Notes:

Correlation and Regression

Key Learning Points

1. Describe the importance of correlation and regression.
2. Explain why to use correlation and regression.
3. Utilize correlation and regression in improvement projects.

What are Correlation and Regression?

Correlation is the measure of strength and direction of the association between two continuous variables. Minitab offers two methods of correlation: the Pearson correlation coefficient and Spearman's rank correlation coefficient. The Pearson correlation (also known as r), which is the most common method, measures the linear relationship between two continuous variables. Spearman correlation is noparametric and does not require a strictly linear relationship.
Regression creates a prediction equation based on the relationship between predictors and a response variable.

Use in Projects

Correlation and regression are used to quantify the influence of a continuous factor on a continuous response. Resulting prediction equations help in the design of effective improvement strategies.

Scatter Diagrams

Scatter diagrams can show complex relationships that help avoid statistical mistakes.

Definitions

Correlation: A technique through which you can quantify the strength of linear association between an output variable and an input variable via the Correlation Coefficient (r).

Regression Equation: A prediction equation, not necessarily linear, which allows the values of inputs to be used to predict a corresponding output.

Coefficient of Determination: R2 represents the adequacy of the regression model or the percentage of total variation in the response variable explained by the regression equation.

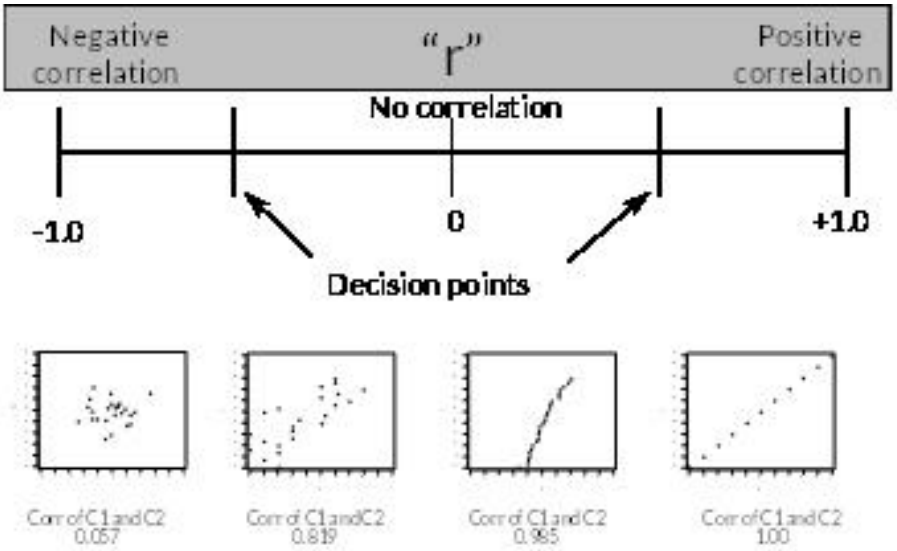
How Much Does X Impact The Y?

The most common way to measure association is using the Pearson Correlation Coefficient, r.

- -1< r < 1, where r=|1| is a perfect correlation.
- If r = 0, there is no correlation between x and y.
- If r < 0, it is a negative correlation, and y decreases as x increases.
- If r > 0, it is a positive correlation, and y increases as x increases.

The Correlation Formula

You can quantify the association between X and Y with the correlation coefficient r.



Correlation Relationship

The correlation relationship decision table is used to show decision points in determining if a relationship is significant. The decision point is the minimum value of the correlation coefficient that would be significant for the given sample size.

Notes:

Notes:

The Correlation Test is based on the hypotheses:

- H_o: There is no relationship between X and Y
- H_a: There is a relationship between X and Y

The P-Value may be used to evaluate the significance of the relationship:
P-Value ≤ α, reject the null (relationship is significant)

Correlation Relationship Decision Table

Guideline:

- If | r | > 0.80, then the relationship is important.
- If | r | < 0.20, then the relationship is not significant.

n	Decision Point	n	Decision Point
5	0.878	18	0.468
6	0.811	19	0.456
7	0.754	20	0.444
8	0.707	22	0.423
9	0.666	24	0.404
10	0.632	26	0.388
11	0.602	28	0.374
12	0.576	30	0.361
13	0.553	40	0.312
14	0.532	50	0.279
15	0.514	60	0.254
16	0.479	80	0.220
17	0.482	100	0.196

Correlation Data Requirements

A correlation study must have two pieces of connected data expressed mathematically in bivariate ordered pairs: X and Y (X, Y).

- The predictor variable, X, is also known as the independent variable.
- The response variable, Y, is also known as the dependent variable.
- If correlated, the value of Y will depend on the level of X.

Note: Correlation studies can also be done with more than one X.

Example: Temperature vs. Mortality

Is there evidence of a relationship between the mean annual temperature in the region and the mortality index?

- **Minitab: Stat > Basic Statistics > Correlation**
  - Variables: Temperature, Mortality

- Method: Pearson correlation
- Display P-Values

Results

- H_o: There is no relationship between X and Y
- H_a: There is a relationship between X and Y

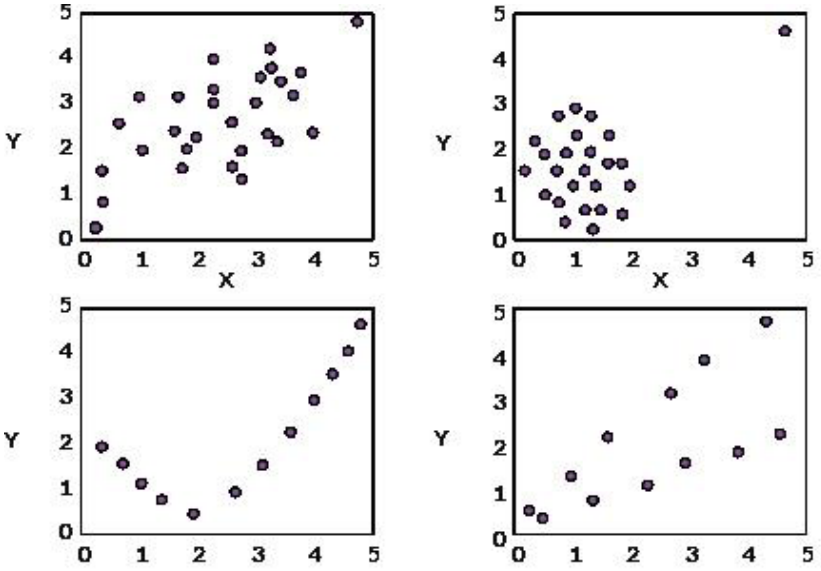
Correlation: Temperature, Mortality

Correlations

- Pearson Correlation: 0.875
- P-value: 0.000

Since the P-Value is 0.000 which is lower than 0.05, you can reject the null hypothesis, and there is a linear relationship between X and Y. The practical conclusion is that there is a significant correlation between Regional Temperature and Mortality.

Identical Values of Correlation Coefficient



Numerical summaries without a graphic picture can be very misleading. This adaptation from Chambers, Cleveland, Kleiner, and Tukey's book Graphical Methods for Data Analysis (Duxbury Press, 1983), shows four very different scatter diagrams that all yield the same value for the correlation coefficient.

What If You Could Determine A Model To Predict and Control Future Performance?

Regression allows you to create an equation based on experience to predict an outcome.

Notes:

Regression Model Example: Time to Close an Order

The time to close an order = (5.2 x size of the order) + (3 x number of service calls this year) - (2.7 x sales rep's experience)

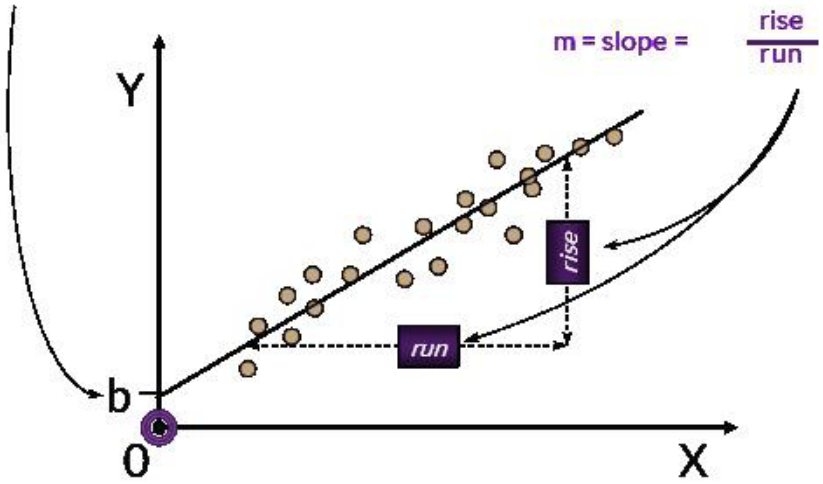
Regression Model Example: Variability of Forecast

The variability of forecast = (3.8 x week of the month) + (2.5 x production change of the past month) + (1.7 x variability of past quarter)

Linear Relationship

A simple linear relationship can be described mathematically by the equation: Y = mX + b

b = Y intercept = the Y value when the line intersects Y axis at X = 0



Residuals

Residuals are the difference between the actual response value and the fitted predicted value from the equation.

A fitted-value line is drawn according to the principle of least squares. This means that the sum of the squares of the distances from the data point to the line (parallel to the y axis) is minimized.

Residuals estimate an inability to predict. They are measured parallel to the Y axis.

In a valid model, the residuals will be normally distributed about a mean of zero, independent of the X and Y values, and randomly distributed over time.

Regression Example

Obtain a prediction equation for mass of fiber as a function of speed of discharge from spinneret.

- **Minitab: Stat > Regression > Fitted Line Plot**

Notes:

- Response Y: Mass
- Response X: Speed
- Type of Regression Model: Linear
- Graphs>Residual Plots: Four In One
- Options>Display Confidence Interval & Prediction Interval

Session Window Results

Regression Analysis: Mass versus Speed

The regression equation is
Mass = 18.00 + 4.784 Speed

Model Summary

S	R-sq	R-sq(adj)
1.35473	69.54%	67.94%

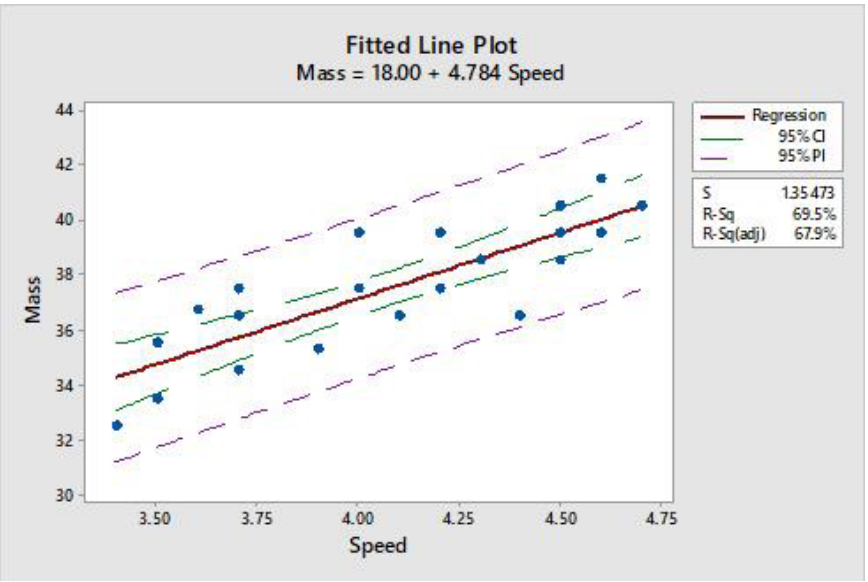
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	79.610	79.6096	43.38	0.000
Error	19	34.870	1.8353		
Total	20	114.480			

Fitted Line: Mass versus Speed

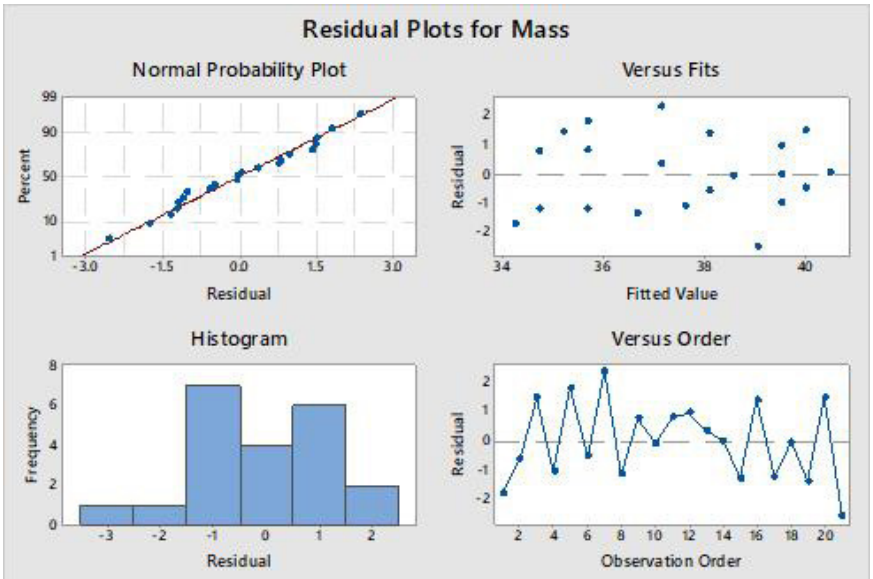
Residual Plots for Mass

Fitted Line Plot



Notes:

Residuals



Multiple Linear Relationship

A multiple-linear relationship allows you to fit points to a plane rather than a simple line. The equation of a plane is Y = b + (m1)(X1) + (m2)(X2)

The best-fit plane is determined by minimizing the sum of the squared differences between the actual data points and the plane.
The best-fit plane is centered in the plotted data with the actual data points dispersed randomly above and below the plane.

Notes

Multiple Regression with more than 2 Xs

- Beyond two Xs (three dimensions), the best-fit shape cannot be visualized; however, the principles of regression still apply.
- Fit the best "hyperplane" with the equation, Y = b + (m1)(X1) + (m2) (X2) + ... + (mn)(Xn)

Multiple Regression with an interaction

- An interaction will change the tilt in the plane and put a kink in the plane.
- Fit the best "hyperplane" with the equation, Y = b + (m1)(X1) + (m2)(X2) + (m3) (X1)(X2).

Multiple Regression with curvature

- Curvature results from higher order terms.
- Fit a curved plane with the equation, Y = b+(m1)X1+(m2)X2+(m3) (X1*X2)+(m4)X1**2+(m5)X2**2.

Notes:

Multiple Regression

You can use multiple regression to extract a lot of information from historical data.

You can gain information on interactions and higher order terms and obtain a predictive model, but the data will likely be suboptimal.

Designed experimentation is the best way to estimate effects and interactions.

Model Building in Multiple Regression

Stepwise regression is a great tool for building multiple regression models one factor at a time. Be sure to use the smallest model that can predict your response. Evaluate your model with:

- P-values for each predictor. Remove all with p>.05
- R^2, the percentage of total variation in Y explained by the model.
- R^2 adj – which is adjusted for the number of terms in the model. If it does not increase with the addition of a term, do not add the term.
- Variation Inflation Factor (VIF) - determines if X variables are correlated and interactions should not be estimated. If VIF>10, the model is compromised. Some predictors are correlated with each other and should be removed.

Extrapolation

A regression equation can be used to predict a response within the range of the variables that were tested. A certain degree of extrapolation should be considered to explore "what if." Consider extrapolations where each individual variable is within its historical range, but the confirmation experiment's points are outside the elliptical region formed by a scatter plot of two Xs.

Multiple Regression Example

Obtain a prediction equation for graduate GPA as a function of undergraduate GPA and GMAT score.

- **Minitab: Stat > Regression > Regression > Fit Regression Model**
  - Responses: Graduate GPA
  - Continuous Predictors: Undergraduate GPA, GMAT Score

Notes:

Session Window Results

Regression Analysis: Graduate GPA versus ... aduate GPA, GMAT Score

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	1.06569	0.53285	24.02	0.001
Undergraduate GPA	1	0.15980	0.15980	7.20	0.031
GMAT Score	1	0.31183	0.31183	14.05	0.007
Error	7	0.15531	0.02219		
Lack-of-Fit	6	0.13531	0.02255	1.13	0.617
Pure Error	1	0.02000	0.02000		
Total	9	1.22100			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.148952	87.28%	83.65%	73.76%

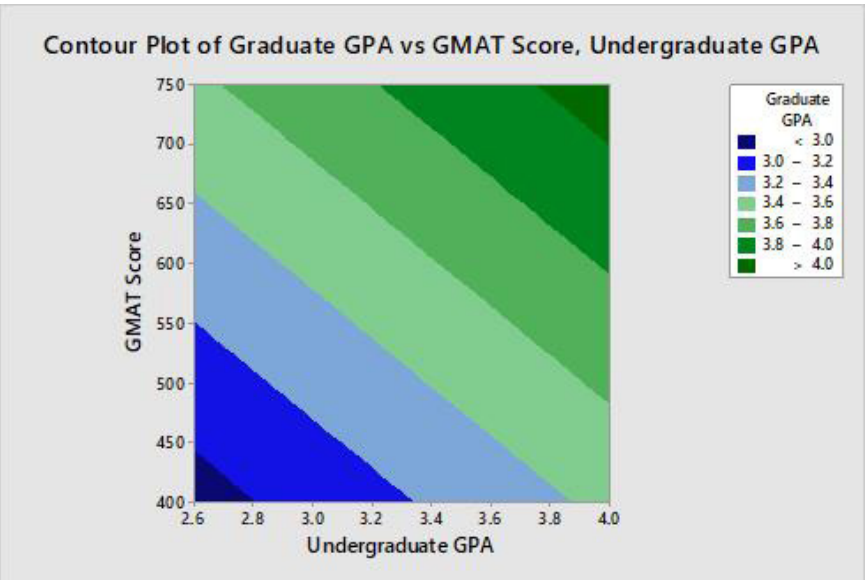
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.210	0.360	3.36	0.012	
Undergraduate GPA	0.376	0.140	2.68	0.031	1.47
GMAT Score	0.001837	0.000490	3.75	0.007	1.47

Regression Equation

Graduate GPA = 1.210 + 0.376 Undergraduate GPA + 0.001837 GMAT Score

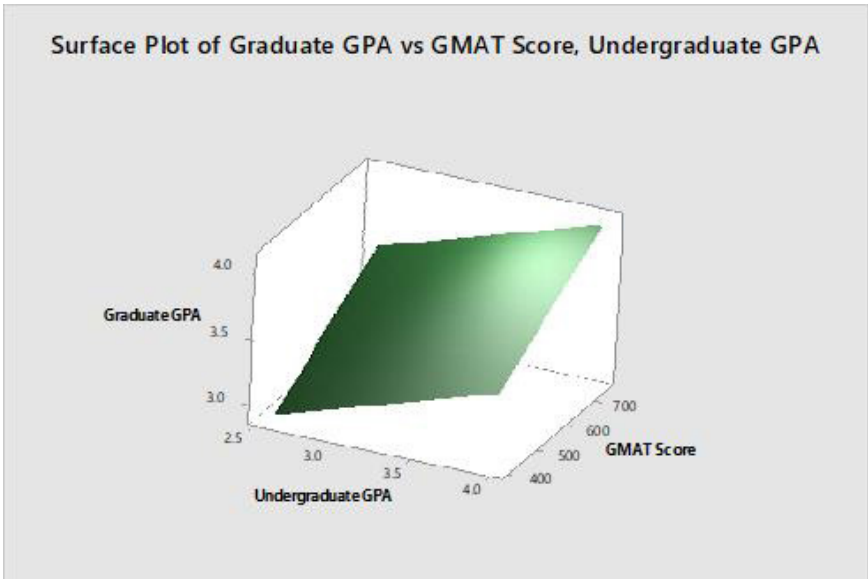
Contour Plot



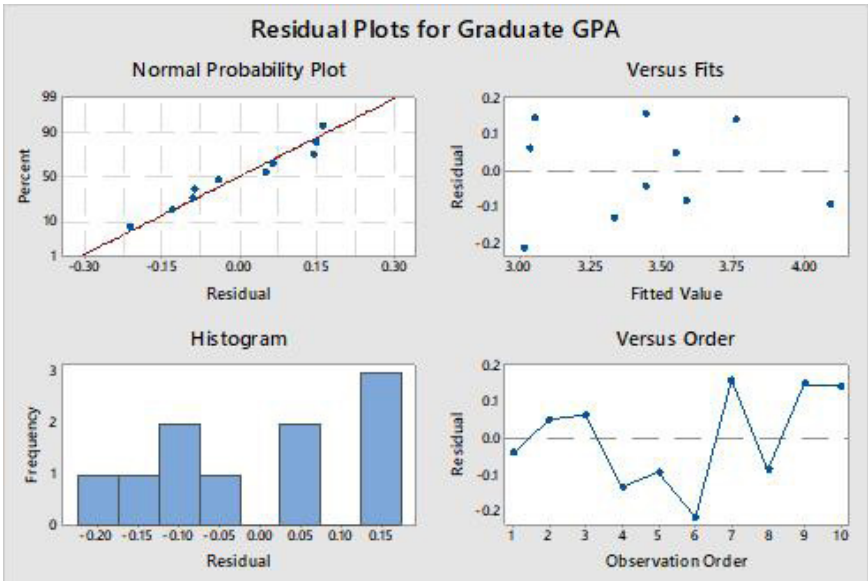
Notes:

Notes:

Surface Plot



Residuals



When Should Regression Be Used?

Use Regression to create prediction models for continuous responses using one or more continuous or categorical predictors.

Pitfalls to Avoid

- Follow the parsimony principle. Use the simplest model that predicts your response.
- Use all of the diagnostic statistics
  - p-values
  - R-square

  - Residual Analysis
  - R-square adjusted (multiple)
  - VIF (multiple)
- Do not extrapolate outside the design region without verifying tests.

Notes: