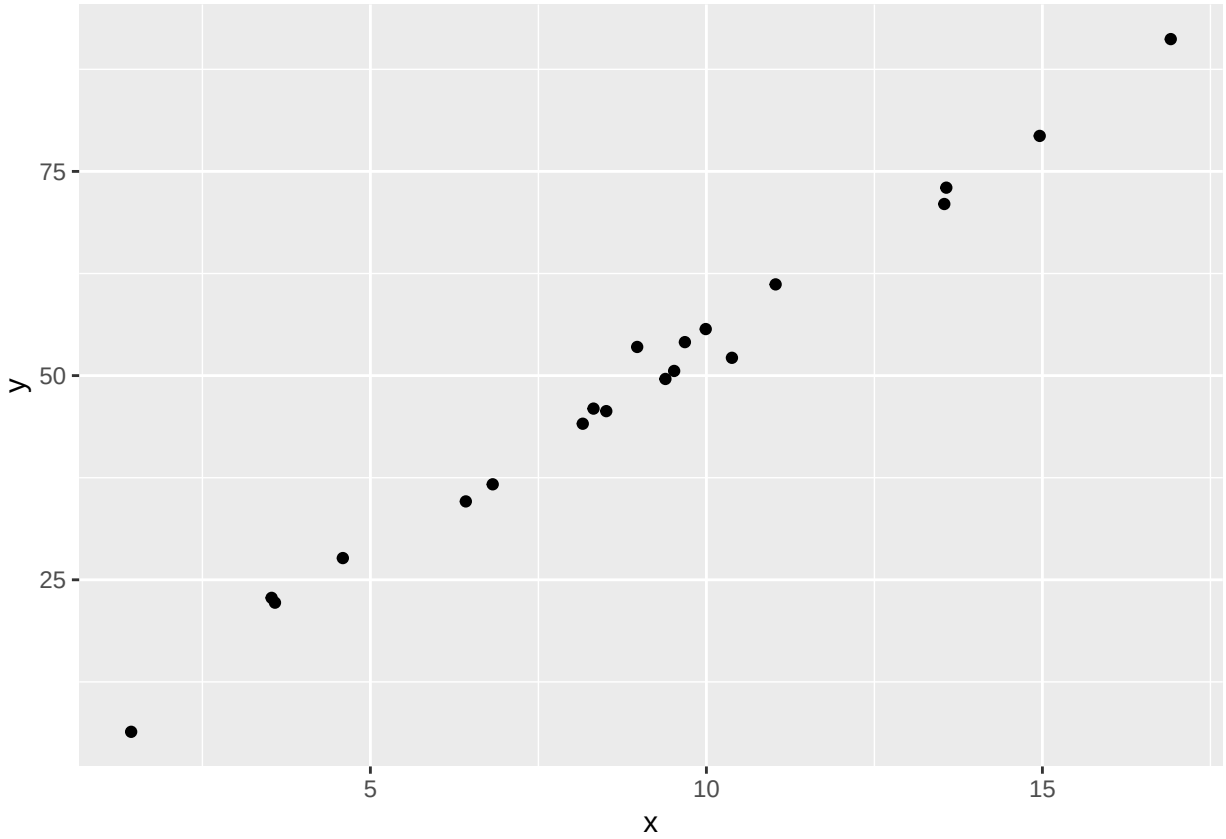
Homework 5

(For questions **1, 2**) Consider the following paired data sets of length 20:

```
x <- c(6.82, 1.44, 9.39,  8.51, 10.38,  4.59, 14.96,  9.68, 13.54,  6.42, 11.03,
       3.53, 16.91,  9.52,  8.16,  8.97,  8.32,  3.58, 13.57,  9.99)
y <- c(36.69, 6.39, 49.59, 45.65, 52.18, 27.66, 79.35, 54.10, 71.01, 34.60, 61.17,
       22.79, 91.20, 50.57, 44.11, 53.51, 45.96, 22.20, 73.01, 55.70)
```

1. (a) Create a scatter plot to visualize the data (*Hint*: you may want to start with making a data frame and then use `geom_point()`). Do you think there is a strong linear association between `x` and `y`?

```
df <- data.frame(x, y)
g <- ggplot(df, aes(x, y)) + geom_point()
print(g)
```



(b) Compute the sample correlation coefficient between `x` and `y`. Is your result consistent with your answer in (a)?

```
cor(x, y)
```

```
## [1] 0.9952905
```

2. (a) Assume that `y` is an outcome in a certain experiment, and `x` is a predictor. Find the best fitting line describing the association between `x` and `y` by specifying its y-intercept (β_0) and slope (β_1).

```r
fit <- lm(y ~ x)
fit$coefficients
```

```
## (Intercept)           x
##    2.222796    5.203190
```

(b) Suppose that a new `x` value came in, say 13. Estimate the corresponding `y` value using the best fitting line you obtained in part (a) above.

```r
fit$coefficients[1] + fit$coefficients[2]*13
```

```
## (Intercept)
##    69.86427
```

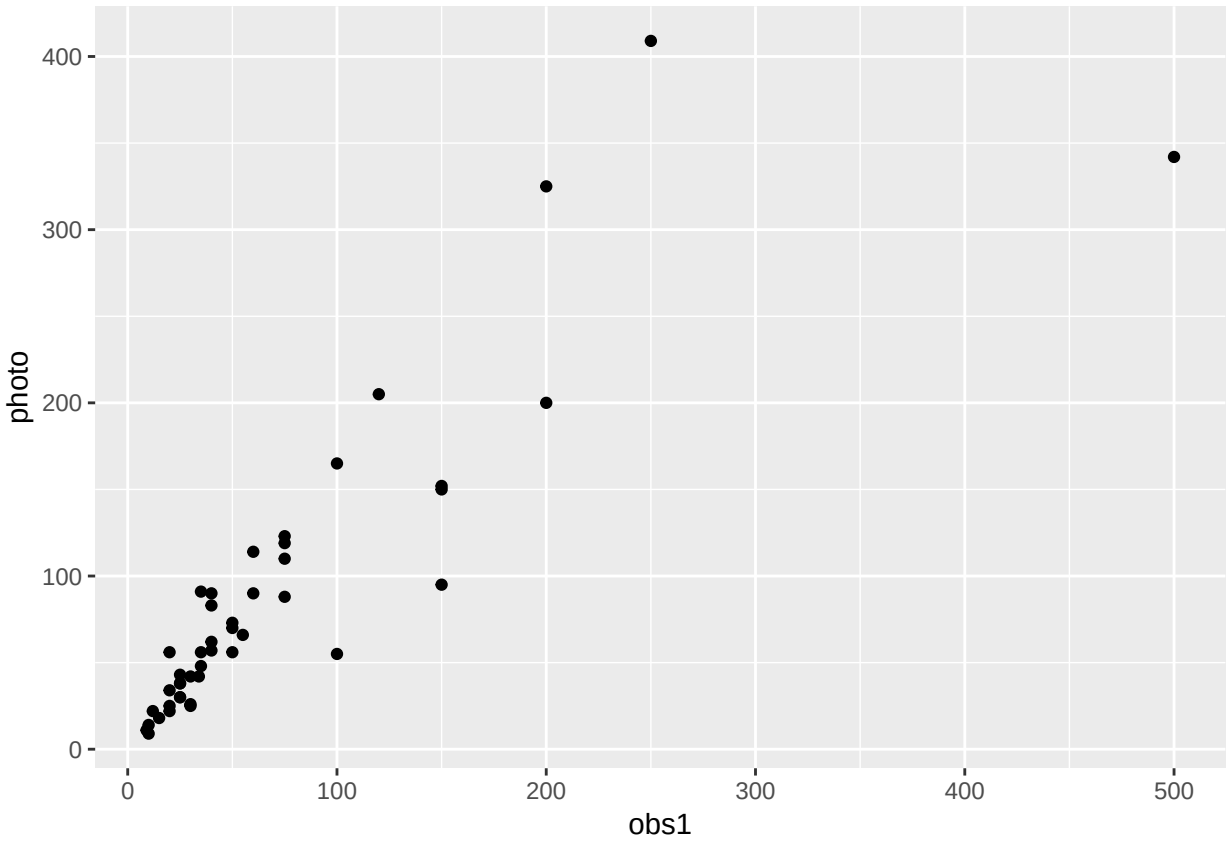
(For questions **3, 4, 5**) Aerial surveys sometimes rely on visual methods to estimate the number of animals in an area. For example, to study snow geese in their summer range areas west of Hudson Bay in Canada, a small aircraft was used to fly over the range, and when a flock of geese was spotted, an experienced person estimated the number of geese in the flock. To investigate the reliability of this method of counting, an experiment was conducted in which an airplane carrying two observers flew over $n = 45$ flocks, and each observer made an independent estimate of the number of birds in each flock. Also, a photograph of the flock was taken so that a more or less exact count of the number of birds in the flock could be made. The resulting data are given in the attached data file `snowgeese.tsv`, which can be downloaded from **HERE**. The three variables in the data sets are `photo` = photo count; `obs1` = aerial count by observer 1; and `obs2` = aerial count by observer 2.

3. Read in the data set and draw the scatter plots of `photo` (response) vs `obs1` (explanatory) and `photo` (response) vs `obs2` (explanatory).

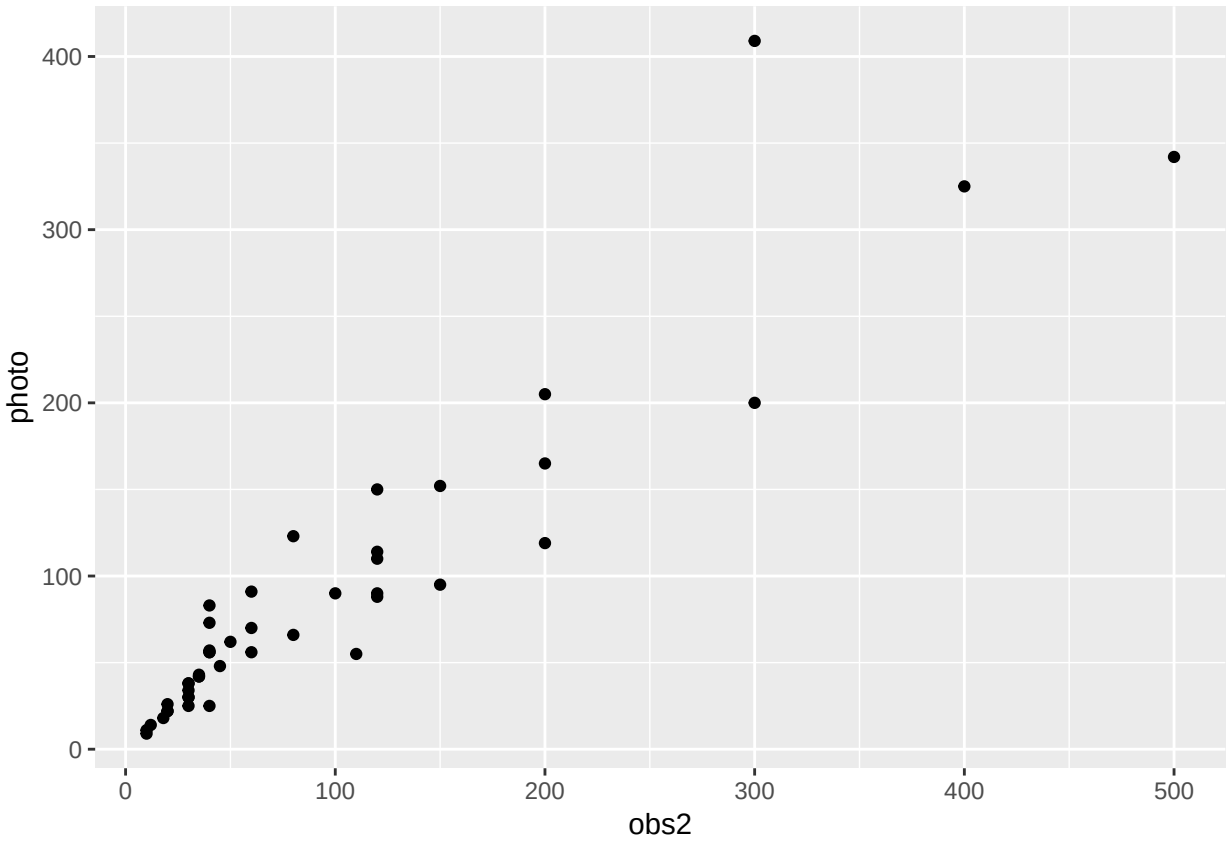
```r
df <- read_tsv("snowgeese.tsv")
```

```
## Parsed with column specification:
## cols(
##   photo = col_double(),
##   obs1 = col_double(),
##   obs2 = col_double()
## )
```

```r
ggplot(df, aes(obs1, photo)) + geom_point()
```

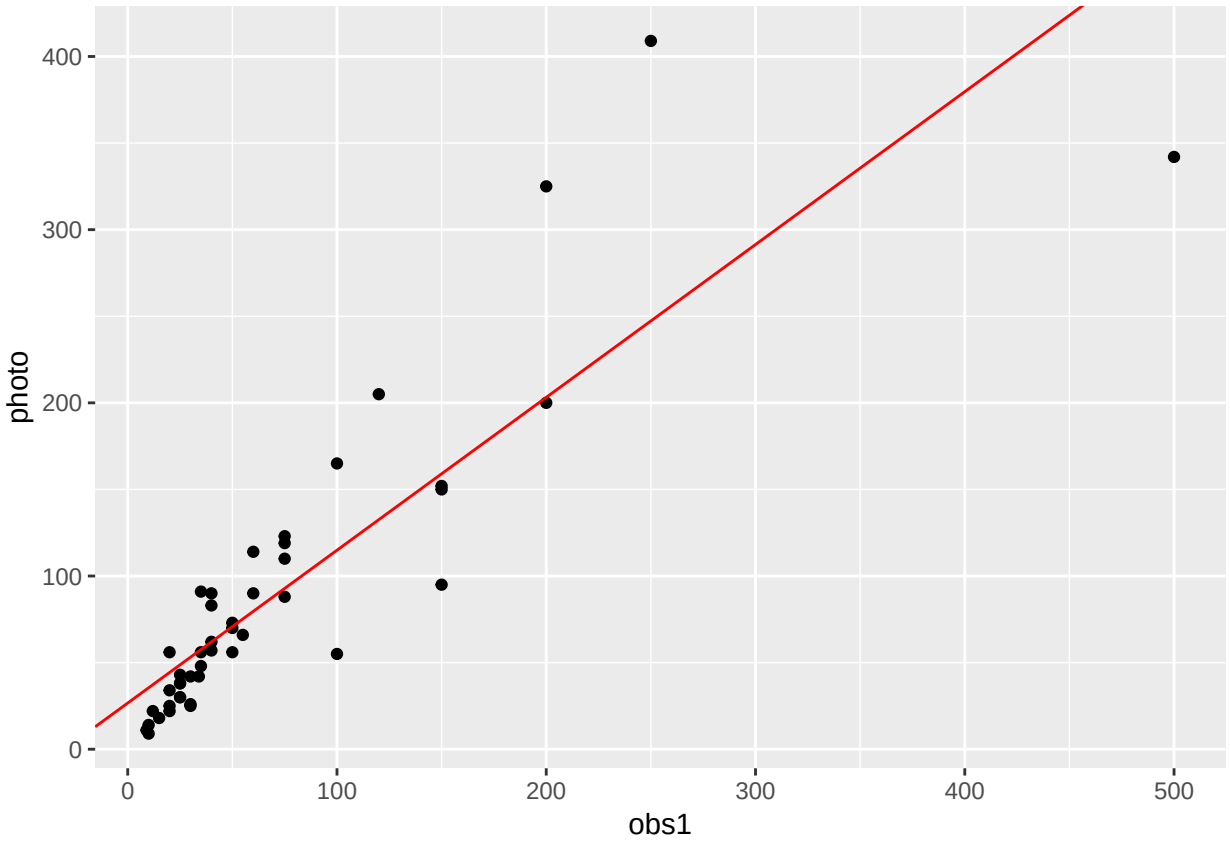



```
ggplot(df, aes(obs2, photo)) + geom_point()
```



4. Compute the regression of `photo` on `obs1` using `lm()` (That is, find the best fitting line). Overlay the best fitting line to the scatter plot.

```
fit1 <- lm(photo ~ obs1, data = df)
ggplot(df, aes(obs1, photo)) +
  geom_point() +
  geom_abline(intercept = fit1$coefficients[1], slope = fit1$coefficients[2], color = "red")
```



5. Suppose that observer 2 made a count of 55 in an additional aerial survey performed later. What is your best guess for the actual count?

```
fit2 <- lm(photo ~ obs2, data = df)
fit2$coefficients[1] + fit2$coefficients[2]*55
```

```
## (Intercept)
##    58.46297
```