STA 130 Tutorial 9

David Veitch

University of Toronto

daveveitch.github.io

Mar 22, 2019

Agenda

1. Poster Grading - Key Points
2. Vocabulary for the Week
   - Dummy Variables
   - What Makes Linear Regression Linear?
   - Correlation $\neq$ a Good Model
   - Kahoot
3. Discussion
4. Writing Exercise

Poster Grading - Key Points

- **Full Rubric**: on Quercus
- **Presentation**: during the poster fair, you will be visited by members of the STA130H1 teaching team and student evaluators. You will give them a 5-minute presentation about your work. Every member of the team is expected to speak.
- **Data Wrangling**: creative use of data wrangling to produce informative variables
- **Statistical Methods** do not do EVERYTHING you have learned, only what makes sense
- **Organization** recommended to use subheaders (e.g. background, methods, results, discussion, conclusion)
- **Poster**: must stand on its own (i.e. TA marking just the poster can understand it)
- **Timing**: do not leave to last minute; data wrangling in particular takes a long time

Dummy Variables

$$\text{Model: } y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- y_i response variable, body mass index, for observation i
- x_i is a covariate, male or female, for observation i
- β's are parameters
- ϵ_i random error for observation i

Since x_i is a category it takes the form:

$$x_i = \mathbb{1}(\text{individual i is male}) = \begin{cases} 0 & \text{observation } i \text{ is female} \\ 1 & \text{observation } i \text{ is male} \end{cases}$$

Note you only need one indicator variable for two categories. In general you need $p-1$ indicators for p categories. The one you do not have an indicator for is the reference category.

What Makes Linear Regression Linear?

The **linear** in linear regression indicates the equation is linear in the parameters (the β's).

Example of linear model:

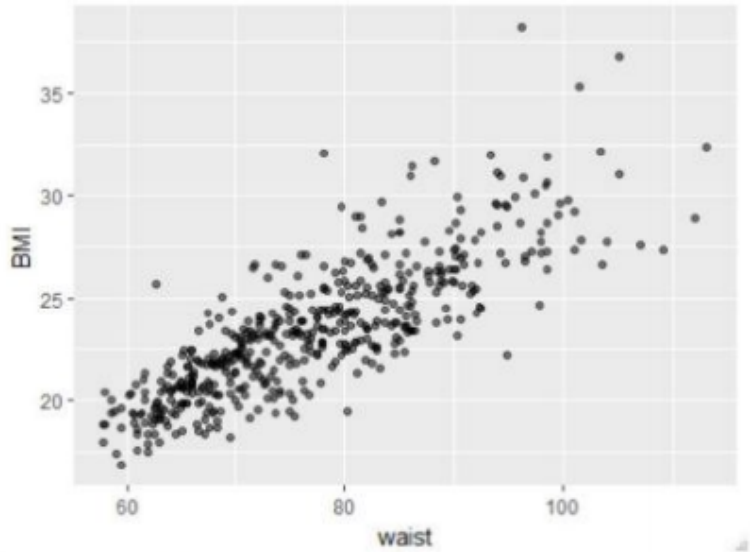
$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}^2 + \beta_3 \log(x_{3,i}) + \epsilon_i$$

Example of a nonlinear model:

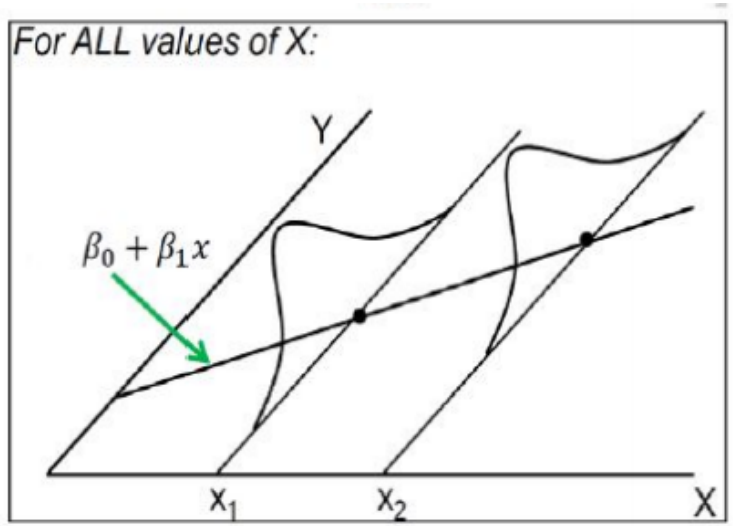
$$y_i = \frac{\beta_0 + \beta_1 x_{1,i}}{1 + \beta_2 x_{2,i}}^{\beta_3} + \epsilon_i$$

Regression Assumptions

- **A1**: Linear pattern between x and y (if you suspect its non-linear you can still do regression if you properly transform your x)
- **A2**: Constant variance in y for all values of x
- **A3**: Independent Observations
- **A4**: Residuals follow a normal distribution



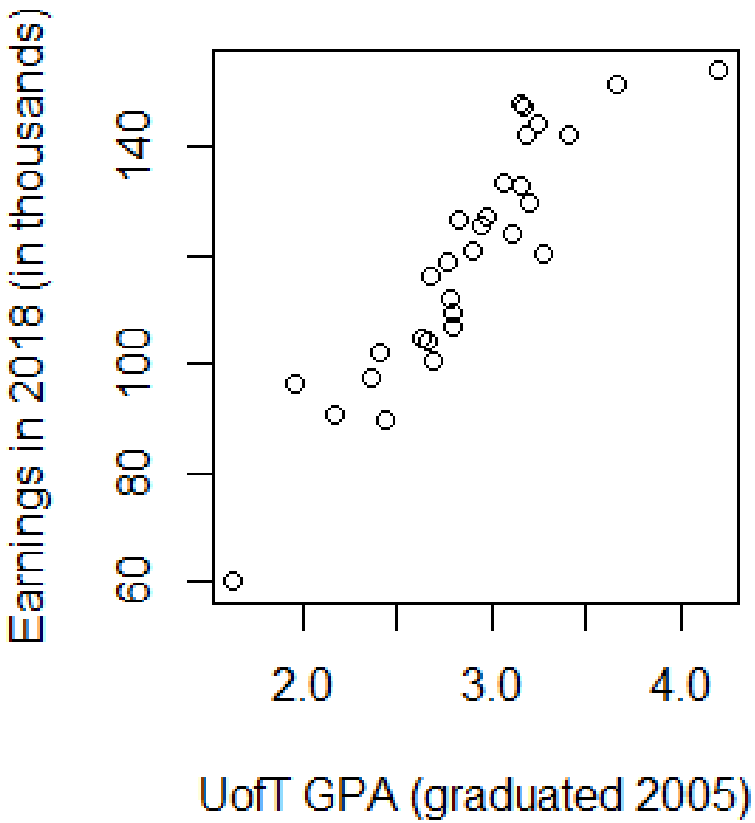
(a) **A2** Violated

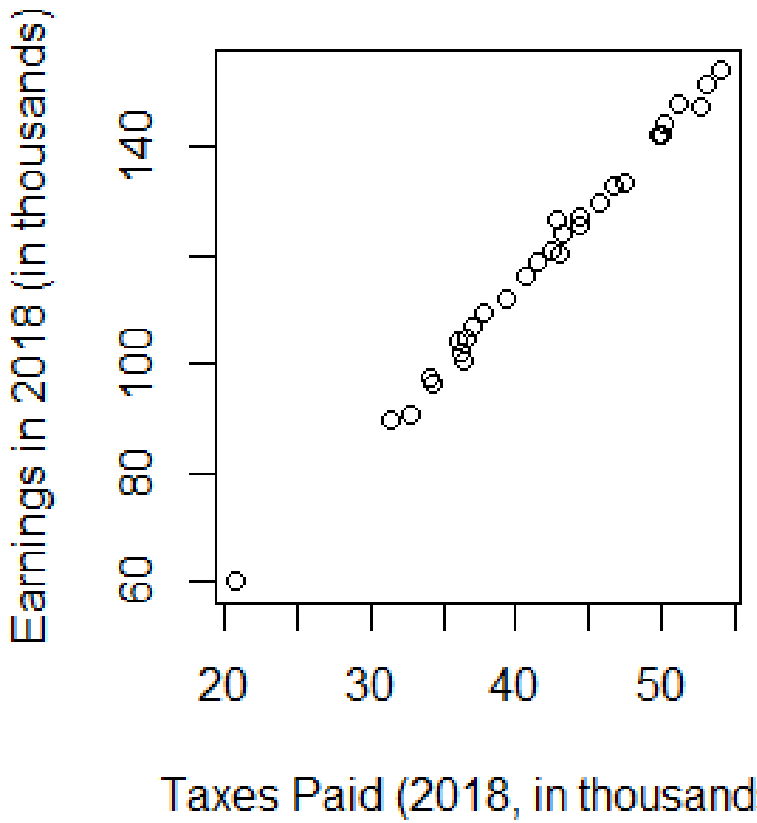


(b) **A3** Met

Correlation $\neq$ a Good Model

A UofT student wants to know the one factor that can help them earn lots of money. You do some analysis and come up with the following plots. What do you tell the student?



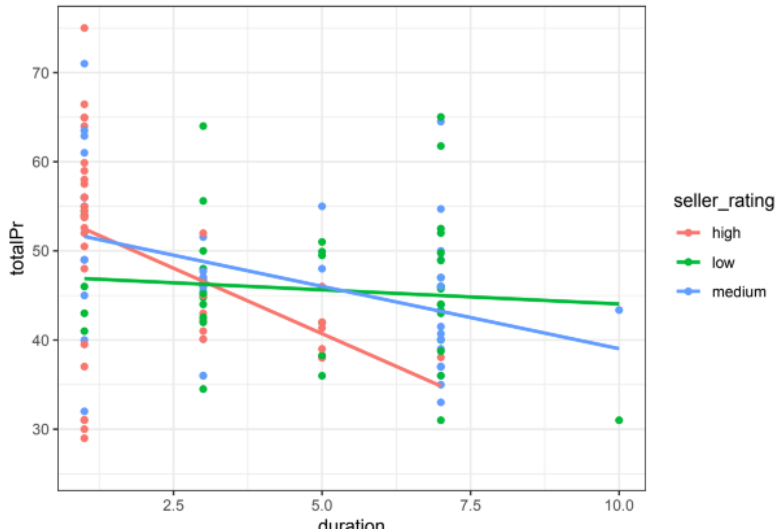


Kahoot

Question 1a ii from the homework

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?

Question 1c from the homework




```
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
```

For the relationship between totalPr and duration is there a difference between: sellers with low and high ratings, sellers with medium and high ratings, sellers with low and medium ratings?

Reminder of What to Include

- Purpose
- Summary of Methods Used
- Summary of the Results
- A **clear** Conclusion

Writing Exercise

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 and the methods you applied.