EDF 5401 Midterm, Part 2: Depression.

2013-10-15

library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.3 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.0  
✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
✔ purrr 1.0.2   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(DescTools)

# Part 1 Depression and the Internet

## Analysis

### Load the data

Load the data. Force category to be an ordered category.

depress <- read\_delim("depression-and-the-internet.txt")

Rows: 169 Columns: 8  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: "\t"  
chr (2): Gender, Age  
dbl (6): Internet use (mean hours per week), DepressionBefore, DepressionAft...  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

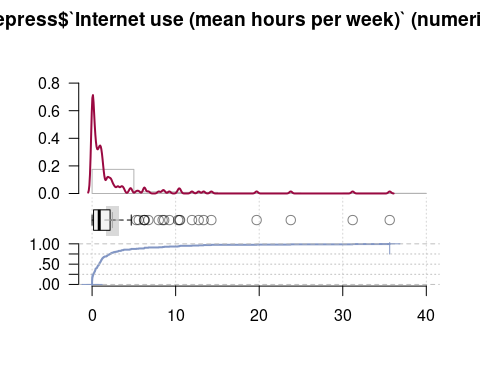
depress$Gender <- factor(depress$Gender)  
depress$Age <- factor(depress$Age)  
depress$Race <- factor(depress$`Race (white = 1, minority = 0)`,0:1,  
 c("minority","white"))  
summary(depress)

Internet use (mean hours per week) DepressionBefore DepressionAfter   
 Min. : 0.0000 Min. :0.0000 Min. :0.0000   
 1st Qu.: 0.1750 1st Qu.:0.3452 1st Qu.:0.3393   
 Median : 0.8554 Median :0.6666 Median :0.5000   
 Mean : 2.4287 Mean :0.7348 Mean :0.6176   
 3rd Qu.: 2.1532 3rd Qu.:1.0000 3rd Qu.:0.7856   
 Max. :35.6186 Max. :2.5332 Max. :3.0000   
 NA's :2 NA's :2 NA's :5   
 Gender Race (white = 1, minority = 0) Age Household income ($000)  
 female:94 Min. :0.0000 Adult:121 Min. : 5.00   
 male :75 1st Qu.:1.0000 Teen : 48 1st Qu.:36.56   
 Median :1.0000 Median :52.50   
 Mean :0.7515 Mean :54.41   
 3rd Qu.:1.0000 3rd Qu.:85.00   
 Max. :1.0000 Max. :85.00   
 NA's :5   
 Household size Race   
 Min. :1.000 minority: 42   
 1st Qu.:3.000 white :127   
 Median :4.000   
 Mean :3.784   
 3rd Qu.:5.000   
 Max. :6.000   
 NA's :2

### One Dimensional Summaries

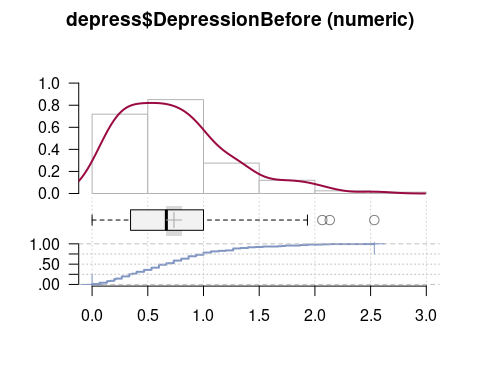
Desc(depress$`Internet use (mean hours per week)`)

------------------------------------------------------------------------------   
depress$`Internet use (mean hours per week)` (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 169 167 2 157 10 2.4286482 1.6743700  
 98.8% 1.2% 5.9% 3.1829264  
   
 .05 .10 .25 median .75 .90 .95  
 0.0000000 0.0112612 0.1750310 0.8554280 2.1531645 6.2751618 10.5071260  
   
 range sd vcoef mad IQR skew kurt  
 35.6186200 4.9370090 2.0328218 1.1668047 1.9781335 4.0803913 19.8697044  
   
lowest : 0.0 (10), 0.000265, 0.000285, 0.000352, 0.001923  
highest: 14.27382, 19.69108, 23.77727, 31.18628, 35.61862  
  
' 95%-CI (classic)



Desc(depress$DepressionBefore)

------------------------------------------------------------------------------   
depress$DepressionBefore (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 169 167 2 38 2 0.734790 0.660560  
 98.8% 1.2% 1.2% 0.809020  
   
 .05 .10 .25 median .75 .90 .95  
 0.133331 0.199982 0.345215 0.666626 1.000000 1.359912 1.723193  
   
 range sd vcoef mad IQR skew kurt  
 2.533203 0.485862 0.661225 0.494170 0.654785 0.945713 0.811305  
   
lowest : 0.0 (2), 0.066666 (5), 0.133331 (6), 0.142853, 0.166656  
highest: 1.866455 (2), 1.933105 (2), 2.066406, 2.133301, 2.533203  
  
' 95%-CI (classic)



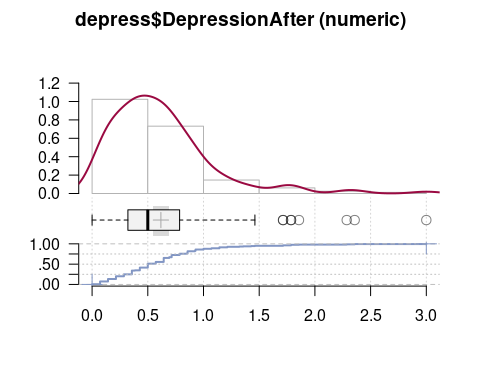
Identify outliers

hbefore <- which (depress$DepressionBefore > 2)  
hbefore

[1] 11 66 135

Desc(depress$DepressionAfter)

------------------------------------------------------------------------------   
depress$DepressionAfter (numeric)  
  
 length n NAs unique 0s mean meanCI'  
 169 164 5 32 1 0.6176045 0.5461457  
 97.0% 3.0% 0.6% 0.6890632  
   
 .05 .10 .25 median .75 .90 .95  
 0.0714260 0.1428530 0.3392642 0.5000000 0.7856450 1.1428220 1.4564821  
   
 range sd vcoef mad IQR skew kurt  
 3.0000000 0.4634394 0.7503821 0.3177123 0.4463808 1.8551112 5.2592718  
   
lowest : 0.0, 0.071426 (11), 0.142853 (10), 0.214264 (11), 0.285706 (8)  
highest: 1.785645 (2), 1.856934, 2.285645, 2.356934, 3.0  
  
heap(?): remarkable frequency (9.8%) for the mode(s) (= 0.5, 0.642822)  
  
' 95%-CI (classic)



Identify outliers

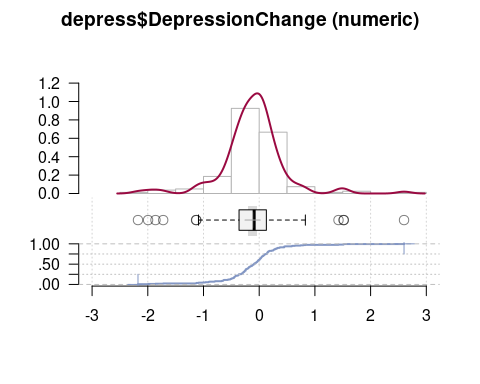
hafter <- which (depress$DepressionAfter > 1.5)  
hafter

[1] 13 18 21 37 47 86 136 169

Subtract before from after to get change.

depress$DepressionChange <- depress$DepressionAfter - depress$DepressionBefore  
Desc(depress$DepressionChange)

------------------------------------------------------------------------------   
depress$DepressionChange (numeric)  
  
 length n NAs unique 0s mean'  
 169 162 7 135 0 -0.1180701  
 95.9% 4.1% 0.0%   
   
 .05 .10 .25 median .75 .90  
 -0.9636996 -0.6419248 -0.3523030 -0.0904540 0.1237792 0.3328587  
   
 range sd vcoef mad IQR skew  
 4.7761230 0.5525114 -4.6795221 0.3424168 0.4760822 0.2044767  
   
 meanCI  
 -0.2037953  
 -0.0323448  
   
 .95  
 0.5901001  
   
 kurt  
 5.6532503  
   
lowest : -2.1760860, -1.9949800, -1.8616790, -1.7236020, -1.1333010  
highest: 0.828491, 1.423706, 1.514129, 1.518982, 2.600037  
  
' 95%-CI (classic)



Identify outliers (There is one point near the lower fence at -1.13, mark anything beyond that as an outlier.)

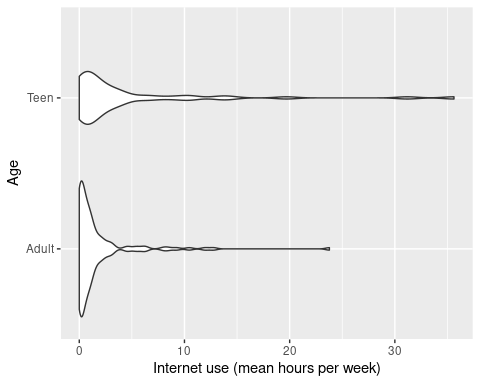
hchange <- which (abs(depress$DepressionChange) > 1.25)  
hchange

[1] 11 13 21 37 60 86 135 149

### Relationships with Age

ggplot(depress,aes(x=`Internet use (mean hours per week)`,y=Age)) +  
 geom\_violin()

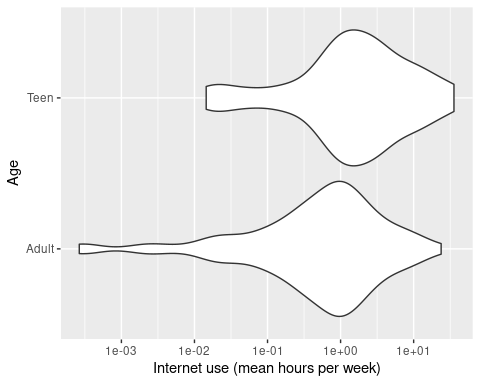
Warning: Removed 2 rows containing non-finite values (`stat\_ydensity()`).



ggplot(depress,aes(x=`Internet use (mean hours per week)`,y=Age)) +  
 geom\_violin() + scale\_x\_log10()

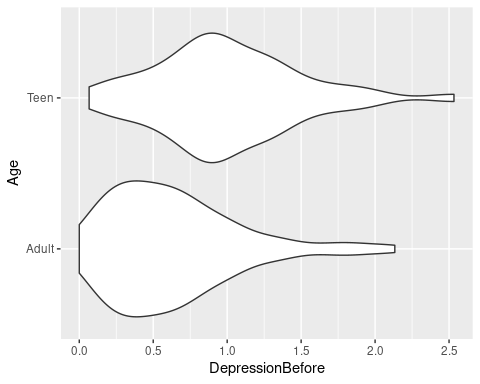
Warning: Transformation introduced infinite values in continuous x-axis

Warning: Removed 12 rows containing non-finite values (`stat\_ydensity()`).



ggplot(depress,aes(x=DepressionBefore,y=Age)) +  
 geom\_violin()

Warning: Removed 2 rows containing non-finite values (`stat\_ydensity()`).



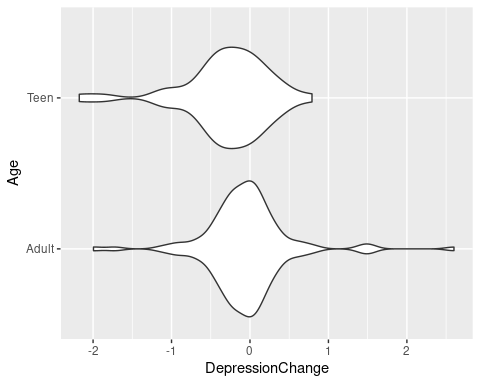
ggplot(depress,aes(x=DepressionAfter,y=Age)) +  
 geom\_violin()

Warning: Removed 5 rows containing non-finite values (`stat\_ydensity()`).



ggplot(depress,aes(x=DepressionChange,y=Age)) +  
 geom\_violin()

Warning: Removed 7 rows containing non-finite values (`stat\_ydensity()`).



## Scatterplots

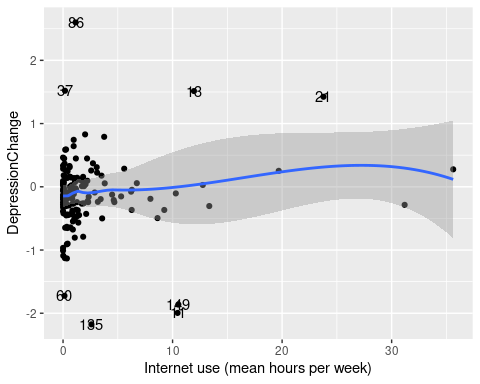
### X, Y

ggplot(depress,aes(x=`Internet use (mean hours per week)`,  
 y=DepressionChange)) +  
 geom\_point() + geom\_smooth()+  
 geom\_text(data=depress[hchange,],aes(label=hchange))

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: Removed 9 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 9 rows containing missing values (`geom\_point()`).



### log(X), Y

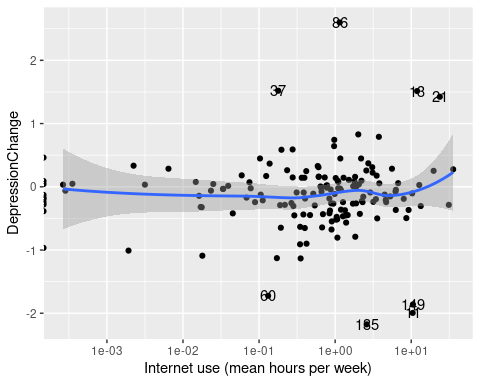
ggplot(depress,aes(x=`Internet use (mean hours per week)`,  
 y=DepressionChange)) +  
 geom\_point() + geom\_smooth()+ scale\_x\_log10() +  
 geom\_text(data=depress[hchange,],aes(label=hchange))

Warning: Transformation introduced infinite values in continuous x-axis  
Transformation introduced infinite values in continuous x-axis

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: Removed 19 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 9 rows containing missing values (`geom\_point()`).



Hmm. Note a number of points piled up on the -axis. The problem is . The solution is to add a small number before taking the logs. Looks like most values are above .001, so add .0001.

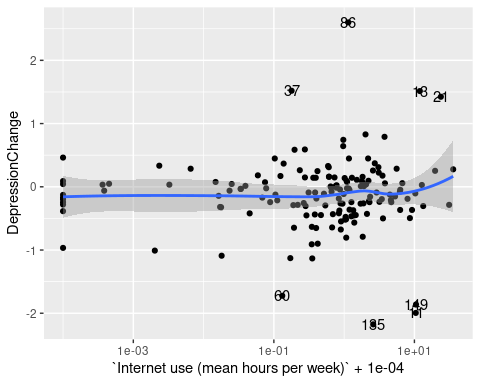
An alternative would be to exclude these people from the model.

ggplot(depress,aes(x=`Internet use (mean hours per week)`+.0001,  
 y=DepressionChange)) +  
 geom\_point() + geom\_smooth()+ scale\_x\_log10() +  
 geom\_text(data=depress[hchange,],aes(label=hchange))

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: Removed 9 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 9 rows containing missing values (`geom\_point()`).



### By Age

ggplot(depress,aes(x=`Internet use (mean hours per week)`,  
 y=DepressionChange, color=Age)) +  
 geom\_point() + geom\_smooth() +  
 geom\_text(data=depress[hchange,],aes(label=hchange))

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: Removed 9 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 9 rows containing missing values (`geom\_point()`).

ggplot(depress,aes(x=`Internet use (mean hours per week)`+.0001,  
 y=DepressionChange,color=Age)) +  
 geom\_point() + geom\_smooth()+ scale\_x\_log10() +  
 geom\_text(data=depress[hchange,],aes(label=hchange))

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: Removed 9 rows containing non-finite values (`stat\_smooth()`).  
Removed 9 rows containing missing values (`geom\_point()`).

|  |  |
| --- | --- |
|  |  |

### By Gender

ggplot(depress,aes(x=`Internet use (mean hours per week)`,  
 y=DepressionChange, color=Gender)) +  
 geom\_point() + geom\_smooth() +  
 geom\_text(data=depress[hchange,],aes(label=hchange))

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: Removed 9 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 9 rows containing missing values (`geom\_point()`).

ggplot(depress,aes(x=`Internet use (mean hours per week)`+.0001,  
 y=DepressionChange,color=Gender)) +  
 geom\_point() + geom\_smooth()+ scale\_x\_log10() +  
 geom\_text(data=depress[hchange,],aes(label=hchange))

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: Removed 9 rows containing non-finite values (`stat\_smooth()`).  
Removed 9 rows containing missing values (`geom\_point()`).

|  |  |
| --- | --- |
|  |  |

### By Race (Whiteness)

Race is coded white/non-white. So this is essentially looking at the effect of “whiteness” (whatever that is).

ggplot(depress,aes(x=`Internet use (mean hours per week)`,  
 y=DepressionChange, color=Race)) +  
 geom\_point() + geom\_smooth() +  
 geom\_text(data=depress[hchange,],aes(label=hchange))

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: Removed 9 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 9 rows containing missing values (`geom\_point()`).

ggplot(depress,aes(x=`Internet use (mean hours per week)`+.0001,  
 y=DepressionChange,color=Race)) +  
 geom\_point() + geom\_smooth()+ scale\_x\_log10() +  
 geom\_text(data=depress[hchange,],aes(label=hchange))

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

Warning: Removed 9 rows containing non-finite values (`stat\_smooth()`).  
Removed 9 rows containing missing values (`geom\_point()`).

|  |  |
| --- | --- |
|  |  |

## Regression Model

dlm <- lm(DepressionChange ~ `Internet use (mean hours per week)`,  
 data=depress)  
summary(dlm)

Call:  
lm(formula = DepressionChange ~ `Internet use (mean hours per week)`,   
 data = depress)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-2.0579 -0.2426 0.0218 0.2285 2.7314   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.141812 0.048747 -2.909 0.00415 \*\*  
`Internet use (mean hours per week)` 0.009103 0.008949 1.017 0.31061   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.5555 on 158 degrees of freedom  
 (9 observations deleted due to missingness)  
Multiple R-squared: 0.006506, Adjusted R-squared: 0.0002183   
F-statistic: 1.035 on 1 and 158 DF, p-value: 0.3106

plot(dlm)

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

### Log X

dllm <- lm(DepressionChange ~ log(`Internet use (mean hours per week)`+.0001),  
 data=depress)  
summary(dllm)

Call:  
lm(formula = DepressionChange ~ log(`Internet use (mean hours per week)` +   
 1e-04), data = depress)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-2.06940 -0.21956 0.01958 0.25194 2.71231   
  
Coefficients:  
 Estimate Std. Error t value  
(Intercept) -0.113204 0.046657 -2.426  
log(`Internet use (mean hours per week)` + 1e-04) 0.006843 0.014907 0.459  
 Pr(>|t|)   
(Intercept) 0.0164 \*  
log(`Internet use (mean hours per week)` + 1e-04) 0.6468   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.5569 on 158 degrees of freedom  
 (9 observations deleted due to missingness)  
Multiple R-squared: 0.001332, Adjusted R-squared: -0.004989   
F-statistic: 0.2108 on 1 and 158 DF, p-value: 0.6468

plot(dllm)

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

### Leverage Points

Linear scale

dfblm <- dfbetas(dlm)  
hlev <- which(abs(dfblm[,2])>2/sqrt(nrow(dfblm)))  
dfblm[hlev,]

(Intercept) `Internet use (mean hours per week)`  
11 -0.05463073 -0.4863923  
13 0.01224088 0.4538574  
21 -0.22203477 0.9756710  
134 0.12886632 -0.4591744  
149 -0.04888997 -0.4548981

Rerun without leverage points.

Linear model.

summary(lm(DepressionChange ~ `Internet use (mean hours per week)`,  
 data=depress,subset=-hlev))

Call:  
lm(formula = DepressionChange ~ `Internet use (mean hours per week)`,   
 data = depress, subset = -hlev)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-2.06071 -0.24484 0.01573 0.22688 2.72857   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.138959 0.046595 -2.982 0.00333 \*\*  
`Internet use (mean hours per week)` 0.009101 0.008696 1.047 0.29697   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.5267 on 153 degrees of freedom  
 (9 observations deleted due to missingness)  
Multiple R-squared: 0.007107, Adjusted R-squared: 0.0006178   
F-statistic: 1.095 on 1 and 153 DF, p-value: 0.297

summary(lm(DepressionChange ~ log(`Internet use (mean hours per week)`+.0001),  
 data=depress,subset=-hlev))

Call:  
lm(formula = DepressionChange ~ log(`Internet use (mean hours per week)` +   
 1e-04), data = depress, subset = -hlev)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-2.0721 -0.2112 0.0128 0.2502 2.7098   
  
Coefficients:  
 Estimate Std. Error t value  
(Intercept) -0.110741 0.045264 -2.447  
log(`Internet use (mean hours per week)` + 1e-04) 0.007048 0.014309 0.493  
 Pr(>|t|)   
(Intercept) 0.0156 \*  
log(`Internet use (mean hours per week)` + 1e-04) 0.6230   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.5281 on 153 degrees of freedom  
 (9 observations deleted due to missingness)  
Multiple R-squared: 0.001583, Adjusted R-squared: -0.004942   
F-statistic: 0.2426 on 1 and 153 DF, p-value: 0.623

## Sensitivity Analyses

### Outliers

Look at sensitivity to the outliers identified above.

summary(lm(DepressionChange ~ `Internet use (mean hours per week)`,  
 data=depress,subset=-hchange))

Call:  
lm(formula = DepressionChange ~ `Internet use (mean hours per week)`,   
 data = depress, subset = -hchange)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-1.00055 -0.23690 0.02088 0.21819 0.95097   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.134920 0.032968 -4.092 6.95e-05 \*\*\*  
`Internet use (mean hours per week)` 0.006176 0.006584 0.938 0.35   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.3694 on 150 degrees of freedom  
 (9 observations deleted due to missingness)  
Multiple R-squared: 0.005832, Adjusted R-squared: -0.0007962   
F-statistic: 0.8799 on 1 and 150 DF, p-value: 0.3497

summary(lm(DepressionChange ~ log(`Internet use (mean hours per week)`+.0001),  
 data=depress,subset=-hchange))

Call:  
lm(formula = DepressionChange ~ log(`Internet use (mean hours per week)` +   
 1e-04), data = depress, subset = -hchange)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-1.01200 -0.19932 0.01898 0.22213 0.93676   
  
Coefficients:  
 Estimate Std. Error t value  
(Intercept) -0.113503 0.032141 -3.531  
log(`Internet use (mean hours per week)` + 1e-04) 0.007466 0.010123 0.738  
 Pr(>|t|)   
(Intercept) 0.000549 \*\*\*  
log(`Internet use (mean hours per week)` + 1e-04) 0.461939   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.3698 on 150 degrees of freedom  
 (9 observations deleted due to missingness)  
Multiple R-squared: 0.003613, Adjusted R-squared: -0.003029   
F-statistic: 0.544 on 1 and 150 DF, p-value: 0.4619

### Internet Vegans

Take a closer look at the people who spend 0 hours per week on the internet. Call them *Internet Vegans*. Look at the sensitivity to them.

vegans <- which(depress$`Internet use (mean hours per week)`<.0005)  
summary(lm(DepressionChange ~ `Internet use (mean hours per week)`,  
 data=depress,subset=-vegans))

Call:  
lm(formula = DepressionChange ~ `Internet use (mean hours per week)`,   
 data = depress, subset = -vegans)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-2.05730 -0.27993 0.02558 0.25470 2.73207   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.142527 0.052922 -2.693 0.00791 \*\*  
`Internet use (mean hours per week)` 0.009160 0.009312 0.984 0.32694   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.5721 on 145 degrees of freedom  
 (9 observations deleted due to missingness)  
Multiple R-squared: 0.006628, Adjusted R-squared: -0.0002224   
F-statistic: 0.9675 on 1 and 145 DF, p-value: 0.3269

summary(lm(DepressionChange ~ log(`Internet use (mean hours per week)`+.0001),  
 data=depress,subset=-vegans))

Call:  
lm(formula = DepressionChange ~ log(`Internet use (mean hours per week)` +   
 1e-04), data = depress, subset = -vegans)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-2.07530 -0.26061 0.02199 0.25812 2.71229   
  
Coefficients:  
 Estimate Std. Error t value  
(Intercept) -0.11417 0.04806 -2.375  
log(`Internet use (mean hours per week)` + 1e-04) 0.01405 0.02518 0.558  
 Pr(>|t|)   
(Intercept) 0.0188 \*  
log(`Internet use (mean hours per week)` + 1e-04) 0.5778   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.5734 on 145 degrees of freedom  
 (9 observations deleted due to missingness)  
Multiple R-squared: 0.002141, Adjusted R-squared: -0.004741   
F-statistic: 0.3112 on 1 and 145 DF, p-value: 0.5778