Multiple Linear Regression

Mini-Assignment - MTH 361 A/B - Spring 2023

Instructions:

- Please provide complete solutions for each problem. If it involves mathematical computations, explanations, or analysis, please provide your reasoning or detailed solutions.
- Note that some problems have multiple solutions or ways to solve it. Make sure that your solutions are clear enough to showcase your work and understanding of the material.
- Creativity and collaborations are encouraged. Use all of the resources you have and what you need to complete the mini-assignment. Each student must take personal responsibility and submit their work individually. Please abide by the University of Portland Academic Honor Principle.
- **Please save your work as one pdf file, don't put your name in any part of the document, and submit it to the Teams Assignments for this course. Your document upload will correspond to your name automatically in Teams.**
- If you have questions or concerns, please feel free to ask the instructor.

I. Introduction to Multiple Regression

Materials

The exercises below are derived from the textbook OpenIntro Statistics (4th edition) by David Diez, Mine Cetinkaya-Rundel, and Christopher Barr.

Exercises

1. **Baby weights, Part I.** The Child Health and Development Studies investigate a range of topics. One study considered all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area. Here, we study the relationship between smoking and weight of the baby. The variable *smoke* is coded 1 if the mother is a smoker, and 0 if not. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, based on the smoking status of the mother. ("Baby Weights Data Set," n.d.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	123.05	0.65	189.60	0.0000
smoke	-8.94	1.03	-8.65	0.0000

The variability within the smokers and non-smokers are about equal and the distributions are symmetric. With these conditions satisfied, it is reasonable to apply the model. (Note that we don't need to check linearity since the predictor has only two levels.)

   a. Write the equation of the regression model.
   b. Interpret the slope in this context, and calculate the predicted birth weight of babies born to smoker and non-smoker mothers.
   c. Is there a statistically significant relationship between the average birth weight and smoking?

The previous table introduces a data set on birth weight of babies. Another variable we consider is *parity*, which is 1 if the child is the first born, and 0 otherwise. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, from *parity*.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	120.07	0.60	199.94	0.0000
parity	-1.93	1.19	-1.62	0.1052

   d. Write the equation of the regression model.
   e. Interpret the slope in this context, and calculate the predicted birth weight of first borns and others.
   f. Is there a statistically significant relationship between the average birth weight and parity?

2. **Baby weights, Part II.** We considered the variables **smoke** and **parity**, one at a time, in modeling birth weights of babies in (1). A more realistic approach to modeling infant weights is to consider all possibly related variables at once. Other variables of interest include length of pregnancy in days (*gestation*), mother's age in years (*age*), mother's height in inches (*height*), and mother's pregnancy weight in pounds (*weight*). Below are three observations from this data set.

	bwt	gestation	parity	age	height	weight	smoke
1	120	284	0	27	62	100	0
2	113	282	0	33	64	135	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1236	117	297	0	38	65	129	0

The summary table below shows the results of a regression model for predicting the average birth weight of babies based on all of the variables included in the data set.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-80.41	14.35	-5.60	0.0000
gestation	0.44	0.03	15.26	0.0000
parity	-3.33	1.13	-2.95	0.0033
age	-0.01	0.09	-0.10	0.9170
height	1.15	0.21	5.63	0.0000
weight	0.05	0.03	1.99	0.0471
smoke	-8.40	0.95	-8.81	0.0000

a. Write the equation of the regression model that includes all of the variables.
b. Interpret the slopes of *gestation* and *age* in this context.
c. The coefficient for *parity* is different than in the linear model shown in (1). Why might there be a difference?
d. Calculate the residual for the first observation in the data set.
e. The variance of the residuals is 249.28, and the variance of the birth weights of all babies in the data set is 332.57. Calculate the R^2 and the adjusted R^2. Note that there are 1,236 observations in the data set.

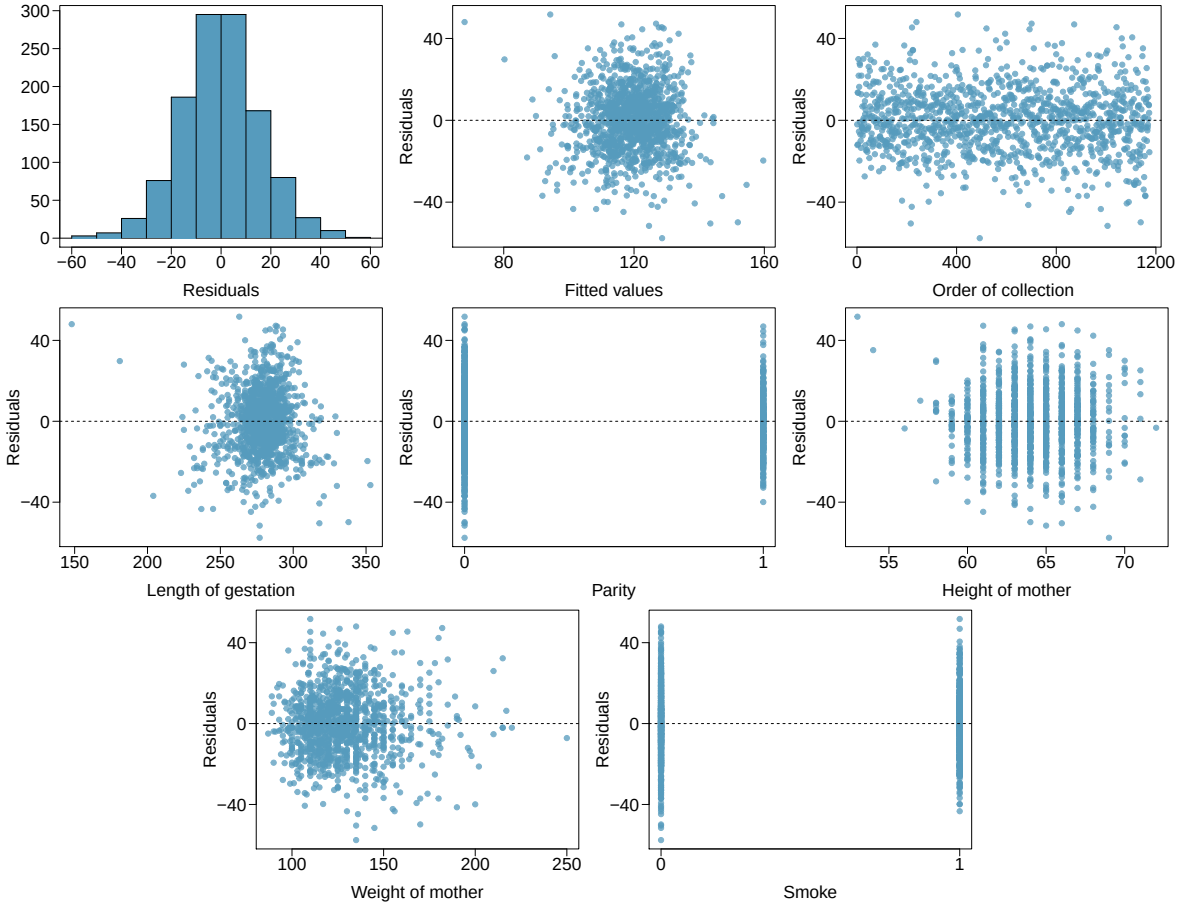
3. (Outstanding Question) **Baby weights, Part III.** Exercises (1) and (2) considers a model that predicts a newborn's weight using several predictors (gestation length, parity, age of mother, height of mother, weight of mother, smoking status of mother). The table below shows the adjusted R-squared for the full model as well as adjusted R-squared values for all models we evaluate in the first step of the backwards elimination process.

	Model	Adjusted R^2
1	Full model	0.2541
2	No gestation	0.1031
3	No parity	0.2492
4	No age	0.2547
5	No height	0.2311
6	No weight	0.2536
7	No smoking status	0.2072

   a. Which, if any, variable should be removed from the model first?

The table shown above presents a regression model for predicting the average birth weight of babies based on length of gestation, parity, height, weight, and smoking status of the mother.

   b. Determine if the model assumptions are met using the plots below. If not, describe how to proceed with the analysis.



References

Baby weights data set. (n.d.). In *Child Health and Development Studies*.