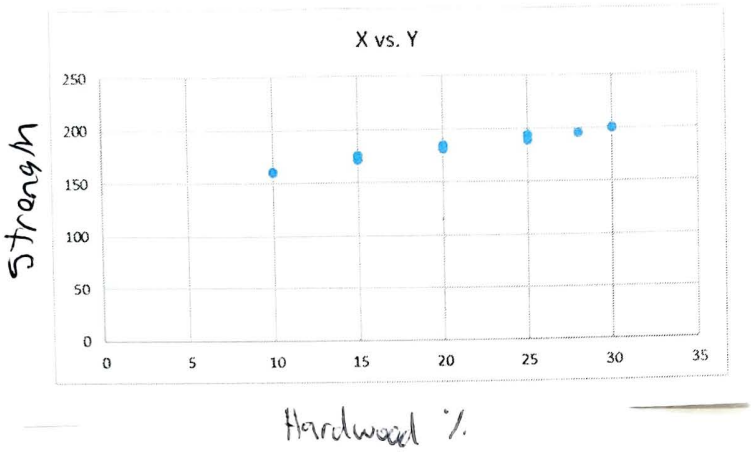
~~HW6~~

PROBLEM 1

1) Problem 4.43

Plot data to get an idea of the relationship:



a) Fit linear regression model:

The data may be able to be fit with a linear regression model: $Y = \beta_0 + \beta_1 x + \epsilon$

The fitted model is given by: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ by:

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{S_{xx}}{S_{xy}}$$

$n = 10 \quad \bar{x} = 20.8 \quad \bar{y} = 182.9$

$\sum_{i=1}^{n} y_i x_i = 38715 \quad \sum_{i=1}^{n} y_i = 1829$

$\sum_{i=1}^{n} x_i = 208 \quad \sum_{i=1}^{n} x_i^2 = 4684$

$$\hat{B}_1 = \frac{38715 - \frac{(1829)(208)}{10}}{4684 - \frac{(208)^2}{10}} = 1.8786$$

$$\hat{B}_0 = 182.9 - (1.8786)(20.8) = 143.82$$

$\therefore \boxed{\hat{y} = 144 + 1.88x}$

b) Test for significance:

Concuct ANOVA

1) Parameter of interest: slope of the linear equation

2) $H_0: \beta_1 = 0$

3) $H_1: \beta_1 \neq 0$

4) Test Statistic: $F_0 = \dfrac{\frac{SS_R}{1}}{\frac{SSE}{n-2}} = \dfrac{MS_R}{MS_E}$

5) Rejection Criteria: Reject if $f_0 > f_{\alpha,1,n-2}$

6) Calculations: $\bar{y} = 182.9$ $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ for all points (see next page)

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = (162.61 - 182.9)^2 + \ldots + (200.18 - 182.9)^2$$

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (160 - 182.9)^2 + \ldots + (200 - 182.9)^2$$

$SS_R = 1262.07$

$SS_T = 1300.90$

$SS_E = SS_T - SS_R = 38.83$

$$F_0 = \left(\frac{1262.067}{38.83/8} \right) = 260$$

$$F_{\alpha,1,n-2} = F_{0.05,1,8} = 5.32$$

7) Conclusions: $F_0 > F_{crit}$ so reject null and conclude that β_1 is not zero

c) See outputs below, and on next page

β_1 @ CI: $1.610 < \beta_1 < 2.147$

CI: $\boxed{1.61 < \beta_1 < 2.15}$

The interval does not cross zero, so it supports the conclusion in part b.

d) $R^2 = \frac{SS_R}{SS_T} = \frac{1262.067}{1300.9} = 0.97$

97% of the variability in the data is accounted for by the model

xi	yi	y(hat)			
% Hardwoc	Strength	Estimated y	(yi-$\bar{y}$)2	(yhat-$\bar{y}$)2	Residuals
10	160	162.6107383	524.41	411.6541	-2.61074
15	171	172.003915	141.61	118.7247	-1.00391
15	175	172.003915	62.41	118.7247	2.996085
20	182	181.3970917	0.81	2.258733	0.602908
20	184	181.3970917	1.21	2.258733	2.602908
20	181	181.3970917	3.61	2.258733	-0.39709
25	188	190.7902685	26.01	62.25634	-2.79027
25	193	190.7902685	102.01	62.25634	2.209732
28	195	196.4261745	146.41	182.9574	-1.42617
30	200	200.1834452	292.41	298.7175	-0.18345

20.8	182.9
$\bar{x}$	$\bar{y}$

yixi	xi2
1600	100
2565	225
2625	225
3640	400
3680	400
3620	400
4700	625
4825	625
5460	784
6000	900

	1300.9	1262.067	38.83277	260.0004
	SST	SSR	SSE	Fo

SUM yixi	38715
SUM yi	1829
SUM xi	208
n	10
SUM xi2	4684
B1	1.878635
B0	143.8244

Regression Equation

Strength = 143.82 + 1.879 x (Percentage Hardwood)

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	143.82	2.52	(138.01, 149.64)	57.04	0.000	
x (Percentage Hardwood)	1.879	0.117	(1.610, 2.147)	16.12	0.000	1.00

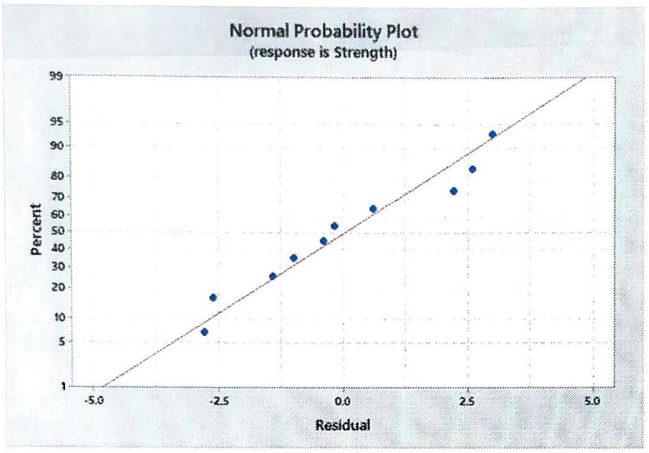
Analysis of Variance

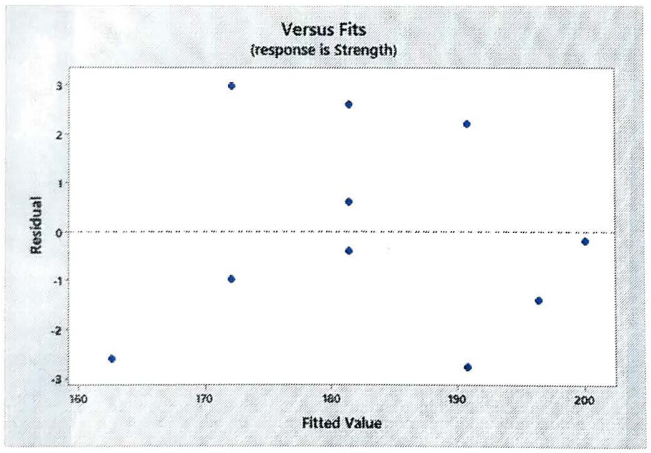
Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1262.07	97.01%	1262.07	1262.07	260.00	0.000
Error	8	38.83	2.99%	38.83	4.85		
Lack-of-Fit	4	13.67	1.05%	13.67	3.42	0.54	0.716
Pure Error	4	25.17	1.93%	25.17	6.29		
Total	9	1300.90	100.00%				

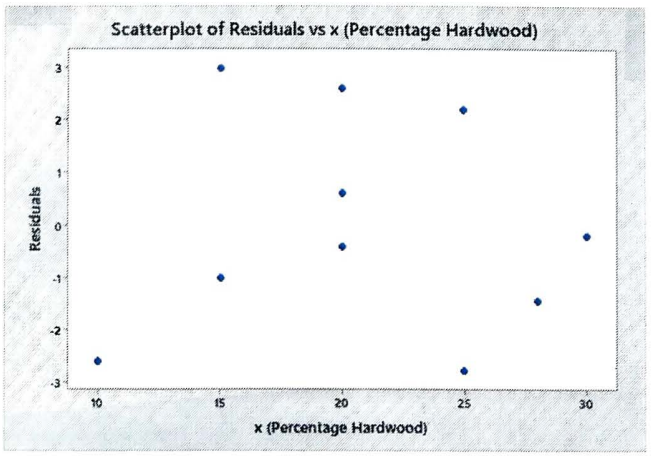
4.45 PROBLEM 2

$e_i = y_i - \hat{y}_i$ (residuals)

As the graphs show below, no patterns emerge and the residuals are normally distributed. The assumptions are valid.



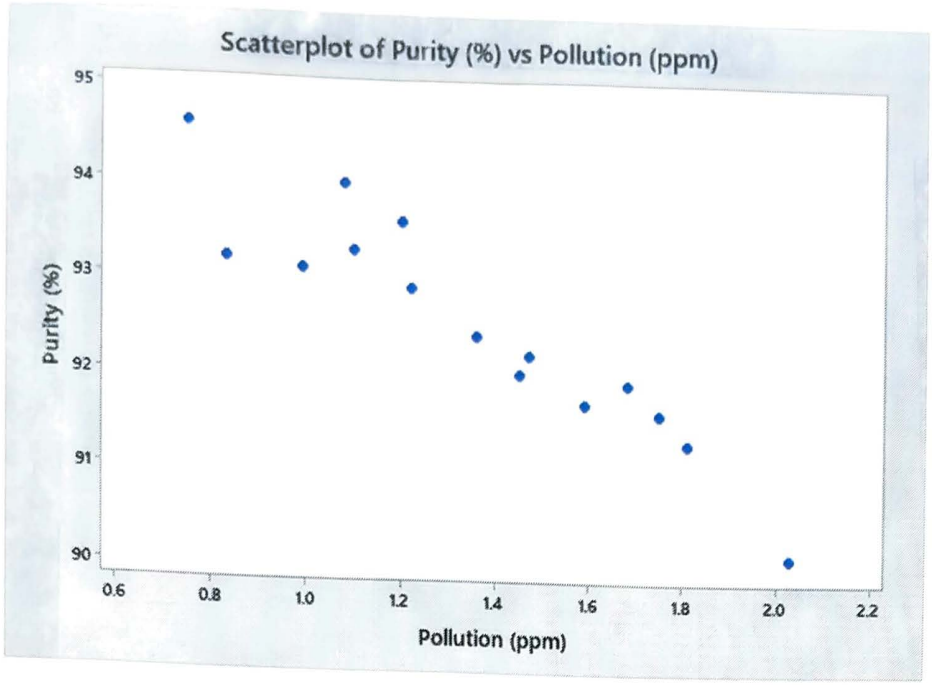




PROBLEM 3

Plotting just the data:

a)



Looks like it could be linear

Regression Equation

Purity (%) = 96.4 - 2.90 Pollution (ppm)

b) ANOVA: $H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	16.491	87.39%	16.491	16.4908	90.13	0.000
Error	13	2.379	12.61%	2.379	0.1830		
Total	14	18.869	100.00%				

The P-value is < 0.05 so we reject the null and conclude that $\beta_1 \neq 0$ Purity and pollution have a significant linear relationship.

The CI is $-3.561 < \beta_1 < -2.241$

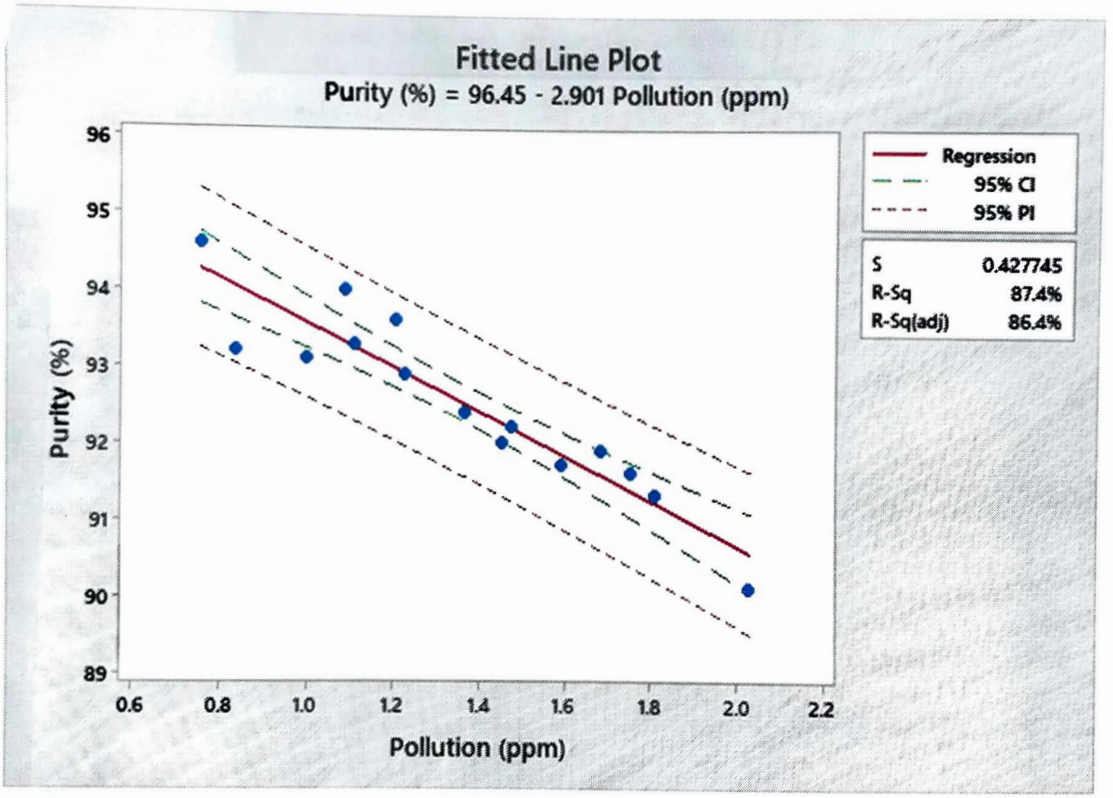
-3.56 < B1 < -2.24 (correct sig figs)

~~[illegible]~~ It doesn't cross 0 so it corroborates the conclusion in part b, We are 95% certain that the true slope lies within this range, or there is a 95% chance that the CI we calculated contains the true slope.

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	96.455	0.428	(95.529, 97.380)	225.24	0.000	
Pollution (ppm)	-2.901	0.306	(-3.561, -2.241)	-9.49	0.000	1.00

d)



CIs: there is 95% chance that the CI we calculated contains the true mean response (purity %) of the population at specified values of pollution count. If you repeat this process many times, you'd expect the confidence interval to capture the true mean purity 95% of the time.

~~response lies within for this specific data set and multiple values of x. The prediction intervals represent the bounds by which we are 95% confident that the true future observation will lie within for multiple values of x. The limits are larger for PIs because the account for the error in the model, and the error associated with future observations. The difference between the two is that CIs are telling you about the response accounting for the error in this specific data set, and the PIs are attempting to predict a future observation and so must account for future error.~~

Regression Equation

Purity (%) = 96.455 - 2.901 Pollution (ppm)

Settings

Variable	Setting
Pollution (ppm)	1

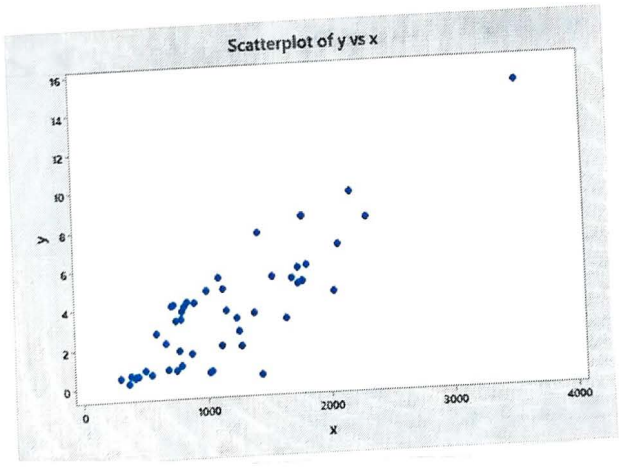
Prediction

Fit	SE Fit	95% CI	95% PI
93.5536	0.154592	(93.2196, 93.8876)	(92.5710, 94.5362)

Predicted Purity (%) @ 1.0 ppm

PROBLEM 4

4): a) The plot does appear as if linear will be OK



Regression Equation

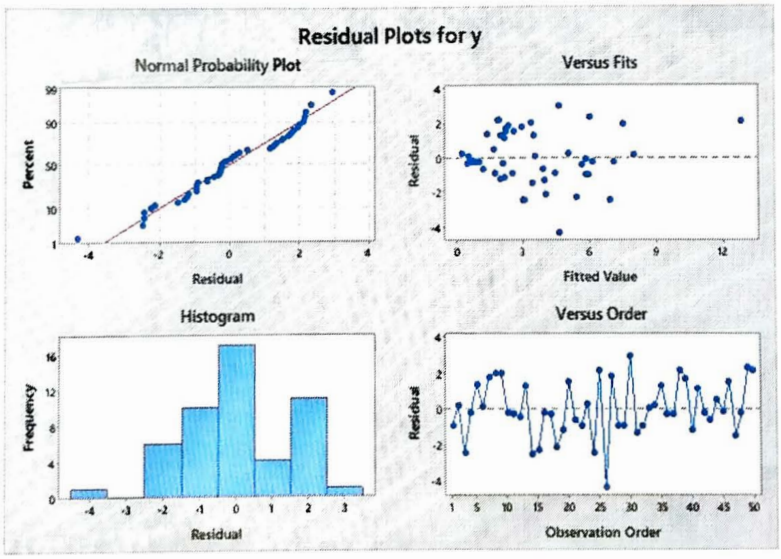
$y = -0.88 + 0.0038\,x$

c) ANOVA: $H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	297.7	71.78%	297.7	297.728	122.11	0.000
Error	48	117.0	28.22%	117.0	2.438		
Total	49	414.8	100.00%				

d) The residuals vs. predicted (fits) show a clear patten. The assumption of equal Variance is not valid.



C shape!

The model seems to fit without violating assumptions

We can conclude for the model

$\hat{B}_0$ $\hat{B}_1$

$\sqrt{y} = 0.60 + 0.00097x$

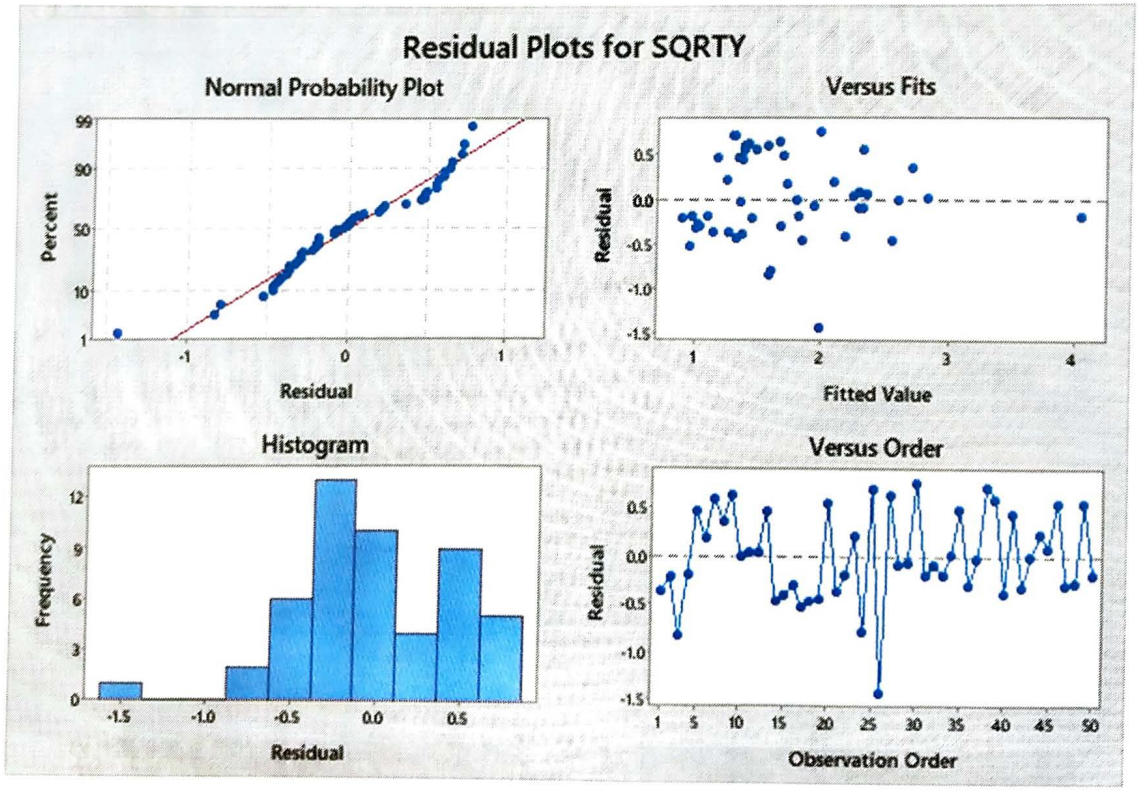
$B_1 \neq 0$ based on ANOVA and validassuptions

Regression Equation

SQRTY = 0.60 + 0.00097 x

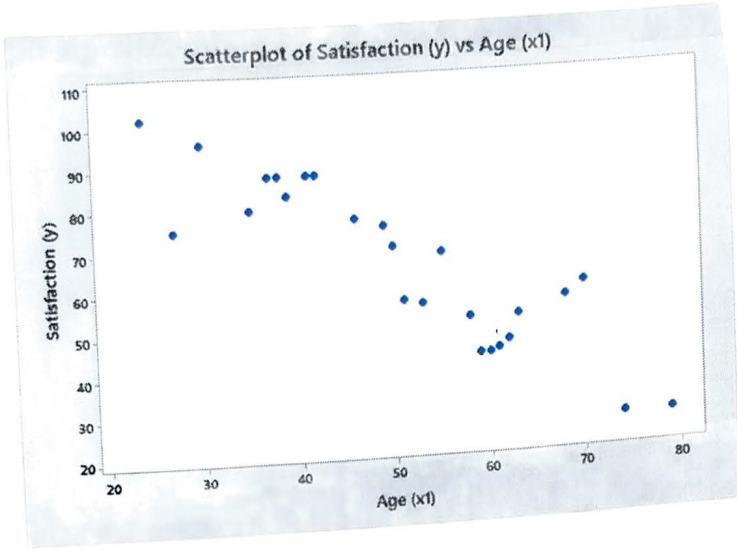
Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	18.98	63.51%	18.98	18.9751	83.56	0.000
Error	48	10.90	36.49%	10.90	0.2271		
Total	49	29.88	100.00%				



PROBLEM 5

a) Always plot out 1st:



Regression Equation

Satisfaction (y) = 130 - 1.3 Age (x1)

b) ANOVA $H_0: \beta_1 = 0$ $\beta_1 \neq 0$
Need to check assumptions but if valid conclude $\beta_1 \neq 0$ because p-value < 0.05

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	8757	81.24%	8757	8756.66	99.63	0.000
Error	23	2022	18.76%	2022	87.89		
Total	24	10778	100.00%				

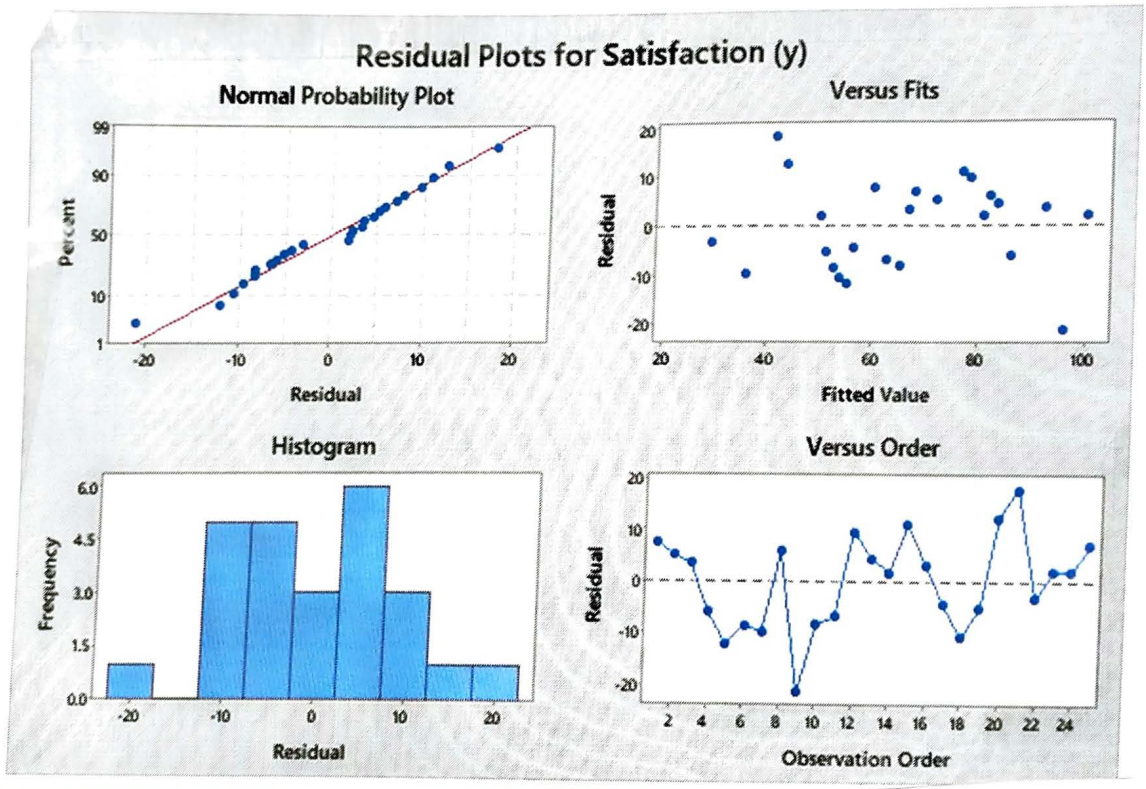
Hypothesis

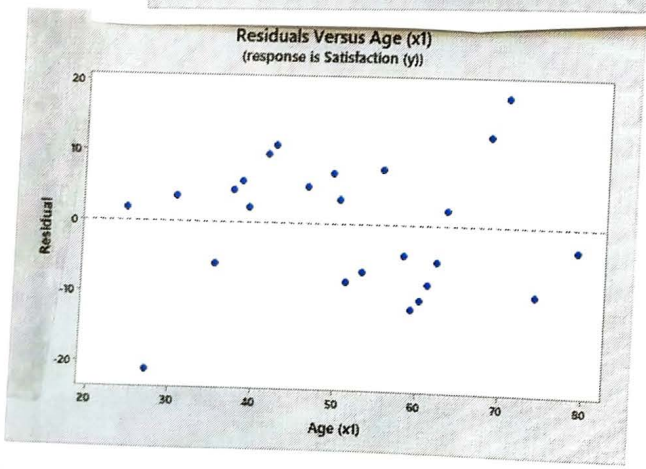
c) 81.2% of the error is from the model (R^2 below) R^2 adjusted for dof is 80.4%.

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
9.37523	81.24%	80.43%	2472.41	77.06%	187.91	190.42

4.50] No apparent patterns emerge and data appear ~ normal. Assumptions appear valid





Make sure you also generate this graph (not automatic)

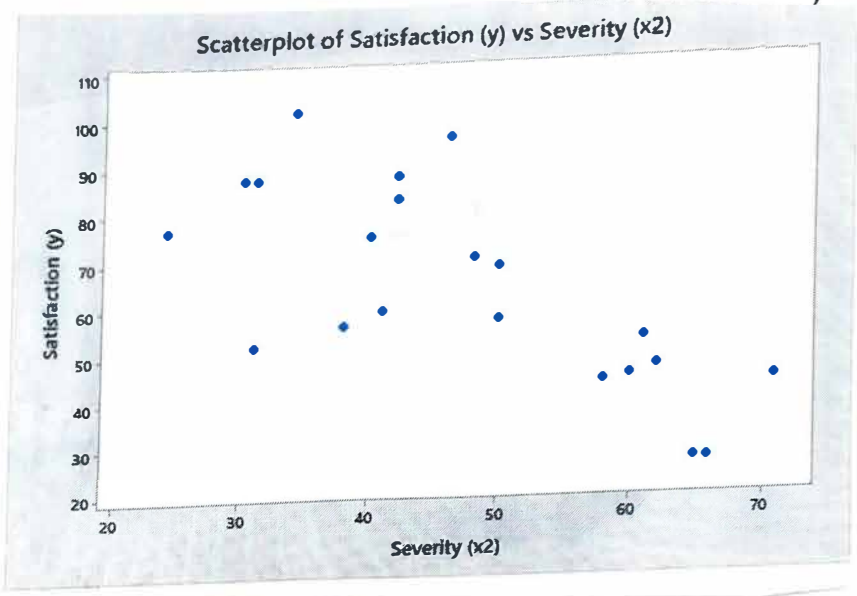
a) If we add severity, we should plot it vs. response. a weaker relationship may exist

$H_0: \beta_1 = \beta_2 = 0$ $H_1: \beta_i \neq 0$ for at least one i

Since p-value < 0.05 at least one $\beta_i \neq 0$ (if assumptions are valid)

Regression Equation

Satisfaction (y) = 140 - 1.0 Age (x1) - 0.55 Severity (x2)



Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	9663.7	89.66%	9663.7	4831.85	95.38	0.000
Age (x1)	1	8756.7	81.24%	4029.4	4029.38	79.54	0.000
Severity (x2)	1	907.0	8.42%	907.0	907.04	17.90	0.000
Error	22	1114.5	10.34%	1114.5	50.66		
Total	24	10778.2	100.00%				

b) Partial ANOVAs show both regressors are significant

Or $\beta_1 \neq 0$ and $\beta_2 \neq 0$

$H_0: \beta_1 = 0$ $H_1: \beta_1 \neq 0$ $H_0: \beta_2 = 0$ $H_1: \beta_2 \neq 0$

Coefficients

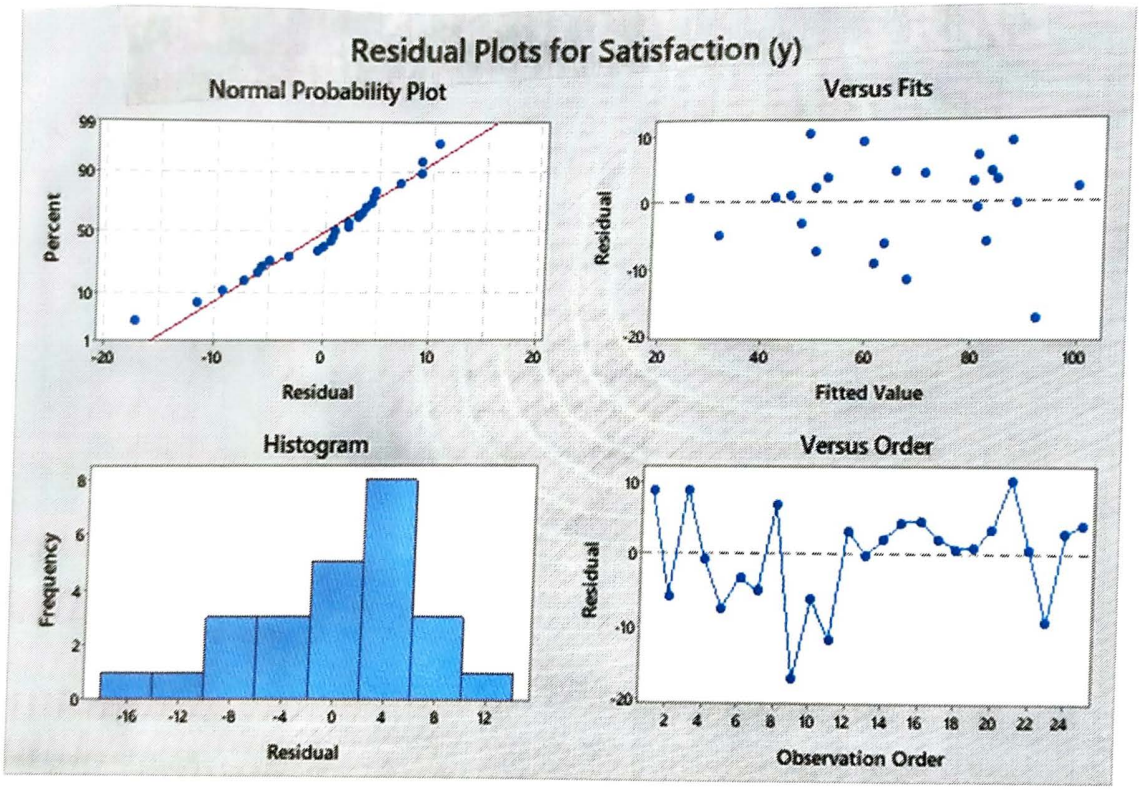
Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	143.47	5.95	(131.12, 155.82)	24.09	0.000	
Age (x1)	-1.031	0.116	(-1.271, -0.791)	-8.92	0.000	1.39
Severity (x2)	-0.556	0.131	(-0.829, -0.284)	-4.23	0.000	1.39

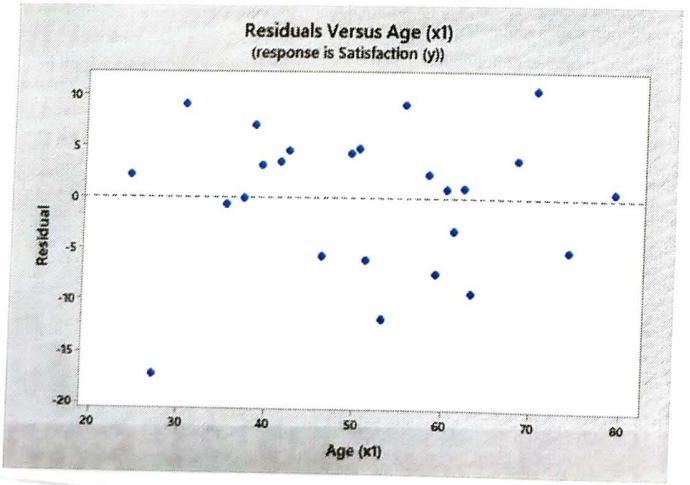
c) Adding the second term increased both R^2 and R^2 adjusted (predicted) so adding 2nd term as increased quality of fit

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
7.11767	89.66%	88.72%	1484.93	86.22%	175.88	178.76

4.52 The model appears adequate





BE SURE TO ALSO GENERATE THIS GRAPH (not automatic)

Your Plots:

This is useful to know because you can now adjust your expectations of satisfaction based on age - or - work on trying to increase satisfaction for elderly patients.

In addition, without including severity, you could come to the wrong conclusions. See below:

40 year old, severity of 30: Satisfaction 80-100

40 " " 60: Satisfaction 60-80

So, if you only based your model in age, you would miss the fact that 40 year olds have varying satisfaction depending on severity.