STAT 448 HW #3

Tianqi Wu

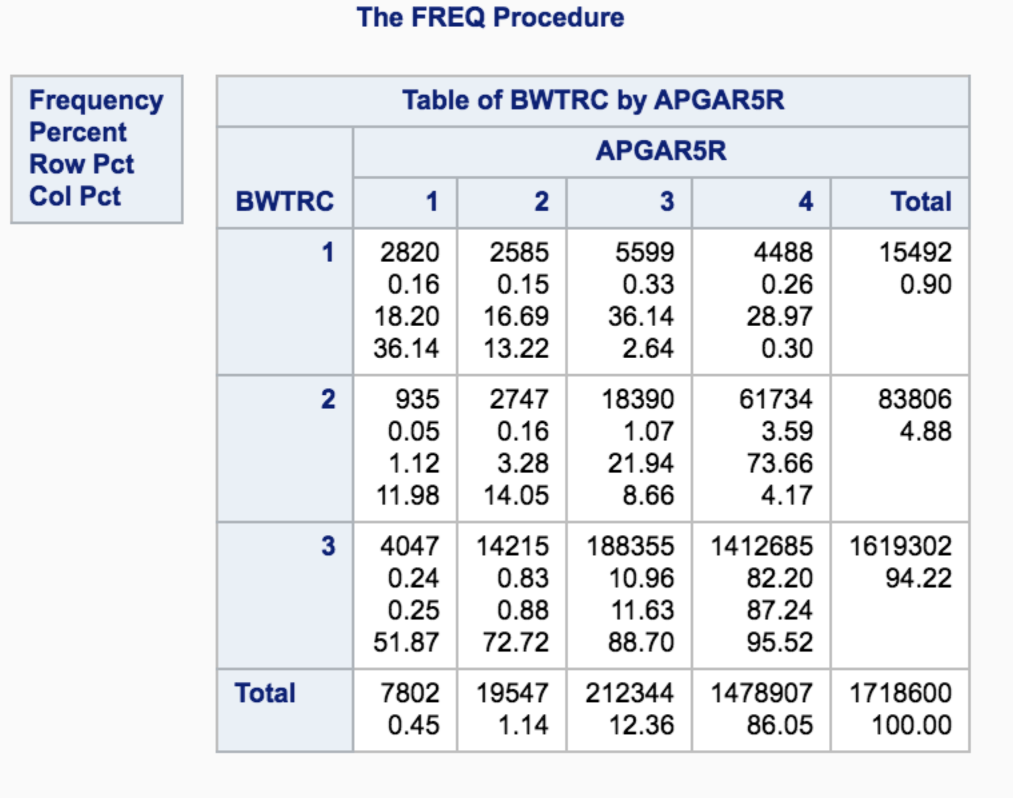
2018/10/03

Problem 1

Here is the code that produces the table:

```
proc freq data = birth2007;
table BWTRC*APGAR5R;
run;
```

From the table, positive association can be observed between APGAR5R and BWTRC. For all three levels of BWTRC, the frequency of the birth weight category tends to increase as apgar score increases looking at the row percent and column percent from the table.



The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of BWTRC by APGAR5R

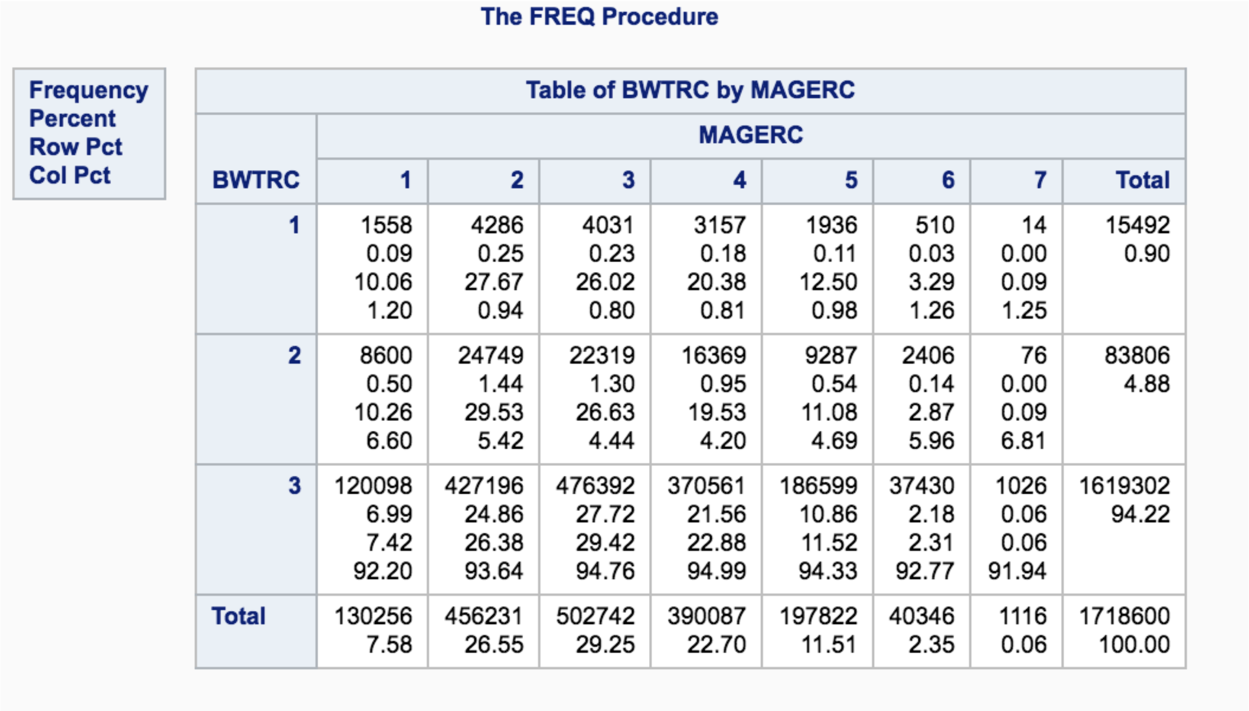
	APGAR5R				
BWTRC	1	2	3	4	Total
1	2820 0.16 18.20 36.14	2585 0.15 16.69 13.22	5599 0.33 36.14 2.64	4488 0.26 28.97 0.30	15492 0.90
2	935 0.05 1.12 11.98	2747 0.16 3.28 14.05	18390 1.07 21.94 8.66	61734 3.59 73.66 4.17	83806 4.88
3	4047 0.24 0.25 51.87	14215 0.83 0.88 72.72	188355 10.96 11.63 88.70	1412685 82.20 87.24 95.52	1619302 94.22
Total	7802 0.45	19547 1.14	212344 12.36	1478907 86.05	1718600 100.00

Problem 2

Here is the code that produces the table:

```
      proc freq data = birth2007;
      table BWTRC*MAGERC;
      run;
```

From the table, there is no obvious correlation between birth weight categories and mother's age categories. Because the distribution of frequency of the birth weight categories does not change much among mother's age categories looking at the row percent and column percent from the table.



The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

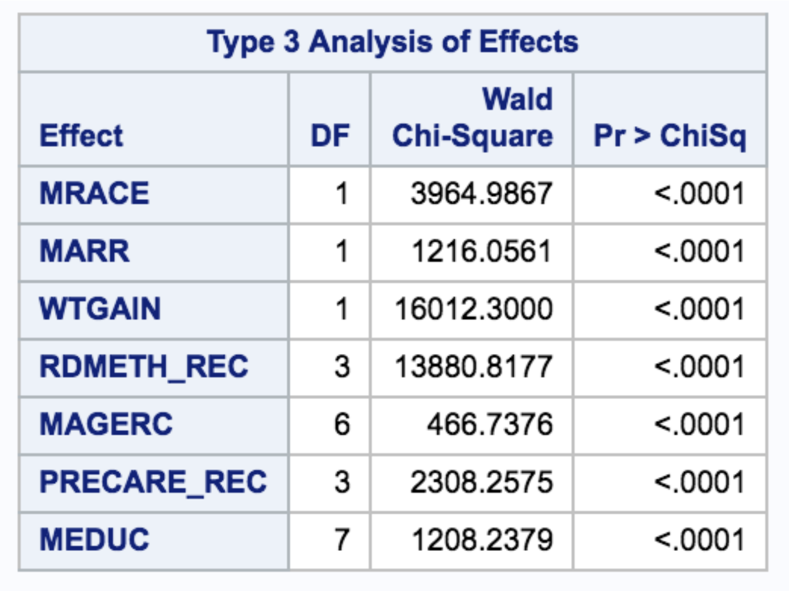
	Table of BWTRC by MAGERC							
	MAGERC							
BWTRC	1	2	3	4	5	6	7	Total
1	1558 0.09 10.06 1.20	4286 0.25 27.67 0.94	4031 0.23 26.02 0.80	3157 0.18 20.38 0.81	1936 0.11 12.50 0.98	510 0.03 3.29 1.26	14 0.00 0.09 1.25	15492 0.90
2	8600 0.50 10.26 6.60	24749 1.44 29.53 5.42	22319 1.30 26.63 4.44	16369 0.95 19.53 4.20	9287 0.54 11.08 4.69	2406 0.14 2.87 5.96	76 0.00 0.09 6.81	83806 4.88
3	120098 6.99 7.42 92.20	427196 24.86 26.38 93.64	476392 27.72 29.42 94.76	370561 21.56 22.88 94.99	186599 10.86 11.52 94.33	37430 2.18 2.31 92.77	1026 0.06 0.06 91.94	1619302 94.22
Total	130256 7.58	456231 26.55	502742 29.25	390087 22.70	197822 11.51	40346 2.35	1116 0.06	1718600 100.00

Problem 3

Here is the code that produces the logistic regression model:

```
data birth2007_1;
set birth2007;
if DBWT <2500 then DBWTB=1;
else if DBWT>=2500 then DBWTB=0;
run;

proc logistic data = birth2007_1 ;
 class MRACE(param=ref ref='1') MARR(param=ref ref='1')
RDMETH_REC(param=ref ref='4')
         MAGERC(param=ref ref='7')  PRECARE_REC(param=ref
ref='4') MEDUC(param=ref ref='8');
 model DBWTB(ref = '0') = MRACE MARR WTGAIN RDMETH_REC
MAGERC PRECARE_REC MEDUC;
run;
```

From the result, all of the predictors are significant with p-value < 0.0001. It means that mother's race, mother's marital status, mother's weight gain, delivery method, mother's age (categorical), beginning prenatal care period, and mother's education level are all significant.



Type 3 Analysis of Effects

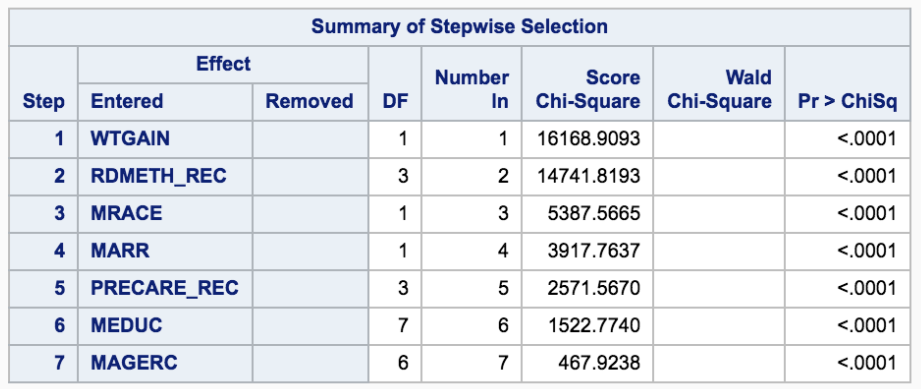
Effect	DF	Wald Chi-Square	Pr > ChiSq
MRACE	1	3964.9867	<.0001
MARR	1	1216.0561	<.0001
WTGAIN	1	16012.3000	<.0001
RDMETH_REC	3	13880.8177	<.0001
MAGERC	6	466.7376	<.0001
PRECARE_REC	3	2308.2575	<.0001
MEDUC	7	1208.2379	<.0001

Problem 4

Here is the code that preforms model selection and detects influential points:

```
proc logistic data = birth2007_1 ;
 class MRACE(param=ref ref='1') MARR(param=ref ref='1')
RDMETH_REC(param=ref ref='2')
         MAGERC(param=ref ref='7')  PRECARE_REC(param=ref
ref='4') MEDUC(param=ref ref='8');
 model DBWTB(ref = '0') = MRACE MARR WTGAIN RDMETH_REC
MAGERC PRECARE_REC MEDUC
 /selection = stepwise sle=0.05 sls=0.05 ;
 output out=birth2007_cbar CBAR = CBAR;
run;

proc print data = birth2007_cbar;
where CBAR>0.5;
run;
```

From the stepwise model selection result, all of the predictors are significant. Hence, mother's race, mother's marital status, mother's weight gain, delivery method, mother's age (categorical), beginning prenatal care period, and mother's education level should all be kept in the model. Analysis of influential points is first done by using IQR rule to detect influential points. However, there is no observation with low infant birth weight after removing the outliers. Hence, influential points are detected again using the criterion Cbar > 0.5. With this cut-off value suggested from the lecture, there is no extreme influential point and we do not need to remove any of the observations.



Summary of Stepwise Selection

	Effect						
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	WTGAIN		1	1	16168.9093		<.0001
2	RDMETH_REC		3	2	14741.8193		<.0001
3	MRACE		1	3	5387.5665		<.0001
4	MARR		1	4	3917.7637		<.0001
5	PRECARE_REC		3	5	2571.5670		<.0001
6	MEDUC		7	6	1522.7740		<.0001
7	MAGERC		6	7	467.9238		<.0001

Problem 5

Here is the code that produces the logistic regression:

```
proc logistic data = birth2007_1 ;
 class MRACE(param=ref ref='1') MARR(param=ref ref='1')
RDMETH_REC(param=ref ref='2')
         MAGERC(param=ref ref='7')  PRECARE_REC(param=ref
ref='4') MEDUC(param=ref ref='8');
 model DBWTB(ref = '0') = MRACE MARR WTGAIN RDMETH_REC
MAGERC PRECARE_REC MEDUC/lackfit rsquare ;
run;
```

From the model result, the predictors captures only 2.46 percent of variation in the response. It means that the model predicts poorly. The Hosmer and Lemeshow Goodness-of-Fit Test significance (p-value<0.0001) rejects the null and we conclude that observed and predicted probabilities are different. The diagnostics do not show any major violations and there is no point beyond the Cbar cutoff of 0.5. From the odds ratio estimates, The confidence interval includes 1 for the following pairs: mother's age (1 vs 7 and 6 vs 7), mother's education level (1 vs 8, 6 vs 8 and 7 vs 8). Hence, except for the pairs above, the levels of the predictors are significant versus their reference levels.

The parameter estimates are be interpreted by odds ratio as follows. The odds of an infant having low birth weight for mother's race = 0 is 0.612 times the odds for mother's race = 1. The odds of an infant having low birth weight for mother's marital status = 0 is 0.763 times the odds for mother's marital status = 1. Rest of parameter estimates can be interpreted in the same way.

To conclude, the odds of an infant having low birth weight increases for infants with following feature: mother's race = not white , mother's marital status = not married, low mother's weight gain, delivery method = primary cesarean, mother's age = 45-49 yrs, beginning prenatal care period = no prenatal care, and mother's education level. = some high school.

R-Square	0.0246	Max-rescaled R-Square	0.0689

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
131.4230	8	<.0001

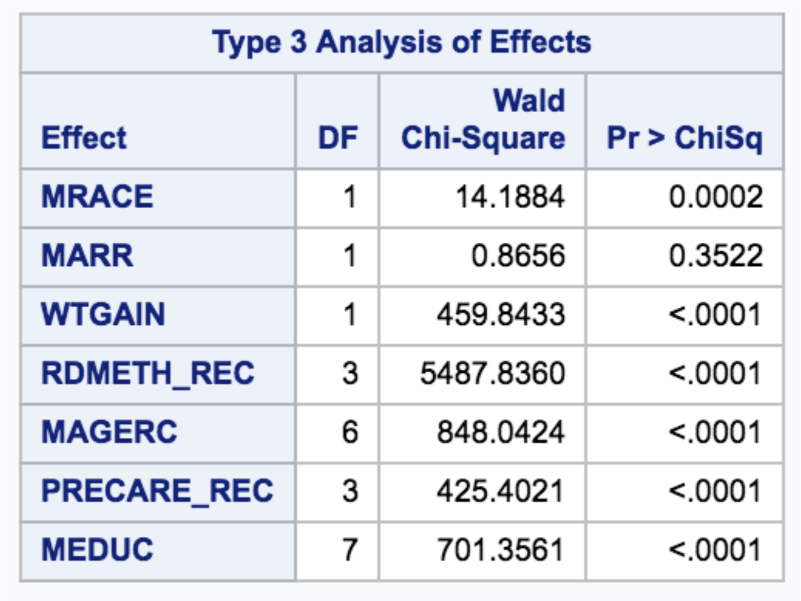
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
MRACE 0 vs 1	0.612	0.603	0.622
MARR 0 vs 1	0.763	0.751	0.774
WTGAIN	0.970	0.970	0.971
RDMETH_REC 1 vs 4	0.811	0.794	0.828
RDMETH_REC 2 vs 4	1.188	1.120	1.260
RDMETH_REC 3 vs 4	1.963	1.920	2.008
MAGERC 1 vs 7	0.849	0.680	1.058
MAGERC 2 vs 7	0.739	0.593	0.921
MAGERC 3 vs 7	0.687	0.551	0.855
MAGERC 4 vs 7	0.687	0.552	0.857
MAGERC 5 vs 7	0.757	0.607	0.943
MAGERC 6 vs 7	0.900	0.720	1.125
PRECARE_REC 1 vs 4	0.446	0.430	0.461
PRECARE_REC 2 vs 4	0.451	0.435	0.468
PRECARE_REC 3 vs 4	0.368	0.352	0.385
MEDUC 1 vs 8	1.037	0.974	1.106
MEDUC 2 vs 8	1.445	1.362	1.534
MEDUC 3 vs 8	1.386	1.308	1.470
MEDUC 4 vs 8	1.280	1.207	1.357
MEDUC 5 vs 8	1.160	1.090	1.234
MEDUC 6 vs 8	1.033	0.973	1.095
MEDUC 7 vs 8	0.982	0.922	1.047

Problem 6

Here is the code that produces the cumulative logit model:

```
proc logistic data = birth2007;
 class MRACE(param=ref ref='1') MARR(param=ref ref='1')
RDMETH_REC(param=ref ref='4')
         MAGERC(param=ref ref='7')  PRECARE_REC(param=ref
ref='4') MEDUC(param=ref ref='8');
 model APGAR5R = MRACE MARR WTGAIN RDMETH_REC MAGERC
PRECARE_REC MEDUC
 /link = clogit;
run;
```

From the result, all of the predictors are significant with p-value < 0.0001 except for MARR with p-value > 0.3522 . It means that mother's race, mother's weight gain, delivery method, mother's age (categorical), beginning prenatal care period, and mother's education level are significant.



Type 3 Analysis of Effects

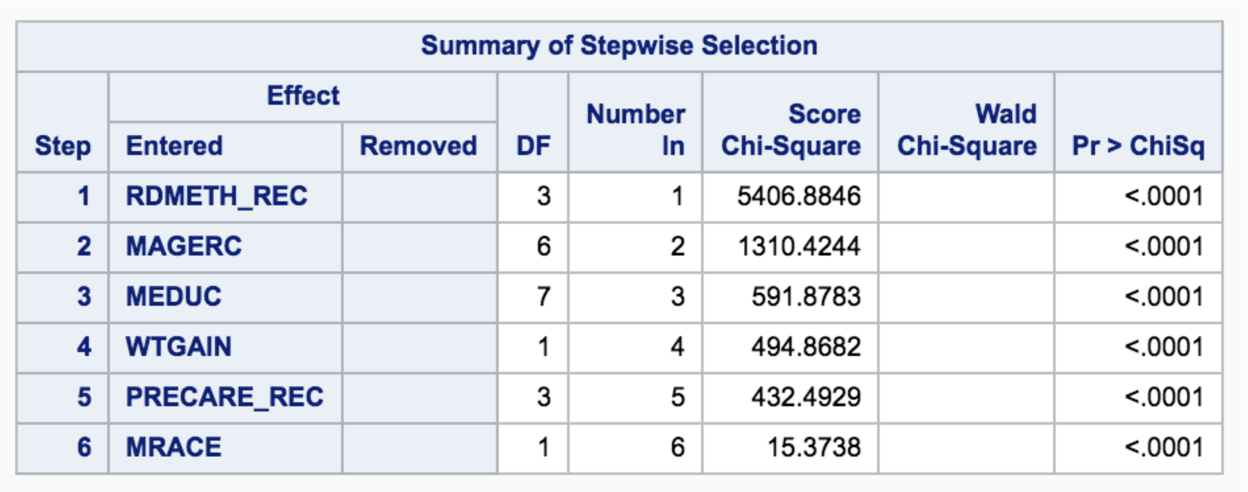
Effect	DF	Wald Chi-Square	Pr > ChiSq
MRACE	1	14.1884	0.0002
MARR	1	0.8656	0.3522
WTGAIN	1	459.8433	<.0001
RDMETH_REC	3	5487.8360	<.0001
MAGERC	6	848.0424	<.0001
PRECARE_REC	3	425.4021	<.0001
MEDUC	7	701.3561	<.0001

Problem 7

Here is the code that preforms model selection:

```
proc logistic data = birth2007;
 class MRACE(param=ref ref='1') RDMETH_REC(param=ref
ref='4')
         MAGERC(param=ref ref='7')  PRECARE_REC(param=ref
ref='4') MEDUC(param=ref ref='8');
 model APGAR5R = MRACE WTGAIN RDMETH_REC MAGERC PRECARE_REC
MEDUC
 /link = clogit lackfit rsquare;
run;
```

From the stepwise model selection result, all of the predictors are significant except for MARR(mother's marital status). Hence, mother's race, mother's weight gain, delivery method, mother's age (categorical), beginning prenatal care period, and mother's education level should be kept in the model. Since Cbar is not available for cumulative logit model, analysis of influential points is ignored.



Summary of Stepwise Selection

	Effect						
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	RDMETH_REC		3	1	5406.8846		<.0001
2	MAGERC		6	2	1310.4244		<.0001
3	MEDUC		7	3	591.8783		<.0001
4	WTGAIN		1	4	494.8682		<.0001
5	PRECARE_REC		3	5	432.4929		<.0001
6	MRACE		1	6	15.3738		<.0001

Problem 8

Here is the code that produces the cumulative logit model:

```
proc logistic data = birth2007;
 class MRACE(param=ref ref='1') MARR(param=ref ref='1')
RDMETH_REC(param=ref ref='4')
         MAGERC(param=ref ref='7')  PRECARE_REC(param=ref
ref='4') MEDUC(param=ref ref='8');
 model APGAR5R = MRACE WTGAIN RDMETH_REC MAGERC PRECARE_REC
MEDUC
 /link = clogit lackfit rsquare;
run;
```

From the model result, the predictors captures only 0.46 percent of variation in the response. The Hosmer and Lemeshow Goodness-of-Fit Test significance (p-value<0.0001) rejects the null and we conclude that observed and predicted probabilities are different. The diagnostics do not show any major violations. The confidence interval includes 1 for the following pairs: mother's age (1-6 vs 7), mother's education level (1 vs 8). Hence, except for the pairs above, the levels of the predictors are significant versus their reference levels.

The parameter estimates are be interpreted by odds ratio as follows. For any fixed level of five minute APGAR as category, the estimated odds that the response for mother's race = 0 are in the lower order direction rather than the higher order direction equal exp(0.977) times the estimated odds for mother's race = 1.Rest of parameter estimates can be interpreted in the same way.

To conclude, the odds of an infant having lower five minute APGAR score indicating infant's health increases for infants with following feature: mother's race = not white , low mother's weight gain, delivery method = primary cesarean, mother's age = 15-19 yrs, beginning prenatal care period = no prenatal care, and mother's education level = associate degree.

R-Square	0.0046	Max-rescaled R-Square	0.0077

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
1797.6843	26	<.0001

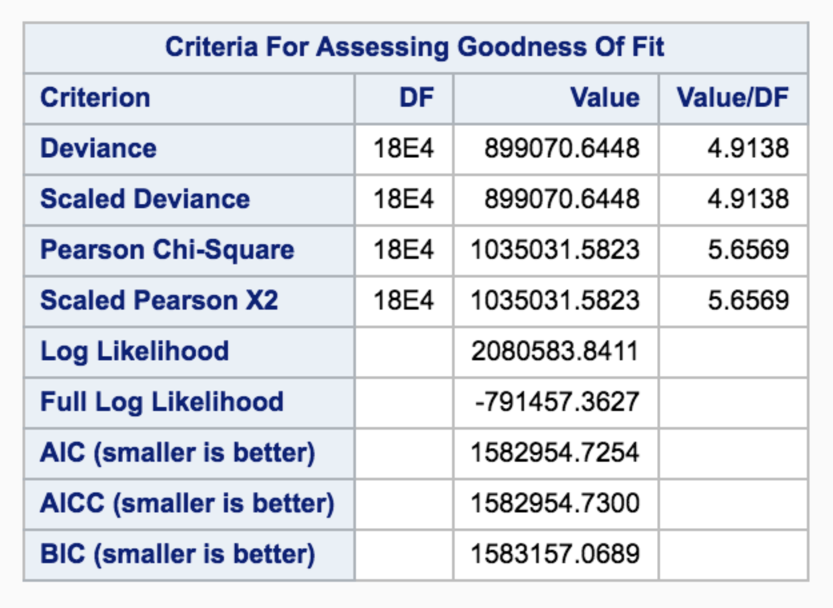
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
MRACE 0 vs 1	0.977	0.966	0.989
WTGAIN	0.997	0.997	0.997
RDMETH_REC 1 vs 4	1.050	1.035	1.065
RDMETH_REC 2 vs 4	1.341	1.288	1.396
RDMETH_REC 3 vs 4	1.537	1.512	1.562
MAGERC 1 vs 7	1.150	0.973	1.359
MAGERC 2 vs 7	1.036	0.877	1.224
MAGERC 3 vs 7	0.954	0.807	1.126
MAGERC 4 vs 7	0.894	0.757	1.056
MAGERC 5 vs 7	0.898	0.760	1.061
MAGERC 6 vs 7	0.964	0.814	1.141
PRECARE_REC 1 vs 4	0.749	0.726	0.773
PRECARE_REC 2 vs 4	0.797	0.772	0.822
PRECARE_REC 3 vs 4	0.791	0.764	0.820
MEDUC 1 vs 8	0.990	0.951	1.031
MEDUC 2 vs 8	1.110	1.069	1.153
MEDUC 3 vs 8	1.148	1.106	1.191
MEDUC 4 vs 8	1.238	1.193	1.284
MEDUC 5 vs 8	1.244	1.196	1.293
MEDUC 6 vs 8	1.158	1.116	1.201
MEDUC 7 vs 8	1.082	1.040	1.125

Problem 9

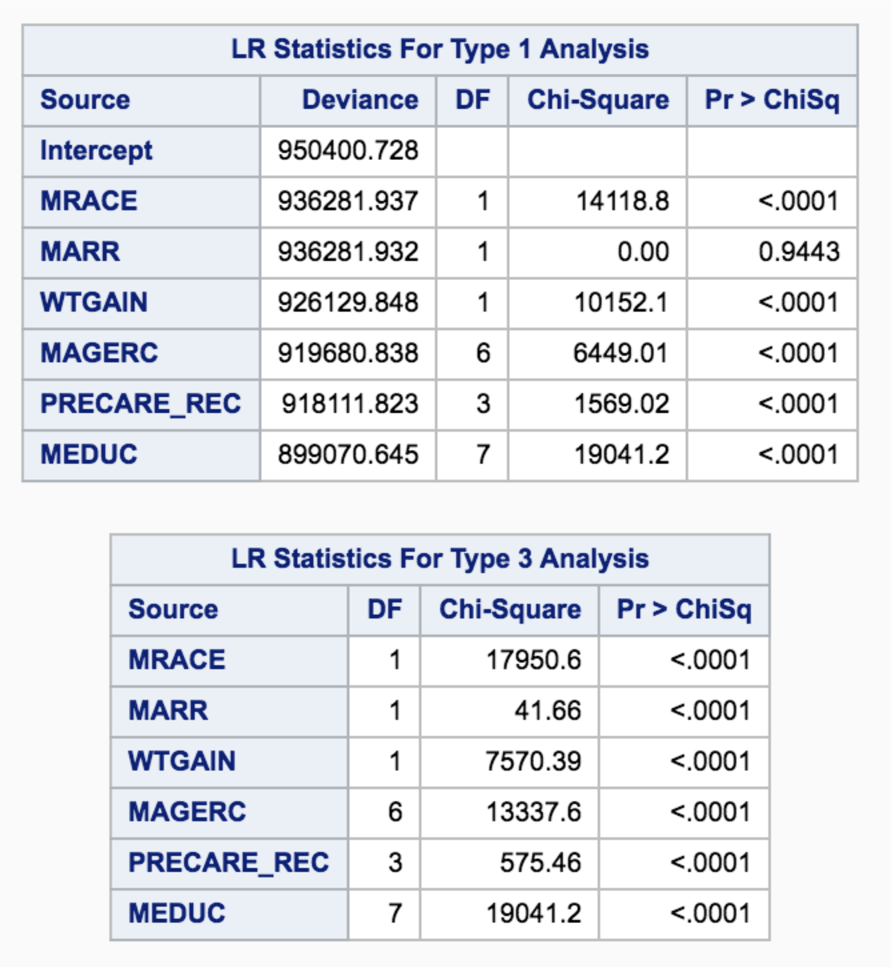
Here is the code that produces the data and Poisson log-linear model:

```
data cigbirth2007;
set birth2007;
where CIG_REC = 'Y';
DAILYCIG_AVG = round((CIG_1+CIG_2+CIG_3)/3);
run;

proc genmod data = cigbirth2007;
 class MRACE(param=ref ref='1') MARR(param=ref ref='1')
         MAGERC(param=ref ref='7')  PRECARE_REC(param=ref
ref='4') MEDUC(param=ref ref='8');
 model DAILYCIG_AVG = MRACE MARR WTGAIN MAGERC PRECARE_REC
MEDUC
 / dist=poisson link=log type1 type3;
run;
```

From the model result, the value/df of scaled deviance is 4.9138. Since the dispersion estimate is larger than 1, we should use an overdispersed model. From type I analysis, all of the predictors are significant with p-value < 0.0001 except for MARR with p-value > 0.9443. From type III analysis, all of the predictors are significant with p-value < 0.0001. Regardless of the order, all predictors are significant based on type III. The predictors are mother's race, mother's marital status, mother's weight gain, mother's age (categorical), beginning prenatal care period, and mother's education level.



Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	18E4	899070.6448	4.9138
Scaled Deviance	18E4	899070.6448	4.9138
Pearson Chi-Square	18E4	1035031.5823	5.6569
Scaled Pearson X2	18E4	1035031.5823	5.6569
Log Likelihood		2080583.8411	
Full Log Likelihood		-791457.3627	
AIC (smaller is better)		1582954.7254	
AICC (smaller is better)		1582954.7300	
BIC (smaller is better)		1583157.0689	



LR Statistics For Type 1 Analysis				
Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	950400.728			
MRACE	936281.937	1	14118.8	<.0001
MARR	936281.932	1	0.00	0.9443
WTGAIN	926129.848	1	10152.1	<.0001
MAGERC	919680.838	6	6449.01	<.0001
PRECARE_REC	918111.823	3	1569.02	<.0001
MEDUC	899070.645	7	19041.2	<.0001

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
MRACE	1	17950.6	<.0001
MARR	1	41.66	<.0001
WTGAIN	1	7570.39	<.0001
MAGERC	6	13337.6	<.0001
PRECARE_REC	3	575.46	<.0001
MEDUC	7	19041.2	<.0001

Problem 10

Here is the code that produces the data and overdispersed Poisson log-linear model:

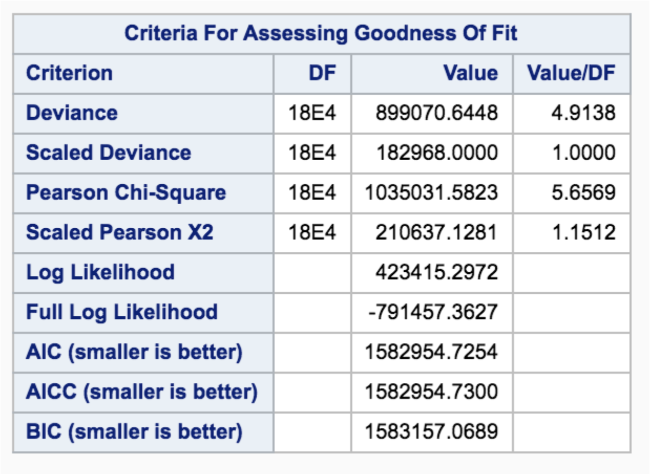
```
proc genmod data = cigbirth2007;
 class MRACE(param=ref ref='1') MARR(param=ref ref='1')
         MAGERC(param=ref ref='7')  PRECARE_REC(param=ref
ref='4') MEDUC(param=ref ref='8');
 model DAILYCIG_AVG = MRACE MARR WTGAIN MAGERC PRECARE_REC
MEDUC
 / dist=poisson link=log scale=deviance type1 type3;
 output out=cigbirth2007_1 cooksd= cook1 ;
run;

proc print data=cigbirth2007_1;
 where cook1 > 1;
run;
```

After trying removing the variables manually, the model performance is not improved. Hence, it optimal to keep all the predictors. The set of best predictors are mother's race, mother's marital status, mother's weight gain, mother's age (categorical), beginning prenatal care period, and mother's education level. Using the cut-off value of 1 for cooks distance, there is no observation with extreme value. We do not need to worry about the influential points. The diagnostics do not show any major violations.

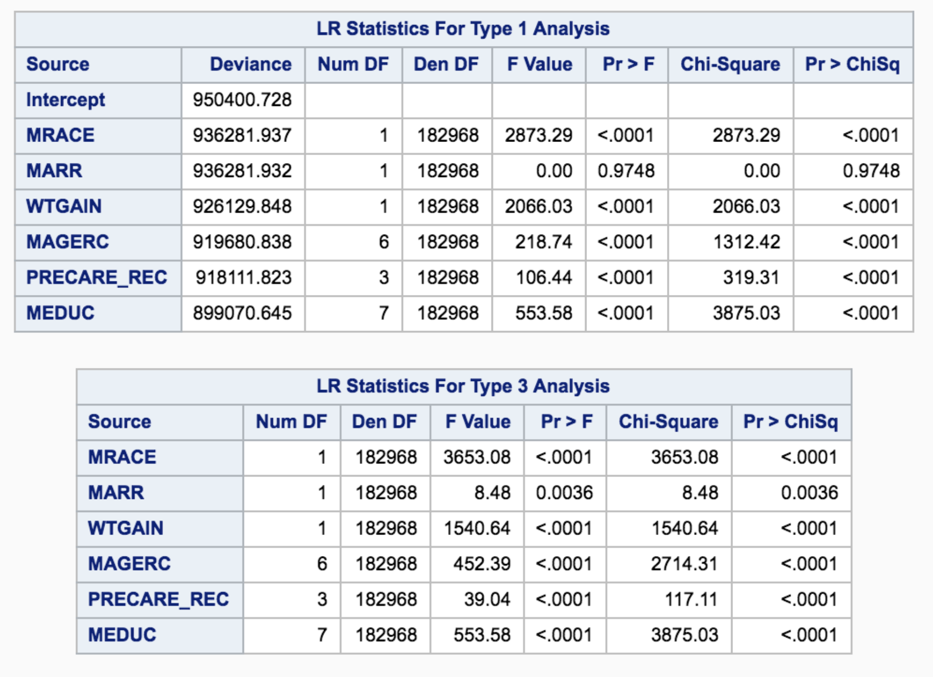
For the goodness of fit, the scaled deviance is 4.9138. If we compare the the deviance of 899070.6448 with its asymptotic chi-sqaure with 180000 degrees of freedom distribution, we will find p-value is < 0.0001. Hence, it indicates the the specified model does not fit the data well. The confidence interval includes 1 for the following pairs: mother's age (3-6 vs 7), beginning prenatal care period (3 vs 4) mother's education level (4-5 vs 8). Hence, except for the pairs above, the levels of the predictors are significant versus their reference levels.

The parameter estimate can be interpreted as follows. The predicted log count of average of the daily cigarette for mother's race = 0 is 0.3493 higher than the predicted log count for mother's race = 1. Rest of parameter estimates can be interpreted in the same way.

To conclude, mothers consuming more average of the daily cigarette have following feature: race = white , marital status = not married, low weight gain, age = 4-44 yrs, beginning prenatal care period = no prenatal care, and mother's education level = middle school.

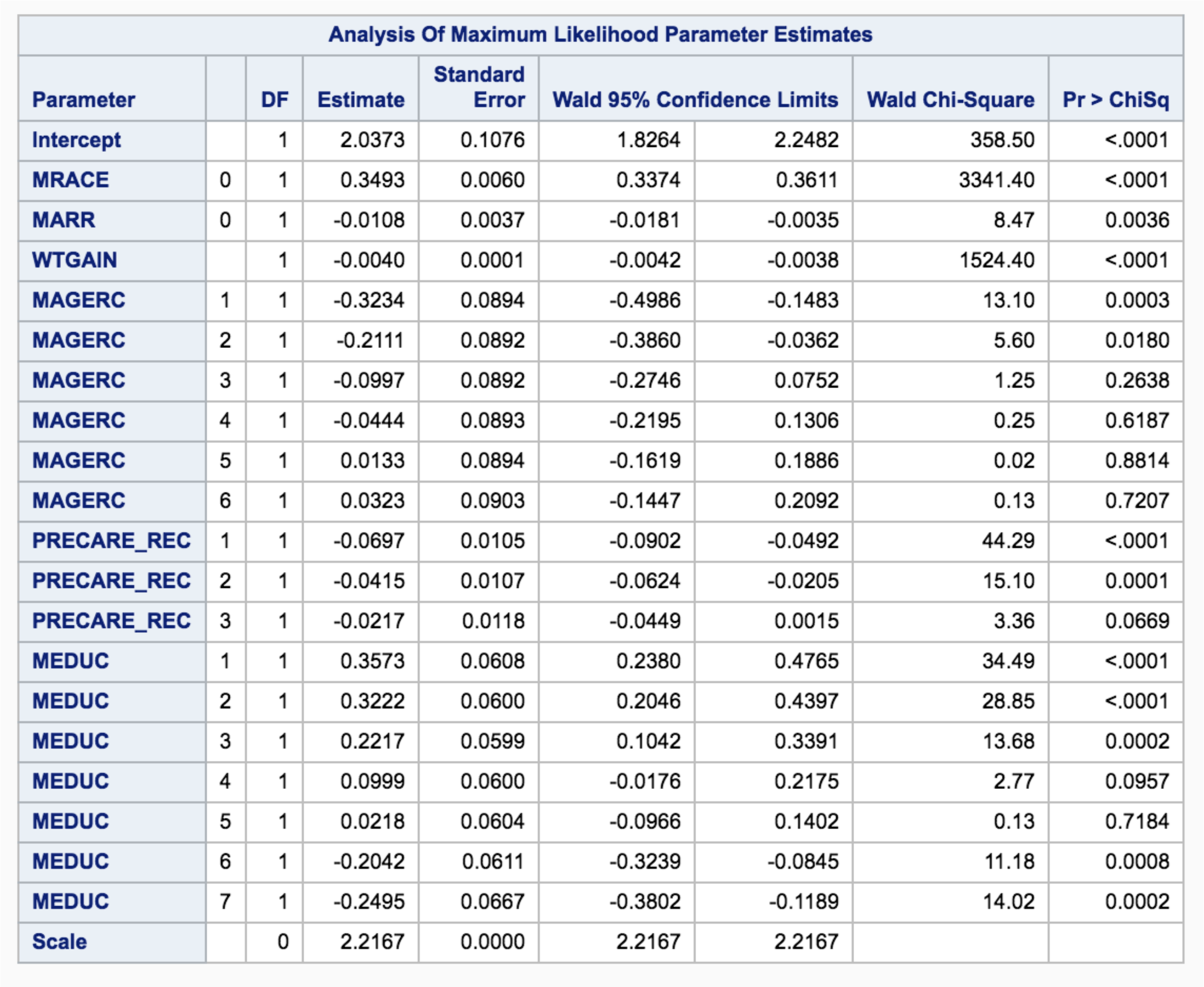


Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	18E4	899070.6448	4.9138
Scaled Deviance	18E4	182968.0000	1.0000
Pearson Chi-Square	18E4	1035031.5823	5.6569
Scaled Pearson X2	18E4	210637.1281	1.1512
Log Likelihood		423415.2972	
Full Log Likelihood		-791457.3627	
AIC (smaller is better)		1582954.7254	
AICC (smaller is better)		1582954.7300	
BIC (smaller is better)		1583157.0689	



LR Statistics For Type 1 Analysis							
Source	**Deviance**	**Num DF**	**Den DF**	**F Value**	**Pr > F**	**Chi-Square**	**Pr > ChiSq**
Intercept	950400.728						
MRACE	936281.937	1	182968	2873.29	<.0001	2873.29	<.0001
MARR	936281.932	1	182968	0.00	0.9748	0.00	0.9748
WTGAIN	926129.848	1	182968	2066.03	<.0001	2066.03	<.0001
MAGERC	919680.838	6	182968	218.74	<.0001	1312.42	<.0001
PRECARE_REC	918111.823	3	182968	106.44	<.0001	319.31	<.0001
MEDUC	899070.645	7	182968	553.58	<.0001	3875.03	<.0001

LR Statistics For Type 3 Analysis						
Source	**Num DF**	**Den DF**	**F Value**	**Pr > F**	**Chi-Square**	**Pr > ChiSq**
MRACE	1	182968	3653.08	<.0001	3653.08	<.0001
MARR	1	182968	8.48	0.0036	8.48	0.0036
WTGAIN	1	182968	1540.64	<.0001	1540.64	<.0001
MAGERC	6	182968	452.39	<.0001	2714.31	<.0001
PRECARE_REC	3	182968	39.04	<.0001	117.11	<.0001
MEDUC	7	182968	553.58	<.0001	3875.03	<.0001



Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	2.0373	0.1076	1.8264	2.2482	358.50	<.0001
MRACE	0	1	0.3493	0.0060	0.3374	0.3611	3341.40	<.0001
MARR	0	1	-0.0108	0.0037	-0.0181	-0.0035	8.47	0.0036
WTGAIN		1	-0.0040	0.0001	-0.0042	-0.0038	1524.40	<.0001
MAGERC	1	1	-0.3234	0.0894	-0.4986	-0.1483	13.10	0.0003
MAGERC	2	1	-0.2111	0.0892	-0.3860	-0.0362	5.60	0.0180
MAGERC	3	1	-0.0997	0.0892	-0.2746	0.0752	1.25	0.2638
MAGERC	4	1	-0.0444	0.0893	-0.2195	0.1306	0.25	0.6187
MAGERC	5	1	0.0133	0.0894	-0.1619	0.1886	0.02	0.8814
MAGERC	6	1	0.0323	0.0903	-0.1447	0.2092	0.13	0.7207
PRECARE_REC	1	1	-0.0697	0.0105	-0.0902	-0.0492	44.29	<.0001
PRECARE_REC	2	1	-0.0415	0.0107	-0.0624	-0.0205	15.10	0.0001
PRECARE_REC	3	1	-0.0217	0.0118	-0.0449	0.0015	3.36	0.0669
MEDUC	1	1	0.3573	0.0608	0.2380	0.4765	34.49	<.0001
MEDUC	2	1	0.3222	0.0600	0.2046	0.4397	28.85	<.0001
MEDUC	3	1	0.2217	0.0599	0.1042	0.3391	13.68	0.0002
MEDUC	4	1	0.0999	0.0600	-0.0176	0.2175	2.77	0.0957
MEDUC	5	1	0.0218	0.0604	-0.0966	0.1402	0.13	0.7184
MEDUC	6	1	-0.2042	0.0611	-0.3239	-0.0845	11.18	0.0008
MEDUC	7	1	-0.2495	0.0667	-0.3802	-0.1189	14.02	0.0002
Scale		0	2.2167	0.0000	2.2167	2.2167		