Lecture 4 - Statistical Approaches to Estimation and Prediction

Adnan Ferdous Ashrafi

Stamford University Bangladesh

Table of Contents

1. STATISTICAL APPROACHES
   - Statistics in Data Mining
2. UNIVARIATE METHODS
   - MEASURES OF CENTER AND SPREAD
     - Mean of a variable
     - Median and Mode
   - Measures of variability
     - Range and Standard Deviation
3. STATISTICAL INFERENCE
   - Definition
   - Population
   - Parameter
   - Sample
   - Population parameters
   - Estimation
   - Proportion
   - Point estimation
4. How confident are we in our estimates?
   - Sampling Error
   - A paradox
   - Confidence Interval Estimate
     - Example 1
     - Example 2
   - Why do we need confidence interval estimation?
5. BIVARIATE METHODS
   - Regression
     - Example
     - Calculation
     - Analysis
     - Estimation
   - DANGERS OF EXTRAPOLATION
     - Example
     - Definition
   - CONFIDENCE INTERVALS
     - PREDICTION INTERVALS
   - MULTIPLE REGRESSION
     - Example
     - Correlation Coefficients
     - Multicollinearity
6. Further Reading

STATISTICAL APPROACHES TO ESTIMATION AND PREDICTION

If estimation and prediction are considered to be data mining tasks, statistical analysts have been performing data mining for over a century. In this lecture we will be examining :

- univariate methods
- statistical estimation
- prediction methods

These methods include point estimation and confidence interval estimation. Next we will consider simple linear regression, where the relationship between two numerical variables is investigated. Finally, we will examine multiple regression, where the relationship between a response variable and a set of predictor variables is modeled linearly.

UNIVARIATE METHODS: MEASURES OF CENTER AND SPREAD

Consider our roles as data miners. We have been presented with a data set with which we are presumably unfamiliar. We have completed the data understanding and data preparation phases and have gathered some descriptive information using exploratory data analysis. Next, we would like to perform univariate estimation and prediction, using numerical field summaries.

Suppose that we are interested in estimating where the center of a particular variable lies, as measured by one of the numerical measures of center, the most common of which are the mean, median, and mode.

Mean

Definition

The mean of a variable is simply the average of the valid values taken by the variable. To find the mean, simply add up all the field values and divide by the sample size.

The sample mean is denoted as $\bar{x}$ (“x-bar”) and is computed as $\bar{x} = \Sigma x/n$, where Σ (capital sigma, the Greek letter “S,” for summation) represents “sum all the values,” and n represents the sample size.

Example

The mean number of customer service calls for this sample of $n = 3333$ customers is given as:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{5209}{3333} = 1.563$$

Median and Mode of a variable

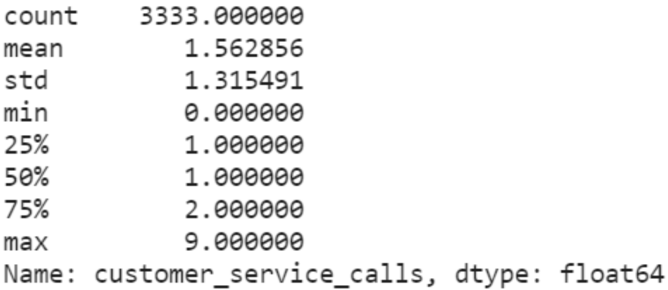
Median

The median is defined as the field value in the middle when the field values are sorted into ascending order. The median is resistant to the presence of outliers.

Mode

Other analysts may prefer to use the mode, which represents the field value occurring with the greatest frequency. The mode may be used with either numerical or categorical data, but is not always associated with the variable center.

Median, Mode, IQR, Range of a variable - Example



```
count    3333.000000
mean        1.562856
std         1.315491
min         0.000000
25%         1.000000
50%         1.000000
75%         2.000000
max         9.000000
Name: customer_service_calls, dtype: float64
```

Figure 1: Statistical summaries of customer service calls.

Median and Mode of a variable

Mean and Median calculation

In Figure 1, the median is 1.0, which means that half of the customers made at least one customer service call; the mode is also 1.0, which means that the most frequent number of customer service calls was 1. The median and mode agree. However, the mean is 1.563, which is 56.3% higher than the other measures. This is due to the mean's sensitivity to the right-skewness of the data.

Measures of variability- Range and Standard Deviation

Range

The range of a variable is simply the minimum and maximum range of values.

Standard Deviation

The standard deviation can be interpreted as the “typical” distance between a field value and the mean, and most field values lie within two standard deviations of the mean. The sample standard deviation is perhaps the most widespread measure of variability and is defined by

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

Because of the squaring involved, the standard deviation is sensitive to the presence of outliers, leading analysts to prefer other measures of spread, such as the mean absolute deviation, in situations involving extreme values.

Measures of variability- Range

Range and Standard Deviation calculation

From Figure 1 we can state that the number of customer service calls made by most customers lies within $2(1.315) = 2.63$ of the mean of 1.563 calls. In other words, most of the number of customer service calls lie within the interval $(-1.067, 4.193)$, that is, $(0, 4)$. This can be verified by examining the histogram of customer service calls.

A more complete discussion of measures of location and variability can be found in any introductory statistics textbook, such as Johnson and Kuby [1].

STATISTICAL INFERENCE

In statistical analysis, estimation and prediction are elements of the field of statistical inference. Statistical inference consists of methods for estimating and testing hypotheses about population characteristics based on the information contained in the sample.

STATISTICAL INFERENCE- Population

Definition

A *population* is the collection of all elements (persons, items, or data) of interest in a particular study.

For example, presumably, the cell phone company does not want to restrict its actionable results to the sample of 3333 customers from which it gathered the data. Rather, it would prefer to deploy its churn model to all of its present and future cell phone customers, which would therefore represent the population.

STATISTICAL INFERENCE- Parameter

Definition

A **parameter** is a characteristic of a population, such as the mean number of customer service calls of all cell phone customers.

STATISTICAL INFERENCE- Sample

Definition

A sample is simply a subset of the population, preferably a representative subset. If the sample is not representative of the population, that is, if the sample characteristics deviate systematically from the population characteristics, statistical inference should not be applied.

A statistic is a characteristic of a sample, such as the mean number of customer service calls of the 3333 customers in the sample (1.563).

STATISTICAL INFERENCE- Population parameters

Note that the values of population parameters are unknown for most interesting problems. Specifically, the value of the population mean is usually unknown. For example, we do not know the true mean number of customer service calls to be made by all of the company's cell phone customers. To represent their unknown nature, population parameters are often denoted with Greek letters. For example, the population mean is symbolized using the Greek lowercase letter (μ), which is the Greek letter for "m" ("mean").

Estimation

The value of the population mean number of customer service calls μ is unknown for a variety of reasons, including the fact that the data may not yet have been collected or warehoused. Instead, data analysts would use estimation.

For example, they would estimate the unknown value of the population mean μ by obtaining a sample and computing the sample mean $\bar{x}$, which would be used to estimate μ. Thus, we would estimate the mean number of customer service calls for all customers to be 1.563, since this is the value of our observed sample mean.

Estimation

An important caveat is that estimation is valid only as long as the sample is truly representative of the population. For example, in the churn data set, the company would presumably implement policies to improve customer service and decrease the churn rate. These policies would, hopefully, result in the true mean number of customer service calls falling to a level lower than 1.563.

Proportion

Analysts may also be interested in proportions, such as the proportion of customers who churn. The sample proportion p is the statistic used to measure the unknown value of the population proportion π. For example, in Lecture 3 we found that the proportion of churners in the data set was $p = 0.145$, which could be used to estimate the true proportion of churners for the population of all customers, keeping in mind the caveats above.

Point estimation

Point estimation refers to the use of a single known value of a statistic to estimate the associated population parameter. The observed value of the statistic is called the point estimate. We may summarize estimation of the population mean, standard deviation, and proportion using Table 1.

Table 1: Use Observed Sample Statistics to Estimate Unknown Population Parameters

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

Point estimation

Estimation need not be restricted to the parameters in Table 1. Any statistic observed from sample data may be used to estimate the analogous parameter in the population. For example, we may use the sample maximum to estimate the population maximum, or we could use the sample 27th percentile to estimate the population 27th percentile. Any sample characteristic is a statistic, which, under the appropriate circumstances, can be used to estimate its appropriate parameter.

Point estimation

More specifically, for example, we could use the sample churn proportion of customers who did select the VoiceMail Plan, but did not select the International Plan, and who made three customer service calls to estimate the population churn proportion of all such customers. Or, we could use the sample 99th percentile of day minutes used for customers without the VoiceMail Plan to estimate the population 99th percentile of day minutes used for all customers without the VoiceMail Plan.

How confident are we in our estimates?

The question is: *How confident can we be in the accuracy of the estimate?*

Do you think that the population mean number of customer service calls made by all of the company's customers is exactly the same as the sample mean $\bar{x} = 1.563$? Probably not. In general, since the sample is a subset of the population, inevitably the population contains more information than the sample about any given characteristic. Hence, unfortunately, our point estimates will nearly always "miss" the target parameter by a certain amount, and thus be in error by this amount, which is probably, though not necessarily, small.

Sampling Error

This distance between the observed value of the point estimate and the unknown value of its target parameter is called sampling error, defined as $|statistic - parameter|$. For example, the sampling error for the mean is $|x - \bar{x}|$, the distance (always positive) between the observed sample mean and the unknown population mean.

Paradox in confidence

Point estimates have no measure of confidence in their accuracy; there is no probability statement associated with the estimate.

All we know is that the estimate is probably close to the value of the target parameter (small sampling error) but that possibly it may be far away (large sampling error).

In fact, point estimation has been likened to a dart thrower, throwing darts with infinitesimally small tips (the point estimates) toward a vanishingly small bull's-eye (the target parameter).

Worse, the bull's-eye is hidden, and the thrower will never know for sure how close the darts are coming to the target.

Confidence Interval Estimate

A confidence interval estimate of a population parameter consists of an interval of numbers produced by a point estimate, together with an associated confidence level specifying the probability that the interval contains the parameter.Most confidence intervals take the general form:

$$\text{point estimate} \pm \text{margin of error}$$

where the margin of error is a measure of the precision of the interval estimate. Smaller margins of error indicate greater precision.

For example, the t-interval for the population mean is given by:

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

where the sample mean $\bar{x}$ is the point estimate and the quantity $t_{\alpha/2}(s/\sqrt{n})$ represents the margin of error. The t-interval for the mean may be used when either the population is normal or the sample size is large.

Confidence Interval Estimate- Example

Usually, finding a large sample size is not a problem for many data mining scenarios. For example, using the statistics in Figure 1, we can find the 95% t-interval for the mean number of customer service calls for all customers as follows:

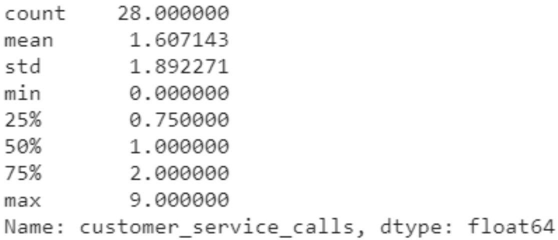
$$\begin{aligned}\bar{x} \pm t_{\alpha/2}(s/\sqrt{n}) &= 1.563 \pm 1.96(1.315\sqrt{3333}) \\ &= 1.563 \pm 0.045 \\ &= (1.518, 1.608)\end{aligned}$$

We are 95% confident that the population mean number of customer service calls for all customers falls between 1.518 and 1.608 calls. Here, the margin of error is 0.045 customer service calls, which is fairly precise for most applications.

* You can calculate $t_{\alpha/2}$, using a t table or a calculator. In our example, we used a t-table from https://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf

Confidence Interval Estimate- Example

However, data miners are often called upon to estimate the behavior of specific subsets of customers instead of the entire customer base, as in the example above. For example, suppose that we are interested in estimating the mean number of customer service calls for customers who have both the International Plan and the VoiceMail Plan and who have more than 220 day minutes. This considerably restricts the sample size, as shown in Figure 2.



```
count    28.000000
mean      1.607143
std       1.892271
min       0.000000
25%       0.750000
50%       1.000000
75%       2.000000
max       9.000000
Name: customer_service_calls, dtype: float64
```

Figure 2: Summary statistics of customers with both the International Plan and VoiceMail Plan and with more than 200 day minutes.

Confidence Interval Estimate- Example

There are only 28 customers in the sample who have both plans and who logged more than 220 minutes of day use. The point estimate for the population mean number of customer service calls for all such customers is the sample mean 1.607. We may find the 95% t-confidence interval estimate as follows:

$$\begin{aligned} \bar{x} \pm t_{\alpha/2}(s/\sqrt{n}) &= 1.607 \pm 2.048(1.892\sqrt{28}) \\ &= 1.607 \pm 0.732 \\ &= (0.875, 2.339) \end{aligned}$$

We are 95% confident that the population mean number of customer service calls for all customers who have both plans and who have more than 220 minutes of day use falls between 0.875 and 2.339 calls. The margin of error for this specific subset of customers is 0.732, which indicates that our estimate of the mean number of customer service calls for this subset of customers is much less precise than for the customer base as a whole.

Why do we need confidence interval estimation?

Confidence interval estimation can be applied to any desired target parameter. The most widespread interval estimates are for the population mean, the population standard deviation, and the population proportion of successes.

BIVARIATE METHODS

So far we have discussed estimation measures for one variable at a time. Analysts, however, are often interested in bivariate methods of estimation, for example, using the value of one variable to estimate the value of a different variable.

To help us learn about regression methods for estimation and prediction, let us get acquainted with a new data set, cereals. The cereals dataset can be found at Data and Story Library website [2].

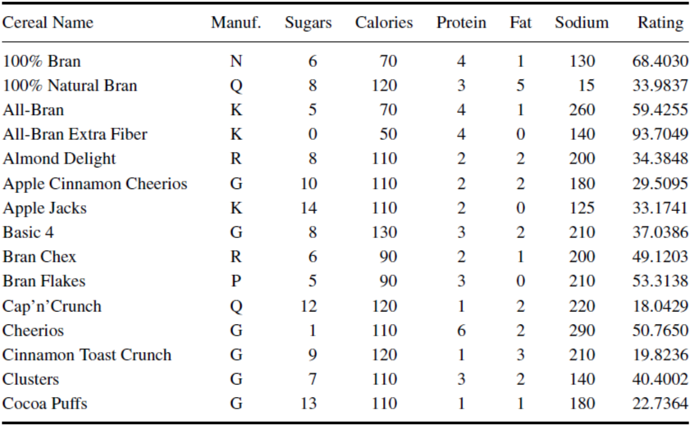
Regression- Example

The cereals dataset [2] contains nutrition information for 77 breakfast cereals and includes the following variables:

- Cereal name
- Cereal manufacturer
- Type (hot or cold)
- Calories per serving
- Grams of protein
- Grams of fat
- Milligrams of sodium
- Grams of fiber
- Grams of carbohydrates
- Grams of sugars
- Milligrams of potassium
- Percentage of recommended daily allowance of vitamins (0% 25%, or 100%)
- Weight of one serving
- Number of cups per serving
- Shelf location (1 = bottom, 2 = middle, 3 = top)
- Nutritional rating, calculated by *Consumer Reports*

Regression- Example

Figure 3 provides a peek at the eight of these fields for the first 16 cereals.We are interested in estimating the nutritional rating of a cereal given its sugar content.



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

Figure 3: Excerpt from Cereals Data Set: Eight Fields, First 16 Cereals

Regression- Example

Figure 4 shows a scatter plot of the nutritional rating versus the sugar content for the 77 cereals, along with the least-squares regression line.

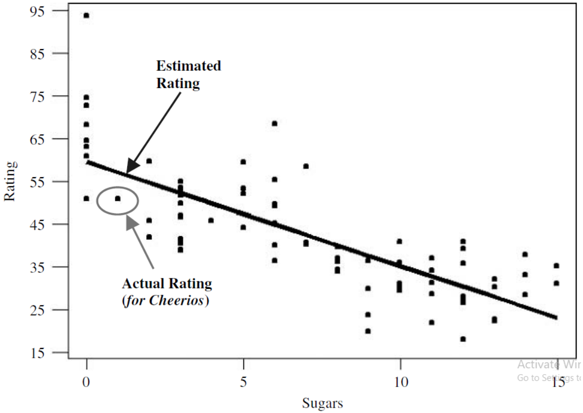


Figure 4: Scatter plot of nutritional rating versus sugar content for 77 cereals.

Regression- Calculation

To calculate b_0 and b_1 we need to calculate the $\bar{x}$, $\bar{y}$,$\sum x^2$,$\sum xy$,σ and co-variance of x and y. In the given scenario, $\bar{x} = 6.92$, $\bar{y} = 42.67$, $\sigma = 4.44$, $\sum x^2 = 5191$, $\sum xy = 19135.91$.

Thus we can calculate the variance of x as:

$\sigma(x) = \sum x^2 - n \times \bar{x} = (\sigma)^2 = 19.76$

Now we can find the co-variance of x and y by using:

Cov(x,y) $= \frac{\sum xy - n \times \bar{x} \times \bar{y}}{n-1} = -47.43$

slope of the regression line is, $b_1 = \frac{Cov(x,y)}{\sigma x^2} = \frac{-47.43}{4.44} = -2.42$

intercept is, $b_0 = \bar{y} - b_1 \times \bar{x} = 59.4$

regression equation, $\hat{y} = 59.4 - 2.42(x)$

Regression- Analysis

The regression line is written in the form $\hat{y} = b_0 + b_1 x$, called the *regression equation* or the *estimated regression equation* (ERE), where:

- $\hat{y}$ is the estimated value of the response variable
- b_0 is the y-intercept of the regression line
- b_1 is the slope of the regression line
- b_0 and b_1, together, are called the regression coefficients

In this case the ERE is given as $\hat{y} = 59.4 – 2.42$(sugars), so that $b_0 = 59.4$ and $b_0 = -2.42$. This estimated regression equation can then be interpreted as: "The estimated cereal rating equals 59.4 minus 2.42 times the sugar content in grams." The regression line and the ERE are used as a linear approximation of the relationship between the x (predictor) and y (response) variables, that is, between sugar content and nutritional rating.

Regression- Estimation

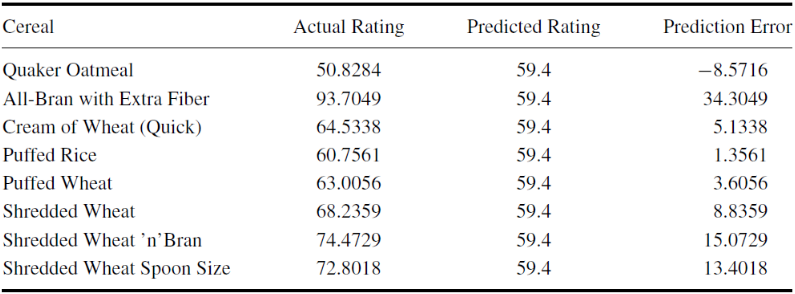
For example, suppose that we are interested in estimating the nutritional rating for a new cereal (not in the original data) that contains x = 1 gram of sugar. Using the ERE, we find the estimated nutritional rating for a cereal with 1 gram of sugar to be $\hat{y} = 59.4 - 2.42(1) = 56.98$. Note that this estimated value for the nutritional rating lies directly on the regression line, at the location $(x = 1, \hat{y} = 56.98)$, as shown in Figure 4. In fact, for any given value of x (sugar content), the estimated value for y (nutritional rating) lies precisely on the regression line.

Regression- Estimation Error

Now, there is one cereal in our data set that does have a sugar content of 1 gram, Cheerios. Its nutrition rating, however, is 50.765, not 56.98 as we estimated above for the new cereal with 1 gram of sugar. Cheerios' point in the scatter plot is located at $(x = 1, \hat{y} = 50.765)$, within the oval in Figure 4. Now, the upper arrow in Figure 4 is pointing to a location on the regression line directly above the Cheerios point. This is where the regression equation predicted the nutrition rating to be for a cereal with a sugar content of 1 gram. The prediction was too high by $56.98 - 50.765 = 6.215$ rating points, which represents the vertical distance from the Cheerios data point to the regression line. This vertical distance of 6.215 rating points, in general $(y - \hat{y})$, is known variously as the *prediction error, estimation error,* or *residual*.

BIVARIATE METHODS- Correlation Coefficient

The correlation coefficient r for rating and sugars is -0.76, indicating that the nutritional rating and the sugar content are negatively correlated. It is not a coincidence that both r and b_1 are both negative. In fact, the correlation coefficient r and the regression slope b_1 always have the same sign.



Cereal	Actual Rating	Predicted Rating	Prediction Error
Quaker Oatmeal	50.8284	59.4	−8.5716
All-Bran with Extra Fiber	93.7049	59.4	34.3049
Cream of Wheat (Quick)	64.5338	59.4	5.1338
Puffed Rice	60.7561	59.4	1.3561
Puffed Wheat	63.0056	59.4	3.6056
Shredded Wheat	68.2359	59.4	8.8359
Shredded Wheat ’n’Bran	74.4729	59.4	15.0729
Shredded Wheat Spoon Size	72.8018	59.4	13.4018

Figure 5: Actual Ratings, Predicted Ratings, and Prediction Errors for Cereals with Zero Grams of Sugar

DANGERS OF EXTRAPOLATION- Example

Suppose that a new cereal called EValley cereal arrives on the market with a very high sugar content of 30 grams per serving. Let us use our estimated regression equation to estimate the nutritional rating for the cereal:

$$\hat{y} = 59.4 - 2.42(\text{sugars}) = 59.4 - 2.42(30) = -13.2.$$

In other words, EValley's cereal has so much sugar that its nutritional rating is actually a negative number, unlike any of the other cereals in the data set (minimum = 18) and analogous to a student receiving a negative grade on an exam. What is going on here?

DANGERS OF EXTRAPOLATION- Definition

Extrapolation is making predictions for x values lying outside this range, can be dangerous, since we do not know the nature of the relationship between the response and predictor variables outside this range.

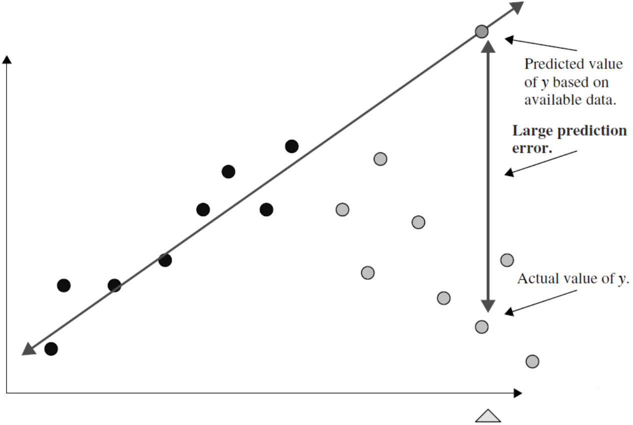


Figure 6: Dangers of extrapolation.

CONFIDENCE INTERVALS FOR THE MEAN VALUE OF y GIVEN x

The confidence interval for the mean value of y for a given value of x is as follows:

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

where,

$$\begin{aligned}
x_p &= \text{the particular value of x for which the prediction is being made} \\
\hat{y}_p &= \text{the point estimate of y for a particular value of x} \\
t_{\alpha/2} &= \text{a multiplier associated with the sample size and confidence level} \\
s &= \sqrt{\text{MSE}} = \sqrt{\text{SSE}/n-1} = \text{the standard error of the estimate} \\
\text{SSE} &= \text{the sum of squared residuals}
\end{aligned}$$

We will see an example of this type of confidence interval below, but first let's know about prediction interval.

PREDICTION INTERVALS

Have you ever considered that it is “easier” to predict the mean value of a variable than it is to predict a randomly chosen value of that variable?

Definition

Prediction intervals are used to estimate the value of a randomly chosen value of y, given x. Clearly, this is a more difficult task than estimating the mean, resulting in intervals of greater width (lower precision) than confidence intervals for the mean with the same confidence level.

The prediction interval for a randomly chosen value of y for a given value of x is as follows:

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

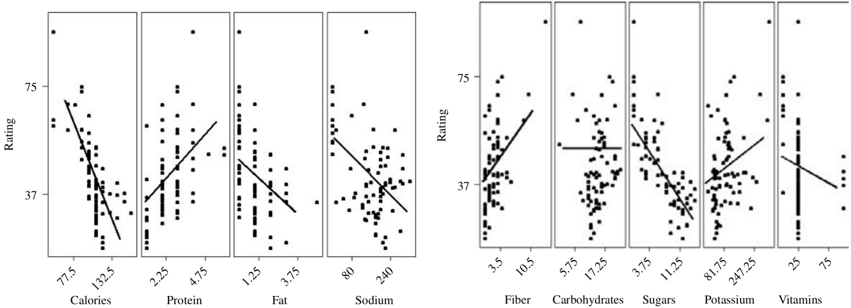
The presence of the “1+” inside the square root ensures that the prediction interval is always wider than the analogous confidence interval.

MULTIPLE REGRESSION

Most data mining applications enjoy a wealth (indeed, a superfluity) of data, with some data sets including hundreds of variables, many of which may have a linear relationship with the target (response) variable.

Multiple regression modeling provides an elegant method of describing such relationships. Multiple regression models provide improved precision for estimation and prediction, analogous to the improved precision of regression estimates over univariate estimates.

MULTIPLE REGRESSION



(a) Draftman's plot of rating versus calories, protein, fat, and sodium.

(b) Draftman's plot of rating versus fiber, carbohydrates, sugars, potassium, and vitamins.

Figure 7: Plot of a response variable against several predictor variables

MULTIPLE REGRESSION

From Figures 7a and 7b, we would expect that

- protein, fiber, and potassium would be positively correlated with a higher nutritional rating
- fat, sodium, sugars, and surprisingly, vitamins are negatively correlated with a higher nutritional rating
- Carbohydrates seem to be uncorrelated with nutritional rating

Correlation Coefficients

We can verify these graphical findings with the correlation coefficients for all the variables, shown in the Figure 8. The first column (in bold) shows the correlation coefficients of the predictor variables with rating. As expected, protein, fiber, and potassium are positively correlated with rating, whereas calories, fat, sodium, and vitamins are negatively correlated.

	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

Figure 8: Correlation Coefficients for All Variables

Multicollinearity

Data analysts need to guard against multicollinearity, a condition where some of the predictor variables are correlated with each other. Multicollinearity leads to instability in the solution space, leading to possible incoherent results. Even if such instability is avoided, inclusion of variables that are highly correlated tends to overemphasize a particular component of the model, since the component is essentially being double counted. Here, potassium is very highly correlated with fiber (r = 0.905). Although more sophisticated methods exist for handling correlated variables, such as principal components analysis, in this introductory example we simply omit potassium as a predictor.

Further Reading

- Chapter 4,5 of Data Mining - Practical Machine Learning Tools and Techniques, Second Edition - Ian H. Witten, Eibe Frank
- Chapter 4 of DISCOVERING KNOWLEDGE IN DATA - An Introduction to Data Mining - DANIEL T. LAROSE
- Chapter 3 of Introduction to Data Mining (Second Edition) - Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar

References

[1] R. Johnson and P. Kuby, *Elementary Statistics*. Cengage Learning, 2011. [Online]. Available: https://books.google.com.bd/books?id=x_QliCVSzToC

[2] “Data and story library,” 2021. [Online]. Available: https://dasl.datadescription.com/

Thank you.
Any Questions?