Biost 518: Applied Biostatistics II
Biost 515: Biostatistics II

Emerson, Winter 2015

Homework #3 Key

January 23, 2015

<u>Written problems:</u> To be submitted as a MS-Word compatible file to the class Catalyst dropbox by 9:30 am on Monday, February 2, 2014. See the instructions for peer grading of the homework that are posted on the web pages.

> *On this (as all homeworks) Stata / R code and unedited Stata / R output is **<u>TOTALLY</u>** unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*
>
> ***Unless explicitly told otherwise in the statement of the problem, in all problems requesting "statistical analyses" (either descriptive or inferential), you should present both***
>
> - ***<u>Methods:</u> A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE Stata OR R CODE.***
> - ***<u>Inference</u>: A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to "Reporting Associations" for details.***

This homework considers pregnancy outcomes in an observational study of women attending a prenatal clinic in South Africa. Questions in this homework focus most closely on association with delivery of babies that are small for gestational age (SGA). The data can be found on the class web page (follow the link to Datasets) in the file labeled pregout.txt (you will not need any of the longitudinal measurements in the file preglong.txt). Documentation is in the file pregnancy.pdf.

1. Provide suitable descriptive statistics relevant to this analysis.

<u>Instructions for grading</u>: This problem is worth 10 points. There must be suitable descriptive statistics and a discussion of what they would mean scientifically. Points to consider in the grading include:

- ***The analyses considered in this homework are associations between SGA and smoking (problems 2-5) and between SGA and maternal age (problems 4-8).***
    - ***For the purposes of "describing materials and methods", we would want to know about the types of women and pregnancies included in the dataset. This would argue that it is of interest to include information beyond the variables explicitly considered in the scientific questions.***
    - ***For the purposes of "assessing validity of assumptions", we would be interested in whether the associations we explore are confounded by other variables. We might also want to consider the other "baseline" variables associations with the POIs. Hence, at a minimum, the descriptive statistics must address associations between age and smoking. To the extent that it is easy to present, we might also want to consider parity and height (the only other maternal factors) and the baby's sex (given that it is***

determined at conception). (Note that birthweight and gestational age are other outcomes, and would not be considered as confounders.)

  - ***For the purposes of "preliminary straightforward estimates of associations" we would want to present information for both of the POIs considered.***

- ***Two basic approaches for presenting the descriptive statistics are to stratify by the "outcome" SGA or to stratify by the predictor of interest for problems 2-4 (smoking behavior) and/or the predictor of interest for problems 6-8 (maternal age). I find it useful to present both stratified descriptive statistics by smoking behavior and SGA. I did not present it stratified by age, thus I am not able to fully assess confounding of the age-SGA association by parity, height, and baby's sex.***
  - ***I note that this could have been placed in a 5 column table, though I did it in separate rows, thereby presenting the overall sample statistics twice.***

<u>Ans:</u> *Methods:* Summary statistics (mean, standard deviation, minimum, and maximum for continuous variables and percentages for binary variables) are presented for selected maternal factors (age, height, parity, smoking behavior) and pregnancy outcomes (baby's sex, birthweight, gestational age, and diagnosis of small for gestational age (SGA)). Descriptive statistics are presented separately within strata defined by smoking behavior and SGA diagnosis. Subjects missing data were omitted only if they were missing data for a variable needed for the specific analysis. Pursuant to this strategy, four subjects missing smoking behavior data are omitted from all aspects of the subtable stratified by smoking status.

***Results:* Data is available on singleton pregnancies (383 boys, 368 girls, 4 missing baby's sex) born live to 755 different mothers. As shown in Table 1 below, the mothers ranged in age from 14 to 43 years (mean 24.8 y, sd 5.39 y) and ranged in height from 106 to 176 cm (mean 157 cm, sd 6.5 cm). A plurality of mothers were nulliparous (38.8%), with 29.4% having had 2 or more prior births. Of the 751 women whose smoking behavior was recorded, 208 (30.8%) smoked during the pregnancy, with only very slight trends for the smokers to be older (means 25.1 vs 24.6 years) and higher parity (means 1.19 vs 1.06 prior live births).**

Estimated gestational age at birth ranged from 30 to 44 weeks (mean 39.2 w, sd 1.5 w), with birth weights ranging from 1,035 to 4,730 g (mean 3,106 g, sd 534 g). One hundred five (105) babies (16.0%) were judged SGA at delivery. Women delivering SGA babies tended to be younger (mean age 23.8 vs 24.9 years), tended to be very slightly smaller (mean height 155 vs 157 cm), were more likely to have had no prior live births (46.7% vs 37.5%), and had higher prevalence of smoking (43.3% vs 28.7%).

(In the tables, I put descriptive statistics for the untransformed smoking variable, only to point out that the mean of a binary variable can be used to estimate a proportion, even when it is not coded 0-1. It would have been easier if it had been, however. I <u>strongly</u> urge you not to include such descriptive statistics in your tables.)

Table 1: Descriptive statistics for selected maternal characteristics and pregnancy outcomes within strata defined by smoking behavior and small for gestational age (SGA) outcomes. Statistics presented for continuous variables are the sample mean (sample SD; minimum – maximum; number of observations with available measurements). Four subjects (one 20 yo, nulliparous woman delivering an SGA baby, and three women ages 21, 34, 38; parity 0, 1, 4 delivering non-SGA babies)are missing data for smoking behavior, gestational age, birthweight and babies' sex. They are thus excluded from all statistics when stratified by smoking status, but are represented in the statistics stratified by SGA.

	Non-smokers	**Smokers**	**All Subjects**
	(n = 520)	**(n=231)**	**(n= 751)**
Maternal age (y)	**24.6 (5.37; 14 - 43; n=520)**	**25.1 (5.35; 15 - 42; n=231)**	**24.8 (5.36; 14 - 43; n=751)**
Height (cm)	**157 (6.2; 127 - 175; n=515)**	**157 (7.2; 106 - 176; n=230)**	**157 (6.5; 106 - 176; n=745)**
Parity	**1.06 (1.19; 0 - 6; n=520)**	**1.19 (1.27; 0 - 6; n=231)**	**1.10 (1.21; 0 - 6; n=751)**
Nulliparous (n (%))	**208 (40.0%)**	**83 (35.9%)**	**291 (38.8%)**
1 prior live birth (n (%))	**166 (31.9%)**	**73 (31.6%)**	**239 (31.8%)**
2 prior live births (n (%))	**90 (17.3%)**	**43 (18.6%)**	**133 (17.7%)**
3+ prior live births (n (%))	**56 (10.8%)**	**32 (13.9%)**	**88 (11.7%)**
Small for Gestational Age	**11.3% (n=520)**	**19.5% (n=231)**	**13.8% (n=751)**
Gestational age (wks)	**39.3 (1.55; 30 - 44; n=520)**	**39.0 (1.36; 33 - 43; n=230)**	**39.2 (1.50; 30 - 44; n=750)**
Birth weight (g)	**3,165 (534; 1,035 - 4,730; n=520)**	**2,972 (512; 1,410 - 4,550; n=231)**	**3,106 (534; 1,035 - 4,730; n=751)**
sex (1=M, 2=F)	1.48 (0.50; 1 - 2; n=520)	1.52 (0.50; 1 - 2; n=231)	1.49 (0.50; 1 - 2; n=751)
Female (%)	**47.7% (n=520)**	**51.9% (n=231)**	**49.0% (n=751)**

	Not Small for Gestational Age	**Small for Gestational Age**	**All Subjects**
	(n = 650)	**(n=105)**	**(n= 755)**
Maternal age (y)	**24.9 (5.45; 14 - 43; n=650)**	**23.8 (4.90; 16 - 35; n=105)**	**24.8 (5.39; 14 - 43; n=755)**
Height (cm)	**157 (6.5; 106 - 176; n=650)**	**155 (5.9; 142 - 172; n=99)**	**157 (6.5; 106 - 176; n=749)**
Parity	**1.13 (1.23; 0 - 6; n=650)**	**0.90 (1.11; 0 - 6; n=105)**	**1.10 (1.21; 0 - 6; n=755)**
Nulliparous (n (%))	**244 (37.5%)**	**49 (46.7%)**	**293 (38.8%)**
1 prior live birth (n (%))	**208 (32.0%)**	**32 (30.5%)**	**240 (31.8%)**
2 prior live births (n (%))	**119 (18.3%)**	**14 (13.3%)**	**133 (17.6%)**
3+ prior live births (n (%))	**79 (12.2%)**	**10 (9.5%)**	**89 (11.8%)**
smoker (1=y,2=n)	1.71 (0.453; 1 - 2; n=647)	1.57 (0.498; 1 - 2; n=104)	1.69 (0.462; 1 - 2; n=751)
Maternal smoking (%)	**28.7% (n=647)**	**43.3% (n=104)**	**30.8% (n=751)**
Gestational age (wks)	**39.4 (1.24; 38 - 44; n=647)**	**37.9 (2.20; 30 - 42; n=103)**	**39.2 (1.50; 30 - 44; n=750)**
Birth weight (g)	**3,246 (402; 2,510 - 4,730; n=647)**	**2,231 (412; 1,035 - 3,780; n=104)**	**3,106 (534; 1,035 - 4,730; n=751)**
sex (1=M, 2=F)	1.48 (0.50; 1 - 2; n=647)	1.58 (0.50; 1 - 2; n=104)	1.49 (0.50; 1 - 2; n=751)
Female (%)	**47.6% (n=647)**	**57.7% (n=104)**	**49.0% (n=751)**

Some important points about the answers to problems 2 – 4

- ***The smoking variable was originally coded as 1=yes, 2= no. As we have learned in lecture, this has no impact on the scientific validity of the analysis. But a student who did not notice this coding could have made major mistakes in the interpretation of the results.***

- ***I personally recoded the variable before fitting a regression model. Such "re-parameterization" of the model has no impact on the scientific validity of the analysis, and it does not even have to be mentioned in the methods, if the regression parameters are interpreted correctly and reported using scientifically important words (e.g., "smokers" vs "nonsmokers" rather than referring to such statistical jargon as "slope parameters" or "a 1 unit difference in the covariate").***
- ***This ability to regard*** <u>*linear transformations of our predictor data*</u> ***as merely a re-parameterization of our regression models is a very important property: We will often find it convenient to do so in order to have the statistical software compute estimates that are more scientifically meaningful. These linear transformations are a very special (and simple) case of the full generality of re-parameterized models:***
  - ***We consider a regression model of response Y on a single covariate X: our general regression model involves some summary measure θ of the distribution of our response variable, a linear predictor η that is a linear combination of specified covariates to be included in our model, and a link function g() that relates the two:***

$$g\left(\theta_{Y|X}\right)=\eta_X\ =\beta_0+\beta_1\times X$$

  - ***We sometimes consider a linear transformation of X in which we are allowed to scale it by an arbitrary constant b and shift it by an arbitrary constant a: W = a + bX. We then can consider a re-parameterized model***

$$g\left(\theta_{Y|X}\right)=\eta_X\ =\gamma_0+\gamma_1\times W$$

  - ***In the above notation for the re-parameterized model, you will note that I claim that the linear predictor is still the same even as I use W instead of X. To see that this is possible, we can substitute the expression for W into this second equation:***

$$\begin{aligned} g\left(\theta_{Y|X}\right)=\eta_X\ &=\gamma_0+\gamma_1\times W \\ &=\gamma_0+\gamma_1\times\left(a+b\times X\right) \\ &=\gamma_0+\gamma_1\times a+\gamma_1\times b\times X \\ &=\left(\gamma_0+\gamma_1\times a\right)+\left(\gamma_1\times b\right)\times X \\ &=\qquad \beta_0 \qquad + \quad \beta_1 \quad \times X \end{aligned}$$

  - ***Thus in this case we have a very straightforward re-parameterization of the model that allows us to easily describe the correspondences between the regression parameters:***

$$\beta_0=\left(\gamma_0+\gamma_1\times a\right) \qquad\qquad \beta_1=\left(\gamma_1\times b\right)$$

  - ***Given that the models are easily re-expressed, it should not be surprising that in these cases, the fitted linear predictors will also be the same. It is that property that truly defines what we regard as a "re-parameterization": For any value of the predictor, we get the same fitted value no matter which model we fit:***

$$\hat{\eta}_X\ =\hat{\beta}_0+\hat{\beta}_1\times W=\hat{\gamma}_0+\hat{\gamma}_1\times W$$

  - ***The above results hold no matter which of our regression models we consider (e.g., linear, logistic, Poisson, proportional hazards, accelerated failure time) and no matter how many covariates we are modeling, so long as the regression model is based on a linear predictor that is "linked" to some summary measure. These linear transformations of our covariates are just changing the units of comparison.***

- ***Now our formal definition of a "re-parameterized" model focuses on the invariance of the fitted values. In the above, I focused on linear transformations of our predictors, though there are many other re-parameterizations that will be important to us as we consider more complicated modeling of our predictors (e.g., dummy variables, polynomials, linear splines) and inclusion of interactions. In those settings, it will sometimes be difficult to convince ourselves that one model is in fact a re-parameterization. In a separate document to be posted under Supplementary Documents on the class web pages, I will describe in greater detail***
    - ***our definition of an "alternative parameterization" of the same model ,***
    - ***some examples of more complicated alternative parameterizations, and***
    - ***what parallels there will be between the regression parameters across two different parameterizations of a model.***
- ***Given our focus on the fitted values to define a "re-parameterized model", it should not be surprising that transformations of our response variable will be more difficult to characterize in this terminology. In particular, commonly used transformations will nearly always change the summary measure θ of the distribution of response variable used in the regression model, so we will in most cases focus on our ability to transform the response variable without affecting our measures of association, rather than our ability to have our fitted values unchanged. Our characterization of our ability to transform the response variable will in general depend on the summary measure being considered, the type of inference (parametric, semi-parametric, or distribution-free), and the type of transformation. There are three levels at which we might describe correspondences among models involving transformed response:***
    - ***settings in which a specific class of transformations of the response variable has no effect whatsoever on the estimates of the association and thus has no effect on any statistical inference about associations,***
    - ***settings in which a specific class of transformations induce straightforward transformations of the fitted values and thus have easily interpretable effects on the estimates of association and no effect on statistical significance about associations, and***
    - ***settings in which a specific class of transformations does not allow defining any correspondence between the fitted values or between estimates of associations, but the statistical significance of the estimated associations are expected to be the same (at least in large samples).***
- ***When analyzing differences in mean response with linear regression***
    - ***Linear transformations Z = c + dY induce linear transformations $\theta_{Z|X}$=c + d $\theta_{Y|X}$ so there are straightforward transformations of the fitted values and estimates of the association. There will be no effect on statistical significance about associations.***

$$
\begin{aligned}
g\left(\theta_{Y|X}\right) &= E\left(Y \mid X\right) = \mu_X = \beta_0 + \beta_1 \times X \\
g\left(\theta_{Z|X}\right) &= E\left(Z \mid X\right) = \eta_X = \gamma_0 + \gamma_1 \times X \\
&= E\left(c + dY \mid X\right) = c + dE\left(Y \mid X\right) = c + d \times \mu_X \\
&= c + d \times \left(\beta_0 + \beta_1 \times X\right) \\
&= \left(c + d \times \beta_0\right) + \left(d \times \beta_1\right) \times X
\end{aligned}
$$

$$
\Rightarrow \qquad \gamma_0 = c + d \times \beta_0 \qquad\qquad \gamma_1 = d \times \beta_1
$$

  - ***With binary response, changing our definition of what is an event (e.g., from survival =1 to death = 1) can be viewed as a very special linear transformation Z = 1 – Y. It should thus be clear that changing that point of reference will have no impact on our scientific interpretation or statistical inference: We can, for instance, easily transform our statements about the difference in probability of death within 5 years to statements about the difference in probability of survival for 5 years, without even fitting another model.***
  - ***All other transformations of the response will generally affect the mean in complicated ways, and thus those other transformations will represent a major change in the scientific question being addressed.***
    - ***(Consider for instance how the log transformation of response changes the scientific question to one about ratios of geometric means, which might show a qualitatively different association—the mean of Group 1 might be higher, while the geometric mean of Group 2 might be higher.)***
- ***When analyzing ratios of geometric mean response with linear regression on log Y***
  - ***Power transformations*** $Z = c\ Y^d$ ***with*** $c > 0$ ***induce power transformations*** $\theta_{Z|X} = c\ \theta^d_{Y|X}$ ***so there are straightforward transformations of the fitted values and estimates of the association. There will be no effect on statistical significance about associations.***

$$
\begin{aligned}
g(\theta_{Y|X}) &= \log(GM(Y \mid X)) = E(\log Y \mid X) = \mu_X = \beta_0 + \beta_1 \times X \\
g(\theta_{Z|X}) &= \log(GM(Z \mid X)) = E(\log Z \mid X) = \eta_X = \gamma_0 + \gamma_1 \times X \\
&= E(\log(c) + d\log(Y) \mid X) = \log(c) + dE(\log Y \mid X) \\
&= \log(c) + d \times \mu_X = \log(c) + d \times (\beta_0 + \beta_1 \times X) \\
&= (\log(c) + d \times \beta_0) + (d \times \beta_1) \times X
\end{aligned}
$$

$$
\Rightarrow \qquad \gamma_0 = \log(c) + d \times \beta_0 \qquad\qquad \gamma_1 = d \times \beta_1
$$

  - ***All other transformations of the response will generally affect the mean in complicated ways, and thus those other transformations will represent a major change in the scientific question being addressed.***
- ***When analyzing ratios of odds of response with logistice regression with a binary response variable Y, the only relevant transformation of response would be changing our definition of what is an event (e.g., from survival =1 to death = 1) which can be viewed as a very special linear transformation Z = 1 - Y***
  - ***This linear transformation Z = 1 - Y induces a reciprocal transformation on the odds*** $\theta_{Z|X} = 1 / \theta_{Y|X}$ ***so there are straightforward transformations of the fitted values and estimates of the association. There will be no effect on statistical significance about associations.***

$$g\left(\theta_{Y|X}\right) = \log\left(Odds\left(Y \mid X\right)\right) = \log\left(\frac{\Pr\left(Y=1 \mid X\right)}{1-\Pr\left(Y=1 \mid X\right)}\right) = \mu_X \;= \beta_0 + \beta_1 \times X$$

$$\begin{aligned} g\left(\theta_{Z|X}\right) = \log\left(Odds\left(Z \mid X\right)\right) = \log\left(\frac{\Pr\left(Z=1 \mid X\right)}{1-\Pr\left(Z=1 \mid X\right)}\right) &= \eta_X \;= \gamma_0 + \gamma_1 \times X \\ &= \log\left(\frac{\Pr\left(Y=0 \mid X\right)}{1-\Pr\left(Y=0 \mid X\right)}\right) = \log\left(\frac{1-\Pr\left(Y=1 \mid X\right)}{\Pr\left(Y=1 \mid X\right)}\right) \\ &= -\log\left(\frac{\Pr\left(Y=1 \mid X\right)}{1-\Pr\left(Y=1 \mid X\right)}\right) = -\mu_X \;= -\beta_0 + -\beta_1 \times X \end{aligned}$$

$$\Rightarrow \qquad \gamma_0 = -\beta_0 \qquad\qquad \gamma_1 = -\beta_1$$

- ***All other transformations of the response will generally affect the mean in complicated ways, and thus those other transformations will represent a major change in the scientific question being addressed.***

- ***When analyzing ratios of mean response with Poisson regression on non-negative Y***
  - ***Scale transformations Z = c Y with c > 0 induce scale transformations $\theta_{Z|X}$=c $\theta_{Y|X}$ so there are straightforward transformations of the fitted values and estimates of the association. There will be no effect on statistical significance about associations providing the mean-variance relationship agrees with the Poisson mean-variance relationship or we use the Huber-White sandwich estimator for the standard error.***

$$\begin{aligned} g\left(\theta_{Y|X}\right) &= \log\left(E\left(Y \mid X\right)\right) = \mu_X \;= \beta_0 + \beta_1 \times X \\ g\left(\theta_{Z|X}\right) &= \log\left(E\left(Z \mid X\right)\right) = \eta_X \;= \gamma_0 + \gamma_1 \times X \\ &= \log\left(E\left(cY \mid X\right)\right) = \log\left(cE\left(Y \mid X\right)\right) = \log(c) + \log\left(E\left(Y \mid X\right)\right) \\ &= \log(c) + \left(\beta_0 + \beta_1 \times X\right) \\ &= \left(\log(c) + \beta_0\right) + \left(\beta_1\right) \times X \end{aligned}$$

$$\Rightarrow \qquad \gamma_0 = \log(c) + \beta_0 \qquad\qquad \gamma_1 = \beta_1$$

  - ***With binary response, changing our definition of what is an event (e.g., from survival =1 to death = 1) can be viewed as a very special linear transformation Z = 1 – Y. Unlike with logistic regression, changing that point of reference will have major impact on the quantification of our scientific measure of association.***

$$\begin{aligned} g\left(\theta_{Y|X}\right) &= \log\left(E\left(Y \mid X\right)\right) = \mu_X \;= \beta_0 + \beta_1 \times X \\ g\left(\theta_{Z|X}\right) &= \log\left(E\left(Z \mid X\right)\right) = \eta_X \;= \gamma_0 + \gamma_1 \times X \\ &= \log\left(E\left(1-Y \mid X\right)\right) = \log\left(1-E\left(Y \mid X\right)\right) \end{aligned}$$

$$\Rightarrow \quad \textit{no easy relationship between} \quad \left(\gamma_0, \gamma_1\right) \textit{and } \left(\beta_0, \beta_1\right)$$

    - ***This is because the ratio $\theta_{Z|X=x+1}$ / $\theta_{Z|X=x}$ = (1- $\theta_{Y|X=x+1}$) /(1- $\theta_{Y|X=x+1}$) is not in general easily related to the ratio $\theta_{Y|X=x+1}$ / $\theta_{Y|X=x}$***

        - ***Furthermore, it is only when we use the Huber-White sandwich estimator for the standard errors that we will tend to get similar p values for tests of association.***
    - ***All other transformations of the response will similarly affect the mean in complicated ways, and thus those other transformations will represent a major change in the scientific question being addressed.***
- ***When analyzing <u>ratios of hazard functions with proportional hazards regression</u>***
    - ***"Order-preserving transformations" Z= h(Y) will have no effect whatsoever on measures of association or statistical inference about those measures.***
        - ***"Order-preserving transformations" are those in which $x_1 > x_2$ implies $h(x_1) > h(x_2)$. These are sometimes called "monotonic transformations") This includes linear transformations in which the rescaling parameter is positive (so Z= c + dX with d > 0) and logarithmic transformation of positive random variables.***
    - ***The fact that inference about associations (hazard ratos) is unaffected by order-preserving transformations is a result of the semi-parametric comparison made at each distinct failure time within risk groups—all that matters is who has an event relative to who is in athe risk set at each point in time. Order-preserving transformations do not affect the risk sets at each point in time.***
    - ***Transformations will affect the estimates of the "baseline survival function", though we rarely estimate that.***
- ***When analyzing <u>ratios of medians (or other quantiles) with parametric accelerated failure time (AFT) regression in the presence of censored data</u>***
    - ***These models look similar to our "distribution-free" models for geometric means, in that they use a log link on the quantiles of the distribution. Their estimation techniques differ from least squares on the log transformed response, because they must account for censored observations, and the way they account for the censoring uses information about the hypothesized shape of the distribution. (Their will not always be an intercept, and its interpretation will be as some quantile of the distribution, but it will not always be the median.)***
    - ***Nonetheless, power transformations $Z = c\, Y^d$ with c > 0 induce power transformations $\theta_{Z|X} = c\, \theta^d_{Y|X}$ so there are straightforward transformations of the fitted values and estimates of the association. There will be no effect on statistical significance about associations.***

$$
\begin{aligned}
g\left(\theta_{Y|X}\right) &= \log\left(GM\left(Y \mid X\right)\right) = E\left(\log Y \mid X\right) = \mu_X = \beta_0 + \beta_1 \times X \\
g\left(\theta_{Z|X}\right) &= \log\left(GM\left(Z \mid X\right)\right) = E\left(\log Z \mid X\right) = \eta_X = \gamma_0 + \gamma_1 \times X \\
&= E\left(\log(c) + d \log(Y) \mid X\right) = \log(c) + dE\left(\log Y \mid X\right) \\
&= \log(c) + d \times \mu_X = \log(c) + d \times \left(\beta_0 + \beta_1 \times X\right) \\
&= \left(\log(c) + d \times \beta_0\right) + \left(d \times \beta_1\right) \times X
\end{aligned}
$$

$$
\Rightarrow \qquad \gamma_0 = \log(c) + d \times \beta_0 \qquad\qquad \gamma_1 = d \times \beta_1
$$

- ***All other transformations of the response will generally affect the quantiles in complicated ways, and thus those other transformations will represent a major change in the scientific question being addressed.***

2. Perform a statistical regression analysis evaluating an association between the odds of delivery of infants who were small for gestational age (SGA) and maternal smoking behavior. (Only give a formal report of the inference where asked to.)

   a. Give full inference regarding the association between SGA and maternal smoking.

***Instructions for grading**: This problem is worth 10 points. Points to consider in the grading include:*

- ***The smoking variable was originally coded as 1=yes, 2= no. As we have learned in lecture, this has no impact on the scientific validity of the analysis. But a student who did not notice this coding could have made major mistakes in the interpretation of the results. Any student who did not properly interpret the regression parameter estimates due to this coding should receive no credit for the entire problem. (I truly hope there were no such students.)***
- ***I personally recoded the variable before fitting a regression model. Such "re-parameterization" of the model has no impact on the scientific validity of the analysis, and it does not even have to be mentioned in the methods, if the regression parameters are interpreted correctly and reported using scientifically important words (e.g., "smokers" vs "nonsmokers" rather than referring to such statistical jargon as "slope parameters" or "a 1 unit difference in the covariate").***

Ans: ***Methods:*** **Any association between delivery of an infant who was small for gestational age (SGA) and maternal smoking behavior (smoker vs nonsmoker) was summarized by the odds ratio comparing the odds of SGA among smoking mothers to the odds among nonsmokers. An estimated odds ratio was computed using the maximum likelihood estimate of the slope from a logistic regression model that only included the binary indicator of smoking. Using the asymptotically normal distribution for Wald statistics derived from the logistic regression model, two-sided 95% confidence intervals for the true odds ratio were computed. Similarly, two-sided P values testing the null hypothesis that there is no true association between SGA and maternal smoking (i.e., that the true odds ratio is equal to 1) were based on the normal approximation for the Wald based Z statistic computed from the logistic regression slope parameter. Four subjects missing data for smoking behavior were omitted from the data analysis.**

Results: **Babies deemed SGA were delivered for 59 of the 520 women who self-reported as nonsmokers (probability of SGA = 11.3%, odds of SGA = 0.128), and there were 45 SGA infants among the 231 women who self-reported as smoking during pregnancy (probability of SGA = 19.5%, odds of SGA = 0.242). Hence, the odds of SGA infants is estimated to be 89.0% higher among smokers than it is among nonsmokers (OR = 1.890). A 95% confidence interval derived from the logistic regression model suggests that the observed data is consistent with a true odds ratio such that the odds of SGA is anywhere between 1.24 to 2.89 times higher among smokers than nonsmokers. Such a result is statistically significant (two-sided P= 0.003), and we can thus reject the null hypothesis of no association between SGA and maternal smoking.**

   b. Use the regression model parameter estimates to provide estimates of both the odds and the probability of delivering a SGA infant separately for smokers and nonsmokers. How do these estimates compare with simple descriptive statistics as you might have reported in problem 1. Explain any differences or similarities.

***Instructions for grading**: This problem is worth 5 points.*

<u>Ans:</u> From the sample statistics, babies deemed SGA were delivered for 59 of the 520 women who self-reported as nonsmokers (probability of SGA = 11.3%, odds of SGA = 0.128), and there were 45 SGA infants among the 231 women who self-reported as smoking during pregnancy (probability of SGA = 19.5%, odds of SGA = 0.242). Because we fit a saturated logistic regression model, the corresponding fitted values from the regression have to agree exactly with those numbers. *(Note that I asked for these probabilities and odds for delivering SGA using scientific words. The students' answers must agree with these numbers no matter which of the four models they fit, and no matter whether they recoded the smoking variable or not. I really do not care whether they showed how to compute the answers from the regression model. Knowing that the saturated model's estimates agreed with the sample estimates is what I am really after. I provide the rest just to show the correspondence.*

From the logistic regression analysis:

- *logit(Pr(SGA=1 | SMOKER) = -2.055861 + 0.6367768 * SMOKER*
- *For nonsmokers SMOKER= 0, so*
  - *log Odds(SGA) = -2.055861;*
  - *Odds(SGA) = exp(-2.015561) = 0.127982593*
  - *Prob(SGA) = Odds(SGA) / (1 + Odds(SGA)) = 0.1279826 / 1.1279826 = 0.11346*
- *For smokers SMOKER= 1, so*
  - *log Odds(SGA) = -2.055861 + 0.6367768 = -1.4190842;*
  - *Odds(SGA) = exp(-1.4190842) = 0.24193548*
  - *Prob(SGA) = Odds(SGA) / (1 + Odds(SGA)) = 0. 24193548 / 1. 24193548 = 0.1948*

c. There were actually four regression analyses that could have been used to answer this question. I am betting that all students would have fit a regression model with SGA as response and the indicator of maternal smoking as the predictor. Presuming that you did indeed fit that model, explain the similarities and differences between the estimates and inference you would have obtained for the following three additional models (You do not need to run these analyses, if you can tell me how they differ without doing so. It is of course okay to run the analyses if it will help you recognize the more general principles.):

i. You create an indicator NONSMOKER that the mother was a nonsmoker, and you fit a logistic regression model of response SGA on predictor NONSMOKER.

ii. You create an indicator NOTSGA that the infant was not small for gestational age, and you fit a logistic regression model of response NOTSGA on predictor SMOKER.

iii. You fit a regression model of response NOTSGA on predictor NONSMOKER.

<u>Instructions for grading</u>: This problem is worth 10 points. Points to consider in grading:

- ***The most important point for them to make is that all fitted values and all inference about associations would be exactly the same when translated into English. That is, if we always express our results as odds of SGA comparing smokers to nonsmokers, we get the exact same estimate, CI, and P value.***
- ***Secondarily, pointing out how the intercept and slopes would differ across models should be mentioned in some way. To the extent that they might have treated my use of the notation SMOKER as being the variable in the data set, that is okay by me, providing they make the right comparisons.***

- ***Using smoker vs NONSMOKER will give the same slopes as each other, but the intercept when using NONSMOKER will be the sum of the intercept and slope when using smoker.***

<u>Ans:</u> These are actually all the same model scientifically. All that differs is the interpretation of the regression parameters. The fitted values will agree for all models. The P values for tests of association will be exactly the same, and when expressing the results from each model in the same scientific terms (e.g. ratio comparing odds of SGA for smokers to that for nonsmokers), we will have the same estimates and the same CI. At a technical statistical level, when using as a reference the parameter estimates from the logistic regression of SGA on SMOKER (coded 0=nonsmoker, 1= smoker):

$$\log\left(Odds\left(SGA \mid SMOKER\right)\right) = \beta_0 + \beta_1 \times SMOKER$$

$$\begin{aligned}\log\left(Odds\left(SGA \mid NONSMOKER\right)\right) &= \gamma_0 + \gamma_1 \times NONSMOKER \\ &= \left(\beta_0 + \beta_1\right) + \left(-\beta_1\right) \times NONSMOKER\end{aligned}$$

$$\begin{aligned}\log\left(Odds\left(NOTSGA \mid SMOKER\right)\right) &= \alpha_0 + \alpha_1 \times SMOKER \\ &= \left(-\beta_0\right) + \left(-\beta_1\right) \times SMOKER\end{aligned}$$

$$\begin{aligned}\log\left(Odds\left(NOTSGA \mid NONSMOKER\right)\right) &= \delta_0 + \delta_1 \times NONSMOKER \\ &= \left(-\beta_0 - \beta_1\right) + \left(\beta_1\right) \times NONSMOKER\end{aligned}$$

- **Fitting response SGA on NONSMOKER will have a log odds intercept equal to the sum of the reference slope and intercept, and it will have a log odds slope opposite in sign of the reference slope.** *(I was using the log odds scale. When expressed on the odds scale, the odds intercept will be the product of the intercept and slope from the reference model, and the OR slope will be the reciprocal of the reference model OR.)*
- **Fitting response NOTSGA on SMOKER will have a log odds intercept equal to the opposite sign of the reference intercept, and it will have a log odds slope opposite in sign of the reference slope.** *(I was using the log odds scale. When expressed on the odds scale, the odds intercept will be the reciprocal of the intercept from the reference model, and the OR slope will be the reciprocal of the reference model OR.)*
- **Fitting response NOTSGA on NONSMOKER will have a log odds intercept equal to the negative of the sum of the reference slope and intercept, and it will have a log odds slope the same as the reference slope.** *(I was using the log odds scale. When expressed on the odds scale, the odds intercept will be the reciprocal of the product of the intercept and slope from the reference model, and the OR slope will be the same the reference model OR.)*

3. Repeat problem 2, except consider a statistical regression analysis evaluating an association between the odds of delivery of infants who were small for gestational age (SGA) and maternal smoking behavior by evaluating the difference in probabilities for SGA across smoking groups.

   a. Give full inference regarding the association between SGA and maternal smoking.

<u>Instructions for grading</u>: This problem is worth 10 points. Points to consider in the grading include:

- ***In the wording of the problem, I carried over the idea that we were looking at odds of delivery of SGA, but explicitly told you to use difference in probabilities. I get away with doing this, because if there is a difference in probabilities, there also has to be a difference in odds, and, furthermore, the group with higher probabilities also has higher odds (it is an "order-preserving transformation" to go from odds to probabilities or vice versa). It would in fact be more common for people to say they were interested in the probabilities and then do an analysis based on OR in order to adjust for variables. But it is the same issue.***
- ***If there is an association between SGA and maternal smoking, there will also be different variances between the groups. Hence for the purposes of estimating a CI, it is probably better (but not best) to use the Huber-White sandwich estimator (best would be to use some sort of score statistic, but I know of no software that does so.) If all we cared about was testing the null hypothesis of no association, it is perfectly valid to use classical linear regression, because in the absence of adjusting for any other covariates, the variances would be equal under the null. (If we were adjusting for other covariates, we would need to consider the heteroscedasticity due to those variables even when there was no association between SGA and smoking.)***
- ***The smoking variable was originally coded as 1=yes, 2= no. As we have learned in lecture, this has no impact on the scientific validity of the analysis. But a student who did not notice this coding could have made major mistakes in the interpretation of the results. Any student who did not properly interpret the regression parameter estimates due to this coding should receive no credit for the entire problem. (I truly hope there were no such students.)***
- ***I personally recoded the variable before fitting a regression model. Such "re-parameterization" of the model has no impact on the scientific validity of the analysis, and it does not even have to be mentioned in the methods, if the regression parameters are interpreted correctly and reported using scientifically important words (e.g., "smokers" vs "nonsmokers" rather than referring to such statistical jargon as "slope parameters" or "a 1 unit difference in the covariate").***

Ans: *Methods:* **Any association between delivery of an infant who was small for gestational age (SGA) and maternal smoking behavior (smoker vs nonsmoker) was summarized by the difference in probabilities of SGA among smoking mothers minus that among nonsmokers. An estimated difference in probabilities was computed using the least squares estimate of the slope from a linear regression model that only included the binary indicator of smoking. Using the asymptotically normal distribution for Wald statistics derived from the linear regression model, two-sided 95% confidence intervals for the true odds ratio were computed using the Huber-White sandwich estimate of the standard error to account for the mean-variance relationship in these binary data. Similarly, two-sided P values testing the null hypothesis that there is no true association between SGA and maternal smoking (i.e., that the true difference in probabilities is equal to 0) were based on the normal approximation for the Wald based Z statistic computed from the regression slope parameter. Four subjects missing data for smoking behavior were omitted from the data analysis.**

Results: **Babies deemed SGA were delivered for 59 of the 520 women who self-reported as nonsmokers (probability of SGA = 11.3%, odds of SGA = 0.128), and there were 45 SGA infants among the 231 women who self-reported as smoking during pregnancy (probability of SGA = 19.5%, odds of SGA = 0.242). Hence, the probability of SGA infants is estimated to be an absolute 0.813 higher among smokers than it is among nonsmokers. A 95% confidence interval derived from the regression model suggests that the observed data is consistent with a true difference such that the probability of SGA is anywhere between 0.0233 to 0.139 higher among smokers than nonsmokers. Such a result is statistically significant (two-sided P= 0.006), and we can thus reject the null hypothesis of no association between SGA and maternal smoking.**

b. Use the regression model parameter estimates to provide estimates of both the odds and the probability of delivering a SGA infant separately for smokers and nonsmokers. How do these estimates compare with simple descriptive statistics as you might have reported in problem 1. Explain any differences or similarities.

***Instructions for grading*: This problem is worth 5 points.**

Ans: From the sample statistics, babies deemed SGA were delivered for 59 of the 520 women who self-reported as nonsmokers (probability of SGA = 11.3%, odds of SGA = 0.128), and there were 45 SGA infants among the 231 women who self-reported as smoking during pregnancy (probability of SGA = 19.5%, odds of SGA = 0.242). Because we fit a saturated linear regression model, the corresponding fitted values from the regression have to agree exactly with those numbers. *(Note that I asked for these probabilities and odds for delivering SGA using scientific words. The students' answers must agree with these numbers no matter which of the four models they fit, and no matter whether they recoded the smoking variable or not. I really do not care whether they showed how to compute the answers from the regression model. Knowing that the saturated model's estimates agreed with the sample estimates is what I am really after. I provide the rest just to show the correspondence.*

From the linear regression analysis:

- *Pr(SGA=1 | SMOKER) = 0.0813437 + .1134615 * SMOKER*
- *For nonsmokers SMOKER= 0, so Pr(SGA) = 0.0813*
- *For smokers SMOKER= 1, so Pr(SGA) = 0.0813437 + .1134615 = 0.1948*

c. There were actually four regression analyses that could have been used to answer this question. I am betting that all students would have fit a regression model with SGA as response and the indicator of maternal smoking as the predictor. Presuming that you did indeed fit that model, explain the similarities and differences between the estimates and inference you would have obtained for the following three additional models (You do not need to run these analyses, if you can tell me how they differ without doing so. It is of course okay to run the analyses if it will help you recognize the more general principles.):

i. You create an indicator NONSMOKER that the mother was a nonsmoker, and you fit a logistic regression model of response SGA on predictor NONSMOKER.

ii. You create an indicator NOTSGA that the infant was not small for gestational age, and you fit a logistic regression model of response NOTSGA on predictor SMOKER.

iii. You fit a regression model of response NOTSGA on predictor NONSMOKER.

***Instructions for grading*: This problem is worth 10 points. Points to consider in grading:**

- ***The most important point for them to make is that all fitted values and all inference about associations would be exactly the same when translated into English. That is, if we always express our results as probability of SGA comparing smokers to nonsmokers, we get the exact same estimate, CI, and P value.***
- ***Secondarily, pointing out how the intercept and slopes would differ across models should be mentioned in some way. To the extent that they might have treated my use of the notation SMOKER as being the variable in the data set, that is okay by me, providing they make the right comparisons.***
  - ***Using smoker vs NONSMOKER will give the same slopes as each other, but the intercept when using NONSMOKER will be the sum of the intercept and slope when using smoker.***

<u>Ans:</u> These are actually all the same model scientifically. All that differs is the interpretation of the regression parameters. The fitted values will agree for all models. The P values for tests of association will be exactly the same, and when expressing the results from each model in the same scientific terms (e.g. difference comparing probability of SGA for smokers to that for nonsmokers), we will have the same estimates and the same CI. At a technical statistical level, when using as a reference the parameter estimates from the linear regression of SGA on SMOKER (coded 0=nonsmoker, 1= smoker):

$$\left(\Pr\left(SGA \mid SMOKER\right)\right) = \beta_0 + \beta_1 \times SMOKER$$

$$\begin{aligned}\left(\Pr\left(SGA \mid NONSMOKER\right)\right) &= \gamma_0 + \gamma_1 \times NONSMOKER \\ &= \left(\beta_0 + \beta_1\right) + \left(-\beta_1\right) \times NONSMOKER\end{aligned}$$

$$\begin{aligned}\left(\Pr\left(NOTSGA \mid SMOKER\right)\right) &= \alpha_0 + \alpha_1 \times SMOKER \\ &= \left(1 - \beta_0\right) + \left(-\beta_1\right) \times SMOKER\end{aligned}$$

$$\begin{aligned}\left(\Pr\left(NOTSGA \mid NONSMOKER\right)\right) &= \delta_0 + \delta_1 \times NONSMOKER \\ &= \left(1 - \beta_0 - \beta_1\right) + \left(\beta_1\right) \times NONSMOKER\end{aligned}$$

- **Fitting response SGA on NONSMOKER will have an intercept equal to the sum of the reference slope and intercept, and it will have a slope opposite in sign of the reference slope.**
- **Fitting response NOTSGA on SMOKER will have an intercept equal to 1 minus the reference intercept, and it will have a slope opposite in sign of the reference slope.**
- **Fitting response NOTSGA on NONSMOKER will have an intercept equal to 1 minus the sum of the reference slope and intercept, and it will have a slope the same as the reference slope.**

4. Repeat problem 2, except consider a statistical regression analysis evaluating an association between the odds of delivery of infants who were small for gestational age (SGA) and maternal smoking behavior by evaluating the ratio of probabilities for SGA across smoking groups.

   a. Give full inference regarding the association between SGA and maternal smoking.

<u>Instructions for grading</u>: This problem is worth 10 points. Points to consider in the grading include:

- ***In the wording of the problem, I carried over the idea that we were looking at odds of delivery of SGA, but explicitly told you to use difference in probabilities. I get away with doing this, because if there is a difference in probabilities, there also has to be a difference in odds, and, furthermore, the group with higher probabilities also has higher odds (it is an "order-preserving transformation" to go from odds to probabilities or vice versa). It would in fact be more common for people to say they were interested in the probabilities and then do an analysis based on OR in order to adjust for variables. But it is the same issue.***
- ***If there is an association between SGA and maternal smoking, there will also be different variances between the groups, and if the event is not rare, it will not agree with the Poisson mean-variance relationship. Hence for the purposes of estimating a CI, it is probably better (but not best) to use the Huber-White sandwich estimator (best would be to use some sort of score statistic, but I know of no software that does so.) If all we cared about was testing the null hypothesis of no association <u>and the probability was sufficiently small</u>, it is perfectly valid to***

use classical Poisson regression, because in the absence of adjusting for any other covariates, the variances for rare events would be well-approximated by the Poisson mean-variance. All things considered, I think it very important to use the sandwich estimator.

- ***The smoking variable was originally coded as 1=yes, 2= no. As we have learned in lecture, this has no impact on the scientific validity of the analysis. But a student who did not notice this coding could have made major mistakes in the interpretation of the results. Any student who did not properly interpret the regression parameter estimates due to this coding should receive no credit for the entire problem. (I truly hope there were no such students.)***
- ***I personally recoded the variable before fitting a regression model. Such "re-parameterization" of the model has no impact on the scientific validity of the analysis, and it does not even have to be mentioned in the methods, if the regression parameters are interpreted correctly and reported using scientifically important words (e.g., "smokers" vs "nonsmokers" rather than referring to such statistical jargon as "slope parameters" or "a 1 unit difference in the covariate").***

<u>Ans:</u> *Methods:* Any association between delivery of an infant who was small for gestational age (SGA) and maternal smoking behavior (smoker vs nonsmoker) was summarized by the ratio comparing the probability of SGA among smoking mothers to the probability among nonsmokers. An estimated risk ratio was computed using the maximum likelihood estimate of the slope from a Poisson regression model that only included the binary indicator of smoking. Using the asymptotically normal distribution for Wald statistics derived from the Poisson regression model, two-sided 95% confidence intervals for the true risk ratio were computed using the Huber-White sandwich estimate of the standard error to account for the mean-variance relationship in these binary data.. Similarly, two-sided P values testing the null hypothesis that there is no true association between SGA and maternal smoking (i.e., that the true risk ratio is equal to 1) were based on the normal approximation for the Wald based Z statistic computed from the Poisson regression slope parameter. Four subjects missing data for smoking behavior were omitted from the data analysis.

***Results:* Babies deemed SGA were delivered for 59 of the 520 women who self-reported as nonsmokers (probability of SGA = 11.3%, odds of SGA = 0.128), and there were 45 SGA infants among the 231 women who self-reported as smoking during pregnancy (probability of SGA = 19.5%, odds of SGA = 0.242). Hence, the probability of SGA infants is estimated to be a relative 71.7% higher among smokers than it is among nonsmokers (RR = 1.72). A 95% confidence interval derived from the Poisson regression model suggests that the observed data is consistent with a true risk ratio such that the probabiity of SGA is anywhere between 1.20 to 2.45 times higher among smokers than nonsmokers. Such a result is statistically significant (two-sided P= 0.003), and we can thus reject the null hypothesis of no association between SGA and maternal smoking.**

b. Use the regression model parameter estimates to provide estimates of both the odds and the probability of delivering a SGA infant separately for smokers and nonsmokers. How do these estimates compare with simple descriptive statistics as you might have reported in problem 1. Explain any differences or similarities.

<u>Instructions for grading</u>: This problem is worth 5 points.

<u>Ans:</u> From the sample statistics, babies deemed SGA were delivered for 59 of the 520 women who self-reported as nonsmokers (probability of SGA = 11.3%, odds of SGA = 0.128), and there were 45 SGA infants among the 231 women who self-reported as smoking during pregnancy (probability of SGA = 19.5%, odds of SGA = 0.242). Because we fit a saturated Poisson regression model, the corresponding fitted values from the regression have to agree exactly with those numbers. *(Note that I asked for these probabilities and odds for delivering SGA using scientific words. The students' answers* *must agree with these numbers no matter which of the four models they fit, and no matter whether they recoded the smoking variable or not. I really do not care whether they showed how to compute the answers from the regression model. Knowing that the saturated model's estimates agreed with the sample estimates is what I am really after. I provide the rest just to show the correspondence.*

From the logistic regression analysis:

- *log(Pr(SGA=1 | SMOKER) = -2.176291 + 0.5405362 * SMOKER*
- *For nonsmokers SMOKER= 0, so*
  - *log Prob(SGA) = -2.176291;*
  - *Prob(SGA) = exp(-2.176291) = 0.1135*
- *For smokers SMOKER= 1, so*
  - *log Prob(SGA) = -2.176291 + 0.5405362 = -1.6357548;*
  - *Prob(SGA) = exp(-1.6357548) = 0.1948*

c. There were actually four regression analyses that could have been used to answer this question. I am betting that all students would have fit a regression model with SGA as response and the indicator of maternal smoking as the predictor. Presuming that you did indeed fit that model, explain the similarities and differences between the estimates and inference you would have obtained for the following three additional models (You do not need to run these analyses, if you can tell me how they differ without doing so. It is of course okay to run the analyses if it will help you recognize the more general principles.):

   i. You create an indicator NONSMOKER that the mother was a nonsmoker, and you fit a logistic regression model of response SGA on predictor NONSMOKER.

   ii. You create an indicator NOTSGA that the infant was not small for gestational age, and you fit a logistic regression model of response NOTSGA on predictor SMOKER.

   iii. You fit a regression model of response NOTSGA on predictor NONSMOKER.

<u>Instructions for grading</u>: This problem is worth 10 points. Points to consider in grading:

- ***The most important points for them to make is that all fitted values and estimates of association for models having the same response variable would be exactly the same when translated into English.***
  - ***That is, when using SGA as the response, if we always express our results as probability of SGA comparing smokers to nonsmokers, we get the exact same estimate, CI, and P value no matter whether we fit SMOKER or NONSMOKER.***
  - ***Similarly, when using NOTSGA as the response, we get the same inference whether we use SMOKER or NONSMOKER.***
- ***Unlike in problems 2 and 3, however, we do not get particularly good correspondence between the models using SGA as response when compared to those using NOTSGA as response.***
  - ***As noted above in the preamble to problem 2-4, the ratio of p_1 / p_0 is not easily converted to a ratio of (p_1 / p_0).***
  - ***(Of course, in this saturated model fitting only two groups, we could compute the ratios from the fitted values, but getting the standard errors and CI would be problematic.)***

- ***It is of course worth noting that if we were only interested in testing hypotheses, we would get very similar p values across the four models providing we use the sandwich estimate of the SE. (Failure to do that will cause the p values to go awry.)***
- ***Secondarily, pointing out how the intercept and slopes would differ across models should be mentioned in some way. To the extent that they might have treated my use of the notation SMOKER as being the variable in the data set, that is okay by me, providing they make the right comparisons.***
  - ***Using smoker vs NONSMOKER will give the same slopes as each other, but the intercept when using NONSMOKER will be the sum of the intercept and slope when using smoker.***

Ans: Because we have a saturated model, in terms of fitted values these are actually all the same model scientifically. However, the contrast across the groups will very much depend upon whether we are looking at ratios of the probability of being SGA or whether we are looking at the probabilites of being not SGA. But within those two choices, how we model the covariate indicating smoking behavior will not matter except for the interpretation of the regression parameters: the inference about ratios of being SGA will agree with each other, and the inference about ratios fo being not SGA will agree with each other. The P values for tests of association will be approximately the same since we used the sandwich estimator for standard errors. When using the sandwich estimator of SE, the p values were .003 when SGA was the response in Poisson regression and .007 when NOTSGA was the response. *(On the other hand, when not using the robust SE, the p values were .006 when using the relatively rarer SGA as the response, and .268 when using NOTSGA as the response.)*

At a technical statistical level, when using as a reference the parameter estimates from the Poisson regression of SGA on SMOKER (coded 0=nonsmoker, 1= smoker):

$$\log(\Pr(SGA \mid SMOKER)) = \beta_0 + \beta_1 \times SMOKER$$

$$\begin{aligned}\log(\Pr(SGA \mid NONSMOKER)) &= \gamma_0 + \gamma_1 \times NONSMOKER \\ &= (\beta_0 + \beta_1) + (-\beta_1) \times NONSMOKER\end{aligned}$$

$$\begin{aligned}\log(\Pr(NOTSGA \mid SMOKER)) &= \alpha_0 + \alpha_1 \times SMOKER \\ &= \left(\log\left(1-e^{\beta_0}\right)\right) + \left(\log\left(\frac{1-e^{\beta_0+\beta_1}}{1-e^{\beta_0}}\right)\right) \times SMOKER\end{aligned}$$

$$\begin{aligned}\log(\Pr(NOTSGA \mid NONSMOKER)) &= \delta_0 + \delta_1 \times NONSMOKER \\ &= (\alpha_0 + \alpha_1) + (-\alpha_1) \times NONSMOKER\end{aligned}$$

- **Fitting response SGA on NONSMOKER will have a log probability intercept equal to the sum of the reference slope and intercept, and it will have a log probability slope opposite in sign of the reference slope.**
- **Fitting response NOTSGA on SMOKER will have a log probability intercept equal to the log of 1 minus the exponentiated reference intercept, and it will have a log odds slope with a**

complicated relationship to the parameters when SGA is response, as given in the displayed equations above.

- **Fitting response NOTSGA on NONSMOKER will have an easy relationship (as given above) to the parameter estimates from fitting NOTSGA on SMOKER, but again a complicated relationship to the models using SGA as response.**
-

(A moral of this story is that looking at risk ratios is best when dealing with low probabilities. The risk ratio is not "invariant" to how you define an event.)

5. How do the analyses performed in problems 2-4 compare to that that would be obtained in a simple two sample comparison of SGA by smoking status (i.e., using methods covered in Biost 517/514.) Explicitly mention where they would be similar or different?

<u>Instructions for grading</u>: This problem is worth 10 points.

<u>Ans:</u> **The analyses presented in problem 2 correspond roughly to a chi square test for association, which is the score test from logistic regression.**

The analyses presented in problem 3 correspond roughly to a t test that allows for unequal variances.

The analyses presented in problem 4 correspond roughly to a 2 sample test of probability ratios, as derived from likelihood theory *(and implemented in cs in Stata).*

Most importantly, in large samples, these tests will all behave similarly.

6. Perform a regression analysis of the distribution of the prevalence of SGA infants across groups defined by the continuous measure of maternal age. In all cases we want formal inference. (Note: In problem 7, I am asking you to plot the estimated probabilities of SGA infants from each of these regression models. Hence, you will want to make sure you estimate those fitted values following each regression.)

    a. Evaluate associations using risk difference (RD: difference in probabilities).

<u>Instructions for grading</u>: This problem is worth 10 points.

<u>Ans:</u> *Methods:* **Any association between delivery of an infant who was small for gestational age (SGA) and maternal age was summarized by the risk difference. The estimated difference in risk of SGA was computed using the estimated slope from a linear regression model that only included an untransformed variable measuring maternal age. Using the asymptotically normal distribution for Wald statistics derived from the logistic regression model, two-sided 95% confidence intervals for the true odds ratio were computed based on the Huber-White sandwich estimator for the standard error. Similarly, two-sided P values testing the null hypothesis that there is no true association between SGA and maternal age (i.e., that the risk difference is equal to 0) were based on the normal approximation for the Wald based Z statistic computed from the linear regression slope parameter. No subjects were missing data for the data analysis.**

Results: **Maternal age averaged 24.8 years (SD 5.4, range 14 – 43) in 755 pregnant women, of whom 13.9% were later observed to give birth to an infant who was small for gestational age (SGA). When comparing groups of women who differed in age, the probability of SGA infants is estimated to be .0226 lower for each 5-year difference in age, with the older women having lower risk of SGA. A 95% confidence interval derived from the linear regression model suggests that the observed data is consistent with a true risk difference such that the probability of SGA for a population of women is anywhere between an absolute 4.37% to 0.143% lower for each 5-year difference in age.**

Such a result is statistically significant (two-sided P= 0.036), and we can thus reject the null hypothesis of no association between SGA and maternal age.

b. Evaluate associations between risk ratio (RR: ratios of probabilities).

Instructions for grading: This problem is worth 10 points.

Ans: ***Methods:*** **Any association between delivery of an infant who was small for gestational age (SGA) and maternal age was summarized by the risk ratio. The estimated ratio of risk of SGA was computed using the estimated slope from a Poisson regression model that only included an untransformed variable measuring maternal age. Using the asymptotically normal distribution for Wald statistics derived from the Poisson regression model, two-sided 95% confidence intervals for the true odds ratio were computed based on the Huber-White sandwich estimator for the standard error. Similarly, two-sided P values testing the null hypothesis that there is no true association between SGA and maternal age (i.e., that the risk ratio is equal to 1) were based on the normal approximation for the Wald based Z statistic computed from the linear regression slope parameter. No subjects were missing data for the data analysis.**

Results: **Maternal age averaged 24.8 years (SD 5.4, range 14 – 43) in 755 pregnant women, of whom 13.9% were later observed to give birth to an infant who was small for gestational age (SGA). When comparing groups of women who differed in age, the probability of SGA infants is estimated to be a relative 16.8% lower for each 5-year difference in age (RR= 0.842), with the older women having lower risk of SGA. A 95% confidence interval derived from the Poisson regression model suggests that the observed data is consistent with a true risk ratio such that the probability of SGA for a population of women is anywhere between a relative 28.9% to 0.303% lower for each 5-year difference in age (95% CI for RR; 0.711 to 0.997). Such a result is statistically significant (two-sided P= 0.046), and we can thus reject the null hypothesis of no association between SGA and maternal age.**

c. Evaluate associations using odds ratio (OR: ratios of odds)

Instructions for grading: This problem is worth 10 points.

Ans: ***Methods:*** **Any association between delivery of an infant who was small for gestational age (SGA) and maternal age was summarized by an odds ratio. The estimated odds ratio was computed using the maximum likelihood estimate of the slope from a logistic regression model that only included an untransformed variable measuring maternal age. Using the asymptotically normal distribution for Wald statistics derived from the logistic regression model, two-sided 95% confidence intervals for the true odds ratio were computed. Similarly, two-sided P values testing the null hypothesis that there is no true association between SGA and maternal age (i.e., that the true odds ratio is equal to 1) were based on the normal approximation for the Wald based Z statistic computed from the logistic regression slope parameter. No subjects were missing data for the data analysis.**

Results: **Maternal age averaged 24.8 years (SD 5.4, range 14 – 43) in 755 pregnant women, of whom 13.9% were later observed to give birth to an infant who was small for gestational age (SGA). When comparing groups of women who differed in age, the odds of SGA infants is estimated to be 18.0% lower for each 5-year difference in age, with the older women having lower risk of SGA (OR = 0.819). A 95% confidence interval derived from the logistic regression model suggests that the observed data is consistent with a true odds ratio such that the odds of SGA for a population of women is anywhere between 33.1%% lower to 0.38% higher for each 5-year difference in age (95% CI for OR: 0.669, 1.004). Such a result is not statistically significant (two-sided P= 0.054), and we cannot reject the null hypothesis of no association between SGA and maternal age.**

d. Using the regression parameter estimates from each of these regressions, provide an estimate of the probability that a 20 year old mother would have a SGA infant. Explain any similarities or differences these estimates might have when compared to the sample proportion of SGA infants among 20 year olds.

***Instructions for grading*: This problem is worth 10 points.**

Ans: There were 40 mothers aged 20 years in the data set, and 3 (7.50%) later delivered SGA infants. Modeled based estimates of the probability of a 20 year old woman delivering an SGA infant were 0.1607 from linear regression, 0.1613 from Poisson regression, and 0.1613 from logistic regression. These model based estimates differ from the simple sample proportion, because information is borrowed across age groups in the regression models, none of which were saturated (there were 29 unique age groups, but each regression model only included two parameters: an intercept and a slope). The similarity between the Poisson and logistic regression estimates will stem from the Poisson approximation to a binomial distribution when the event rate is relatively low. All regression models agreed fairly closely in the central part of the age range, which contains the bulk of the data (see problem 7 below).

7. Produce a plot of the estimated probability of an SGA infant by age as derived by each of the following methods. Comment on the similarity and difference among the various fitted values form the various analyses performed in problem 6. (Note that Stata allows you to specify multiple Y variables for a single X variable: `scatter y1 y2 y3 y4 age`)

   a. Sample proportions within each unique age: This can be obtained in Stata using the command `egen` *`varname`*`= mean(sga), by(age).`

   b. Estimated probabilities for each age in the data as derived from each of the regression analyses. In Stata, this can be obtained using the simple "post-estimation" command: `predict` *`varname.`* (But use a different variable name for each fitted value.)

      i. After performing a linear regression, the default action of the "predict" function is to create a variable that contains the estimated "linear predictor", which corresponds to the regression based estimate of the mean. With a binary response variable, the mean response is the proportion.

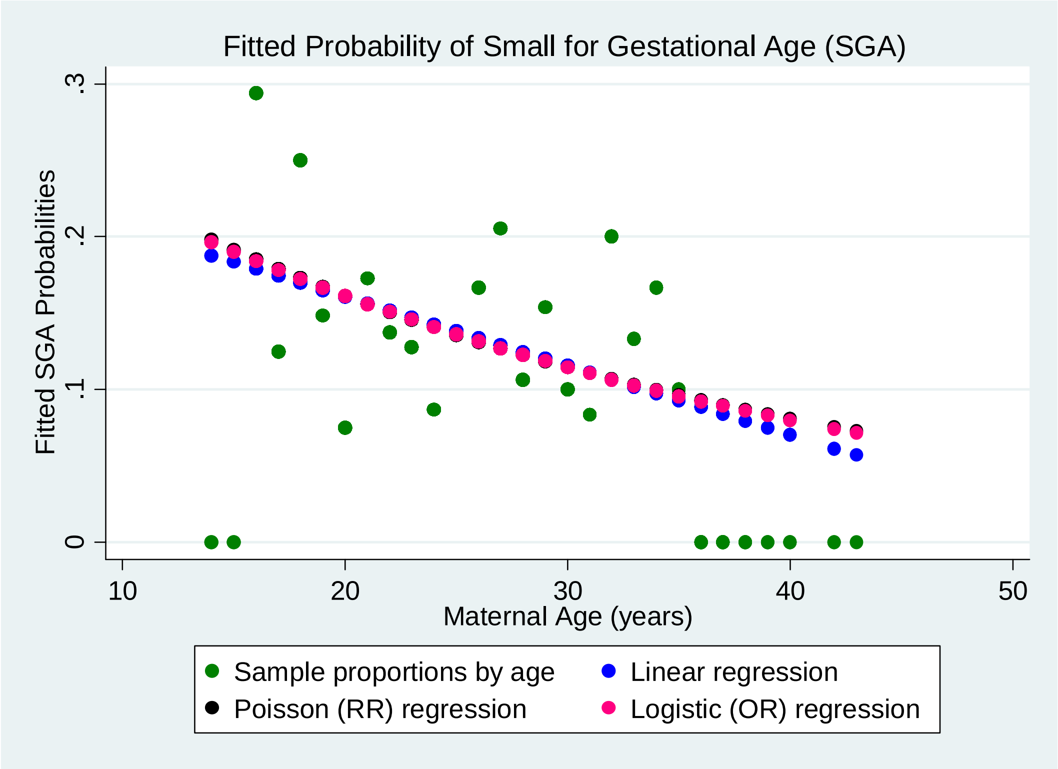
      ii. After performing a Poisson regression, the default action of the "predict" function is to create a variable that contains the exponentiated estimated "linear predictor", which corresponds to the regression based estimate of the mean. With a binary response variable, the mean response is the proportion. (The linear predictor in Poisson regression corresponds to the log "rate", because Poisson regression uses a log link function.

      iii. In logistic regression, the estimated "linear predictor" corresponds to the log odds. Exponentiating that would correspond to the odds. By default, Stata figures that you would really rather have the estimated probability, which is computed as prob = odds / (1 + odds). So, after performing a logistic regression, the default action of the "predict" function is to create a variable that contains the the regression based estimate of the mean.

***Instructions for grading*: This problem is worth 10 points.**

Ans: In the figure shown below, we see that the estimated probability of SGA in each age group is quite variable owing to large sample sizes. The fitted values computed from the three regression models are quite similar where the bulk of the data lies. Note that the fitted probabilities for the linear regression model lie on a straight line as expected for the RD model. The Poisson and logistic regression models each show a curvilinear pattern, as we fit models expecting either the log

probability (Poisson regression) or log odds (logistic regression) to be linear in age. Owing to the relatively low probability of SGA, the Poisson and logistic regression models produce quite similar fitted values.



8. Perform a logistic regression analyses of the distribution of the prevalence of SGA infants across groups defined by the logarithmically transformed maternal age.

    a. Provide formal inference for associations using odds ratio (OR: ratios of odds) and log transformed age.

***Instructions for grading*: This problem is worth 10 points.**

Ans: *Methods:* **Any association between delivery of an infant who was small for gestational age (SGA) and maternal age was summarized by an odds ratio. The estimated odds ratio was computed using the maximum likelihood estimate of the slope from a logistic regression model that only included an logarithmic transformation of maternal age. Using the asymptotically normal distribution for Wald statistics derived from the logistic regression model, two-sided 95% confidence intervals for the true odds ratio were computed. Similarly, two-sided P values testing the null hypothesis that there is no true association between SGA and maternal age (i.e., that the true odds ratio is equal to 1) were based on the normal approximation for the Wald based Z statistic computed from the logistic regression slope parameter. No subjects were missing data for the data analysis.**

Results: **Maternal age averaged 24.8 years (SD 5.4, range 14 – 43) in 755 pregnant women, of whom 13.9% were later observed to give birth to an infant who was small for gestational age (SGA). When comparing groups of women with one group being 50% older than the other (e.g., 24 vs 16 years old or 30 vs 20 years old), the odds of SGA infants is estimated to be 32.1% lower for each 1.5 fold difference in age, with the older women having lower risk of SGA (OR = 0.679). A 95% confidence interval derived from the logistic regression model suggests that the observed data is consistent with a true odds ratio such that the odds of SGA for a population of women is anywhere**

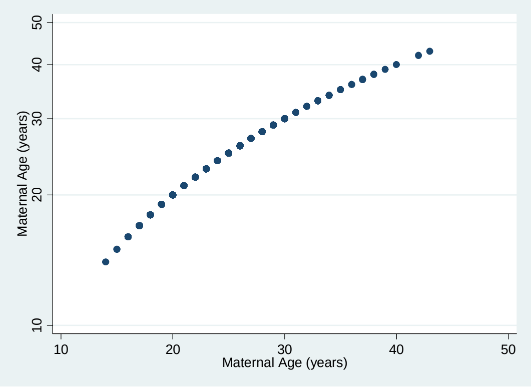
between 54.5% lower to1.38% higher for each 1.5 fold difference in age (95% CI for OR: 0.455, 1.014). Such a result is not statistically significant (two-sided P= 0.058), and we cannot reject the null hypothesis of no association between SGA and maternal age.

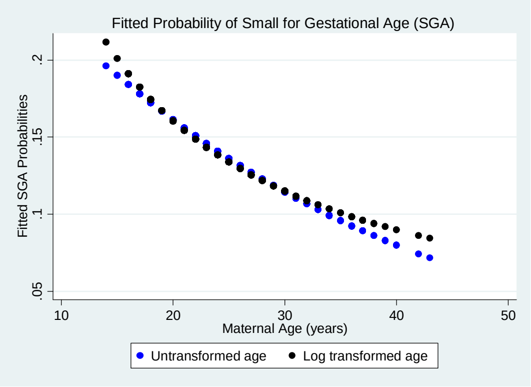
b. Why might it be reasonable or silly to have performed such an analysis rather than the analysis in problem 6c?

***Instructions for grading*: This problem is worth 5 points.**

Ans: It is not very common to consider ages of people on a multiplicative scale, and perhaps even less common when considering the range of ages of pregnant women. Over a short interval, the log function is well approximated by a straight line, so there is little difference in model precision to be expected, even if the true trend were multiplicative. Thus we are sacrificing clarity of communication for no reason.

The following graphs display the relationship between a log transformed age and untransformed age over the range of our data (left panel) and the fitted values that result from the model fit in the problem compared to that in problem 6c. I see little value in using a log transformed variable, because it would be expected to behave just like an untransformed variable.


Maternal Age (years)
Maternal Age (years)


Fitted Probability of Small for Gestational Age (SGA)
Fitted SGA Probabilities
Maternal Age (years)
● Untransformed age
● Log transformed age