EDA A5 Q3

Data

Set up the following:

```
## Set this up for your own directory
#imageDirectory <- "MyAssignmentDirectory/img"  # e.g. in current "./img"
#dataDirectory <- "MyAssignmentDirectory/data"  # e.g. in current "./data"
#path_concat <- function(path1, ..., sep="/") paste(path1, ..., sep = sep)
```

The full data set is then read in as:

```
labData <- read.csv("/Users/rudranibhadra/Downloads/labData.csv")
```

The data can be subsetted according to the three different experimental plans.

a. *(1 mark)* Select that subset of the data corresponding to the observational plan. Assign it to the variable `observational`. Show your code.

```
observational<-labData[labData$type=='observational',]
```

Analysis

b. *(4 marks)* Plot the (x, y) pairs from all of the observational data

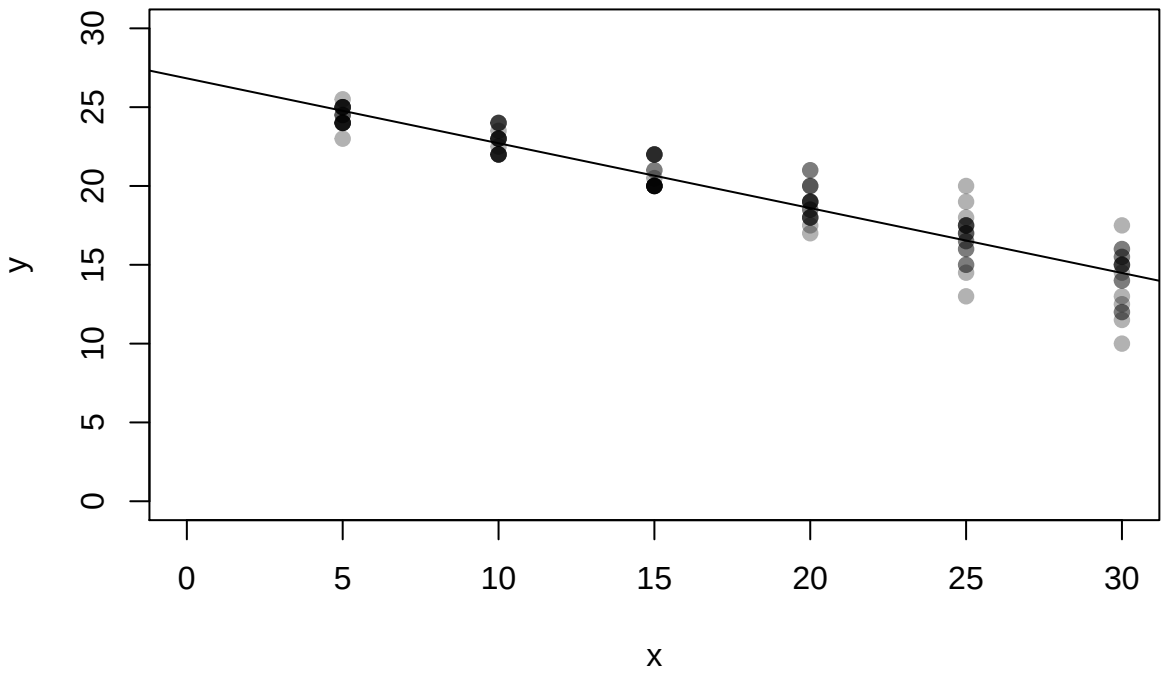
Use `xlim = c(0, 30), ylim = c(0,30), pch = 19, col = adjustcolor("black", 0.3)` in the call to `plot()`.

Label the plot meaningfully.

Fit a straight line model of y on x and add this fitted line to the plot. Save the fit object. Report the value of the slope estimate.

Show your code.

```
plot(observational$x,observational$y,xlim = c(0, 30), ylim = c(0,30), pch = 19,
     col = adjustcolor("black", 0.3),xlab='x',ylab='y' )
fit<-lm(observational$y~observational$x)
abline(fit)
```




```
#slope estimate
fit$coefficients['observational$x']
```

```
## observational$x
##      -0.4115873
```

c. **The measuring system.** Recall the model fitted in part (b).

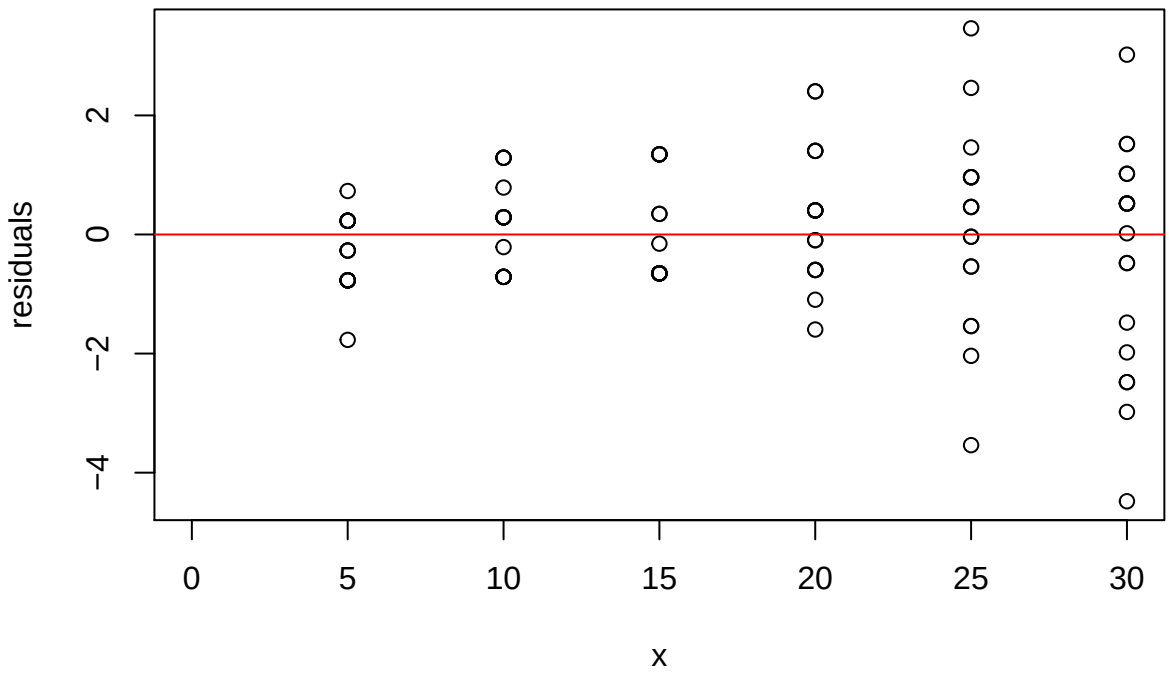
d. *(3 marks)* Plot the residuals (on the vertical axis) against the x values (on the horizontal axis).

Use `xlim = c(0, 30)`.

Make sure the plot is meaningfully labelled. Add a horizontal line at 0.

Show your code.

```
plot(observational$x, fit$residuals,ylab='residuals',xlab='x',xlim = c(0, 30))
abline(h=0,col='red')
```



ii. *(2 marks)* Based on what you see in the above residual plot, what would you say about the measuring system for y? Justify your answer.

For all the x values, the residuals are mostly scattered around zero but as the values of x increase, the range of residuals also seem to increase. When x=25 and x=30, the range of residuals is approximately 2 to -4. Hence, the measuring system of y is not good because when the values of x increase, the variability of y increases.

d. **Learning from repetition** Each team executed the same plan. To gain a better appreciation of the qualities of that plan, we investigate the individual team estimates of β.

e. *(4 marks)* First fit a separate line for each team's data. Capture the slope estimate of each fit and collect these into a single vector. Report the average of the estimated slopes.

Show your code.

```
s<-c()
for(i in 1:18){
  r<-observational[observational$team==i,]
  f<-lm(r$y~r$x)
  s[i]<-(f$coefficients['r$x'])
}
mean(s)
```

```
## [1] -0.4115873
```

ii. *(3 marks)* Recall that in fitting a straight line model of y on x, the least-squares estimate of the slope coefficient for x, using n points is

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

You will have already observed that the average of the slopes for the individual team data is identical to the slope estimate of the observational data combined (i.e. ignoring the teams).

Must this be so? A little more notation may help.

Let the teams be indexed $j = 1, \ldots, J$, and each team has exactly m pairs of observations which will be indexed by i. Then, the slope estimate for the jth team's data pairs (x_{ij}, y_{ij}), for $i = 1, \ldots, m$ is just

$$\widehat{\beta}_j = \frac{\sum_{i=1}^{m}(x_{ij} - \bar{x}_j)y_{ij}}{\sum_{i=1}^{m}(x_{ij} - \bar{x}_j)^2}$$

where $\bar{x}_j = (\sum_{i=1}^{m} x_{ij})/m$ is the average of the xs from team j.

Explain mathematically, or otherwise prove, why in this study we have $\frac{\sum_{j=1}^{J}\widehat{\beta}_j}{J} = \widehat{\beta}$ from above with $n = m \times J$.

$$\sum_{j=1}^{J}\widehat{\beta}_j$$

$$= \sum_{j=1}^{J}\frac{\sum_{i=1}^{m}(x_{ij} - \bar{x}_j)y_{ij}}{\sum_{i=1}^{m}(x_{ij} - \bar{x}_j)^2}$$

since the x values in each team are the same, $\bar{x}_j = \bar{x}$

$$= \sum_{j=1}^{J}\frac{\sum_{i=1}^{m}(x_{ij} - \bar{x})y_{ij}}{\sum_{i=1}^{m}(x_{ij} - \bar{x})^2}$$

$$= \sum_{j=1}^{J}\frac{\sum_{i=1}^{m}(x_{ij} - \bar{x})y_{ij}}{(\sum_{i=1}^{m} x_{ij}^2 + \sum_{i=1}^{m}\bar{x}^2 - 2\bar{x}\sum_{i=1}^{m} x_{ij})}$$

$$= \sum_{j=1}^{J}\frac{\sum_{i=1}^{m}(x_{ij} - \bar{x})y_{ij}}{(\sum_{i=1}^{m} x_{ij}^2 + \sum_{i=1}^{m}\bar{x}^2 - 2\bar{x}m\bar{x}))}$$

$$= \sum_{j=1}^{J}\frac{\sum_{i=1}^{m}(x_{ij} - \bar{x})y_{ij}}{(\sum_{i=1}^{m} x_{ij}^2 + m\bar{x}^2 - 2m\bar{x}^2))}$$

$$= \sum_{j=1}^{J}\frac{\sum_{i=1}^{m}(x_{ij} - \bar{x})y_{ij}}{(\sum_{i=1}^{m} x_{ij}^2 - \frac{n}{J}\bar{x}^2))}$$

$$= J\sum_{j=1}^{J}\frac{\sum_{i=1}^{m}(x_{ij} - \bar{x})y_{ij}}{J\sum_{i=1}^{m} x_{ij}^2 - n\bar{x}^2}$$

$$= J\sum_{j=1}^{J}\frac{\sum_{i=1}^{m}(x_{ij} - \bar{x})y_{ij}}{J\sum_{i=1}^{m} x_{ij}^2 - n\bar{x}^2}$$

$$= J\Big(\frac{\sum_{i=1}^{m}(x_{i1}-\bar{x})y_{i1}}{J\sum_{i=1}^{m}x_{i1}^2 - n\bar{x}^2} + \frac{\sum_{i=1}^{m}(x_{i2}-\bar{x})y_{i2}}{J\sum_{i=1}^{m}x_{i2}^2 - n\bar{x}^2} + ... + \frac{\sum_{i=1}^{m}(x_{ij}-\bar{x})y_{ij}}{J\sum_{i=1}^{m}x_{ij}^2 - n\bar{x}^2}\Big)$$

since x values in each team are same, $\sum_{i=1}^{m} x_{ij}^2$ of each team are equal

$$= J\Big(\frac{\sum_{i=1}^{m}(x_{i1}-\bar{x})y_{i1}}{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2} + \frac{\sum_{i=1}^{m}(x_{i2}-\bar{x})y_{i2}}{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2} + ... + \frac{\sum_{i=1}^{m}(x_{ij}-\bar{x})y_{ij}}{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}\Big)$$

$$= J\Big(\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}\Big)$$

$$= J\Big(\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}\Big)$$

$$= J\Big(\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\Big)$$

$$= J\widehat{\beta}$$

therefore

$$\frac{\sum_{j=1}^{J}\widehat{\beta}_j}{J}$$

$$= \frac{J\widehat{\beta}}{J}$$

$$= \widehat{\beta}$$

iii. *(3 marks)* Draw a meaningfully labelled histogram of the individual slope coefficient estimates for all teams.

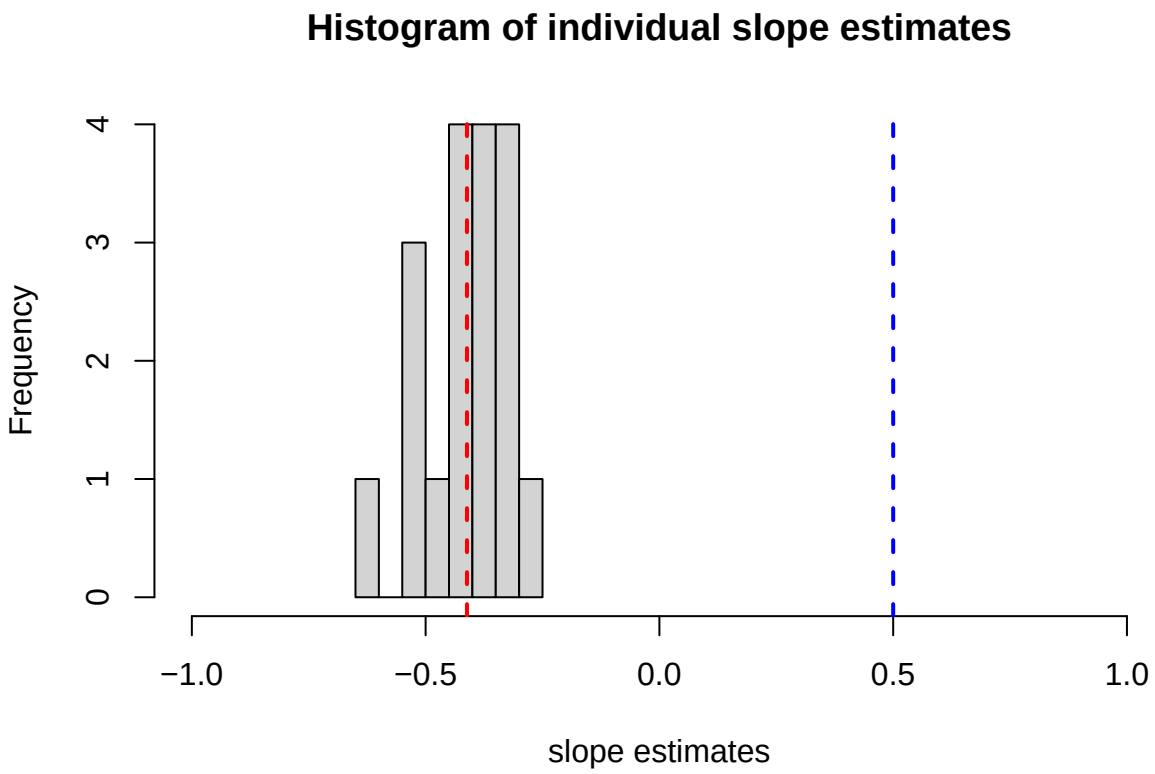
Show your code.

Use `xlim = c(-1, 1), col = "lightgrey"` in `hist()` and an appropriate `main` title and `xlab`.

Add a vertical red dashed line at the average of the slope estimates.

Add a vertical blue dashed line at the true value of β.

```
hist(s,xlim = c(-1, 1), col = "lightgrey" ,main = 'Histogram of individual slope estimates',xlab='slope
abline(v=mean(s),col='red',lwd=2,lty=2)
abline(v=0.5,col='blue',lwd=2,lty=2)
```



Conclusion

e. *(3 marks)* What do you conclude about the quality of team slope estimates from the observational study?

The total average slope estimates is negative and there is almost a difference of 1 when compared to the true slope estimate in this case. This shows that the quality of slope estimates from the observational study is not accurate.

f. *(2 marks)* What effect, if any, has been produced by a lurking variable? Explain.

Here we are analysing the relationship between x and y taking type as observational. So here a lurking variable might have produced a negative relationship between x and y since the true beta is positive (0.5), which implies that the relationship between x and y should be postive rather than negative.