w271 Lab 1: Investigation of the 1989 Space Shuttle Challenger Accident

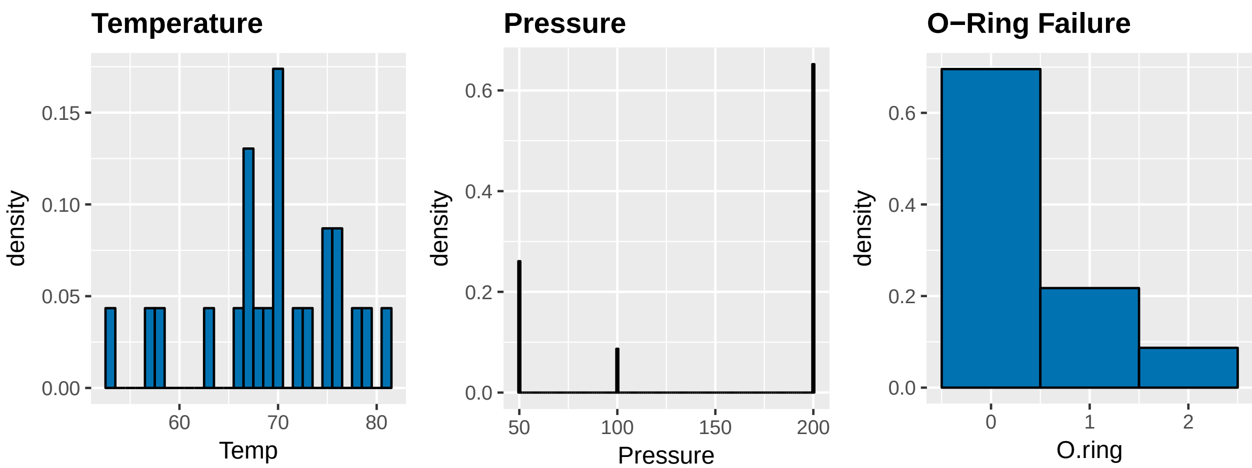
Jessica Hays Fisher, Alice Lam, Marshall Ratliff, Paul Varjan

5/20/2018

Exploratory Data Analysis

A First Look at the Individual Factors

```
temp.plt <- ggplot(df, aes(x = Temp)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill="#0072B2", colour="black") +
  ggtitle("Temperature") +
  theme(plot.title = element_text(lineheight=1, face="bold"))

pres.plt <- ggplot(df, aes(x = Pressure)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill="#0072B2", colour="black") +
  ggtitle("Pressure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))

oring.plt <- ggplot(df, aes(x = O.ring)) +
  geom_histogram(aes(y = ..density..), binwidth = 1, fill="#0072B2", colour="black") +
  ggtitle("O-Ring Failure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))

grid.arrange(temp.plt, pres.plt, oring.plt, ncol=3)
```



Basic Summary Data

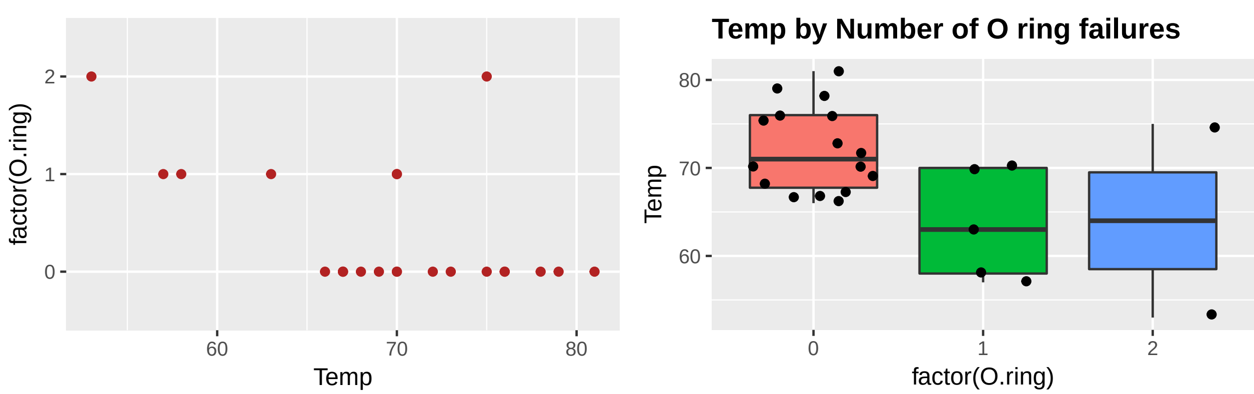
```
summary(df)
```

```
##      Flight          Temp          Pressure          O.ring      
##  Min.   : 1.0   Min.   :53.00   Min.   : 50.0   Min.   :0.0000  
##  1st Qu.: 6.5   1st Qu.:67.00   1st Qu.: 75.0   1st Qu.:0.0000  
##  Median :12.0   Median :70.00   Median :200.0   Median :0.0000  
##  Mean   :12.0   Mean   :69.57   Mean   :152.2   Mean   :0.3913  
##  3rd Qu.:17.5   3rd Qu.:75.00   3rd Qu.:200.0   3rd Qu.:1.0000  
##  Max.   :23.0   Max.   :81.00   Max.   :200.0   Max.   :2.0000  
##      Number 
##  Min.   :6  
##  1st Qu.:6  
##  Median :6  
##  Mean   :6  
##  3rd Qu.:6  
##  Max.   :6
```

```
# describe(df) # too long to fit everything else
```

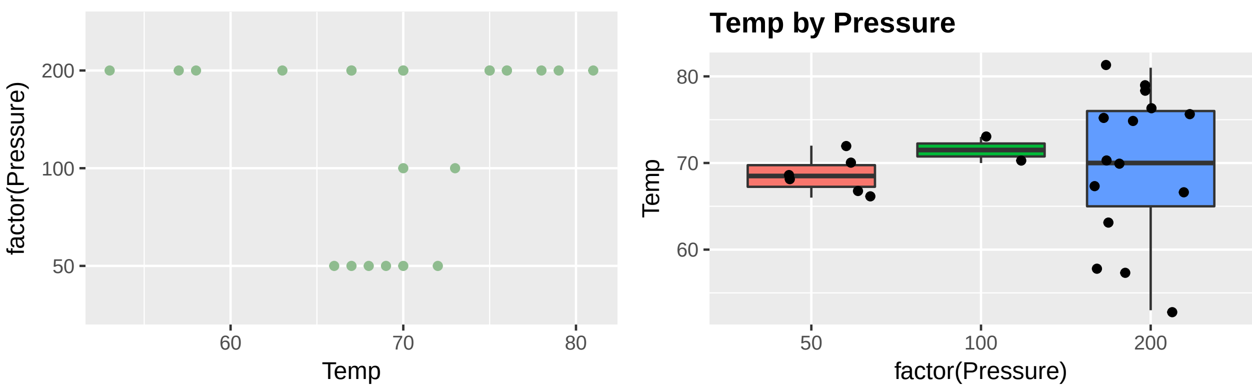
Relationships Between Time Series

```
otemp.plt <- ggplot(df, aes(Temp, factor(O.ring))) + geom_point(color="firebrick")
# was plot(O.ring ~ Temp, data = df)

otemp.box <- ggplot(df, aes(factor(O.ring), Temp)) +
  geom_boxplot(aes(fill = factor(O.ring))) +
  geom_jitter() + guides(fill=FALSE) +
  ggtitle("Temp by Number of O ring failures") +
  theme(plot.title = element_text(lineheight=1, face="bold"))

grid.arrange(otemp.plt, otemp.box, ncol=2)
```

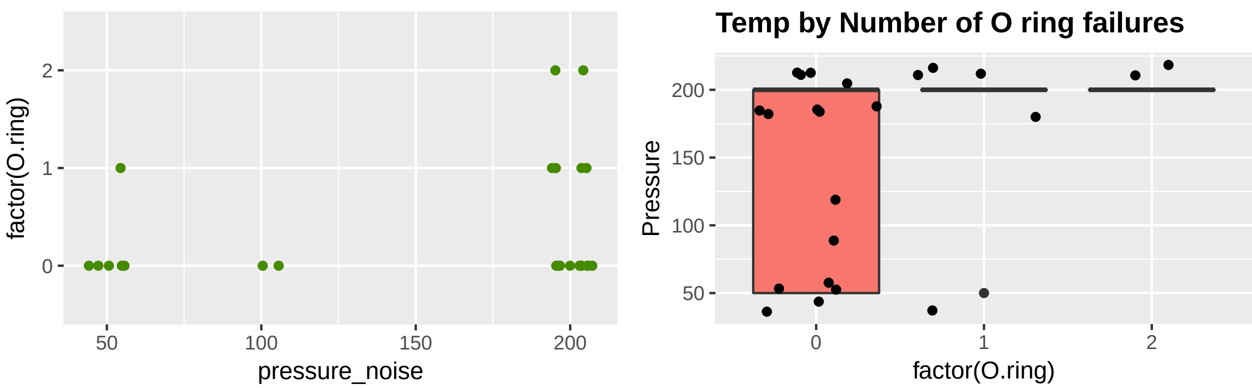



```
tpres.plt <- ggplot(df, aes(Temp, factor(Pressure))) + geom_point(color="darkseagreen")
# was plot(Pressure ~ Temp, data = df)
```

```r
tpres.box <- ggplot(df, aes(factor(Pressure), Temp)) +
  geom_boxplot(aes(fill = factor(Pressure))) +
  geom_jitter() + guides(fill=FALSE) +
  ggtitle("Temp by Pressure") +
  theme(plot.title = element_text(lineheight=1, face="bold"))

grid.arrange(tpres.plt, tpres.box, ncol=2)
```

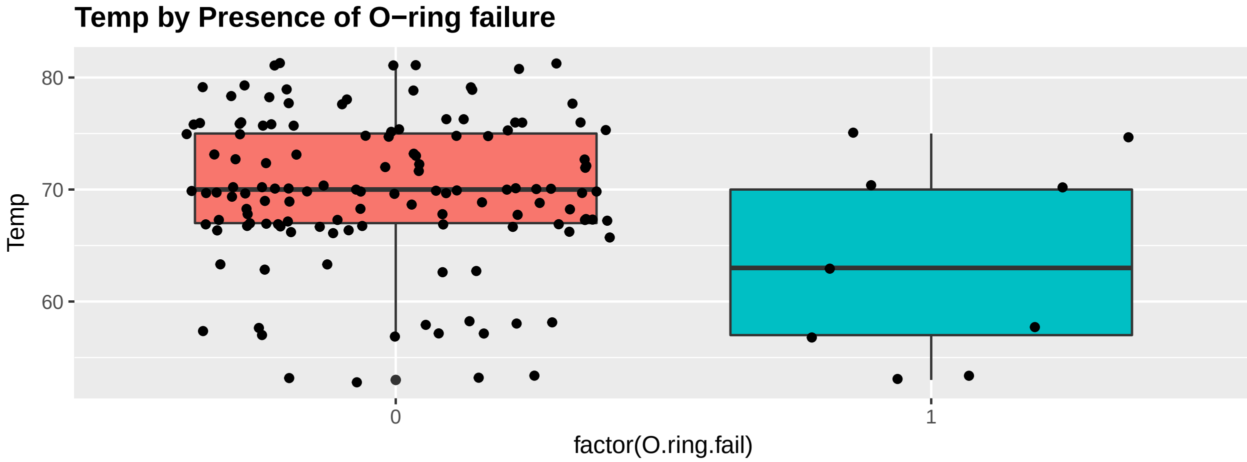



```r
noise <- runif(length(df$Pressure), min=-8, max = 8)
pressure_noise <- df$Pressure + noise
opres.plt <- ggplot(df, aes(pressure_noise, factor(O.ring))) + geom_point(color="chartreuse4")
# was opres.plt <- plot(pressure_noise, df$O.ring)

opres.box <- ggplot(df, aes(factor(O.ring), Pressure)) +
  geom_boxplot(aes(fill = factor(O.ring))) +
  geom_jitter() + guides(fill=FALSE) +
  ggtitle("Temp by Number of O ring failures") +
  theme(plot.title = element_text(lineheight=1, face="bold"))

grid.arrange(opres.plt, opres.box, ncol=2)
```




```r
ggplot(df2, aes(factor(O.ring.fail), Temp)) +
  geom_boxplot(aes(fill = factor(O.ring.fail))) +
  geom_jitter() + guides(fill=FALSE) +
  ggtitle("Temp by Presence of O-ring failure") +
```

```
  theme(plot.title = element_text(lineheight=1, face="bold"))
```



Answer to questions 4 and 5 on Chapter 2 (page 129 and 130) of Bilder and Loughin's *"Analysis of Categorical Data with R"*

4)

a. The authors assume that for each trial, the probability of failure for each of the 6 O-rings is independent. This is necessary to validate the use of the binomial distribution for the probability of failure. The binomial distribution assumes that the success/failure of each trial is independent, and in this case trials correspond to different O-rings in the same test. If the binomial distribution is not accurate, then this means the interpretation of the logistic regression implying the odds of success/failure for each O-ring is invalid. Conceivably, the failure of one O-ring may contribute to some structural damage that causes other O-rings to fail, violating the independence assumption. On the other side, the success of the primary O-ring may diminish the likelihood of failure of the second O-ring, if it does not experience the same conditions. Furthermore, there may be other variables that influence the quality of the O-rings or their likelihood of failure, for example related to their production, that violates the independence of O-rings on a given flight or different flights.

b. Base model of probability of single O.ring failures modeled on linear relationship of temperature and pressure. In df, we model each observation outcome using the count of O-ring failures, in the `O.ring` var, over the total number of O-rings, in the `Number` var which is always 6, as a binomial random variable. Similarly we check that this is the same as counting each O-ring as its own observation representing a bernoulli random variable with probability representing probability of its failure. These are indeed identical.

```
model1.binom <- glm(O.ring/Number ~ Temp + Pressure, data = df, family = binomial,
                    weights=Number)
```

```r
model1.bern <- glm(O.ring.fail ~ Temp + Pressure, data = df2, family = binomial)
summary(model1.binom)
```

```
##
## Call:
## glm(formula = O.ring/Number ~ Temp + Pressure, family = binomial,
##     data = df, weights = Number)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0361  -0.6434  -0.5308  -0.1625   2.3418
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.520195   3.486784   0.723   0.4698
## Temp        -0.098297   0.044890  -2.190   0.0285 *
## Pressure     0.008484   0.007677   1.105   0.2691
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 16.546  on 20  degrees of freedom
## AIC: 36.106
##
## Number of Fisher Scoring iterations: 5
```

```r
summary(model1.bern)
```

```
##
## Call:
## glm(formula = O.ring.fail ~ Temp + Pressure, family = binomial,
##     data = df2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7940  -0.3670  -0.2500  -0.2162   2.8127
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.520195   3.486822   0.723   0.4698
## Temp        -0.098297   0.044890  -2.190   0.0285 *
## Pressure     0.008484   0.007677   1.105   0.2691
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 66.540  on 137  degrees of freedom
## Residual deviance: 58.856  on 135  degrees of freedom
## AIC: 64.856
##
## Number of Fisher Scoring iterations: 6
```

Thus, we stick with the first as our model1:

```r
model1 <- glm(O.ring/Number ~ Temp + Pressure, data = df, family = binomial,
              weights=Number)
summary(model1)
```

```
##
## Call:
## glm(formula = O.ring/Number ~ Temp + Pressure, family = binomial,
##     data = df, weights = Number)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0361  -0.6434  -0.5308  -0.1625   2.3418
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.520195   3.486784   0.723   0.4698
## Temp        -0.098297   0.044890  -2.190   0.0285 *
## Pressure     0.008484   0.007677   1.105   0.2691
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 16.546  on 20  degrees of freedom
## AIC: 36.106
##
## Number of Fisher Scoring iterations: 5
```

c. We perform likelihood ratio tests using this model as our alternative hypothesis and the two reduced models setting the coeffs for temp and pressure respectively to zero, then conducting the ANOVA tests using the chi-squared distribution as follows:

```r
ha <- model1
h0 <- glm(O.ring/Number ~ Pressure, data = df, family = binomial, weights = Number)
anova(h0, ha, 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 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
h0 <- glm(O.ring/Number ~ Temp, data = df, family = binomial, weights = Number)
anova(h0, ha, 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
```

Thus we see that the inclusion of `Temp` in the model is significant at the alpha=0.05 level, whereas the inclusion of `Pressure` is not even marginally significant.

d. Given the lack of statistical significance of the pressure variable in the model here it certainly validates the authors decision to remove this variable from the model, however it is also reasonable to suggest that further testing may have still been warranted. It is important to keep in mind that we are assuming that the relationship with pressure is linear here, but some transformation may be relevant here, e.g. a log transformation or a translation given the note in the paper that the puddy covers pressure of 50 PSI and thus it may be that only additional pressure should be considered relevant to O-ring failure.

5)

a. The model on `Temp` alone corresponds to the second `h0` model above:

```
model2 <- h0
summary(model2)
```

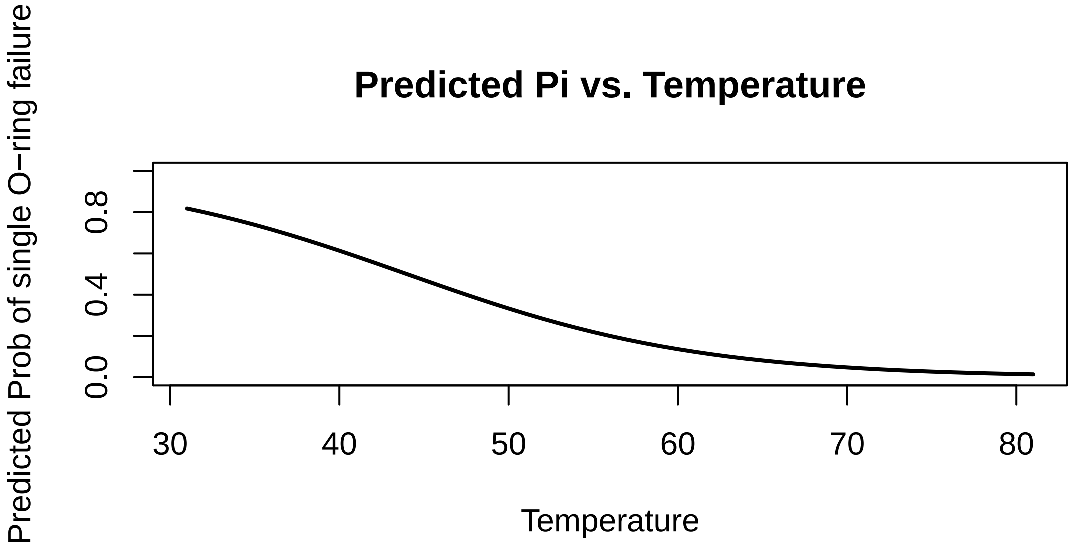
```
##
## Call:
## glm(formula = O.ring/Number ~ Temp, family = binomial, data = df,
##     weights = Number)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
```

```
## -0.95227  -0.78299  -0.54117  -0.04379   2.65152
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.08498    3.05247   1.666   0.0957 .
## Temp        -0.11560    0.04702  -2.458   0.0140 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 18.086  on 21  degrees of freedom
## AIC: 35.647
##
## Number of Fisher Scoring iterations: 5
```

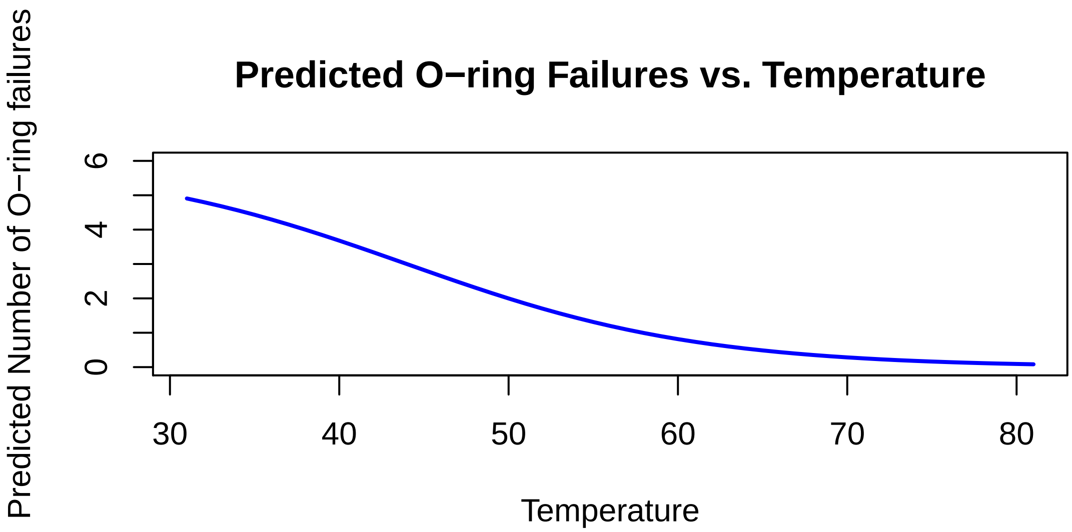
Using only a linear predictor on the `Temp` variable for the log-odds of yields an intercept of 5.085 and a coefficient for `Temp` of -0.116, which is significant at the 0.05 level.

b. Plot

```r
# pi vs. Temp  ## QUESTION - the books says to plot _pi_ vs temp, not pi hat. I
# assume they're looking for the latter, though?
newdf <- data.frame(Temp = seq(from = 31, to = 81, by = 1))

lp.hat <- predict.glm(model2, newdata = newdf, type = "link", se.fit = TRUE)
lp.hat.mean <- lp.hat$fit
pi.hat <- exp(lp.hat.mean) / (1 + exp(lp.hat.mean))

plot(newdf$Temp, pi.hat, ylim = range(c(0,1)),
     xlab = "Temperature", ylab = "Predicted Prob of single O-ring failure",
     main = "Predicted Pi vs. Temperature", type = 'l', col = 'black', lwd = 2)
```



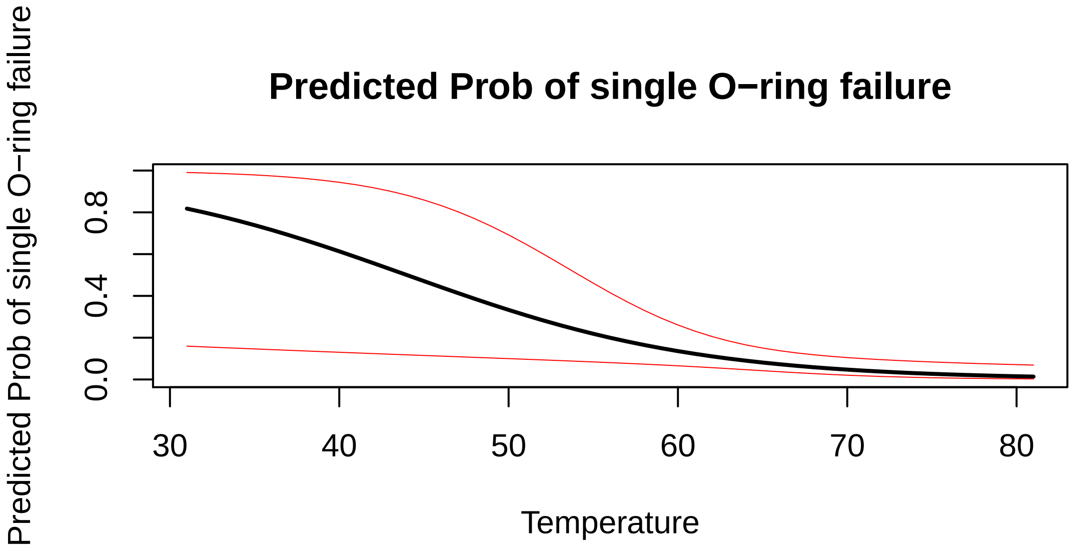

```r
#expected number of failures vs. Temp

plot(newdf$Temp, pi.hat * 6, ylim = range(c(0,6)),
     xlab = "Temperature", ylab = "Predicted Number of O-ring failures",
     main = "Predicted O-ring Failures vs. Temperature", type = 'l', col = 'blue',
     lwd = 2)
```



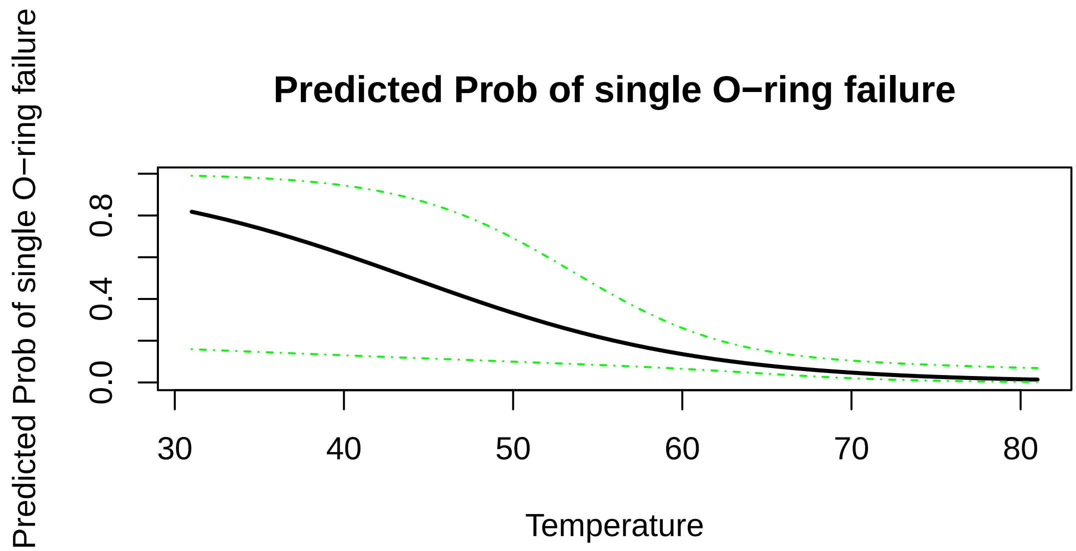
c. Plot. The bands are wider for lower temperature because there are very few observations in this region.

```r
#jeff's way
lp.hat.lci <- lp.hat$fit - 1.96 * lp.hat$se.fit
lp.hat.uci <- lp.hat$fit + 1.96 * lp.hat$se.fit


pi.hat.lci <- exp(lp.hat.lci) / (1 + exp(lp.hat.lci))
```

```
pi.hat.uci <- exp(lp.hat.uci) / (1 + exp(lp.hat.uci))

### Plot predicted probabilities
plot(newdf$Temp, pi.hat, ylim = range(c(pi.hat.lci, pi.hat.uci)),
     xlab = "Temperature", ylab = "Predicted Prob of single O-ring failure",
     main= "Predicted Prob of single O-ring failure", type = 'l', col = 'black',
     lwd = 2)
lines(newdf$Temp, pi.hat.lci, col = 'red', lwd = 0.5)
lines(newdf$Temp, pi.hat.uci, col = 'red', lwd = 0.5)
```




```
#book way
ci.pi <- function(newdata, mod.fit.obj, alpha){
  linear.pred <- predict(object = mod.fit.obj, newdata = newdata, type = "link",
                         se = TRUE)
  CI.lin.pred.lower <- linear.pred$fit - qnorm(p = 1-alpha/2)*linear.pred$se
  CI.lin.pred.upper <- linear.pred$fit + qnorm(p = 1-alpha/2)*linear.pred$se
  CI.pi.lower <- exp(CI.lin.pred.lower) / (1 + exp(CI.lin.pred.lower))
  CI.pi.upper <- exp(CI.lin.pred.upper) / (1+ exp(CI.lin.pred.upper))
  list(lower = CI.pi.lower, upper = CI.pi.upper)
}

plot(newdf$Temp, pi.hat, ylim = range(c(pi.hat.lci, pi.hat.uci)),
     xlab = "Temperature", ylab = "Predicted Prob of single O-ring failure",
     main= "Predicted Prob of single O-ring failure", type = 'l', col = 'black',
     lwd = 2)
curve(expr = ci.pi(newdata = data.frame(Temp = x), mod.fit.obj = model2,
                   alpha = 0.05)$lower, col = "green", lty = "dotdash", add = TRUE, xlim = c(3
curve(expr = ci.pi(newdata = data.frame(Temp = x), mod.fit.obj = model2,
                   alpha = 0.05)$upper, col = "green", lty = "dotdash", add = TRUE, xlim = c(3
```



d. Key assumption being made here is that there is a linear relationship between the temperature and the log-likelihood of O-ring failure. It is possible that either assumption is invalid, i.e. the logit is not the proper link-function for this relationship or there is a nonlinear relationship between temperature and the logit of the probability of O-ring failure.

The temperature was 31 at launch for the Challenger in 1986. Estimate the probability of an O-ring failure using this temperature, and compute a corre- sponding confidence interval. Discuss what assumptions need to be made in order to apply the inference procedures.

```
# Prob(failure) ~ temp = 31
model2.pred31 <- model2$coefficients[1] + model2$coefficients[2]*31
model2.pred31
```

```
## (Intercept)
##    1.501341
```

```
exp(model2.pred31)/(1+exp(model2.pred31))
```

```
## (Intercept)
##   0.8177744
```

```
# Another way to do it
predict.data<-data.frame(Temp=31)
predict(object = model2, newdata = predict.data, type = "link")
```

```
##        1
## 1.501341
```

```
predict(object = model2, newdata = predict.data, type = "response")
```

```
##         1
```

```
## 0.8177744
```

```r
# Wald CI
pred31 <- predict(object = model2, newdata = predict.data, type = "link", se = TRUE)
pred31
```

```
## $fit
##        1
## 1.501341
##
## $se.fit
## [1] 1.613565
##
## $residual.scale
## [1] 1
```

```r
pi.hat31 <- exp(pred31$fit) / (1 + exp(pred31$fit))
alpha <- 0.05
CI.pred31 <- pred31$fit + qnorm(p = c(alpha/2, 1-alpha/2))* pred31$se
CI.pi <- exp(CI.pred31)/(1 + exp(CI.pred31))
#CI.pi
data.frame(predict.data, pi.hat31, lower = CI.pi[1], upper = CI.pi[2])
```

```
##   Temp  pi.hat31     lower     upper
## 1   31 0.8177744 0.1596025 0.9906582
```

```r
# Profile Likelihood Ratio Interval
K <- matrix(data = c(1,31), nrow = 1, ncol = 2)
model2.combo <- mcprofile(object = model2, CM = K)
ci.logit.profile <- confint(object = model2.combo, level = 0.95)
#ci.logit.profile
exp(ci.logit.profile$confint)/(1 + exp(ci.logit.profile$confint))
```

```
##       lower     upper
## 1 0.1418508 0.9905217
```

At temperature of 31, the model predicted that the probability of O-ring failure is 0.8178. The 95% Wald interval for π is 0.1596 $< \pi <$ 0.9907. Since we have only 23 data points, which is $<$ 40, Wald CI generally does not work well. We therefore also check the profile likelihood ratio interval, the 95% interval for π is 0.1419 $< \pi <$ 0.9905. Despite small sample size, the profile likelihood ratio interval is not too far away from the Wald interval, thus we opt to report the profile likelihood ratio interval.

Key assumption being made here is that there is a linear relationship between the temperature and the log-likelihood of O-ring failure. It is possible that either assumption is invalid, i.e. the logit is not the proper link-function for this relationship or there is a nonlinear relationship between temperature and the logit of the probability of O-ring failure. As the range of data we have for Temp is only 28 degrees (from 53 to 81), 31 degree is 22 degree lower than the minimum Temp we observe, which is almost as far away as the range of data we observe. A slightly non-linear relationship may not be as obvious with a range of 28 degrees difference, but at 31 degree the deviance from linear relationship would be much more prominent.

CAN BE REMOVED ### From Async: ### So we have learned that the Wald confidence intervals do not necessarily perform well. And so we want to estimate a profile likelihood ration interval, as well. However, computationally, the profile value ratio is a lot more difficult, because it involves a large number of parameters.

And so we don't do that manually. What we recommend you guys to do is to use this mcprofile package– that is a user-written package, but it's available in R– and just the following steps. Earlier versions of this particular package do not provide very stable results. But the current version actually provides results that are much more stable and don't have the problems associated with early versions.

And so we recommend the following. So when it comes to calculating confidence intervals, we recommend that we first calculate a Wald interval. And then we calculate a profile likelihood ratio interval using the mcprofile package.

And as long as the profile likelihood ratio interval not too far away from the Wald interval, we recommend reporting the profile likelihood ration interval.

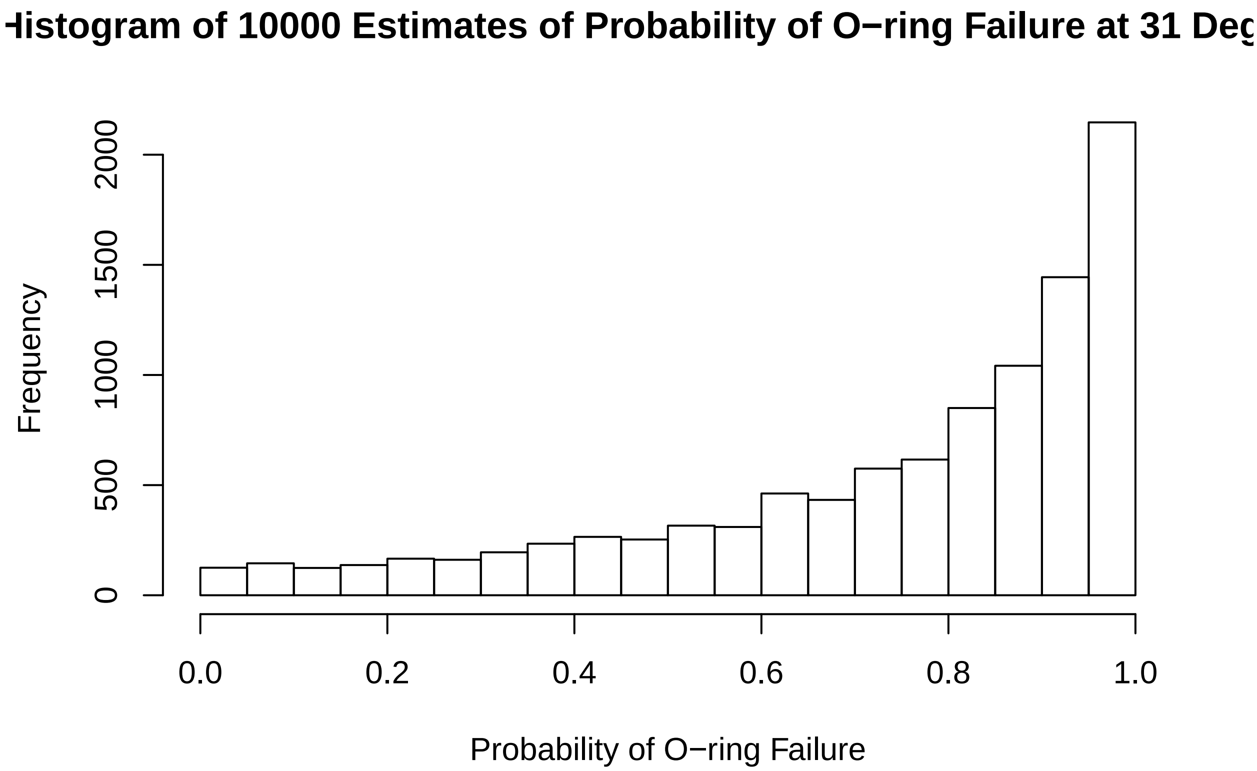
e. Bootstrap

```
#define sigmoid function for computing values of pi
sigmoid = function(x) {
  1 / (1 + exp(-x))
}

#start with the parameter estimates from our model and our Temp data
beta0 = model2$coefficients[1]
beta1 = model2$coefficients[2]
x <- df$Temp
```

```r
pi <- sigmoid(beta0 + beta1*x)
weights <- df$Number

#simulate new O.ring failure counts to estimate new model parameters
sim <- function(){
  #simulate new O.ring failure counts as binomial random variable with n=6
  #trials and p=pi probability of success
  y <- rbinom(n = length(x), size = 6, prob = pi)

  #fit a new regression model on the simulated O.ring failure counts
  mod.fit <- glm(y/weights ~ x, family = binomial, weights = weights)
  beta0.star = mod.fit$coefficients[1]
  beta1.star = mod.fit$coefficients[2]

  #use new model to compute predicted probability of O.ring failure at Temp = 31
  #and 72 degrees
  pi_star.31degrees <- sigmoid(beta0.star + beta1.star*31)
  pi_star.72degrees <- sigmoid(beta0.star + beta1.star*72)
  return(c(pi_star.31degrees,pi_star.72degrees))
}

#run simulation 10000 times
n=10000
sim_vals <- replicate(n,sim())
```

```
## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

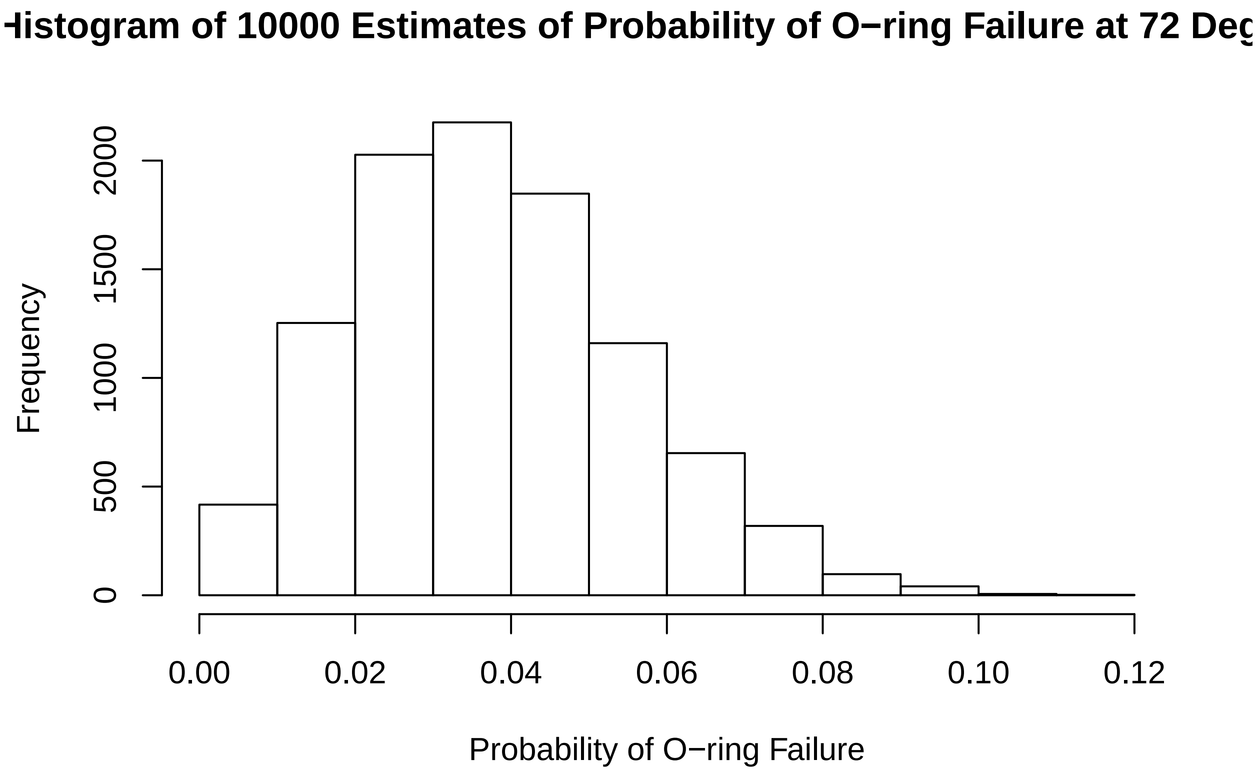
```r
#plot distribution of computed pi values and return the 90% conf interval for
#Temp = 31 degrees
hist(sim_vals[1,], freq = T, xlab = "Probability of O-ring Failure",
     main = "Histogram of 10000 Estimates of Probability of O-ring Failure at 31 Degrees")
```




```r
quantile(sim_vals[1,],c(0.05,0.95))
```

```
##        5%       95%
## 0.1900356 0.9892855
```

```r
#plot distribution of computed pi values and return the 90% conf interval for
#Temp = 72 degrees
hist(sim_vals[2,], freq = T, xlab = "Probability of O-ring Failure",
     main = "Histogram of 10000 Estimates of Probability of O-ring Failure at 72 Degrees")
```




```
quantile(sim_vals[2,],c(0.05,0.95))
```

```
##          5%         95%
## 0.01127681 0.06930500
```

f. We include the quadratic term on temperature and run a LRT using the chi-squared distribution to determine if its inclusion is statistically significant, as follow:

```
model3 <- glm(O.ring/Number ~ Temp + I(Temp^2), data = df, family = binomial,
              weights = Number)
summary(model3)
```

```
##
## Call:
## glm(formula = O.ring/Number ~ Temp + I(Temp^2), family = binomial,
##     data = df, weights = Number)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.84320  -0.72385  -0.61980  -0.01335   2.52101
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 22.126148  23.794426   0.930    0.352
## Temp        -0.650885   0.740756  -0.879    0.380
```

```
## I(Temp^2)    0.004141   0.005692   0.727    0.467
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 17.592  on 20  degrees of freedom
## AIC: 37.152
##
## Number of Fisher Scoring iterations: 5
```

```
ha <- model3
anova(h0, ha, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: O.ring/Number ~ Temp
## Model 2: O.ring/Number ~ Temp + I(Temp^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        21     18.086
## 2        20     17.592  1   0.4947   0.4818
```

Testing other Models before moving on

Temp and Pressure with an Interaction term - an interaction term does not help improve the model.

```
model2 <- glm(O.ring/Number ~ Temp + Pressure + Temp:Pressure, data = df,
              family = binomial, weights = Number)
summary(model2)
```

```
##
## Call:
## glm(formula = O.ring/Number ~ Temp + Pressure + Temp:Pressure,
##     family = binomial, data = df, weights = Number)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0279  -0.6600  -0.5141  -0.1744   2.3724
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -21.765547  46.285607  -0.470    0.638
## Temp            0.252463   0.662989   0.381    0.703
## Pressure        0.130857   0.232643   0.562    0.574
## Temp:Pressure  -0.001769   0.003336  -0.530    0.596
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 16.257  on 19  degrees of freedom
## AIC: 37.817
##
## Number of Fisher Scoring iterations: 6
```

The quadratic term addition to the model is not statistically significant, suggesting that either it shouldn't be included or some other variable transformations or terms should be conducted/tested first.

3. In addition to the questions in Question 4 and 5, answer the following questions:

a. Interpret the main result of your final model in terms of both odds and probability of failure

```r
df$bin.Temp = df$Temp<65
model4 <- glm(O.ring/Number ~ bin.Temp, data = df, family = binomial,
              weights = Number)
summary(model4)
```

```
##
## Call:
## glm(formula = O.ring/Number ~ bin.Temp, family = binomial, data = df,
##     weights = Number)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6547  -0.6547  -0.6547  -0.2582   2.4591
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.3142     0.5090  -6.511 7.46e-11 ***
## bin.TempTRUE   1.9792     0.7153   2.767  0.00566 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 16.911  on 21  degrees of freedom
## AIC: 34.471
##
## Number of Fisher Scoring iterations: 5
```

```r
exp(1.9792)
```

```
## [1] 7.236951
```

```r
exp(-0.1156)
```

```
## [1] 0.8908315
```

After eliminating other potential covariates like order of launch, pressure, cross terms, square terms and log terms, we tested one more thing. Using a visual cue from the original Temp vs. failure chart in the EDA, we replaced the continuous variable Temp with a binary variable Temp<65. Our coefficient for the binary variable had an estimate of 1.9792, and while it had a relatively large standard error, was still highly significant. Exp(1.9792) = 7.327, which means that if the temperature is below the 65 degree threshold a failure is 6.327 times as likely to occur as it would if the temperature is above the 65 degree threshold. This is a nice tidy answer, is reflective of our observations in EDA and has a great p-value but it reeks of p-hacking, and it would not be robust to further declining temperatures.

As a result it is ultimately best to go back to the basic O.ring ~ Temp single factor, loglinear(is this accurate?) regression model. That model is not as dramatic in terms of statistical significance but is still around a 95% confidence level and is far less forced. It implies that with every one degree increase in temperature the likelihood of an o-ring failure decreases 11% from what it was, and vice versa.

b. With the same set of explanatory variables in your final model, estimate a linear regression model. Explain the model results; conduct model diagnostic; and assess the validity of the model assumptions. Would you use the linear regression model or binary logistic regression in this case. Why? Or, why not?

THESE ARE MODELS FROM OTHER .RMD FILES JUST TO PREVENT LOSS OF USEFUL THINGS - SOME OF THESE DONT RUN

```r
df$bin.Temp = df$Temp<67
model4 <- glm(O.ring/Number ~ bin.Temp, data = df, family = binomial,
              weights = Number)
summary(model4)
```

```
##
## Call:
## glm(formula = O.ring/Number ~ bin.Temp, family = binomial, data = df,
##     weights = Number)
##
## Deviance Residuals:
##    Min       1Q   Median       3Q      Max
## -1.479   -0.673   -0.673    0.000    2.418
##
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.2581     0.5095  -6.394 1.61e-10 ***
## bin.TempTRUE   1.6487     0.7068   2.332   0.0197 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.23  on 22  degrees of freedom
## Residual deviance: 18.94  on 21  degrees of freedom
## AIC: 36.501
##
## Number of Fisher Scoring iterations: 5
```

```r
df$bin.Temp = df$Temp<65
model4 <- glm(O.ring/Number ~ bin.Temp, data = df, family = binomial,
              weights = Number)
summary(model4)
```

```
##
## Call:
## glm(formula = O.ring/Number ~ bin.Temp, family = binomial, data = df,
##     weights = Number)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6547  -0.6547  -0.6547  -0.2582   2.4591
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.3142     0.5090  -6.511 7.46e-11 ***
## bin.TempTRUE   1.9792     0.7153   2.767  0.00566 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24.230  on 22  degrees of freedom
## Residual deviance: 16.911  on 21  degrees of freedom
## AIC: 34.471
##
## Number of Fisher Scoring iterations: 5
```

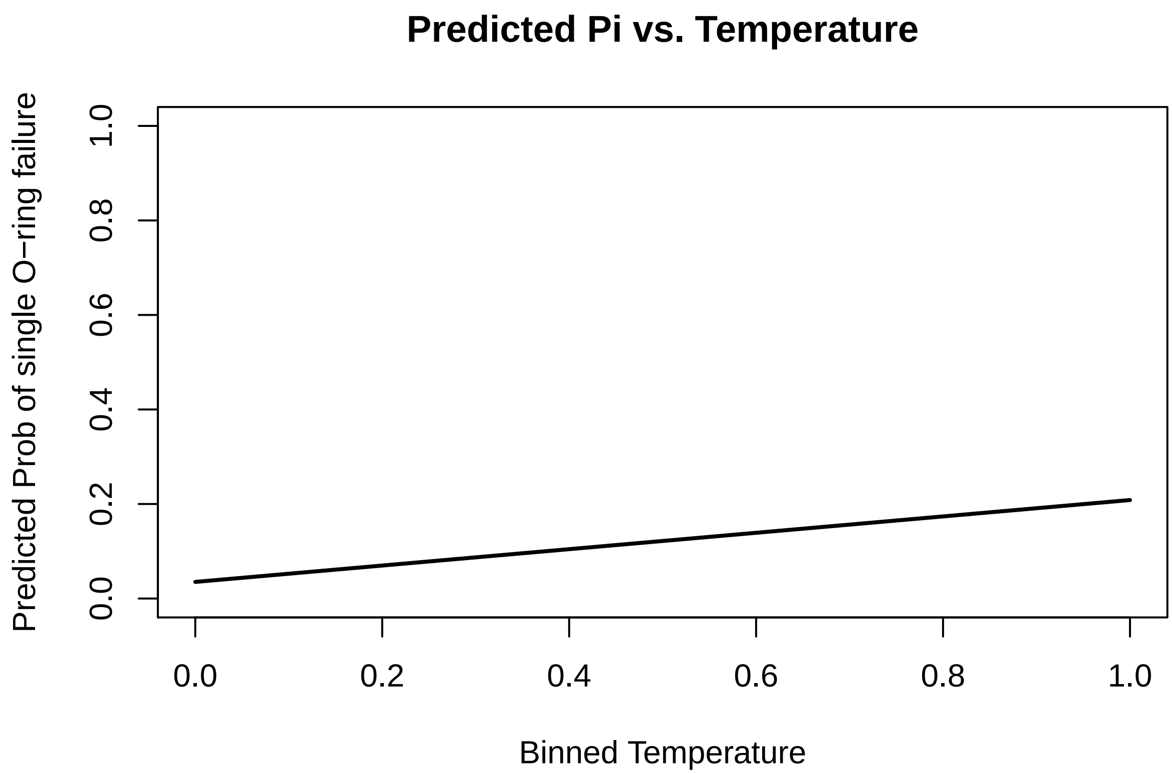
```r
exp(1.9792)
```

```
## [1] 7.236951
```

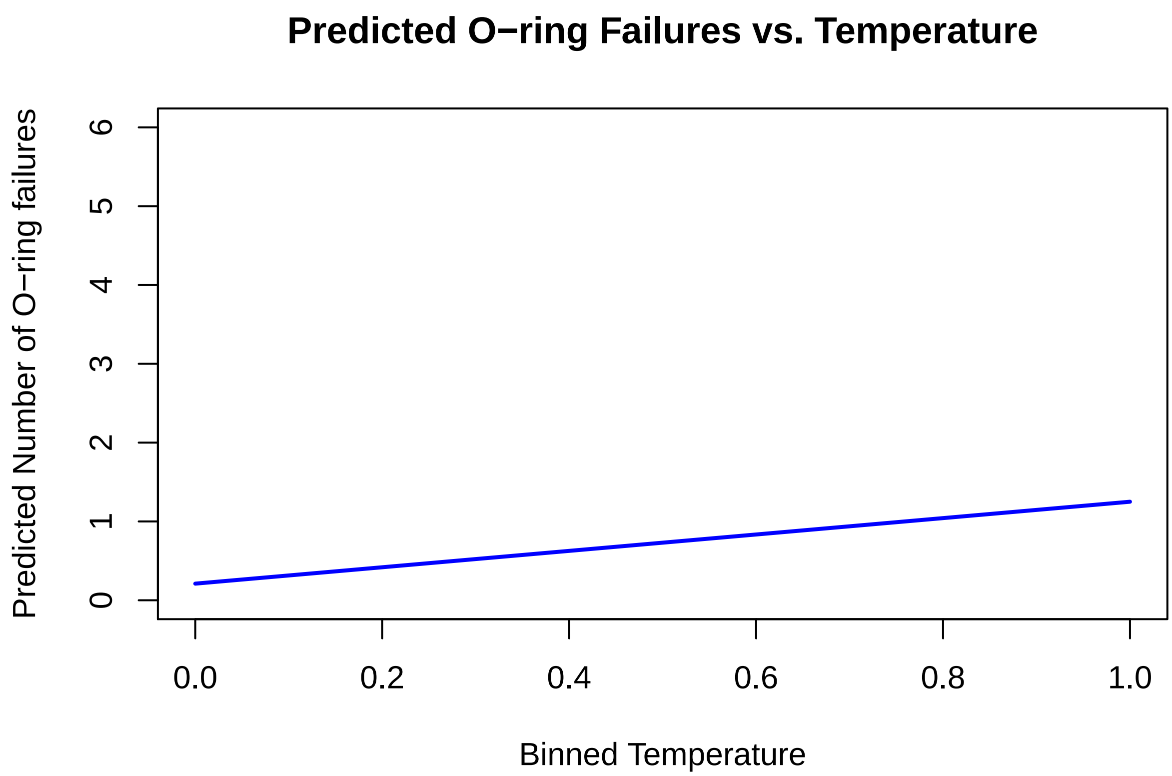
```r
exp(-0.1156)
```

```
## [1] 0.8908315
```

```r
#pi vs. Temp  ## QUESTION - the books says to plot _pi_ vs temp, not pi hat. I
#assume they're looking for the latter, though?
newdf <- data.frame(bin.Temp = c(TRUE,F))

lp.hat <- predict.glm(model4, newdata = newdf, type = "link", se.fit = TRUE)
lp.hat.mean <- lp.hat$fit
pi.hat <- exp(lp.hat.mean) / (1 + exp(lp.hat.mean))

plot(newdf$bin.Temp, pi.hat, ylim = range(c(0,1)),
     xlab = "Binned Temperature", ylab = "Predicted Prob of single O-ring failure",
     main = "Predicted Pi vs. Temperature", type = 'l', col = 'black', lwd = 2)
```




```r
#expected number of failures vs. Temp

plot(newdf$bin.Temp, pi.hat * 6, ylim = range(c(0,6)),
     xlab = "Binned Temperature", ylab = "Predicted Number of O-ring failures",
     main = "Predicted O-ring Failures vs. Temperature", type = 'l', col = 'blue',
     lwd = 2)
```




```
{r} model4 <- glm(O.ring/Number ~ Temp + log.Pressure, data =
df, family = binomial, weights = Number) summary(model4) ha <-
model4 anova(h0, ha, test = "Chisq")
```

```
{r} model4 <- glm(O.ring/Number ~ Temp + translate.Pressure,
data = df, family = binomial, weights = Number) summary(model4)
ha <- model4 anova(h0, ha, test = "Chisq")
```