MKT 6337.002 : **Predictive Analytics using SAS**
Homework 3 Solutions
Group 11

Question1: -

1. **Run a regression model and interpret the coefficients. Comment on the model fit.**

Answer: Before that we converted kids to 4 dummy variables as kids1 having 1 kid kids2 having 2 kids and so on. 0 kids is our base group. Also, we checked for any missing values, multicollinearity and outliers before running regression.

Missing values: No missing values were found.

The SAS System

The MEANS Procedure

Variable	Label	N	N Miss
miles	miles traveled per year	200	0
income	annual income ($1000)	200	0
age	average age of adult members of household	200	0
kids	number of children in household	200	0

Multicollinearity: To check for multicollinearity, we included VIF (Variance Inflation Factor) and Collin.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	Intercept	1	-415.83801	177.05720	-2.35	0.0199	0	0
income	annual income ($1000)	1	14.10027	1.81930	7.75	<.0001	0.46765	1.06871
age	average age of adult members of household	1	16.04028	3.83878	4.18	<.0001	0.27128	1.23725
kids1		1	-26.52142	94.26413	-0.28	0.7787	-0.01976	1.44780
kids2		1	-162.09012	95.97043	-1.69	0.0928	-0.12370	1.57471
kids3		1	-215.52586	95.83431	-2.25	0.0256	-0.16693	1.61724
kids4		1	-351.69115	148.90030	-2.36	0.0192	-0.15147	1.20719

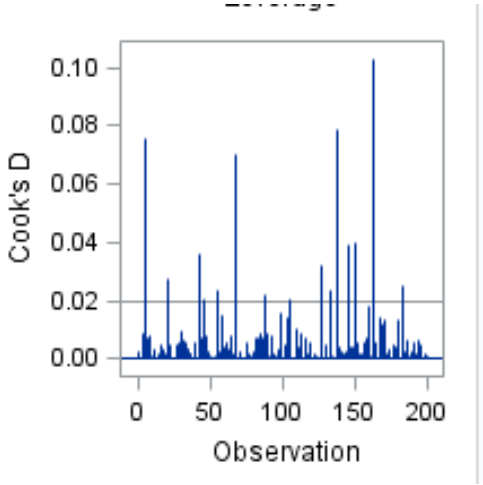
From the above output we can conclude that there is no multicollinearity 'Variance Inflation' does not have any value greater than 10, therefore independent variables are not highly correlated.

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation						
			Intercept	income	age	kids1	kids2	kids3	kids4
1	3.73426	1.00000	0.00230	0.00480	0.00256	0.00869	0.00938	0.00983	0.00381
2	1.00240	1.93011	0.00001939	0.00001948	0.00005890	0.38898	0.04729	0.04178	0.07173
3	1.00034	1.93209	0.00000494	0.00001466	1.130304E-7	0.00576	0.03727	0.21667	0.35147
4	1.00000	1.93242	0	0	0	0.00279	0.25060	0.05006	0.29265
5	0.19547	4.37080	0.01127	0.04730	0.00773	0.56471	0.59407	0.62740	0.23773
6	0.04723	8.89230	0.08248	0.90437	0.20895	0.01143	0.04858	0.03023	0.03634
7	0.02029	13.56583	0.90393	0.04350	0.78070	0.01764	0.01281	0.02403	0.00627

Also Using Collin, we can infer that as **condition index** is less than 100 and so variables do not have high proportion of variance and variables are linearly independent.

Outliers:



There are 5 outlier in the data, concluded by analyzing the plot between Cook's D and observations. Threshold is 3*average.

Regression:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	20828253	3471376	16.76	<.0001
Error	193	39983516	207168		
Corrected Total	199	60811769			

Root MSE	455.15764	R-Square	0.3425
Dependent Mean	1054.23000	Adj R-Sq	0.3221
Coeff Var	43.17442		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-415.83801	177.05720	-2.35	0.0199
income	annual income ($1000)	1	14.10027	1.81930	7.75	<.0001
age	average age of adult members of household	1	16.04028	3.83878	4.18	<.0001
kids1		1	-26.52142	94.26413	-0.28	0.7787
kids2		1	-162.09012	95.97043	-1.69	0.0928
kids3		1	-215.52586	95.83431	-2.25	0.0256
kids4		1	-351.69115	148.90030	-2.36	0.0192

Variables age, income, kids3, kids4 are significant variables.

Interpretations:

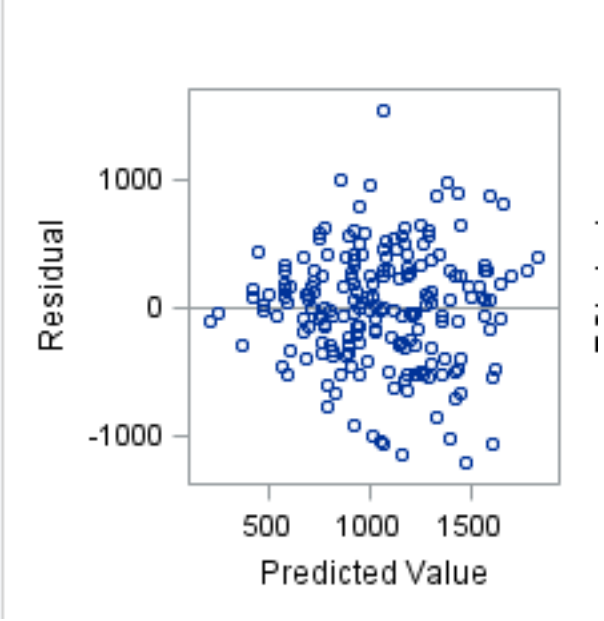
1) **Income**: For every $1000 increase in annual income, the estimated miles traveled for vacation per year increases by 14.10027 miles keeping all the other parameters constant.
2) **Age**: For every 1-year increase in average age of adult members of the household, the estimated miles traveled for vacation per year increases by 16.04028 miles keeping all the other parameters constant.
3) **Kids3**: For households with 3 kids, the estimated miles traveled for vacation per year decreases by 215.525 miles with reference to household with no kids keeping all the other parameters constant.
4) **Kids4**: For households with 4 kids, the estimated miles traveled for vacation per year decreases by 351.6911 miles with reference to household with no kids keeping all the other parameters constant.

Model Fit: -

The model has an adjusted R2 =0.3222, which means 32.22% of variance in the dependent variable miles per year is explained by the explanatory variables. Adjusted R2 lies in the range 0 to 1 with values closer to 1 indicating a very good model fit. Our model has a reasonable fit. In practical scenario there could be other factors that affect miles travelled per year.

2. **Check whether there is heteroscedasticity in the model using White test**.

Answer: To check for heteroskedasticity, we analyzed the plot of residuals Vs predictions.



From the plot we can conclude that variance of residuals increases with corresponding increase in predicted y, therefore heteroskedasticity is present.

Checking with White correction test:

White Test Hypothesis: -

Null hypothesis i.e. H0: - Errors are Homoscedastic, variance of error terms is constant

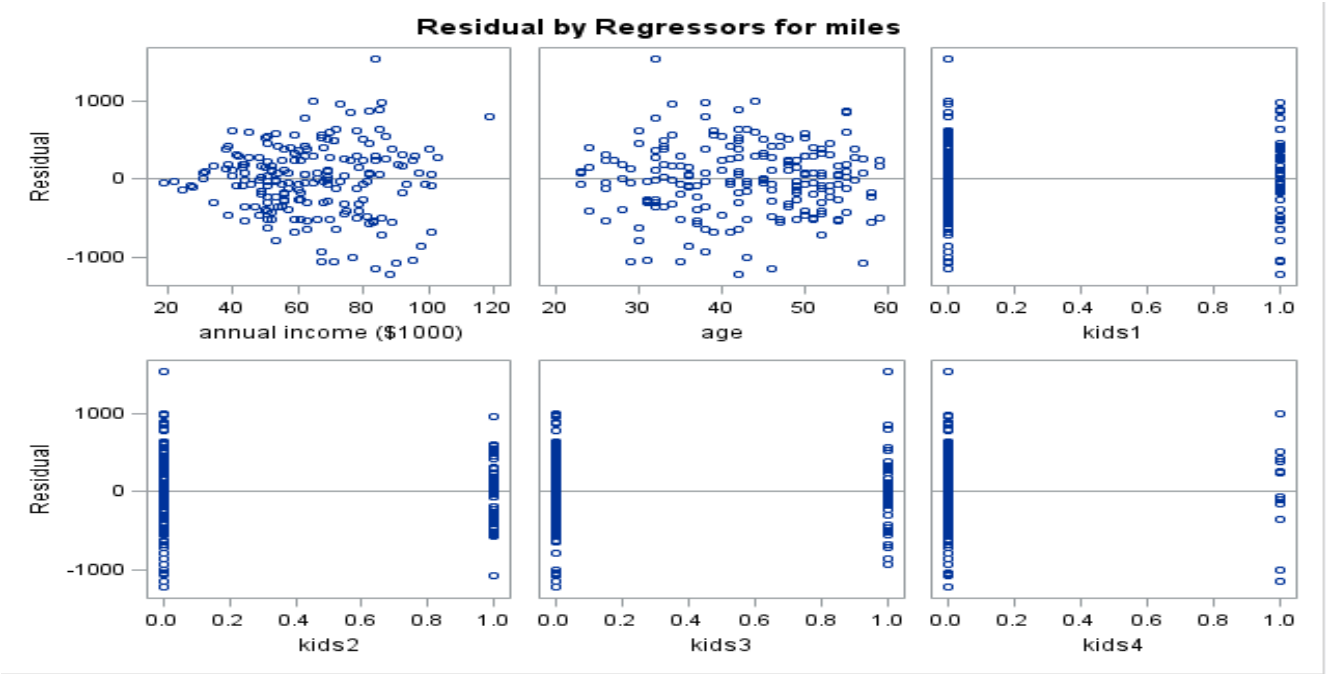
Alternative hypothesis i.e. H1: - Errors are Heteroscedastic, variance of error terms is not constant

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
miles	White's Test	54.07	17	<.0001	Cross of all vars

Since the p value is less than 0.05, we reject the null hypothesis that errors are homoscedastic. So, we can conclude that there is heteroscedasticity in the model.

3. **Run a weighted Least squares (WLS) regression. Discuss your results in a paragraph. (Comment on model fit, significance of coefficients, and the effect of doing WLS.)**

Answer: Below output shows the plot of residuals against all the independent variables showing income variable has high variance and therefore signs of heteroscedasticity. So, we used income as weight in our WLS.



Weighted Least Squares Regression:

Below is the output of WLS with income variable as the weight. We can see that the values of R2 and adjusted R2 have decreased significantly. The R2 has reduced to 0.1473 adjusted R2 has decreased to 0.1208, which means the model had a better fit before the WLS treatment. We can also see that all the variables are still significant except families with 1 kid (Kids1 dummy variable) and families with kids 2 (Kids2 dummy variable).

Nonlinear OLS Summary of Residual Errors

Equation	DF Model	DF Error	SSE	MSE	Root MSE	R-Square	Adj R-Sq
wmiles	7	193	8965.1	46.4510	6.8155	0.1473	0.1208

Nonlinear OLS Parameter Estimates

Parameter	Estimate	Approx Std Err	t Value	Approx Pr > |t|
b0	-439.36	132.5	-3.32	0.0011
b1	16.84299	3.0825	5.46	<.0001
b2	-58.736	73.9447	-0.79	0.4280
b3	-156.657	81.5188	-1.92	0.0561
b4	-226.478	79.5331	-2.85	0.0049
b5	-300.709	114.8	-2.62	0.0095
b6	14.02298	1.5195	9.23	<.0001

Number of Observations		Statistics for System	
Used	200	Objective	44.8253
Missing	0	Objective*N	8965

Heteroscedasticity Test

Equation	Test	Statistic	DF	Pr > ChiSq	Variables
wmiles	White's Test	20.24	17	0.2620	Cross of all vars

We also tried taking **income square** as weight in our model, although it increased the adjusted R-squared to 0.4407, the error terms were heteroscedastic in this case**.**

Nonlinear OLS Summary of Residual Errors								
Equation	DF Model	DF Error	SSE	MSE	Root MSE	R-Square	Adj R-Sq	Label
miles	7	193	8965.1	46.4510	6.8155	0.4576	0.4407	miles traveled per year

Nonlinear OLS Parameter Estimates				
Parameter	Estimate	Approx Std Err	t Value	Approx Pr > |t|
b0	-439.36	132.5	-3.32	0.0011
b1	14.02298	1.5195	9.23	<.0001
b2	16.84299	3.0825	5.46	<.0001
b3	-58.736	73.9447	-0.79	0.4280
b4	-156.657	81.5188	-1.92	0.0561
b5	-226.478	79.5331	-2.85	0.0049
b6	-300.709	114.8	-2.62	0.0095

Number of Observations		Statistics for System	
Used	200	Objective	44.8253
Missing	0	Objective*N	8965
Sum of Weights	0.0671		

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
miles	White's Test	55.75	17	<.0001	Cross of all vars

Question2: - BASS MODEL PROBLEM:

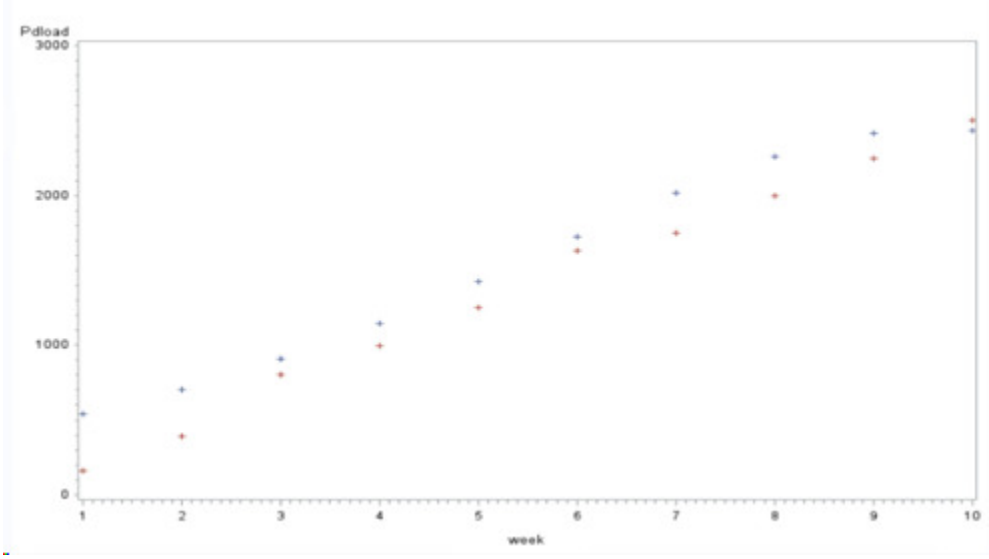
1. **Using SAS and regression, estimate the Bass model. Save the regression parameters using option OUTEST. Find p, q, and M and compute peak sales and the time when that peak will occur.**
   Using the regression model on SAS, the estimated bass model is:
   St = 539.73 + 0.3101*N(t-1) -0.000012*(N(t-1)) ^2
   The estimated value of p = 0.020581
   The estimated value of q = 0.3307
   The estimated value of M = 26225.01
   The estimated peak sales value is 2446.51
   The estimated peak time when the peak sales occur is 7.90474
   From the above values, we can infer that 2.05% of people who will buy the product on their own, 33.07% of people who will buy the product after hearing about the product from others and the total market potential for the product is 26225(maximum number of people that will ever buy this product). The peak sales for the product is 2446 units of goods sold and the time when the peak sales occur is 8th week approximately.

2. **Predict sales in each period using only the model parameters p, q, and M and the fact that sales at time period 0=0 .**

Obs	week	sales	Pdload
1	1	160	539.74
2	2	390	703.45
3	3	800	905.80
4	4	995	1147.96
5	5	1250	1425.14
6	6	1630	1723.00
7	7	1750	2014.72
8	8	2000	2260.86
9	9	2250	2415.17
10	10	2500	2437.60

As per the above scenario, initially, the sales are observed to be high and constantly grows over time and towards the 9th week, begins to show a very stunted growth and that could signify the decline in sales of the product.

3. **Plot a graph of actual versus predicted sales. (SAS code given to you in the slides)**



Question3: - CONJOINT ANALYSIS PROBLEM:

1. **Find the importance weights and part-worths for each respondent using PROC TRANSREG**.

PART-WORTHS: -

- **Util_complete, util_wave and util_smile** are the levels of attribute Brand
- **Util_u, util_fresh and util_lemon** are the levels of attribute Scent.
- **Util_y and util_n** are the levels of attribute Soft.
- **Util_small, util_med and util_lar** are levels of attribute OZ.
- **Util_high, util_mid and util_low** are levels of attribute Price.

Obs	VARIABLES	RESP1	RESP2	RESP3	RESP4	RESP5
1	util_complete	-2.06057	-0.3410	-0.15008	-0.00386	-0.39680
2	util_smile	-1.92863	0.5756	0.03742	0.26003	0.51640
3	util_wave	3.98920	-0.2346	0.11265	-0.25617	-0.11960
4	util_lemon	0.21142	0.2654	0.11265	0.22531	-2.39738
5	util_u	0.22762	0.2562	-0.55633	0.14660	1.89313
6	util_fresh	-0.43904	-0.5216	0.44367	-0.37191	0.50424
7	util_small	0.07137	-3.4799	-0.62924	-1.55478	-0.26138
8	util_med	0.60610	1.0478	0.18326	-0.18904	-0.45235
9	util_lar	-0.67747	2.4321	0.44599	1.74383	0.71373
10	util_low	0.67207	1.7006	0.22145	1.81327	1.17091
11	util_mid	-0.06867	0.3858	1.62886	1.03549	0.43017
12	util_high	-0.60340	-2.0864	-1.85031	-2.84877	-1.60108
13	util_y	-0.07292	0.0417	-0.98958	-0.14583	-0.05729
14	util_n	0.07292	-0.0417	0.98958	0.14583	0.05729
15	BrandMaxMinDiff	6.04977	0.9167	0.26273	0.51620	0.91319
16	ScentMaxMinDiff	0.66667	0.7870	1.00000	0.59722	4.29051
17	SoftMaxMinDiff	0.14583	0.0833	1.97917	0.29167	0.11458
18	OZMaxMinDiff	1.28356	5.9120	1.07523	3.29861	1.16609
19	PriceMaxMinDiff	1.27546	3.7870	3.47917	4.66204	2.77199
20	Total	9.42130	11.4861	7.79630	9.36574	9.25637
21	RelBrand	0.64214	0.0798	0.03370	0.05512	0.09866
22	RelScent	0.07076	0.0685	0.12827	0.06377	0.46352
23	RelSoft	0.01548	0.0073	0.25386	0.03114	0.01238
24	RelOZ	0.13624	0.5147	0.13792	0.35220	0.12598
25	RelPrice	0.13538	0.3297	0.44626	0.49778	0.29947

Max-Min: -

- **BrandMaxMinDiff**: - Difference between Max & Min utility values in attribute Brand
- **ScentMaxMinDiff**: - Difference between Max & Min utility values in attribute Scent
- **SoftMaxMinDiff**: - Difference between Max & Min utility values in attribute Soft
- **OZMaxMinDiff**: - Difference between Max & Min utility values in attribute OZ
- **PriceMAxMinDiff**: - Difference between Max & Min utility values in attribute Price

IMPORTANCE WEIGHTS: -

- **RelBrand**: - Relative importance of attribute Brand compared to total of all attributes
- **RelScent:** - Relative importance of attribute Scent compared to total of all attributes
- **RelSoft**: - Relative importance of attribute Soft compared to total of all attributes
- **RelOZ**: - Relative importance of attribute OZ compared to total of all attributes
- **RelPrice**: - Relative importance of attribute Price compared to total of all attributes

2. Predict the choice (using logit rule) for each respondent (s1-s5) for each of the following combinations using your estimates in question 1 above.

Using the utility values of each level across the attributes and taking their sum provides the predicted value for utility for the combination of attributes under observation.

LOGIT RULE: -

As per the theory of logit rule, Probability of choosing combination A = exp (UA) / sum (exp (Ui)
PR_A, PR_B, PR_C, PR_D & PR_E are the respective probabilities of choosing the combination by each respondent.
Util_A, Util_B, Util_C, Util_D, Util_E are the predicted utility values for each respondent for respective combination.

Obs	UTILITY	RESP11	RESP21	RESP31	RESP41	RESP51
1	PR_A	0.00064	0.43496	0.12221	0.65753	0.00737
2	PR_B	0.00137	0.12404	0.15783	0.06819	0.10413
3	PR_C	0.00127	0.07250	0.23720	0.05262	0.19910
4	PR_D	0.99362	0.12009	0.06260	0.06835	0.22109
5	PR_E	0.00309	0.24841	0.42017	0.15331	0.46831
6	Util_A	-1.92747	4.09877	-0.35957	3.63272	-0.96682
7	Util_B	-1.16242	2.84414	-0.10378	1.36651	1.68191
8	Util_C	-1.23650	2.30710	0.30363	1.10725	2.33005
9	Util_D	5.42207	2.81173	-1.02855	1.36883	2.43480
10	Util_E	-0.34992	3.53858	0.87539	2.17670	3.18538

Add up all the probabilities for all combinations and divide by M, the number of people, to get market share.

Obs	UTILITY	RESP11	RESP21	RESP31	RESP41	RESP51	MS
1	PR_A	0.00064	0.43496	0.12221	0.65753	0.00737	0.24454
2	PR_B	0.00137	0.12404	0.15783	0.06819	0.10413	0.09111
3	PR_C	0.00127	0.07250	0.23720	0.05262	0.19910	0.11254
4	PR_D	0.99362	0.12009	0.06260	0.06835	0.22109	0.29315
5	PR_E	0.00309	0.24841	0.42017	0.15331	0.46831	0.25866

This table show the market share of each combination of attributes under consideration (A to E).

- Highest market share is for the combination D
- Lowest market share is for the combination B

RESPONDENT CHOICES:

- Respondent 1 will most likely choose combination D
- Respondent 2 will most likely choose combination A
- Respondent 3 will most likely choose combination E
- Respondent 4 will most likely choose combination A
- Respondent 5 will most likely choose combination E