Linear regression

Helen Burn

2019-04-15

Activities

- Introducing linear regression
- Describing relationship patterns
- How much is explained by regression?

Learning objectives connected to linear regression

- Create scatterplots for bivariate data using graphing technology where appropriate. Lesson: point plots
- Sensibly choose which variable should be the response and which the explanatory variable, and know when it does and doesn't matter. (Other nomenclature for explanatory/response: predictor/response or independent/dependent.) Lesson: response and explanatory variables
  - External evidence of which direction causation goes, e.g. hours work explains total pay.
  - Why are you making a prediction:
    * Deduce from something easy to measure, something that would be hard to measure. Often, it's a future value that we're interested in. Sometimes it can be a past or present value that we just don't know yet.
    * Hypothesis formation.
- Determine whether a straight-line model is appropriate for describing a given relationship. Lesson: flexibility
  - Students can distinguish between situations where the relationship is approximately linear and when it is not. Examples: Height versus age, BMI vs weight, BMI versus height (which has a crazy, upside-down whistle-shaped cloud)
  - residual, e.g. heteroscedasticity
  - covariate
- Interpret the correlation coefficient in terms of pos/neg/null and strength of correlation
- Use appropriately terms such as **equation**, **function**, **model**, **formula**
- Interpret the slope of the regression in terms of the relationship between incremental change in x and the corresponding incremental change in y.
- Translate a difference in the input to the corresponding difference in the output. (Rule of 4 from calculus reform.)
  - from the graph

  - from the regression b coefficient
  - Effect size,
  - What's a big change in input? (A couple of SD of x), What's a strong relationship: results in a big change in the output (e.g. SD of y). Correlation coefficient is directly in terms of translation of SD in input to SD in output.
- Identify the residual of a point given the location of the point and the regression function.
- Use the regression equation for prediction
  - plug in an input to get an output
  - recognize extrapolation as unsafe
  - proper prediction includes the residual variation around the model.
- Use technology to find linear regression models and correlation coefficients for a pair of variables Lesson: relationship-patterns
- Understand the pitfalls of extrapolation
- Be able to make a point plot using technology and to relate the location of each point to the corresponding row in a data table.
- Develop an intuition for how a mathematical function can describe the pattern in a point-plot cloud.
- Recognize settings and variables for which regression is an appropriate technique.
- Be able to use the slope as a concise description of a relationship.
- Recognize what residuals from a regression model have to say.
- Understand how a regression model can be used for prediction.
- Insofar as the correlation coefficient is topic in your course (and it need not be!) . . . establish the connections between regression slopes and correlation coefficients.

Additional resources

-
- Instructor orientation
- Role in statistical practice
- Classroom discussion
- Assessment
- Tips for an active classroom
- Student pre-requisites
- Looking forward
- Pitfalls

Orientation for instructors

Linear regression is one of the oldest and most widely used statistical techniques. It is used to describe or *model* a connection or relationship between a quantitative *response variable* and one or more *explanatory variables*.

Many, perhaps most, introductory statistics courses cover *simple regression*, which is a special case of linear regression in which the response variable, y is modeled as a straight-line function of the explanatory variable x, that is, $y = f(x) = ax + b$. The slope m and intercept b, constitute a concise but very limited way of describing important features of the relationship between the response and explanatory variables.

Role in statistical practice

It's fair to say that simple regression is too simple to support contemporary research and has been for some decades. It is uncommon for there to be just a single explanatory variable. A more general technique, *multiple regression*, supports the use of multiple explanatory variables.

Conceptual pitfalls

There are many potential pitfalls in teaching about simple regression. One has to do with nomenclature. Mathematicians describe a and b as "coefficients" or "parameters." In statistics, the meaning of "parameter" is different (referring to a population) and the values of a and b generated by regression are "statistics" (referring to a sample from the population). And a "coefficient" in a formula like $a + bx$ is not particularly similar to a "correlation coefficient."

Usually, the slope parameter b is the quantity of interest. The slope parameter is not, in general a number. Instead, it is a quantity expressed in units. Modeling spending versus age? Then b will have units like dollars-per-year.

Many instructors are tempted to use Greek-like notation in teaching regression. If you're going to use sophisticated mathematical notation to convey concepts, you are assuming your students know something about that notation. This might include:

- Greek letters and their Roman equivalents, e.g. distinguishing among β and B and b or between μ and m and remember that μ is not cognate to u.
- The different meanings of subscripts and superscripts, e.g. the distinct meanings of β^2 (exponentiation) and β_2 (identifying one in a series).
- The various (inconsistent and sometimes contradictory) notations for distinguishing between estimates and population parameters:
  - Parameters: β, μ, σ, and informally b, m, s
  - Estimates: $\hat{\beta}$, b, $\hat{b}$, $\hat{\mu}$, m, $\bar{m}$, s, $\hat{\sigma}$ It's unlikely that you intend for your students to have to deal with such complexity, so try to keep the notation as simple as possible. We suggest:

- b the slope of the regression line as estimated by data.
- R^2 the coefficient of determination
- r the correlation coefficient
- s_x and s_y standard deviations of the x and y variables

The *correlation coefficient* is a pure number that combines three pieces of information: b and the standard deviations s_x and s_y of the x and y variables. The relationship is

$$r = b\frac{s_y}{s_x}$$

Note that s_y has the same units as y and s_x the same units as x. Thus, the ratio $s_y/s + x$ cancels out the units of b.

In multiple regression, it makes sense to describe a model using the unit-ful coefficients like b, but there is no equivalent to the relationship between r and b in simple regression. Given the importance of multiple regression, it seems sensible to teach simple regression in terms of the unit-ful coefficient b rather than the unitless r.

Almost all statistics textbooks present r as a means to quantify the "strength" of the relationship between two quantitative variables. It is that, but it is equally applicable to situations where one or both of the variables are binomial, for instance yes/no or win/lose or A/B.

The analog to r in multiple regression is $\sqrt{R^2}$, where R^2, the *coefficient of determination* presents the fraction of the variance in the response variable that is captured by the model. R^2 ("R-squared") is an important summary description of a model. It makes sense, then to prepare students for R^2 by using it as a descriptive statistic even in simple regression. You might be tempted to refer to this as r^2, but do recall that R^2 is a more generally applicable statistic that encompasses the special case of r^2 in simple regression.

When we use coefficients like b to quantify a relationship, we set up an interpretation of b as as a kind of translation factor from x units to y units. That is, a one-unit increase in x is associated with a b-unit increase in y.

R^2 (or, r^2 if you insist) has a central role in statistical inference. The ratio

$$F = (n-1)\frac{R^2}{1-R^2}$$

is an informative quantity with respect to p-values and confidence intervals. For the p-value, an F of 3.84 corresponds to p = 0.05, an F of around 7 corresponds to p = 0.01, and an F of 12 to p = 0.001. (You can read off the p-value by looking up the quantile of F in the F distribution with 1 and $n-1$ degrees of freedom.)

The 95% confidence interval on b is

$$CI_{95} = b(1 \pm \sqrt{3.84/F})$$

Note that the t-statistic on b is simply $t = \sqrt{F}$. A reason to use F instead of t is that F generalizes to multiple regression while t does not.

The F statistic also generalizes to nonlinear formulas $y = f(x)$. Roughly speaking, for a quadratic shaped model, the $n-1$ term in F should be replaced by $\frac{n-2}{2}$.

Sometimes simple regression is presented as a way to *predict* a value of y given the value of x. This use is seriously misleading. Clearly you can plug a value of the explanatory variable x into the $a + bx$ regression formula. The number you get out will be the *most likely* single value. This is not a proper statistical prediction. The prediction should not be in the form of a single number. Instead, the prediction should take the form of a probability assigned to each possible outcome. In the case of simple regression, a meaningful prediction is that the output y for any given x is predicted to have the form of a normal distribution with mean $a + bx$ and a standard deviation corresponding roughly to the standard deviation of the residuals of the y-values from the corresponding model value.

Student pre-requisites

Students will need some background knowledge in order to follow lessons on simple regression.

- Variable types: quantitative and categorical Lesson: variable types
- Point plot: (The term “scatter plot” has traditionally been used.) Lesson: point plots
  - each axis corresponds to a variable
  - each row is one dot.
- Mathematical functions:
  - translate a given input to an output by plugging the input into an arithmetic formula
  - in writing the formula, we often use symbols, like m and b to represent quantities.
  - the straight-line function
    - slope (primary importance here)
    - intercept
- Understand distinctions between various reasons for examining relationship. Lesson: response and explanatory variables
  - to make a prediction of the unknown value of a variable given the known values of other variables
  - to anticipate the result of an intervention (This is a form of prediction that assumes a specific causal relationship)
  - to demonstrate that two variables are connected in some way.

  - to explore data in order to frame hypotheses about how the system works.
- Standard deviations if using r. This is not central if focusing on slope and intercept.

Creating an active classroom

See the document on general tips for creating an active classroom.
Some specific discussion topics/themes for linear regression:

1. BMI (from NHANES2) as a response variable. It's important for students to know what this 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. wage (from CPS85)
   - age
   - education
3. mother's age (from Births_2014)
   - father's age. Moderate size correlation. Ask what it means
4. **Open-ended exploring**
5. Consider **systolic blood pressure** from the NHANES2 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. Determine three explanatory variables that are predictive of systolic blood pressure.
     2. For each of the three, list the strength of the relationship both as a fraction of the variation explained as as the change in systolic blood pressure per unit change of the explanatory variable.
     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?
6. **Diamonds** similar to the above, but predict the price of a diamond.

Assessment items

- Point plot and functions. In which we'll ask students to sketch out some functions from prior knowledge (e.g. height versus age) and then indicate the range of values around the function. Then turn this around so that you deduce the function and range of residuals from the point plot.
- Explanatory vs response variable: prediction versus intervention vs description vs hypothesis formation.
- From data to function.
- Slopes and differences.
  - Don't use $y = mx + b$ except as a reminder of what a slope is. Instead ...
    * read the slope off a graph. Don't worry about the intercept.
    * read the slope off a regression report.
    * interpret the slope as the "effect size" of x on y."
  - Differences: if the input changes, how much does the output change?
- With the app: Can we predict something hard to measure from something easy.
  - systolic blood pressure from height?
  - income from BMI
- With the app: f(x) is not destiny. Predict BMI from education. The averages differ, but there is a big range around the line. Can't predict for an individual, could say something about averages in a group.
- With the app: How much variation is explained?

Looking forward

Understand the different settings in which regression is used in practice. A good topic for discussion in the workshop. Use examples from the different settings. - causation - classification - exploration: what might explain body mass index?

Defining big in terms of a the individual variables, e.g. a couple of standard deviations. This relates to the discussion of "interpreting slope."

A commonly used tricotomy for describing relationships between two variables is "negative" vs "zero"/"none" vs "positive". In the context of simple regression, these correspond to the sign of the slope b. This can be misleading, since a zero value of b can occur even when there is a strong (nonlinear) relationship between y and x.

The slope b is a physical quantity that has dimension and units. For instance if y is a person's height in cm, and x is a person's weight in kg, the units of b will be cm/kg. (The "dimension" of this is L/M – length over mass.) Many mathematical educators prefer to de-emphasize physical units, preferring to regard b as a pure number. This is a mistake from a statistical point of view. The size of physical quantities is important. Interpreting b as large or small needs to be understood in the context of the problem.

The correlation coefficient r is a scaled version of b. The scaling is by the ratio of the standard deviations of the x and y variables, that is, $r = \frac{\sigma_x}{\sigma_y} b$. This scaling results in r being a pure number since the units of σ_x/σ_y cancel out the units of b.

The slope b can be any numerical quantity. In contrast, the correlation coefficient must always be $-1 \leq r \leq 1$. Many mathematics educators believe that this means that r describes the "strength" of the relationship between y and x. Whether or not this is true depends on what one means by "strength." In scientific research, the intuition behind strength corresponds better to the slope b and includes the physical units of b. In statistics, when "strength" is taken to refer to how compelling the evidence is for a claim, an appropriate measure is the *confidence interval* on b. Another statistical quantity, the *p-value* on the slope, refers to a quantifying the evidence for a particular but very weak sort of claim, that b is anything but zero.

Although students are often drilled in the fact that $-1 \leq r \leq 1$, the reason why r is bounded in this way is subtle. It's misleading to conclude that the bounds on r suggest that a "strong" relationship is one where $|r| \approx 1$. The correlation coefficient r predates the distinction between descriptive and inferential statistics and mixes together aspects of both. This leads to pedagogical challenges that could be avoided if relationships are described using b and inferences made using the confidence interval on b.

- Too much is made of the "optimality" of the estimates of the slope and intercept. See the sum of squares Little App.
- Categorical explanatory variables can also be used. ANOVA is a general procedure in linear regression. Almost every statistical method covered in intro stats – proportions, differences in proportions, means, differences in means, ANOVA – can be presented quite naturally as a linear regression problem.
- Robust statistical methods are available to deal automatically with outliers, without having to handle them as special cases.
- r is meaningless in multiple regression. R^2 is more general.
- Although y and x are conventional names given to the variables involved when discussing statistical and mathematical theory, in statistical practice, both x and y are variables with *names*, and those names should be used explicitly. This is one reason why a regression table is the conventional format for describing a linear regression, not a formula.

Author info