Computer workshop 2 (NBS8186)

Edu Gonzalo Almorox

Introduction

These are sample solutions for the second computer lab of NBS8186. The data used for the analysis correspond to `clothing.csv`. You should have downloaded this data in a specific folder in your computer using `setwd()`.

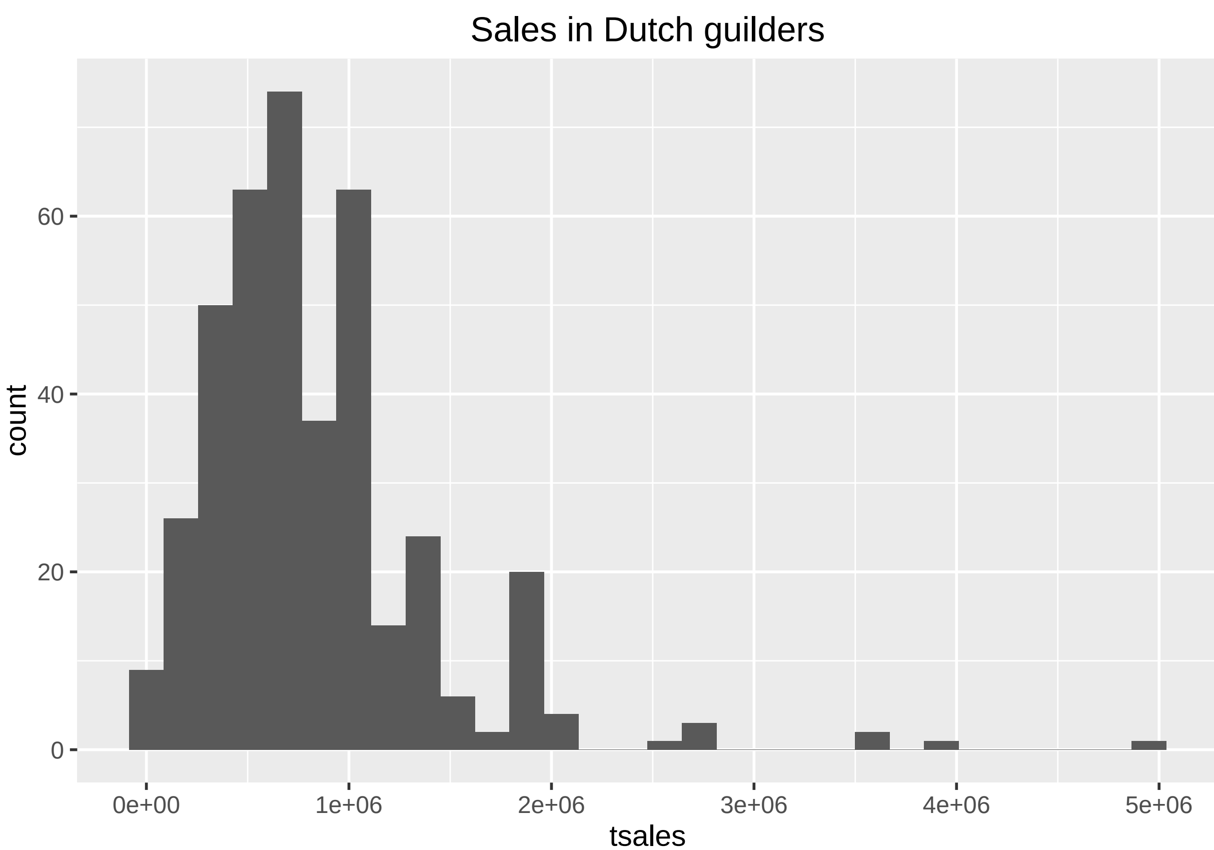
Question A

Plot a histogram for `tsales`

First we load the data setting the working directory and loading the dataset. It will be called "clothing". Unlike last computer lab this time we will use `import()` from `rio` package.

Creating a histogram is done by using `ggplot2()` which uses aesthetics of the graphics.

```
# Histogram
library(ggplot2)


m <- ggplot(clothing, aes(x = tsales))
m + geom_histogram() + ggtitle("Sales in Dutch guilders")
```



What are the mean and the median?

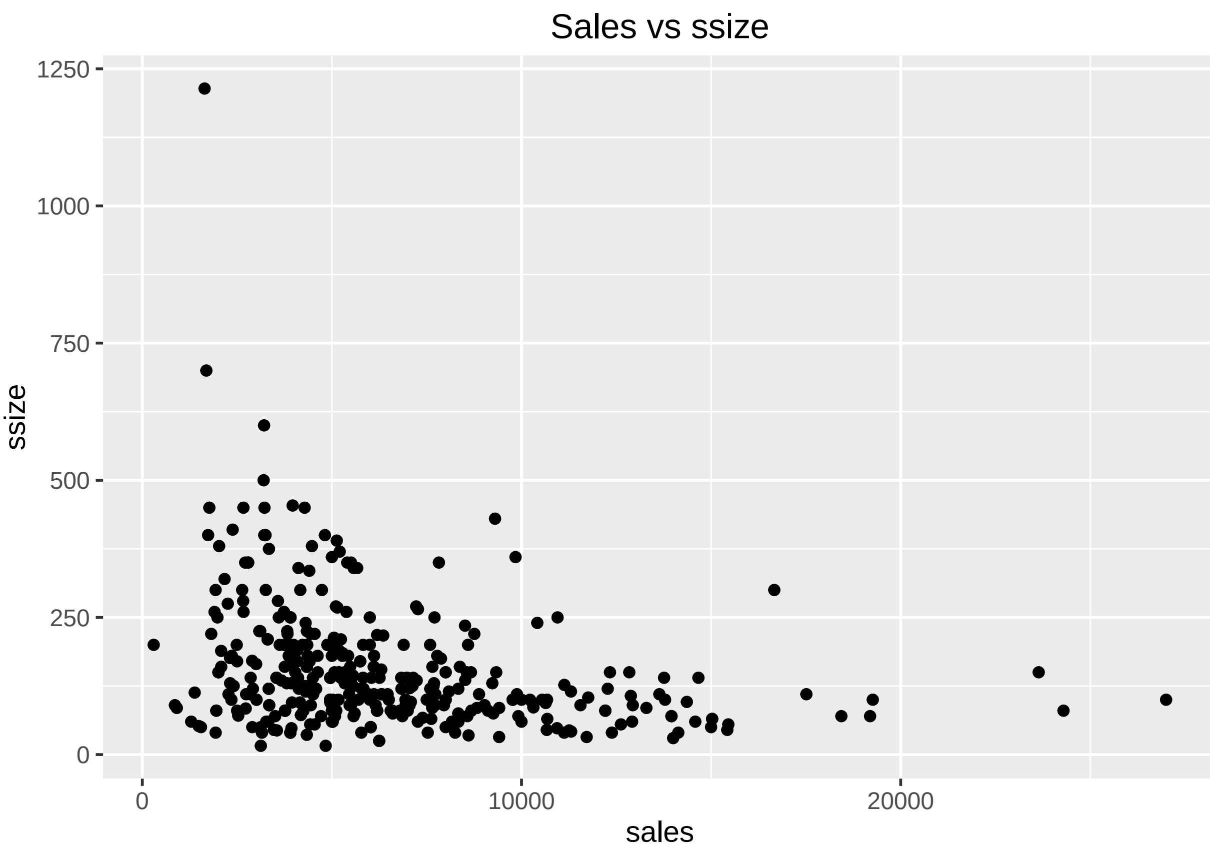
We have several options. We could use `summary()` and calling `tsales`. Alternatively, we could calculate the mean and median applying directly the `mean()` and `median()` functions on `tsales`.

```
# mean and median

summary(clothing$tsales)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   50000  495300  694200  833600  976800 5000000
```

Plot `sales` against `ssize`

```
# Histogram

library(ggplot2)
p <- ggplot(clothing, aes(sales, ssize))
p + geom_point() + ggtitle("Sales vs ssize")
```

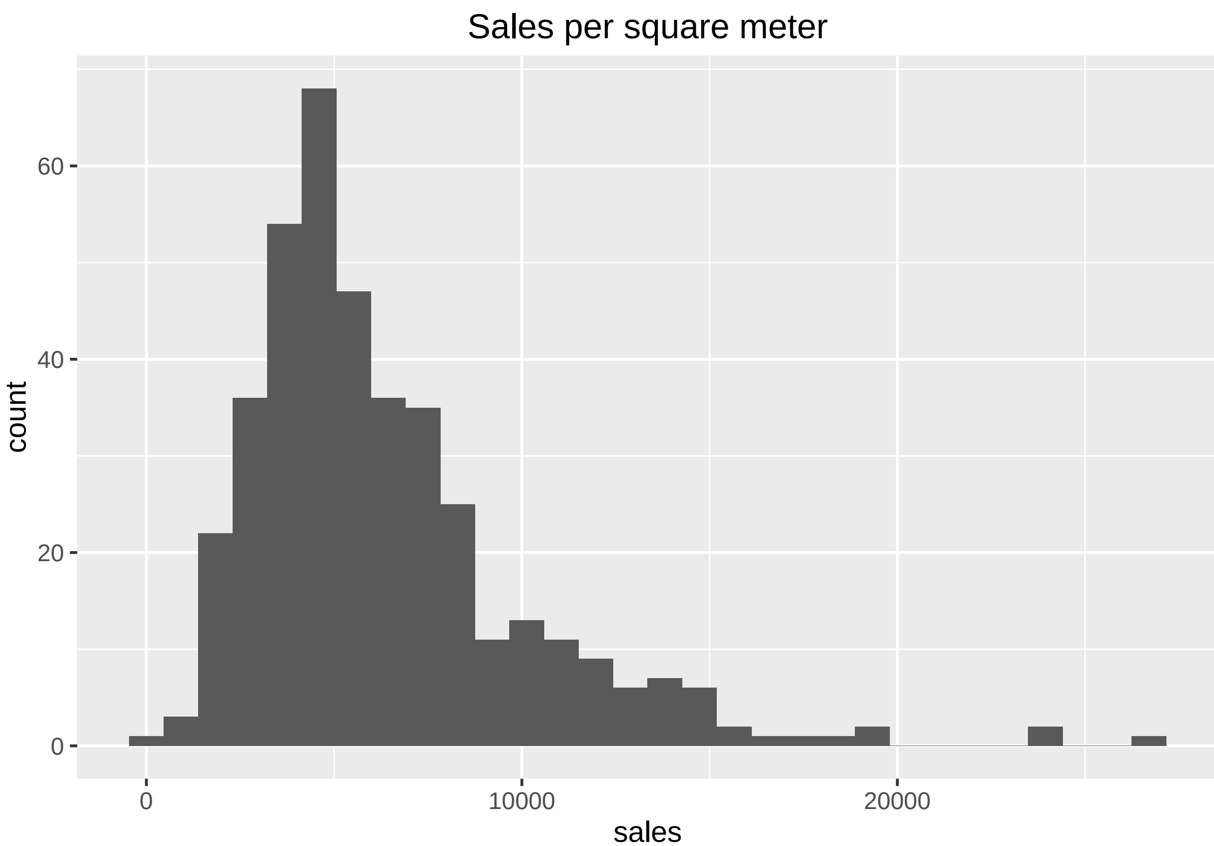


Question B

Redo a) considering `sales`

```
# Histogram
library(ggplot2)
```

```r
m <- ggplot(clothing, aes(x = sales))
m + geom_histogram() + ggtitle("Sales per square meter")
```



The `mean` and the `median` are calculated using the functions `mean` and `median`[1]

```r
# mean

mean(clothing$sales)
```

```
## [1] 6334.751
```

```r
# median

median(clothing$sales)
```

```
## [1] 5278.935
```

Question C

Regress sales on ssize. Interpret.

[1] Results can be checked using `summary()`

```r
mod1 = lm(sales ~ ssize, clothing)
summary(mod1)
```

```
##
## Call:
## lm(formula = sales ~ ssize, data = clothing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6071.3 -2194.5  -813.2  1139.8 20166.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7809.808    299.817  26.049  < 2e-16 ***
## ssize         -9.765      1.593  -6.132  2.1e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3579 on 398 degrees of freedom
## Multiple R-squared:  0.08631,    Adjusted R-squared:  0.08402
## F-statistic:  37.6 on 1 and 398 DF,  p-value: 2.097e-09
```

This simple model is a simple linear regression where we analyse the relationship between to variables. A dependent variables `sales` and a regressor `ssize`. We want to see to what extent the sales floor space of the store is related to the number of sales per square meter.

The can see that there is a negative relationship so that an additional square meter in the floor space supposes almost 10 sales less (9.76 exactly). This negative effect is statistically significant.

Question D

Regress sales on ssize and ssize squared. Interpret. Is there evidence for a nonlinear relationship? If yes, what type of extremum do you find?

First we have to create the variable ssize squared `ssize2`. Then we run the model with the new variable created. Since there is more than one regressor, we are fitting a multiple linear regression.

By adding the squared regressor we are assuming that the relationship between that regressor and the dependent variable is going to change *-wears off-* at some point. The value of the estimate of the squared term indicates actually the turning point of the relationship.

```r
clothing$ssize2 = clothing$ssize^2

mod2 = lm(sales ~ ssize + ssize2, clothing)

summary(mod2)
```

```
##
## Call:
## lm(formula = sales ~ ssize + ssize2, data = clothing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -6217.3 -2104.5  -710.9  1243.9 20086.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.419e+03  3.957e+02  21.277  < 2e-16 ***
## ssize       -1.598e+01  3.092e+00  -5.168 3.75e-07 ***
## ssize2       9.312e-03  3.979e-03   2.340   0.0198 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3559 on 397 degrees of freedom
## Multiple R-squared:  0.09875,    Adjusted R-squared:  0.09421
## F-statistic: 21.75 on 2 and 397 DF,  p-value: 1.089e-09
```

If the sign of the squared regressor is poitive, the relationship is a convex model (so it is a *minimum*) and conversely if the sign is negative then the curve is concave (and therefore a *maximum*).

In our case, the sign of the squared variable is positive so it would suppose a minimum and it would only significant at 0.05 level of significance.

Question E

Regress sales on nown, nfull, npart, naux, inv1, inv2, ssize and ssize squared

```r
mod3 =lm( sales~nown+nfull+npart+naux+inv1+inv2+ssize+ssize2, clothing)
summary(mod3)
```

```
##
## Call:
## lm(formula = sales ~ nown + nfull + npart + naux + inv1 + inv2 +
##     ssize + ssize2, data = clothing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6809.2 -2095.8  -218.7  1398.9 19043.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.729e+03  8.358e+02   4.462 1.07e-05 ***
## nown         9.331e+02  2.558e+02   3.648 0.000301 ***
## nfull        1.298e+03  1.752e+02   7.413 7.72e-13 ***
## npart        5.727e+02  2.551e+02   2.245 0.025319 *
## naux         4.968e+02  4.250e+02   1.169 0.243127
## inv1         6.110e-04  1.754e-03   0.348 0.727738
## inv2         1.079e-03  4.472e-03   0.241 0.809517
## ssize       -2.111e+01  2.961e+00  -7.130 4.87e-12 ***
## ssize2       7.250e-03  3.685e-03   1.968 0.049813 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3238 on 391 degrees of freedom
## Multiple R-squared:  0.2653, Adjusted R-squared:  0.2502
## F-statistic: 17.65 on 8 and 391 DF,  p-value: < 2.2e-16
```

(i) Interpret your results.

All the variables have a positive association with sales excepting the space of the floor for sales.

(ii) Is the regression significant

Yes because of the F-Statistic.

(iii) Test whether $\beta_{inv} = 0$

It is. `Pr(>|t|)` is greater than any other value of significance.

(iv) Test whether $\beta_{nown} = 1000$

A general procedure to test the value of a coefficient against an alternative value to 0 consists of calculating the density function of the *t statistic*

$$t = \frac{\hat{\beta} - \beta_{H_0}}{s.e(\hat{\beta})} \tag{1}$$

`tidy()` creates a `data.frame` with the results of the regression.

```
library(broom)
mod3.tidy = tidy(mod3)

t = (mod3.tidy[2, 2] - 1000)/mod3.tidy[2,3]
pt(t, df = 391)
```

```
## [1] 0.3969924
```

We cannot reject the H_0 at a level of significance $\alpha > 0.1$ so that we would say that $\beta_{nown} = 1000$.

(v) Test whether $\beta_{nfull} = 2\beta_{npart}$

Applying (1) we can calculate the following

$$t.1 = \frac{\hat{\beta}_n full - \beta_{H_0}}{s.e(\hat{\beta})} \tag{2}$$

```
library(broom)
mod3.tidy = tidy(mod3)

t.1 = (mod3.tidy[3,2] - 2*(mod3.tidy[4,2]))/mod3.tidy[3,2]
pt(t.1, df = 391)
```

```
## [1] 0.5468764
```

We cannot reject the H_0 at a level of significance $\alpha > 0.1$ so that we would say that $\beta_{nown} = 1000$.

(vi) Use a Chow test to see whether the relationship is the same for stores with start $>= 40$ *and* $<= 40$.

The Chow test tests the implicit assumption of β are constant over the whole sample. Essentially what we are testing is whether there are structural breaks $H_0 : \beta_{ur1} = \beta_{ur2}$ and $H_1 : \beta_{ur1} \neq \beta_{ur2}$

The procedure consists of various steps

- Run a restricted regression

- Divide the sample into tow groups that are detemined by the breakpoint ($sales >= 40$)
- Run an "unrestricted" regression on each of your subsamples. You will run two "unrestricted" regressions with a single breakpoint.
- Calculate the Chow F-statistic as follows

$$\frac{SSR_r - SSR_u/k}{SSR_u/(n-2k)} = F_{k,n-2k} \tag{3}$$

```r
# Step 1: Create the regression
mod3.1 =lm(sales~nown+nfull+npart+naux+inv1+inv2+ssize+ssize2, subset(clothing,
                                                                     start <= 40))
summary(mod3.1)
```

```
##
## Call:
## lm(formula = sales ~ nown + nfull + npart + naux + inv1 + inv2 +
##     ssize + ssize2, data = subset(clothing, start <= 40))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6076.9 -1968.0  -193.6  1175.0 19793.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.595e+03  1.306e+03   2.753 0.006440 **
## nown         1.119e+03  3.202e+02   3.494 0.000584 ***
## nfull        1.117e+03  3.082e+02   3.626 0.000365 ***
## npart        4.277e+02  5.438e+02   0.787 0.432439
## naux         3.259e+02  5.498e+02   0.593 0.553999
## inv1         1.097e-03  2.424e-03   0.453 0.651384
## inv2        -3.489e-05  6.736e-03  -0.005 0.995873
## ssize       -1.971e+01  6.870e+00  -2.869 0.004560 **
## ssize2       9.329e-03  1.361e-02   0.685 0.493839
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3350 on 202 degrees of freedom
## Multiple R-squared:  0.2166, Adjusted R-squared:  0.1856
## F-statistic: 6.982 on 8 and 202 DF,  p-value: 4.035e-08
```

```r
mod3.2 <- lm(sales~nown+nfull+npart+naux+inv1+inv2+ssize+ssize2, subset(clothing,
                                                                       start>40))
summary(mod3.2)
```

```
##
## Call:
## lm(formula = sales ~ nown + nfull + npart + naux + inv1 + inv2 +
##     ssize + ssize2, data = subset(clothing, start > 40))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5743.2 -1965.5  -221.2  1504.8 14737.5
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.521e+03  1.314e+03   3.441  0.00072 ***
## nown         6.133e+02  4.605e+02   1.332  0.18458
## nfull        1.363e+03  2.171e+02   6.279 2.48e-09 ***
## npart        5.232e+02  2.856e+02   1.832  0.06865 .
## naux         9.508e+02  7.062e+02   1.346  0.17988
## inv1        -2.194e-04  2.731e-03  -0.080  0.93604
## inv2         2.910e-03  6.167e-03   0.472  0.63756
## ssize       -2.502e+01  3.847e+00  -6.504 7.50e-10 ***
## ssize2       9.534e-03  4.136e-03   2.305  0.02229 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3070 on 180 degrees of freedom
## Multiple R-squared:  0.3521, Adjusted R-squared:  0.3233
## F-statistic: 12.23 on 8 and 180 DF,  p-value: 6.521e-14
```

```r
# Step 2: Create the residuals
SSR = NULL
SSR$r = mod3$residuals^2
SSR$ur1 = mod3.1$residuals^2
SSR$ur2 = mod3.2$residuals^2

K = mod3$rank

# Step 3: Compute the Chow
numerator = (sum(SSR$r) - (sum(SSR$ur1) + sum(SSR$ur2)) ) / K
denominator = (sum(SSR$ur1) + sum(SSR$ur2))/(nrow(clothing) - 2*K)

chow = numerator / denominator
chow
```

```
## [1] 1.452919
```

```r
# Step 4: Compute the p-value

pchow = 1-pf(chow, K, (nrow(clothing) - 2*K))
pchow
```

```
## [1] 0.1637574
```

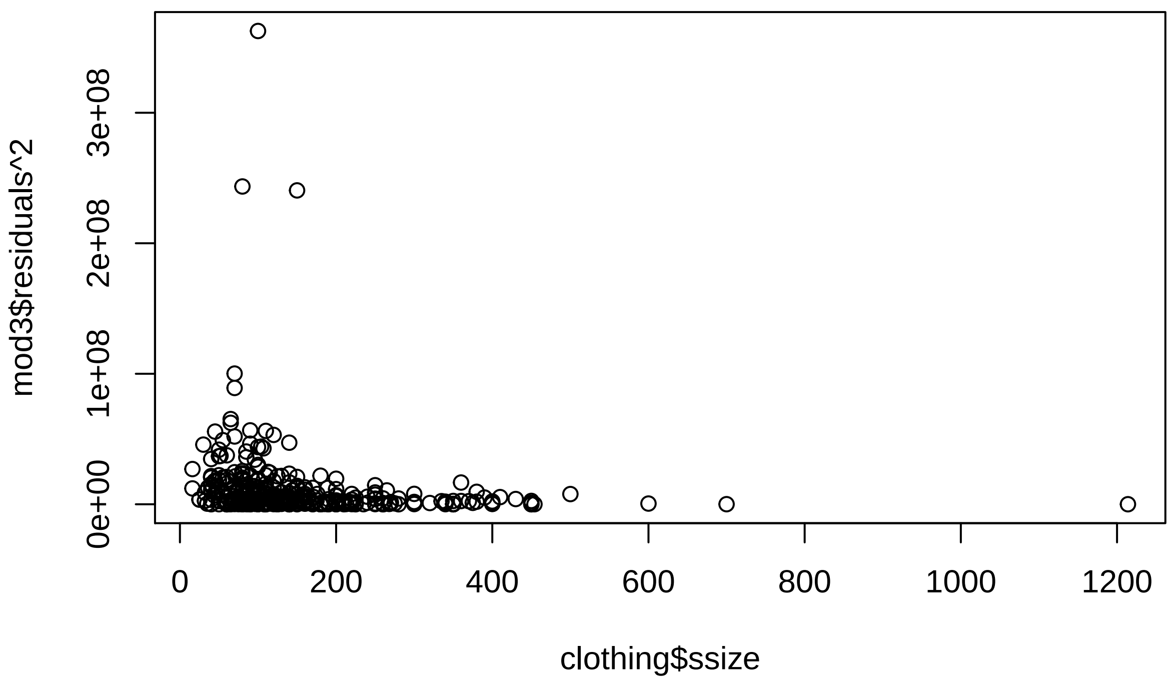
We cannot reject the H_0 so we can conclude that the relationship is the same with stores that started before the 40s and after.

Question F

Plot the squared residuals from the original regression in (e) against the explanatory variables. Do you find evidence of heteroskedasticity? How could you test for heteroskedasticity using a regression?

Homocedasticity is an assumption of the classical (linear) model. Under homocedasticity the error terms are constant and have the value of the variance ($V(\epsilon_i) = \sigma^2$). In case the former does not hold, then we have heterokedasticity. You can have a visual analysis of the heterokedasticity by plotting the residuals of the model against the explanatory variable

```r
# Pattern of heterokedasticity
plot(clothing$ssize, mod3$residuals^2 )
```



In case you want to use regression, then you have to regress the squared residuals against the explanatory variable. The estimates are considered how much the dependent variable changes under changes of the dependent variables. If $\beta = 0$ then it does not change with additional units and therefore there is homocedasticity.

```r
m4 <- lm( mod3$residuals^2 ~ ssize, clothing)   # slope is significant -> heteroskedasticity
summary(m4)
```

```
##
## Call:
## lm(formula = mod3$residuals^2 ~ ssize, data = clothing)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -14506454  -9597740  -5881929    732269 350444830
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16063578    2276515   7.056 7.64e-12 ***
## ssize         -38503      12093  -3.184  0.00157 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27170000 on 398 degrees of freedom
## Multiple R-squared:  0.02484,    Adjusted R-squared:  0.02239
## F-statistic: 10.14 on 1 and 398 DF,  p-value: 0.001567
```