TSS	42,519,984

Estimation Model

$$CPL = 1320 + 1.529\ (Dist) + 0.06\ (RefDist)$$

	Coefficient	Std Error (s_{bi})	t-stat	p-value	Lower CI (95%)	Upper CI (95%)
Intercept (b_0)	1,283	92.267	13.906	<0.0001	1,099.967	1,466.216
Distance (b_1)	1.495	0.078	19.191	<0.0001	1.341	1.65
RefFlag(b_2)	0.059	0.045	1.304	0.195	-0.031	0.149

Interpretation:

- Model is good – b_0, and b_1 are fine.
- b_2 is problematic – sign and magnitude are reasonable – but p-value is bad.
- Should we keep it? This is where we get more art than science. If important, then OK – but always state the p-value so the user of the model understands its strength/weakness.

Model 6: $CPL = b_0 + b_1(Dist) + b_2(SameDay)$

R^2	0.828
adj R^2	0.824
s_e	274.655
RSS	7,317,250
TSS	42,519,984

SameDay = 1 if LdTime=0, =0 o.w.

Estimation Model

CPL = 1238 + 1.552 (Dist) + 233 (SameDay)

	Coefficient	Std Error (s_{bi})	t-stat	p-value	Lower CI (95%)	Upper CI (95%)
Intercept (b_0)	1,238	92.308	13.407	<0.0001	1,054	1,420
Distance (b_1)	1.552	0.072	21.602	<0.0001	1.409	1.694
SameDay(b_2)	233	96.609	2.408	0.018	40.9	424.4

Interpretation & Insights:

- There are many ways to model the Lead Time effect:
  - Continuous – each day adds a linear cost
  - OverWeek = 1 if LdTime>7 days, =0 o.w.
- Quantified potential financial benefit for changing practice (SameDay tenders)

Model Validation II

Regression Assumptions

1. There is a population regression line that joins the mean of the dependent variables.
2. This implies that the mean of the error is 0.
3. The variance of the dependent variable is constant for all values of the explanatory variables (Homoscedasticity)
4. The dependent variable is normally distributed for any value of the explanatory variables.
5. The error terms are probabilistically independent.

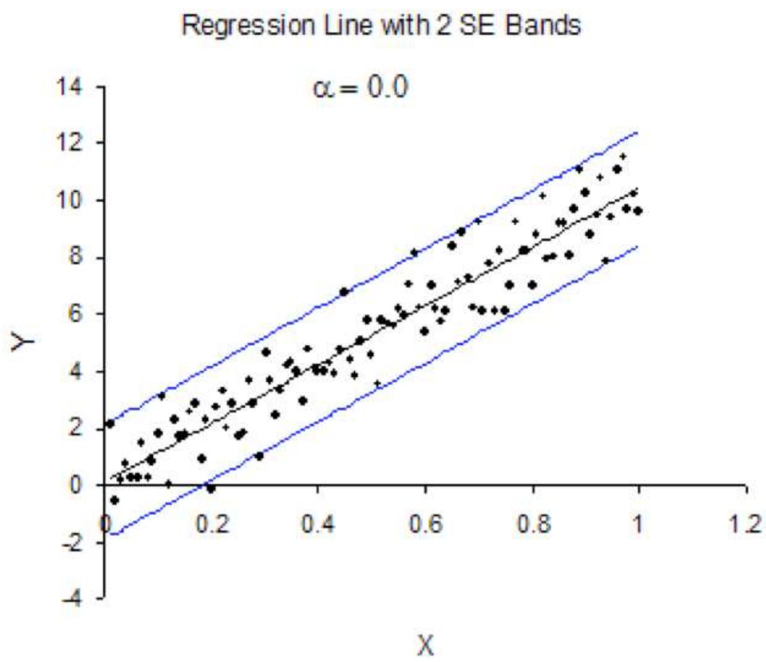
Model Validation – Multi-Collinearity & Autocorrelation

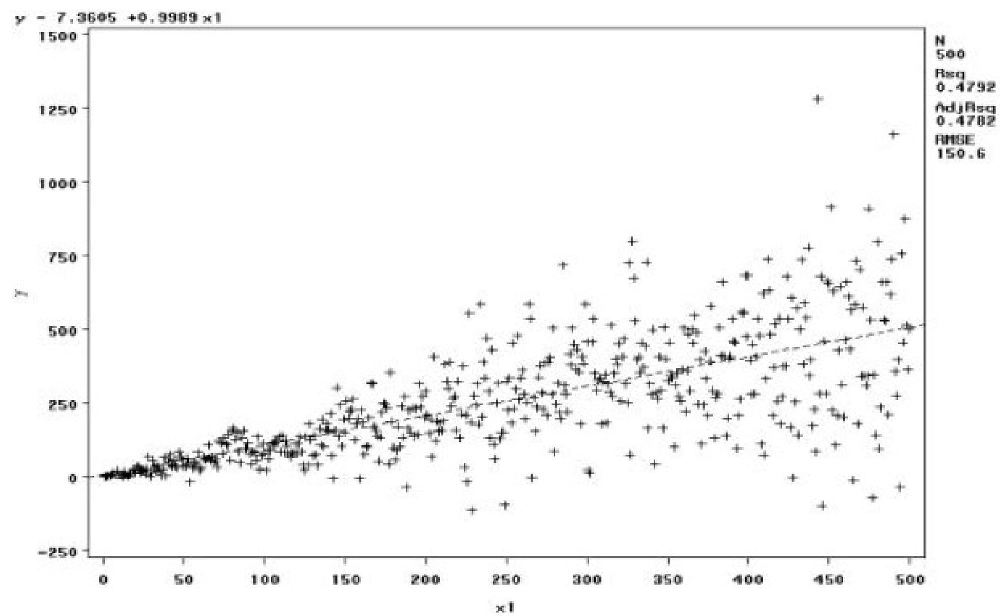
- Multi-Collinearity – Are the independent variables correlated?
  - When two or more independent variables are highly correlated
  - Model might have high R^2 but explanatory variables might fail t-test
  - Can often result in strange results for correlated variables
  - Check for correlation and remove correlated independent variables

- Autocorrelation – Are the residuals not independent?
  - Errors are supposed to be identical and independently distributed (iid)
  - Typically a time series issue – plot variables over time to see trend
  - If they are not independent, they are autocorrelated

Model Validation - Heterscedasticity

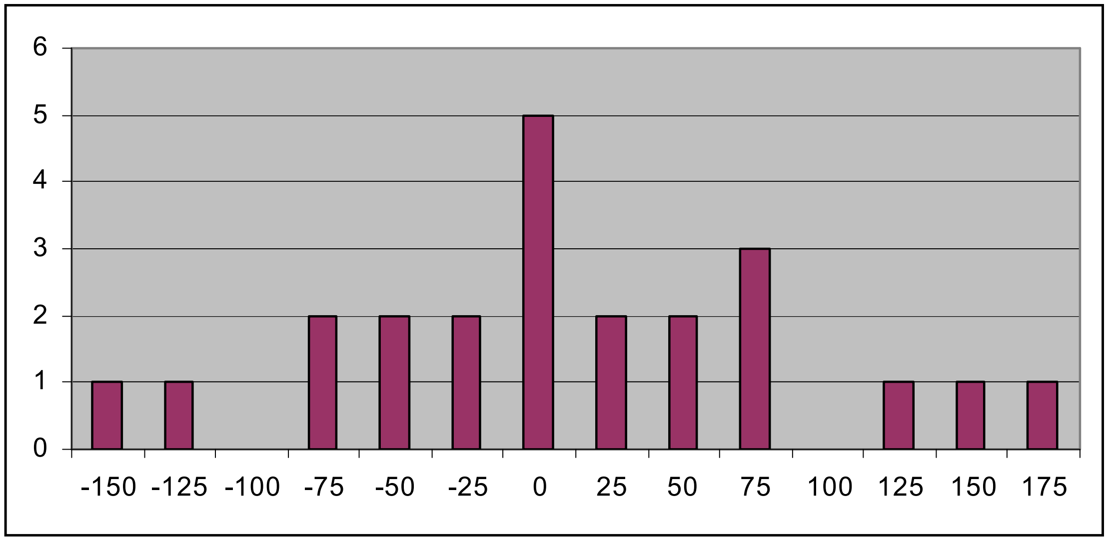
- Heteroscedasticity – Does the standard deviation of the error terms differ for different values of the independent variables?
  - Observations are supposed to have the same variance
  - How do the residuals behave over the ind. var.'s?
  - Examine scatter plots and look for "fan-shaped" distributions
  - Fixes (weighted least squares regression, others – beyond scope)





Model Validation

- Linearity – Is the dependent variable linear with independent variables?
  - With one ind. variable, scatter plots work
  - More than one ind. variable – look at R^2
- Normality of Error Terms – Are the error terms distributed Normally?
  - We have assumed that $e \sim N(0,\sigma)$
  - Look at a histogram of the residuals
  - There are more formal tests; e.g., Chi-Square or Kolmogorov–Smirnov tests



Key Points from Lesson

The Practice of Regression

1. Choose which independent variables to include in the model, based on common sense and context specific knowledge.
2. Collect data (create dummy variables in necessary).
3. Run regression -- the easy part.
4. Analyze the output and make changes in the model -- this is where the action is.
   - Using fewer independent variables is better than using more
   - Always be able to explain or justify inclusion (or exclusion) of a variable
   - Always validate individual explanatory variables (p-value)
   - There is more art than science to these models

Regression Analysis Checklist

- **Linearity**: scatter plot, common sense, and knowing your problem
- **Signs of Regression Coefficients**: do they agree with intuition?
- **t-statistics**: are the coefficients significantly different from zero?
- **adj R^2**: is it reasonably high in the context?
- **Normality**: plot histogram of the residuals
- **Heteroscedasticity**: plot residuals with each x variable
- **Autocorrelation**: "time series plot"
- **Multicollinearity**: compute correlations of the x variables
  - If |corr|>.70 – you might want to remove one of the variables

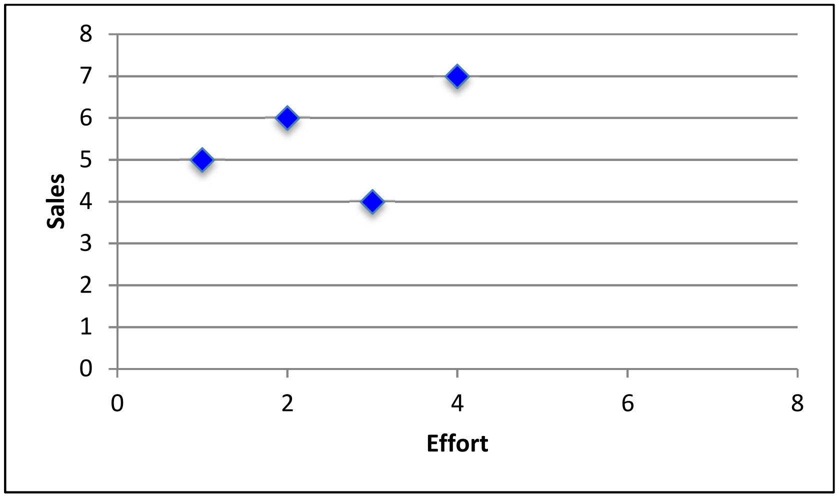
Practical Concerns – Beware Over-Fitting

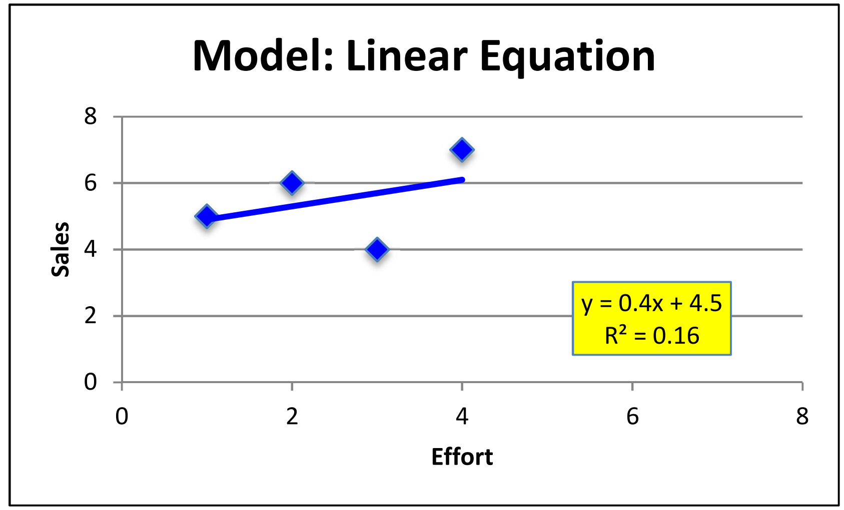
- Suppose you are given the following data,

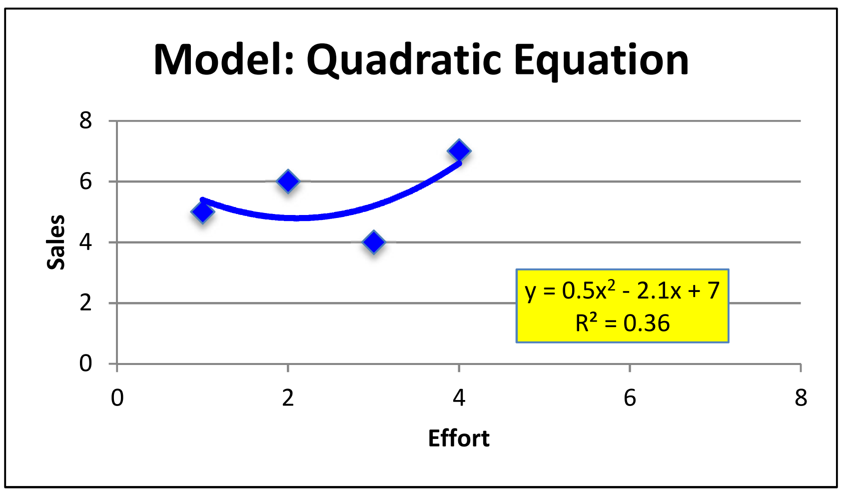
Sales	Effort
5	1
4	3
7	4
6	2

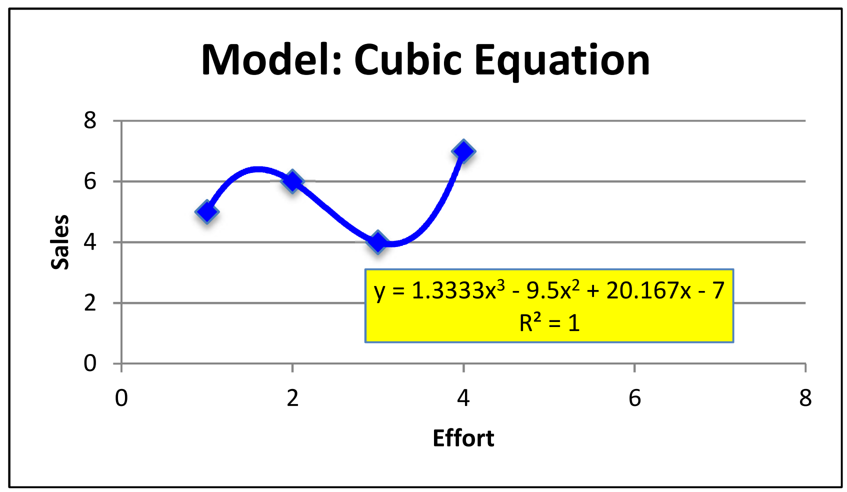
- How would you find the “best” fit curve assuming Sales = f(Effort)?

Results – which is the best fit?









41

Modeling Issues and Tips

- Don't confuse causality and relationship
  - Statistics find and measure relationships – not causality
  - User must try to explain the causality
- Don't be a slave to R^2 – model must make sense
  - Look at adjusted R^2 to compare between models
- Simple is better (avoid over-specifying)

  Rule of thumb $n \geq 5(k+2)$ where
  n= num obs and k=num of ind variables
- Avoid extrapolating out of observed range
- Non-linear relationships can be modeled through data transformations (ln, sqrt, 1/x, Multiply)

Questions, Comments, Suggestions?
Use the Discussion Forum!

"Wilson and Dexter waiting patiently to regress their way out of their holiday costumes"

caplice@mit.edu
ctl.mit.edu