

## Lesson 1: Least Squares Regression Line -- Teacher's Notes

Questions to ask students as they complete the lab:

1. When you complete the first scatterplot using Stay and Risk, what does this plot actually represent? (It is a comparison of the length of hospital stay and the risk of acquiring an infection while in the hospital.)
2. Does it matter which variable is the explanatory variable and which is the response variable? (Yes, because the explanatory variable is the "input" variable and we expect it to be a predictor variable for the response variable.)
3. Here is the output from the first linear regression of Stay vs Risk:

*Coefficients:*

*(Intercept) stay*

0.7437 0.3742

Ask students to complete the blanks with their interpretation of the coefficients. Point out to them that Intercept means the y-intercept of the regression line and *stay* represents the slope of the regression line.

4. After the students type the command **model1\$fitted.values**, they will get a data dump of the fitted values of the regression line. These are the  $\hat{y}$  values. The next few steps walk the students through the process of finding the residuals. The residual plot should show students that while there's a regression model, it's not the best regression model. There are lots of residuals that are greater than 2 units away from the regression line.
5. The Pearson's correlation coefficient,  $r$ , is 0.5329. This is a good point to ask students about the difference between the correlation coefficient and the coefficient of determination. The  $r^2 = 0.284$  which is pretty low. The coefficient of determination measures the percent of variation in the response variable, risk, that can be explained by the least squares linear model based on the length of hospital stay.