

## Regression diagnostics with R\*

```
> sat =  
read.table("http://www.utstat.utoronto.ca/~brunner/302f14/code_n_data/lectu  
re/sat.data")  
> head(sat)  
  VERBAL MATH GPA  
1    623  509 2.6  
2    454  471 2.3  
3    643  700 2.4  
4    585  719 3.0  
5    719  710 3.1  
6    693  643 2.9  
> mod1 = lm(GPA ~ VERBAL+MATH, data=sat); summary(mod1)
```

```
Call:  
lm(formula = GPA ~ VERBAL + MATH, data = sat)
```

```
Residuals:  
      Min       1Q   Median       3Q      Max  
-2.24875 -0.35113  0.04659  0.38745  1.03527
```

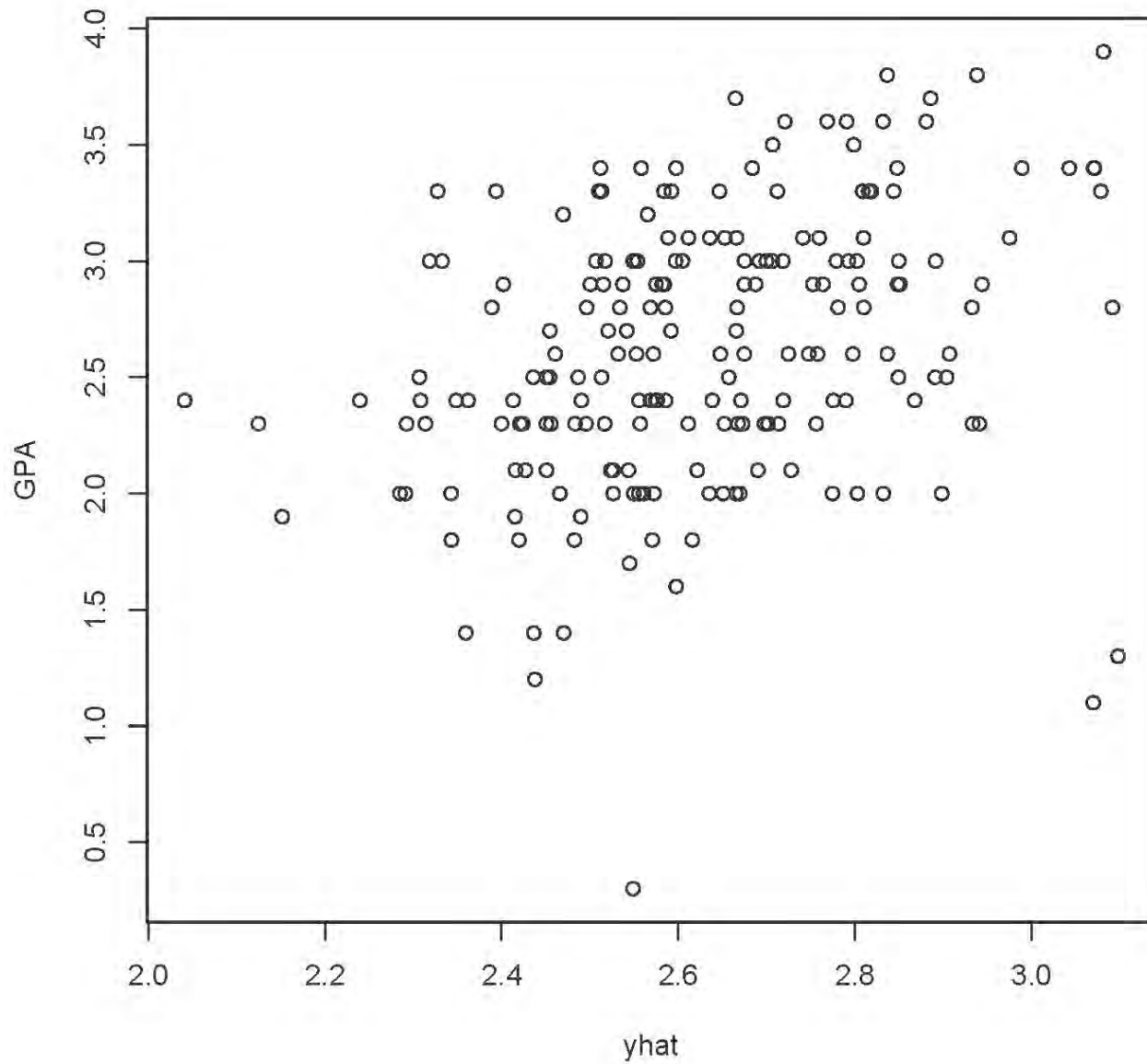
```
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) 0.6062975  0.4414062   1.374   0.171  
VERBAL       0.0023072  0.0005522   4.178 4.42e-05 ***  
MATH         0.0009999  0.0006093   1.641   0.102
```

```
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

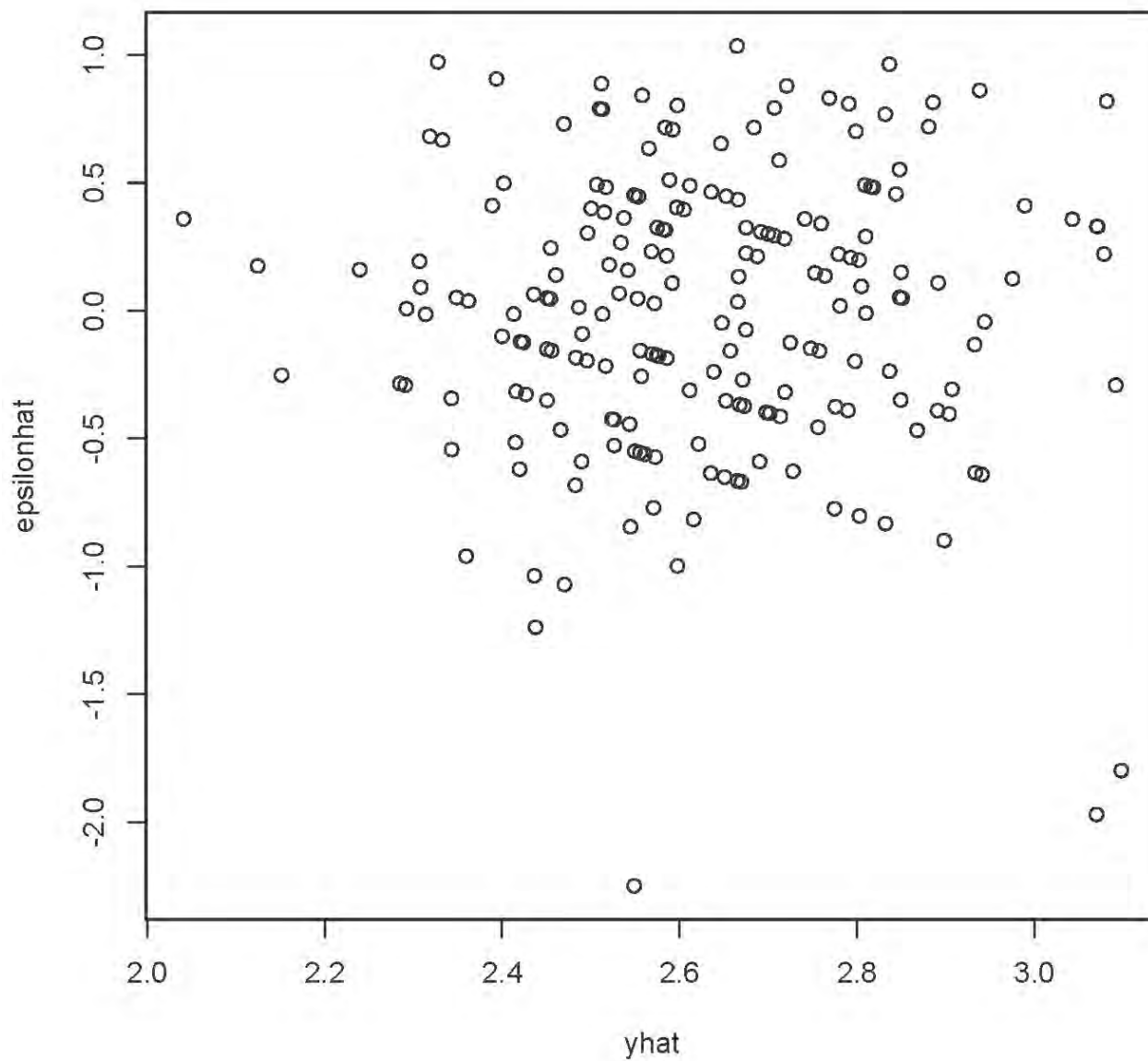
```
Residual standard error: 0.5484 on 197 degrees of freedom  
Multiple R-squared: 0.1161, Adjusted R-squared: 0.1071  
F-statistic: 12.93 on 2 and 197 DF, p-value: 5.284e-06
```

```
> attach(sat) # Make variable names accessible  
>
```

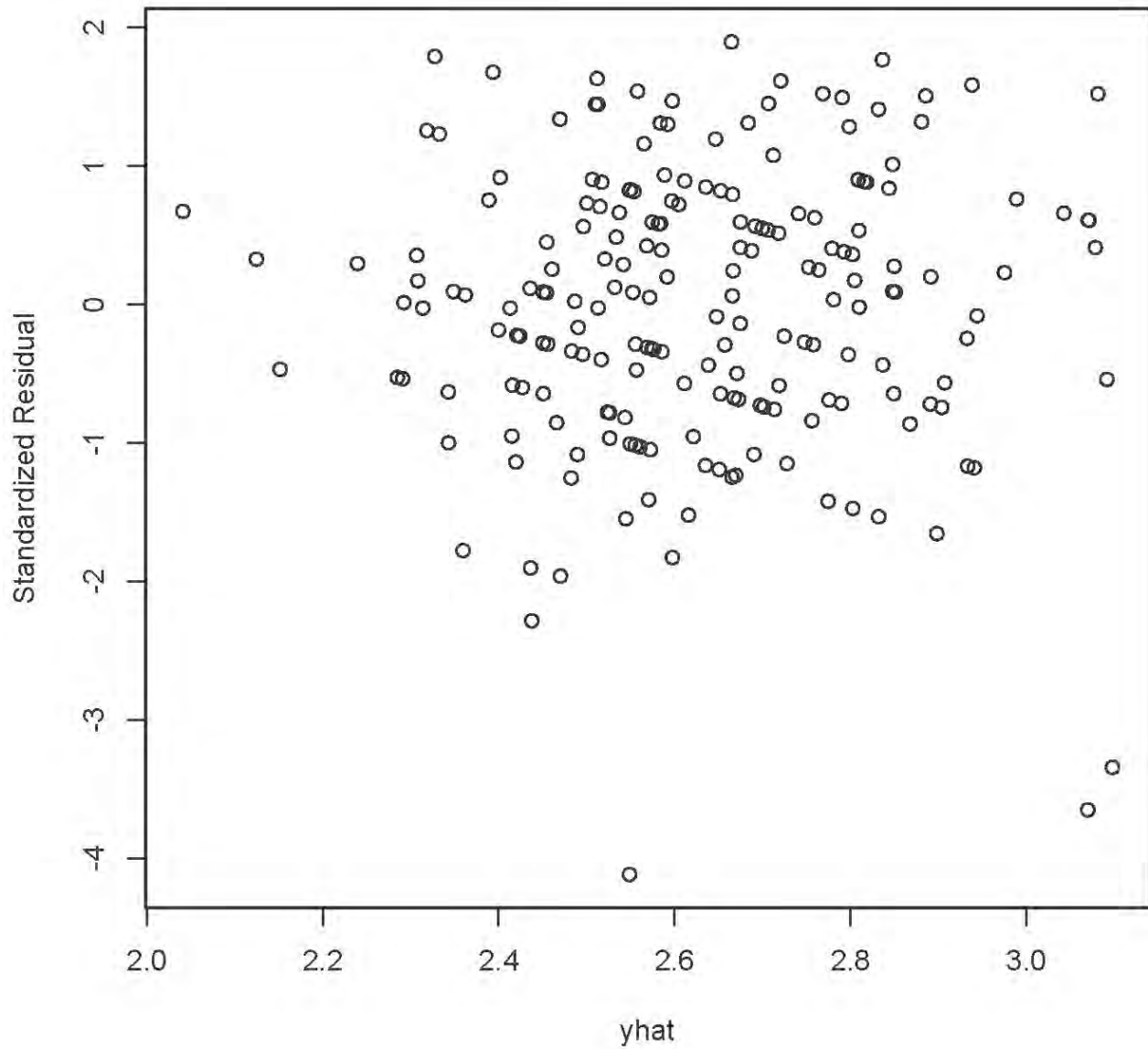
```
> # Plot y-hat versus y
> yhat = mod1$fitted.values
> plot(yhat,GPA)
>
```



```
> # Plot y-hat versus residuals  
> epsilonhat = mod1$residuals  
> plot(yhat,epsilonhat)
```



```
> # Compare plot of standardized residuals  
> sr = rstandard(mod1)  
> plot(yhat,sr,ylab='Standardized Residual')
```



```

> # Three look like possible outliers: Investigate
> id = 1:200
> suspect = id[sr < -3]
> cbind(sat[suspect,],yhat[suspect],epsilonhat[suspect])
      VERBAL MATH GPA yhat[suspect] epsilonhat[suspect]
121    780  692 1.3      3.097791      -1.797791
131    578  609 0.3      2.548754      -2.248754
136    760  710 1.1      3.069645      -1.969645

> # Studentized deleted residuals are t-statistics
> sdr = rstudent(mod1) # Studentized deleted residuals
> # Bonferroni critical value for n=200 tests, at joint alpha = 0.05 level
> dfe = mod1$df.residual; dfe
[1] 197
> alpha = 0.05; a = alpha/200; bcrit = qt(1-a/2,dfe-1); bcrit
[1] 3.730706
> sdr[abs(sdr)>bcrit]
      131      136
-4.293141 -3.768640
>

```

I feel that all three suspicious points are worthy of investigation.

# Trees Data

```
> rm(list=ls()) # Remove everything to start
> head(trees)
  Girth Height Volume
1   8.3    70   10.3
2   8.6    65   10.3
3   8.8    63   10.2
4  10.5    72   16.4
5  10.7    81   18.8
6  10.8    83   19.7
> attach(trees)
> mod1 = lm(Volume ~ Girth + Height)
> summary(mod1)
```

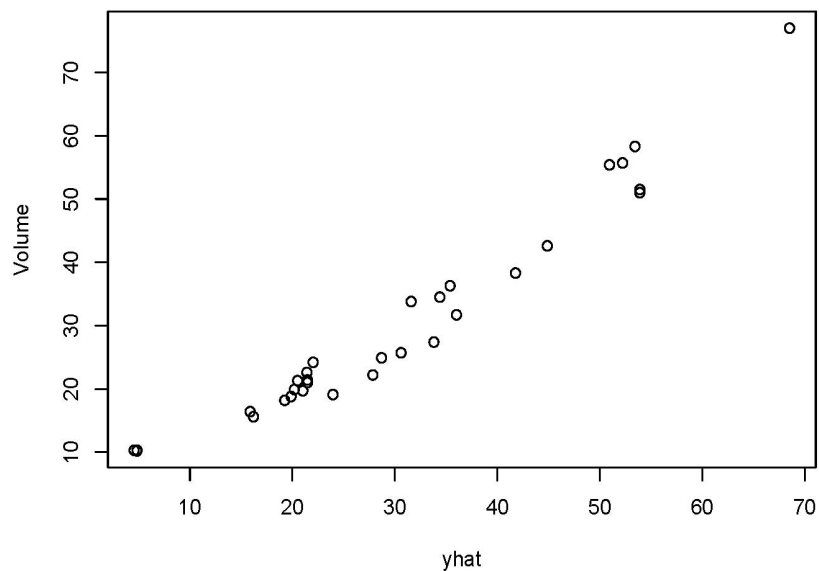
```
Call:
lm(formula = Volume ~ Girth + Height)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847
```

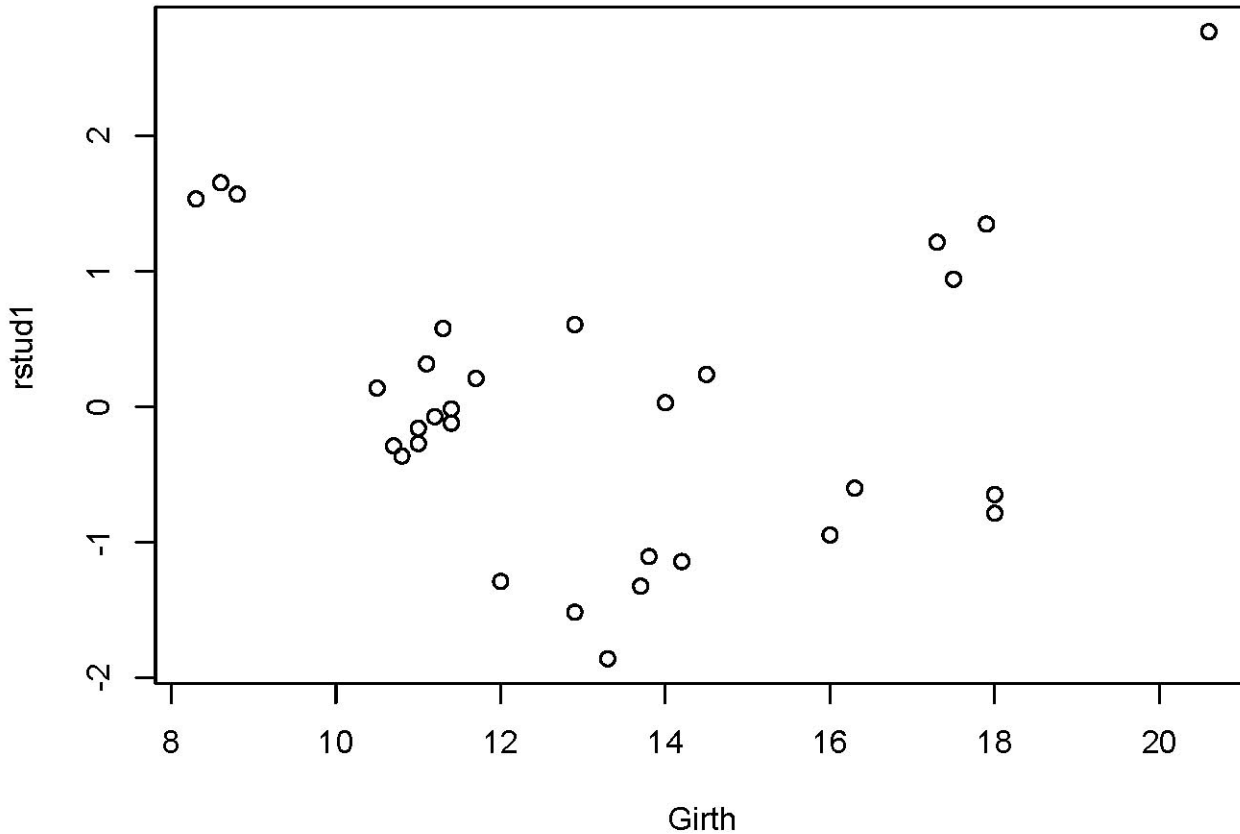
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Girth         4.7082     0.2643  17.816 < 2e-16 ***
Height        0.3393     0.1302   2.607  0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948, Adjusted R-squared:  0.9442
F-statistic: 255 on 2 and 28 DF, p-value: < 2.2e-16
```

```
>
> plot(yhat, Volume)
```



```
> rstud1 = rstudent(mod1) # Studentized deleted residuals
> plot(Girth,rstud1)
```



```
> plot(Height,rstud1) # No pattern
> Girthsq = Girth^2 # Polynomial term -- literally square it
> mod2 = lm(Volume ~ Girth + Girthsq + Height); summary(mod2)
```

```
Call:
lm(formula = Volume ~ Girth + Girthsq + Height)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-4.2928 -1.6693 -0.1018  1.7851  4.3489
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.92041    10.07911  -0.984  0.333729
Girth        -2.88508     1.30985  -2.203  0.036343 *
Girthsq       0.26862     0.04590   5.852  3.13e-06 ***
Height        0.37639     0.08823   4.266  0.000218 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.625 on 27 degrees of freedom
Multiple R-squared:  0.9771, Adjusted R-squared:  0.9745
F-statistic: 383.2 on 3 and 27 DF, p-value: < 2.2e-16
```

```

> a = (0.9771-0.948)/(1-0.948); a # Proportion of remaining variation
[1] 0.5596154

> rstud2 = rstudent(mod2)
> plot(Girth,rstud2) # Nothing
> plot(Height,rstud2) # Nothing
>
> # Look at prediction intervals
> cbind(Volume[1:5],predict(mod1,interval='predict')[1:5,])
      fit      lwr      upr
1 10.3  4.837660 -3.561809 13.23713
2 10.3  4.553852 -3.962908 13.07061
3 10.2  4.816981 -3.809144 13.44311
4 16.4 15.874115  7.690594 24.05764
5 18.8 19.869008 11.451358 28.28666
Warning message:
In predict.lm(mod1, interval = "predict") :
  predictions on current data refer to _future_ responses

>
> cbind(Volume[1:5],predict(mod2,interval='predict')[1:5,])
      fit      lwr      upr
1 10.3 10.985950  4.902269 17.06963
2 10.3  9.600406  3.566753 15.63406
3 10.2  9.205421  3.163767 15.24708
4 16.4 16.501775 10.954762 22.04879
5 18.8 20.451204 14.746331 26.15608
Warning message:
In predict.lm(mod2, interval = "predict") :
  predictions on current data refer to _future_ responses

```