R Notebook

##Interactions

A homeowner in England recorded his weekly natural gas consumption, in thousands of cubic feet, during two winter heating seasons. For the second season, cavity wall insulation had been installed.

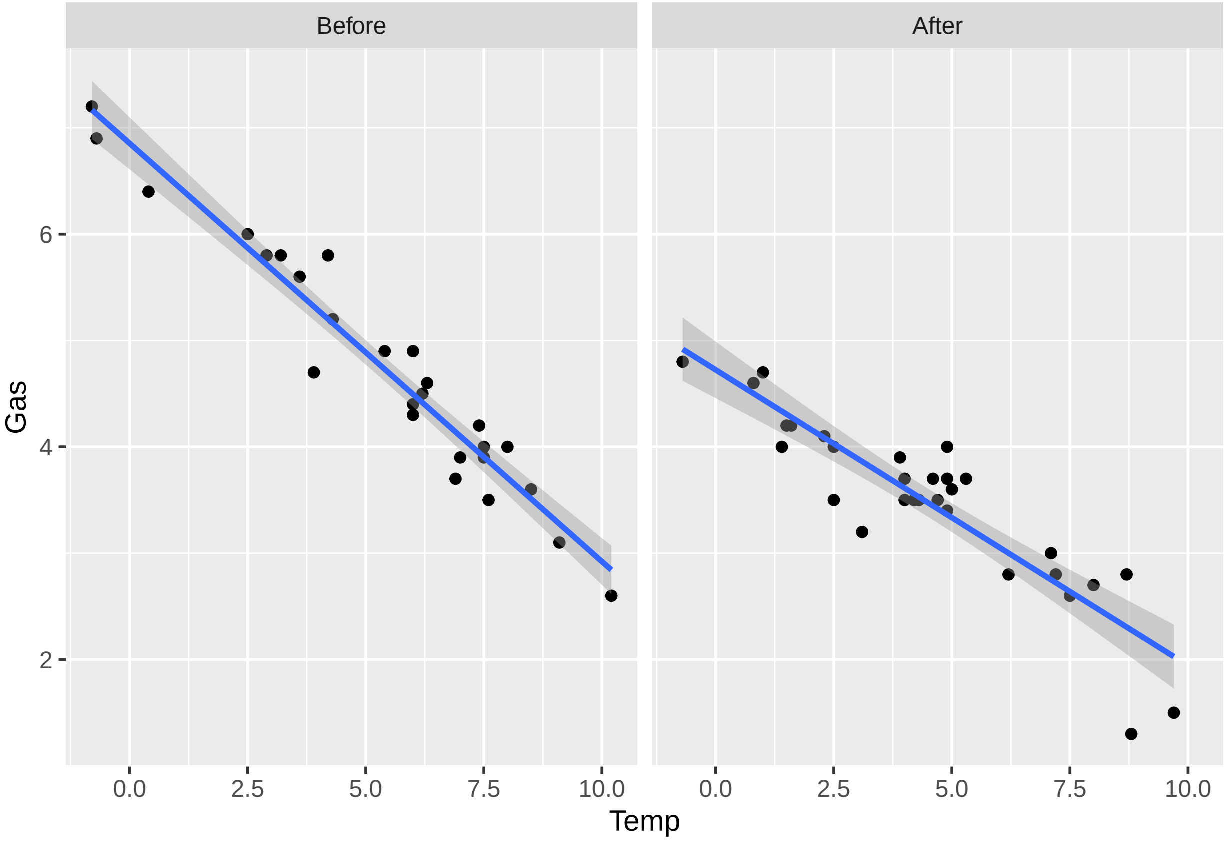
The homeowner also recorded the average weekly temperature in degrees Celsius because this would also affect gas consumption. The data may be found in the MASS package

```
data(whiteside,package="MASS")
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
ggplot(aes(x=Temp,y=Gas),data=whiteside)+geom_point()+facet_grid(~ Insul)+geom_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



We can see that less gas is used after the insulation is installed but the difference varies by temperature. The relationships appear linear so we fit a model:

```
lmod <- lm(Gas ~ Temp*Insul, whiteside)
summary(lmod)
```

```
##
## Call:
## lm(formula = Gas ~ Temp * Insul, data = whiteside)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97802 -0.18011  0.03757  0.20930  0.63803
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.85383    0.13596  50.409  < 2e-16 ***
## Temp            -0.39324    0.02249 -17.487  < 2e-16 ***
## InsulAfter      -2.12998    0.18009 -11.827 2.32e-16 ***
## Temp:InsulAfter  0.11530    0.03211   3.591 0.000731 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.323 on 52 degrees of freedom
## Multiple R-squared:  0.9277, Adjusted R-squared:  0.9235
## F-statistic: 222.3 on 3 and 52 DF,  p-value: < 2.2e-16
```

We would predict that the gas consumption would fall by 0.393 for each 1°C increase in temperature before insulation. After insulation, the fall in consumption per degree is only 0.393 - 0.115 = 0.278. But the interpretation for the other two parameter estimates is more problematic since these represent predicted consumption when the temperature is zero. This is on the lower edge of the observed range of temperatures and would not represent a typical difference. For other datasets, a continuous predictor value of zero might be far outside the range and so these parameters would have little practical meaning.The solution is to center the temperature predictor by its mean value and recompute the linear model:

```r
mean(whiteside$Temp)
```

```
## [1] 4.875
```

```r
whiteside$ctemp <- whiteside$Temp - mean(whiteside$Temp)
lmodc <- lm(Gas ~ ctemp*Insul, whiteside)
summary(lmodc)
```

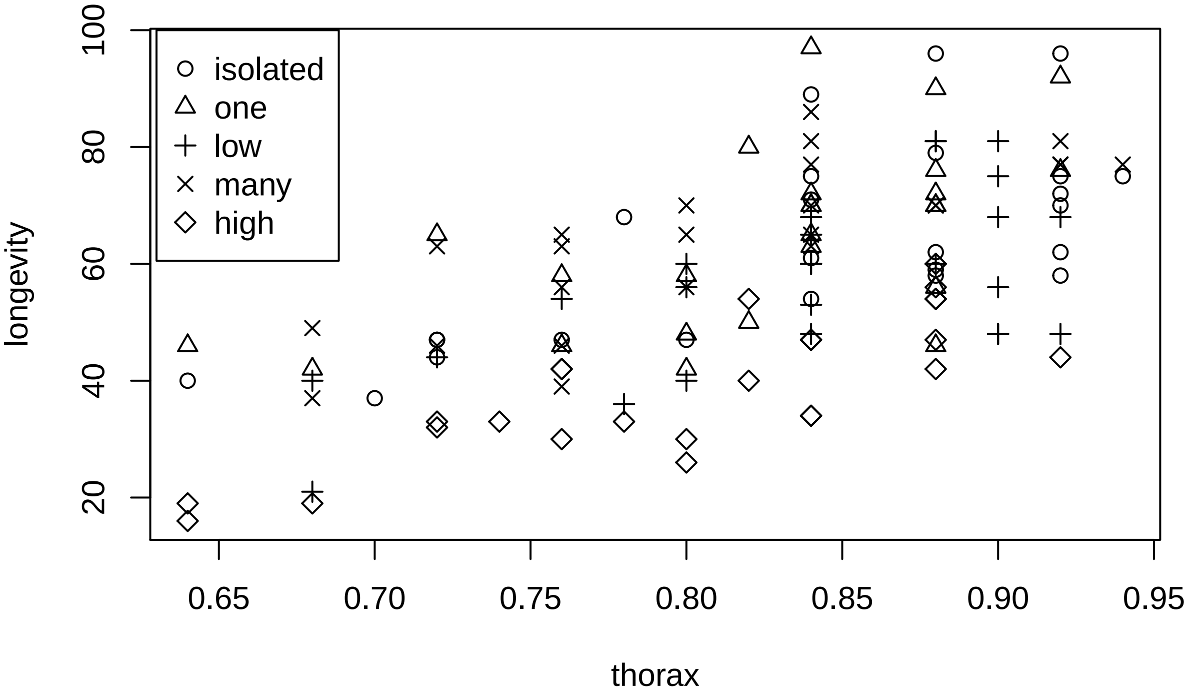
```
##
## Call:
## lm(formula = Gas ~ ctemp * Insul, data = whiteside)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97802 -0.18011  0.03757  0.20930  0.63803
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.93679    0.06424  76.848  < 2e-16 ***
## ctemp            -0.39324    0.02249 -17.487  < 2e-16 ***
## InsulAfter       -1.56787    0.08771 -17.875  < 2e-16 ***
## ctemp:InsulAfter  0.11530    0.03211   3.591 0.000731 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.323 on 52 degrees of freedom
## Multiple R-squared:  0.9277, Adjusted R-squared:  0.9235
## F-statistic: 222.3 on 3 and 52 DF,  p-value: < 2.2e-16
```

Now we can say that the average consumption before insulation at the average temperature was 4.94 and 4.94 - 1.57 = 3.37 afterwards. The other two coefficients are unchanged and their interpretation remains the same. Thus we can see that centering allows a more natural interpretation of the parameter estimates in the presence of interaction.

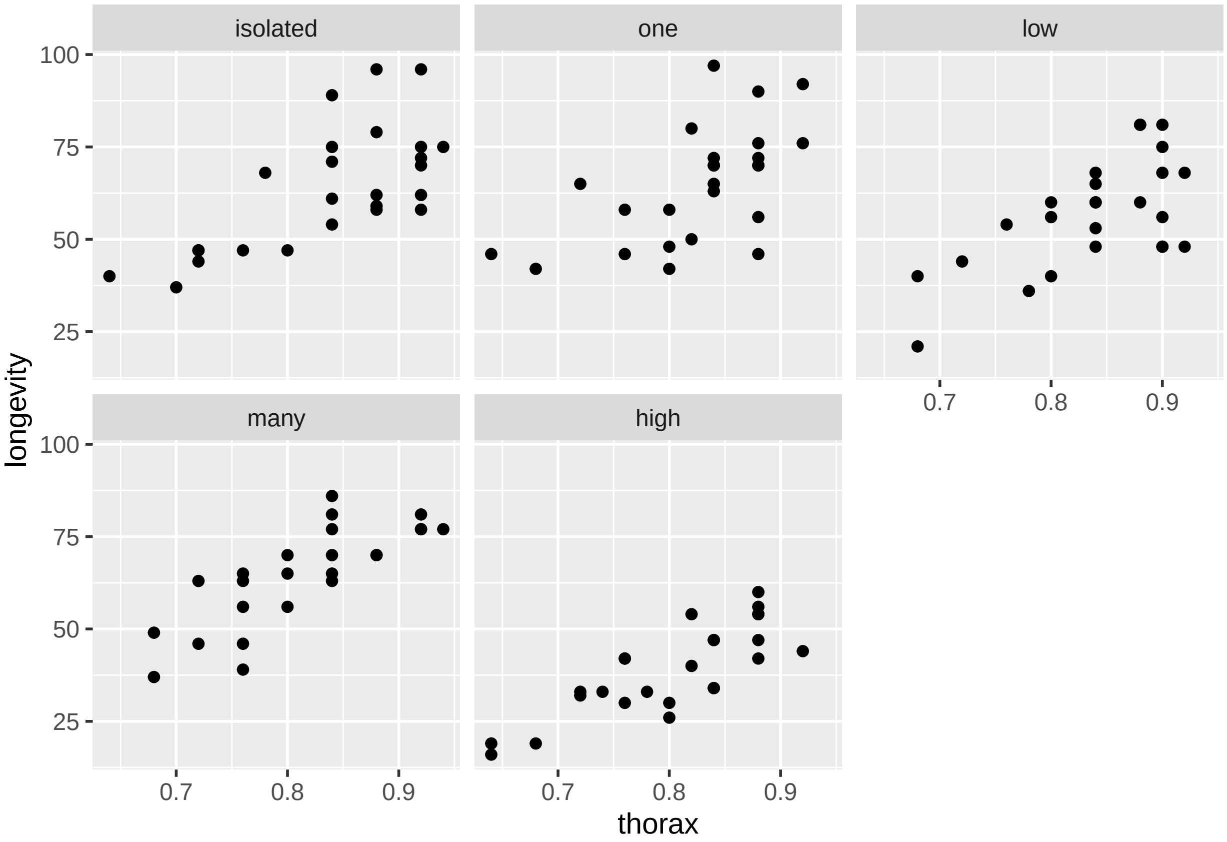
##Factors With More Than Two Levels

With multiple levels, it can be hard to distinguish the groups. Sometimes it is better to plot each level separately. This can be achieved nicely with the help of the ggplot2 package:

```
data(fruitfly,package="faraway")
plot(longevity ~ thorax, fruitfly, pch=unclass(activity))
legend(0.63,100,levels(fruitfly$activity),pch=1:5)
```



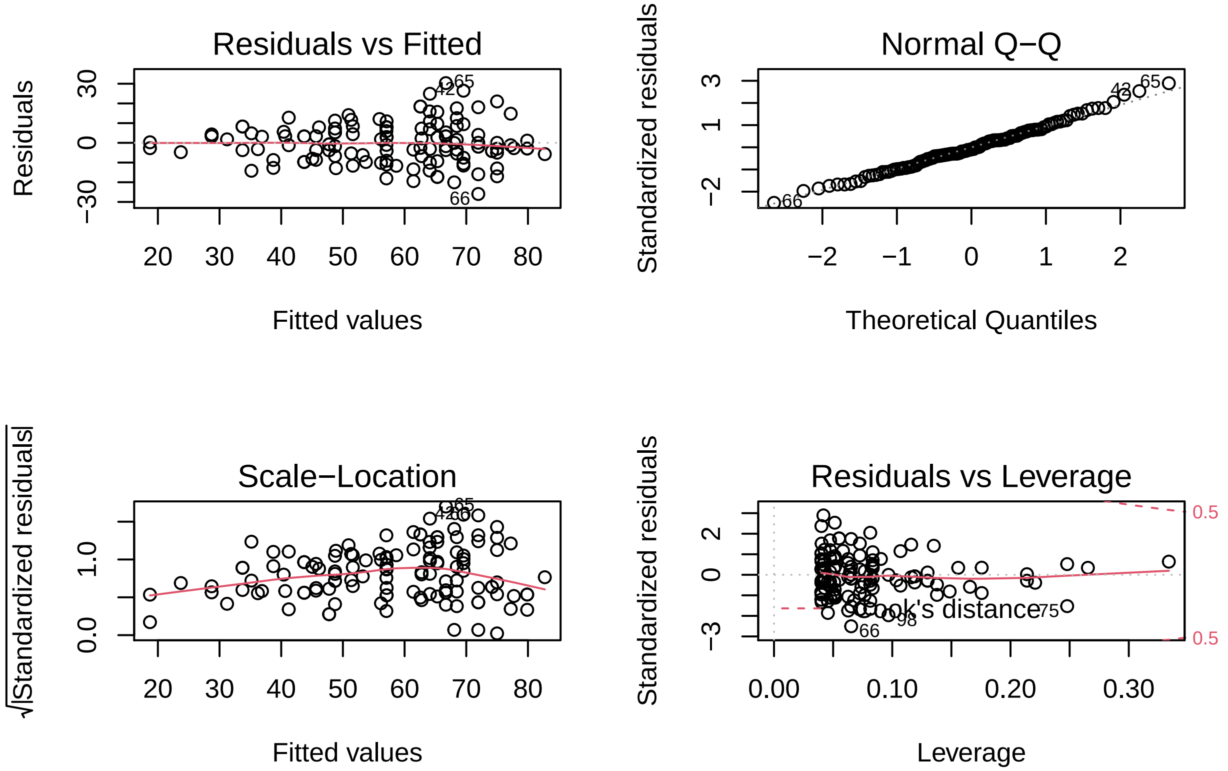
```
require(ggplot2)
ggplot(aes(x=thorax,y=longevity),data=fruitfly) + geom_point() + facet_wrap( ~ activity)
```



```r
lmod <- lm(longevity ~ thorax*activity, fruitfly)
summary(lmod)
```

```
##
## Call:
## lm(formula = longevity ~ thorax * activity, data = fruitfly)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.9509  -6.7296  -0.9103   6.1854  30.3071
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -50.2420    21.8012  -2.305    0.023 *
## thorax               136.1268    25.9517   5.245 7.27e-07 ***
## activityone            6.5172    33.8708   0.192    0.848
## activitylow           -7.7501    33.9690  -0.228    0.820
## activitymany          -1.1394    32.5298  -0.035    0.972
## activityhigh         -11.0380    31.2866  -0.353    0.725
## thorax:activityone    -4.6771    40.6518  -0.115    0.909
## thorax:activitylow     0.8743    40.4253   0.022    0.983
## thorax:activitymany    6.5478    39.3600   0.166    0.868
## thorax:activityhigh -11.1268    38.1200  -0.292    0.771
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.71 on 114 degrees of freedom
## Multiple R-squared:  0.6534, Adjusted R-squared:  0.626
## F-statistic: 23.88 on 9 and 114 DF,  p-value: < 2.2e-16
```

```
#model.matrix(lmod)
par(mfrow=c(2,2))
plot(lmod)
```



```
anova(lmod)
```

```
## Analysis of Variance Table
##
## Response: longevity
##                 Df  Sum Sq Mean Sq F value    Pr(>F)
## thorax           1 15003.3 15003.3 130.733 < 2.2e-16 ***
## activity         4  9634.6  2408.6  20.988 5.503e-13 ***
## thorax:activity  4    24.3     6.1   0.053    0.9947
## Residuals      114 13083.0   114.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(1,1))
```

The plot makes it clearer that longevity for the high activity group is lower. Since “isolated” is the reference level, the fitted regression line within this group is longevity= -50.2 + 136.1*thorax. For “many,” it is longevity= (-50.2 -1.1) + (136.1 + 6.5)*thorax. Similar calculations can be made for the other groups.

```
head(model.matrix(lmod))
```

```
##   (Intercept) thorax activityone activitylow activitymany activityhigh
## 1           1   0.68           0           0            1            0
## 2           1   0.68           0           0            1            0
## 3           1   0.72           0           0            1            0
## 4           1   0.72           0           0            1            0
## 5           1   0.76           0           0            1            0
## 6           1   0.76           0           0            1            0
##   thorax:activityone thorax:activitylow thorax:activitymany thorax:activityhigh
## 1                  0                  0                0.68                   0
## 2                  0                  0                0.68                   0
## 3                  0                  0                0.72                   0
## 4                  0                  0                0.72                   0
## 5                  0                  0                0.76                   0
## 6                  0                  0                0.76                   0
```

There is perhaps some heteroscedasticity, but we will let this be until later for ease of presentation. Now we see whether the model can be simplified. The model summary output is not suitable for this purpose because there are four t-tests corresponding to the interaction term while we want just a single test for this term.

```
anova(lmod)
```

```
## Analysis of Variance Table
##
## Response: longevity
##                  Df  Sum Sq Mean Sq F value    Pr(>F)
## thorax            1 15003.3 15003.3 130.733 < 2.2e-16 ***
## activity          4  9634.6  2408.6  20.988 5.503e-13 ***
## thorax:activity   4    24.3     6.1   0.053    0.9947
## Residuals       114 13083.0   114.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is a sequential analysis of variance (ANOVA) table. Starting from a null model, terms are added and sequentially tested. The interaction term thorax:activity is not significant, indicating that we can fit the same slope within each group. No further simplification is possible.

We notice that the F-statistic for the test of the interaction term is very small and its p-value close to one. For these data, the fitted regression lines to the five groups happen to be very close to parallel. This can, of course, just happen by chance. In some other cases, unusually large p-values have been used as evidence that data have been tampered with or “cleaned” to improve the fit.

```
lmodp <- lm(longevity ~ thorax+activity, fruitfly)
drop1(lmodp,test="F")
```

```
## Single term deletions
##
## Model:
## longevity ~ thorax + activity
##          Df Sum of Sq   RSS    AIC F value    Pr(>F)
## <none>                13107 589.92
## thorax    1   12368.4 25476 670.32 111.348 < 2.2e-16 ***
## activity  4    9634.6 22742 650.25  21.684 1.974e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

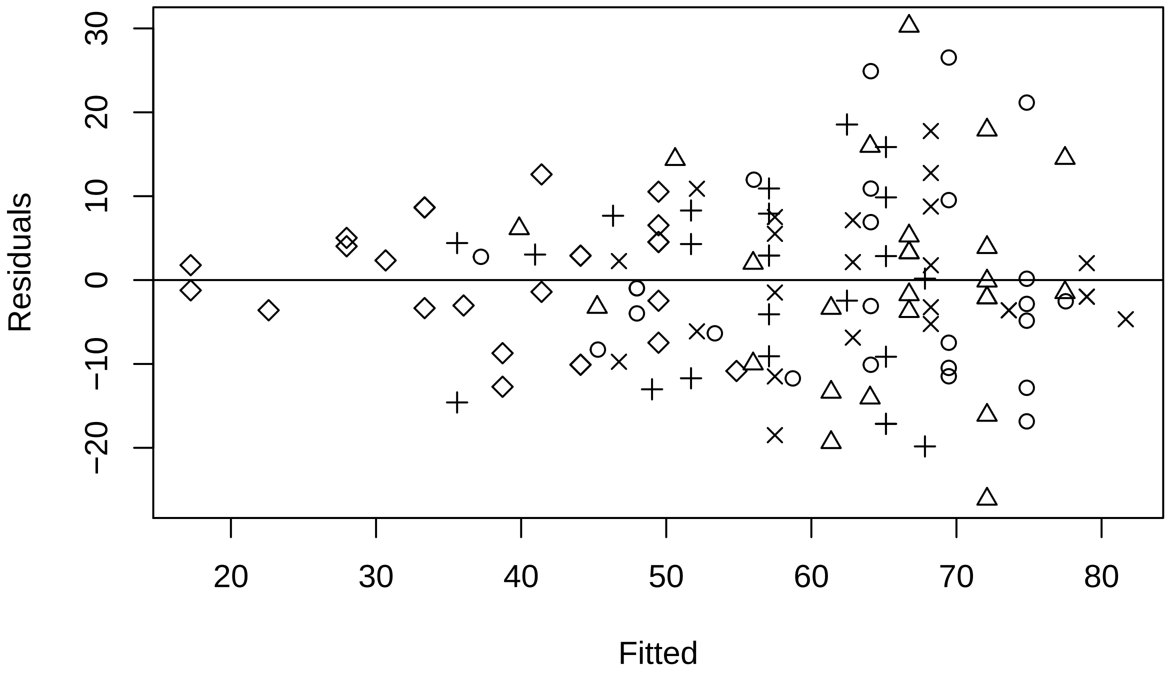
```
summary(lmodp)
```

```
##
## Call:
## lm(formula = longevity ~ thorax + activity, data = fruitfly)
##
## Residuals:
```

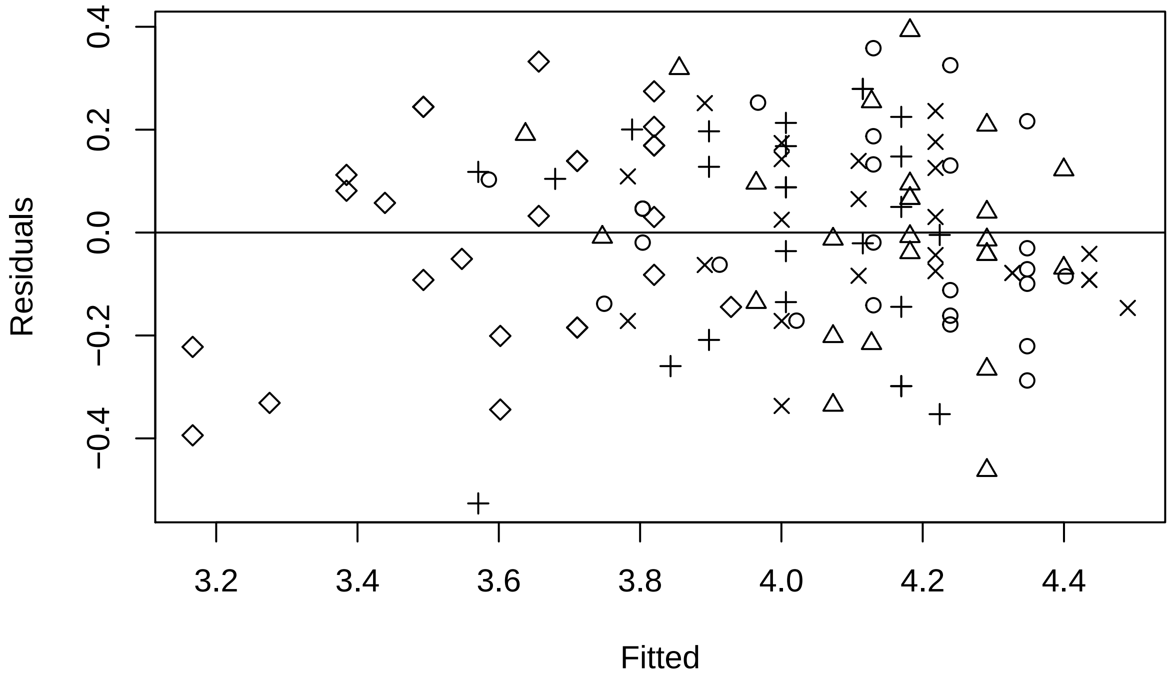
```
##      Min       1Q   Median       3Q      Max
## -26.108  -7.014  -1.101   6.234  30.265
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -48.749     10.850  -4.493 1.65e-05 ***
## thorax        134.341     12.731  10.552  < 2e-16 ***
## activityone     2.637      2.984   0.884   0.3786
## activitylow    -7.015      2.981  -2.353   0.0203 *
## activitymany    4.139      3.027   1.367   0.1741
## activityhigh  -20.004      3.016  -6.632 1.05e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.54 on 118 degrees of freedom
## Multiple R-squared:  0.6527, Adjusted R-squared:  0.638
## F-statistic: 44.36 on 5 and 118 DF,  p-value: < 2.2e-16
```

Returning to the diagnostics: A log transformation can remove the heteroscedasticity:

```
plot(residuals(lmodp) ~fitted(lmodp),pch=unclass(fruitfly$activity),xlab="Fitted",ylab="Residuals")
abline(h=0)
```



```
lmodl <- lm(log(longevity) ~ thorax+activity, fruitfly)
plot(residuals(lmodl) ~ fitted(lmodl),pch=unclass(fruitfly$activity), xlab="Fitted",ylab="Residuals")
abline(h=0)
```



```
summary(lmodl)
```

```
##
## Call:
## lm(formula = log(longevity) ~ thorax + activity, data = fruitfly)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52641 -0.13629 -0.00823  0.13918  0.39273
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.84421    0.19882   9.276 1.04e-15 ***
## thorax        2.72146    0.23329  11.666  < 2e-16 ***
## activityone   0.05174    0.05468   0.946   0.3459
## activitylow  -0.12387    0.05463  -2.268   0.0252 *
## activitymany  0.08791    0.05546   1.585   0.1156
## activityhigh -0.41925    0.05527  -7.586 8.35e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1931 on 118 degrees of freedom
## Multiple R-squared:  0.7025, Adjusted R-squared:  0.6899
## F-statistic: 55.72 on 5 and 118 DF,  p-value: < 2.2e-16
```

```
exp(coef(lmodl)[3:6])
```

```
##  activityone  activitylow activitymany activityhigh
##    1.0531064    0.8834971    1.0918894    0.6575384
```

```
lmodh <- lm(thorax ~ activity, fruitfly)
anova(lmodh)
```

```
## Analysis of Variance Table
```

```
##
## Response: thorax
##            Df  Sum Sq  Mean Sq F value Pr(>F)
## activity    4 0.02555 0.006388  1.1092 0.3555
## Residuals 119 0.68532 0.005759
```

```
lmodu <- lm(log(longevity) ~ activity, fruitfly)
summary(lmodu)
```

```
##
## Call:
## lm(formula = log(longevity) ~ activity, data = fruitfly)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95531 -0.13187  0.03108  0.19814  0.49222
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.11935    0.05644  72.986  < 2e-16 ***
## activityone   0.02344    0.07982   0.294    0.770
## activitylow  -0.11951    0.07982  -1.497    0.137
## activitymany  0.02396    0.08065   0.297    0.767
## activityhigh -0.51722    0.07982  -6.480 2.17e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2822 on 119 degrees of freedom
## Multiple R-squared:  0.3594, Adjusted R-squared:  0.3378
## F-statistic: 16.69 on 4 and 119 DF,  p-value: 6.963e-11
```

```
#different coding
contr.treatment(4)
```

```
##   2 3 4
## 1 0 0 0
## 2 1 0 0
## 3 0 1 0
## 4 0 0 1
```

```
contr.helmert(4)
```

```
##   [,1] [,2] [,3]
## 1   -1   -1   -1
## 2    1   -1   -1
## 3    0    2   -1
## 4    0    0    3
```

```
contr.sum(4)
```

```
##   [,1] [,2] [,3]
## 1    1    0    0
## 2    0    1    0
## 3    0    0    1
## 4   -1   -1   -1
```

```
data(sexab,package="faraway")
#help(sexab,package="faraway")
```

```r
contrasts(sexab$csa) <- contr.sum(2)
summary(lm(ptsd ~ csa, sexab))
```

```
##
## Call:
## lm(formula = ptsd ~ csa, data = sexab)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0451 -2.3123  0.0951  2.1645  7.0514
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.3185     0.4053  20.526  < 2e-16 ***
## csa1          3.6226     0.4053   8.939 2.17e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.473 on 74 degrees of freedom
## Multiple R-squared:  0.5192, Adjusted R-squared:  0.5127
## F-statistic:  79.9 on 1 and 74 DF,  p-value: 2.172e-13
```