STA130H1S TUT0109
W10: Multiple Linear Regression

Mar 22, 2019

(Materials used in this presentation are provided by the UofT Statistical Sciences Department)
This presentation was prepared by Sonia Chhay.

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

Overview

- Vocabulary
- This week’s material/homework
- About the poster project
- Group work and writing activity

Vocabulary for this Week

Multiple linear regression	Categorical predictor
Dummy variable	Baseline value
Interaction	Modification/ Modifier
Simpson's paradox	Extrapolation

Multiple Linear Regression

- Attempts to model relationship between 2/+ explanatory variables (x) and a response variable (y)
  - Does so by fitting a linear equation to the observed data
- Formally, the model for multiple linear regression, given *n* observations is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$ For *i* = 1, 2, ..., n

- x_i = Categorical predictors encoded as indicator variables
  - (aka dummy variables)
  - Takes value of 0 or 1 to indicate absence or presence of some categorical effect
  - The level corresponding to x=0 is the baseline value

Multiple Linear Regression

```
> summary(lm(BMI ~ gender, data=body))$coefficients
             Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 22.27793  0.1886217 118.109055 0.000000e+00
genderMale   2.43330  0.2702385   9.004269 4.435794e-18
```

- In R output, the level which is missing from the output is the baseline
- E.g. Looking at gender and given only the levels: male and female (where x=0 if individual is female)

Dummy Coding

- Regression is used often for continuous variables
  - E.g. Using hours spent studying to predict GPA
  - Increased study times correspond with increased GPA
- Dummy coding allows us to incorporate categorical variables into regression analysis

Homework Q3 - Dummy Coding

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

where y_{1i} is the BMI of individual i,
x_{1i} is the waist circumference of individual i and
x_{2i} = I(individual i is male).

- Consider a regression model used to predict BMI on waist circumference and gender (where female is the baseline value)
- Baseline value is not included in the model
  - Doing so would give the regression redundant information

```
> mod1 <- lm(BMI ~ waist + gender, data = body)
> summary(mod1)$coefficients
               Estimate  Std. Error    t value      Pr(>|t|)
(Intercept)  0.7454759 0.651078332   1.144987  2.527582e-01
waist        0.3084726 0.009205015  33.511364 2.358772e-130
genderMale  -2.1104205 0.202610484 -10.416146  3.845116e-23
```

The fitted regression line was: $\hat{y} = 0.745 + 0.308x_1 - 2.110x_2$

Homework Q3 - Dummy Coding

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_4 x_{4i} + \epsilon_i$$

where y_{1i} is the BMI of individual i,
x_{1i} is the waist circumference of individual i and
$x_{4i} = I(\text{individual i is female})$.

- Suppose that instead of using “female” as the baseline value, we decided to use “male” as the baseline. Carefully define a new predictor x_4, write down the regression model, and the new fitted regression line.

```
> mod1 <- lm(BMI ~ waist + gender, data = body)
> summary(mod1)$coefficients
               Estimate  Std. Error    t value      Pr(>|t|)
(Intercept)  0.7454759 0.651078332   1.144987  2.527582e-01
waist        0.3084726 0.009205015  33.511364 2.358772e-130
genderMale  -2.1104205 0.202610484 -10.416146  3.845116e-23
```

Estimation vs Prediction

- Estimation: Uses data to guess at a parameter
  - Interpolation
- Prediction: Uses fitted model to predict *y* for values of predictors that are *not* part of the dataset
  - Extrapolation
  - Potentially dangerous because we don’t know if the fitted regression model predicts *y* for other predictor variables

Assumptions Regression Does *Not* Make

- The response variable, Y, doesn't need to be normal as this is not an assumption of multiple regression
  - In fact, Y will rarely be normal
  - What must be true is that the ***errors*** around a prediction Y^ must be ***normal***
    - This can be checked with a normal plot of the residuals
  - Errors must also have a ***constant variance***
    - Can check with a predicted by residual plot (more on this in future stats classes)
- Regression does not assume that the regressors have any distribution
  - However, you should use box plots to check for outliers in the regressors
  - You need a very good reason to remove an outlier!

Simpson’s Paradox

- Don’t worry too much about it yet
- Confounding (and ethics) will be covered next week

Poster Project

Things to note, marking + details (rubric)

POSTER PROJECT: Things to Note

- Example poster project question:
  - With respect to the metrics outlined, it would be interesting to look at how (or whether) the November 21 launch of request expiry has changed user behaviour.
- Required to give a 5-minute (max) presentation on your project
- Will be marked by 2 TAs and your peers
- There will be time for questions after the presentation
- Poster board with velcro strips to hang up your printed sheets will be provided

Marking for Oral Presentations

The oral presentation will be marked on several aspects:

- Preparedness
- Speech clarity
- Content clarity
- Use of transitional words/ phrases
  - (Refer to the Resources uploaded to Quercus - Under Course Syllabus and Help)
- Use of statistical vocabulary
- Delivery
- The WOW factor. Extra points for doing something very impressive!

This + more important information available at: https://q.utoronto.ca/courses/78086/pages/project

Rubric for Oral Presentation

		Excellent	Good	Adequate	Poor
Content	Reasonable scope	The scope of the analysis is clear and questions can be fully addressed using the available data.	The scope of the analysis is clear and questions can be reasonably addressed using the available data.	The scope of the analysis is less clear, the questions can somewhat be addressed using the available data with slight modifications.	The questions are beyond the scope, cannot be reasonably addressed with the available data; need to resort to additional data or complete modification.
	Data wrangling	Creative use of data wrangling to produce informative variables.	Appropriate use of data wrangling to create sensible variables.	Some use of data wrangling to create new variables.	No evidence of data wrangling to create any variables.
	Graphical display	Choice of graphs are appropriate and creative; graphs reveal useful information and tell a story. Meaningful captions and titles.	Choice of graphs are appropriate; graphs reveal useful information, but are not self-sufficient. Might require some explaining.	Choice of graphs are appropriate; graphs reveal some useful information. Might require some explaining and minor changes to titles/axes/labels, etc.	A lack of visual aid; graphs are inappropriate, reveal no information.
	Statistical methods	The choice of methods is appropriate; analyses are complete; diverse and creative use of more than one approach.	The choice of methods is appropriate; some non-essential analyses are missing.	The choice of methods is somewhat appropriate; some analyses are missing.	The choice of methods is inappropriate; essential analyses are missing.
	Appropriate conclusion	Results are clearly and completely summarized. Appropriate limitations and concerns are clearly stated.	Results are completely summarized. Some limitations and concerns are stated.	Some results are summarized. The conclusion is not appropriate and no mentioning of any limitations.	Results are not summarized and conclusion is missing.
Writing	Organization	Contents are very well organized under the appropriate section and subsection headings.	Contents are organized under the appropriate section and subsection headings.	Contents are somewhat organized under section and subsection headings.	Contents are poorly organized under section and subsection headings.
	Overall Writing	Very polished and well written.	Few errors in spelling, punctuation, and/or grammar. Mostly clear and understandable.	Partly unclear, but mostly understandable. Several errors in spelling, punctuation, and/or grammar.	Too many errors in spelling, punctuation, and/or grammar, which make it unclear and difficult to follow.
Wow factor	Has this poster impressed you? Comment on the following (all that apply): - Overall creativity - Use of additional sources of data - Creation of additional variables (for example, through categorizing numerical variables)				

REMINDER:

- Every group member is expected to speak.
- Members who do not speak will be penalized.
- See the poster page on Quercus for further details.

Marking Details

Your poster **(rmd and pdf)** will also be marked on the following aspects:

Reasonable Scope	Research question should be **clear and answerable**
Data Wrangling	Creative use of data wrangling to produce **informative variables**. - Think about if your variables make sense and how they are useful for answering your research question.
Graphical Display	Use **appropriate (and creative) figures** to tell your story. - Remember, these need to have **clear and meaningful titles** (i.e., don't label with an R variable name if it doesn't make sense). - CRITICALLY, your **figures need to stand alone**. Note that providing pages of data tables is not very meaningful for your reader. - It's your job as a data scientist to turn this into something your reader can easily understand.
Statistical Methods	The choice of methods should be **appropriate, complete and creative**. - You shouldn't do EVERYTHING you've learned this term, but pick the methods that make the most sense. - You will be penalized if you use EVERY method – because this doesn't make sense. Also, think about what a "method" is. - **Using an R command is NOT a method.** - Developing a linear regression model to study the relationship between X and Y, for example, IS a method.

Marking Details

Your poster **(rmd and pdf)** will also be marked on the following aspects:

Appropriate Conclusion	Your **conclusion should clearly follow the work** you've done and **your results**. - State any limitations. - For example, maybe there is something you'd like to study but it is not available in the data.
Organization	The contents should be **ordered logically**, use of **sub-headers** is recommended. - E.g. Background, Methods, Results, Discussion, Conclusions.
Overall Writing	The #1 mistake we see if that people do not **proofread** their work.
WOW Factor	Bonus points for doing something very creative, - E.g. using an additional data source, creating a new (meaningful) variable, etc.

REMEMBER:

- The poster needs to stand on its own (i.e. Clearly explains the information)
- Your TA will be marking the poster and will not be attending your presentation

- Your rmd file will also be marked, so make sure it runs smoothly
- We strongly suggest annotating your code (using comments) so that your TA can navigate it better

Other Things to Note

- Pay attention to the **deadlines** for submitting your work. These are **STRICT**.
- Submit **1 poster per group** through Quercus (Quercus groups to be made soon).
- Each of you will also be **acting as a peer reviewer** (rubrics will be available on poster day and to be submitted **BY THE STUDENT** through Quercus).
- Arrive on time to put up and take down your poster. Any work left will be discarded.
- **Arrive early** for your scheduled poster presentation time. **If you arrive late, you will not be marked.** This is a large class which requires a lot of coordination. As such, you have a designated presentation time for a reason.

Other Things to Note

- We will have extra TAs in the Stats Aid center the week before the poster to help you. **Come prepared.**
  - It is NOT the TA's job to work on your project for you!
- We can also answer questions through Piazza all week, but this is not an appropriate forum for extensive programming help.
- **DO NOT LEAVE THE PROJECT UNTIL THE LAST MINUTE.** It takes time to do this properly and if you run into any last-minute issues, we will not be there to help you.
- **PRACTICE, PRACTICE, PRACTICE!** :D Ideally, **practice as a group** as many times as you can so that your transitions seem more natural.

Group Discussion

Homework Q1a(ii)
15 min

Homework Q1c
15 min

Homework Q1a(ii)

Sellers on eBay have the option to include a stock photo as the illustration of the product for sale. Does this choice affect the selling price?

ii. Carry out an hypothesis test to investigate whether the mean selling price is the same for sellers who do and do not use stock photos. Assuming the conditions necessary for the inference procedure to be valid are reasonable in this situation, What do you conclude? How could you apply a method from earlier in the term to carry out this hypothesis test?

For your reference:

```{r}
library(openintro)
marioKart2 <- marioKart %>% filter(totalPr < 100)
summary(lm(totalPr ~ stockPhoto, data=marioKart2))$coefficients
```

```
                Estimate Std. Error   t value     Pr(>|t|)
(Intercept)   44.327222   1.493540 29.679305 5.092241e-62
stockPhotoyes  4.169159   1.730739  2.408889 1.731116e-02
```

Homework Q1c

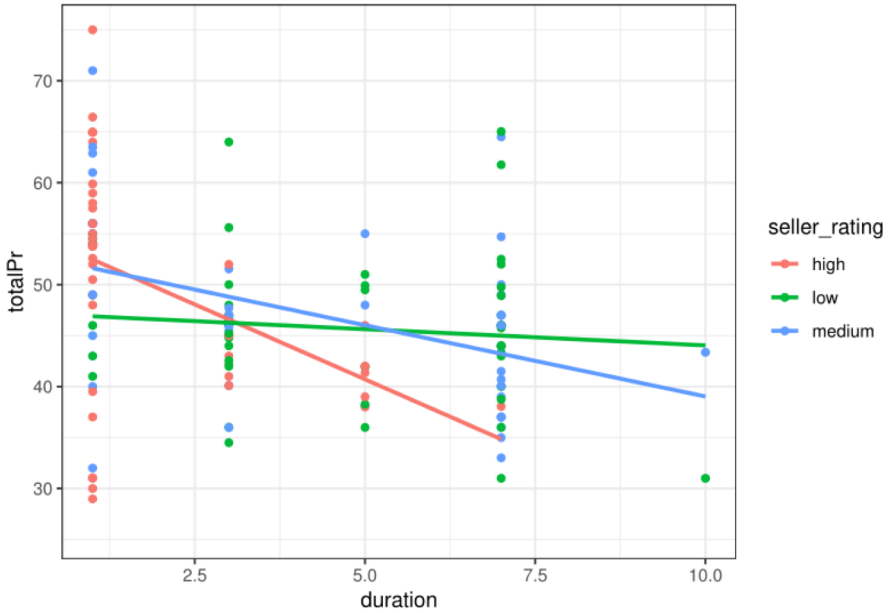
(c) Now produce and appropriate plot and fit an appropriate regression line to examine whether `seller_rating` has an effect on the relationship between `totalPr` and `duration`.

The regression model is

$$totalPr = \beta_0 + \beta_1 seller\_rating\_is\_low + \beta_2 seller\_rating\_is\_medium + \beta_3 duration + \beta_4 seller\_rating\_is\_low \times duration + \beta_5 seller\_rating\_is\_medium \times duration + \epsilon$$

i. Is there evidence of a difference between sellers with low and high ratings in the relationship between 'totalPr' and 'duration'?
ii. Is there evidence of a difference between sellers with medium and high ratings in the relationship between 'totalPr' and 'duration'?
iii. Is there evidence of a difference between sellers with low and medium ratings in the relationship between 'totalPr' and 'duration'? iv. Briefly explain a way to obtain p-values for the tests needed to address iii?

For your reference:




```
summary(lm(totalPr ~ seller_rating*duration, data=marioKart2))$coefficients
```

```
##                                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)                  55.399199  1.9003593 29.1519610 4.101299e-60
## seller_ratinglow             -8.185758  3.6192132 -2.2617507 2.531113e-02
## seller_ratingmedium          -2.387931  3.0064661 -0.7942651 4.284351e-01
## duration                     -2.937082  0.7652626 -3.8380058 1.897989e-04
## seller_ratinglow:duration     2.620252  0.9533562  2.7484504 6.807181e-03
## seller_ratingmedium:duration  1.538756  0.8856835  1.7373654 8.460333e-02
```

1-Paragraph Written Response

Based on Practice Problem #2
30 min

Writing = Clear, Concise, Cohesive

1. The purpose. What are you studying? Why should we care about the analysis you've done?
2. A summary of the methods used. What did you do? Why did you do it this way?
3. A summary of the results. We don't need to know everything you found – only the most critical things relating to your purpose! Remember, sometimes less is more! If they include a figure, it should be clear and able to stand on its own (e.g. contain proper titles).
4. A conclusion. What is your take away message? Remember, a conclusion is not the place to present new findings.

Writing Activity

Using R, you fit two regression models that you defined in (2c). Which model do you think would best explain the association between temperature and yield? Think about the shape of the association and any model statistics that may be relevant. Are there any limitations to these models?

Remember to mention:

- Your research question (introduction)
- The methods you applied
- Summary of results
- Conclusion

(a) Plot the data and superimpose the best (straight line) linear predictor for yield, based only on the temperature. Does this line capture the relationship between temperature and yield?

```
yield %>% ggplot(aes(x=temperature, y=yield)) + geom_point() +
  geom_smooth(method="lm", se=FALSE)
```