Project 2

##Audrey Malcolm aem3886

##Introduction

```
data(package=.packages(all.available=TRUE))
```

```
## Warning in data(package = .packages(all.available = TRUE)): datasets have
## been moved from package 'base' to package 'datasets'
```

```
## Warning in data(package = .packages(all.available = TRUE)): datasets have
## been moved from package 'stats' to package 'datasets'
```

```
library(datasets)
library(tidyverse)
```

```
## -- Attaching packages ---------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

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

```
infert<-as.data.frame(infert)
?infert
```

```
## starting httpd help server ...
```

```
##  done
```

The infertility dataset looks at infertility rates in women after having a spontaneous or an induced abortion as compared to a control group of women. The dataset looks at other factors like a woman's education, age, number of prior pregnancies, and number of prior abortions. Parity is the number of pregnancies that have been brought to gestation age and is the fertility measure. Women in case 1 were experiencing infertility issues while women in case 0 were not experiencing infertility issues.

##Manova, Anova, T Tests

```
man1<-manova(cbind(spontaneous,induced,age, parity)~case, data=infert)
summary(man1)
```

```
##            Df  Pillai approx F num Df den Df    Pr(>F)
## case        1 0.20608   15.769      4    243 1.732e-11 ***
## Residuals 246
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
##  Response spontaneous :
##              Df  Sum Sq Mean Sq F value   Pr(>F)
## case          1  17.561 17.5614  37.572 3.48e-09 ***
## Residuals   246 114.983  0.4674
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response induced :
##              Df  Sum Sq Mean Sq F value Pr(>F)
## case          1   0.039 0.03944  0.0721 0.7886
## Residuals   246 134.654 0.54737
##
##  Response age :
##              Df Sum Sq Mean Sq F value Pr(>F)
## case          1    0.1  0.0849  0.0031 0.9559
## Residuals   246 6811.9 27.6907
##
##  Response parity :
##              Df Sum Sq Mean Sq F value Pr(>F)
## case          1   0.03 0.03072  0.0195  0.889
## Residuals   246 386.84 1.57250
```

```
pairwise.t.test(infert$case, infert$spontaneous, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  infert$case and infert$spontaneous
##
##   0       1
## 1 0.00027 -
## 2 4e-08   0.01162
##
## P value adjustment method: none
```

```
.05/16
```

```
## [1] 0.003125
```

A MANOVA test was conducted in order to test the effect of infertility on number of spontaneous and induced abortions, parity, and age of the women. Because the infertility p value is significant, the null hypothesis is rejected. So, for at least one dependent variable (spontaneous abortions, induced abortions, parity, and age), the infertility mean is significantly different with p<1.353e-08. From there, the ANOVA test for each dependent variable was conducted. The p value for spontaneous abortions is significant with p=3.48e-09. However, the mean age, parity, and induced abortions across parity were not significant. Therefore, for spontaneous abortions, at least of mean differs, but for parity, age and, induced abortions, the mean does not vary across case. If unadjusted, the probability of a type 1 error is 0.05. The Bonferroni method was used to control Type 1 error rates. Because 1 manova, 3 anovas, and (4x3)=12 t tests were conducted, the adjusted p value is 0.05/16 = 0.0031. Still, with the adjusted p value, the manova and anova for induced and spontaneous abortions are still significant and the previous analyses stand. In order to determine which groups within the number of induced and spontaneous abortions were different, a pairwise t test was conducted. It shows that women with 1 or spontaneous abortion face infertility issues in their second pregnancy.

##Randomization Test <- make graph for null statistic

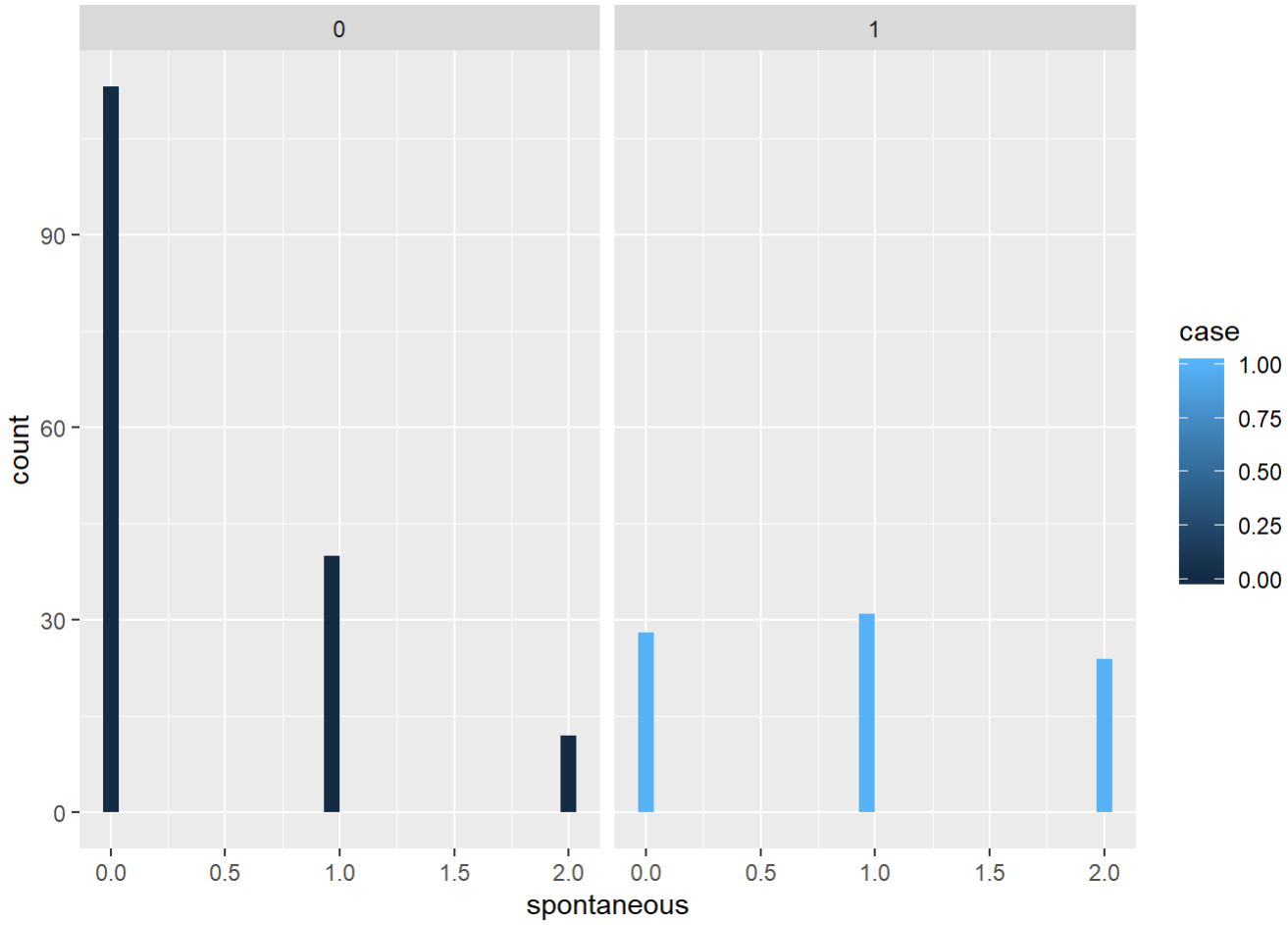
```
infert%>%group_by(case)%>%summarize(mean=mean(spontaneous))%>%summarize(diff(mean))
```

```
## # A tibble: 1 x 1
##   `diff(mean)`
##          <dbl>
## 1        0.564
```

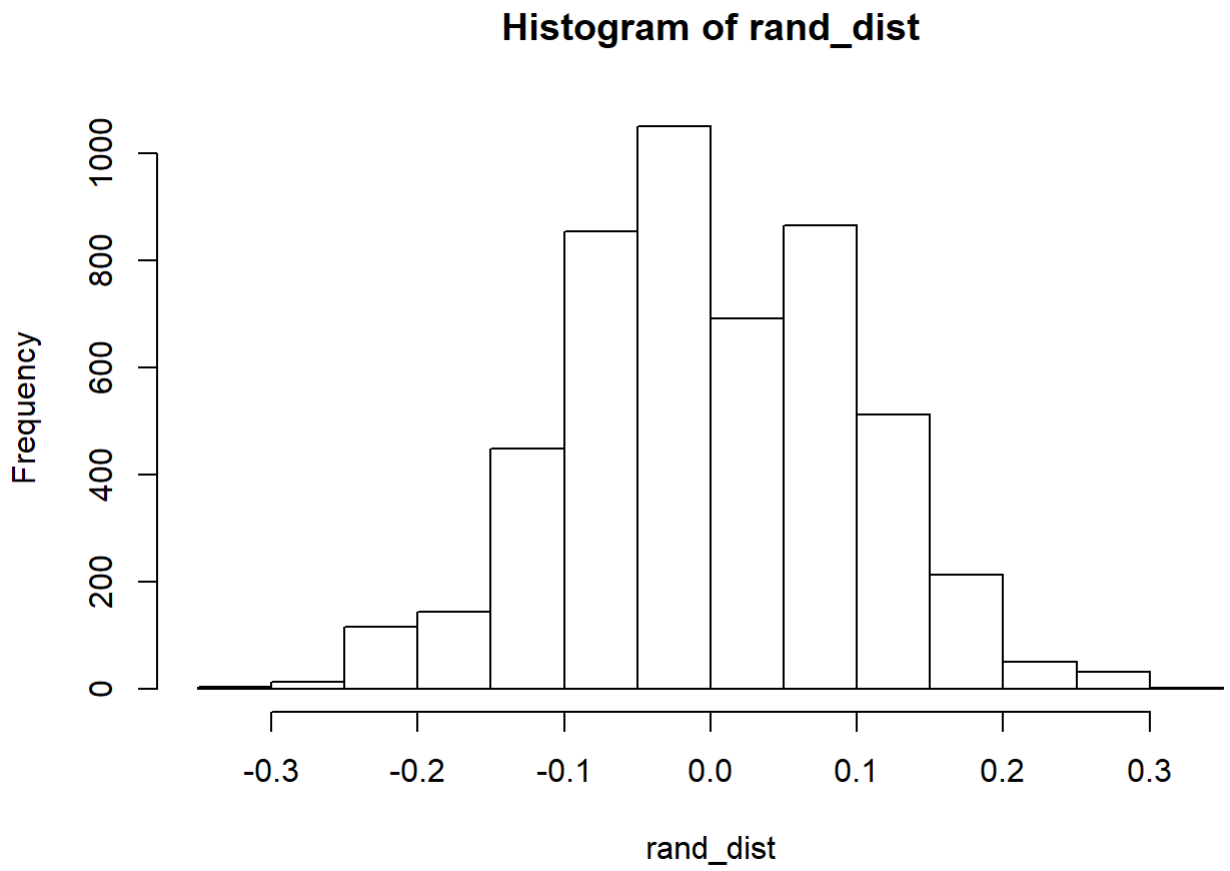
```
rand_dist<-vector()
for(i in 1:5000){
  new<-data.frame(spontaneous=sample(infert$spontaneous),case=infert$case)
  rand_dist[i]<-mean(new[new$case=="0",]$spontaneous)-mean(new[new$case=="1",]$spontaneous)
}

ggplot(infert,aes(spontaneous,fill=case))+geom_histogram()+facet_wrap(~case, ncol=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
hist(rand_dist)
```




```
mean(rand_dist>0.5639284)*2
```

```
## [1] 0
```

H0: Mean number of spontaneous abortions is the same for women with fertility issues and without. H1: The mean number of spontaneous abortions is different between women with fertility issues and without.

The p value of the randomized t test is 0, which is significant because p<0.0031. This makes sense because if a woman naturally failed to have a pregnancy to term in the past, she is more likely to have infertility issues with future pregnancies due to underlying health problems.

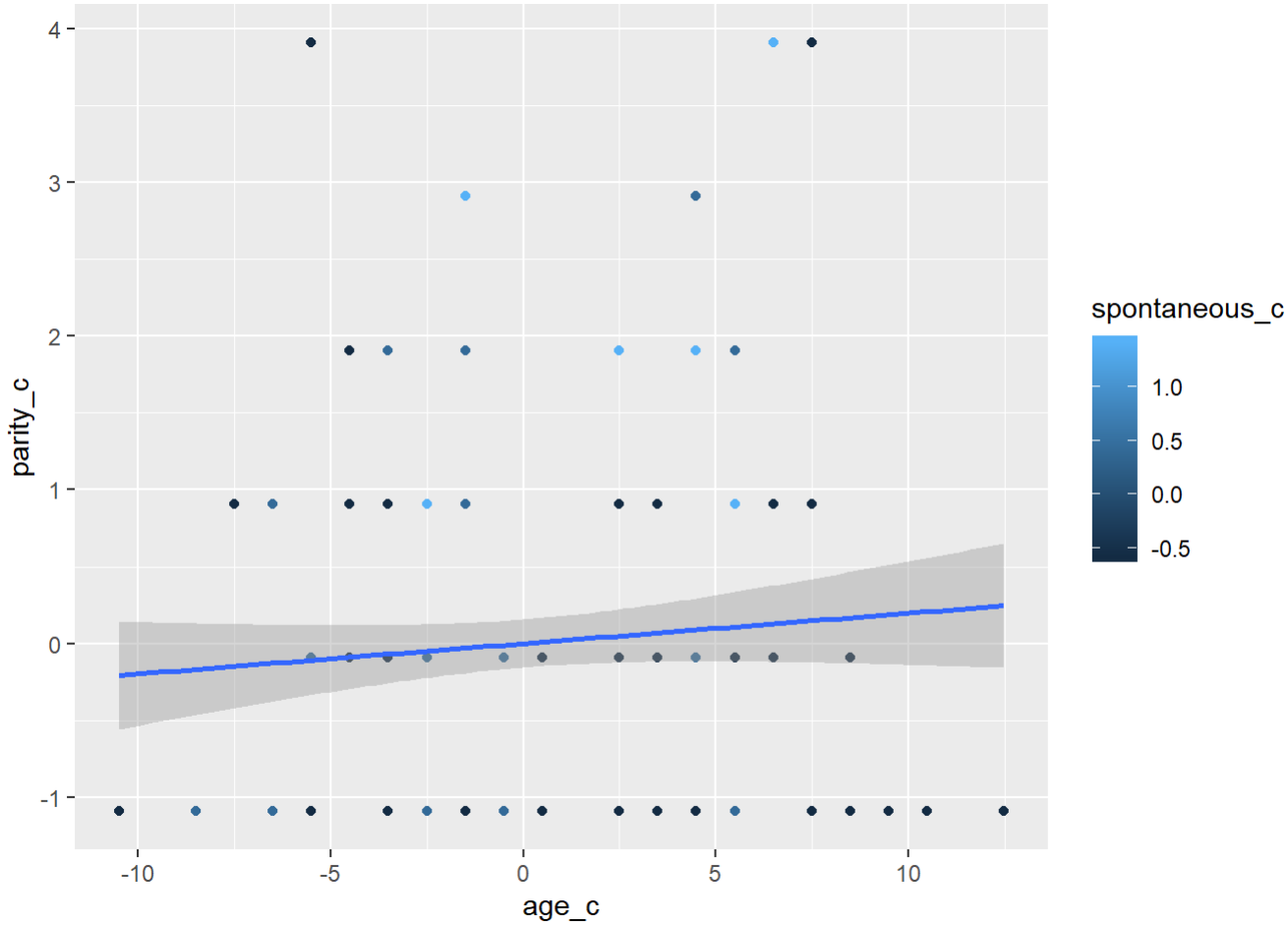
##Continuous Linear Regression

```
infert_ce<-infert%>%mutate(induced_c=induced-mean(induced,rm.na=T),spontaneous_c=spontaneous-mea
n(spontaneous,rm.na=T), parity_c=parity-mean(parity,rm.na=T),age_c=age-mean(age,rm.na=T))

fit<-lm(parity_c~spontaneous_c*age_c, data=infert_ce)
summary(fit)
```

```
##
## Call:
## lm(formula = parity_c ~ spontaneous_c * age_c, data = infert_ce)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8839 -0.8191 -0.3114  0.2493  4.2493
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.01177    0.07528   0.156   0.8759
## spontaneous_c        0.54109    0.10304   5.251 3.29e-07 ***
## age_c                0.02861    0.01442   1.984   0.0484 *
## spontaneous_c:age_c  0.03644    0.02082   1.750   0.0814 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.181 on 244 degrees of freedom
## Multiple R-squared:  0.1206, Adjusted R-squared:  0.1098
## F-statistic: 11.16 on 3 and 244 DF,  p-value: 6.919e-07
```

```
ggplot(infert_ce, aes(age_c,parity_c))+geom_point(aes(color=spontaneous_c))+geom_smooth(method="lm")
```



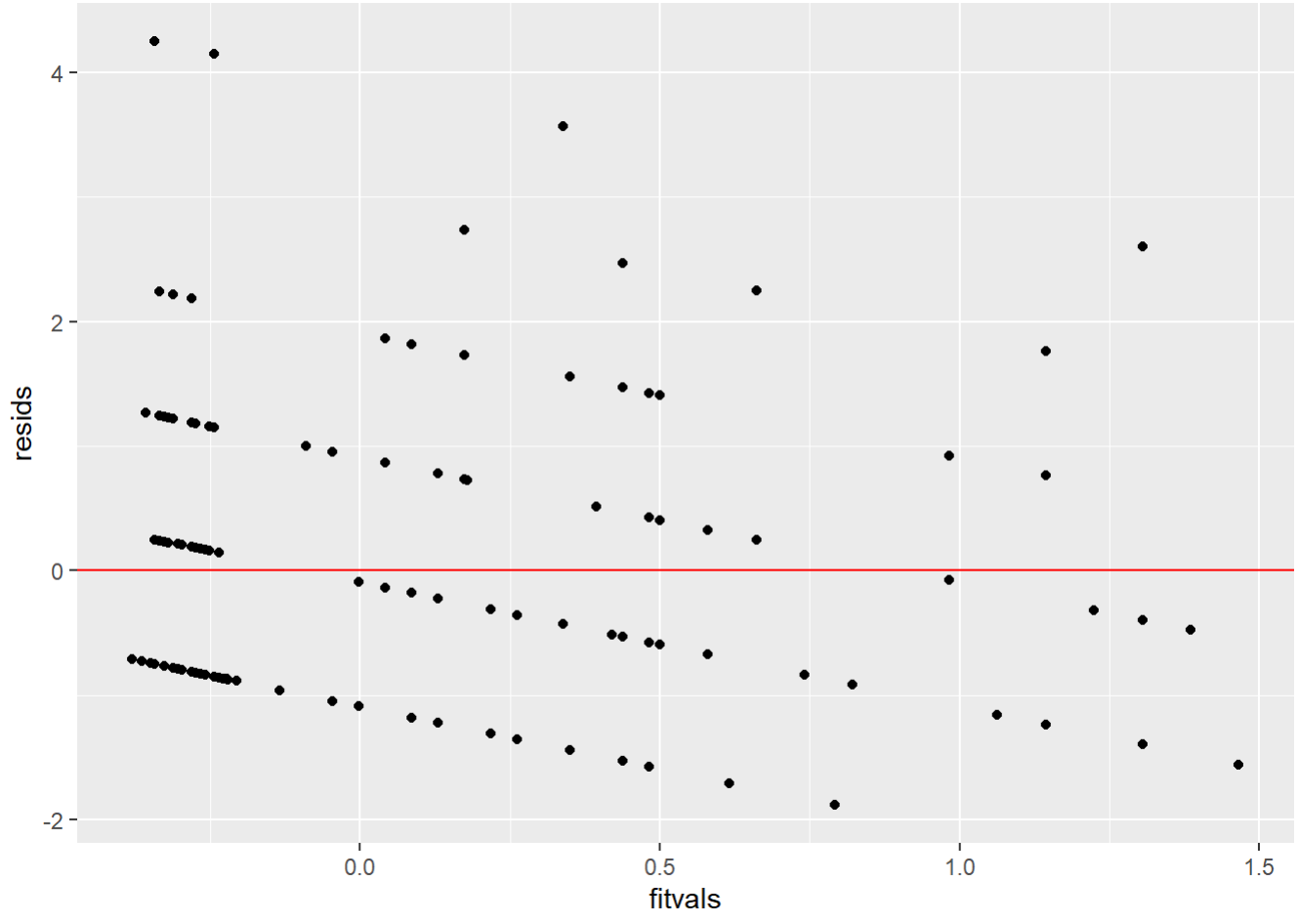
```
#Testing for linearity, normality, homoskedsaticity
library(lmtest)
```

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

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

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

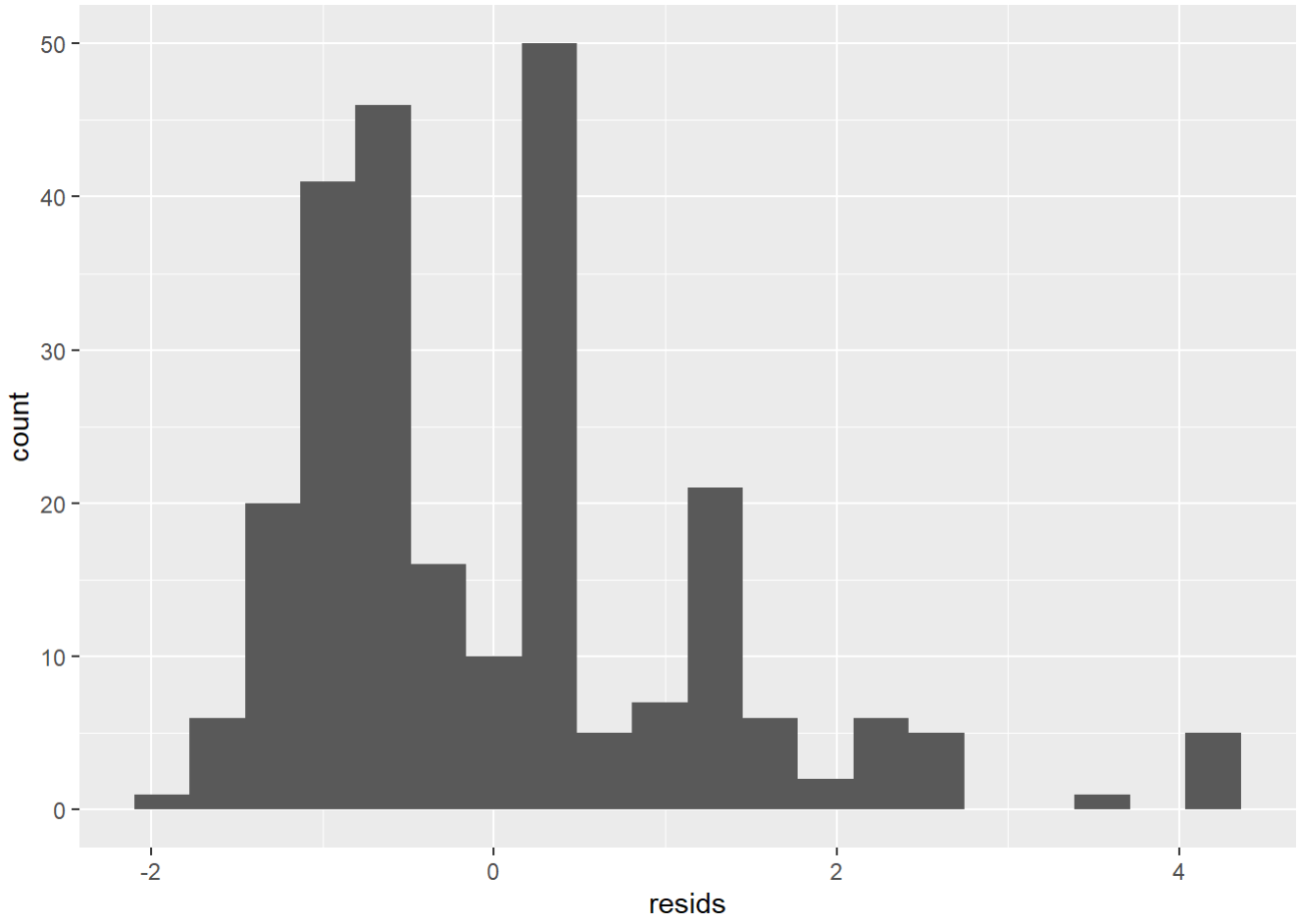
```
resids<-fit$residuals
fitvals<-fit$fitted.values
ggplot()+geom_point(aes(fitvals,resids))+geom_hline(yintercept=0, color='red')
```




```
bptest(fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 3.3784, df = 3, p-value = 0.3369
```

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




```
shapiro.test(resids)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resids
## W = 0.88128, p-value = 5.306e-13
```

```
#Recompute regression
??vcovHC
library(sandwich)
coeftest(fit,vcov=vcovHC(fit))
```

```
##
## t test of coefficients:
##
##                     Estimate Std. Error t value  Pr(>|t|)
## (Intercept)         0.011765   0.076635  0.1535   0.87811
## spontaneous_c       0.541088   0.109452  4.9436 1.427e-06 ***
## age_c               0.028614   0.016015  1.7867   0.07523 .
## spontaneous_c:age_c 0.036437   0.024466  1.4893   0.13771
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(sum((infert$parity-mean(infert$parity))^2)-sum(fit$residuals^2))/sum((infert$parity-mean(infert
$parity))^2)
```

```
## [1] 0.1206122
```

```
fit2<-lm(parity_c~spontaneous_c+age_c, data=infert_ce)
summary(fit2)
```

```
##
## Call:
## lm(formula = parity_c ~ spontaneous_c + age_c, data = infert_ce)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6527 -0.8674 -0.2336  0.3486  4.3683
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.343e-15  7.530e-02   0.000   1.0000
## spontaneous_c 5.495e-01  1.034e-01   5.316 2.38e-07 ***
## age_c         2.619e-02  1.442e-02   1.817   0.0705 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.186 on 245 degrees of freedom
## Multiple R-squared:  0.1096, Adjusted R-squared:  0.1023
## F-statistic: 15.08 on 2 and 245 DF,  p-value: 6.692e-07
```

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

```
## Analysis of Variance Table
##
## Model 1: parity_c ~ spontaneous_c * age_c
## Model 2: parity_c ~ spontaneous_c + age_c
##   Res.Df    RSS Df Sum of Sq Pr(>Chi)
## 1    244 340.21
## 2    245 344.48 -1   -4.2691  0.08015 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient for spontaneous abortions is the difference between the predicted values of parity when age=0. So, as parity increases, the number of spontaneous abortions increases slightly by 0.54109 when holding all other variables constant. Similarly, as parity increases, age increases slightly by 0.02861. The interaction of age and spontaneous abortions increases slightly with parity as well by 0.03644.

Based on the assumptions test for the regression, the data is not normally distributed and is homoscedastic.

Comparing the original regression standard errors to the robust standard errors, the error is slightly lower for the robust standard errors but not much different from the original regression. The spontaneous abortion and age significantly interact with the parity number in both the new and old linear regression.

The proportion of variance in parity that is explained by the overall model is 0.1206122.

The likelihood ratio test for the regression without interactions compared to the interaction model is not significant, so the interaction model is better at predicting the number of kids a woman will have.

##Bootstrapped SEs

```
boot_dat<-replicate(5000,{
  boot_dat<-infert_ce[sample(nrow(infert_ce),replace=TRUE),]
  fit<-lm(parity_c~spontaneous_c*age_c, data=boot_dat)
  coef(fit)
})

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

```
##   (Intercept) spontaneous_c      age_c spontaneous_c:age_c
## 1  0.07560647     0.1083194 0.01568314           0.0239632
```

The bootstrapped standard errors are about the same as the initial regression model with interaction.

##Binary Logistic Regression

```
fit3<-glm(case~spontaneous+induced, data=infert, family = binomial(link="logit"))
infert$prob<-predict(fit3,type="response")
summary(fit3)
```

```
##
## Call:
## glm(formula = case ~ spontaneous + induced, family = binomial(link = "logit"),
##     data = infert)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6678  -0.8360  -0.5772   0.9030   1.9362
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.7079     0.2677  -6.380 1.78e-10 ***
## spontaneous   1.1972     0.2116   5.657 1.54e-08 ***
## induced       0.4181     0.2056   2.033    0.042 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 316.17  on 247  degrees of freedom
## Residual deviance: 279.61  on 245  degrees of freedom
## AIC: 285.61
##
## Number of Fisher Scoring iterations: 4
```

```
table(predicted=as.numeric(infert$prob>.5),true=infert$case)%>%addmargins()
```

```
##          true
## predicted   0   1 Sum
##       0   149  55 204
##       1    16  28  44
##       Sum 165  83 248
```

```
#accuracy, sensitivity, specificity, recall
(28+149)/248
```

```
## [1] 0.7137097
```

```
28/83
```

```
## [1] 0.3373494
```

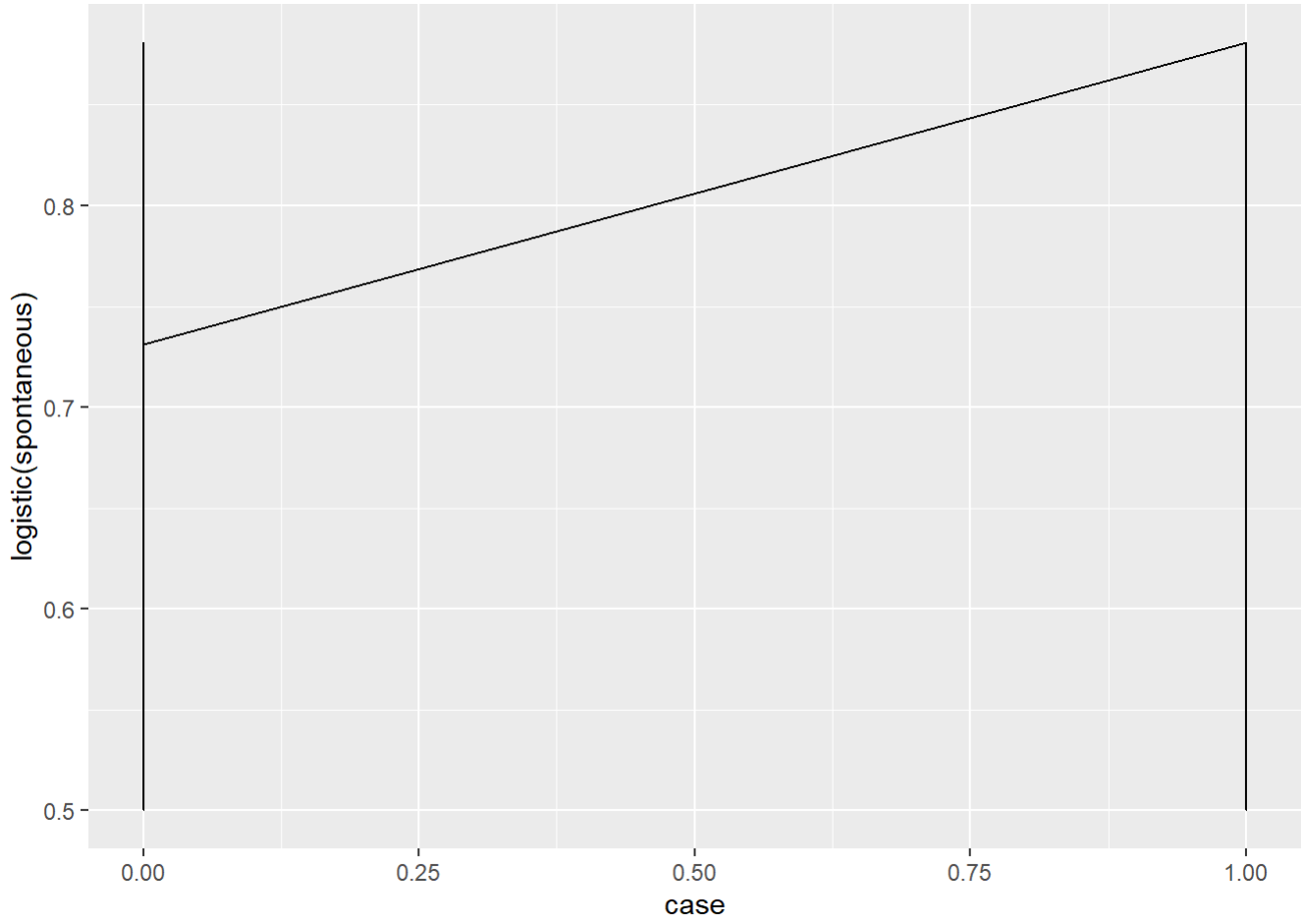
```
149/165
```

```
## [1] 0.9030303
```

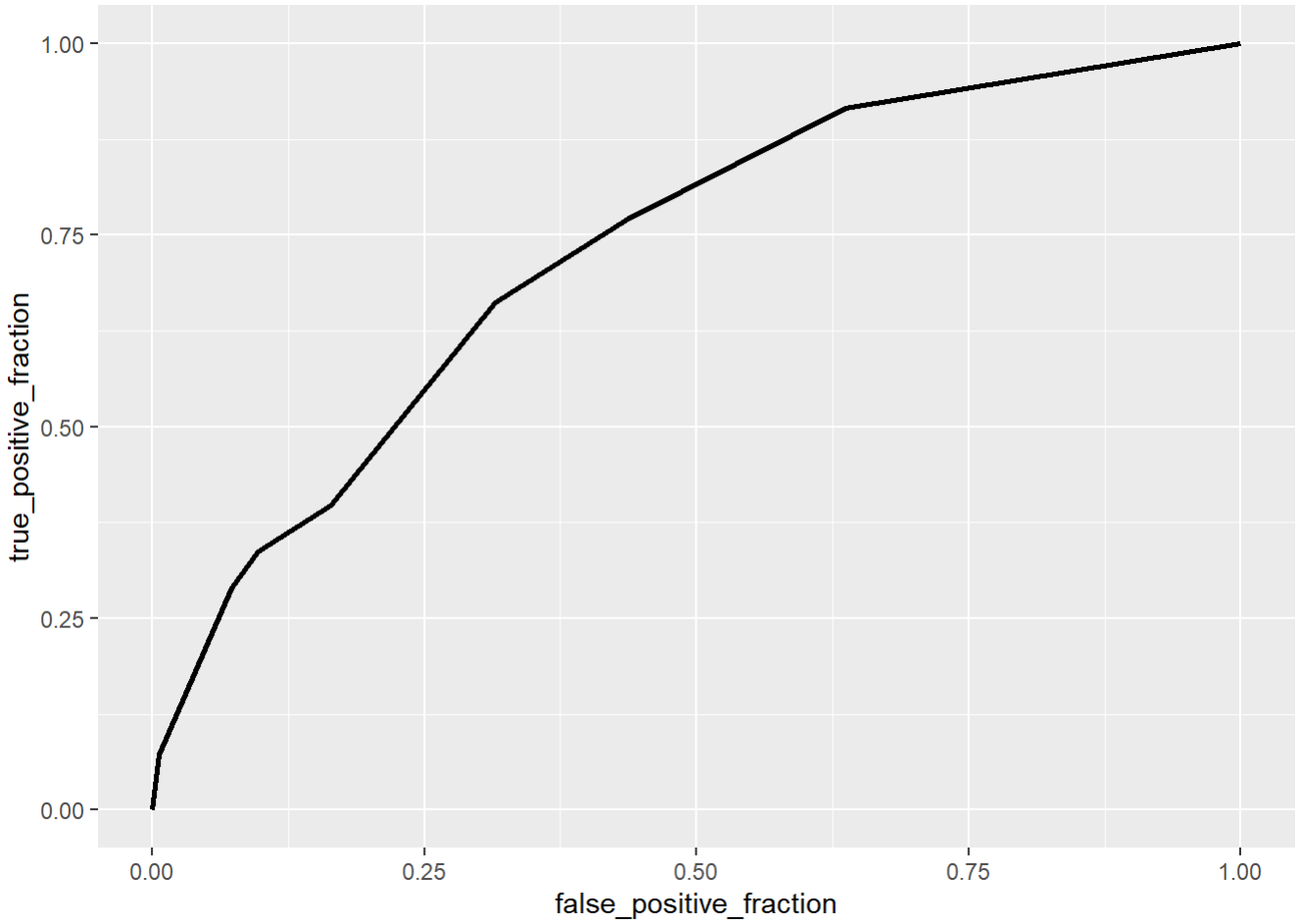
```
28/44
```

```
## [1] 0.6363636
```

```
#Graph
logistic<-function(x){exp(x)/(1+exp(x))}
infert_l<-infert%>%mutate(logistic(spontaneous))
ggplot(data=infert)+geom_line(aes(x=case,y=logistic(spontaneous)))
```




```
#ROC
library(plotROC)
ROCplot<-ggplot(infert)+geom_roc(aes(d=case,m=prob),n.cuts=0)
ROCplot
```




```
calc_auc(ROCplot)
```

```
##   PANEL group       AUC
## 1     1    -1 0.7285506
```

```
#10-fold CV
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
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
 #CALCULATE EXACT AUC
 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<-infert[sample(nrow(infert)),]
folds<-cut(seq(1:nrow(infert)),breaks=k,labels=F)
diags<-NULL
for(i in 1:k){
  train<-data[folds!=i,]
  test<-data[folds==i,]
  truth<-test$case
  fit3<-glm(case~spontaneous+induced, data=infert, family = "binomial")
  probs<-predict(fit3,newdata=test,type="response")
  diags<-rbind(diags,class_diag(probs,truth))
}
apply(diags,2,mean)
```

```
##       acc      sens      spec       ppv       auc
## 0.7140000 0.3276365 0.9034115 0.6509524 0.7354819
```

Controlling for induced abortions, there is a significant difference between the number of spontaneous abortions and fertility. Similarly, there is a significant difference between the number of induced abortions and fertility when controlling for spontaneous abortions.

Looking at the confusion matrix, the accuracy of prediction is 0.713. The sensitivity is 0.337 and the specificity is 0.903. The recall is 0.636.

The AUC calculated from the ROC plot is 0.7285506 meaning that, which means the model is an ok predictor of whether a woman will have infertility issues.

Calculating the 10-fold CV, the AUC is 0.735, which is slightly better than the AUC from the ROC plot but ultimately the same. The average out of sample accuracy is 0.714, the sensitivity is 0.328 and the recall is 0.651.

##LASSO

```
library(glmnet)
```

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

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

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

```
## Loaded glmnet 3.0-1
```

```
y<-as.matrix(infert$case)
x<-infert%>%select_if(is.numeric)%>%dplyr::select(-case,-prob)%>%mutate_all(scale)%>%as.matrix
cv<-cv.glmnet(x,y)
lasso1<-glmnet(x,y,lambda=cv$lambda.1se)
coef(lasso1)
```

```
## 7 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept)     0.33467742
## age             .
## parity         -0.04467257
## induced         0.05001560
## spontaneous     0.17504826
## stratum         .
## pooled.stratum -0.01094999
```

```
set.seed(1234)
k=10
data<-infert[sample(nrow(infert)),]
folds<-cut(seq(1:nrow(infert)),breaks=k,labels=F)
diags<-NULL
for(i in 1:k){
  train<-data[folds!=i,]
  test<-data[folds==i,]
  truth<-test$case
  fit4<-glm(case~spontaneous+parity, data=infert, family = "binomial")
  probs<-predict(fit4,newdata=test,type="response")
  diags<-rbind(diags,class_diag(probs,truth))
}
diags%>%summarize_all(mean)
```

```
##     acc      sens      spec      ppv       auc
## 1 0.746 0.5340193 0.8634247 0.659881 0.6945263
```

Parity and spontaneous abortions are the most important predictors if infertility. The 10-fold CV for parity and spontaneous shows the AUC is 0.695. However, the accuracy is slightly improved at 0.746. This model is a slightly better predictor of infertility than the model of induced and spontaneous abortions, but ultimately both are very similar.