LOGISTIC REGRESSION FINAL

PREDICTIVE ANALYTICAL MODEL ON LOW BIRTH WEIGHT OUTCOME

CHARIS ACKERSON, RN, BSN, BS, MPH, RNC-NIC
CHARIS.ACKERSON@GMAIL.COM
650-303-0310

Variable #	Variable Name	Description	Codes	Values	Data Manipulation Plan
1	ID	Identification code	1-189	continuous	Ignore, it's just a placeholder
2	LOW (outcome variable)	Low birth weight	0 1	≥2500g <2500g	
3	AGE	Age of mother	years	continuous	categorize
4	LWT	Weight of mother last menstrual period	pounds	continuous	categorize
5	RACE	Race	1 2 3	White Black Other	categorize
6	SMOKE	Smoking status during pregnancy	0 1	No Yes	
7	PTL	History of premature labor	0 1 2	None 1 ≥2	categorize
8	HT	History of hypertension	0 1	No Yes	
9	UI	Presence of uterine irritability	0 1	No Yes	
10	FTV	Number of physician visits during the 1st trimester	0 1 2	None 1 ≥2	categorize
11	BWT	Recorded birth weight	grams	continuous	Ignore, since LOW is essentially the same variable. This is an outcome variable

Continuous variables	=	purple
Dichotomous variables	=	blue
Categorical variables	=	green
Outcome variables	=	orange

Table 1 - LOWBWT Data Dictionary

1. **Using the lowbwt data), build a model based on goal 1. Consider collapsing categories and deleting some extreme values. Show all steps of your model building process.**

Cross tabulate, descriptive statistics, table 1
- Present (for each level of the outcome variable)
  - Variable name and description
  - Mean
  - Standard deviation
  - Minimum
  - First quartile
  - Median
  - Third quartile
  - Maximum
  - Extreme values
- Look for incorrect values and outliers

```
*Make permanent copy, but don't overwrite original dataset;
data logistic.lowbwt;
        set desktop.lowbwt;
run;

*Univariate, to check the distribution;
proc univariate data=lowbwt2;
        var age lwt race ptl ftv smoke ui ht;
run;
```

	AGE		LWT	
	Low birth weight (LOW=1) <2500g	**Not low birth weight (LOW=0) ≥2500g**	**Low birth weight (LOW=1) <2500g**	**Not low birth weight (LOW=0) ≥2500g**
Mean	22.31	23.66	122.14	133.3
Std	4.51	5.58	26.56	31.72
Min	14	14.0	80	85.0
1st quartile (25%)	19.0	19.0	103.0	113.0
2nd quartile Median (50%)	22.0	23.0	120.0	123.5
3rd quartile (75%)	25.0	28.0	130.0	147.0
4th quartile Maximum	34.0	45.0	200.0	250.0
Lowest values	14, 14, 15, 15, 16	14, 15, 16, 16, 16	80, 85, 89, 91, 92	85, 90, 90, 90, 95
Highest values	29, 30, 31, 32, 34	35, 35, 36, 36, 45	165, 187, 187, 190, 200	215, 229, 235, 241, 250

Table 1 - Descriptive statistics of the continuous study variables by LOW status

- AGE: There is one outlier (45). Age distribution depends on the outcome (low birthweight/not low birth weight). 45 is possible physiologically with in vitro fertilization, so I'll keep it.
- LWT: There are no obvious outliers. Weight of mother last menstrual period seems to depend on the outcome (low birth weight/ not low birth weight)

```
*Find quartiles for AGE;
proc univariate data=lowbwt;
        var age;
run;

*Make dummy variables for: age-->age2, age-->age3, race-->race2, lwt-->lwt2, lwt-->lwt3, ptl-->ptl2, ftv-->ftv2;
*Name dataset lowbwt2 so as to not mess up original dataset;
data lowbwt2;
        set lowbwt;
                *make age categorical by quartiles, min/max: age-->age2;
                if 14<=age<19 then age2=0;
                        else if 19<=age<23 then age2=1;
                        else if 23<=age<26 then age2=2;
                        else if 26<=age<46 then age2=3;

                *categorize age → age3;
                if 14<=age<30 then age3=0;
                        else if 30<=age<=45 then age3=1;

                *combine race 2&3 values into 2 categories
                if race=1 then race2=0;
                        else if race in (2,3) then race2=1;

                *make lwt categorical by quartiles, min/max;
                if 80<=lwt<110 then lwt2=0;
                        else if 110<=lwt<121 then lwt2=1;
                        else if 121<=lwt<140 then lwt2=2;
                        else if 140<=lwt<251 then lwt2=3;

                *categorize lwt → lwt3;
                if 80<=lwt<109 then lwt3=0;
                        else if 110<=lwt<251 then lwt3=1;

                *combine ptl 1, 2 & 3 values into 2 categories;
                if ptl=0 then ptl2=0;
                        else if ptl in (1,2,3) then ptl2=1;

                *collapse ftv;
                if ftv=0 then ftv2=0;
                        else if ftv=1 then ftv2=1;
                        else if ftv in (2,3,4,5,6) then ftv2=2;

run;
```

- Cross-tabulate outcome variable vs categorical study variables (Table 2, Ex 4_4)

```
*Frequency tables for categorical & dichotomous variables*;
proc freq data=lowbwt2;
        tables (age lwt race ptl ftv smoke ui ht)*low/norow nocol;
run;
```

- For each level of each study variable, present
  - # and % of subjects with the outcome
- Look for evidence of sparse or zero cells

		LOW = 0 (not low birth weight)		LOW = 1 (low birth weight)	
	Original Codes	**Frequency**	**Percent**	**Frequency**	**Percent**
RACE	1	73	38.62	23	12.17
Race of mother	2	15	7.94	11	5.82
(3=other race, 2=black race, 1=white race)	3	42	22.22	25	13.23
	Total	130	68.78	59	31.22
PTL	0	118	62.43	41	21.69
History of premature labor	1	8	4.23	16	8.47
(3=3preterm labor, 2=preterm labor, 1=1 preterm labor, 0=none)	2	3	1.59	2	1.06
	3	1	0.53	0	0.00
	Total	130	68.78	59	31.22
FTV	0	64	33.86	36	19.05
Number of physician visits during the 1st trimester	1	36	19.05	11	5.82
(1=1 visit, 0=no physician visits during the 1st trimester)	2	23	12.17	7	3.70
	3	3	1.59	4	2.12
	4	3	1.59	1	0.53
	6	1	0.53	0	0.00
	Total	130	68.78	59	31.22
SMOKE	0	86	45.50	29	15.34
Smoking status during pregnancy	1	44	23.28	30	15.87
SMOKE (1=yes, 0=no)	Total	130	68.78	59	31.22
UI	0	116	61.38	45	23.81
Presence of uterine irritability	1	14	7.41	14	7.41
(1=yes. 0=no presence of uterine irritability)	Total	130	68.78	59	31.22
HT	0	125	66.14	52	27.51
History of hypertension	1	5	2.65	7	3.70
(0=no history of hypertension, 1=yes)	Total	130	68.78	59	31.22

Table 2 - Frequencies and relative frequencies of the categorical & dichotomous study variables by LOW status

Yellow means sparse cells

- Collapse/remove categories if necessary and possible/meaningful

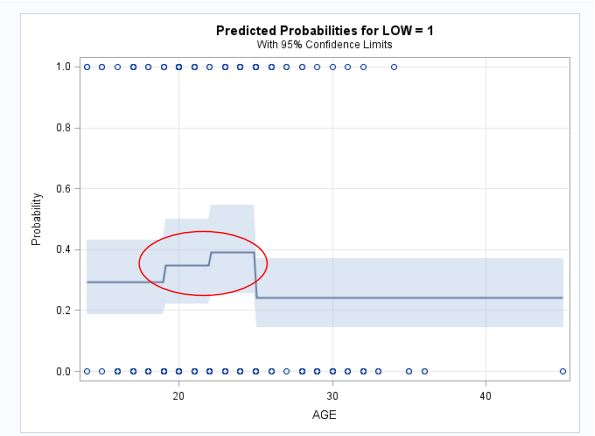
Categorical/dichotomous variables:

- RACE: Borderline sparse cells, combine 2 &3. Distribution depends on outcome (low birth weight/not low birth weight).
- PTL: Sparse cells in 1,2,3. → rewrite code to combine 1,2,3. It makes sense to combine 2 or more preterm birth histories with just 1, so I did that. Distribution depends on outcome (low birth weight/not low birth weight).
- FTV: Many sparse cells. Combine 2-6 into a new category. Distribution depends on outcome (low birth weight/not low birth weight).
- SMOKE: No sparse cells. Distribution depends on outcome (low birth weight/not low birth weight).
- UI: No sparse cells. Distribution depends on outcome (low birth weight/not low birth weight).
- HT: Sparse cells, value = 7. Can't combine since it doesn't make sense physiologically to do so (can't combine no exposure with exposure).

- Rough determination of scale

```
**********SPLINES**********;
ods trace on;

*Non-transformed data - constant connection -3 knots(quartiles);
proc logistic descending data=lowbwt2;
 effect ages=spline(age/knotmethod=list(19 22 25) basis=tpf(noint) degree=0);
 model low=ages;
effectplot;
run;
```

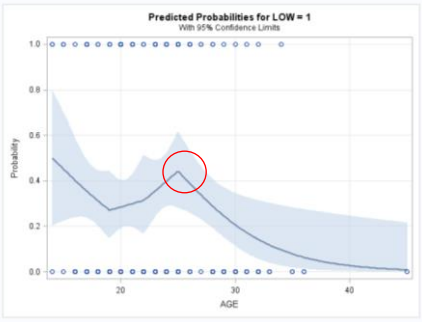


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	236.672	239.865
SC	239.914	252.832
-2 Log L	234.672	231.865

Deviance

Figure 2 - Non-transformed data - constant connection -3 knots(quartiles)

```
*Non-transformed data - linear connection -3 knots(quartiles);
proc logistic descending data=lowbwt2;
 effect ages=spline(age/knotmethod=list(19 22 25) basis=tpf(noint) degree=1);
 model low=ages;
effectplot;
run;
```

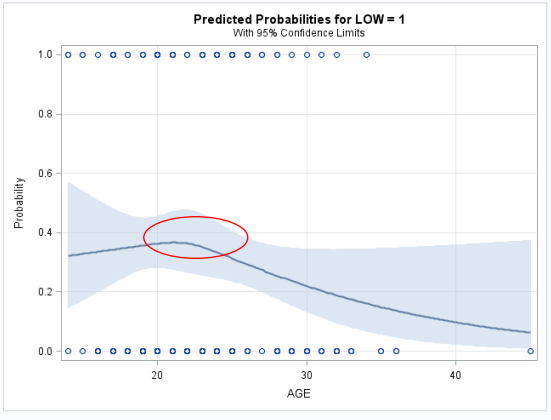


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	236.672	236.532
SC	239.914	252.740
-2 Log L	234.672	226.532

Figure 3 - Non-transformed data - linear connection -3 knots(quartiles)

Could consider transforming AGE at 25 based on splines.

```
*Non-transformed data - cubic connection -3 knots(quartiles);
proc logistic descending data=lowbwt2;
 effect ages=spline(age/knotmethod=list(19 22 25) basis=tpf(noint) naturalcubic);
 model low=ages;
effectplot;
run;
```

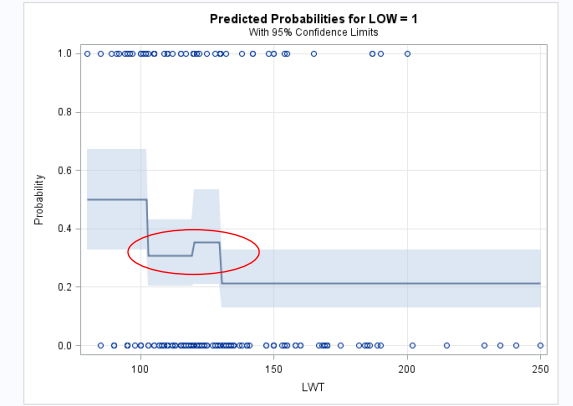


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	236.672	236.903
SC	239.914	246.628
-2 Log L	234.672	230.903

Figure 4 - Non-transformed data - cubic connection -3 knots(quartiles)

Don't do splines for categorical variables, so don't do AGE2 or AGE3, because they have been categorized (it doesn't make any sense!).

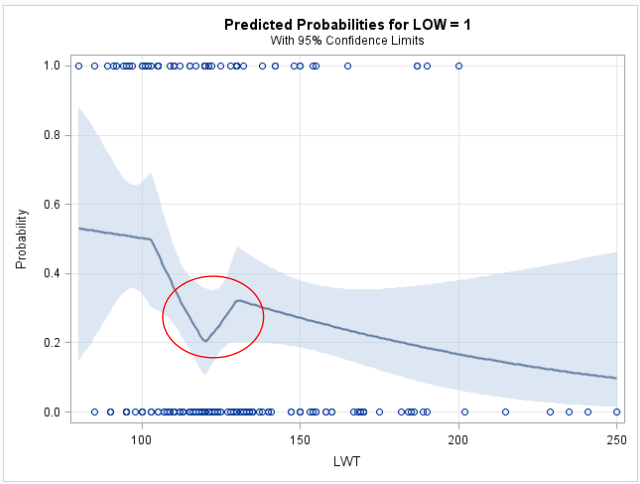
```
********************** LWT **************************;
*Non-transformed data  LWT variable - constant connection -3 knots(quartiles);
proc logistic descending data=lowbwt2;
 effect lwts=spline(lwt/knotmethod=list(103 120 130) basis=tpf(noint) degree=0);
 model low=lwts;
effectplot;
run;
```



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	236.672	234.537
SC	239.914	247.504
-2 Log L	234.672	226.537

Figure 5 - Non-transformed data LWT variable: constant connection – 3 knots (quartiles)

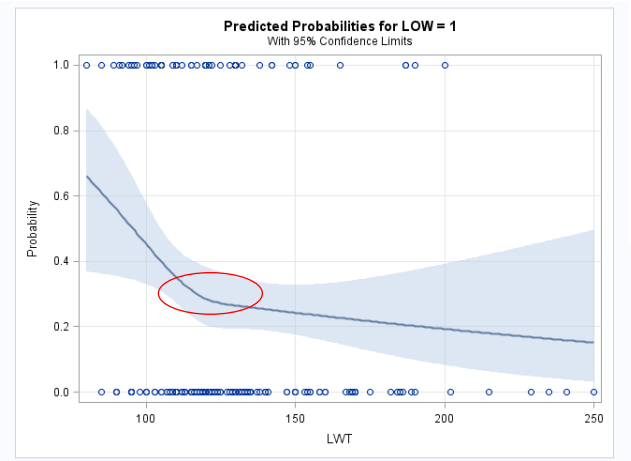
```
*Non-transformed data LWT variable : linear connection 3 knots;
proc logistic descending data=lowbwt2;
 effect lwts=spline(lwt/knotmethod=list(103 120 130) basis=tpf(noint) degree=1);
 model low=lwts;
effectplot;
run;
```



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	236.672	234.363
SC	239.914	250.571
-2 Log L	234.672	224.363

Figure 6 - Non-transformed data LWT variable: linear connection – 3 knots (quartiles)

```
*Non-transformed data LWT variable - cubic connection -3 knots(quartiles);
proc logistic descending data=lowbwt2;
 effect lwts=spline(lwt/knotmethod=list(103 120 130) basis=tpf(noint) naturalcubic);
 model low=lwts;
effectplot;
run;
```



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	236.672	232.405
SC	239.914	242.131
-2 Log L	234.672	226.405

Figure 7 - Non-transformed data LWT variable: cubic connection – 3 knots (quartiles)

Cutpoints identified from Spline Plots

Consider transforming LWT at 120 based on splines.

Consider transforming AGE at 25. So, create the following 4-level categorical variable LWT3 and AGE2 for now.

New Variable	New Code	Original Code	Units
AGE2	0	14-18	years
Age of mother	1	19-22	years
	2	23-25	years
	3	26-46	years
LWT2	0	80-109	pounds
Weight of mother last menstrual period	1	110-120	pounds
	2	121-139	pounds
	3	140-250	pounds

Table 2b – AGE2 and LWT3 Variables comparing new code

Design Variable Plots:

```
************************ DESIGN VARIABLE PLOTS**************************;

************** Design variable plot AGE – Univariate******************;
 *Design variables plots for continuous model covariates;
proc univariate  data=lowbwt2;
        var age2;
run;

proc logistic descending data=lowbwt2;
        class age2/param=ref ref=first;
        model low=age2;
run;
```

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
AGE2 1 vs 0	0.841	0.345	2.052
AGE2 2 vs 0	1.227	0.480	3.136
AGE2 3 vs 0	0.608	0.238	1.550

Figure 8 – PROC UNIVARIATE to get continuous variable coefficients for AGE

```
*Design variable plots for AGE;
*(0=14-<19, 1=19<23, 2=23-<26, 3=26-<46 years,);
data dvplot_age;
        input mp coeff;
        cards;
16.5            0
21              0.841
24.5            1.227
36              0.608
run;

axis1 minor=none label=(f=swiss h=2.5 'age');
axis2 minor=none label=(f=swiss h=2.5 a=90 'ln(or) ');
goptions ftext=swissb htext=2.0 hsize=6 in vsize= 6 in;
symbol1 c=black v=dot i=stepjc;
proc gplot data=dvplot_age;
        plot coeff*mp/haxis=axis1 vaxis=axis2;
        run;
quit;
```

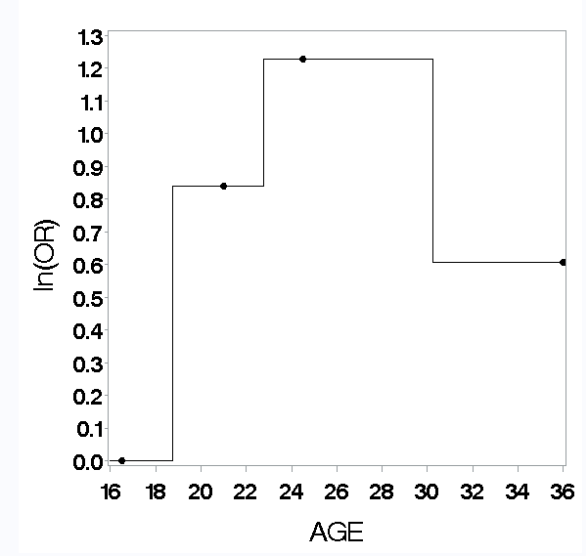


Figure 9 – Design variable plot for AGE

```
********** Design variable plot for LWT*************;
proc univariate  data=lowbwt2;
        var lwt2;
run;
proc logistic descending data=lowbwt2;
        class lwt2/param=ref ref=first;
        model low=lwt2;
run;
```

Odds Ratio Estimates

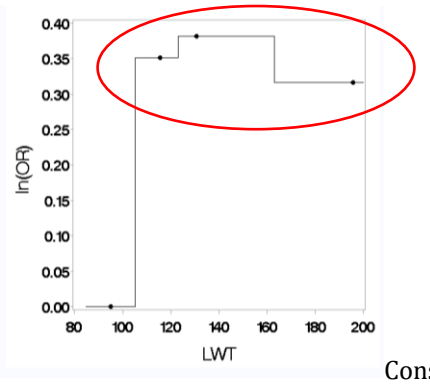
Effect	Point Estimate	95% Wald Confidence Limits	
LWT2 1 vs 0	0.351	0.147	0.843
LWT2 2 vs 0	0.382	0.159	0.922
LWT2 3 vs 0	0.316	0.130	0.767

Figure 10 – PROC UNIVARIATE to get continuous variable coefficients for LWT

```
*(0-80-<110, 1=110-<121, 2=121-<140, 3=140- <251);
*Design variable plots for LWT;
data dvplot_lwt;
        input mp coeff;
        cards;
95              0
115.5           0.351
130.5           0.382
195.5           0.316
run;
```

```
axis1 minor=none label=(f=swiss h=2.5 'lwt');
axis2 minor=none label=(f=swiss h=2.5 a=90 'ln(or) ');
goptions ftext=swissb htext=2.0 hsize=6 in vsize= 6 in;
symbol1 c=black v=dot i=stepjc;

proc gplot data=dvplot_lwt;
        plot coeff*mp/haxis=axis1 vaxis=axis2;
        run;
quit;
```



Consider collapsing these groups

Figure 11 – Design variable plot for LWT

Variable	Comparison	Coefficient	p-value	Interval midpoints
AGE2	AGE (0=14-<19, 1=19<23, 2=23-<26, 3=26-<46 years,)	0		16.5
	AGE2 1 vs 0	0.841	0.7042	21
	AGE2 2 vs 0	1.227	0.6696	24.5
	AGE 2 3 vs 0	0.606	0.2972	36
LWT2	LWT (0=80-<110, 1=110-<121, 2=121-<140, 3=140-<251)	0		95
	LWT2 1 vs 0	0.351	0.0191	115.5
	LWT2 2 vs 0	0.382	0.0323	130.5
	LWT2 3 vs 0	0.316	0.0109	195.5

Table 2d – (AB table 6) Univariate Design Variable Plots for AGE and LWT
Yellow = p<0.05 (significant)

Based on the design plots, keep age categorical as age2. There is no good way to recategorize age other than age2. lwt2 should be recategorized based on figure 11 design variable plot. So categorize lwt2 as lwt3.

<u>lwt3:</u>
lwt3=0 if lwt is 80-109 pounds
lwt3=1 if lwt is 110-251 pounds

<u>age2:</u>
age2=0 if age is 14-18 years
age2=1 if age is 19-22 years
age2=2 if age is 23-25 years
age2=3 if age is 26-46 years
so i'm keeping age2 and lwt3 categorical.

New Variable	New Code	Original Code	Units
AGE2	0	14-18	years
Age of mother	1	19-22	years
	2	23-25	years
	3	26-46	years
LWT3	0	80-109	pounds
Weight of mother last menstrual period	1	110--251	pounds

*Do UNIVARIATE FP (Fractional polynomials) Macro here:

```
**************************fp for LWT************************;
%macro fp1(dset,death,var,lb,p1);
 %do %until(&p1=7);
 %put ***** &p1 *****;
 ODS output FitStatistics = mfs;
 data fpdat; set &dset; if &var>&lb; pc=&p1/2;
   if pc ne 0 then F1=&var**pc; else if pc = 0 then F1=log(&var);
 run;
 proc logistic descending data=fpdat; model &death=F1; run;
 data mfs; set mfs; if criterion='-2 Log L'; drop Criterion InterceptOnly; run;
 proc append data=mfs base=tres; run;
 proc datasets; delete fpdat mfs; run;
 quit;
 %let p1=%eval(&p1+1);
 %end;
%mend fp1;
%fp1(lowbwt2,low,LWT,0,-4);

data pvals; do p1=-4 to 6; output; end; run;
data pvals; set pvals; p1=p1/2; run;
data tres; merge pvals tres; if p1 in (-1.5, 1.5, 2.5) then delete; run;
proc sort data=tres; by InterceptAndCovariates; run;
data tres; set tres; if _N_=1 or p1=1; run;

%macro fp2(dset,death,var,lb,p1,p2);
 %do %until(&p1=7);
  %do %until(&p2=7);
 %put ***** &p1 &p2 *****;
 ODS output FitStatistics = mfs;
 data fpdat; set &dset; if &var>&lb; pc1=&p1/2; pc2=&p2/2;
   if pc1 ne 0 then F1=&var**pc1; else if pc1 = 0 then F1=log(&var);
   if pc1 ne pc2 then do; if pc2 ne 0 then F2=&var**pc2;
                  else if pc2 = 0 then F2=log(&var); end;
   if pc1=pc2 then F2=F1*log(&var);
 run;

 proc logistic descending data=fpdat; model &death=F1 F2; run;
 data mfs; set mfs; if criterion='-2 Log L'; drop Criterion InterceptOnly; run;
 proc append data=mfs base=tres2; run;
 proc datasets; delete fpdat mfs; run;
 quit;
  %let p2=%eval(&p2+1);
```

```
 %end;
 %let p2=%eval(-4);
 %let p1=%eval(&p1+1);
 %end;
%mend fp2;
%fp2(LOWBWT2,LOW,LWT,0,-4,-4);

data pvals2; do p1=-4 to 6; do p2=-4 to 6; output;end; end; run;
data pvals2; set pvals2; p1=p1/2; p2=p2/2; run;
data tres2; merge pvals2 tres2;
 if p1 in (-1.5, 1.5, 2.5) or p2 in (-1.5, 1.5, 2.5) then delete; run;
proc sort data=tres2; by InterceptAndCovariates; run;
data tres2; set tres2; if _N_=1; run;

data comb; set tres tres2; run;

data c1; set comb; if p1=1 and p2=.; rename InterceptAndCovariates=Dev_linear;
                   drop p1 p2; run;
data c2; set comb; if p1 ne 1 and p2=.; rename InterceptAndCovariates=Dev_fp1;
                     rename p1=e_fp1; drop p2; run;
data c3; set comb; if p2 ne .; rename InterceptAndCovariates=Dev_fp2;
                rename p1=e1_fp2; rename p2=e2_fp2; run;

data c;
 merge c1 c2 c3;
 diff_lin_fp1=Dev_linear-Dev_fp1;
 diff_lin_fp2=Dev_linear-Dev_fp2;
 diff_fp1_fp2=Dev_fp1-Dev_fp2;

 p_lin_fp1=1-probchi(diff_lin_fp1,1);
 p_lin_fp2=1-probchi(diff_lin_fp2,3);
 p_fp1_fp2=1-probchi(diff_fp1_fp2,2);
run;

proc print noobs data=c;
 var Dev_linear e_fp1 Dev_fp1 e1_fp2 e2_fp2 Dev_fp2 p_lin_fp1 p_lin_fp2 p_fp1_fp2;
 format p_lin_fp1 p_lin_fp2 p_fp1_fp2 6.4;
run;

proc datasets; delete tres tres2 pvals pvals2 comb c c1 c2 c3; run; quit;
**********************************END FP FOR LWT**********************
```

UNIVARIATE FP for LWT:

The SAS System

Dev_linear	e_fp1	Dev_fp1	e1_fp2	e2_fp2	Dev_fp2	p_lin_fp1	p_lin_fp2	p_fp1_fp2
228.691	-2	227.276	-2	-2	227.163	0.2343	0.6760	0.9451

Figure 12 - UNIVARIATE FP for LWT

$x^{-2} = 1/x^2$ $x^{-2} = 1/x^2$ $x^{-2} = 1/x^2$

UNIVARIATE FP for AGE:

The SAS System

Dev_linear	e_fp1	Dev_fp1	e1_fp2	e2_fp2	Dev_fp2	p_lin_fp1	p_lin_fp2	p_fp1_fp2
231.912	3	230.982	3	3	229.868	0.3348	0.5632	0.5729

Figure 13 - UNIVARIATE FP for AGE

x^3 x^3 x^3

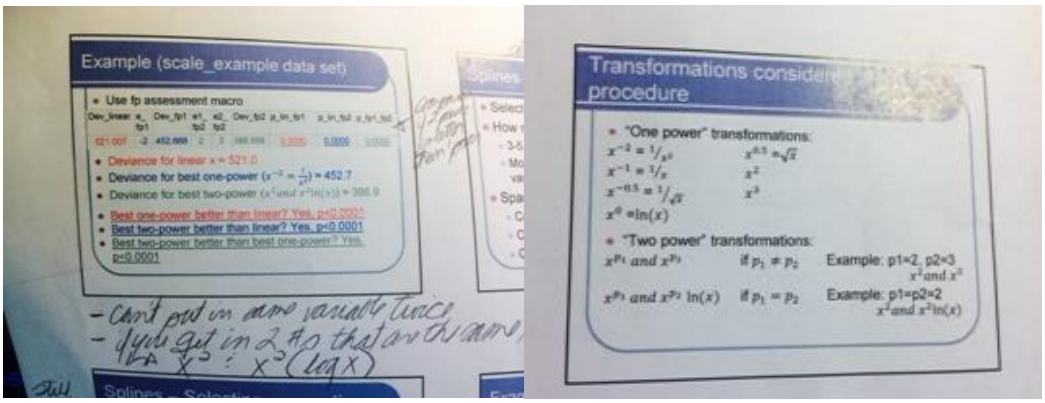


Figure 6 – FP Macro transformations (Source: Bachand)

Variable	Transformation	Deviance	Comparison	p-value	FP Macro Variable	FP Macro Value
AGE	Linear				Dev_linear	
	Best 1-power	231.912	Best 1-power vs linear	0.3348	e_fp1	3
	Best 2-power $1/x^2$	230.982	Best 2-power vs linear	0.5632	e1_fp2	3
	$p^1 = p^2$ so, x^3 and $x^3\ln(x)$	229.868	Best 2-power vs best 1-power	0.5729	e2_fp2	3
LWT	Linear				Dev_linear	
	Best 1-power: x^3	228.691	Best 1-power vs linear	0.2343	e_fp1	-2
	Best 2-power: $1/x^2$	227.276	Best 2-power vs linear	0.6760	e1_fp2	-2
	$p^1 = p^2$ so, $1/x^2$ and $(1/x^2)\ln(x)$	227.163	Best 2-power vs best 1-power	0.9451	e2_fp2	-2

Table 2c (AB table 5)– Results of FP procedure for variables AGE and LWT

FP for variable AGE:

- Best one-power better than linear?
  - No, p=0.3348
- Best two-power better than linear?
  - No, p=0.5632
- Best two-power better than best one-power?
  - No, p=0.5729

FP for variable LWT:

- Best one-power better than linear?
  - No, p=0.2343
- Best two-power better than linear?
  - No, p=0.6760
- Best two-power better than best one-power?
  - No, p=0.9451

No transformations suggested, confirm with design variable plots. The design variable plot in table 2d for LWT2 demonstrates that LWT2 is statistically significant and that the coefficients are very similar to eachother for LWT2 1 vs 0 (0.351), 2 vs 0 (0.382), and 3 vs 0 (0.316), therefore recategorize as LWT3.

Variable	Code	pounds
LWT2	0	80-<110
	1	110-<121
	2	121-<140
	3	140-<251
LWT3	**0**	80-109
	1	110--251

LWT = weight of mother last menstrual period

All p-values are non-significant when checking AGE and LWT scale using the FP macro, so AGE and LWT will be made categorical. For LWT and AGE, $p^1 = p^2$ the p-value was insignificant when checking the scale of LWT and AGE. Note that if I make everything categorical, I lose power in my model.

Consider for inclusion:

AGE2
RACE2
SMOKE
PTL2
HT
UI
LWT3
FTV2

Univariate modeling:

```
* Univariate models - Don't forget to use 0.25 cutoff!*;
*Checking variables for inclusion first;
proc logistic descending data=lowbwt2;
        class age2/param=ref ref=first;
        model low=age2;
run;

proc logistic descending data=lowbwt2;
        class race2/param=ref ref=first;
        model low=race2;
run;

proc logistic descending data=lowbwt2;
        class smoke/param=ref ref=first;
        model low=smoke;
run;

proc logistic descending data=lowbwt2;
        class ptl2/param=ref ref=first;
        model low=ptl2;
run;

proc logistic descending data=lowbwt2;
        class ht/param=ref ref=first;
        model low=ht;
run;

proc logistic descending data=lowbwt2;
        class ui/param=ref ref=first;
        model low=ui;
run;

proc logistic descending data=lowbwt2;
        class lwt3/param=ref ref=first;
        model low=lwt3;
run;

proc logistic descending data=lowbwt2;
        class ftv2/param=ref ref=first;
        model low=ftv2;
run;
```

Variable	Comparison/Unit/Description	OR	95% CI		p-value
SMOKE (smoking status during pregnancy)	SMOKE 1 vs 0 (1=yes vs 0=no)	2.022	1.081	3.783	0.0276
HT (history of hypertension)	HT 1 vs 0 (0=no history of hypertension, 1=yes)	3.365	1.021	11.088	0.0461
UI (presence of uterine irritability)	UI 1 vs 0 (1=yes vs 0=no presence of uterine irritability)	2.578	1.139	5.834	0.0231
AGE2 (Age of mother)	AGE2 1 vs 0 (1=19-22, 0=14-18 years)	0.841	0.345	2.052	0.7042
	AGE2 2 vs 0 (2=23-25, 0=14-18 years)	1.227	0.480	3.136	0.6696
	AGE2 3 vs 0 (3=26-46, 0=14-18 years)	0.608	0.238	1.550	0.2972
RACE2 (race of mother)	RACE2 1 vs 0 (1=black+other race vs 0=not white, not other race)	2.004	1.070	3.754	0.0299
LWT3 (weight of mother last menstrual period)	LWT3 1 vs 0 (1=110-251 pounds vs 0=80-109 pounds)	0.349	0.169	0.717	0.0042
PTL2 (history of premature labor)	PTL2 1 vs 0 (1=≥2 vs 0 = history of premature labor)	4.317	1.916	9.726	0.0004
FTV2 (# of MD visits during 1st trimester)	FTV2 1vs 0	0.543	0.247	1.196	0.1297
	FTV2 2 vs 0 (2 or greater than 2 visits)	0.711	0.325	1.558	0.3941

Table 3 – Univariate logistic regression results

Dichotomous variables	=	blue
Categorical variables	=	green

Grey means p-value is below 0.25 cutoff
Red means wide CI width

Univariate Logistic Regression Results:

Significant variables: SMOKE, HT, UI, RACE2, LWT3, PTL2

Non-significant variables: AGE2, FTV2

Note: FTV2 has some issues. When I create design variables, I get insignificant p-values. I'm going to throw FTV2 out of the model for now.

Try AGE2 and FTV2 univariately and multivariately because they are biologically significant and may be confounders or have interaction.

Do univariate modeling with variables that weren't included. I can't here since I have all variables covered at this point. I've already decided if I'm going linear or categorical with each variable. It wouldn't make sense to do this now.
I could do likelihood ratio test for AGE2 and FTV2 in vs out of the model if I felt like it.

Consider for Inclusion
(LOWBWT2 dataset)
PTL2
LWT3
RACE2
SMOKE
HT
UI
AGE2
FTV2

1st Multivariate Model
(LOWBWT2 dataset)
PTL2
LWT3
RACE2
SMOKE
HT
UI

Variables thrown out (Univariate)
AGE2
FTV2
LWT2

Multivariate Modeling:

```
**********Multivariate Models**********;
*You can't put the same variable in twice, include AGE as continuous OR categorical variable, not both;
*HT and UI have yes/no, so can use CLASS statement here;
*CLASS statement makes design variables for categorical variables, can use it for dichotomous, it won't hurt
the model, but it's not necessary. Don't use it for continuous variables.
*Only use UNITS statement if linear, not if categorical;

proc logistic descending data=lowbwt2;
        class lwt3 race2 smoke ht ui ptl2/param=ref ref=first;
                model low= lwt3 race2 smoke ht ui ptl2/clodds=wald;
run;
```

Variable	Comparison/Unit	OR	95% CI		p-value
RACE2 (race of mother)	RACE2 1 vs 0 (1=black+other race vs 0=not white, not other race)	2.581	1.196	5.572	0.0157
SMOKE (smoking status during pregnancy)	SMOKE 1 vs 0 (1=yes vs 0=no smoking during pregnancy)	2.480	1.143	5.381	0.0215
HT (history of hypertension)	HT 1 vs 0 (0=no history of hypertension, 1=yes)	4.347	1.212	15.601	0.0242
UI (presence of uterine irritability)	UI 1 vs 0 (1=yes vs 0=no presence of uterine irritability)	2.066	0.833	5.125	0.1173
PTL2 (history of premature labor)	PTL2 1 vs 0 (1=≥2 vs 0=history of premature labor)	3.193	1.325	7.694	0.0097
LWT3 (weight of mother last menstrual period)	LWT3 1 vs 0 (0=80-109 pounds 1=110-251 pounds	0.454	0.207	0.997	0.0490

Table 4a – Multivariate Logistic Regression Results with all variables that are significant at the 0.25 level in univariate analysis

Grey means p-value is below 0.05 cutoff
Red means wide 95% CI

Significant variables: RACE2, SMOKE, HT, UI, PTL2, LWT3 (barely)

HT and PTL2 both have wide CIs, so watch them, and they both don't include 1.0 in the CI.
Model seems stable, no ridiculously wide Cis, don't need to check scale of LWT or AGE.

I don't need to do FP or design variable plots multivariately since AGE and LWT (continuous variables) have been excluded from the model.

```
proc logistic descending data=lowbwt2;
        class lwt3 race2 smoke ht ptl2/param=ref ref=first;
                model low= lwt3 race2 smoke ht ptl2/clodds=wald;
run;
```

Variable	Comparison/Unit	OR	95% CI		p-value
RACE2 (race of mother)	RACE2 1 vs 0 (1=black+other race vs 0=not white, not other race)	2.560	1.195	5.483	0.0156
SMOKE (smoking status during pregnancy)	SMOKE 1 vs 0 (1=yes vs 0=no smoking during pregnancy)	2.472	1.147	5.327	0.0209
HT (history of hypertension)	HT 1 vs 0 (0=no history of hypertension, 1=yes)	3.876	1.085	13.844	0.0370
PTL2 (history of premature labor)	PTL2 1 vs 0 (1=≥2 vs 0=history of premature labor)	3.515	1.474	8.385	0.0046
LWT3 (weight of mother last menstrual period)	LWT3 1 vs 0 (0=80-109 pounds 1=110-251 pounds	0.415	0.191	0.901	0.0261

Table 4b – Multivariate Logistic Regression Results with all variables that are significant at the 0.25 level in univariate analysis (UI removed)

Grey means p-value is below 0.05 cutoff
Red means wide 95% CI

Significant variables: RACE2, SMOKE, HT, UI, PTL2, LWT3

Final Main Effects Model:
RACE2
SMOKE
HT
PTL2
LWT3

```
*main effects model: lwt categorized;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3/param=ref ref=first;
        model low= race2 smoke ht ptl2 lwt3/clodds=wald;
run;
```

Variable	Comparison/Unit	Coefficient	Standard Error	Wald Chi-Square	P-value	OR	95% CI	
RACE2 (race of mother)	RACE2 1 vs 0 (1=black+other race vs 0=not white, not other race)	0.9398	0.3887	5.8463	0.0156	2.560	1.195	5.483
SMOKE (smoking status during pregnancy)	SMOKE 1 vs 0 (1=yes vs 0=no smoking during pregnancy)	0.9050	0.3918	5.3356	0.0209	2.472	1.147	5.327
HT (history of hypertension)	HT 1 vs 0 (0=no history of hypertension, 1=yes)	1.3547	0.6496	4.3492	0.0370	3.876	1.085	13.844
PTL2 (history of premature labor)	PTL2 1 vs 0 (1=≥2 vs 0=history of premature labor)	1.2571	0.4435	8.0329	0.0046	3.515	1.474	8.385
LWT3 (weight of mother last menstrual period)	LWT3 1 vs 0 LWT2 1 vs 0 (0=80-109 pounds 1=110-251 pounds	-0.8803	0.3957	4.9494	0.0261	0.415	0.191	0.901

Table 7: Final main effects model

Grey means p-value is below 0.05 cutoff
Red means wide 95% CI

Significant variables: RACE2, SMOKE, HT, PTL2, LWT3

Multiplicative Interaction:

Variables					
	RACE2	*SMOKE*	*HT*	*PTL2*	*LWT3*
RACE2		RACE2 * SMOKE	RACE2 * HT	RACE2 * PTL2	RACE 2 * LWT3
SMOKE			SMOKE * HT	SMOKE * PTL2	SMOKE * LWT3
HT				HT * PTL2	HT * LWT3
PTL2					PTL2 * LWT3
LWT3					

Table 7b– Variables to test in multiplicative interaction

```
**********multiplicative interaction**********;
*final model: race2 smoke ht ptl2 lwt3;


**********multiplicative interaction**********;
*final model: race2 smoke ht ptl2 lwt3;

*multiplicative interaction, one at a time with race2 * smoke;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3 race2 * smoke;
run;

*multiplicative interaction, one at a time with race2 * ht;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3 race2 * ht;
run;

*multiplicative interaction, one at a time with race2 * ptl2;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3 race2 * ptl2;
run;

*multiplicative interaction, one at a time with race2 * lwt3;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3 race2 * lwt3;
run;

*multiplicative interaction, one at a time with smoke * ht;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3 smoke * ht;
run;

*multiplicative interaction, one at a time with race2 * smoke;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3 race2 * smoke;
run;

*multiplicative interaction, one at a time with smoke * ptl2;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3 smoke * ptl2;
run;

*multiplicative interaction, one at a time with smoke * ht;
proc logistic descending data=lowbwt2;
        class race2 smoke ht  ptl2 lwt3/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3 smoke * ht;
```

```
run;

*multiplicative interaction, one at a time with smoke * lwt3;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3 smoke * lwt3;
run;

*multiplicative interaction, one at a time with smoke * lwt3;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3 smoke * lwt3;
run;

*multiplicative interaction, one at a time with ht * ptl2;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3 ht * ptl2;
run;

*multiplicative interaction, one at a time with ht * lwt3;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3 ht * lwt3;
run;

*multiplicative interaction, one at a time with ptl2 * lwt3;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3 ptl2 * lwt3;
run;
```

Interaction	1st interaction variable	2nd interaction variable	p-value
RACE2 * SMOKE	Race of mother [RACE2 1 vs 0 (1=black+ other race vs 0=not white, not other race)]	Smoking status during pregnancy, SMOKE 1 vs 0 (1=yes vs 0=no)	0.4158
RACE2 * HT	Race of mother [RACE2 1 vs 0 (1=black+ other race vs 0=not white, not other race)]	History of hypertension, HT 1 vs 0 (0=no history of hypertension, 1=yes)	0.6762
RACE2 * PTL2	Race of mother [RACE2 1 vs 0 (1=black+ other race vs 0=not white, not other race)]	History of premature labor, PTL2 1 vs 0 (1=≥2 vs 0 = history of premature labor)	0.4683
RACE 2 * LWT3	Race of mother [RACE2 1 vs 0 (1=black+ other race vs 0=not white, not other race)]	Weight of mother last menstrual period, LWT3 1 vs 0 (1=110-251 pounds vs 0=80-109 pounds)	0.2375
SMOKE * HT	Smoking status during pregnancy, SMOKE 1 vs 0 (1=yes vs 0=no)	History of hypertension, HT 1 vs 0 (0=no history of hypertension, 1=yes)	0.7450
SMOKE * PTL2	Smoking status during pregnancy, SMOKE 1 vs 0 (1=yes vs 0=no)	History of premature labor, PTL2 1 vs 0 (1=≥2 vs 0 = history of premature labor)	0.3668
SMOKE * LWT3	Smoking status during pregnancy, SMOKE 1 vs 0 (1=yes vs 0=no)	Weight of mother last menstrual period, LWT3 1 vs 0 (1=110-251 pounds vs 0=80-109 pounds)	0.2643
HT * PTL2	History of hypertension, HT 1 vs 0 (0=no history of hypertension, 1=yes)	History of premature labor, PTL2 1 vs 0 (1=≥2 vs 0 = history of premature labor)	0.9896
HT * LWT3	History of hypertension, HT 1 vs 0 (0=no history of hypertension, 1=yes)	Weight of mother last menstrual period, LWT3 1 vs 0 (1=110-251 pounds vs 0=80-109 pounds)	0.2431
PTL2 * LWT3	History of premature labor, PTL2 1 vs 0 (1=≥2 vs 0 = history of premature labor)	Weight of mother last menstrual period, LWT3 1 vs 0 (1=110-251 pounds vs 0=80-109 pounds)	0.7571

Table 8 - Multiplicative interaction terms added in one at a time, use <0.1 as cutoff

Blue means borderline interaction
Yellow means interaction

Checked multiplicative interaction with the variables that were in original multivariate model. There are no statistically significant interactions at the 0.1 level.

Final Model
RACE2
SMOKE
HT
PTL2
LWT3

```
*final model;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3/clodds=wald;
run;
```

Variable	Comparison/Unit	Coefficient	Standard Error	Wald Chi-Square	p-value
RACE2	RACE2 1 vs 0 (1=black+other race vs 0=not white, not other race)	2.560	0.3887	5.8463	0.0156
SMOKE	SMOKE 1 vs 0 (1=yes vs 0=no smoking during pregnancy)	2.472	0.3918	5.3356	0.0209
HT	HT 1 vs 0 (0=no history of hypertension, 1=yes)	3.876	0.6496	4.3492	0.0370
PTL2	PTL2 1 vs 0 (1=≥2 vs 0=history of premature labor)	3.515	0.4435	8.0329	0.0046
LWT3	LWT3 1 vs 0 LWT2 1 vs 0 (0=80-109 pounds 1=110-251 pounds	0.415	0.3957	4.9494	0.0261

Table 9 – Final Model

Grey means p-value is below 0.05 cutoff
Red means wide 95% CI

There are no ridiculously wide Cis in my final model, therefore my model is stable. All p-values are statistically significant.

2. **Perform best subsets selection to determine if you missed any important model covariates (you can use collapsed variables and scale assessment results from question 1). If necessary, make changes to your final model from question 1.**

Final Model
RACE2
SMOKE
HT
PTL2
LWT3

```
*BEST SUBSETS SELECTION;
*best subsets makes sure you didn't miss any variables;
*class statement doesn't work with best subsets;
*all variables must be dichotomous or continuous (no need to use class statement);
*if you have a categorical variable with more than 2 levels you can create your own design variables.
None of my variables have more than one level, so I don't need to.

*with lwt3;
proc logistic descending data=lowbwt2;
   model low=age smoke ht ui race2 ptl2 ftv2 lwt3/selection=score start=3 stop=7 best=4;
run;

*without lwt3;
proc logistic descending data=lowbwt2;
   model low=age smoke ht ui race2 ptl2 ftv2/selection=score start=3 stop=7 best=4;
run;
```

LWT allowed

Regression Models Selected by Score Criterion		
Number of Variables	Score Chi-Square	Variables Included in Model
3	24.8684	HT PTL2 LWT3
3	23.6579	SMOKE RACE2 PTL2
3	23.0645	RACE2 PTL2 LWT3
3	22.6245	AGE PTL2 LWT3
4	27.7454	SMOKE RACE2 PTL2 LWT3
4	27.4409	HT RACE2 PTL2 LWT3
4	27.4063	SMOKE HT RACE2 PTL2
4	27.2836	HT UI PTL2 LWT3
5	31.8146	SMOKE HT RACE2 PTL2 LWT3
5	30.5447	SMOKE HT UI RACE2 PTL2
5	29.7440	HT UI RACE2 PTL2 LWT3
5	29.3047	AGE HT UI PTL2 LWT3
6	33.9598	SMOKE HT UI RACE2 PTL2 LWT3
6	32.7579	AGE SMOKE HT RACE2 PTL2 LWT3
6	31.8604	SMOKE HT RACE2 PTL2 FTV2 LWT3
6	31.6047	AGE SMOKE HT UI RACE2 PTL2
7	34.7161	AGE SMOKE HT UI RACE2 PTL2 LWT3
7	33.9833	SMOKE HT UI RACE2 PTL2 FTV2 LWT3
7	32.7583	AGE SMOKE HT RACE2 PTL2 FTV2 LWT3
7	31.6081	AGE SMOKE HT UI RACE2 PTL2 FTV2

LWT Not allowed

Regression Models Selected by Score Criterion		
Number of Variables	Score Chi-Square	Variables Included in Model
3	23.5284	SMOKE RACE2 PTL2
3	21.9618	AGE HT PTL2
3	21.9023	HT RACE2 PTL2
3	21.6981	HT UI PTL2
4	27.1756	SMOKE HT RACE2 PTL2
4	25.8020	SMOKE UI RACE2 PTL2
4	25.2820	HT UI RACE2 PTL2
4	25.2114	AGE SMOKE RACE2 PTL2
5	30.1964	SMOKE HT UI RACE2 PTL2
5	28.8456	AGE SMOKE HT RACE2 PTL2
5	27.3578	AGE HT UI RACE2 PTL2
5	27.2818	SMOKE HT RACE2 PTL2 FTV2
6	31.5073	AGE SMOKE HT UI RACE2 PTL2
6	30.2620	SMOKE HT UI RACE2 PTL2 FTV2
6	28.8499	AGE SMOKE HT RACE2 PTL2 FTV2
6	27.4001	AGE HT UI RACE2 PTL2 FTV2
7	31.5081	AGE SMOKE HT UI RACE2 PTL2 FTV2

Figure 12 – Best Subsets Selection Results: LWT Allowed and LWT Not Allowed

Variable	Purposeful selection	Model 1 Best subsets selection (LWT allowed)	Model 2 Best subsets selection (LWT not allowed)
AGE			X
SMOKE	X	X	X
HT	X	X	X
UI		X	X
RACE2	X	X	X
PTL2	X	X	X
FTV2			
LWT3	X	X	

Table 10 – Final Main Effects Model

Model 1 (LWT allowed): may consider adding UI
Model 2 (LWT not allowed): may consider adding AGE

My final model has 5 variables (SMOKE, HT, LWT3, RACE2 and PTL2), which the best subsets selection also selected with 5 variables. This shows that I am likely missing important confounders during the purposeful model selection. AGE and UI might be considered without LWT allowed, and UI might be considered with LWT allowed. I'm going to include AGE and UI from now on, because AGE and UI are likely confounders.

Best subsets identified several models, but I chose this one because it makes sense biologically that UI (uterine irritability) would be correlated with low birth weight. An irritable uterus could potentially be a confounder for having a preterm or low birth weight infant. Because of this, I'm including UI in my model, which best subsets identified. In addition, age of the mother (AGE) could also be a confounder since teenage or advanced maternal age mothers may have a higher frequency of low birth weight infants. This is of biological significance.

I decided that I don't have interactions in my model, so I'm not including the table for "interaction terms we are including".

So… My new model will be:

Final Model (after best subsets)
RACE2
SMOKE
HT
PTL2
LWT3
UI
AGE

Need to test if they actually are confounders in my model or are statistically significant.

```
*final model after best subsets;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3 ui/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3 ui age/clodds=wald;
        units age=10;
run;
```

Variable	Comparison/Unit	Coefficient	Standard Error	Wald Chi-Square	p-value	OR	95% CI	
RACE2	RACE2 1 vs 0 (1=black+other race vs 0=not white, not other race)	0.9162	0.3978	5.3031	0.0213	2.500	1.146	5.452
SMOKE	SMOKE 1 vs 0 (1=yes vs 0=no smoking during pregnancy)	0.8488	0.3969	4.5728	0.0325	2.337	1.073	5.087
HT	HT 1 vs 0 (0=no history of hypertension, 1=yes)	1.4321	0.6459	4.9158	0.0266	4.187	1.181	14.851
PTL2	PTL2 1 vs 0 (1=≥2 vs 0=history of premature labor)	1.2662	0.4620	7.5113	0.0061	3.547	1.434	8.774
LWT3	LWT3 1 vs 0 LWT2 1 vs 0 (0=80-109 pounds 1=110-251 pounds	-0.7271	0.4014	3.2811	0.0701	0.483	0.220	1.061
AGE	10 years (age of mother)	-0.0464	0.0373	1.5400	0.2146	0.629	0.303	1.308
UI	UI 1 vs 0 1=yes 0=no uterine irritability	0.6497	0.4677	1.9300	0.1648	1.915	0.766	4.789

Table 11 – Final Model after Best Subsets Selection

Grey means p-value is below 0.05 cutoff
Red means wide 95% CI

There are no ridiculously wide Cis in my final model, therefore my model is stable.

Variable	Comparison/Unit	p-value before best subsets	p-value After best subsets (UI + AGE added)	Greater than 10% change when adding UI & AGE?
RACE2	RACE2 1 vs 0 (1=black+other race vs 0=not white, not other race)	0.0156	0.0213	yes
SMOKE	SMOKE 1 vs 0 (1=yes vs 0=no smoking during pregnancy)	0.0209	0.0325	no
HT	HT 1 vs 0 (0=no history of hypertension, 1=yes)	0.0370	0.0266	yes
PTL2	PTL2 1 vs 0 (1=≥2 vs 0=history of premature labor)	0.0046	0.0061	no
LWT3	LWT3 1 vs 0 LWT2 1 vs 0 (0=80-109 pounds 1=110-251 pounds	0.0261	0.0701	no
AGE	10 years (age of mother)	N/A	0.2146	N/A
UI	UI 1 vs 0 1=yes 0=no uterine irritability	N/A	0.1648	N/A

Table 12 – Check for confounding and statistical significance when adding UI & AGE to model after best subsets selection

Grey means p-value is below 0.05 cutoff
Blue means evidence of confounding when adding UI & AGE

Race and hypertension meet the >10% rule for confounding, which means they are potential confounders. AGE had a horrible p-value after adding it into my model after best subsets, so I'm excluding it. I'm going to try the model without AGE next, to make sure that UI will be statistically significant. UI also had a horrible p-value, so I'm going to run the model without AGE, and then consider UI for exclusion after I exclude AGE.

```
*final model after best subsets, age removed;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3 ui/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3 ui/clodds=wald;
run;
```

Variable	Comparison/Unit	Coefficient	Standard Error	Wald Chi-Square	p-value	OR	95% CI	
RACE2	RACE2 1 vs 0 (1=black+other race vs 0=not white, not other race)	0.9875	0.3900	6.4108	0.0028	2.685	1.250	5.766
SMOKE	SMOKE 1 vs 0 (1=yes vs 0=no smoking during pregnancy)	0.8708	0.3929	4.9109	0.0267	2.389	1.106	5.160
HT	HT 1 vs 0 (0=no history of hypertension, 1=yes)	1.4420	0.6515	4.8996	0.0269	4.229	1.180	15.163
PTL2	PTL2 1 vs 0 (1=≥2 vs 0=history of premature labor)	1.1499	0.4479	6.5898	0.0103	3.158	1.313	7.597
LWT3	LWT3 1 vs 0 LWT2 1 vs 0 (0=80-109 pounds 1=110-251 pounds	-0.7679	0.3998	3.6900	0.0547	0.464	0.212	1.016
UI	UI 1 vs 0 1=yes 0=no uterine irritability	0.7072	0.4616	2.3466	0.1256	2.028	0.821	5.013

Table 12 – Check for confounding and statistical significance when adding UI to model after best subsets selection, AGE removed

Grey means p-value is below 0.05 cutoff

UI had a high p-value, so I can't leave it in my model after best subsets. I'm going to exclude it next, which brings me back to the same model before best subsets selection. At least I've proved why I'm not adding AGE or UI.

```
*final model after best subsets, age & ui removed;
proc logistic descending data=lowbwt2;
        class race2 smoke ht ptl2 lwt3/param=ref ref=first;
        model low=race2 smoke ht ptl2 lwt3/clodds=wald;
run;
```

Variable	Comparison/Unit	Coefficient	Standard Error	Wald Chi-Square	p-value	OR	95% CI	
RACE2	RACE2 1 vs 0 (1=black+other race vs 0=not white, not other race)	0.9794	0.3864	6.4252	0.0113	2.663	1.249	5.679
SMOKE	SMOKE 1 vs 0 (1=yes vs 0=no smoking during pregnancy)	0.8679	0.3898	4.9583	0.0260	2.382	1.110	5.113
HT	HT 1 vs 0 (0=no history of hypertension, 1=yes)	1.3317	0.6494	4.2056	0.0403	3.788	1.061	13.524
PTL2	PTL2 1 vs 0 (1=≥2 vs 0=history of premature labor)	1.254	0.428	7.9088	0.0049	3.474	1.459	8.276
LWT3	LWT3 1 vs 0 LWT2 1 vs 0 (0=80-109 pounds 1=110-251 pounds	-0.8574	0.3943	4.7298	0.0296	0.424	0.196	0.919

Table 13 – Final Model after best subsets, after removing UI and AGE

Grey means p-value is below 0.05 cutoff
Red means wide 95% CI

There are no ridiculously wide Cis in my final model, therefore my model is stable. All p-values are statistically significant.

3. **For your final model (from question 1 or, if you made changes, from question2), use the Pearson chi-square test, deviance test, Hosmer-Lemeshow testand Osius-Rojek test to assess overall model fit. Comment on the appropriatenessof each test and draw conclusions regarding model fit.**

Final Model
RACE2
SMOKE
HT
PTL2
LWT3

```
*Pearson chi-square, deviance and Hosmer-Lemeshow test **;
proc logistic descending data=lowbwt2;
   model low=race2 smoke ht ptl2 lwt3/scale=n aggregate lackfit;
run;
```

Perform Pearson Chi-square and deviance test

Model by covariate Pattern rather than by subject

Perform Hosmer-Lemeshow test

Number of Observations Read	189
Number of Observations Used	189

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	14.3965	16	0.8998	0.5692
Pearson	12.9545	16	0.8097	0.6761

Number of unique profiles: 22

Figure 13 – Goodness of fit for Deviance and Pearson chi-square

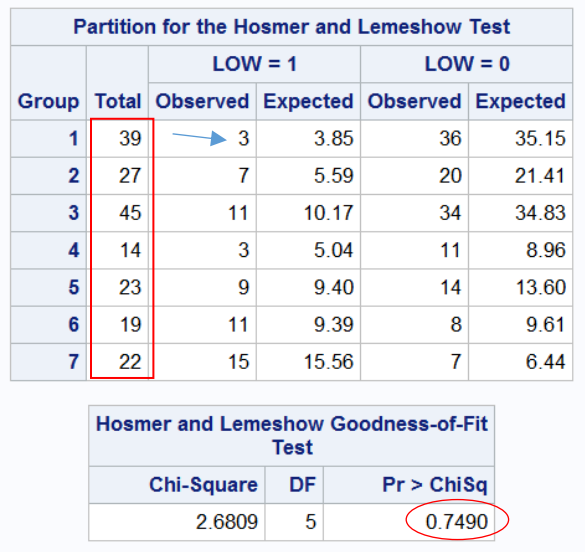
J = # of covariate patterns = 22
n = sample size = 189

Pearson χ^2 test:
"Use when there are few different covariate patterns, J, i.e. when J<<n"
The p-value is 0.6761, which is greater than 0.05. This indicates evidence of the model fitting. J = 22, and n = 189, therefore J<<n. The Pearson χ^2 is an appropriate test to run here, because the J<<n requirement is satisfied.

Bachand's Notes:
The Pearson Chi-square test (Use when J<<n):

- "Calculate the difference between the observed and the predicted value for each covariate pattern"
- "Standardize and square each difference"
- "Add the squared standardized differences over all covariate patterns"
- "If J<<n, the resulting test statistic is χ^2 distributed with J – p – 1 degrees of freedom"
- "(J = # of covariate patterns, p = # of model covariates)"
- "p ≤ 0.05 → evidence of lack of model fit"
- "p ≥ 0.05 → evidence of model fit"

Deviance test:
"Use when there are few different covariate patterns, J, i.e. when J<<n"
The p-value is 0.5692, which is greater than 0.05. This indicates evidence for model fit. J = 22, and n = 189, therefore J<<n. The Deviance test is an appropriate test to run here.
Since J≉n, the χ^2 test is *not* violated.



Partition for the Hosmer and Lemeshow Test

		LOW = 1		LOW = 0	
Group	Total	Observed	Expected	Observed	Expected
1	39	3	3.85	36	35.15
2	27	7	5.59	20	21.41
3	45	11	10.17	34	34.83
4	14	3	5.04	11	8.96
5	23	9	9.40	14	13.60
6	19	11	9.39	8	9.61
7	22	15	15.56	7	6.44

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
2.6809	5	0.7490

Figure 14 – Hosmer and Lemeshow Test for Goodness of Fit

J<<n now

p > 0.25 at 0.7490 → evidence of model fit
The assumptions are still violated because J is not that Much smaller than n, resulting in sparse cells.

There are 2 groups with <5 observed cell frequency.
There is 1 group with <5 expected cell frequency.

There are 7 covariate groups (I don't know why it's not 10 this time), The p-value is 0.7490, which is greater than 0.25. This indicates evidence of the model fitting. Since J <<N, this test should not be used (it is not appropriate), because Hosmer and Lemeshow Test should be used only "when there are many different covariate patterns, J, i.e. when J≈n". Note that the degrees of freedom here are = 5. The Hosmer and Lemeshow test for GOF test doesn't handle ties well. Ties occur "when J<n and a covariate pattern can be shared by several study subjects. The other issue is that "this test is not very powerful, especially for n<400", and since n=189 here, this is the case. Regarding "deciles of risk method: even after grouping, the expected cell frequencies may still be small".
Note: If I'm going to delete anything, it must be the entire covariate pattern, not individuals.

```
****  Osius-Rojek g-o-f test ******;
proc sort data=lowbwt2;
        by race2 smoke ht ptl2 lwt3;
run;
proc means n sum noprint data=lowbwt2;
        by race2 smoke ht ptl2 lwt3;
        var low;
        output out=odat n=m_j sum=y_j;
run;
proc logistic descending data=odat;
 model y_j/m_j=race2 smoke ht ptl2 lwt3;
 output out=pdat p=p_j;
run;

data pdat;
 set pdat;

 v_j=m_j*p_j*(1-p_j);
 c_j=(1-2*p_j)/v_j;
 chisq_j=(y_j-m_j*p_j)**2/v_j;

 m_j_inv=1/m_j;
run;

proc means sum noprint data=pdat;
 var chisq_j m_j_inv;
 output out=cdat sum=chisq m_inv;
run;

proc reg data=pdat outest=ss;
 model c_j=RACE2 SMOKE HT PTL2 LWT3;
 weight v_j;
run;

data zdat;
 merge cdat (keep=_freq_ chisq m_inv) ss (keep=_rmse_);
 *MUST BE CHANGED TO ACTUAL NUMBER OF VARIABLES IN MODEL;
 ************_*****************;
 rss=(_freq_-5-1)*_rmse_**2;

 A=2*(_freq_-m_inv);
 *MUST BE CHANGED TO ACTUAL NUMBER OF VARIABLES IN MODEL;
 ****************_*****************;
 z=(chisq-(_freq_-5-1))/sqrt(A+rss);
 z=abs(z);
 pval=(1-probnorm(z))*2;
run;

proc print noobs data=zdat;
run;
```

FREQ	chisq	m_inv	_RMSE_	rss	A	z	pval
22	12.9545	9.26135	0.90656	13.1497	25.4773	0.49002	0.62412

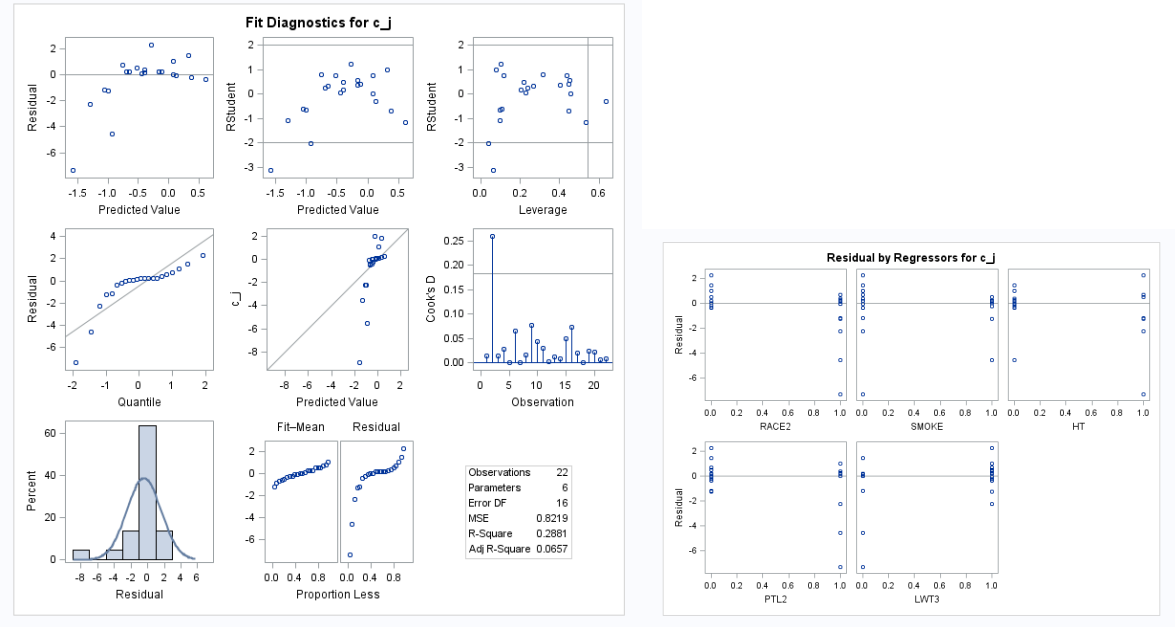


Figure 15 – Osius-Rojek test (I know you said not to include it, but it's so pretty!)

The p-value is 0.62412 which shows evidence of model fit because p-value is > 0.05, and a large p-value is evidence of the model fitting. Here, J<<n. The sample size assumption might be violated here as well, because "none of the goodness-of-fit tests are very powerful for sample sizes of less than ~400". Here n=189, so the results may be incorrect when the sample size is small. This is not an appropriate test to use here because J<<n, and this test is only appropriate when J<n.

4. Use the Stukel test to test the tails assumption. Draw conclusions.

Final Model (after best subsets)

RACE2
SMOKE
HT
PTL2
LWT3

```
**********4**********;
***** Stukel test of logistic regression model assumption ***;
*change outcome and independent variable names in model statement!;
ods output globaltests=gt1;
proc logistic descending data=lowbwt2;
 model low=race2 smoke ht ptl2 lwt3;
 output out=pdat2 xbeta=g_j p=p_j;
run;

data pdat2;
 set pdat2;
  if p_j>=0.5 then ind1=1;
              else ind1=0;
  if p_j<0.5 then ind2=1;
              else ind2=0;
              z1_j=0.5*g_j**2*ind1;
              z2_j=-0.5*g_j**2*ind2;
run;

*change outcome and independent variable names!;
ods output globaltests=gt2;
proc logistic descending data=pdat2;
 model low=race2 smoke ht ptl2 lwt3 z1_j z2_j;
run;

data pval;
 merge gt1(rename=(chisq=chisq1))
              gt2(rename=(chisq=chisq2));
              if _n_=1;
 drop test df probchisq;
              lr=chisq2-chisq1;
              pval_lr=(1-probchi(lr,2));
run;

proc print noobs data=pval;
run;
```

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.8292	0.9718	0.7280	0.3935
RACE2	1	0.4959	0.8462	0.3434	0.5579
SMOKE	1	0.4621	0.7495	0.3800	0.5376
HT	1	0.8218	1.0209	0.6480	0.4208
PTL2	1	0.7926	0.8409	0.8884	0.3459
LWT3	1	-0.5043	0.6724	0.5624	0.4533
z1_j	1	0.9138	1.8395	0.2468	0.6194
z2_j	1	0.4364	0.7038	0.3845	0.5352

Upper Tail

Lower Tail

ChiSq1	ChiSq2	lr	pval_lr
32.1924	32.6283	0.43593	0.80415

Figure 16 – Stukel test

- p > 0.05 → evidence of model fitting

The overall p-value is 0.80415, which is greater than 0.05 and thus indicates evidence of the model fitting. We can also accept the tails assumption to be true. The Stukel test is not a Goodness of fit test, but tests "whether the model produces more or fewer small or large pi-hat than the standard logistic regression model assumes. If neither extra parameter is significantly different from 0, the standard logistic regression model is fine." The shape of the lower tail exceeds 0.05 at 0.5352, which indicates model fit in the lower tail. The shape of the upper tail exceeds 0.05 at 0.6194, which indicates model fit in the upper tail.

Does the model fit overall?

➔ Summary goodness-of-fit tests: Overall I'm happy with the model, but I need to be careful proceeding since my sample size isn't ideal, I'd like it to be greater than 400, also some of my GOF tests weren't appropriate to use: the Osius-Rojek, for example.

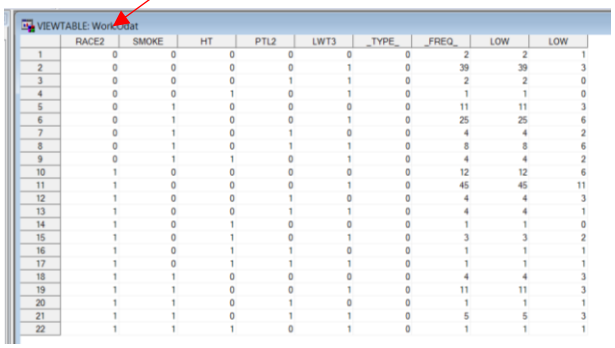
Are there any individual observations that don't fit?

➔ Logistic regression diagnostics

5. **Use all 4 logistic regression diagnostics to identify outliers. Describe why they are outliers in the model and show the effect their deletion has on the model coefficients and p-values. Propose how to deal with the outliers.**

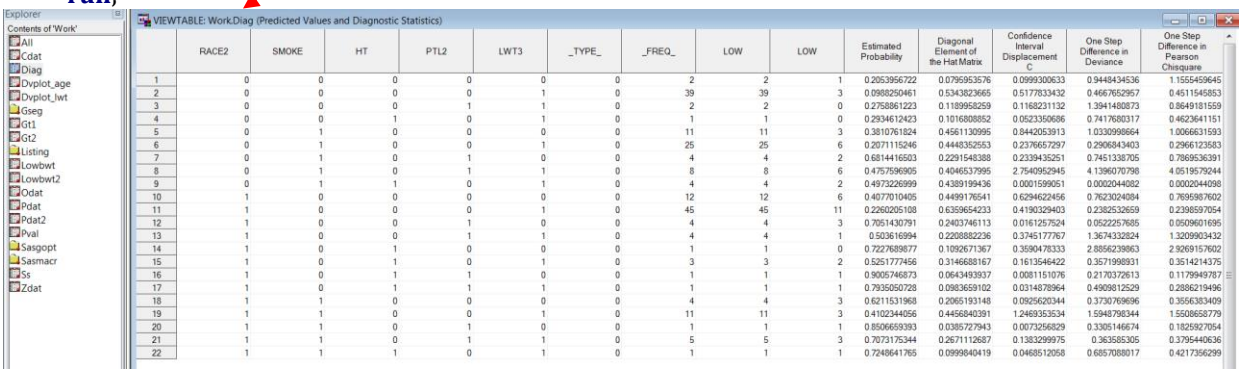
```
***** diagnostics ****;
proc sort data=lowbwt2;
 by race2 smoke ht ptl2 lwt3;
run;

proc means n sum noprint data=lowbwt2;
 by race2 smoke ht ptl2 lwt3;
 var low;
 output out=odat n=m_j sum=y_j;
run;
```



	RACE2	SMOKE	HT	PTL2	LWT3	_TYPE_	_FREQ_	LOW	LOW
1	0	0	0	0	0	0	2	2	1
2	0	0	0	0	1	0	39	39	3
3	0	0	0	1	1	0	2	2	0
4	0	0	1	0	1	0	1	1	0
5	0	1	0	0	0	0	11	11	3
6	0	1	0	0	1	0	25	25	6
7	0	1	0	1	0	0	4	4	2
8	0	1	0	1	1	0	8	8	6
9	0	1	1	0	1	0	4	4	2
10	1	0	0	0	0	0	12	12	6
11	1	0	0	0	1	0	45	45	11
12	1	0	0	1	0	0	4	4	3
13	1	0	0	1	1	0	4	4	1
14	1	0	1	0	0	0	1	1	0
15	1	0	1	0	1	0	3	3	2
16	1	0	1	1	0	0	1	1	1
17	1	0	1	1	1	0	1	1	1
18	1	1	0	0	0	0	4	4	3
19	1	1	0	0	1	0	11	11	3
20	1	1	0	1	0	0	1	1	1
21	1	1	0	1	1	0	5	5	3
22	1	1	1	0	1	0	1	1	1

```
proc logistic descending data=odat;
    model y_j/m_j = race2 smoke ht ptl2 lwt3;
    output out=diag h=h difchisq=difchisq difdev=difdev c=db p=pihat;
run;
```

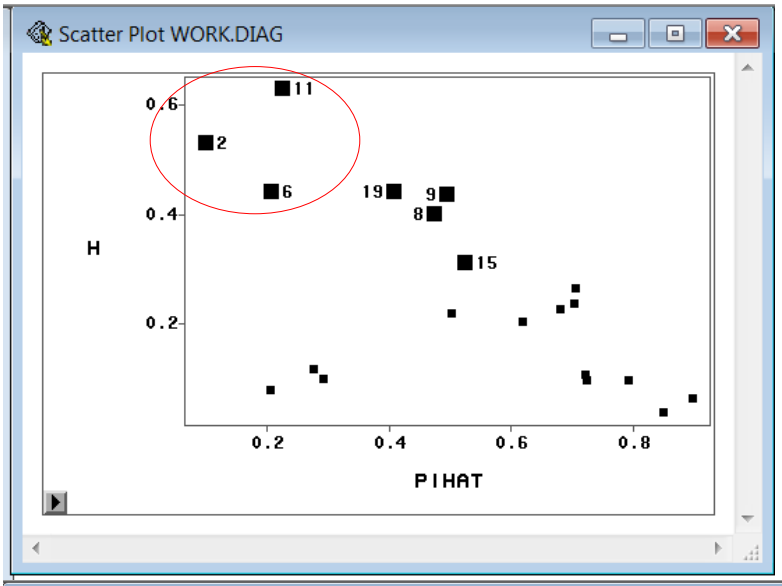


	RACE2	SMOKE	HT	PTL2	LWT3	_TYPE_	_FREQ_	LOW	LOW	Estimated Probability	Diagonal Element of the Hat Matrix	Confidence Interval Displacement C	One Step Difference in Deviance	One Step Difference in Pearson Chisquare
1	0	0	0	0	0	0	2	2	1	0.2053956722	0.0796953576	0.0999300633	0.9448434536	1.1555459645
2	0	0	0	0	1	0	39	39	3	0.0988250461	0.5343823665	0.5177833432	0.4667652957	0.4511545853
3	0	0	0	1	1	0	2	2	0	0.2758861223	0.1189958259	0.1168231132	1.3941480873	0.8649181559
4	0	0	1	0	1	0	1	1	0	0.2934612423	0.1016808852	0.0523350686	0.7417680317	0.4623641151
5	0	1	0	0	0	0	11	11	3	0.3810761824	0.4561130995	0.8442053913	1.0330998664	1.0666631593
6	0	1	0	0	1	0	25	25	6	0.2071115246	0.4483852553	0.2376657297	0.2906843403	0.2966123583
7	0	1	0	1	0	0	4	4	2	0.6814416503	0.2291548388	0.2339435251	0.7451338705	0.7869536391
8	0	1	0	1	1	0	8	8	6	0.4757596905	0.4046537995	2.7540952945	4.1396070798	4.0519579244
9	0	1	1	0	1	0	4	4	2	0.4973226999	0.4389199436	0.0001599051	0.0002044082	0.0002044098
10	1	0	0	0	0	0	12	12	6	0.4077010405	0.4499176541	0.6294622456	0.7623024084	0.7695987602
11	1	0	0	0	1	0	45	45	11	0.2260205108	0.6359654233	0.4190329403	0.2382532659	0.2398597054
12	1	0	0	1	0	0	4	4	3	0.7051430791	0.2403746113	0.0161257524	0.0522257685	0.0509601695
13	1	0	0	1	1	0	4	4	1	0.503616994	0.2208882236	0.3745177767	1.3674332824	1.3209903432
14	1	0	1	0	0	0	1	1	0	0.7227689677	0.1092671367	0.3590478333	2.8856239863	2.9269157602
15	1	0	1	0	1	0	3	3	2	0.5251777456	0.3146688167	0.1613546422	0.3571998931	0.3514214375
16	1	0	1	1	0	0	1	1	1	0.9005746873	0.0643493937	0.0081151076	0.2170372613	0.1179949787
17	1	0	1	1	1	0	1	1	1	0.7935050728	0.0983659102	0.0314878964	0.4909812529	0.2886219496
18	1	1	0	0	0	0	4	4	3	0.6211531968	0.2065193148	0.0925620344	0.3730769696	0.3556383409
19	1	1	0	0	1	0	11	11	3	0.4102344056	0.4456840391	1.2469353534	1.5948798344	1.5508658779
20	1	1	0	1	0	0	1	1	1	0.8506659393	0.0386727943	0.0073256829	0.3305146674	0.1825927054
21	1	1	0	1	1	0	5	5	3	0.7073175344	0.2671112687	0.1383299975	0.363585305	0.3795440636
22	1	1	1	0	1	0	1	1	1	0.7248641765	0.0999840419	0.0468512058	0.6857088017	0.4217356299

m_j = number of observations with covariate pattern j
y_j = number of observations with low=1 in covariate pattern j

```
*evaluate outliers;
proc sort dta=lowbwt;
by race2 smoke ht ptl2 lwt3;
run;
proc sort data=diag;
by race2 smoke ht ptl2 lwt3;
run;
data all;
merge lowbwt2 diag;
by race2 smoke ht ptl2 lwt3;
run;
```

```
* Plot diagnostics vs. pihat *;
proc insight data=diag;
 scatter h*pihat;
run;
quit;
```



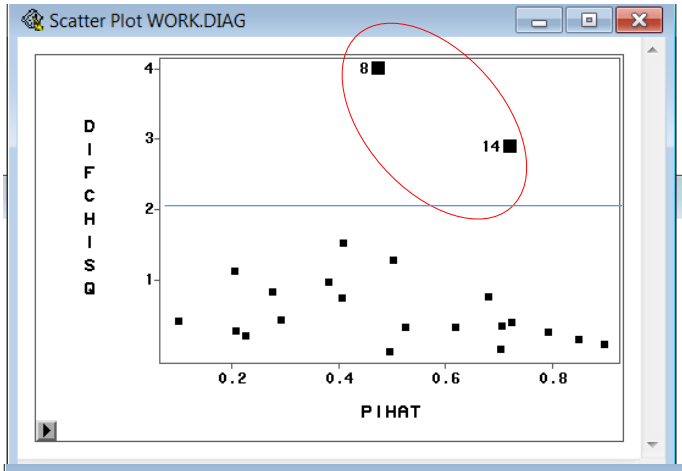
WORK.DIAG

	RACE2	SMOKE	HT	PTL2	LWT3	_TYPE_	_FREQ_	m_j	y_j	PIHAT	H	DB	DIFDEV	DIFCHISQ
1	0	0	0	0	0	0	2	2	1	0.2054	0.0796	0.0999	0.9448	1.1555
2	0	0	0	0	1	0	39	39	3	0.0988	0.5344	0.5178	0.4668	0.4512
3	0	0	0	1	1	0	2	2	0	0.2759	0.1190	0.1168	1.3941	0.8649
4	0	0	1	0	1	0	1	1	0	0.2935	0.1017	0.0523	0.7418	0.4624
5	0	1	0	0	0	0	11	11	3	0.3811	0.4561	0.8442	1.0331	1.0067
6	0	1	0	0	1	0	25	25	6	0.2071	0.4448	0.2377	0.2907	0.2966
7	0	1	0	1	0	0	4	4	2	0.6814	0.2292	0.2339	0.7451	0.7870
8	0	1	0	1	1	0	8	8	6	0.4758	0.4047	2.7541	4.1396	4.0520
9	0	1	1	0	1	0	4	4	2	0.4973	0.4389	0.0002	0.0002	0.0002
10	1	0	0	0	0	0	12	12	6	0.4077	0.4499	0.6295	0.7623	0.7696
11	1	0	0	0	1	0	45	45	11	0.2260	0.6360	0.4190	0.2383	0.2399
12	1	0	0	1	0	0	4	4	3	0.7051	0.2404	0.0161	0.0522	0.0510
13	1	0	0	1	1	0	4	4	1	0.5036	0.2209	0.3745	1.3674	1.3210
14	1	0	1	0	0	0	1	1	0	0.7228	0.1093	0.3590	2.8856	2.9269
15	1	0	1	0	1	0	3	3	2	0.5252	0.3147	0.1614	0.3572	0.3514
16	1	0	1	1	0	0	1	1	1	0.9006	0.0643	0.0081	0.2170	0.1180
17	1	0	1	1	1	0	1	1	1	0.7935	0.0984	0.0315	0.4910	0.2886
18	1	1	0	0	0	0	4	4	3	0.6212	0.2065	0.0926	0.3731	0.3556
19	1	1	0	0	1	0	11	11	3	0.4102	0.4457	1.2469	1.5949	1.5509
20	1	1	0	1	0	0	1	1	1	0.8507	0.0386	0.0073	0.3305	0.1826
21	1	1	0	1	1	0	5	5	3	0.7073	0.2671	0.1383	0.3636	0.3795
22	1	1	1	0	1	0	1	1	1	0.7249	0.1000	0.0469	0.6857	0.4217

Figure 17 – h vs Pihat

Covariate pattern outliers:
2
6
11

```
proc insight data=diag;
 scatter difchisq*pihat;
run;
quit;
```



Covariate pattern #8 & #14 is >2, so I'll change my code to reflect that. 2.0 is a good cutpoint for difchisq

WORK.DIAG

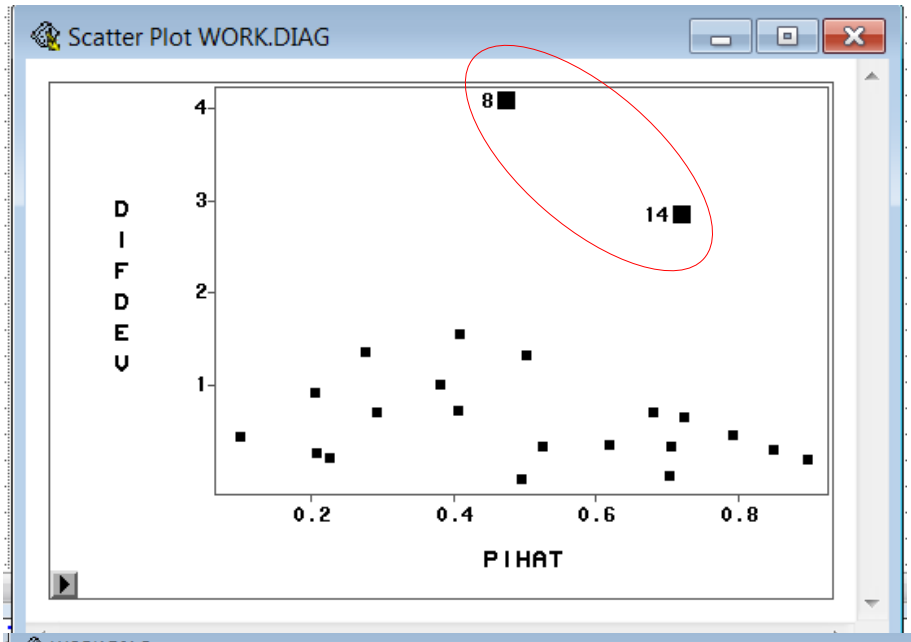
	RACE2	SMOKE	HT	PTL2	LWT3	_TYPE_	_FREQ_	m_j	y_j	PIHAT	H	DB	DIFDEV	DIFCHISQ
1	0	0	0	0	0	0	2	2	1	0.2054	0.0796	0.0999	0.9448	1.1555
2	0	0	0	0	1	0	39	39	3	0.0988	0.5344	0.5178	0.4668	0.4512
3	0	0	0	1	1	0	2	2	0	0.2759	0.1190	0.1168	1.3941	0.8649
4	0	0	1	0	1	0	1	1	0	0.2935	0.1017	0.0523	0.7418	0.4624
5	0	1	0	0	0	0	11	11	3	0.3811	0.4561	0.8442	1.0331	1.0067
6	0	1	0	0	1	0	25	25	6	0.2071	0.4448	0.2377	0.2907	0.2966
7	0	1	0	1	0	0	4	4	2	0.6814	0.2292	0.2339	0.7451	0.7870
8	0	1	0	1	1	0	8	8	6	0.4758	0.4047	2.7541	4.1396	4.0520
9	0	1	1	0	1	0	4	4	2	0.4973	0.4389	0.0002	0.0002	0.0002
10	1	0	0	0	0	0	12	12	6	0.4077	0.4499	0.6295	0.7623	0.7696
11	1	0	0	0	1	0	45	45	11	0.2260	0.6360	0.4190	0.2383	0.2399
12	1	0	0	1	0	0	4	4	3	0.7051	0.2404	0.0161	0.0522	0.0510
13	1	0	0	1	1	0	4	4	1	0.5036	0.2209	0.3745	1.3674	1.3210
14	1	0	1	0	0	0	1	1	0	0.7228	0.1093	0.3590	2.8856	2.9269
15	1	0	1	0	1	0	3	3	2	0.5252	0.3147	0.1614	0.3572	0.3514
16	1	0	1	1	0	0	1	1	1	0.9006	0.0643	0.0081	0.2170	0.1180
17	1	0	1	1	1	0	1	1	1	0.7935	0.0984	0.0315	0.4910	0.2886
18	1	1	0	0	0	0	4	4	3	0.6212	0.2065	0.0926	0.3731	0.3556
19	1	1	0	0	1	0	11	11	3	0.4102	0.4457	1.2469	1.5949	1.5509
20	1	1	0	1	0	0	1	1	1	0.8507	0.0386	0.0073	0.3305	0.1826
21	1	1	0	1	1	0	5	5	3	0.7073	0.2671	0.1383	0.3636	0.3795
22	1	1	1	0	1	0	1	1	1	0.7249	0.1000	0.0469	0.6857	0.4217

Figure 18 – difchisq vs pihat

Covariate pattern outliers:
8
14

DB =Confidence Interval Displacement C
Pihat =Estimated Probability
H =Diagonal Element of the Hat Matrix
Divdev =One Step Difference in Deviance

```
proc insight data=diag;
 scatter difdev*pihat;
run;
quit;
```



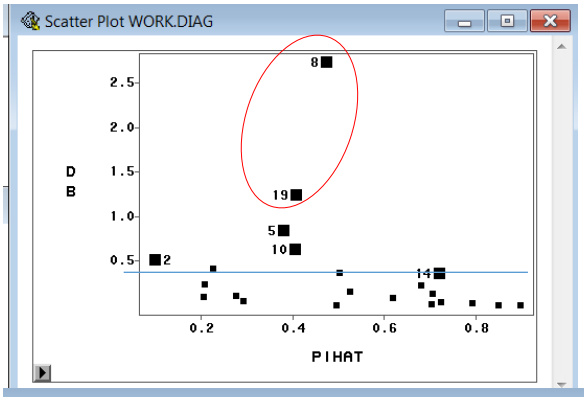
WORK.DIAG

	RACE2	SMOKE	HT	PTL2	LWT3	_TYPE_	_FREQ_	m_j	y_j	PIHAT	H	DB	DIFDEV	DIFCHISQ
1	0	0	0	0	0	0	2	2	1	0.2054	0.0796	0.0999	0.9448	1.1555
2	0	0	0	0	1	0	39	39	3	0.0988	0.5344	0.5178	0.4668	0.4512
3	0	0	0	1	1	0	2	2	0	0.2759	0.1190	0.1168	1.3941	0.8649
4	0	0	1	0	1	0	1	1	0	0.2935	0.1017	0.0523	0.7418	0.4624
5	0	1	0	0	0	0	11	11	3	0.3811	0.4561	0.8442	1.0331	1.0067
6	0	1	0	0	1	0	25	25	6	0.2071	0.4448	0.2377	0.2907	0.2966
7	0	1	0	1	0	0	4	4	2	0.6814	0.2292	0.2339	0.7451	0.7870
8	0	1	0	1	1	0	8	8	6	0.4758	0.4047	2.7541	4.1396	4.0520
9	0	1	1	0	1	0	4	4	2	0.4973	0.4389	0.0002	0.0002	0.0002
10	1	0	0	0	0	0	12	12	6	0.4077	0.4499	0.6295	0.7623	0.7696
11	1	0	0	0	1	0	45	45	11	0.2260	0.6360	0.4190	0.2383	0.2399
12	1	0	0	1	0	0	4	4	3	0.7051	0.2404	0.0161	0.0522	0.0510
13	1	0	0	1	1	0	4	4	1	0.5036	0.2209	0.3745	1.3674	1.3210
14	1	0	1	0	0	0	1	1	0	0.7228	0.1093	0.3590	2.8856	2.9269
15	1	0	1	0	1	0	3	3	2	0.5252	0.3147	0.1614	0.3572	0.3514
16	1	0	1	1	0	0	1	1	1	0.9006	0.0643	0.0081	0.2170	0.1180
17	1	0	1	1	1	0	1	1	1	0.7935	0.0984	0.0315	0.4910	0.2886
18	1	1	0	0	0	0	4	4	3	0.6212	0.2065	0.0926	0.3731	0.3556
19	1	1	0	0	1	0	11	11	3	0.4102	0.4457	1.2469	1.5949	1.5509
20	1	1	0	1	0	0	1	1	1	0.8507	0.0386	0.0073	0.3305	0.1826
21	1	1	0	1	1	0	5	5	3	0.7073	0.2671	0.1383	0.3636	0.3795
22	1	1	1	0	1	0	1	1	1	0.7249	0.1000	0.0469	0.6857	0.4217

Figure 19 – difdev vs pihat

Covariate pattern outliers:
8
14

```
proc insight data=diag;
 scatter db*pihat;
run;
quit;
```



WORK.DIAG

	RACE2	SMOKE	HT	PTL2	LWT3	_TYPE_	_FREQ_	m_j	y_j	PIHAT	H	DB	DIFDEV	DIFCHISQ
1	0	0	0	0	0	0	2	2	1	0.2054	0.0796	0.0999	0.9448	1.1555
2	0	0	0	0	1	0	39	39	3	0.0988	0.5344	0.5178	0.4668	0.4512
3	0	0	0	1	1	0	2	2	0	0.2759	0.1190	0.1168	1.3941	0.8649
4	0	0	1	0	1	0	1	1	0	0.2935	0.1017	0.0523	0.7418	0.4624
5	0	1	0	0	0	0	11	11	3	0.3811	0.4561	0.8442	1.0331	1.0067
6	0	1	0	0	1	0	25	25	6	0.2071	0.4448	0.2377	0.2907	0.2966
7	0	1	0	1	0	0	4	4	2	0.6814	0.2292	0.2339	0.7451	0.7870
8	0	1	0	1	1	0	8	8	6	0.4758	0.4047	2.7541	4.1396	4.0520
9	0	1	1	0	1	0	4	4	2	0.4973	0.4389	0.0002	0.0002	0.0002
10	1	0	0	0	0	0	12	12	6	0.4077	0.4499	0.6295	0.7623	0.7696
11	1	0	0	0	1	0	45	45	11	0.2260	0.6360	0.4190	0.2383	0.2399
12	1	0	0	1	0	0	4	4	3	0.7051	0.2404	0.0161	0.0522	0.0510
13	1	0	0	1	1	0	4	4	1	0.5036	0.2209	0.3745	1.3674	1.3210
14	1	0	1	0	0	0	1	1	0	0.7228	0.1093	0.3590	2.8856	2.9269
15	1	0	1	0	1	0	3	3	2	0.5252	0.3147	0.1614	0.3572	0.3514
16	1	0	1	1	0	0	1	1	1	0.9006	0.0643	0.0081	0.2170	0.1180
17	1	0	1	1	1	0	1	1	1	0.7935	0.0984	0.0315	0.4910	0.2886
18	1	1	0	0	0	0	4	4	3	0.6212	0.2065	0.0926	0.3731	0.3556
19	1	1	0	0	1	0	11	11	3	0.4102	0.4457	1.2469	1.5949	1.5509
20	1	1	0	1	0	0	1	1	1	0.8507	0.0386	0.0073	0.3305	0.1826
21	1	1	0	1	1	0	5	5	3	0.7073	0.2671	0.1383	0.3636	0.3795
22	1	1	1	0	1	0	1	1	1	0.7249	0.1000	0.0469	0.6857	0.4217

Figure 20 – db vs pihat (this is the most important plot!)

Covariate pattern outliers:
8
19
I found 14, but it's not a huge outlier, I can play with it to see it's removal effect

So with 2 of the logistic regression diagnostics, the same covariate patterns emerge;
8
14
I'm going to play with those and 4 others

```
* List extreme values *;
proc univariate data=diag;
 var h difchisq difdev db;
run;
```

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.000159905	9	0.517783	2
0.007325683	20	0.629462	10
0.008115108	16	0.844205	5
0.016125752	12	1.246935	19
0.031487896	17	2.754095	8

Figure 21 – Extreme values

```
*Create list of outliers - EVALUATE OUTLIERS;
proc sort data=diag;
        by race2 smoke ht ptl2 lwt3;
run;

data all;
        merge lowbwt2 diag;
                by race2 smoke ht ptl2 lwt3;
run;
```

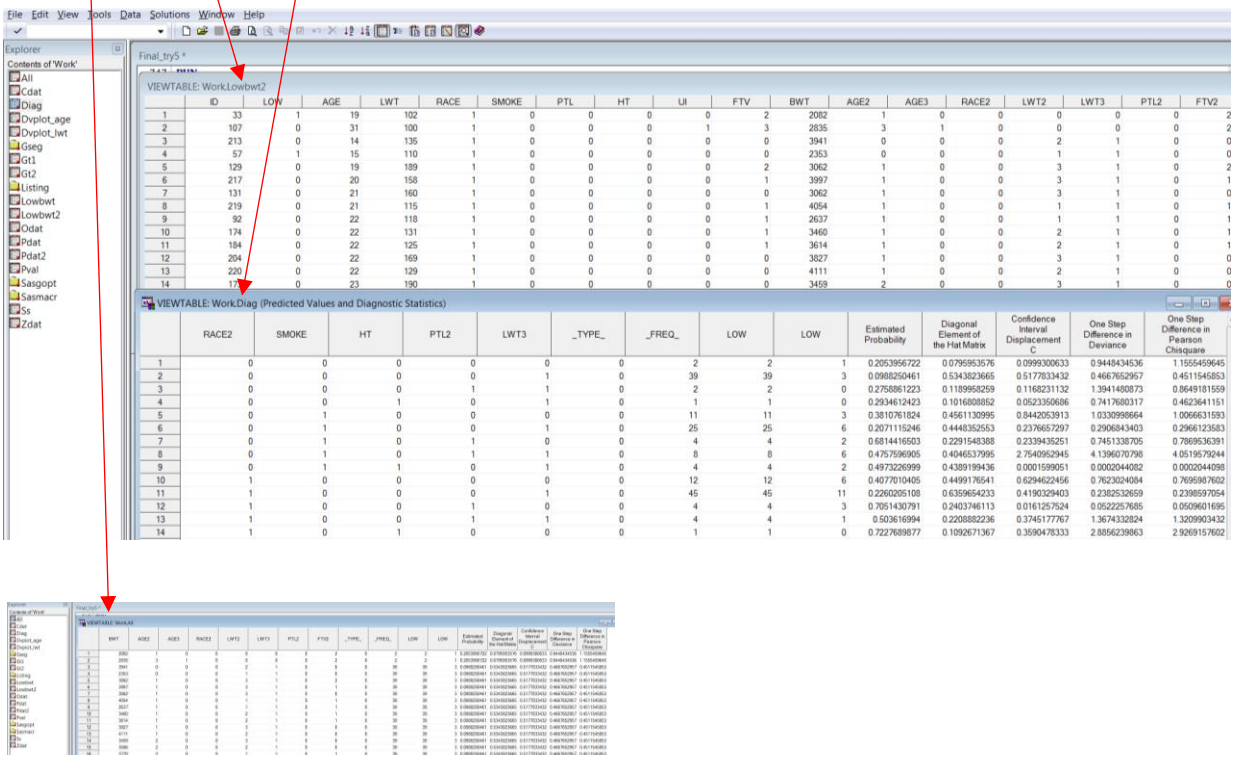
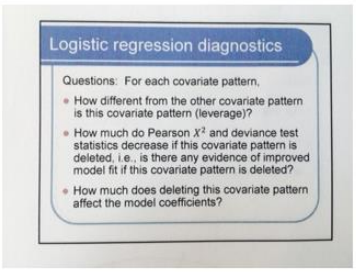
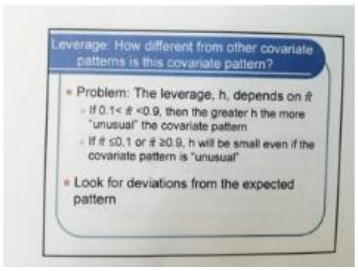
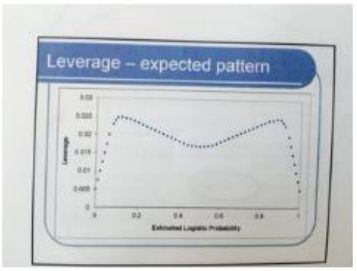
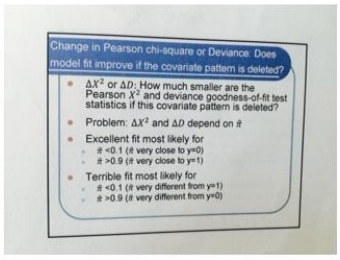


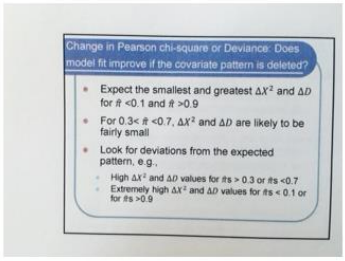
Figure 22 – How SAS is merging tables (this is just for me to learn which kind of merge SAS is doing)

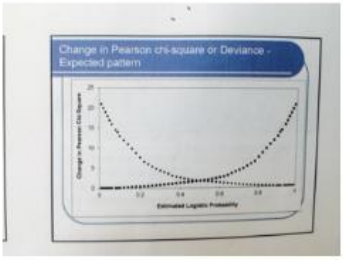


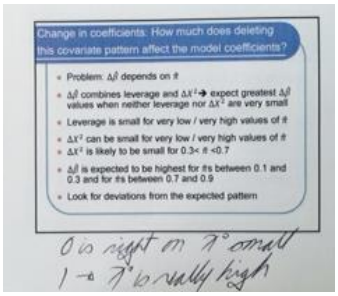


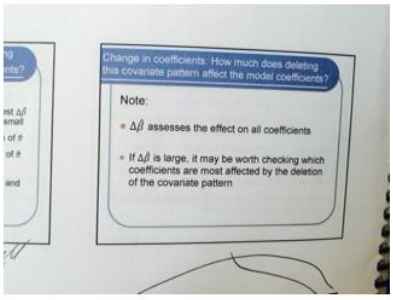


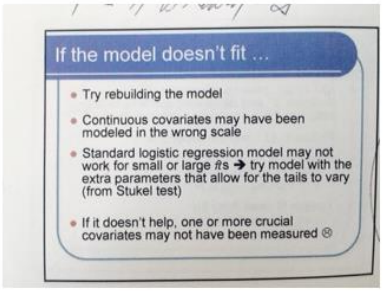






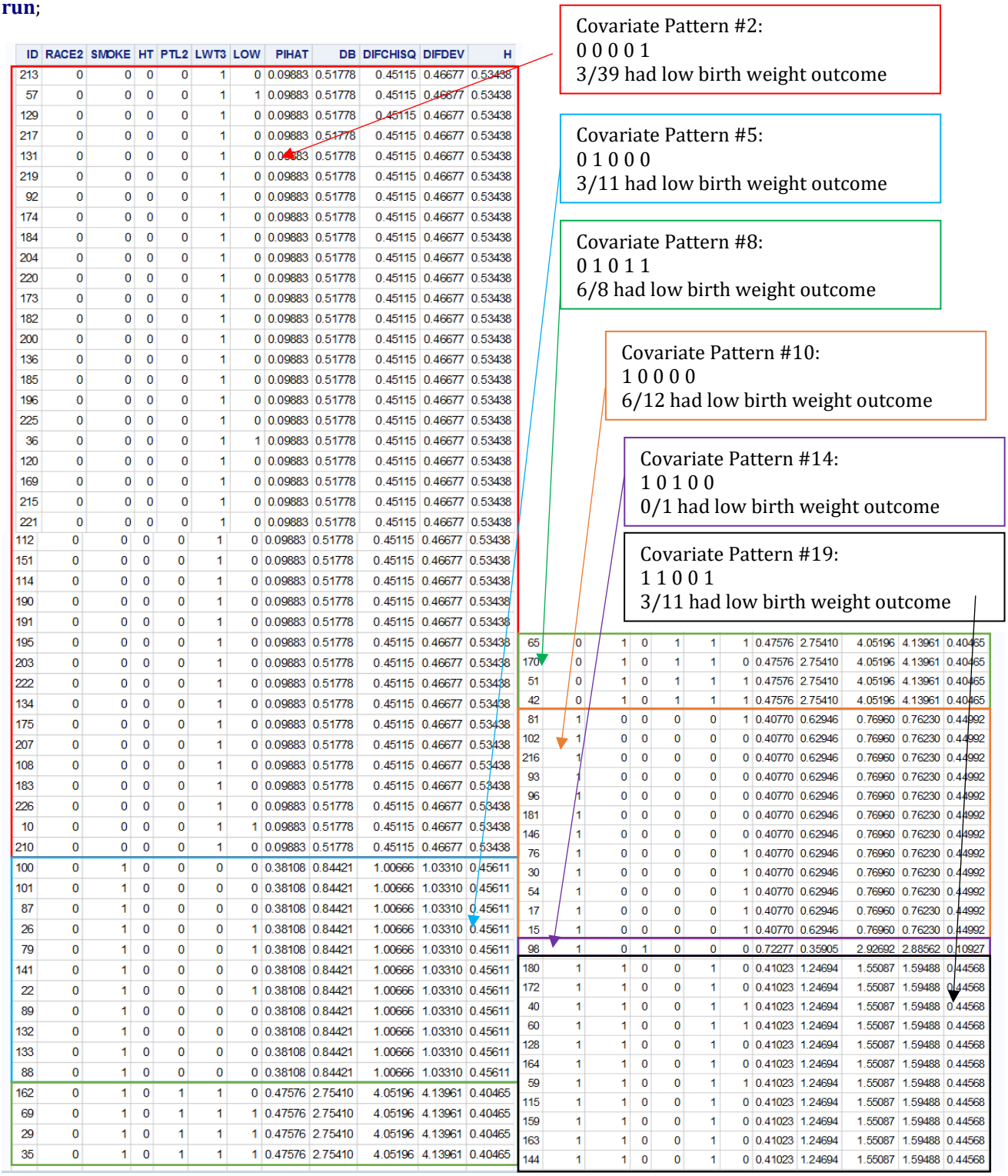






- m_j = number of observations with covariate pattern j
- y_j = number of observations with low=1 in covariate pattern j

```
*change db and difchisq based on graphs;
proc print data=all noobs;
        where db>0.5 or difchisq>2.0;
                var id race2 smoke ht ptl2 lwt3 low pihat db difchisq difdev h;
run;
```



ID	RACE2	SMOKE	HT	PTL2	LWT3	LOW	PIHAT	DB	DIFCHISQ	DIFDEV	H
213	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
57	0	0	0	0	1	1	0.09883	0.51778	0.45115	0.46677	0.53438
129	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
217	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
131	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
219	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
92	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
174	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
184	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
204	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
220	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
173	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
182	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
200	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
136	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
185	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
196	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
225	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
36	0	0	0	0	1	1	0.09883	0.51778	0.45115	0.46677	0.53438
120	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
169	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
215	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
221	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
112	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
151	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
114	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
190	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
191	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
195	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
203	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
222	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
134	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
175	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
207	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
108	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
183	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
226	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
10	0	0	0	0	1	1	0.09883	0.51778	0.45115	0.46677	0.53438
210	0	0	0	0	1	0	0.09883	0.51778	0.45115	0.46677	0.53438
100	0	1	0	0	0	0	0.38108	0.84421	1.00666	1.03310	0.45611
101	0	1	0	0	0	0	0.38108	0.84421	1.00666	1.03310	0.45611
87	0	1	0	0	0	0	0.38108	0.84421	1.00666	1.03310	0.45611
26	0	1	0	0	0	1	0.38108	0.84421	1.00666	1.03310	0.45611
79	0	1	0	0	0	1	0.38108	0.84421	1.00666	1.03310	0.45611
141	0	1	0	0	0	0	0.38108	0.84421	1.00666	1.03310	0.45611
22	0	1	0	0	0	1	0.38108	0.84421	1.00666	1.03310	0.45611
89	0	1	0	0	0	0	0.38108	0.84421	1.00666	1.03310	0.45611
132	0	1	0	0	0	0	0.38108	0.84421	1.00666	1.03310	0.45611
133	0	1	0	0	0	0	0.38108	0.84421	1.00666	1.03310	0.45611
88	0	1	0	0	0	0	0.38108	0.84421	1.00666	1.03310	0.45611
162	0	1	0	1	1	0	0.47576	2.75410	4.05196	4.13961	0.40465
69	0	1	0	1	1	1	0.47576	2.75410	4.05196	4.13961	0.40465
29	0	1	0	1	1	1	0.47576	2.75410	4.05196	4.13961	0.40465
35	0	1	0	1	1	1	0.47576	2.75410	4.05196	4.13961	0.40465
65	0	1	0	1	1	1	0.47576	2.75410	4.05196	4.13961	0.40465
170	0	1	0	1	1	0	0.47576	2.75410	4.05196	4.13961	0.40465
51	0	1	0	1	1	1	0.47576	2.75410	4.05196	4.13961	0.40465
42	0	1	0	1	1	1	0.47576	2.75410	4.05196	4.13961	0.40465
81	1	0	0	0	0	1	0.40770	0.62946	0.76960	0.76230	0.44992
102	1	0	0	0	0	0	0.40770	0.62946	0.76960	0.76230	0.44992
216	1	0	0	0	0	0	0.40770	0.62946	0.76960	0.76230	0.44992
93	1	0	0	0	0	0	0.40770	0.62946	0.76960	0.76230	0.44992
96	1	0	0	0	0	0	0.40770	0.62946	0.76960	0.76230	0.44992
181	1	0	0	0	0	0	0.40770	0.62946	0.76960	0.76230	0.44992
146	1	0	0	0	0	0	0.40770	0.62946	0.76960	0.76230	0.44992
76	1	0	0	0	0	1	0.40770	0.62946	0.76960	0.76230	0.44992
30	1	0	0	0	0	1	0.40770	0.62946	0.76960	0.76230	0.44992
54	1	0	0	0	0	1	0.40770	0.62946	0.76960	0.76230	0.44992
17	1	0	0	0	0	1	0.40770	0.62946	0.76960	0.76230	0.44992
15	1	0	0	0	0	1	0.40770	0.62946	0.76960	0.76230	0.44992
98	1	0	1	0	0	0	0.72277	0.35905	2.92692	2.88562	0.10927
180	1	1	0	0	1	0	0.41023	1.24694	1.55087	1.59488	0.44568
172	1	1	0	0	1	0	0.41023	1.24694	1.55087	1.59488	0.44568
40	1	1	0	0	1	1	0.41023	1.24694	1.55087	1.59488	0.44568
60	1	1	0	0	1	1	0.41023	1.24694	1.55087	1.59488	0.44568
128	1	1	0	0	1	0	0.41023	1.24694	1.55087	1.59488	0.44568
164	1	1	0	0	1	0	0.41023	1.24694	1.55087	1.59488	0.44568
59	1	1	0	0	1	1	0.41023	1.24694	1.55087	1.59488	0.44568
115	1	1	0	0	1	0	0.41023	1.24694	1.55087	1.59488	0.44568
159	1	1	0	0	1	0	0.41023	1.24694	1.55087	1.59488	0.44568
163	1	1	0	0	1	0	0.41023	1.24694	1.55087	1.59488	0.44568
144	1	1	0	0	1	0	0.41023	1.24694	1.55087	1.59488	0.44568

Figure 22 – List of outlier covariate patterns (6 patterns)

None of the covariate values are unreasonable.
List of six covariate patterns:

Covariate Pattern #2:
0 0 0 0 1
3/39 had low birth weight outcome

#2:	RACE2	Race=0	black+other race
	SMOKE	Smoke=0	no smoking during pregnancy
	HT	Hypertension=0	no history of hypertension
	PTL2	Premature labor=0	0 prior premature labors
	LWT3	weight at LMP=1	110-251 pounds

Pihat 0.09883 0.1<pihat<0.9, the greater h, the more unusual the covariate pattern

Risk profile:
low risk profile due to black race, no smoking, no hypertension, and no prior premature labors, low pihat, 3/39 had outcome LOW (low birth weight infants)

Covariate Pattern #5:
0 1 0 0 0
3/11 had low birth weight outcome

#5:	RACE2	Race=0	black+other race
	SMOKE	Smoke=1	smoking during pregnancy
	HT	Hypertension=0	no history of hypertension
	PTL2	Premature labor=0	0 prior premature labors
	LWT3	weight at LMP=0	80-109 pounds

Pihat 0.38108 0.1<pihat<0.9, the greater h, the more unusual the covariate pattern

Risk profile:
lower risk profile due to black race, smoking, no hypertension, and no prior premature labors, low maternal weight, medium pihat, 3/11 had outcome LOW (low birth weight infants)

Covariate Pattern #8:
0 1 0 1 1
6/8 had low birth weight outcome

#8:	RACE2	Race=0	black+other race
	SMOKE	Smoke=1	smoking during pregnancy
	HT	Hypertension=0	no history of hypertension
	PTL2	Premature labor=1	≥2 prior premature labors
	LWT3	weight at LMP=1	110-251 pounds

Pihat 0.47576 0.1<pihat<0.9, the greater h, the more unusual the covariate pattern

Risk profile:
lower risk profile due to black race, smoking, no hypertension, ≥2 prior premature labors, medium pihat, 6/8 had outcome LOW (low birth weight infants)

Covariate Pattern #10:
1 0 0 0 0
6/12 had low birth weight outcome

#10:	RACE2	Race=1	not white, not other race
	SMOKE	Smoke=0	no smoking during pregnancy
	HT	Hypertension=0	no history of hypertension
	PTL2	Premature labor=0	0 prior premature labors
	LWT3	weight at LMP=0	80-109 pounds

Pihat 0.40770 0.1<pihat<0.9, the greater h, the more unusual the covariate pattern

Risk profile:
higher risk profile due to not white, not other race, smoking, no hypertension, no premature labors, low maternal rate, medium pihat, 6/12 had outcome LOW (low birth weight infants)

Covariate Pattern #14:
1 0 1 0 0
0/1 had low birth weight outcome

#14:	RACE2	Race=1	not white, not other race
	SMOKE	Smoke=0	no smoking during pregnancy
	HT	Hypertension=1	history of hypertension
	PTL2	Premature labor=0	0 prior premature labors
	LWT3	weight at LMP=0	80-109 pounds

Pihat 0.72277 0.1<pihat<0.9, the greater h, the more unusual the covariate pattern

Risk profile:
lower risk profile due to not white, not other race, no smoking, history of hypertension, no premature labors, low maternal weight, higher pihat, 0/1 had outcome LOW (low birth weight infants)

Covariate Pattern #19:
1 1 0 0 1
3/11 had low birth weight outcome

#19:	RACE2	Race=1	not white, not other race
	SMOKE	Smoke=1	smoking during pregnancy
	HT	Hypertension=0	no history of hypertension
	PTL2	Premature labor=0	0 prior premature labors
	LWT3	weight at LMP=1	110-251 pounds

Pihat 0.41023 0.1<pihat<0.9, the greater h, the more unusual the covariate pattern

Risk profile:
lower risk profile due to not white, not other race, smoking, no history of hypertension, no premature labors, regular maternal weight, medium pihat, 3/11 had outcome LOW (low birth weight infants)

```
* Print data for outliers *;
data diag1;
 set diag;
 if _N_ in (2,5,8,10,14,19);
 cov_pat=_N_;
run;

proc print noobs data=diag1;
 var cov_pat RACE2 SMOKE HT PTL2 LWT3 m_j y_j PIHAT DB DIFCHISQ DIFDEV H;
run;
```

- m_j = number of observations with covariate pattern j
- y_j = number of observations with low=1 in covariate pattern j
- m = predicted
- y = how many had the outcome

cov_pat	RACE2	SMOKE	HT	PTL2	LWT3	m_j	y_j	PIHAT	DB	DIFCHISQ	DIFDEV	H
2	0	0	0	0	1	39	3	0.09883	0.51778	0.45115	0.46677	0.53438
5	0	1	0	0	0	11	3	0.38108	0.84421	1.00666	1.03310	0.45611
8	0	1	0	1	1	8	6	0.47576	2.75410	4.05196	4.13961	0.40465
10	1	0	0	0	0	12	6	0.40770	0.62946	0.76960	0.76230	0.44992
14	1	0	1	0	0	1	0	0.72277	0.35905	2.92692	2.88562	0.10927
19	1	1	0	0	1	11	3	0.41023	1.24694	1.55087	1.59488	0.44568

Figure 23 – List of covariate patterns

```
************DELETING OUTLIERS****************;
*Covariate patterns: 2,5,8,10,14,19;
*Model before deletion;
proc logistic descending data=lowbwt2;
        model low=race2 smoke ht ptl2 lwt3/scale=n aggregate lackfit;
run;


****  osius-rojek g-o-f test ******;
proc sort data=lowbwt2;
        by race2 smoke ht ptl2 lwt3;
run;
proc means n sum noprint data=lowbwt2;
        by race2 smoke ht ptl2 lwt3;
        var low;
        output out=odat n=m_j sum=y_j;
run;
proc logistic descending data=odat;
 model y_j/m_j=race2 smoke ht ptl2 lwt3;
 output out=pdat p=p_j;
run;

data pdat;
 set pdat;

 v_j=m_j*p_j*(1-p_j);
 c_j=(1-2*p_j)/v_j;
```

```
 chisq_j=(y_j-m_j*p_j)**2/v_j;

 m_j_inv=1/m_j;
run;

proc means sum noprint data=pdat;
 var chisq_j m_j_inv;
 output out=cdat sum=chisq m_inv;
run;

proc reg data=pdat outest=ss;
 model c_j=RACE2 SMOKE HT PTL2 LWT3;
 weight v_j;
run;

data zdat;
 merge cdat (keep=_freq_ chisq m_inv) ss (keep=_rmse_);
 *MUST BE CHANGED TO ACTUAL NUMBER OF VARIABLES IN MODEL;
 ************_*****************;
 rss=(_freq_-5-1)*_rmse_**2;

 A=2*(_freq_-m_inv);
 *MUST BE CHANGED TO ACTUAL NUMBER OF VARIABLES IN MODEL;
 ****************_*****************;
 z=(chisq-(_freq_-5-1))/sqrt(A+rss);
 z=abs(z);
 pval=(1-probnorm(z))*2;
run;

proc print noobs data=zdat;
run;

***** stukel test of logistic regression model assumption ***;
*change outcome and independent variable names in model statement!;
ods output globaltests=gt1;
proc logistic descending data=all;
 model low=race2 smoke ht ptl2 lwt3;
 output out=pdat2 xbeta=g_j p=p_j;
run;

data pdat2;
 set pdat2;
        if p_j>=0.5 then ind1=1;
                else ind1=0;
        if p_j<0.5 then ind2=1;
                else ind2=0;
                z1_j=0.5*g_j**2*ind1;
                z2_j=-0.5*g_j**2*ind2;
run;

*change outcome and independent variable names!;
ods output globaltests=gt2;
proc logistic descending data=pdat2;
 model low=race2 smoke ht ptl2 lwt3 z1_j z2_j;
run;
```

```
data pval;
 merge gt1(rename=(chisq=chisq1))
          gt2(rename=(chisq=chisq2));
                    if _n_=1;
 drop test df probchisq;
                    lr=chisq2-chisq1;
                    pval_lr=(1-probchi(lr,2));
run;

proc print noobs data=pval;
run;


*dataset ALL*;
```

Covariate pattern #2:
(213,57,129,217,131,219,92,174,184,204,220,173,182,200,136,185,196,225,36,120,169,215,221,112,151,
114,190,191,195,203,222,134,175,207,108,226,10,210)

```
**********DELETE Covariate pattern #2**********;
*I did this a different way than you did, but it worked :)   ;
data all_2;
          set all;
if id not in
(213,57,129,217,131,219,92,174,184,204,220,173,182,200,136,185,196,225,36,120,169,215,221,112,1
51,114,190,191,195,203,222,134,175,207,108,226,10,210)
          then output all_2;
run;
```

```
NOTE: There were 189 observations read from the data set WORK.ALL.
NOTE: The data set WORK.ALL_2 has 151 observations and 27 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      cpu time            0.01 seconds
```

VIEWTABLE: Work.All_2

	ID	LOW	AGE	LWT	RACE	SMOKE	PTL	HT	UI	FTV	BWT	AGE2	AGE3	RACE2	LWT2
131	82	1	23	94	3	1	0	0	0	0	2495	2	0	1	
132	61	1	24	105	2	1	0	0	0	0	2381	2	0	1	
133	44	1	20	80	3	1	0	0	1	0	2211	1	0	1	
134	180	0	17	120	3	1	0	0	0	0	3572	0	0	1	
135	172	0	20	121	2	1	0	0	0	0	3444	1	0	1	
136	40	1	20	120	2	1	0	0	0	3	2126	1	0	1	
137	60	1	20	122	2	1	0	0	0	0	2381	1	0	1	
138	128	0	21	185	2	1	0	0	0	2	3042	1	0	1	
139	164	0	23	115	3	1	0	0	0	1	3331	2	0	1	
140	59	1	23	187	2	1	0	0	0	1	2367	2	0	1	
141	115	0	26	168	2	1	0	0	0	0	2920	3	0	1	
142	159	0	28	250	3	1	0	0	0	6	3303	3	0	1	
143	163	0	31	150	3	1	0	0	0	2	3321	3	1	1	
144	144	0	21	110	3	1	0	0	1	0	3203	1	0	1	
145	78	1	14	101	3	1	1	0	0	0	2466	0	0	1	
146	50	1	18	110	2	1	1	0	0	0	2296	0	0	1	
147	154	0	26	133	3	1	2	0	0	0	3260	3	0	1	
148	119	0	35	121	2	1	1	0	0	1	2948	3	1	1	
149	37	1	17	130	3	1	1	0	1	0	2125	0	0	1	
150	4	1	28	120	3	1	1	0	1	0	709	3	0	1	
151	11	1	34	187	2	1	0	1	0	0	1135	3	1	1	

I created a new dataset "ALL_2", which excludes the IDs identified in covariate pattern #2. I know this is a roundabout way to do this, but that's ok because it worked. I just have to be careful to draw from dataset "ALL" in my next covariate pattern deletion.

```
proc logistic descending data=all_2;
        model low=race2 smoke ht ptl2 lwt3/scale=n aggregate lackfit;
run;

****  osius-rojek g-o-f test covariate pattern #2******;
proc sort data=all_2;
        by race2 smoke ht ptl2 lwt3;
run;

proc means n sum noprint data=all_2;
        by race2 smoke ht ptl2 lwt3;
        var low;
        output out=odat n=m_j sum=y_j;
run;

proc logistic descending data=odat;
 model y_j/m_j=race2 smoke ht ptl2 lwt3;
 output out=pdat p=p_j;
run;

data pdat;
 set pdat;
                v_j=m_j*p_j*(1-p_j);
                c_j=(1-2*p_j)/v_j;
                chisq_j=(y_j-m_j*p_j)**2/v_j;
                m_j_inv=1/m_j;
```

```
run;

proc means sum noprint data=pdat;
 var chisq_j m_j_inv;
 output out=cdat sum=chisq m_inv;
run;

proc reg data=pdat outest=ss;
 model c_j=race2 smoke ht ptl2 lwt3;
 weight v_j;
run;

data zdat;
 merge cdat (keep=_freq_ chisq m_inv) ss (keep=_rmse_);
 *must be changed to actual number of variables in model;
 ***********_*****************;
 rss=(_freq_-5-1)*_rmse_**2;

 a=2*(_freq_-m_inv);
 *must be changed to actual number of variables in model;
 ***************_*****************;
 z=(chisq-(_freq_-5-1))/sqrt(a+rss);
 z=abs(z);
 pval=(1-probnorm(z))*2;
run;

proc print noobs data=zdat;
run;

***** stukel test of logistic regression model assumption for deletion of covariate pattern #2***;
*change outcome and independent variable names in model statement!;
ods output globaltests=gt1;
proc logistic descending data=all_2;
 model low=race2 smoke ht ptl2 lwt3;
 output out=pdat2 xbeta=g_j p=p_j;
run;

data pdat2;
 set pdat2;
        if p_j>=0.5 then ind1=1;
                else ind1=0;
        if p_j<0.5 then ind2=1;
                else ind2=0;
                z1_j=0.5*g_j**2*ind1;
                z2_j=-0.5*g_j**2*ind2;
run;

*change outcome and independent variable names!;
ods output globaltests=gt2;
proc logistic descending data=pdat2;
 model low=race2 smoke ht ptl2 lwt3 z1_j z2_j;
run;

data pval;
 merge gt1(rename=(chisq=chisq1))
```

```
        gt2(rename=(chisq=chisq2));
                if _n_=1;
 drop test df probchisq;
                lr=chisq2-chisq1;
                pval_lr=(1-probchi(lr,2));
run;

proc print noobs data=pval;
run;


covariate pattern #5:
(100,101,87,26,79,141,22,89,132,133,88)

**********covariate pattern #5**********:
*(100,101,87,26,79,141,22,89,132,133,88);
data all_5;
        set all;
if id not in (100,101,87,26,79,141,22,89,132,133,88)
        then output all_5;
run;

proc logistic descending data=all_5;
        model low=race2 smoke ht ptl2 lwt3/scale=n aggregate lackfit;
run;

****  osius-rojek g-o-f test covariate pattern #2*****;
proc sort data=all_5;
        by race2 smoke ht ptl2 lwt3;
run;

proc means n sum noprint data=all_5;
        by race2 smoke ht ptl2 lwt3;
        var low;
        output out=odat n=m_j sum=y_j;
run;

proc logistic descending data=odat;
 model y_j/m_j=race2 smoke ht ptl2 lwt3;
 output out=pdat p=p_j;
run;

data pdat;
 set pdat;
                v_j=m_j*p_j*(1-p_j);
                c_j=(1-2*p_j)/v_j;
                chisq_j=(y_j-m_j*p_j)**2/v_j;
                m_j_inv=1/m_j;
run;

proc means sum noprint data=pdat;
 var chisq_j m_j_inv;
 output out=cdat sum=chisq m_inv;
run;
```

```
proc reg data=pdat outest=ss;
 model c_j=race2 smoke ht ptl2 lwt3;
 weight v_j;
run;

data zdat;
 merge cdat (keep=_freq_ chisq m_inv) ss (keep=_rmse_);
 *must be changed to actual number of variables in model;
 ************_*****************;
 rss=(_freq_-5-1)*_rmse_**2;

 a=2*(_freq_-m_inv);
 *must be changed to actual number of variables in model;
 ****************_*****************;
 z=(chisq-(_freq_-5-1))/sqrt(a+rss);
 z=abs(z);
 pval=(1-probnorm(z))*2;
run;

proc print noobs data=zdat;
run;

***** stukel test of logistic regression model assumption for deletion of covariate pattern #5***;
*change outcome and independent variable names in model statement!;
ods output globaltests=gt1;
proc logistic descending data=all_5;
 model low=race2 smoke ht ptl2 lwt3;
 output out=pdat2 xbeta=g_j p=p_j;
run;

data pdat2;
 set pdat2;
        if p_j>=0.5 then ind1=1;
                else ind1=0;
        if p_j<0.5 then ind2=1;
                else ind2=0;
                z1_j=0.5*g_j**2*ind1;
                z2_j=-0.5*g_j**2*ind2;
run;

*change outcome and independent variable names!;
ods output globaltests=gt2;
proc logistic descending data=pdat2;
 model low=race2 smoke ht ptl2 lwt3 z1_j z2_j;
run;

data pval;
 merge gt1(rename=(chisq=chisq1))
        gt2(rename=(chisq=chisq2));
                if _n_=1;
 drop test df probchisq;
                lr=chisq2-chisq1;
                pval_lr=(1-probchi(lr,2));
run;
```

```
proc print noobs data=pval;
run;


covariate pattern #8:
(162,69,29,35,65,170,51,42)

**********covariate pattern #8**********:
(162,69,29,35,65,170,51,42);

data all_8;
        set all;
if id not in (162,69,29,35,65,170,51,42)
        then output all_8;
run;
NOTE: There were 189 observations read from the data set WORK.ALL.
NOTE: The data set WORK.ALL_8 has 181 observations and 27 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      cpu time            0.01 seconds
proc logistic descending data=all_8;
        model low=race2 smoke ht ptl2 lwt3/scale=n aggregate lackfit;
run;

****  osius-rojek g-o-f test covariate pattern #2******;
proc sort data=all_8;
        by race2 smoke ht ptl2 lwt3;
run;

proc means n sum noprint data=all_8;
        by race2 smoke ht ptl2 lwt3;
        var low;
        output out=odat n=m_j sum=y_j;
run;

proc logistic descending data=odat;
 model y_j/m_j=race2 smoke ht ptl2 lwt3;
 output out=pdat p=p_j;
run;

data pdat;
 set pdat;
                v_j=m_j*p_j*(1-p_j);
                c_j=(1-2*p_j)/v_j;
                chisq_j=(y_j-m_j*p_j)**2/v_j;
                m_j_inv=1/m_j;
run;

proc means sum noprint data=pdat;
 var chisq_j m_j_inv;
 output out=cdat sum=chisq m_inv;
run;

proc reg data=pdat outest=ss;
 model c_j=race2 smoke ht ptl2 lwt3;
```

```
 weight v_j;
run;

data zdat;
 merge cdat (keep=_freq_ chisq m_inv) ss (keep=_rmse_);
 *must be changed to actual number of variables in model;
 ***********_*****************;
 rss=(_freq_-5-1)*_rmse_**2;

 a=2*(_freq_-m_inv);
 *must be changed to actual number of variables in model;
 ****************_*****************;
 z=(chisq-(_freq_-5-1))/sqrt(a+rss);
 z=abs(z);
 pval=(1-probnorm(z))*2;
run;

proc print noobs data=zdat;
run;

***** stukel test of logistic regression model assumption for deletion of covariate pattern #8***;
*change outcome and independent variable names in model statement!;
ods output globaltests=gt1;
proc logistic descending data=all_8;
 model low=race2 smoke ht ptl2 lwt3;
 output out=pdat2 xbeta=g_j p=p_j;
run;

data pdat2;
 set pdat2;
        if p_j>=0.5 then ind1=1;
                else ind1=0;
        if p_j<0.5 then ind2=1;
                else ind2=0;
                z1_j=0.5*g_j**2*ind1;
                z2_j=-0.5*g_j**2*ind2;
run;

*change outcome and independent variable names!;
ods output globaltests=gt2;
proc logistic descending data=pdat2;
 model low=race2 smoke ht ptl2 lwt3 z1_j z2_j;
run;

data pval;
 merge gt1(rename=(chisq=chisq1))
        gt2(rename=(chisq=chisq2));
                if _n_=1;
 drop test df probchisq;
                lr=chisq2-chisq1;
                pval_lr=(1-probchi(lr,2));
run;

proc print noobs data=pval;
run;
```

covariate pattern #10:
(81,102,216,93,96,181,146,76,30,54,17,15)

```
note: there were 189 observations read from the data set work.all.
note: the data set work.all_10 has 177 observations and 27 variables.
note: data statement used (total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
```

```
**************************covariate pattern #10*****************:
(81,102,216,93,96,181,146,76,30,54,17,15);

data all_10;
        set all;
if id not in (81,102,216,93,96,181,146,76,30,54,17,15)
        then output all_10;
run;

proc logistic descending data=all_10;
        model low=race2 smoke ht ptl2 lwt3/scale=n aggregate lackfit;
run;

****  osius-rojek g-o-f test covariate pattern #10******;
proc sort data=all_10;
        by race2 smoke ht ptl2 lwt3;
run;

proc means n sum noprint data=all_10;
        by race2 smoke ht ptl2 lwt3;
        var low;
        output out=odat n=m_j sum=y_j;
run;

proc logistic descending data=odat;
 model y_j/m_j=race2 smoke ht ptl2 lwt3;
 output out=pdat p=p_j;
run;

data pdat;
 set pdat;
                v_j=m_j*p_j*(1-p_j);
                c_j=(1-2*p_j)/v_j;
                chisq_j=(y_j-m_j*p_j)**2/v_j;
                m_j_inv=1/m_j;
run;

proc means sum noprint data=pdat;
 var chisq_j m_j_inv;
 output out=cdat sum=chisq m_inv;
run;

proc reg data=pdat outest=ss;
```

```
 model c_j=race2 smoke ht ptl2 lwt3;
 weight v_j;
run;

data zdat;
 merge cdat (keep=_freq_ chisq m_inv) ss (keep=_rmse_);
 *must be changed to actual number of variables in model;
 ***********_*****************;
 rss=(_freq_-5-1)*_rmse_**2;

 a=2*(_freq_-m_inv);
 *must be changed to actual number of variables in model;
 ****************_*****************;
 z=(chisq-(_freq_-5-1))/sqrt(a+rss);
 z=abs(z);
 pval=(1-probnorm(z))*2;
run;

proc print noobs data=zdat;
run;

***** stukel test of logistic regression model assumption for deletion of covariate pattern #10***;
*change outcome and independent variable names in model statement!;
ods output globaltests=gt1;
proc logistic descending data=all_10;
 model low=race2 smoke ht ptl2 lwt3;
 output out=pdat2 xbeta=g_j p=p_j;
run;

data pdat2;
 set pdat2;
        if p_j>=0.5 then ind1=1;
                else ind1=0;
        if p_j<0.5 then ind2=1;
                else ind2=0;
                z1_j=0.5*g_j**2*ind1;
                z2_j=-0.5*g_j**2*ind2;
run;

*change outcome and independent variable names!;
ods output globaltests=gt2;
proc logistic descending data=pdat2;
 model low=race2 smoke ht ptl2 lwt3 z1_j z2_j;
run;

data pval;
 merge gt1(rename=(chisq=chisq1))
        gt2(rename=(chisq=chisq2));
                if _n_=1;
 drop test df probchisq;
                lr=chisq2-chisq1;
                pval_lr=(1-probchi(lr,2));
run;

proc print noobs data=pval;
```

```
run;



covariate pattern #14:
(98)

**************************covariate pattern #14*****************:
*covariate pattern #14:
(98);

data all_14;
        set all;
if id not in (98)
        then output all_14;
run;


note: there were 189 observations read from the data set work.all.
note: the data set work.all_14 has 188 observations and 27 variables.
note: data statement used (total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds

proc logistic descending data=all_14;
        model low=race2 smoke ht ptl2 lwt3/scale=n aggregate lackfit;
run;

****  osius-rojek g-o-f test covariate pattern #14******;
proc sort data=all_14;
        by race2 smoke ht ptl2 lwt3;
run;

proc means n sum noprint data=all_14;
        by race2 smoke ht ptl2 lwt3;
        var low;
        output out=odat n=m_j sum=y_j;
run;

proc logistic descending data=odat;
 model y_j/m_j=race2 smoke ht ptl2 lwt3;
 output out=pdat p=p_j;
run;

data pdat;
 set pdat;
                v_j=m_j*p_j*(1-p_j);
                c_j=(1-2*p_j)/v_j;
                chisq_j=(y_j-m_j*p_j)**2/v_j;
                m_j_inv=1/m_j;
run;

proc means sum noprint data=pdat;
 var chisq_j m_j_inv;
```

```
 output out=cdat sum=chisq m_inv;
run;

proc reg data=pdat outest=ss;
 model c_j=race2 smoke ht ptl2 lwt3;
 weight v_j;
run;

data zdat;
 merge cdat (keep=_freq_ chisq m_inv) ss (keep=_rmse_);
 *must be changed to actual number of variables in model;
 ************_*****************;
 rss=(_freq_-5-1)*_rmse_**2;

 a=2*(_freq_-m_inv);
 *must be changed to actual number of variables in model;
 ****************_*****************;
 z=(chisq-(_freq_-5-1))/sqrt(a+rss);
 z=abs(z);
 pval=(1-probnorm(z))*2;
run;

proc print noobs data=zdat;
run;

***** stukel test of logistic regression model assumption for deletion of covariate pattern #14***;
*change outcome and independent variable names in model statement!;
ods output globaltests=gt1;
proc logistic descending data=all_14;
 model low=race2 smoke ht ptl2 lwt3;
 output out=pdat2 xbeta=g_j p=p_j;
run;

data pdat2;
 set pdat2;
        if p_j>=0.5 then ind1=1;
                else ind1=0;
        if p_j<0.5 then ind2=1;
                else ind2=0;
                z1_j=0.5*g_j**2*ind1;
                z2_j=-0.5*g_j**2*ind2;
run;

*change outcome and independent variable names!;
ods output globaltests=gt2;
proc logistic descending data=pdat2;
 model low=race2 smoke ht ptl2 lwt3 z1_j z2_j;
run;

data pval;
 merge gt1(rename=(chisq=chisq1))
        gt2(rename=(chisq=chisq2));
                if _n_=1;
 drop test df probchisq;
                lr=chisq2-chisq1;
```

```
                    pval_lr=(1-probchi(lr,2));
run;

proc print noobs data=pval;
run;
```

covariate pattern #19:
(108,172,40,60,128,164,59,115,159,163,144)

```
**************************covariate pattern #19*****************:
*covariate pattern #19:
(108,172,40,60,128,164,59,115,159,163,144);

data all_19;
        set all;
if id not in (108,172,40,60,128,164,59,115,159,163,144)
        then output all_19;
run;
```

```
note: there were 189 observations read from the data set work.all.
note: the data set work.all_19 has 178 observations and 27 variables.
note: data statement used (total process time):
      real time           0.00 seconds
      cpu time            0.01 seconds
```

```
proc logistic descending data=all_19;
        model low=race2 smoke ht ptl2 lwt3/scale=n aggregate lackfit;
run;


****  osius-rojek g-o-f test covariate pattern #19******;
proc sort data=all_19;
        by race2 smoke ht ptl2 lwt3;
run;

proc means n sum noprint data=all_19;
        by race2 smoke ht ptl2 lwt3;
        var low;
        output out=odat n=m_j sum=y_j;
run;

proc logistic descending data=odat;
 model y_j/m_j=race2 smoke ht ptl2 lwt3;
 output out=pdat p=p_j;
run;

data pdat;
 set pdat;
```

```
                    v_j=m_j*p_j*(1-p_j);
                    c_j=(1-2*p_j)/v_j;
                    chisq_j=(y_j-m_j*p_j)**2/v_j;
                    m_j_inv=1/m_j;
run;

proc means sum noprint data=pdat;
 var chisq_j m_j_inv;
 output out=cdat sum=chisq m_inv;
run;

proc reg data=pdat outest=ss;
 model c_j=race2 smoke ht ptl2 lwt3;
 weight v_j;
run;

data zdat;
 merge cdat (keep=_freq_ chisq m_inv) ss (keep=_rmse_);
 *must be changed to actual number of variables in model;
 ************_*****************;
 rss=(_freq_-5-1)*_rmse_**2;

 a=2*(_freq_-m_inv);
 *must be changed to actual number of variables in model;
 ****************_*****************;
 z=(chisq-(_freq_-5-1))/sqrt(a+rss);
 z=abs(z);
 pval=(1-probnorm(z))*2;
run;

proc print noobs data=zdat;
run;

***** stukel test of logistic regression model assumption for deletion of covariate pattern #19***;
*change outcome and independent variable names in model statement!;
ods output globaltests=gt1;
proc logistic descending data=all_19;
 model low=race2 smoke ht ptl2 lwt3;
 output out=pdat2 xbeta=g_j p=p_j;
run;

data pdat2;
 set pdat2;
          if p_j>=0.5 then ind1=1;
                    else ind1=0;
          if p_j<0.5 then ind2=1;
                    else ind2=0;
                    z1_j=0.5*g_j**2*ind1;
                    z2_j=-0.5*g_j**2*ind2;
run;

*change outcome and independent variable names!;
ods output globaltests=gt2;
proc logistic descending data=pdat2;
 model low=race2 smoke ht ptl2 lwt3 z1_j z2_j;
```

```
run;

data pval;
 merge gt1(rename=(chisq=chisq1))
       gt2(rename=(chisq=chisq2));
                  if _n_=1;
 drop test df probchisq;
                  lr=chisq2-chisq1;
                  pval_lr=(1-probchi(lr,2));
run;

proc print noobs data=pval;
run;
```

	All Data	Covariate Pattern #2 deleted	Covariate Pattern #5 deleted	Covariate Pattern #8 deleted	Covariate Pattern #10 deleted	Covariate Pattern #14 deleted	Covariate Pattern #19 deleted
RACE2	0.9794	0.8084	0.8661	1.1654	0.9205	1.0073	1.1648
SMOKE	0.8679	0.7159	0.9617	0.7739	0.9566	0.8298	1.0812
HT	1.3317	1.2666	1.2900	1.4353	1.3667	1.6694	1.2502
PTL2	1.2454	1.2059	1.1312	0.7268	1.3114	1.2389	1.1491
LWT3	-0.8574	-0.8074	-1.1037	-1.0383	-0.6279	-0.9721	-0.7615

Table 14 – Model coefficients before and after deletion of outliers

p-values	All Data	Covariate Pattern #2 deleted	Covariate Pattern #5 deleted	Covariate Pattern #8 deleted	Covariate Pattern #10 deleted	Covariate Pattern #14 deleted	Covariate Pattern #19 deleted
RACE2	0.0113	0.0907	0.0302	0.0040	0.0194	0.0096	0.0101
SMOKE	0.0260	0.1217	0.0159	0.0520	0.0177	0.0348	0.0212
HT	0.0403	0.0524	0.0469	0.0281	0.0346	0.0162	0.0560
PTL2	0.0049	0.0065	0.0135	0.1516	0.0033	0.0056	0.0115
LWT3	0.0296	0.0434	0.0180	0.0099	0.1839	0.0153	0.0608

Table 15 – p-values before and after deletion of outliers

Grey = $p<0.05$

GOF statistics	All Data	Covariate Pattern #2 deleted	Covariate Pattern #5 deleted	Covariate Pattern #8 deleted	Covariate Pattern #10 deleted	Covariate Pattern #14 deleted	Covariate Pattern #19 deleted
Pearson χ^2	12.9545	12.3173	12.0911	9.5748	12.0879	10.1087	12.5997
Deviance	14.3965	14.1992	13.3642	10.2750	13.6351	11.5066	14.1781
Hosmer-Lemeshow Test	2.6809	3.6834	3.6937	2.9049	6.8166	6.1065	1.2867
GOF p-value	**All Data**	**Covariate Pattern #2 deleted**	**Covariate Pattern #5 deleted**	**Covariate Pattern #8 deleted**	**Covariate Pattern #10 deleted**	**Covariate Pattern #14 deleted**	**Covariate Pattern #19 deleted**
Pearson χ^2	0.6761	0.7219	0.6721	0.8456	0.6724	0.8128	0.7018
Deviance	0.5692	0.5839	0.5742	0.8021	0.5534	0.7159	0.5855
Hosmer-Lemeshow Test	0.7490	0.8154	0.5943	0.7147	0.3381	0.4114	0.9363
Osius-Rojek Test	0.62412	0.5469	0.63394	0.36053	0.62137	0.46046	0.57409
Stukel Test	0.80415	0.95062	0.75977	0.93625	0.88571	0.65271	0.87112

Table 16 – GOF Statistics and p-values before and after deletion of outliers

Table 15 shows the p-value that is changed by deleting covariate patterns 2,5,8,10,14, or 19. Deleting covariate pattern #2 would be a bad idea, because it changes the p-values from all variables being significant, to most non-significant. It would change the model vastly. This is probably due to there being 37 observations within covariate pattern #2. So I'm not deleting it.

Covariate pattern #5 would be ok to delete, it doesn't have a huge effect on p-values or coefficients. It does change some of the GOF tests (Hosmer-Lemeshow). If I delete it, all the p-values remain significant. So I'm going to try deleting it.
Covariate patterns #8, #10, and #19 have an effect on the p-values and GOF statistics, I'm not going to delete those.
Covariate pattern #14 has little effect on p-values and does a have a big effect on the GOF statistics. That is probably because it contains only one observation. I will not delete it because it changes the GOF too much.

In summary, I would try deleting covariate pattern #5 only, which is in dataset ALL_5. The model stability would be ok by deleting it. Ooooo, I could try bootstrapping, like Austin et al.

6. **Keeping all outliers in the data set, determine how well your model predicts the outcome.**
   a. **Plot the ROC curve and determine the area under the ROC curve**

- "Receiver Operating Characteristic (ROC)
- Area under the ROC curve measures receiver's ability to discriminate between true and false signals."

```
*** cross classification and roc curve ****;

proc logistic descending data=all;
 model low=race2 smoke ht ptl2 lwt3
            /ctable pprob=(0.2) pevent=0.2 0.3 outroc=rocdat;
run;


symbol v=none i=join;

axis1 order=(0 to 1 by 0.1) label=(f=swiss h=1.5 '1-specificity');
axis2 order=(0 to 1 by 0.1) label=(a=90 f=swiss h=1.5 'sensitivity');

footnote1 ' ';

proc gplot data=rocdat;
 plot _sensit_*_1mspec_/haxis=axis1 vaxis=axis2 hminor=0 vminor=0;
run;


data rocdat;
 set rocdat;
 spec=1-_1mspec_;
run;

symbol1 c=black v=dot h=0.5 i=join;
symbol2 c=black v=diamond h=1.0 i=join;

axis1 order=(0 to 1 by 0.1) label=(f=swiss h=1.5 'probability cutoff');
axis2 order=(0 to 1 by 0.1) label=(a=90 f=swiss h=1.5 'sensitivity,specificity');

footnote1 c=black f=special h=1 'j j j' f=swissb h=1.5 '   sensitivity'
      c=black f=special h=1 '     d d d' f=swissb h=1.5 '   specificity';

proc gplot data=rocdat;
 plot (_sensit_ spec)*_prob_/overlay haxis=axis1 vaxis=axis2 hminor=0 vminor=0;
run;

quit;
end;
```

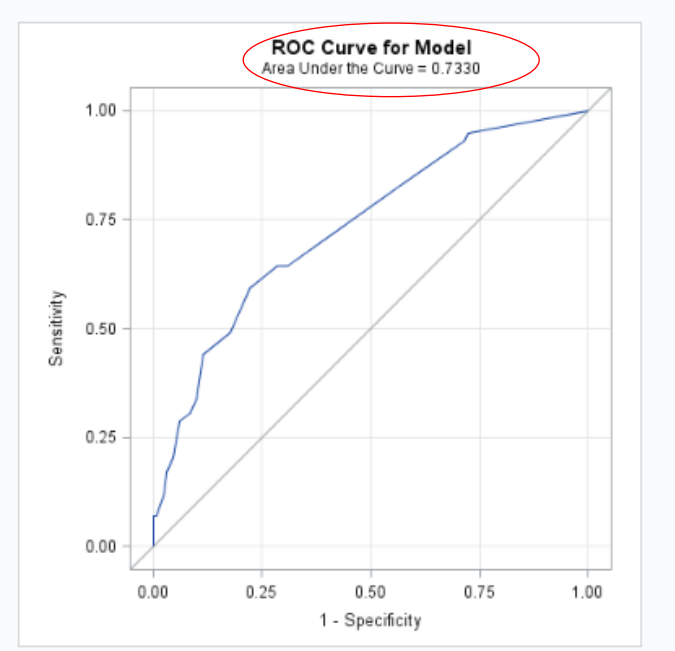


Figure 51 – ROC Curve

b. Plot sensitivity and specificity vs. possible cutpoints and select the"best" cutpoint

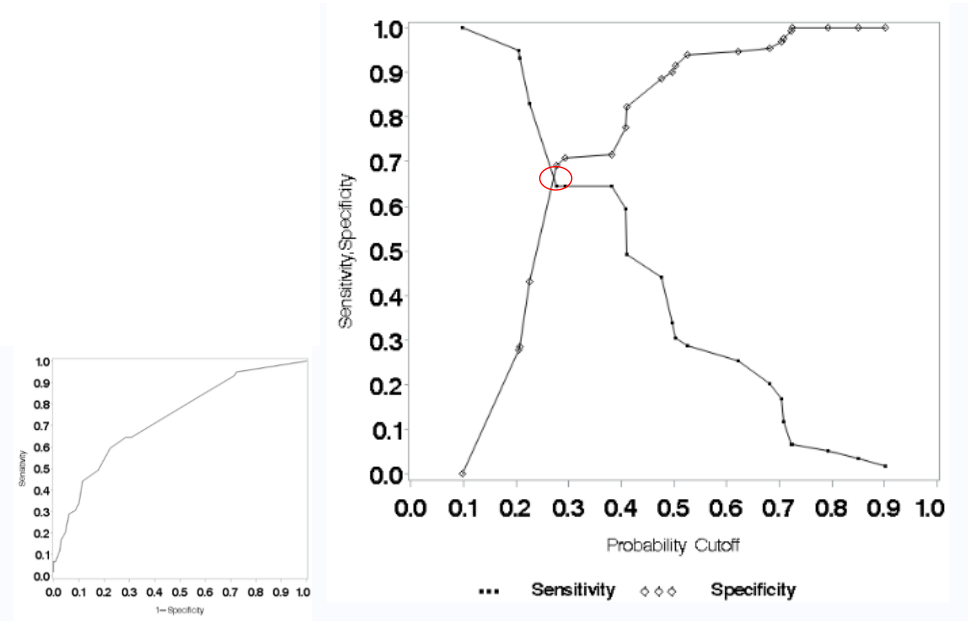


Figure 52 – Sensitivity & Specificity

- Se and Sp cross at pihat≈0.3
- Se higher for pihat<0.25, but lower for pihat >0.25
- Sp higher for pihat >0.25, but lower for pihat <0.25

```
*b.        Plot sensitivity and specificity vs. possible cutpoints and select the "best" cutpoint;
proc gplot data=rocdat;
 plot _sensit_*_1mspec_/haxis=axis1 vaxis=axis2 hminor=0 vminor=0;
run;

*Calculates specificity based on 1 minus specificity;
data rocdat;
 set rocdat;
 spec=1-_1mspec_;
run;

symbol1 c=black v=dot h=0.5 i=join;
symbol2 c=black v=diamond h=1.0 i=join;

axis1 order=(0 to 1 by 0.1) label=(f=swiss h=2.5 'Probability Cutoff');
axis2 order=(0 to 1 by 0.1) label=(a=90 f=swiss h=2.5 'Sensitivity,Specificity');

footnote1 c=black f=special h=1 'J J J' f=swissb h=1.5 '    Sensitivity'
      c=black f=special h=1 '      D D D' f=swissb h=1.5 '    Specificity';

proc gplot data=rocdat;
 plot (_sensit_ spec)*_prob_/overlay haxis=axis1 vaxis=axis2 hminor=0 vminor=0;
run;
QUIT;
```

c. Based on this cutpoint, calculate sensitivity and specificity

VIEWTABLE: Work.Rocdat

	Probability Level	No. of Correctly Predicted Events	No. of Correctly Predicted Nonevents	No. of Nonevents Predicted as Events	No. of Events Predicted as Nonevents	Sensitivity	1 - Specificity	spec
1	0.9005746873	1	130	0	58	0.0169491525	0	1
2	0.8506659393	2	130	0	57	0.0338983051	0	1
3	0.7935050728	3	130	0	56	0.0508474576	0	1
4	0.7248641765	4	130	0	55	0.0677966102	0	1
5	0.7227689877	4	129	1	55	0.0677966102	0.0076923077	0.9923076923
6	0.7073175344	7	127	3	52	0.1186440678	0.0230769231	0.9769230769
7	0.7051430791	10	126	4	49	0.1694915254	0.0307692308	0.9692307692
8	0.6814416503	12	124	6	47	0.2033898305	0.0461538462	0.9538461538
9	0.6211531968	15	123	7	44	0.2542372881	0.0538461538	0.9461538462
10	0.5251777456	17	122	8	42	0.2881355932	0.0615384615	0.9384615385
11	0.503616994	18	119	11	41	0.3050847458	0.0846153846	0.9153846154
12	0.4973226999	20	117	13	39	0.3389830508	0.1	0.9
13	0.4757596905	26	115	15	33	0.4406779661	0.1153846154	0.8846153846
14	0.4102344056	29	107	23	30	0.4915254237	0.1769230769	0.8230769231
15	0.4077010405	35	101	29	24	0.593220339	0.2230769231	0.7769230769
16	0.3810761824	38	93	37	21	0.6440677966	0.2846153846	0.7153846154
17	0.2934612423	38	92	38	21	0.6440677966	0.2923076923	0.7076923077
18	0.2758861223	38	90	40	21	0.6440677966	0.3076923077	0.6923076923
19	0.2260205108	49	56	74	10	0.8305084746	0.5692307692	0.4307692308
20	0.2071115246	55	37	93	4	0.9322033898	0.7153846154	0.2846153846
21	0.2053956722	56	36	94	3	0.9491525424	0.7230769231	0.2769230769
22	0.0988250461	59	0	130	0	1	1	0

```
proc logistic descending data=all;
        model low=race2 smoke ht ptl2 lwt3/ctable pprob=(0.3) pevent=0.2
        outroc=rocdat;
run;

*relevant sas code:;
proc logistic descending data=all;
        model low=race2 smoke ht ptl2 lwt3;
        output out=pdat p=pihat;
run;

proc univariate data=pdat;
run;
```

Classification Table										
		Correct		Incorrect		Percentages				
Prob Event	Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.200	0.300	38	92	38	21	69.5	64.4	70.8	64.5	11.2

Figure 53 – Se, Sp, PPV, NPV

Sensitivity = 64.4%
Specificity = 70.8%

Se and Sp are not indicators of goodness-of-fit and don't depend on prevalence.

d. Calculate the positive and negative predictive value assuming the prevalence of the outcome in the population of interest is 10%.

PPV & NPV with 0.3 cut-off 0.20 probablity event:
PPV = 100% - false positives=100% - 64.5%
PPV = 35.5%

NPV=100% - false negatives = 100% - 11.2%
NPV = 88.8%

Sensitivity & PPV are very low. Specificity is sub-optimal.

Positive predictive value [edit]

The positive predictive value (PPV) is defined as

$$\mathrm{PPV} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false positives}} = \frac{\text{number of true positives}}{\text{number of positive calls}}$$

where a "true positive" is the event that the test makes a positive prediction, and the subject has a positive result under the gold standard, and a "

standard.

The PPV can also be computed from sensitivity, specificity, and prevalence:

$$\mathrm{PPV} = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}$$

The complement of the PPV is the false discovery rate (FDR):

$$\mathrm{FDR} = 1 - \mathrm{PPV} = \frac{\text{number of false positives}}{\text{number of true positives} + \text{number of false positives}} = \frac{\text{number of false positives}}{\text{number of positive calls}}$$

Negative predictive value [edit]

The negative predictive value is defined as:

$$\mathrm{NPV} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false negatives}} = \frac{\text{number of true negatives}}{\text{number of negative calls}}$$

where a "true negative" is the event that the test makes a negative prediction, and the subject has a negative result under the gold standard, and gold standard.

The NPV can also be computed from sensitivity, specificity, and prevalence:

$$\mathrm{NPV} = \frac{\text{specificity} \times (1 - \text{prevalence})}{(1 - \text{sensitivity}) \times \text{prevalence} + \text{specificity} \times (1 - \text{prevalence})}$$

Figure 52 – Formulas for NPV and PPV, just to help me interpret this ☺ (Source: http://en.wikipedia.org/wiki/Positive_and_negative_predictive_values)

7. **For your final model, present a table containing appropriate ORs and 95% confidence intervals.**

```
*******************7*******************;
proc logistic descending data=all;
   model low=race2 smoke ht ptl2 lwt3/scale=n aggregate lackfit;
run;
```

Variable	Comparison/Unit	OR	95% CI		p-value
RACE2	RACE2 1 vs 0 (1=black+other race vs 0=not white, not other race)	2.663	1.249	5.679	0.0113
SMOKE	SMOKE 1 vs 0 (1=yes vs 0=no smoking during pregnancy)	2.382	1.110	5.113	0.0260
HT	HT 1 vs 0 (0=no history of hypertension, 1=yes)	3.788	1.061	13.524	0.0403
PTL2	PTL2 1 vs 0 (1=≥2 vs 0=history of premature labor)	3.474	1.459	8.276	0.0049
LWT3	LWT3 1 vs 0 LWT2 1 vs 0 (0=80-109 pounds 1=110-251 pounds	0.424	0.196	0.919	0.0296

Table 23 – Final model

Green means p-value is below 0.05 cutoff
Red means wide 95% CI

8. **Interpret the ORs and 95% confidence intervals.**

RACE2: I collapsed black and other race into value 1, and 0 was white race. Therefore the way to interpret this is that the risk of having a low birth weight baby is 2.663 times as likely in a population of black or other race mothers than in a population of white mothers. We are 95% confident that the true population odds ratio lies between 1.249 and 5.679. This confidence interval width is fairly wide and thus the model has doesn't really have high power. Since a one is not contained within the confidence interval, is is a statistically significant confidence interval. Of note, the p-value was 0.0113, which is less than an alpha of 0.05, and therefore statistically significant, but I already talked about that with the CI and epidemiologists like CIs better than p-values because CI's also have a power calculation in them (the CI width reflects power). Race is a known risk factor for low birth weight, and can be attributed to socioeconomic factors, access to healthcare, cultural norms, and nutrition.

SMOKE: With the dichotomous variable smoking, the risk of having a low birth weight baby is 2.382 times as likely in a population of mothers that smoke than in a population of mothers who do not smoke. The royal "we" are 95% confident that the true population odds ratio lies between 1.110 and 5.113. This confidence interval width is fairly wide and thus the model has lower power. Since a one is not contained within the confidence interval, is is a statistically significant confidence interval. Of note, the p-value was 0.0260, which is less than an alpha of 0.05, and therefore statistically significant. Smoking is also a known risk factor for low birth

weight, it makes sense physiologically since smoking vasoconstricts blood vessels to the placenta, thus depriving the fetus of blood flow, which results in lower weight infants. Hey, there's benefits to being an old NICU nurse. ☺

HT: With the dichotomous variable hypertension, the risk of having a low birth weight baby is 3.788 times as likely in a population of mothers that have hypertension than in a population of mothers who do not have hypertension. We are 95% confident that the true population odds ratio lies between 1.061 and 13.524. This confidence interval width is very wide and thus the model has low power. Since a one is not contained within the confidence interval, is is a statistically significant confidence interval. Of note, the p-value was 0.0403, which is slightly less than an alpha of 0.05, and therefore statistically significant. It's a pretty high p-value and wide CI, which tells me I have low power. I could fix that by bootstrapping, correct (increase my power)? Hypertension is also a known risk factor for low birth weight, it makes sense physiologically since hypertension constricts blood vessels to the placenta, thus depriving the fetus of blood flow, which results in lower weight infants. It's the same reason as smoking.

PTL2: I collapsed history of premature birth for variable PTL2. It's history of premature birth (PTL2=1), which can be 1 or greater preterm births, or (PTL2=0), which means no history of preterm births. Therefore the risk of having a low birth weight infant is 3.474 times as likely in a population of mothers who have had a prior history of preterm births than in a population of mothers who have never had a previous preterm birth. This makes sense physiologically and is a known risk factor for low birth weight. We are 95% confident that the true population odds ratio lies between 1.459 and 8.276. This is a fairly wide confidence interval, in fact "terrible". That makes PTL2 lower in power as a predictor. Since a one is not contained within the confidence interval, is is a statistically significant confidence interval. Of note, the p-value was 0.0049, which is way less than an alpha of 0.05, and therefore statistically significant.

LWT3: I categorized weight of mother (which was a continuous variable originally) to the now categorical variable LWT3. It's weight of mother (LWT3=1), which means 110-251 pounds, or (LWT3=0), which means 80-109 pounds. Therefore the risk of having a low birth weight infant is 0.424 times as likely in a population of mothers who who weight 110-251 pounds than in a population of mothers who weight 80-109 pounds, which means the higher weight (LWT3=1) is protective against having a low weight infant. This makes sense physiologically and is a known risk factor for low birth weight. We are 95% confident that the true population odds ratio lies between 0.196 and 0.919. This is a narrow confidence interval.. That makes LWT3 higher in power as a predictor. Since a one is not contained within the confidence interval, is is a statistically significant confidence interval. Of note, the p-value was 0.0296, which is less than an alpha of 0.05, and therefore statistically significant.