Doctor Visits

Charles Costanzo, Afia Fosu, Logan Loftus, Mason Rocco

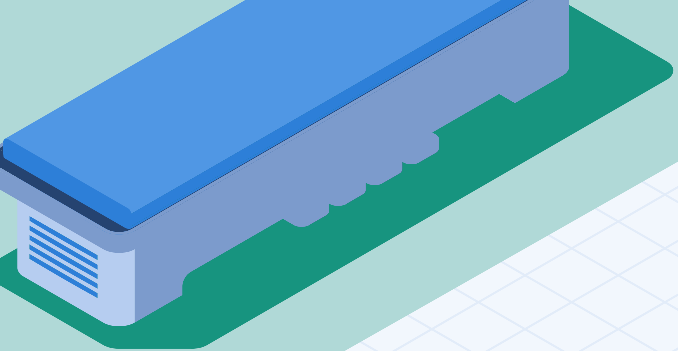
DoctorVisits from AER package

Australian Health Service Utilization Data

- Cross-section data originating from the 1977–1978 Australian Health Survey.
- A data frame containing 5,190 observations on 12 variables.
- Source: Journal of Applied Econometrics Data Archive

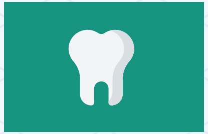
Variable	Description
visits	Number of doctor visits in past 2 weeks.
gender	Factor indicating gender.
age	Age in years divided by 100.
income	Annual income in tens of thousands of dollars.
illness	Number of illnesses in past 2 weeks.
reduced	Number of days of reduced activity in past 2 weeks due to illness or injury.
health	General health questionnaire score using Goldberg's method.
private	Factor. Does the individual have private health insurance?
freepoor	Factor. Does the individual have free government health insurance due to low income?
freerepat	Factor. Does the individual have free government health insurance due to old age, disability or veteran status?
nchronic	Factor. Is there a chronic condition not limiting activity?
lchronic	Factor. Is there a chronic condition limiting activity?

*Note: this data restricts “gender” to a binary “male” or “female” variable.
A more appropriate (but still imperfect) variable name may be “sex”, as the use of “gender” does not consider other genders. However, to avoid overcomplicating our analysis, we will keep the variable as is.



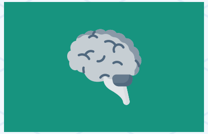
Our Questions

What we will investigate



Q1

Is there an association between gender and the number of doctor visits?



Q2

Is there an association between income and number of doctor visits?

Q3

Is there an association between insurance type and number of doctor visits?

What we will investigate

Q4

Is there an association between income and the probability of having free government insurance due to low income?

Q5

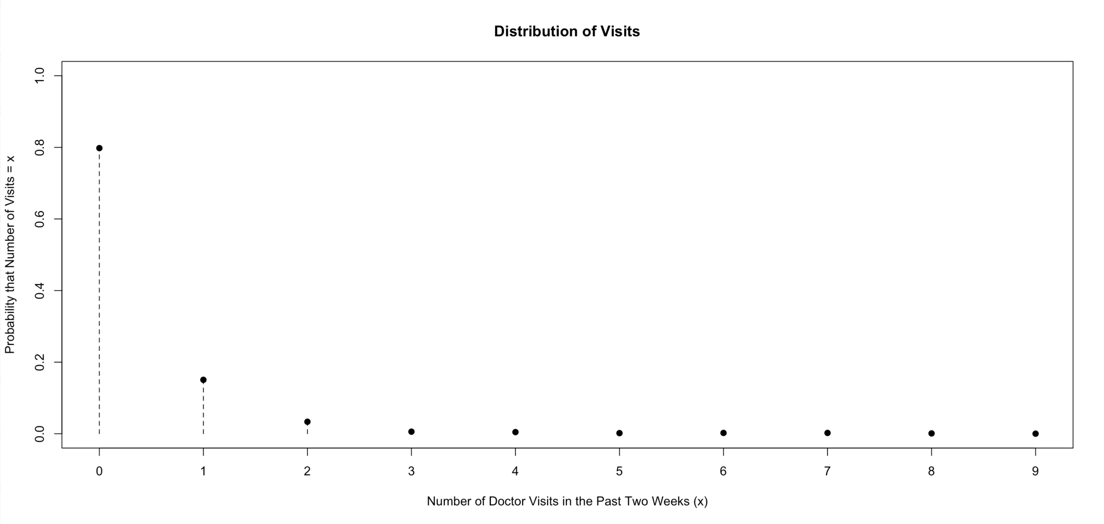
Is there an association between overall health and the probability of having free government insurance due to low income?

Q6

Is there an association between age and the probability of having free government insurance due to low income?

Data Exploration

Density of Doctor Visits



Poisson

Discrete distribution.
Models counts of “rare” events.

Zero Inflated?

Large probability when number of visits is 0.

Modeling

A count outcome suggests a Poisson General Linear Model may be appropriate.

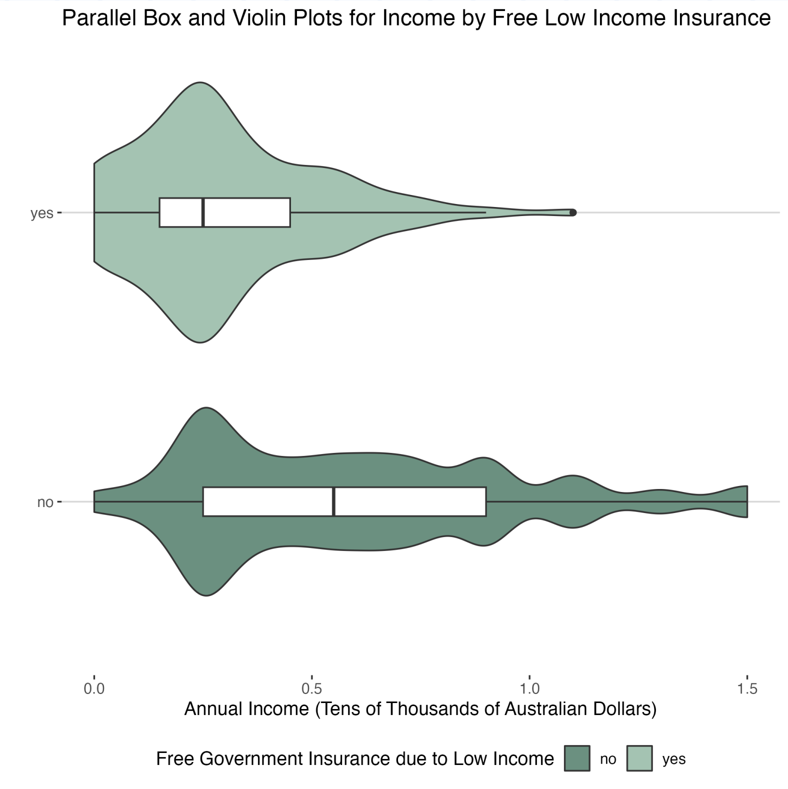
Mean Number of Doctor Visits by Sex

Sex	Mean Doctor Visits (past 2 weeks)
Male	0.2363344
Female	0.3619541

Mean Number of Doctor Visits by Insurance Type

Insurance Type	Mean Doctor Visits (past 2 weeks)
Private	0.295
Free Government Insurance due to Low Income	0.158
Free Government Insurance due to Old Age, Disability, or Veteran Status	0.467
No Insurance	0.218

Free Government Insurance due to Low Income and Income



Income

It appears that those with free insurance tend to have a lower mean annual income than those without free insurance.

$10,000 Australian dollars in 1977 is roughly equal to $45,000 U.S. dollars in 2023.

Outliers

There are a few outliers (very high income) in the no private insurance group.

Modeling

Generalized Linear Model for Poisson Data

Model Description

Let $\mathbf{Y}$ be our vector of data containing the number of doctor visits within the past two weeks, `visits`.

$$\text{Assume } \{Y_1 \ldots, Y_{n=5190}\} \overset{ind}{\sim} Poisson(\lambda_i)$$

We will use the log link.

Poisson GLM:

$$\eta_i = g(\lambda_i) = log(\lambda_i) = \mathbf{x}_i^\intercal \boldsymbol{\beta}, \quad i = 1, \ldots, n = 5190$$

where

- η_i is the estimated log mean number of doctor visits in the past two weeks for individual i (according to their specific covariate values $\mathbf{x}_i$)
- $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p=12})^\intercal$ are the p covariates (including an intercept) for individual i
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p=12})$ is our unknown coefficient vector.

Full Poisson Model

- Assumes the mean is equal to the variance.
- Takes the dispersion parameter to be 1.

```
##
## Call:
## glm(formula = visits ~ ., family = "poisson", data = doctor)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9502  -0.6858  -0.5747  -0.4852   5.7055
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.097821   0.101554 -20.657  < 2e-16 ***
## genderfemale   0.156490   0.056139   2.788  0.00531 **
## age            0.279123   0.165981   1.682  0.09264 .
## income        -0.187416   0.085478  -2.193  0.02834 *
## illness        0.186156   0.018263  10.193  < 2e-16 ***
## reduced        0.126690   0.005031  25.184  < 2e-16 ***
## health         0.030683   0.010074   3.046  0.00232 **
## privateyes     0.126498   0.071552   1.768  0.07707 .
## freepooryes   -0.438462   0.179799  -2.439  0.01474 *
## freerepatyes   0.083640   0.092070   0.908  0.36365
## nchronicyes    0.117300   0.066545   1.763  0.07795 .
## lchronicyes    0.150717   0.082260   1.832  0.06692 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5634.8  on 5189  degrees of freedom
## Residual deviance: 4380.1  on 5178  degrees of freedom
## AIC: 6735.7
##
## Number of Fisher Scoring iterations: 6
```

Analysis of Deviance Table

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: visits
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       5189     5634.8
## gender     1    68.64      5188     5566.2 < 2.2e-16 ***
## age        1   118.90      5187     5447.3 < 2.2e-16 ***
## income     1    12.42      5186     5434.9 0.0004239 ***
## illness    1   354.29      5185     5080.6 < 2.2e-16 ***
## reduced    1   673.63      5184     4406.9 < 2.2e-16 ***
## health     1     9.57      5183     4397.4 0.0019818 **
## private    1     3.95      5182     4393.4 0.0468745 *
## freepoor   1     7.96      5181     4385.5 0.0047840 **
## freerepat  1     1.25      5180     4384.2 0.2644339
## nchronic   1     0.74      5179     4383.5 0.3897203
## lchronic   1     3.34      5178     4380.1 0.0676851 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Used the `anova()` function in `R`

Test Whether the Full Model is Needed

Conduct a Chi-Square Test at the $\alpha = 0.05$ level to demonstrate that the `freerepat`, `nchronic`, and `lchronic` variables should be removed from the model.

H_0 : Reduced Model (without `freerepat`, `nchronic`, and `lchronic`) is good enough vs.

H_1 : Full Model (including all predictor variables) is Needed

- $D_{full} = 4380.1$
- $D_{reduced} = 4385.5$

Test Statistic:

$$\Delta D = D_{reduced} - D_{full} = 4385.5 - 4380.1$$
$$\Rightarrow \Delta D = 5.4$$

Under $H_0, \Delta D \sim \chi^2_{pFull-pReduced=12-9}$

$$\Rightarrow \Delta D \sim \chi^2_3$$

Rejection Rule: Reject if $\Delta D >$ `qchisq(1-0.05,3)` $= 7.814728$

Since our Test Statistic $\Delta D = 5.4 < 7.8$, we fail to reject the null hypothesis at the $\alpha = 0.05$ significance level.

p value: `pchisq(5.4, 3, lower.tail = FALSE)` $= 0.1447436$

Conclusion:

- **The reduced model is good enough!**
  - The model that does not include `freerepat`, `nchronic`, and `lchronic` provides a better fit to the data set than the full model that includes those three covariates.
  - We will now drop those covariates and fit a reduced model.

Check for Overdispersion

```
##
##  Overdispersion test
##
## data:  model1
## z = 6.5386, p-value = 3.105e-11
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   1.415602
```

There is indeed overdispersion, so run a quasipoisson model

*Used the `dispersiontest()` function from the `AER` package in `R`.

(Reduced) quasipoisson Model

- Same coefficient estimates as before
- But no longer assumes the mean is equal to the variance
- Dispersion parameter now 1.32

```
##
## Call:
## glm(formula = visits ~ gender + age + income + illness + reduced +
##     health + private + freepoor, family = "quasipoisson", data = doctor)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0180  -0.6811  -0.5772  -0.4916   5.6590
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.072446   0.115325 -17.970  < 2e-16 ***
## genderfemale  0.167591   0.064003   2.618  0.00886 **
## age           0.437894   0.157775   2.775  0.00553 **
## income       -0.203978   0.096926  -2.104  0.03539 *
## illness       0.196366   0.020262   9.692  < 2e-16 ***
## reduced       0.127994   0.005645  22.672  < 2e-16 ***
## health        0.032854   0.011465   2.865  0.00418 **
## privateyes    0.087156   0.061583   1.415  0.15705
## freepooryes  -0.465788   0.203005  -2.294  0.02180 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.324931)
##
##     Null deviance: 5634.8  on 5189  degrees of freedom
## Residual deviance: 4385.5  on 5181  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

Count Model Summaries

	Dependent variable:		
	visits		
	Poisson		*glm: quasipoisson* *link = log*
	(1)	(2)	(3)
genderfemale	0.156^{***}	0.168^{***}	0.168^{***}
	(0.056)	(0.056)	(0.064)
age	0.279^{*}	0.438^{***}	0.438^{***}
	(0.166)	(0.137)	(0.158)
income	-0.187^{**}	-0.204^{**}	-0.204^{**}
	(0.085)	(0.084)	(0.097)
illness	0.186^{***}	0.196^{***}	0.196^{***}
	(0.018)	(0.018)	(0.020)
reduced	0.127^{***}	0.128^{***}	0.128^{***}
	(0.005)	(0.005)	(0.006)
health	0.031^{***}	0.033^{***}	0.033^{***}
	(0.010)	(0.010)	(0.011)
privateyes	0.126^{*}	0.087	0.087
	(0.072)	(0.054)	(0.062)
freepooryes	-0.438^{**}	-0.466^{***}	-0.466^{**}
	(0.180)	(0.176)	(0.203)
freerepatyes	0.084		
	(0.092)		
nchronicyes	0.117^{*}		
	(0.067)		
lchronicyes	0.151^{*}		
	(0.082)		
Constant	-2.098^{***}	-2.072^{***}	-2.072^{***}
	(0.102)	(0.100)	(0.115)
Observations	5,190	5,190	5,190
Log Likelihood	-3,355.850	-3,358.512	
Akaike Inf. Crit.	6,735.701	6,735.024	
Note:			*p<0.1; **p<0.05; ***p<0.01

Model Interpretation

Coefficients from (reduced) quasipoisson Model

$\hat{\beta}_{female} = 0.168 \quad => \quad e^{0.168} = 1.183$

For a female, vs. male, the estimated mean number of doctor visits within the past two weeks increases 1.183 times (18.3% increase), keeping all other variables constant.

$\hat{\beta}_{income} = -0.204 \quad => \quad e^{0.204} = 0.816$

For an increase of annual income by ten thousand dollars, the estimated mean number of doctor visits within the past two weeks decreases 0.816 times (18.4% decrease), keeping all other variables constant.

$\hat{\beta}_{privateyes} = 0.087 \quad => \quad e^{0.087} = 1.091$

For an individual with private health insurance, vs. no insurance, the estimated mean number of doctor visits within the past two weeks increases 1.091 times (9.1% increase), keeping all other variables constant.

$\hat{\beta}_{freepooryes} = -0.466 \quad => \quad e^{0.466} = 0.628$

For an individual with free government health insurance due to low income, vs. no insurance, the estimated mean number of doctor visits within the past two weeks decreases 0.628 times (37.2% decrease), keeping all other variables constant.

$\hat{\beta}_{age} = 0.438 \quad => \quad e^{0.438} = 1.55$

For an increase in age by one year, the estimated mean number of doctor visits within the past two weeks increases 1.55 times (55% increase), keeping all other variables constant.

$\hat{\beta}_{illness} = 0.196 \quad => \quad e^{0.196} = 1.217$

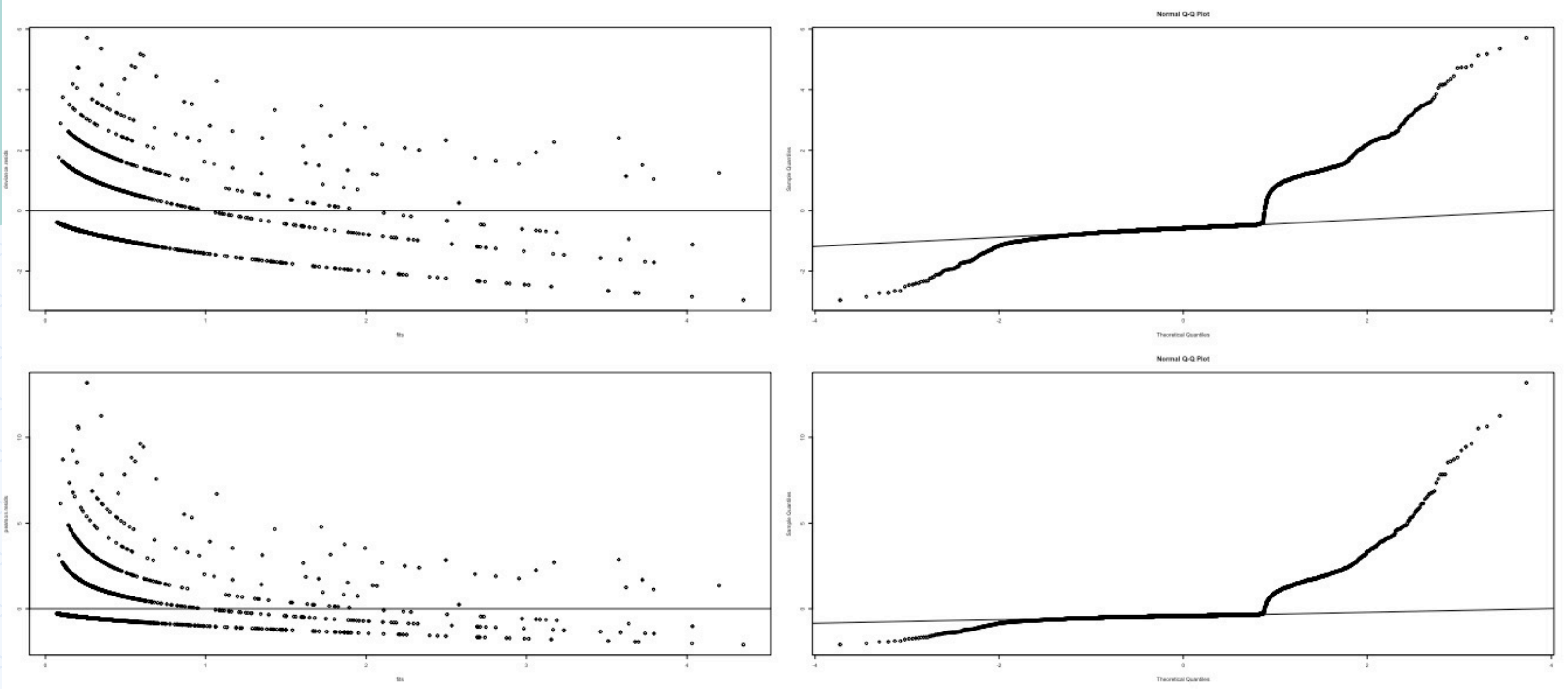
For an increase of one illness within the past two weeks, the estimated mean number of doctor visits within the past two weeks increases 1.217 times (21.7% increase), keeping all other variables constant.

Reference categories: Gender = male. Insurance type = No insurance

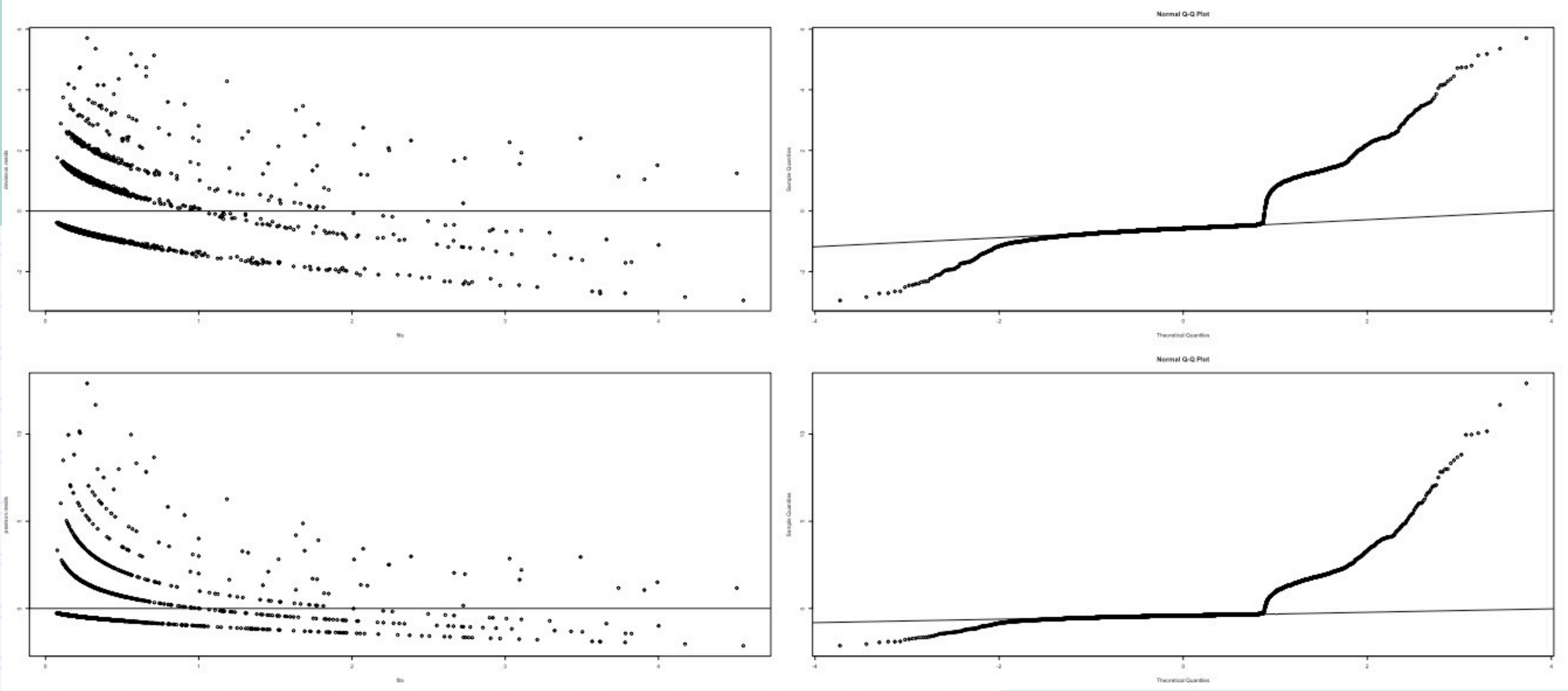
	Dependent variable:		
	visits		
	Poisson		*glm: quasipoisson* *link = log*
	(1)	(2)	(3)
genderfemale	0.156***	0.168***	0.168***
	(0.056)	(0.056)	(0.064)
age	0.279*	0.438***	0.438***
	(0.166)	(0.137)	(0.158)
income	-0.187**	-0.204**	-0.204**
	(0.085)	(0.084)	(0.097)
illness	0.186***	0.196***	0.196***
	(0.018)	(0.018)	(0.020)
reduced	0.127***	0.128***	0.128***
	(0.005)	(0.005)	(0.006)
health	0.031***	0.033***	0.033***
	(0.010)	(0.010)	(0.011)
privateyes	0.126*	0.087	0.087
	(0.072)	(0.054)	(0.062)
freepooryes	-0.438**	-0.466***	-0.466**
	(0.180)	(0.176)	(0.203)
freerepatyes	0.084		
	(0.092)		
nchronicyes	0.117*		
	(0.067)		
lchronicyes	0.151*		
	(0.082)		
Constant	-2.098***	-2.072***	-2.072***
	(0.102)	(0.100)	(0.115)
Observations	5,190	5,190	5,190
Log Likelihood	-3,355.850	-3,358.512	
Akaike Inf. Crit.	6,735.701	6,735.024	
Note:			*p<0.1; **p<0.05; ***p<0.01

Count Model Diagnostics

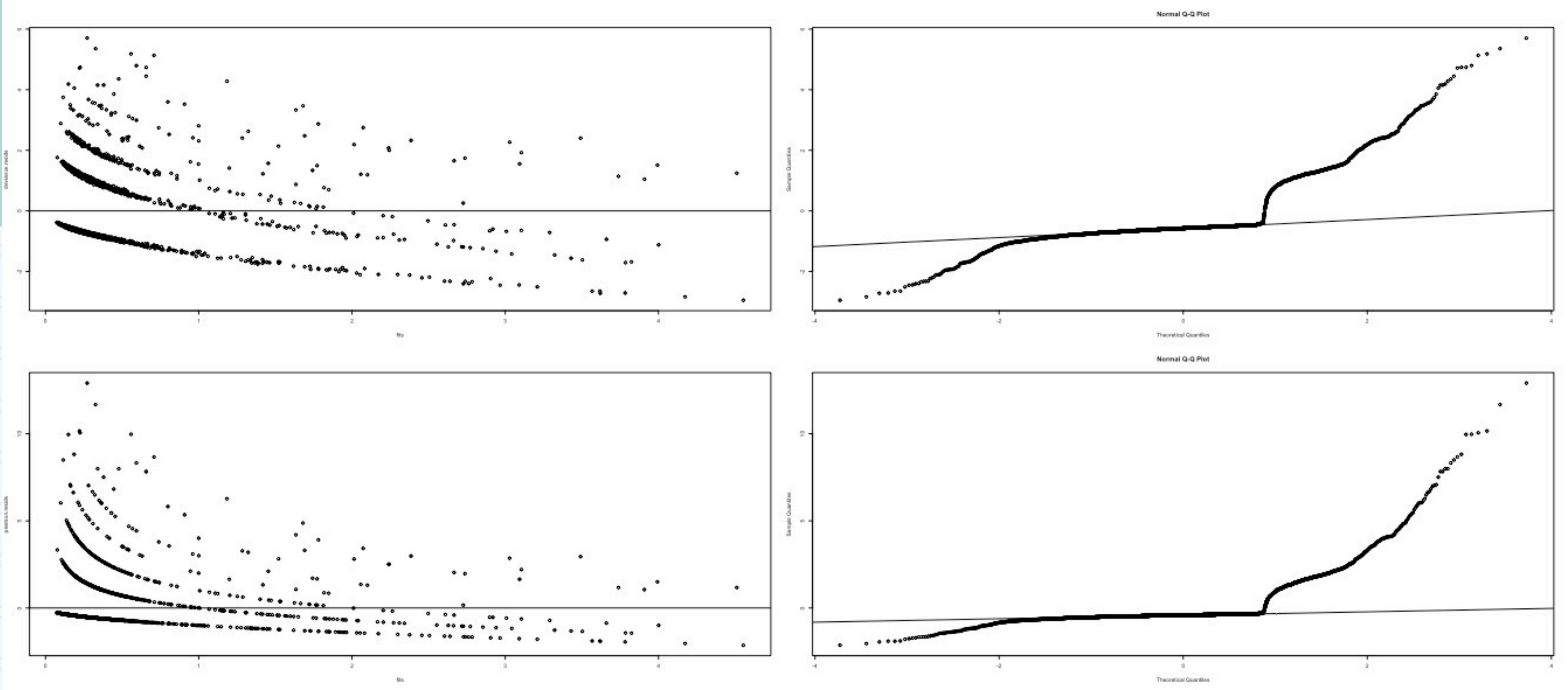
Poisson Model



Reduced Poisson Model

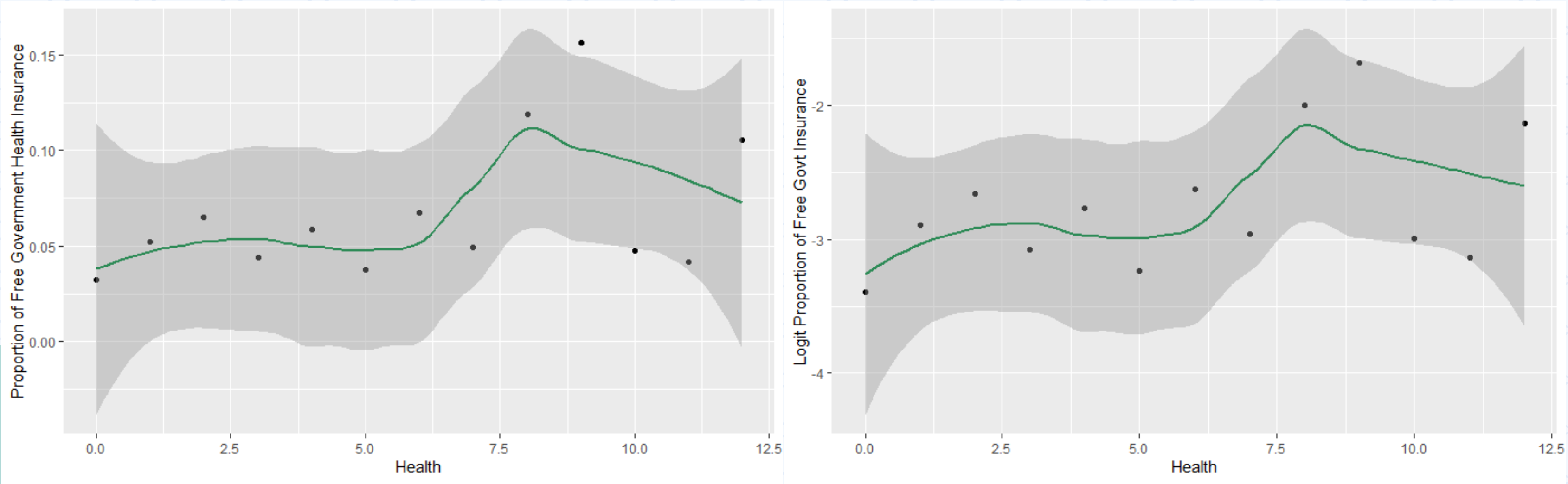


Reduced quasipoisson Model

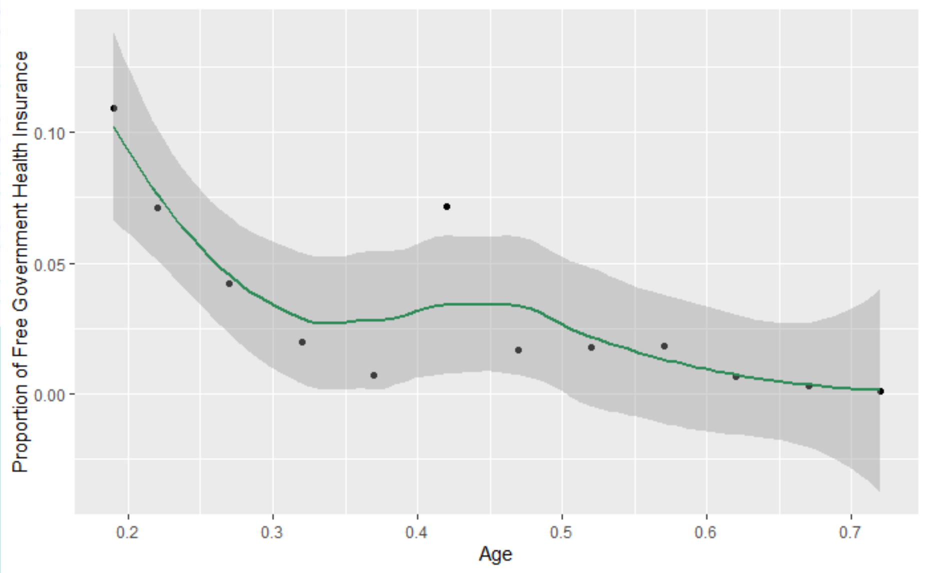


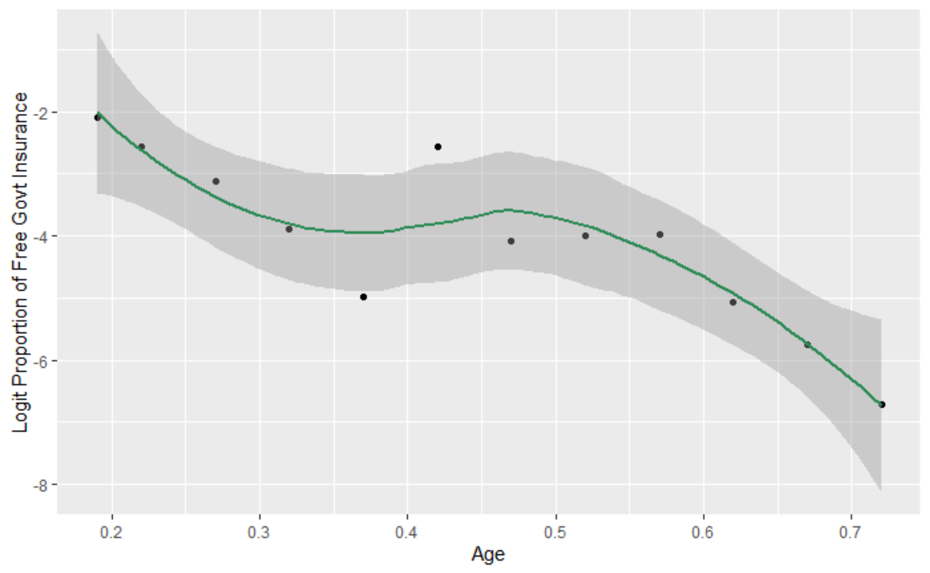
Logistic Regression Model

Logistic Regression

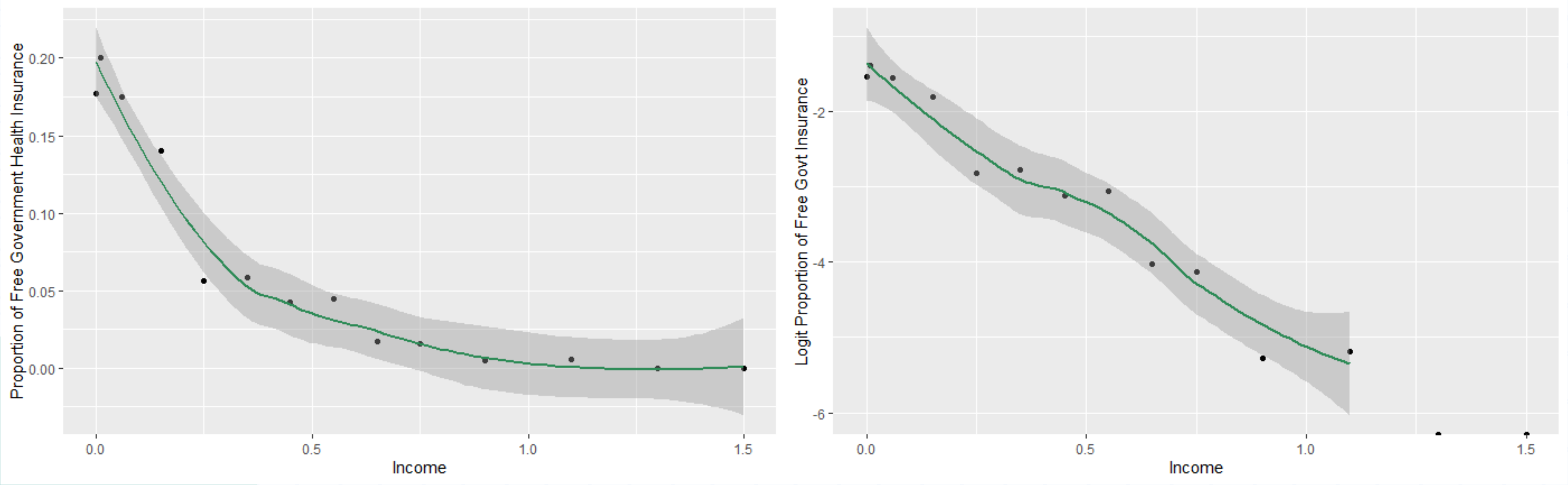


Logistic Regression





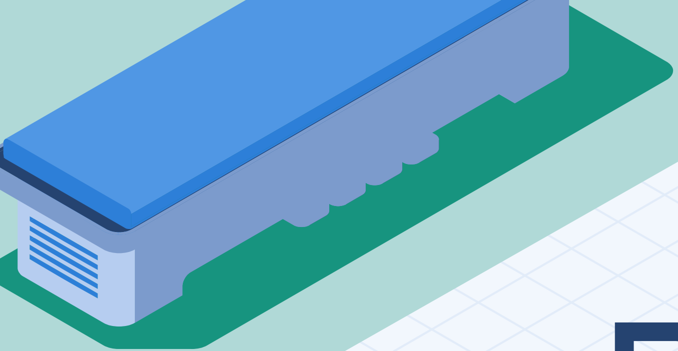
Logistic Regression



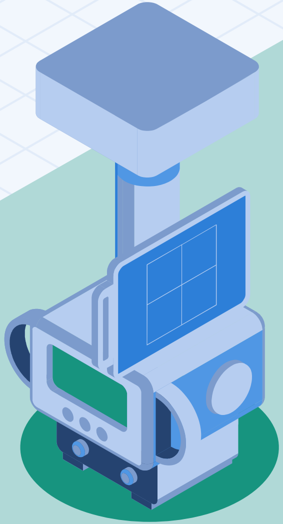
Logistic Regression

```
Call:
glm(formula = freepoor ~ income, family = "binomial", data = DoctorVisits)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6052  -0.3977  -0.2357  -0.1268   3.3281

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6045     0.1233  -13.01   <2e-16 ***
income       -3.5725     0.3282  -10.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1833.8  on 5189  degrees of freedom
Residual deviance: 1652.7  on 5188  degrees of freedom
AIC: 1656.7

Number of Fisher Scoring iterations: 7
```

$$e^{\hat{B}_1} = 0.028$$



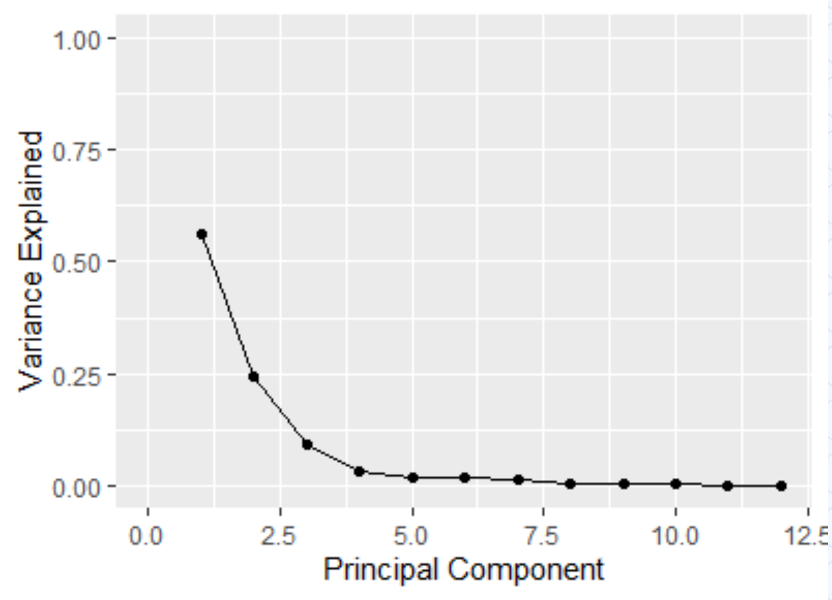
Principal Component Analysis



Principal Component Analysis

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	3.0549	2.0121	1.25296	0.7110	0.58733	0.53288	0.4685	0.33089	0.3105	0.25677	0.18186	0.13856
Proportion of Variance	0.5612	0.2435	0.09441	0.0304	0.02074	0.01708	0.0132	0.00658	0.0058	0.00396	0.00199	0.00115
Cumulative Proportion	0.5612	0.8047	0.89909	0.9295	0.95023	0.96731	0.9805	0.98709	0.9929	0.99686	0.99885	1.00000

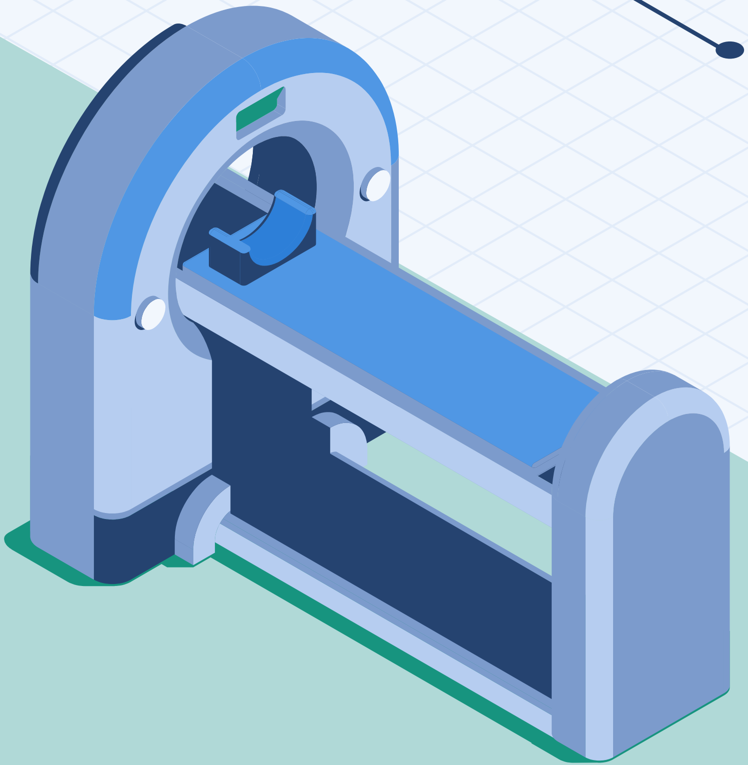


- PC1 alone explains over 56% of the variation in the dataset
- PC1 through PC5 explain 95.02% of the variation in the dataset
- The scree plot does not show an explicit elbow as a cutoff point

Principal Component Analysis

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
visits	-0.1194663436	-0.0102635660	0.08760976	-0.981402194	0.091781669	0.06608277	0.032320845	-0.030056058	0.000107392	0.001817089	-0.0057854249	-0.0045957496
gender	-0.0106555384	0.0147818904	0.05881913	-0.076062110	-0.356031357	-0.62439551	-0.552787701	-0.356642964	-0.197357245	0.041858729	-0.0052280572	-0.0172316718
age	-0.0072508914	0.0004944458	0.03924577	-0.029140372	-0.149556630	-0.09568960	0.008555681	0.309960638	-0.084294002	-0.087285078	-0.1009330699	0.9192244873
income	0.0095807424	-0.0155973337	-0.04102244	0.041818673	0.373933302	0.03162882	0.120624497	-0.149256708	-0.882736275	-0.066516765	-0.1866945418	0.0087347933
illness	-0.1641574477	0.2710898654	0.92858799	0.115264848	0.103425274	0.08318614	-0.043507700	-0.047121948	0.008947543	-0.046619029	-0.0041684623	0.0002849057
reduced	-0.9063915583	-0.4039723925	-0.05432020	0.108676563	-0.005604531	-0.01492437	0.008548157	-0.007223474	0.005186363	-0.009491390	-0.0001624671	-0.0017208099
health	-0.3687111903	0.8728954755	-0.31805558	0.006082140	-0.002856182	-0.01997792	0.017221603	0.005629343	-0.004707777	-0.010897502	0.0032841970	0.0046853504
private	0.0072713088	-0.0103933043	-0.01350832	0.001279423	0.618668447	-0.51167991	-0.200009827	0.479395425	0.174534976	-0.145568816	-0.1347351933	-0.1243294226
freepoor	-0.0003916768	0.0053365898	-0.01063745	0.012169331	-0.028340279	0.06362150	0.006445899	-0.174180396	0.205723214	0.194846912	-0.9403044764	-0.0043617032
freerepat	-0.0149313396	0.0117931013	0.06961109	-0.053876314	-0.533850026	0.04749895	0.011092990	0.578602662	-0.258029391	-0.352627361	-0.2170507425	-0.3629011910
nchronic	-0.0038777872	0.0175076952	0.11582646	-0.016051215	-0.152427636	-0.53933529	0.736653231	0.038162957	-0.029534410	0.347432028	0.0305344535	-0.0725275961
lchronic	-0.0271590179	0.0148531756	0.02643808	-0.004767052	-0.010212177	0.17047557	-0.306053198	0.393508178	-0.185354337	0.824383730	0.0670761734	-0.0466116140

PC1: Income and private health insurance (wealth)
PC2: Gender, number of recent illnesses, health score, free health insurance, and chronic condition (general wellness)
PC3: Number of doctor visits, gender, age, number of recent illnesses, free health insurance, and chronic condition (general wellness)
PC4: Income, number of recent illnesses, number of reduced activity days, and free health insurance
PC5: Number of doctor visits, income, number of recent illnesses, and private health insurance
PC6-PC12: Not significant



Thank You!

Questions?