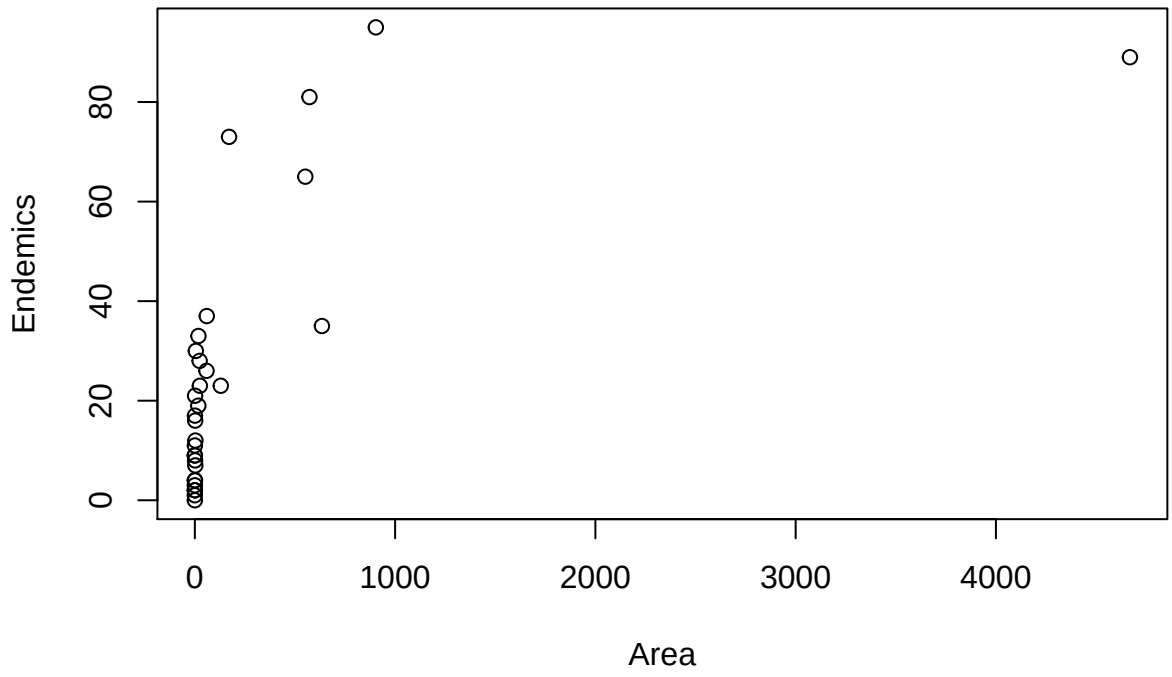
Homework1

Gulishana Adilijiang

Problem 1

(a)

```
library(faraway)
data(gala)
x = gala$Area
y = gala$Endemics
plot(x,y, xlab = "Area", ylab = "Endemics")
```

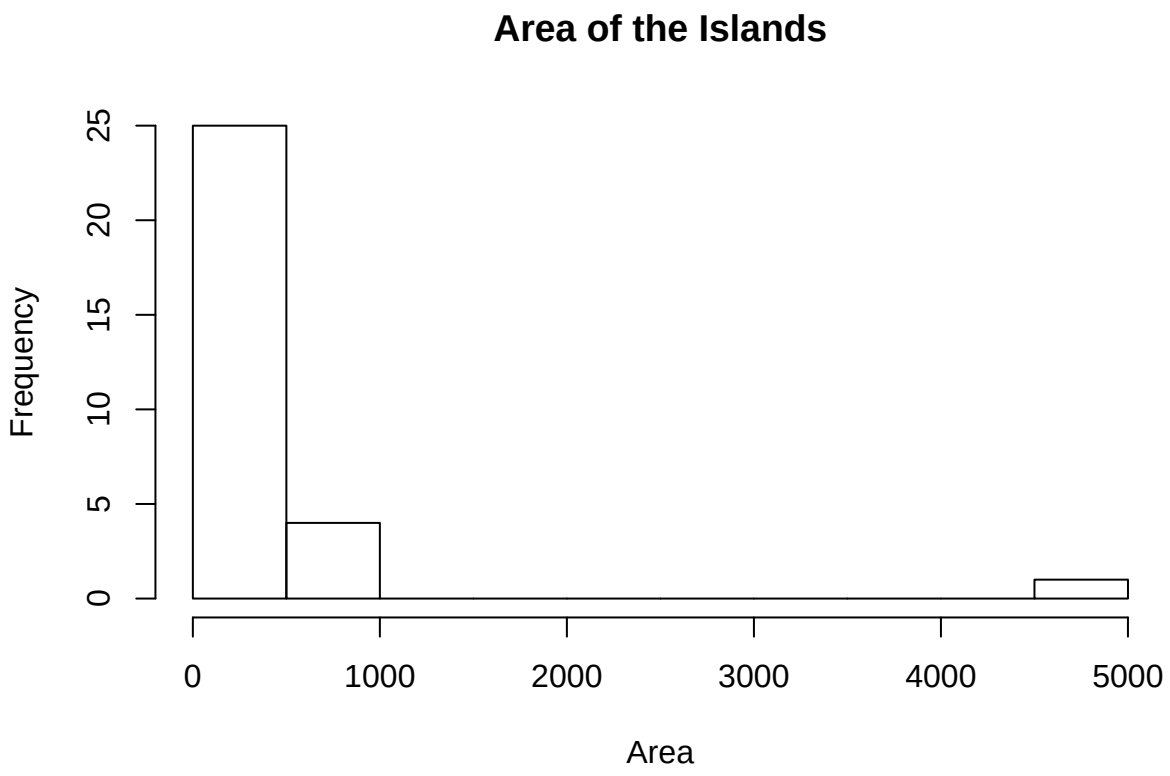


Answer: According to the scatterplot, a linear model is not appropriate for this data set. Most of the islands have a small size of area while some islands have a quite larger size of area.

```
summary(x)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.010    0.258    2.590  261.709   59.238 4669.320
```

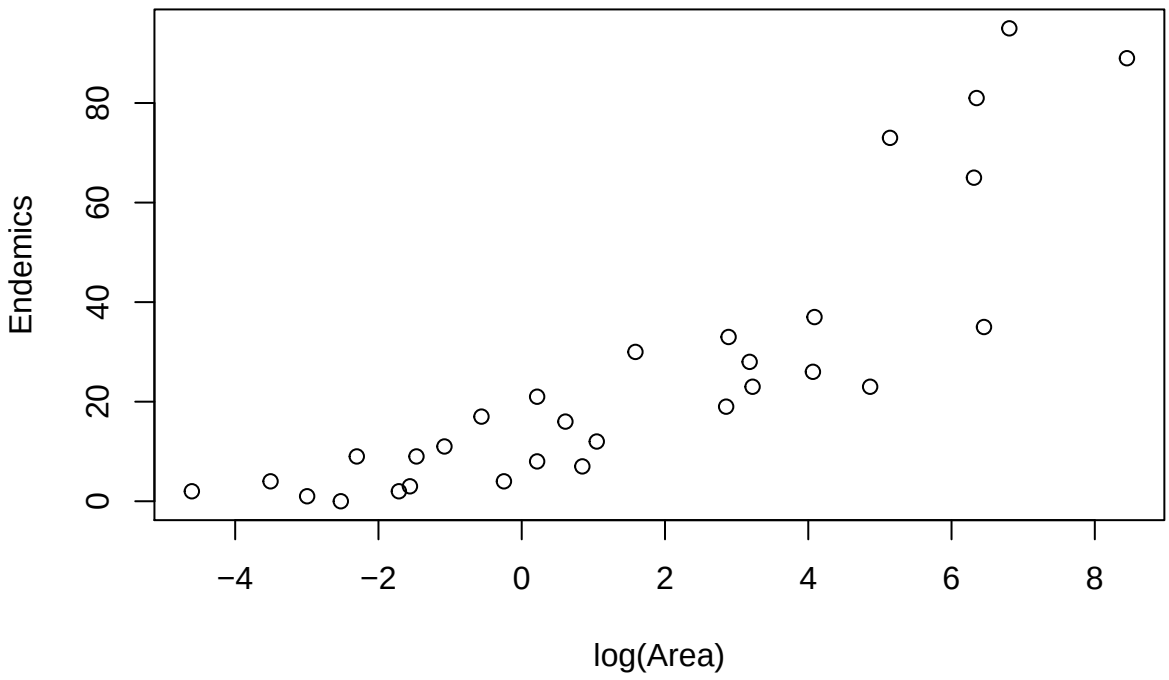
```
hist(x, main = "Area of the Islands", xlab = "Area", breaks = 10)
```



Answer: There is an island which has a very large size of area comparing with the most of other islands. It looks like an outlier data point.

(b)

```
x = log(gala$Area)
plot(x,y, xlab = "log(Area)", ylab = "Endemics")
```



Answer: According to the scatterplot, a linear model seems to be appropriate for this data set.

(c)

```
SXY = sum((x-mean(x))*(y-mean(y)))
SXX = sum((x-mean(x))^2)
beta1_hat = SXY/SXX; beta1_hat
```

```
## [1] 6.697806
```

```
beta0_hat = mean(y)-beta1_hat*mean(x); beta0_hat
```

```
## [1] 15.69099
```

```
y_hat = beta0_hat + beta1_hat * x
residuals = y - y_hat
n = length(x)
variance_hat = sum((residuals^2)) / (n-2) ; variance_hat
```

```
## [1] 204.2586
```

Answer: So $\hat{\beta_0} = 15.69099$, $\hat{\beta_1} = 6.697806$, $\hat{\sigma^2} = 204.2586$

(d)

```
x_new = log(2.0)
y_new = beta0_hat + beta1_hat * x_new ; y_new
```

```
## [1] 20.33355
```

Answer: So the predicted number of species is 20.

Problem 2

(a)

```
cor(y_hat, residuals)
```

```
## [1] 1.880957e-16
```

Answer: The correlation between the vector of fitted values and the vector of residuals is close to zero. This is consistent with the OLS property that the covariance between the fitted values and the residuals is zero.

(b)

```
cor(residuals, gala$Nearest)
```

```
## [1] -0.4099413
```

Answer: The vector of residuals and the variable gala$Nearest has a negative correlation which equals to -0.4. This means when the island is farther from any other island, the residuals become smaller. It makes sense because the distance from one island to its nearest island is correlated with the area of the island and the number of endemic species on the island, so it affects the relationship of these two variables.

Problem 4

(a)

```
empirical_vector = NULL
for (i in 1:1000){
    # generate simulated data set
    x = runif(n = 100, min = -1, max = 1)
    error = rnorm(n = 100, mean = 0, sd = 1)
    y = 1 + x + error

    # compute OLS estimate of beta1_hat
    model = lm(y ~ x)
    summary(model)
    beta1_hat = summary(model)$coefficients[2]

    # compute the 90% confidence interval for true coefficient beta1
    std_beta1_hat = summary(model)$coefficients[4]
    alpha = 1- 0.9
    t_value = qt(p = 1-alpha/2, df=100-2)
    confidence_interval_lower = beta1_hat - std_beta1_hat * t_value
    confidence_interval_upper = beta1_hat + std_beta1_hat * t_value

    # measure if the confidence interval actually contains true beta1 (= 1)
    empirical = (1>=confidence_interval_lower & 1<=confidence_interval_upper)
    empirical_vector[i] = empirical
}

coverage_rate = sum(empirical_vector) / length(empirical_vector) ; coverage_rate
```

```
## [1] 0.904
```

(b)

```r
empirical_vector = NULL
for (i in 1:1000){
    # generate simulated data set
    x = runif(n = 100, min = -1, max = 1)
    error = rnorm(n = 100, mean = 0, sd = abs(x))
    y = 1 + x + error

    # compute OLS estimate of beta1_hat
    model = lm(y ~ x)
    summary(model)
    beta1_hat = summary(model)$coefficients[2]

    # compute the 90% confidence interval for true coefficient beta1
    std_beta1_hat = summary(model)$coefficients[4]
    alpha = 1- 0.9
    t_value = qt(p = 1-alpha/2, df=100-2)
    confidence_interval_lower = beta1_hat - std_beta1_hat * t_value
    confidence_interval_upper = beta1_hat + std_beta1_hat * t_value

    # measure if the confidence interval actually contains true beta1 (= 1)
    empirical = (1>=confidence_interval_lower & 1<=confidence_interval_upper)
    empirical_vector[i] = empirical
}

coverage_rate = sum(empirical_vector) / length(empirical_vector) ; coverage_rate
```

```
## [1] 0.79
```

(c)

```r
empirical_vector = NULL
for (i in 1:1000){
    # generate simulated data set
    x = runif(n = 100, min = -1, max = 1)
    error = rnorm(n = 100, mean = 0, sd = abs(1-abs(x)) )
    y = 1 + x + error

    # compute OLS estimate of beta1_hat
    model = lm(y ~ x)
    summary(model)
    beta1_hat = summary(model)$coefficients[2]

    # compute the 90% confidence interval for true coefficient beta1
    std_beta1_hat = summary(model)$coefficients[4]
    alpha = 1- 0.9
    t_value = qt(p = 1-alpha/2, df=100-2)
    confidence_interval_lower = beta1_hat - std_beta1_hat * t_value
    confidence_interval_upper = beta1_hat + std_beta1_hat * t_value

    # measure if the confidence interval actually contains true beta1 (= 1)
    empirical = (1>=confidence_interval_lower & 1<=confidence_interval_upper)
    empirical_vector[i] = empirical
}
```

```
coverage_rate = sum(empirical_vector) / length(empirical_vector) ; coverage_rate
```

```
## [1] 0.992
```

(d)

Answer: If the linear regression model have non-constant error variance, it will break the Gauss-Markov theorem so that the OLS estimators will not be the Best Linear Unbiased Estimators and their variances will not be the lowest of all the other unbiased estimators. In this case, the OLS coefficient estimates are not biased, but the OLS estimates of the standard errors of coefficients are biased which lead to biased inference of coefficients.

3. $\because \tilde{X}_i = c(X_i+d)$ $\therefore \bar{\tilde{X}} = c(\bar{X}+d)$ $\Rightarrow \tilde{X}_i - \bar{\tilde{X}} = c(X_i+d) - c(\bar{X}+d) = c(X_i - \bar{X})$

$$\therefore \tilde{\beta}_1 = \frac{S_{\tilde{X}Y}}{S_{\tilde{X}\tilde{X}}} = \frac{\sum_i (\tilde{X}_i - \bar{\tilde{X}})(Y_i - \bar{Y})}{\sum_i (\tilde{X}_i - \bar{\tilde{X}})^2} = \frac{\sum_i c(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i c^2 (X_i - \bar{X})^2} = \frac{1}{c} \cdot \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{1}{c}\hat{\beta}_1$$

$$\therefore \tilde{\beta}_0 = \bar{Y} - \tilde{\beta}_1 \bar{\tilde{X}} = (\hat{\beta}_0 + \hat{\beta}_1 \bar{X}) - \frac{1}{c}\hat{\beta}_1 \cdot c(\bar{X}+d) = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 \bar{X} - d\hat{\beta}_1 = \hat{\beta}_0 - d\hat{\beta}_1$$

$$\Rightarrow \tilde{\beta}_1 = \frac{1}{c}\hat{\beta}_1 \quad \tilde{\beta}_0 = \hat{\beta}_0 - d\hat{\beta}_1$$

prove: for model $Y_i = \beta_0 + \beta_1 \tilde{X}_i + \varepsilon_i$, $RSS = \sum_i (Y_i - (\beta_0 + \beta_1 \tilde{X}_i))^2$

$$\frac{\partial RSS}{\partial \beta_0} = -2\sum_i (Y_i - (\beta_0 + \beta_1 \tilde{X}_i)) = -2\sum_i Y_i + 2n\beta_0 + 2\beta_1 \sum_i \tilde{X}_i = -2n(\bar{Y} - \beta_0 - \beta_1 \bar{\tilde{X}})$$

$\Rightarrow$ so when $\tilde{\beta}_0 = \bar{Y} - \tilde{\beta}_1 \bar{\tilde{X}}$, we have $\frac{\partial RSS}{\partial \beta_0} = 0$

$$\frac{\partial RSS}{\partial \beta_1} = -2\sum_i \tilde{X}_i (Y_i - \beta_0 - \beta_1 \tilde{X}_i) = -2\sum_i \tilde{X}_i (Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 \tilde{X}_i) = -2\sum_i \tilde{X}_i [(Y_i - \bar{Y}) - \tilde{\beta}_1 (\tilde{X}_i - \bar{\tilde{X}})]$$

$$= -2\sum_i (\tilde{X}_i - \bar{\tilde{X}})\left((Y_i - \bar{Y}) - \tilde{\beta}_1 (\tilde{X}_i - \bar{\tilde{X}})\right)$$

$\Rightarrow$ so when $\tilde{\beta}_1 = \frac{S_{\tilde{X}Y}}{S_{\tilde{X}\tilde{X}}}$, we have $\frac{\partial RSS}{\partial \beta_1} = 0$

$\Rightarrow$ so $\tilde{\beta}_0, \tilde{\beta}_1 = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 \tilde{X}_i))^2 \Rightarrow \tilde{\beta}_0, \tilde{\beta}_1$ are OLS estimators

5. $$\begin{cases} \bar{Y}_P = \hat{\beta}_0 + \hat{\beta}_1 \cdot \bar{X}_P = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 = \hat{\beta}_0 \\ \bar{Y}_T = \hat{\beta}_0 + \hat{\beta}_1 \cdot \bar{X}_T = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 = \hat{\beta}_0 + \hat{\beta}_1 \end{cases} \Rightarrow \begin{cases} \hat{\beta}_0 = \bar{Y}_P \\ \hat{\beta}_1 = \bar{Y}_T - \bar{Y}_P \end{cases}$$