Q1 (21 points)

Q1(21 points). Given a data set with four potential independent variables. For the following table.

ith model	R^2	Adj. R^2	C(k)	$\hat{\sigma}^2$ MS_{res}	SS_{Res}	I.V. in Model	
1	0.9694	0.967	0.7339	1.75298	22.78876	x2	
2	0.7308	0.7101	92.1947	15.41676	200.41791	x4	$k=2$
3	0.3486	b. 0.2988	238.6936	37.30295	484.93832	x3	
4	0.0085	-0.0678	369.0349	56.77528	738.07869	x1	
5	0.9724	0.9678	1.5826	1.71273	20.55278	x1x2	
6	0.9724	0.9678	1.5851	1.71314	e. 20.5577	x2x3	
7	0.9705	0.9656	2.3164	1.8315	21.97797	x2x4	$k=3$
8	a 0.7378	0.694	91.5351	16.27105	195.25264	x3x4	
9	0.731	0.6862	94.1047	16.68692	200.24308	x1x4	
10	0.3669	0.2614	233.6663	39.27421	471.29049	x1x3	
11	0.9736	0.9664	3.1122	d. 1.7854	19.63927	x1x2x3	
12	0.9731	0.9658	3.3	1.81854	20.00392	x1x2x4	$k=4$
13	0.9726	0.9652	c. 3.4870	1.85155	20.36709	x2x3x4	
14	0.7381	0.6667	93.3883	17.72432	194.96757	x1x3x4	
15	0.9739	0.9635	5	1.94213	19.42134	x1x2x3x4	$k=5$

1. (5 points) Fill the blank a, b, c, d, e in the above table.
2. (2 points) Calculate the total variation.
3. (5 points) Fill the Analysis of variance table for the last model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$.

Source	DF	sum of sqaure	Mean square	F-value
Model				
Error				
Total				

4. (2 points) Find the best model based on adjusted R^2.
5. (2 points) Find the best model based on $\hat{\sigma}^2$.
6. (5 points) Let $\alpha = 0.05$, implement an F-test: $H_0 : \beta 1 = \beta 3 = \beta 4 = 0, H_a$: At least one of $\beta 1, \beta 3$, and $\beta 4$ are not zero.

1. (5 points) Fill the blank a, b, c, d, e in the above table.

(a) $R^2 = 1 - \frac{SS_{Res}}{SS_T}$,

$\boxed{\text{1st model}}$: $R^2 = 1 - \frac{SS_{Res}}{SS_T}$

$$0.9694 = 1 - \frac{22.78876}{SS_T}$$

$$\frac{22.78876}{SS_T} = 1 - 0.9694$$

$$SS_T = \frac{22.78876}{1 - 0.9694}$$

$$SS_T = 744.7307$$

$\boxed{\text{For a}}$: $R^2 = 1 - \frac{SS_{Res}}{SS_T}$

$$= 1 - \frac{195.25264}{744.7307}$$

$$= \boxed{0.7378}$$

(b) Adj. R^2 :

$\boxed{\text{1st model}}$: $\bar{R}^2 = 1 - \frac{\hat{\sigma}^2 (n-1)}{SS_T}$

$$0.967 = 1 - \frac{1.75298(n-1)}{744.7307}, \quad \hat{\sigma}^2 = MS_{Res} = 1.75298$$

$$1.75298(n-1) = 744.7307(1 - 0.967)$$

$$n = \frac{744.7307(1-0.967)}{1.75298} + 1$$

$$n = 15$$

$\boxed{\text{For b}}$: $\bar{R}^2 = 1 - \frac{\hat{\sigma}^2 (n-1)}{SS_T}$

$$= 1 - \frac{37.30295(15-1)}{744.7307}$$

$$= \boxed{0.2988}$$

(C) C statistics

$$C = \frac{SS_{Res}}{\hat{\sigma}_p^2} - (n-2k)$$

$$= \frac{20.36709}{1.94213} - (15-2(4))$$

$$= \boxed{3.4870}$$

(d) $$MS_{Res} = \frac{SS_{Res}}{n-k}$$

$$= \frac{19.63927}{15-4}$$

$$= \boxed{1.7854}$$

(e) $$SS_{Res} = MS_{Res}(n-k)$$

$$= 1.71314(15-3)$$

$$= \boxed{20.5577}$$

2. (2 points) Calculate the total variation.

$\boxed{\text{1st model}}$: $R^2 = 1 - \frac{SS_{Res}}{SS_T}$

$$0.9694 = 1 - \frac{22.78876}{SS_T}$$

$$\frac{22.78876}{SS_T} = 1 - 0.9694$$

$$SS_T = \frac{22.78876}{1-0.9694}$$

$$SS_T = \boxed{744.7307}$$

3. (5 points) Fill the Analysis of variance table for the last model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$.

$\hat{\sigma}^2 = MS_{Res}$

$n = 15$, $k = 5$: $k-1$, $n-k$, $n-1$

Source	DF	sum of sqaure	Mean square	F-value
Model	4	$SS_R = 725.3094$	$MS_R = 181.3273$	93.3650
Error	10	$SS_{Res} = 19.42134$	$MS_{Res} = 1.94213$	
Total	14	$SS_T = 744.7307$		

$$SS_R = SS_T - SS_{Res} = 744.7307 - 19.42134 = 725.3094$$

$$MS_R = \frac{SS_R}{DF_{model}} = \frac{725.3094}{4} = 181.3273$$

$$F(model) = \frac{SS_R/(k-1)}{SS_{Res}/(n-k)} = \frac{725.3094/(5-1)}{19.42134/(15-5)} = 93.3650$$

4. (2 points) Find the best model based on adjusted R^2.

According to the 1st table provided, the adjusted R^2 of models 5 and 6 are both approximately equal to 0.9678. However, since the adjusted R^2 is a function of sigma_hat, a smaller sigma_hat leads to a larger adjusted R^2. The value of MS_res (sigma_hat_square) of model 5 is slightly smaller than the value in model 6 (1.71273 < 1.71314 indicating sqrt(1.71273) < sqrt(1.71314)). Hence, model 5 is the best based on the value of adjusted R^2 and the sigma_hat.

This can be proved by recalculating the adjusted R^2 for both models and then leaving more decimal places to compare which one is larger than the other.

$\boxed{\text{model 5}}$

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}^2(n-1)}{SS_T} = 1 - \frac{1.71213(15-1)}{744.7307} = 0.967814$$

$\boxed{\text{model 6}}$

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}^2(n-1)}{SS_T} = 1 - \frac{1.71314(15-1)}{744.7307} = 0.967795$$

Since, $0.967814 > 0.967795$, model 5 is the best.

5. (2 points) Find the best model based on $\hat{\sigma}^2$.

Model 5 has the smallest MS_res (sigma_hat_square) meaning that it has the smallest sigma_hat. This suggests that model 5 is likely to be the best model. Since the value of adjusted R^2 of model 5 is also the largest, so we can conclude that model 5 is for sure the best model.

6. (5 points) Let $\alpha = 0.05$, implement an F-test: $H_0 : \beta1 = \beta3 = \beta4 = 0, H_a$: At least one of $\beta1, \beta3$, and $\beta4$ are not zero.

Given the complete model : $\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon.}$ $\begin{cases} k = 5 \\ p = 4 \end{cases}$

Reduced - model : $y = \beta_0 + \beta_2 x_2 + \varepsilon$

The # of kept IV is $g = 1$, so $p - g = 4 - 1 = 3$, $n = 15$, $n - k = 15 - 5 = 10$

$SS_{Res}(C) = 19.42134$

$SS_{Res}(R) = 22.78876$ given by 1st model from table 1

$$\text{T.S. } F(x_1, x_3, x_4 \mid x_2) = \frac{SS_{drop} / (p-g)}{SS_{Res}(C) / (n-k)}$$

$$= \frac{[SS_{Res}(R) - SS_{Res}(C)] / (p-g)}{SS_{Res}(C) / (n-k)}$$

$$= \frac{(22.78876 - 19.42134) / 3}{19.42134 / 10}$$

$$= 0.5780$$

$F_\alpha(p-g, n-k) = F_{0.05}(3, 10) = 3.71 > F(x_1, x_3, x_4 \mid x_2) = 0.5780$

Since T.S. < R.P, we do not reject H_0; Hence, we conclude that there is evidence to drop x_1, x_3, and x_4 from the model.

Q2 (14 points)

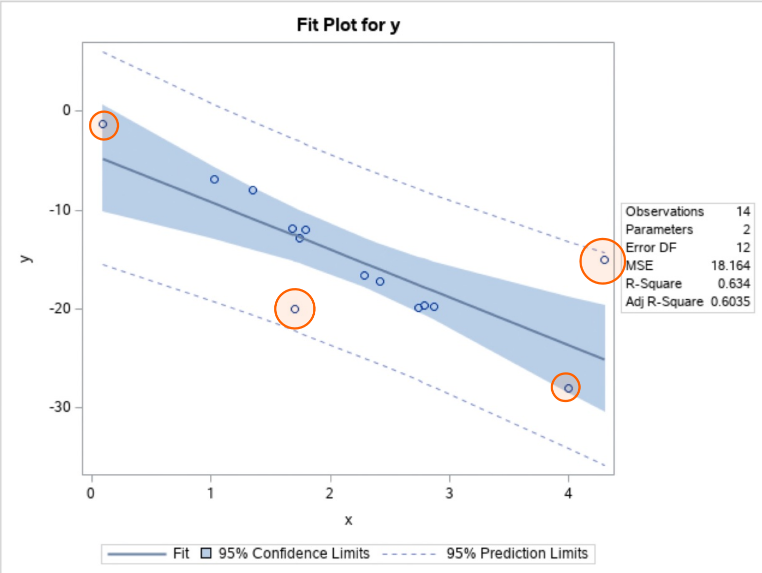
Q2(14 points) For the following data.

	x	y
1	2.29	-16.55
2	2.79	-19.62
3	2.42	-17.25
4	1.74	-12.80
5	1.35	-8.00
6	2.87	-19.75
7	1.03	-6.83
8	1.79	-12.02
9	1.68	-11.91
10	2.74	-19.93
11	0.09	-1.33
12	1.70	-20.00
13	4.00	-28.00
14	4.30	-15.00

Apply SAS to work on the following, code and output and explanation all requested.

1. (2 points) plot the scatter plot y vs x, and specify suspect outliers and/or influential points on the plot.
2. (2 points) detect any outliers with respect to x using the leverage value.
3. (2 points) detect any outliers with respect to y using the R student. Use the criterion $t_{0.025}^{(n-k-1)}$.
4. (1 points) Is there any points, x_i, which would substantially change the point prediction $\hat{y}_i$ if it is removed from the data set? Use the larger criterion 2.
5. (3 points) detect any influential points using Cook's distance measure.
6. (1 points) Is there any points which would substantially change the β_1 if it is removed from the data set? Use the larger criterion 2.
7. (3 points) Is there any points which would significantly damage or enhance the precision of the least square estimates?

1. (2 points) plot the scatter plot y vs x, and specify suspect outliers and/or influential points on the plot.



4 suspect points

2. (2 points) detect any outliers with respect to x using the leverage value.

The REG Procedure
Model: MODEL1
Dependent Variable: y

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	Cook's D	t_i RStudent	Hat Diag H h_{ii}	Cov Ratio	DFFITS	DFBETAS Intercept	DFBETAS x
1	-16.55	-15.3653	1.1431	-1.1847	4.106	-0.289	0.003	-0.2772	0.0719	1.2646	-0.0772	-0.0280	-0.0065
2	-19.62	-17.7761	1.2991	-1.8439	4.059	-0.454	0.011	-0.4387	0.0929	1.2673	-0.1404	0.0065	-0.0675
3	-17.25	-15.9921	1.1627	-1.2579	4.100	-0.307	0.004	-0.2949	0.0744	1.2657	-0.0836	-0.0210	-0.0168
4	-12.80	-12.7133	1.2383	-0.0867	4.078	-0.021	0.000	-0.0203	0.0844	1.2997	-0.0062	-0.0047	0.0024
5	-8.00	-10.8329	1.4506	2.8329	4.007	0.707	0.033	0.6914	0.1158	1.2362	0.2503	0.2256	-0.1550
6	-19.75	-18.1618	1.3419	-1.5882	4.045	-0.393	0.008	-0.3783	0.0991	1.2873	-0.1255	0.0127	-0.0663
7	-6.83	-9.2900	1.6813	2.4600	3.916	0.628	0.036	0.6115	0.1556	1.3183	0.2625	0.2517	-0.1931
8	-12.02	-12.9544	1.2185	0.9344	4.084	0.229	0.002	0.2195	0.0817	1.2847	0.0655	0.0478	-0.0233
9	-11.91	-12.4240	1.2645	0.5140	4.070	0.126	0.001	0.1210	0.0880	1.3015	0.0376	0.0296	-0.0163
10	-19.93	-17.5350	1.2745	-2.3950	4.067	-0.589	0.017	-0.5722	0.0894	1.2325	-0.1793	0.0018	-0.0805
11	-1.33	-4.7576	2.5048	3.4276	3.448	0.994	0.261	0.9935	0.3454	1.5310	0.7217	0.7216	-0.6427
12	-20.00	-12.5205	1.2555	-7.4795	4.073	-1.836	0.160	-2.0737	0.0868	0.6736	-0.6392	-0.4965	0.2689
13	-28.00	-23.6103	2.2191	-4.3897	3.639	-1.206	0.271	-1.2322	0.2711	1.2607	-0.7515	0.4096	-0.6449
14	-15.00	-25.0568	2.4967	10.0568	3.454	2.912	2.215	5.1453	0.3432	0.1561	3.7192	-2.2261	3.3096

Leverage h_{ii} : $2(\frac{k}{n}) = 2(\frac{2}{14})$; k = 2 (1 indep. variable)

$= 0.2857 < 0.3454$ (11th) and 0.3432 (14th)

Hence, 11th and 14th are leverage points (outliers with respect to x)

3. (2 points) detect any outliers with respect to y using the R student. Use the criterion $t_{0.025}^{(n-k-1)}$.

$n-k-1 = 14-2-1 = 11$

$t_{0.025}(n-k-1) = t_{0.025}(11) = 2.201$

$i=14$, $|t_{14}| = 5.1453 > t_{0.025}(n-k-1) \Rightarrow$ 14th is outlier W.R.T. y

4. (1 points) Is there any points, x_i, which would substantially change the point prediction $\hat{y}_i$ if it is removed from the data set? Use the larger criterion 2.

DFFITS :

$i=14$, $|DFFITS_{14}| = 3.7192 > 2 \Rightarrow$ remove 14th obs. would substantially change the prediction of y_{14}

5. (3 points) detect any influential points using Cook's distance measure.

Cook's distance D :

$F_{0.5}(k, n-k) = F_{0.5}(2, 14) = 0.7348 \approx 1$

when $i=14$, $D_{14} = 2.215 > F_{0.5}(n, n-k) \Rightarrow$ 14th is influential

6. (1 points) Is there any points which would substantially change the β_1 if it is removed from the data set? Use the larger criterion 2.

DFBETAS :

$i=14$, $|DFBETAS_{1,14}| = 3.396 > 2 \Rightarrow$ remove 14th obs. could substantially change the estimate of β_1 $(\hat{\beta}_1)$

7. (3 points) Is there any points which would significantly damage or enhance the precision of the least square estimates?

$1+\frac{3k}{n} = 1+\frac{6}{14} = 1.4286$

$1-\frac{3k}{n} = 1-\frac{6}{14} = 0.5714$

$i=11 > 1+\frac{3k}{n} \Rightarrow$ enhance precision with it

$i=14 < 1-\frac{3k}{n} \Rightarrow$ change precision with it