EU regional models

Bodo Balazs, Daniel Antal, CFA

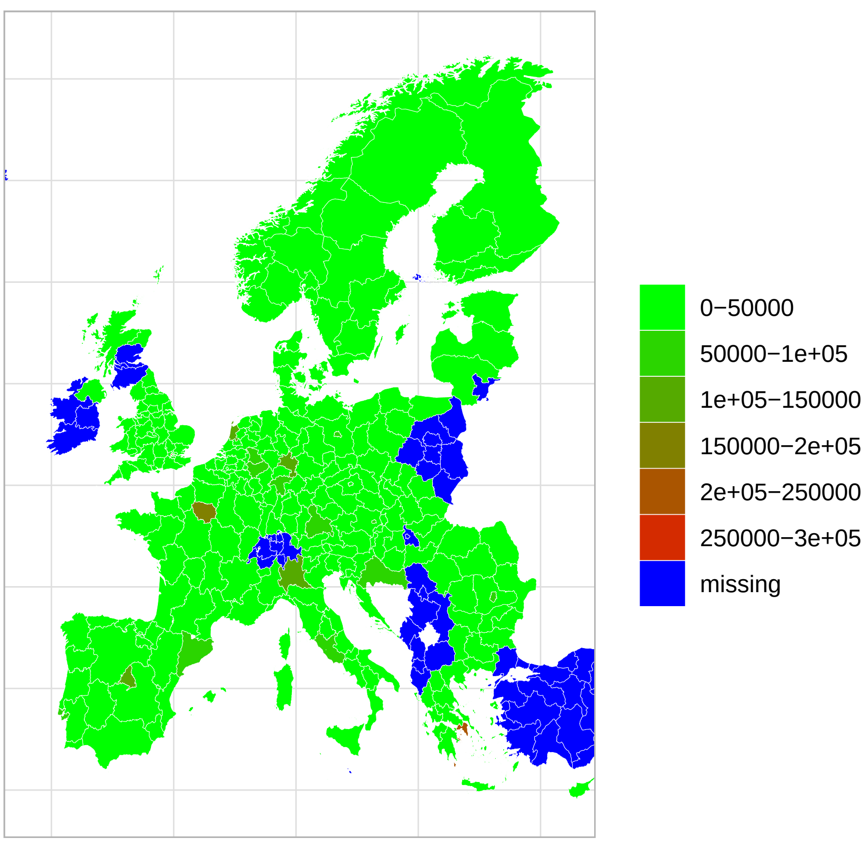
04/05/2020

Introduction

The following document contains the analysis of illegal library downloads on an European NUTS2 regional level. We used EUROSTAT and EUROBAROMETER data sources to compile two data sets. The dataset which only contains 17 explanatory variables data from the EUROSTAT database covers 265 NUTS2 regions, while the second dataset, which also includes ˆ additional variables from the EUROBAROMETER database covers 217 NUTS2 regions. We describe the dataset used in the analysis in a separate document in this repository.

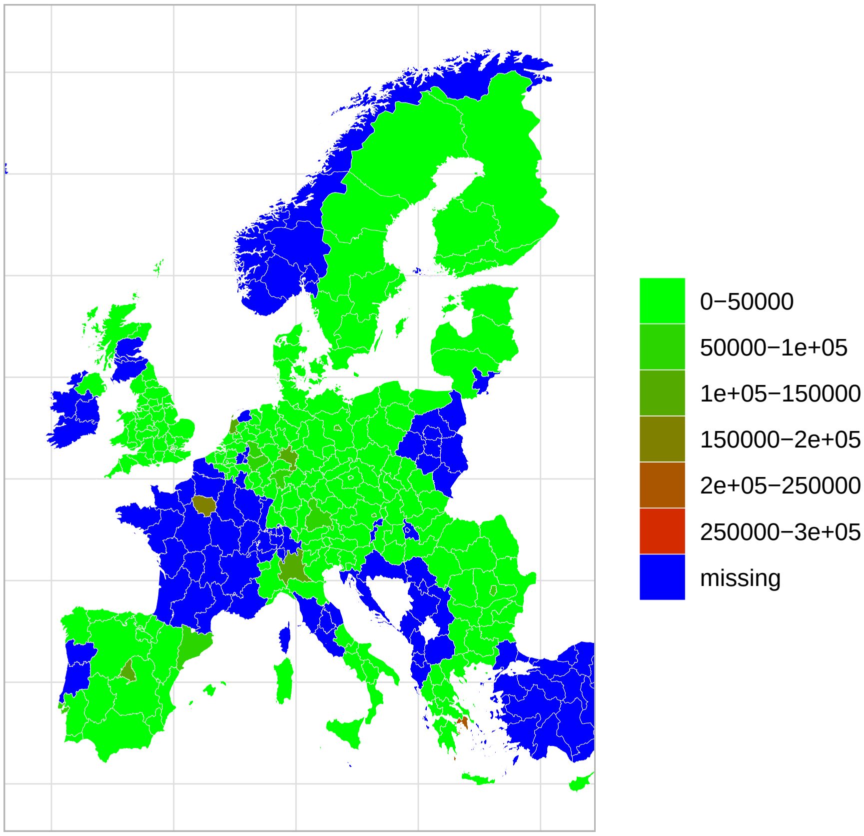
The complete cases for the eurostat dataset

The map below shows the total number of downloads per region, and the completeness of that dataset.



The complete cases for the eurostat+eurobarometer dataset

The map below shows the total number of downloads per region, and the completeness of the second, richer, but smaller dataset.



Analysis

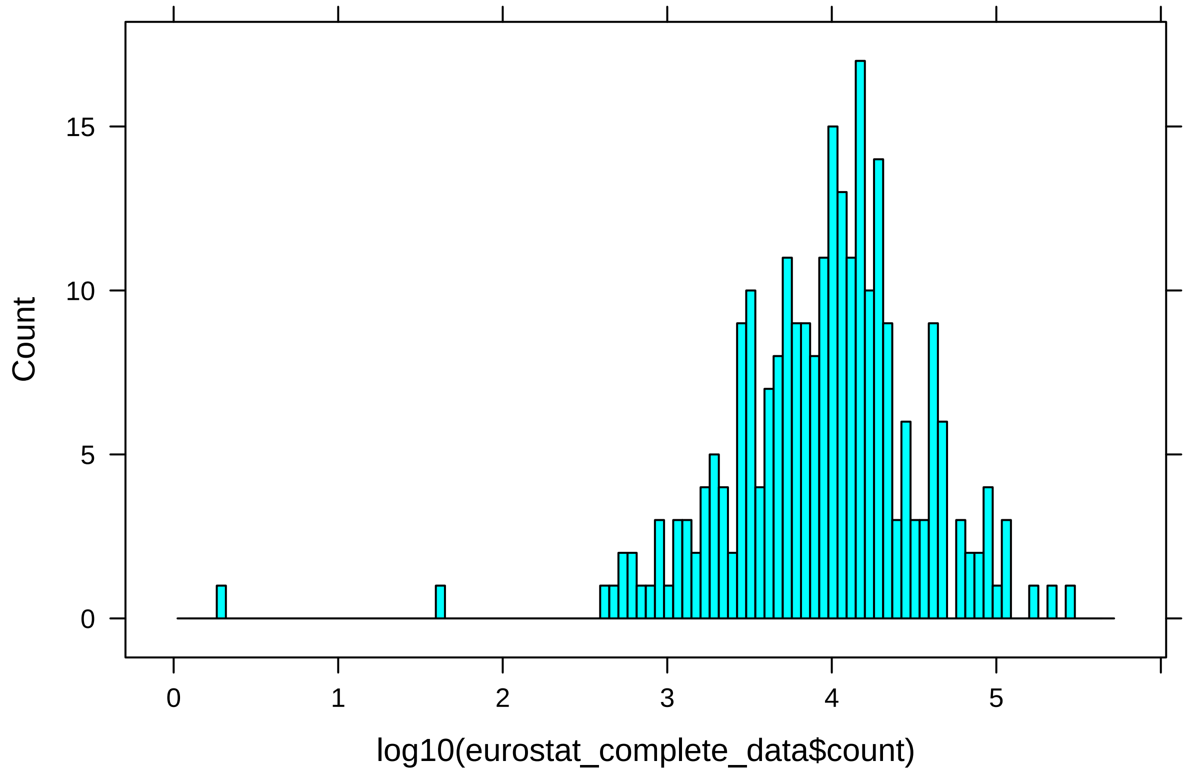
Descriptives, and general concerns

The median number of downloads is 10k, the mean is higher, 18648. With some extreme outliers.

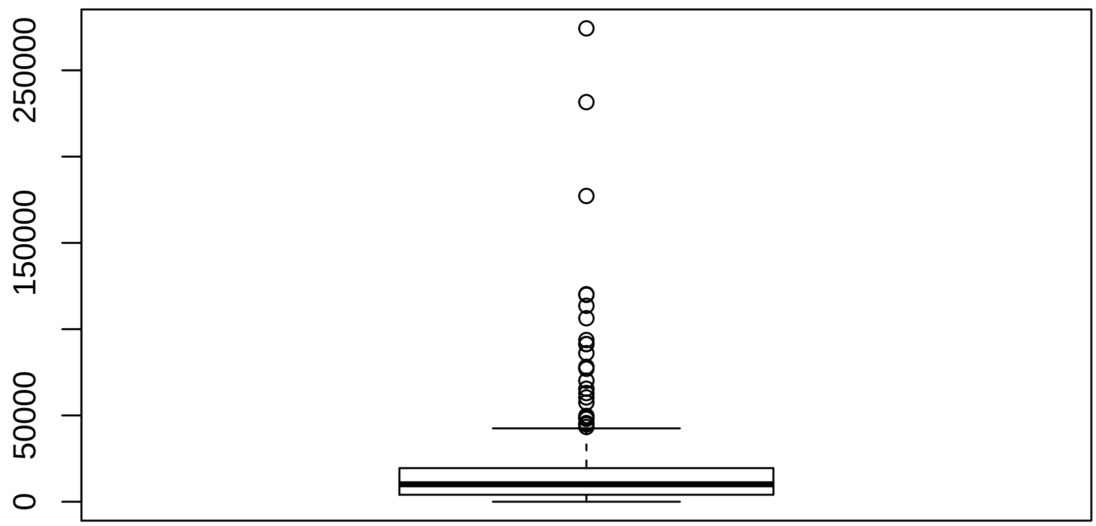
```
summary(eurostat_complete_data$count)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    4038   10095   18648   19477  274309
```

```
histogram(log10(eurostat_complete_data$count),
          nint=100,
          type="count")
```



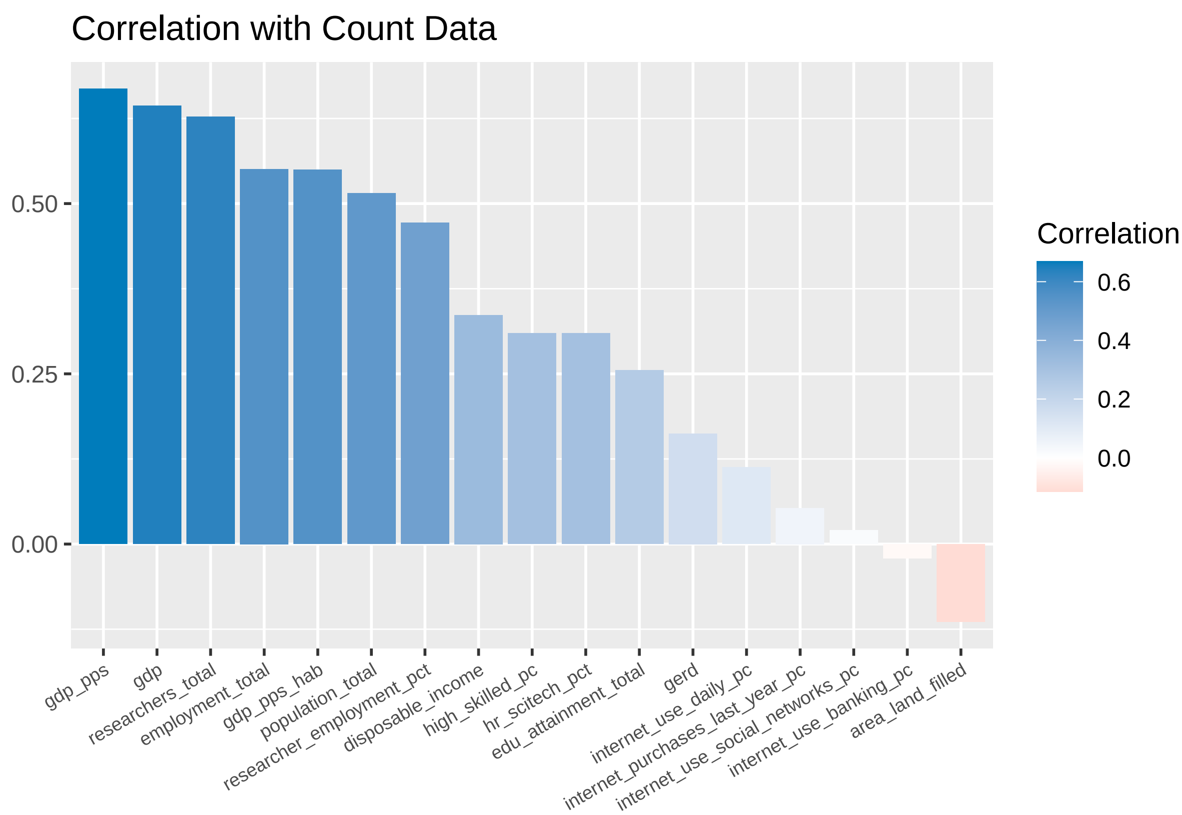

```
boxplot(eurostat_complete_data$count)
```



Some of the other regions (see the maps above) are as expected, big, metropolitan regions, such as inner London, with large populations, and a strong concentration of knowledge-intensive activities. There are, however, exceptions to this rule, where regions without significant urban centers, or educational, research capacities demonstrate unusually high download volumes. We identified a number of possible reasons for these anomalies:

- there might be issues with the translation of IP addresses to geolocation. We used the MaxMind service to assign coordinates to IP addresses, and the accuracy of the service, may vary for different countries, or internet service providers. In this latter case, it is possible that lacking better information, whole IP ranges resolve to, for example, the HQ address of the provider, rather than to the approximate location of the user. This is a well-known issue in general, and a potential source of noise in our case as well.
- We did our best to identify Virtual Private Networks, TOR exit nodes, and other traffic sources, which may mask the true location of the downloader. However, such information may not always be available, therefore it is possible that we failed to identify traffic sources as VPNs. In such cases, we incorrectly associate substantial foreign traffic with a particular geographic location.
- Last, but not least, though we tried our best to identify bots and other automated traffic sources in the dataset. For example, we filtered repeated downloads from the same IP of the same book within a given time-window. However, it is possible that we did not identified all the automatic scraping, which does not represent human downloaders. In other parts of the dataset we have evidence for such automated, scraper-generated traffic, and on smaller scale, this might also produce unexpected outliers in the European dataset.

That being said, if we look at the degrees of correlation between or ultimate dependent variable, the number of downloads in a region, and other variables, we see strong correlations, especially with wealth (measured by GDP), the number of researchers, population, and knowledge intensive economic activities.



Spatial auto-correlation

As a first step in the analysis, we consider the spatial distribution of the data. If the spatial geography of the environment is relevant to the data, then we should see a level autocorrelation by nearer territorial units. We have examined the spatial autocorrelation using the `spdep` package of Bivand, Pebesma and Gomez-Rubio (**???**).

```
## Loading required package: sp
```

Moran's I statistic takes the value of 0.042 with a p-value of 0.094, so we can only reject the randomness of downloads at a 90% significance level. The positive z value means that the downloads are clustering, i.e. NUTS2 regions with high download numbers tend to be neighbors of NUTS2 regions with high download numbers.

```
##
##  Monte-Carlo simulation of Moran I
##
## data:  moran_i_spdf %>% dplyr::select(count) %>% unlist() %>% as.numeric()
## weights: ww
## number of simulations + 1: 1000
##
## statistic = 0.042747, observed rank = 904, p-value = 0.096
## alternative hypothesis: greater
```

Similarly, running the same test for GDP adjusted by purchasing power standard, we see a very similar level of spatial autocorrelation.

```
##
```

```
##  Monte-Carlo simulation of Moran I
##
## data:  moran_analysis_spdf_gdp_pps %>% dplyr::select(gdp_pps) %>% unlist() %>%
## weights: ww2
## number of simulations + 1: 1000
##     as.numeric()
## weights: ww2
## number of simulations + 1: 1000
##
## statistic = 0.044017, observed rank = 923, p-value = 0.077
## alternative hypothesis: greater
```

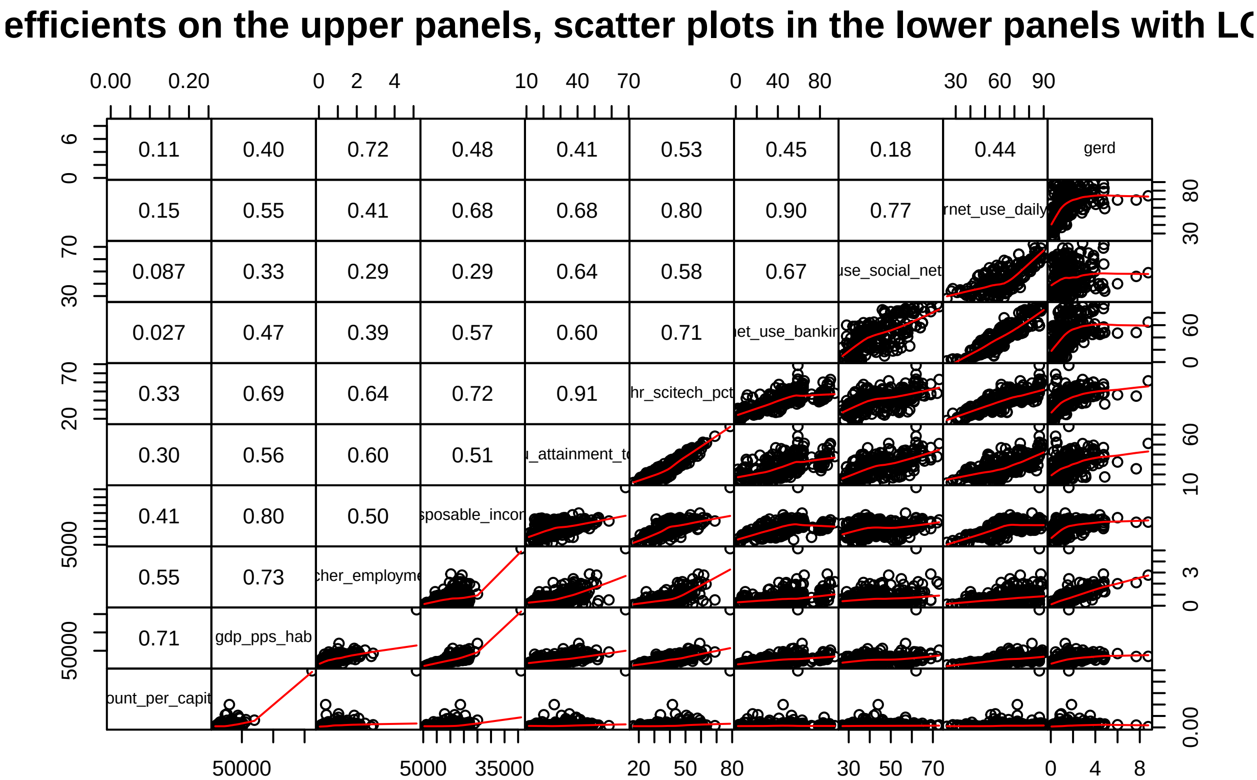
Interaction with environmental variables

Next, we pursued the following modeling approach:

- first, we tried to test on the European dataset, the same hypothesis that we tested on the global data, namely that lower income regions compensate their infrastructural shortcomings by the more extensive use of piratical resources. We control for wealth, and knowledge intensive macro-economic variables, such as R7D sending, or the share of researchers in the active population. This would be a limited model in terms of independent variables, but which, in return, would allow us to include the most NUTS2 regions in the analysis.
- Second, we used alternative modeling techniques, such as random forest methods, to check if we could find additional variables which we could be included in our models. In this step we run this analysis on the narrower, but more larger EUORUSTAT dataset.
- Third, we add the EUROBAROMETER variables, at the expense of reducing somewhat the size of the dataset, and use the random forest approach to identify if there are important new explanatory variables among the newly added ones,
- lastly, we re-run any liner regression models if the random forest identified new variables.

In each of these steps we test the models for three dependent variables: (1) the raw download count, (2) the download count normalized by the population, and (3) the download count normalized by the number of researchers. To be able to use Poisson and quasipoisson models we normalized the count variables per million inhabitants or researchers, and rounded the results.

Hypothesis testing through simple linear regressions



per capital download models

```r
#poisson with scitech hr
percapita_plm1 <- glm (count_per_million ~
                 gdp_pps_hab +
                 researcher_employment_pct +
                 disposable_income +
                 edu_attainment_total +
                 hr_scitech_pct +
                 internet_use_banking_pc,
                 data = eurostat_complete_data,
                 family = poisson )


#poisson withoutr scitech hr
percapita_plm2 <- glm (count_per_million ~
                         gdp_pps_hab +
                         researcher_employment_pct +
                         disposable_income +
                         edu_attainment_total +
                         internet_use_banking_pc,

                        data = eurostat_complete_data,
                        family = poisson )
```

```r
#Switch to log(gdp_pps)

percapita_plm3 <- glm (count_per_million ~
                         log(gdp_pps) +
                         researcher_employment_pct +
                         disposable_income +
                         edu_attainment_total +
                         internet_use_banking_pc,

                         data = eurostat_complete_data,
                         family = poisson )

#quasipoisson

percapita_qplm3 <- glm (count_per_million ~
                          log(gdp_pps) +
                          researcher_employment_pct +
                          internet_use_banking_pc,

                        data = eurostat_complete_data,
                        family = quasipoisson)

##try online shopping instead of banking
percapita_qplm4 <- glm (count_per_million ~
                          log(gdp_pps) +
                          researcher_employment_pct +
                          internet_purchases_last_year_pc,

                          data = eurostat_complete_data,
                          family = quasipoisson)

##try disposable income and edu level
percapita_qplm5 <- glm (count_per_million ~
                          log(gdp_pps) +
                          researcher_employment_pct +
                          internet_use_banking_pc +
                          log(disposable_income) ,

                        data =eurostat_complete_data,
                        family = quasipoisson)

percapita_qplm6 <- glm (count_per_million ~
                          log(gdp_pps) +
                          researcher_employment_pct +
                          internet_use_banking_pc +
                          gerd,

                        data =eurostat_complete_data,
                        family = quasipoisson)

percapita_qplm7 <- glm (count_per_million ~
                          log(gdp_pps) +
                          internet_use_banking_pc +
```

```r
                          gerd,

                        data =eurostat_complete_data,
                        family = quasipoisson)
vif(percapita_qplm7)
```

```
##             log(gdp_pps) internet_use_banking_pc                    gerd
##                 1.119251                1.229634                1.338946
```

```r
#summary(percapita_plm3)
#summary(percapita_qplm3)

export_summs(percapita_qplm3, percapita_qplm4, percapita_qplm5, percapita_qplm6,percapita_qplm7,
             digits=3,
             statistics = c("null.deviance", "deviance")
             )
```

```
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.

## Warning in if (statistics == "all") {: the condition has length > 1 and only the
## first element will be used

## Registered S3 methods overwritten by 'broom.mixed':
##   method         from
##   augment.lme    broom
##   augment.merMod broom
##   glance.lme     broom
##   glance.merMod  broom
##   glance.stanreg broom
##   tidy.brmsfit   broom
##   tidy.gamlss    broom
##   tidy.lme       broom
##   tidy.merMod    broom
##   tidy.rjags     broom
##   tidy.stanfit   broom
##   tidy.stanreg   broom
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
```

	Model 1	Model 2	Model 3	M
(Intercept)	6.438 ***	6.295 ***	-0.143	
	(0.794)	(0.838)	(2.457)	
log(gdp_pps)	0.247 **	0.242 **	0.175 *	
	(0.077)	(0.081)	(0.075)	
researcher_employment_pct	0.697 ***	0.683 ***	0.570 ***	
	(0.057)	(0.063)	(0.068)	
internet_use_banking_pc	-0.011 ***		-0.015 ***	
	(0.003)		(0.003)	
internet_purchases_last_year_pc		-0.006		
		(0.003)		
log(disposable_income)			0.792 **	
			(0.280)	
gerd				
null.deviance	2990524.371	2990524.371	2990524.371	29905
deviance	1415805.433	1507337.393	1343129.996	13968

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

```
#summary (percapita_qplm4)
```

We first developed a number of models with download per million as the dependent variable. We found that adding the sci-tech employment variable causes serious multicollinearity issues, so we decided to drop it. We also switched gdp_pps_hab to the logarithmic form of GDP_pps because it also created multicollinearity issues. We report here only the results of four quasipoission models.

The models offer the following findings:

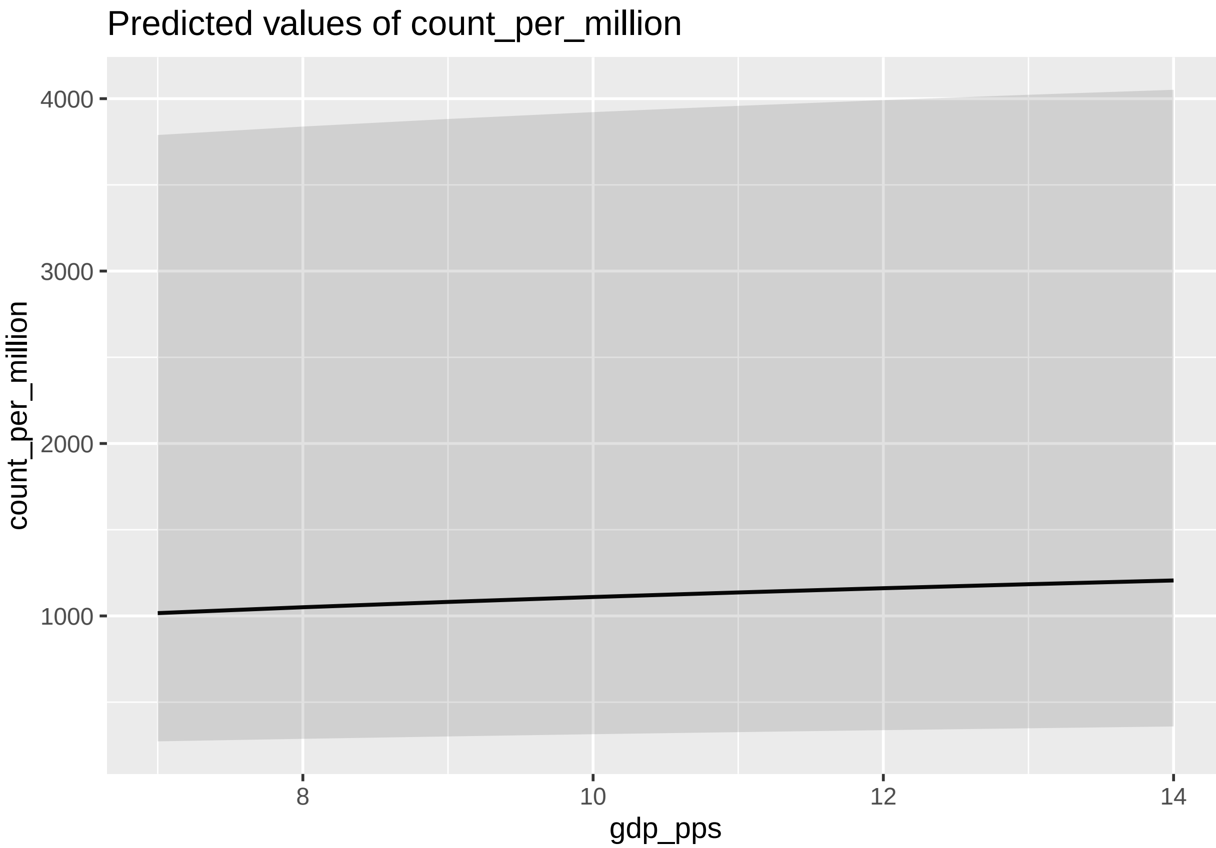
- gdp is has a significant positive effect on the per capita downloads. wealthier regions download more.
- the per capita downloads grow with higher percentages of researchers in the labor pool. Researchers are a primary source of download traffic
- higher disposable income also leads to higher download activity.

These three effects point to different forms of structural demand effect. Economic activity, research activity drives demand for scientific literature. The disposable income points to an individual demand effect: higher disposable income does not lower piracy, but actually creates more demand.

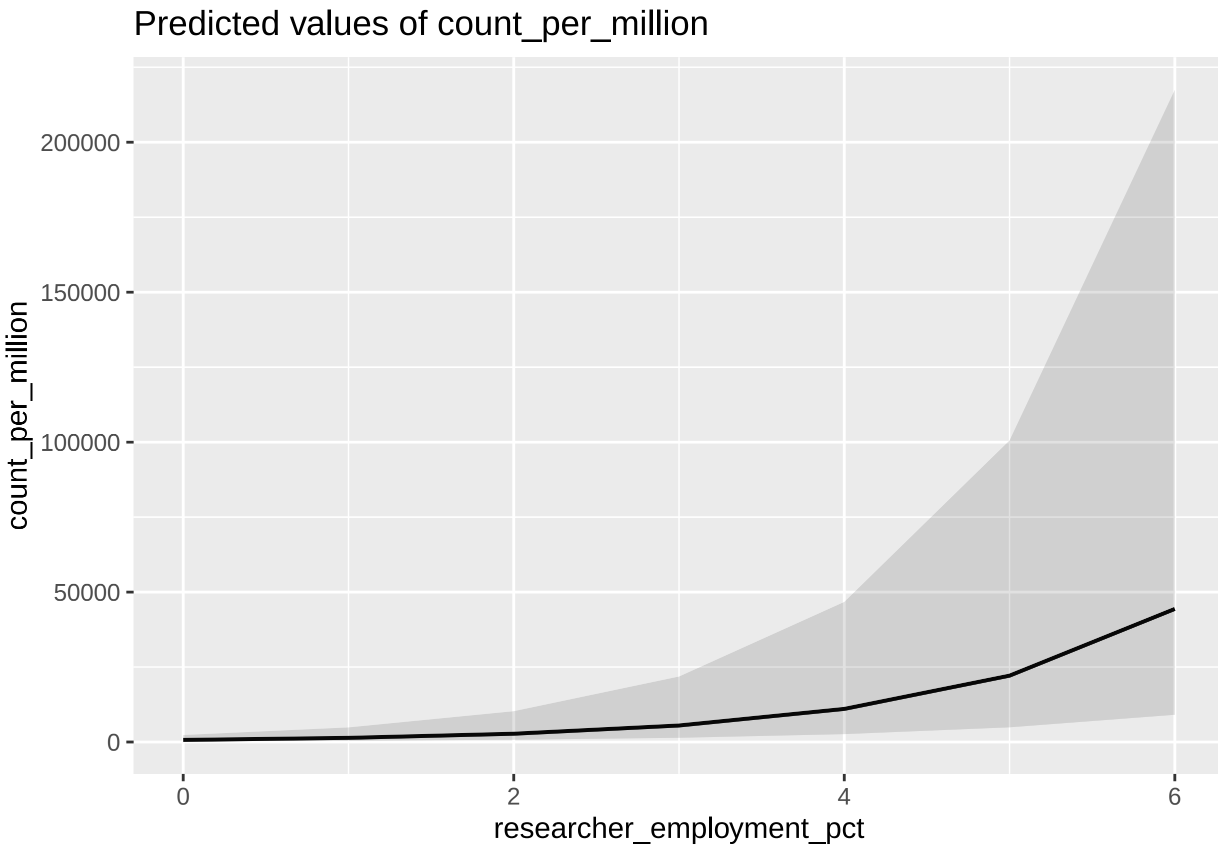
- on the other hand, per capita downloads are moderated by better online skills. The negative effect of online banking use may point to a higher use of legal sources, such as online and offline purchases, but we should also consider that online proficiency provides the skills to hide the online traces of illegal activities via the use of VPNs and Tor browsing.

```
## Model has log-transformed predictors. Consider using `terms="gdp_pps [exp]"` to back-transform scale
## Model has log-transformed predictors. Consider using `terms="gdp_pps [exp]"` to back-transform scale
## Model has log-transformed predictors. Consider using `terms="gdp_pps [exp]"` to back-transform scale
```

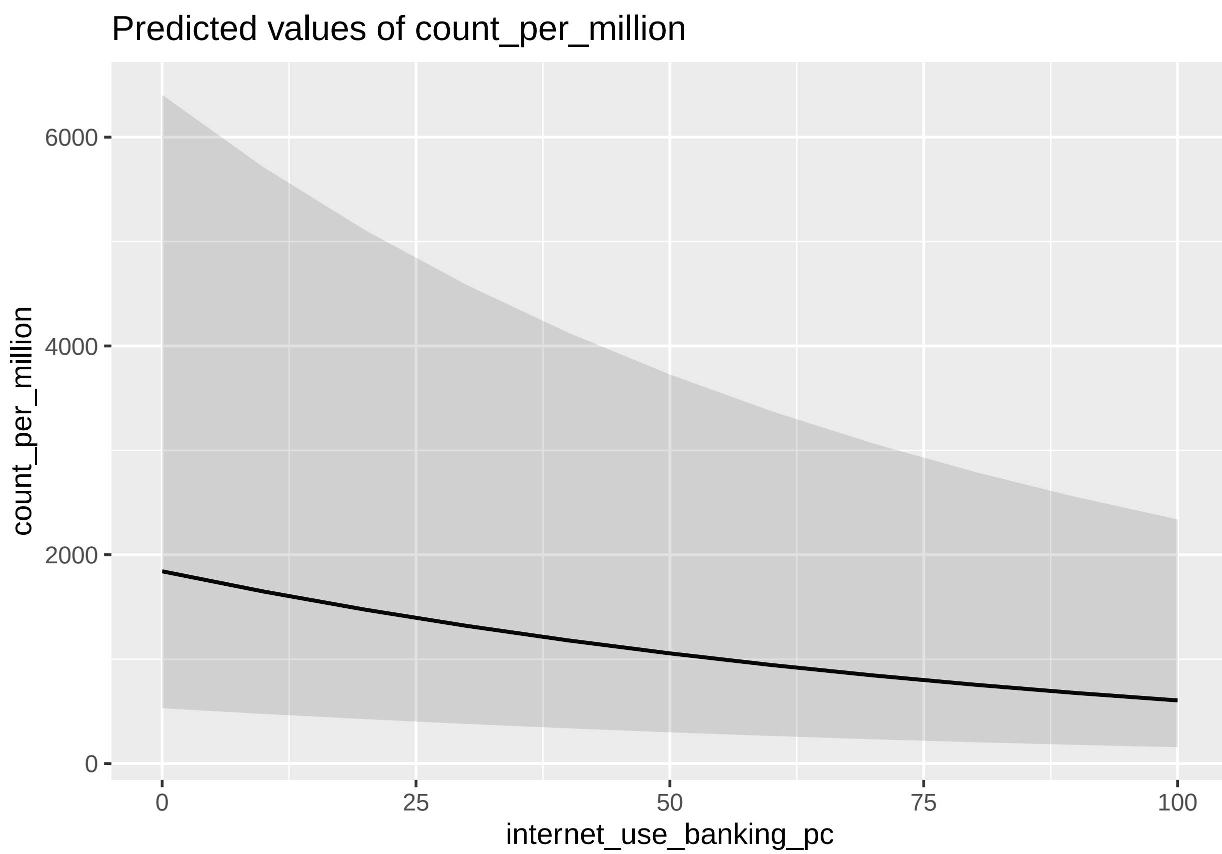
```
## $gdp_pps
```

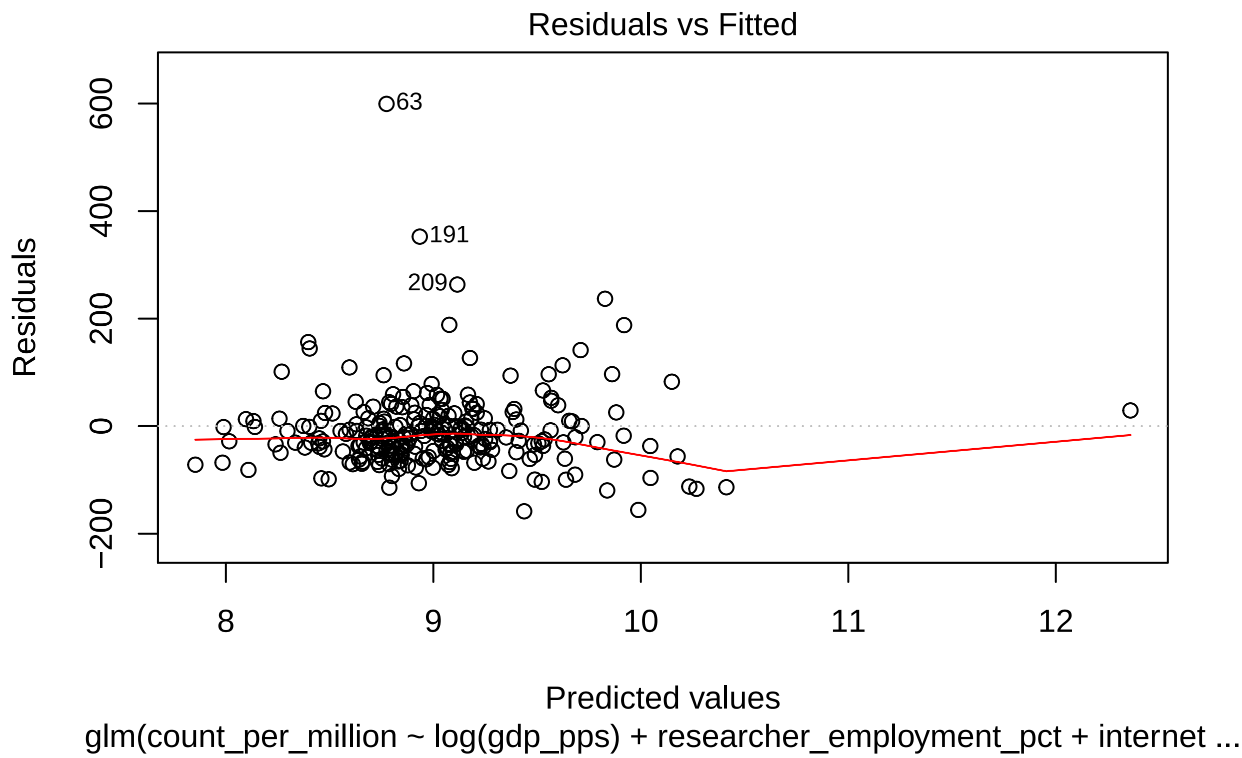



```
##
## $researcher_employment_pct
```

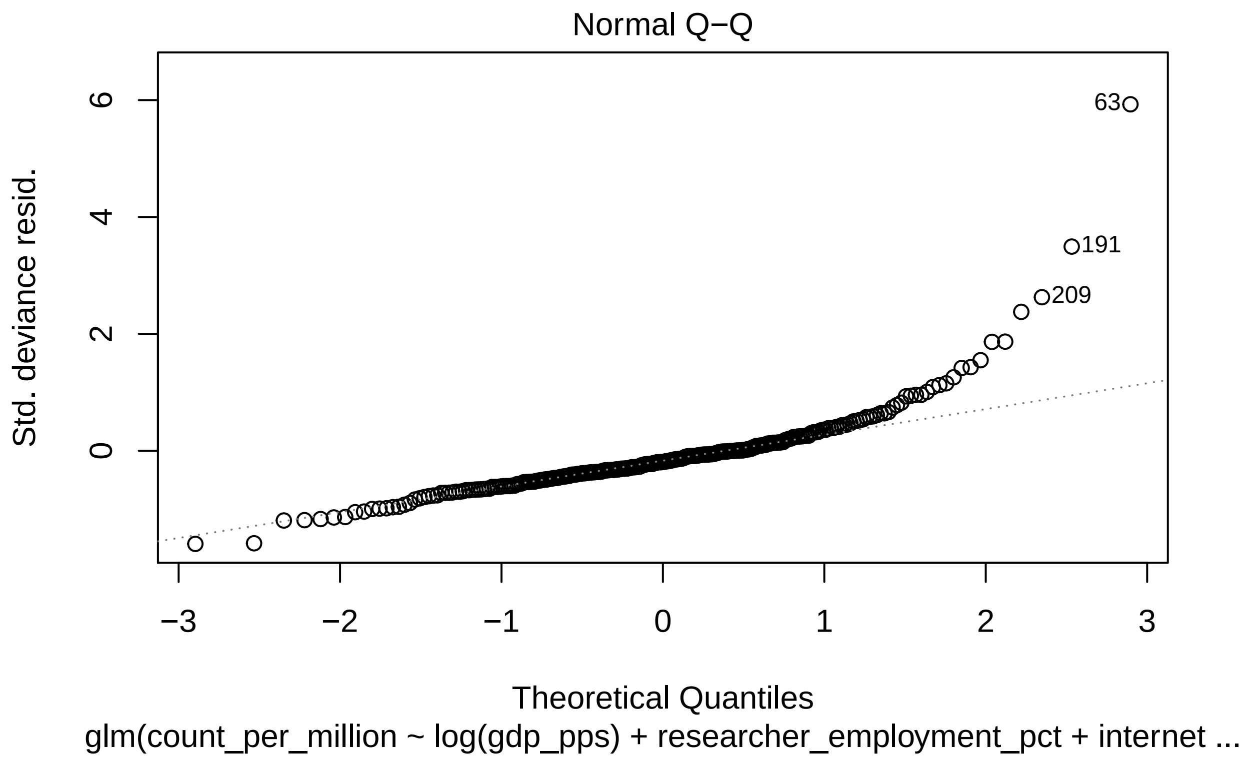


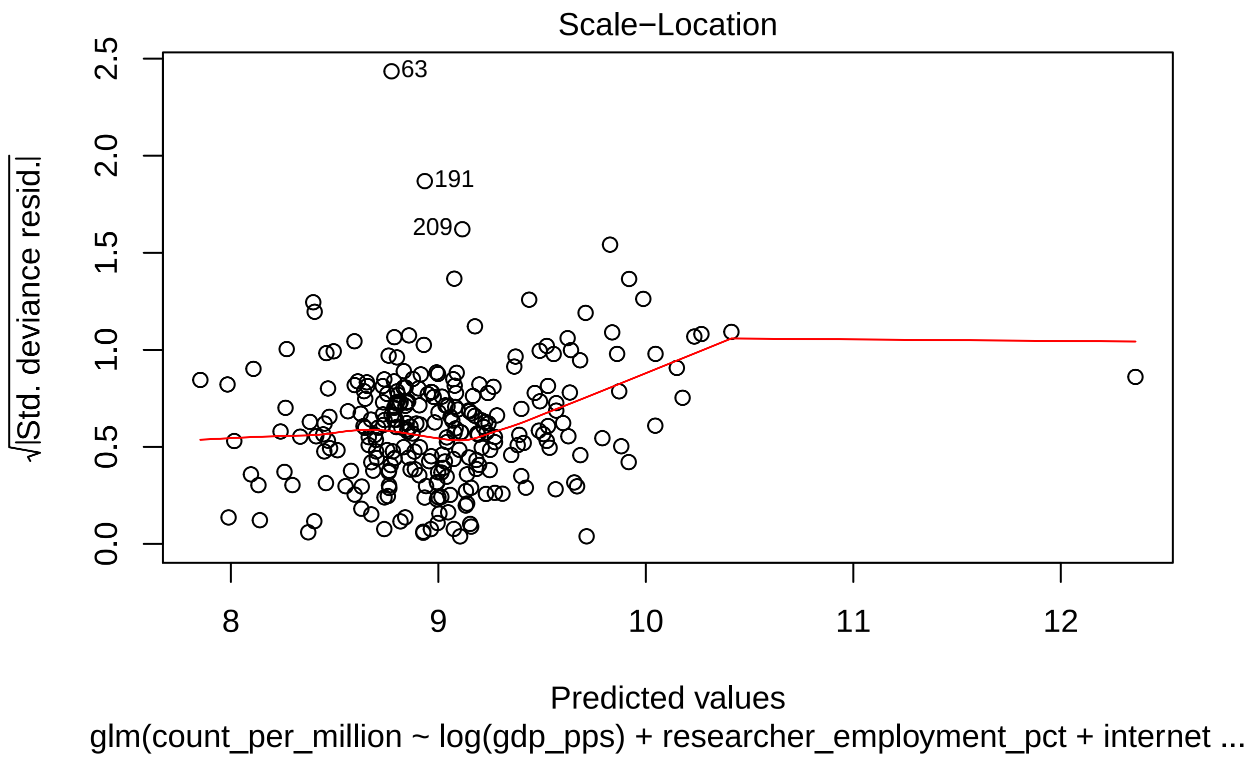

```
##
## $internet_use_banking_pc
```

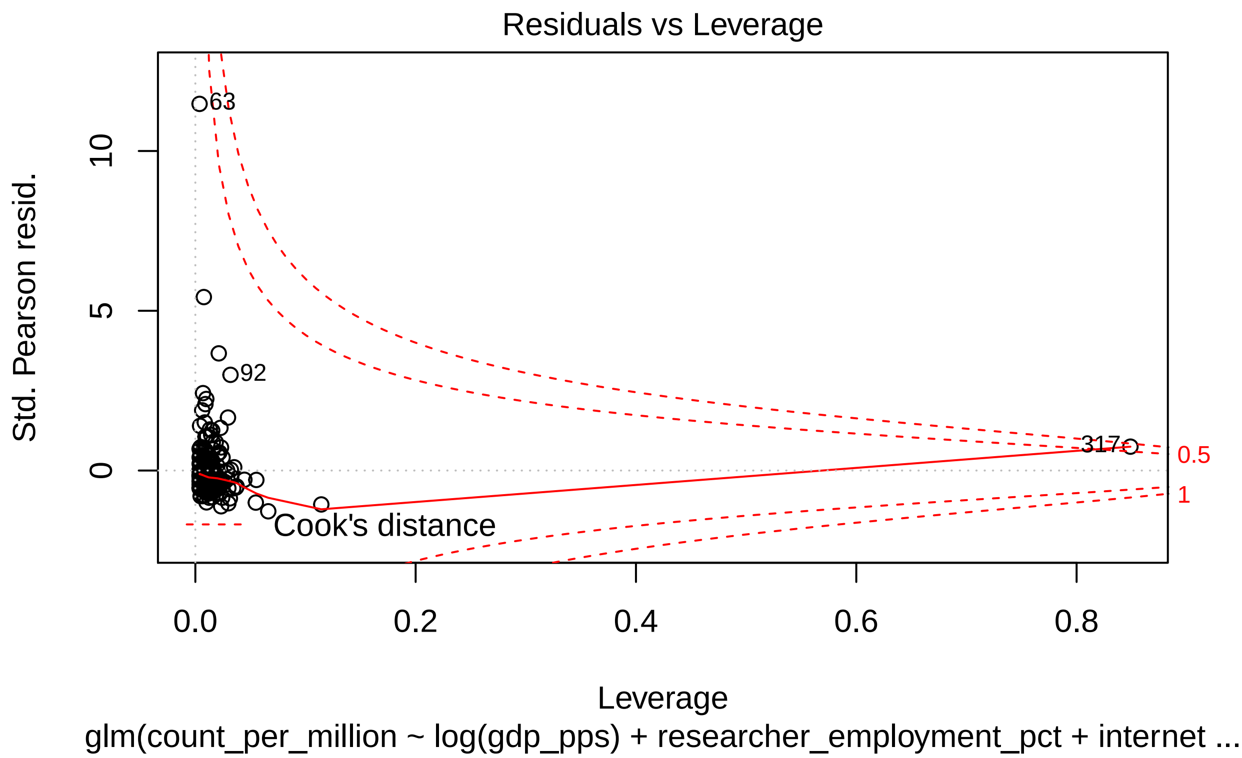

Predicted values of count_per_million
count_per_million
0
2000
4000
6000
0
25
50
75
100
internet_use_banking_pc



Residuals vs Fitted
Residuals
600
400
200
0
−200
63
191
209
8
9
10
11
12
Predicted values
glm(count_per_million ~ log(gdp_pps) + researcher_employment_pct + internet ...


Normal Q−Q
Std. deviance resid.
6
4
2
0
63
191
209
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(count_per_million ~ log(gdp_pps) + researcher_employment_pct + internet ...


Scale−Location
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
2.0
2.5
63
191
209
8
9
10
11
12
Predicted values
glm(count_per_million ~ log(gdp_pps) + researcher_employment_pct + internet ...



per researcher download models

```
#with scitech
perresearcher_plm2 <- glm (count_per_thousand_researchers ~
                             gdp_pps_hab  +
                             disposable_income +
                             edu_attainment_total + gerd +
                             internet_use_banking_pc+hr_scitech_pct,
                           data = eurostat_complete_data,
                           family = poisson )

#without
perresearcher_plm1 <- glm (count_per_thousand_researchers ~
                             log(gdp_pps)  +
                             log(disposable_income) +
                             edu_attainment_total + gerd +
                             internet_use_banking_pc,
                           data = eurostat_complete_data,
                           family = poisson )

#qpoisson model


perresearcher_qplm1 <- glm (count_per_thousand_researchers ~
                              log(gdp_pps)  + log(disposable_income) +
```

```r
                              edu_attainment_total + gerd +
                              internet_use_banking_pc,
                            data = eurostat_complete_data,
                            family = quasipoisson )

perresearcher_qplm2 <- glm (count_per_thousand_researchers ~
                              log(gdp_pps)  + log(disposable_income) +
                              edu_attainment_total + gerd +
                              internet_purchases_last_year_pc,
                            data = eurostat_complete_data,
                            family = quasipoisson )

perresearcher_qplm3 <- glm (count_per_thousand_researchers ~
                              log(gdp_pps)*gerd,
                            data = eurostat_complete_data,
                            family = quasipoisson )

vif(perresearcher_qplm1)
```

```
##            log(gdp_pps)  log(disposable_income)    edu_attainment_total
##                1.346839                2.345741                1.644007
##                    gerd internet_use_banking_pc
##                1.625264                2.379288
```

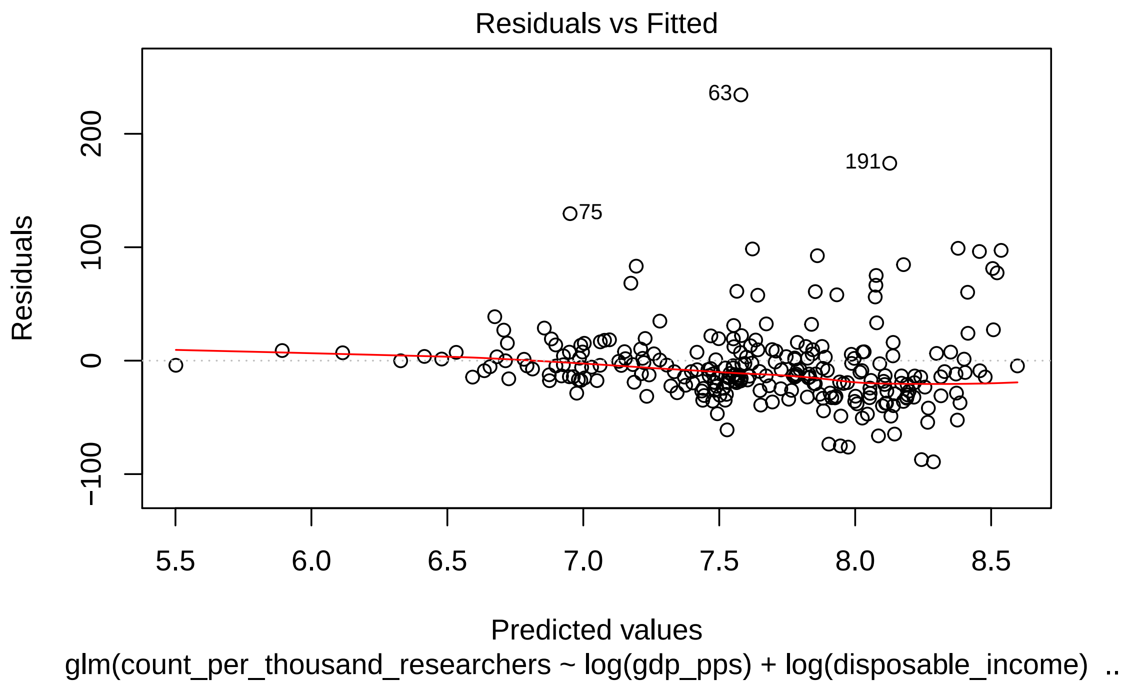
```r
summary(perresearcher_qplm1)
```

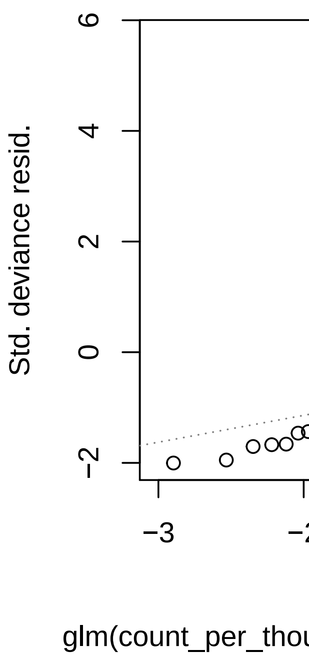
```
##
## Call:
## glm(formula = count_per_thousand_researchers ~ log(gdp_pps) +
##     log(disposable_income) + edu_attainment_total + gerd + internet_use_banking_pc,
##     family = quasipoisson, data = eurostat_complete_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -89.166  -22.578  -10.432    6.938  234.282
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              5.529855   2.239821   2.469  0.01420 *
## log(gdp_pps)             0.160907   0.070548   2.281  0.02337 *
## log(disposable_income)   0.148007   0.254805   0.581  0.56184
## edu_attainment_total     0.007557   0.007681   0.984  0.32613
## gerd                    -0.253286   0.079302  -3.194  0.00158 **
## internet_use_banking_pc -0.018276   0.004194  -4.358  1.9e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 2078.842)
##
##     Null deviance: 495799  on 264  degrees of freedom
## Residual deviance: 362062  on 259  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

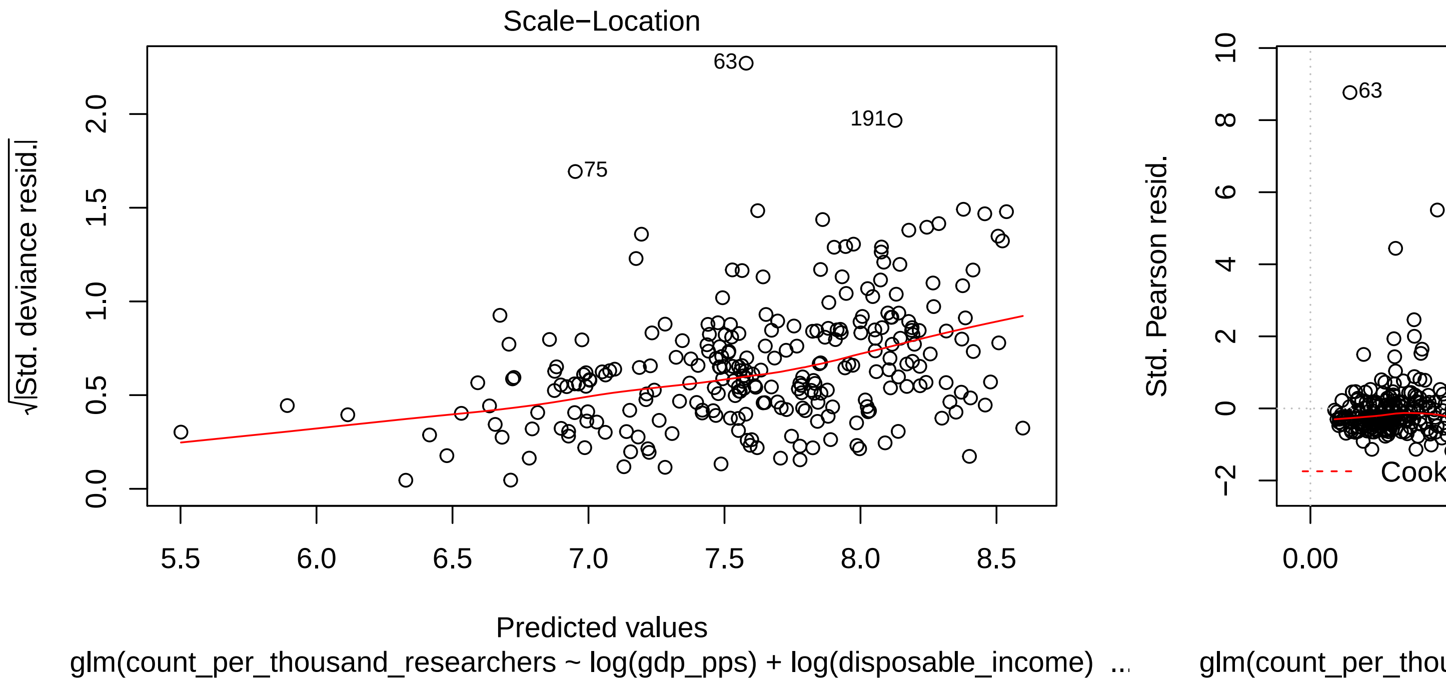
```r
export_summs(perresearcher_qplm1, perresearcher_qplm2,perresearcher_qplm3,
             digits=3,
             statistics = c("null.deviance", "deviance")
             )
```

```
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.

## Warning in if (statistics == "all") {: the condition has length > 1 and only the
## first element will be used

## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
```

```r
plot(perresearcher_qplm1)
```







since the VIF check points to a high multicollinearity with hr_scitech_pct, we remove that variable from the analysis. Since the Poisson models shows high overdispersion, we run a quasipoisson model. there the effects of the GDP and disposable income become non-significant.

These results point to similar conclusions. the the per researcher download volume:

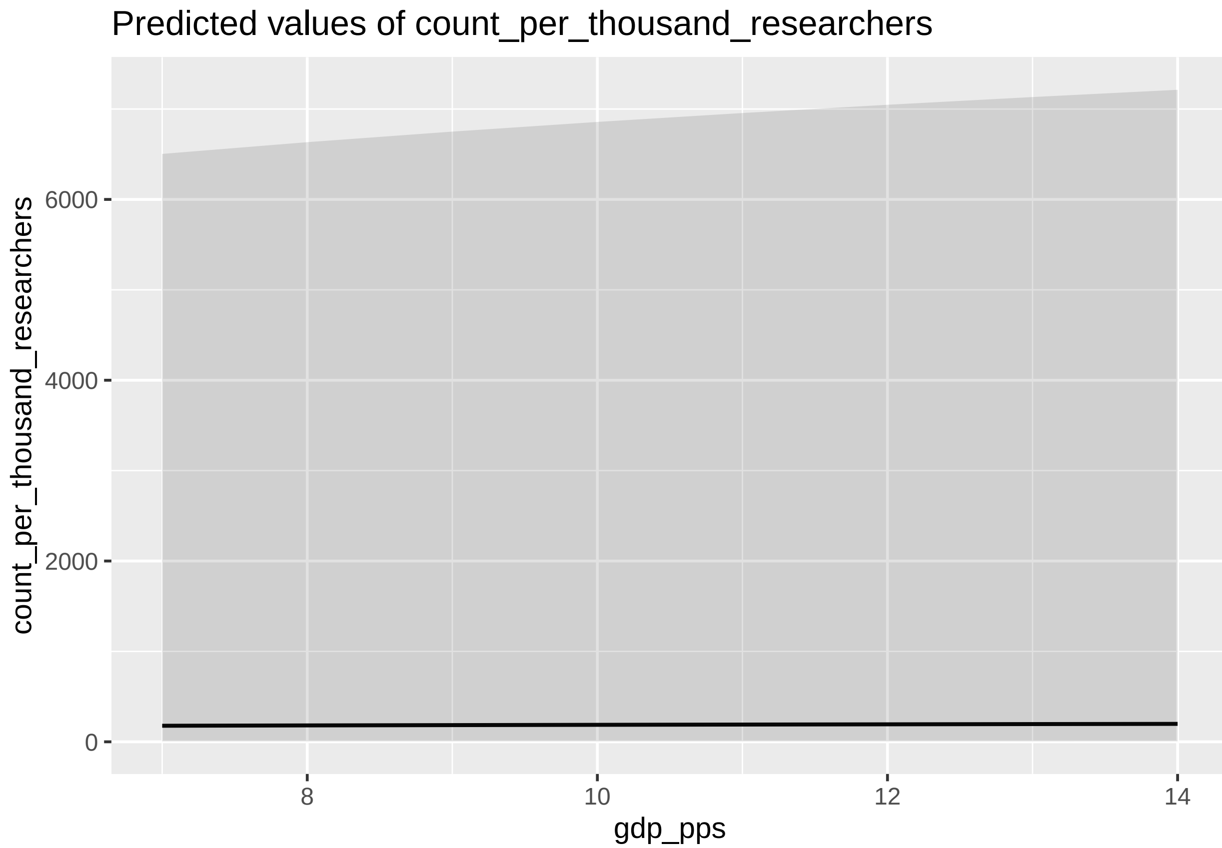
- still grows with wealth, but
- is moderated by the R&D expenditure, and internet proficiency.

Regions with more internet proficiency populations, and with higher R&D spending download less per researcher.

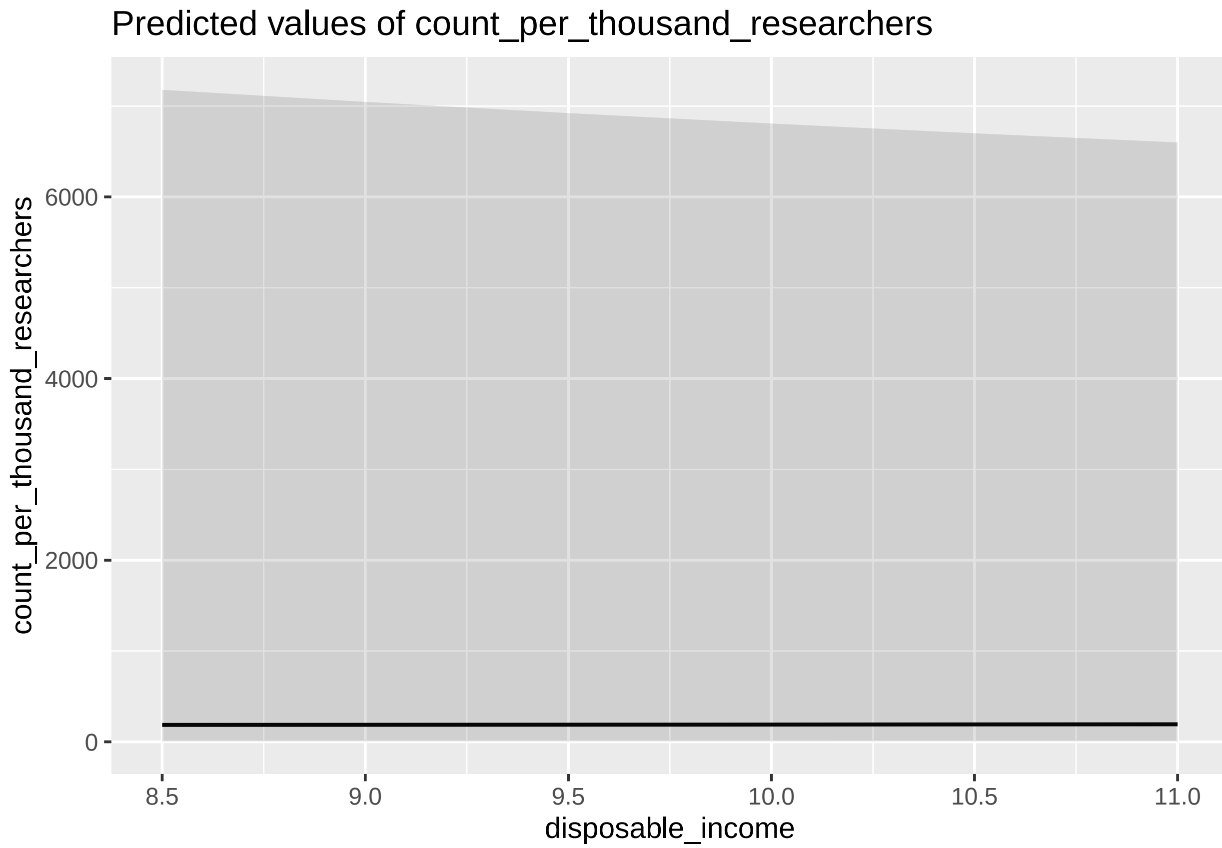
```r
plot_model(perresearcher_qplm1, type = "pred")
```

```
## Model has log-transformed predictors. Consider using `terms="gdp_pps [exp]"` to back-transform scale
## Model has log-transformed predictors. Consider using `terms="gdp_pps [exp]"` to back-transform scale
## Model has log-transformed predictors. Consider using `terms="gdp_pps [exp]"` to back-transform scale
## Model has log-transformed predictors. Consider using `terms="gdp_pps [exp]"` to back-transform scale
## Model has log-transformed predictors. Consider using `terms="gdp_pps [exp]"` to back-transform scale
```

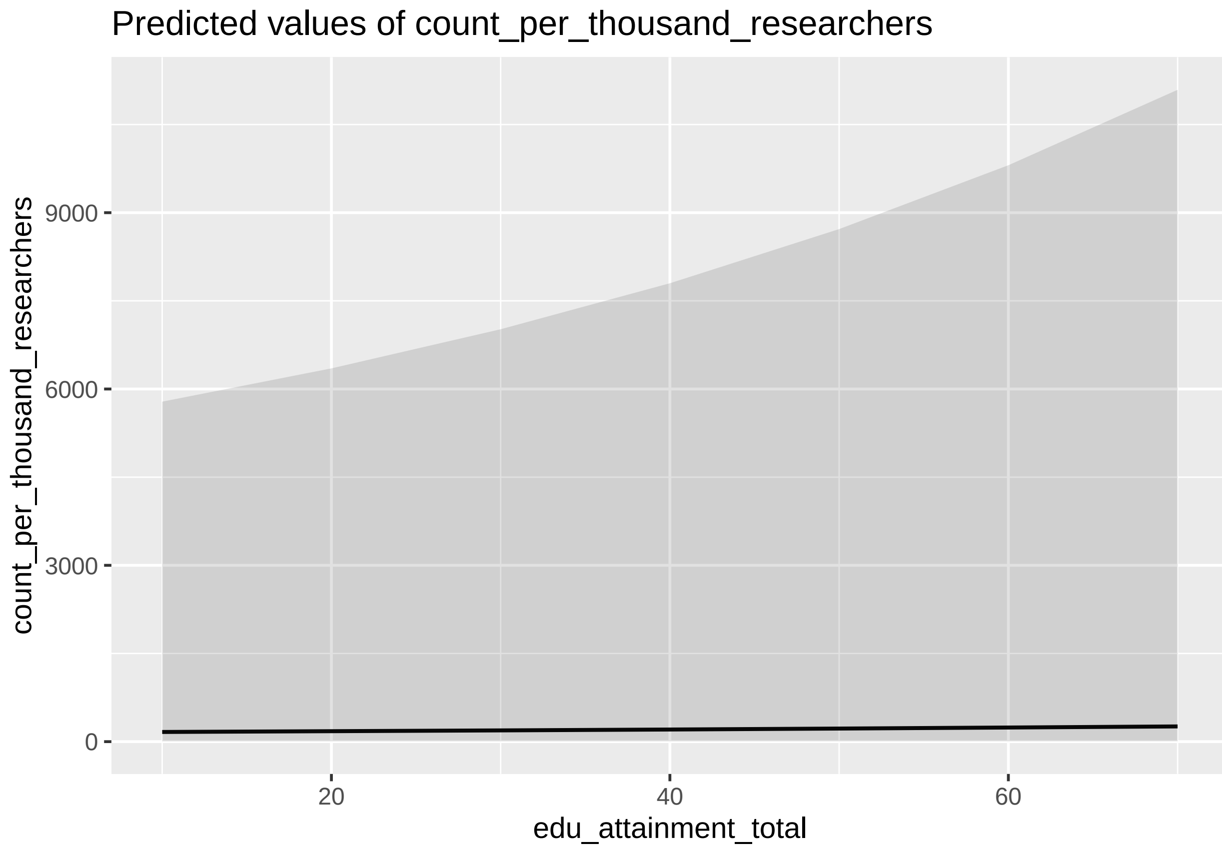
```
## $gdp_pps
```



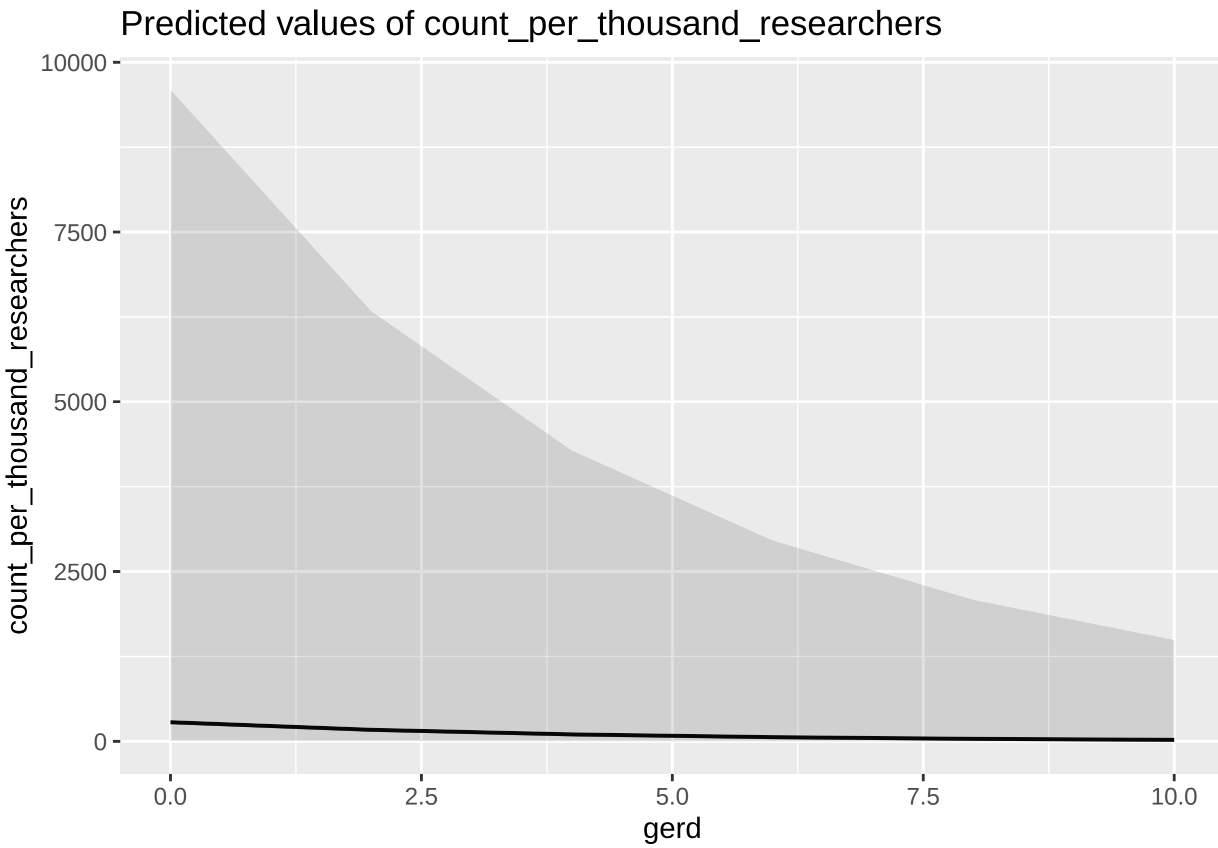

```
##
## $disposable_income
```



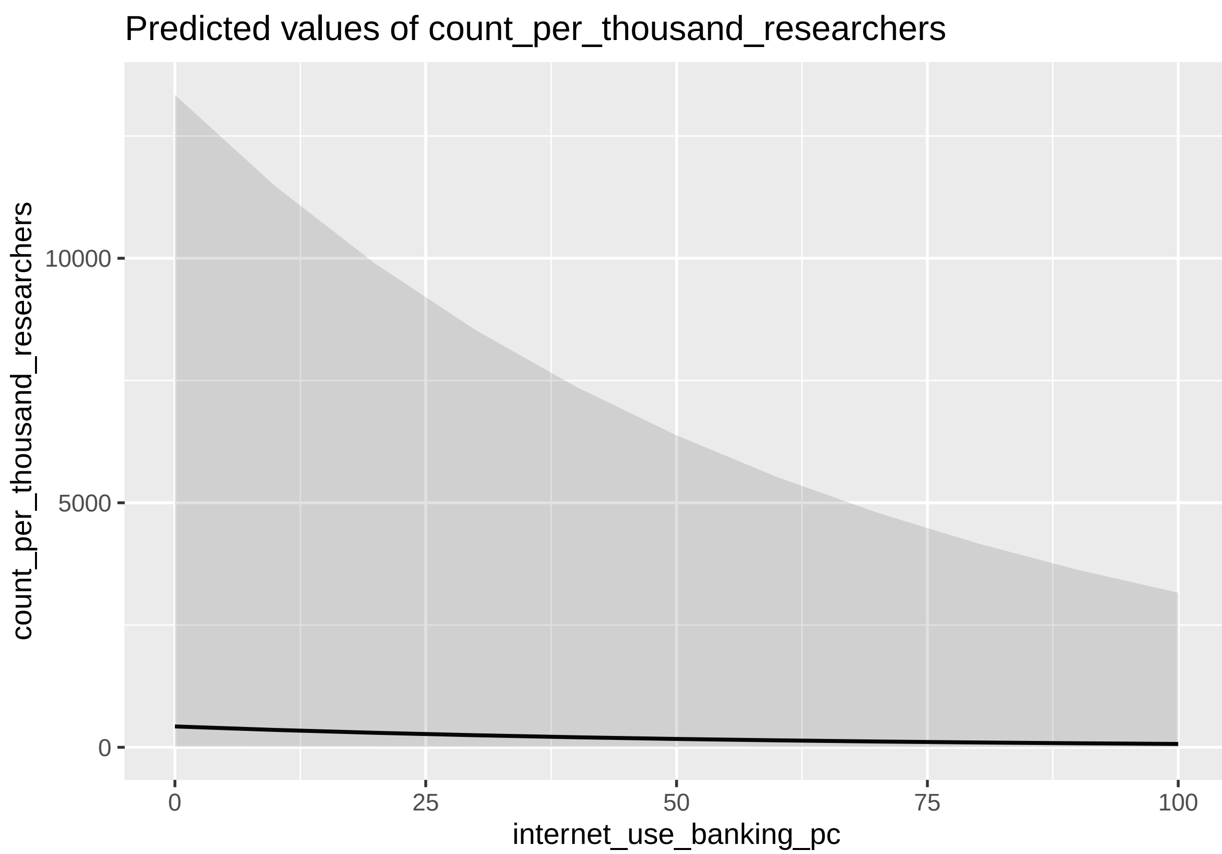

```
##
## $edu_attainment_total
```


Predicted values of count_per_thousand_researchers
count_per_thousand_researchers
0
3000
6000
9000
20
40
60
edu_attainment_total

```
##
## $gerd
```




```
##
## $internet_use_banking_pc
```



Count Regression Models

```
#count without R&D
count_plm2 <- glm (count ~ log(gdp_pps) +  researcher_employment_pct +
                 disposable_income +
                 edu_attainment_total +
                 internet_use_banking_pc,
                 data = eurostat_complete_data, poisson )

#with R&D
count_plm3 <- glm (count ~ log(gdp_pps) +  researcher_employment_pct +
                 disposable_income +
                 edu_attainment_total +
                 internet_use_banking_pc +
                  gerd
                , data = eurostat_complete_data, poisson )

count_plm4 <- glm (count ~ log(gdp_pps) +
                    researcher_employment_pct +
                    disposable_income +
                    edu_attainment_total +
                    internet_use_banking_pc +
                    gerd+
                    internet_purchases_last_year_pc
                    , data = eurostat_complete_data, poisson )
```

```r
vif(count_plm4)
```

```
##                 log(gdp_pps)       researcher_employment_pct
##                     1.479542                        7.264413
##            disposable_income             edu_attainment_total
##                     5.850430                        4.948335
##       internet_use_banking_pc                             gerd
##                     4.042030                        1.767706
## internet_purchases_last_year_pc
##                     6.105920
```

```r
summary (count_plm4)
```

```
##
## Call:
## glm(formula = count ~ log(gdp_pps) + researcher_employment_pct +
##     disposable_income + edu_attainment_total + internet_use_banking_pc +
##     gerd + internet_purchases_last_year_pc, family = poisson,
##     data = eurostat_complete_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -359.24   -50.92   -24.84    18.49   600.08
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -3.669e-01  5.997e-03  -61.18   <2e-16 ***
## log(gdp_pps)                    9.834e-01  5.882e-04 1671.93   <2e-16 ***
## researcher_employment_pct       4.190e-01  1.056e-03  396.82   <2e-16 ***
## disposable_income              -2.145e-05  1.545e-07 -138.82   <2e-16 ***
## edu_attainment_total            1.004e-02  7.646e-05  131.25   <2e-16 ***
## internet_use_banking_pc        -2.297e-02  4.640e-05 -495.06   <2e-16 ***
## gerd                           -4.835e-02  4.640e-04 -104.20   <2e-16 ***
## internet_purchases_last_year_pc  6.351e-03  4.688e-05  135.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 7192467  on 264  degrees of freedom
## Residual deviance: 1966657  on 257  degrees of freedom
## AIC: 1969514
##
## Number of Fisher Scoring iterations: 5
```

```r
summary (count_plm3)
```

```
##
## Call:
## glm(formula = count ~ log(gdp_pps) + researcher_employment_pct +
##     disposable_income + edu_attainment_total + internet_use_banking_pc +
##     gerd, family = poisson, data = eurostat_complete_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -361.15   -49.67   -23.02    18.05   610.94
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -2.742e-01  5.927e-03  -46.27   <2e-16 ***
## log(gdp_pps)               9.632e-01  5.643e-04 1706.91   <2e-16 ***
## researcher_employment_pct  3.657e-01  9.788e-04  373.67   <2e-16 ***
## disposable_income         -1.130e-05  1.352e-07  -83.56   <2e-16 ***
## edu_attainment_total       1.361e-02  7.089e-05  192.04   <2e-16 ***
## internet_use_banking_pc   -1.872e-02  3.352e-05 -558.44   <2e-16 ***
## gerd                      -2.595e-02  4.324e-04  -60.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 7192467  on 264  degrees of freedom
## Residual deviance: 1985062  on 258  degrees of freedom
## AIC: 1987918
##
## Number of Fisher Scoring iterations: 5
```

```r
#qpoisson
count_qplm2 <- glm (count ~ log(gdp_pps) +  researcher_employment_pct +
                 log(disposable_income) +
                 edu_attainment_total +
                 internet_use_banking_pc,
                 data = eurostat_complete_data, quasipoisson)

count_qplm3<- glm (count ~ log(gdp_pps) +  researcher_employment_pct +
                 log(disposable_income) +
                 edu_attainment_total +
                 internet_use_banking_pc +
                gerd,
                data = eurostat_complete_data, quasipoisson )

#is there multicollinearity? yes.
vif(count_qplm3)
```

```
##              log(gdp_pps) researcher_employment_pct    log(disposable_income)
##                  1.452575                  4.324325                  3.176083
##      edu_attainment_total   internet_use_banking_pc                      gerd
##                  4.304273                  2.274701                  1.479112
```

```r
count_qplm3a<- glm (count ~ log(gdp_pps) +
                 log(disposable_income) +
                 edu_attainment_total +
                 internet_use_banking_pc +
                gerd,
                data = eurostat_complete_data, quasipoisson )

count_qplm3b<- glm (count ~ log(gdp_pps) +
                 log(disposable_income) +
                 internet_use_banking_pc +
                gerd,
```

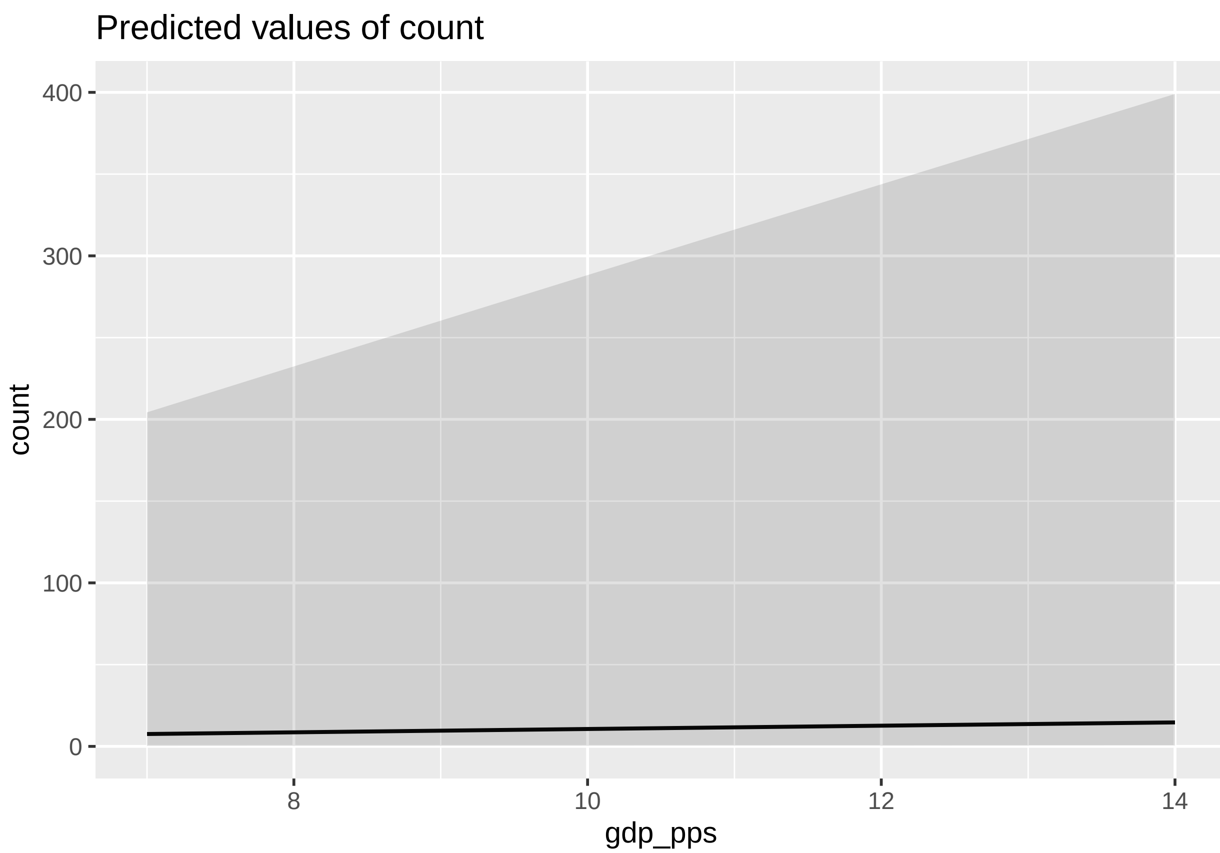
```r
                data = eurostat_complete_data, quasipoisson )

# is gerd sinificant if we remove researcher pct? no. education attainment?
summary(count_qplm3a)
```

```
##
## Call:
## glm(formula = count ~ log(gdp_pps) + log(disposable_income) +
##     edu_attainment_total + internet_use_banking_pc + gerd, family = quasipoisson,
##     data = eurostat_complete_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -355.13   -54.44   -22.30    17.23   575.64
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -3.098086   1.974263  -1.569    0.118
## log(gdp_pps)             0.963964   0.063842  15.099  < 2e-16 ***
## log(disposable_income)   0.261247   0.235443   1.110    0.268
## edu_attainment_total     0.032279   0.005526   5.841 1.55e-08 ***
## internet_use_banking_pc -0.023900   0.003638  -6.569 2.77e-10 ***
## gerd                     0.028054   0.044957   0.624    0.533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 12110.63)
##
##     Null deviance: 7192467  on 264  degrees of freedom
## Residual deviance: 2155840  on 259  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

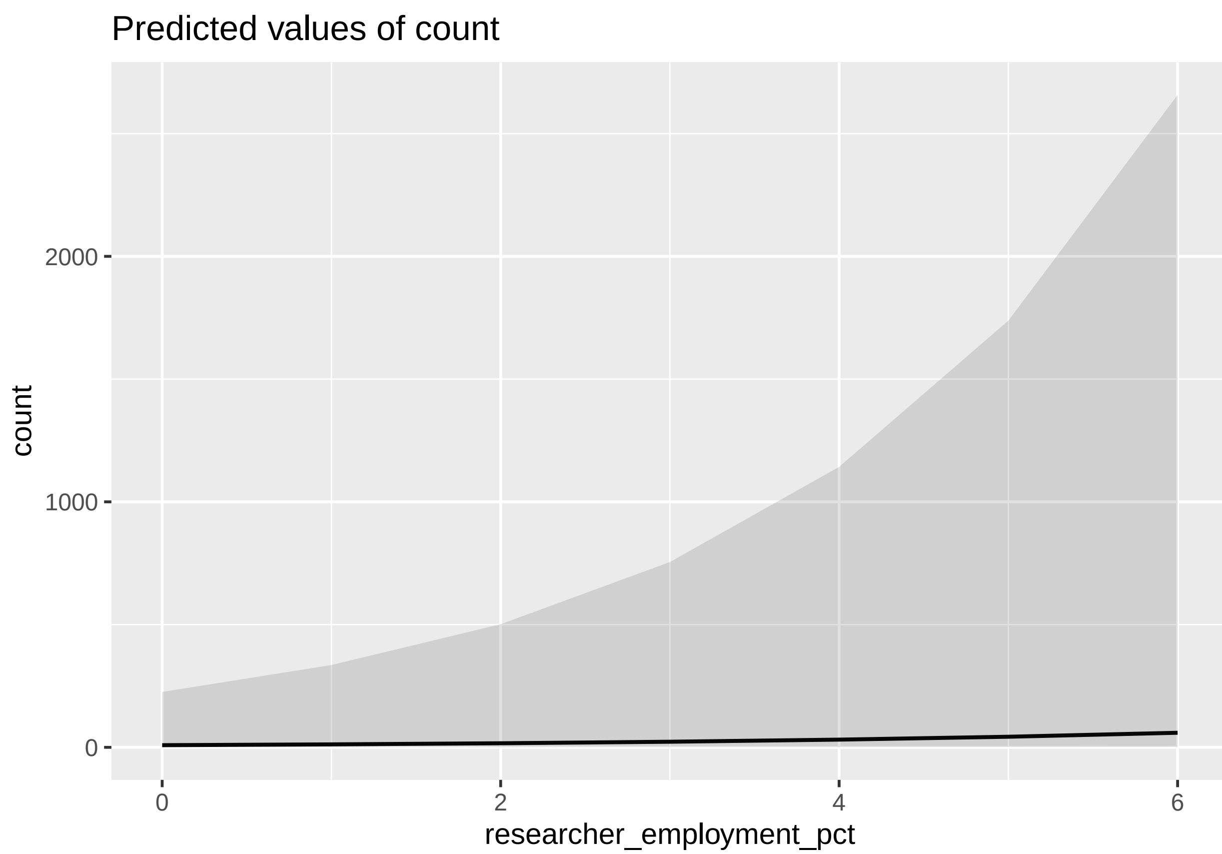
```r
count_qplm4 <- glm (count ~ log(gdp_pps) +
                    researcher_employment_pct +
                    log(disposable_income) +
                    edu_attainment_total +
                    gerd +
                    internet_purchases_last_year_pc,
                    data = eurostat_complete_data, quasipoisson )

export_summs(count_qplm2,count_qplm3,count_qplm4, digits=10, statistics = "all")
```

```
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
```

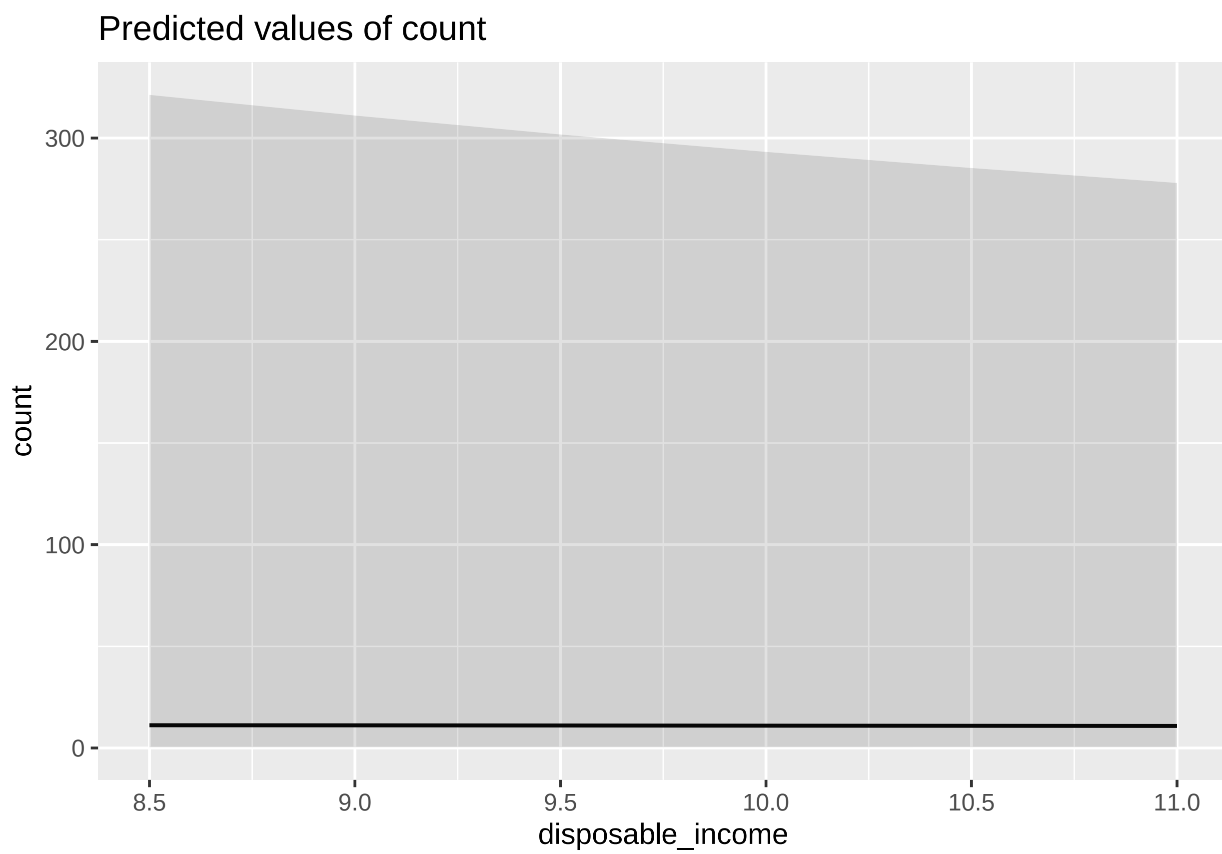
```
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.

## Model has log-transformed predictors. Consider using `terms="gdp_pps [exp]"` to back-transform scale
## Model has log-transformed predictors. Consider using `terms="gdp_pps [exp]"` to back-transform scale
## Model has log-transformed predictors. Consider using `terms="gdp_pps [exp]"` to back-transform scale
## Model has log-transformed predictors. Consider using `terms="gdp_pps [exp]"` to back-transform scale
## Model has log-transformed predictors. Consider using `terms="gdp_pps [exp]"` to back-transform scale

## $gdp_pps
```



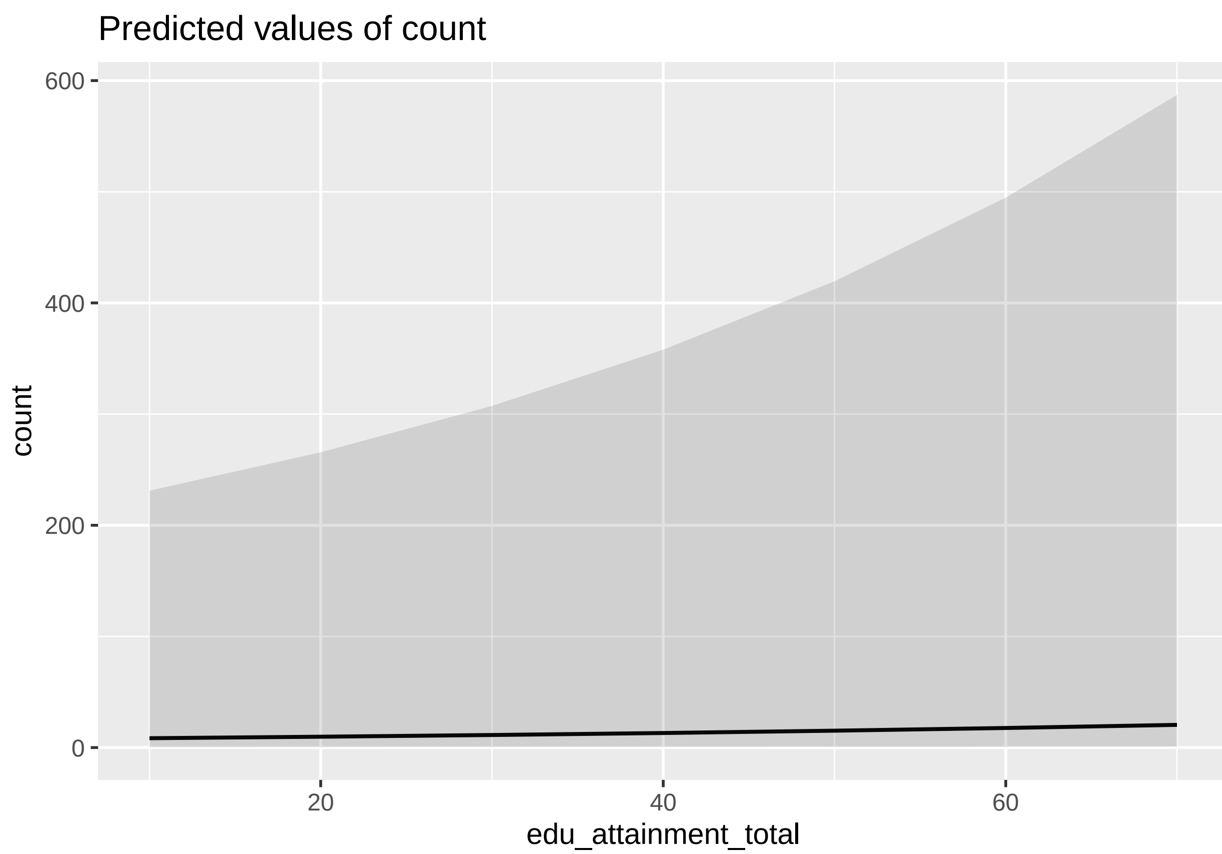

```
##
## $researcher_employment_pct
```



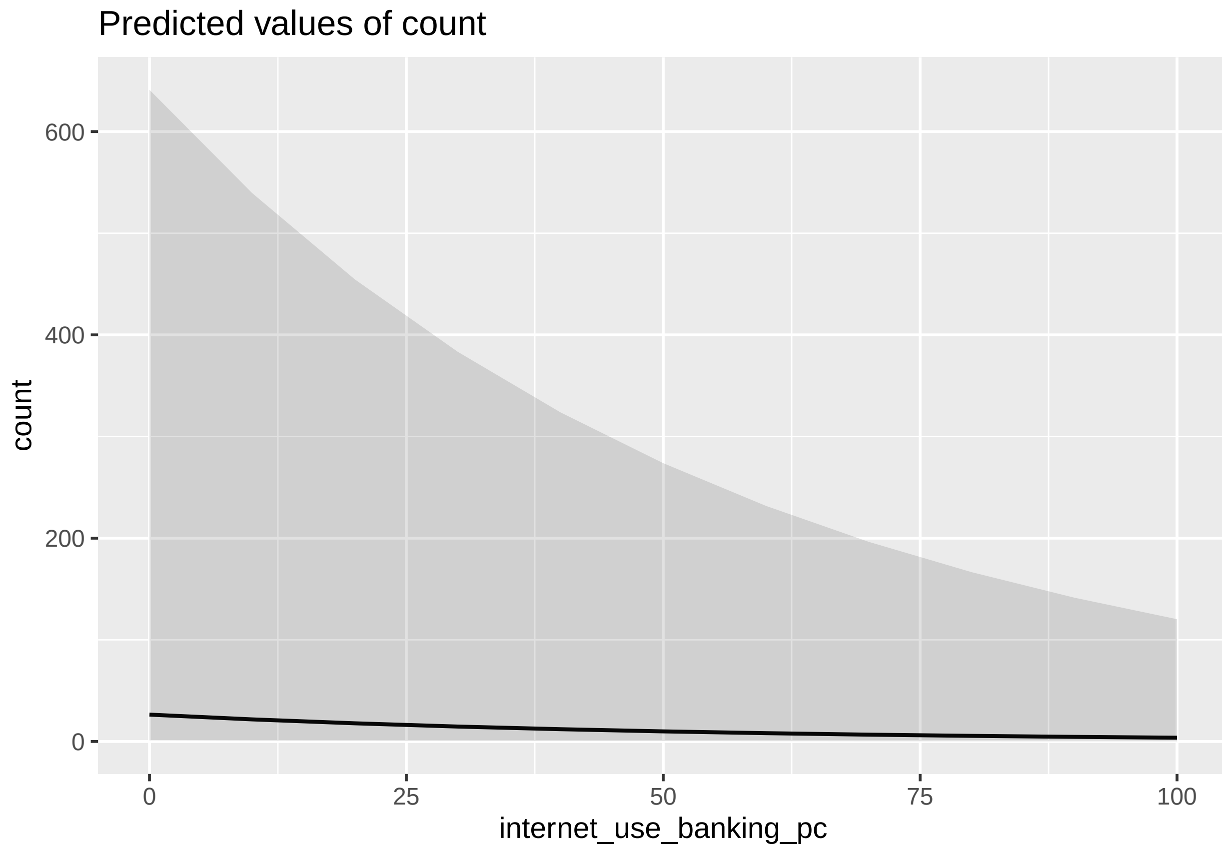

```
##
## $disposable_income
```




```
##
## $edu_attainment_total
```




```
##
## $internet_use_banking_pc
```



in the total count model multicollinearity forces us to use GDP_PPS and drop the sci sci-tech variables. Adding R&D or online purchases does not make the model much better and it is not significant in the quasipoisson models.

Comparing the Three Models

If we compare the three models (per capita, per researcher, and absolute download counts, all modeled as quasipoisson), we find that consistent results. wealth, researcher employment has a positive effect, internet proficiency has a negative effect. In the model per researcher model we see that higher R&D expenditure can lower download counts.

```
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
```

To check the robustness of these models, we also did simple linear regression for all 3 dependent variables. They do not yield different results from the quasipoisson models, but their error terms are much uglier.

Regions with higher gdp and higher researcher share in the workforce download more, while higher online proficiency lowers download numbers, probably due to the positive effects of e-commerce, and the negative effects of better hiding.

interaction models

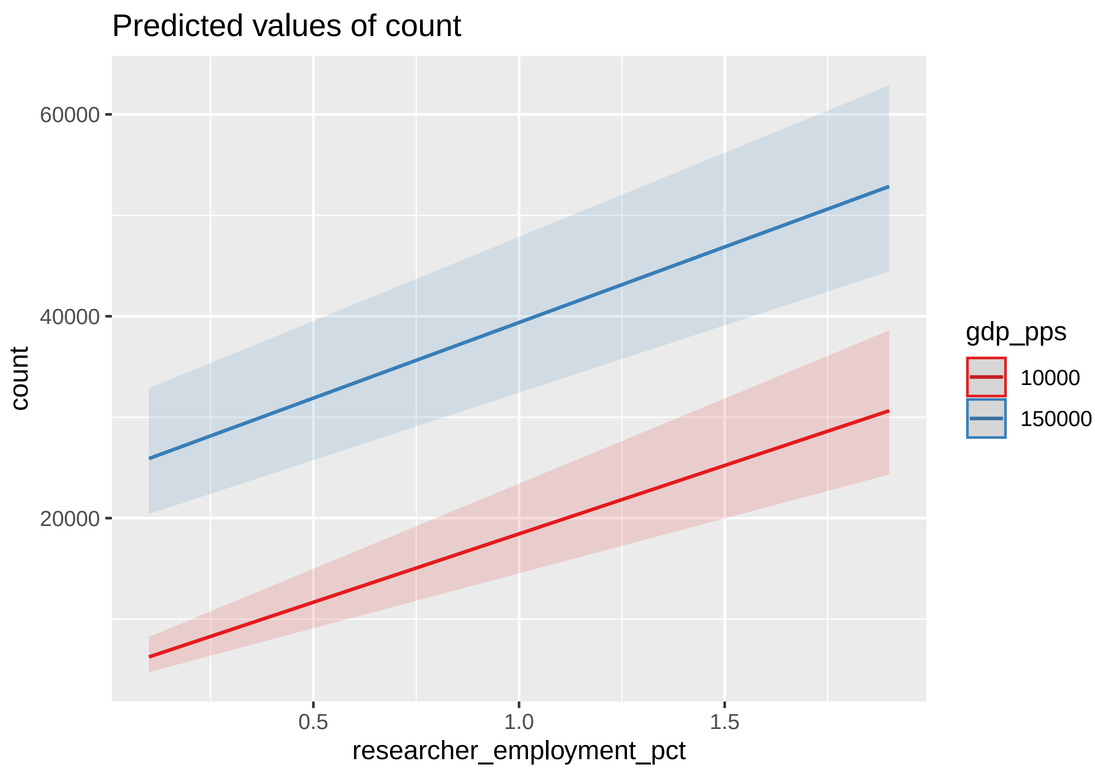
Finally, we check the interaction of wealth (GDP_PPS) and researcher employment with a simple interaction model.

simple count variable, GDP and researcher share

```
interaction_qplm1<- glm (count  ~ gdp_pps * researcher_employment_pct,
                         data = eurostat_complete_data,
                         family = quasipoisson )

summary (interaction_qplm1)
```

```
##
## Call:
## glm(formula = count ~ gdp_pps * researcher_employment_pct, family = quasipoisson,
##     data = eurostat_complete_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -345.53   -84.90   -37.05    17.52   765.65
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       8.546e+00  1.588e-01  53.830  < 2e-16 ***
## gdp_pps                           1.051e-05  1.279e-06   8.216 9.87e-15 ***
## researcher_employment_pct         9.183e-01  1.159e-01   7.926 6.60e-14 ***
## gdp_pps:researcher_employment_pct -3.479e-06  7.195e-07  -4.836 2.27e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 21449.96)
##
##     Null deviance: 7192467  on 264  degrees of freedom
## Residual deviance: 3556374  on 261  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
plot_model(interaction_qplm1,
           type = "eff", terms=c("researcher_employment_pct[0.1,1.9]",
                                 "gdp_pps [10000,150000]"))
```



A simple interaction at the count model shows that in richer regions download more even if they have a the same share of researchers as poorer regions, and the count grows faster as the share grows. This confirms our original hypothesis.

```
#per capita model does not yield meaningful interaction. only researcher share significant
interaction_qplm2 <- glm (count_per_million  ~
                           gdp_pps * researcher_employment_pct,
                         data = eurostat_complete_data,
                         family = quasipoisson )
summary (interaction_qplm2)
```

```
##
## Call:
## glm(formula = count_per_million ~ gdp_pps * researcher_employment_pct,
##     family = quasipoisson, data = eurostat_complete_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -201.32   -54.51   -18.54    18.18   596.39
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      8.540e+00  1.567e-01  54.492  < 2e-16 ***
## gdp_pps                          5.049e-07  1.654e-06   0.305    0.760
## researcher_employment_pct        6.444e-01  1.276e-01   5.052 8.23e-07 ***
## gdp_pps:researcher_employment_pct 3.566e-07  8.692e-07   0.410    0.682
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 11203.89)
```

```
##
##     Null deviance: 2990524  on 264  degrees of freedom
## Residual deviance: 1602050  on 261  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```r
#per researcher models yield little.  gdp is not relevant
interaction_qplm3 <- glm (count_per_thousand_researchers ~
                            gdp_pps * internet_use_banking_pc,
                          data = eurostat_complete_data,
                          family = quasipoisson )

summary (interaction_qplm3)
```

```
##
## Call:
## glm(formula = count_per_thousand_researchers ~ gdp_pps * internet_use_banking_pc,
##     family = quasipoisson, data = eurostat_complete_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -84.632  -23.533  -11.553    5.234  228.824
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     8.512e+00  1.452e-01  58.603  < 2e-16 ***
## gdp_pps                        -7.355e-07  2.909e-06  -0.253    0.801
## internet_use_banking_pc        -1.921e-02  3.439e-03  -5.586 5.83e-08 ***
## gdp_pps:internet_use_banking_pc  1.898e-08  5.846e-08   0.325    0.746
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 2177.066)
##
##     Null deviance: 495799  on 264  degrees of freedom
## Residual deviance: 392699  on 261  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

Other interaction models with per capita and per researcher dependent variables, and wealth and researcher employment and online proficiency do not yield significant results.

Inductive models

We took a first look at this data with two methods. First, we created all possible linear regression equations and two-variable multiple linear regression between the count data and the socio-economic variables. We also used the random forest algorithm to rank the importance of socio-economic environmental variables in explaining the difference in the level of book piracy. The logic of the two approaches is similar. We use a well-defined searching algorithm to find a relationship between the levels of socio-economic environmental variables and download count numbers. We did the inductive approach twice: first on the larger, but narrower dataset used so far, and second for a smaller dataset which includes new variables from the EUROBAROMETER dataset.

Linear regression models

To understand the interaction of environmental variables and count data, we created all possible linear regressions 'explaining' the variability of count per capita data in the following steps:

- We created the initial linear regression
- We checked for outliers, and removed them
- Re-run the regression model, and selected those whose coefficients were significant on 1.96 level
- We ordered the remaining 38 models by adjusted R squares.

The following table shows the results of this approach for the smaller, dataset with additional Eurobarometer variables. We get very results consistent with these, if we run the analysis on the larger dataset.

This approach shows results which are consistent with the deductive models. Wealth, the percentage of knowledge workers in the workforce has positive effects, online proficiency (in all forms) has a negative effect. What is new are the emergence of tow of the new EUROBAROMETER variables: the share of population who visited a public library at least once in the last 12 months (with a negative sign), and the percentage of students in he population (with a positive one). It is equally informative that neither the open science attitudinal variable (weighted sum of yes answer options to the QD 17 Do you think that the results of publicly funded research should be made available online free of charge?), nor the library inadequacy variable (a weighted sum of the responses that chose from the question block QB2 why you haven't Visited a public library or haven't done it more often in the last 12 months? … answered with the option: Limited or poor quality of this activity in the place where you live.) emerged as significant variables. We should note, that in this latter case, the number of respondents is rather low and this is not a very reliable statistic on regional level.

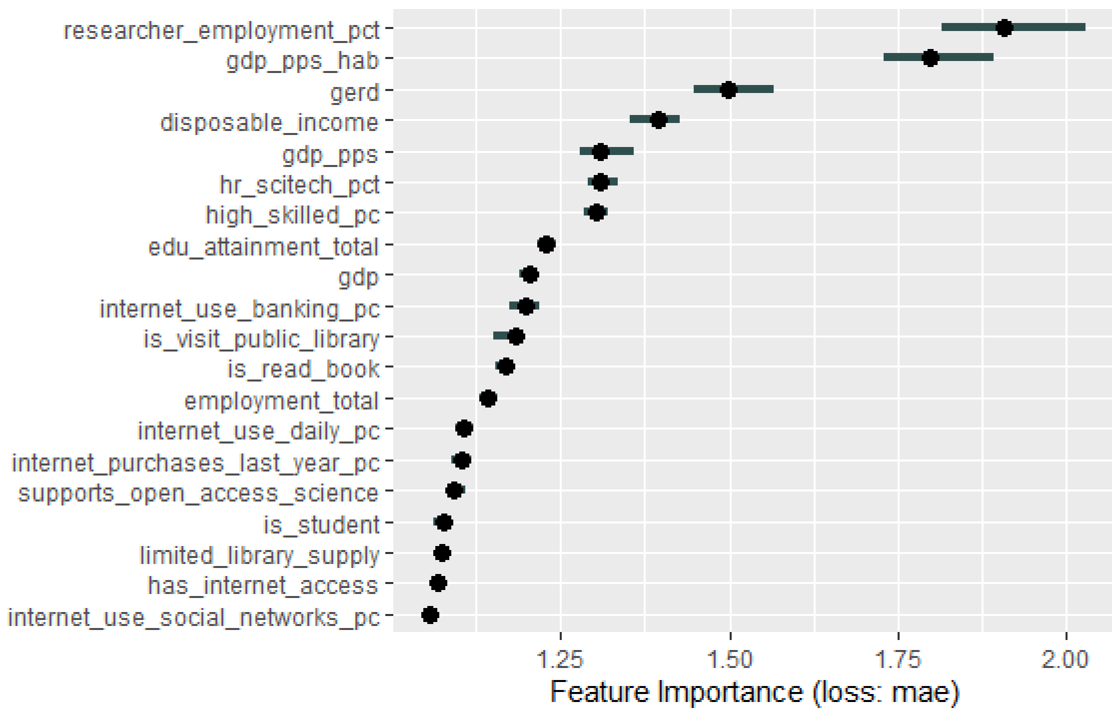
Random forest models

In the following we created three sets of analysis: one that uses count per researcher as the dependent variable, One that uses download per capita as download variable, and one that uses raw count as a dependent variable. The reason for that is that if i use dl/researcher variable as dependent, I may not be able to capture and explain the downloads that possibly come from non-researchers, while if I use dl/capita, then i may find independent variables that account for the professional (researcher downloads) and others that are more characteristic for no-professionals. In the next step we scaled the variables to unit variance, so that they have equal weight in the variable selection process.

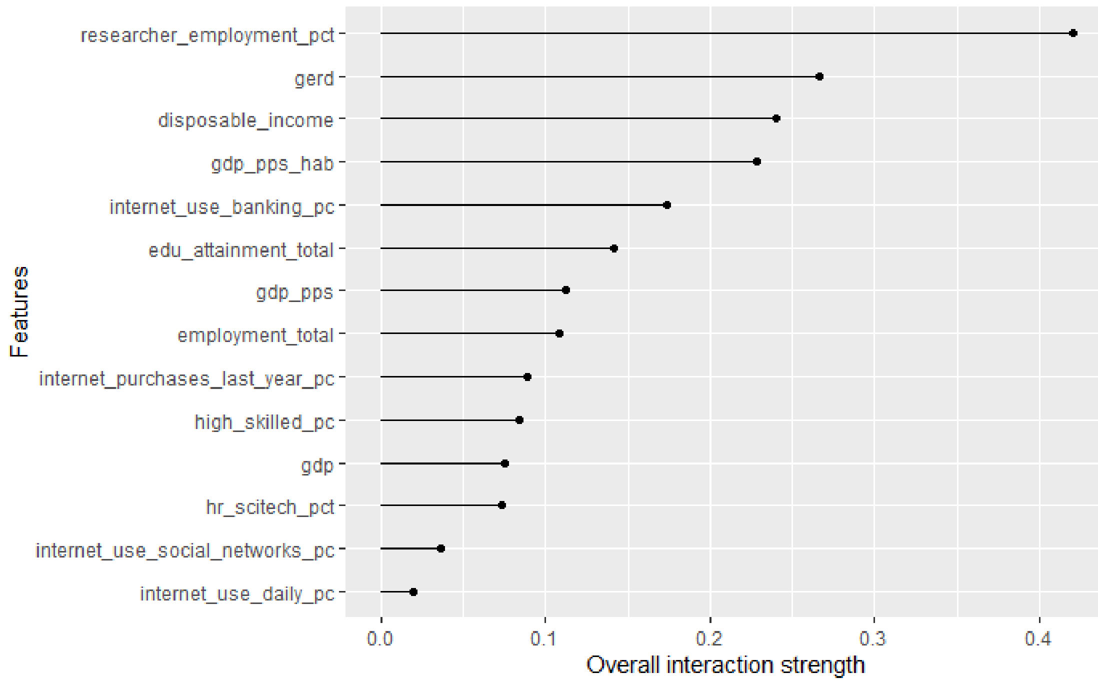
count per capita -without eurobarometer

Educational attainment, disposable income may be relevant for the whole population, beyond just researchers.

```
#count per capita
cpc_predictor_var_select = Predictor$new(
  cpc_var_select.rf,
  data = dplyr::select ( cpc_var_select_df , -geo, - count_per_capita),
  y = as.numeric(cpc_var_select_df$count_per_capita))

cpc_imp <- FeatureImp$new(cpc_predictor_var_select, loss = "mae", n.repetitions = 10)
plot(cpc_imp)
```




```r
#This code takes several minutes to run if uncommented.
run_in_function <- function() {
  interact <- Interaction$new(predictor_var_select)
  plot(interact) #causes knitr issues, something is wrong with the plot, maybe the size, I saved it and

}
#run_in_function()
#I saved the result here:
knitr::include_graphics('images_graphs/percapita_interactions_0416.png')
```



count per capita - with eurobarometer

Educational attainment, disposable income may be relevant for the whole population, beyond just researchers.

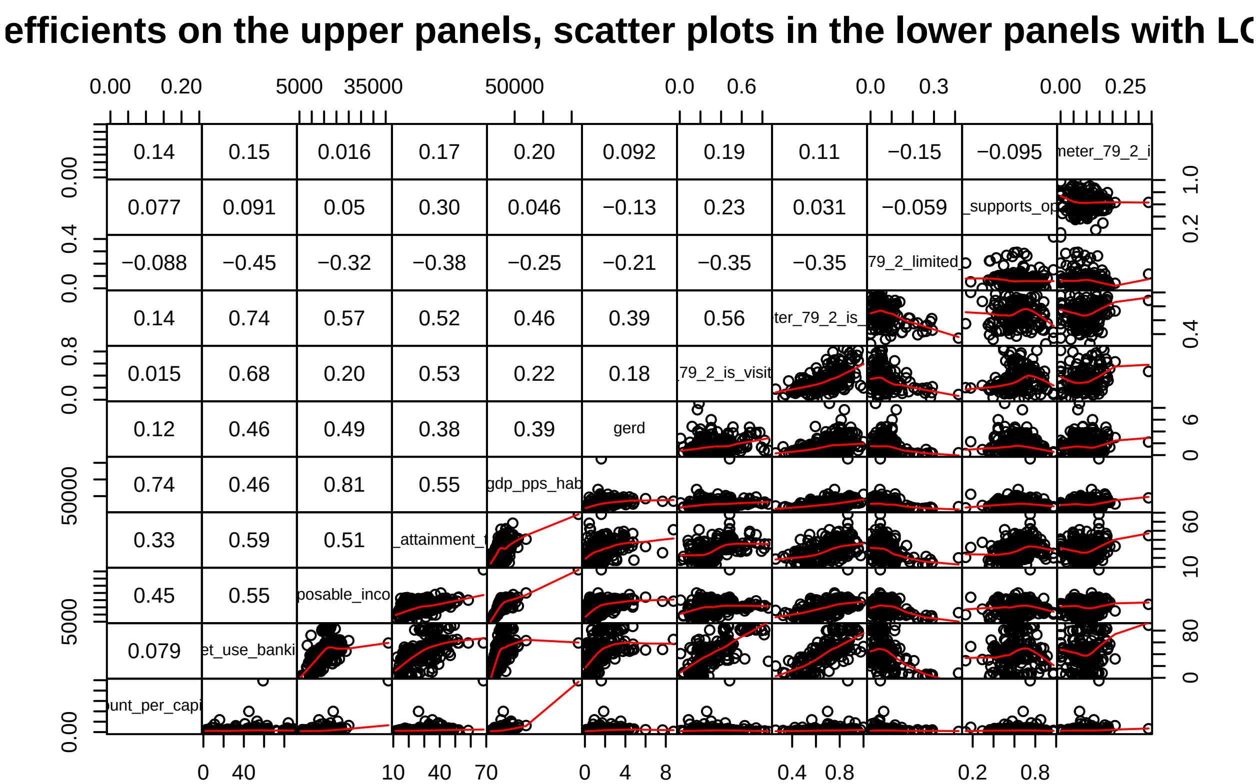
```
#count per capita
cpc_predictor_var_select = Predictor$new(
  cpc_var_select.rf,
  data = dplyr::select ( cpc_var_select_df , -geo, - count_per_capita),
  y = as.numeric(cpc_var_select_df$count_per_capita))

cpc_imp <- FeatureImp$new(cpc_predictor_var_select, loss = "mae", n.repetitions = 100)
plot(cpc_imp)
```

Count per researcher - with eurobarometer

Here we model the per researcher download counts as dependent variables.

Linear Regression with EUROBAROMETER

Since the random forest indicated that the EUROBAROMETER variables might be relevant, we tried to enrich our earlier, simplest models with them.




```
percapita_qplm3_1 <- glm (count_per_million ~
                         log(gdp_pps) +
                         researcher_employment_pct +
                         internet_use_banking_pc+
                       erobarometer_79_2_is_visit_public_library,
                       data = eurostat_eurobarometer_complete_data,
                       family = quasipoisson)


percapita_qplm3_2 <- glm (count_per_million ~
                         log(gdp_pps) +
                         researcher_employment_pct +
                         internet_use_banking_pc+
                                        erobarometer_79_2_is_read_book,
                       data = eurostat_eurobarometer_complete_data,
                       family = quasipoisson)

percapita_qplm3_3 <- glm (count_per_million ~
                         log(gdp_pps) +
                         researcher_employment_pct +
                         internet_use_banking_pc+
                                        eurobarometer_79_2_is_student,
                       data = eurostat_eurobarometer_complete_data,
                       family = quasipoisson)

percapita_qplm3_4 <- glm (count_per_million ~
```

```r
                          log(gdp_pps) +
                          researcher_employment_pct +
                          internet_use_banking_pc+
                                         erobarometer_79_2_limited_library_supply,

                        data = eurostat_eurobarometer_complete_data,
                        family = quasipoisson)

percapita_qplm3_5 <- glm (count_per_million ~
                          log(gdp_pps) +
                          researcher_employment_pct +
                          internet_use_banking_pc+
                                         erobarometer_79_2_supports_open_access_science,
                        data = eurostat_eurobarometer_complete_data,
                        family = quasipoisson)


export_summs(percapita_qplm3_1, percapita_qplm3_2, percapita_qplm3_3, percapita_qplm3_4,percapita_qplm3_
             digits=3,
             statistics = c("null.deviance", "deviance")
             )
```

```
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
```

```
## Warning in if (statistics == "all") {: the condition has length > 1 and only the
## first element will be used
```

```
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
## Note: Pseudo-R2 for quasibinomial/quasipoisson families is calculated by
## refitting the fitted and null models as binomial/poisson.
```

```r
perresearcher_qplm3_1 <- glm (count_per_thousand_researchers ~
                          log(gdp_pps) +
                          researcher_employment_pct +
                          internet_use_banking_pc+
                        erobarometer_79_2_is_visit_public_library,
                        data = eurostat_eurobarometer_complete_data,
                        family = quasipoisson)
```

```r
summary(perresearcher_qplm3_1)
```

```
##
## Call:
## glm(formula = count_per_thousand_researchers ~ log(gdp_pps) +
##     researcher_employment_pct + internet_use_banking_pc + erobarometer_79_2_is_visit_public_library,
##     family = quasipoisson, data = eurostat_eurobarometer_complete_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -84.276  -24.920  -11.238    6.816  221.363
##
## Coefficients:
##                                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                7.090654   0.805511   8.803 4.76e-16
## log(gdp_pps)                               0.135173   0.079629   1.698 0.091063
## researcher_employment_pct                 -0.167086   0.144926  -1.153 0.250247
## internet_use_banking_pc                   -0.016385   0.004261  -3.846 0.000159
## erobarometer_79_2_is_visit_public_library  0.138979   0.537682   0.258 0.796289
##
## (Intercept)                               ***
## log(gdp_pps)                              .
## researcher_employment_pct
## internet_use_banking_pc                   ***
## erobarometer_79_2_is_visit_public_library
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 2004.342)
##
##     Null deviance: 350304  on 216  degrees of freedom
## Residual deviance: 282840  on 212  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```r
rawcount_qplm3_1 <- glm (count ~
                          log(gdp_pps) +
                          researcher_employment_pct +
                          internet_use_banking_pc+
                        erobarometer_79_2_is_visit_public_library,
                        data = eurostat_eurobarometer_complete_data,
                        family = quasipoisson)

rawcount_qplm3_2 <- glm (count ~
                          log(gdp_pps) +
                          researcher_employment_pct +
                          internet_use_banking_pc+
                        erobarometer_79_2_limited_library_supply,
                        data = eurostat_eurobarometer_complete_data,
                        family = quasipoisson)
summary(rawcount_qplm3_2)
```

```
##
```

```
## Call:
## glm(formula = count ~ log(gdp_pps) + researcher_employment_pct +
##     internet_use_banking_pc + erobarometer_79_2_limited_library_supply,
##     family = quasipoisson, data = eurostat_eurobarometer_complete_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -350.26   -52.32   -22.59    12.41   587.98
##
## Coefficients:
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               0.069472   0.744248   0.093    0.926
## log(gdp_pps)                              0.935546   0.063825  14.658  < 2e-16
## researcher_employment_pct                 0.388654   0.053117   7.317 5.16e-12
## internet_use_banking_pc                  -0.013836   0.003216  -4.302 2.58e-05
## erobarometer_79_2_limited_library_supply -0.919495   1.275537  -0.721    0.472
##
## (Intercept)
## log(gdp_pps)                             ***
## researcher_employment_pct                ***
## internet_use_banking_pc                  ***
## erobarometer_79_2_limited_library_supply
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 12337.44)
##
##     Null deviance: 6134520  on 216  degrees of freedom
## Residual deviance: 1673143  on 212  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

We found no significant effect of any of the eurobarometer variables for the per capita or per researcher downloads, only in the raw count model is the share the library users has a marginally significant, negative effect, which is not strong enough to speak of a replacement effect, especially given that scholarly pirates are most probably avid readers and library users as well.

Other - simple random models, just for the sake of doing it

Here are the results of the per capita, per researchers, and raw count simple linear models.

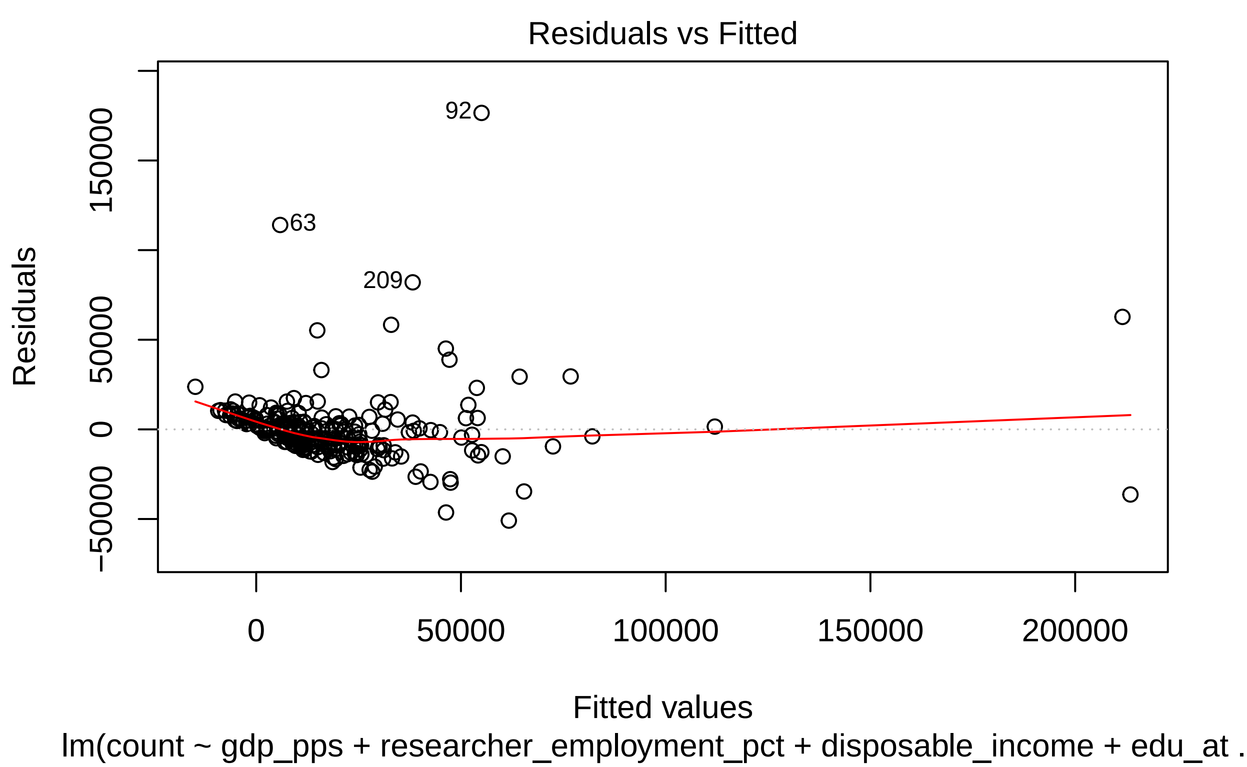
```
percapita_lm <- lm (count_per_million ~ gdp_pps +
                      researcher_employment_pct +
                      disposable_income +
                      edu_attainment_total +
                      internet_use_banking_pc,
                    data = eurostat_complete_data)

perresearcher_lm <- lm (count_per_thousand_researchers ~
                          gdp_pps  + disposable_income +
                          edu_attainment_total +
                          internet_use_banking_pc,
                        data = eurostat_complete_data)
```

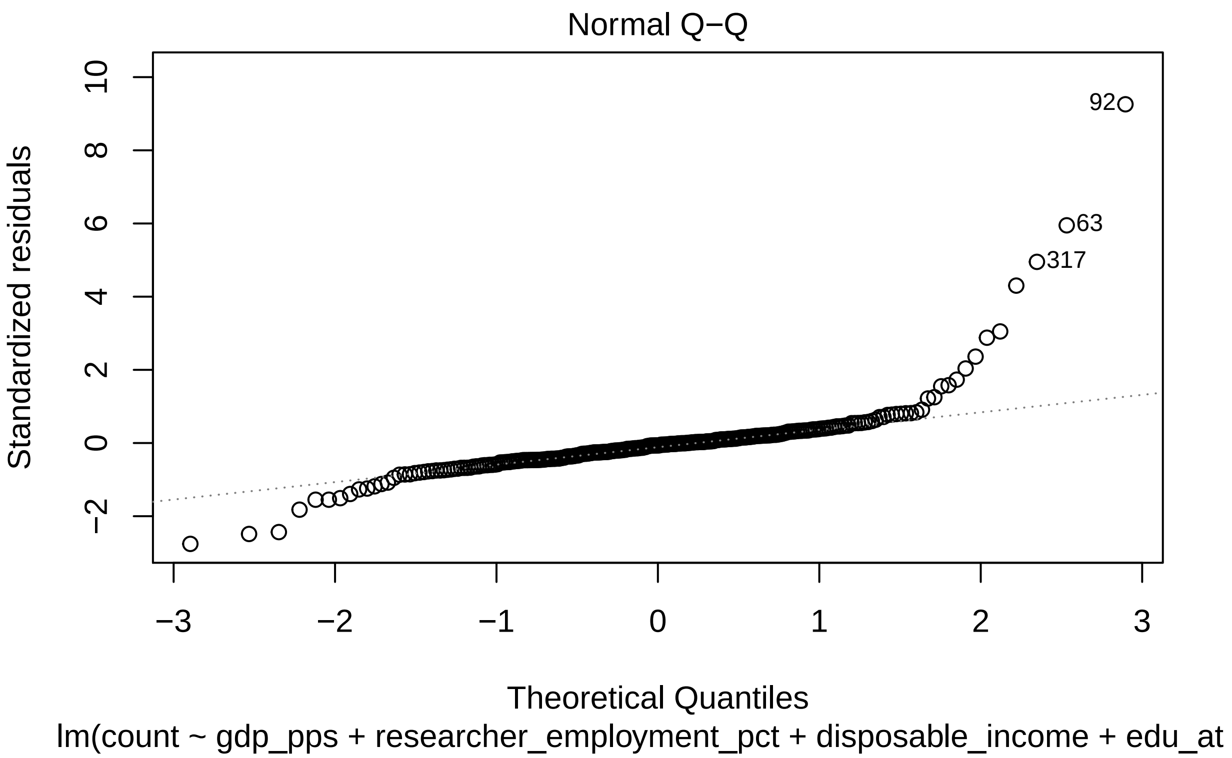
```r
count_lm <- lm (count ~ gdp_pps +
                    researcher_employment_pct +
                    disposable_income +
                    edu_attainment_total +
                    internet_use_banking_pc +
                    gerd+internet_purchases_last_year_pc,
                data = eurostat_complete_data)
vif(count_lm)
```

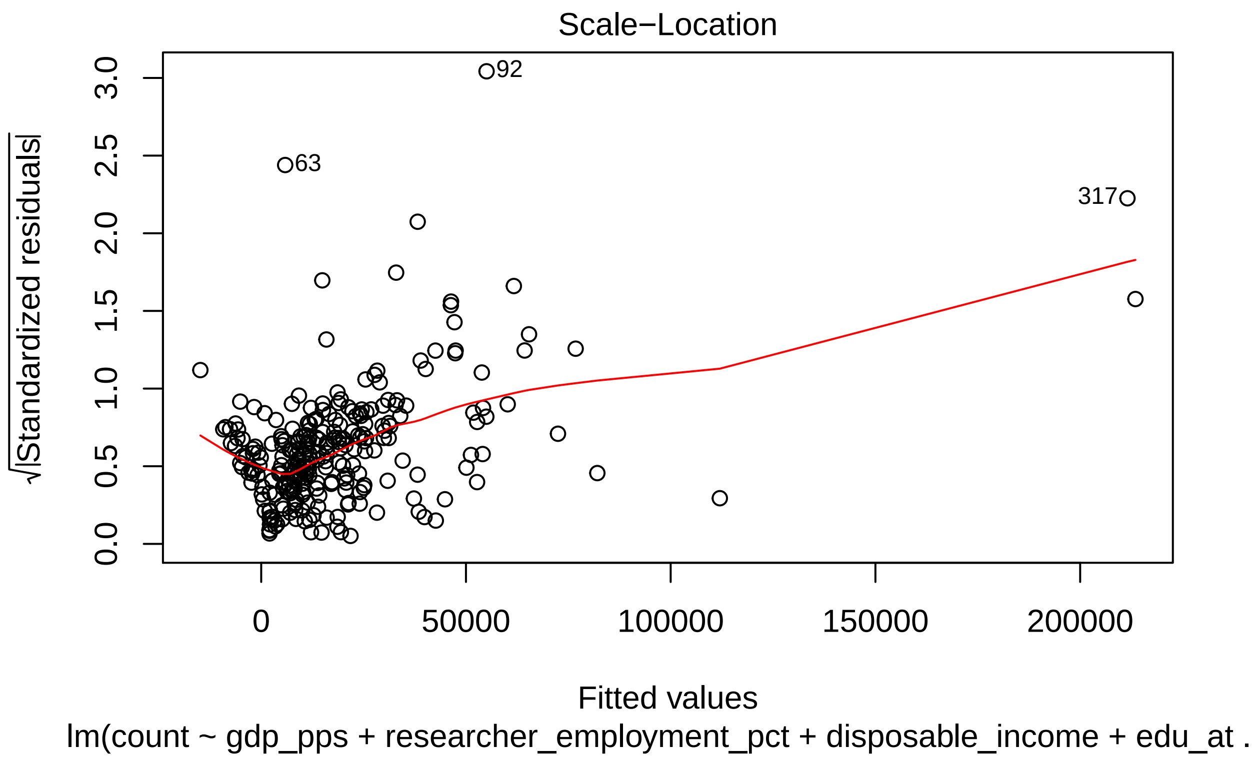
```
##                         gdp_pps       researcher_employment_pct
##                        1.235016                        3.182476
##               disposable_income            edu_attainment_total
##                        2.381141                        2.502546
##         internet_use_banking_pc                            gerd
##                        3.318307                        2.439704
## internet_purchases_last_year_pc
##                        4.715426
```

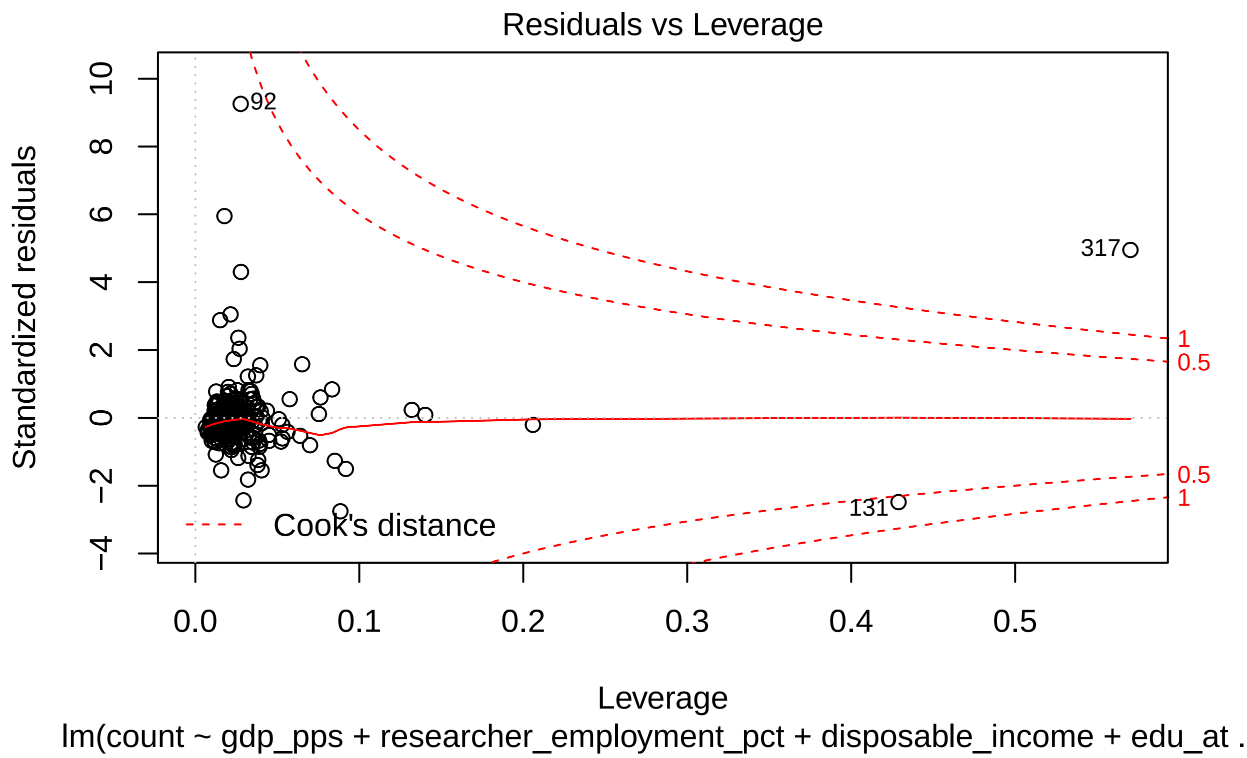
```r
export_summs(percapita_lm,perresearcher_lm,count_lm,
             digits=10, statistics = "all",
             model.names=c('percapita', 'perresearcher','count')
             )
```

```r
plot(count_lm)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(count ~ gdp_pps + researcher_employment_pct + disposable_income + edu_at .
92
63
317
−2
0
2
4
6
8
10
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
50000
100000
150000
200000
92
63
317
Fitted values
lm(count ~ gdp_pps + researcher_employment_pct + disposable_income + edu_at .

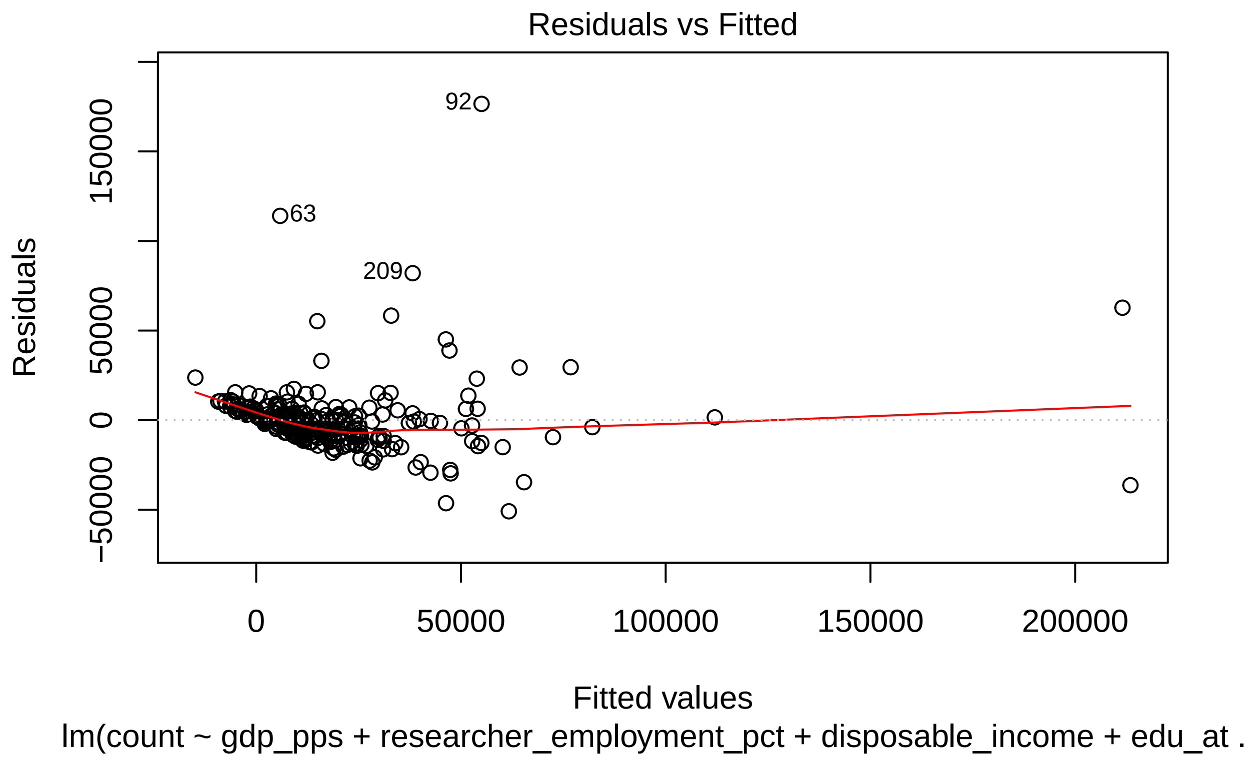


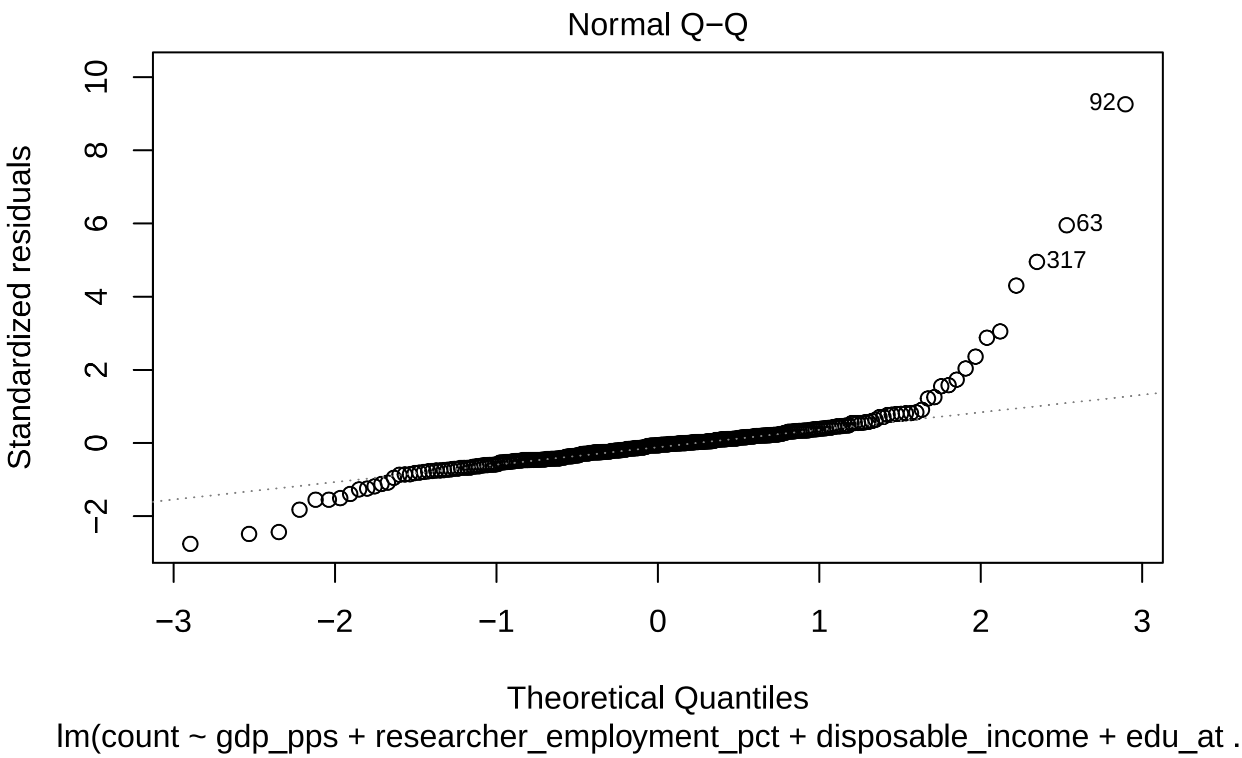
The standard linear model for the count variable shows the same as the Poisson and quasipoisson models:

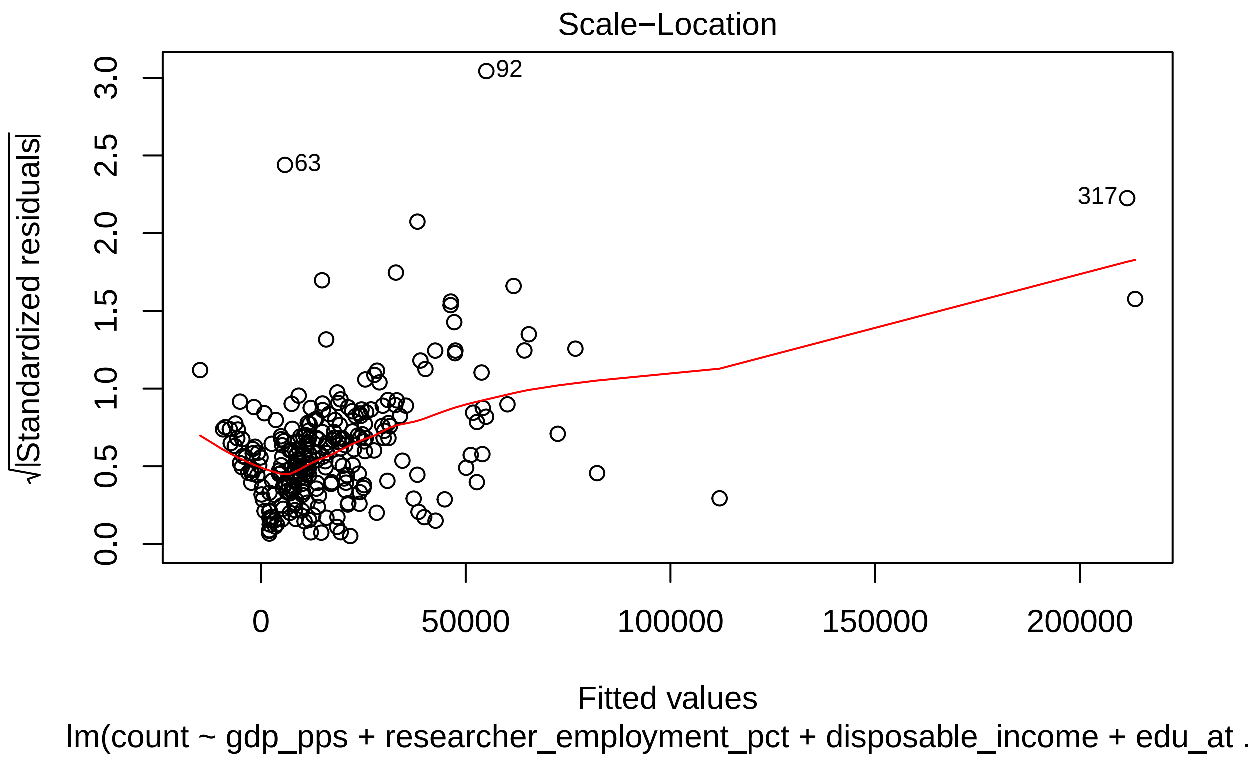
- downloads grow with wealth, researcher share, and disposable income, but are moderated by online proficiency and R&D investment.
- online purchases are not significant, neither is the level of education. the model fit is comparable to the quasipoisson model fits.

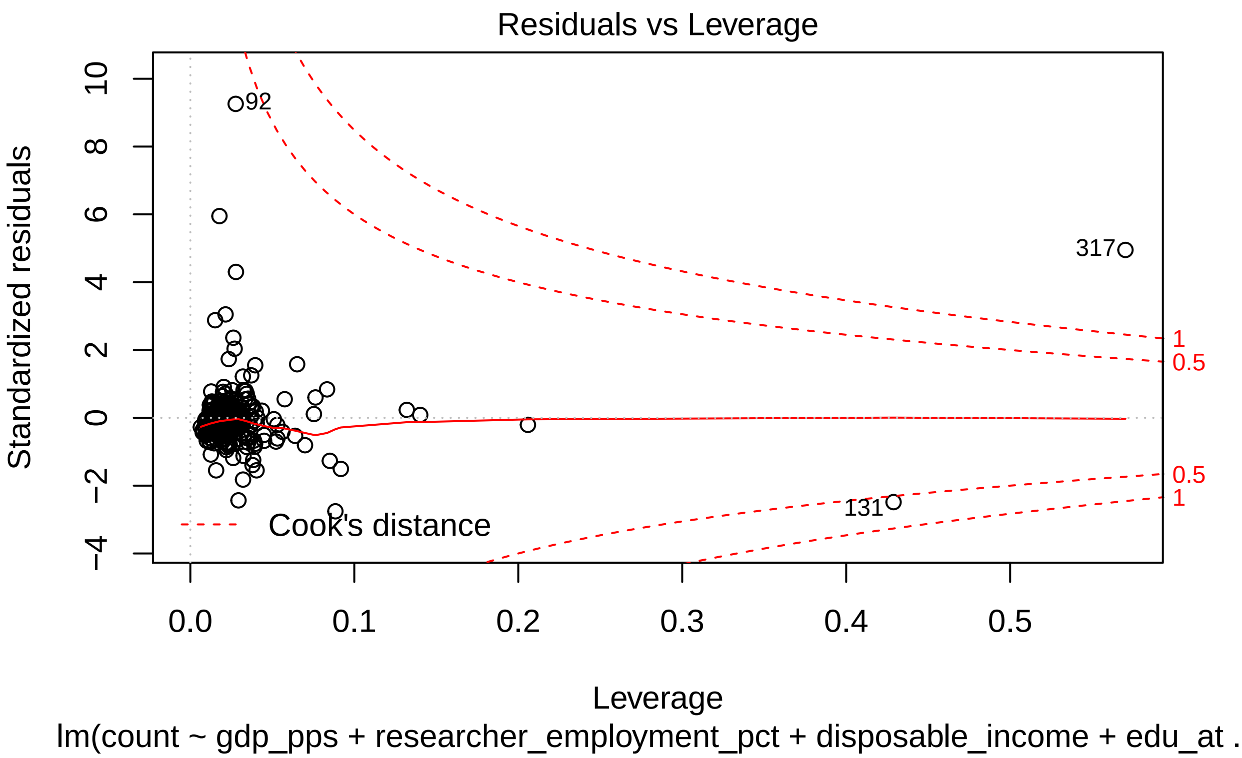
we also have to note that linear models behave extremely bad in high count regions, worse than quasipoisson models.

```
plot(count_lm)
```


Residuals vs Fitted
Residuals
Fitted values
150000
50000
0
−50000
200000
100000
92
63
209
lm(count ~ gdp_pps + researcher_employment_pct + disposable_income + edu_at .


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(count ~ gdp_pps + researcher_employment_pct + disposable_income + edu_at .
92
63
317
10
8
6
4
2
0
−2
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
92
63
317
0
50000
100000
150000
200000
Fitted values
lm(count ~ gdp_pps + researcher_employment_pct + disposable_income + edu_at .


Residuals vs Leverage
Standardized residuals
10
8
6
4
2
0
−2
−4
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
lm(count ~ gdp_pps + researcher_employment_pct + disposable_income + edu_at .
92
317
131
1
0.5
0.5
1
Cook's distance

Parameter1	Parameter2	r	CI_low	CI_high	t	df	p	Method	n_O
area_land_fille	count	-0.114	-0.232	0.00637	-1.86	263	1	Pearson	2
area_land_filled	disposable_income	-0.16	-0.276	-0.0406	-2.63	263	0.875	Pearson	2
area_land_fille	edu_attainmer	-0.0404	-0.16	0.0805	-0.656	263	1	Pearson	2
area_land_filled	employment_total	0.081	-0.0399	0.2	1.32	263	1	Pearson	2
area_land_fille	gdp	-0.000454	-0.121	0.12	-0.00736	263	1	Pearson	2
area_land_filled	gdp_pps	0.00738	-0.113	0.128	0.12	263	1	Pearson	2
area_land_fille	gdp_pps_hab	-0.147	-0.263	-0.0273	-2.42	263	1	Pearson	2
area_land_filled	gerd	-0.0109	-0.131	0.11	-0.176	263	1	Pearson	2
area_land_fille	high_skilled_p	-0.134	-0.25	-0.0137	-2.19	263	1	Pearson	2
area_land_filled	hr_scitech_pct	-0.134	-0.25	-0.0137	-2.19	263	1	Pearson	2
area_land_fille	internet_purch	-0.099	-0.217	0.0217	-1.61	263	1	Pearson	2
area_land_filled	internet_use_banking_pc	0.091	-0.0298	0.209	1.48	263	1	Pearson	2
area_land_fille	internet_use_c	-0.0867	-0.205	0.0342	-1.41	263	1	Pearson	2
area_land_filled	internet_use_social_networks_pc	-0.0237	-0.144	0.097	-0.385	263	1	Pearson	2
area_land_fille	population_tot	0.13	0.00957	0.247	2.12	263	1	Pearson	2
area_land_filled	researcher_employment_pct	-0.0515	-0.171	0.0695	-0.836	263	1	Pearson	2
area_land_fille	researchers_tot	-0.0144	-0.135	0.106	-0.234	263	1	Pearson	2
area_land_filled	count_per_million	-0.147	-0.263	-0.0271	-2.41	263	1	Pearson	2
area_land_fille	count_per_cap	-0.147	-0.263	-0.0271	-2.41	263	1	Pearson	2
area_land_filled	count_per_area	-0.0711	-0.19	0.0498	-1.16	263	1	Pearson	2
area_land_fille	count_per_tho	-0.0597	-0.179	0.0613	-0.969	263	1	Pearson	2
area_land_filled	count_per_researcher	-0.0597	-0.179	0.0613	-0.969	263	1	Pearson	2
area_land_fille	count_per_po	0.638	0.56	0.704	13.4	263	2.62e-29	Pearson	2
count	disposable_income	0.336	0.225	0.439	5.79	263	3.28e-06	Pearson	2
count	edu_attainmer	0.255	0.139	0.365	4.28	263	0.00357	Pearson	2
count	employment_total	0.551	0.461	0.63	10.7	263	4.19e-20	Pearson	2
count	gdp	0.644	0.567	0.709	13.6	263	4.72e-30	Pearson	2
count	gdp_pps	0.668	0.596	0.73	14.6	263	2.6e-33	Pearson	2
count	gdp_pps_hab	0.55	0.46	0.629	10.7	263	5.43e-20	Pearson	2
count	gerd	0.162	0.0427	0.277	2.67	263	0.827	Pearson	2
count	high_skilled_p	0.31	0.197	0.415	5.28	263	4.22e-05	Pearson	2
count	hr_scitech_pct	0.31	0.197	0.415	5.28	263	4.22e-05	Pearson	2
count	internet_purch	0.053	-0.068	0.172	0.86	263	1	Pearson	2
count	internet_use_banking_pc	-0.021	-0.141	0.0998	-0.34	263	1	Pearson	2
count	internet_use_c	0.113	-0.00755	0.23	1.85	263	1	Pearson	2
count	internet_use_social_networks_pc	0.0202	-0.101	0.14	0.327	263	1	Pearson	2

	Model 1	Model 2	Model 3
(Intercept)	5.530 *	8.034 **	6.486 ***
	(2.240)	(2.546)	(1.147)
log(gdp_pps)	0.161 *	0.174 *	0.175
	(0.071)	(0.078)	(0.113)
log(disposable_income)	0.148	-0.143	
	(0.255)	(0.291)	
edu_attainment_total	0.008	-0.000	
	(0.008)	(0.009)	
gerd	-0.253 **	-0.310 ***	-0.155
	(0.079)	(0.088)	(0.901)
internet_use_banking_pc	-0.018 ***		
	(0.004)		
internet_purchases_last_year_pc		-0.006	
		(0.004)	
log(gdp_pps):gerd			-0.024
			(0.084)
null.deviance	495798.787	495798.787	495798.787
deviance	362061.621	398536.668	414631.765

*** p < 0.001; ** p < 0.01; * p < 0.05.

	Model 1	Model 2	
(Intercept)	0.6211211100	0.5730664396	
	(2.1132612052)	(2.1172749655)	(
log(gdp_pps)	0.9551262540 ***	0.9590985572 ***	0
	(0.0632768252)	(0.0640742590)	(
researcher_employment_pct	0.3181504930 ***	0.3271627382 ***	0
	(0.0863298463)	(0.0890344766)	(
log(disposable_income)	-0.1028660123	-0.0997598068	-
	(0.2460385841)	(0.2461695466)	(
edu_attainment_total	0.0147511373 *	0.0138489820	
	(0.0074391023)	(0.0077696817)	(
internet_use_banking_pc	-0.0195794874 ***	-0.0189299950 ***	
	(0.0034621068)	(0.0038077620)	
gerd		-0.0190133191	-
		(0.0466327558)	(
internet_purchases_last_year_pc			-(
			(
nobs	265	265	
null.deviance	7192467.3824574202	7192467.3824574202	719246
df.null	264.0000000000	264.0000000000	26
logLik			
AIC			
BIC			
deviance	1992084.2207974601	1990052.0420572299	222881
df.residual	259.0000000000	258.0000000000	25
pseudo.r.squared	1.0000000000	1.0000000000	
pseudo.r.squared.mcfadden	0.7227462984	0.7230287294	

*** p < 0.001; ** p < 0.01; * p < 0.05.

	percapita	perresearcher	coun
(Intercept)	6.4383579762 ***	5.5298545753 *	0.62112
	(0.7937306511)	(2.2398209742)	(2.11326
log(gdp_pps)	0.2465563796 **	0.1609069163 *	0.955126
	(0.0769748083)	(0.0705484186)	(0.06327
researcher_employment_pct	0.6966753816 ***		0.318150
	(0.0565692560)		(0.08632
internet_use_banking_pc	-0.0111376705 ***	-0.0182764133 ***	-0.019579
	(0.0032270824)	(0.0041940523)	(0.00346
log(disposable_income)		0.1480069549	-0.10286
		(0.2548048700)	(0.24603
edu_attainment_total		0.0075568225	0.01475
		(0.0076812247)	(0.00743
gerd		-0.2532864784 **	
		(0.0793017980)	
nobs	265	265	265
null.deviance	2990524.3713359898	495798.7867977770	7192467.38245
df.null	264.0000000000	264.0000000000	264.00000
logLik			
AIC			
BIC			
deviance	1415805.4334292000	362061.6209839690	1992084.22079
df.residual	261.0000000000	259.0000000000	259.00000
pseudo.r.squared	1.0000000000	1.0000000000	1.00000
pseudo.r.squared.mcfadden	0.5260853485	0.2684410421	0.72274

*** p < 0.001; ** p < 0.01; * p < 0.05.

rowname	model	r.squared	adj.r.squared	(Intercept)	names	values
21	gdp_pps_hab+i	0.156	0.15	-14.8	gdp_pps_hab	0.0012
21	gdp_pps_hab+internet_use_banking_pc	0.156	0.15	-14.8	internet_use_banking_pc	-0.227
39	internet_use_ba	0.156	0.15	-14.8	gdp_pps_hab	0.0012
39	internet_use_banking_pc+gdp_pps_hab	0.156	0.15	-14.8	internet_use_banking_pc	-0.227
20	gdp_pps_hab+i	0.144	0.138	-14.6	gdp_pps_hab	0.00113
20	gdp_pps_hab+internet_purchases_last_year_pc	0.144	0.138	-14.6	internet_purchases_last_year_pc	-0.169
36	internet_purchas	0.144	0.138	-14.6	gdp_pps_hab	0.00113
36	internet_purchases_last_year_pc+gdp_pps_hab	0.144	0.138	-14.6	internet_purchases_last_year_pc	-0.169
23	high_skilled_pc-	0.112	0.104	-26.7	high_skilled_pc	1.08
23	high_skilled_pc+erobarometer_79_2_is_visit_public_library	0.112	0.104	-26.7	erobarometer_79_2_is_visit_public_	-28.6
30	hr_scitech_pct+	0.112	0.104	-26.7	hr_scitech_pct	1.08
30	hr_scitech_pct+erobarometer_79_2_is_visit_public_library	0.112	0.104	-26.7	erobarometer_79_2_is_visit_public_	-28.6
10	erobarometer_79	0.112	0.104	-26.7	high_skilled_pc	1.08
10	erobarometer_79_2_is_visit_public_library+high_skilled_pc	0.112	0.104	-26.7	erobarometer_79_2_is_visit_public_	-28.6
11	erobarometer_79	0.112	0.104	-26.7	hr_scitech_pct	1.08
11	erobarometer_79_2_is_visit_public_library+hr_scitech_pct	0.112	0.104	-26.7	erobarometer_79_2_is_visit_public_	-28.6
26	high_skilled_pc-	0.109	0.103	-28.6	high_skilled_pc	1.17
26	high_skilled_pc+internet_use_banking_pc	0.109	0.103	-28.6	internet_use_banking_pc	-0.279
33	hr_scitech_pct+	0.109	0.103	-28.6	hr_scitech_pct	1.17
33	hr_scitech_pct+internet_use_banking_pc	0.109	0.103	-28.6	internet_use_banking_pc	-0.279
40	internet_use_ba	0.109	0.103	-28.6	high_skilled_pc	1.17
40	internet_use_banking_pc+high_skilled_pc	0.109	0.103	-28.6	internet_use_banking_pc	-0.279
41	internet_use_ba	0.109	0.103	-28.6	hr_scitech_pct	1.17
41	internet_use_banking_pc+hr_scitech_pct	0.109	0.103	-28.6	internet_use_banking_pc	-0.279
37	internet_purchas	0.105	0.099	-28.9	high_skilled_pc	1.21
37	internet_purchases_last_year_pc+high_skilled_pc	0.105	0.099	-28.9	internet_purchases_last_year_pc	-0.277
38	internet_purchas	0.105	0.099	-28.9	hr_scitech_pct	1.21
38	internet_purchases_last_year_pc+hr_scitech_pct	0.105	0.099	-28.9	internet_purchases_last_year_pc	-0.277
25	high_skilled_pc-	0.105	0.099	-28.9	high_skilled_pc	1.21
25	high_skilled_pc+internet_purchases_last_year_pc	0.105	0.099	-28.9	internet_purchases_last_year_pc	-0.277
32	hr_scitech_pct+	0.105	0.099	-28.9	hr_scitech_pct	1.21
32	hr_scitech_pct+internet_purchases_last_year_pc	0.105	0.099	-28.9	internet_purchases_last_year_pc	-0.277
42	internet_use_da	0.0964	0.0899	-16.4	high_skilled_pc	1.18
42	internet_use_daily_pc+high_skilled_pc	0.0964	0.0899	-16.4	internet_use_daily_pc	-0.399
43	internet_use_da	0.0964	0.0899	-16.4	hr_scitech_pct	1.18
43	internet_use_daily_pc+hr_scitech_pct	0.0964	0.0899	-16.4	internet_use_daily_pc	-0.399

variable	IncNodePurity
researcher_employment_pct	0.0131
edu_attainment_total	0.0121
gdp_pps_hab	0.0108
disposable_income	0.0106
high_skilled_pc	0.0103
hr_scitech_pct	0.0101
gdp	0.0023
gdp_pps	0.00215
gerd	0.00171
internet_use_banking_pc	0.00168
employment_total	0.00137
internet_use_daily_pc	0.00127
internet_purchases_last_year_]	0.00108
internet_use_social_networks_pc	0.000694

IncNodePurity
0.0124
0.011
0.0104
0.0103
0.00973
0.00844
0.0023
0.00199
0.00123
0.00114
0.000849
0.000834
0.000737
0.000726
0.000723
0.000574
0.000527
0.000505
0.000439
0.000379

IncNodePurity
125
109
108
106
80.1
71.8
70.3
56
55.6
51.3
46
43
41.1
38.2
37.4
36.4
35.8
25.4
15.7

	Model 1	Model 2	M
(Intercept)	6.196 ***	5.734 ***	
	(0.849)	(0.903)	
log(gdp_pps)	0.266 **	0.248 **	
	(0.082)	(0.086)	
researcher_employment_pct	0.678 ***	0.652 ***	
	(0.057)	(0.058)	
internet_use_banking_pc	-0.003	-0.013 **	
	(0.004)	(0.005)	
erobarometer_79_2_is_visit_public_library	-0.900		
	(0.544)		
erobarometer_79_2_is_read_book		1.173	
		(0.732)	
eurobarometer_79_2_is_student			
erobarometer_79_2_limited_library_supply			
erobarometer_79_2_supports_open_access_science			
null.deviance	2553172.101	2553172.101	2553
deviance	1058039.618	1058447.425	1074

*** p < 0.001; ** p < 0.01; * p < 0.05.

	percapita	perresearcher
(Intercept)	-12763.7378913799 ***	3710.7625502644 ***
	(3523.2692575711)	(599.9872808217)
gdp_pps	-0.0175480164	-0.0014413293
	(0.0162491675)	(0.0028579102)
researcher_employment_pct	15885.7284859612 ***	
	(1945.7957061601)	
disposable_income	1.5368260722 ***	0.0307481696
	(0.2577285755)	(0.0452299255)
edu_attainment_total	103.3661973273	12.0436603325
	(126.9985173642)	(20.4993615895)
internet_use_banking_pc	-328.6762527926 ***	-47.9116783495 ***
	(50.4213514487)	(9.0531345735)
gerd		
internet_purchases_last_year_pc		
nobs	265	265
r.squared	0.4311475504	0.1360504936
adj.r.squared	0.4201658429	0.1227589627
sigma	13526.0906886901	2429.8313802648
statistic	39.2605202459	10.2358783885
p.value	0.0000000000	0.0000001035
df	6.0000000000	5.0000000000
logLik	-2893.7637993945	-2439.3227683843
AIC	5801.5275987890	4890.6455367686
BIC	5826.5857075710	4912.1239157245
deviance	47385378493.5353012085	1535060939.4951200485
df.residual	259.0000000000	260.0000000000

*** p < 0.001; ** p < 0.01; * p < 0.05.