STA 138 Discussion 8 – solutions

Fall 2020

glm(·, family = binomial)

For this discussion we will explore logistic regression models using `wine.csv`, containing data regarding the quality of wines.

He have here three variables, "quality," "SO2," and "pH," recorded for each of 1599 wine samples:

- quality (binary): 1 if high quality, 0 otherwise
- SO2 (binary): 1 if high sulfur dioxide levels, 0 o.w.
- pH (numeric): pH of the wine

Let's let

$$x_1 = \begin{cases} 1 & \text{if SO2} = 1 \\ 0 & \text{if SO2} = 0 \end{cases},$$

$$x_2 = \text{pH} ,$$

and

$$x_3 = x_1 \cdot x_2 .$$

1. Consider the model π = P(high quality)

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 . \quad \Rightarrow \pi = \frac{e^{\alpha+\beta_1 x_1}}{1+e^{\alpha+\beta_1 x_1}}$$

x_1 is binary

   (a) What are the estimated parameters for this model?

   (b) Interpret the parameters.

   The estimated parameters here are

$$\hat{\alpha} = -1.6288362$$

   and

$$\hat{\beta}_1 = -1.2094792 .$$

   SO2 here is binary, and low SO2 is the 'baseline' case. The estimated log-odds for low SO2 wine according to this model are $\hat{\alpha}$, which corresponds to estimated odds of 0.1961577, and estimated probability of 0.1639899 of high quality for a low-SO2 wine.

$x_1 = 0$, $\log\left(\frac{\pi}{1-\pi}\right) = \hat{\alpha}$

$\frac{\pi}{1-\pi} = e^{\hat{\alpha}}$ $\quad \hat{\pi} = \frac{e^{\hat{\alpha}}}{1+e^{\hat{\alpha}}}$

   The estimated log-odds ratio of high quality for high SO2 vs. low SO2 wines is $\hat{\beta}_1$. Thus, estimated odds of high quality for high SO2 wines under this model are 0.2983526 times those of low SO2 wines.

$e^{\hat{\beta}_1}$

$\hat{\beta}_1 = \log\left(\frac{P(y=1 \mid x_1=1)/P(y=0 \mid x_1=1)}{P(y=1 \mid x_1=0)/P(y=0 \mid x_1=0)}\right)$

   So, this model suggests that wines with low SO2 are more likely to be high quality.

2. Consider the model

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_2 x_2 .$$

x_2 is continuous

$P(y=1 \mid x_1=1)/P(y=0 \mid x_1=1) = \frac{e^{\alpha+\beta_1}}{1+e^{\alpha+\beta_1}} \Big/ \frac{1}{1+e^{\alpha+\beta_1}} = e^{\alpha+\beta_1}$

$P(y=1 \mid x_1=0)/P(y=0 \mid x_1=0) = \frac{e^{\alpha}}{1+e^{\alpha}} \Big/ \frac{1}{1+e^{\alpha}} = e^{\alpha}$

   (a) What are the estimated parameters for this model?

   (b) Interpret the parameters.

   (c) Plot both the fitted log-odds and fitted probability of high quality for wines as a function of pH.

   The estimated parameters here are

$$\hat{\alpha} = 1.7774224$$

   and

$$\hat{\beta}_2 = -1.0990887 .$$

   pH here is continuous, and pH= 0 is the 'baseline' case. The estimated log-odds for pH= 0 wine according to this model are $\hat{\alpha}$. This is not really interpretable in itself, because pH 0 wines would be really bad for the digestion and generally not advisable to drink, and maybe more importantly they don't exist.

too acidic to drink

The estimated log-odds ratio of high quality for wines with a one-unit difference in pH's is $\hat{\beta}_2$. Thus, the estimated odds of high quality for a wine with pH that is one higher than that of another one under this model are 0.3331746 times those of the other.

So, this model suggests that wines with higher pH's (i.e. wines that are less acidic) are less likely to be high quality than those with lower pH's (i.e. those that are more acidic) (see Figures 1 and 2 below).

3. Consider the model

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 \ .$$

(a) What are the estimated parameters for this model?

(b) Interpret the parameters.

(c) Plot both the fitted log-odds and fitted probability of high quality for wines as a function of pH and SO2.

The estimated parameters here are

$$\hat{\alpha} = 2.6421678 \ ,$$
$$\hat{\beta}_1 = -1.2395354 \ ,$$

and

$$\hat{\beta}_2 = -1.2918775 \ .$$

pH= 0 and low SO2 is the 'baseline' case. The estimated log-odds for such wine according to this model are $\hat{\alpha}$. This is, again, not really interpretable in itself, because pH 0 wines are still not a thing.

All other things equal, the estimated log-odds ratio of high quality for high SO2 vs. low SO2 wines is $\hat{\beta}_1$. Thus, estimated odds of high quality for high SO2 wines under this model are 0.2895187 times those of low SO2 wines, holding pH constant.

All other things equal, the estimated log-odds ratio of high quality for wines with a one-unit difference in pH's is $\hat{\beta}_2$. Thus, the estimated odds of high quality for a wine with pH that is one higher than that of another one under this model are 0.2747544 times those of the other, holding SO2 constant.

So, this model suggests that wines with higher pH's (i.e. wines that are less acidic) are less likely to be high quality than those with lower pH's (i.e. those that are more acidic) (see Figures 3 and 4 below).

4. Consider the model

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ .$$

(a) What are the estimated parameters for this model?

(b) Interpret the parameters.

(c) Plot both the fitted log-odds and fitted probability of high quality for wines as a function of pH and SO2.

The estimated parameters here are

$$\hat{\alpha} = 4.5924092 \ ,$$
$$\hat{\beta}_1 = -15.4897825 \ ,$$
$$\hat{\beta}_2 = -1.8843085 \ ,$$

and

$$\hat{\beta}_3 = 4.3077793 \ .$$

For low SO2 wines, we have fitted model

$$\log\left(\frac{\pi}{1-\pi}\right) = \hat{\alpha} + \hat{\beta}_2 x_2 = 4.5924092 - 1.8843085 x_2 \ ,$$

while for high SO2 wines we have fitted model

$$\log\left(\frac{\pi}{1-\pi}\right) = (\hat{\alpha} + \hat{\beta}_1) + (\hat{\beta}_2 + \hat{\beta}_3) x_2 = -10.8973733 + 2.4234708 x_2 \ .$$

So, under this model, the chances of high quality are increasing in pH for high SO2 wines, while they are decreasing in pH for low SO2 wines (see Figures 5 and 6).

Figure 1: Log odds linear in pH

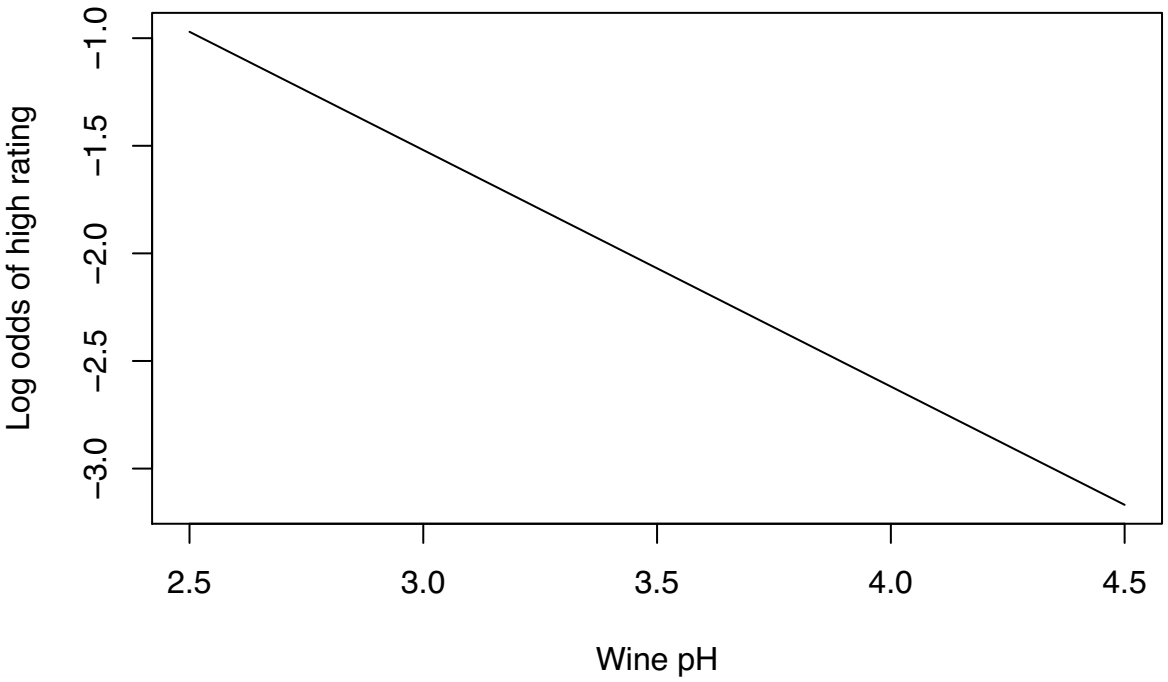


Figure 2: Log odds linear in pH

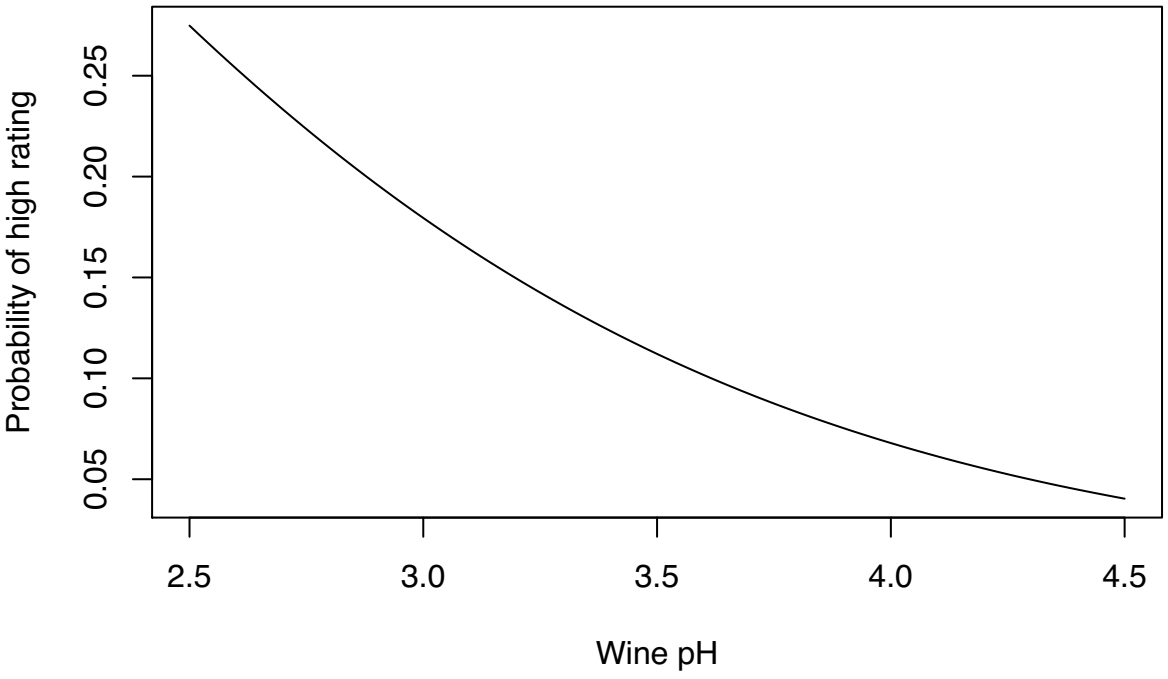


Figure 3: Log odds linear in pH and SO2

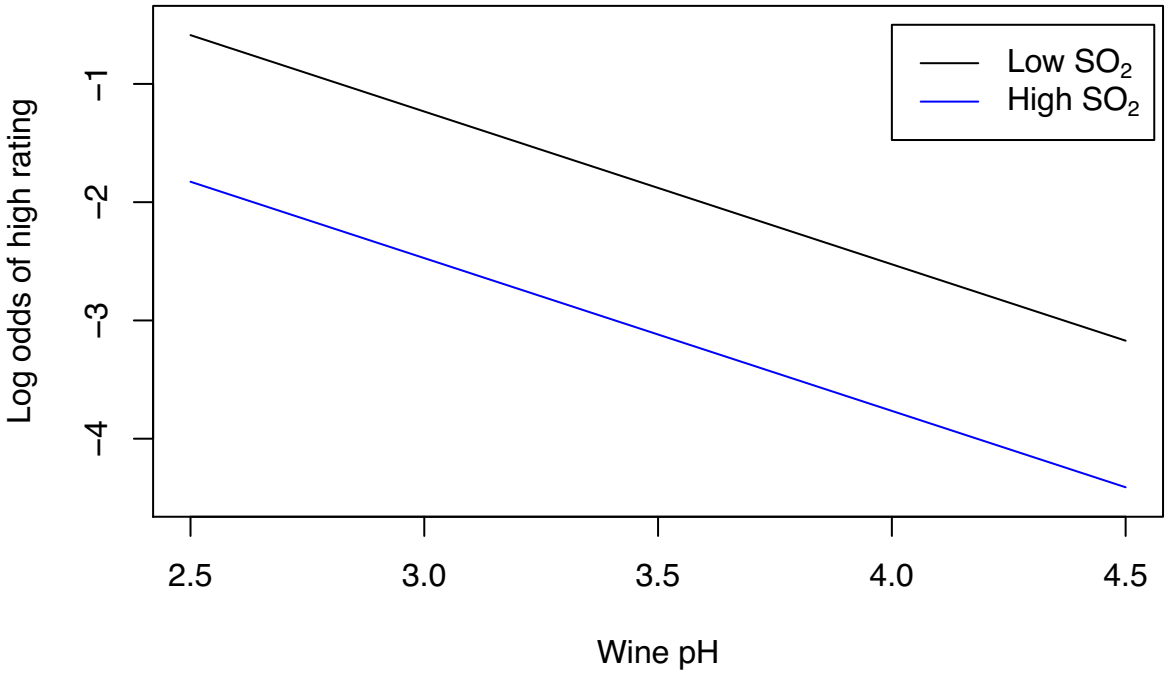


Figure 4: Log odds linear in pH and SO2

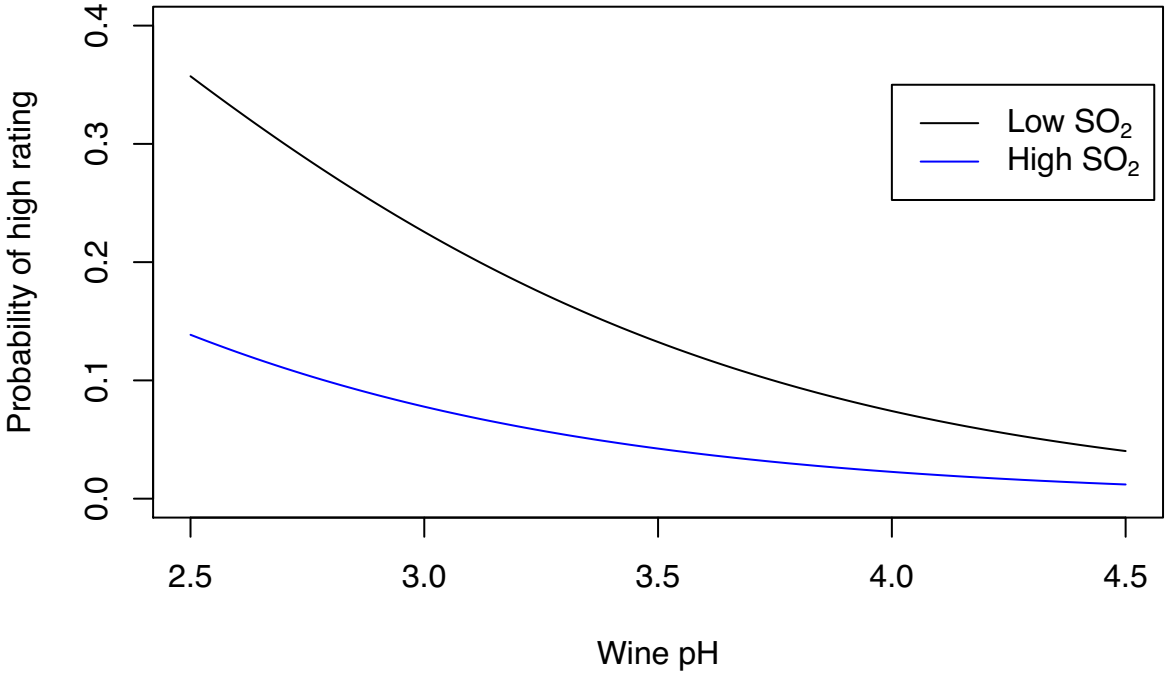


Figure 5: Log odds linear in pH, SO2, and interaction

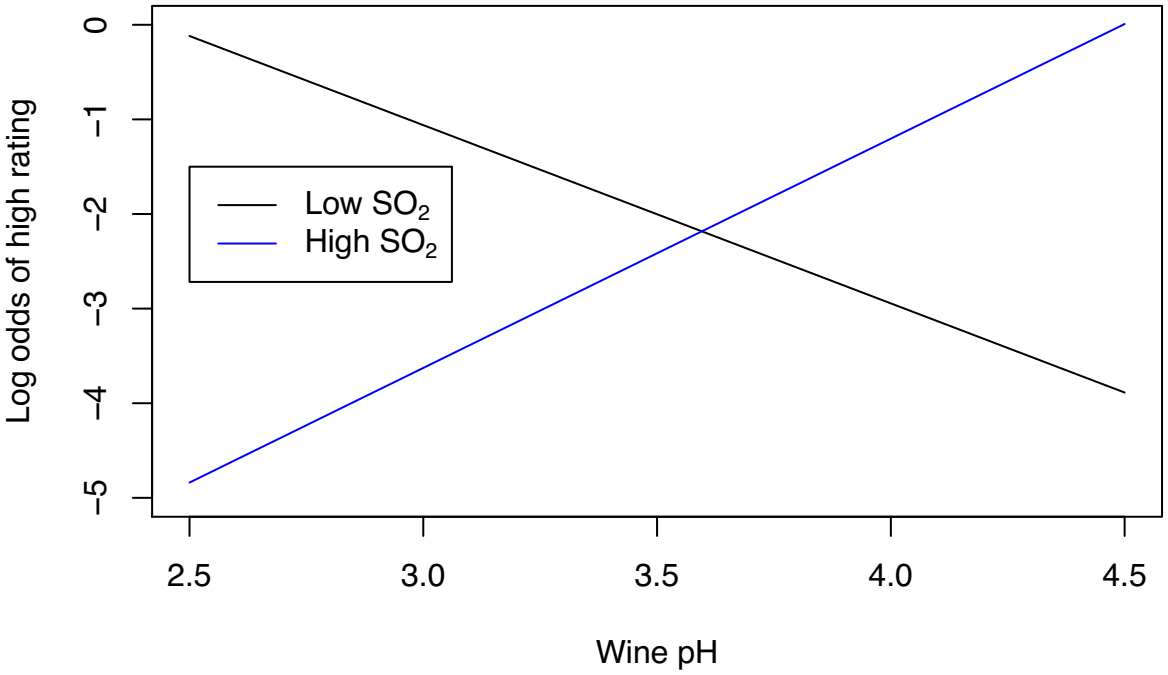
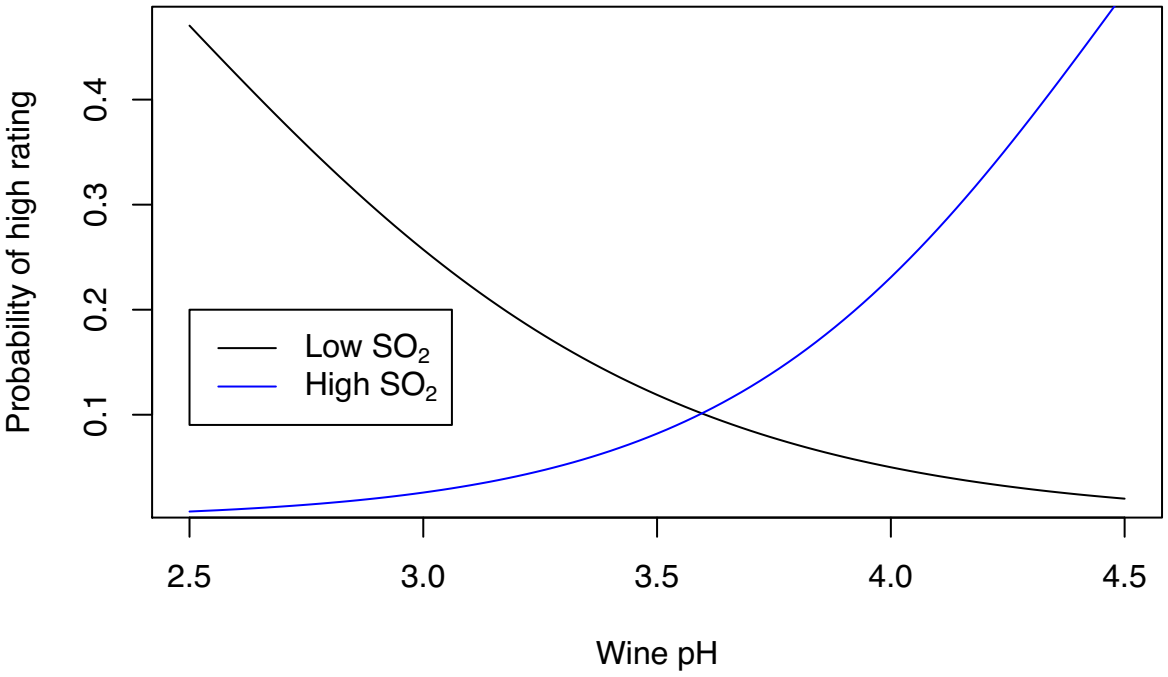


Figure 6: Log odds linear in pH, SO2, and interaction



Appendix: R Script

```
wine <- read.csv("wine.csv")
#### problem 1
fittedModel <- glm(quality~SO2, family=binomial, data=wine)
#### problem 2
fittedModel2 <- glm(quality~pH, family=binomial, data=wine)
logOdds_2 <- function(x){predict(fittedModel2,data.frame(pH=x))}
pi_2 <- function(x){plogis(predict(fittedModel2,data.frame(pH=x)))}
## note:
## In R, the logistic function is evaluated by plogis()
## (CDF of logistic distribution)
## and the logit function is evaluated by qlogis()
## (quantile funciton of logistic distribution)
## Furthermore, predict() here will give fitted log-odds.
#### problem 3
fittedModel3 <- glm(quality~SO2+pH, family=binomial, data=wine)
logOdds0_3 <- function(x){predict(fittedModel3,data.frame(pH=x,SO2=0))}
logOdds1_3 <- function(x){predict(fittedModel3,data.frame(pH=x,SO2=1))}
pi0_3 <- function(x){plogis(predict(fittedModel3,data.frame(pH=x,SO2=0)))}
pi1_3 <- function(x){plogis(predict(fittedModel3,data.frame(pH=x,SO2=1)))}
#### problem 4
fittedModel4 <- glm(quality~SO2+pH+SO2:pH, family=binomial, data=wine)
logOdds0_4 <- function(x){predict(fittedModel4,data.frame(pH=x,SO2=0))}
logOdds1_4 <- function(x){predict(fittedModel4,data.frame(pH=x,SO2=1))}
pi0_4 <- function(x){plogis(predict(fittedModel4,data.frame(pH=x,SO2=0)))}
pi1_4 <- function(x){plogis(predict(fittedModel4,data.frame(pH=x,SO2=1)))}
plot(logOdds_2,
     from=2.5,
     to=4.5,
     ylab="Log odds of high rating",
     xlab="Wine pH",
     main="Figure 1: Log odds linear in pH")
## note we should plot over a range of values of pH that are represented in the sample
plot(pi_2,
     from=2.5,
     to=4.5,
     ylab="Probability of high rating",
     xlab="Wine pH",
     main="Figure 2: Log odds linear in pH")
plot(logOdds0_3,
     from=2.5,
     to=4.5,
     ylim=c(-4.5,-0.5),
     ylab="Log odds of high rating",
     xlab="Wine pH",
     main="Figure 3: Log odds linear in pH and SO2")
plot(logOdds1_3,
     from=2.5,
     to=4.5,
     col="blue",
     add=TRUE)
legend(4,-0.5,
       legend=c(
         expression('Low SO'[2]),
         expression('High SO'[2])
         ),
       col=c("black","blue"),
       lwd=1)
plot(pi0_3,
```

```
     from=2.5,
     to=4.5,
     ylim=c(0,0.4),
     ylab="Probability of high rating",
     xlab="Wine pH",
     main="Figure 4: Log odds linear in pH and SO2")
plot(pi1_3,
     from=2.5,
     to=4.5,
     col="blue",
     add=TRUE)
legend(4,0.35,
       legend=c(
         expression('Low SO'[2]),
         expression('High SO'[2])
         ),
       col=c("black","blue"),
       lwd=1)
plot(logOdds0_4,
     from=2.5,
     to=4.5,
     ylim=c(-5,0),
     ylab="Log odds of high rating",
     xlab="Wine pH",
     main="Figure 5: Log odds linear in pH, SO2, and interaction")
plot(logOdds1_4,
     from=2.5,
     to=4.5,
     col="blue",
     add=TRUE)
legend(2.5,-1.5,
       legend=c(
         expression('Low SO'[2]),
         expression('High SO'[2])
         ),
       col=c("black","blue"),
       lwd=1)
plot(pi0_4,
     from=2.5,
     to=4.5,
     ylab="Probability of high rating",
     xlab="Wine pH",
     main="Figure 6: Log odds linear in pH, SO2, and interaction")
plot(pi1_4,
     from=2.5,
     to=4.5,
     col="blue",
     add=TRUE)
legend(2.5,0.2,
       legend=c(
         expression('Low SO'[2]),
         expression('High SO'[2])
         ),
       col=c("black","blue"),
       lwd=1)
```