[AM-03-063] Regression Fundamentals

Abstract

Regression analysis is a common statistical tool used to model relationships between variables and to explore the influencing factors underlying observed spatial data patterns. This entry focuses on the most basic form of regression model: linear regression. The notations, inference, assumptions, and diagnostics of linear regression are introduced, and interpretations of linear regression results are demonstrated using an empirical example in R software. The entry concludes with a brief discussion of the challenges of applying standard linear regression to spatial data.

Keywords: geospatial analysis, spatial statistics, statistical learning

Author & citation

Li, Z. (2024). Regression Fundamentals. The Geographic Information Science & Technology Body of Knowledge (2024 Edition). John P. Wilson (ed.). DOI: 10.22224/gistbok/2024.1.11.

Explanation

1. Introduction
2. Regression Formulation
3. Linear Regression
4. Inference in Linear Regression
5. Linear Regression Software and an Empirical Example
6. Regression Assumptions and Checks
7. When Linear Regression Meets Spatial Data

1. Introduction

Regression analysis is a fundamental statistical tool used by geographers and spatial scientists to study and model relationships in spatial data. Regression allows for the exploration of how one or more factors are correlated with an outcome of interest. The outcome of interest is commonly referred to as the dependent variable in a regression model, while the potential influencing factors are referred to as independent variables. Regression models are estimated from data in a tabular structure, providing insights into the processes and relationships behind spatial data. These models help researchers identify significant factors, quantify the strength and direction of relationships, and predict outcomes given new data. For example, geographers used regression analysis to study social determinants of spatial health disparities (e.g., Anderson et al., 2023), examine physical and human factors associated with wildfire occurrence (e.g., Oliveira et al., 2012), explore socio-demographic factors in determining voting behavior (e.g., Fotheringham et al., 2021), and predict house prices using property and neighborhood-level attributes (e.g, Bourassa et al., 2007), among other applications. Regression modeling is arguably the most widely applied research method in quantitative geography and other disciplines.

2. Regression Formulation

A general regression model takes the form of:

$$y = f(X) + \epsilon$$

where y is the dependent variable and usually the outcome of interest to be explained or predicted. X is one or a set of independent variables that are presumed to correlate with the dependent variable. The function $f(X)$ represents the relationship between y and X and can take many forms, such as linear, polynomial, logarithmic, or more complex structures. The error term ϵ represents random variation or noise that cannot be captured by the model. Regression modeling involves estimating the function f using data at hand to provide insights into the relationships between the independent variables and the dependent variable for explanatory or prediction tasks.

3. Linear Regression

The most basic form of a regression model is simple linear regression. Simple linear regression has one dependent variable and one independent variable. The function f is represented by a straight line with an intercept and a slope:

$$y = \beta_0 + \beta_1 X + \epsilon$$

where β_0 is the intercept and β_1 is the slope of the regression line. As shown in Figure 1, the intercept is the value of y when X equals zero, and the slope measures the change in y when X changes by one unit. The intercept and the slope of the regression line are estimated using the sampled data and can be determined based on the least squares criterion, which minimizes the sum of the squared differences between the observed values and the values predicted by the line.

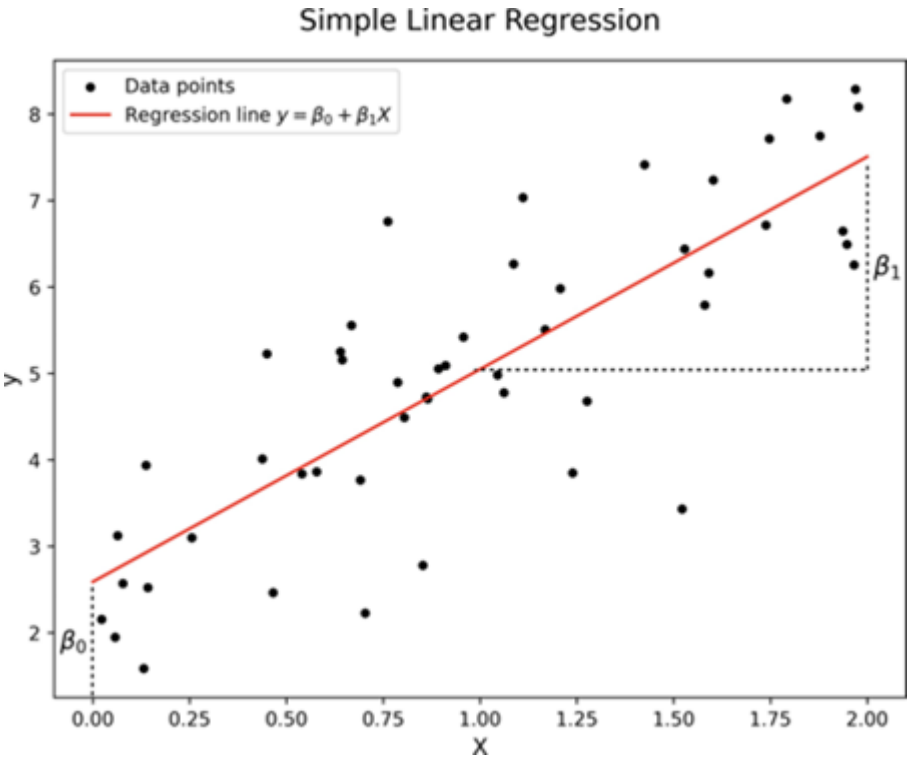


Figure 1. A regression line is determined by an intercept and slope. Source: author.

Simple linear regression can be extended to the multivariate case when there are multiple (k) independent variables. Multiple linear regression is formulated by

$$y = \beta_0 + \beta_1 X + \cdots + \beta_k X_k + \epsilon$$

where the intercept β_0 is the value of y when all independent variables equal zero. Each independent variable is associated with a slope coefficient that quantifies the change in y for a one-unit change in the corresponding independent variable, while holding all other variables constant. Following the same least squares criterion, all regression coefficients can be estimated by

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

where X is an Invalid Math Expression matrix of the independent variables, including a column of ones for the intercept term. X^T is the transpose of the X matrix, $(X^T X)^{-1}$ and is the inverse of the $X^T X$ matrix. The resulting $\hat{\beta}$ is a (k +1) x 1 vector of the estimated coefficients where the last element is the intercept estimate.

4. Inference in Linear Regression

The data we collect and observe is a sample of a population and is subject to sampling variation. As a result, the regression coefficients estimated based on the collected sample will also change when independent sampling from the population is repeated. Therefore, performing statistical inferences such as hypothesis testing is critical to understanding the

true underlying relationships in the population. The most common task is to determine the statistical significance of independent variables, suggesting their strength associated with the dependent variable. To test the significance of an individual coefficient β_j, the following hypothesis test is performed:

$H_0 : \beta_j = 0$ (null hypothesis)

$H_1 : \beta_j \neq 0$ (alternative hypothesis)

The null hypothesis H_0 states that the coefficient of the independent variable β_j is equal to zero, which indicates that there is no relationship between the independent variable and the dependent variable in the population. The test statistic t-value for β_j is given by:

$$t = \frac{\hat{\beta}_j}{SE\left(\hat{\beta}_j\right)}$$

where $\hat{\beta}_j$ is the estimated coefficient and SE ($\hat{\beta}_j$) is the standard error of $\hat{\beta}_j$. This test statistic follows a t-distribution with n−k−1 degrees of freedom, where n is the sample size and k is the number of independent variables. A p-value, which is the probability of obtaining the observed t-value when the null hypothesis is true, can be calculated from the area under the t-distribution curve that is more extreme than the observed t-value. If the p-value is smaller than a specified threshold (0.05 is a commonly referenced value), then we can reject the null hypothesis and conclude that the coefficient β_j is significantly different from zero and there is a statistically significant association between independent variable X_j and the dependent variable.

In addition to testing an individual coefficient's significance, another common task is to evaluate the goodness-of-fit of a linear regression model. The most commonly used goodness-of-fit metric is the Coefficient of Determination R^2 which measures the proportion of the variance in the dependent variable that is explained by the independent variables. It is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the overall mean of the dependent variable. An R^2 value of 1 indicates a model that perfectly explains the dependent variable. An R^2 value that is equal to 0 indicates a model that has no explanatory power. A higher R^2 value indicates the model has more explanatory power

when model assumptions are satisfied. R^2 will increase if adding more independent variables, in order to penalize model complexity and potential overfitting, an adjustment to R^2 is often used to account for the number of independent variables k and the number of observations n involved in the calculation:

$$R^2_{adj}$$

Another commonly used test statistic for overall model significance is the F-statistic test, given by:

$$F = \frac{(R^2/k)}{(1-R^2)\,/\,(n-k-1)}$$

This F-statistic follows an F-distribution with k and n − k − 1 degrees of freedom. The null hypothesis of the F-test is that all regression coefficients are zero, indicating that the model with the included independentvariables is not significantly different from a model with only an intercept. Accordingly, a p-value can be calculated based on the observed F statistic and F-distribution. A small p-value than a specified threshold (e.g., 0.05) will reject the null hypothesis and indicate the model provides a better fit to the data than a model that contains no independent variables.

5. Linear Regression Software and an Empirical Example

Linear regression as one of the most fundamental statistical tools is available in popular GIS and statistical software, including ArcGIS Pro, R, Python, MATLAB, Microsoft Excel, Stata, SAS, among others. Here, an example of fitting a linear regression model is demonstrated using the open-source R programming language. The R code and R output can be seen in Figure 2. A voting data set of the county-level 2020 Presidential Election from Fotheringham and Li (2023) is used as an example. The voting data is loaded as an R data frame. The dependent variable used in the model is the percentage of people who voted for the Democratic party (pct_dem), and three independent variables are selected to create a simple model: the percentage of people who have a Bachelor’s degree or higher (pct_bach), the ratio of males to females (sex_ratio), and the log-transformed population density (log_pop_den). A linear model then can be fitted using the lm() function.

```
> voting <- read.csv("voting_data.csv")
> model <- lm("pct_dem ~ 1 + pct_bach + sex_ratio + log_pop_den", data = voting)
> summary(model)

Call:
lm(formula = "pct_dem ~ 1 + pct_bach + sex_ratio + log_pop_den",
    data = voting)

Residuals:
    Min      1Q  Median      3Q     Max
-30.803  -8.952  -2.497   6.441  70.370

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.072487   2.428483   2.501   0.0125 *
pct_bach    0.637875   0.026685  23.904   <2e-16 ***
sex_ratio   0.009763   0.021572   0.453   0.6509
log_pop_den 3.372961   0.150770  22.372   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.75 on 3104 degrees of freedom
Multiple R-squared:  0.3871,    Adjusted R-squared:  0.3865
F-statistic: 653.5 on 3 and 3104 DF,  p-value: < 2.2e-16
```

Figure 2. R code example and output of the linear regression model. Source: author.

Once the model is fitted, the user can call a summary() function to the fitted model object to output a regression summary. The Residuals section provides descriptive statistics of the model residuals. The Coefficients section provides the estimated coefficients for the intercept and each independent variable, along with their standard errors, t-values, and p-values. The intercept of the model is 6.072487, suggesting that when all independent variables are zero, the baseline percentage of Democratic voters is approximately 6.07% for all counties. The coefficient for pct_bach is 0.637875, indicating that for each 1% increase in the percentage of people with a Bachelor’s degree, holding all other variables constant, the percentage of Democratic votes increases by about 0.64%. The p-value associated with pct_bach is less than 2e-16 (i.e., 2 x 10^{-16}), indicating that the coefficient is statistically different from zero. Similarly, the coefficient for log_pop_den is 3.372961, meaning a one-unit increase in the log-transformed population density is associated with a 3.37% increase in Democratic votes, holding all other variables constant, also with a highly significant p-value of less than 2e-16. In contrast, the sex_ratio coefficient is 0.009763, and with a p-value of 0.6509, it is not statistically significant, indicating that the county-level sex ratio does not significantly associate with the Democratic vote share.

In summary, the multiple R-square value is 0.3871, and the adjusted R-square value is 0.3865. Both R^2 values indicate that approximately 38.7% of the variance in pct_dem is explained by the model. The F-statistic of the model is 653.5 with a p-value less than

2.2e-16, suggesting that the included independent variables collectively have a significant relationship with the percentage of Democratic votes.

6. Regression Assumptions and Checks

In order to ensure that linear regression results (i.e., coefficient estimates and inference) are valid (meaning the OLS estimators are unbiased and have the lowest sampling variance according to the Gauss-Markov theorem when assumptions hold), it is important to verify that certain statistical assumptions are met. For linear regression, it is convenient to refer to the ”**LINE**” assumptions. These assumptions are:

- **L**inearity, where the relationship between the dependent variable and independent variables is linear;
- **I**ndependence of errors, meaning the model residuals are independent of each other;
- **N**ormality of errors, indicating that the residuals are normally distributed with a mean of zero; and
- **E**qual Variance (also known as homoscedasticity, whichis in oppose to heteroscedasticity) of errors, meaning that the variance of the residuals is the same for all values of X.

A couple of visual plots are helpful in diagnosing assumption violations. For example, a residuals vs. fitted values plot (Figure 3) is useful to check for non-linearity and heteroscedasticity. A well-behaved plot should have the residuals ”bounce randomly” and roughly form a ”horizontal band” around the 0 line, suggesting that the relationship is linear and the variance is equal.

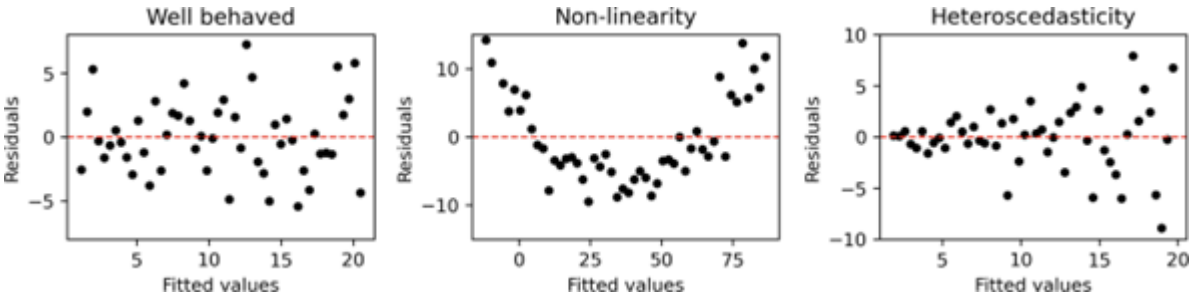


Figure 3. Examples of well-behaved and ill-conditioned residual vs. fit plots. Source: author.

Plotting histograms and Quantile-Quantile (Q- Q) plots of residuals can check for residual normality (Figure 4) . A Q-Q plot shows the quantiles of the residuals against the quantiles of the theoretical normal distribution. Often, a diagonal reference line is plotted on a Q-Q plot, and if the residuals come from a normal distribution, all the points should fall along the reference line. Dependence of errors often occurs in time-series and spatial data due to temporal and spatial dependency. A plot of the residuals following the temporal order or spatial order (i.e., a map) will help to provide insights into the degree of dependency.

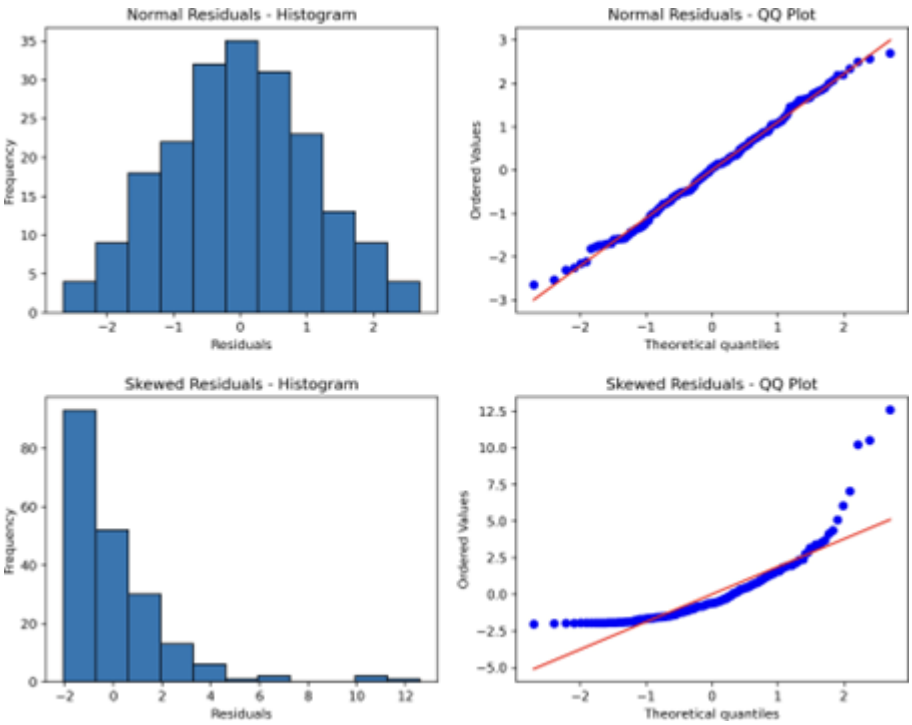


Figure 4. Histograms and Q-Q plots of normal and skewed residuals. Source: author.

When multiple independent variables are used in a regression model, it is important to check for correlations among these variables to avoid the issue of multicollinearity, which occurs when one variable can be linearly (or almost linearly) explained by others. This can happen, for example, when some independent variables sum to a constant, such as including all percentages of racial groups or land cover classes of a geographic unit in one model. The consequence of multicollinearity is that estimated regression coefficients will have large uncertainties and reduced precision. Common checks include calculating bi-variate correlation coefficients to identify and remove variables that are highly correlated (e.g., > 0.8) with others, and using the Variance Inflation Factor (VIF) to remove independent variables with high VIF values (e.g., > 10).

7. When Linear Regression Meets Spatial Data

The challenge of modeling spatial data using standard linear regression methods arises due to potential spatial effects that govern the data-generating processes, namely spatial autocorrelation and heterogeneity. Failing to account for these spatial effects will result in model residuals that are spatially correlated and heteroskedastic, which violates the Independence and Equal Variance assumptions mentioned in the above section. Consequently, the regression coefficients may be biased and have inflated variances (Anselin and Bera, 1998; Dormann et al., 2007). A common diagnostic is to calculate a spatial autocorrelation measure of the regression residuals, such as Moran's I (its calculation can be found in AM-03-022 Global Measures of Spatial Association). If Moran's I indicates substantial spatial autocorrelation in the residuals, spatial statistical models should be used instead, such as various forms of spatial econometric models (Anselin, 1988) and geographically weighted regression models (Fotheringham et al., 2023). Additionally, linear regression is the most basic form of a supervised machine learning model, and for more complicated non-linear processes, one should resort to more advanced statistical or machine learning methods. More details can be found in these

entries: AM-32-032- Spatial Autoregressive Models, AM-34-034 - The Geographically Weighted Regression Framework, and AM-08-094 - Machine Learning Approaches.

References

Anderson, T., Herrera, D., Mireku, F., Barner, K., Kokkinakis, A., Dao, H., ... Pierobon, M. (2023). Geographical variation in social determinants of female breast cancer mortality across US counties. JAMA Network Open, 6(9), e2333618-e2333618.

Anselin, L. (1988). Spatial econometrics: Methods and models. Dordrecht: Kluwer Academic Publishers.

Anselin, L., & Bera, A. K. (1998). Spatial dependence in linear regression models with an introduction to spatial econo met- rics. In A. Ullah (Ed.), Handbook of applied economic statistics (pp. 237–290). Boca Raton, FL: CRC Press.

Bourassa, S. C., Cantoni, E., Hoesli, M. (2007). Spatial dependence, housing submarkets, and house price prediction. The Journal of Real Estate Finance and Economics, 35, 143-160.

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G, Davies, R.G., Hirzel, A., Jets, W., Kissling, W.D., Kühn, I, Ohlemüller, R., Peres-Neto, P.R., Reineking, B., Schröder, B., Schurr, F.M. & Wilson, R. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography, 30(5), 609-628.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons.

Fotheringham, A. S., Li, Z., Wolf, L. J. (2021). Scale, context, and heterogeneity: A spatial analytical perspective on the 2016 US presidential election. Annals of the American Association of Geographers, 111(6), 1602-1621.

Fotheringham, A. S., Oshan, T. M., & Li, Z. (2023). Multiscale Geographically Weighted Regression: Theory and Practice. CRC Press.

Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A., Pereira, J. M. (2012). Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. Forest Ecology and Management, 275, 117-129.