40.017 PROBABILITY & STATISTICS

Homework 5

Michael Hoon
1006617
Section 2

April 16, 2024

Question 1

(a)

Since we have 4 brands of spark plugs, with 5 plugs of each brands being tested, we have $N = 4 \times 5 = 20$, and $k = 4$ for each brand. The degrees of freedom are given by: $N - k = 20 - 4 = 16$, $k - 1 = 3$, and $N - 1 = 19$. Figure 1 below shows the Excel screenshot of the missing entries.

ANOVA								
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	*N*	*k*
Between Groups	75081.720	3	25027.240	1.701	0.207	3.239	20	4
Within Groups	235419.040	16	14713.690					
Total	310500.760	19						

Figure 1: Excel missing entries

The MSE is calculated using the relation $\text{MSE} = \frac{\text{SSE}}{N-k}$, and the SSA is calculated with the ANOVA Identity $\text{SST} = \text{SSA} + \text{SSE}$. Lastly, MSA is obtained with the relation $\text{MSA} = \frac{\text{SSA}}{k-1}$. The answers are given to 3 decimal places.

(b)

Let μ_i denote the mean performance of the ith brand of spark plugs, $\forall i \in 1, 2, 3, 4$. We define the hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_1 : \text{at least 1 pair of the } \mu_i\text{'s are different}$$

(c)

From Figure 1, since the p-value $= 0.206927 > \alpha = 0.05$, we do not have sufficient evidence to reject H_0 at the 95% significance level, and hence we conclude that the mean performance of the 4 brands of spark plugs are equal.

Question 2

(a)

Using the Bonferroni Method, we have $m = \binom{k}{2}$ pairs involved in testing. Since we have 4 sites in total, then we have $m = \binom{4}{2} = 6$ pairs.

(b)

	Water salinity at four sites			
	site I	**site II**	**site III**	**site IV**
	37.54	40.17	39.04	38.94
	37.01	40.8	39.21	39.53
	36.71	39.76	39.05	39.18
	37.03	39.7	38.24	38.38
	37.32	40.79	38.53	38.92
	37.01	40.44	38.71	38.65
	37.03	39.79	38.89	39.96
	37.7	39.38	38.66	38.65
	37.36		38.51	39.38
	36.75		40.08	
	37.45			
	x			
Sum	408.91 + x	320.83	388.92	351.59
Mean	(408.91 + x) / 12	40.10375	38.892	39.065556
Variance		0.2823696	0.2609289	0.2475028
n	12	8	10	9
N	39			
k	4			
Grand Mean	((408.91 + x) / 12 + 118.0613056)/39			

Anova: Single Factor

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	40.46382259	3	13.48794086	47.91241022	1.76215E-12	2.874187484
Within Groups	9.852936389	35	0.281512468			
Total	50.31675897	38				

118.0613056

Figure 2: Excel values

The formula for SSA is given by

$$\begin{aligned} \text{SSA} &= \sum_i (\bar{y}_i - \bar{\bar{y}})^2 \\ &= \sum_{i=1}^{4} n_i (\bar{y}_i - \bar{\bar{y}})^2 \end{aligned} \tag{1}$$

From the table, we find that SSA $= 40.46382259$. From Equation 1, we use the calculated $\bar{y}_i$'s as well as the grand mean to form a quadratic in x. We have the relation:

$$40.46382259 = 12\left(\frac{408.91 + x}{12} - \bar{\bar{y}}\right)^2 + 8\left(40.10375 - \bar{\bar{y}}\right)^2 + 10\left(38.892 - \bar{\bar{y}}\right)^2 + 9\left(39.065556 - \bar{\bar{y}}\right)^2$$

where $\bar{\bar{y}} = (1470.25 + x)/39$. Restructuring the equation, we have

$$\begin{aligned} 12\left(\frac{408.91 + x}{12} - \bar{\bar{y}}\right)^2 + 8\left(40.10375 - \bar{\bar{y}}\right)^2 + 10\left(38.892 - \bar{\bar{y}}\right)^2 \\ +9\left(39.065556 - \bar{\bar{y}}\right)^2 - 40.46382259 = 0 \end{aligned} \tag{2}$$

We can now solve for x in Equation 2 by using an online solver. First, plotting the equation in desmos we get the following graph:

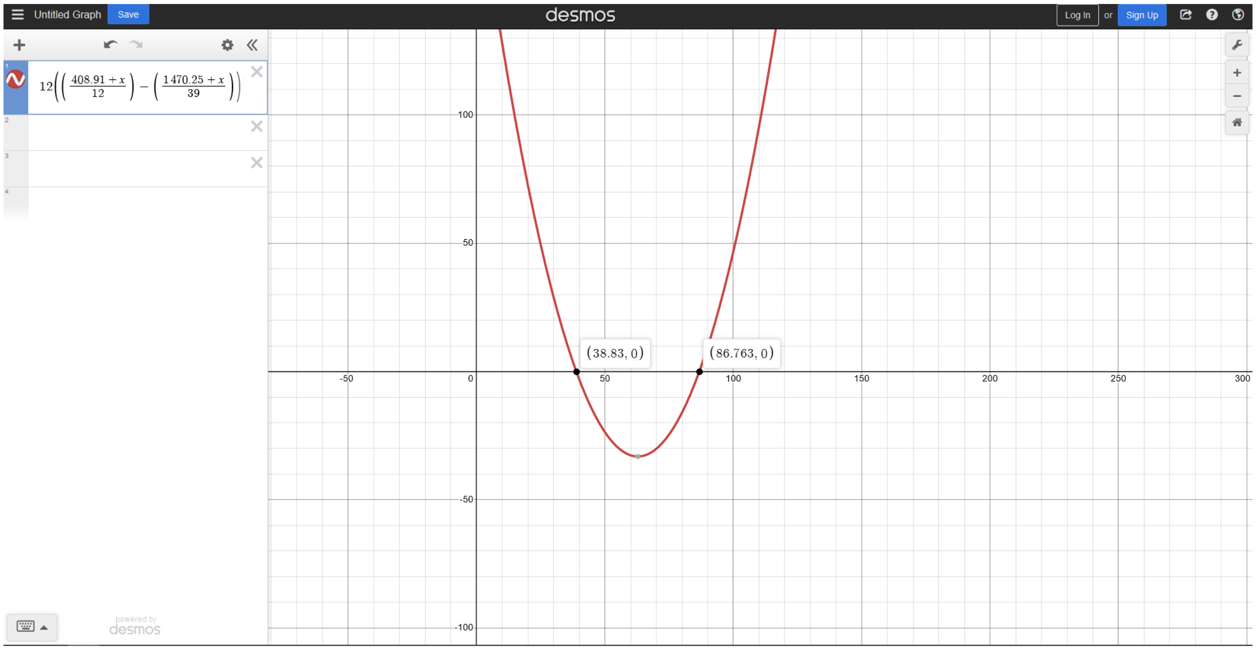


Figure 3: Desmos Graph

From here, we can see that the two values of x are 38.83 and 86.763. To determine which of the two is the correct solution, we plug these values into the Excel file and run the ANOVA test again:

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Column 1	12	447.74	37.31166667	0.322542424		
Column 2	8	320.83	40.10375	0.282369643		
Column 3	10	388.92	38.892	0.260928889		
Column 4	9	351.59	39.06555556	0.247502778		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	40.46382259	3	13.48794086	47.91241022	1.762E-12	2.8741875
Within Groups	9.852936389	35	0.281512468			
Total	50.31675897	38				

Figure 4: ANOVA for $x = 38.83$

We see that for $x = 38.83$, we obtain an ANOVA table with values which correspond to the original ANOVA table, and we can conclude that the correct solution for x is indeed 38.83.

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Column 1	12	495.673	41.30608333	205.0193295		
Column 2	8	320.83	40.10375	0.282369643		
Column 3	10	388.92	38.892	0.260928889		
Column 4	9	351.59	39.06555556	0.247502778		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	40.4629008	3	13.4876336	0.208739113	0.889663	2.8741875
Within Groups	2261.517595	35	64.61478842			
Total	2301.980495	38				

Figure 5: ANOVA for $x = 86.763$

On the other hand, the ANOVA table for $x = 86.763$ does not match the values given in the original ANOVA table, hence we **reject this solution**.

Question 3

The 90% Prediction Interval for the winning distance corresponding to the year 1940 is given by:

$$\left(\hat{y}^*_{1940} - t_{n-2,\,\alpha/2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x^*_{1940} - \bar{x})^2}{(n-1)s_x^2}},\ \hat{y}^*_{1940} + t_{n-2,\,\alpha/2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x^*_{1940} - \bar{x})^2}{(n-1)s_x^2}}\right)$$

where s represents the Mean Squared Error (MSE) and we have $n = 29$ observations, with $\alpha = 0.10$. The values for each variable used is calculated in Excel in Figure 6.

year	winning distance (m)
1896	13.71
1900	14.47
1904	14.35
1908	14.92
1912	14.76
1920	14.51
1924	15.53
1928	15.21
1932	15.72
1936	16.00
1948	15.40
1952	16.22
1956	16.35
1960	16.81
1964	16.85
1968	17.39
1972	17.35
1976	17.29
1980	17.35
1984	17.25
1988	17.61
1992	18.17
1996	18.09
2000	17.71
2004	17.79
2008	17.67
2012	17.81
2016	17.86
2021	17.98

x*	1940
x_bar	1960.58621
s_x^2	1474.82266
s_y^2	1.82949
s_xy^2	2486.934603
s_xy	49.86917
B_1 hat	0.033813676
B_0 hat	-49.80738445
y hat	15.79114606
t	1.703288446
s	0.385404865

Upper CI	15.12016
Lower CI	16.46213

Data Analysis Package

Regression Statistics	
Multiple R	0.960056982
R Square	0.921709408
Adjusted R Square	0.918809757
Standard Error	0.385404865
Observations	29

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	47.21528274	47.21528	317.869	1.83551E-16
Residual	27	4.010496575	0.148537		
Total	28	51.22577931			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	-49.80738445	3.719072514	-13.3924	1.93E-13	-57.43829092	-42.17647797	-56.14203769	-43.47273121
X Variable 1	0.033813676	0.001896567	17.82888	1.84E-16	0.029922241	0.03770511	0.030583274	0.037044077

Figure 6: Excel Calculations for variables

With this, we obtain the 90% Confidence Interval in 3 decimal places:

$$(14.983, 16.599)$$

Question 4

(a)

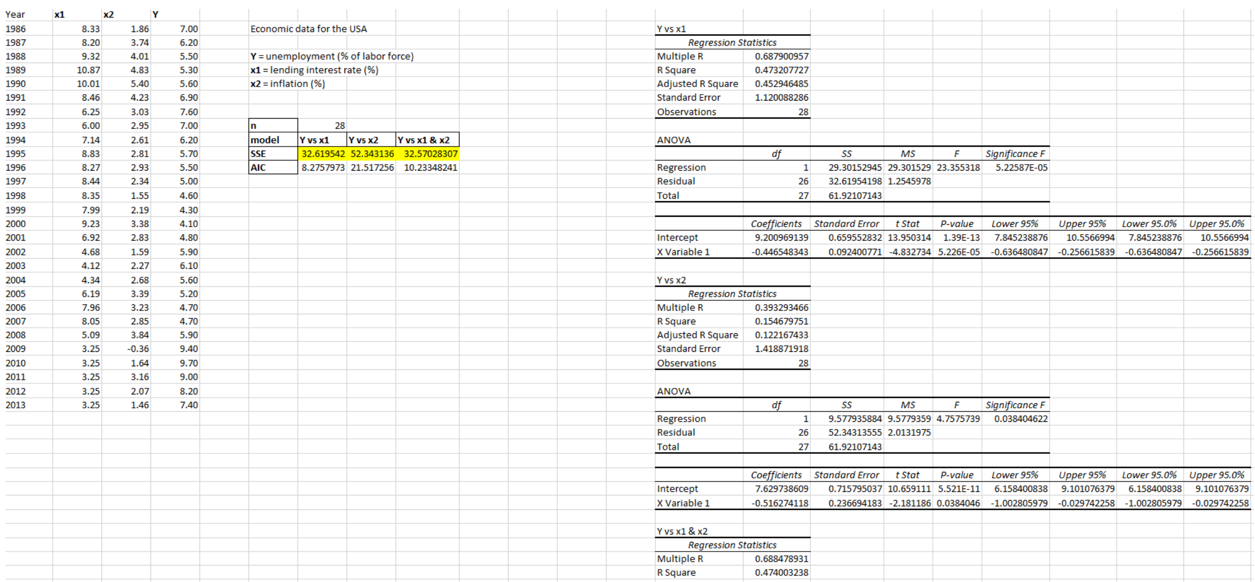
The formula for SSE for Regression is given by:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Let n be the number of rows of data, and k be the total number of predictors. AIC is defined as follows:

$$\text{AIC} = n \ln \left(\frac{\text{SSE}}{n} \right) + 2(m+1)$$

with m representing the subset of number of predictors in the model. The Excel calculated values are given in Figure below.



Year	x1	x2	Y
1986	8.33	1.86	7.00
1987	8.20	3.74	6.20
1988	9.32	4.01	5.50
1989	10.87	4.83	5.30
1990	10.01	5.40	5.60
1991	8.46	4.23	6.90
1992	6.25	3.03	7.60
1993	6.00	2.95	7.00
1994	7.14	2.61	6.20
1995	8.83	2.81	5.70
1996	8.27	2.93	5.50
1997	8.44	2.34	5.00
1998	8.35	1.55	4.60
1999	7.99	2.19	4.30
2000	9.23	3.38	4.10
2001	6.92	2.83	4.80
2002	4.68	1.59	5.90
2003	4.12	2.27	6.10
2004	4.34	2.68	5.60
2005	6.19	3.39	5.20
2006	7.96	3.23	4.70
2007	8.05	2.85	4.70
2008	5.09	3.84	5.90
2009	3.25	-0.36	9.40
2010	3.25	1.64	9.70
2011	3.25	3.16	9.00
2012	3.25	2.07	8.20
2013	3.25	1.46	7.40

Economic data for the USA

Y = unemployment (% of labor force)
x1 = lending interest rate (%)
x2 = inflation (%)

n	28		
model	Y vs x1	Y vs x2	Y vs x1 & x2
SSE	32.619542	52.343136	32.57028307
AIC	8.2757973	21.517256	10.23348241

Y vs x1

Regression Statistics	
Multiple R	0.687900957
R Square	0.473207727
Adjusted R Square	0.452946485
Standard Error	1.120088286
Observations	28

ANOVA

	df	SS	MS	F	Significance F
Regression	1	29.30152945	29.301529	23.355318	5.22587E-05
Residual	26	32.61954198	1.2545978		
Total	27	61.92107143			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	9.200969139	0.659552832	13.950314	1.39E-13	7.845238876	10.5566994	7.845238876	10.5566994
X Variable 1	-0.446548343	0.092400771	-4.832734	5.226E-05	-0.636480847	-0.256615839	-0.636480847	-0.256615839

Y vs x2

Regression Statistics	
Multiple R	0.393293466
R Square	0.154679751
Adjusted R Square	0.122167433
Standard Error	1.418871918
Observations	28

ANOVA

	df	SS	MS	F	Significance F
Regression	1	9.577935884	9.5779359	4.7575739	0.038404622
Residual	26	52.34313555	2.0131975		
Total	27	61.92107143			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	7.629738609	0.715795037	10.659111	5.521E-11	6.158400838	9.101076379	6.158400838	9.101076379
X Variable 1	-0.516274118	0.236694183	-2.181186	0.0384046	-1.002805979	-0.029742258	-1.002805979	-0.029742258

Y vs x1 & x2

Regression Statistics	
Multiple R	0.688478931
R Square	0.474003238

Figure 7: SSE and AIC Values

The SSE and AIC values for each model is given in Table below (3 decimal places).

Model	Y vs x_1	Y vs x_2	Y vs $x_1,\ x_2$
SSE	32.620	52.343	32.570
AIC	8.276	21.517	10.233

Table 1: Table of SSE and AIC Values

The AIC values are computed as follows (3 decimal places):

$$\mathrm{AIC}_{x_1} = 28 \ln\left(\frac{32.61954}{28}\right) + 2(1+1) \approx 8.276$$
$$\mathrm{AIC}_{x_2} = 28 \ln\left(\frac{52.34314}{28}\right) + 2(1+1) \approx 21.517$$
$$\mathrm{AIC}_{x_1,\, x_2} = 28 \ln\left(\frac{32.57028307}{28}\right) + 2(2+1) \approx 10.233$$

(b)

The best model corresponds to the one with the lowest AIC, which is given by

$$\min\{8.276, 21.517, 10.233\} = 8.276$$

This value corresponds to model (1), which is Y vs x_1.

Question 5

The Python code is modified to conduct the bootstrap resampling 10^4 times, and the distribution is obtained from the means of each resampling process. Then, the values are sorted in ascending order into a Python list (using list comprehension), and a Histogram of the distribution is plotted, with the corresponding 'L' and 'U' markers to denote the Lower and Upper bounds of the bootstrap Confidence Interval for the true mean.

Since we are finding the 99% Confidence Interval, only the lowest $\frac{0.01\times10^4}{2} = 50$ and highest 50 values were excluded from the interval. Figure 8 below shows the Python Code used.

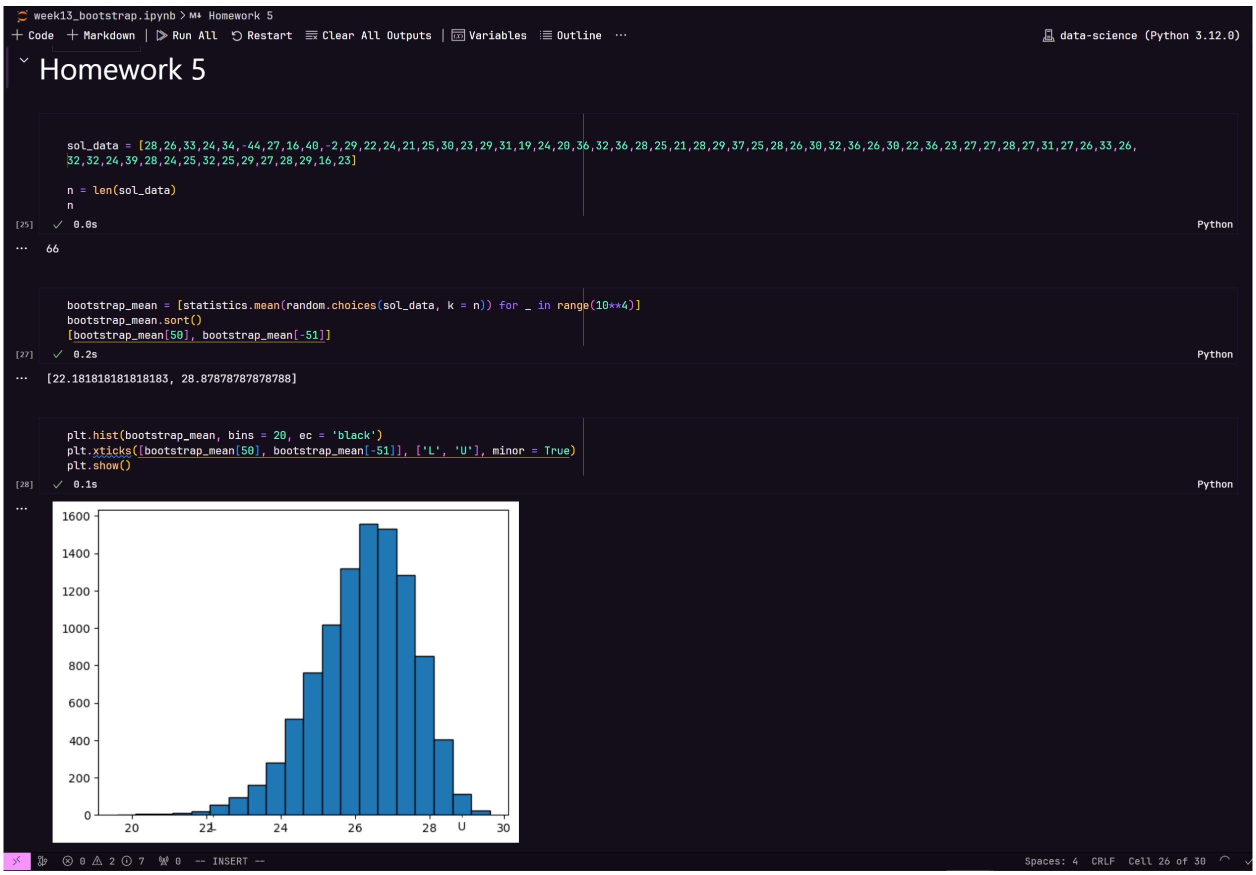


Figure 8: Python Code for Bootstrap

From the Python code, we obtain a 99% Confidence Interval (3 decimal places) of

$$[22.182, 28.879]$$