UCLA STAT 110B

Applied Statistics for Engineering and the Sciences

- **Instructor:** **Ivo Dinov,**
  Asst. Prof. In Statistics and Neurology
- **Teaching Assistants:** Brian Ng, UCLA Statistics

University of California, Los Angeles, Spring 2003

http://www.stat.ucla.edu/~dinov/courses_students.html

Linear Regression Analysis

Correlation Coefficient

Correlation coefficient ($-1 <= R <= 1$): a measure of linear association, or clustering around a line of multivariate data.

Relationship between two variables (X, Y) can be summarized by: (μ_X, σ_X), (μ_Y, σ_Y) and the correlation coefficient, R. $R=1$, perfect positive correlation (straight line relationship), $R=0$, no correlation (random cloud scatter), $R=-1$, perfect negative correlation.

Computing R(X,Y): (standardize, multiply, average)

$$R(X,Y)=\frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

X={$x_1, x_2, \ldots, x_N$,}
Y={$y_1, y_2, \ldots, y_N$,}
(μ_X, σ_X), (μ_Y, σ_Y)
sample mean / SD.

Correlation Coefficient

Example:

$$R(X,Y)=\frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

Student i	Height x_i	Weight y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	167	60	6	4.67	36	21.8089	28.02
2	170	64	9	8.67	81	75.1689	78.03
3	160	57	-1	1.67	1	2.7889	-1.67
4	152	46	-9	-9.33	81	87.0489	83.97
5	157	55	-4	-0.33	16	0.1089	1.32
6	160	50	-1	-5.33	1	28.4089	5.33
Total	966	332	0	≈ 0	216	215.3334	195.0

Correlation Coefficient

Example:

$$R(X,Y)=\frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

$$\mu_x=\frac{966}{6}=161\,\text{cm},\qquad \mu_y=\frac{332}{6}=55\,\text{kg},$$

$$\sigma_x=\sqrt{\frac{216}{5}}=6.573,\qquad \sigma_y=\sqrt{\frac{215.3}{5}}=6.563,$$

$$Corr(X,Y)=R(X,Y)=0.904$$

Correlation Coefficient - Properties

Correlation is invariant w.r.t. linear transformations of X or Y

$$R(X,Y)=\frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)=$$

$$R(aX+b, cY+d), \qquad \text{since}$$

$$\left(\frac{ax_k+b-\mu_{ax+b}}{\sigma_{ax+b}}\right)=\left(\frac{ax_k+b-(a\mu_x+b)}{|a|\times\sigma_x}\right)=$$

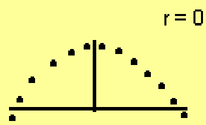
$$\left(\frac{a(x_k-\mu)+b-b}{a\times\sigma_x}\right)=\left(\frac{x_k-\mu_x}{\sigma_x}\right)$$

Correlation Coefficient - Properties

Correlation is Associative

$$R(X,Y)=\frac{1}{N}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)=R(Y,X)$$

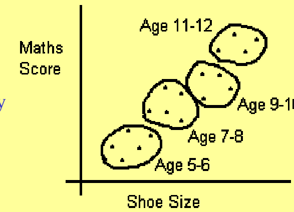
Correlation measures linear association, NOT an association in general!!! So, Corr(X,Y) could be misleading for X & Y related in a non-linear fashion.



Correlation Coefficient - Properties

$$R(X,Y)=\frac{1}{N}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)=R(Y,X)$$

1. R measures the extent of linear association between two continuous variables.
2. Association does not imply causation - both variables may be affected by a third variable – age was a confounding variable.



Recall the correlation coefficient…

Another form for the correlation coefficient is:

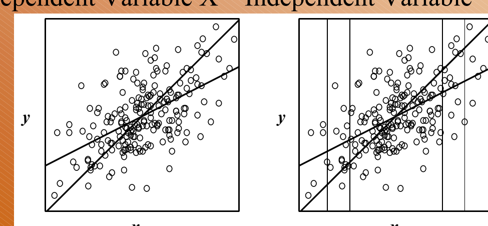
$$R(X;Y)=Corr(X;Y)=\boxed{\frac{\sum_{i=1}^{n}\left[(x_i-\bar{x})(y_i-\bar{y})\right]}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\times\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}}}=$$

$$\frac{\sum_{i=1}^{n}\left[y_i x_i+\bar{x}\bar{y}-\bar{y}x_i-\bar{x}y_i\right]}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}}=\boxed{\frac{\sum_{i=1}^{n}\left(y_i x_i\right)\;-\;n\times\bar{x}\times\bar{y}}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\times\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}}}$$

Linear Regression Analysis (ch. 12)

Observe a response Y and one or more predictors X. Formulate a model that relates the mean response E(Y) to X.

Y – Dependent Variable X – Independent Variable



Deterministic Model

- Y = f(x) ; Once we know the value of x, the value of Y is completely satisfied
- Simplest (Straight Line)Model:
  $Y=\beta_o+\beta_1 x$
- β_1 = Slope of the Line
- β_o = Y-intercept of the Line

Probabilistic Model

- $Y = f(x) + \varepsilon$; The value of Y is a R.V.
- Model for Simple Linear Regression:
  $Y_i = \beta_o + \beta_1 x_i + \varepsilon_i$, i=1,..,n
- $Y_1,\ldots,Y_n$ – Observed Value of the Response
- $x_1,\ldots,x_n$ – Observed Value of Predictor
- β_o,β_1 – Unknown Parameters to be Estimated from the Data
- $\varepsilon_1,\ldots,\varepsilon_n$ – Unknown Random Error Terms – Usually iid $N(0,\sigma^2)$ Random Variables

Interpretation of Model

For each value of x, the observed Y will fall above or below the line $Y = \beta_o + \beta_1 x$ according to the error term ε. For each fixed x

$$Y \sim N(\beta_o + \beta_1 x, \sigma^2)$$

Questions

1. How do we estimate β_o, β_1, and σ^2?
2. Does the proposed model fit the data well?
3. Are the assumptions satisfied?

Plotting the Data

A scatter plot of the data is a useful first step for checking whether a linear relationship is plausible.

Example (12.4)

A study to assess the capability of subsurface flow wetland systems to remove **biochemical oxygen demand** (BOD) and other various chemical constituents resulted in the following scatter plot of the data where x = BOD mass loading and y = BOD mass removal. Does the plot suggest a linear relationship?

x	3	8	10	11	13	16	27	30	35	37	38	44	103	142
y	4	7	8	8	10	11	16	26	21	9	31	30	75	90

Example (12.5)

An experiment conducted to investigate the stretchability of mozzarella cheese with temperature resulted in the following scatter plot where x = temperature and y = % elongation at failure. Does the scatter plot suggest a linear relationship?

Estimating β_o and β_1

Consider an arbitrary line $y = b_0 + b_1 x$ drawn through a scatter plot. We want the line to be as close to the points in the scatter plot as possible. The vertical distance from (x,y) to the corresponding point on the line $(x, b_0 + b_1 x)$ is $y-(b_0 + b_1 x)$.

Possible Estimation Criteria

- Eyeball Method
- L_1 Estimation - Choose β_o,β_1 to minimize
  $\Sigma \mid y_i - \beta_o x - \beta_1 x_i \mid$
- Least Squares Estimation - Choose β_o,β_1 to minimize $\Sigma(y_i - \beta_o - \beta_1 x_i)^2$

* We use Least Squares Estimation in practice since it is difficult to mathematically manipulate the other options*

Least Squares Estimation

Take derivatives with respect to b_0 and b_1, and set equal to zero. This results in the "normal equations" (based on right angles – not the Normal distribution)

Formulas for Least Squares Estimates

Solving for b_0 and b_1 results in the L.S. estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

Example (12.12)

Refer to the previous example (12.4). Obtain the expression for the Least Squares line

$$n = 14 \qquad \sum \mathrm{x}_i = 517$$
$$\sum \mathrm{y}_i = 346 \qquad \sum \mathrm{x}_i^2 = 39{,}095$$
$$\sum \mathrm{y}_i^2 = 17{,}454 \qquad \sum \mathrm{x}_i\,\mathrm{y}_i = 25{,}825$$

Estimating σ^2

Residual = Observed – Predicted

$$e_i = y_i - \hat{y}_i$$

Recall the definition of sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Estimating σ^2 Cont'd

- The minimum value of the squared deviation is

$D = \Sigma(y_i - \beta_o x - \beta_1 x_i)^2 = \Sigma(y_i - \hat{y}_i)^2 = SSE$

- Divide the SSE by it's degrees of freedom (n-2) to estimate σ^2

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2}$$

Example (12.12) Cont'd

Predict the value of BOD mass removal when BOD loading is 35. Calculate the residual. Calculate the SSE and a point estimate of σ^2

Examining the Overall Fit of the Model

Recall from previous lecture:

- Linear Regression Model:

  $Y_i = \beta_o + \beta_1 x_i + \varepsilon_i$, i=1,..,n

- Assumptions:

  $\varepsilon_i \sim N(0,\sigma^2) \Rightarrow Y \sim N(\beta_o + \beta_1 x, \sigma^2)$

Review Cont'd

- L.S. estimate of β_1:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} = \frac{\sum_{i=1}^{n}\left[(x_i - \bar{x})(y_i - \bar{y})\right]}{\sum_{i=1}^{n}(x_i - \bar{x})^2};$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- L.S. estimate of β_0:

Another Notation for the Slope of the LS line

1. Note that there is a slight difference in the formula for the slope of the Least Squares Best-Linear Fit line:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left[(x_i - \bar{x})(y_i - \bar{y})\right]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}; \quad \hat{\beta}_o = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = Corr(X;Y) \times \frac{SD(Y)}{SD(X)}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Review Cont'd

- Predicted Values: $\hat{y}_i = \hat{\beta}_o + \hat{\beta}_1 x_i$
- Residuals: $e_i = y_i - \hat{y}_i$
- Sum of Squares Error:

  $SSE = \Sigma e_i^2 = \Sigma (y_i - \hat{y}_i)^2$

- Sample Variance: $s^2 = \frac{SSE}{n-2}$

Examining Fit Cont'd

Total Sum of Squares:

Error Sum of Squares:

Regression Sum of Squares:

Examining Fit Cont'd

Decomposition of SST:

Degrees of Freedom:

Demos

RegressionApplet.html

C:/Ivo.dir/UCLA_Classes/others.dir/JSci/examples/CurveFitter/SOCRCurveFitter.html

Coefficient of Determination

A useful measure of overall fit

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Properties:

1. $0 \le r^2 \le 1$
2. If all the data lies in a straight line, $r^2 = 1$

3. No Linear Relationship, $r^2 = 0$
4. r^2 is the proportion of variation of y "explained" by the linear relationship with x.

Testing for a Linear Relationship

Inference about β_1 is more important that β_o in that β_1 measures the effect on E[Y] of changing x by one unit.

Hypothesis Test:

H_o: $\beta_1 = 0$

H_a: $\beta_1 \neq 0$

Test Statistic:

$$F = \frac{MSR}{MSE} = (n-2)\frac{r^2}{1-r^2}$$

Rejection Region:

$F > F_{\alpha,1,n-1}$

Mean and Variance of $\hat{\beta}_1$

$$E[\hat{\beta}_1] = \beta_1$$

$$Var[\hat{\beta}_1] = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

Under the assumptions of Linear Regression

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \sim t_{n-2}$$

where $s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$

C.I. for β_1

$$\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1}$$

Hypothesis Testing

Hypothesis Test:

H_o: $\beta_1 = \beta_{1o}$

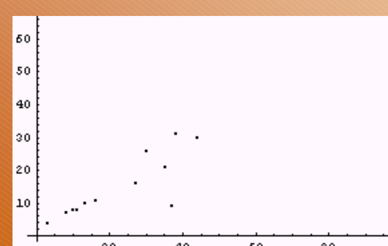
H_a: β_1 ($\neq$,>,<) β_{1o}

$$t = \frac{\hat{\beta}_1 - \beta_{1o}}{s_{\hat{\beta}_1}} \sim t_{n-2}$$

The inequality test when $\beta_{1o} = 0$ is referred to as the "model utility" test and is equivalent to the ANOVA test shown previously

Example 12.4, Cont'd

data = {{3, 4}, {8, 7}, {10, 8}, {11, 8}, {13, 10}, {16, 11}, {27, 16}, {30, 26}, {35, 21}, {37, 9}, {38, 31}, {44, 30}, {103, 75}, {142, 90}}



		Estimate	SE	TStat	PValue
ParameterTable →	1	0.62614	2.13541	0.293218	0.774364
	x	0.65229	0.0404095	16.142	1.67475×10^{-9}

AdjustedRSquared → 0.952305, RSquared → 0.955974,

EstimatedVariance → 32.6634, ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	1	8510.9	8510.9	260.564	1.67475×10^{-9}
Error	12	391.961	32.6634		
Total	13	8902.86			

Linear Correlation (12.5)

In the regression analysis that we have considered so far, we assume that x is a controlled independent variable and Y is an observed Random Variable. What if both X and Y are observed Random Variables (i.e., we observe both X and Y together)? A correlation analysis may be used to study the relationship between these two R.V.'s

• Regression Analysis – We wish to form a model to estimate $\mu_{y \cdot x}$ or to predict Y for a given value of x

•Correlation Analysis – We wish to study the relationship between X and Y

A measure of the linear relationship between X and Y is the population covariance

$$Cov(X,Y) = E[(X- \mu_X)(Y- \mu_Y)]$$

The computed sample covariance is given by

$$\frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})$$

The measure of covariance is affected by the units of the measurement of X&Y. The correlation coefficient, however, is not affected by the measurement unit of X&Y

The population correlation coefficient for X&Y is given by

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

The computed correlation coefficient is given by

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Remarks about ρ:

1. $-1 \leq \rho \leq 1$
2. $\rho = \pm 1$ if the distribution of X&Y is concentrated on a straight line
3. ρ near 0 indicated no linear relationship
4. $\rho > 0$ indicates that Y has a tendency to increase as X increases
5. $\rho < 0$ indicates that Y has a tendency to decrease as X increases
6. r has a similar interpretation for the scatter plot of (x,y)

Testing for a Linear Relationship

Assume that X&Y are distributed as a bivariate normal distribution. The parameters of this distribution are μ_X, μ_Y, σ_X^2, σ_Y^2, and ρ.

Hypothesis:

H_o: $\rho = 0$

H_a: $\rho \neq 0$

Test Statistic: $t = \dfrac{r}{\sqrt{\dfrac{1-r^2}{n-2}}}$

Rejection Region:

$|t| > t_{\alpha/2,\, n-2}$

Example (12.59)

Toughness and Fibrousness of asparagus are major determinants of quality. A journal article reported the accompanying data on x = sheer force (kg) and y = percent fiber dry weight

x	46	48	55	57	60	72	81	85	94
y	2.18	2.1	2.13	2.28	2.34	2.53	2.28	2.62	2.63

x	109	121	132	137	148	149	184	185	187
y	2.5	2.66	2.79	2.8	3.01	2.98	3.34	3.49	3.26

1. Calculate the sample correlation coefficient. How would you describe the nature of the relationship between these two variables?
2. If sheer force were to be expressed in pounds, what happens to the value of r?
3. If simple linear regression model were to be fit to this data, what proportion of observed variation in percent dry fiber weight could be explained by the model relationship?
4. Test at a 0.01 los for a positive linear correlation between these populations.

Example 12.52

x	1.5	1.5	2	2.5	2.5	3	3.5	3.5	4
y	23	24.5	25	30	33.5	40	40.5	47	49

X = Chlorine Flow

Y = Etch Rate

Model Residual Plots

The linear model for regression:

$$Y_i = \beta_o + \beta_1 x_i + \varepsilon_i \text{ , } i=1,..,n$$

where $\varepsilon_1, \ldots, \varepsilon_n \sim N(0,\sigma^2)$

This model yields the following assumptions:

1. Linear relationship between x and Y:

$$\mu_{Y \cdot x} = \beta_o + \beta_1 x$$

2. Equal variance for errors
3. Normally distributed errors
4. Independent errors

The estimated error (residual) may be used to test whether these assumptions are satisfied (i.e., the model is appropriate)

Recall:

$$e_i = y_i - \hat{y}_i$$
$$= \hat{\beta}_o + \hat{\beta}_1 x_i$$

Expectation and Variance of e_i

This leads to the standardized residual

$$e_i^* = \frac{y_i - \hat{y}_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}}}$$

If the assumptions are correct, the residuals should behave like normally distributed random variables and the standardized residuals like standard normal random variables.

To check the linearity and equal variance assumptions, plot e_i or e_i* against x_i or $\hat{y}_i$

The use of standardized residuals e_i* in these plots additionally provides some information about the normality assumption.

Good Residual Plots

Residual Plots w/ Nonlinear Data

Residual Plots w/ Unequal Varaiances

Residual Plots w/ Autocorrelation

To check the Independence assumption – In general, this is difficult to check. A plot of the residual vs. time of observation may be used.

To check the Normality Assumption – A Normal Probability Plot (NPP) of the residuals may be used. Recall, a linear plot indicates that the normal distribution is consistent with the data (residuals).

Forming an NPP for the residuals:

1. Order the residuals: $e_{(1)},\ldots,e_{(n)}$
2. Compute the normal percentiles:

$$P_i = \Phi^{-1}\left(\frac{i-.5}{n}\right)$$

3. Plot the (P_i, $e_{(i)}$) pairs

What If Some of the Assumptions Are Violated?

- Residual plot shows non-linearity – Fit a non-linear function (polynomial regression) or use a transformation to linearize (if possible)
- Residual plot supports linearity, but shows a violation of the equal variances assumption – Use weighted least squares (WLS); give less weight to observation with larger variance. Consult the text Applied Linear Regression Models as referenced in Lecture 17.

- The residuals support linearity and equal variances, but one of the standardized residuals is much greater (less) than +2 (-2) – This point is an outlier. If an assignable cause for this point may be found, throw it out and recalculate the regression parameters. If no assignable cause may be found, a MAD (minimum absolute deviation) approach may be used in place of L.S. (Least Squares). This approach, however, may be tedious.

- A plot of the residuals vs. time show a violation of the independence assumption – A transformation may be used (if possible) or the time variable may be included in the model via multiple regression. See Applied Linear Regression Models.
- A plot of the residuals vs. an independent variable not included in the model exhibits a definite pattern – Include this independent variable in a multiple regression analysis

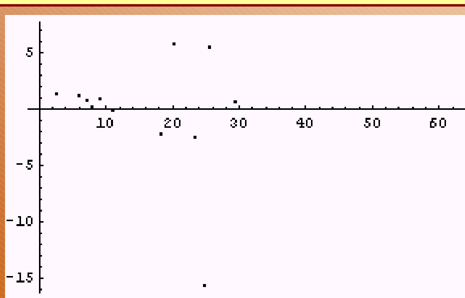
Example: (12.4) Cont'd

```
data = {{3, 4}, {8, 7}, {10, 8}, {11, 8}, {13, 10}, {16, 11},
  {27, 16}, {30, 26}, {35, 21}, {37, 9}, {38, 31}, {44, 30},
  {103, 75}, {142, 90}}
```

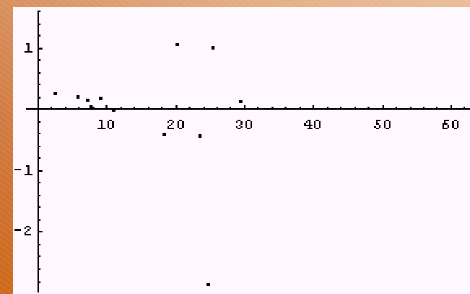
```
0.62614 + 0.65229 x
```

```
FitResiduals → {1.41699, 1.15554, 0.850958,
  0.198667, 0.894087, -0.0627837, -2.23798, 5.80515,
  -2.4563, -15.7609, 5.58683, 0.673091, 7.18797, -3.25135},
 StandardizedResiduals → {0.265657, 0.214712, 0.157622,
  0.0367446, 0.164908, -0.0115369, -0.407448, 1.05545,
  -0.446053, -2.86182, 1.01447, 0.122383, 1.49226, -0.926964}}
```

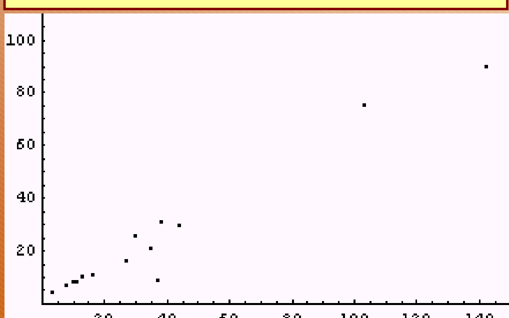
Residuals vs. Predicted



Standardized Residual vs. Predicted



Scatter Plot

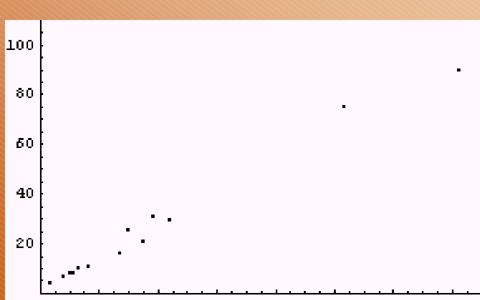


Revised Data Set – Outlier Omitted

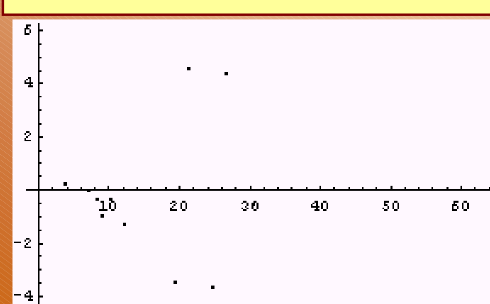
```
data2 = {{3, 4}, {8, 7}, {10, 8}, {11, 8}, {13, 10}, {16, 11},
  {27, 16}, {30, 26}, {35, 21}, {38, 31}, {44, 30}, {103, 75}, {142, 90}}
```

```
FitResiduals →
  {0.20667, -0.0550843, -0.359786, -1.01214, -0.316838, -1.27389,
   -3.44975, 4.5932, -3.66856, 4.37439, -0.539713, 5.97159, -4.47009},
 StandardizedResiduals → {0.0660442, -0.0174454, -0.113586,
   -0.319061, -0.0995999, -0.398958, -1.07037, 1.42319,
   -1.13533, 1.35368, -0.167239, 2.11475, -2.1816}}
```

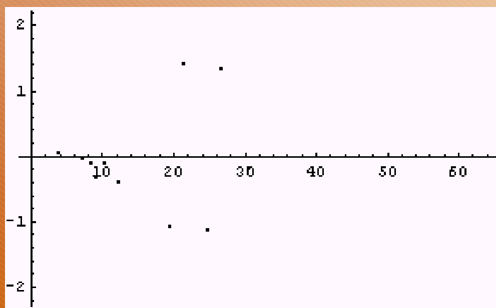

Scatter Plot



Residuals vs. Predicted



Standardized Residuals vs. Predicted



Multiple Regression

The objective of multiple regression is to build a probabilistic model that relates a dependent (response) variable y to more than one independent (predictor) variables x_i

Example: A particular steel company uses multiple regression to relate the dependent variable y = strength of hardened steel (psi) to the independent variables x_1= temperature of heat treatment (^{o}C) and x_2= length of time treatment was applied (hours)

General Multiple Regression Model

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$$

Mean Response:

$$\mu_{Y \cdot x_1^*,\ldots,x_n^*} = \beta_0 + \beta_1 x_1^* + \ldots + \beta_k x_k^*$$

Two Variable Models

First Order Model:

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

First Order Model with Interactions:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Two Variable Models Cont'd

Second Order Model:

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

Second Order Model with Interactions:

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

Data from Multiple Regression Model:

n observations: $(y_1,x_{11},\ldots,x_{k1})$, $(y_2,x_{12},\ldots,x_{k2})$, … , $(y_n,x_{1n},\ldots,x_{kn})$

Estimation of β's: Take partial derivatives of D wrt $b_0,\ldots,b_k$ to obtain k+1 equations with k+1 unknowns. The solution yields L.S. estimates of the β's

$$D = \sum_{i=1}^{n} \left[y_i - (b_o + b_1 x_{i1} + \cdots + b_k x_{ki}) \right]^2$$

Obtaining the ANOVA Table

Overall Measure of Fit

Coefficient of Determination:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Adjusted R^2:

$$adj.\, R^2 = \frac{(n-1)R^2 - k}{n-1-k}$$

Model Utility Test

To test the fit of the overall model, we can test

H_o:β_1=…= β_k=0 versus H_a: at least one $\beta_j \neq 0$

Use the ANOVA table for regression. The rejection region is

$$F = \frac{MSR}{MSE} = \frac{n-(k+1)}{k}\frac{R^2}{1-R^2} > F_{\alpha,k,n-(k+1)}$$

Inference Concerning β_j

To test H_o: $\beta_j = \beta_{jo}$ use the test statistic

$$t = \frac{\hat{\beta}_j - \beta_{j0}}{s_{\hat{\beta}_j}}$$

Under H_0, this test statistic is distributed as a t with n-(k+1) degrees of freedom. A test of H_o: $\beta_j = 0$ is used to see whether x_j should be included in the model.

Testing a set of β_j's

Formulate Two Models:

Full Model:

$$Y = \beta_0 + \ldots + \beta_l x_l + \ldots + \beta_k x_k + \varepsilon$$

Reduced Model:

$$Y = \beta_0 + \ldots + \beta_l x_l + \varepsilon$$

Testing a set of β_j's Cont'd

To choose between these models, we test

H_o: β_{l+1}=…= β_k= 0 versus

H_a: at least one β_{l+1} ,…, $\beta_k \neq 0$

Calculate the SSE for the Full and Reduced Models. (SSE_k and SSE_l respectively). The test statistic and rejection region are given by

$$F = \frac{\frac{SSE_l - SSE_k}{k-l}}{MSE_k} > F_{\alpha,k-l,n-(k+1)}$$

Confidence Intervals for the parameters β_j and the mean response $\mu_{Y \cdot x_1^*, \ldots, x_n^*}$, and Prediction Intervals for future Y at x=x* are calculated in the usual manner. Consult page 583 of the text for the specific form of these intervals.

Picking a Regression Model – Variable Selection

1. Use Scientific Knowledge of the Problem
2. (Full Enumeration) Use a summary measure of fit on a possible regression models (R^2, adj.R^2, and SSE). Select the model with the "best" measures comparatively.

3. (Backward Selection) Fit a model with all possible predictors included. Use t-tests for H_o: $\beta_j = 0$ to suggest candidate x_j predictors to omit. Eliminate the "least significant" predictor and fit a new model. Continue until all variables are needed. Note: One cannot eliminate more than one variable at a time on this basis

3. (Forward Selection) Build a model starting with the predictor most highly correlated with the response. Then find the best two-predictor model including this predictor, and so forth

Multicollinearity

Multicollinearity among the predictor variables is said to exist when these variables are highly correlated amongst themselves.

Effects of Multicollinearity:

1. In general, data that exhibits multicollinearity does not inhibit our ability to obtain a good fit or affect inferences about the mean response and future observation

2. In the presence of multicollinearity, The information obtained about the regression parameters, however, is imprecise. Hence the usual interpretation about these parameters in unwarranted (i.e. the effect of varying one parameter while holding the others constant).

Consult "Applied Linear Regression Models" for a detailed discussion of multicollinearity and possible remedies.

Detecting Multicollinearity

1. The value of R^2 is large, yet the t statistics for a particular β_j is small even though the predictor are known to significantly affect the response
2. The sign of a particular β_j is opposite to what intuition would suggest.

Multiple Regression Example

A hospital administrator wished to study the relation between patient satisfaction (Y) and the patient's age (X_1), severity of illness (X_2), and anxiety level (X_3). The administrator randomly selected 23 patients a collected the following data where larger values of Y, X_2, and X_3 are, respectively, associated with more satisfaction, increased severity of illness, and more anxiety. The data is of the form (X_1, X_2, X_3,Y).


```
{{50.0, 51.0, 2.3, 48}, {36.0, 46.0, 2.3, 57}, {40.0, 48.0, 2.2, 66},
 {41.0, 44.0, 1.8, 70}, {28.0, 43.0, 1.8, 89}, {49.0, 54.0, 2.9, 36},
 {42.0, 50.0, 2.2, 46}, {45.0, 48.0, 2.4, 54}, {52.0, 62.0, 2.9, 26},
 {29.0, 50.0, 2.1, 77}, {29.0, 48.0, 2.4, 89}, {43.0, 53.0, 2.4, 67},
 {38.0, 55.0, 2.2, 47}, {34.0, 51.0, 2.3, 51}, {53.0, 54.0, 2.2, 57},
 {36.0, 49.0, 2.0, 66}, {33.0, 56.0, 2.5, 79}, {29.0, 46.0, 1.9, 88},
 {33.0, 49.0, 2.1, 60}, {55.0, 51.0, 2.4, 49}, {29.0, 52.0, 2.3, 77},
 {44.0, 58.0, 2.9, 52}, {43.0, 50.0, 2.3, 60}}
```


Backward Elimination

{ParameterTable →

	Estimate	SE	TStat	PValue
1	189.105	237.11	0.79754	0.436814
x_1	-5.9021	2.81042	-2.10008	0.0519326
x_2	1.8482	12.616	0.146497	0.885359
x_3	-7.60405	128.813	-0.0590318	0.953658
x_1^2	0.0577719	0.0344071	1.67907	0.112558
x_2^2	-0.0253057	0.120932	-0.209257	0.836889
x_3^2	0.141337	26.9073	0.00525276	0.995874

,

RSquared → 0.721785, AdjustedRSquared → 0.617455, EstimatedVariance → 106.856,

ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	6	4435.53	739.255	6.91826	0.000916386
Error	16	1709.69	106.856		
Total	22	6145.22			

}

{ParameterTable →

	Estimate	SE	TStat	PValue
1	189.576	212.91	0.890404	0.385677
x_1	-5.90183	2.72606	-2.16497	0.0449073
x_2	1.79928	8.25637	0.217926	0.830081
x_3	-6.93054	11.9702	-0.57898	0.570195
x_1^2	0.0577692	0.033376	1.73086	0.101585
x_2^2	-0.0248369	0.0791632	-0.313743	0.757533

,

RSquared → 0.721785, AdjustedRSquared → 0.639957, EstimatedVariance → 100.57,

ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	5	4435.53	887.105	8.82076	0.000282143
Error	17	1709.69	100.57		
Total	22	6145.22			

{ParameterTable →

	Estimate	SE	TStat	PValue
1	253.763	57.4641	4.41603	0.000333435
x_1	-5.78281	2.63106	-2.1979	0.0412808
x_2	-0.77882	0.782546	-0.995238	0.332812
x_3	-6.99413	11.6649	-0.599586	0.556255
x_1^2	0.0563325	0.0322218	1.74828	0.0974511

,

RSquared → 0.720174, AdjustedRSquared → 0.657991, EstimatedVariance → 95.5328,

ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	4	4425.63	1106.41	11.5814	0.0000787018
Error	18	1719.59	95.5328		
Total	22	6145.22			

ParameterTable →	Estimate	SE	TStat	PValue
1	259.236	55.7702	4.64828	0.000175241
x_1	-5.94775	2.57216	-2.31236	0.0321275
x_2	-1.12356	0.521828	-2.15311	0.0443675
x_1^2	0.0578661	0.031574	1.83271	0.0825647

RSquared → 0.714585, AdjustedRSquared → 0.66952, EstimatedVariance → 92.3124,

ANOVATable →	DF	SumOfSq	MeanSq	FRatio	PValue
Model	3	4391.28	1463.76	15.8566	0.0000208921
Error	19	1753.94	92.3124		
Total	22	6145.22			

ParameterTable →	Estimate	SE	TStat	PValue
1	166.591	24.9084	6.68815	1.64798×10^{-6}
x_1	-1.26046	0.289186	-4.35864	0.000304217
x_2	-1.08932	0.551389	-1.97559	0.0621629

RSquared → 0.664129, AdjustedRSquared → 0.630542, EstimatedVariance → 103.2,

ANOVATable →	DF	SumOfSq	MeanSq	FRatio	PValue
Model	2	4081.22	2040.61	19.7734	0.0000182692
Error	20	2064.	103.2		
Total	22	6145.22			

ParameterTable →	Estimate	SE	TStat	PValue
1	121.832	11.0422	11.0333	3.37134×10^{-10}
x_1	-1.52704	0.272881	-5.59598	0.0000148907

RSquared → 0.598585, AdjustedRSquared → 0.57947, EstimatedVariance → 117.466,

ANOVATable →	DF	SumOfSq	MeanSq	FRatio	PValue
Model	1	3678.44	3678.44	31.315	0.0000148907
Error	21	2466.78	117.466		
Total	22	6145.22			

Forward Selection

ParameterTable →	Estimate	SE	TStat	PValue
1	121.832	11.0422	11.0333	3.37134×10^{-10}
x_1	-1.52704	0.272881	-5.59598	0.0000148907

RSquared → 0.598585, AdjustedRSquared → 0.57947, EstimatedVariance → 117.466,

ANOVATable →	DF	SumOfSq	MeanSq	FRatio	PValue
Model	1	3678.44	3678.44	31.315	0.0000148907
Error	21	2466.78	117.466		
Total	22	6145.22			

ParameterTable →	Estimate	SE	TStat	PValue
1	173.614	33.8724	5.12553	0.0000445882
x_2	-2.21072	0.664581	-3.32649	0.00320519

RSquared → 0.345091, AdjustedRSquared → 0.313905, EstimatedVariance → 191.646,

ANOVATable →	DF	SumOfSq	MeanSq	FRatio	PValue
Model	1	2120.66	2120.66	11.0655	0.00320519
Error	21	4024.56	191.646		
Total	22	6145.22			

ParameterTable →	Estimate	SE	TStat	PValue
1	137.432	22.1878	6.19402	3.81763×10^{-6}
x_3	-33.1427	9.58524	-3.45768	0.00235594

RSquared → 0.362778, AdjustedRSquared → 0.332434, EstimatedVariance → 186.47,

ANOVATable →	DF	SumOfSq	MeanSq	FRatio	PValue
Model	1	2229.35	2229.35	11.9555	0.00235594
Error	21	3915.87	186.47		
Total	22	6145.22			

ParameterTable →

	Estimate	SE	TStat	PValue
1	166.591	24.9084	6.68815	1.64798×10^{-6}
x_1	-1.26046	0.289186	-4.35864	0.000304217
x_2	-1.08932	0.551389	-1.97559	0.0621629

RSquared → 0.664129, AdjustedRSquared → 0.630542, EstimatedVariance → 103.2,

ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	2	4081.22	2040.61	19.7734	0.0000182692
Error	20	2064.	103.2		
Total	22	6145.22			

ParameterTable →

	Estimate	SE	TStat	PValue
1	147.075	16.7334	8.78929	2.6445×10^{-8}
x_1	-1.24336	0.29612	-4.19884	0.000441918
x_3	-15.8906	8.2556	-1.92483	0.0685932

RSquared → 0.661324, AdjustedRSquared → 0.627457, EstimatedVariance → 104.062,

ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	2	4063.98	2031.99	19.5268	0.0000198536
Error	20	2081.24	104.062		
Total	22	6145.22			

ParameterTable →

	Estimate	SE	TStat	PValue
1	209.232	55.1213	3.79585	0.00113345
x_1	-6.02522	2.79593	-2.155	0.0435276
x_1^2	0.0554324	0.0343023	1.616	0.12176

RSquared → 0.644946, AdjustedRSquared → 0.60944, EstimatedVariance → 109.094,

ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	2	3963.33	1981.67	18.1647	0.0000318383
Error	20	2181.89	109.094		
Total	22	6145.22			

ParameterTable →

	Estimate	SE	TStat	PValue
1	162.876	25.7757	6.31898	4.59181×10^{-6}
x_1	-1.21032	0.301452	-4.01497	0.000740441
x_2	-0.665906	0.820997	-0.811094	0.427356
x_3	-8.61303	12.2413	-0.703607	0.490211

RSquared → 0.672659, AdjustedRSquared → 0.620973, EstimatedVariance → 105.873,

ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	3	4133.63	1377.88	13.0145	0.0000748239
Error	19	2011.58	105.873		
Total	22	6145.22			

ParameterTable →

	Estimate	SE	TStat	PValue
1	259.236	55.7702	4.64828	0.000175241
x_1	-5.94775	2.57216	-2.31236	0.0321275
x_2	-1.12356	0.521828	-2.15311	0.0443675
x_1^2	0.0578661	0.031574	1.83271	0.0825647

RSquared → 0.714585, AdjustedRSquared → 0.66952, EstimatedVariance → 92.3124,

ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	3	4391.28	1463.76	15.8566	0.0000208921
Error	19	1753.94	92.3124		
Total	22	6145.22			

ParameterTable →

	Estimate	SE	TStat	PValue
1	148.57	147.628	1.00638	0.326878
x_1	-1.26127	0.296651	-4.25169	0.000431349
x_2	-0.564215	4.27428	-0.132002	0.89637
x_2^2	-0.0000643321	0.000519054	-0.123941	0.902664

RSquared → 0.664401, AdjustedRSquared → 0.611411, EstimatedVariance → 108.544,

ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	3	4082.89	1360.96	12.5384	0.000094299
Error	19	2062.33	108.544		
Total	22	6145.22			

Reduced Sets of β_j's

{ParameterTable →

	Estimate	SE	TStat	PValue
1	189.105	237.11	0.79754	0.436814
x_1	-5.9021	2.81042	-2.10008	0.0519326
x_2	1.8482	12.616	0.146497	0.885359
x_3	-7.60405	128.813	-0.0590318	0.953658 ,
x_1^2	0.0577719	0.0344071	1.67907	0.112558
x_2^2	-0.0253057	0.120932	-0.209257	0.836889
x_3^2	0.141337	26.9073	0.00525276	0.995874

RSquared → 0.721785, AdjustedRSquared → 0.617455,
EstimatedVariance → 106.856, ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	6	4435.53	739.255	6.91826	0.000916386
Error	16	1709.69	106.856		
Total	22	6145.22			

}

{ParameterTable →

	Estimate	SE	TStat	PValue
1	162.876	25.7757	6.31898	4.59181×10^{-6}
x_1	-1.21032	0.301452	-4.01497	0.000740441 ,
x_2	-0.665906	0.820997	-0.811094	0.427356
x_3	-8.61303	12.2413	-0.703607	0.490211

RSquared → 0.672659, AdjustedRSquared → 0.620973,
EstimatedVariance → 105.873, ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	3	4133.63	1377.88	13.0145	0.0000748239
Error	19	2011.58	105.873		
Total	22	6145.22			

}

{ParameterTable →

	Estimate	SE	TStat	PValue
1	121.832	11.0422	11.0333	3.37134×10^{-10},
x_1	-1.52704	0.272881	-5.59598	0.0000148907

RSquared → 0.598585, AdjustedRSquared → 0.57947,
EstimatedVariance → 117.466, ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	1	3678.44	3678.44	31.315	0.0000148907
Error	21	2466.78	117.466		
Total	22	6145.22			

}

All "Possible" Models; X_1,X_2 Only

{ParameterTable →

	Estimate	SE	TStat	PValue
1	162.361	129.875	1.25013	0.22643
x_1	-1.15699	3.12827	-0.369851	0.715584,
x_2	-1.00566	2.5807	-0.389685	0.701103
$x_1 x_2$	-0.00202929	0.061078	-0.0332245	0.973842

RSquared → 0.664149, AdjustedRSquared → 0.61112,
EstimatedVariance → 108.625, ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	3	4081.34	1360.45	12.5242	0.0000949578
Error	19	2063.88	108.625		
Total	22	6145.22			

}

{ParameterTable →

	Estimate	SE	TStat	PValue
1	166.591	24.9084	6.68815	1.64798×10^{-6}
x_1	-1.26046	0.289186	-4.35864	0.000304217 ,
x_2	-1.08932	0.551389	-1.97559	0.0621629

RSquared → 0.664129, AdjustedRSquared → 0.630542,
EstimatedVariance → 103.2, ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	2	4081.22	2040.61	19.7734	0.0000182692
Error	20	2064.	103.2		
Total	22	6145.22			

}

{ParameterTable →

	Estimate	SE	TStat	PValue
1	121.832	11.0422	11.0333	3.37134×10^{-10},
x_1	-1.52704	0.272881	-5.59598	0.0000148907

RSquared → 0.598585, AdjustedRSquared → 0.57947,
EstimatedVariance → 117.466, ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	1	3678.44	3678.44	31.315	0.0000148907
Error	21	2466.78	117.466		
Total	22	6145.22			

}

		Estimate	SE	TStat	PValue
{ParameterTable →	1	173.614	33.8724	5.12553	0.0000445882,
	x_2	-2.21072	0.664581	-3.32649	0.00320519

RSquared → 0.345091, AdjustedRSquared → 0.313905,

EstimatedVariance → 191.646, ANOVATable →

	DF	SumOfSq	MeanSq	FRatio	PValue
Model	1	2120.66	2120.66	11.0655	0.00320519
Error	21	4024.56	191.646		
Total	22	6145.22			

}

Multicollinearity Example

The following data is a portion of that from a study of the relation of the amount of body fat (Y) to the predictor variables (X_1) Tricep skinfold thickness, (X_2) Thigh circumference, and (X_3) Midarm circumference based on a sample of 20 healthy females 25-34 years old.

Subject	Triceps	Thigh	Midarm	Body Fat
1	19.5	43.1	29.1	11.9
2	24.7	49.8	28.2	22.8
3	30.7	51.9	37	18.7
…	…	…	…	…
18	30.2	58.6	24.6	25.4
19	22.7	48.2	27.1	14.8
20	25.2	51	27.5	21.1

The L.S. regression coefficients for X_1 and X_2 of various models are given in the table

Variables in Model	b1	b2
X1	0.8572	…
X2	…	0.8565
X1, X2	0.224	0.6594
X1, X2, X3	4.334	-2.857

Hence, the regression coefficient of one variable depends upon which other variables are in the model and which ones are not. Therefore, a regression coefficient does not reflect any inherent effect of particular predictor variable on the response variable (Only a partial effect, given what other variables are included)