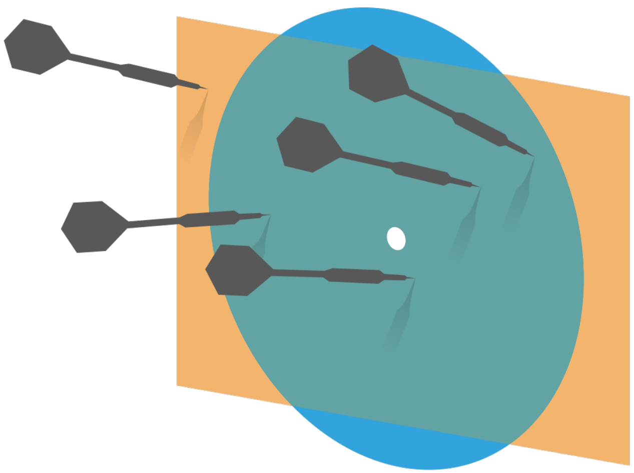
Probability and Statistics for Computer Science



Credit: wikipedia

"All models are wrong, but some models are useful"--- George Box

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 11.17.2020

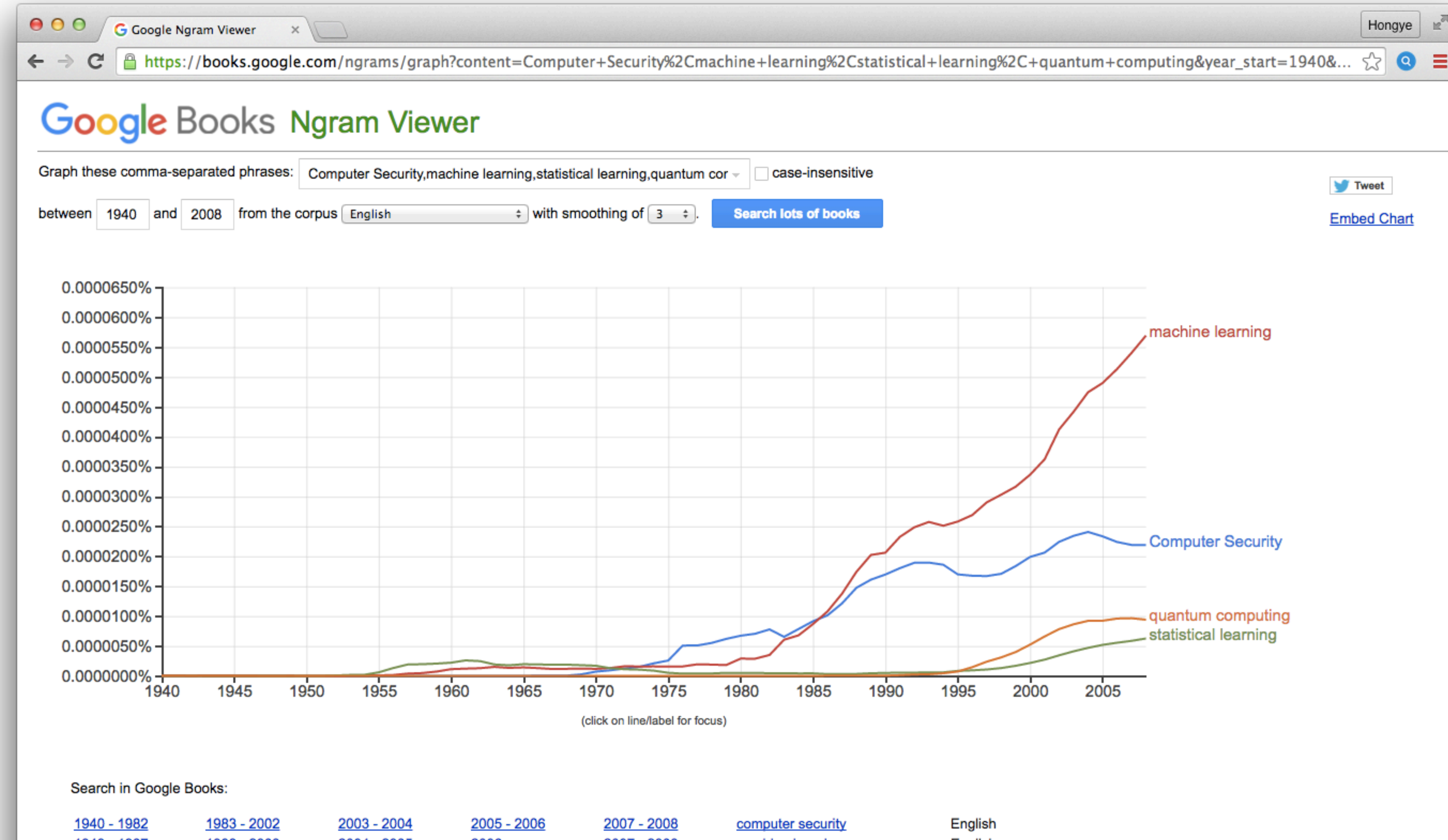
Last time

- Stochastic Gradient Descent
- Naïve Bayesian Classifier

classifier

Regression

Some popular topics in Ngram



Objectives

* Linear regression defition.
* The least square solution
* Training and prediction
* R-squared for evaluating the fit.

Regression models are Machine learning methods

- Regression models have been around for a while
- Dr. Kelvin Murphy’s Machine Learning book has 3+ chapters on regression

The regression problem

Classification

$x^{(1)}$	$x^{(2)}$	$\cdots$	$x^{(d)}$	$y^{(\text{label})}$
0.5	1		*	0
				1
$x_0^{(1)}$	$x_0^{(2)}$			y ?

Regression

$x^{(1)}$	$x^{(2)}$	$\cdots$	$x^{(d)}$	y
0.5	1		*	0.5
				0.6
				1.9
$x_0^{(1)}$	$x_0^{(2)}$			y^P?

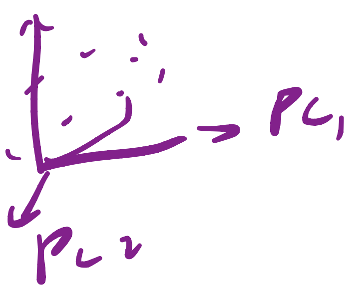
Chicago social economic census

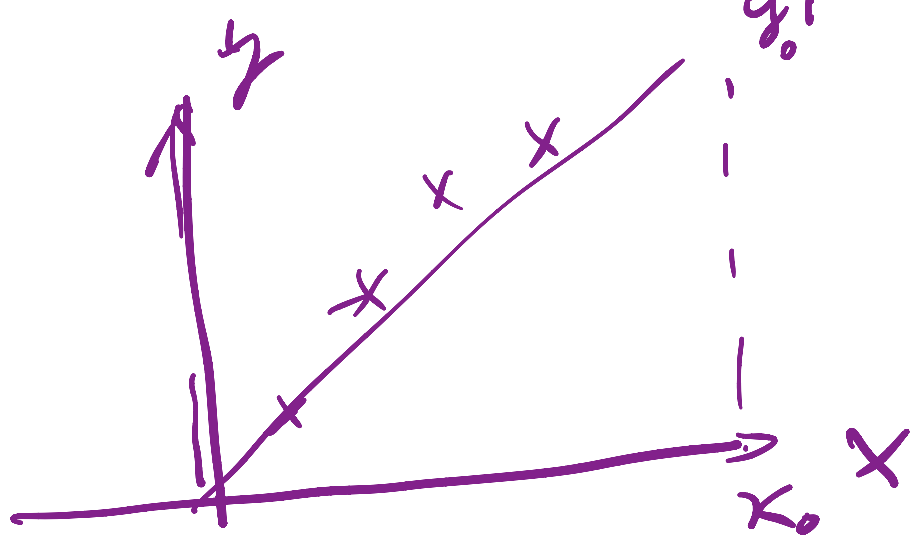
- The census included 77 communities in Chicago
- The census evaluated the average hardship index of the residents
- The census evaluated the following parameters for each community:
  - PERCENT_OF_**HOUSING_CROWDED**
  - PERCENT_**HOUSEHOLD_BELOW_POVERTY**
  - PERCENT_**AGED_16p_UNEMPLOYED**
  - PERCENT_**AGED_25p_WITHOUT_HIGH_SCHOOL_DIPLOMA**
  - PERCENT_**AGED_UNDER_18_OR_OVER_64**
  - PER_CAPITA_**INCOME**

Given a new community and its parameters,
can you predict its average hardship index with all these parameters?

Wait, have we seen the linear regression before?

Correlation X^T





$$\hat{y} = r \cdot \hat{x}$$

It's about *Relationship* between data features

- Example: Is the height of people related to their weight?

IDNO	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT
1	12.6	1.0708	23	154.25	67.75
2	6.9	1.0853	22	173.25	72.25
3	24.6	1.0414	22	154.00	66.25
4	10.9	1.0751	26	184.75	72.25
5	27.8	1.0340	24	184.25	71.25
6	20.6	1.0502	24	210.25	74.75
7	19.0	1.0549	26	181.00	69.75
8	12.8	1.0704	25	176.00	72.50
9	5.1	1.0900	25	191.00	74.00
10	12.0	1.0722	23	198.25	73.50

- x : HIGHT, y: WEIGHT

Some terminology

- Suppose the dataset $\{(\mathbf{x}, y)\}$ consists of N labeled items $(\mathbf{x}_i, y_i)$

- If we represent the dataset as a table
  - The d columns representing $\{\mathbf{x}\}$ are called **explanatory variables** $\mathbf{x}^{(j)}$
  - The numerical column y is called the **dependent variable**

d

N

$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	y
1	3	0
2	3	2
3	6	5

Variables of the Chicago census

[1] "PERCENT_OF_HOUSING_CROWDED"
[2]"PERCENT_HOUSEHOLDS_BELOW_POVERTY"
[3] "PERCENT_AGED_16p_UNEMPLOYED"
[4]"PERCENT_AGED_25p_WITHOUT_HIGH_SCHOOL_DIPLOMA"
[5] "PERCENT_AGED_UNDER_18_OR_OVER_64"
[6]"PER_CAPITA_INCOME"
[7] "HardshipIndex"

Which is the dependent variable in the census example?

A. "PERCENT_OF_HOUSING_CROWDED"
B. "PERCENT_AGED_25p_WITHOUT_HIGH_SCHOOL_DIPLOMA"
C. "HardshipIndex"
D. "PERCENT_AGED_UNDER_18_OR_OVER_64"

Linear model

- We begin by modeling y as a linear function of $\mathbf{x}^{(j)}$ plus randomness

$$y = \mathbf{x}^{(1)}\beta_1 + \mathbf{x}^{(2)}\beta_2 + \ldots + \mathbf{x}^{(d)}\beta_d + \xi$$

Where ξ is a zero-mean random variable that represents model error

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_d \end{bmatrix}$$

$$x^T = [x^{(1)} \; x^{(2)} \ldots x^{(d)}]$$

- In vector notation:

$$y = \mathbf{x}^T \boldsymbol{\beta} + \xi$$

Where $\boldsymbol{\beta}$ is the d-dimensional vector of coefficients that we train

$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	y
1	3	0
2	3	2
3	6	5

Each data item gives an equation

- The model: $y = \mathbf{x}^T \boldsymbol{\beta} + \xi = \mathbf{x}^{(1)}\beta_1 + \mathbf{x}^{(2)}\beta_2 + \xi$

$y = 0 = 1 \times \beta_1 + 3 \times \beta_2 + \xi_1$

$2 = 2 \times \beta_1 + 3 \times \beta_2 + \xi_2$

$5 = 3 \times \beta_1 + 6 \times \beta_2 + \xi_3$

Training data

$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	y
1	3	0
2	3	2
3	6	5

Which together form a matrix equation

- The model $y = \mathbf{x}^T \boldsymbol{\beta} + \xi = \mathbf{x}^{(1)}\beta_1 + \mathbf{x}^{(2)}\beta_2 + \xi$

$E[\xi] = 0$

Training data

$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	y
1	3	0
2	3	2
3	6	5

X

$$\begin{bmatrix} 0 \\ 2 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 2 & 3 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix}$$

$$y = X \cdot \beta + e$$

unknown

Which together form a matrix equation

- The model $y = \mathbf{x}^T \boldsymbol{\beta} + \xi = \mathbf{x}^{(1)} \beta_1 + \mathbf{x}^{(2)} \beta_2 + \xi$

d

Training data

$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	y
1	3	0
2	3	2
3	6	5

N

$$\begin{bmatrix} 0 \\ 2 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 2 & 3 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix}$$

$$\mathbf{y} = X \cdot \boldsymbol{\beta} + \mathbf{e}$$

Q. What’s the dimension of matrix X?

A. $N \times d$

B. $d \times N$

C. $N \times N$

D. $d \times d$

Training the model is to choose β

- Given a training dataset $\{(\mathbf{x}, y)\}$, we want to fit a model $y = \mathbf{x}^T \boldsymbol{\beta} + \xi$

- Define $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$ and $X = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix}$ and $\mathbf{e} = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_N \end{bmatrix}$

- To train the model, we need to choose $\boldsymbol{\beta}$ that makes $\mathbf{e}$ small in the matrix equation $\mathbf{y} = X \cdot \boldsymbol{\beta} + \mathbf{e}$

(1) Least Square = (2) MLE

Loss function

Textbook pg 309

Training using least squares

- In the least squares method, we aim to minimize $\|\mathbf{e}\|^2$

Loss / Cost

$$\|\mathbf{e}\|^2 = \|\mathbf{y} - X\boldsymbol{\beta}\|^2 = \boxed{(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})}$$

scalar

$\|v\|^2 = v^T v$

- Differentiating with respect to $\boldsymbol{\beta}$ and setting to zero

$$X^T X\boldsymbol{\beta} - X^T\mathbf{y} = 0$$

$X^T X \hat{\beta} = X^T y$

- If $X^T X$ is invertible, the least squares estimate of the coefficient is:

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$$

$\hat{\beta} = \arg\min_{\beta} \|e\|^2$

$y = X\beta + e$

$X^T X$

$X^T \sim d \times N$

$X \sim N \times d$

$X^T_{(d \times N)} \cdot X_{(N \cdot d)}$ λ_i

$= X^T X \sim \underline{d \times d}$

symmetric. real valued λ.

for $X^T X$, we have $\underline{\lambda_i \geq 0}$

Derivation of least square solution

$$\|e\|^2 = (y - X\beta)^T (y - X\beta)$$
$$= y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta \quad (1)$$

useful derivative involving vector/matrix

a, b are vectors ; A is a square matrix

$$\frac{\partial (a^T A a)}{\partial a} = (A + A^T) a \Rightarrow \frac{\partial (\beta^T X^T X \beta)}{\partial \beta} = 2X^T X\beta$$

$$\frac{\partial (b^T a)}{\partial a} = b \Rightarrow \frac{\partial (y^T X \beta)}{\partial \beta} = X^T y$$

since $b^T a$ is scalar

$$\frac{\partial (b^T a)}{\partial a} = \frac{\partial (b^T a)^T}{\partial a} = \frac{\partial (a^T b)}{\partial a} = b \Rightarrow \frac{\partial (\beta^T X^T y)}{\partial \beta} = X^T y$$

$\because$ $X^T X$ is symmetric

$\because$ $X^T X = (X^T X)^T$

$\therefore X^T X + (X^T X)^T = 2X^T X$

Note $\|e\|^2$ is scalar, all items in (1) are scalar

$$\frac{\partial \|e\|^2}{\partial \beta} = 0 - X^T y - X^T y + 2X^T X\beta = 0$$

$$\Rightarrow X^T X \beta = X^T y$$

$$\boxed{\Rightarrow \beta = (X^T X)^{-1} X^T y}$$

here y is vector

Derivation of least square solution

$$X^T y = X^T X \hat{\beta}$$

$$\Rightarrow X^T(y - X\hat{\beta}) = 0$$

$$\Rightarrow X^T e = 0 \quad (d \times 1)$$

$$\Rightarrow e^T X = 0 \quad (1 \times d)$$

$$\Rightarrow e^T X \hat{\beta} = 0 \quad (1 \times 1)$$

$$e \perp X\hat{\beta}$$

uncorrelated !!

$X^T \sim d \times N$

$e \sim N \times 1$

$(X^T e)^T = 0$

Least square Loss function

$$\|e\|^2 = f(\beta) = \sum_{j=1}^{K} Q_j(\beta) = \sum_{j=1}^{K} \underbrace{(x_j^T\beta - y_j)}_{e_i}{}^2$$

$$Q_j(\beta) = (x_j^T\beta - y_j)^2$$

in the final project

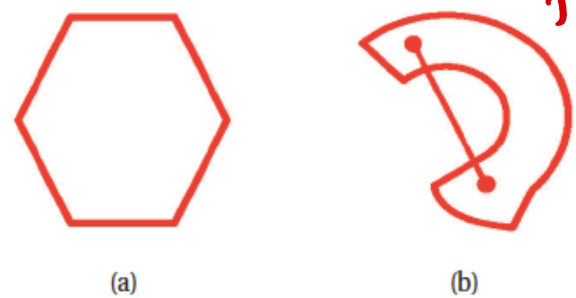
$$Q_j(\theta) = |x_j^T\theta - y_j|^\gamma$$

$$\nabla Q_j = ? \qquad \frac{\partial Q_j}{\partial \theta} = ?$$

Convex set and convex function

- If a set is convex, any line connecting two points in the set is completely included in the set

$f(\theta) = \theta^2$



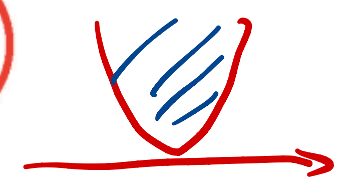
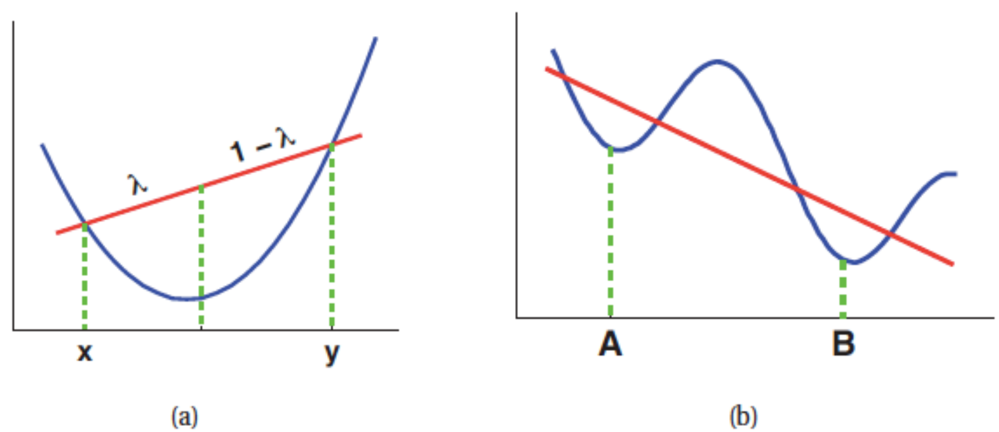


Figure 7.4 (a) Illustration of a convex set. (b) Illustration of a nonconvex set.

- A convex function: the area above the curve is convex

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y)$$



- The least square function is **convex**

Credit: Dr. Kelvin Murphy

What's the dimension of matrix X^TX?

A. $N \times d$
B. $d \times N$
C. $N \times N$
D. $\boxed{\text{D.}}$ $d \times d$

$X \sim N \times d$

$X^T \sim d \times N$

$X^T X \sim d \times d$

$d \rightarrow$ # of features / explanatory var.

Is this statement true?

If the matrix $\mathbf{X^TX}$ does NOT have zero valued eigenvalues, it is invertible.

$\boxed{\text{A.}}$ TRUE
B. FALSE

$\lambda_i \geq 0$

if $\lambda_i \neq 0$ for all i

$\lambda > 0$

Training using least squares example

- Model: $y = \mathbf{x}^T\boldsymbol{\beta} + \xi = \mathbf{x}^{(1)}\beta_1 + \mathbf{x}^{(2)}\beta_2 + \xi$

X

Training data

$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	y
1	3	0
2	3	2
3	6	5

$$\widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{y} = \begin{bmatrix} 2 \\ -\frac{1}{3} \end{bmatrix}$$

$$\widehat{\boldsymbol{\beta}}_1 = 2$$
$$\widehat{\boldsymbol{\beta}}_2 = -\frac{1}{3}$$

Prediction

- If we train the model coefficients $\widehat{\boldsymbol{\beta}}$, we can predict y_0^p from $\mathbf{x}_0$

$$y_0^p = \mathbf{x}_0^T \widehat{\boldsymbol{\beta}}$$

- In the model $y = \mathbf{x}^{(1)}\beta_1 + \mathbf{x}^{(2)}\beta_2 + \xi$ with $\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 2 \\ -\frac{1}{3} \end{bmatrix}$

  - The prediction for $\mathbf{x}_0 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ is y_0^p $= 2 \times \beta_1 + 1 \times \beta_2$

    $= 4 - 1 \times \frac{1}{3}$

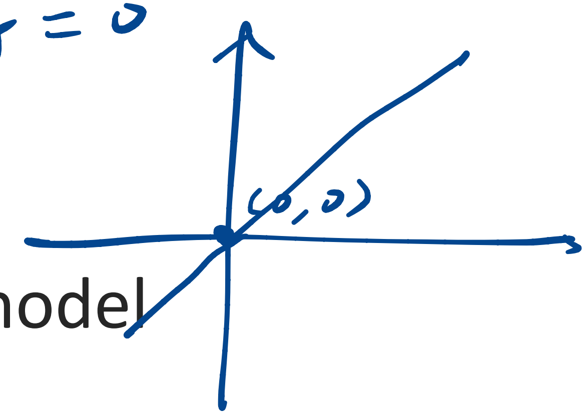
  - The prediction for $\mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ is y_0^p

A linear model with constant offset

- The problem with the model $y = \mathbf{x}^{(1)}\beta_1 + \mathbf{x}^{(2)}\beta_2 + \xi$ is:

when $x^{(1)} = 0,\ x^{(2)} = 0$

$y = 0$



- Let's add a constant offset β_0 to the model

$$y = \beta_0 + \mathbf{x}^{(1)}\beta_1 + \mathbf{x}^{(2)}\beta_2 + \xi$$

$1 \times \beta_0 + x^{(1)} \cdot \beta_1 \ - \ - \ \cdot$

Training and prediction with constant offset

- The model $y = \beta_0 + \mathbf{x}^{(1)}\beta_1 + \mathbf{x}^{(2)}\beta_2 + \xi = \mathbf{x}^T\boldsymbol{\beta} + \xi$

intercept

- Training data:

$$\underbrace{\begin{bmatrix} 1 & x^{(1)} & x^{(2)} \end{bmatrix}}$$

β_0

$\mathbf{1}$	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	y
1	1	3	0
1	2	3	2
1	3	6	5

$$\widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{y} = \begin{bmatrix} -3 \\ 2 \\ \frac{1}{3} \end{bmatrix}$$

- For $\mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$$y_0^p = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} -3 \\ 2 \\ \frac{1}{3} \end{bmatrix} = -3$$

Comparing our example models

$$y = \mathbf{x}^{(1)}\beta_1 + \mathbf{x}^{(2)}\beta_2 + \xi$$

$$y = \beta_0 + \mathbf{x}^{(1)}\beta_1 + \mathbf{x}^{(2)}\beta_2 + \xi$$

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 2 \\ -\frac{1}{3} \end{bmatrix}$$

$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	y	$\mathbf{x}^T\widehat{\boldsymbol{\beta}}$
1	3	0	1
2	3	2	3
3	6	5	4

$y - x^T\beta$

$= -1$

-1

1

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} -3 \\ 2 \\ \frac{1}{3} \end{bmatrix}$$

1	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	y	$\mathbf{x}^T\widehat{\boldsymbol{\beta}}$
1	1	3	0	0
1	2	3	2	2
1	3	6	5	5

Variance of the linear regression model

- The least squares estimate satisfies this property

$$var(\{y_i\}) = var(\{\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}\}) + var(\{\xi_i\})$$

$y = X\hat{\beta} + e$

$e \perp \underline{X\hat{\beta}}$

- The random error is uncorrelated to the least square solution of linear combination of explanatory variables.

$X^T y = X^T X \hat{\beta}$

Variance of the linear regression model: proof

- The least squares estimate satisfies this property

$y = X\beta + e$

$$var(\{y_i\}) = var(\{\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}\}) + var(\{\xi_i\})$$

Proof:

$var(y) = var(X\beta) + var(e)$
$+ 2\,cov(X\beta, e)$

$\because \; X\beta \perp e$

$cov(X\beta, e) = 0$

Variance of the linear regression model: proof

- The least squares estimate satisfies this property

$$var(\{y_i\}) = var(\{\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}\}) + var(\{\xi_i\})$$

Proof:

$$var[y] = (1/N)([X\hat{\beta} - \overline{X\hat{\beta}}] + [\mathbf{e} - \overline{\mathbf{e}}])^T([X\hat{\beta} - \overline{X\hat{\beta}}] + [\mathbf{e} - \overline{\mathbf{e}}])$$

$$var[y] = (1/N)([X\hat{\beta} - \overline{X\hat{\beta}}]^T[X\hat{\beta} - \overline{X\hat{\beta}}] + 2[\mathbf{e} - \overline{\mathbf{e}}]^T[X\hat{\beta} - \overline{X\hat{\beta}}] + [\mathbf{e} - \overline{\mathbf{e}}]^T[\mathbf{e} - \overline{\mathbf{e}}])$$

Because $\overline{\mathbf{e}} = 0$; $\mathbf{e}^T X\widehat{\boldsymbol{\beta}} = 0$ and $\mathbf{e}^T\mathbf{1} = 0$ ← Due to Least square minimized

$$var[y] = (1/N)([X\hat{\beta} - \overline{X\hat{\beta}}]^T[X\hat{\beta} - \overline{X\hat{\beta}}] + [\mathbf{e} - \overline{\mathbf{e}}]^T[\mathbf{e} - \overline{\mathbf{e}}])$$

$$\boxed{var[y] = var[X\widehat{\beta}] + var[\mathbf{e}]}$$

Evaluating models using R-squared

- The least squares estimate satisfies this property

$$var(\{y_i\}) = var(\{\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}\}) + var(\{\xi_i\})$$

- This property gives us an evaluation metric called R-squared

$$R^2 = \frac{var(\{\mathbf{x}_i^T \widehat{\boldsymbol{\beta}}\})}{var(\{y_i\})}$$

- We have $0 \leq R^2 \leq 1$ with a larger value meaning a better fit.

Q: What is R-squared if there is only one explanatory variable in the model?

if $X = N \times 1$

$d = 1$

$R^2 \rightarrow r^2$

r is corr.

Q: What is R-squared if there is only one explanatory variable in the model?

$$\hat{y} = r\hat{x} + \varepsilon$$

$$var(\hat{y}) = r^2 var[\hat{x}] + var[\varepsilon]$$

$$R^2 = \frac{r^2 var[\hat{x}]}{var[\hat{y}]}$$

$$var[\hat{x}] = 1$$

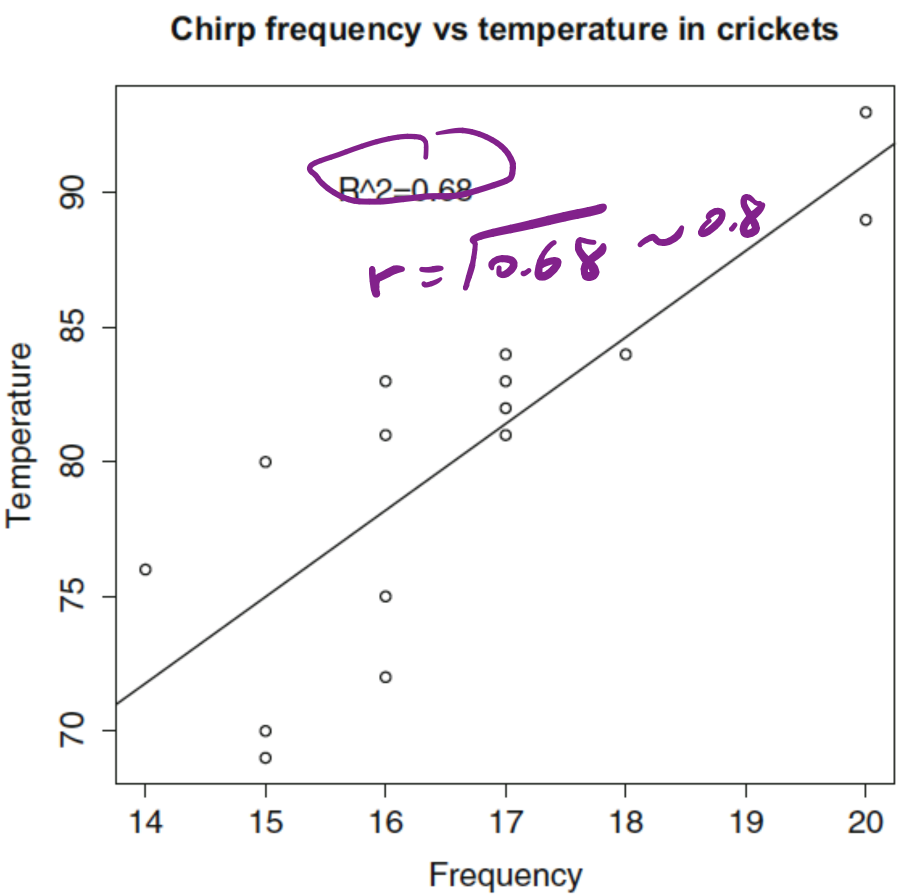
$$var[\hat{y}] = 1$$

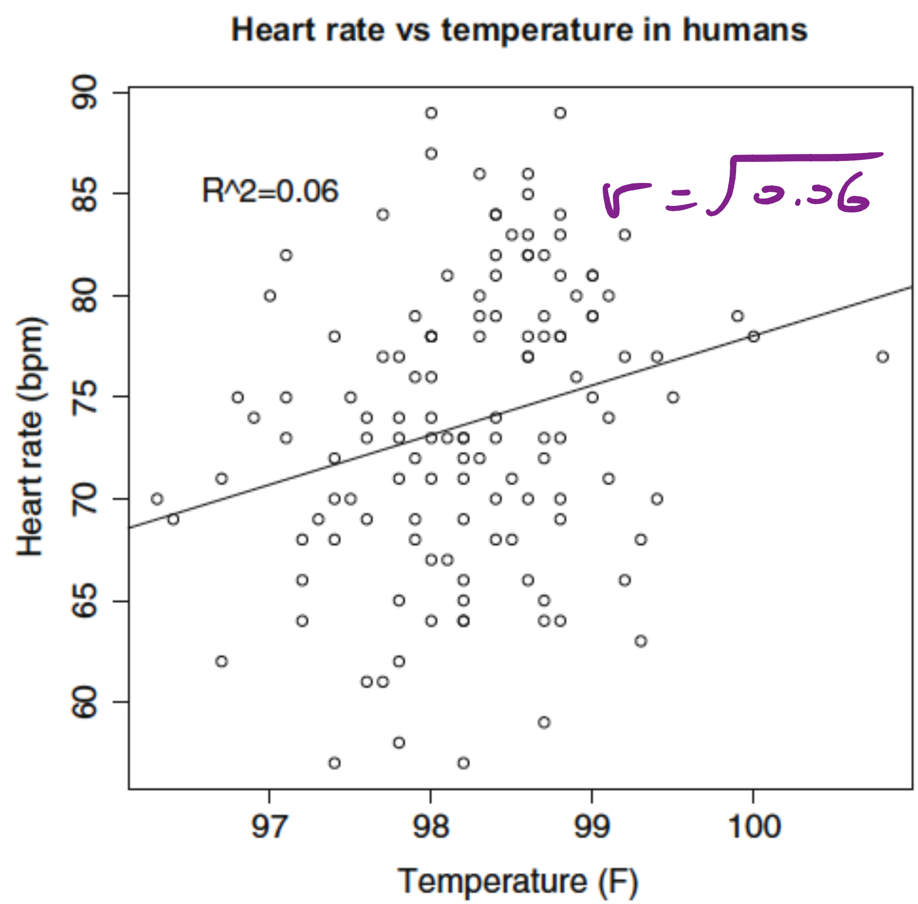
$$= r^2$$

Q: What is R-squared if there is only one explanatory variable in the model?

R-squared would be **the correlation coefficient squared** (textbook pgs 43-44)

R-squared examples





Linear regression model for the Chicago census data

```
Call:
lm(formula = HardshipIndex ~ ., data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-15.7157  -1.9230   0.1301   1.9810   8.6719

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               105.1394    37.3622   2.814 0.006346 **
PERCENT_OF_HOUSING_CROWDED                  0.7189     0.2753   2.612 0.011014 *
PERCENT_HOUSEHOLDS_BELOW_POVERTY            0.6665     0.0781   8.534 1.90e-12 ***
PERCENT_AGED_16p_UNEMPLOYED                 0.8023     0.1350   5.941 9.93e-08 ***
PERCENT_AGED_25p_WITHOUT_HIGH_SCHOOL_DIPLOMA 0.7751    0.1063   7.293 3.64e-10 ***
PERCENT_AGED_UNDER_18_OR_OVER_64            0.4807     0.1202   3.998 0.000156 ***
PER_CAPITA_INCOME                         -11.8819     3.1888  -3.726 0.000391 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.9 on 70 degrees of freedom
Multiple R-squared:  0.983, Adjusted R-squared:  0.9815
F-statistic: 673.9 on 6 and 70 DF,  p-value: < 2.2e-16
```

$df = 70$

$= N - d^*$

$d^* \rightarrow$ # of explanatory + intercept

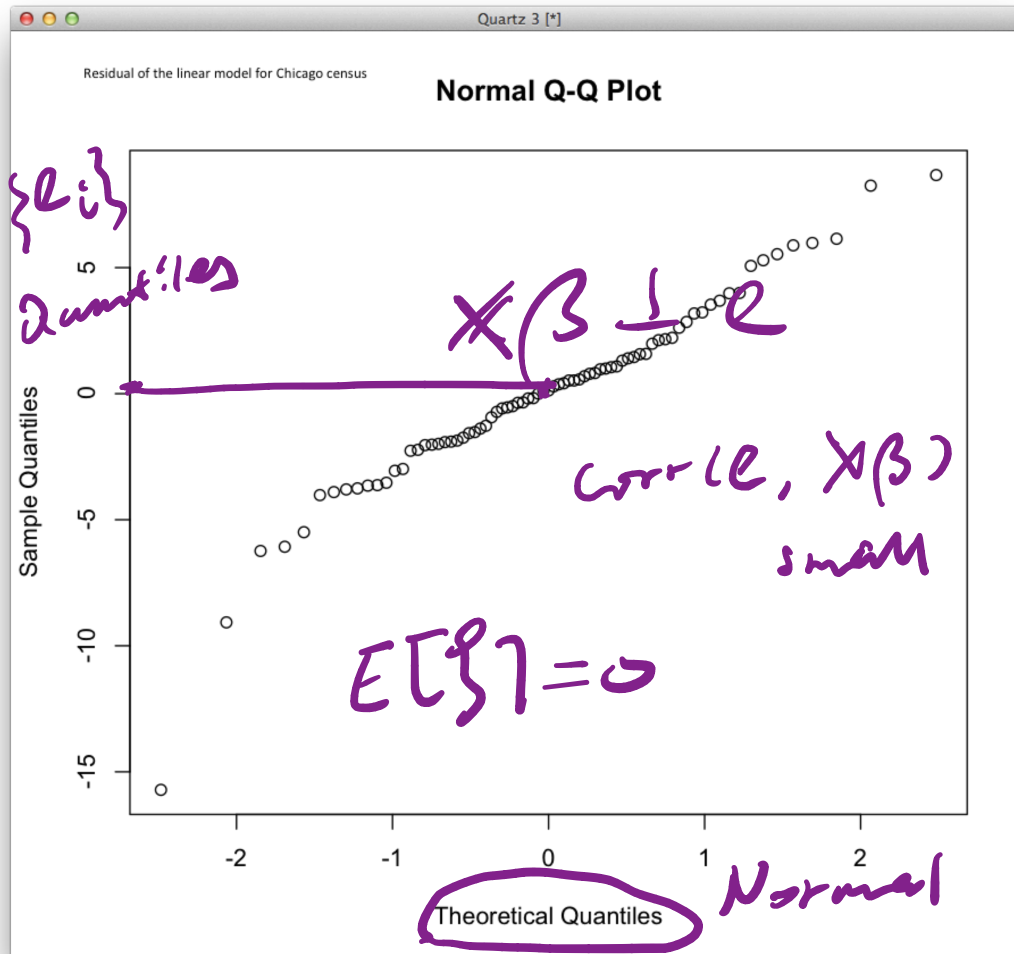
Residual is normally distributed?

The Q-Q plot of the residuals is roughly normal

e $\{e_i\}$

$$y_0 = x_0^T \beta + e_0$$

$\{e_i\}$



Prediction for another community

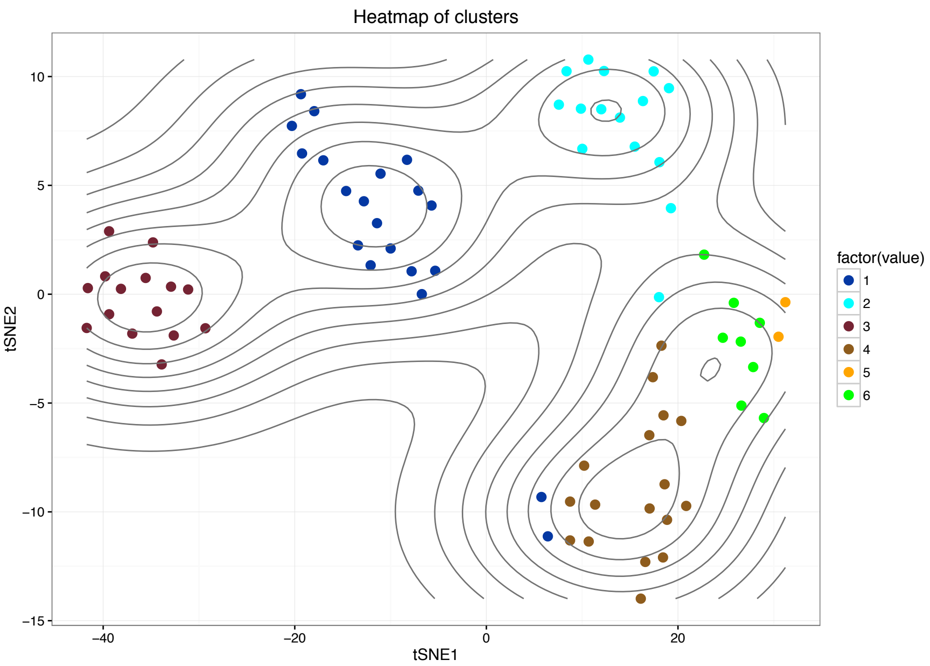
[1] "PERCENT_OF_HOUSING_CROWDED"	**4.7**
[2]"PERCENT_HOUSEHOLDS_BELOW_POVERTY"	**19.7**
[3] "PERCENT_AGED_16p_UNEMPLOYED"	**12.9**
[4]"PERCENT_AGED_25p_WITHOUT_HIGH_SCHOOL_DIPLOMA"	**19.5**
[5] "PERCENT_AGED_UNDER_18_OR_OVER_64"	**33.5**
[6]"PER_CAPITA_INCOME"	**Log(28202)**

Predicted hardship index: **41.46038**

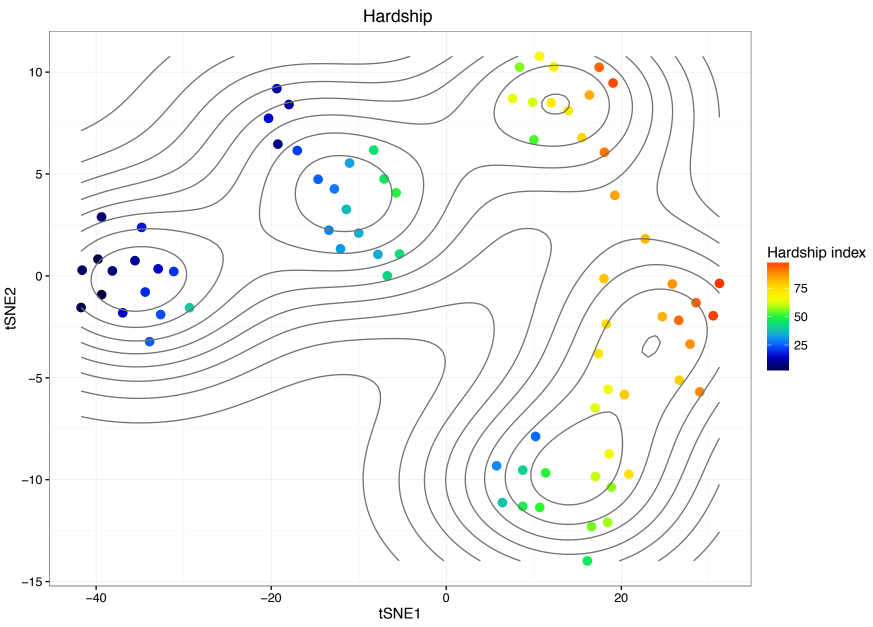
Note: maximum of hardship index in the training data is 98, minimum is 1

The clusters of the Chicago communities: clusters and hardship

Clusters of community

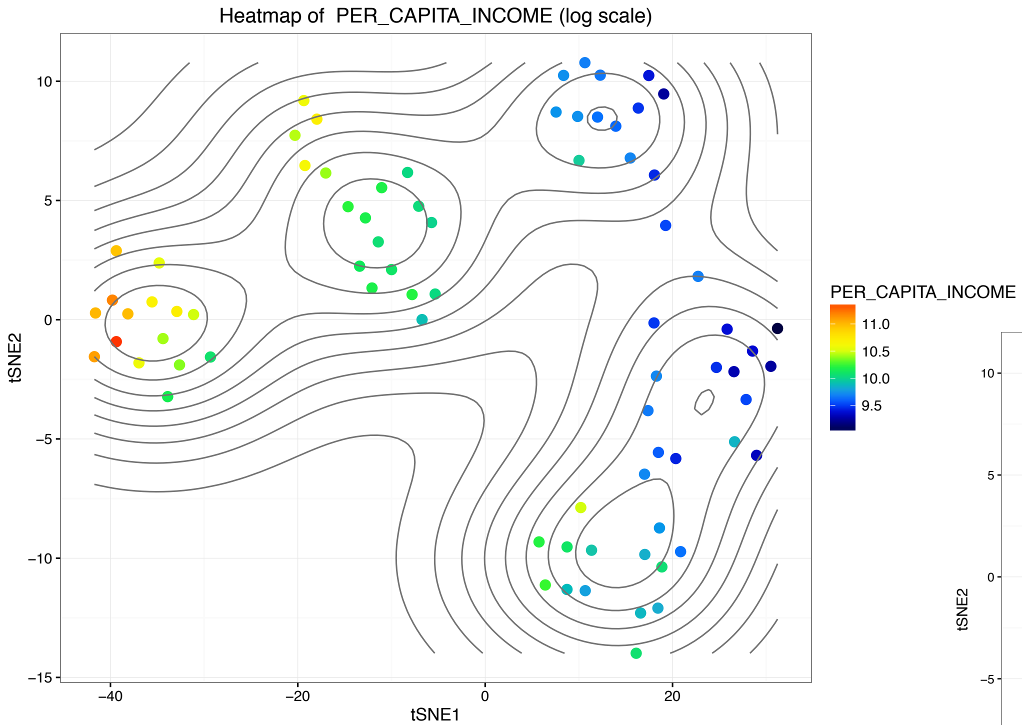


Hardship index of communities

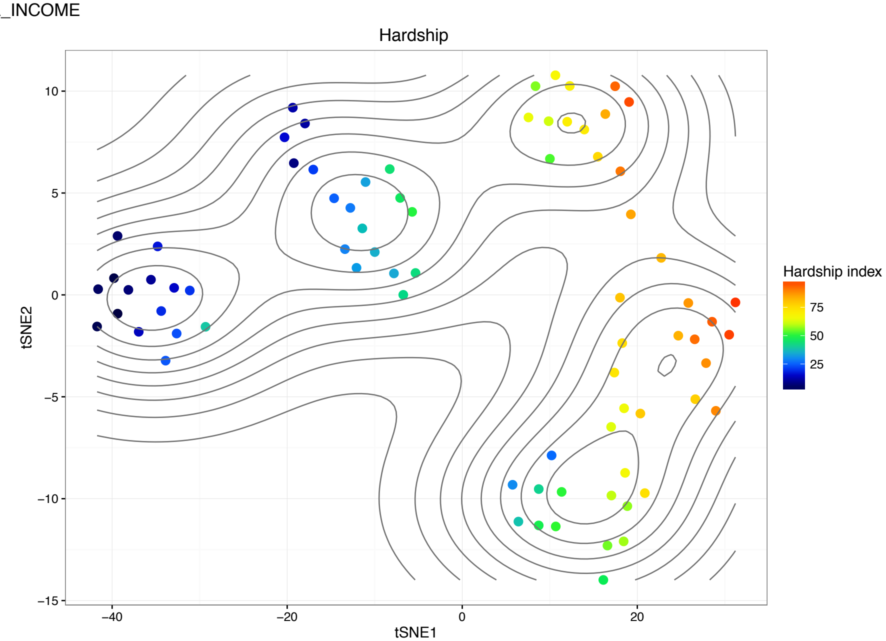


The clusters of the Chicago communities: per capital income and hardship

PER_CAPITAL_INCOME

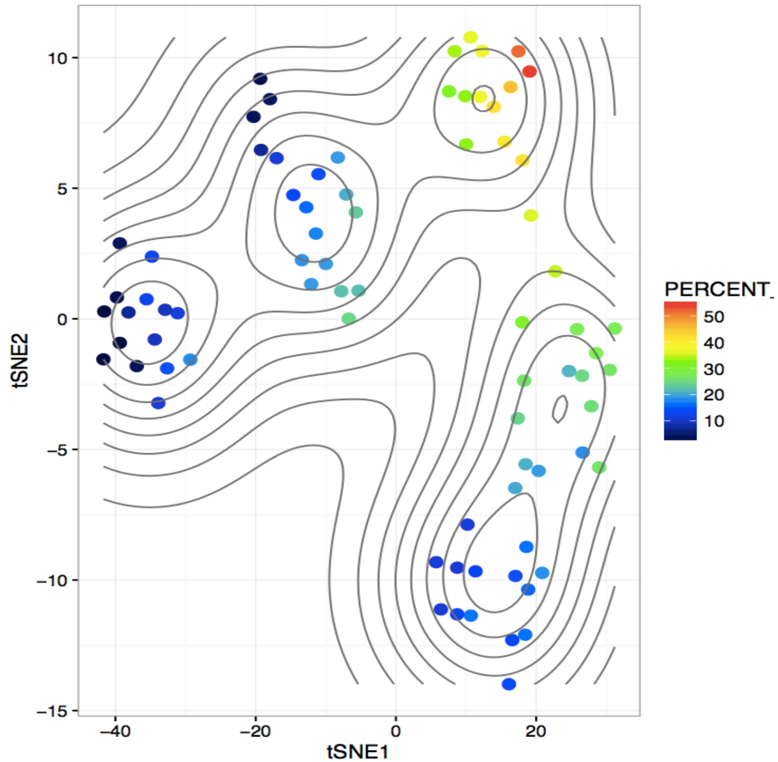


Hardship index of communities

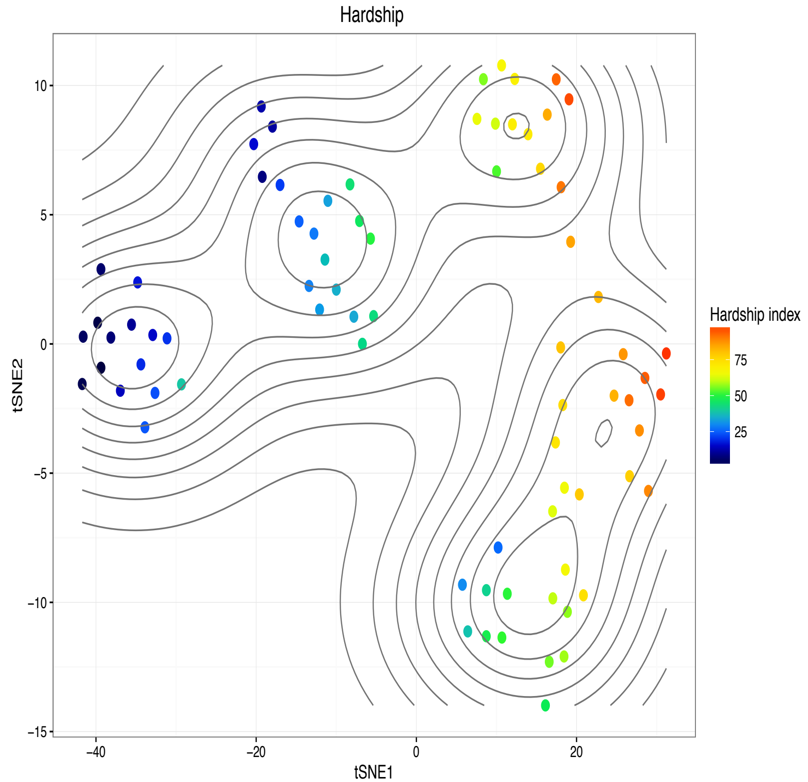


The clusters of the Chicago communities: without diploma and hardship

PERCENT_AGED_25p_WITHOUT_HIGH_SCHOOL_DIPLOMA



Hardship index of communities



Assignments

- Read Chapter 13 of the textbook
- Next time: More on linear regression

Additional References

- Robert V. Hogg, Elliot A. Tanis and Dale L. Zimmerman. “Probability and Statistical Inference”
- Kelvin Murphy, “Machine learning, A Probabilistic perspective”

See you next time

See You!