Name:_____ Date:_____

LAB 4C: Cross-Validation
Response Sheet

Directions: Record your responses to the lab questions in the spaces provided.

What is cross-validation?

Step 1: train-test split

(1) First, fill in the blanks below to randomly select which rows of `arm_span` will go into the `training` set.

```
set.seed(123)

training_rows <- sample(1:___________, size = 68)
```

(2) Second, use the `slice` function to create two dataframes: one called `training` consisting of the `training_rows`, and another called `test` consisting of the remaining rows of `arm_span`.

```
training <- slice(arm_span, ___________)

test <- slice(___________, - ___________)
```

(3) Explain these lines of code and describe the `training` and `test` datasets.

Aside: set.seed

Aside: training-test ratio

Step 2: training the model

(4) Write and run code fitting a line of best fit model to our `training` data and assign it the name `best_training`.

Step 3: test the model

(5) Fill in the blanks below to add predicted heights to our `test` data:

```
test <- mutate(test, ___________ = predict(best_training, newdata = ___________))
```

Name:_____ Date:_____

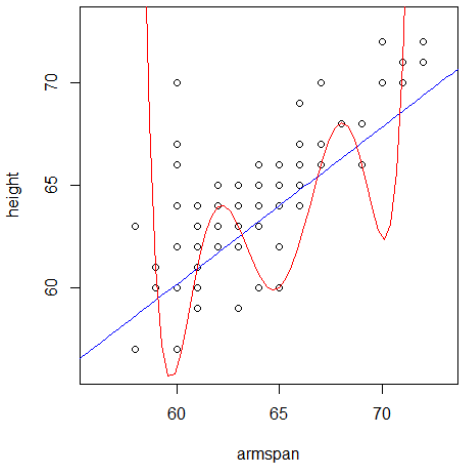
LAB 4C: Cross-Validation
Response Sheet

(6) Calculate the *test MSE* in the same way as you did in the previous lab (test MSE is simply MSE of the predictions on the test data).

Recap

Why cross-validate?

Example of overfitting



(7) Which model does a better job of predicting the 7 `training` points?

(8) Which model do you think will do a better job of predicting the rest of the data?

Example of overfitting, continued

(9) Which model does a better job of generalizing to the rest of the `arm_span` dataset?