Solution for Assignment 1, MATH3805

1. Exercise 3.22 (3' for each sub-question)

   (a) x is the unflooded area ratio; y is heat transfer enhancement value. The regression model is

   $$y = \beta_0 + \beta_1 x + \varepsilon$$

   where,

   $$\begin{aligned} \hat{\beta}_1 &= \frac{SS_{xy}}{SS_{xx}} \\ &= \frac{\sum_{i=1}^{24} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{24} (x_i - \bar{x})^2} \\ &= \frac{9.592}{3.9532} = 2.4264 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 0.2134 \end{aligned}$$

   So the least square line to the data is

   $$\text{heat} = 0.2134 + 2.4264\text{ratio}$$

   (b) Figure 1

   (c)

   $$\begin{aligned} SSE &= \textstyle\sum_{i=1}^{24} (y_i - \hat{y}_i)^2 = 4.5311 \\ s^2 &= \tfrac{SSE}{n-2} = 0.2060 \end{aligned}$$

   (d)

   $$s = \sqrt{0.2060} = 0.4539$$

   Interpretation of s: We expect most (approximately 95%) of the observed y-values to lie within $2s$ of their respective least squares predicted values, $\hat{y}$.

   SAS output(Figure 2).

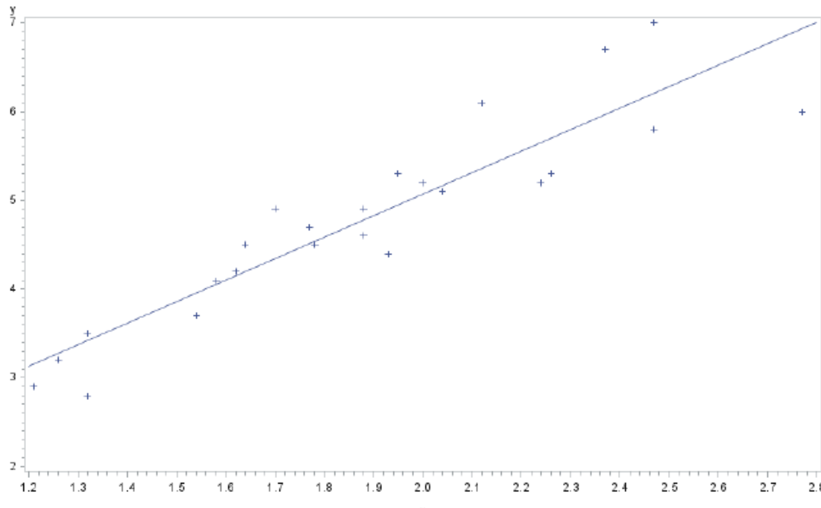

Figure 1: the scatterplot and the regression line

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: y1

Number of Observations Read	24
Number of Observations Used	24

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	23.27392	23.27392	113.00	<.0001
Error	22	4.53108	0.20596		
Corrected Total	23	27.80500			

Root MSE	0.45383	R-Square	0.8370
Dependent Mean	4.77500	Adj R-Sq	0.8296
Coeff Var	9.50421		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.21339	0.43900	0.49	0.6317
x1	1	2.42639	0.22825	10.63	<.0001

Figure 2: SAS output for Exercise 3.22

2. Exercise 3.24 (3' for each sub-question)

   (a) $H_0 : \beta_1 = 0 \quad H_1 : \beta_1 > 0.$

   $$\text{Test statistic: } t = \hat{\beta}_1 / s_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{s/\sqrt{\text{SS}_{xx}}} = 38.132$$
   $$p - value < 0.00005(one\, side)$$

   Given that p-value is smaller than 0.05, there is sufficient evidence to indicate that β_1 is positive, i.e. there is a positive linear relationship between x and y.

   (b) 95% confidence interval for the slope is $(1.335, 1.482)$. It means that when appraised properties value increase 1 unit, the increment of sale price y will fall into $[1.335, 1.482]$ with probability 95%.

   (c) remove the intercept.(Figure 4)
   collect more observations.
   SAS output(Figure 3).

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: y2

Number of Observations Read	76
Number of Observations Used	76

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6874034	6874034	1454.06	<.0001
Error	74	349833	4727.46974		
Corrected Total	75	7223866			

Root MSE	68.75660	R-Square	0.9516
Dependent Mean	431.69342	Adj R-Sq	0.9509
Coeff Var	15.92718		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	1.35868	13.76817	0.10	0.9217	-26.07501	28.79237
x2	1	1.40827	0.03693	38.13	<.0001	1.33468	1.48186

Figure 3: SAS output for Exercise 3.24

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: y2

Number of Observations Read	76
Number of Observations Used	76

Note: No intercept in model. R-Square is redefined.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	21037287	21037287	4509.55	<.0001
Error	75	349879	4665.05064		
Uncorrected Total	76	21387166			

Root MSE	68.30118	R-Square	0.9836
Dependent Mean	431.69342	Adj R-Sq	0.9834
Coeff Var	15.82169		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
x2	1	1.41126	0.02102	67.15	<.0001	1.36939	1.45312

Figure 4: SAS output without intercept for Exercise 3.24

3. Exercise 3.58 (10')

(1) We hypothesize a straight-line probabilistic model: $y = \beta_0 + \beta_1 x + \varepsilon$

(2) We collect the (x, y) values for each of the $n = 223$ experimental units in the sample.

(3) Next, we enter the data into a computer and use statistical software to estimate the unknown parameters in the deterministic component of the hypothesized model.

$$\hat{\beta}_0 = 0.35255, \quad \hat{\beta}_1 = 0.11644$$

Thus

$$\hat{y} = 0.35 + 0.12x$$

(4) Now, we specify the probability distribution of the random error component ε. The assumptions about the distribution are:

(1) $E(\varepsilon) = 0$
(2) $\operatorname{Var}(\varepsilon) = \sigma^2$ is constant for all x -values
(3) ε has a normal distribution
(4) ε 's are independent

$$s^2 = 26.18066$$

(5) We can now check the utility of the hypothesized model,

(a) Test of model utility:

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

$$p = .7821 > 0.05$$

Given that p-value is larger than 0.05, it is not sufficient to support the linear relationship between x and y.

(b) Confidence interval for slope: $(-0.70754, 0.940424)$

$$\hat{\beta}_1 \pm \left(t_{\alpha/2}\right) s_{\hat{\beta}_1} = 0.11644 \pm 1.96 \times 0.42040$$

(c) Numerical descriptive measures of model adequacy

$$r^2 = 0.0003$$

(6) $\hat{y} = 1.284$ if $x = 8.00$, confidence interval is $(-8.78654, 11.35454)$

$$\hat{y} \pm \left(t_{\alpha/2}\right) s \sqrt{1 + \frac{1}{n} + \frac{\left(x_{\mathrm{p}} - \bar{x}\right)^2}{\mathrm{SS}_{xx}}} = 1.284 \pm 1.96 \times 5.11670 \sqrt{1 + \frac{1}{223} + \frac{8 - 7.4267}{148.1339}}$$

The prediction is unreliable since the evidence for β_1 is insufficient.

SAS output(Figure5-6)

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: y3

Number of Observations Read	223
Number of Observations Used	223

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.00842	2.00842	0.08	0.7821
Error	221	5785.92676	26.18066		
Corrected Total	222	5787.93519			

Root MSE	5.11670	R-Square	0.0003
Dependent Mean	1.21731	Adj R-Sq	-0.0042
Coeff Var	420.32899		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.35255	3.14094	0.11	0.9107	-5.83749	6.54258
x3	1	0.11644	0.42040	0.28	0.7821	-0.71207	0.94495

Figure 5: SAS output for Exercise 3.58

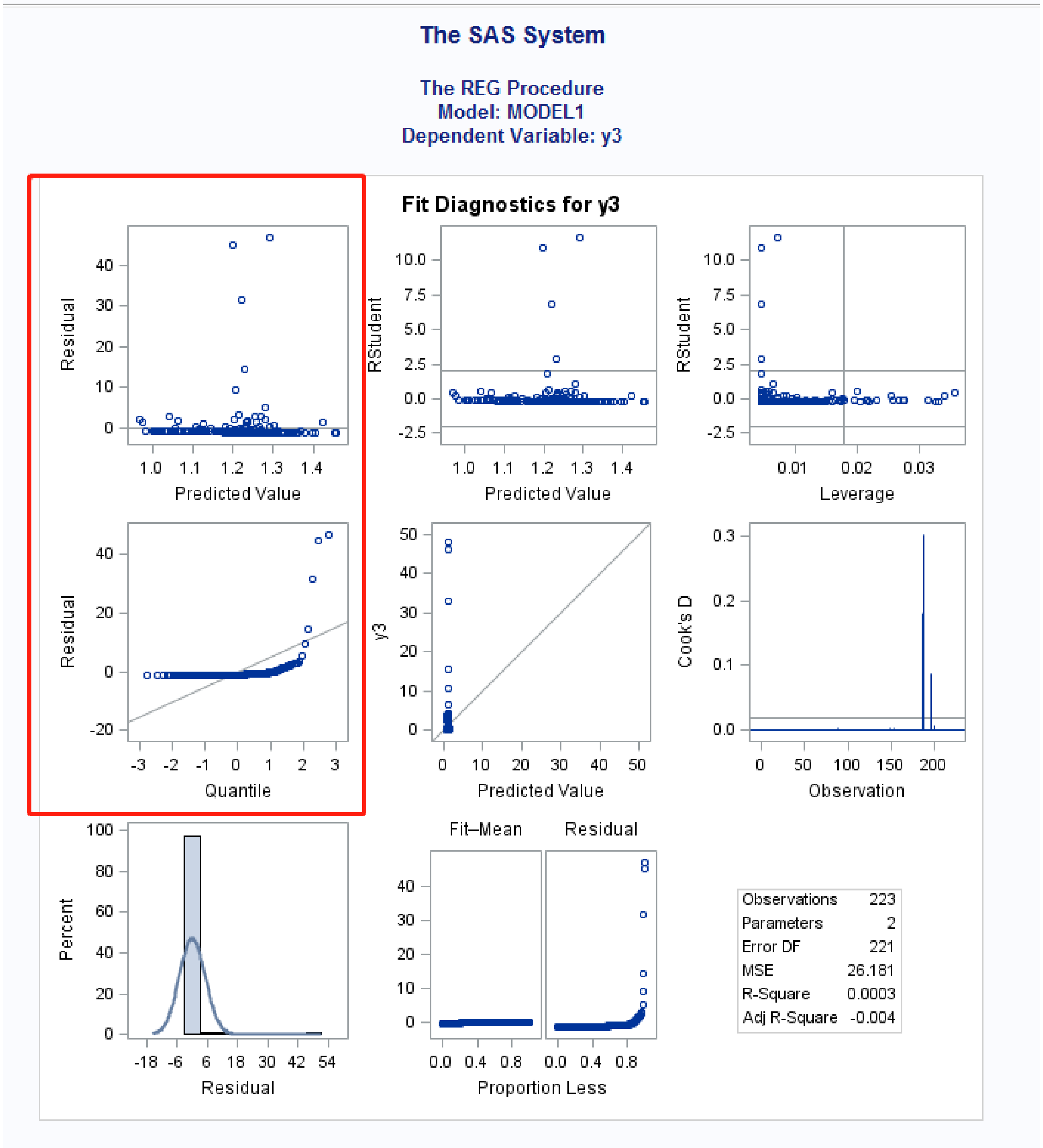

Figure 6: test for Exercise 3.58

4. Exercise 3.64 (3' for each sub-question)

   (a) A straight-line model through the origin is

$$y = \beta_1 x + \varepsilon$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{158400}{33020} = 0.2085$$

   (b)

$$\begin{aligned} \text{SSE} &= \sum (y_i - \hat{y})^2 = 22.66414 \\ s^2 &= \frac{\text{SSE}}{n-1} = \frac{22.66414}{9} = 2.518238 \\ s &= \sqrt{s^2} = 1.586896 \end{aligned}$$

   (c)

$$\begin{aligned} t &= \frac{\hat{\beta}_1}{s/\sqrt{\sum x_i^2}} = \frac{0.20846}{1.586896/\sqrt{158400}} = 52.28 \\ p - value &< 0.00005(\text{one side}) \end{aligned}$$

   Given that p-value is smaller than 0.05, so we can think the evidence is sufficient to support the linear relationship between x and y.

   (d) 95% confidence interval for β_1 is $(0.19944, 0.21748)$

$$\hat{\beta}_1 \pm (t_{\alpha/2})\, s_{\hat{\beta}_1} = \hat{\beta}_1 \pm (t_{\alpha/2}) \left(\frac{s}{\sqrt{\sum x_i^2}} \right) = 0.20846 \pm 1.833 \times 0.00399$$

   (e) 95% confidence interval for $E(y)$ when $x = 125$ is $(24.9300, 27.1849)$

$$\hat{y} \pm (t_{\alpha/2})\, s_{\hat{y}} = \hat{y} \pm (t_{\alpha/2})\, s \left(\frac{x_{\text{p}}}{\sqrt{\sum x_i^2}} \right) = 26.00575 \pm 1.833 \times 1.586896 \frac{125}{\sqrt{33020}}$$

   (f) 95% confidence interval for y when $x = 125$ is $(22.47521, 29.53629)$

$$\hat{y} \pm (t_{\alpha/2})\, s_{(y-\hat{y})} = \hat{y} \pm (t_{\alpha/2})\, s \sqrt{1 + \frac{x_{\text{p}}^2}{\sum x_i^2}} = 26.00575 \pm 1.833 \times 1.586896 \sqrt{1 + \frac{125^2}{33020}}$$

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: y4

Number of Observations Read	10
Number of Observations Used	10

Note: No intercept in model. R-Square is redefined.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6883.33586	6883.33586	2733.39	<.0001
Error	9	22.66414	2.51824		
Uncorrected Total	10	6906.00000			

Root MSE	1.58690	R-Square	0.9967
Dependent Mean	23.60000	Adj R-Sq	0.9964
Coeff Var	6.72413		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
x4	1	0.20846	0.00399	52.28	<.0001	0.19944	0.21748

Figure 7: SAS output for Exercise 3.64

5. Exercise 3.68 (3' for each of (a)-(g) and 2' for (h) and (i))

   (a) Figure 8

   (b) Figure 8

   (c) The parameter β_1 reflect the relationship between the number of factors per patient and the patient's length of stay. β_1 can be interpreted by, when the number of factors increase 1 , the average of patient's length of stay will increase β_1.

$$y = \beta_0 + \beta_1 x$$
$$\beta_1 = 0.01475$$

   (d)

$$\begin{aligned} &H(0) : \beta_1 = 0 \\ &H(1) : \beta_1 \neq 0 \qquad\qquad p - value < 0.0001 \end{aligned}$$

   Given that p-value is small enough, it is sufficient to support the linear relationship between x and y.

   (e) We have 95% confidence that (0.00922, 0.02029) covers β_1.

   (f) $R^2 = 0.3740 \Rightarrow r = \sqrt{R^2} = 0.6116$. It shows that the patient's length of stay is positive related to the number of factors per patient.

   (g) The result of ANOVA is shown in Figure 9. R-square is 0.374. We can find that the f-value is 28.68 and the p-value of the test is smaller than 0.05. So the model we constructed in part b is useful.

   (h) The 95% prediction interval at $x = 231$ is $(2.4480, 10.9808)$.

   (i) MSE is large. Other variables or information are needed to improve the performance.

   SAS output(8-10)

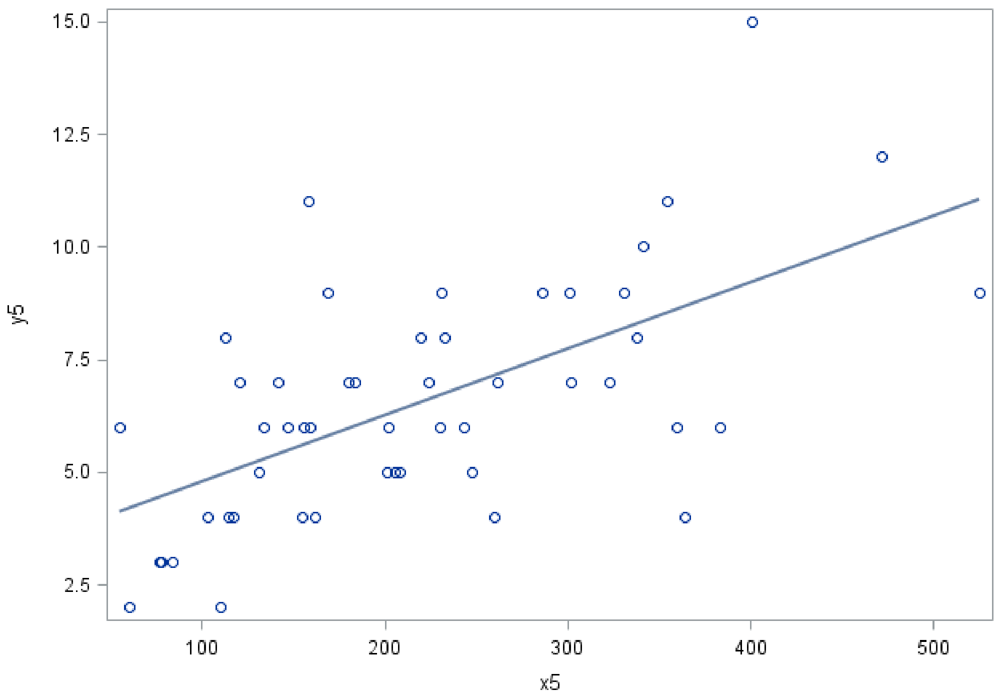

Figure 8: scatter plot and regression line for Exercise 3.68

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: y5

Number of Observations Read	50
Number of Observations Used	50

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	126.58393	126.58393	28.68	<.0001
Error	48	211.83607	4.41325		
Corrected Total	49	338.42000			

Root MSE	2.10077	R-Square	0.3740
Dependent Mean	6.54000	Adj R-Sq	0.3610
Coeff Var	32.12193		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	3.30603	0.67297	4.91	<.0001	1.95293	4.65914
x5	1	0.01475	0.00276	5.36	<.0001	0.00922	0.02029

Figure 9: SAS output for Exercise 3.68

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: y5

Output Statistics						
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	9	6.7144	0.2989	2.4480	10.9808	2.2856
2	7	8.0718	0.4124	3.7673	12.3764	-1.0718
3	8	4.9733	0.4169	0.6671	9.2796	3.0267
4	5	6.3750	0.2987	2.1087	10.6414	-1.3750
5	4	5.6963	0.3363	1.4187	9.9740	-1.6963
6	4	5.0323	0.4093	0.7290	9.3357	-1.0323
7	6	5.6521	0.3402	1.3731	9.9310	0.3479
8	9	5.7996	0.3277	1.5246	10.0746	3.2004

Figure 10: Predict value for Exercise 3.68

8. Exercise 3.80 (2' for each sub-question)

(a)

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Calculation steps:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{SS_{xy}}{SS_{xx}} \\
&= \frac{\sum_{i=1}^{24} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{24} (x_i - \bar{x})^2} \\
&= \frac{-15.728}{297.716} = -0.05283 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = -13.4903
\end{aligned}$$

Intepretation: For the slope, when temperature increases 1 unit, the average of proportion of impurity will decrease 0.0528. The regression line passes (0,-13.4903) while this point is not meaningful here since this case only focuses on the quite low temperature around -260^oC.

(b)

$$\begin{aligned}
H(0) &: \beta_1 = 0 \\
H(1) &: \beta_1 > 0
\end{aligned}$$

$$\hat{\beta}_1 \pm t_{2/\alpha} s_{\hat{\beta}_1} = -0.05283 \pm 1.96 \times 0.00773$$

95% confidence interval for β_1 is $(-0.07065, -0.03501)$. The interval supports the hypothesis that temperature contributes information about the proportion of impurity.

(c) $R^2 = 0.8538$, So the linear model can explain 85.38% variation of response.

(d) The prediction interval is

$$\hat{y} \pm \left(t_{\alpha/2}\right) \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} = (0.5987, 1.2653)$$

The proportion will appear in this interval with probability 95%.

(e) Because y is bounded in $[0, 1]$ and we do not take this constrain into consideration. If we transform y to $\log(-\log(y))$ or $\log(y/(1-y))$, prediction interval will be more reliable.

SAS output(Figure13)

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: PROPPASS

Number of Observations Read	10
Number of Observations Used	10

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.83089	0.83089	46.73	0.0001
Error	8	0.14225	0.01778		
Corrected Total	9	0.97315			

Root MSE	0.13335	R-Square	0.8538
Dependent Mean	0.68260	Adj R-Sq	0.8356
Coeff Var	19.53519		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-13.49035	2.07377	-6.51	0.0002	-18.27247	-8.70822
TEMP	1	-0.05283	0.00773	-6.84	0.0001	-0.07065	-0.03501

Figure 13: SAS output for Exercise 3.80