ECHE313 Homework 7 - Due 03/27/25 — Trevor Swan (tcs94)

4.26 in book

1.

Table 4E.6 Tensile Strength Data for Exercise 4.26

Strength	Percentage Hardwood	Strength	Percentage Hardwood
160	10	181	20
171	15	188	25
175	15	193	25
182	20	195	28
184	20	200	30

a) $n=10$

$\Sigma x_i = 160+171+\dots+200 = 1829$

$\Sigma x_i^2 = 160^2+171^2+\dots+200^2 = 335825$

$\Sigma y_i = 10+15+\dots+30 = 208$

$\Sigma y_i^2 = 10^2+15^2+\dots+30^2 = 4684$

$\Sigma x_i y_i = 160(10)+171(15)+\dots+200(30) = 38715$

x = strength ; y = % Hardwood

$$S_{xy} = \Sigma x_i y_i - \frac{(\Sigma y_i)(\Sigma x_i)}{n} = 38715 - \frac{(208)(1829)}{10} = 671.8$$

$$S_{xx} = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = 335825 - \frac{(1829)^2}{10} = 1300.9$$

$$\tilde{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{671.8}{1300.9} = 0.5164 \;;\quad \bar{x} = \frac{1}{10}(\Sigma x_i) = \frac{1}{10}(1829) = 182.9$$

$$\bar{y} = \frac{1}{10}(\Sigma y_i) = \frac{1}{10}(208) = 20.8$$

$$\beta_0 = \bar{y} - \tilde{\beta}_1 \cdot \bar{x} = 20.8 - 0.5164(182.9) = -73.65$$

$$\hat{y} = 0.516x - 73.7 \quad \text{where} \begin{cases} x = \text{Strength} \\ \hat{y} = \text{Percentage Hardwood} \end{cases}$$

b) ANOVA for linear regression

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

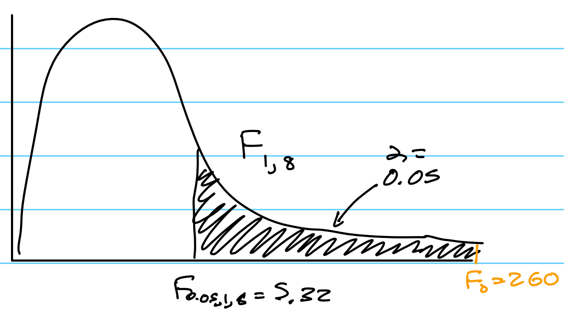
Test Statistic:

$$SS_R = \Sigma(\hat{y}_i - \bar{y})^2 = (8.97-20.8)^2 + (14.65-20.8)^2 + \dots + (29.63-20.8)^2 = 346.93$$

$$SS_E = \Sigma(y_i - \hat{y}_i)^2 = (10-8.97)^2 + \dots + (30-29.63)^2 = 10.67$$

$$F_0 = \frac{SS_R/1}{SS_E/(n-2)} = \frac{346.93}{10.67/(8)} = 260$$

Reject if: $F_0 > F_{0.05,1,8}$



$260 > 5.32 \rightarrow$ Reject H_0

We have sufficient evidence to support the claim that there is a significant linear relationship between percentage hardwood (y) and strength (x).

c)

Regression Equation

Percentage Hardwood = -73.65 + 0.5164 Strength

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-73.65	5.87	(-87.19, -60.12)	-12.55	0.000	
Strength	0.5164	0.0320	(0.4426, 0.5903)	16.12	0.000	1.00

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
1.15513	97.01%	96.64%	17.4802	95.11%	39.03	35.94

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	1	346.93	97.01%	346.93	346.925	260.00	0.000
Strength	1	346.93	97.01%	346.93	346.925	260.00	0.000
Error	8	10.67	2.99%	10.67	1.334		
Total	9	357.60	100.00%				

95% CI on β_1 (slope)
(0.4426, 0.5903)
In repeated sampling, the interval calculated captures the slope β_1 95% of the time.

Also calculations from previous page (excel assisted) match the minitab output

d) $R^2 = \frac{SS_R}{SS_T} = \frac{SS_R}{SS_R + SS_E} = \frac{346.93}{346.93 + 10.67} = 97.0\%$

97% of the variability in the target variable is accounted for by the regression model.

4.28 in book

Data from problem ①

2. **Table 4E.6** Tensile Strength Data for Exercise 4.26

Strength	Percentage Hardwood	Strength	Percentage Hardwood
160	10	181	20
171	15	188	25
175	15	193	25
182	20	195	28
184	20	200	30

x = strength; y = % Hardwood

Residuals

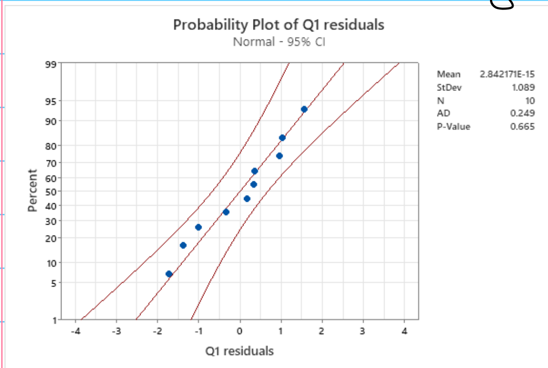
$e_1 = 10 - 8.97 = 1.03$

$e_2 = 15 - 14.65 = .345$

$\vdots$

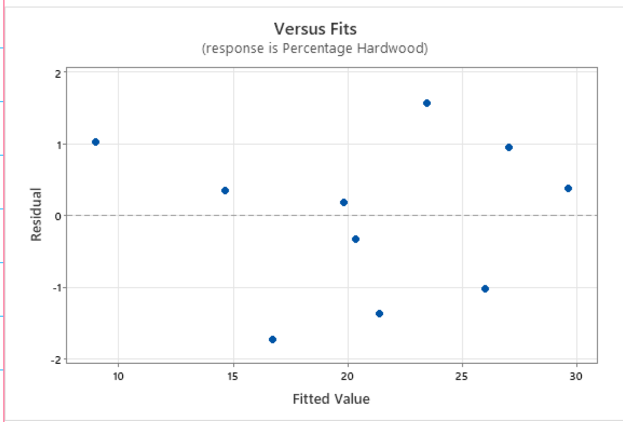
$e_{10} = 30 - 29.63 = 0.369$

a) Residual Normal Probability Plot



Insufficient evidence to reject the claim that the data is not normal. As pvalue = 0.665 > 0.05 = α, we can support the normality assumption of the model.

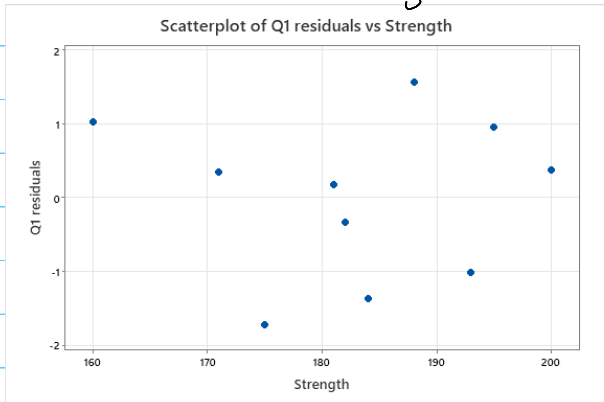
b) Residuals vs. Fitted Value



The residuals are evenly distributed about $y = 0$. There is no clear pattern, and thus the equal variance assumption is valid here.

If there was a pattern, I would apply a transformation

c) Residuals vs. Strength values (x)



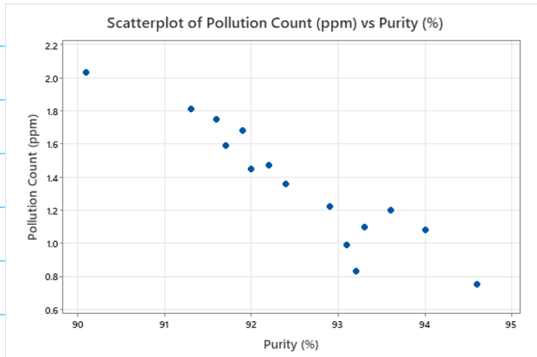
The residuals show no patterns and are evenly distributed about 0. The equal variance assumption is again valid.

4.27 in book

3.

Purity (%)		93.3	92.0	92.4	91.7	94.0	94.6	93.6
Pollution count (ppm)		1.10	1.45	1.36	1.59	1.08	0.75	1.20
Purity (%)	93.1	93.2	92.9	92.2	91.3	90.1	91.6	91.9
Pollution count (ppm)	0.99	0.83	1.22	1.47	1.81	2.03	1.75	1.68

a)



Without a fit there is a clear negative linear relationship between Pollution Count (y) and Purity Percent (x). I'm convinced!

Regression:

$\hat{y} = 29.23 - .303x$

$\left\{ \begin{array}{l} \hat{y} = \text{Pollution Count (ppm)} \\ x = \text{Purity Percent} \end{array} \right.$

b) ANOVA test for linear regression

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

Minitab p-value $\approx 0 < \alpha = 0.05$

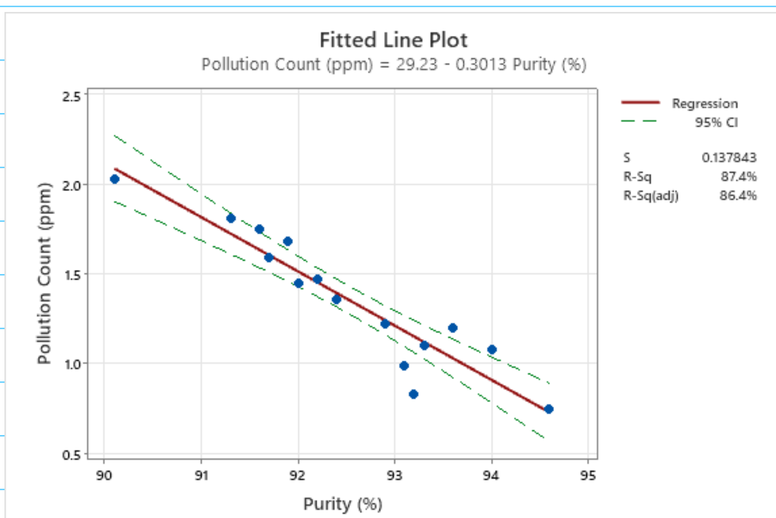
p-value $< \alpha$, so we have significant evidence to reject the claim that there is no linear relationship. We can accept there is a significant linear relationship between Pollution Count (y) and Purity Percent (x).

c) 95% CI for: Purity % = $(-0.3698, -.2327)$ ← slope

Pollution Count (Constant) = $(22.89, 35.57)$ ← intercept

In repeated sampling, the intervals calculated capture the true slope or intercept 95% of the time.

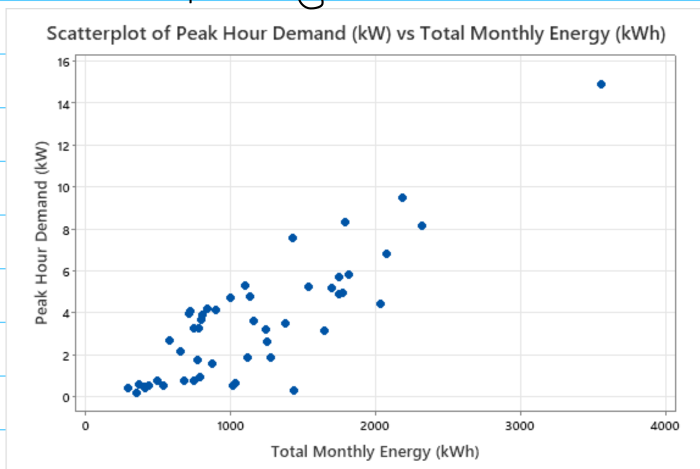
d)



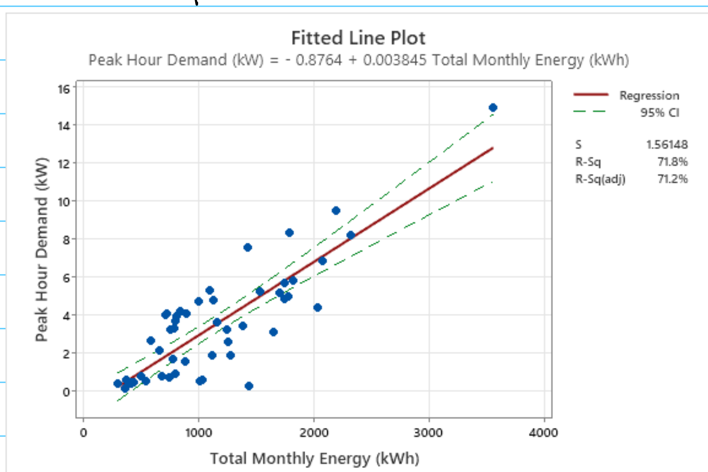
In repeated experiments there is a 95% chance that the CI on the responses calculated contain the true responses.

4. use $\begin{cases} x = \text{Total Monthly Energy Used (kWh)} \\ y = \text{Peak Hour Demand (kW)} \end{cases}$

a) Scatter plot of y vs x



b) Linear Fit



$\hat{y} = -0.8764 + .003845x$

$\begin{cases} x = \text{Total Monthly Energy Used (kWh)} \\ y = \text{Peak Hour Demand (kW)} \end{cases}$

c) $H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$

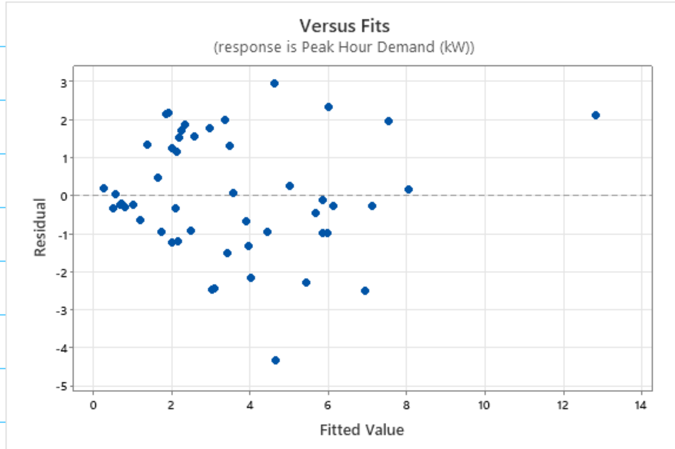
Analysis of Variance

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
Total Monthly Energy (kWh)	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%				

p-value $\approx 0 < \alpha = 0.05$

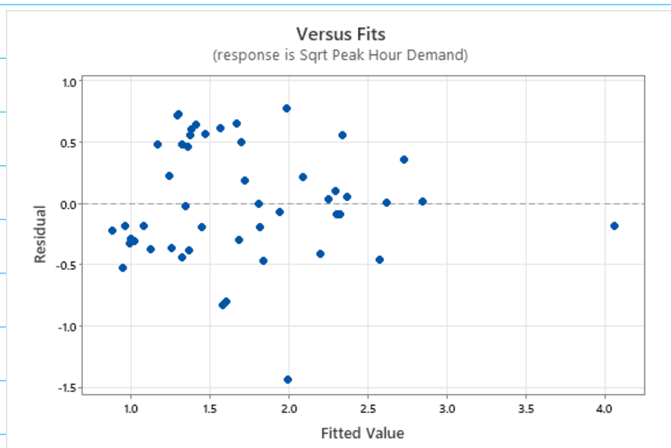
p-value $< \alpha$, so we have significant evidence to reject the claim that there is no linear relationship. We can accept there is a significant linear relationship between the Total Monthly Energy Used (x), and the Peak Hour Demand (y).

d)



There seems to be a feather like pattern, despite being roughly displayed about 0. At lower fitted values, residuals are closer together. Thus, the equality of variance may not be a good assumption and we should apply a data transformation to remedy this

e)



Yes, the data is more evenly distributed about zero and there is not a pattern corresponding to the fitted values. With this transformation, the equality of variance assumption appears to be a valid assumption.

5. a) Problem 4.30

(i) Satisfaction $(y) = 131.10 - 1.240 \cdot \text{Age } (x_1)$

$\left\{ \begin{array}{l} y = \text{Satisfaction} \\ x_1 = \text{Age} \end{array} \right.$

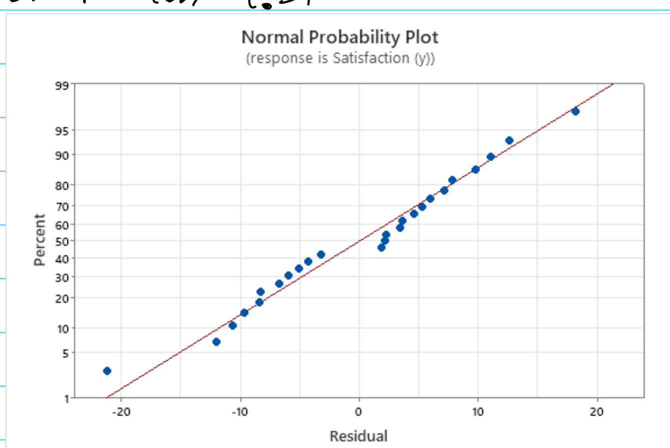
(ii) $H_0: \beta_1 = 0$ $\quad H_a: \beta_1 \neq 0$

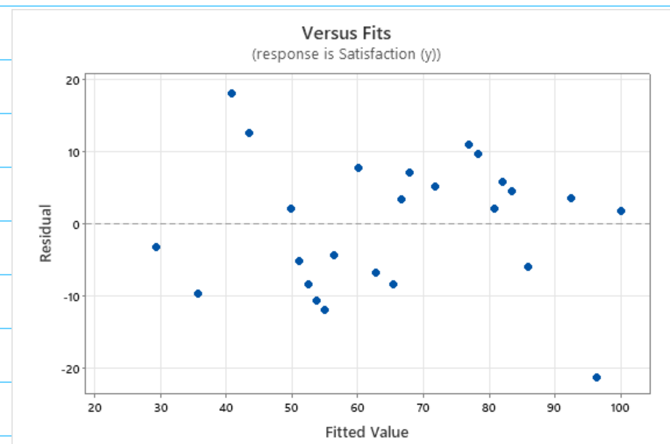
Minitab → p-value $= 0 < \alpha = 0.05$

we have significant evidence to reject the claim that there is no linear relationship. we can accept there is a significant linear relationship between the Age of the patient (x_1), and their satisfaction (y).

(iii) 81.24% of the variability is accounted for by the regressor variable age.

b) Problem 4.31





The normal probability plot of the residuals is linear with no outstanding variability, so the normality assumption is valid. The residuals vs. Fitted value is evenly distributed about 0 and no clear pattern is seen. Thus, the equality of variance assumption is valid. The model seems to be a good fit.

c) Problem 4.32

(i) Satisfaction $(y) = 143.47 - 1.031 \cdot$ Age $(x_1) - 0.556 \cdot$ Severity (x_2)

$\{ \; y =$ Satisfaction $\qquad x_1 =$ Age
$\qquad\qquad\qquad\qquad\qquad x_2 =$ Severity

$H_0: B_1 = B_2 = 0 \qquad H_a: B_i \neq 0$ for at least one i

Minitab $\rightarrow$ p-value $= 0 < \alpha = 0.05$

we have significant evidence to reject the claim that there is no linear relationship. We can accept there is a significant linear relationship between the Age of the patient (x_1), Severity of their illness (x_2), and their satisfaction (y).

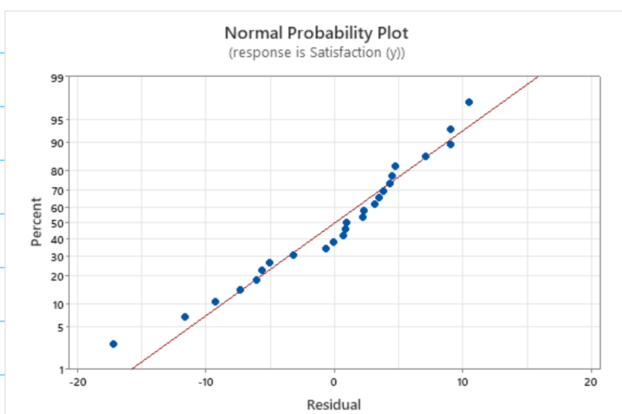
(ii) $R^2(x_1) = 81.24\%$ ← Age
$R^2(x_2) = 8.042\%$ ← Severity

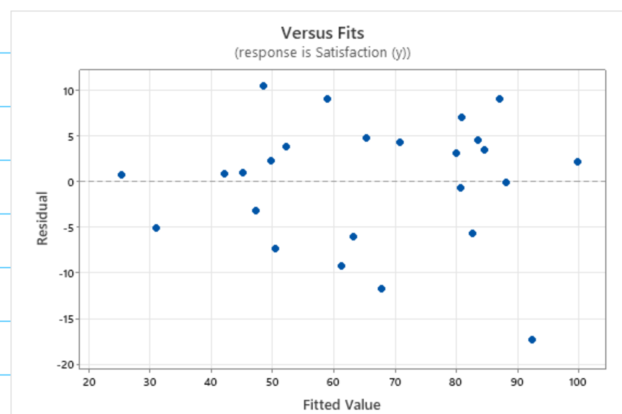
The portion of variability is significant $(> \alpha = 5\%)$ in both cases. Both regressors are needed.

(iii) $R^2_{mult.} = 88.72\%$ (adj)
$R^2_{single} = 80.43\%$ (adj)

The adjusted R^2 value, which accounts for overfitting effects, improved with the full fit expression. Thus we can conclude adding severity improved the model's quality.

d) Problem 4.33





The normal probability plot of the residuals is linear with no outstanding variability, so the normality assumption is valid. The residuals vs. Fitted value is evenly distributed about 0 and no clear pattern is seen. Thus, the equality of variance assumption is valid. The model seems to be a good fit.