MSDS 6372 Project 3

Team members:

Killion, Kyle and Taylor, Celia

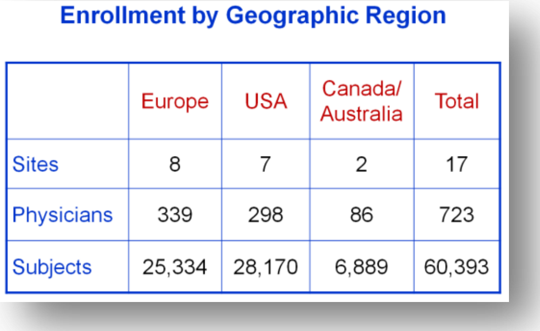
Date: December 9, 2016

Bone Density – Osteoporosis in Women ages 55 and older

Introduction

Between 2005 and 2014, the Global Longitudinal Study of Osteoporosis in Women (GLOW) conducted a study to further an understanding of ways to prevent and minimize the probabilities of fracture related incidences associated with osteoporosis. This disease heavily impacts roughly 200 million people and life expectancies continue to keep growing. It is vital to gather further knowledge and awareness to improve quality of life within the golden years.

The design of experiment considered a wide variety of diverse observations. Among 10 countries sampled (Australia, Belgium, Canada, France, Germany, Italy, Netherlands, Spain, UK, and USA), Figure 1 shows how the study sample is broken down by Geographic Regions. These patients were sampled through physician practices and were a representative sample of the practice within its region. These physicians possessed a high level of expertise over the field of osteoporosis. The sites were filtered by its ability to supervise assessments and conduct treatment options which would be the data supplied for comparisons with the GLOW sample. By doing this, the available sites became somewhat limited when selecting them for the analysis (2010, University of Massachusetts Medical School).

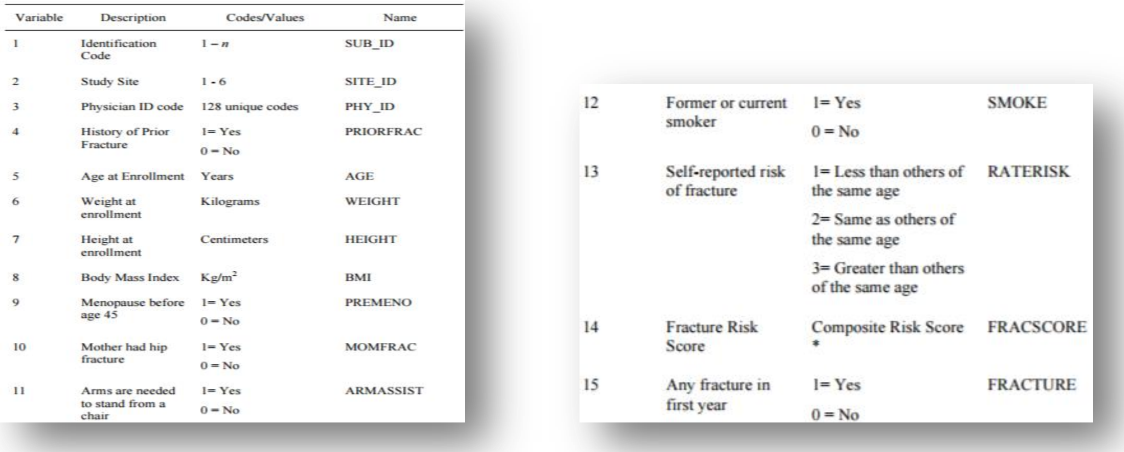

Enrollment by Geographic Region

	Europe	USA	Canada/ Australia	Total
Sites	8	7	2	17
Physicians	339	298	86	723
Subjects	25,334	28,170	6,889	60,393

Figure 1 - (2010, University of Massachusetts Medical School, Slide 16)

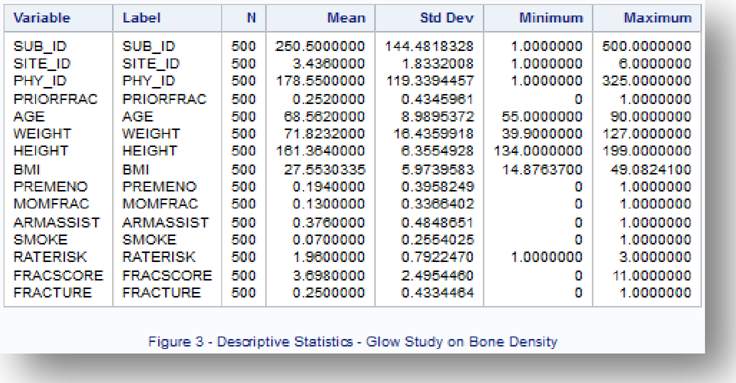
Our intentions are to explore this data set with logistic regression techniques to manufacture a model that provides insight over subject characteristics in relation with osteoporosis. Determining odds ratios, probabilities, and accurately discerning which lifestyle factors that are more threatening can further assist the fight against osteoporosis.

Descriptive Statistics - Initial Survey, Correlation, and Variance Inflation



Variable	Description	Codes/Values	Name
1	Identification Code	$1 - n$	SUB_ID
2	Study Site	1 - 6	SITE_ID
3	Physician ID code	128 unique codes	PHY_ID
4	History of Prior Fracture	1= Yes 0 = No	PRIORFRAC
5	Age at Enrollment	Years	AGE
6	Weight at enrollment	Kilograms	WEIGHT
7	Height at enrollment	Centimeters	HEIGHT
8	Body Mass Index	Kg/m^2	BMI
9	Menopause before age 45	1= Yes 0 = No	PREMENO
10	Mother had hip fracture	1= Yes 0 = No	MOMFRAC
11	Arms are needed to stand from a chair	1= Yes 0 = No	ARMASSIST
12	Former or current smoker	1= Yes 0 = No	SMOKE
13	Self-reported risk of fracture	1= Less than others of the same age 2= Same as others of the same age 3= Greater than others of the same age	RATERISK
14	Fracture Risk Score	Composite Risk Score *	FRACSCORE
15	Any fracture in first year	1= Yes 0 = No	FRACTURE

Figure 2 **- Data Overview (2004, University of Massachusetts Amherst, pg.1-2)**



Variable	Label	N	Mean	Std Dev	Minimum	Maximum
SUB_ID	SUB_ID	500	250.5000000	144.4818328	1.0000000	500.0000000
SITE_ID	SITE_ID	500	3.4360000	1.8332008	1.0000000	6.0000000
PHY_ID	PHY_ID	500	178.5500000	119.3394457	1.0000000	325.0000000
PRIORFRAC	PRIORFRAC	500	0.2520000	0.4345961	0	1.0000000
AGE	AGE	500	68.5620000	8.9895372	55.0000000	90.0000000
WEIGHT	WEIGHT	500	71.8232000	16.4359918	39.9000000	127.0000000
HEIGHT	HEIGHT	500	161.3640000	6.3554928	134.0000000	199.0000000
BMI	BMI	500	27.5530335	5.9739583	14.8763700	49.0824100
PREMENO	PREMENO	500	0.1940000	0.3958249	0	1.0000000
MOMFRAC	MOMFRAC	500	0.1300000	0.3366402	0	1.0000000
ARMASSIST	ARMASSIST	500	0.3760000	0.4848651	0	1.0000000
SMOKE	SMOKE	500	0.0700000	0.2554025	0	1.0000000
RATERISK	RATERISK	500	1.9600000	0.7922470	1.0000000	3.0000000
FRACSCORE	FRACSCORE	500	3.6980000	2.4954460	0	11.0000000
FRACTURE	FRACTURE	500	0.2500000	0.4334464	0	1.0000000

Figure 3 - Descriptive Statistics - Glow Study on Bone Density

During the exporatory data analysis (EDA), we gathered variable information from variable descriptions in Figure 2, descriptive statistics in Figure 3, and scatter and histogram in Figure 4. Throughout the exporatory data analysis (EDA), 7 categorical variables were observed which included Prior History of Fracture (PRIORFRAC), Menopause before age 45 (PREMENO), Mother had hip fracture (MOMFRAC), Arm rests are needed in order to stand from chair (ARMASSIT), Former or current Smoker (SMOKE), Self-reported risk of fracture (RATERISK), and any fracture in the first year (FRACTURE). The dependent variable was deemed the FRACTURE variable in order to conduct the study utilizing logistic regression. Within the study and considering the age of the subjects, the average age of women came to 68.562 years, and 61% of the women are 65 years of age or older out of the 60,393 women who registered to participate in the study (2010, University of Massachusetts Medical School, Slide 14). The data set had also been broken down statistically when emphasizing the women and their BMI.

> *"Researchers evaluating fracture history and body mass index (BMI) among 44,534 women in the multinational, prospective GLOW study found that 23.4% of the subjects were obese (body mass index [BMI] over 30 kg/m2), 74.9% were not obese (BMI 18.5-29.9 kg/m2), and 1.7% were underweight (BMI under 18.5 kg/m2)"* (2011, September 20, Melville, N. A.).

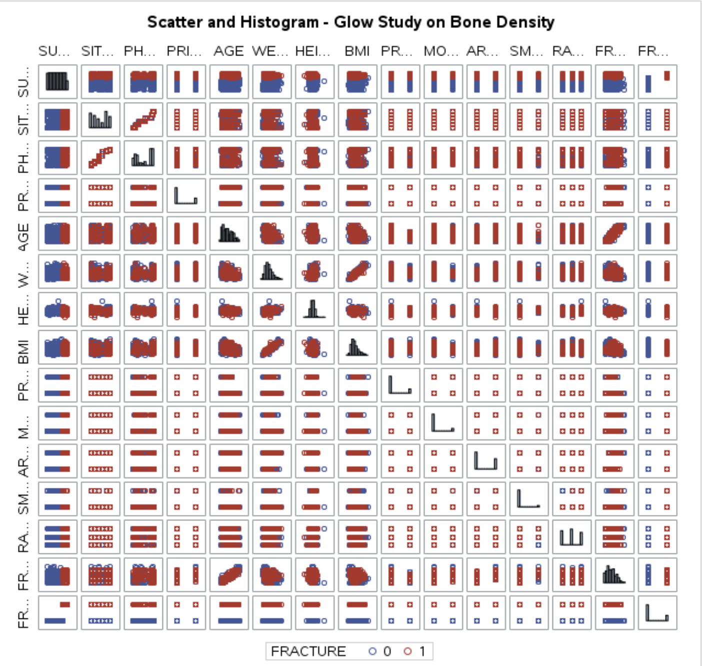
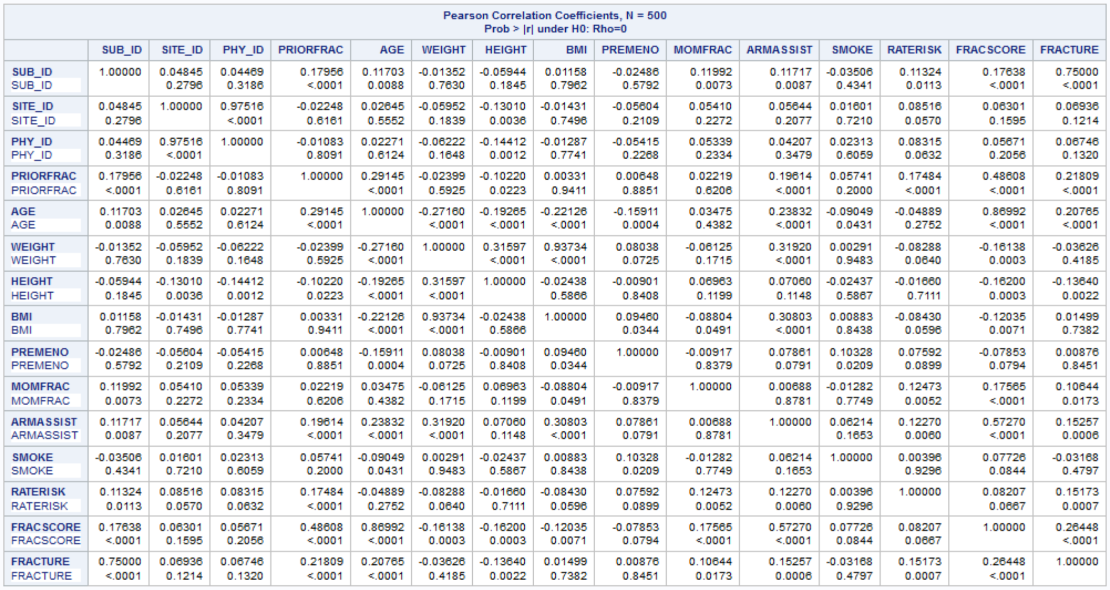


Figure 4 - Scatter and Histogram - Glow Study on Bone Density

The first steps in understanding how to build our model was to do initial survey tests of all the data set variables. After these preliminary steps of EDA were conducted, the analysis then turned to the correlation of independent variables, variance of inflation, and logistic regression. Figure 5 displays that PRIORFRAC, AGE, HEIGHT, MOMFRAC, ARMASSIST, RATERISK, and FRACSCORE have statistically significant correlation to FRACTURE. This started to give us more awareness with the data set and what variables needed to be utilized further for the analysis.

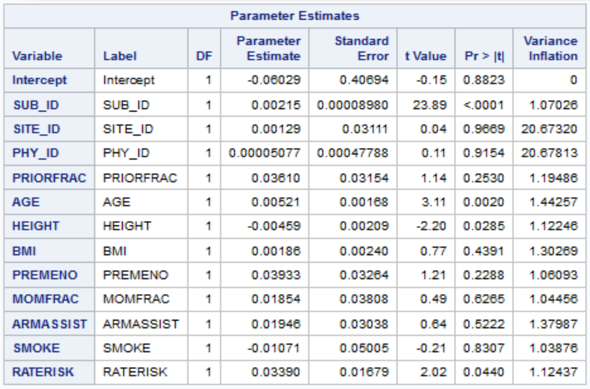


Pearson Correlation Coefficients, N = 500
Prob > |r| under H0: Rho=0

	SUB_ID	SITE_ID	PHY_ID	PRIORFRAC	AGE	WEIGHT	HEIGHT	BMI	PREMENO	MOMFRAC	ARMASSIST	SMOKE	RATERISK	FRACSCORE	FRACTURE
SUB_ID SUB_ID	1.00000	0.04845 0.2796	0.04469 0.3186	0.17956 <.0001	0.11703 0.0088	-0.01352 0.7630	-0.05944 0.1845	0.01158 0.7962	-0.02486 0.5792	0.11992 0.0073	0.11717 0.0087	-0.03506 0.4341	0.11324 0.0113	0.17638 <.0001	0.75000 <.0001
SITE_ID SITE_ID	0.04845 0.2796	1.00000	0.97516 <.0001	-0.02248 0.6161	0.02645 0.5552	-0.05952 0.1839	-0.13010 0.0036	-0.01431 0.7496	-0.05604 0.2109	0.05410 0.2272	0.05644 0.2077	0.01601 0.7210	0.08516 0.0570	0.06301 0.1595	0.06936 0.1214
PHY_ID PHY_ID	0.04469 0.3186	0.97516 <.0001	1.00000	-0.01083 0.8091	0.02271 0.6124	-0.06222 0.1648	-0.14412 0.0012	-0.01287 0.7741	-0.05415 0.2268	0.05339 0.2334	0.04207 0.3479	0.02313 0.6059	0.08315 0.0632	0.05671 0.2056	0.06746 0.1320
PRIORFRAC PRIORFRAC	0.17956 <.0001	-0.02248 0.6161	-0.01083 0.8091	1.00000	0.29145 <.0001	-0.02399 0.5925	-0.10220 0.0223	0.00331 0.9411	0.00648 0.8851	0.02219 0.6206	0.19614 <.0001	0.05741 0.2000	0.17484 <.0001	0.48608 <.0001	0.21809 <.0001
AGE AGE	0.11703 0.0088	0.02645 0.5552	0.02271 0.6124	0.29145 <.0001	1.00000	-0.27160 <.0001	-0.19265 <.0001	-0.22126 <.0001	-0.15911 0.0004	0.03475 0.4382	0.23832 <.0001	-0.09049 0.0431	-0.04889 0.2752	0.86992 <.0001	0.20765 <.0001
WEIGHT WEIGHT	-0.01352 0.7630	-0.05952 0.1839	-0.06222 0.1648	-0.02399 0.5925	-0.27160 <.0001	1.00000	0.31597 <.0001	0.93734 <.0001	0.08038 0.0725	-0.06125 0.1715	0.31920 <.0001	0.00291 0.9483	-0.08288 0.0640	-0.16138 0.0003	-0.03626 0.4185
HEIGHT HEIGHT	-0.05944 0.1845	-0.13010 0.0036	-0.14412 0.0012	-0.10220 0.0223	-0.19265 <.0001	0.31597 <.0001	1.00000	-0.02438 0.5866	-0.00901 0.8408	0.06963 0.1199	0.07060 0.1148	-0.02437 0.5867	-0.01660 0.7111	-0.16200 0.0003	-0.13640 0.0022
BMI BMI	0.01158 0.7962	-0.01431 0.7496	-0.01287 0.7741	0.00331 0.9411	-0.22126 <.0001	0.93734 <.0001	-0.02438 0.5866	1.00000	0.09460 0.0344	-0.08804 0.0491	0.30803 <.0001	0.00883 0.8438	-0.08430 0.0596	-0.12035 0.0071	0.01499 0.7382
PREMENO PREMENO	-0.02486 0.5792	-0.05604 0.2109	-0.05415 0.2268	0.00648 0.8851	-0.15911 0.0004	0.08038 0.0725	-0.00901 0.8408	0.09460 0.0344	1.00000	-0.00917 0.8379	0.07861 0.0791	0.10328 0.0209	0.07592 0.0899	-0.07853 0.0794	0.00876 0.8451
MOMFRAC MOMFRAC	0.11992 0.0073	0.05410 0.2272	0.05339 0.2334	0.02219 0.6206	0.03475 0.4382	-0.06125 0.1715	0.06963 0.1199	-0.08804 0.0491	-0.00917 0.8379	1.00000	0.00688 0.8781	-0.01282 0.7749	0.12473 0.0052	0.17565 <.0001	0.10644 0.0173
ARMASSIST ARMASSIST	0.11717 0.0087	0.05644 0.2077	0.04207 0.3479	0.19614 <.0001	0.23832 <.0001	0.31920 <.0001	0.07060 0.1148	0.30803 <.0001	0.07861 0.0791	0.00688 0.8781	1.00000	0.06214 0.1653	0.12270 0.0060	0.57270 <.0001	0.15257 0.0006
SMOKE SMOKE	-0.03506 0.4341	0.01601 0.7210	0.02313 0.6059	0.05741 0.2000	-0.09049 0.0431	0.00291 0.9483	-0.02437 0.5867	0.00883 0.8438	0.10328 0.0209	-0.01282 0.7749	0.06214 0.1653	1.00000	0.00396 0.9296	0.07726 0.0844	-0.03168 0.4797
RATERISK RATERISK	0.11324 0.0113	0.08516 0.0570	0.08315 0.0632	0.17484 <.0001	-0.04889 0.2752	-0.08288 0.0640	-0.01660 0.7111	-0.08430 0.0596	0.07592 0.0899	0.12473 0.0052	0.12270 0.0060	0.00396 0.9296	1.00000	0.08207 0.0667	0.15173 0.0007
FRACSCORE FRACSCORE	0.17638 <.0001	0.06301 0.1595	0.05671 0.2056	0.48608 <.0001	0.86992 <.0001	-0.16138 0.0003	-0.16200 0.0003	-0.12035 0.0071	-0.07853 0.0794	0.17565 <.0001	0.57270 <.0001	0.07726 0.0844	0.08207 0.0667	1.00000	0.26448 <.0001
FRACTURE FRACTURE	0.75000 <.0001	0.06936 0.1214	0.06746 0.1320	0.21809 <.0001	0.20765 <.0001	-0.03626 0.4185	-0.13640 0.0022	0.01499 0.7382	0.00876 0.8451	0.10644 0.0173	0.15257 0.0006	-0.03168 0.4797	0.15173 0.0007	0.26448 <.0001	1.00000

Figure 5 - Pearson Correlation Analysis 1 - All Variables - Glow Study on Bone Density

The variance in inflation test showed that both SITE_ID and PHY_ID have VIF values of greater than 20 as depicted in Figure 6. Any VIF value over 10 makes a variable a candidate for elimination out of the model. One variable was removed at a time until all VIF values are under 10, starting with WEIGHT and FRACSCORE. We chose SITE_ID as the next variable because while SITE_ID and PHY_ID both give location information, PHY_ID also gives physician information. We removed SITE_ID, and PHY_ID's VIF value went down to 1.04.



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-0.06029	0.40694	-0.15	0.8823	0
SUB_ID	SUB_ID	1	0.00215	0.00008980	23.89	<.0001	1.07026
SITE_ID	SITE_ID	1	0.00129	0.03111	0.04	0.9669	20.67320
PHY_ID	PHY_ID	1	0.00005077	0.00047788	0.11	0.9154	20.67813
PRIORFRAC	PRIORFRAC	1	0.03610	0.03154	1.14	0.2530	1.19486
AGE	AGE	1	0.00521	0.00168	3.11	0.0020	1.44257
HEIGHT	HEIGHT	1	-0.00459	0.00209	-2.20	0.0285	1.12246
BMI	BMI	1	0.00186	0.00240	0.77	0.4391	1.30269
PREMENO	PREMENO	1	0.03933	0.03264	1.21	0.2288	1.06093
MOMFRAC	MOMFRAC	1	0.01854	0.03808	0.49	0.6265	1.04456
ARMASSIST	ARMASSIST	1	0.01946	0.03038	0.64	0.5222	1.37987
SMOKE	SMOKE	1	-0.01071	0.05005	-0.21	0.8307	1.03876
RATERISK	RATERISK	1	0.03390	0.01679	2.02	0.0440	1.12437

Figure 6 - Regression for Variance Inflation and Lack of Fit Analysis 1 - Glow Study on Bone Density

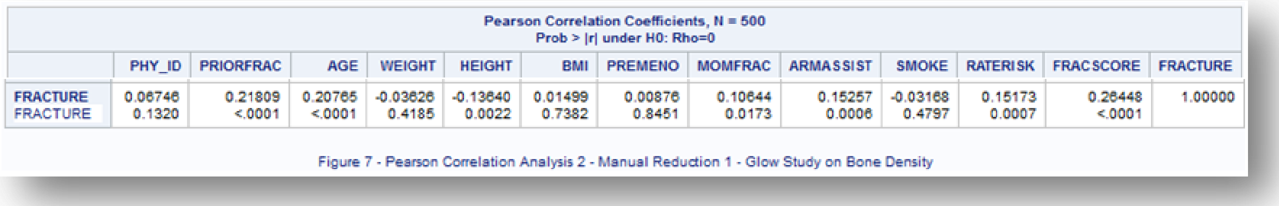
The first run of the logistic regression on all of the variables yielded an interesting result. The statistical software, SAS, reported, "WARNING: There is a complete separation of data points. The maximum likelihood estimate does not exist…Validity of the model fit is questionable." Research of this message indicated that there was perfect predictability between one or more of the variables and the dependent variable, FRACTURE. A survey of the data set revealed that if

SUB_ID <= 375, then FRACTURE = 0. If SUB_ID > 376, then FRACTURE = 1. SUB_ID is really just a subject identifier and does not give necessary medical information.

Logistic Regression Analysis

Manual Reduction

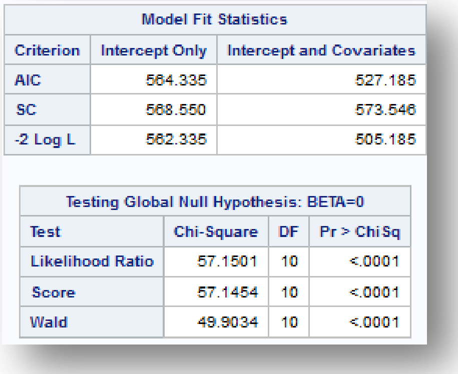
These two tests designated WEIGHT, FRACSCORE, SUB_ID and SITE_ID for removal. We executed the three tests again without these four variables as the first manual reduction and verified a stable status for the reduced model. The results of the correlation and variance in inflation tests are given in Figures 7 and 8 in the appendix with a condensed version of Figure 7 given here. Now six variables correlated to FRACTURE as statistically significant. All VIF values ranged between 1 and 2.



Pearson Correlation Coefficients, N = 500
Prob > |r| under H0: Rho=0

	PHY_ID	PRIORFRAC	AGE	WEIGHT	HEIGHT	BMI	PREMENO	MOMFRAC	ARMASSIST	SMOKE	RATERISK	FRACSCORE	FRACTURE
FRACTURE FRACTURE	0.06746 0.1320	0.21809 <.0001	0.20765 <.0001	-0.03626 0.4185	-0.13640 0.0022	0.01499 0.7382	0.00876 0.8451	0.10644 0.0173	0.15257 0.0006	-0.03168 0.4797	0.15173 0.0007	0.26448 <.0001	1.00000

Figure 7 - Pearson Correlation Analysis 2 - Manual Reduction 1 - Glow Study on Bone Density



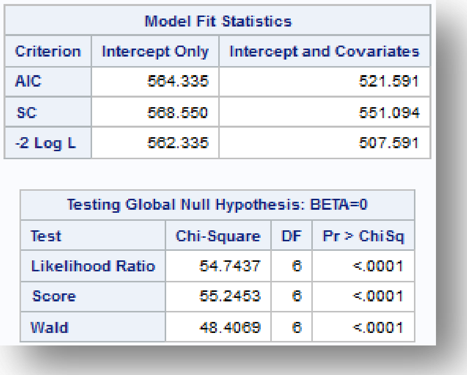
Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	564.335	527.185
SC	568.550	573.546
-2 Log L	562.335	505.185

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	57.1501	10	<.0001
Score	57.1454	10	<.0001
Wald	49.9034	10	<.0001

Figure 9 - Logistic Regression 2 - Manual Reduction 1 - Glow Study on Bone Density



Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	564.335	521.591
SC	568.550	551.094
-2 Log L	562.335	507.591

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	54.7437	6	<.0001
Score	55.2453	6	<.0001
Wald	48.4069	6	<.0001

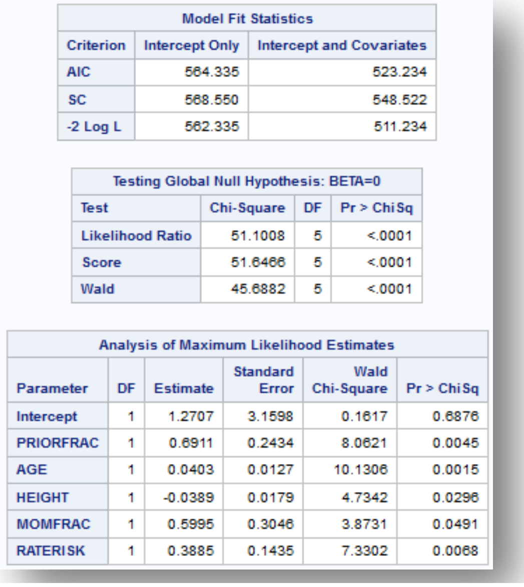
Figure 10 – Logistic Regression 3 - Manual Reduction 2 - Six Interest Variables - Glow Study on Bone Density

The logistic regression test reported, “Convergence criterion (GCONV=1E-8) satisfied.” This message means the perfect predictability was removed with SUB_ID. The logistic regression values, including a Likelihood Ratio = 57.15 with DF=10 and p-value (Pr <ChiSq) <.0001 as shown here in the Model Fit Statistics and Testing Global Null Hypothesis: BETA=0 tables of Figure 9, became the baseline logistic regression values (i.e. the new full model) for comparison with other manual and automatic regression selection methods during subsequent iterations. Additionally, we noticed the AIC and -2 Log L values for the Intercept and Covariates are smaller than for the Intercept alone, so we conducted additional iterations in search of the best model.

Because the same six variables, PRIORFRAC, AGE, HEIGHT, MOMFRAC, ARMASSIST, and RATERISK remained statistically significantly correlated to FRACTURE, we selected these variables to be the second manual reduction model candidate. The Model Fit Statistics and Testing Global Null Hypothesis: BETA=0 values, as depicted in tables of Figure 10, showed improvement in the smaller AIC(521.59) and SC(551.09) values in the Intercept and Covariates column. Correspondingly, the Likelihood Ratio became 54.74 with DF=6 and a p-value (Pr <ChiSq) <.0001.

Automatic Selection Reduction

In further effort to reduce the number of variables, we employed the automatic selection methods of stepwise (Figure 11), forward (Figure 12 in appendix), and backward (Figure 13 in appendix) in Proc Logistic. We specified the same 10 variables in the selection as in the baseline or "new full model" of Manual Reduction 1 to see what model Proc Logistic offered. All three automatic selection methods reported the exact same results in the Model Fit Statistics and Testing Global Null Hypothesis: BETA=0 tables. The automatic selections improved over the manual iterations with smaller SC (548.52) and Likelihood Ratio (51.10) with DF 5 and a p-value (Pr <ChiSq) <.0001.



Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	564.335	523.234
SC	568.550	548.522
-2 Log L	562.335	511.234

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	51.1008	5	<.0001
Score	51.6466	5	<.0001
Wald	45.6882	5	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.2707	3.1598	0.1617	0.6876
PRIORFRAC	1	0.6911	0.2434	8.0621	0.0045
AGE	1	0.0403	0.0127	10.1306	0.0015
HEIGHT	1	-0.0389	0.0179	4.7342	0.0296
MOMFRAC	1	0.5995	0.3046	3.8731	0.0491
RATERISK	1	0.3885	0.1435	7.3302	0.0068

Figure 81 - Logistic Regression 4 - Stepwise Automatic Selection - FRACTURE = PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE RATERISK

Final Model Generation and Context

During the model generation process, we iteratively removed variables while carefully comparing that the smaller subset model's test values improved over the full model as well as the intercept only model. The comparison was based on AIC, SC, -2 Log L, and Likelihood Ratio values. Figure 14 gives a summary of the process, which was a "Drop in Deviance" progression that focused on the Likelihood Ratio Test.

The general form of LRT statistic is: Likelihood Ratio or LRT = $2(LL_{\text{full model}} – LL_{\text{reduced model}})$

Chi-Square for Likelihood Ratio Test	Number of variables	DF	Pr>ChiSq
57.15	10(new full model)	10	<.0001
54.74	6(subset)	6	<.0001
51.10	5(subset)	5	<.0001

Figure 14 – Logistic Regression LRT Comparison

SAS generates LRT automatically using a different algebraic form.

Likelihood Ratio Test (Chi-Square) value

= -2Log L value for intercept only - (-2Log L for intercept and covariates value)

where the deviance, $\Delta G^2 = G^2$ for smaller model − G^2 for larger model or

$\Delta\ X^2 = X^2$ for smaller model − X^2 for larger model

LRT has approximate χ^2 distribution with degrees of freedom equal to difference between numbers of parameters in full and reduced models, under null hypothesis assuming reduced model is correct. We apply the Hypothesis Test of

H_0 : smaller model (intercept only) is true versus H_1 : larger model is true

Because the p-value (Pr <ChiSq) was < .0001 for the Likelihood Ratio Test, we reject the hypothesis that the H_0 : smaller model (intercept only) is true in favor of the H_1 : larger model is true.

Because the p-values (Pr <ChiSq) of all intercept and covariate models were consistently <.0001, we determined to accept the most parsimonious model with the lowest LRT value = 51.10 as denoted in Figure 11, Analysis of Maximum Likelihood Estimate table. (Showing good corroborating support with values < .0001, are the p-values (Pr <ChiSq) for the Score and Wald test.) This model has PRIORFRAC=X_1, AGE=X_2, HEIGHT=X_3, MOMFRAC=X_4, and RATERISK=$X_{5.}$ The logit model equation in regards to the dependent variable FRACTURE is as follows:

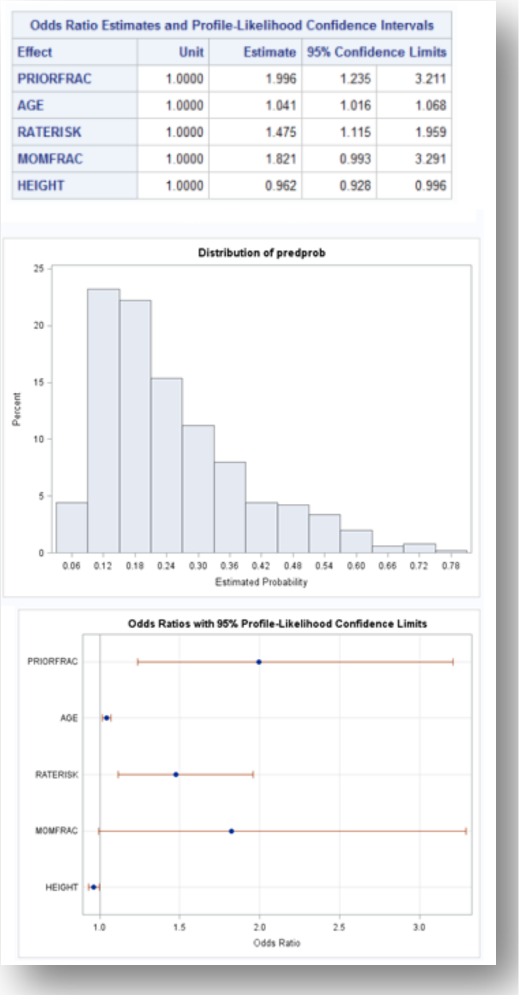
$$\text{logit}(\pi) = \ln(\pi/1-\pi)=\beta_0+\beta_1X_1+\beta_2X2-\beta_3X_3+\beta_4X_4+\beta_5X_5$$

$$\text{logit}(\pi) = \ln(\pi/1-\pi)=1.2707+0.6911X_1+0.0403X_2-0.389X_3+0.5995X_4+0.3885X_5.$$

The women of the study provided important information about the fracture risk according to this logit model. Figure 15 shows that under the Profile-Likelihood and Odds Ratio Estimates, the PRIORFRAC variable indicates the odds ratio increase of an estimated 2 to 1 of another fracture and a 95% confidence interval that ranges from 1.235 to 3.211 if the categorical variable registers. This is an obvious risk identified within women and translates into more awareness gained of fractures in the demographic. The RATERISK odds ratio measuring an estimated 1.475 to 1 with a smaller range within its confidence limits which remains above 1 in its lower limit of 1.115. However, the MOMFRAC variable, even though indicating a larger estimated odds ratio, possesses 1 within its very wide 95% confidence limits. This indicates that even odds can exist within the variable but also increase up to 3.291 to 1 odds in its upper limit.

The AGE and HEIGHT variables encompass tighter odds and closer to 1. The HEIGHT variable is peculiar given that it's the only variable that lowers your odds of fracture and also produces a -.0389 coefficient within the Logistic Regression.

The women within this demographic begin to fall in the following distribution of probabilities. In respect to the histogram, you can see how the predicted estimated probabilities undergo a positive skew of 1.15, mean of 25%, median of 21%, and mode of 9.4% within the model. So, most women have a 9.4% of fracture but on average will carry a 25% probability. The impactful variables of RATERISK, PRIORFRAC, and MOMFRAC can be key indicators of increasing a woman's odds and probability for fracture in the future.



Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
PRIORFRAC	1.0000	1.996	1.235	3.211
AGE	1.0000	1.041	1.016	1.068
RATERISK	1.0000	1.475	1.115	1.959
MOMFRAC	1.0000	1.821	0.993	3.291
HEIGHT	1.0000	0.962	0.928	0.996

Figure 15 - Odds Ratios, Confidence Limits, and Histogram of Estimated Probability

Conclusion

The quest for improvement in quality of life in later years will continue to be an uphill battle. The greater awareness of the substantial risks that a person may fall under can be mitigated by acting accordingly. Discovery of these risks aids in reducing the endangerment of diseases like osteoporosis and can point us in the right direction as we navigate further in our increasing life expectancies. This boost in knowledge can then begin to grow and gain momentum on women and osteoporosis. As we continue with this quest, other factors might need to be further explored such as diet and nutrition during these phases of our lives. These predictive models and data wrangling can start to show the fruits of the labor from the data collection effort conducted with this GLOW study.

Programs, institutions, and non-profits can begin to confidently interact with the general public and start to sway factors in our favor to produce more productive and enjoyable lives. The resulting product from these efforts can increase a larger appetite to further studies and gain priority funding. Physicians within the field can continue or alter treatment strategies to help them provide strong diagnosis and effective therapies. Effective change can range from fueling vitamin D supplementations to altering national policies. These changes give hope that osteoporosis and musculoskeletal disease will no longer burden an already struggling healthcare system and families needing to care for their loved ones.

Bibliography

University of Massachusetts Medical School. (2010). Retrieved from https://www.outcomes-umassmed.org/GLOW/publicfiles/GLOW_Overview.ppt

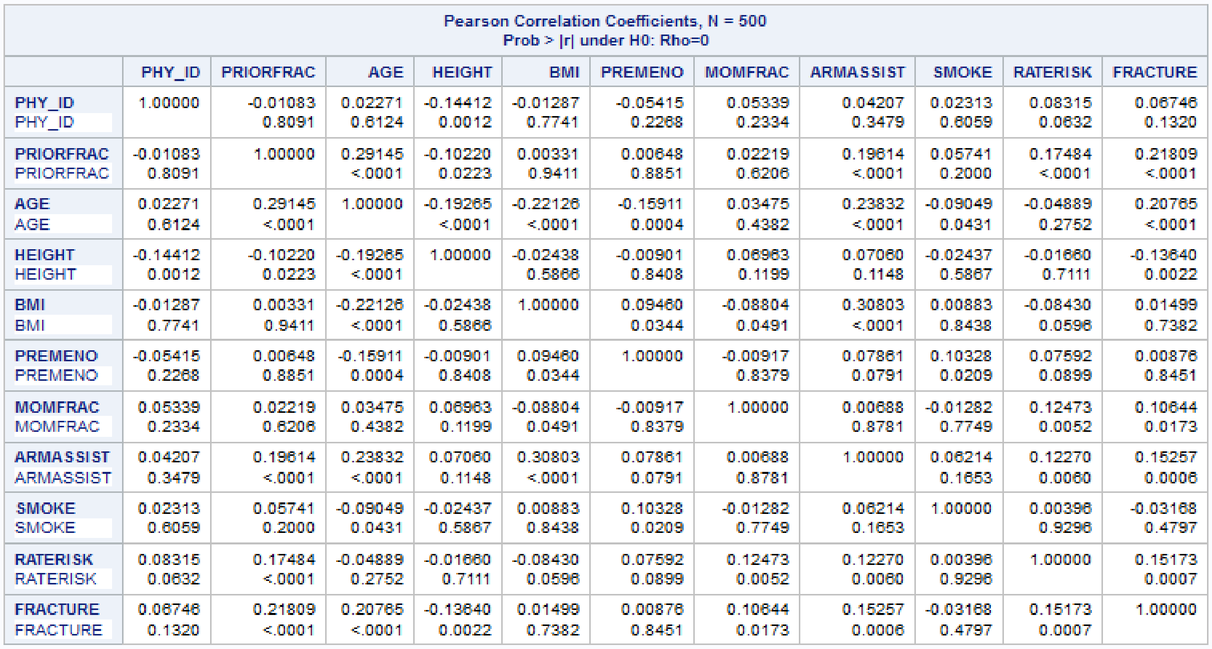
Melville, N. A. (2011, September 20). *Obesity a Risk Factor for Fractures in Postmenopausal Women.* Retrieved from Docguide.com: http://www.docguide.com/obesity-risk-factor-fractures-postmenopausal-women?hash=8bbf84a9&eid=22233&alrhash=23da12-64067ca5d68ac0733415244ac15db20b

University of Massachusetts Amherst. (2004). *Software - Statistical Consulting Center - UMass Amherst.* Retrieved from Statistical Software Information: https://www.umass.edu/statdata/statdata/data/glow/glow.pdf; https://www.umass.edu/statdata/statdata/data/glow;

References

https://onlinecourses.science.psu.edu/stat504/node/220
http://www2.sas.com/proceedings/forum2008/360-2008.pdf

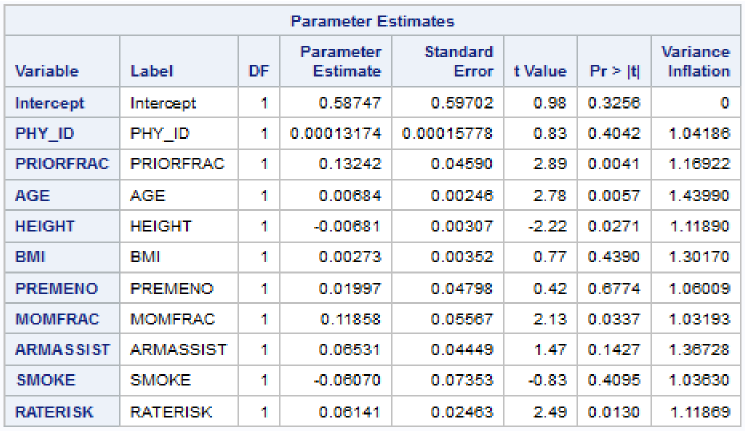
Table of Figures



Pearson Correlation Coefficients, N = 500
Prob > |r| under H0: Rho=0

	PHY_ID	PRIORFRAC	AGE	HEIGHT	BMI	PREMENO	MOMFRAC	ARMASSIST	SMOKE	RATERISK	FRACTURE
PHY_ID PHY_ID	1.00000	-0.01083 0.8091	0.02271 0.6124	-0.14412 0.0012	-0.01287 0.7741	-0.05415 0.2268	0.05339 0.2334	0.04207 0.3479	0.02313 0.6059	0.08315 0.0632	0.06746 0.1320
PRIORFRAC PRIORFRAC	-0.01083 0.8091	1.00000	0.29145 <.0001	-0.10220 0.0223	0.00331 0.9411	0.00648 0.8851	0.02219 0.6206	0.19614 <.0001	0.05741 0.2000	0.17484 <.0001	0.21809 <.0001
AGE AGE	0.02271 0.6124	0.29145 <.0001	1.00000	-0.19265 <.0001	-0.22126 <.0001	-0.15911 0.0004	0.03475 0.4382	0.23832 <.0001	-0.09049 0.0431	-0.04889 0.2752	0.20765 <.0001
HEIGHT HEIGHT	-0.14412 0.0012	-0.10220 0.0223	-0.19265 <.0001	1.00000	-0.02438 0.5866	-0.00901 0.8408	0.06963 0.1199	0.07060 0.1148	-0.02437 0.5867	-0.01660 0.7111	-0.13640 0.0022
BMI BMI	-0.01287 0.7741	0.00331 0.9411	-0.22126 <.0001	-0.02438 0.5866	1.00000	0.09460 0.0344	-0.08804 0.0491	0.30803 <.0001	0.00883 0.8438	-0.08430 0.0596	0.01499 0.7382
PREMENO PREMENO	-0.05415 0.2268	0.00648 0.8851	-0.15911 0.0004	-0.00901 0.8408	0.09460 0.0344	1.00000	-0.00917 0.8379	0.07861 0.0791	0.10328 0.0209	0.07592 0.0899	0.00876 0.8451
MOMFRAC MOMFRAC	0.05339 0.2334	0.02219 0.6206	0.03475 0.4382	0.06963 0.1199	-0.08804 0.0491	-0.00917 0.8379	1.00000	0.00688 0.8781	-0.01282 0.7749	0.12473 0.0052	0.10644 0.0173
ARMASSIST ARMASSIST	0.04207 0.3479	0.19614 <.0001	0.23832 <.0001	0.07060 0.1148	0.30803 <.0001	0.07861 0.0791	0.00688 0.8781	1.00000	0.06214 0.1653	0.12270 0.0060	0.15257 0.0006
SMOKE SMOKE	0.02313 0.6059	0.05741 0.2000	-0.09049 0.0431	-0.02437 0.5867	0.00883 0.8438	0.10328 0.0209	-0.01282 0.7749	0.06214 0.1653	1.00000	0.00396 0.9296	-0.03168 0.4797
RATERISK RATERISK	0.08315 0.0632	0.17484 <.0001	-0.04889 0.2752	-0.01660 0.7111	-0.08430 0.0596	0.07592 0.0899	0.12473 0.0052	0.12270 0.0060	0.00396 0.9296	1.00000	0.15173 0.0007
FRACTURE FRACTURE	0.06746 0.1320	0.21809 <.0001	0.20765 <.0001	-0.13640 0.0022	0.01499 0.7382	0.00876 0.8451	0.10644 0.0173	0.15257 0.0006	-0.03168 0.4797	0.15173 0.0007	1.00000

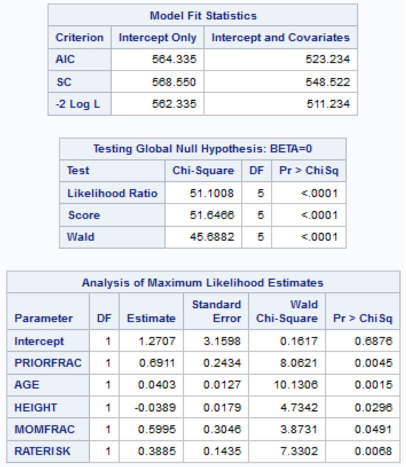
Figure 7 - Pearson Correlation Analysis 2 - Manual Reduction 1 - Glow Study on Bone Density



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	0.58747	0.59702	0.98	0.3256	0
PHY_ID	PHY_ID	1	0.00013174	0.00015778	0.83	0.4042	1.04186
PRIORFRAC	PRIORFRAC	1	0.13242	0.04590	2.89	0.0041	1.16922
AGE	AGE	1	0.00684	0.00246	2.78	0.0057	1.43990
HEIGHT	HEIGHT	1	-0.00681	0.00307	-2.22	0.0271	1.11890
BMI	BMI	1	0.00273	0.00352	0.77	0.4390	1.30170
PREMENO	PREMENO	1	0.01997	0.04798	0.42	0.6774	1.06009
MOMFRAC	MOMFRAC	1	0.11858	0.05567	2.13	0.0337	1.03193
ARMASSIST	ARMASSIST	1	0.06531	0.04449	1.47	0.1427	1.36728
SMOKE	SMOKE	1	-0.06070	0.07353	-0.83	0.4095	1.03630
RATERISK	RATERISK	1	0.06141	0.02463	2.49	0.0130	1.11869

Figure 8 - Regression for Variance Inflation and Lack of Fit Analysis 2 - Manual Reduction - Glow Study on Bone Density

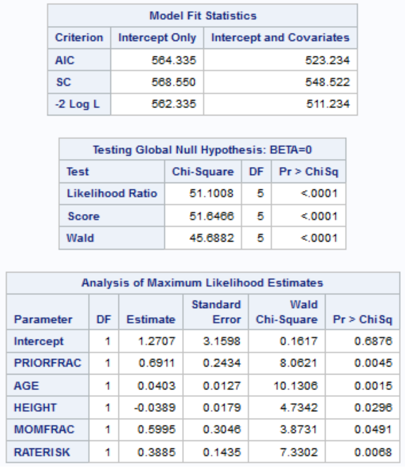


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	564.335	523.234
SC	568.550	548.522
-2 Log L	562.335	511.234

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	51.1008	5	<.0001
Score	51.6466	5	<.0001
Wald	45.6882	5	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.2707	3.1598	0.1617	0.6876
PRIORFRAC	1	0.6911	0.2434	8.0621	0.0045
AGE	1	0.0403	0.0127	10.1306	0.0015
HEIGHT	1	-0.0389	0.0179	4.7342	0.0296
MOMFRAC	1	0.5995	0.3046	3.8731	0.0491
RATERISK	1	0.3885	0.1435	7.3302	0.0068

Figure 11 - Logistic Regression 4 - Stepwise Automatic Selection - FRACTURE = PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE RATERISK

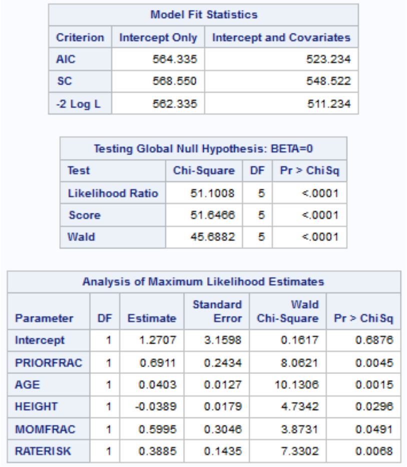


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	564.335	523.234
SC	568.550	548.522
-2 Log L	562.335	511.234

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	51.1008	5	<.0001
Score	51.6466	5	<.0001
Wald	45.6882	5	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.2707	3.1598	0.1617	0.6876
PRIORFRAC	1	0.6911	0.2434	8.0621	0.0045
AGE	1	0.0403	0.0127	10.1306	0.0015
HEIGHT	1	-0.0389	0.0179	4.7342	0.0296
MOMFRAC	1	0.5995	0.3046	3.8731	0.0491
RATERISK	1	0.3885	0.1435	7.3302	0.0068

Figure 12 - Logistic Regression 5 - Forward Automatic Selection - FRACTURE = PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE RATERISK



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	564.335	523.234
SC	568.550	548.522
-2 Log L	562.335	511.234

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	51.1008	5	<.0001
Score	51.6466	5	<.0001
Wald	45.6882	5	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.2707	3.1598	0.1617	0.6876
PRIORFRAC	1	0.6911	0.2434	8.0621	0.0045
AGE	1	0.0403	0.0127	10.1306	0.0015
HEIGHT	1	-0.0389	0.0179	4.7342	0.0296
MOMFRAC	1	0.5995	0.3046	3.8731	0.0491
RATERISK	1	0.3885	0.1435	7.3302	0.0068

Figure 13 - Logistic Regression 6 - Backward Automatic Selection - FRACTURE = PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE RATERISK

Appendix

```
PROC IMPORT OUT= bones
/*
Citrix access
DATAFILE= '\\Client\C$\Users\Celia
Taylor\Documents\SASUniversityEdition\myfoldersSASDATA\glow500.xls'
*/
          DATAFILE= '/folders/myfolders/SASDATA/glow500.xls' DBMS=XLS REPLACE;
    SHEET="GLOW500.TAB.XLS";
    GETNAMES=YES;

run;
data bones;
set bones;
numObs = _N_;
run;
title "Descriptive Statistics - Glow Study on Bone Density";
proc means data=bones;
footnote "Figure 3 - Descriptive Statistics - Glow Study on Bone Density";
run;
title "Scatter and Histogram - Glow Study on Bone Density";
proc sgscatter data=bones;
matrix SUB_ID SITE_ID PHY_ID PRIORFRAC AGE WEIGHT HEIGHT BMI PREMENO
MOMFRAC
    ARMASSIST SMOKE RATERISK FRACSCORE FRACTURE / diagonal=(histogram)
group=FRACTURE;
footnote "Figure 4 - Scatter and Histogram - Glow Study on Bone Density";
run;
/*
We conduct initial survey first by executing proc corr, proc reg, proc logistic.
```

```
*/
title "Pearson Correlation Analysis 1 - Glow Study on Bone Density";
title "Univariate of Bone Density Variables";
proc univariate data=bones plots normal;
var SUB_ID SITE_ID PHY_ID PRIORFRAC AGE WEIGHT HEIGHT BMI PREMENO
MOMFRAC
      ARMASSIST SMOKE RATERISK FRACSCORE FRACTURE;
footnote "Observation 0 - Univariate of Bone Density Variables";
run;

proc corr PEARSON data=bones;
var SUB_ID SITE_ID PHY_ID PRIORFRAC AGE WEIGHT HEIGHT BMI PREMENO
MOMFRAC
    ARMASSIST SMOKE RATERISK FRACSCORE FRACTURE;
footnote "Figure 5 - Pearson Correlation Analysis 1 - All Variables - Glow Study on Bone
Density";
run;
/*
Proc corr shows as statistically significant the following variables:
PRIORFRAC           0.21809
                              <.0001
AGE                           0.20765
                              <.0001
HEIGHT                        -0.13640
                              0.0022
MOMFRAC                       0.10644
                              0.0173
ARMASSIST           0.15257
                              0.0006
RATERISK            0.15173
                              0.0007
FRACSCORE           0.26448
                              <.0001 .
These seven variables will be of the most interest while running the remaining procedures.

*/
title "Regression for Variance Inflation and Lack of Fit Analysis 1 - Glow Study on Bone
Density";
proc reg data=bones;
model FRACTURE = SUB_ID SITE_ID PHY_ID PRIORFRAC AGE WEIGHT HEIGHT BMI
PREMENO MOMFRAC ARMASSIST SMOKE RATERISK FRACSCORE/ lackfit VIF;
/*
model FRACTURE = SUB_ID SITE_ID PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO
MOMFRAC ARMASSIST SMOKE RATERISK FRACSCORE/ lackfit VIF;
model FRACTURE = SUB_ID SITE_ID PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO
MOMFRAC ARMASSIST SMOKE RATERISK/ lackfit VIF;
model FRACTURE = SUB_ID PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO MOMFRAC
ARMASSIST SMOKE RATERISK/ lackfit VIF;
*/
footnote "Figure 6 - Regression for Variance Inflation and Lack of Fit Analysis 1 - Glow Study on
Bone Density";
```

```
run;
/*
Both SITE_ID and PHY_ID hav VIF > 20.
SITE_ID 1	0.01811	0.04578	0.40	0.6926	20.66261
PHY_ID	1	-0.00013943	0.00070342	-0.20	0.8430	20.67239
BMI	1	0.02140	0.02558	0.84	0.4033	148.08801
WEIGHT	WEIGHT	1	-0.00739	0.00972	-0.76	0.4475
	161.95196
FRACSCORE	FRACSCORE	1	0.01083	0.03205	0.34	0.7357
	40.55584
AGE	AGE	1	0.00310	0.00649	0.48	0.6333	21.54780
We remove one variable with a VIF score above 10 at a time until all VIFs are below 10.  We
removed WEIGHT, FRACSCORE, SUB_ID, SITE_ID
If we remove either SITE_ID or PHY_ID,
then we get the same result.  The remaining variables have a VIF that is close to one.  Both give
information on location,
but PHY_ID gives information on location (including medical facility through association) and
physician.
We removed SITE_ID, and PHY_ID VIF went down to 1.04186.  Neither SITE_ID or PHY_ID
were in proc corr's
seven statistically significant variables.  FRACSCORE unfortunately was.  So now there are six
interest variables.
*/
title "Logistic Regression 1 - All Variables- Glow Study on Bone Density";
proc logistic data=bones outest= fracAll1;
model FRACTURE (event='1') =  SUB_ID SITE_ID PHY_ID PRIORFRAC AGE WEIGHT
HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE RATERISK FRACSCORE /
risklimits lackfit ctable clodds=both;
output out = bonesOut1 predprobs=I p=predprob resdev=resdev reschi=pearres;
footnote "Observation 1 - Logistic Regression 1 - All Variables- Glow Study on Bone Density";
run;
/*
If we run PROC Logistic with all of the variables, we get the following message in the log file:
 "NOTE: PROC LOGISTIC is modeling the probability that FRACTURE='1'.
  WARNING: There is a complete separation of data points. The maximum likelihood estimate
does not exist.
  WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown
are based on the last maximum likelihood
         iteration. Validity of the model fit is questionable."
If we remove SUB_ID, then the message goes away because there is perfect predictability
between SUB_ID and FRACTURE.  If SUB_ID <= 375, then FRACTURE = 0.  IF SUB_ID >
376, then FRACTURE = 1.
After removing SUB_ID, the log file reports,
 "NOTE: PROC LOGISTIC is modeling the probability that FRACTURE='1'.
  NOTE: Convergence criterion (GCONV=1E-8) satisfied.
  NOTE: There were 500 observations read from the data set WORK.BONES."
SUB_ID was not in proc corr's seven statistically significant variables.
*/
/*
Then, we reduce variables manually according to the initial survey results from proc corr, proc
reg, proc logistic.
```

```
We removed WEIGHT, FRACSCORE, SUB_ID and SITE_ID.
*/
title "Pearson Correlation Analysis 2 - Manual Reduction 1 - Glow Study on Bone Density";
proc corr PEARSON data=bones;
var PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO MOMFRAC
    ARMASSIST SMOKE RATERISK FRACTURE;
footnote "Figure 7 - Pearson Correlation Analysis 2 - Manual Reduction 1 - Glow Study on Bone
Density";
run;
title "Regression for Variance Inflation and Lack of Fit Analysis 2 - Manual Reduction - Glow
Study on Bone Density";
proc reg data=bones;
model FRACTURE = PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO MOMFRAC
ARMASSIST SMOKE RATERISK/ lackfit VIF;
footnote "Figure 8 - Regression for Variance Inflation and Lack of Fit Analysis 2 - Manual
Reduction - Glow Study on Bone Density";
run;
title "Logistic Regression 2 - Manual Reduction 1  - Glow Study on Bone Density";
proc logistic data=bones outest= fracMan1;
model FRACTURE (event='1') =  PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO
MOMFRAC
    ARMASSIST SMOKE RATERISK / risklimits lackfit ctable clodds=both;
output out = bonesOut2 predprobs=I p=predprob resdev=resdev reschi=pearres;
footnote "Figure 9 - Logistic Regression 2 - Manual Reduction 1 - Glow Study on Bone Density";
run;
proc print data=bonesOut2; run;
/*Look at residuals to see if anything is out of the norm for any high leverage points*/
/*
title "GPlot of output from LR 2 vs Observations - Manual Reduction 1 - Glow Study on Bone
Density";
proc gplot data=bonesOut2;
plot resdev * numObs;
plot pearres * numObs;
plot predprob * numObs;
footnote "Observation 2 - GPlot of output from LR2 vs Observations - Manual Reduction 1 -
Glow Study on Bone Density";
run;
quit;
*/
title "Logistic Regression 3 - Manual Reduction 2 - Six Interest Variables - Glow Study on Bone
Density";
proc logistic data=bones outest= fracMan2;
model FRACTURE (event='1') =  PRIORFRAC AGE HEIGHT MOMFRAC ARMASSIST
RATERISK/
risklimits lackfit ctable clodds=both;
output out = bonesOut3 predprobs=I p=predprob resdev=resdev reschi=pearres;
footnote "Figure 10 - Logistic Regression 3 - Manual Reduction 2 - Six Interest Variables - Glow
Study on Bone Density";
run;
proc print data=bonesOut3; run;
/*Look at residuals to see if anything is out of the norm for any high leverage points*/
```

```
/*
title "GPlot 1 of output from LR vs Observations - Manual Reduction 2 - Six Interest Variables - Glow Study on Bone Density";
proc gplot data=bonesOut3;
plot resdev * numObs;
plot pearres * numObs;
plot predprob * numObs;
footnote "Observation 3 - GPlot of output from LR vs Observations - Manual Reduction 2 - Six Interest Variables - Glow Study on Bone Density";
run;
quit;
*/
/*Checking automatic selection method*/
title "Logistic Regression 4 - Stepwise Automatic Selection 1 - FRACTURE = PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE RATERISK";
proc logistic data=bones outest= fracAuto1;
model FRACTURE (event='1') =  PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE RATERISK /
selection=stepwise sle=.05 sls=.05 corrb cl details;
output out = bonesOut4 predprobs=I p=predprob resdev=resdev reschi=pearres;
footnote "Figure 11 - Logistic Regression 4 - Stepwise Automatic Selection - FRACTURE = PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE RATERISK";
run;
/*
title "GPlot 2 of output from LR4 vs Observations - Stepwise Automatic Selection - Glow Study on Bone Density";
proc gplot data=bonesOut4;
plot resdev * numObs;
plot pearres * numObs;
plot predprob * numObs;
footnote "Observation 4 - GPlot 2 of output from LR4 vs Observations - Stepwise Automatic Selection - Glow Study on Bone Density";
run;
QUIT;
*/
title "Logistic Regression 5 - Forward Automatic Selection - FRACTURE = PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE RATERISK";
proc logistic data=bones outest= fracAuto2;
model FRACTURE (event='1') =  PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE RATERISK /
selection=forward sle=.05 sls=.05 corrb cl details;
output out = bonesOut5 predprobs=I p=predprob resdev=resdev reschi=pearres;
footnote "Figure 12 - Logistic Regression 5 - Forward Automatic Selection - FRACTURE = PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE RATERISK";
run;
/*
title "GPlot 3 of output from LR5 vs Observations - Forward Automatic Selection - Glow Study on Bone Density";
proc gplot data=bonesOut5;
```

```
plot resdev * numObs;
plot pearres * numObs;
plot predprob * numObs;
footnote "Observation 5 - GPlot 3 of output from LR5 vs Observations - Forward Automatic
Selection - Glow Study on Bone Density";
run;
QUIT;
*/
title "Logistic Regression 6 - Backward Automatic Selection - FRACTURE = PHY_ID
PRIORFRAC AGE HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE RATERISK";
proc logistic data=bones outest= fracAuto3;
model FRACTURE (event='1') =  PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO
MOMFRAC ARMASSIST SMOKE RATERISK /
selection=backward sle=.05 sls=.05 corrb cl details;
output out = bonesOut6 predprobs=I p=predprob resdev=resdev reschi=pearres;
footnote "Figure 13 - Logistic Regression 6 - Backward Automatic Selection - FRACTURE =
PHY_ID PRIORFRAC AGE HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE
RATERISK";
run;
/*
title "GPlot 4 of output from LR6 vs Observations - Backward Automatic Selection - Glow Study
on Bone Density";
proc gplot data=bonesOut6;
plot resdev * numObs;
plot pearres * numObs;
plot predprob * numObs;
footnote "Observation 6 - GPlot 4 of output from LR6 vs Observations - Backward Automatic
Selection - Glow Study on Bone Density";
run;
QUIT;
*/
/*
Intercept only
*/
proc logistic data=bones;
model FRACTURE (event='1') = / scale=none;
run;

data bones;
set bones;
numObs = _N_;
run;
/*
and then ran univariate for the probabilities histogram
title 'GPlot of output from Logistic Regression 2 vs Observations';
proc gplot data=bonesOut4;
plot resdev * numObs;
plot pearres * numObs;
plot predprob * numObs;
footnote 'Figure 15';
run;
```

```
proc univariate data=bonesOut4;
var predprob;
histogram predprob;
run;
*/
```