STA 471 - Homework 7

Richard McCormick

2023-11-20

2. Problem H - Develop a suitable fitted equation using these data and compare its form with the form of the one fitted by Anderson and Bancroft.

```
X1 <- c( 1.76, 1.55, 2.73, 2.73, 2.56, 2.80, 2.80, 1.84, 2.16, 1.98, 0.59, 0.80, 0.80, 1.05, 1.80, 1.80

X2 <- c( 0.070, 0.070, 0.070, 0.070, 0.070, 0.070, 0.070, 0.070, 0.070, 0.020, 0.020, 0.020, 0.020, 0.0

X3 <- c( 7.8, 8.9, 8.9, 7.2, 8.4, 8.7, 7.4, 8.7, 8.8, 7.6, 6.5, 6.7, 6.2, 7.0, 7.3, 6.5, 7.6, 8.2, 7.6,

Y <- c( 110.4, 102.8, 101.0, 108.4, 100.7, 100.3, 102.0, 93.7, 98.9, 96.6, 99.4, 96.2, 99.0, 88.4, 75.3

turnip.data <- data.frame( X1, X2, X3, Y )

anderson.bancroft.model <- lm( Y ~ X1 + X2 + X3 + I(X1*X2), data=turnip.data )
summary( anderson.bancroft.model )
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + I(X1 * X2), data = turnip.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.155  -3.770   1.458   5.503  16.983
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   82.173     20.338   4.040 0.000547 ***
## X1             2.463      4.722   0.522 0.607190
## X2           -75.378     39.144  -1.926 0.067168 .
## X3             1.584      3.122   0.507 0.616997
## I(X1 * X2)    -1.374     21.265  -0.065 0.949058
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.1 on 22 degrees of freedom
## Multiple R-squared:  0.7549, Adjusted R-squared:  0.7103
## F-statistic: 16.94 on 4 and 22 DF,  p-value: 1.784e-06
```

```r
model.1 <- lm( Y ~ X1 + X2 + X3 + I(X1*X2) + I(X1^2) + I(X2^2) + I(X3^2) + I(X1^2 * X2^2), data=turnip.d
summary( model.1 )
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + I(X1 * X2) + I(X1^2) + I(X2^2) +
##     I(X3^2) + I(X1^2 * X2^2), data = turnip.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2210  -3.0367  -0.4704   1.5922  13.3598
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      130.9966   156.4226   0.837  0.41332
## X1               -15.1051    13.6463  -1.107  0.28291
## X2               399.9180   199.7515   2.002  0.06058 .
## X3                -5.8891    42.1669  -0.140  0.89048
## I(X1 * X2)        64.0103    85.3727   0.750  0.46308
## I(X1^2)            2.3519     3.6191   0.650  0.52400
## I(X2^2)        -1034.4271   318.6840  -3.246  0.00449 **
## I(X3^2)            0.1358     2.7160   0.050  0.96066
## I(X1^2 * X2^2)   -33.7152    57.4428  -0.587  0.56454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.427 on 18 degrees of freedom
## Multiple R-squared:  0.9187, Adjusted R-squared:  0.8826
## F-statistic: 25.44 on 8 and 18 DF,  p-value: 2.704e-08
```

```r
model.2 <- lm( Y ~ X1 + X2 + X3 + I(X1*X2) + I(X1^2) + I(X2^2) + I(X1^2 * X2^2), data=turnip.data )
summary( model.2 )
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + I(X1 * X2) + I(X1^2) + I(X2^2) +
##     I(X1^2 * X2^2), data = turnip.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2535  -3.0717  -0.4978   1.6649  13.3255
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      123.209     14.432   8.537  6.3e-08 ***
## X1               -15.372     12.222  -1.258  0.22371
## X2               400.105    194.403   2.058  0.05356 .
## X3                -3.784      2.421  -1.563  0.13452
## I(X1 * X2)        64.575     82.371   0.784  0.44273
## I(X1^2)            2.398      3.407   0.704  0.49011
## I(X2^2)        -1036.417    307.778  -3.367  0.00323 **
## I(X1^2 * X2^2)   -33.892     55.809  -0.607  0.55085
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.256 on 19 degrees of freedom
## Multiple R-squared:  0.9187, Adjusted R-squared:  0.8888
## F-statistic: 30.69 on 7 and 19 DF,  p-value: 4.63e-09
```

```
model.3 <- lm( Y ~ X1 + X2 + X3 + I(X1*X2) + I(X2^2), data=turnip.data )
summary( model.3 )
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + I(X1 * X2) + I(X2^2), data = turnip.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4000  -2.9165  -0.1146   2.5611  13.0970
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   119.599     13.723   8.715 2.03e-08 ***
## X1             -5.145      3.114  -1.652    0.113
## X2            536.508    101.078   5.308 2.91e-05 ***
## X3             -4.700      2.146  -2.190    0.040 *
## I(X1 * X2)     12.064     13.078   0.922    0.367
## I(X2^2)     -1239.333    198.997  -6.228 3.54e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.125 on 21 degrees of freedom
## Multiple R-squared:  0.9139, Adjusted R-squared:  0.8934
## F-statistic: 44.59 on 5 and 21 DF,  p-value: 1.757e-10
```

```
model.3 <- lm( Y ~ X1 + X2 + X3 + I(X2^2), data=turnip.data )
summary( model.3 )
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + I(X2^2), data = turnip.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8890  -3.4896  -0.6321   2.7720  13.9565
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   119.571     13.676   8.743 1.31e-08 ***
## X1             -3.367      2.438  -1.381   0.1811
## X2            542.504    100.526   5.397 2.03e-05 ***
## X3             -5.026      2.109  -2.383   0.0263 *
## I(X2^2)     -1209.047    195.603  -6.181 3.20e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 6.104 on 22 degrees of freedom
## Multiple R-squared:  0.9104, Adjusted R-squared:  0.8941
## F-statistic:  55.9 on 4 and 22 DF,  p-value: 3.282e-11
```

```
model.4 <- lm( Y ~ X2 + X3 + I(X2^2), data=turnip.data )
summary( model.4 )
```

```
##
## Call:
## lm(formula = Y ~ X2 + X3 + I(X2^2), data = turnip.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9663  -3.4432  -0.8141   4.2950  13.2652
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   120.627     13.922   8.665 1.06e-08 ***
## X2            490.414     95.006   5.162 3.12e-05 ***
## X3             -5.716      2.089  -2.736   0.0118 *
## I(X2^2)     -1107.853    184.910  -5.991 4.14e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.223 on 23 degrees of freedom
## Multiple R-squared:  0.9027, Adjusted R-squared:   0.89
## F-statistic: 71.09 on 3 and 23 DF,  p-value: 8.747e-12
```

The appropriate model generated via backwards elimination is:

$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_{22} X_2^2$$

Model Analysis:

1. P Value

The P-Value of the appropriate model is $8.747 * 10^{-12}$. This is statistically significant, and shows a decrease from the Anderson-Bancroft model.

2. R^2 Value

The R^2 value of the appropriate model is **0.9027**. This shows a strong, positive linear relationship. This value is significantly higher than the R^2 value of the Anderson-Bancroft model.

3. R_a^2 Value

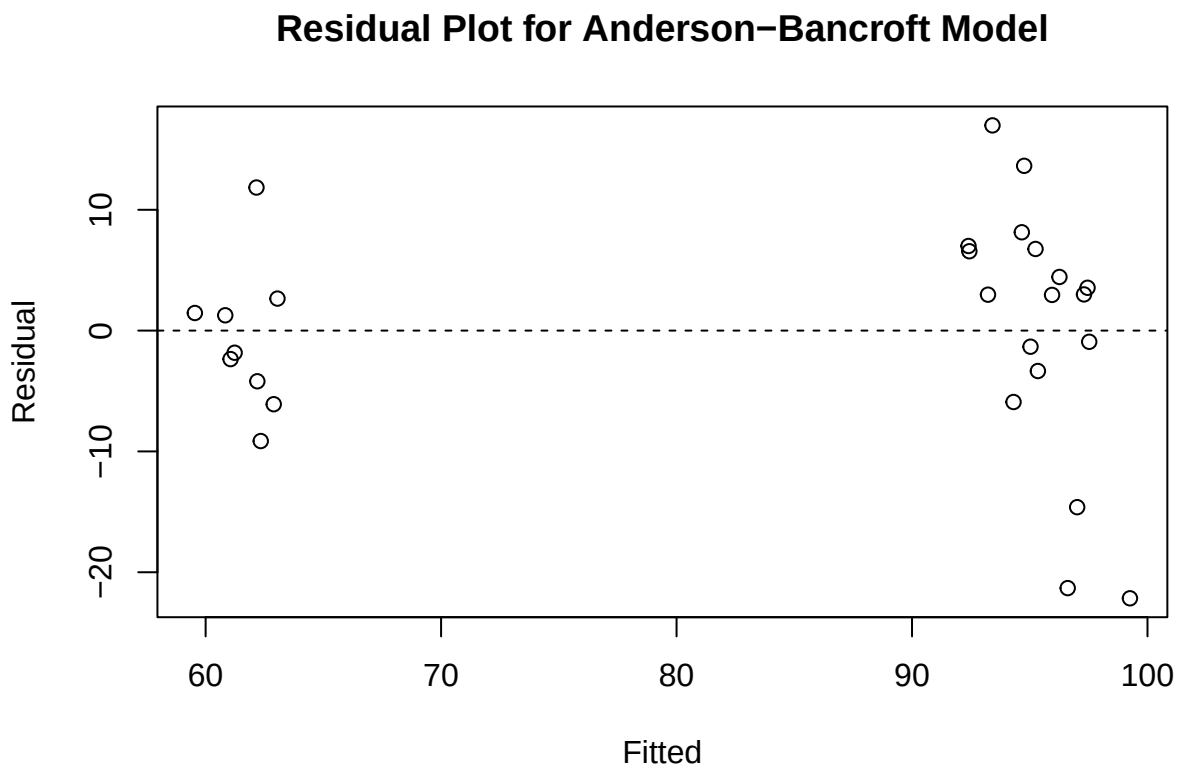
The R_a^2 value of the appropriate model is **0.89**. This value also shows a strong, positive linear relationship. The value is significantly higher than the R_a^2 value of the Anderson-Bancroft model.

4. Residual Standard Error (S Value)

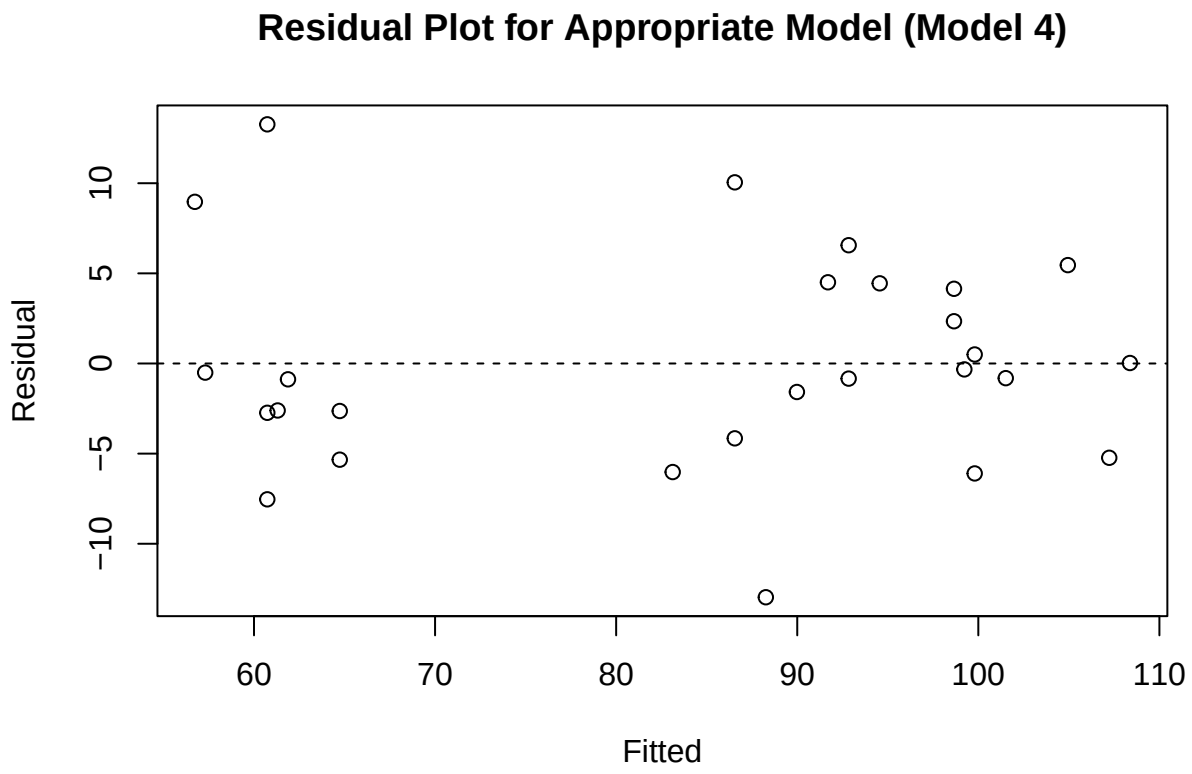
The residual standard error for the appropriate model is **6.223**. This is a significant decrease compared to the Anderson-Bancroft model.

5. Residual Scatter Plot

```
plot( fitted(anderson.bancroft.model), resid(anderson.bancroft.model), xlab='Fitted', ylab='Residual', m
abline(0, 0, lty = 2)
```



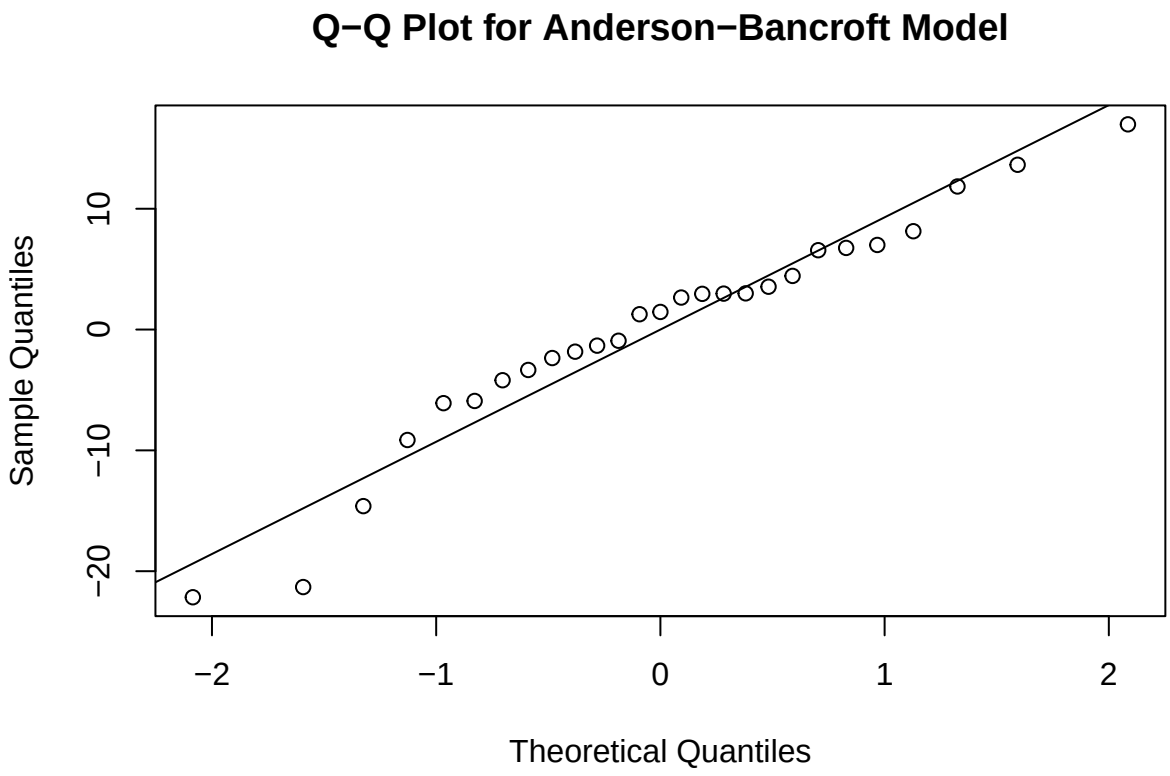

```
plot( fitted(model.4), resid(model.4), xlab='Fitted', ylab='Residual', main='Residual Plot for Appropria
abline(0, 0, lty = 2)
```



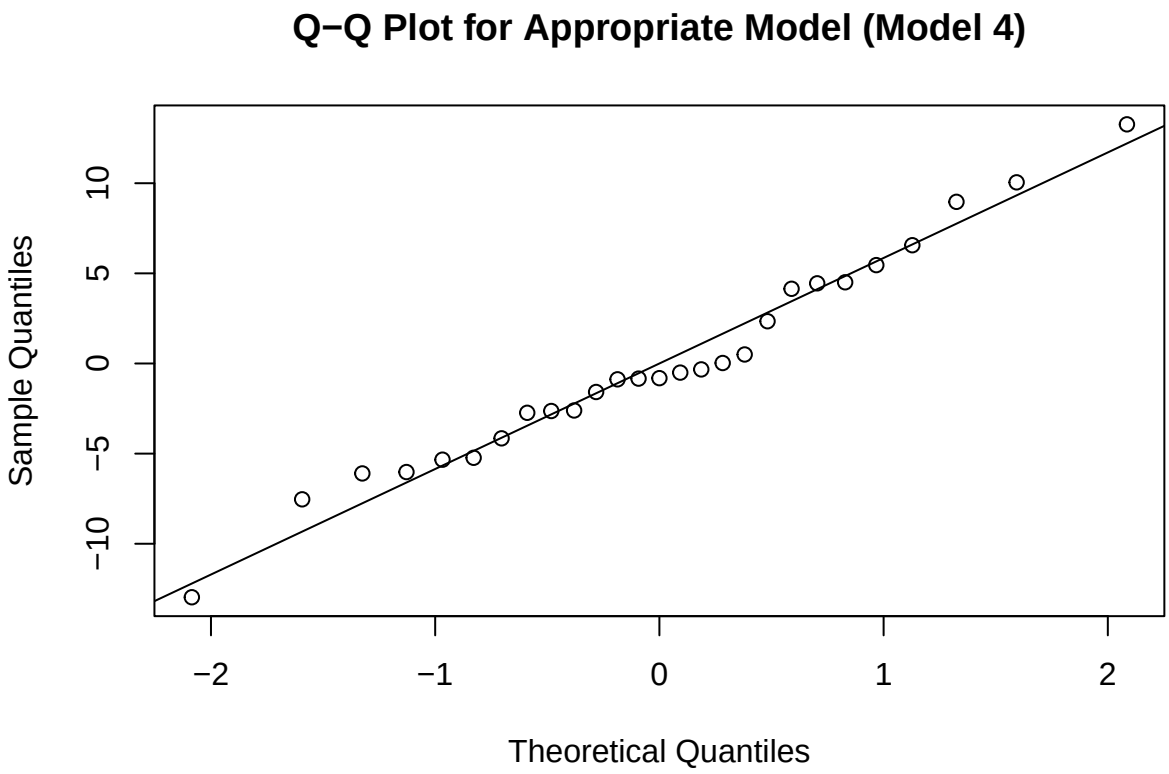
The appropriate model has a much more uniform distribution of the residuals compared to the Anderson-Bancroft model.

6. Residual Q-Q Plot

```
qqnorm( resid( anderson.bancroft.model ),
        main="Q-Q Plot for Anderson-Bancroft Model" )
abline( mean( resid( anderson.bancroft.model ) ),
        sd( resid( anderson.bancroft.model ) ) )
```




```
qqnorm( resid( model.4 ),
        main="Q-Q Plot for Appropriate Model (Model 4)")
abline( mean( resid( model.4 ) ),
        sd( resid( model.4 ) ) )
```



The appropriate model shows a stronger linear pattern with less deviation than the Anderson-Bancroft model.

7. Shapiro & Wilk Test

```
shapiro.test( resid( anderson.bancroft.model ) )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(anderson.bancroft.model)
## W = 0.94751, p-value = 0.1864
```

```
shapiro.test( resid( model.4 ) )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model.4)
## W = 0.97879, p-value = 0.834
```

The p-value of the Shapiro and Wilk test is significantly higher for the appropriate model, compared to the Anderson-Bancroft model. Thus, the residuals of the appropriate model are much more normally distributed.

Conclusion

Of all statistics, plots, and tests performed on the appropriate model and the Anderson-Bancroft model, the appropriate model performs better. The appropriate model is therefor better suited to represent the data than the Anderson-Bancroft model.