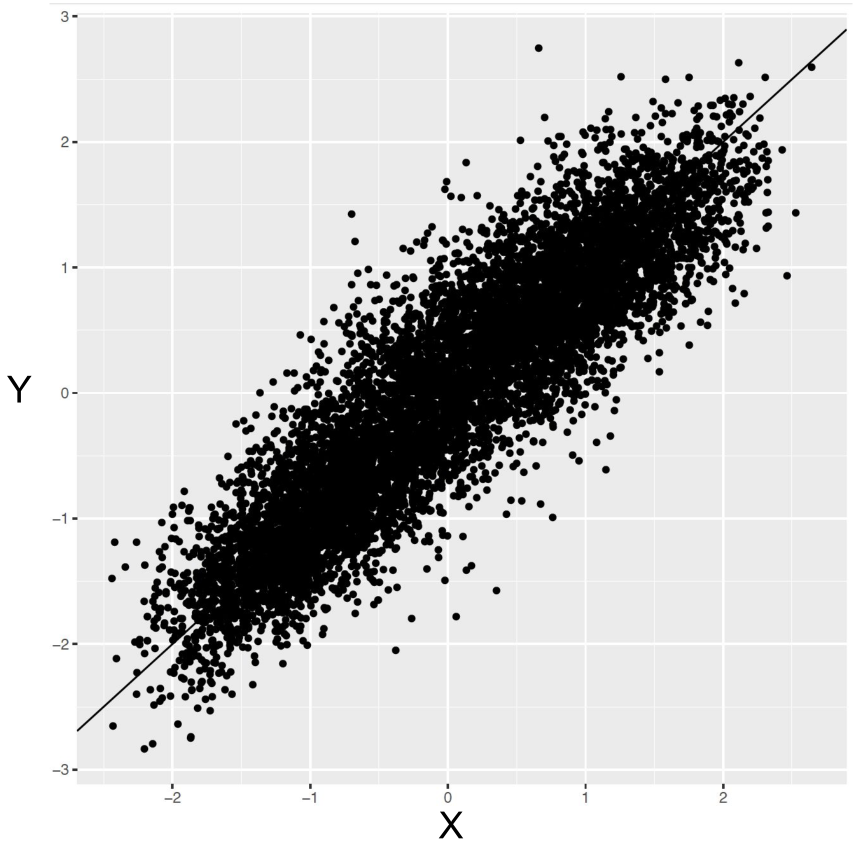
Probabilistic Modeling

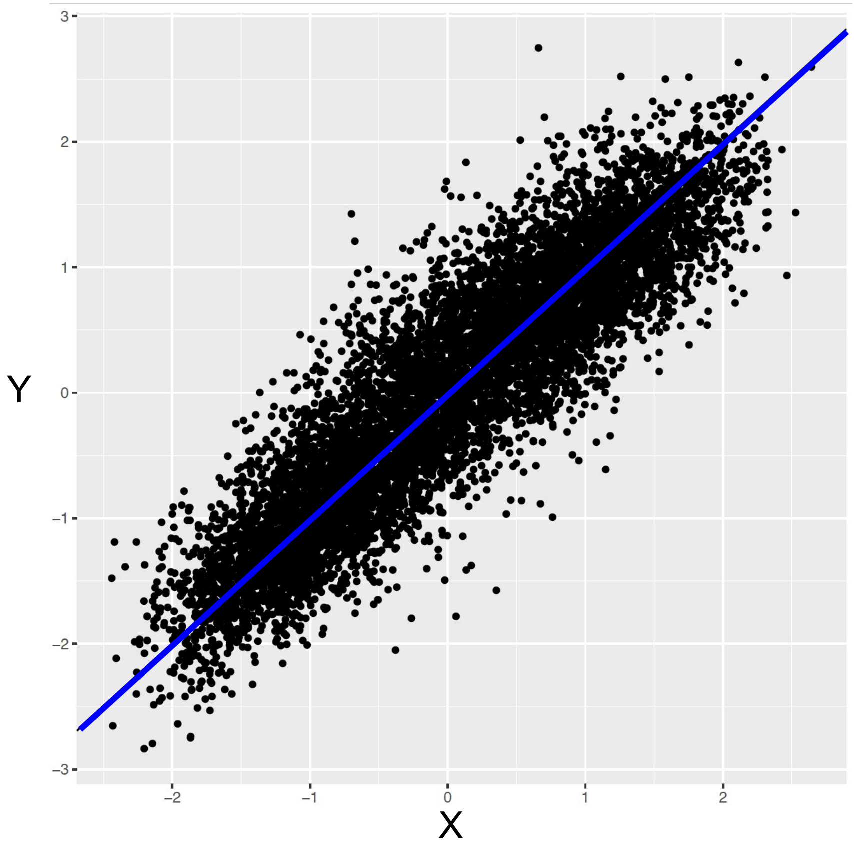
IntroML

Natalie Davidson
natalie.davidson@inf.ethz.ch

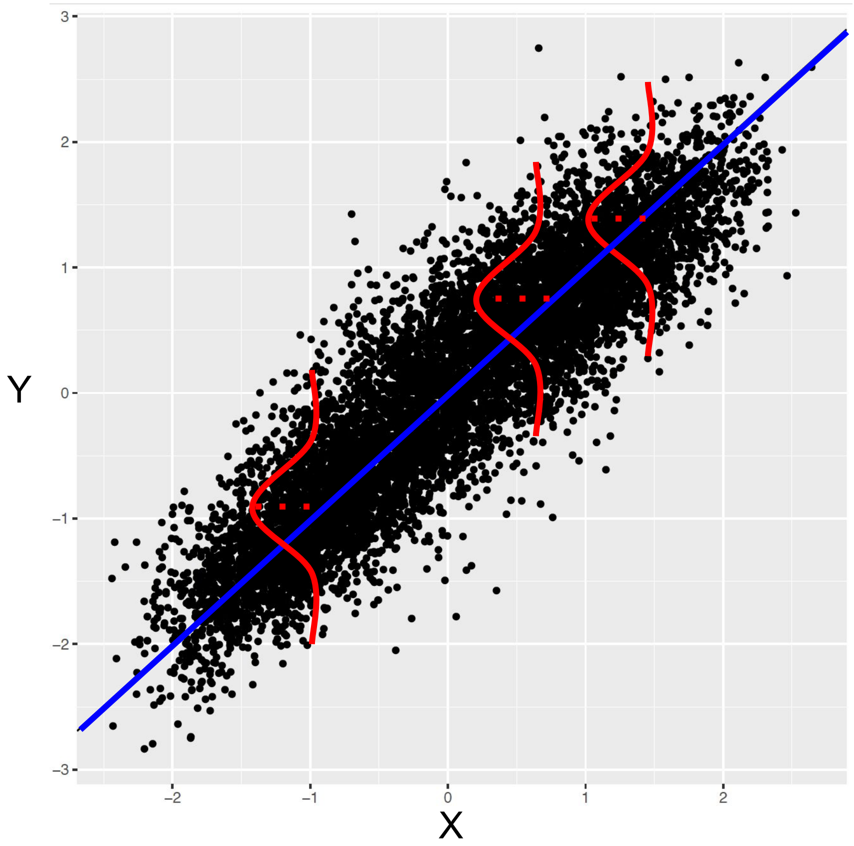
Motivation



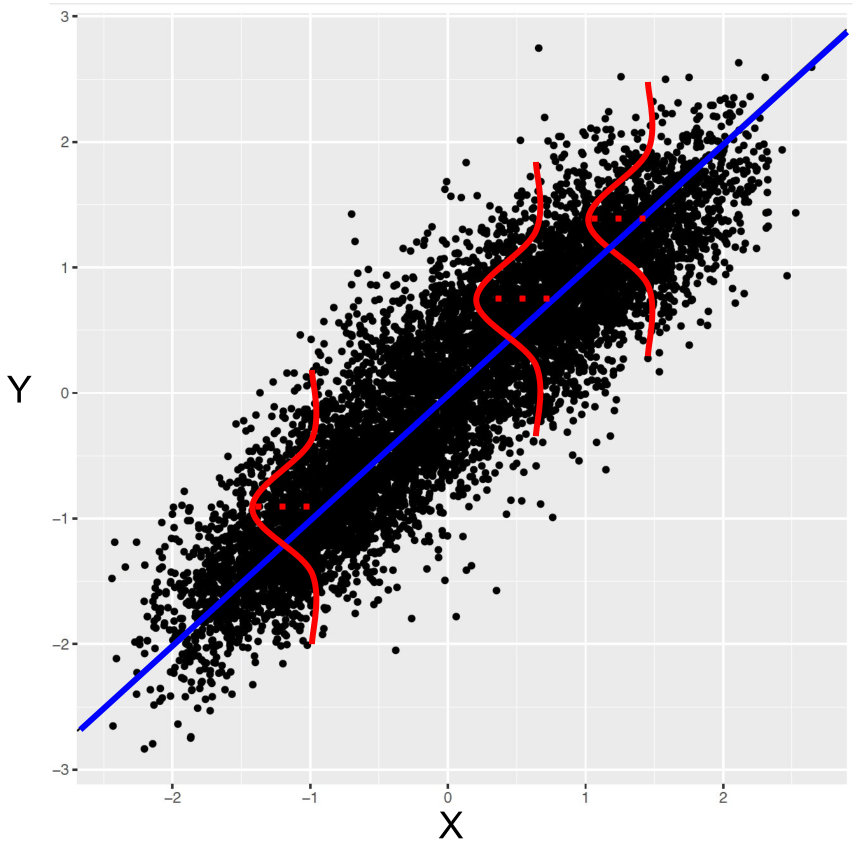
Motivation



Motivation



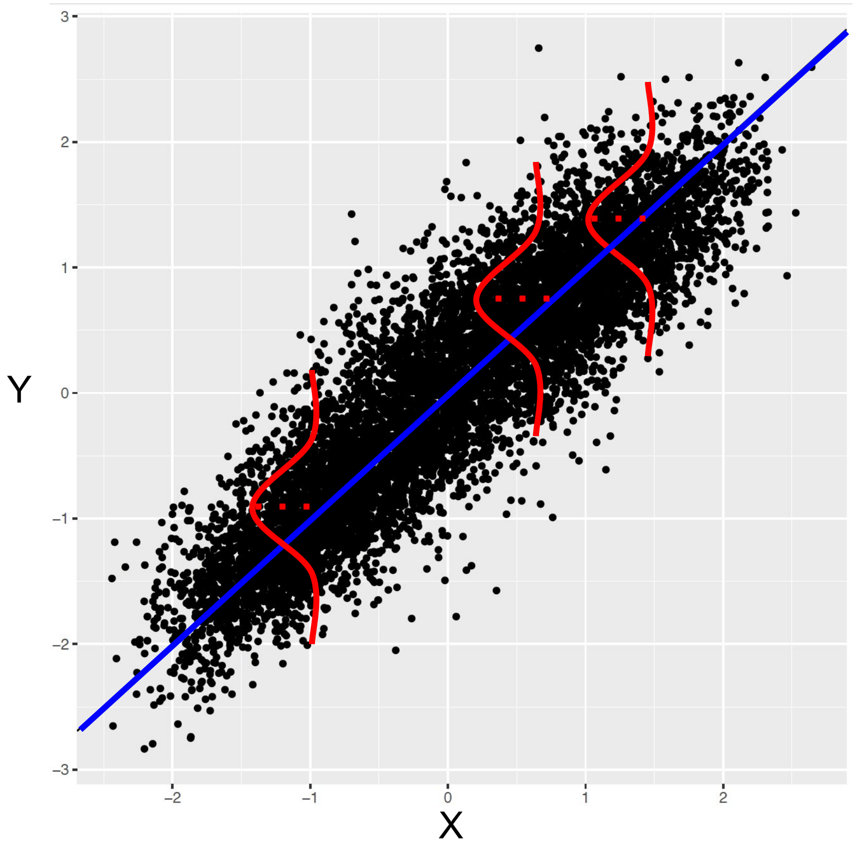
Minimizing least squares error



The hypothesis that minimizes the risk is given by the conditional mean.

---- shown on board ----

Minimizing least squares error

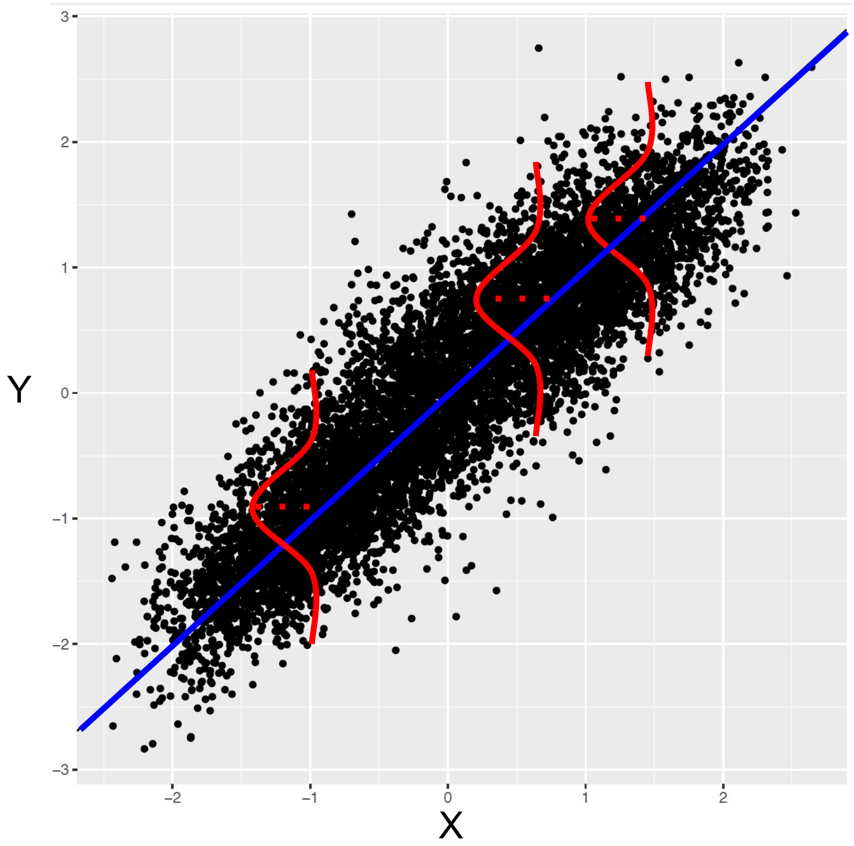


The hypothesis that minimizes the risk is given by the conditional mean.

---- shown on board ----

How do we identify the appropriate conditional distribution? What do you assume it is for this example?

Maximum Likelihood Estimation



The hypothesis that minimizes the risk is given by the conditional mean.

---- shown on board ----

How do we identify the appropriate conditional distribution? What do you assume it is for this example?

Make a statistical assumption about your data to define P(Y|X=x).

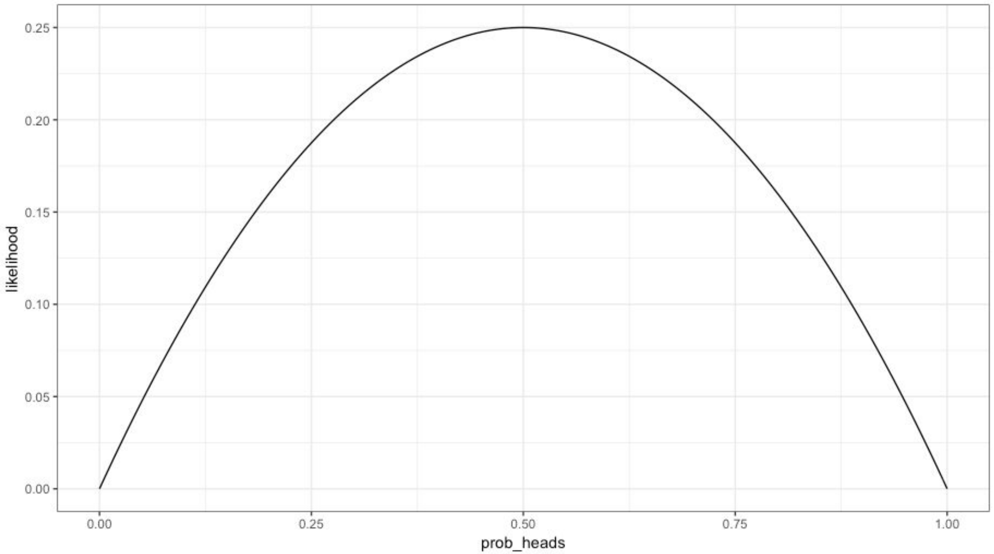
Use MLE to identify most likely parameters for the estimated conditional distribution.

Likelihood Function

- The likelihood gives the probability of the observations given the parameters.
  - $P(Y=y \mid \Theta, X=x)$
- We wants to find Θ such that $P(Y=y \mid \Theta, X=x)$ is maximized
  - Maximum Likelihood Estimator
  - $\Theta^* = \mathrm{argmax}_{\Theta} P(x \mid \Theta)$

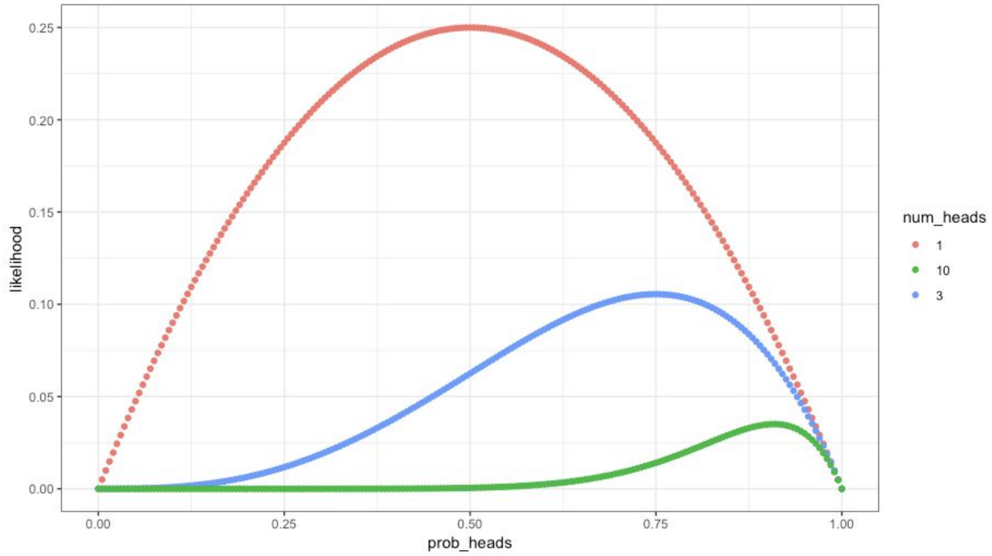
Likelihood Function - coin example

- We flip a coin twice times and get HT. We assume the coin is fair.
  - θ = P(Heads) = P(Tails) = 0.5
- Likelihood
  - P(HT | θ) = (0.5 * (1-0.5)) = 0.25



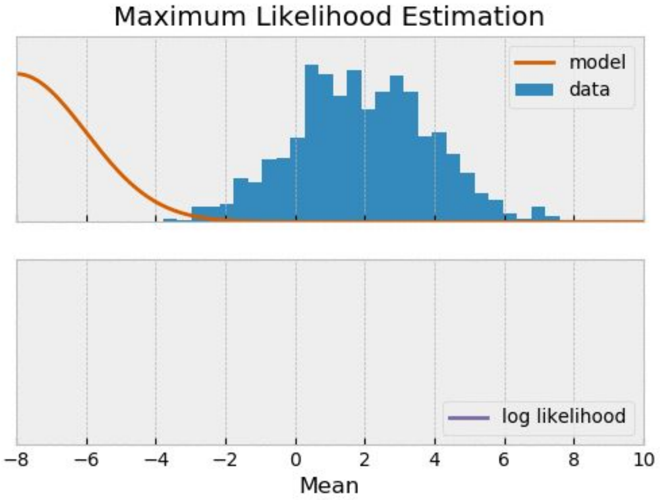
Likelihood Function - coin example

- We flip a coin twice times and get HT. We assume the coin is fair.
  - θ = P(Heads) = P(Tails) = 0.5
- Likelihood
  - P(HT | θ) = (0.5 * (1-0.5)) = 0.25



Likelihood Function - coin example

- We flip a coin twice times and get HT. We assume the coin is fair.
  - θ = P(Heads) = P(Tails) = 0.5
- Likelihood
  - P(HT | θ) = (0.5 * (1-0.5)) = 0.25



https://towardsdatascience.com/maximum-likelihood-estimation-984af2dcfcac

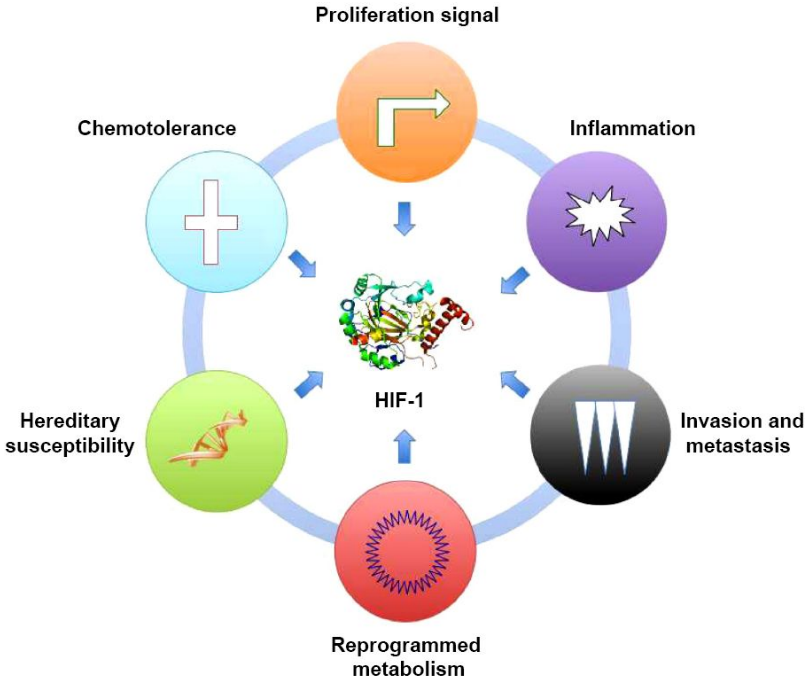
Real-world example - Predicting gene expression

- Does amount of oxygen in a cell predict the expression of the gene HIF?

https://www.dovepress.com/targeting-the-hypoxia-pathway-to-treat-pancreaticnbspcancer-peer-reviewed-fulltext-article-DDDT

Real-world example - Predicting gene expression

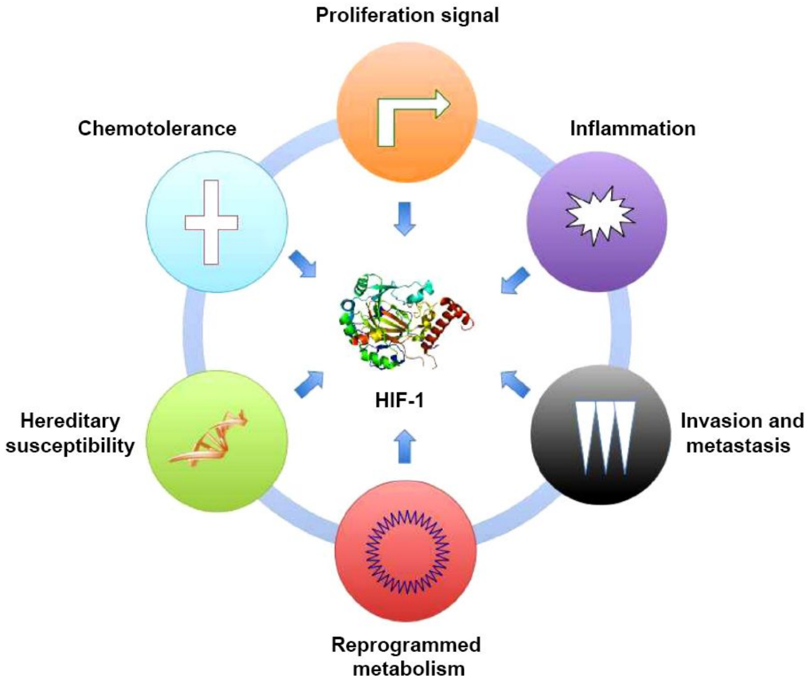
- Does amount of oxygen in a cell predict the expression of the gene HIF?



https://www.dovepress.com/targeting-the-hypoxia-pathway-to-treat-pancreaticnbspcancer-peer-reviewed-fulltext-article-DDDT

Real-world example - Predicting gene expression

- Does amount of oxygen in a cell predict the expression of the gene HIF?



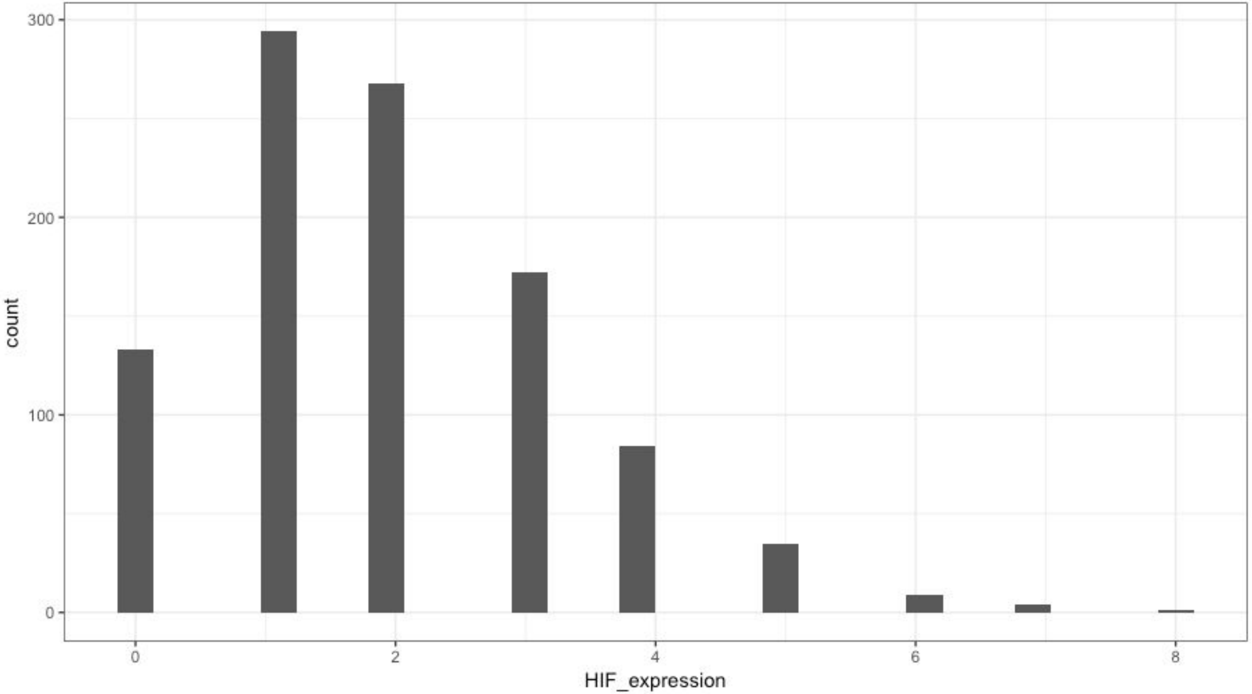
Oxygen level in a cell → HIF expression

(simplified model)

https://www.dovepress.com/targeting-the-hypoxia-pathway-to-treat-pancreaticnbspcancer-peer-reviewed-fulltext-article-DDDT

Real-world example - Predicting gene expression

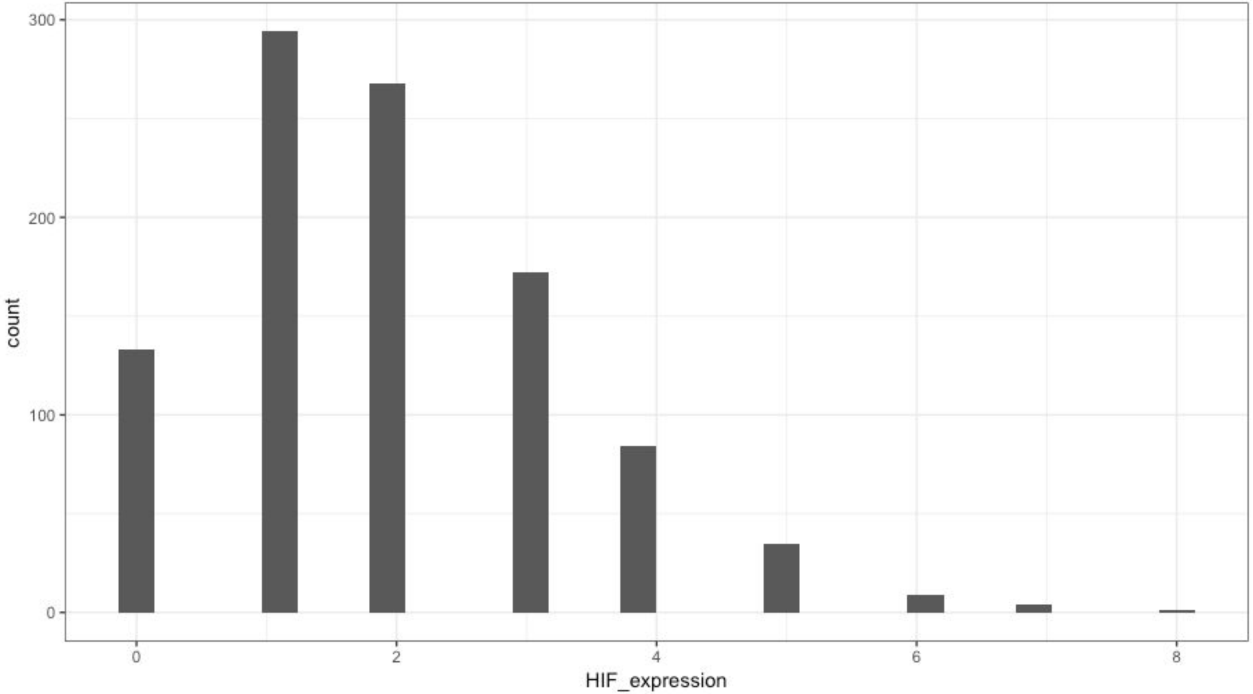
- Does amount of oxygen in a cell predict the expression of the gene HIF?



- This is the observed levels of HIF when oxygen is at 15%.
- What is the distributional assumption?

Real-world example - Predicting gene expression

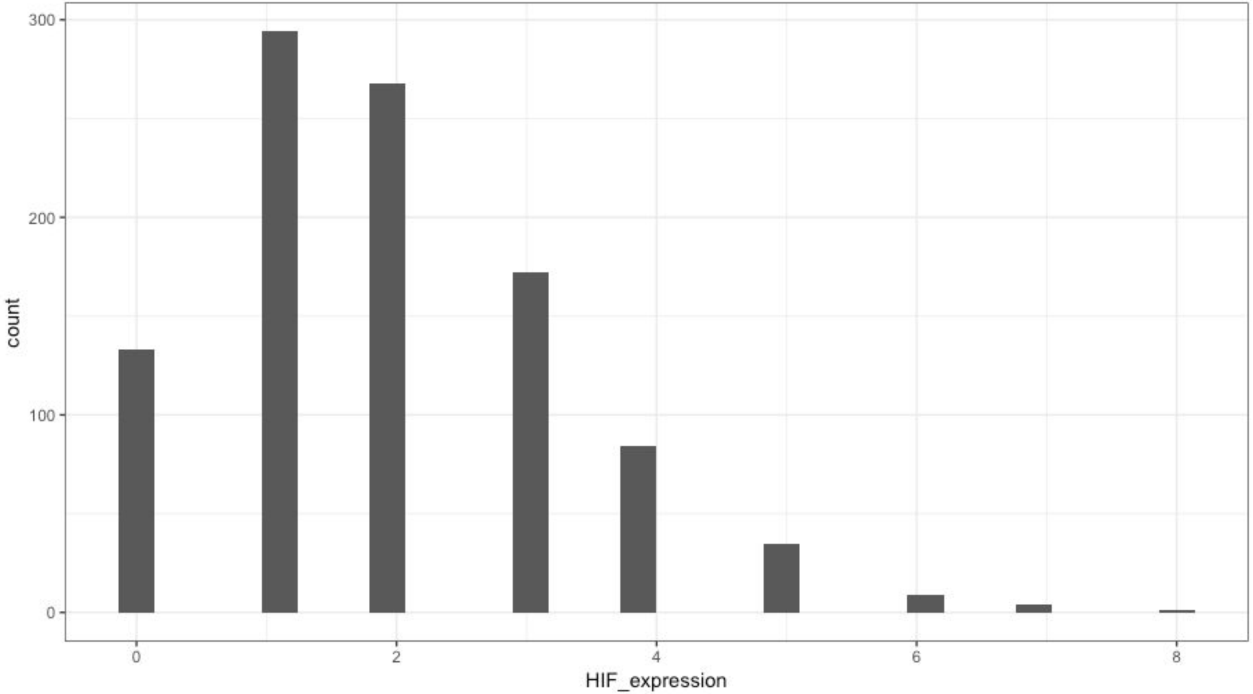
- Does amount of oxygen in a cell predict the expression of the gene HIF?



- This is the observed levels of HIF when oxygen is at 15%.
- What is the distributional assumption?
    - Positive
    - Discrete

Real-world example - Predicting gene expression

- Does amount of oxygen in a cell predict the expression of the gene HIF?

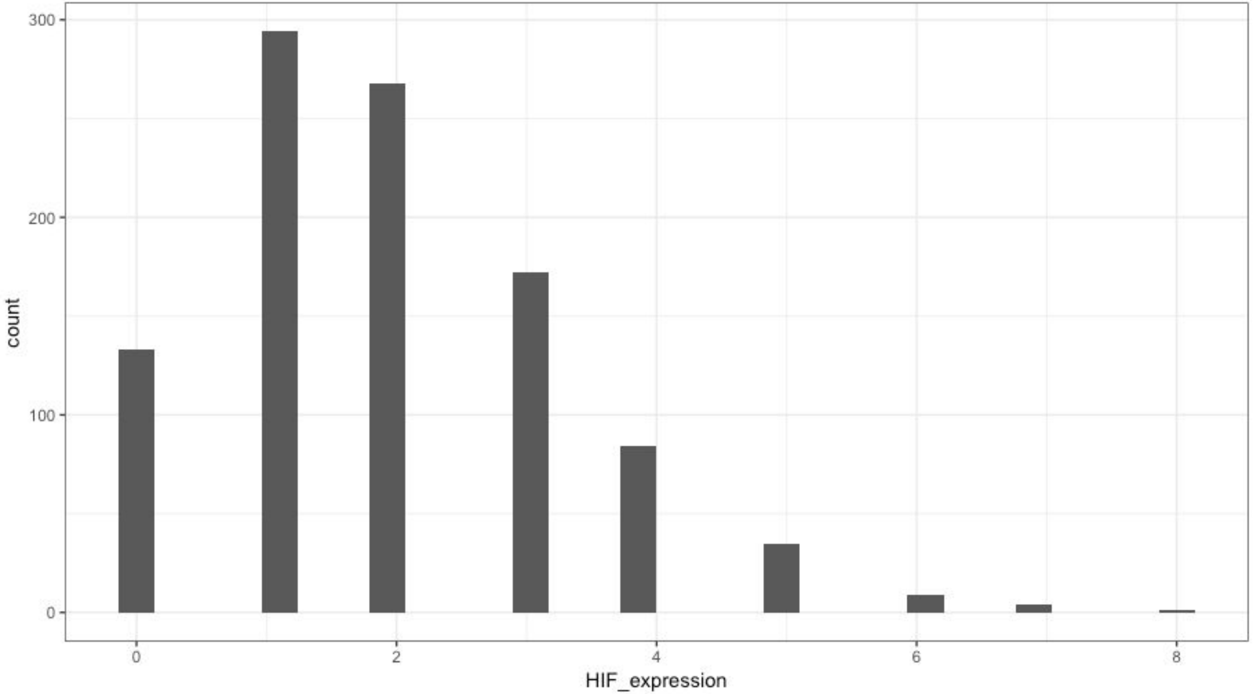


- This is the observed levels of HIF when oxygen is at 15%.
- What is the distributional assumption?

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Real-world example - Predicting gene expression

- Does amount of oxygen in a cell predict the expression of the gene HIF?



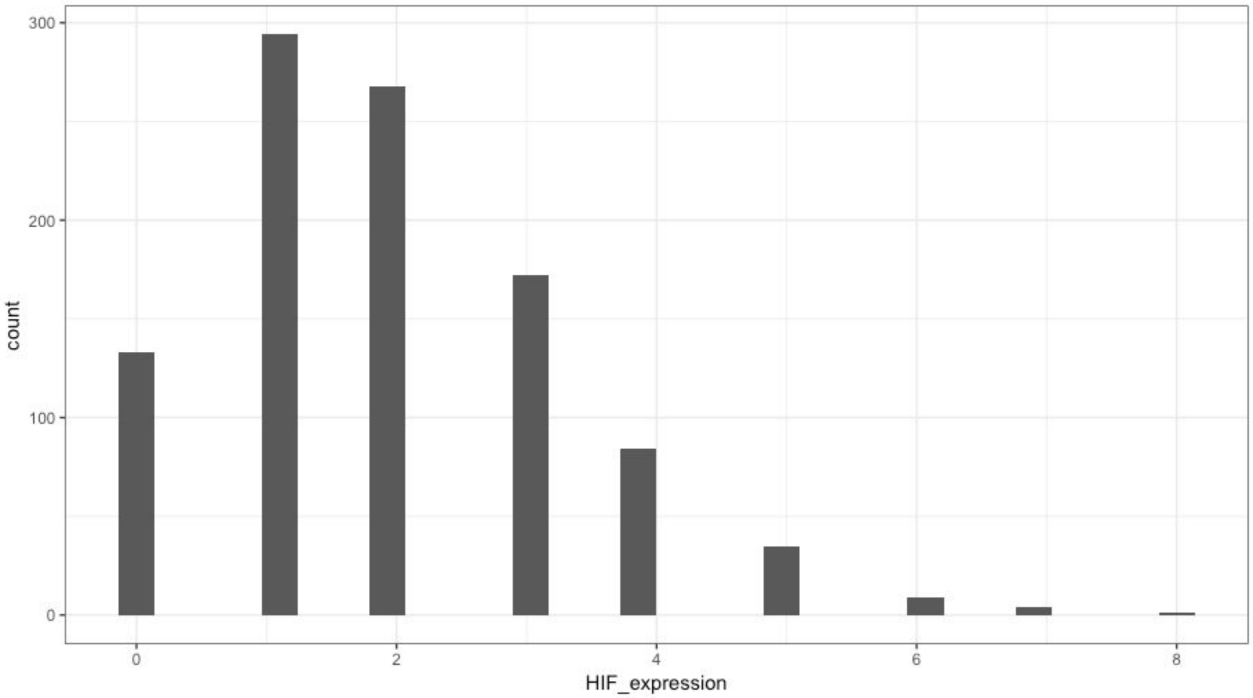
- This is the observed levels of HIF when oxygen is at 15%.
- What is the distributional assumption?

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

- MLE estimator for mean of poisson shown on board

Real-world example - Predicting gene expression

- Does amount of oxygen in a cell predict the expression of the gene HIF?



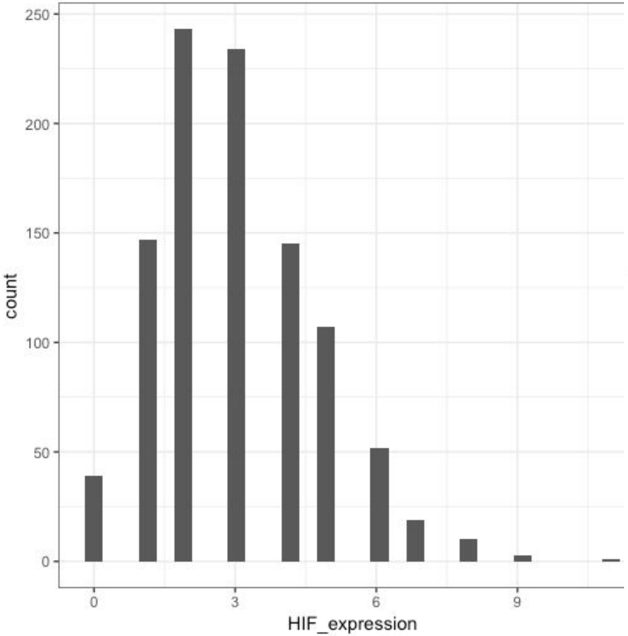
- This is the observed levels of HIF when oxygen is at 15%.
- What is the distributional assumption?

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

- Key assumption:
  - mean = variance

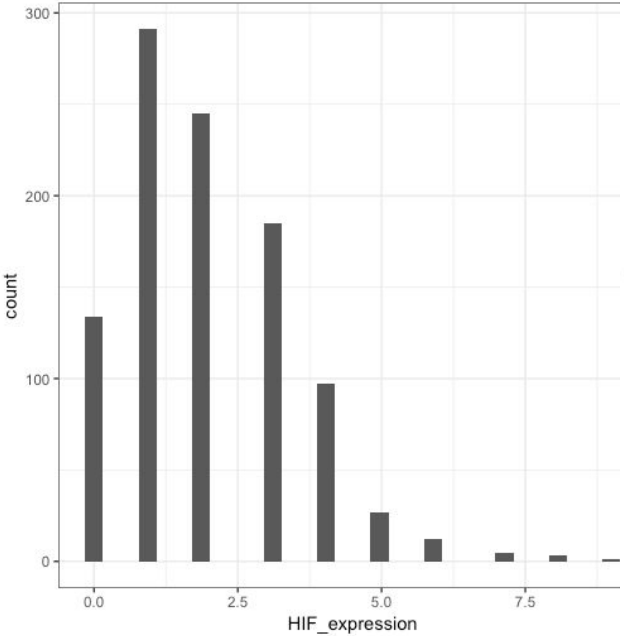
Real-world example - Predicting gene expression

Oxygen-level = 10%



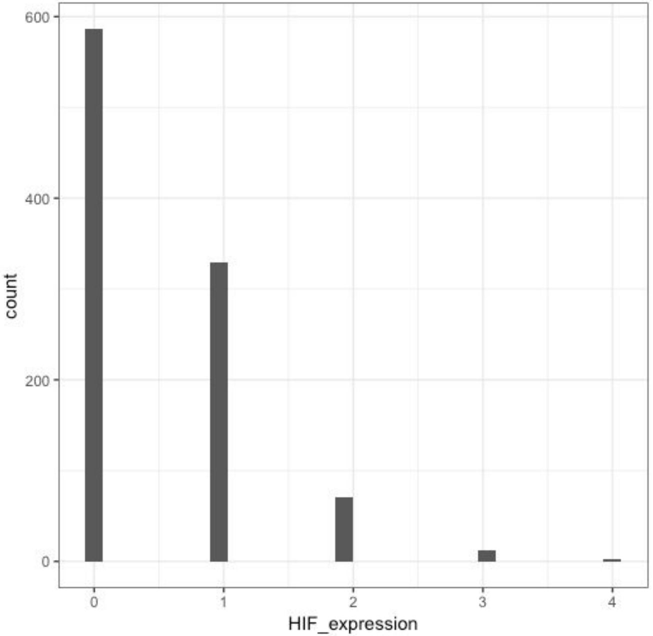
Mean = 3.03
Variance = 3.1

Oxygen-level = 15%



Mean = 2
Variance = 1.85

Oxygen-level = 20%



Mean = 0.535
Variance = 0.52

Real-world example - Predicting gene expression

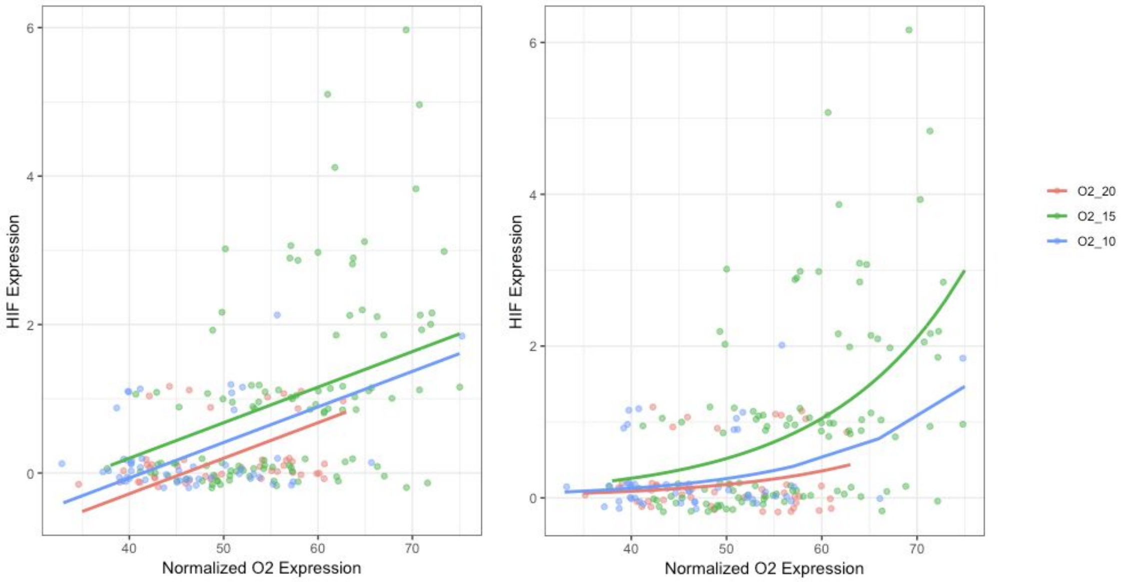
- Does amount of oxygen in a cell predict the expression of the gene HIF?
    - We want to model the conditional distribution of P(HIF expression level | Amount of Oxygen)

Real-world example - Predicting gene expression

- Does amount of oxygen in a cell predict the expression of the gene HIF?
  - We want to model the conditional distribution of P(HIF expression level | Amount of Oxygen)
  - Can we simply model conditional distribution with $w^T x$?
    - Remember: our estimator $w^T x$ models the mean of our conditional distribution.

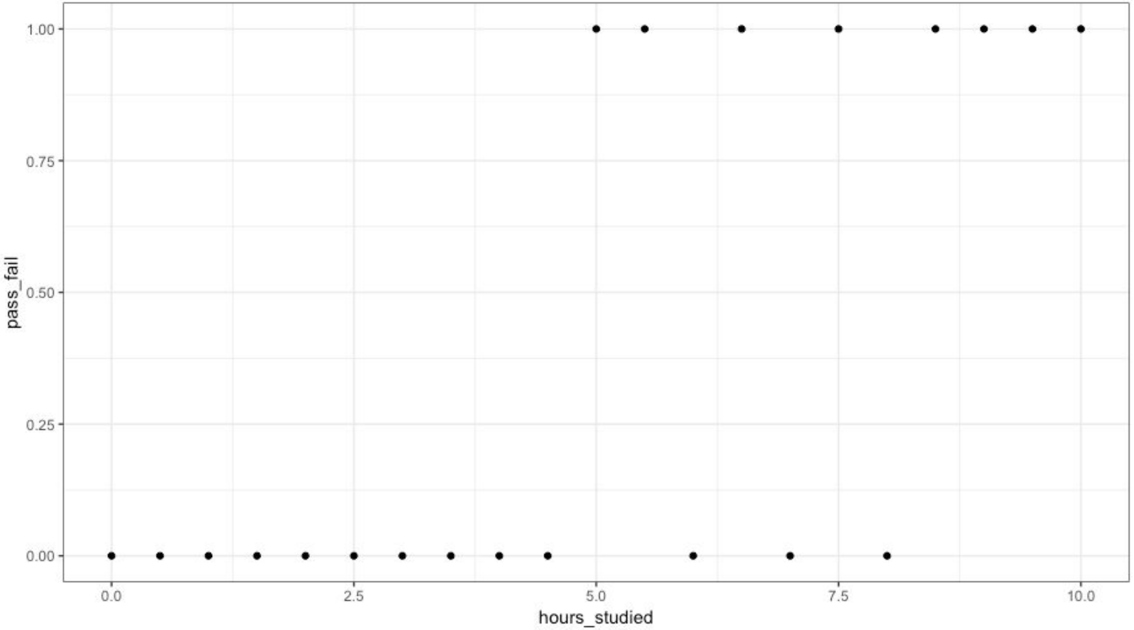
Real-world example - Predicting gene expression

- Does amount of oxygen in a cell predict the expression of the gene HIF?
  - We want to model the conditional distribution of P(HIF expression level | Amount of Oxygen)
  - Can we simply model conditional distribution with $w^T x$?
    - Remember: our estimator $w^T x$ models the mean of our conditional distribution.



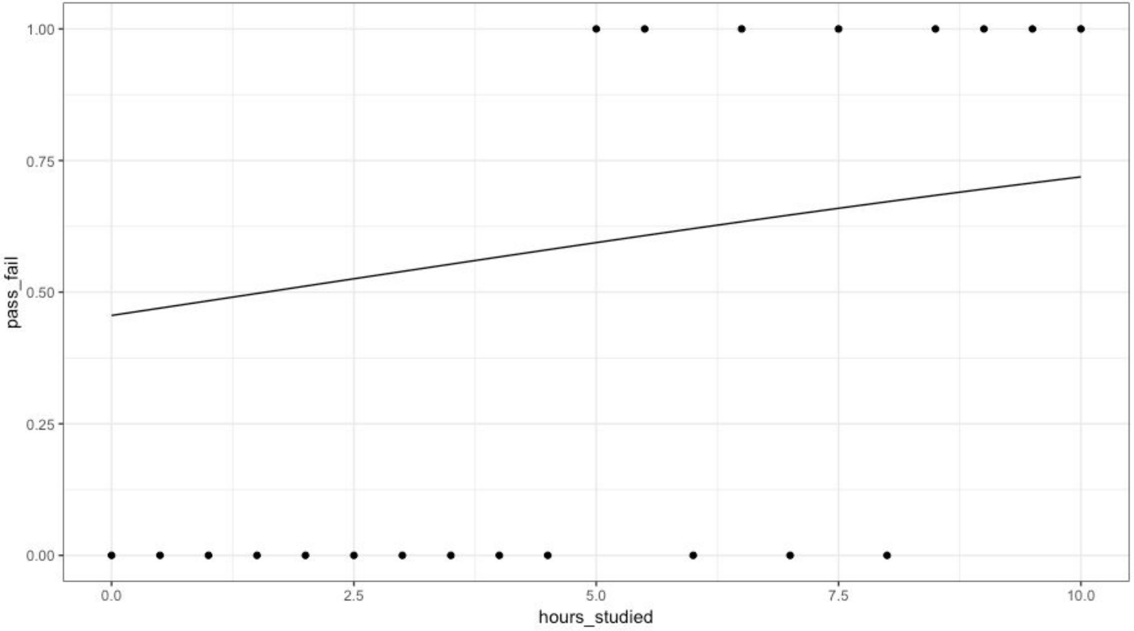
Link Function

- Remember from logistic regression, our conditional distribution is Bernoulli



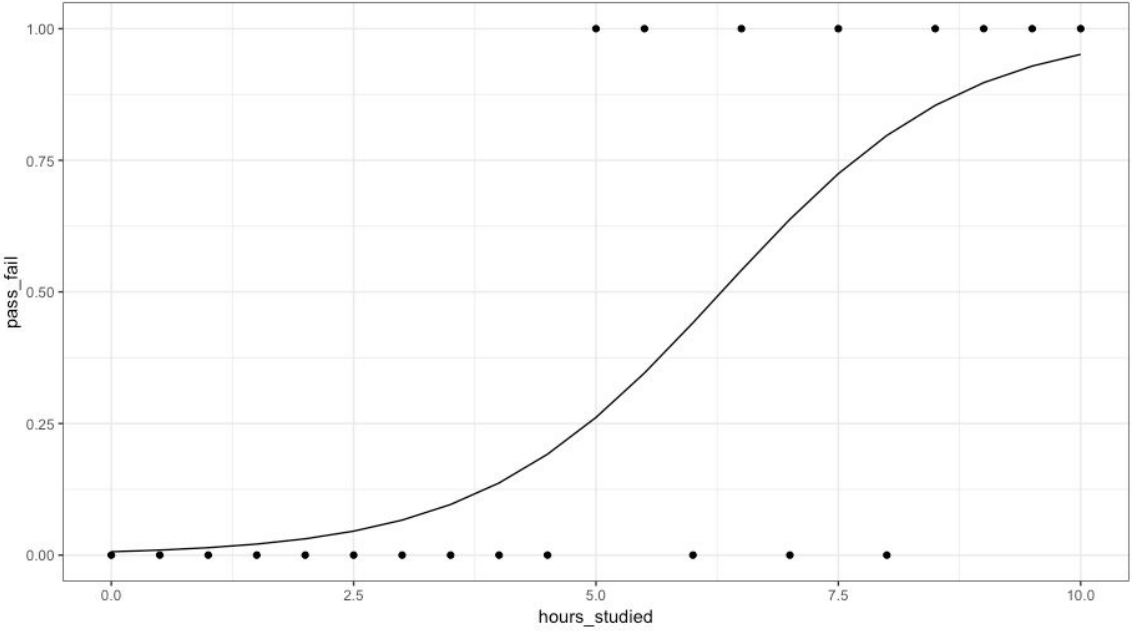
Link Function

- Remember from logistic regression, our conditional distribution is Bernoulli



Link Function

- Remember from logistic regression, our conditional distribution is Bernoulli



Canonical Link Functions

Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\beta = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\beta = \mu$	$\mu = \mathbf{X}\beta$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\beta = -\mu^{-1}$	$\mu = -(\mathbf{X}\beta)^{-1}$
Gamma					
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\beta = \mu^{-2}$	$\mu = (\mathbf{X}\beta)^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\beta = \ln(\mu)$	$\mu = \exp(\mathbf{X}\beta)$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\beta = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\beta)}{1+\exp(\mathbf{X}\beta)} = \frac{1}{1+\exp(-\mathbf{X}\beta)}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences			
Categorical	integer: $[0, K)$	outcome of single K-way occurrence			
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1				
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences			

https://en.wikipedia.org/wiki/Generalized_linear_model#Link_function

Canonical Link Functions

Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\beta = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\beta = \mu$	$\mu = \mathbf{X}\beta$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\beta = -\mu^{-1}$	$\mu = -(\mathbf{X}\beta)^{-1}$
Gamma					
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\beta = \mu^{-2}$	$\mu = (\mathbf{X}\beta)^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\beta = \ln(\mu)$	$\mu = \exp(\mathbf{X}\beta)$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\beta = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\beta)}{1+\exp(\mathbf{X}\beta)} = \frac{1}{1+\exp(-\mathbf{X}\beta)}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences			
Categorical	integer: $[0, K)$	outcome of single K-way occurrence			
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1				
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1 .. *K*) out of *N* total *K*-way occurrences			

https://en.wikipedia.org/wiki/Generalized_linear_model#Link_function

Real-world example - Predicting gene expression

- Does amount of oxygen in a cell predict the expression of the gene HIF?
  - We want to model the conditional distribution of P(HIF expression level | Amount of Oxygen)
  - Can we simply model conditional distribution with $w^T x$?
    - Remember: our estimator $w^T x$ models the mean of our conditional distribution.

- Now we have all the elements to find an optimal w
  - Conditional Distribution: Poisson
  - Link function: Log

Real-world example - Predicting gene expression

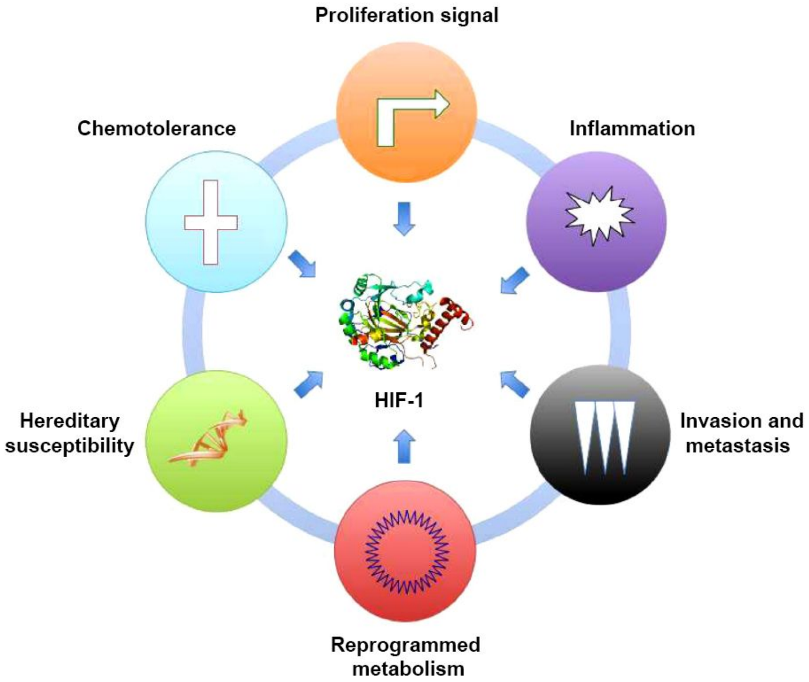
- Does amount of oxygen in a cell predict the expression of the gene HIF?
  - We want to model the conditional distribution of P(HIF expression level | Amount of Oxygen)
  - Can we simply model conditional distribution with $w^T x$?
    - Remember: our estimator $w^T x$ models the mean of our conditional distribution.

- Now we have all the elements to find an optimal w
  - Conditional Distribution: Poisson
  - Link function: Log

MLE estimate on board

Real-world example - Predicting gene expression

- Does amount of oxygen in a cell predict the expression of the gene HIF?



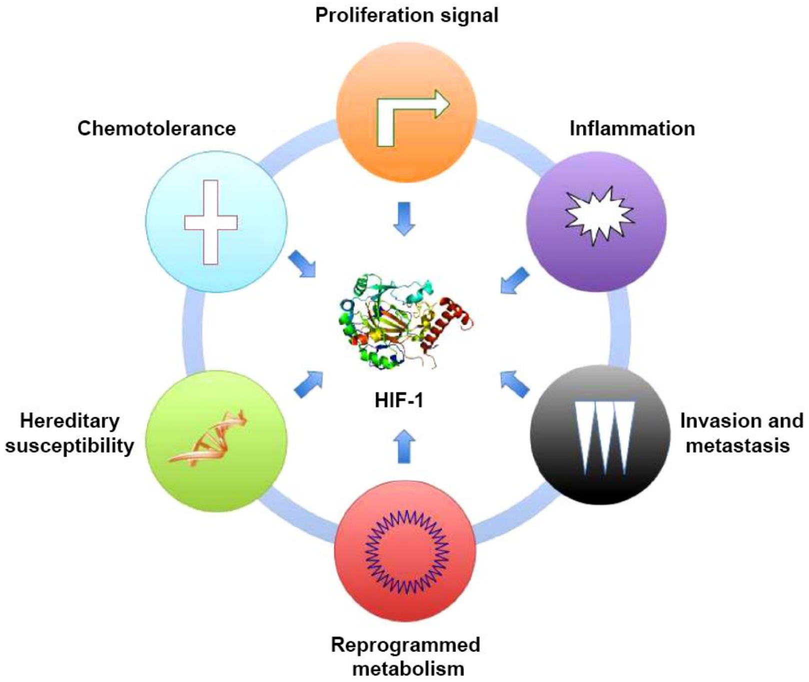
Oxygen level in a cell → HIF expression

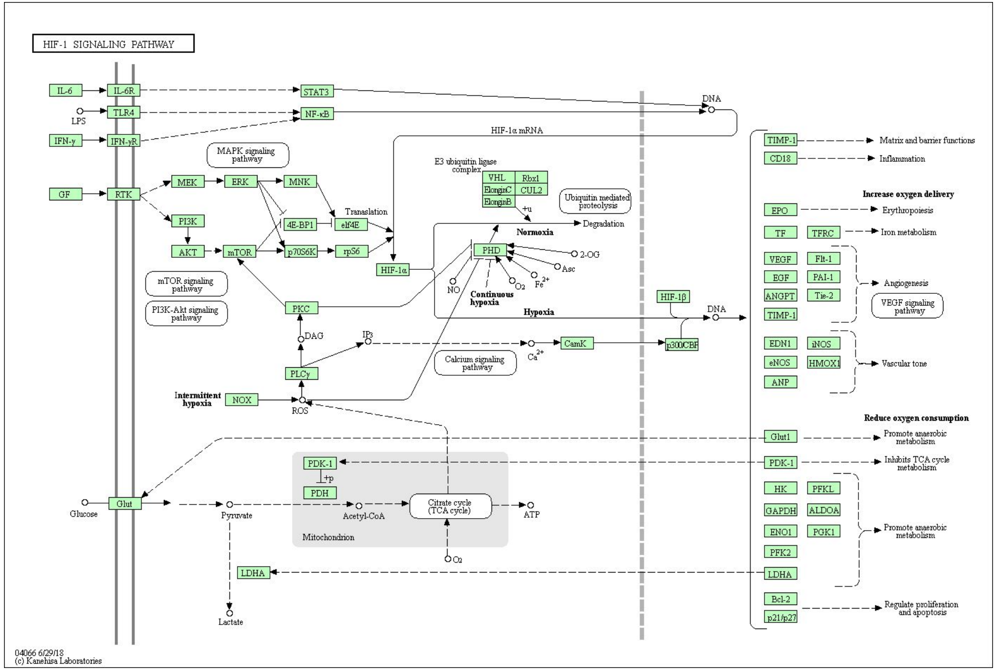
(simplified model)

https://www.dovepress.com/targeting-the-hypoxia-pathway-to-treat-pancreaticnbspcancer-peer-reviewed-fulltext-article-DDDT

Real-world example - Predicting gene expression

- Does amount of oxygen in a cell predict the expression of the gene HIF?

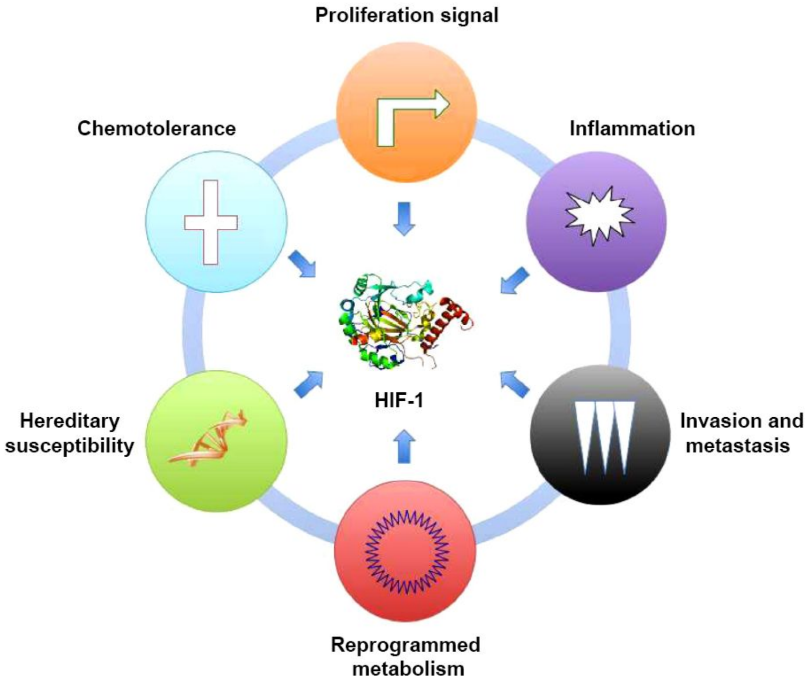


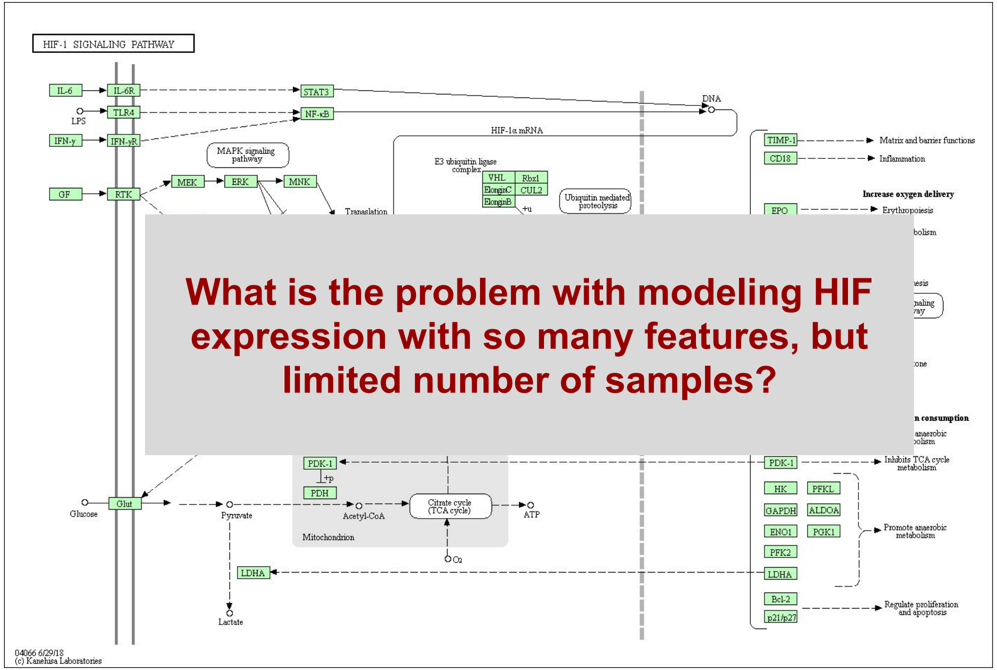


https://www.dovepress.com/targeting-the-hypoxia-pathway-to-treat-pancreaticnbspcancer-peer-reviewed-fulltext-article-DDDT

Real-world example - Predicting gene expression

- Does amount of oxygen in a cell predict the expression of the gene HIF?





What is the problem with modeling HIF expression with so many features, but limited number of samples?

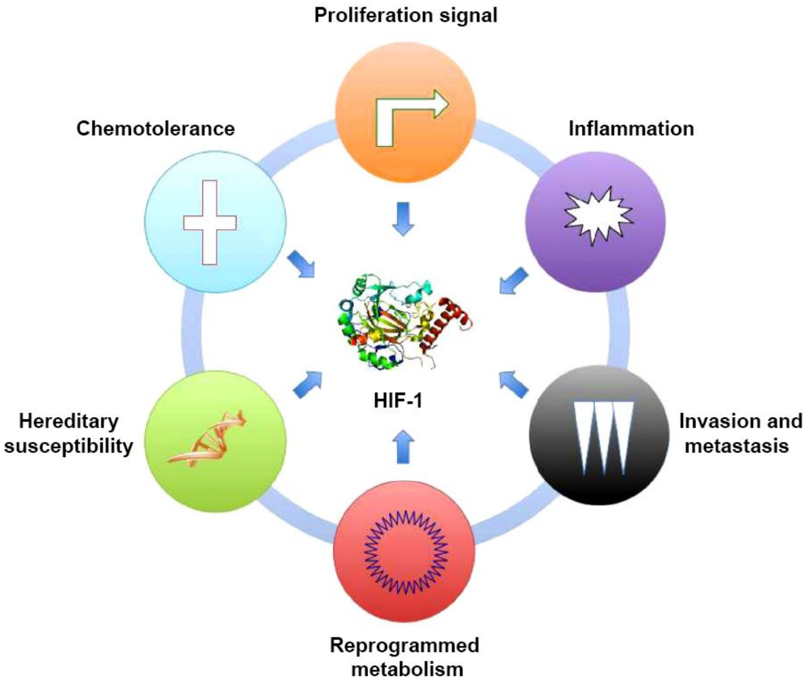
https://www.dovepress.com/targeting-the-hypoxia-pathway-to-treat-pancreaticnbspcancer-peer-reviewed-fulltext-article-DDDT

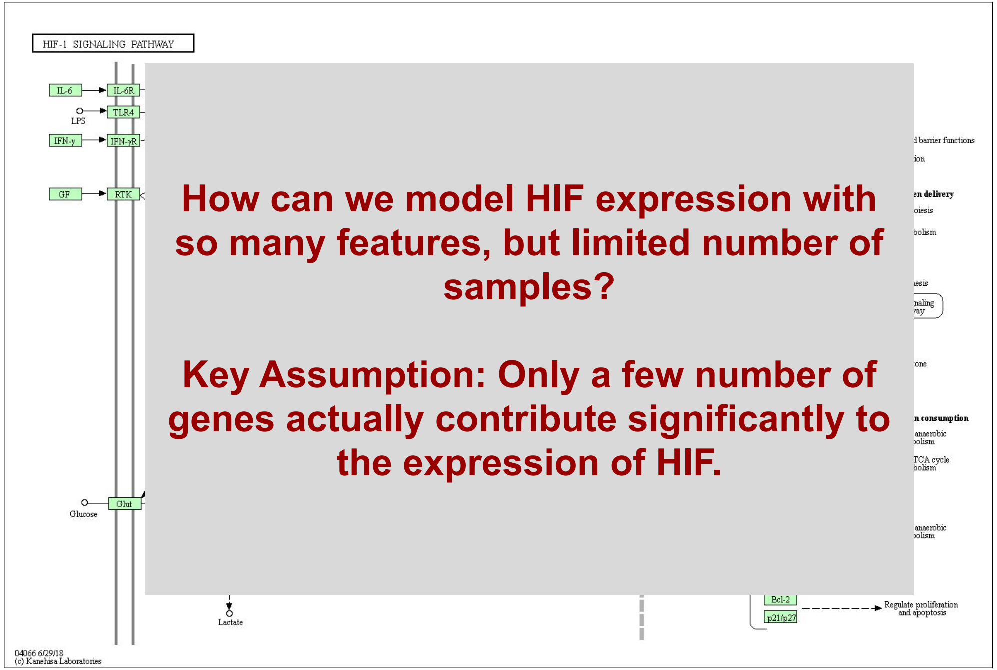
Bias-Variance Trade-off

- Ridge Regression Demo
  - https://cscheid.net/writing/data_science/regularization/

Real-world example - Predicting gene expression

- Does amount of oxygen in a cell predict the expression of the gene HIF?





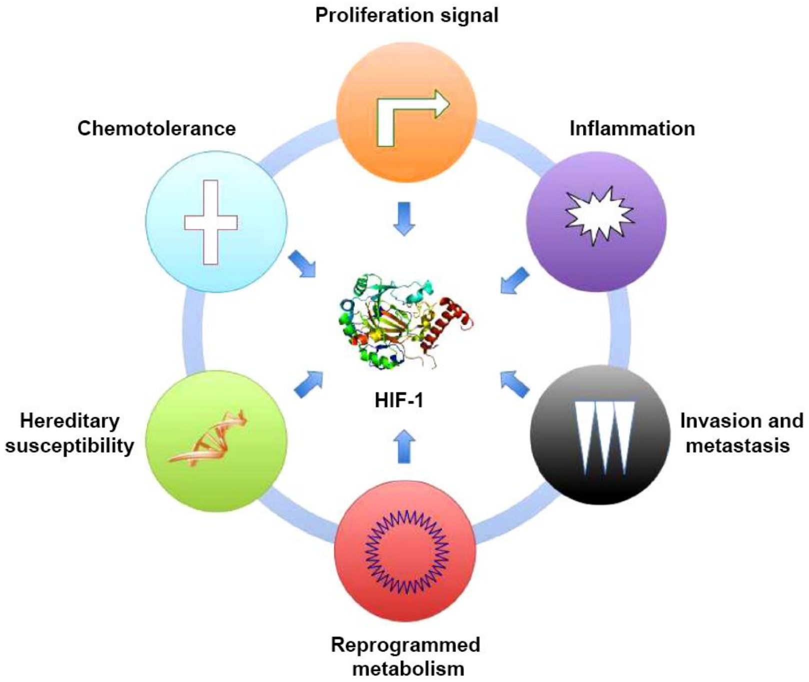
How can we model HIF expression with so many features, but limited number of samples?

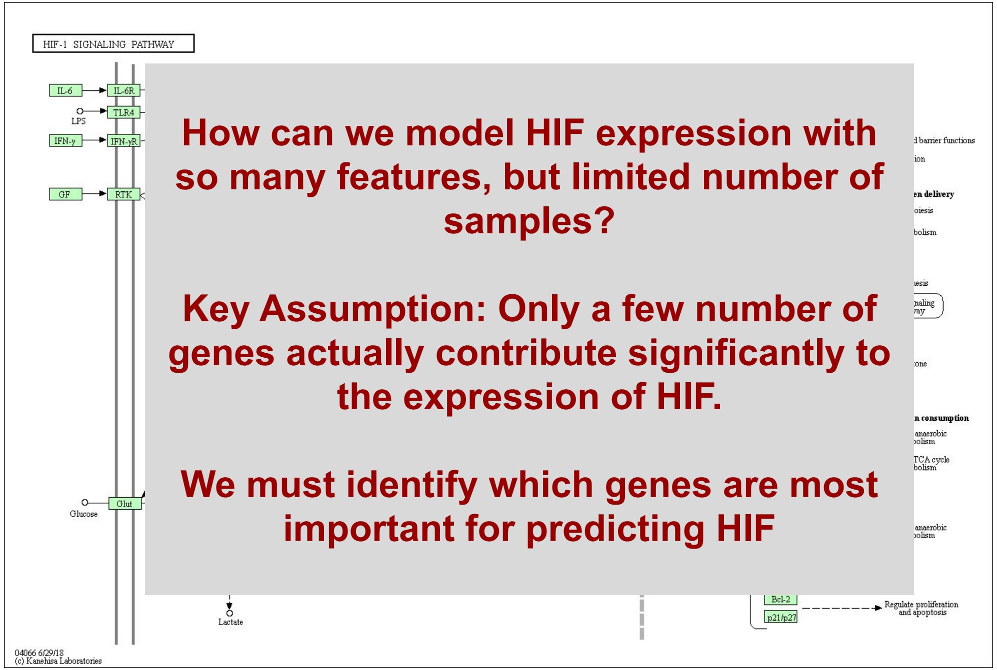
Key Assumption: Only a few number of genes actually contribute significantly to the expression of HIF.

https://www.dovepress.com/targeting-the-hypoxia-pathway-to-treat-pancreaticnbspcancer-peer-reviewed-fulltext-article-DDDT

Real-world example - Predicting gene expression

- Does amount of oxygen in a cell predict the expression of the gene HIF?





How can we model HIF expression with so many features, but limited number of samples?

Key Assumption: Only a few number of genes actually contribute significantly to the expression of HIF.

We must identify which genes are most important for predicting HIF

https://www.dovepress.com/targeting-the-hypoxia-pathway-to-treat-pancreaticnbspcancer-peer-reviewed-fulltext-article-DDDT

MAP Estimation

- We have a prior assumption about how our weightings should be distributed
- Bayes Rule
  - $P(\theta, X=x \mid Y=y) \propto P(Y=y \mid \theta, X=x) * P(\theta)$
  - Posterior $\propto$ Likelihood * Prior

MAP Estimation

- We have a prior assumption about how our weightings should be distributed
- Bayes Rule
  - $P(\theta, X=x \mid Y=y) \propto P(Y=y \mid \theta, X=x) * P(\theta)$
  - Posterior $\propto$ Likelihood * Prior
  - We can place a distributional assumption on θ

MAP Estimation

- We have a prior assumption about how our weightings should be distributed
- Bayes Rule
  - $P(\theta, X=x \mid Y=y) \propto P(Y=y \mid \theta, X=x) * P(\theta)$
  - Posterior $\propto$ Likelihood * Prior
  - We can place a distributional assumption on θ
- Elements of Maximum A Posteriori Estimation
  - What is the conditional distribution $P(Y=y \mid \theta, X=x)$?
  - How do we model the relationship between mean of the conditional distribution with our predictor (link function)?
  - What is the distribution on our parameters?

MAP Estimation

- We have a prior assumption about how our weightings should be distributed
- Bayes Rule
  - $P(\theta, X=x \mid Y=y) \propto P(Y=y \mid \theta, X=x) * P(\theta)$
  - Posterior $\propto$ Likelihood * Prior
  - We can place a distributional assumption on θ
- Elements of Maximum A Posteriori Estimation
  - What is the conditional distribution $P(Y=y \mid \theta, X=x)$? **Poisson**
  - How do we model the relationship between mean of the conditional distribution with our predictor (link function)? **log**
  - What is the distribution on our parameters? **Conjugate Priors**

Conjugate Priors

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters[note 1]	Posterior predictive[note 2]
Bernoulli	p (probability)	Beta	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^n x_i,\ \beta + n - \sum_{i=1}^n x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$p(\tilde{x} = 1) = \frac{\alpha'}{\alpha' + \beta'}$
Binomial	p (probability)	Beta	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^n x_i,\ \beta + \sum_{i=1}^n N_i - \sum_{i=1}^n x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$\operatorname{BetaBin}(\tilde{x} \mid \alpha', \beta')$ (beta-binomial)
Negative binomial with known failure number, r	p (probability)	Beta	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^n x_i,\ \beta + rn$	$\alpha - 1$ total successes, $\beta - 1$ failures[note 1] (i.e., $\frac{\beta - 1}{r}$ experiments, assuming r stays fixed)	
Poisson	λ (rate)	Gamma	$k,\ \theta$	$k + \sum_{i=1}^n x_i,\ \frac{\theta}{n\theta + 1}$	k total occurrences in $\frac{1}{\theta}$ intervals	$\operatorname{NB}(\tilde{x} \mid k', \theta')$ (negative binomial)
			$\alpha,\ \beta$[note 3]	$\alpha + \sum_{i=1}^n x_i,\ \beta + n$	α total occurrences in β intervals	$\operatorname{NB}\left(\tilde{x} \mid \alpha', \frac{1}{1 + \beta'}\right)$ (negative binomial)
Categorical	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + (c_1, \ldots, c_k)$, where c_i is the number of observations in category i	$\alpha_i - 1$ occurrences of category i[note 1]	$p(\tilde{x} = i) = \frac{{\alpha_i}'}{\sum_i {\alpha_i}'} = \frac{\alpha_i + c_i}{\sum_i \alpha_i + n}$
Multinomial	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + \sum_{i=1}^n \mathbf{x}_i$	$\alpha_i - 1$ occurrences of category i[note 1]	$\operatorname{DirMult}(\tilde{\mathbf{x}} \mid \boldsymbol{\alpha}')$ (Dirichlet-multinomial)
Hypergeometric with known total population size, N	M (number of target members)	Beta-binomial[4]	$n = N, \alpha,\ \beta$	$\alpha + \sum_{i=1}^n x_i,\ \beta + \sum_{i=1}^n N_i - \sum_{i=1}^n x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	
Geometric	p_0 (probability)	Beta	$\alpha,\ \beta$	$\alpha + n,\ \beta + \sum_{i=1}^n x_i - n$	$\alpha - 1$ experiments, $\beta - 1$ total failures[note 1]	

https://en.wikipedia.org/wiki/Conjugate_prior

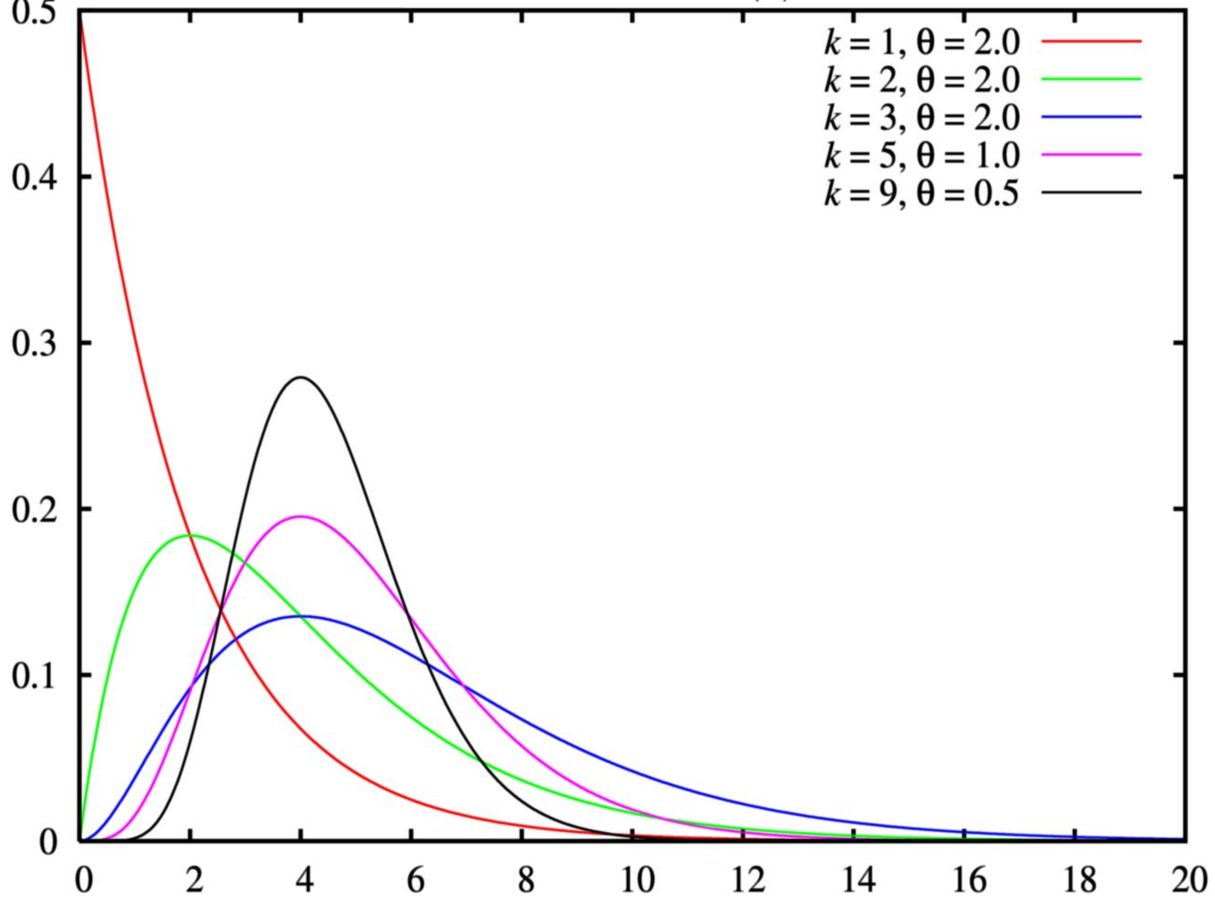
Conjugate Priors

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters[note 1]	Posterior predictive[note 2]
Bernoulli	p (probability)	Beta	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + n - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$p(\tilde{x} = 1) = \frac{\alpha'}{\alpha' + \beta'}$
Binomial	p (probability)	Beta	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$\operatorname{BetaBin}(\tilde{x} \vert \alpha', \beta')$ (beta-binomial)
Negative binomial with known failure number, r	p (probability)	Beta	$\alpha,\ \beta$	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + rn$	$\alpha - 1$ total successes, $\beta - 1$ failures[note 1] (i.e., $\frac{\beta - 1}{r}$ experiments, assuming r stays fixed)	
Poisson	λ (rate)	Gamma	$k,\ \theta$	$k + \sum_{i=1}^{n} x_i,\ \frac{\theta}{n\theta + 1}$	k total occurrences in $\frac{1}{\theta}$ intervals	$\operatorname{NB}(\tilde{x} \mid k', \theta')$ (negative binomial)
			$\alpha,\ \beta$[note 3]	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + n$	α total occurrences in β intervals	$\operatorname{NB}\left(\tilde{x} \mid \alpha', \frac{1}{1 + \beta'}\right)$ (negative binomial)
Categorical	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + (c_1, \ldots, c_k)$, where c_i is the number of observations in category i	$\alpha_i - 1$ occurrences of category i[note 1]	$p(\tilde{x} = i) = \frac{\alpha_i{}'}{\sum_i \alpha_i{}'} = \frac{\alpha_i + c_i}{\sum_i \alpha_i + n}$
Multinomial	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + \sum_{i=1}^{n} \mathbf{x}_i$	$\alpha_i - 1$ occurrences of category i[note 1]	$\operatorname{DirMult}(\tilde{\mathbf{x}} \mid \boldsymbol{\alpha}')$ (Dirichlet-multinomial)
Hypergeometric with known total population size, N	M (number of target members)	Beta-binomial[4]	$n = N, \alpha,\ \beta$	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	
Geometric	p_0 (probability)	Beta	$\alpha,\ \beta$	$\alpha + n,\ \beta + \sum_{i=1}^{n} x_i - n$	$\alpha - 1$ experiments, $\beta - 1$ total failures[note 1]	

https://en.wikipedia.org/wiki/Conjugate_prior

Conjugate Priors

$$f(x;k,\theta) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k\Gamma(k)} \quad \text{for } x > 0 \text{ and } k,\theta > 0.$$



https://en.wikipedia.org/wiki/Gamma_distribution

Prior - Likelihood - Posterior Demo

- https://rpsychologist.com/d3/bayes/

MAP Estimation

- We have a prior assumption about how our weightings should be distributed
- Bayes Rule
  - $P(\theta, X=x \mid Y=y) \propto P(Y=y \mid \theta, X=x) * P(\theta)$
  - Posterior $\propto$ Likelihood * Prior
  - We can place a distributional assumption on θ
- Elements of Maximum A Posteriori Estimation
  - What is the conditional distribution $P(Y=y \mid \theta, X=x)$? **Poisson**
  - How do we model the relationship between mean of the conditional distribution with our predictor (link function)? **log**
  - What is the distribution on our parameters? **Conjugate Priors - Gamma**

MAP derivation written on board

References + online guides

- https://www.statlect.com/fundamentals-of-statistics/Poisson-distribution-maximum-likelihood
- https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/ppt/22-MAP.pdf
- http://statmath.wu.ac.at/courses/heather_turner/glmCourse_001.pdf
- https://www.cs.indiana.edu/~predrag/classes/2015springb555/4.pdf
- https://www.irit.fr/~Herwig.Wendt/data/EstDect_TD1_compl.pdf
- https://www4.stat.ncsu.edu/~reich/ABA/notes/PoissonGamma.pdf
  - https://www4.stat.ncsu.edu/~reich/ABA/derivations4.pdf