Midterm Q2

Name______________________________________Course_________________

1 Q1 in other file

2 Starwars Characters

Use R or Python for data analysis. Show and interpret the outputs.

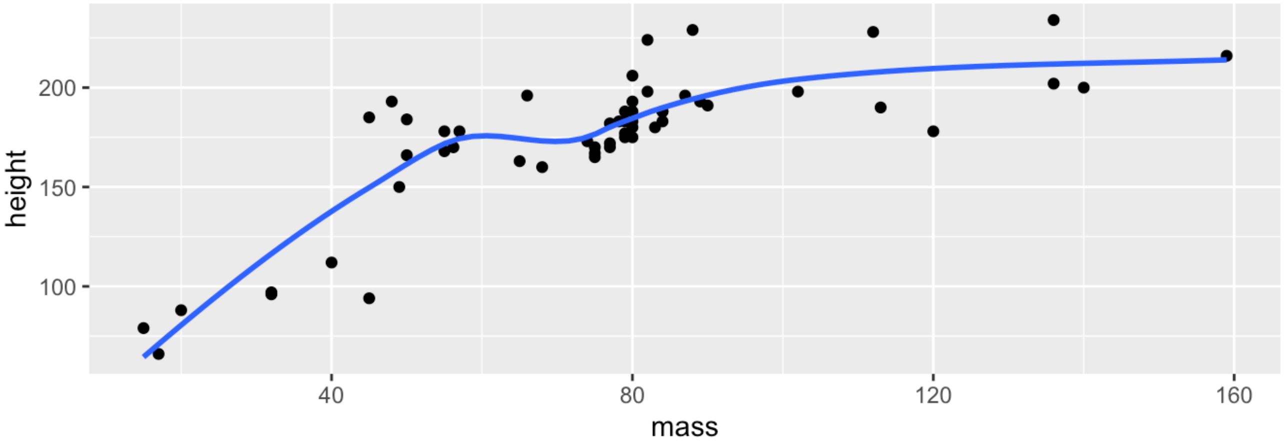
The {dplyr} package contains a data frame with attributes of characters from the `starwars` series. See help or `?dplyr::starwars` for information on the variables.

We want to see if we can predict a characters height based on their mass and possibly other attributes.

Due to the presence of NAs and an extreme value (Jabba the Hut is eight times as massive as the next closest character), we will filter out those rows.

(a) Filter the data. Create a scatter plot showing `height` vs `mass` with a non-linear smoother. Interpret the plot in one sentence.

~~~
```{r}
#| include: true
#| message: false
#| fig-height: 2.5
library(tidyverse)
starwars |>
  filter(mass < 500) |>
  drop_na(sex, mass, height) ->
  sw
ggplot(sw, aes(mass, height)) +
  geom_point() +
  geom_smooth(se=FALSE)
```
~~~



(b) Fit a linear regression model to predict height based on mass.

- Identify the proposed model regression equation and then insert the estimated coefficients into the regression equation
- Is mass significant in this prediction? Justify your answer using the p-value.
- Is the model useful for prediction? Justify your answer using the adjusted r-squared.

(c) Test if the modeled relation between height and mass` is *truly linear* or could a non-linear model could be better.

- Describe your approach and why it will help answer the question
- State the test statistic, the p-value, and your conclusion.

(d) Add a term for $mass^2$ to the data to add some nonlinearity. Rerun the model with both `mass` and $mass^2$.

~~~
```{r}
#| include: true
sw$mass2 <- sw$mass^2
```
~~~

- Is the model more useful?
- Justify your answer with the results from a statistical test comparing the original model, with just `mass`, to the model with both `mass` and $mass^2$.
- Would it be valid to use the same statistical test to compare two models where one has just `mass` and one has just $mass^2$?

(e) With just `height` and `mass` (not `mass`^{2}), use a validation-set approach (split at 60%), LOOCV, and K-fold cross-validation (**K = 11**) to estimate the **prediction mean squared error** for each method.

- Treat these as independent runs and set the random number seed to 124 when appropriate to ensure repeatability.
- Summarize the answers at the end.

(f) Can we improve prediction accuracy of the model by including the sex of the character (variable `sex`)?

- Calculate the prediction MSE of the expended model with `mass` and `sex` using K-fold cross-validation where $K = 11$.
- Note that `sex` is a categorical variable.
- Set the random number seed to 124 when appropriate to ensure repeatability.

(g) Interpreting Interactions: Include `mass`, `sex`, *and their interactions* in the linear model with all the data.

- Compare the summaries for both the model without and with interactions. What do you observe?
- For the same value of mass, which sex is expected to be the tallest and why?
- Does the answer depend on the value of mass? Explain why or why not.
- Suggest looking at a plot or the data as well as the coefficients and p-values.

(h) Extra Credit 1 PT

Describe how a plot of the fitted values for a model with a quantitative response, a quantitative predictor, and a categorical predictor with three levels might appear if the predictors and the interaction affects are significant. How you would interpret it.