Andrew Curtis

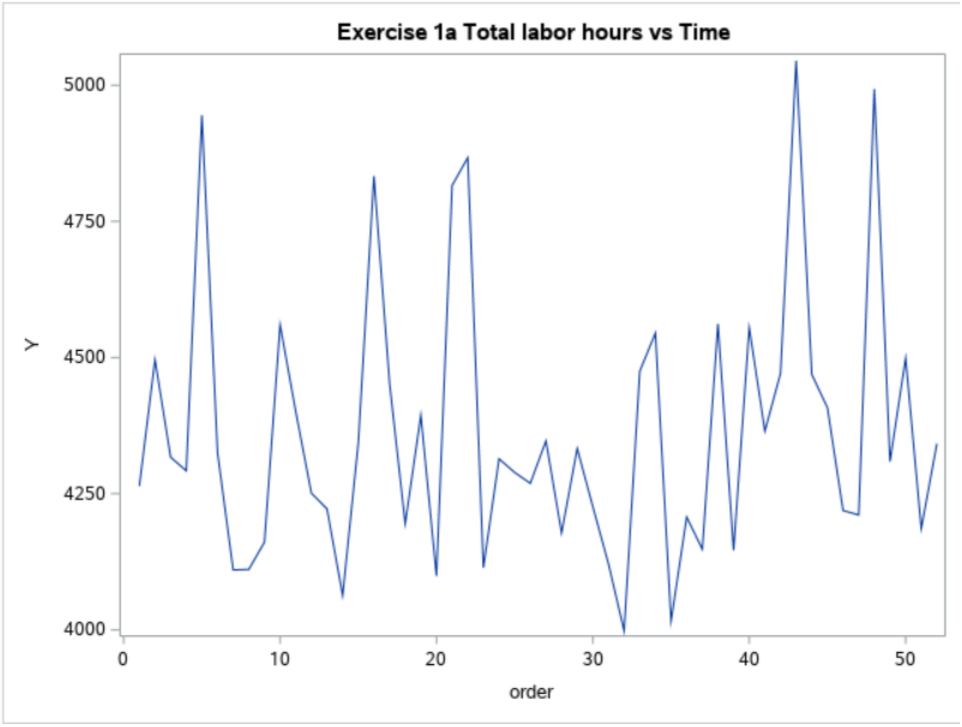
STAT 5100

Spring 2022

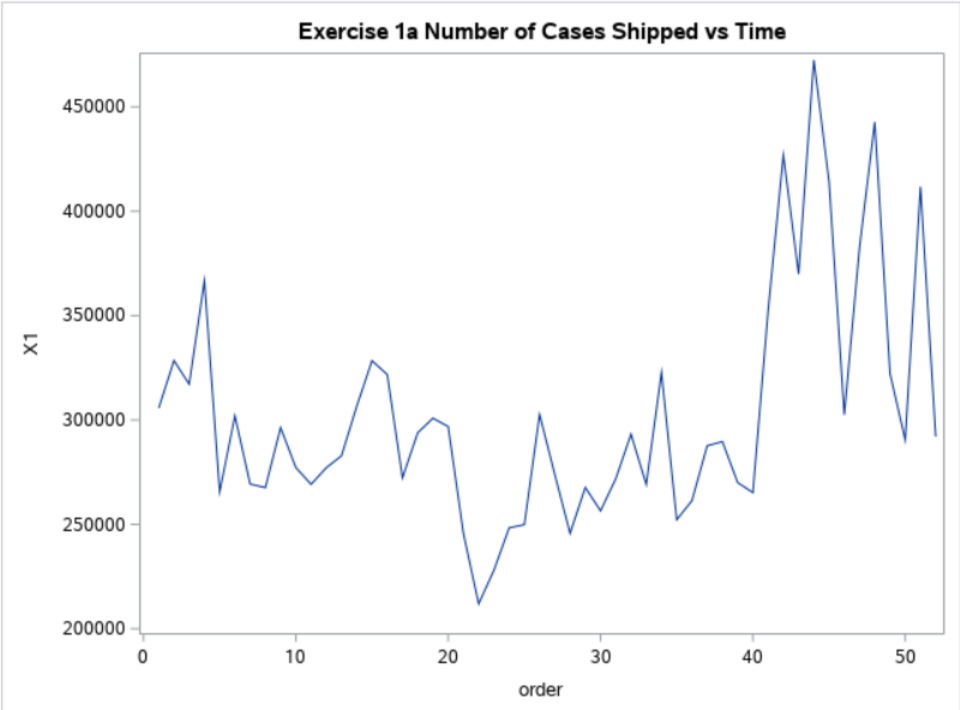
Homework #4

Exercise 1

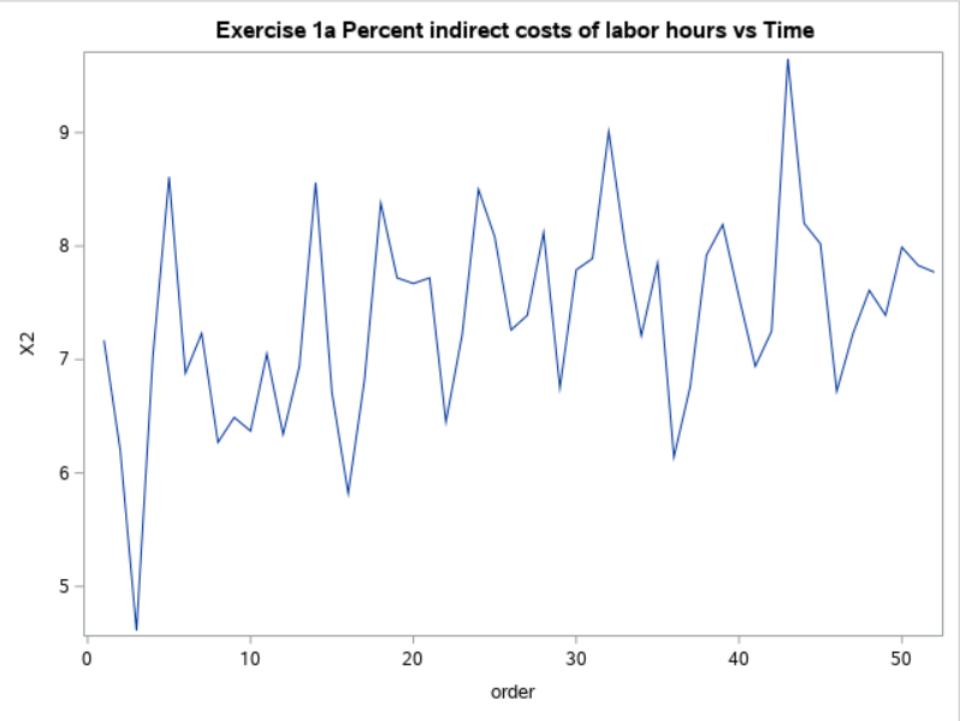
a) Provide sequence plots for all three predictors and the response, and comment on any patterns observed.



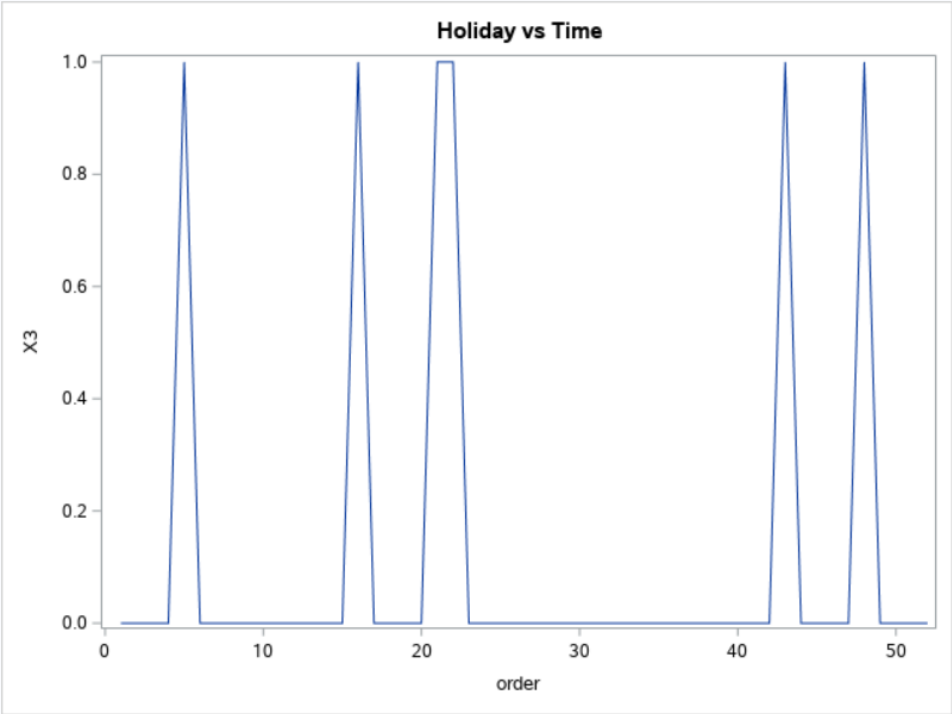
The response, total labor hours, seems to bounce around a mean with some high and low outliers throughout the year.



The number of cases shipped is fairly constant except for some larger weeks towards the end of the year.

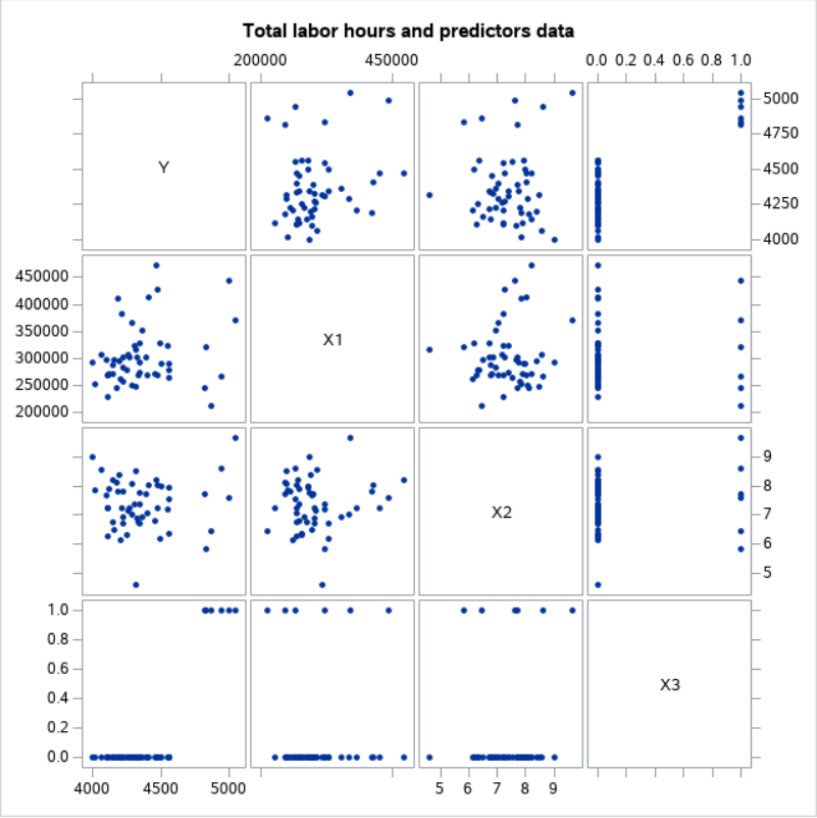


The indirect costs of labor hours as a percentage looks like it might have a linear correlation over the course of a year. There is some variability however.



Holiday vs time should show where the holiday weeks lie. 1 if a holiday, 0 otherwise.

b) Provide a scatterplot matrix and correlation matrix for all three predictors and the response, and comment on what these suggest.



The x1 variable may have some positive correlation with the y response, but linearity is not readily apparent. X2 does not appear to have a strong correlation with y. x3 looks as though it does have some positive correlation with y.

Pearson Correlation Coefficients, N = 52
Prob > |r| under H0: Rho=0

	Y	X1	X2	X3
Y	1.00000	0.20766 0.1396	0.06003 0.6725	0.81058 <.0001
X1	0.20766 0.1396	1.00000	0.08490 0.5496	0.04566 0.7479
X2	0.06003 0.6725	0.08490 0.5496	1.00000	0.11337 0.4236
X3	0.81058 <.0001	0.04566 0.7479	0.11337 0.4236	1.00000

The x3 variable has much higher correlation with y than the other predictor variables. It is the only values with p-value < alpha = 0.05 as well.

Exercise 2

a) Fit the multiple regression and report the estimated regression function. Provide an interpretation of each of the estimated regression coefficients.

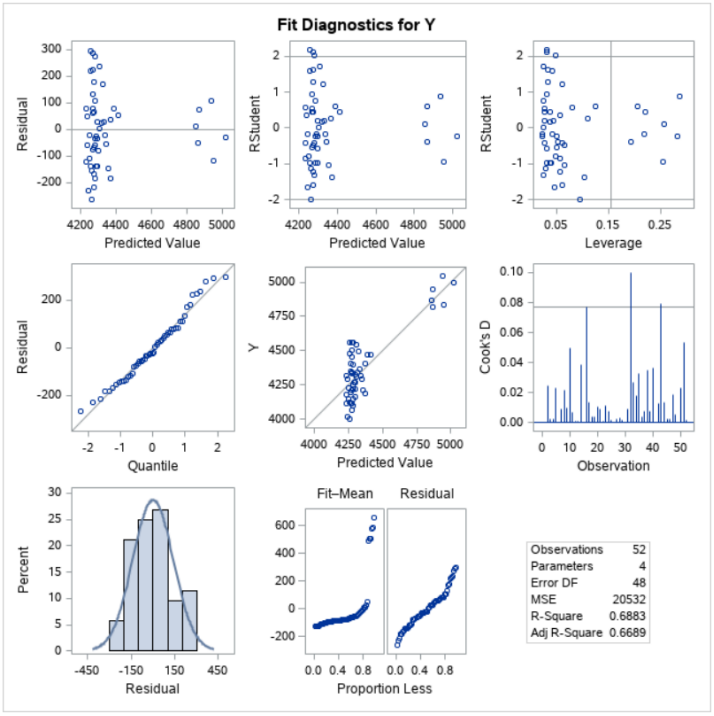
$$Y = 4149.887 + 0.00078708X_1 - 13.16602X_2 + 623.55448X_3 + \varepsilon$$

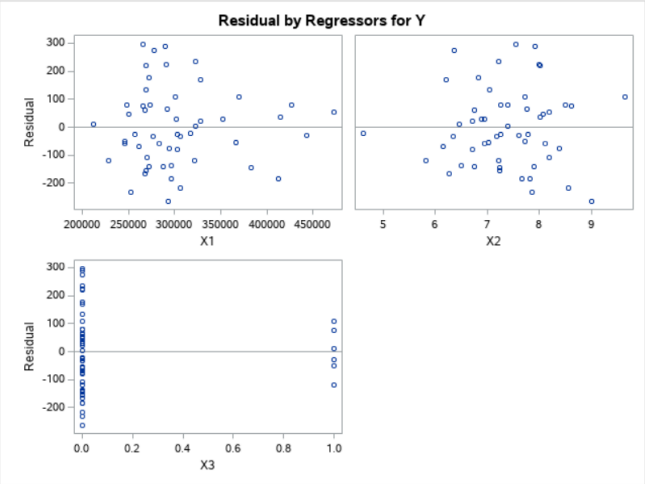
B_1 is very small and there does not appear to be a strong correlation between it and the response variable.

B_2 is negative and may have some negative correlation with the response variable.

B_3 is large and positive and seems to have the strongest correlation of the predictors with the response variable.

b) Report graphical (histogram, normal prob. plot, sequence plot, residual plot) and numerical checks (Brown-Forsythe and correlation test of normality) of model assumptions, and comment briefly on what these suggest.





P-value for Brown-Forsythe test of constant variance in residual vs. predicted

Obs	t_BF	BF_pvalue
1	2.66204	0.010418

Output for correlation test of normality of residual (Check text Table B.6 for threshold)

The CORR Procedure

2 Variables:	resid expectNorm

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
resid	52	0	139.01120	0	-264.05302	295.75182	residual
expectNorm	52	0	0.98204	0	-2.25836	2.25836	

Pearson Correlation Coefficients, N = 52 Prob > |r| under H0: Rho=0		
	resid	expectNorm
resid residual	1.00000	0.99087 <.0001
expectNorm	0.99087 <.0001	1.00000

For the graphical diagnostics, the residuals appear to show normality. Notice, the histogram looks fairly gaussian distributed and residuals do not show any trends in residual plots. The p-value for test of constant variance is above threshold so fail to reject that there is constant variance. As for test of normality of residuals, the expect norm value is above the related table value, so fail to reject normality.

Exercise 3

Text problem 6.13. (Report both Bonferroni and Scheffe intervals and indicate which is more efficient).

Simultaneous 90% intervals of individual prediction at three x-profiles, using Scheffe and Bonferroni

Obs	X1	X2	X3	Yhat	S_lower	S_upper	B_lower	B_upper
53	230000	7.5	0	4232.17	3853.64	4610.70	3909.45	4554.89
54	250000	7.3	0	4250.55	3875.18	4625.91	3930.52	4570.57
55	280000	7.1	0	4276.79	3903.80	4649.78	3958.79	4594.79
56	340000	6.9	0	4326.65	3951.61	4701.69	4006.91	4646.39

The Bonferrroni prediction intervals are more efficient for the test values.

Exercise 4

a)

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Squared Partial Corr Type I
Intercept	1	4149.88721	195.56541	21.22	<.0001	989877440	.
X1	1	0.00078708	0.00036455	2.16	0.0359	136366	0.04312
X3	1	623.55448	62.64095	9.95	<.0001	2033565	0.67208
X2	1	-13.16602	23.09173	-0.57	0.5712	6674.58809	0.00673

SSR(X1) = 136,366

$SSR(X_3|X_1) = SSE(X_1) - SSE(X_1, X_3)$

$SSR(X_3|X_1) = 2{,}033{,}565$

$SSR(X_2|X_1, X_3) = SSE(X_1, X_3) - SSE(X_1, X_2, X_3)$

$SSR(X_2|X_1, X_3) = 6{,}674$

b)

Subset F-test, automatically

The REG Procedure
Model: MODEL1

Test subsetcheck Results for Dependent Variable Y

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	6674.58809	0.33	0.5712
Denominator	48	20532		

For the F-test with $\beta_2 = 0$, the F-value is equal to 0.33, which is small. This leads to a large p-value. This means fail to reject null hypothesis which is that $\beta_2 = 0$. Meaning that predictor is likely to not be affecting the model.

c)

X1 X2 Regression
(X1 then X2 model)

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Number of Observations Read	52
Number of Observations Used	52

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	142092	71046	1.15	0.3242
Error	49	3020044	61634		
Corrected Total	51	3162136			

Root MSE	248.26104	R-Square	0.0449
Dependent Mean	4363.03846	Adj R-Sq	0.0060
Coeff Var	5.69010		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Squared Partial Corr Type I
Intercept	1	3995.47867	337.76602	11.83	<.0001	989877440	.
X1	1	0.00091916	0.00063120	1.46	0.1517	136366	0.04312
X2	1	12.12052	39.76555	0.30	0.7618	5725.92181	0.00189

X2 X1 Regression
(X2 then X1 model)

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Number of Observations Read	52
Number of Observations Used	52

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	142092	71046	1.15	0.3242
Error	49	3020044	61634		
Corrected Total	51	3162136			

Root MSE	248.26104	R-Square	0.0449
Dependent Mean	4363.03846	Adj R-Sq	0.0060
Coeff Var	5.69010		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Squared Partial Corr Type I
Intercept	1	3995.47867	337.76602	11.83	<.0001	989877440	.
X2	1	12.12052	39.76555	0.30	0.7618	11395	0.00360
X1	1	0.00091916	0.00063120	1.46	0.1517	130697	0.04148

SSR(X1)+SSR(X2|X1)=SSR(X2)+SSR(X1|X2)? This statement is true and should be true every time. The order does not matter as long as both are considered. This could be considered a cumulative property.

Check of statement, 136,366 + 5,725 = 11,395 + 130,697

Exercise 5

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4149.88721	195.56541	21.22	<.0001	0
X1	1	0.00078708	0.00036455	2.16	0.0359	1.00860
X2	1	-13.16602	23.09173	-0.57	0.5712	1.01960
X3	1	623.55448	62.64095	9.95	<.0001	1.01436

Collinearity Diagnostics						
			Proportion of Variation			
Number	Eigenvalue	Condition Index	Intercept	X1	X2	X3
1	3.13677	1.00000	0.00102	0.00308	0.00133	0.02004
2	0.83411	1.93923	0.00032068	0.00091374	0.00035401	0.97010
3	0.02283	11.72211	0.03448	0.88968	0.16309	0.00033831
4	0.00629	22.32696	0.96418	0.10632	0.83523	0.00953

The variance inflation numbers for each parameter are low and near 1. This indicates that maybe no multicollinearity is taking place. As for the condition indexes, all are less than 30, this also indicates that probably no multicollinearity taking place.

Appendix SAS code

See attached pdf of SAS code to homework submission