Review of Bivariate Linear Regression

Contents

1 The Classic Bivariate Least Squares Model

1.1 The Setup

The Setup

Data Setup

- You have data on two variables, x and y, where at least y is continuous
- You want to characterize the relationship between x and y

The Setup

Theoretical Goals

- Describe the relationship between x and y
- Predict y from x
- Decide whether x causes y
- *The above goals are not mutually exclusive!*

1.2 An Example – Predicting Kids IQ

Predicting Kids IQ

Example 1 (Predicting Kids IQ). The goal is to predict cognitive test scores of three- and four-year-old children given characteristics of their mothers, using data from a survey of adult American women and their children (a subsample from the National Longitudinal Survey of Youth).

Predicting Kids IQ

Two Potential Predictors

One potential predictor of a child's test score (`kid.score`) is the mother's IQ score (`mom.iq`). Another potential predictor is whether or not the mother graduated from high school (`mom.hs`). In this case, both (`kid.score`) and (`mom.iq`) are continuous, while the second predictor variable (`mom.hs`) is *binary*.

Questions

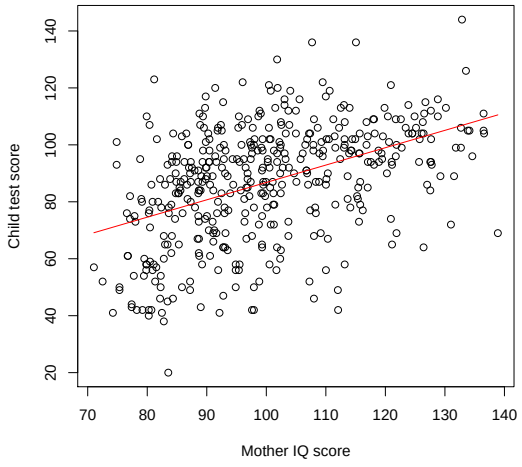
Would you expect these two potential predictors `mom.hs` and `mom.iq` to be correlated? Why?

Predicting Kids IQ

Least Squares Scatterplot

- The plot on the next slide is a standard two-dimensional scatterplot showing Kid's IQ vs. Mom's IQ
- We have superimposed the line of best *least squares* fit on the data
- Least squares linear regression finds the line that minimizes the sum of squared distances from the points to the line in the up-down direction

Scatterplot Kid's IQ vs. mom's IQ



Fitting the Linear Model with R

The Fixed-Regressor Linear Model

- When we fit a straight line to the data, we were fitting a very simple "linear model"
- The model is that $y = b_1 x + b_0 + \epsilon$, with the ϵ term having a normal distribution with mean 0 and variance σ_e^2
- b_1 is the slope of the line and b_0 is its y-intercept
- We can write the model in matrix "shorthand" in a variety of ways
- One way is to say that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- Another way or at the level of the individual observation, $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i$
- Note that in the above notations, y, $\boldsymbol{X}$ and ϵ have a finite number of rows, and the scores in $\boldsymbol{X}$ are considered as fixed constants, not random variables

Fitting the Linear Model with R The model

Using the `lm` function

- R has an `lm` function
- You define the linear model using a simple syntax
- In the model $y = b_1 x + b_0 + \epsilon$, y is a linear function of x To fit this model with `kid.score` as the y variable and `mom.iq` as the x variable, we simply enter the R command shown on the following slide

Fitting the Linear Model with R

```
> lm(kid.score~mom.iq)


Call:
lm(formula = kid.score ~ mom.iq)

Coefficients:
(Intercept)       mom.iq
      25.80         0.61
```

Comment

The intercept of 0.61 and slope of 25.8, taken literally, would seem to indicate that the child's IQ is definitely related to the mom's IQ, but that mom's with IQs around 100 have children with IQs averaging about 87.

Fitting the Linear Model with R

Saving a Fit Object

- R is an *object oriented language*
- You save the results of `lm` computation in *fit objects*
- Fit objects have well-defined ways of responding when you apply certain functions to them
- In the code that follows, we save the linear model fit in a fit object called `fit.1`
- Then, we apply the `summary` function to the object, and get a more detailed output summary

Fitting the Linear Model with R

```
> fit.1 ← lm(kid.score~mom.iq)
> summary(fit.1)


Call:
lm(formula = kid.score ~ mom.iq)

Residuals:
    Min      1Q  Median      3Q     Max
-56.753 -12.074   2.217  11.710  47.691

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.79978    5.91741    4.36 1.63e-05 ***
mom.iq       0.60997    0.05852   10.42  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.27 on 432 degrees of freedom
Multiple R-squared: 0.201,        Adjusted R-squared: 0.1991
F-statistic: 108.6 on 1 and 432 DF,  p-value: < 2.2e-16
```

Interpreting Regression Output

Key Quantities

- In the preceding output, we saw the *estimates*, their (estimated) *standard errors*, and their associated t-statistics, along with the Multiple R^2, adjusted R^2, and an overall test statistic

- Under the assumptions of the linear model (which are almost certainly only an approximation), the estimates divided by their standard errors have a Student-t distribution with $N - k$ degrees of freedom, where k is the number of parameters estimated in the linear model (in this case 2)

Iterpreting Regression Output

Key Quantities – Continued

- Since the parameter estimates have a distribution that is approximately normal, we can construct an approximate 95% confidence interval by taking the estimate ± 2 standard errors
- If we take the t distribution assumption seriously, we can calculate exact 2-sided probability values for the hypothesis test that a model coefficient is zero.
- For example, the coefficient b_1 has a value of 0.61, and a standard error of 0.0585
- The t-statistic has a value of $0.61/0.0585 = 10.4$
- The approximate confidence interval for b_1 is 0.61 ± 0.117

Iterpreting Regression Output

Key Quantities – Continued

- The multiple R^2 value is an estimate of the proportion of variance accounted for by the model
- When N is not sufficiently large or the number of predictors is large, multiple R^2 can be rather positively biased
- The "adjusted" or "shrunken" R^2 value attempts to compensate for this, and is an approximation to the known unbiased estimator
- *The adjusted R^2 does not fully correct the bias in R^2, and of course it does not correct at all for the extreme bias produced by post hoc selection of predictor(s) from a set of potential predictor variables*

Fitting the Linear Model with R

The `display` function

- The `summary` function produces output that is somewhat cluttered
- Often this is more than we need

- The `display` function (provided by Gelman and Hill in the `arm` library), pares things down to the essentials
- In general, if a coefficient is larger in absolute value than about two standard errors, it is significantly different from zero
- By taking the coefficient plus or minus two standard errors, you can get a quick (approximate) 95% confidence interval

Fitting the Linear Model with R

```
> fit.1 <- lm(kid.score~mom.iq)
> display(fit.1)

lm(formula = kid.score ~ mom.iq)
            coef.est coef.se
(Intercept) 25.80     5.92
mom.iq       0.61     0.06
---
n = 434, k = 2
residual sd = 18.27, R-Squared = 0.20
```

2 Evaluating and Extending the Model

2.1 Interpreting the Regression Line

Basic theoretical orientation

Basic theoretical orientation

When we obtain the best-fitting regression line and try to evaluate what it means, we first have to consider our basic theoretical orientation. There are three fundamental approaches:

- Descriptive
- Predictive
- Counterfactual

Interpreting the regression line — 3 approaches

Regression as description

One approach to regression is purely descriptive:

- We have a set of data
- We wish to describe the relationship between variables in a way that is mathematically succinct
- We concentrate on the data at hand, and resist generalizing to what might happen in new, as yet unmeasured, data sets

Interpreting the regression line — 3 approaches

Regression as prediction
Regression can be *predictive* in two senses.

One sense, used by Gelman and Hill, p. 34, is similar to the descriptive approach described previously. It considers how the criterion variable changes, on average, between two groups of scores that differ by 1 on a predictor variable while being identical on all other predictors. In the kids IQ example, we could say that, "all other things being equal, children with moms having IQs of 101 have IQs that are .61 points higher than children whose moms have IQs of 100" Another sense, employed frequently in marketing and data mining, obtains a regression equation in the hope of using it on new data to predict the criterion value *in advance* from values of the predictor that have already been obtained.

Interpreting the regression line — 3 approaches

Counterfactual interpretation

- The counterfactual or causal interpretation attempts to analyze how the criterion variable would change if the predictor variable were changed by one unit
- Suppose, for example, we found a linear relationship with a negative slope b_1 between size of classroom and standardized achievement scores
- We might then seek to conclude that decreasing class size by 1 would increase a child's achievement score by $-b_1$ units

Interpreting the regression line — quantitative aspects

Interpreting a regression fit
Key numerical aspects of a simple linear regression analysis include

- The slope
- The intercept
- How well the line fits the points, i.e., whether the variance of the errors is large or small, or, alternatively, whether the correlation coefficient is high in absolute value

Interpreting the Regression Line – Quantitative Aspects Regression Slope

Interpreting regression slope
Depending on whether the basic orientation is descriptive, predictive, or counterfactual, the slope might be interpreted as

- The difference in conditional mean on criterion variable y observed in groups of observations that differ by one unit on predictor variable x
- The difference in average value that will be observed in the future on y if you select an observation that is currently one unit higher on x
- The amount of change in y you will produce by increasing x by one unit

Interpreting the Regression Line – Quantitative Aspects Regression Intercept

Interpreting regression intercept

- Technically, the regression intercept is the average value of criterion variable y observed for those observational units with a value of 0 on predictor variable x
- Often this interpretation is nonsensical or at least very awkward

Example 2. Suppose you examine the relationship between height and weight for a group of individuals, and plot the linear regression line with height as the predictor variable x. The intercept represents the average weight of individuals with heights of zero!

2.2 Extending the Model

Can the model be improved?

- Maybe a simple linear regression doesn't predict the kids' IQ scores that well
- Perhaps we can do better
- There are numerous ways we might proceed

Extending and Improving the Model Adding Predictors

Selecting and adding predictors

- Perhaps mom's IQ, by itself, is simply inadequate for predicting a child's IQ
- In that case, we might consider additional variables in our data set
- *But we have to be careful!*

Extending and Improving the Model Adding Predictors

Dangers of overfitting

- If we have a long list of potential predictors, we *could* scan through the list and pick out variables that correlate highly with the criterion
- In fact, many standard regression programs (such as the module in SPSS) will do this for us automatically
- But this can be very dangerous
- Why?

Extending and Improving the Model Modeling Interaction

Interaction terms

- Once we have more than one predictor, we have an additional option
- We can add *interaction terms* to our model
- Variables *interact* if the effect of one varies depending on the value of the other(s)
- *Interaction effects can be very important in a number of contexts!*

Extending and Improving the Model Transforming the data

Linear and nonlinear transforms

- In some cases, simply transforming the variables linearly will make the meaning of the regression line clearer
- In other cases, a nonlinear transform may be necessary
- For example, when positive data have a huge range and a non-normal distribution, a log transformation may be very useful

Extending and Improving the Model Fitting a Nonlinear Model

Nonlinear Models

- An interaction model is nonlinear, but there are many other kinds of nonlinear models
- For example, we might fit the *polynomial* model $y = b_0 + b_1x + b_2x^2 + b_3x^3 + \epsilon$
- Or, we might fit a *piecewise regression* model, where different straight lines are fit to different ranges of predictor values
- Of course, this barely scratches the surface of what is available