Kevin Scharnhorst
Med Inf 409
Spring 2011

Final Exam

DIRECTIONS:
The below attached object contains the original instructions for the final as well as the problems answered within the completed exam.

Final Exam
052611.pdf

Problem 1

Hypothesis

The problem considers a research study where a researcher is interested in explaining the effectiveness of Mannitol as a treatment for symptoms of Ciguatera poisoning. To consider the drug treatments effectiveness, the researcher studies the effects with 10 subjects (so, N=10). The dependent or outcome variable in our study is the change in observed number of neurological signs and symptoms. Our independent variables are time and more specifically the difference in times observed. The first measurement is captured as a Baseline and then 2.5 hours later our research subjects are tested again. Between the two collected measurements we are dealing with the same test subjects so we have cross over to consider.

The research and comparison between our two time intervals calls for a 95% confidence level. So to reject the null hypothesis, we would have to show a significant statistical difference exists between our baseline data collection and the collection taken 2.5 hours after. More specifically a significance factor of below 5% would need to exist. The hypothesis that we are testing then is below, where H_0 is the null hypothesis and H_A is the alternative hypothesis.

H_0 – Time does not play a significant factor in reducing the number of neurological signs and symptoms of ciguatera poisoning.
H_A – Time does play a significant factor.

Statistical Procedure, Tests and Assumptions

The conditions and parameters of our study from our hypothesis set the stage for considering what type of test would help us determine if our data suggest that there is any change between our two observation times in signs and symptoms of Ciguatera poisoning. Because cross over exists between our two observations with the same subjects, the correct test to perform is a **Paired Samples T-Test**. We can further gain comfort with the chosen test method by considering the assumptions that the test expects. Those assumptions are as follows.

Assumptions:

- Normally distributed population
- Our observations are independent
- Simple random sample of 10 subjects were taken from the larger normally distributed population
- Equality of variances exist between our groups
- Number of neurological signs and symptoms measured (dependent variable) is continuous (interval or ratio)
- Observation time (independent variable) is categorical, with 2 levels (Baseline & 2.5 Hours Later)

After considering the expected assumptions, a Paired Samples T-Test fits all.

Results

In evaluating the results of our descriptive statistics from our Paired Samples T-Test from the table below, we can see a comparison of means across our two observation points show a clear difference with the trend declining in number of signs and symptoms over time with 2.5 less recorded over 2.5 hours (highlighted).

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Baseline	**8.10**	10	3.035	.960
	2.5 hours	**5.60**	10	2.797	.884

By further examining the results of the Pair Samples T-Test results, we can conclude with the significance factor value from the 2-tailed test being .021 or 2.1 that time does play a significant factor in the number of observed neurological signs and symptoms in ciguatera poisoning. 2.1% is lower than our 5% requirement, so we would **therefore reject the null hypothesis and accept the alternative**.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Baseline - 2.5 hours	2.500	2.838	.898	.470	4.530	2.785	9	**.021**

Inference

The study sought to explain if time showed affected the number of neurological signs and symptoms in ciguatera poisoning. The resulting analysis does show that time plays a significant factor. We cannot assume or infer that lower number of signs are better or worse, but can say that time is a significant parameter to consider in the number of signs and symptoms observed with Mannitol as a treatment for Ciguatera Poisoning.

Problem 2

Hypothesis

The problem considers a research study that seeks to understand bank teller wait times. The bank manager is our researcher and is responsible for sampling a random sample of customers across three random times of the day consisting of morning, afternoon and evening. The study seeks to simply explain if there is significant evidence to show that there is a difference in wait times that a customer might experience in waiting to see a bank teller based on the time of day they visit the bank.

The research and comparison of correlation of our variables calls for a 95% confidence level. So to reject the null hypothesis, we would have to show a significant statistical difference exists to show that wait time is correlated with wait time to see a bank teller. More specifically a significance factor of below 5% would need to exist. The hypothesis that we are testing then is below, where H_0 is the null hypothesis and H_A is the alternative hypothesis.

H_0 – No significant relationship exists between the wait time to see a bank teller and the time of day.
H_A – Time of day does make a difference in the average wait time to see a bank teller.

Statistical Procedure, Tests and Assumptions

On the surface, determining the appropriate type of test to perform on the data to determine if there is a difference in wait times at different parts of the day is a tough one. An analysis of variance (ANOVA) seems appropriate because you have what appears to be a categorical variable with time of day where we have categorized observations into buckets consisting of morning, afternoon and evening. However, ANOVA does not consider the continuity of time. So, eliminating time of day as a categorical variable leaves us to consider it as a continuous. Wait Time is our other variable and is clearly continuous. With two continuous variables to consider, Linear Regression appears to be the appropriate choice. In further considering the application of Linear Regression it is appropriate where we seek to explain the correlation of variables. That is, does the time spent waiting for a bank teller change based on the time of day? Through regression analysis we can see if an inverse relationship exists between the time of day and wait time experienced. If visited earlier or later, does the wait time increase or decrease on average? Our study also fits an experimental method where we are seeking to explain if X (time of day) causes Y (longer or shorter wait times).

All of the stated reasons more comfortably match with what Linear Regression is built to explain. We can further match the assumptions that are made with Linear Regression by making sure our study exhibits the following

Assumptions:

- Normality
- Independence
- Random Sampling
- Constant Variance
- Linearity

All of our assumptions seem to hold true so we proceed with confidence in using **Linear Regression** to explain our correlation of wait time on the time of day.

Results

In reviewing the model summary from our Linear Regression test, we can see that only 6.5% (highlighted) of our variability is explained by our model. This is a very weak correlation.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.254[a]	**.065**	.062	1.05718

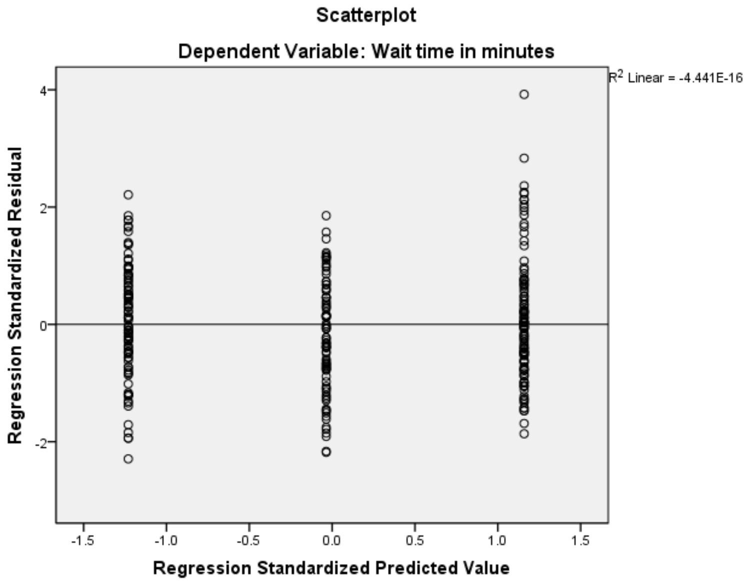
a. Predictors: (Constant), Time of Day

b. Dependent Variable: Wait time in minutes

To help visualize our correlation, we can examine a scatter plot with standardized residuals over standardized predicted values and see that we have inverse relationship. Meaning, that the time of day, inversely affects the amount of wait time to see a bank teller. This is also shown in the correlations table and we see that the inverse relationship is -.254 (highlighted).

Correlations

		Wait time in minutes	Time of Day
Pearson Correlation	Wait time in minutes	1.000	-.254
	Time of Day	**-.254**	1.000
Sig. (1-tailed)	Wait time in minutes	.	.000
	Time of Day	.000	.
N	Wait time in minutes	299	299
	Time of Day	299	299



Our ANOVA table from the Linear Regression results shows that we have a significance factor of .000 which is less than 5%. We **would reject the null hypothesis and accept the alternative** which accepts that time of day does play a significant factor in wait time to see a teller.

The horizontal line shown in our scatter plot represents the prediction equation. We can create our prediction equation from our coefficients table below.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	6.086	.157		38.870	.000	5.778	6.394
	Time of Day	-.332	.073	-.254	-4.535	.000	-.476	-.188

a. Dependent Variable: Wait time in minutes

Our equation then is Wait time (in minutes) = 6.086 + (-.332 * time of day)

<u>Inference</u>

From the above interpretation of our results, we can infer that the relationship between time and day and expected wait time is weak, but that the relationship is significant enough to be considered. We can use the stated prediction equation to estimate how much less time a person can expect to wait as the time of day continues. Given the weak relationship, the results of this study might encourage the bank manager to further consider other factors such as the number of tellers at each counter at various times of the day.

Doing so might help him/her find other significant factors that might help lead to better efficiencies and greater customer satisfaction.

Problem 3

Hypothesis

The study considered, seeks to explain the effects nitrogen dioxide (NO_2) on lung damage. To explain the effects, the researcher used mice as test subjects. Ten mice were exposed directly to NO_2 and the same subjects observed over three different periods of time (10, 12 and 14 days) of exposure. The percentage of serum fluorescence reading serves as our measurement in determining the severity of damage after exposure time.

The research and comparison between our three time intervals calls for a 95% confidence level. So to reject the null hypothesis, we would have to show a significant statistical difference exists between two or more means across the three observations. More specifically a significance factor of below 5% would need to exist. The hypothesis that we are testing then is below, where H_0 is the null hypothesis and H_A is the alternative hypothesis.

H_0 – The mean of serum fluorescence values are the same across our three observed times ($M_1 = M_2 = M_3$)
H_A – Two or more means differ across serum fluorescence values measured in our three observed times

Statistical Procedure, Tests and Assumptions

In consideration of the study parameters and conditions stated, we observe that we have a continuous dependent variable (Lung Damange) and a categorical independent variable (Exposure Time) having more than 2 levels (10, 12 and 14 days). This leads me to believe that analysis of variance (ANOVA) is the appropriate test to perform. By considering further the assumptions that ANOVA expects, it is safe to conclude that ANOVA will be appropriate to compare the mean pollutant readings and determine its effects on lung damage. Because we have cross over with the same test subjects being measured across three collections, a **randomized block ANOVA** is the best choice for test method.

Assumptions:

- Normally distributed population
- Our groups are dependent (crossover)
- Simple random sample of 10 subjects were taken from the larger normally distributed population
- Equality of variances exist between our groups
- Lung Damage (dependent variable) is continuous (interval or ratio)
- Exposure Time (independent variable) is categorical, with 2 or more groups to compare (10, 12 and 14 days)

Results

After setting up our data and running the randomized block ANOVA test on it, we can evaluate the results. In considering the Test of Between-Subjects Effects table below we can immediately see that our days factor has a significance factor of .013 or 1.3% (highlighted). Since 1.3% is less than our required 5%, we can confidently **reject the null hypothesis and accept the alternative** which assumes that the mean averages differ across two or more observations.

Tests of Between-Subjects Effects

Dependent Variable:Reading

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	20819.233[a]	11	1892.658	1.747	.142
Intercept	463514.700	1	463514.700	427.749	.000
Days	11993.600	2	5996.800	5.534	.013
Block	8825.633	9	980.626	.905	.541
Error	19505.067	18	1083.615		
Total	503839.000	30			
Corrected Total	40324.300	29			

a. R Squared = .516 (Adjusted R Squared = .221)

Realizing that our mean average of serum fluorescence measurements were statistically significantly different, we can further analyze the rest of our results. The Turkey post-hoc results can show us that our difference exists between days 10 and 12 of being exposed to our pollutant. We could use the multiple comparisons table to highlight this or the Homogeneous Subsets does the same. From the table below, we can see that we have two subsets to consider. The table demonstrates that days 14 and 10 are statistically equivalent. As well, equivalents exist between days 10 and 12, but days 14 and 12 are statistically different in means comparison.

Homogeneous Subsets (Reading)

Tukey HSD[a,b]

Days	N	Subset 1	Subset 2
14 Days	10	98.70	
10 Days	10	126.70	126.70
12 Days	10		147.50
Sig.		.167	.355

Means for groups in homogeneous subsets are displayed.

Based on observed means.

The error term is Mean Square(Error) = 1083.615.

a. Uses Harmonic Mean Sample Size = 10.000.

b. Alpha = .05.

Inference

In evaluating our results and concluding on our hypothesis we can infer that the effect of lung damage does depend on exposure time to NO_2. The mean measurement reading of Serum Fluorescence increased from 10 days of exposure to 12, but then declines from 12 to 14. The researcher would have to determine the cause of the drop, but the test can safely infer that the exposure in days played a significant factor in showing a difference in mean averages compared across or three observed days of exposure.

Problem 4

Hypothesis

In this study, the researcher seems to be seeking some validation on his/her collected data by explaining if the mean time to accelerate from 0 to 60 is equivalent to 15 seconds across all cases. The comparison value is known.

The objective is to compare the means of the cars measured to see if their average mean equals 15 seconds. Our study for a 95% confidence level. So to reject the null hypothesis, we would have to show a significant statistical difference exists between in the means comparison. More specifically a significance factor of below 5% would need to exist. The hypothesis that we are testing then is below, where H_0 is the null hypothesis and H_A is the alternative hypothesis.

H_0 – The mean acceleration from 0 to 60 is equal to 15 seconds.
H_A – The mean acceleration from 0 to 60 is not equal to 15 seconds.

Statistical Procedure, Tests and Assumptions

The parameters of our study suggest the appropriate test method. Our parameters include a known comparison mean of 15 seconds. A random sample of acceleration times among a random sample of cars (N=406). These factors strongly suggest that we want to perform a One Samples T-Test. We can further validate the appropriateness of our chosen test by checking our assumptions of the following conditions.

Assumptions:

- Normally distributed population
- Outcome variable (average acceleration time) is continuous (interval or ratio)
- We wish to test if our sample observations (average acceleration time) were drawn from a population of a specific mean value (15 seconds)
- Acceleration Time (dependent variable) is continuous (interval or ratio)
- Car or vehicle (independent variable) is categorical, with 2 or more groups to compare (4 Devices)

All of the above assumptions hold true and we can safely proceed with a **One Samples T-Test**.

Results

After performing our One Samples T-Test, we can evaluate the results and determine from the descriptive statistics that the mean acceleration from 0 to 60 is 15.50 seconds (highlighted).

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Time to Accelerate from 0 to 60 mph (sec)	406	**15.50**	2.821	.140

That is different from 15 seconds, so we have to evaluate the results further to interpret our acceptance or rejection of the null hypothesis. The One-Sample Test results below shows that our significance factor is .000 or 0%. We would therefore **reject the null hypothesis and accept the alternative** which assumes that the mean acceleration times were statistically different from 15 seconds.

One-Sample Test

	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Time to Accelerate from 0 to 60 mph (sec)	110.677	405	**.000**	15.495	15.22	15.77

<u>Inference</u>

Although the mean acceleration time from 0 to 60 only differed by a half a second, our significance factor shows that that is enough to show that it is significantly different from our predetermined value of 15 seconds. So, we know from this that our mean in the cars sampled for mean acceleration times are significantly different from 15 seconds.

Problem 5

Hypothesis

The study considered, continues from the research study data considered in problem 4. We have a random sample of 406 cars (so, N=406) and we are seeking to develop a prediction model that can calculate the time to accelerate from 0 to 60 mph. Our predictive values are Miles Per Gallon, Engine Displacement, Horsepower and Number of Cylinders. In order to build a predictive equation our model must be considered to ensure that our predictors fit and are significantly related in determining acceleration time. The study as setup would consider the below hypothesis where H_0 is the null hypothesis and H_A is the alternative hypothesis.

H_0 – There is no significant relationship in predicting acceleration time from 0 to 60 from the four predictor variables.
H_A – There is a significant relationship in predicting the acceleration time from 0 to 60 from the four predictor variables.

Statistical Procedure, Tests and Assumptions

Given the objective of our study is to produce a predictive equation to estimate acceleration from 0 to 60 from miles per gallon, engine displacement, horsepower and number of cylinders, it is pretty straight forward to choose Linear Regression as our test method. Our dependent and independent variables are all continuous. Linear regression will explain to us the strength and relationship of our predictive values in estimating our dependent variable of acceleration time. After understanding the correlation we can better understand its relationship in calculating our outcome variable. The Linear Regression model comes with assumptions and to ensure we've picked the proper test method we need be comfortable with them. Those assumptions are as follows.

Assumptions:

- *Normality*
- *Independence*
- *Random Samples*
- *Constant Variance*
- *Linearity*

Based on what has been stated and after further considering the assumptions that we are making from above, there is no reservation in using **Linear Regression** to develop our predictive equation to determine acceleration time from 0 to 60 mph.

Results

To determine the fit of our model in developing a predictive equation to determine acceleration from 0 to 60 mph we can start with the model summary from our produced results.

Model Summary[b]

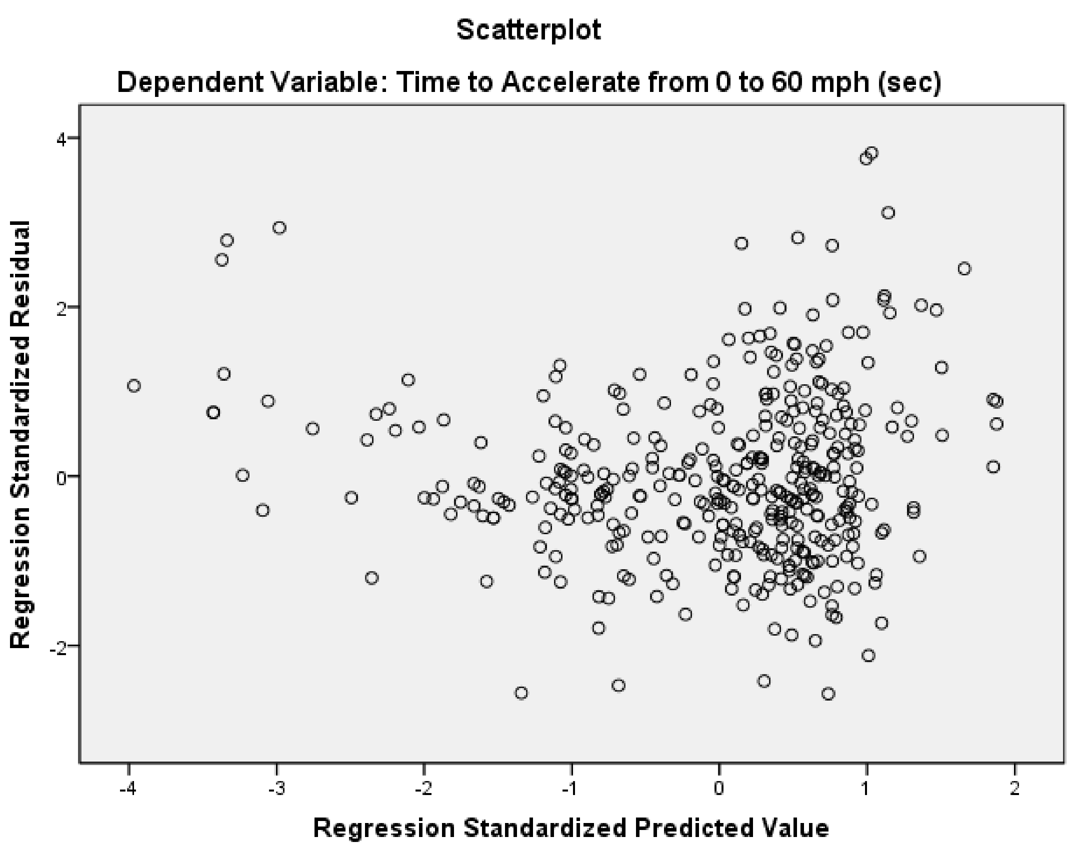
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.735[a]	.541	.536	1.879

a. Predictors: (Constant), Number of Cylinders, Miles per Gallon, Horsepower, Engine Displacement (cu. inches)

b. Dependent Variable: Time to Accelerate from 0 to 60 mph (sec)

From the R Square value above of .541 or 54.1% (highlighted) we can tell that 54% of our variability is explained by our model. This is considered to be a moderate amount of explanation and so it is appropriate to continue evaluating our results further.

With Linear Regression models, scatter plots help to show non-significant linear regression that might be hiding in our collected data points. We can use a plot showing the Standardized Residual (y axis) over Standardized Predicted (x axis) to show any peculiarities with our data. The plot is demonstrated below.



No clear patterns are apparent, so it is further okay to continue with our analysis. The Coefficients tables will show the significance of our predictors in determining our predictive equation to calculate acceleration time from 0 to 60 mph (sec). From the below, we can see that all but one of our considered four variables are significant. Miles per Gallon (.1%), Engine Displacement (.8%) and Horsepower (0%) are all lower than our required 5%. Number of Cylinders has a significance of 60.8% which demonstrates no significance in predicting acceleration time from 0 to 60.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	24.426	.996		24.522	.000	22.467	26.384
	Miles per Gallon	-.072	.021	-.203	-3.411	.001	-.114	-.031
	Engine Displacement (cu. inches)	.010	.004	.369	2.648	.008	.003	.017
	Horsepower	-.082	.006	-1.142	-14.157	.000	-.094	-.071
	Number of Cylinders	-.093	.182	-.058	-.513	.608	-.451	.264

a. Dependent Variable: Time to Accelerate from 0 to 60 mph (sec)

We could further examine our ANOVA table from below and see that it too suggests that our model is significant with a value of .000 or 0% (highlighted). Since 0% is less than our required 5% and we know that we had one or more predictive variables show a significant relationship in predicting acceleration time, we would **reject the null hypothesis and accept the alternative.**

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1603.854	4	400.963	113.519	.000[a]
	Residual	1363.401	386	3.532		
	Total	2967.254	390			

a. Predictors: (Constant), Number of Cylinders, Miles per Gallon, Horsepower, Engine Displacement (cu. inches)

b. Dependent Variable: Time to Accelerate from 0 to 60 mph (sec)

Being that we identified Number of Cylinders as not being significant, we would remove that from our model and run the same analysis. After doing so we can see that our Model Summary below shows that our R Square value went down slightly to .531 or 53.1% (highlighted). This shows that our eliminated variable did little to explain our variability in the overall model.

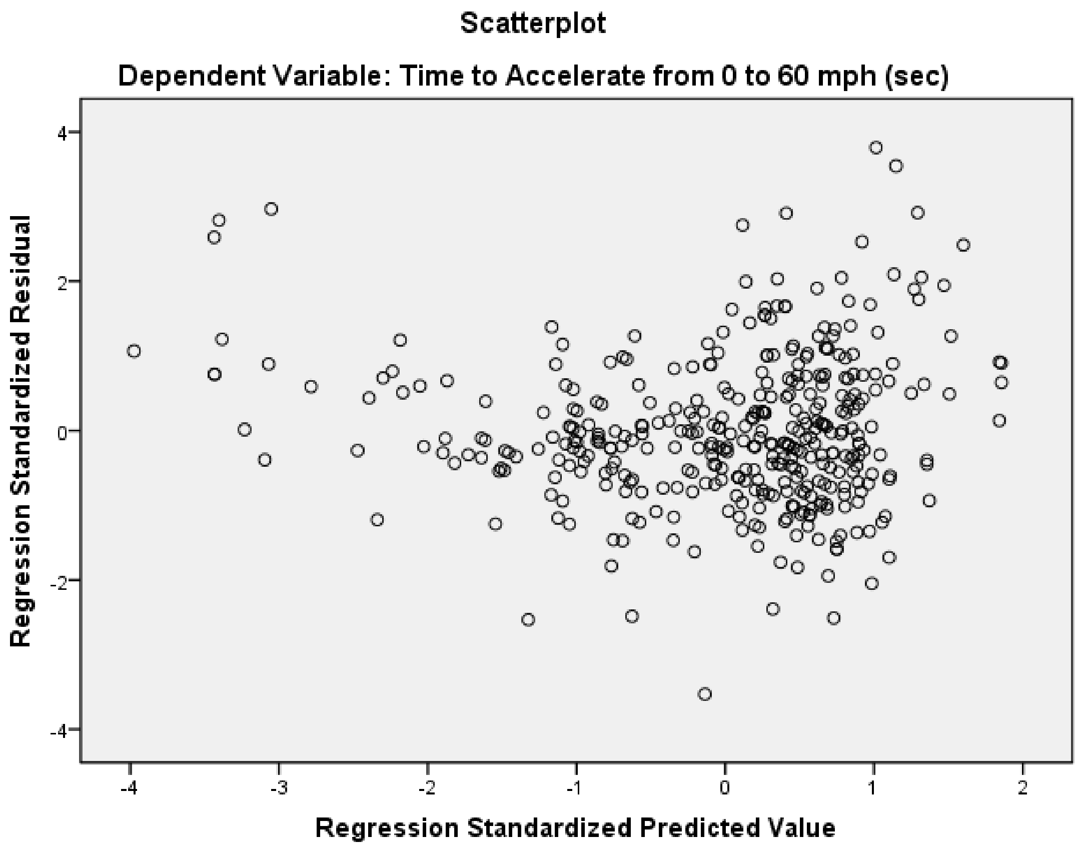
Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.729[a]	.531	.528	1.909

a. Predictors: (Constant), Horsepower, Miles per Gallon, Engine Displacement (cu. inches)

b. Dependent Variable: Time to Accelerate from 0 to 60 mph (sec)

In examining our scatter plot again, it still looks good with no apparent pattern to show non-linear regression.



Our Coefficients table will tell us the significance of our predictors in calculating our outcome variable of acceleration time. With Number of Cylinders eliminated, all the remaining variables show they are statistically significant in predicting acceleration time. We can then proceed to the goal of our model and develop a predictive equation for our dependent variable.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	23.479	.837		28.047	.000	21.833	25.125
	Miles per Gallon	-.052	.021	-.145	-2.496	.013	-.092	-.011
	Engine Displacement (cu. inches)	.010	.002	.392	4.697	.000	.006	.015
	Horsepower	-.084	.006	-1.157	-14.369	.000	-.096	-.073

a. Dependent Variable: Time to Accelerate from 0 to 60 mph (sec)

Time to Accelerate from 0 to 60 mph (sec) = 23.479 + (-.052 * Miles per Gallon) + (.010 * Engine Displacement) + (-.084 * Horsepower)

Inference

In setting up the model and evaluating it for fit, we can conclude that after eliminating Number of Cylinders that Miles per Gallon, Engine Displacement and Horsepower are all effective predictors in determining acceleration time from 0 to 60 mph (sec). The equation is **Time to Accelerate from 0 to 60 mph (sec) = 23.479 + (-.052 * Miles per Gallon) + (.010 * Engine Displacement) + (-.084 * Horsepower)**. Both Horsepower and Miles per Gallon have an inverse relationship to acceleration time and Engine displacement is positive. In other words, the more gas and horsepower you have with low engine displacement (piston producing power by mixing air and gasoline), the quicker you will get from 0 to 60 mph (sec). Seems pretty intuitive and the model helps explain it. This was an enjoyable problem!

Problem 6

Hypothesis

The study seeks to develop a predictive equation for determine self-esteem from three predictor variables. The predictor variables are locus of control, alienation and social ability. The study considers 56 subjects (so, N=56) and each were given an assessment to measure their self-esteem, locus of control, alienation and social ability. The study as setup would consider the below hypothesis where H_0 is the null hypothesis and H_A is the alternative hypothesis.

H_0 – There is no significant relationship in predicting self-esteem from the three predictor variables.
H_A – There is a significant relationship in predicting the self-esteem from the three predictor variables.

Statistical Procedure, Tests and Assumptions

Given the objective of our study is to produce a predictive equation to self-esteem from locus of control, alienation and social ability, Linear Regression is a strong candidate for a test method. In further examining the data from the study, our dependent variable (self-esteem) as well as the three predictors all appear to be continuous in nature. Linear regression will explain to us the strength and relationship of our predictive values in estimating our dependent variable of self-esteem. After understanding the correlation we can better understand its relationship in calculating our outcome variable. The Linear Regression model comes with assumptions and to ensure we've picked the proper test method we need be comfortable with them. Those assumptions are as follows.

Assumptions:

- *Normality*
- *Independence*
- *Random Samples*
- *Constant Variance*
- *Linearity*

Now that we have evaluated the conditions of our study and examined the data for our variables, it is clear that **Linear Regression** is the correct test to perform to evaluate the correlation of our predictor variables in predicting self-esteem.

Results

To examine the initial fit of our model, we can first look at the model summary. In considering the R Square value of .531 we can see that 53.1% of the variability can be explained by the model. 53.1% (highlighted) is a moderate amount and so we continue to look at the results.

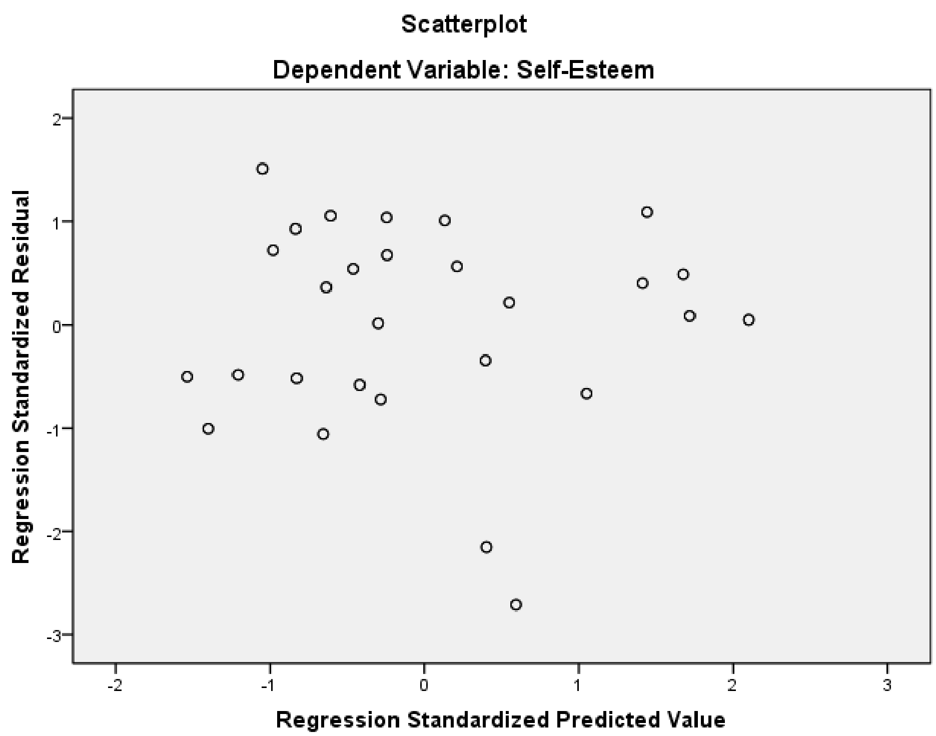
Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.729[a]	.531	.504	2.77751

a. Predictors: (Constant), Social Ability, Alienation, Locus of Control

b. Dependent Variable: Self-Esteem

It is important with Linear Regression to visualize our data and to see if any apparent patterns exist. From the scatter plot below we have plotted the Standardized Residual (y axis) over Standardized Predicted (x axis) to show any peculiarities with our data. The plots are not concerning and all center around zero so it is okay to continue further in examining our results.



We can look at the Coefficients table below to determine the significance of our predictors. Quickly, we can see from below that both Locus of Control and Social Ability have values that are greater than our required 5%. Locus of Control being 60.8% and Social Ability being 18.3% (highlighted in yellow). Alienation remains and has a significance factor of .000 or 0% (indicated in red). Since our null hypothesis assumed that none of our predictors would be significant in estimating self-esteem and we have one that is, we would **reject the null hypothesis and accept the alternative**.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8.649	6.201		1.395	.169
	Locus of Control	-.107	.208	-.073	-.517	.608
	Alienation	5.368	1.015	.603	5.288	.000
	Social Ability	.087	.065	.169	1.349	.183

a. Dependent Variable: Self-Esteem

Our ANOVA table also shows that the model is significant (highlighted) with a value of .000 or 0%. This further validates our decision to reject the null hypothesis and accept the alternative.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	455.057	3	151.686	19.662	.000[a]
	Residual	401.157	52	7.715		
	Total	856.214	55			

a. Predictors: (Constant), Social Ability, Alienation, Locus of Control

b. Dependent Variable: Self-Esteem

To continue in developing our predictive equation, we now want to eliminate our non-significant predictors of Locus of Control and Social Ability and interpret the results again to see how our model fits. In looking at our R Square value again, we see that it went down to .491 or 49.1% (highlighted), but only slightly. This shows that the eliminated variables were not showing significant explanation for our variability. 49% is still moderate and so it is appropriate to continue and plot our data points to look for any patterns.

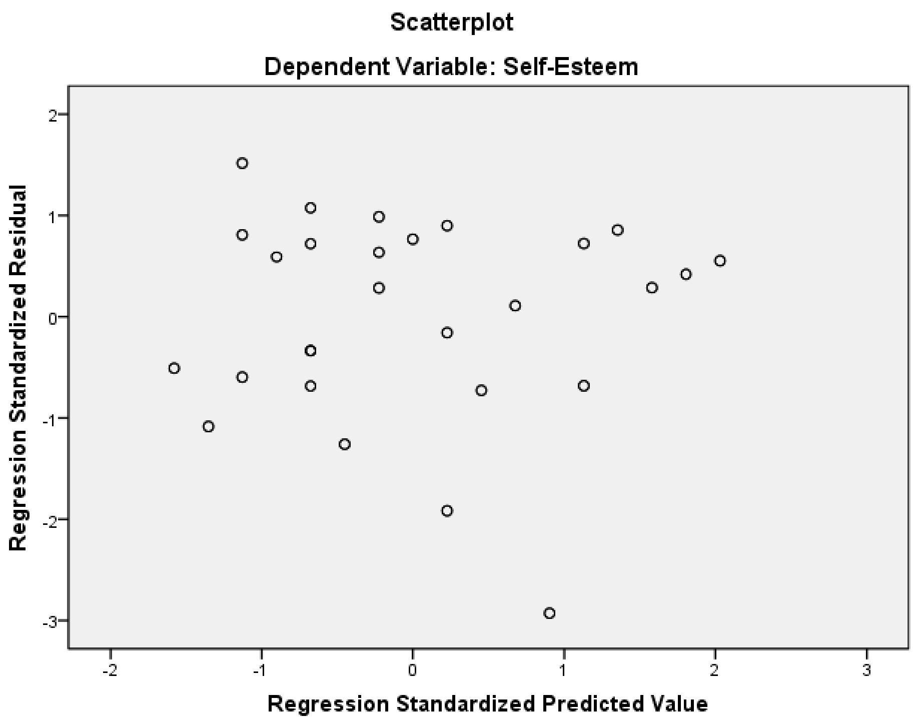
Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.701[a]	.491	.482	2.84015

a. Predictors: (Constant), Alienation

b. Dependent Variable: Self-Esteem

From the below plot, our data still looks okay with no apparent pattern existing and all data centered on or around zero. All data being within two standard deviations and the majority within one.



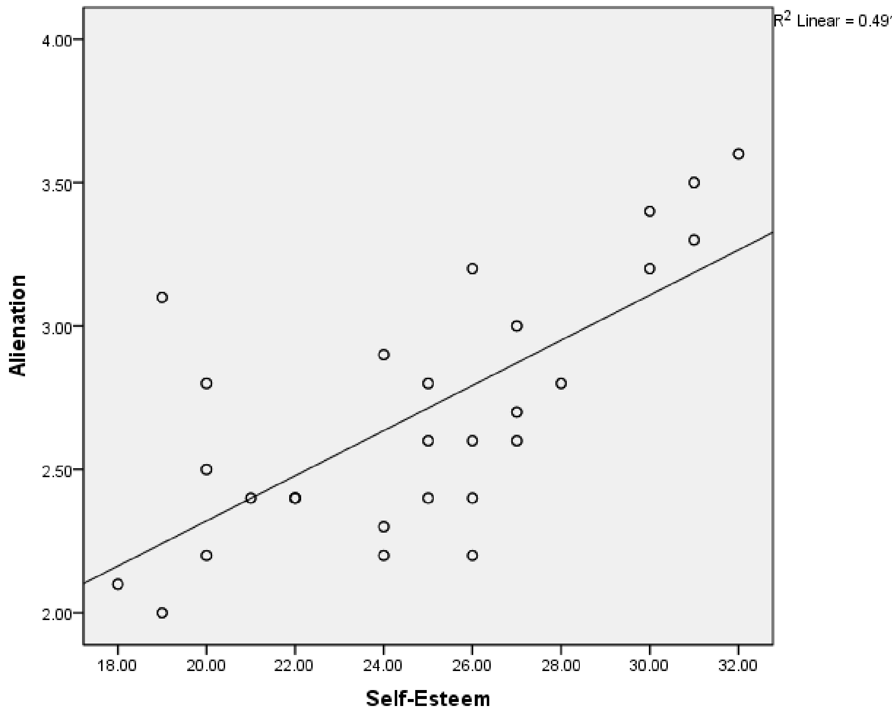
We can then continue to develop our predictive equation by evaluating the Coefficients table below. We can see that Alienation still remains significant at 0% (highlighted).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.971	2.364		3.372	.001
	Alienation	6.241	.864	.701	7.221	.000

a. Dependent Variable: Self-Esteem

Our predictive equation then is plotted and stated below.



Self-Esteem = 7.971 + (6.241 * Alienation)

<u>Inference</u>

From the study and in interpreting our results we can infer that Alienation is a significant predictor in determine self-esteem measurements but Locus of Control and Social Ability are not. Our predictive equation is **Self-Esteem = 7.971 + (6.241 * Alienation).** From the predictive equation we can see that a positive slope exists. That is to say, that, as a person experiences a unit increase of alienation factor their self-esteem factor increases. This seems counter intuitive, but maybe the results that being around one peers makes one question their self-esteem as a person struggles to fit in. The study does not to explain why, but merely shows that a significant statistical relationship exists with alienation being able to predict self-esteem.

Problem 7

Hypothesis

The study considers 210 women that were selected by a researcher to determine the relationship between ductal carcinoma and family history of breast cancer. The study seeks to explain the relationship between our two variables and to also calculate and interpret an odds ratio of having ductal carcinoma for woman that have a family history of having it. Such a study might prove useful to women who have a family history of breast cancer and wish to predict their statistical risks of having ductal carcinoma.

The study calls for a 95% confidence level. So to reject the null hypothesis, we would have to show a significant statistical difference exists between ductal carcinoma and family history of breast cancer. More specifically a significance factor of below 5% would need to exist. The hypothesis that we are testing then is below, where H_0 is the null hypothesis and H_A is the alternative hypothesis.

H_0 – Ductal carcinoma and family history of breast cancer are independent of each other with no relationship existing.
H_A – There is a significant relationship that exists between ductal carcinoma and family history of breast cancer

Statistical Procedure, Tests and Assumptions

The way in which the study is setup and seeks to establish if a relationship exists between our two independent observations strongly suggest that a Chi-Square test is the appropriate test to perform. Chi-Square is used to analyze data that is recorded on a nominal scale. Both of our variables are a frequency count of subjects having either a family history of breast cancer or ductal carcinoma with values indicated as “yes” or “no” for either. A Chi-Square test make the following assumptions, so we must be willing to accept these.

Assumptions:

- Simple random sample
- Independent observations
- Each subject contributes data to only one cell
- Expected cell frequencies are ≥ 5

The parameters of our test , assumptions above and nature of what we are trying to prove in outcome fits well with what a **Chi-Square test** can provide. Specifically we will be using the Chi-Square to perform an independence test. The independence test seeks to explain if there is a relationship between two variables.

Results

From the below Chi-Square Test we can see that a significant relationship is shown between ductal carcinoma and a family history of breast cancer. The Pearson Chi-Square significance factor is .029 or 2.9% (highlighted). Since 2.9% is less than our required 5% the relationship is significant. The null hypothesis would have suggested that there was no relationship so we reject the null hypothesis and accept the alternative which assumes that there is.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4.786[a]	1	**.029**		
Continuity Correction[b]	3.258	1	.071		
Likelihood Ratio	3.799	1	.051		
Fisher's Exact Test				.045	.045
Linear-by-Linear Association	4.764	1	.029		
N of Valid Cases	210				

a. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 2.14.

b. Computed only for a 2x2 table

Calculating an odds ratio would give us a better feeling for the relationship between ductal carcinoma and a family history of breast cancer. To calculate an odds ratio we can consider the below cross table. When a family history of breast cancer exists, the odds of ductal carcinoma occurring was 5 to 25 or 20% chance. When there were no family history of breast cancer odds were reduced to ductal carcinoma occurring was 10 to 170 or 5.9% chance.

Family History of Breast Cancer * Ductal Carcinoma Crosstabulation

			Ductal Carcinoma		
			Yes	No	Total
Family History of Breast Cancer	Yes	Count	5	25	30
		% within Family History of Breast Cancer	16.7%	83.3%	100.0%
	No	Count	10	170	180
		% within Family History of Breast Cancer	5.6%	94.4%	100.0%
Total		Count	15	195	210
		% within Family History of Breast Cancer	7.1%	92.9%	100.0%

Within our results, we are able to produce a risks estimate table and from it we can see that the odds ratio for family history of breast cancer is 3.4 (highlighted). If computed manually the odds ratio is figured by taking the considered odds above by taking $\frac{5 \div 25}{10 \div 170} = 3.4$ The below risk estimate table (produced by SPSS) shows the same matching odds factor (highlighted in red). Our risk factor would be 1.818 (highlighted in green).

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Family History of Breast Cancer (Yes / No)	3.400	1.074	10.767
For cohort Ductal Carcinoma = Yes	3.000	1.102	8.167
For cohort Ductal Carcinoma = No	.882	.749	1.040
N of Valid Cases	210		

Inference

The study sought to explain if a relationship existed between ductal carcinoma and a family history of breast cancer. We can infer that a relationship does exist from the considered results. With our calculated odds ratio, we can also say that with women having a family history of breast cancer, they are 3.4 times more likely to get ductal carcinoma than with women that don't have a family history of breast cancer.