Homework 15 Solution

STAT 509 Statistics for Engineers
Summer 2017 Section 001
Instructor: Tahmidul Islam

1. For the `teengamb` dataset, use R to calculate the 95% two-sided confidence interval and prediction interval for gamble when income is 2, and make a detailed interpretation about these two intervals by the context of the problem. Show me your R code and R output.

   The 95% confidence interval and prediction interval can be calculated by the following code

```
> library(faraway)
> data(teengamb)
> fit <- lm(gamble ~ income, data=teengamb)
> predict(fit, data.frame(income=2), confidence=0.95, interval="confidence")
       fit       lwr      upr
1 4.716412 -4.454029 13.88685
> predict(fit, data.frame(income=2), confidence=0.95, interval="prediction")
       fit       lwr      upr
1 4.716412 -46.36182 55.79464
```

   A 95% confidence interval is $(-4.45, 13.89)$. It means when the income is 2, we are 95% confident that the mean expenditure on gambling is less than 13.89 pounds per year. *(Remark: negative expenditure is impossible, which should be excluded from the interpretation.)*

   A 95% prediction interval is $(-46.36, 55.79)$. It means when the income is 2, we are 95% confident that the expenditure on gambling for one people in Britain is less than 55.79 pounds per year.

2. There is a `gala` dataset in faraway package. It concerns the number of species of tortoise on the various Galapagos Islands. There are 30 cases (Islands) and 7 variables in the dataset, including

   - **Species** The number of species of tortoise found on the island
   - **Endemics** The number of endemic species
   - **Elevation** The highest elevation of the island (m)
   - **Nearest** The distance from the nearest island (km)
   - **Scruz** The distance from Santa Cruz island (km)
   - **Adjacent** The area of the adjacent island (km^2)

   Fit a simple linear regression model with **Species** as response and **Elevation** as explanatory variable. Show me the output.

```
> fit <- lm(Species ~ Elevation, data=gala)
> summary(fit)

Call:
lm(formula = Species ~ Elevation, data = gala)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-218.319  -30.721  -14.690    4.634  259.180

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.33511   19.20529   0.590     0.56
Elevation    0.20079    0.03465   5.795 3.18e-06 ***
---

Residual standard error: 78.66 on 28 degrees of freedom
Multiple R-squared:  0.5454,    Adjusted R-squared:  0.5291
F-statistic: 33.59 on 1 and 28 DF,  p-value: 3.177e-06
```

(a) Calculate $\hat{Y}$ (a vector) and $\bar{Y}$ (a number).

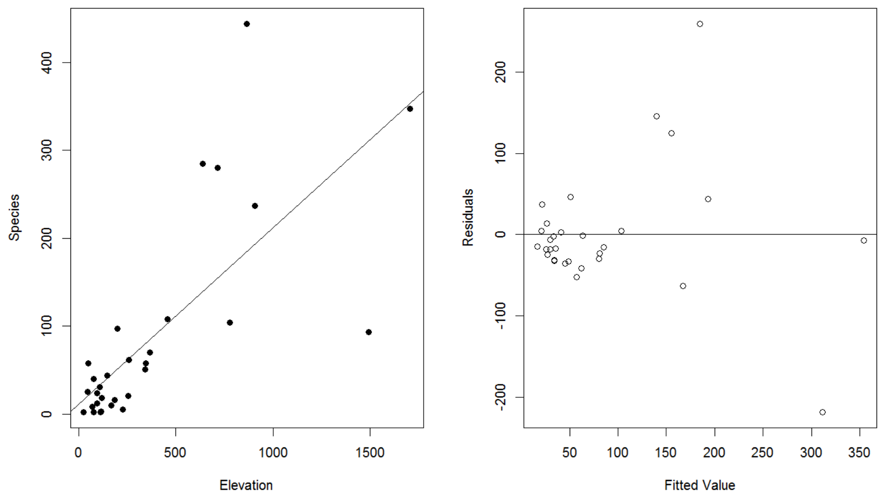
```
> yhat <- predict(fit)
> yhat
      Baltra    Bartolome     Caldwell     Champion      Coamano Daphne.Major
    80.80921     33.22146     34.22542     20.57155     26.79611     35.22938
Daphne.Minor       Darwin         Eden     Enderby     Espanola   Fernandina
    30.00879     45.06820     25.59136     33.82384     51.09197    311.31865
    Gardner1     Gardner2     Genovesa      Isabela     Marchena       Onslow
    21.17393     56.91494     26.59532    354.08739     80.20684     16.35492
       Pinta      Pinzon   Las.Plazas       Rabida SanCristobal  SanSalvador
   167.35065    103.29794     30.20958     85.02585    155.10232    193.25284
   SantaCruz      SantaFe   SantaMaria      Seymour      Tortuga         Wolf
   184.81957     63.34029    139.84212     40.85157     48.68246     62.13554
> ybar <- mean(gala$Species)
> ybar
[1] 85.23333
```

(b) Calculate SSTO and SSE.

```
> SSTO <- sum((gala$Species - ybar)^2)
> SSTO
[1] 381081.4
> SSE <- sum((gala$Species - yhat)^2)
> SSE
[1] 173253.9
```

(c) Draw the scatter plot (with the regression line) and residual plot. Do you think the equal variance assumption holds?

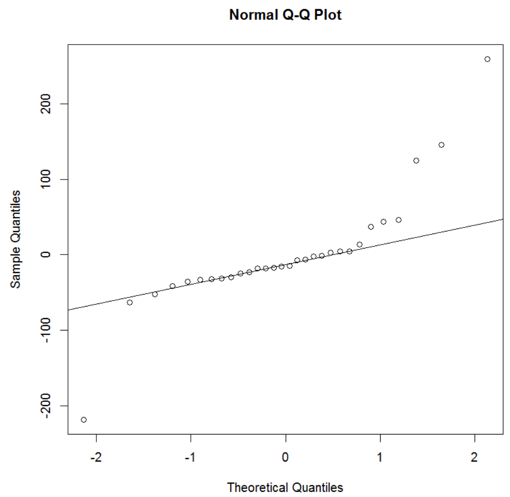
Both scatter plot and the residual plot indicate that the variance of the error term ϵ increases as the fitted value ($\hat{Y}$) increases. The equal variance assumption clearly breaks.




```
par(mfrow=c(1,2))
plot(gala$Elevation, gala$Species, xlab="Elevation", ylab="Species", pch=16)
abline(fit)
plot(yhat, residuals(fit), xlab="Fitted Value", ylab="Residuals")
abline(h=0)
```

(d) Use qq plot to check whether the normality assumption holds.

It is clear that the tail part of the qq plot doesn't pass the fat-pencil test. Therefore, we suspect the normality assumption doesn't hold perfectly here.

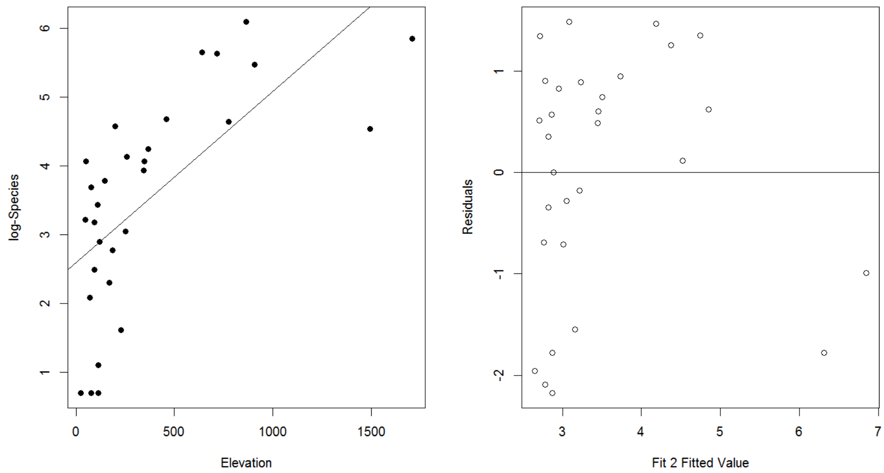



```
qqnorm(residuals(fit))
qqline(residuals(fit))
```

(e) Re-fit the model with the transformation $\log Y$, and draw the scatter plot, residual plot, and qq plot. Make comments to each plot. Does the transformation make your model better?

After transformation, the scatter plot looks better in the way that not all points are concentrated in the corner (meaning extreme large values of Species are relatively smaller due to log transformation.) and magnitude of the variance is more similar.

From the residual plot, we can confirm this observation. Overall, the model is better than the original one.



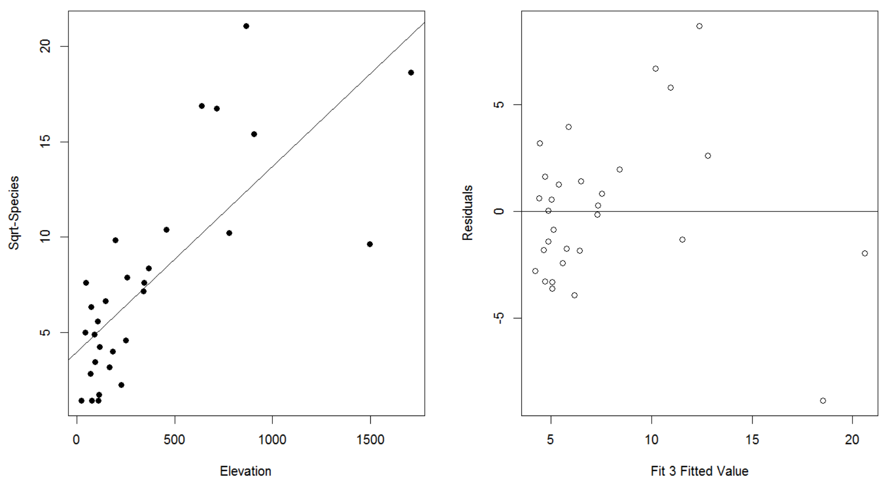

```
> fit2 <- lm(log(Species) ~ Elevation, data=gala)
> summary(fit2)
(Intercept) 2.5913986  0.2879198   9.000 9.33e-10 ***
Elevation   0.0024895  0.0005194   4.793 4.88e-05 ***

Residual standard error: 1.179 on 28 degrees of freedom
Multiple R-squared:  0.4507,    Adjusted R-squared:  0.4311
F-statistic: 22.97 on 1 and 28 DF,  p-value: 4.885e-05

par(mfrow=c(1,2))
plot(gala$Elevation, log(gala$Species), xlab="Elevation",
  ylab="log-Species", pch=16)
abline(fit2)
plot(predict(fit2), residuals(fit2), xlab="Fit 2 Fitted Value", ylab="Residuals
abline(h=0)
```

(f) Re-fit the model with the transformation $\sqrt{Y}$, and draw the scatter plot, residual plot, and qq plot. Make comments to each plot. Does the transformation make your model better?

The megaphone shape of the variance still exist observing from the scatter plot and residual plot. It means the variance goes large when the fitted value goes large. The squre root transformation doesn't make the model any better.




```
> fit3 <- lm(sqrt(Species) ~ Elevation, data=gala)
> summary(fit3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.954656   0.874615   4.522 0.000102 ***
Elevation   0.009753   0.001578   6.181 1.13e-06 ***

Residual standard error: 3.582 on 28 degrees of freedom
Multiple R-squared:  0.5771,    Adjusted R-squared:  0.562
F-statistic: 38.21 on 1 and 28 DF,  p-value: 1.125e-06

par(mfrow=c(1,2))
plot(gala$Elevation, sqrt(gala$Species), xlab="Elevation",
  ylab="Sqrt-Species", pch=16)
abline(fit3)
plot(predict(fit3), residuals(fit3), xlab="Fit 3 Fitted Value", ylab="Residuals
abline(h=0)
```

(g) Compare the coefficient of determination in the original regression model and the model with $\sqrt{Y}$ transformation. Make comments.

From `summary(fit)` we find the coefficient of determination in the original model is 0.5454, and from `summary(fit3)`, the one in sqare root transferred model is 0.5771. Even though the sqre root transformation doesn't solve the unequal variance assumption problem, it slightly increases the R^2. The interpretation for the transferred model is: Elevation explains the 57.74% variability of the $\sqrt{\text{Species}}$.

Note: if you have problem loading faraway package, download the `gala` dataset from the course webpage and save it in D drive. Run the following code to load.

```
gala <- read.table("D:/galadata.txt", sep="\t")
```