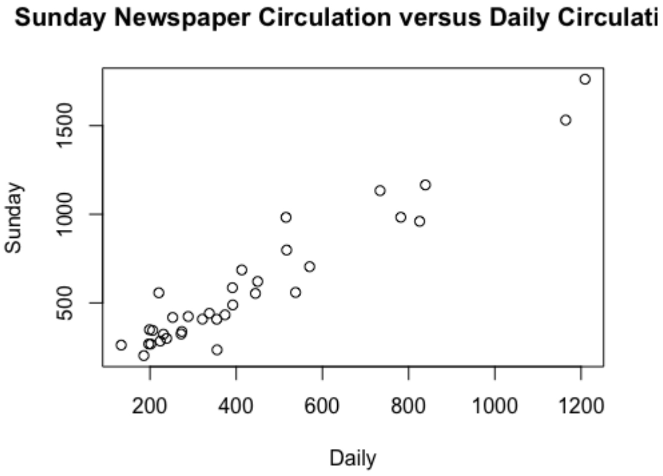
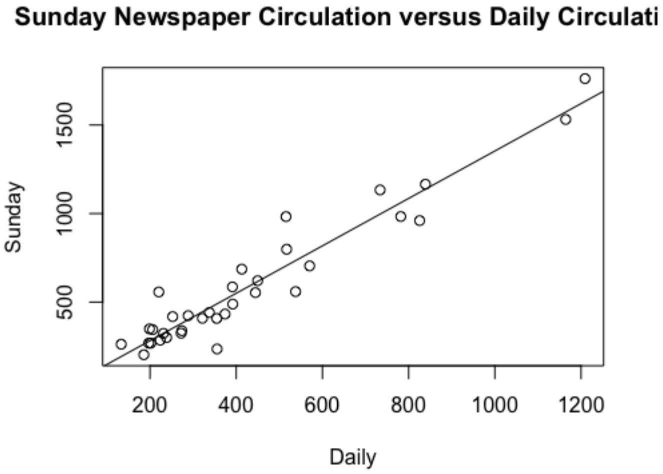
1 Newspaper Circulation

a. Below is a scatter plot of Sunday circulation versus daily circulation. Below that is the R code that generated the plot. This plot indeed suggests a linear relationship between daily and Sunday circulation. This relationship is plausible because for every person who gets a daily newspaper, they are likely to also be a Sunday subscriber, plus a handfull of other customers who only want Sunday papers. Therefore, we would expect Sunday subscriptions to increase linearly with daily subscriptions.



```
plot(df$Daily, df$Sunday, xlab = "Daily", ylab = "Sunday",
main="Sunday␣Newspaper␣Circulation␣versus␣Daily␣Circulation")
```

b. Below is the plot showing the regression line, with the regression equation and R code following.



$$\widehat{Sunday} = 13.836 + 1.340(Daily) \tag{1}$$

R Code:

```
model <- lm(Sunday ~ Daily, data=df)
abline(model)
summary(model)
```

c. The confidence intervals for β_0 and β_1 are as follows:

$$\beta_0 : (-59.095, 86.766)$$
$$\beta_1 : (1.196, 1.484)$$

R Code:

```
confint(model, level = .95)
```

d. If we assume that there is no relationship between daily circulation and Sunday circulation, the chances of seeing these results is 2e−16. This probability is extremely low, so we can conclude that there is a relationship between Sunday circulation and daily circulation. To conclude this, I observed the summary table in R, which showed that the t-value for my null hypothesis is 18.935, which corresponds to a p-value of 2e−16.

R Code:

```
summary(model)
```

e. 91.55% of the variability in Sunday circulation is explained by daily circulation.

R Code:

```
summary(model)
```

f. The interval estimate with a 95% level for the average Sunday circulation of newspapers with daily circulation of 500, 000 is (644.195, 723.191).

R Code:

```
predict(model, newdata = list(Daily=500), interval='confidence', level = .95)
```

g. The interval estimate with a 95% level for the predicted Sunday circulation of newspapers with daily circulation of 500, 000 is (457.337, 910.039).

R Code:

```
predict(model, newdata = list(Daily=500), interval='prediction', level = .95)
```

The prediction interval is the same as the confidence interval, except the variance is increased by σ^2. This makes sense because a confidence interval accounts for uncertainty about the population mean, while a prediction interval must account for the uncertainty of the population mean *and* variation for one given survey result.

2 Galaxy Distance and Velocity

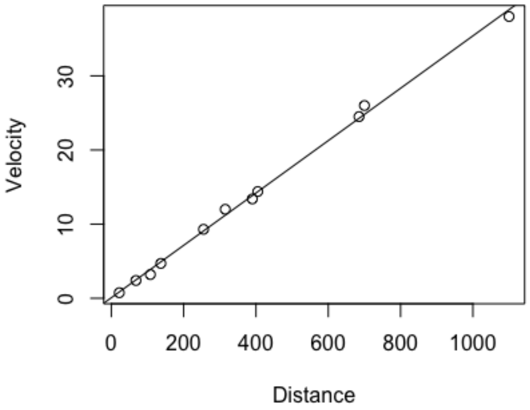
a. If we assume $\beta_0 = 0$, then we must minimize

$$\sum_{i=1}^{n}(Y_i - (\beta_1 X_i))^2$$

We can take the derivative with respect to β_1 and set it equal to 0 to find when the expression is minimized:

$$0 = \sum_{i=1}^{n} 2(Y_i - (\beta_1 X_i))X_i$$
$$0 = \sum_{i=1}^{n} Y_i X_i - \beta_1 X_i^2$$
$$\beta_1 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}$$

b. Below is the graph of Velocity versus Distance of the given galaxy clusters, along with the usual simple regression line:

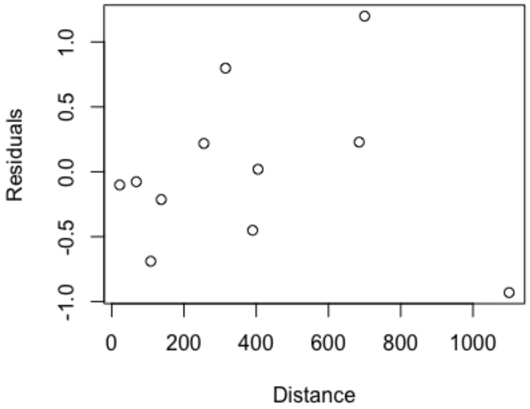


This linear regression yielded the following equation:

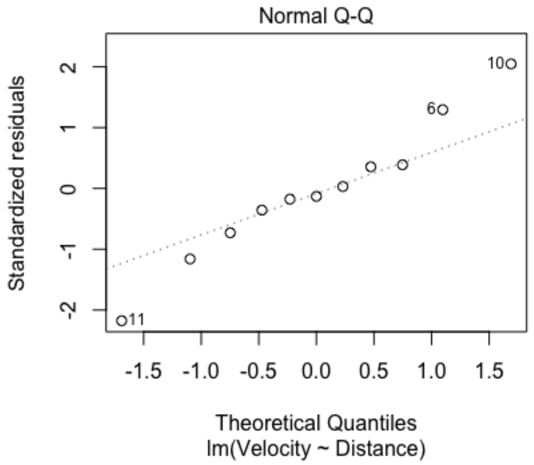
$$\widehat{Velocity} = .075 + .035(Distance) \tag{2}$$

If we assume that $\beta_0 = 0$, the chances of getting this result is .813, so there is not enough evidence to suggest that $\beta_0 \neq 0$. Therefore, Hubble's model is appropriate given this data.
However, it is worth noting that the conditions for linear regression were not met, so conclusions made from this linear regression may not be accurate. Examining our necessary assumptions, we see:

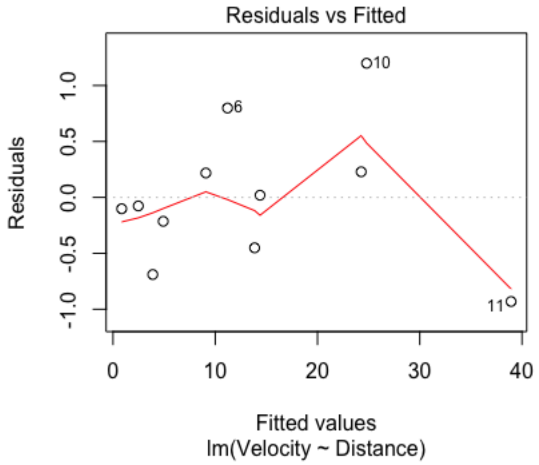
b.i. Linearity: The data from the above chart indeed appears linear. However, if we examine the plot of residuals versus distance, we see that the scatter does not appear random. In fact, the residuals appear to spread out more as distance gets larger, suggesting that the data is not linear.



b.ii. Normality: If we examine the qq-plot for this regression, we see that the data appears to be heavy-tailed, which indicates a departure from normality.



b.iii. Homoscedasticity: Looking at the residuals versus fitted values, we see that there is not an even spread. In fact, there are three clear outliers (points 6, 10, and 11). Therefore, this data set does not meet the requirement of homoscedasticity.



It is important to note that two of these outliers, points 10 and 11, are the two galaxies with the greatest distance. Without accessing more data, it is impossible to know why the far away galaxies behave as outliers.

R Code:

```
galData <- read.table("GalacticClusters2.txt", header = TRUE) #read in data
galLine2 <- lm(Velocity ~ Distance, data=galData) #regular linear regression
summary(galLine2) #summary of results
abline(galLine2) #graph regression
plot(galLine2) #plots for normality and homoscedasticity
galResiduals <- resid(galLine2)
plot(galResiduals ~ galData$Distance, ylab = "Residuals", xlab = "Distance") #plots fo
```

c. If we assume $\beta_0 = 0$, then after performing regression we derive:

$$\widehat{Velocity} = .035(Distance) \tag{3}$$

This result provides $\beta_1 = .03544$. This is Hubble's constant. From Hubble's constant, we can derive the age of the universe. The units of β_1 are $\frac{\text{thousands of miles}}{\text{seconds}\times\text{millions of light years}}$. Therefore, by simple algebra, we see that $\frac{1.86\text{e}8}{\beta_1}$ will yield the age of the universe. From this calculation, we find that the age of the universe is 5.25 billion years old. However, we know that this is slightly less than half of the true age of the universe, suggesting that we perhaps need a larger sample of data to do an accurate calculation. Alternatively, this method for calculating the age of the universe could be invalid.

R Code:

```
galLine <- lm(Velocity ~ Distance + 0, data=galData)
```