BUS 735: Business Decision Making and Research
Instructor: Dr. James Murray
Take Home Exam 1 - Fall 2014
Answer Key

Directions: Type up answers to all of the following questions. Include in your document only the relevant SPSS output that you need to answer the question. Please copy and paste this SPSS output; do not include in your submission any other files except a single Microsoft Word document or PDF document that includes all your answers with the relevant SPSS output accompanying each answer. Every time you conduct a hypothesis test, indicate what statistical test you are using, what are the null and alternative hypotheses, what is your p-value, and a plain English description of what is your conclusion.

1. The dataset `wage1D.sav` contains the following variables including wage and background information for 526 individuals:

   - `wage`: average hourly earnings
   - `educ`: years of education
   - `exper`: years of experience
   - `tenure`: years with current employer
   - `nonwhite`: Dummy variable = 1 if employee is non-white.
   - `female`: Dummy variable = 1 if employee is female.

   (a) Estimate a regression that explains average hourly earnings using all the variables in the dataset. What is your estimated regression equation?

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.603[a]	.364	.358	2.960

a. Predictors: (Constant), female, nonwhite, exper, educ, tenure

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2603.752	5	520.750	59.427	.000[b]
	Residual	4556.662	520	8.763		
	Total	7160.414	525			

a. Dependent Variable: wage

b. Predictors: (Constant), female, nonwhite, exper, educ, tenure

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.540	.732		-2.103	.036
	educ	.570	.050	.428	11.507	.000
	exper	.025	.012	.093	2.188	.029
	tenure	.141	.021	.276	6.660	.000
	nonwhite	-.116	.427	-.010	-.271	.786
	female	-1.812	.265	-.245	-6.835	.000

a. Dependent Variable: wage

$$\mathbf{Wage_i = -1.54 + 0.570(Educ_i) + 0.025(Exper_i) + 0.141(Tenure_i) - 0.116(Nonwhite_i) - 1.812(Female_i) + e_i}$$

   (b) What percentage of the variability in average hourly earnings is explained by education, experience with employer, total experience, race, and gender?

   36.4% (the value of R-square)

(c) Controlling for the other variables in the model, is there evidence that female employees receive lower wages on average than male employees? If so, what is the estimated difference in hourly earnings?

$H_0: \beta_{female} = 0$
$H_a: \beta_{female} < 0$
P-value = 0.000
We found sufficient statistical evidence that women earn smaller hourly wages than men. On average women make \$1.81 less than men ($b_{female} = -1.812$).

(d) Controlling for the other variables in the model, is there evidence that more total experience leads to an increase in average hourly earnings? If so, what is the estimated difference in average hourly earnings between someone who has 8 years of experience, and someone who has 10 years of total experience?

$H_0: \beta_{exper} = 0$
$H_a: \beta_{exper} > 0$
P-value = 0.015 (= 0.029 / 2)
We found sufficient statistical evidence that experience leads to higher hourly wages. On average, each year experience adds \$0.025 to hourly wages ($b_{exper} = 0.025$). Therefore, an additional 2 years experience leads to an average hourly wage that is 5 cents more per hour.

(e) What is the predicted wage for a white male with 12 years of education, 10 years of experience, and 8 year experience with his current employer?

$\hat{Wage}_i = -1.54 + 0.570(12) + 0.025(10) + 0.141(8) - 0.116(0) - 1.812(0) = \$6.68$

2. The following questions use the dataset `cex.sav`. This is recent (2010:Q2) consumer income data from the Current Population Survey. The variables included in this SPSS file include:

   - Age (in years)
   - Relationship to head-of-household: 1=head of household, 2=spouse, 3=child or adopted child of head, 4=grandchild of head, 5=in-law of head, 6=brother/sister of head, 7=mother/father of head, 8=other relatives, 9=unrelated individual, 0=na.
   - Education: 00=Never attended school, 1-11 1st grade through 11th grade, 38=Twelfth grade no degree, 39=High school graduate, 40=Some college no degree, 41=Associate's degree (occupational/vocational), 42=Associate's degree (academic), 43=Bachelor's degree, 44=Master's degree, 45 =Professional degree, 46=Doctorate degree
   - Race: 1=White, 2=Black/African American, 3=American Indian or Aleut Eskimo, 4=Asian or Pacific Islander, 5=other
   - Gender: 1=male, 2=female
   - Marital Status: 1=married, 2=widowed, 3=divorced, 4=separated, 5=never married
   - Employee Status: 1=member worked full time for a full year, 2=member worked part time for a full year, 3=member worked full time for part of year, 4=member worked part time for part of year, blank if member did not work.
   - Employee Type: 1=private company, 2=government employee, 3=self-employed, 4=working without pay.
   - Hours worked per week
   - Weeks worked per year
   - Occupation: 01=managerial and professional specialty occupation, 02=technical, sales, and administrative support occupations, 03=service occupations, 04=farming, forestry, and fishing occupations, 05=precision production, craft, and repair occupations, 06=operators, fabricators, and laborers, 07=armed forces, 08=self employed, 09=not working, 10=retired, 11=other, including not reported.
   - Total Income: in dollars.

   Transform Education into a new variable with the following categories:

   (a) =1 if High School graduate or less

   (b) =2 if Any college below a Bachelor's degree (40<EDU<43)

   (c) =3 if Bachelor degree

   (d) =4 if Masters, professional, or doctorate degree

   For all of the following questions, use the results from a single analysis that controls for how all of the following variables affect income: employee type, gender, race, age, hours worked per week, and education level (as defined by your new variable).

(a) Comment on all the factors for which you find statistical evidence that influence average income. Report explicitly your hypothesis tests for these variables.

ANCOVA with TotalInc as the dependent variable; age and hours per week as explanatory covariates; and employee type, gender, race, and education as fixed factors.

Tests of Between-Subjects Effects

Dependent Variable: TotalInc

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	8.482E+11[a]	95	8928635780	7.881	.000
Intercept	1.082E+10	1	1.082E+10	9.549	.002
Age	3.382E+10	1	3.382E+10	29.852	.000
hrsweek	2.098E+11	1	2.098E+11	185.166	.000
emptype	870331528.4	3	290110509.5	.256	.857
Gender	21389006.20	1	21389006.20	.019	.891
Race	1790727206	4	447681801.4	.395	.812
Education	2.437E+10	3	8124719855	7.171	.000
emptype * Gender	1056351141	2	528175570.3	.466	.627
emptype * Race	6173021430	9	685891270.0	.605	.793
emptype * Education	1.380E+10	7	1970754563	1.740	.095
Gender * Race	7068974324	4	1767243581	1.560	.182
Gender * Education	267557005.9	3	89185668.64	.079	.972
Race * Education	9209640636	12	767470053.0	.677	.775
emptype * Gender * Race	2930689392	6	488448232.0	.431	.859
emptype * Gender * Education	3979161271	6	663193545.1	.585	.742
emptype * Race * Education	2.516E+10	14	1797367065	1.586	.075
Gender * Race * Education	1.002E+10	9	1113306427	.983	.452
emptype * Gender * Race * Education	2530818368	6	421803061.3	.372	.897
Error	3.109E+12	2744	1132924296		
Total	5.725E+12	2840			
Corrected Total	3.957E+12	2839			

a. R Squared = .214 (Adjusted R Squared = .187)

Age, hours per week, and education influence average income.

Null: Average income is equal over education categories
Alt: Average income is not equal over education categories
p-value = 0.000. Reject the null hypothesis.
Found statistical evidence that average income does depend on education level.

Null: Age does not affect average income.
Alt: Age does affect average income.
p-value = 0.000. Reject the null hypothesis.
Found statistical evidence that average income does depend on age.

Null: Hours per week does not affect average income.
Alt: Hours per week does affect average income.
p-value = 0.000. Reject the null hypothesis.
Found statistical evidence that average income does depend on hours per week.

(b) For the categorical variables that you identify above as influencing average income, comment on the relationship between the factor and average income. What categories of individuals do you find statistical evidence for having higher income?

Education

Estimates

Dependent Variable: TotalInc

Education	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1.00	12208.137[a,b]	3808.363	4740.590	19675.685
2.00	18898.143[a,b]	4152.441	10755.917	27040.369
3.00	24557.798[a,b]	4757.458	15229.237	33886.359
4.00	47472.158[a,b]	4796.116	38067.796	56876.520

a. Covariates appearing in the model are evaluated at the following values: Age = 40.37, hrsweek = 38.95.

b. Based on modified population marginal mean.

Pairwise Comparisons

Dependent Variable: TotalInc

(I) Education	(J) Education	Mean Difference (I-J)	Std. Error	Sig.[d]	95% Confidence Interval for Difference[d]	
					Lower Bound	Upper Bound
1.00	2.00	-6690.006[a,b]	5635.574	.235	-17740.402	4360.389
	3.00	-12349.7[a,b,*]	6093.119	.043	-24297.225	-402.097
	4.00	-35264.0[a,b,*]	6125.451	.000	-47274.983	-23253.059
2.00	1.00	6690.006[a,b]	5635.574	.235	-4360.389	17740.402
	3.00	-5659.655[a,b]	6315.854	.370	-18043.963	6724.654
	4.00	-28574.0[a,b,*]	6342.567	.000	-41010.703	-16137.326
3.00	1.00	12349.66[a,b,*]	6093.119	.043	402.097	24297.225
	2.00	5659.655[a,b]	6315.854	.370	-6724.654	18043.963
	4.00	-22914.4[a,b,*]	6753.360	.001	-36156.543	-9672.177
4.00	1.00	35264.02[a,b,*]	6125.451	.000	23253.059	47274.983
	2.00	28574.01[a,b,*]	6342.567	.000	16137.326	41010.703
	3.00	22914.36[a,b,*]	6753.360	.001	9672.177	36156.543

Based on estimated marginal means

Of all the categorical variables, there is only statistical evidence that education influences income. The post-hoc tests on Education reveal that the highest level of education (Master's degree or above) has the highest level of income, and is statistically significantly greater than all the others. Also, statistical evidence is found for those with a bachelor have a higher average income than those with high school or less.

(c) For the interval/ratio variables that you identify above as influencing average income, comment on the relationship between the factor and average income.

Correlations

Correlations

		TotalInc	Age	hrsweek
TotalInc	Pearson Correlation	1	.208**	.507**
	Sig. (2-tailed)		.000	.000
	N	6953	6953	6953
Age	Pearson Correlation	.208**	1	.225**
	Sig. (2-tailed)	.000		.000
	N	6953	6953	6953
hrsweek	Pearson Correlation	.507**	.225**	1
	Sig. (2-tailed)	.000	.000	
	N	6953	6953	6953

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

Review of the Pearson correlation coefficients show that age and hours per week are *positively* related to income.

(d) For what *pairs* of categorical factors do you find statistical evidence for an interaction effect? Conduct these hypothesis tests and comment on the nature of the interaction effect.

None of the pairwise interaction effects are statistically significant at the 5% level.

3. The dataset `LoanApplications.sav` contains data collected as part of the Housing Mortgage Disclosure Act on all mortgage loans in the state of Wisconsin for 2013 for first-mortgages for owner occupied housing (more than 58,000 observations). The variables include the following:

   - `LoanAmt`: Amount of the loan request (in thousands of dollars)
   - `PreapproveRequest`: Whether or not pre-approval was requested (=1 if requested, =2 if not, and =3 if not applicable)
   - `ActionType`: Whether the loan was approved by the lending institution and accepted by the borrower (=1 if approved and accepted, =2 if approved but not accepted, =3 if denied).
   - `County`: County in Wisconsin, number instead of descriptive text.
   - `Ethinicity`: Ethnicity of the applicant (first applicant in the case of co-applicants). =1 if non-Hispanic / non-Latino/a, =2 if Hispanic or Latina
   - `Race`: Race of the applicant (first applicant in the case of co-applicants). =1 if White, =2 Black / African American, ...
   - `Sex`: Sex of the applicant (first applicant in the case of co-applicants). =1 if Male, =2 if Female
   - `Income`: Combined household income of the applicants and co-applicants (in thousands of dollars)

   (a) Not controlling for any other variables, is there a relationship between the loan amount and the income of the applicant(s)? If so, describe the relationship.

   Pearson Correlation for Income and Loan Amount
   H_0 : Correlation is equal to zero.
   H_A : Correlation is not equal to zero.

Correlations

Correlations

		TotalInc	Age	hrsweek
TotalInc	Pearson Correlation	1	.208**	.507**
	Sig. (2-tailed)		.000	.000
	N	6953	6953	6953
Age	Pearson Correlation	.208**	1	.225**
	Sig. (2-tailed)	.000		.000
	N	6953	6953	6953
hrsweek	Pearson Correlation	.507**	.225**	1
	Sig. (2-tailed)	.000	.000	
	N	6953	6953	6953

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

p-value = 0.000 Reject the null hypothesis.
There is statistical evidence that income and loan amount are correlated. The positive Pearson correlation coefficient indicates it is a positive relationship.

(b) Not controlling for any other variables, is there a relationship between whether pre-approval was sought on the loan and the subsequent action taken on the loan.

Chi-square test of independence for two categorical variables: PreapproveRequest and ActionType

H_0 : PreapproveRequest and ActionType are independent (not related)

H_0 : PreapproveRequest and ActionType are not independent (they are related)

PreapproveRequest * ActionType Crosstabulation

Count

		ActionType			
		1	2	3	Total
PreapproveRequest	1	9362	287	864	10513
	2	19709	909	3017	23635
	3	19512	937	3490	23939
Total		48583	2133	7371	58087

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	314.244[a]	4	.000
Likelihood Ratio	336.699	4	.000
Linear-by-Linear Association	277.874	1	.000
N of Valid Cases	58087		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 386.05.

p-value = 0.000 Reject the null hypothesis.

There is statistical evidence that requesting preapproval is related to the subsequent action taken on the loan.

(c) Not controlling for any other variables, is there a relationship between race and the action taken on the loan?

Chi-square test of independence for two categorical variables: Race and ActionType

H_0 : Race and ActionType are independent (not related)

H_0 : Race and ActionType are not independent (they are related)

Race * ActionType Crosstabulation

Count

		ActionType			
		1	2	3	Total
Race	1	221	12	96	329
	2	1080	73	219	1372
	3	800	31	259	1090
	4	65	2	26	93
	5	45731	1933	6579	54243
	6	686	82	192	960
Total		48583	2133	7371	58087

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	385.189[a]	10	.000
Likelihood Ratio	317.621	10	.000
Linear-by-Linear Association	113.076	1	.000
N of Valid Cases	58087		

a. 1 cells (5.6%) have expected count less than 5. The minimum expected count is 3.42.

p-value = 0.000 Reject the null hypothesis.

There is statistical evidence that race is related to the action taken on the loan.

(d) Not controlling for any other variables, is there a difference in the proportion of men approved for a loan versus the proportion of women approved?

Need to create a new variable for Approval. Approval = 1 if ActionType = 1 or 2 (approved), Approval = 0 if ActionType = 3 (denied)

Independent Samples T-test, compute mean of Approval (proportion approved) for men versus women.
H_0 : Proportion of men approved for loan is equal proportion of women approved for a loan.
H_0 : Proportion of men approved for loan is not equal proportion of women approved for a loan.

T-Test

Group Statistics

	Sex	N	Mean	Std. Deviation	Std. Error Mean
Approved	1	41006	.8773	.32814	.00162
	2	17081	.8631	.34373	.00263

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Approved	Equal variances assumed	85.808	.000	4.665	58085	.000	.01414	.00303
	Equal variances not assumed			4.577	30668.518	.000	.01414	.00309

p-value = 0.000. Reject the null hypothesis.
There is statistical evidence that the proportion of people approved for a loan is different for men and women. The approval rate for men is 1.4% more than women.

4. Use the same dataset as the previous question, and answer the following questions by estimating a logistic regression that predicts whether or not an applicant is approved for a loan, based sex, ethnicity, whether or not the person is white, and the ratio of amount of their loan request to their income.

   (a) Write down the estimated logistic regression equation.
   Recode variables so that approval, sex, ethnicity, and race are all dummy variables. Also, compute a new variable for the ratio of loan amount to income.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Sex	.100	.027	13.477	1	.000	1.105
	Ethnicity	.773	.048	263.018	1	.000	2.166
	White	-1.002	.125	63.918	1	.000	.367
	Ratio	-.193	.011	310.068	1	.000	.825
	Constant	1.603	.056	815.617	1	.000	4.968

a. Variable(s) entered on step 1: Sex, Ethnicity, White, Ratio.

$$\mathbf{l_i = 1.6 + 0.100(Sex_i) + 0.773(Ethinicity_i) - 1.002(Race_i) - 0.193(Ratio_i) + e_i}$$

   (b) Accounting for the controls in your logistic regression, is there evidence that sex influences whether or not someone will be approved for a loan?

   T-test on regression coefficient on sex
   H_0 : Sex does not influence loan approval (coefficient=0)
   H_A : Sex does influence loan approval (coefficient≠0)
   p-value = 0.000 Reject the null hypothesis.
   There is statistical evidence that sex does influence whether or not someone is approved of a loan.

   (c) Accounting for the controls in your logistic regression, is there evidence that race influences whether or not someone will be approved for a loan?

   T-test on regression coefficient on race
   H_0 : Race does not influence loan approval (coefficient=0)
   H_A : Race does influence loan approval (coefficient≠0)
   p-value = 0.000 Reject the null hypothesis.
   There is statistical evidence that race does influence whether or not someone is approved of a loan.

   (d) Predict the probability that the following person will be approved for a loan. The person is male, white, non-Hispanic / non-Latino, and the ratio of their loan request to their income is 2.5.
   $\mathbf{\hat{l}_i = 1.6 + 0.100(1) + 0.773(1) - 1.002(1) - 0.193(2.5) = 0.9885}$
   $\mathbf{P(\hat{y_i} = 1) = \frac{1}{1+e^{-0.9885}} = 0.729}$
   This person has a 72.9% chance of being approved for a loan.

   (e) How much more or less likely would the person above be approved for a loan if he was not white?
   $\mathbf{\hat{l}_i = 1.6 + 0.100(1) + 0.773(1) - 1.002(0) - 0.193(2.5) = 1.9905}$
   $\mathbf{P(\hat{y_i} = 1) = \frac{1}{1+e^{-1.9905}} = 0.880}$
   This person has an 88% chance of being approved for a loan. This person is 15% more likely.