Diagnostic and Remedial Measures

Paweł Polak

September 20, 2017

Linear Regression Models - Lecture 4

Content: ALRM Book Chapter 3 (Sec. 3.1-3.11)

- Diagnostic for Predictor Variable
- Residuals
- Diagnostic for Residuals
- Overview of Tests Involving Residuals
- Kolomogorov-Smirnov Test and Correlation Test for Normality
- Test for Constancy of Error Variance
- F Test for Lack of Fit
- Overview of Remedial Measures
- Transformations
- Exploration of Shape of Regression Function

Remedial Measures

- How do we know that the (linear) regression function is a good explainer of the observed data?
  - Plotting
  - Tests
- What if it is not? What can we do about it?
  - Transformation of variables

Graphical Diagnostics for the Predictor Variable

Goal: We need diagnostic information about X to see if there are any outlying X_i values that could influence the appropriateness of the fitted regression function.

Tools:

- Dot Plot

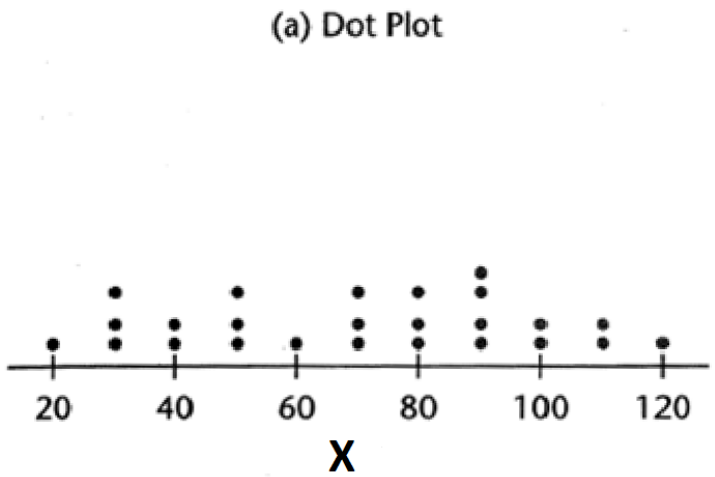
  - Useful for visualizing distribution of inputs

- Sequence Plot

  - Useful for visualizing dependencies between error terms

- Box Plot - Useful for visualizing distribution of inputs

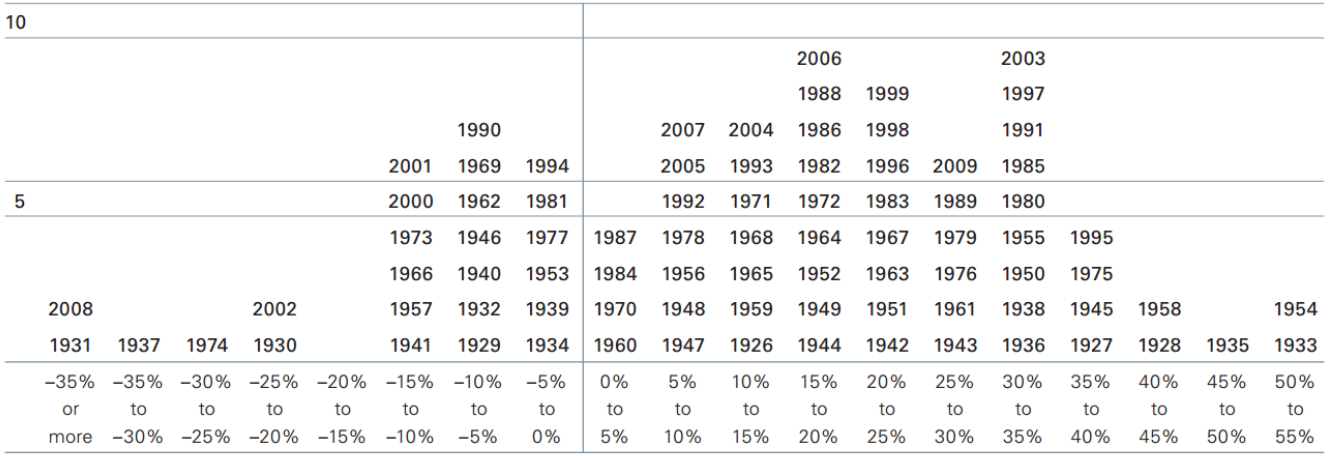
Dot Plot



- How many observations per input value?
- Range of inputs?

Dot Plot

Figure 1. Range of calendar-year returns for U.S. stocks: 1926 through 2009

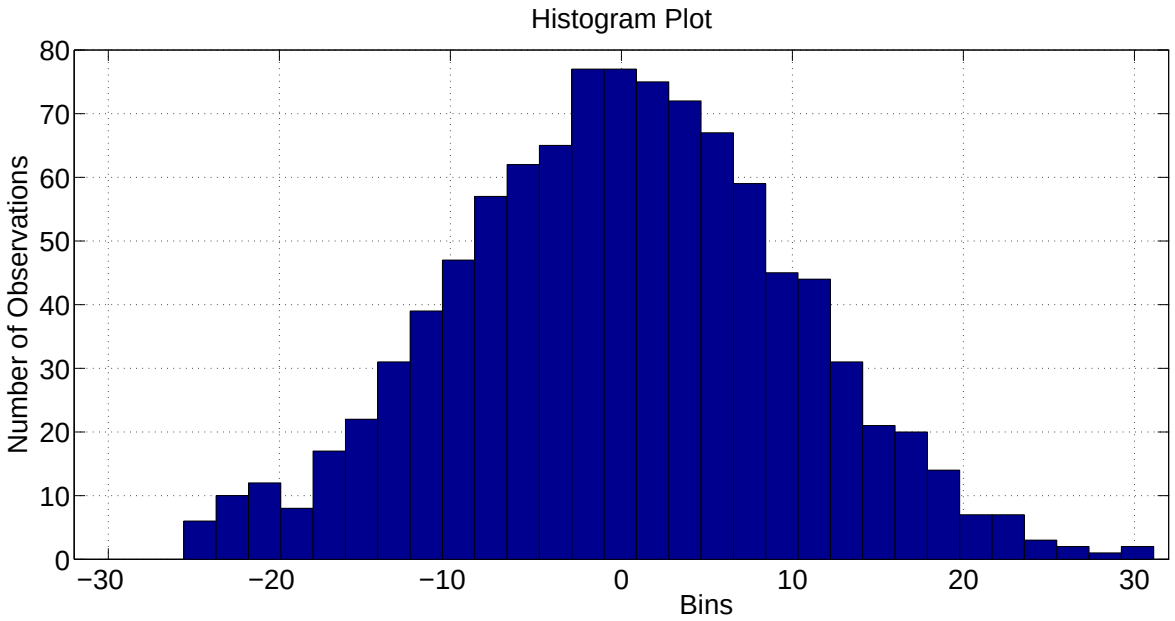


Notes: All returns are in nominal U.S. dollars. For benchmark data, see box on page 2.

Sources: Vanguard calculations, using data from Standard & Poor's, Wilshire, and MSCI.

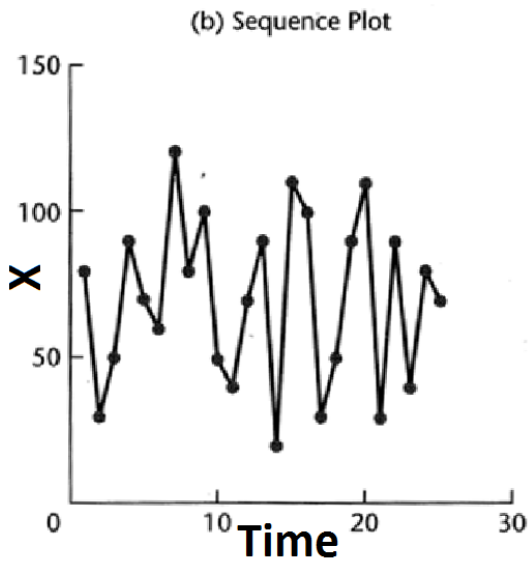
- How many observations per input value?
- Range of inputs?

Histogram Plot



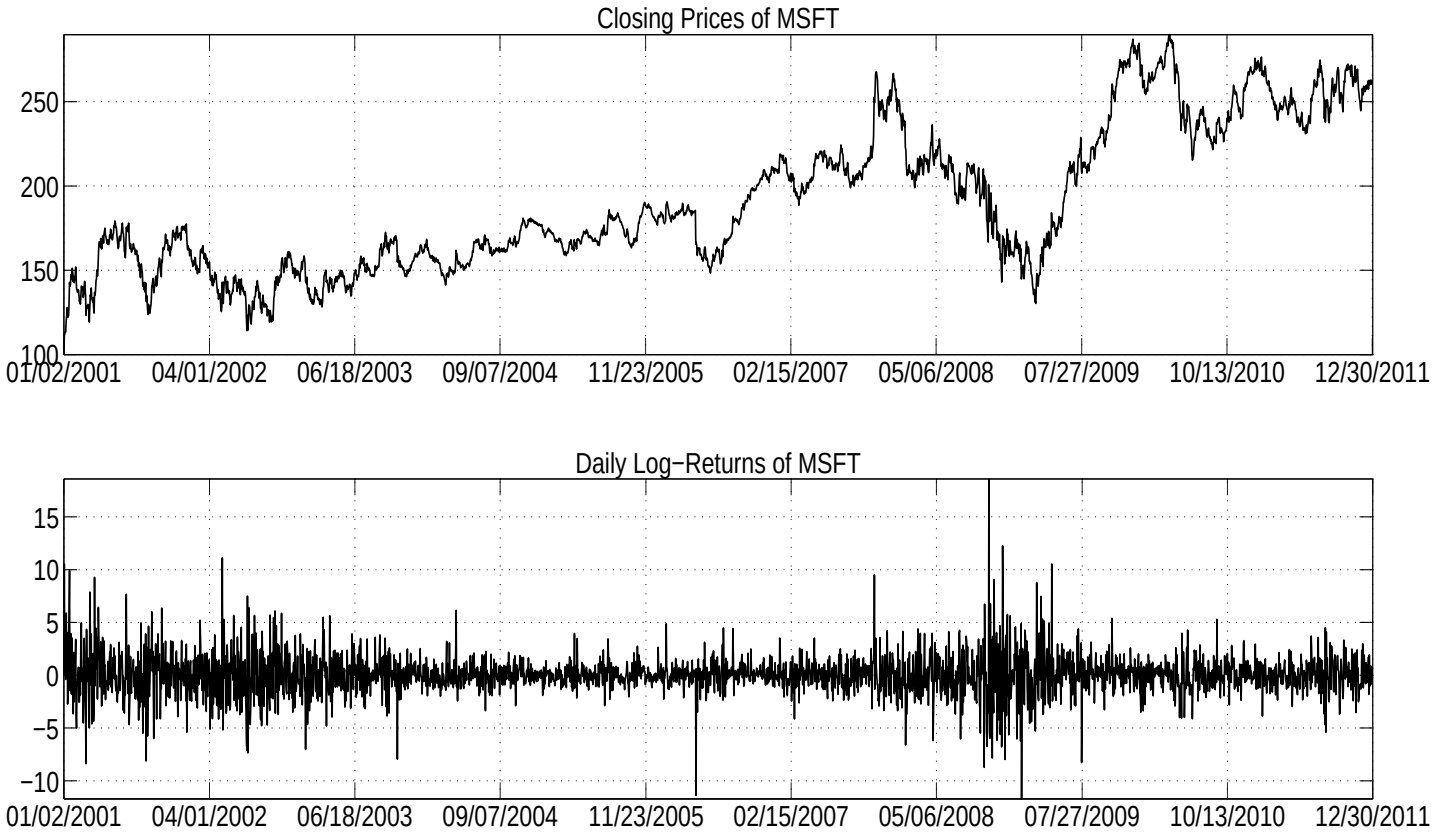
- How many observations per input value?
- Range of inputs?

Sequence Plot



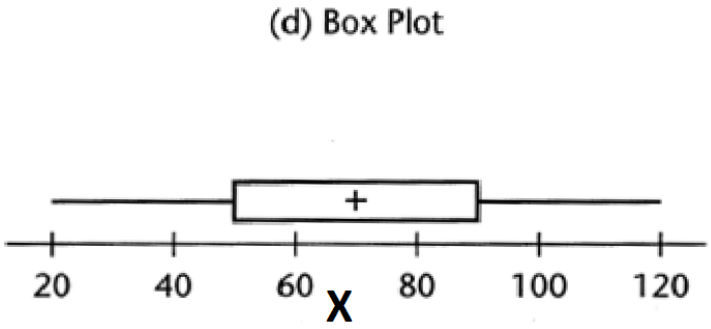
- If observations are made over time, is there a correlation between input and position in observation sequence?

Sequence Plot



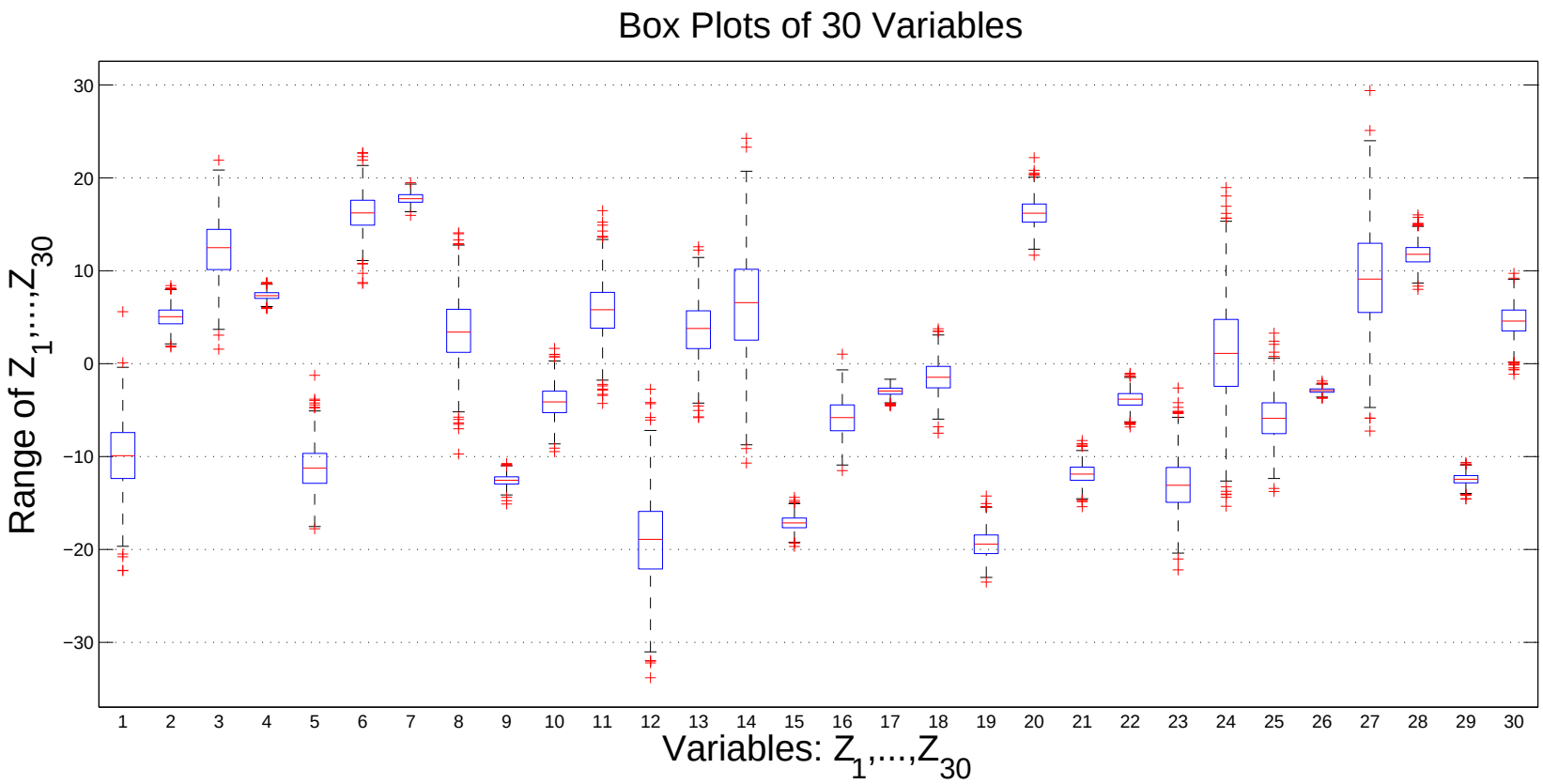
- If observations are made over time, is there a correlation between input and position in observation sequence?
- Visualize the effect of the transformation, here prices vs. returns $R_t = 100\left(\log\left(P_t\right) - \log\left(P_{t-1}\right)\right)$.
- Is the data homoscedastic (constant variance)?

Box Plot



- Shows
  - Median
  - 1st and 3rd quartiles
  - Maximum and minimum
  - Symmetry or not
  - tails

Box Plots



Box plots are useful to visualize the marginal distributions when there are many variables in the model.

Residuals

Diagnostics for the response variable are usually carried out indirectly through the examination of the residuals.

Why? Because the values of Y are influenced by the values of X and we want to focus on what is not explained by X.

- Recall the definition of residuals:

$$e_i = Y_i - \hat{Y}_i$$

- And the difference between that and the unknown true error

$$\varepsilon_i = Y_i - E(Y_i)$$

- In a normal regression model the ε_i's are assumed to be iid $N(0, \sigma^2)$ random variables. The observed residuals e_i should reflect these properties.

Residuals Properties

- Mean

$$\bar{e} = \frac{\sum_{i=1}^{N} e_i}{N} = 0 \quad \text{(always!)}$$

  - so it provides no information if the unobserved ε's have mean 0.

- Variance

$$s^2 = \frac{\sum_{i=1}^{N} (e_i - \bar{e})^2}{N-2} = \frac{\sum_{i=1}^{N} e_i^2}{N-2} = \frac{SSE}{N-2} = MSE$$

  - if the model is appropriate, then MSE is an unbiased estimator of σ^2.

Nonindependence of Residuals

- The residuals e_i are not independent random variables - The fitted values $\hat{Y}_i$ are based on the same fitted regression line.

  - The residuals are subject to two constraints

    1. - Sum of the e_i's equals 0
    2. - Sum of the products $X_i e_i$'s equals 0

- When the sample size is large in comparison to the number of parameters in the regression model, the dependency effect among the residuals e_i can be safely ignored.

Definition: semistudentized residuals

- It may be useful sometimes to look at a standardized set of residuals, for instance in outlier detection.
- Like usual, since the standard deviation of ε_i is σ (itself estimated by square root of MSE) a natural form of standardization to consider is

$$e_i^* = \frac{e_i}{\sqrt{MSE}}$$

- This is called a semistudentized residual, because the residuals, unlike the errors, do not all have the same variance.

$$\text{Var}(e_i) = \sigma^2 \left(1 - \frac{1}{N} - \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{N} (X_i - \bar{X}_i)^2}\right).$$

Recall, in Lecture 4 we derived the prediction variance $\sigma^2\{\text{pred}\} = \sigma^2\left\{Y_{h(\text{new})} - \hat{Y}_h\right\}$. It is equal to

$$\sigma^2\{\text{pred}\} = \sigma^2 \left(1 + \frac{1}{N} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{N} (X_i - \bar{X}_i)^2}\right).$$

Departures from Model...

There are 6 important departures from Simple Linear Regression model with normal errors which can be studied by residuals:

- Regression function is not linear
- Error terms do not have constant variance
- Error terms are not independent
- Model fits all but one or a few outlier observations
- Error terms are not normally distributed
- One or more predictor variables have been omitted from the model

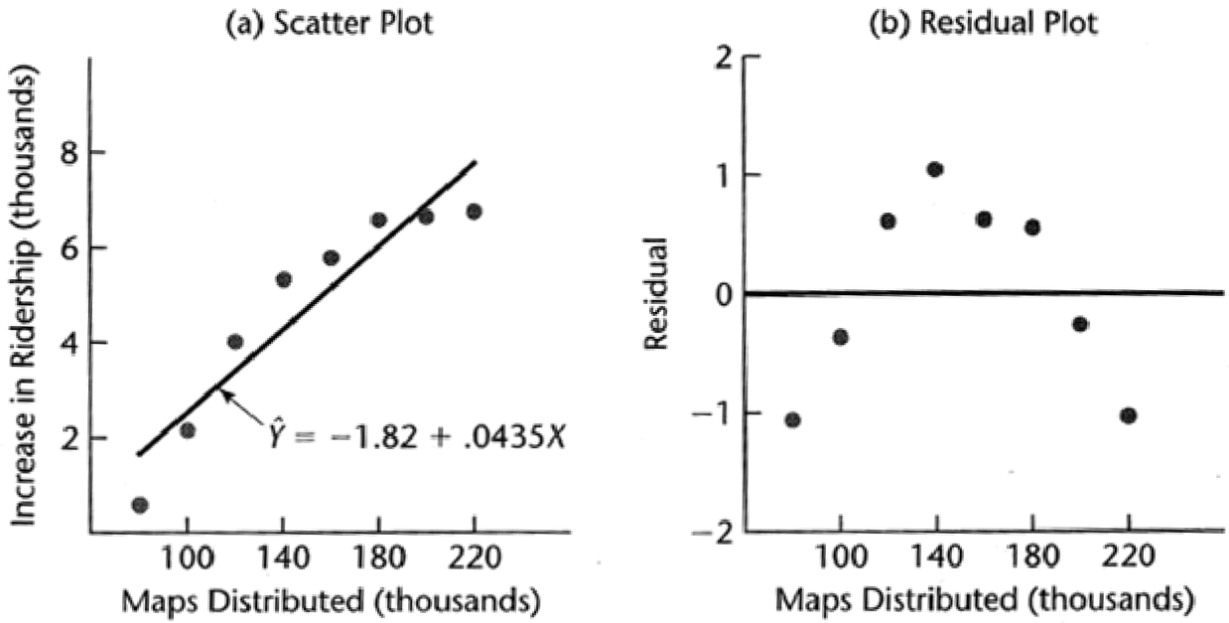
Diagnostics for Residuals

They provide information on whether any of the six types of departure from the Simple Linear Regression Model, given in the previous slide, is present

- Plot of residuals against predictor variable
- Plot of absolute or squared residuals against predictor variable
- Plot of residuals against fitted values
- Plot of residuals against time or other sequence
- Plot of residuals against omitted predictor variables
- Box plot of residuals
- Normal probability plot of residuals

1. Test for nonlinearity of Regression Function: Residual Plot against the predictor variable

Figure: Transit example : ridership increase vs. num. maps distributed

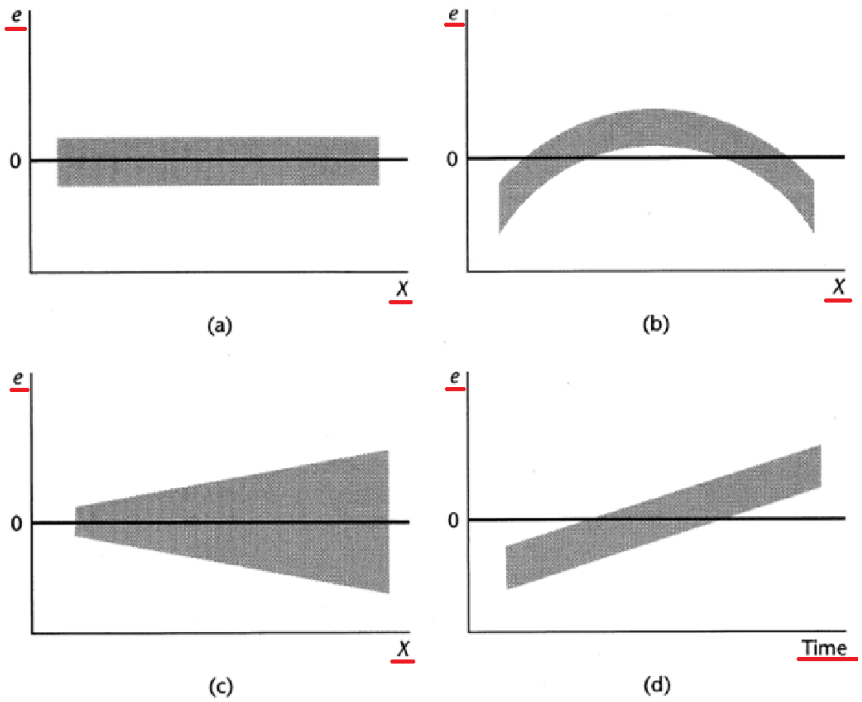


Should be no systematic relationship between residual and predictor variable if it is linearly related.

Here we plot residuals as a function of the predictor X in a multivariate regression with more predictors we can plot the residuals as a function of $\widehat{Y}$. Since $\widehat{Y}$ is a linear combination of X's we will have the same information.

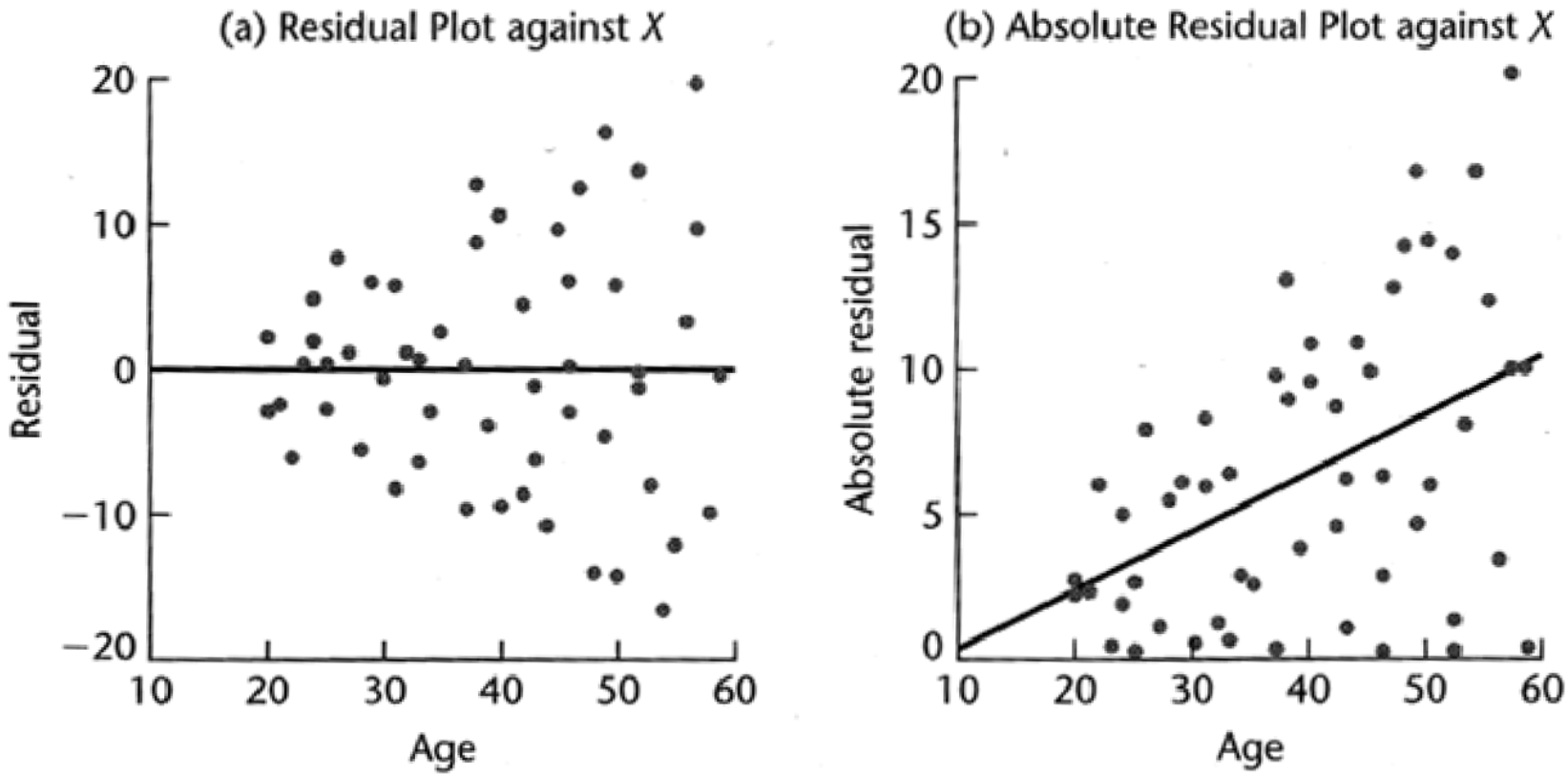
Prototype Residual Plots

Figure: Indicate residual plots



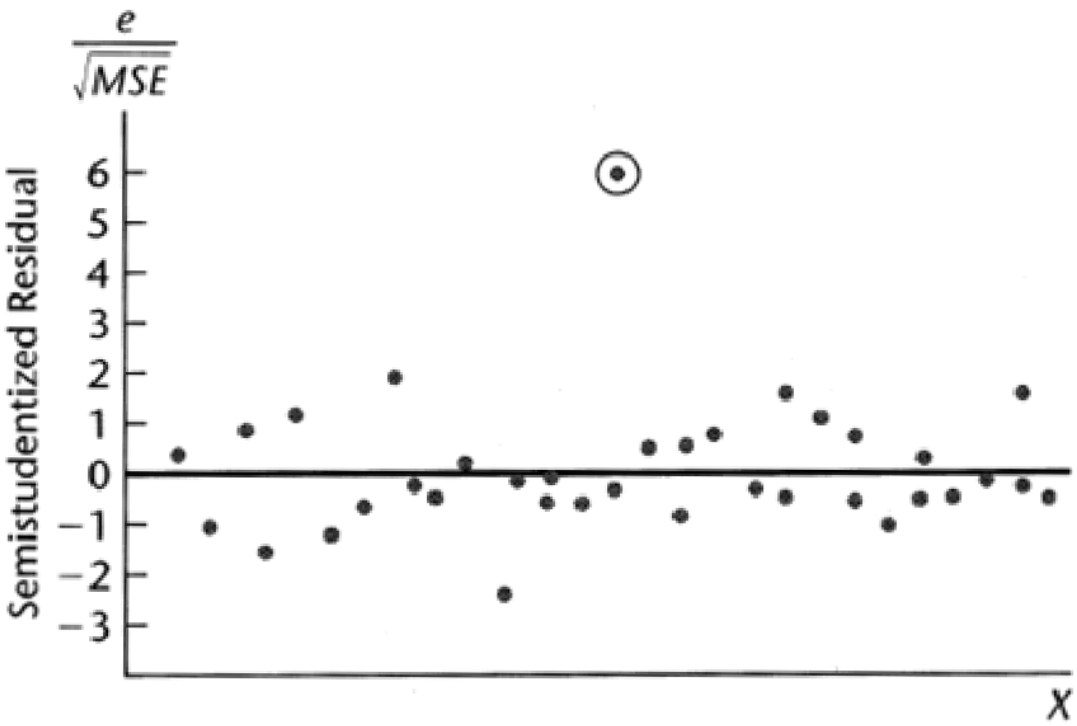
(a) appropriate; (b) need for curvielinear regression function;
(c) heteroskedasticity; (d) nonindependence of the error terms.

2. Nonconstancy of Error Variance



Y - blood preasure of a healthy, adult woman vs. X - her age.

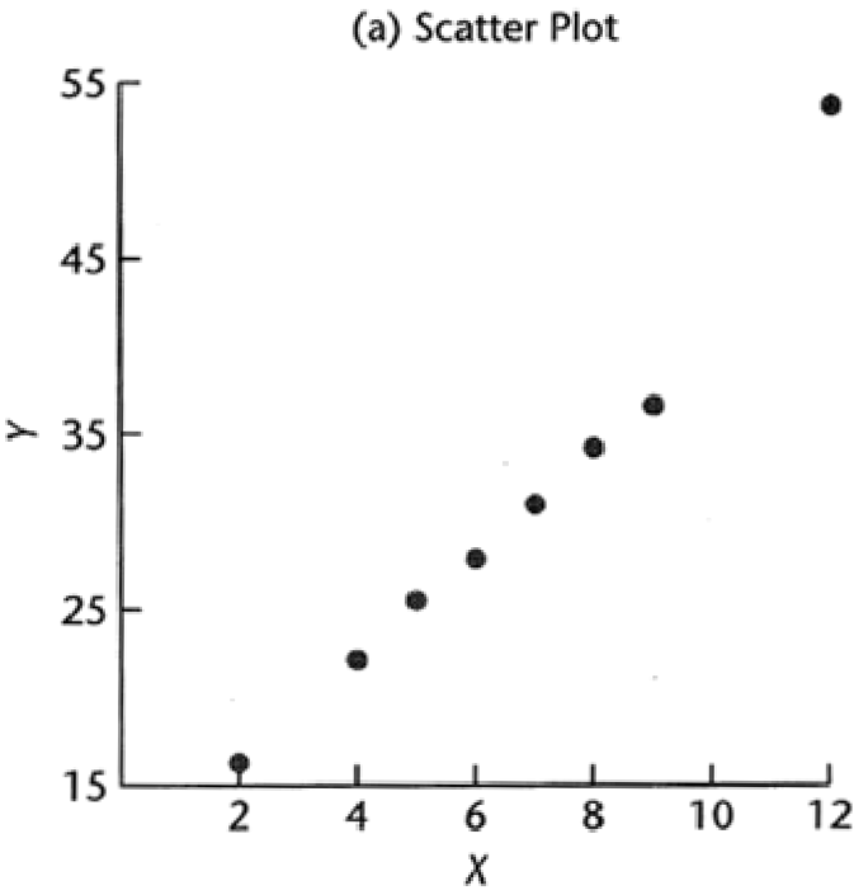
3. Presence of Outliers

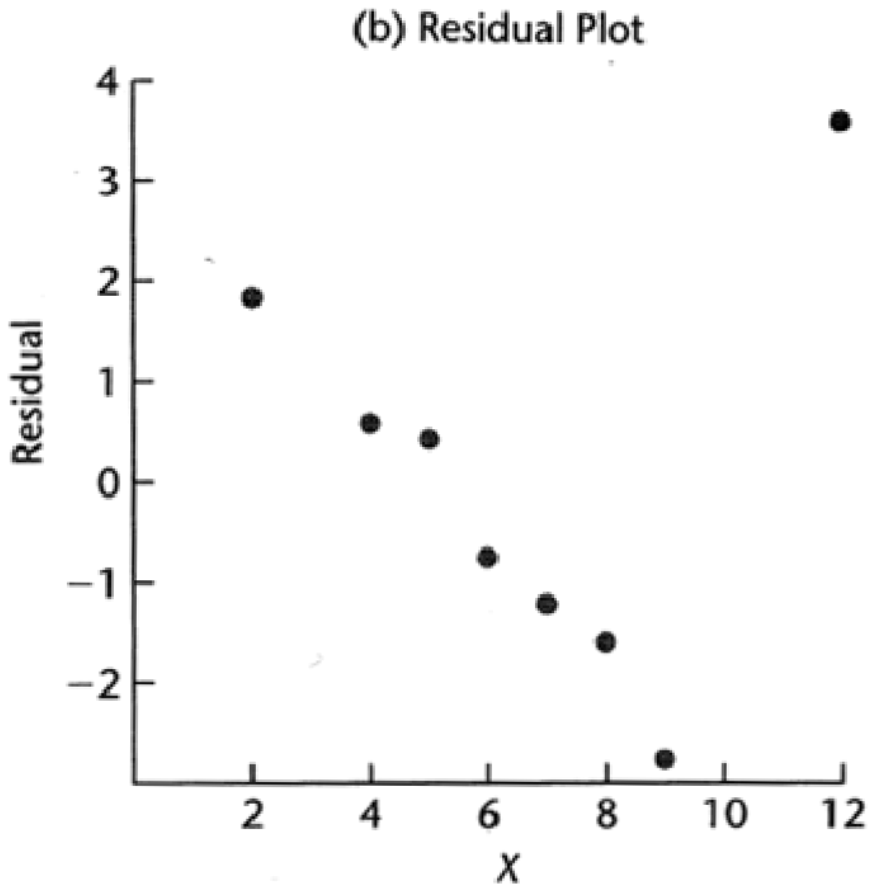


Outliers can strongly effect the fitted values of the regression line.
Rule of thumb:

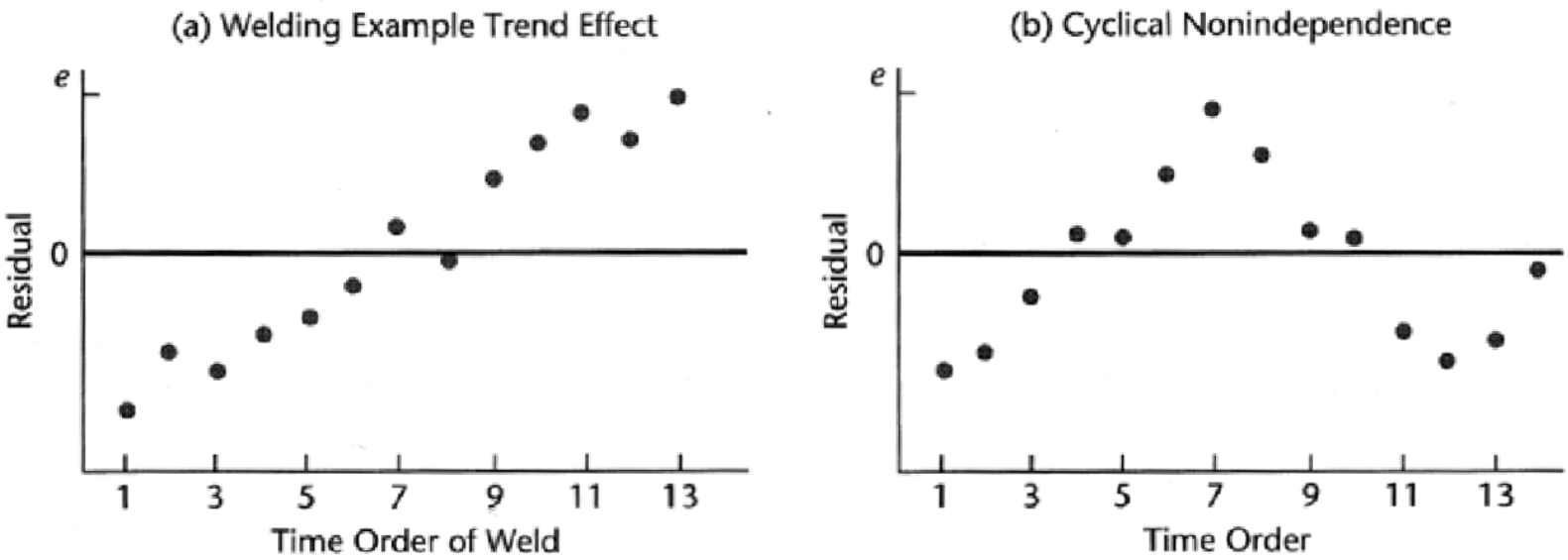
If $\frac{|e_i|}{\sqrt{MSE}} \geq 4$, say it is an outlier.

Outlier effect on residuals





4. Nonindependence of Error Terms

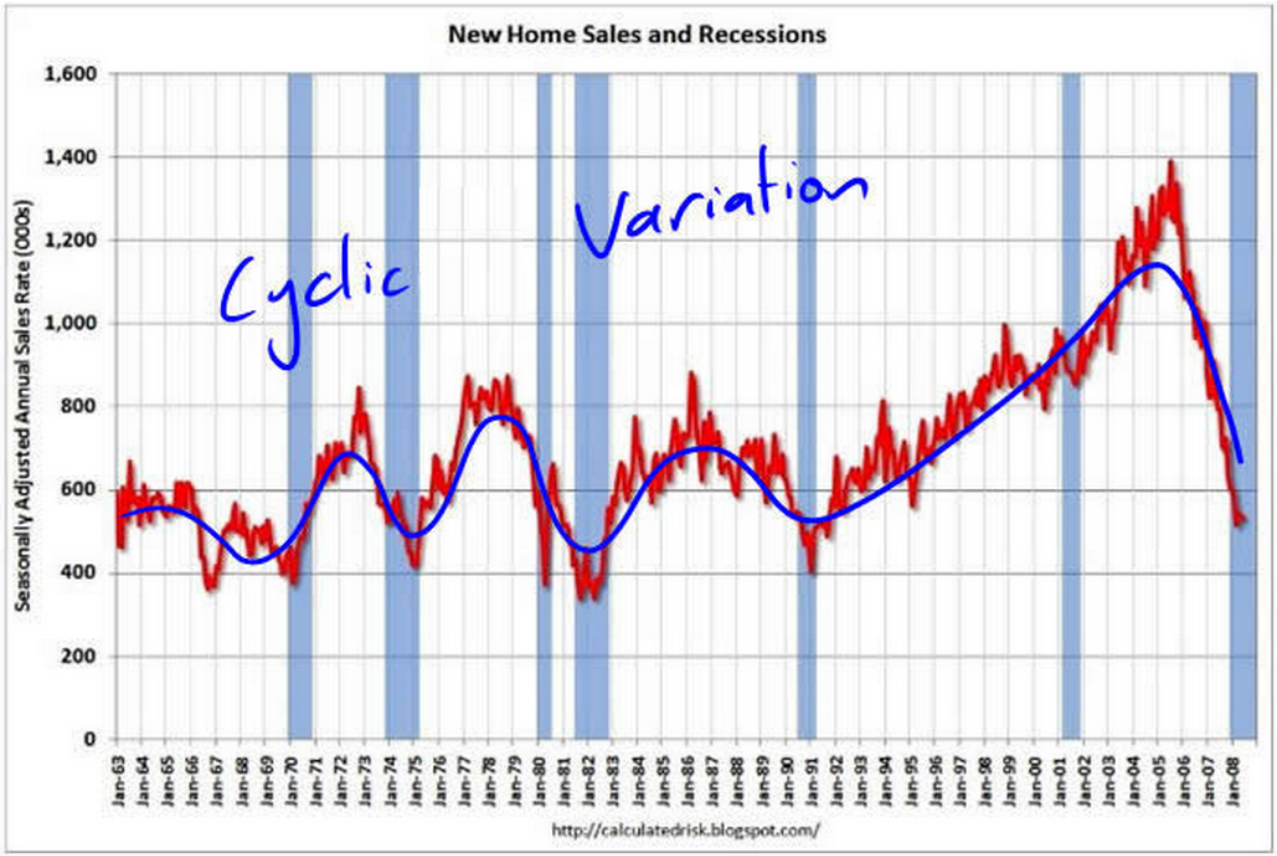


(a) X-diameter of a weld vs. Y-the shear strength of the weld. (some effect connected with time was present, e.g., learning effect of the subject)

(b) a simulation based example of cyclical behaviour (e.g. time of the year effect and the price of electricity)

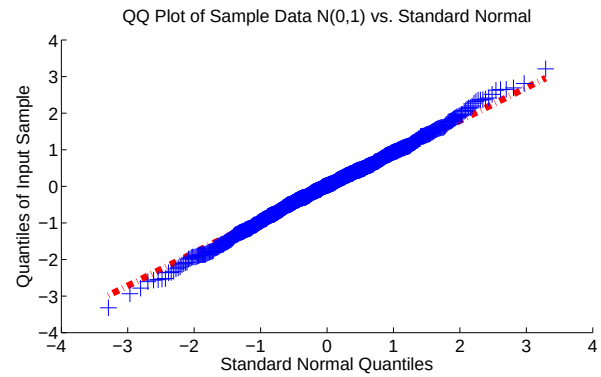
Sequential observations can exhibit observable trends in error distribution. Application: Time series to detect different time trends.

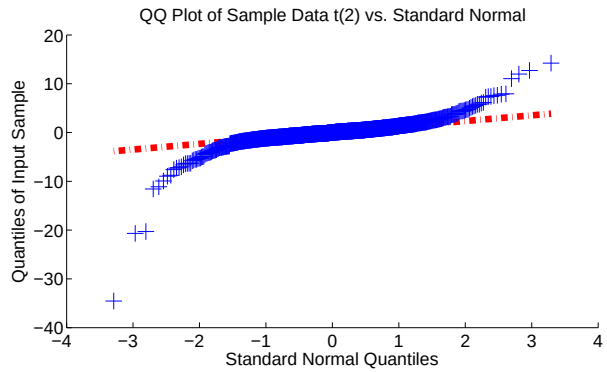
Seasonal Behaviour and Cycles

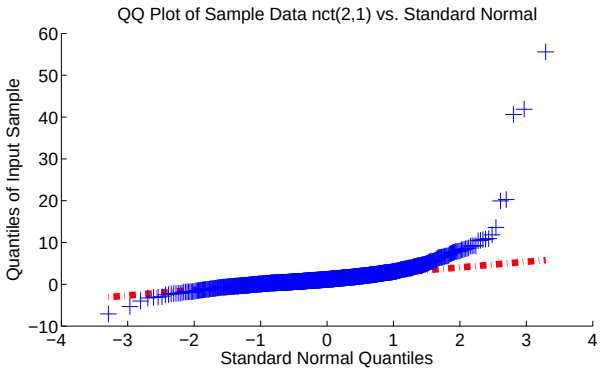


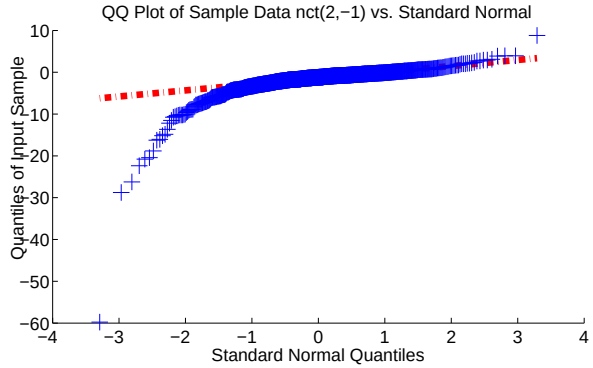
5. Non-normality of Error Terms

- Distribution plots. (e.g., boxplot of the residuals - information about possible asymmetry and residuals)
- Comparison of Frequencies. (e.g., 68 percent of the residuals fall between $\pm\sqrt{MSE}$ or about 90 percent between $\pm 1.645\sqrt{MSE}$.
- Normal probability plot, i.e., $Q-Q$ plot with numerical quantiles on the horizontal axis.



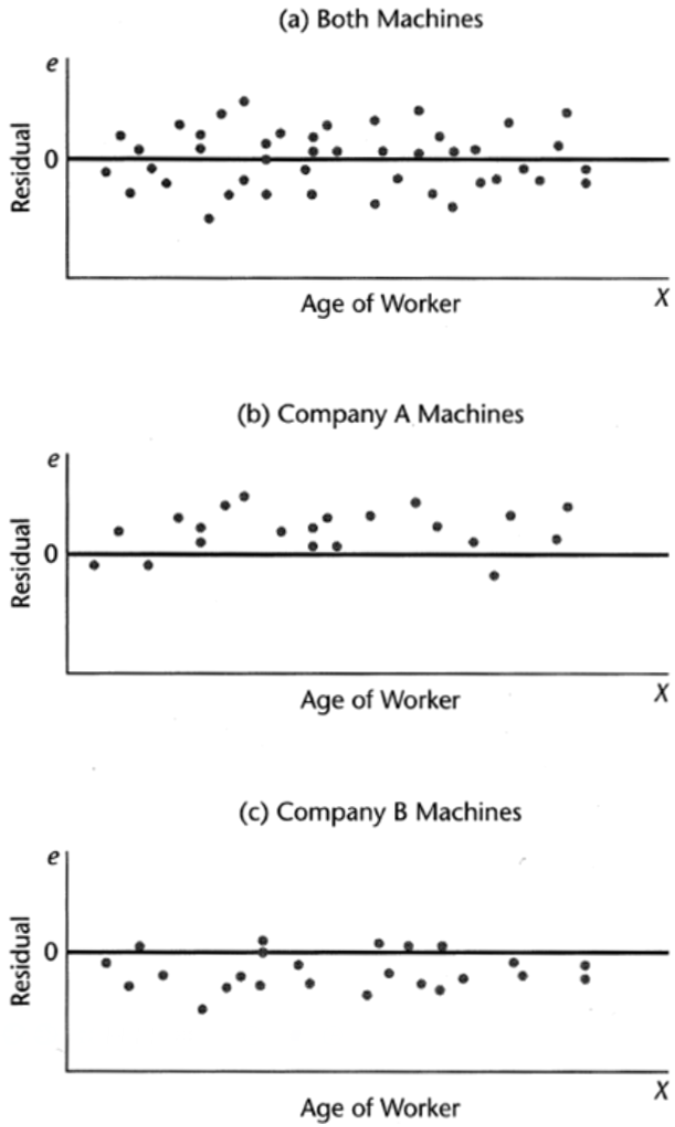






6. Omission of Important Predictor Variables

- Example
  - Y production output
  - X age of worker
  - omitted is the qualitative variable company A vs. B
- Partitioning data can reveal dependence on omitted variable(s)
- Works for quantitative variables as well
  - one has to plot the residuals against the omitted quantitative variable
- Can suggest that inclusion of other inputs is important



Tests Involving Residuals

- Tests for randomness (run test, Durbin-Watson test, Chapter 12)
- Tests for constancy of variance (Brown-Forsythe test, Breusch-Pagan test, Section 3.6)
- Tests for outliers (fit a new regression line to the other $N-1$ observations. Details in Chapter 10)
- Tests for normality of error distribution (will discuss now.)

Kolmogorov-Smirnov Test

The KS test is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K-S test).

The KS statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution. It is given by

$$D_N = \sup_x \left| F_N(x) - F(x) \right|,$$

where $F_N(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{(-\infty, x]}(X_i)$ is the empirical distribution function of N iid observations X_i.

The goodness-of-fit test or the KS test is constructed by using the critical values of the Kolmogorov distribution. The null hypothesis is rejected at level α if $\sqrt{N} D_N > K_\alpha$, where K_α is found from $\mathbb{P}(K \leq K_\alpha) = 1 - \alpha$.

Correlation Test for Normality of Error Distribution

For one way to run this correlation test let

$$e = \{e_{\sigma 1}, ..., e_{\sigma N}\}$$

be the ordered sequence (from smallest to the largest) of the observed errors. Let

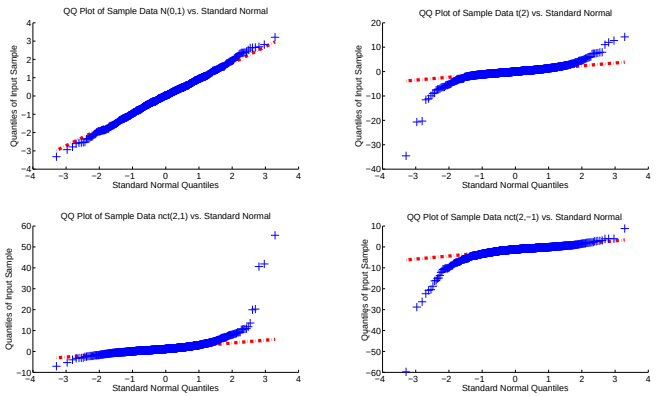
$$r = \{r_1, ..., r_k, ..., r_N\},$$

where r_k is the expected value of the k^{th} residual under the normality assumption, i.e.

$$r_k \approx \sqrt{MSE}[z(\frac{k - .375}{N + .25})]$$

Then compute the sample correlation between e and r.

Recall the Q-Q plot



Correlation Test for Normality of Error Distribution

- formal test for the normality of the error terms can be developed in terms of the coefficient of correlation between the residuals e_i and their expected values under normality. High value indicates normality!
- Tables (B.6 in the book) gives critical values for the null hypothesis (normally distributed errors).
- Less than the critical value, reject the null hypothesis!

Tests for Constancy of Error Variance

- Brown-Forsythe test does not depend on normality of error terms. The Brown-Forsythe test is applicable to simple linear regression when
  - The variance of the error terms either increases or decreases with X ("megaphone" residual plot)
  - Sample size is large enough to ignore dependencies between the residuals
- The Brown-Forsythe test is essentially a t-test for testing whether the means of two normally distributed populations are the same where the populations are the absolute deviations between the prediction and the observed output space in two non-overlapping partitions of the input space.

Brown-Forsythe Test

- Divide X into X_1 (the low values of X) and X_2 (the high values of X)
- Let e_{i1} be the i-th residual for X_1 and vice versa
- $\tilde{e}_1$ and $\tilde{e}_2$ denote the median of the residuals.
- Let $N = N_1 + N_2$
- The Brown-Forsythe test uses the absolute deviations of the residuals around their group median

$$d_{i1} = |e_{i1} - \tilde{e}_1|, d_{i2} = |e_{i2} - \tilde{e}_2|$$

Brown-Forsythe Test

- The test statistic for comparing the means of the absolute deviations of the residuals around the group medians is

$$t^*_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

where the pooled variance

$$s^2 = \frac{\sum_{i=1}^{N}(d_{i1} - \bar{d}_1)^2 + \sum_{i=1}^{N}(d_{i2} - \bar{d}_2)^2}{N - 2}$$

Brown-Forsythe Test

- If N_1 and N_2 are not extremely small

$$t^*_{BF} \sim t(N-2)$$

approximately

- From this confidence intervals and tests can be constructed.

Breusch-Pagan Test

- Another test for the constancy of error variance.
- Assume error terms are independently and normally distributed and

$$\log \sigma_i^2 = \gamma_0 + \gamma_1 X_i$$

- Then the problem reduces to

$$H_0 : \gamma_1 = 0 \text{ v.s. } H_1 : \gamma_1 \neq 0$$

- Test statistics X^2_{BP} is derived in the following two steps
  1. Regress Y on X, get SSE and residual vector e.
  2. Regress e^2 on X, get SSR^*.
- Under H_0, when N is reasonably large,

$$X^2_{BP} = \frac{SSR^2}{2(\frac{SSE}{N})^2} \sim \chi^2(1)$$

- Decision rule:

  If $X^2_{BP} \leq \chi^2(1-\alpha;1)$, then conclude H_0, i.e., constant error variance

  If $X^2_{BP} > \chi^2(1-\alpha;1)$, then conclude H_1, i.e., not constant error variance

- Direct function in R: ncvTest in the package car.

F test for lack of fit

- Formal test for determining whether a specific type of regression function adequately fits the data.
- Assume the observations $Y|X$ are
  1. independent
  2. normally distributed
  3. same variance σ^2
- Requires: repeat observations at one or more X levels (called replicates)

Example

- 11 similar branches of a bank offered gifts for setting up money market accounts
- Minimum initial deposits were specific to qualify for the gift
- Value of gift was proportional to the specified minimum deposit
- Interested in: relationship between specified minimum deposit and number of new accounts opened

F Test Example Data and ANOVA Table

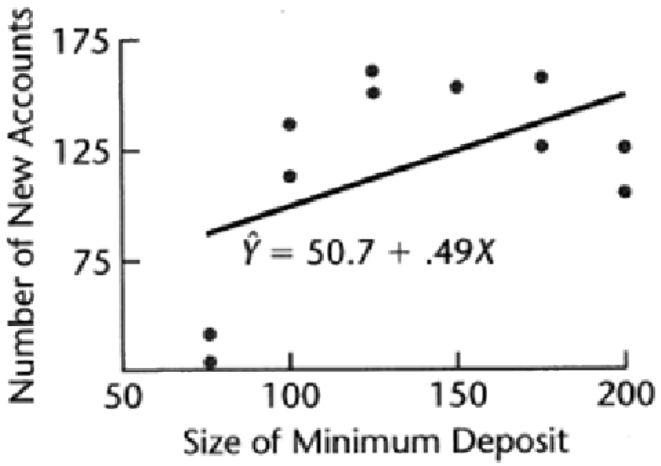
(a) Data

Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i	Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i
1	125	160	7	75	42
2	100	112	8	175	124
3	200	124	9	125	150
4	75	28	10	200	104
5	150	152	11	100	136
6	175	156			

(b) ANOVA Table

Source of Variation	*SS*	*df*	*MS*
Regression	5,141.3	1	5,141.3
Error	14,741.6	9	1,638.0
Total	19,882.9	10	

Fit


Number of New Accounts
175
125
75
$\hat{Y} = 50.7 + .49X$
50
100
150
200
Size of Minimum Deposit

Data Arranged To Highlight Replicates

	Size of Minimum Deposit (dollars)					
Replicate	$j=1$ $X_1=75$	$j=2$ $X_2=100$	$j=3$ $X_3=125$	$j=4$ $X_4=150$	$j=5$ $X_5=175$	$j=6$ $X_6=200$
$i=1$	28	112	160	152	156	124
$i=2$	42	136	150		124	104
Mean $\bar{Y}_j$	35	124	155	152	140	114

- The observed value of the response variable for the i-th replicate for the j-th level of X is Y_{ij}
- The mean of the Y observations at the level $X = X_j$ is $\bar{Y}_j$

Full Model vs. Regression Model

- The full model is

$$Y_{ij} = \mu_j + \epsilon_{ij}$$

where

1. μ_j are parameters, $j = 1, ..., c$
2. ϵ_{ij} are iid $N(0, \sigma^2)$

- Since the error terms have expectation zero

$$E(Y_{ij}) = \mu_j$$

Full Model

- In the full model there is a different mean (a free parameter) for each X_i
- In the regression model the mean responses are constrained to lie on a line

$$E(Y) = \beta_0 + \beta_1 X$$

Fitting the Full Model

- The estimators of μ_j are simply

$$\hat{\mu}_j = \bar{Y}_j$$

- The error sum of squares of the full model therefore is

$$SSE(F) = \sum_{i=1}^{N} \sum_{j=1}^{N} (Y_{ij} - \bar{Y}_j)^2 = SSPE$$

SSPE: Pure Error Sum of Squares

Degrees of Freedom

- Ordinary total sum of squares had N-1 degrees of freedom.
- Each of the j terms is a ordinary total sum of squares

  - Each then has $N_j - 1$ degrees of freedom

- The number of degrees of freedom of SSPE is the sum of the component degrees of freedom

$$df_F = \sum_j (N_j - 1) = \sum_j N_j - c = N - c$$

General Linear Test

- Remember: the general linear test proposes a reduced model

$$H_0 : E(Y) = \beta_0 + \beta_1 X \, (\textit{Normal regression model})$$

$$H_1 : E(Y) \neq \beta_0 + \beta_1 X \, (\textit{The full model, one independent mean for each level of } X)$$

SSE For Reduced Model

The SSE for the reduced model is as before:

- Remember

$$SSE = \sum_i \sum_j [Y_{ij} - (b_0 + b_1 X_j)]^2 = \sum_i \sum_j \left(Y_{ij} - \hat{Y}_{ij}\right)^2$$

- and has $N - 2$ degrees of freedom $dfR = N - 2$

SSE(R)

(a) Data

Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i	Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i
1	125	160	7	75	42
2	100	112	8	175	124
3	200	124	9	125	150
4	75	28	10	200	104
5	150	152	11	100	136
6	175	156			

(b) ANOVA Table

Source of Variation	*SS*	*df*	*MS*
Regression	5,141.3	1	5,141.3
Error	14,741.6	9	1,638.0
Total	19,882.9	10	

F Test Statistic

From the general linear test approach

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$
$$= \frac{SSE - SSPE}{(N-2) - (N-c)} \div \frac{SSPE}{N-c}$$

Lack of fit sum of squares:

$$SSLF = SSE - SSPE$$

Then

$$F^* = \frac{SSLF}{(N-2) - (N-c)} \div \frac{SSPE}{N-c} = \frac{MSLF}{MSPE}$$

F Test Rule

- From the F test we know that large values of F^* lead us to reject the null hypothesis:

  If $F^* \leq F(1-\alpha; c-2, N-c)$, *conclude* H_0

  If $F^* > F(1-\alpha; c-2, N-c)$, *conclude* H_a

- For this example we have

$$SSPE = 1{,}148.0 \qquad n - c = 11 - 6 = 5$$

$$SSE = 14{,}741.6$$

$$SSLF = 14{,}741.6 - 1{,}148.0 = 13{,}593.6 \qquad c - 2 = 6 - 2 = 4$$

$$F^* = \frac{13{,}593.6}{4} \div \frac{1{,}148.0}{5}$$

$$= \frac{3{,}398.4}{229.6} = 14.80$$

Variance decomposition

$$SSE = SSPE + SSLF.$$

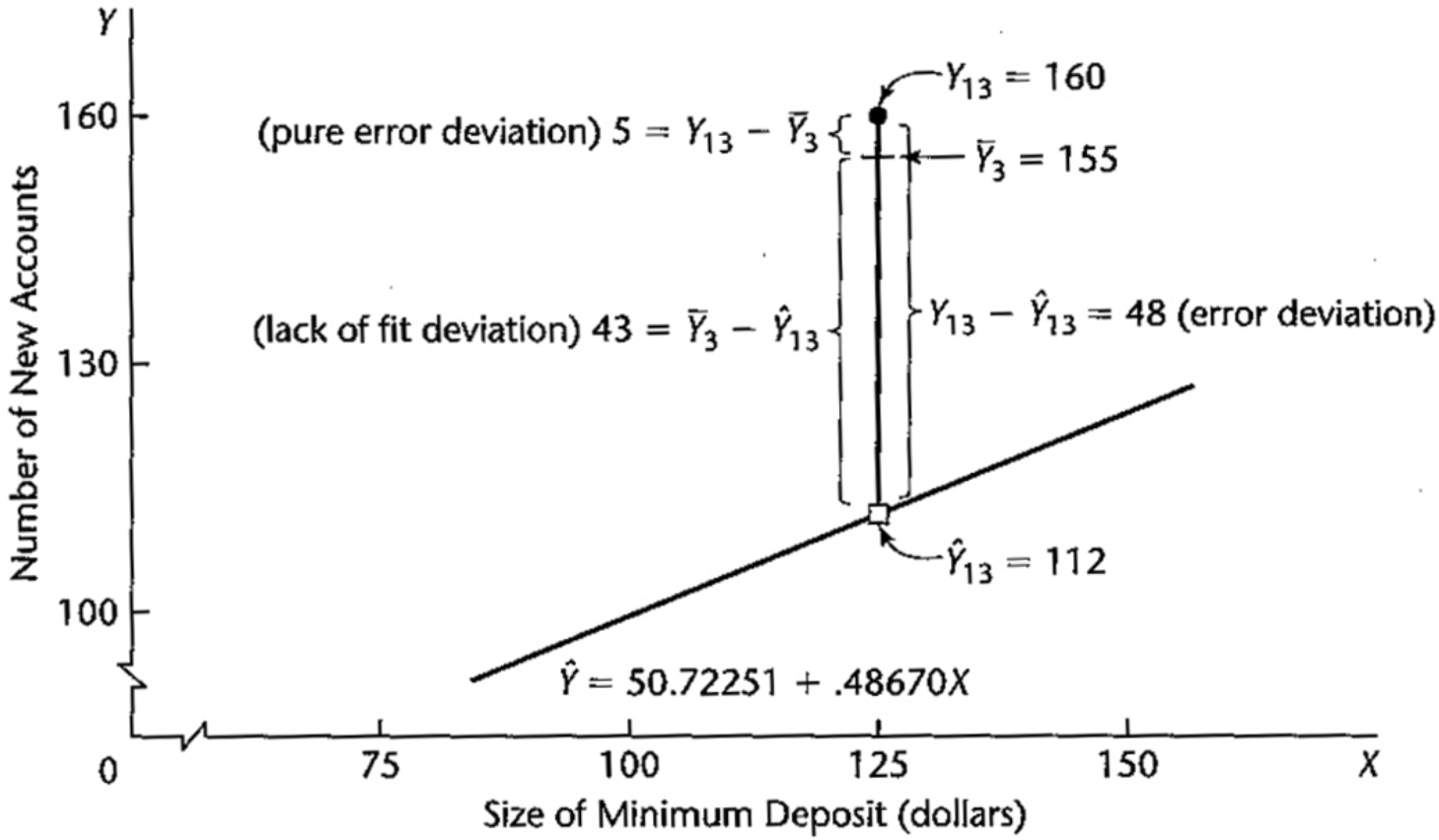
$$\sum_i \sum_j (Y_{ij} - \hat{Y}_{ij})^2 = \sum_i \sum_j (Y_{ij} - \bar{Y}_j)^2 + \sum_i \sum_j (\bar{Y}_j - \hat{Y}_{ij})^2$$

Since $\hat{Y}_{ij} = \hat{Y}_j$ for all i because, for given j, X_{ij} are the same, then

$$SSLF = SSE - SSPE = \sum_i \sum_j (\bar{Y}_j - \hat{Y}_{ij})^2 = \sum_j N_j (\bar{Y}_j - \hat{Y}_j)^2.$$

If the linear regression function is appropriate, then the means $\bar{Y}_j$ will be near the fitted values $\hat{Y}_j$ calculated from the estimated regression function and $SSLF$ would be small. Hence, $SSLF$ measures the lack of fit.

Example decomposition



$$\underbrace{Y_{ij} - \hat{Y}_{ij}}_{\text{Error deviation}} = \underbrace{Y_{ij} - \bar{Y}_j}_{\text{Pure error deviation}} + \underbrace{\bar{Y}_j - \hat{Y}_{ij}}_{\text{Lack of fit deviation}}$$

ANOVA for Lack of Fit of SLR

ANOVA Table for Testing Lack of Fit of Sample Linear Regression Function

(a) General

Source of Variation	*SS*	*df*	*MS*
Regression	SSR	1	MSR
Error	SSE	$n-2$	MSE
Lack of fit	$SSLF$	$c-2$	$MSLF$
Pure error	$SSPE$	$n-c$	$MSPE$
Total	$SSTO$	$n-1$	

(b) Bank Example

Source of Variation	*SS*	*df*	*MS*
Regression	$SSR = 5{,}141.3$	1	$MSR = 5{,}141.3$
Error	$SSE = 14{,}741.6$	9	$MSE = 1{,}638.0$
Lack of fit	$SSLF = 13{,}593.6$	4	$MSLF = 3{,}398.4$
Pure error	$SSPE = 1{,}148.0$	5	$MSPE = 229.6$
Total	$SSTO = 19{,}882.9$	10	

Example Conclusion

- If we set the significance level to $\alpha = .01$
- And look up the value of the F inv-cdf $F(.99, 4, 5) = 11.4$
- We can conclude that the null hypothesis should be rejected.

A review

- Graphical procedures for determining appropriateness of regression fit
  - Various Residual plots
- Tests to determine
  - Constancy of error variance
  - Lack of fit test
- Next topic: what do we do if we determine(through testing or otherwise) that the linear regression fit is not good?

How to fix

If simple regression model is not appropriate then there are two choices:

1. Abandon simple regression model and develop and use a more appropriate model

   e.g., logistic regression, nonparametric regression, etc...

   may yield better insights, but a more complex model lead to more complex procedures for estimating the parameters. (Later in the course)

2. Employ some transformation of the data so that the simple regression model is appropriate for the transformed data. (This chapter)

Fixes For...

- Nonlinearity of regression function – Transformation(s) (today)
- Nonconstancy of error variance – Weighted least squares (Chapter 11) and transformations
- Nonindependence of error terms – Directly model correlation or use first differences (Chapter 12)
- Non-normality of error terms – Transformation(s) (today)
- Omission of Important Predictor Variables – Modify the model Multiple regression analysis, Chapter 6 and later on.
- Outlying observations – Robust regression (Chapter 11)

Nonlinearity of regression function

Direct approach

- Modify regression model by altering the nature of the regression function. For instance, a quadratic regression function might be used

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2$$

- or an exponential function

$$E(Y) = \beta_0 \beta_1^X$$

- Such approaches employ a tranformation to (approximately) linearize a regression function

Quick Questions

- How would you fit such models?
- How does the exponential regression function relate to regular linear regression?
- Where did the error terms go?

Transformations

Transformations for nonlinearity relation only

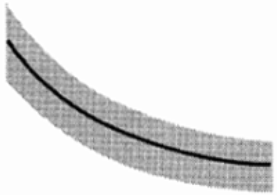
- Appropriate when the distribution of the error terms is reasonably close to a normal distribution

- In this situation

1. transformation of X should be attempted;

2. transformation of Y should not be attempted because it will materially effect the distribution of the error terms.

Prototype Regression Patterns

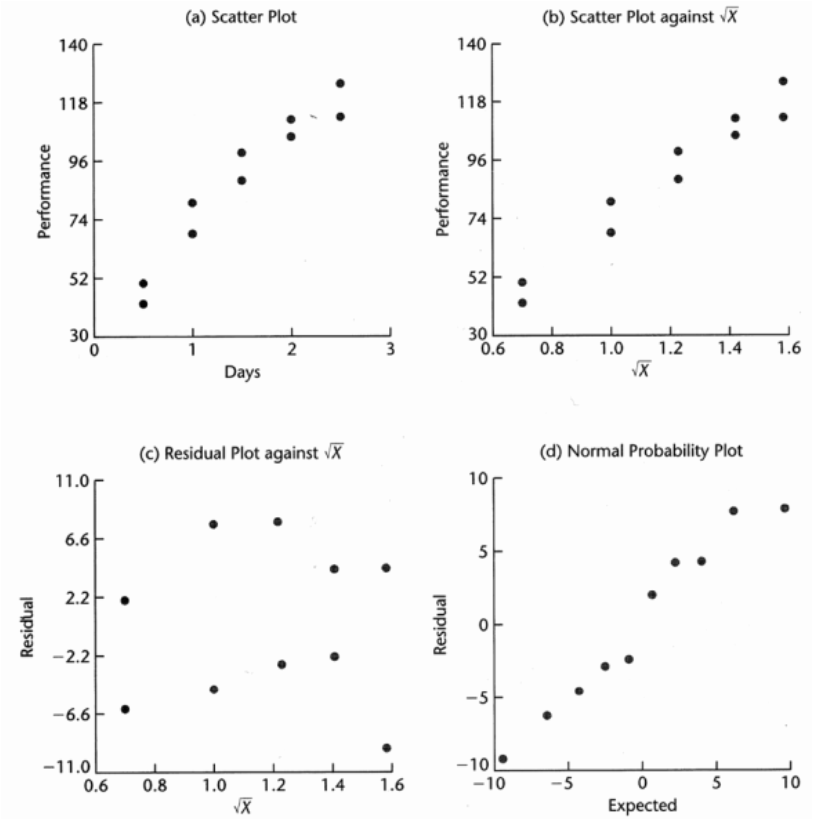
	Prototype Regression Pattern	Transformations of X	
(a)		$X' = \log_{10} X$	$X' = \sqrt{X}$
(b)		$X' = X^2$	$X' = \exp(X)$
(c)		$X' = 1/X$	$X' = \exp(-X)$

Example

Experiment

- X: days of training received
- Y: sales performance(score)

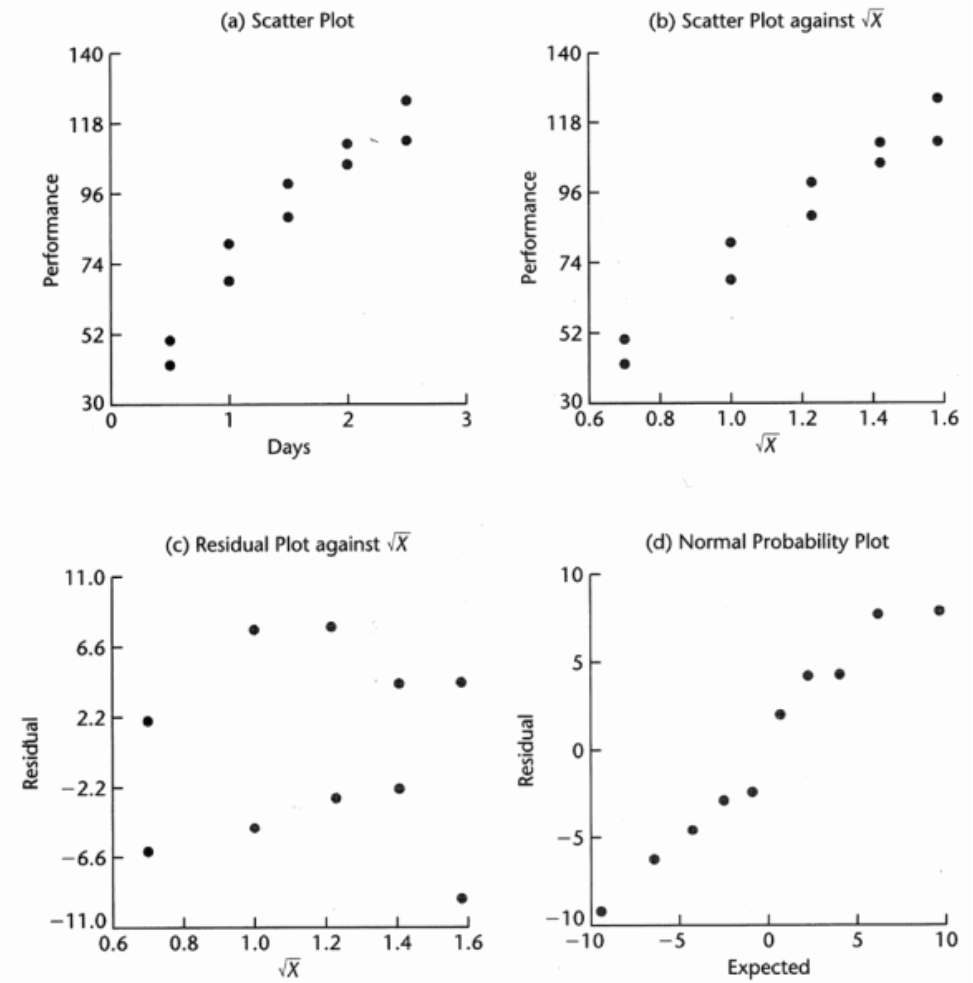
$X' = \sqrt{X}$


(a) Scatter Plot
Performance
Days
(b) Scatter Plot against $\sqrt{X}$
Performance
$\sqrt{X}$
(c) Residual Plot against $\sqrt{X}$
Residual
$\sqrt{X}$
(d) Normal Probability Plot
Residual
Expected

Example Data Transformation

Sales Trainee i	(1) Days of Training X_i	(2) Performance Score Y_i	(3) $X_i' = \sqrt{X_i}$
1	.5	42.5	.70711
2	.5	50.6	.70711
3	1.0	68.5	1.00000
4	1.0	80.7	1.00000
5	1.5	89.0	1.22474
6	1.5	99.6	1.22474
7	2.0	105.3	1.41421
8	2.0	111.8	1.41421
9	2.5	112.3	1.58114
10	2.5	125.7	1.58114

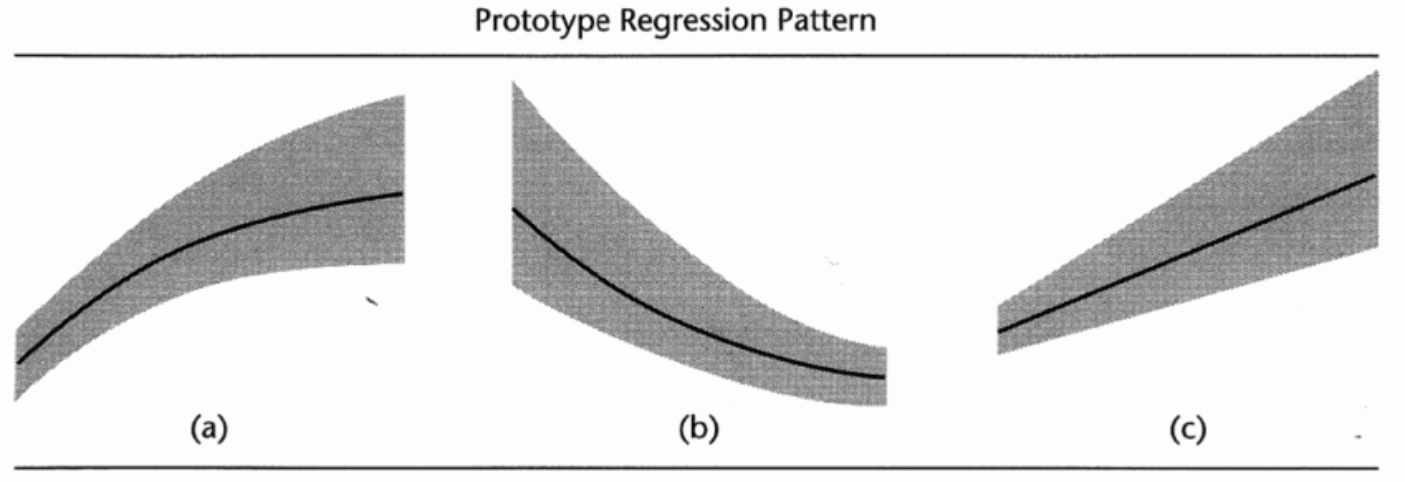
Graphical Residual Analysis



Transformations on Y

- Non-normality and unequal variances of error terms frequently appear together
- To remedy these in the normal regression model we need a transformation on Y
- This is because
  - Shapes and spreads of distributions of Y need to be changed
  - May help linearize a curvilinear regression relation
- Can be combined with transformation on X

Prototype Regression Patterns and Y Transformations



Transformations on Y:

$Y' = \sqrt{Y}$

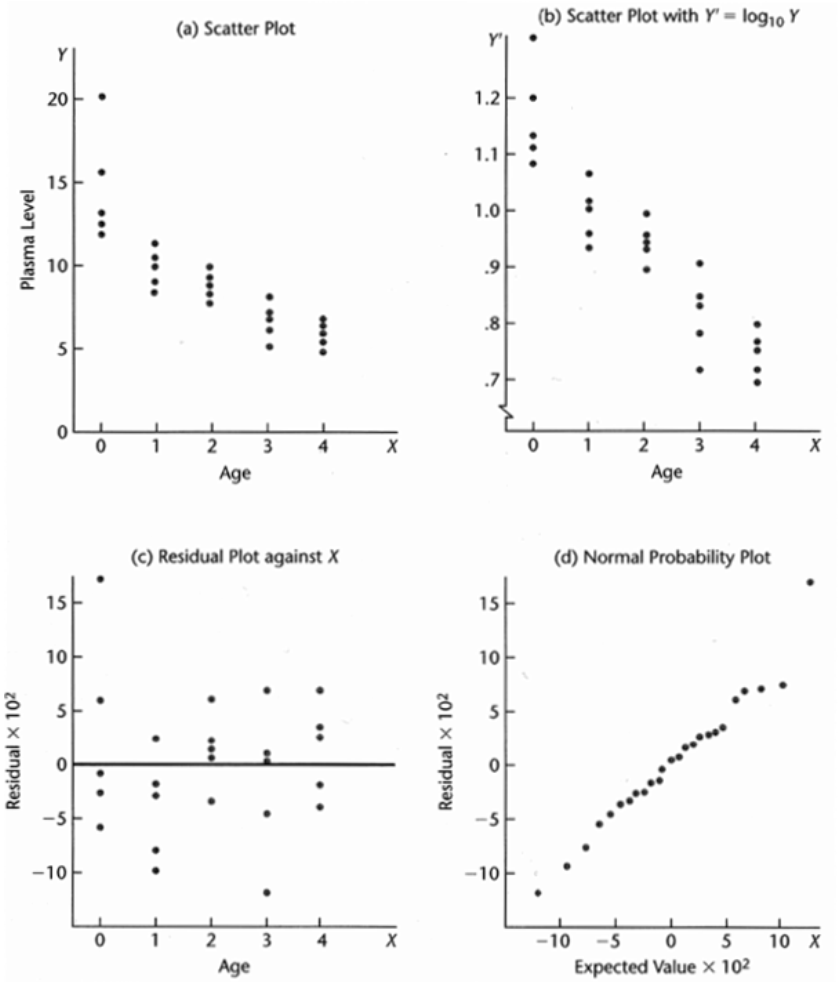
$Y' = log_{10} Y$ (logarithmic transformation of Y is used to linearize regression relations and stabilize error variance)

$Y' = 1/Y$

Example

- Use of logarithmic transformation of Y to linearize regression relations and stabilize error variance.
- Data on age(X) and plasma level of a polyamine (Y) for a portion of the 25 healthy children in a study. Younger children exhibit greater variability than older children.

Plasma Level vs. Age



Associated Data

Child i	(1) Age X_i	(2) Plasma Level Y_i	(3) $Y_i' = \log_{10} Y_i$
1	0 (newborn)	13.44	1.1284
2	0 (newborn)	12.84	1.1086
3	0 (newborn)	11.91	1.0759
4	0 (newborn)	20.09	1.3030
5	0 (newborn)	15.60	1.1931
6	1.0	10.11	1.0048
7	1.0	11.38	1.0561
...	...	...	...
19	3.0	6.90	.8388
20	3.0	6.77	.8306
21	4.0	4.86	.6866
22	4.0	5.10	.7076
23	4.0	5.67	.7536
24	4.0	5.75	.7597
25	4.0	6.23	.7945

Associated Data (Cont')

- if we fit a simple linear regression line to the log transformed Y data we obtain:

$$\hat{Y}' = 1.135 - .1023X$$

- And the coefficient of correlation between the ordered residuals and their expected values under normality is .981(for $\alpha = .05$ Table $B.6$ in the book shows a critical value of .959)
- Normality of error terms supported, regression model for transformed Y data appropriate.

Box-Cox Transformations

- It can be difficult to graphically determine which transformation of Y is most appropriate for correcting
  - skewness of the distributions of error terms
  - unequal variances
  - nonlinearity of the regression function
- The Box-Cox procedure automatically identifies a transformation from the family of power transformations on Y

Box-Cox Transformations

- This family is of the form

$$Y' = Y^{\lambda}$$

- Examples include

$$\begin{aligned}
\lambda = 2 \qquad & Y' = Y^2 \\
\lambda = .5 \qquad & Y' = \sqrt{Y} \\
\lambda = 0 \qquad & Y' = lnY \text{ (by definition)} \\
\lambda = -.5 \qquad & Y' = \frac{1}{\sqrt{Y}} \\
\lambda = -1 \qquad & Y' = \frac{1}{Y}
\end{aligned}$$

Box-Cox Cont.

- The normal error regression model with the response variable a member of the family of power transformations becomes

$$Y_i^{\lambda} = \beta_0 + \beta_1 X_i + \epsilon_i$$

- This model has an additional parameter that needs to be estimated
- Maximum likelihood is a way to estimate this parameter

Box-Cox Maximum Likelihood Estimation

- Before setting up MLE, the observations are further standardized so that the magnitude of the error sum of squares does not depend on the value of λ
- The transformation is give by

$$W_i = K_1(Y_i^\lambda - 1) \qquad \lambda \neq 0$$

$$or\ K_2(log_e\, Y_i) \qquad \lambda = 0$$

where

$$K_2 = (\prod Y_i)^{1/N}$$

$$K_1 = \frac{1}{\lambda K_2^{\lambda-1}}$$

Box-Cox Maximum Likelihood Estimation

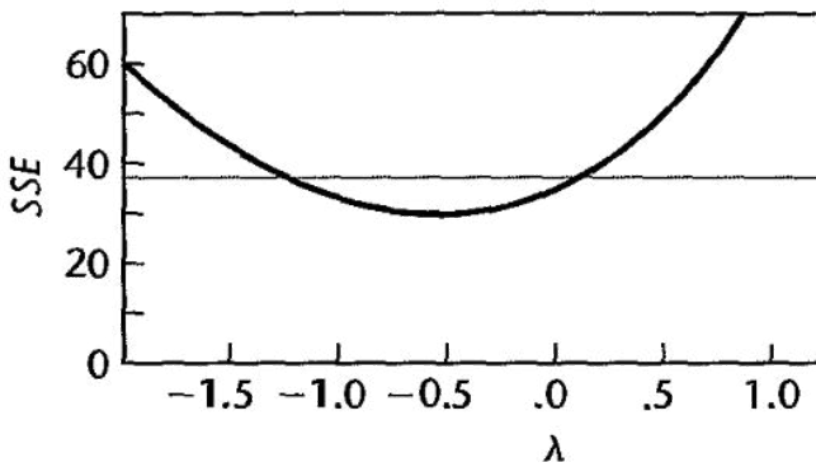
- Maximize
  $log(L(X, Y, \sigma, \lambda, b_1, b_0)) = -\sum_i \frac{(W_i - (b_1 X_i + b_0))^2}{2\sigma^2} - nlog(\sigma)$
  w.r.t λ σ b_1 b_0
- How?
  - Take partial derivatives
  - Solve
  - or... gradient ascent methods

Box-Cox Maximum Likelihood Estimation

- Maximize
  $log(L(X, Y, \sigma, \lambda, b_1, b_0)) = -\sum_i \frac{(W_i - (b_1 X_i + b_0))^2}{2\sigma^2} - nlog(\sigma)$
  w.r.t λ σ b_1 b_0
- How?
  - Take partial derivatives
  - Solve
  - or... gradient ascent methods
- Not easy task, so we work on a grid of λ values, say $\lambda = -2, -1.75, ..., 1.75, 2$. And calculate a list of MSE according to different λ values.

The plasma levels example

λ	SSE	λ	SSE
1.0	78.0	−.1	33.1
.9	70.4	−.3	31.2
.7	57.8	−.4	30.7
.5	48.4	−.5	30.6
.3	41.4	−.6	30.7
.1	36.4	−.7	31.1
0	34.5	−.9	32.7
		−1.0	33.9



Comments on Box-Cox

- The Box-Cox procedure is ordinarily used only to provide a guide for selecting a transformation
- At time, theoretical considerations or prior information can be utilized to help in choosing an appropriate transformation
- It is important to perform residual analysis after the transformation to ensure the transformation is appropriate
- When transformed models are employed, b_0 and b_1 obtained via least squares have the least squares property w.r.t the transformed observations not the original ones.
- Usually take a nearby λ value for which the power transformation is easier to understand. Say $\hat{\lambda} = 0.03$, we may just use $\hat{\lambda} = 0$. Of course, one should examine the flatness of likelihood function in the neighborhood of $\hat{\lambda}$.
- When λ near 1, no transformation of Y is needed.

Exploration of Shape of Regression Function: Nonparametric Regression Curves

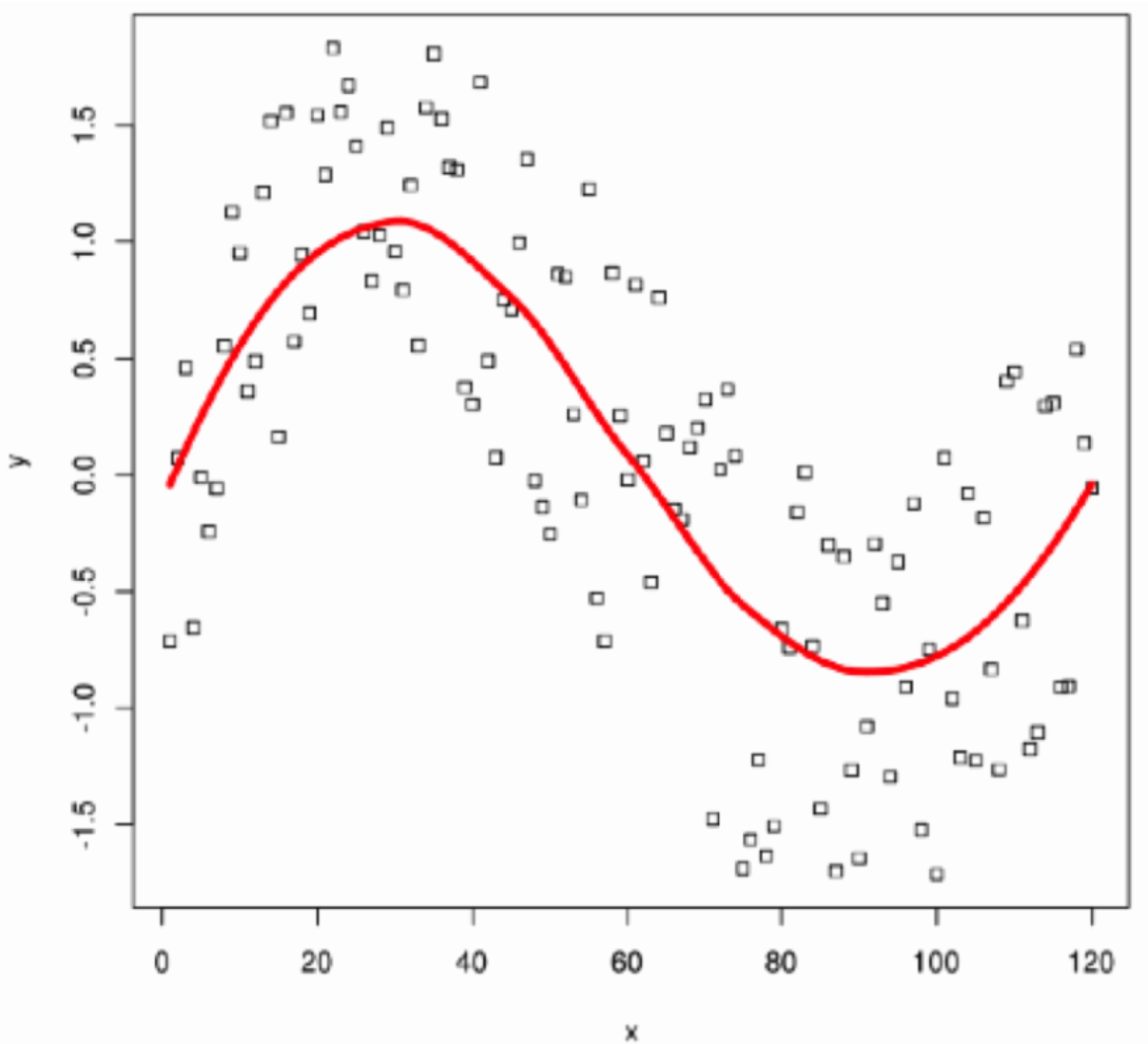
- So far: parametric regression approaches
  - Linear
  - Linear with transformed inputs and outputs
  - etc.
- Other approaches
  - Method of moving averages : interpolate between mean outputs at adjacent inputs
  - Lowess: "LOcally WEighted Scatterplot Smoothing"

Lowess Method

- Intuition
  - Fit low-order polynomial (linear) regression models to points in a neighborhood
    - The neighborhood size is a parameter
    - Determining the neighborhood is done via a Cross Validation

    Produce predictions by weighting the regressors by how far the set of points used to produce the regressor is from the input point for which a prediction is wanted
- While somewhat ad-hoc, it is a method of producing a nonlinear regression function for data that might seem otherwise difficult to regress

Lowess Method Example



R example

```
require(graphics)

plot(cars, main = "lowess(cars)")
lines(lowess(cars), col = 2)
lines(lowess(cars, f=.2), col = 3)
legend(5, 120, c(paste("f = ", c("2/3", ".2"))),
lty = 1, col = 2:3)
```

Conditional Probability Distribution

Theorem

Let $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *be a multivariate normal vector with mean* $\boldsymbol{\mu}$ *and covariance matrix* $\boldsymbol{\Sigma}$*. Consider partitioning* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$ *into*

$$\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2]' \quad \text{and} \quad \boldsymbol{\Sigma} = [\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{12};\ \boldsymbol{\Sigma}_{21}, \boldsymbol{\Sigma}_{22}]$$

$$\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2]'$$

Then $\mathbf{Y}_1 \mid \mathbf{Y}_2 = \mathbf{y}_2$*, the conditional distribution of the first partition given the second, is* $N\left(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}}\right)$*, with mean*

$$\bar{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{y}_2 - \boldsymbol{\mu}_2\right),$$

and covariance matrix

$$\bar{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

Conditional Probability Distribution - Proof 1/3

Before we prove the theorem, we start with an intuition behind the proof. The idea is to find a linear combination $Z = C_1\mathbf{Y}_1 + C_2\mathbf{Y}_2$ of the whole vector that is uncorrelated with $\mathbf{Y}_2$. If we find such a vector combination, then we can use two facts (i) $\text{Var}(Z \mid \mathbf{Y}_2) = \text{Var}(Z)$ and (ii) $\mathbb{E}(Z \mid \mathbf{Y}_2) = \mathbb{E}(Z)$.

In order to be scale invariant we can set $C_1 = \mathbb{I}$ and all we need is a matrix C_2.

Let us first consider a bivariate case of $\mathbf{Y} = (Y_1, Y_2)'$, where Y_1, and Y_2 are univariate random variables. We are looking for C_2 such that

$$\text{Cov}(Y_1 + C_2 Y_2, Y_2) = 0$$

I.e.

$$\text{Cov}(Y_1 + C_2 Y_2, Y_2) = \text{Cov}(Y_1, Y_2) + C_2\text{Var}(Y_2) = \sigma_{12} + C_2\sigma_{22} = 0.$$

Therefore, $C_2 = -\sigma_{12}\sigma_{22}^{-1}$. Hence, in the multivariate case, we conjecture that a good candidate for a linear combination is

$$Z = \mathbf{Y}_1 - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{Y}_2.$$

Conditional Probability Distribution - Proof 2/3

Now we can write

$$\begin{aligned}\text{Cov}\left(\mathbf{Z}, \mathbf{Y}_2\right) &= \text{Cov}\left(\mathbf{Y}_1, \mathbf{Y}_2\right) - \text{Cov}\left(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Y}_2, \mathbf{Y}_2\right) \\ &= \boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\text{Var}\left(\mathbf{Y}_2, \mathbf{Y}_2\right) \\ &= \boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{22} \\ &= \boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}_{12} = 0\end{aligned}$$

So, indeed $\mathbf{Z}$ and $\mathbf{Y}_2$ are uncorrelated (the same as in the bivariate case).
Since they are jointly Gaussian, they must be independent.
Now, using $\mathbf{Z}$, we compute the conditional mean of $\mathbf{Y}_1$ given $\mathbf{Y}_2$

$$\begin{aligned}\mathbb{E}\left(\mathbf{Y}_1 \mid \mathbf{Y}_2 = \mathbf{y}_2\right) &= \mathbb{E}\left(\mathbf{Z} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Y}_2 \mid \mathbf{Y}_2 = \mathbf{y}_2\right) \\ &= \mathbb{E}\left(\mathbf{Z} \mid \mathbf{Y}_2 = \mathbf{y}_2\right) + \mathbb{E}\left(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Y}_2 \mid \mathbf{Y}_2 = \mathbf{y}_2\right) \\ &= \mathbb{E}\left(\mathbf{Z}\right) + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{y}_2 \\ &= \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{y}_2 \\ &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{y}_2 - \boldsymbol{\mu}_2\right),\end{aligned}$$

which proves the conditional mean part.

Conditional Probability Distribution - Proof 3/3

For the covariance matrix

$$\begin{aligned}\text{Var}\left(\mathbf{Y}_1 \mid \mathbf{Y}_2 = \mathbf{y}_2\right) &= \text{Var}\left(\mathbf{Z} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Y}_2 \mid \mathbf{Y}_2 = \mathbf{y}_2\right) \\ &= \text{Var}\left(\mathbf{Z} \mid \mathbf{Y}_2 = \mathbf{y}_2\right) + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\text{Var}\left(\mathbf{Y}_2 \mid \mathbf{Y}_2 = \mathbf{y}_2\right)\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ &+ \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\text{Cov}\left(\mathbf{Z}, \mathbf{Y}_2\right) + \text{Cov}\left(\mathbf{Y}_2, \mathbf{Z}\right)\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ &= \text{Var}\left(\mathbf{Z} \mid \mathbf{Y}_2 = \mathbf{y}_2\right) = \text{Var}\left(\mathbf{Z}\right),\end{aligned}$$

since $\boldsymbol{\Sigma}_{22} = \boldsymbol{\Sigma}_{22}'$, $\boldsymbol{\Sigma}_{12}' = \boldsymbol{\Sigma}_{21}$, $\text{Var}\left(\mathbf{Y}_2 \mid \mathbf{Y}_2 = \mathbf{y}_2\right) = 0$, and $\text{Cov}\left(\mathbf{Z}, \mathbf{Y}_2\right) = 0$. Now,

$$\begin{aligned}\text{Var}\left(\mathbf{Y}_1 \mid \mathbf{Y}_2\right) = \text{Var}\left(\mathbf{Z}\right) &= \text{Var}\left(\mathbf{Y}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{Y}_2\right) \\ &= \text{Var}\left(\mathbf{Y}_1\right) + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\text{Var}\left(\mathbf{Y}_2\right)\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ &- \text{Cov}\left(\mathbf{Y}_1, \mathbf{Y}_2\right)\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\text{Cov}\left(\mathbf{Y}_2, \mathbf{Y}_1\right) \\ &= \boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} - 2\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ &= \boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} - 2\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \text{ which completes the proof.}\end{aligned}$$