Homework 03 - Regression and Correlation.

Observations made by the astronomer Edwin Hubble showed that the universe is expanding. If v is the galaxy's recession from the Milky Way and d is the distance to that galaxy, Hubble's law is the linear relationship.

$$v = H_0 d.$$

H_0 is known as **Hubble's constant**. To estimate H, we use the data below, Distance in measured in millions of light years. Velocity is measured in thousands of miles per second.

Cluster	Distance	Velocity
Virgo	22	0.8
Pegasus	68	2.4
Perseus	108	3.2
Coma Berenices	137	4.7
Ursa Major #1	255	9.3
Leo	315	12.0
Corona Borealis	390	13.4
Gemini	405	14.4
Bootes	685	24.5
Ursa Major #2	700	26.0
Hydra	1100	38.0

Problem 1

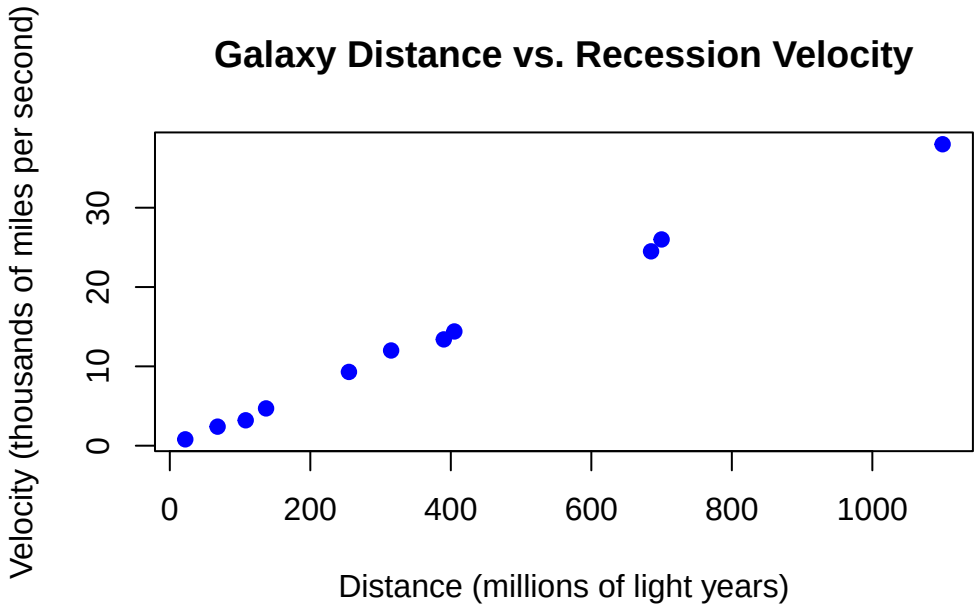
```
dist <- c(22, 68, 108, 137, 255, 315, 390, 405, 685, 700, 1100)

velo <- c(0.8, 2.4, 3.2, 4.7, 9.3, 12.0, 13.4, 14.4, 24.5, 26.0, 38.0)
```

Part a)

Provide a scatterplot of the data using d as the explanatory variable

```
plot(dist, velo, main = "Galaxy Distance vs. Recession Velocity",
     xlab = "Distance (millions of light years)",
     ylab = "Velocity (thousands of miles per second)",
     col = "blue", pch = 19)
```



Part b)

Determine corresponding the regression line, and report your estimate of Hubble's constant. (Note that the y-intercept is 0.)

```
lin_reg <- lm(velo ~ dist - 1)

summary(lin_reg)
```

```
Call:
lm(formula = velo ~ dist - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98437 -0.28853  0.02031  0.24304  1.19176

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
dist 0.0354403  0.0003765   94.13 4.48e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.6169 on 10 degrees of freedom
Multiple R-squared:  0.9989,    Adjusted R-squared:  0.9988
F-statistic:  8861 on 1 and 10 DF,  p-value: 4.482e-16
```

The estimated Hubble's constant is approximately 0.0354 thousands of miles per second per million light years.

Part c)

Give the sum of the residuals. Note that it is not equal to zero. What would lead this to be the case?

```r
reg_resid <- resid(lin_reg)

sum(reg_resid)
```

```
[1] 0.3821927
```

The sum of the residuals is not zero because the regression line is forced through the origin (no intercept term). In a standard regression with an intercept, the residuals always sum to zero by construction. When we remove the intercept, this property no longer holds.

Part d)

Determine the equation for the regression line using v as the explanatory variable.

```r
lin_reg2 <- lm(dist ~ velo - 1)

summary(lin_reg2)
```

```
Call:
lm(formula = dist ~ velo - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-32.800  -6.320  -0.548   8.429  28.984
```

```
Coefficients:
     Estimate Std. Error t value Pr(>|t|)
velo  28.1846     0.2994   94.13 4.48e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.4 on 10 degrees of freedom
Multiple R-squared:  0.9989,    Adjusted R-squared:  0.9988
F-statistic:  8861 on 1 and 10 DF,  p-value: 4.482e-16
```

The regression line using v as the explanatory variable is $d = 28.1846 \cdot v$.

Part e)

Find the product of the slopes of the two regression lines.

```r
slope1 <- lin_reg$coefficients[1]
slope2 <- lin_reg2$coefficients[1]

slope1 * slope2
```

```
     dist
0.9988727
```

The product of the two slopes is 0.9989.

Part f)

How close are they are being the same line? Explain your answer.

If the two regression lines were identical, the product of their slopes would equal exactly 1 (since one slope would be the reciprocal of the other). The product we computed is very close to 1, which indicates the two lines are nearly the same. This makes sense because the data points lie very close to a straight line through the origin, meaning the correlation between distance and velocity is very strong. The closer the correlation is to ± 1, the closer the product of the slopes is to 1, and the more the two regression lines coincide.