Chapter 4:

Statistical Approaches to Estimation and Prediction

Univariate
- Point Estimation
- Confidence Interval Estimation

Bivariate: Linear Regression

Multivariate: Multiple Regression

Review: Data Mining Tasks

Task	We Learn about This Task in
Description	Chapter 3: Exploratory Data Analysis
Estimation	Chapter 4: Statistical Approaches to Estimation and Prediction
Prediction	Chapter 4: Statistical Approaches to Estimation and Prediction
Classification	Chapter 5: *k*-Nearest Neighbor Algorithm
	Chapter 6: Decision Trees
	Chapter 7: Neural Networks
Clustering	Chapter 8: Hierarchical and *k*-Means Clustering
	Chapter 9: Kohonen Networks
Association	Chapter 10: Association Rules

Descriptive model

2

Introduction

- Numerical variables
- statistical analysis have been performing data mining for **over a century**
- Univariate, based on one variable
  - Point Estimation, measure of center (mean, median, mode), measure of location (percentile, quantile)
  - Measure of spread or measure of variability
  - Confidence Interval Estimation

3

Measure of Spread

Price/Earnings Ratios for Five Stocks in Each of Two Portfolios

Stock Portfolio A	Stock Portfolio B
1	7
11	8
11	11
11	11
16	13

For both: The mean P/E ratio is 10, the median is 11, and the mode is 11

The measures of variability for portfolio A should be larger than those of B

4

Measure of Spread

- Range
  - (maximum - minimum)
- Standard deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

- Mean absolute deviation

$$MAD = \frac{\sum |X_i - \bar{X}|}{n}$$

- Interquartile range
  - IQR = Q3 – Q1

5

Statistical Inference

Statistical inference consists of methods for estimating and testing hypotheses about **population** characteristics based on the information contained in the **sample**.

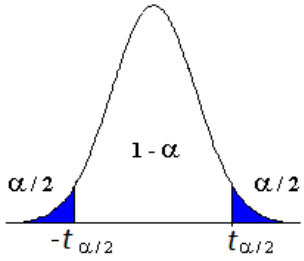
Use Observed Sample Statistics to Estimate Unknown Population Parameters

	Sample Statistic	...Estimates...	Population Parameter
Mean	$\bar{x}$	⟶	μ
Standard deviation	s	⟶	σ
Proportion	p	⟶	π

6

How confidence are we in our estimates?

- Confidence Interval
  - point estimate ± margin of error
- Example:
  - t-interval is used when the population is normal or the sample size is large

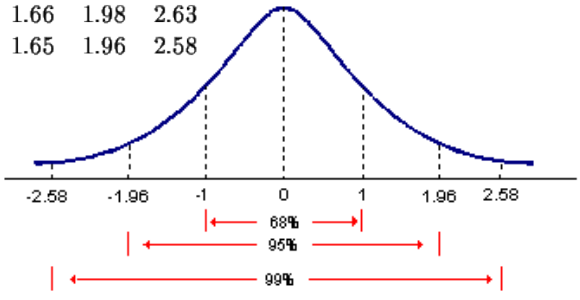


$$\bar{x} \pm t_{\alpha/2}(s/\sqrt{n})$$

7

T-interval

$1-\alpha$	.80	.90	.95	.99
t : n-1=5	1.48	2.02	2.57	4.03
t : n-1=15	1.34	1.75	2.13	2.95
t : n-1=25	1.32	1.71	2.06	2.79
t : n-1=35	1.31	1.69	2.03	2.72
t : n-1=50	1.30	1.68	2.01	2.68
t : n-1=100	1.29	1.66	1.98	2.63
t : n-1=500	1.28	1.65	1.96	2.58



8

T-interval
Example: Customer Service Calls

Customer Service Calls

Statistics

Count	3333
Mean	1.563
Sum	5209.000
Standard Deviation	1.315
Standard Error of Mean	0.023
Median	1.000
Mode	1.000

$$\bar{x} \pm t_{\alpha/2}(s/\sqrt{n})$$
$$1.563 \pm 1.96(1.315/\sqrt{3333})$$
$$1.563 \pm 0.045$$
$$(1.518, 1.608)$$

We are 95% confident that the population mean number of customer service calls for all customers falls between 1.518 and 1.608 calls.

Percentage $(1-\alpha)$	.68	.80	.90	.950	.954	.99
t	1.00	1.28	1.64	1.96	2.00	2.58

9

TOC

- Univariate
  - Point Estimation
  - Confidence Interval Estimation
- **Bivariate: Linear Regression**
- Multivariate: Multiple Regression

10

Bivariate Method: Linear Regression

- Bivariate:
  - Using the value of one variable to estimate the value of a different variable
- Linear Regression:
  - An approach for modeling the linear relationship between predicator (x) and response (y) variables.

11

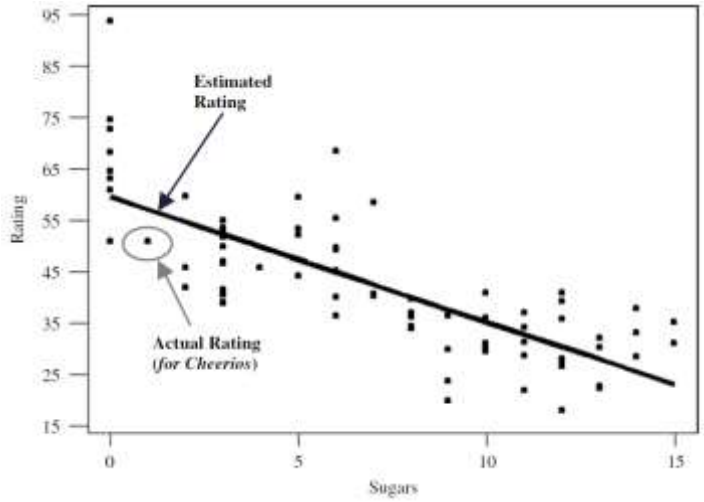
Cereal dataset

(from *Data and Story Library,* www.lib.stat.cmu.edu/DASL**)**

Cereal Name	Manuf.	Sugars	Calories	Protein	Fat	Sodium	Rating
100% Bran	N	6	70	4	1	130	68.4030
100% Natural Bran	Q	8	120	3	5	15	33.9837
All-Bran	K	5	70	4	1	260	59.4255
All-Bran Extra Fiber	K	0	50	4	0	140	93.7049
Almond Delight	R	8	110	2	2	200	34.3848
Apple Cinnamon Cheerios	G	10	110	2	2	180	29.5095
Apple Jacks	K	14	110	2	0	125	33.1741
Basic 4	G	8	130	3	2	210	37.0386
Bran Chex	R	6	90	2	1	200	49.1203
Bran Flakes	P	5	90	3	0	210	53.3138
Cap'n'Crunch	Q	12	120	1	2	220	18.0429
Cheerios	G	1	110	6	2	290	50.7650
Cinnamon Toast Crunch	G	9	120	1	3	210	19.8236
Clusters	G	7	110	3	2	140	40.4002
Cocoa Puffs	G	13	110	1	1	180	22.7364

12

Scatter plot of nutritional rating versus sugar content



13

Regression Analysis: Rating versus Sugars

The regression equation is
Rating = 59.4 - 2.42 Sugars

Predictor	Coef	SE Coef
Constant	59.444	1.951
Sugars	-2.4193	0.2376

S = 9.162 R-Sq = 58.0%

Unusual Observations

Obs	Sugars	Rating	Fit	Residual
2	0.0	93.70	59.44	34.26
32	6.0	68.40	44.93	23.48

Predicted Values for New Observations

New Obs	Fit	95.0% CI	95.0% PI
1	57.02	(53.53,60.52)	(38.44,75.61)

14

Regression Analysis: Rating versus Sugars

- The estimated regression equation (ERE): $\hat{y}$ = 59.4 – 2.42(*sugars*).
- *coefficients*:
  - Y-intercept or constant b_0 = 59.4
  - *Slope* b_1 = –2.42
- *SE coef:* the standard errors of the coefficients, a measure of the variability of the coefficients.
- *S*, the standard error of the estimate, indicates a measure of the size of the "typical" error in prediction.
- *R-squared* is a measure of how closely the linear regression model fits the data, (90 to 100%: a very nice fit)
- Fit: estimated rating using the regression equation
- Residual, prediction error or estimation error: (y-$\hat{y}$)
- CI: confidence Interval
- PI: Prediction Interval

15

Confidence Intervals for the Mean Value of *Y* Given *X*

$$\text{point estimate} \pm \text{margin of error} = \hat{y}_p \pm t_{\alpha/2}(s)\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

where

x_p = the particular value of x for which the prediction is being made

$\hat{y}_p$ = the point estimate of y for a particular value of x

$t_{\alpha/2}$ = a multiplier associated with the sample size and confidence level

$s = \sqrt{\text{MSE}} = \sqrt{\text{SSE}/n - 1}$ = the standard error of the estimate

SSE = the sum of squared residuals

16

Prediction Intervals for a randomly Chosen Value of *Y* Given *X*

- It is "easier" to predict the class average on an exam than it is to predict a randomly chosen student's score.
- *Prediction intervals* are used to estimate the value of a randomly chosen value of *y*, given *x*.
- Clearly, this is a more difficult task
- => intervals of greater width (lower precision) than confidence intervals for the mean

$$\text{point estimate} \pm \text{margin of error} = \hat{y}_p \pm t_{\alpha/2}(s)\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

17

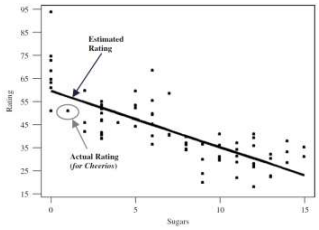
Regression Analysis: Rating versus Sugars

```
The regression equation is
Rating = 59.4 - 2.42 Sugars

Predictor      Coef   SE Coef
Constant     59.444     1.951
Sugars      -2.4193     0.2376

S = 9.162     R-Sq = 58.0%

Unusual Observations
Obs    Sugars    Rating      Fit    Residual
  2       0.0     93.70    59.44       34.26
 32       6.0     68.40    44.93       23.48

Predicted Values for New Observations

New Obs    Fit        95.0% CI          95.0% PI
1        57.02  ( 53.53,60.52)   ( 38.44,75.61)
```

18

Danger of Extrapolation

- *Extrapolation*: making predictions for *x* values lying outside the range of regression,
- Example: sugar = 30 grams
- Estimated Regression Equation:
  - *Rating*: $\hat{y}$ = 59.4 – 2.42(*sugars*)
  - =59.4 – 2.42(30) = –13.2.
- While minimum rating is 18 !!!
- Negative rating !!



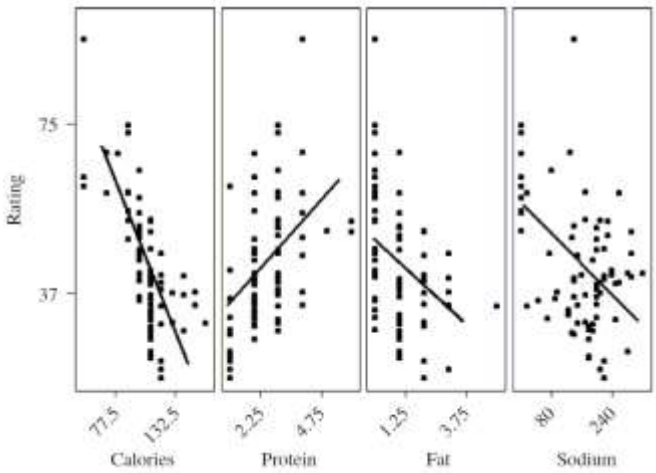
19

TOC

- Univariate
  - Point Estimation
  - Confidence Interval Estimation
- Bivariate: Linear Regression
- **Multivariate: Multiple Regression**

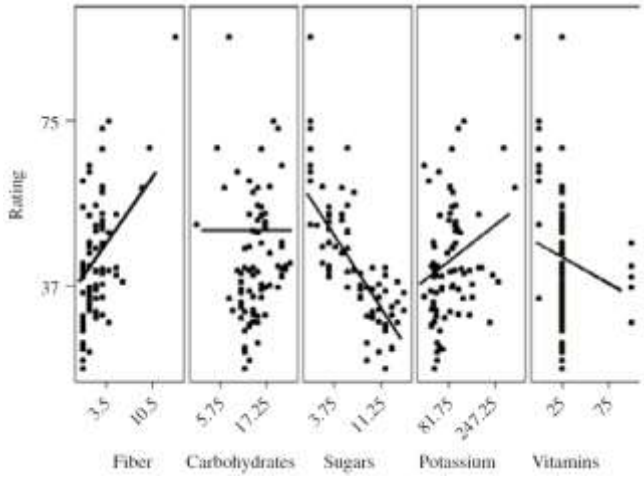
20

Scatter plot of *rating* versus *calories, protein, fat,* and *sodium*



21

Scatter plot of *rating* versus *fiber, carbohydrates, sugars, potassium,* and *vitamins*



22

Cross correlation

Correlation Coefficients for All Variables

	Rating	Calories	Protein	Fat	Sodium	Fiber	Carbohydrates	Sugars	Potassium
Calories	−0.689								
Protein	0.471	0.019							
Fat	−0.409	0.499	0.208						
Sodium	−0.401	0.301	−0.055	−0.005					
Fibre	0.577	−0.291	0.506	0.026	−0.071				
Carbos	0.050	0.255	−0.125	−0.315	0.357	−0.357			
Sugars	−0.762	0.564	−0.324	0.257	0.096	−0.137	−0.351		
Potass	0.380	−0.067	0.549	0.193	−0.033	0.905	−0.354	0.22	
Vitamins	−0.241	0.265	0.007	−0.031	0.361	−0.036	0.257	0.122	0.021

Negative correlation: negative slope in regression line
Zero correlation: uncorrelated
One correlation: high correlated (multicollinearity)

23

Multicollinearity

- Definition: a condition where some of the predictor variables are correlated with each other. it leads to
  - instability in the solution space,
  - leading to possible incoherent results.
  - Overemphasizing a particular component of the model, like it's double counted.
- Solutions:
  - Simply elimination (*potassium* in Cereal dataset)
  - Supervised dimension reduction approaches like LDA
  - Unsupervised dimension reduction approaches such as PCA, ICA, …

24

Multiple regression

The regression equation is

Rating = 55.9 − 0.225 Calories + 2.88 Protein − 2.00 Fat − 0.0546 Sodium
+ 2.57 Fiber + 1.08 Carbos − 0.823 Sugars − 0.0514 Vitamins

Predictor	Coef	SE Coef
Constant	55.9047	0.8421
Calories	−0.22456	0.01551
Protein	2.8824	0.1626
Fat	−2.0048	0.1857
Sodium	−0.054647	0.001609
Fiber	2.57151	0.06505
Carbos	1.07504	0.06093
Sugars	−0.82343	0.06189
Vitamins	−0.051422	0.005802

S = 1.015 R-Sq = 99.5% R-Sq(adj) = 99.5%

25

Comparisons

	R-sq	SE
Bivariate Regression	58.0%	9.162
Multiple Regression	99.5%	1.015

- **The uncorrelated variable (Carbohydrates) did not eliminated**
- **Elimination of Carbohydrate leads to an increase in SE and decrease in R-sq**

We can trust the results only when we verify our assumptions

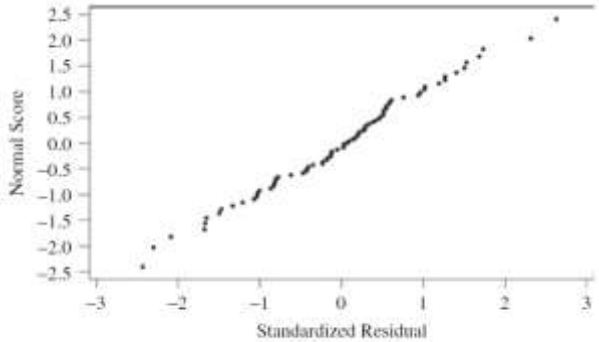
26

Verifying Model Assumptions

- Assumptions such as:
  - Linearity
  - Independence
  - Normality
  - Constant Variance
  - …
- Examples:
  - Checking normal distribution by normal plot of the residual
  - Checking linear relationship and constant variance in error term by plotting standardized residual vs. fitted (predicted value)

27

Normal Plot of the Residual

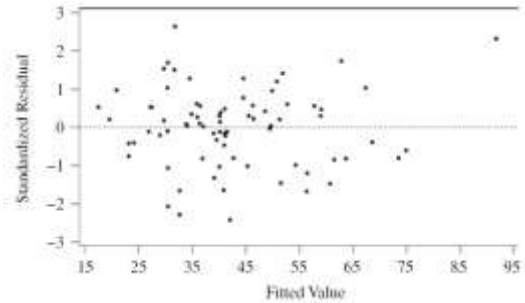


Normal plot of the residuals.

We do not detect systematic deviations from linearity in the normal plot of the standardized residuals
=> our normality assumption is intact.

28

Checking linear relationship and constant variance



- If obvious curvature exists in the scatter plot, the linearity assumption is violated.
- If the vertical spread of the points in the plot is systematically nonuniform, the constant variance assumption is violated.
- We detect no such patterns in Figure
- => the linearity and constant variance assumptions are intact

29

Homework 2: Cleaning *Churn* Dataset,

(Due Date is 95/01/20)

- *get the churn* data set from http://www.sgi.com/tech/mlc/db/
  - Explore whether there are missing values for any of the variables.
  - Compare the area code and state fields. Discuss any apparent abnormalities.
  - Use a graph to determine visually whether there are any outliers among the number of calls to customer service.
  - Transform the *day minutes* attribute using min–max normalization. Verify using a graph that all values lie between zero and 1.
  - Transform the *night minutes* attribute using *Z*-score standardization. Using a graph, describe the range of the standardized values.
  - Apply as much as preprocessing and descriptive techniques you have learned from Chapter 2 and 3.

30