Introducing linear regression

Helen Burn

6/03/2021

Showing relationships for discussion

Open up the linear regression Little App. (See footnote[1]). In this activity we'll explore three different data sets available through the app. For each data set, the variable indicated below plays the role of the response variable. Some explanatory variables are listed as sub-points.

1. In the Data Tab in the top toll bar, set Source Package to `Little Apps`. For Data set choose `NHANES2`. Response variable: BMI. It's important for students to know what BMI is. Explanation from the CDC & BMI calculator for students.
    - age (r = 0.5 reasonable scatterplot to assume linearity)
    - income (r = -0.07) shows a very diffuse scatter plot but also helps demo the app to students.
    - pulse: weak relationship
    - systolic: weak-to-moderate relationship
    - diastolic: has outliers
    - sleep_hour: weak-to-moderate. But has a negative relationship
2. Set Source Package to `mosaic`, and then Data set choose `CPS85`. Response variable: wage
    - age
    - education
3. Set Source Package to `Little Apps`. For Data set `Natality_2014` Response variable: `mager`, mother's age
    - `fagecomb`, father's age. Moderate size correlation. Ask what real-world phenomenon accounts for the correlation.

Open-ended exploring

Systolic blood pressure from the `NHANES` data.

Background: Explain to students what is the difference between the systolic and diastolic blood pressure. Each time the heart beats, the blood pressure in the arteries goes up. It quickly rises to a maximum and then decays until the next beat. Systolic is the maximum blood pressure each beat, diastolic the minimum. The "pulse pressure" is the difference between the two. See this site on blood pressure.

Tasks

1. Determinine three explanatory variables that are predictive of systolic blood pressure.

Write down the names of the explanatory variables here . . .

2. In the Graph tab in the upper tool bar, make sure the "Show auxiliary graph" box is checked.

The primary plot (the large one on the left of the tab) shows a scatter plot of the response variable versus explanatory variable and the best fitting model. Imagine that each of the scattered points were raised or lowered to fall exactly on the best fitting model. This vertical position corresponds to the **model values**.

[1] https://maa-statprep.shinyapps.io/Little__App__Regression/

The auxiliary graph is on the right. It shows two clouds of points. The right cloud gives the **raw values** of the response variable. The vertical position of each point in the raw cloud is identical to the vertical position of the corresponding data point in the primary graph.

The left cloud in the auxiliary graph gives the **model values**. Again, the vertical position of each point in the model-value cloud is identical to the corresponding model value in the primary plot.

Both the raw and model-value clouds in the auxiliary graph are marked with an I-shaped interval. This vertical interval covers approximately 95% of the points in its cloud. The center of the interval is the mean points in the cloud, the ends are plus-or-minus 2 standard deviations away from this.

One helpful way to describe a relationship between two variables is to quantify how much of the *variation* in the response variable can be accounted for by the explanatory variable. A standard way to quantify this is with a statistic called R-squared, which always falls between 0 and 1. Zero means no relationship and 1 is a "perfect" relationship where the explanatory variable exactly accounts for the response variable. Think of R-squared as measuring the **strength of the relationship**. More precisely, R-squared is the fraction of the variance of the response variable accounted for by the explanatory variable.

You can estimate the strength of the relationship, R-squared, from the auxiliary graph.

> R is the ratio of the length of the model value interval to the length of the raw values interval.
> Square R to get R-squared.

Another way to describe the relationship between the explanatory and response variables is with the **effect size**. Whereas R-squared is always on the scale 0 to 1, the effect size reflects the actual units of the explanatory and response variables. It is the change in model value ("rise") per unit change in the explanatory variable ("run"). The ratio is rise over run, in other words, the slope of the model line.

For each of the three variables, list the strength of the relationship both as a fraction of the variation explained (R-squared) and as the change in systolic blood pressure per unit change of the explanatory variable (slope of model line).

Fill in the table with your answers. . . .

variable name	fraction of variation	change of response per unit change in explanatory

3. Then check whether those three explanatory variables explain diastolic blood pressure as well.

Which of systolic or diastolic blood pressure is better explained by the explanatory variables? . . .

Housing Prices

Set Source Package to `mosaic`, and then Data set choose `SaratogaHouses`. Response variable: `price`

a Determinine three explanatory variables that are predictive of house price.

Write down the names of the explanatory variables here . . .

b For each of the three variables , list the strength of the relationship both as a fraction of the variation explained and as the change in price of the explanatory variable.

Fill in the table with your answers. . . .

variable name	fraction of variation	change of response per unit change in explanatory

Version 0.4, 2021-06-03