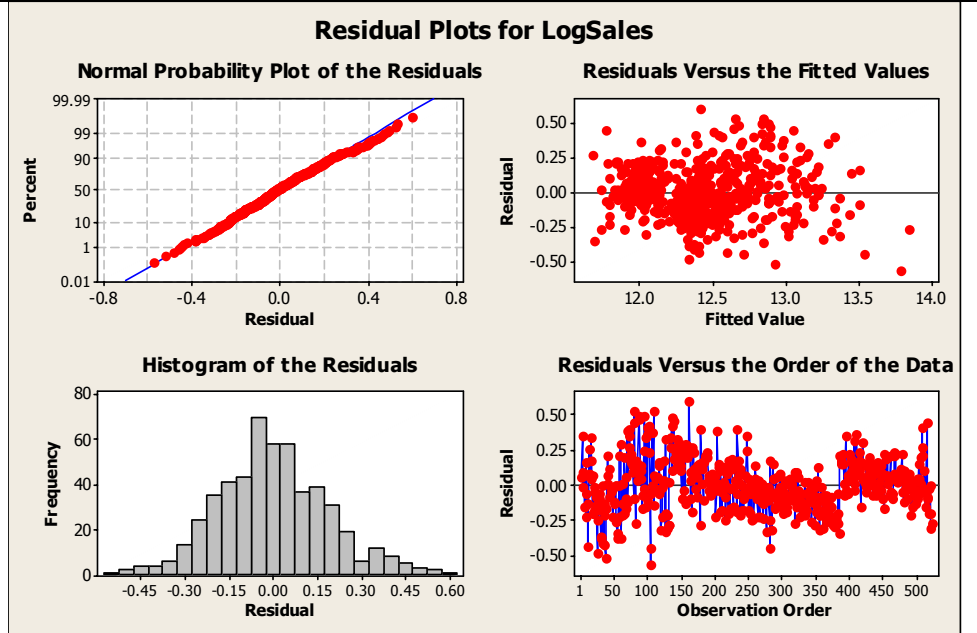


EXAMPLE OF “BEST SUBSETS” REGRESSION (Using Minitab)

Best Subsets Regression: LogSales versus SqFt/100, AC, ...

Response is LogSales

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	S	B	A	C	S	L	H	G	A	Q
1	70.5	70.4	285.5	0.23472	X									
1	62.0	61.9	517.3	0.26644										X
2	78.5	78.4	69.5	0.20056	X									X
2	73.6	73.5	203.7	0.22235	X						X			
3	79.7	79.6	39.3	0.19515	X		X							X
3	79.5	79.3	45.5	0.19624	X					X				X
4	80.5	80.3	19.7	0.19149	X			X		X				X
4	80.2	80.1	26.8	0.19276	X		X	X						X
5	80.9	80.7	9.3	0.18942	X		X	X		X				X
5	80.7	80.5	15.3	0.19051	X	X		X		X				X
<b>6</b>	<b>81.1</b>	<b>80.9</b>	<b>6.4</b>	<b>0.18871</b>	<b>X</b>	<b>X</b>		<b>X</b>	<b>X</b>	<b>X</b>				<b>X</b>
6	81.0	80.8	9.5	0.18928	X		X	X		X	X			X
<b>7</b>	<b>81.2</b>	<b>80.9</b>	<b>6.8</b>	<b>0.18861</b>	<b>X</b>	<b>X</b>		<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>			<b>X</b>
7	81.1	80.9	7.8	0.18879	X	X		X	X	X	X			X
<b>8</b>	<b>81.2</b>	<b>80.9</b>	<b>8.3</b>	<b>0.18869</b>	<b>X</b>	<b>X</b>		<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>		<b>X</b>
8	81.2	80.9	8.5	0.18873	X	X	X	X	X	X	X			X
<b>9</b>	<b>81.2</b>	<b>80.9</b>	<b>10.0</b>	<b>0.18882</b>	<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>



Above are the diagnostic plots for the model chosen, which is the one shown in bold on the left. The “residuals versus order of the data” plot isn’t useful in this example, but the other three plots are. See note #3 below.

NOTES:

1. All of the highlighted models have acceptable Mallows’s Cp. I chose the model (in bold) with good Cp and smallest number of variables to get best R-Sq(adj), which stays the same for the rest of the models, at 80.9%.
2. That model has the variables in bold as predictors. They include SqFt/100, AC, Bathrooms, Lot size, Garage size and Quality. Bedrooms, near highway and pool are not included.
3. The diagnostic plots for the chosen model are shown on the right. They look good. The normal probability plot and the histogram of residuals show that the residuals are approximately normal, and the plot of residuals versus fitted values looks like random scatter, as it should.
4. The final model is:

$$\text{LogSales} = 11.9 + 0.0283 \text{ SqFt/100} + 0.0552 \text{ AC} + 0.0418 \text{ Bathrooms} + 0.000004 \text{ LotSize} + 0.0643 \text{ GarageSize} - 0.206 \text{ Quality}$$

## EXAMPLE OF STEPWISE REGRESSION (Using Stata)

Forward selection:

**stepwise, pe(.2): regress LnPrice SqFtHdrc AC Bedrooms Bathrooms LotSize Hwy Garage Pool Quality**

```

begin with empty model
p = 0.0000 < 0.2000 adding SqFtHdrc
p = 0.0000 < 0.2000 adding Quality
p = 0.0000 < 0.2000 adding LotSize
p = 0.0000 < 0.2000 adding Garage
p = 0.0005 < 0.2000 adding Bathrooms
p = 0.0276 < 0.2000 adding AC
    
```

Source	SS	df	MS			
Model	78.7422093	6	13.1237015	Number of obs =	522	
Residual	18.3407146	515	.035613038	F( 6, 515) =	368.51	
				Prob > F =	0.0000	
				R-squared =	0.8111	
				Adj R-squared =	0.8089	
				Root MSE =	.18871	
Total	97.0829239	521	.186339585			

LnPrice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
SqFtHdrc	.0283202	.001959	14.46	0.000	.0244716	.0321688
Quality	-.2064869	.0203222	-10.16	0.000	-.2464114	-.1665623
LotSize	4.00e-06	7.32e-07	5.46	0.000	2.56e-06	5.44e-06
Garage	.0643159	.0158714	4.05	0.000	.0331351	.0954966
Bathrooms	.0417736	.0126665	3.30	0.001	.0168893	.0666579
AC	.0552223	.0249893	2.21	0.028	.0061289	.1043158
_cons	11.85661	.0877624	135.10	0.000	11.68419	12.02902

. estat ic

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	522	-301.6573	133.2841	7	-252.5682	-222.7645

Backward selection:

**stepwise, pr(.2): regress LnPrice SqFtHdrc AC Bedrooms Bathrooms LotSize Hwy Garage Pool Quality**

```

begin with full model
p = 0.5895 >= 0.2000 removing Bedrooms
p = 0.4668 >= 0.2000 removing Hwy
p = 0.2075 >= 0.2000 removing Pool
    
```

Source	SS	df	MS			
Model	78.7422093	6	13.1237015	Number of obs =	522	
Residual	18.3407146	515	.035613038	F( 6, 515) =	368.51	
				Prob > F =	0.0000	
				R-squared =	0.8111	
				Adj R-squared =	0.8089	
				Root MSE =	.18871	
Total	97.0829239	521	.186339585			

LnPrice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
SqFtHdrc	.0283202	.001959	14.46	0.000	.0244716	.0321688
AC	.0552223	.0249893	2.21	0.028	.0061289	.1043158
Quality	-.2064869	.0203222	-10.16	0.000	-.2464114	-.1665623
Bathrooms	.0417736	.0126665	3.30	0.001	.0168893	.0666579
LotSize	4.00e-06	7.32e-07	5.46	0.000	2.56e-06	5.44e-06
Garage	.0643159	.0158714	4.05	0.000	.0331351	.0954966
_cons	11.85661	.0877624	135.10	0.000	11.68419	12.02902

estat ic

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	522	-301.6573	133.2841	7	-252.5682	-222.7645