Biostats HW#3

Total marks: 137/160

1.) 10/10
**Methods:** Descriptive statistics are provided for mothers with infants that were small for gestational age (SGA), mothers with babies that were not small for gestational age, and for all mothers combined. For continuous variables, the mean, standard deviation, minimum, and maximum are provided. For categorical variables, percentages are provided. For subjects with missing values, these subjects were excluded only from aspects of the analysis that included the variable for which the value is missing.

**Results:**

There were a total of 755 mothers in the study. The number of missing values for each variable are as follows: mother’s height 6, gestational age 4, maternal smoking status 4, infant birth weight 4, and infant sex 4. There were no missing values for the following variables: small for gestational age, mother’s age, and parity. As can be seen in the table below, mothers with normal size infants were slightly taller on average, and also had slightly higher variability in height and a wider range in heights. Mothers with infants of normal size were slightly older. Mothers with normal size infants were more likely to be multiparous. Mothers with infants small for gestational age were more likely to be smokers. Mothers with normal size infants were more likely to have a boy. Finally, mothers with normal size infants had a higher mean gestational age, less variability in gestational age, a higher minimum gestational age, and a higher maximum gestational age.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Mothers with infants Small For Gestational Age****(n=105)** | **Mothers with infants Normal Size for Gestational Age (n=650)** | **All mothers****(n=755)** |
| **Maternal Height (cm)1** | 154.6 (5.87; 142-172) | 157.0 (6.54; 106-176) | 156.7 (6.50; 106-176) |
| **Maternal Age (yrs)1** | 23.8 (4.90; 16-35) | 24.9 (5.45; 14-43) | 24.8 (5.39; 14-43) |
| **Maternal Parity (%)** |  |  |  |
|  **0** | 46.7% | 37.5% | 38.8% |
|  **1** | 30.5% | 32.0% | 31.8% |
|  **2** | 13.3% | 18.3% | 17.6% |
|  **3** | 7.62% | 6.77% | 6.89% |
|  **4** | 0.95% | 3.38% | 3.05% |
|  **5** | 0.00% | 1.23% | 1.06% |
|  **6** | 0.95% | 0.77% | 0.79% |
| **Maternal smokers (%)** | 43.3% | 28.8% | 30.8% |
| **Infant birth weight (grams)1** | 2231 (411.6; 1035-3780) | 3246 (402.1; 2510-4730) | 3105 (534.4; 1035-4730) |
| **Male infant (%)** | 42.3% | 52.4% | 51.0% |
| **Gestational Age at Delivery (weeks)1** | 37.9 (2.20; 30-42) | 39.4 (1.24; 38-44) | 39.2 (1.50; 30-44) |

 1Descriptive statistics are: mean (standard deviation; minimum-maximum).

Percentages are column percentages.

2.) 10/10

**Methods:** I performed logistic regression analysis, with maternal smoking status as the predictor and log odds of giving birth to a small or normal for gestational age infant as the response. Parameters were exponentiated to calculate an odds ratio comparing maternal smokers versus non-smokers. 95% confidence intervals and 2 sided p-values were calculated using Wald statistics. Subjects with missing smoking status (n=4) were excluded from analysis.

A.) **Results:** From logistic regression analysis on 751 subjects, I estimate that the odds of giving birth to an infant that is small for gestational age is 1.89 higher in smoking mothers compared to non-smoking mothers. In other words, the odds are 89% higher in smoking mothers compared to non-smoking mothers. A 95% confidence interval for the odds ratio suggests that our data would not be surprising if the odds in the population are truly 1.24 to 2.89 times higher for smoking mothers compared to non-smoking mothers. A 2 sided p-value of 0.003 allows us to reject the null hypothesis that the odds of delivering an infant that is small for gestational age is the same in smoking and non-smoking mothers, in favor of the alternative hypothesis that the odds are higher in smoking mothers.

B.) 4/5

Odds of delivering a small for gestational infant in non-smokers:

Odds =eB0 = e-2.055861= 0.128

Probability of delivering a small for gestational infant in non-smokers:

Probability = odds / (1+odds) = 0.128 / 1.128 = 0.113= 11.3%

Odds of delivering a small for gestational infant in smokers:

Odds= eB0 + eB1= 0.128 x e.6367768= 0.128 x 1.890 = 0.241

Probability of delivering a small for gestational infant in smokers:

Probability = odds / (1+odds) = 0.241/1.241 = 0.194 = 19.4%

While I chose to condition on SGA status in my descriptive statistics rather than on smoking status, I could have also chosen to condition on smoking status. If I had done so, a portion of the table would have included the following information:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Smoker | Non-smoker | All  |
| Infant born small for gestational age (%)  | 19.5% | 11.4%  | 13.9% |

As can be seen in the above table, the described frequency of delivering a SGA baby for smokers and non-smokers are exactly the same as the calculated probabilities from the logistic regression model. This is expected, as a saturated model should yield values equivalent to the sample values.

(describe scientifically)

C 5/10

i.) In this model, the estimates for the odds ratio and 95% confidence interval would be the inverse of the estimates from the original model (which inverse? Multiplicative? Additive?). The p value would not change. The intercept would be different (different how?).

Cii.) In this model, the estimates for the odds ratio, 95% confidence interval, and intercept would be the inverse of the estimates in the original model (which inverse?). The p value would not change.

Ciii.) In this model, the estimates for the odds ratio, 95% confidence interval, and p value would all be the same. The intercept would be different (how so?).

3.)

**Methods:** I performed linear regression analysis, with maternal smoking status as the predictor and probability of giving birth to a small for gestational age infant as the response. Robust standard errors were calculated using the Huber-White sandwich estimator to allow for unequal variance. A point estimate for the difference in probabilities across smoking groups was estimated from the slope parameter of the regression model. A 95% confidence interval and 2 sided p-value were calculated using Wald statistics. Estimated odds were calculated from the obtained estimates of probabilities. Subjects with missing smoking status (n=4) were excluded from analysis.

A.) 10/10 **Results:** From linear regression on 751 subjects, the probability of delivering a small for gestational age baby is 0.0813 higher in smokers compared to non-smokers. A 95% confidence interval suggests that our data would not be surprising if the true difference in probability was 0.0233 to 0.139 higher in smokers. These probability values correspond to an estimated odds of 0.0885 higher in smokers (95% CI 0.0239 to 0.161). A 2 sided p value of 0.006 allows us to reject the null hypothesis of no difference in probabilities (or odds) between groups, in favor of the alternative hypothesis that the probability (or odds) is higher in smokers.

B.) 4/5

Probability of delivering a small for gestational infant in non-smokers, determined from the intercept of the regression model:

Probability = 0.113

Odds of delivering a small for gestational infant in non-smokers:

Odds = 0.113 / (1-0.113) = 0.127

Probability of delivering a small for gestational infant in smokers:

Probability = 0.113 + 0.0813 = 0.194

Odds of delivering a small for gestational infant in smokers:

Odds= 0.194 / (1-0.194) = 0.241

The fitted values for the probability of delivering a SGA are exactly the same as the actual proportion of SGA infants in the sample. This is expected, because this is a saturated model. (describe scientifically)

C 8/10

i.) In this model, the intercept is the value obtained by adding the intercept and slope in the original model. The point estimate and 95% confidence interval is the same for the slope coefficient, except the sign is changed (from positive to negative) for the values. The p value is the same.

Cii.) In this model, the intercept is the value obtained from subtracting the original model intercept from 1. The point estimate and 95% confidence interval is the same for the slope coefficient, except the sign is changed (from positive to negative) for the values. The p value is the same.

Ciii.) In this model, the intercept is the value obtained from subtracting the original model intercept and the original model slope from 1. The point estimate and 95% confidence interval is the same for the slope coefficient The p value is the same.

4.)

**Methods:** I performed Poisson regression analysis, with maternal smoking status as the predictor and the log rate (or probability) of giving birth to a small for gestational age infant as the response. Parameters were exponentiated to calculate a rate ratio comparing maternal smokers versus non-smokers. Robust standard errors were calculated using the Huber-White sandwich estimator to allow for unequal variance. A point estimate for the ratio in rates (or probabilities) between smoking mothers and non-smoking mothers smoking was estimated from the slope parameter of the regression model. A 95% confidence interval and 2 sided p-value were calculated using Wald statistics. The probability of giving birth to a small for gestational age infant in non-smokers was estimated from the regression intercept, and the probability of giving birth to a small for gestational age infant in smokers was estimated using this estimate for non-smokers and the estimated rate ratio. Estimated odds were calculated from the obtained estimates of probabilities. Subjects with missing smoking status (n=4) were excluded from analysis.

A. 9/10

**Results:** From Poisson regression on 751 subjects, the probability ratio of delivering a small for gestational age infant in smoking mothers compared to non-smoking mothers is 1.72. (also include percentage terms, ie. “the risk is 72% higher in…”) A 95% confidence interval suggests that our data would not be surprising if the true probability ratio is between 1.20 and 2.45. A 2 sided p value of 0.003 allows us to reject the null hypothesis that the rate ratio is 1, in favor of the alternative hypothesis that the ratio is greater than 1 (with the rate being higher in smoking mothers).

B. 4/5

Probability of delivering a small for gestational infant in non-smokers, determined from the intercept of the regression model:

Probability = 0.113

Odds of delivering a small for gestational infant in non-smokers:

Odds = 0.113 / (1-0.113) = 0.127

Probability of delivering a small for gestational infant in smokers:

Probability = RR X probability in non-smokers = 1.716 X 0.113 = 0.194

Odds of delivering a small for gestational infant in smokers:

Odds= 0.194 / (1-0.194) = 0.241

The fitted values for the probability of delivering a SGA are exactly the same as the actual proportion of SGA infants in the sample. This is expected, because this is a saturated model. (explain scientifically)

C.) 6/10

 For part C, and am comparing the two models assuming that values were already exopentiated (ie comparing stata output for poisson sga smoker, irr) .

Ci.) In this model, the intercept is the product of the original model intercept times the original model rate ratio. The point estimate and 95% confidence interval for the slope coefficient is the inverse (which inverse?) of the original model values. The p value is the same.

Cii.) In this model, the intercept is the value obtained from subtracting the original model point estimate for the intercept from 1 (not exactly). The point estimate and 95% confidence interval for the slope coefficient are different, as is the obtained p value.

Ciii.) In this model, the intercept, point estimate and slope for the slope coefficient, and p value are all different from the original model.

5.) 10/10

The logistic regression analysis in problem 2 corresponds to the chi squared test. The chi squared statistic and p value obtained from the two tests should be the same. The chi squared test does not directly give an estimate of the odds ratio, or a 95% confidence interval for this ratio.

The linear regression analysis using robust standard errors in problem 3 corresponds to the t test that allows for equal variances. The point estimate for the difference in proportions should be the same, but the p value and 95% confidence intervals will be slightly different due to differences in calculating standard errors.

The Poisson regression analysis corresponds to the two sample test of poisson rates. The point estimate for the difference in rates, p value, and 95% confidence interval should all be the same.

6a.) 10/10

**Methods:** I performed linear regression analysis, with maternal age as the predictor and probability of giving birth to a small for gestational age infant as the response. Robust standard errors were calculated using the Huber-White sandwich estimator to allow for unequal variance. A point estimate for the difference in probabilities between women differing by 1 year in age was estimated from the slope parameter of the regression model. A 95% confidence interval and 2 sided p-value were calculated using Wald statistics.

**Results:** From linear regression on 755 subjects, for each 1 year increase in maternal age, the probability of giving birth to a small for gestational age infant decreases by 4.52%. A 95% confidence interval suggests that our data would not be surprising if the true decrease in probability for each 1 year increase in maternal age is between 0.0286% and 8.74%. A two sided p value of 0.036 allows us to reject the null hypothesis of no difference in probability of giving birth to a small for gestational age infant, in favor of the null hypothesis that the probability of giving birth to a small for gestational age is lower in older mothers.

6b.) 10/10

**Methods:** I performed Poisson regression analysis, with maternal age as the predictor and the log rate (or probability) of giving birth to a small for gestational age infant as the response. Robust standard errors were calculated using the Huber-White sandwich estimator to allow for unequal variance. A point estimate for the ratio of rates (or probabilities) between mothers differing in age by 1 year was estimated from the exponentiated slope parameter of the regression model. A 95% confidence interval and 2 sided p-value were calculated using Wald statistics.

**Results:** From Poisson regression on 755 subjects, the RR of giving birth to a small for gestational age infant for mothers differing in age by 1 year is 0.966, with older mothers having the lower probability. A 95% confidence interval suggests that our data would not be surprising if the true RR is between 0.934 and 0.999. A two sided p value of 0.046 allows us to reject the null hypothesis that the RR is 1, in favor of the alternative hypothesis that the RR is lower than 1, with the probability in older mothers being lower.

6c.) 8/10

**Methods:** I performed logistic regression analysis, with maternal age as the predictor and log odds of giving birth to a small or normal for gestational age infant as the response. Robust standard errors were calculated using the Huber-White sandwich estimator to allow for unequal variance. Parameters were exponentiated to calculate an odds ratio comparing mothers that differ in age by 1 year. 95% confidence intervals and 2 sided p-values were calculated using Wald statistics.

**Results:** From logistic regression on 755 subjects, the OR of giving birth to a small for gestational age infant for mothers differing in age by 1 year is 0.961, with older mothers having the lower odds. A 95% confidence interval suggests that our data would not be surprising if the true OR is between 0.924 and 0.999. A two sided p value of 0.046 allows us to reject the null hypothesis that the OR is 1, in favor of the alternative hypothesis that the OR is lower than 1, with the odds in older mothers being lower. (p-value should be higher)

6d.) Estimates of the probability that a 20 year old mother would have a SGA infant:

9/10

Estimate from linear regression:

Probability = B0 + B1X = 0.251 + 20(-0.00452) = 0.1607

Estimate from Poisson regression:

Probability= eB0 X (RR20)= 0.321 X 0.96620= 0.1613

Estimate from logistic regression: Odds = eB0 X eB1X20= 0.426 X e-0.795572= 0.191

Probability = odds / (1+odds) = 0.1613

These values are all very similar, but differ from the sample proportion of 3/40=0.075. The fitted values from the models are not expected to exactly match the sample proportion because these are not saturated models.

7a.) 10/10 As seen in the below graph, the fitted values for estimated probability of SGA for each regression model are very similar, with slight differences that are more pronounced at the extremes of maternal age.



8a.) 6/10

**Methods:** I performed logistic regression analysis, with log maternal age as the predictor and log odds of giving birth to a SGA infant as the response variable. Robust standard errors were calculated using the Huber-White sandwich estimator to allow for unequal variance. Parameters were exponentiated in order to obtain an estimate of the odds ratio comparing mothers that differ in age by log 1 year. 95% confidence intervals and 2 sided p-values were calculated using Wald statistics.

**Results:** From logistic regression on 755 subjects, the estimated OR of giving birth to a small for gestational age infant for mothers differing in is 0.385, (when log-transforming, express inference on multiplicative scale NOT log-additive scale; see key for details) with older mothers having lower odds. A 95% confidence interval suggests that our data would not be surprising if the true OR is between 0.1455and 1.010. A two sided p value of 0.052 does not allow us to reject the null hypothesis that the OR is 1.

8b.) 4/5

 It would be reasonable if you think the association between maternal age and probability of delivering a baby that is SGA is multiplicative. It would be silly if you think the association is additive. It is important to note that the range of ages in our sample is not very large, so it may be difficult to use this data to make inference about the effect of age on a multiplicative scale (why?). This is another reason this analysis may be deemed “silly.”