MATH 392 Problem Set 7

Exercises from the book

9.1: 4

9.2: 2, 3, 4, 5

Additional Exercises

1. Consider the following dataset:

```
set.seed(32)
n <- 10
x <- rnorm(n)
y <- -1 + 1.3 * x + rnorm(n, .3)
df <- data.frame(x, y)
```

We can find the least squares regression line by running `lm()` (which uses the normal equations), and extract the coefficient estimates.

```
m1 <- lm(y ~ x, data = df)
coef(m1)
```

```
## (Intercept)           x
##  -0.4488683   1.2495758
```

A more general approach to finding the estimates that optimize a loss function is to use a numerical optimization technique. Here we use `optim()` to minimize the RSS. By default this function uses the Nelder-Mead algorithm), but you can also toggle to another algorithm such as BFGS or select an entirely different optimization function/package.

```
RSS <- function(par, x, y) {
  beta_0 <- par[1]
  beta_1 <- par[2]
  sum((y - (beta_0 + beta_1 * x))^2)
}
opt <- optim(par = c(0, 0), fn = RSS, x = x, y = y)
```

The `par` argument is the set of values of the two parameters that you want to initialize the algorithm at. You can try several different values and see if the final estimates agree. The final estimates are found in the `opt` object.

```
opt$par
```

```
## [1] -0.4488259  1.2495164
```

Which agree very closely with the analytical solutions from the normal equations.

   a. Using numerical optimization, find the estimates that minimize two additional loss functions: a) the absolute deviation in the `y` and b) the squared deviation in the x.

   b. Plot all three lines on top of a scatterplot of the data. Add an `annotate()` layer or legend to make it clear which line is which.

   c. Create a second scatterplot that again shows the least squares regression line. Add to this plot pairs of lines that represent each of the following intervals:

- A confidence interval on β_1.
- A confidence interval on $E(Y|X = x)$.
- A prediction interval on $[Y|X = x]$.

2. Ecological Fallacy refers to a situation where one draws inferences on the individual level from data that was collected at the group level.

```
# install.packages("resampledata")
library(resampledata)
data(corrExerciseB)
```

a. Create a scatter plot of all of the data, which each group plotted in a different color. Add in the group means for each.

b. Compute two sample correlations: one for the group means, the other for all of the data. Under which conditions, stated informally, will the correlation at the group level exceed that at the individual level? Do you expect that this is the a more common or less common feature of aggregated data in the real world?

c. In a setting such as this, what is the consequence for your data analysis of committing an ecological fallacy?