Chapter 2: Introduction to Regression Analysis

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

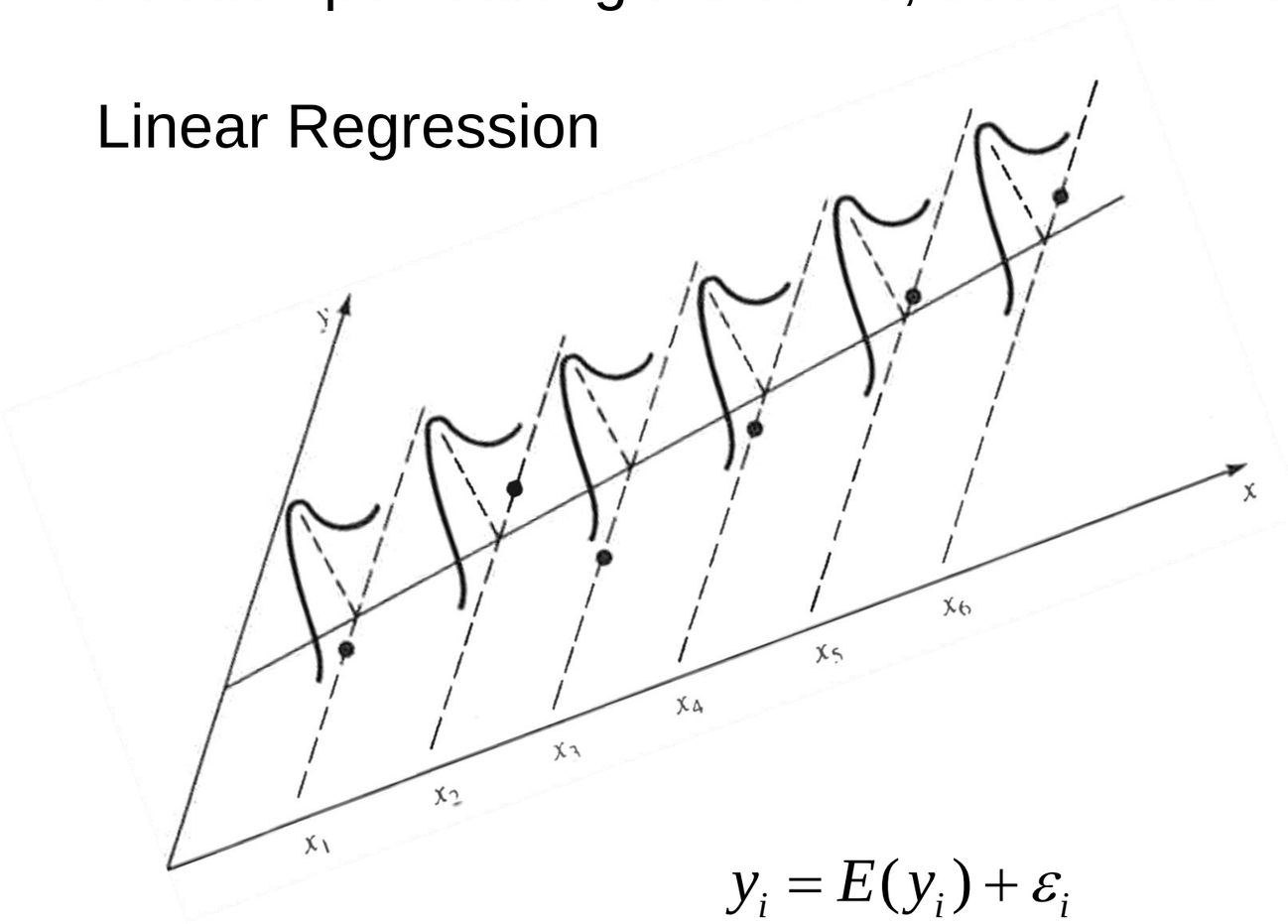
Introduction to Regression Analysis

Modeling a Response

$$\varepsilon \sim N(0, \sigma^2)$$

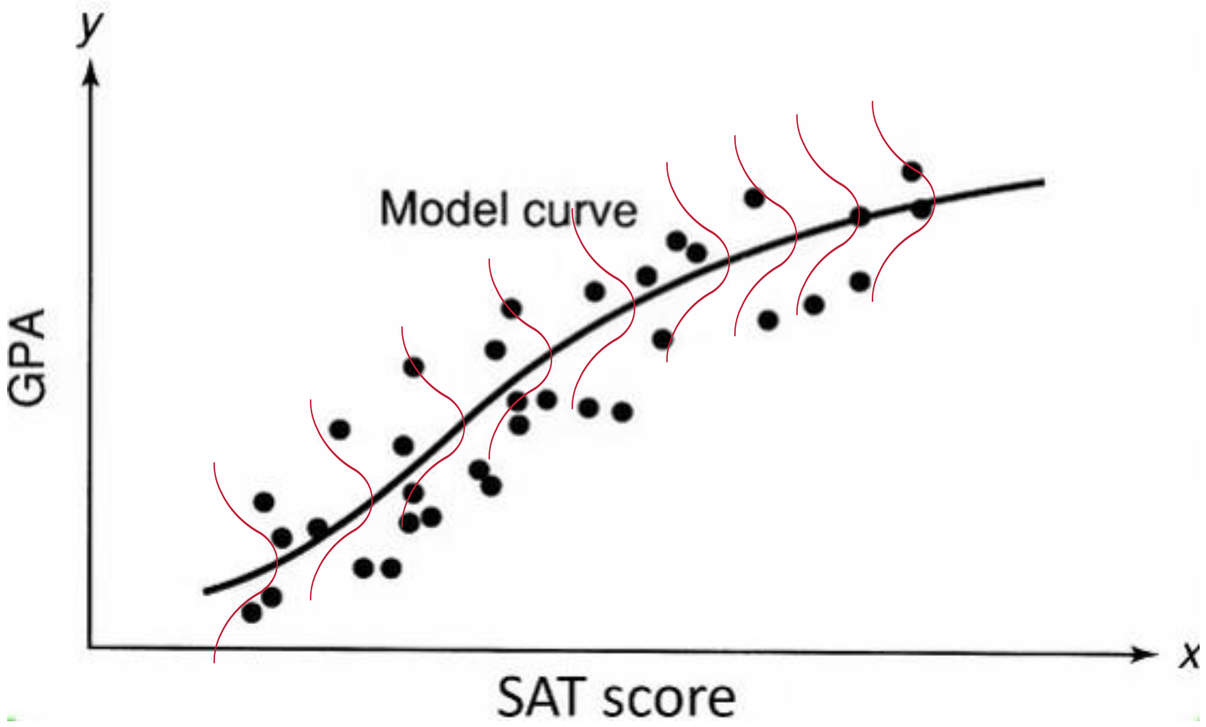
At each point along the curve, observations have additive normal error ε.

Linear Regression



$$y_i = E(y_i) + \varepsilon_i$$

Non-Linear Regression



Introduction to Regression Analysis

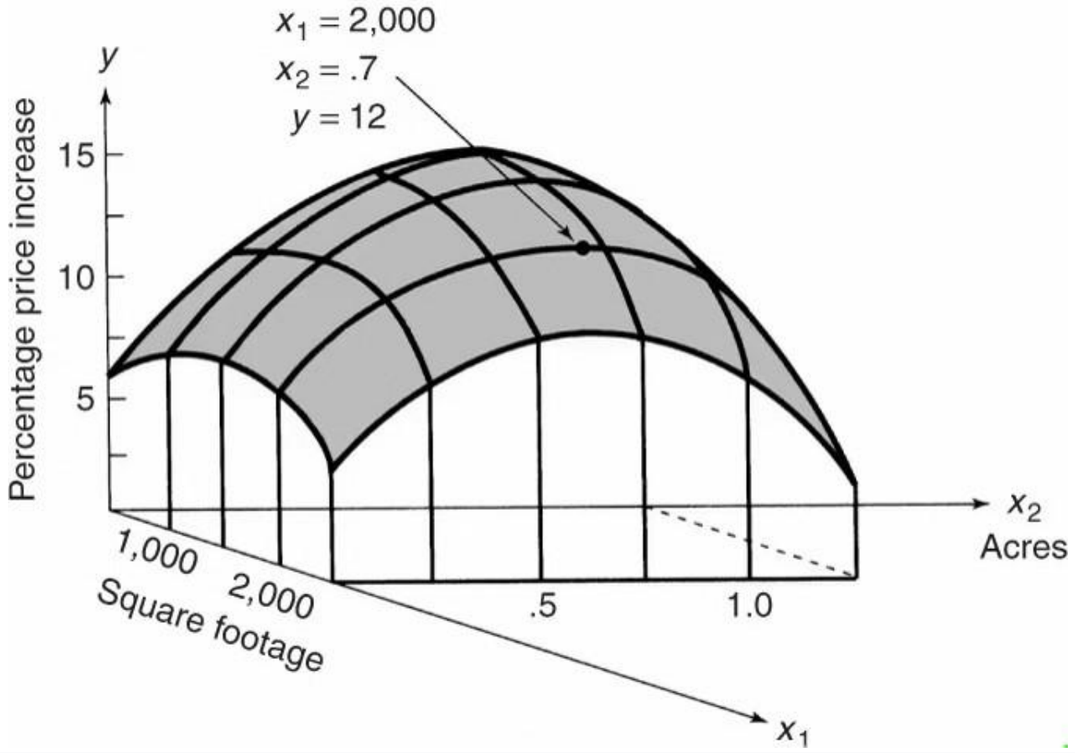
Overview of Regression Analysis

Example: A property appraiser might like to relate percentage price increase y of residential properties to the two quantitative independent variables x_1, square footage of heated space, and x_2, lot size.

This model could be represented by a response surface that traces the mean percentage price increase $E(y \mid x_1, x_2)$ for various combinations of x_1 and x_2.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2 + \hat{\beta}_4 x_1^2 + \hat{\beta}_5 x_2^2$$



Introduction to Regression Analysis

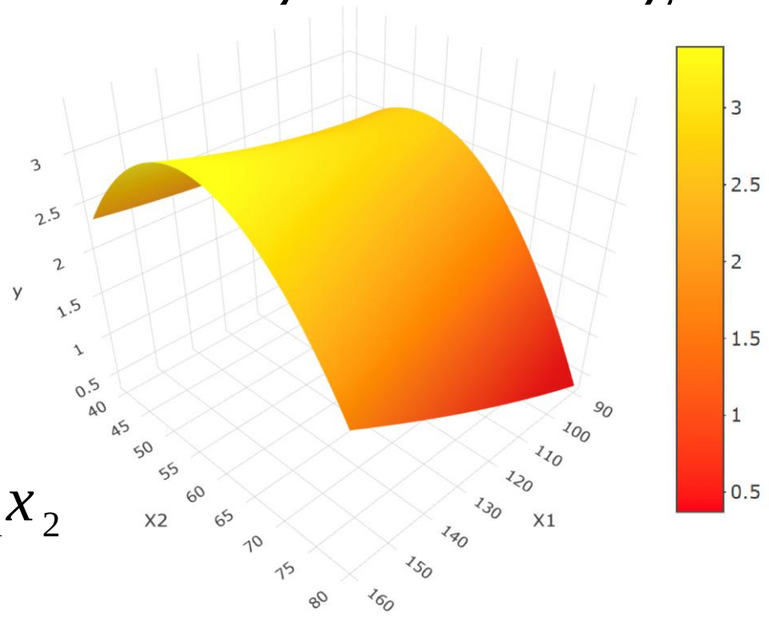
Collecting the Data for Regression

Definition 2.4

If the values of the independent variables (x's) in regression are **controlled** using a designed experiment (i.e., set in advance before the value of y is observed), the data are experimental.

Temperature, x_1	Pressure, x_2	Impurity, y
100	50	2.7
	60	2.4
	70	2.9
125	50	2.6
	60	3.1
	70	3.0
150	50	1.5
	60	1.9
	70	2.2

$$\hat{y} = -7.894 + 0.207x_1 - 0.061x_2$$
$$-0.001x_1^2 + 0.002x_2^2 + 0.005x_1x_2$$

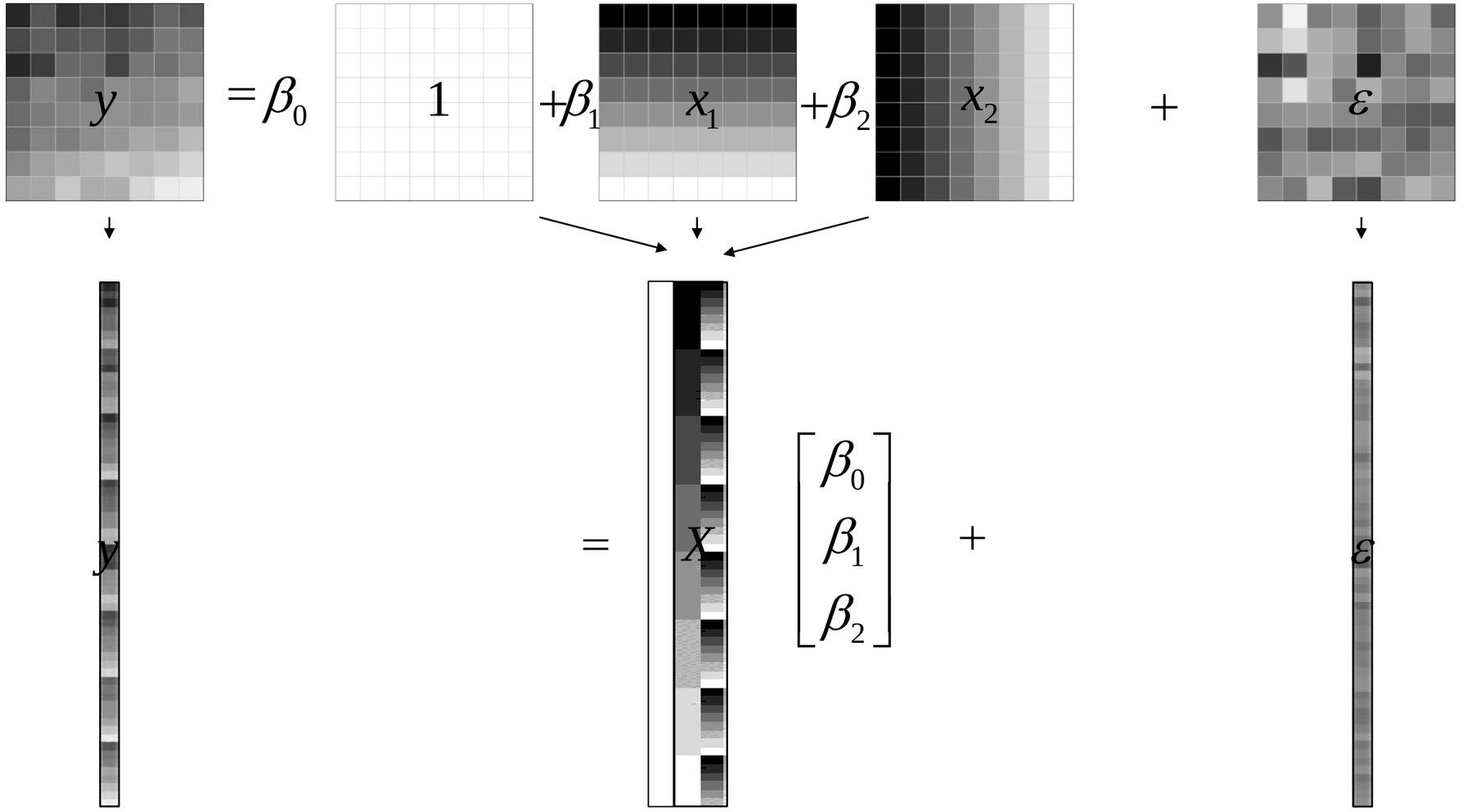


Think of x as dial settings for your science experiment. Every time you fix an x, you run the experiment to get a y. In regression, x is fixed and known.

Introduction to Regression Analysis

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

Example: We can use regression to fit a surface to (x,y) data.

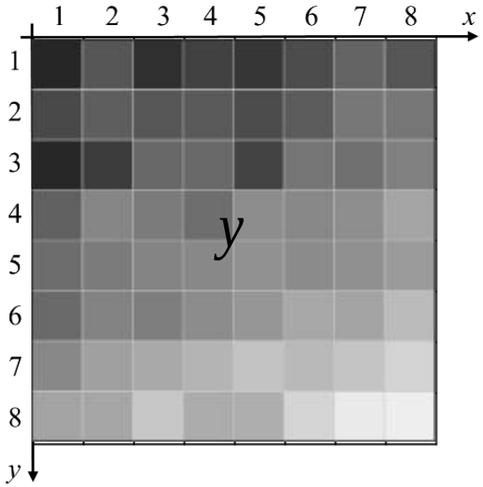


x1	x2	y
-1	-1	86.0753
-1	-0.7143	91.5249
-1	-0.4286	86.1966
-1	-0.1429	95.2958
-1	0.1429	97.0661
-1	0.4286	96.6703
-1	0.7143	101.2757
-1	1	105.6852
-0.7143	-1	93.5854
-0.7143	-0.7143	94.8246
-0.7143	-0.4286	89.4431
-0.7143	-0.1429	101.0698
-0.7143	0.1429	99.308
-0.7143	0.4286	100.5882
-0.7143	0.7143	105.0009
-0.7143	1	106.0186
-0.4286	-1	87.6089
-0.4286	-0.7143	93.6937
-0.4286	-0.4286	96.3895
-0.4286	-0.1429	99.263
-0.4286	0.1429	100.6287
-0.4286	0.4286	99.7279
-0.4286	0.7143	106.4345
-0.4286	1	111.1176
-0.1429	-1	90.2635
-0.1429	-0.7143	94.2122
-0.1429	-0.4286	96.4538
-0.1429	-0.1429	97.2503
-0.1429	0.1429	101.302
-0.1429	0.4286	101.9969
-0.1429	0.7143	108.2054
-0.1429	1	106.9916
0.1429	-1	88.5765
0.1429	-0.7143	91.9524
0.1429	-0.4286	90.54
0.1429	-0.1429	102.1625
0.1429	0.1429	102.7932
0.1429	0.4286	103.4901
0.1429	0.7143	110.5977
0.1429	1	107.2913
0.4286	-1	91.9384
0.4286	-0.7143	94.5171
0.4286	-0.4286	98.4956
0.4286	-0.1429	101.34
0.4286	0.1429	101.8417
0.4286	0.4286	106.3685
0.4286	0.7143	108.956
0.4286	1	113.3983
0.7143	-1	95.758
0.7143	-0.7143	98.6471
0.7143	-0.4286	97.5584
0.7143	-0.1429	102.2976
0.7143	0.1429	102.5718
0.7143	0.4286	105.6301
0.7143	0.7143	110.7006
0.7143	1	116.6367
1	-1	93.4607
1	-0.7143	98.5999
1	-0.4286	100.2631
1	-0.1429	105.8061
1	0.1429	104.2504
1	0.4286	109.3508
1	0.7143	113.2479
1	1	117.2012

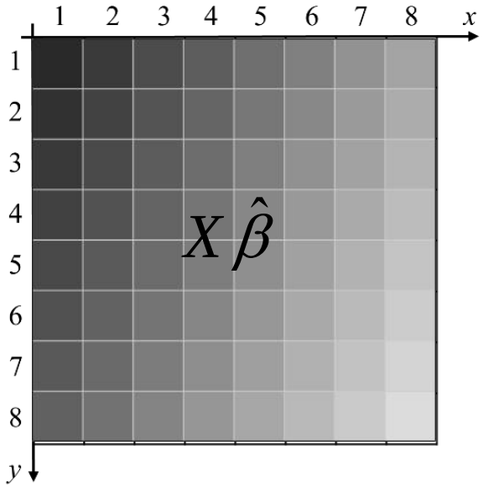
Introduction to Regression Analysis

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

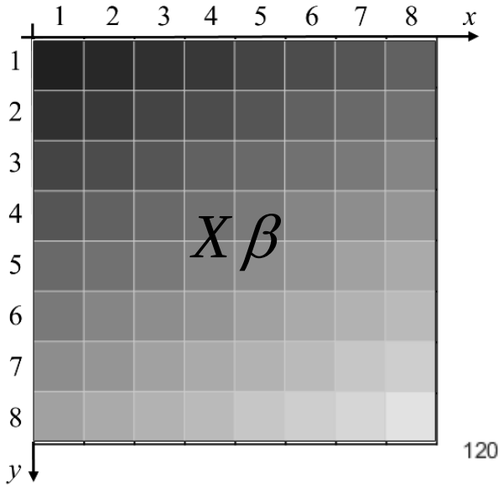
Example: We can use regression to fit a surface to (x,y) data.



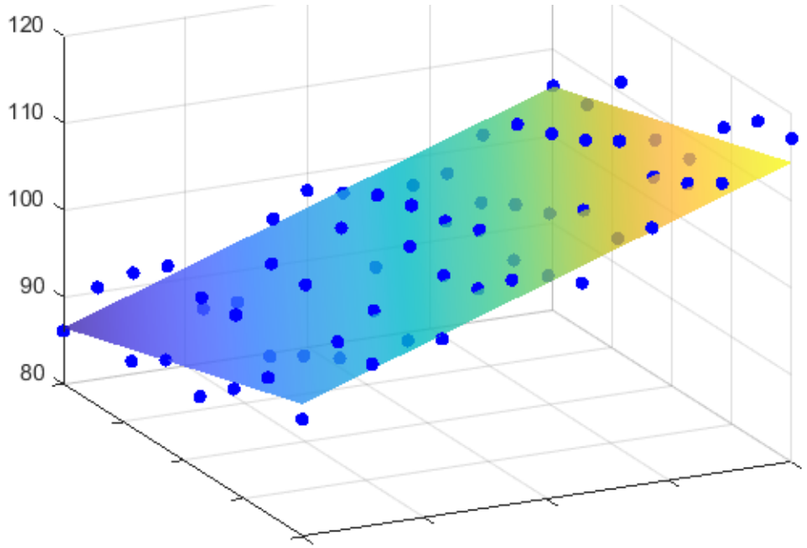
Observed



Estimated



True



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

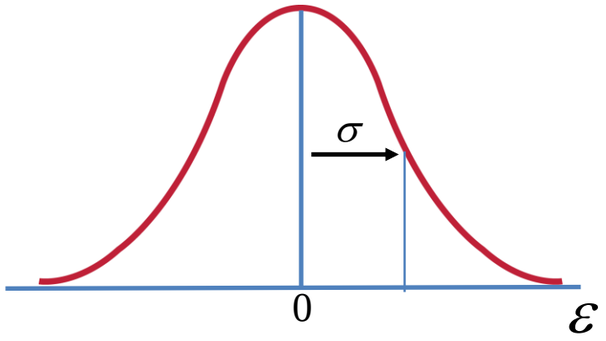
Introduction to Regression Analysis

Questions?

Chapter 3: Simple Linear Regression

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Simple Linear Regression
Model Assumptions



The probabilistic (linear) model relating y to x is

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

Assumption 1 The mean of the probability distribution of ε is 0. $E(\varepsilon) = 0$

Assumption 2 The variance of the probability distribution of is constant. $\text{var}(\varepsilon) = \sigma^2$

Assumption 3 The probability distribution of ε is normal. $\varepsilon \sim N(0, \sigma^2)$

Assumption 4 The errors associated with any two observations are independent.

$$f(\varepsilon_i, \varepsilon_j) = f(\varepsilon_i) f(\varepsilon_j)$$

Simple Linear Regression
Assessing the Utility of the Model

```
# R Code
x=c(1,2,3,4,5)
y=c(1,1,2,2,4)
model=lm(y~x)
summary(model)
```

Hypothesized probabilistic model

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

Wish to test to see if β_1 is statistically significant.

$$H_0: \beta_1 = 0 \quad \xrightarrow{?} \quad y = \beta_0 + \varepsilon$$
$$H_a: \beta_1 \neq 0$$

If the errors are normally distributed, $\varepsilon \sim N(0, \sigma^2)$, then $\hat{\beta}_1 \sim N(\beta_1, \sigma^2 / SS_{xx})$.

$$t = \frac{\hat{\beta}_1 - Hypothesized\ Value}{s / \sqrt{SS_{xx}}}$$

$$t = \frac{\hat{\beta}_1 - 0}{s / \sqrt{SS_{xx}}}$$ has a Student-t distribution with n-2 degrees of freedom.

Simple Linear Regression

Assessing the Utility of the Model

```
# R Code
x=c(1,2,3,4,5)
y=c(1,1,2,2,4)
model=lm(y~x)
confint(model, level=0.95)
```

A 100(1-α)% Confidence Interval for the Simple Linear Regression Slope β_1

$$\hat{\beta}_1 \pm t_{\alpha/2} \frac{s}{\sqrt{SS_{xx}}}$$

and $t_{\alpha/2}$ is based on a Student-t distribution with $(n\text{-}2)$ df

Simple Linear Regression

The Coefficient of Correlation

Pearson product moment coefficient of correlation *r* is

Wish to test to see if ρ is statistically significant.

$H_0: \rho = 0$

$H_a: \rho \neq 0$

If the errors are normally distributed, then

$$t = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}}$$ has a Student-t distribution with n-2 degrees of freedom.

Simple Linear Regression

Using the Model for Estimation and Prediction

A 100(1-α)% Confidence Interval for the Mean Value of y for $x=x_p$

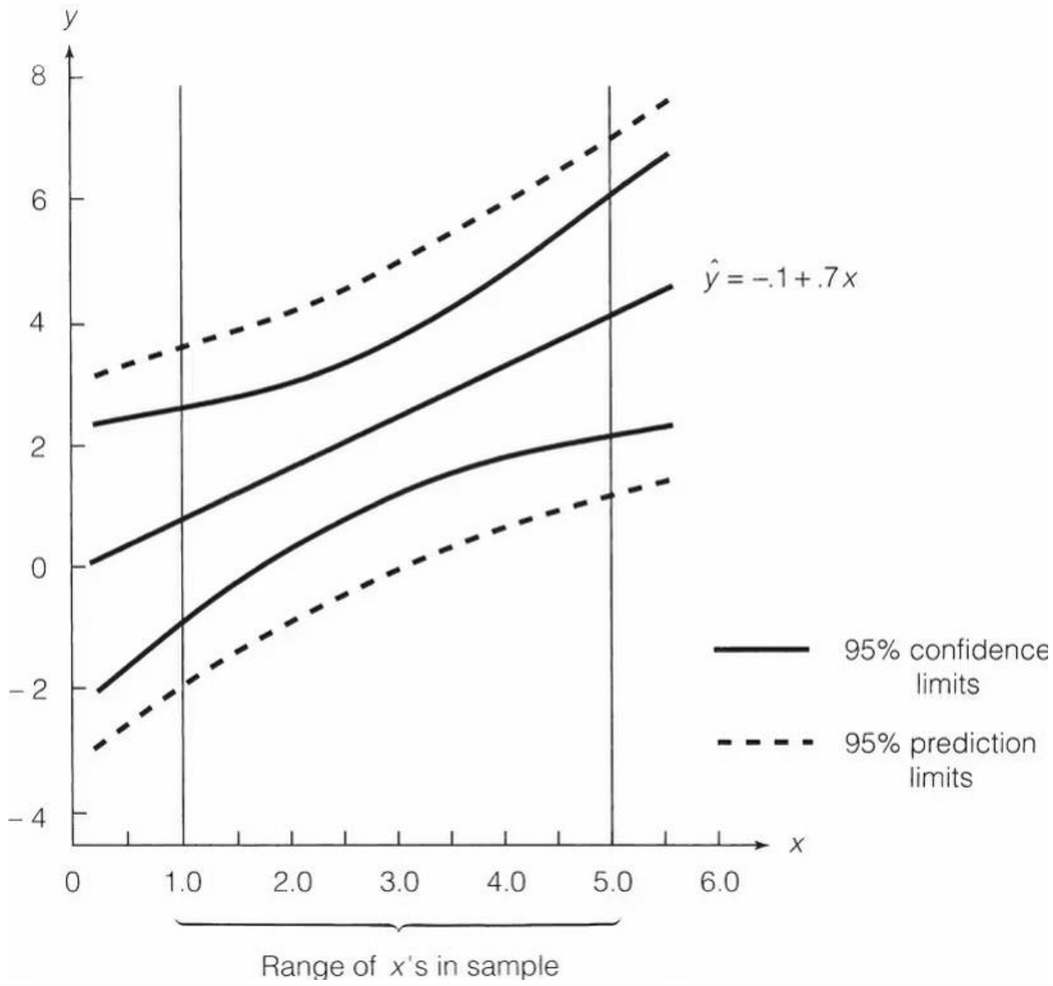
$$\sigma_{\hat{y}} = \sigma\sqrt{\frac{1}{n}+\frac{(x_p-\bar{x})^2}{SS_{xx}}}$$

$$\hat{y} \pm t_{\alpha/2} s\sqrt{\frac{1}{n}+\frac{(x_p-\bar{x})^2}{SS_{xx}}}$$

A 100(1-α)% Prediction Interval for an Individual y for $x=x_p$

$$\sigma_{(y-\hat{y})} = \sigma\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{SS_{xx}}}$$

$$\hat{y} \pm t_{\alpha/2} s\sqrt{1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{SS_{xx}}}$$



Simple Linear Regression

Questions?

Chapter 4: Multiple Regression Models

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Multiple Regression Models

Model Assumptions

The multiple regression model

$$y = \overbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}^{\text{Deterministic Portion}} + \underbrace{\varepsilon}_{\text{Random Error}}$$

Assumptions About the Random Error ε

1. For any given set of values of $x_1, \ldots, x_k$, the error ε has a normal probability distribution with mean equal to 0 [i.e., $E(\varepsilon)=0$] and variance equal to σ^2 [i.e., $\text{var}(\varepsilon)=\sigma^2$]. (Normal only needed for inferences, CIs and HTs).
2. The random errors are independent (in a probabilistic sense). (For normal errors, independent and uncorrelated are the same.)

Model is called first order if of $x_1, \ldots, x_k$, are all quantitative variables that are not functions of other independent variables.

Multiple Regression Models

Fitting the Model: The Method of Least Squares

Observation	*y* Value	x_1	x_2	$\cdots$	x_k
1	y_1	x_{11}	x_{21}		x_{k1}
2	y_2	x_{12}	x_{22}		x_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	y_n	x_{1n}	x_{2n}		x_{kn}

The multiple regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i \quad i = 1, \ldots, n$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{kn} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ 1 & x_{13} & x_{23} & \cdots & x_{k3} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad E = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y = X\beta + E \qquad SSE = E'E = (Y - X\beta)'(Y - X\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki})^2$$

$$(y - X\beta)'(y - X\beta) = (y - X\hat{\beta})'(y - X\hat{\beta}) + (\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta})$$ ← algebra

$\hat{\beta} = (X'X)^{-1}X'y$ minimizes *SSE* and $s^2 = (y - X\hat{\beta})'(y - X\hat{\beta})/(n - k - 1)$.

Multiple Regression Models

Testing the Utility of a Model: The Analysis of Variance F-Test

For the general multiple linear regression model, $E(y|x_1,...,x_k)=\beta_0+\beta_1x_1+\ldots+\beta_kx_k$, we may test

H_0: $\beta_1=\beta_2=\ldots=\beta_k=0$

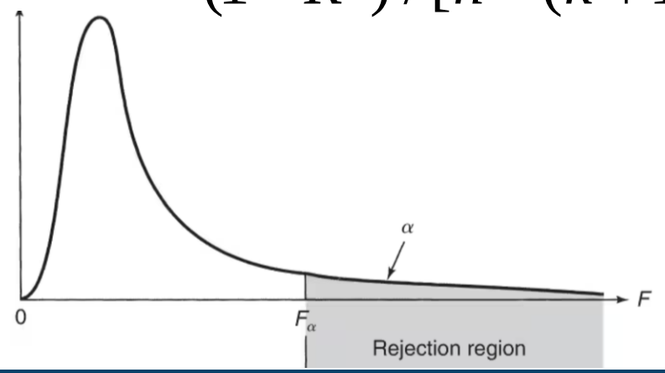
vs.

H_a: At least one of the coefficients is nonzero.

The test statistic is an F statistic,

$$Test\ \ Statistic:\ \ F=\frac{(SS_{yy}-SSE)/k}{SSE/[n-(k+1)]}=\frac{Mean\ \ Square\ (Model)}{MSE}=\frac{R^2/k}{(1-R^2)/[n-(k+1)]}$$

Rejection region: $F>F_\alpha$, where F is based on k numerator and n-$(k+1)$ denominator df or $\alpha>p\text{-}value$, where $p\text{-}value=P(F>F_\alpha)$.



Multiple Regression Models

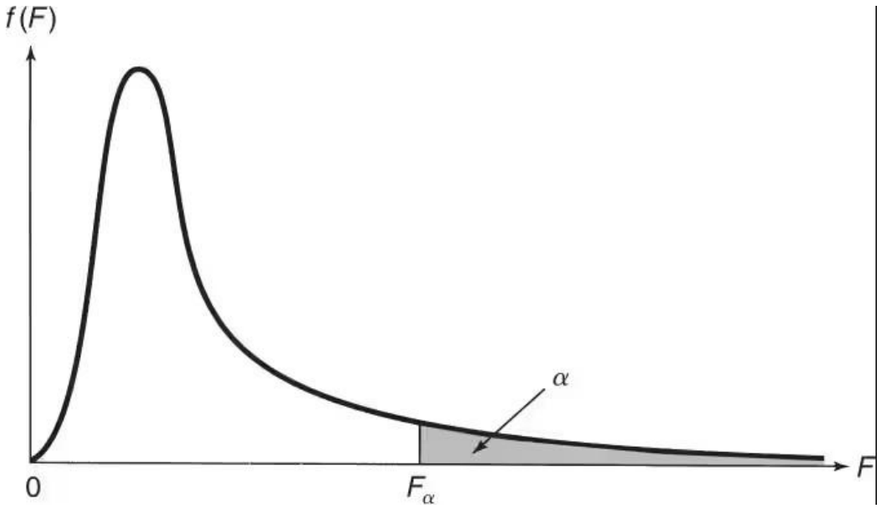
A Test for Comparing Nested Models

F-Test for Comparing Nested Models

Reduced model: $E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g$

Complete model: $E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g + \beta_{g+1} x_{g+1} + \cdots + \beta_k x_k$

H_0: $\beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$



H_a: At least one of the β parameters being tested in nonzero.

$$F = \frac{\text{Dropin SSE/Number of } \beta \text{ parameters being tested}}{s^2 \text{for larger model}} = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/[n-(k+1)]} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

Multiple Regression Models

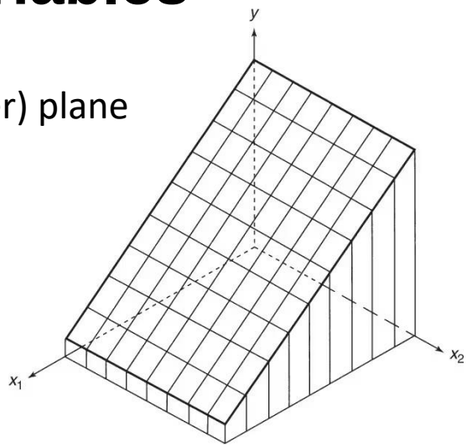
Questions?

Chapter 5: Principles of Model Building

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Principles of Model Building

First-Order Models with Two or More Quantitative Independent Variables



First-Order Model in k Quantitative Independent Variables

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

where $\beta_0, \ldots, \beta_k$ are unknown parameters that must be estimated.

Interpretation of model parameters

β_0: y-intercept of $(k+1)$-dimensional surface; the value of $E(y)$ when $x_1 = \ldots = x_k = 0$

β_1: Change in $E(y)$ for a 1-unit increase in x_1, when $x_2, x_3, \ldots, x_k$ are held fixed.

β_2: Change in $E(y)$ for a 1-unit increase in x_2, when $x_1, x_3, \ldots, x_k$ are held fixed.

$\vdots$

β_k: Change in $E(y)$ for a 1-unit increase in x_k, when $x_1, x_2, \ldots, x_{k-1}$ are held fixed.

Principles of Model Building

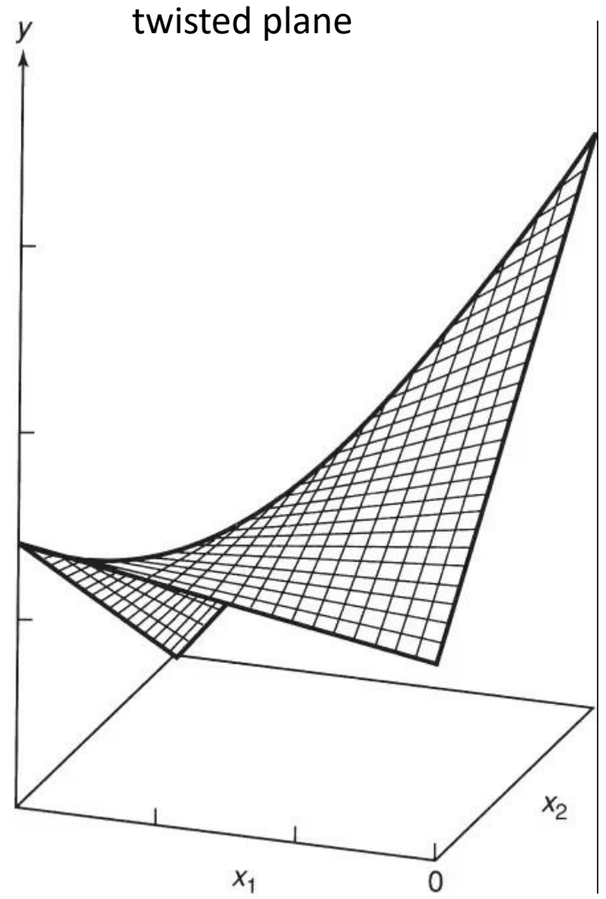
Second-Order Models with Two or More Quantitative Independent Variables

Second-order term accounts for interaction between two variables

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

model traces a twisted plane in a three-dimensional space.

The second-order term $\beta_3 x_1, x_2$ is called the **interaction term**, and it permits the contour lines to be nonparallel.



Principles of Model Building

Second-Order Models with Two or More Quantitative Independent Variables

Interaction (Second-Order) Model with Two Independent Variables

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Interpretation of Model Parameters

β_0: y-intercept; the value of $E(y)$ when x_1=x_2=0

β_1 and β_2: Changing β_1 and β_2 causes the surface to shift along the x_1 and x_2 axes

β_3: Controls the rate of twist in the ruled surface

$\beta_1 + \beta_3 x_2$: Change in $E(y)$ for a 1-unit increase in x_1, when x_2 is held fixed

$\beta_2 + \beta_3 x_1$: Change in $E(y)$ for a 1-unit increase in x_2, when x_1 is held fixed

Principles of Model Building

Second-Order Models with Two or More Quantitative Independent Variables

Interaction (Second-Order) Model with Two Independent Variables

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$

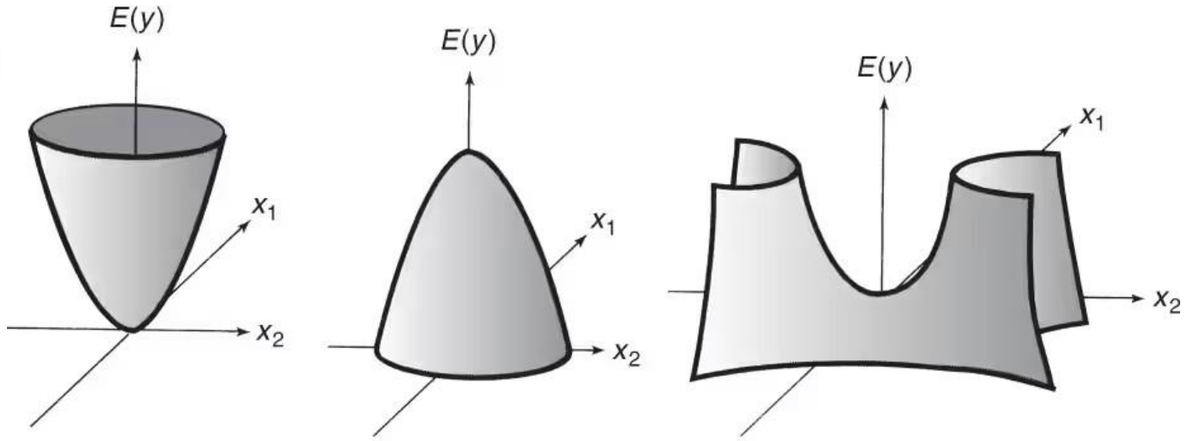
Interpretation of Model Parameters

β_0: y-intercept; the value of $E(y)$ when x_1=x_2=0

β_1 and β_2: Changing β_1 and β_2 causes the surface to shift along the x_1 and x_2 axes

β_3: The value of β_3 controls the surface rotation

β_4 and β_5: Signs and values of these parameters control the type of surfaces the rate of curvature



Principles of Model Building

Questions?

Chapter 6: Variable Screening Methods

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Variable Screening Methods

Introduction: Why Use a Variable Screening Method?

Researches often will collect a data set with a **large** number of independent variables, each of which is a potential predictor of some dependent variable, y.

The problem of deciding to include multiple regression model for $E(y)$ is common. Suppose y depends on 10 x's. 7 quantitative and 3 qualitative yield 288 terms.

$$E(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_7 x_7 + \beta_8 x_1 x_2 + \ldots + \beta_{28} x_6 x_7$$
$$+ \beta_{29} x_1^2 + \ldots + \beta_{35} x_7^2 + \beta_{36} x_8 + \ldots + \beta_{38} x_{10}$$
$$+ \beta_{39} x_8 x_9 + \ldots + \beta_{42} x_8 x_9 x_{10} + \beta_{43} x_1 x_8 + \ldots + \beta_{77} x_7^2 x_8$$
$$+ \beta_{78} x_1 x_9 + \ldots + \beta_{112} x_7^2 x_9 + \ldots + \beta_{113} x_1 x_{10} + \ldots + \beta_{147} x_7^2 x_{10}$$
$$+ \beta_{148} x_1 x_8 x_9 + \ldots + \beta_{182} x_7^2 x_8 x_9 + \beta_{183} x_1 x_8 x_{10} + \ldots + \beta_{217} x_7^2 x_8 x_{10}$$
$$+ \beta_{218} x_1 x_9 x_{10} + \ldots + \beta_{252} x_7^2 x_9 x_{10} + \beta_{253} x_1 x_8 x_9 x_{10} + \ldots + \beta_{287} x_7^2 x_8 x_9 x_{10}$$

Too complex to be practicably useful.

Variable Screening Methods

Stepwise Regression (Forward Selection)

Stepwise Regression: The user identifies the set of potentially important independent variables x's that influence the dependent (response) variable y.

Step 1: Fit all possible one-variable models of the form $E(y)=\beta_0+\beta_1 x_i$, $i=1,\ldots,k$.

Perform the t-test H_0: $\beta_1=0$ vs. H_a: $\beta_1\neq 0$.

$t = \hat{\beta}_i / s\sqrt{W_{ii}}$, W_{ii} is the i^{th} diagonal element of $W=(X'X)^{-1}$.

Select the best one variable model (largest $|t|$ statistic). Call it x_1

Step 2: Fit all two variable models with remaining x's, $E(y)=\beta_0+\beta_1 x_1+\beta_2 x_i$, $i\neq 1$.

Perform the t-test H_0: $\beta_2=0$ vs. H_a: $\beta_2\neq 0$.

$t = \hat{\beta}_i / s\sqrt{W_{ii}}$, W_{ii} is the i^{th} diagonal element of $W=(X'X)^{-1}$.

Select the best two variable model (largest $|t|$ statistic). Call it x_2

Go back and check the t-value of $\hat{\beta}_1$ after $\hat{\beta}_2$ has been added to the model.

Variable Screening Methods

Stepwise Regression (Forward Selection)

Step 3: Fit all three variable models with remaining x's, $E(y)=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_i$, $i\neq 1,2$.

Perform the t-test H_0: $\beta_3=0$ vs. H_a: $\beta_3\neq 0$.

$t=\hat{\beta}_i / s\sqrt{W_{ii}}$, W_{ii} is the i^{th} diagonal element of $W=(X'X)^{-1}$.

Select the best two variable model (largest $|t|$ statistic). Call it x_2

Go back and check the t-values of $\hat{\beta}_1, \hat{\beta}_2$ after $\hat{\beta}_3$ has been added.

This procedure is continued until no further independent variables can be found that yield significant t-values (at the specified α level) in the presence of the variables already in the model.

Variable Screening Methods

All-Possible-Regressions Selection Procedure

There are several criteria that can be used.

1. $R^2 = 1 - \dfrac{SSE}{SS(Total)}$

   R-Square

2. $R_a^2 = 1 - (n-1)\left[\dfrac{MSE}{SS(Total)}\right]$

   Adjusted R-Square

3. $C_p = \dfrac{SSE_p}{MSE_k} + 2(p+1) - n$

   Mallow's C_p

4. $PRESS = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2$

   Predictive Sum of Squares

1. R^2 criterion
   Looking for a simple model that is as good as, or nearly as good as, the model with all k independent variables.
2. Adjusted R^2 or MSE criterion
   Prefer the model with largest, or near largest, adjusted R^2.
3. Mallow's C_p Criterion
   Prefer a small value of C_p and a value of C_p near p+1.
4. $PRESS$ Criterion
   Desire a model with a small $PRESS$.

Variable Screening Methods

Questions?

Chapter 7: Some Regression Pitfalls

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Some Regression Pitfalls

Observational Data versus Designed Experiments

When an experiment has been designed, the experimental units have an equal chance of receiving unusually high (or low) readings.

This averages out any variation within the experimental units and statistically significant difference between sample means implies that you can infer that the population means differ.

More importantly, you can infer that this difference was from the settings of the predictor x variables. Thus, you can infer a ***cause-and-effect*** relationship.

If the data are observational, a statistically significant relationship between x and y does not imply a cause-and-effect relationship. It simply means that x contributes information for the prediction of y, and nothing more.

Some Regression Pitfalls

Multicollinearity

Detecting Multicollinearity in the Regression Model

1. Significant correlations between pairs of independent variables in the model

2. Nonsignificant t-tests for all (or nearly all) the individual β parameters when the F-test for overall model adequacy H_0: $\beta_1=\beta_2=\ldots=\beta_k=0$ is significant

3. Opposite signs (from what is expected) in the estimated parameters

4. A variance inflation factor (VIF) for a β parameter greater than 10, where $R_i^2 > 0.90$

$$(VIF)_i = \frac{1}{1-R_i^2},\ i=1,\ldots,k$$

and R_i^2 is the multiple coefficient of determination for the model

$E(x_i)=\alpha_0+\alpha_1 x_1+\alpha_2 x_2+\ldots+\alpha_{i-1}x_{i-1}+\alpha_3 x_{i+1}+\ldots+\alpha_k x_k$.

Some Regression Pitfalls

Multicollinearity

Solutions to Some Problems Created by Multicollinearity

1. Drop one or more of the correlated x's. Stepwise regression is helpful in dropping.

2. If you decide to keep all the independent variables in the model:
   a. Avoid making inferences about the individual parameters.
   b. Restrict inferences about $E(y)$ and future y-values to the experimental region.

3. To establish cause-and-effect between y and the x's, use a designed experiment.

Causal Inference

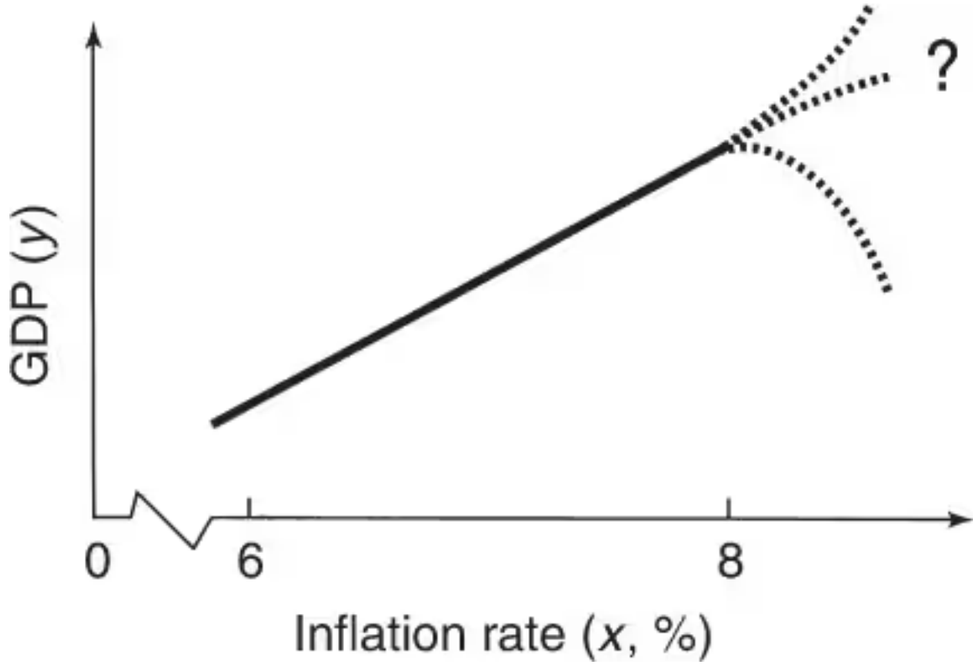
4. To reduce rounding errors in polynomial regression, code the x variables so that 1^{st}, 2^{nd}, and higher-order terms for a particular x are not highly correlated.

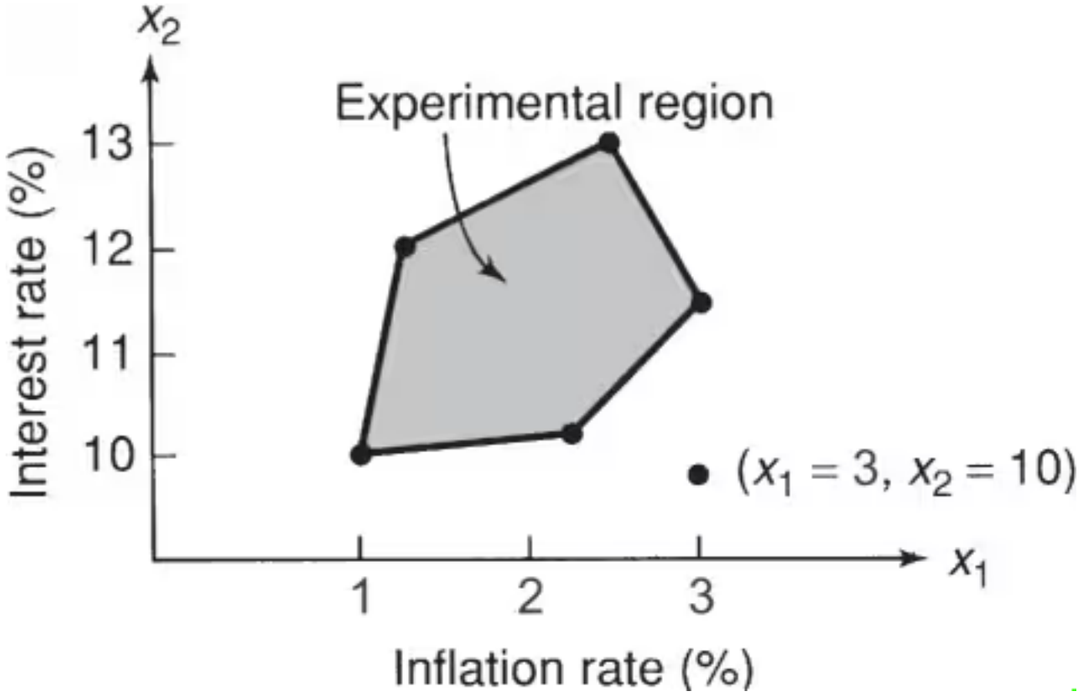
5. To reduce rounding errors and stabilize the regression coefficients, use ridge regression to estimate the β parameters.

Some Regression Pitfalls

Extrapolation: Predicting Outside the Experimental Region

Quite often when we develop statistical models, we want to not just interpolate between observations we have, but to forecast (extrapolate) additional observations.



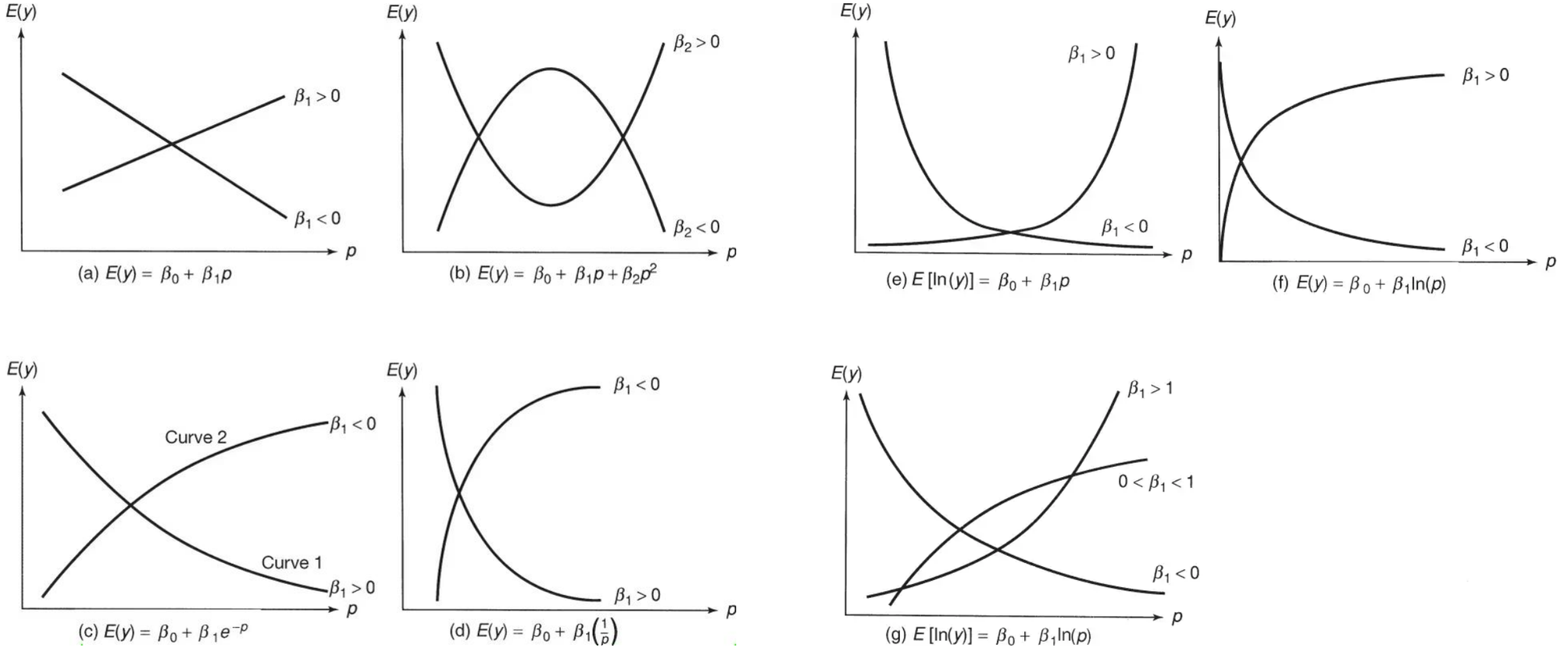


Some Regression Pitfalls

Variable Transformations

Transformations are performed on the y-values to make them to make them resemble $E(y)$ and satisfy the linear regression model assumptions.

$y=E(y)+\varepsilon$



Transforming y and/or the x's in a model can provide a better model fit.

Some Regression Pitfalls

Questions?

Chapter 8: Residual Analysis A

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Residual Analysis

Introduction

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

When we test a hypothesis about a regression coefficient or a set of regression coefficients, or when we form a prediction interval for a future value of y, we must assume that

(0) need to get $E(y)$ correct or we have lack of fit
(2) ε has a mean of 0, $E(\varepsilon)=0$
(3) the variance of ε is σ^2 is constant, $var(\varepsilon)= \sigma^2$ and
(4) all pairs of error terms are uncorrelated $cor(\varepsilon_i,\varepsilon_j)=0$
(1) ε is normally distributed (for CIs and HTs)

Graphical tools and statistical tests that will aid in identifying significant departures from the assumptions.

Residual Analysis

Detecting Lack of Fit

Detecting Model Lack of Fit with Residuals

Plot the residuals, $\hat{\varepsilon}$, on the vertical axis against each of the independent variables, $x_1,\ldots,x_n$ on the horizontal axis.

Plot the residuals, $\hat{\varepsilon}$, on the vertical axis against the predicted value, $\hat{y}$ on the horizontal axis.

In each plot, look for trends, dramatic changes in variability, and/or more than 5% of residuals that lie outside $1.96s$ of 0. Any of these patterns indicates a problem with model fit.

Residual Analysis
Detecting Lack of Fit

An alternative method of detecting lack of fit in models with more than one independent variable is to construct a partial residual plot.

The set of partial regression residuals for the jth independent variable x_j is calculated as follows:

$$\hat{\varepsilon}^* = y - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_{j-1} x_{j-1} + \hat{\beta}_{j+1} x_{j+1} + \ldots + \hat{\beta}_k x_k)$$

$$\hat{\varepsilon}^* = \hat{\varepsilon} + \hat{\beta}_j x_j$$

where $\hat{\varepsilon} = y - \hat{y}$ is the usual regression residual.

Residual Analysis

Detecting Unequal Variances

One of the regression assumptions is that the variance σ^2 is constant, $var(\varepsilon)= \sigma^2$.

When $var(\varepsilon_i)=\sigma^2$, $i=1,\ldots,n$, the errors are called homoscedastic.
When $var(\varepsilon_i)=\sigma_i^2$, $i=1,\ldots,n$, the errors are called heteroscedastic.

When data fail to be homoscedastic, the reason is often that the variance of the response y is a function of its mean $E(y)$.
i.e. Unmodeled signal gets put into the variance.

Residual Analysis

Detecting Unequal Variances

When the variance of y is a function of its mean, we can often satisfy the least squares assumption of homoscedasticity by transforming the response to some new response that has a constant variance.
These are called **variance-stabilizing transformations**.

Type of Response	Variance	Stabilizing Transformation
Poisson	$E(y)$	$\sqrt{y}$
Binomial proportion	$\frac{E(y)[1-E(y)]}{n}$	$\sin^{-1}\sqrt{y}$
Multiplicative	$[E(y)]^2\sigma^2$	$\ln(y)$

Residual Analysis

Detecting Unequal Variances

to detect non-constant variance is to split
the data in two parts. Fit
A more quantitative way the regression model to each part.
Perform a hypothesis test for non-equality of variances.

H_0: $\frac{\sigma_1^2}{\sigma_2^2}=1$ (Assumption of equal variances satisfied)

H_1: $\frac{\sigma_1^2}{\sigma_2^2}\neq 1$ (Assumption of equal variances not satisfied)

σ_1^2 =Variance of the random error term, ε, for subpopulation 1 (i.e. $x<20$)
σ_2^2 =Variance of the random error term, ε, for subpopulation 1 (i.e. $x\geq 20$)

$$F=\frac{Larger\ s^2}{Smaller\ s^2}=\frac{Larger\ MSE}{Smaller\ MSE}$$

Reject H_0 if $F>F_{\alpha,n/2\text{-}k\text{-}1,n/2\text{-}k\text{-}1}$.

Residual Analysis

Questions?

Chapter 8: Residual Analysis B

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Residual Analysis

Checking the Normality Assumption

Hypothesis tests are available to check the normality assumption.

Shapiro-Wilk test: A hypothesis test that's often used for small samples

Kolmogorov-Smirnov test: A non-parametric test that's often used for large samples

Lilliefors test: Based on the K-S test, adjusted to also estimate mean and variance

Anderson-Darling test: A test that's often used for heavier-tailed distributions

D'Agostino-Pearson test: A test that assesses normality based on skewness

However, at this time we will be qualitatively assessing normality.

Residual Analysis

Checking the Normality Assumption

Graphical methods to assess normality.

1. Histogram for the residuals. Inspect for looking like normal curve.
2. Stem-and- leave plot of the residuals. Inspect for looking like normal curve.
3. Normal probability plot (QQ plot). Plot ε_i vs. $E(\varepsilon_i)$ assuming normality.

$$z_i = \Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right) \quad , \qquad \text{for } i=1,\ldots,n.$$

R uses

$$z_i = \Phi^{-1}\left(\frac{i-a}{n+1-2a}\right) \quad , \qquad \text{for } i=1,\ldots,n \text{ where } a=3/8 \text{ if } n\leq 10,\ a=1/2 \text{ if } n>10.$$

Inspect for points looking like fit a line indicating normality satisfied.

https://en.wikipedia.org/wiki/Normal_probability_plot

Residual Analysis

Detecting Outliers and Identifying Influential Observations

The standardized residual, denoted z_i, for the ith observation is the residual for the observation divided by s, that is,

$$z_i = \hat{\varepsilon}_i / s = (y_i - \hat{y}_i) / s$$

An observation with a residual that is larger than $3s$ (in absolute value)
or, equivalently, a standardized residual that is larger than 3 (in absolute value)
is considered to be an **outlier**.

The studentized residual, denoted z_i*, for the ith observation is

$$z_i^* = \frac{\hat{\varepsilon}_i}{s\sqrt{1-h_i}} = \frac{(y_i - \hat{y}_i)}{s\sqrt{1-h_i}}$$

where h_i (called leverage). h_i=ith diagonal element of $X(X'X)X'$.

Residual Analysis

Detecting Outliers and Identifying Influential Observations

The **leverage** of the ith observation is h_i, associated with y_i in the equation

$$\hat{y}_i = h_1 y_1 + h_2 y_2 + \cdots + h_i y_i + \cdots + h_n y_n$$

where $h_1, h_2, h_3, \ldots, h_n$ are functions of only the x's in the model.
The leverage, h_i, measures the influence of y_i on its predicted value $\hat{y}_i$.
Hat matrix $H = X(X'X)^{-1}X'$ with ith diagonal element h_i.

Rule of Thumb: $h_i > 2(k+1)/n$ is outlier

Cook's Distance: A large value of indicates that the observed y_i value has strong influence on the estimated β coefficients. Compare D_i to the F distribution with $v_1 = k+1$ and $v_2 = n-k-1$.

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(k+1)MSE}\left[\frac{h_i}{(1-h_i)^2}\right]$$

Residual Analysis

Detecting Residual Correlation: The Durbin-Watson Test

We can test for temporal autocorrelation with the **Durbin-Watson** statistic.
Once we fit a regression model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k ,$$

we calculate the residuals

$$\hat{\varepsilon}_i = \hat{y}_i - \hat{\beta}_{0i} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_k x_{ki}$$

and test for positive correlation, H_0: $\rho \leq 0$ vs. H_a: $\rho > 0$ via

$$d = \frac{\sum_{t=2}^{n} (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2} \approx \underbrace{2(1-\hat{\rho})}_{\text{Large } n}$$

and reject for $d << 2$. $0 \leq d \leq 4$

H_0: $\rho \leq 0$ vs. H_a: $\rho > 0$
H_0: $\rho \geq 0$ vs. H_a: $\rho < 0$
H_0: $\rho = 0$ vs. H_a: $\rho \neq 0$

$$\sum_{t=2}^{n} (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2$$

~~$\hat{\varepsilon}_1$~~	$\hat{\varepsilon}_2$	$\hat{\varepsilon}_3$	$\hat{\varepsilon}_4$	$\hat{\varepsilon}_5$	$\hat{\varepsilon}_6$	$\hat{\varepsilon}_7$	$\hat{\varepsilon}_8$	$\hat{\varepsilon}_9$	$\hat{\varepsilon}_{10}$
$\hat{\varepsilon}_1$	$\hat{\varepsilon}_2$	$\hat{\varepsilon}_3$	$\hat{\varepsilon}_4$	$\hat{\varepsilon}_5$	$\hat{\varepsilon}_6$	$\hat{\varepsilon}_7$	$\hat{\varepsilon}_8$	$\hat{\varepsilon}_9$	~~$\hat{\varepsilon}_{10}$~~

Residual Analysis

Questions?

Chapter 9: Special Topics in Regression A

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Special Topics in Regression

Piecewise Linear Regression

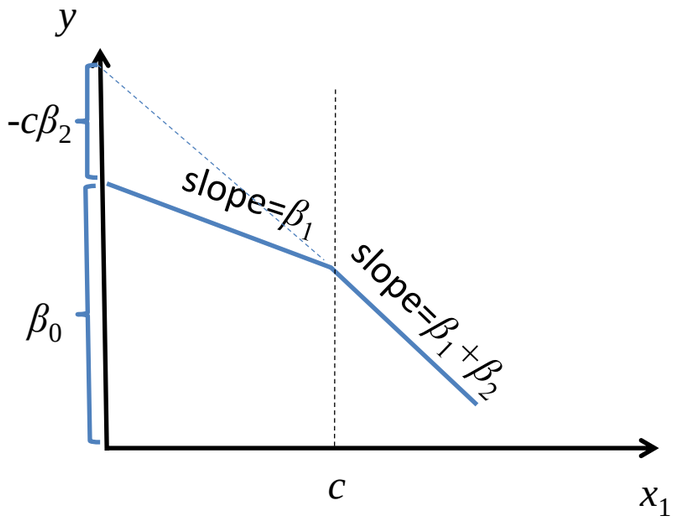
Occasionally a single line model

$E(y) = \beta_0 + \beta_1 x_1$ is not sufficient for our data,

and a continuous two-line or changepoint model

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 (x_1 - c) x_2$ is appropriate

$$x_2 = \begin{cases} 1 \ if \ x_1 > c \\ 0 \ if \ x_1 \le c \end{cases} \qquad x_2^* = (x_1 - c) x_2$$



c is called the knot value.

Special Topics in Regression

Piecewise Linear Regression

Three straight lines are also possible,

and a continuous three-line or changepoint model is

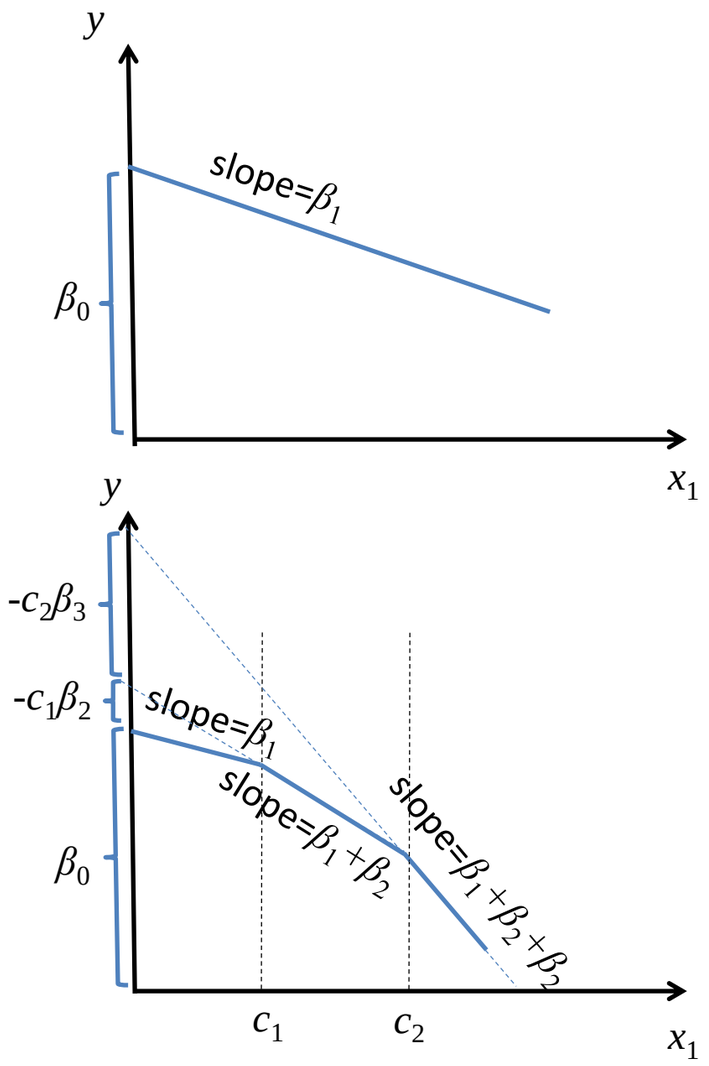
$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 (x_1 - c_1) x_2 + \beta_3 (x_1 - c_2) x_3$$

$$x_2 = \begin{cases} 1 \ if \ x_1 > c_1 \\ 0 \ if \ x_1 \le c_1 \end{cases} \qquad x_3 = \begin{cases} 1 \ if \ x_1 > c_2 \\ 0 \ if \ x_1 \le c_2 \end{cases}$$

c_1 and c_2 are the knot values.

$$x_2^* = (x_1 - c_1) x_2$$

$$x_3^* = (x_1 - c_2) x_3$$

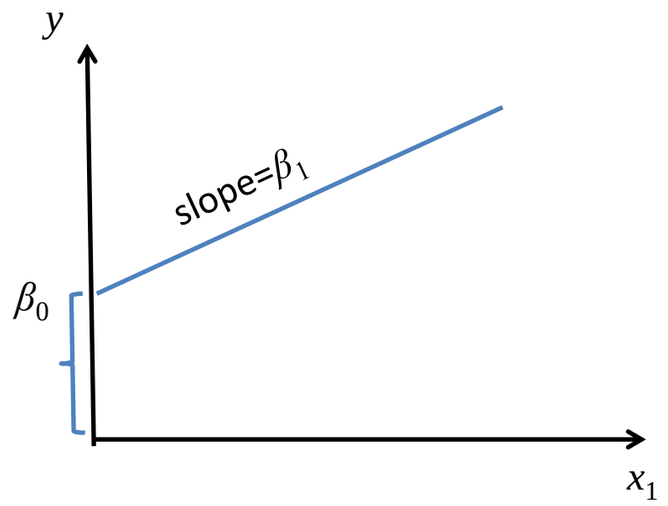


Special Topics in Regression

Piecewise Linear Regression

Occasionally a single line model

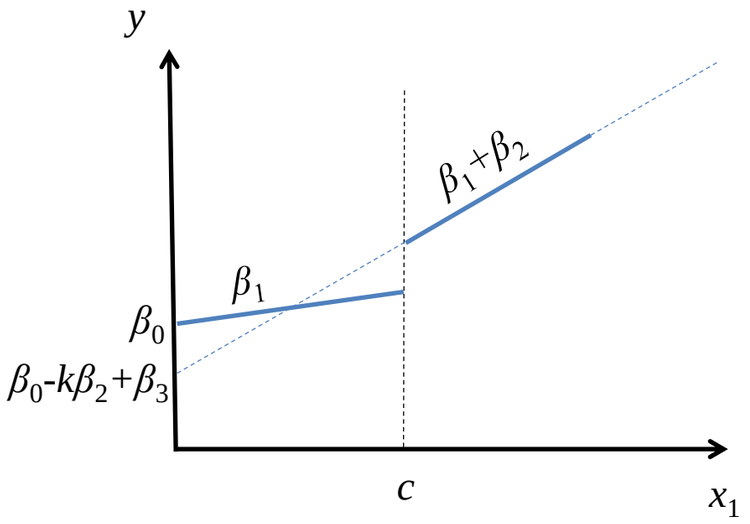
$E(y) = \beta_0 + \beta_1 x_1$ is not sufficient for our data,



and a discontinuous two-line model

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 (x_1 - c) x_2 + \beta_3 x_3$ is appropriate

$$x_2 = \begin{cases} 1 \ if \ x_1 > c \\ 0 \ if \ x_1 \le c \end{cases} \qquad x_2^* = (x_1 - c) x_2$$



c is called the knot (discontinuity) value.

Special Topics in Regression

Weighted Least Squares

Let's reconsider the problem of heteroscedastic errors, nonconstant variance.

Many times, transformations ($\sqrt{y}$, $\log(y)$, $1/y$ and $1/\sqrt{y}$) are not effective in stabilizing the error variance so we use weighted least squares.

Weighted Least Squares Properties

1. Stabilizing the variance of to satisfy the standard regression assumption of homoscedasticity.

2. Limiting the influence of outlying observations in the regression analysis.

3. Giving greater weight to more recent observations in time series analysis.

Special Topics in Regression

Weighted Least Squares

For weighted least squares, the residuals are

$$r_i^* = \sqrt{w_i}(y_i - \hat{y}_i) \qquad WSSE = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} w_i (y - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 - \cdots - \hat{\beta}_k x_k)^2$$

$$WSSE = \sum_{i=1}^{n} (r_i^*)^2$$

Generally weights are determined by

$$w_i = \frac{1}{\sigma_i^2}$$

The variance at observation i is unknown and often modeled as proportional to x,

$\sigma_i^2 = cx_i$, and the weight becomes $w_i = \frac{1}{cx_i}$, but it has been shown that can use c=1.

Special Topics in Regression

Weighted Least Squares

Determining the Weights in Weighted Simple Least Squares Regression

1. Divide the data into several groups according to the values of the independent variable, x. The groups should have approximately equal sample sizes.
   a. If the data is replicated and balanced, then create one group for each value of x.
   b. If the data is not replicated, group the data according into ranges of x

2. Determine the sample mean $\bar{x}$ and variance s^2 of the residuals in each group.
3. For each group, compare the residual variance s^2 to different functions of $\bar{x}$ by calculating the ratio $s^2/f(\bar{x})$
4. Find the function of $\bar{x}$ for which the ratio is nearly constant across groups.

5. The appropriate weights for the groups are $1/f(\bar{x})$.

Special Topics in Regression

Questions?

Chapter 9: Special Topics in Regression B

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Special Topics in Regression

Ridge and LASSO Regression

Sometimes we have multicollinearity, we don't get a very good estimate of our regression coefficients β due to the inversion of the ill conditioned matrix $(X'X)$.

$$\hat{\beta}_{LS} = (X'X)^{-1}X'y$$

There is a theorem in Linear Algebra that describes a matrix A such that $(X'X+A)$ is well conditioned and invertible. Bayesian origins.

$$\hat{\beta}_{R} = (X'X + A)^{-1}X'y$$

Ridge regression uses $A=cI_{k+1}$, $c\geq 0$. However, $\hat{\beta}_R$ is biased in that $E(\hat{\beta}_R)\neq\beta$.

Choose c such that $MSE_R < MSE_{LS}$.

Special Topics in Regression
Ridge and LASSO Regression

Consider the Bayes regression estimator

$$\hat{\beta}_{Bayes} = (X'X + cI_{k+1})^{-1}(c\delta + X'X\hat{\beta}_{MLE})$$

If c=0, we get the LS estimator

$$\hat{\beta}_{LS} = (X'X)^{-1}X'y$$

and if c>>0, we get the prior estimator

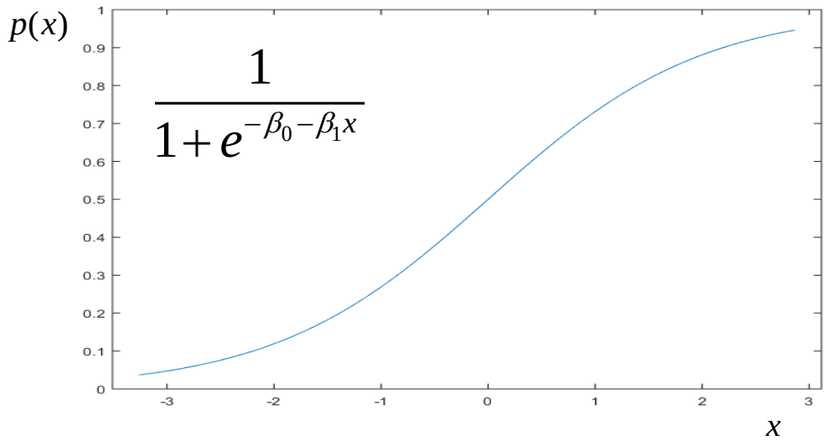
$\hat{\beta}_{prior} = \delta$ which is 0 in Ridge Regression.

So we want to select c as small as possible that allows $(X'X+cI)^{-1}$.

Special Topics in Regression
Logistic Regression

This dependency of a probability $p(x)$, $0 \leq p(x) \leq 1$, on an independent variable x, $-\infty < x < \infty$, is generally described through the logistic mapping function

$$p = p(x) = \frac{1}{1+e^{-\beta_0-\beta_1 x}} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad .$$



If the event E occurs, then we say $y=1$ and if not $y=0$.
$P(y=1)=p$ and $P(y=0)=1-p$. …

This is a Binomial trial with $n=1$ and whose probability of success depends on x.

Verhulst, 1838; Ostwald, 1883; .., Fisher,1935,….

Special Topics in Regression
Logistic Regression

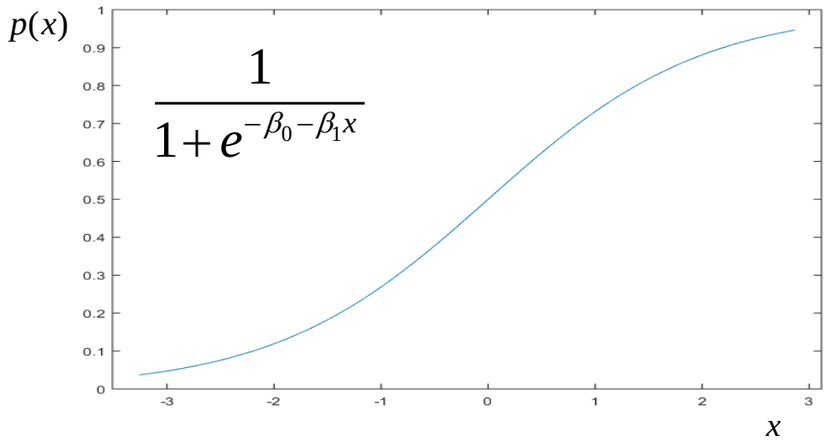
Sometimes the logistic regression is written as log odds

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p$$

and it looks like we can then use Linear Regression to estimate the coefficients. It turns out that we need to find the coefficient values that maximize

$$LL = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} ln(1 + e^{\beta_0 + \beta_1 x_i}) .$$

We need to use a software package such as ***R***.



Special Topics in Regression

Questions?

Chapter 10: Introduction to Time Series Modeling and Forecasting A

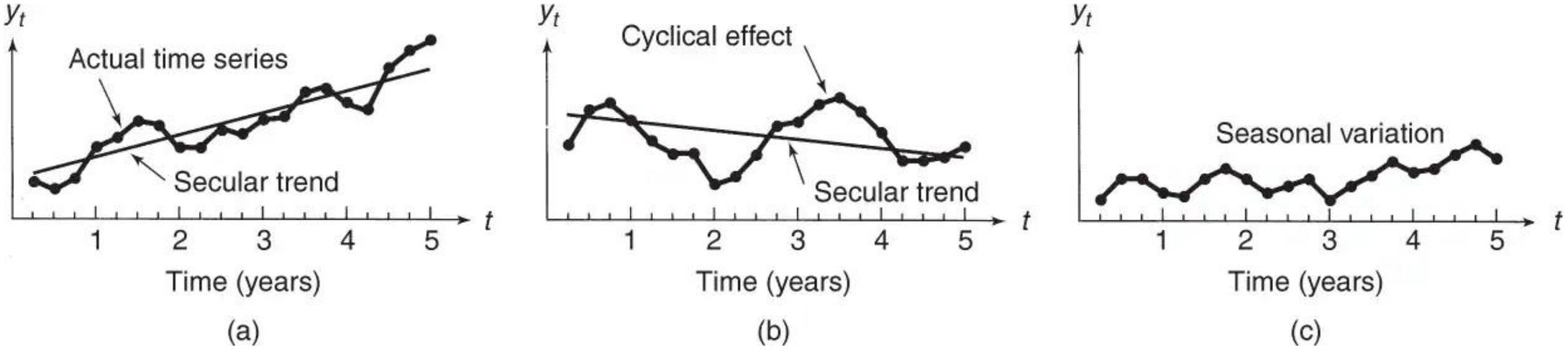
Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

Intro to Time Series Modeling & Forecasting

Time Series Components

If repeated observations on a variable produce a time series, the variable is called a **time series variable**. We use y_t to denote the value of the variable at time t.

Researchers often describe a time series y_t by four components
(1) secular trend, (2) cyclical effect, (3) seasonal variation, and (4) residual effect.



the residual effect, is what remains after other components have been removed.

Intro to Time Series Modeling & Forecasting

Time Series Components

The **secular trend** (T_t) is the tendency of the series to increase or decrease over a long period of time. It is also known as the long-term trend.

The **cyclical fluctuation** (C_t) is the wavelike or oscillating pattern about the secular trend. It is also known as a business cycle.

The **seasonal variation** (S_t) describes the fluctuations that recur during specific portions of the year (e.g., monthly or seasonally).

The **residual effect** (R_t) is what remains after the secular, cyclical, and seasonal components have been removed.

Additive model $y_t = T_t + C_t + S_t + R_t$

Intro to Time Series Modeling & Forecasting

Autocorrelation and Autoregressive Error Models

A property commonly observed for autocorrelated residuals is that the size of the autocorrelation between values of the residual R at two different points in time diminishes rapidly as the distance between the time points increases.

Thus, the autocorrelation between R_t and R_{t+1} becomes smaller (i.e., weaker) as the distance m between the time points becomes larger.

First-order autoregressive error model: $R_t=\phi R_{t+m}+\varepsilon_t$, $-1<\phi<1$.

A consequence of which is that $\mathrm{AC}(R_t,R_{t+m})=\phi^m$.

Intro to Time Series Modeling & Forecasting

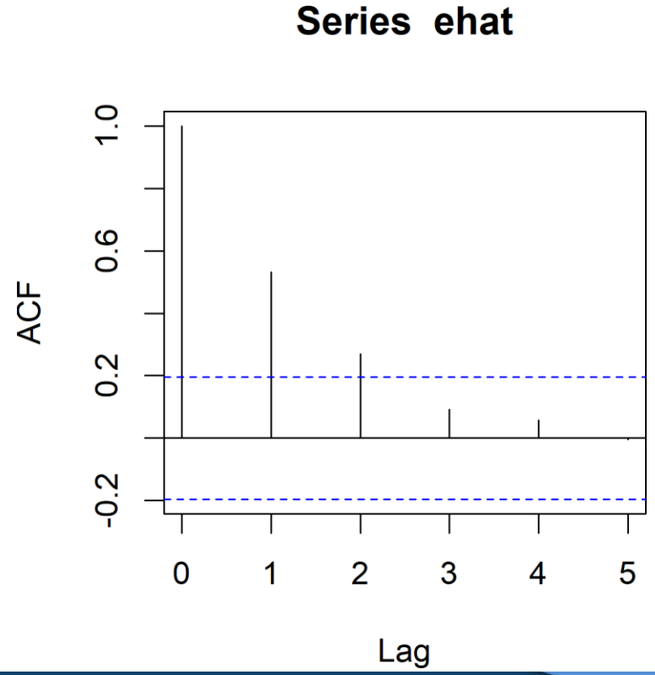
Autocorrelation and Autoregressive Error Models

First-order autoregressive error model: $R_t=\phi R_{t+1}+\varepsilon_t$, $-1<\phi<1$.

A consequence of which is that $\mathrm{AC}(R_t,R_{t+m})=\phi^m$.

Recall: Chapter 8b worksheet with

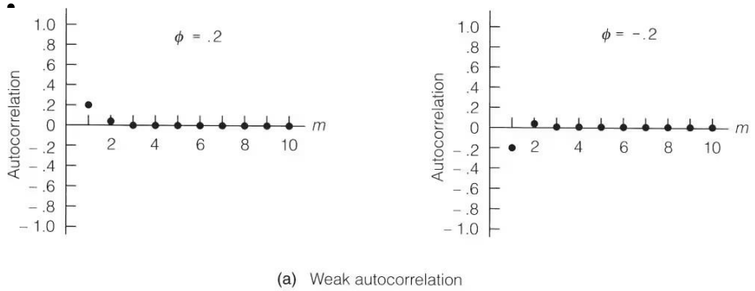
ϕ =0.5. y[i]=b0+b1*t[i]+phi*y[i-1]+ e[i]



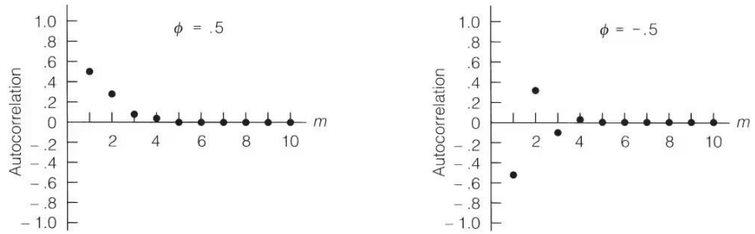
$\mathrm{AC}(R_t,R_{t+m})=(0.5)^m$, m=0,1,2,3,4,5.

1.0, 0.5, 0.25, 0.125, 0.0625, 0.0312

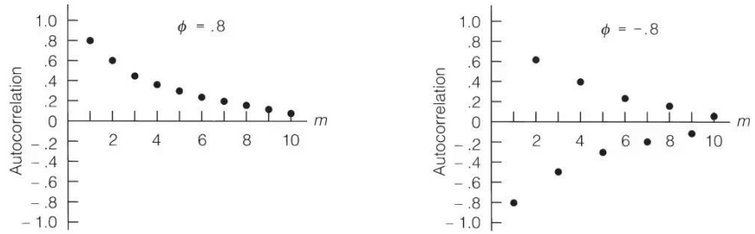
Durbin-Watson Test



(a) Weak autocorrelation



(b) Moderate autocorrelation



(c) Strong autocorrelation

Intro to Time Series Modeling & Forecasting

Autocorrelation and Autoregressive Error Models

First-order autoregressive error model: $R_t=\phi R_{t+1}+\varepsilon_t$, $-1<\phi<1$.

A consequence of which is that $\text{AC}(R_t,R_{t+m})=\phi^m$.

Recall: Chapter 8b worksheet with

$$d=\sum_{t=2}^{n}(\hat{\varepsilon}_t-\hat{\varepsilon}_{t-1})^2 \Big/ \sum_{t=1}^{n}\hat{\varepsilon}_t^2 \approx \underbrace{2(1-\hat{\phi})}_{\text{Large } n} \quad , 0\le d\le 4.$$

H_0: $\phi\le 0$ vs. H_a: $\phi>0$, Reject $d<d_{L,\alpha}$

H_0: $\phi\ge 0$ vs. H_a: $\phi<0$, Reject $(4\text{-}d)<d_{L,\alpha}$

H_0: $\phi=0$ vs. H_a: $\phi\ne 0$, Reject $d<d_{L,\alpha/2}$ or $(4\text{-}d)<d_{L,\alpha/2}$

Intro to Time Series Modeling & Forecasting

Questions?

Chapter 10: Introduction to Time Series Modeling and Forecasting B

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

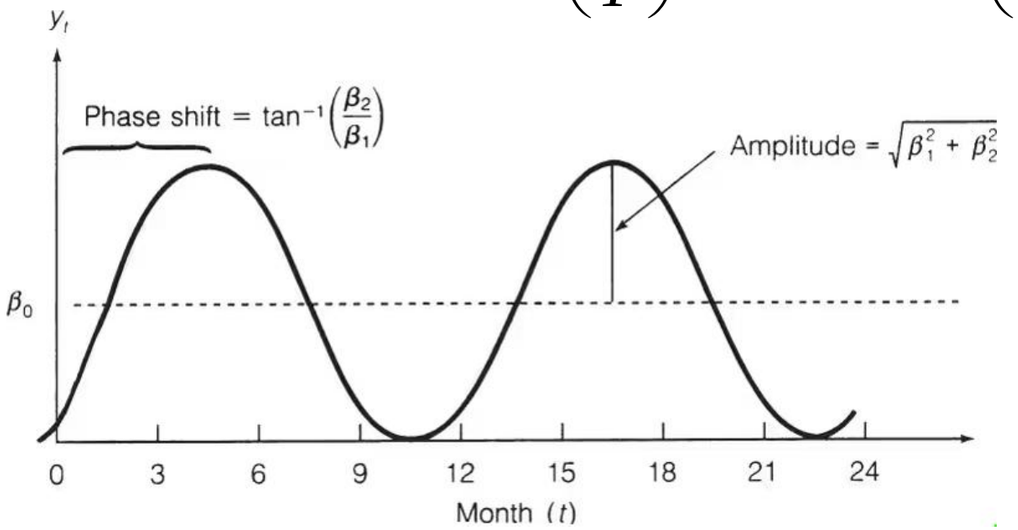
Intro to Time Series Modeling & Forecasting

Constructing Time Series Models

Quite often, the deterministic component has distinct seasonal patterns, which can be modeled with

cosines and sines, T=12 months

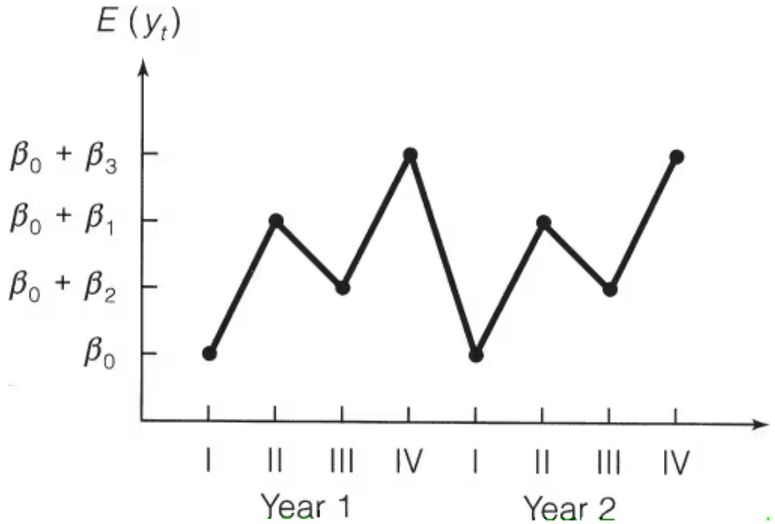
$$E(y_t) = \beta_0 + \beta_1 \cos 2\pi\left(\frac{1}{T}\right)t + \beta_2 \sin 2\pi\left(\frac{1}{T}\right)t$$



$$E(y_t) = \beta_0 + \beta_1 \sin(2\pi t/T + \theta)$$

Annual four seasons

$$E(y_t) = \beta_0 + \beta_1 S_1 + \beta_2 S_2 + \beta_3 S_3$$



$$Q_1 = \begin{cases} 1 \text{ if spring} \\ 0 \text{ if not} \end{cases}$$

$$Q_2 = \begin{cases} 1 \text{ if summer} \\ 0 \text{ if not} \end{cases}$$

$$Q_3 = \begin{cases} 1 \text{ if fall} \\ 0 \text{ if not} \end{cases}$$

Intro to Time Series Modeling & Forecasting

Constructing Time Series Models

We can “whiten” or uncorrelate our residuals by “differencing”

$$R_t = \phi R_{t-1} + \varepsilon_t$$

$$y_t = \beta_0 + \beta_1 t + R_t$$

$$y_{t-1} = \beta_0 + \beta_1 (t-1) + R_{t-1}$$

$$\phi y_{t-1} = \phi\beta_0 + \phi\beta_1 (t-1) + \phi R_{t-1}$$

$$y_t - \phi y_{t-1} = \beta_0 - \phi\beta_0 + \beta_1 [t - \phi(t-1)] + R_t - \phi R_{t-1}$$

$y_t^* = \beta_0^* + \beta_1^* t^* + \varepsilon_t$, $E(\varepsilon_t)=0$, $Var(\varepsilon_t)=\sigma^2$, $Cor(\varepsilon_t,\varepsilon_{t-1})=0$.

$\hat{\phi} = \left[\sum \varepsilon_t^2 / (n-1)\right] / s_\varepsilon$, $\tilde{\beta}^* = (X^{*\prime} X^*)^{-1} X^{*\prime} y^*$, transform back $\tilde{\beta}_0 = \tilde{\beta}_0^* / (1-\hat{\phi})$ and $\tilde{\beta}_1 = \tilde{\beta}_1^*$

Intro to Time Series Modeling & Forecasting

Forecasting with Time Series Autoregressive Models

We can use the regression time series model

$$y_t = \beta_0 + \beta_1 x_t + R_t \quad , \quad R_t = \phi R_{t-1} + \varepsilon_t$$

to forecast future observations, once we have estimated β_0, β_1, ϕ by least squares.
Want to forecast at $t=n+1, n+2, \ldots$

$$y_{t+1} = \beta_0 + \beta_1 x_{t+1} + R_{t+1} \quad , \quad R_{t+1} = \phi R_t + \varepsilon_{t+1}$$

$$y_{t+1} = \beta_0 + \beta_1 x_{t+1} + \phi R_t + \varepsilon_{t+1}$$

which is

$$F_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1} + \hat{\phi}\hat{R}_n \quad , \qquad \hat{R}_n = y_n - (\hat{\beta}_0 + \hat{\beta}_1 x_n)$$

$$F_{t+2} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+2} + \hat{\phi}\hat{R}_{n+1} \quad , \quad \hat{R}_{n+1} = \phi \hat{R}_n \quad , \quad \ldots$$

Intro to Time Series Modeling & Forecasting

Forecasting with Time Series Autoregressive Models

Approximate 95% PI Forecasting Limits

$$\hat{y}_{n+1} \pm 1.96\sqrt{MSE}$$

$$\hat{y}_{n+2} \pm 1.96\sqrt{MSE(1+\hat{\phi}^2)}$$

$$\hat{y}_{n+3} \pm 1.96\sqrt{MSE(1+\hat{\phi}^2+\hat{\phi}^4)}$$

.

.

.

$$\hat{y}_{n+m} \pm 1.96\sqrt{MSE(1+\hat{\phi}^2+\hat{\phi}^4\cdots+\hat{\phi}^{2(m-1)})}$$

Better forecast than would be obtained using the standard least squares procedure.

Intro to Time Series Modeling & Forecasting

Questions?