Statistical mathematics project work

Trieu Huynh Ba Nguyen

Task 1:

a) Geometric distribution represents the probability of the number of successive failures before the first success. It is calculated with the function:

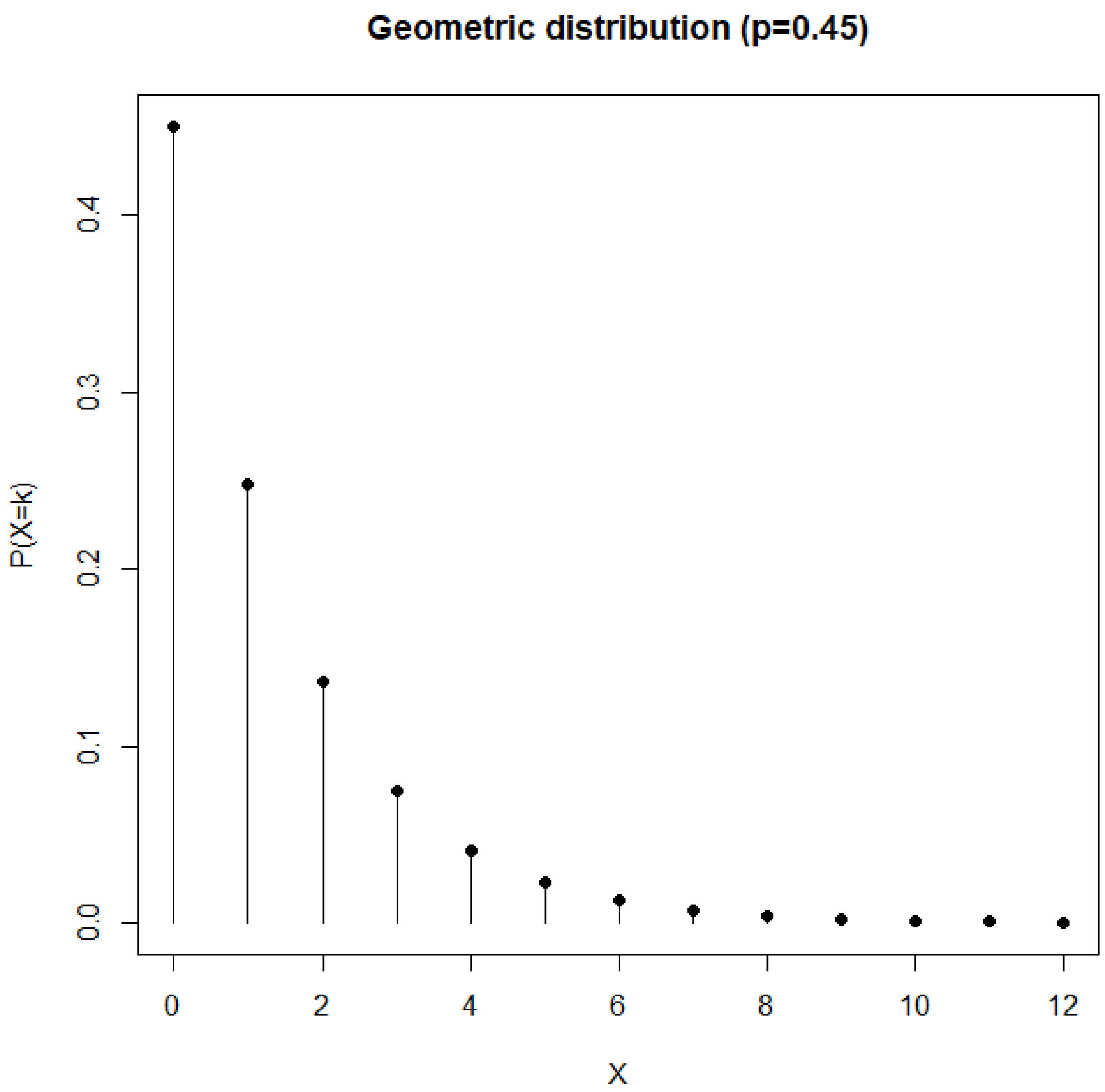
$$P_X(k) = p(1-p)^{k-1}$$

with k: number of trials (k = 1, 2, 3…)

p: probability of each success

b) R code:

```
X <- 0:12
Y <- dgeom(X, p = 0.45)
plot(X, Y, type = "h", main = "Geometric distribution (p=0.45)", ylab =
"P(X=k)")
points(X, Y, pch = 16)
```



Task 2:

a) Binominal distribution represents the probability of the number of successes or failures occurring during a number of trials. It is calculated with the function:

$$P_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

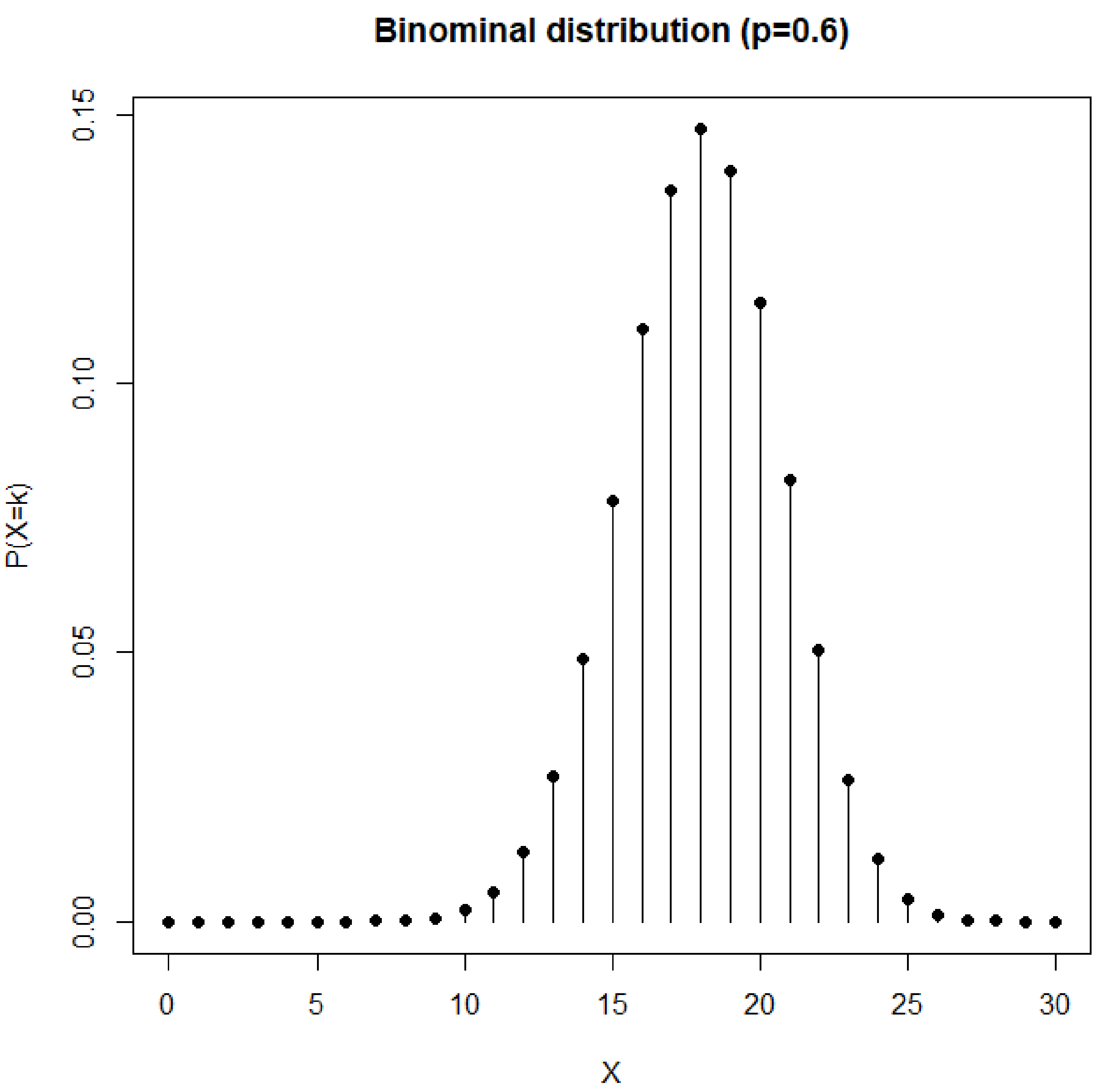
with k: number of successes or failures (k = 1, 2, 3…n)

n: total number of trials

p: probability of each success

b) R code:

```
X <- 0:30
Y <- dbinom(X, size = 30, prob = .6)
plot(X, Y, type = "h", main = "Binominal distribution (p=0.6)", ylab =
"P(X=k)")
points(X, Y, pch = 16)
```



Task 3:

a) Poisson distribution represents the probability of a given number of events occurring in an interval of time. It is calculated with the function:

$$P_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

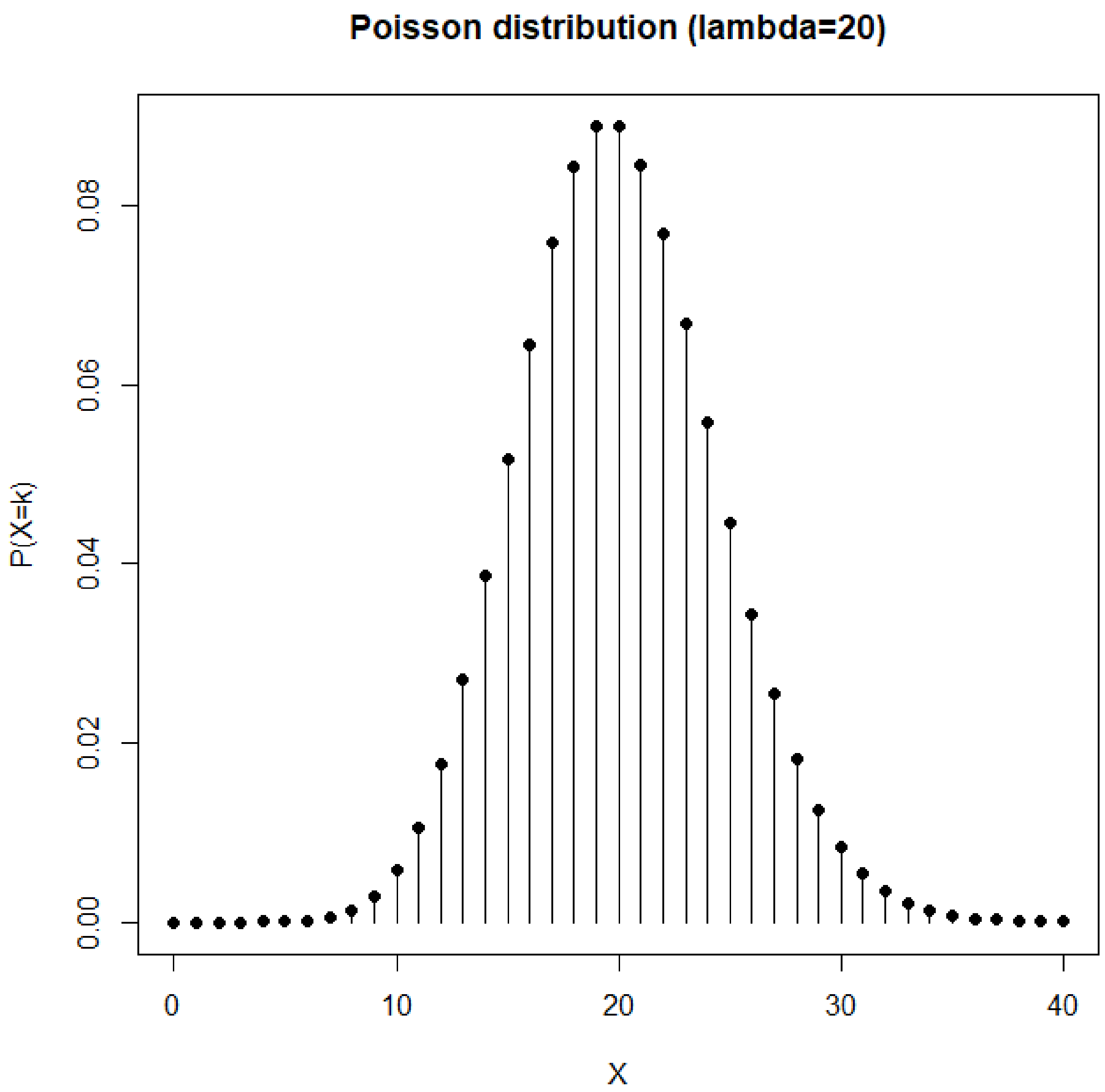
with k: the number of occurrences (k = 1, 2, 3…)

e: Euler's number (e=2.71828…)

λ: lambda – total number of events divided by the number of units in the data

b) R code:

```
X <- 0:40
Y <- dpois(X, lambda = 20)
plot(X, Y, type = "h", main = "Poisson distribution (lambda=20)", ylab =
"P(X=k)")
points(X, Y, pch = 16)
```



Task 4:

1. There is one column named “Val”.
2. There are 1029 rows.
3. Min = 4.193534

```
min(data_set1)
```

4. Max = 109.379

```
max(data_set1)
```

5. Mean = 50.49665

```
mean(data_set1$Val)
```

6. Median = 50.52415

```
median(data_set1$Val)
```

7. Variance = 218.7175

```
var(data_set1$Val)
```

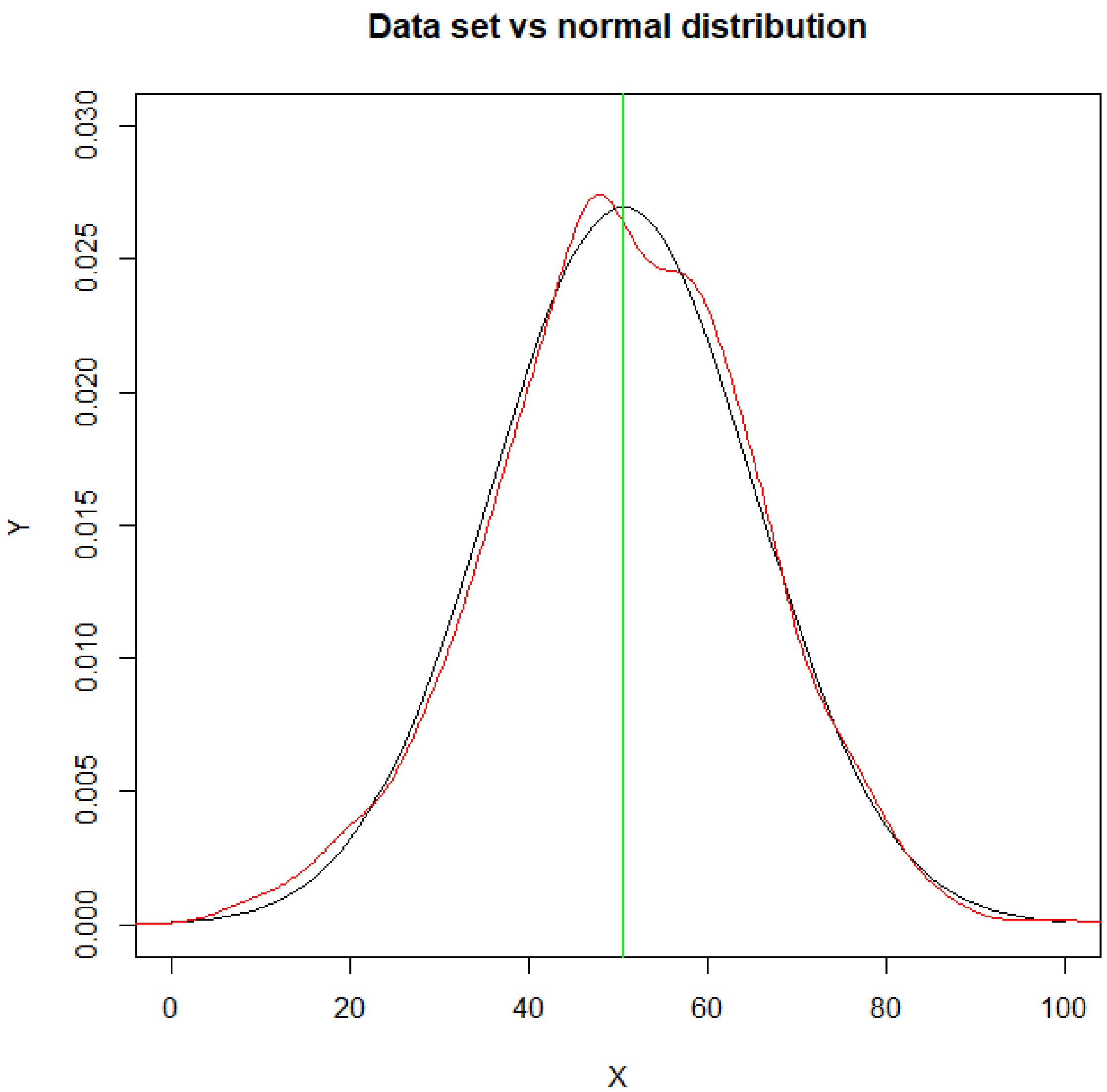
8. Standard deviation = 14.7891

```
sd(data_set1$Val)
```

Task 5:

R code:

```
library(readr)
data_set1 <- read_csv("data_set1.csv")
X <- 0:100
Y <- dnorm(X, mean = mean(data_set1$Val), sd = sd(data_set1$Val))
plot(X, Y, type = "l", ylim = c(0, 0.03), main = "Data set vs normal
distribution")
d <- density(data_set1$Val, bw = 3)
points(d, col = "red", type = "l")
abline(v = mean(data_set1$Val), col = "green")
```



Task 6:

4 variables most correlated with hp: mpg, cyl, disp, carb.

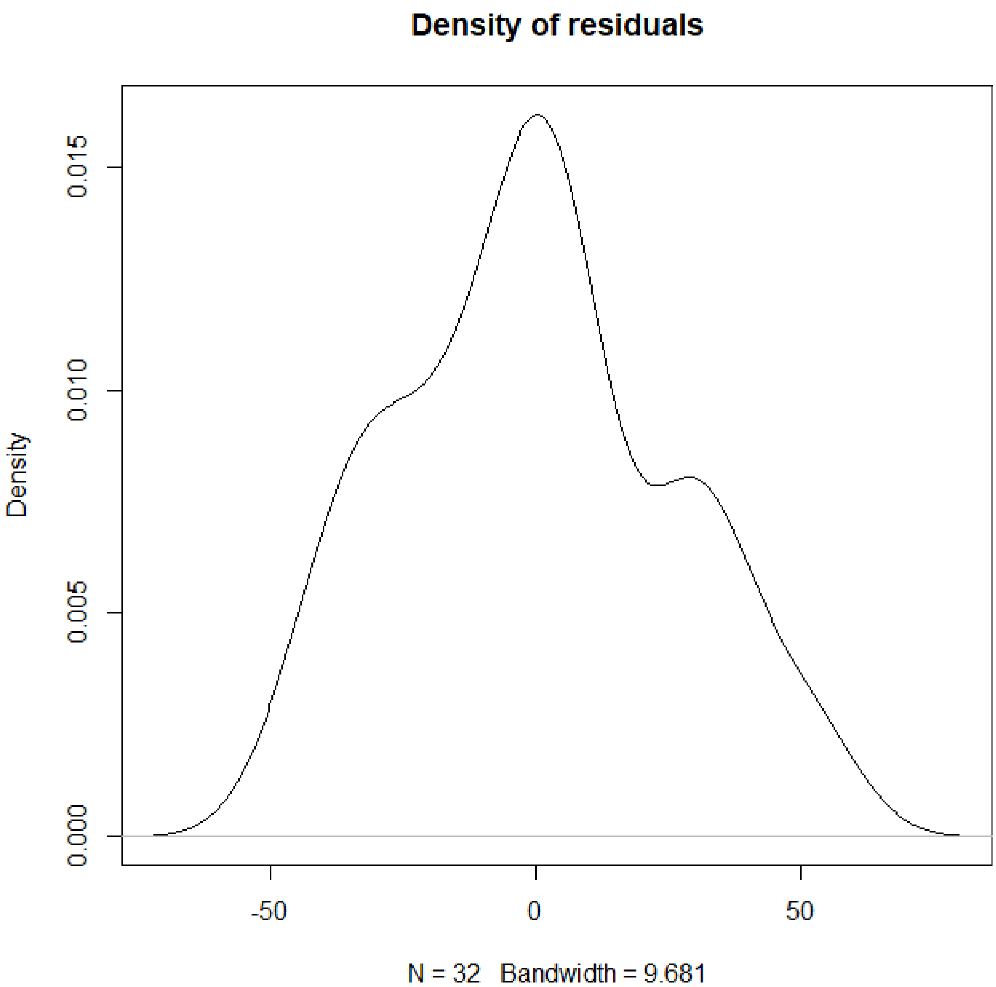
```
cars <- mtcars
round(cor(cars), digits = 2)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.00	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.60	0.48	-0.55
cyl	-0.85	1.00	0.90	0.83	-0.70	0.78	-0.59	-0.81	-0.52	-0.49	0.53
disp	-0.85	0.90	1.00	0.79	-0.71	0.89	-0.43	-0.71	-0.59	-0.56	0.39
hp	-0.78	0.83	0.79	1.00	-0.45	0.66	-0.71	-0.72	-0.24	-0.13	0.75
drat	0.68	-0.70	-0.71	-0.45	1.00	-0.71	0.09	0.44	0.71	0.70	-0.09
wt	-0.87	0.78	0.89	0.66	-0.71	1.00	-0.17	-0.55	-0.69	-0.58	0.43
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1.00	0.74	-0.23	-0.21	-0.66
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1.00	0.17	0.21	-0.57
am	0.60	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1.00	0.79	0.06
gear	0.48	-0.49	-0.56	-0.13	0.70	-0.58	-0.21	0.21	0.79	1.00	0.27
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1.00

Task 7:

R code:

```
model <- lm(hp ~ cyl + disp + carb + mpg, data = mtcars)
hp_hat <- predict(model)
residuals <- mtcars$hp - hp_hat
hpplot <- density(residuals)
plot(hpplot, main = "Density of residuals")
summary(model)$r.squared
```



The plot has a bell shape.

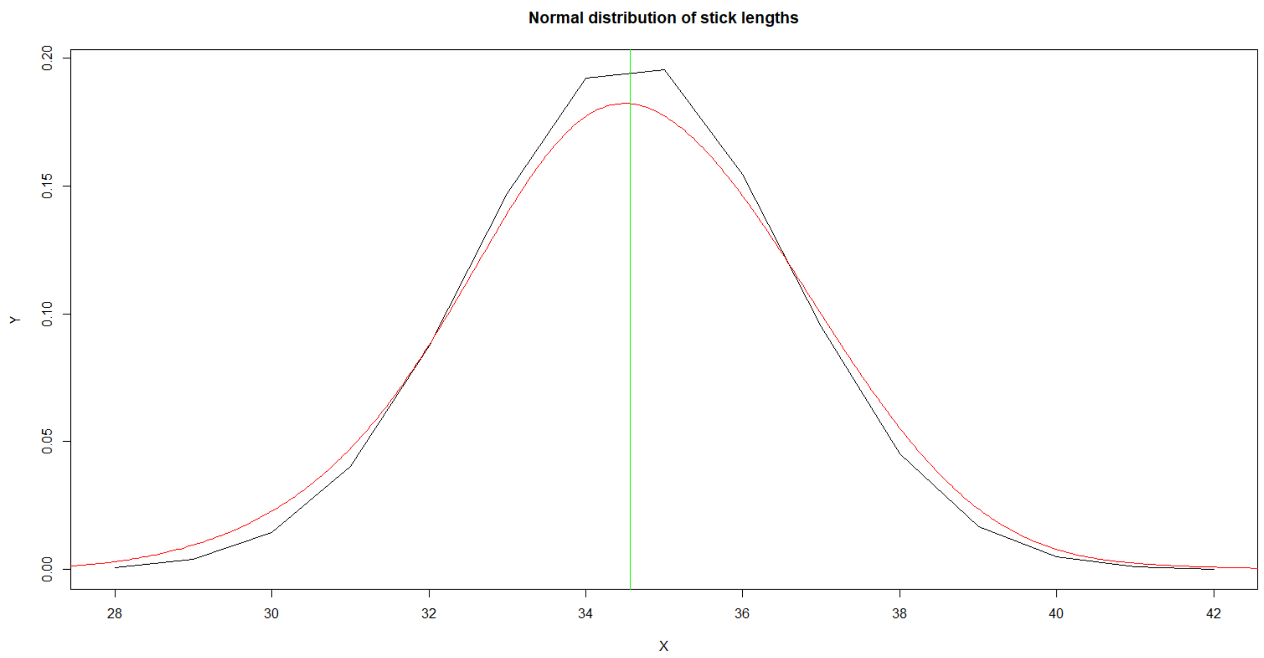
R-squared = 0.8594845

Therefore, the model is correct and accurate.

Task 8:

R code:

```
library(readr)
data_set2 <- read_csv("data_set2.csv")
X <- min(data_set2):max(data_set2)
Y <- dnorm(X, mean = mean(data_set2$Val), sd = sd(data_set2$Val))
plot(X, Y, type = "l", main = "Normal distribution of stick lengths")
d <- density(data_set2$Val, bw = 1)
points(d, col = "red", type = "l")
abline(v = mean(data_set2$Val), col = "green")
```



The length of the sticks is not acceptable as the mean and most values are much higher than the null hypothesis of μ = 30.