STA130 T0209

Week 9: Simple Linear Regression

(Materials used in this presentation are provided by the U of T Statistical Sciences Department.

This presentation was prepared by Vivian Ngo.)

Agenda

- Reminder: final project
  - Information is posted
  - Tell me your groups
- Material, vocabulary, homework discussion
- Group work and presentations
- Remaining time: start preparing for the final poster project

Material and Vocabulary Review

- Linear Relationship
- Approximately linear
- Non-linear
- Slope
- Intercept
- (Simple) Linear Regression
- Regression model
- Parameter
- Regression coefficients
- Fitted regression line
- Explanatory/independent variable
- Dependent variable
- Measure of model fit
- Coefficient of determination
- Root mean square error
- Error
- Residual
- Prediction error
- Least squares
- Least squares estimator

Correlation coefficient (r)

- Between -1 and +1
- Abs value =1 => perfect correlation
- 0 = no correlation
- Positive, negative
- Measures **linear** relationship
- Examples?

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

Simple linear regression

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$\boldsymbol{Y_i}$ = response variable (dependent variable, target variable, ...) for i^{th} individual selected

x_i = independent variable (predictor, covariate, feature, input, ...) for i^{th} individual selected

$\boldsymbol{\beta_0}$ = intercept parameter

$\boldsymbol{\beta_1}$ = slope parameter

$\boldsymbol{\varepsilon_i}$ = random error term for i^{th} individual selected

E.g. Question 1. Interested in brain weight (grams) as our dependent variable and head size (cm^3) as our independent variable.

- $\hat{Y} = 325.57342 + 0.26343\ X_i$
  - What does this mean?

Simple linear regression: Estimating the coefficients

Setting the equations to 0 and solving for β_0 and β_1 give estimates:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r\frac{s_y}{s_x}$$

where $\bar{x}$ and $\bar{y}$ are sample means and s_x and s_y are sample standard deviations for the x and y observations, respectively.

$\hat{\beta}_0$ and $\hat{\beta}_1$ are the **least squares estimators** of β_0 and β_1

Aside: $$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad and \quad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad and \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Some important notes

- What is $\hat{Y}$ and how is it different from Y?
- What is the difference between error and residual?
- How do we get the coefficient estimates? (i.e. what are we minimizing?)
- what are the model assumptions for SLR? (next slide)

Assumptions of linear regression

1. There must be a **linear relationship** between the outcome variable and the independent variables. Scatterplots can show whether there is a linear or curvilinear relationship.
2. **Multivariate Normality–**Multiple regression assumes that the residuals are normally distributed. (More on this next week)
3. **No (or little) Multicollinearity**—Multiple regression assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values. (more on this next week)
4. **Homoscedasticity**–This assumption states that the variance of error terms are similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.

Measures of correlation

- Measures of correlation:
- **R^2**: a **relative** measure of fit, ranging from 0 to 1
  - E.g. $R^2 = 0.80$ means that 80% of the variation in your outcome can be explained by your model.
- **RMSE:** an **absolute** measure of fit
  - As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance
  - Useful property: in the same units as the response variable
  - Lower values of RMSE indicate better fit

Measures of correlation

spread of y's around their mean

spread of predicted values around mean of y

spread of y's around their predicted values

$$\sum (y-\bar{y})^2 = \sum (\hat{y}-\bar{y})^2 + \sum (y-\hat{y})^2$$

TOTAL variation

"EXPLAINED"

NOT "EXPLAINED"

$$R^2 = \frac{"Explained" Variation}{Total\ Variation} = \frac{\sum (\hat{y}-\bar{y})^2}{\sum (y-\bar{y})^2} = 1 - \frac{\sum (y-\hat{y})^2}{\sum (y-\bar{y})^2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i{}^2}$$

Group presentation

- Prepare a 5 minute oral presentation based on the following topics (next slide)
- When not presenting:
  - One person from each group evaluate other students, upload rubric to Quercus
  - Write down any questions you have
  - Presentation rubric on Quercus and Github

Oral presentations

- GROUP 1: Questions 1a and 1c
  - Describe your plot produced in question 1a. Make sure to note the x- and y-axis and to describe the association you observe, if any. E.g. the association linear, positive, negative, strong, weak, etc.?
  - What is the correlation between head size and brain weight? Make sure to explain how you calculated this value and what it means; i.e., provide an interpretation of the value.
  - Does this make sense based on your prior expectations? Are there any other variables you think may be important factors influencing brain weight?
  - Do there appear to be many outliers? Why might this matter?

Oral presentations

- GROUP 2: Questions 1d-f
  - Provide a simple linear regression equation for the association between head size and brain weight. Explain what each part of the model means in lay terms.
  - Based on your answer to part e, report the estimated values of your model and provide an interpretation of these values.
  - How well does your model fit the data? Explain what the coefficient of determination means and provide an interpretation.

Oral presentations

- GROUP 3: Question 2c
  - Present your regression model of msrp on year based on the training set.
  - What is the model equation and estimated values? What is the coefficient of determination? Explain what these values mean and an interpretation in lay terms.
  - How well does your model perform?

Oral presentations

- GROUP 4: Question 2d
  - What is your predicted 2013 msrp for a 2010 model hybrid vehicle? Make sure to present your regression equation, including all coefficients.
  - Suppose the actual 2013 msrp for this 2010 model hybrid vehicle was $27,000. What is the residual? Provide an interpretation in lay terms. Is this a large difference? Based on previous work done in this question, why do you think this may be the case? Hint: Think about how well the model fits the data, if there may be other important factors, etc.

	4 (Excellent)	3 (Good)	2 (Adequate)	1 (Poor)
Context	The context and connection to the problem are clear.	Some context was provided and all variables/concepts were mentioned. Some aspects were not clear.	Very little context was provided and only some variables/ concepts were mentioned.	No context and mentioning of any variables/ concepts covering in this week's materials.
Structure	Well organized, follows a logical structure.	The organization follows some logical structure.	Some structure but difficult to follow.	There is no structure, very difficult to follow.
Conclusion	There is a clear central idea and the conclusion is correct.	A central idea or conclusion is present. The conclusion might be incorrect.	The central idea or conclusion is weak and not supported.	The central idea or conclusion is missing. Incorrect conclusion.
Transitions	The progression is logical. Effective use of transitions.	The progression is controlled. The use of transitions is mostly meaningful.	Minor disruptions in flow and weak transitions.	Weak progression and lack of transitions.
Vocabulary	Good use of statistical terms and appropriate choice of words.	Use of statistical terms and phrases mostly correct, demonstrates understanding of concepts.	Some use of statistical terms/ phrases and some understanding of concepts demonstrated.	Inaccurate or incorrect use of statistical terms or phrases and a lack of understanding statistical concepts.
Presentation Skills	Regular eye contact with all parts of the audience. The audience was engaged. The presenter held the audience's attention. Appropriate speaking volume & body language. Good pace.	Somewhat regular eye contact or eye contact with some of the audience The audience was mostly engaged. The presenter mostly spoke at a suitable volume. Spoke too quietly at times. Some fidgeting. Going too fast/slow.	Focused on only one or two members of the audience. Sporadic eye contact. The audience was not engaged. Speaker could be heard by only some of the audience. Body language was distracting.	Minimal (or no) eye contact. The audience was never engaged. The presenter did not speak clearly. Presenter was very difficult to hear.
Preparedness/ Participation	Extremely prepared and rehearsed. The presenter was confident.	Mostly prepared but some dependence on or reading off of notes. The presenter seemed fairly confident.	The presenter was not well prepared. The presenter did not seem confident.	Evident lack of preparation/rehearsal. Complete dependence on notes.