SF2930 - Regression analysis

KTH Royal Institute of Technology, Stockholm

Lecture 13 – Models with a binary response & an introduction to logistic regression. (MPV 13.1-13.2, 4-5)

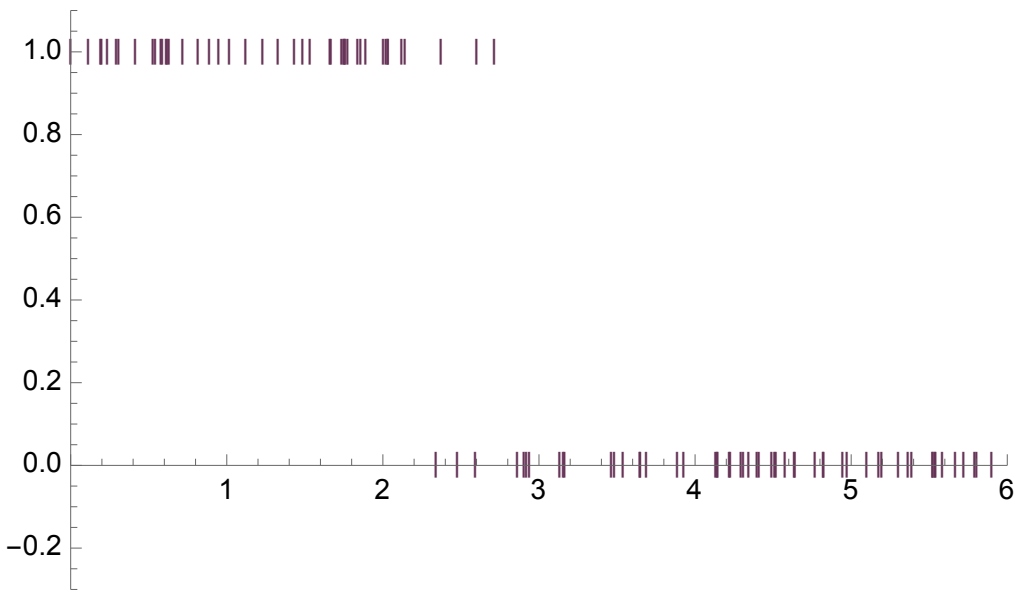
February 18, 2022

Todays lecture

- Logistic regression

Motivation

In previous lectures we have always made assumptions on the model. In particular, in most lectures, we have assumed that the response has been continuous with a normal distribution (this follows from the error having this form). When this is not the case, it can sometimes be remedied by a transform of the response variable, but in some cases it makes more sense to construct a model without this assumption.

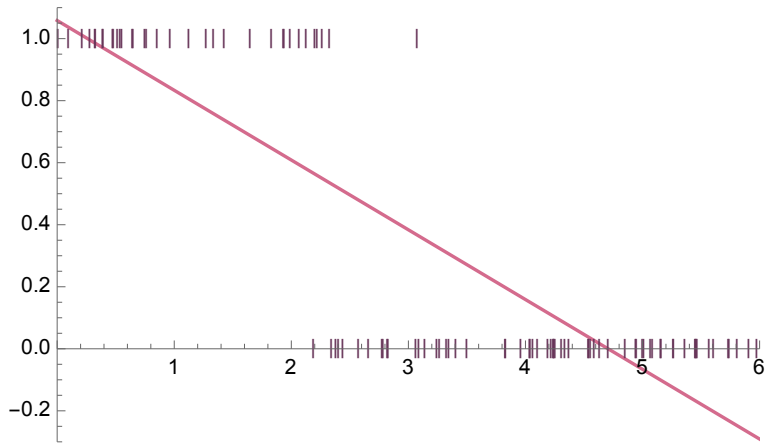


The logistic model

We want a model where the response y_j is binary, while the regressors $\mathbf{x}_j$ are allowed to be continuous.

First attempt at a model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$$



- Since $y_i \in \{0, 1\}$, we must have $y_i \sim Bernoulli(\pi_i)$ for some $\pi_i \in [0, 1]$.
- We must have $\varepsilon_i \in \{1 - \mathbf{x}_i^T \boldsymbol{\beta}, -\mathbf{x}_i^T \boldsymbol{\beta}\}$. In particular, the error would not be independent of $\mathbf{x}_i$. Moreover, since $y_i \sim Bernoulli(\pi_i)$, $\operatorname{Var} y_i = \pi_i(1 - \pi_i)$, and hence the variance is not generally constant.

This suggests considering another response function than $\mathbf{x}_i^T \boldsymbol{\beta}$.

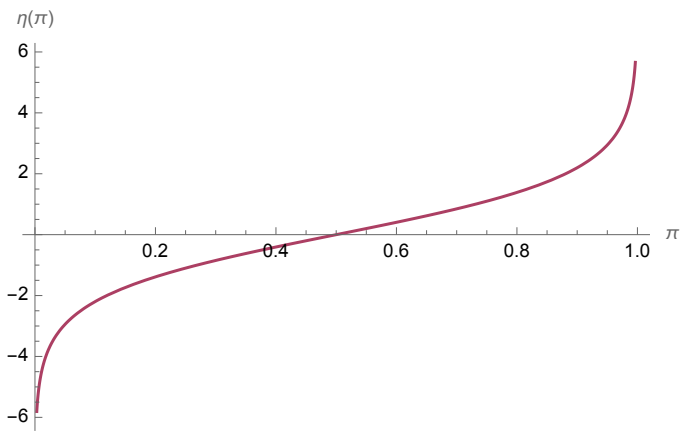
The logistic model

We want a model where the response y_j is binary, while the regressors $\mathbf{x}_j$ are allowed to be continuous.

Idea

We want to apply a function to $\mathbb{E}[\mathbf{y} \mid \mathbf{x}]$ so that the transformed values take their values in all of $\mathbb{R}$. Such a function is called a *link function*.
A link function which is often used when $y_i \in \{0, 1\}$ is the so called *logit transformation, given by*

$$\eta \colon \pi_i = P(y_i = 1 \mid \mathbf{x}_i) \mapsto \log \frac{P(y_i = 1 \mid \mathbf{x}_i)}{P(y_i = 0 \mid \mathbf{x}_i)} = \log \underbrace{\frac{\pi_i}{1 - \pi_i}}_{\text{the odds ratio}} =: \eta_i.$$

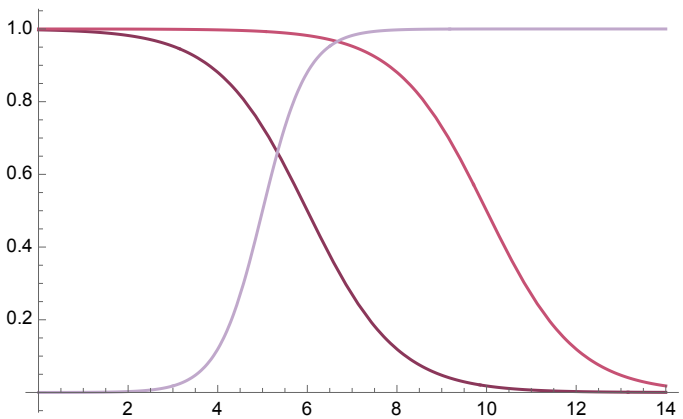


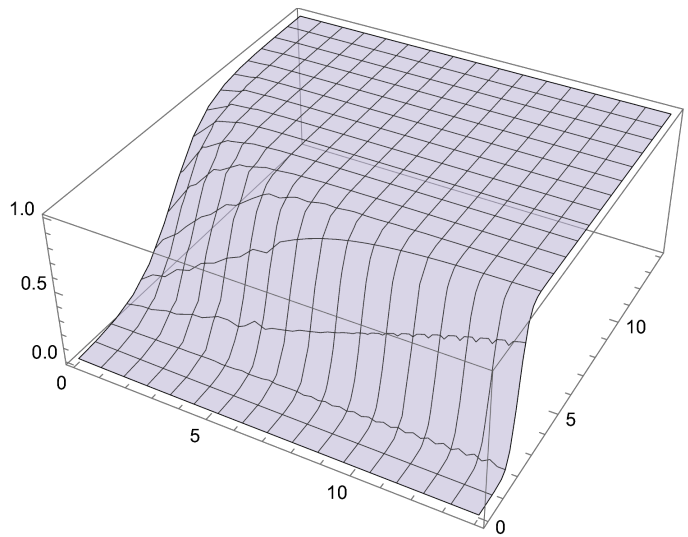
Using the logit transformation, we can define a model $\mathbf{x}_i^T \boldsymbol{\beta} = \eta_i$ which thus models the logarithm of the odds ratio as a linear function.

The logistic model

Note that

$$\mathbf{x}_i^T\boldsymbol{\beta} = \eta_i = \frac{\pi_i}{1-\pi_i} \Leftrightarrow \boldsymbol{\pi} = \underbrace{\frac{\exp(\mathbf{x}^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}^T\boldsymbol{\beta})}}_{\text{the logistic response function}}.$$



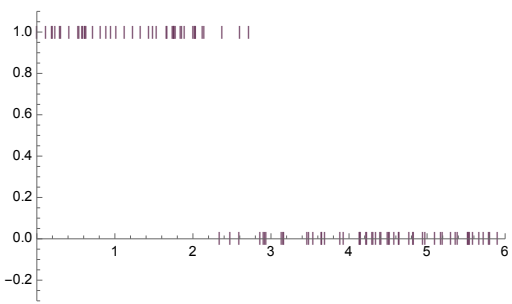


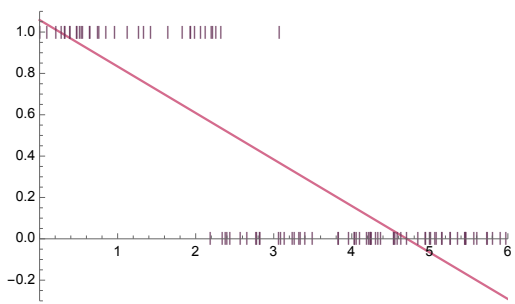
The logistic model

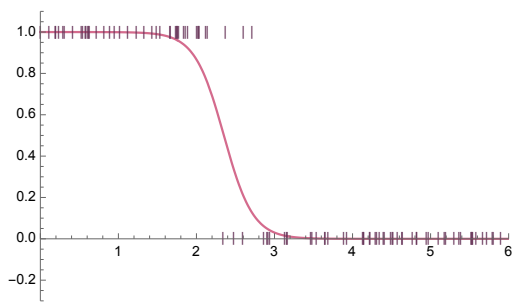
The logistic regression model

$$\boldsymbol{\pi} = \mathbb{E}[\mathbf{y}] = \frac{\exp(\mathbf{x}^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}^T\boldsymbol{\beta})} + \varepsilon,$$

where $y_i \sim Bernoulli(\pi_i)$, and $\pi = E(y) = \frac{\exp(\mathbf{x}^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}^T\boldsymbol{\beta})}$. Note that the logistic model models the probability of observing a one at a certain sample point, and not a $\{0, 1\}$-valued function.







MLE estimation of $\boldsymbol{\beta}$

The log-likelihood function

Assume that the observations are independent. Then

$$L(\mathbf{y}, \boldsymbol{\beta}) = \prod_i f_i(y_i) = \prod_i \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

and thus

$$\log L(\mathbf{y}, \boldsymbol{\beta}) = \sum_i \log \pi_i^{y_i} (1 - \pi_i)^{1-y_i} = \sum_i y_i \underbrace{\log \frac{\pi_i}{1 - \pi_i}}_{\eta_i = X^T \boldsymbol{\beta}(i)} + \sum_i \underbrace{\log(1 - \pi_i)}_{(1+\exp X^T \boldsymbol{\beta}(i))^{-1}} .$$

The binomial log-likelihood function

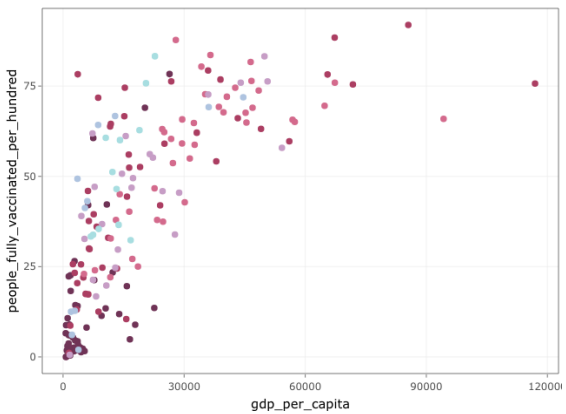
When the regressors are not continuous, the same observation of the regressors could occur multiple times, i.e. we might have $\mathbf{x}_i = \mathbf{x}_j$ (and hence $\pi_i = \pi_j$). We let $(\tilde{\mathbf{x}}_j)$ be the unique observations of the regressors, and let $n_j := \#\{i \colon \mathbf{x}_i = \tilde{\mathbf{x}}_j\}$ and $y^j := \sum_{i \colon \mathbf{x}_i = \tilde{\mathbf{x}}_j} y_i$. Then we can write

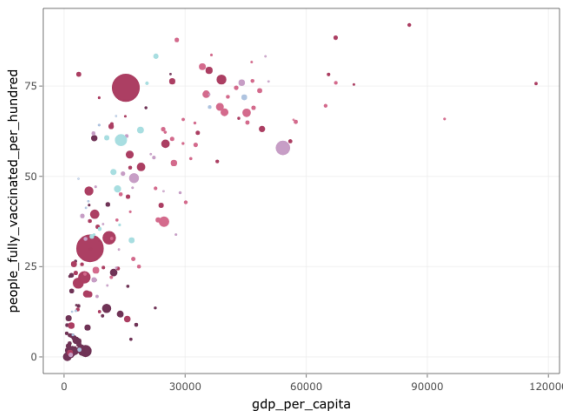
$$\log L(\mathbf{y}, \boldsymbol{\beta}) = \sum_j y^j \log \pi_j + \sum (n_j - y^j) \log(1 - \pi_j)$$

Log likelihood estimates of β

$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{MLE}$ is almost always found numerically.

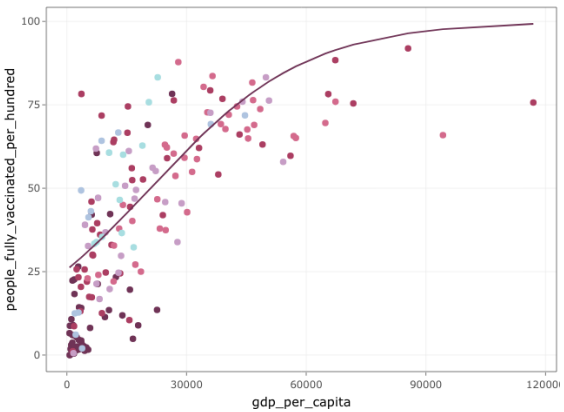
Example

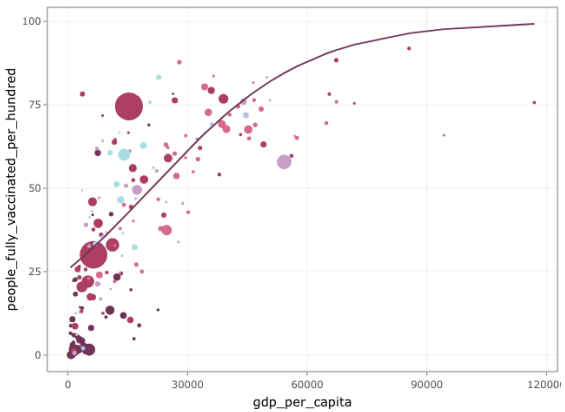


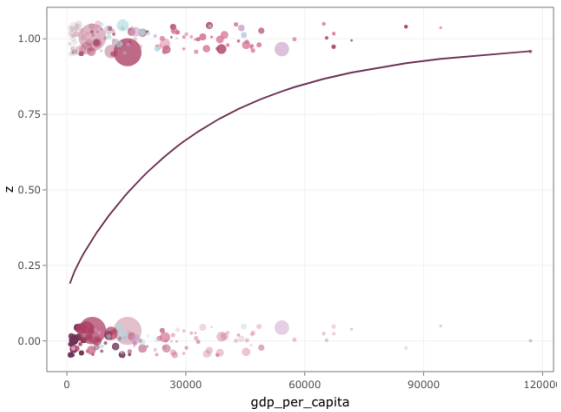



```
df00.logmodel <- glm(prop_vaccinated ~ gdp_per_capita,
    weights=population, family=binomial(link='logit'), data=
    df00)
```

Example


people_fully_vaccinated_per_hundred
gdp_per_capita
0
25
50
75
100
0
30000
60000
90000


people_fully_vaccinated_per_hundred
gdp_per_capita
0
25
50
75
100
0
30000
60000
90000


z
gdp_per_capita
0.00
0.25
0.50
0.75
1.00
0
30000
60000
90000

Example

```
summary(df00.logmodel)
```

```
Call:
glm(formula = prop_vaccinated ~ gdp_per_capita, family =
    binomial(link = "logit"),
    data = df00, weights = population)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-11614.4   -1635.8    -341.4      97.1   24407.0

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.067e+00  3.733e-05  -28586   <2e-16 ***
gdp_per_capita  5.099e-05  1.987e-09   25666   <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: 2417622288  on 186  degrees of freedom
Residual deviance: 1599152072  on 185  degrees of freedom
AIC: 1599154948

Number of Fisher Scoring iterations: 5
```

Properties of the MLE estimates

One can show that the MLE estimates $\hat{\boldsymbol{\beta}}$ satisfies

- $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$
- $\mathrm{Var}[\hat{\boldsymbol{\beta}}] = (\tilde{X}^T V \tilde{X})^{-1}$, where $\tilde{X}$ contains the *unique* samples of the regressors, and $V = \mathrm{diag}(n_j \hat{\pi}_j (1 - \hat{\pi}_j))$.

Interpretation of the regression coefficients

Let

$$\hat{\eta}(\mathbf{x}) \coloneqq \mathbf{x}^T\hat{\boldsymbol{\beta}} \qquad \text{and} \qquad \hat{y}(\mathbf{x}) \coloneqq \hat{\pi}(\mathbf{x}) \coloneqq \frac{\exp \mathbf{x}^T\hat{\boldsymbol{\beta}}}{1+\exp \mathbf{x}^T\hat{\boldsymbol{\beta}}}.$$

Then

$$\log \frac{\hat{\pi}(\mathbf{x})}{1-\hat{\pi}(\mathbf{x})} = \hat{\eta}(\mathbf{x}) \qquad \text{and} \qquad \hat{\eta} = \mathbf{x}^T\hat{\boldsymbol{\beta}}.$$

Let $\delta > 0$ be small. Then

$$\delta\hat{\boldsymbol{\beta}}_j = (\mathbf{x}+\delta\mathbf{e}_j)^T\hat{\boldsymbol{\beta}} - \mathbf{x}^T\hat{\boldsymbol{\beta}} = \hat{\eta}(\mathbf{x}+\delta\mathbf{e}_j) - \hat{\eta}(\mathbf{x}) = \log \frac{\mathsf{odds}_{\mathbf{x}+\delta\mathbf{e}_j}}{\mathsf{odds}_{\mathbf{x}}}$$

and hence

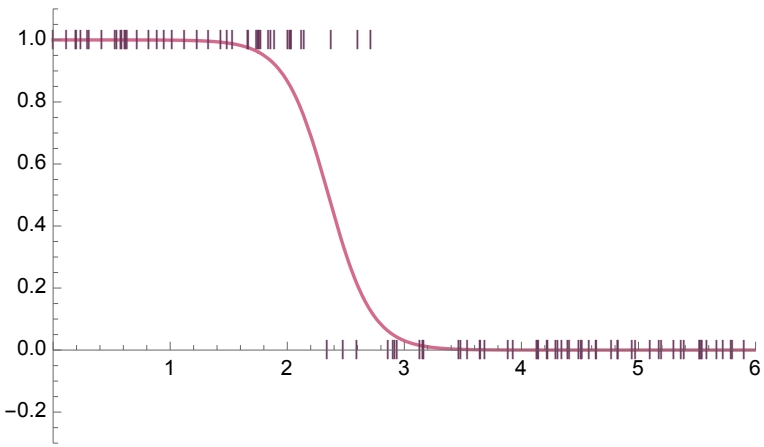
$$\frac{\mathsf{odds}_{\mathbf{x}+\delta\mathbf{e}_j}}{\mathsf{odds}_{\mathbf{x}}} = e^{\delta\hat{\boldsymbol{\beta}}_j}$$

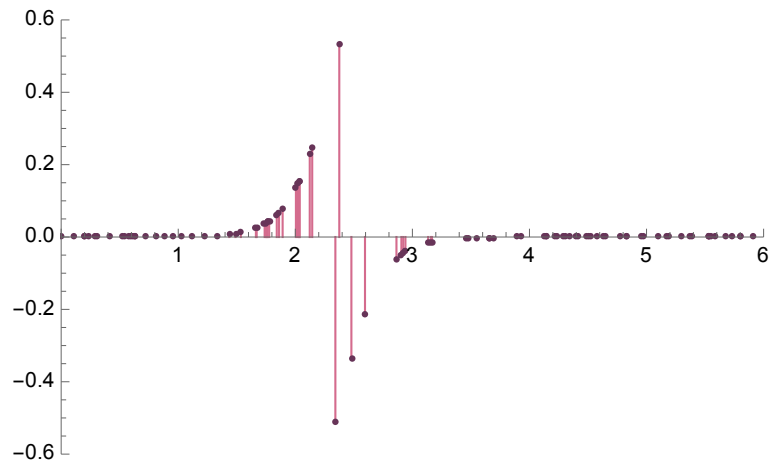
```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.067e+00  3.733e-05  -28586   <2e-16 ***
gdp_per_capita   5.099e-05  1.987e-09   25666   <2e-16 ***
```

Residuals

Residuals

$$e_i := y^i - \hat{y}^i = y^i - n_i \hat{\pi}_i$$



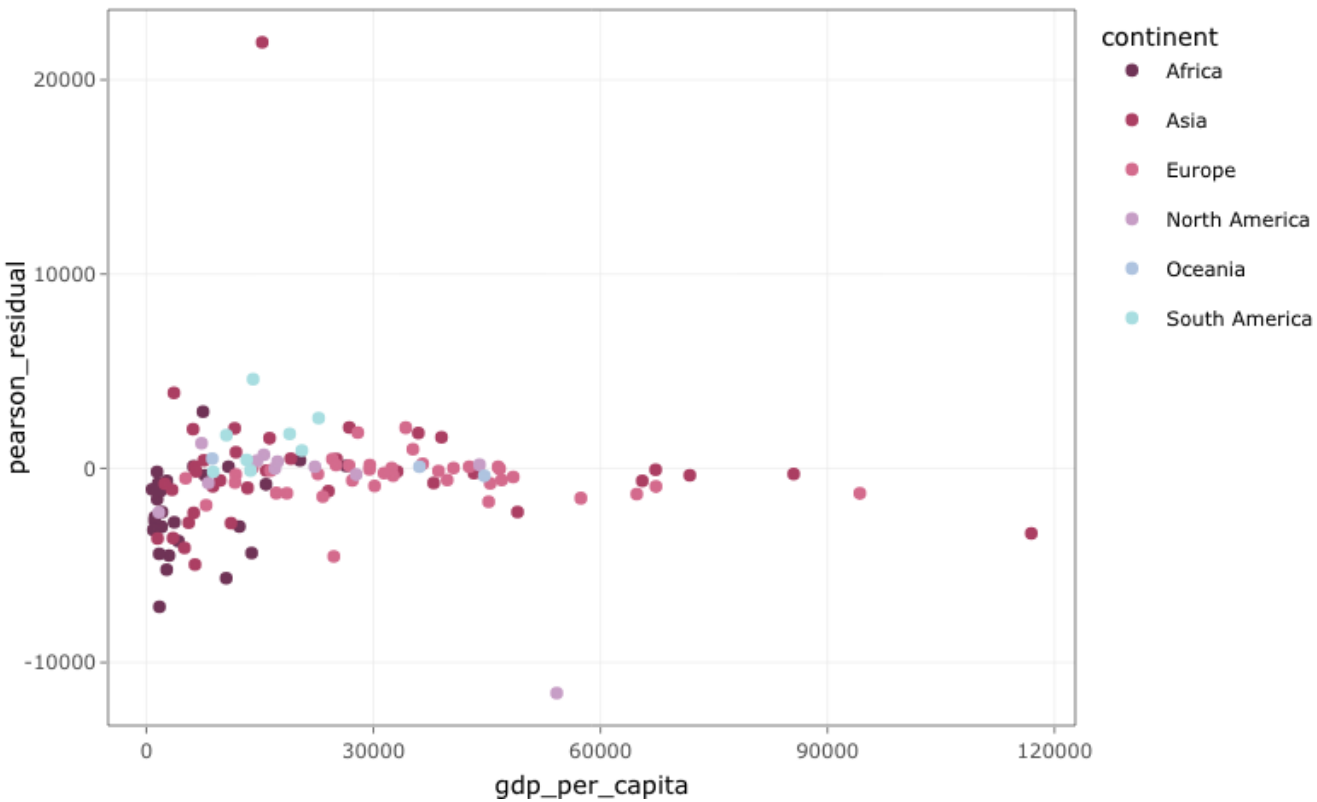


Note that if we do not have repeated samples at the same regressors, then the squared residuals are not expected to be small!

Residuals

Pearson residuals

$$\frac{\overbrace{(y^i - n_i\hat{\pi}_i)}^{\text{\#successes - } \mathbb{E}[\text{\#successes}]}{}^2}{n_i\hat{\pi}_i} + \frac{\overbrace{\left((n_i - y^i) - n_i(1-\hat{\pi}_i)\right)}^{\text{\#failures - } \mathbb{E}[\text{\#failures}]}{}^2}{n_i(1-\hat{\pi}_i)} = \overbrace{\left(\frac{y^i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}\right)}^{\text{Pearson residual } r_i}{}^2 =: r_i^2$$



Residuals

Pearson chi-square statistic

$$\chi^2 := \sum r_i^2 = \sum \frac{(y^i - n_i \hat{\pi}_i)^2}{n_i \hat{\pi}_i (1 - \pi_i)}, \qquad \chi^2 \approx \chi^2_{n-k-1}$$

```
sum(residuals(df00.logmodel1, type = "pearson")^2)
```

```
[1] 1121264528
```

Residuals

Recall the maximum of the binomial log-likelihood function is given by

$$\log L(\mathbf{y}, \hat{\boldsymbol{\beta}}) = \sum_j y^j \log \hat{\pi}_j + \sum (n_j - y^j) \log(1 - \hat{\pi}_j)$$

The *saturated model* (SM) is the model where each choice of regressor gets its own predictor y^j/n_j instead of $\hat{\pi}_j$. We have

$$\begin{aligned}
\log \frac{L(SM)}{L(\mathbf{y}, \hat{\boldsymbol{\beta}})} &= \log SM - \log L(\mathbf{y}, \boldsymbol{\beta}) \\
&= \sum_j y^j \log \frac{y^j/n_j}{\hat{\pi}_j} + \sum (n_j - y^j) \log \frac{1 - y^j/n_j}{1 - \hat{\pi}_j} \\
&= \sum_j y^j \log \frac{y^j}{n_j \hat{\pi}_j} + \sum (n_j - y^j) \log \frac{n_j - y^j}{n_j (1 - \hat{\pi}_j)}
\end{aligned}$$

Deviance

$$D := \log \frac{L(SM)}{L(\mathbf{y}, \hat{\boldsymbol{\beta}})} \sim \chi^2_{n-p}$$

Deviance residuals

$$d_i = \operatorname{sgn} e_i \cdot \left| 2\left(y^i \log \frac{y^i}{n_i \hat{\pi}_i} + (n_i - y^i) \log \frac{n_i - y^i}{n_i (1 - \hat{\pi}_i)} \right) \right|^{1/2}.$$

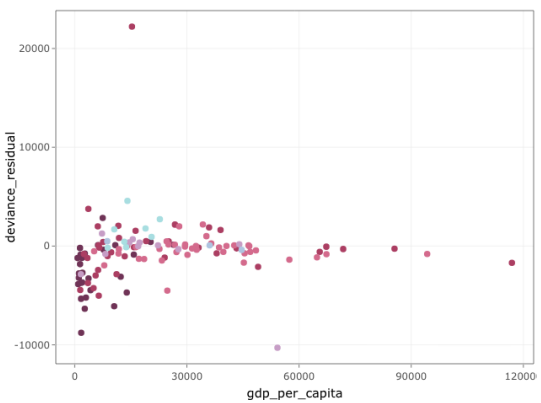
Residuals

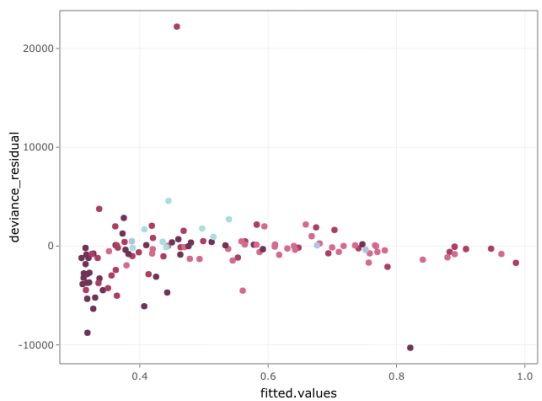
Deviance residuals

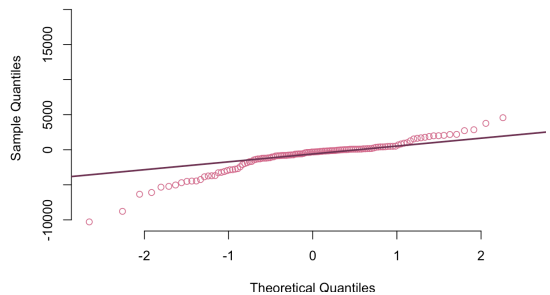
$$d_i = \operatorname{sgn} e_i \cdot \left| 2\left(y^i \log \frac{y^i}{n_i \hat{\pi}_i} + (n_i - y^i) \log \frac{n_i - y^i}{n_i(1-\hat{\pi}_i)} \right) \right|^{1/2}.$$

```
df00.logmodel1$deviance
```

```
[1] 1205910872
```







Likelihood ratio tests for model selection

Test for significance of model

H_0 : a reduced model is correct $\quad$ H_1 : the reduced model is incorrect

If the sample is very large and H_0 is correct, then

$$LR := 2 \log \frac{L(\text{full model})}{L(\text{reduced model})} \approx \chi^2_{\#\text{removed regressors}}$$

Reject H_0 if $LR > \chi^2_{\alpha, \#\text{removed regressors}}$.

```
df00.logmodel1 <- glm(prop_vaccinated~gdp_per_capita ,
    weights=population,family=binomial(link='logit'), data=
    df00)
df00.logmodel2 <- glm(prop_vaccinated~gdp_per_capita +
    hospital_beds_per_thousand ,weights=population,family=
    binomial(link='logit'), data=df00)
```

```
  #Df      LogLik Df      Chisq Pr(>Chisq)
1   2 -602956454
2   3 -503625652  1 198661604  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
    1
```

Likelihood ratio tests for model selection

Test for significance of regression

This test works also if the reduced model contains no regressors, i.e. to test for the significance of regression. In detail, in this case, the model is $E(y) = \pi = e^{\beta_0}/(1 + e^{\beta_0})$, where $\hat{\beta}_0 = \bar{\mathbf{y}}$, and

$$\log L(\mathbf{y}, \hat{\beta}_0) = \sum_i \log \pi_i^{y_i} (1 - \pi_i)^{1-y_i} = n\bar{\mathbf{y}} \log \bar{\mathbf{y}} + (n - n\bar{\mathbf{y}}) \log \bar{\mathbf{y}}$$

```
df00.logmodel0 <- glm(prop_vaccinated~1 ,weights=population,
    family=binomial(link='logit'), data=df00)
df00.logmodel1 <- glm(prop_vaccinated~gdp_per_capita ,
    weights=population,family=binomial(link='logit'), data=
    df00)

library(lmtest)
lrtest(df00.logmodel0, df00.logmodel1)
```

```
  #Df       LogLik Df      Chisq Pr(>Chisq)
1   1 -1208812580
2   2  -799577472  1 818470217  < 2.2e-16 ***
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Likelihood ratio tests for model selection

McFadden's R^2

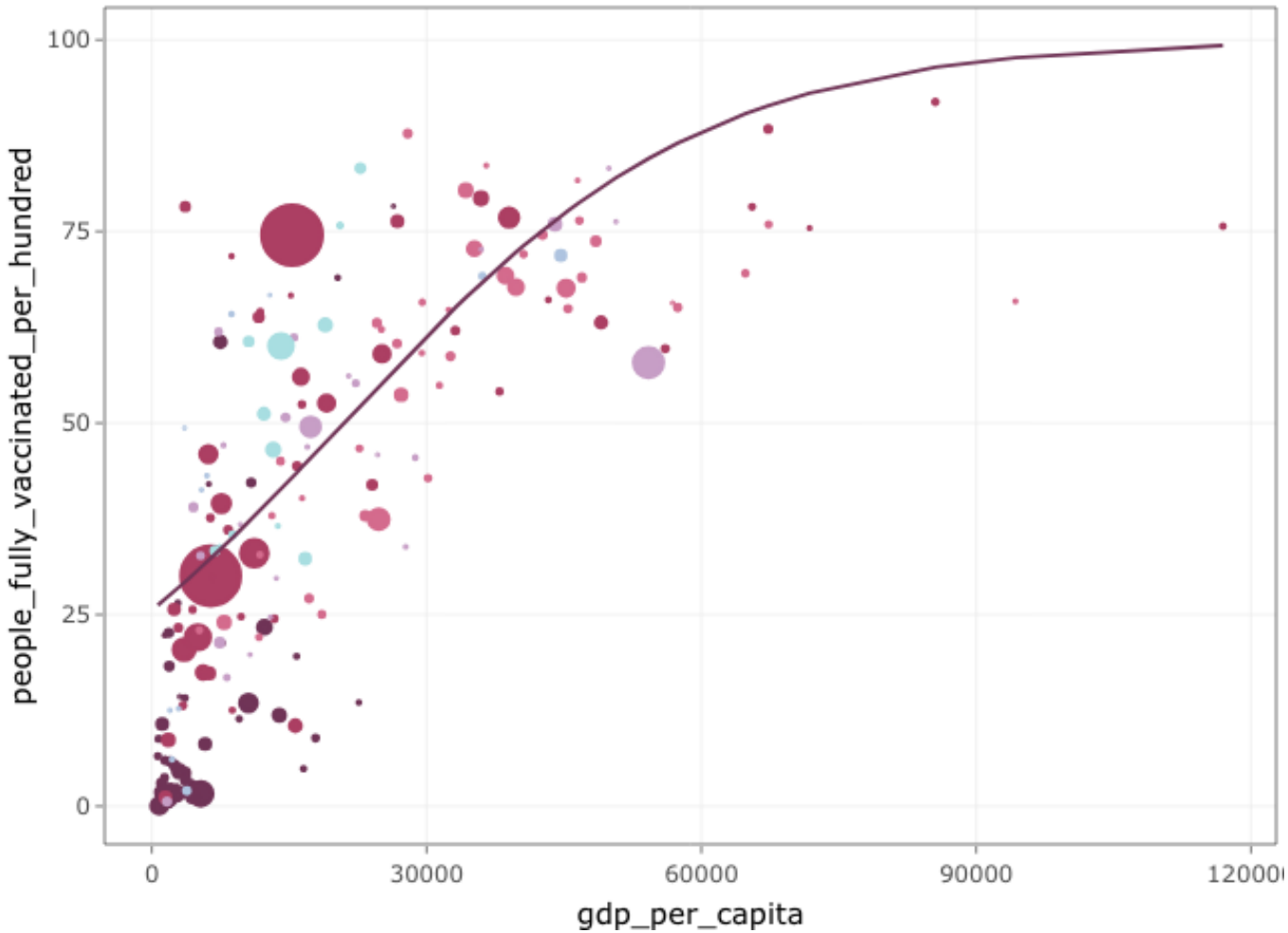
$$R^2_{dev} := 1 - \log L(\text{fill model}) / \log L(\text{intercept model})$$

```
t0 <- lrtest(df00.logmodel0, df00.logmodel1)
1 - t0$LogLik[2]/t0$LogLik[1]
```

```
[1] 0.3207144
```

Example (Attempt 1)

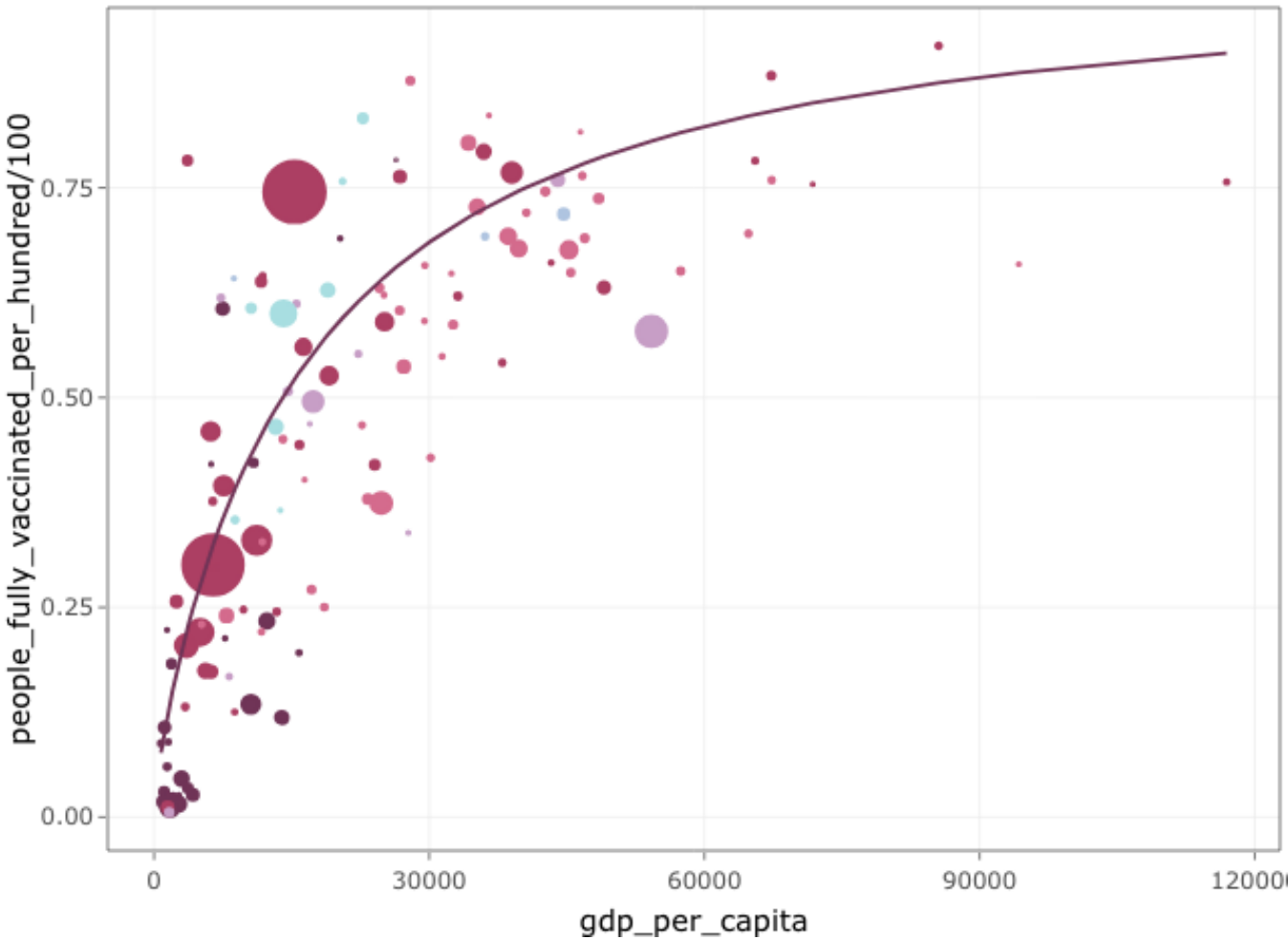
Model: $\eta = \beta_0 + \beta_1 x$ $\quad (R^2_{dev} = 0.32)$



- Wrong power? $\rightarrow$ best power algorithm
- Wrong regressor? $\rightarrow$ best subsets regression/step-wise forward/step-wise backwards?

Example (better power?)

Model: $\eta = \beta_0 + \beta_1 x^{0.1}$ $\quad (R^2_{dev} = 0.56)$



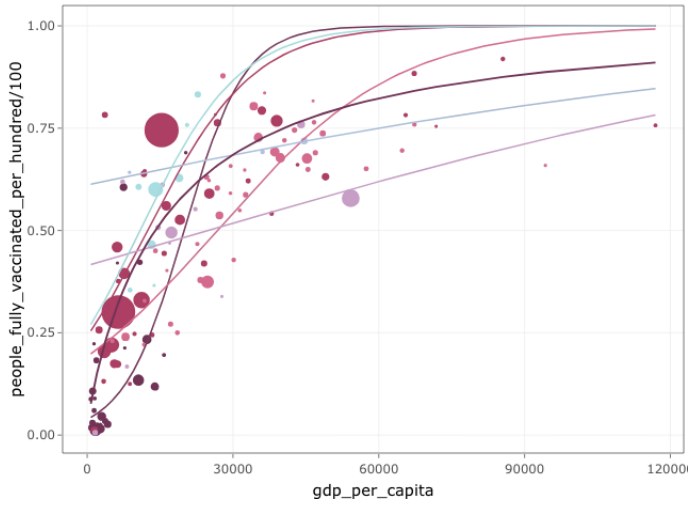
- More regressors? $\rightarrow$ best subsets regression
- Wrong regressor? $\rightarrow$ best subsets regression
- Wrong dataset? $\rightarrow$ divide data into natural groups
- Big data problems? $\rightarrow$ aggregate data

Example (divide data into groups?)

Model: $\eta = \beta_0 + \beta_1 x^{0.1}$

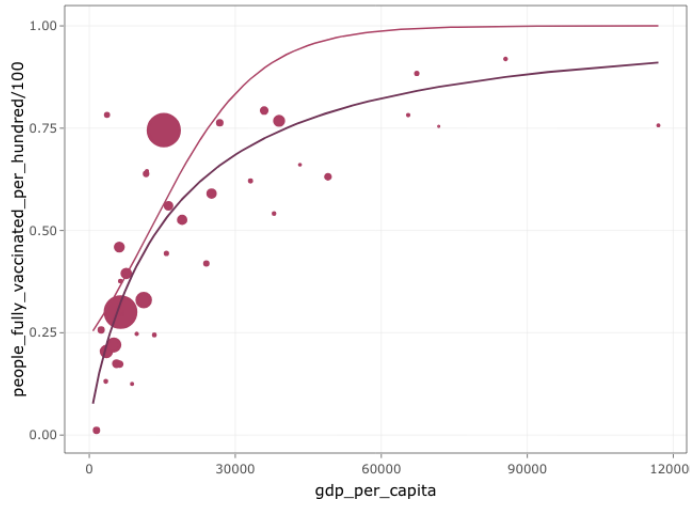
All countries

$R^2_{dev} = 0.5600723$



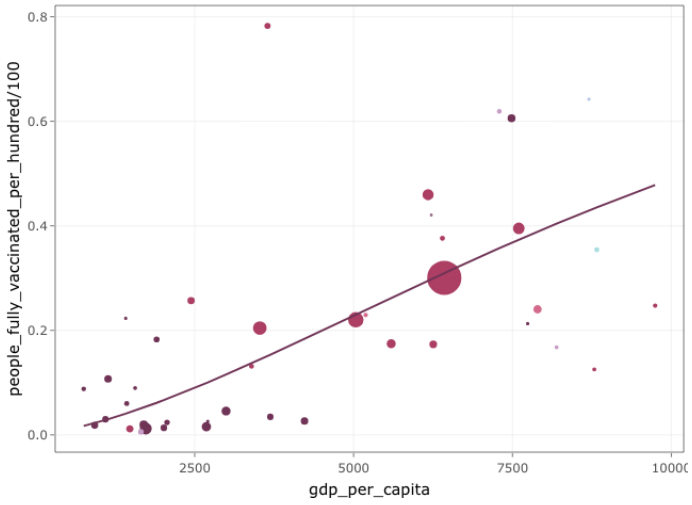
Asia

$R^2_{dev} = 0.7129612$



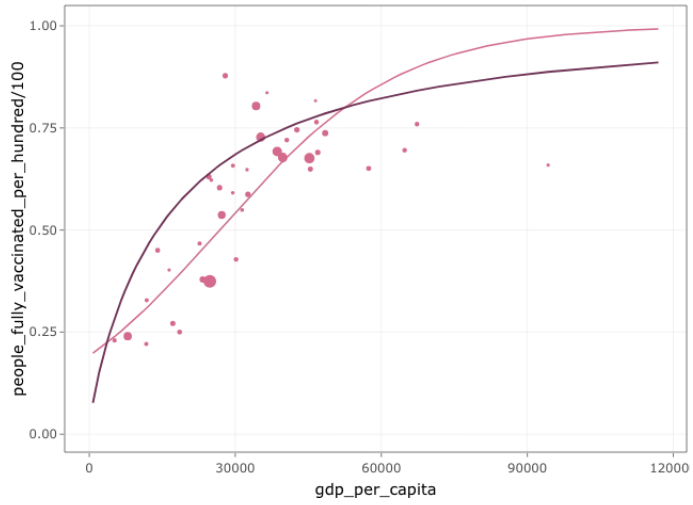
Poorer countries

$R^2_{dev} = 0.6104746$



Europe

$R^2_{dev} = 0.7129612$



Example (more regressors?)

Stepwise forward

- With step-wise forward all regressors are included (with ***), and $R^2 = 0.8753796$
- Inclusion order: `life_expectancy`, `population`, `male_smokers`, `aged_65_older`, `median_age`, ...

Regressor	R^2	D/n
`+ life_expectancy`	0.7254661	0.1858378
`+ population`	0.8307854	0.02697892
`+ male_smokers`	0.8409417	0.002601678
`+ aged_65_older`	0.8466356	0.001458577
`+ median_age`	0.8662862	0.005033745

$\rightarrow$ Multicollinearity?

$\rightarrow$ Is the model too complicated?

$\rightarrow$ Will R^2 change much if we remove some regressors?

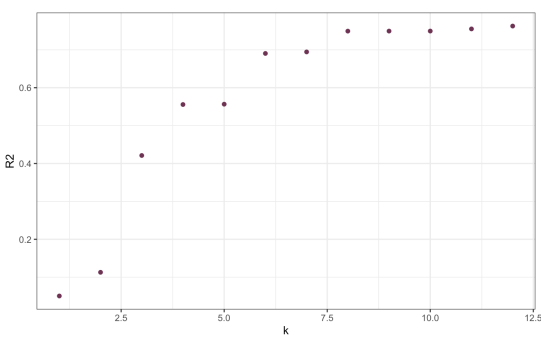
$\rightarrow$ Can changing some powers make the model better?

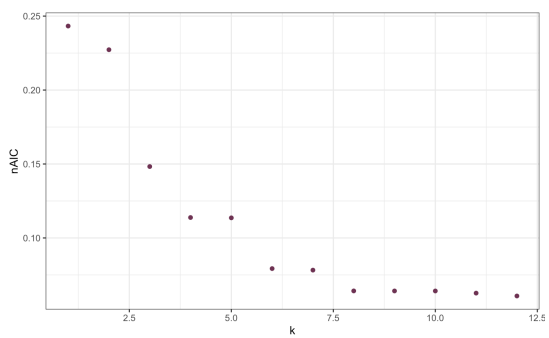
$\rightarrow$ Residuals? Are they normal? Influential points?

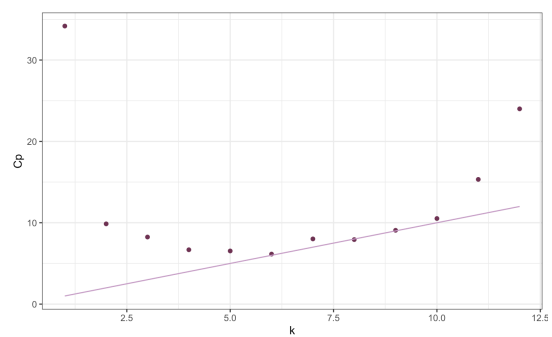
Example (more regressors?)

Best subsets regression

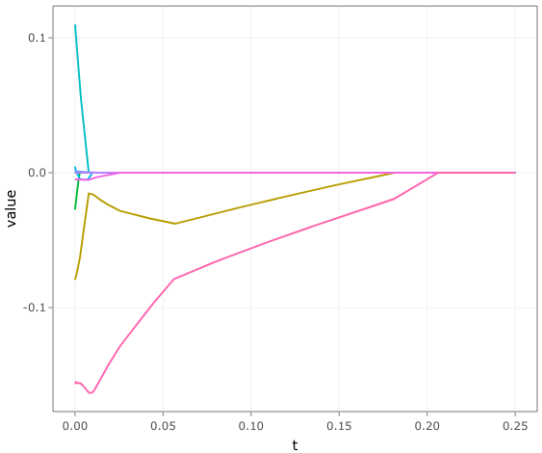
- 1 regressor
  - `gdp_per_capita`
- 2 regressors
  - `total_cases_per_million`
  - `cardiovasc_death_rate`
- 3 regressors
  - `gdp_per_capita`
  - `cardiovasc_death_rate`
  - `aged_65_older`
- 4 regressors
  - `gdp_per_capita`
  - `hospital_beds_per_thousand`
  - `cardiovascular_death_rate`
  - `male_smokers`
- 5 regressors
  - `gdp_per_capita`
  - `hospital_beds_per_thousand`
  - `cardiovascular_death_rate`
  - `male_smokers`
  - `population_density`

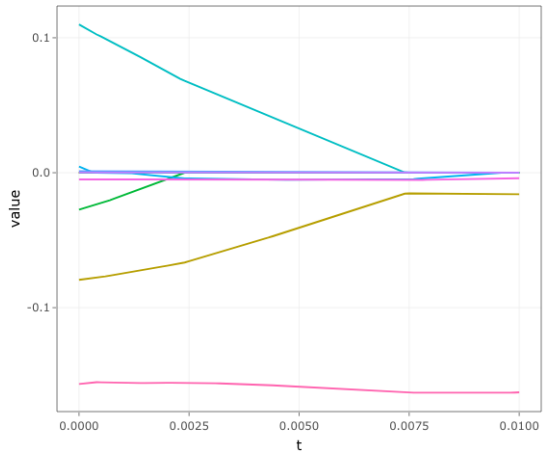


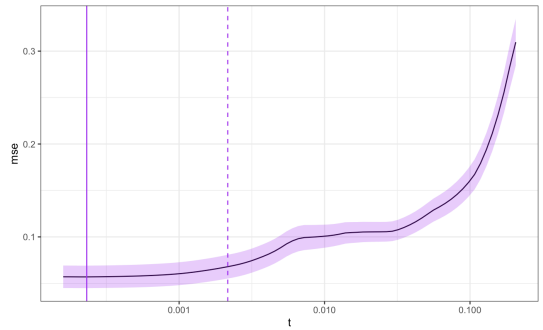




Example (more regressors?)







At $\lambda = .01$, we include: `median_age` and `life_expectancy` ($R^2_{dev} = 0.6748349$)

Bootstrap confidence intervals:

	.05%	99.5%
$\beta_{\texttt{median\_age}}$	-0.03672829	-0.03669084
$\beta_{\texttt{life\_expectancy}}$	-0.07667731	-0.07662509

Example (tests?)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.367e+01	1.008e-03	-13565.3	<2e-16	***
total_cases_per_	-5.132e-07	1.530e-09	-335.4	<2e-16	***
total_deaths_per_	8.930e-05	6.921e-08	1290.3	<2e-16	***
median_age	7.966e-02	1.773e-05	4492.1	<2e-16	***
gdp_per_capita	4.234e-06	4.076e-09	1038.6	<2e-16	***
hospital_beds_p	2.740e-02	2.179e-05	1257.4	<2e-16	***
population	4.776e-10	8.099e-14	5897.7	<2e-16	***
aged_65_older	-1.101e-01	1.759e-05	-6256.0	<2e-16	***
diabetes_prev	-4.579e-03	1.512e-05	-302.9	<2e-16	***
cardiovasc_death	-1.126e-03	4.658e-07	-2416.7	<2e-16	***
population_dens	-1.518e-04	8.936e-08	-1699.1	<2e-16	***
male_smokers	4.990e-03	2.658e-06	1877.6	<2e-16	***
life_expectancy	1.563e-01	1.685e-05	9277.7	<2e-16	***

- What is happening?

Example (aggregate data)

Big data problems

In the past, data was necessarily small and statisticians worked to extract the most value from a little information.
What is true is that trivially small effects can be found to be "significant" with very large sample sizes.

1. Randomness in sample vs. error in measurement
2. Confidence intervals become extremely narrow and we would reject almost any hypothesis for a simple model.

Data aggregation

- Data per country and logistic regression without weights?
- Data per 1m region?