# Regression Diagnostics Application in R

```
library(car)
library(ggplot2)
library(effects)
```

## Illustration

The data are from an RCT for adult inpatients recruited from an detox facility. Eligible subjects were adults, who spoke Spanish or English, reported alcohol, heroin, or cocaine as their first or second drug of choice, and either resided in proximity to the primary care clinic to which they would be referred, or were homeless. Patients with no primary care physician were randomized to receive a multidisciplinary assessment and a brief motivational intervention or usual care, with the goal of linking them to primary medical care. Subjects were interviewed at baseline during their detoxification stay, and follow-up interviews were undertaken every 6 months for 2 years so there are five measurement occasions. The data are in a file called help.csv in the Data folder on Canvas. More information about the data and variables are available in the codebook, which is also in the Data folder on Canvas.

Read in the data:

```
helpdata <- read.csv("help.csv")
```

The variable `substance` indicates their drug of choice, which has three categories. So let's regress the maximum number of drinks per day (past 30 days) on substance, age, and the interaction between substance and age. Recall R will automatically create the dummy variables for us if the variable is coded as a factor. We do not have to do it ourselves. You can check using the function `str()`.

```
lm1 <- lm(i1 ~ substance*age, data=helpdata)
summary(lm1)
```

```
##
## Call:
## lm(formula = i1 ~ substance * age, data = helpdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.653  -9.625  -4.832   5.576 102.891
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             7.9130     6.7925   1.165   0.2447
## substancecocaine        7.8539    10.1649   0.773   0.4401
## substanceheroin        -2.6009     9.6617  -0.269   0.7879
## age                     0.5571     0.1744   3.195   0.0015 **
## substancecocaine:age   -0.6625     0.2770  -2.391   0.0172 *
## substanceheroin:age    -0.4504     0.2653  -1.698   0.0902 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.7 on 447 degrees of freedom
```

```
## Multiple R-squared:  0.2268, Adjusted R-squared:  0.2181
## F-statistic: 26.22 on 5 and 447 DF,  p-value: < 2.2e-16
```

```
confint(lm1)
```

```
##                             2.5 %      97.5 %
## (Intercept)           -5.436213 21.26224963
## substancecocaine     -12.123107 27.83086562
## substanceheroin      -21.588828 16.38712406
## age                    0.214371  0.89978250
## substancecocaine:age  -1.206893 -0.11804406
## substanceheroin:age   -0.971717  0.07087019
```

We see that overall, the model accounts for 22.7% of the variance in number of drinks and that this is statistically significant, $F(5, 447) = 26.22, p < .001$.

Let's write out the equations for $\widehat{Y}$ for each of the substances:

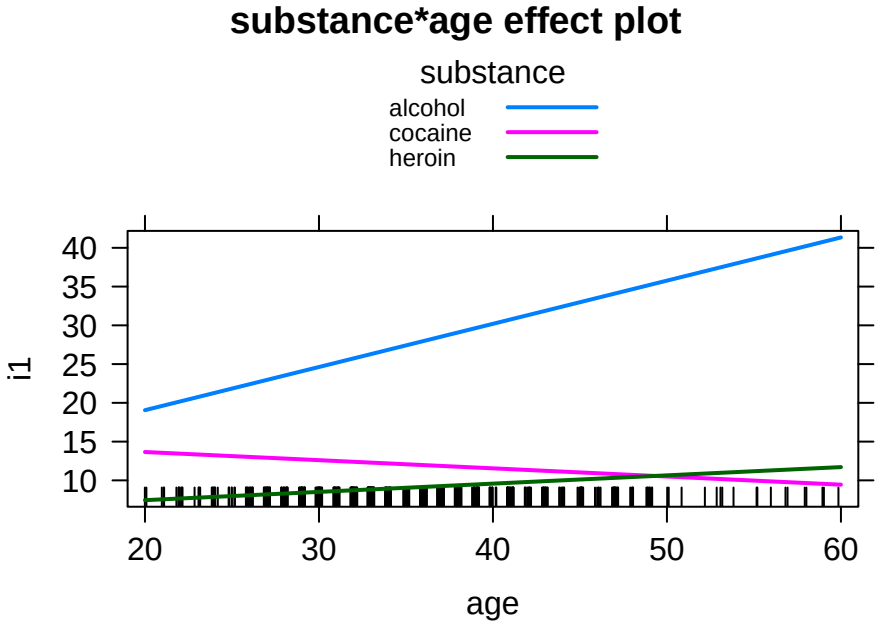
For alcohol, which is the reference group: $\widehat{Y} = 7.91 + .557age$

For cocaine: $\widehat{Y} = (7.91 + 7.85) + (.557 - .663)age$

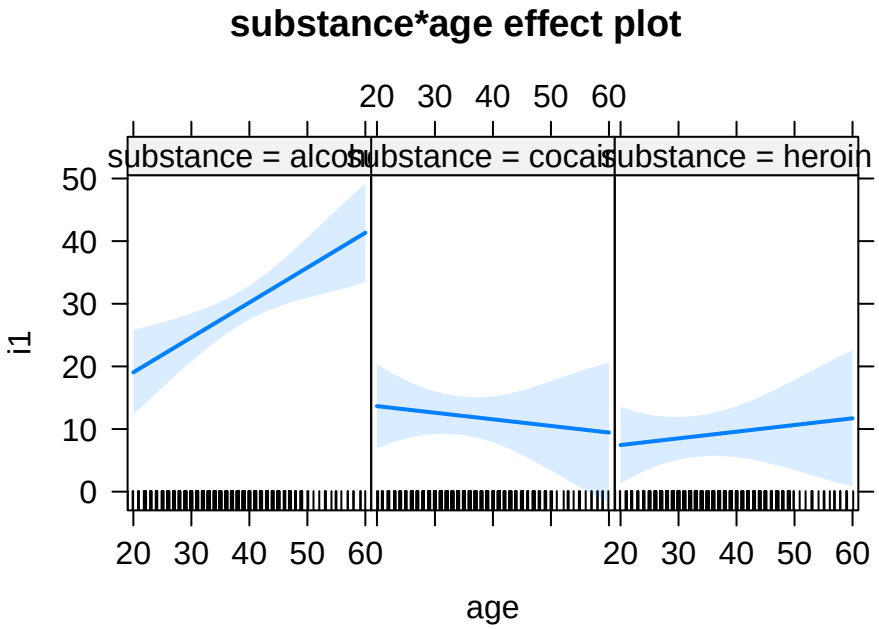
For heroin: $\widehat{Y} = (7.91 - 2.60) + (.557 - .450)age$

It is also helpful to obtain a plot of the effects:

```
plot(effect("substance*age", lm1, xlevels=list()), multiline=TRUE)
```




```
plot(allEffects(lm1))
```




For those whose primary substance is alcohol, they are predicted to on average consume 7.91 drinks at age=0, which doesn't make a lot of sense but that is what it means. More importantly, for every one year increase in age, they consume an additional .557 drinks (per day per year) and this effect is statistically significant, $t(447) = 3.195, p = .002$.

For those whose primary substance is cocaine, for every one year increase in age, they consume -0.106 fewer drinks (per day per year). This effect is also statistically significant, $t(447) = -2.391, p = .017$. And for those whose primary substance is heroin, for every one year increase in age, they consume 0.107 more drinks (per day per year) but this effect is not statistically significant, $t(447) = -1.698, p = .09$.

Next, we will examine regression diagnostics for assessing normality, linearity, homoscedasticity, and detecting influential observations.

### Outliers & Influential Observations

```
outlierTest(lm1)
```

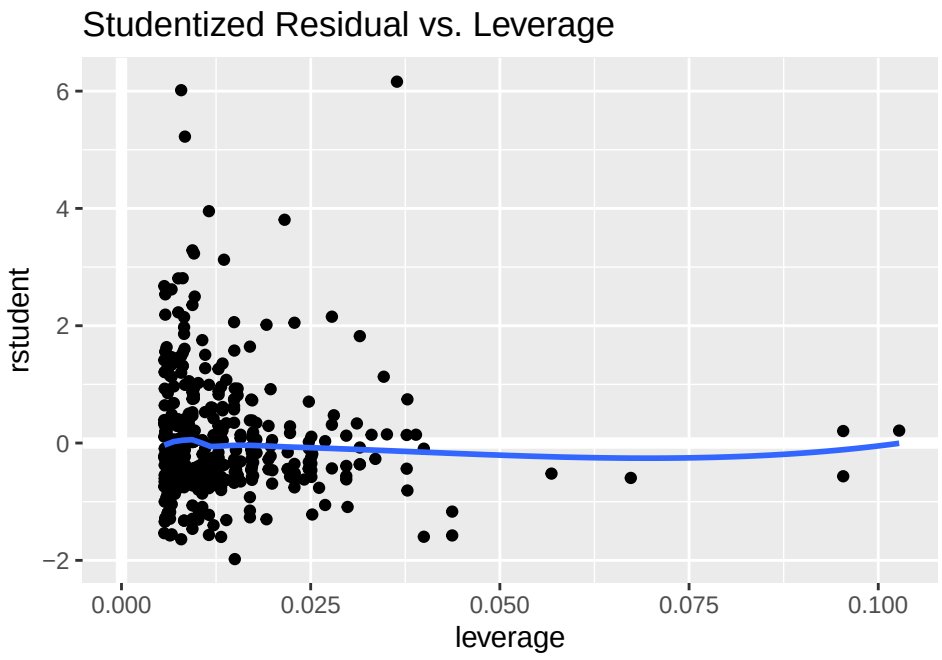
```
##     rstudent unadjusted p-value Bonferroni p
## 286 6.160754         1.6190e-09   7.3342e-07
## 244 6.015853         3.7328e-09   1.6910e-06
## 74  5.224211         2.6900e-07   1.2186e-04
## 177 3.952927         8.9748e-05   4.0656e-02
```

```
# add residuals and predicted values to data frame
helpdata$fitted <- fitted(lm1)
helpdata$resid <- resid(lm1)
helpdata$rstandard <- rstandard(lm1)
helpdata$rstudent <- rstudent(lm1)
helpdata$leverage <- hatvalues(lm1)
helpdata$cooksD <- cooks.distance(lm1)
```

```
ggplot(helpdata, aes(leverage, rstudent)) +
  geom_vline(size = 2, colour = "white", xintercept = 0) +
  geom_hline(size = 2, colour = "white", yintercept = 0) +
  geom_point() + geom_smooth(se = FALSE) +
  ggtitle("Studentized Residual vs. Leverage")
```

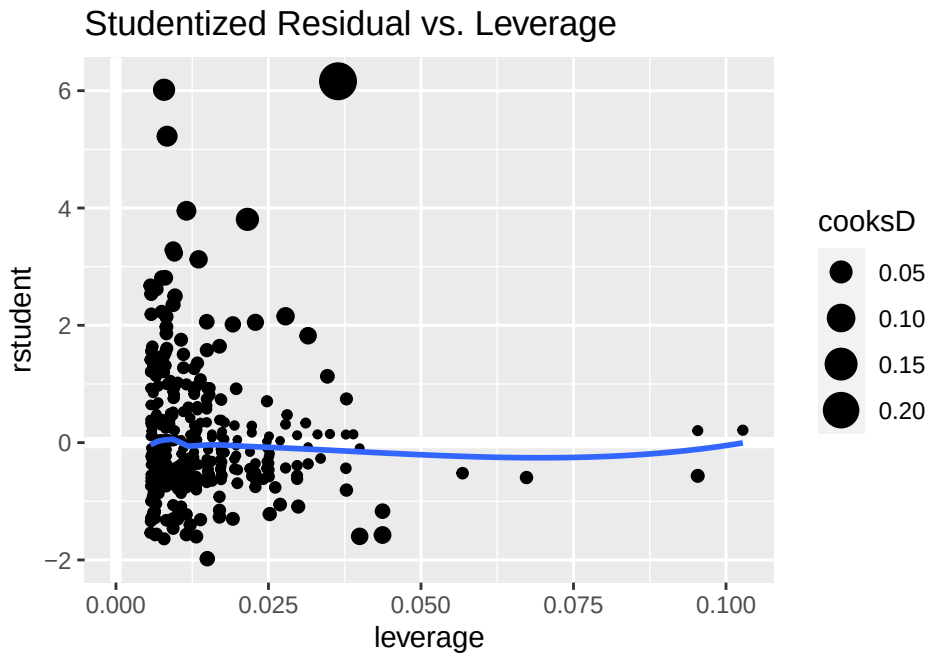
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

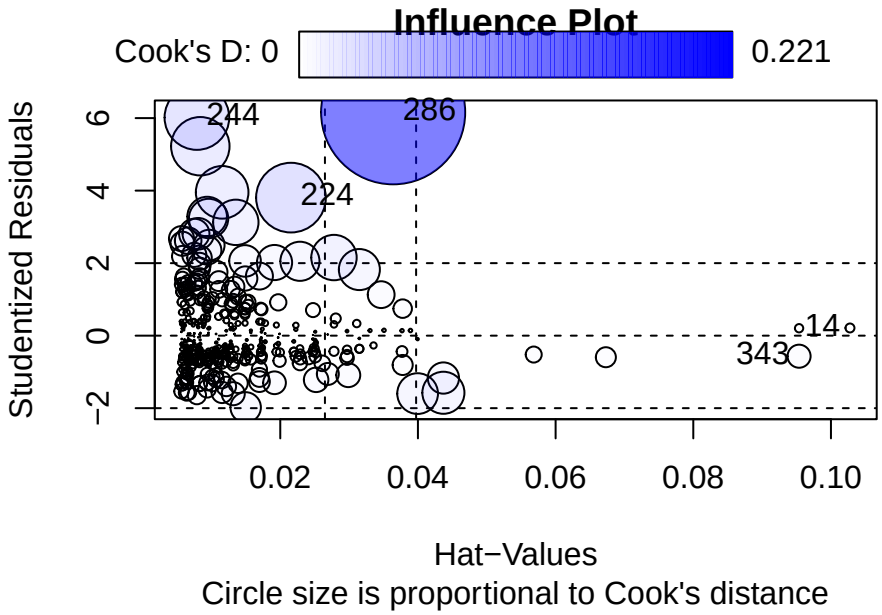



```
ggplot(helpdata, aes(leverage, rstudent)) +
  geom_vline(size = 2, colour = "white", xintercept = 0) +
  geom_hline(size = 2, colour = "white", yintercept = 0) +
  geom_point(aes(size = cooksD)) + geom_smooth(se = FALSE) +
  ggtitle("Studentized Residual vs. Leverage")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

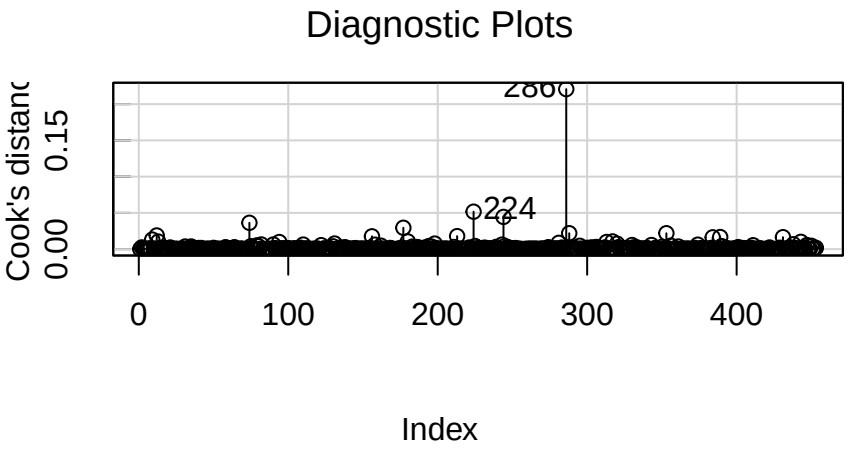



```
influencePlot(lm1, main="Influence Plot",
              sub="Circle size is proportional to Cook's distance")
```



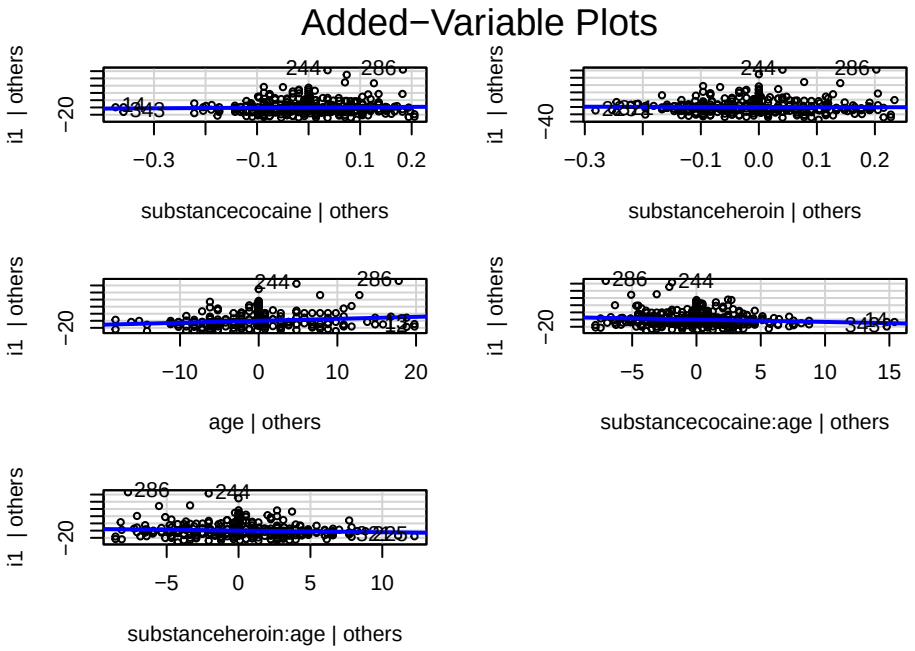

```
##        StudRes         Hat        CookD
## 14   0.2118724 0.102762596 0.0008587255
## 224  3.8067363 0.021552745 0.0516423104
## 244  6.0158535 0.007887397 0.0444533867
## 286  6.1607545 0.036400545 0.2207143273
## 343 -0.5666826 0.095368568 0.0056509600
```

```
influenceIndexPlot(lm1, vars="Cook")
```



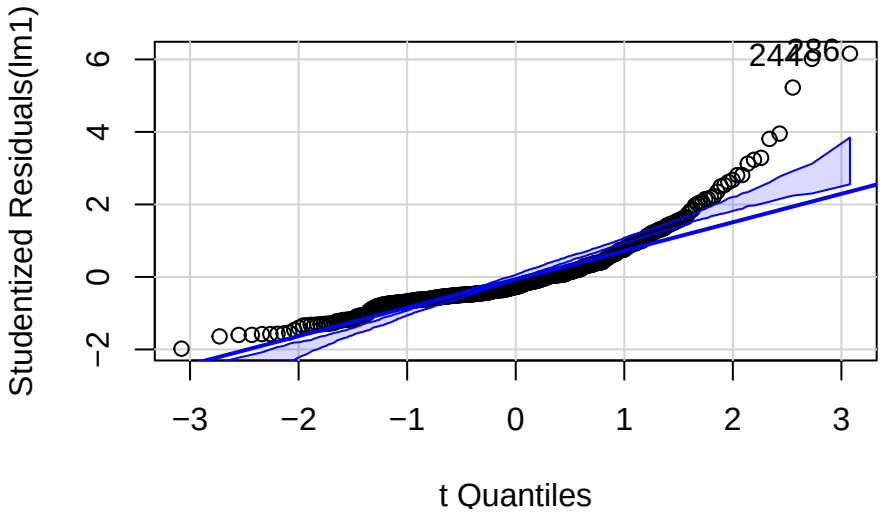

### Jointly influential observations

```
avPlots(lm1)
```




### Normally distributed residuals

```
qqPlot(lm1)
```



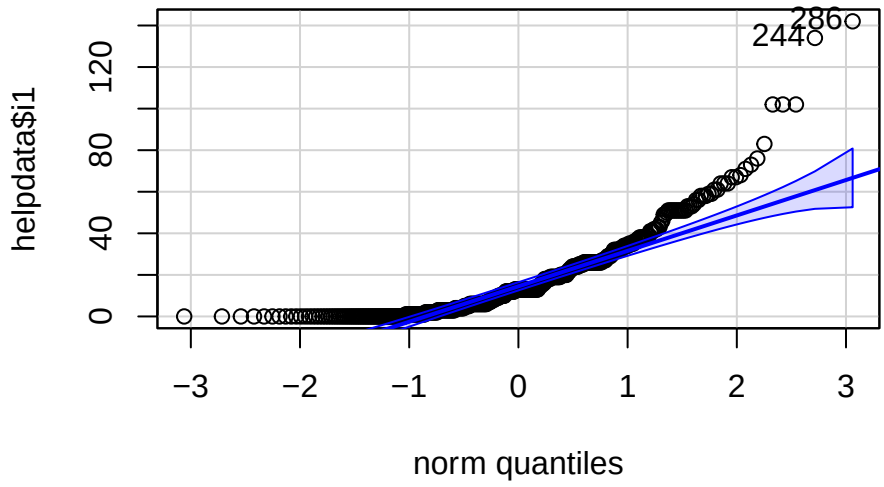

```
## [1] 244 286
```

The distribution of the residuals should be $N(0, \sigma_\epsilon)$.

The function `qqPlot()` uses the studentized residuals (i.e., those obtained from a jacknife) against a $t$ distribution with $n \check{} k \check{} 1$ degrees of freedom, where $n$ is the sample size and $k$ is the number of regression parameters (including the intercept).

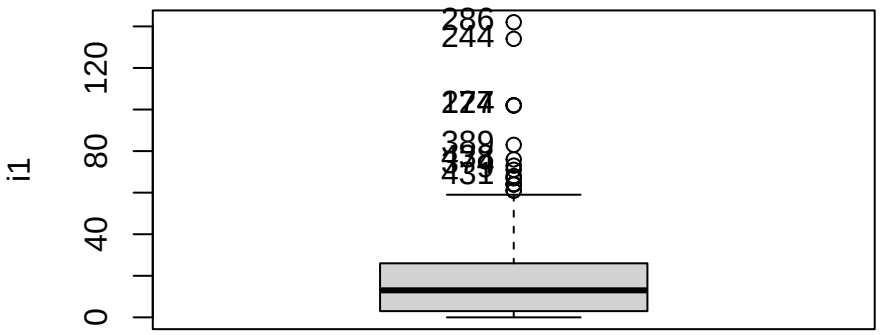
It looks like we have lighter tails to the left of the distribution and heavy tails to the right. Let's look at the distribution of the number of drinks outcome.

```
qqPlot(helpdata$i1)
```




```
## [1] 286 244
```

```
Boxplot(~ i1, data=helpdata)
```




```
##  [1] "286" "244" "74"  "177" "224" "389" "438" "374" "9"   "431"
```

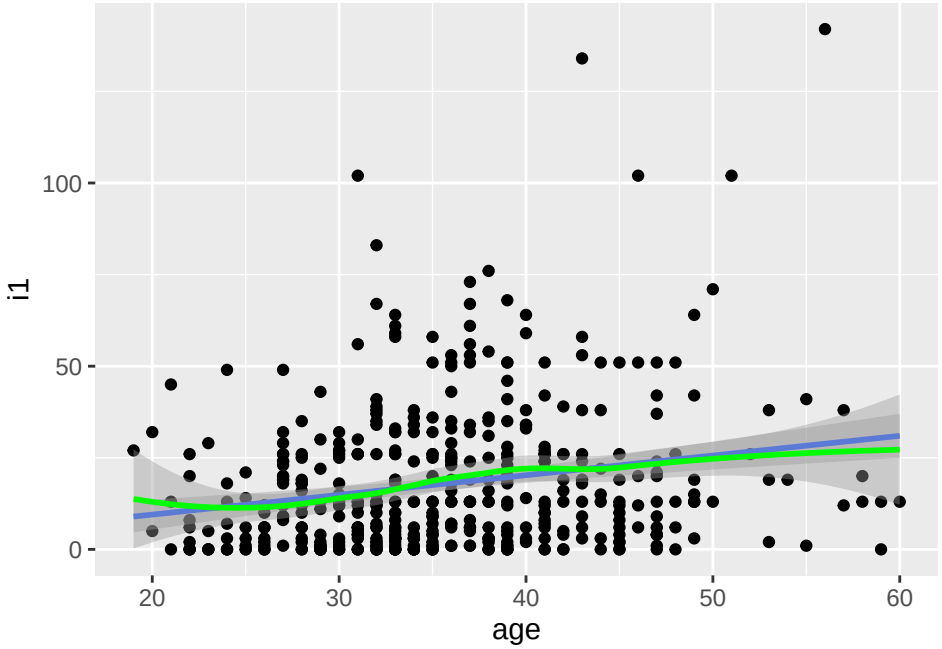
We could also have looked at a histogram or density plot.

So we should have looked at this first! But let's keep going.

We can also look at a scatterplot of age and number of drinks.

```
ggplot(data=helpdata, aes(x=age, y=i1)) +
  geom_point() +
  stat_smooth(method = lm) + stat_smooth(method = loess, color = "green")
```

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

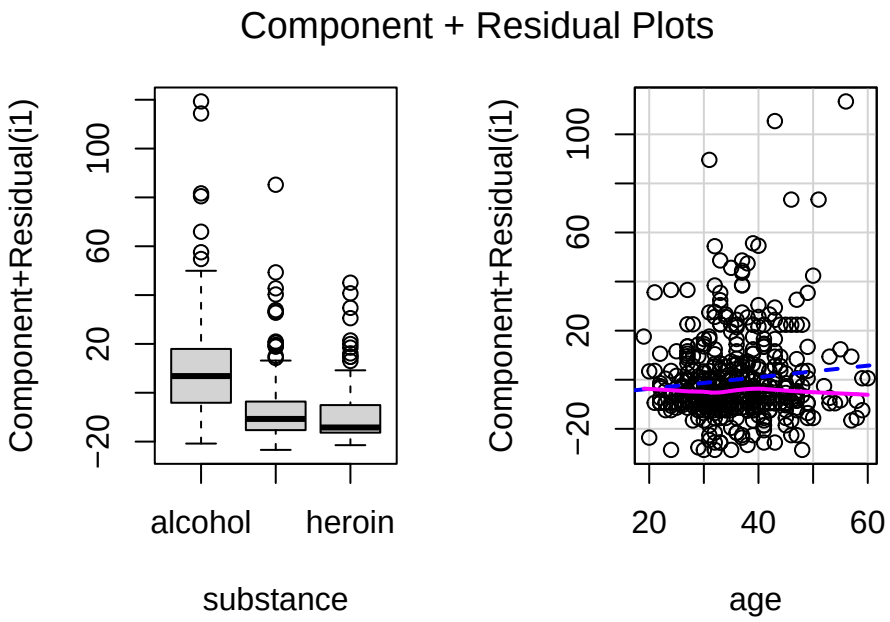



There does appear to be a weak relationship (although this is across all substance preferences).

### Linearity

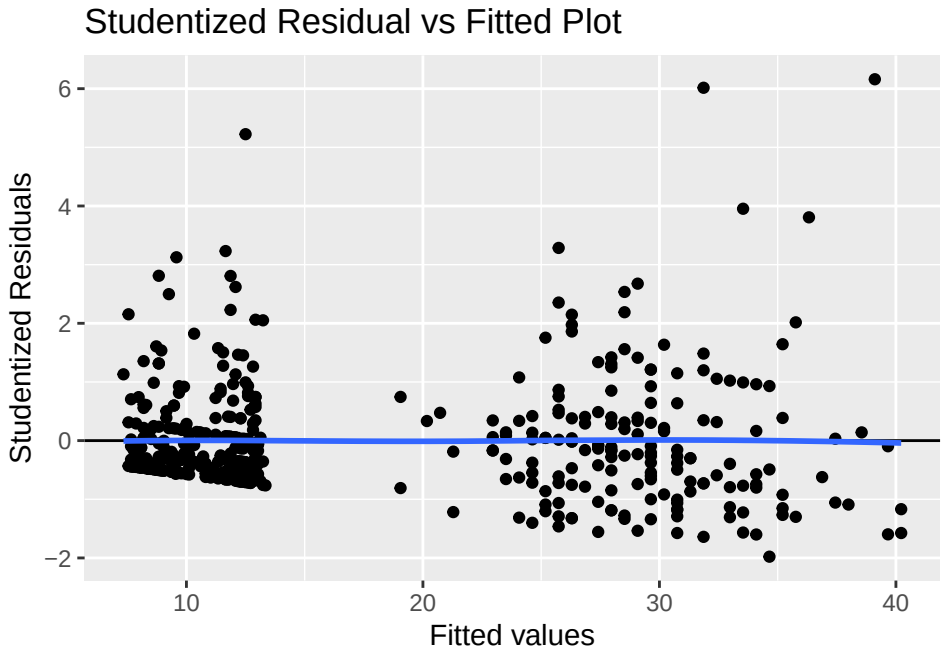
Component residual plots are not available for models with interactions, so let's refit the model without the interaction term. The `crPlots()` function in the `car` package will generate Component+Residual plots.

```
lm0 <- lm(i1 ~ substance + age, data=helpdata)
crPlots(lm0)
```




```
ggplot(helpdata , aes(fitted, rstudent)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  geom_smooth(se=FALSE) +
  xlab("Fitted values") + ylab("Studentized Residuals") +
  ggtitle("Studentized Residual vs Fitted Plot")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




If the outcome variable is linearly related to the predictor variables, there should be no systematic relationship, and actually there isn't but we can begin to see a potential problem concerning the next assumption.

### Non-constant error variance

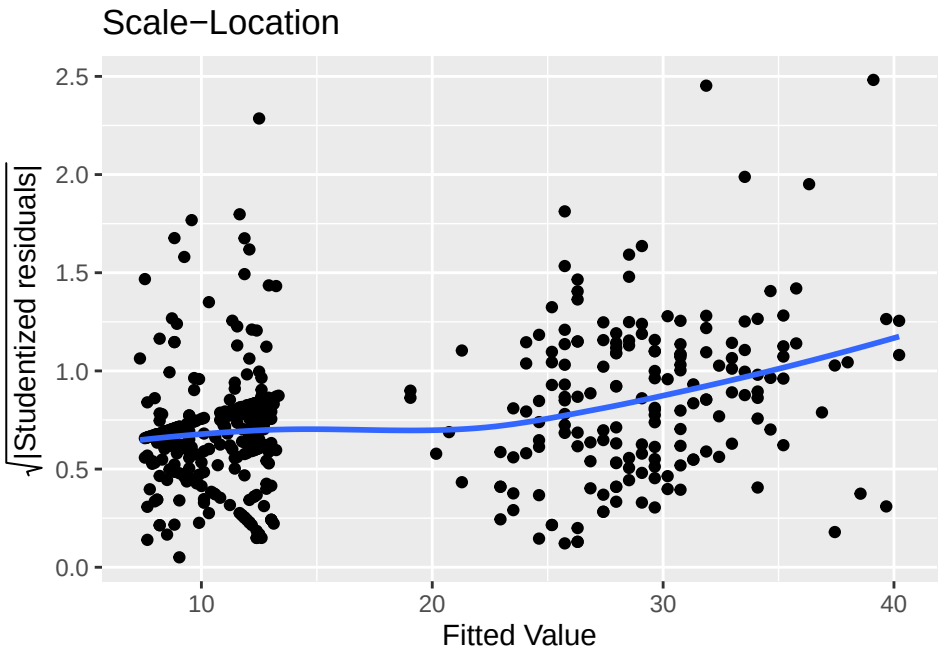
```
ncvTest(lm1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 79.19882, Df = 1, p = < 2.22e-16
```

We clearly reject the null hypothesis of constant error variance, $\chi^2(1) = 79.2, p < .001$.

```
ggplot(helpdata, aes(fitted, sqrt(abs(rstudent)))) +
  geom_point(na.rm=TRUE) +
  stat_smooth(method="loess", se=FALSE, na.rm = TRUE) +
  xlab("Fitted Value") +
  ylab(expression(sqrt("|Studentized residuals|"))) +
  ggtitle("Scale-Location")
```

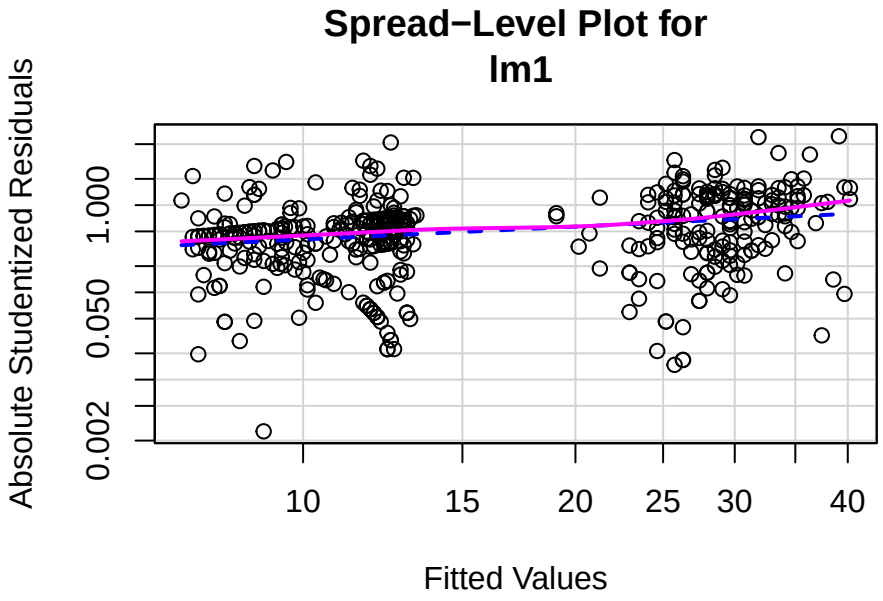
```
## `geom_smooth()` using formula = 'y ~ x'
```




There should be no discernible trends in this plot but there is.

The `spreadLevelPlot()` function in the `car` package will create the spread level plot, which is like the Scale-Location plot but does not use the square root of the absolute studentized residuals, and it also suggests a variance stabilizing transformation.

```
spreadLevelPlot(lm1)
```




```
##
## Suggested power transformation:  0.5130973
```

We have all kinds of problems here. So let's try a few things.

```
helpdata[c(286,244,224), ]
```

```
##      id e2b1 g1b1 i11     pcs1     mcs1 cesd1 indtot1 drugrisk1 sexrisk1 pcrec1
## 286 338   NA    1  NA 38.19467 29.68435    42     108        11         7      2
## 244 289    3    0 216 50.23810 40.80059    12      75         0         5      2
## 224 262    2    0  64 22.43385 32.74409    34      61         0         1      0
##     e2b2 g1b2 i12     pcs2     mcs2 cesd2 indtot2 drugrisk2 sexrisk2 pcrec2
## 286   NA    1  NA 38.80085 32.74532    30       NA         0         0      2
## 244   NA   NA  NA       NA       NA    NA       NA        NA        NA     NA
## 224   NA   NA  NA       NA       NA    NA       NA        NA        NA     NA
##     e2b3 g1b3 i13     pcs3     mcs3 cesd3 indtot3 drugrisk3 sexrisk3 pcrec3
## 286   NA    0  NA 39.63049 31.05753    31       21         0         0      2
## 244    6    0 256 46.22470 55.54596     9       82         0         4      2
## 224   NA   NA  NA       NA       NA    NA       NA        NA        NA     NA
##     e2b4 g1b4 i14     pcs4     mcs4 cesd4 indtot4 drugrisk4 sexrisk4 pcrec4
## 286   NA    0  NA 46.78941 34.62001    36        5         0         0      2
## 244    4    0 102 53.36348 50.41734     8       69         0         7      2
## 224   NA   NA  NA       NA       NA    NA       NA        NA        NA     NA
##     a15a a15b d1 e2b f1a f1b f1c f1d f1e f1f f1g f1h f1i f1j f1k f1l f1m f1n
## 286    5   49 36  11   2   3   1   2   3   1   0   0   1   2   3   1   3   3
## 244   90   14  4   4   0   1   0   3   1   0   1   3   2   2   0   3   2   0
## 224   35   20  6   4   2   3   1   1   1   3   3   1   3   1   3   1   3   3
##     f1o f1p f1q f1r f1s f1t g1b  i1  i2 age treat homeless      pcs      mcs
## 286   0   1   2   2   1   2   0 142   0  56     0        1 25.92422 34.41272
## 244   0   3   1   1   1   0   0 134 140  43     1        1 32.58783 55.99100
## 224   0   2   1   3   1   1   1 102 102  51     1        1 25.61815 27.80811
##     cesd indtot pss_fr drugrisk sexrisk satreat drinkstatus daysdrink
## 286   37     37      5        3       8       0           1         0
## 244   12     42     11        0      13       1           1        11
## 224   39     44      7        0       1       0           1         4
##     anysubstatus daysanysub linkstatus dayslink female substance racegrp
```

```
## 286            1          0          0      412      0   alcohol   white
## 244            1         11          1      236      0   alcohol   black
## 224            1          4          1      272      0   alcohol   white
##       fitted     resid rstandard rstudent    leverage     cooksD
## 286 39.10932 102.89068  5.920866 6.160754 0.036400545 0.22071433
## 244 31.86732 102.13268  5.792175 6.015853 0.007887397 0.04445339
## 224 36.32393  65.67607  3.750558 3.806736 0.021552745 0.05164231
```

Let's remove these observations. I will call the new data set `helpdata1`

```
helpdata1 <- helpdata[-c(286,244,224), ]
```

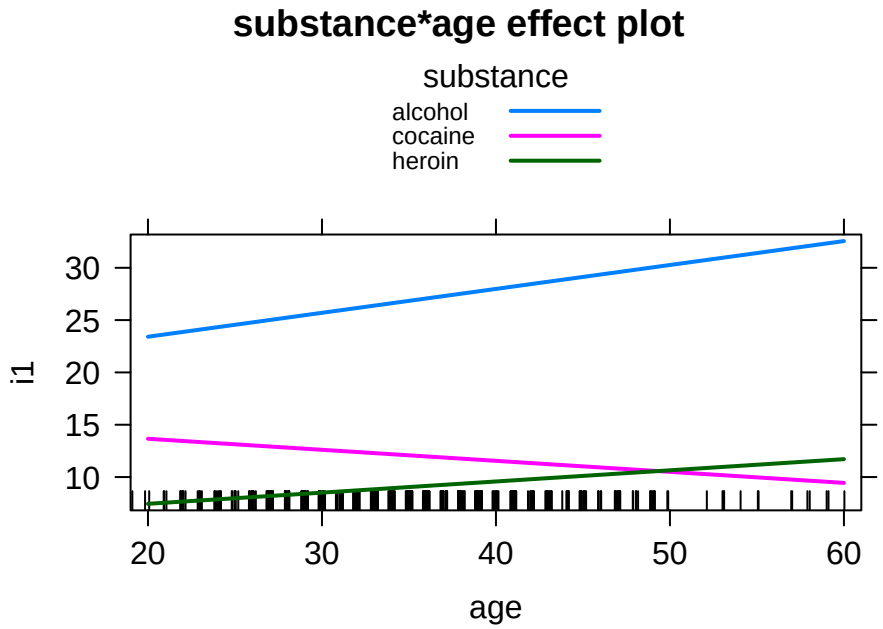
```
lm2 <- lm(i1 ~ substance*age, data=helpdata1)
summary(lm2)
```

```
##
## Call:
## lm(formula = i1 ~ substance * age, data = helpdata1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.810  -9.420  -4.590   6.064  89.500
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           18.8399     6.2503   3.014  0.00272 **
## substancecocaine      -3.0730     9.2532  -0.332  0.73997
## substanceheroin      -13.5277     8.8034  -1.537  0.12509
## age                    0.2285     0.1614   1.416  0.15746
## substancecocaine:age  -0.3339     0.2525  -1.322  0.18672
## substanceheroin:age   -0.1219     0.2420  -0.504  0.61477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.97 on 444 degrees of freedom
## Multiple R-squared:  0.2187, Adjusted R-squared:  0.2099
## F-statistic: 24.86 on 5 and 444 DF,  p-value: < 2.2e-16
```

```
confint(lm2)
```

```
##                           2.5 %     97.5 %
## (Intercept)            6.55604922 31.1237086
## substancecocaine     -21.25848624 15.1125233
## substanceheroin      -30.82913519  3.7737103
## age                   -0.08864486  0.5457198
## substancecocaine:age  -0.83020623  0.1623480
## substanceheroin:age   -0.59751252  0.3537443
```

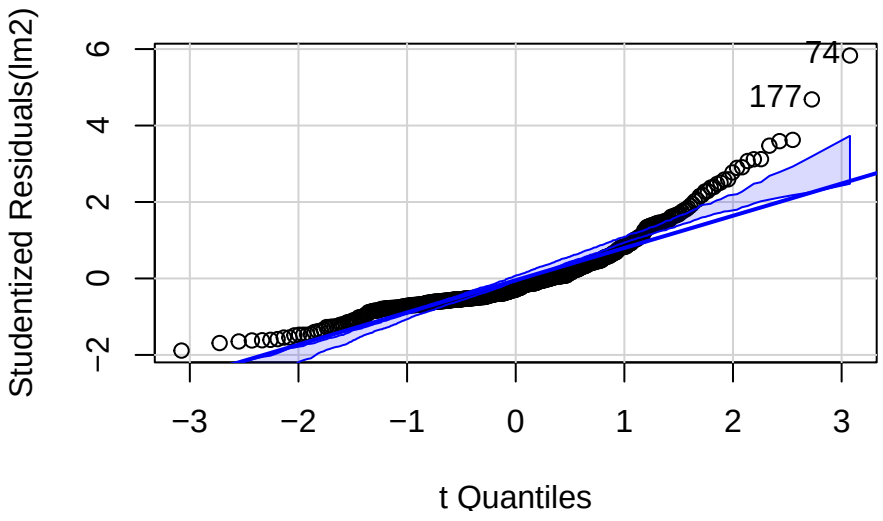
```
plot(effect("substance*age", lm2, xlevels=list()), multiline=TRUE)
```




Overall, the model still accounts for about the same proportion of the variance $R^2 = .22$ and is statistically significant $F(5, 444) = 24.86, p < .001$, but the coefficients have changed - none are statistically significant, except for the intercept.

But there are still some problems with this model in terms of normally distributed residuals.

```
qqPlot(lm2)
```



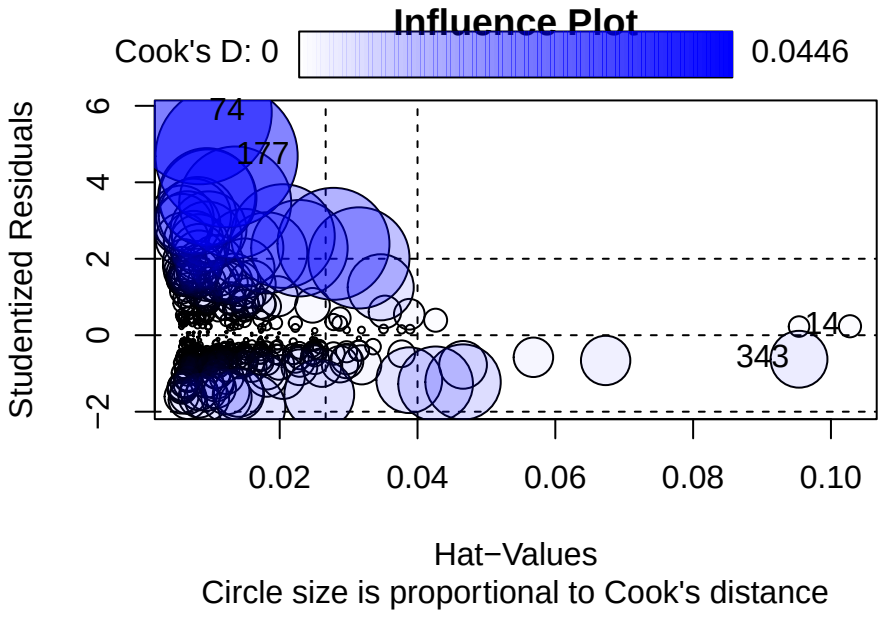

```
## [1]  74 177
```

There are perhaps more outliers and influential observations that we should have removed.

```
outlierTest(lm2)
```

```
##     rstudent unadjusted p-value Bonferroni p
## 74  5.832442         1.0547e-08   4.7460e-06
## 177 4.683072         3.7638e-06   1.6937e-03
```

```
influencePlot(lm2, main="Influence Plot",
              sub="Circle size is proportional to Cook's distance")
```




```
##        StudRes         Hat       CookD
## 14   0.2348234 0.102762596 0.001054835
## 74   5.8324416 0.008383206 0.044613323
## 177  4.6830715 0.012290467 0.043435398
## 343 -0.6281143 0.095368568 0.006941485
```

There's also still a problem with non-constant error variance.

```
ncvTest(lm2)
```

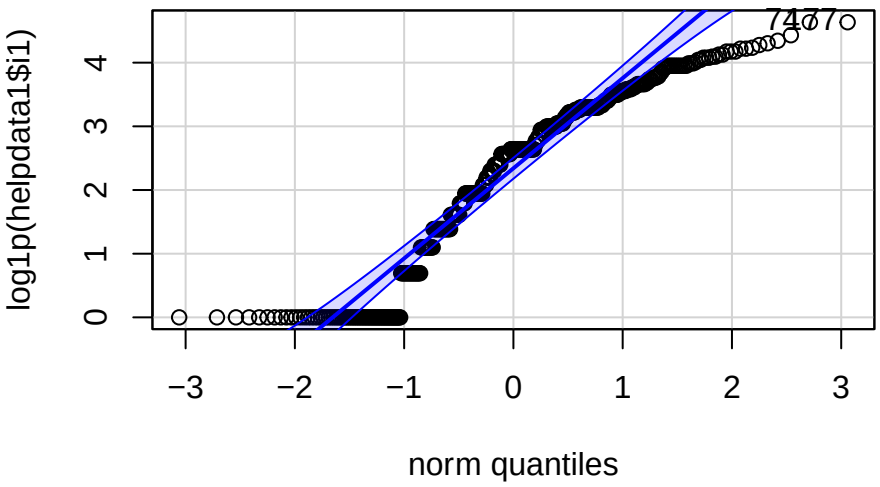
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 19.39394, Df = 1, p = 1.0634e-05
```

Let's look at a log transformation for `i1`. The function `log1p()` computes $log(1 + x)$. Thus, $s$, the "start" is 1 is used because we have values of 0 for `i1`.

```
summary(helpdata1$i1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    3.00   13.00   17.19   26.00  102.00
```

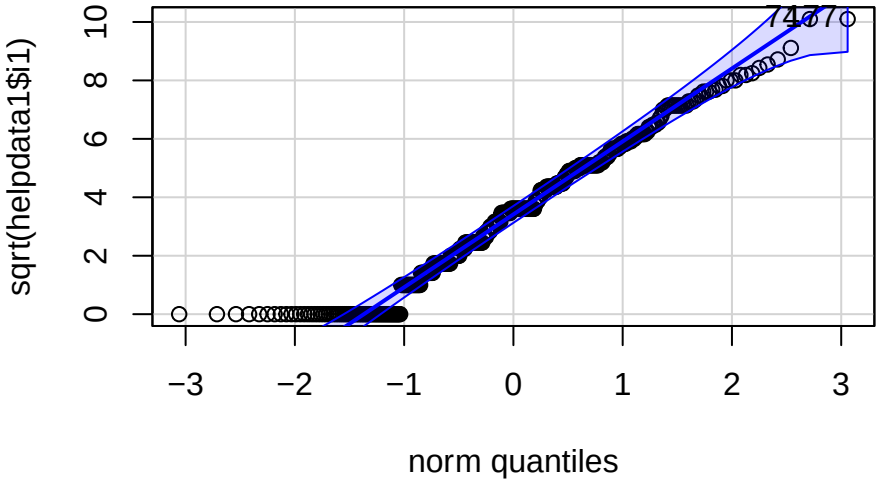
```
qqPlot(log1p(helpdata1$i1))
```




```
## [1]  74 177
```

That doesn't help.

```
qqPlot(sqrt(helpdata1$i1))
```




```
## [1]  74 177
```

Nor does that. The problem here is that there are a lot of people with a value of zero. In addition, number of drinks cannot go below zero. This is a count variable and count variables are best fit using the Poisson distribution and Poisson regression.

### An Aside on Numeric Cutoffs

It is generally more effective to examine the distributions of these quantities directly to locate unusual values but researchers like to have specific numerical criteria for identifying noteworthy observations on the basis of measures of leverage and influence. Numerical cutoffs are guidelines or suggestions only.

Nevertheless, numerical cutoffs can be of some use, as long as they are not given too much weight, and especially when they are employed to enhance graphical displays.

Cutoffs for a diagnostic statistic may be derived from statistical theory, or they may result from examination of the sample distribution of the statistic.

Cutoffs may be absolute, or they may be adjusted for sample size.

For some diagnostic statistics, such as measures of influence, absolute cutoffs are unlikely to identify noteworthy observations in large samples. In part, this characteristic reflects the ability of large samples to absorb discrepant data without changing the results substantially, but it is still often of interest to identify relatively influential points, even if no observation has strong absolute influence.

## Summary

Unusual data are problematic in linear models fit by least squares because they can substantially influence the results of the analysis, and because they may indicate that the model fails to capture important features of the data.

Observations with unusual combinations of explanatory variables values have high leverage in a least-squares regression. The hat-values provide a measure of leverage. A rough cutoff for noteworthy hat-values is $h_i > 2\overline{h} = 2(k+1)/n$.

A regression outlier is an observation with an unusual response variable value given its combination of explanatory variable values.

Studentized residuals $\epsilon_i^*$ can be used to identify outliers through graphical examination or a Bonferroni test for the largest absolute $\epsilon_i^*$. If the model is correct (and there are no true outliers), then each studentized residual follows a $t$-distribution with $n-k-2$ degrees of freedom.

Observations that combine high leverage with a large studentized residual exert substantial influence on the regression coefficients. Cook's $D$ provides a summary index of influence on the coefficients. A rough cutoff is $D_i > 4/(n-k-1)$.

Subsets of observations can be jointly influential. Added-variable plots are useful for detecting joint influence on the regression coefficients. The added-variable plot for the regressor $X_j$ is formed using the residuals from the least-squares regressions of $X_j$ and $Y$ on all of the other $X$'s.

Outlying and influential data should not be ignored, but they also should not simply be deleted without investigation. 'Bad' data can often be corrected. 'Good' observations that are unusual may provide insight into the structure of the data, and may motivate respecification of the statistical model used to summarize the data.

Heavy-tailed errors threaten the efficiency of least-squares estimation; skewed and multimodal errors compromise the interpretation of the least-squares fit.

Non-normality can often be detected by examining the distribution of the studentized residuals, and frequently can be corrected by transforming the data.

It is common for the variance of the errors to increase with the level of the response variable. This pattern of non-constant error variance can often be detected in a plot of studentized residuals against fitted values.

A rough rule of thumb is that non-constant error variance seriously degrades the least-squares estimator only when the ratio of the largest to smallest variance is about 10 or more.

These four plots are important diagnostic tools in assessing whether the linear model is appropriate.

**Studentized Residuals vs. Fitted**

When a linear model is appropriate, we expect

1. the studentized residuals will have constant variance when plotted against fitted values; and
2. the studentized residuals and fitted values will be uncorrelated.

If there are clear trends in this plot, or the plot looks like a funnel, these are clear indicators that the given linear model is inappropriate.

**Normal QQ plot**

You can use a linear model for prediction even if the underlying normality assumptions do not hold. However, in order for the p-values to be valid, the studentized residuals from the regression must look approximately normally distributed.

**Scale-location plot**

This is another version of the studentized residuals vs fitted plot. There should be no discernible trends in this plot.

**Studentized Residuals vs Leverage**

Leverage is a measure of how much an observation influenced the model fit. It is a one-number summary of how different the model fit would be if the given observation was excluded, compared to the model fit where the observation is included. Points with a *high residual* (poorly described by the model) and *high leverage* (high influence on model fit) are potentially influential.

**Component-Residual Plots**

Simple forms of nonlinearity can often be detected in component+residual plots.

Component+residual plots adequately reflect nonlinearity when the explanatory variables are themselves not strongly nonlinearly related.