Regression Analysis Student Project （Fall 2013）

Introduction

The objective of this project is to analyze the expenditure for food with respect to two variables using a regression model.
The explanatory variables used for this analysis are: personal income after tax and sex.

Data Source

The data for this analysis was taken from Bureau of Labor Statistics (http://stats.bls.gov/cex/csxcross.htm)
All explanatory and response variable data were taken from 2000 to 2001. And the units of variables are U.S. dollars.
We divided the data into several groups by ages, and they are shown in appendix. In each group, we calculate the average personal income after tax and the average expenditure for food. The number of persons in each group can reach 1000+.

Chosen response and explanatory variables

Response variable (Y): Expenditure for food
Explanatory variable(X_1): Personal income after tax
Explanatory variable (D_1): Sex
Explanatory variable (D_1*X_1): Personal income after tax*Sex

Results and Analysis

Model1

The fitted model is: $Y = \alpha + \beta_1 * X_1$

Regression in Excel:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.944485
R Square	0.892052
Adjusted R Squ	0.881257
Standard Error	208.2664
Observations	12

ANOVA

	df	SS	MS	F	ignificance F
Regression	1	3584369	3584369	82.63697	3.78175E-06
Residual	10	433748.9	43374.89		
Total	11	4018118			

	Coefficient	tandard Er	t Stat	P-value	*Lower 95%*	*Upper 95%*
Intercept	1301.443	188.4733	6.905185	4.17E-05	881.4983731	1721.388
X Variable 1	0.062454	0.00687	9.090488	3.78E-06	0.047146044	0.077762

Fitted Model:

$Y_i = 1301.44 + 0.06 * X_i$

Observations:

- An adjusted R^2 value is 0.88, which means that 88% of the expenditure for food can be explained by the personal income after tax.
- The coefficient of X_1 0.06 means that if the personal income after tax increases by $1, the expenditure for food will increase by $0.06. The relationship between these two variables is positive. We also know that the P-value is 3.78*E-06, which means that this variable is significant.
- The intercept of 1301.44 means that if the personal income after tax is zero, the expenditure for food will be 1301.44. So we can know that the expenditure for food is a rigid demand, which is reasonable.

We know that the expenditure of food for male and female can be different, and usually man eat more. So we set up another model to test whether male and female has the same expenditure for food under the same personal income after tax.

Model2

The fitted model is: $Y = \alpha + \beta_1 * X_1 + \gamma_1 * D_1$

Regression in Excel:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.963541
R Square	0.928411
Adjusted R Square	0.912502
Standard Error	178.7778
Observations	12

ANOVA

	df	SS	MS	F	Significance F
Regression	2	3730465	1865232	58.35873	7.02766E-06
Residual	9	287653.5	31961.5		
Total	11	4018118			

	Coefficient	tandard Er	t Stat	P-value	*Lower 95%*	*Upper 95%*
Intercept	1506.245	188.0192	8.011124	2.19E-05	1080.916047	1931.574
X Variable 1	0.058979	0.006117	9.641242	4.85E-06	0.045140513	0.072817
X Variable 2	-228.905	107.0656	-2.13799	0.061221	-471.1040418	13.29452

Fitted Model:

$$Y_i = 1506.25 + 0.06 * X_i - 228.91 * D_i$$

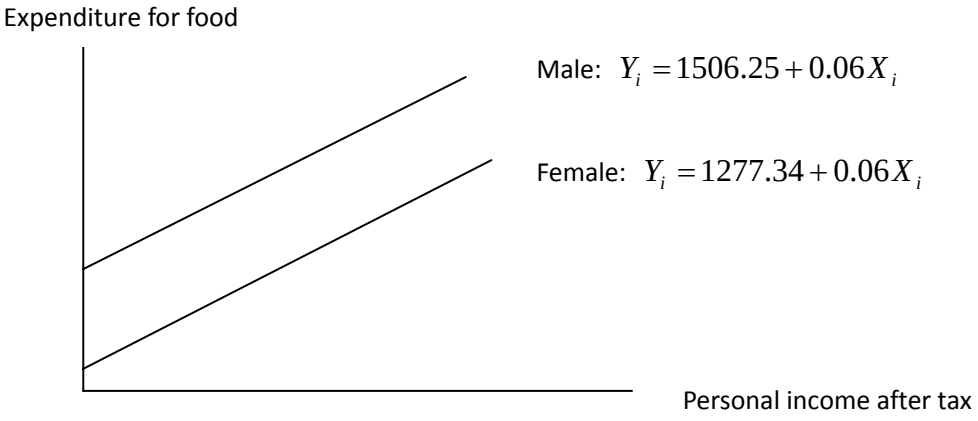
0	male
1	female

Observations:

- The variable X_1 is significant because p-value is almost zero. And the coefficient is 0.06. It means that if the personal income after tax increases by \$1, the expenditure for food will increase by \$0.06. And the result is same as the model 1.
- The variable X_2 is significant, which P-value is 0.06. And the coefficient is -228.91.If we set D_1 equals to 0 (male), the expenditure for food will be \$1506.25. And if we set D_2 equals to 1(female), the expenditure for food will be \$1277.34. The difference is 228.91, which is the difference of expenditure of food between male and female. What's more, the regression can be expressed as below.

  Male: $Y_i = 1506.25 + 0.06 * X_i$

  Female: $Y_i = 1277.34 + 0.06 * X_i$



- In this model, the adjusted R^2 value is 0.91, which is higher the previous model.

In the next model, we will consider another variable, that is interaction dummy variable. The purpose of this model is to test whether X_i and D_i interact with each other.

Model3

The fitted model is: $Y = \alpha + \beta_1 * X_1 + \gamma_1 * D_1 + \delta_1 * X_1 * D_1$

Regression in Excel:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.964597
R Square	0.930448
Adjusted R Squa	0.904366
Standard Error	186.9049
Observations	12

ANOVA

	df	SS	MS	F	gnificance F
Regression	3	3738651	1246217	35.67406	5.6E-05
Residual	8	279467.4	34933.43		
Total	11	4018118			

	Coefficient	tandard Er	t Stat	P-value	*Lower 95%*	*Upper 95%*
Intercept	1432.657	248.4899	5.765455	0.000421	859.6385	2005.676
X Variable 1	0.061577	0.008349	7.375432	7.8E-05	0.042324	0.08083
X Variable 2	-67.9731	350.7863	-0.19377	0.851183	-876.888	740.9417
X Variable 3	-0.00629	0.012988	-0.48408	0.641303	-0.03624	0.023663

Fitted Model:

$$Y_i = 1432.66 + 0.06 * X_i - 67.97 * D_i - 0.00629 * D_i * X_i$$

Observations:

- In this model, the adjusted R^2 value is 0.90, which is slightly lower than Model 2.
- We can see that the interaction dummy is not significant, and the P value is 0.64.
- We can also find that the D_1 variable is not significant, and the P value is 0.85.

Summary

A comparison of the models is shown here

	Adjusted R^2	Standard Error	Highest P value
Model 1	0.88	208.27	4.17E-05
Model 2	0.91	178.78	0.06
Model 3	0.90	186.90	0.85

Among the three models, I think the best model is Model 2. The reasons are listed as follows.

- Model two has the highest adjusted R^2, and with the lowest SE.
- All the coefficients are significant.
- We added the variable D_1*X_1 in model 3, but the adjusted R^2 doesn't increase. We think that the variable D_1*X_1 doesn't provide additional explanatory information to the model. In addition, if we add the variable D_1*X_1 in the model, we may make a mistake.

Conclusion

In my conclusion that the best model for determining expenditure for food is model 2, which is summarized as below.

$$Y_i = 1506.25 + 0.06 * X_i - 228.91 * D_i$$

This model shows a positive correlation between expenditure for food and personal income after tax. Also male and female has significant difference. Male has more expenditure for food than female. And the difference can reach $229. It is reasonable, because male always eat more than female. Finally, the adjusted R^2 is 0.91, meaning that 91% of the expenditure for food can be explained by personal income after tax and sex.

Appendix

Age	*expenditure for food_female*	*personal income after tax_female*	*expenditure for food_male*	*personal income after tax_male*
<25	1983	11557	2230	11589
25-34	2987	29387	3757	33328
35-44	2993	31463	3821	36151
45-54	3156	29554	3291	35448
55-64	2706	25137	3429	32998
>65	2217	14952	2533	20437