SAS Analytics Guide:

How to perform logistic regression

Data version: Registered Tier *All of Us* Curated Data Repository (CDR) v7 (R2022Q4R9)
Analysis tool: SAS Studio
Authors:
Bassent Abdelbary, Ph.D., MPH (✉)
University of Texas Rio Grande Valley College of Health Professions

Please note: This guide aims to demonstrate the utilization of **PROC LOGISTIC** using SAS Studio. However, it is important to note that this guide is not comprehensive and does not encompass all facets of the scientific process which researchers are required to undertake. Specifically, it only does not delve into data cleaning and verification, assumption validation, model diagnostics, potential follow-up analyses, or any other possible approaches for performing a logistic regression with forward selection.

Table of Contents

- Introduction
- Description of the dataset
- Statistical analysis procedures
- Additional resources

Introduction

This guide includes examples of statistical analysis processes which were used as part of an exploratory study designed to understand differences in prescriptions of newer generation diabetes medications such as glucagon-like peptide 1 (GLP-1) agonists and sodium-glucose transport protein 2 (SGLT2) inhibitors by looking at patterns within the electronic health records (EHR) and survey data from the *All of Us* Research Program.

For this project, a cohort from the *All of Us* dataset was selected using all participants aged 18 to 75 years old with documented cases of type 2 diabetes in their EHR. Demographic data were combined with prescription data from EHR and select survey questions regarding health insurance type and trends in prescription refills.

- ***Outcome variable:*** Prescription of newer generation diabetes medications such as GLP-1 agonists and SGLT-2 inhibitors (binary variable Yes/No).
- ***Independent variable:*** Several independent variables were evaluated as a part of a stepwise method. For this demonstration guide, we will use health insurance (Yes/No) as our independent variable.
- ***Confounding variable:*** For this guide, we will use race and gender as the confounding variable.

In this analysis, **PROC LOGISTIC** models the probability of no GLP-1/SGLT-2 initial prescription based on chronology of prescription history. The Wald test is used to display the chi^2 and p-value when using **PROC LOGISTIC**. Parameter estimates and odds ratio could be displayed as well with the additional options listed below.

1. The **CLASS** statement names the classification variables to be used as explanatory variables in the analysis.
2. The **EXPB** option displays the Exp (Est) column containing the exponentiated parameter estimates.
3. The **SELECTION=FORWARD** option is specified to carry out the forward selection.
4. For the ods graphics, The **ODDSRATIO** statements compute the odds ratios for the covariates, and the **NOOR** option suppresses the default odds ratio table. **CONTRAST** statements provide another method of producing the odds ratios.

Description of the dataset

For this example, we will use the following criteria to define the cohort for this analysis.

1. Created cohort as defined by the following measures.
    - Inclusion criteria:
        i. Completed The Basics survey including questions about demographics
        ii. Adults (Participants aged 18-75 years old)
        iii. Had the following conditions and lab measurements:
            1. Type 2 diabetes mellitus (diabetes mellitus without complications or type 2 diabetes mellitus without complications)
            2. Lab measurements of an A1C >8
    - Exclusion criteria:
        i. Had the following conditions:
            1. Type 1 diabetes mellitus
            2. Type 1 diabetes mellitus without complications
2. Created a concept set by medication name to include all diabetic medications using medication concept ID. These medications were reconciled and combined into a new variable defining medication class based on mechanism of action/pharmaceutical class classification. Finally, a binary variable was created (Yes/No) based on the medication classes to reflect having an initial prescription of newer medications to be used as our outcome variable.
3. Three tables were loaded using the Cohort Builder tool; demographics (386,404 Observations), medications (854,047 observations with repeated entries over time) and select survey questions. Data cleaning and re-coding were done to individual tables then the three tables were merged using the *person_id* variable (unique participant identifier).
4. The final dataset had 386,404 participants where only 386,317 had completed health insurance status (only 176.906 had a health utilization questionnaire on file) and 44,960 had their EHR prescription data on file.

Confounding variables used in this guide were re-coded to the following:

- Race (White: 1, Black or African American: 2, Asian: 3, Other: 4, more than one race: 5)
- Ethnicity (Hispanic or Latino: 1, non-Hispanic or Latino: 2, Other: 3)
- Gender (Female: 1, Male: 2, Not man only, not woman only, prefer not to answer, or skipped: 3)

Statistical analysis procedures

Step 1: Describe the variables to be evaluated for further analysis.

Example code:

```
proc freq data=demosurvey;
table insurance race;
run;
```

Example results:

The FREQ Procedure

insurance	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Frequency Missing = 1748				
.	12573	3.27	12573	3.27
1	345648	89.86	358221	93.13
2	26435	6.87	384656	100.00

89.6% have health insurance.

race, 5 gr (white/black/Asian/none of the above/more than one)

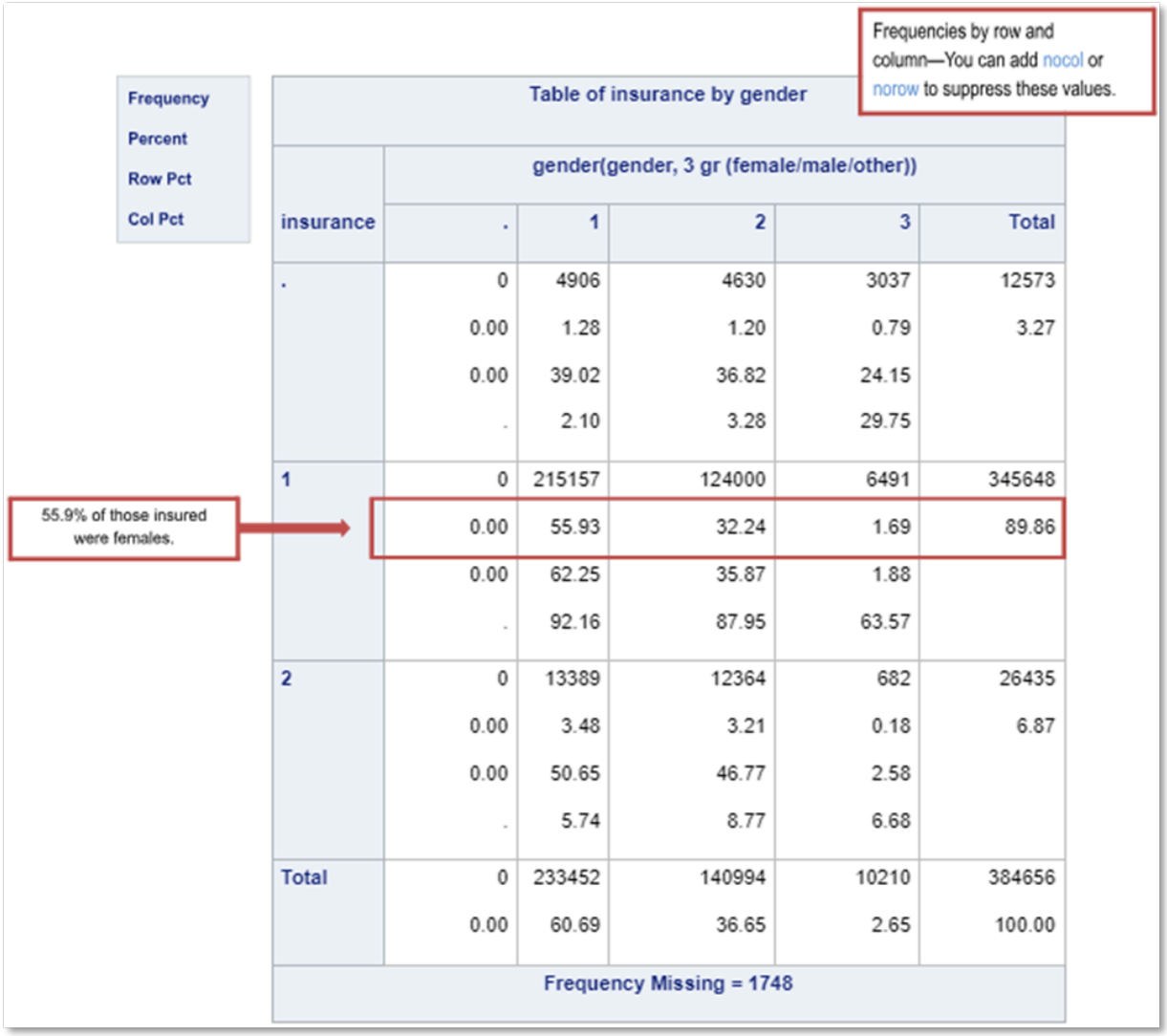
race	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	73063	18.91	73063	18.91
1	209263	54.16	282326	73.06
2	75797	19.62	358123	92.68
3	13762	3.56	371885	96.24
4	6999	1.81	378884	98.05
5	7520	1.95	386404	100.00

Step 2: Using cross tabulation and Pearson's Chi-square test to examine differences in distribution or interactions with our confounders.

Example code:

```
proc freq data=demosurvey;
table insurance*gender / chisq;
run;
```

Example results:



Frequencies by row and column—You can add nocol or norow to suppress these values.

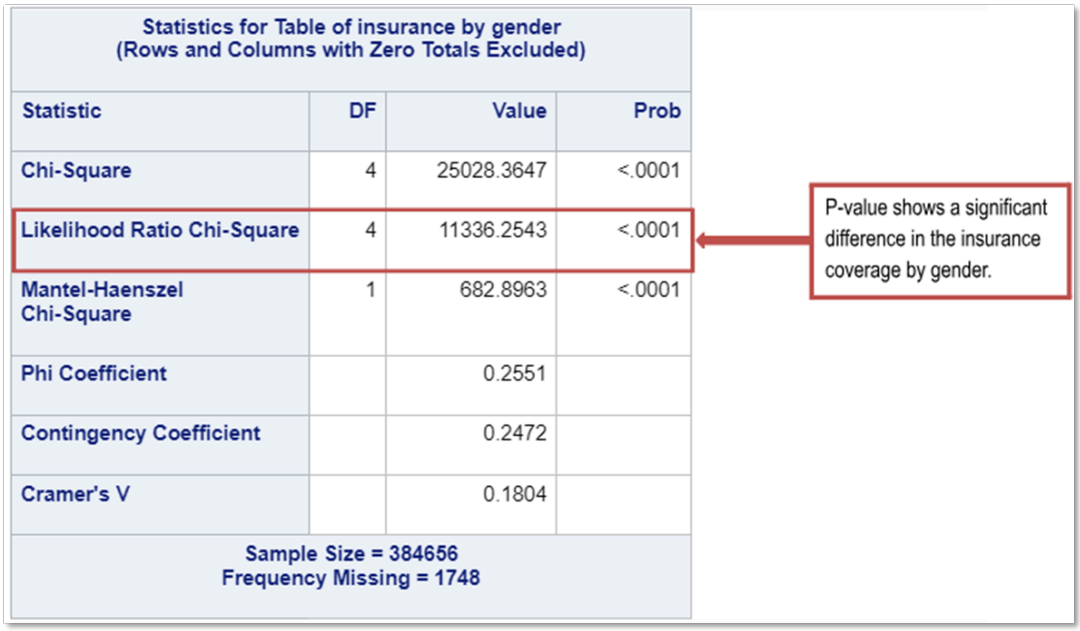
Frequency
Percent
Row Pct
Col Pct

Table of insurance by gender

insurance	gender(gender, 3 gr (female/male/other)) .	1	2	3	Total
.	0	4906	4630	3037	12573
	0.00	1.28	1.20	0.79	3.27
	0.00	39.02	36.82	24.15	
	.	2.10	3.28	29.75	
1	0	215157	124000	6491	345648
	0.00	55.93	32.24	1.69	89.86
	0.00	62.25	35.87	1.88	
	.	92.16	87.95	63.57	
2	0	13389	12364	682	26435
	0.00	3.48	3.21	0.18	6.87
	0.00	50.65	46.77	2.58	
	.	5.74	8.77	6.68	
Total	0	233452	140994	10210	384656
	0.00	60.69	36.65	2.65	100.00

Frequency Missing = 1748

55.9% of those insured were females.



Statistics for Table of insurance by gender
(Rows and Columns with Zero Totals Excluded)

Statistic	DF	Value	Prob
Chi-Square	4	25028.3647	<.0001
Likelihood Ratio Chi-Square	4	11336.2543	<.0001
Mantel-Haenszel Chi-Square	1	682.8963	<.0001
Phi Coefficient		0.2551	
Contingency Coefficient		0.2472	
Cramer's V		0.1804	

Sample Size = 384656
Frequency Missing = 1748

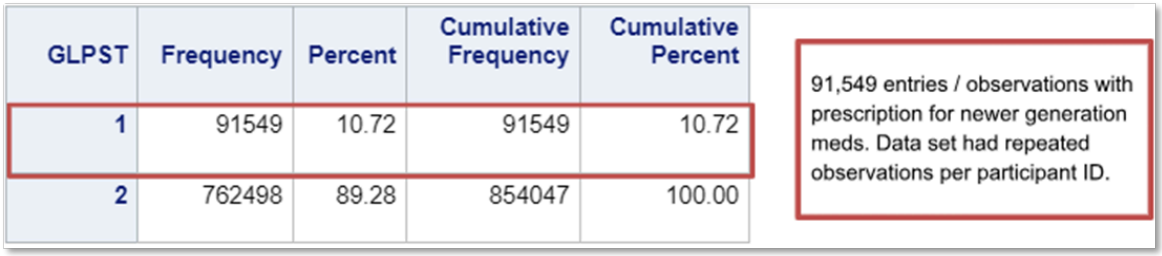
Step 3: Recoding the outcome variable.

Example code:

```
data meds6;/*5=SGLT2/6=GLP1/3=ddp4 and 7 are combined new formula*/
set meds5;
if medsclass=5 then GLPST=1;
else if medsclass=6 then GLPST=1;
else if medsclass=3 then GLPST=1;
else if medsclass=7 then GLPST=1;
else GLPST=2;
run;
proc freq data=meds6;
table GLPST;
run;
```

This recoding was done in the original medications table with the repeated observation and was used later coupled with the *person_ID* to drop duplicates keeping the first entry to highlight those with initial or first prescription type.

Example results:



GLPST	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	91549	10.72	91549	10.72
2	762498	89.28	854047	100.00

After dropping duplicate entries by person_ID and GLPST, we had 9,469 participants with newer initial prescription meds (21% of the total participants with EHR information on file).

Step 4: Use the **LOGISTIC** procedure to fit a two-way logit model for the effect of having insurance with race and gender as covariates.

Example code:

```
proc logistic data=demomedsurvey;/*running a demo logistic regression*/
class insurance (ref="2") race (ref="1") gender;
model GLPST= insurance race gender/ expb;
run;
```

Example results:
In this analysis, **PROC LOGISTIC** models the probability of no newer generation prescription medication (***GLPST***=No).

The LOGISTIC Procedure

Number of Observations Read	394691
Number of Observations Used	44765

Response Profile

Ordered Value	GLPST	Total Frequency
1	1	9469
2	2	35296

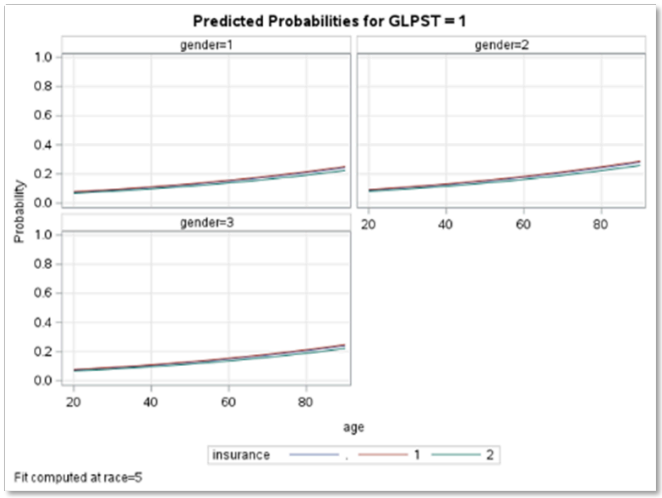
Probability modeled is GLPST=1.
Note:349926 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	46197.030	46051.282
SC	46205.739	46138.374
-2 Log L	46195.030	46031.282



Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	163.7480	9	<.0001
Score	163.8551	9	<.0001
Wald	163.1166	9	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
insurance	2	18.9366	<.0001
race	5	28.2035	<.0001
gender	2	128.5001	<.0001

Both independent variable and confounders are significant.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	-1.4463	0.0449	1035.6526	<.0001	0.235
insurance	.	1	0.0505	0.0552	0.8353	0.3607	1.052
insurance	1	1	0.1021	0.0334	9.3264	0.0023	1.108
race	.	1	0.0760	0.0346	4.8318	0.0279	1.079
race	2	1	0.0818	0.0334	5.9801	0.0145	1.085
race	3	1	0.1173	0.0675	3.0238	0.0821	1.124
race	4	1	0.0378	0.0759	0.2480	0.6185	1.039
race	5	1	-0.2860	0.0972	8.6631	0.0032	0.751
gender	1	1	-0.0913	0.0299	9.3565	0.0022	0.913
gender	2	1	0.1773	0.0304	34.0660	<.0001	1.194

Step 5: The following **PROC LOGISTIC** statements illustrate the use of forward selection to identify the effects that differentiate the two GLPST responses. The option **SELECTION=FORWARD** is specified to carry out the forward selection. Notice that we did not specify a reference to illustrate the effect coding.

Example code:

```
proc logistic data=demomedsurvey;/*running same model with forward selection*/
class insurance race gender;
model GLPST= insurance race gender/ selection=forward expb;
run;
```

Example results:

Forward Selection Procedure

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

-2 Log L	=	46195.030

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
163.8551	9	<.0001

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Variable Label
1	gender	2	1	121.1584	<.0001	gender, 3 gr (female/male/other)
2	race	5	2	23.5876	0.0003	race, 5 gr (white/black/asian/none of the above/more than one)
3	insurance	2	3	19.0377	<.0001	

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
insurance	2	18.9366	<.0001
race	5	28.2035	<.0001
gender	2	128.5001	<.0001

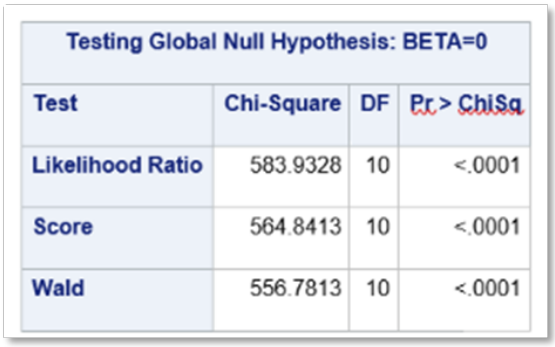
All three variables are statistically significant at the 0.05 level with forward selection.

Step 6: Finally, the following statements refit the previously selected model, except that reference coding is used for the **CLASS** variables instead of effect coding.

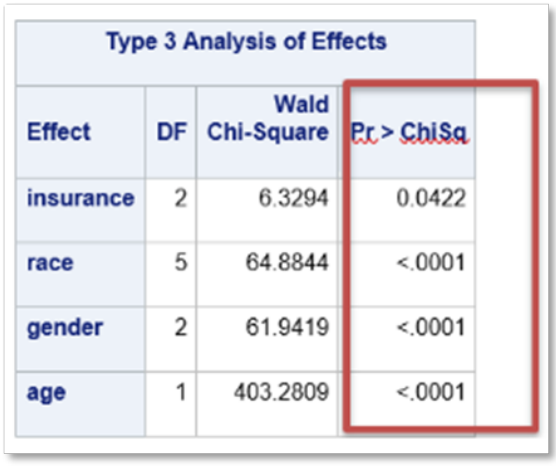
Example code:

```
ods graphics on;
proc logistic data=demomedsurvey plots(only maxpoints=none)=(oddsratio(range=clip));
    class insurance race gender /param=ref;
    model GLPST= insurance race gender Age / noor;
    oddsratio insurance;
    oddsratio race;
    contrast 'insurance 1 vs 2' insurance 1 2 / estimate=exp;
    contrast 'race 1 vs 2' race 1 - 2 / estimate=exp;
    contrast 'race 1 vs 3' race 1-3 / estimate=exp;
    contrast 'race 1 vs 4' race 1 -4 / estimate=exp;
    effectplot / at(gender=all) noobs;
    effectplot slicefit(sliceby=gender plotby=insurance / noobs;
run;
```

Example results:



Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	583.9328	10	<.0001
Score	564.8413	10	<.0001
Wald	556.7813	10	<.0001

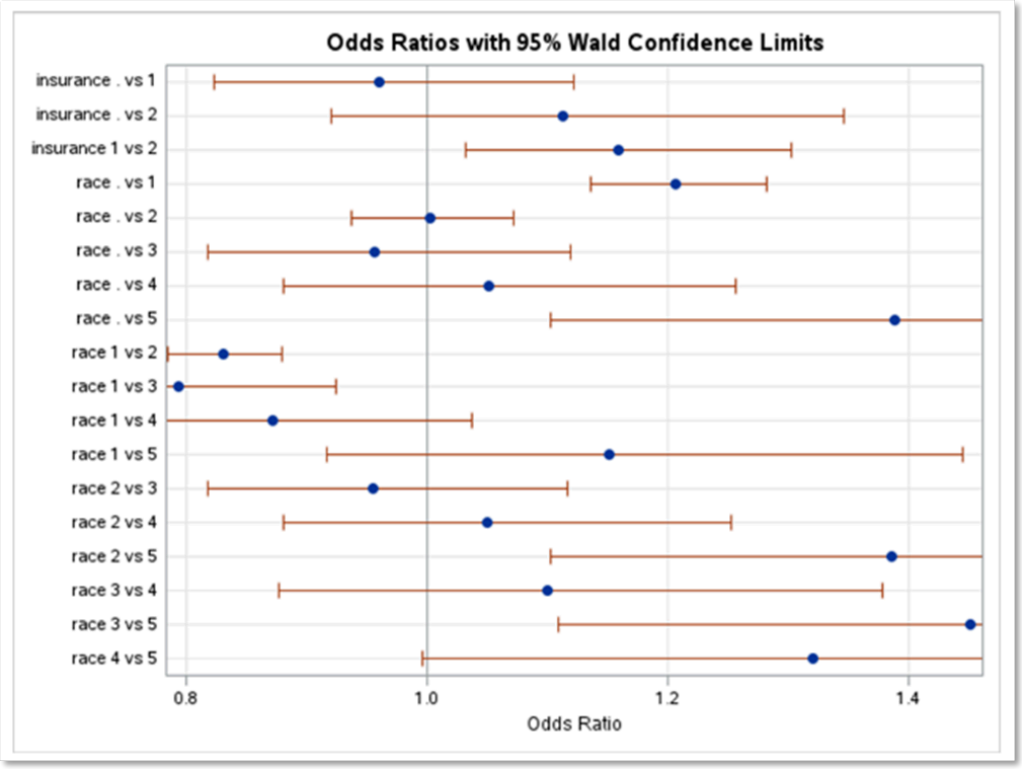


Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
insurance	2	6.3294	0.0422
race	5	64.8844	<.0001
gender	2	61.9419	<.0001
age	1	403.2809	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.0230	0.1616	350.0462	<.0001
insurance	.	1	0.1067	0.0969	1.2138	0.2706
insurance	1	1	0.1473	0.0593	6.1649	0.0130
race	.	1	0.3275	0.1175	7.7662	0.0053
race	1	1	0.1399	0.1161	1.4519	0.2282
race	2	1	0.3260	0.1170	7.7582	0.0053
race	3	1	0.3721	0.1375	7.3184	0.0068
race	4	1	0.2776	0.1440	3.7167	0.0539
gender	1	1	0.0177	0.0841	0.0444	0.8331
gender	2	1	0.2066	0.0848	5.9428	0.0148
age		1	0.0196	0.000976	403.2809	<.0001

Odds Ratio Estimates and Wald Confidence Intervals			
Odds Ratio	Estimate	95% Confidence Limits	
insurance . vs 1	0.960	0.822	1.121
insurance . vs 2	1.113	0.920	1.345
insurance 1 vs 2	1.159	1.032	1.302
race . vs 1	1.206	1.135	1.282
race . vs 2	1.002	0.937	1.071
race . vs 3	0.956	0.818	1.119
race . vs 4	1.051	0.880	1.256
race . vs 5	1.388	1.102	1.747
race 1 vs 2	0.830	0.784	0.879
race 1 vs 3	0.793	0.681	0.923

Odds Ratio Estimates and Wald Confidence Intervals			
Odds Ratio	Estimate	95% Confidence Limits	
race 1 vs 4	0.871	0.732	1.037
race 1 vs 5	1.150	0.916	1.444
race 2 vs 3	0.955	0.817	1.116
race 2 vs 4	1.050	0.880	1.252
race 2 vs 5	1.385	1.101	1.743
race 3 vs 4	1.099	0.877	1.378
race 3 vs 5	1.451	1.108	1.900
race 4 vs 5	1.320	0.995	1.750



Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
insurance 1 vs 2	1	4.3853	0.0363
race 1 vs 2	1	0.1505	0.6981
race 1 vs 3	1	0.1515	0.6971
race 1 vs 4	1	0.4339	0.5101

Overall Wald test for each CONTRAST statement

Contrast Estimation and Testing Results by Row

Contrast	Type	Row	Estimate	Standard Error	Alpha	Confidence Limits		Wald Chi-Square	Pr > ChiSq
insurance 1 vs 2	EXP	1	1.4938	0.2863	0.05	1.0260	2.1749	4.3853	0.0363
race 1 vs 2	EXP	1	1.0488	0.1289	0.05	0.8244	1.3344	0.1505	0.6981
race 1 vs 3	EXP	1	0.9119	0.2162	0.05	0.5730	1.4511	0.1515	0.6971
race 1 vs 4	EXP	1	0.7928	0.2795	0.05	0.3973	1.5820	0.4339	0.5101

Estimates and tests of individual contrast rows

Additional resources

For additional information about using SAS Studio in the Researcher Workbench, explore the following articles: Exploring All of Us data using SAS Studio and How to run SAS in the Researcher Workbench.

Updated April 9, 2024