Eric_Hirsch_605_Discussion_12

Eric Hirsch

11/10/2021

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.0     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Discussion 12 - Using Quadratic, Dichotomous, and Dichotomous*quantitative terms

For this exercise I looked for data that would benefit from the introduction of a quadratic term, i.e. that would have at least one relationship which was parabolic. The dataset I chose is from kaggle and looks at drug use by age category - “https://www.kaggle.com/fivethirtyeight/fivethirtyeight-drug-use-by-age-dataset (https://www.kaggle.com/fivethirtyeight/fivethirtyeight-drug-use-by-age-dataset).” I reasoned that drug use would be lowest at the ends of the age spectrum and highest in the middle.

The dataset had some problems - but because it was for an exercise and not a work project I was willing to make some compromises. I converted age categories to the mean age in each category. I also did not concern myself with the fact that it was aggregated data and the categories had different n’s. Finally, I could not construct a dichotomous term that wasn’t correlated with the other independent variable (alcohol use), so I chose the one least correlated (stimulant use).

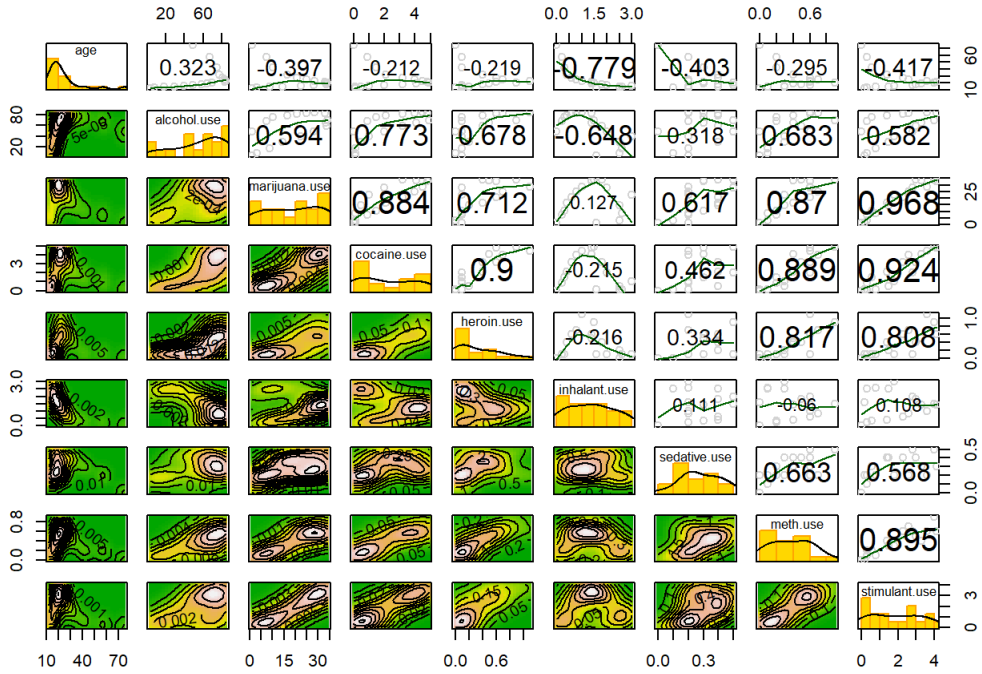
```
dfT <- as.data.frame(read.delim("D:\\RStudio\\CUNY_605\\12\\druguse.csv", header = TRUE, stringsAsFactors = FALSE, sep=","))

dfDrugs <- dfT %>% mutate(age=recode(age,
                         `22-23`='22.5',
                         `24-25`='24.5',
                         '26-29' = '27.5',
                         '30-34'='32',
                         '35-49'='42',
                         '50-64'='57',
                         '65+'='75')) %>%
                  mutate(age = as.numeric(age))
```

```
dfSelectedDrugs <- dfDrugs %>%
  select(age, alcohol.use, marijuana.use, cocaine.use, heroin.use, inhalant.use, sedative.use, meth.use, stimulant.use )
library(ResourceSelection)
```

```
## Warning: package 'ResourceSelection' was built under R version 4.0.5
```

```
## ResourceSelection 0.3-5   2019-07-22
```

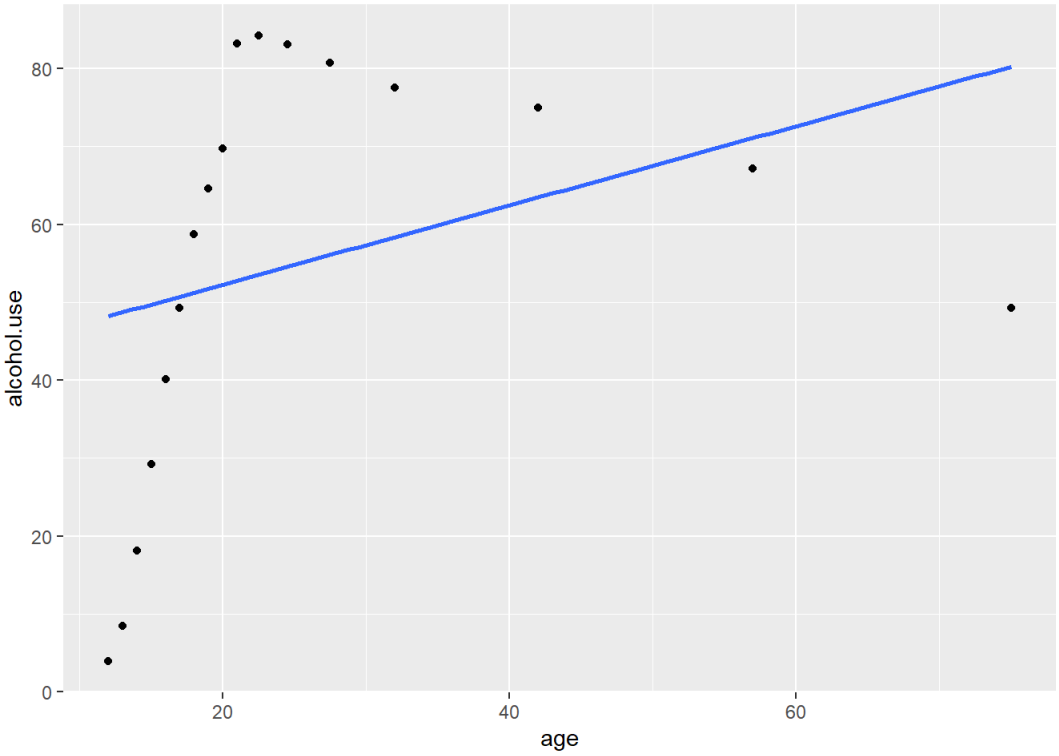
```
kdepairs(dfSelectedDrugs)
```



Below is my analysis of age vs % of individuals that age who use alcohol. The overall p is .2 and the R2 is .04. This regression shows no apparent relationship. The scatterplot, however, shows that they are highly correlated - the plot is not quite parabolic, however - there may be a log transformation needed as well.

```
ggplot(dfDrugs, aes(age, alcohol.use)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE)
```

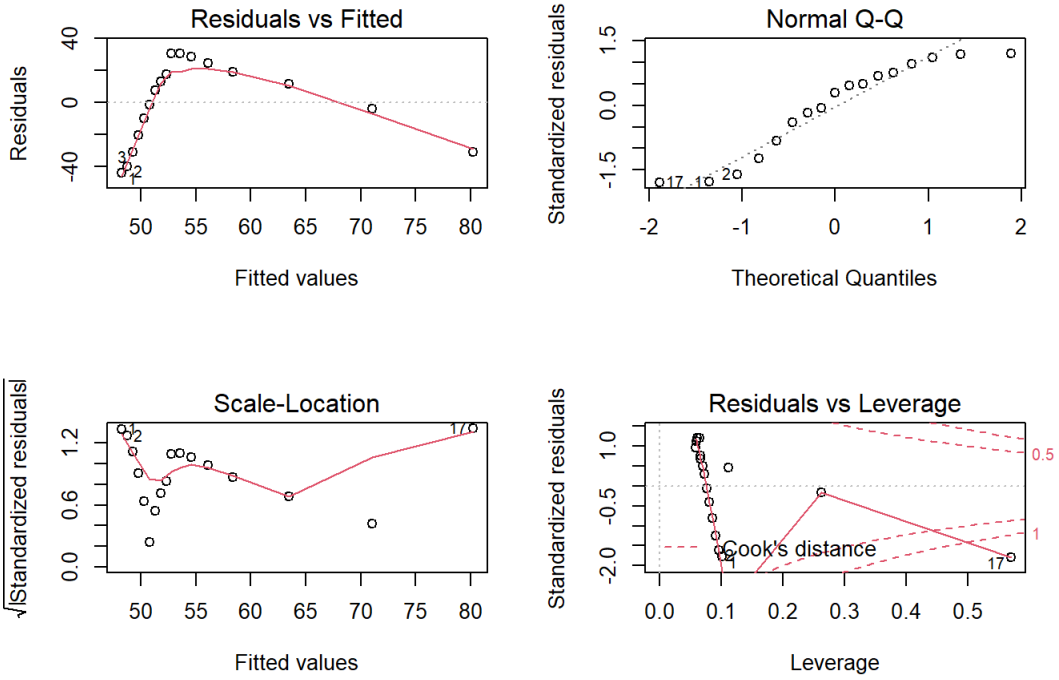
```
## `geom_smooth()` using formula 'y ~ x'
```




```
m1 <- lm(alcohol.use ~ age, data = dfDrugs)
summary(m1)
```

```
##
## Call:
## lm(formula = alcohol.use ~ age, data = dfDrugs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.305 -20.530   7.444  19.124  30.655
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  42.1019    11.9356   3.527  0.00305 **
## age           0.5086     0.3851   1.321  0.20643
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.27 on 15 degrees of freedom
## Multiple R-squared:  0.1042, Adjusted R-squared:  0.04443
## F-statistic: 1.744 on 1 and 15 DF,  p-value: 0.2064
```

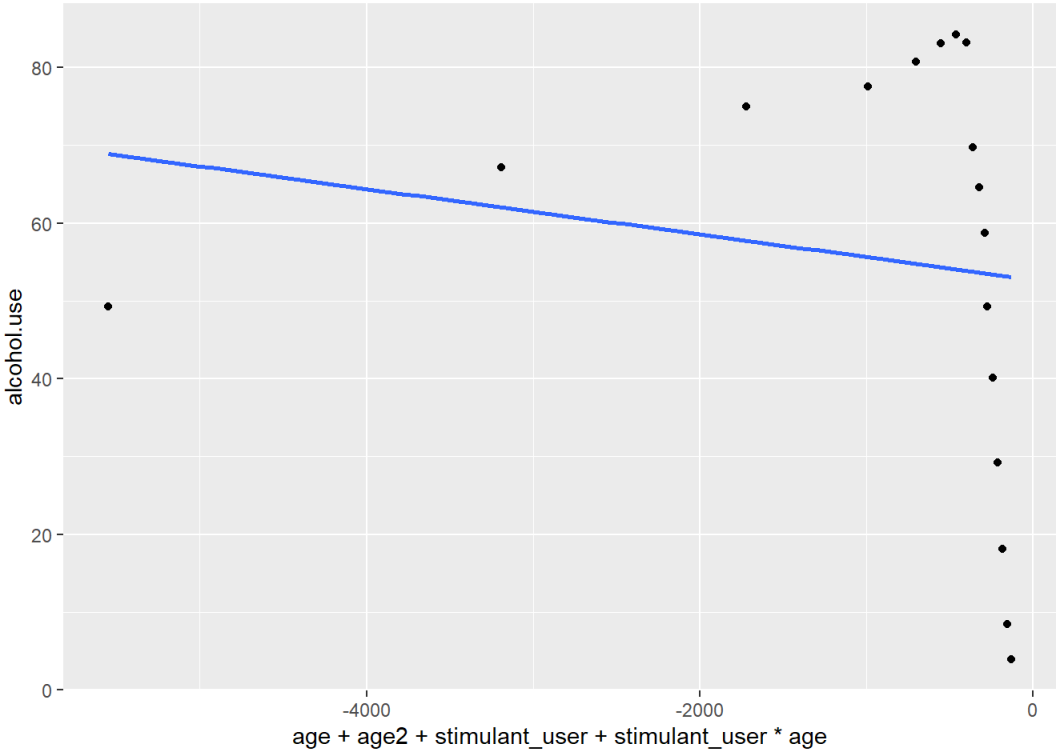
```
par(mfrow=c(2,2))
plot(m1)
```



In this version we add our new terms - age squared, stimulant_user, and age*stimulant_user. The difference is dramatic - p approaches 0, and the R2 is .88. Age and age squared are both significant with p near 0, stimulant_user is not. The residuals are closer to normal, and there is a bit less heteroskedasticity, but there is still quite a lot.

```
stimulant_mean <- mean(dfDrugs$cocaine.use)

dfDrugs2 <- dfDrugs %>% mutate(age2 = -1*age^2, age3=age^3) %>%
  mutate(stimulant_user = case_when(cocaine.use < stimulant_mean ~ 0, cocaine.use >= stimulant_mean ~ 1))


ggplot(dfDrugs2, aes(age + age2 + stimulant_user + stimulant_user*age, alcohol.use)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE)
```

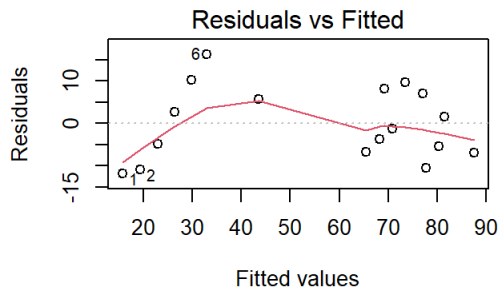
```
## `geom_smooth()` using formula 'y ~ x'
```

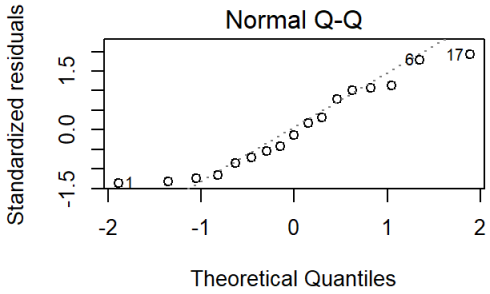


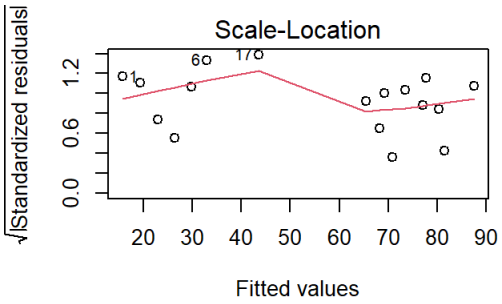

```
m3 <- lm(alcohol.use ~ age + age2 + stimulant_user + stimulant_user*age,  data = dfDrugs2)
summary(m3)
```

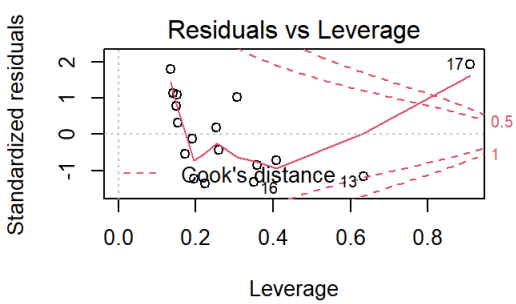
```
##
## Call:
## lm(formula = alcohol.use ~ age + age2 + stimulant_user + stimulant_user *
##     age, data = dfDrugs2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.816  -6.741  -1.177   7.113  16.311
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -36.333221  11.251306  -3.229  0.00723 **
## age                  4.960556   0.755126   6.569 2.65e-05 ***
## age2                 0.051929   0.009115   5.697 9.96e-05 ***
## stimulant_user      34.161240  27.564854   1.239  0.23892
## age:stimulant_user  -0.269555   1.257423  -0.214  0.83386
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.823 on 12 degrees of freedom
## Multiple R-squared:  0.8998, Adjusted R-squared:  0.8664
## F-statistic: 26.95 on 4 and 12 DF,  p-value: 6.466e-06
```

```
par(mfrow=c(2,2))
plot(m3)
```



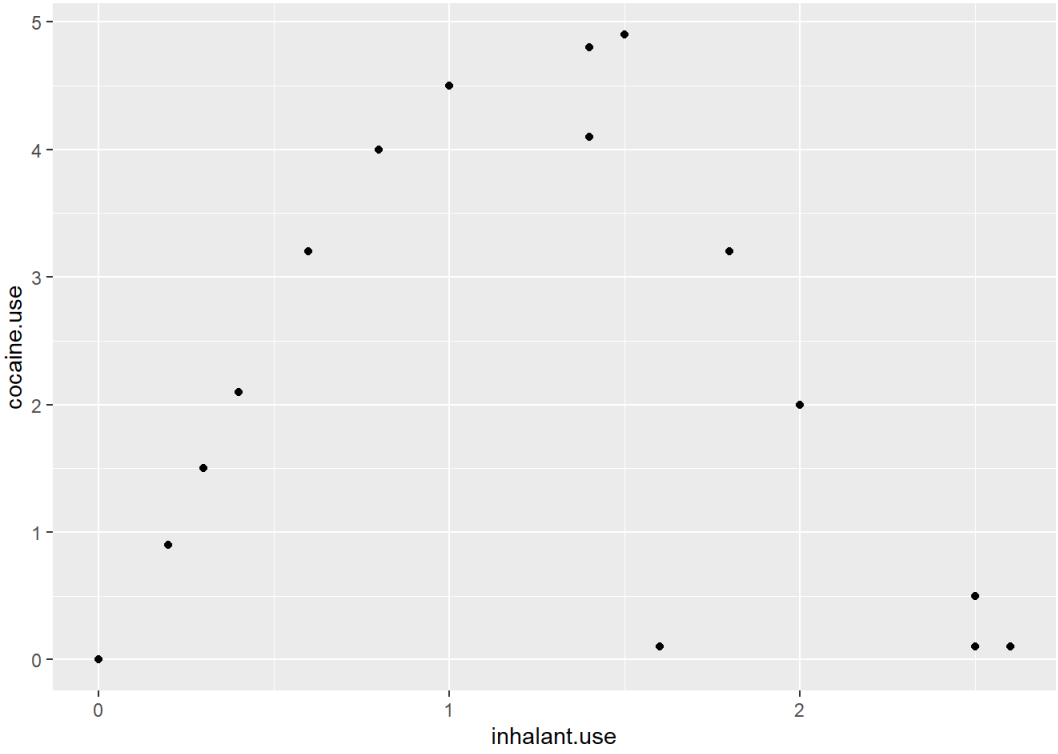






Noting from the kde plots above that cocaine use and inhalant use are quite parabolically related (as inhalant use goes up, so does cocaine use, but as inhalant use continues to go up, cocaine use begins to fall), I tried the analysis with these variables. Here everything works perfectly with the quadratic - the residuals normalize, heteroskedasticy disappears, R2 rises from .02 to .68, overall p from .56 to near 0 and all coefficients highly significant.

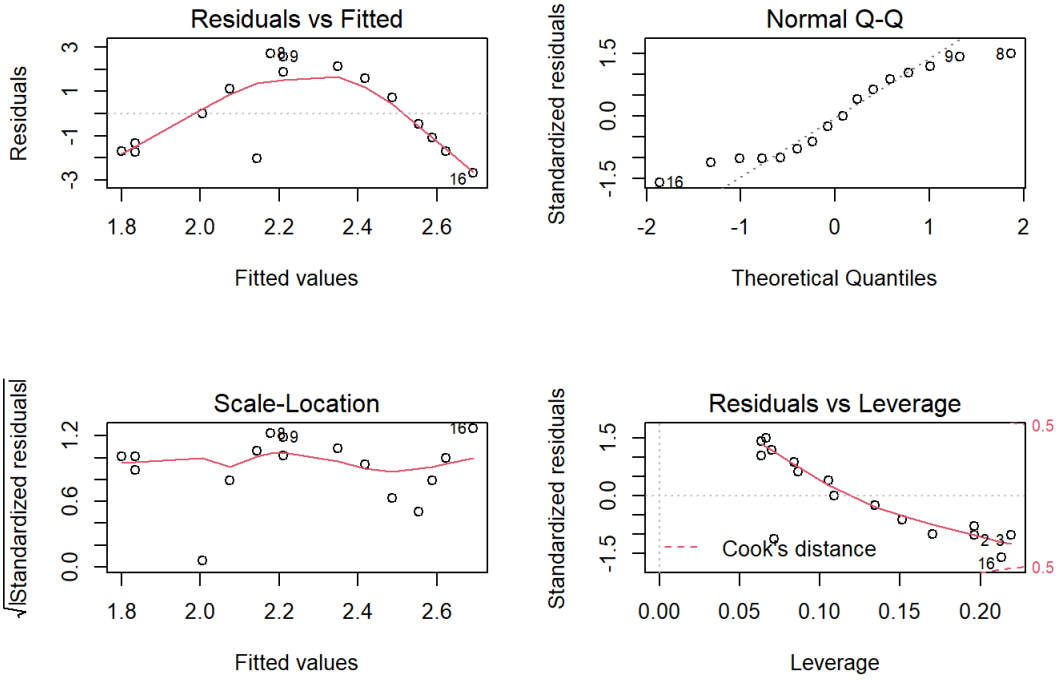
```
dfDrugs3 <- dfDrugs %>%
  mutate(inhalant2 = inhalant.use^2) %>%
  filter(!row_number()==5) %>%
  mutate(inhalantPlus = inhalant.use + inhalant2)


ggplot(dfDrugs3, aes(y=cocaine.use, x=inhalant.use)) +
  geom_point()
```




```
m3 <- lm(cocaine.use ~ inhalant.use, dfDrugs3)
summary(m3)
```

```
##
## Call:
## lm(formula = cocaine.use ~ inhalant.use, data = dfDrugs3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.691 -1.706 -0.230  1.659  2.723
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.6914     0.8731   3.083  0.00811 **
## inhalant.use  -0.3428     0.5701  -0.601  0.55727
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.891 on 14 degrees of freedom
## Multiple R-squared:  0.02517,    Adjusted R-squared:  -0.04446
## F-statistic: 0.3615 on 1 and 14 DF,  p-value: 0.5573
```

```
par(mfrow=c(2,2))
plot(m3)
```




```
m3 <- lm(cocaine.use ~ inhalant.use + inhalant2, dfDrugs3)
summary(m3)
```

```
##
## Call:
## lm(formula = cocaine.use ~ inhalant.use + inhalant2, data = dfDrugs3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4896 -0.1901  0.1548  0.5366  1.1697
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.1749     0.7155   0.245 0.810654
## inhalant.use   5.9107     1.2655   4.671 0.000438 ***
## inhalant2     -2.3603     0.4601  -5.131 0.000193 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.128 on 13 degrees of freedom
## Multiple R-squared:  0.6777, Adjusted R-squared:  0.6281
## F-statistic: 13.67 on 2 and 13 DF,  p-value: 0.0006361
```

```
par(mfrow=c(2,2))
plot(m3)
```