1 Introduction to Regression

Regression analysis is used to create an empirical model of the relationship between process inputs (x's) and outputs (y's).

- It is the method for analyzing designed experiments.
- It can also be used with historical data to help identify some factors for an experiment, or to develop an empirical model with that data.

Topics:

- Terminology
- Purposes of regression analysis
- Data collection for use in regression analysis
- The line of best fit
- Simple Regression

Notes

Terminology

- The term ***correlation*** is often used any time we speak of relating one variable to another
  - Correlation is a measure of the relationship
  - An input/output relationship between the two variables is not required (for example, two variables measured at the same point in a process)
  - As a result, unrelated things can be "correlated." Remember, correlation *does not* prove causation.
- ***Regression*** analysis yields a model equation of the input-output relationship, $Y = f(X)$, which can be useful in prediction
  - In the dataset, a series of inputs and their resulting output measures are aligned
  - Regression is used to investigate and model the relationship

Notes

Purposes of regression analysis

The result of regression analysis is an empirical model, created from the data/observations, that can be used to:

- Understand and describe the relationship between Y and X's
- Predict Y from X's
- Determine best setting for X's (optimization)
- Reduce variation in Y by controlling X's

Notes

Data collection for use in regression analysis

Regression analysis is only as good as the data used.

Three basic sources of data are:

- Historical data (data that exists in routine collection systems)
- An observational study (data collected from uncontrolled processes for a specific purpose)
- A designed experiment (data from structured and controlled tests)

Regression analysis is a very big statistical topic and is commonly the analysis type for data from all three sources listed above.

Designs of experiments (DOEs) is the best strategy for many problems we are trying to solve as it is constructed to eliminate many of the problems that exist with the first two sources. However, historical and observational data is often easier to get and can still give powerful insights, although care must be taken with the analysis and conclusions drawn.

Notes

Considerations when using historical data

Historical data is often plentiful and easily accessible.

- It may be useful in identifying some variables that are critical to our process

However, there are several potential issues in using it:

- Some relevant data is not available, such as values of critical x's that are not recorded as part of the on-going process
- Reliability of the data is often questionable, including data being missing or lost
- The nature of the data is not helpful in solving the problem, as in situations when an x variable is controlled, so its impact cannot be seen in the regression analysis
- Often, data is used in ways that were not intended, such as using available data as a surrogate for what was really needed

Caution: We will not be able to cover the many aspects of creating and validating regression models from historical data in this course. If you choose to do this, proceed with caution! Better yet, get additional help.

Notes

Considerations when using an observational study

In an observational study, we would observe the process, with as little interaction or disturbance as possible, in order to obtain the data.

- With adequate planning, an observational study can yield accurate, complete, reliable data
- These studies can lead to ideas on what might be impacting the process
- However, these studies often provide limited information about specific relationships of interest, such as the impact of a variable that is tightly controlled in normal operation

Notes

"Simple" regression

Simple linear regression refers to the case when there is only one regressor (variable) x used.

- In simple regression, the model equation is for a best-fit line
- The form of the model equation created is:

$$Y = b_0 + b_1 x_1 + error$$

where b_0 is the intercept and b_1 is the slope of the line.

- This may remind you of your early algebra days, when you learned the equation for a line between two points:

$$Y = mx + b$$

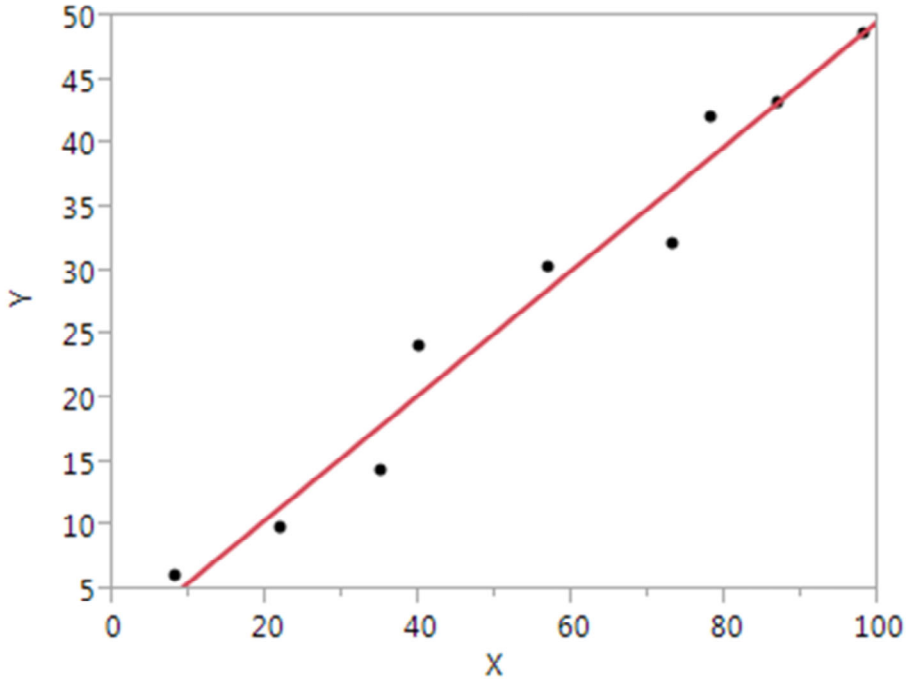
- Because there is variation (and more than two points to create the line), there will be scatter around the best-fit line determined by regression analysis.

Notes

Simple regression (cont'd)

Intercept ↓ Slope ↓

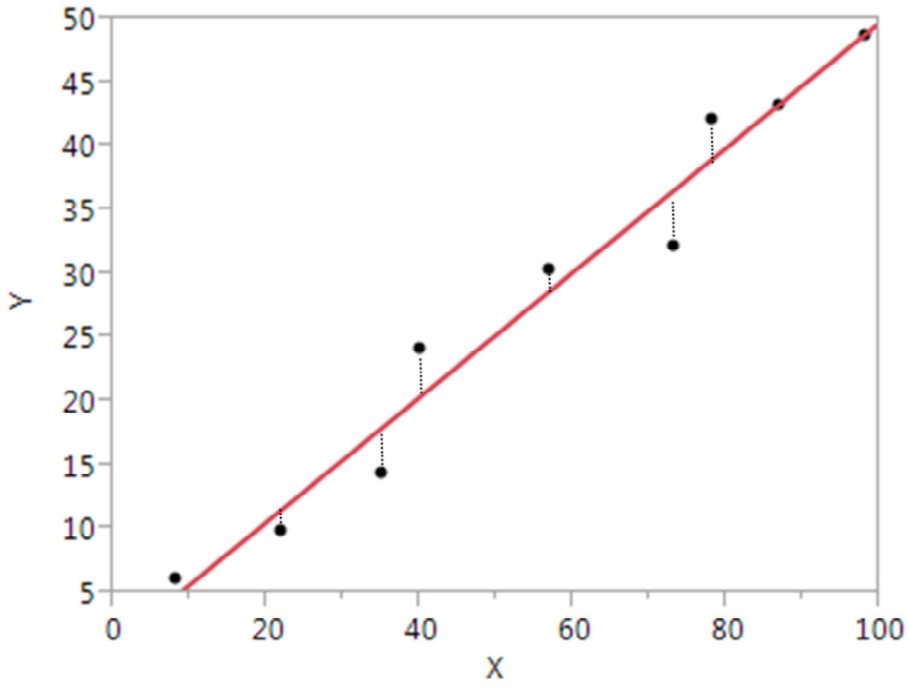
$$Y = 0.8387 + 0.4891\,X + \text{“Error”}$$



Notes

The line of best fit

The best-fitting line is the one that minimizes the sum of the squared "errors"



Notes

The line of best fit (cont'd)

- **"Errors" are the vertical distances between each Y data value and the fitted line**
- The line of best fit is the one that minimizes the sum of the squared errors
- This is the simplest example of *least-squares model fitting*
- The fitted line is often referred to as the predicted Y value

Notes

LSSV2 student files \ ANOVA linear fit

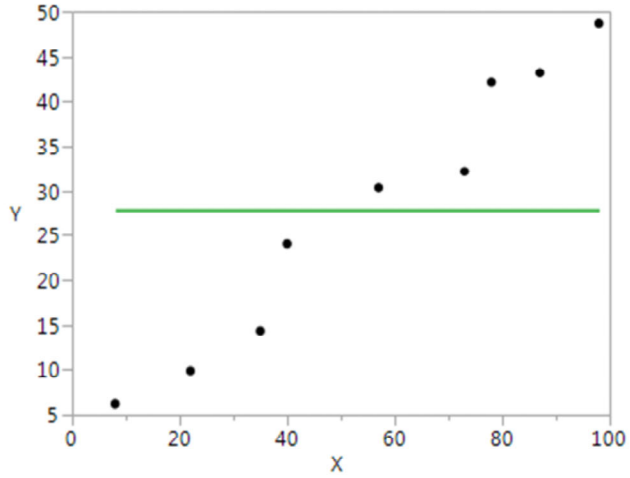
Worksheet \ *Prediction & error 1*

	X data		Y data		Prediction		Error
							Y = 27.9033 + 0.0000 X
	8		6.16		27.90		-21.74
	22		9.88		27.90		-18.02
	35		14.35		27.90		-13.55
	40		24.06		27.90		-3.84
	57	=	30.34	=	27.90	+	2.44
	73		32.17		27.90		4.27
	78		42.18		27.90		14.28
	87		43.23		27.90		15.33
	98		48.76		27.90		20.86
Sum of squares (SS)			8901.3	=	7007.4	+	1893.9
Degrees of freedom (DF)			9	=	1	+	8
Root mean square error (RMSE)							15.39
Average Y	27.90						
STDEV of Y	15.39						

Notes

Finding the line of best fit (cont'd)

In this worksheet we ignore the X variable completely, and use the average value of Y as the prediction. This is just the calculation of the mean and standard deviation of the Y variable. (The values in cells I14 and E17 are the same.)



The sum of the squared errors (cell I12) can be dramatically reduced by using the X variable to "explain" more of the variation in the Y variable.

Notes

Finding the line of best fit (cont'd)

Worksheet \ *Prediction & error 2*

	X data		Y data		Prediction		Error
	8		6.16	=	4.75	+	1.41
	22		9.88	=	11.60	+	-1.72
	35		14.35	=	17.96	+	-3.61
	40		24.06	=	20.40	+	3.66
	57		30.34	=	28.72	+	1.62
	73		32.17	=	36.54	+	-4.37
	78		42.18	=	38.99	+	3.19
	87		43.23	=	43.39	+	-0.16
	98		48.76	=	48.77	+	-0.01
Sum of squares (SS)			8901.3	=	8838.0	+	63.3
Degrees of freedom (DF)			9	=	2	+	7
Root mean square error (RMSE)							3.007

Y = 0.8387 + 0.4891 X

Average Y 27.90
STDEV of Y 15.39
Adjusted R square 0.962

Proportion of total Y variation caused by ("explained by") X variation

Notes

Degrees of freedom for regression

N = total sample size

G = number of parameters in the equation
= DF for the prediction column

N – G = DF for the error column

- The *Error* DF is more important than the *Prediction* DF
- It determines the accuracy of the predicted values
- When DF is mentioned without a qualifier, it usually means *Error* DF

Notes

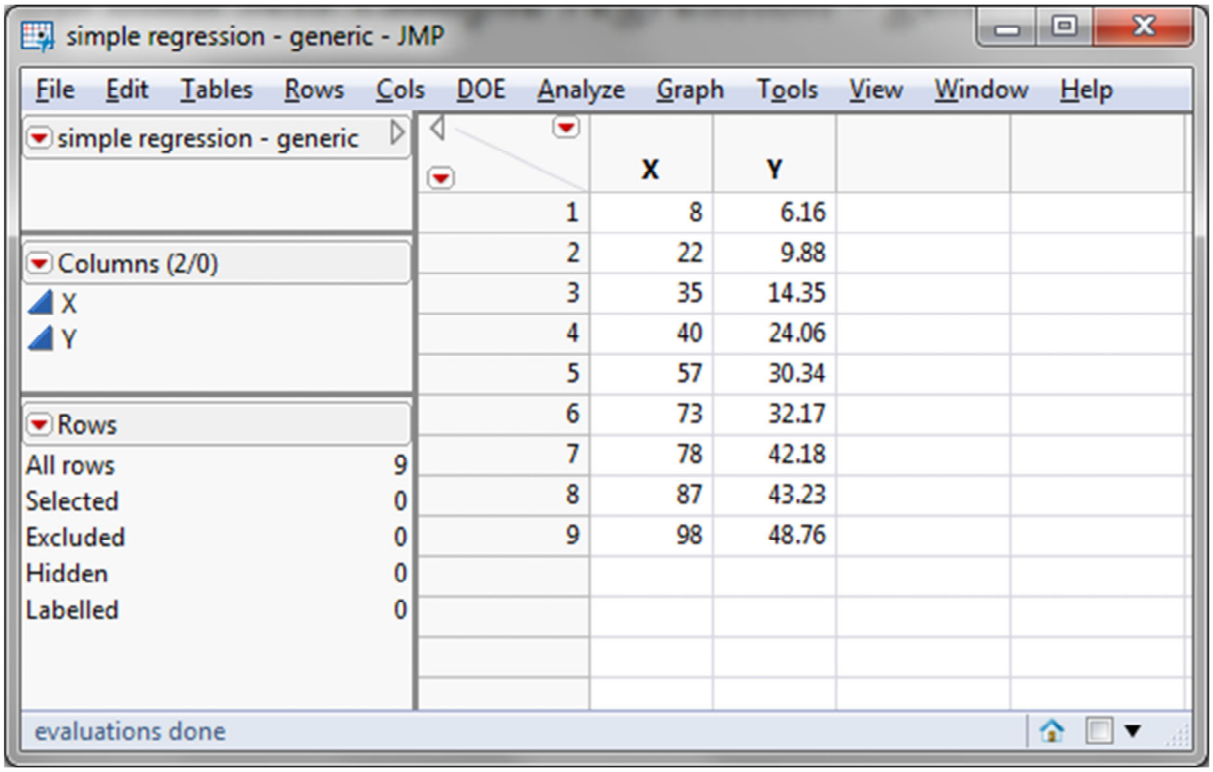
Steps in Simple Regression

1. Run Analyze > Fit Model in JMP to investigate the relationship between y and x
2. Check the p-value for the fit to determine whether the regression is significant. If not, then no need to go further.
3. If the regression is significant, determine the strength of the relationship, using the $Adjusted\ R^2$
4. Check model adequacy by reviewing the residuals plots
   - Residual Normal Quantile Plot
   - Residual by Predicted Plot
   - Studentized Residuals (in run order)

We'll go through these steps and additional analysis details, for simple regression in the following example.

Notes

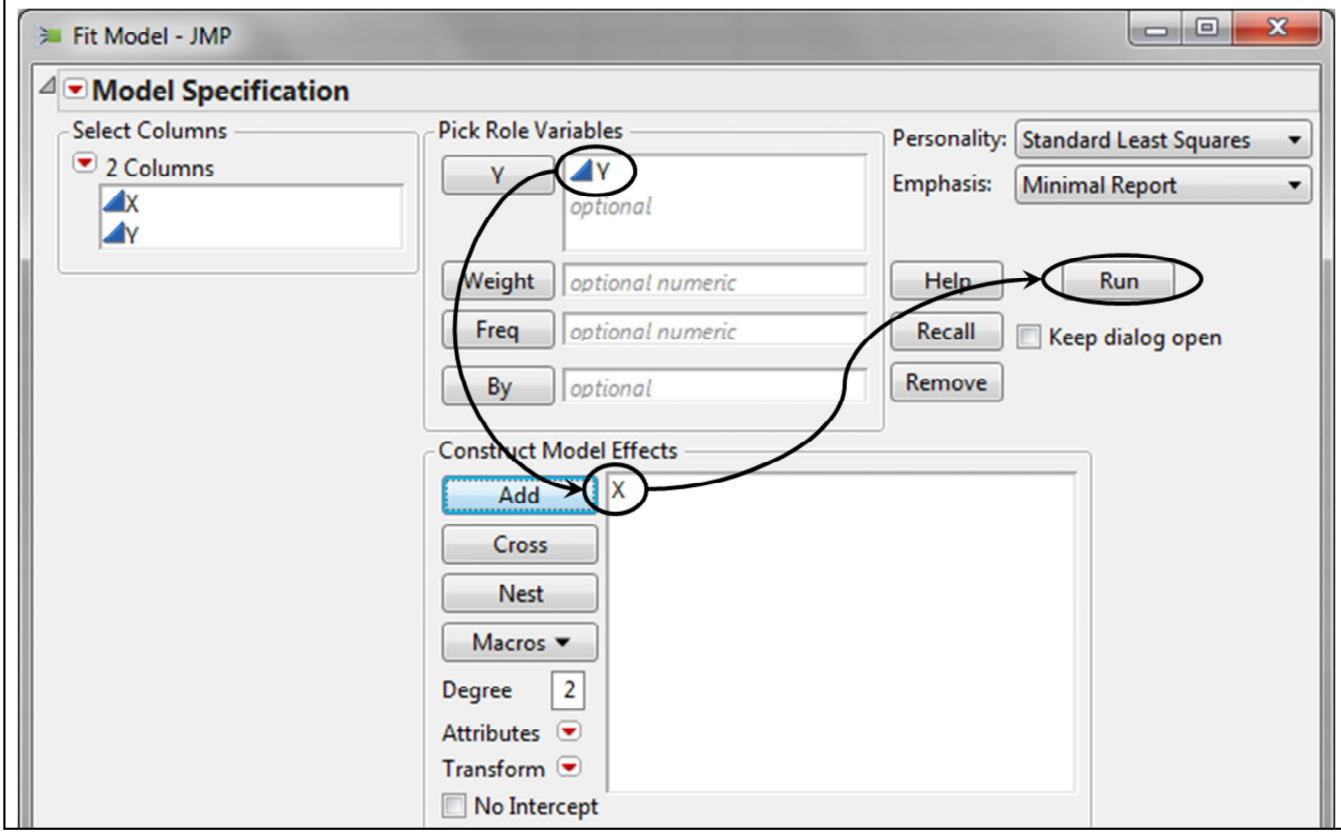
Open: Data sets \ simple regression - generic



Notes

Simple Regression in JMP (cont'd)

Analyze → *Fit Model* → Set up as shown → Run

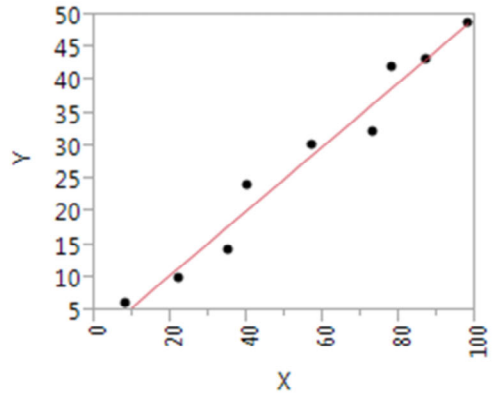


Notes

Analysis details

Response Y

Regression Plot



Summary of Fit

RSquare	0.966581
RSquare Adj	0.961807
Root Mean Square Error	3.006984
Mean of Response	27.90333
Observations (or Sum Wgts)	9

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1830.6557	1830.66	202.4624
Error	7	63.2937	9.04	Prob > F
C. Total	8	1893.9494		<.0001*

- The **Root Mean Square Error (RMSE)** is the standard deviation of Y caused by factors other than X
- It can be thought of as the **standard deviation** about the fitted line (or model)
- Also known as the "error" or "residual" standard deviation
- Smaller is better

- **P-value** indicates whether the regression is significant
- This low p-value shows that it is significant

Notes

Analysis details (cont'd)

Summary of Fit

RSquare	0.966581
RSquare Adj	0.961807
Root Mean Square Error	3.006984
Mean of Response	27.90333
Observations (or Sum Wgts)	9

R^2 "Coefficient of Determination"

Analysis of Variance

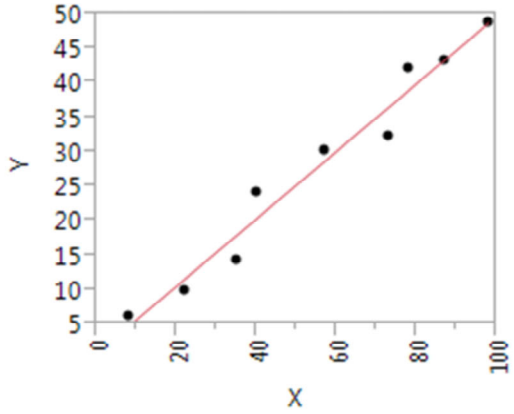
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1830.6557	1830.66	202.4624
Error	7	63.2937	9.04	Prob > F
C. Total	8	1893.9494		<.0001*

- Proportion of the variation in Y that is "explained by" variation in X.
- Varies from 0 to 1.
- Larger is better
- Unitless

Notes

Analysis details (cont'd)

Regression Plot



Summary of Fit

RSquare	0.966581
RSquare Adj	0.961807
Root Mean Square Error	3.006984
Mean of Response	27.90333
Observations (or Sum Wgts)	9

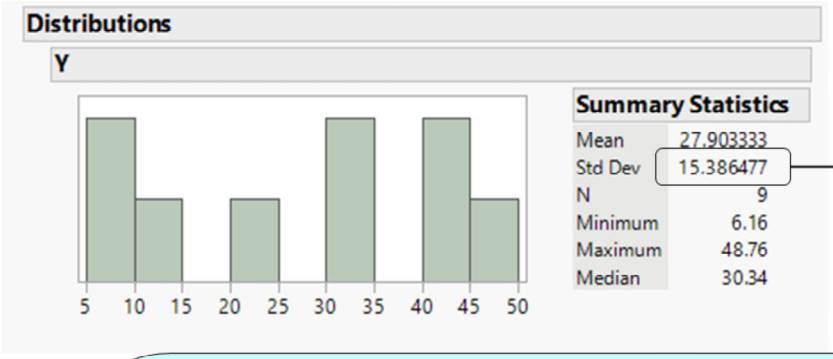
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1830.6557	1830.66	202.4624
Error	7	63.2937	9.04	Prob > F
C. Total	8	1893.9494		<.0001*

- **Adjusted R^2** also gives us the proportion of Y variation *explained by* the model (a line in simple regression)
- Varies from 0 to 1
- Larger is better
- **Always use the Adjusted R^2 value, not R^2**
- Adjusted R^2 takes the number of model terms into account and penalizes for including insignificant terms
- In this example, the simple regression model explains much of the variation in Y.

Notes

How R^2 and R^2_{Adj} are calculated



Standard Deviation (STDEV) of the data set

$$R^2 = 1 - \frac{SS_{Error}}{SS_{Total}}$$

$$R^2_{Adj} = 1 - \frac{SS_{Error}/(n-p)}{SS_{Total}/(n-1)} = 1 - \left(\frac{RMSE}{STDEV}\right)^2$$

p = number of terms in the model (including the intercept)

n = sample size (number of measurements in the data set)

SS_{Total} is the sum of squares of the data (measurements in the data set)

SS_{Error} is the sum of squares of the Errors or residuals

We saw the sum of squares calculations earlier, in the ANOVA

Notes

Why use Adjusted R^2 ?

There is a potential problem with R^2:

- R^2 always increases when terms are added to a model, even when the terms are not significant
- **This is particularly a problem in multiple regression**, as it can lead to "overfitting," giving false confidence in using the model, especially for prediction.
- Adjusted R^2 corrects for this by considering the number of terms in the model
- Adjusted R^2 can actually decrease if non-significant terms are added to a model

Adjusted R^2 is the recommended statistic for determining the proportion of variation in Y explained by the model

Notes

P-values for the ANOVA and individual model parameters

Red triangle next to *Response Y* → *Regression Reports* → *Parameter Estimates*

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1830.6557	1830.66	202.4624
Error	7	63.2937	9.04	Prob > F
C. Total	8	1893.9494		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.8386661	2.150023	0.39	0.7081
X	0.4891205	0.034375	14.23	<.0001*

- In regression of Y on a single X, the Analysis of Variance P-value is the same as the P-value for the slope of the line.
- The P-value for the slope of the line indicates the evidence of a correlation between Y and X.
- Significance of individual model terms are determined by testing whether their regression coefficient is equal to 0, using the t statistic. Hypotheses are:

$$H_0: b_i = 0$$
$$H_1: b_i \neq 0$$

- This is a test of the contribution of the model term, given the other terms in the model.

Notes

P-values for individual model parameters

Estimates and P-values for the slope and intercept

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.8386661	2.150023	0.39	0.7081
X	0.4891205	0.034375	14.23	<.0001*

Model: $\boldsymbol{Y = 0.84 + 0.50X + error}$

- In this example, the P-value for the slope of the line indicates very strong evidence of a correlation between Y and X.
- The P-value for the Intercept indicates that it is not significant.
  - **Best practice is to leave the Intercept in the model, whether or not the P-value indicates that it is significant**
  - Regression equations are developed, and are only valid, over the region of the regressor variables (x's) contained in the data set
  - Forcing the model to pass through (0, 0) by removing the intercept, can create problems in the region being modeled

Notes

Using Adjusted R^2 and p-values

Both the Adjusted R^2 and the p-values must be considered, in order to understand what has been learned in the analysis.

When the resulting model has:

- **High Adjusted R^2 and significant model term p-values,** this is ideal. Factors driving the response have been identified and the variation is largely explained. A decent model has been created.
- **Low Adjusted R^2 and significant model term p-values,** more work must be done. Some significant factors influencing the response have been identified, but the low Adjusted R^2 indicates that other important factors exist. These need to be found, for the model to be useful.
- **High R^2 and insignificant model terms**, this is usually due to the data violating the assumptions of the regression analysis. There is more information on this scenario in upcoming slides.
- **Low Adjusted R^2 and insignificant model terms,** no relationship between X and Y variables have been found. Usually this means that new ideas about which factors influence Y must be developed, although it can occasionally be due to missing higher order terms.

Notes

This page intentionally left blank

2 Checking Model Adequacy

In least squares fit regression (continuous Y), the analysis methods used to calculate regressor coefficients and their p-values, depend on certain assumptions being met.

Assumptions:

- Errors (residuals) are normally and independently distributed with mean zero and constant variance (σ^2)
- Observations are adequately described by the model

Whether performing regression from "file cabinet" data or analyzing the results of a designed experiment, these assumptions must be validated.

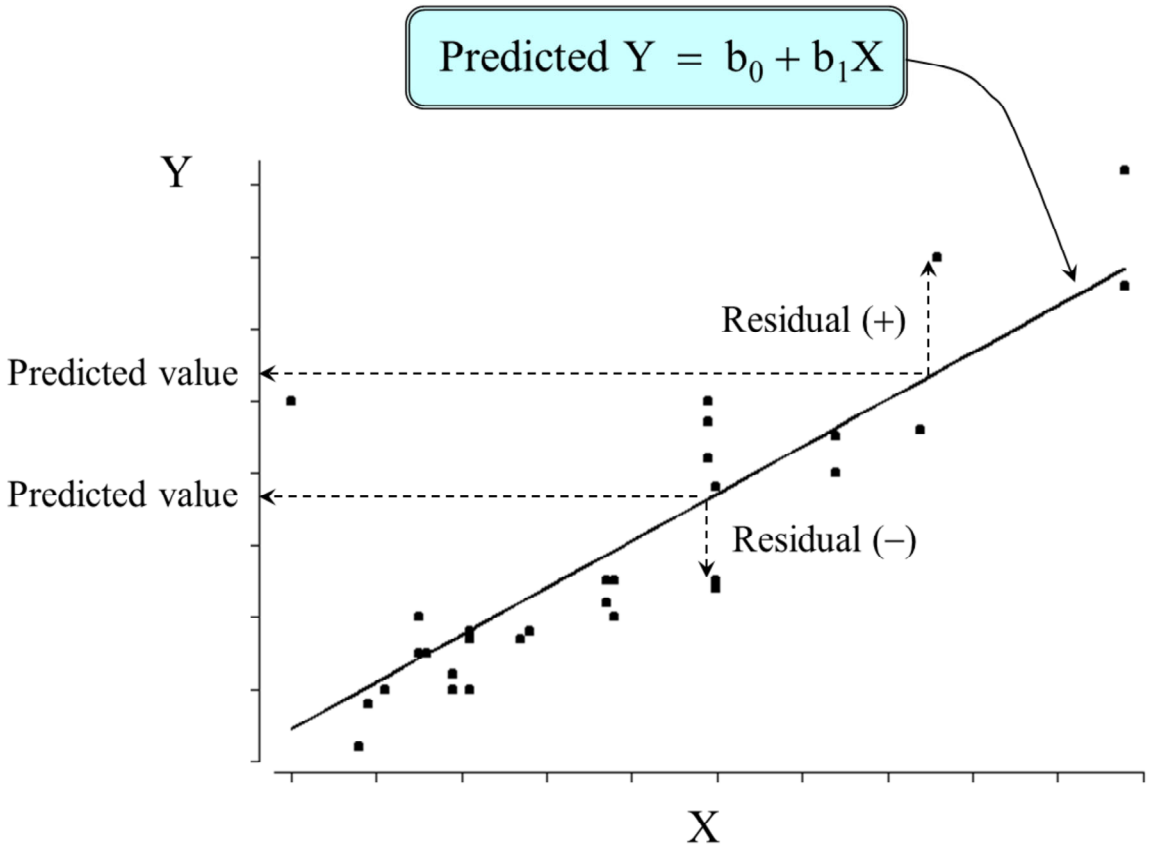
Notes

To validate that these assumptions have been met, the *residuals* are examined:

1. **Normal Probability Plot of Residuals**
   - Validate that the residuals are normally distributed
   - In JMP, this is the *Residual Normal Quantile Plot*
2. **Residuals vs. Predicted (or Fitted) Values**
   - Validate constant variance and mean 0
   - In JMP, this is the *Residual by Predicted Plot*
3. **Residuals vs. Run Order**
   - Verify independence of errors
   - There should be no patterns over the timeframe of the data
   - In JMP, the best graph to use is *Studentized Residuals*
   - The JMP data table must be in run order for *Studentized Residuals* to graph the residuals in run order

Notes

Residuals Review



Notes

A fitted model, the equation generated during regression, gives the predicted mean value of the response variable as a function of the predictor variables. These predicted mean values are also called *predicted values*, or just *predicted* for short. The *residual value* is the data (observation) value minus the predicted value. Residual values are called *residuals* for short.

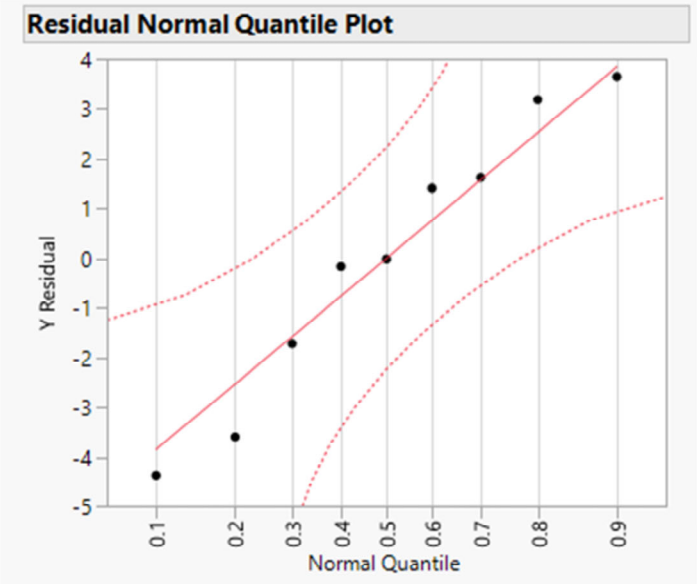
These terms are easiest to visualize in the simple linear model shown above. A predicted value is the fitted line evaluated at some X value. A residual is the difference between a measured (observed) Y value and the predicted value at the corresponding X.

Residuals contain information about the magnitude and direction of variability in the data relative to the fitted model.

- An unusually large residual might signal a measurement error, data entry error or some other type of outlier.
- A systematic trend or pattern in the residuals might signal an inadequacy in the fitted model.

Notes

Residual Analysis



In viewing the Residual Normal Quantile Plot for the *simple regression-generic*, we can see whether the residuals are normally distributed.

Notes

Notes

If residuals are normally distributed, the plot will be approximately a straight line.

Emphasis should be on the central values of the plot, rather than the ends

It is common for plots to bend upward at the high end and downward at the low end.

Small sample sizes, such as from experiments, often appear more non-normal

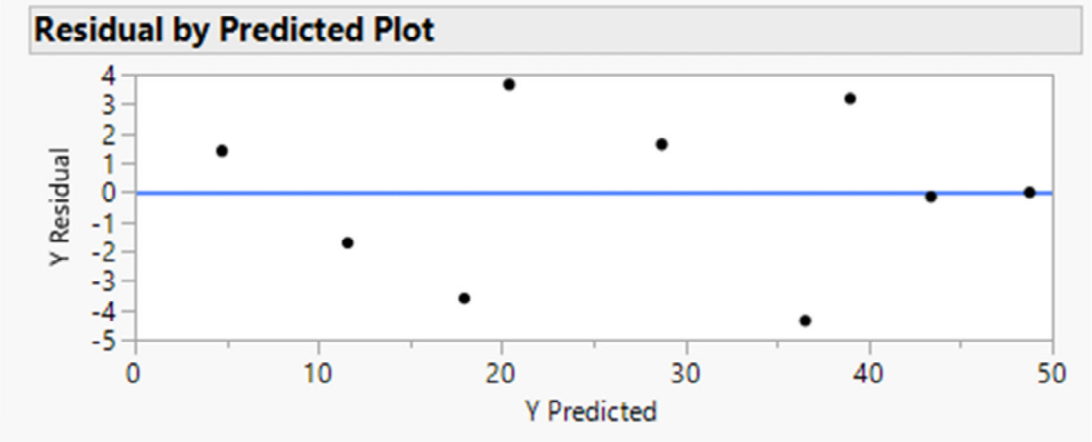
Use the "Fat Pencil" Rule: If a "fat pencil" placed over the central points would cover them on the plot, then the residuals are approximately normal (good enough). Hyperbolic bands displayed in JMP plots give these bounds.

A curve throughout the plot is a strong indication of non-normality. In this case, a transformation would be needed.

The plot above shows an error (residuals) distribution that is approximately normal, so it is not concerning.

Notes

Residual Analysis (cont'd)



In viewing the Residual by Predicted Plot for the *simple regression-generic*, we can see whether the residuals have constant variance and mean 0.

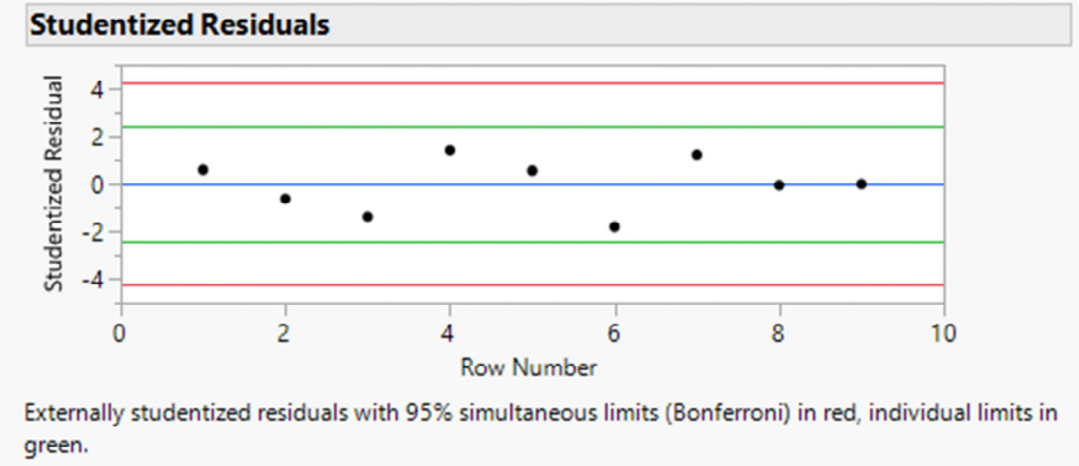
Notes

Here the residuals are plotted against the predicted values. This is a good all-around diagnostic plot.

"Healthy" residuals look like random scatter around 0. There should be no obvious patterns. The amount of "scatter" or variance (how high and low the plot goes) should be consistent across the graph. This verifies the assumption of constant variance. If the variance is increasing or decreasing across the graph, a transformation is needed.

Notes

Residual Analysis (cont'd)



In viewing the Studentized Residuals for the *simple regression-generic*, the best form for checking residuals by run order, we can see whether there are any patterns over the timeframe of the data.

Note that the data table must be in run order for this plot.

Notes

Again, on this graph, healthy residuals look like a random scatter around 0.

Runs (points in a row) of positive-negative-positive-negative residuals indicate correlation between runs. This implies that the assumption of independence has been violated. **In designed experiments, randomization protects against this! Do it every time!**

This plot can also show a change in variance over the time span of the experiment. This could be due to increased skill as the experiment progresses, a process drift, operator fatigue, tool wear, etc. This type of problem would show as an increase or decrease in spread or "scatter" of the residuals across the graph. Increasing or decreasing variance indicates the need for a transformation.

Notes

3 Using the Model: RMSE and Prediction Profiler

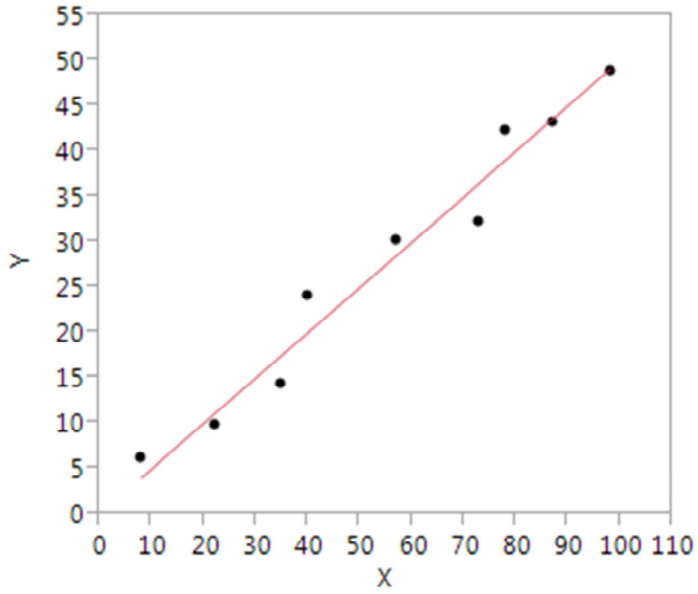
In this section, we'll see how we can:

- Use the Root Mean Square Error (RMSE) in predicting our future process variation,
- Use JMP's Prediction Profiler to help us optimize our process, and
- Estimate our future % defective, using the t distribution calculator.

Notes

Using the Model (cont'd)

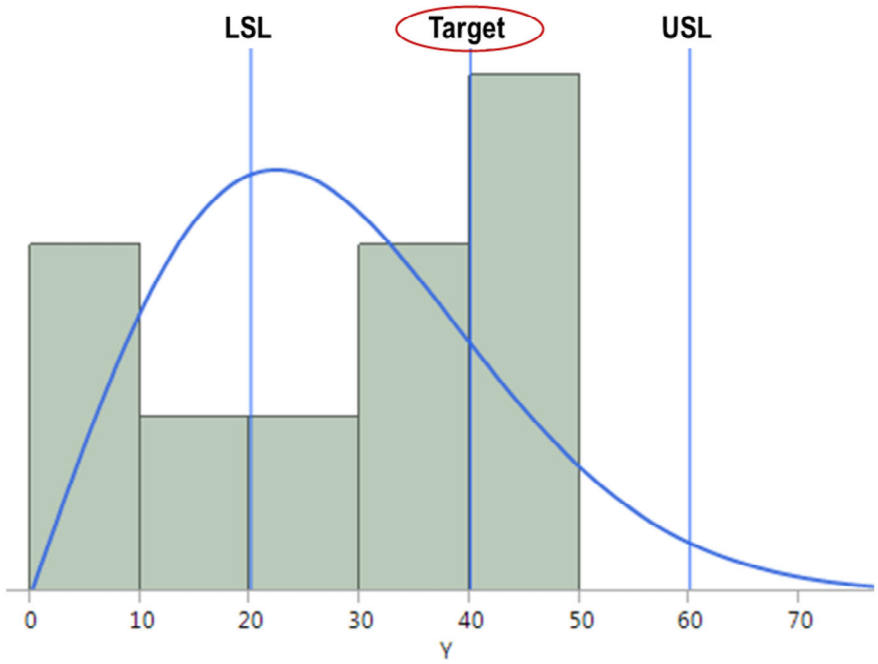
When Y is correlated with a controllable X variable,



how can we use the regression to improve the Y capability?

Notes

Using the Root Mean Square Error (RMSE)



Suppose we are not happy with our current process capability

Mean = 27.9, Std dev = 15.4

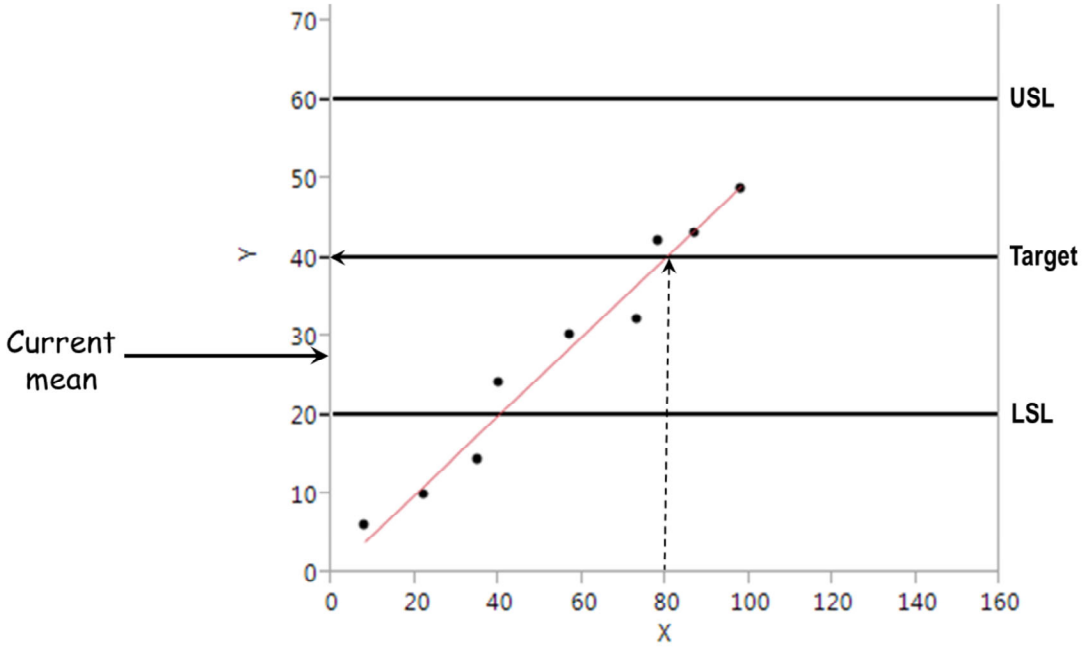
Defective in the data: 33.3%

Predicted from distribution curve: 35.8%

Notes

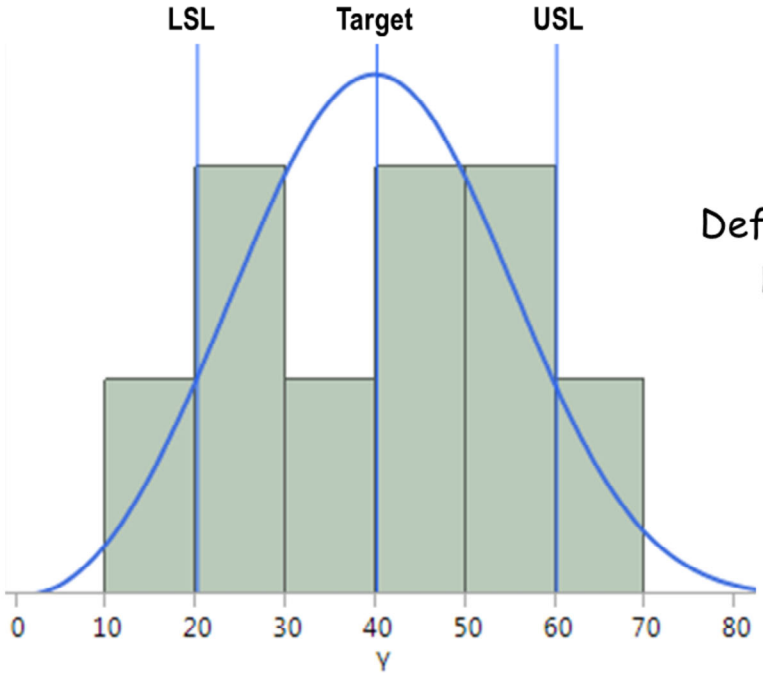
RMSE (cont'd)

If we control X at 80, the mean will change from 27.9 to 40



Notes

RMSE (cont'd)



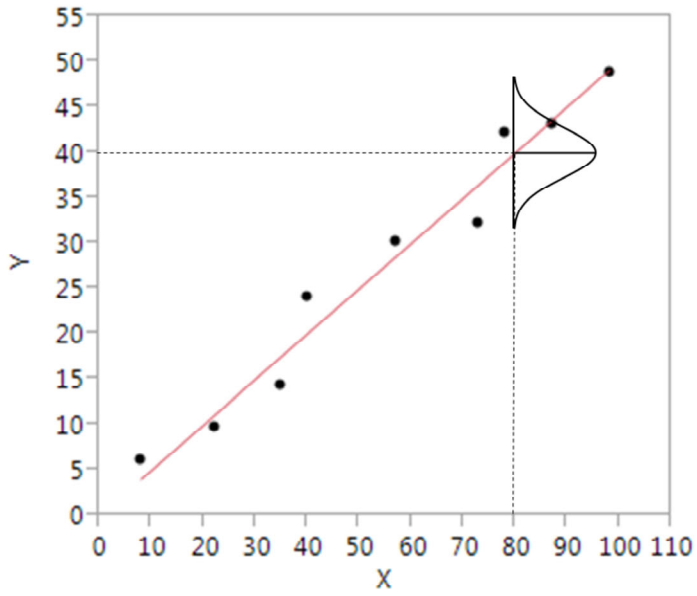
Mean = 40.0
Std dev = 15.4
Defective in the data: 22.2%
Distribution curve: 15.9%

- Moving mean Y to the center of the spec range does reduce % defective
- Is the mean the only thing that changes when we control X at 80?

Notes

RMSE (cont'd)

By definition, RMSE is the standard deviation of Y that would result from eliminating the variation in X

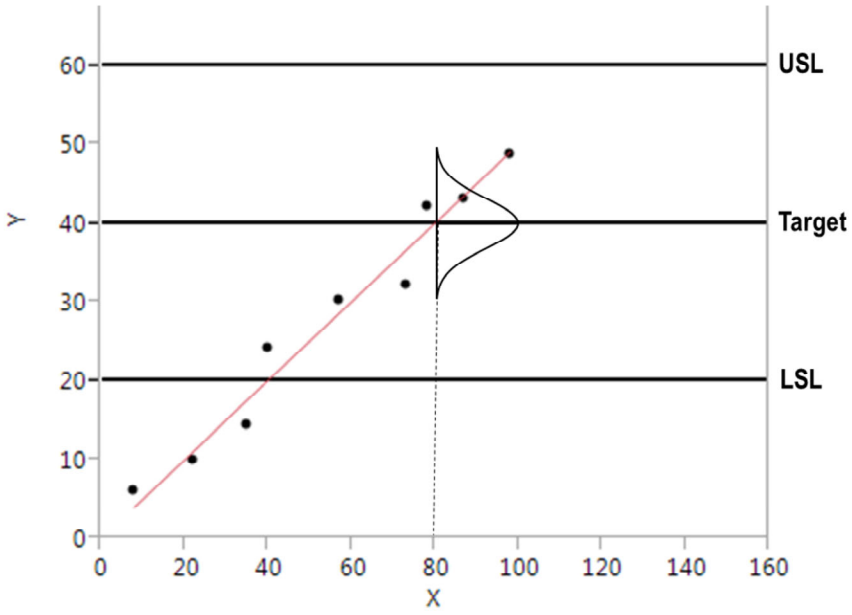


σ = RMSE
= 2.84

Notes

RMSE (cont'd)

When we control X at 80, we don't just move the mean from 27.9 to 40 — we also reduce the standard deviation from 15.4 to 2.84 !

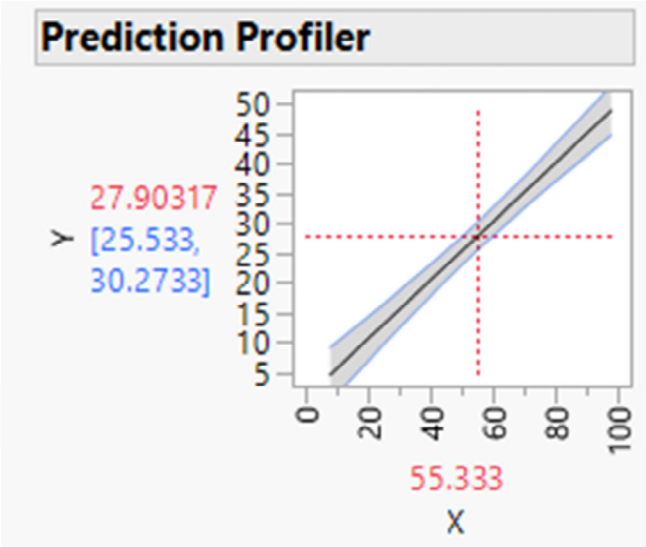


Notes

This page intentionally left blank

4. Introduction to the Prediction Profiler

JMP's Prediction Profiler helps us use our regression model to make predictions and optimize our process.



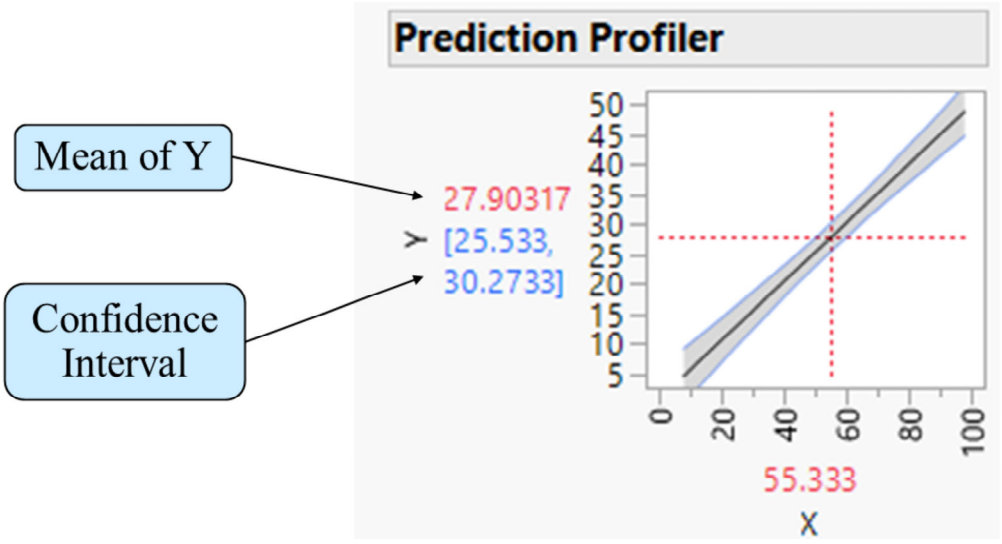
Follow these steps to access the prediction profiler:

- Analyze > Fit Model > Y = Y, Model Effects = X > Run > Red Triangle > Factor Profiling > Profiler

Notes

Introduction to the Prediction Profiler (cont'd)

JMP's Prediction Profiler helps us use our regression model to make predictions and optimize our process.



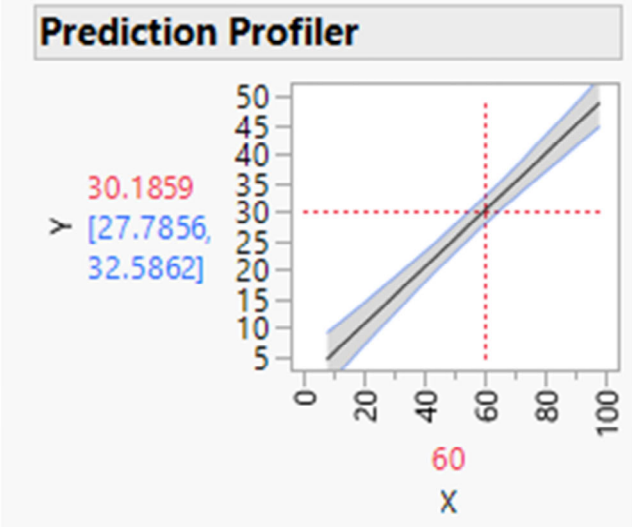
- Calculates predicted ***mean Y*** as a function of X
- Calculates **confidence intervals** for predicted **means**

Notes

Simple example of prediction of Mean Y

Continuing with the *simple regression-generic* data:

- Suppose we are interested in the predicted mean Y for X = 60
- Click on the 55.333, change it to 60

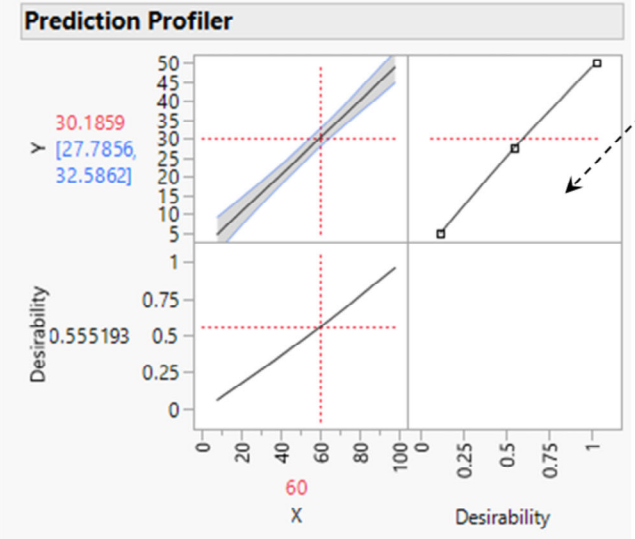


- Predicted mean Y (based on the data) is 30.19
- With 95% confidence, the population mean lies between 27.79 and 32.59

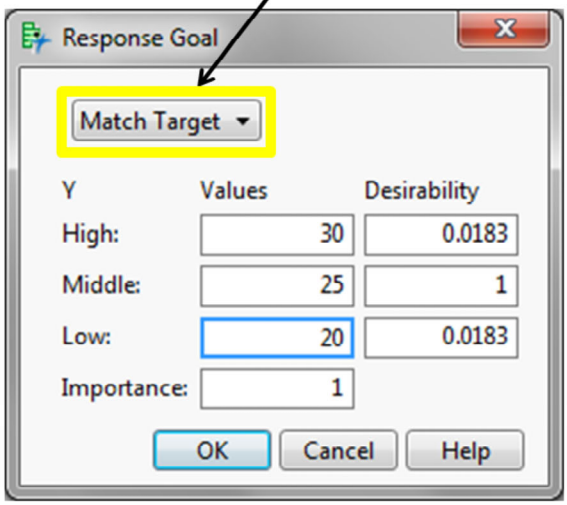
Notes

Simple example of optimization

- Suppose we want to find the X value that predicts a mean Y value of 25
- Red triangle next to *Prediction Profiler* → *Optimization and Desirability* → *Desirability Functions*



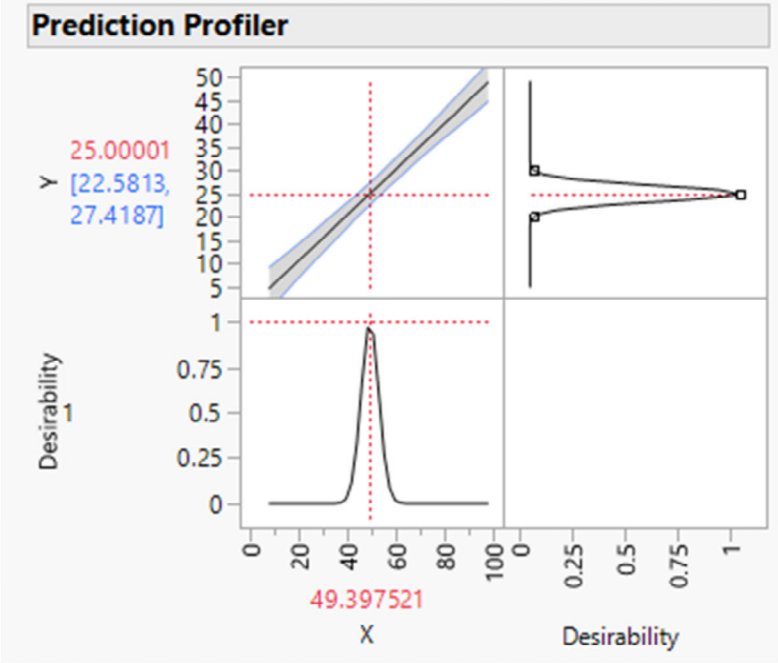
- Double click in here (don't touch the line plot)
- Modify the **Response Goal** dialog as shown below
- Click OK



Notes

Optimization (cont'd)

Red triangle
next to
Prediction Profiler
↓
Optimization and Desirability
↓
Maximize Desirability



- Predicted mean Y of 25 is achieved when X = 49.4
- With 95% confidence, this population mean lies between 22.6 and 27.4

Notes

Confidence Intervals and Prediction Intervals

- The ***95% Confidence Interval on the Mean Response*** gives the range which will contain the "true" mean, μ, 95% of the time
  - For a sample, the confidence interval is calculated:

$$\bar{Y} - t_{.025,n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{Y} + t_{.025,n-1}$$

  - For a regression, calculation of the confidence interval is similarly structured, but considerably more complicated, involving matrix math.

- A ***95% Prediction Interval*** gives the range which will contain future individual response observations 95% of the time.
  - The prediction interval is wider than the confidence interval, because it is to contain individual measurements, not averages.
  - Calculation of this interval is complicated, involving matrix math.

Notes

Exercise 4.1

a) Continuing with *simple regression-generic,* find the X value that predicts a mean Y value of 35. Give the confidence limits for the predicted mean.

b) The overall standard deviation of Y is 15.39. The RMSE from the regression is 2.84. Which of these would be the standard deviation of Y if we controlled X to a constant value?

c) Save your script, close and save the data table.

Notes

Exercise 4.2

Data sets \ production vs capacity.

(a) Fit a regression for *Production qty* as a function of *Capacity utilized* (%) (using *Fit Model*, of course). Is there a correlation? Give the appropriate P-value and strength of evidence.

(b) For this exercise, we will not review the residuals plots. Use your model to find the capacity utilization level that predicts a mean daily production quantity of 3500. Give the confidence limits.

(c) The overall standard deviation of *Production qty* is 733.5 (not shown in Fit Model output—calculated in Distribution Platform). The RMSE from the analysis in (a) is 409.732. Which of these would be the standard deviation if capacity utilization was held constant?

(d) Save your scripts, close and save the data table.

Notes

Estimating Improved % Defective

Once we determine the level at which we want to control our x, we can use the root mean square error (RMSE) and other regression results to estimate the % defective in the improved process.

Remember that by definition, the RMSE is the standard deviation of the improved process, with x's held at desired levels.

The *t distribution calculator* helps us calculate the future % defective.

Notes

1. Enter the quantities in the YELLOW cells.

2. The other values are calculated for you.

LSL	20
USL	60
Mean	40
Standard deviation	3.006984
Degrees of freedom	7

	LSL	USL	Total
Population % out of spec	0.015	0.015	0.029
Population PPM out of spec	145.1	145.1	290.2

PPM defective = 290

These calculations can be sensitive to round-off error. Don't round off the mean and standard deviation when you enter them into the calculator.

Error DF from the Analysis of Variance

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1830.6557	1830.66	202.4624
Error	7	63.2937	9.04	Prob > F
C. Total	8	1893.9494		<.0001*

Notes

Exercise 4.3

Data sets \ production vs capacity.jmp.

In this process data, on 75% of the days production quantity fell below 3000.

Based on the best fit distribution, the Lognormal, the expected % of days that production quantity will fall below 3000 is 71.8%.

a) We found earlier that capacity utilization 52.1% gives a mean daily production quantity of 3500. The RMSE was 409.7, the error degrees of freedom was 34. Assuming 52.1% capacity utilization, use the *t distribution calculator* to find the predicted % of days on which production quantity will be less than 3000.

b) Save your scripts, close and save the data table.

Notes

Exercise 4.4

Open *Data sets \ outgassing process*. *Current* (the Y variable) is the current required to heat a filament to a target temperature. *Resist* (the X variable) is the electrical resistance of the filament. *Machine* is the processing unit. This example shows how to reduce % defective by separate optimization of each machine.

a) For this process, the % of *Current* data values that fall outside the interval (1.9, 2.1) is 8.87%.

b) Fit a regression for *Current* as a function of *Resist*, using *Machine* as the *By variable*. For each machine, give the RMSE, the error degrees of freedom, and the resistance that predicts a mean current of 2.

Machine	RMSE	DF	Resistance	% Outside
A				
B				
C				

c) Assuming we use the indicated resistance values, use the *t distribution calculator* to find for each machine the % of *Current* values predicted to fall outside the interval (1.9, 2.1).

d) Save your scripts, close and save the data table.

Notes

5 Multiple Regression

- Multiple regression model
- Examples
- Fitting regression models
- Interactive effects
- Predicted values and uncertainty
- Modeling and optimization

Notes

Multiple regression model

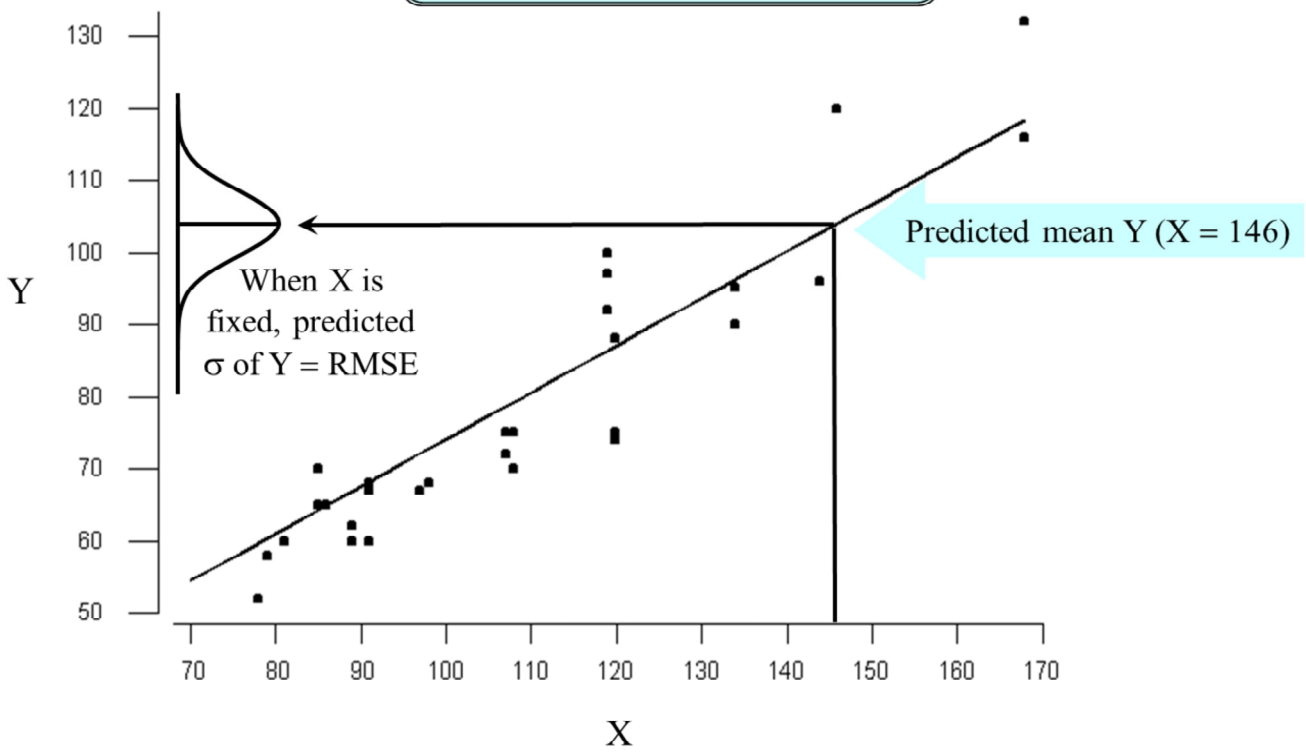
$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k + \text{“error”}$$

Y	$X_1, X_2, \ldots, X_k$	b_0	$b_1, b_2, \ldots, b_k$	“Error”
Dependent variable	Independent variables	Intercept	Regression coefficients	Residuals Mean = 0
Response variable	Explanatory variables	Parameter	Parameters	Standard deviation = σ (RMSE)
Output	Inputs Predictors Regressors Factors (in DOE)			Distribution = Assumed to be Normal

Notes

Model and error components, one X

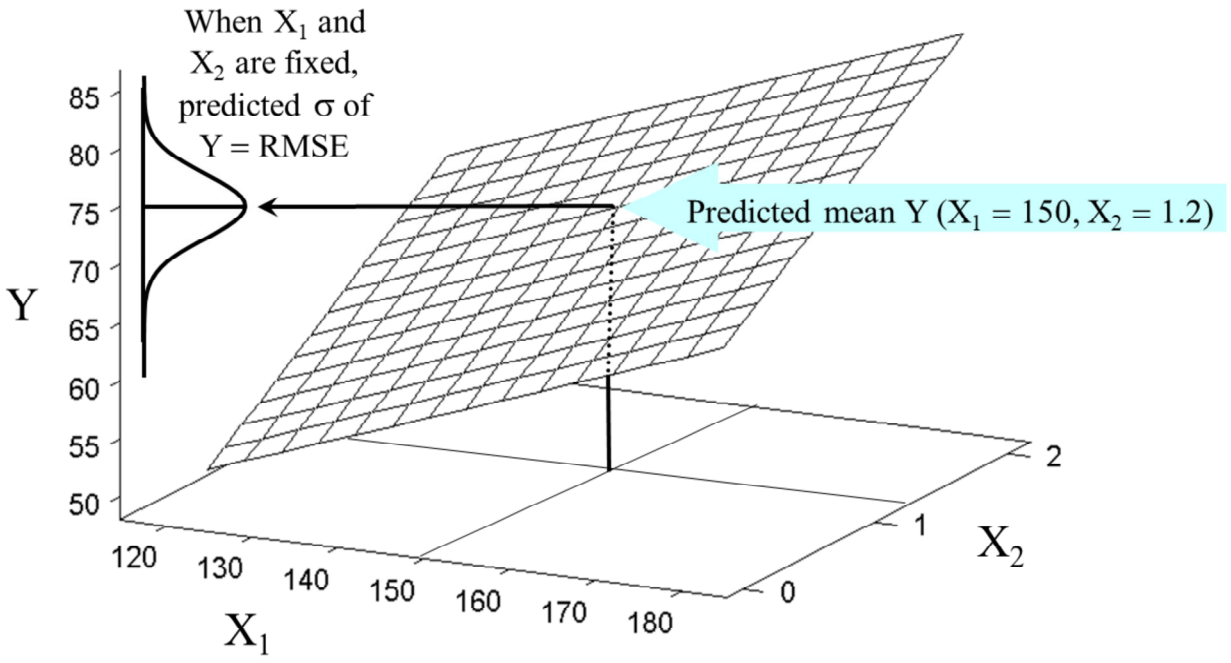
$$Y = b_0 + b_1X + \text{“error”}$$



Notes

Model and error components, two Xs

$$Y = b_0 + b_1X_1 + b_2X_2 + \text{“error”}$$



Notes

Multiple regression examples

Y	X_1	X_2	X_3	X_4	X_5
Life of cutting tool	RPM	Tool type	Material	Feed rate	
MPG	Displace-ment	Horsepower	Weight		
Salary	Education	Experience	Performance	Seniority	Gender
Vending machine service time	Amount of product stocked	Distance from truck to machine			

Fill in examples of interest to you

Notes

Regression model equations

Y	X_1	X_2	X_3	X_4	X_5
MPG	Displacement (D)	Horsepower (H)	Weight (W)		

$$MPG = b_0 + b_1D + b_2H + b_3W + error$$

Y	X_1	X_2	X_3	X_4	X_5
Bond strength	Temperature (T)	Dwell time (D)	$T \times D$	T^2	D^2

$$Bond = b_0 + b_1T + b_2D + b_3TD + b_4T^2 + b_5D^2 + error$$

Response surface model (RSM) with two continuous Xs.

TD is the interaction term for T and D, T^2 and D^2 show curvature.

Notes

Linearizing nonlinear models

Nonlinear model	Equivalent linear model
$Y = b_0 (X_1)^{b_1} (X_2)^{b_2}$	$\log(Y) = \log(b_0) + b_1 \log(X_1) + b_2 \log(X_2)$
$Y = b_0 (b_1)^{X_1} (b_2)^{X_2}$	$\log(Y) = \log(b_0) + \log(b_1) X_1 + \log(b_2) X_2$

- In many cases, log(Y) transformations can successfully linearize nonlinear regression models
- This greatly extends the application of standard multiple regression models

Notes

Data sets \ teenage growth

Y	X_1	X_2
Height	Age	Gender
Weight	Age	Gender

Teenage growth - JMP

File Edit Tables Rows Cols DOE Analyze Graph Tools View Window

Teenage growth
Source

Columns (5/0)
Name
Age
Gender
Height
Weight

Rows
All rows 40
Selected 0
Excluded 0
Hidden 0
Labelled 0

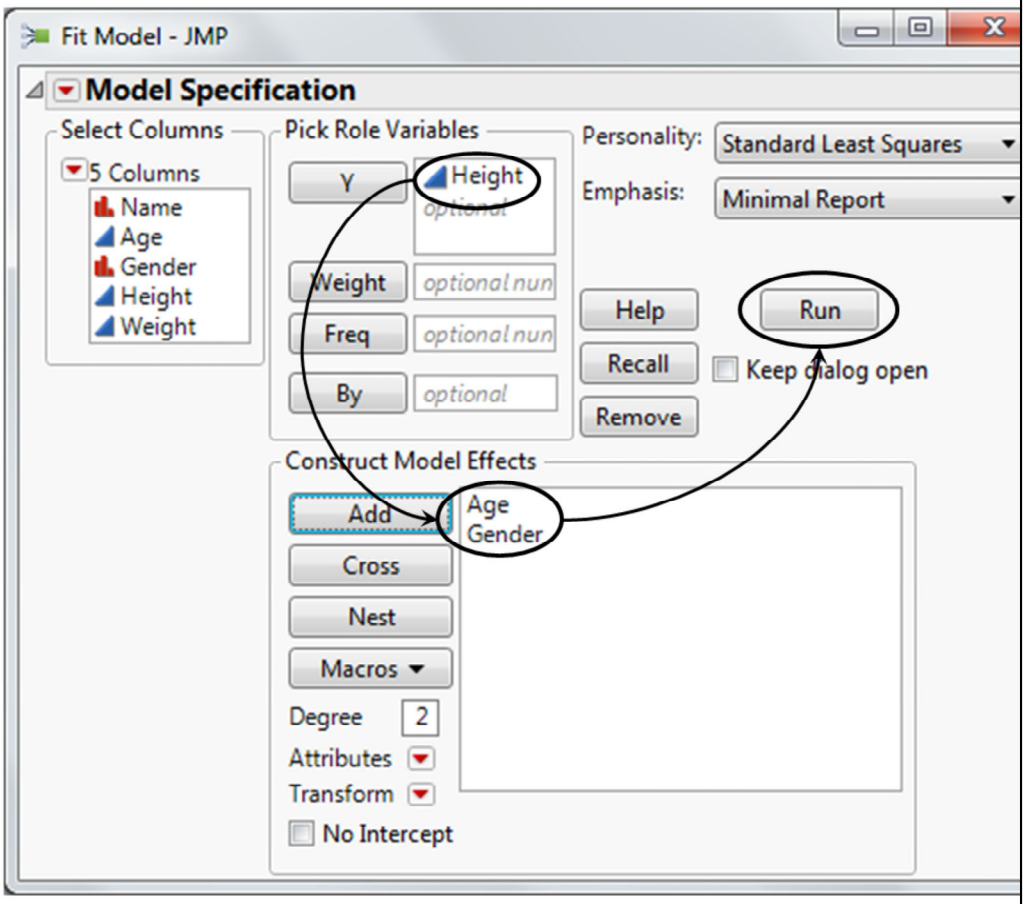
	Name	Age	Gender	Height	Weight
1	ALICE	13	F	61	107
2	AMY	15	F	64	112
3	BARBARA	13	F	60	112
4	CAROL	14	F	63	84
5	ELIZABETH	14	F	62	91
6	JACLYN	12	F	66	145
7	JANE	12	F	55	74
8	JUDY	14	F	61	81
9	KATIE	12	F	59	95
10	LESLIE	14	F	65	142
11	LILLIE	12	F	52	64
12	LINDA	17	F	62	116
13	LOUISE	12	F	61	123
14	MARION	16	F	60	115
15	MARTHA	16	F	65	112
16	MARY	15	F	62	92
17	PATTY	14	F	62	85
18	SUSAN	13	F	56	67
19	ALFRED	14	M	64	99
20	CHRIS	14	M	64	99
21	CLAY	15	M	66	105
22	DANNY	15	M	66	106
23	DAVID	13	M	59	79
24	EDWARD	14	M	68	112

Notes

Fitting models (cont'd)

Say we want to model *Height* as a function of *Age* and *Gender*

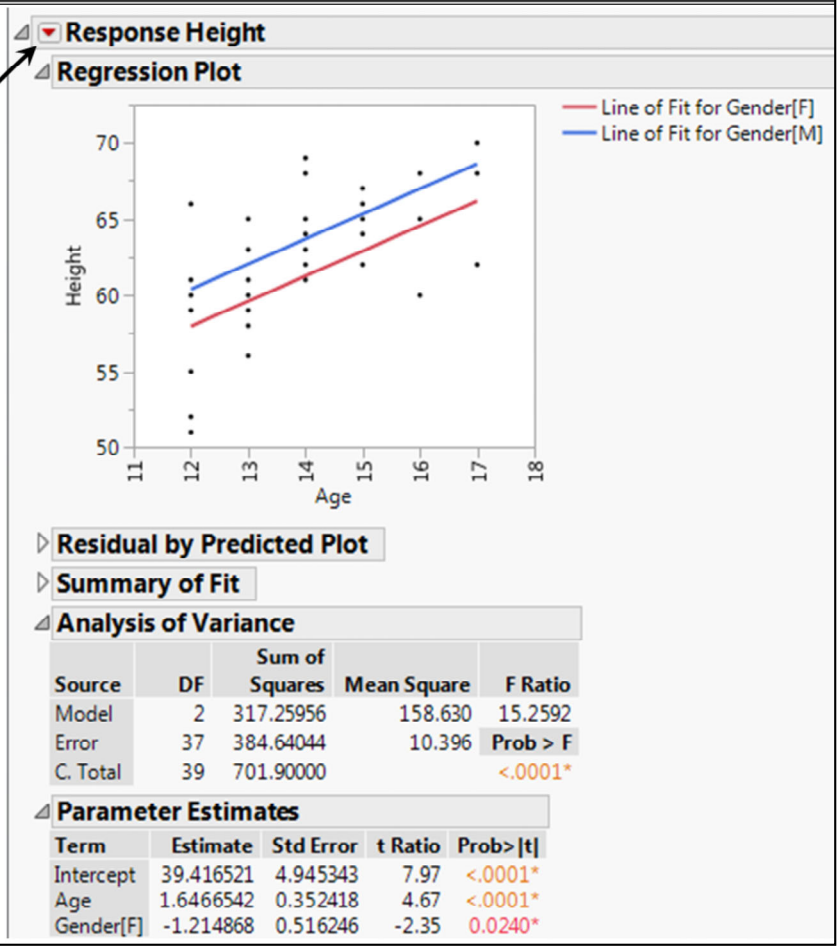
Analyze
↓
Fit Model



Notes

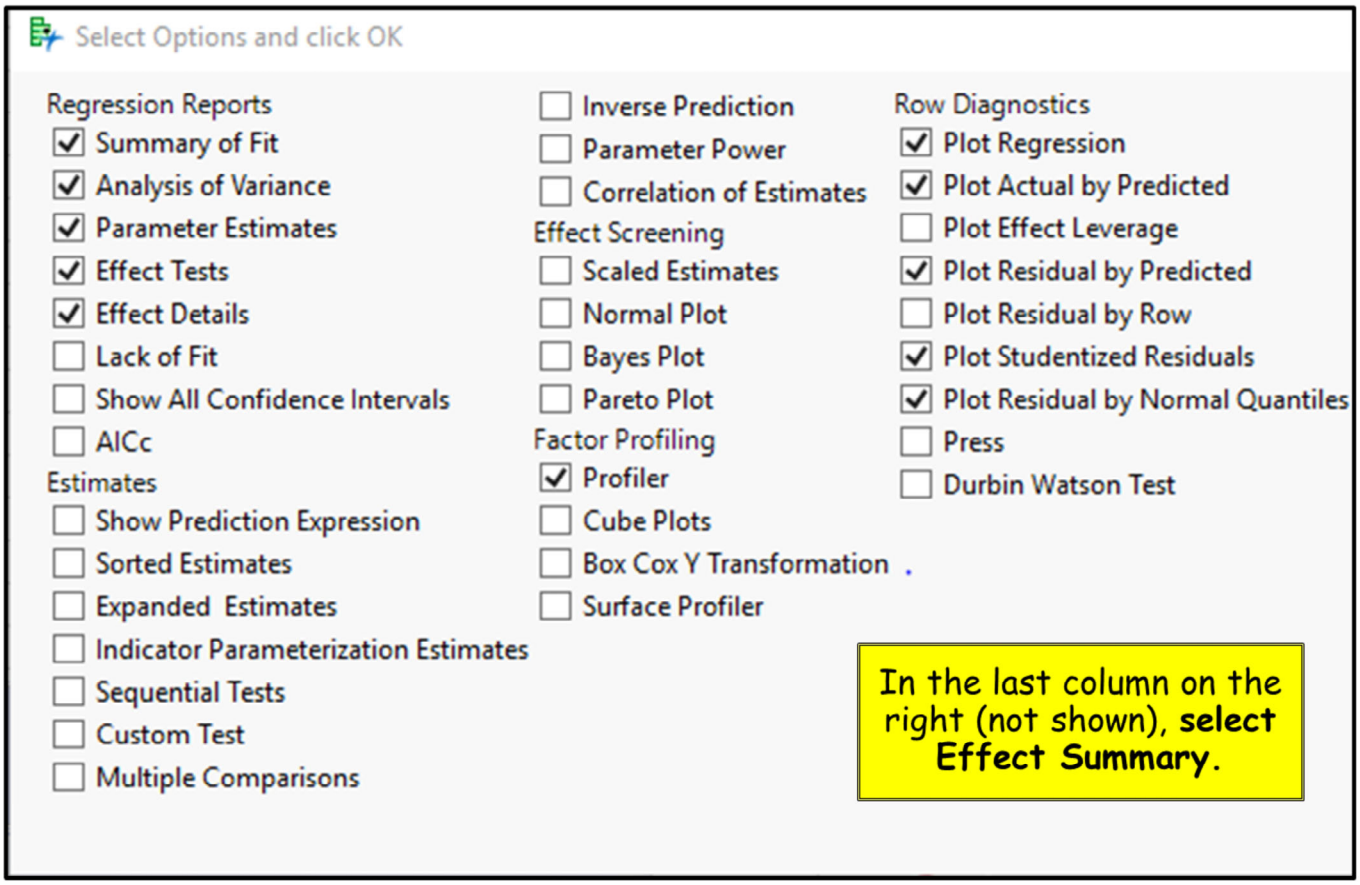
How to change options (for *Fit Model*) during analysis

- Alt-click on *Response Height* red triangle (This technique works for may JMP platforms)
- Set up as shown on next slide



Notes

Default options for *Fit Model* (cont'd)



Notes

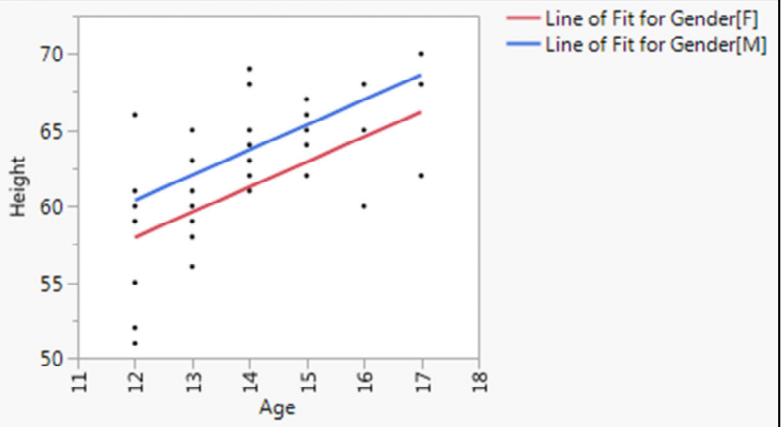
Handling categorical X variables in the model

"Indicator" or "dummy" variables are used to represent categorical variables in regression.

Indicator variable representing the effect of *Gender* in the equation

Response Height

Regression Plot



Residual by Predicted Plot

Summary of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	317.25956	158.630	15.2592
Error	37	384.64044	10.396	Prob > F
C. Total	39	701.90000		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	39.416521	4.945343	7.97	<.0001*
Age	1.6466542	0.352418	4.67	<.0001*
Gender[F]	-1.214868	0.516246	-2.35	0.0240*

Notes

Numeric coding for two-level categorical X

In JMP, two-level categorical factors are coded +1 and -1

$$\text{Gender[F]} = \begin{cases} +1 & \text{if Gender is F} \\ -1 & \text{if Gender is M} \end{cases}$$

$$\text{Height} = b_0 + b_1\text{Age} + b_2\text{Gender[F]}$$

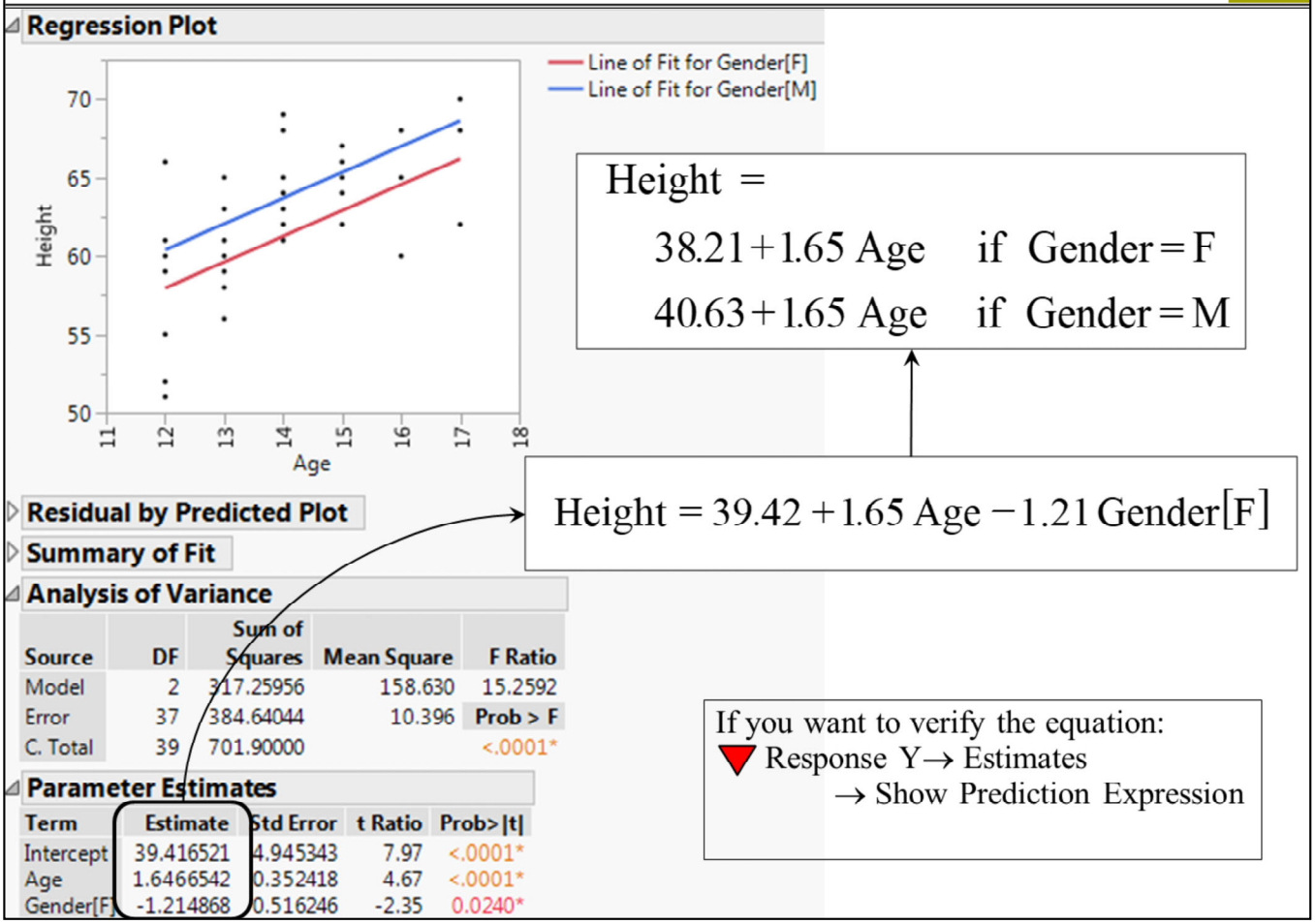
$$= \begin{cases} b_0 + b_2 + b_1\text{Age} & \text{if Gender is F} \\ b_0 - b_2 + b_1\text{Age} & \text{if Gender is M} \end{cases}$$

This results in one equation for Females and one equation for Males, with equal slopes (b_1) and different intercepts ($b_0 + b_2$ and $b_0 - b_2$).

An additional indicator variable is added for each additional level of a categorical variable.

Notes

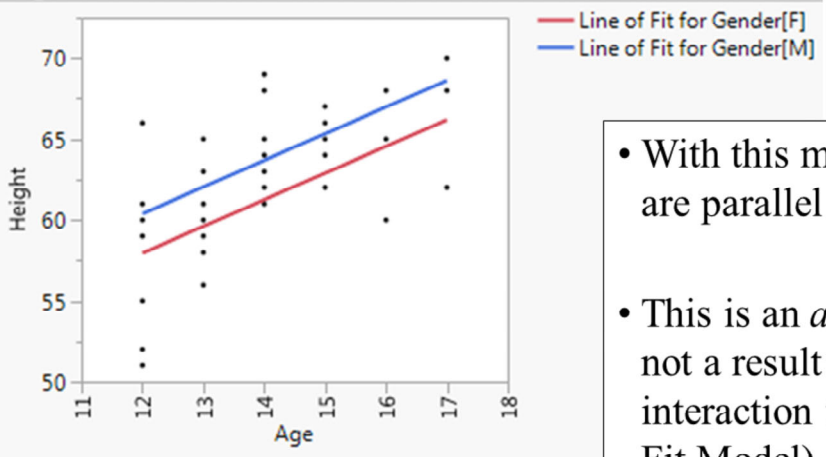
Constructing the model equation



Notes

The need for interaction effects

Regression Plot



Residual by Predicted Plot

Summary of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	317.25956	158.630	15.2592
Error	37	384.64044	10.396	Prob > F
C. Total	39	701.90000		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	39.416521	4.945343	7.97	<.0001*
Age	1.6466542	0.352418	4.67	<.0001*
Gender[F]	-1.214868	0.516246	-2.35	0.0240*

- With this model, the growth curves are parallel
- This is an *assumption* of the model, not a result of the analysis (no interaction terms were included in Fit Model)
- How do we *test* for parallel curves?

Notes

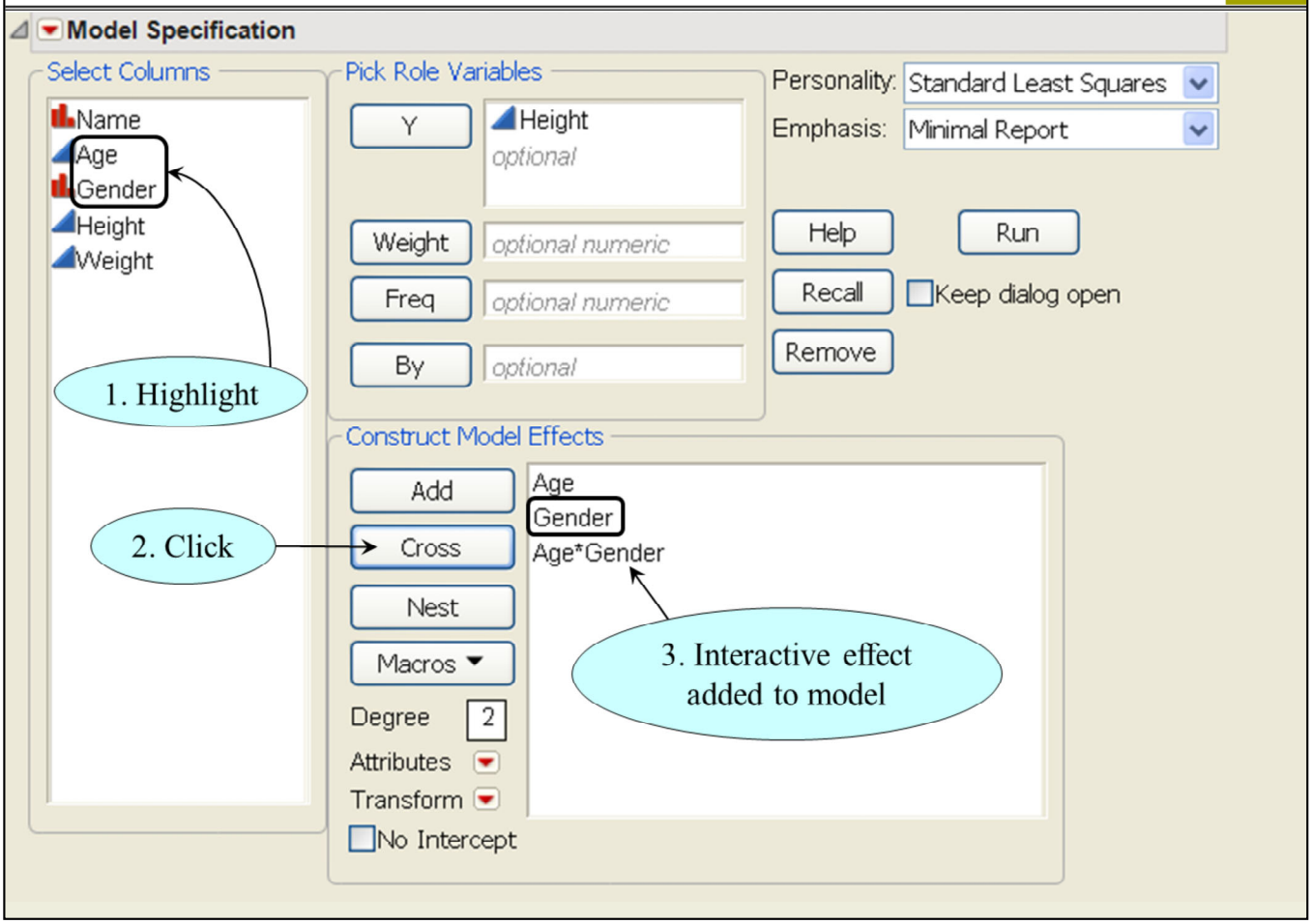
Interaction effects (cont'd)

$$\text{Height} = b_0 + b_1\text{Age} + b_2\text{Gender}[F] + b_3\text{Age} * \text{Gender}[F]$$

This product term allows different slopes for M and F

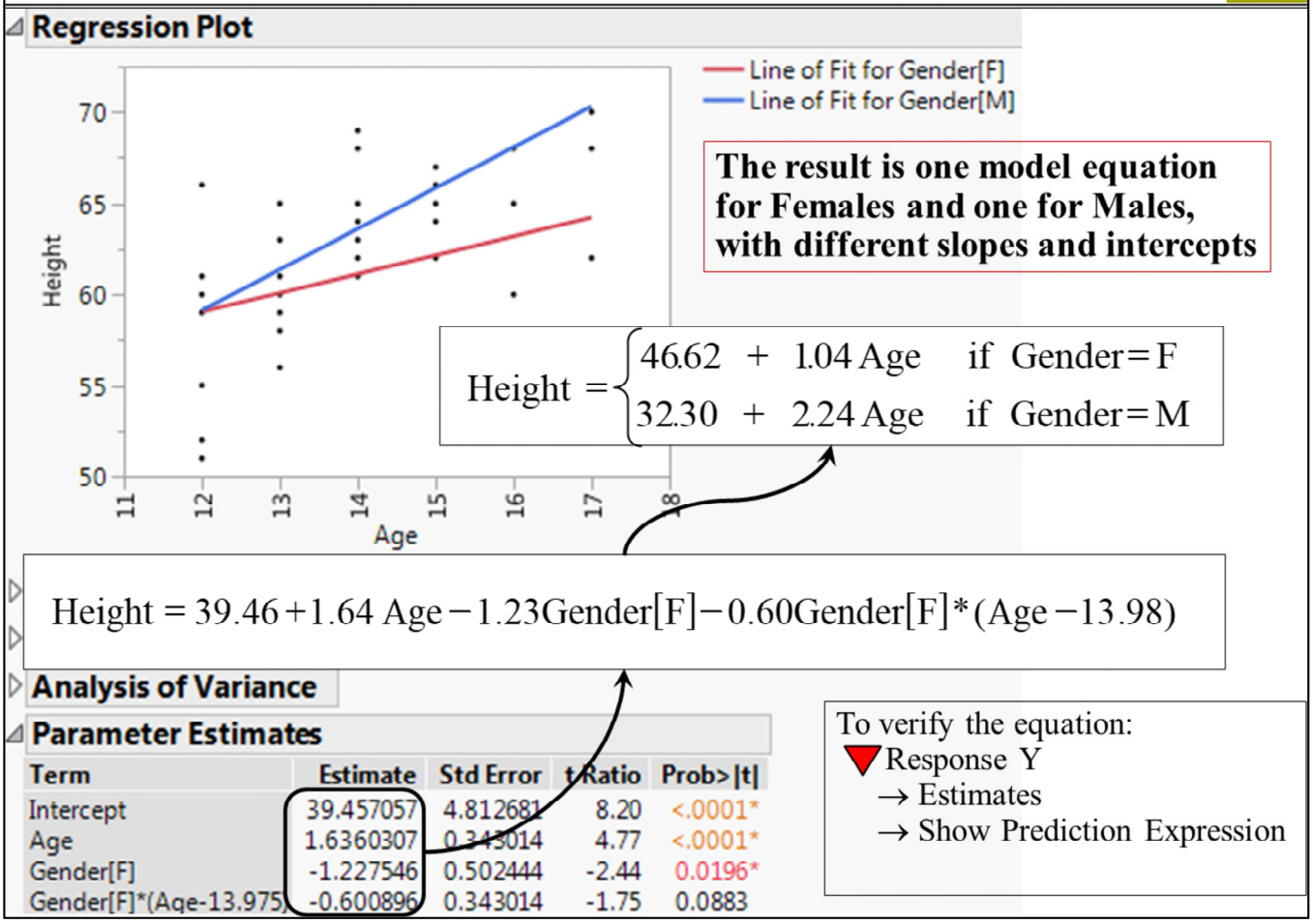
Notes

Adding an interaction effect



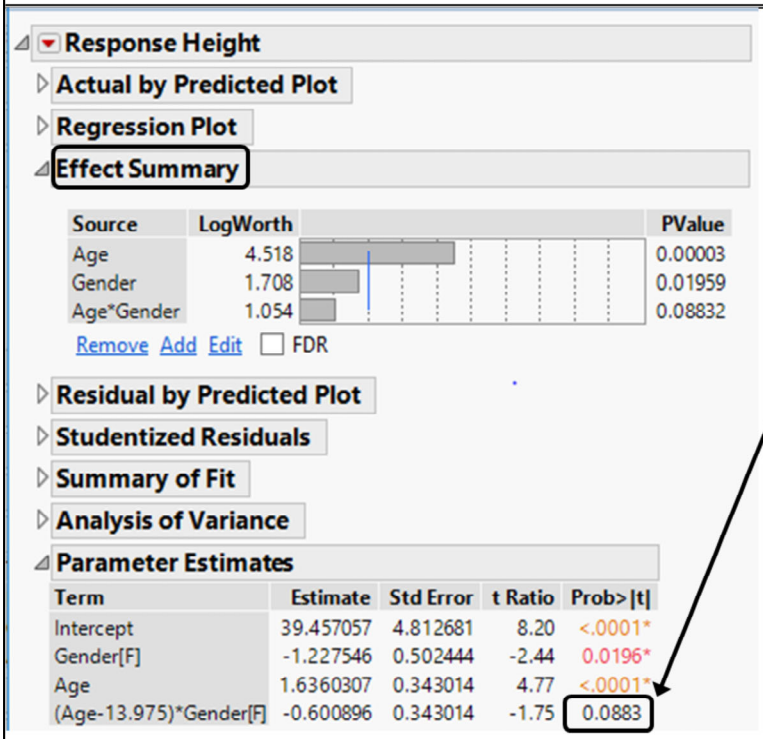
Notes

Non-parallel growth curves



Notes

Testing the interaction effect



Response Height

Actual by Predicted Plot

Regression Plot

Effect Summary

Source	LogWorth	PValue
Age	4.518	0.00003
Gender	1.708	0.01959
Age*Gender	1.054	0.08832

Remove Add Edit FDR

Residual by Predicted Plot

Studentized Residuals

Summary of Fit

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	39.457057	4.812681	8.20	<.0001*
Gender[F]	-1.227546	0.502444	-2.44	0.0196*
Age	1.6360307	0.343014	4.77	<.0001*
(Age-13.975)*Gender[F]	-0.600896	0.343014	-1.75	0.0883

The p-value for Gender*Age indicates some evidence that growth curves for girls and boys have different slopes

- From now on we will use *Effect Summary* to find P-values. It gives the same information and allows model modification.

Summary of Fit without Interaction

RSquare	0.452001
RSquare Adj	0.42238
Root Mean Square Error	3.224234

✓ Adjusted R^2 went up

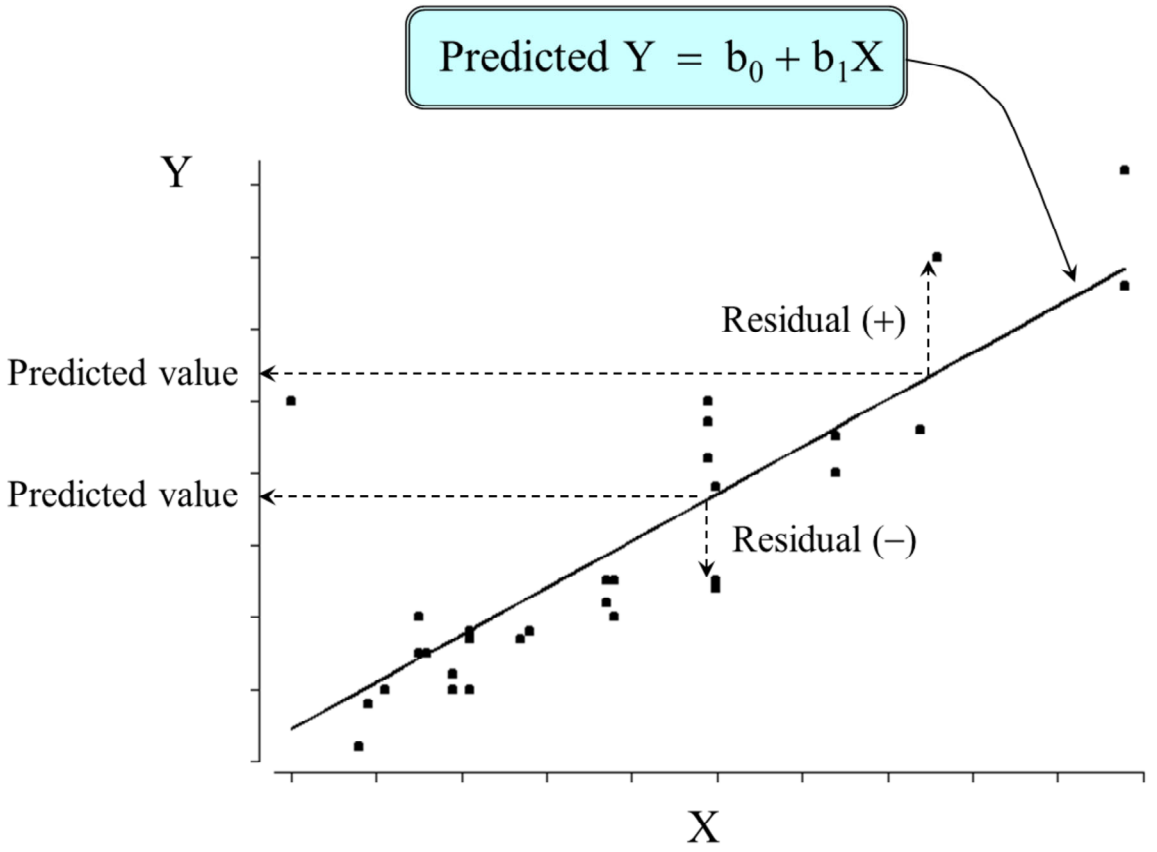
✓ RMSE went down

Summary of Fit with Interaction

RSquare	0.495046
RSquare Adj	0.452967
Root Mean Square Error	3.137706

Notes

Residuals Review



Notes

A fitted model, the equation generated during regression, gives the predicted mean value of the response variable as a function of the predictor variables. These predicted mean values are also called *predicted values*, or just *predicted* for short. The *residual value* is the data (observation) value minus the predicted value. Residual values are called *residuals* for short.

These terms are easiest to visualize in the simple linear model shown above. A predicted value is the fitted line evaluated at some X value. A residual is the difference between a measured (observed) Y value and the predicted value at the corresponding X.

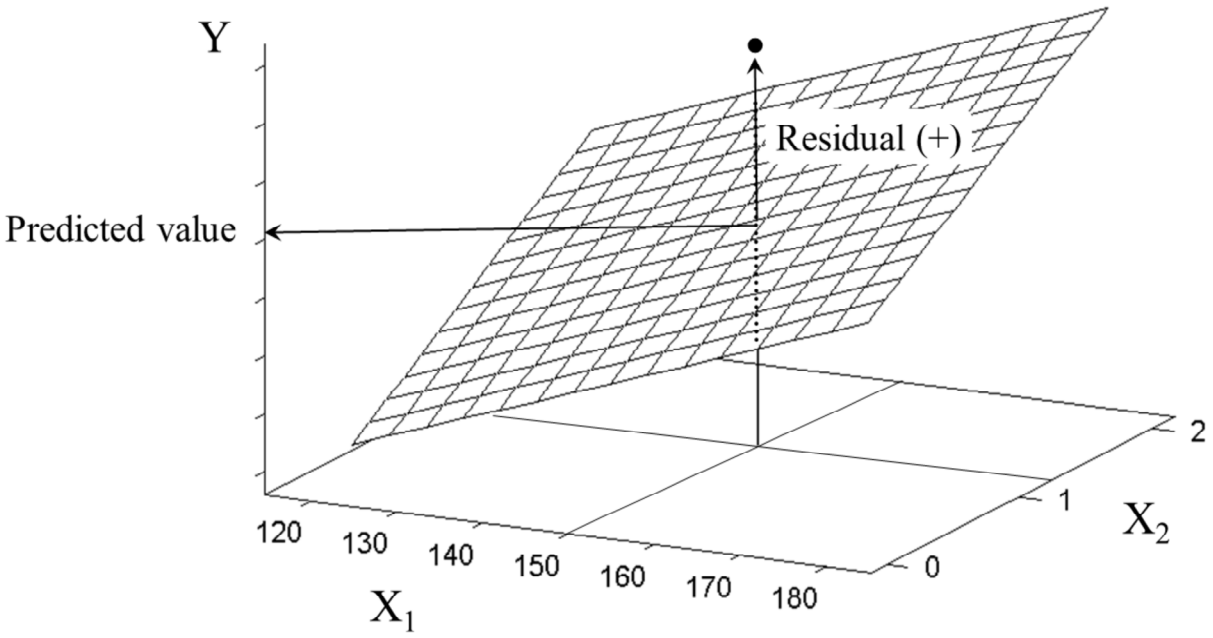
Residuals contain information about the magnitude and direction of variability in the data relative to the fitted model.

- An unusually large residual might signal a measurement error, data entry error or some other type of outlier.
- A systematic trend or pattern in the residuals might signal an inadequacy in the fitted model.

Notes

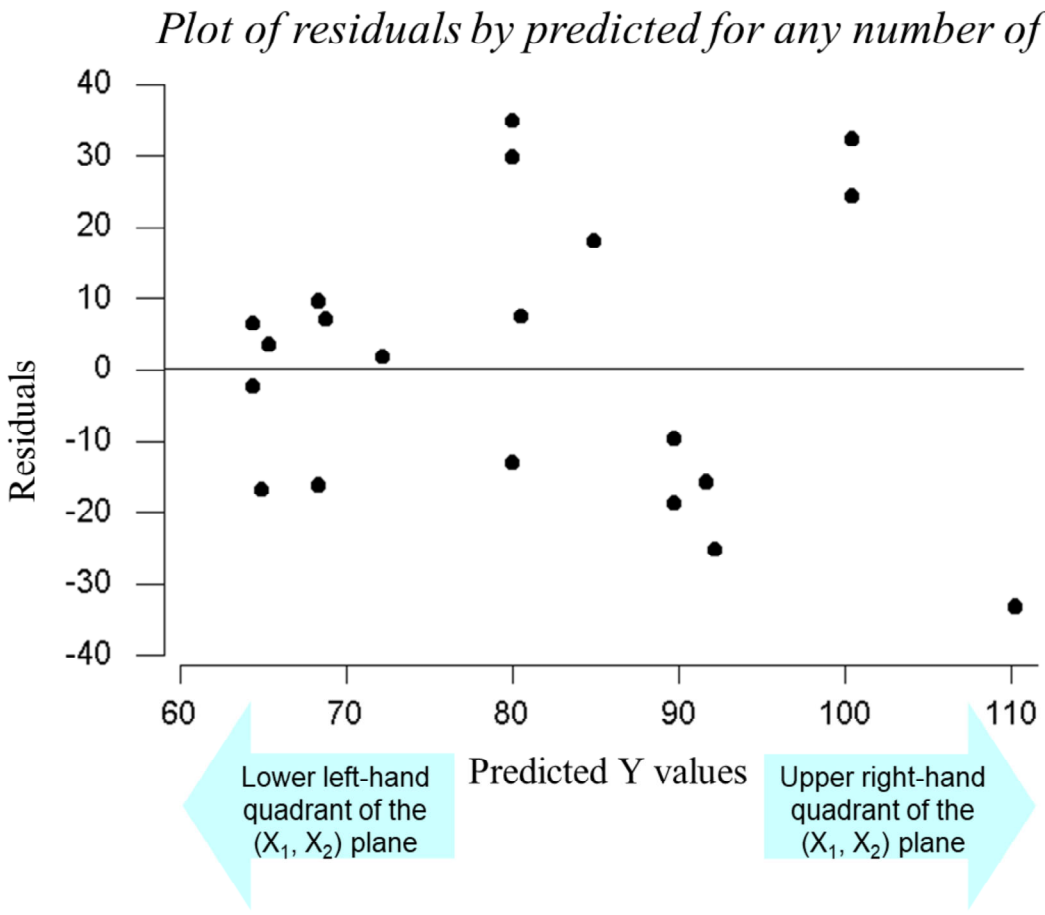
Residuals Review: Same thing for any number of X's

$$\text{Predicted } Y = b_0 + b_1X_1 + b_2X_2$$

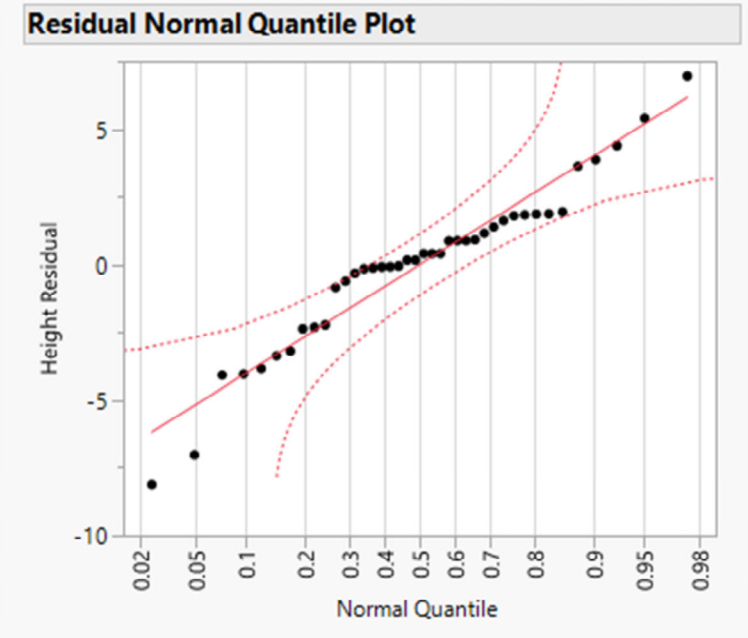


Notes

Same thing (cont'd)



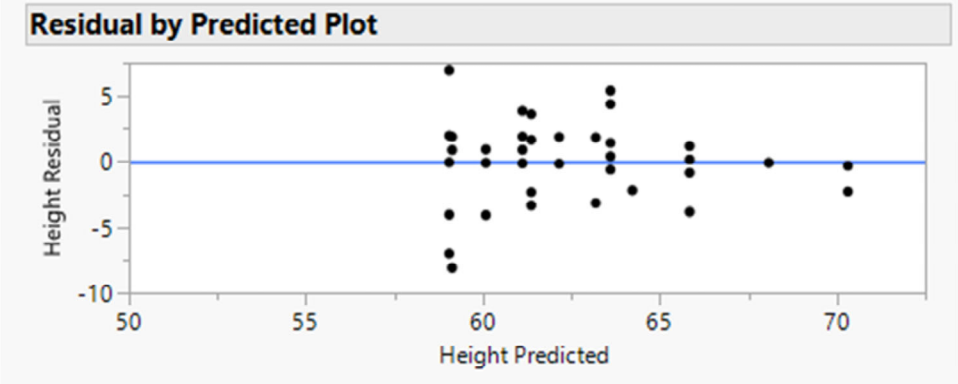
Notes



We can see points on the hyperbolic bands here, but there is not an obvious curve through the data. Given the small sample size, this is not too concerning.

Notes

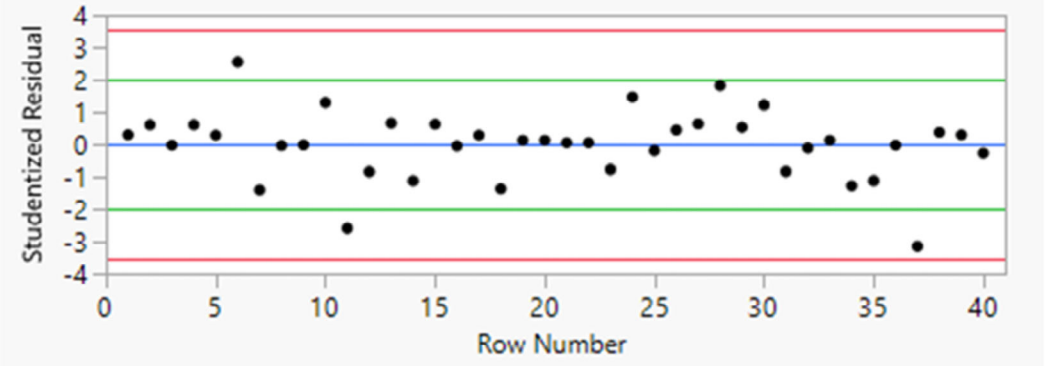
Checking model adequacy (cont'd)



In this plot, we can see that the variance in the residuals is decreasing as height increases. This indicates the need for a transformation. We will see how to do this a little later in the course.

Notes

Checking model adequacy (cont'd)



There are no obvious patterns in residuals in run order, and they scatter about zero. There is no concern here.

(Points outside the red limits are considered outliers, and should be investigated. Points outside the green limits but inside the red limits are possibly outliers, but with less certainty.)

Notes

Variance Inflation Factor (VIF)

When historical or observational data is used to generate a regression model, an additional test is needed:

- The variance inflation factor (VIF) must be checked
- The VIF indicates whether the regressors (i.e. Xs or predictors) are correlated with each other
  - $VIF = 1$: regressor is independent of all other regressors
  - $1 \geq VIF \geq 5$: regressor is moderately correlated to other regressors
  - $VIF > 5$: regressor is highly correlated with other regressors
- VIFs in the final model need to be less than 5
  - When X variables are correlated (high VIFs), the analysis makes statistical determinations based on the noise between the correlated variables. This will often result in high R^2 values but insignificant p values.
  - VIFs are often lowered when insignificant terms are removed from the model, and terms should be removed one at a time. The first term removed should be the one with the highest p value unless theory implies removing a different one.
  - High VIFs are not an issue in designed experiments, as the designs prevent high correlation between terms/regressors

Notes

Parameter Estimates

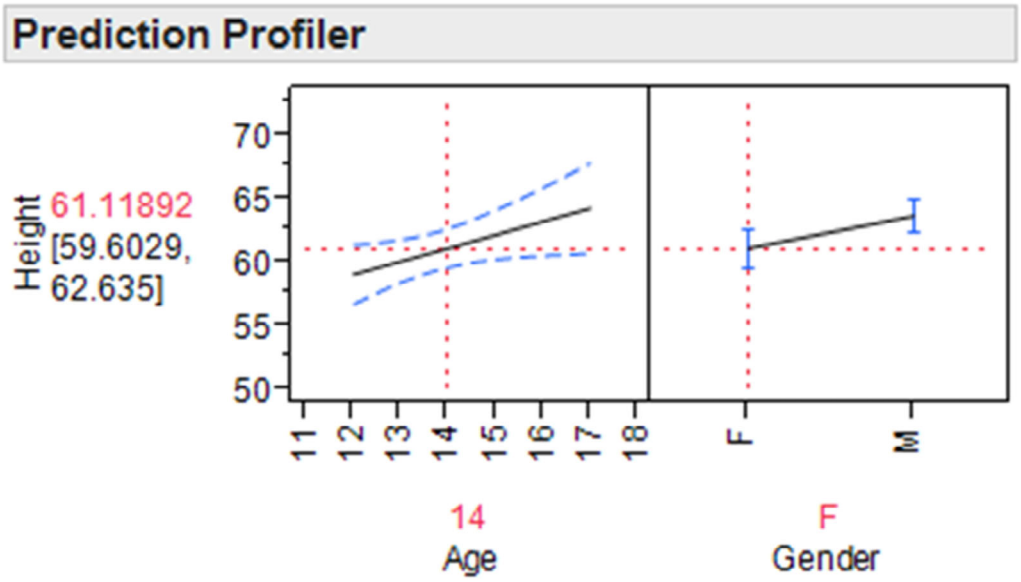
Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	39.457057	4.812681	8.20	<.0001*	.
Gender[F]	-1.227546	0.502444	-2.44	0.0196*	1.0154192
Age	1.6360307	0.343014	4.77	<.0001*	1.0155259
(Age-13.975)*Gender[F]	-0.600896	0.343014	-1.75	0.0883	1.0004648

The variance inflation factors for all terms in the model are below 5. There is no concerning level of correlation between model terms.

To display the VIFs, right click in the Parameter Estimates section, click Columns, then VIF.

Notes

Predicted values and associated uncertainty



Predicted avg. height in the population of 14 year old girls	61.12
95% confidence interval for avg. height of 14 year old girls	[59.60, 62.64] 61.12 ± 1.52

Notes

The model without interaction gave 61.25 ± 1.55 (slightly larger margin of error).

Notes

Steps in Multiple Regression (backward elimination method)

1. Run Analyze > Fit Model in JMP to investigate the relationship between y and x's. Use the Response Surface Model (all factors, all interactions, quadratic terms for continuous variables/factors)
2. Check model adequacy by reviewing the residuals plots:
   - Residual Normal Quantile Plot
   - Residual by Predicted Plot
   - Studentized Residuals (in run order)
3. Transform the data and resolve other issues, if needed.
4. Verify all VIFs < 5. Address the issue if any are over 5.
5. Remove insignificant terms from the model, that are not needed to maintain model hierarchy (main effects must be included if a higher order term of that variable remains in the model).
6. Use $Adjusted\ R^2$ to determine the amount of variation in Y that is explained by the model.

Notes

Notes

Your instructor will go through Exercise 5.4 as an example.

Notes

Exercise 5.1

a) In the table below, record the Adjusted R^2 and RMSE from the analysis of *Height* in this section. Also, record the P-values from *Effects Tests*. Run the same analysis for *Weight* and record the corresponding results.

			P-values		
Response	**Adj. R^2**	**RMSE**	Age	Gender	Age*Gender
Height					
Weight					

b) Which variable (*Height* or *Weight*) has the greater proportion of variation explained by *Age* and *Gender*?

b) Explain why it wouldn't make sense to compare the two models in terms of RMSE.

Notes

Exercise 5.1 (cont'd)

d) Both *Age* and *Gender* were statistically significant for predicting *Height*. Is this true for *Weight*?

e) For *Height* we found evidence that the growth curves for girls and boys have different slopes. Is this true for *Weight* as well? Give the P-value that is relevant to this question and explain what it means.

f) Give the predicted average *Weight* in the population of 15-year-old boys. Give a 95% confidence interval for this average.

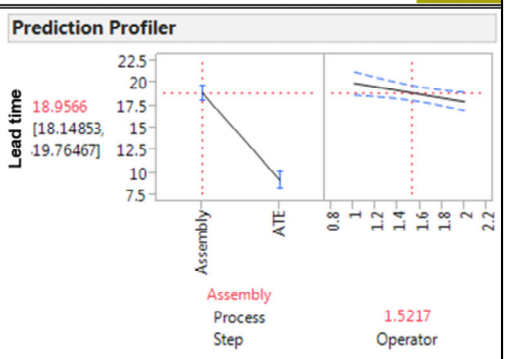
g) Save your scripts, close and save the data table.

Notes

Exercise 5.2

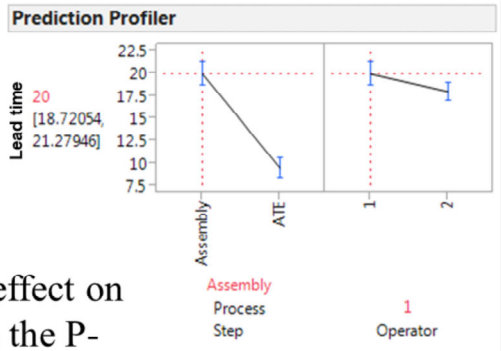
Data sets \ lead time 2.

a) Fit a model for *Lead time* including the terms *Process Step*, *Operator*, and their interactive effect. **Be sure you have the correct modeling type for *Operator*.** (If you got the upper right profiler, the modeling type for Operator is not correct. The lower right profiler is correct.)



b) Note anything concerning in the residuals plots.

c) Remove terms under *Effect Summary* with P-values exceeding 0.15 (*Remove* button). Which terms are left? Any issues with VIFs?



d) Based on the profiler, which factor has the larger effect on lead time (steeper slope)? Does this correlate with the P-values? Please explain.

e) Save your script, close and save the data table.

Notes

Exercise 5.3

Data sets \ number and size of defects.jmp.

a) Fit a model for *Max size* including the terms *Welder*, *# Defects*, their interactive effect, and the quadratic effect for *# Defects* (cross it with itself). This is the *Response Surface Model (RSM)* for one categorical factor and one continuous factor.

b) Do you see anything concerning in the residuals plots?

c) Using the *Effect Summary,* remove terms with P-values exceeding 0.15 (use the *Remove* button). Which terms are left in the model? Do all remaining terms have VIFs < 5?

d) Based on the profiler, which factor has the larger effect on *Max size*? Does this correlate with the P-values? Please explain.

e) Save your script, close and save the data table.

Notes

Exercise 5.4 [Instructor to demonstrate]

In this example you will analyze data from an optimization experiment concerning the removal of excess metal from castings by belt grinding.

The belt supplier had been recommending that belts be discarded when they are "50% used up." This rule was based on tests conducted by the supplier to define the usage point at which the total of labor and belt costs will be minimized. One of the grinders thought the supplier's rule caused grinders to discard belts too soon. Aside from being suspicious that the supplier just wanted to sell more belts, he argued that the supplier's tests did not take into account the time lost to belt changes.

This grinder developed a new standard under which belts would be discarded only after they were "75% used up." He wanted to do a comparative study to show that his method was cheaper overall. After he explains the study with his fellow grinders, 3 additional factors are added to the experiment.

Each casting in the experiment was weighed before and after the grinding operation. A technician kept track of how many belts were used and how long it took the grinder to complete each casting. From this information the total cost per unit of metal removed was calculated for each casting.

Data sets \ belt grinding.

Notes

Exercise 5.4 (cont'd) [Instructor to demonstrate]

- Y variable: *cost per unit of metal removed*
- X variables:
  - Contact wheel land-groove ratio (LGR): Low or High
  - Contact wheel material (MATL): Steel or Rubber
  - Belt usage limit (USAGE): "50%" or "75%"
  - Belt grit size (GRIT): 30 or 50
- **Run the *Fit Model* script provided in the left panel,** by clicking on the green triangle. This is the response surface model for 4 categorical X variables.
- Check the residuals plots. Any problems?
- Using the *Effect Summary,* remove insignificant terms not needed to maintain model hierarchy, starting with the group of terms with $P > 0.20$, then one at a time. Which terms are left in the model?
- Use the *Prediction Profiler* to find the minimum cost factor settings.
- What do you expect the mean and standard deviation of *Cost* to be after implementing the optimal factor settings?
- Save your script, close and save the data table.

Notes

Exercise 5.5

In this example you will analyze data from an optimization experiment concerning the bond strength of potato chip bags.

Chips 'R' Us was receiving customer complaints about stale chips, especially from customers on airplanes. They traced the problem to the bag sealing process. The current process involved a temperature of 150°C, a pressure of 100 psi and a dwell time of 1.1 secs. The current average bond strength was about 85 psi.

Process Engineer Chip Kettle ran an experiment to increase the bond strength. Production Manager Justin Thyme reminded Chip that he would very much like to avoid an increase in the dwell time.

Justin is able to free up a bag sealer for only so much time each shift. Chip realizes he will need two shifts to complete the experiment. He decides to include *Shift* as an additional variable in the analysis just in case there is an operator and/or equipment effect.

Data sets \ heat sealing 1.

Notes

Exercise 5.5 (cont'd)

- Y variable: *bond strength*
- X variables and feasible ranges:

Temperature (TEMP):	120 to 180
Pressure (PRESS):	50 to 150
Dwell time (DWELL):	0.2 to 2.0
Shift:	1 or 2

- **Run the *Fit Model* script provided in the left panel**. This is the response surface model (RSM) for 3 continuous X's. Is anything concerning in the residuals plots?
- Remove from the model insignificant terms that are not needed to maintain model hierarchy ($P > 0.15$), using the *Effect Summary*. Which terms are left?
- Use the *Prediction Profiler* to maximize the average bond strength. If your solution requires a long dwell time, manually move things around in the profiler to find another solution with a short dwell time.
- What do you expect the mean and standard deviation of *bond* to be after implementing the optimal factor settings?
- Save your script, close and save the data table.

Notes

Exercise 5.6

Data sets \ outgassing process. Current (the Y variable) is the electrical current required to heat a filament to a specified temperature. *Resist* (one of the X variables) is the electrical resistance of the filament. *Machine* (the other X variable) identifies which of three processing units was used. We want to develop a model for *Current* as a function of *Resist* and *Machine*.

a) Fit a response surface model for *Current*. (The terms will be *Resist*, *Machine*, the interaction term *Resist*Machine*, and the quadratic term *Resist*Resist*. To get the quadratic term, highlight *Resist* both under Select Columns and under Construct Model Effects, then click Cross.)

b) Do you see anything concerning in the residuals plots?

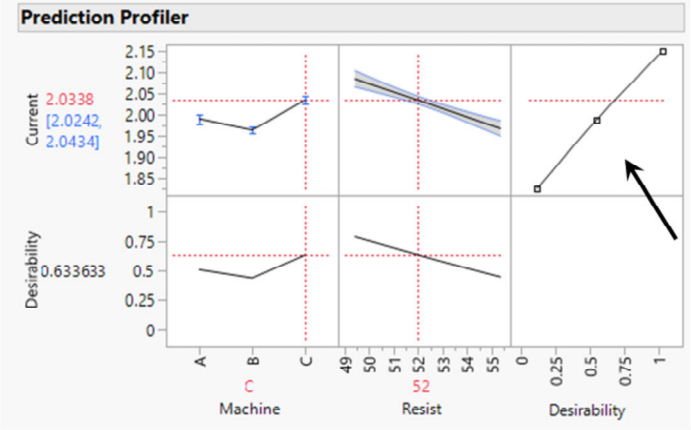
c) Remove any terms under *Effect Summary* with P value exceeding 0.15. (Use the *Remove* button.) Record the RMSE.

d) Use the *Prediction Profiler* to find the predicted average *Current* for each machine if we always use filaments with resistance 52.

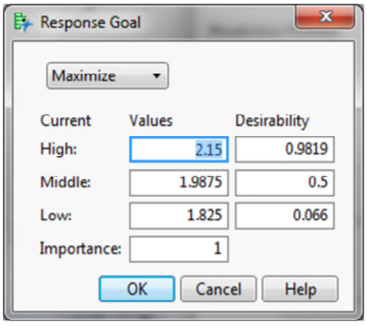
Notes

Exercise 5.6 (cont'd)

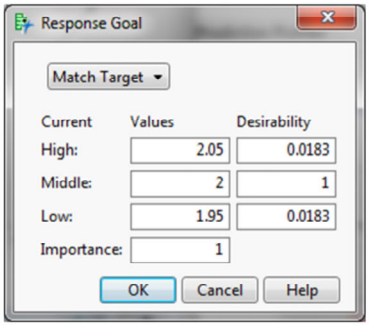
e) The target value for *Current* is 2. For each machine, we want to find the resistance for which the average current is 2. On the *Prediction Profiler* red triangle, select *Desirability Functions*. It should look like this:



f) Double click in the upper right hand panel of the profiler. (Try to avoid the plotted line.) You should get the dialog shown below.

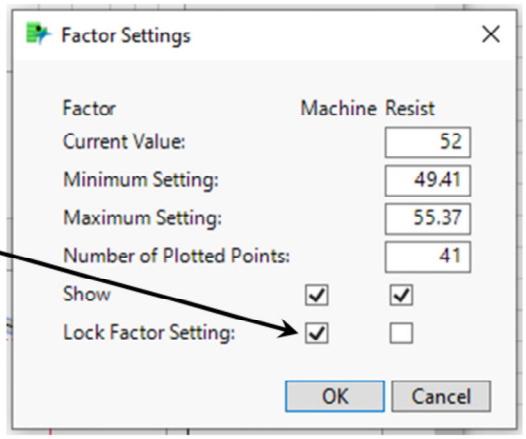


g) Modify the dialog as shown to the right, then select OK. Proceed to the next slide.

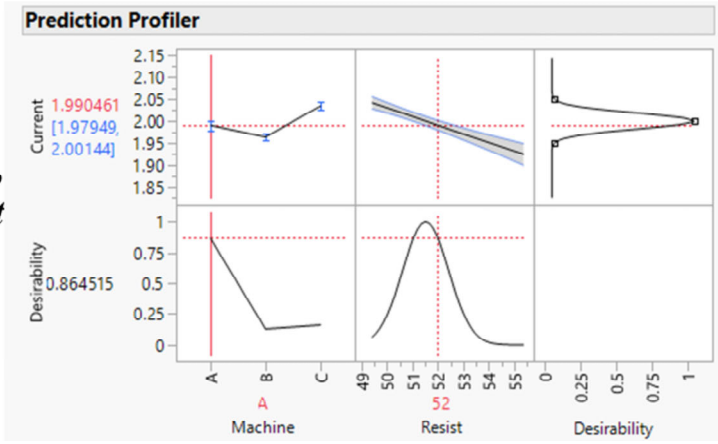


Notes

h) On the *Prediction Profiler* red triangle, select *Reset Factor Grid*. We want to lock the factor setting for *Machine*, so check the *Lock Factor Setting* box as shown here.



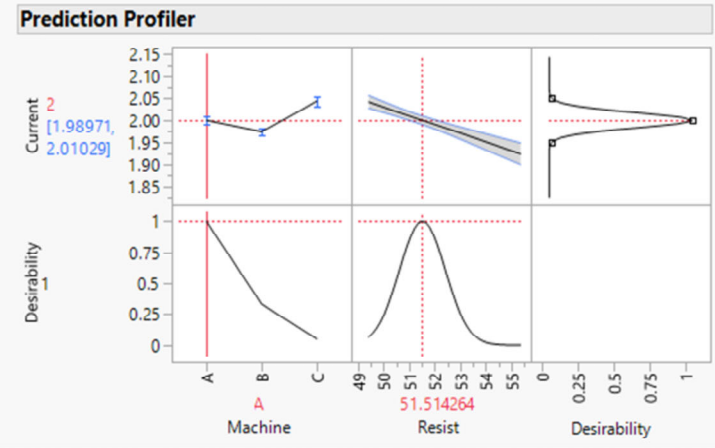
i) The vertical line for Machine should now be solid instead of dotted. **This will hold the machine setting in place during *Maximize Desirability*, which allows you to optimize *Resist* separately for each machine.** On the *Prediction Profiler* red triangle, select *Maximize Desirability*. Proceed to the next slide.



Notes

Exercise 5.6 (cont'd)

j) The optimal resistance value for Machine A is 51.5. Drag the solid vertical line across to B, then click *Maximize Desirability* to find the optimal resistance value for Machine B. Do the same for Machine C.



k) What will the average current be if we always use the optimal resistance values for each machine?

l) What will the standard deviation of current be if we always use the optimal resistance values?

m) Save your scripts, close and save the data table.

Notes

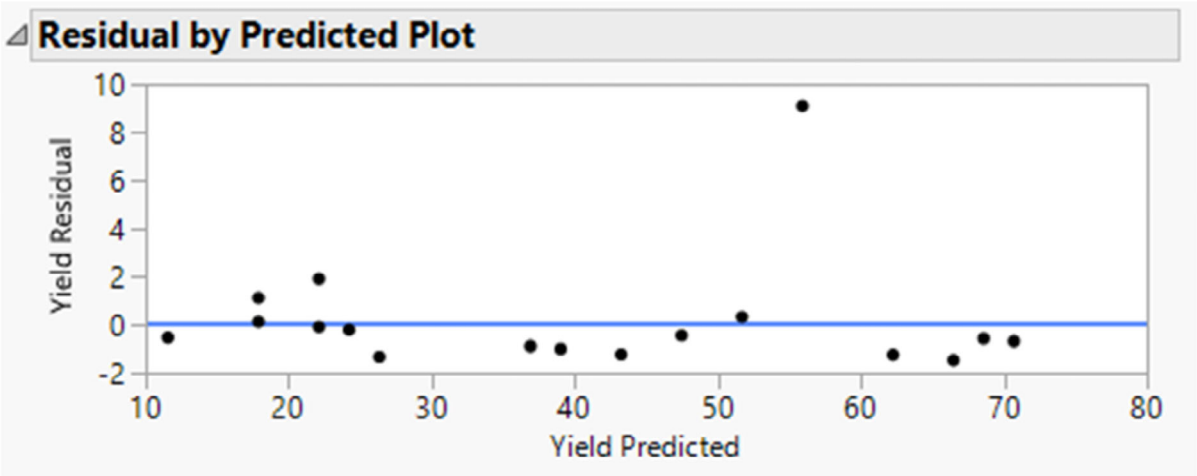
6 Dealing with Model Adequacy Issues

In this section, we will cover the most common model adequacy issues:

- Outliers
- Pattern in run order plot of residuals
- Multicollinearity (VIFs over 5)
- Unequal variance and non-normal residuals

Notes

Outliers can easily be seen on the Residual by Predicted and Studentized Residuals (residuals by run order) plots



Remember, healthy residuals look like random scatter about zero.

Here, it looks like there might be a suspicious data point.

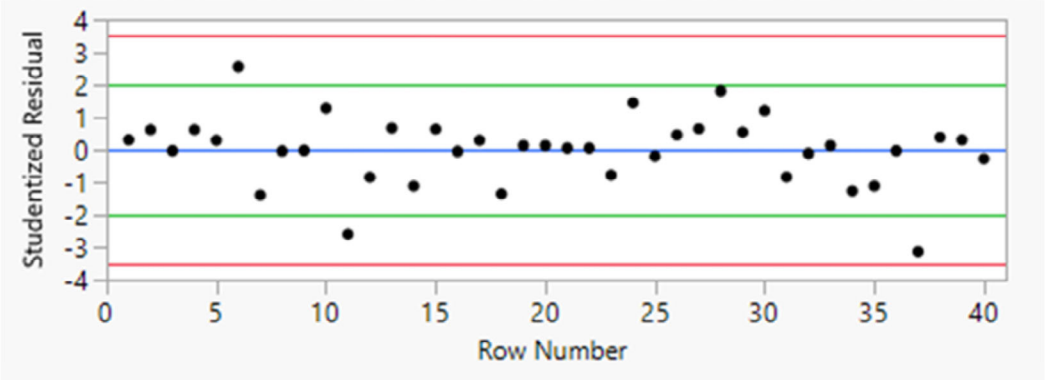
Notes

Issue: Outliers (cont'd)

- Investigate the data point.
  - If it turns out to be just a data entry error, we simply enter the correct value, then all is well. Most of the time it's not that simple.
- If you have an outlier of unknown origin:
  - Run the analysis with and without the questionable data point.
  - If you're lucky, the results will be pretty much the same both ways, hence no worries. Leave the data point in.
- If excluding the outlier does make a significant difference in the results, then you have a hard decision to make.
  - The official rule is: leave the data point in unless you can identify the cause. The idea is to throw it out only if you can demonstrate that it does not come from the population you want to study. This is the "pure" approach.
  - This should be tempered with the following practical consideration: you don't want your results to be unduly influenced by one extreme outlier, even if you can't explain it.

Notes

Issue: Pattern in run order of residuals



Remember, healthy residuals look like random scatter about zero.

There are no patterns of concern here.

Notes

Issue: Pattern in run order of residuals (cont'd)

- Runs (points in a row) of positive-negative-positive-negative residuals indicate correlation between runs in an experiment.
  - This implies that the assumption of independence has been violated.
  - **Randomization of an experiment protects against this! Do it every time!**
- This plot can show changes in variance over the time span of the experiment or data collection.
  - This could be due to increased skill as the experiment progresses, a process drift, operator fatigue, tool wear, etc.
  - This type of problem would show as an increase or decrease in spread or "scatter" of the residuals across the graph.
  - If there is x data available to support it, one remedy is to add a factor (time since tool change, number of hours of operator work, etc.)
  - Increasing or decreasing variance can also indicate the need for a transformation.

Notes

Issue: Multicollinearity (VIFs > 5)

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	4.868125	0.157585	30.89	<.0001*	.
LGR[Low]	0.616875	0.157585	3.91	0.0035*	1
Material[Rubber]	1.145625	0.157585	7.27	<.0001*	1
Usage[50%]	1.054375	0.157585	6.69	<.0001*	1
Grit[30]	-0.048125	0.157585	-0.31	0.7670	1
LGR[Low]*Grit[30]	-0.316875	0.157585	-2.01	0.0752	1
Usage[50%]*Grit[30]	0.395625	0.157585	2.51	0.0333*	1

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	14.044944	0.291958	48.11	<.0001*	.
Process Step[Assembly]	4.8792135	0.298829	16.33	<.0001*	1.0478749
Operator[1]	0.6713483	0.296556	2.26	0.0349*	1.0478749

Remember, VIF < 5 is not concerning.

- One aspect of factorial design experiments (often called DOEs) is that they are orthogonal designs. This results in the model terms being completely uncorrelated.
- Regressors that are completely uncorrelated with others have VIF = 1.
- High correlation is only a potential issue when using historical or observational data in regression analysis.

Notes

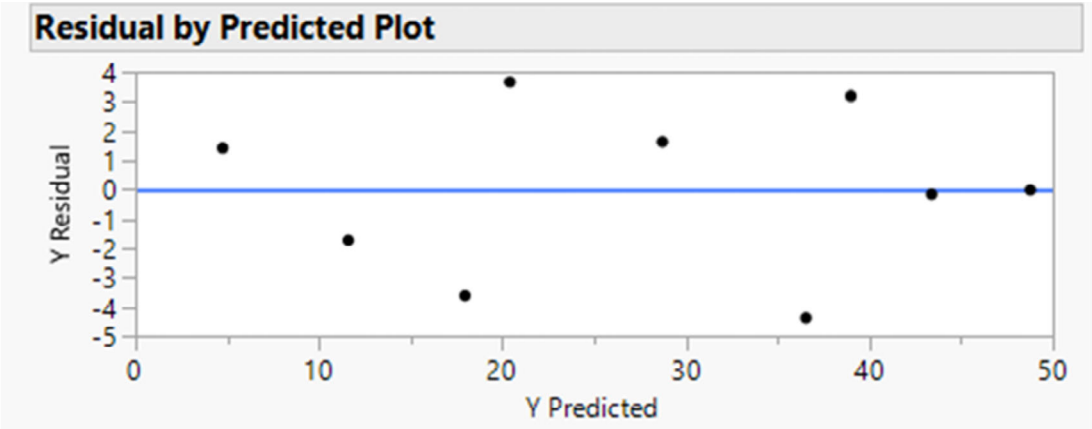
Issue: Multicollinearity (cont'd)

Several strategies can be tried for resolving multicollinearity, but they may not be satisfactory, especially if the model will be used for prediction.

- Collect additional data in a way that breaks up the multicollinearity.
  - Historical data may contain only certain combinations of x-variables, for example, only low levels of x_1 when x_2 is at a low level and only high levels of x_1 when x_2 is at a high level
  - Note: it may not be feasible or possible to collect this additional data.
  - In some cases, the factors (x's) are inherently correlated, for example as may be the case with household income and house size.
- Respecifying the model, can help.
  - If x_1 and x_2 are nearly linearly dependent, use one term, $x = x_1 + x_2$, which preserves the information content of the original variables
  - Try removing the term with the highest p-value, and look at that model. Then, replace it and remove the term with the highest VIF. See which gives the better model.
- Use ridge or principal-component regression (way beyond the scope of this course)

Notes

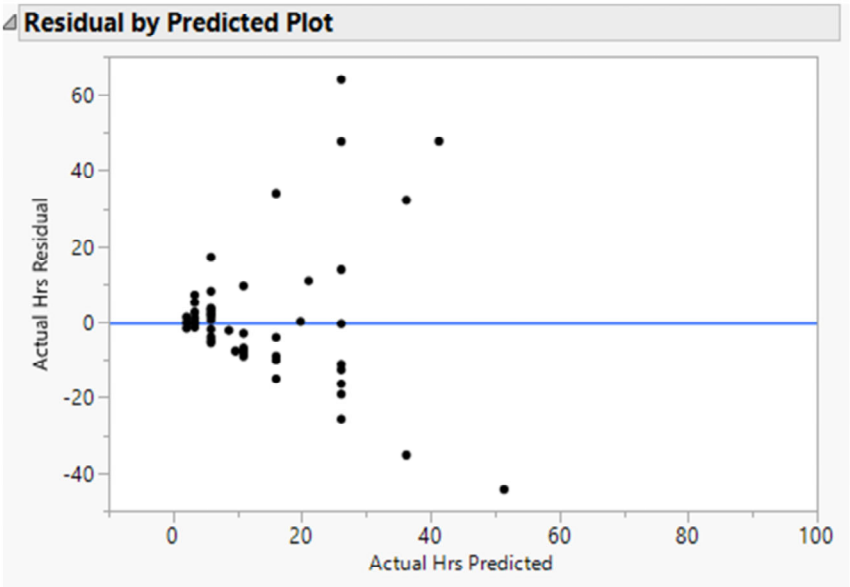
Issue: Unequal variance and non-normal residuals



Remember, the variation in the residuals should be fairly constant across the Residual by Predicted Plot. There is no issue here.

Notes

Issue: Unequal variance and non-normal residuals (cont'd)



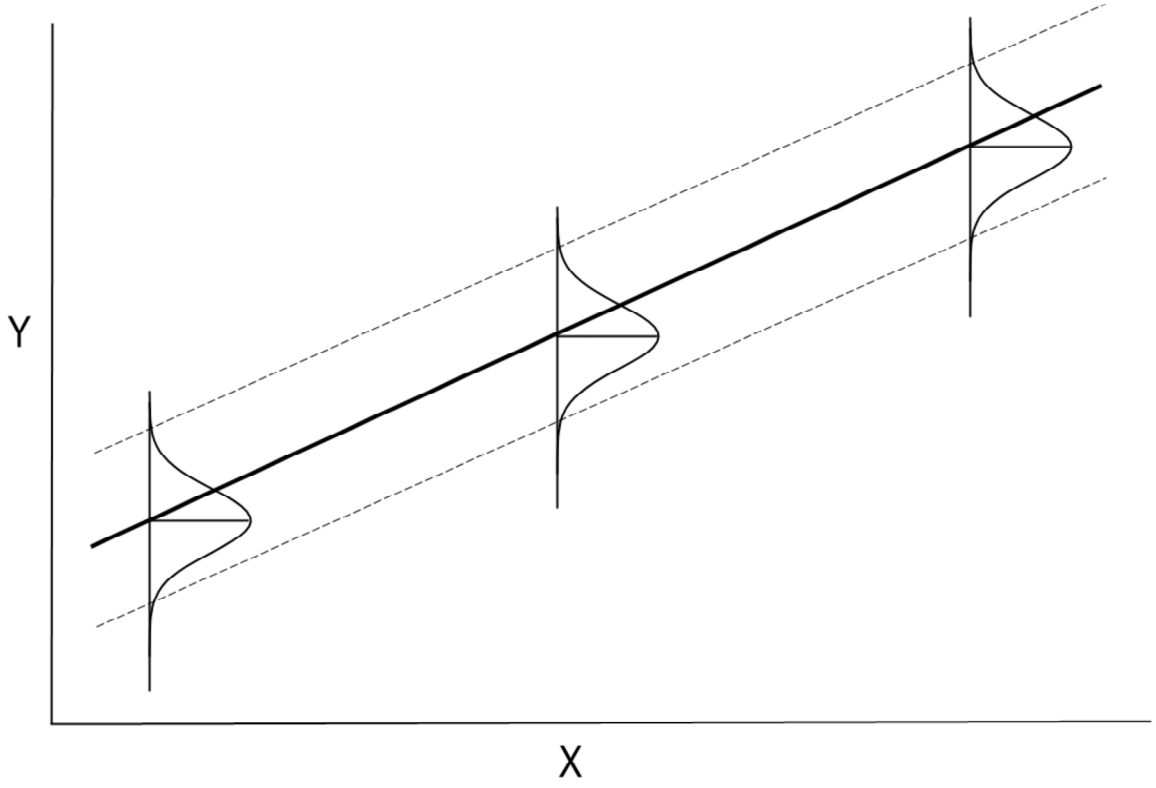
In this plot, we can see an issue.

σ_Y^2 proportional to mean Y $\rightarrow$ "sideways V"

Notes

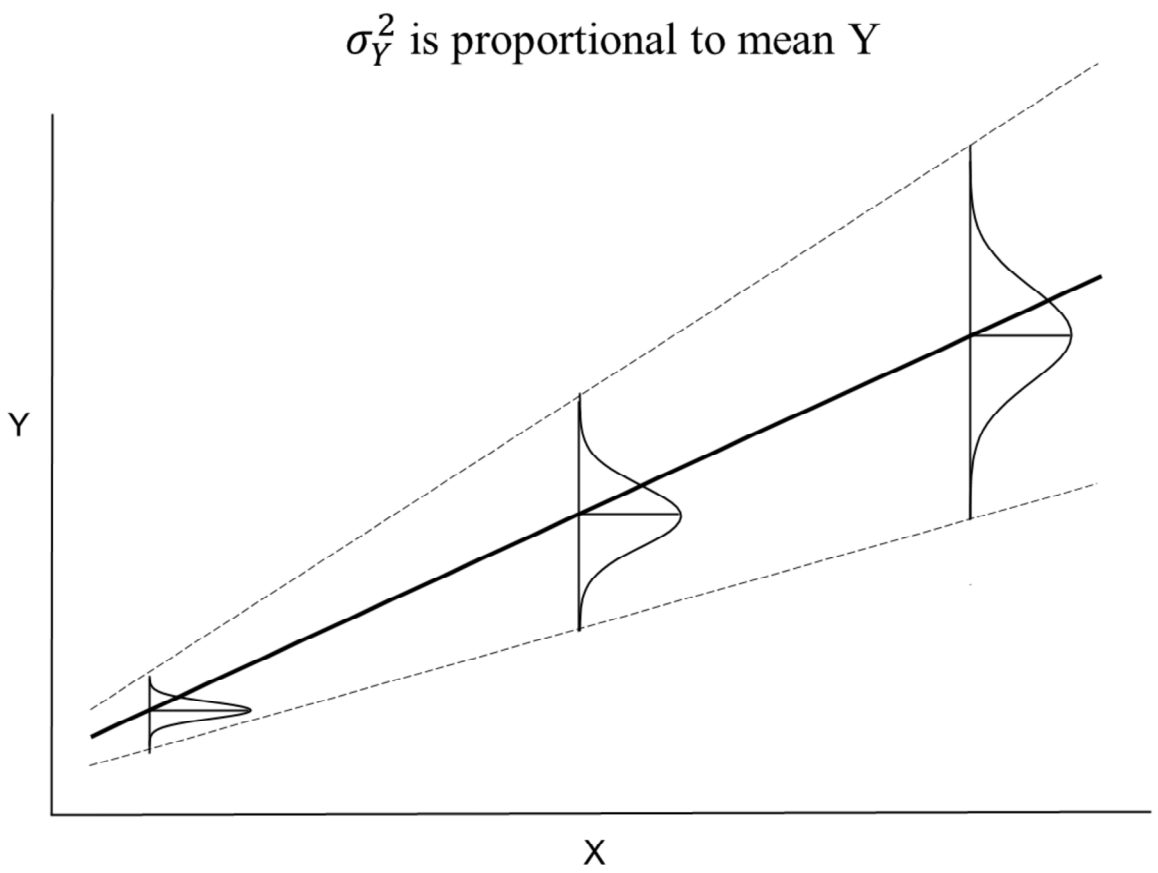
Basic model assumption: constant variance

σ_Y^2 is constant (does not depend on the X's)



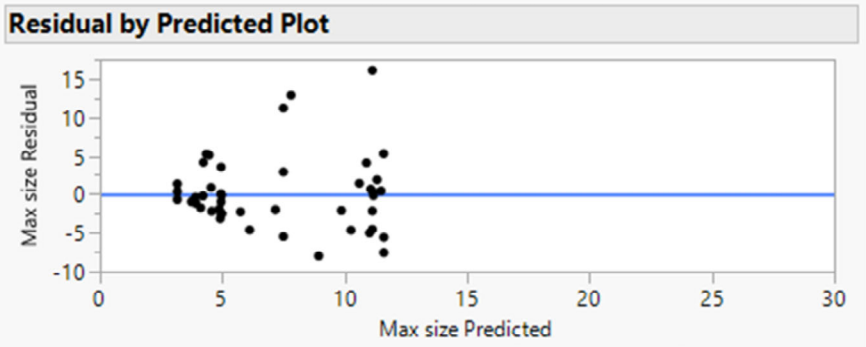
Notes

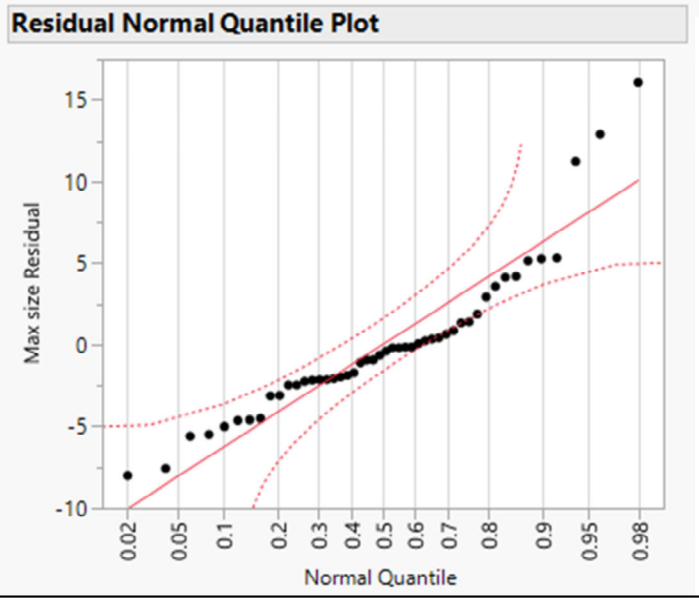
Most common violation of the basic assumption



Notes

Issue: Unequal variance and non-normal residuals (cont'd)





- Often, when there is an issue with constant variance, there is also the issue of non-normal residuals.
- This can be seen in these two plots
- Fortunately, they usually both resolve with the same treatment—a transformation.

Notes

The standard assumption in all comparison and correlation analyses involving a quantitative Y variable is that the noise (unexplained/error/residual) variation follows a Normal distribution with mean 0 and a standard deviation that does not depend on the X variables.

This simple model has served us well. However, when Normality or constant σ is grossly violated, something must be done. The most common remedy is to use log(Y) or sqrt(Y) as the dependent variable instead of Y. This is a transformation. This "trick of the trade" is simple and, in most cases, effective.

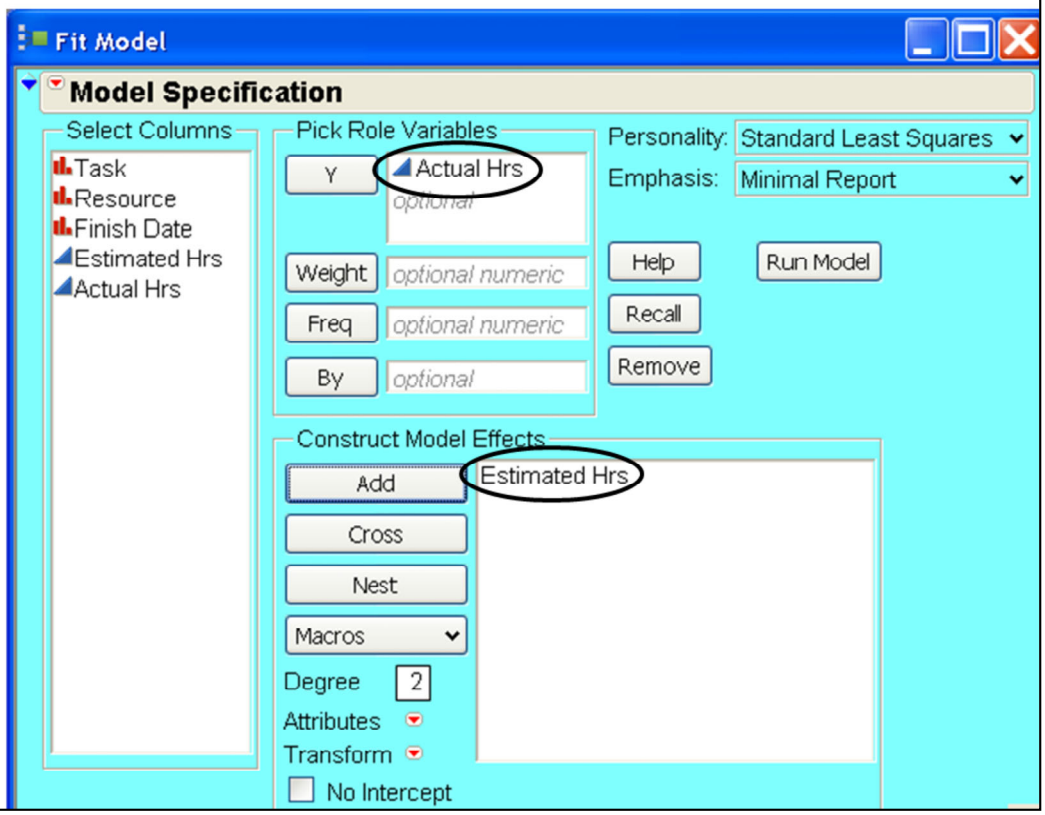
Notes

Transforming the Y variable

Data sets \ actual vs estimated

We want to see how accurately we can estimate the time it takes to do certain tasks

Analyze
↓
Fit Model

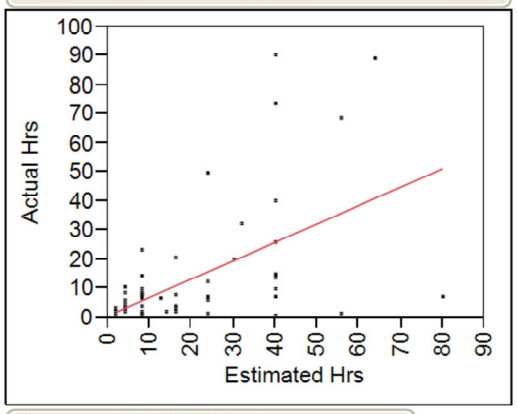


Notes

Transforming Y (cont'd)

Response Actual Hrs

Regression Plot



Summary of Fit

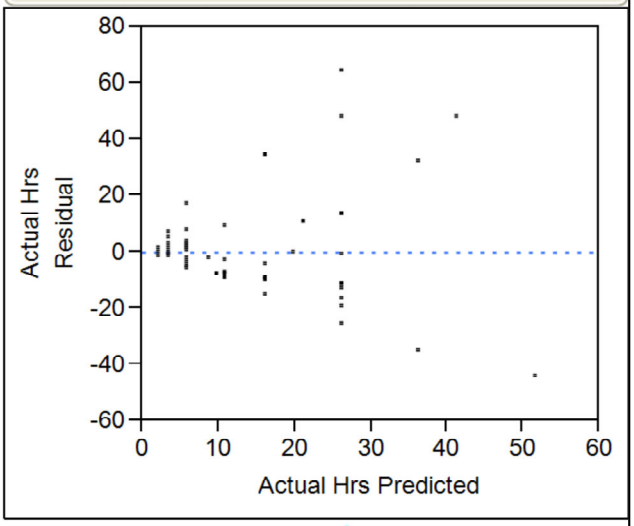
RSquare	0.307347
RSquare Adj	0.296176
Root Mean Square Error	16.95281
Mean of Response	12.23828
Observations (or Sum Wgts)	64

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.8352064	3.035964	0.28	0.7842
Estimated Hrs	0.6321871	0.120529	5.25	<.0001*

$Y = 0.835 + 0.632\ X$

Residual by Predicted Plot

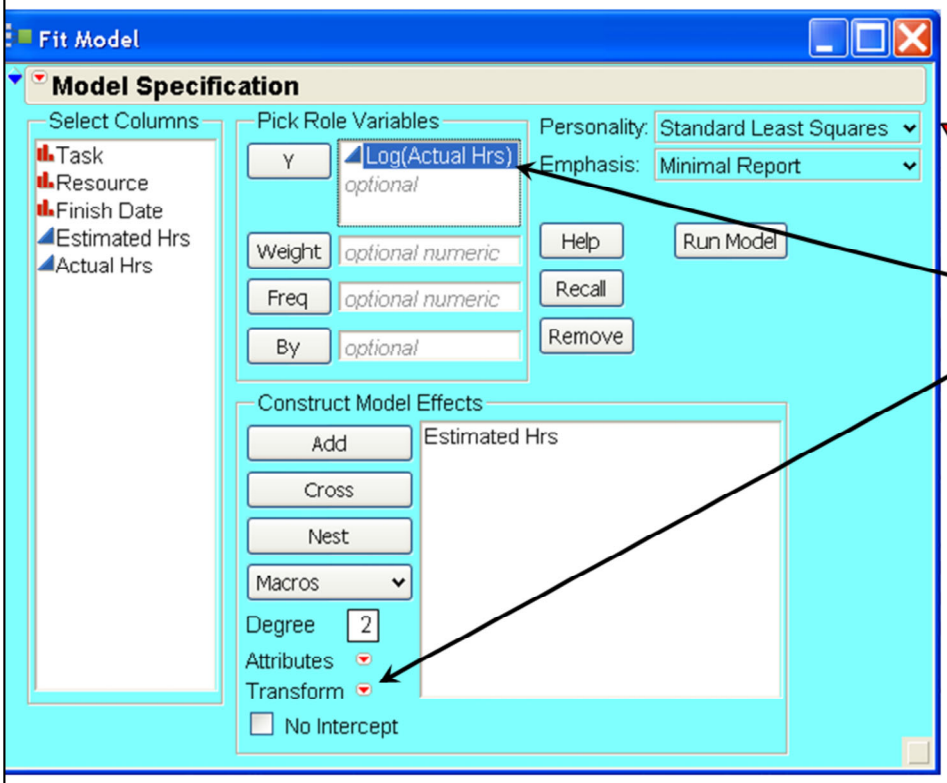


Variation increases as average *Actual Hrs* increases

Notes

Transforming Y (cont'd)

σ_Y proportional to mean Y ⟺ $\sigma_{Log(Y)}$ constant



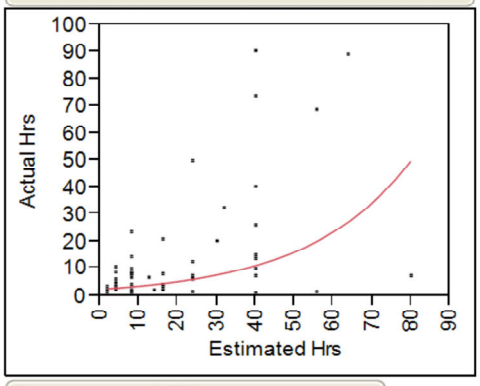
- Response Actual Hrs > Model Dialog
- Click on *Actual Hrs*
- Click on *Transform* red triangle
- Select *Log*
- Run the model

Notes

Effects of log transformation

Response Log(Actual Hrs)

Regression Plot



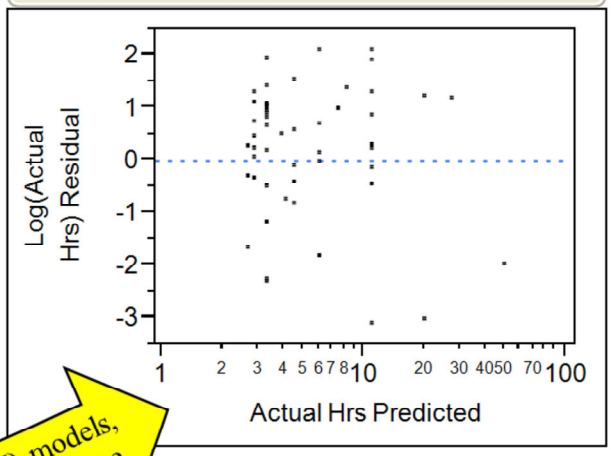
Summary of Fit

RSquare	0.233276
RSquare Adj	0.22091
Root Mean Square Error	1.217933
Mean of Response	1.576584
Observations (or Sum Wgts)	64

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.8982207	0.218111	4.12	0.0001*
Estimated Hrs	0.0376085	0.008659	4.34	<.0001*

Residual by Predicted Plot



For Log(Y) models, use Log scale here

Nonlinear model for Y

$$\text{Log}(Y) = 0.898 + 0.038X \rightarrow Y = \exp(0.898 + 0.038X) = e^{0.898}\left(e^{0.038}\right)^X = 2.45(1.04)^X$$

Notes

Note on JMP notation, and impacts of the Log transformation

JMPs notation regarding Logs requires some clarification:

- Although JMP expresses the logarithm as "Log", it is actually base e, or the natural log, which is usually written as Ln. It is not a base 10 logarithm.
- However, the plots that use a log transformed X-axis display use base 10 log for the X-axis. This does not change the interpretation of the chart.

The impact of transformation on R^2 and p-values:

- In the previous example, a transformation was required because the residuals variance wasn't constant over the range of the predicted values.
- After the transformation, the R^2 value went down. This can lead to a belief that the non-transformed model was "better". However,
- Residuals showing this condition (heteroscedasticity) can cause p-values and R^2 to be over or under stated.
- When this condition occurs, the problem must be corrected. The resulting model, even if R^2 is lower or p-values are higher, is the more "real" model.

Notes

Steps in Multiple Regression (backward elimination method)

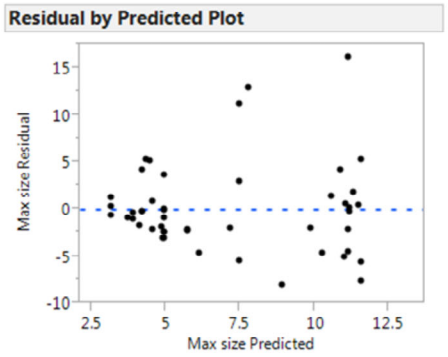
1. Run Analyze > Fit Model in JMP to investigate the relationship between y and x's. Use the Response Surface Model (all factors, all interactions, quadratic terms for continuous variables/factors)
2. Check model adequacy by reviewing the residuals plots:
   - Residual Normal Quantile Plot
   - Residual by Predicted Plot
   - Studentized Residuals (in run order)
3. Transform the data and resolve other issues, if needed.
4. Verify all VIFs < 5. Address the issue if any are over 5.
5. Remove insignificant terms from the model, that are not needed to maintain model hierarchy (main effects must be included if a higher order term of that variable remains in the model).
6. Use $Adjusted\ R^2$ to determine the amount of variation in Y that is explained by the model.

Notes

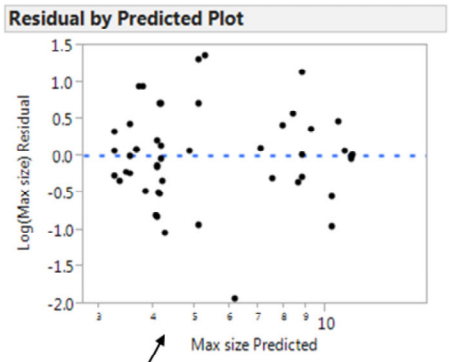
Exercise 6.1

Data sets \ number and size of defects.jmp

a) Fit a model for *Max size* including the terms *Welder*, *# Defects*, their interactive effect, and the quadratic effect for *# Defects* (*response surface model* for one continuous factor and one categorical factor). You should see a distinct sideways V. Do you see issues in any other residuals plots?



b) Select *Model Dialog* on the *Response* red triangle menu, apply a Log transformation to *Max size*, re-run the model. The sideways V isn't completely gone, but close enough. Did other residuals plots improve?



c) Use *Effect Summary* to remove terms with P > 0.15.

Remember to change the x-axis on the plot, as well.

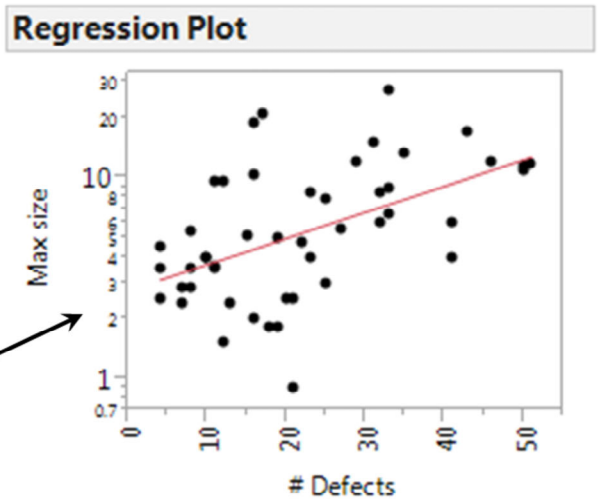
Notes

Exercise 6.1 (cont'd)

d) Which terms are left in the model?

e) Now we have a log-linear simple regression.

When you use a Log or square root transformation on Y, it is helpful to use same scale for the Y axes of the plots

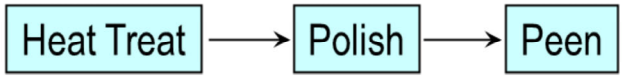


f) Save your script, close and save the data table.

Notes

Exercise 6.2

An aerospace manufacturer uses integral castings as structural components of jet engines. Integral castings give design engineers more flexibility and simplify the assembly process. Defect-free castings are known to have long cycle fatigue life, but defects often arise in the casting process and must be weld repaired. The engine manufacturer's metallurgical team has proposed a finishing process of the following type to ensure adequate cycle fatigue life of weld-repaired castings:



The team wants to optimize the first two steps in this process to achieve maximum cycle fatigue life. Also, though other applications of similar processes have included peening, they would like to see if it can be omitted to reduce processing time and cost.

Due to project time constraints and limited availability of test fixtures, the team can perform at most 12 cycle fatigue tests for their experiment.

Notes

Exercise 6.2 (cont'd)

- Y variable: *Cycles* (to failure)
- X variables:
  - Heat treat: Anneal or Solution/age
  - Polish: Chemical or Mechanical
  - Peen: Yes or No
- *Data sets \ weldment fatigue.jmp.*
- Run the *Model* script provided in the left panel, run the model.
- Notice the extreme sideways V on the *Residual by Predicted Plot*. Are there issues in any of the other residuals plots? If yes, what are they?
- Rerun the model using a Log transformation on *Cycles*. Did residuals plots improve?
- Remove insignificant terms from the model ($P > 0.15$) that are not needed to maintain model heirarchy.
- Use the *Prediction Profiler* to maximize the cycle fatigue life.

Notes

Exercise 6.3

A Black Belt wants to minimize the *leak rate* in plastic containers ultrasonically welded together. The X variables and ranges are:

- Force: 70 to 150
- Energy: 275 to 325
- Amplitude: 70 to 90

- *Data sets \ ultrasonic welding 1.jmp.*
- Run the *Model* script provided in the left panel.
- What problems do you see in the residuals plots?

Notes

Exercise 6.3 (cont'd)

- Rerun the model using the Log transformation on *leak rate*. (Be sure to change the x-scale to Log on the Residual by Predicted Plot.)
- Rerun the model using the Sqrt transformation on *leak rate*. (Be sure to change the x-scale to Sqrt on the Residual by Predicted Plot.)
- Which set of residuals plots looks better? Use whichever transformation looks like it worked better, going forward.
- Remove insignificant term(s) from the model ($P > 0.15$), while maintaining model hierarchy.
- Use the *Prediction Profiler* to minimize the leak rate.

Notes

This page intentionally left blank

7 Simple Regression with Pass/Fail Y

When the response variable, Y, is binary (pass/fail, yes/no, success/failure, etc.), the regression model used for a continuous Y-variable *cannot* be used.

- A *logistic response function* must be used
- The resulting analysis yields an equation that allows us to calculate **event probability:**

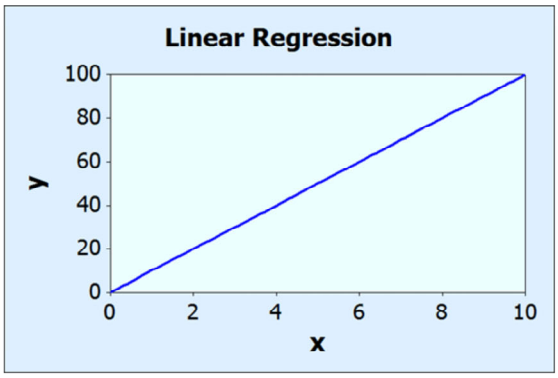
$$P_{event} = f(x_1, x_2, \ldots, x_n)$$

- This equation is used to answer questions such as:
  - What is the probability of being in spec (at various levels of x)?
  - What is the probability of getting the contract?
  - What is the probability of a defect?

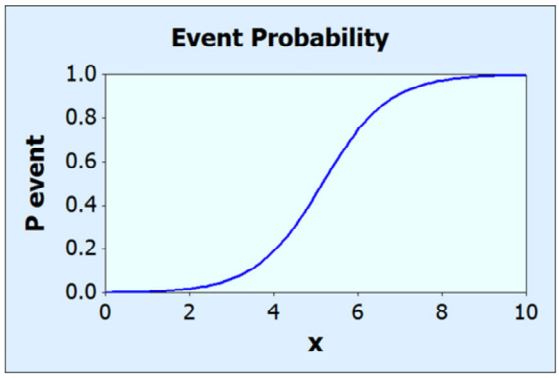
Notes

Probability Function for Pass/Fail Y

This probability function, the *logistic response function*, has a much different behavior than a linear regression function:



- The *y* values of a linear regression can have any values



- The logistic response function is an S-shaped function that can only have values between 0 and 1

To be useful in prediction, the logistic response function must be transformed into an unbounded linear function

Notes

Transforming the logistic response function (cont'd)

The *logit transformation* is used to linearize the model:

$$logit(P_{event}) = \ln\left(\frac{P_{event}}{1 - P_{event}}\right) = b_0 + b_1x_1 + \ldots + b_nx_n$$

$$\frac{P_{event}}{1-P_{event}} = e^{b_0+b_1x_1+\ldots+b_nx_n}$$

$$P_{event} = \frac{1}{1 + e^{-(b_0+b_1x_1+\ldots+b_nx_n)}}$$

- This is the form of the final equation in the regression analysis
- The *maximum likelihood* method is used to estimate the parameters in this probability equation . . . JMP does this work for us
- We can use this equation (model) to predict the probability of an event for various levels of $x_1, x_2, \ldots, x_n$

Notes

Using JMP for Simple Regression with Pass/Fail Y

We will see how to use JMP do the regression analysis when we have:

a) Raw data – each row represents one part or transaction

b) Tabulated data – each row represents multiple parts or transactions

Notes

Raw data

2/0 Cols / 25/0	Target speed	Result
1	200	Hit
2	205	Hit
3	210	Hit
4	215	Hit
5	220	Hit
6	225	Miss
7	230	Hit
8	235	Hit
9	240	Miss
10	245	Hit
11	250	Hit
12	255	Hit
13	260	Hit
14	265	Miss
15	270	Miss
16	275	Hit
17	280	Miss
18	285	Miss
19	290	Miss
20	295	Miss
21	300	Hit
22	305	Miss
23	310	Miss
24	315	Miss
25	320	Miss

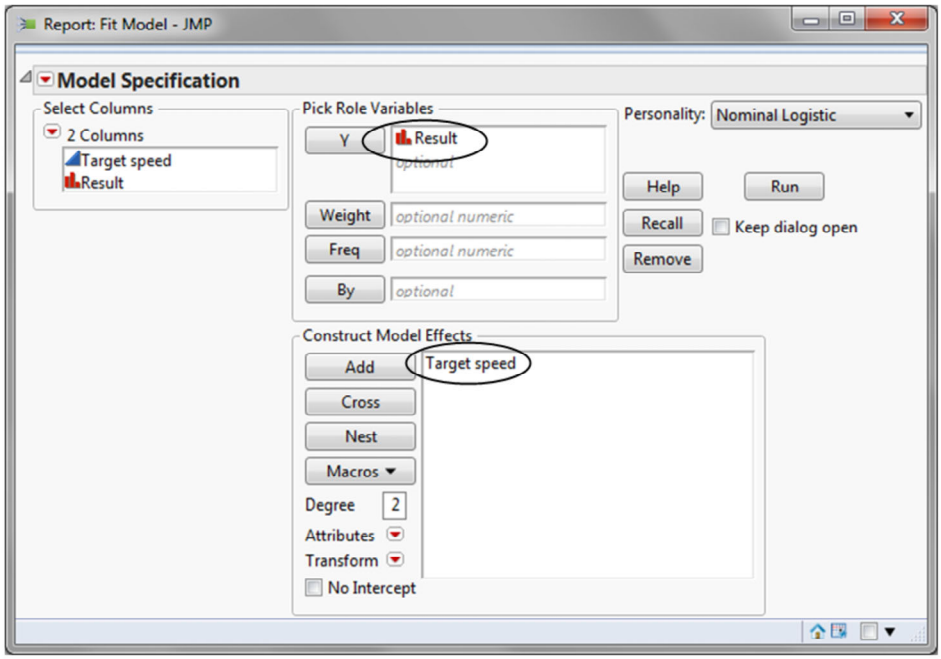
Data sets \ target practice

↓

Fit Model

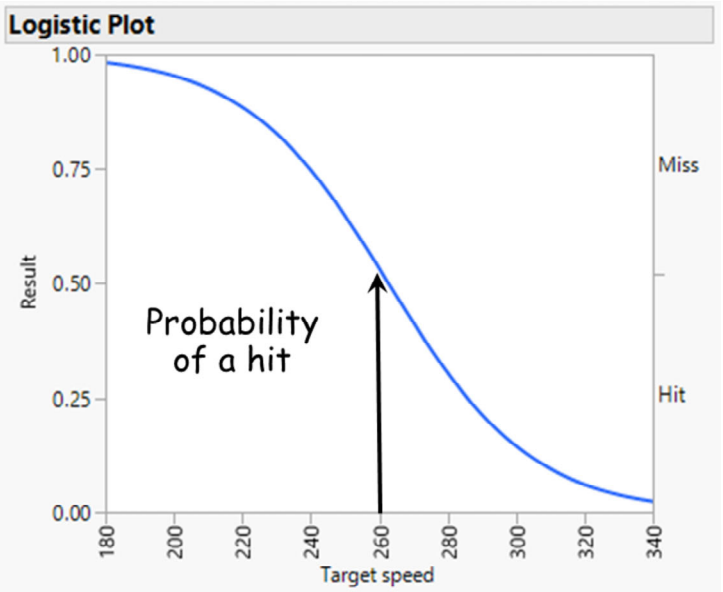
↓

Set up as shown



Notes

Analysis output



Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	12.5297022	4.7931154	6.83	0.0089*
Target speed	-0.0476836	0.0181939	6.87	0.0088*

For log odds of Hit/Miss

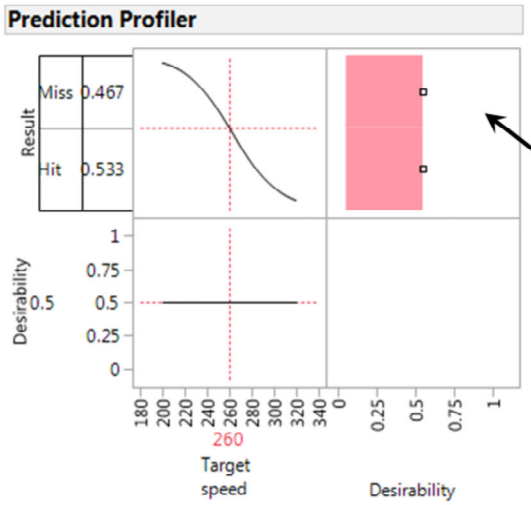
Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Target speed	1	1	11.1939322	0.0008*

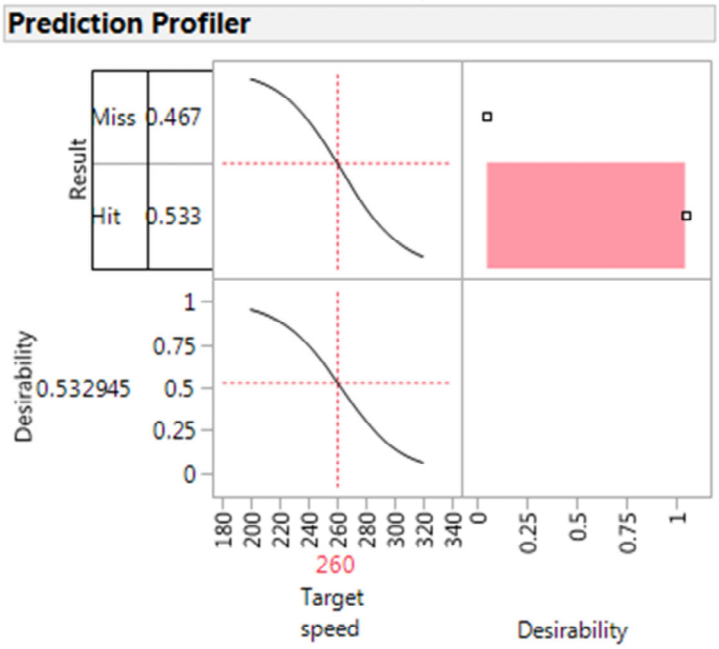
- P-value for correlation (this is the one that matters)
- Very strong evidence of a negative correlation between the speed of the target and the probability of hitting it

Notes

The prediction profiler



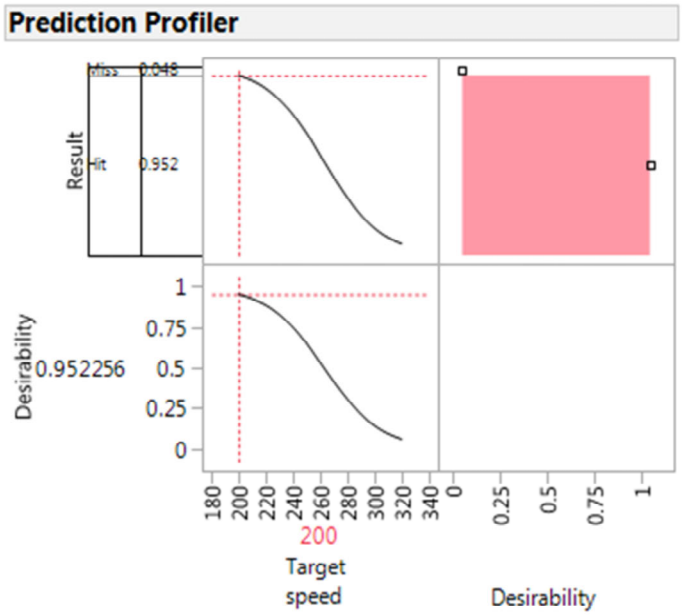
- *Red Triangle* → Profiler → *Prediction Profiler* red triangle → *Optimization and Desirability* → *Desirability Functions*
- Double-click in the blank area, enter 1 for *Hit* and 0 for *Miss* → OK → OK → next slide



Notes

Prediction profiler (cont'd)

Prediction Profiler red triangle → *Optimization and Desirability* → *Maximize Desirability*



- The target speed of 200 produces the maximum hit probability of 0.952
- The corresponding miss probability is 0.048
- The target speed of 320 produces the minimum hit probability of 0.061
- The corresponding miss probability is 0.939

Notes

Exercise 7.1

Open *Data sets \ quotation process.jmp*.

a) Fit *PO* by *TAT*. Which P-value in the output is the most reliable?

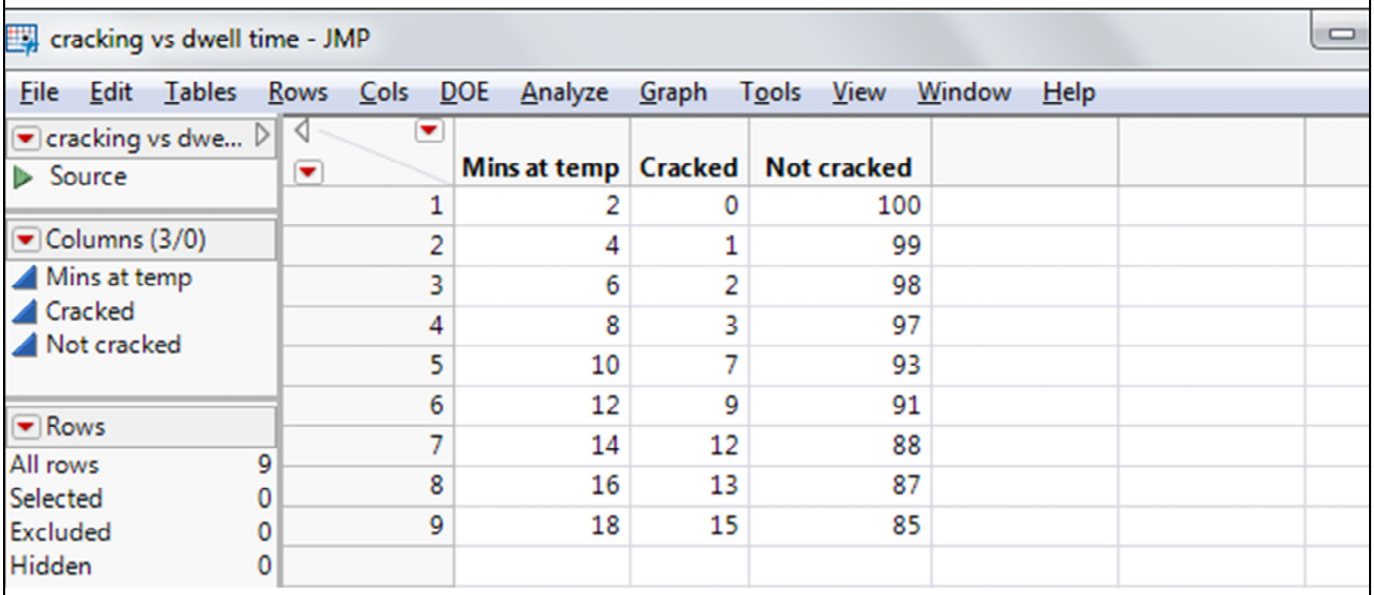
b) Does the PO hit rate increase or decrease as the TAT increases?

c) Find the PO hit rates for 3 day and 15 day turnarounds.

d) Save your script, close and save the data table.

Notes

Data sets \ cracking vs dwell time

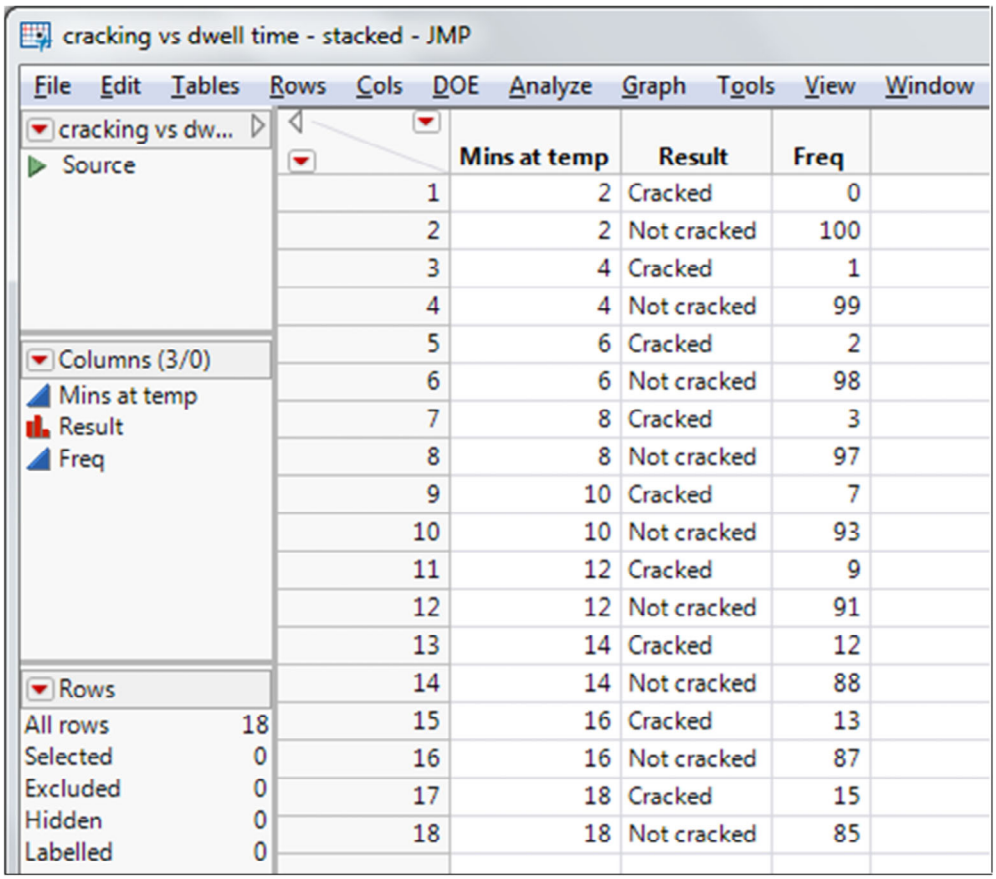


	Mins at temp	Cracked	Not cracked
1	2	0	100
2	4	1	99
3	6	2	98
4	8	3	97
5	10	7	93
6	12	9	91
7	14	12	88
8	16	13	87
9	18	15	85

1) Tables → Stack
2) Use *Cracked* and *Not cracked* as the stack columns
3) Change *Label* to *Result*, change *Data* to *Freq* → OK
4) Save as *cracking vs dwell time stacked*

Notes

Stacked format

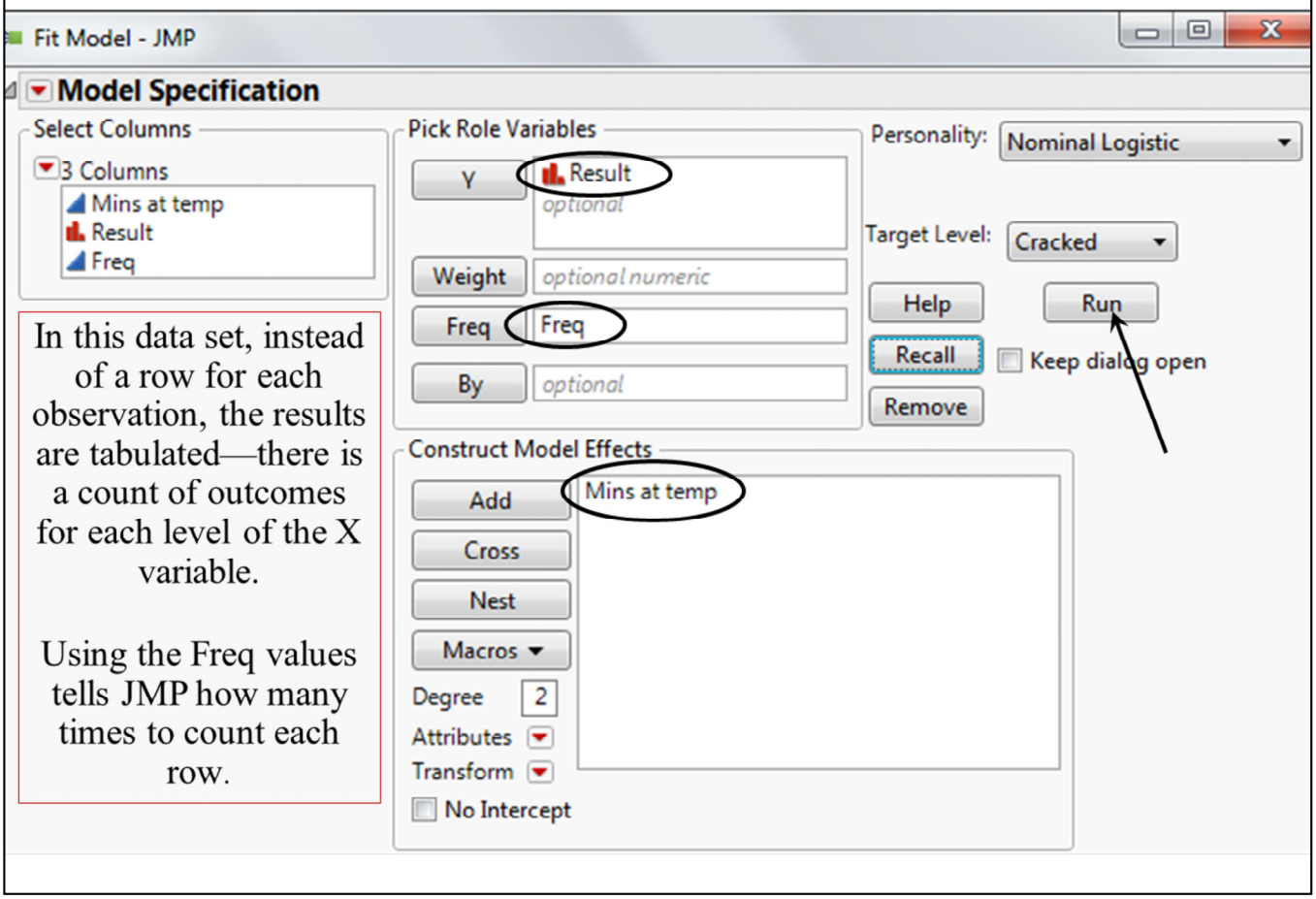


	Mins at temp	Result	Freq
1	2	Cracked	0
2	2	Not cracked	100
3	4	Cracked	1
4	4	Not cracked	99
5	6	Cracked	2
6	6	Not cracked	98
7	8	Cracked	3
8	8	Not cracked	97
9	10	Cracked	7
10	10	Not cracked	93
11	12	Cracked	9
12	12	Not cracked	91
13	14	Cracked	12
14	14	Not cracked	88
15	16	Cracked	13
16	16	Not cracked	87
17	18	Cracked	15
18	18	Not cracked	85

Analyze
↓
Fit Model
↓
See next slide
↓
Set up as shown

Notes

Fit Model



Notes

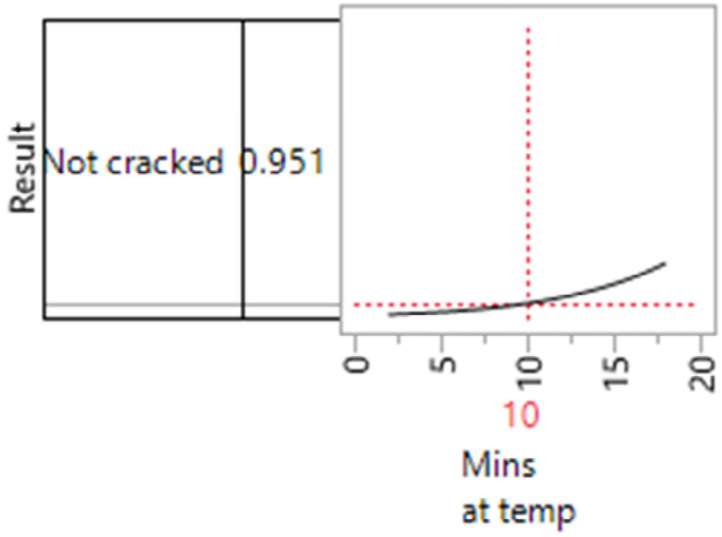
Analysis output

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Mins at temp	1	1	41.5372498	<.0001*

Very strong evidence of positive correlation between dwell time and probability of cracking

Prediction Profiler



Dwell time (mins)	Probability of cracking
5	0.020
10	0.049
15	0.114

Notes

This page intentionally left blank

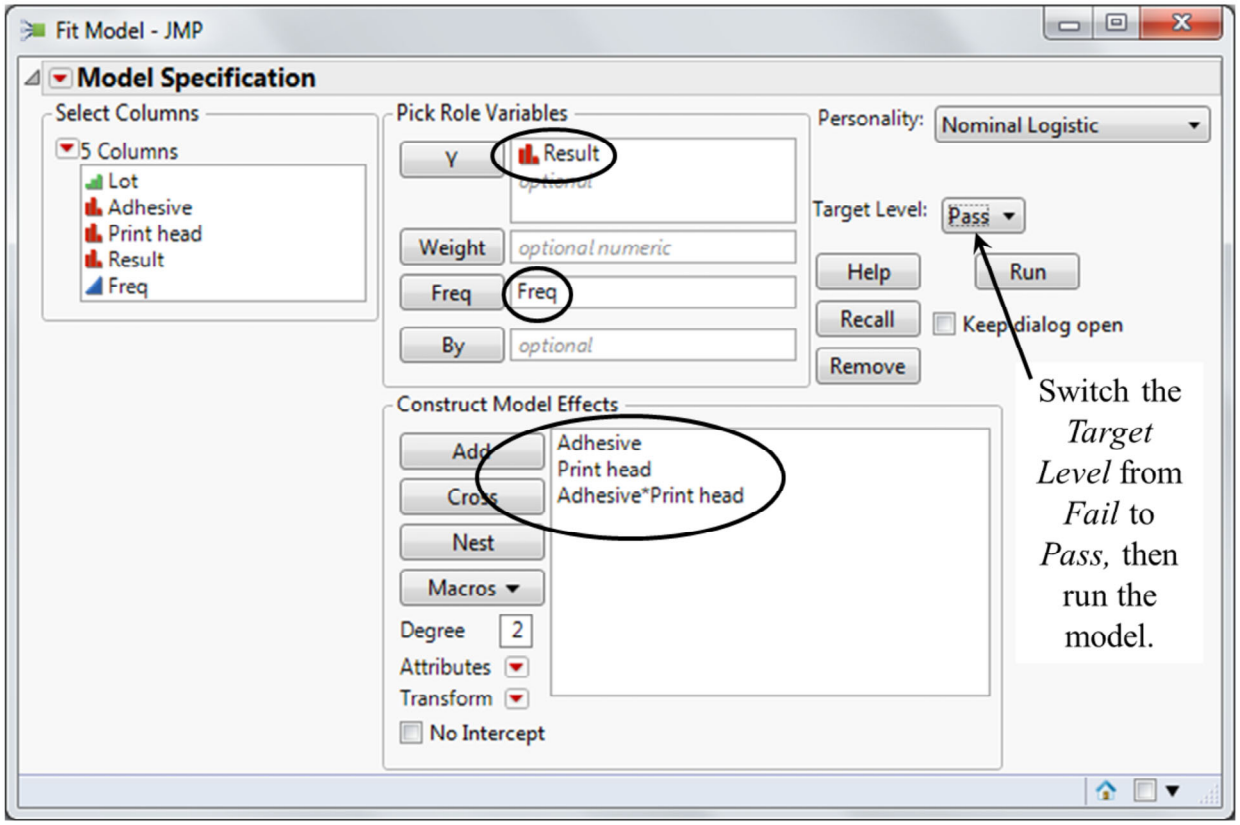
8 Multiple Regression with Pass/Fail Y

- Project to reduce clogged nozzles in print heads
- Comparison of four types of adhesive and two print head designs
- Each lot = 60 print cartridges
- "Pass" = no customer detectable print defects
- *Data sets \ clogging pass-fail*
- Run the *Model* script. If necessary, bring the *Model Specification* to the front.

5/0 Cols / 32/0	Lot	Adhesive	Print head	Result	Freq
1	1	A4	D2	Fail	2
2	1	A4	D2	Pass	58
3	2	A4	D1	Fail	1
4	2	A4	D1	Pass	59
5	3	A2	D2	Fail	13
6	3	A2	D2	Pass	47
7	4	A1	D2	Fail	11
8	4	A1	D2	Pass	49
9	5	A3	D2	Fail	4
10	5	A3	D2	Pass	56
11	6	A4	D1	Fail	5
12	6	A4	D1	Pass	55
13	7	A1	D2	Fail	8
14	7	A1	D2	Pass	52
15	8	A2	D1	Fail	3
16	8	A2	D1	Pass	57
17	9	A3	D2	Fail	1
18	9	A3	D2	Pass	59
19	10	A2	D2	Fail	13
20	10	A2	D2	Pass	47
21	11	A2	D1	Fail	1
22	11	A2	D1	Pass	59
23	12	A1	D1	Fail	1
24	12	A1	D1	Pass	59
25	13	A3	D1	Fail	7

Notes

Example (cont'd)



Notes

Example (cont'd)

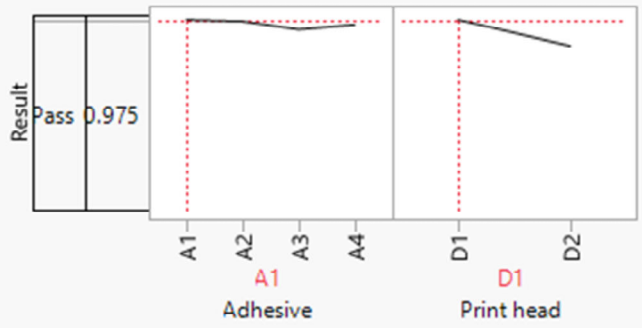
Effect Summary

Source	LogWorth	PValue
Adhesive*Print head	3.721	0.00019
Print head	2.254	0.00557 ^
Adhesive	0.410	0.38926 ^

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Adhesive	3	3	3.01536048	0.3893
Print head	1	1	7.68556658	0.0056*
Adhesive*Print head	3	3	19.7623242	0.0002*

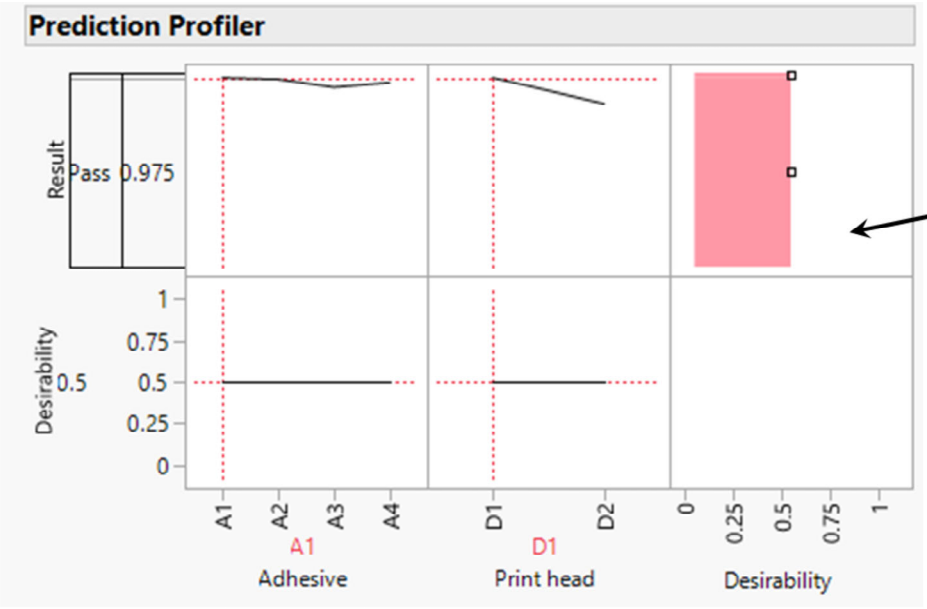
Prediction Profiler



- The *Adhesive* factor was insignificant, but we left it in the model to preserve model hierarchy (Adhesive*Print head is significant)
- On the *Prediction Profiler* red triangle select *Optimization and Desirability* → *Desirability Functions*
- See next slide

Notes

Example (cont'd)

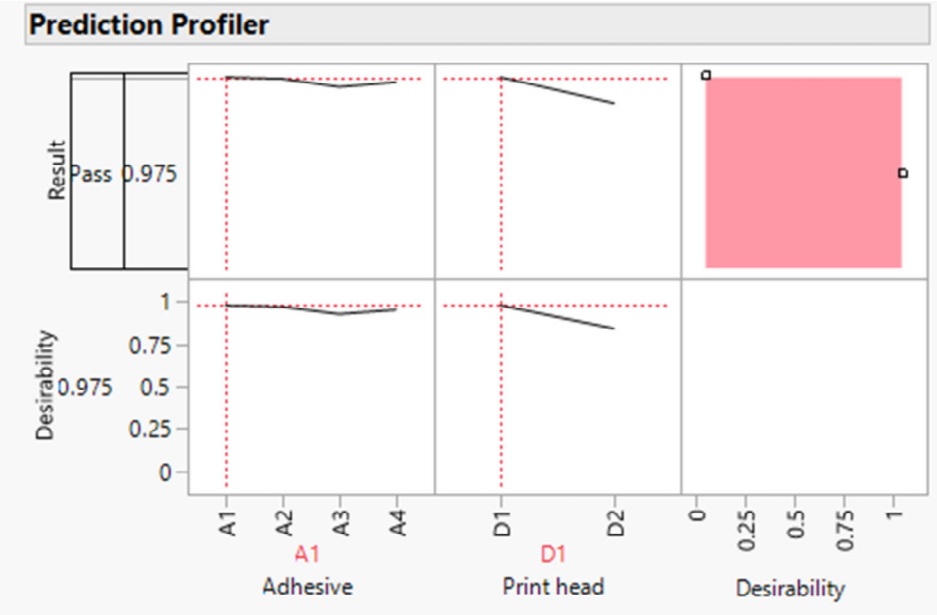


- Double-click in the blank area
- Enter 1 for *Pass* and 0 for *Fail* → OK → OK

Notes

Example (cont'd)

- *Prediction Profiler* red triangle → *Optimization and Desirability* → *Maximize Desirability*
- The failure rate predicted from the optimization was 0.025 or 2.5% (current state failure rate was 20% or more)
- Best combination was D1 with A1



Notes

Exercise 8.1

A Black Belt wants to minimize the occurrence of bubbles and ripples in the urethane coating on truck nameplates. The X variables and ranges are:

- Badge temp: 20 to 40
- Mixing ratio: 92.6 to 94.6
- Curing temp: 30 to 55

- *Data sets \ urethane coating pass-fail*
- Run the *Model* script in the left panel. In the *Model Specification*, switch the *Target Level* from *Fail* to *Pass*, then run the model.
- Remove insignificant terms from the *Effect Summary* (P > 0.15).
- Use the *Prediction Profiler* to find a factor combination that maximizes the yield.
- The current state yield was about 95%. What is the predicted yield for the improved process?

Notes