BUS 735: Business Decision Making and Research
Instructor: Dr. James Murray
Fall 2010
Take Home Exam 1

1. Develop a model that predicts students' weight based on their gender, age, height, the number of hours they watch TV, their participation in sports, and whether they used to be a smoker.

Linear Regression
Dependent variable: weight (kg)
Explanatory variables: gender (dummy), age (years), height (cm), hours watching TV, past smoker (dummy).

a) What variables, if any, lead to statistically signicant increases in weight? Report the relevant statistics and p-values.

The variables with positive and statistically significant coefficients are age, height, and spending more time watching TV lead to an increase in weight. See table and hypothesis tests below.

(b) What variables, if any, lead to statistically signicant decreases in weight? Report the relevant statistics and p-values.

The only variable with a negative and statistically significant coefficient is gender; females have lower weight than males. See table and hypothesis tests below.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-61.078	17.719		-3.447	.001
	age	.324	.069	.251	4.720	.000
	sex	-5.146	2.104	-.183	-2.446	.015
	height in cm	.682	.097	.530	7.012	.000
	hours watching TV/week	.159	.075	.110	2.111	.036
	Participation in sport	.711	.528	.072	1.347	.180
	Past smoker?	.013	1.307	.001	.010	.992

a. Dependent Variable: weight in kg

Age:
Null: Age does not affect weight (coefficient is equal to zero)
Alternative: Age does affect weight (coefficient is not equal to zero)
Coefficient = 0.324
P-value = 0.000
Reject the null hypothesis.
There is sufficient statistical evidence that age affects weight.

Gender:
Null: Gender does not affect weight (coefficient is equal to zero)
Alternative: Gender does affect weight (coefficient is not equal to zero)
Coefficient = -5.146
P-value = 0.015
Reject the null hypothesis.
There is sufficient statistical evidence that gender affects weight.

Height:
Null: Height does not affect weight (coefficient is equal to zero)
Alternative: Height does affect weight (coefficient is not equal to zero)
Coefficient = 0.682
P-value = 0.000
Reject the null hypothesis.
There is sufficient statistical evidence that height affects weight.

Television:
Null: Time watching television does not affect weight (coefficient is equal to zero)
Alternative: Time watching television does affect weight (coefficient is not equal to zero)
Coefficient = 0.159
P-value = 0.036
Reject the null hypothesis.
There is sufficient statistical evidence that watching television affects weight.

Participation in sports:
Null: Participation in sports does not affect weight (coefficient is equal to zero)
Alternative: Participation in sports does affect weight (coefficient is not equal to zero)
Coefficient = 0.711
P-value = 0.180
Fail to reject the null hypothesis.
Failed to find statistical evidence that height affects weight.

Past smoker:
Null: Being a past smoker does not affect weight (coefficient is equal to zero)
Alternative: Being a past smoker does affect weight (coefficient is not equal to zero)
Coefficient = 0.013
P-value = 0.992
Failed to reject the null hypothesis.
Failed to find statistical evidence that height affects weight.

(c) Cite a statistic and report its value that gives information on how well the independent variables explain weight. Do you think your model explains the data well?

R-squared is 0.739. Therefore, 73.9% of the variability in weight is explained by variables in the model. This is a well fitting model.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.739[a]	.546	.531	8.092

a. Predictors: (Constant), Past smoker?, sex, age, hours watching TV/week, Participation in sport, height in cm

(d) Suppose someone is 28 years old, is male, is 180 centimeters tall, watches TV 10 hours per week, does not participate in sports and has smoked in the past. What is his predicted weight?

Predicted weight = -61.078 + 0.324 (Age) – 5.146 (Gender) + 0.682 (Height) + 0.159 (Hours TV) + 0.711 (Sports Participation) + 0.013 (Smoker)

Predicted weight = -61.078 + 0.324 (28) – 5.146 (0) + 0.682 (180) + 0.159 (10) + 0.711 (0) + 0.013 (1) = 72.36 kg

(e) What is the marginal predicted impact on weight when the person in the problem above becomes one year older?

Coefficient = 0.324

On average, people gain 0.324 pounds every year.

(f) Accounting for all the other variables in the model, what is the predicted dierence in weight between men and women?

Coefficient = -5.146. On average, men weight 5.146 more pounds than women.

2. Create a new variable called 'BMI' (Body Mass Index) according to the following formula: BMI = (Weight in Kg)/(Height in Meters). People with a BMI over 35 are considered overweight. Create another variable called 'OVERWEIGHT' that is equal to 1 if BMI is over 35 and 0 otherwise. Develop a model that predicts whether or not someone is overweight based on the same explanatory variables as in #1.

Logistic Regression
Dependent variable: overweight (dummy)
Explanatory variables: gender (dummy), age (years), height (cm), hours watching TV, past smoker (dummy).

(a) How many people in your sample are overweight?

59.5% of the people in the same are overweight.

overweight

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	90	40.5	40.5	40.5
	1.00	132	59.5	59.5	100.0
	Total	222	100.0	100.0	

(b) What variables, if any, lead to statistically signicant increases in the probability of being overweight? Report the relevant statistics and p-values.

Age is only variable that has a positive coefficient and is statistically significantly different from zero. See hypothesis tests below

(c) What variables, if any, lead to statistically signicant decreases in the probability of being overweight? Report the relevant statistics and p-values.

Gender is the only variable that has a negative coefficient and is statistically significantly different from zero. Women are less likely to be overweight. See hypothesis tests below.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	sex	-1.444	.641	5.074	1	.024	.236
	age	.049	.019	6.750	1	.009	1.051
	height	.051	.027	3.682	1	.055	1.053
	hourstv	.009	.021	.181	1	.670	1.009
	parsport	.075	.141	.282	1	.595	1.078
	pastsmok	.216	.353	.374	1	.541	1.241
	Constant	-8.842	4.857	3.314	1	.069	.000

a. Variable(s) entered on step 1: sex, age, height, hourstv, parsport, pastsmok.

Age:
Null: Age does not affect being overweight (coefficient is equal to zero)
Alternative: Age does affect being overweight (coefficient is not equal to zero)
Coefficient = 0.049
P-value = 0.009
Reject the null hypothesis.
There is sufficient statistical evidence that age affects being overweight.

Gender:
Null: Gender does not affect being overweight (coefficient is equal to zero)
Alternative: Gender does affect being overweight (coefficient is not equal to zero)
Coefficient = -1.444
P-value = 0.024
Reject the null hypothesis.
There is sufficient statistical evidence that gender affects being overweight.

Height:
Null: Height does not affect being overweight (coefficient is equal to zero)
Alternative: Height does affect being overweight (coefficient is not equal to zero)
Coefficient = 0.051
P-value = 0.055
Fail to reject the null hypothesis (borderline though).
Failed to find sufficient statistical evidence that height affects being overweight.

Television:
Null: Time watching television does not affect being overweight (coefficient is equal to zero)
Alternative: Time watching television does affect being overweight (coefficient is not equal to zero)
Coefficient = 0.009
P-value = 0.670
Failed to reject the null hypothesis.
Failed to find sufficient statistical evidence that watching television affects being overweight.

Participation in sports:
Null: Participation in sports does not affect being overweight (coefficient is equal to zero)
Alternative: Participation in sports does affect being overweight (coefficient is not equal to zero)
Coefficient = 0.075
P-value = 0.595
Fail to reject the null hypothesis.
Failed to find statistical evidence that height affects being overweight.

Past smoker:
Null: Being a past smoker does not affect being overweight (coefficient is equal to zero)
Alternative: Being a past smoker does affect being overweight (coefficient is not equal to zero)
Coefficient = 0.216
P-value = 0.541
Failed to reject the null hypothesis.
Failed to find statistical evidence that height affects being overweight.

(d) How do your answers to questions (b) and (c) compare to your answers in #1(b) and #1(c)?

Taller people (height variable higher) and people who watch more TV (hours TV variable is higher) do lead to increases in weight (#1), but are not more likely to make someone *overweight* (#2).

(e) Cite statistics that report how well your independent variables explain whether or not someone is overweight. Do you think your model explains the data well?

The model seems to describe the data well. The model correctly predicts people will be overweight 68.2% of the time, and correctly predicts people will have a normal weight 69% of the time. Overall, the model correctly predicts whether or not a person will be overweight 68.6% of the time.

Classification Table[a]

Observed			Predicted		
			overweight		
			.00	1.00	Percentage Correct
Step 1	overweight	.00	58	26	69.0
		1.00	34	73	68.2
	Overall Percentage				68.6

a. The cut value is .500

(f) Suppose someone is 28 years old, is male, is 180 centimeters tall, watches TV 10 hours per week, does not participate in sports and has smoked in the past. What is the probability he is overweight?

Lhat = -8.842 – 1.444(Gender) + 0.049(Age) + 0.051(Height) + 0.009(HoursTV) + 0.075(Sports) + 0.216(Smoker)

Lhat = -8.842 – 1.444(0) + 0.049(28) + 0.051(180) + 0.009(10) + 0.075(0) + 0.216(1) = 2.016

Prob(y=1) = 1 / (1 + exp(-1*2.016)) = 0.882

88.2% chance of being overweight.

(g) What is the marginal predicted impact on the probability of being overweight when the person in the problem above becomes one year older?

Lhat = -8.842 – 1.444(Gender) + 0.049(Age) + 0.051(Height) + 0.009(HoursTV) + 0.075(Sports) + 0.216(Smoker)

Lhat = -8.842 – 1.444(0) + 0.049(29) + 0.051(180) + 0.009(10) + 0.075(0) + 0.216(1) = 2.065

Prob(y=1) = 1 / (1 + exp(-1*2.065)) = 0.887

88.7% chance of being overweight. This is 0.5% higher.

(h) Suppose there is another person identical to the person in part (f) except she is female. Is she more or less likely to be overweight than the gentleman in problem (f)? How much more or less?

Lhat = -8.842 – 1.444(Gender) + 0.049(Age) + 0.051(Height) + 0.009(HoursTV) + 0.075(Sports) + 0.216(Smoker)

Lhat = -8.842 – 1.444(0) + 0.049(29) + 0.051(180) + 0.009(10) + 0.075(0) + 0.216(1) = 0.572

Prob(y=1) = 1 / (1 + exp(-1*0.572)) = 0.639

63.9% chance of being overweight. This is 23.4% lower.

3. Develop a model that predicts average pulse while at rest, after moderate exercise, and after vigorous exercise while taking into account whether someone was a past smoker or not?

Repeated Measures Analysis of Variance with Pulse Rate (At rest / After Moderate Exercise / After Vigorous Exercise) at the within-factor variable and Past Smoker as the between-factor variable.

(a) Using this model, is there statistical evidence that pulse rate is different while at rest, after moderate exercise, and after vigorous exercise?

The test for Sphericity indicates that this assumption is violates (p-value 0.000). Therefore we will use either Greenhouse-Geisser test or Huynh-Feldt test.

Null: There is no difference in pulse at rest, after moderate exercise, and after vigorous exercise.
Alternative: There is a difference in pulse at rest, after moderate exercise, and after vigorous exercise.
P-value = 0.000. (From pulse_exercise rows in "Tests of Within-Subjects Effects" table below)
Reject Null Hypothesis.
Found sufficient statistical evidence that pulse is different at rest, after moderate exercise, and after vigorous exercise.

Mauchly's Test of Sphericity[b]

Measure:MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
pulse_exercise	.476	26.718	2	.000	.656	.689	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept + pastsmok
Within Subjects Design: pulse_exercise

Tests of Within-Subjects Effects

Measure:MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
pulse_exercise	Sphericity Assumed	63624.923	2	31812.461	595.195	.000
	Greenhouse-Geisser	63624.923	1.312	48479.515	595.195	.000
	Huynh-Feldt	63624.923	1.378	46165.900	595.195	.000
	Lower-bound	63624.923	1.000	63624.923	595.195	.000
pulse_exercise * pastsmok	Sphericity Assumed	123.384	2	61.692	1.154	.321
	Greenhouse-Geisser	123.384	1.312	94.014	1.154	.304
	Huynh-Feldt	123.384	1.378	89.527	1.154	.306
	Lower-bound	123.384	1.000	123.384	1.154	.290
Error(pulse_exercise)	Sphericity Assumed	3955.214	74	53.449		
	Greenhouse-Geisser	3955.214	48.559	81.452		
	Huynh-Feldt	3955.214	50.993	77.564		
	Lower-bound	3955.214	37.000	106.898		

(b) Using this model, is there evidence that being a past smoker influences pulse rate?

Null hypothesis: There is no difference in pulse rate between past smokers and those who have not ever smoked.
Alternative hypothesis: There is a difference in pulse rate between past smokers and those who have not ever smoked.

P-value = 0.278.
Fail to Reject Null Hypothesis.
Failed to find evidence that having smoked in the past effects pulse rate.

Tests of Between-Subjects Effects

Measure:MEASURE_1
Transformed Variable:Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	1188592.974	1	1188592.974	2542.052	.000
pastsmok	566.308	1	566.308	1.211	.278
Error	17300.171	37	467.572		

(c) Using this model, is there evidence of an interaction effect between type of exercise and whether or someone smoked in the past?

Null: There is no interaction effect between type of exercise and whether someone smoked in the past.
Alternative: There is an interaction effect between type of exercise and whether someone smoked in the past.
P-value = 0.304 (From pulse_exercise * pastsmoke) rows in "Tests of Within-Subjects Effects" table below)
Fail to Reject Null.
Failed to find statistical evidence for an interaction effect between type on exercise and whether or not someone is a past smoker.

4. Create a new variable called 'WEIGHTCAT' (for weight category) that is equal to 1 if BMI 35 (normal weight), 2 if 35 < BMI 40 (overweight), and 3 if BMI > 40 (obese). Develop a model that predicts average pulse while at rest taking into account gender and weight category.

Two-way Analysis of Variance. Factors = Gender and Weight. Dependent variable = Weight category

(a) Is there evidence that males and females have a different average pulse? If so, which gender has a higher pulse?

Null: There is no difference in average pulse between males and females.
Alternative: There is a difference in average pulse between males and females.
P-value = 0.001 ('Sex' variable in "Test of Between-Subjects Effects" table below)
Reject Null Hypothesis.
We found statistical evidence that there is a difference in average pulse between males and females.

(b) Is there evidence that weight category influences average pulse? If so, conduct post-hoc tests to indicate which weight levels lead to higher pulse rates.

Null: There is no difference in average pulse between people in different weight categories
Alternative: There is a difference in average pulse between people in different weight categories
P-value = 0.537 ('weightcat' variable in "Test of Between-Subjects Effects" table below)
Fail to Reject Null Hypothesis.
We failed to find statistical evidence that there is a difference in average pulse between people in different weight categories

Tests of Between-Subjects Effects

Dependent Variable:Pulse rate - resting (beats/min)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2963.405[a]	5	592.681	7.316	.000
Intercept	478310.431	1	478310.431	5904.596	.000
sex	1017.450	1	1017.450	12.560	.001
weightcat	101.174	2	50.587	.624	.537
sex * weightcat	513.885	2	256.942	3.172	.045
Error	12961.035	160	81.006		
Total	857423.000	166			
Corrected Total	15924.440	165			

a. R Squared = .186 (Adjusted R Squared = .161)

(c) Is there evidence for an interaction between gender and weight category? If so, examine the means (no hypothesis testing) and comment on the relationship between gender and weight.

Null: There is no interaction between weight category and gender.
Alternative: There is an interaction between weight category and gender.
P-value = 0.045 ('sex*weightcat' variable in "Test of Between-Subjects Effects" table above)
Reject Null Hypothesis.
We found statistical evidence that there is an interaction between weight category and gender.

Estimated Marginal Means

sex * weightcat

Dependent Variable:Pulse rate - resting (beats/min)

sex	weightcat	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
male	1.00	68.400	4.025	60.451	76.349
	2.00	70.333	1.964	66.455	74.212
	3.00	64.000	1.480	61.078	66.922
female	1.00	74.712	1.248	72.247	77.176
	2.00	72.469	1.591	69.327	75.611
	3.00	75.158	2.065	71.080	79.236

From the table above, we see there is are differences in pulse rate among men for different weight categories, but little difference among women. Obese men tend to have lower pulse rates than other men.