BST 210 HW 7

Marina Cheng, Shuhei Miyasaka, Melissa Zhao

November 6, 2017

1. A study was performed to identify risk factors associated with giving birth to a low birth weight baby (< 2500 grams). More information is included in the background file. To control for the important factor of mother's age, one case (with low birth weight) was matched to three controls (without low birth weight), matching on mother's age. STRATUM is the matching variable, with LOW as the outcome variable. Focus on looking at the (linear) effects of other factors, namely LWT, SMOKE, HT, UI, and PTD to predict low birth weight, appropriately adjusting for matching using conditional logistic regression.

a). Perform a backward elimination model selection method by hand using all three controls matched with each case. What important predictor variable(s) are you left with? Write a one or two sentence summary of your findings, suitable for inclusion in a manuscript.

We are left with just the ptd (history of premature labor) covariate in our model. After performing a backward elimination model selection with alpha = 0.05, we are left with a conditional logistic regression model with a single covariate for premature labor after matching cases and controls. Our model suggests that history of premature labor is a significant predictor of low birthweight, after controlling for mother's age.

```
library(survival)
mod1 <- clogit(low ~ lwt + smoke + ht + ui + ptd + strata(stratum), data = lowbwt)
summary(mod1)
```

```
## Call:
## coxph(formula = Surv(rep(1, 116L), low) ~ lwt + smoke + ht +
##     ui + ptd + strata(stratum), data = lowbwt, method = "exact")
##
##   n= 116, number of events= 29
##
##            coef exp(coef)  se(coef)      z Pr(>|z|)
## lwt   -0.005482  0.994533  0.009147 -0.599   0.5490
## smoke  0.553584  1.739477  0.488930  1.132   0.2575
## ht     0.098300  1.103293  1.394870  0.070   0.9438
## ui     0.525356  1.691060  0.549192  0.957   0.3388
## ptd    1.532963  4.631882  0.638241  2.402   0.0163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## lwt      0.9945     1.0055   0.97686     1.013
## smoke    1.7395     0.5749   0.66718     4.535
## ht       1.1033     0.9064   0.07168    16.983
## ui       1.6911     0.5913   0.57635     4.962
## ptd      4.6319     0.2159   1.32583    16.182
##
## Rsquare= 0.127   (max possible= 0.5 )
## Likelihood ratio test= 15.73  on 5 df,   p=0.007671
## Wald test            = 12.47  on 5 df,   p=0.02889
## Score (logrank) test = 17.01  on 5 df,   p=0.004482
```

```r
mod2 <- clogit(low ~ lwt + smoke + ui + ptd + strata(stratum), data = lowbwt)
summary(mod2)
```

```
## Call:
## coxph(formula = Surv(rep(1, 116L), low) ~ lwt + smoke + ui +
##     ptd + strata(stratum), data = lowbwt, method = "exact")
##
##   n= 116, number of events= 29
##
##            coef exp(coef)  se(coef)      z Pr(>|z|)
## lwt   -0.005228  0.994785  0.008400 -0.622   0.5336
## smoke  0.558682  1.748366  0.483577  1.155   0.2480
## ui     0.528152  1.695796  0.547822  0.964   0.3350
## ptd    1.531214  4.623787  0.637735  2.401   0.0163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## lwt      0.9948     1.0052    0.9785     1.011
## smoke    1.7484     0.5720    0.6777     4.511
## ui       1.6958     0.5897    0.5795     4.962
## ptd      4.6238     0.2163    1.3248    16.138
##
## Rsquare= 0.127   (max possible= 0.5 )
## Likelihood ratio test= 15.72  on 4 df,   p=0.003417
## Wald test            = 12.47  on 4 df,   p=0.01419
## Score (logrank) test = 17.01  on 4 df,   p=0.001928
```

```r
mod3 <- clogit(low ~ smoke + ui + ptd + strata(stratum), data = lowbwt)
summary(mod3)
```

```
## Call:
## coxph(formula = Surv(rep(1, 116L), low) ~ smoke + ui + ptd +
##     strata(stratum), data = lowbwt, method = "exact")
##
##   n= 116, number of events= 29
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## smoke 0.6002    1.8225   0.4795 1.252   0.2107
## ui    0.6061    1.8333   0.5405 1.121   0.2622
## ptd   1.5670    4.7921   0.6359 2.464   0.0137 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## smoke     1.823     0.5487    0.7120     4.665
## ui        1.833     0.5455    0.6355     5.289
## ptd       4.792     0.2087    1.3779    16.666
##
## Rsquare= 0.124   (max possible= 0.5 )
## Likelihood ratio test= 15.32  on 3 df,   p=0.001565
## Wald test            = 12.18  on 3 df,   p=0.00678
## Score (logrank) test = 16.57  on 3 df,   p=0.0008642
```

```r
mod4 <- clogit(low ~ smoke + ptd + strata(stratum), data = lowbwt)
summary(mod4)
```

```
## Call:
## coxph(formula = Surv(rep(1, 116L), low) ~ smoke + ptd + strata(stratum),
##     data = lowbwt, method = "exact")
##
##   n= 116, number of events= 29
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## smoke 0.5979    1.8183   0.4757 1.257   0.2088
## ptd   1.7328    5.6564   0.6114 2.834   0.0046 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## smoke     1.818     0.5500    0.7158     4.619
## ptd       5.656     0.1768    1.7064    18.749
##
## Rsquare= 0.114   (max possible= 0.5 )
## Likelihood ratio test= 14.08  on 2 df,   p=0.0008743
## Wald test            = 11.69  on 2 df,   p=0.00289
## Score (logrank) test = 15.53  on 2 df,   p=0.0004252
```

```r
mod5 <- clogit(low ~ ptd + strata(stratum), data = lowbwt)
summary(mod5)
```

```
## Call:
## coxph(formula = Surv(rep(1, 116L), low) ~ ptd + strata(stratum),
##     data = lowbwt, method = "exact")
##
##   n= 116, number of events= 29
##
##       coef exp(coef) se(coef)     z Pr(>|z|)
## ptd 1.9371    6.9389   0.5914 3.275  0.00105 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## ptd     6.939     0.1441     2.177     22.12
##
## Rsquare= 0.102   (max possible= 0.5 )
## Likelihood ratio test= 12.5  on 1 df,   p=0.0004066
## Wald test            = 10.73  on 1 df,   p=0.001055
## Score (logrank) test = 14.08  on 1 df,   p=0.0001749
```

b). Compare your results of the model you end up with above (using all three controls per case) to that of the model using just control = 2 and case = 1 (i.e., a matched pair analysis, ignoring controls 3 and 4). Which model seems to perform better? Why? Which model would you prefer? Why?

The model with three controls per case seems to perform better. The regression coefficient for PTD is more statistically significant for the model with three controls per case than the model with matched pair analysis (p-value 0.00105 vs. 0.0371). We also have a smaller standard error estimate for PTD using the model with three controls per case than the matched pair model (0.5914 vs. 1.054). In general, having more observations will improve the power of our model.

```
## Call:
## coxph(formula = Surv(rep(1, 58L), low) ~ lwt + smoke + ht + ui +
##     ptd + strata(stratum), data = lowbwt2, method = "exact")
##
##   n= 58, number of events= 29
##
##          coef exp(coef) se(coef)     z Pr(>|z|)
## lwt   0.01541   1.01553  0.01338 1.152   0.2493
## smoke 1.37070   3.93810  0.83985 1.632   0.1027
## ht    0.63140   1.88024  1.60773 0.393   0.6945
## ui    1.07497   2.92990  0.99607 1.079   0.2805
## ptd   2.10389   8.19796  1.18529 1.775   0.0759 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## lwt       1.016     0.9847   0.98925     1.043
## smoke     3.938     0.2539   0.75928    20.425
## ht        1.880     0.5318   0.08048    43.925
## ui        2.930     0.3413   0.41591    20.640
## ptd       8.198     0.1220   0.80313    83.681
##
## Rsquare= 0.169   (max possible= 0.5 )
## Likelihood ratio test= 10.77  on 5 df,   p=0.05622
## Wald test            = 5.35  on 5 df,   p=0.3747
## Score (logrank) test = 8.5  on 5 df,   p=0.1308

## Call:
## coxph(formula = Surv(rep(1, 58L), low) ~ lwt + smoke + ui + ptd +
##     strata(stratum), data = lowbwt2, method = "exact")
##
##   n= 58, number of events= 29
##
##          coef exp(coef) se(coef)     z Pr(>|z|)
## lwt   0.01502   1.01513  0.01329 1.130   0.2583
## smoke 1.30191   3.67631  0.82102 1.586   0.1128
## ui    1.03014   2.80146  0.98391 1.047   0.2951
## ptd   2.10052   8.17045  1.17749 1.784   0.0744 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## lwt       1.015     0.9851    0.9890     1.042
## smoke     3.676     0.2720    0.7355    18.377
## ui        2.801     0.3570    0.4073    19.270
## ptd       8.170     0.1224    0.8128    82.135
##
## Rsquare= 0.167   (max possible= 0.5 )
## Likelihood ratio test= 10.61  on 4 df,   p=0.03131
## Wald test            = 5.3  on 4 df,   p=0.2579
## Score (logrank) test = 8.43  on 4 df,   p=0.07712

## Call:
## coxph(formula = Surv(rep(1, 58L), low) ~ lwt + smoke + ptd +
##     strata(stratum), data = lowbwt2, method = "exact")
```

```
##
##   n= 58, number of events= 29
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## lwt   0.009016  1.009056 0.011196 0.805   0.4207
## smoke 0.845362  2.328821 0.657076 1.287   0.1983
## ptd   2.048369  7.755246 1.111973 1.842   0.0655 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## lwt       1.009     0.9910    0.9872     1.031
## smoke     2.329     0.4294    0.6424     8.442
## ptd       7.755     0.1289    0.8772    68.566
##
## Rsquare= 0.15   (max possible= 0.5 )
## Likelihood ratio test= 9.4  on 3 df,   p=0.02444
## Wald test            = 5.61  on 3 df,   p=0.1325
## Score (logrank) test = 8.02  on 3 df,   p=0.04555
```

```
## Call:
## coxph(formula = Surv(rep(1, 58L), low) ~ smoke + ptd + strata(stratum),
##     data = lowbwt2, method = "exact")
##
##   n= 58, number of events= 29
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## smoke 0.6989    2.0115   0.6148 1.137   0.2556
## ptd   1.8962    6.6604   1.0828 1.751   0.0799 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## smoke     2.011     0.4972    0.6028     6.712
## ptd       6.660     0.1501    0.7976    55.616
##
## Rsquare= 0.14   (max possible= 0.5 )
## Likelihood ratio test= 8.73  on 2 df,   p=0.0127
## Wald test            = 5.41  on 2 df,   p=0.06693
## Score (logrank) test = 7.59  on 2 df,   p=0.02254
```

```
## Call:
## coxph(formula = Surv(rep(1, 58L), low) ~ ptd + strata(stratum),
##     data = lowbwt2, method = "exact")
##
##   n= 58, number of events= 29
##
##      coef exp(coef) se(coef)     z Pr(>|z|)
## ptd 2.197     9.000    1.054 2.084   0.0371 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## ptd         9     0.1111      1.14     71.04
##
```

```
## Rsquare= 0.119   (max possible= 0.5 )
## Likelihood ratio test= 7.36  on 1 df,   p=0.006664
## Wald test            = 4.35  on 1 df,   p=0.03712
## Score (logrank) test = 6.4  on 1 df,   p=0.01141
```

```
mod5 <- clogit(low ~ ptd + strata(stratum), data = lowbwt2)
summary(mod5)
```

```
## Call:
## coxph(formula = Surv(rep(1, 58L), low) ~ ptd + strata(stratum),
##     data = lowbwt2, method = "exact")
##
##   n= 58, number of events= 29
##
##      coef exp(coef) se(coef)     z Pr(>|z|)
## ptd 2.197     9.000    1.054 2.084   0.0371 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## ptd         9     0.1111      1.14     71.04
##
## Rsquare= 0.119   (max possible= 0.5 )
## Likelihood ratio test= 7.36  on 1 df,   p=0.006664
## Wald test            = 4.35  on 1 df,   p=0.03712
## Score (logrank) test = 6.4  on 1 df,   p=0.01141
```

c). Go back to using all of the controls for the rest of this problem. It was thought that mother's age would be an important confounding variable. In your model above from part (a) (using all controls), add in the effects of AGE. What do you find? Does that make sense? Briefly explain.

We find that the age covariate does not produce any coefficients in the model, because age is already accounted for inherently in our model through matching cases to controls by age in our study.

```
mod6 <- clogit(low ~ ptd + age + strata(stratum), data = lowbwt)
```

```
## Warning in coxph(formula = Surv(rep(1, 116L), low) ~ ptd + age +
## strata(stratum), : X matrix deemed to be singular; variable 2
```

```
summary(mod6)
```

```
## Call:
## coxph(formula = Surv(rep(1, 116L), low) ~ ptd + age + strata(stratum),
##     data = lowbwt, method = "exact")
##
##   n= 116, number of events= 29
##
##       coef exp(coef) se(coef)     z Pr(>|z|)
## ptd 1.9371    6.9389   0.5914 3.275  0.00105 **
## age     NA        NA   0.0000    NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## ptd     6.939     0.1441     2.177     22.12
## age        NA         NA        NA        NA
```

```
##
## Rsquare= 0.102   (max possible= 0.5 )
## Likelihood ratio test= 12.5  on 1 df,   p=0.0004066
## Wald test            = 10.73  on 1 df,   p=0.001055
## Score (logrank) test = 14.08  on 1 df,   p=0.0001749
```

d). Another investigator suggests that if integer mother's age was matched on, one could use AGE (rather than STRATUM) as the matching variable. Do you agree or not? Briefly explain. And, if you did that, how do your results change? Which do you prefer (and why)?

Yes, you can use age as the matching variable. However, the results are different between the model matched on stratum and the model matched on age because when we match on age, we are no longer fitting a model with three controls per case (e.g., there are instances of 2 cases per 6 controls or 3 cases per 9 controls.) Our matched groups are different when we match on stratum vs age. We prefer to match on stratum because the study was designed in that way.

```
mod7 <- clogit(low ~ ptd + strata(age), data = lowbwt)
summary(mod7)
```

```
## Call:
## coxph(formula = Surv(rep(1, 116L), low) ~ ptd + strata(age),
##     data = lowbwt, method = "exact")
##
##   n= 116, number of events= 29
##
##       coef exp(coef) se(coef)     z Pr(>|z|)
## ptd 1.8061    6.0866   0.5394 3.349 0.000812 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##     exp(coef) exp(-coef) lower .95 upper .95
## ptd     6.087     0.1643     2.115     17.52
##
## Rsquare= 0.098   (max possible= 0.562 )
## Likelihood ratio test= 11.96  on 1 df,   p=0.0005446
## Wald test            = 11.21  on 1 df,   p=0.0008124
## Score (logrank) test = 13.81  on 1 df,   p=0.0002026
```

e). Overall, do we have any statistical evidence that it was important to adjust for (matched) AGE or STRATUM? Why or why not? Briefly explain. (However, if one designs a study using matching, one should analyze the study using matching. And, if one matches on strata based on age, one cannot estimate the influence of age on the outcome.)

We cannot obtain any statistical evidence of whether age is a statistically meaningful confounder as we are matching upon age (thus its effects are embedded in the alpha coefficients, which we cannot measure). However, by definition, age is associated with history of premature labor, associated with birth weight regardless of history of premature labor, and is not a downstream consequence of history of premature labor and birth weight. Thus, age is a potential confounder of the association between history of premature labor and birth weight, therefore it is a good idea to match on it.

f). Can we assess whether or not age is an effect modifier of any of the other variables you have found to be statistically significant above? If so, assess potential effect modification, or if not, briefly explain why not.

Age is not an effect modifier of ptd on low birthweight, as the ptd x age interaction term has a P value of 0.535, which is not significant at the alpha = 0.05 level.

```r
mod7 <- clogit(low ~ ptd + ptd*age + strata(stratum), data = lowbwt)
```

```
## Warning in coxph(formula = Surv(rep(1, 116L), low) ~ ptd + ptd * age +
## strata(stratum), : X matrix deemed to be singular; variable 2
```

```r
summary(mod7)
```

```
## Call:
## coxph(formula = Surv(rep(1, 116L), low) ~ ptd + ptd * age + strata(stratum),
##     data = lowbwt, method = "exact")
##
##   n= 116, number of events= 29
##
##             coef exp(coef) se(coef)      z Pr(>|z|)
## ptd      3.77468  43.58360  3.06153  1.233    0.218
## age           NA        NA  0.00000     NA       NA
## ptd:age -0.07768   0.92526  0.12519 -0.621    0.535
##
##         exp(coef) exp(-coef) lower .95 upper .95
## ptd       43.5836    0.02294    0.1080 17591.534
## age            NA         NA        NA        NA
## ptd:age    0.9253    1.08078    0.7239     1.183
##
## Rsquare= 0.105   (max possible= 0.5 )
## Likelihood ratio test= 12.89  on 2 df,   p=0.001591
## Wald test            = 10.61  on 2 df,   p=0.004976
## Score (logrank) test = 14.63  on 2 df,   p=0.0006656
```

g). Finally, compare your results to that where you use (unconditional) logistic regression on the whole sample, and you adjust for age as a covariate. Is this a good approach to use? Why or why not?

We get similar beta coefficients for PTD between the conditional logistic regression (1.94, p-value: 0.001) and unconditional logistic regression model (1.97, p-value: 0.0004). In addition, our standard errors around PTD is similar between the conditional logistic regression (0.591) and logistic regression models (0.554). However, running an unconditional logistic regression is not a good approach because this was a case-control study. We need to take into account the case-control study design in our modeling.

```r
mod8 <- glm(low ~ ptd + age, data = lowbwt, family = "binomial")
summary(mod8)
```

```
##
## Call:
## glm(formula = low ~ ptd + age, family = "binomial", data = lowbwt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4063  -0.6568  -0.6292  -0.2046   1.9202
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.0631     1.2132  -0.876  0.38090
## ptd           1.9667     0.5546   3.546  0.00039 ***
## age          -0.0190     0.0528  -0.360  0.71891
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 130.46  on 115  degrees of freedom
## Residual deviance: 117.40  on 113  degrees of freedom
## AIC: 123.4
##
## Number of Fisher Scoring iterations: 4
```

2. A large study was performed looking at the dose-response effects of cigarette smoking on lung cancer incidence in British male physicians. The data to be analyzed were presented in Frome (Biometrics, 1983) and originally were collected by Doll and Hill. The data are given in the fromelungcancer file. We'll be fitting a variety of models and making model comparisons using likelihood ratio tests, Akaike's information criteria, assessment of goodness of fit, and related methods.

smokedur = smoking duration (in years, set equal to age - 20) cigpday = average cigarette smoking per day CASES = # of cases of lung cancer MANYEARS = man-years of follow-up

a). Our main interest is in the effects of cigpday on lung cancer incidence. Do we have any evidence that (linear) smokedur (a surrogate for both age and smoking duration, given it is coded as age - 20, under the assumption that most smokers started smoking around age 20) is a confounder or an effect modifier of the effects of (linear) cigpday on (log of) lung cancer incidence? Justify your responses, and summarize your overall findings briefly (e.g., in terms of incidence rate ratios).

Smoking duration is associated with cigarettes/day, associated with lung cancer regardless of cigarettes/day, and is not a downstream consequence of cigarettes/day or lung cancer. Thus, smoking duration meets the definition of a confounder. However, we do not have evidence that smoking duration is a meaningful confounder or significant effect modifier of linear cigarettes/day on log of lung cancer incidence. Adding smoking duration into our model does not change the beta coefficient of cigarettes/day by more than 10% (from 0.07 to 0.067), and adding the interaction term of cigarettes/day and smoking duration into our model does not yield a beta coefficient that is statistically significant (p = 0.68). However, smoking duration is a significant predictor of lung cancer (p-value <0.0001). Therefore, we prefer the model with covariates for smoking duration and cigarettes per day.

Using the model with smoking duration and cigarettes/day, our overall finding is that the incidence rate of lung cancer is 1.07 times higher for every one unit increase in cigarettes per day, holding smoking duration constant. The incidence rate of lung cancer is 1.12 times higher with every one unit increases in smoking duration, holding cigarettes per day constant.

```
mod9 <- glm(cases ~ cigpday, offset = log(manyears), data = lungca, family = "poisson")
summary(mod9)
```

```
##
## Call:
## glm(formula = cases ~ cigpday, family = "poisson", data = lungca,
##     offset = log(manyears))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8684  -1.6542  -0.5111   1.8415   4.8937
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.159268   0.175063  -46.61   <2e-16 ***
## cigpday      0.070359   0.006468   10.88   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 445.10  on 62  degrees of freedom
## Residual deviance: 324.71  on 61  degrees of freedom
## AIC: 448.55
##
## Number of Fisher Scoring iterations: 6
```

```
mod10 <- glm(cases ~ cigpday + smokedur, offset = log(manyears), data = lungca, family = "poisson")
summary(mod10)
```

```
##
## Call:
## glm(formula = cases ~ cigpday + smokedur, family = "poisson",
##     data = lungca, offset = log(manyears))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2706  -1.1369  -0.4592   0.6675   2.4863
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.215635   0.377854  -32.33   <2e-16 ***
## cigpday       0.066712   0.006323   10.55   <2e-16 ***
## smokedur      0.111457   0.007508   14.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 445.099  on 62  degrees of freedom
## Residual deviance:  82.243  on 60  degrees of freedom
## AIC: 208.09
##
## Number of Fisher Scoring iterations: 5
```

```
mod11 <- glm(cases ~ cigpday + smokedur + cigpday*smokedur, offset = log(manyears), data = lungca, famil
summary(mod11)
```

```
##
## Call:
## glm(formula = cases ~ cigpday + smokedur + cigpday * smokedur,
##     family = "poisson", data = lungca, offset = log(manyears))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3352  -1.1180  -0.4311   0.6454   2.4962
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.247e+01  7.386e-01 -16.887  < 2e-16 ***
## cigpday           7.784e-02  2.796e-02   2.784  0.00537 **
## smokedur          1.172e-01  1.598e-02   7.335 2.21e-13 ***
```

```
## cigpday:smokedur -2.516e-04  6.152e-04  -0.409  0.68256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 445.099  on 62  degrees of freedom
## Residual deviance:  82.076  on 59  degrees of freedom
## AIC: 209.92
##
## Number of Fisher Scoring iterations: 5
```

b). Consider the model looking at the effects of (linear) cigpday on lung cancer incidence, adjusting for (linear) smokedur (with no interaction). What is a point estimate and 95% CI for the IRR for the effects of 20 cigarettes/day, adjusting for smoking duration? Also, does this model show evidence of lack of fit? Considering how many cases of lung cancer occurred in the dataset and the number of covariate patterns, do you trust a goodness-of-fit test? Briefly comment.

The incidence rate of lung cancer is 3.8 times higher among those who smoke 20 cigarettes/day to those who don't smoke any cigarettes/day, adjusting for smoking duration. With 95% confidence, the incidence rate ratio comparing those who smoke 20 cigarettes/day to those who don't smoke any cigarettes/day is between 2.96 and 4.87.

Based on the LRT test between the intercept only model and model with linear effects of cigpday adjusting for linear smokedur, we conclude that the latter model is preferred (p-value < 0.0001). The goodness-of-fit was also assessed by looking at the pearson chi square statistic (Lab 11, pg 9):

Pearson chi square statistic: 64.3, p-value=.328

However, we don't trust the goodness-of-fit test since we have 63 covariate patterns for small number of cases. Consequently, we won't have enough observation in each covariate pattern. Therefore, we don't think that our calculated Pearson chi square statistic follows a chi square distribution with 60 degrees of freedom.

Since we have a large number of covariate patterns relative to the sample size, we instead decided to do a more empirical check on the goodness-of-fit:

Deviance / (J - (p+1)) = 1.37 Chi2 statistics / (J - (p+1)) = 1.07

However, we don't think both the Pearson chi square test statistic and the empirical checks are appropriate for this scenario. Therefore, we should still be concerned about overdispersion.

c). Consider a model including linear and quadratic effects of both cigpday and smokedur. Does this model show improvements relative to the model including only linear covariates? Using this model, calculate a point estimate and 95% CI for the IRR for the effects of 20 vs. 0 cigarettes/day, and for the effects of 40 vs. 20 cigarettes/day, adjusting for linear and quadratic smokedur. Note that these point estimates and confidence intervals are not the same due to the quadratic effects of cigpday included in this model.

A likelihood ratio test was conducted to assess whether a model with linear and quadratic effects of both cigpday and smokedur is preferred over the model including only linear covariates. Based on the LRT, we conclude that the model with linear and quadratic effects of both cigpday and smokedur is preferred over the model including only linear covariates (p-value $<$ 0.0001). The point estimate and 95% CI for the IRR for the effects of 20 vs. 0 cigarettes/day are 10.45 and (5.25, 20.8), respectively. The point estimate and 95% CI for the IRR for the effects of 40 vs. 20 cigarettes/day are 2.22 and (1.5, 3.28), respectively.

```
mod12 <- glm(cases ~ cigpday + smokedur + I(cigpday^2) + I(smokedur^2), offset = log(manyears), data = l
summary(mod12)
```

```
##
## Call:
## glm(formula = cases ~ cigpday + smokedur + I(cigpday^2) + I(smokedur^2),
##     family = "poisson", data = lungca, offset = log(manyears))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1705  -0.9117  -0.4424   0.6165   1.8939
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.607e+01  1.358e+00 -11.836  < 2e-16 ***
## cigpday        1.561e-01  2.799e-02   5.576 2.46e-08 ***
## smokedur       2.667e-01  6.520e-02   4.090 4.32e-05 ***
## I(cigpday^2)  -1.938e-03  5.565e-04  -3.483 0.000496 ***
## I(smokedur^2) -1.855e-03  7.735e-04  -2.398 0.016503 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 445.099  on 62  degrees of freedom
## Residual deviance:  61.369  on 58  degrees of freedom
## AIC: 191.21
##
## Number of Fisher Scoring iterations: 5
```

```
anova(mod10, mod12)
```

```
## Analysis of Deviance Table
##
## Model 1: cases ~ cigpday + smokedur
## Model 2: cases ~ cigpday + smokedur + I(cigpday^2) + I(smokedur^2)
##   Resid. Df Resid. Dev Df Deviance
## 1        60     82.243
## 2        58     61.369  2   20.875
```

```
#IRR point estimates
lIRR_20_0 <- 20*coef(mod12)[2] + 400*coef(mod12)[4]
lIRR_40_20 <- 20*coef(mod12)[2] + 1200*coef(mod12)[4]
exp(lIRR_20_0)
```

```
##  cigpday
## 10.45361
```

```
exp(lIRR_40_20)
```

```
##  cigpday
## 2.217759
```

```
#CIs of estimates
se_20_0 <- sqrt((20^2)*vcov(mod12)[2,2] + (400^2)*vcov(mod12)[4,4] + 2*(20*400)*vcov(mod12)[2,4])
se_40_20 <- sqrt((20^2)*vcov(mod12)[2,2] + (1200^2)*vcov(mod12)[4,4] + 2*(20*1200)*vcov(mod12)[2,4])
exp(lIRR_20_0 + c(-1,1)*qnorm(.975)*se_20_0)
```

```
## [1]  5.253286 20.801819
```

```r
exp(lIRR_40_20 + c(-1,1)*qnorm(.975)*se_40_20)
```

```
## [1] 1.497791 3.283804
```

d). The model in (c) does not include any interaction terms. Run two interaction models, each including the linear and quadratic effects of both cigpday and smokedur as main effects. In the first interaction model, just add the cigpday*smokedur interaction term (one parameter). In the second interaction model, add in interactions between the linear and quadratic effects of cigpday and smokedur (so four interaction parameters needed). Do we have any evidence that effect modification is occurring? Justify your response.

No, we do not have any evidence that effect modification is occurring. The interaction term in the model with just one interaction term of cigpday x smokedur is not statistically significant (p-value = 0.4).

Assessing the fit of the second interaction model with interactions between the linear and quadratic effects of cigpday and smokedur also led to the same conclusion. A LRT was conducted to assess whether the model with the 4 interactions terms was preferred over the model with no interaction terms (model from part [c]). The results from the LRT suggest that we prefer the model with no interaction terms (p-value = 0.05366) at the 0.05 significance level.

```r
mod13 <- glm(cases ~ cigpday + smokedur + I(cigpday^2) + I(smokedur^2) + cigpday*smokedur, offset = log
summary(mod13)
```

```
##
## Call:
## glm(formula = cases ~ cigpday + smokedur + I(cigpday^2) + I(smokedur^2) +
##     cigpday * smokedur, family = "poisson", data = lungca, offset = log(manyears))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1292  -0.8399  -0.4244   0.5668   2.1257
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.688e+01  1.719e+00  -9.820  < 2e-16 ***
## cigpday           1.903e-01  5.073e-02   3.751 0.000176 ***
## smokedur          2.854e-01  7.017e-02   4.067 4.76e-05 ***
## I(cigpday^2)     -1.998e-03  5.656e-04  -3.532 0.000412 ***
## I(smokedur^2)    -1.880e-03  7.797e-04  -2.411 0.015918 *
## cigpday:smokedur -7.090e-04  8.513e-04  -0.833 0.404985
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 445.099  on 62  degrees of freedom
## Residual deviance:  60.666  on 57  degrees of freedom
## AIC: 192.51
##
## Number of Fisher Scoring iterations: 6
```

```r
mod14 <- glm(cases ~ cigpday + smokedur + I(cigpday^2) + I(smokedur^2) + cigpday*smokedur + I(cigpday^2
summary(mod14)
```

```
##
## Call:
## glm(formula = cases ~ cigpday + smokedur + I(cigpday^2) + I(smokedur^2) +
```

```
##     cigpday * smokedur + I(cigpday^2) * I(smokedur^2), family = "poisson",
##     data = lungca, offset = log(manyears))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0959  -0.8462  -0.4605   0.5916   1.7230
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -1.472e+01  2.286e+00  -6.437 1.22e-10 ***
## cigpday                      5.214e-02  1.177e-01   0.443   0.6579
## smokedur                     2.130e-01  8.743e-02   2.436   0.0148 *
## I(cigpday^2)                -4.077e-04  1.369e-03  -0.298   0.7659
## I(smokedur^2)               -1.373e-03  8.597e-04  -1.598   0.1101
## cigpday:smokedur             2.457e-03  2.648e-03   0.928   0.3534
## I(cigpday^2):I(smokedur^2) -8.144e-07  6.627e-07  -1.229   0.2191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 445.099  on 62  degrees of freedom
## Residual deviance:  59.186  on 56  degrees of freedom
## AIC: 193.03
##
## Number of Fisher Scoring iterations: 6
```

e). Because quadratic effects seem to be statistically significant, we might also want to run models that were even more complicated than quadratic. Given the small number of (effectively categorical) cigpday and smokedur values, using generalized additive models or restricted cubic splines does not seem appealing. Those methods are more effective when you have a truly continuous covariate. Instead, fit a model with categorical cigpday and smokedur, but no interaction. Using this model, calculate a point estimate and 95% CI for the IRR for the effects of 20.4 vs. 0 cigarettes/day, and for the effects of 40.8 vs. 20.4 cigarettes/day, adjusting for categorical smokedur. Note that these point estimates and confidence intervals are not the same due to the categorical (rather than linear) effects of cigpday included in this model.

The point estimate and 95% CI for the IRR for the effects of 20.4 vs. 0 cigarettes/day (adjusting for categorical smokedur) are 18.19 and (5.66, 58.45), respectively. The point estimate and 95% CI for the IRR for the effects of 40.8 vs. 20.4 cigarettes/day (adjusting for categorical smokedur) are 2.02 and (1.29, 3.17), respectively.

```
mod14 <- glm(cases ~ as.factor(cigpday) + as.factor(smokedur), offset = log(manyears), data = lungca, fa
summary(mod14)
```

```
##
## Call:
## glm(formula = cases ~ as.factor(cigpday) + as.factor(smokedur),
##     family = "poisson", data = lungca, offset = log(manyears))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8329  -0.8560  -0.3808   0.4241   2.1762
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -12.5784     1.1475 -10.961  < 2e-16 ***
```

```
## as.factor(cigpday)5.2     1.2200     0.7073   1.725 0.084547 .
## as.factor(cigpday)11.2    2.0991     0.6363   3.299 0.000971 ***
## as.factor(cigpday)15.9    2.3089     0.6327   3.649 0.000263 ***
## as.factor(cigpday)20.4    2.9009     0.5956   4.870 1.11e-06 ***
## as.factor(cigpday)27.4    3.1162     0.5947   5.240 1.61e-07 ***
## as.factor(cigpday)40.8    3.6059     0.6048   5.962 2.49e-09 ***
## as.factor(smokedur)22.5   0.9469     1.1548   0.820 0.412202
## as.factor(smokedur)27.5   1.7016     1.0805   1.575 0.115284
## as.factor(smokedur)32.5   3.2029     1.0204   3.139 0.001695 **
## as.factor(smokedur)37.5   3.2423     1.0242   3.166 0.001547 **
## as.factor(smokedur)42.5   4.2088     1.0137   4.152 3.30e-05 ***
## as.factor(smokedur)47.5   4.4476     1.0171   4.373 1.23e-05 ***
## as.factor(smokedur)52.5   4.9048     1.0201   4.808 1.52e-06 ***
## as.factor(smokedur)57.5   5.4134     1.0239   5.287 1.24e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 445.099  on 62  degrees of freedom
## Residual deviance:  51.471  on 48  degrees of freedom
## AIC: 201.31
##
## Number of Fisher Scoring iterations: 6
```

```
#IRR
lIRR_20.4_0 <- coef(mod14)[5]
lIRR_40.8_20.4 <- coef(mod14)[7] - coef(mod14)[5]
exp(lIRR_20.4_0)
```

```
## as.factor(cigpday)20.4
##               18.18981
```

```
exp(lIRR_40.8_20.4)
```

```
## as.factor(cigpday)40.8
##               2.023964
```

```
#CI
se_20.4_0 <- sqrt(vcov(mod14)[5,5])
se_40.8_20.4 <- sqrt(vcov(mod14)[7,7] + vcov(mod14)[5,5] -2*vcov(mod14)[5,7])
exp(lIRR_20.4_0 + c(-1,1)*qnorm(0.975)*se_20.4_0)
```

```
## [1]  5.66044 58.45294
```

```
exp(lIRR_40.8_20.4 + c(-1,1)*qnorm(0.975)*se_40.8_20.4)
```

```
## [1] 1.293136 3.167826
```

f). Consider whether one of the extensions to Poisson regression modeling would be helpful with this data analysis. Choosing either (your choice, pick one) quadratic or categorical effects of cigpday and smokedur in a model (with no interaction terms), suggest whether or not you feel your new model is helpful.

We decided the model with linear and quadratic effects of cigpday and smokedur is better than the categorical model based on the AIC score (191 vs. 201). We decided to compare our variance estimates from the Poisson regression model with linear and quadratic effects of cigpday and smokedur against the variance estimated using the robust variance estimation method. We found that the standard errors do not differ much. We also compared our model

standard errors to robust standard errors. The results are tabulated below:

```
##                Estimate Model SE Robust SE  Ratio
## (Intercept)    -16.073    1.358     1.209  0.890
## cigpday          0.156    0.028     0.027  0.950
## smokedur         0.267    0.065     0.057  0.880
## I(cigpday^2)    -0.002    0.001     0.001  0.907
## I(smokedur^2)   -0.002    0.001     0.001  0.906
```

The table results show that the standard errors estimated from the Poisson regression are similar to the standard errors estimated using robust estimation method. Hence, the assumption we made in our Poisson regression, that the mean and variance are equal, seems to be a reasonable one.

We also tested a negative binomial model with categorical effects of cigpday and smokedur and found that it produces the same betas as our poisson model with categorical effects of cigpday and smokedur, thus, it is not very helpful.

```
#Robust variance with quadratic cigpday and smokedur
mod16 <- glm(cases ~ cigpday + smokedur + I(cigpday^2) + I(smokedur^2),
                  data = lungca, family="quasipoisson")
summary(mod16)
```

```
##
## Call:
## glm(formula = cases ~ cigpday + smokedur + I(cigpday^2) + I(smokedur^2),
##     family = "quasipoisson", data = lungca)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8269  -0.9448  -0.4328   0.4736   2.5958
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -9.9324956  2.2134176  -4.487 3.47e-05 ***
## cigpday        0.2264671  0.0520492   4.351 5.56e-05 ***
## smokedur       0.3948714  0.1035101   3.815 0.000333 ***
## I(cigpday^2)  -0.0038160  0.0009999  -3.816 0.000331 ***
## I(smokedur^2) -0.0043563  0.0012272  -3.550 0.000772 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 2.514775)
##
##     Null deviance: 253.288  on 62  degrees of freedom
## Residual deviance:  78.837  on 58  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

```
#Negative Binomial with categorical cigpday and smokedur
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
```

```
##     select
```

```
mod16 <- glm.nb(cases ~ as.factor(cigpday) + as.factor(smokedur) + offset(log(manyears)), data = lungca
summary(mod16)
```

```
##
## Call:
## glm.nb(formula = cases ~ as.factor(cigpday) + as.factor(smokedur) +
##     offset(log(manyears)), data = lungca, link = log, init.theta = 5028.906533)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8326  -0.8560  -0.3808   0.4241   2.1761
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -12.5784     1.1475 -10.961  < 2e-16 ***
## as.factor(cigpday)5.2      1.2200     0.7074   1.725 0.084565 .
## as.factor(cigpday)11.2     2.0991     0.6363   3.299 0.000971 ***
## as.factor(cigpday)15.9     2.3089     0.6328   3.649 0.000263 ***
## as.factor(cigpday)20.4     2.9008     0.5957   4.870 1.12e-06 ***
## as.factor(cigpday)27.4     3.1163     0.5948   5.240 1.61e-07 ***
## as.factor(cigpday)40.8     3.6058     0.6049   5.961 2.50e-09 ***
## as.factor(smokedur)22.5    0.9469     1.1548   0.820 0.412236
## as.factor(smokedur)27.5    1.7015     1.0805   1.575 0.115314
## as.factor(smokedur)32.5    3.2029     1.0204   3.139 0.001696 **
## as.factor(smokedur)37.5    3.2421     1.0242   3.165 0.001549 **
## as.factor(smokedur)42.5    4.2088     1.0138   4.152 3.30e-05 ***
## as.factor(smokedur)47.5    4.4475     1.0171   4.373 1.23e-05 ***
## as.factor(smokedur)52.5    4.9049     1.0202   4.808 1.53e-06 ***
## as.factor(smokedur)57.5    5.4134     1.0239   5.287 1.24e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(5028.907) family taken to be 1)
##
##     Null deviance: 444.862  on 62  degrees of freedom
## Residual deviance:  51.455  on 48  degrees of freedom
## AIC: 203.33
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  5029
##           Std. Err.:  20348
## Warning while fitting theta: alternation limit reached
##
##  2 x log-likelihood:  -171.331
```

g). Review the various models you have run above (plus any others you may decide to run) and determine which model you think fits the data best. Briefly describe your reasoning in choosing your model, and briefly describe your findings in a few sentences, and possibly including a small table, if you think that is appropriate.

From part (f) we decided that the assumption of the mean and the variance being equal was a reasonable one. Hence, the Poisson regression model is an appropriate model to use to evaluate the dose-response effects of cigarette smoking on lung cancer incidence. We also decided from

part (f) that the model with linear and quadratic effects of cigpday and smokedur is better than the categorical model based on the AIC score (191 vs. 201). We also determined in part (d) that the we prefer the model with linear and quadratic effects of cigpday and smokedur with no interaction terms. When comparing the AIC scores between the model with and without the interaction terms, we notice that the AIC score is slightly better for the model with the interaction term (190 vs 191). However, the difference is so small that our conclusion doesn't change, and we prefer the more parsimonious model without the interaction term.