

Project #4: Simple Linear Regression

For this assignment use the data set **ncbirth200.sav**. Recall the variables from the North Carolina birth data set are:

The variables examined are:

Variable Label	Description
plurality	Number of children born of the pregnancy
sex	Sex of child (1=Male, 2=Female)
mage	Age of mother (years)
weeks	Completed Weeks of Gestation (weeks)
marital	Marital status (1=married, 2=not married)
racemom	Race of Mother (0=Other Non-white, 1=White, 2=Black 3=American Indian, 4=Chinese, 5=Japanese, 6=Hawaiian, 7=Filipino, 8=Other Asian or Pacific Islander)
hispmom	Mother of Hispanic origin (C=Cuban, M=Mexican, N=Non-Hispanic, O=Other and Unknown Hispanic, P=Puerto Rican, S=Central/South American, U=Not Classifiable)
gained	Weight gained during pregnancy (pounds)
smoke	0=mother did not smoke during pregnancy 1=mother did smoke during pregnancy
drink	0=mother did not consume alcohol during pregnancy 1=mother did consume alcohol during pregnancy
tounces	Weight of child (ounces)
tgrams	Weight of child (grams)
low	0=infant was not low birth weight 1=infant was low birth weight
Premie	0=infant was not premature 1=infant was premature premature defined at 36 weeks or sooner

Answer the following for the variables **tounces** (dependent variable) and **mage** (independent variable).

- a. Make a scatterplot of this data. Fit the regression line. Report the parameter estimates (the estimates of the intercept and slope).
- b. Is **mage** useful in predicating **tounces**? Why? Report the level of significance (P-value).
- c. What percentage of the variation in **tounces** is explained by **mage**? Is that high or low?
- d. What is the predicted value for **tounces** when **mage** is 35? What if **mage** is 17?
- e. Make a residual plot. Comment on the fit of the model.
- f. Determine how many outliers there are and identify the observation numbers.

Answer the following for the variables **tounces** (dependent variable) and **gained** (independent variable).

- a. Make a scatterplot of this data. Fit the regression line. Report the parameter estimates.
- b. Is **gained** useful in predicating **tounces**? Why? Report the P-value.
- c. What percentage of the variation in **tounces** is explained by **gained**? Is that high or low?
- d. What is the predicted value for **tounces** when **gained** is 10? What if **gained** is 30?
- e. Make a residual plot. Comment on the fit of the model.
- f. Determine how many outliers there are and identify the observation numbers.

Answer the following for the variables **tounces** (dependent variable) and **weeks** (independent variable).

- a. Make a scatterplot of this data. Fit the regression line. Report the parameter estimates.
- b. Is **weeks** useful in predicating **tounces**? Why? Report the P-value.
- c. What percentage of the variation in **tounces** is explained by **weeks**? Is that high or low?
- d. What is the predicted value for **tounces** when **weeks** is 35? What if **weeks** is 40?
- e. Make a residual plot. Comment on the fit of the model.
- f. Determine how many outliers there are and identify the observation numbers.

Make a scatter plot of **tgrams** (dependent variable) and **tounces** (independent variable). What happens? Are you surprised? Why or why not? Explain why the R-squared value is 1.0 when **tounces** predicts **tgrams**.

Write a paragraph summarizing your conclusions of the four analyses done above. Be sure to interpret the meaning of each. Which results are most useful, which are most surprising?

Suppose a 30 year old woman, who carried her baby 38 weeks, and gained 20 pounds during pregnancy wanted you to predict her child's birth weight. What answer would you give and why?