Regression Diagnostics

Dragos Ailoae
dailoae@gradcenter.cuny.edu

Advanced Economics and Business Statistics
ECON-4400w - Spring 2022

Brooklyn College
Apr 25, 2022

Regression User's Guide (1 of 2)

What Can Go Wrong?	What Are the Consequences?	How Can It Be Detected?	How Can It Be Corrected?
Omitted Variable			
The omission of a relevant independent variable	Bias in the coefficient estimates (the $\hat{\beta}$s) of the included Xs.	Theory, significant unexpected signs, or surprisingly poor fits.	Include the omitted variable or a proxy.
Irrelevant Variable			
The inclusion of a variable that does not belong in the equation	Decreased precision in the form of higher standard errors, lower *t*-scores and wider confidence intervals.	1. Theory 2. *t*-test on $\hat{\beta}$ 3. $\bar{R}^2$ 4. Impact on other coefficients if X is dropped.	Delete the variable if its inclusion is not required by the underlying theory.
Incorrect Functional Form			
The functional form is inappropriate	Biased estimates, poor fit, and difficult interpretation.	Examine the theory carefully; think about the relationship between X and Y.	Transform the variable or the equation to a different functional form.

Functional form (SW 8.2)

The best way to choose a functional form for a regression model is to select the specification that best matches the underlying theory of the equation. In a majority of cases, the linear form will be adequate, and for most of the rest, common sense will point out a fairly easy choice from the following alternatives:

Functional Form	Equation (one X only)	The Change in Y when X Changes
Linear	$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$	If X increases by one unit, Y will change by β_1 units.
Double-log	$\ln Y_i = \beta_0 + \beta_1 \ln X_i + \epsilon_i$	If X increases by one percent, Y will change by β_1 percent. (Thus β_1 is the elasticity of Y with respect to X.)
Semilog (lnX)	$Y_i = \beta_0 + \beta_1 \ln X_i + \epsilon_i$	If X increases by one percent, Y will change by $\beta_1/100$ units.
Semilog (lnY)	$\ln Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$	If X increases by one unit, Y will change by roughly $100\beta_1$ percent.
Polynomial	$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$	If X increases by one unit, Y will change by $(\beta_1 + 2\beta_2 X)$ units.

Logarithms refresher

What the heck is a log? If e (a constant equal to 2.71828) to the "*b*th power" produces x, then b is the log of x:

$$\text{b is the log of x to the base e if:} \quad e^b = x$$

Thus, a **log** (or logarithm) is the exponent to which a given base must be taken in order to produce a specific number. While logs come in more than one variety, we'll use only **natural logs** (logs to the base e) in this text. The symbol for a natural log is "ln," so $\ln(x) = b$ means that $(2.71828)^b = x$ or, more simply,

$$\ln(x) = b \qquad \text{means that} \qquad e^b = x$$

For example, since $e^2 = (2.71828)^2 = 7.389$, we can state that:

$$\ln(7.389) = 2$$

Thus, the natural log of 7.389 is 2! Two is the power of e that produces 7.389. Let's look at some other natural log calculations:

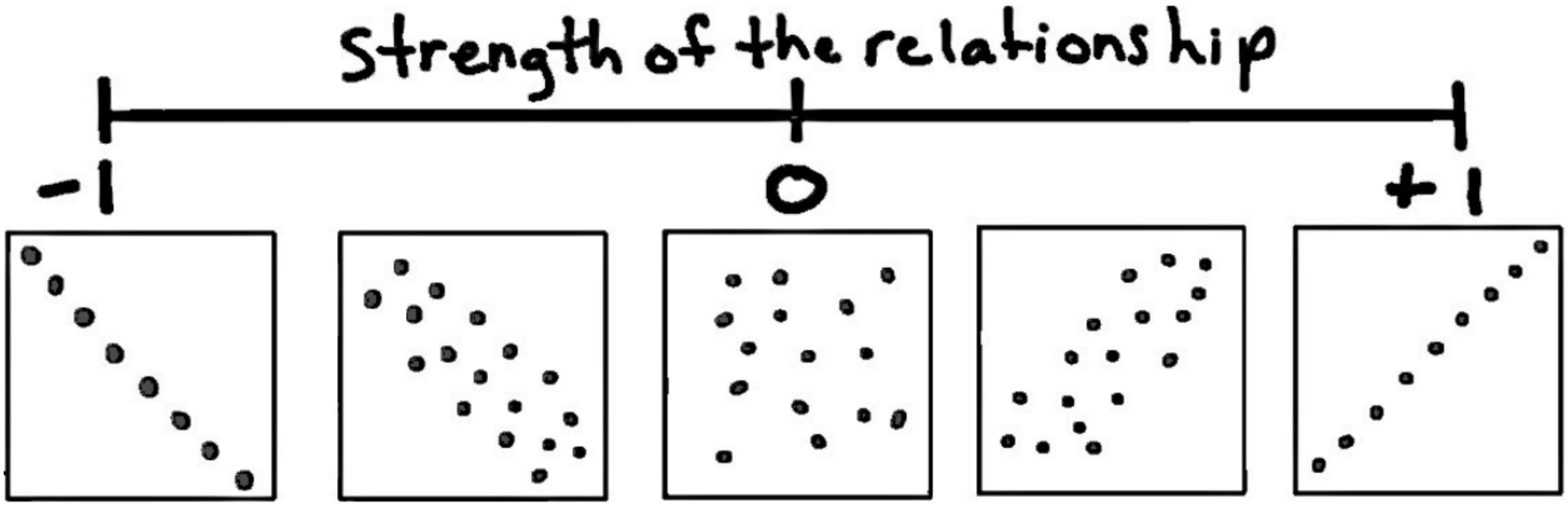
$$\begin{aligned} \ln(100) &= 4.605 \\ \ln(1000) &= 6.908 \end{aligned}$$

Regression User’s Guide (2 of 2)

What Can Go Wrong?	What Are the Consequences?	How Can It Be Detected?	How Can It Be Corrected?
Multicollinearity Some of the independent variables are (imperfectly) correlated	No biased $\hat{\beta}$s, but estimates of the separate effects of the Xs are not reliable, i.e., high SE($\hat{\beta}$)s and low *t*-scores.	Pairwise correlations or scatterplots	Drop redundant variables, but to drop others might introduce bias. Often doing nothing is best.
Serial Correlation Observations of the error term are correlated, as in: $\epsilon_t = \rho\epsilon_{t-1} + u_t$	No biased $\hat{\beta}$s, but OLS no longer is minimum variance, and hypothesis testing and confidence intervals are unreliable.	Use residual plots	If impure, fix the specification.
Heteroskedasticity The variance of the error term is not constant for all observations, as in: $VAR(\epsilon_i) = \sigma^2 Z_i$	Same as for serial correlation.	Use residual plots	If impure, fix the specification. Otherwise, use robust std. errors or reformulate the variables.

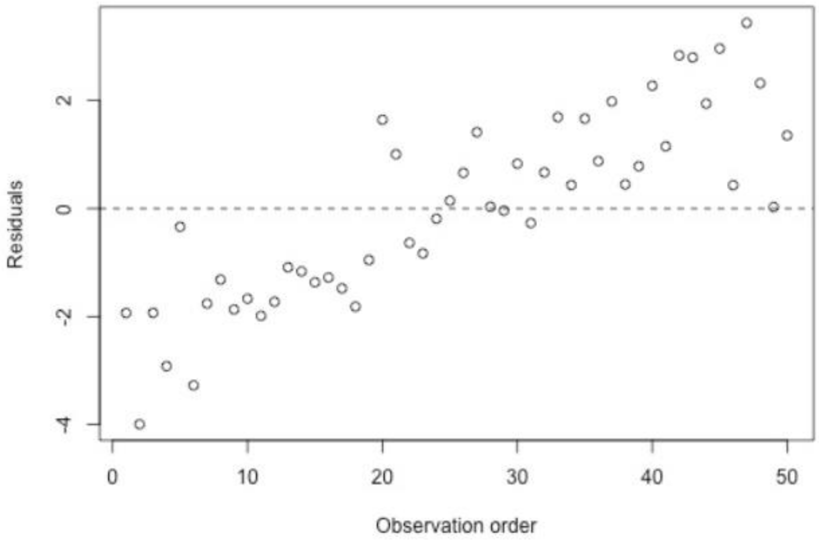
Multicollinearity

Check pairwise correlations and scatterplots of the suspected independent variables

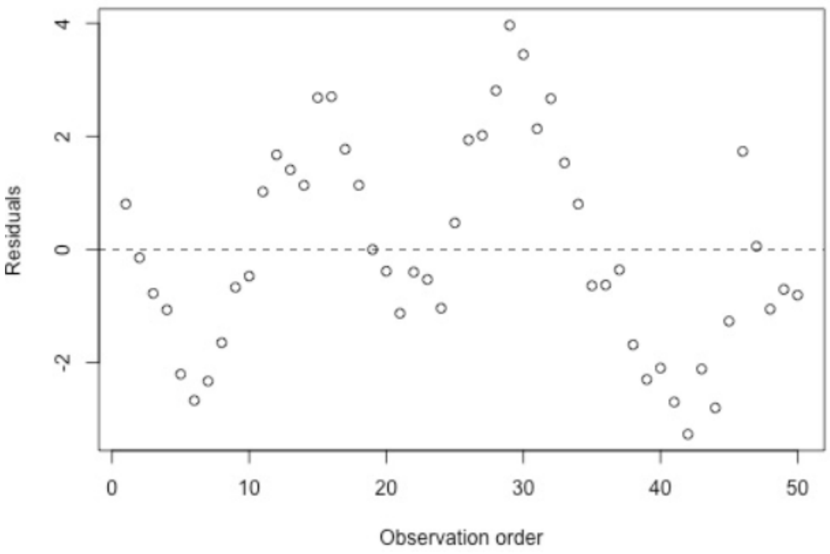


Serial correlation

A residuals vs. order plot that exhibits (positive) trend suggests that some of the variation in the response is due to time

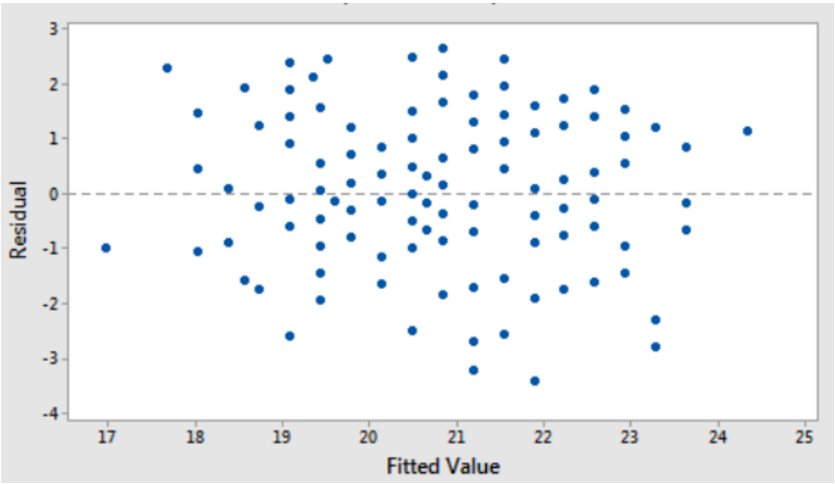


A residuals vs. order plot that suggests that there is "positive serial correlation" among the error terms. The plot suggests that the assumption of independent error terms is violated.

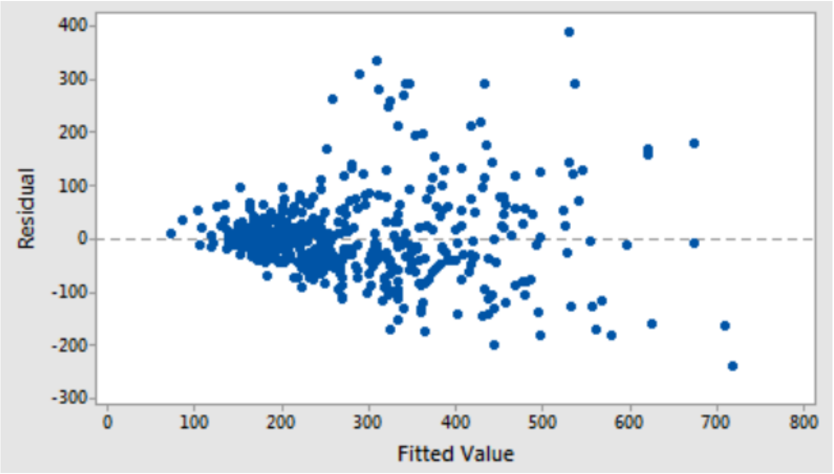


Heteroskedasticity

A Good Residual Plot



Indications that Assumption of Constant Variance is Not Valid



Presentation of regression results

TABLE 7.1 Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28 (0.52) [−3.30, −1.26]	−1.10 (0.43) [−1.95, −0.25]	−1.00 (0.27) [−1.53, −0.47]	−1.31 (0.34) [−1.97, −0.64]	−1.01 (0.27) [−1.54, −0.49]
Control variables					
Percentage English learners (X_2)		−0.650 (0.031)	−0.122 (0.033)	−0.488 (0.030)	−0.130 (0.036)
Percentage eligible for subsidized lunch (X_3)			−0.547 (0.024)		−0.529 (0.038)
Percentage qualifying for income assistance (X_4)				−0.790 (0.068)	0.048 (0.059)
Intercept	698.9 (10.4)	686.0 (8.7)	700.2 (5.6)	698.0 (6.9)	700.4 (5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
$\bar{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

These regressions were estimated using the data on K–8 school districts in California, described in Appendix 4.1. Heteroskedasticity-robust standard errors are given in parentheses under coefficients. For the variable of interest, the student–teacher ratio, the 95% confidence interval is given in brackets below the standard error.