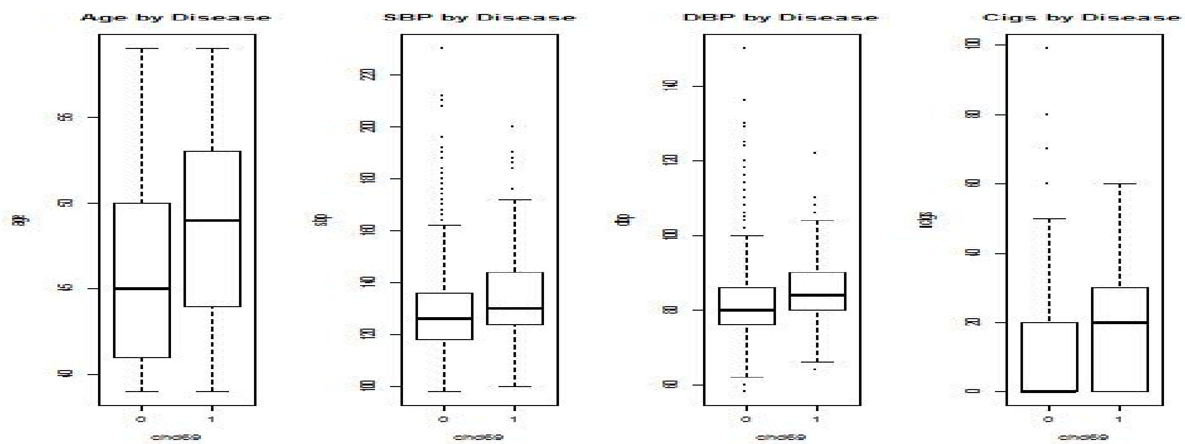
Stat 6021: Guided Question Set 9 Solutions

1. People who developed heart disease tend to be older, have higher SBP and DBP, as well as smoke more cigarettes. There is high variability in a lot of these predictors for each group (those without heart disease and those with heart disease). The number of cigarettes smoked appears to be the biggest factor in whether one develops heart disease as their distributions are most different. Among those with no heart disease, 50% of them did not smoke. Among those with heart disease, 25% of them did not smoke.



2.
```
> result<-glm(chd69 ~ age + sbp + dbp + ncigs, family="binomial")
> summary(result)

Call:
glm(formula = chd69 ~ age + sbp + dbp + ncigs, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2006  -0.4427  -0.3491  -0.2733   2.7644

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.119553   0.749692 -12.164  < 2e-16 ***
age          0.066602   0.011733   5.677 1.37e-08 ***
```

```
sbp          0.019464   0.006133   3.173  0.00151 **
dbp          0.007867   0.010057   0.782  0.43406
ncigs        0.024193   0.004136   5.850 4.92e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1781.2  on 3153  degrees of freedom
Residual deviance: 1670.0  on 3149  degrees of freedom
```

The estimated logistic regression equation is

$$\log \frac{\hat{\pi}}{1-\hat{\pi}} = -9.1196 + 0.0666 age + 0.0195 sbp + 0.0079 dbp + 0.0242 ncigs.$$

3. Coefficient for ncigs is 0.0242. A few interpretations:

   - For an additional cigarette smoked per day (on average), the estimated log odds of developing coronory heart disease increases by 0.0242, while holding age, sbp, and dbp constant.
   - For an additional cigarette smoked per day (on average), the estimated odds ratio of developing coronory heart disease increases by exp(0.0242) = 1.0245, while holding age, sbp, and dbp constant.
   - For an additional cigarette smoked per day (on average), the estimated odds of developing coronory heart disease gets multiplied by a factor of exp(0.0242) = 1.0245, while holding age, sbp, and dbp constant.

4. The estimated odds is 0.0324. The estimated probability is 0.0313.

$$\begin{aligned}
\log \frac{\hat{\pi}}{1-\hat{\pi}} &= -9.1196 + 0.0666 age + 0.0195 sbp + 0.0079 dbp + 0.0242 ncigs \\
&= -9.1196 + 0.0666(45) + 0.0195(110) + 0.0079(70) + 0.0242(0) \\
&= -3.430695
\end{aligned}$$

Exponentiating both sides, we get the estimated odds

$$\frac{\hat{\pi}}{1-\hat{\pi}} = \exp(-3.430695) = 0.0323644.$$

So the estimated probability is

$$\hat{\pi} = \frac{0.0323644}{1 + 0.0323644} = 0.03134982.$$

Notice that with a rare event (having coronary heart disease among middle-aged men), the odds and probability are approximately the same.

Using R,

```
> ##make prediction for log odds
> newdata<-data.frame(age=45, sbp=110, dbp=70, ncigs=0)
> predict(result,newdata)
       1
-3.430695
> ##note predict gives the log odds, need to exponentiate to get odds
> odds<-exp(predict(result,newdata))
> odds
         1
0.03236444
> predict(result,newdata, type="response")
         1
0.03134982
> ##need to add type="response" for probability instead of log odds
```

5.

$$
\begin{aligned}
H_0 &: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \\
H_a &: \text{ at least one of the coefficients in the null is not zero.}
\end{aligned}
$$

The test statistic is $\Delta G^2 = \text{ null deviance - residual deviance } = 111.2295$. The p-value is 0. Our data supports the claim that our logistic regression model is useful in estimating the log odds of developing coronoray heart disease.

```
> deltaG2<-result$null.deviance-result$deviance
> deltaG2
[1] 111.2295
> 1-pchisq(deltaG2,4)
[1] 0
```

6. Diastolic blood pressure is not a significant predictor of heart disease, when the other three predictors are already in the model, since its p-value is large.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.119553   0.749692 -12.164  < 2e-16 ***
age          0.066602   0.011733   5.677 1.37e-08 ***
sbp          0.019464   0.006133   3.173  0.00151 **
dbp          0.007867   0.010057   0.782  0.43406
ncigs        0.024193   0.004136   5.850 4.92e-09 ***
```

7.

$$
\begin{aligned}
H_0 &: \beta_2 = \beta_3 = 0 \\
H_a &: \text{ at least one of the coefficients in the null is not zero.}
\end{aligned}
$$

The test statistic is $\Delta G^2 = 35.8458$. The p-value is 0. Our data supports going with the more complicated model with 4 predictors.

```
> reduced<-glm(chd69 ~ age + ncigs, family="binomial")
> deltaG2_partial<-reduced$deviance-result$deviance
> deltaG2_partial
[1] 35.84578
> 1-pchisq(deltaG2_partial,2)
[1] 1.645085e-08
```

8. I would go with age, sbp, and ncigs as the predictors, dropping dbp from the model. The answer from question 7 supports a model with all 4 predictors over a model minus the two blood pressure predictors, and the answer from question 6 supports dropping one of the blood pressure predictors to result in a model with 3 predictors: age, sbp, ncigs.