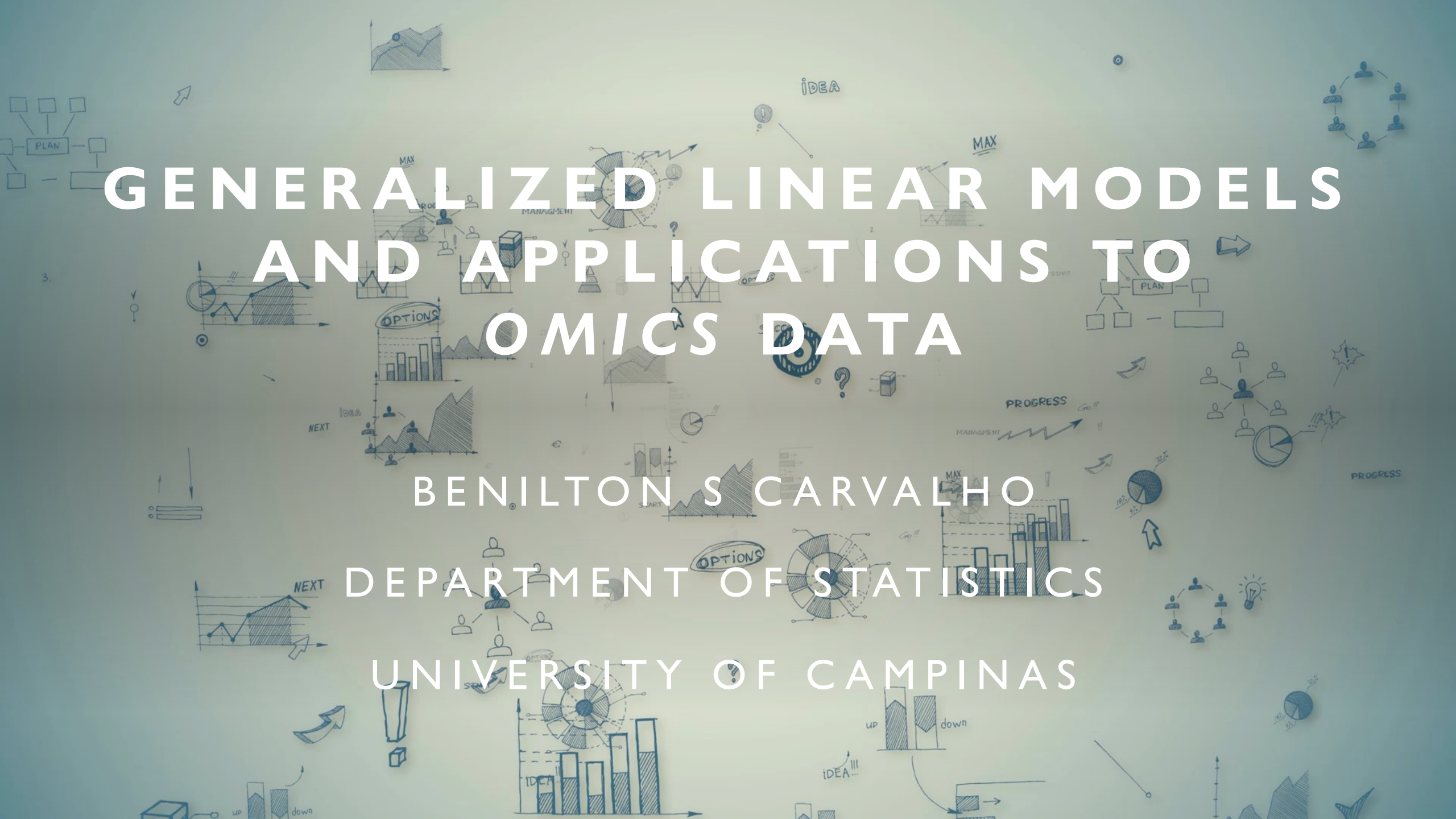

GENERALIZED LINEAR MODELS AND APPLICATIONS TO *OMICS* DATA

BENILTON S CARVALHO

DEPARTMENT OF STATISTICS

UNIVERSITY OF CAMPINAS

LINEAR REGRESSION MODEL

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

- The coefficient for each variable is the average increment on the response when the variable is increased by 1 unit;
- The response is a continuous (as in measurement) variable;

INTRODUCTION TO GLM

- Generalized Linear Models (GLMs) extend the linear modelling framework to variables that do not follow the Normal distribution;
- Linear regression won't work when:
  - The range of the response Y is restricted
  - The variance of Y depends on the mean
- GLMs are often used to model binary and count data;

STRUCTURE OF A GLM

- Linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$$

- Link function

$$g(\mu_i) = \eta_i$$

- Variance function

$$V(Y_i) = \phi V(\mu_i)$$

Note: Linear regression is a particular case of the above

ONE IMPORTANT INFORMATION

- The link function is on the scale of the predictors

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$$

- To estimate the link function (linear predictor), we need to estimate the βs;
- To compare profiles (i.e., combinations of predictions), we make contrasts of the linear predictors;
- To have the estimates at the scale of the mean, we need to apply the inverse (link) function to the linear predictors;

BINARY DATA

- Responses are of the form: Success/Fail, 1/0, T/F, H/T

$$
\begin{aligned}
p_i &\in [0, 1] \\
Y_i &\sim \text{Bernoulli}(p_i) \\
E(Y_i) &\sim p_i \\
V(Y_i) &\sim p_i(1 - p_i)
\end{aligned}
$$

- How to map probabilities to (-∞, + ∞)?

LOGISTIC REGRESSION

$$
\begin{aligned}
p_i &\in [0, 1] \\
\frac{p_i}{1 - p_i} &\in [0, +\infty) \\
\log\left(\frac{p_i}{1 - p_i}\right) &\in (-\infty, +\infty)
\end{aligned}
$$

FAMUSS DATASET

```
> head(data1[, c("gain50", "actn3.r577x")])
  gain50 actn3.r577x
1  FALSE          CC
2  FALSE          CT
3  FALSE          CT
4   TRUE          CT
5  FALSE          CC
6   TRUE          CT
```

RUNNING THE LOGISTIC REGRESSION

- Does genotype affect the ability to gain more than 50% in force on the non-dominant arm?

```
> fit = glm(gain50 ~ actn3.r577x, data = data1, family=binomial)
```

```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.6587     0.1604  -4.107 4.01e-05 ***
actn3.r577xCT     0.2945     0.2039   1.445   0.1485
actn3.r577xTT     0.5716     0.2250   2.541   0.0111 *
```

PREDICTION WITH LOGISTIC REGRESSION

```
> pred1 = data.frame(actn3.r577x = c("CC", "CT", "TT"))
> pred1$p = predict(fit, newdata = pred1, type="response")
```

actn3.r577x	p
CC	0.3410405
CT	0.4099617
TT	0.4782609

COUNT DATA

- Responses are non-negative integers: 0, 1, 2, …
- The Poisson model is usually applied in these cases:

$$\begin{aligned} \lambda_i &> 0 \\ Y_i &\sim \text{Poisson}(\lambda_i) \\ E(Y_i) &\sim \lambda_i \\ V(Y_i) &\sim \lambda_i \end{aligned}$$

- How to map the mean to (-∞, + ∞)?

POISSON REGRESSION

$$\begin{aligned} \lambda_i &\in (0, +\infty) \\ \log(\lambda_i) &\in (-\infty, +\infty) \end{aligned}$$

BOTTOMLY DATASET

- The data used here corresponds to the first gene, ENSMUSG00000000001, from the "*Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays*" paper.

```
> fit = glm(gene1 ~ strain, data=dat2, family=poisson)
```

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.23989    0.01396 446.839  < 2e-16 ***
strainDBA/2J  0.04983    0.01907   2.613  0.00898 **
```

PREDICTION WITH POISSON REGRESSION

```
> pred2$mu = predict(fit, newdata=pred2, type="response")
> pred2
```

strain	mu
C57BL/6J	512.8
DBA/2J	539.0

STILL ON COUNT DATA

- The Poisson model forces the variance to be equal to the mean;
- In most cases, this statement does not hold;
- As variance changes as a function of the mean, one approach is to use the Negative Binomial model;

NEGATIVE BINOMIAL

$$
\begin{aligned}
\alpha_i &> 0 \\
\beta_i &> 0 \\
Y_i &\sim \text{Negative Binomial}\left(\alpha_i, \frac{1}{1+\beta_i}\right) \\
E(Y_i) &= \frac{\alpha_i}{\beta_i} \\
V(Y_i) &= \frac{\alpha_i(1+\beta_i)}{\beta_i^2} = \frac{\alpha_i}{\beta_i} + \frac{1}{\beta_i}\frac{\alpha_i}{\beta_i} \\
&= \frac{\alpha_i}{\beta_i} + \frac{1}{\alpha_i}\left(\frac{\alpha_i}{\beta_i}\right)^2 = \mu_i + \frac{1}{\alpha_i}\mu_i^2
\end{aligned}
$$

NEGATIVE BINOMIAL REGRESSION

- The link function is, again, log()

$$
\begin{aligned}
\mu_i &= \frac{\alpha_i}{\beta_i} \\
\mu_i &> 0 \\
\log(\mu_i) &\in (-\infty, +\infty)
\end{aligned}
$$

BOTTOMLY DATASET

- The data used here corresponds to the first gene, ENSMUSG00000000001, from the "*Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays*" paper.

```
> library(MASS)
> fit = glm.nb(gene1 ~ strain, data=dat2)
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	6.23989	0.12052	51.774	<2e-16	***
strainDBA/2J	0.04983	0.16650	0.299	0.765	

Theta: 6.98
Std. Err.: 2.13

PREDICTION WITH NEGATIVE BINOMIAL REGRESSION

```
> pred3 = data.frame(strain = c("C57BL/6J", "DBA/2J"))
> pred3$mu = predict(fit, newdata=pred3, type="response")
```

strain	mu
C57BL/6J	512.8
DBA/2J	539.0

THANK YOU...

- benilton@unicamp.br