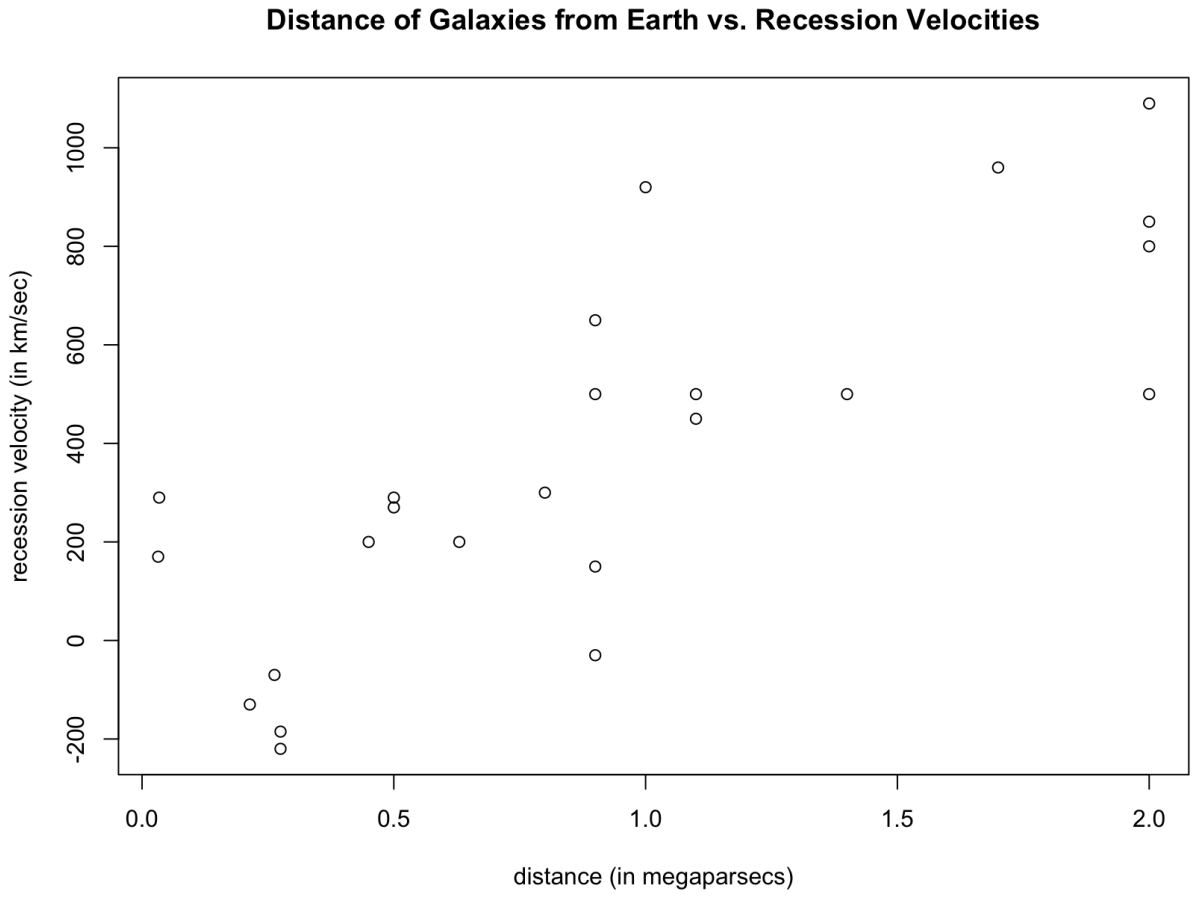
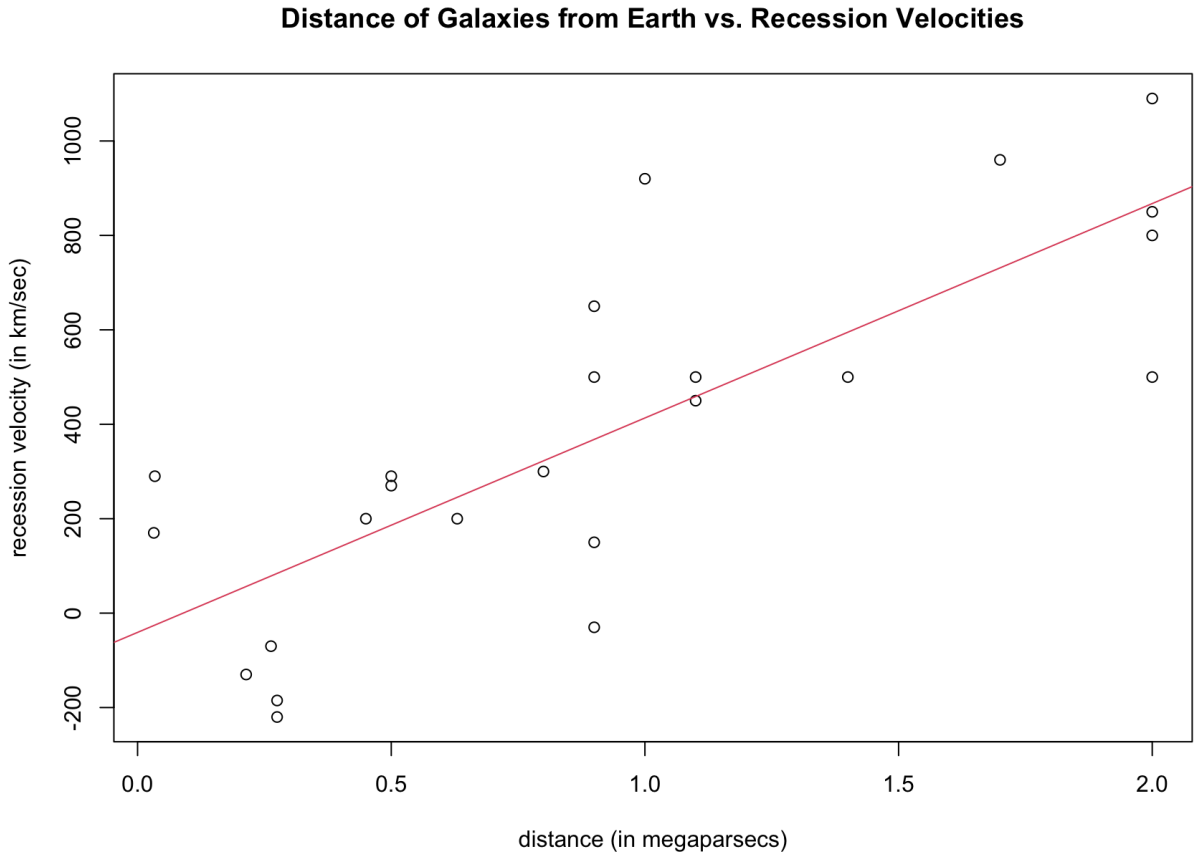
STAT 463 - Assignment 2
Due Date: Saturday February 25th, 2023

Data Description and Background
In 1929, Edwin Hubble investigated the relationship between the distance of a galaxy from the earth and the velocity with which it appears to be receding. Galaxies appear to be moving away from us no matter which direction we look. This is thought to be the result of the "Big Bang". Hubble hoped to provide some knowledge about how the universe was formed and what might happen in the future. The data collected includes distances (in megaparsecs) to 24 galaxies and their recession velocities (in km/sec). Note: 1 parsec = 3.26 light years. In a short write-up address the following:

1) Visually display the data in an appropriate graph and comment on anything that may be of note. Be sure to adequately label the graph, indicating in a title what is being displayed as well as what each axis represent (including units of measurement). – 2 points



2) It is desired to ascertain whether galaxies that are farther away tend to be receding faster from Earth (i.e. they have higher recession velocities). Based on the plot you constructed in part (1), would a linear regression model with recession velocity as the response variable and distance as the explanatory variable be appropriate for this dataset? – 2 points



A linear regression model with recession velocity as the response variable and distance as the explanatory variable would be appropriate for this dataset because we are observing the effect distance from earth has on recession velocities of galaxies.

3) What is the nature of the relationship between galaxies' distances from earth and their recession velocities? Specifically, as the distance increases, how is the recession velocity affected? Quantify it and interpret this quantity. – 2 points

```
Call:
lm(formula = recession_velocity ~ distance, data = hubble.data)

Residuals:
    Min      1Q  Median      3Q     Max
-397.96 -158.10  -13.16  148.09  506.63

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -40.78      83.44  -0.489     0.63
distance      454.16      75.24   6.036 4.48e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 232.9 on 22 degrees of freedom
Multiple R-squared: 0.6235, Adjusted R-squared: 0.6064
F-statistic: 36.44 on 1 and 22 DF, p-value: 4.477e-06

Based on the linear regression model summary, the slope term indicates that for every 1 megaparsec increase in distance, the recession velocity will increase by 454.16 km/sec. As distance increases the recession velocity also increases, indicating a positive linear relationship between the two variables.

4) Does the data suggest that there is a statistically significant relationship between galaxies' recession velocities and their distances from earth? – 3 points

The p value of 4.48e-06 from the summary indicates that there is a statistically significant relationship between galaxies' recession velocities and their distances from earth becuase the pvalue is less than the significance value of 0.001 therefore we can reject the null hypothesis (no relationship exists) which allows us to conclude that there is a relationship between galaxies' recession velocities and their distances from earth.

5) How well do galaxies' distances from earth explain their recession velocities? Quantify this and interpret this quantity. – 2 points

The R-squared value from the summary is equal to 0.6235 and as the coefficent of determination, this indicates that 62.35% of the variance found in the response variable (recession velocities) can be explained by the explanatory variable (galaxies' distances)

6) Andromeda, the nearest large spiral galaxy to our home galaxy the Milky Way, is 0.77 megaparsecs away. Give an estimate of its recession velocity. Additionally, provide a range of value that will encompass Andromeda's true recession velocity with probability 0.95. – 4 points

An estimate of Andromeda's recession velocity is 308.9184 km/sec and the interval from (-184.5625, 802.3992) will encompass Andromeda's true recession velocity with probability 0.95.

Code I ran:

```
plot(hubble.data[,1],hubble.data[,2],main="Distance of Galaxies from Earth vs. Recession Velocities",xlab="distance (in megaparsecs)",ylab="recession velocity (in km/sec)")
hub.fit<-lm(recession_velocity~distance,data=hubble.data)
abline(hub.fit$coe[1], hub.fit$coe[2], col = 2)
summary(hub.fit)
predict(hub.fit, data.frame(distance= 0.77))
predict(hub.fit, data.frame(distance= 0.77), interval='prediction',level = 0.95)
```