Homework2.R

dbarron

Mon Feb 05 11:06:26 2018

```
#########################################################
### Solutions to week 2 homework
##################################################
library(car)
library(effects)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'carData'
```

```
## The following objects are masked from 'package:car':
##
##     Guyer, UN, Vocab
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
data(SLID)

## This gives information about the variables
help(SLID)
```

```
## starting httpd help server ...
```

```
##  done
```

```
str(SLID)
```

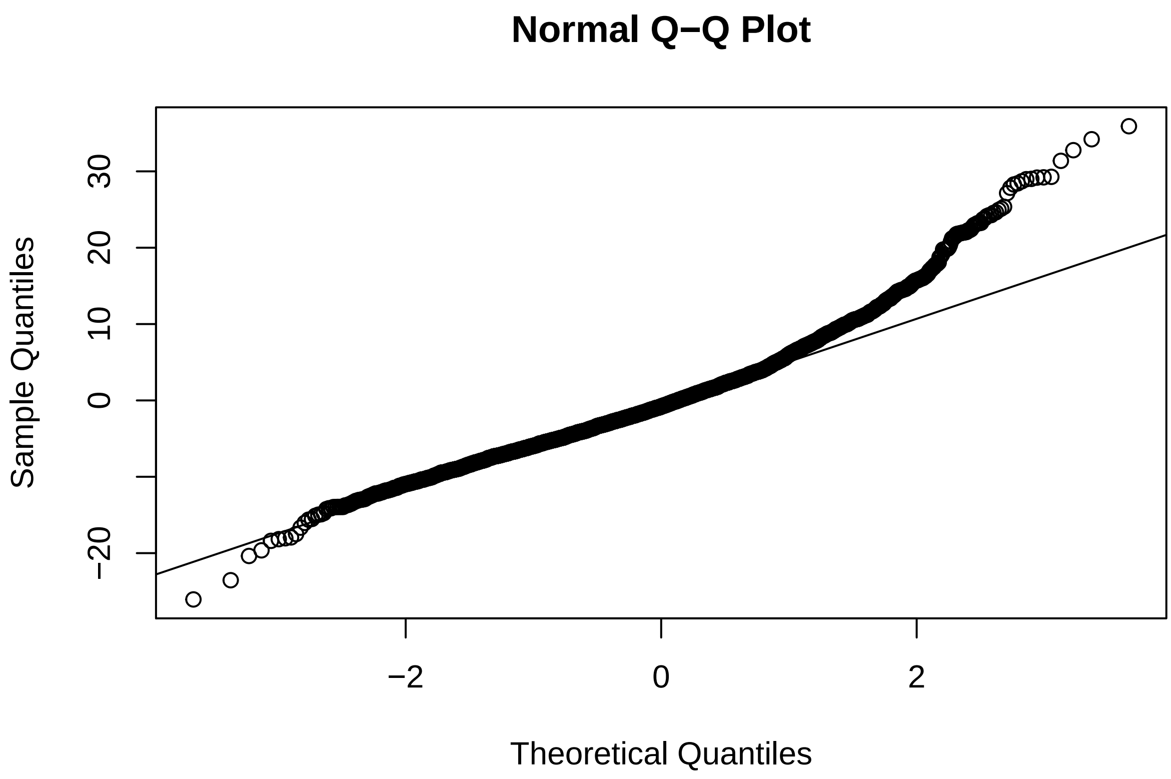
```
## 'data.frame':    7425 obs. of  5 variables:
##  $ wages    : num  10.6 11 NA 17.8 NA ...
##  $ education: num  15 13.2 16 14 8 16 12 14.5 15 10 ...
##  $ age      : int  40 19 49 46 71 50 70 42 31 56 ...
##  $ sex      : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 1 1 1 2 1 ...
##  $ language : Factor w/ 3 levels "English","French",..: 1 1 3 3 1 1 1 1 1 1 ...
```

```
## Shows two numeric variables (with wages having missing cases), one integer
# variable and two factors

# First attempt

l1 <- lm(wages ~ education + age + sex + language, data=SLID)
summary(l1)
```

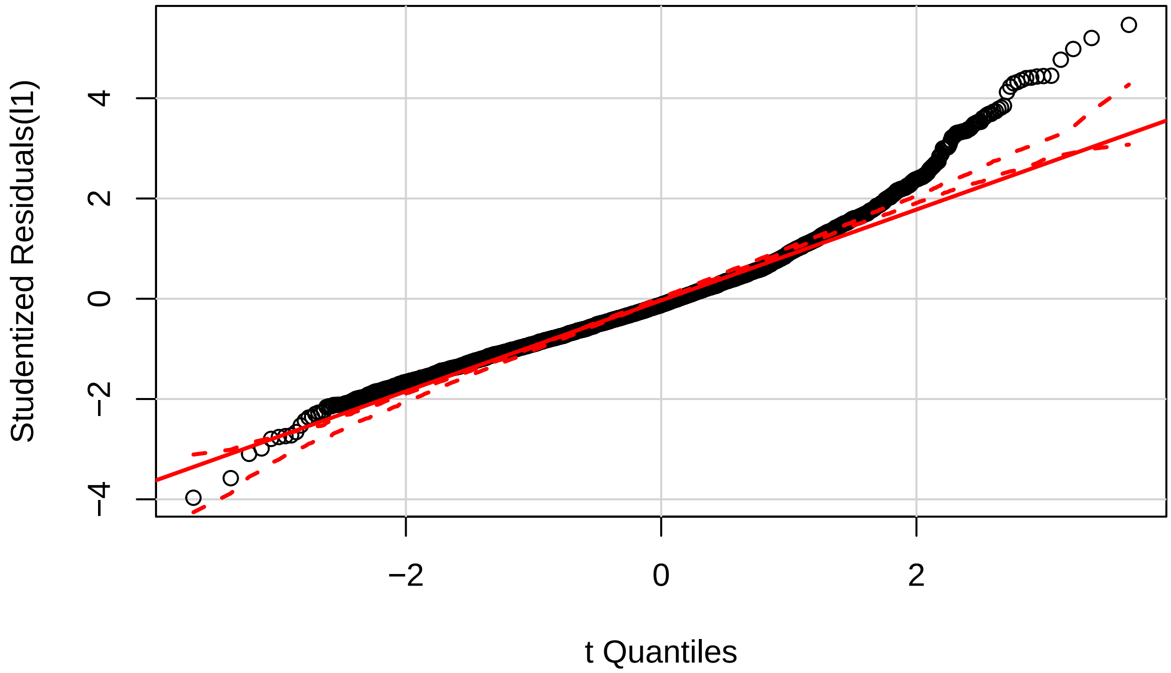
```
##
## Call:
## lm(formula = wages ~ education + age + sex + language, data = SLID)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.062  -4.347  -0.797   3.237  35.908
##
```

```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -7.888779   0.612263 -12.885   <2e-16 ***
## education        0.916614   0.034762  26.368   <2e-16 ***
## age              0.255137   0.008714  29.278   <2e-16 ***
## sexMale          3.455411   0.209195  16.518   <2e-16 ***
## languageFrench -0.015223   0.426732  -0.036    0.972
## languageOther    0.142605   0.325058   0.439    0.661
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.6 on 3981 degrees of freedom
##   (3438 observations deleted due to missingness)
## Multiple R-squared:  0.2973, Adjusted R-squared:  0.2964
## F-statistic: 336.8 on 5 and 3981 DF,  p-value: < 2.2e-16
```

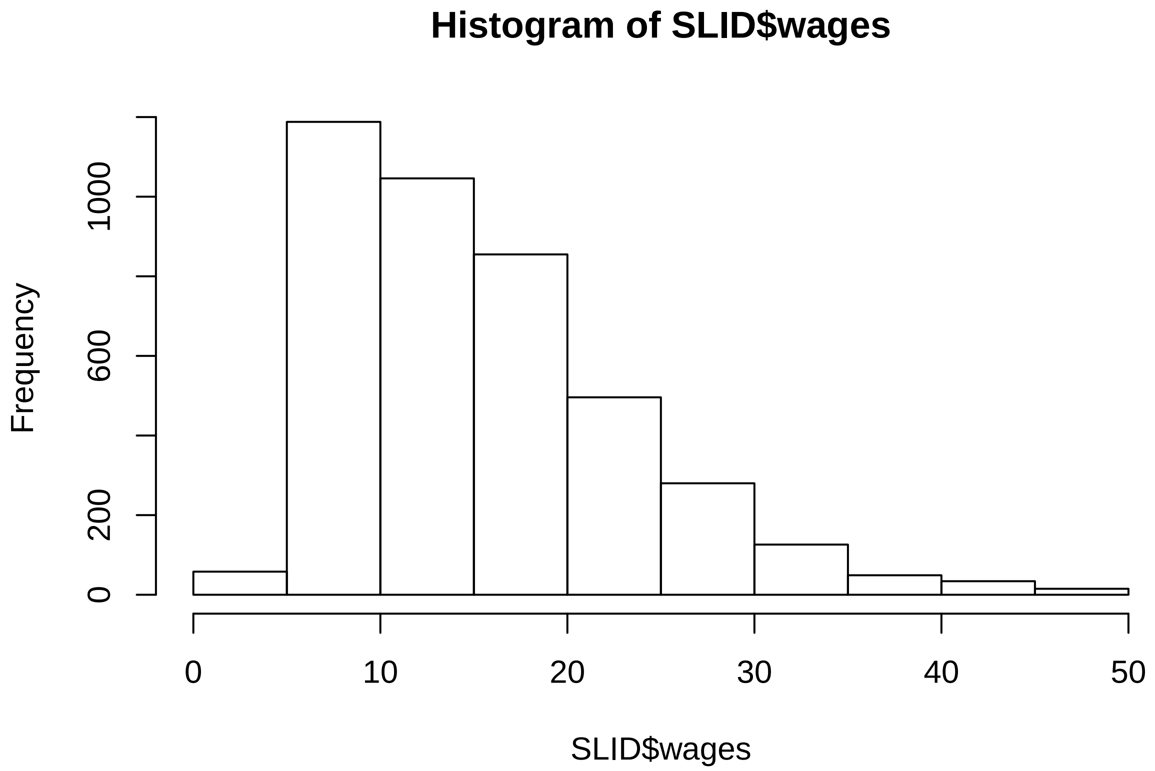
```r
# Normality check
qqnorm(residuals(l1))
qqline(residuals(l1))
```



```r
# Using car
qqPlot(l1)
```



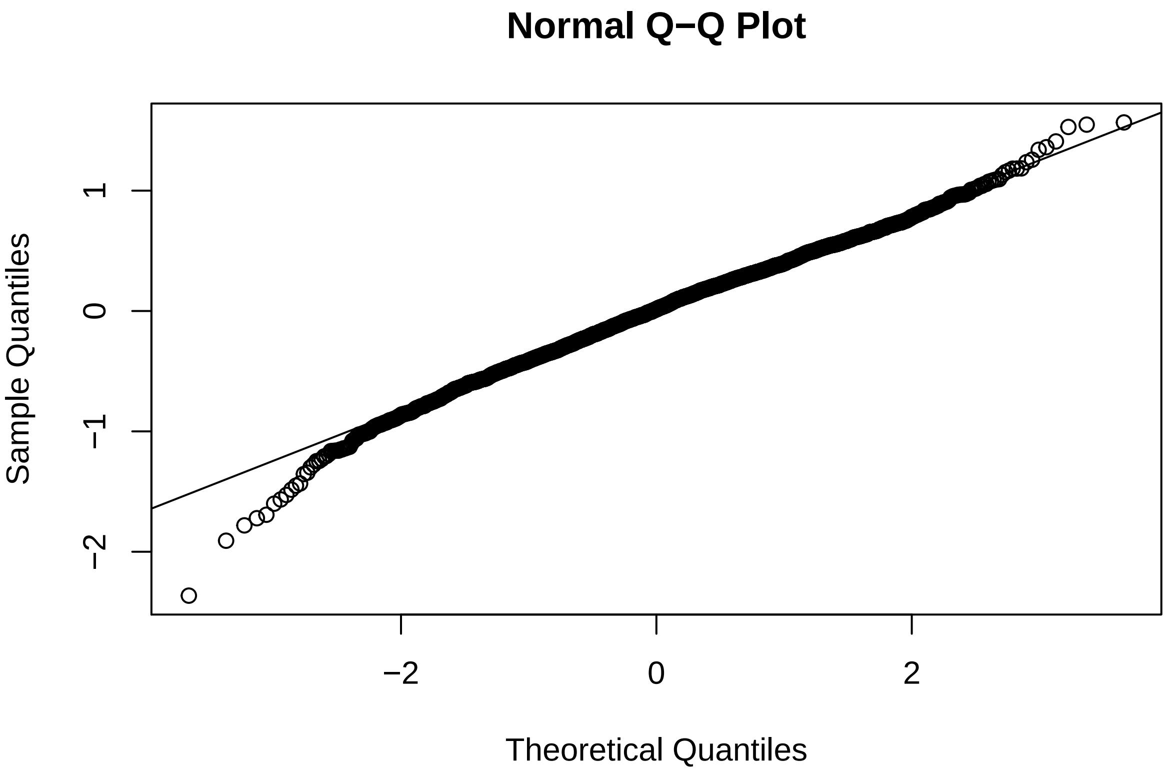
```
## Doesn't look good!
hist(SLID$wages)
```



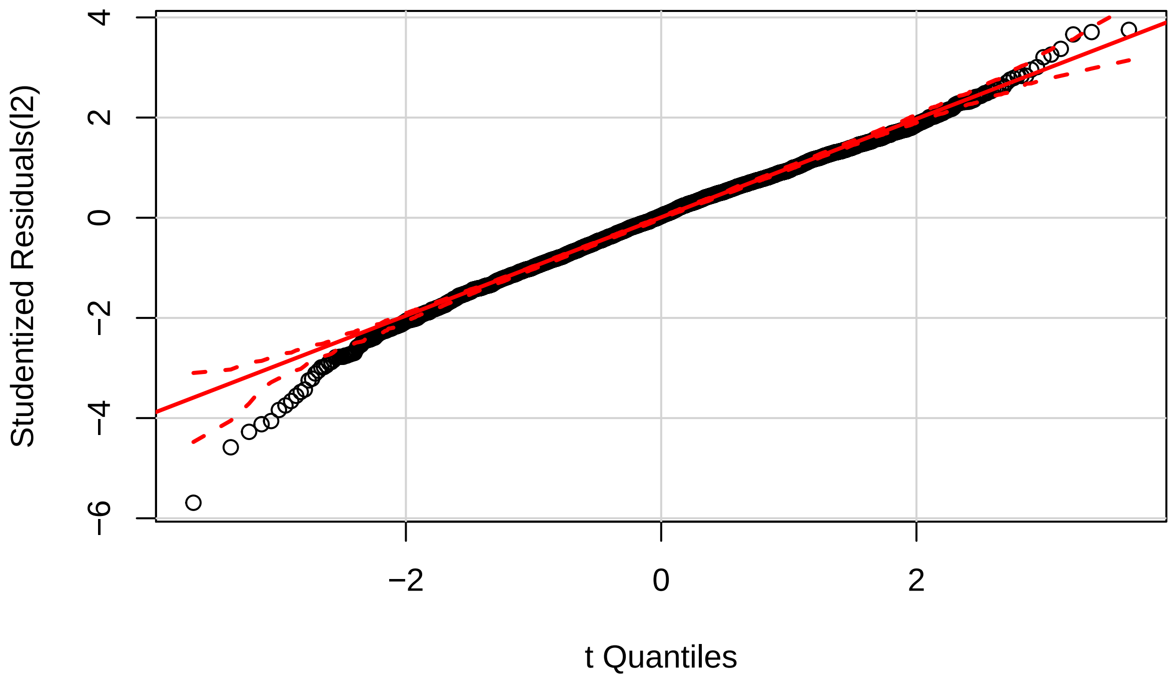
```r
## Try log wages

l2 <- update(l1, log(wages) ~ .)
summary(l2)
```

```
##
## Call:
## lm(formula = log(wages) ~ education + age + sex + language, data = SLID)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36456 -0.27684  0.01459  0.28443  1.56665
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.1184129  0.0388270  28.805   <2e-16 ***
## education      0.0550354  0.0022045  24.966   <2e-16 ***
## age            0.0176215  0.0005526  31.888   <2e-16 ***
## sexMale        0.2242585  0.0132662  16.905   <2e-16 ***
## languageFrench 0.0049223  0.0270615   0.182    0.856
## languageOther  0.0099270  0.0206137   0.482    0.630
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4186 on 3981 degrees of freedom
##   (3438 observations deleted due to missingness)
## Multiple R-squared:  0.3095, Adjusted R-squared:  0.3086
## F-statistic: 356.9 on 5 and 3981 DF,  p-value: < 2.2e-16
```

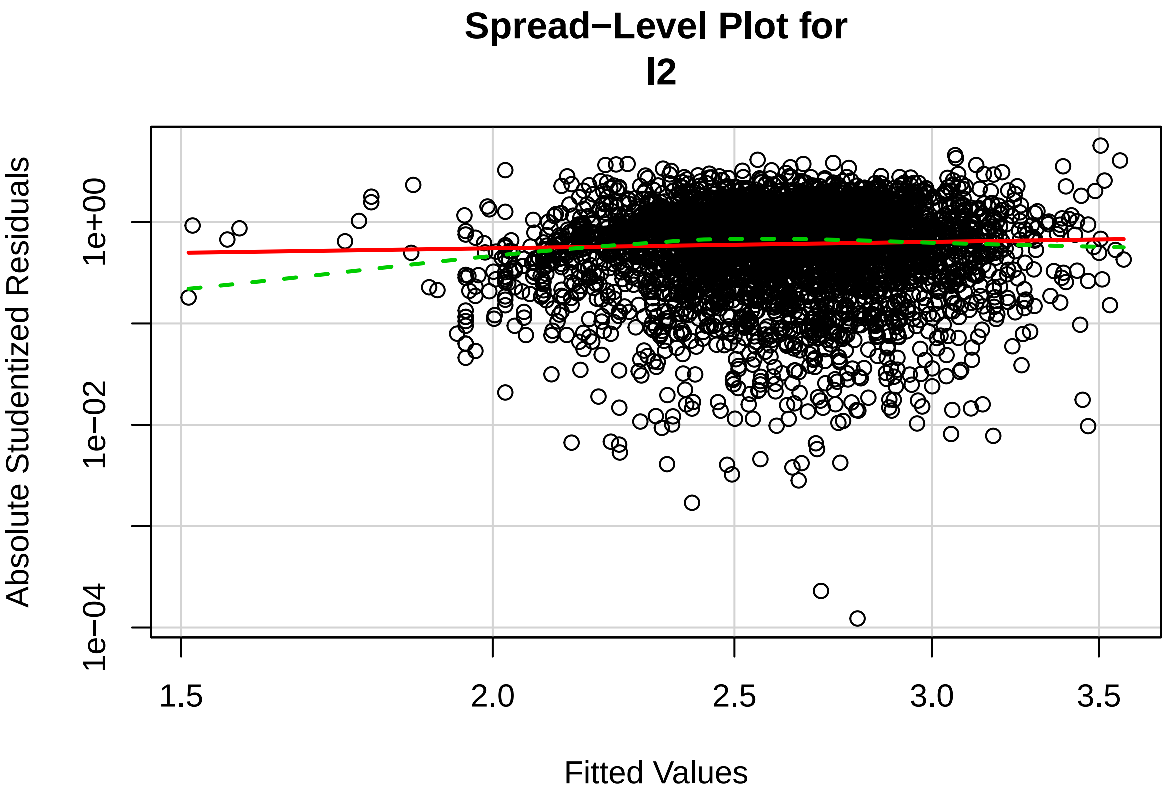
```
qqnorm(residuals(l2))
qqline(residuals(l2))
```



```
qqPlot(l2)
```



```
## Much better!

# Check for non-constant variance
spreadLevelPlot(l2)## Doesn't look too bad
```



```
##
## Suggested power transformation:  0.6420316
```

```
ncvTest(l2)  # But this is highly signficant
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 27.42581    Df = 1     p = 1.632382e-07
```

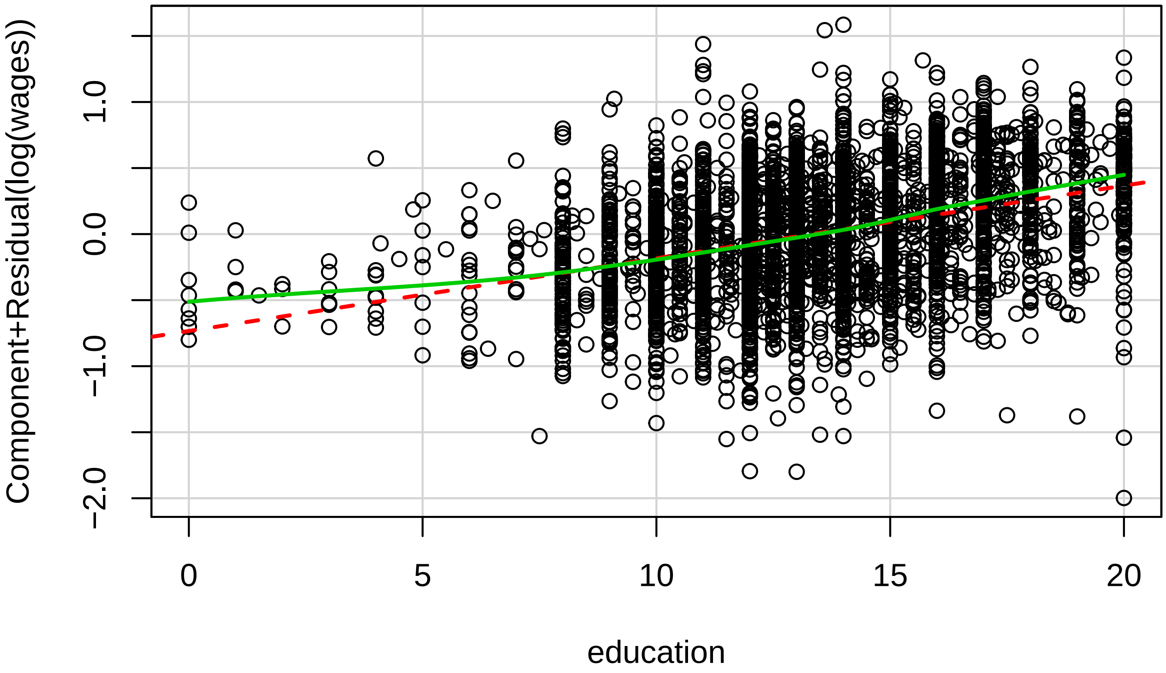
```
sqrt(diag(hccm(l2)))
```

```
##    (Intercept)      education            age        sexMale languageFrench
##   0.0394208573   0.0022620181   0.0005913719   0.0132655482   0.0296852045
##  languageOther
##   0.0201622208
```

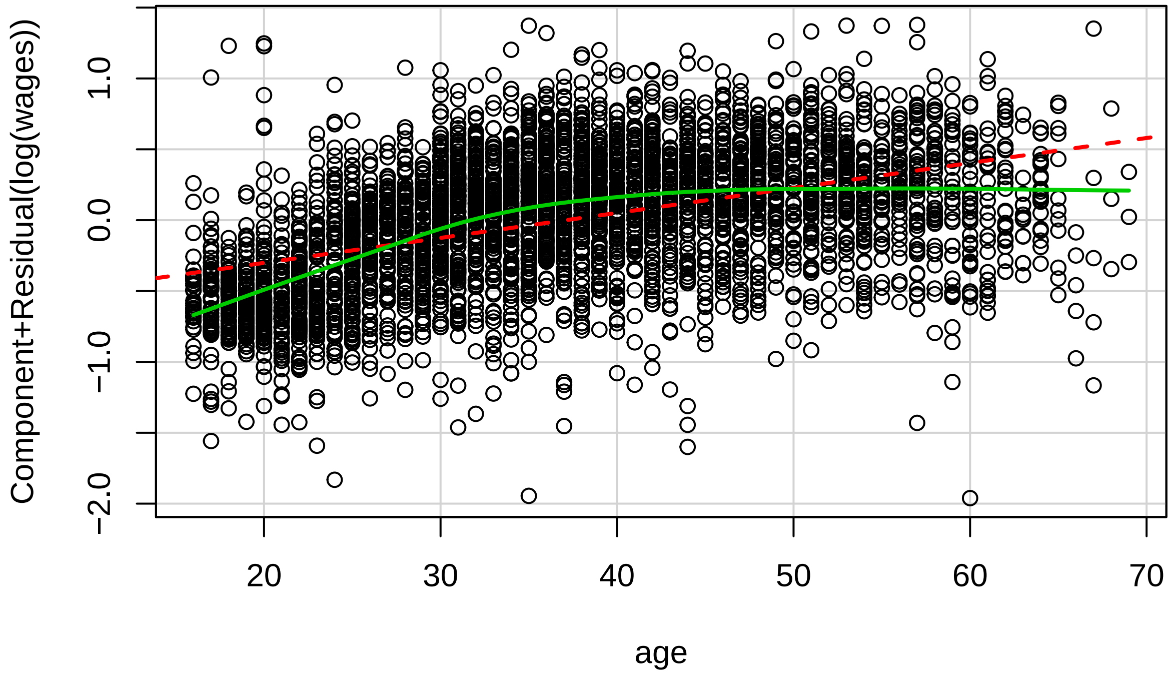
```
summary(l2)
```

```
##
## Call:
## lm(formula = log(wages) ~ education + age + sex + language, data = SLID)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36456 -0.27684  0.01459  0.28443  1.56665
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.1184129  0.0388270  28.805   <2e-16 ***
## education       0.0550354  0.0022045  24.966   <2e-16 ***
## age             0.0176215  0.0005526  31.888   <2e-16 ***
## sexMale         0.2242585  0.0132662  16.905   <2e-16 ***
## languageFrench  0.0049223  0.0270615   0.182    0.856
## languageOther   0.0099270  0.0206137   0.482    0.630
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4186 on 3981 degrees of freedom
##   (3438 observations deleted due to missingness)
## Multiple R-squared:  0.3095, Adjusted R-squared:  0.3086
## F-statistic: 356.9 on 5 and 3981 DF,  p-value: < 2.2e-16
```

```
## These are very similar to regular standard errors, so I think this can be ignored

# Linearity
crPlot(l2, 'education')
```



```
crPlot(l2, 'age')
```



```
l3 <- update(l2, . ~ . + poly(age, 2, raw = TRUE) - age)
summary(l3)
```

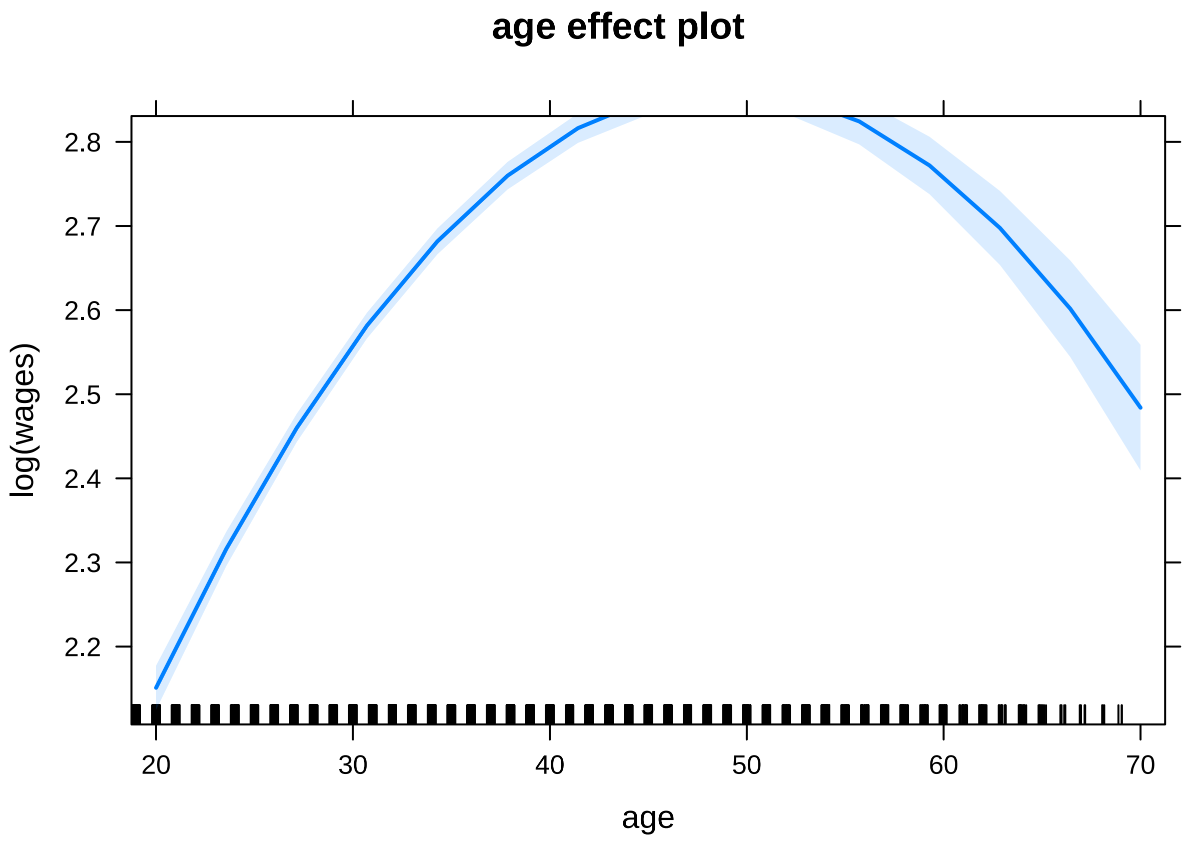
```
##
## Call:
## lm(formula = log(wages) ~ education + sex + language + poly(age,
##     2, raw = TRUE), data = SLID)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.02933 -0.23950  0.02141  0.25442  1.77906
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.0921787  0.0604474   1.525    0.127
## education                  0.0464076  0.0021266  21.823   <2e-16 ***
## sexMale                    0.2240107  0.0125652  17.828   <2e-16 ***
## languageFrench            -0.0100608  0.0256410  -0.392    0.695
## languageOther              0.0058475  0.0195253   0.299    0.765
## poly(age, 2, raw = TRUE)1  0.0834855  0.0031231  26.731   <2e-16 ***
## poly(age, 2, raw = TRUE)2 -0.0008536  0.0000399 -21.392   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3964 on 3980 degrees of freedom
##   (3438 observations deleted due to missingness)
## Multiple R-squared:  0.3807, Adjusted R-squared:  0.3798
## F-statistic: 407.8 on 6 and 3980 DF,  p-value: < 2.2e-16
```

```r
anova(l2, l3)
```

```
## Analysis of Variance Table
##
## Model 1: log(wages) ~ education + age + sex + language
## Model 2: log(wages) ~ education + sex + language + poly(age, 2, raw = TRUE)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   3981 697.45
## 2   3980 625.53  1     71.92 457.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(effects::effect('age', l3))
```

```
## NOTE: age does not appear in the model
```



```r
##  Interactions

l4 <- update(l3, . ~ . + education * sex)
summary(l4)
```

```
##
## Call:
## lm(formula = log(wages) ~ education + sex + language + poly(age,
##     2, raw = TRUE) + education:sex, data = SLID)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96436 -0.23874  0.02438  0.25620  1.75801
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -6.062e-02  6.741e-02  -0.899    0.369
## education                  5.765e-02  3.073e-03  18.764  < 2e-16 ***
## sexMale                    5.029e-01  5.657e-02   8.891  < 2e-16 ***
## languageFrench            -1.040e-02  2.556e-02  -0.407    0.684
## languageOther              6.198e-03  1.947e-02   0.318    0.750
## poly(age, 2, raw = TRUE)1  8.356e-02  3.114e-03  26.838  < 2e-16 ***
## poly(age, 2, raw = TRUE)2 -8.542e-04  3.978e-05 -21.472  < 2e-16 ***
## education:sexMale         -2.091e-02  4.134e-03  -5.056 4.46e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3952 on 3979 degrees of freedom
##   (3438 observations deleted due to missingness)
## Multiple R-squared:  0.3847, Adjusted R-squared:  0.3836
## F-statistic: 355.3 on 7 and 3979 DF,  p-value: < 2.2e-16
```

```
anova(l3, l4)
```

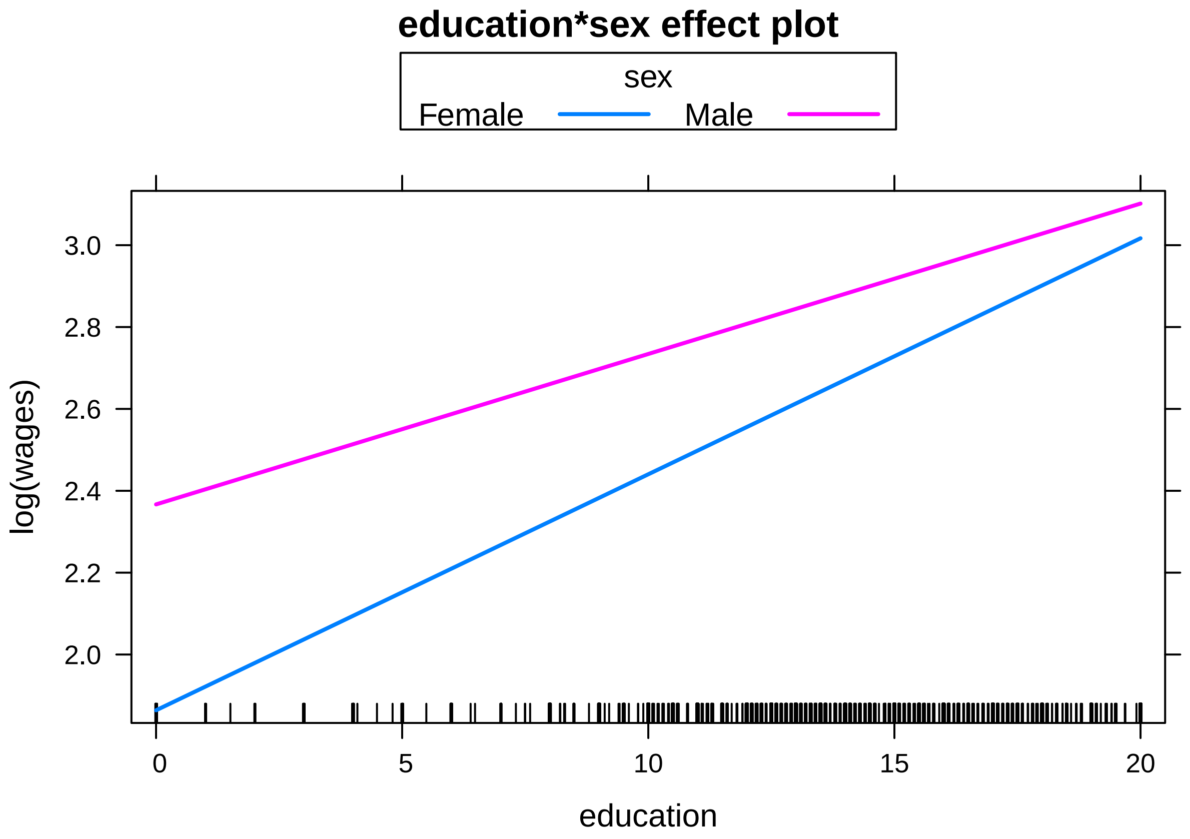
```
## Analysis of Variance Table
##
## Model 1: log(wages) ~ education + sex + language + poly(age, 2, raw = TRUE)
## Model 2: log(wages) ~ education + sex + language + poly(age, 2, raw = TRUE) +
##     education:sex
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   3980 625.53
## 2   3979 621.53  1    3.9938 25.568 4.463e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(effect(c('education', 'sex'), l4), multiline = TRUE)
```

```
## Warning in term == terms: longer object length is not a multiple of shorter
## object length
```

```
## Warning in term == names: longer object length is not a multiple of shorter
## object length
```

```
## NOTE: educationsex is not a high-order term in the model
```



```
vif(l4)
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## education                2.223547  1        1.491156
## sex                     20.417479  1        4.518570
## language                 1.023941  2        1.005932
## poly(age, 2, raw = TRUE)  1.062368  2        1.015240
## education:sex           21.336194  1        4.619112
```