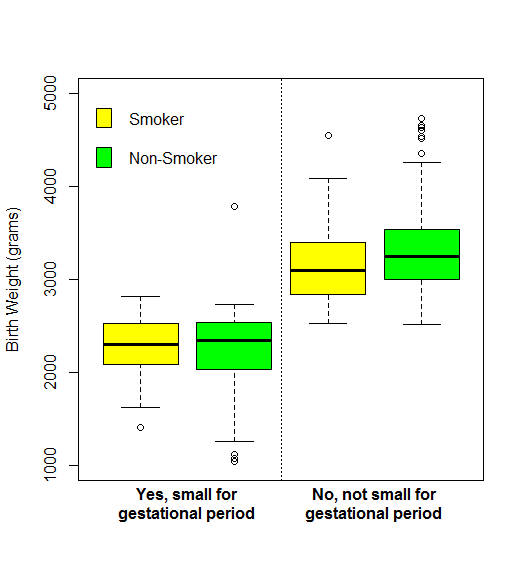
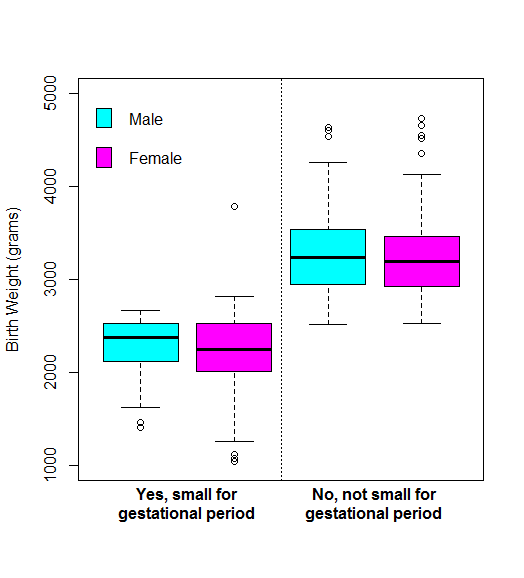
**Total Grade: 74/160 : Question 1**

*Methods*: The following analyzes are interested in the variables associated with delivery of babies that are small for gestational age (SGA). Therefore, the data was divided into babies that were small for gestational age and babies that were not small for gestational age. Descriptive statistics were performed on these groups, as well as the dataset as a whole.

*Inference*: The dataset had data for 755 babies: 650 that were not small and 105 that were small for their gestational age. The birth weight for these babies that were small for their gestational age was lower than those not small for gestational age. Similarly, these babies that were small had lower means for mother’s height and age. The small babies also had a higher percentage of mothers that smoked. The majority of these babies were female, whereas the majority of the not small babies were male. There were also less number of prior deliveries for this small group and they had a lower mean value for gestational age at delivery. This is further depicted below in the table and illustrated in the boxplots.

|  |  |  |  |
| --- | --- | --- | --- |
|  | ***Small for Gestational Age*** | | |
|  | **No**  (n = 650) | **Yes**  (n = 105) | **All**  (n = 755) |
| *Mothers Height (cm) 1* | 157 (6.54;  106-176) | 154.6 (5.87;  142-172) NA’s = 6 | 156.7 (6.50;  106-176) NA’s = 6 |
| *Mothers Age (years) 1* | 24.94 (5.45;  14.0-43.00) | 23.85 (4.90;  16-35) | 24.79 (5.38;  14-43) |
| *Number of Prior Deliveries 1* | 1.132 (1.23;  0-6) | 0.90 (1.11;  0-6) | 1.10 (1.21;  0-6) |
| *Smoker* | 28.7% | 43.3% | 30.8% |
| *Birth Weight (gms) 1* | 3246 (402.13;  2510-4730) NA’s = 3 | 2231 (411.60;  1035-3780) NA’s = 1 | 3106 (534.46;  1035-4730) NA’s = 4 |
| *Male* | 52.4% | 42.3% | 51.0% |
| *Gestational Age at Delivery (weeks) 1* | 39.38 (1.24;  38-44) NA’s = 3 | 37.92 (2.20;  30-42) NA’s = 2 | 39.18 (1.50;  30-44) NA’s = 5 |

Note: 1 Mean (SD; min-max)

**Question 2a**

*Methods*: A logistic regression model was run on the binary variable SGA (dependent) and independent binary variable smoker. The confidence interval was calculated using 1.96 \* the standard error.

*Inference*: The equation for delivery of infants who are small for gestational age (SGA) is: log odds SGA = -2.0559 + 0.6368 \* Smoker. The odds of delivering a SGA infant are 89.0% higher for smokers than non-smokers. A 95% CI suggests that this observation is not unusual if smokers have the odds of delivering a small baby anywhere from 46.6% and 131.4% higher than non-smokers, for p = 0.003.

**Question 2b**

*Methods*: The logistic regression equation from above was used to determine the odds for smokers and nonsmokers. These odds were then used to calculate the probabilities of the two groups.

*Inference*: The odds of SGA for smokers is 0.2419 (e^-.2.0559 \* e^0.6368); the probability is 0.1947 (0.2419/.2419+1). The odds of SGA for non-smokers is 0.1279 (e^-2.0559); the probability is 0.1133 (.1279/.1279+1). These estimates are similar to the descriptive statistics from above, which showed a larger proportion of mothers who smoked has babies SGA. Furthermore, the descriptive statistics suggested that smokers had on average babies with lower birth weights than mothers who did not smoke – where birth weight has been suggested to be associated with SGA.

**Question 2c**

*Methods*: Four regression equations were created and compared, which compared the different ways to fit a regression model on SGA as a response and maternal smoking as the predictor.

*Inference*: The four models provide the same odds, however through different means. The method used in 2a and iii both use nonsmokers as the default estimate – where the intercept provides the odds. Whereas ii and i use smoker as the default, where the intercept provides the odds for the smokers. However, 2a and i predict for the odds of SGA, while ii and iii predict the odds for not having SGA.

**Question 3**

*Methods*: A similar regression was performed as in part 2, however this evaluates the difference in probabilities for SGA across smoking groups, which is a risk difference problem using linear regression.

*Inference*: The difference in probabilities for SGA between smokers and nonsmokers was 8.13%, where smokers had a higher probability of SGA. For nonsmokers, the probability was estimated at 11.34%, while it was 19.47% for smokers. This was significant at p < 0.001.

**Question 4**

*Methods*: A similar regression as problem 2 was performed, however evaluating for an association between the odds of SGA by the ratio of probabilities across smoking groups. This was done using the Poisson regression.

*Inference*: The Poisson regression for SGA by the ratio of probabilities yielded an intercept of -2.1763 and slope of 0.5405 for the smoker variable. This means that the probability for smokers was 19.47% higher for SGA than non-smokers.

**Question 5**

The analyses performed in problems 2-4 yield estimates for the differences between the two groups (smokers vs non-smokers), where as the two sample comparison just tells whether or not a difference exists. Problems 2-4 not only shows a difference exists, but quantifies this difference.

**Question 6a**

*Methods*: A regression analysis of the distribution of the prevalence of SGA defined by maternal age was evaluating using risk difference, which uses linear regression.

*Inference*: The prevalence of SGA decreased by 0.45% for each year increase in maternal age, with an intercept at 25%. This association with age was significant at p=0.0537. This implies that the data would not be unusual for a 95% CI of -0.45 +/- 0.45 %.

**Question 6b**

*Methods*: A regression analysis of the distribution of the prevalence of SGA defined by maternal age was evaluating using risk ratio, which uses Poisson regression.

*Inference*: The Poisson regression for SGA by the maternal age yielded an intercept of -1.1359 and slope of -0.03442 for the age variable. This means that the probability for each one year increase in maternal age was associated with 3.4% decrease in SGA. However, age is only significant at p=0.0738.

**Question 6c**

*Methods*: A regression analysis of the distribution of the prevalence of SGA defined by maternal age was evaluating using odds ratio, which uses logistic regression.

*Inference*: The log odds of SGA = -0.85316 – 0.03978 \* age. Therefore the odds of delivering a SGA baby is 3.89% lower for each increase in age. However, age is only significant at p=0.0545. From this p, a 95% CI suggests that this observation is not unusual if the increase in age is associated with a decrease of 3.89% +/- 4% odds of SGA.

**Question 6d**

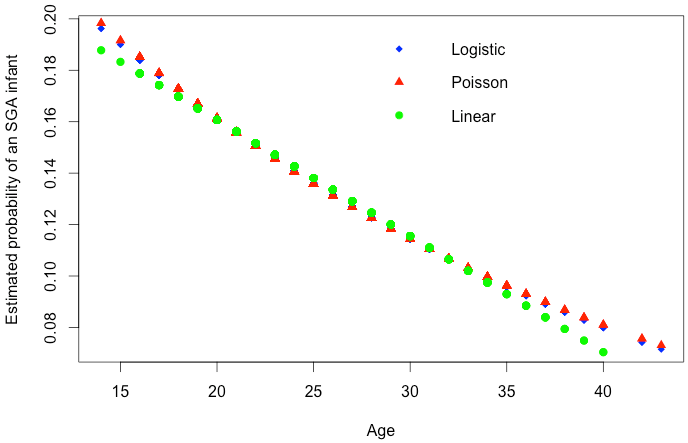
*Methods*: Each of the equations from a-c were used, with 20 years old in the age variable. Then these were compared.

*Inference*: For the linear regression, the predicted probability for a 20 year old mother to have a SGA infant is 16.07%. For the Poisson regression, this was 16.13%. For the logistic regression, it was 19.22%. The dataset included 40 mothers that were 20 years old. The proportion was heavy on the non SGA for these women, and had twice as many non-smokers as smokers. However, this was somewhat comparable to the spread across other ages, which is probably why these three regression estimates are similar.

**Question 7a**

*Methods*: A plot was created using the three types of regressions performed in problem 6. A sample proportion within each unique age was obtained by plotting the fitted values for each age, based on the mean obtained through each regression.

*Inference*: The three regressions yield similar fitted values, except differ the most at the ends of the age spectrum. For younger ages, the linear regression estimates lower probabilities, while the Poisson estimates higher probabilities of SGA. Similarly, for the older ages, the linear regression estimates lower probabilities, while the Poisson estimates higher probabilities. The logistic regression is fitted between these two regressions. The plot is on the following page.



**Question 7b**

*Methods*: The estimated probabilities for each age included in the dataset was calculated using the predict function in R based on the regression analyses fitted in problem 6.

*Inference*: The table below shows the predicted values for each age. In all cases, the younger mothers have higher probabilities of having an infant with SGA. The logistic regression has the highest predicted proabilities, followed by the Poisson, and the linear last.

|  |  |  |  |
| --- | --- | --- | --- |
| **Age** | **Linear** | **Logistic** | **Poisson** |
| 14 | 0.188 | 0.244 | 0.198 |
| 15 | 0.183 | 0.235 | 0.192 |
| 16 | 0.179 | 0.225 | 0.185 |
| 17 | 0.174 | 0.217 | 0.179 |
| 18 | 0.170 | 0.208 | 0.173 |
| 19 | 0.165 | 0.200 | 0.167 |
| 20 | 0.161 | 0.192 | 0.161 |
| 21 | 0.156 | 0.185 | 0.156 |
| 22 | 0.152 | 0.178 | 0.151 |
| 23 | 0.147 | 0.171 | 0.145 |
| 24 | 0.143 | 0.164 | 0.141 |
| 25 | 0.138 | 0.158 | 0.136 |
| 26 | 0.134 | 0.151 | 0.131 |
| 27 | 0.129 | 0.146 | 0.127 |
| 28 | 0.125 | 0.140 | 0.122 |
| 29 | 0.120 | 0.134 | 0.118 |
| 30 | 0.116 | 0.129 | 0.114 |
| 31 | 0.111 | 0.124 | 0.110 |
| 32 | 0.107 | 0.119 | 0.107 |
| 33 | 0.102 | 0.115 | 0.103 |
| 34 | 0.097 | 0.110 | 0.100 |
| 35 | 0.093 | 0.106 | 0.096 |
| 36 | 0.088 | 0.102 | 0.093 |
| 37 | 0.084 | 0.098 | 0.090 |
| 38 | 0.079 | 0.094 | 0.087 |
| 39 | 0.075 | 0.090 | 0.084 |
| 40 | 0.070 | 0.087 | 0.081 |
| 42 | 0.061 | 0.080 | 0.076 |
| 43 | 0.057 | 0.077 | 0.073 |

**Question 8a**

*Methods*: The maternal age was logarithmically transformed and then used to create a logistic regression analysis of the distribution of the prevalence of SGA infants across groups defined by this transformed age variable.

*Inference*:

The log odds of SGA = 1.202 – 0.9536 \* log(age). Therefore the odds of delivering a SGA baby is 3.85% lower for each increase in log of age. However, age is only significant at p=0.0584.

**Question 8b**

The untransformed variable for age was more appropriate. Although age was slightly skewed in our dataset, see the plots below; the use of untransformed age like in problem 6 c is optimal. This is a more typical application of age, as opposed to log transformed. Also, this distribution yields better to the logistic regression application.

