Project 2: Horse Colic

Melodie Irvin, mki77

Colic is a cause for many horse emergencies in veterinary medicine due to the severity of the abdominal obstructions that may occur. Colic itself is not a diagnosis, but rather a symptom of a larger problem that is usually found in the Gastrointestinal tract. This dataset, found on Kaggle, consists of different parameters measuring the health of 209 horses with colic. The main variables I will be focusing on are surgery (yes or no), outcome (lived or died), abdomen, total protein, and packed cell volume. Total Protein is a measure of the total amount of Albumin and Globulin in the blood, and increased levels typically indicate dehydration in the mammal. Packed cell volume is a measure of the amount of red blood cells by volume in blood, and like total protein counts, an increase in levels may indicate dehydration. Dehydration that results in colic is a common problem that can be caused by a change in feed, in addition to a loss of water intake. This combination results in obstructions that occur in the abdomen. The variable for abdomen measures possible indexes of abdominal presentation. Distended large and small intestines typically indicate an obstruction present in the GI tract due to a mechanical impaction which often requires immediate surgery. Less severe abdominal findings, however, may not need surgery and can be fixed with the use of different types of medications.

```
options(repos="https://cran.rstudio.com" )
library(readxl);library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------------------------------ tidy
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
horse <- read_excel("C:/Users/Melodie/Desktop/horse.xlsx")
man1<-manova(cbind(total_protein,packed_cell_volume)~outcome, data=horse)
summary(man1)
```

```
##            Df  Pillai approx F num Df den Df   Pr(>F)
## outcome     1 0.23698   31.989      2    206 7.97e-13 ***
## Residuals 207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary.aov(man1)
```

```
##  Response total_protein :
##              Df Sum Sq Mean Sq F value   Pr(>F)
## outcome       1   6949  6949.3  10.076 0.001731 **
## Residuals   207 142766   689.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response packed_cell_volume :
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## outcome       1  4472.8  4472.8  51.754 1.132e-11 ***
## Residuals   207 17889.5    86.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
horse%>%group_by(outcome)%>%summarize(mean(total_protein),mean(packed_cell_volume))
```

```
## # A tibble: 2 x 3
##   outcome `mean(total_protein)` `mean(packed_cell_volume)`
##   <chr>                   <dbl>                      <dbl>
## 1 died                     15.6                       52.1
## 2 lived                    27.4                       42.6
```

```
pairwise.t.test(horse$total_protein,horse$outcome,
                p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  horse$total_protein and horse$outcome
##
##       died
## lived 0.0017
##
## P value adjustment method: none
```

```
pairwise.t.test(horse$packed_cell_volume,horse$outcome,
                p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  horse$packed_cell_volume and horse$outcome
##
##       died
## lived 1.1e-11
##
## P value adjustment method: none
```

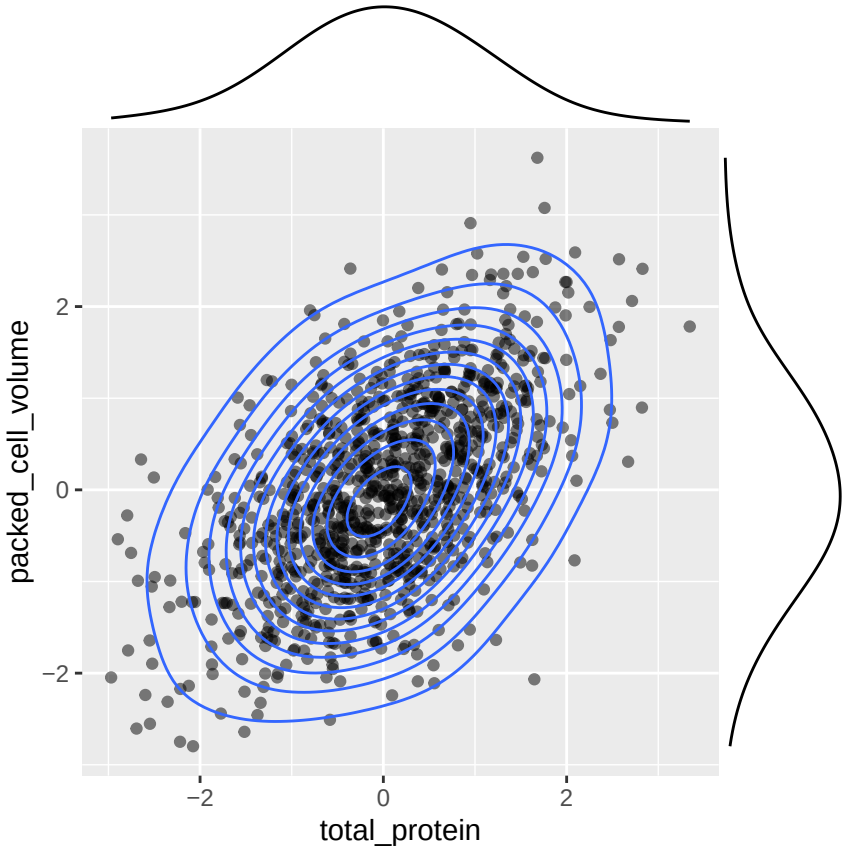
```
1-.95^5
```

```
## [1] 0.2262191
```

```
.05/5
```

```
## [1] 0.01
```

```
library(mvtnorm); library(ggExtra)
df<-rmvnorm(1000,mean=c(0,0),sigma=matrix(c(1,.5,.5,1),ncol=2,byrow=T))
df<-data.frame(df)%>%rename(total_protein=X1,packed_cell_volume=X2)
p<-ggplot(df, aes(total_protein,packed_cell_volume))+geom_point(alpha=.5)+geom_density_2d(h=2)+coord_fi
ggMarginal(p,type="density",xparams = list(bw=.5), yparams=list(bw=.5))
```

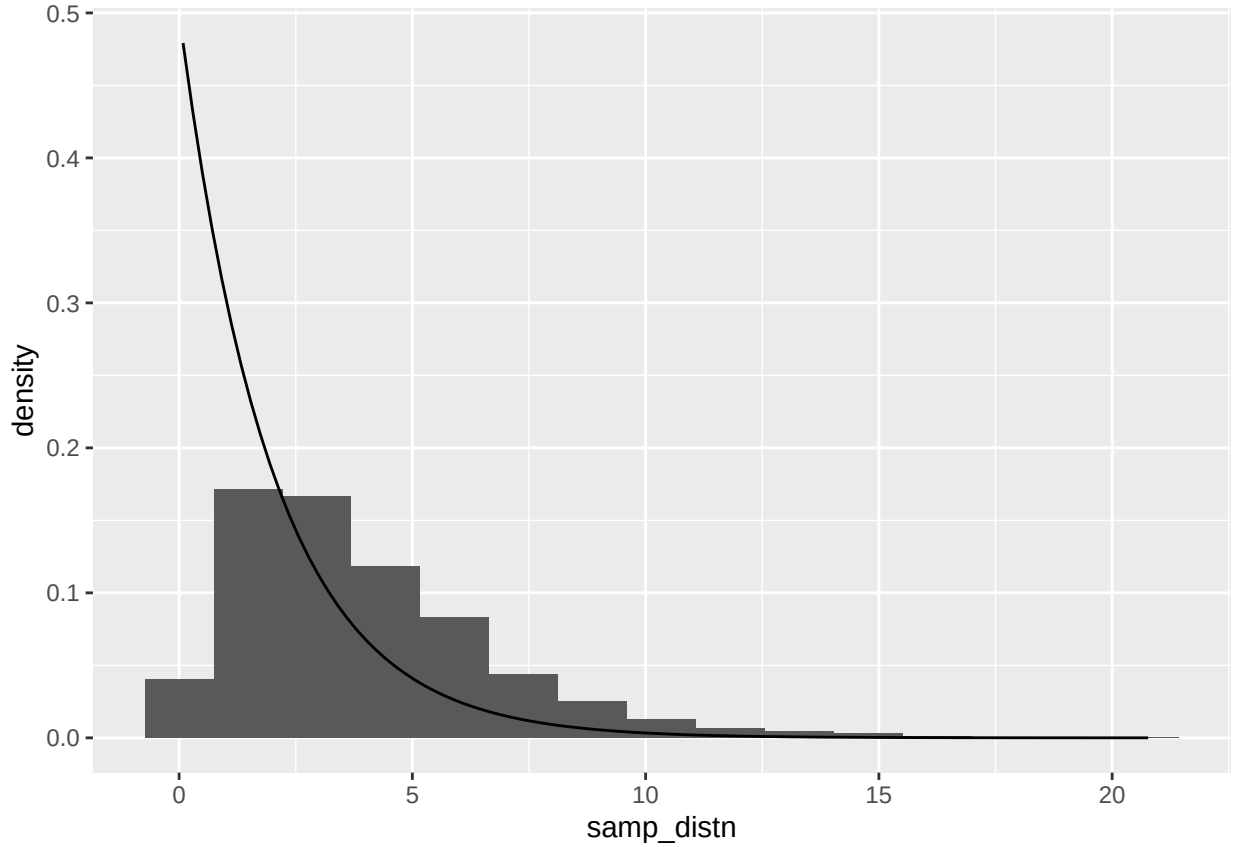



```
cov(df)
```

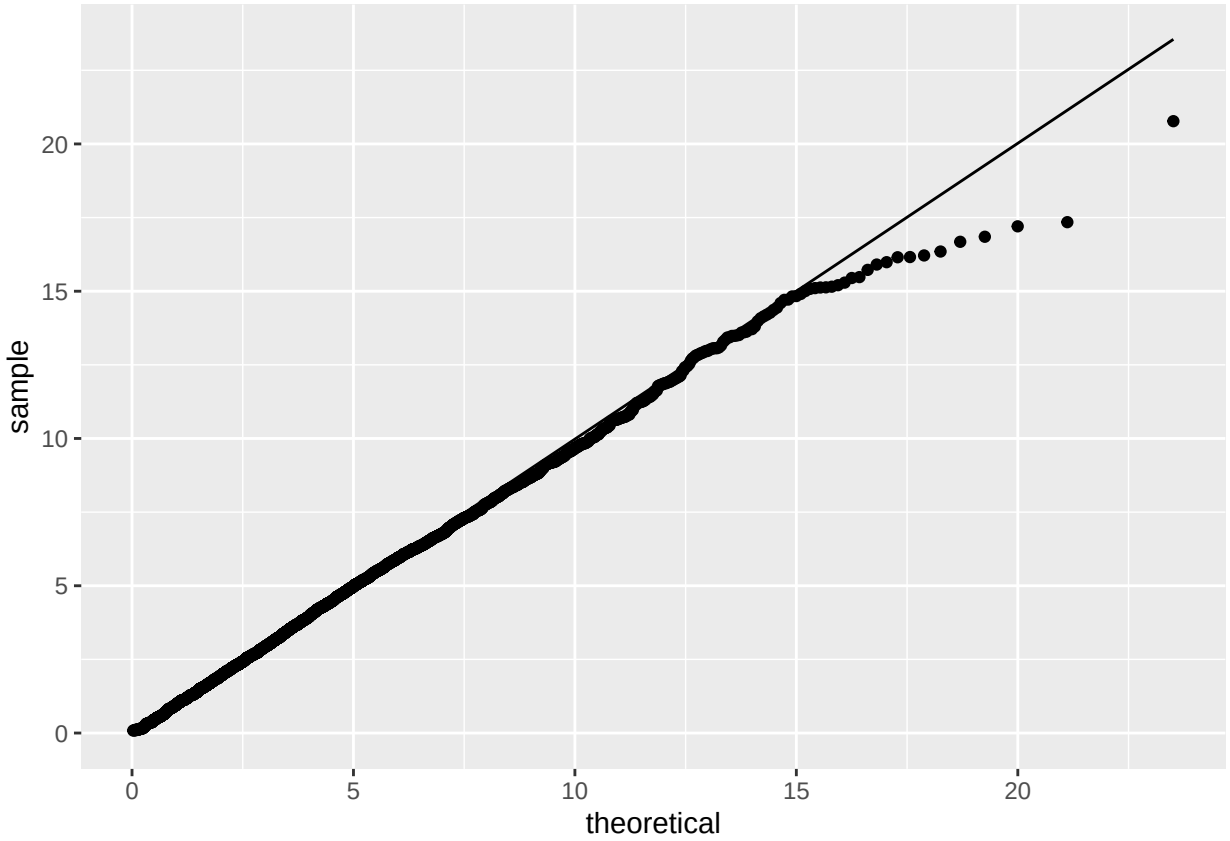
```
##                    total_protein packed_cell_volume
## total_protein          0.9965721          0.5625676
## packed_cell_volume     0.5625676          1.0503993
```

Given the large sample size and further confirmed with the ggplot, the data meets the assumption of multivariate normality. The data also appears to meet the assumption of homogenaity of covariances from the values presented in the table. Given the significance of the overall MANOVA, additional tests were performed to test significance for each variable using univariate ANOVAs. After these were found to be significant, pairwise t.tests were performed to test p-values of each combination of tested variables. A total of five tests were performed, and as such, the probability of a type-one error was given to be 22.6%. However, because of these multiple comparisons a bonferroni's coorection was determined, giving a new significant p-value of .01. Even with this adjustment, all p-values remained significant, signifying that both outcomes were found to differ significantly from one another in terms of total protein and packed cell volume values.

```r
samp_distn<-vector()
for(i in 1:5000){
  horse$pain<-sample(horse$pain)
  obs<-table(horse$pain,horse$surgery)
  exp<-outer(rowSums(obs),colSums(obs),"*")/sum(obs)
  samp_distn[i]<-sum((obs-exp)^2/exp)
}
data.frame(samp_distn)%>%
  ggplot(aes(samp_distn))+geom_histogram(aes(y=..density..),bins = 15)+
  stat_function(fun=dchisq,args=list(df=2),geom="line")
```




```r
data.frame(samp_distn)%>%
  ggplot(aes(sample=samp_distn)) +
  stat_qq(distribution = qchisq, dparams = list(df=4)) +
  stat_qq_line(distribution = qchisq, dparams = list(df=4))
```



HO: Whether or not the horse had surgery is independent of the level of its pain. HA: Whether or not the horse had surgery is not independent of the level of its pain. A simulation of the chi square test of independence was performed to test whether surgery and pain presentation were independent of one another. The distribution of 5,000 draws based on the ggplot, does not closely match the true chi square distribution as seen on the overlaid line. This discrepancy is also seen on the qqplot, as the sample and theoretical values are not a very good match. As a result, it can be concluded that we reject the null hypothesis, finding that whether or not the horse had surgery depends on the level of its pain.

```
library(ggplot2);library(lmtest);library(sandwich)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

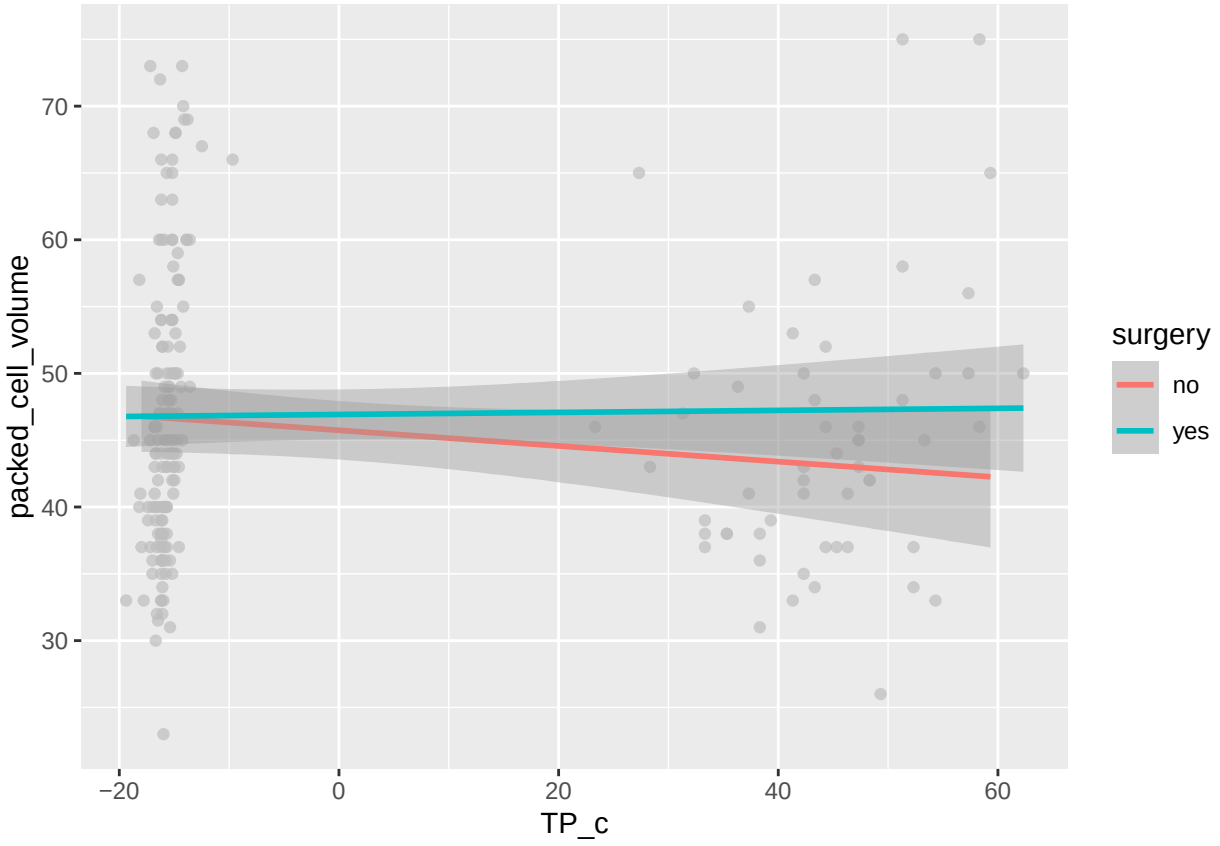
```
horse1<-horse
horse1$TP_c <- horse1$total_protein - mean(horse1$total_protein)
fit<-lm(packed_cell_volume~age+TP_c+abdominal_distention+surgery:TP_c, data=horse1)
summary(fit)
```

```
##
```

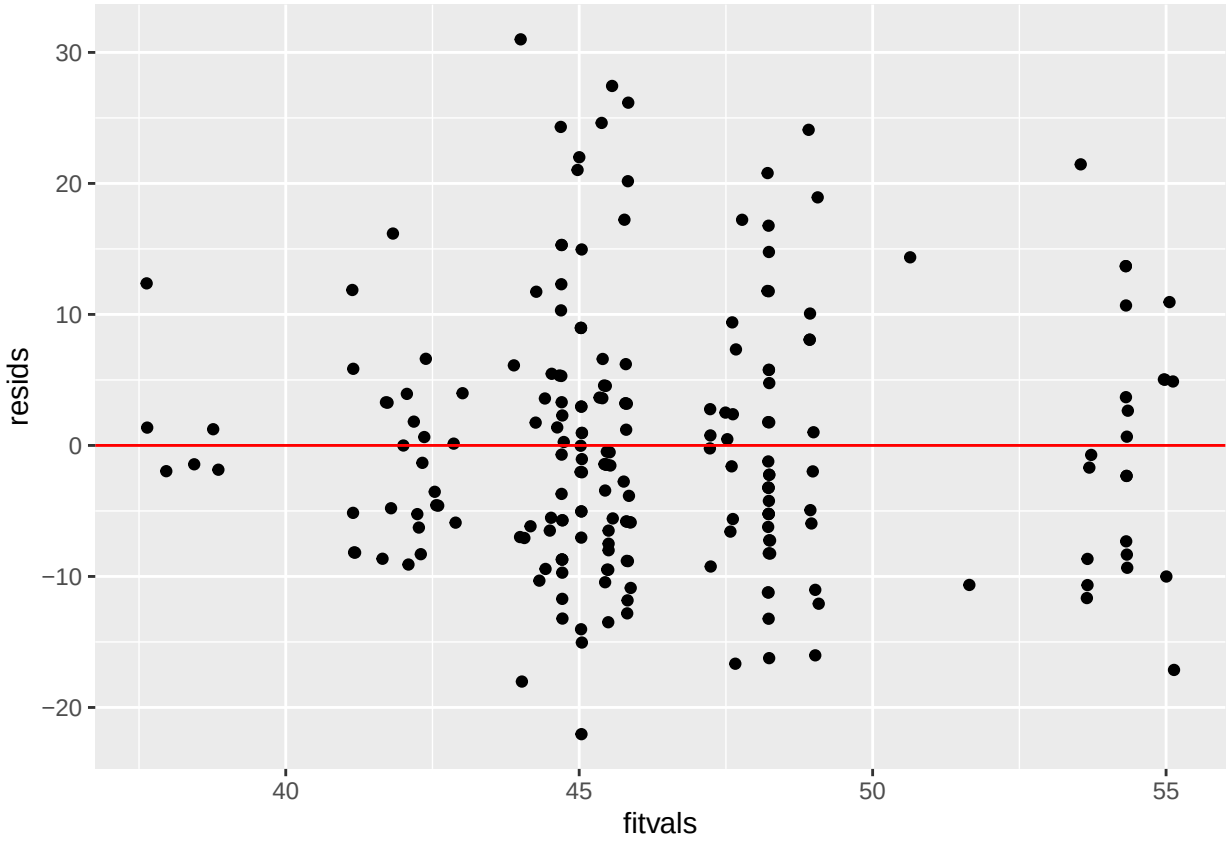
```
## Call:
## lm(formula = packed_cell_volume ~ age + TP_c + abdominal_distention +
##     surgery:TP_c, data = horse1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.038  -6.992  -1.442   4.883  30.997
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 48.06257    1.32706  36.217  < 2e-16 ***
## ageyoung                    -7.08464    2.70467  -2.619  0.00948 **
## TP_c                        -0.05931    0.03995  -1.485  0.13920
## abdominal_distentionnone    -3.19217    1.81260  -1.761  0.07973 .
## abdominal_distentionsevere   6.09516    2.25733   2.700  0.00752 **
## abdominal_distentionslight -3.52117    1.83149  -1.923  0.05594 .
## TP_c:surgeryyes              0.04881    0.05194   0.940  0.34843
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.843 on 202 degrees of freedom
## Multiple R-squared:  0.1248, Adjusted R-squared:  0.09878
## F-statistic:   4.8 on 6 and 202 DF,  p-value: 0.0001337
```

```
fit %>%
  ggplot() +
  aes(x = TP_c, y = packed_cell_volume, group = surgery, color = surgery) +
  geom_point(color = "grey", alpha = .7) +
  geom_smooth(method = "lm")
```

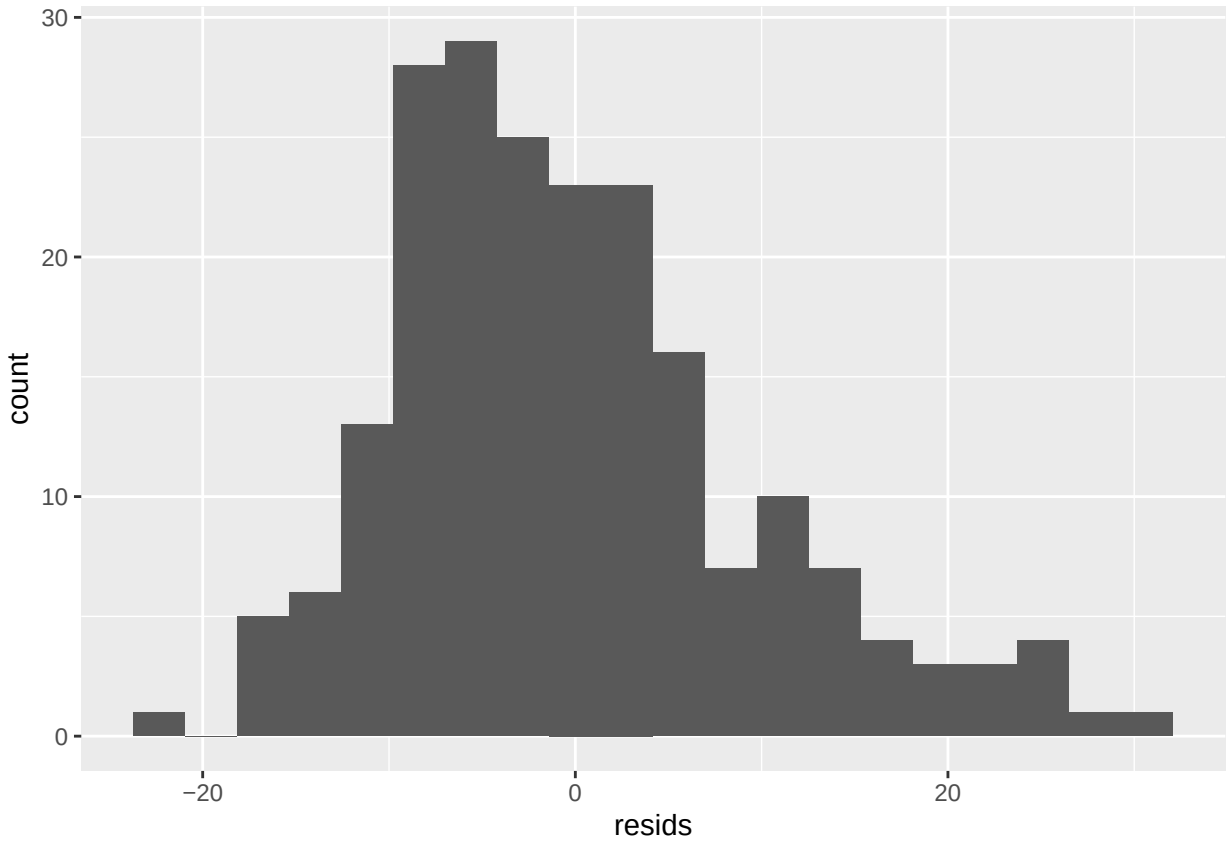
```
## `geom_smooth()` using formula 'y ~ x'
```



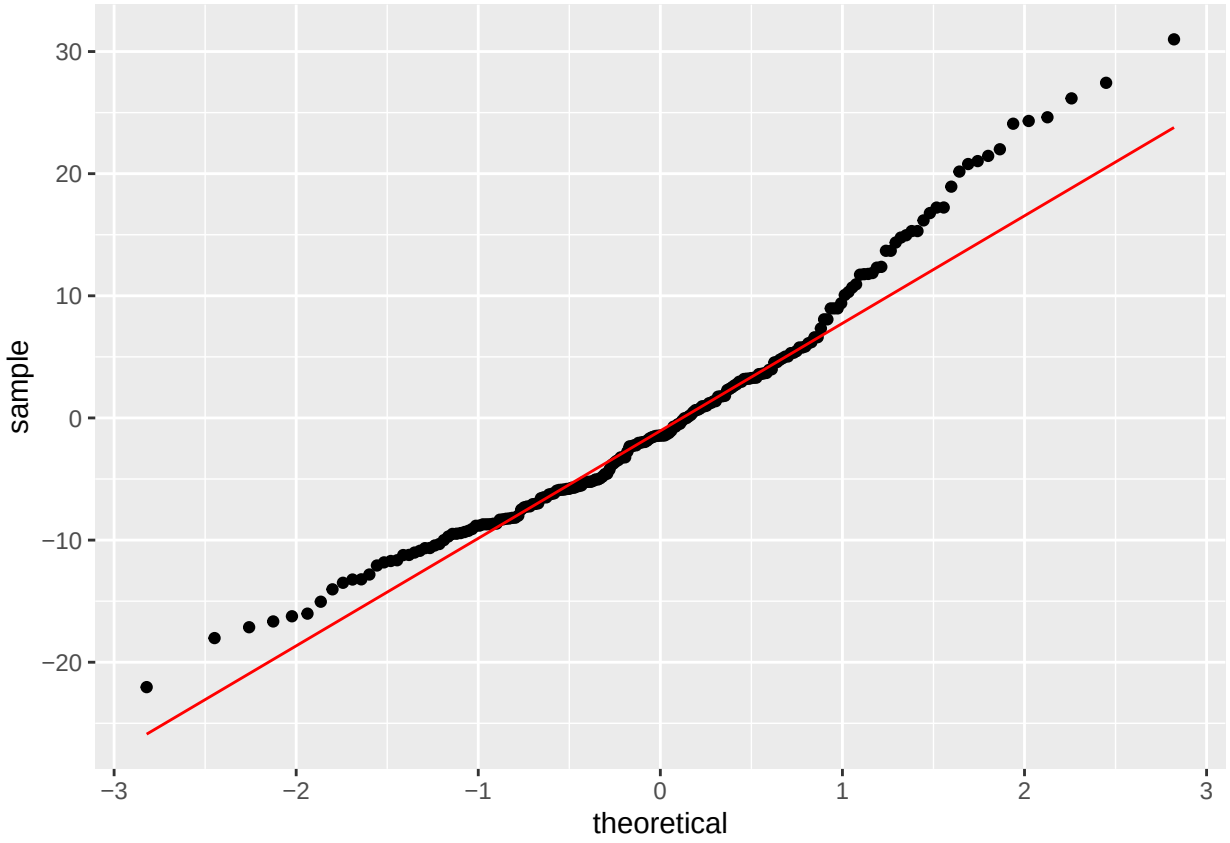

```
resids<-fit$residuals; fitvals<-fit$fitted.values
ggplot()+geom_point(aes(fitvals,resids))+geom_hline(yintercept=0, col="red")
```




```
ggplot()+geom_histogram(aes(resids),bins=20)
```




```
ggplot()+geom_qq(aes(sample=resids))+geom_qq_line(aes(sample=resids), color='red')
```




```
bptest(fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 5.9423, df = 6, p-value = 0.4297
```

```
coeftest(fit, vcov = vcovHC(fit))[,1:2] #Robust SEs. Get normal SEs from fit
```

```
##                               Estimate Std. Error
## (Intercept)                 48.06256706 1.33324160
## ageyoung                    -7.08463957 2.00070523
## TP_c                        -0.05931106 0.03612691
## abdominal_distentionnone    -3.19217178 1.81330103
## abdominal_distentionsevere   6.09516031 2.24806217
## abdominal_distentionslight -3.52117451 1.89147462
## TP_c:surgeryyes              0.04881339 0.05190732
```

```
coeftest(fit)[,1:2] # NOrmal SEs
```

```
##                               Estimate Std. Error
## (Intercept)                 48.06256706 1.32705980
## ageyoung                    -7.08463957 2.70467241
```

```
## TP_c                         -0.05931106 0.03995003
## abdominal_distentionnone     -3.19217178 1.81259738
## abdominal_distentionsevere    6.09516031 2.25733249
## abdominal_distentionslight -3.52117451 1.83149172
## TP_c:surgeryyes               0.04881339 0.05193841
```

Interpretations of coefficient estimates: (Intercept): Predicted packed cell volume for an adult horse with an average total protein count and no surgery is 48.062 microliters. (TP_c): Controlling for age young, horses with no surgery show a decrease of .059 microliter in packed cell volume for every one unit increase in total protein on average. (Abdominal_distention_none:) Controlling for age, a horse with average total protein has a packed cell volume that is 3.19 microliters lower for horses with no abdominal distention compared to horses with abdominal distention. (Abdominal_distention_severe:) Controlling for age, a horse with average total protein has a packed cell volume that is 6.10 micoliters higher for horses with severe abdominal distention compared to horses without abdominal distention. (Abdominal_distention_slight:) Controlling for age, a horse with average total protein has a packed cell volume that is 3.52 microliters lower for horses with slight abdominal distention compared to horses without abdominal distention. (Age_young:) Controlling for Abdominal distention, a horse with average total protein has a packed cell volume that is 7.08 microliters lower for young horses compared to adult horses.

The null hypothesis for a Breusch-Pagan test states that the data is homoskedastic. Based on a p-value of .42, we fail to reject the null hypothesis concluding that the data homoskedastic. Normality was tested using ggplot to plot residuals, and was found to be an okay indicator of normality.

The assumption for homoskedasticity was confirmed with the Breush-Pagan test confirming equal variances. However, the Normal Standard Errors did not predictably increase when Robust Standard Errors were calculated. It is normal for standard errors to increase when testing for robust standard errors due to the lack of need for meeting assumptions. Overall, the standard errors did not change much, and some did increase with Robust standard errors, but those that decreased may have been becasue heteroskedasticity may make normal standard errors upward biased.

Based on the Adjusted R squared value given from the coeftest, the model explains 9.87% of the variation in the outcome.

```
fit<-lm(packed_cell_volume~age+TP_c+abdominal_distention+surgery:TP_c, data=horse1)
boot_dat<- sample_frac(horse, replace=T)

samp_distn<-replicate(5000, {
  boot_dat <- sample_frac(horse1, replace=T)
  fit <- lm(packed_cell_volume~age+TP_c+abdominal_distention+surgery:TP_c, data=boot_dat)
  coef(fit)
})

samp_distn %>% t %>% as.data.frame %>% summarize_all(sd)
```

```
##   (Intercept) ageyoung       TP_c abdominal_distentionnone
## 1    1.311489 1.940694 0.03531027                 1.778944
##   abdominal_distentionsevere abdominal_distentionslight TP_c:surgeryyes
## 1                   2.212293                   1.842452      0.05069666
```

```
fit<-lm(packed_cell_volume~age+TP_c+abdominal_distention+surgery:TP_c, data=horse1)
resids<-fit$residuals
fitted<-fit$fitted.values
resid_resamp<-replicate(5000,{
  new_resids<-sample(resids,replace=TRUE)
```

```
  horse1$new_y<-fitted+new_resids
  fit<-lm(new_y~age+TP_c+abdominal_distention+surgery:TP_c, data=horse1)
  coef(fit)
})

resid_resamp%>%t%>%as.data.frame%>%summarize_all(sd)
```

```
##   (Intercept) ageyoung       TP_c abdominal_distentionnone
## 1    1.290959 2.628944 0.03957525                 1.798316
##   abdominal_distentionsevere abdominal_distentionslight TP_c:surgeryyes
## 1                    2.22021                   1.770094      0.05143016
```

The Bootstrapped standard errors caluculated were unexpectedly lower than all values from the original standard errors.This is an unusual finding given the Bootstrapped standard errors do not require assumptions to be met and thus a higher SE is expected. In addition, some Robust standard errors also did not increase, but rather decreased from the original. These decreases may have been becasue some violations in the data were met. The expected p-values for both Bootstrapped and Robust standard errors would likely be higher than the p-value for original, because they do not require assumptions to be met.

```
log_horse<-horse%>%mutate(y=ifelse(outcome=="died",1,0))%>%drop_na(abdominal_distention)
head(log_horse)
```

```
## # A tibble: 6 x 16
##   surgery age   temp_of_extremi~ peripheral_pulse mucous_membrane
##   <chr>   <chr> <chr>            <chr>            <chr>
## 1 no      adult cool             reduced          NA
## 2 yes     adult NA               NA               pale_cyanotic
## 3 no      adult normal           normal           pale_pink
## 4 yes     young cold             normal           dark_cyanotic
## 5 yes     adult normal           normal           normal_pink
## 6 no      adult cool             absent           pale_pink
## # ... with 11 more variables: capillary_refill_time <chr>, pain <chr>,
## #   peristalsis <chr>, abdominal_distention <chr>, nasogastric_tube <chr>,
## #   rectal_exam_feces <chr>, abdomen <chr>, packed_cell_volume <dbl>,
## #   total_protein <dbl>, outcome <chr>, y <dbl>
```

```
fit2<-glm(y~surgery+abdominal_distention, family="binomial", data=log_horse)
coeftest(fit2)
```

```
##
## z test of coefficients:
##
##                             Estimate Std. Error z value  Pr(>|z|)
## (Intercept)                  0.45192    0.37345  1.2101 0.2262326
## surgeryyes                   0.11318    0.32993  0.3431 0.7315608
## abdominal_distentionnone    -1.87279    0.42502 -4.4064 1.051e-05 ***
## abdominal_distentionsevere   0.10837    0.47794  0.2267 0.8206252
## abdominal_distentionslight -1.49305    0.41330 -3.6125 0.0003032 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
exp(coef(fit2))
```

```
##                (Intercept)                  surgeryyes
##                  1.5713193                   1.1198354
##   abdominal_distentionnone abdominal_distentionsevere
##                  0.1536938                   1.1144582
## abdominal_distentionslight
##                  0.2246866
```

```r
prob<-predict(fit2, type="response")
pred<-ifelse(prob > .5,1,0)
table(prediction=pred, truth=log_horse$y)%>%addmargins
```

```
##           truth
## prediction   0   1 Sum
##        0    94  29 123
##        1    31  55  86
##        Sum 125  84 209
```

```r
#accuracy
(94+55)/209
```

```
## [1] 0.7129187
```

```r
#Specificity
94/125
```

```
## [1] 0.752
```

```r
#Sensitivity
55/84
```

```
## [1] 0.6547619
```

```r
odds<-function(p)p/(1-p)
p<-seq(0,1,by=.1)
cbind(p, odds=odds(p))%>%round(4)
```

```
##          p   odds
##  [1,] 0.0 0.0000
##  [2,] 0.1 0.1111
##  [3,] 0.2 0.2500
##  [4,] 0.3 0.4286
##  [5,] 0.4 0.6667
##  [6,] 0.5 1.0000
##  [7,] 0.6 1.5000
##  [8,] 0.7 2.3333
##  [9,] 0.8 4.0000
## [10,] 0.9 9.0000
## [11,] 1.0    Inf
```

```
logit<-function(p)log(odds(p))
cbind(p, odds=odds(p),logit=logit(p))%>%round(4)
```

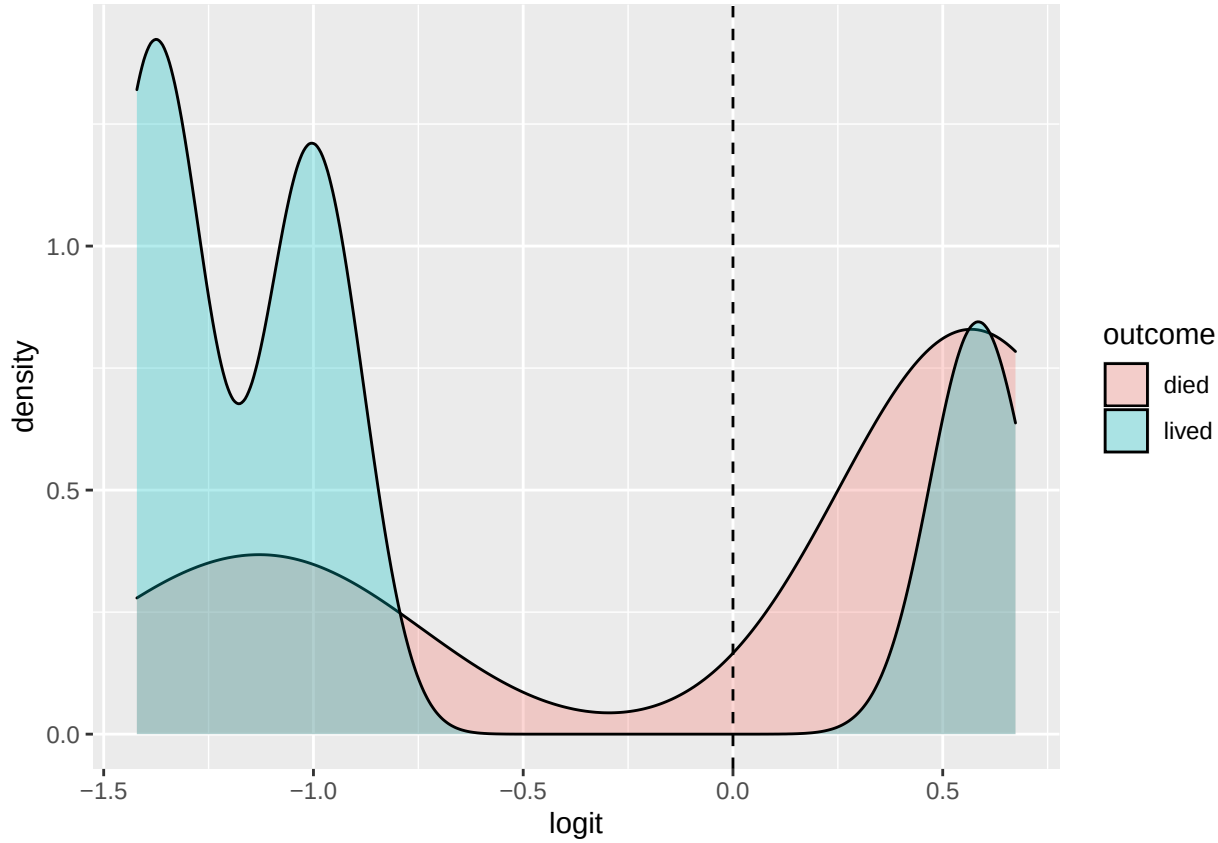
```
##         p   odds   logit
##  [1,] 0.0 0.0000    -Inf
##  [2,] 0.1 0.1111 -2.1972
##  [3,] 0.2 0.2500 -1.3863
##  [4,] 0.3 0.4286 -0.8473
##  [5,] 0.4 0.6667 -0.4055
##  [6,] 0.5 1.0000  0.0000
##  [7,] 0.6 1.5000  0.4055
##  [8,] 0.7 2.3333  0.8473
##  [9,] 0.8 4.0000  1.3863
## [10,] 0.9 9.0000  2.1972
## [11,] 1.0    Inf     Inf
```

```
fit3<-glm(y~surgery+abdominal_distention,data=log_horse,family=binomial(link="logit"))
coeftest(fit3)
```

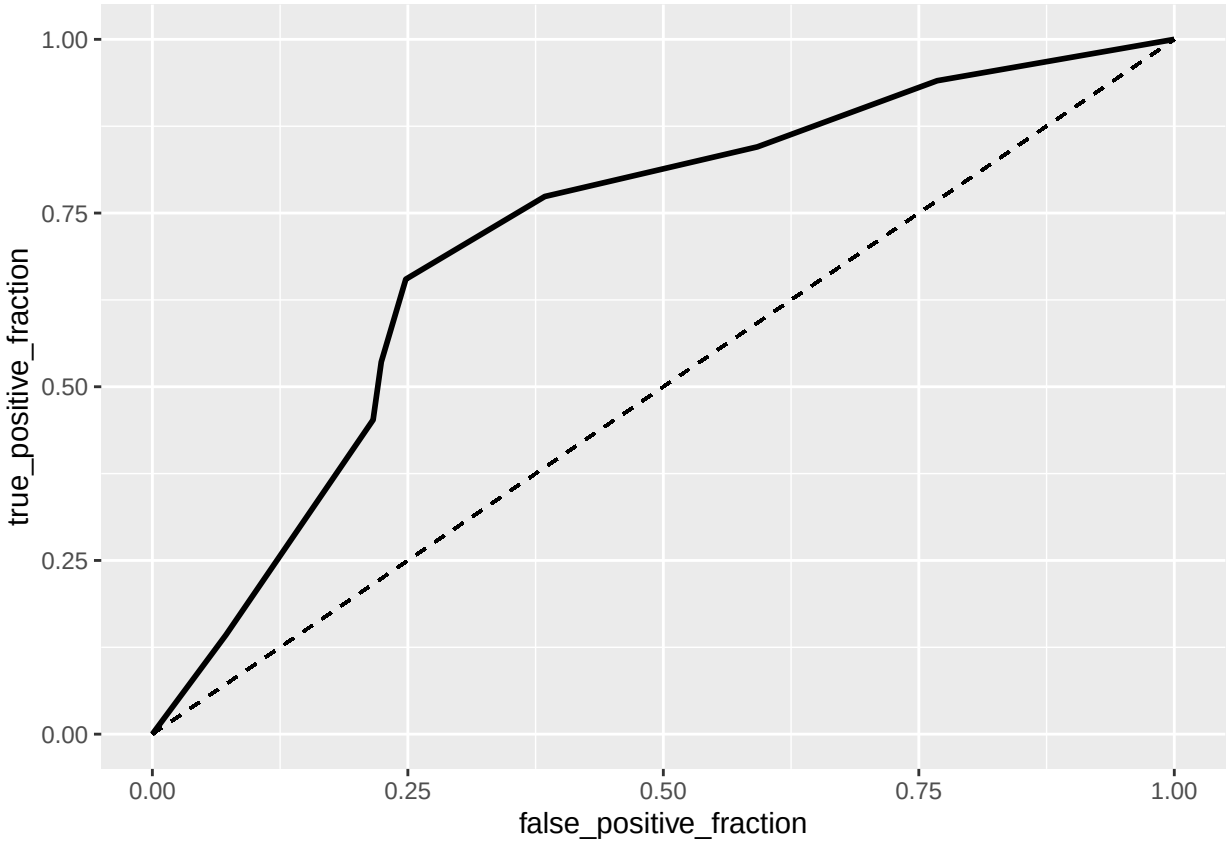
```
##
## z test of coefficients:
##
##                             Estimate Std. Error z value  Pr(>|z|)
## (Intercept)                  0.45192    0.37345  1.2101 0.2262326
## surgeryyes                   0.11318    0.32993  0.3431 0.7315608
## abdominal_distentionnone    -1.87279    0.42502 -4.4064 1.051e-05 ***
## abdominal_distentionsevere   0.10837    0.47794  0.2267 0.8206252
## abdominal_distentionslight -1.49305    0.41330 -3.6125 0.0003032 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
log_horse$logit<-predict(fit3)
log_horse$outcome<-factor(log_horse$outcome,levels=c("died","lived"))

ggplot(log_horse,aes(logit, fill=outcome))+geom_density(alpha=.3)+
  geom_vline(xintercept=0,lty=2)
```




```
library(plotROC)
prob2<-predict(fit2,type="response")
pred2<-ifelse(prob>.5,1,0)
ROCplot<-ggplot(log_horse)+geom_roc(aes(d=y,m=prob2), n.cuts=0)+
  geom_segment(aes(x=0,xend=1,y=0,yend=1),lty=2)
ROCplot
```




```
calc_auc(ROCplot)
```

```
##   PANEL group   AUC
## 1     1    -1 0.714
```

```
class_diag <- function(probs,truth){

  tab<-table(factor(probs>.5,levels=c("FALSE","TRUE")),truth)
  acc=sum(diag(tab))/sum(tab)
  sens=tab[2,2]/colSums(tab)[2]
  spec=tab[1,1]/colSums(tab)[1]
  ppv=tab[2,2]/rowSums(tab)[2]
  if(is.numeric(truth)==FALSE & is.logical(truth)==FALSE) truth<-as.numeric(truth)-1

  ord<-order(probs, decreasing=TRUE)
  probs <- probs[ord]; truth <- truth[ord]
  TPR=cumsum(truth)/max(1,sum(truth))
  FPR=cumsum(!truth)/max(1,sum(!truth))
  dup<-c(probs[-1]>=probs[-length(probs)], FALSE)
  TPR<-c(0,TPR[!dup],1); FPR<-c(0,FPR[!dup],1)
  n <- length(TPR)
  auc<- sum( ((TPR[-1]+TPR[-n])/2) * (FPR[-1]-FPR[-n]) )
  data.frame(acc,sens,spec,ppv,auc)
}
set.seed(1234)
```

```
k=10
data<-log_horse[sample(nrow(log_horse)),]
folds<-cut(seq(1:nrow(log_horse)),breaks=k,labels=F)
diags<-NULL
for(i in 1:k){
  train<-data[folds!=i,]
  test<-data[folds==i,]
  truth<-test$y
  fit4<-glm(y~surgery+abdominal_distention, family="binomial", data=train)
  probs<-predict(fit4,newdata = test,type="response")
  diags<-rbind(diags,class_diag(probs,truth))
}
summarize_all(diags,mean)
```

```
##        acc      sens    spec       ppv       auc
## 1 0.712619 0.6353824 0.75219 0.6193326 0.6814749
```

Interpretation of Coefficient Estimates: (Intercept:) The odds of death for no surgery and moderate abdominal distention is 1.571. (Surgeryyes:) With moderate abdominal distention, odds of death for horses that have surgery is 1.119 times odds for no surgery. (Abdominal_distention_none:) Odds of death for a horse with no surgery and no abdominal distention is .1537 time odds of moderate abdominal distention (84.63% less). (Abdominal_distention_severe:) Odds of death for a horse with no surgery and severe abdominal distention is 1.114 times odds of moderate abdominal distention. (Abdominal_distention_slight:) Odds of death for a horse with no surgery and slight abdominal distention is .2246 times odds of moderate abdominal distention (77.53% less).

A test was run to see how well the model could predict whether or not a horse died. The model had an accuracy of .7129, representing the proportion of correctly classified deaths. The sensitivity, or true positive rate was determined to be .6547, while the specificity, or true negative rate was calculated at .752.

The ROC curve tests the true positive rate and false positive rate for predictions from the model. As seen in the ROC curve, it does not form a perfect square which would have indicated an AUC of one, or a TPR of 100%. Instead, the curve is close to the diagonal threshold, which would indicate that the model is an extremely bad predictor. While the AUC was not this bad, it was considered "fair" with a calculated value of .714. An AUC of this value indicates that the model is fairly predicting the outcome of horses.

A 10-fold cross validation was performed to see how well the model could generalize to fit the dataset. The average out-of-sample AUC was found to be "poor" with .681, average accuracy was .712, average sensitivity was .635, and average specificity was

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 3.0-2
```

```
y<-as.matrix(log_horse$surgery)
x<-model.matrix(surgery~.,data=log_horse)[,-1]
x<-scale(x)
head(x)%>%glimpse()
```

```
##  num [1:6, 1:46] -0.277 -0.277 -0.277 3.588 -0.277 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:6] "1" "2" "3" "4" ...
##   ..$ : chr [1:46] "ageyoung" "temp_of_extremitiescool" "temp_of_extremitiesNA" "temp_of_extremitiesr
```

```
cv<-cv.glmnet(x,y,family="binomial")
lasso<-glmnet(x,y,family="binomial",lambda=cv$lambda.1se)
coef(lasso)
```

```
## 47 x 1 sparse Matrix of class "dgCMatrix"
##                                          s0
## (Intercept)                       1.20340579
## ageyoung                          0.12448823
## temp_of_extremitiescool          -0.15723023
## temp_of_extremitiesNA             0.28408046
## temp_of_extremitiesnormal         .
## temp_of_extremitieswarm          -0.18273833
## peripheral_pulseincreased        -0.42766578
## peripheral_pulseNA                .
## peripheral_pulsenormal           -0.01119876
## peripheral_pulsereduced           .
## mucous_membranebright_red         0.10153509
## mucous_membranedark_cyanotic      0.37151257
## mucous_membraneNA                 .
## mucous_membranenormal_pink       -0.10106804
## mucous_membranepale_cyanotic      0.27389494
## mucous_membranepale_pink          0.06716799
## capillary_refill_timeless_3_sec   0.07719376
## capillary_refill_timemore_3_sec   .
## capillary_refill_timeNA           0.00285114
## paindepressed                    -0.18950163
## painextreme_pain                  .
## painmild_pain                    -0.12741253
## painsevere_pain                   0.01535363
## peristalsishypermotile           -0.07070429
## peristalsishypomotile             0.12502675
## peristalsisNA                    -0.16761720
## peristalsisnormal                 0.32693794
## abdominal_distentionnone         90.60441010
## abdominal_distentionsevere       -4.07437760
## abdominal_distentionslight       70.28434886
## nasogastric_tubenone             -0.07428491
## nasogastric_tubesignificant       0.03394785
## nasogastric_tubeslight            0.10081939
## rectal_exam_fecesdecreased       -0.06401515
## rectal_exam_fecesincreased        0.07024860
## rectal_exam_fecesNA              -0.03197841
## rectal_exam_fecesnormal          -0.14022922
```

```
## abdomendistend_small             0.23060073
## abdomenfirm                     -0.35883060
## abdomenNA                       -0.01025819
## abdomennormal                   -0.22710651
## abdomenother                    -0.34110594
## packed_cell_volume              -0.21076517
## total_protein                    0.03739615
## outcomelived                     .
## y                                .
## logit                           92.73969426
```

```
horse2<-log_horse%>%mutate(Warm_temp=ifelse(temp_of_extremities=="warm",1,0))%>%
  mutate(Normal_pink=ifelse(mucous_membrane=="normal_pink",1,0))%>%
  mutate(Hypermotile=ifelse(peristalsis=="hypermotile",1,0))%>%
  mutate(Distend_small=ifelse(abdomen=="distend_small",1,0))%>%
  mutate(Firm=ifelse(abdomen=="firm",1,0))%>%
  mutate(Ab_other=ifelse(abdomen=="other",1,0))%>%mutate(z=ifelse(surgery=="yes",1,0))

set.seed(1234)
k=10
data <- horse2[sample(nrow(horse2)),]
folds <- cut(seq(1:nrow(horse2)),breaks=k,labels=F)
diags<-NULL
for(i in 1:k){
  train <- data[folds!=i,]
  test <- data[folds==i,]
  truth <- test$z
  fit5 <- glm(z~Warm_temp+Normal_pink+Hypermotile+Distend_small+Firm+Ab_other,data=train, family="binom
  probs <- predict(fit5, newdata=test, type="response")
  diags<-rbind(diags,class_diag(probs,truth))
}
diags%>%summarize_all(mean)
```

```
##         acc      sens      spec       ppv       auc
## 1 0.7171429 0.8763487 0.5151407 0.7094968 0.7580151
```

The out-of-sample AUC from this model is .758 while the out-of-sample AUC from the logisitic regression model was .681. This LASSO model provided greater accuracy using only the predictors that showed the most importance. The variables retained and run through the cross-validation were: temperature of extremeties-warm, mucous membrane-normal pink, peristalsis-hypermotile, abdomen-distended small, abdomen-firm, and abdomen-other.