Case Study:

A food product company is trying to gauge the public interest in their new product through market survey. The market survey involves distributing a questionnaire to a randomly selected group of people. The public interest in the new product is being measured through "Willingness"- a scale variable with values from 1 to 11. "1" indicates least willingness to purchase the product. "11" indicates maximum willingness to purchase the product. The other variables in the questionnaire includes:

1. Age (of the respondent) – scale variable measured in years
2. Gender (of respondent) named as Female – Female =1 and Male = 0
3. Income group (of respondent) – categorical variable with 3 categories: High Income indicated by 1; Middle Income indicated by 2 and Low Income indicated by 3.

The descriptive statistics of the variables are as follows:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Willingness	100	1.0	11.0	6.350	2.9555
Age	100	20.0	58.0	35.710	9.8947
middleclass	100	0	1	.37	.485
highincome	100	0	1	.32	.469
female	100	.0	1.0	.510	.5024
Valid N (listwise)	100				

The number of respondents in the study is 100. A multiple linear regression with "Willingness" as the dependent and the remainder variables as the independents is run and the partial results are given as follows:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1		.793		1.3741	1.593

a. Predictors: (Constant), female, middleclass, Age, highincome

b. Dependent Variable: Willingness

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	685.379		171.345		
	Residual	179.371		1.888		
	Total	864.750				

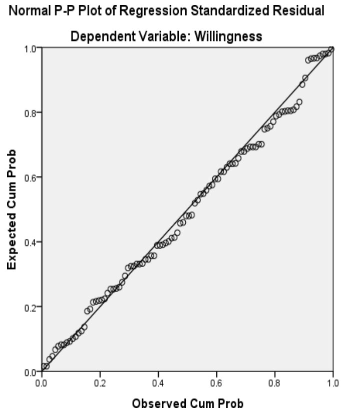
a. Dependent Variable: Willingness

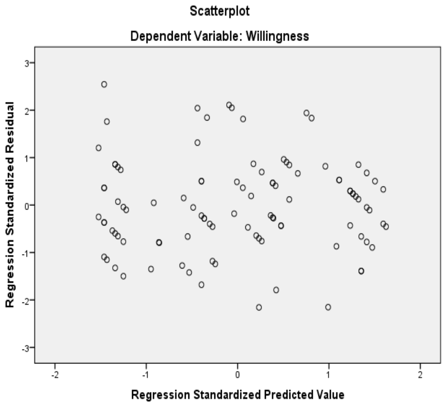
b. Predictors: (Constant), female, middleclass, Age, highincome

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	.756	.612				-.458	1.970		
	Age	.079	.021				.038	.121	.454	
	middleclass	2.010	.381	.330			1.254	2.767	.558	
	highincome	3.527	.541	.559			2.452	4.601	.296	
	female	1.736	.329	.295			1.082	2.389	.698	

a. Dependent Variable: Willingness





Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.058	100	.200*	.984	100	.268

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Question 1

Can you find a goodness of fit measure for the above regression model, other than the given R-square? What is the difference between your new measure and the given R-square? Explain clearly with all the workings.

Question 2

Can you find the independent variables that have a statistically significant linear relationship with the dependent variable – "Willingness"? Also, is the intercept term in the above model statistically significant? Explain clearly with all the workings.

Question 3

In the above regression model, the Income group of the respondents is represented by the two dummy (0-1) variables – “middleclass” (for category 2) and “highincome” (for category 1). Do you think this is the correct approach? Explain clearly with proper reasons. Also how do you interpret the coefficient of the dummy variable – “middleclass”?

Question 4

From the ANOVA table can you conclude that the overall model is significant i.e. there is linear relationship between some of the independent variables and “Willingness”? Explain clearly with all the workings.

Question 5

Do you think the above regression model suffers from the problem of multi-collinearity? Explain clearly with all the workings.

Question 6

Can you find a 90% confidence interval for the coefficients of “Age” and “female”? Show all your workings / steps.

Question 7

Do you think that the interpretation from the given P-P (probability-probability) plot is supported by the results of other statistical tests given? Explain clearly with proper steps/ reasoning.

Question 8

Among the independent variables which has the largest impact on “Willingness”? Explain clearly with all the workings.

Question 9

Please comment on the validity and reliability of the above regression model.

The previous regression model is augmented by the addition of a new independent variable. Variable named “agemiddle” computed as age*middleclass (product of age & middleclass) is created and the regression model is rerun after the addition of the new variable. The partial regression results are as follows.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.898[a]	.807	.797	1.3315	1.554

a. Predictors: (Constant), agemiddle, Age, female, highincome, middleclass

b. Dependent Variable: Willingness

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	698.108	5	139.622	78.758	.000[b]
	Residual	166.642	94	1.773		
	Total	864.750	99			

a. Dependent Variable: Willingness

b. Predictors: (Constant), agemiddle, Age, female, highincome, middleclass

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	1.765	.702		2.514	.014	.371	3.159		
	Age	.041	.025	.139	1.685	.095	-.007	.090	.303	
	middleclass	-1.792	1.466	-.294	-1.222	.225	-4.702	1.119	.035	
	highincome	4.115	.569	.653	7.236	.000	2.986	5.245	.252	
	female	1.882	.323	.320	5.819	.000	1.240	2.524	.678	
	agemiddle	.117	.044	.679			.030	.203	.032	

a. Dependent Variable: Willingness

Question 10

How do you interpret the coefficient of the new variable “agemiddle”? Is the coefficient of this new variable statistically significant?

Question 11

How does the addition of the new variable- “agemiddle” impact the first regression model? On comparing the two regression models do you see any positive effects of the addition of “agemiddle”? Do you also observe any negative effects of this addition? Please explain clearly supported by workings (if required).

Question 12

Using the first regression model can you predict the “willingness to purchase the product” of a 40 years old gentleman belonging to the middle class. Point prediction will suffice. Show your workings clearly.

The following values may be useful in your above analysis:

P (t > 1.235, df=95) = 0.11 (approx.) P (t > 3.762, df =95) = 0.00015 (approx.) P (t > 5.276, df =95) = 0.000 (approx.) P (t > 6.52, df=95) = 0.000 (approx.) P (t> 5.277, df=95) = 0.000 (approx.) P (F > 90.755, df numerator=4, df denominator =95) =0.000 (approx.) P (t > 2.659, df=94) = 0.0046 (approx.) P (t> 3.06, df =94) = 0.0014 (approx.) P (t<-6.532, df=94) = 0.000 (approx.)	**T critical (alpha = 0.025, df=95) = -1.98525 T critical (alpha = 0.05, df =95) = -1.66105 F critical (alpha=0.05, df numerator=4, df denominator=95) = 2.47 (approx.) T critical (alpha = 0.025, df=94) = -1.9855**