Assignment

null

1.

(a) Combining the information from the two tables results in the following table

	D	not D
E	45	40
not E	20	40

The unadjusted estimated odds ratio is $\widehat{\text{OR}} = \frac{45\times 40}{40\times 20} = 2.25$. The corresponding estimated variance for $\log \widehat{\text{OR}}$ is $\sqrt{\widehat{\text{var}}(\log \widehat{\text{OR}})} = \frac{1}{45} + \frac{1}{40} + \frac{1}{20} + \frac{1}{40} = 0.1222$. Finally, the 95% confidence interval for the OR is

$$\left(e^{\log 2.25 - 1.96\times\sqrt{0.1222}}, e^{\log 2.25 + 1.96\times\sqrt{0.1222}}\right) = (1.133, 4.468).$$

The unadjusted estimated odds ratio is 2.25 and the 95% CI is $(1.133, 4.468)$, thus indicating a positive association between consumption of substance E and bladder cancer.

(b) The estimated OR for smokers is $\widehat{\text{OR}}_{\text{smokers}} = \frac{35\times 10}{20\times 5} = 3.5$. The corresponding estimated variance for $\log \widehat{\text{OR}}_{\text{smokers}} = \frac{1}{35} + \frac{1}{20} + \frac{1}{5} + \frac{1}{10} = 0.3785714$, which leads to the following 95% confidence interval

$$\left(e^{\log 3.5 - 1.96\times\sqrt{0.3785714}}, e^{\log 3.5 + 1.96\times\sqrt{0.3785714}}\right) = (1.049, 11.683).$$

In turn, the estimated OR for non-smokers is $\widehat{\text{OR}}_{\text{smokers}} = \frac{10\times 30}{20\times 15} = 1$. The corresponding estimated variance for $\log \widehat{\text{OR}}_{\text{non-smokers}} = \frac{1}{10} + \frac{1}{20} + \frac{1}{15} + \frac{1}{30} = 0.25$, which leads to the following 95% confidence interval

$$\left(e^{\log 1 - 1.96\times\sqrt{0.25}}, e^{\log 1 + 1.96\times\sqrt{0.25}}\right) = (0.375, 2.664).$$

For smokers, there is evidence of positive association between consumption of substance E and bladder cancer. This was the case also for the unadjusted estimate, but here the association is 'stronger'. For non-smokers, there is no evidence of an association. The two odds ratio, for smokers and non-smokers, are quite different and there is little overlap in the corresponding 95% CIs, thus indicating that smoking is an effect modifier of the $E - D$ association.

2. A possible explanation for the discrepancy mentioned is that a significant number of individuals with high stress levels already had high blood pressure at the beginning of the study (and so they would not be included in the incidence calculation) and they continued suffering from high blood pressure troughout the duration of the study. Another possible explanation is if stress levels affect the way high blood pressure responds to treatment, i.e., if high blood pressure responds better to treatment in subjects who have low stress levels than in subjects who have high stress levels. This would imply that a larger number of cases of high blood pressure, that were developed during the study, were 'cured' (return to normal levels) amongst low stress levels individuals than in high stressed subjects.

Bonus question. (a) Here X is modelled through a single numeric variable and the result logistic regression model is given by

$$\log\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 x, \quad p_x = \Pr(\text{CHD} \mid X = x), \quad x \in \{0,1,2,3,4,5,6,7\}.$$

Here β_0 is the log odds of CHD for those with $X = 0$ (so SBP < 117 mm Hg). Also, β_1 is the log odds ratio for a unit increase in X, i.e., the log odds ratio comparing those with $X = 1$ to $X = 0$, those with $X = 2$ to $X = 1$, those with $X = 3$ to $X = 2$, and so on and so forth.

(b) By modelling X as a continuous variable, there is a linear trend imposed. This model assumes that the log odds ratios of CHD comparing $X = 1$ to $X = 0$, $X = 2$ to $X = 1$, $X = 3$ to $X = 2$, etc, are the same (and given by β_1). This is by opposition to a model that would use indicator variables (seven in this case) to model the effect of the different SBP categories, which imposes no structural relationships and which allows the log odds ratio of CHD to be distinct for the different SBP categories.

(c) Below we fit the model specified in (a).

```r
x <- c(0:7)
bp.y <- c(3, 17, 12, 16, 12, 8, 16, 8)
bp.n <- c(153, 235, 272, 255, 127, 77, 83, 35)
data_ex2 <- data.frame("x" = x, "yes" = bp.y, "no" = bp.n)
res_ex2 <- glm(cbind(yes, no) ~ x,
               data = data_ex2, family = "binomial")
summary(res_ex2)
```

```
##
## Call:
## glm(formula = cbind(yes, no) ~ x, family = "binomial", data = data_ex2)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.42264    0.22065 -15.512  < 2e-16 ***
## x            0.26943    0.05504   4.895 9.81e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30.0226  on 7  degrees of freedom
## Residual deviance:  6.3948  on 6  degrees of freedom
## AIC: 43.096
##
## Number of Fisher Scoring iterations: 4
```

```r
exp(res_ex2$coefficients)
```

```
## (Intercept)           x
##  0.03262632  1.30921939
```

```r
exp(confint.default(res_ex2))
```

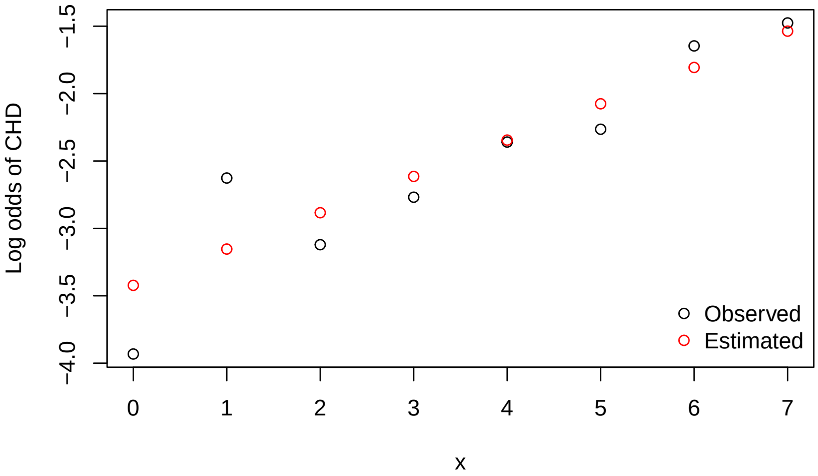
```
##                  2.5 %     97.5 %
## (Intercept) 0.02117161 0.05027848
## x           1.17534480 1.45834262
```

From the results of the fitted model to the data, it is evident that there is a significant effect of SBP on CHD. The estimated odds ratio for a unit increase in X is 1.309 $(1.175, 1.478)$.

(d) The required odds ratio is given by $e^{3\beta-\beta} = e^{2\beta}$ and its estimate is thus $e^{2\times0.26943} = 1.714$.

(e) Below we provide the code to construct such a plot.

```r
plot(x, log(bp.y/bp.n),
     xlab = expression(x),
     ylab = "Log odds of CHD")
points(x, res_ex2$coefficients[1] + res_ex2$coefficients[2]*x,
       col = "red")
legend("bottomright", col = c("black", "red"), pch = c(1, 1),
       legend = c("Observed", "Estimated"), bty = "n")
```



Overall, apart from the first two SBP categories, the fit is quite reasonable, as can be observed from the above plot.