1.a.

> summary(gdata.lmdur)

Call:

lm(formula = waittime ~ duration, data = gdata)

Residuals:

Min 1Q Median 3Q Max

-17.4729 -3.8285 -0.2377 3.3466 17.5271

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 33.5060 3.4332 9.759 <2e-16 \*\*\*

duration 13.8561 0.8646 16.027 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.826 on 239 degrees of freedom

Multiple R-squared: 0.518, Adjusted R-squared: 0.516

F-statistic: 256.9 on 1 and 239 DF, p-value: < 2.2e-16

Fitted response function:

Y= 33.5060+13.8562 X1

b. From our initial model, an eruption lasting X1=4 minutes would have a waiting time of 88.9305 minutes. (33.5060+13.8561\*4).

Prediction interval at X=4:

> predict(gdata.lmdur,geyser4waittime,interval="predict")

fit lwr upr

1 88.9305 77.43018 100.4308

Confidence interval at X=4:

> predict(gdata.lmdur,geyser4waittime,interval="confidence")

fit lwr upr

1 88.9305 88.18581 89.67519

c. From our initial model, an eruption lasting X1=2 minutes would have a waiting time of 61.21827 minutes. (33.5060+13.8561\*2).

Confidence interval at X=2

> predict(gdata.lmdur,geyser2waittime,interval="confidence")

fit lwr upr

1 61.21827 57.82049 64.61605

Prediction interval at X=2

> predict(gdata.lmdur,geyser2waittime,interval="predict")

fit lwr upr

1 61.21827 49.24966 73.18688

Compared to our intervals found in part b, our confidence interval seems less precise but our prediction interval also seems slightly more precise. This is in part due to having significantly less observations near X=2 from our data set than X=4. Intervals at X=4 will be more accurate because they are more significant than intervals at X=2.

d.

> summary(geyser.lmheight)

Call:

lm(formula = waittime ~ duration + height, data = gdata)

Residuals:

Min 1Q Median 3Q Max

-17.8734 -3.9960 -0.2936 3.0799 17.7915

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 39.28081 5.36405 7.323 3.73e-12 \*\*\*

duration 13.94437 0.86514 16.118 < 2e-16 \*\*\*

height -0.04433 0.03168 -1.399 0.163

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.814 on 238 degrees of freedom

Multiple R-squared: 0.5219, Adjusted R-squared: 0.5179

F-statistic: 129.9 on 2 and 238 DF, p-value: < 2.2e-16

As seen above, the height variable may not improve the predictive ability of the model as it does not have a P value that hits our threshold, however the F statistic is still significant so a further test must be done.

|  |
| --- |
| > app.step.fw=step(fit2, direction="forward", scope=list(upper=fit1, lower=fit2))  Start: AIC=1025.29  waittime ~ 1  Df Sum of Sq RSS AIC  + duration 1 8717.2 8111.2 851.41  <none> 16828.4 1025.29  + height 1 1.7 16826.7 1027.27  Step: AIC=851.41  waittime ~ duration  Df Sum of Sq RSS AIC  <none> 8111.2 851.41  + height 1 66.179 8045.1 851.43 |
|  |
| |  | | --- | | > | |

Using stepwise regression, we see that height does not add to the predictive ability of our model.

ii.

> summary(gdata.lmvol)

Call:

lm(formula = waittime ~ volume, data = gdata)

Residuals:

Min 1Q Median 3Q Max

-16.7133 -4.4636 -0.1901 4.4930 19.8746

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 56.918726 3.283712 17.33 <2e-16 \*\*\*

volume 0.057331 0.005959 9.62 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.124 on 239 degrees of freedom

Multiple R-squared: 0.2791, Adjusted R-squared: 0.2761

F-statistic: 92.55 on 1 and 239 DF, p-value: < 2.2e-16

Compared to models in a and b, this model has significant p-values and thus the potential to be a predictive candidate. However, the multiple R-squared values are lower which could mean that a smaller proportion of our model fits this regression function.

> summary(gdata.lmall3)

Call:

lm(formula = waittime ~ duration + height + volume, data = gdata)

Residuals:

Min 1Q Median 3Q Max

-17.926 -4.075 -0.356 3.164 18.049

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13.29219 44.89093 0.296 0.7674

duration 20.47436 11.23187 1.823 0.0696 .

height 0.15242 0.33889 0.450 0.6533

volume -0.04942 0.08475 -0.583 0.5604

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

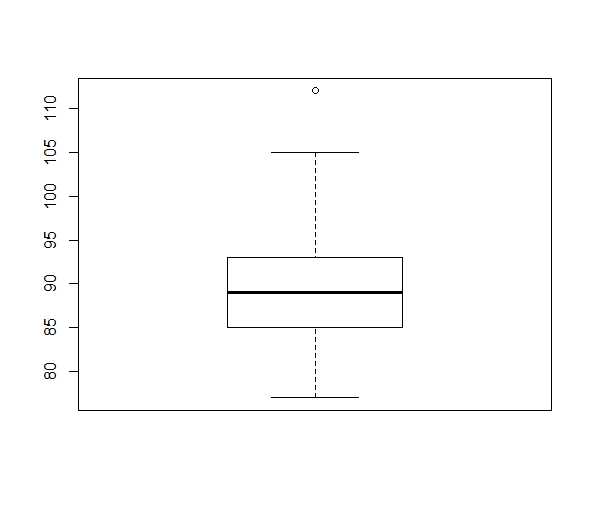
Residual standard error: 5.822 on 237 degrees of freedom

Multiple R-squared: 0.5226, Adjusted R-squared: 0.5166

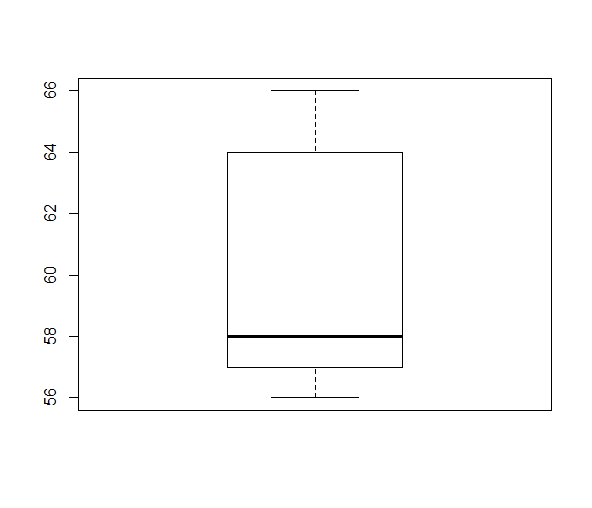
F-statistic: 86.49 on 3 and 237 DF, p-value: < 2.2e-16

We see a higher R^2 value with this model than the previous one, which could mean to a larger proportion of waiting times being correlated with our X variables. However, a large concern is that none of the P-values are significant, which could lead to the removal of multiple predictor variables to have a better fit.

e. boxplot of waiting times with durations greater than 3 minutes



Boxplot of waiting times for durations less than 3 minutes



> summary(data.less3)

Duration\_min waitingtime\_min Height

Min. :1.800 Min. :56.00 Min. :120.0

1st Qu.:1.900 1st Qu.:57.00 1st Qu.:125.0

Median :2.020 Median :58.00 Median :130.0

Mean :2.100 Mean :60.20 Mean :131.0

3rd Qu.:2.315 3rd Qu.:63.75 3rd Qu.:133.8

Max. :2.500 Max. :66.00 Max. :150.0

summary(data.more3)

Duration\_min waitingtime\_min Height

Min. :3.330 Min. : 77.00 Min. :100.0

1st Qu.:3.920 1st Qu.: 85.00 1st Qu.:130.0

Median :4.030 Median : 89.00 Median :140.0

Mean :4.027 Mean : 89.41 Mean :138.4

3rd Qu.:4.130 3rd Qu.: 93.00 3rd Qu.:150.0

Max. :4.570 Max. :112.00 Max. :170.0

> summary(gdata.lmless)

Call:

lm(formula = wless ~ xless, data = data.less3)

Residuals:

Min 1Q Median 3Q Max

-2.0863 -1.6804 0.1608 1.2529 3.0157

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 30.965 5.145 6.018 0.000317 \*\*\*

xless 13.922 2.433 5.721 0.000443 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.904 on 8 degrees of freedom

Multiple R-squared: 0.8036, Adjusted R-squared: 0.7791

F-statistic: 32.73 on 1 and 8 DF, p-value: 0.0004433

> summary(gdata.lmmore)

Call:

lm(formula = wmore ~ xmore, data = data.more3)

Residuals:

Min 1Q Median 3Q Max

-15.6966 -3.8039 -0.4935 3.2862 19.3034

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 53.923 7.756 6.953 3.70e-11 \*\*\*

xmore 8.812 1.923 4.581 7.59e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.832 on 229 degrees of freedom

Multiple R-squared: 0.08396, Adjusted R-squared: 0.07996

F-statistic: 20.99 on 1 and 229 DF, p-value: 7.586e-06

The fit for values less than 3 seems to be a better fit than the original model as we achieve a higher R2 value as well as very significant p-values. The second model achieves small p-values, however the R2 value suggests minimal proportion of correlation covered at R=.08