Chap 1. 19 (a)
The sample consists of 16,402 people who are 18-30 years old in the Rayong and Chon Buri provinces in Thailand. Note that the sample given in the contingency table has 7 people removed because they had HIV-1 infection at the time the study began. The population is all 18-30 year old healthy people in the Rayong and Chon Buri provinces in Thailand. While this is the population for this study, the hope is the results for this population would extend to the entire world. Notice that this trial is not focused on people with high risk of HIV infection. This is what the âĂIJcommunity-basedâĂİ part of the trial means.

Chap 1. 19 (b)
The estimated odds ratio is 0.69, and the corresponding 95% confidence interval is 0.4805 to 0.9832. The score test p-value is 0.0390 for a two-sided test. The estimated relative risk is 0.69, and the corresponding 95% confidence interval is 0.4831 to 0.9834. The upper bounds on the two confidence intervals are less than 1 (although quite close to 1) and the p-value is less than Îś = 0.05 (although quite close to 0.05). Due to these measures being close to their boundaries of being declared "non-significant," we are hesitant to make an absolute judgment that indeed the treatment type affects HIV status. Rather, we would prefer to say that there is marginal evidence the treatment type affects HIV status. One could also perform a one-sided test or calculate a one-sided confidence interval. For example, the one-sided test leads to a p-value of 0.0195 where the alternative hypothesis is that the vaccine reduces the probability of HIV infection. However, in a different clinical trial (HVTN 502) a few years prior to this one, the reverse effect than what was intended (vaccine led to a higher HIV prevalence) occurred, so two-sided tests and confidence intervals may be better to examine.

Chap 1. 19 (c)
Omitted.

Chap 1. 20 (a)
Intent-to-treat: The population is the same as in the previous problem, but with the added assumption that they were all HIV-negative prior to the study. Per-protocol: The population is the same as in the previous problem, but includes only those individuals who complete all treatments.

Chap 1. 20 (b)
Odds ratios and score tests: The p-values are now greater than $\alpha = 0.05$, and

	95 % confidence interval for the odds ratio	Two-sided score test p-value
Intent-to-treat	(0.52, 1.04)	0.0803
Per-protocol	(0.48, 1.14)	0.1694

the intervals contain 1. Using a strict $\alpha = 0.05$ level, one would conclude there is not sufficient evidence that the vaccine works. However, with respect to the intent-to-treat data, our "marginal evidence" conclusion from the previous problem would not change.

Chap 1. 21 (a)
We conjecture that the results would have been less publicized. It may be more interesting to read about a success than something which is somewhat inconclusive.

Chap 1. 21 (b)
Yes, if strict α -levels were used, the conclusions from the clinical trials could change. For example, using $\alpha = 0.01$ would lead to saying "there is not sufficient evidence to conclude the vaccine is effective."

Chap 1. 21 (c)
One can only speculate here. A potential reason is that the media prefers to highlight "significant" results. The intent-to-treat analysis should have received the same level of coverage as the modified intent-to-treat analysis. We would not be surprised if many people only saw the original headlines for the modified intent-to-treat analysis.

Chap 1. 21 (d)
Our previous report would work here as well. The key is to not use a strict $\alpha = 0.05$ level when developing conclusions.

```
> options(continue=" ", prompt="   ", digit = 4)
```

4 (a) Logistic regression assumes that each response has a binomial distribution, and independence of trials is required for the binomial (see Chapter 1). Because three O-rings are on each rocket, there may be dependencies (e.g., installed by the same workers, failure in one perhaps could lead to failure in another, ...) in their success or failure.

```
  data = read.csv("~/Documents/stat475/challenger.csv")
  fit <- glm(O.ring / Number ~ Temp + Pressure,
family = binomial, data = data, weights = Number)
  c = fit$coefficients
  c = round(c, 4)
```

4 (b)

$$\log \frac{\pi}{1-\pi} = 2.5202 + -0.0983\text{Temp} + 0.0085\text{Pressure}$$

```
  fit1 <- glm(O.ring / Number ~ Pressure,
family = binomial, data = data, weights = Number)
  fit2 <- glm(O.ring / Number ~ Temp,
family = binomial, data = data, weights = Number)
  anova(fit1, fit, test="Chisq")

Analysis of Deviance Table
```

```
Model 1: O.ring/Number ~ Pressure
Model 2: O.ring/Number ~ Temp + Pressure
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        21     21.730
2        20     16.546  1   5.1838   0.0228 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

```
   anova(fit2,fit, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: O.ring/Number ~ Temp
Model 2: O.ring/Number ~ Temp + Pressure
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        21     18.086
2        20     16.546  1   1.5407   0.2145
```

4 (c) p-value for Temp is 0.023; p-value for Pressure is 0.21;

4 (d) Because the p-value corresponding to Pressure is large, there is not sufficient evidence to indicate that the variable is important given that Temp is in the model. A potential problem is that the explanatory variable still may be important with respect to an interaction term or a transformation. Chapter 5 discusses other potential problems with this model selection approach.

```
c = fit2$coefficients
c = round(c, 4)
```

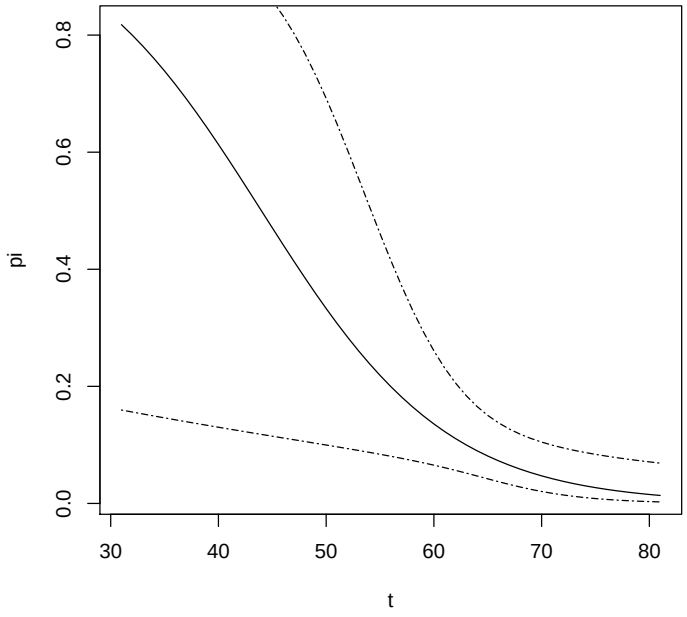
5 (a)

$$\log \frac{\pi}{1-\pi} = 5.085 + -0.1156 \text{ Temp}$$

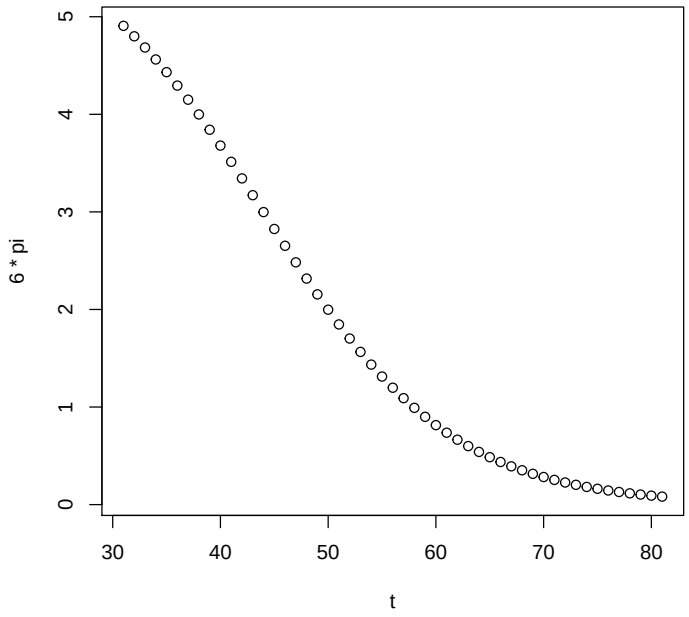
5 (b)

```
t = seq(31, 81, by = 1)
y =   c[1] +  c[2] *t
pi = exp(y)/ (1 + exp(y))
plot(t,pi, type='l')
var = vcov(fit2)
b = fit2$coefficients
b = c(b)
X = cbind(1, t)
var_z = diag(X%*% var %*% t(X))
hat_z = X %*% b
CI_upper = hat_z + 1.96 * var_z ** (1/2)
```

```
CI_lower = hat_z - 1.96 * var_z ** (1/2)
ci_lower = exp(CI_lower)/(1 + exp(CI_lower))
ci_upper = exp(CI_upper) / ( 1 + exp(CI_upper))
lines(t, ci_lower, lty=18)
lines(t, ci_upper, lty=18)
```




```
t = seq(31, 81, by = 1)
y =   c[1] +  c[2] *t
pi = exp(y)/ (1 + exp(y))
plot(t,6 * pi)
```



5 (c) There are fewer observations (or none at all!) for the lower temperatures. Thus, there is more uncertainty about the estimates which is reflected by having wider intervals.

5 (d) The estimated probability at 31° is 0.8178, and the CI is $[0.1596, 0.9907]$.

```
  fit3 <- glm(O.ring / Number ~ Pressure+ Temp + I(Temp^2),
family = binomial, data = data, weights = Number)
  anova(fit3, fit, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: O.ring/Number ~ Pressure + Temp + I(Temp^2)
Model 2: O.ring/Number ~ Temp + Pressure
  Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
1        19     16.494
2        20     16.546 -1 -0.051661   0.8202
```

```
  fit3 <- glm(O.ring / Number ~ Pressure+ Temp + I(Pressure^2),
family = binomial, data = data, weights = Number)
  anova(fit3, fit, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: O.ring/Number ~ Pressure + Temp + I(Pressure^2)
Model 2: O.ring/Number ~ Temp + Pressure
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        19     15.875
2        20     16.546 -1  -0.6705   0.4129
```

Because the p-value is large, there is not sufficient evidence of a quadratic relationship.

6

```
  fit_prob <- glm(O.ring / Number ~ Pressure+ Temp,
family = binomial(link = probit), data = data, weights = Number)
  fit_llog<- glm(O.ring / Number ~ Pressure+ Temp,
family = binomial(link = cloglog), data = data, weights = Number)
```