Regression analysis

Tiago M. Magalhães

Department of Statistics - ICE-UFJF

Juiz de Fora, August 21, 2026

Outline

1. Motivation

2. Simple linear regression model

3. Application

4. References

Outline

1. Motivation
2. Simple linear regression model
3. Application
4. References

Motivation

The first concepts of regression were proposed by Galton (1889), applied to Anthropometry, where the objective was to establish the dependence relationships between the height of the parents (**explanatory variable**, predictor variable, or covariate)

Motivation

The first concepts of regression were proposed by Galton (1889), applied to Anthropometry, where the objective was to establish the dependence relationships between the height of the parents (**explanatory variable**, predictor variable, or covariate) and the height of the children (variable of interest, **response variable**, or outcome).

Motivation

The first concepts of regression were proposed by Galton (1889), applied to Anthropometry, where the objective was to establish the dependence relationships between the height of the parents (**explanatory variable**, predictor variable, or covariate) and the height of the children (variable of interest, **response variable**, or outcome).

Motivation

However, these concepts could be applied in any context, such as:

- the need for a company to analyze the factors that may affect sales;

Motivation

However, these concepts could be applied in any context, such as:

- the need for a company to analyze the factors that may affect sales;
- predicting a person's mortality upon contracting a disease, based on their characteristics, such as age and pre-existing diseases.

Motivation

However, these concepts could be applied in any context, such as:

- the need for a company to analyze the factors that may affect sales;
- predicting a person's mortality upon contracting a disease, based on their characteristics, such as age and pre-existing diseases.

Motivation

- classifying a customer as a good or bad payer, based on the maximum number of days overdue in the last 6 months, whether or not they made a cash withdrawal using the card, and the length of time the customer has had the card;
- distance from a galaxy to Earth, based on the grayscale values of each pixel in a photograph of the galaxy being observed.

Motivation

- classifying a customer as a good or bad payer, based on the maximum number of days overdue in the last 6 months, whether or not they made a cash withdrawal using the card, and the length of time the customer has had the card;
- distance from a galaxy to Earth, based on the grayscale values of each pixel in a photograph of the galaxy being observed.

Motivation

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. In fact, regression analysis may be the most widely used statistical technique (Montgomery et al., 2021).

Motivation

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. In fact, regression analysis may be the most widely used statistical technique (Montgomery et al., 2021).

Motivation

Regression models gained even greater prominence with the popularization of Machine Learning, specifically, Supervised Learning, in which the focus is predictive, that is, given input values (covariates), the goal is to predict the expected outcome (response variable).

Motivation

Regression models gained even greater prominence with the popularization of Machine Learning, specifically, Supervised Learning, in which the focus is predictive, that is, given input values (covariates), the goal is to predict the expected outcome (response variable).

Outline

1 Motivation

2 Simple linear regression model

3 Application

4 References

Simple linear regression model

As we have seen, there are phenomena (response variable, Y) that can be described by a predictor variable (x), this is done as follows:

$$Y \sim \mathcal{D}(\mu, \sigma^2),$$

Simple linear regression model

As we have seen, there are phenomena (response variable, Y) that can be described by a predictor variable (x), this is done as follows:

$$Y \sim \mathcal{D}(\mu, \sigma^2),$$

where $\mu = \beta_1 + \beta_2 x$ and σ^2 are, respectively, the mean and variance of the distribution of Y, x is a known value, and β_1, β_2, and σ^2 are unknown parameters.

Simple linear regression model

As we have seen, there are phenomena (response variable, Y) that can be described by a predictor variable (x), this is done as follows:

$$Y \sim \mathcal{D}(\mu, \sigma^2),$$

where $\mu = \beta_1 + \beta_2 x$ and σ^2 are, respectively, the mean and variance of the distribution of Y, x is a known value, and β_1, β_2, and σ^2 are unknown parameters. In addition, β_1 and β_2 are called regression coefficients.

Simple linear regression model

As we have seen, there are phenomena (response variable, Y) that can be described by a predictor variable (x), this is done as follows:

$$Y \sim \mathcal{D}(\mu, \sigma^2),$$

where $\mu = \beta_1 + \beta_2 x$ and σ^2 are, respectively, the mean and variance of the distribution of Y, x is a known value, and β_1, β_2, and σ^2 are unknown parameters. In addition, β_1 and β_2 are called regression coefficients.

Simple linear regression model

It is convenient to write the relationship between the response and predictor variables as follows:

$$Y = \beta_1 + \beta_2 x + \varepsilon,$$

where $\varepsilon \sim \mathcal{D}(0, \sigma^2)$,

Simple linear regression model

It is convenient to write the relationship between the response and predictor variables as follows:

$$Y = \beta_1 + \beta_2 x + \varepsilon,$$

where $\varepsilon \sim \mathcal{D}(0, \sigma^2)$, with ε being referred to as the error.

Simple linear regression model

It is convenient to write the relationship between the response and predictor variables as follows:

$$Y = \beta_1 + \beta_2 x + \varepsilon,$$

where $\varepsilon \sim \mathcal{D}(0, \sigma^2)$, with ε being referred to as the error. Note: this formulation is always possible for location-scale families.

Simple linear regression model

It is convenient to write the relationship between the response and predictor variables as follows:

$$Y = \beta_1 + \beta_2 x + \varepsilon,$$

where $\varepsilon \sim \mathcal{D}(0, \sigma^2)$, with ε being referred to as the error. Note: this formulation is always possible for location-scale families.

Interpretation of the model parameters

Parameter β_1

It is the intercept of the regression line.

Interpretation of the model parameters

Parameter β_1

It is the intercept of the regression line. It is the expected value of the response variable when the predictor variable is zero.

Interpretation of the model parameters

Parameter β_1

It is the intercept of the regression line. It is the expected value of the response variable when the predictor variable is zero.

Interpretation of the model parameters

Parameter β_2

It is the slope of the regression line, the angular coefficient.

Interpretation of the model parameters

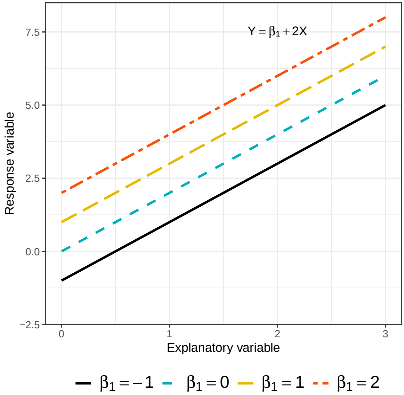
Parameter β_2

It is the slope of the regression line, the angular coefficient. It indicates the expected increase in the response variable when the predictor variable increases by one unit.

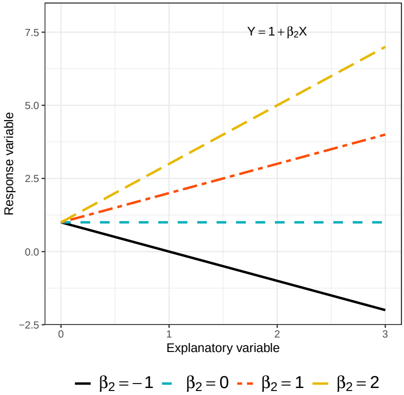
Interpretation of the model parameters

Parameter β_2

It is the slope of the regression line, the angular coefficient. It indicates the expected increase in the response variable when the predictor variable increases by one unit.



(a) Free β_1 and $\beta_2 = 2$.



(b) $\beta_1 = 1$ and free β_2.

Figure 1: Regression lines as a function of β_1 and β_2.

Linear regression model

In order to estimate β_1, β_2, and σ^2, and to make the relationship between the variables explicit, it is necessary to draw an independent sample of size n from the pair (Y, x), i.e., $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$,

Linear regression model

In order to estimate β_1, β_2, and σ^2, and to make the relationship between the variables explicit, it is necessary to draw an independent sample of size n from the pair (Y, x), i.e., $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$, with

$$Y_\ell \sim \mathcal{D}(\mu_\ell, \sigma^2),$$

Linear regression model

In order to estimate β_1, β_2, and σ^2, and to make the relationship between the variables explicit, it is necessary to draw an independent sample of size n from the pair (Y, x), i.e., $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$, with

$$Y_\ell \sim \mathcal{D}(\mu_\ell, \sigma^2),$$

where $\mu_\ell = \beta_1 + \beta_2 x_\ell$, $\ell = 1, 2, \ldots, n$.

Linear regression model

In order to estimate β_1, β_2, and σ^2, and to make the relationship between the variables explicit, it is necessary to draw an independent sample of size n from the pair (Y, x), i.e., $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$, with

$$Y_\ell \sim \mathcal{D}(\mu_\ell, \sigma^2),$$

where $\mu_\ell = \beta_1 + \beta_2 x_\ell$, $\ell = 1, 2, \ldots, n$.

Linear regression model

Alternatively,

$$Y_\ell = \beta_1 + \beta_2 x_\ell + \varepsilon_\ell,$$

with $\varepsilon_\ell \sim \mathcal{D}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$. In other words, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent random variables with mean zero and common variance σ^2.

Linear regression model

Alternatively,

$$Y_\ell = \beta_1 + \beta_2 x_\ell + \varepsilon_\ell,$$

with $\varepsilon_\ell \sim \mathcal{D}(0, \sigma^2)$, $\ell = 1, 2, \ldots, n$. In other words, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent random variables with mean zero and common variance σ^2.

Assumptions

In summary, the **assumptions** of a simple linear regression model are:

- The relationship between the response and predictor variables is linear;

Assumptions

In summary, the **assumptions** of a simple linear regression model are:

- The relationship between the response and predictor variables is linear;
- The errors

Assumptions

In summary, the **assumptions** of a simple linear regression model are:

- The relationship between the response and predictor variables is linear;
- The errors
  - have mean zero;

Assumptions

In summary, the **assumptions** of a simple linear regression model are:

- The relationship between the response and predictor variables is linear;
- The errors
  - have mean zero;
  - have constant variance;

Assumptions

In summary, the **assumptions** of a simple linear regression model are:

- The relationship between the response and predictor variables is linear;
- The errors
  - have mean zero;
  - have constant variance;
  - are uncorrelated.

Assumptions

In summary, the **assumptions** of a simple linear regression model are:

- The relationship between the response and predictor variables is linear;
- The errors
  - have mean zero;
  - have constant variance;
  - are uncorrelated.

Estimation

By the Least Squares Method, the estimators of β_1 and β_2 are given, respectively, by

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X} \text{ e } \hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \quad (1)$$

Estimation

By the Least Squares Method, the estimators of β_1 and β_2 are given, respectively, by

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \text{ e } \hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \quad (1)$$

The least squares estimators are unbiased.

Estimation

By the Least Squares Method, the estimators of β_1 and β_2 are given, respectively, by

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X} \text{ e } \hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \tag{1}$$

The least squares estimators are unbiased.

Variance estimation

The estimator of the variance, σ^2, is given by

$$\hat{\sigma}^2 = \frac{\sum_{\ell=1}^{n}\left(Y_\ell - \hat{\beta}_1 - \hat{\beta}_2 X_\ell\right)^2}{n-2}. \tag{2}$$

Variance estimation

The estimator of the variance, σ^2, is given by

$$\hat{\sigma}^2 = \frac{\sum_{\ell=1}^{n} \left(Y_\ell - \hat{\beta}_1 - \hat{\beta}_2 X_\ell \right)^2}{n-2}. \tag{2}$$

The estimator $\hat{\sigma}^2$, given in (2), is also called the **residual mean square** (MSRes) and is an unbiased estimator of σ^2.

Variance estimation

The estimator of the variance, σ^2, is given by

$$\hat{\sigma}^2 = \frac{\sum_{\ell=1}^{n} \left(Y_\ell - \hat{\beta}_1 - \hat{\beta}_2 X_\ell\right)^2}{n-2}. \tag{2}$$

The estimator $\hat{\sigma}^2$, given in (2), is also called the **residual mean square** (MSRes) and is an unbiased estimator of σ^2.

Outline

1 Motivation

2 Simple linear regression model

3 Application

4 References

Application

Since 2003, each soccer team in the Brazilian men's football championship, the Brasileirão, plays the others twice (a double round-robin system) and, at the end, the club with the most points is crowned champion (Magalhães et al., 2025).

The Brasileirão was nominated as the World's Strongest National League in 2021 and 2022 by the International Federation of Football History & Statistics (IFFHS, 2022, 2023).

Application

Here, our interest may be in assessing whether there is a linear relationship between points percentage and goal difference among the champion teams.

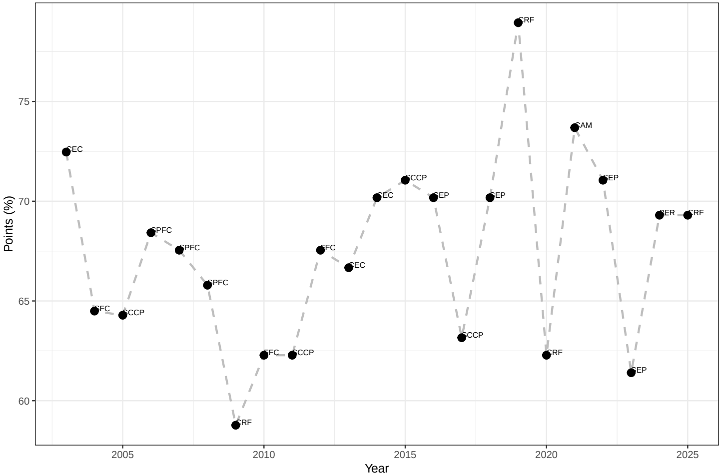


Figure 2: Points (%) of the champion teams over the years.

Application

Table 1: States of the champion teams.

Minas Gerais	Rio de Janeiro	São Paulo
4	7	12
(17.4%)	(30.4%)	(52.2%)
2	3	4

Application

Table 2: Points (%) of the champion teams.

Min	Q1	Md	Mean	Q3	Max	SD	CV
58.8	63.7	67.5	67.4	70.2	78.9	4.7	7.0

Application

Table 3: Points (%) of the champion teams, by state.

	Min	Q1	Md	Mean	Q3	Max	SD	CV
Minas Gerais	66.7	69.3	71.3	70.7	72.8	73.7	3.1	4.4
Rio de Janeiro	58.8	62.3	67.5	66.9	69.3	78.9	6.7	9.9
São Paulo	61.4	64.0	66.7	66.7	70.2	71.1	3.5	5.3

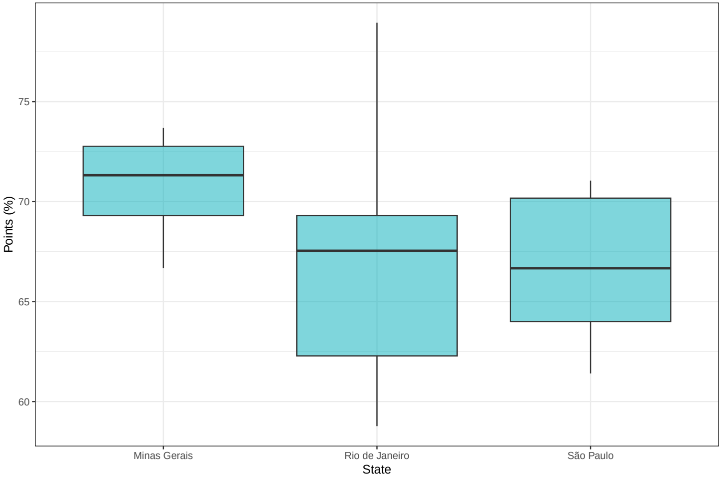


Figure 3: Boxplot of points (%) of the champion teams by state.

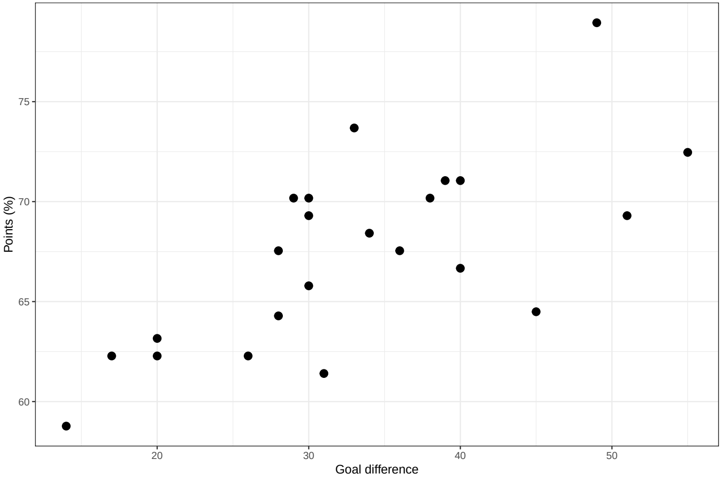


Figure 4: Scatterplot of points (%) and goal difference of the champion teams.

After fitting the model (parameter estimation), using (1) and (2), we have the following estimated model,

$$\hat{Y}_\ell = 57.2 + 0.3x_{\ell 2},$$

Application

After fitting the model (parameter estimation), using (1) and (2), we have the following estimated model,

$$\hat{Y}_\ell = 57.2 + 0.3x_{\ell 2},$$

where $\hat{Y}_\ell$: expected points (%) of the ℓth champion team, $x_{\ell 2}$: goal difference of the ℓth champion team, $\ell = 1, 2, \ldots, 23$ and MSRes $= 11.9$.

Application

After fitting the model (parameter estimation), using (1) and (2), we have the following estimated model,

$$\hat{Y}_\ell = 57.2 + 0.3x_{\ell 2},$$

where $\hat{Y}_\ell$: expected points (%) of the ℓth champion team, $x_{\ell 2}$: goal difference of the ℓth champion team, $\ell = 1, 2, \ldots, 23$ and $\text{MSRes} = 11.9$.

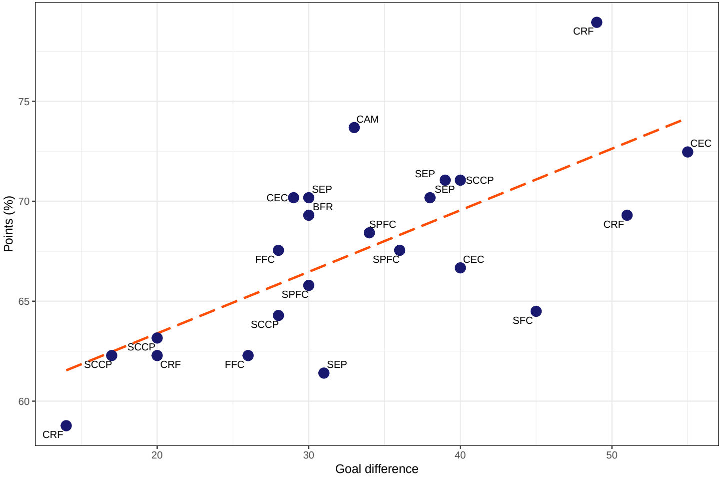


Figure 5: Scatterplot of points (%) and goal difference of the champion teams, with the estimated regression line.

Application

The regression coefficients can be interpreted as follows:

- $\hat{\beta}_1 = 57.2$:

Application

The regression coefficients can be interpreted as follows:

- $\hat{\beta}_1 = 57.2$: if the goal difference were zero, the champion team's expected points percentage would be 57.2%;

Application

The regression coefficients can be interpreted as follows:

- $\hat{\beta}_1 = 57.2$: if the goal difference were zero, the champion team's expected points percentage would be 57.2%;
- $\hat{\beta}_2 = 0.3$:

Application

The regression coefficients can be interpreted as follows:

- $\hat{\beta}_1 = 57.2$: if the goal difference were zero, the champion team's expected points percentage would be 57.2%;
- $\hat{\beta}_2 = 0.3$: for each additional goal in goal difference, the expected percentage of points earned increases by 0.3 percentage points.

Application

The regression coefficients can be interpreted as follows:

- $\hat{\beta}_1 = 57.2$: if the goal difference were zero, the champion team's expected points percentage would be 57.2%;
- $\hat{\beta}_2 = 0.3$: for each additional goal in goal difference, the expected percentage of points earned increases by 0.3 percentage points.

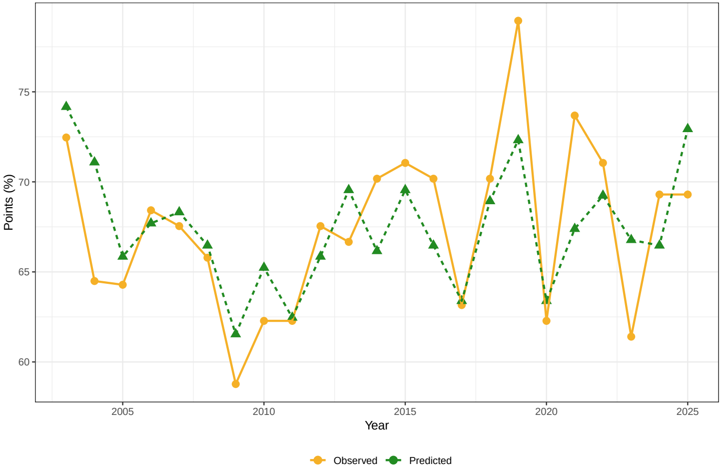


Figure 6: Observed and predicted points (%) from the model.

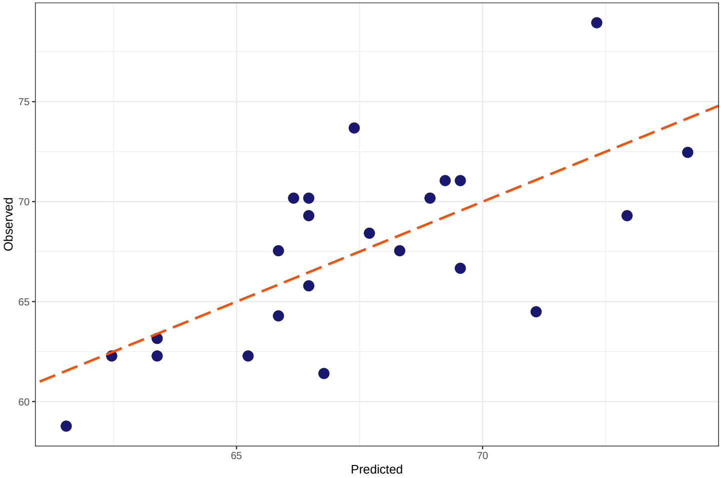


Figure 7: Observed and predicted points (%) from the model.

Outline

1 Motivation

2 Simple linear regression model

3 Application

4 References

References

Galton, F. (1889). *Natural Inheritance*. Macmillan, London.

IFFHS (2022). The strongest national league in the world 2021 by IFFHS.
`https://iffhs.com/en/news/`
`the-strongest-national-league-in-the-world-2021-by-iffhs-16`
Accessed March 20, 2026.

IFFHS (2023). IFFHS men's strongest national league in the world 2022.
`https://iffhs.com/en/news/`
`iffhs-mens-strongest-national-league-in-the-world-2022-2483`
Accessed March 20, 2026.

References

Magalhães, T. M., Gallardo, D. I., and Diniz, M. A. (2025). Inferential and predictive procedures for inverse gamma regression model. *Journal of Statistical Computation and Simulation*, 95(11):2327–2342.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to linear regression analysis*. Wiley, New York, 6th edition.

Thank you!

✉ tiago.magalhaes@ufjf.br

ufjf.br/tiago_magalhaes

Dept. of Statistics, Room 319