B4M36SAN BE4M36SAN Seminar/Tutorial

Alikhan Anuarbekov

Ali

anuarali@fel.cvut.cz

- You have 3 weeks left to form a team (deadline 11.11.2024)
  - 4 people maximum
  - We strongly recommend you start forming and preparing for Final assignment

- If you have any problems or questions regarding the Final assignment NOW, feel free to write an email / arrange the consultation

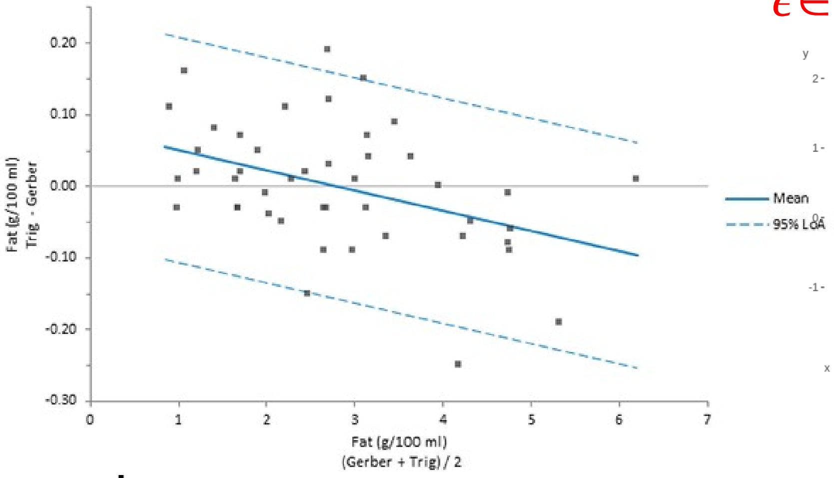
- After the deadline you won't have enough time to discuss your topic!
  - 1 week to come up with an idea, 2 week to prepare a detailed plan

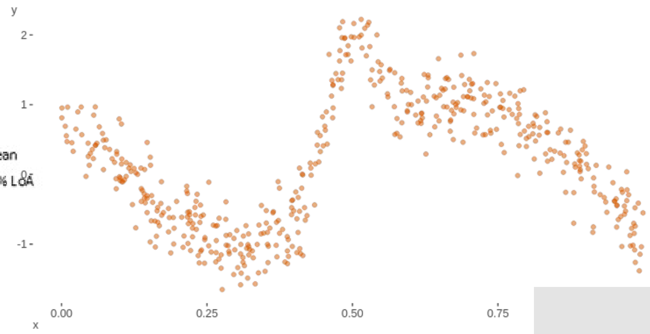
- Also, after 3 weeks you will have a midterm test
  - You can prepare by looking into these files (Courseware – SAN - exam)

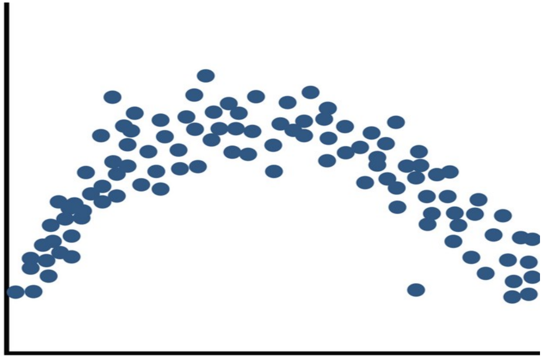
5. Solved exercises:
   - san_solved.pdf.
6. Sample questions pertaining to prerequisites:
   - stat_min.pdf, stat_min_eng.pdf.

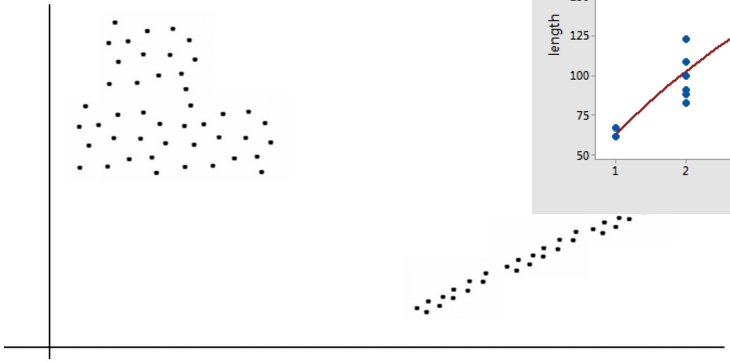
- **Up until now we have used a variety of models and tricks that allowed us to predict the response/dependent Y variable from a set of independent predictor X variables**

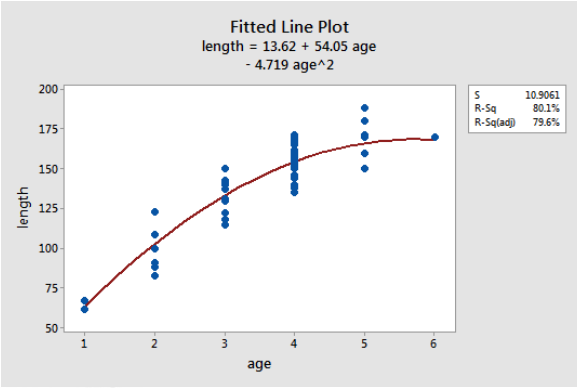
$$Y=\beta_0+\beta_1\cdot X_1+\beta_2\cdot X_2+\ldots+\beta_P\cdot X_P+\epsilon,$$
$$\epsilon\in N\left(0,\sigma^2\right)$$



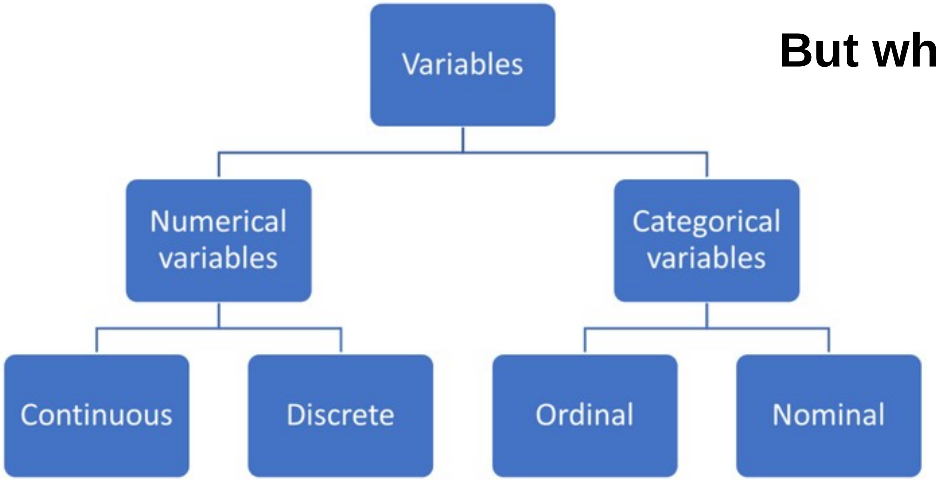




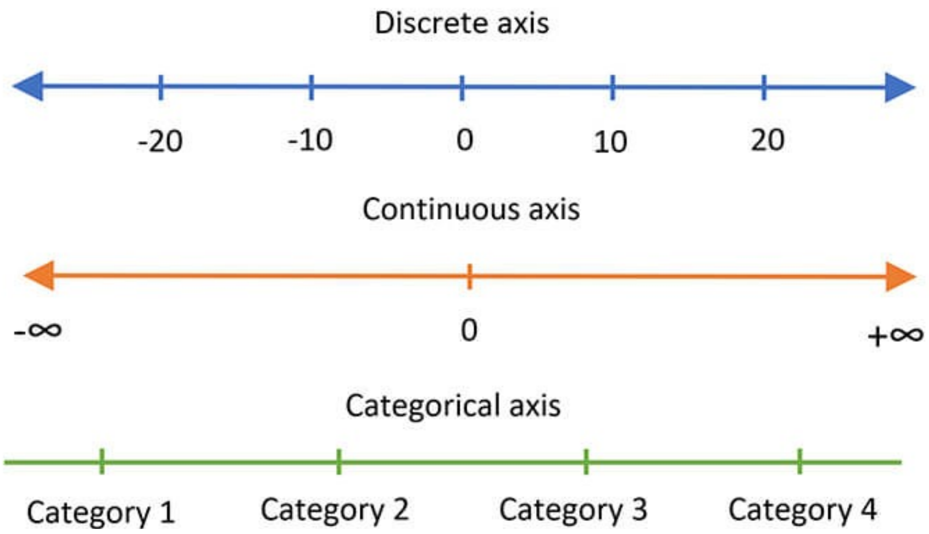




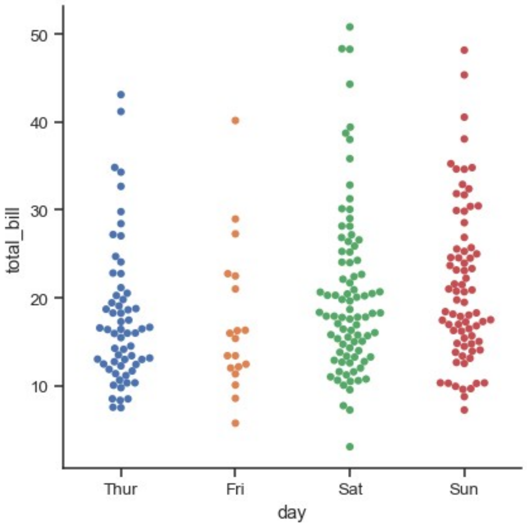
- **However, the problem is that in all of previous cases, we assumed that response variable Y is not constrained and is a continuous variable….**

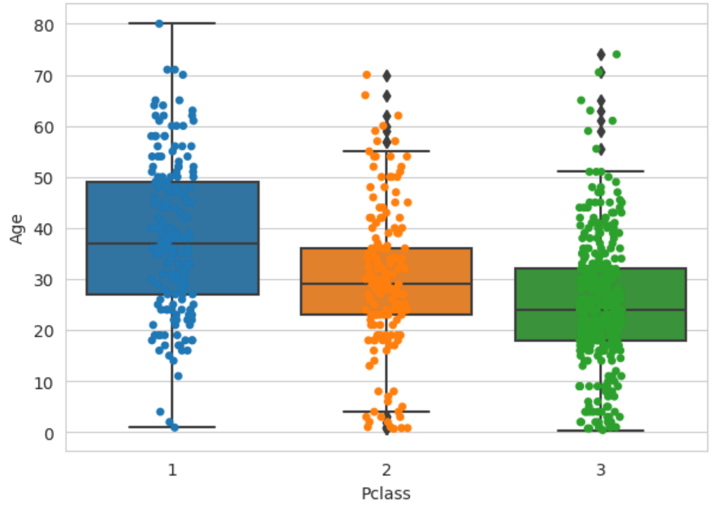


But what if it is not the case anymore?

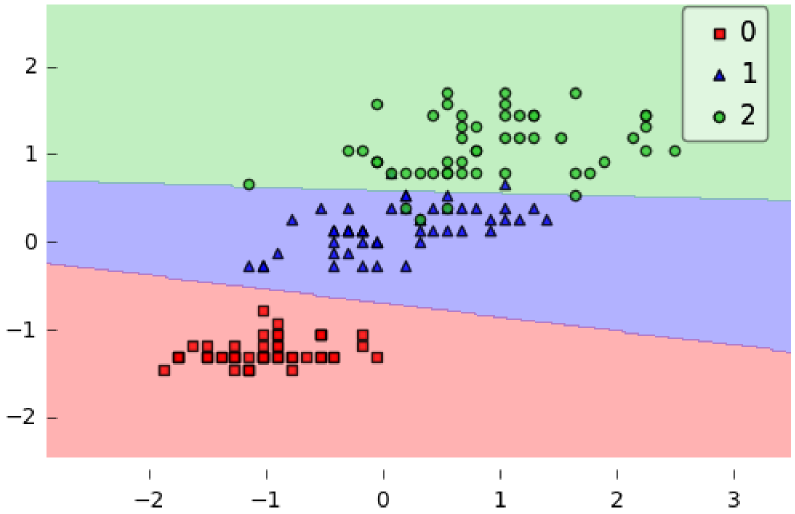


- **For simplicity, last seminar you were working with categorical predictors**





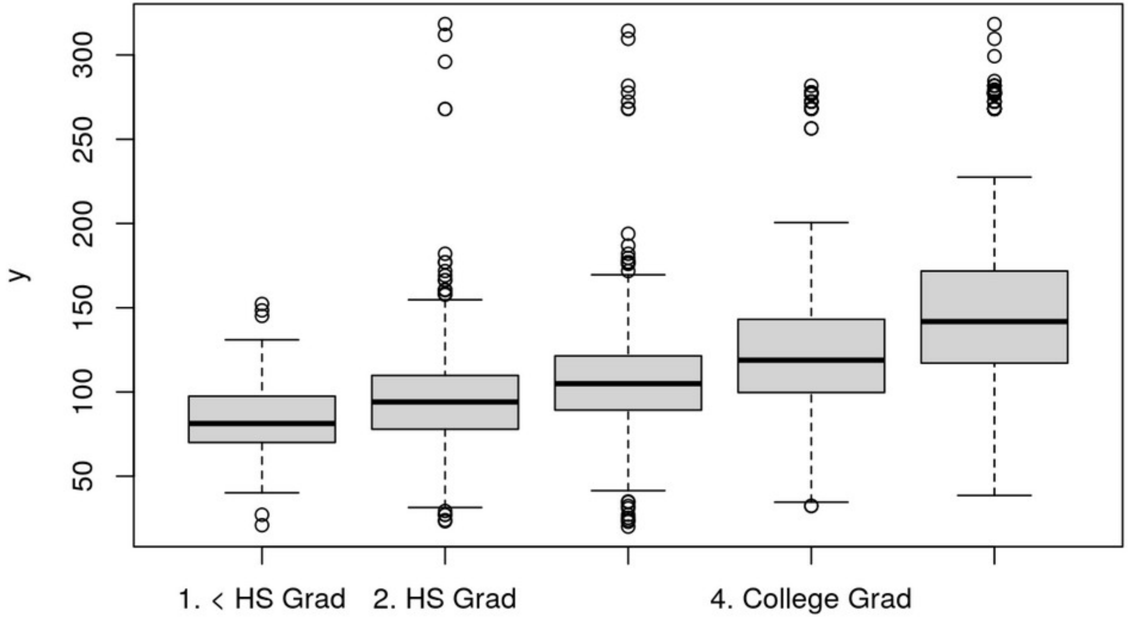
But what if your task was actually to predict the category, not the value? (rotate the graph)



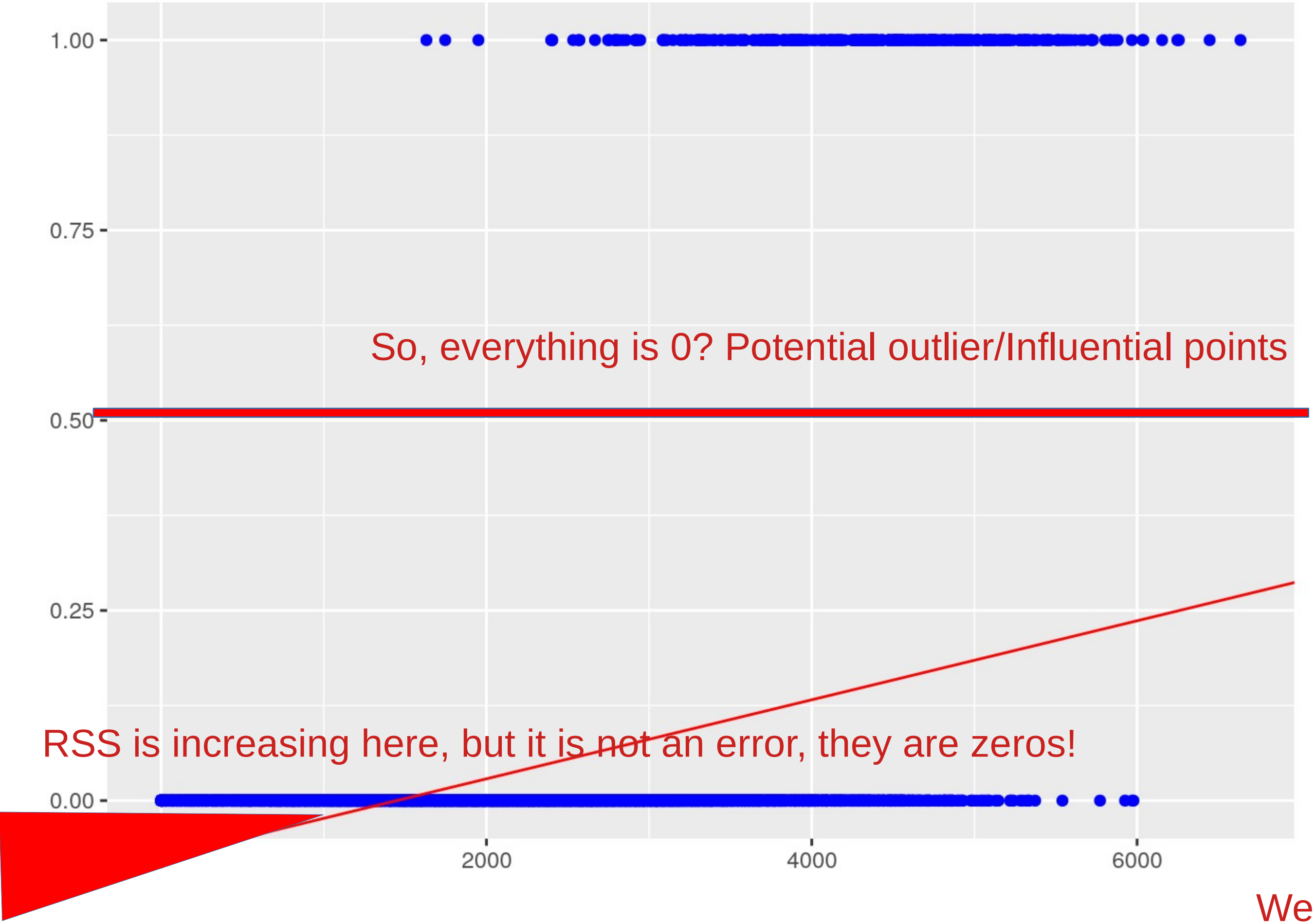
- **Some practical example: consider the same Wage dataset you have been working with last time and two variables = Wage and Education**
- **We will flip the task**

Example: assume you work in an advertisement agency and you need to tune the target advertisement. You collected the data from online shop and estimated the wage that customers have based on their purchases.

Now: your advertisement is targeted for different age and education groups (Scholars people like comic books/games, Students watch Netflix and Adults watch Football You want to predict the group based on your collected data to use targeted advertisement



- **Simple idea = use the same linear regression (lets say with 2 groups):**



We can't just say that:

$$\hat{y}\left(5000\right)=0.23$$

Our Y is categorical, it can only have limited number of values: 0 and 1

We need to define a procedure to classify based on the 0.23 value

Maybe:

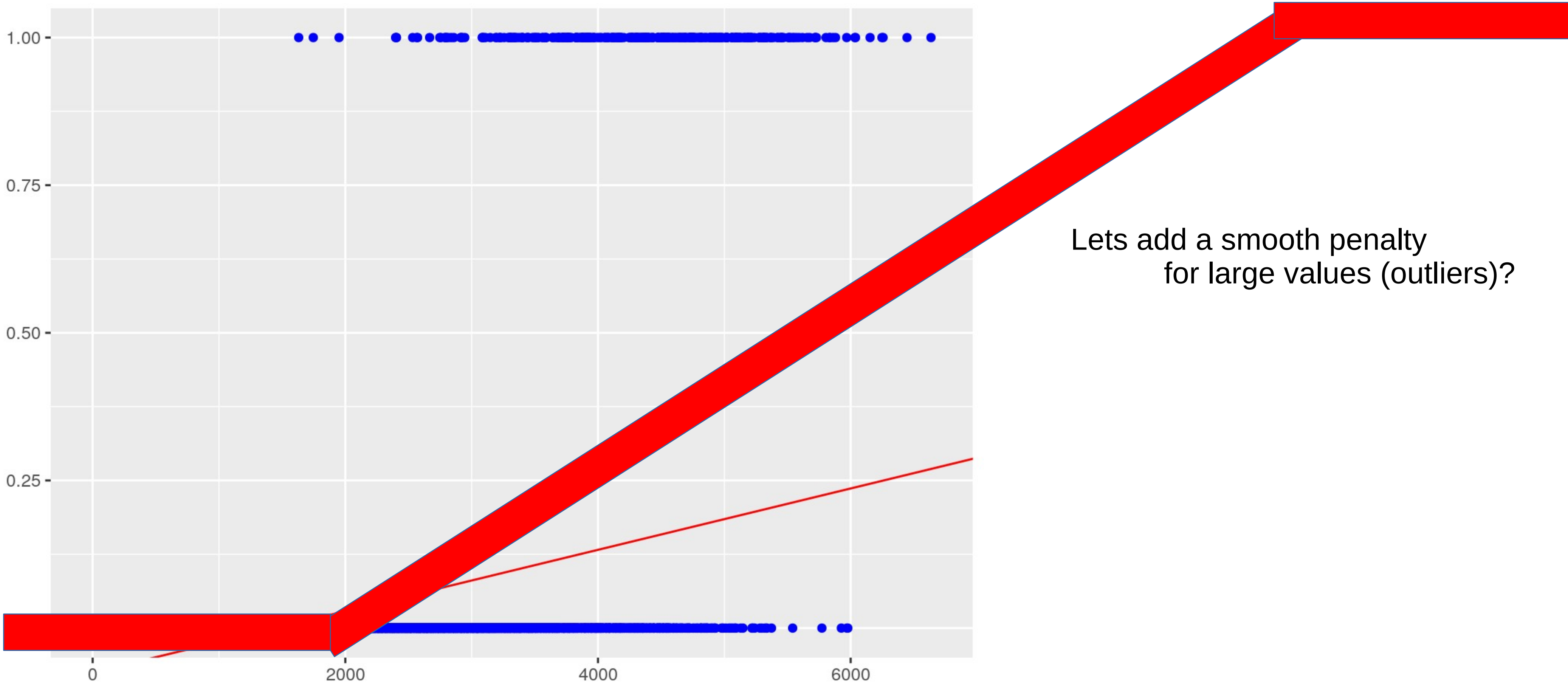
0 if $y < 0.5$

1 if $y > 0.5$

We can't just apply linreg, any solution?

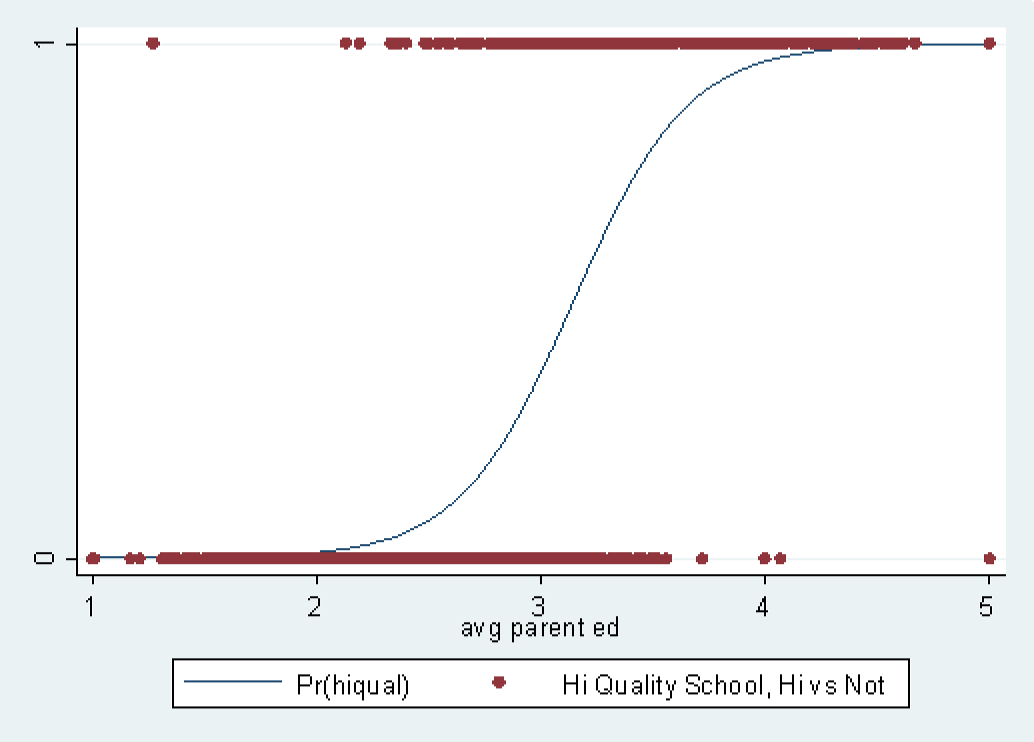
- **Correct RSS penalization when correct:**

Maybe cutoff beyond 0 and 1? A little bit better, but still suffers from outliers and no optimum can be find….



Lets add a smooth penalty
for large values (outliers)?

- **To mitigate extreme (even not outlier) values we can add weight by distance**



$$\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$

It works now! We applied a **link function**

Name	Link Function	Mean	Range of Mean
Identity	$z = \mu$	$\mu = z$	$(-\infty, +\infty)$
Log	$z = \log(\mu)$	$\mu = \exp(z)$	$(0, +\infty)$
Inverse	$z = 1/\mu$	$\mu = \frac{1}{z}$	$(-\infty, +\infty)$
Inverse Squared	$z = 1/\mu^2$	$\mu = \frac{1}{\sqrt{z}}$	$(0, +\infty)$
Square root	$z = \sqrt{\mu}$	$\mu = z^2$	$(0, +\infty)$

If link function applied = **Generalized Linear Models (GLMs)**
glm = g & lm = link function & lm

this particular model is called **Logistic regression**

Why does it work with linear regression? Because it removes the boundaries on Y!

But what does it changed about our regression line?

$$\hat{y}(x)=\frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \qquad \frac{e^{\beta_0+\beta_1X_1+\ldots\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\ldots\beta_pX_p}}=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\ldots\beta_pX_p)}}$$

$$\textit{then}: \quad 1+e^{-(\beta_0+\beta_1 x)}=\frac{1}{\hat{y}(x)}$$

$$e^{-(\beta_0+\beta_1 x)}=\frac{1}{\hat{y}(x)-1}=\frac{1-\hat{y}(x)}{\hat{y}(x)}$$

$$-(\beta_0+\beta_1 x)=\ln\left[\frac{1-\hat{y}(x)}{\hat{y}(x)}\right]$$

$$(\beta_0+\beta_1 x)=-\ln\left[\frac{1-\hat{y}(x)}{\hat{y}(x)}\right]=\text{negative log is invert}=\ln\left[\frac{\hat{y}(x)}{1-\hat{y}(x)}\right]$$ **Log odds**

We just converted: $y_{original}\in[0,1]$ ➔ $y_{logit}\in[-\infty,\infty]$ **Logit transform**

A little numeric problem with logistic regression (mainly for more X predictors)

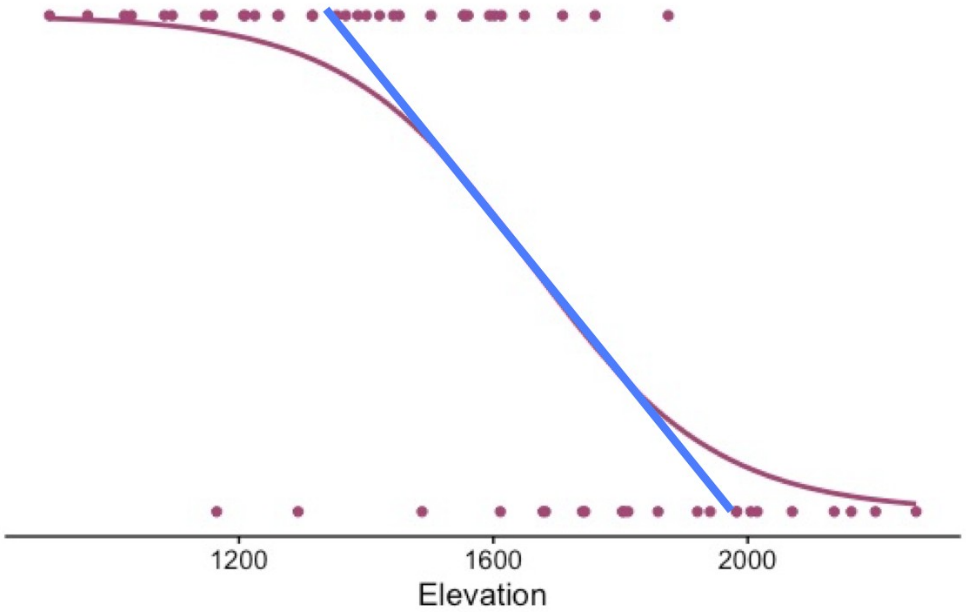
But how do we deal with infinity-far points?! Just do **Gradient Descent**!

> Still converges to the global minimum because it is convex(concave)

$$\ln\left[\frac{1}{1-1}\right]=\ln[\infty]=\infty$$

$$\ln\left[\frac{0.5}{1-0.5}\right]=\ln[1]=0$$

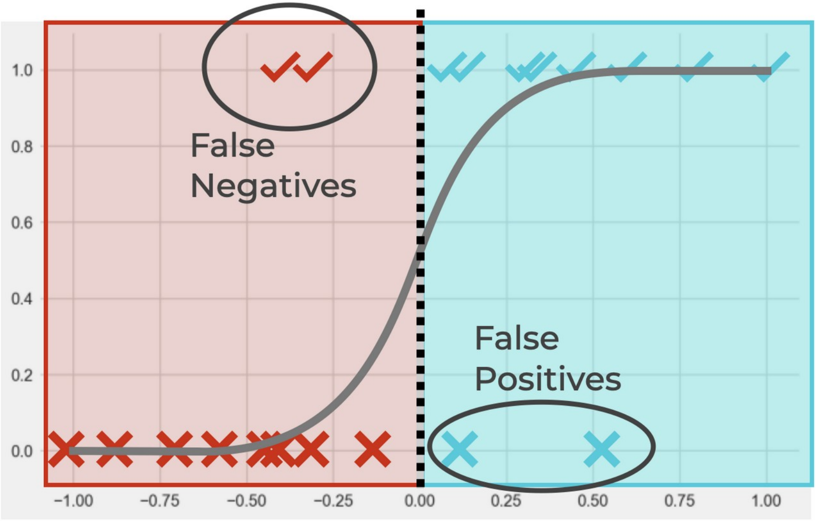
$$\ln\left[\frac{0}{1-0}\right]=\ln[0]=-\infty$$

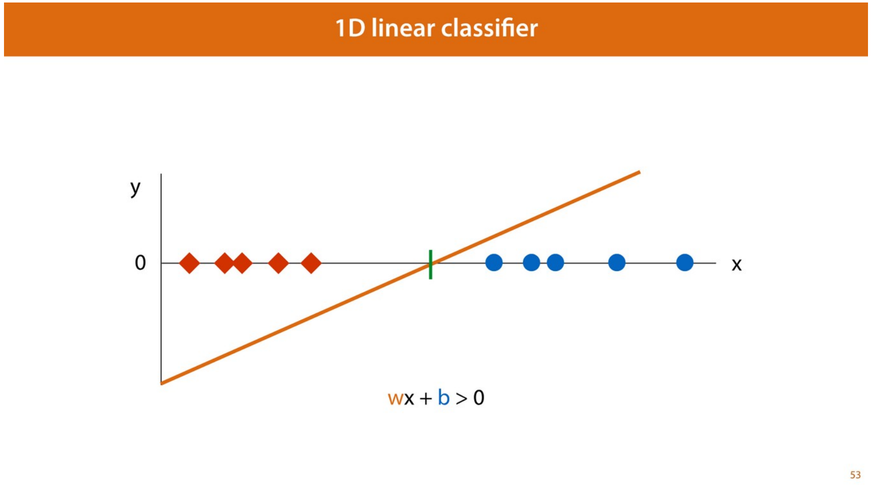


$$\ln\left[\frac{0.5}{1-0.5}\right]=0=\beta_0+\beta_1 x \quad \rightarrow \quad x_{0.5}=\frac{-\beta_0}{\beta_1}$$

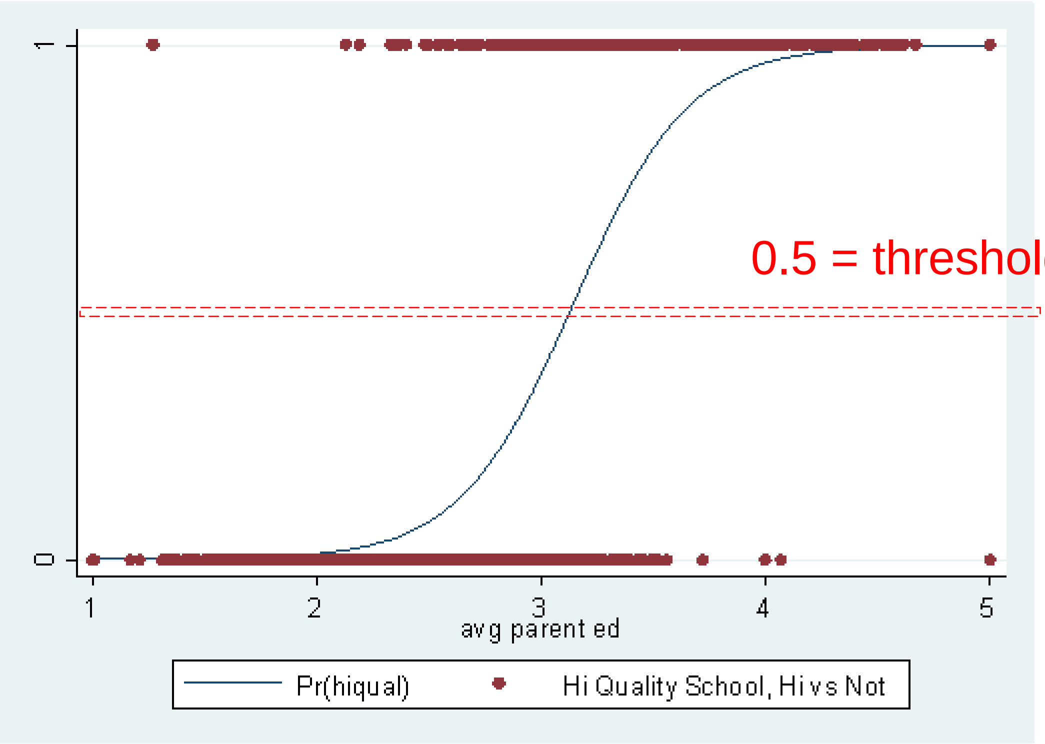
- Up until now we have not named the task properly. Actually, it is a well known **classification task**:

  - You may have seen its version in the right form with linear classifier





Finally, how to decide about cutoff threshold?



0.5 = threshold

	Coefficient	Std. error	Z-statistic	p-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
balance	0.0057	0.0002	24.74	< 0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	-0.6468	0.2362	-2.74	0.0062

$$\hat{y}(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} = 0.5$$

> When doing a categorical (logistic) regression,
we assume that we get enough samples,
(only 2 groups, not much samples needed)

e.g. T-test is converging to N(0,1):
normalized Normal distrib = Z-test

> In other words = you can just use p-values from table!

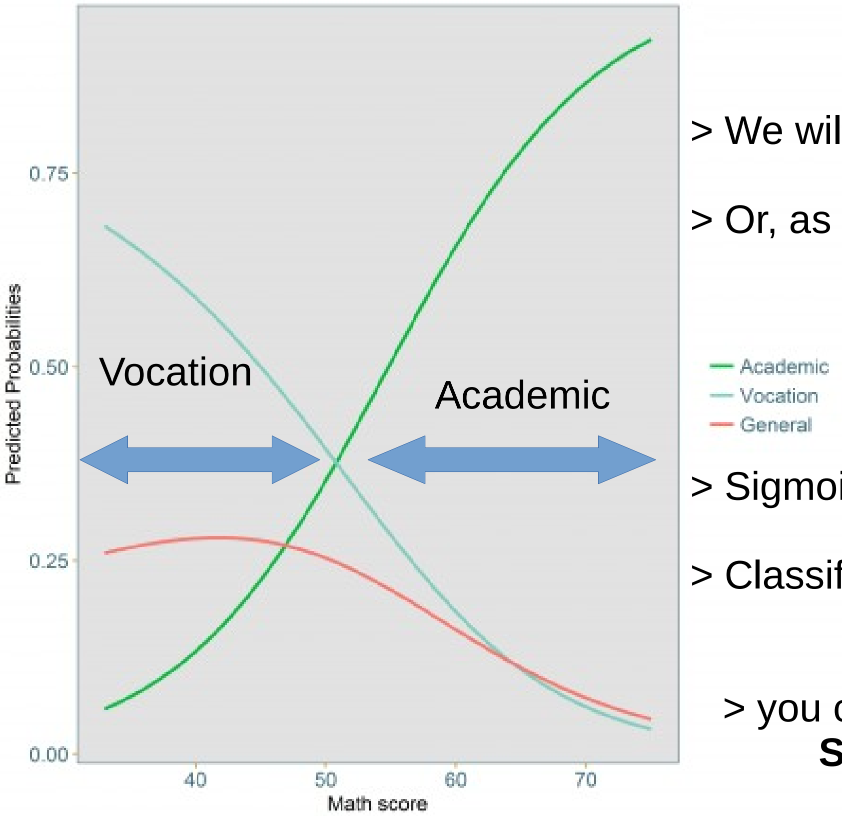
$$\beta_0+\beta_1 x = \ln\left[\frac{\hat{y}(x)}{1-\hat{y}(x)}\right] = \ln\left[\frac{0.5}{1-0.5}\right] = \ln 1 = 0$$

- **If you have more than 2 groups, just do multiple logistic regressions**

Multinomial regression:

- e.g., $ethnicity \in$ {Asian, Caucasian, African American} could be captured by

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases} \qquad x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$



> We will learn K-1 logistic regression curves = K-1 Linear (logistic) regression

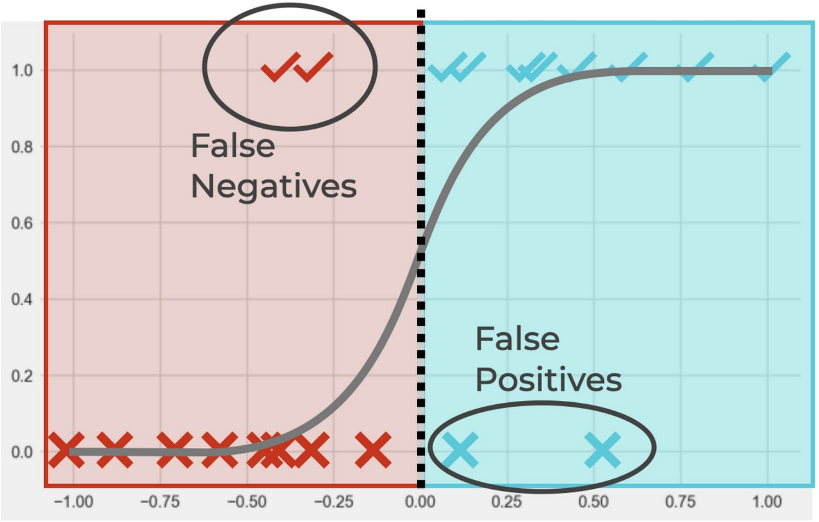
> Or, as in the image on the left, can do easier K logistic regression
(1 categ x others) x K

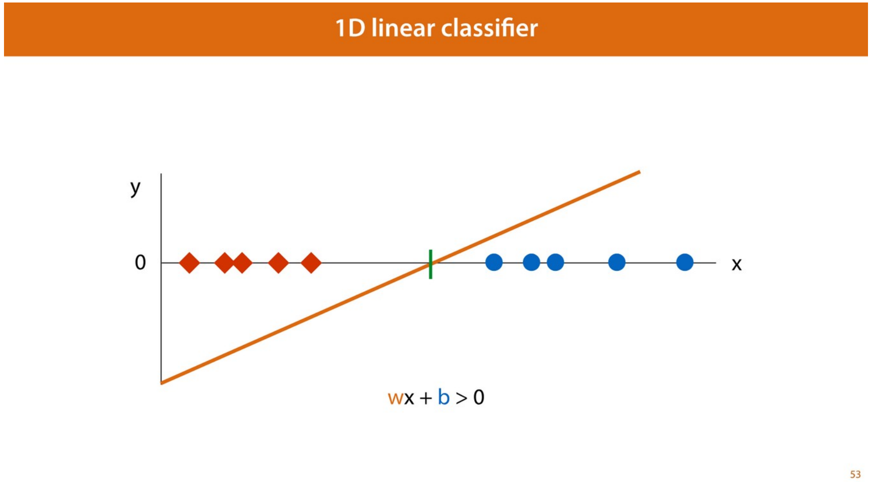
> Sigmoid can be inverted if the trend is decreasing! (**Vocation/General**)

> Classification is that we choose the **largest log odds** out of K

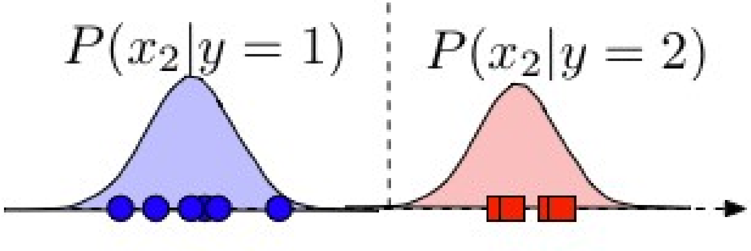
> you can see a problem, we don't predict **GENERAL** at all!
Some other alternative model?

- **Remember first seminar?**





- **Model as distribution directly:**

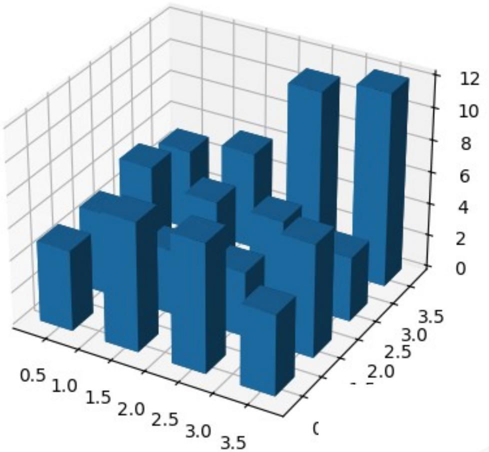


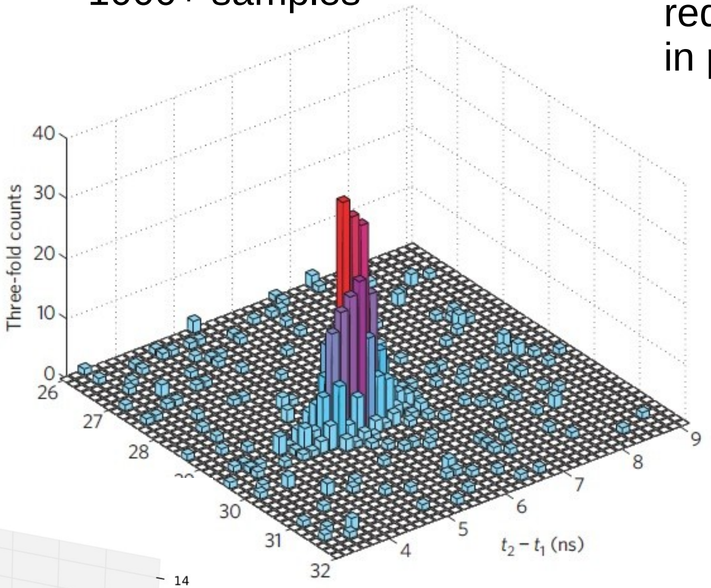
Problem of distribution-only modeling for more variables

CURSE OF DIMENSIONALITY!

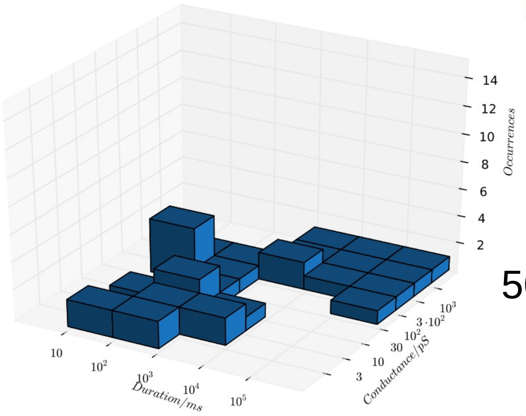
1000+ samples

1D Gaussian normal required just > 30 samples in practice





300 samples

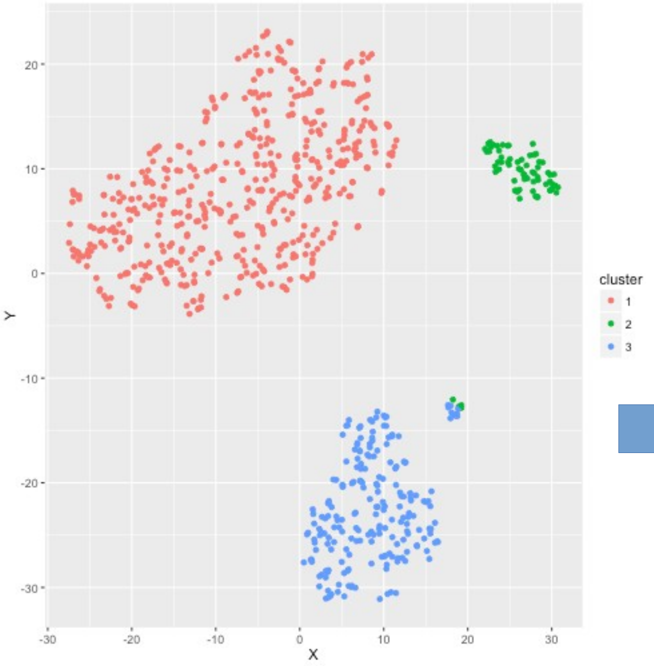


50 samples

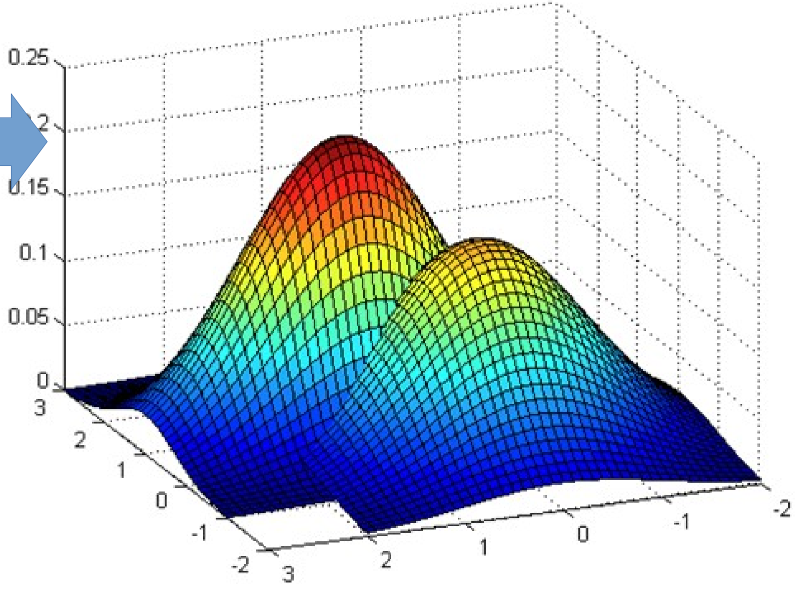
But what if we actually have enough samples?

Discrimination Analysis

But what if we have a categorical Y?



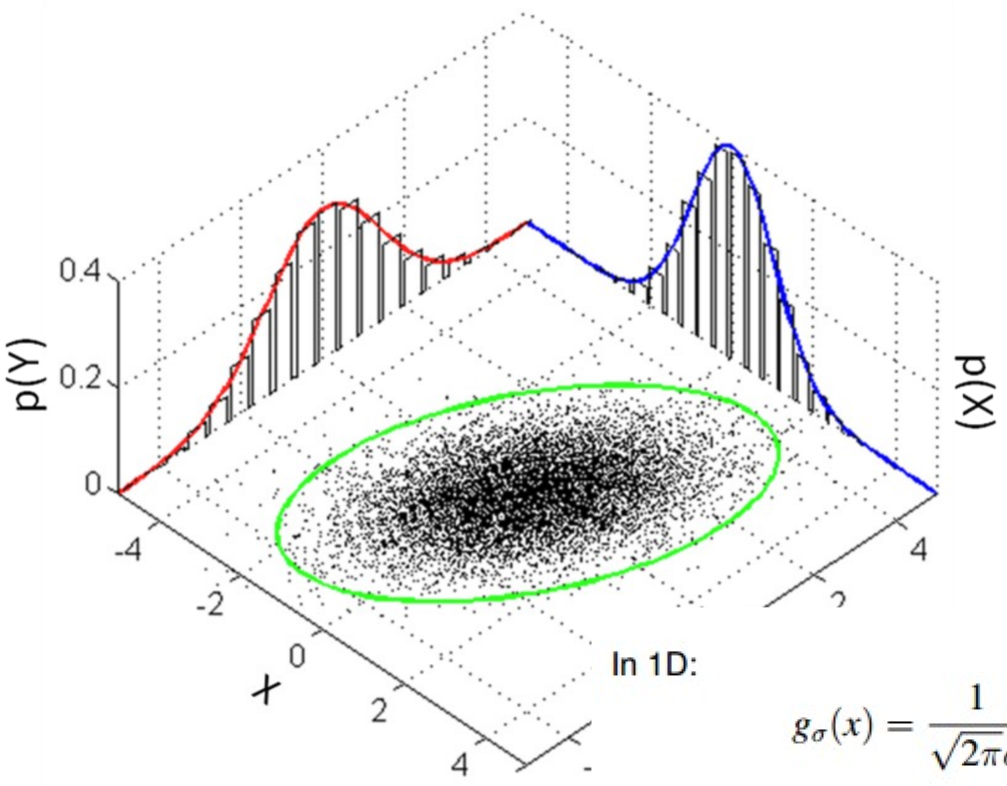
Model as high dimensional distribution
(lets say **Gaussian normal**
very strong assumption!)
for every category (remember multinomial):

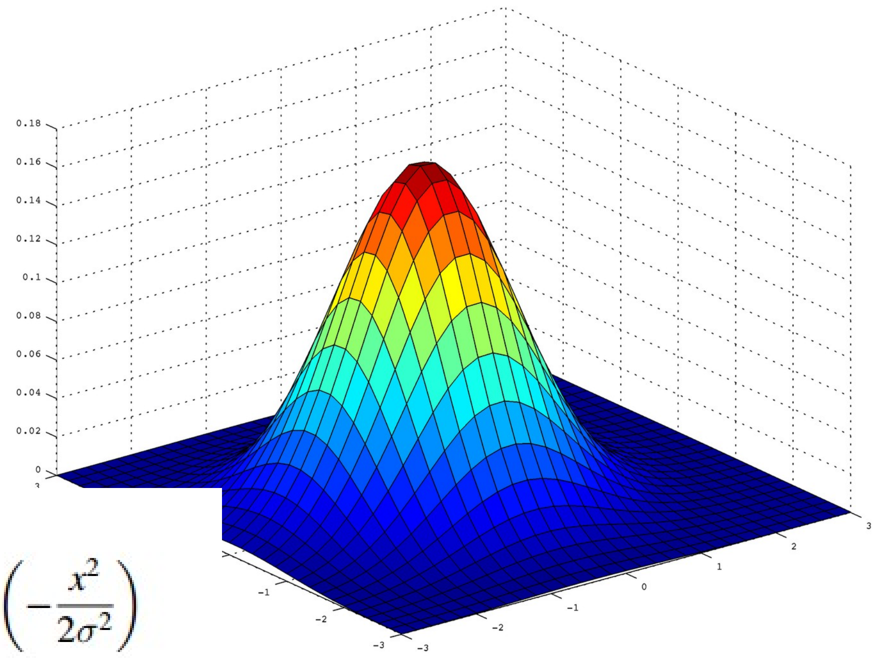


H = Height, **W** = Width,
C=Color
All are random variables

I can see that each color of cat has significantly different shape

Multivariate Gaussian Normal:





In 1D:

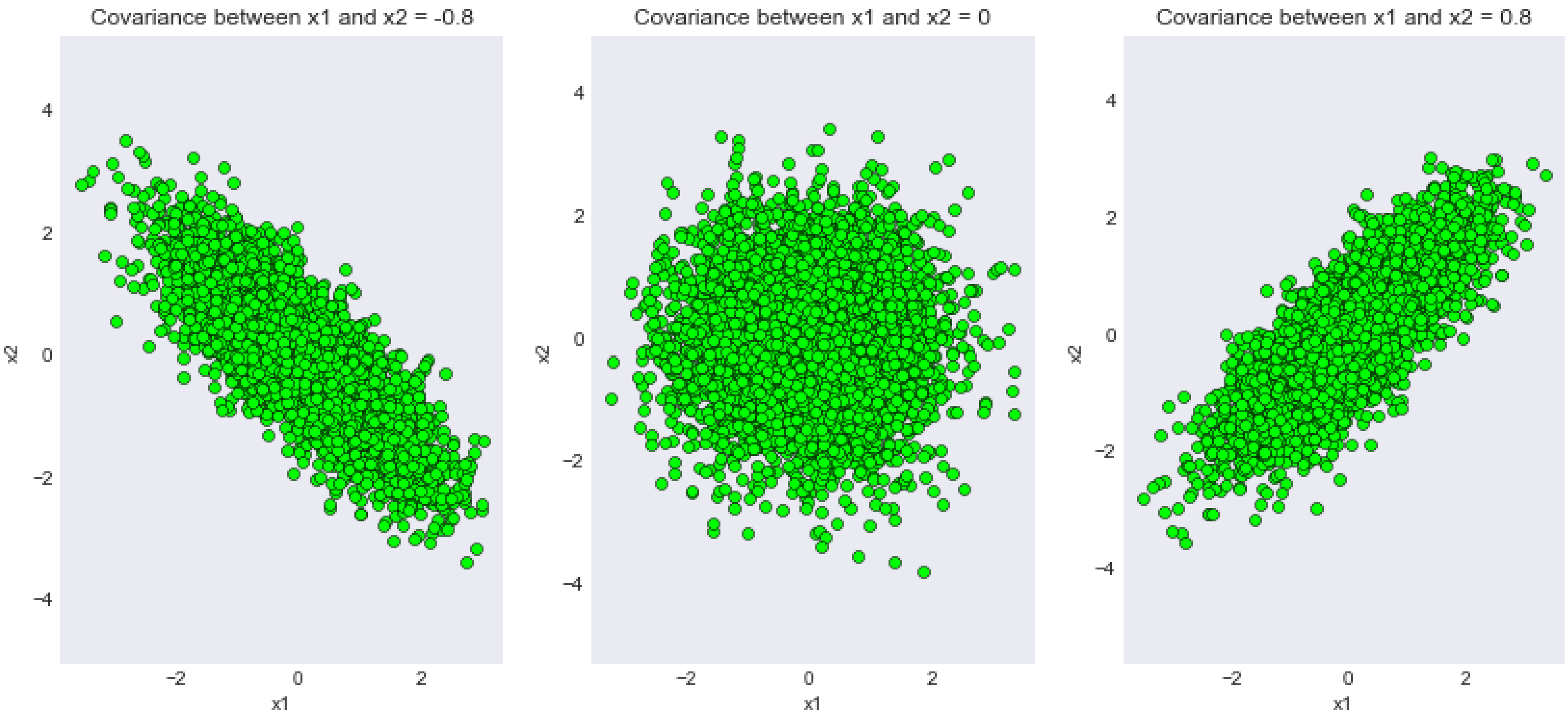
$$g_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

In 2D:

$$G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right)$$

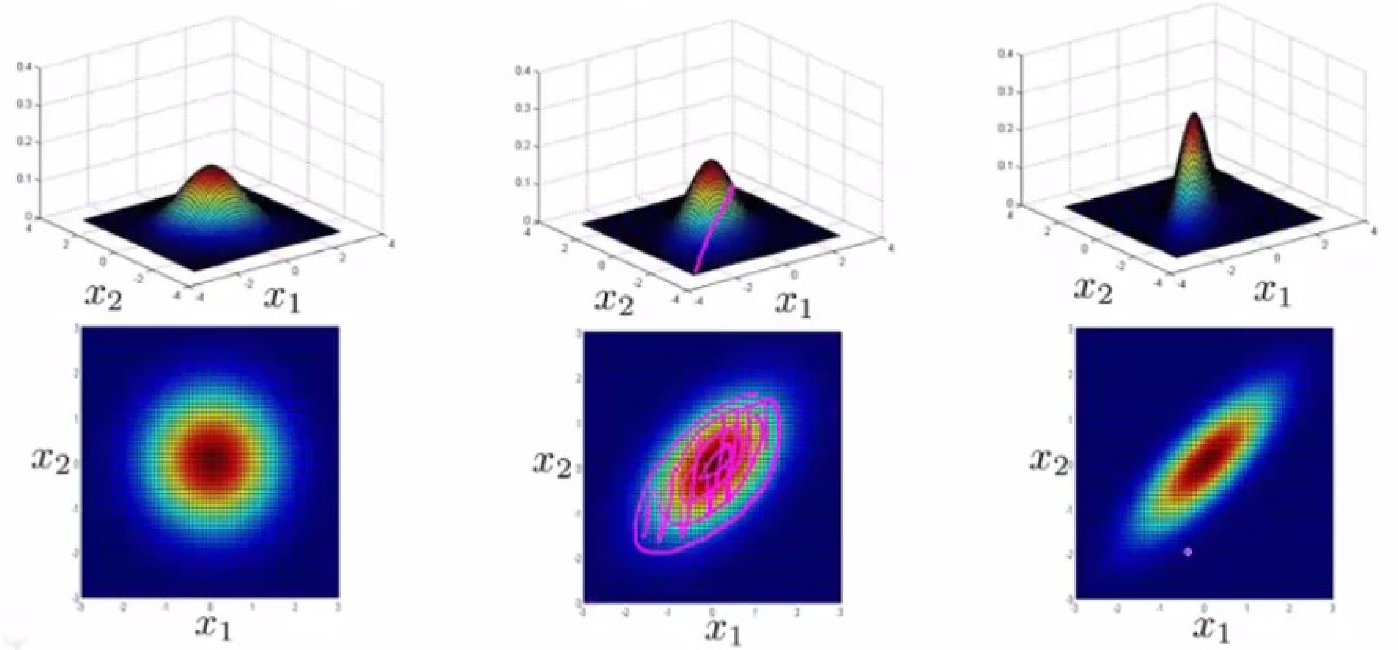
* Ne vzdycky!

Multivariate Gaussian Normal:



Multivariate Gaussian (Normal) examples

$\mu = \begin{bmatrix}0\\0\end{bmatrix} \; \Sigma = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$ $\quad$ $\mu = \begin{bmatrix}0\\0\end{bmatrix} \; \Sigma = \begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}$ $\quad$ $\mu = \begin{bmatrix}0\\0\end{bmatrix} \; \Sigma = \begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}$



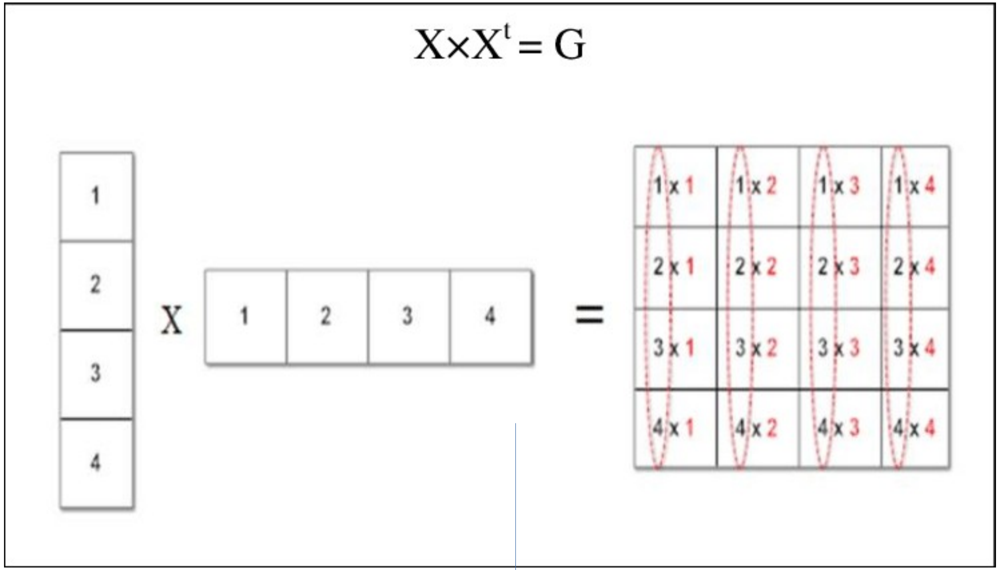
Andrew

Example: the Bivariate Normal distribution

$$f(x_1, x_2) = \frac{1}{(2\pi)|\Sigma|^{1/2}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})' \Sigma^{-1} (\vec{x}-\vec{\mu})}$$

with $\vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ and

$$\underset{2\times 2}{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$



In more than 1D now our formulas become in matrix form

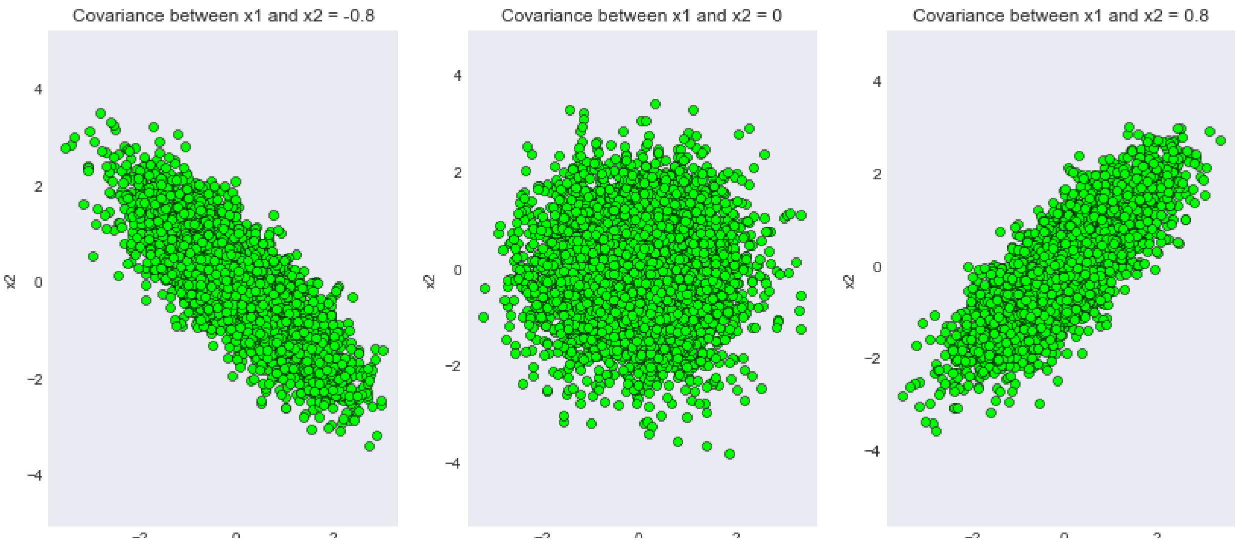
But this is the same formula!

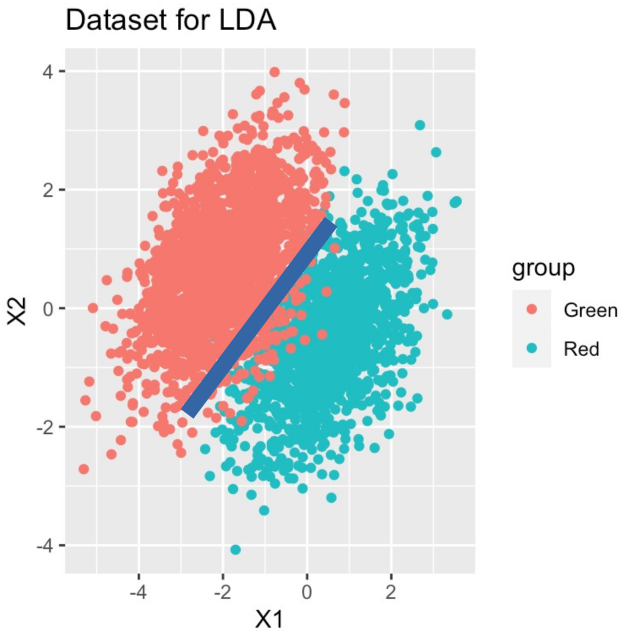
Import Result (MLE of μ and Σ)

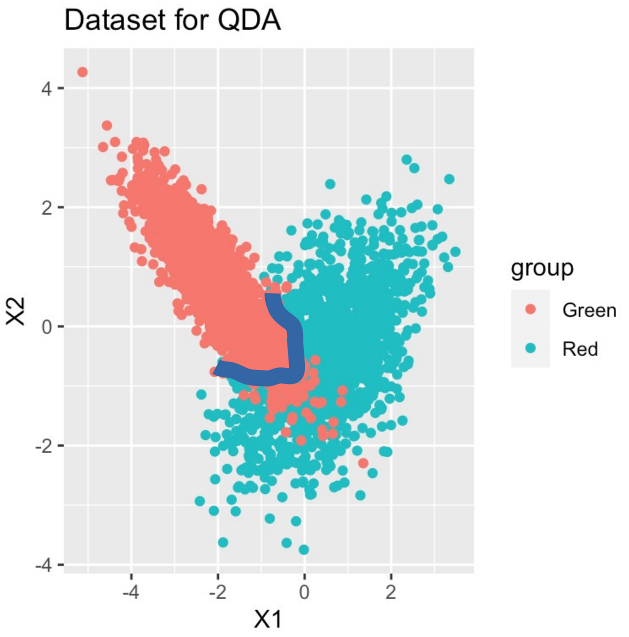
Let $X_1, X_2, \ldots, X_n \sim N(\mu, \Sigma)$ be a random sample from a multivariate normal population. Then,

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\Sigma} = \frac{\sum_{j=1}^{n} (X_j - \bar{X})(X_j - \bar{X})^T}{n} = \frac{(n-1)S}{n}$$

are the maximum likelihood estimators of μ and Σ, respectively,







Linear classification curve

- **Estimate covariance matrix once**

$$LDA:$$
$$\sum_1 = \sum_2$$

$$QDA:$$
$$\sum_1 \neq \sum_2$$

Polynomial classification curve

- **Estimate covariance matrix twice**

Bayesian theorem:
AGAIN! Like in linear regression, we have an assumption of Gaussian normality!

$$Pr(Y = k|\mathbf{X} = \mathbf{x}) = \frac{Pr(\mathbf{X} = \mathbf{x}|Y = k)Pr(Y = k)}{Pr(\mathbf{X} = \mathbf{x})}$$

when we use normal (Gaussian) distributions for each class

- this option leads to **linear or quadratic discriminant analysis**

$$Pr(Y = k|\mathbf{X} = \mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{j=1}^{K} \pi_j f_j(\mathbf{x})}$$

- where $f_k(\mathbf{x}) = Pr(\mathbf{X} = \mathbf{x}|Y = k)$ is the density for $\mathbf{X}$ in class k,
- where $\pi_k = Pr(Y = k)$ is is the marginal or prior probability for class k.

> Pr(Y = k) is scale coefficient based on number of samples = more data is larger Gaussian

How to compute the equation for decision boundary (threshold line)?

Use the definition – largest Bayesian posterior probability means that we predict the corresponding group

$$LDA/QDA: for\ a\ given\ \vec{x}=(x_1, x_2)$$

$$\pi_1 \cdot N_1(\vec{x}|\mu_1, \textstyle\sum_1) > \pi_2 \cdot N_2(\vec{x}|\mu_2, \textstyle\sum_2), \text{ then classify x as 1}$$

$$\pi_1 \cdot N_1(\vec{x}|\mu_1, \textstyle\sum_1) < \pi_2 \cdot N_2(\vec{x}|\mu_2, \textstyle\sum_2) \quad \text{then classify x as 2}$$

$$\color{red}{Then:}\ \pi_1 \cdot N_1(\vec{x}|\mu_1, \textstyle\sum_1) = \pi_2 \cdot N_2(\vec{x}|\mu_2, \textstyle\sum_2) \quad \text{is a decision boundary}$$

$$f(x_1, x_2) = \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^T |\Sigma|^{-1} (\vec{x}-\vec{\mu})\right]$$

or identical:
$$\frac{\pi_1 N_1(\vec{x}|\mu_1, \sum_1)}{\pi_2 N_2(\vec{x}|\mu_2, \sum_2)} = 1$$

$$(x-\mu_1)^T \textstyle\sum_1^{-1} (x-\mu_1) - (x-\mu_2)^T \textstyle\sum_2^{-1} (x-\mu_2) = C$$

$$C = \log[|\textstyle\sum_0|] - \log[|\textstyle\sum_1|] + 2\log[p_1] - 2\log[p_0]$$

$$\pi_1/\pi_2 \text{ either computed or are both } \quad \frac{1}{2} = 0.5$$

$$also:\ \pi_1 + \pi_2 = P(y=1) + P(y=2) = 1$$

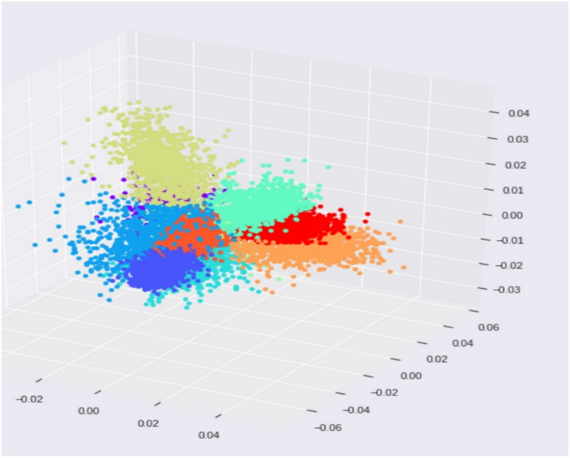
But what if we have more than 2 X variables? We can't even visualize them in dimension larger than 3D

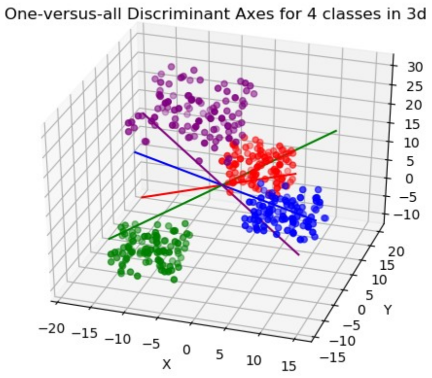
If we assume that all classes have same

$$LDA: \Sigma_1 = \Sigma_2$$

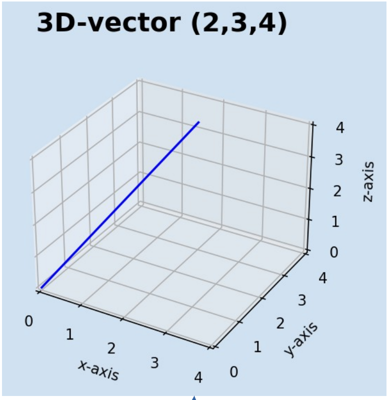
Fisher's discriminant plot

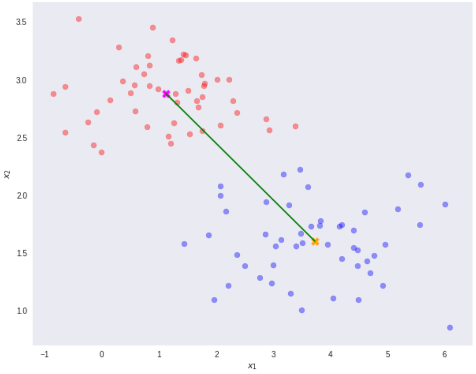
Then we can take every pair and visualize their decision boundary:





Eigenvector formula (linear algebra)





For every pair, take their mean centers
and connect with a vector

Each connection is a line
in D dimensional space

If we project along 1-2 largest eigenvectors,
Then we can get a clean projection in 2D/3D

Assignment 2 = implement this part

Assignment 2: LDA (Fischer) implementation

5	21.10.	JB, AA, JK	Discriminant analysis	san_lda.zip, assignment2.zip, reading for assignment 2

Introduction

The aim of this assignment is to get familiar with Linear Discriminant Analysis (LDA). LDA and Principal Component Analysis (PCA) are two techniques for dimensionality reduction. PCA can be described as an unsupervised algorithm

Small help:

- Read the assignment 2 manual in this tutorial's files
- 2 more weeks to implement

Comparison of algorithms

Logistic regression	LDA	QDA
Weak assumption: Linear threshold	Strong assumption: Normal distribution per group	
Only numeric iterative gradient to optimum	Always has an explicit global solution	
Outlier/Nonnormality are solved	Outlier/Nonnormality breaks the algorithm	
Linear (line) threshold	$LDA:$ $\sum_1 = \sum_2$ Linear (line) threshold	$QDA:$ $\sum_1 \neq \sum_2$ Quadratic (parabola) threshold
Does not matter on group diff	If groups have large differences in sample number, breaks the algorithm	
Has problems with more classes (Missing prediction for some)	Theoretically can model any number of classes with enough samples	

	LDA	**LR**
Assumptions	Normality, Absence of outliers, HOCV, Linearity, Absence of Multicollinearity and Singularity, Independence of observations	No Assumptions (minimal sample size requirement)
Predictor Variables	Continuous	Continuous, Categorical
Outcome Variable	Categorical	Categorical
Decision Rule	Highest Group Score	Cut Score (probability, generally 0.5)

Seminar 5: LDA_weight_height_gender.html

Independent work

1. Which method is expected to work best on test data in this task (LDA, QDA or LR)? Answer without testing first. Use the knowledge of the individual methods assumptions.
2. Experimentally verify your answer. Note that you may need to deal with different (and sometimes very small) train sets to see any difference …

Small help:

- Compare different models (Logistic reg, LDA, QDA) on classification data
  - Same as previous seminar homework
- Re-run with different parameters
- Find parameters that result in needed result in most of runs

- Send by email or in **Bonusy** BRUTE assignment

Seminar 5: LDA_weight_height_gender.html

Big hint: Maybe you can go for a very little training data size (lets say 10-200) to see significant differences