Sui Kun Guan

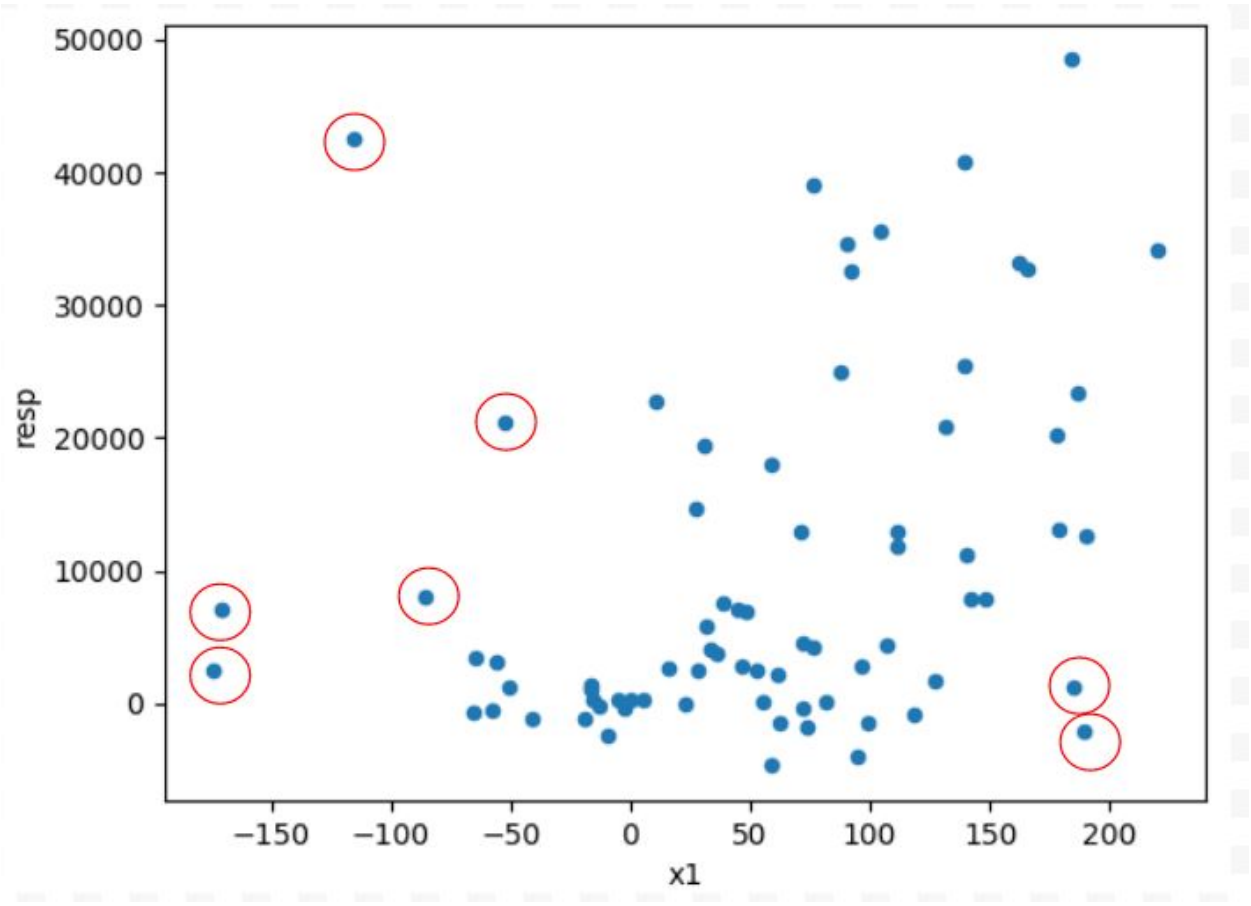
CS 286

Nov. 6 Class Ex.

4a. Descriptive statistics for each variable in the data set:

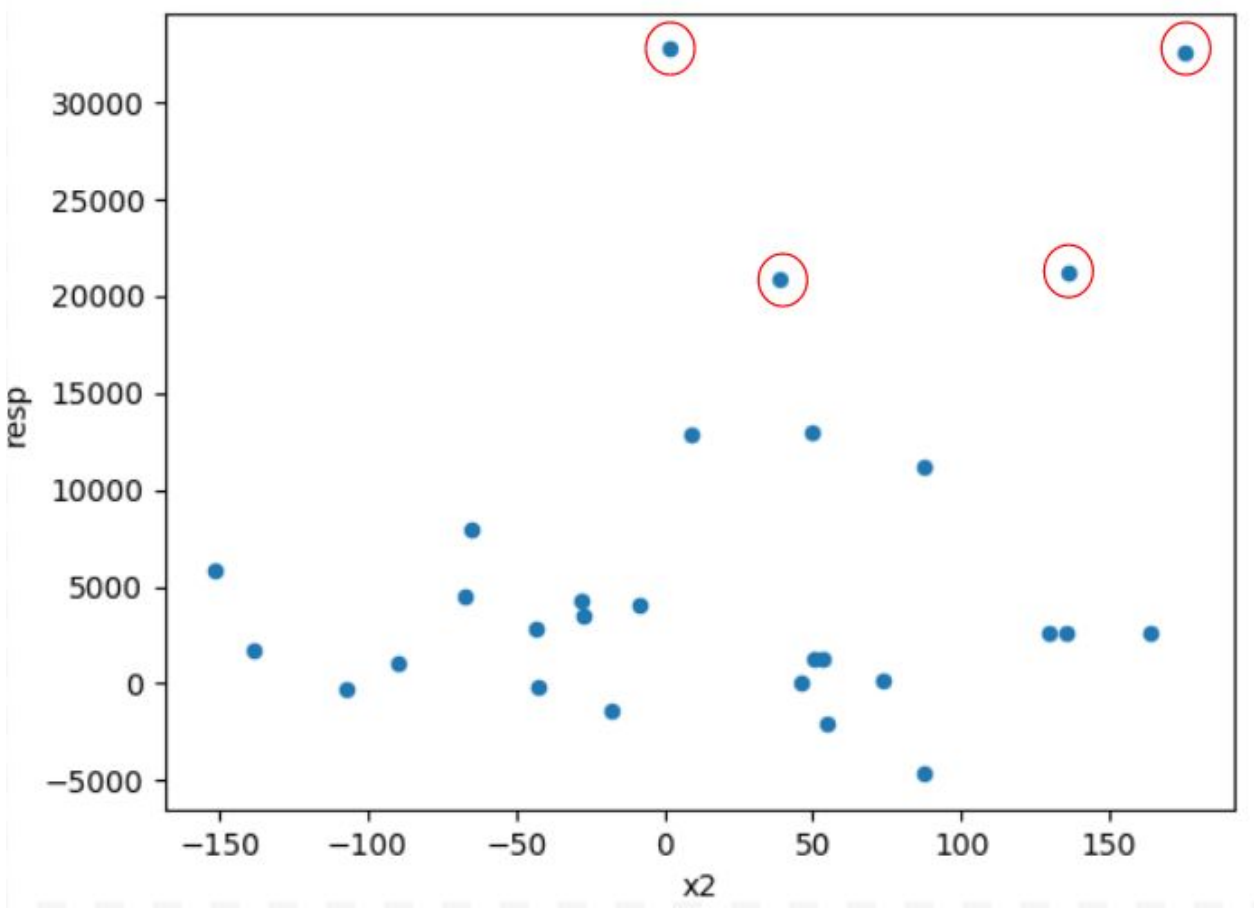
```
Descriptive statistics:
                resp          x1          x2          x3          x4
count      75.000000   75.000000   28.000000   28.000000   56.000000
mean     9988.586102   55.694617   18.108467   49.262491   86.782752
std     13190.846109   84.903342   88.967309   96.720908   96.697312
min     -4672.127047 -174.689621 -151.938749 -190.530177 -127.439196
25%       241.853001   -1.345912  -42.805844    7.061486   24.241707
50%      4061.197536   58.614509   24.035193   34.106820   85.689088
75%     16373.618170  111.301922   77.180591  107.045923  145.037815
max     48486.250496  220.344097  175.782436  235.856357  311.827542

SE of the mean:
resp    1523.147710
x1         9.803793
x2        16.813241
x3        18.278534
x4        12.921722
dtype: float64
```

4b. Scatter plot of resp vs x1



From the scatter plot, it shows there could be a positive linear relationship between the variable “x1” and “resp”. I can also identify several outliers from the scatter plot. I have circled them above.

4c. Scatter plot of resp vs x2



From the scatter plot, it shows that there is no/weak relationship between the variable “x2” and “resp”. There are several outliers circled.

4d. Linear Regression & Anova. Software output:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   resp   R-squared:                       0.591
Model:                            OLS   Adj. R-squared:                  0.519
Method:                 Least Squares   F-statistic:                     8.294
Date:                Tue, 06 Nov 2018   Prob (F-statistic):           0.000270
Time:                        14:57:19   Log-Likelihood:                -283.66
No. Observations:                  28   AIC:                             577.3
Df Residuals:                      23   BIC:                             584.0
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -3064.5092   2230.739     -1.374      0.183   -7679.144    1550.126
x1            42.5859     16.431      2.592      0.016       8.595      76.577
x2            31.2160     14.849      2.102      0.047       0.499      61.933
x3            45.6740     13.599      3.359      0.003      17.541      73.806
x4            43.6550     14.919      2.926      0.008      12.792      74.518
==============================================================================
Omnibus:                        0.197   Durbin-Watson:                   2.255
Prob(Omnibus):                  0.906   Jarque-Bera (JB):                0.104
Skew:                          -0.127   Prob(JB):                        0.949
Kurtosis:                       2.841   Cond. No.                         275.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

ANOVA results
             df        sum_sq       mean_sq          F    PR(>F)
x1          1.0  1.818785e+08  1.818785e+08   4.048331  0.056071
x2          1.0  2.762906e+08  2.762906e+08   6.149796  0.020900
x3          1.0  6.476995e+08  6.476995e+08  14.416776  0.000930
x4          1.0  3.846678e+08  3.846678e+08   8.562103  0.007596
Residual   23.0  1.033316e+09  4.492679e+07        NaN       NaN
```

From the program output, the p-value and F for each variables are:

Variable	p-value	F
x1	0.016	4.048
x2	0.047	6.150
x3	0.003	14.417
x4	0.008	8.562
Intercept	0.183	N/A

R^2 = 0.591. I think R^2 is not very good, but OK. First, R^2 > 0.5, so it indicates there are more than 50% of the variation in the “resp” variable can be explained by the explanatory variables x1, x2, x3, and x4. However, on the other hand, the R^2 < 0.6 shows that, there are still more than 0.4 variation in the “resp” variable cannot be

explained by the linear regression model. So I think this R^2 is not very good but kind of OK.

4e. Regression Equation:

resp = -3064.5092 + (42.5859)x1 + (31.2160)x2 + (45.6740)x3 + (43.6550)x4

4f. P-values for each regression coefficients and whether they are statistically significant at alpha = 0.05 significant level:

Variable	p-value	Is statistically significant at α = 0.05 ?
x1	0.016	Yes
x2	0.047	Yes (but pretty close to 0.05)
x3	0.003	Yes
x4	0.008	Yes
Intercept	0.183	No

I would consider keeping variables x1, x3, and x4 because their p-values are much less than the significance level 0.05, indicating that they are statistically significant that they are not equal to 0.

I would consider discarding the other variables x2 and Intercept. The intercept has pretty high p-value of 0.183, which is shows that it is not statistically significant than 0. Even though the p-value for x2 is 0.047 < 0.05, but it is pretty close to 0.05. In fact, if we round up its p-value to the 2nd decimal place or increase the significant level to 0.04, we would have concluded that the coefficient for x2 is NOT statistically significant. The

p-value 0.047 pretty close to 0.05 indicates its contribution to variation of the “resp” variable is pretty weak.

4g. Conclusion

From the linear regression analysis above, we can see that, the “resp” variable indeeds depends on some of the “X_i” explanatory variables. On the other hand, we can see that, some explanatory variables have no/weak relationship with “resp”. Through the analysis of the p-value of each coefficients, we can identify that x1, x3, and x4 variables indeed have positive linear relationships with the “resp” dependent variable, while the intercept and x2 do not show significant relationship with the “resp” variable. Therefore, when I re-run the linear regression with only the x1, x3, and x4 variables, the R^2 improved to 0.655. (previous is 0.591) , which matches the above analysis.

Re-Run Linear Regression with x1, x3, and x4 only:

OLS Regression Results

Dep. Variable:	resp	R-squared:	0.655
Model:	OLS	Adj. R-squared:	0.614
Method:	Least Squares	F-statistic:	15.82
Date:	Tue, 06 Nov 2018	Prob (F-statistic):	5.67e-06
Time:	15:36:14	Log-Likelihood:	-286.67
No. Observations:	28	AIC:	579.3
Df Residuals:	25	BIC:	583.3
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	29.7981	15.971	1.866	0.074	-3.095	62.691
x3	44.7717	14.007	3.196	0.004	15.925	73.619
x4	33.8267	12.277	2.755	0.011	8.541	59.113

Omnibus:	3.994	Durbin-Watson:	2.186
Prob(Omnibus):	0.136	Jarque-Bera (JB):	2.494
Skew:	0.683	Prob(JB):	0.287
Kurtosis:	3.520	Cond. No.	2.03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

ANOVA results

	df	sum_sq	mean_sq	F	PR(>F)
x1	1.0	9.198705e+08	9.198705e+08	17.950518	0.000269
x3	1.0	1.123030e+09	1.123030e+09	21.915016	0.000085
x4	1.0	3.890006e+08	3.890006e+08	7.591027	0.010783
Residual	25.0	1.281120e+09	5.124479e+07	NaN	NaN