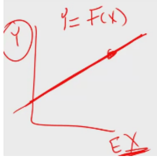
Session 6: Categorical and Linear Models

6.1 – Introduction to Linear Models

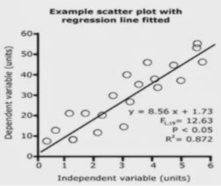
Bringing Models to Data:

- **Categorical**

- **Linear Models**
  $\boldsymbol{(x) = ax + b)},$



- **Fitting Lines to Data**



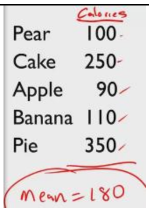
- **Non Linear Models**

- **Big Coefficient**
  $y = a_1X_1 + a_2X_2$ Does a_1 or a_2 dominate?
  Only works where we have the data. Often we don't have.

- **New Reality**, where we don't have the data but things are a lot better than we have today.

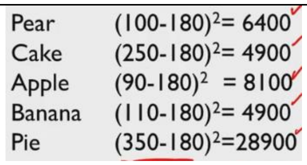
6.2 – Categorical Models

Bin Reality into different categories:
Example: Amazon early on – Info company or Delivery company.
"Lump to Live" people categorize to simplify the world, making faster decision.
Example: What to eat? Green or Not Green.
Broccoli, Grasshopper, Banana, Candy bar, Orange , Asparagus, Pear, Strawberry.
Simple rule, not always selective enough.



	Calories
Pear	100
Cake	250
Apple	90
Banana	110
Pie	350

Mean = 180



Pear	$(100-180)^2 = 6400$
Cake	$(250-180)^2 = 4900$
Apple	$(90-180)^2 = 8100$
Banana	$(110-180)^2 = 4900$
Pie	$(350-180)^2 = 28900$

Point here is that to reduce the large deviations from the mean, we should apply categorization to the grouping to reduce total variation. *Total variation* is the sum of square of the differences. $(\sum variation\ above\ = \ 53{,}200)$

Quiz: 5 students take an exam: Amy scores 100; Ben scores 90; Carla scores 70; Demitri scores 65; Emile scores 85. What is the Total Variation of these scores? (1) 82, (2) 324, (3) 830, (4) 940

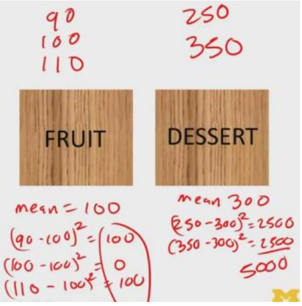
Analysis: $\mu = \frac{1}{5}(100 + 90 + 70 + 65 + 85) = 82$ and

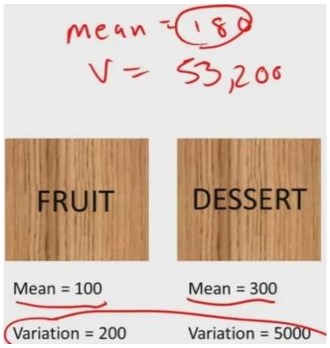
$$\sigma^2 = \sum\{(100 - 82)^2 + \ 90 - 82)^2 + \ 70 - 82)^2 + \ 65 - 82)^2 + \ 85 - 82)^2\} = 830$$

Ans: (3) 830

Breaking into categories by food type:
Fruit (Pear, Apple, Banana) and
Dessert (Cake, Pie)

In the chart:
Fruit (mean=100, variation = 200)
Dessert (mean=300, variation=5,000)





How much did I explain?
Combined fruit & dessert variation = 53,200. Fruit variation = 200, Dessert variation = 5,000
Explanation:

$$\frac{53{,}200 - \ 5{,}000 + 200)}{53{,}200} =$$

$$\frac{48{,}000}{53{,}200} = 0.90226$$

R-squared: % variation explained

$$1 - \frac{total\ subgroup\ variation}{total\ grouped\ variation} = 1 - \frac{5200}{53{,}200} = 90.2\%$$

- R-squared near 1 ➔ model explains a lot
- R-squared near 0 ➔ model explains little.
- Data quality impacts a good model's R-squared

Many subcategories: (fruit, dessert, vegetables, grains).
Experts have a lot of 'right' categories.

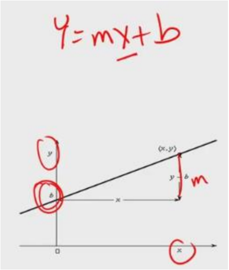
Finally: Even if you have a lot of explanation, it doesn't always follow you have a good model. ***"Correlation is not causation".***
Example: Schools with equestrian teams happen to do better. But it may be due to other factors such as parental support in these districts as opposed to the criteria that was used (money, class sizes, etc.)

Summary: Simplest method of explanation is by categorization and the R-squared metric. More categories means there are more possible explanations. The next section on linear models basically creates a box for each value of x.

6.3 – Linear Models

Linear Model:
Variable X (independent) and variable Y (dependent) that is a function of x,
$y = F(x)$
Used for:

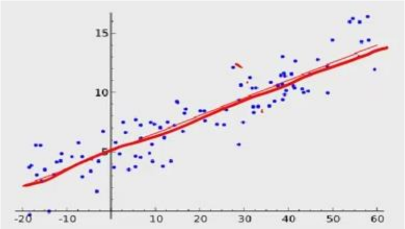
- Prediction
- Understand data



Example: Purchase TV where x = length of diagonal, and y = cost of TV.
Linear Model: $Cost = 15 * (length) + 100$
Concerns:
Sign – (does y increase or decrease in x?),
Magnitude – (how much does y increase for each one unit increase in x?)
How much would a 30-in TV cost? From model
$C = 15(30) + 100 = \$550$

Quiz: According to a psychiatrist, the probability that one developers haa mental disorder is linearly related to the number of stressful life events one experiences. This psychiatrist uses an equation to predict mental disorders: P(mental disorder)=0.1(# stressful life events)+.005. If Jeff has had 5 stressful life events, what is the probability he develops a mental disorder, according to this equation? (1) 50.5%, (2) 60%, (3) 49.5%, (4) 5.05%
Analysis: $P = 0.1(5) + 0.005 = 0.505 = 50.5\%$
Ans: (1) 50.5%

Linear Model:
Best linear fit to a data set. Offset from line is a measure of variation.



Robyn Dawes 1979: *"The Robust Beauty of Improper Linear Models in Decision Making"*

Example: 43 bank loan officers predict which 30 of 60 firms would go bankrupt. They see the financial statements. Bankers 75% accurate, Linear Model – ratio of assets to liabilities 80% accurate.

Mehl (1954) 20 studies of clinicians, Sawyer (1966) 45 studies of predictions in the social world.

Experts NEVER did significantly better than the linear models.

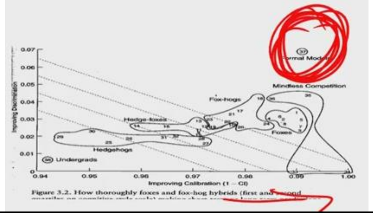
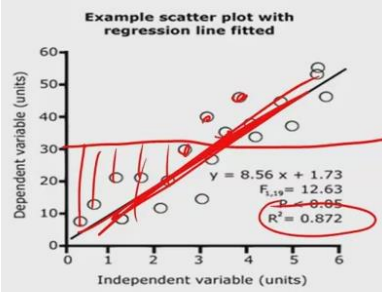


Figure 3.2. How thoroughly foxes and fox-hog hybrids (first and second

Recall Tetlock: Formal models do better than experts.

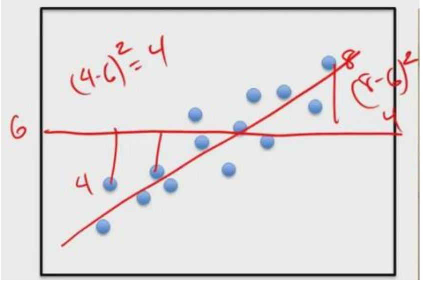
6.4 – Fitting Lines to Data

Recall R-squared
Note mean of this data and the variation compared to variation of best fit line $R^2 = 0.872$ Explains 87.2% of variation.

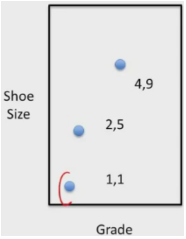


Data:
mean = 6
Find variation from mean.

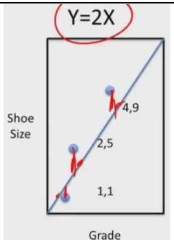
Compare with variation from line.



Example: (grade, shoe size)
Note variation of dependent variable (shoe size, mean=5)
(1,**1**) - $(1-5)^2 = 16$
(2,**5**) - $(5-5)^2 = 0$
(4,**9**) - $(9-5)^2 = 16$
Variation = 32



Find best linear fit: Assume y=2x.
(x,y,2x)
(1,**1**,**2**) - $(2-1)^2 = 1$
(2,**5**,**4**) - $(4-5)^2 = 1$
(4,**9**,**8**) - $(8-9)^2 = 1$
Variation = 3 so model explains
$1-\frac{Var_l}{Var_m} = 1-\frac{3}{32} = 90.625\ \%$



Best Line calculation:

$y = mx + b$

$x = 1$: $(m+b-1)^2 = m^2 + 2mb + b^2 - 2m - 2b + 1$

$x = 2$: $(2m+b-5)^2 = 4m^2 + 4mb + b^2 - 20m - 10b + 25$

$x = 4$: $(4m+b-9)^2 = 16m^2 + 8mb + b^2 - 72m - 18b + 81$

$sum = 21m^2 + 14mb + 3b^2 - 94m - 30b + 107$

Find b, m:
$b = -1$, & $m = {}^{8}/_{3}$
(x, y, model) $y = \frac{8}{3}x - 1$
$\left(1,1,{}^{5}/_{3}\right)$: $\left(\frac{5}{3}-1\right)^2 = \left({}^{2}/_{3}\right)^2$
$\left(2,5,{}^{13}/_{3}\right)$: $\left(\frac{13}{3}-\frac{15}{3}\right)^2 = \left({}^{2}/_{3}\right)^2$
$\left(4,9,{}^{29}/_{3}\right)$: $\left(\frac{29}{3}-\frac{27}{3}\right)^2 = \left({}^{2}/_{3}\right)^2$
$sum = 3\left(\frac{4}{9}\right) = {}^{4}/_{3}$

R-squared:
$1 - \frac{\left({}^{4}/_{3}\right)}{32} = 95.8\%$

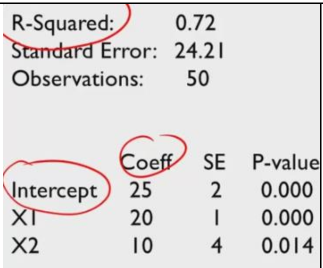
Multiple variables: $y = ax_1 + bx_2 + c$
Recall: Sign and magnitude of x coefficients.
Example:
Y = Test Score, Q = IQ, T = Teacher Quality, Z = Class Size,
$Y = a + bQ + cT + dZ$
Note: We expect: $b > 0$, $c > 0$, $d < 0$, but may be wrong
Studies:
Class Size (78 studies, d>0 in 4, d<0 in 13, d~0 in 61)
Kindergarten Teacher (teacher quality matters a lot – Chetty (2010))

Summary: Linear models help us explain variation in data. Sign and Magnitude are indicators of the influence of corresponding independent variable.

6.5 – Reading Regression Output

Understanding the Data:
Multiple variables:
$y = m_1x_1 + m_2x_2 + b$

Chart data implies:
$y = 20x_1 + 10x_2 + 25$



```
R-Squared:        0.72
Standard Error:  24.21
Observations:       50

             Coeff   SE   P-value
Intercept     25      2    0.000
X1            20      1    0.000
X2            10      4    0.014
```

SE=standard error of coefficient, P-value indicates probability of sign error (1.1% in chart for m_2).

Example:
Y = Test Score, T = Teacher Quality, Z = Class Size,
$Y = cT + dZ + b$ and expect $c > 0$, $d < 0$

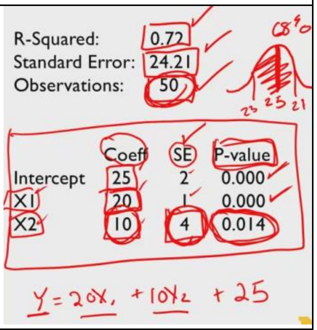
SE – standard error in coefficients
'Intercept', SE =2 ➔ b∈ [23,27] @ 68% confidence
m_1, SE=1 ➔ $m_1 \in [19,21]$ @ 68%
m_2, SE=4 ➔ $m_2 \in [6,16]$ @ 68%, note >0 but large error

Quiz: A researcher attempts to use a linear model to understand tennis. In her model, points the dependent variable is "points won" and there are several independent variables, such as: serve speed, backhand speed, accuracy, player height and dominant hand. The researcher finds that left-handedness has a positive coefficient and a p-value of 0.20. What does this mean? (1) The coefficient indicates that left-handedness is a disadvantage to tennis players, and the p-value confirms this. (2) The coefficient indicates that left-handedness is an advantage to tennis players, but the p-value indicates that there is some chance that left-handedness is a disadvantage. (3) The coefficient indicates that left-handedness is always an advantage to tennis players, and the p-value confirms this. (4) The coefficient indicates that left-handedness has the greatest influence on points won of any variable, but the p-value indicates that there is some chance that this is not the case.

Ans: (2)

Explanation: *The positive coefficient means that left-handedness increases the points won. In other words, being left-handed helps the tennis player. The p-value means there is a 20% chance that this is wrong, or in other words, that being left-handed actually has a negative effect.*

Summary:
R-Squared – how much of the data is explained,
Observations – how many data points,
Standard Error – How much variation in the data to begin with
Linear Model – Intercept (sign and magnitude, SE and probability of sign error)
Linear Model – Coefficients of independent variables (sign and magnitude, SE & probability of sign error)
Recall: *Models are generally better than we are!*



R-Squared: 0.72
Standard Error: 24.21
Observations: 50

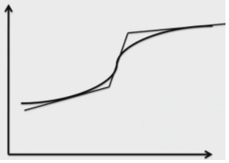
	Coeff	SE	P-value
Intercept	25	2	0.000
X1	20	1	0.000
X2	10	4	0.014

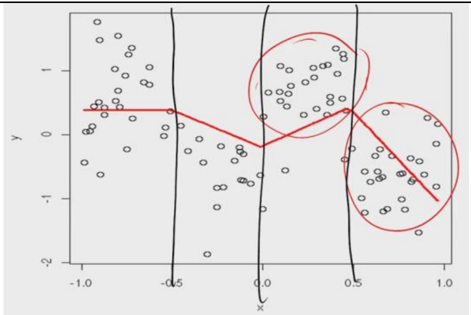
6.6 – From Linear to Nonlinear

Neuman: *Study of non-linear functions is like the study of non elephants.*

How to handle nonlinear functions?

1. ***Piecewise linear approximations.***





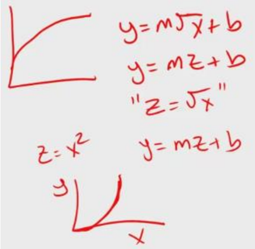
Example:
break clustered data into groups and piecewise linear fit in each section (spline fit).

2. ***Add non-linear terms:***

Can substitute to create a linear form of a non-linear relationship.

e.g., in $y = a\sqrt{x} + b$ let $z = \sqrt{x}$

Then we have a linear form $y = az + b$



6.7 – The Big Coefficient vs The New Reality

Big Coefficient: does one coefficient dominate in $y = a_1x_1 + a_2x_2 + b$? if so, then that has the major influence in explaining y.

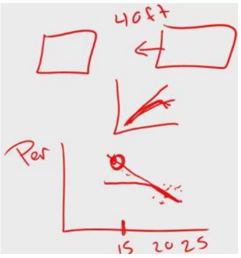
Evidence Based: (1) medicine, (2) philanthropy, (3) education, (4) management
In each, look for dominant influence.

Process:
(1) construct model, (2) gather data, (3) identify important variables, (4) change those variables to test and/or optimize.

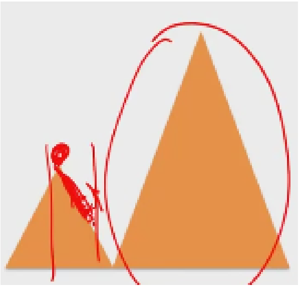
Big Data approach: (1) gather data, (2) find pattern(s), (3) identify important variables, (4) change those variables

Observations: (1) Appears to ignore model step, BUT big data does not obviate the uses of models. Identification of a pattern does not explain where it came from, the why. Correlation is not Causation, (2) Linear models tell sign and magnitude of changes in independent variables within data range.

Examples:
Feedbacks: (1) antilock brakes on cars. People adapt and may shorten safe distance between cars. (2) Class size-Performance: data in the range of [20,30] students may not extrapolate to classes of smaller sizes.



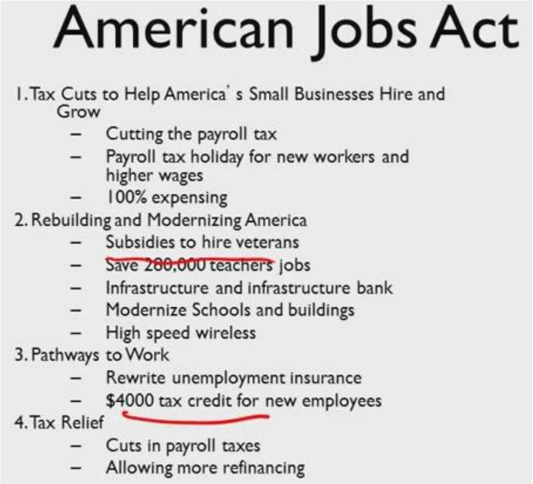
Multiple Peaks: Local maximum (or minimum) can mislead an optimization process.



The New Reality: Need to consider that some big change is necessary to avoid local maximum trap.

Big Coefficient:
Tax Cigarettes, Increase HOV lanes, Oat Bran Pretzels
New Reality:
Universal Health Care, Huge Rail System, Fitness Regime

Big Coefficient Logic: $447 B



American Jobs Act

1. Tax Cuts to Help America' s Small Businesses Hire and Grow
   - Cutting the payroll tax
   - Payroll tax holiday for new workers and higher wages
   - 100% expensing
2. Rebuilding and Modernizing America
   - Subsidies to hire veterans
   - ~~Save 280,000 teachers~~ jobs
   - Infrastructure and infrastructure bank
   - Modernize Schools and buildings
   - High speed wireless
3. Pathways to Work
   - Rewrite unemployment insurance
   - $4000 tax credit for new employees
4. Tax Relief
   - Cuts in payroll taxes
   - Allowing more refinancing

New Reality Logic: 1956 - US Highway system with 41,000 miles of highways for $25B (CPI: $207B today or at $10 million/mile equivalent in today's dollar of $410B).

Evidence-based methods really useful. Put your money on the big coefficients but remember these can blind you to new reality opportunities. ***Models can help optimize in the box as well as think outside the box.***

Quiz: Big coefficients may be limiting if they prevent us from achieving a new, better reality. However, programs intended to create new realities are not always successful. Which of the following is a possible explanation? (1) New realities often occur when the coefficients of linear regressions are miscalculated. (2) New realities often don't have a lot of evidentiary support since they are a new way of doing things. (3) New realities focus on existing empirical data rather than trying something completely different. (4) None of the above, new realities are always better than using big coefficients.
Ans: (2)
Explanation: *The correct response is that we don't have much empirical support for new realities, by definition. The third option, that new realities fail to try something completely different, is a limitation of big coefficients rather than of new realities. The other two options are incorrect as well: there is no evidence that miscalculated coefficients somehow often lead to new realities; nor can we say that new realities are always better than big coefficients.*