# Linear Regression with Normal errors

## Statistics in Medical Research Fall Series

Marcio Augusto Diniz, Ph.D.
Biostatistics and Bioinformatics Research Center
Cedars Sinai Medical Center

October 18, 2022

# Summary

# What are regression models?

**Scientific Model**

It is a simplified representation of an idea, an object or even a process or a system that is used to describe and understand phenomena.

# What are regression models?


Figure: Solar system model

# What are regression models?


Figure: Solar system model

Figure: DNA model

# What are regression models?


Figure: Solar system model

Figure: DNA model

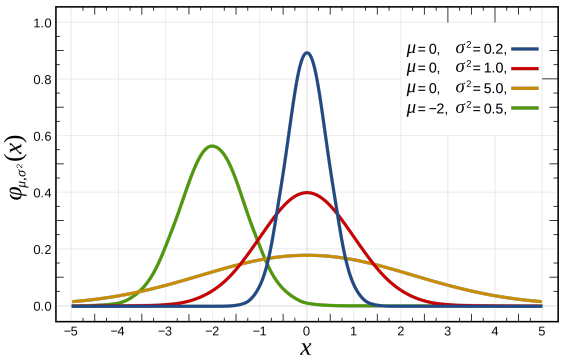



Figure: Normal probabilistic model

# What are regression models?

## Probabilistic Model

- It is a mathematical representation for the probabilities of all possible results of an experiment;
- Each probabilistic model is totally defined by a set of parameters.

# What are regression models?

## Probabilistic Model

- It is a mathematical representation for the probabilities of all possible results of an experiment;
- Each probabilistic model is totally defined by a set of parameters.

## Example

- Experiment: Tossing a fair coin;
- Results: heads, tails;
- Prob(heads) = Prob(tails) = 1/2.

# What are regression models?

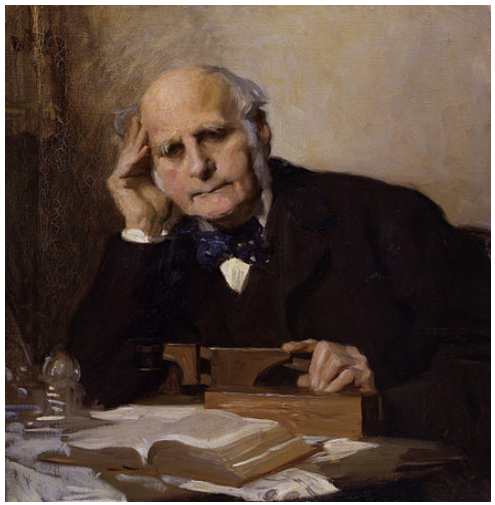


Figure: Francis Galton, 1854

## Regression?

- The term was coined by Francis Galton in $19^{th}$ to describe the biological phenomenon that the heights of descendants of tall ancestors tend to regress towards a normal average;
- It was later extended by Udny Yule and Karl Pearson to describe the relationship between two variables.

# The impact of Manuka honey dressings on the surface pH of chronic wounds

Georgina T Gethin, Seamus Cowman, Ronan M Conroy

Gethin GT, Cowman S, Conroy RM. The impact of Manuka honey dressings on the surface pH of chronic wounds. Int Wound J 2008;5:185–194.

# What are regression models?

## Example


Figure: $ 54.99

# What are regression models?

## Example

The impact of Manuka honey dressings on the surface pH of chronic wounds

- Hypothesis: The acidic pH of Manuka honey makes it a potential treatment for lowering wound pH.
- Aim: Analyze the changes in surface pH and size of non-healing ulcers following application of Manuka honey dressing after 2 weeks;
- Variable of interest: Percent of wound size change defined as

$$100 \times \frac{\text{Wound size at baseline} - \text{Wound size after two weeks}}{\text{Wound size at baseline}}$$

# What are regression models?

## The impact of Manuka honey dressings on the surface pH of chronic wounds

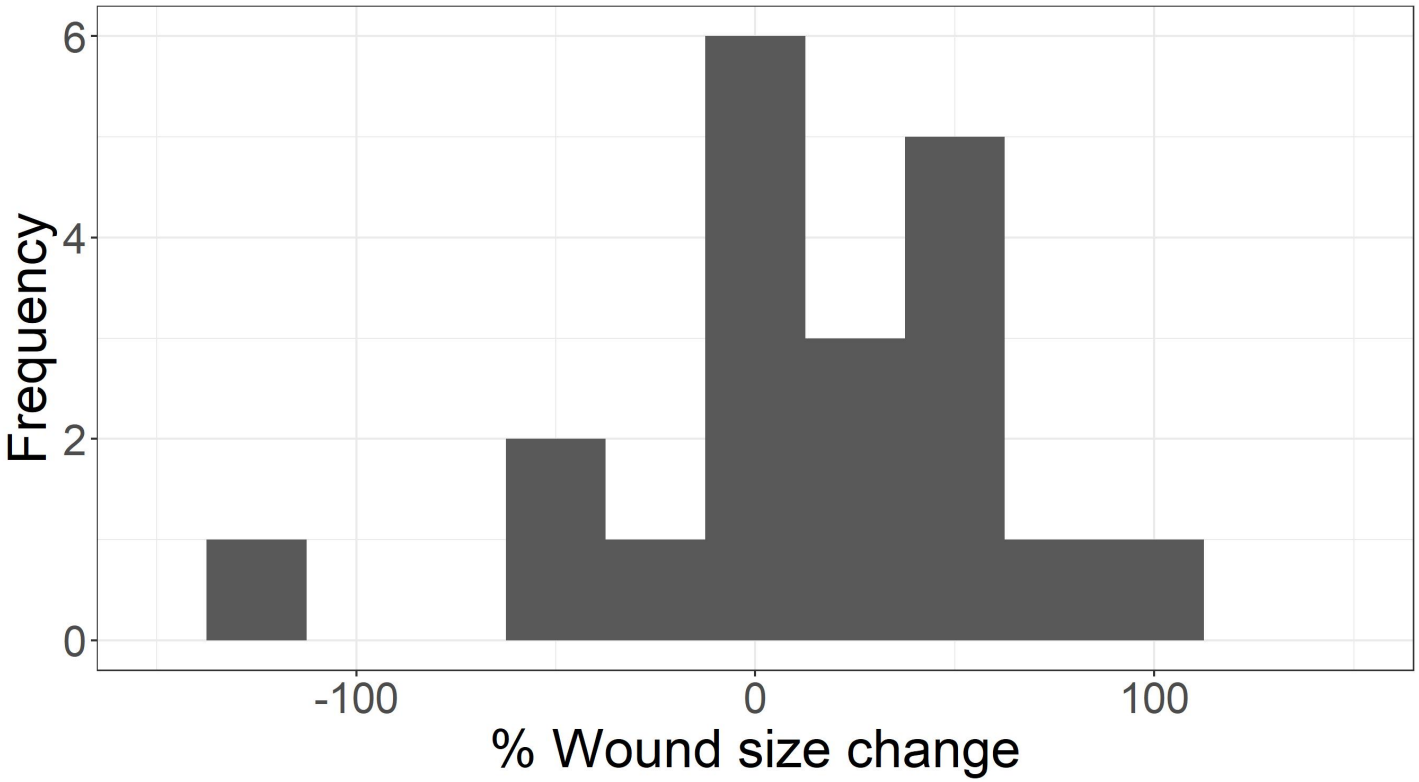



Figure: Histogram of percent of wound size change

# What are regression models?

## The impact of Manuka honey dressings on the surface pH of chronic wounds

### Probabilistic model

- Let be $Y$ the percent of change of wound size;
- $Y$ is a continuous measure and it is random;
- $Y \sim Normal(\mu, \sigma^2)$, where $\mu$ and $\sigma$ represent the mean and the standard deviation, respectively, of the *Normal* distribution.

# What are regression models?

## The impact of Manuka honey dressings on the surface pH of chronic wounds

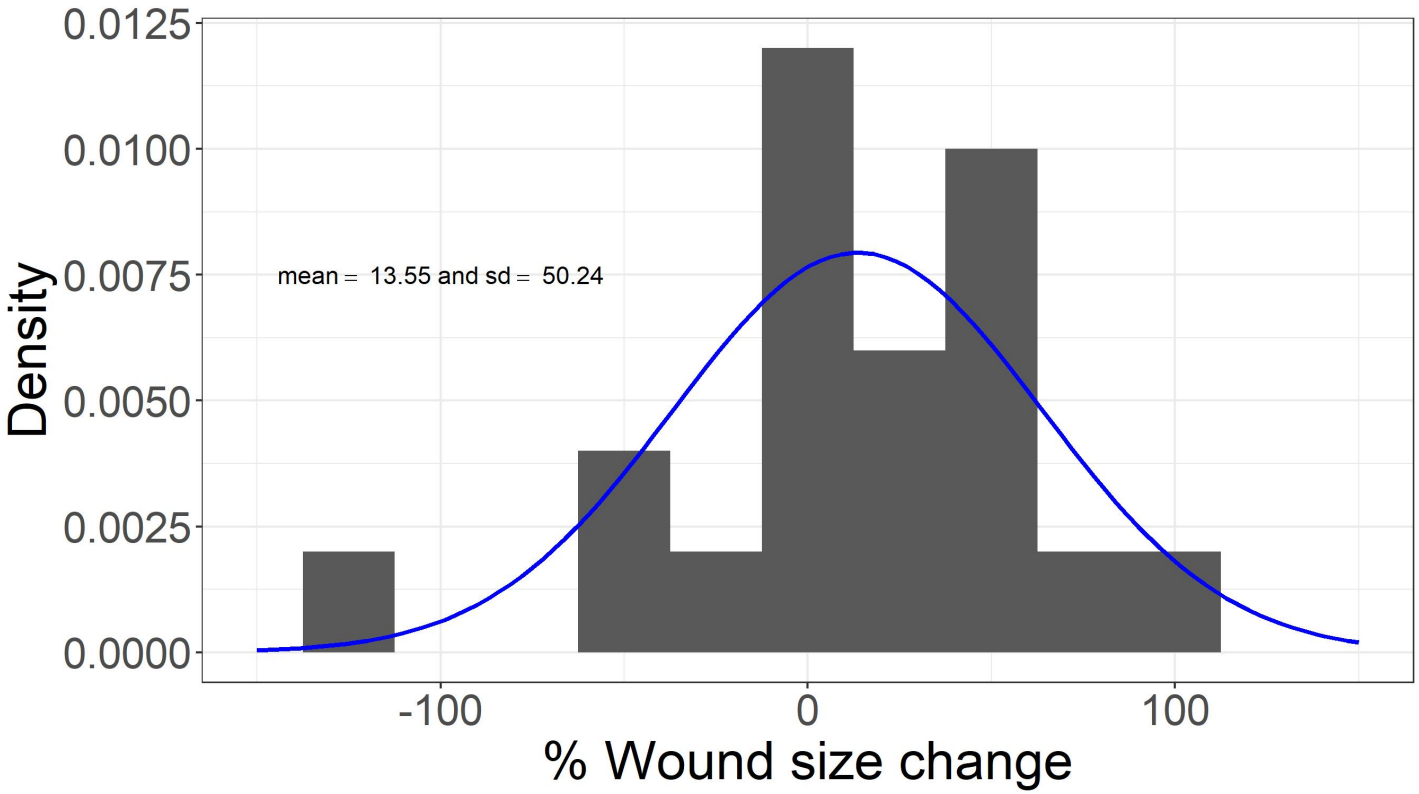



Figure: Histogram of percent of wound size change

# What are regression models?

## The impact of Manuka honey dressings on the surface pH of chronic wounds

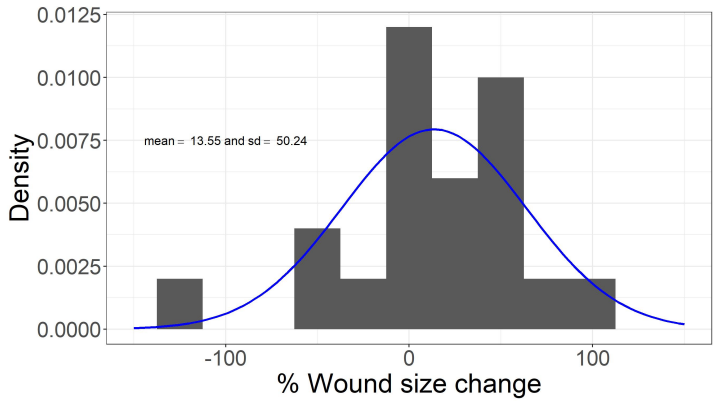



Figure: Histogram of percent of wound size change

### Conclusion - Probabilistic model

- 95% Confidence Interval for the average percent of wound size change is [-8.46% ; 35.57%].

# What are regression models?

## The impact of Manuka honey dressings on the surface pH of chronic wounds

- Is initial pH an important factor to explain wound size change?

# What are regression models?

## The impact of Manuka honey dressings on the surface pH of chronic wounds

- Is initial pH an important factor to explain wound size change?

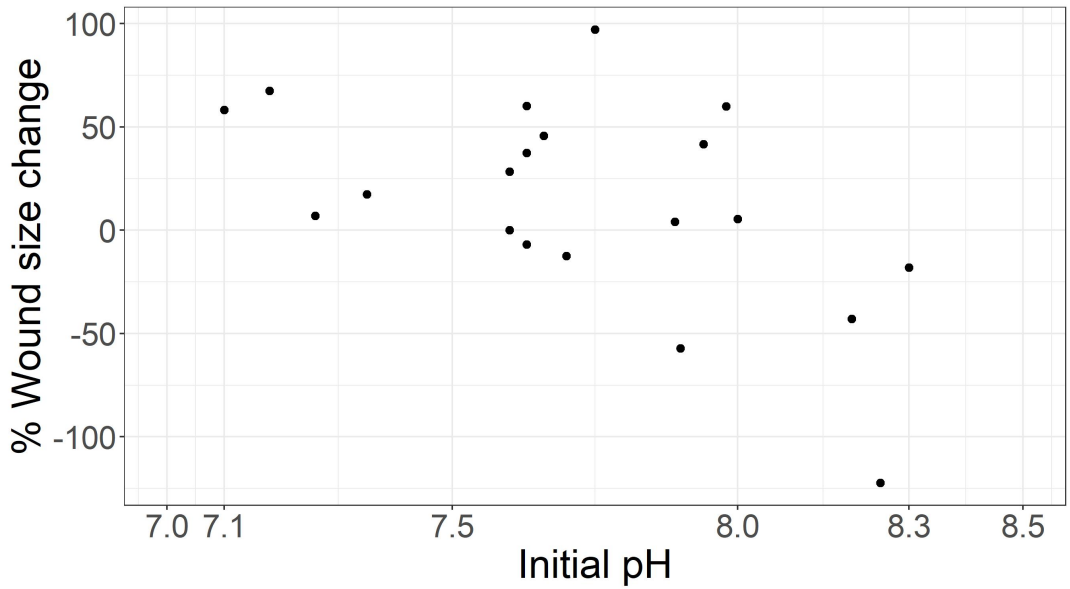



Figure: Scatterplot between percent of wound size change and initial wound pH

# What are regression models?

## What is correlation?

- It is a measure of how strongly pairs of variables are related;

# What are regression models?

## What is correlation?

- It is a measure of how strongly pairs of variables are related;
- There are several measures of correlation. The most common are Pearson and Spearman;

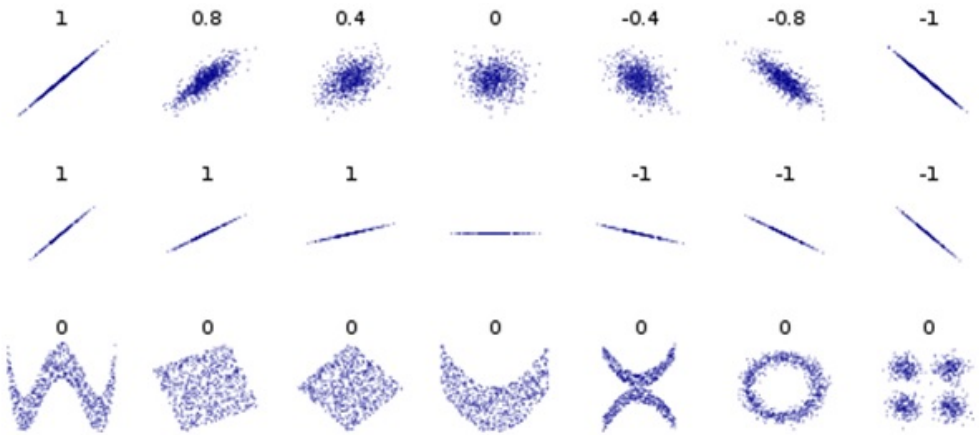



Figure: Pearson Correlation examples

# What are regression models?

## What is correlation?

- It is a measure of how strongly pairs of variables are related;
- There are several measures of correlation. The most common are Pearson and Spearman;
- Correlation does not imply causation.

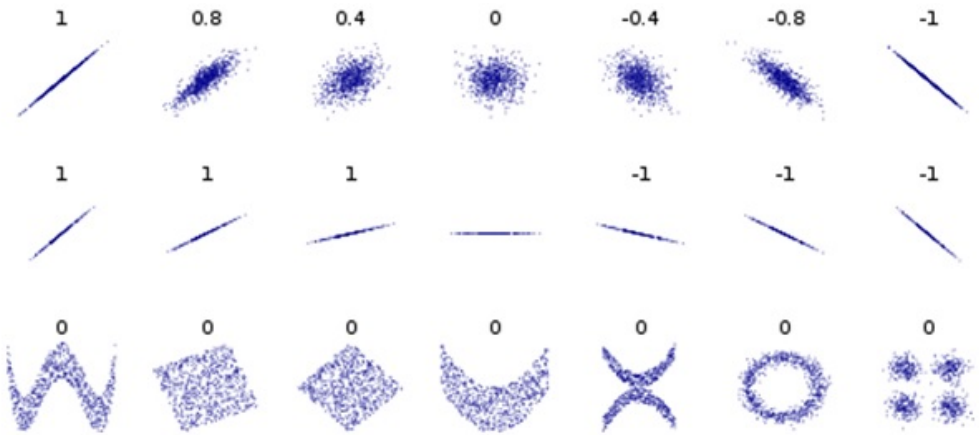



Figure: Pearson Correlation examples

## Attention

Correlation does not imply causation.

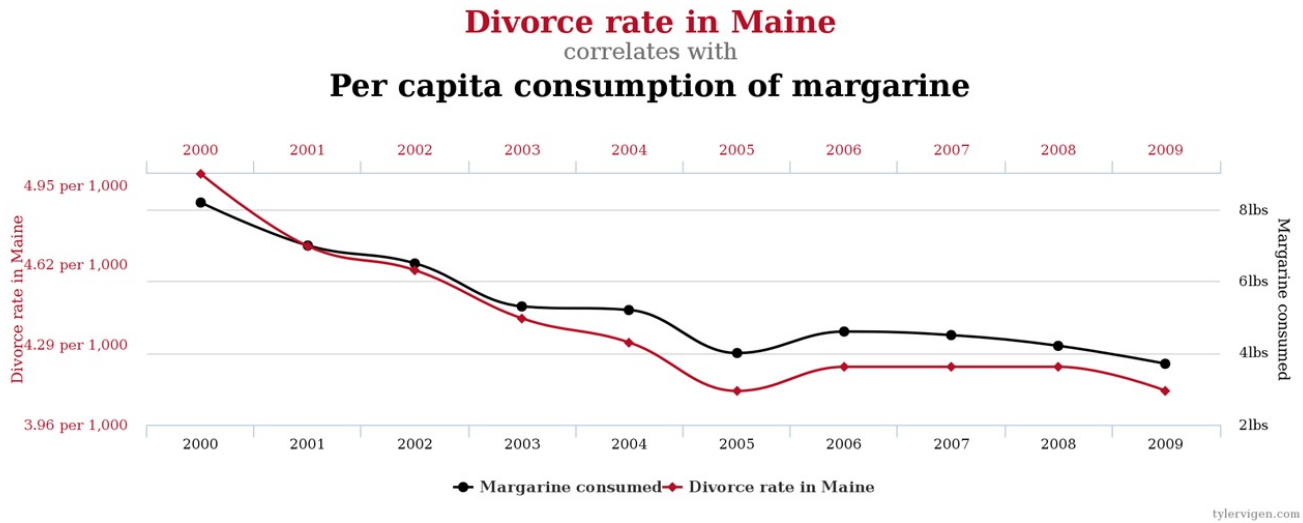



Figure: Data sources: U.S. Bureau of Transportation Statistics and Centers for Disease Control & Prevention

- $r = 0.992$

## Attention

Correlation does not imply causation.

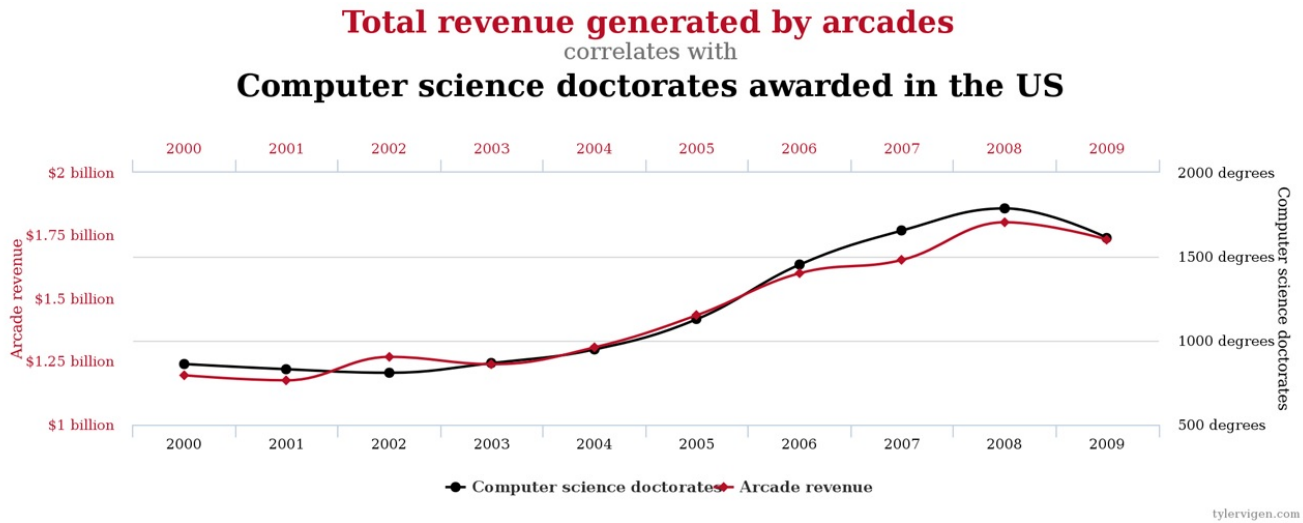



Figure: Data sources: U.S. Bureau of Transportation Statistics and Centers for Disease Control & Prevention

- $r = 0.985$

## Attention

Correlation does not imply causation.

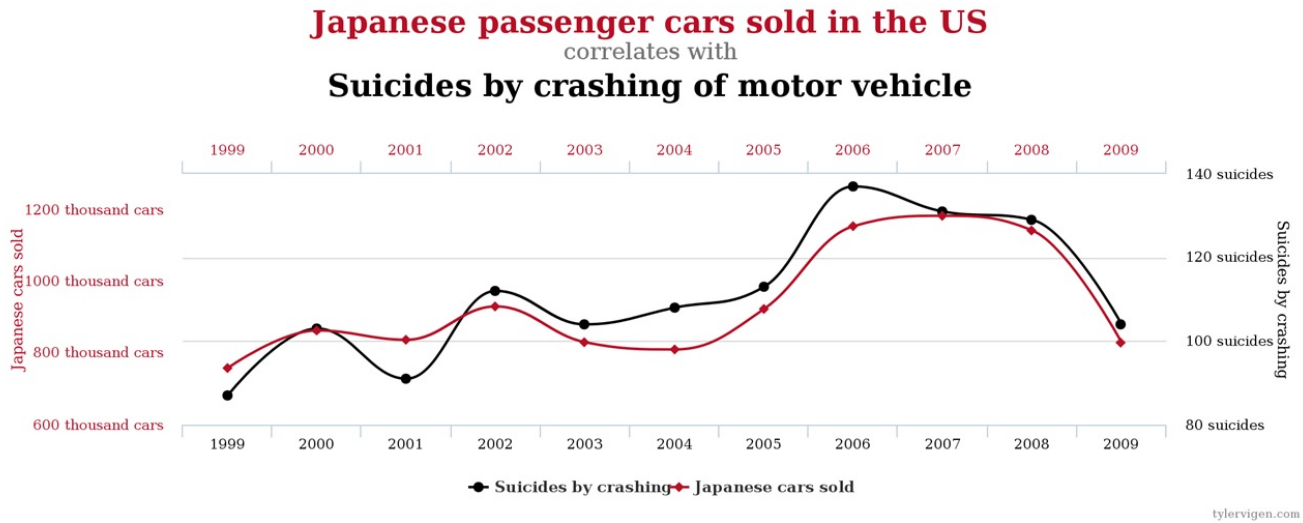



Figure: Data sources: U.S. Bureau of Transportation Statistics and Centers for Disease Control & Prevention

- $r = 0.93$

# What are regression models?

## The impact of Manuka honey dressings on the surface pH of chronic wounds

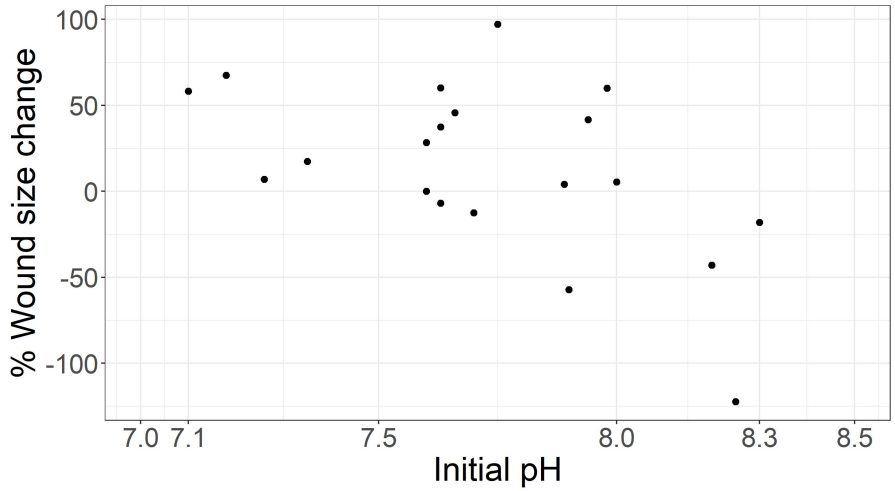



Figure: Scatterplot between percent of wound size change and initial wound pH

- Pearson correlation $r = -0.539$ with 95% confidence interval [-0.792 ; -0.127].

# What are regression models?

## The impact of Manuka honey dressings on the surface pH of chronic wounds

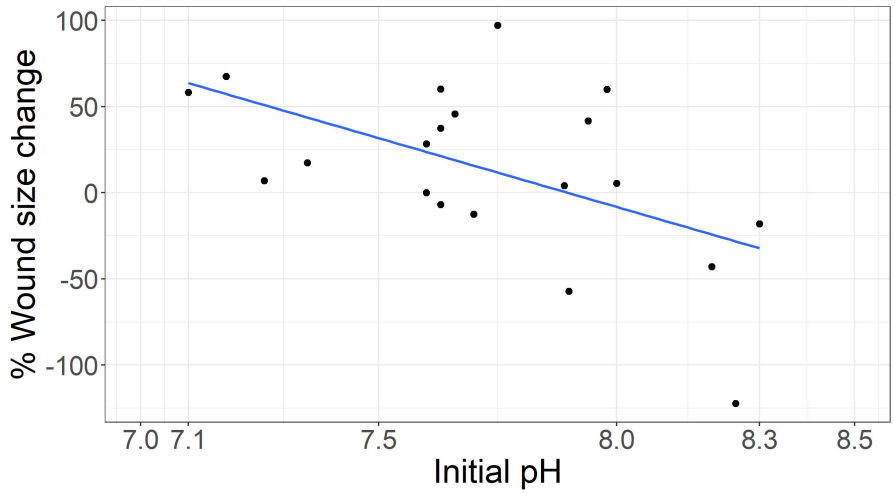



Figure: Linear regression allows us to make inferences about the relationship between percent of wound size change and initial wound pH

- Why we do not draw a line?

# What are regression models?

## Probabilistic model

- Let $Y$ be the percent of change of wound size;
- $Y$ is a continuous measure and it is random;
- $Y \sim Normal(\mu, \sigma^2)$ where $\mu$ and $\sigma$ represent the mean and the standard deviation, respectively, of the *Normal* distribution.
  - $\mu$ and $\sigma$ are constants.

# What are regression models?

## Regression model with normal errors

- Let $Y$ be the percent of wound size change;
- $Y$ is a continuous measure and it is random;
- $Y \sim Normal(\mu, \sigma^2)$ where $\mu$ and $\sigma$ represent the mean and the standard deviation, respectively, of the *Normal* distribution.
  - $\mu$ is a function of the initial pH;
  - $\sigma$ is constant.

# What are regression models?

## Regression model

- It has two components:
  - a probabilistic model for the response variable;
  - a function which describes the relationship between the parameters of a probabilistic model and explaining variables.
- Aims:
  - Study of the association between the response variables and possible explaining variables;
  - Predictions.

# Summary

# Linear regression

## The impact of Manuka honey dressings on the surface pH of chronic wounds

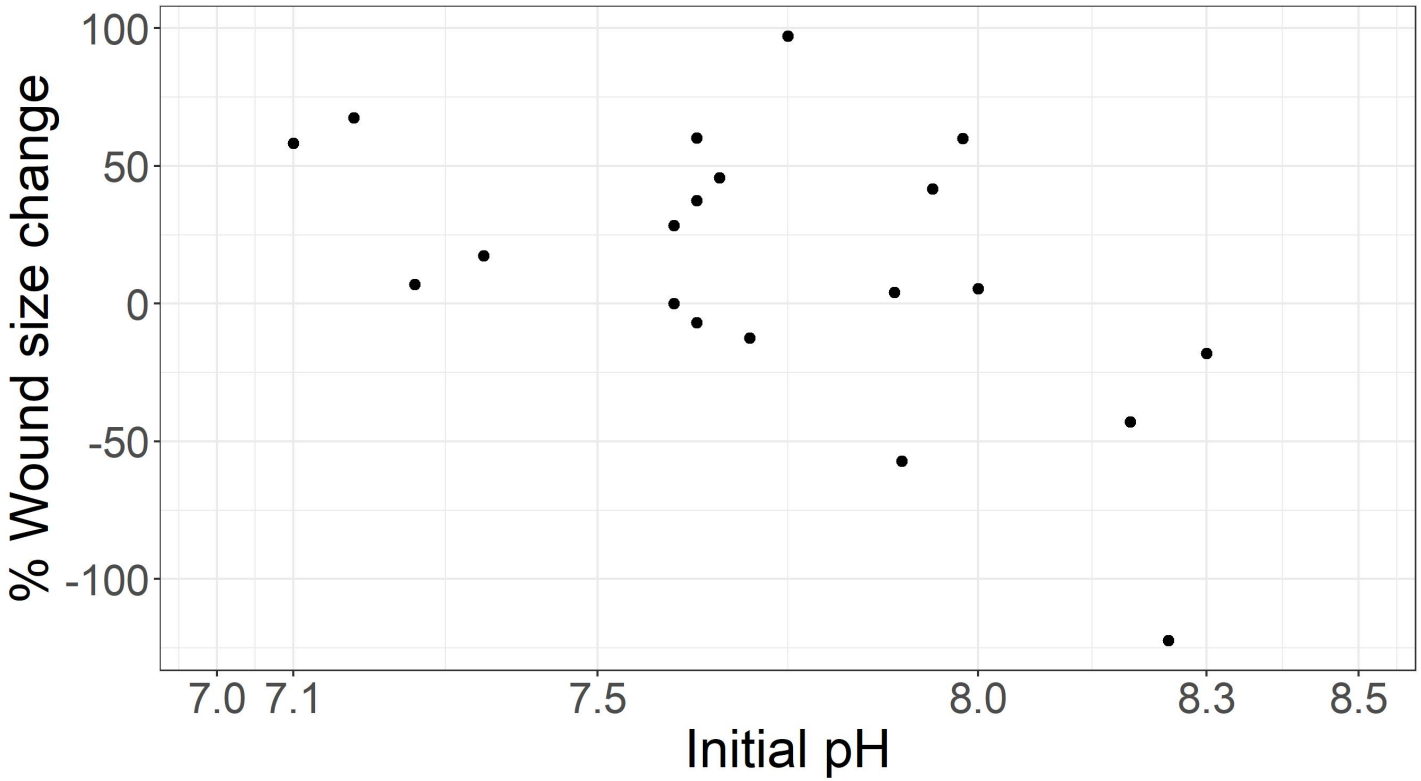



Figure: Scatterplot between percent of wound size change and initial wound pH

# Linear regression

## The impact of Manuka honey dressings on the surface pH of chronic wounds

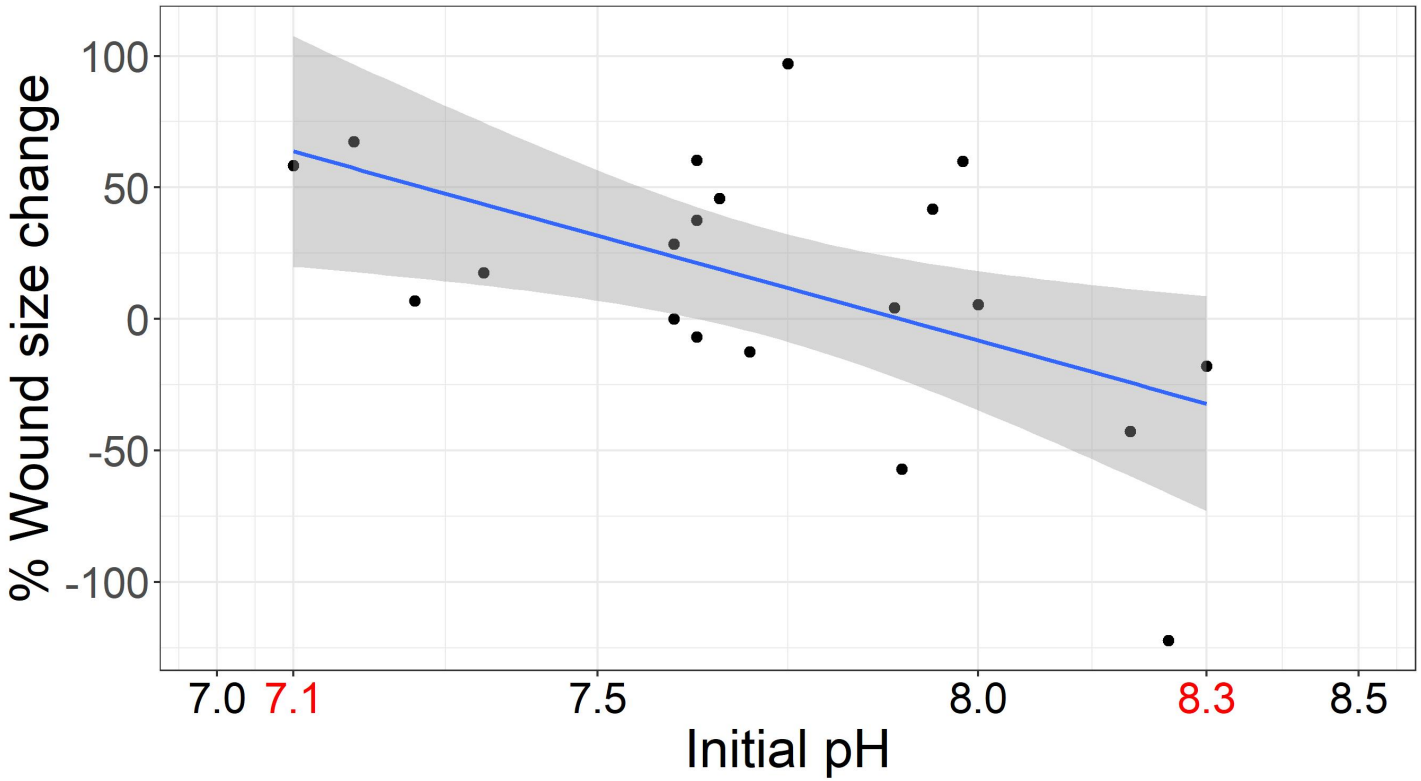



Figure: Scatterplot and regression line of average percent of wound size change ($\mu$) as fucntion of initial wound pH

## Univariable linear regression

- Let $Y$ be the percent of wound size change;
- $Y \sim Normal(\mu, \sigma^2)$;
- $\mu = \beta_0 + \beta_1 \times$ initial pH;
- $\sigma^2$ is constant.

# Linear regression

## The impact of Manuka honey dressings on the surface pH of chronic wounds

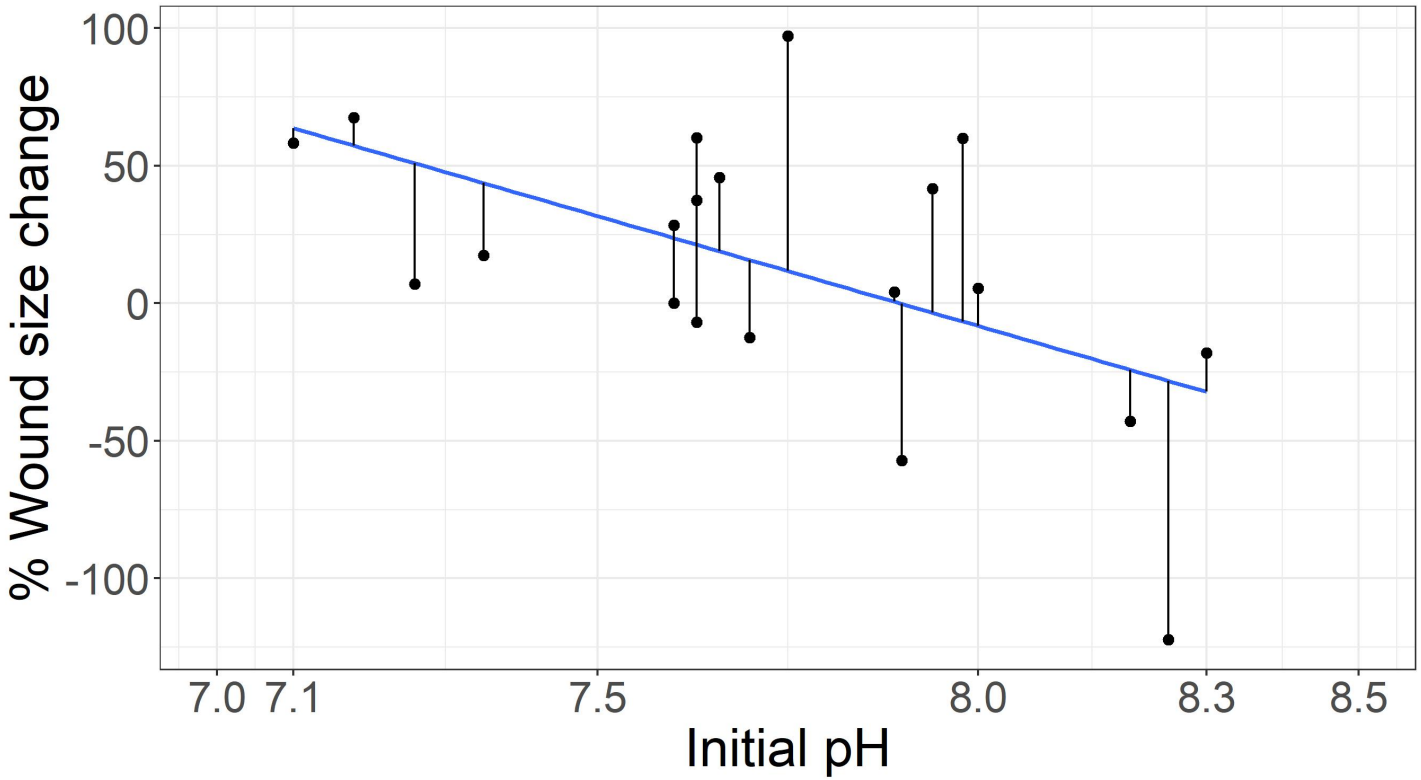



Figure: Fitted univariable linear model for the percent of wound size change as function of initial wound pH is the regression line which minimizes the average squared distance between the line and all the points

| Coefficients | Estimate | Std. Error | t value | p value |
|---|---|---|---|---|
| (Intercept) $\beta_0$ | 631.01 | 227.47 | 2.774 | 0.0125 |
| (Initial pH) $\beta_1$ | -79.9 | 29.41 | -2.717 | 0.0141 |

Table: Univariable linear fitted model

What do these p values mean?

- $H_0 : \beta_0 = 0 \qquad H_1 : \beta_0 \neq 0,$
- $H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0.$

| Coefficients | Estimate | Std. Error | t value | p value |
|---|---|---|---|---|
| (Intercept) $\beta_0$ | 631.01 | 227.47 | 2.774 | 0.0125 |
| (Initial pH) $\beta_1$ | -79.9 | 29.41 | -2.717 | 0.0141 |

Table: Univariable linear fitted model

What if $\beta_1$ is equal to zero?

- $Y \sim Normal(\mu, \sigma^2)$;
- $\mu = \beta_0 + 0 \times \text{initial pH}$;

| Coefficients | Estimate | Std. Error | t value | p value |
|---|---|---|---|---|
| (Intercept) $\beta_0$ | 631.01 | 227.47 | 2.774 | 0.0125 |
| (Initial pH) $\beta_1$ | -79.9 | 29.41 | -2.717 | 0.0141 |

Table: Univariable linear fitted model

## How to interpret the coefficients?

- If there are two patients such that their initial pH values are 7.2 and 8.2 then the percent of wound size change for the two patients would be, respectively,

$$\begin{aligned} \hat{Y}_1 &= 631.01 - 79.9 \times 7.2 = 55.73 \\ \hat{Y}_2 &= 631.01 - 79.9 \times 8.2 = -24.17 \end{aligned}$$

| Coefficients | Estimate | Std. Error | t value | p value |
|---|---|---|---|---|
| (Intercept) $\beta_0$ | 631.01 | 227.47 | 2.774 | 0.0125 |
| (Initial pH) $\beta_1$ | -79.9 | 29.41 | -2.717 | 0.0141 |

Table: Univariable linear fitted model

### How to interpret the coefficients?

- The percent of wound size change after manuka honey dressing for two weeks reduces around 80% of the wound size for each decrease of 1 unit of initial pH, i.e., reduces 8% of the wound size for each decreased 0.1 unit of initial pH.

# Linear regression

## The impact of Manuka honey dressings on the surface pH of chronic wounds

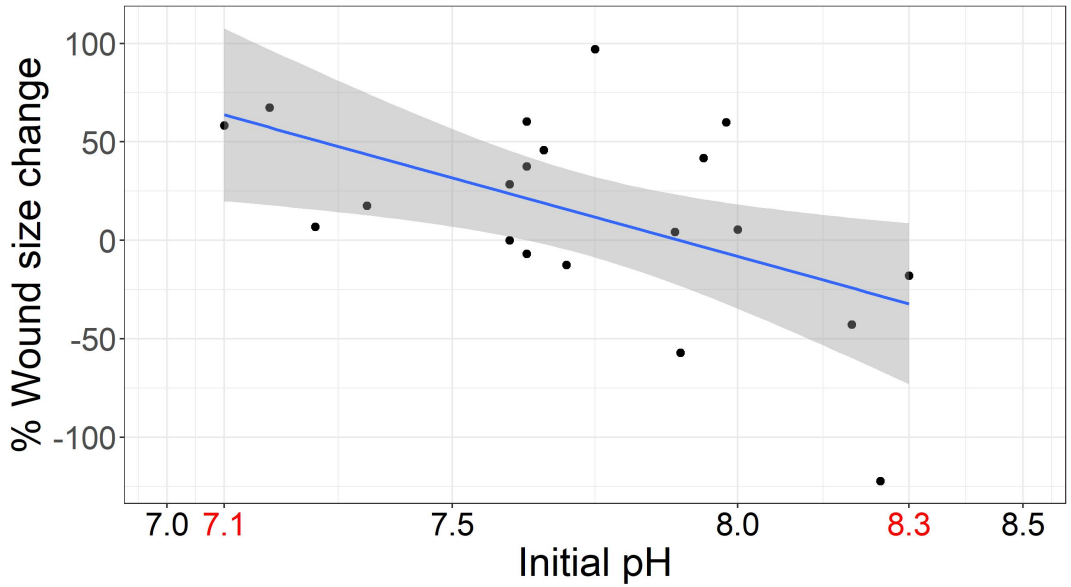



Figure: Fitted univariable linear model for the percent of wound size change as function of initial wound pH ranging from 7.1 to 8.3

**Attention**

- Do not make inferences outside the range of your data!

## Multivariable linear regression

- We also could explain the percent wound size change considering other covariates;

## Multivariable linear regression

- We also could explain the percent wound size change considering other covariates;
- Let be $Y$ the percent of wound size change;
- $Y \sim Normal(\mu, \sigma^2)$;
- $\mu = \beta_0 + \beta_1 \times \text{initial pH} + \beta_2 \times \text{venous ulcer}$;
- $\sigma^2$ is constant.

## Multivariable linear regression

- We also could explain the percent wound size change considering other covariates;
- Let be $Y$ the percent of wound size change;
- $Y \sim Normal(\mu, \sigma^2)$;
- $\mu = \beta_0 + \beta_1 \times \text{initial pH} + \beta_2 \times \text{venous ulcer}$;
- $\sigma^2$ is constant.

## Attention

Hidalgo, B. and Goodman, M., 2013. Multivariate or multivariable regression?. American journal of public health, 103(1), pp.39-40.

# What are regression models?

## The impact of Manuka honey dressings on the surface pH of chronic wounds

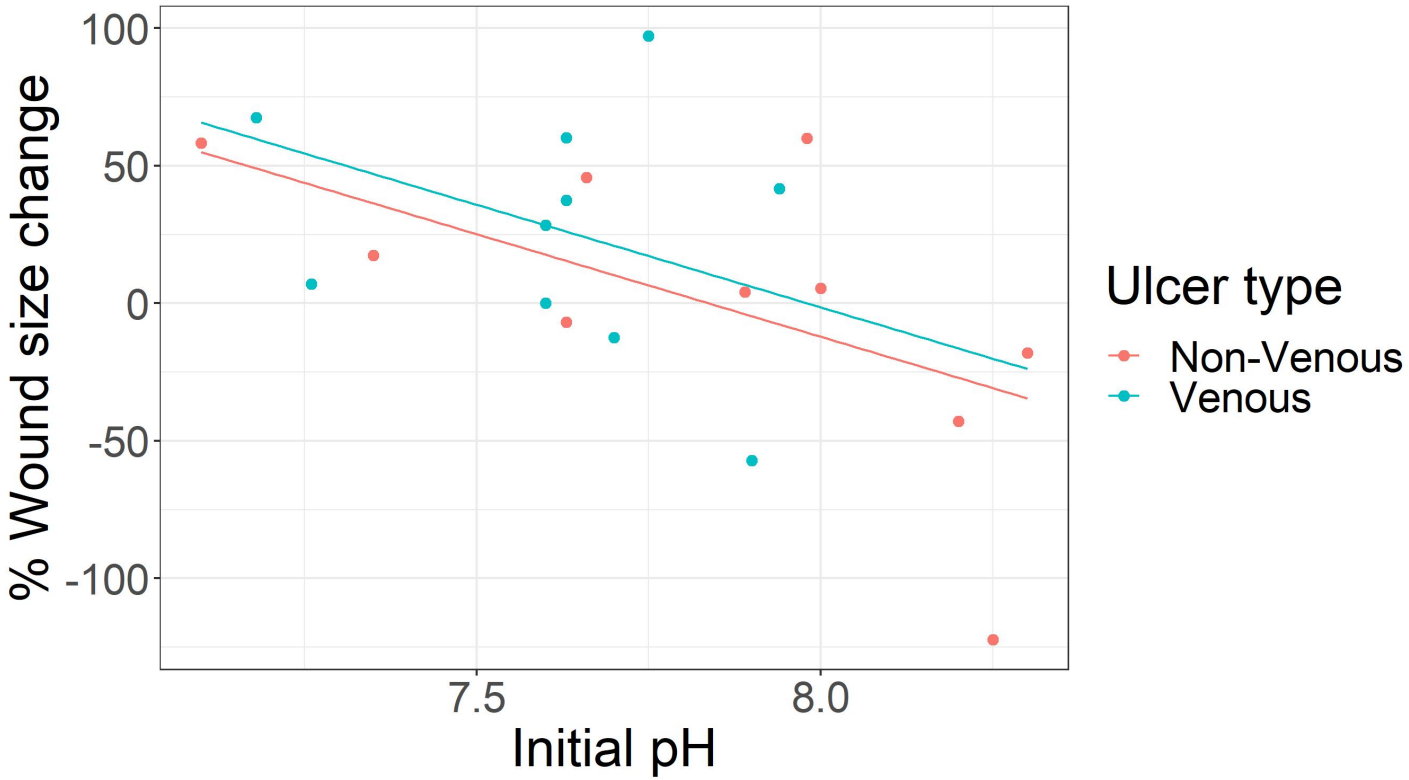



Figure: Fitted multivariable linear model between percent of wound size change, ulcer type and initial wound pH

# Linear regression

## The impact of Manuka honey dressings on the surface pH of chronic wounds

| Coefficients | Estimate | Std. Error | t value | p value |
|---|---|---|---|---|
| (Intercept) $\beta_0$ | 584.67 | 248.56 | 2.343 | 0.0316 |
| (Initial pH) $\beta_1$ | -74.6 | 31.80 | -2.346 | 0.0314 |
| (Ulcer type: Venous) $\beta_2$ | 10.68 | 21.02 | 0.508 | 0.6178 |

Table: Multivariable fitted model

| Coefficients | Estimate | Std. Error | t value | p value |
|---|---|---|---|---|
| (Intercept) $\beta_0$ | 584.67 | 248.56 | 2.343 | 0.0316 |
| (Initial pH) $\beta_1$ | -74.6 | 31.80 | -2.346 | 0.0314 |
| (Ulcer type: Venous) $\beta_2$ | 10.68 | 21.02 | 0.508 | 0.6178 |

Table: Multivariable fitted model

### How to interpret the coefficients?

- If there are two patients such that their initial pH value is the same given by 7.2, but their lesion types are venous and non-venous, respectively. Then the percent of would size change for the two patients would be,

$$\begin{aligned} \hat{Y}_1 &= 584.67 - 74.6 \times 7.2 + 10.68 = 58.23 \\ \hat{Y}_2 &= 584.67 - 74.6 \times 7.2 = 47.55 \end{aligned}$$

| Coefficients | Estimate | Std. Error | t value | p value |
|---|---|---|---|---|
| (Intercept) $\beta_0$ | 584.67 | 248.56 | 2.343 | 0.0316 |
| (Initial pH) $\beta_1$ | -74.6 | 31.80 | -2.346 | 0.0314 |
| (Ulcer type: Venous) $\beta_2$ | 10.68 | 21.02 | 0.508 | 0.6178 |

Table: Multivariable fitted model

### How to interpret the coefficients?

- If there are two patients such that their initial pH values are 7.2 and 8.2 (disregarding if the ulcer is venous or non-venous), the percent of wound size change after manuka honey dressing for two weeks reduces 74.6% for each decrease of 1 unit of initial pH, i.e., reduces 7.46% the wound size for each decrease of 0.1 unit of initial pH.

# Linear regression

## The impact of Manuka honey dressings on the surface pH of chronic wounds

| Coefficients | Estimate | Std. Error | t value | p value |
|---|---|---|---|---|
| (Intercept) $\beta_0$ | 584.67 | 248.56 | 2.343 | 0.0316 |
| (Initial pH) $\beta_1$ | -74.6 | 31.80 | -2.346 | 0.0314 |
| (Ulcer type: Venous) $\beta_2$ | 10.68 | 21.02 | 0.508 | 0.6178 |

Table: Multivariable fitted model

| Coefficients | Estimate | Std. Error | t value | p value |
|---|---|---|---|---|
| (Intercept) $\beta_0$ | 584.67 | 248.56 | 2.343 | 0.0316 |
| (Initial pH) $\beta_1$ | -74.6 | 31.80 | -2.346 | 0.0314 |
| (Ulcer type: Venous) $\beta_2$ | 10.68 | 21.02 | 0.508 | 0.6178 |

Table: Multivariable fitted model

How much of the variability of the response variable (% wound size change) is explained by the covariables (initial pH, ulcer type)?

- R-Squared = 0.3014;
- Adjusted R-Squared = 0.2193.
  - 21.9% of the observed variability of the percent wound size change is explained by initial pH.

International Wound Journal ISSN 1742-4801

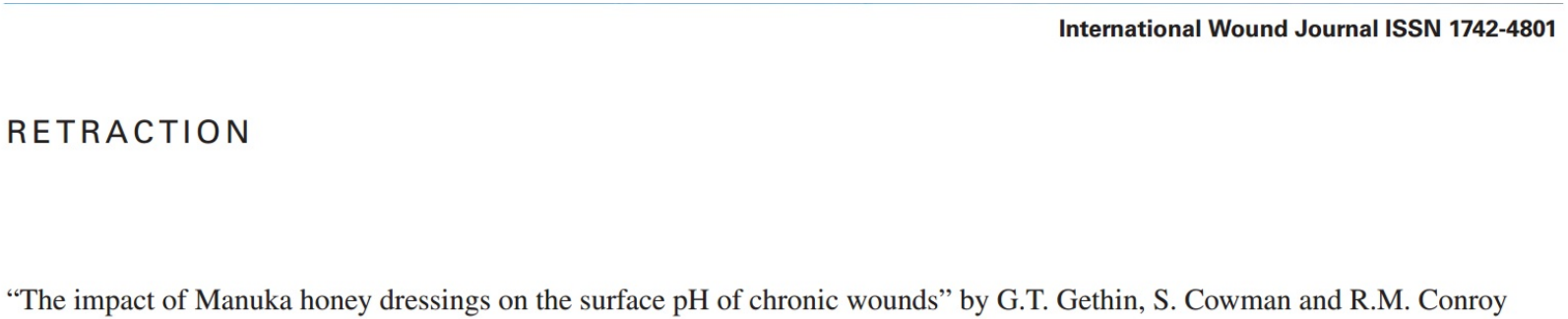
RETRACTION

"The impact of Manuka honey dressings on the surface pH of chronic wounds" by G.T. Gethin, S. Cowman and R.M. Conroy

Figure: Why?

# Linear regression

## The impact of Manuka honey dressings on the surface pH of chronic wounds



International Wound Journal ISSN 1742-4801

RETRACTION

"The impact of Manuka honey dressings on the surface pH of chronic wounds" by G.T. Gethin, S. Cowman and R.M. Conroy

Figure: Why?

### Retraction

The authors did not verify the assumptions of the regression model as well as the goodness of fit. If these steps are performed, then their conclusions are no longer statistically significant.

## Regression Diagnostics

Every regression model has assumptions that need to be verified. Otherwise, the inferences based on such model can be not valid.

- Unusual and Influential Data:
  - Outlier: High values for the response;
  - Leverage: High values for the covariables;
  - Influence: High values for the response and covariables.
- Heteroscedasticity: $\sigma^2$ is not constant;
- Normality;
- Multicolinearity: Several covariates are correlated.

# Linear regression

## The impact of Manuka honey dressings on the surface pH of chronic wounds

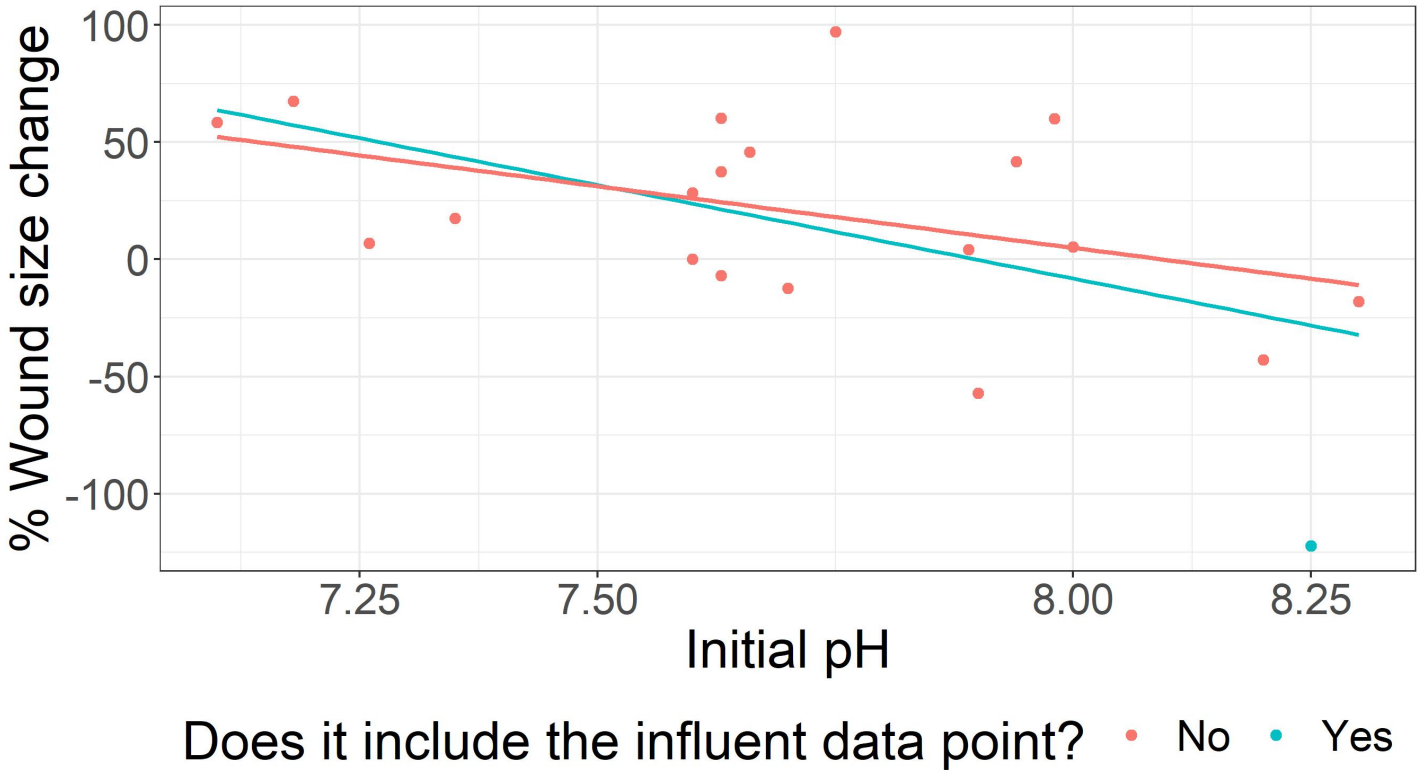



Figure: Fitted simple linear model between percent of wound size change and initial wound pH considering or not an influential point.

| Coefficients | Estimate | Std. Error | t value | p value |
|---|---|---|---|---|
| (Intercept) $\beta_0$ | 631.01 | 227.47 | 2.774 | 0.0125 |
| (Initial pH) $\beta_1$ | -79.9 | 29.41 | -2.717 | 0.0141 |

Table: Fitted simple linear model with complete data

| Coefficients | Estimate | Std. Error | t value | p value |
|---|---|---|---|---|
| (Intercept) $\beta_0$ | 426.08 | 207.11 | 2.057 | 0.0553 |
| (Initial pH) $\beta_1$ | -52.65 | 26.87 | -1.959 | 0.0667 |

Table: Fitted simple linear model with deleting the influential observation

# Regression models

## Summary

- Models are sometimes useful approximations of reality;
- Linear regression can be used to unravel associations and make predictions;
- Diagnostics to verify the assumptions are always a requirement.

# Regression models

## Beyond linear regression

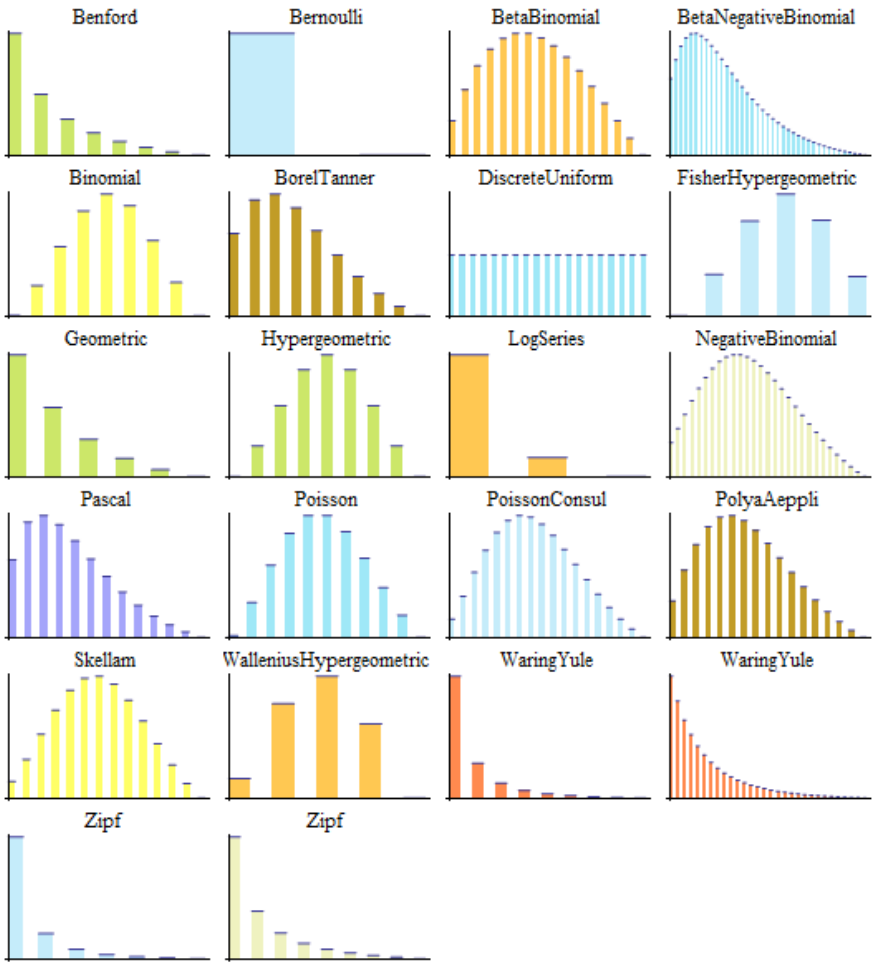



Figure: Discrete probabilistic models

# Regression models

## Beyond linear regression

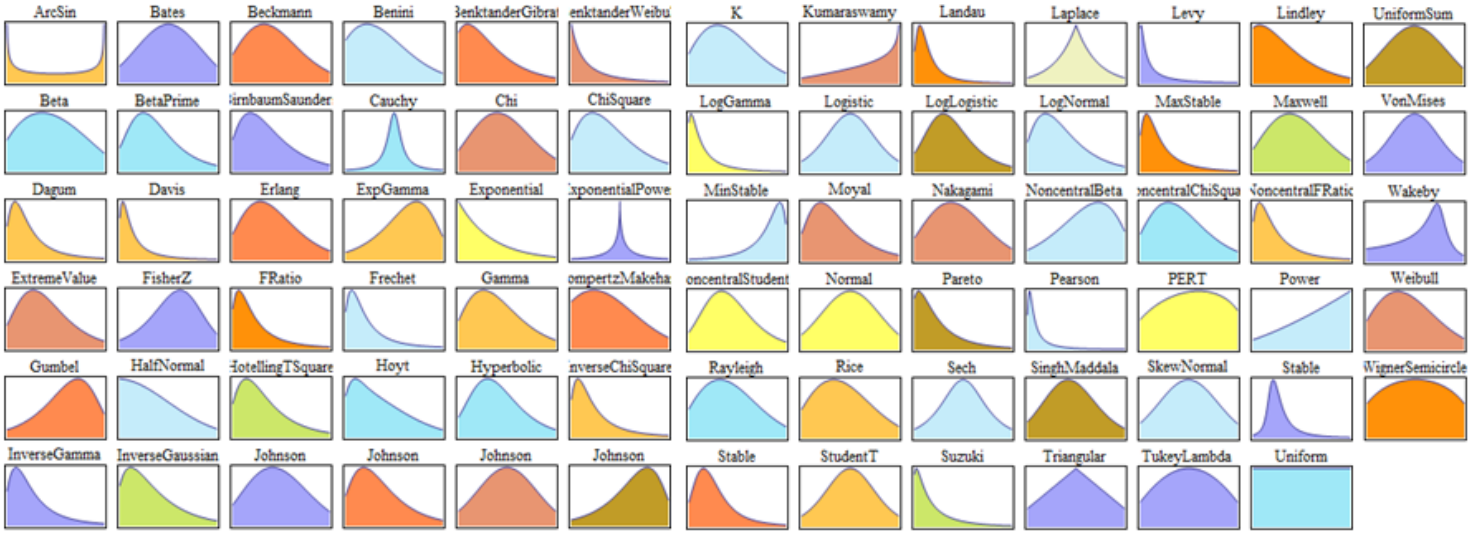



Figure: Continuous probabilistic models

- Nature - Points of Significance: Linear regression
- Nikolaou V. Statistical analysis: a practical guide for psychiatrists. BJPsych Advances. 2016 Jul 1;22(4):251-9.
- Worster A, Fan J, Ismaila A. Understanding linear and logistic regression analyses. CJEM. 2007 Mar 1;9(02):111-3;
- Generalized Additive Models for Location, Scale and Shape.

Questions?
marcio.diniz@cshs.org
Biostatistics Core Request Form