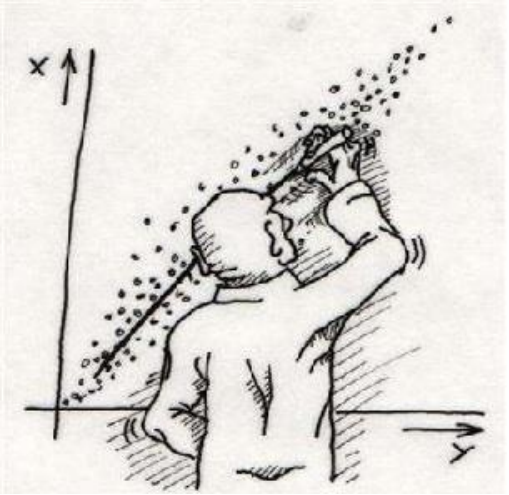
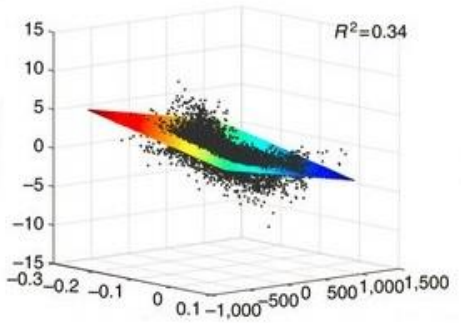
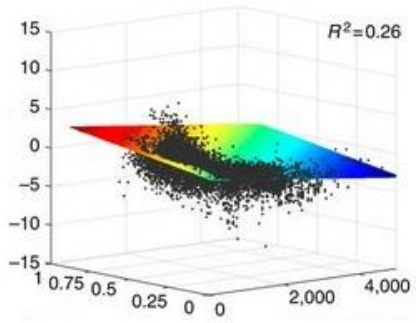


DATA 621 - Business Analytics and Data Mining





Meetup #1

Department of Data Analytics and IS
CUNY School of Professional Studies
The City University of New York

Dr. Nathan Bastian

Course Introduction

Dr. Nathan Bastian (Prof. B)

*An **operations researcher**, **data scientist**, **decision analyst** and **industrial engineer** who discovers and translates data-driven, actionable insights into effective decisions using mathematics, statistics, engineering, economics, and computational science to develop decision-support models for descriptive, predictive and prescriptive analytics.*

Current Employment:

- Healthcare Operations Research Analyst, **U.S. Army Medical Department**
- Managing Partner and Chief Decision Scientist, **Brewery Analytics Consulting, LLC**
- Adjunct Faculty, **The City University of New York, Northwestern University, Worcester Polytechnic Institute**

Academic Training:

- **Ph.D.**, *Industrial Engineering and Operations Research*, Pennsylvania State University
- **M.Eng.**, *Industrial Engineering*, Pennsylvania State University
- **M.S.**, *Econometrics and Operations Research*, Maastricht University, The Netherlands
- **B.S.**, *Engineering Management with Honors*, U.S. Military Academy at West Point

Email: nathaniel.bastian@sps.cuny.edu

What is Business Analytics?

- **Business analytics** refers to the skills, technologies, practices for continuous iterative exploration, and investigation of past business performance to gain insight and drive business planning.

- It focuses on developing **new insights** and understanding of business performance based on data and statistical methods.

- One component is **predictive analytics**, which deals with predictive modeling through statistical inference, data mining, and machine learning techniques.

What is Data Mining?

- **Data mining** is a set of methods that *automatically* detect patterns in data so that they can be understood. The key premise is ***learning*** from data!!

- These uncovered patterns are then used to **predict** future data, or to perform other kinds of decision-making under uncertainty.

- Tools, methodologies, and theories from statistics, computer science, etc. for revealing patterns in data – critical step in **knowledge discovery**.

Course Description

- This course develops the foundations of **predictive modeling** by introducing the key concepts of applied regression modeling and its extensions.

- The course is heavily weighted towards **practical application** using the R statistical programming language and data sets containing missing values and outliers.

- The course also addresses **issues** of exploratory data analysis, data preparation, model development, model validation, and model deployment.

Relevance to Data Scientists

- **Regression modeling skills** are crucial, high-value skills in today's data-driven business environment where real-world decision-making processes are complex.

- The ability to leverage rapidly expanding data sets to obtain new insights is at the heart of **predictive data analytics**.

How This Course Works

- The course is conducted **entirely online** via Blackboard.
- **Each week**, you will complete assigned readings from the required textbooks, watch lecture videos, complete optional (but recommended) textbook exercises, complete homework assignments, and conduct a final group project.
- Students are expected to complete all deliverables by their assigned due dates. **No late work** is accepted.

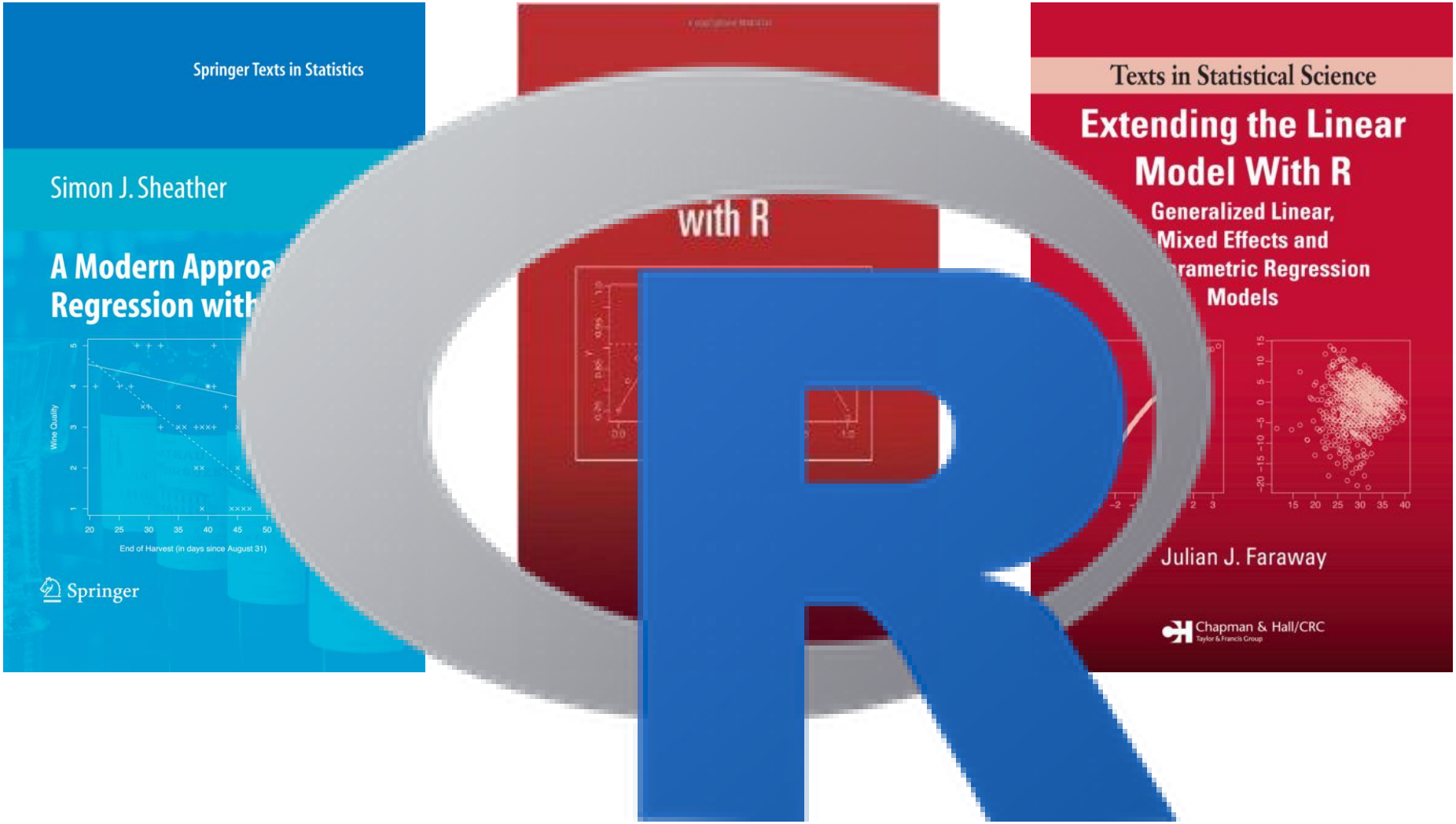
Assignments and Grading

Homework Assignments - There will be 5 homework assignments (15% each, or 150 points each) used to re-enforce course concepts and provide implementation experience. - Students will work on the homework assignments in collaboration with the members of their Critical Thinking Group. Only one homework solution should be submitted per group.	**75%**	**750 points**
Final Group Project - Students will form a Critical Thinking Group and conduct a final course project using regression modeling techniques covered in class to solve a real-world problem. - Only one project report should be submitted per group.	**25%**	**250 points**
TOTAL	**100%**	**1000 points**

Course Outline

Unit	Meetup	Topic	Readings	Key Tasks
Week #1 6/6 – 6/12	6/7 8-9pm	Simple Linear Regression: Estimation, Inference, Prediction and Explanation	*MARR* – Ch. 1, 2 *LMR* – Ch. 1 – 5	- HW #1 Assigned
Week #2 6/13 – 6/19	None	Simple Linear Regression: Diagnostics and Transformations / Multiple Linear Regression: Missing Data, Diagnostics and Transformations	*MARR* – Ch. 3, 5, 6 *LMR* – Ch. 6, 7, 9, 13, 14	- HW #1 Due - HW #2 Assigned
Week #3 6/20 – 6/26	6/21 8-9pm	Variable Selection, Shrinkage Methods, and Binary Logistic Regression	*MARR* – Ch. 7, 8 *LMR* – Ch. 10, 11 *ELMR* – Ch. 2	- HW #2 Due - HW #3 Assigned
Week #4 6/27 – 7/3	None	Weighted Least Squares, Robust Regression, and Generalized Least Squares	*MARR* – Ch. 4, 9 *LMR* – Ch. 8	- HW #3 Due - HW #4 Assigned
Week #5 7/4 – 7/10	None	Count Regression and Multinomial Logistic Regression	*ELMR* – Ch. 3, 5	- HW #4 Due - HW #5 Assigned
Week #6 7/11 – 7/17	7/12 8-9pm	Generalized Linear Models and Panel Regression	*ELMR* – Ch. 6, 7, 9	- HW #5 Due
Week #7 7/18 – 7/21	7/19 8-9pm	Nonparametric Regression	*ELMR* – Ch. 11	- Project Report Due

Textbooks and Software



Any questions?

Linear Regression

Overview: Linear Regression

- Linear regression is a simple approach to supervised learning, as it assumes that the dependence of Y on X_1, X_2,…,X_p is linear.

- Most modern machine learning approaches can be seen as generalizations or extensions of linear regression.

- When augmented with kernels or other forms of basis function expansion (which replace X with some non-linear function of the inputs), it can also model non-linear relationships.

- Goal: predict Y from X by $f(X)$

Review: Expectation

- The expectation of a random variable is its “average” value under its distribution.
- The expectation of a random variable X, denoted E[X], is its Lebesgue integral with respect to its distribution.
- If X takes values in some countable numeric set χ, then

$$E(X) = \sum_{x \in \chi} xP(X = x)$$

Review: Expectation (cont.)

- If $X \in \mathbb{R}^m$ has a density p, then $E(X) = \int_{\mathbb{R}^m} xp(x)dx$
- Expectation is linear: $E(aX + b) = aE(X) + b$
- Also, $E(X + Y) = E(X) + E(Y)$
- Expectation is monotone: if $X \geq Y$, then $\mathrm{E}(X) \geq \mathrm{E}(Y)$

Review: Variance

- The variance of a random variable X is:

$$Var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

- The variance obeys the following $a, b \in \mathbb{R}$:

$$Var(aX + b) = a^2 Var(X)$$

Review: Frequentist Basics

- The training data $X_1,\ldots,X_n$ is generally assumed to be *independent and identically distributed (iid).*

- We want to estimate some unknown value θ associated with the distribution from which the data was generated.

- In general, our estimate will be a function of the data:

$$\hat{\theta} = f(X_1, X_2, \ldots, X_n)$$

Review: Parameter Estimation

- In practice, we often seek to select a distribution (model) corresponding to our data.
- If the model is parameterized by some set of values, then this problem is that of parameter estimation.
- In general, we typically use maximum likelihood estimation (MLE) to obtain parameter estimates.

Review: Parameter Estimation (cont.)

- Given that the training data are *iid* and come from the probability density function *p*, to use MLE we first specify the joint density function (which is also the *likelihood* function):

$$\mathcal{L}(\boldsymbol{\theta}; X_1, \ldots, X_n) = \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(x_i)$$

- In practice, it is more convenient to work with the logarithm of the likelihood function, called the *log-likelihood*:

$$\ell(\boldsymbol{\theta}) = \ln \mathcal{L}(\boldsymbol{\theta}; X_1, \ldots, X_n) = \sum_{i=1}^{n} \ln p_{\boldsymbol{\theta}}(x_i)$$

Review: Parameter Estimation (cont.)

- Using the method of MLE: $\widehat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$

- Instead of maximizing the log-likelihood, we can equivalently minimize the *negative log-likelihood* (NLL):

$$NLL(\boldsymbol{\theta}) = -\ln \mathcal{L}(\boldsymbol{\theta}; X_1, \ldots, X_n) = -\sum_{i=1}^{n} \ln p_{\boldsymbol{\theta}}(x_i)$$

$$\widehat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} NLL(\boldsymbol{\theta})$$

- This formulation is sometimes more convenient, since many optimization software packages are designed to find the minima of functions, rather than maxima.

Statistical Decision Theory

- Let $X \in \mathbb{R}^p$ denote a real valued random input vector.
- Let $Y \in \mathbb{R}$ denote a real valued random output variable, with joint distribution Pr(X, Y).
- We seek a function *f*(*X*) for predicting *Y* given values of the input *X*.
- *Loss function L*(*Y, f*(*X*)) → penalizing errors in prediction.

Statistical Decision Theory (cont.)

- *Squared error loss*: $L\big(Y, f(X)\big) = (Y - f(X))^2$

- This leads us to a criterion for choosing *f*,

$$\mathrm{EPE}(f) = E(Y - f(X))^2 = \int [y - f(x)]^2 \Pr(dx, dy)$$

which is the *expected (squared) prediction error*. By conditioning on *X*, we can write EPE as

$$\mathrm{EPE}(f) = E_X E_{Y|X}([Y - f(X)]^2 | X)$$

Statistical Decision Theory (cont.)

- This suffices to minimize EPE as follows:

$$f(x) = \text{argmin}_c E_{Y|X}([Y - c]^2 | X = x)$$

- The solution is $f(x) = E(Y|X = x)$, which is the conditional expectation, also known as the *regression* function.

- The best prediction of *Y* at any point *X* = *x* is the conditional mean, when best is measured by average squared error.

- A linear regression model assumes that the regression function E(*Y* | *X*) is linear in the inputs X_1, X_2,…,X_p.

Linear Regression Model

- Input vector: $X^T = (X_1, X_2, \ldots, X_p)$
- Output Y is real-valued (quantitative response) and ordered
- We want to predict Y from X.
- Before we actually do the prediction, we have to *train* the function $f(X)$.
- By the end of training, we have a function $f(X)$ to map every X into an estimated Y (aka $\hat{Y}$).

Linear Regression Model (cont.)

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

- This is a linear combination of the measurements that are used to make predictions, plus a constant.

- No matter the source of the X_j, the model is linear in the parameters.

- β_0 is the intercept and β_j is the slope for the *j*th variable X_j, which is the **average** increase in Y when X_j is increased by one unit and all other X's are held constant.

Assumptions of the Linear Regression Model

1. **Linearity:** The model specifies a linear relationship between the response variable *y* and the predictor variables ***x*** (i.e., linear in the parameters and the disturbance).

2. **Full column rank:** There is no exact linear relationship among any of the predictors.
   - This assumption is necessary for estimation of the parameters of the model (i.e., taking an inverse).

Assumptions of the Linear Regression Model (cont.)

3. **Exogeneity of the independent variables:** The expected value of the disturbance (error term) at observation *i* in the sample is NOT a function of the predictors observed at any observation.
   - $\mathrm{E}[\varepsilon_i|\mathbf{X}] = 0 \; \forall \; i = 1,2,\ldots,n$
   - The predictors will not carry useful information for prediction of the error terms (i.e., no correlation).

Assumptions of the Linear Regression Model (cont.)

4. **Homoscedasticity and nonautocorrelation:** Each error term has the SAME finite variance, $\mathrm{Var}[\varepsilon_i|\mathbf{X}] = \sigma^2 \; \forall \; i = 1,2,\ldots,n$, and is uncorrelated with every other error term, $\mathrm{Cov}[\varepsilon_i, \varepsilon_j|\mathbf{X}] = 0 \; \forall \; i \neq j$

5. **Data generation:** The data may be any mixture of constants and random variables.

6. **Normal distribution:** The disturbances are normally distributed → $\boldsymbol{\varepsilon}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

Ordinary Least Squares Estimation

- Typically we have a set of *training data* $(X_1, Y_1)\ldots (X_n, Y_n)$ from which to estimate the parameters β.

- Each X_i is a vector of feature measurements for the ith case.

- We can apply the method of MLE to the linear regression setting (using the definition of the Gaussian), where the log-likelihood function is given by:

$$\ell(\boldsymbol{\theta}) = \frac{-1}{2\sigma^2} RSS(\boldsymbol{\beta}) - \frac{n}{2} \ln 2\pi\sigma^2$$

OLS Estimation (cont.)

- Note that RSS stands for *residual sum of squares*:

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n}(Y_i - f(X_i))^2 = \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j)^2$$

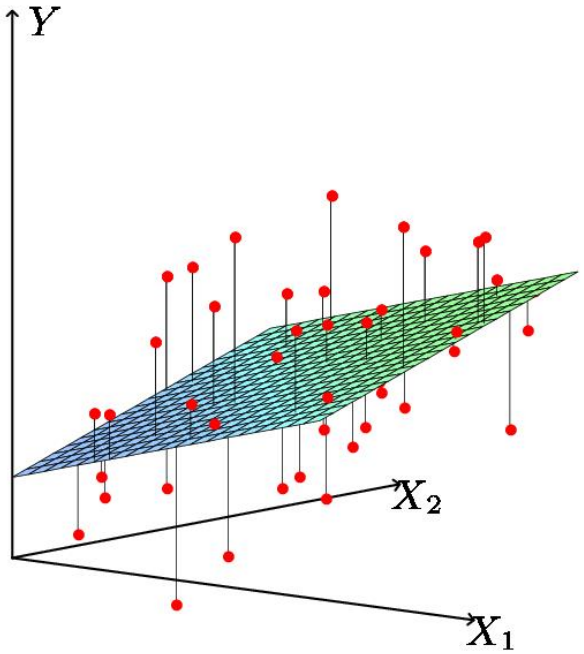
- The RSS is also called the *sum of squared errors* (SSE), where

$$MSE = \frac{SSE}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

- We see that the MLE for $\boldsymbol{\beta}$ is the one that minimizes the RSS. Thus, we estimate the parameters using *ordinary least squares* (OLS), which is identical to the MLE, to choose $\hat{\beta}_0$ through $\hat{\beta}_p$ as to minimize the RSS.

OLS Estimation (cont.)

- We illustrate the geometry of OLS fitting, where we seek the linear function of *X* that minimizes the sum of squared residuals from *Y*.



- The predictor function corresponds to a plane (hyper plane) in the 3D space.
- For accurate prediction, hopefully the data will lie close to this hyper plane, but they won't lie exactly in the hyper plane.

OLS Estimation (cont.)

For the *simple* and *multiple* linear regression model:

- Let $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{bmatrix}$ be the *n*-vector of outputs in the training set.
- Let $\boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1p} \\ 1 & X_{21} & \dots & X_{2p} \\ 1 & \dots & \dots & \dots \\ 1 & X_{n1} & \dots & X_{np} \end{bmatrix}$ be the *n* x (*p* + 1) matrix of inputs.

where there are *n* observations and *p* predictors in the training data.

OLS Estimation (cont.)

- Let $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_p \end{bmatrix}$, so the $RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X\beta})^T \ (\boldsymbol{Y} - \boldsymbol{X\beta})$
- So, we must solve the following quadratic minimization problem:

$$\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} RSS(\boldsymbol{\beta})$$

- This minimization problem has a unique solution, provided that $\boldsymbol{X}$ has full column rank (i.e. the p columns of $\boldsymbol{X}$ are linearly independent), given by solving the normal equations:

$$(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{Y}$$

OLS Estimation (cont.)

The unique solution to the normal equations yields the vector $\widehat{\boldsymbol{\beta}}$ (i.e. the OLS estimates of the parameters): $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$

Note for ***simple*** OLS linear regression (intercept and one predictor):

$$\text{Let } \boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

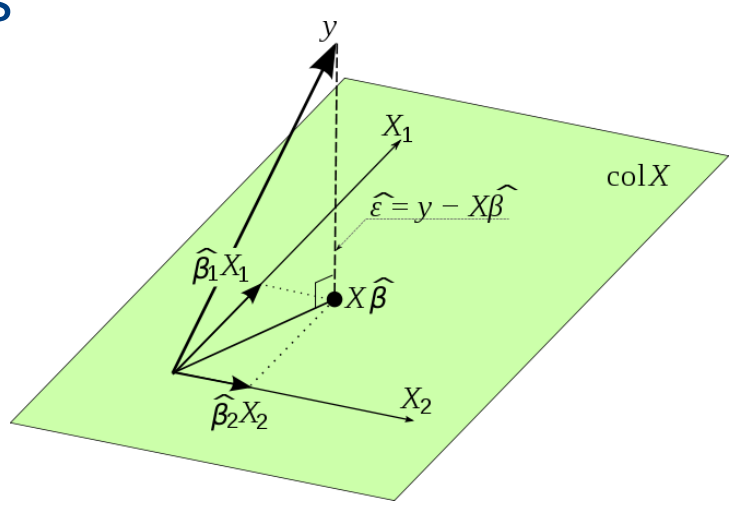
$$\text{Let } \boldsymbol{X}^T\boldsymbol{Y} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$$

$$\text{Let } (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \frac{1}{\det(\boldsymbol{X}^T\boldsymbol{X})} \begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix}$$

OLS Estimation (cont.)

- The fitted values at the training inputs (i.e. vector of the OLS predictions) are:

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



- In geometric representation, this corresponds to an orthogonal projection of $\boldsymbol{Y}$ onto the column space of $\boldsymbol{X}$.

- The matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is the projection matrix, which is called the *hat* matrix because it puts a hat on $\boldsymbol{Y}$.

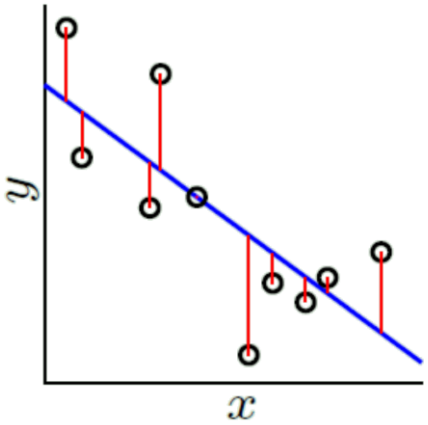
OLS Estimation (cont.)

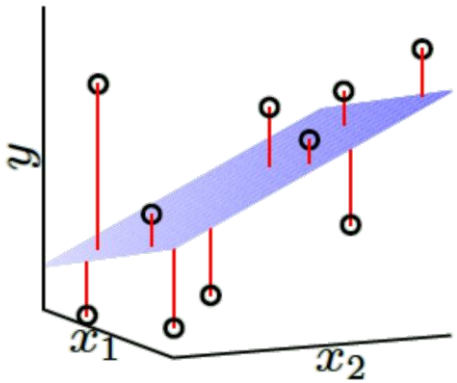
OLS Linear Regression Algorithm:

1. From the training data set, construct the input matrix $\boldsymbol{X}$ and the output vector $\boldsymbol{Y}$

2. Assuming $\boldsymbol{X}^T\boldsymbol{X}$ is invertible (positive definite and non-singular), compute $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$

3. Return $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$





OLS Estimation (cont.)

If the linear model is true (i.e. if the conditional expectation of *Y* given *X* indeed is a linear function of the X_j's), and *Y* is the sum of that linear function and an independent Gaussian noise, then we have the following properties for OLS estimation:

1. The OLS estimation of β is unbiased → $E[\hat{\beta}_j] = \beta_j \; \forall \; j = 0,1, \ldots, p$

2. To draw inferences about β, further assume:
   $Y = E(Y|X) + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$ and independent of X.

OLS Estimation (cont.)

- The estimation accuracy (variance) of $\widehat{\boldsymbol{\beta}}$ is: $Var(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2$

- Typically one estimates the variance σ^2 (unbiased) by

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$$

- When σ^2 is higher, the variance of $\widehat{\boldsymbol{\beta}}$ is higher. Also, the variance-covariance matrix tells us the variance for every individual beta and also the covariance for any pair of betas.

- **Gauss-Markov Theorem**: The OLS estimates of the parameters $\boldsymbol{\beta}$ have the smallest variance (i.e. smallest mean squared error) among all linear unbiased estimates.

Population vs. OLS Lines

- Population Regression Line: $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon$
- OLS Regression Line: $\hat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j X_j$

- We would like to know the population parameters, but we only know the OLS estimates $\hat{\beta}_0$ through $\hat{\beta}_p$.

- Further, we use $\hat{Y}_i$ as an estimate for Y_i.

Accuracy of Coefficient Estimates

- Let's consider a simple linear regression model with $\hat{\beta}_0$ and $\hat{\beta}_1$. How close are $\hat{\beta}_0$ and $\hat{\beta}_1$ to the true values β_0 and β_1, respectively?

- We can answer this by computing the standard errors associated with $\hat{\beta}_0$ and $\hat{\beta}_1$.

- These SEs can be used to compute confidence intervals (CIs), prediction intervals (PIs), and perform hypothesis tests on the coefficients.

Accuracy of the Model: RSE

- The residual standard error (RSE) is an estimate of the standard deviation of ε.

- In other words, RSE is the average amount that the response will deviate from the true regression line:

$$RSE = \sqrt{\frac{RSS}{n - p - 1}}$$

where *p* is the number of predictors (slopes) in the regression model (not including the intercept).

Accuracy of the Model: R^2

- The proportion of variability in *Y* that can be explained using *X:*

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

where total sum of squares (TSS) measures the total variance in the response *Y*. It is thought of as the amount of variability inherent in the response before the regression is performed.

- Note that RSS measures the amount of variability that is left unexplained after performing the regression.
- Always between 0 (no fit) and 1 (perfect fit).

Two Key Questions

1. Is $\beta_j = 0$ or not?

2. Is at least one of the predictors useful in predicting the response?

Is $\beta_j = 0$ or not?

- H_0: There is no relationship between X_j and Y ($\beta_j = 0$).
- H_a: There is some relationship between X_j and Y ($\beta_j \neq 0$).

- Compute the *t-statistic*: $t = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)}$

- If *t* is large (and the *p-value* is small, typically < α = 0.05), then we reject H_0 and declare that there is relationship.

- We use the Regression Output table to get the beta coefficients, standard errors, t-statistics and p-values.

Are all regression coefficients 0?

- H_0: all slopes equal 0 ($\beta_1 = \beta_2 = \cdots = \beta_p = 0$).
- H_a: at least one slope $\neq 0$.

- Compute the *F-statistic:* $F = \frac{(TSS-RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$

- We use the ANOVA table to get the *F-statistic* and its corresponding p-value.

- If *p-value* < 0.05, reject H_0. Otherwise, all of the slopes equal 0 and none of the predictors are useful in predicting the response.

Deciding on Important Variables

- *Best Subset Selection*: we compute the OLS fit for all possible subsets of predictors and then choose between them based on some criterion that balances training error with model size.

- There are 2^p possible models, so can't examine them all.

- We use an automated approach that searches through a subset of all the models.
  - Forward Selection
  - Backward Selection

Overview: Forward Selection

- We begin with the *null model*, a model containing an intercept but no predictors.
- We fit *p* simple linear regressions and add to the null model that variable resulting in the lowest RSS.
- We add to that model the variable that results in the lowest RSS amongst all two-variable models.
- The algorithm continues until some stopping rule is satisfied (i.e. all remaining variables have a *p-value* greater than some threshold).

Overview: Backward Selection

- We begin with all variables in the model.
- We remove the variable with the largest *p-value* (i.e. least statistically significant).
- The new (p – 1)-variable model is fit, and the variable with the largest *p-value* is removed.
- The algorithm continues until a stopping rule is reached.

Qualitative Predictors

- Some predictors are not quantitative but are *qualitative*, taking a discrete set of values.
- These are known as categorical variables, which we can code as indicator variables (dummy variables).
- Examples: gender, student status, marital status, ethnicity

Qualitative Predictors (cont.)

- When a qualitative predictor has more than two levels, a single dummy variable cannot represent all possible variables.

- Thus, there will always be one fewer dummy variables than the number of levels in the factor.
  - Factor = Ethnicity
  - Levels = Asian, Caucasian, African American
  - # of Dummy Variables = 3 – 1 = 2

- The level with no dummy variable is the *baseline*.

Qualitative Predictors (cont.)

- Suppose we want to regress the quantitative response variable Y (such as balance) on both a quantitative variable (such as income) and a qualitative variable (such as gender).
- There are two levels of gender: $Gender_i = \begin{cases} 1 \text{ if female} \\ 0 \text{ if male} \end{cases}$
- The regression model (without interaction) is:

$$Balance_i \approx \beta_0 + \beta_1 Income_i + \beta_2 Gender_i$$
$$= \begin{cases} \beta_0 + \beta_1 Income_i + \beta_2 & \text{if female} \\ \beta_0 + \beta_1 Income_i & \text{if male} \end{cases}$$

- β_2 is the average extra balance that females have for a given income level; note that males are *baseline* (coded as 0).

Extensions of the Linear Model

- Allow for *interaction effects.* Note that if interaction is included in the model, all of the *main effects* should be include as well (even if not statistically significant).

$$\begin{aligned} \texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon. \end{aligned}$$

- Accommodate non-linear relationships using *polynomial regression*. For example, you can include transformed versions of the predictors in the model.

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

Residual Tests

- Residuals are supposed to:
  - be normally distributed (with zero mean)
  - have a constant variance
  - be uncorrelated (if it is a time series)
- There are many formal tests, but here are a few simple ones.

Simple Diagnostic Tests

Check residuals for outliers

- Standardized residual > $\pm$ 3 in list of residuals
- Unusual points on plot of Y_{actual} and Y_{fitted}

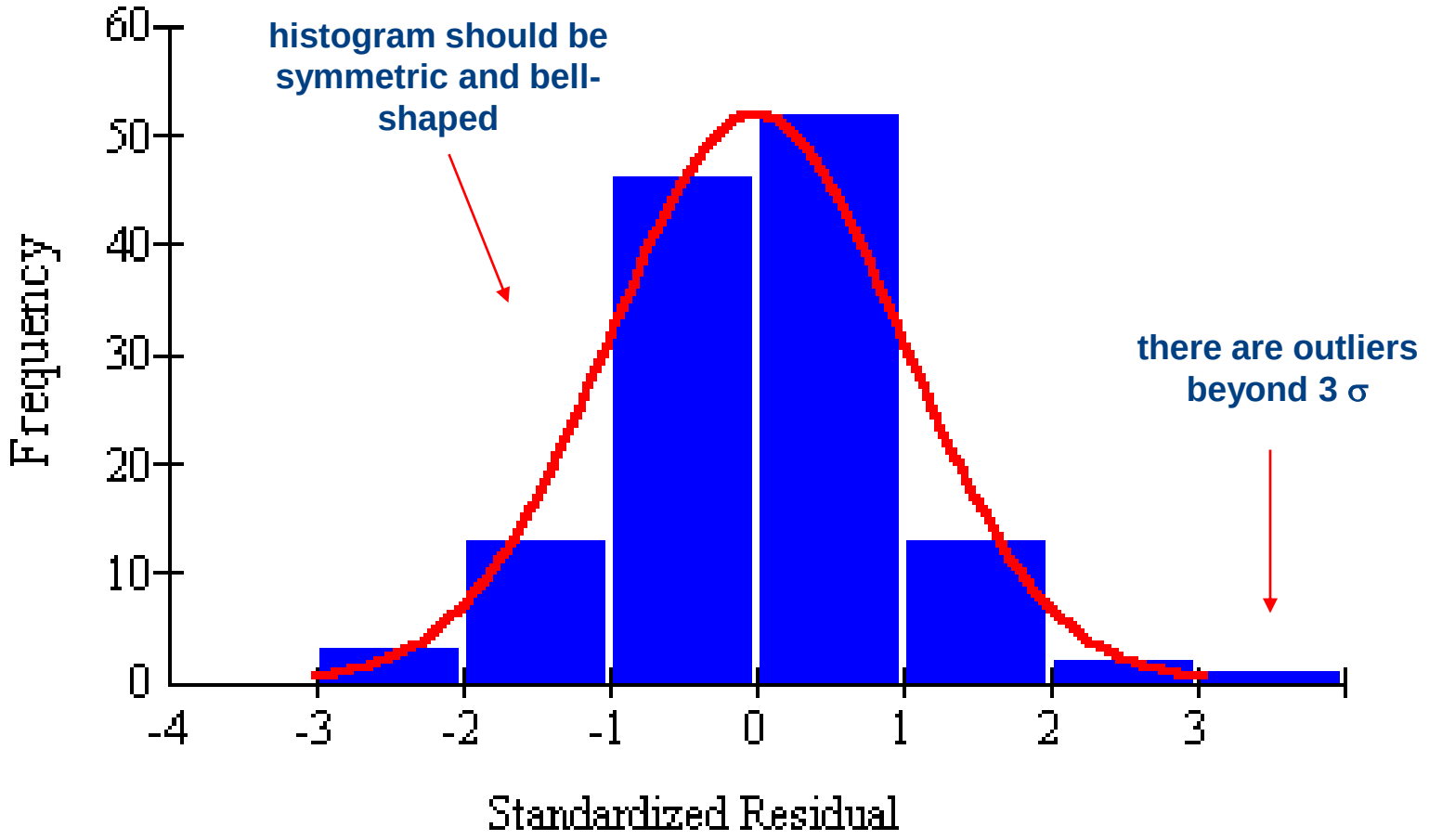
Check residuals for normality

- Histogram should be bell-shaped
- Probability plots should be linear

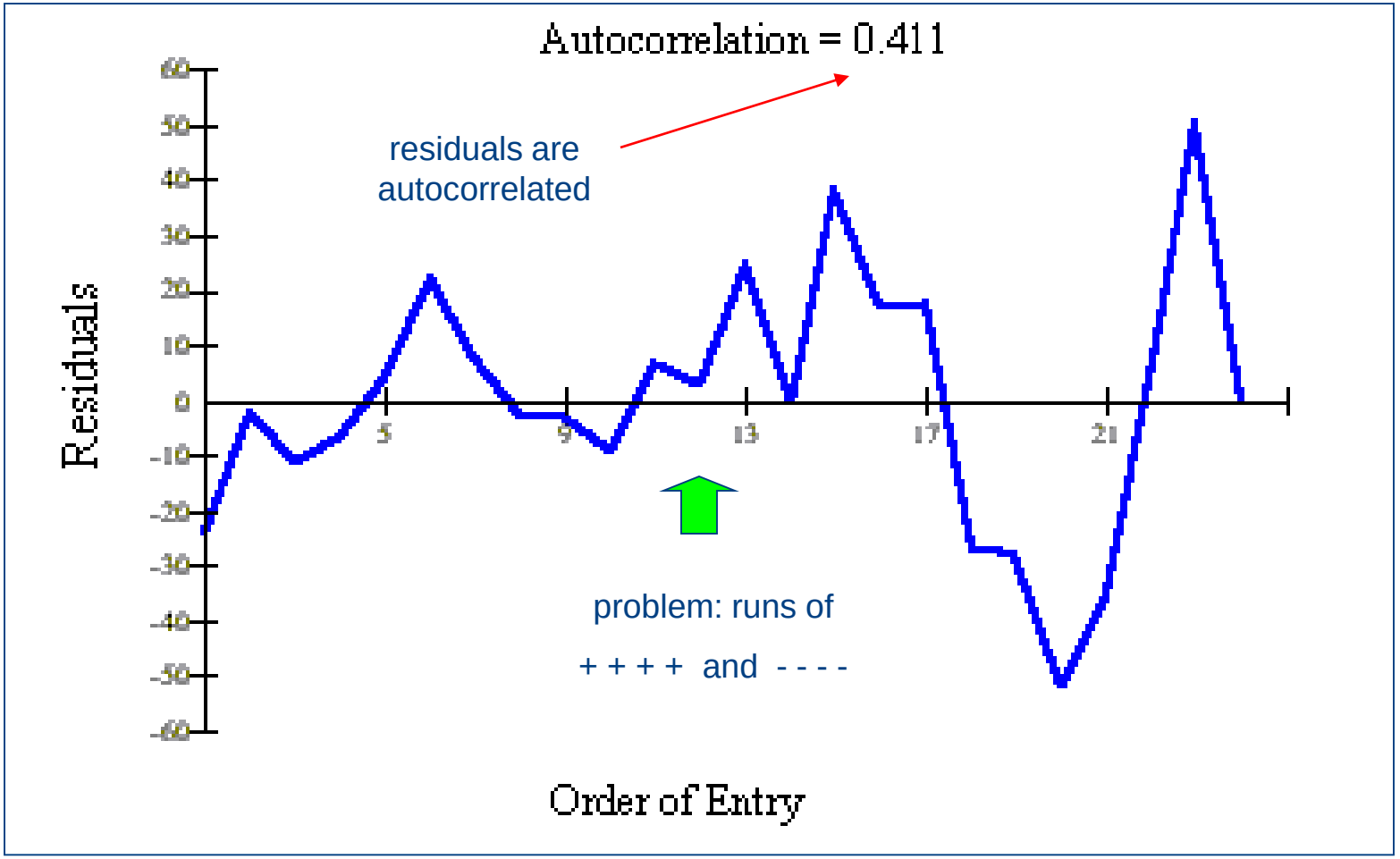
Plot residuals over time (if time series data)

- Look for patterns (cycles or oscillation)
- Durbin-Watson statistic (DW $\cong$ 2 is ideal)

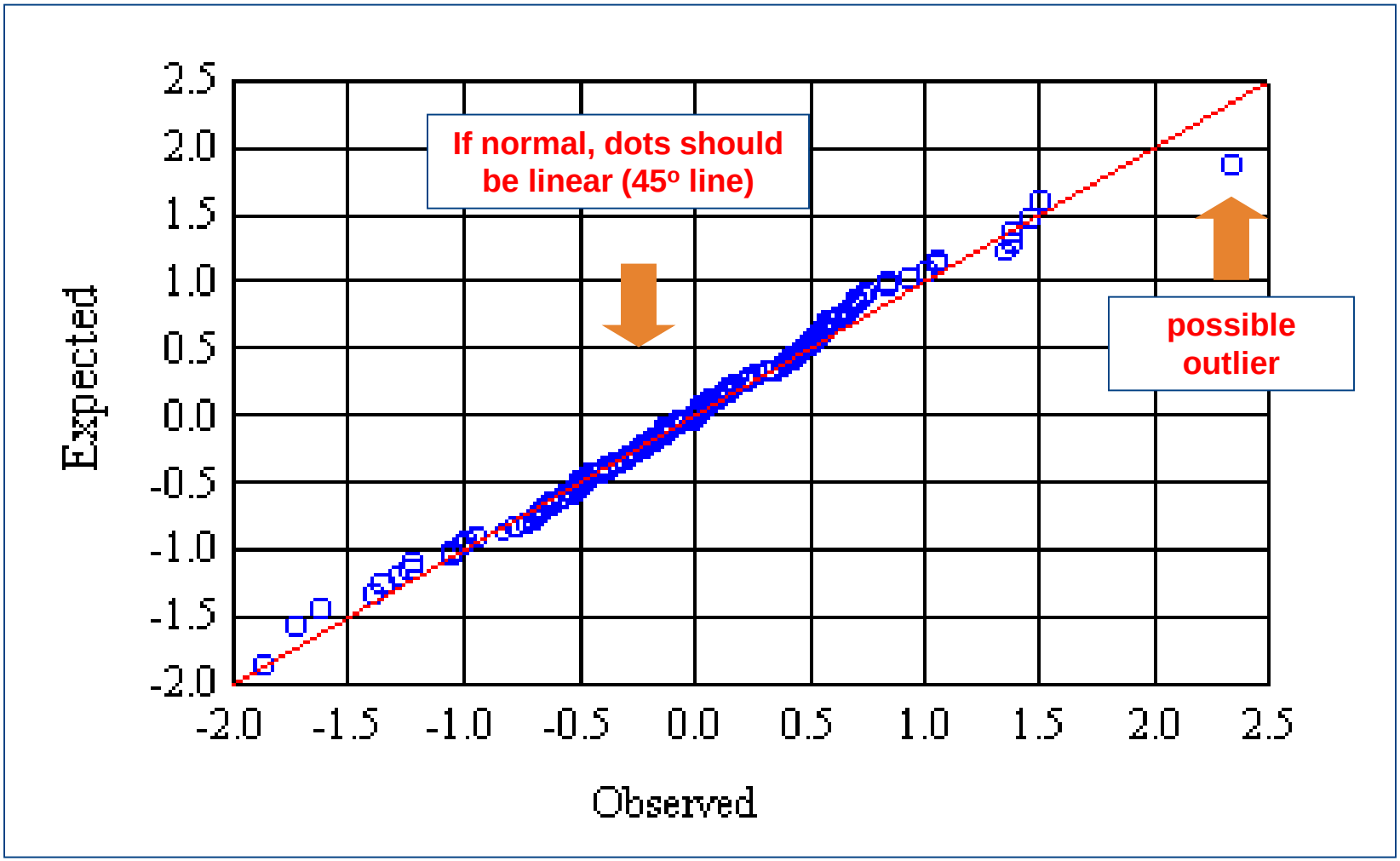
Residual Histogram



Residual Time Plot



Residual Probability Plot



Nonlinearity Tests

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 + \varepsilon$$

If we can reject $\beta_2 = 0$ it would suggest a non-linear relationship between Y and X_1

If we can reject $\beta_4 = 0$ it would suggest a non-linear relationship between Y and X_2

Pro

- **Adding X^2 terms allows curvilinear relationships**
- **Can still be estimated by OLS**
- **Significant X^2 terms would suggest nonlinearity**

Con

- **More predictors causes loss of d.f.**
- **Introduces collinearity between X and X^2 terms (high VIF)**

Tests for Interaction

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon_i$$

If we can reject $\beta_3 = 0$ there is a significant interaction effect

Pro

- **Detects interaction between any two predictors**
- **Multiple interactions are possible (e.g., $X_1X_2X_3$)**

Con

- **Becomes complex if many predictors**
- **Difficult to interpret the coefficient**

Regression Assumptions

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i$$

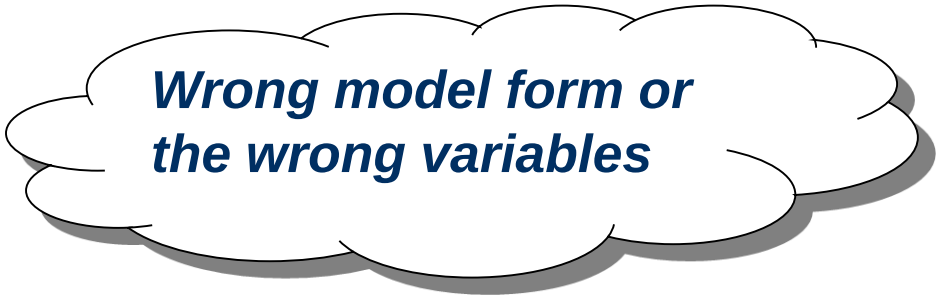
- Correct model specified (no variables omitted)
- Appropriate model form (e.g., linear)
- Predictors are non-stochastic and independent
- Errors (disturbances) are random $N(0,\sigma^2)$
  - zero mean
  - normally distributed
  - homoscedastic (constant variance)
  - mutually independent (non-autocorrelated)

Potential Problems

There are several potential problems that may occur when fitting a linear regression model, particularly when the six standard OLS assumptions are violated.

1. **Relevant predictors were omitted**
2. **Wrong model form specified (e.g., linear)**
3. **Collinear predictors (i.e., correlated Xj and Xk)**
4. **Non-normal errors (e.g., skewed, outliers)**
5. **Heteroscedastic errors (non-constant variance)**
6. **Autocorrelated errors (non-independent)**

What Is Specification Bias?



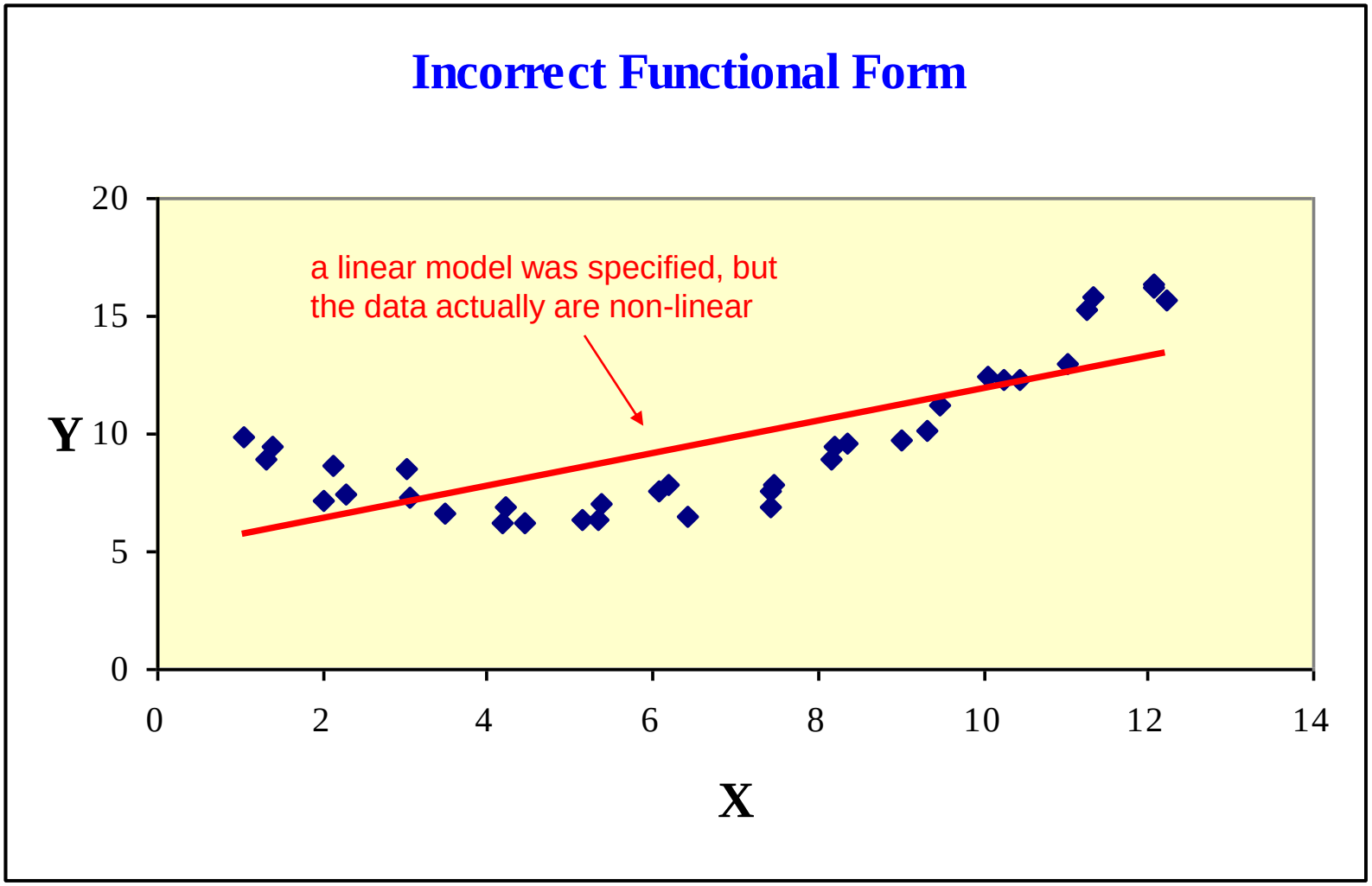
Example of Wrong Model Form:

You said $Y = a + bX$, but actually $Y = a + bX + cX^2$

Example of Wrong Variables:

You said $Y = a + bX$ but actually $Y = a + bX + cZ$

Specification Bias



Detecting Specification Bias

In a bivariate model:

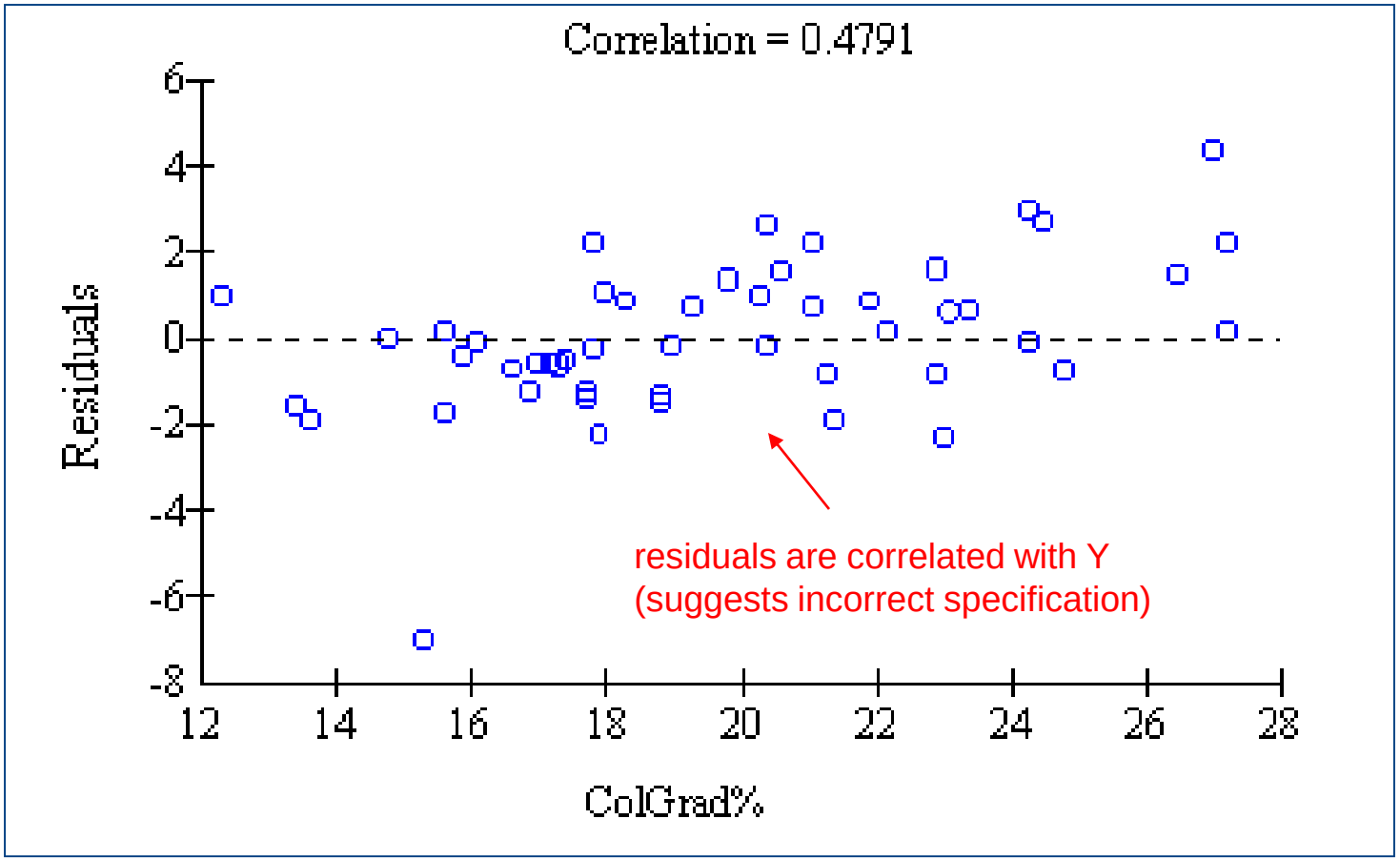
- **Plot *Y* against *X***
- **Plot residuals against estimated *Y***

In a multivariate model:

- **Plot residuals against estimated *Y***
- **Plot residuals against actual *Y***
- **Plot fitted *Y* against actual *Y***

Look for patterns (there should be none)

Residuals Plotted on Y



Non-linearity of the Data

- The linear regression model assumes that there is a straight-line relationship between predictors and the response.
- If the true relationship is non-linear then conclusions are suspect.
- Examine the *residual plots*, as strong patterns (U-shape) in the residuals indicate non-linearity in the data.
- If there are non-linear associations in the data, then use non-linear transformations of the predictors (e.g., log X), which can be found optimally using Box-Cox transformations.

What Is Multicollinearity?

Collinearity:

Correlation between any two predictors

Multicollinearity:

Relationship among several predictors

Effects of Multicollinearity

- Estimates may be unstable
- Standard errors may be misleading
- Confidence intervals generally too wide
- High R^2 yet t statistics insignificant

Variance Inflation Factor

VIFs give a simple multicollinearity test. Each predictor has a VIF. For predictor *j*, the VIF is

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination when predictor *j* is regressed against all the other predictors.

Variance Inflation Factor (cont.)

Example A: If R_j^2 =.00 then VIF_j = 1:

$$VIF_j = \frac{1}{1 - R_j^2} = \frac{1}{1 - 0} = 1$$

Example B: If R_j^2 = .90 then VIF_j = 10:

$$VIF_j = \frac{1}{1 - R_j^2} = \frac{1}{1 - .90} = 10$$

Evidence of Multicollinearity

- **Any VIF > 10**
- **High correlation for pairs of predictors X_j and X_k**
- **Unstable estimates**
  (i.e., the remaining coefficients change sharply when a suspect predictor is dropped from the model)

Example: Estimating Body Fat

The regression equation is

Fat%1 = 18.6 + 0.0685 Age - 0.197 Height - 0.765 Neck - 0.051 Chest
+ 0.943 Abdomen - 0.731 Hip + 0.530 Thigh

Predictor	Coef	StDev	T	P	VIF
Constant	18.63	12.44	1.50	0.14	
Age	0.06845	0.09268	0.74	0.46	1.7
Height	-0.197	0.1087	-1.81	0.08	1.3
Neck	-0.765	0.3836	-1.99	0.05	4.4
Chest	-0.0514	0.1865	-0.28	0.78	10.9
Abdomen	0.9426	0.1731	5.45	0.00	17.6
Hip	-0.7309	0.2281	-3.20	0.00	15.9
Thigh	0.5299	0.2886	1.84	0.07	10.5

S = 4.188 R-Sq = 81.8% R-Sq(adj) = 78.7%

Problem: Several VIFs exceed 10.

Correlation Matrix of Predictors

	Age	Height	Neck	Chest	Abdomen	Hip
Height	-0.276					
Neck	0.176	0.201				
Chest	0.376	0.014	0.820			
Abdomen	0.442	-0.052	0.781	0.942		
Hip	0.314	-0.045	0.804	0.911	0.942	
Thigh	0.219	-0.037	0.823	0.859	0.890	0.938

Age and *Height* are relatively independent of other predictors.

Problem: *Neck*, *Chest*, *Abdomen*, *Hip*, and *Thigh* are highly correlated.

Solution: Eliminate Some Predictors

The regression equation is

Fat%1 = 0.8 + 0.0927 Age - 0.184 Height - 0.842 Neck + 0.637 Abdomen

Predictor	Coef	StDev	T	P	VIF
Constant	0.79	10.35	0.08	0.94	
Age	0.0927	0.09199	1.01	0.32	1.4
Height	-0.1837	0.1133	-1.62	0.11	1.2
Neck	-0.8418	0.3516	-2.39	0.02	3.2
Abdomen	0.63659	0.0846	7.52	0.00	3.6

S = 4.542 R-Sq = 77.0% R-Sq(adj) = 75.0%

R^2 is reduced slightly, but all VIFs are now below 10.

Stability Check for Coefficients

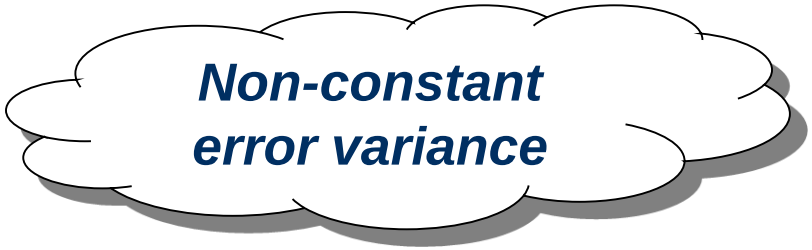
Variable	Run 1	Run 2	Run 3	Run 4	%Chg
Constant	18.63	17.67	19.89	0.79	-95.8%
Age	0.06845	0.0689	0.0200	0.0927	35.4%
Height	-0.1970	-0.1978	-0.2387	-0.1837	-6.8%
Neck	-0.7650	-0.8012	-0.5717	-0.8418	10.0%
Chest	-0.0514				
Abdomen	0.9426	0.9158	0.9554	0.6366	-32.5%
Hip	-0.7309	-0.7408	-0.5141		
Thigh	0.5299	0.5406			
Std Err	4.188	4.143	4.266	4.542	8.5%
R-Sq	81.8%	81.7%	80.2%	77.0%	-5.9%
R-Sq(adj)	78.7%	79.2%	77.9%	75.0%	-4.7%

There are large changes in estimated coefficients as high VIF predictors are eliminated, revealing that the original estimates were unstable. But the "fit" deteriorates when we eliminate predictors.

Remedies for Multicollinearity

- Drop one or more predictors
- But this may create specification error
- Transform some variables (e.g., log *X*)
- Enlarge the sample size (if you can)
- Perform Factor Analysis and replace previous correlated predictors with "latent" variable.
- Perform principal components regression or partial least squares regression

What Is Heteroscedasticity?



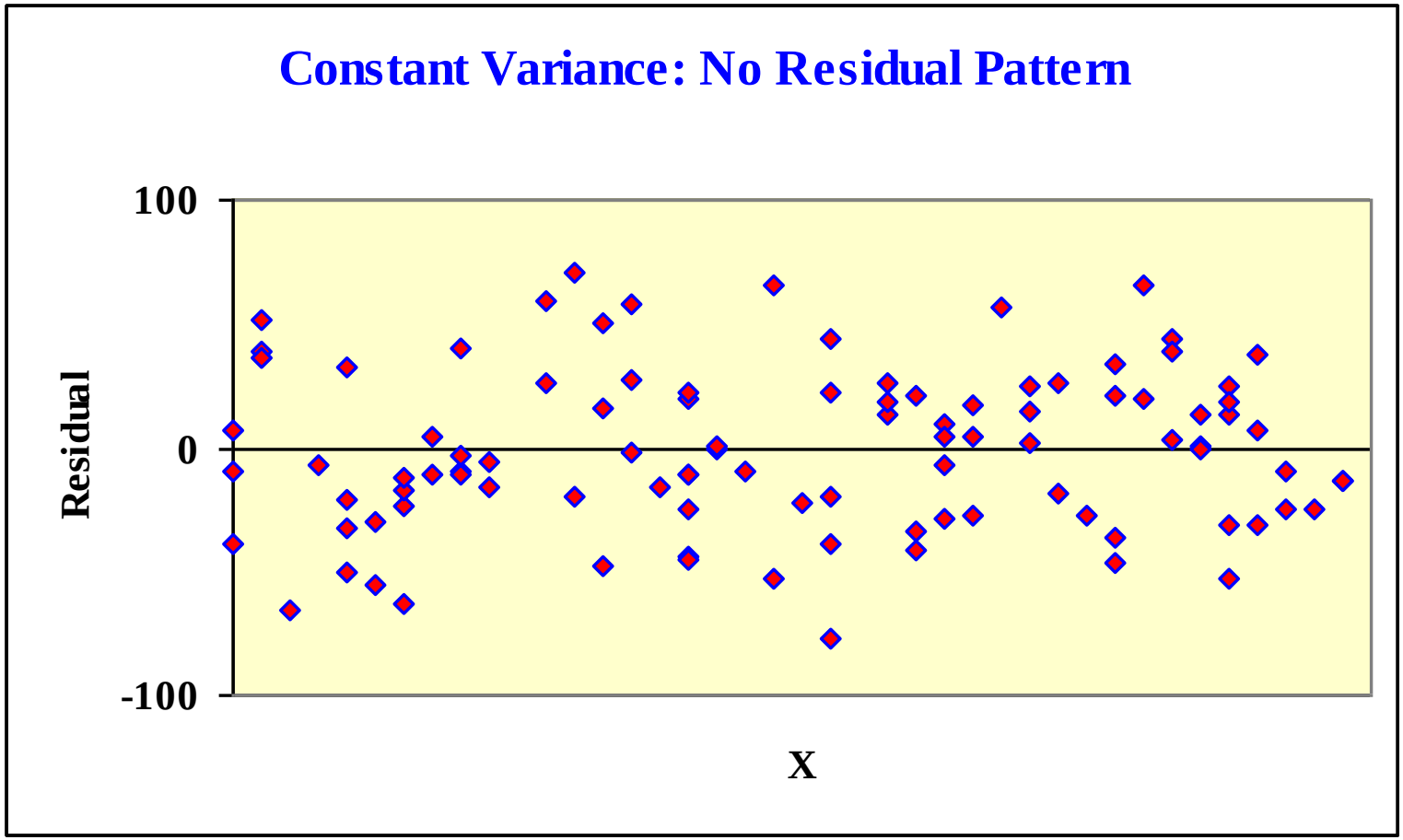
Homoscedastic:
Errors have the same variance for all values of the predictors (or *Y*)

Heteroscedastic:
Error variance changes with the values of the predictors (or *Y*)

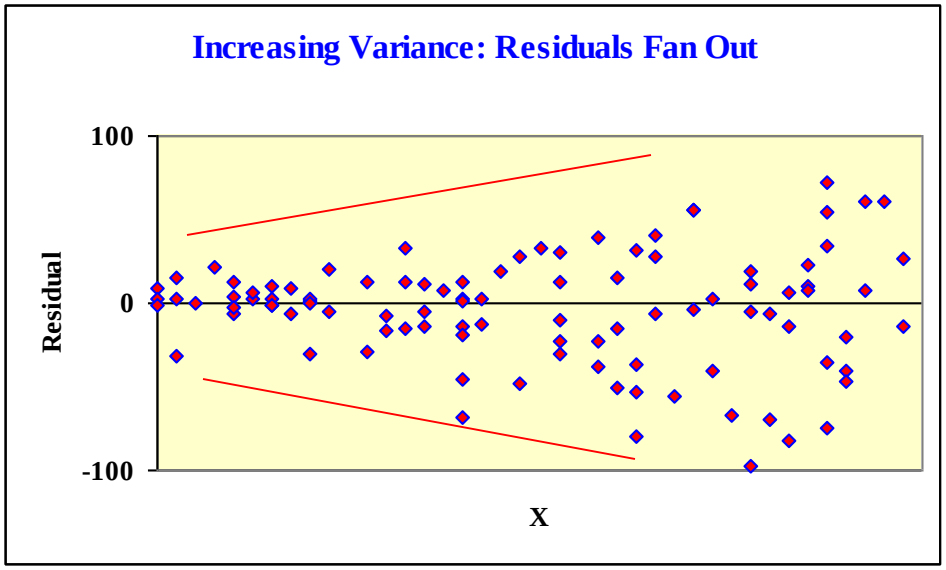
How to Detect Heteroscedasticity

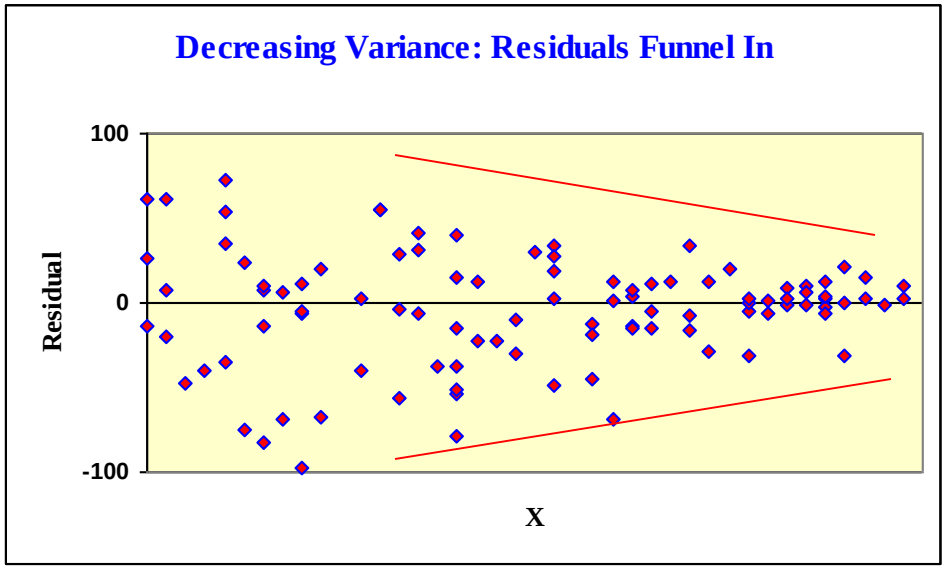
- Plot residuals against each predictor (a bit tedious).
- Plot residuals against estimated *Y* (quick check).
- There are more general tests, but they are complex (Breusch-Pagan, Modified Levene, or Special White's Test).

Homoscedastic Residuals



Heteroscedastic Residuals





To detect heteroscedasticity, we plot the residuals against each predictor. Some predictors may show a problem, while others are O.K. A quick overall test is to plot the residuals only against estimated *Y*.

Effects of Heteroscedasticity

Happily ...

- OLS coefficients b_j are still unbiased
- OLS coefficients b_j are still consistent

But ...

- Std errors of *b*'s are biased (bias may be + or -)
- *t* values and CI for *b*'s may be unreliable
- May indicate incorrect model specification

Remedies for Heteroscedasticity

1. Transform the response variable Y.

2. Adjust the standard errors, *t*, *F* (use robust SEs).

3. Perform Feasible Generalized Least Squares (FGLS), particularly using weighted least squares estimation.

What Is Autocorrelation?

Independent errors:

e_t does not depend on e_{t-1} ($\rho = 0$)

Autocorrelated errors:

e_t depends on e_{t-1} ($\rho \neq 0$)

Good News! **Autocorrelation is a worry in time-series models (the subscript $t = 1, 2, ..., n$ denotes time) but generally not in cross-sectional data.**

What Is Autocorrelation?

A Simple Model: $e_t = \rho\, e_{t-1} + u_t$

where u_t is assumed non-autocorrelated (white noise)

Independent errors:

e_t does not depend on e_{t-1} $(\rho = 0)$

Autocorrelated errors:

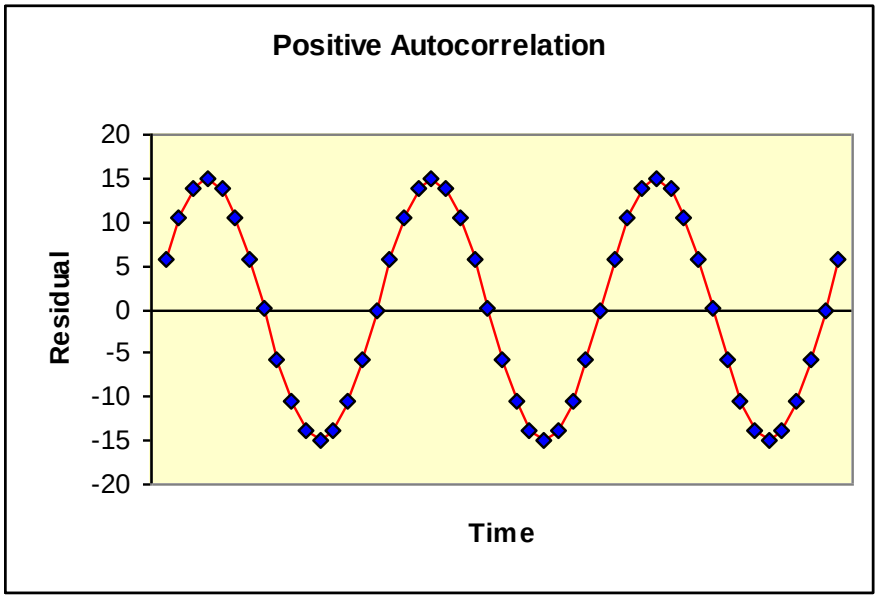
e_t depends on e_{t-1} $(\rho \neq 0)$

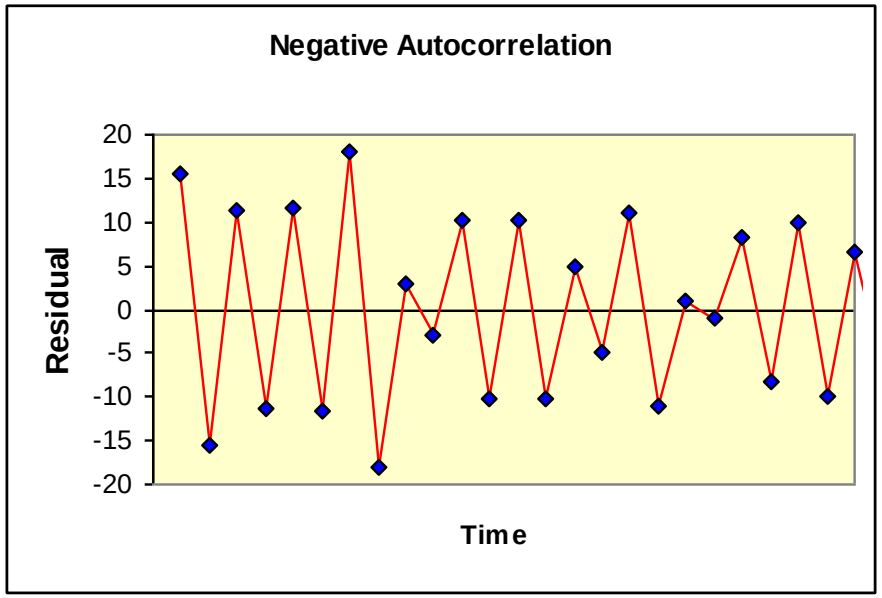
$\therefore$ ***The residuals will show a pattern over time***

Autocorrelated Residuals

Common

Rare





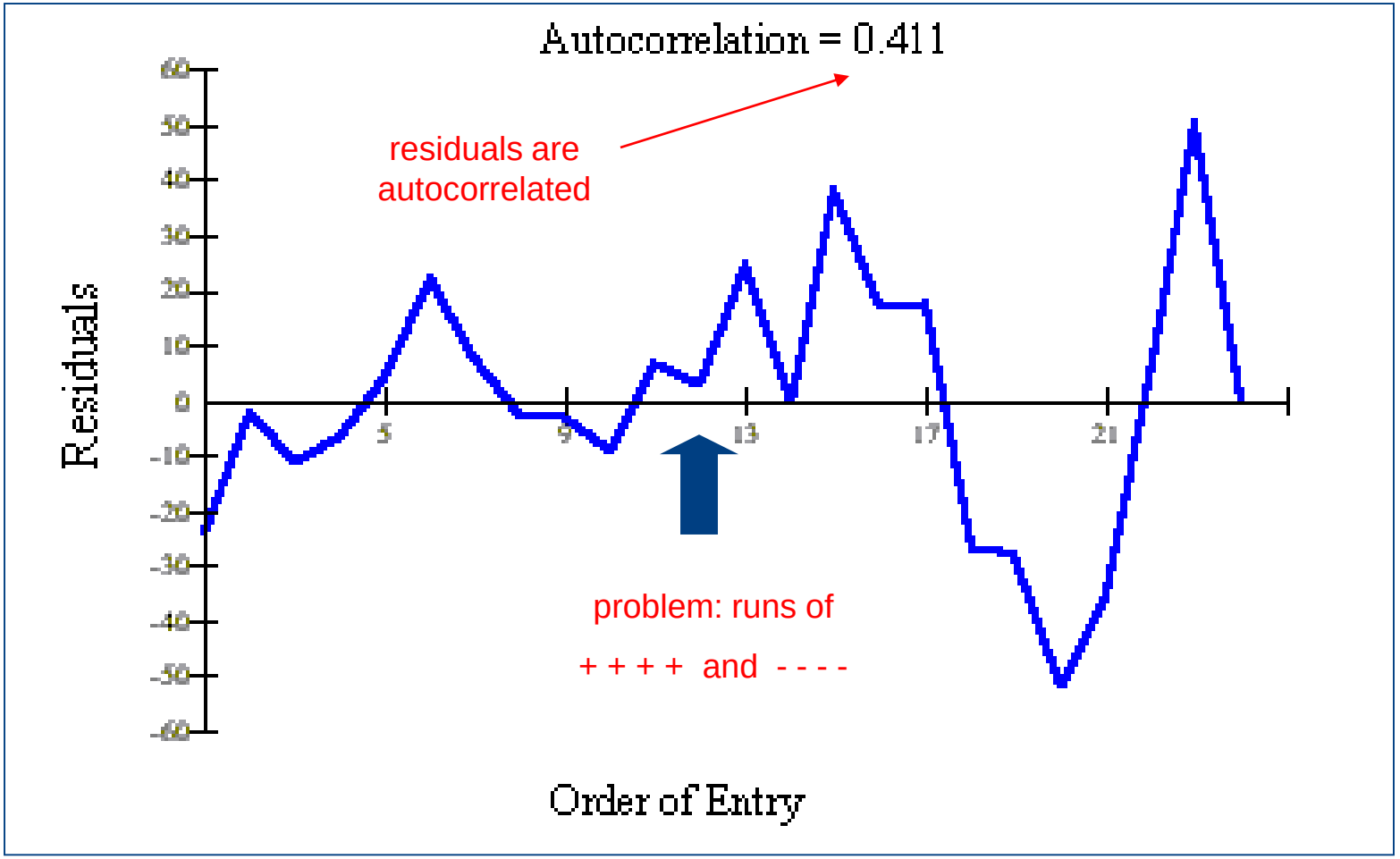
When a residual tends to be followed by another of the same sign, we have positive autocorrelation.

When a residual tends to be followed by another of opposite sign, we have negative autocorrelation.

How to Detect Autocorrelation

- Look for pattern in residuals plotted against time
  - ✓ Look for cycles of + + + + followed by - - - -
  - ✓ Look for alternating + - + - pattern
- Calculate the correlation between e_t and e_{t-1}
  - ✓ This is called the "autocorrelation coefficient"
  - ✓ It should not differ significantly from 0
- Check Durbin-Watson statistic
  - ✓ DW = 2 suggests no autocorrelation
  - ✓ DW < 2 suggests positive autocorrelation (common)
  - ✓ DW > 2 suggests negative autocorrelation (rare)

Residual Time Plot



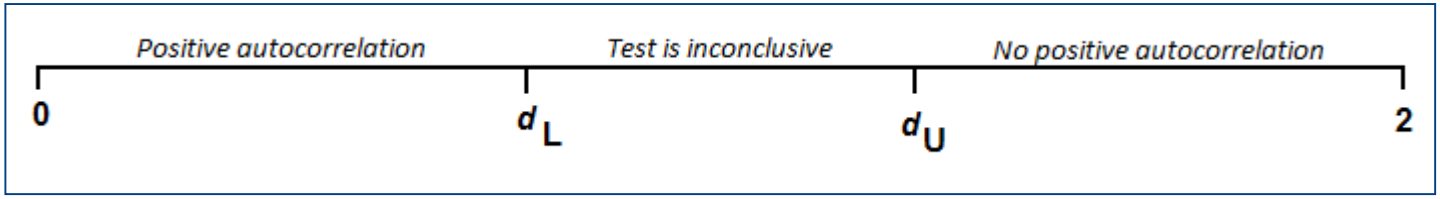
Durbin-Watson Test

To test for *positive* autocorrelation (most common) the hypotheses are:

H_0: $\rho = 0$ (errors are not autocorrelated)
H_1: $\rho > 0$ (errors are positively autocorrelated)

The interpretation of the *DW* test for positive autocorrelation is expressed in words and visually (a table is required for d_L and d_U):

If $DW < d_L$ conclude H_1 (errors are positively autocorrelated)
If $DW > d_U$ conclude H_0 (errors are not positively autocorrelated)
If $d_L \leq DW \leq d_U$ the test is inconclusive



Common Types of Autocorrelation

- **Errors are autocorrelated (relatively minor)**

First Order Autocorrelation

$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \varepsilon_t$ where

$\varepsilon_t = \rho e_{t-1} + u_t$ and u_t is $N(0,\sigma^2)$

- **Lagged Y used as predictor (OK if large *n*)**

Lagged Predictor

$Y_t = \beta_0 + \beta_1 X_{t-1} + \beta_2 Y_{t-1} + \varepsilon_t$

Effects of Simple First-Order Autocorrelation

- OLS coefficients b_j are still unbiased
- OLS coefficients b_j are still consistent
- If r > 0 (the typical situation) then the
  - ✓ *standard errors of b_j is underestimated*
  - ✓ *computed t values will be too high*
  - ✓ *CI and PI for b_j will be too narrow*

Data Transformations for Autocorrelation

- **Use first differences: $\Delta Y = f(\Delta X_1, \Delta X_2)$**

$$\Delta Y_t = \gamma_0 + \beta_1 \Delta X_1 + \beta_2 \Delta X_2 + \varepsilon_t$$

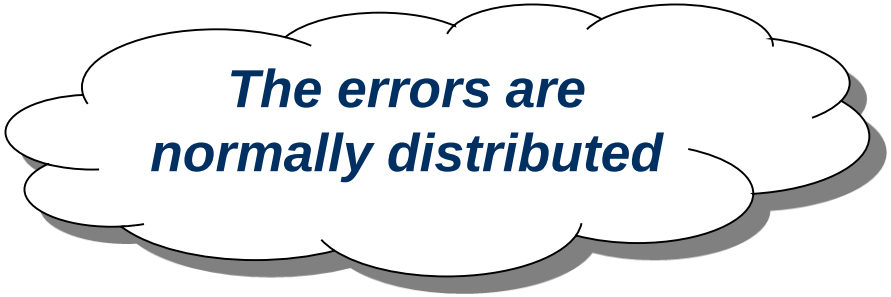
Comment: **Simple, but only suffices when ρ is near 1.**

- **Use Cochrane-Orcutt transformation**

$$Y_t^* = Y_t - \rho Y_{t-1}$$
$$X_t^* = X_t - \rho X_{t-1}$$

Comment: **We must estimate the sample autocorrelation coefficient and use it to estimate ρ.**

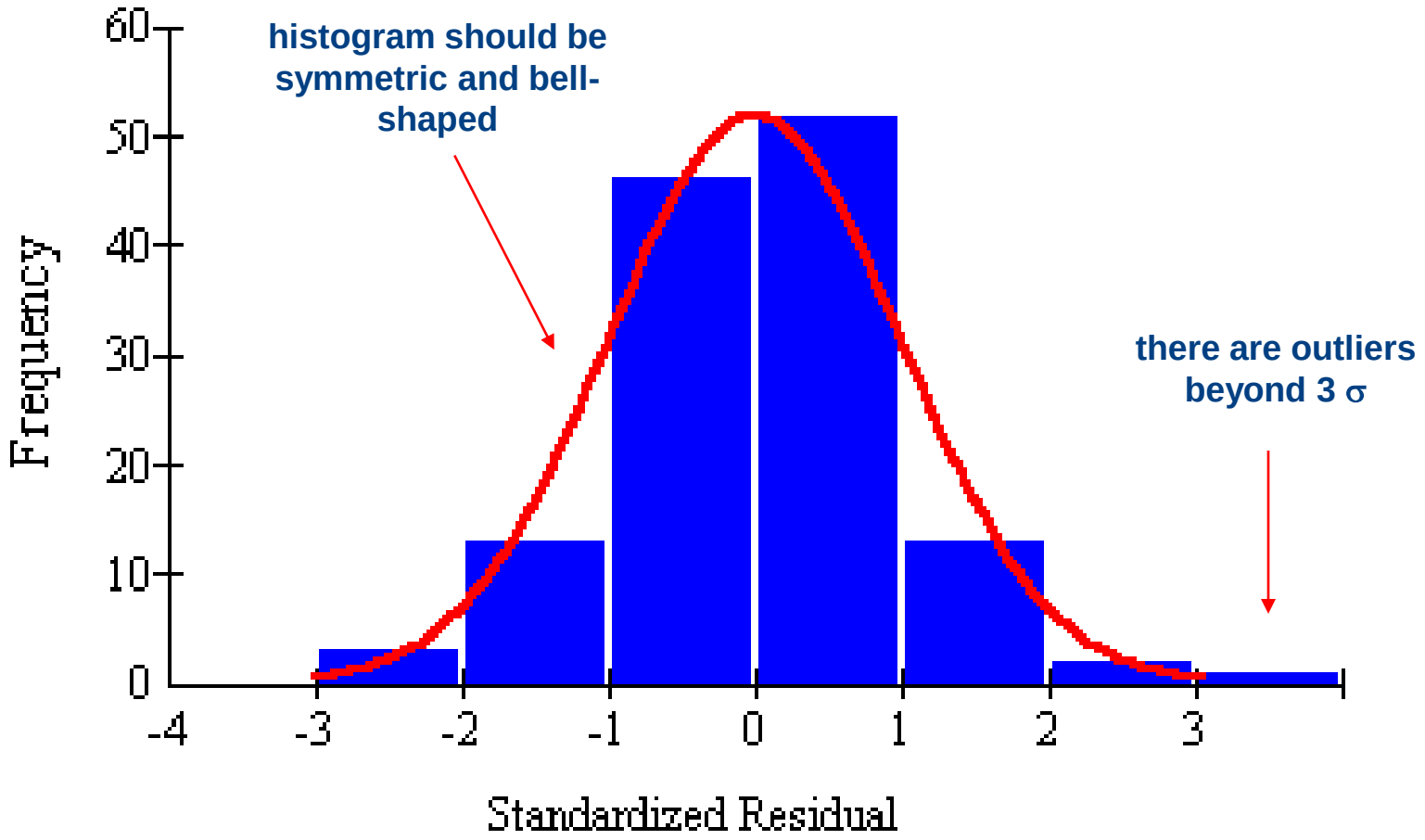
What Is Non-Normality?



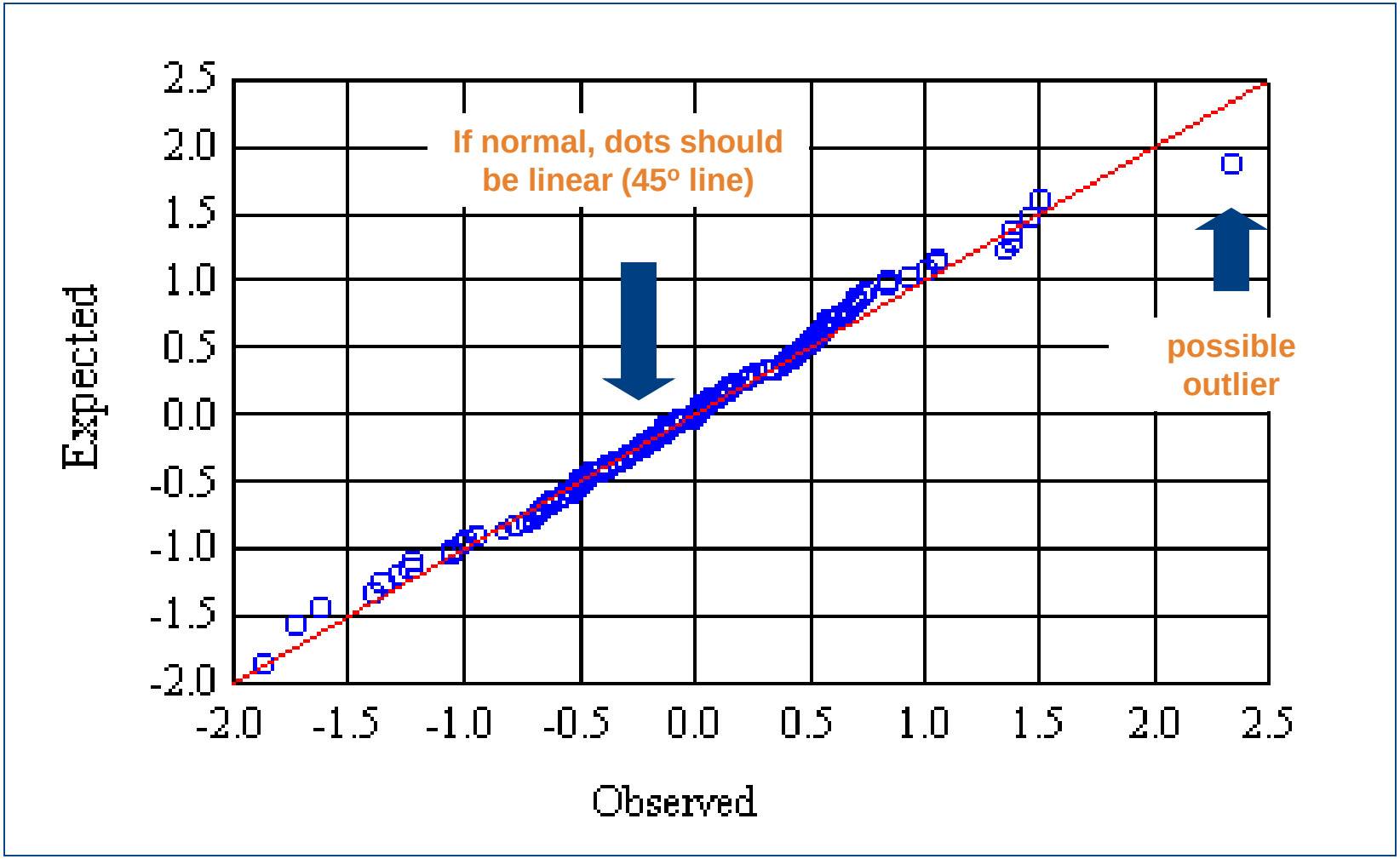
- **Normal errors:**
  - ✓ The histogram of residuals is "bell-shaped"
  - ✓ There are no outliers in the residuals
  - ✓ The probability plot is linear

- **Non-normal errors**
  - ✓ Any violations of the above

Residual Histogram



Residual Probability Plot



Effects of Non-Normal Errors

- Confidence intervals for *Y* may be incorrect
- May indicate outliers
- May indicate incorrect model specification

Detection of Non-Normal Errors

- Look at histogram of residuals
  - ✓ Should be symmetric
  - ✓ Should be bell-shaped
- Look for outliers or asymmetry
  - ✓ Outliers are a serious violation
  - ✓ Mild asymmetry is common
- Look at probability plot of residuals
  - ✓ Should be linear
  - ✓ Look for outliers
- Anderson-Darling Test

Remedies for Non-Normal Errors

- **Avoid aggregated data**
- **Transform the response/predictor variables (Box-Cox)**
- **Enlarge the sample (asymptotic normality)**