

## Regression and Correlation - II

Failure of Assumptions  
Nonparametric Regression  
Regression Diagnostics  
Correlation  
Nonparametric Correlation

---

---

---

---

---

---

---

## Regression Assumptions

**Linearity**  
A linear relationship exists between Y and the X's.

**Constant Variance**  
Variance of the e's is constant for all values of the X's.

**Normality**  
The e's are normally distributed.

**Independence**  
The e's are independent of x and each other.

**Special Causes**  
Outliers & problem points have been removed.

---

---

---

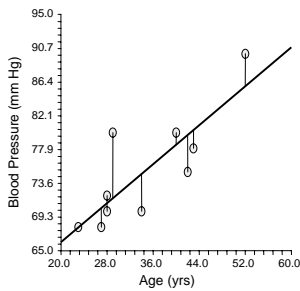
---

---

---

---

## Assumption: Linearity



Shows up with early scatterplots of the data.

Will also show up with various residual plots later in analysis.

Must switch models if the data are not linear.

---

---

---

---

---

---

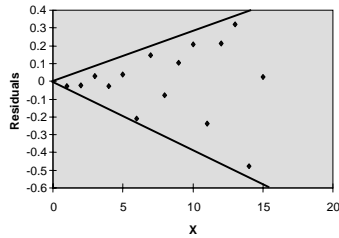
---

Data Set:

X	Y
1	0.10
2	0.20
3	0.35
4	0.39
5	0.55
6	0.40
7	0.85
8	0.72
9	1.00
10	1.20
11	0.85
12	1.40
13	1.60
14	0.90
15	1.50

### Assumption: Variance

Shows up most clearly in residual plot. A wedge (or bowtie) indicates increasing variance with increasing  $x$ . Attempt a transformation to correct (e.g.,  $\log_{10}$ ).




---

---

---

---

---

---

---

---

---

---

---

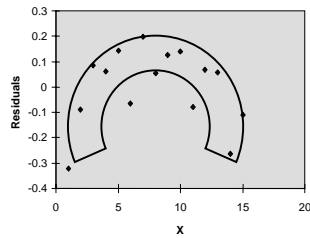
---

Transform Y

X	$\log_{10} y + 1$
1	1.000
2	1.301
3	1.477
4	1.602
5	1.699
6	1.778
7	1.845
8	1.903
9	1.954
10	2.000
11	2.041
12	2.079
13	2.114
14	2.146
15	2.176

### Assumption: Variance

A  $\log_{10}+1$  transformation of the  $y$ 's creates another problem: a horseshoe effect (indicates wrong transformation). Continue by trying other transforms.




---

---

---

---

---

---

---

---

---

---

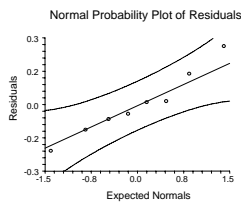
---

---

### Assumption: Normality

Data Set:

X	Y
1	1.9
2	2.3
3	2.6
4	2.9
5	2.7
6	3.0
7	3.1
8	3.3



The  $\epsilon$ 's are normally distributed based on D'Agostino tests and normal probability plot.

Assumption	Value	Prob	Dec(5%)
Skewness	0.8142	0.415520	Accepted
Kurtosis	0.4544	0.649546	Accepted
Omnibus	0.8694	0.647451	Accepted

---

---

---

---

---

---

---

---

---

---

---

---

## Nonparametric Regression

If the data are suspected to be linear and you are unable to correct either a variance or normality assumption (via transformation), it may be appropriate to conduct a **nonparametric regression**.

There are a variety of options, one of which is **Kendall's robust line-fit method**.

---

---

---

---

---

---

---

---

## Kendall's Robust Line-fit Method

- Procedure -

The method is fairly straightforward:

Rank order the  $x, y$  pairs based on the  $x$ 's.

Compute a slope ( $S_{ji}$ ) for EVERY pair of  $x$ -values, for  $i = 1$  to  $n - 1$  and  $j > i$  :

$$S_{ji} = \frac{Y_j - Y_i}{X_j - X_i}$$

There will be  $n(n-1)/2$  slope estimates per sample.

---

---

---

---

---

---

---

---

## Kendall's Robust Line-fit Method

- Procedure -

After computing all possible  $S_{ji}$ , the nonparametric estimate  $b$  of the slope  $\beta$  is the MEDIAN of the of the  $S_{ji}$  values.

The Kendall's Rank Correlation Coefficient (an ordering test) can be used to test  $b$  for sig.

To estimate the  $y$ -intercept, compute the  $n$ -values of  $Y_i - bX_i$  and again choose the MEDIAN.

---

---

---

---

---

---

---

---

### Kendall's Robust Line-fit Method

- Example -

Data Set:

X	Y
0.0	8.98
12.0	8.14
29.5	6.67
43.0	6.08
53.0	5.90
62.5	5.83
75.5	4.68
85.0	4.20
93.0	3.72

Calculate  $S_{ij}$ 's:

$S_{21} = (8.14 - 8.98)/(12-0) = -0.07000$
$S_{32} = (6.67 - 8.14)/(29.5-12) = -0.08400$
.
$S_{31} = (6.67 - 8.98)/(29.5 - 0) = -0.07831$
.
$S_{91} = (3.72 - 8.98)/(93.0 - 0) = -0.05656$
.
Median of the 36 slopes: $b = -0.05436$

---

---

---

---

---

---

---

---

---

---

---

---

### Kendall's Robust Line-fit Method

- Example -

To estimate the y-intercept, compute the 9 values of  $a_i$  using  $Y_i - bX_i$ .

For  $i = 1$ :  $a_1 = 8.98 - (-0.05436)(0.0) = 8.980$

For  $i = 2$ :  $a_2 = 8.14 - (-0.05436)(12.0) = 8.792$

MEDIAN of the  $n$  intercepts:  $a = 8.783$

THEREFORE:  $Y = 8.783 - 0.05436 X$

---

---

---

---

---

---

---

---

---

---

---

---

### Assumption: Independence

Data Set:

X	Y
1	1.1
2	1.2
3	1.1
4	1.3
5	1.2
6	1.4
7	1.3
8	1.5
9	1.4
10	1.6

#### Serial-Correlation Section

Lag	Correlation
1	-0.868116
2	0.750198
3	-0.647760
4	0.522266
5	-0.427207

Above serial correlations significant if their absolute values are greater than 0.632458

Durbin-Watson Value 3.6387

D-W is a test of first order serial correlation.

D-W ranges from 0-4.

D-W = 2 means no autocorrelation.

---

---

---

---

---

---

---

---

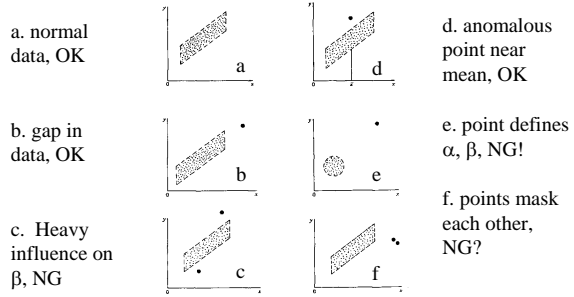
---

---

---

---

### Assumption: Special Causes




---

---

---

---

---

---

---

---

### Regression Diagnostics

In addition to residual analysis, regression diagnostics provide a relatively objective method of identifying individual points that may be affecting estimates of  $\alpha$  &  $\beta$ , their standard errors, and/or test statistics.

Influential observations are those which have a marked effect on the inference process.

See: Belsley, Kuh, and Welsch. (1980)  
Regression Diagnostics, John Wiley, New York.

---

---

---

---

---

---

---

---

### Regression Diagnostics

These statistics flag observations that exert three types of influence:

Outliers in Residual Space  
Studentized Residual, RStudent, CovRatio

Outliers in X-space  
Hat Diagonals

Parameter Estimates and Fit  
Dffitts, Cook's D

---

---

---

---

---

---

---

---

## Outliers in Residual Space

### Studentized Residual

An “internally” standardized residual that eliminates the effect of location of the  $i_{th}$  data point in  $x$ -space.

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

### RStudent

An “externally” standardized residual that has the impact of a *single observation removed from the mean square error*.

$$t_i = \frac{e_i}{\sqrt{MSE_i(1-h_{ii})}}$$

---

---

---

---

---

---

---

---

## Outliers in Residual Space

### Covratio

The Covratio flags observations that have a major impact on the generalized variance of the regression coefficients.

A value  $< 1.0$  is increasing the estimated generalized variance and this is unfavorable.

$$Covratio_i = \frac{|(X_{-i} X_{-i})^{-1} MSE_i|}{|(X' X)^{-1} MSE|}$$

---

---

---

---

---

---

---

---

## Outliers in X-Space

### Hat Diagonal

The hat diagonal,  $h_{ii}$ , captures an observation's remoteness in  $x$ -space.

Some authors refer to the hat diagonal as a measure of leverage in the  $x$ -space.

Hat diagonals greater than two times the number of coefficients in the model divided by the number of observations are considered influential (i.e.,  $h_{ii} > 2p/n$ ).

---

---

---

---

---

---

---

---

## Parameter Estimates and Fit

### Dffits

A standardized difference between the prediction with and without the  $i$ th observation.

Represents the numbers of estimated SEs that the fitted value changes if the  $i$ th obs is removed.

Dffits  $> 1.0$  are problematic.

$$dffits = \frac{\hat{y}_i - \hat{y}_{i,-i}}{\sqrt{MSE_i(h_{ii})}}$$

---

---

---

---

---

---

---

---

## Parameter Estimates and Fit

### Cook's D

Cook's distance is an overall measure of how an observation impacts the regression coefficients.

$D$  is related to the  $i$ th studentized residual ( $r_i$ ) as well as the hat diagonal ( $h_{ii}$ ).

Cook's  $D > 1.0$  means that obs has great influence on the coefficients as a whole.

$$D_i = \frac{r_i^2 h_{ii}}{1 - h_{ii}}$$

---

---

---

---

---

---

---

---

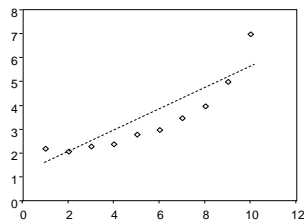
## Regression Diagnostics

- Example -

Data Set:

X	Y
1	2.2
2	2.1
3	2.3
4	2.4
5	2.8
6	3.0
7	3.5
8	4.0
9	5.0
10	7.0

Q. Linear?  
A. Maybe.



Independent Variable	Regression Coefficient	Standard Error	T-Value (Ho: B=0)	Prob Level
Intercept	0.9133334	0.5104949	1.7891	0.111390
C1	0.4575758	0.0822739	5.5616	0.000534

R-Squared 0.794512

---

---

---

---

---

---

---

---

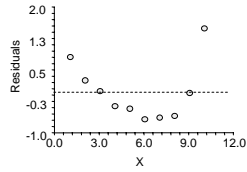
## Regression Diagnostics

- Example -

### Residual Analysis

Q. Variance homogeneous?

A. Maybe.



#### Residual Report

Row	Actual	Predicted	Residual	Percent Error	MSEI
1	2.2	1.370909	0.8290909	37.69	0.4881905
2	2.1	1.828485	-0.2715152	12.93	0.6242028
3	2.3	2.286061	1.3939E-02	0.61	0.6381828
4	2.4	2.743636	-0.3436364	14.32	0.6188869
5	2.8	3.201212	-0.4012121	14.33	0.6125792
6	3	3.658788	-0.6587879	21.96	0.5690946
7	3.5	4.116364	-0.6163636	17.61	0.5760298
8	4	4.573939	-0.5739394	14.35	0.5811239
9	5	5.031515	-3.1515E-02	0.63	0.6380277
10	7	5.489091	1.510909	21.58	0.1399762

---

---

---

---

---

---

---

---

---

---

---

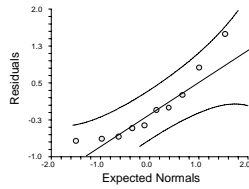
---

## Regression Diagnostics

- Example -

Q.  $\epsilon$ 's normal?

A. Yes.



#### Normality Tests Section

Assumption	Value	Probability
Skewness	1.8556	0.063514
Kurtosis	0.9655	0.334277
Omnibus	4.3754	0.112174

---

---

---

---

---

---

---

---

---

---

---

---

## Regression Diagnostics

- Example -

Obs. 10 has extreme influence. Delete & redo.

#### Regression Diagnostics Section

Row	Studentized Residual	Rstudent	Hat Diagonal	Cook's D	Dffits	Covratio
1	1.371337	1.466688	0.345455	0.496260	1.065524	1.167580
2	0.419119	0.396426	0.248485	0.029041	0.227952	1.662500
3	0.020546	0.019220	0.175758	0.000045	0.008875	1.584467
4	-0.492234	-0.467578	0.127273	0.017667	-0.178559	1.407317
5	-0.566888	-0.541258	0.103030	0.018457	-0.183442	1.341512
6	-0.930826	-0.922071	0.103030	0.049762	-0.312506	1.157815
7	-0.882896	-0.869311	0.127273	0.056839	-0.331974	1.219156
8	-0.845961	-0.829286	0.175758	0.076301	-0.382943	1.313804
9	-0.048648	-0.045513	0.248485	0.000391	-0.026170	1.736957
10	2.499081	4.991618	0.345455	1.648094	3.626326	0.095988

---

---

---

---

---

---

---

---

---

---

---

---

## Regression Diagnostics - Example -

Regression Equation Section (Original)				
Independent Variable	Regression Coefficient	Standard Error	T-Value (Ho: B=0)	Prob Level
Intercept	0.9133334	0.5104949	1.7891	0.111390
C1	0.4575758	0.0822739	5.5616	0.000534

R-Squared 0.794512

Regression Equation Section (Revised)				
Independent Variable	Regression Coefficient	Standard Error	T-Value (Ho: B=0)	Prob Level
Intercept	1.375	0.271802	5.0588	0.001465
CS	0.3316667	4.830E-02	6.8667	0.000238

R-Squared 0.870734

---

---

---

---

---

---

---

---

---

---

## Regression Diagnostics - Example -

### Regression Diagnostics Section

Row	Studentized Residual	Rstudent	Hat Diagonal	Cook's D	Dffits	Covratio
1	1.671632	1.996640	0.377778	0.848286	1.555770	0.789619
2	0.191749	0.177993	0.261111	0.006497	0.105810	1.822805
3	-0.206337	-0.191614	0.177778	0.004603	-0.089099	1.635330
4	-0.863350	-0.845593	0.127778	0.054597	-0.323650	1.245872
5	-0.661494	-0.632513	0.111111	0.027348	-0.223627	1.345795
6	-1.044605	-1.052636	0.127778	0.079929	-0.402895	1.111908
7	-0.579708	-0.550072	0.177778	0.036331	-0.255778	1.500273
8	-0.088101	-0.081611	0.261111	0.001371	-0.048515	1.838022
9	2.168603	3.504775	0.377778	1.427648	2.730899	0.235578



Obs-9 is still a problem via diagnostics.

---

---

---

---

---

---

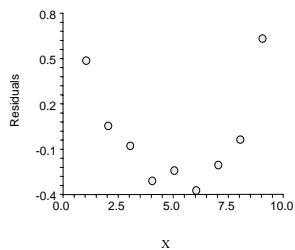
---

---

---

---

## Regression Diagnostics - Example -



### Conclusion:

Diagnostics were useful to probe data and to justify the deletion of a possible observation with undue influence.

However, deletion did not improve residual problems, and data are likely not linear.

Conclusion:  
*Switch to another model.*

---

---

---

---

---

---

---

---

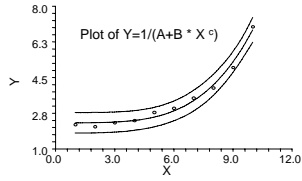
---

---

## Regression Diagnostics

- Example -

### Best Fit Model



#### Model Estimation Section

Parameter Name	Parameter Estimate	Asymptotic Standard Error	Lower 95% C.L.	Upper 95% C.L.
A	2.281144	9.732569E-02	2.051005	2.511282
B	4.22E-04	3.675781E-04	-4.46E-04	1.29E-03
C	4.032616	0.3737149	3.148921	4.916311

Dependent C2  
 Independent C1  
 Model C2=1/(A+B\*(C1)^C)  
 R-Squared 0.988491  
 Iterations 57  
 Estimated Model 1/((2.281144)+(4.228083E-04)\*(C1)^(4.032616))

---

---

---

---

---

---

---

---

---

---

---

---

## Correlation

There are many purposes to regression, but the main one is for prediction. Thus, the goal is to determine the **NATURE** of the relationship between two variables.

Often as a next step, or for other reasons, one wishes just to determine the **STRENGTH** of the relationship, one would do a correlation analysis.

The product is a correlation coefficient,  $r$ .

NB: Regression and correlation are different, but not mutually exclusive, techniques.

---

---

---

---

---

---

---

---

---

---

---

---

## Correlation

The correlation coefficient is determined using the same sample statistics as used in regression:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

In all cases,  $-1 \leq r \leq +1$

$r = 0$  is no relationship

$r = 1$  is a perfect relationship (pos. or neg.)

---

---

---

---

---

---

---

---

---

---

---

---

## Coefficient of Determination

To demonstrate the inter-relatedness of correlation & regression, let's return to regression momentarily to tie up a loose end.

When we discuss the variability in  $y$  which is explained by the linear association between  $x$  and  $y$ , we frequently refer to the coefficient of determination,  $R^2$ :

$$R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

---

---

---

---

---

---

---

---

## Coefficient of Determination

If  $R^2$  is large (close to 1.0) virtually all of the variability is explained by the relationship. Knowledge of  $x$  permits knowledge of  $y$ .

If  $R^2$  is close to 0, there is no relationship and a knowledge of  $x$  permits no insight in to  $y$ .

$R^2$  is sometimes symbolized as  $r^2$  (since it is the square of the correlation coefficient) but, to clearly differentiate the two, use a capital  $R$  for regression and a lowercase  $r$  for correlation.

---

---

---

---

---

---

---

---

## Correlation

The assumptions for correlation are different than for regression:

Subjects are sampled at random.

Both  $x$  and  $y$  contain sampling variability.

For each value of  $x$  there is a normal dist. of  $y$ 's.

For each value of  $y$  there is a normal dist. of  $x$ 's.

The  $x$  distributions have the same variance.

The  $y$  distributions have the same variance.

The joint distribution of  $x$  and  $y$  is bivariate normal.

---

---

---

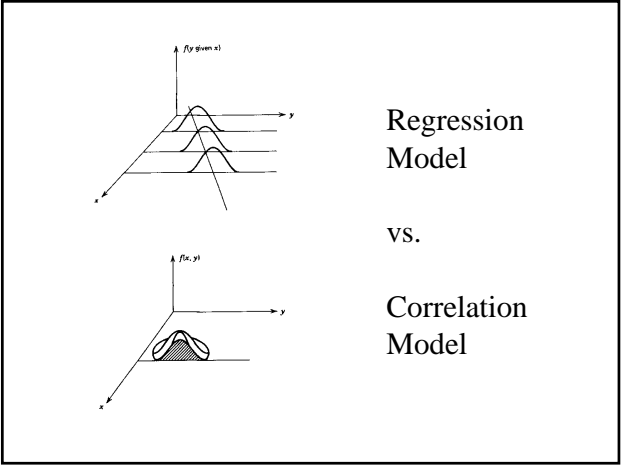
---

---

---

---

---



Regression Model

vs.

Correlation Model

---

---

---

---

---

---

---

---

### Correlation

Like the previous statistics we have used, there are confidence intervals for estimates of  $r$  as well as significance tests:

$$CI_{0.95} : z_r - z_{\alpha/2} \left( \frac{1}{\sqrt{n-3}} \right) \leq z_r \leq z_r + z_{\alpha/2} \left( \frac{1}{\sqrt{n-3}} \right)$$

where  $z_r = \log_e \sqrt{(1+r)/(1-r)}$  (transformation)

and

to test  $H_0 : \rho = 0 \quad t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$

Use z- and t-tables, respectively.

---

---

---

---

---

---

---

---

### Nonparametric Correlation

When we discuss the "correlation coefficient" it is usually assumed that we are referring to the parametric correlation coefficient which is most correctly referred to as the **Pearson Product Moment Correlation Coefficient**.

However, recognize that there are MANY types of correlation coefficients to deal with different situations. One, most often used for the failure of parametric assumptions is the nonparametric **Spearman's Rank Correlation Coefficient**.

---

---

---

---

---

---

---

---

## Spearman's Rank Correlation

- Procedure -

$H_0: E(r_s) = 0$  (i.e., the ranking of the  $x$  variable is independent of the ranking of the  $y$  variable)

$H_a: E(r_s) \neq 0, E(r_s) > 0, E(r_s) < 0$

$$r_s = 1 - \frac{6 \sum d^2}{N(N^2 - 1)} \quad \text{with } d = r_x - r_y \text{ (diff. in } x, y \text{ ranks)}$$

Test statistic:  $z = r_s \sqrt{n-1}$

Test can be performed on continuous data converted to ranks, or ordinal data.

---

---

---

---

---

---

---

---

## Spearman's Rank Correlation

- Example -

Rank of rat health by two observers at the end of an endurance experiment.

Rat	Obs-1	Obs-2	$d$	$d^2$	
1	4	4	0	0	$r_s = 1 - \frac{6 \sum d^2}{N(N^2 - 1)}$
2	1	2	-1	1	
3	6	5	1	1	$r_s = 1 - (6)(8)/(7)(48)$
4	5	6	-1	1	
5	3	1	2	4	$r_s = 0.857$
6	2	3	-1	1	$z = 2.099, P = 0.018$
7	7	7	0	0	
			Sum $d^2=8$		reject $H_0$

---

---

---

---

---

---

---

---

# The End

Stay tuned for Analysis of Variance (ANOVA).

---

---

---

---

---

---

---

---