Stat 539 Homework 3

Kenny Flagg

February 14, 2017

1. *Recall from equation (4.15) on p. 133 that the deviance of a model M is equal to:*

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n} w_i \left[y_i \left(\tilde{\theta}_i - \hat{\theta}_i \right) - b\left(\tilde{\theta}_i \right) + b\left(\hat{\theta}_i \right) \right].$$

Let $\hat{\mu}_i$ be the fitted values for model M. Use this expression of the deviance to show that

(a) *the deviance for a Normal GLM with identity link is equal to:*

$$\sum_{i=1}^{n} \left(y_i - \hat{\mu}_i \right)^2.$$

The Normal distribution has $\theta_i = \mu_i$, $b(\theta_i) = \frac{\theta_i^2}{2} = \frac{\mu_i^2}{2}$, and canonical link $g(\mu_i) = \mu_i$. Also, $a(\phi) = \frac{\phi}{w_i} = \sigma^2$ so $\phi = \sigma^2$ and $w_i = 1$. In the saturated model, $\tilde{\theta}_i = y_i$, so

$$\begin{aligned}
D\left(\mathbf{y};\hat{\boldsymbol{\mu}}\right) &= 2\sum_{i=1}^{n}(1)\left[y_i \left(y_i - \hat{\mu}_i \right) - b(y_i) + b\left(\hat{\mu}_i \right) \right] \\
&= 2\sum_{i=1}^{n}\left[y_i^2 - y_i\hat{\mu}_i - \frac{y_i^2}{2} + \frac{\hat{\mu}_i^2}{2} \right] \\
&= \sum_{i=1}^{n}\left(y_i^2 - 2y_i\hat{\mu}_i + \hat{\mu}_i^2 \right) \\
&= \sum_{i=1}^{n}\left(y_i - \hat{\mu}_i \right)^2.
\end{aligned}$$

(b) *the deviance for a Poisson GLM with log link is equal to:*

$$2\sum_{i=1}^{n}\left(y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right)-(y_i-\hat{\mu}_i)\right).$$

The Poisson distribution has $\theta_i = \log(\mu_i)$, $b(\theta_i) = \exp(\theta_i) = \mu_i$, and canonical link $g(\mu_i) = \log(\mu_i)$. And $a(\phi) = \dfrac{\phi}{w_i} = 1$ so $\phi = 1$ and $w_i = 1$. In the saturated model, $\tilde{\theta}_i = \log(y_i)$, so

$$\begin{aligned}
D\left(\mathbf{y};\hat{\boldsymbol{\mu}}\right) &= 2\sum_{i=1}^{n}(1)\left[y_i\left(\log(y_i)-\log(\hat{\mu}_i)\right)-b(y_i)+b\left(\hat{\mu}_i\right)\right] \\
&= 2\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right)-y_i+\hat{\mu}_i\right] \\
&= 2\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right)-(y_i-\hat{\mu}_i)\right].
\end{aligned}$$

2. *Consider the simple case where $nY \sim Bin(n,\pi)$ and we would like to test the null hypothesis $H_0 : \pi = \pi_0$ versus $H_a : \pi \neq \pi_0$.*

(a) *Show that the chi-squared forms of the test statistics are:*

$$\text{Likelihood-ratio: } -2(\ell_0-\ell_1) = -2\log\left[\frac{\pi_0^{ny}(1-\pi_0)^{n(1-y)}}{y^{ny}(1-y)^{n(1-y)}}\right].$$

$$\text{Wald: } z^2 = \frac{(y-\pi_0)^2}{[y(1-y)/n]}$$

$$\text{Score: } z^2 = \frac{(y-\pi_0)^2}{[\pi_0(1-\pi_0)/n]}$$

Likelihood-ratio: The log-likelihood is

$$\ell(\pi;y) = \log\left(\binom{n}{ny}\pi^{ny}(1-\pi)^{n(1-y)}\right) = \log\binom{n}{ny} + \log\left(\pi^{ny}(1-\pi)^{n(1-y)}\right).$$

The MLE under H_a is $\hat{\pi} = y$ so the LRT statistic is

$$\begin{aligned}
-2\left(\ell_0-\ell_1\right) &= -2\left(\ell(\pi_0;y)-\ell(y;y)\right) \\
&= -2\left(\log\binom{n}{ny}+\log\left(\pi_0^{ny}(1-\pi_0)^{n(1-y)}\right)\right. \\
&\qquad\left.-\log\binom{n}{ny}-\log\left(y^{ny}(1-y)^{n(1-y)}\right)\right) \\
&= -2\log\left(\frac{\pi_0^{ny}(1-\pi_0)^{n(1-y)}}{y^{ny}(1-y)^{n(1-y)}}\right).
\end{aligned}$$

Wald: The variance function is $var(Y) = v(\pi) = \dfrac{\pi(1-\pi)}{n}$, which is also the variance of $\hat{\pi} = y$. Then the estimated variance of $\hat{\pi}$ is $\widehat{var}(\hat{\pi}) = \dfrac{y(1-y)}{n}$ so the Wald statistic is

$$z^2 = \frac{(\hat{\pi} - \pi_0)^2}{\widehat{var}(\hat{\pi})} = \frac{(y - \pi_0)^2}{y(1-y)/n}.$$

Score: The variance of $\hat{\pi} = y$ under H_0 is $var(\hat{\pi}) = \frac{\pi_0(1-\pi_0)}{n}$ so the score statistic is

$$z^2 = \frac{(\hat{\pi} - \pi_0)^2}{var(\hat{\pi})} = \frac{(y - \pi_0)^2}{\pi_0(1-\pi_0)/n}.$$

(b) *A recent study examined expressions of commitment between two partners in a committed romantic relationship. One aspect of the study involved 47 heterosexual couples who are part of an online pool of people willing to participate in surveys. These 47 couples were asked about which person was the first to say "I love you." In 26 of the 47 couples, the male said "I love you" first. Set up a hypothesis test to test the null hypothesis that males and females in a committed romantic relationship are equally likely to say "I love you" first, defining the parameter of interest in context. Calculate the likelihood-ratio test statistic, the Wald test statistic, and the score test statistic, and the corresponding p-value for each. Do the three tests yield different conclusions? Write a conclusion of the study in context of the problem (using the LRT p-value).*

The parameter of interest is π, the proportion of committed heterosexual couples where the male said "I love you" first. We are testing

- H_0: $\pi = \frac{1}{2}$, the male and female are equally likely to say "I love you" first;
- H_a: $\pi \neq \frac{1}{2}$, the male and female are not equally likely to say "I love you" first.

In $n = 47$ couples, we observed $y = \frac{26}{47}$.

Likelihood-ratio: Under H_0, the LRT statistic follows a χ^2_1 distribution. The observed value of the statistic is

$$-2\log\left[\frac{\left(\frac{1}{2}\right)^{26}\left(\frac{1}{2}\right)^{21}}{\left(\frac{26}{47}\right)^{26}\left(\frac{21}{47}\right)^{21}}\right] \approx 0.533$$

with a p-value of 0.4654.

Wald: Under H_0, the Wald statistic follows a χ^2_1 distribution. The observed value of the statistic is

$$\frac{\left(\frac{26}{47} - \frac{1}{2}\right)^2}{\left(\frac{26}{47}\right)\left(\frac{21}{47}\right)/47} \approx 0.538$$

with a p-value of 0.4633.

Score: Under H_0, the Score statistic follows a χ^2_1 distribution. The observed value of the statistic is

$$\frac{\left(\frac{26}{47} - \frac{1}{2}\right)^2}{\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)/47} \approx 0.532$$

with a p-value of 0.4658.

In this case, all three tests yield similar test statistics and p-values. There is no evidence ($\chi^2_1 = 0.533$, p-value $= 0.4654$) that the true proportion of committed heterosexual couples in this pool of survey-takers where the male said "I love you" first is different from $\frac{1}{2}$.

3. *Agresti Exercise 4.16 (p. 161) Find the form of the deviance residual for an observation in:*

   (a) *a Binomial GLM*

   For the Binomial GLM with logit link, $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, $b(\theta_i) = \log\left(1+\exp(\theta_i)\right) = \log\left(\frac{1}{1-\pi_i}\right)$, and $a(\phi) = \frac{1}{n_i}$ so $w_i = n_i$. In the saturated model, $\tilde{\theta} = \log\left(\frac{y_i}{1-y_i}\right)$. Then

$$\begin{aligned}
d_i &= 2n_i\left[y_i\left(\log\left(\frac{y_i}{1-y_i}\right) - \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)\right) - \log\left(\frac{1}{1-y_i}\right) + \log\left(\frac{1}{1-\hat{\pi}_i}\right)\right] \\
&= 2n_i\left[y_i \log\left(\frac{y_i/(1-y_i)}{\hat{\pi}_i/(1-\hat{\pi}_i)}\right) + \log\left(\frac{1-y_i}{1-\hat{\pi}_i}\right)\right] \\
&= 2\log\left(\frac{y_i^{n_i y_i}(1-y_i)^{n_i(1-y_i)}}{\hat{\pi}_i^{n_i y_i}(1-\hat{\pi})^{n_i(1-y_i)}}\right)
\end{aligned}$$

   so the deviance residuals are

$$\sqrt{2\log\left(\frac{y_i^{n_i y_i}(1-y_i)^{n_i(1-y_i)}}{\hat{\pi}_i^{n_i y_i}(1-\hat{\pi})^{n_i(1-y_i)}}\right)} \times \operatorname{sign}\left(y_i - \hat{\pi}_i\right).$$

   (b) *a Poisson GLM*

   For the Poisson GLM, from 1(b), $d_i = 2\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]$ so the deviance residuals are

$$\sqrt{2\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]} \times \operatorname{sign}\left(y_i - \hat{\mu}_i\right).$$

4. *Consider again the Framingham heart study discussed in Lecture 7. A description of the data set can be found here:* `http://www.ics.uci.edu/~staceyah/111-202/data/framingham.html`*. Read the data into R using the following command:*

```
fram <- read.table('http://www.math.montana.edu/shancock/courses/stat539/data/Framingham.txt',
                   header = TRUE)
```

(a) *Write an R function to run the iterated weighted least squares algorithm for a generalized linear model that uses the logit link. (You may just modify the IWLS R code from class.) Use this function to calculate the maximum likelihood estimates of the coefficients for a logistic regression model with response* `chdfate` *and predictors* `sex` *and age. (Be sure to first re-code sex as a factor.) Turn in a well-commented .R script file with your IWLS R function and the R commands used to run the IWLS algorithm on the Framingham data to the "Homework 3 R Code" Assignment submission folder in D2L.*

The table below shows the estimated coefficients that my IWLS function finds in 5 iterations using initial values of $\beta_0^{(0)} = \beta_1^{(0)} = \beta_2^{(0)} = 0$ and a convergence criterion of $\left|\beta_j^{(t+1)} - \beta_j^{(t)}\right| < 0.00001$.

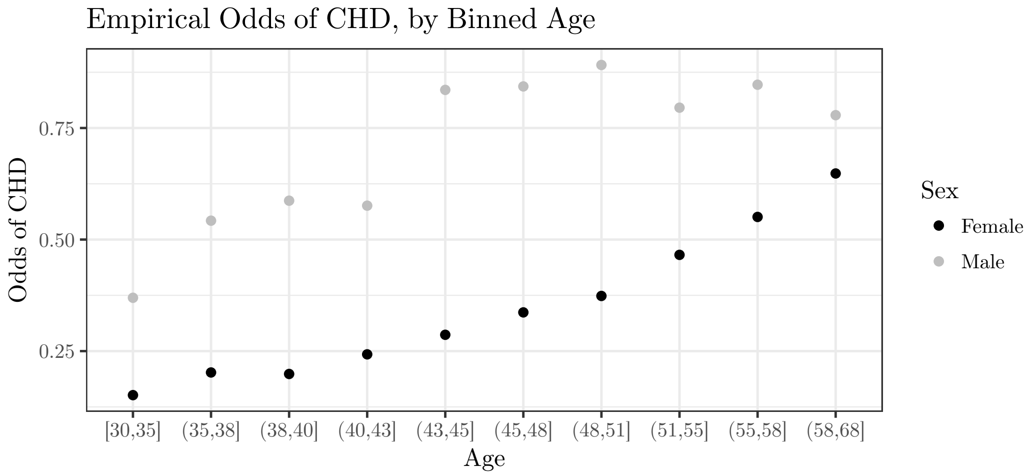
	Estimate
(Intercept)	-2.220076
female	-0.760819
age	0.039550

(b) *For each of the following questions,*

- *Create at least one well labeled, informative plot that helps illuminate the question of interest.*
- *Clearly state the reduced model and full model for a model comparison test that addresses the question.*
- *Use the* `anova` *function in R to carry out the appropriate likelihood ratio test. Write a conclusion of the test that addresses the question.*
- *Calculate and interpret an approximate 95% confidence interval(s) that addresses the question.*

i. *How do the odds of coronary heart disease differ between men and women, adjusting for age?*

I first binned the ages with cutpoints at every 10th percentile so that the bins have similar sizes. The plot on the next page shows the observed odds of CHD for males and females in each age bin.



The appropriate reduced model is

$$\begin{gathered} Y_i \sim \text{Binomial}(1, \mu_i); \\ \text{logit}(\mu_i) = \beta_0 + \beta_1 \text{age}_i \end{gathered}$$

and the full model is

$$\begin{gathered} Y_i \sim \text{Binomial}(1, \mu_i); \\ \text{logit}(\mu_i) = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{female}_i. \end{gathered}$$

```
prob_i_reduced <- glm(chdfate ~ age, data = fram, family = binomial)
prob_i_full <- glm(chdfate ~ age + female, data = fram, family = binomial)
anova(prob_i_reduced, prob_i_full, test = 'LRT')
```

```
# Analysis of Deviance Table
#
# Model 1: chdfate ~ age
# Model 2: chdfate ~ age + female
#   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
# 1      4697     5743.5
# 2      4696     5603.4  1   140.07 < 2.2e-16
```

With an LRT statistic of $\chi^2_1 = 140.07$ and a p-value < 0.0001, there is very strong evidence that the odds of CHD differ between males and females after accounting for age.

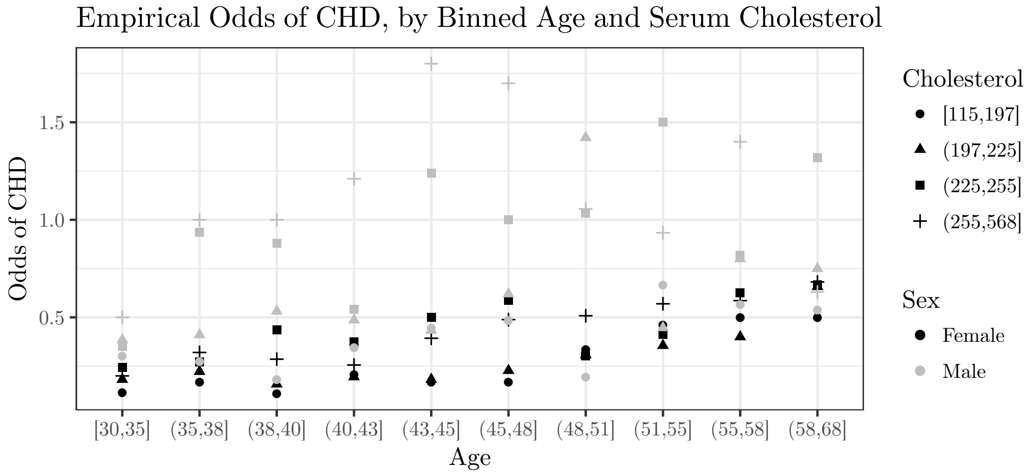
```
summary(prob_i_full)$coefficients
```

```
#                Estimate  Std. Error   z value     Pr(>|z|)
# (Intercept) -2.22007620 0.182101845 -12.19140 3.454409e-34
# age          0.03955003 0.003807717  10.38681 2.847218e-25
# female      -0.76081942 0.064771499 -11.74621 7.386061e-32
```

We are 95% confident that the odds of CHD for a female are between $1-\exp(-0.761+1.96 \times 0.0648) = 46.9\%$ and $1-\exp(-0.761-1.96 \times 0.0648) = 58.843\%$ lower than the odds of CHD for males of the same age.

ii. *Controlling for age and sex, what effect does serum cholesterol level have on the odds of coronary heart disease?*

I removed 33 individuals with missing serum cholesterol values from the dataset and binned cholesterol level into quartiles. The plot below shows the empirical odds by age and serum cholesterol bin.



The reduced model is

$$Y_i \sim \text{Binomial}(1, \mu_i);$$
$$\text{logit}(\mu_i) = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{female}$$

and the full model is

$$Y_i \sim \text{Binomial}(1, \mu_i);$$
$$\text{logit}(\mu_i) = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{female}_i + \beta_3 \text{scl}.$$

```
prob_ii_reduced <- glm(chdfate ~ age + female, data = fram,
                       subset = !is.na(scl), family = binomial)
prob_ii_full <- glm(chdfate ~ age + female + scl, data = fram,
                    subset = !is.na(scl), family = binomial)
anova(prob_ii_reduced, prob_ii_full, test = 'LRT')

# Analysis of Deviance Table
#
# Model 1: chdfate ~ age + female
# Model 2: chdfate ~ age + female + scl
#   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
# 1      4663     5570.5
# 2      4662     5466.7  1   103.82 < 2.2e-16
```

With an LRT statistic of $\chi^2_1 = 103.82$ and a p-value < 0.0001, there is very strong evidence that serum cholesterol level is associated with the odds of CHD after accounting for age and sex.

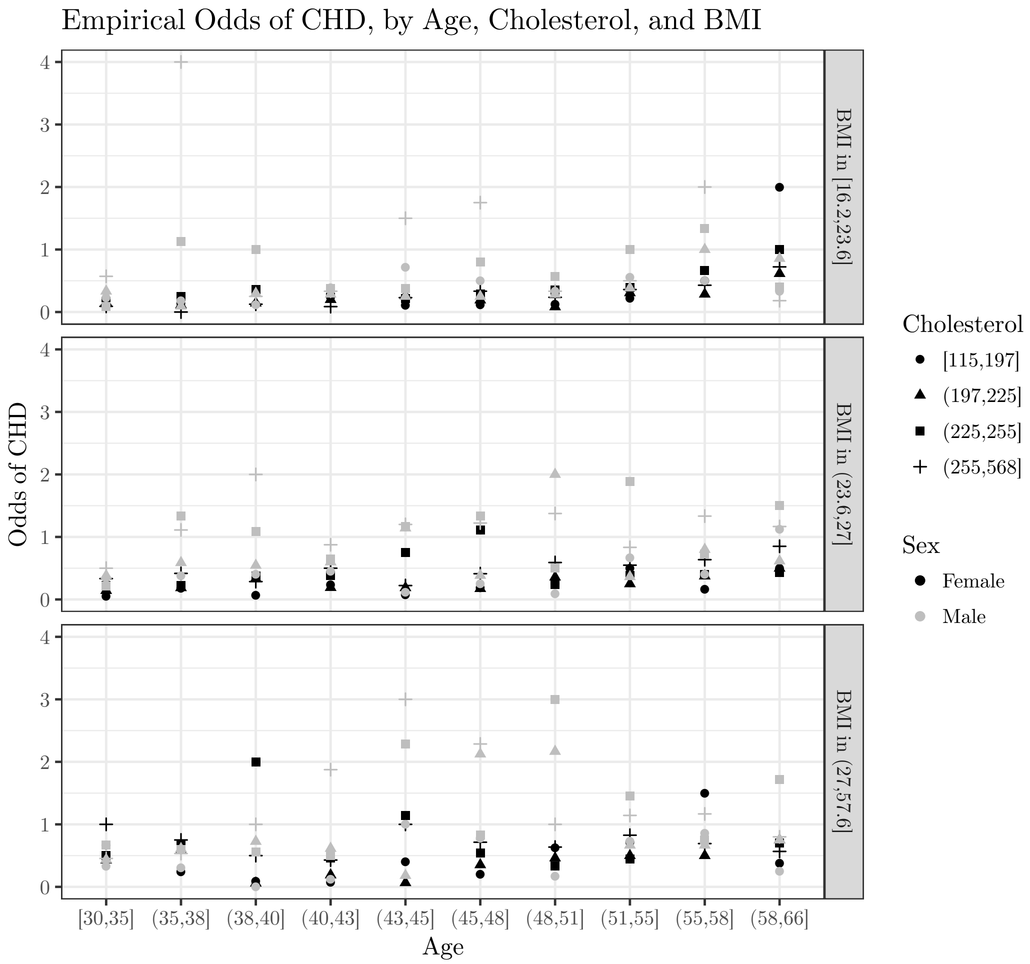
```
summary(prob_ii_full)$coefficients

#                  Estimate   Std. Error    z value     Pr(>|z|)
# (Intercept) -3.542214349 0.2290175452 -15.467000 5.794532e-54
# age          0.030282707 0.0039716085   7.624796 2.444197e-14
# female      -0.798480336 0.0660631446 -12.086623 1.242910e-33
# scl          0.007663466 0.0007639518  10.031347 1.109921e-23
```

We are 95% confident that the odds of CHD for individuals with a given serum cholesterol level are between $\exp(0.00766 - 1.96 \times 0.000764) - 1 = 0.619\%$ and $\exp(0.00766 + 1.96 \times 0.000764) - 1 = 0.920\%$ higher that the odds of CHD for those of the same age and sex with serum cholesterol one unit lower.

iii. *Controlling for age and sex, does the effect of serum cholesterol level on the odds of coronary heart disease differ among different levels of body mass index? If so, how?*

I omitted the 41 individuals who were missing cholesterol or BMI values, and binned BMI into thirds. The plot shows the empirical odds of CHD by age, paneled vertically by BMI.



The reduced model is

$$\begin{aligned} Y_i &\sim \text{Binomial}(1, \mu_i); \\ \text{logit}(\mu_i) &= \beta_0 + \beta_1\text{age}_i + \beta_2\text{female}_i + \beta_3\text{scl}_i + \beta_4\text{bmi}_i \end{aligned}$$

and the full model is

$$\begin{aligned} Y_i &\sim \text{Binomial}(1, \mu_i); \\ \text{logit}(\mu_i) &= \beta_0 + \beta_1\text{age}_i + \beta_2\text{female}_i + \beta_3\text{scl}_i + \beta_4\text{bmi}_i + \beta_5\text{scl}_i \times \text{bmi}_i. \end{aligned}$$

```
prob_iii_reduced <- glm(chdfate ~ age + female + scl + bmi, data = fram,
                       subset = !is.na(scl) & !is.na(bmi), family = binomial)
prob_iii_full <- glm(chdfate ~ age + female + scl * bmi, data = fram,
                    subset = !is.na(scl) & !is.na(bmi), family = binomial)
anova(prob_iii_reduced, prob_iii_full, test = 'LRT')

# Analysis of Deviance Table
#
# Model 1: chdfate ~ age + female + scl + bmi
# Model 2: chdfate ~ age + female + scl * bmi
#   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
# 1      4653     5402.5
# 2      4652     5402.4  1 0.028592   0.8657
```

With an LRT statistic of $\chi^2_1 = 0.03$ and a p-value 0.8657, there is no evidence that the association between serum cholesterol level and the odds of CHD depend on BMI after accounting for age and sex.

```
summary(prob_iii_full)$coefficients

#                    Estimate  Std. Error     z value     Pr(>|z|)
# (Intercept) -5.07082277311 1.147980571  -4.4171678 1.000026e-05
# age          0.02669261546 0.004033900   6.6170748 3.663763e-11
# female      -0.77496250072 0.066564139 -11.6423425 2.510249e-31
# scl          0.00802631572 0.004935549   1.6262254 1.039017e-01
# bmi          0.06908031308 0.043773893   1.5781167 1.145388e-01
# scl:bmi     -0.00003153701 0.000186399  -0.1691908 8.656465e-01
```

We are 95% confident that, among people of a given age, sex, and BMI, the percent difference in the odds of CHD between individuals with a given serum cholesterol level and individuals with one unit less of serum cholesterol is between $1 - \exp(-0.0000315 - 1.96 \times 0.00019) = 0.0397\%$ lower and $\exp(-0.0000315 + 1.96 \times 0.000186) - 1 = 0.0334\%$ higher than the percent difference in the odds of CHD between individuals with a given serum cholesterol level and individuals with one unit less of serum cholesterol for people of the same age and sex, with one unit lower BMI.

Appendix: Code for the Plots in 4(b)

```
library(dplyr)
library(ggplot2)
theme_set(theme_bw())

# Plot for 4(b)i.
ggplot(fram %>%
         mutate(
           Age = cut(age, quantile(age, seq(0, 1, 0.1)), include.lowest = TRUE),
           Sex = ifelse(sex == 1, 'Male', ifelse(sex == 2, 'Female', NA))
         ) %>%
         group_by(Age, Sex) %>%
         summarise(`Odds of CHD` = mean(chdfate) / (1 - mean(chdfate))) %>%
         ungroup,
       aes(y = `Odds of CHD`, x = Age, col = Sex)) +
  geom_point() +
  scale_color_manual(values = c('black', 'grey')) +
  ggtitle('Empirical Odds of CHD, by Binned Age')

# Plot for 4(b)ii.
ggplot(fram %>% filter(!is.na(scl)) %>%
         mutate(
           Age = cut(age, quantile(age, seq(0, 1, 0.1)), include.lowest = TRUE),
           Cholesterol = cut(scl, quantile(scl, seq(0, 1, 0.25)), include.lowest = TRUE),
           Sex = ifelse(sex == 1, 'Male', ifelse(sex == 2, 'Female', NA))
         ) %>%
         group_by(Age, Sex, Cholesterol) %>%
         summarise(`Odds of CHD` = mean(chdfate) / (1 - mean(chdfate))) %>%
         ungroup,
       aes(y = `Odds of CHD`, x = Age, col = Sex, shape = Cholesterol)) +
  geom_point() +
  scale_color_manual(values = c('black', 'grey')) +
  ggtitle('Empirical Odds of CHD, by Binned Age and Serum Cholesterol')

# Plot for 4(b)iii.
ggplot(fram %>% filter(!is.na(scl), !is.na(bmi)) %>%
         mutate(
           Age = cut(age, quantile(age, seq(0, 1, 0.1)), include.lowest = TRUE),
           Cholesterol = cut(scl, quantile(scl, seq(0, 1, 0.25)), include.lowest = TRUE),
           BMI = paste('BMI in', cut(bmi, quantile(bmi, seq(0, 1, 1/3)), include.lowest = TRUE)),
           Sex = ifelse(sex == 1, 'Male', ifelse(sex == 2, 'Female', NA))
         ) %>%
         group_by(Age, Sex, Cholesterol, BMI) %>%
         summarise(`Odds of CHD` = mean(chdfate) / (1 - mean(chdfate))) %>%
         ungroup,
       aes(y = `Odds of CHD`, x = Age, col = Sex, shape = Cholesterol)) +
  geom_point() +
  facet_grid(BMI ~ .) +
  scale_color_manual(values = c('black', 'grey')) +
  ggtitle('Empirical Odds of CHD, by Age, Cholesterol, and BMI')
```