Mice Data Analysis

Importing the Data

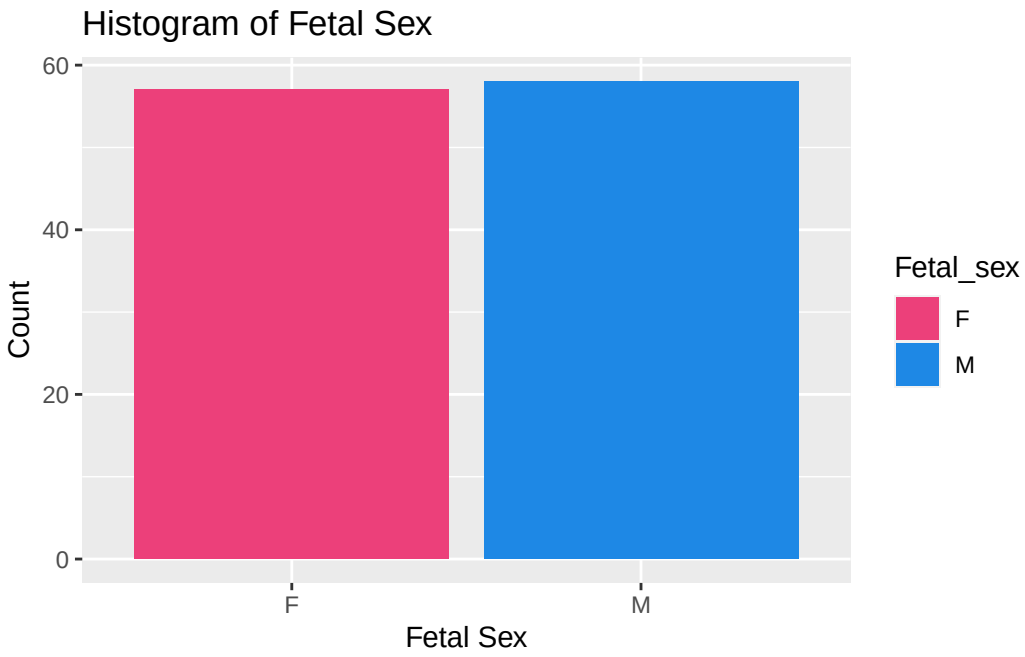
```
fetus_full <- read.csv("data/fetus-cleaned.csv", header=T)

fetus <- fetus_full %>%
  filter(Fetus_genotype != "resorp") %>%
  na.omit() # One empty row. Not sure why...
```

We drop all resorp fetuses. This is because they do not provide information. Of course, the data are not missing at random in this case. However, we can conduct an entirely separate analysis of resorp vs living fetuses later on if that's required.

Assessing Normality of Features and Dependent Variables

First we assess the proportion of male v. female fetuses to make sure our data are not skewed.

```
# Define colors for each sex
color_female <- "#ec407a"  # pink
color_male <- "#1e88e5"    # blue

ggplot(fetus, aes(x = Fetal_sex, fill = Fetal_sex)) +
  geom_bar() +
  scale_fill_manual(values = c(color_female, color_male)) +
  xlab("Fetal Sex") +
  ylab("Count") +
  ggtitle("Histogram of Fetal Sex")
```

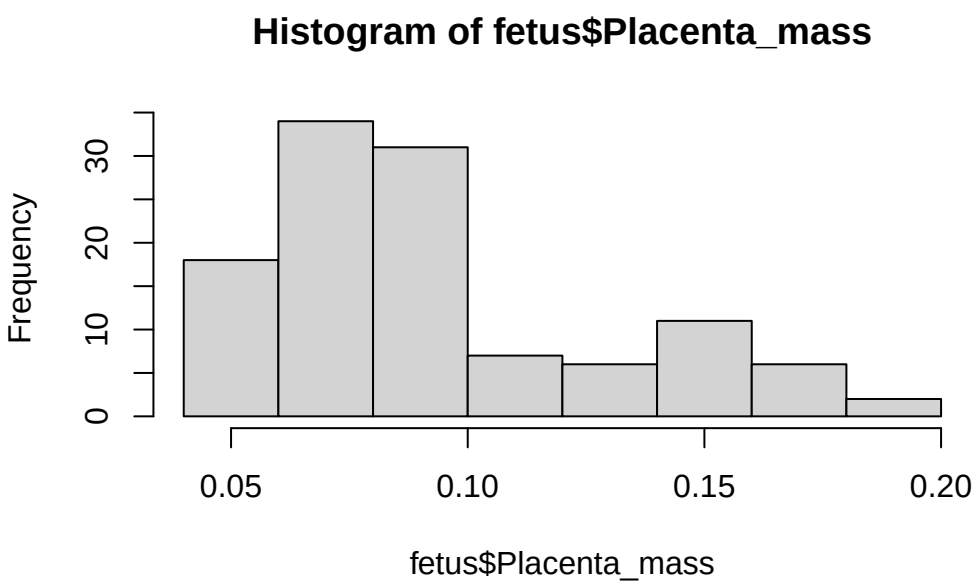



```
mean(fetus$isFemale)
```

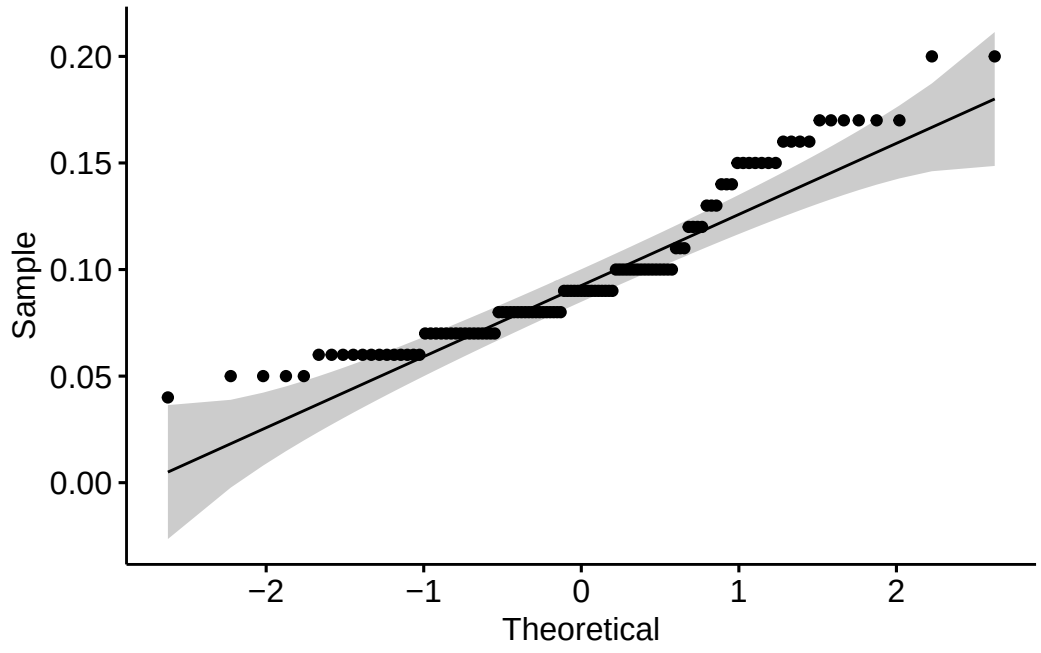
```
[1] 0.4956522
```

Approximately 50% of surviving fetuses are female, so no worries about composition of the data.

```
hist(fetus$Placenta_mass)
```




```
ggqqplot(fetus$Placenta_mass)
```

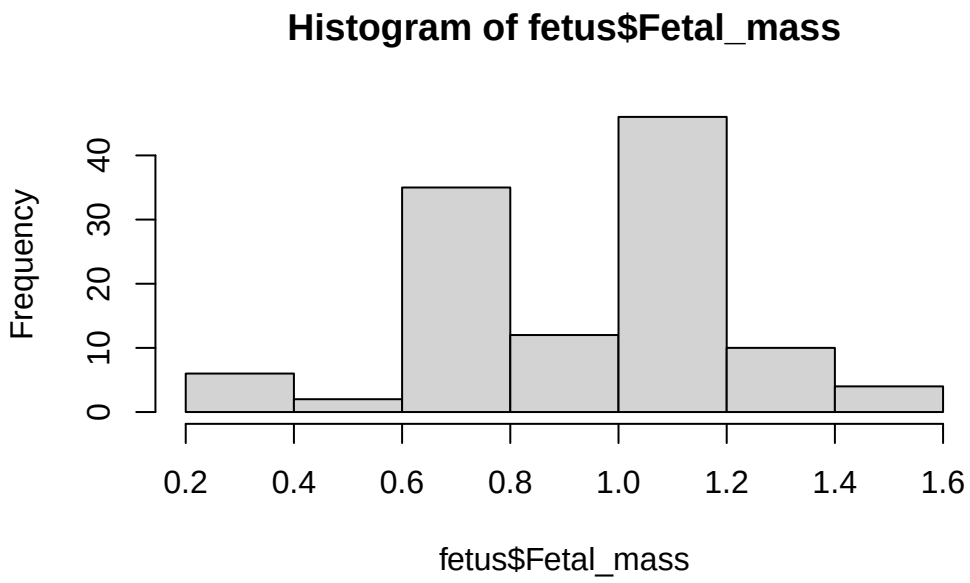


```
shapiro.test(fetus$Placenta_mass)
```

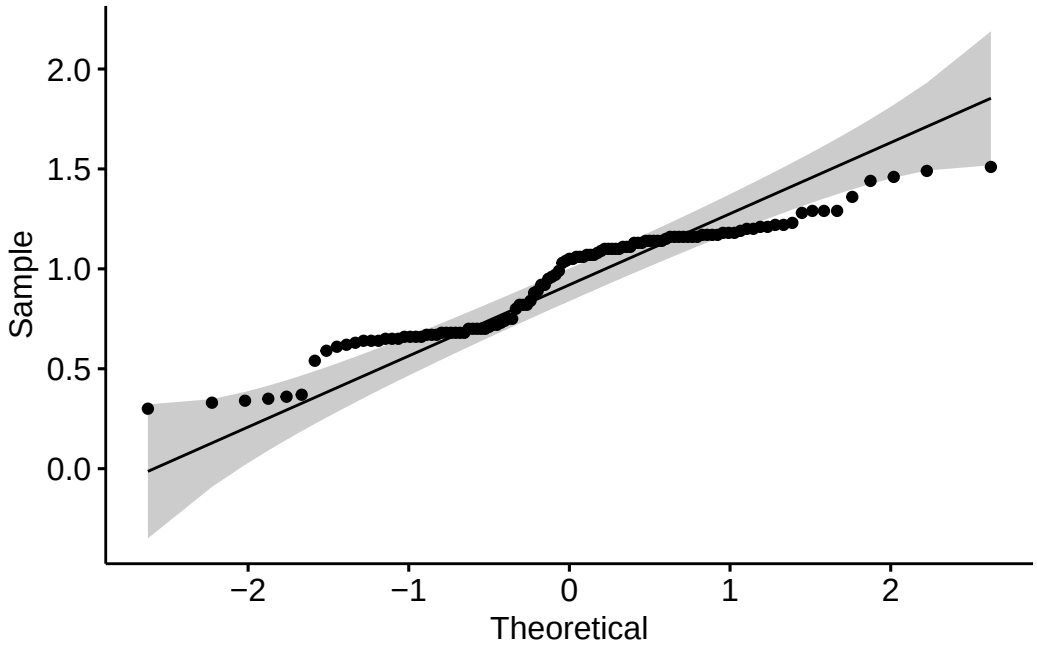
```
    Shapiro-Wilk normality test

data:  fetus$Placenta_mass
W = 0.90156, p-value = 3.798e-07
```

Placenta mass is not normally distributed (right-skewed)

```
hist(fetus$Fetal_mass)
```




```
ggqqplot(fetus$Fetal_mass)
```

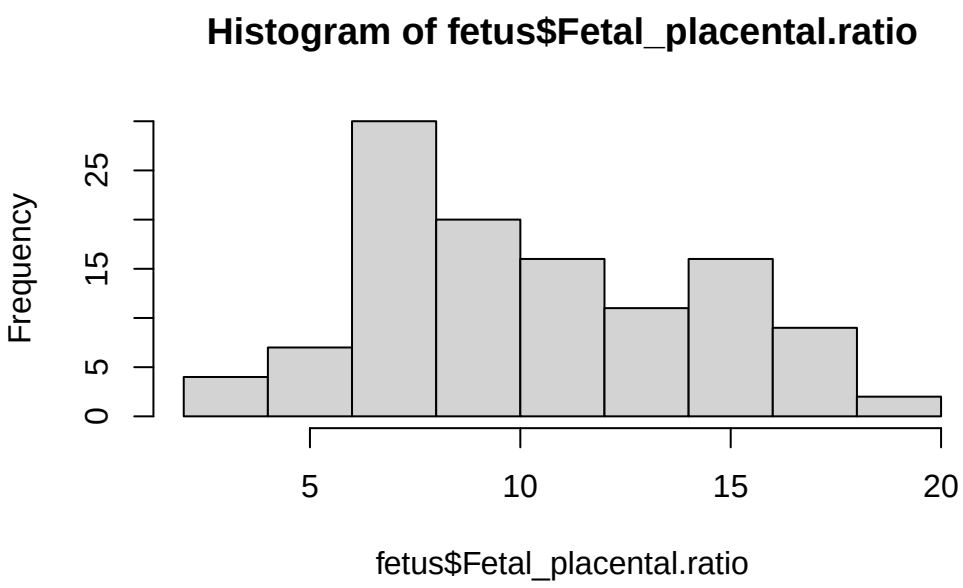



```
shapiro.test(fetus$Fetal_mass)
```

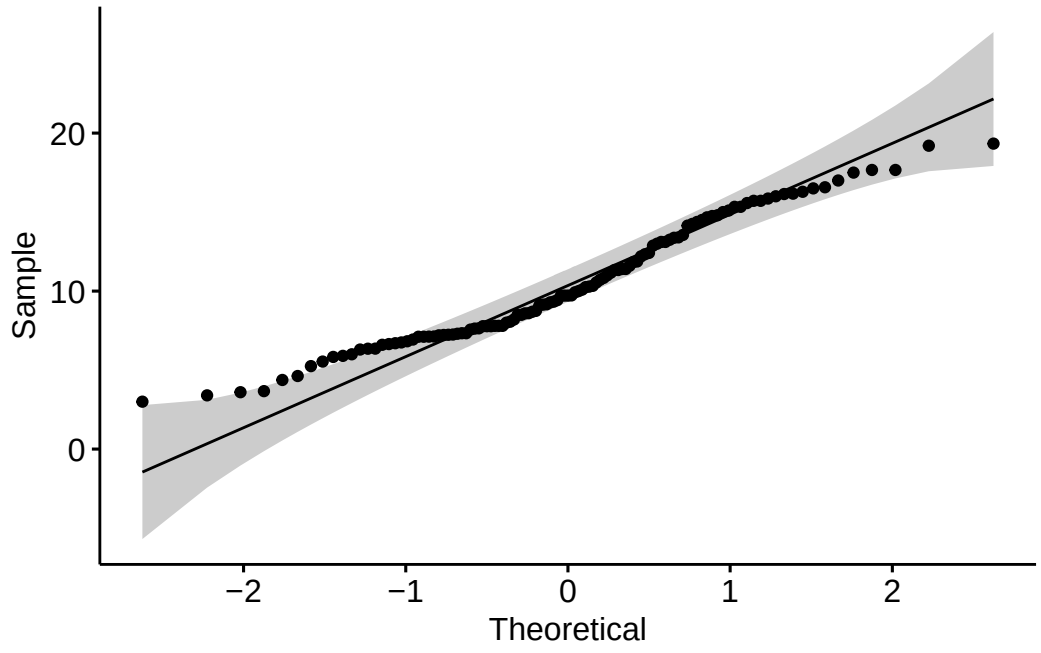
```
    Shapiro-Wilk normality test

data:  fetus$Fetal_mass
W = 0.93942, p-value = 5.661e-05
```

Fetal mass not normally distributed (bimodal). Perhaps bimodality has to do with the sex of the fetus? Worth investigating because, if not, could be related to the genotype.

```
hist(fetus$Fetal_placental.ratio)
```




```
ggqqplot(fetus$Fetal_placental.ratio)
```



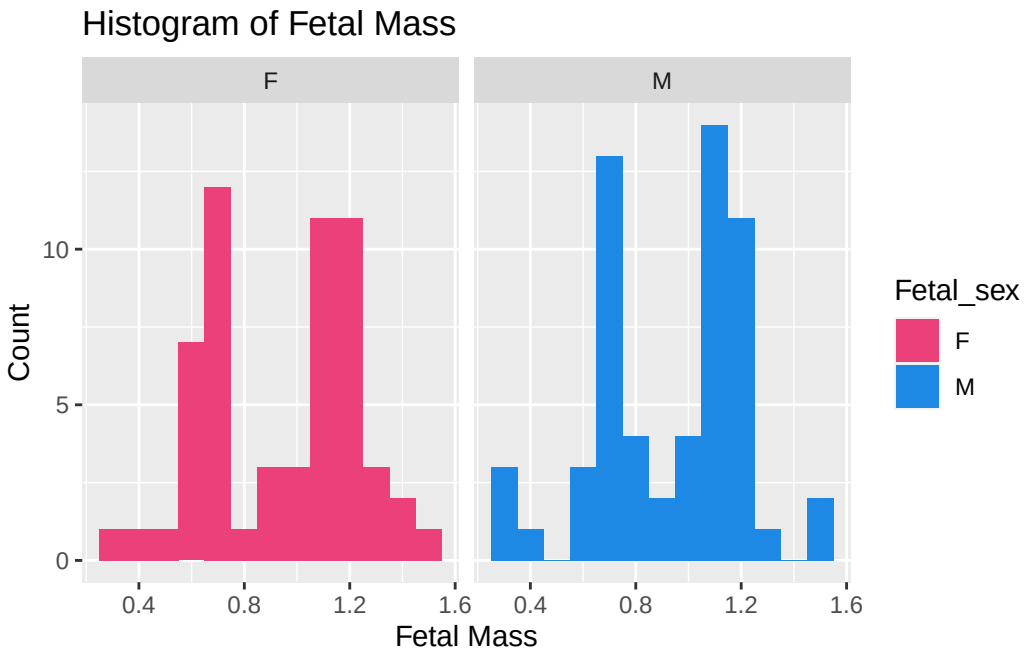
```
shapiro.test(fetus$Fetal_placental.ratio)
```

```
    Shapiro-Wilk normality test

data:  fetus$Fetal_placental.ratio
W = 0.96431, p-value = 0.003709
```

Fetal/placental ratio is not normally distributed (right skew). This makes sense since it's a transformation that is not a sum of two non-normal random variables.

Let's return to assessing fetal mass. Maybe it is related to gender?

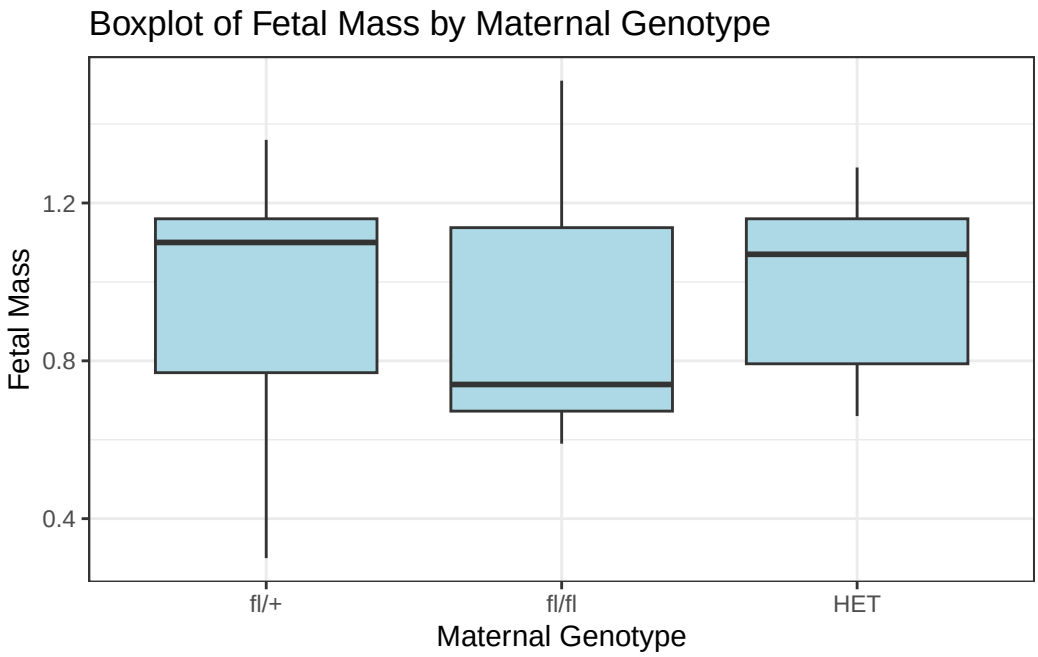
```
# Plot side by side histograms of fetal mass faceted by fetal sex with different colors
ggplot(fetus, aes(x = Fetal_mass, fill = Fetal_sex)) +
  geom_histogram(binwidth = 0.1) +
  scale_fill_manual(values = c(color_female, color_male)) +
  xlab("Fetal Mass") +
  ylab("Count") +
  ggtitle("Histogram of Fetal Mass") +
  facet_wrap(~ Fetal_sex, ncol = 2)
```



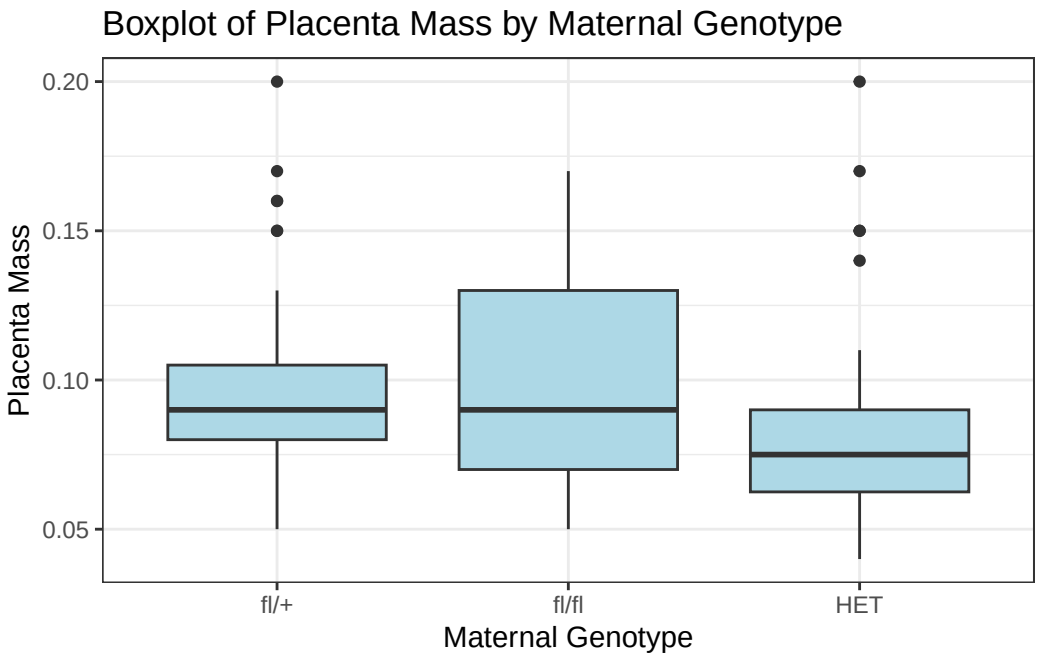
This did not help the bimodality. There is more going on here. Maybe it is related to the genotypes?

Now we visually assess the relationship between our variables of interest and the genotypes using boxplots.

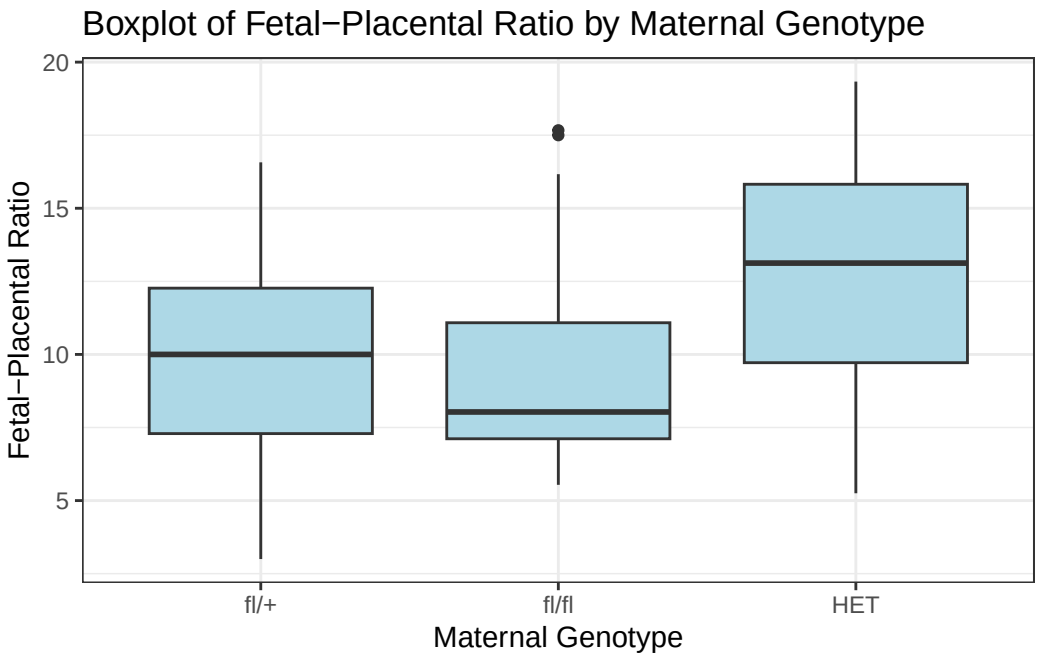
```
# Boxplot of Fetal_mass by Maternal_genotype
ggplot(fetus, aes(x = Maternal_genotype, y = Fetal_mass)) +
  geom_boxplot(fill = "lightblue") +
  xlab("Maternal Genotype") +
  ylab("Fetal Mass") +
  ggtitle("Boxplot of Fetal Mass by Maternal Genotype") +
  theme_bw()
```



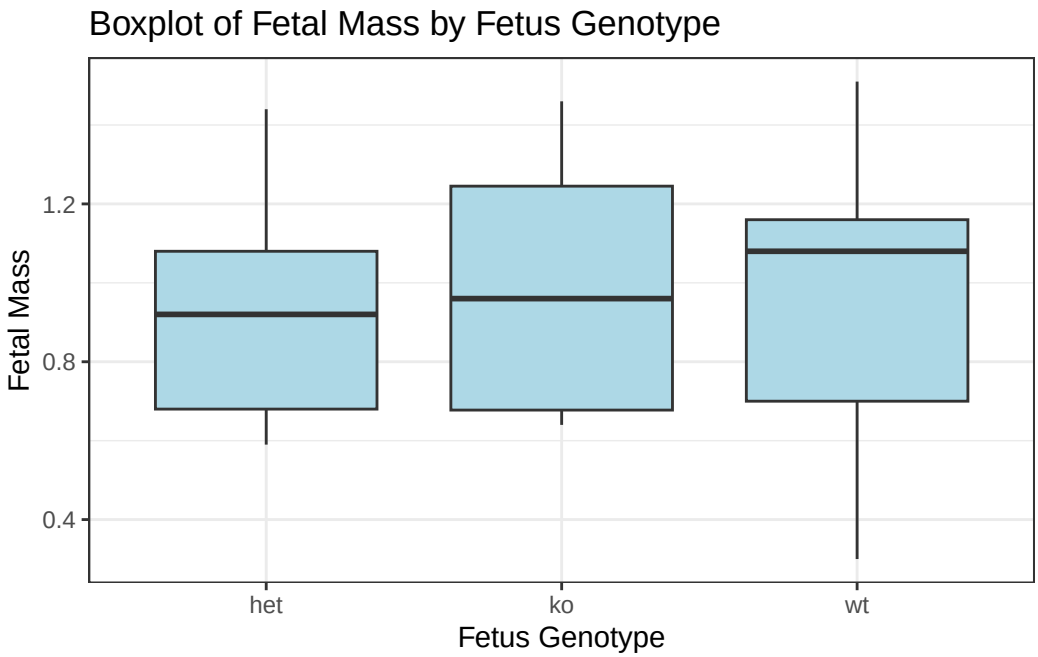

```
# Boxplot of Placenta_mass by Maternal_genotype
ggplot(fetus, aes(x = Maternal_genotype, y = Placenta_mass)) +
  geom_boxplot(fill = "lightblue") +
  xlab("Maternal Genotype") +
  ylab("Placenta Mass") +
  ggtitle("Boxplot of Placenta Mass by Maternal Genotype") +
  theme_bw()
```



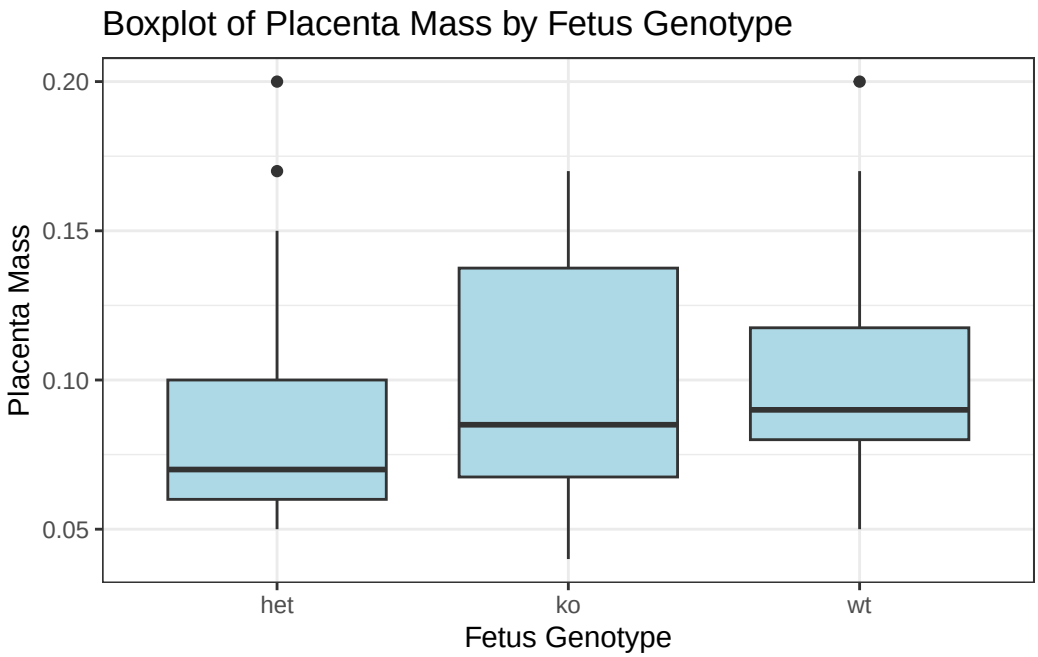

```
# Boxplot of Fetal_placental.ratio by Maternal_genotype
ggplot(fetus, aes(x = Maternal_genotype, y = Fetal_placental.ratio)) +
  geom_boxplot(fill = "lightblue") +
  xlab("Maternal Genotype") +
  ylab("Fetal-Placental Ratio") +
  ggtitle("Boxplot of Fetal-Placental Ratio by Maternal Genotype") +
  theme_bw()
```



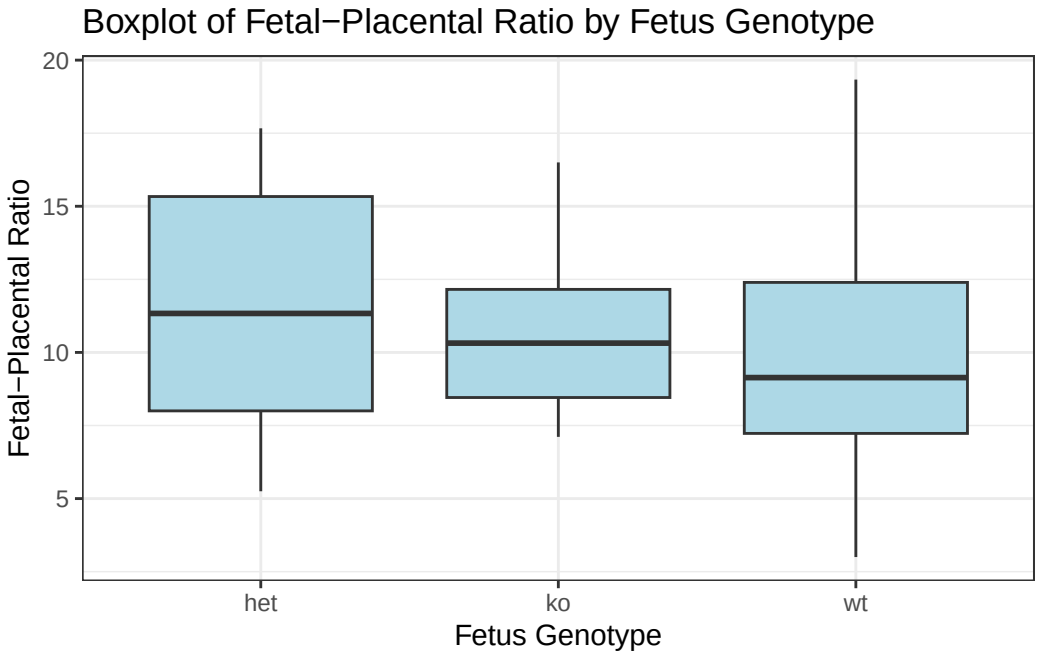

```
# Boxplot of Fetal_mass by Fetus_genotype
ggplot(fetus, aes(x = Fetus_genotype, y = Fetal_mass)) +
  geom_boxplot(fill = "lightblue") +
  xlab("Fetus Genotype") +
  ylab("Fetal Mass") +
  ggtitle("Boxplot of Fetal Mass by Fetus Genotype") +
  theme_bw()
```




```
# Boxplot of Placenta_mass by Fetus_genotype
ggplot(fetus, aes(x = Fetus_genotype, y = Placenta_mass)) +
  geom_boxplot(fill = "lightblue") +
  xlab("Fetus Genotype") +
  ylab("Placenta Mass") +
  ggtitle("Boxplot of Placenta Mass by Fetus Genotype") +
  theme_bw()
```




```
# Boxplot of Fetal_placental.ratio by Fetus_genotype
ggplot(fetus, aes(x = Fetus_genotype, y = Fetal_placental.ratio)) +
  geom_boxplot(fill = "lightblue") +
  xlab("Fetus Genotype") +
  ylab("Fetal-Placental Ratio") +
  ggtitle("Boxplot of Fetal-Placental Ratio by Fetus Genotype") +
  theme_bw()
```

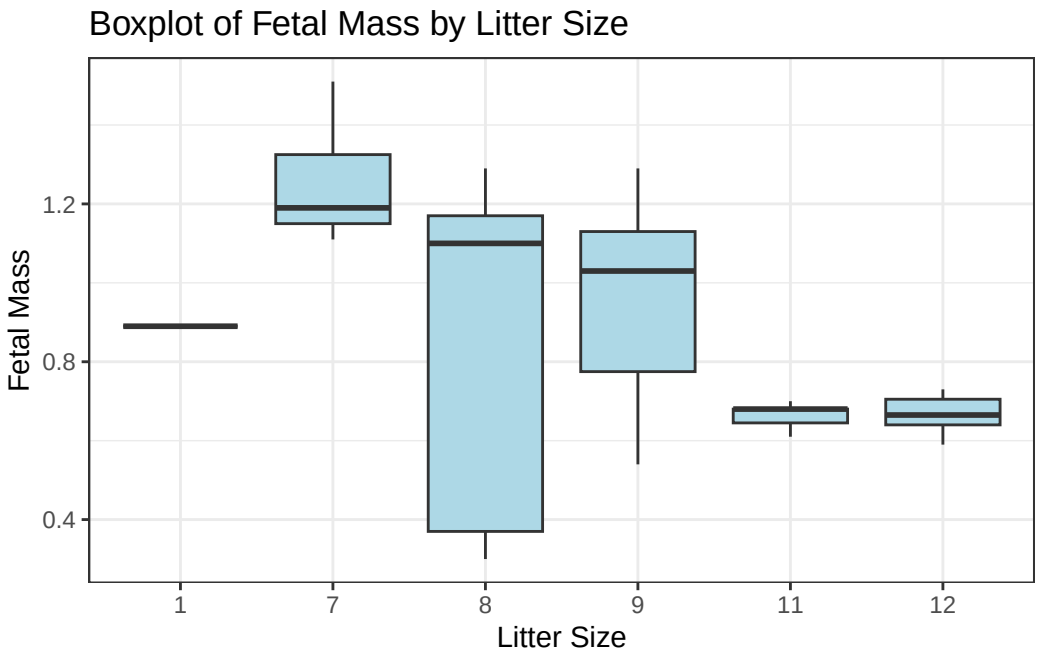


In all cases, it looks like the maternal/fetal genotype has no impact on our dependent variables. In all honesty, it's likely that masses have much more to do with the mass of the mother/father rats since it is genetic. It also might have to do with the day of conception. Presumably, if a fetus is a few days older, it will weigh more on average. I have limited subject expertise though, so I am not sure.

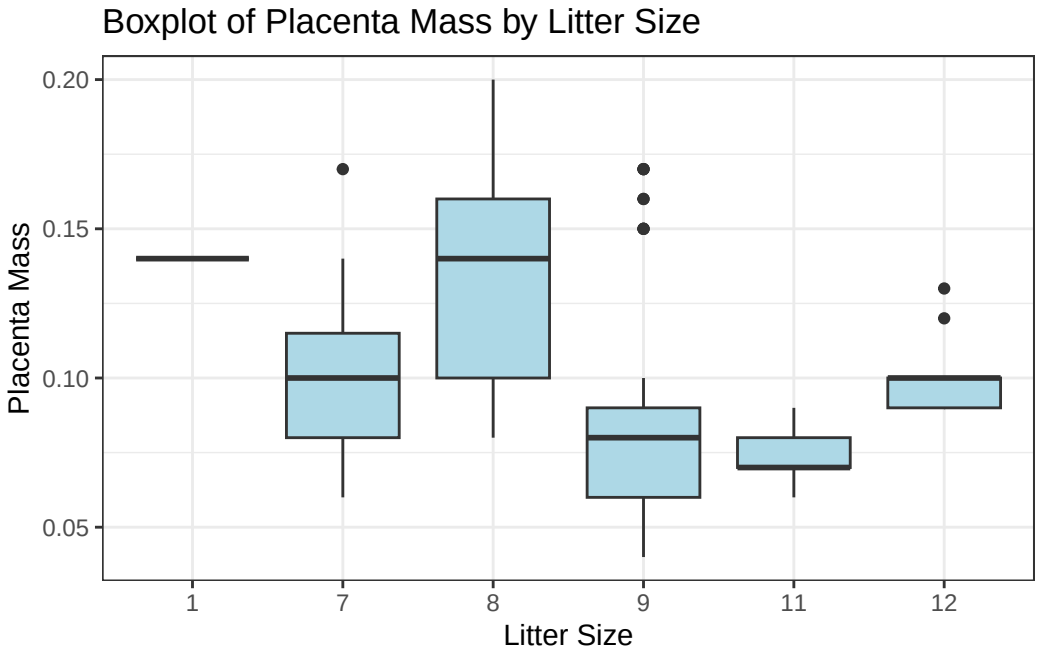
This could be assessed by linking the mother dataset to the fetus dataset, but there is no joining variable (dam id missing from mother dataset?)

Let's look at the last variables of interest for modeling, Litter_size and Fetal_sex.

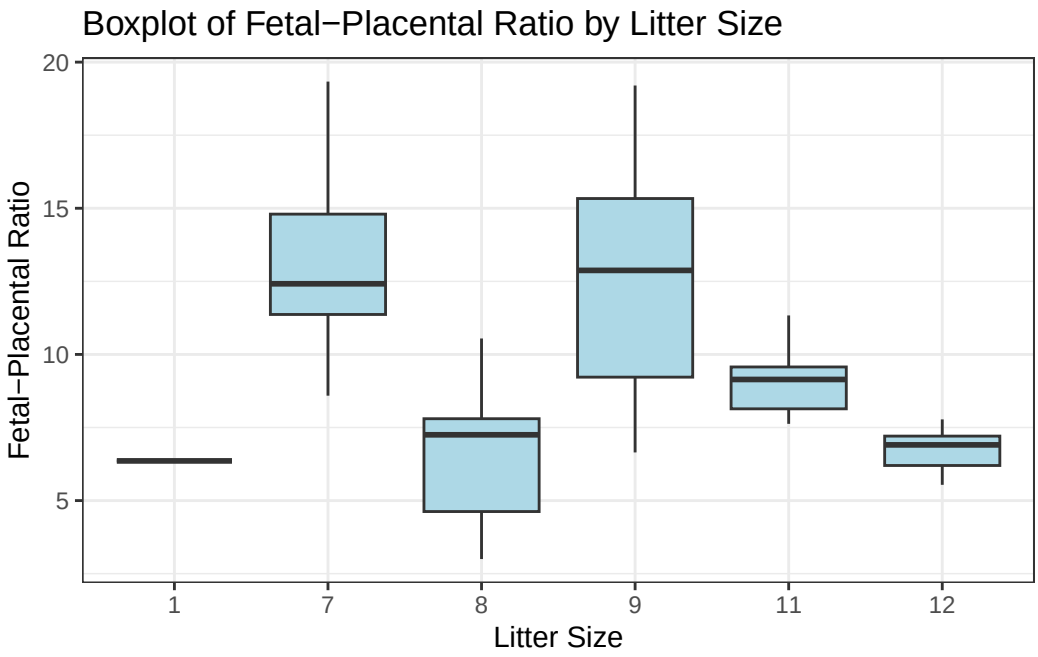
```
# Boxplot of Fetal_mass by Litter_size
ggplot(fetus, aes(x = as.factor(Litter_size), y = Fetal_mass)) +
  geom_boxplot(fill = "lightblue") +
  xlab("Litter Size") +
  ylab("Fetal Mass") +
  ggtitle("Boxplot of Fetal Mass by Litter Size") +
  theme_bw()
```



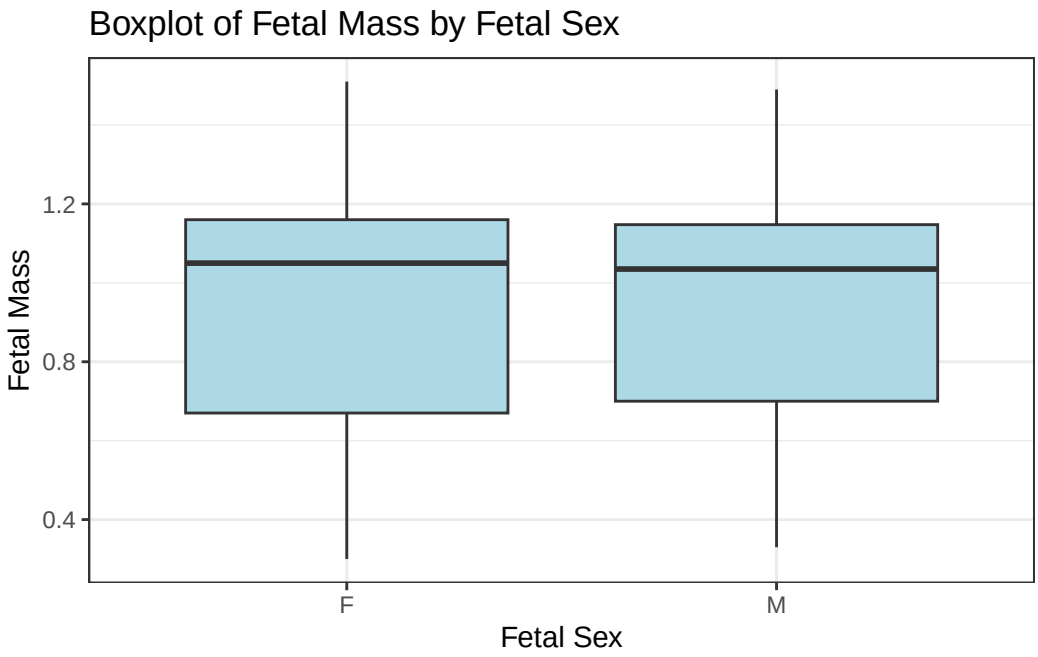

```
# Boxplot of Placenta_mass by Litter_size
ggplot(fetus, aes(x = as.factor(Litter_size), y = Placenta_mass)) +
  geom_boxplot(fill = "lightblue") +
  xlab("Litter Size") +
  ylab("Placenta Mass") +
  ggtitle("Boxplot of Placenta Mass by Litter Size") +
  theme_bw()
```



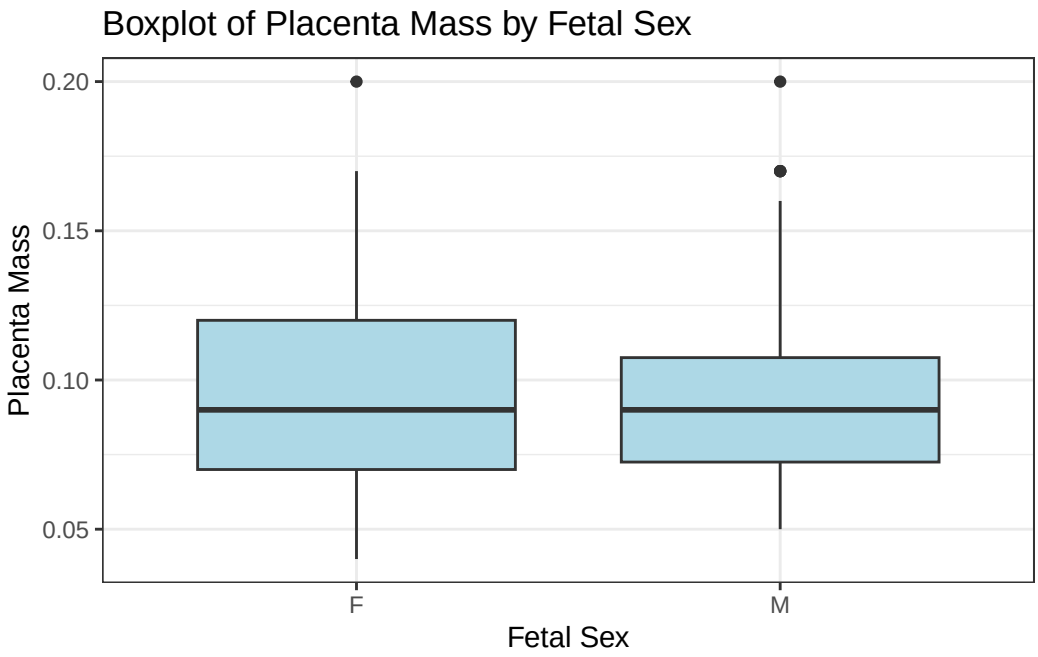

```
# Boxplot of Fetal_placental_ratio by Litter_size
ggplot(fetus, aes(x = as.factor(Litter_size), y = Fetal_placental.ratio)) +
  geom_boxplot(fill = "lightblue") +
  xlab("Litter Size") +
  ylab("Fetal-Placental Ratio") +
  ggtitle("Boxplot of Fetal-Placental Ratio by Litter Size") +
  theme_bw()
```



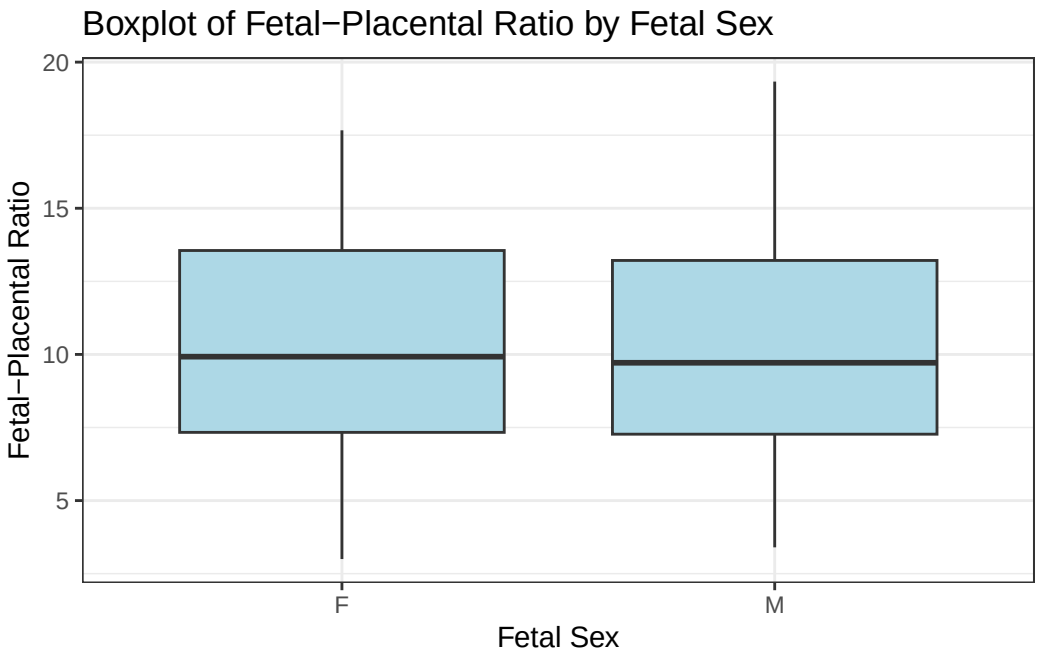

```
# Boxplot of Fetal_mass by Fetal_sex
ggplot(fetus, aes(x = Fetal_sex, y = Fetal_mass)) +
  geom_boxplot(fill = "lightblue") +
  xlab("Fetal Sex") +
  ylab("Fetal Mass") +
  ggtitle("Boxplot of Fetal Mass by Fetal Sex") +
  theme_bw()
```




```
# Boxplot of Placenta_mass by Fetal_sex
ggplot(fetus, aes(x = Fetal_sex, y = Placenta_mass)) +
  geom_boxplot(fill = "lightblue") +
  xlab("Fetal Sex") +
  ylab("Placenta Mass") +
  ggtitle("Boxplot of Placenta Mass by Fetal Sex") +
  theme_bw()
```

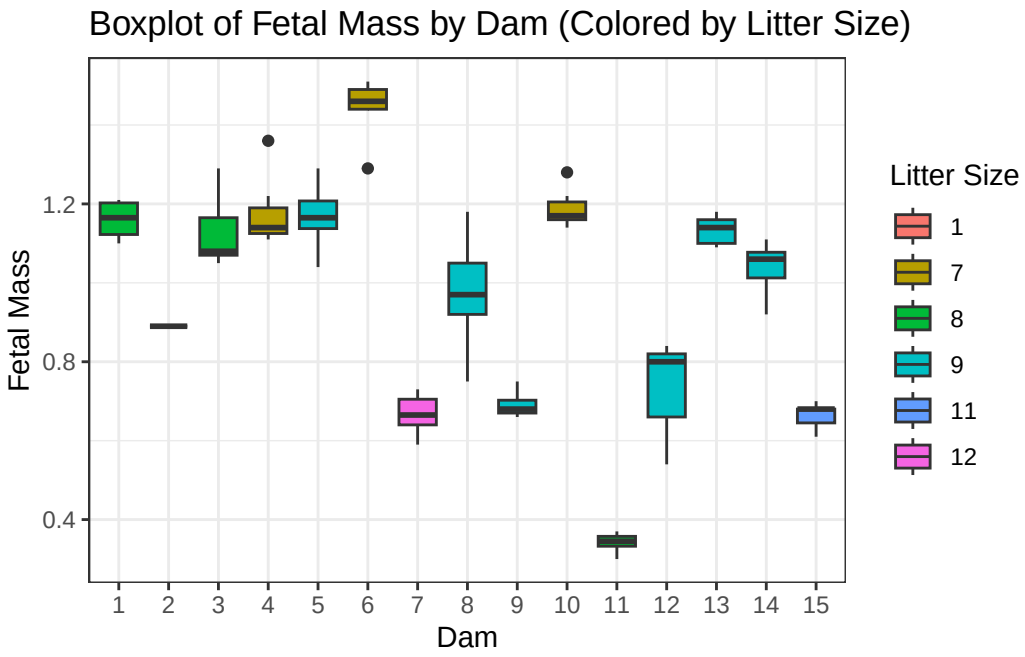



```
# Boxplot of Fetal_placental_ratio by Fetal_sex
ggplot(fetus, aes(x = Fetal_sex, y = Fetal_placental.ratio)) +
  geom_boxplot(fill = "lightblue") +
  xlab("Fetal Sex") +
  ylab("Fetal-Placental Ratio") +
  ggtitle("Boxplot of Fetal-Placental Ratio by Fetal Sex") +
  theme_bw()
```

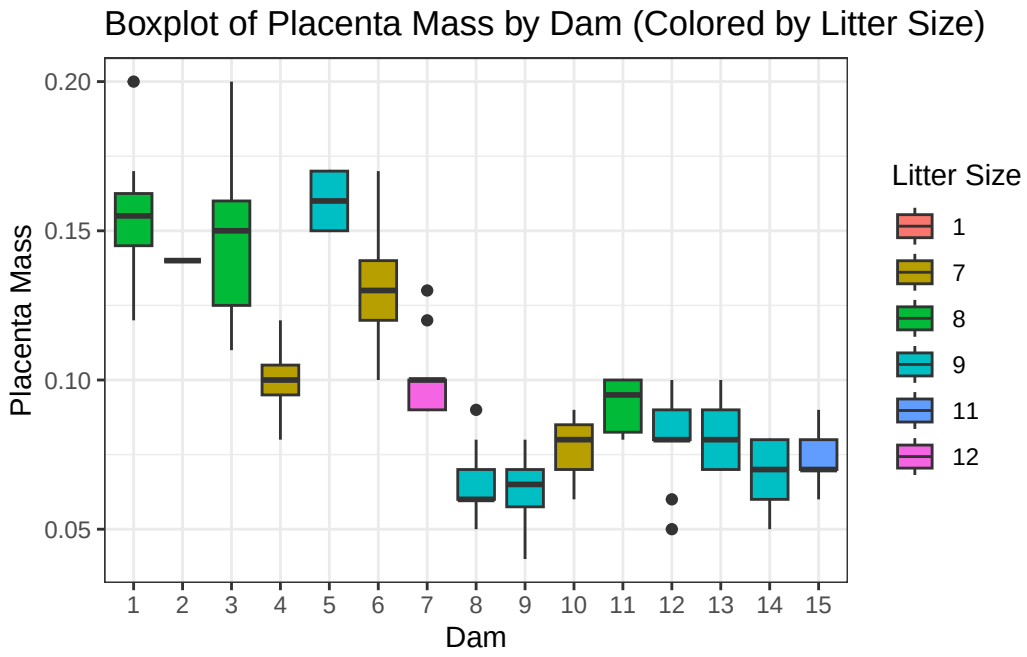


Litter size could have an effect. But, what is far more likely, is that litter size captures the variance of Dam.

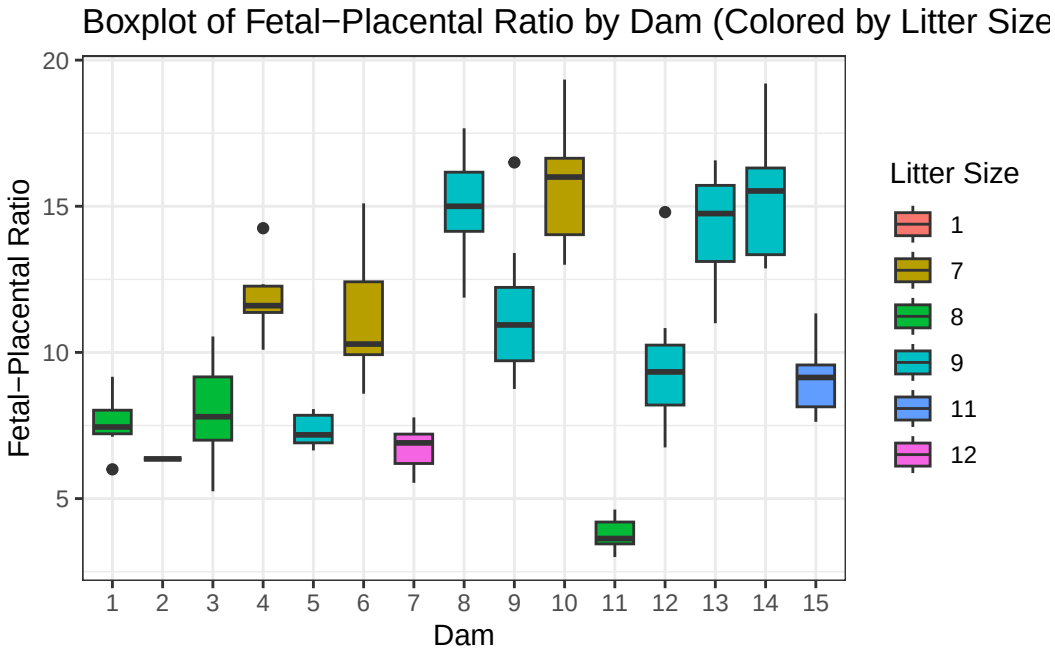
```
# Boxplot of Fetal_mass by Dam with box color by Litter_size
ggplot(fetus, aes(x = as.factor(Dam), y = Fetal_mass, fill = as.factor(Litter_size))) +
  geom_boxplot() +
  xlab("Dam") +
  ylab("Fetal Mass") +
  labs(fill = "Litter Size") +
  ggtitle("Boxplot of Fetal Mass by Dam (Colored by Litter Size)") +
  theme_bw()
```




```
# Boxplot of Placenta_mass by Dam with box color by Litter_size
ggplot(fetus, aes(x = as.factor(Dam), y = Placenta_mass, fill = as.factor(Litter_size))) +
  geom_boxplot() +
  xlab("Dam") +
  ylab("Placenta Mass") +
  labs(fill = "Litter Size") +
  ggtitle("Boxplot of Placenta Mass by Dam (Colored by Litter Size)") +
  theme_bw()
```




```r
# Boxplot of Fetal_placental_ratio by Dam with box color by Litter_size
ggplot(fetus, aes(x = as.factor(Dam), y = Fetal_placental.ratio, fill = as.factor(Litter_s
  geom_boxplot() +
  xlab("Dam") +
  ylab("Fetal-Placental Ratio") +
  labs(fill = "Litter Size") +
  ggtitle("Boxplot of Fetal-Placental Ratio by Dam (Colored by Litter Size)") +
  theme_bw()
```



While these plots are a bit cluttered, what they do show is that there is little correlation between litter size and any of our variables of interest. It also shows that these variables are highly dependent on Dam. Babies from the same mother looks similar (hence why the boxes are not tall and have short tails on average).

Modeling of the Dependent Variables

Dam is obviously the most important factor here. If we were to train a regression without it, we'd certainly get awful results. In fact, we can see this here. We only do this for Fetal_mass.

```
model <- lm(Fetal_mass ~ Litter_size + isFemale + Maternal_genotype + Fetus_genotype, data

summary(model)
```

```
Call:
lm(formula = Fetal_mass ~ Litter_size + isFemale + Maternal_genotype +
    Fetus_genotype, data = fetus)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-0.90129 -0.10817  0.02673  0.15308  0.31921

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.727050   0.151470  11.402  < 2e-16 ***
Litter_size             -0.098878   0.015497  -6.380 4.51e-09 ***
isFemale                -0.011653   0.045971  -0.253   0.8004
Maternal_genotypefl/fl   0.142784   0.065461   2.181   0.0313 *
Maternal_genotypeHET     0.086907   0.066400   1.309   0.1934
Fetus_genotypeko         0.002518   0.083843   0.030   0.9761
Fetus_genotypewt         0.031989   0.060187   0.531   0.5962
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2392 on 108 degrees of freedom
Multiple R-squared:  0.2931,    Adjusted R-squared:  0.2539
F-statistic: 7.465 on 6 and 108 DF,  p-value: 1.055e-06
```

Notice that nothing except for the intercept and Litter_size (which before we proposed might just indirectly be capturing the variance of Dam ID) are significant. With Bayesian model selection, we can confirm this is the case.

```r
y<-fetus$Fetal_mass
fetus_encoded <- fetus %>%
  mutate(
    Maternal_genotype = as.factor(Maternal_genotype),
    Fetus_genotype = as.factor(Fetus_genotype)
  )

# Create the design matrix X
X <- model.matrix(Fetal_mass ~ isFemale + Litter_size + Maternal_genotype + Fetus_genotype

lmratio.gprior<-function(z0,z1,y,X,g=dim(X)[1],nu0=1,
                 s200=mean( lm(y~-1+X[,z0==1])$res^2),
                 s201=mean( lm(y~-1+X[,z1==1])$res^2) )
{
  n<-dim(X)[1]

  X0<-X[,z0==1]
  X1<-X[,z1==1]

  H0<- (g/(g+1)) * X0%*%solve(t(X0)%*%X0)%*%t(X0)
```

```
  SS0<- t(y)%*%( diag(1,nrow=n)  - H0 ) %*%y
  p0<-sum(z0==1)

  H1<- (g/(g+1)) * X1%*%solve(t(X1)%*%X1)%*%t(X1)
  SS1<- t(y)%*%( diag(1,nrow=n)  - H1 ) %*%y
  p1<-sum(z1==1)

   -.5*(p1-p0)*log( 2*pi*(1+g))  +
    .5*nu0*log(s201/s200) + .5*(nu0+n)*log( (nu0*s200+SS0)/(nu0+s201+SS1) )
}

lpy.X<-function(y,X,g=length(y),nu0=1,s20=try(summary(lm(y~-1+X))$sigma^2,silent=TRUE))
{
  n<-dim(X)[1] ; p<-dim(X)[2]
  if(p==0) { s20<-mean(y^2) }
  H0<-0 ; if(p>0) { H0<- (g/(g+1)) * X%*%solve(t(X)%*%X)%*%t(X) }
  SS0<- t(y)%*%( diag(1,nrow=n)  - H0 ) %*%y

  -.5*n*log(2*pi) +lgamma(.5*(nu0+n)) - lgamma(.5*nu0)  - .5*p*log(1+g) +
   .5*nu0*log(.5*nu0*s20) -.5*(nu0+n)*log(.5*(nu0*s20+SS0))
}
```

```
#### Bayesian model selection
p<-dim(X)[2]
S<-1000
z<-rep(1,p)
Z<-matrix(NA,S,p)
lpy.c<-lpy.X(y,X[,z==1,drop=FALSE])
for(s in 1:S)
{
  for(j in sample(1:p))
  {
    zp<-z ; zp[j]<-1-zp[j]
    lpy.p<-lpy.X(y,X[,zp==1,drop=FALSE])
    r<- (lpy.p - lpy.c)*(-1)^(zp[j]==0)
    z[j]<-rbinom(1,1,1/(1+exp(-r)))
    if(z[j]==zp[j]) {lpy.c<-lpy.p}
  }
  Z[s,]<-z
}
```

```
means <- colMeans(Z)
matrix(means, nrow = 1, ncol = ncol(Z), dimnames = list(NULL, colnames(X)))
```

```
     (Intercept) isFemale Litter_size Maternal_genotypefl/fl
[1,]           1    0.097           1                  0.304
     Maternal_genotypeHET Fetus_genotypeko Fetus_genotypewt
[1,]                0.113            0.092            0.102
```

Only the intercept and Litter Size are probably features. We will thus switch over to a mixed model to see if there is any value to the genotypes.

```
fetal_mass_model<-lmer(Fetal_mass ~ Litter_size + isFemale*Fetus_genotype + (1|Dam) + (1|M
```

```
boundary (singular) fit: see help('isSingular')
```

```
summary(fetal_mass_model)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: Fetal_mass ~ Litter_size + isFemale * Fetus_genotype + (1 | Dam) +
    (1 | Maternal_genotype)
   Data: fetus

REML criterion at convergence: -189.1

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.81303 -0.48478 -0.05347  0.56638  2.72874

Random effects:
 Groups            Name        Variance Std.Dev.
 Dam               (Intercept) 0.082115 0.28656
 Maternal_genotype (Intercept) 0.000000 0.00000
 Residual                      0.004689 0.06848
Number of obs: 115, groups:  Dam, 15; Maternal_genotype, 3

Fixed effects:
                          Estimate Std. Error       df t value Pr(>|t|)
(Intercept)                1.16725    0.27595 13.97137   4.230 0.000844 ***
```

```
Litter_size               -0.03086    0.03217 13.65742  -0.959 0.354159
isFemale                   0.03554    0.02693 94.99296   1.320 0.190007
Fetus_genotypeko           0.03493    0.03413 95.14271   1.023 0.308707
Fetus_genotypewt           0.06588    0.02709 95.43547   2.432 0.016887 *
isFemale:Fetus_genotypeko -0.01959    0.05207 95.06774  -0.376 0.707546
isFemale:Fetus_genotypewt -0.05356    0.03322 95.20069  -1.612 0.110204
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
                 (Intr) Lttr_s isFeml Fts_gntypk Fts_gntypw isFml:Fts_gntypk
Litter_size      -0.960
isFemale         -0.061  0.000
Fets_gntypk      -0.051  0.002  0.508
Fts_gntypwt      -0.077  0.006  0.638  0.487
isFml:Fts_gntypk  0.025  0.002 -0.502 -0.602     -0.252
isFml:Fts_gntypw  0.040  0.013 -0.838 -0.421     -0.741       0.396
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

```
placenta_mass_model<-lmer(Placenta_mass ~ Litter_size + isFemale*Fetus_genotype + (1|Dam)
```

```
boundary (singular) fit: see help('isSingular')
```

```
summary(placenta_mass_model)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: Placenta_mass ~ Litter_size + isFemale * Fetus_genotype + (1 |
    Dam) + (1 | Maternal_genotype)
   Data: fetus

REML criterion at convergence: -522.4

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.29513 -0.58310 -0.01651  0.52652  3.05477

Random effects:
 Groups            Name        Variance  Std.Dev.
```

```
 Dam               (Intercept) 0.0011418 0.03379
 Maternal_genotype (Intercept) 0.0000000 0.00000
 Residual                      0.0002511 0.01584
Number of obs: 115, groups:  Dam, 15; Maternal_genotype, 3

Fixed effects:
                            Estimate Std. Error        df t value Pr(>|t|)
(Intercept)                 0.149157   0.034623 16.993231   4.308 0.000477 ***
Litter_size                -0.005834   0.003981 15.749549  -1.466 0.162446
isFemale                    0.005232   0.006227 95.410430   0.840 0.402842
Fetus_genotypeko            0.010757   0.007885 95.954771   1.364 0.175679
Fetus_genotypewt            0.001410   0.006244 97.004093   0.226 0.821820
isFemale:Fetus_genotypeko -0.016641   0.012034 95.687772  -1.383 0.169936
isFemale:Fetus_genotypewt -0.009003   0.007670 96.161517  -1.174 0.243413
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
                (Intr) Lttr_s isFeml Fts_gntypk Fts_gntypw isFml:Fts_gntypk
Litter_size     -0.956
isFemale        -0.113  0.000
Fets_gntypk     -0.096  0.005  0.508
Fts_gntypwt     -0.144  0.011  0.639  0.489
isFml:Fts_gntypk  0.046  0.004 -0.502 -0.603     -0.255
isFml:Fts_gntypw  0.076  0.023 -0.838 -0.421     -0.741       0.397
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

```
fpratio_model<-lmer(Fetal_placental.ratio ~ Litter_size + isFemale*Fetus_genotype + (1|Dam
summary(fpratio_model)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: Fetal_placental.ratio ~ Litter_size + isFemale * Fetus_genotype +
    (1 | Dam) + (1 | Maternal_genotype)
   Data: fetus

REML criterion at convergence: 489.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
```

```
-1.9672 -0.5152 -0.1121  0.4116  3.0492

Random effects:
 Groups            Name        Variance Std.Dev.
 Dam               (Intercept) 15.0830  3.8837
 Maternal_genotype (Intercept)  0.1468  0.3832
 Residual                       2.8984  1.7025
Number of obs: 115, groups:  Dam, 15; Maternal_genotype, 3

Fixed effects:
                          Estimate Std. Error       df t value Pr(>|t|)
(Intercept)                8.71167    3.94680 15.49467   2.207   0.0428 *
Litter_size                0.14759    0.45407 12.39835   0.325   0.7506
isFemale                   0.15960    0.66913 95.17382   0.239   0.8120
Fetus_genotypeko          -1.15735    0.84746 95.65656  -1.366   0.1752
Fetus_genotypewt           0.47085    0.67146 96.45555   0.701   0.4848
isFemale:Fetus_genotypeko  2.45928    1.29330 95.41909   1.902   0.0602 .
isFemale:Fetus_genotypewt -0.06081    0.82446 95.84057  -0.074   0.9414
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
                 (Intr) Lttr_s isFeml Fts_gntypk Fts_gntypw isFml:Fts_gntypk
Litter_size      -0.955
isFemale         -0.107  0.000
Fets_gntypk      -0.090  0.005  0.508
Fts_gntypwt      -0.135  0.011  0.639  0.489
isFml:Fts_gntypk  0.043  0.003 -0.502 -0.603     -0.254
isFml:Fts_gntypw  0.071  0.022 -0.838 -0.421     -0.741       0.397
```

We've noticed that a fetus of genotype WT might have, on average, higher fetal mass. We investigate this claim on 4 mothers with the largest litter sizes.

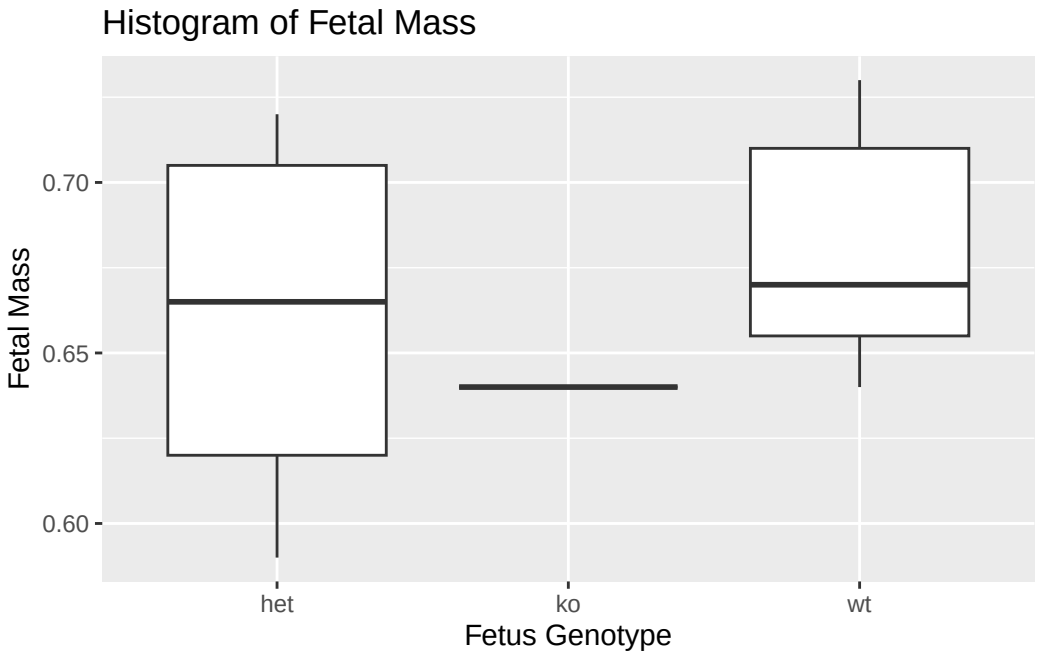
filtered_fetus

```r
# Filter the dataset to include observations from the four Dams with the largest litter si
dam_values <- unique(fetus %>%
  group_by(Dam) %>%
  arrange(desc(Litter_size)) %>%
  top_n(4, Dam) %>%
  pull(Dam)
)[1:4]
```

```
# Filter the dataset to include observations where Dam is one of the specified values
dam1 <- fetus %>%
  filter(Dam == dam_values[1])
dam2 <- fetus %>%
  filter(Dam == dam_values[2])
dam3 <- fetus %>%
  filter(Dam == dam_values[3])
dam4 <- fetus %>%
  filter(Dam == dam_values[4])
```

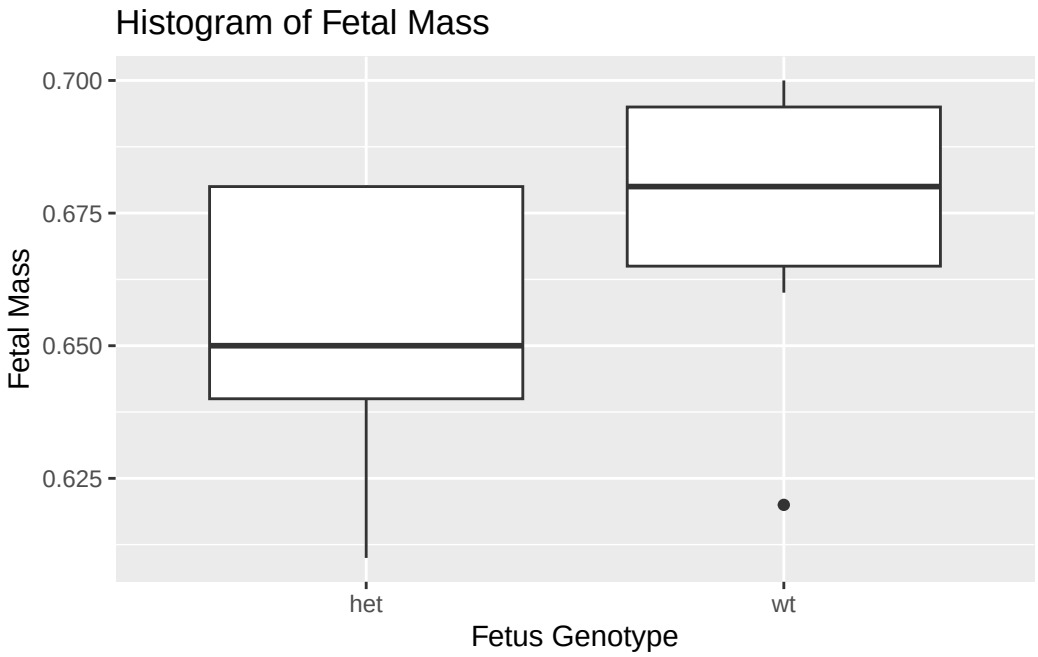
The plots below show the actual values.

```
ggplot(dam1, aes(y = Fetal_mass, x = Fetus_genotype)) +
  geom_boxplot() +
  ylab("Fetal Mass") +
  xlab("Fetus Genotype") +
  ggtitle("Histogram of Fetal Mass")
```

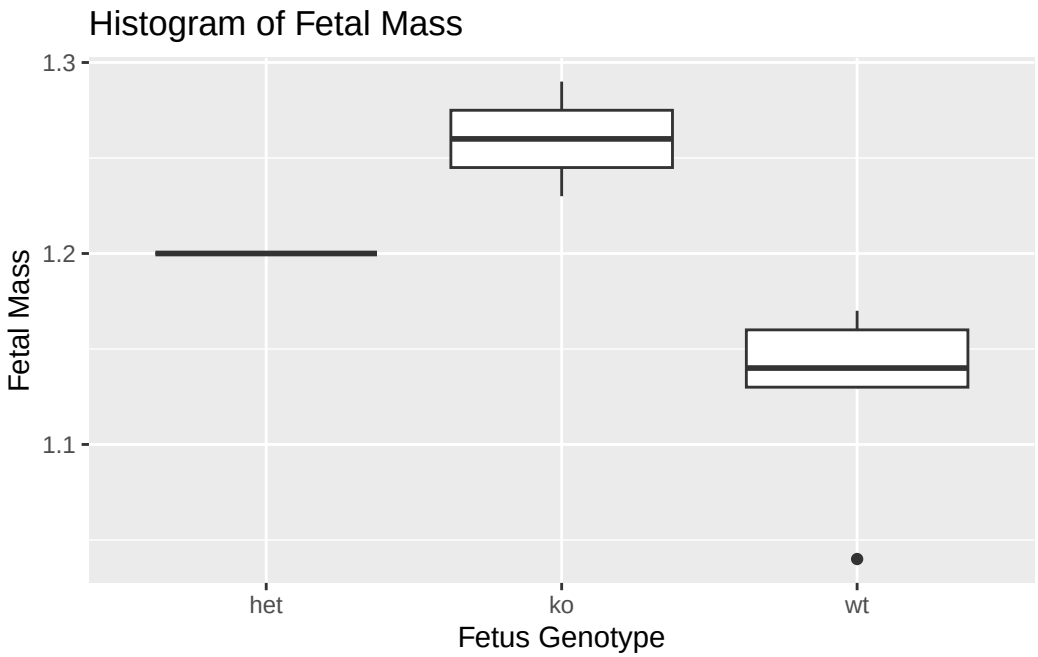



```
ggplot(dam2, aes(y = Fetal_mass, x = Fetus_genotype)) +
  geom_boxplot() +
  ylab("Fetal Mass") +
```

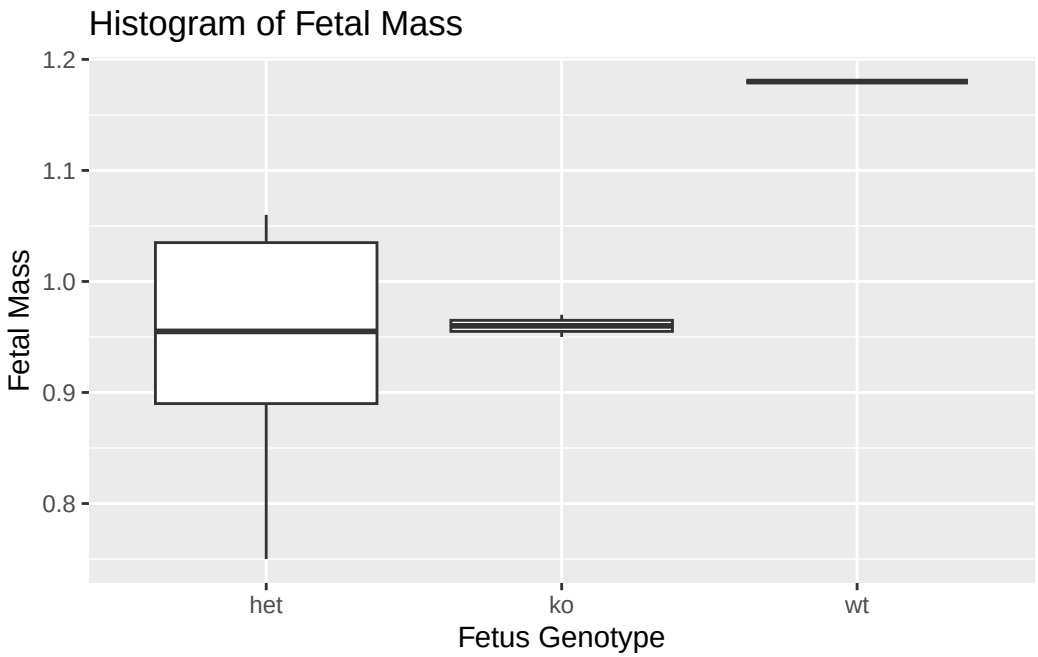
```
  xlab("Fetus Genotype") +
  ggtitle("Histogram of Fetal Mass")
```



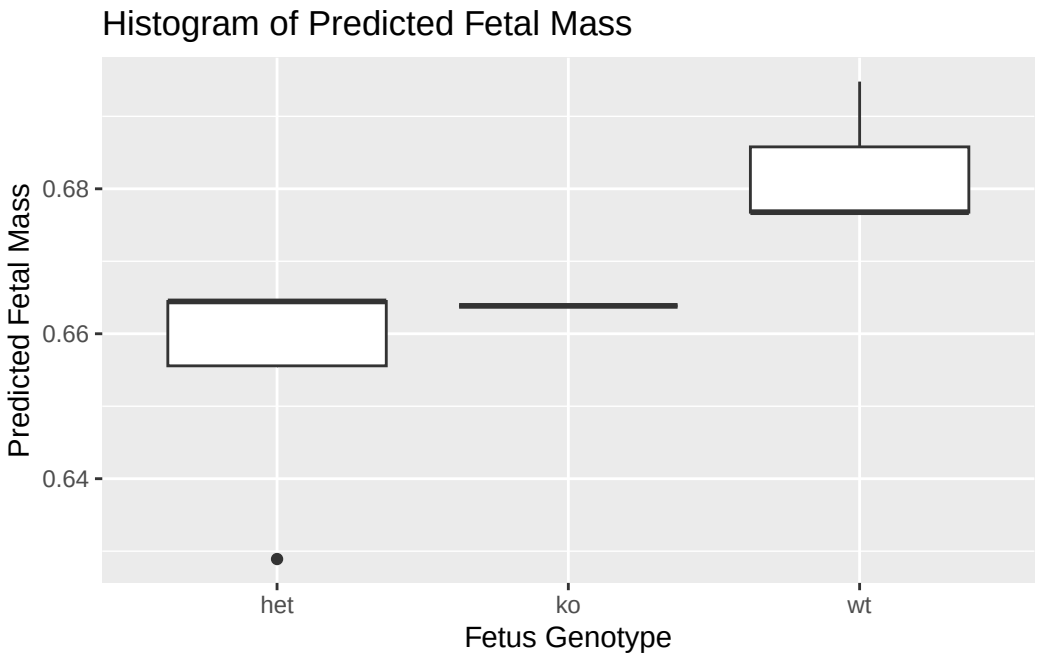

```
ggplot(dam3, aes(y = Fetal_mass, x = Fetus_genotype)) +
  geom_boxplot() +
  ylab("Fetal Mass") +
  xlab("Fetus Genotype") +
  ggtitle("Histogram of Fetal Mass")
```



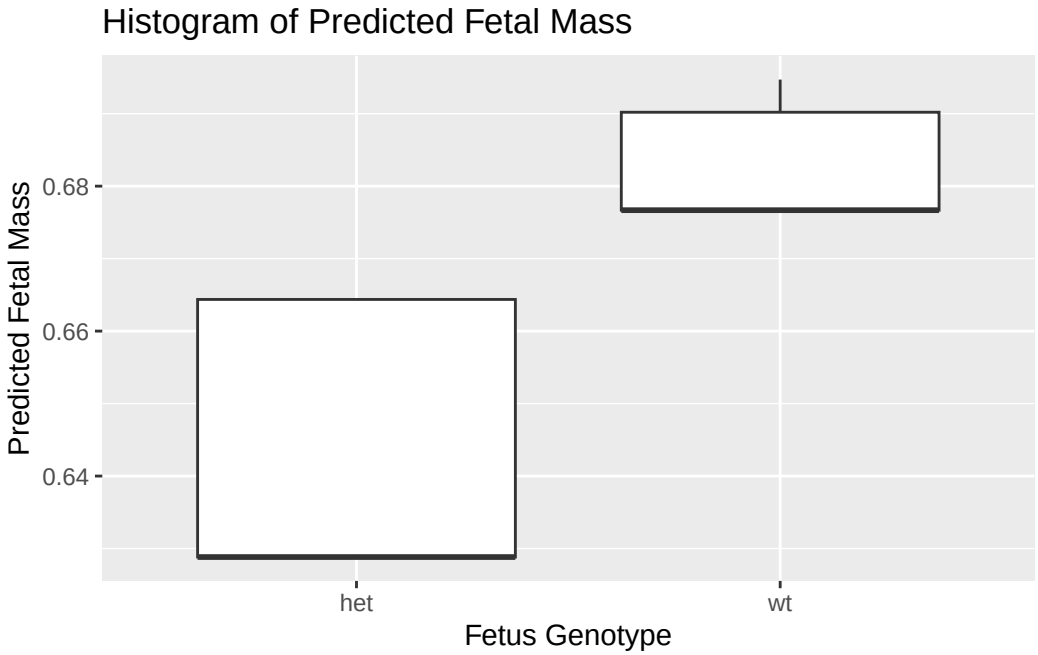

```
ggplot(dam4, aes(y = Fetal_mass, x = Fetus_genotype)) +
  geom_boxplot() +
  ylab("Fetal Mass") +
  xlab("Fetus Genotype") +
  ggtitle("Histogram of Fetal Mass")
```



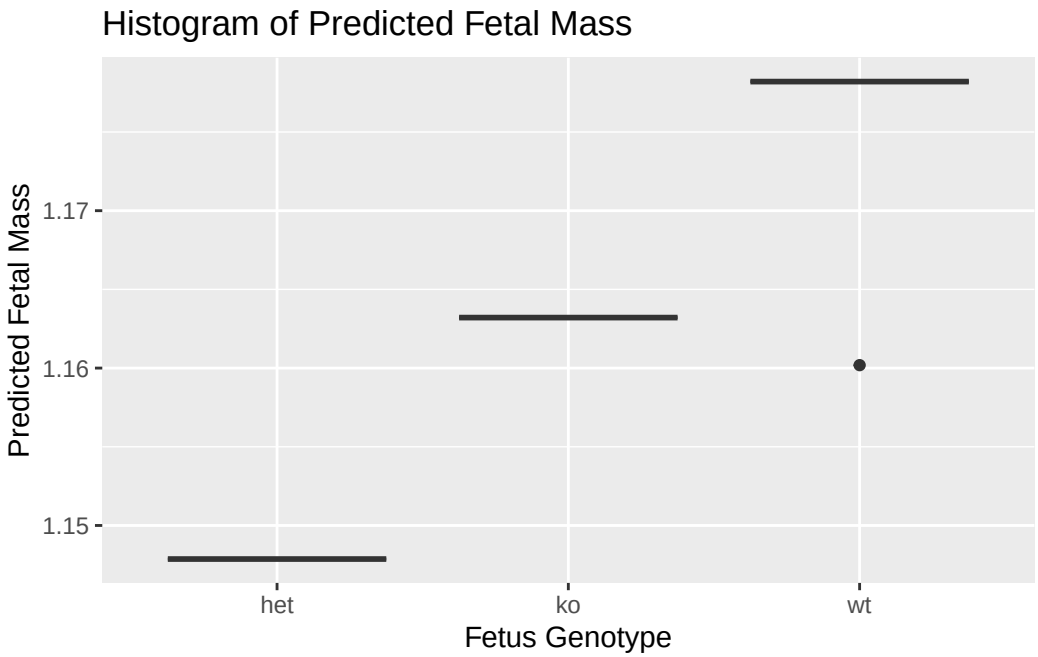

```
dam1$predicted = predict(fetal_mass_model, newdata=dam1)
dam2$predicted = predict(fetal_mass_model, newdata=dam2)
dam3$predicted = predict(fetal_mass_model, newdata=dam3)
dam4$predicted = predict(fetal_mass_model, newdata=dam4)

ggplot(dam1, aes(y = predicted, x = Fetus_genotype)) +
  geom_boxplot() +
  ylab("Predicted Fetal Mass") +
  xlab("Fetus Genotype") +
  ggtitle("Histogram of Predicted Fetal Mass")
```



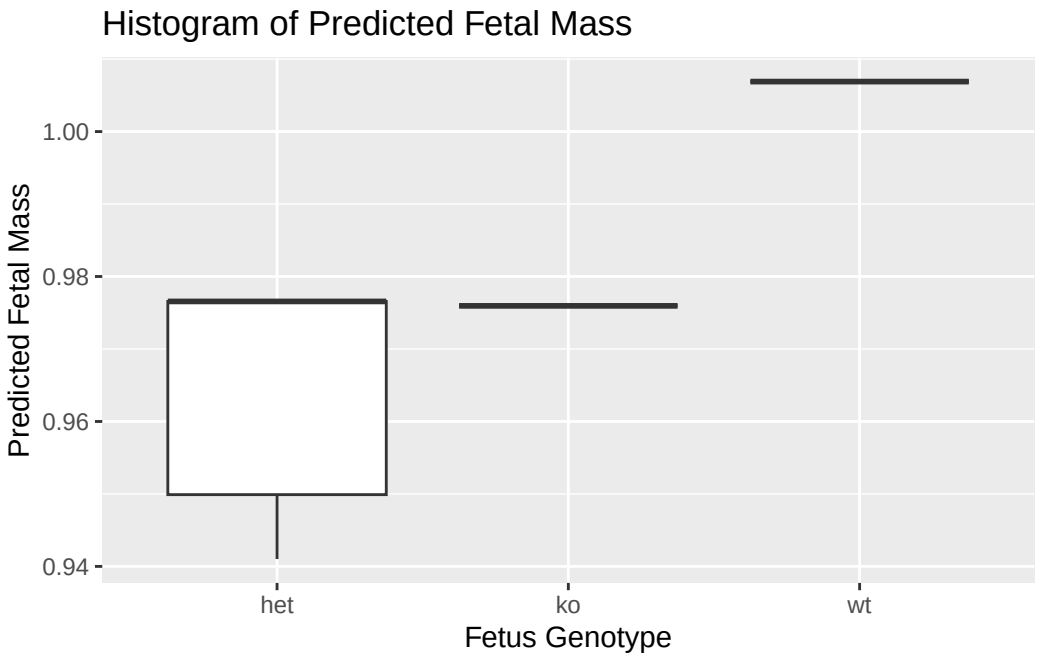

```
ggplot(dam2, aes(y = predicted, x = Fetus_genotype)) +
  geom_boxplot() +
  ylab("Predicted Fetal Mass") +
  xlab("Fetus Genotype") +
  ggtitle("Histogram of Predicted Fetal Mass")
```




```
ggplot(dam3, aes(y = predicted, x = Fetus_genotype)) +
  geom_boxplot() +
  ylab("Predicted Fetal Mass") +
  xlab("Fetus Genotype") +
  ggtitle("Histogram of Predicted Fetal Mass")
```




```
ggplot(dam4, aes(y = predicted, x = Fetus_genotype)) +
  geom_boxplot() +
  ylab("Predicted Fetal Mass") +
  xlab("Fetus Genotype") +
  ggtitle("Histogram of Predicted Fetal Mass")
```



Mother Data

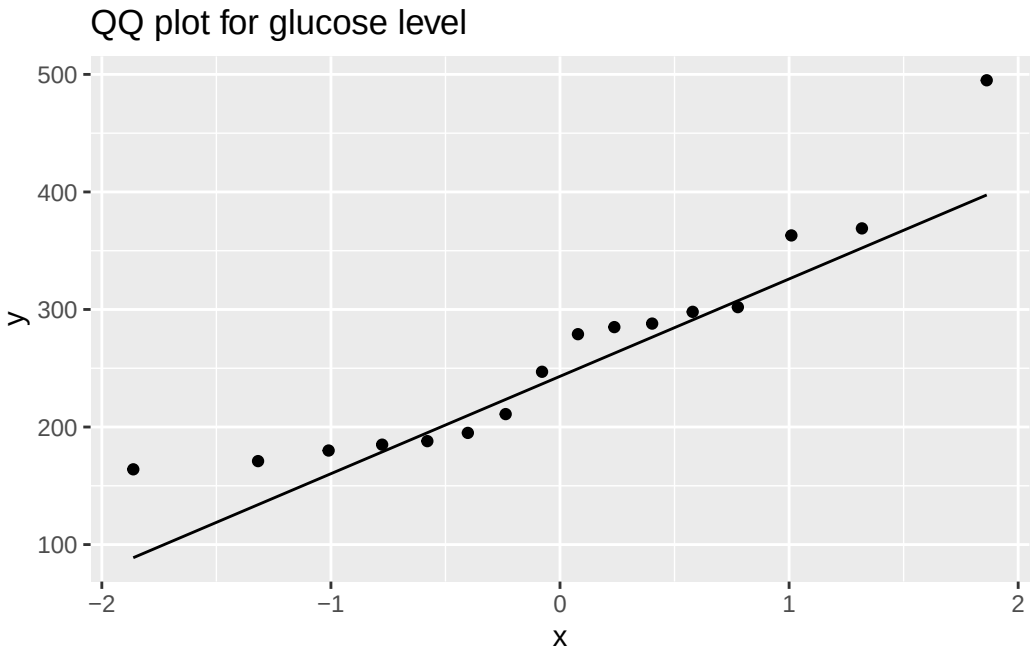
```
mother <- read.csv("data/mother-data.csv", header=T)
```

Now we'd like to assess the differences in mean blood glucose levels between maternal genotypes. For each time period, we will check if there is normality and constant variance, then conduct an ANCOVA analysis to see if the genotype makes a difference.

```
# Manual Computation of CIs and Pvals. Will be used for Pairwise Comparison

compute_CI_pval <- function(mu1, mu2, conf_level, n1, n2, N, K, MSw) {
  # Pairwise CI computation of the difference between mu1 and mu2 (0 being no diff)
  mu = mu1 - mu2
  sd = sqrt(MSw * (1/n1 + 1/n2))

  bonf = 1 - ((1-conf_level) / (K*2))
  t = qt(bonf, df = N - K)

  interval = c(mu - t*sd, mu + t*sd)

  # Using a normal dist. Maybe incorrect?
```

```
  pval = pnorm(0, mean = mu, sd = sd)

  return(c(interval, pval))
}

nplus = 5 ; nfl = 7 ; nhet = 4 ; N = 16 ; K = 3
```

```
# Checking visually
ggplot(mother, aes(sample=gluc_60)) +
  stat_qq() +
  stat_qq_line() +
  ggtitle("QQ plot for glucose level")
```




```
# Checking mathematically
shapiro.test(mother$gluc_60)
```

```
    Shapiro-Wilk normality test

data:  mother$gluc_60
W = 0.88957, p-value = 0.05483
```

Normality is satisfied.

```
mother %>%
  group_by(Maternal_genotype) %>%
  summarise(n = n(), mean = mean(gluc_60), sd = sd(gluc_60))
```

```
# A tibble: 3 x 4
  Maternal_genotype     n  mean    sd
  <chr>             <int> <dbl> <dbl>
1 fl/+                  5  223.  60.7
2 fl/fl                 7  327.  95.3
3 HET                   4  204.  34.5
```

Constant variance is satisfied ($95/34 \approx 2.8$ and sample size is small, so this is fine).

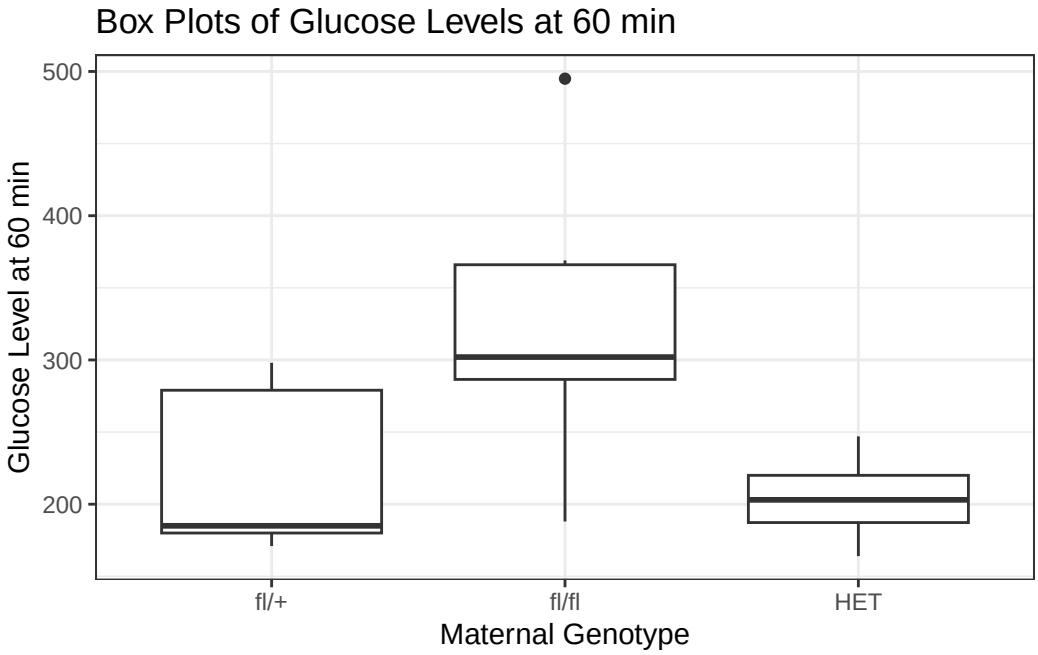
```
# First, construct a model including covariates
model <- lm(gluc_60 ~ Maternal_genotype + num_fetus + percent_body_weight_gain, data = mot

# Then perform ANCOVA using the model
anova(model)
```

```
Analysis of Variance Table

Response: gluc_60
                         Df Sum Sq Mean Sq F value  Pr(>F)
Maternal_genotype         2  50758 25379.1  6.2132 0.01564 *
num_fetus                 1  27732 27732.4  6.7894 0.02445 *
percent_body_weight_gain  1    187   186.8  0.0457 0.83456
Residuals                11  44932  4084.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We find that only at 60 minutes are we 95% confident that there is a difference in average blood glucose level.

```
# Group Comparison Glucose 60
ggplot(data = mother, aes(x = Maternal_genotype, y = gluc_60)) +
  geom_boxplot() +
  xlab("Maternal Genotype") +
  ylab("Glucose Level at 60 min") +
```

```
  ggtitle("Box Plots of Glucose Levels at 60 min") +
  theme_bw()
```




```
# Glucose 60
MSw = 4084.7
muplus = 222.6 ; mufl = 327.1429 ; muhet = 204.25
compute_CI_pval(mufl, muplus, 0.95, nfl, nplus, N, K, MSw)
```

```
[1] 1.782089e+00 2.073037e+02 2.606575e-03
```

```
compute_CI_pval(muhet, muplus, 0.95, nhet, nplus, N, K, MSw)
```

```
[1] -136.0772993   99.3772993    0.6656766
```

```
compute_CI_pval(mufl, muhet, 0.95, nfl, nhet, N, K, MSw)
```

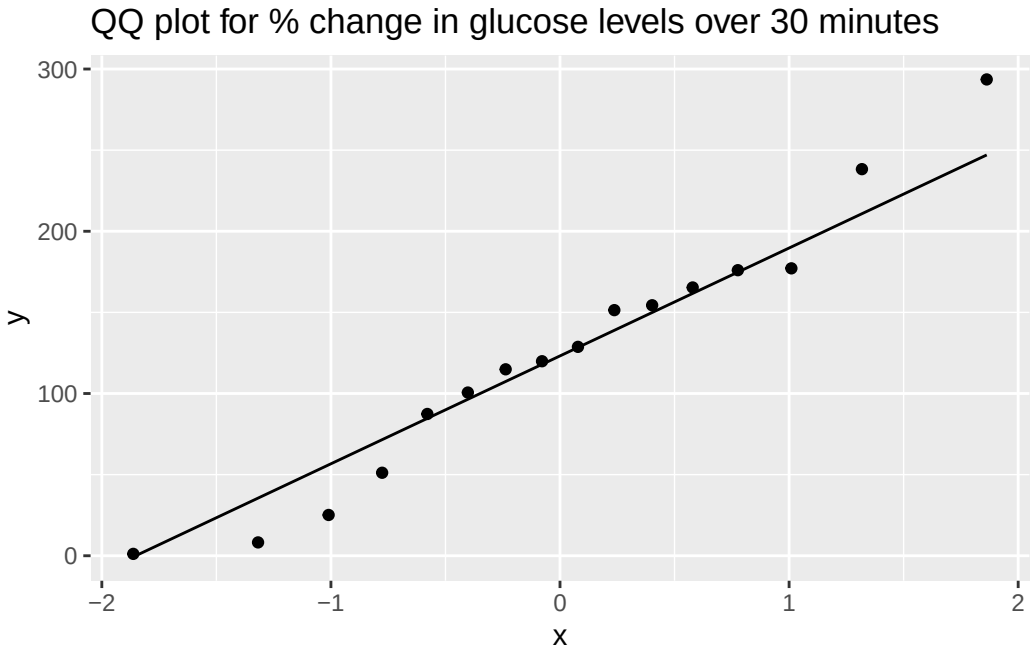
```
[1] 1.289402e+01 2.328918e+02 1.078145e-03
```

We are 95% confident that mean glucose levels at 60 min differs between:

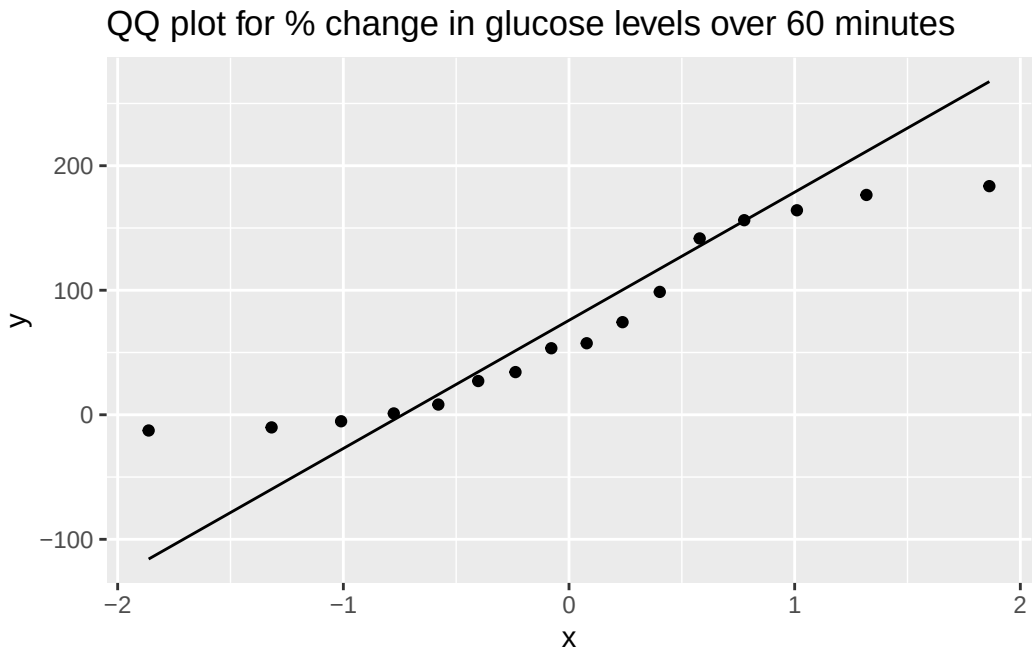
- fl/+ and fl/fl (pval = 0.00261)
- fl/fl and HET (pval = 0.00108)

Now we repeat this for % change in blood glucose levels

```
# Checking visually
ggplot(mother, aes(sample=chng_base_30)) +
  stat_qq() +
  stat_qq_line() +
  ggtitle("QQ plot for % change in glucose levels over 30 minutes")
```




```
ggplot(mother, aes(sample=chng_base_60)) +
  stat_qq() +
  stat_qq_line() +
  ggtitle("QQ plot for % change in glucose levels over 60 minutes")
```




```
# Checking mathematically
shapiro.test(mother$chng_base_30)
```

```
    Shapiro-Wilk normality test

data:  mother$chng_base_30
W = 0.96662, p-value = 0.7814
```

```
shapiro.test(mother$chng_base_60)
```

```
    Shapiro-Wilk normality test

data:  mother$chng_base_60
W = 0.88822, p-value = 0.0522
```

Normality is satisfied.

```r
mother %>%
  group_by(Maternal_genotype) %>%
  summarise(n = n(), mean = mean(chng_base_30), sd = sd(chng_base_30))
```

```
# A tibble: 3 x 4
  Maternal_genotype     n  mean    sd
  <chr>             <int> <dbl> <dbl>
1 fl/+                  5  85.1  70.1
2 fl/fl                 7 186.   59.7
3 HET                   4  67.0  51.7
```

```r
mother %>%
  group_by(Maternal_genotype) %>%
  summarise(n = n(), mean = mean(chng_base_60), sd = sd(chng_base_60))
```

```
# A tibble: 3 x 4
  Maternal_genotype     n  mean    sd
  <chr>             <int> <dbl> <dbl>
1 fl/+                  5  33.9  49.3
2 fl/fl                 7 131.   59.8
3 HET                   4  16.3  29.8
```

Constant variance is satisfied.

```r
# First, construct a model including covariates
model_base_30 <- lm(chng_base_30 ~ Maternal_genotype + num_fetus + percent_body_weight_gai
model_base_60 <- lm(chng_base_60 ~ Maternal_genotype + num_fetus + percent_body_weight_gai

# Then perform ANCOVA using the model
anova(model_base_30)
```

```
Analysis of Variance Table

Response: chng_base_30
                         Df Sum Sq Mean Sq F value  Pr(>F)
Maternal_genotype         2  47244 23621.9  5.4670 0.02247 *
num_fetus                 1     38    37.9  0.0088 0.92709
percent_body_weight_gain  1   1499  1499.2  0.3470 0.56773
Residuals                11  47529  4320.8
```

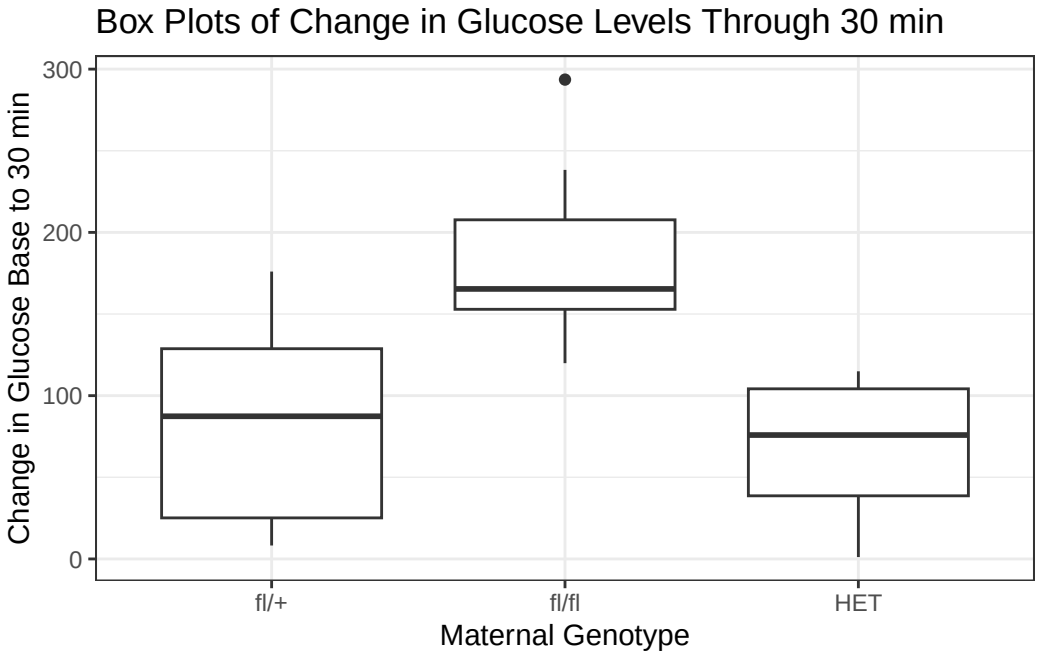
```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model_base_60)
```

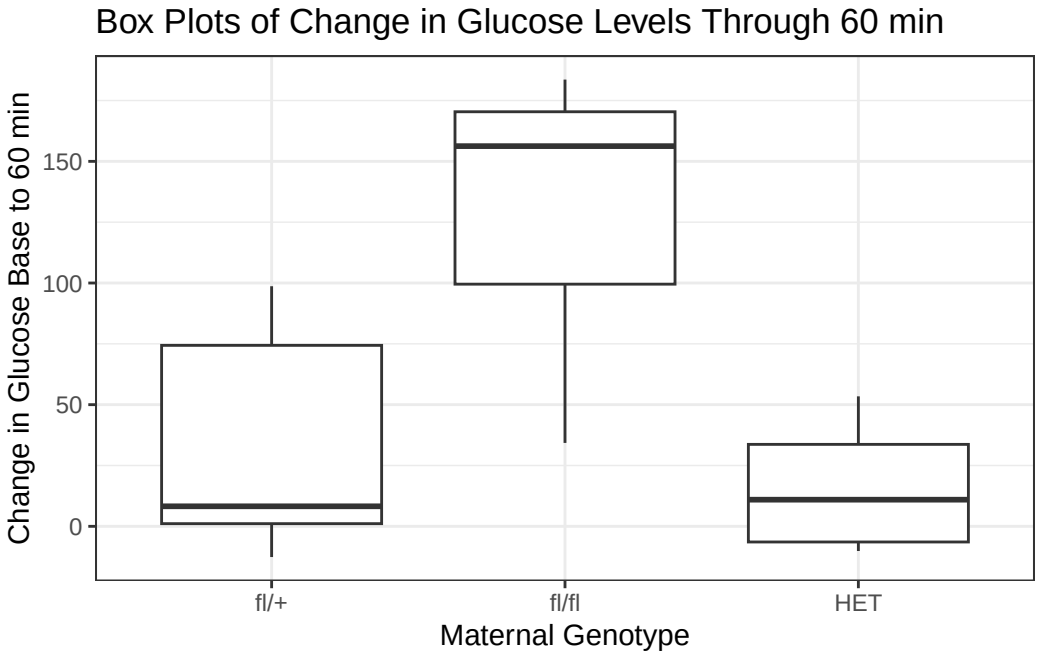
```
Analysis of Variance Table

Response: chng_base_60
                         Df Sum Sq Mean Sq F value   Pr(>F)
Maternal_genotype         2  43659 21829.3  9.7512 0.003663 **
num_fetus                 1   7821  7821.0  3.4937 0.088443 .
percent_body_weight_gain  1   1396  1396.3  0.6237 0.446353
Residuals                11  24625  2238.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

base-30 and base-60 have significant p-values to suggest a difference among maternal genotype groups.

```r
# Group Comparison Change Base to 30 Glucose
ggplot(data = mother, aes(x = Maternal_genotype, y = chng_base_30)) +
  geom_boxplot() +
  xlab("Maternal Genotype") +
  ylab("Change in Glucose Base to 30 min") +
  ggtitle("Box Plots of Change in Glucose Levels Through 30 min") +
  theme_bw()
```




```
# Group Comparison Change Base to 60 Glucose
ggplot(data = mother, aes(x = Maternal_genotype, y = chng_base_60)) +
  geom_boxplot() +
  xlab("Maternal Genotype") +
  ylab("Change in Glucose Base to 60 min") +
  ggtitle("Box Plots of Change in Glucose Levels Through 60 min") +
  theme_bw()
```




```
# Change in Glucose, Base to 30m
MSw = 4320.8
muplus = 85.09862 ; mufl = 185.72052 ; muhet = 66.95435
compute_CI_pval(mufl, muplus, 0.95, nfl, nplus, N, K, MSw)
```

```
[1]  -5.06703570 206.31083570   0.00447065
```

```
compute_CI_pval(muplus, muhet, 0.95, nhet, nplus, N, K, MSw)
```

```
[1] -102.9376170  139.2261570    0.3403595
```

```
compute_CI_pval(mufl, muhet, 0.95, nfl, nhet, N, K, MSw)
```

```
[1] 5.632923e+00 2.318994e+02 1.971667e-03
```

```
# Change in Glucose, Base to 60m
MSw = 2238.6
muplus = 33.94821 ; mufl = 130.56350 ; muhet = 16.29601
```

```
compute_CI_pval(mufl, muplus, 0.95, nfl, nplus, N, K, MSw)
```

```
[1] 2.054139e+01 1.726892e+02 2.438761e-04
```

```
compute_CI_pval(muplus, muhet, 0.95, nhet, nplus, N, K, MSw)
```

```
[1] -69.5014034 104.8058034   0.2890487
```

```
compute_CI_pval(mufl, muhet, 0.95, nfl, nhet, N, K, MSw)
```

```
[1] 3.283524e+01 1.956997e+02 5.830143e-05
```

We are 95% confident that change in glucose levels between base and 30 min differs between:

- fl/+ and fl/fl (pval = 0.00447)
- fl/fl and HET (pval = 0.00197)

We are 95% confident that change in glucose levels between base and 60 min differs between:

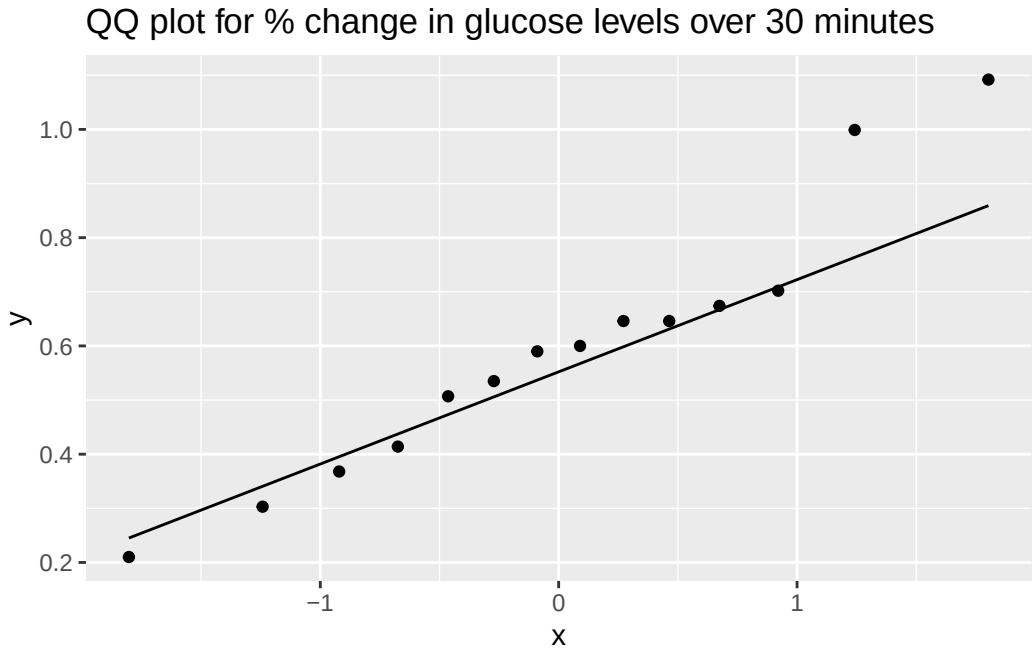
- fl/+ and fl/fl (pval = 0.000244)
- fl/fl and HET (pval = 0.0000583)

We repeat this again for insulin levels.

```
# Checking visually
ggplot(mother, aes(sample=insulin_60)) +
  stat_qq() +
  stat_qq_line() +
  ggtitle("QQ plot for % change in glucose levels over 30 minutes")
```

```
Warning: Removed 2 rows containing non-finite values (`stat_qq()`).
```

```
Warning: Removed 2 rows containing non-finite values (`stat_qq_line()`).
```




```
# Checking mathematically
shapiro.test(mother$insulin_60)
```

```
    Shapiro-Wilk normality test

data:  mother$insulin_60
W = 0.94495, p-value = 0.4854
```

```
mother %>%
  group_by(Maternal_genotype) %>%
  summarise(n = n(), mean = mean(insulin_60, na.rm=TRUE), sd = sd(insulin_60, na.rm=TRUE))
```

```
# A tibble: 3 x 4
  Maternal_genotype     n  mean     sd
  <chr>             <int> <dbl>  <dbl>
1 fl/+                  5 0.802 0.226
2 fl/fl                 7 0.417 0.172
3 HET                   4 0.590 0.0555
```

```r
# First, construct a model including covariates
model_insulin <- lm(insulin_60 ~ Maternal_genotype + num_fetus + percent_body_weight_gain,

# Then perform ANCOVA using the model
anova(model_insulin)
```

```
Analysis of Variance Table

Response: insulin_60
                         Df  Sum Sq  Mean Sq F value  Pr(>F)
Maternal_genotype         2 0.40398 0.201989  6.1403 0.02081 *
num_fetus                 1 0.00505 0.005048  0.1534 0.70438
percent_body_weight_gain  1 0.05706 0.057062  1.7346 0.22037
Residuals                 9 0.29606 0.032896
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

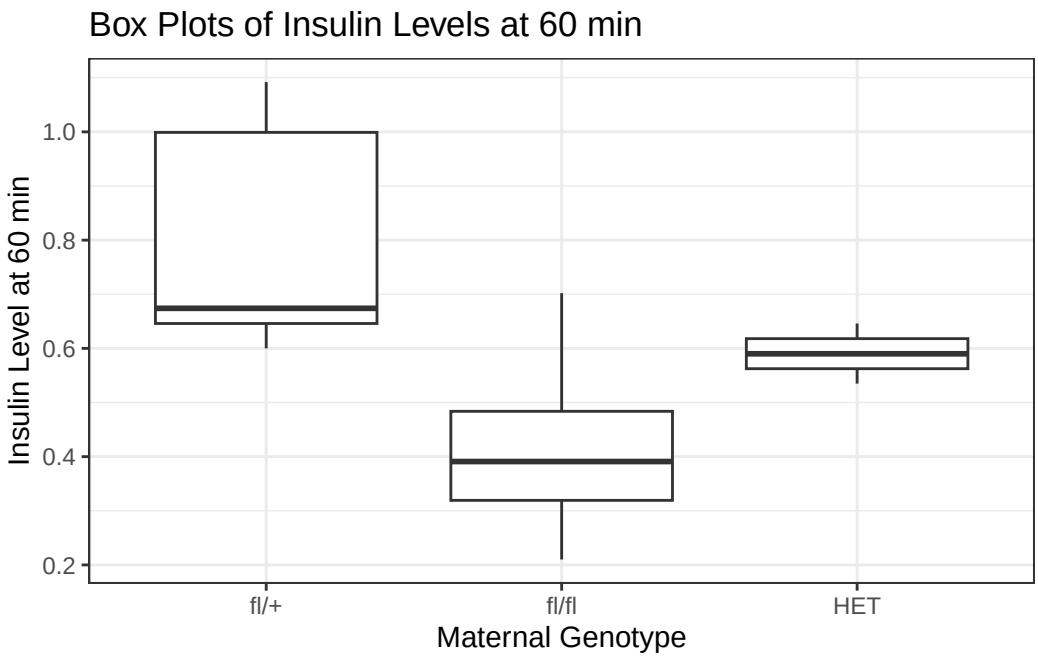
```r
tidy(aov(model_insulin)) %>%
  kable(digits = 3,
        col.names = c("term", "degrees of freedom", "sum of squares",
                      "mean squares", "test statistic", "p-value"))
```

term	degrees of freedom	sum of squares	mean squares	test statistic	p-value
Maternal_genotype	2	0.404	0.202	6.140	0.021
num_fetus	1	0.005	0.005	0.153	0.704
percent_body_weight_gain	1	0.057	0.057	1.735	0.220
Residuals	9	0.296	0.033	NA	NA

The insulin level at 60 minutes proves to be statistically significant. We are 95% confident there is a difference in insulin levels at 60 minutes between the genotypes.

```r
# Group Comparison Insulin 60
ggplot(data = mother, aes(x = Maternal_genotype, y = insulin_60)) +
  geom_boxplot() +
  xlab("Maternal Genotype") +
  ylab("Insulin Level at 60 min") +
  ggtitle("Box Plots of Insulin Levels at 60 min") +
  theme_bw()
```

```
Warning: Removed 2 rows containing non-finite values (`stat_boxplot()`).
```




```
# Insulin 60
MSw = 0.032896
mufl = 0.41733333 ; muplus = 0.8022 ; muhet = 0.5903333
compute_CI_pval(muplus, mufl, 0.95, nfl, nplus, N, K, MSw)
```

```
[1] 0.0932455397 0.6764878003 0.0001450691
```

```
compute_CI_pval(muplus, muhet, 0.95, nhet, nplus, N, K, MSw)
```

```
[1] -0.12222728  0.54596068  0.04081153
```

```
compute_CI_pval(muhet, mufl, 0.95, nfl, nhet, N, K, MSw)
```

```
[1] -0.13916180  0.48516174  0.06402979
```

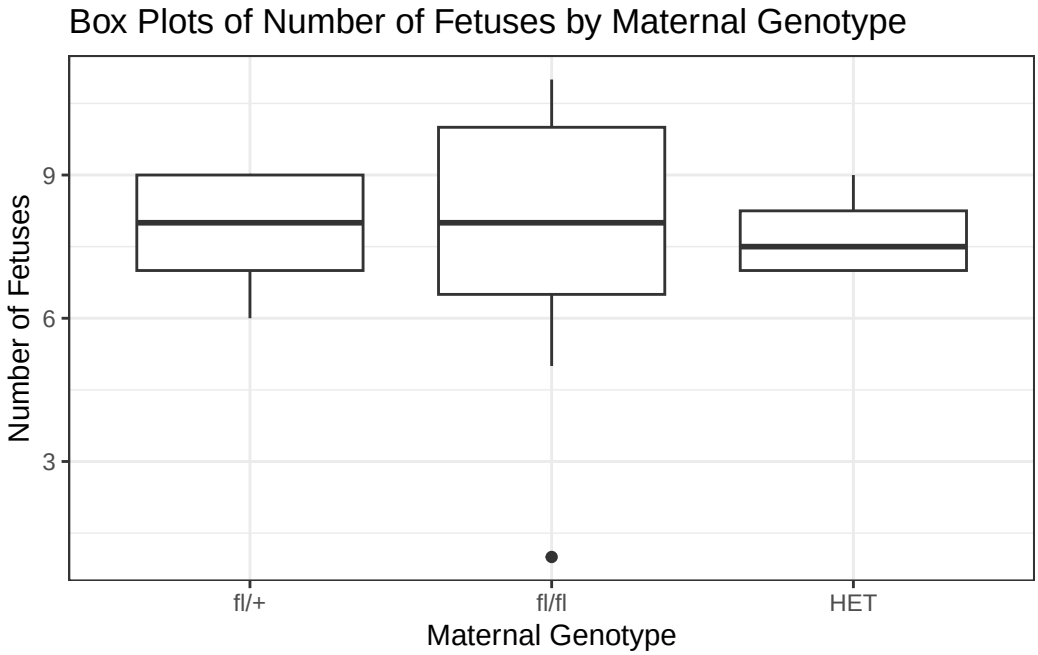
95% Confident that mean insulin differs at 60 minutes between:

- fl/+ and fl/fl (pval = 0.000145)

Now we are interested in assessing the impact of Maternal Genotype on the # of fetuses, # of absorptions, body weight gain, and % body weight gain.

We first do this visually and with linear models.

```
# Num Fetuses
ggplot(data = mother, aes(x = Maternal_genotype, y = num_fetus)) +
  geom_boxplot() +
  xlab("Maternal Genotype") +
  ylab("Number of Fetuses") +
  ggtitle("Box Plots of Number of Fetuses by Maternal Genotype") +
  theme_bw()
```



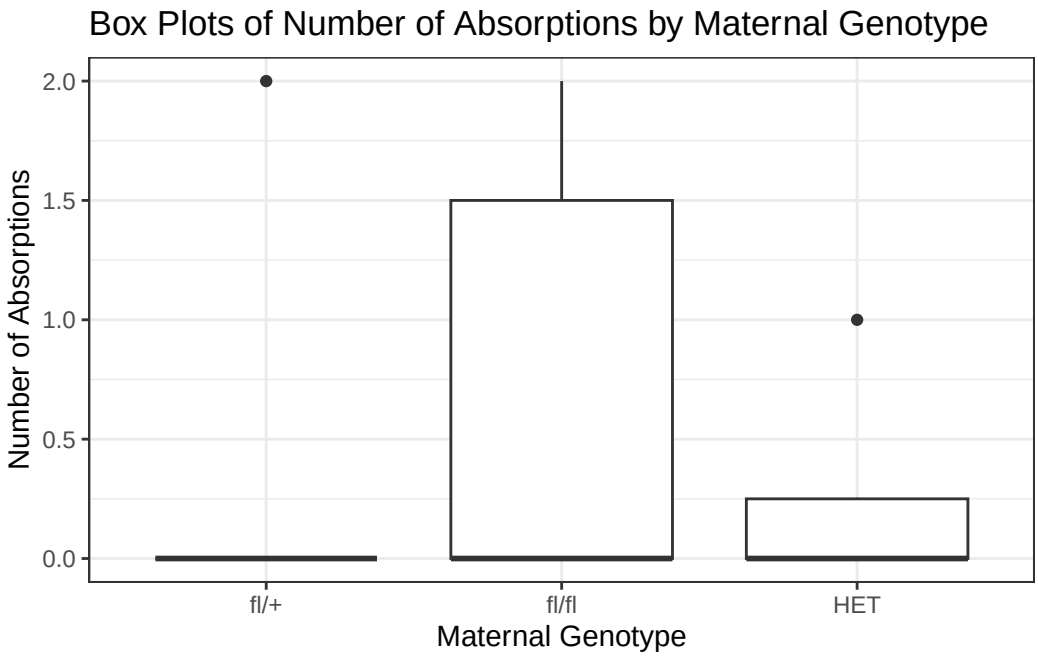

```
pairwiseCI(num_fetus ~ Maternal_genotype, data = mother, conf.level = 1 - (.05/3),
           var.equal = TRUE) %>%
  kable(digits = 3)
```

estimate	lower	upper	comparison
-0.229	-5.056	4.599	fl/fl-fl/+

estimate	lower	upper	comparison
-0.050	-2.501	2.401	HET-fl/+
0.179	-5.250	5.607	HET-fl/fl

No apparent relationship between maternal genotype and # of fetus

```
# Num Absorptions
ggplot(data = mother, aes(x = Maternal_genotype, y = absorptions)) +
  geom_boxplot() +
  xlab("Maternal Genotype") +
  ylab("Number of Absorptions") +
  ggtitle("Box Plots of Number of Absorptions by Maternal Genotype") +
  theme_bw()
```



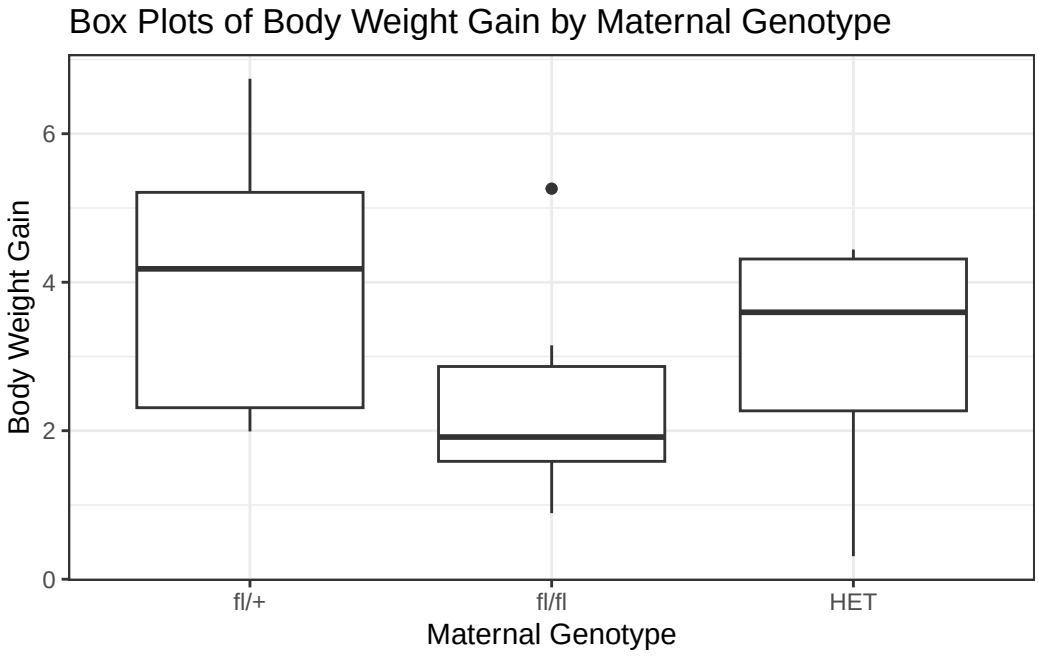

```
pairwiseCI(absorptions ~ Maternal_genotype, data = mother, conf.level = 1 - (.05/3),
           var.equal = TRUE) %>%
  kable(digits = 3)
```

estimate	lower	upper	comparison
0.314	-1.247	1.875	fl/fl-fl/+
-0.150	-1.726	1.426	HET-fl/+
-0.464	-1.988	1.059	HET-fl/fl

No attributable effect of maternal genotype on number of absorptions, almost meaningless analysis though since there's very little data

```
# Body weight gain
ggplot(data = mother, aes(x = Maternal_genotype, y = body_weight_gain)) +
  geom_boxplot() +
  xlab("Maternal Genotype") +
  ylab("Body Weight Gain") +
  ggtitle("Box Plots of Body Weight Gain by Maternal Genotype") +
  theme_bw()
```

```
Warning: Removed 1 rows containing non-finite values (`stat_boxplot()`).
```

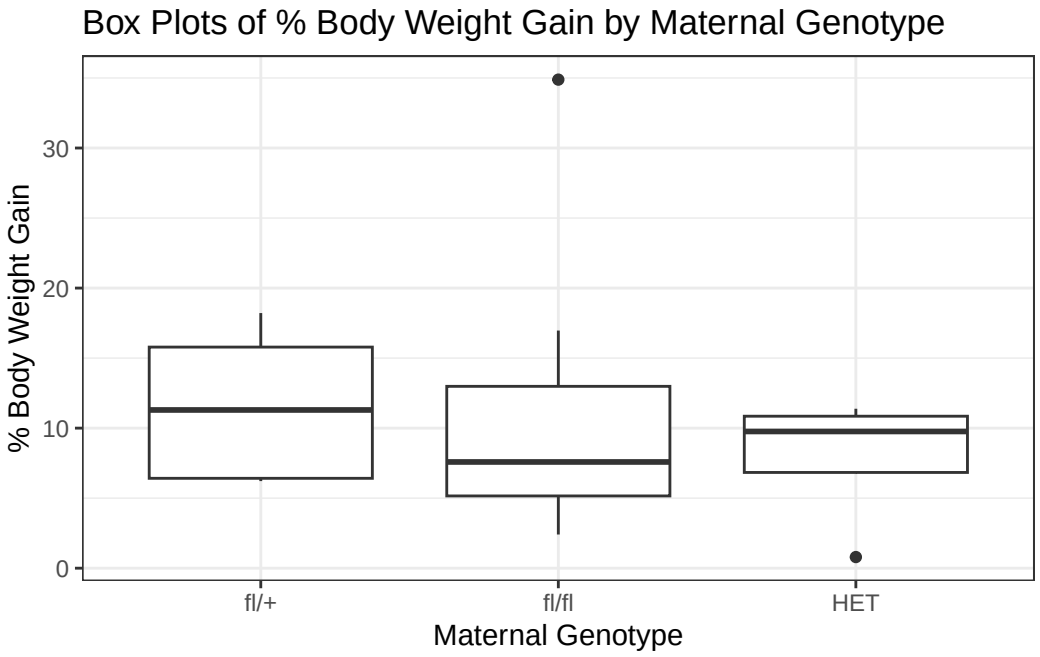


```
pairwiseCI(body_weight_gain ~ Maternal_genotype, data = mother, conf.level = 1 - (.05/3),
           var.equal = TRUE) %>%
  kable(digits = 3)
```

estimate	lower	upper	comparison
-1.646	-4.788	1.496	fl/fl-fl/+
-1.101	-5.205	3.003	HET-fl/+
0.545	-2.772	3.862	HET-fl/fl

No significant effect of maternal genotype on body weight gain.

```
# % body weight gain
ggplot(data = mother, aes(x = Maternal_genotype, y = percent_body_weight_gain)) +
  geom_boxplot() +
  xlab("Maternal Genotype") +
  ylab("% Body Weight Gain") +
  ggtitle("Box Plots of % Body Weight Gain by Maternal Genotype") +
  theme_bw()
```



```
pairwiseCI(percent_body_weight_gain ~ Maternal_genotype, data = mother, conf.level = 1 - (
           var.equal = TRUE) %>%
  kable(digits = 3)
```

estimate	lower	upper	comparison
0.007	-15.751	15.765	fl/fl-fl/+
-3.662	-14.547	7.224	HET-fl/+
-3.668	-21.359	14.022	HET-fl/fl

No significant effect of % bw gain.