

In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use `set.seed(1)` prior to starting part (a) to ensure consistent results.

- (a) Using the `rnorm()` function, create a vector, `x`, containing 100 observations drawn from a  $N(0, 1)$  distribution. This represents a feature,  $X$ .
- (b) Using the `rnorm()` function, create a vector, `eps`, containing 100 observations drawn from a  $N(0, 0.25)$  distribution i.e. a normal distribution with mean zero and variance 0.25.
- (c) Using `x` and `eps`, generate a vector `y` according to the model

$$Y = -1 + 0.5X + \epsilon.$$

What is the length of the vector `y`? What are the values of  $\beta_0$  and  $\beta_1$  in this linear model?

- (d) Create a scatterplot displaying the relationship between `x` and `y`. Comment on what you observe.
- (e) Fit a least squares linear model to predict `y` using `x`. Comment on the model obtained. How do  $\hat{\beta}_0$  and  $\hat{\beta}_1$  compare to  $\beta_0$  and  $\beta_1$ ?
- (f) Display the least squares line on the scatterplot obtained in (d). Draw the population regression line on the plot, in a different color. Use the `legend()` command to create an appropriate legend.
- (g) Now fit a polynomial regression model that predicts `y` using `x` and `x2`. Is there evidence that the quadratic term improves the model fit? Explain your answer.
- (h) Repeat (a)–(f) after modifying the data generation process in such a way that there is *less* noise in the data. The model (3.39) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term  $\epsilon$  in (b). Describe your results.

- (i) Repeat (a)–(f) after modifying the data generation process in such a way that there is *more* noise in the data. The model (3.39) should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term  $\epsilon$  in (b). Describe your results.
- (j) What are the confidence intervals for  $\beta_0$  and  $\beta_1$  based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.