Logistic Regression

Statistics in Medical Research Fall Series

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

October 25, 2022

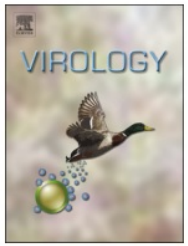
Summary

1. Introduction
2. Relative effects
3. Logistic Regression
4. Predictive Model

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro



Interferon lambda and hepatitis C virus core protein polymorphisms associated with liver cancer

João Paulo Moreira [a,b], Fernanda de Mello Malta [a,b,*], Márcio Augusto Diniz [b,c], Luciana Kikuchi [b], Aline Lopes Chagas [b], Livia de Souza Botelho Lima [a,b], Michele Soares Gomes-Gouvêa [a,b], Vanessa Fusco Duarte de Castro [d], Rubia Anita Ferraz Santana [d], Nairo Massakazu Sumita [e], Denise Cerqueira Paranagua Vezozzo [b], Flair José Carrilho [b], João Renato Rebello Pinho [a,b,d]

Example

- Hypothesis: Polymorphisms could be associated to liver damage in chronic hepatitis C;
- Groups: HCC (Hepatocellular Carcinoma) and Cirrhotic without HCC;
- Polymorphisms: rs12980275 (AA/ AG+GG);
- Design: Case-Control.

rs12980275	HCC	Cirrhotic without HCC	Total
AG + GG	43	29	72
AA	16	21	37
Total	59	50	109

HCC: Hepatocellular Carcinoma

Table: Distribution genotype by group

How should we compare the groups?

rs12980275	HCC	Cirrhotic without HCC	Total
AG + GG	43	29	72
AA	16	21	37
Total	59	50	109

HCC: Hepatocellular Carcinoma

Table: Distribution genotype by group

How should we compare the groups?

- Our first attempt would be to calculate the proportions $43/72 = 0.59$ and $16/37 = 0.43$.

rs12980275	HCC	Cirrhotic without HCC	Total
AG + GG	43	29	72
AA	16	21	37
Total	59	50	109

HCC: Hepatocellular Carcinoma

Table: Distribution genotype by group

How should we compare the groups?

- Our first attempt would be to calculate the proportions $43/72 = 0.59$ and $16/37 = 0.43$.
- However, they cannot be calculated. Why?

rs12980275	HCC	Cirrhotic without HCC	Total
AG + GG	43	29	72
AA	16	21	37
Total	59	50	109

HCC: Hepatocellular Carcinoma

Table: Distribution genotype by group

How should we compare the groups?

- Our first attempt would be to calculate the proportions $43/72 = 0.59$ and $16/37 = 0.43$.
- However, they cannot be calculated. Why?
- The study is retrospective.

Summary

1 Introduction

2 Relative effects

3 Logistic Regression

4 Predictive Model

What type of study do you have?

Prospective

- Prospective studies are carried out from the present time into the future;

Retrospective

- Retrospective cohort studies are carried out at the present time and look to the past to examine medical events or outcomes;

Song JW, Chung KC. Observational studies: cohort and case-control studies. Plastic and reconstructive surgery. 2010 Dec;126(6):2234.

What type of study do you have?

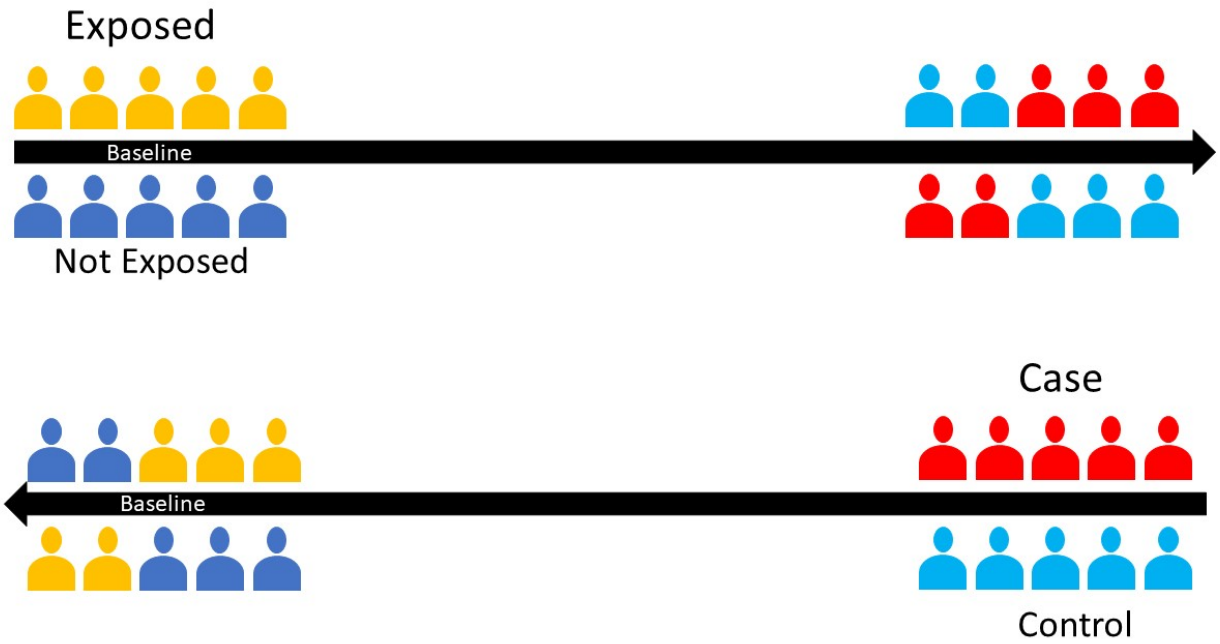


Figure: Prospective Cohort and Case-Control

Relative risk

What is relative risk?

Group	Disease	Non-Disease	Total
Exposed	a	b	E
Non-Exposed	c	d	NE
Total	D	ND	n

Table: 2×2

Prospective study

- The number of exposed (E) and non-exposed (NE) patients are defined in advance then it is possible to calculate the probabilities

$$P(disease|exposed) = \frac{a}{E}, \qquad P(disease|non-exposed) = \frac{c}{NE};$$

Relative risk

What is relative risk?

Group	Disease	Non-Disease	Total
Exposed	a	b	E
Non-Exposed	c	d	NE
Total	D	ND	n

Table: 2×2

Prospective study

- Then the effect measure relative risk (RR) is straightforward calculated by

$$RR = \frac{P(disease|exposed)}{P(disease|non-exposed)} = \frac{\frac{a}{E}}{\frac{c}{NE}} = \frac{a \times NE}{E \times c}.$$

Odds ratio

What is odds ratio?

Group	Disease	Non-Disease	Total
Exposed	a	b	E
Non-exposed	c	d	NE
Total	D	ND	n

Table: 2×2

Retrospective study

- The number exposed and non-exposed patients are not defined in advance, therefore the probabilites

$$P(disease|exposed) = \frac{a}{E}, \qquad P(disease|non-exposed) = \frac{c}{NE};$$

cannot be calculated and compared using relative risk.

Odds ratio

What is odds ratio?

Group	Disease	Non-Disease	Total
Exposed	a	b	E
Non-exposed	c	d	NE
Total	D	ND	n

Table: 2×2

Retrospective study

- We can calculate the odds for exposed patients

$$\begin{aligned} Odds(disease|exposed) &= \frac{P(disease|exposed)}{P(non-disease|exposed)}; \\ &= \frac{\frac{a}{E}}{\frac{b}{E}} = \frac{a}{b}; \end{aligned}$$

Odds ratio

What is odds ratio?

Group	Disease	Non-Disease	Total
Exposed	a	b	E
Non-exposed	c	d	NE
Total	D	ND	n

Table: 2×2

Retrospective study

- We can calculate the odds for non-exposed patients

$$\begin{aligned} Odds(disease|non-exposed) &= \frac{P(disease|non-exposed)}{P(non-disease|non-exposed)}; \\ &= \frac{\frac{c}{NE}}{\frac{d}{NE}} = \frac{c}{d}; \end{aligned}$$

Odds ratio

What is odds ratio?

Group	Disease	Non-Disease	Total
Exposed	a	b	E
Non-exposed	c	d	NE
Total	D	ND	n

Table: 2×2

Retrospective study

- Then, the effect measure odds ratio (OR) is calculated by

$$OR = \frac{Odds(disease|exposed)}{Odds(disease|non-exposed)} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \times d}{b \times c}.$$

Odds ratio and Relative risk

Example

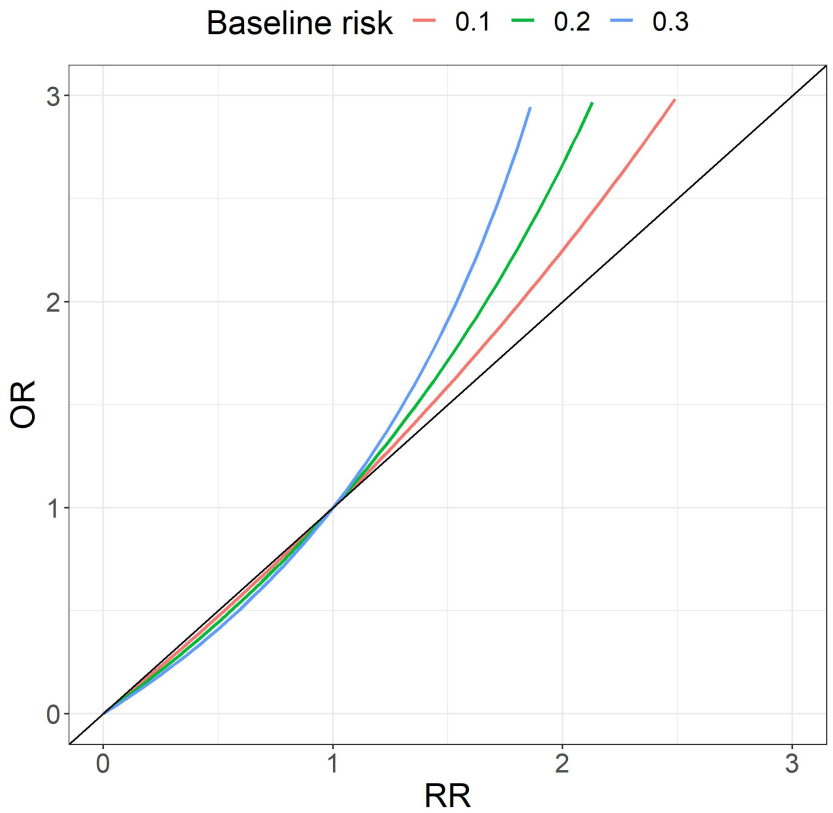


Figure: Relationship between odds ratio and relative risk for several baseline risks

Relative effects

Attention

- Relative effects are calculated based on the risk or odds of a reference group;
- If the risk/odds of a reference group is small, large values of OR and RR could not be meaningful;
- It is always possible to calculate absolute risk in a prospective study, but not in a retrospective study.

Odds ratio

Interferon lambda and hepatitis C virus core protein polymorphisms associated with liver cancer

rs12980275	HCC	Cirrhotic without HCC	Total
AG + GG	43	29	72
AA	16	21	37
Total	59	50	109

HCC: Hepatocellular Carcinoma

Table: Distribution of genotype by group

How should we compare the groups?

- Now, we can compare the groups:

$$Odds(HCC|AG + GG) = \frac{43}{29}, Odds(HCC|AA) = \frac{16}{21};$$

Odds ratio

Interferon lambda and hepatitis C virus core protein polymorphisms associated with liver cancer

rs12980275	HCC	Cirrhotic without HCC	Total
AG + GG	43	29	72
AA	16	21	37
Total	59	50	109

HCC: Hepatocellular Carcinoma

Table: Distribution of genotype by group

How should we compare the groups?

- Now, we can compare the groups:

$$OR(HCC|AG + GG : AA) = \frac{43 \times 21}{29 \times 16} = 1.94;$$

Odds ratio

Interferon lambda and hepatitis C virus core protein polymorphisms associated with liver cancer

rs12980275	HCC	Cirrhotic without HCC	Total
AG + GG	43	29	72
AA	16	21	37
Total	59	50	109

HCC: Hepatocellular Carcinoma

Table: Distribution of genotype by group

How should we compare the groups?

- H_0: there is no association between genotype and HCC;
- H_1: there is association between genotype and HCC;

Odds ratio

Interferon lambda and hepatitis C virus core protein polymorphisms associated with liver cancer

rs12980275	HCC	Cirrhotic without HCC	Total
AG + GG	43	29	72
AA	16	21	37
Total	59	50	109

HCC: Hepatocellular Carcinoma

Table: Distribution of genotype by group

How should we compare the groups?

- H_0: OR $= 1$;
- H_1: OR $\neq 1$;
- Chi-square test p value $= 0.11$.

Summary

1. Introduction
2. Relative effects
3. Logistic Regression
4. Predictive Model

Logistic regression

Interferon lambda and hepatitis C virus core protein polymorphisms associated with liver cancer

Regression model

- Let be Y the presence of HCC;
- Y is a categorical measure;
- $Y \sim Bernouli(p)$ where p is the probability of the patient having HCC;
- p is a function of the SNP rs12980275:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 SNP : AG + GG.$$

Logistic regression

Interferon lambda and hepatitis C virus core protein polymorphisms associated with liver cancer

Coefficients	Estimate	Std. Error	t value	p value
(Intercept) β_0	-0.27	0.33	-0.819	0.413
(rs12980275) β_1	0.66	0.41	1.625	0.104

Table: Fitted Simple Logistic model

What do these p values mean?

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

Logistic regression

Interferon lambda and hepatitis C virus core protein polymorphisms associated with liver cancer

Coefficients	Estimate	Std. Error	t value	p value
(Intercept) β_0	-0.27	0.33	-0.819	0.413
(rs12980275) β_1	0.66	0.41	1.625	0.104

Table: Fitted Simple Logistic model

How to interpret the coefficients?

- We calculate the odds ratio,

$$OR(HCC|AG + GG : AA) = \exp\{\beta_1\} = 1.94.$$

Logistic regression

Interferon lambda and hepatitis C virus core protein polymorphisms associated with liver cancer

Coefficients	Estimate	Std. Error	t value	p value
(Intercept) β_0	-0.27	0.33	-0.819	0.413
(rs12980275) β_1	0.66	0.41	1.625	0.104

Table: Fitted Simple Logistic model

How to interpret the coefficients?

- The group with genotype AG+GG is 1.94 times (95% CI: 0.87 ; 4.39) more likely to be associated with HCC than the genotype AA.

Logistic regression

Interferon lambda and hepatitis C virus core protein polymorphisms associated with liver cancer

Coefficients	Estimate	Std. Error	t value	p value
(Intercept) β_0	-0.27	0.33	-0.819	0.413
(rs12980275) β_1	0.66	0.41	1.625	0.104

Table: Fitted Simple Logistic model

What are the advantages of a logistic regression from the Table 2×2?

- Odds ratios adjusted by confounding variables can be calculated and continuous covariables can be incorporated without cut-offs.

Logistic regression

Interferon lambda and hepatitis C virus core protein polymorphisms associated with liver cancer

Regression model

- Let be Y the presence of HCC;
- Y is a categorical measure;
- $Y \sim Bernouli(p)$ where p is the probability of the patient having HCC;
- p is a function of the SNP rs12980275:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 age + \beta_2 gender + \beta_3 SNP : AG + GG.$$

Logistic regression

Interferon lambda and hepatitis C virus core protein polymorphisms associated with liver cancer

Coefficients	Estimate	Std. Error	t value	p value
(Intercept) β_0	-0.27	0.33	-0.819	0.413
(Age) β_1	0.07	0.02	2.684	0.007
(Gender) β_2	0.78	0.43	1.802	0.071
(SNP:AG+GG) β_3	0.66	0.44	2.019	0.043

Table: Fitted Multivariable Logistic model

Logistic regression

Interferon lambda and hepatitis C virus core protein polymorphisms associated with liver cancer

How to interpret the coefficients?

- We calculated the odds ratio followed by its confidence interval,

$$OR(HCC|AG + GG : AA) = exp(0.66) = 2.41,$$
$$95\% CI[1.04; 5.82].$$

- The group with genotype AG+GG is 2.41 times (95% CI: 1.04 ; 5.82) more likely to be associated with HCC than the genotype AA.

Logistic regression

Interferon lambda and hepatitis C virus core protein polymorphisms associated with liver cancer

Coefficients	Estimate	Std. Error	t value	p value
(Intercept) β_0	-0.27	0.33	-0.819	0.413
(Age) β_1	0.07	0.02	2.684	0.007
(Gender) β_2	0.78	0.43	1.802	0.071
(SNP:AG+GG) β_3	0.66	0.44	2.019	0.043

Table: Fitted Multivariable Logistic model

Logistic regression

Interferon lambda and hepatitis C virus core protein polymorphisms associated with liver cancer

How to interpret the coefficients?

- We calculated the odds ratio followed by its confidence interval,

$$OR(HCC|(x+1):x) = exp(0.07) = 1.07, 95\%CI[1.02; 5.12].$$

- A patient of age **x+1** has odds 1.07 (95% CI: 1.02 ; 5.12) of having HCC times higher than a patient of age **x**.

Logistic regression

Interferon lambda and hepatitis C virus core protein polymorphisms associated with liver cancer

How to interpret the coefficients?

- We calculated the odds ratio followed by its confidence interval,

$$OR(HCC|(x+5):x) = exp(5 \times 0.07) = 1.07^5,$$
$$95\% CI[1.02^5; 5.12^5].$$

How to interpret the coefficients?

- We calculated the odds ratio followed by its confidence interval,

$$OR(HCC|(x+5):x) = 1.40, 95\%CI[1.10; 3518.437].$$

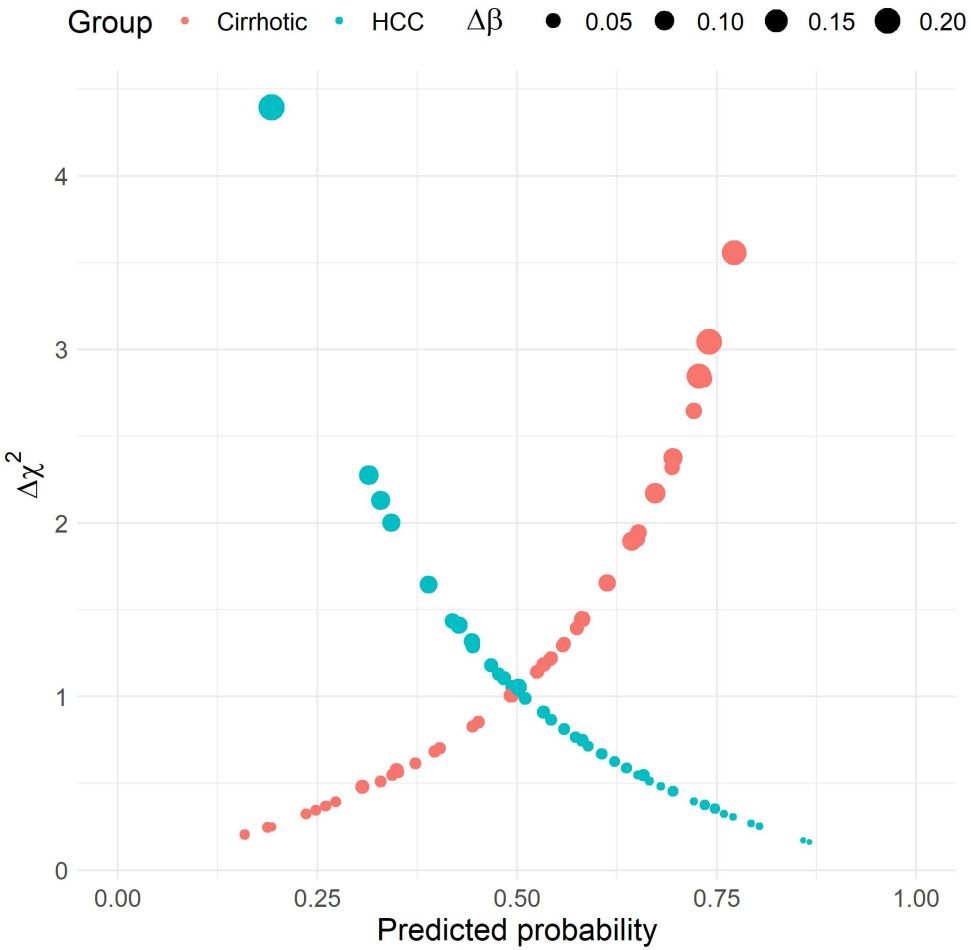
- A patient of age **x+5** has odds 1.40 (95% CI: 1.10 ; 3518.437) of having HCC times higher than a patient of age **x**.

Diagnostics

- Similarly to linear regression, logistic regression also requires diagnostic methods;
- Typical measures are $\Delta\beta$ and $\Delta\chi^2$ that measures the change in the regression coefficients and Chi-square statistic when an observation is removed;
- If the a fitted model suffers drastic changes on the estimates, then the observation is influential and results should be interpreted carefully.

Logistic regression

Diagnostics



Summary

1. Introduction
2. Relative effects
3. Logistic Regression
4. Predictive Model

Logistic regression

Stroke Trajectory Project

- Objective: To develop prediction models to advice patients on quality of life (QOL) and caregiving needs.
- Study population: 1495 stroke patients discharged from acute care hospital are available in the database.
- Joint work with Sungjin Kim, Pamela Roberts and Harriet Aronow.

Outcomes

- Functional Independence Measure (< 80 vs. $>= 80$);
- Eating (All Others vs. 6/7);
- Dressing Upper (All Others vs. 6/7);
- Dressing Lower (All Others vs. 6/7);
- Toileting (All Others vs. 6/7);
- Walking (All Others vs. 6/7).

Covariates

- Gender;
- Age at admission;
- Marital status;
- Race (White, Black/AA, or Other);
- Modified Rankin at DC;
- NIHSS;
- Impairment Group Code;
- Diagnosis;
- DC Destination (Home vs. Institution);
- Length of Stay.

Prediction vs Inference

- Inference: you want to evaluate the effect of covariables on the response variable:

Prediction vs Inference

- Inference: you want to evaluate the effect of covariables on the response variable:
  - Effect sizes (OR, RR, etc) are relevant;
  - Confidence intervals are essential and p-values are useful.

Prediction vs Inference

- Inference: you want to evaluate the effect of covariables on the response variable:
  - Effect sizes (OR, RR, etc) are relevant;
  - Confidence intervals are essential and p-values are useful.
- Prediction: you want to predict the response variable of new patients based on the their covariables:

Prediction vs Inference

- Inference: you want to evaluate the effect of covariables on the response variable:
  - Effect sizes (OR, RR, etc) are relevant;
  - Confidence intervals are essential and p-values are useful.
- Prediction: you want to predict the response variable of new patients based on the their covariables:
  - Discrimination and calibration are important;
  - p-values could be a possible guide to select predictors. They are not essential.

Prediction vs Inference

- Inference: you want to evaluate the effect of covariables on the response variable:
  - Effect sizes (OR, RR, etc) are relevant;
  - Confidence intervals are essential and p-values are useful.
- Prediction: you want to predict the response variable of new patients based on the their covariables:
  - Discrimination and calibration are important;
  - p-values could be a possible guide to select predictors. They are not essential.
- van Diepen, M., et al. (2017). Prediction versus aetiology: common pitfalls and how to avoid them. Nephrology Dialysis Transplantation, 32(suppl2), ii1-ii5

Discrimination

- It is also known as predictive performance;
- It measures the ability to separate different responses;
- Statistical tools: Area Under the Curve (AUC), Net Reclassification Index (NRI).

Calibration

- It is also known as goodness-of-fit;
- It measures the ability to make unbiased estimates for the probability of the event of interest;
- Statistical tools: Calibration plot, calibration-in-large and calibration slope.

References

- Nature - Points of Significance: Logistic regression
- Diniz MA, Magalhães TM (2020) Logistic Regression and Related Methods. In: Piantadosi S., Meinert C. (eds) Principles and Practice of Clinical Trials. Springer, Cham.
- Anderson RP, Jin R, Grunkemeier GL. Understanding logistic regression analysis in clinical reports: an introduction. The Annals of thoracic surgery. 2003 Mar 1;75(3):753-7.
- Worster A, Fan J, Ismaila A. Understanding linear and logistic regression analyses. CJEM. 2007 Mar 1;9(02):111-3.