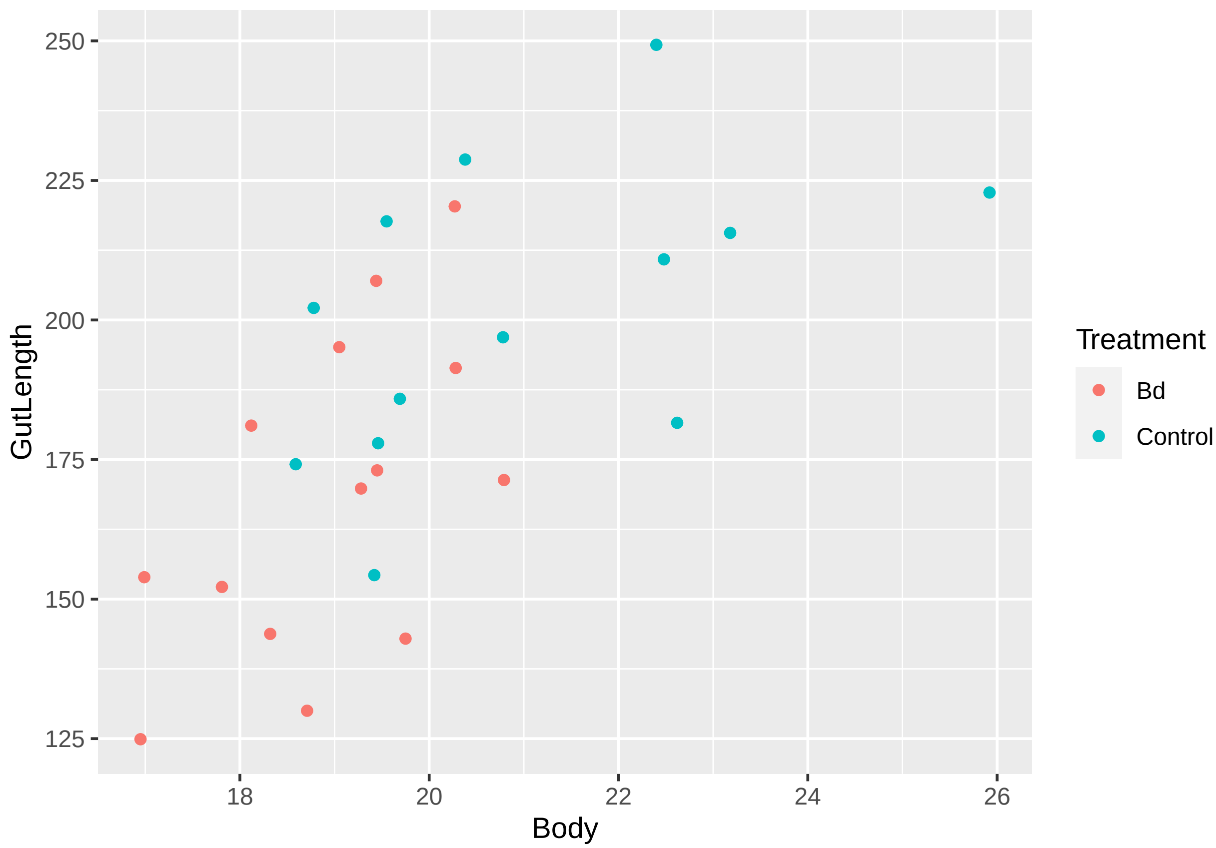
PSET3

Peter

2023-09-20

```r
#3.12a: X1=BodySize, X2 = Ifgp, Y = Mrate
#Mrate = B0 + B1 * BodySize + B2 * Ifgp + epsilon
#Free growth: Mrate = B0 + B1 * BodySize + B2 * Ifgp (1) + epsilon
#No free growth: Mrate = B0 + B1 * BodySize + epsilon

#3.12b: As seen above...
#Mrate = B0 + B1 * BodySize + B2 * Ifgp + epsilon

#3.12c:
#Full model: Mrate = B0 + B1 * BodySize + B2 * Ifgp + B3 * (BodySize * Ifgp) + epsilon
#Reduced model: Mrate = B0 + B1 *BodySize + epsilon
#The nested F Test would compare the SSE for the models, if the full model had less
#error that the reduced then two regression lines are probably necessary, but if the
#error was similar then the reduced model is probably fine.
```

```r
#3.26
library(Stat2Data)
library(ggplot2)
data("Tadpoles")
#3.26a: The plot seems to be somewhat positively linear, the control points have a
#larger gutlength and longer body on average than Bd.
ggplot(Tadpoles, aes(Body, GutLength, color = Treatment)) +
  geom_point()
```




```r
#3.26b: REGRESSION LINE: GutLength = -20.764 + 10.280 * Body
#Ho: B1 = 0, Ha: B1 != 0
#With a t value of 4.204, a pvalue of .000293<0.05, 25 df, and an r^2 of .4141,
#we have significant evidence to reject the null hypothesis. We can conclude that
#there is a significant positive linear relationship between Body length (mm) and
#gut length (mm) in tadpoles.
Gutbod <- lm(GutLength~Body, data = Tadpoles)
summary(Gutbod)
```

```
##
## Call:
## lm(formula = GutLength ~ Body, data = Tadpoles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.575 -22.245   0.027  17.815  39.998
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -20.764     49.002  -0.424 0.675384
## Body          10.280      2.445   4.204 0.000293 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.83 on 25 degrees of freedom
## Multiple R-squared:  0.4141, Adjusted R-squared:  0.3907
```

```
## F-statistic: 17.67 on 1 and 25 DF,  p-value: 0.0002931
```

```r
#3.26c:
as.factor(Tadpoles$Treatment)
```

```
##  [1] Bd      Bd      Bd      Bd      Bd      Bd      Bd      Bd      Bd
## [10] Bd      Bd      Bd      Bd      Bd      Control Control Control Control
## [19] Control Control Control Control Control Control Control Control Control
## Levels: Bd Control
```

```r
Gut_treat_bod <- lm(GutLength~Body+Treatment, data = Tadpoles)
summary(Gut_treat_bod)
```

```
##
## Call:
## lm(formula = GutLength ~ Body + Treatment, data = Tadpoles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.462 -15.006  -2.545  19.117  41.298
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         15.47      53.81   0.288  0.77618
## Body                 8.07       2.82   2.862  0.00859 **
## TreatmentControl    16.28      11.03   1.476  0.15286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.26 on 24 degrees of freedom
## Multiple R-squared:  0.4629, Adjusted R-squared:  0.4181
## F-statistic: 10.34 on 2 and 24 DF,  p-value: 0.0005762
```

```r
#as.numeric(Tadpoles$Treatment)
as.factor(Tadpoles$Treatment)
```

```
##  [1] Bd      Bd      Bd      Bd      Bd      Bd      Bd      Bd      Bd
## [10] Bd      Bd      Bd      Bd      Bd      Control Control Control Control
## [19] Control Control Control Control Control Control Control Control Control
## Levels: Bd Control
```

```r
#Regression lines: Bd = 1, control = 2
#Bd Model: GutLength = B0 + B1 * Body + B2 * Treatment (1, so just + B2) + epsilon
#Bd Reg Line: GutLength = 15.47 + 8.07 * Body + 16.28
#Control Model: GutLength = B0 + B1 * Body + B2 * Treatment + epsilon
#Control Reg Line: GutLength = 15.47 + 8.07 * Body + 16.28 * Treatment


#3.26d:
g_t_m_bod <- lm(GutLength~Body+Treatment+MouthpartDamage, data = Tadpoles)
summary(g_t_m_bod)
```

```
##
## Call:
## lm(formula = GutLength ~ Body + Treatment + MouthpartDamage,
##     data = Tadpoles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.422 -17.701  -6.771  16.338  40.877
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -20.258     53.070  -0.382   0.7062
## Body                  6.442      2.746   2.346   0.0280 *
## TreatmentControl     25.412     11.177   2.274   0.0326 *
## MouthpartDamage      96.839     45.839   2.113   0.0457 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.68 on 23 degrees of freedom
## Multiple R-squared:  0.5502, Adjusted R-squared:  0.4915
## F-statistic: 9.378 on 3 and 23 DF,  p-value: 0.0003092
```

```r
#Mouthpart model: GutLength = B0 + B1 * Body + B2 * Treatment + B3
#* MouthpartDamage + epsilon
#Regression: GutLength = -20.258 + 6.442 * body + 25.412 * treatment (1 or 2)
#+ 96.839 * MouthpartDamage
#More MouthpartDamage means a reduced food consumption. This could confound the
#variables so it is important to account for this in our model. The R^2 and pvalues
#are fairly solid here.
```

```r
#3.38:
data("Diamonds")

#3.38a: R^2 = .04748, adjusted R^2 = .042, t value for depth = .229 with a pvalue
#of .819>0.05 slope = 766.369, t value for depth^2 is -.130 with a p value of .897>0.05
#slope = -3.233
modle <- lm(TotalPrice~Depth + I(Depth^2), data = Diamonds)
summary(modle)
```

```
##
## Call:
## lm(formula = TotalPrice ~ Depth + I(Depth^2), data = Diamonds)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9323  -4251  -2676   2134  45513
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28406.783 112211.790  -0.253    0.800
## Depth          766.369   3353.222   0.229    0.819
## I(Depth^2)      -3.233     24.869  -0.130    0.897
##
```

```
## Residual standard error: 7616 on 348 degrees of freedom
## Multiple R-squared:  0.04748,    Adjusted R-squared:  0.042
## F-statistic: 8.673 on 2 and 348 DF,  p-value: 0.0002111
```

```r
#3.38b: R^2 = 0.8704, adjusted R^2 = 0.8696, tvalue for carat = 47.006 with a p value
#of 0<0.05 and slope = 15087.01, for depth the tvalue = -4.364 with a pvalue of 0<0.05
#and slope = -134.94.
modle2 <- lm(TotalPrice~Carat+Depth, data = Diamonds)
summary(modle2)
```

```
##
## Call:
## lm(formula = TotalPrice ~ Carat + Depth, data = Diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9234.7 -1223.7  -274.3  1161.0 16368.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1059.24    1918.36   0.552    0.581
## Carat       15087.01     320.96  47.006  < 2e-16 ***
## Depth        -134.94      30.92  -4.364 1.68e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2809 on 348 degrees of freedom
## Multiple R-squared:  0.8704, Adjusted R-squared:  0.8696
## F-statistic:  1168 on 2 and 348 DF,  p-value: < 2.2e-16
```

```r
#3.38c: R^2 = 0.89, adjusted R^2 = 0.889
#Carat: slope = -11827.73, t = -3.442, pval = 00.000648 < 0.05
#Depth: slope = -598.18, t = -9.137, pval = < 2e-16 < 0.05
#interaction term: slope = 408.45, t = 7.861, pval = 4.84e-14 < 0.05.
modle3 <- lm(TotalPrice~Carat*Depth, data = Diamonds)
summary(modle3)
```

```
##
## Call:
## lm(formula = TotalPrice ~ Carat * Depth, data = Diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8254.4 -1311.5  -157.2  1131.8 14513.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31171.41    4219.58   7.387 1.13e-12 ***
## Carat       -11827.73    3436.47  -3.442 0.000648 ***
## Depth         -598.18      65.47  -9.137  < 2e-16 ***
## Carat:Depth    408.45      51.96   7.861 4.84e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2592 on 347 degrees of freedom
## Multiple R-squared:   0.89,  Adjusted R-squared:  0.889
## F-statistic: 935.7 on 3 and 347 DF,  p-value: < 2.2e-16
```

```r
#3.38d: R^2 = 0.9313, adjusted R^2 = 0.9304
#Carat: slope = 7573.620, t = 2.491, pval = 0.0132<0.05
#Depth: slope = -728.700, t = -0.806, pval = 0.4210>0.05
#Carat^2: slope = 4761.592, t = 14.418, pval = <2e-16<0.05
#Depth^2: slope = 5.276, t = 0.784, pval = 0.4333>0.05
#Carat:Depth: slope = -83.891, t = -1.567, pval = 0.1180>0.05
modle4 <- lm(TotalPrice~Carat*Depth + I(Carat^2) + I(Depth^2), data = Diamonds)
summary(modle4)
```

```
##
## Call:
## lm(formula = TotalPrice ~ Carat * Depth + I(Carat^2) + I(Depth^2),
##     data = Diamonds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12196.1   -652.7    -38.5    485.7  10582.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24338.820  30297.912   0.803   0.4223
## Carat        7573.620   3040.787   2.491   0.0132 *
## Depth        -728.700    904.439  -0.806   0.4210
## I(Carat^2)   4761.592    330.246  14.418   <2e-16 ***
## I(Depth^2)      5.276      6.727   0.784   0.4333
## Carat:Depth   -83.891     53.530  -1.567   0.1180
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2053 on 345 degrees of freedom
## Multiple R-squared:  0.9313, Adjusted R-squared:  0.9304
## F-statistic: 936.1 on 5 and 345 DF,  p-value: < 2.2e-16
```

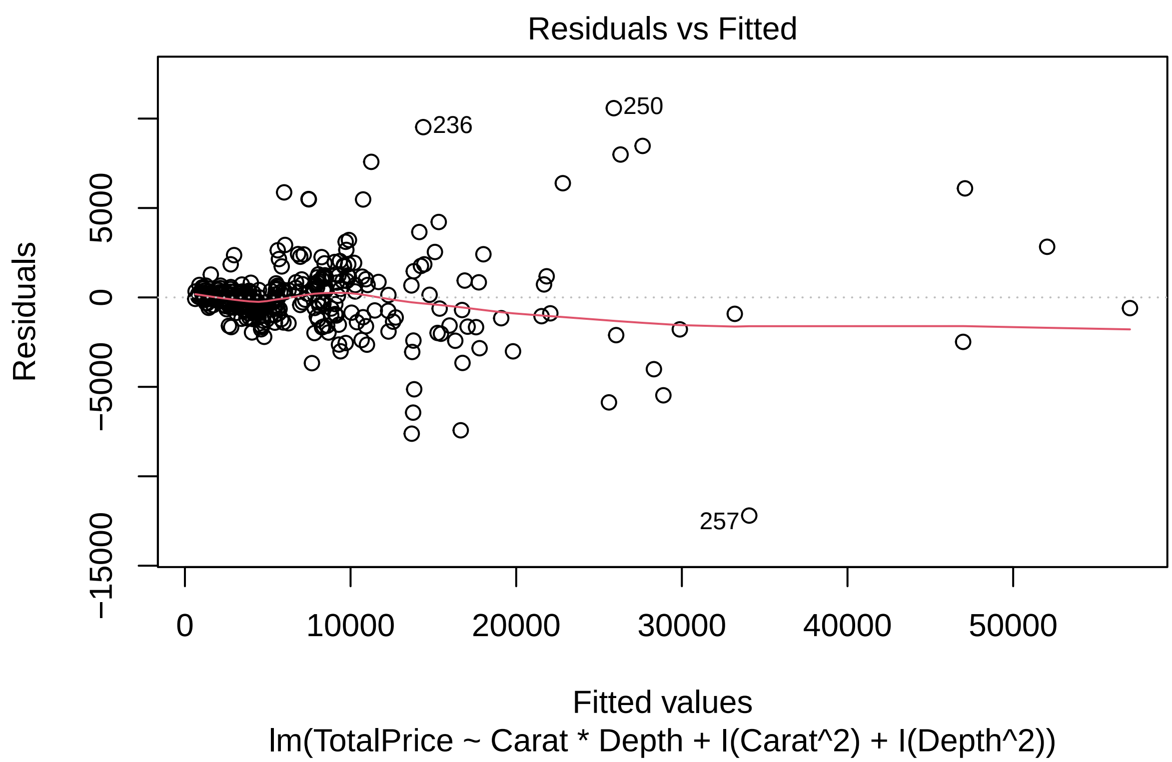
```r
#I think that the model in part d is the best fit. With the quadratic formula for
#carat and depth with the interaction term.
```

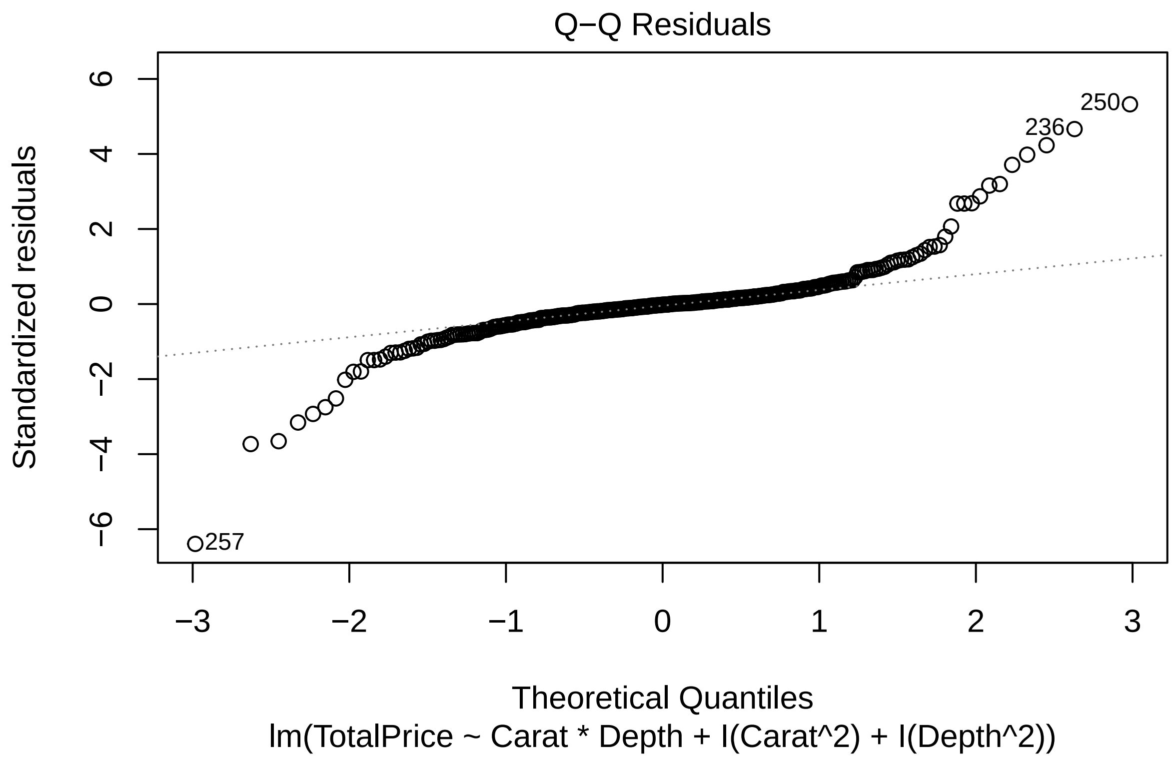
```r
#3.39a:
modle4 <- lm(TotalPrice~Carat*Depth + I(Carat^2) + I(Depth^2), data = Diamonds)
summary(modle4)
```

```
##
## Call:
## lm(formula = TotalPrice ~ Carat * Depth + I(Carat^2) + I(Depth^2),
##     data = Diamonds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -12196.1   -652.7    -38.5    485.7  10582.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24338.820  30297.912   0.803   0.4223
## Carat        7573.620   3040.787   2.491   0.0132 *
## Depth        -728.700    904.439  -0.806   0.4210
## I(Carat^2)   4761.592    330.246  14.418   <2e-16 ***
## I(Depth^2)      5.276      6.727   0.784   0.4333
## Carat:Depth   -83.891     53.530  -1.567   0.1180
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2053 on 345 degrees of freedom
## Multiple R-squared:  0.9313, Adjusted R-squared:  0.9304
## F-statistic: 936.1 on 5 and 345 DF,  p-value: < 2.2e-16
```

```r
plot(modle4, 1:2)
```



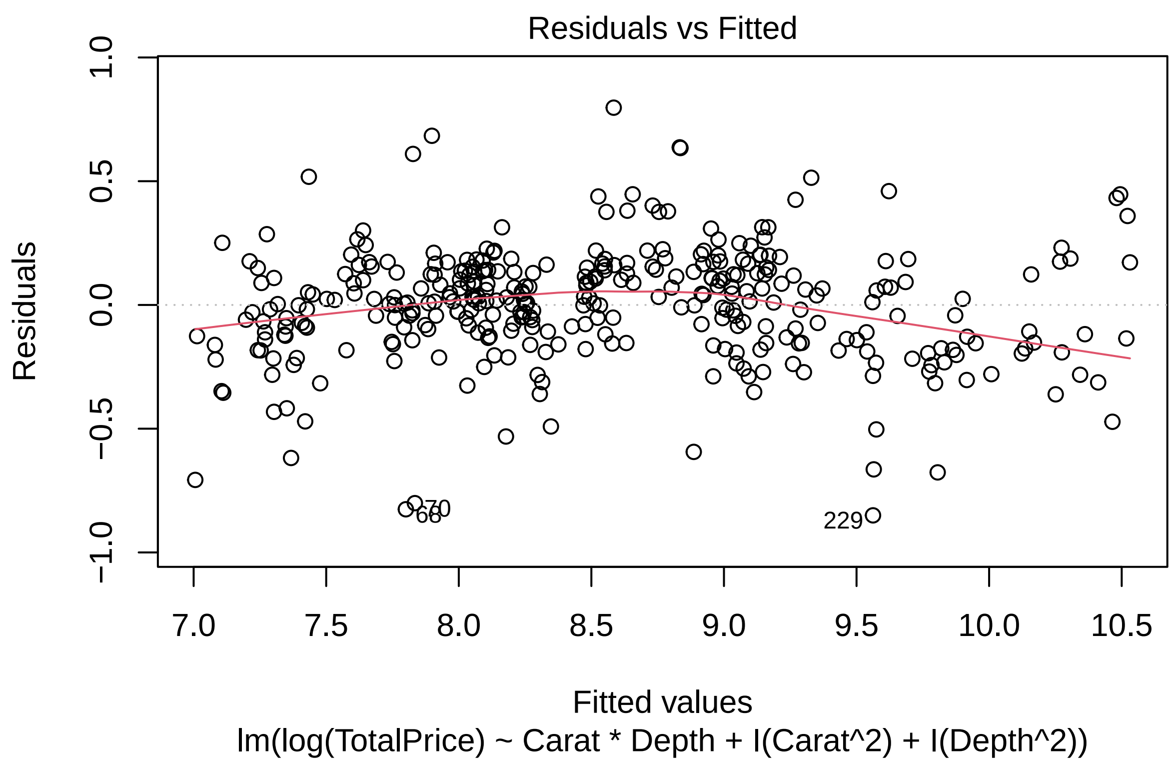


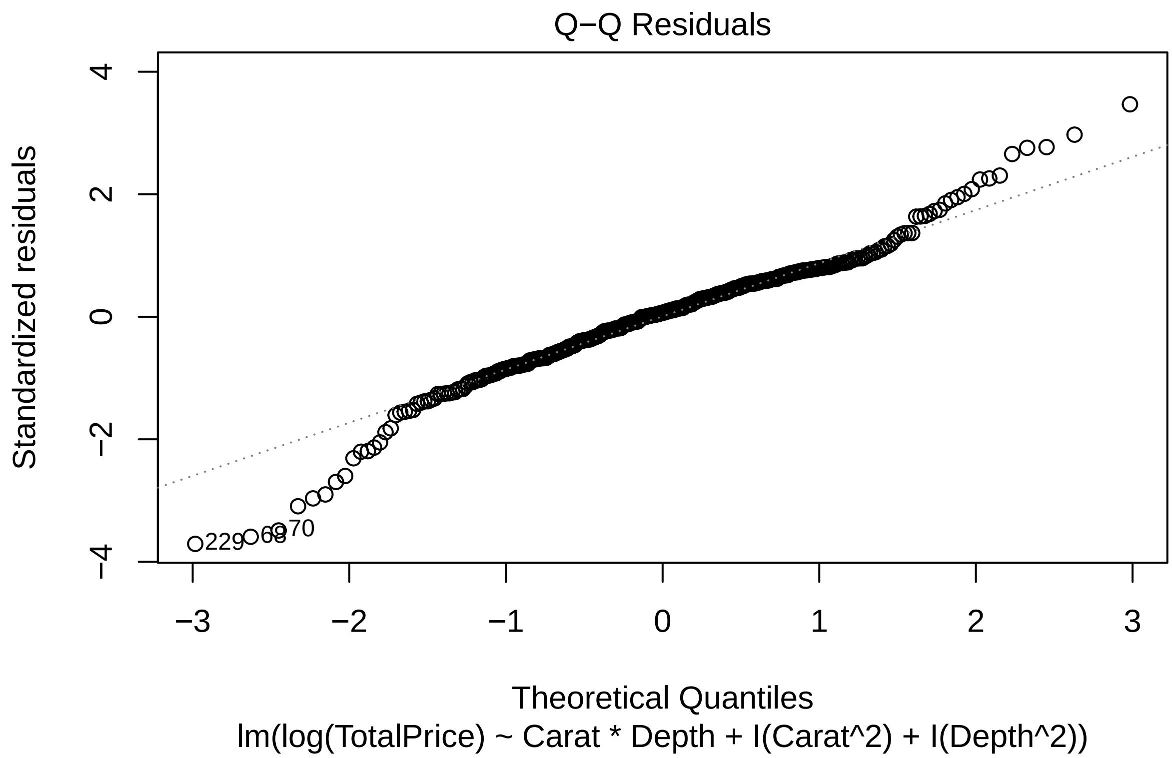

```r
#constant variance doesn't look great according to the resids vs fitted graph.
#There is a heavy weighting on the left side of the plot, and the variance is
#much greater in the middle.
#the QQ plot shows a very non-linear fit, telling us that normality is not met either.

#3.39b: This model still looks fairly solid. Our pvalues for certain predictors
#like depth, depth^2 and the interaction term are worse than before, but only
#depth^2 and carat:depth are no longer significant. The adjusted R^2 value is still
#high, at 0.9292. The t values are slightly worse, but still solid. This is still
#very reasonable for being a good predictor of log(TotalPrice)
modle5 <- lm(log(TotalPrice)~Carat*Depth + I(Carat^2) + I(Depth^2), data = Diamonds)
summary(modle5)
```

```
##
## Call:
## lm(formula = log(TotalPrice) ~ Carat * Depth + I(Carat^2) + I(Depth^2),
##     data = Diamonds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.85021 -0.13209  0.01441  0.13613  0.79710
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.5049624  3.4020467   3.970 8.76e-05 ***
```

```
## Carat        2.5863485  0.3414393   7.575 3.33e-13 ***
## Depth       -0.2027689  0.1015563  -1.997   0.0467 *
## I(Carat^2)  -0.5714071  0.0370821 -15.409  < 2e-16 ***
## I(Depth^2)   0.0013384  0.0007553   1.772   0.0773 .
## Carat:Depth  0.0095943  0.0060107   1.596   0.1114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2306 on 345 degrees of freedom
## Multiple R-squared:  0.9302, Adjusted R-squared:  0.9292
## F-statistic: 919.9 on 5 and 345 DF,  p-value: < 2.2e-16
```

```r
#3.39c:
plot(modle5, 1:2)
```






```r
#log(TotalPrice) looks much better for the residual plots. Constant variance looks
#great, the data is very evenly dispersed on the resids vs fits plot and our QQ plot
#has less of a pattern, looks much more normal.
```

```r
#3.46:
model2 <- lm(TotalPrice~Carat*Depth + I(Carat^2) + I(Depth^2), data = Diamonds)
#model from 3.38d

model1 <- lm(TotalPrice ~ Carat + Depth, data = Diamonds) #simpler model to
#compare against

anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: TotalPrice ~ Carat + Depth
## Model 2: TotalPrice ~ Carat * Depth + I(Carat^2) + I(Depth^2)
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    348 2746468485
## 2    345 1454702094  3 1291766391 102.12 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#model1 has 348 df, model2 has 345.
#For the Residuals sums of squared, model 2 has a lower RSS value, and therefore
#explains more of the variance in the response variable. So model 2 has a better
#fit in terms of RSS.
#The f stat is 102.12, which is high in this context. So model 2 explains alot
#more of the variance than model 1 does.
#The pval = 0<0.05, which means that the difference in fit btwn the two models
#IS significant.

#The ANOVA table suggests that Model 2, full model with interaction and quadratic
#terms, provides a significantly better fit to the data compared to the reduced
#Model 1. Given the extremely low p-value, there is very strong evidence that the
#additional terms in Model 2 help explain more of the variance in TotalPrice.
```