Six Sigma – Week 11

John Fico, Adjunct Professor
Fall 2020

Agenda: Week 11

- Syllabus check – progress status/timeline
  - Remaining tools to cover in class
    - Week 11: Regression
    - Week 12: Process capability
    - Week 13: Control charts
    - Weeks 14-15:
      - Stakeholder analysis
      - Communication plan
      - Kano model
      - Obtaining customer data
  - Remaining deliverables (assignments)
    - Case Study 2
    - Out of class assignments based on remaining topics
    - Final Team Project – team presentations during final exams week
- Review of LS Text Section 48 (Simple Linear and Multiple Regression)
  - Operations and transactional examples
- Review of LSSM Text
  - Tool application – examples using Minitab
  - Video of residuals – posted in Canvas Discussions (Khan Academy)

LS Text: $Y = f(X_1......X_n)$

X Data

Y Data	Attribute	Continuous
Attribute	Chi-Square	Logistic Regression
Continuous	ANOVA Means/ Medians Tests	Regression

Simple linear regression:
One continuous Y, one continuous X

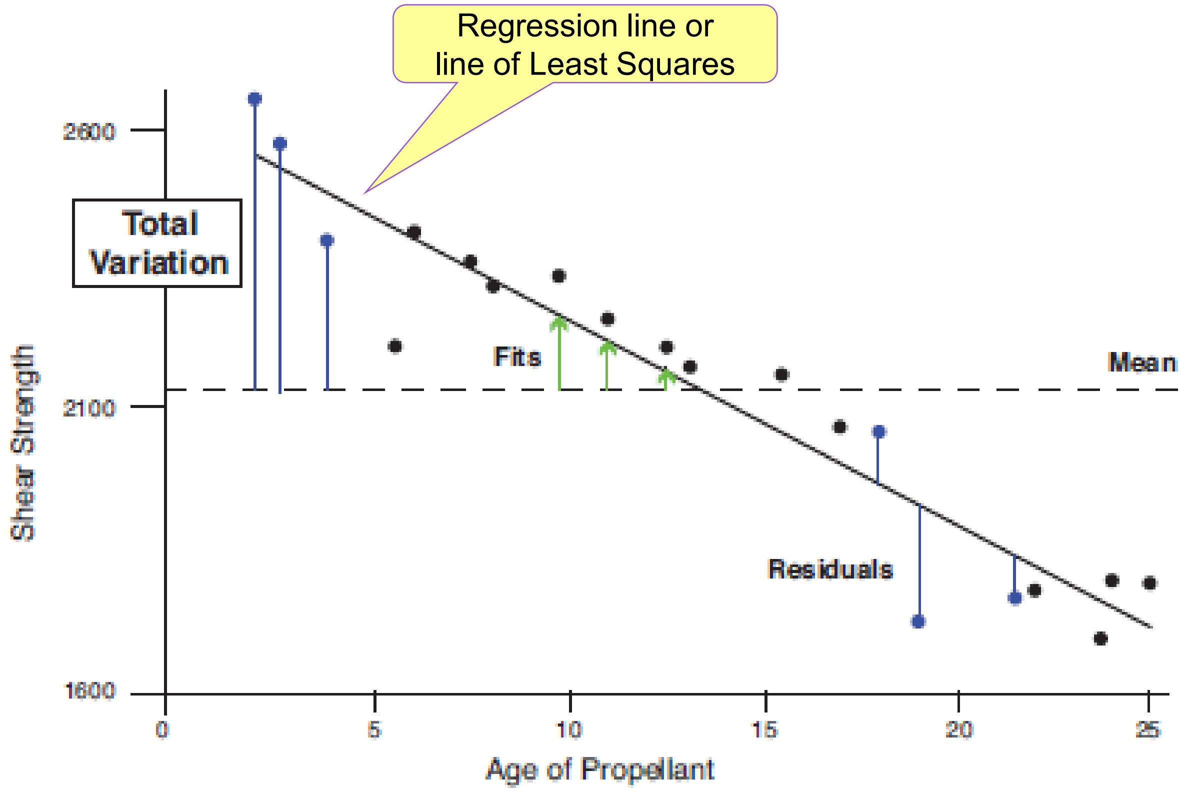
Multiple linear regression:
One continuous Y, more than 1 continuous X

Understanding the statistical significance of relationship between the Y and X(s) is key to finding root causes of problems in Six Sigma projects. Tool use depends on type of data we have.

Highlights of Simple and Multiple Regression

- Regression is a powerful tool and is broad in scope
- Regression has significant application in Six Sigma and finding root causes in business problems
- Generally, 30 or more data points are required for the X and corresponding value of Y at that point
- Regression outputs
  - Fitted line plot – graphical representation of statistical relationship of X to Y
  - Mathematical formula uses “Least Squares” method, which minimizes the total squares of the distances from the regression line

Simple Linear Regression: LS Text pages 464-465



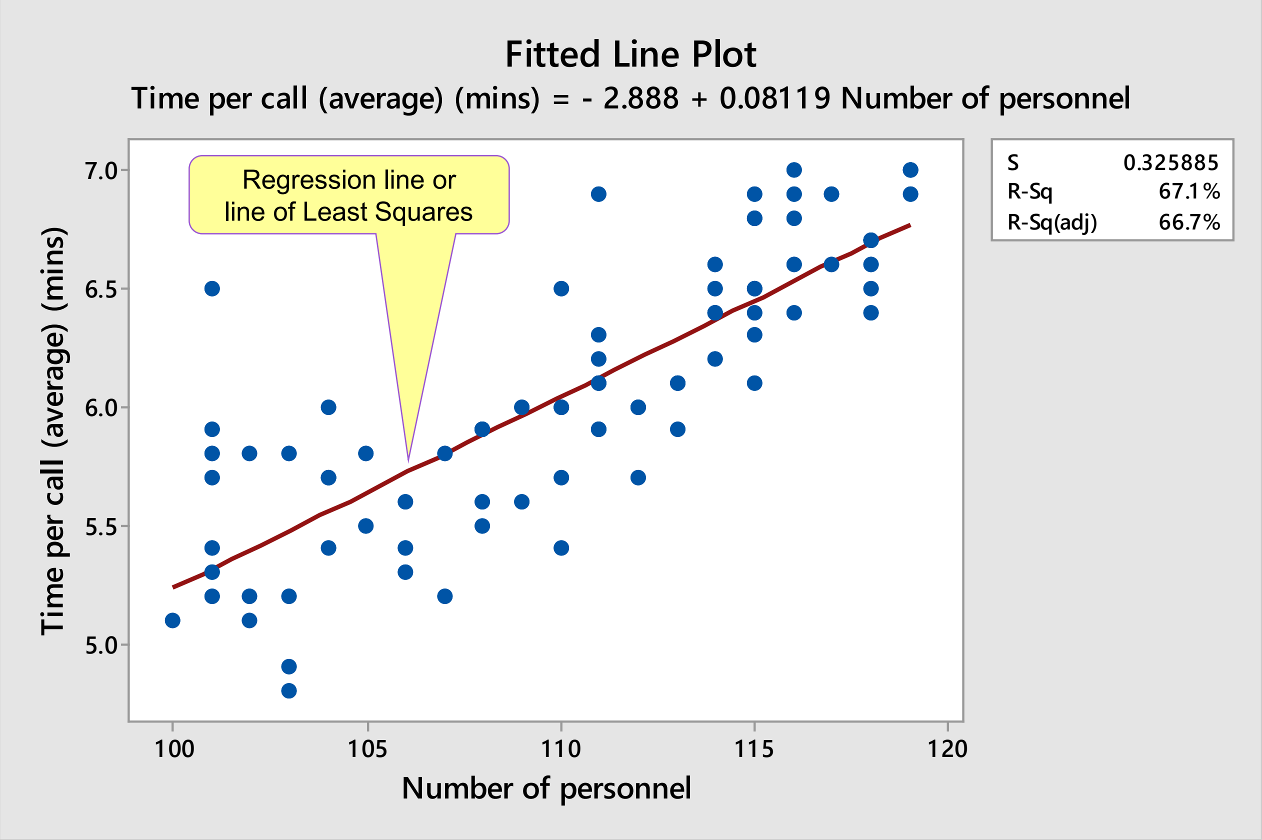
Variation explained by the model is the *Regression.* Variation not explained by the model is the *Residual Error.*

<u>Regression</u>
Calculated by taking the square of the distance from where the line predicts a point *should be* (the Fits) from the mean for every data point, then summing all of the squares.

<u>Residual Error</u>
Calculated by taking the square of the distance of each data point from the Regression line and then summing all of the squares. This is what is “left over” after the line has been fitted.

Example: A technical support call center wants to determine if the number of personnel has an effect on the time per call.

LSSM Text: Simple Linear Regression



Regression Analysis:
Time per call (average) (mins) versus Number of personnel

```
Regression Equation
Time per call (mins) = -2.888 + 0.08119 Number of personnel

S = 0.3259    R-Sq = 67.3%    R-Sq(adj) = 66.70%
```

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	16.914	16.9143	159.27	0.000
Number of personnel	1	16.914	16.9143	159.27	0.000
Error	78	8.284	0.1062		
Lack-of-Fit	18	3.072	0.1707	1.97	0.027
Pure Error	60	5.211	0.0869		
Total	79	25.198			

Interpreting p-values (95% confidence)

- If p>0.05, personnel does not influence time per call (H_0)
- If p<0.05, personnel *does* influence time per call (H_a)

Interpreting regression equation constants

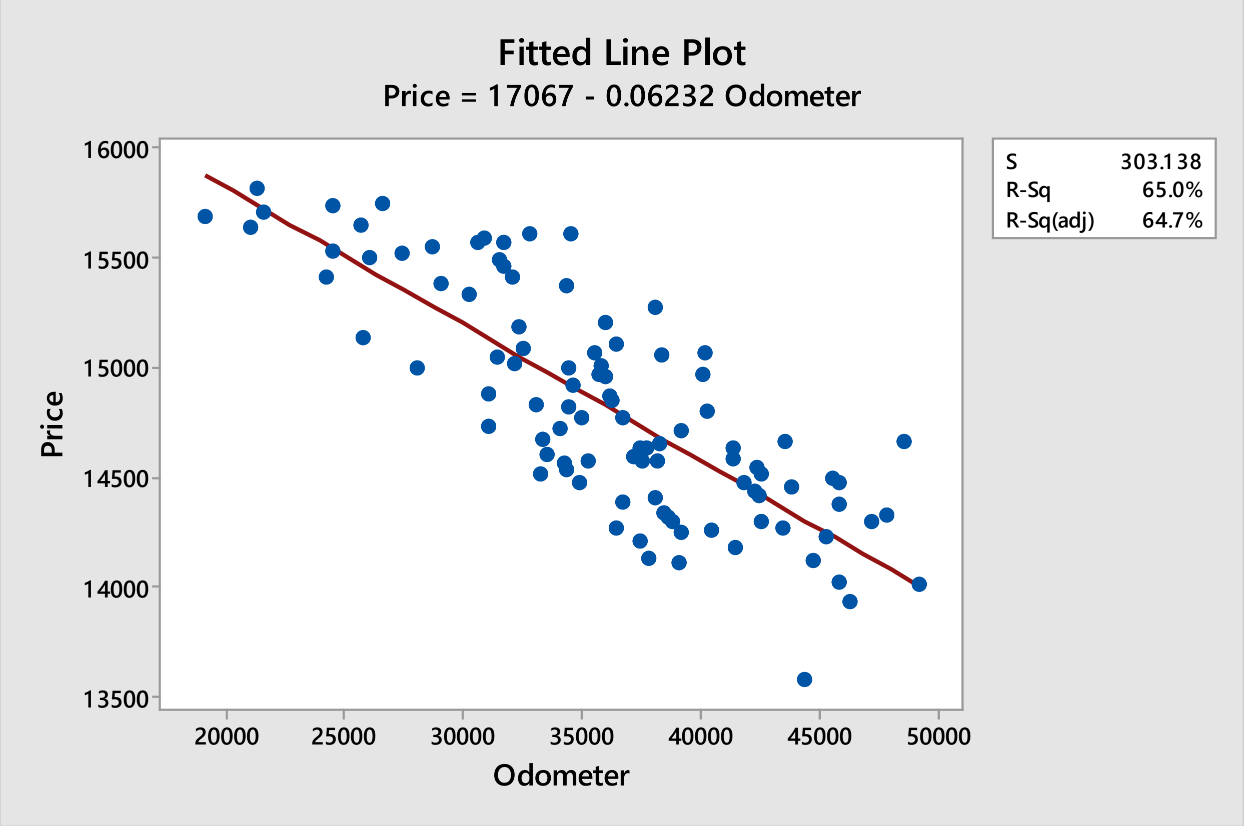
For increase of one person in no. of personnel, we have an avg. increase in time per call of 0.081 minutes, or 4.8 sec.

Interpreting R-Squared (Coefficient of Determination)

67.3% of variation in time per call is explained by the number of personnel. Remaining 32.7% is unexplained.

Example: A technical support call center wants to determine if the number of personnel has an effect on the time per call.

Simple Linear Regression



Regression Analysis: Price versus Odometer

```
The regression equation is
Price = 17067 - 0.06232 Odometer

S = 303.138   R-Sq = 65.0%   R-Sq(adj) = 64.7%

Analysis of Variance

Source      DF        SS        MS       F      P
Regression   1  16734111  16734111  182.11  0.000
Error       98   9005450     91892
Total       99  25739561
```

Fitted Line: Price versus Odometer

Interpreting p-values (95% confidence)

- If p>0.05, odometer reading does not influence sell price (H_0)
- If p<0.05, odometer reading *does* influence sell price (H_a)

Interpreting regression equation constants

For each additional mile on odometer, sell price decreases by $0.0623 (6.23 cents)

Interpreting R-Squared (Coefficient of Determination)

64.7% of variation in sell price is explained by variation in odometer reading. Remaining 35.3% is unexplained.

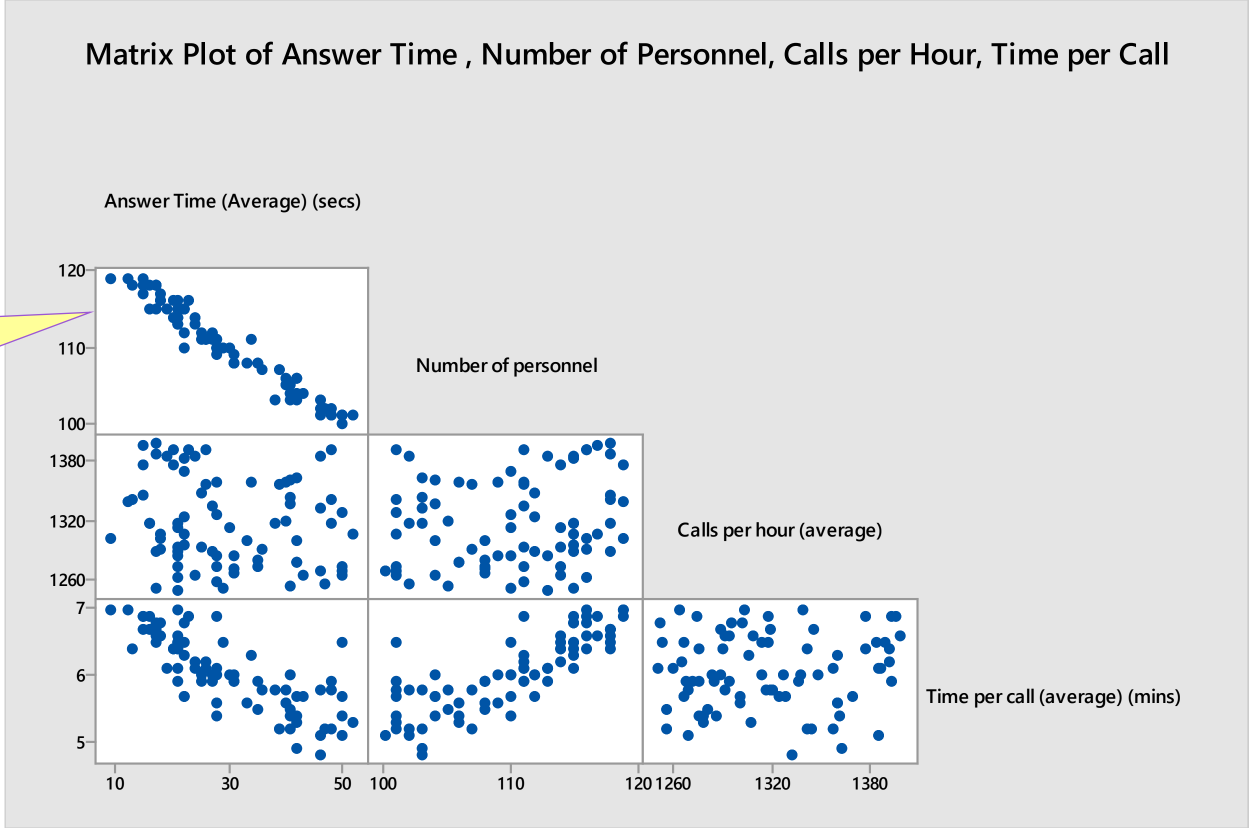
Example: A car dealer wants to find the regression line (relationship between sell price & odometer reading) based on selling 100 three-year old Ford Tauruses at an auction during a one-month period. Prices are based on 2004 data.

Multiple Regression Analysis

- Purpose: Identify the critical Xs (inputs having significant effect on process) and mathematically model their relationship with the process output
  - Y = f(x1, x2, x3……)
  - Example: answer time to call center (Y) depends on several factors (no. of personnel (X1), calls per hour (X2), time per call (X3).
  - If values of factors are set, then repeatedly observe the “Y”
  - “Y” output will not always be the same with repeated observations of cycles

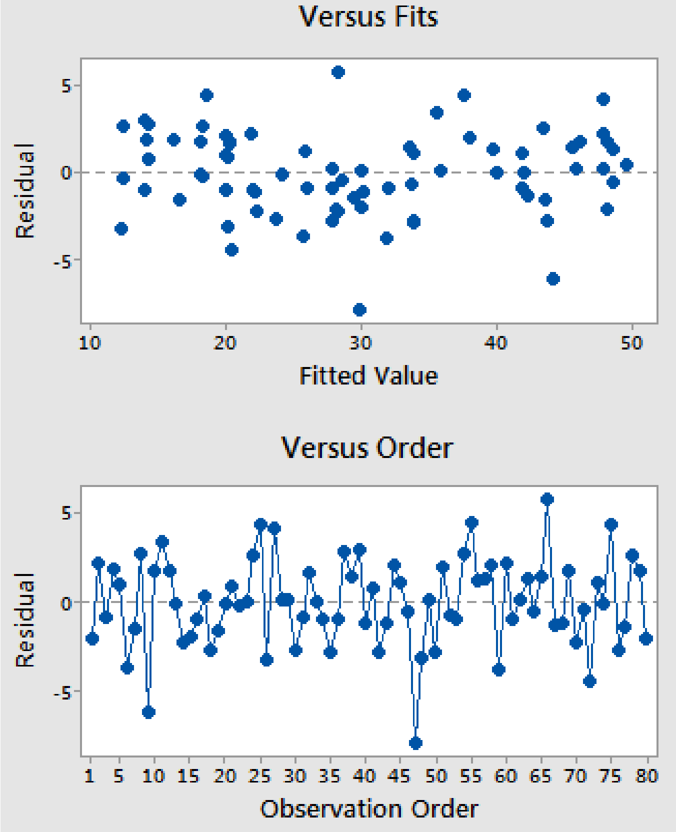
Text 2: Multiple Linear Regression (Matrix Plot): Transactional

Number of personnel appears to have an effect



In a call center, matrix plot can illustrate the effects of number of personnel, avg. calls per hour, and time per call have on answer time.

LSSM Text: Multiple Linear Regression: Transactional



Residual errors are normally distributed and random.
Good model fit

```
Regression Equation
Answer Time (Average) (secs) = 246.91 + 0.00274 Calls per hour (average)
                               - 2.0470 Number of personnel
                               + 0.768 Time per call (average) (mins)
```

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	246.91	9.44	26.16	0.000	
Calls per hour (average)	0.00274	0.00661	0.41	0.680	1.06
Number of personnel	-2.0470	0.0889	-23.02	0.000	3.22
Time per call (average) (mins)	0.768	0.884	0.87	0.388	3.12

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.51002	95.46%	95.28%	94.98%

Number of personnel has significant effect on answer time.
Calls per hour and time per call have little effect (p-values).
Answer time decreases by 2 seconds for each extra person

<u>Interpreting R-Squared (Coefficient of Determination)</u>
95.4% of variation in answer time is explained by variation in input variables. Remaining 4.6% is unexplained.

In reality, the output of a process rarely has a simple relationship with just one input. Several factors likely influence the output.

Text 2: Multiple Linear Regression: Best Subsets

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
2.49113  95.41%     95.35%      95.19%

Regression Equation

Answer Time (Average) (secs) = 247.88 -
 1.9808 Number of personnel
```

FLORIDA POLYTECHNIC
UNIVERSITY

Best Subsets Regression: Answer Time versus Number of pe, Calls per ho, ...

Response is Answer Time (Average) (secs)

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	Number of personnel	Calls per hour (average)	Time per call (average) (mins)
1	95.4	95.4	95.2	0.8	2.4911	X		
1	62.2	61.8	60.6	556.5	7.1477			X
2	95.5	95.3	95.1	2.2	2.4965	X		X
2	95.4	95.3	95.1	2.8	2.5060	X	X	
3	95.5	95.3	95.0	4.0	2.5100	X	X	X

Number of Personnel is best predictor of Answer Time, as seen from Matrix Plot

LSSM Text: Binary Logistic Regression: Transactional

- Output of a process is measured in attribute data such as Pass/Fail or Yes/No
- Example:
  - Quality Approval status of a call center and the average Wrap-up Time (per call) are critical inputs in generating Customer Satisfaction (process output) across a number of call centers.
  - Customer satisfaction has been measured in a survey as Low or High.
  - Analyze data from Minitab worksheet (data in columns representing variables).

LSSM Text: Binary Logistic Regression: Transactional

Variable	Value	Count	
Customer Satisfaction	Low	70	(Event)
	High	22	
	Total	92	

70 *Low* results out of 92 customers. *Low* defined as reference event

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	2	7.574	3.787	7.57	0.023
Average Wrap Up Time (seconds)	1	4.629	4.629	4.63	0.031
Quality Approved	1	4.737	4.737	4.74	0.030
Error	89	93.640	1.052		
Total	91	101.214			

Wrap-up Time & *Quality Approved* status have significant effect on *Customer Satisfaction*

Coefficients

Term	Coef	SE Coef	VIF
Constant	-1.99	1.68	
Average Wrap Up Time (seconds)	0.0250	0.0123	1.12
Quality Approved			
Yes	-1.193	0.553	1.12

Coef. Of 0.025 indicates as wrap-up times increase, chances of lower *Customer Satisfaction* increase
Coef of -1.193 indicates call centers that are Quality Approved tend to have higher *Customer Satisfaction.* Coef is negative – as *Quality Approved* changes from No to Yes, *Customer Satisfaction* tends to move away from the reference (Low to High)

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
Average Wrap Up Time (seconds)	1.0253	(1.0010, 1.0503)

Odds Ratios for Categorical Predictors

Level A	Level B	Odds Ratio	95% CI
Quality Approved			
Yes	No	0.3033	(0.1026, 0.8966)

Odds ratio for level A relative to level B

Odds ratio of 0.30 indicates odds of a *Quality Approved* call center *having Low Customer Satisfaction* are 30% of the odds of the odds of a *Non-Quality Approved* call center having *Low Customer Satisfaction*

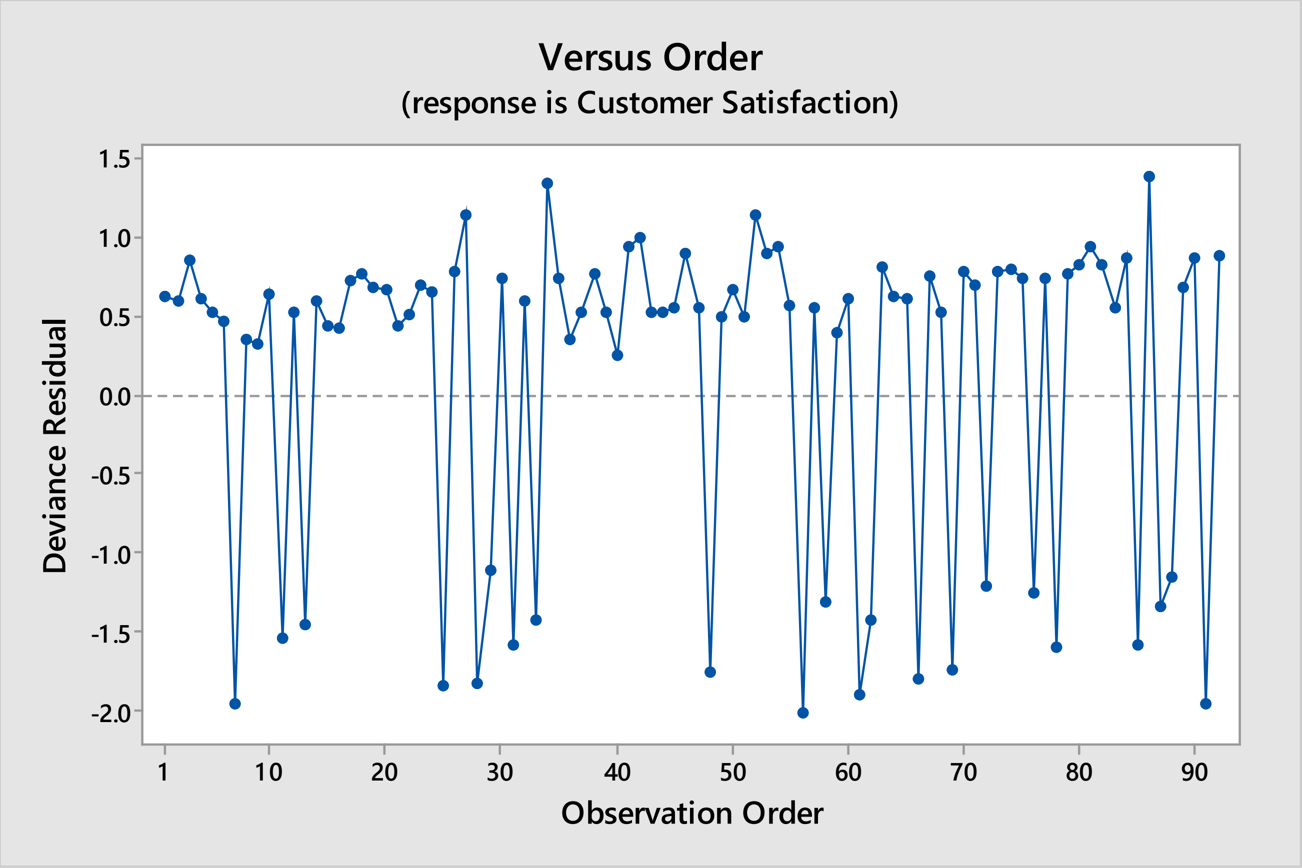
Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	89	93.64	0.348
Pearson	89	88.63	0.491
Hosmer-Lemeshow	8	4.75	0.784

Used to check if model is a reasonable fit for data. We do not want the tests to reject the model (as a good fit) so we are looking for p-values >0.05, which we have in this model.

LSSM Text: Binary Logistic Regression: Transactional (Residuals)

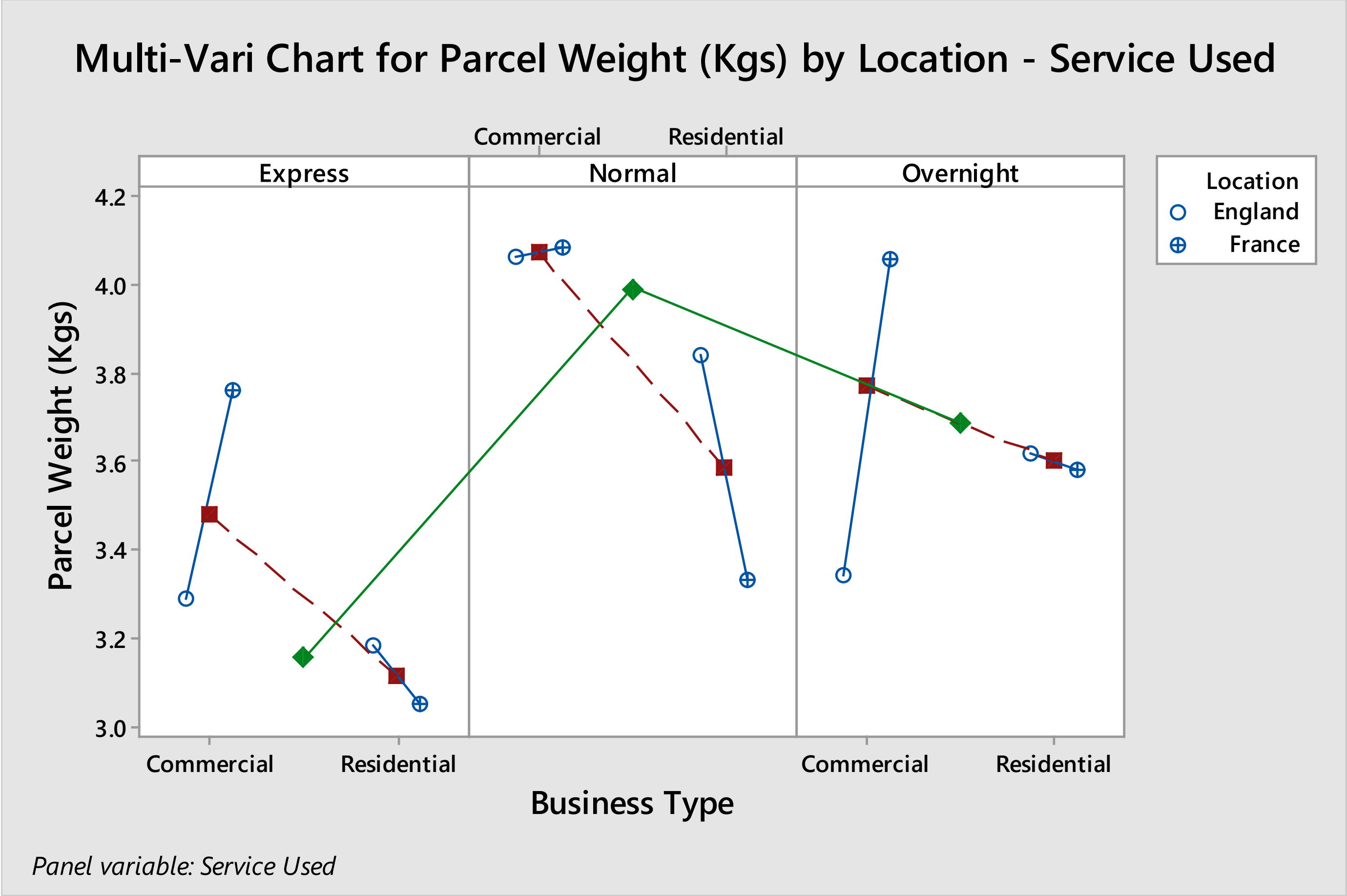
Residuals are random over time – indicating a good model



Text 2: Multi-Vari Charts

- Useful for an initial look at data that has been stratified by several different factors. Box Plots and Individual Value Plots can then focus on specific factors in more detail.
- Example:
  - Data from parcel weights from a courier process.
  - Categorical information on the Location, Business and Service was collected along with the weight of each parcel.
  - This can be used for further understanding of process.
  - Analyze data from Minitab worksheet (data in columns representing variables).

LSSM Text: Mult-Vari Charts: Transactional (Logistics example)



On average:

- *Residential* parcel weights are lower than for *Commercial*
- *Express* parcel weights are lowest, and *Normal* highest
- *French* parcel weights are higher than *English* for *Commercial* parcels, but lower for *Residential* parcels