STAT2402: Analysis of Observations

R. Nazim Khan
Department of Mathematics and Statistics
nazim.khan@uwa.edu.au

The University of Western Australia

1. Linear Statistical Model—Continued

Contents

Analysis of cost of power data

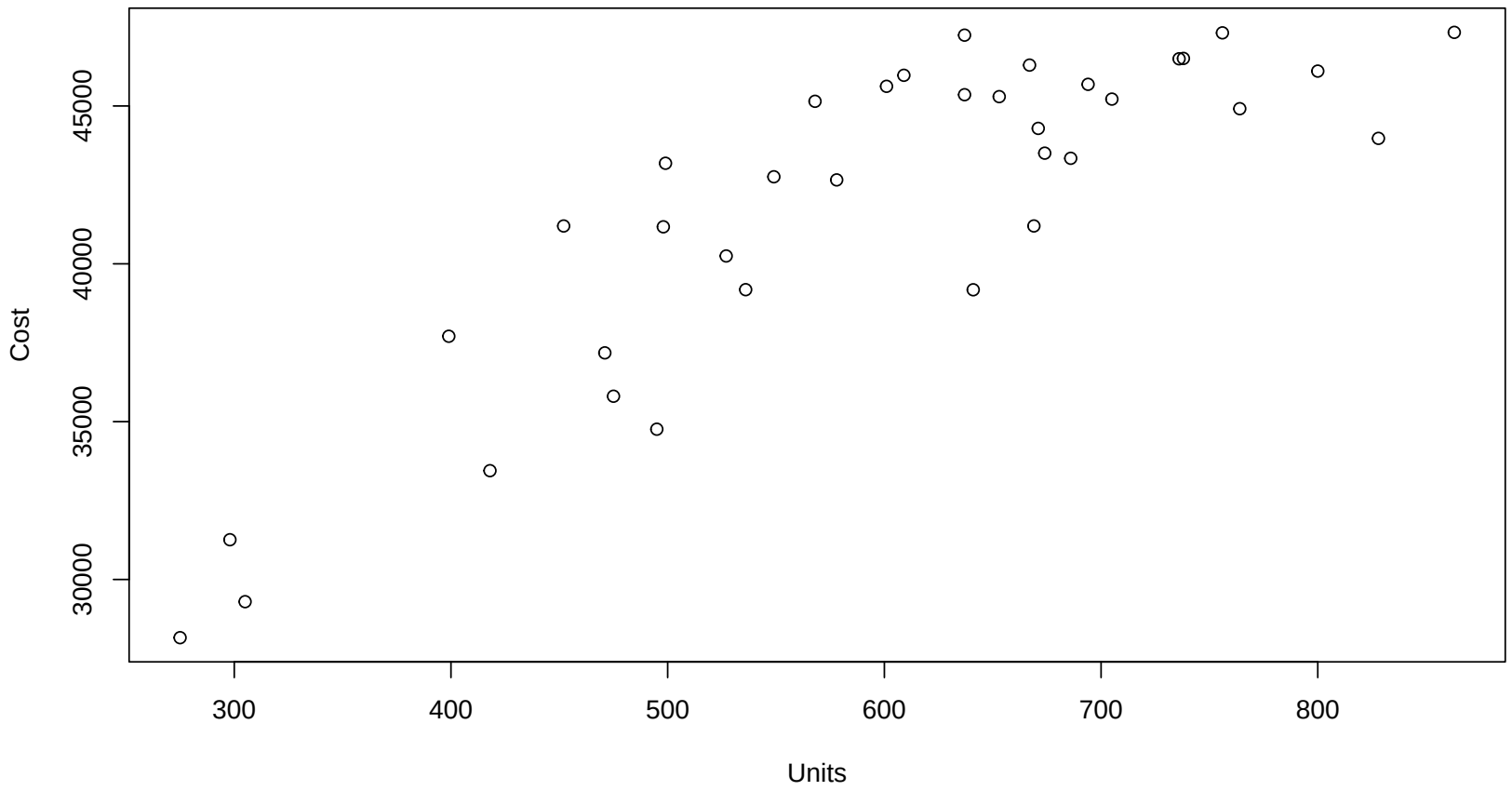
We analyse the data `power.txt` to build a model for cost of power based on the usage. The data contains the cost of power and the usage in units. First read in the data and plot it.

```
power <- read.table("../Data/power.txt", header = T, sep = "\t", stringsAsFactors = T)
summary(power)
```

```
##      Month            Cost          Units
##  Min.   : 1.00   Min.   :28157   Min.   :275.0
##  1st Qu.: 9.75   1st Qu.:39180   1st Qu.:497.2
##  Median :18.50   Median :43424   Median :623.0
##  Mean   :18.50   Mean   :41778   Mean   :593.7
##  3rd Qu.:27.25   3rd Qu.:45639   3rd Qu.:688.0
##  Max.   :36.00   Max.   :47332   Max.   :863.0
```

Plot of data

```
with(power, plot(Cost ~ Units))
```



The cost appears to increase with usage, but seems to flatten off as cost increases. We start by fitting a simple linear model

Linear model

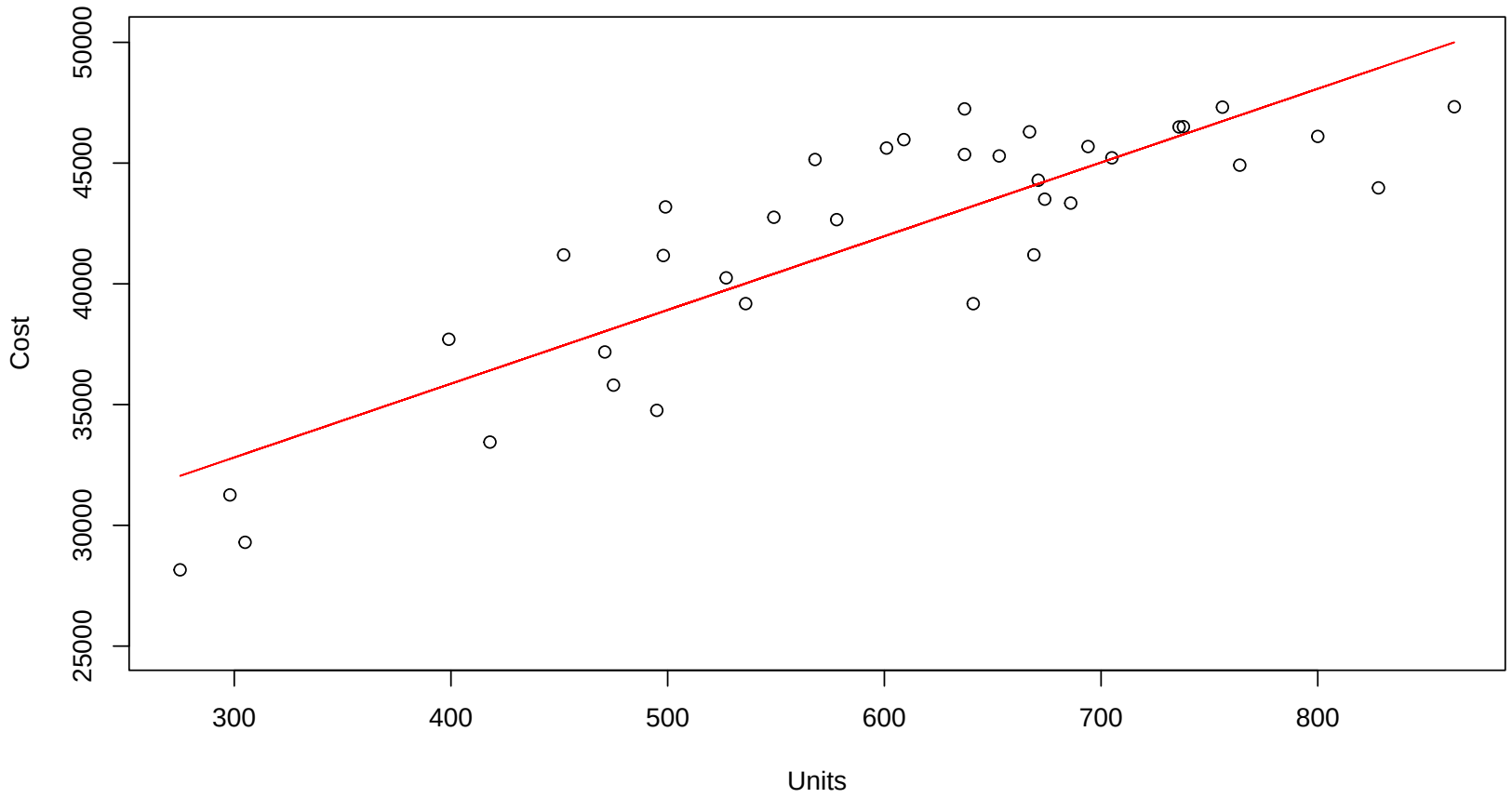
```
plm <- lm(Cost ~ Units, data = power)
summary(plm)
```

```
##
## Call:
## lm(formula = Cost ~ Units, data = power)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4958.9 -2136.0   236.4  2261.4  4297.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23651.489   1917.137  12.337 4.17e-14 ***
## Units          30.533      3.137   9.734 2.32e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2734 on 34 degrees of freedom
## Multiple R-squared:  0.7359,Adjusted R-squared:  0.7282
## F-statistic: 94.75 on 1 and 34 DF,  p-value: 2.317e-11
```

Model diagnostics

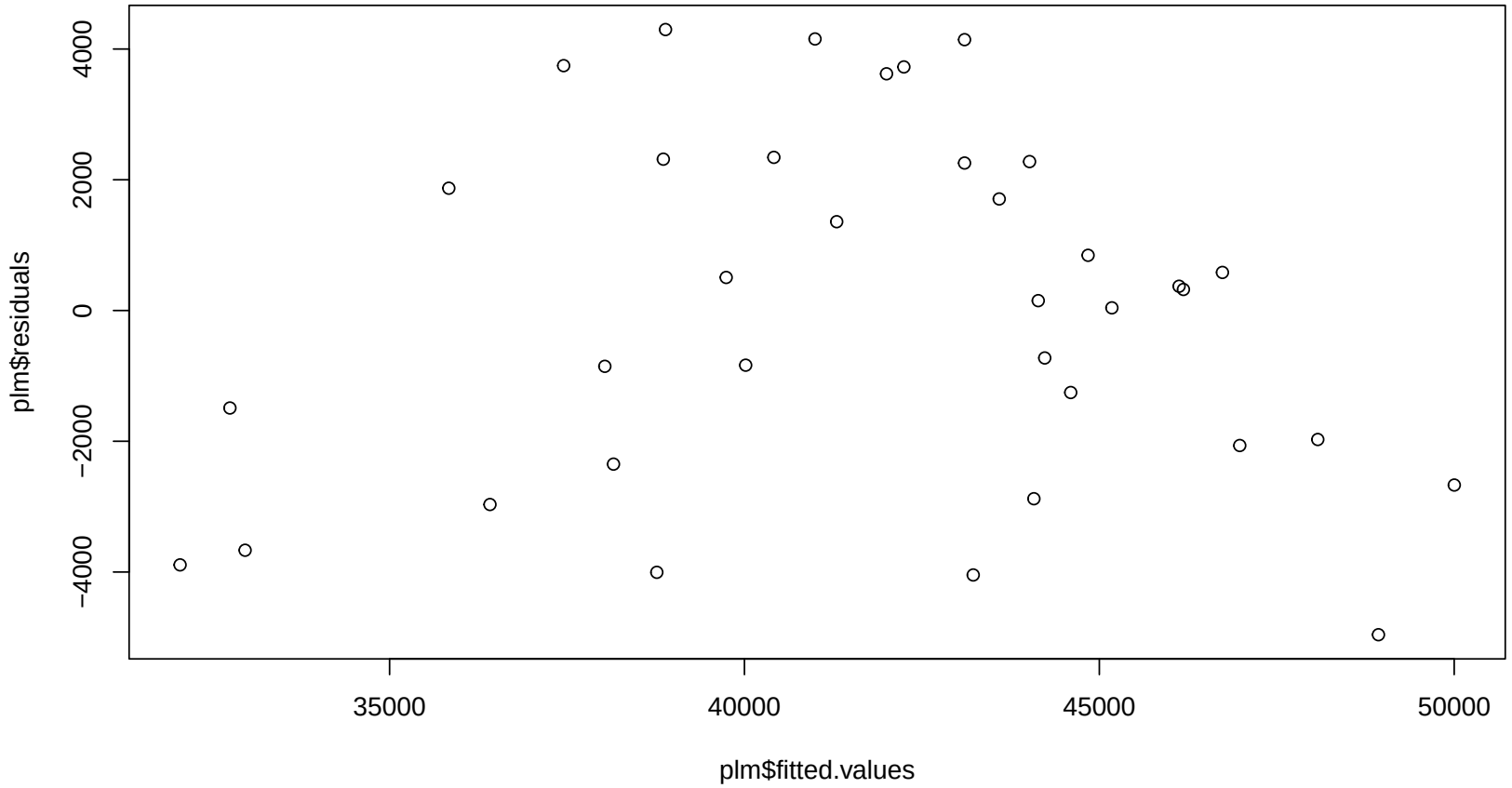
Some diagnostics now. First the plot of the fit.

```
with(power, plot(Cost ~ Units, ylim = c(25000, 50050)))
lines(predict.lm(plm) ~ power$Units, col = "red")
```



Next step

```
plot(plm$residuals ~ plm$fitted.values)
```



Nothing too clear here! But a closer look does indicate some curvature and a quadratic trend. Let us include a quadratic term

Log model

```
plm2 <- lm(Cost ~ I(log(Units)), data = power)
summary(plm2)
```

```
##
## Call:
## lm(formula = Cost ~ I(log(Units)), data = power)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4573.6 -1439.2   184.7  1758.1  3716.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -63993       9144  -6.998 4.49e-08 ***
## I(log(Units))    16654       1438  11.578 2.41e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2393 on 34 degrees of freedom
## Multiple R-squared:  0.7977,Adjusted R-squared:  0.7917
## F-statistic:   134 on 1 and 34 DF,  p-value: 2.409e-13
```

```
plot(plm2$residuals ~ plm2$fitted.values)
with(power, plot(Cost ~ Units, ylim = c(25000, 50050)))
lines(sort(predict.lm(plm2)) ~ sort(power$Units), col = "red")
AIC(plm1)
```

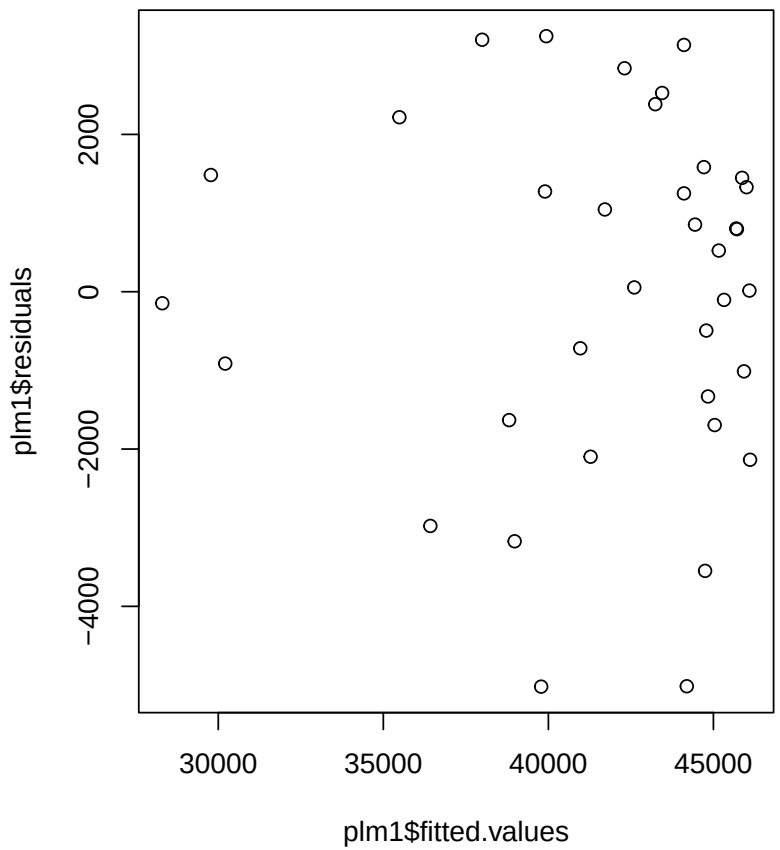
```
## [1] 663.7554
```

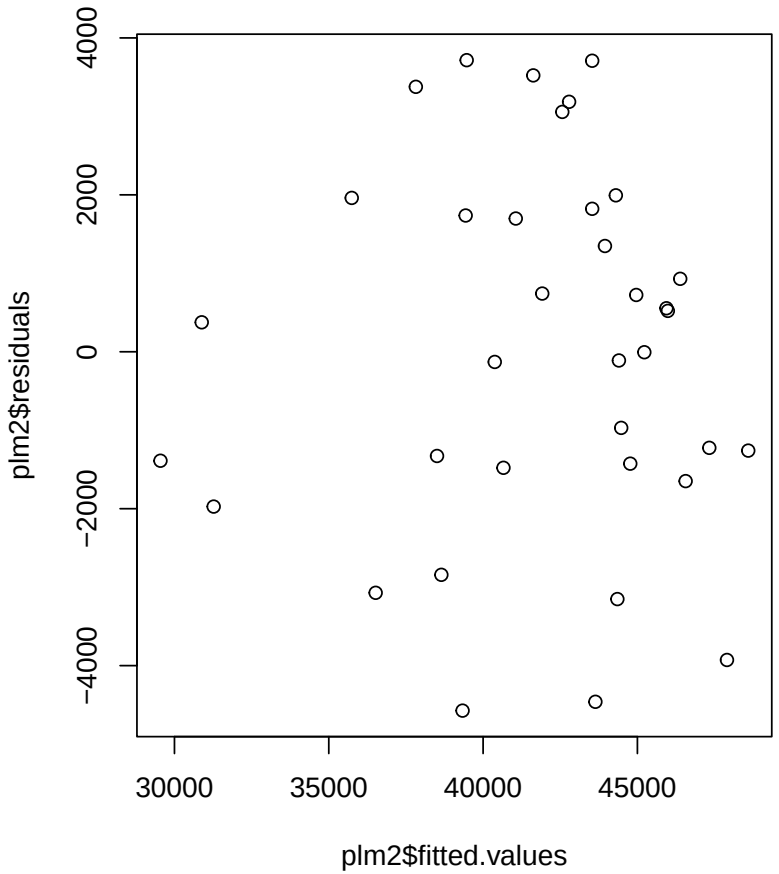
```
AIC(plm2)
```

```
## [1] 666.2827
```

Diagnostics

```
oldpar <- par(mfrow = c(1, 2))
plot(plm1$residuals ~ plm1$fitted.values)
plot(plm2$residuals ~ plm2$fitted.values)
par(oldpar)
```

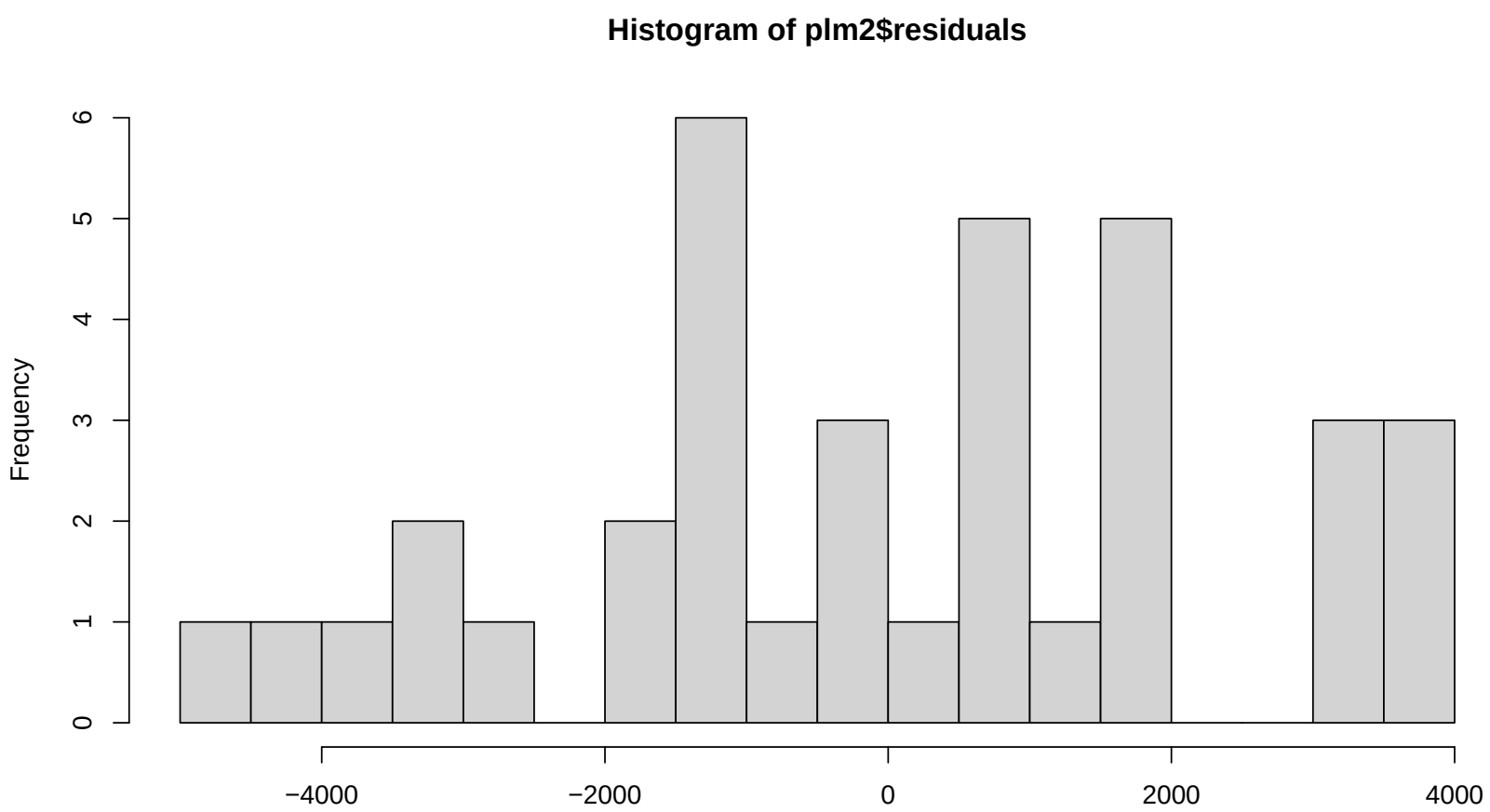




Normality

Let us look at normality assumption.

```
hist(plm2$residuals, nclass = 20)
plm2.stdres = rstandard(plm2)
hist(plm2.stdres)
qqnorm(plm2.stdres, ylab = "Standardized Residuals", xlab = "Normal Scores", main = "Normal Probability pl
qqline(plm2.stdres)
```



Conclusion

The histogram of residuals does not look to be from a normal distribution. The normal probability plot is expected to be close to a straight line. In the given plot the departures from straight line are not severe, so there is not reason to doubt the normality assumption. The departures are at either end. At both ends the plot flattens off, indicating that the normal scores continue but the residuals stop. This indicates a "cliff", that is, a short tail for the data.

Can this model be improved?

Let us try an extra term in the model.

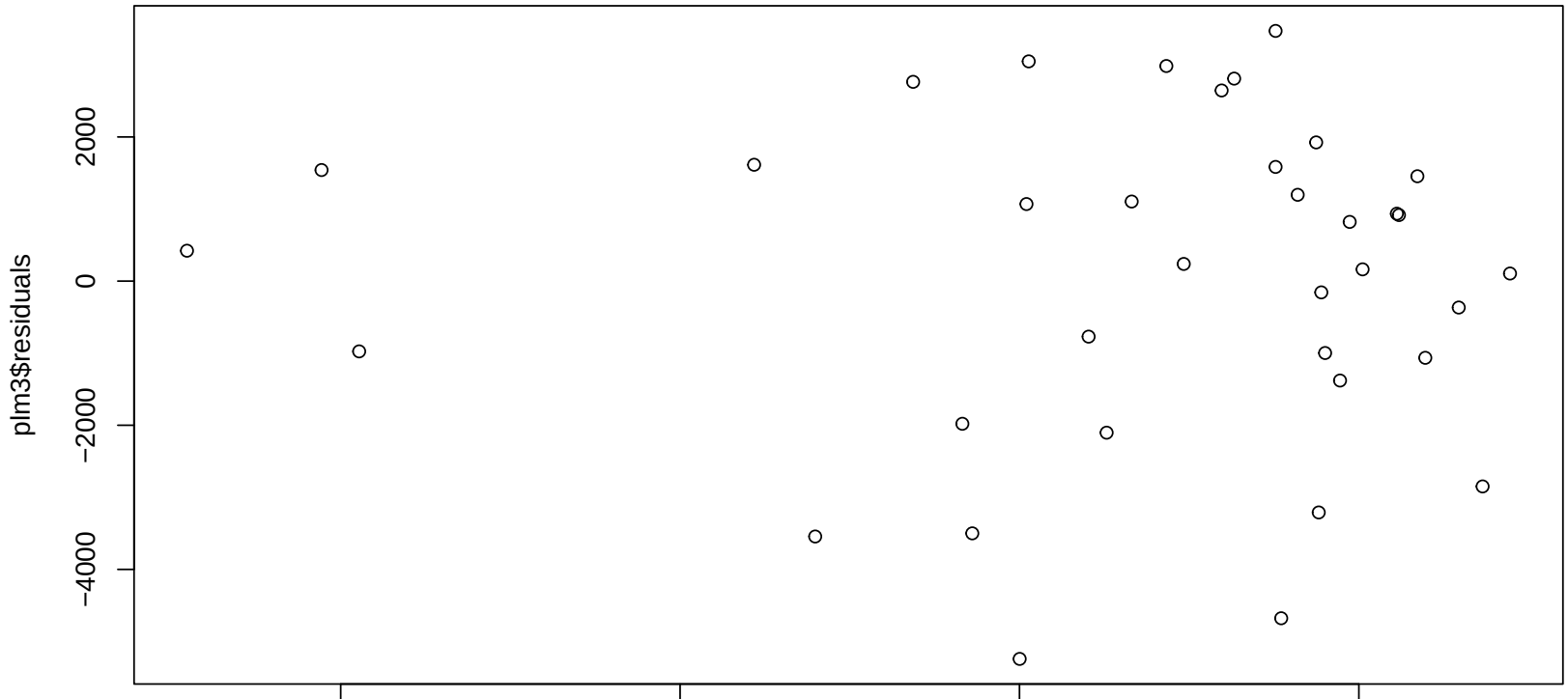
```
plm3 <- lm(Cost ~ I(log(Units)) + I(sqrt(log(Units))), data = power)
summary(plm3)
```

```
##
## Call:
## lm(formula = Cost ~ I(log(Units)) + I(sqrt(log(Units))), data = power)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5241.7 -1143.2   329.8  1551.9  3471.1
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1117903     624539  -1.790   0.0826 .
## I(log(Units))        -153338     100736  -1.522   0.1375
## I(sqrt(log(Units)))   846805     501759   1.688   0.1009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2330 on 33 degrees of freedom
## Multiple R-squared:  0.8138,Adjusted R-squared:  0.8025
## F-statistic: 72.09 on 2 and 33 DF,  p-value: 9.048e-13
```

MOdel fit

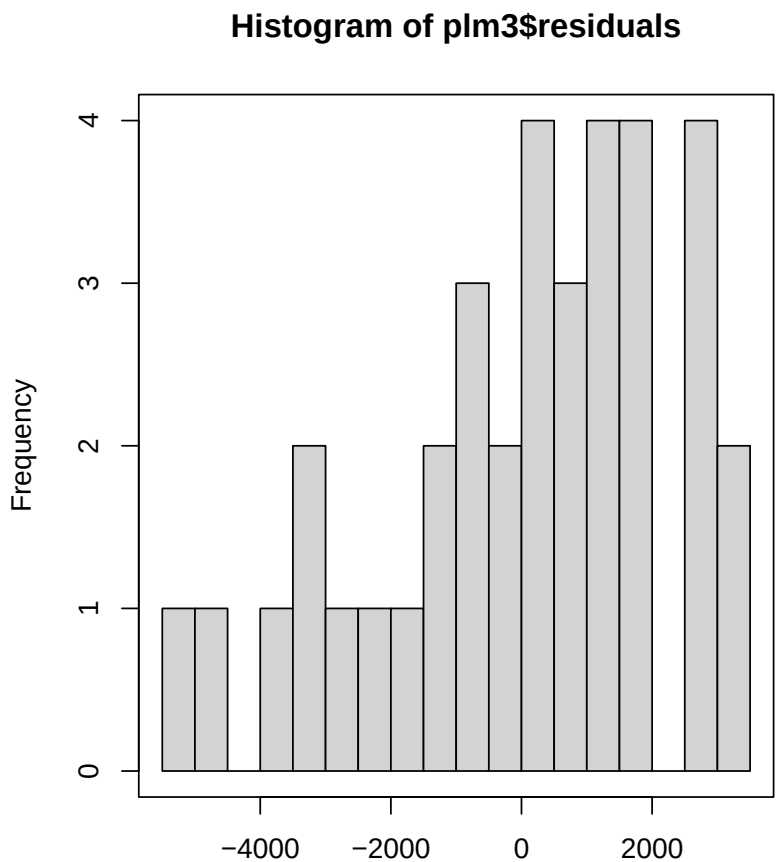
```
plot(plm3$residuals ~ plm3$fitted.values)
with(power, plot(Cost ~ Units, ylim = c(25000, 50050)))
lines(sort(predict.lm(plm3)) ~ sort(power$Units), col = "red")
AIC(plm1, plm2, plm3)
```

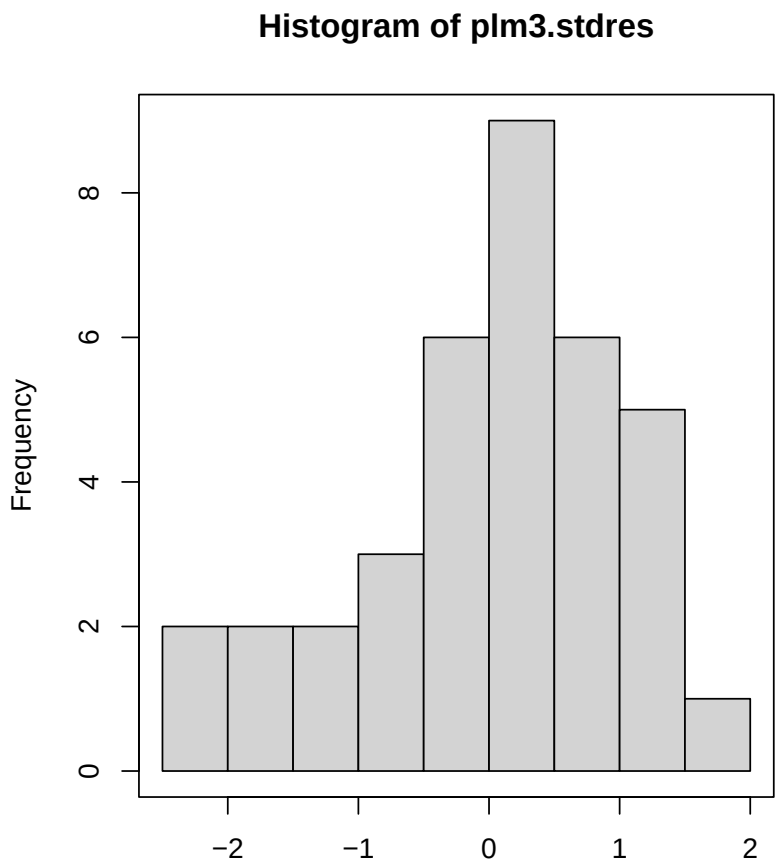
```
##      df      AIC
## plm1  4 663.7554
## plm2  3 666.2827
## plm3  4 665.3024
```



Diagnostics

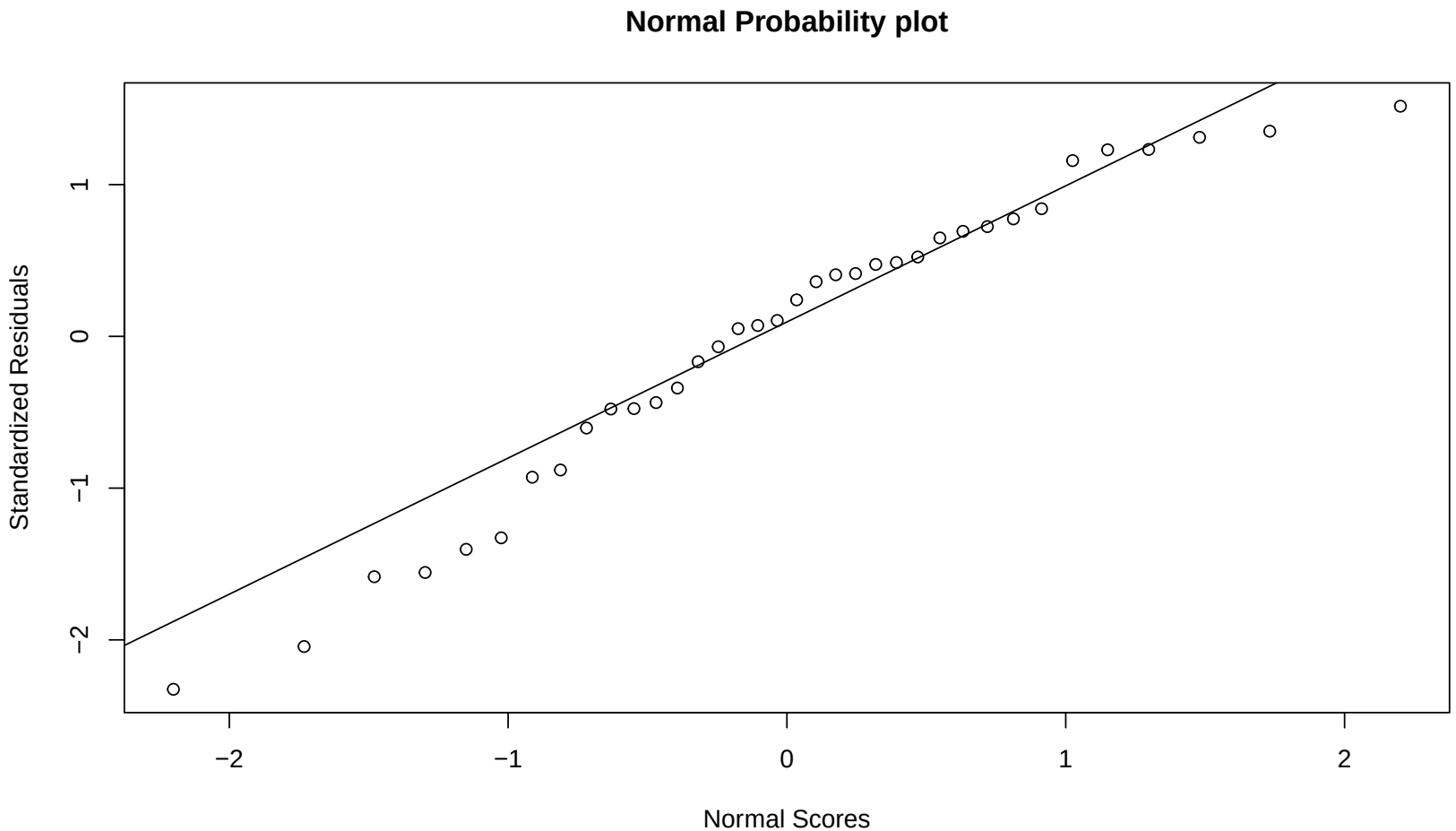
```
oldpar <- par(mfrow = c(1, 2))
hist(plm3$residuals, nclass = 20)
box()
plm3.stdres = rstandard(plm3)
hist(plm3.stdres)
box()
par(oldpar)
```





Diagnostics (ctd)

```
qqnorm(plm3.stdres, ylab = "Standardized Residuals", xlab = "Normal Scores", main = "Normal Probability pl
qqline(plm3.stdres)
```



Is this better?

Looks better than the previous model, but harder to interpret! I will be happy with just the log model.
Any other ways of improving the model? Well, if you examine the plot of residuals against fitted values for model 2, you will see some evidence of heterogenous variance. That topic is covered in a third year unit, STAT3401.

1.13 Example: Urchin data

Constable (1993) compared the inter-radial suture widths of urchins maintained on one of three food regimes.

- Initial: no additional food supplied above what was in the initial sample
- low: food supplied periodically
- high: food supplied freely

To control for variation in urchin sizes, the initial body volume of each urchin was measured.

Reference: A. J. Constable 1993) The role of sutures in shrinking of the test in *Heliocidaris erythrogramma* (Echinoidea: Echinometridae). *Marine biology*, **117**, 423–430.

Data exploration

The aim of the analysis is to determine how the suture width depends on the food regime, while adjusting for size or urchin.

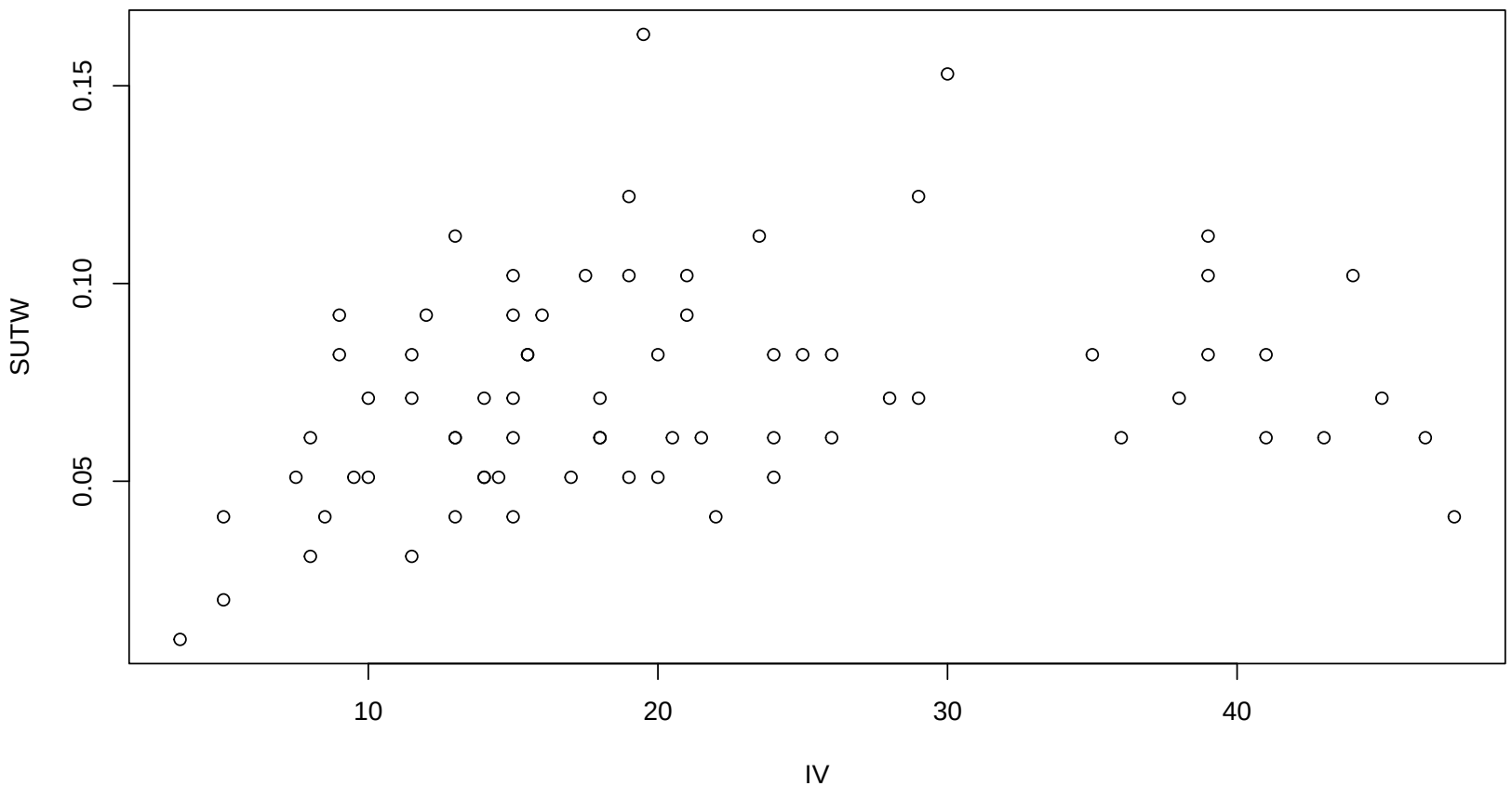
```
constable <- read.table("Data/constable.csv", header = T, sep = ",", stringsAsFactors = T)
summary(constable)
```

```
##      TREAT          IV              SUTW
##  High   :24   Min.   : 3.50   Min.   :0.01000
##  Initial:24   1st Qu.:13.00   1st Qu.:0.05100
##  Low    :24   Median :18.00   Median :0.07100
##               Mean   :20.88   Mean   :0.07237
##               3rd Qu.:26.00   3rd Qu.:0.08450
##               Max.   :47.50   Max.   :0.16300
```

Data exploration

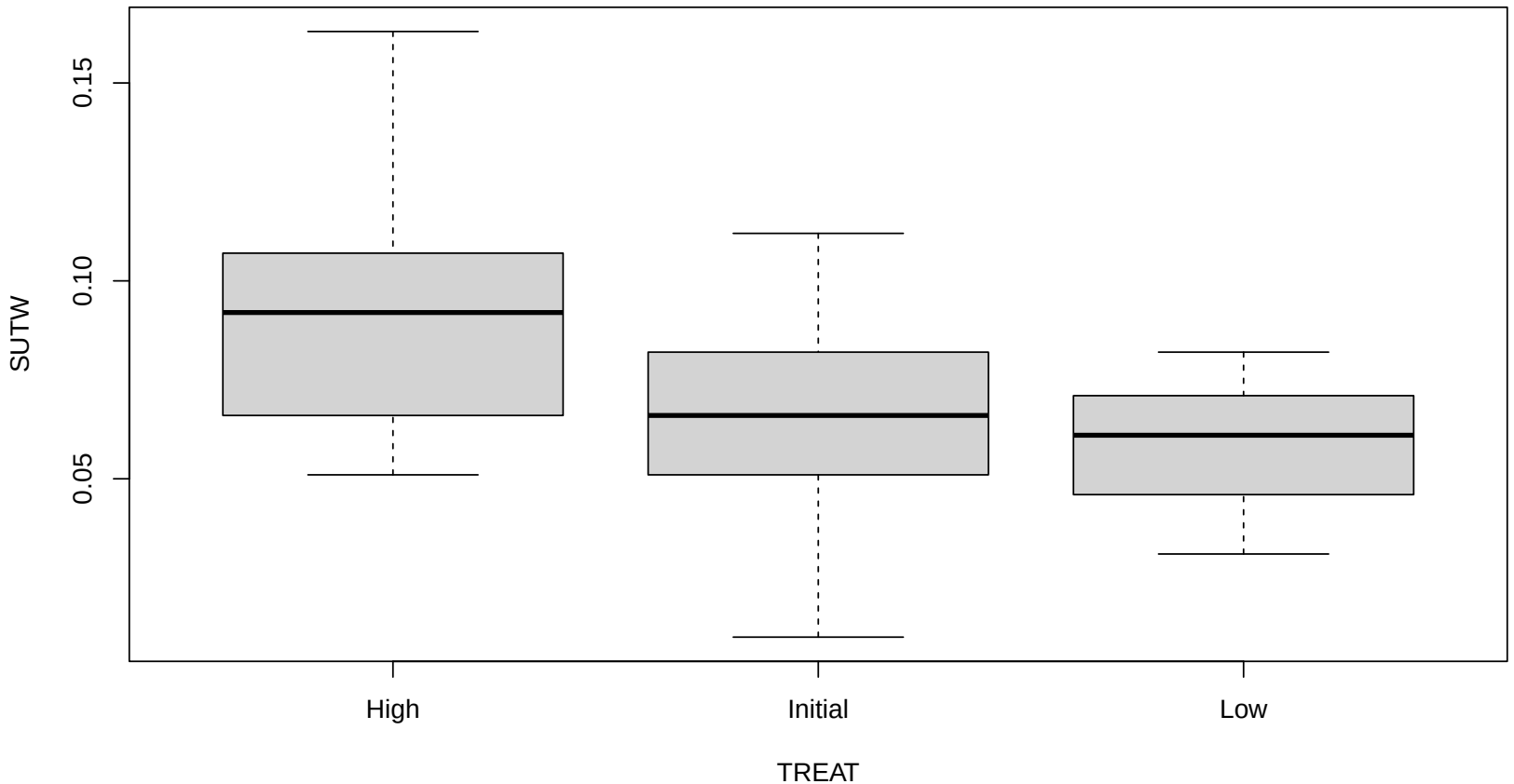
The aim of the analysis is to determine how the suture width depends on the food regime, while adjusting for size or urchin.

```
plot(constable$SUTW ~ constable$IV, xlab = "IV", ylab = "SUTW")
```



Data exploration

```
plot(constable$SUTW ~ constable$TREAT, xlab = "TREAT", ylab = "SUTW")
```



What do you see?

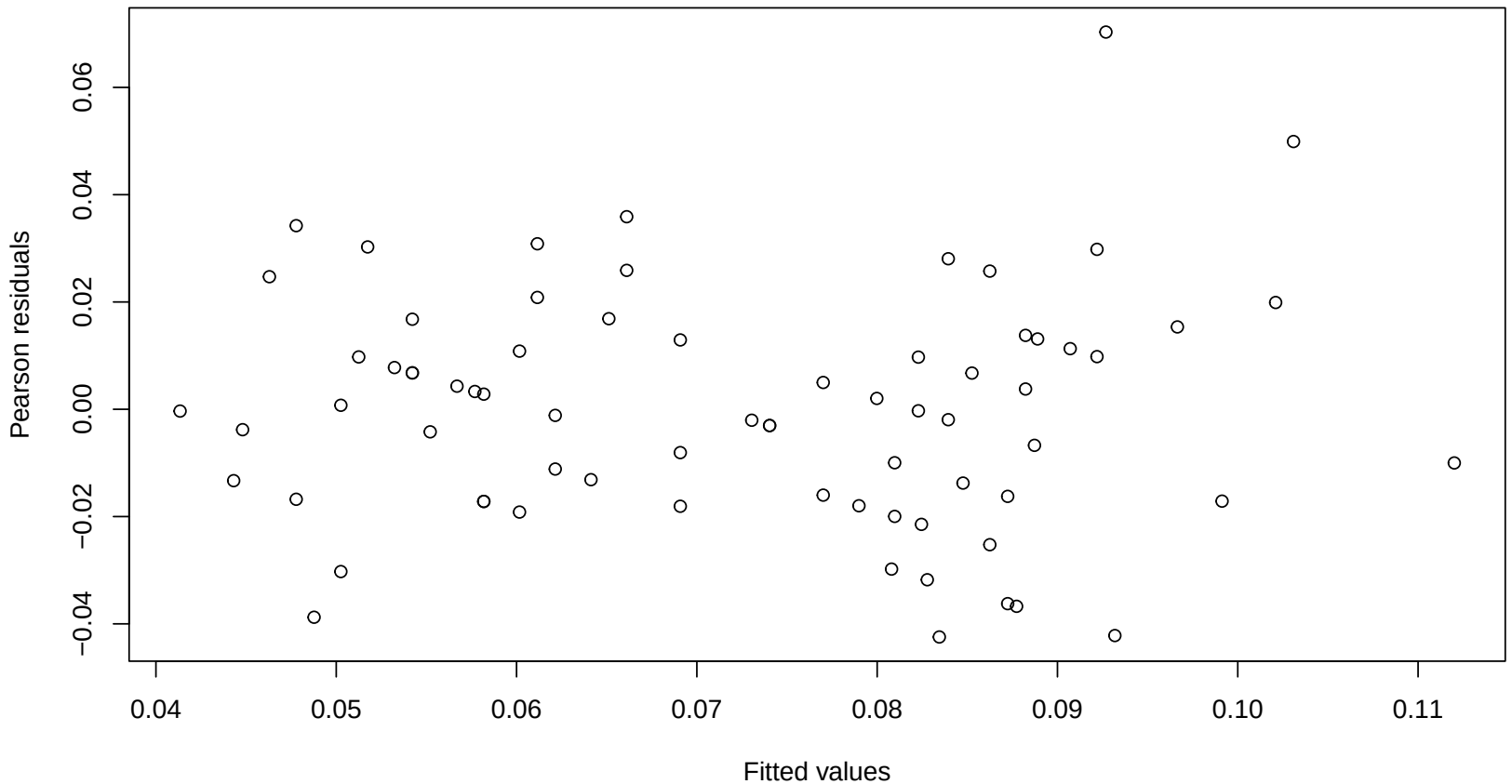
MOdel 1

```
sut.lm1 <- lm(SUTW ~ TREAT + IV, data = constable)
summary(sut.lm1)
```

```
##
## Call:
## lm(formula = SUTW ~ TREAT + IV, data = constable)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.042448 -0.016374 -0.000312  0.012964  0.070312
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0733675  0.0062029  11.828  < 2e-16 ***
## TREATInitial -0.0280645  0.0065525  -4.283 5.93e-05 ***
## TREATLow     -0.0369822  0.0066553  -5.557 4.96e-07 ***
## IV            0.0009908  0.0002423   4.089 0.000117 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02241 on 68 degrees of freedom
## Multiple R-squared:  0.3783,Adjusted R-squared:  0.3509
## F-statistic: 13.79 on 3 and 68 DF,  p-value: 4.014e-07
```

Residual plot

```
plot(residuals(sut.lm1, type = "pearson") ~ sut.lm1$fitted.values, xlab = "Fitted values",
    ylab = "Pearson residuals")
```



MOdel 2

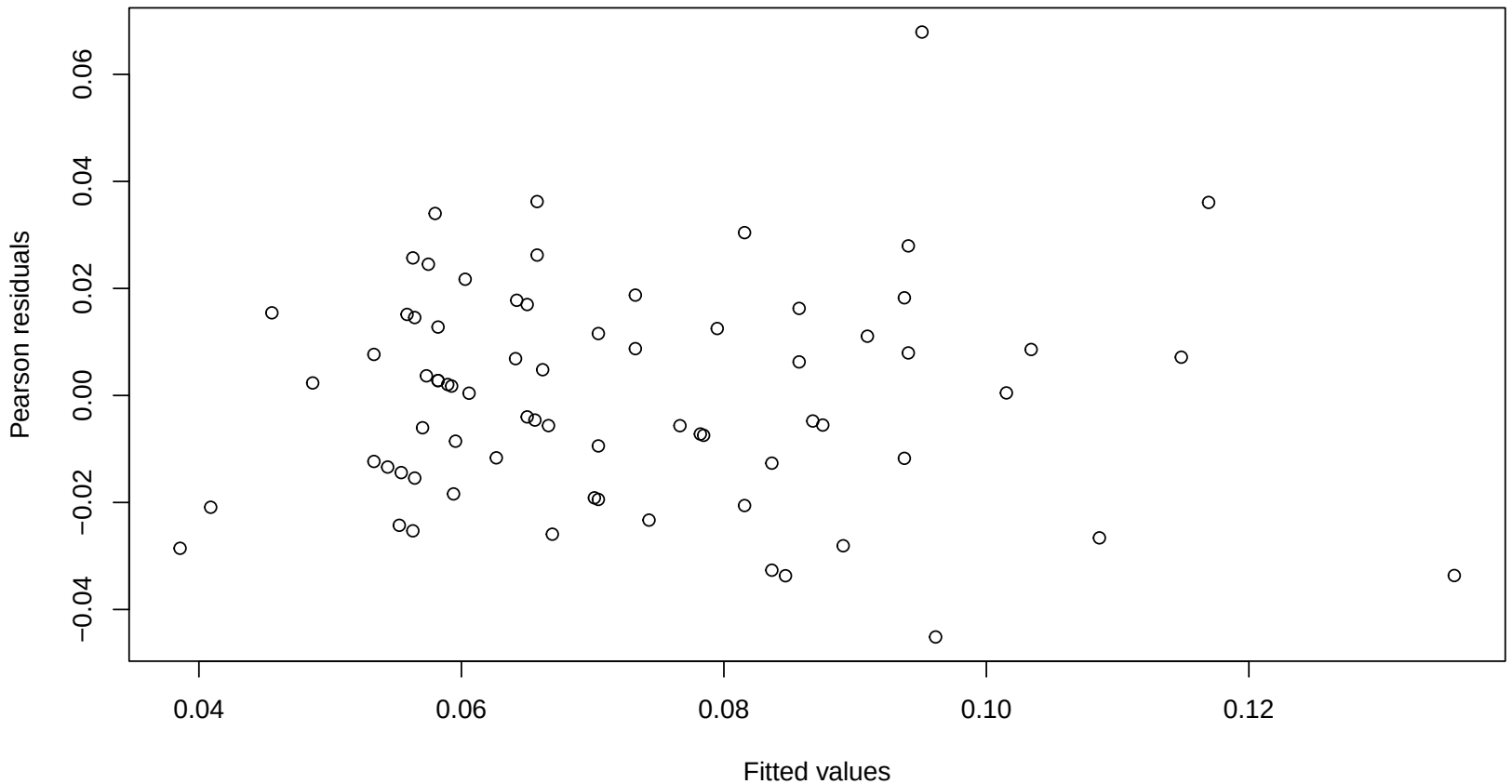
Add an interaction term.

```
sut.lm2 <- lm(SUTW ~ TREAT * IV, data = constable)
summary(sut.lm2)
```

```
##
## Call:
## lm(formula = SUTW ~ TREAT * IV, data = constable)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.045133 -0.013639  0.001111  0.013226  0.067907
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.0545327  0.0108929   5.006 4.38e-06 ***
## TREATInitial    -0.0214111  0.0145318  -1.473 0.145397
## TREATLow        -0.0016287  0.0139583  -0.117 0.907463
## IV               0.0020800  0.0005783   3.597 0.000617 ***
## TREATInitial:IV -0.0005254  0.0007020  -0.748 0.456836
## TREATLow:IV     -0.0017848  0.0006607  -2.701 0.008764 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02115 on 66 degrees of freedom
## Multiple R-squared:  0.4622,Adjusted R-squared:  0.4215
## F-statistic: 11.35 on 5 and 66 DF,  p-value: 6.424e-08
```

Residual plot

```
plot(residuals(sut.lm2, type = "pearson") ~ sut.lm2$fitted.values, xlab = "Fitted values",
    ylab = "Pearson residuals")
```



```
library(lattice)
print(with(constable, xyplot(SUTW ~ IV, groups = TREAT, type = c("p", "r"), lty = 1, col = 1:3,
    par.settings = list(superpose.symbol = list(pch = 1:3, col = 1), superpose.line = list(lty = 1:3)),
    key = list(space = "right", lty = 1, lines = T, points = T, pch = 1:3, col = 1:3,
        text = list(levels(TREAT))))))
```