Bivariate Regression

Joel S Steele

Bivariate example (Regression)

Within the regression framework, we are most interested in using a linear combination of parameters and variables to explain variance in our outcome of interest. The basic model takes the form of a line.

$$y = ax + b$$

or a more common expression in regression,

$$y_i = b_0 + b_1 x_i + \epsilon_i$$

Where b_0 and b_1 represent the intercept and slope respectively.

Parameter estimation

As you may remember from an earlier statistics course, we can use the *least squares* criteria to find the optimal estimates of both the intercept and slope. However, it may be instructive to see a small example of exactly such a function is *minimized*.

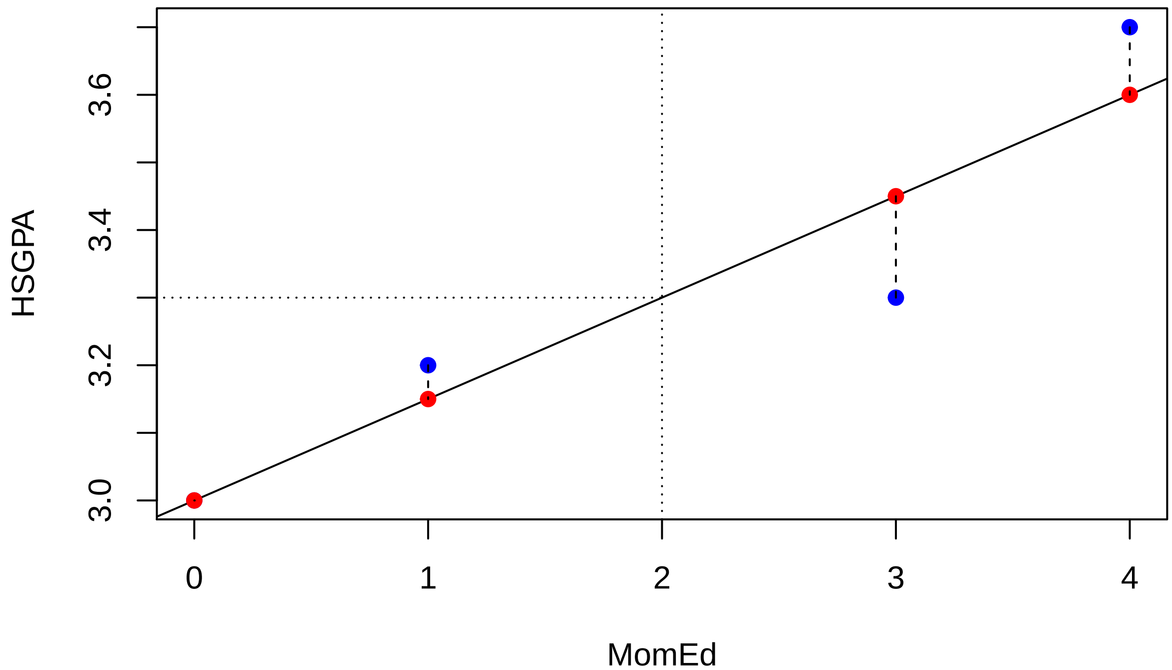
Hand computation with Calculus

Example:

Say that you are interested in whether or not a mother's level of education relates to her child's high school GPA.

The data:

- Mother's education: $X = [0, 1, 3, 4]$
- HS GPA: $Y = [3.0, 3.2, 3.3, 3.7]$
- point $1 = (0, 3.0)$
- point $2 = (1, 3.2)$
- point $3 = (3, 3.3)$
- point $4 = (4, 3.7)$



We know that the equation for a line is $y = ax + b$ Thus, we need to define the error term, we will use expected, $ax + b$, minus observed y.

The error equation: $\epsilon = ax + b - y$

To minimize the sum of squared error we take this function and square it

$$\sum_i \epsilon_i^2 = \sum_i (ax_i + b - y_i)^2$$

Using our data this sum of squared errors can now be expressed as:

$$\begin{aligned} SS_e &= [(0a + b - 3.0)^2 && \text{values from point 1} \\ &+(1a + b - 3.2)^2 && \text{values from point 2} \\ &+(3a + b - 3.3)^2 && \text{values from point 3} \\ &+(4a + b - 3.7)^2] && \text{values from point 4} \end{aligned}$$

Simplify and expand

$$\begin{aligned} SS_e &= [(b - 3.0)(b - 3.0)+ \\ &(a + b - 3.2)(a + b - 3.2)+ \\ &(3a + b - 3.3)(3a + b - 3.3)+ \\ &(4a + b - 3.7)(4a + b - 3.7)] \end{aligned}$$

Multiply through

$$\begin{aligned} SS_e &= [(b^2 - 6b + 9)+ \\ &(a^2 + ab - 3.2a + ab + b^2 - 3.2b - 3.2a - 3.2b + 10.24)+ \\ &(9a^2 + 3ab - 9.9a + 3ab + b^2 - 3.3b - 9.9a - 3.3b + 10.89)+ \\ &(16a^2 + 4ab - 14.8a + 4ab + b^2 - 3.7b - 14.8a - 3.7b + 13.69)] \end{aligned}$$

Collect similar terms within each subexpression

$$\begin{aligned} SS_e &= [(b^2 - 6b + 9)+ \\ &(a^2 + 2ab - 6.4a + b^2 - 6.4b + 10.24)+ \\ &(9a^2 + 6ab - 19.8a + b^2 - 6.6b + 10.89)+ \\ &(16a^2 + 8ab - 29.6a + b^2 - 7.4b + 13.69)] \end{aligned}$$

Combine all subexpressions and collect common terms

$$\begin{aligned} SS_e &= a^2 + 9a^2 + 16a^2 \\ &+b^2 + b^2 + b^2 + b^2 \\ &-6.4a - 19.8a - 29.6a \\ &-6b - 6.4b - 6.6b - 7.4b \\ &+2ab + 6ab + 8ab \\ &+9 + 10.24 + 10.89 + 13.69 \end{aligned}$$

Simplify common terms

$$\begin{aligned} SS_e &= 26a^2 \\ &+4b^2 \\ &-55.8a \\ &-26.4b \\ &+16ab \\ &+43.82 \end{aligned}$$

This is the equation for the sum of squared errors for our four observed points

$$SS_e = 26a^2 + 4b^2 - 55.8a - 26.4b + 16ab + 43.82$$

Take the partial derivative of this equation with respect to each parameter. For example, taking the partial derivative of the function SS_e with respect to a is presented below. It is important to note that since we are differentiating the equation based on the parameter a we only need to consider those terms that have and a in them. We will be using the power rule, which states $\frac{d}{dx} = (x^n) = n \cdot x^{n-1}$.

$$\begin{aligned} SS_e\big|_a &= 26a^2 & -55.8a & & +16ab \\ \frac{\partial SS_e}{\partial a} &= 26\left(2 \cdot a^1\right) & -55.8\left(1 \cdot a^0\right) & & +16b\left(1 \cdot a^0\right) \\ \frac{\partial SS_e}{\partial a} &= 26(2 \cdot a) & -55.8(1 \cdot 1) & & +16b(1 \cdot 1) \end{aligned}$$

$$\frac{\partial SS_e}{\partial a} = 52a - 55.8 + 16b$$

We rearrange it to look like our equation for a line and set this equal to zero, this gives us the minimum point for the equation, or where the change stops.

$$\begin{aligned} \frac{\partial SSe}{\partial a} &= 52a + 16b - 55.8 \\ 0 &= 52a + 16b - 55.8 \end{aligned}$$

Repeat for the parameter b

$$\begin{aligned} SS_e\big|_b &= 4b^2 - 26.4b + 16ab \\ \frac{\partial SSe}{\partial b} &= 8b - 26.4 + 16a \\ \frac{\partial SSe}{\partial b} &= 16a + 8b - 26.4 \\ 0 &= 16a + 8b - 26.4 \end{aligned}$$

Equation 1 (how the function changes with respect to a)

$$0 = 52a + 16b - 55.8$$

Equation 2 (how the function changes with respect to b)

$$0 = 16a + 8b - 26.4$$

Solve for a in Equation 1

$$\frac{(55.8 - 16b)}{52} = a$$

Plug a into Equation 2 and solve for b

$$\begin{aligned} 0 &= 16 \cdot \left(\frac{(55.8-16b)}{52}\right) + 8b - 26.4 \\ 0 &= 16 \cdot \frac{55.8}{52} - 16 \cdot \frac{16b}{52} + 8b - 26.4 \\ &\text{move all of the } b \text{ terms to one side of the equation} \\ 26.4 - 16 \cdot \frac{55.8}{52} &= -16 \cdot \frac{16b}{52} + \frac{416b}{52} \\ 9.23077 &= \frac{160b}{52} \\ 9.23077 \cdot 52 &= 160b \\ 480 &= 160b \\ \frac{480}{160} &= b \\ &\text{the intercept estimate} \\ 3 &= b \end{aligned}$$

Plug b into our equation for a from above

$$\begin{aligned} \frac{(55.8-16 \cdot 3)}{52} &= a \\ \frac{(55.8-48)}{52} &= a \\ \frac{7.8}{52} &= a \\ 0.15 &= a \end{aligned}$$

So the best fitting line is

$$y = 0.15x + 3$$

Let's confirm our findings.

```r
lm(HSGPA ~ MomEd)
```

```

Call:
lm(formula = HSGPA ~ MomEd)

Coefficients:
(Intercept)        MomEd
       3.00         0.15
```