Regression Analysis and Forecasting

1. Introduction

Regression analysis is one of the most widely used statistical tools for modeling relationships between variables and for making forecasts. At its core, regression examines how a **response variable** (often called the dependent variable, denoted y) changes as a function of one or more **predictor variables** (also called independent or regressor variables, denoted x).

The ultimate goal is often **prediction or forecasting**: once a relationship is established, it can be used to estimate or forecast new outcomes. For example:

- Predicting patient satisfaction (response) using patient age and severity of illness (predictors).
- Forecasting electricity demand (response) using temperature, time of day, and season (predictors).

There are two common data settings:

- **Cross-sectional data**: observations are collected at a single point in time (e.g., survey data).
- **Time-series data**: observations are collected sequentially over time (e.g., monthly sales figures).

Regression is fundamental in both settings, but additional considerations apply when the data are time-dependent (e.g., autocorrelation).

2. The Linear Regression Model

2.1 Simple Linear Regression

The simplest regression model involves one predictor variable:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- y: Response variable.
- x: Predictor variable.
- β_0: Intercept, the expected value of y when $x = 0$.

- β_1: Slope, the change in the mean of y for a one-unit increase in x.
- ε: Error term, representing random deviations.

Assumptions

1. **Linearity**: The mean of y is a linear function of x.
2. **Independence**: Errors are independent of each other.
3. **Homoscedasticity**: Constant variance of errors, $Var(\varepsilon) = \sigma^2$.
4. **Normality**: Errors are normally distributed $\varepsilon \sim N(0, \sigma^2)$.

2.2 Multiple Linear Regression

When multiple predictors are included, the model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Each coefficient β_j is a **partial regression coefficient**, showing the expected change in y for a one-unit increase in x_j, holding all other predictors constant.

2.3 Linear in Parameters

A model is "linear" if it is linear in its coefficients, even if the relationship between y and x is nonlinear. For example:

- Quadratic model: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$.
- Seasonal model: $y_t = \beta_0 + \beta_1 sin\left(2\pi \frac{t}{d}\right) + \beta_2 cos\left(2\pi \frac{t}{d}\right) + \varepsilon_t$.

Both are still linear regression models because coefficients enter linearly.

3. Least Squares Estimation

The most common method to estimate regression coefficients is **Ordinary Least Squares (OLS)**.

3.1 Idea

Choose $\beta_0, \beta_1, \ldots, \beta_k$ to minimize the **sum of squared errors (SSE):**

$$SSE = \sum_{\{i=1\}}^{n} (y_i - \hat{y}_i)^2 = \sum_{\{i=1\}}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})\right)^2$$

3.2 Normal Equations

Taking partial derivatives wrt to each β_i and setting them equal to zero leads to a system of equations (*normal equations*). In matrix form, the multiple linear regression model is:

$$y = X\beta + \varepsilon$$

- y: $n \times 1$ vector of responses.
- X: $n \times p$ matrix of predictors (with first column of 1's for intercept).
- β: $p \times 1$ vector of coefficients.
- ε: $n \times 1$ error vector.

The least squares function is

$$L = (y - X\beta)^T(y - X\beta)$$

$$L = y^T y - 2\beta^T X^T y + \beta^T X^T X\beta$$

To minimize the least squares, the least squares estimator must satisfy

$$\frac{\partial L}{\partial \beta} = -2X^T y + 2(X^T X)\hat{\beta} = 0$$

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

Thus the least squares estimator is given by:

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

3.3 Properties

- **Unbiased**: $E[\hat{\beta}] = \beta$.
- **Variance**: $Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$.
- **Gauss-Markov theorem**: OLS is the **Best Linear Unbiased Estimator (BLUE)**.

3.4 Variance of Model Errors

Estimated using:

$$\sigma^2 = \frac{SS_E}{n - p}$$

where SS_E is the sum of squares of the residuals, $p = k + 1$ is the number of parameters, n number of residuals.

3.5 Residuals

$e_i = y_i - \widehat{y_i}$. Residuals are key for diagnostics.

Example – Patient Satisfaction

A hospital is implementing a program to improve quality and productivity. As part of this program, the hospital is attempting to measure and evaluate patient satisfaction. Table 3.2 contains some of the data that has been collected for a random sample of 25 recently discharged patients. The “severity” variable is an index that measures the severity of the patient’s illness, measured on an increasing scale (i.e., more severe illnesses have higher values of the index), and

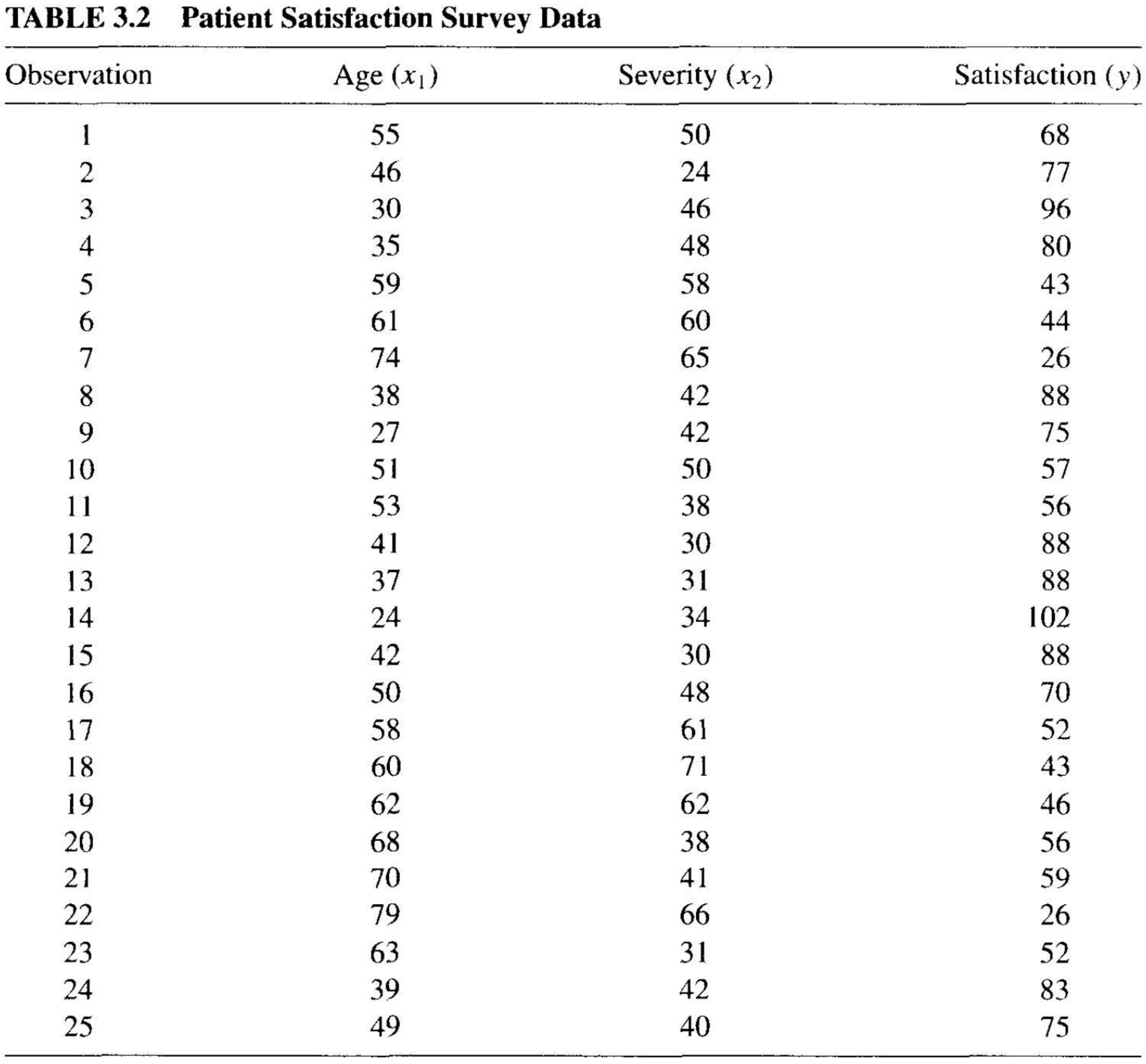


TABLE 3.2 Patient Satisfaction Survey Data

Observation	Age (x_1)	Severity (x_2)	Satisfaction (y)
1	55	50	68
2	46	24	77
3	30	46	96
4	35	48	80
5	59	58	43
6	61	60	44
7	74	65	26
8	38	42	88
9	27	42	75
10	51	50	57
11	53	38	56
12	41	30	88
13	37	31	88
14	24	34	102
15	42	30	88
16	50	48	70
17	58	61	52
18	60	71	43
19	62	62	46
20	68	38	56
21	70	41	59
22	79	66	26
23	63	31	52
24	39	42	83
25	49	40	75

the response satisfaction is also measured on an increasing scale, with larger values indicating greater satisfaction.

We will fit a multiple linear regression model to the patient satisfaction data. The model is

$$y = \beta_0 + \beta_1 x_1 + \ \beta_2 x_2 + \varepsilon$$

where y = patient satisfaction, x_1 = patient age, and x_2 = illness severity. To solve the least squares normal equations, we will need to set up the $X^T X$ matrix and the X'y vector

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 55 & 46 & \cdots & 49 \\ 50 & 24 & \cdots & 40 \end{bmatrix} \begin{bmatrix} 1 & 55 & 50 \\ 1 & 46 & 24 \\ \vdots & \vdots & \vdots \\ 1 & 49 & 40 \end{bmatrix} = \begin{bmatrix} 25 & 1271 & 1148 \\ 1271 & 69881 & 60814 \\ 1148 & 60814 & 56790 \end{bmatrix}$$

and

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 55 & 46 & \cdots & 49 \\ 50 & 24 & \cdots & 40 \end{bmatrix} \begin{bmatrix} 68 \\ 77 \\ \vdots \\ 75 \end{bmatrix} = \begin{bmatrix} 1638 \\ 76487 \\ 70426 \end{bmatrix}$$

we can find the least squares estimates of the parameters in the regression model as

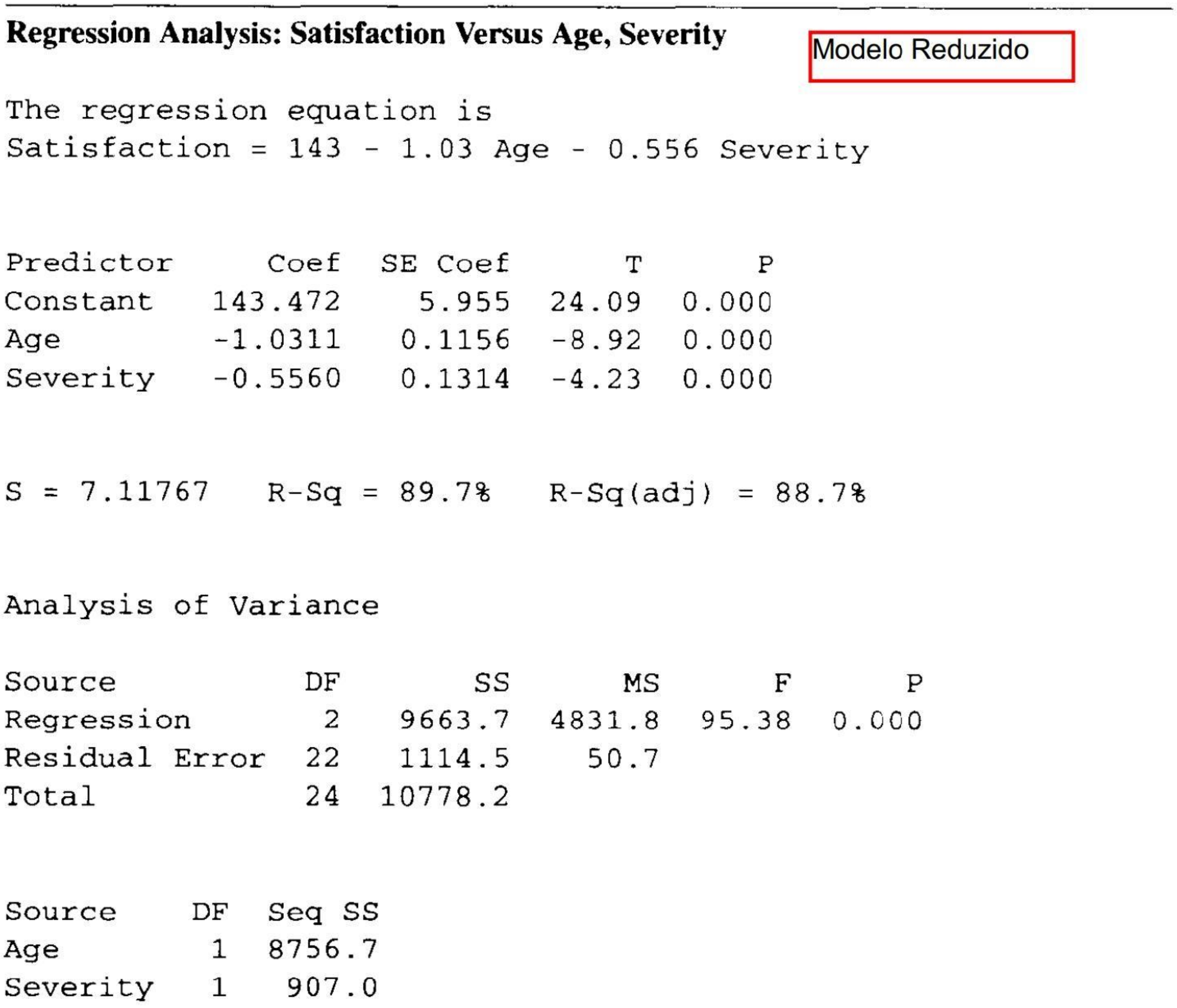
$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \begin{bmatrix} 25 & 1271 & 1148 \\ 1271 & 69881 & 60814 \\ 1148 & 60814 & 56790 \end{bmatrix}^{-1} \begin{bmatrix} 1638 \\ 76487 \\ 70426 \end{bmatrix} \\ &= \begin{bmatrix} 0.699946097 & -0.006128086 & -0.007586982 \\ -0.006128086 & 0.00026383 & -0.000158646 \\ -0.007586982 & -0.000158646 & 0.000340866 \end{bmatrix} \begin{bmatrix} 1638 \\ 76487 \\ 70426 \end{bmatrix} \\ &= \begin{bmatrix} 143.4720118 \\ -1.031053414 \\ -0.55603781 \end{bmatrix} \end{aligned}$$

Therefore the regression model of patient satisfaction on age (x_1) and severity (x_2) is:

$$\hat{y} = 143.47 - 1.03x_1 - 0.556x_2$$

Interpretation: Holding severity constant, each additional year of age reduces satisfaction score by about 1.03.

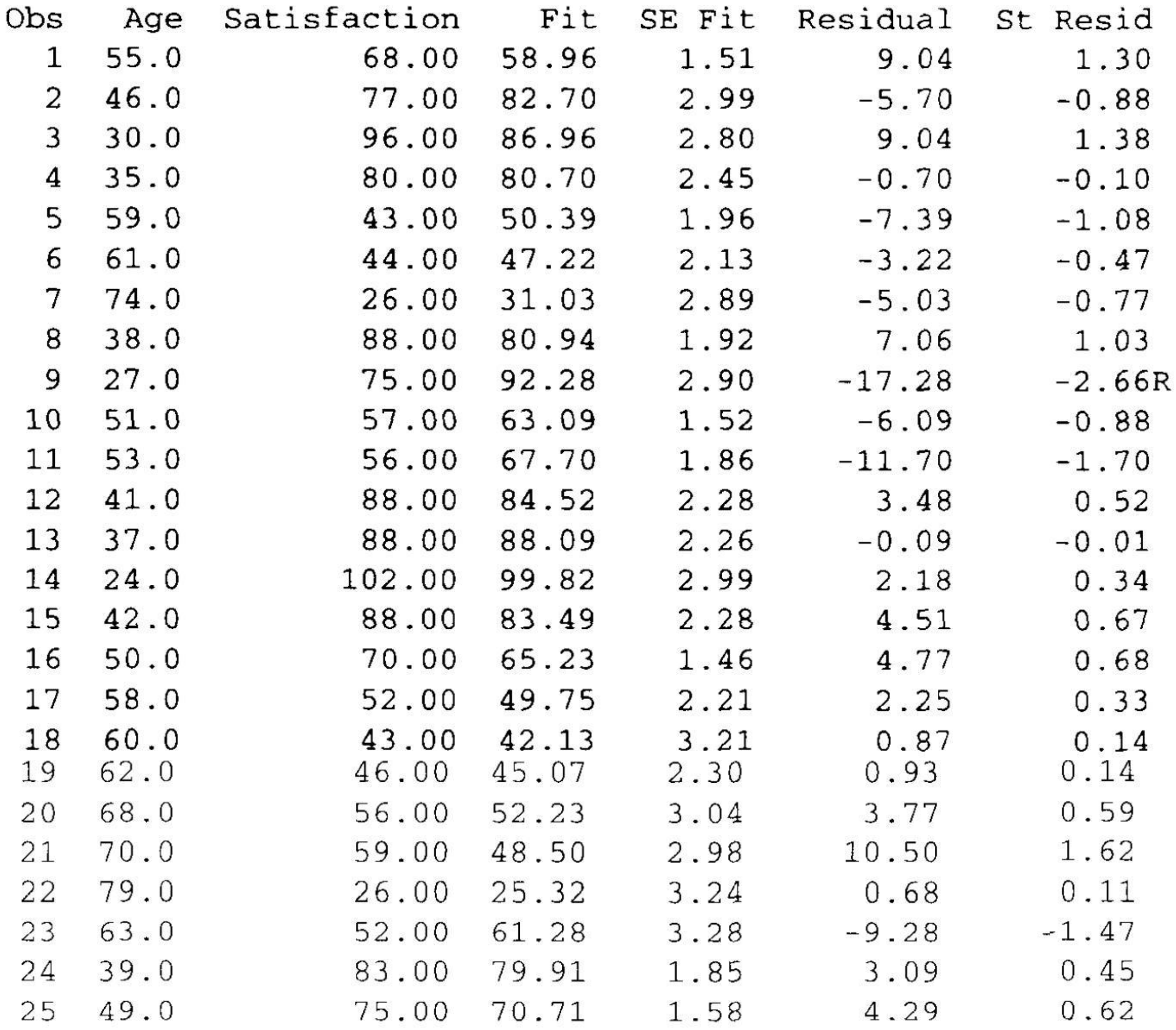
TABLE 3.3 Minitab Regression Output for the Patient Satisfaction Data in Table 3.2



Regression Analysis: Satisfaction Versus Age, Severity

Modelo Reduzido

```
The regression equation is
Satisfaction = 143 - 1.03 Age - 0.556 Severity


Predictor      Coef  SE Coef      T      P
Constant    143.472    5.955  24.09  0.000
Age         -1.0311   0.1156  -8.92  0.000
Severity    -0.5560   0.1314  -4.23  0.000


S = 7.11767   R-Sq = 89.7%   R-Sq(adj) = 88.7%


Analysis of Variance

Source          DF       SS      MS      F      P
Regression       2   9663.7  4831.8  95.38  0.000
Residual Error  22   1114.5    50.7
Total           24  10778.2


Source    DF  Seq SS
Age        1  8756.7
Severity   1   907.0
```



```
Obs   Age  Satisfaction     Fit  SE Fit  Residual   St Resid
  1  55.0         68.00   58.96    1.51      9.04       1.30
  2  46.0         77.00   82.70    2.99     -5.70      -0.88
  3  30.0         96.00   86.96    2.80      9.04       1.38
  4  35.0         80.00   80.70    2.45     -0.70      -0.10
  5  59.0         43.00   50.39    1.96     -7.39      -1.08
  6  61.0         44.00   47.22    2.13     -3.22      -0.47
  7  74.0         26.00   31.03    2.89     -5.03      -0.77
  8  38.0         88.00   80.94    1.92      7.06       1.03
  9  27.0         75.00   92.28    2.90    -17.28      -2.66R
 10  51.0         57.00   63.09    1.52     -6.09      -0.88
 11  53.0         56.00   67.70    1.86    -11.70      -1.70
 12  41.0         88.00   84.52    2.28      3.48       0.52
 13  37.0         88.00   88.09    2.26     -0.09      -0.01
 14  24.0        102.00   99.82    2.99      2.18       0.34
 15  42.0         88.00   83.49    2.28      4.51       0.67
 16  50.0         70.00   65.23    1.46      4.77       0.68
 17  58.0         52.00   49.75    2.21      2.25       0.33
 18  60.0         43.00   42.13    3.21      0.87       0.14
 19  62.0         46.00   45.07    2.30      0.93       0.14
 20  68.0         56.00   52.23    3.04      3.77       0.59
 21  70.0         59.00   48.50    2.98     10.50       1.62
 22  79.0         26.00   25.32    3.24      0.68       0.11
 23  63.0         52.00   61.28    3.28     -9.28      -1.47
 24  39.0         83.00   79.91    1.85      3.09       0.45
 25  49.0         75.00   70.71    1.58      4.29       0.62
```

note on the large residual in observation 9.

3.6 Trend Adjustment

One way to forecast time series data that contains a linear trend is with a trend adjustment procedure. This involves fitting a model with a linear trend term in time, subtracting the fitted values from the original observations to obtain a set of residuals that are trend-free, then forecast the residuals, and compute the forecast by adding the forecast of the residual value(s) to the estimate of trend.

Consider the annual U.S. production of blue and gorgonzola cheeses shown in Figure 1.4. There is clearly a positive. Nearly linear trend. The trend analysis plot in Figure 2.17 shows the original time series with the fitted line. Plots of the residuals from this model indicate that. In addition to an underlying trend, there is additional structure. The normal probability plot (Figure 2.18a) and histogram (Figure 2.18c) indicate the residuals are approximately normally distributed. However, the plots of residuals versus fitted values (Figure 2.18b) and versus observation order (Figure 2.18d) indicate nonconstant variance in the last half of the time series.

line.

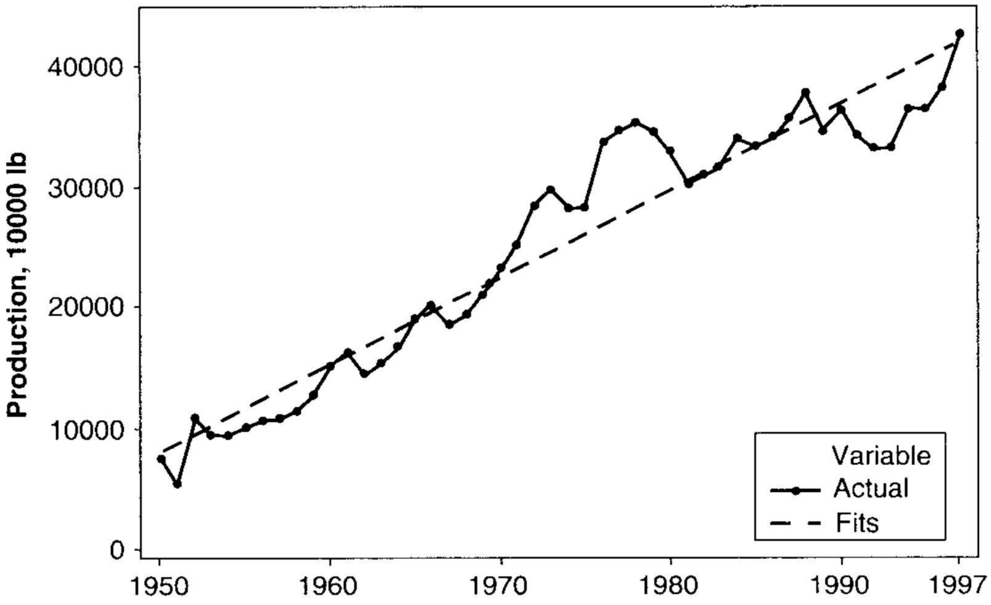
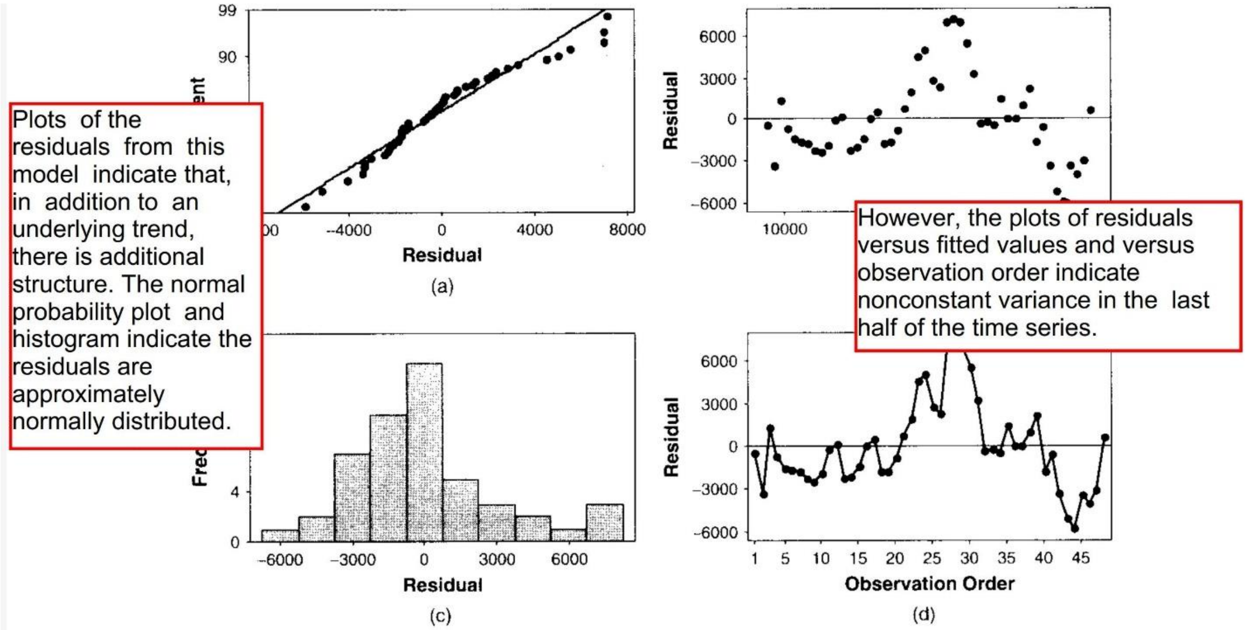


FIGURE 2.17 Blue and gorgonzola cheese production, with fitted regression line. (*Source*: USDA–NASS.)



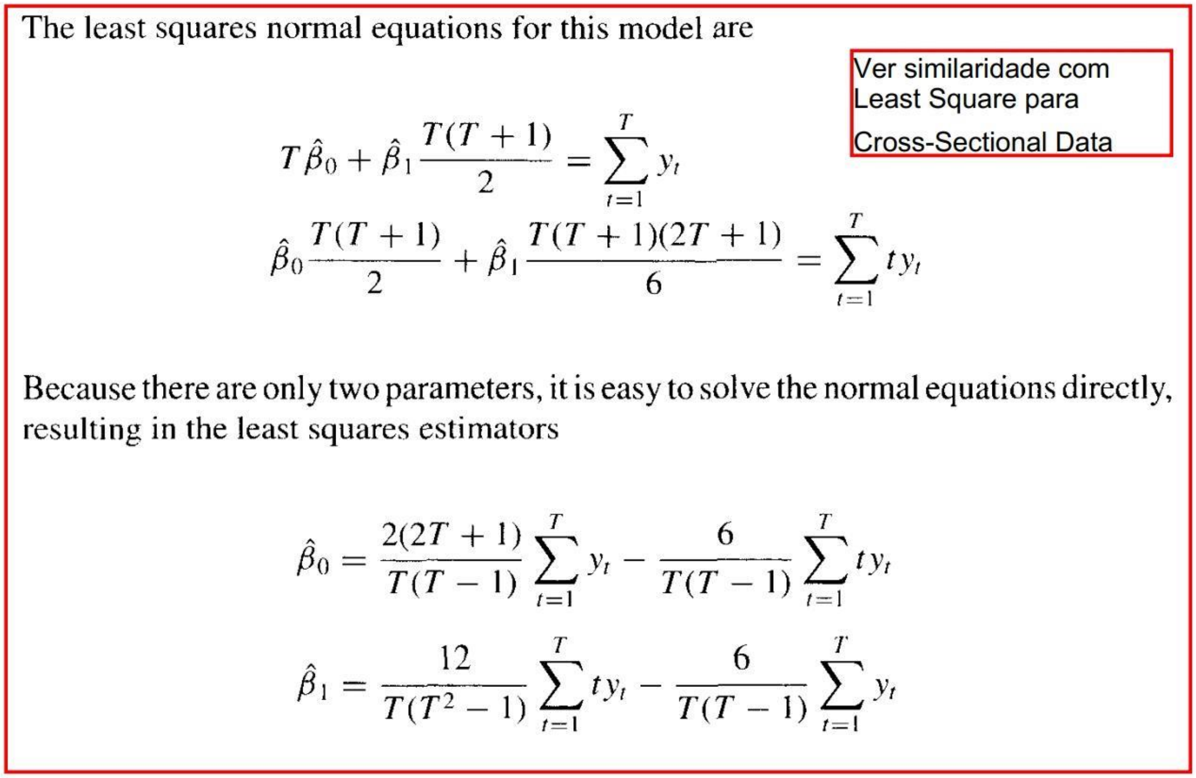
Plots of the residuals from this model indicate that, in addition to an underlying trend, there is additional structure. The normal probability plot and histogram indicate the residuals are approximately normally distributed.

However, the plots of residuals versus fitted values and versus observation order indicate nonconstant variance in the last half of the time series.

FIGURE 2.18 Residual plots for simple linear regression model of blue and gorgonzola cheese production.

The basic trend adjustment model is

$$y_t = \beta_0 + \beta_1 t + \varepsilon, \qquad t = 1,2,\ldots,T$$



The least squares normal equations for this model are

Ver similaridade com Least Square para Cross-Sectional Data

$$T\hat{\beta}_0 + \hat{\beta}_1 \frac{T(T+1)}{2} = \sum_{t=1}^{T} y_t$$

$$\hat{\beta}_0 \frac{T(T+1)}{2} + \hat{\beta}_1 \frac{T(T+1)(2T+1)}{6} = \sum_{t=1}^{T} t y_t$$

Because there are only two parameters, it is easy to solve the normal equations directly, resulting in the least squares estimators

$$\hat{\beta}_0 = \frac{2(2T+1)}{T(T-1)} \sum_{t=1}^{T} y_t - \frac{6}{T(T-1)} \sum_{t=1}^{T} t y_t$$

$$\hat{\beta}_1 = \frac{12}{T(T^2-1)} \sum_{t=1}^{T} t y_t - \frac{6}{T(T-1)} \sum_{t=1}^{T} y_t$$

The least squares estimates obtained from this trend adjustment model depend on the point in time at which they were computed, That is, T. Sometimes it may be convenient to keep track of the period of computation and denote the estimates as functions of time, say, $\beta_0(T), \beta_1(T)$. The model can be used to predict the next observation by predicting the point on the trend line in period T + 1, which is $\beta_0(T) + \beta_1(T)(T + 1)$, and adding to the trend a forecast off the next residual, say, $\varepsilon_{T+1}(1)$. If the residuals are structureless and have average value zero, the forecast of the next residual would be zero. When a new observation becomes available, the parameter estimates $\beta_0(T), \beta_1(T)$ could be updated to reflect the new information.

4. Statistical Inference in Regression

4.1 Significance of Regression (Overall F-test)

Hypotheses:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \qquad H_1: At\ least\ one\ \ \beta_j \neq 0$

Test statistic:

$F_0 = \frac{\frac{SSR}{k}}{\frac{SSE}{n-p}}$

where:

- $SSE = \left(y - X\hat{\beta}\right)^T (y - X\hat{\beta})$: Error sum of squares.
- $SSR = \hat{\beta}^T X^T y - n\bar{y}^2$: Regression sum of squares.
- n-p: Degrees of freedom for error.

If F_0 is large, reject H_0, meaning predictors explain significant variation. These results are summarized in the ANOVA table. It also shows the statistic $R^2 = \frac{SSR}{SSR+SSE}$, which is a measure of the amount of reduction in the variability of y obtained by using the predictor variables $x_1, \dots, x_k$ In the model.

However, a large value of R^2 does not necessarily imply that the regression model is a good one. Adding a variable to the model will never cause a decrease in R^2, even in situations where the additional variable is not statistically significant. In almost all cases, when a variable is added to the regression model R^2 increases. As a result, over reliance on R^2 as a measure of model adequacy often results in overfitting: that is, putting too many predictors in the model.

$$Adj\ R^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SSR + SSE}{n-1}}$$

In general, the adjusted R^2 statistic will not always increase as variables are added to the model. In fact, if unnecessary regressors are added, the value of the adjusted R^2 statistic will often decrease. Consequently, models with a large value of the adjusted R^2 statistic are usually considered good regression models.

4.2 Individual Coefficient Tests (t-tests)

Hypotheses:

$H_0: \beta_j = 0 \quad vs \quad H_1: \beta_j \neq 0$

Test statistic:

$t_0 = \frac{\widehat{\beta_j}}{se(\widehat{\beta_j})}$

where $se(\widehat{\beta_j}) = \sqrt{\widehat{\sigma^2}\ C_{jj}}$ and C_{jj} is the j-th diagonal element of $(X^T X)^{-1}$.

4.3 Partial F-tests

Used to test groups of coefficients simultaneously (extra sum of squares principle).

Hypotheses:

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$$

where $\beta_1 = [\beta_1, \dots, \beta_r]^T$ are the coefficients of the variables/predictors you want to test.

Test statistic:

$$F_0 = \frac{\hat{\beta}^T X^T y - \widehat{\beta_2}^T X_2^T y}{se(\hat{\beta}_j)}$$

where $\beta_2 = [\beta_{r+1}, \dots, \beta_k]^T$ are the coefficients and X_2 the variables/predictors you do not want to test.

4.4 Confidence Intervals

- For coefficients:

$$\widehat{\beta_J} \pm t_{\frac{\alpha}{2},n-p} \times se(\widehat{\beta_J})$$

- For mean response at x_0:

$$\hat{y}(x_0) \pm t_{\frac{\alpha}{2},n-p} \sqrt{\widehat{\sigma^2}\, x_0^T (X^T X)^{-1} x_0}$$

5. Forecasting with Regression

5.1 Prediction of New Observations

Point prediction: $\hat{y}(x_0) = x_0^T \hat{\beta}$.

Prediction Interval (PI):

$$\hat{y}(x_0) \pm t_{\frac{\alpha}{2},n-p} \sqrt{\widehat{\sigma^2}\, (1 + x_0^T (X^T X)^{-1} x_0)}$$

Difference from CI: PI is wider because it accounts for both uncertainty in mean estimate and natural variability of future observations.

Example – Patient Age 75, Severity 60

Predicted satisfaction: 32.78
95% CI for mean: (26.99, 38.57)
95% PI for individual: (16.93, 48.64)

6. Model Adequacy Checking

6.1 Residual Analysis

1. Residual vs Fitted Plot

- Purpose: Checks homoscedasticity (constant variance of errors).
- Ideal Pattern: Random scatter of points around 0, with no systematic shape.
- Problems to Watch For:
    - Funnel shape (widening/narrowing): Variance increases or decreases with fitted values → heteroscedasticity.
    - Curvature: Indicates the model might be missing nonlinear terms.

2. Residual vs Predictor Plot

- Purpose: Checks if the assumed functional form of predictors is correct.
- Ideal Pattern: Random scatter around 0, no obvious trend with predictor values.
- Problems to Watch For:
    - Systematic curve (e.g., U-shape): Suggests adding a quadratic or nonlinear term for that predictor.
    - Step changes: May indicate a categorical variable not properly modeled.

3. Residual vs Time (for time series data)

- Purpose: Checks for autocorrelation (correlation between errors over time).
- Ideal Pattern: Residuals appear randomly scattered, no pattern over time.
- Problems to Watch For:
    - Trends: Indicates missing time trend component in the model.
    - Cyclic patterns: Suggests seasonality not captured.
    - Runs of positive/negative residuals: Evidence of autocorrelation.

4. Normal Probability Plot (Q-Q Plot)

- Purpose: Checks if residuals are normally distributed (important for inference).
- Ideal Pattern: Points fall approximately along a straight diagonal line.
- Problems to Watch For:
    - S-shaped curve: Indicates skewness (non-normality).
    - Heavy tails (points deviate at ends): Indicates outliers or heavy-tailed distribution.

  - Outliers far from line: Suggest potential influential points.

6.2 Scaled Residuals

- **Standardized residuals**: $d_i = \frac{e_i}{\hat{\sigma}}$. Values beyond ±3 may indicate outliers.
- **Studentized residuals**: $r_i = \frac{e_i}{\sqrt{\widehat{\sigma^2}(1-h_{ii})}}$ adjust for varying variances. (h_{ii} is the ith diagonal element of the hat matrix $H = X(X^TX)^{-1}X^T$)
- The elements h_{ij} of the hat matrix H may be interpreted as the amount of leverage exerted by the observation y_j on the predicted value $\widehat{y_\iota}$. Thus inspection of the elements of H can reveal points that are potentially influential by virtue of their location in x-space.

6.3 PRESS Statistic

Prediction Residual Sum of Squares, used for model validation by leave-one-out predictions.

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} \left(y_i - \widehat{y_{(\iota)}}\right)^2$$

($\widehat{y_{(\iota)}}$ predicted when ith variable witheld) Generally, small values of PRESS imply that the regression model will be useful in predicting new observations.

7. Variable Selection Methods

Including too many predictors may cause overfitting. Common methods:

- Forward Selection: Start with none, add predictors one by one if significant.
- Backward Elimination: Start with all, remove non-significant predictors.
- Stepwise Regression: Combination of forward and backward procedures.
- Information Criteria: Choose models based on AIC or BIC.

TABLE 3.8 Minitab Forward Selection for the Patient Satisfaction Data in Table 3.6

Stepwise Regression: Satisfaction Versus Age, Severity, . . .

```
Forward selection.  Alpha-to-Enter: 0.25


Response is Satisfaction on 4 predictors, with N = 25


Step                1       2       3
Constant        131.1   143.5   143.9

Age             -1.29   -1.03   -1.11
T-Value         -9.98   -8.92   -8.40
P-Value         0.000   0.000   0.000

Severity                -0.56   -0.58
T-Value                 -4.23   -4.43
P-Value                 0.000   0.000

Anxiety                           1.3
T-Value                          1.23
```

TABLE 3.9 Minitab Backward Elimination for the Patient Satisfaction Data

Stepwise Regression: Satisfaction Versus Age, Severity, . . .

```
Backward elimination.  Alpha-to-Remove: 0.1


Response is Satisfaction on 4 predictors, with N = 25


Step                1       2       3
Constant        143.9   143.9   143.5

Age             -1.12   -1.11   -1.03
T-Value         -8.08   -8.40   -8.92
P-Value         0.000   0.000   0.000

Severity        -0.59   -0.58   -0.56
T-Value         -4.32   -4.43   -4.23
P-Value         0.000   0.000   0.000

Surg-Med          0.4
T-Value          0.14
P-Value         0.892

Anxiety           1.3     1.3
T-Value          1.21    1.23
P-Value         0.242   0.233

S                7.21    7.04    7.12
R-Sq            90.36   90.35   89.66
R-Sq(adj)       88.43   88.97   88.72
```

TABLE 3.10 Minitab Stepwise Regression Applied to the Patient Satisfaction Data

Stepwise Regression: Satisfaction Versus Age, Severity, . . .

```
  Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15


Response is Satisfaction on 4 predictors, with N = 25


Step             1      2
Constant     131.1  143.5

Age          -1.29  -1.03
T-Value      -9.98  -8.92
P-Value      0.000  0.000

Severity            -0.56
T-Value             -4.23
P-Value             0.000

S             9.38   7.12
R-Sq         81.24  89.66
R-Sq(adj)    80.43  88.72
```

TABLE 3.11 Minitab All Possible Regressions Algorithm Applied to the Patient Satisfaction Data

Best Subsets Regression: Satisfaction Versus Age, Severity, . . .

Ver Best Subsets

```
Response is Satisfaction


                                           S S
                                           e u A
                                           v r n
                                           e g x
                                           r - i
                                         A i M e
                       Mallows           g t e t
Vars  R-Sq  R-Sq(adj)      C-p       S   e y d y
   1  81.2       80.4     17.9  9.3752   X
   1  52.3       50.2     78.0  14.955     X
   2  89.7       88.7      2.5  7.1177   X X
   2  81.3       79.6     19.7  9.5626   X     X
   3  90.4       89.0      3.0  7.0371   X X   X
   3  89.7       88.2      4.5  7.2846   X X X
   4  90.4       88.4      5.0  7.2074   X X X X
```

7.2 Generalized and Weighted Least Squares

OLS assumptions may fail when errors are correlated or heteroscedastic.

Generalized Least Squares (GLS)

- Used when error variance-covariance matrix is not $\sigma^2 \; I$ (unequal variance and correlated).
- Transformation applied so that errors become homoscedastic and uncorrelated.
- Estimator: $\hat{\beta}_{GLS} \; = \; (X^T V^{-1} X)^{-1} X^T V^{-1} y$ where V is the covariance matrix of errors.

Weighted Least Squares (WLS)

- Special case of GLS when errors have unequal variances but uncorrelated.
- Each observation is weighted inversely proportional to its variance.
- Useful for heteroscedastic data (e.g., variance increasing with x).
- Estimator: $\hat{\beta}_{WLS} = (X^T W X)^{-1} X^T W y,$ wHere $W = V^{-1}, \; V$ diagonal matrix of variance of each variable.

Discounted Least Squares (DLS)

- Used in forecasting to give more weight to recent observations.
- Weights decrease for older data points, making the model adapt to changes over time.

$$w_t \; = \lambda^{T-t}, \qquad 0 \; < \lambda < \; 1$$

where T is the most recent period.

- If $\lambda \; = \; 1$: reduces to standard OLS (equal weights).
- If lambda < 1: older data have geometrically smaller weights.
- Estimator: $\widehat{\beta_{DLS}} \; = \; (X^T W X)^{-1} \; X^T W y, \;\; W \; = \; diag(w_1, \ldots, w_T).$

Choice of λ:

- Determines memory of the model.
- Monthly data: λ ∈ [0.90, 0.99].
- Daily/weekly data: λ ∈ [0.95, 0.995].
- Selected via cross-validation, information criteria, or by minimizing one-step-ahead forecast error.

Applications:

- Financial econometrics: market regimes shift rapidly.
- Sales forecasting: changing customer preferences.
- Macroeconomics: adaptive policy response.
- Engineering & process control: emphasis on latest sensor readings.

- Online learning / machine learning pipelines: streaming environments.

8. Regression Models for General Time Series Data

Many time series require regression models that combine deterministic trends, seasonality, interventions, and external regressors with stochastic error processes.

The presence of autocorrelation in the errors has several effects on the ordinary least squares regression procedure. These are summarized as follows:

1. The ordinary least squares (OLS) regression coefficients are still unbiased, but they are no longer minimum-variance estimates.

2. When the errors are positively autocorrelated, the residual mean square may seriously underestimate the error variance. Consequently, the standard errors of the regression coefficients may be too small. As a result, confidence and prediction intervals are shorter than they really should be, and tests of hypotheses on individual regression coefficients may be misleading in that they may indicate that one or more predictor variables contribute significantly to the model when they really do not. Generally, underestimating the error variance gives the analyst a false impression of precision of estimation and potential forecast accuracy.

3. The confidence intervals, prediction intervals, and tests of hypotheses based on the t and F distributions are, strictly speaking. no longer exact procedures.

8.1 Deterministic Components

- Trend: linear, quadratic, spline, or segmented.
- Seasonality: dummy variables, Fourier terms.
- Events: holidays, promotions, shocks.
- Exogenous regressors: weather, macroeconomic indicators.

There are three approaches to dealing with the problem of autocorrelation. If autocorrelation is present because of one or more omitted predictors and if those predictor variable(s) can be identified and included in the model, the observed autocorrelation should disappear. Alternatively, the weighted least squares or generalized least squares methods could be used if there were sufficient knowledge of the autocorrelation structure. Finally, if these approaches cannot be used. the analyst must tum to a model that specifically incorporates the autocorrelation structure. These models usually require special parameter estimation techniques.

8.2 Stochastic Error Structures

- AR, MA, ARMA, ARIMA errors.
- ARIMAX/dynamic regression: regression plus ARIMA error.
- Motivation: correct SEs, improve forecast accuracy.

8.3 Distributed-Lag & Transfer Function Models

- Capture delayed effects of regressors.
- Finite lags (y_t depends on $x_t, x_{t-1}, ...$).
- Infinite distributed lags approximated with rational functions.
- Useful for policy, advertising, environmental impacts.

8.4 Applications

- Economics: consumption vs income, with lagged effects.
- Marketing: sales vs advertising.
- Energy: demand vs temperature.
- Public health: cases vs interventions.

9. Detecting Autocorrelation: The Durbin–Watson Test

9.1 Definition

$$DW = \frac{\Sigma(e_t - e_{t-1})^2}{\Sigma\, e_t^2}.$$

- Range: [0,4].
- ≈2 → no autocorr; <2 → positive; >2 → negative.

9.2 Test Procedure

- H₀: no autocorrelation.
- Compare DW with lower/upper bounds d_L, d_U.
- Decision: $DW < d_L$ reject H₀; $DW > d_U$ fail to reject; else inconclusive.

10. Estimating Parameters in Time Series Regression Models

When regression errors are autocorrelated, OLS estimators remain unbiased but lose efficiency, and standard errors become unreliable. To address this, advanced estimation methods such as the Cochrane–Orcutt procedure, the Maximum Likelihood (ML) approach, and proper handling of forecasting and prediction intervals are required.

10.1 Cochrane–Orcutt Method

Problem setup:
Consider the regression model with AR(1) errors:

$$y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + \varepsilon_t,$$

$$\varepsilon_t = \rho\, \varepsilon_{t-1} + u_t, \quad u_t \sim \ i.i.d.\ (0, \sigma^2), \qquad |\rho| < 1.$$

- OLS estimates of β are unbiased but inefficient.
- The error variance is underestimated if autocorrelation is ignored.
- Confidence intervals and hypothesis tests become invalid.

Idea: Estimate ρ and transform the model so the errors become approximately uncorrelated. Then apply OLS to the transformed system.

Procedure (iterative feasible GLS):

1. Initial OLS fit: Run OLS regression, obtain residuals $\widehat{e_t}$.
2. Estimate autocorrelation: Regress $\widehat{e_t}$ on $\widehat{e_{t-1}}$ to estimate $\hat{\rho}$.
3. Transform the data:

$$y_t^* = y_t - \hat{\rho}\, y_{t-1}, \qquad x_{jt}^* = x_{jt} - \hat{\rho}\, x_{j,t-1}, \ j = 1, \ldots, k.$$

4. Refit regression: Run OLS on $y_t^* \ vs \ x_t^*$, obtain new estimates $\hat{\beta}$.
5. Repeat: Re-estimate residuals, update $\hat{\rho}$, and iterate until convergence.

Notes:

- Provides consistent estimates of coefficients and more reliable standard errors.
- Best suited for AR(1) error structures.
- If higher-order autocorrelation exists, generalized Cochrane–Orcutt (or full GLS) may be needed.

10.2 Maximum Likelihood Approach

Model framework:
Suppose regression errors follow an ARMA(p,q) process:

$$y_t = x_t'\beta + n_t, \quad n_t = \phi_1\, n_{t-1} + \cdots + \phi_p\, n_{t-p} + \theta_1\, a_{t-1} + \cdots + \theta_q\, a_{t-q} + a_t,$$

where $a_t \sim N(0, \sigma^2)$.

Steps:

1. Construct likelihood function: Under normality, residuals can be expressed in terms of innovations a_t.

$$L(\beta, \phi, \theta, \sigma^2) \propto \sigma^{-T}\ exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T} a_t^2\right)$$

2. Profile likelihood: Concentrate out σ^2:

$$\widehat{\sigma^2} = \frac{1}{T}\sum_{t=1}^{T} a_t^2$$

   Substitute into the log-likelihood for maximization over β,ϕ,θ.

3. Numerical optimization: Use algorithms (Newton–Raphson, BFGS, or Kalman filter recursions) to maximize likelihood.
4. Inference: Compute standard errors using the Hessian (observed information). Perform likelihood ratio tests or Wald tests for hypothesis testing.