Simple Linear Regression

Lecture 6 Concept Question Solutions

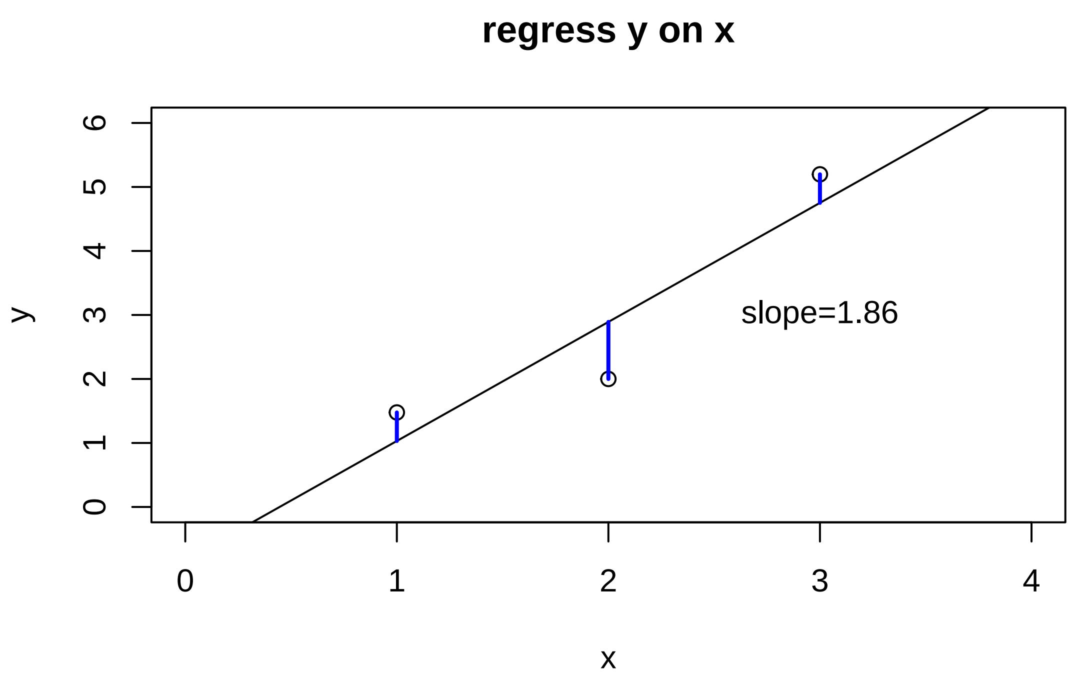
Statistics 139

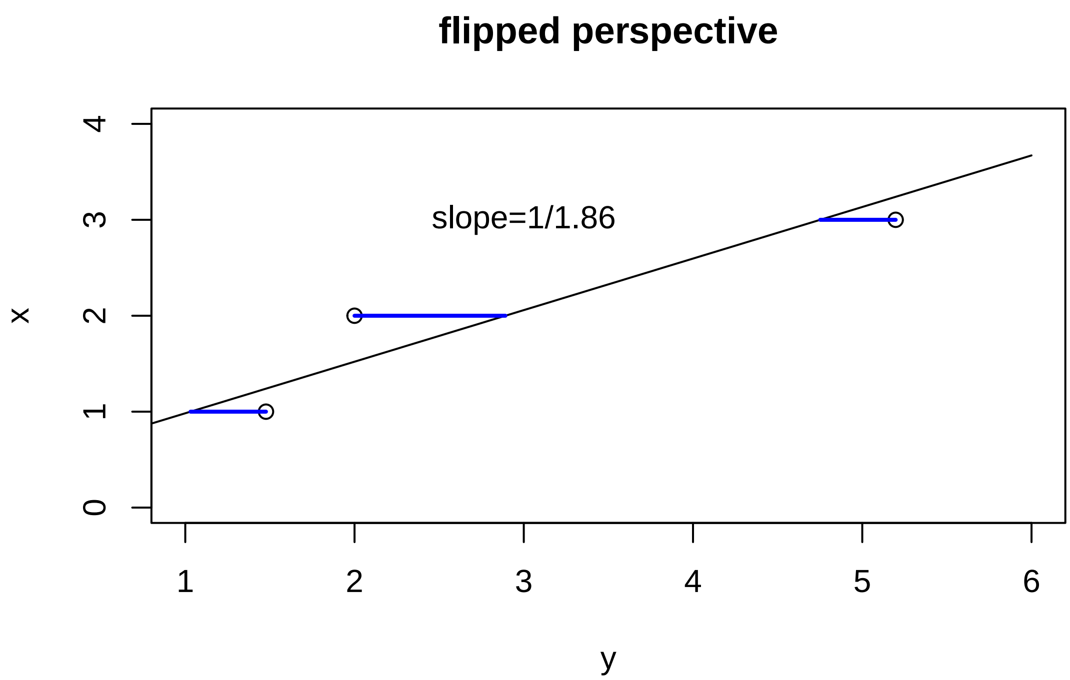
Concept Checks

a) If the predictor and the response were switched in a simple regression model, would the new estimated slope just be the reciprocal of the original? Why or why not?

Algebraically: Based on the formula $\hat{\beta}_1 = r_{xy}\frac{s_x}{s_y}$, if the X and Y variables are flipped, correlation is unchanged, but the ratio of standard deviations does flip. Thus the slope is **not** simply the reciprocal (unless the estimated correlation is 1 or -1. . . a perfectly fit line).

Geometrically: by flipping the response and predictor variable, the minimization perspective changes: instead of minimizing vertical distances, it's like we are minimizing horizontal distances. See the following figures.





b) When can R^2 be negative? Can an OLS (ordinary least squares) regression model have an R^2 less than zero? Why or why not?

R^2 can be negative for a model in general if it performs worse than the horizontal line at $\bar{y}$. An example: if the scatterplot displays a positive association but our model has a negative slope. This will never happen under OLS: the worst case setting is a slope of zero and an intercept at $\bar{y}$.

c) The OLS estimate of variance is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2}{n-2} = \frac{SSE}{df_E}$$

.

Why the $n-2$ (where does this come from)? What is the sampling distribution of $\hat{\sigma}^2$?

$\hat{\sigma}^2 \sim \left(\frac{\sigma^2}{n-2}\right)\chi^2_{n-2}$. $n-2$ can be justified from many perspectives. This leads to the estimate being unbiased for the true σ. This is the result because of the *degree of freedom* that the sum of squares in the numerator has: in order to calculate this sum of squares error (around the line), the slope and intercept has to first be calculated. This results in the vector of Y being *anchored* at these two estimates, *eating up* those 2 degree of freedom. More conceptually: a line fit to $n = 2$ is guaranteed to go through the 2 points exactly (unless they have the same value of x), and thus it is impossible to estimate variability around the line. This formula agrees with that intuition: $\hat{\sigma}^2$ is undefined unless $n \geq 3$.