UNIVERSITY TUNKU ABDUL RAHMAN

FACULTY OF SCIENCE

ACADEMIC YEAR 2019 / 2020

UDPS2223 APPLIED REGRESSION ANALYSIS

ASSIGNMENT

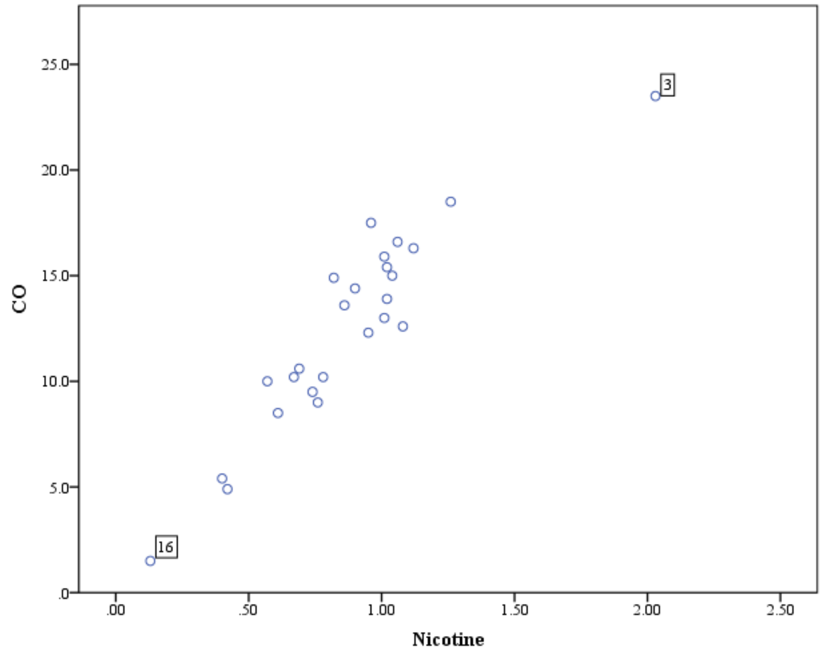
Name **:** Ngu Yi Hui

ID **:** 18ADB01438

Course **:** SCOR

Section A

a) **Scatter Plot of CO against Nicotine**



Observation 3 (Bull Durham) is considered as an unusual point based on the scatter plot. Besides that, observation 16 (Now) can also be considered as a potential unusual point. This is because they are far away from other points. Since the data set is small, the unusual points might affect the slope of the regression line and ultimately cause poor estimation.

b) **Correlation between CO and Nicotine**

Correlations

		Nicotine	CO
Nicotine	Pearson Correlation	1	.926**
	Sig. (2-tailed)		.000
	N	25	25
CO	Pearson Correlation	.926**	1
	Sig. (2-tailed)	.000	
	N	25	25

**. Correlation is significant at the 0.01 level (2-tailed).

The correlation between CO and Nicotine is 0.926. This indicates that they have a very strong positive linear correlation. From the table above, there is a significant correlation between CO and Nicotine at 0.01 level.

c) **Simple linear regression (SLR) model**

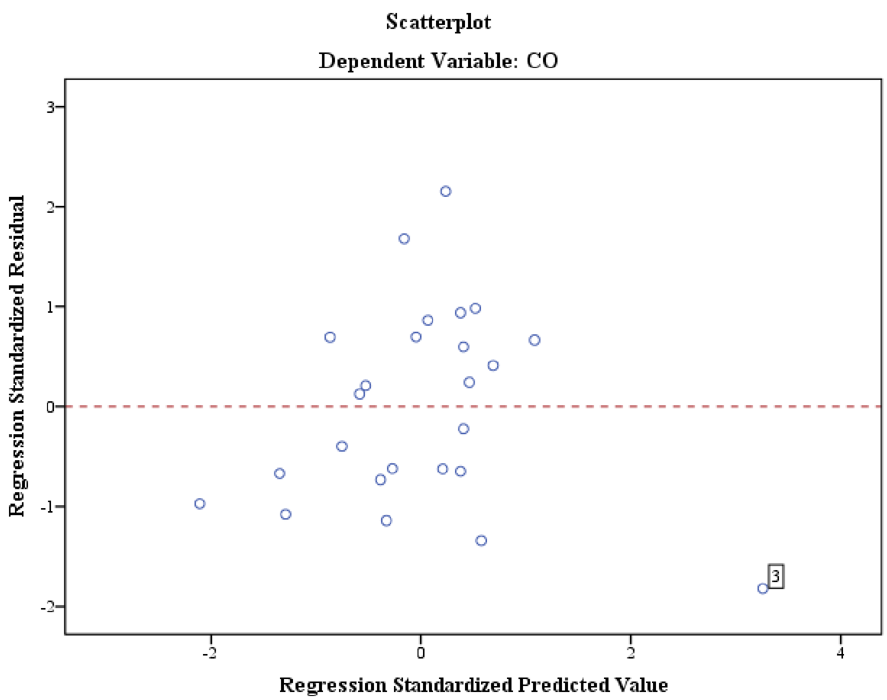
Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	1.665	.994		1.675	.107	-.391	3.720
	Nicotine	12.395	1.054	.926	11.759	.000	10.215	14.576

a. Dependent Variable: CO

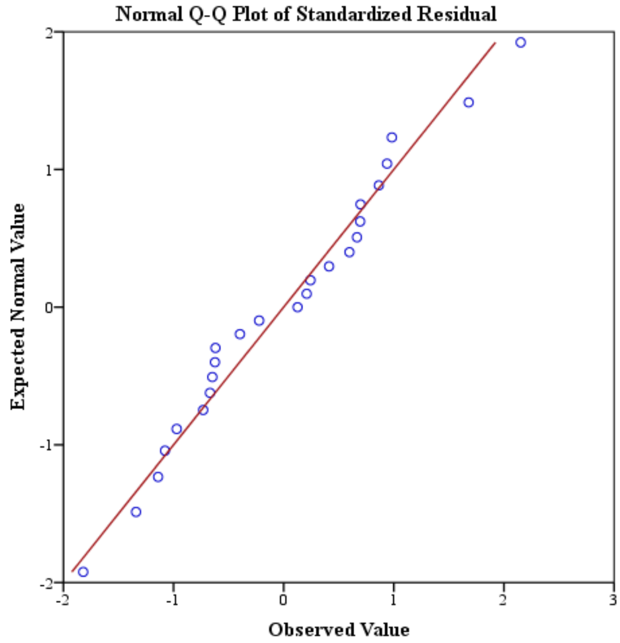
Let $\hat{y}$ = CO, x = Nicotine,
$\hat{y} = 1.665 + 12.395x$

Plot of the residuals against the fitted values



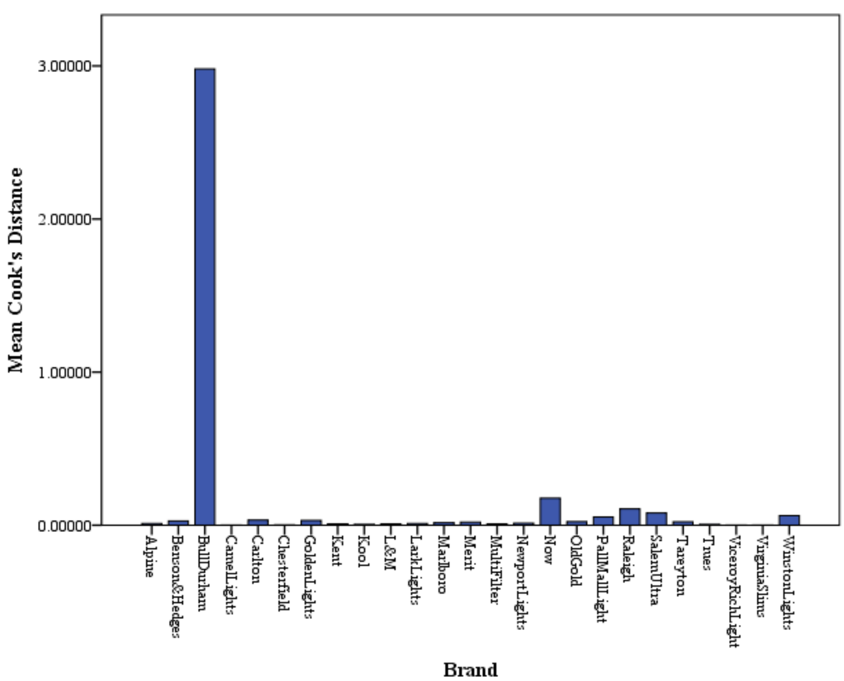
From the scatter plot above, we can observe some pattern. For the smaller nicotine value and larger nicotine value, the standardized residuals tend to be negative values. This shows that the error variance is not constant. Moreover, the regression model might not be linear. The normality of the error terms can be further determined through the normal Q-Q plot. Furthermore, we can observe one outlier in this plot which is the observation 3 (Bull Durham).

Normal Q-Q plot of the standardized residuals



The normal Q-Q plot of the residuals shows some minor up-and-down in the curve. However, the points are not greatly deviated from a straight line, thus the normality assumption is considered not violated.

Bar plot of Cook's distances



The 3rd observation (Bull Durham) has a very high Cook's distance value compared with others. Thus, Bull Durham is an unusual data point.

d) The 3rd observation (Bull Durham) is removed from the data.

Simple linear regression (SLR) model

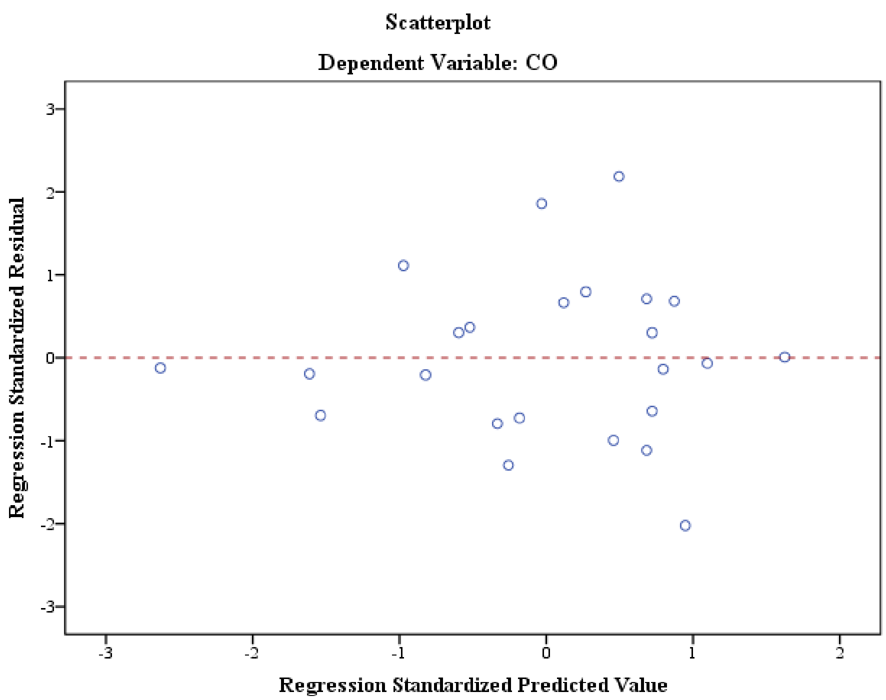
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.238	1.083		-.220	.828	-2.484	2.007
	Nicotine	14.860	1.247	.931	11.916	.000	12.274	17.446

a. Dependent Variable: CO

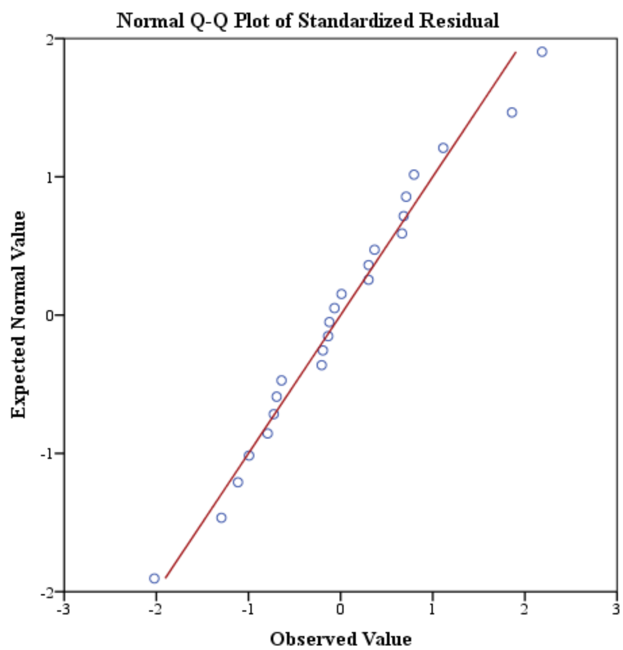
Let $\hat{y}$ = CO, x = Nicotine,
$\hat{y} = -0.238 + 14.860x$

Plot of the residuals against the fitted values



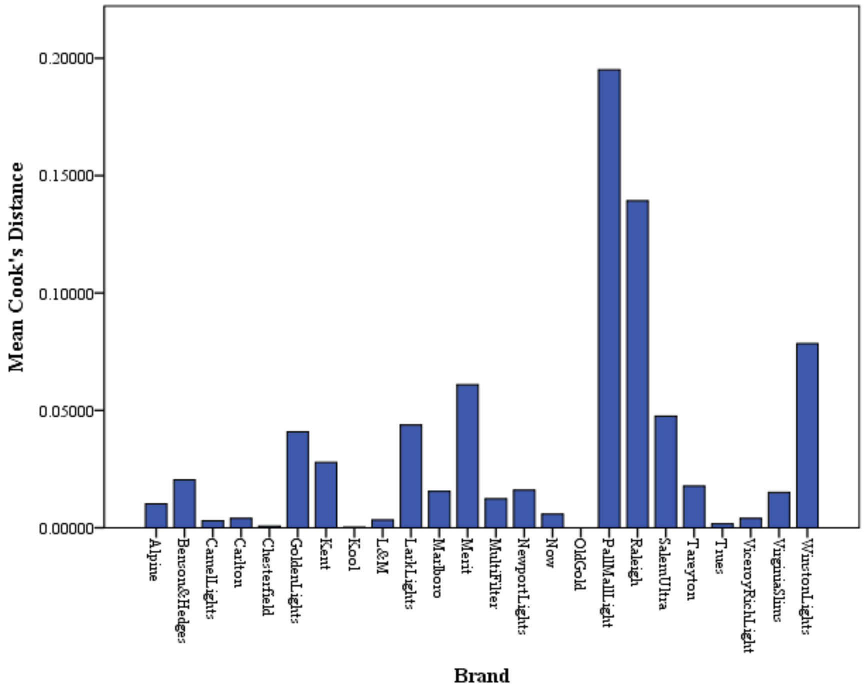
Notice the funnel shape appears in the scatter plot above indicating a problem with the constant variance assumption. The error variance is larger for the larger nicotine value. Hence, the problem of non-constancy of error variance has not been solved after removing the outlier. On the other hand, the linearity of the regression model is met based on the plot above.

Normal Q-Q plot of the standardized residuals



From the normality probability plot above, we find that there is no strong indications of substantial departures from normality are indicated. As compared with previous plot in part (c), we found that it has a better normality after removing the unusual observation.

Bar plot of Cook's distances



From the Cook's distance plot above, we observe there is no potential outlier in the dataset after removing the brand Bull Durham. The problem of having outliers is solved.

Section B

a) **ANOVA (Analysis of Variance) table**

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
	Regression	358.242	1	358.242	141.986	.000[b]
1	Residual	55.508	22	2.523		
	Total	413.750	23			

a. Dependent Variable: CO
b. Predictors: (Constant), Nicotine

Critical F-value = 4.3 < 141.986

At α = 0.05, there is statistically significance evidence that the model using Nicotine as predictor variable is useful for estimating CO.

b) **Coefficient of determination**

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.931[a]	.866	.860	1.5884

a. Predictors: (Constant), Nicotine
b. Dependent Variable: CO

R^2 = 0.866, indicates that 86.6% of the variation in CO can be explained by Nicotine, that is using Nicotine to predict CO.

c) Let $\hat{y}$ = CO, x_1 = Tar, x_2 = Nicotine, x_3 = Weight

Firstly, we are interested to know whether the model changes depending on the order when we include the same explanatory variables. From the output run by SPSS (*Output 3*), we found out that the model does not change depending on the order when we include the same explanatory variables. Same regression model and ANOVA table will be obtained when the same explanatory variables are used. Thus, the predicted value ($\hat{y}$) cannot be different with different orders of the predictors, as long as the same predictors are used every time.

However, the sequential analysis of variance, which is adding one term after the other, will cause the result to be different.

CO against Tar
$\hat{y} = 2.743 + 0.801x_1$

Critical F-value = 4.279 < F = 253.370

At α = 0.05, there is statistically significance evidence that the model using Tar as predictor variable is useful for estimating CO.

CO against Nicotine
$\hat{y} = 1.665 + 12.395x_2$

Critical F-value = 4.279 < F = 138.266

At α = 0.05, there is statistically significance evidence that the model using Nicotine as predictor variable is useful for estimating CO.

CO against Weight
$\hat{y} = -11.795 + 25.068x_3$

Critical F-value = 4.279 < F = 6.309

At α = 0.05, there is statistically significance evidence that the model using Weight as predictor variable is useful for estimating CO.

CO against Tar and Nicotine
$\hat{y} = 3.090 + 0.962x_1 - 2.646x_2$

Critical F-value = 3.443 < F = 124.110

At α = 0.05, there is statistically significance evidence that the model using Tar and Nicotine as predictor variables is useful for estimating CO.

CO against Tar and Weight
$\hat{y} = 3.114 + 0.804x_1 - 0.423x_3$

Critical F-value = 3.443 < F = 121.251

At α = 0.05, there is statistically significance evidence that the model using Tar and Weight as predictor variables is useful for estimating CO.

CO against Nicotine and Weight
$\hat{y} = 1.614 + 12.388x_2 + 0.059x_3$

Critical F-value = 3.443 < F = 66.128

At α = 0.05, there is statistically significance evidence that the model using Nicotine and Weight as predictor variables is useful for estimating CO.

CO against Tar, Nicotine and Weight
$\hat{y} = 3.202 + 0.963x_1 - 2.632x_2 - 0.130x_3$

Critical F-value = 3.072 < F = 78.984

At α = 0.05, there is statistically significance evidence that the model using Tar, Nicotine and Weight as predictor variables is useful for estimating CO.

Correlations

		CO	Tar	Nicotine	Weight
Pearson Correlation	CO	1.000	.957	.926	.464
	Tar	.957	1.000	.977	.491
	Nicotine	.926	.977	1.000	.500
	Weight	.464	.491	.500	1.000
Sig. (1-tailed)	CO	.	.000	.000	.010
	Tar	.000	.	.000	.006
	Nicotine	.000	.000	.	.005
	Weight	.010	.006	.005	.

From the ANOVA tables developed, we observed that although the model using Weight as predictor variable is statistically significant, it is always insignificant when Tar or Nicotine is already given in the model at α = 0.05. This indicates that Weight might not be a useful explanatory variable for CO when Tar or Nicotine is already in the model at α = 0.05.

Besides that, from the correlation table above, Tar and Nicotine are highly correlated with Pearson Correlation coefficient 0.977, which means they might be redundant independent variables in this dataset.

In addition, we observed that when Tar and Nicotine is used together in the model to estimate CO, the coefficient of Nicotine will become nearly insignificant at α = 0.05. After considering the high correlation between Tar and Nicotine, I would like to exclude Nicotine from the model.

Therefore, I will choose Tar as the only explanatory variable for CO, and finally obtain the equation $CO = 2.743 + 0.801\ Tar$.

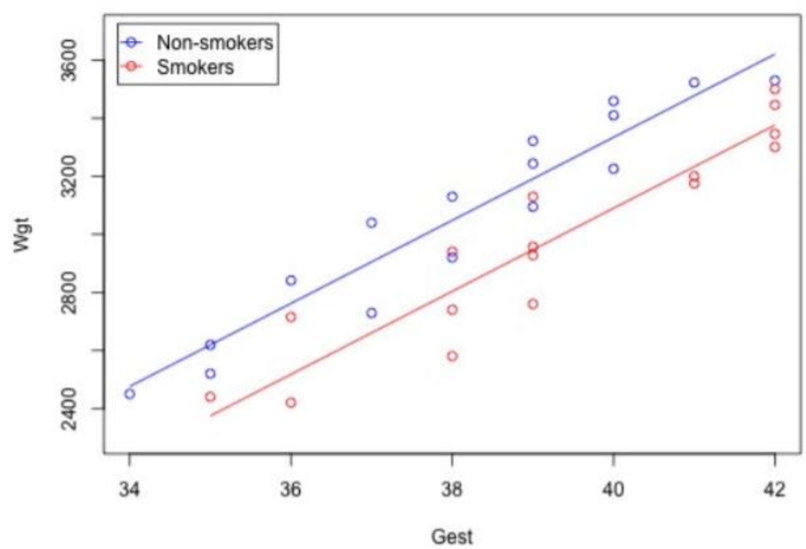
Section C

Categorical variables can be applied in building regression model. In this essay, we are going to look at the example on “Birth weight and Smoking during pregnancy”. We hope to know the relationship between the smoking status, the length of gestation, with baby birth weight.

A random sample of n = 32 births is collected. *Table 1* shows the dataset collected.

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$
where:
y_i is the birth weight of baby i,
x_{i1} is the length of gestation of baby i,
x_{i2} is the smoking status of mother of baby i, it is coded as 1 if she smoked during pregnancy, and 0 if she did not.

Based on the sample data, the plot of the estimated regression function looks like:



The blue circles represent the data on non-smoking mothers, while the red circles represent the data on smoking mothers. Meanwhile, the blue line represents the estimated linear relationship between length of gestation and birth weight for non-smoking mothers, while the red line represents the estimated linear relationship for smoking mothers.

A regression equation is obtained after performing calculation:
$\hat{y}_i = -2390 + 143\, x_{i1} - 245\, x_{i2}$

Therefore, the estimated regression equation for non-smoking mothers (x_{i2} = 0) is:
$\hat{y}_i = -2390 + 143\, x_{i1}$ ---- *equation 1*
and the estimated regression equation for smoking mothers (x_{i2} = 1) is:
$\hat{y}_i = -2635 + 143\, x_{i1}$ ---- *equation 2*

The regression equations above help us to predict the birth weight of the baby. When the smoking status of the mother is given as non-smoking, we can use *equation 1* above to estimate the birth weight. On the other hand, when the smoking status of the mother is given as smoking, we can use *equation 2* above to predict the birth weight.

Upon analyzing the data, the output:

Regression Analysis: Birth Weight versus Gestation, Smoke

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	3348720	1674360	125.45	0.000
Gest	1	3280270	3280270	245.76	0.000
Smoke	1	452881	452881	33.93	0.000
Error	29	387070	13347		
Lack-of-Fit	12	52383	4365	0.22	0.994
Pure Error	17	334687	19687		
Total	31	3735790			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
115.530	89.64%	88.92%	87.6%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-2390	349	-6.84	0.000	
Gest	143.10	9.13	15.68	0.000	1.06
Smoke	-244.5	42.0	-5.83	0.000	1.06

From the result, we can conclude that:

The *p*-value for the analysis of variance F-test ($p < 0.001$) suggests that the model containing length of gestation and smoking status is more useful in predicting birth weight than not taking into account the two predictors. The length of gestation and the smoking status of the mother in the model successfully explain 89.64% of the variation in the birth weights of babies.

Furthermore, the p-values for the t-tests appearing in the table of estimates suggest that the slope parameters for the length of gestation ($p < 0.001$) and the smoking status of mothers ($p < 0.001$) are significantly different from 0. This indicates that there is statistically significance evidence that the length of gestation is related to the birth weights of babies in the model, given that the smoking status of mothers already in the model at $\alpha = 0.5$. The smoking status of mothers is also related to the birth weights of babies in the model, given that the length of gestation already in the model at $\alpha = 0.5$.

Table 1. Birth and Smokers dataset

Birth Weight (gram)	Gestation (week)	Smoke
2940	38	yes
3130	38	no
2420	36	yes
2450	34	no
2760	39	yes
2440	35	yes
3226	40	no
3301	42	yes
2729	37	no
3410	40	no
2715	36	yes
3095	39	no
3130	39	yes
3244	39	no
2520	35	no
2928	39	yes
3523	41	no
3446	42	yes
2920	38	no
2957	39	yes
3530	42	no
2580	38	yes
3040	37	no
3500	42	yes
3200	41	yes
3322	39	no
3459	40	no
3346	42	yes
2619	35	no
3175	41	yes
2740	38	yes
2841	36	no