slides originally by
Dr. Richard Burns,
modified by
Dr. Stephanie Schwartz

MULTIVARIATE LINEAR REGRESSION

CSCI 452: Data Mining

Multivariate Linear Regression

- In practice, often have more than one *predictor*
- Option: run three separate simple linear regressions for the Advertising dataset
  - However, it's unclear how to make single prediction of sales given all three predictor values
  - Media may be correlated with each other, but each regression equation ignores the other two media

Multivariate Linear Regression Model

- Extend the simple linear regression model for each predictor
  - Response variable Y is numeric (continuous)
- For p predictor variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

- Since error ε has mean zero, variance σ^2, with normal distribution, we usually omit it.

- A one-unit change in any predictor variable x_i will change the expected mean response by β_i units.

Advertising Dataset

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \varepsilon$$

Estimating the Parameters $\beta_0 \beta_1 \beta_2 \ldots$

- Parameters (regression coefficients) are typically estimated through the method of <u>least squares</u>
  - Just like with simple linear regression
  - Automatic in R

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

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_p x_{ip})^2$$

We want to minimize the RSS

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

Advertising Dataset

- R Syntax:

```
… ← lm(Y ~ X1 + X2 + … + Xp, …)
```

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \varepsilon$$

*Sales = 2.938889 + 0.045765 * TV + 0.188530 * radio + -0.001037 * newspaper*

Simple and Multiple Linear Regression Coefficients can be Quite Different

```
lm(formula = Sales ~ TV, data = Advertising)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
```

```
lm(formula = Sales ~ Radio, data = Advertising)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.31164    0.56290  16.542   <2e-16 ***
Radio        0.20250    0.02041   9.921   <2e-16 ***
```

```
lm(formula = Sales ~ Newspaper, data = Advertising)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
Newspaper    0.05469    0.01658    3.30  0.00115 **
```

```
lm(formula = Sales ~ TV + Radio + Newspaper, data
= Advertising)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16
***
TV           0.045765   0.001395  32.809   <2e-16
***
Radio        0.188530   0.008611  21.893   <2e-16
***
Newspaper   -0.001037   0.005871  -0.177     0.86
```

Simple and Multiple Linear Regression Coefficients can be Quite Different

Slope term (*newspaper* coefficient) represents the average effect of a $1,000 increase in newspaper advertising, ignoring other predictors (*TV* and *radio*).

```
lm(formula = Sales ~ Newspaper, data = Advertising)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
Newspaper    0.05469    0.01658    3.30  0.00115 **
```

```
lm(formula = Sales ~ TV + Radio + Newspaper, data =

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 *
TV           0.045765   0.001395  32.809   <2e-16 *
Radio        0.188530   0.008611  21.893   <2e-16 *
Newspaper   -0.001037   0.005871  -0.177     0.86
```

Coefficient for *newspaper* represents the average effect of increasing newspaper spending by $1,000 while holding *TV* and *radio* fixed.

Correlation Matrix

- Correlation between *radio* and *newspaper* is 0.35
  - Barely any correlation (or "not correlated") for TV/radio and TV/newspaper
- Reveals tendency to spend more on *Newspaper* advertising in markets where more is spent on *Radio* advertising.
- *Sales* higher in markets where more is spent on *Radio*, but more also tends to be spend on *Newspaper*.
- In Simple LM: *Newspaper* "gets credit" for effect of *Radio* on *Sales*.

Correlation Matrix

	X	TV	Radio	Newspaper	Sales
X	1.00000000	0.01771469	-0.11068044	-0.15494414	-0.05161625
TV	0.01771469	1.00000000	0.05480866	0.05664787	0.78222442
Radio	-0.11068044	0.05480866	1.00000000	0.35410375	0.57622257
Newspaper	-0.15494414	0.05664787	0.35410375	1.00000000	0.22829903
Sales	-0.05161625	0.78222442	0.57622257	0.22829903	1.00000000

Advertising Dataset

- *Goal:* What marketing plan for next year will result in high product sales?
- *Questions:*
  1. Is there a relationship between advertising budget and sales?

Yes, hypothesis testing shows that we can reject the null hypothesis that $B_{TV} = B_{RADIO} = B_{NEWSPAPER} = 0$

Advertising Dataset

- *Goal*: What marketing plan for next year will result in high product sales?
- *Questions*:

2. How strong is the relationship between advertising budget and sales?

- RSE is 1,681 units while the mean value of the response is 14,022, indicating a percentage error of 12%.
- Via R^2, the predictors explain almost 90% of the variance in *sales*.

Advertising Dataset

- *Goal*: What marketing plan for next year will result in high product sales?
- *Questions:*
  3. Which media contribute to sales?
     - Need to separate the effects of each medium

The *p*-values associated with *TV* and *radio* are low, while *newspaper* is not, suggesting that only *TV* and *radio* are related to *sales*.

Advertising Dataset

- *Goal:* What marketing plan for next year will result in high product sales?
- *Questions:*
  4. How accurately can we estimate the effect of each medium on sales?
     - For every dollar spent on advertising in a particular medium, by what amount will sales increase? How accurately can we predict this increase?

- The 95% confidence intervals for each medium are as follows: (0.043, 0.049) for *TV*, (0.172, 0.206) for *radio*, and (-0.013, 0.011) for *newspaper*. The confidence intervals for *TV* and *radio* are far from zero, providing evidence that these media are related to *sales*.
- Accuracy depends if we wish to predict an individual response (will use prediction interval), or the average response (use confidence interval). Prediction intervals are always wider because they account for the uncertainty associated with ε, the irreducible error.

Advertising Dataset

- *Goal:* What marketing plan for next year will result in high product sales?
- *Questions:*
  5. Is the relationship linear?
     - If the relationship between advertising budget and sales is a straight-line, then linear regression seems appropriate.
     - If not, all is not lost yet. (Variable Transformation)

- Residual plots can be used in order to identify non-linearity. If the relationships are linear, then the residual plots should display no pattern.

Advertising Dataset

- *Goal:* What marketing plan for next year will result in high product sales?
- *Questions:*
  6. Is there any <u>interaction effect</u>? *(called "synergy" in business)*
     - *Example:* spending 50k on TV ads + 50k on radio ads results in more sales than spending 100k on only TV

- The standard linear regression model assumes an <u>additive relationship</u> between the predictors and the response. The effect of each predictor on the response is unrelated to the values of the other predictors.
- The additive assumption may be unrealistic for certain datasets.
- Can extend linear model to include *interaction term*.

Create an Interaction Term

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- The third term captures possible interaction between the two parameters
- If we do this with Radio and TV, we see a better fit:
  - RSE is lower (.9435 vs 1.686)
  - R^2 is .9678 (vs .8973)

R: Cereal Dataset

- http://lib.stat.cmu.edu/DASL/Datafiles/Cereals.html
- From CMU Data and Story Library
- 77 cereals
- 15 Attributes: calories, sugar content, protein, etc.
- Target: Consumer Reports "Health Rating" (continuous)

R: Cereal Dataset

- Regression Model:

  Rating = B0 + B1 x Sugar

- "For each increase of 1 gram in sugar content, the estimated nutritional rating <u>decreases</u> by 2.4008 rating points."

R: Cereal Dataset

- Correlation coefficient (r): -0.760
  - Correlation coefficient *r* and regression slope *b1* will always have the same sign
- RSE: "57.7% of the variability in nutritional rating is accounted for by the linear relationship between rating and sugars alone, without looking at other variables (such as sodium)."

Dangers of Extrapolation

- Extrapolation should be avoided if possible.
- Analysts should confine the estimates and predictions made using the regression equation to values of the predictor variable contained within the range of the values of x in the dataset.

Dangers of Extrapolation

- *Cereal Example*: range of any value of x (sugar) between 0 and 15 grams is appropriate
  - New cereal has 30 grams of sugar

$$\hat{y} = 59.284 - 2.4008(sugars) = 59.284 - 2.4008(30) = -12.74$$

  - Predicted nutritional rating is a negative number, unlike any of the other cereals in the dataset.

Dangers of Extrapolation

- If predictions outsides the given x range must be performed, the end user should be informed that no x-data is available to support such a prediction.
- Also possible that the relationship between x and y is linear within the range of x, but may no longer be linear beyond that range.

R: Cereal Dataset

- Example of multiple linear regression
  - *Predictors*: sodium, sugars
  - *Target*: nutritional score

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\hat{y} = 67.322 - 2.2958(sugars) - 0.05489(sodium)$$

R: Cereal Dataset

- Example of multiple linear regression
  - *Predictors*: sodium, sugars
  - *Target*: nutritional score

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$
$$\hat{y} = 67.322 - 2.2958(sugars) - 0.05489(sodium)$$

"For each additional milligram of sodium, the estimated decrease in nutritional rating is 0.05489 points, <u>when sugars are held constant</u>.

R: Cereal Dataset

# of predictors	1	2
Regression Equation	y = 59.9 – 2.46 (sugars)	y = 69.1 – 2.39 (sugars) – 0.06 (sodium)
Standard Error of the Estimate (RSE)	9.2	8.0
R^2	57.7% (0.577)	68.3%

- The addition of sodium information to the model has reduced our typical prediction errors to 8.0 points.
- The proportion of the variability in nutritional rating that is explained by our regression model is now over 68%.

References

- *Data Mining and Business Analytics in R*, 1st edition, Ledolter
- *An Introduction to Statistical Learning*, 1st edition, James et al.
- *Discovering Knowledge in Data*, 2nd edition, Larose et al.