Exercise

To help understand and explore new concepts, you can simulate fake datasets in R. The advantage of this is that you "play God" because you actually know the underlying truth, and you get to see how good your model is at recovering the truth.

Once you've better understood what's going on with your fake dataset, you can then transfer your understanding to a real one. We'll show you how to simulate a fake dataset here, then we'll give you some ideas for how to explore it further:

```
# Simulating fake data
x_1 <- rnorm(1000,5,7) # from a normal distribution simulate
                       # 1000 values with a mean of 5 and
                       #  standard deviation of 7
hist(x_1, col="grey") # plot p(x)
true_error <- rnorm(1000,0,2)
true_beta_0 <- 1.1
true_beta_1 <- -8.2

y <- true_beta_0 + true_beta_1*x_1 + true_error
hist(y) # plot p(y)
plot(x_1,y, pch=20,col="red") # plot p(x,y)
```

1. Build a regression model and see that it recovers the true values of the βs.

2. Simulate another fake variable x_2 that has a Gamma distribution with parameters you pick. Now make the truth be that y is a linear combination of both x_1 and x_2. Fit a model that only depends on x_1. Fit a model that only depends on x_2. Fit a model that uses both. Vary the sample size and make a plot of mean square error of the training set and of the test set versus sample size.

3. Create a new variable, z, that is equal to x_1^2. Include this as one of the predictors in your model. See what happens when you fit a model that depends on x_1 only and then also on z. Vary the sample size and make a plot of mean square error of the training set and of the test set versus sample size.

4. Play around more by (a) changing parameter values (the true β s), (b) changing the distribution of the true error, and (c) including more predictors in the model with other kinds of probability distributions. (`rnorm()` means randomly generate values from a normal distribution. `rbinom()` does the same for binomial. So look up these functions online and try to find more.)

5. Create scatterplots of all pairs of variables and histograms of single variables.