HW3

Getong Zhong

2022-10-28

Problem 5

(a)

The analyst's conclusion is made basically on the general hypothesis test and the coefficient of determination. In most of the cases, that's far not enough to give a thorough and unbiased conclusion to determine whether the model is valid or not. We should also consider the assumptions of regression while analysis the models.

(b)

First, they are a lot of bad leverage points (outliers) that could be removed. Second, variance are not random distributed, a common way to fix this is to redefine the dependent variable, for example use a rate for the dependent variable, rather than the raw value, this will be helpful especially when categorical variable are included. Third, the residuals are not follow a normal distribution, we probably need to fit and tranform the model in to a different one.

(c)

Yes, the model (3.11) is a better model than model (3.10). First, model (3.11) has less ourliers than model (3.10). Second variance in model (3.11) are more random-distributed than variance in model (3.10), and no apparent pattern. Third, the standard residuals in model (3.11) are more fitted to the normal distributed line than the standard residuals in model (3.10).

(d)

The intercept ($\beta 0$) is -61.904248 and the estimator for $\beta 1$ is 1.088841 which means

$$log(y) = -61.904248 + 1.088841 * x_i + e$$

that is,

$$y = e^{-61.904248} + e^{1.088841} * x_i + e$$

(e)

First, the p value of intercept show it is not statistically significant. Second, the standard residuals not are strictly fitted to the normal distribution. Third, there are still outliers could be potentially removed.

Problem 6

Although e is normally distributed, x is highly skewed distributed, therefore (λ) is not 0 in this case.

Problem 7

Let y be the random variable with mean (μ) and variance g(μ) we are trying to find function h(y) such that variance of h(y) is constant

$$h(y) = h(\mu) + h'(\mu)(y - \mu)$$

let var(h(y)) equals to constant c,

$$Var(h(y)) = [h'(\mu)]^2 Var(y) = [h'(\mu)]^2 g(\mu) = c$$

if g(μ) = (μ) then

$$h'(\mu) = \sqrt{\frac{c}{\mu}}, h(\mu) = \sqrt{c} * (-\frac{1}{2}) * \sqrt{\mu}$$

if g(μ) = (μ^2) then

$$h'(\mu) = \sqrt{\frac{c}{\mu}}, h(\mu) = \sqrt{c} * log(\mu)$$

Problem 8

Part 1

```
diamond <- read.table("C:/Users/tonyg/Desktop/Academic/Grad/HUDM 5126/diamonds.txt", header = TRUE)
mod1<-lm(Price ~ Size, data = diamond)
summary(mod1)
```

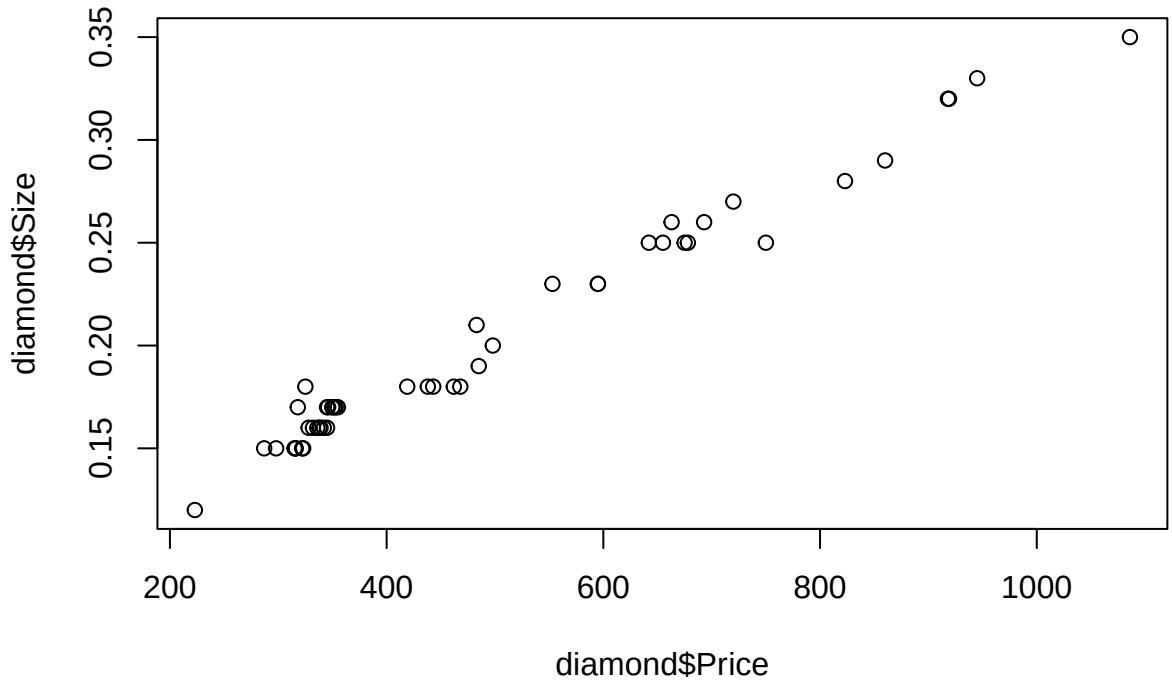
(a)

```
##
## Call:
## lm(formula = Price ~ Size, data = diamond)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.654 -21.503  -1.203  16.797  79.295
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -258.05      16.94  -15.23   <2e-16 ***
## Size         3715.02      80.41   46.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.6 on 47 degrees of freedom
## Multiple R-squared:  0.9785, Adjusted R-squared:  0.978
## F-statistic:  2135 on 1 and 47 DF,  p-value: < 2.2e-16
```

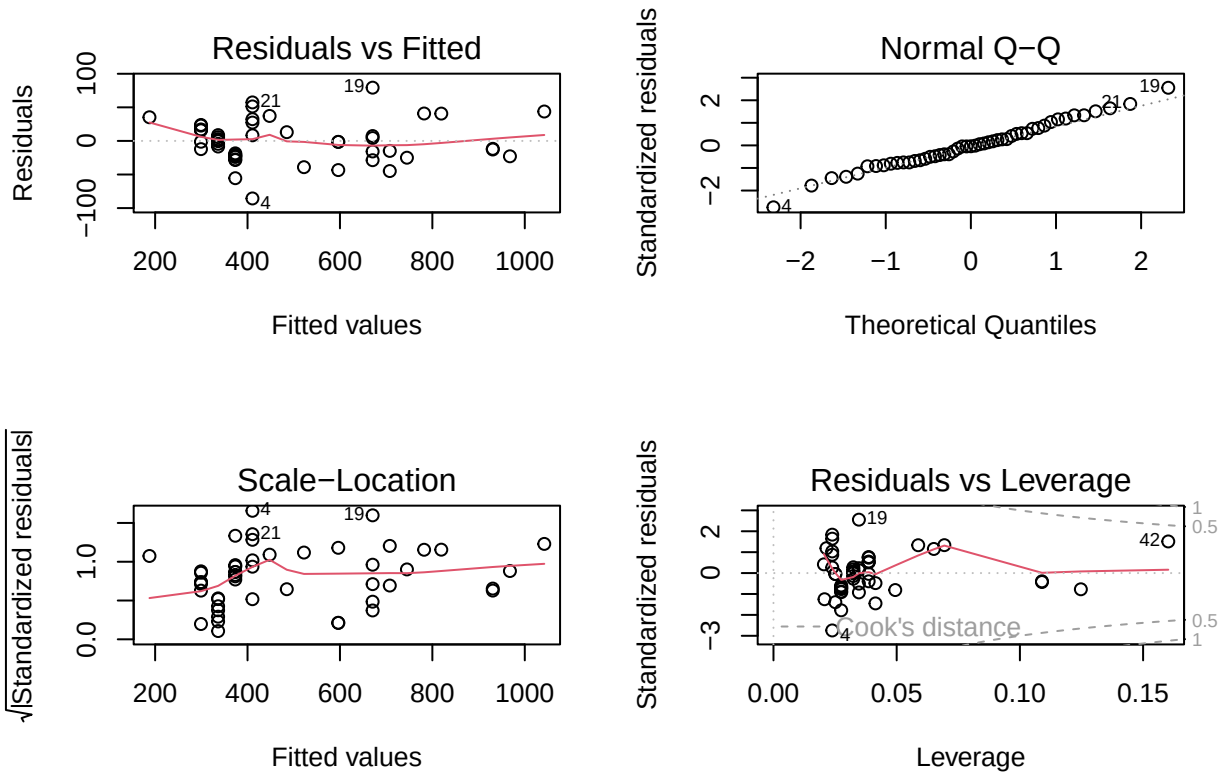
Price = -258.05+3715.02*size

(b) There are several outliers that could be potentially removed. Variance more not strictly non-patterned.

```
plot(diamond$Price,diamond$Size)
```




```
par(mfrow = c(2, 2))
plot(mod1)
```



Part 2

```
crPrice <- (diamond$Price) ^ (1/3)
crSize <- (diamond$Size) ^ (1/3)
mod2<-lm(crPrice ~ crSize)
summary(mod2)
```

(a)

```
##
## Call:
## lm(formula = crPrice ~ crSize)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52630 -0.11993  0.00142  0.08128  0.36974
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.9551     0.2785  -14.20   <2e-16 ***
## crSize       20.1139     0.4757   42.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.1708 on 47 degrees of freedom
## Multiple R-squared:  0.9744, Adjusted R-squared:  0.9738
## F-statistic:  1788 on 1 and 47 DF,  p-value: < 2.2e-16
```

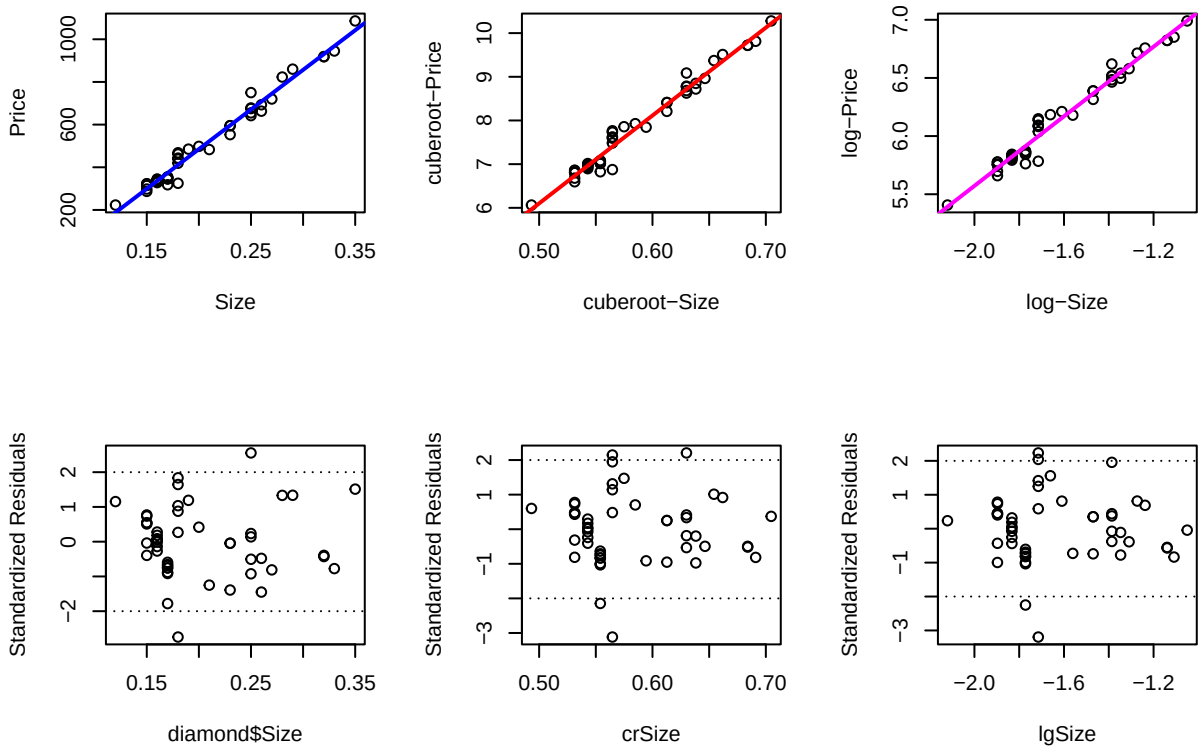
```
lgPrice <- log(diamond$Price)
lgSize <- log(diamond$Size)
mod3<-lm(lgPrice ~ lgSize)
summary(mod3)
```

```
##
## Call:
## lm(formula = lgPrice ~ lgSize)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21460 -0.04646 -0.00274  0.03001  0.15005
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.56317    0.06221  137.65   <2e-16 ***
## lgSize       1.49566    0.03772   39.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06796 on 47 degrees of freedom
## Multiple R-squared:  0.971,  Adjusted R-squared:  0.9704
## F-statistic:  1572 on 1 and 47 DF,  p-value: < 2.2e-16
```

based on the R-square, mod2 (cuberoot) has a higher R-square than mod3 (Log). From the graph below, we find they have very similar behavior in outliers but mod2 fit to the regression line a little bit better than mod3 thus the model after transformation we selected is mod2.

```
par(mfrow = c(2, 3))
plot( diamond$Size,diamond$Price,
ylab = "Price", xlab = "Size")
abline(coef(mod1), lwd = 2, col = "blue")
plot(crSize, crPrice,
ylab = "cuberoot-Price", xlab = "cuberoot-Size")
abline(coef(mod2), lwd = 2, col = "red")
plot(lgSize, lgPrice,
ylab = "log-Price", xlab = "log-Size")
abline(coef(mod3), lwd = 2, col = "magenta")
plot(diamond$Size, rstandard(mod1),
ylab = "Standardized Residuals")
abline(h = 2, lty = 3)
abline(h = -2, lty = 3)
plot(crSize, rstandard(mod2),
ylab = "Standardized Residuals")
abline(h = 2, lty = 3)
abline(h = -2, lty = 3)
plot(lgSize, rstandard(mod3),
ylab = "Standardized Residuals")
```

```
abline(h = 2, lty = 3)
abline(h = -2, lty = 3)
```



(b) The weakness of mod2 is there are a few of outliers so that some of the points are not perfectly fit to the regression line as shown.

Part 3

I will choose model from part A over model from part B. First we can see that model from part A has a higher R-squared value than model from part B. Second from the above graphs we can the that after transformation of the both variables, there are more outliers in mod2 than mod1. Third, from the fitted line we can see that points in mod1 fitted to the regression line better than mod2. Therefore, I will choose the model in Part A as a better model.