The Bill of Progress

Case Study for Linear Models

Bram Otten, Gergely Talyigás, Joost Boelema Robertus, and Rohit Nekkanti

24 November 2020

Contents

1 Introduction

In this case study we model the relationship between demographic indicators of the past to economic outcomes in the present. While it would be important to discuss the selection of indicators and measurements if we would take our results to imply anything, here we care only about the statistical modelling.

Every author mentioned above contributed equally to this case study.

We merge two seperate data-sets that contain our desired variables and are split by county of the United States of America – 10 demographic indicators in 1992,[1] and economic outcomes in 2017.[2] Table 1 provides a brief description of our variables. The 1992 data-set also contains the average incomes and amounts of savings per county, which are not considered to make the model at least *potentially* interesting.

Variable	Description
`clinton`	Percentage of population voting for Bill Clinton
`age92`	Median age of population
`poverty92`	Percentage of population living in poverty[3]
`veterans92`	Percentage of population that is a veteran
`female92`	Percentage of population that is female
`density92`	Mean number of people per square mile
`nursing92`	Percentage of population living in nursing homes
`crime92`	Mean reported number of crimes per 10000 inhabitants
`income17`	Median income in 2017

Table 1: Description of variables used in this case study

The merging process is inherently imperfect, as county borders and names have changed in the intervening 25 years. The data-set from 2017 recognises 3242 counties, whereas there are only 2704 in 1992; $n = 2647$ of these match after stripping county names of their special characters and bits like 'St.' or 'County'.[4]

[1] From `http://users.stat.ufl.edu/~winner/data/clinton1.dat`, with description at `http://users.stat.ufl.edu/~winner/data/clinton1.txt`

[2] From `https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/`.

[3] According to the U.S Census Bureau, the average poverty threshold in 1992 was $7143 dollars if you were living alone and $14335 dollars for a family of four.

[4] A comma seperated file should be provided with this report or can be found at `https://pastebin.com/dRz14Efq`.

2 Data Description

Table 2 summarises some descriptive statistics of the considered variables, and figure 1 contains their histograms.

Variable	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
`clinton` (%)	9.55	32.73	38.8	39.45	45.39	84.64
`age92`	20	32.50	34.44	34.51	36.50	55.40
`poverty92` (%)	1.90	11.40	14.70	15.91	19.10	49.90
`veteran92` (%)	2.78	9.89	11.48	11.45	13.05	27.29
`female92` (%)	37.53	50.38	51.12	50.99	51.84	55.39
`density92`	0.4	19.1	43.2	179.3	104.0	32,360
`nursing92` (%)	0.08	5.17	7.89	9.50	11.96	59.22
`crime92`	0.0	150	260	302.8	409	2,792
`income17`	25344	43052	49446	51498	56989	136,191

Table 2: Descriptive statistics of variables used in this case study

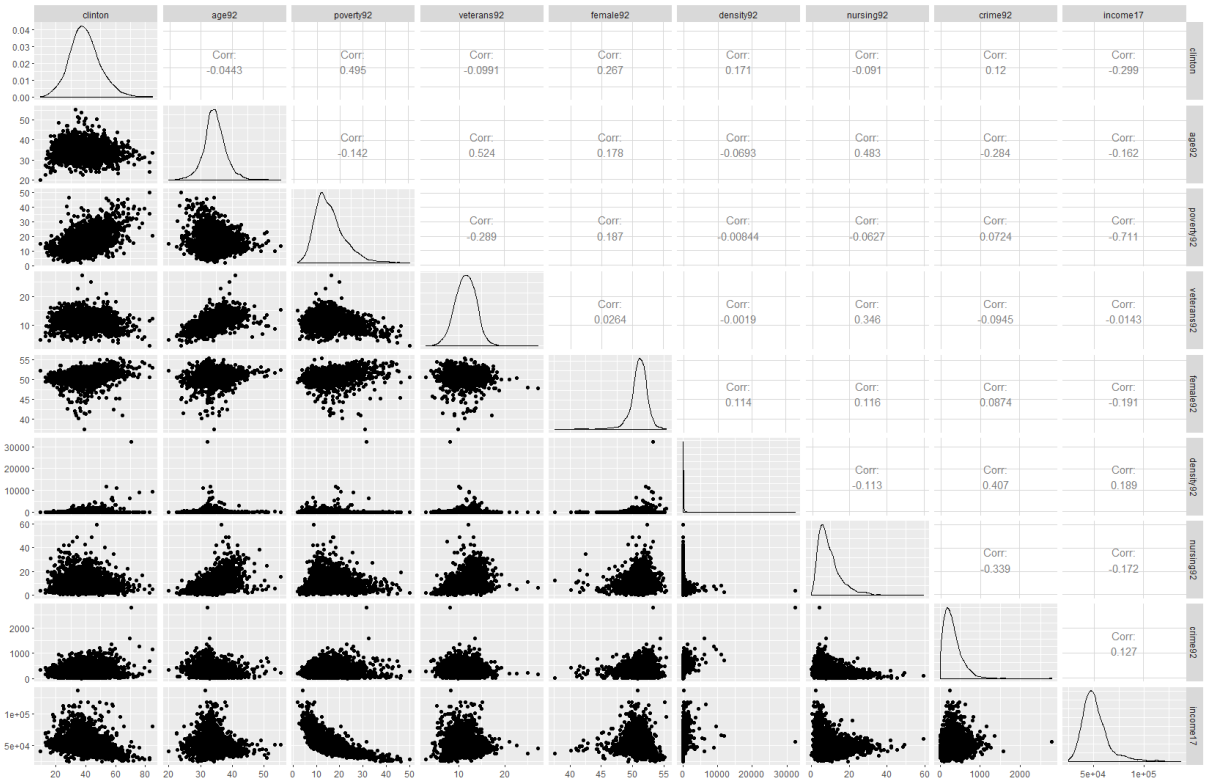


Figure 1: Correlation matrix and histograms of variables used in this case study

A high degree of a linear relationship between independent variables[5] is called multicollinearity, and could be problematic for coefficient estimation. To visualise the phenomenon, a correlation matrix of all variables is also shown in figure 1. The highest correlation between (supposedly) independent variables is between `age92` and `veteran92` ($r = 0.52$). This relationship makes intuitive sense – one expects veterans to be of older age. Others relatively high correlations are between `clinton92` and `poverty92` and between `nursing` and `age`. To quantify the severity of multicollinearity in our model we give the variance inflation factor (VIF) values for our independent variables in table 3. VIF is an index that measures how much the variance of our estimated regression coefficient is increased because of multicollinearity. These values are far below commonly proposed threshold such as 4 or even 10 and thus there is no cause for worry.

[5] In this case study, we may at times refer to our independent variables as explanatory variables or regressors, and to our dependent variable the response.

Variable	*VIF*
`Clinton (%)`	1.44
`Age`	1.72
`Poverty (%)`	1.48
`Veteran (%)`	1.53
`Female (%)`	1.16
`Population density`	1.25
`Nursing (%)`	1.44
`Crime rate`	1.41

Table 3: VIF table for independent variables

3 Linear Models

3.1 Initial Model

The linear model can be written as $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where

- $\boldsymbol{y}$ is the $n \times 1$ vector representing outcomes;
- $\boldsymbol{X}$ the $n \times (k+1)$ matrix of regressors with an initial columns of 1s
- $\boldsymbol{\beta}$ is a $(k+1) \times 1$ vector of parameters to be estimated; and
- $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I}_n)$ (i.e., also an $n \times 1$ vector),

and where we will find a $(k+1) \times 1$ vector $\boldsymbol{b}$ with fitted coefficients for an $n \times 1$ vector $\hat{\boldsymbol{y}}$ of fitted values and $n \times 1$ vector $\boldsymbol{e}$ of residuals – i.e., $\hat{\boldsymbol{y}} = \boldsymbol{Xb} + \boldsymbol{e}$ with a minimised $\boldsymbol{e'e}$.

Based mostly on figure 1, we should not expect `income17` to be a simple linear combination of all of our explanatory variables. However, this is where we start our journey towards a useful model – or in any case, where we can diagnose *a* model.

We first introduce a few metrics that we may refer to at later points.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})}$$

denotes the proportion of variance of $\boldsymbol{y}$ that is explained by the model – higher is in principle better.

$$R^2_{\text{adj}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2 (n-1)}{\sum(y_i - \bar{y})^2 (n-(k+1))}$$

is very similar to R^2 but penalizes a large amount of parameters.

$$\text{Root mean squared error (RMSE)} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}}$$

denotes the standard deviation of the residuals – it is the average error of an element of $\hat{\boldsymbol{y}}$. A low value is desirable.

$$\text{Residual standard error (RSE)} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-(k+1)}},$$

where k refers to the number of fitted parameters, is a variant of the RMSE adjusted for the number of predictors in the model. The lower the RSE, the better the model. In practice, the difference between RMSE and RSE is minimal, particularly for large n.

$$\text{Mean absolute error (MAE)} = \frac{\sum |y_i - \hat{y}_i|}{n}$$

is comparable to RMSE but less sensitive to very wrong fitted values.

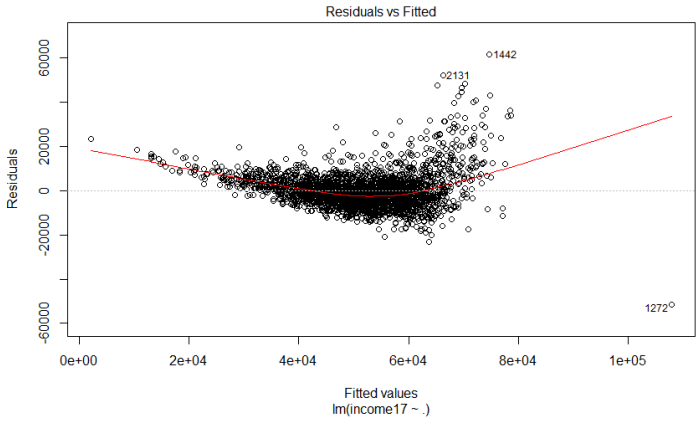
Anyway, in this first model all explanatory variables are fit to the response variable using a linear regression model – i.e., `fm1 = lm(income17 ~ ., data)`. The summary of this linear model is shown in table 4.

	RSE	df	Multiple R^2	R^2_{adj}
	8027	2368	0.6252	0.6241
Coefficients	*Estimate*	*Standard error*	*t*	*significance*
(constant)	1.190e+05	5.617e+03	21.193	<2e-16
clinton	6.211e+01	1.838e+01	3.378	7.4e-04
age	-5.095e+02	5.81e+01	-8.777	<2e-16
poverty	-1.591e+03	2.905e+01	-54.77	<2e-16
density	2.119e+00	2.002e-01	10.582	<2e-16
veteran	-7.803e+02	8.340e+01	-9.356	<2e-16
female	-3.654e+02	1.146e+02	-3.110	1.890e-3
crime	3.070e+00	8.641e-01	3.553	3.880e-4
nursing	-1.358e+02	2.966e+01	-4.578	4.92e-06

Table 4: Summary of `fm1 = lm(income17 ~ ., data)`.

Linear regression makes several assumptions about the data – it is necessary to diagnose these to check if the model works well.

One important assumption is that the residuals are independent of each other, as well as normally distributed with mean 0 and some constant variance. To check whether this assumption holds, we check the Residual vs. Fitted plot of the given model. Residual vs. Fitted plots show whether the residual have non-linear patterns – these would result from a non-linear relationship between predictor variables and the outcome variable. If the residuals are spread equally around 0 without distinct patterns, that is a good indication that non-linear relationships do not exist.

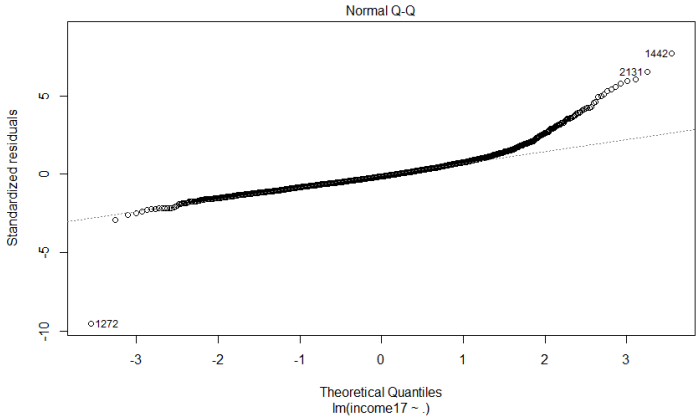


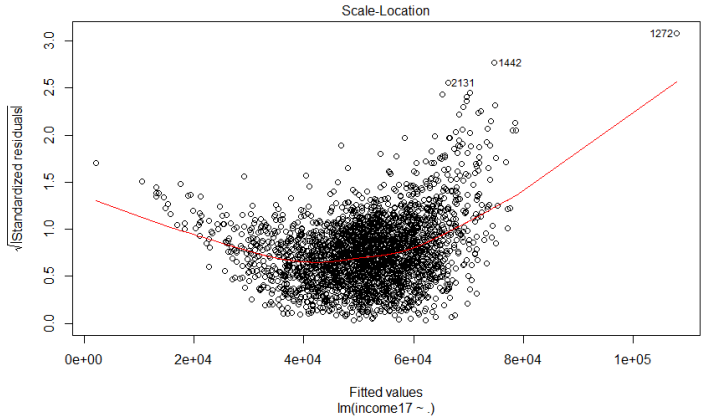
However, the residuals of this model are spread out in a parabolic way, indicating that the model could not explain a non-linear relationship.

To check whether the residuals conform to the normality assumption, we inspect the normal QQ-plot on the residuals. Residuals following a straight line indicate they are normally distributed (a bit more technically, that the emprical cdf of the residuals is somewhat similar to a normal cdf).

The high residuals seem a bit *too* high – the tails are a bit too fat to say the normality assumption is completely met.

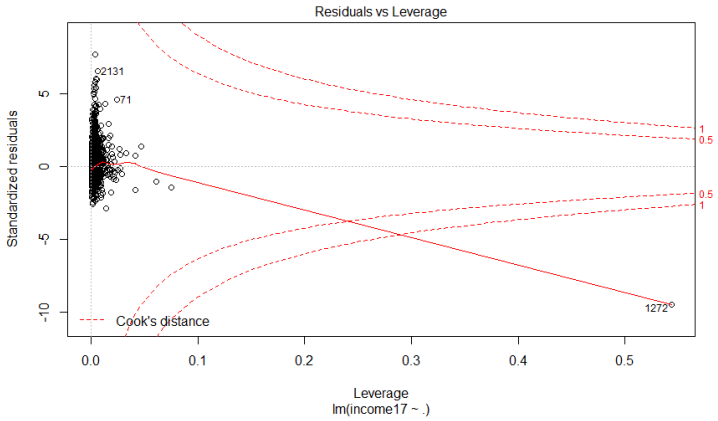
The next assumption is that the residuals are assumed to have constant variance (homoscedasticity) across the range. To check for homoscedasticity, we inspect the Scale-Location plots. Scale-location plots, also known as the Spread-location plot, show if residuals are spread equally along with the ranges of the predictors. The assumption of equal variance (homoscedasticity) holds if the residuals are spread equally along the horizontal line.



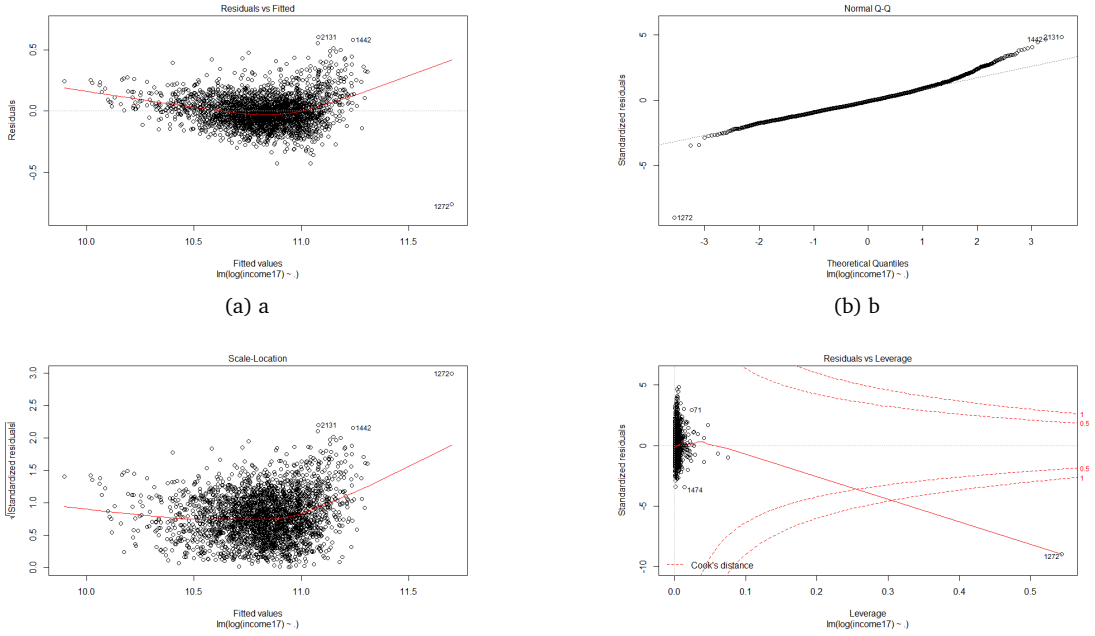


Here, the spread of the residuals increases along with fitted value, implying that the assumption of constant variance (homoscedasticity) does not hold.

Next, we look into the presence of influential values in data that can be either/both outliers or high-leverage points. Residual vs. Leverage plots aid in finding influential cases in our dataset. It is necessary to note that not all outliers are influential in linear regression – even an extreme outlier might not be influential in determining a regression model; that would require high leverage. Conversely, there can be cases where very influential points can appear to be within a reasonable range of values; these will have that high leverage. Therefore, instances lying on the upper or lower right corners of Residual vs. Leverage plots can be influential.



The instance 1272 seems to be an outlier and a high-leverage point – it has a lot of influence on the model. Carefully examining the point from the dataset indicates that this instance represents values from Kings



(a) a (b) b

(c) c (d) d

Figure 2: Caption

County, New York (also known as Brooklyn). Its exceptionally high population density and crime rate may have influenced the results. Although removing this point has a noticeable impact on the model, we do not exclude it from further analysis since we see no reason to doubt the values of the demographic indicators particularly in this case.

3.2 Modified Full Models

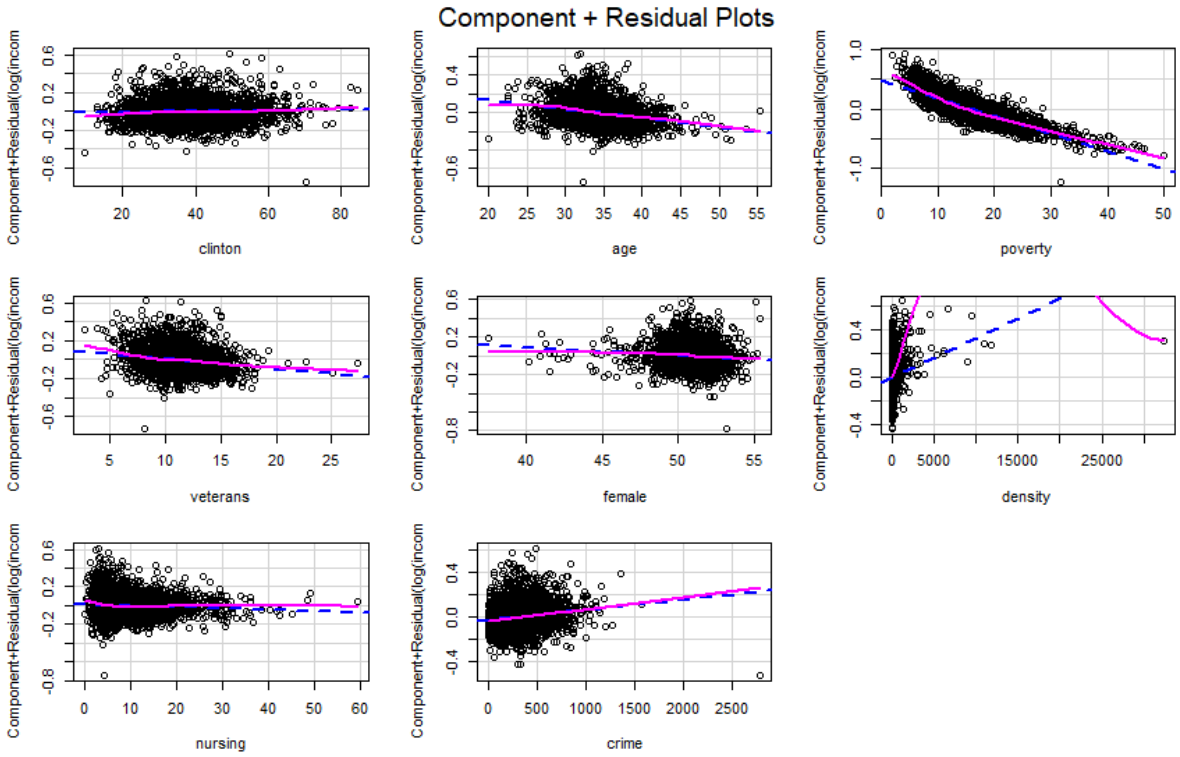
	RSE	df	Multiple R^2	R^2_{adj}
	0.1253	2638	0.7135	0.7126
Coefficients	*Estimate*	*Standard error*	*t*	*significance*
(constant)	1.216e+01	8.771e-02	138.600	<2e-16
clinton	5.177e-04	2.871e-04	1.803	0.07147
age	-9.906e-03	9.064e-04	-10.929	<2e-16
poverty	-3.004e-02	4.536e-04	-66.216	<2e-16
density	3.306e-05	3.127e-06	10.572	<2e-16
veteran	-1.052e-02	1.302e-03	-8.074	1.02e-06
female	-8.537e-03	1.789e-03	-4.771	1.93e-06
crime	9.066e-05	1.349e-05	6.719	2.23e-11
nursing	-1.483e-03	4.632e-04	-3.201	0.00138

Table 5: Summary of `fm2 = lm(log(income17) ~ ., data)`.

Taking a log on our dependent variable (for `fm2 = lm(log(income17) ~ ., data)`) results in approximately a .1 increase of R^2. We again look at some diagnostic plots in figure 2, and see they do not appear very different from those of `fm1` discussed before.

We look at Component + Residual plots to understand partial non-linear effects from each exploratory variable. From the standard residual plots, it is evident that the model can not explain the non-linear relationships between the response variable and other exploratory variables adequately. The Component + Residual plots define the marginal relationship ('ignoring' the other independent variables) between the variables instead of the partial relationships('controlling for' the other independent variables).

It is apparent from the Component + Residual plots that the poverty and density have a noticeable contribution towards the non-linearity the model does not explain. Thus transforming poverty using quadratic transformation and density using the log transformation might result in the model that better



captures these nonlinear partial relationships.

	RSE	df	Multiple R^2	R^2_{adj}
	0.1196	2637	0.7393	0.7384
Coefficients	*Estimate*	*Standard error*	*t*	*significance*
(constant)	1.235e+01	8.497e-02	145.335	<2e-16
clinton	-2.180e-04	2.885e-04	-0.756	0.450
age	-6.475e-03	8.812e-04	-7.348	2.66e-03
poverty	-4.897e-02	1.444e-03	-33.910	<2e-16
I(povertyˆ2)	5.418e-04	3.480e-05	15.569	<2e-16
log(density)	2.251e-02	2.208e-03	10.192	<2e-16
veteran	-6.209e-02	1.265e-03	-4.909	9.69e-07
female	-1.421e-03	1.800e-03	-7.896	4.19e-15
crime rate	1.205e-04	1.305e-05	9.228	<2e-16
nursing	-3.182e-04	4.578e-04	-0.695	0.450

Table 6: Summary of `fm3 <- lm(log(income17) ~ poverty92 + I(poverty92 ^ 2) + log(density92) + age92 + veterans92 + crime92 + female92+clinton+nursing, data)`.

Inclusion of the quadratic term for poverty and applying log transformation on the density (for `fm3 <- lm(log(income17) ~ poverty92 + I(poverty92 ^ 2) + log(density92) + age92 + veterans92 + crime92 + female92, data)`) results in a further increase of R^2 by approximately .02. Although this might not be much larger, looking at the daignostic plots again for this model is cause for a bit more optimism – take a look at figure 3.

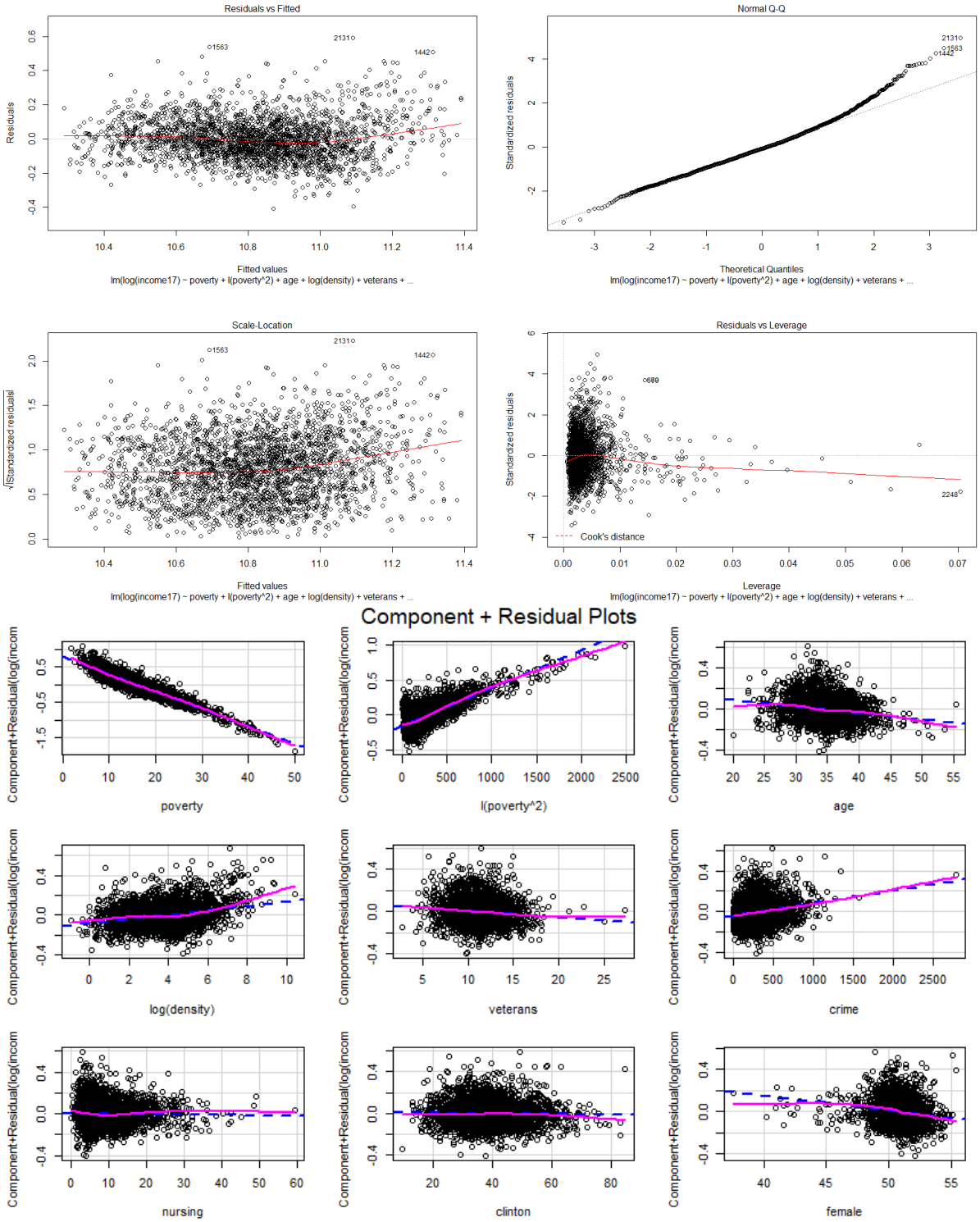


Figure 3: Some diagnostics for `fm3 <- lm(log(income17) ~ poverty92 + I(poverty92 ^ 2) + log(density92) + age92 + veterans92 + crime92 + female92 + clinton + nursing, data)`

3.3 Smaller Models

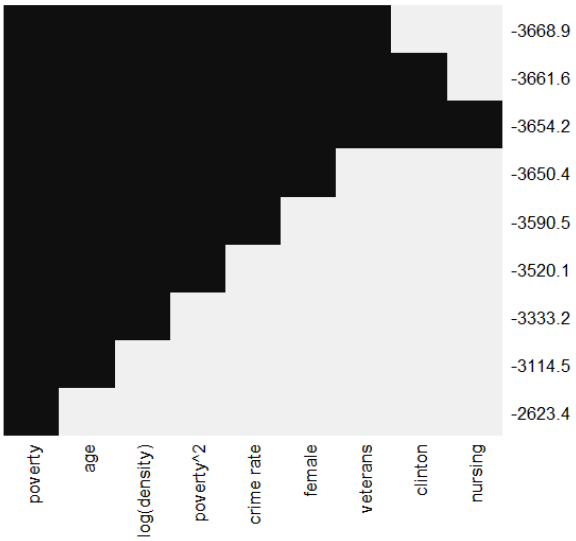


Figure 4: Best models for 2-10 parameters based on BIC criteria. On the vertical axis the BIC values can be seen, and it is sorted from lowest (best) to highest BIC values. The black squares indicates which variables (horizontal axis) are included in each model, and the variables are sorted based on how many models are they appear in.

The goal of model selection is to find the best model from a set of models. If the models differ only in their exploratory variables, we call this approach variable selection. Our motive for variable selection is twofold – we want to increase prediction power or help interpretation by finding the 'true' underlying model without distraction by unnecessary variables.

Model selection methods can be divided into two main types: stepwise testing approaches, and criterion based approaches. In the first case, in each step we add (forward) or remove (backward) a variable based on hypothesis tests. There are some disadvantages of these methods: they often miss the optimal model, and p-values are treated too literally with multiple tests (simultaneous inference). With criterion based procedures, we can usually test all possible models or try to find it with optimisation algorithms.

We use the Bayesian information criterion (BIC) for variable selection. BIC belongs to family of penalised model-fit statistics alongside with Akaike information criterion (AIC). Here, we prefered BIC because it penalises free parameters more then AIC does. The BIC defines as

$$-2 \log_e L(\hat{\theta}) + s \log_e n,$$

where $L(\hat{\theta})$ is the maximized likelihood, s is the number of parameters, and n is the number of data-points as above. In our case – that of a linear model – the equation simplifies to

$$-2n \log_e \hat{\sigma}_\epsilon^2 + s \log_e n$$

where $\hat{\sigma}_\epsilon^2$ is maximum likelihood estimator of the error variance. The absolute value of BIC is not informative but we can compare models with their differences. Given the definition, the log-Bayes factor for any pair of models M_1 and M_2 is approximated by the difference in theirs BICs:

$$2 \log_e \frac{p(y|M_2)}{p(y|M_1)} \approx \text{BIC}_{M_1} - \text{BIC}_{M_2}.$$

In other words the difference of BIC values express the relative support in the data for model M_2 compared to M_1. Table 7 shows the interpretation of difference of BIC values

In the last model we had 9 explanatory variables (including `poverty92^2`), so it is feasible to calculate all the 2^9 linear models and the corresponding BIC values. Figure 4 shows the best models for all numbers of parameters (intercept always included). The lowest BIC value corresponds to the model without `clinton` and `nursing92`. Including `clinton` in the model increases the BIC value by 7.3, so we have strong evidence against this (worth mentioning, this is the second best model). Furthermore, we also have strong evidence

Difference in BIC	Bayes Factor	$p(M_2\|y)$	Evidence for M_2
0-2	1-3	.50-.75	"Weak"
2-6	3-20	.75-.95	"Positive"
6-10	20-150	.95-.99	"Strong"
>10	>150	>.99	"Conclusive"

Table 7: Relative support for M_2 in BIC differences according to Kass and Raftery (1995)

	RSE	df	Multiple R^2	R^2_{adj}
	0.1196	2639	0.7392	0.7385
Coefficients	*Estimate*	*Standard error*	*t*	*significance*
(constant)	1.236e+01	8.310e-02	148.788	<2e-16
age	-6.672e-03	8.511e-04	-7.839	6.53e-15
poverty	-4.910e-02	1.418e-03	-34.636	<2e-16
I(povertyˆ2)	5.405e-04	3.480e-05	15.552	<2e-16
log(density)	2.233e-02	1.990e-03	11.220	<2e-16
veteran	-6.417e-03	1.248e-03	-5.141	2.94e-07
female	-1.451e-03	1.770e-03	-8.201	3.69e-16
crime rate	1.224e-04	1.291e-05	9.483	<2e-16

Table 8: Summary of `fm4 <- lm(log(income17) ~ poverty92 + I(poverty92 ^ 2) + log(density92) + age92 + veterans92 + crime92 + female92, data)`.

for including all the other variables. We can also see the predictive power of these variables – `poverty` has by far the most, followed by `age92` and `log(density92)`.

Our final model shall be `fm4 <- lm(log(income17) ~ poverty92 + I(poverty92 ^ 2) + log(density92) + age92 + veterans92 + crime92 + female92, data)`, and its summary is shown in table 8.

To check how the models generalise to unseen data, we have split our dataset into a training and a test set (ratio 80:20). After using the training set to fit all the three regression models referenced in the previous section, the value of some metrics for the test set is given in table 9. Curiously, reducing the number of variables (as done in going from `fm3` to `fm4`) does not seem to improve generalisation much.

Model	mean absolute error (MAE)
`fm1`	5711 ± 182
`fm2`	5210 ± 256
`fm3`	4922 ± 172
`fm4`	4916 ± 172

Table 9: Model comparison on a random 20% of data when fitting on other 80% of data (mean ± sd over 100 runs)

4 Discussion

In our final model (`fm4`) we used six variables from 1992 to predict the median income in 2017 by county.

- `poverty92` explained the biggest part of the variance, but unfortunately it is the hardest to interpret since we also used its square. The effect of 1% higher `poverty92` is a 4.11% decrease in median `income17` at the median of `poverty92` (which is 14.7%) and the coefficient shrinks by 0.054% with every 1% of `poverty92`.
- An extra year of `age92` in a county decreases `income17` by .67%.
- One unit of the $\log_e$ `density92` of a county increases `income17` by 2.23% which means doubling population density causes a 1.55% increase if everything else is held constant.
- Increasing `veterans92` by 1% decrease `income17` by a factor of 1.4%.
- Increasing `female92` by 1% decrease `income17` by a factor of 1.4%.
- `crime92` may have the most surprising coefficient because it is positive and for every committed crime by 10,000 people the `income17` 25 years later grows by a factor of .012%.

We had to reject our initial hypothesis that the percentage of votes on Clinton in 1992 influences the long-term economic outcomes of a county based on this data. On the whole though, the demographic indicators we worked with – and, of course, in particular `poverty92` – were surprisingly predictive of economic outcomes 25 years later with an $R^2_{\text{adj}} \approx .74$

One of the main limitations of this study is that we merged two data sets imperfectly. This issue is hard to work around, as many new counties have been created been 1992 and 2017. We must hope the effects of this are random and independent, and do not bias our models too much. We have made some errors as well – the District of Columbia is in $\boldsymbol{x}$ twice, and `income17` actually refers to median *household* income. The former is easy to alleviate (but would require new plots), and the latter is not necessarily an issue for statistical purposes.

Furthermore, we are using indicators that are very simplistic and in no way take into account policy decisions or economic events like factories closing down. This is common practice and can be useful, but if, e.g., Apple decides to build some 'campus' in a particular county, that will have a far larger economic impact than the large contingent of retirees that has flocked around the golf course for more than 25 years.

Besides potential problems, we have potential improvements. Of course, the 1992 income level that we chose to disregard comes to mind. But we can also be a bit more creative and consider taking into account population numbers per county. These numbers could be considered a proxy for 'reliability' and be used as such in weighted least squares or something similar. Something else to consider is a classification of counties into rural and urban, although the exact decision boundary for this classification would be a bit arbitrary or at least subjective. It might also be interesting to use all available data from 1992, and predict something like income growth. And in all of these cases, the state a county is in could be considered an independent variable (as factor) as well.

Overall, we have modelled this hard problem better than expected. We admit that we hoped for a greater influence of the `clinton` variable, but this outcome should lead to no less armchair political philosophy. Even when explicitly not considering income and savings, and by looking at median incomes instead of averages or savings 25 years later, we have still only managed to produce a piece of Soviet propaganda – or the advice to move out of impoverished counties. Hard work pays off.

References

Kass, R. E. and A. E. Raftery (1995). 'Bayes Factors'. In: *Journal of the American Statistical Association* 90.430, pages 773–795. DOI: 10.1080/01621459.1995.10476572.