

Regression Analysis in Linguistics

语言学中的回归分析

Concepts, Methods, and Applications

概念、方法与应用

guoshujian@tongji.edu.cn

2025年11月26日

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

Outline / 提纲

- 1 Concepts (基本概念)
- 2 Methods (分析方法)
- 3 Example (实例分析)
- 4 Exercises (习题)

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

What is Regression Analysis? / 什么是回归分析?

English Definition

A statistical method to model the relationship between a **dependent variable** (response) and one or more **independent variables** (predictors).

- Goal: Prediction and Explanation.

中文定义

一种用于建模**因变量**（响应变量）与一个或多个**自变量**（预测变量）之间关系的统计方法。

- 目标：预测数值变化与解释变量关系。

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

Variables in Linguistics / 语言学中的变量

In linguistics, we often deal with specific types of data:

- **Continuous Variables** 连续变量:
 - Reaction Time (RT) / 反应时 (ms)
 - Formant Frequency (F1, F2) / 共振峰频率 (Hz)
 - Voice Onset Time (VOT) / 嗓音起始时间
- **Categorical Variables** 分类变量:
 - Native Language (L1 vs. L2) / 母语背景
 - Grammaticality (Grammatical vs. Ungrammatical) / 语法性
 - Part of Speech (Noun, Verb) / 词性

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

Common Types of Regression Models / 常见的回归模型类型

Different research questions require different regression models:

不同的研究问题需要不同的回归模型：

- **Simple Linear Regression (简单线性回归):**
One predictor (X) $\rightarrow$ Continuous outcome (Y).
- **Multiple Linear Regression (多元线性回归):**
Multiple predictors ($X_1, X_2 \dots$) $\rightarrow$ Continuous outcome (Y).
- **Logistic Regression (逻辑回归):**
Used for binary outcomes (e.g., Correct/Incorrect). Models the log-odds of an event.
- **Linear Probability Model (线性概率模型):**
Using OLS linear regression on binary data. Simple to interpret, but predictions can fall outside $[0, 1]$.

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

Simple Linear Regression / 简单线性回归

The fundamental equation / 基本公式:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

y_i Response variable (e.g., Vowel Duration) / 因变量

x_i Predictor variable (e.g., Speech Rate) / 自变量

β_0 **Intercept:** Value of y when $x = 0$ / 截距

β_1 **Slope:** Change in y for 1 unit change in x / 斜率

ϵ_i Error term (Residuals) / 误差项

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

Diagnostic Checks: Assumptions (L.I.N.E.) / 回归诊断：假设检验

After fitting a simple linear model, we must check four assumptions (L.I.N.E.):
拟合简单线性模型后，必须检验四个核心假设 (L.I.N.E.):

- **Linearity (线性):** The relationship between X and Y is linear.
- **Independence (独立性):** Observations are independent (often violated in linguistics!).
- **Normality (正态性):** Residuals (ϵ) are normally distributed.
- **Equal Variance (同方差性):** Variance of residuals is constant (Homoscedasticity).

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

Check 1: Normality of Residuals / 检验1: 残差正态性

Method / 方法:

- Visual: **Q-Q Plot**. Points should fall on the diagonal line.
- Test: **Shapiro-Wilk Test**.

Reference Value / 参考值:

- **p-value** > 0.05 .
- If $p < 0.05$, residuals are NOT normal.

Significance / 意义

If residuals are not normal, standard errors and p-values for your coefficients (β) may be biased.

如果残差非正态，系数的对数误差和P值可能有偏，导致结论不可靠。

Check 2: Homoscedasticity & Outliers / 检验2: 同方差性与异常值

Homoscedasticity (同方差性):

- **Test:** Breusch-Pagan Test. Reference: **p-value** > 0.05.
- **Visual:** Residuals vs. Fitted Plot. Should be a random cloud, not a "funnel" shape.

Influential Points (强影响点):

- **Metric:** Cook's Distance (Cook's D).
- **Threshold:** Generally worry if $D > \frac{4}{n}$ (where n is sample size).
- **Significance:** A single outlier can drastically change your slope (β_1).
意义: 单个异常值可能会剧烈改变斜率, 导致错误的语言学结论。

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

The Standard in Linguistics: LMER / 语言学标准：混合效应模型

Language data is rarely independent. We repeat items across different subjects.
语言数据很少是独立的，因为我们在不同受试者之间重复使用相同的语言材料。

Linear Mixed-Effects Regression (LMER)

We must account for **Random Effects** (随机效应):

- **By-Subject:** Some people speak faster than others.
- **By-Item:** Some words are recognized faster than others.

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + (u_{0j} + \epsilon_{ij})$$

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

Example Case / 实例分析

Research Question: Does Word Frequency affect Reaction Time?

研究问题: 词频是否影响反应时?

Code Implementation (R w/ lme4):

```
library(lme4)
library(lmerTest)

# Model: RT predicted by Frequency
# Random intercepts for Subject and Item
model <- lmer(RT ~ Frequency + (1|Subject) + (1|Item),
              data = lexical_decision_data)

summary(model)
```

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

Interpreting Results / 结果解读

Suppose the output for 'Frequency' is:
假设 'Frequency' 的输出结果如下:

Term	Estimate (β)	Std. Error	t-value	p-value
(Intercept)	600.5	10.2	58.8	< .001
Frequency	-15.3	2.1	-7.2	< .001

- **Coefficient (-15.3):** For every 1 unit increase in Frequency, Reaction Time decreases by 15.3 ms.
- **中文解读:** 词频每增加一个单位, 反应时减少 15.3 毫秒 (负相关, 说明高频词认知更快)。

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

Exercise 1 / 习题 1

Concept Check / 概念检查

In a study of second language acquisition, we measure test scores (0-100). We use "Age of Acquisition" (AoA) as a predictor.

在二语习得研究中，我们测量测试分数（0-100）。我们使用“习得年龄”（AoA）作为预测变量。

If $\beta_{AoA} = -2.5$, what does this mean?

如果 $\beta_{AoA} = -2.5$ ，这意味着什么？

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

Exercise 1 / 习题 1

Concept Check / 概念检查

In a study of second language acquisition, we measure test scores (0-100). We use "Age of Acquisition" (AoA) as a predictor.

在二语习得研究中，我们测量测试分数（0-100）。我们使用“习得年龄”（AoA）作为预测变量。

If $\beta_{AoA} = -2.5$, what does this mean?

如果 $\beta_{AoA} = -2.5$ ，这意味着什么？

Answer / 答案

It suggests a negative correlation. As the Age of Acquisition increases by 1 year, the predicted test score decreases by 2.5 points.

这表示负相关。习得年龄每增加1岁，预测的测试分数下降2.5分。

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

Methodology / 方法论

Why is a simple linear regression often insufficient for psycholinguistic experiments?
为什么简单线性回归通常不足以处理心理语言学实验数据？

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

Exercise 2 / 习题 2

Methodology / 方法论

Why is a simple linear regression often insufficient for psycholinguistic experiments?
为什么简单线性回归通常不足以处理心理语言学实验数据？

Answer / 答案

It violates the assumption of independence. Data points from the same subject or using the same words are correlated. We need Mixed Effects Models to handle this variation.
因为它违反了独立性假设。来自同一受试者或使用同一单词的数据点是相关的。我们需要混合效应模型来处理这种变异。

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn

Thank You!

谢谢!

Q & A

(C) <https://languages.lol/> | guoshujian@tongji.edu.cn