Exercise Set 3-1

1) a) Here's one way you could use equations 3.6 and 3.7 in R.

```
> x <- anscombe$x1
> y <- anscombe$y1
> n <- length(x) #number of pairs of observations
> #calculate the slope and intercept.
> b <- (sum(x*y) - (1/n) * sum(x) * sum(y))/(sum(x^2) - (1/n) *
sum(x)^2)
> a <- (sum(y) - b*sum(x))/n
> a
[1] 3.000091
> b
[1] 0.5000909
```

I assigned values values to `x`, `y`, and `n` to make the code a little easier to read, but you could calculate `a` and `b` without doing that, just replacing `x` with `anscombe$x1`, replacing `y` with `anscombe$y1`, and replacing `n` with `length(anscombe$x1)`.

To use equations 3.8 and 3.9, you could write

```
> x <- anscombe$x1
> y <- anscombe$y1
> b <- sum((x - mean(x))*(y - mean(y))) / sum((x - mean(x))^2)
> a <- mean(y) - b*mean(x)
> a
[1] 3.000091
> b
[1] 0.5000909
```

2) a) Start by plugging the expression for $\tilde{a}$ into the expression for $\tilde{b}$.

$$\tilde{b} = \frac{\sum_{i=1}^{n} x_i y_i - \tilde{a}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i y_i - \left(\frac{\sum_{i=1}^{n} y_i - \tilde{b}\sum_{i=1}^{n} x_i}{n}\right)\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}$$

We can distribute the $\sum_{i=1}^{n} x_i$ on the right and then split up the expression as follows:

$$\tilde{b} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i + \tilde{b}(\sum_{i=1}^{n} x_i)^2}{n}}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2} + \frac{\tilde{b}(\sum_{i=1}^{n} x_i)^2}{n\sum_{i=1}^{n} x_i^2}$$

Subtract $\frac{\tilde{b}(\sum_{i=1}^{n} x_i)^2}{n\sum_{i=1}^{n} x_i^2}$ from both sides.

$$\tilde{b} - \frac{\tilde{b}(\sum_{i=1}^{n} x_i)^2}{n\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2}$$

Factor out the $\tilde{b}$ on the left.

$$\tilde{b}\left(1 - \frac{1}{n}\frac{(\sum_{i=1}^{n} x_i)^2}{\sum_{i=1}^{n} x_i^2}\right) = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2}$$

Divide both sides by $\left(1 - \frac{1}{n}\frac{(\sum_{i=1}^{n} x_i)^2}{\sum_{i=1}^{n} x_i^2}\right)$ to isolate the $\tilde{b}$

$$\tilde{b} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2\left(1 - \frac{1}{n}\frac{(\sum_{i=1}^{n} x_i)^2}{\sum_{i=1}^{n} x_i^2}\right)},$$

and simplify the denominator by distributing the $\sum_{i=1}^{n} x_i^2$, giving equation 3.7:

$$\tilde{b} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}.$$

b) Equation 3.6 can be split into

$$\tilde{a} = \frac{1}{n}\sum_{i=1}^{n} y_i - \tilde{b}\frac{1}{n}\sum_{i=1}^{n} x_i,$$

Which, by the definitions of $\bar{x}$ and $\bar{y}$, is equivalent to equation 3.8.

c) Following the hint in the problem, we work backwards, starting with the denominator of equation 3.9, which is $\sum_{i=1}^{n}[(x_i - \bar{x})^2]$. Expanding the square term gives

$$\sum_{i=1}^{n}[(x_i - \bar{x})^2] = \sum_{i=1}^{n}[x_i^2 - 2x_i\bar{x} + \bar{x}^2].$$

Breaking this into three sums and pulling out the terms that do not depend on i, we have

$$\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \bar{x}^2\sum_{i=1}^{n} 1.$$

$\sum_{i=1}^{n} 1 = n$, and by definition, $\sum_{i=1}^{n} x_i = n\bar{x}$, so we have

$$\sum_{i=1}^{n}(x_i^2) - 2\bar{x}n\bar{x} + n\bar{x}^2 = \sum_{i=1}^{n}(x_i^2) - n\bar{x}^2,$$

Where the last step comes from noticing that $\bar{x}n\bar{x} = n\bar{x}^2$. Finally, to get the denominator of equation 3.9, replace $\bar{x}$ with its definition, $(1/n)\sum_{i=1}^{n} x_i$, giving

$$\sum_{i=1}^{n}[(x_i - \bar{x})^2] = \sum_{i=1}^{n} x_i^2 - n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2.$$

Thus, the denominators of equations 3.7 and 3.9 are equivalent. Next, we need to show that the numerators of equations 3.7 and 3.9 are equivalent. The procedure is similar. We start with the numerator of equation 3.9,

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Expanding the product inside the sum gives

$$\sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}).$$

Breaking this statement into four separate sums and pulling out the terms that do not depend on i gives

$$\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\bar{y}$$

Recalling the definitions of $\bar{x}$ and $\bar{y}$ gives

$$\sum_{i=1}^{n}(x_i y_i) - n\bar{y}\bar{x} - n\bar{x}\bar{y} + n\bar{x}\bar{y} = \sum_{i=1}^{n}(x_i y_i) - n\bar{y}\bar{x}.$$

Finally, to get the numerator of equation 3.7, replace $\bar{x}$ and $\bar{y}$ with their definitions,

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i y_i) - n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \sum_{i=1}^{n}(x_i y_i) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right).$$

Having shown that the numerators and denominators of the expressions in equations 3.7 and 3.9 are equal, we have shown that the two expressions are equal.

Exercise Set 3-2

1) a) Suppose that we already know $\tilde{b}$. Then the sum of squared errors as a function of a is

$$g(a) = \sum_{i=1}^{n}(y_i^2 - 2ay_i - 2\tilde{b}x_iy_i + a^2 + 2a\tilde{b}x_i + \tilde{b}^2x_i^2).$$

The derivative with respect to a is

$$g'(a) = \sum_{i=1}^{n}(-2y_i + 2a + 2\tilde{b}x_i).$$

We find $\tilde{a}$ by finding any values of a for which $g'(a) = 0$. Through a set of steps parallel to those used for $\tilde{b}$ in the main text, we find the sole value of a for which $g'(a) = 0$:

$$\tilde{a} = \frac{\sum_{i=1}^{n} y_i - \tilde{b}\ \sum_{i=1}^{n} x_i}{n}.$$

b) If $x = \bar{x}$, then $\tilde{y} = \tilde{a} + \tilde{b}\bar{x}$. Replacing $\tilde{a}$ with $\bar{y} - \tilde{b}\bar{x}$ gives $\tilde{y} = (\bar{y} - \tilde{b}\bar{x}) + \tilde{b}\bar{x} = \bar{y}$, so when $x = \bar{x}$, the y-coordinate of the line is $\bar{y}$.

2) The slope is

$$\acute{b} = \frac{\sum_{i=1}^{n} x_iy_i}{\sum_{i=1}^{n} x_i^2}.$$

One way to get the slope is to minimize the squared line errors for a line of the form $y_i = \acute{b}x_i$. A second way is to set $\tilde{a} = 0$ in equation 3.5, which works because we know that equation 3.5 gives the "best" (in the sense of minimizing the squared line errors) value of b given any provided value of a, including 0.

3) One could re-do the minimization of squared line errors, or one could simply switch x and y in the expressions for $\tilde{a}$ and $\tilde{b}$ to get

$$\tilde{c} = \frac{\sum_{i=1}^{n} x_i - \tilde{d}\sum_{i=1}^{n} y_i}{n}$$

and

$$\tilde{d} = \frac{\sum_{i=1}^{n} x_iy_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2}$$

In general, $\tilde{c}$ and $\tilde{d}$ are not equal to $\tilde{a}$ and $\tilde{b}$. (Try `lm(anscombe$y1 ~ anscombe$x1)` and `lm(anscombe$x1 ~ anscombe$y1)`.) But they are equal if the x and y values are rescaled so that their sums are equal and the sums of their squares are equal. Sometimes people *standardize* variables—subtracting each variable's mean and dividing by the standard

deviation—after which the least-squares slope is the same regardless of which variable is on which side of the linear equation.

4) a) Use

```
install.packages("quantreg")
library(quantreg)
```

b) Use

```
mod.fit.L1  <- rq(anscombe$y1 ~ anscombe$x1)
summary(mod.fit.L1)
```

The intercept is 3.24, and the slope is 0.48.

c) The least-squares and L1 lines are similar in this case.

e) The L1 line goes through almost all the points and misses the one that falls off the line. The least-squares line, in contrast, is pulled toward the outlying point. One way to understand this is to note that the sum of the squared line errors is sensitive to large line errors. As such, the least-squares line will be pulled toward individual outlying points in order to avoid making large errors. In contrast, the L1 line weights small line errors more heavily and large line errors less heavily than the least-squares line.