STA130H1S TUT0109
W9: Simple Linear Regression

Mar 15, 2019

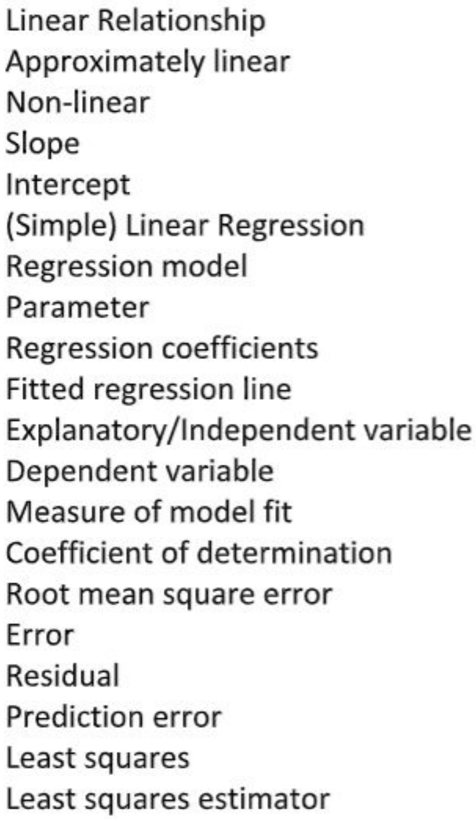
(Materials used in this presentation are provided by the UofT Statistical Sciences Department)

Email: sonia.chhay@mail.utoronto.ca
Website: soniachhay.github.io

Overview

- Group signup
- Material, vocabulary, homework discussion
- Group work and presentations
    - From today onwards, in-tutorial activities will be done with your final project group members

Vocabulary for this Week



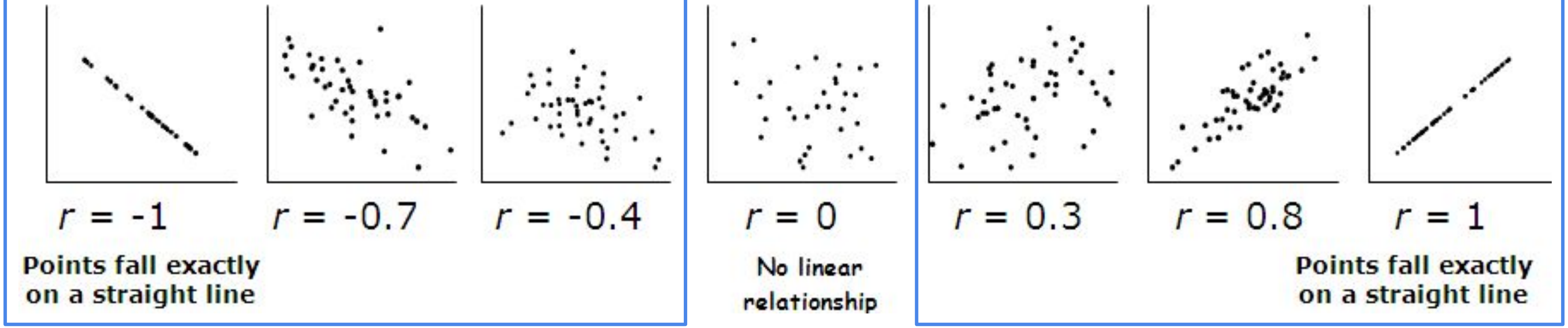
Linear Relationship
Approximately linear
Non-linear
Slope
Intercept
(Simple) Linear Regression
Regression model
Parameter
Regression coefficients
Fitted regression line
Explanatory/Independent variable
Dependent variable
Measure of model fit
Coefficient of determination
Root mean square error
Error
Residual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

Correlation Coefficient (r)

- The value of r is in between -1 and +1
- No correlation = 0
- Perfect correlation = -1 or +1

Negative Correlation
As x increases, y decreases

Positive Correlation
As x increases, y increases



Standard Linear Regression Equation

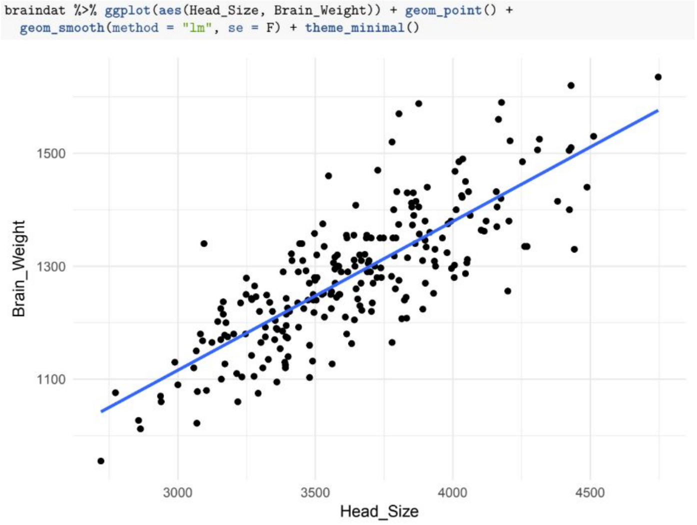
$$\boxed{Y = B_0 + B_1X_i + e_i}$$

- Y = Linear outcome
- B_0 = Intercept
- B_1 = Regression coefficient
- X_i = Explanatory variable
- e_i = Error
- i = # of individuals in the sample

Homework Q1

Brainhead.csv dataset contains sample data on:

- Brain weight (in grams)
- Head size (in cm^3)
- Gender (Male = 1, Female = 2)
- Age range (1=20-45, 2=46+)

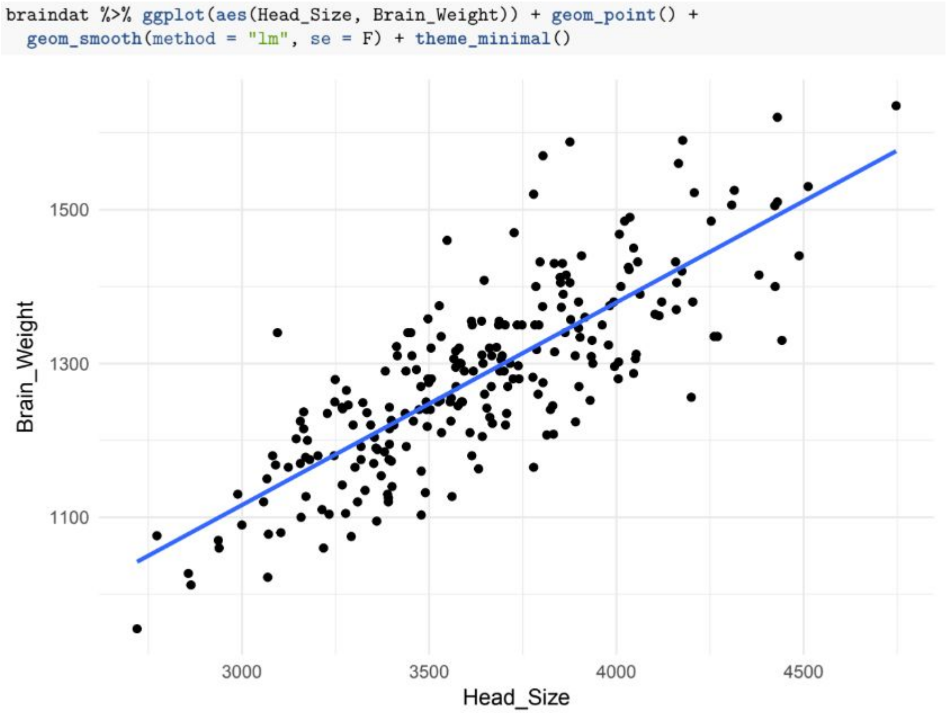


```
braindat %>% ggplot(aes(Head_Size, Brain_Weight)) + geom_point() +
  geom_smooth(method = "lm", se = F) + theme_minimal()
```

We’re interested in the relationship between Head_Size and Brain_Weight

- Head size (in cm^3) = Independent variable
- Brain weight (in grams) = Dependent variable

Homework Q1 (cont’d)



```
braindat %>% ggplot(aes(Head_Size, Brain_Weight)) + geom_point() +
  geom_smooth(method = "lm", se = F) + theme_minimal()
```

Describe this association:

- Linearity
  - Strength
- Interpretation/Interesting patterns
  - E.g. Variation

Homework Q1 (cont'd)

Fitted regression line: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

- '^' means it's an estimated value

E.g. **$\hat{Y} = 325.57342 + 0.26343 X_i$**

- When head size is 0 cm^3, the average brain weight is 325.57 grams
    - NOTE: The intercept in this case is not very informative since the interpretation does not make sense
- Estimated slope is 0.26
    - If an individual's head is one cm^3 larger, the average brain weight will be 0.26 grams heavier, on average.

Measures of Correlation

- **R^2** = Relative measure of fit
  - Ranges 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
  - Larger R^2 = Higher agreement between the observed and predicted values, using your model
- **RMSE (Root mean square error)** = Absolute measure of fit
  - The square root of variance
  - Can be interpreted as the standard deviation of the unexplained variance
  - Has the useful property of being in the same units as the response variable

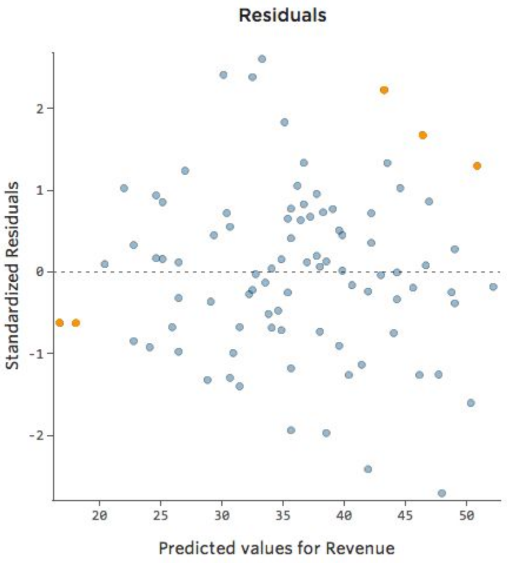
Measures of Correlation

How well does your model fit the data?

- Lower values of RMSE indicate better fit**
- If RMSE between training and test sets are large and differ a little, this may be a sign of *overfitting*

Assumptions of Linear Regression

- 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
- Homoscedasticity
  - 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



Assumptions of Linear Regression

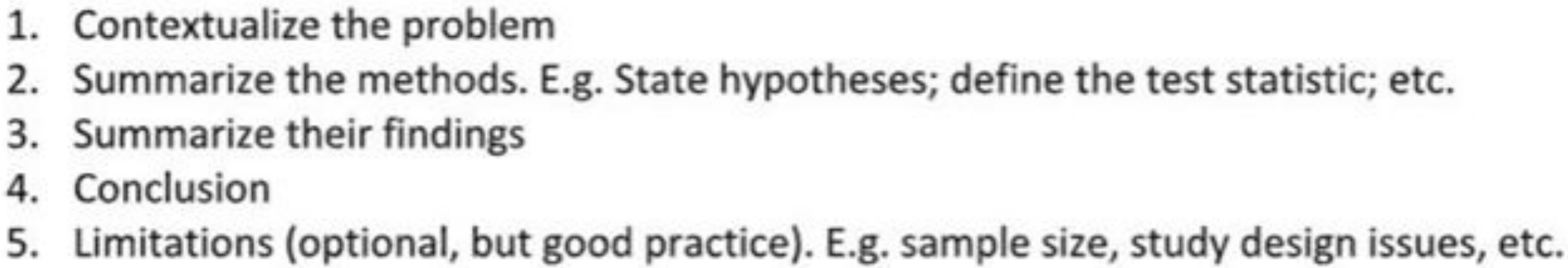
- Multivariate Normality
    - Multiple regression assumes that the residuals are normally distributed
    - (More on this next week)
- No (or little) Multicollinearity
    - Multiple regression assumes that the independent variables are not highly correlated with each other
    - This assumption is tested using Variance Inflation Factor (VIF) values (More on this next week)

Classification Trees vs Linear Regression

- How do classification trees differ from linear regression (LR)?
    - Classification trees are limited to binary variables (or dichotomizing categorical or continuous variables)
    - LR requires you to have a continuous outcome with some sort of linear relationship (refer to assumptions of LR)

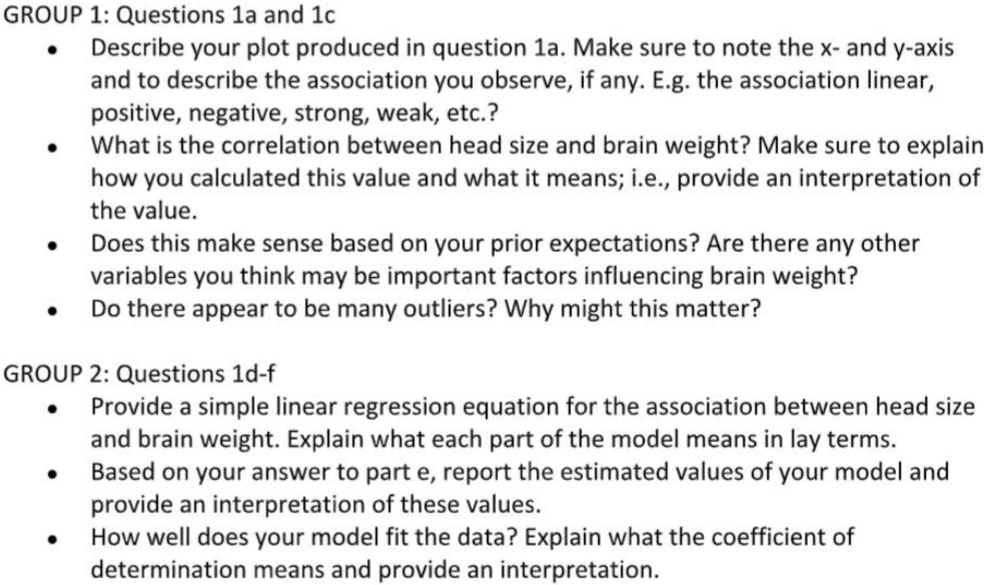
- In what circumstances might you use one vs the other?

Like a written summary, the presentation should include:



1. Contextualize the problem
2. Summarize the methods. E.g. State hypotheses; define the test statistic; etc.
3. Summarize their findings
4. Conclusion
5. Limitations (optional, but good practice). E.g. sample size, study design issues, etc.

Oral Presentations

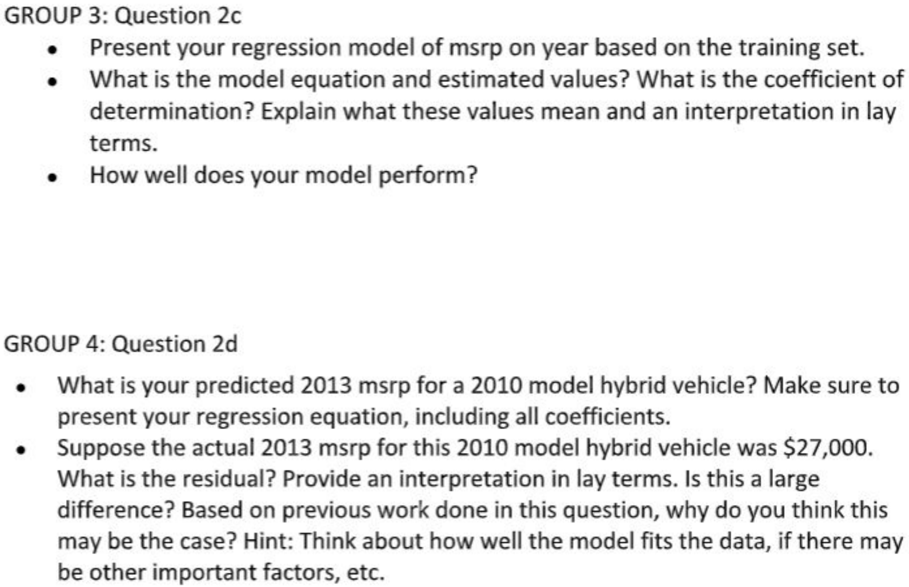


GROUP 1: Questions 1a and 1c

- 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?

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.

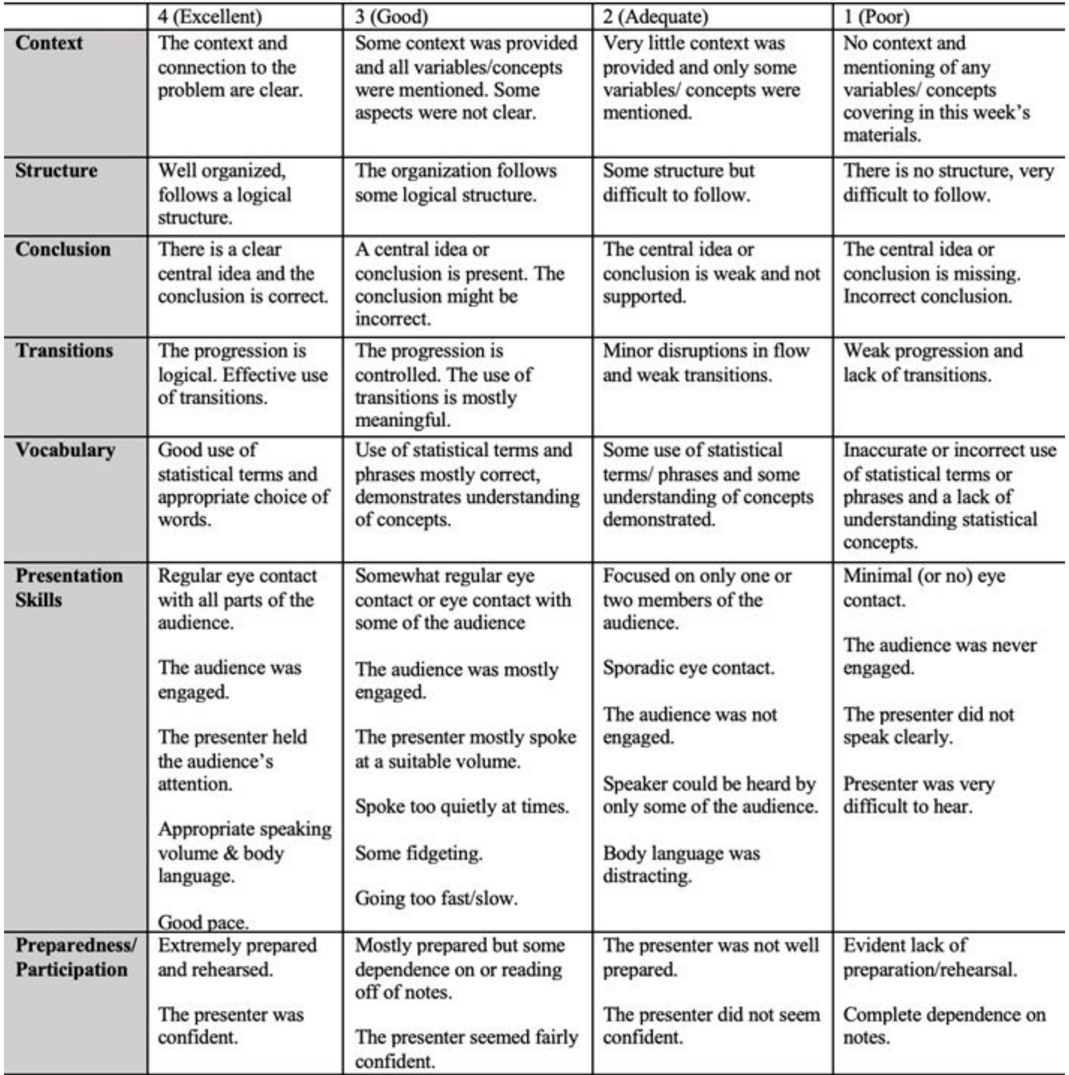


GROUP 3: Question 2c

- 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?

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.