Data Analysis #3

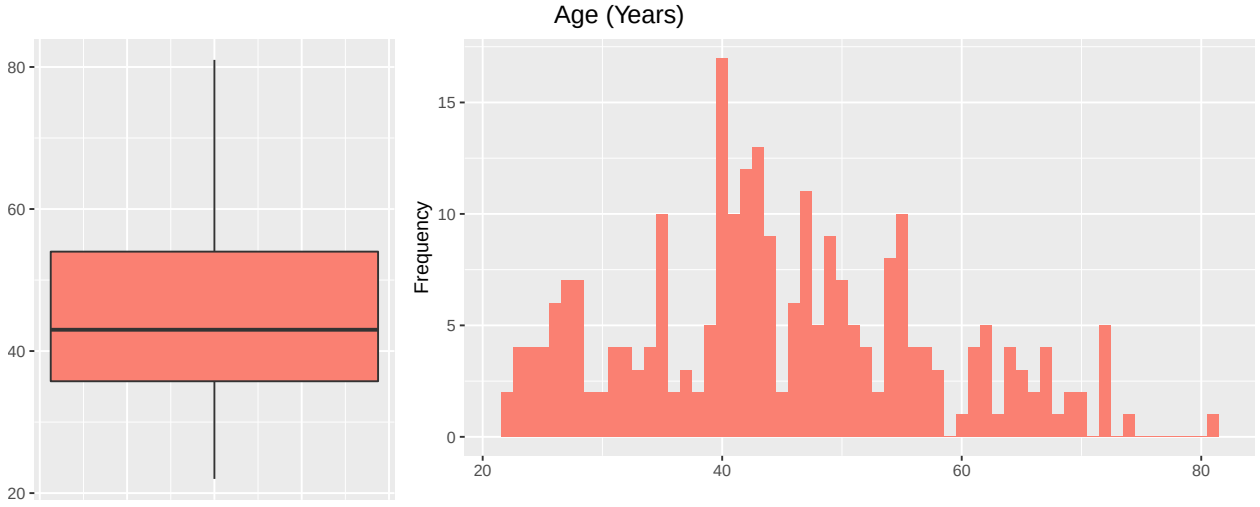
Dustin Leatherman

November 10, 2018

Introduction

In order to determine factors that contribute to high body fat percentages, measurements of 252 men were taken using an underwater weighing technique. The goal of this study is to determine if there is a correlation between average body fat percentage and age in men. The sample contains calculations for two body fat percentage algorithms, Brozek and Siri. Prior to the invention of Dual-energy X-ray Absorptiometry (DXA), body fat percentages were calculated using these algorithms. These are still used to approximate body fat percentage even though DXA exists.

```
# ------------------------------------------------
# age plots
# ------------------------------------------------

bplot_age <- ggplot(fat, aes(y = age, fill = I("salmon"))) +
  geom_boxplot() +
  ylab("") +
  xlab("") +
  theme(axis.text.x = element_blank(), axis.title.x = element_blank(), axis.ticks.x = element_blank())

hist_age <- ggplot(fat, aes(x = age, fill = I("salmon"))) +
  geom_histogram(binwidth = 1) +
  xlab("") +
  ylab("Frequency")

grid.arrange(bplot_age, hist_age,
             widths = c(1, 2),
             top = textGrob("Age (Years)", gp=gpar(fontsize=14,font=1),just=c("center")))
```

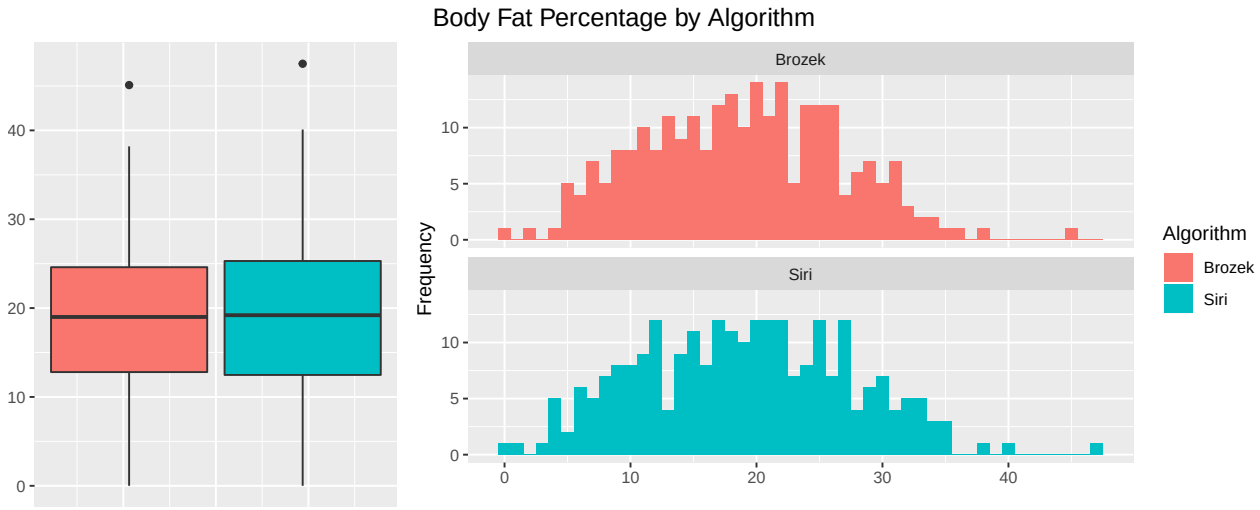



```
kable(fat %>%
        summarise_at(c("age"),
             funs(
```

```
                mean,
                sd,
                min,
                "25%"=quantile(age, probs = 0.25),
                median,
                "75%"=quantile(age, probs = 0.75),
                max,
                length)
            ),
      col.names = c("Mean", "Std. Dev", "Min", "1st Quartile", "Median", "3rd Quartile", "Max", "Sample
      align = c('l'),
      digits = 4
)
```

Mean	Std. Dev	Min	1st Quartile	Median	3rd Quartile	Max	Sample Size
44.8849	12.602	22	35.75	43	54	81	252

```
# ------------------------------------------------
# BodyFat plots
# ------------------------------------------------
# Pivot brozek and siri calculations for ease of display
# Capitalize the first letter of the Algorithm for display purposes
gatherBodyFat <- fat %>%
  gather(Algorithm, BodyFat, brozek:siri) %>%
  mutate(Algorithm=capitalize(Algorithm))

bplot_bodyfat <- ggplot(gatherBodyFat, aes(y = BodyFat, fill = Algorithm)) +
  geom_boxplot(show.legend = FALSE) +
  ylab("") +
  xlab("") +
  theme(axis.text.x = element_blank(), axis.title.x = element_blank(), axis.ticks.x = element_blank())

hist_bodyfat <- ggplot(gatherBodyFat, aes(x = BodyFat, fill = Algorithm)) +
  geom_histogram(binwidth = 1) +
  xlab("") +
  ylab("Frequency") +
  facet_wrap(~Algorithm, ncol = 1)

grid.arrange(bplot_bodyfat, hist_bodyfat,
             widths = c(1, 2),
             top = textGrob("Body Fat Percentage by Algorithm", gp=gpar(fontsize=14,font=1),just=c("cen
```




```
kable(gatherBodyFat %>% group_by(Algorithm) %>% summarise_at(c("BodyFat"),
              funs(
                 mean,
                 sd,
                 min,
                 "25%"=quantile(BodyFat, probs = 0.25),
                 median,
                 "75%"=quantile(BodyFat, probs = 0.75),
                 max,
                 length)),
      col.names = c("","Mean", "Std. Dev", "Min", "1st Quartile", "Median", "3rd Quartile", "Max", "Samp
      align = c('l'),
      digits = 4
)
```

	Mean	Std. Dev	Min	1st Quartile	Median	3rd Quartile	Max	Sample Size
Brozek	18.9385	7.7509	0	12.800	19.0	24.6	45.1	252
Siri	19.1508	8.3687	0	12.475	19.2	25.3	47.5	252

Observations

Brozek & Siri

The Brozek and Siri calculations are fairly close in this sample. The Brozek histogram appears to be slightly more normal than Siri. This is confirmed with a Shapiro-Wilk tests for Brozek (p-value = 0.2747) and Siri (p-value = 0.1649) calculations. From the casual observer, there doesn't appear to be a strong reason to use one over the other for approximations. In the interest of covering all of our bases, both calculations will be presented

Outliers

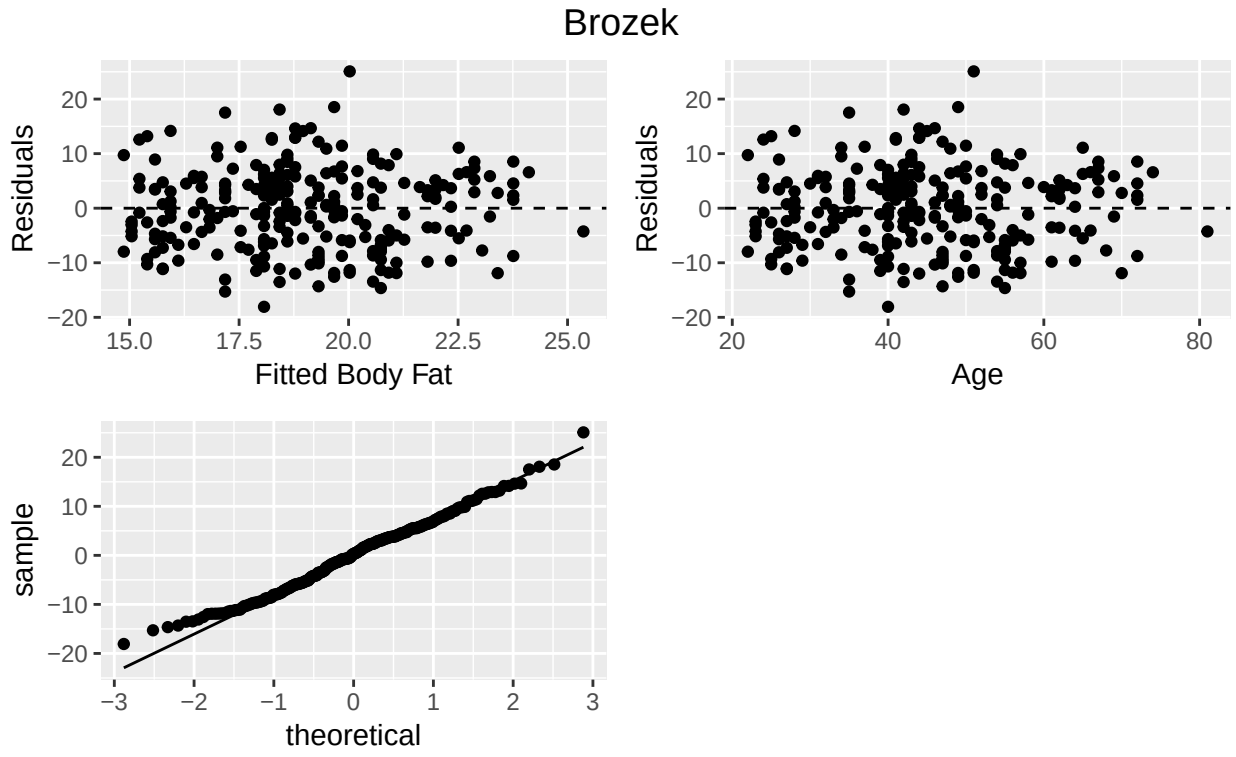
There is a body fat percentage of 0 which is impossible. Male body fat percentages under 2 cause multiple organ failures which would make the subject unlikely to be able to partake in the study. There is not any indication whether this value is an error though it is suspect; thus, this value will be removed and modeled separately from the full dataset.

Methods

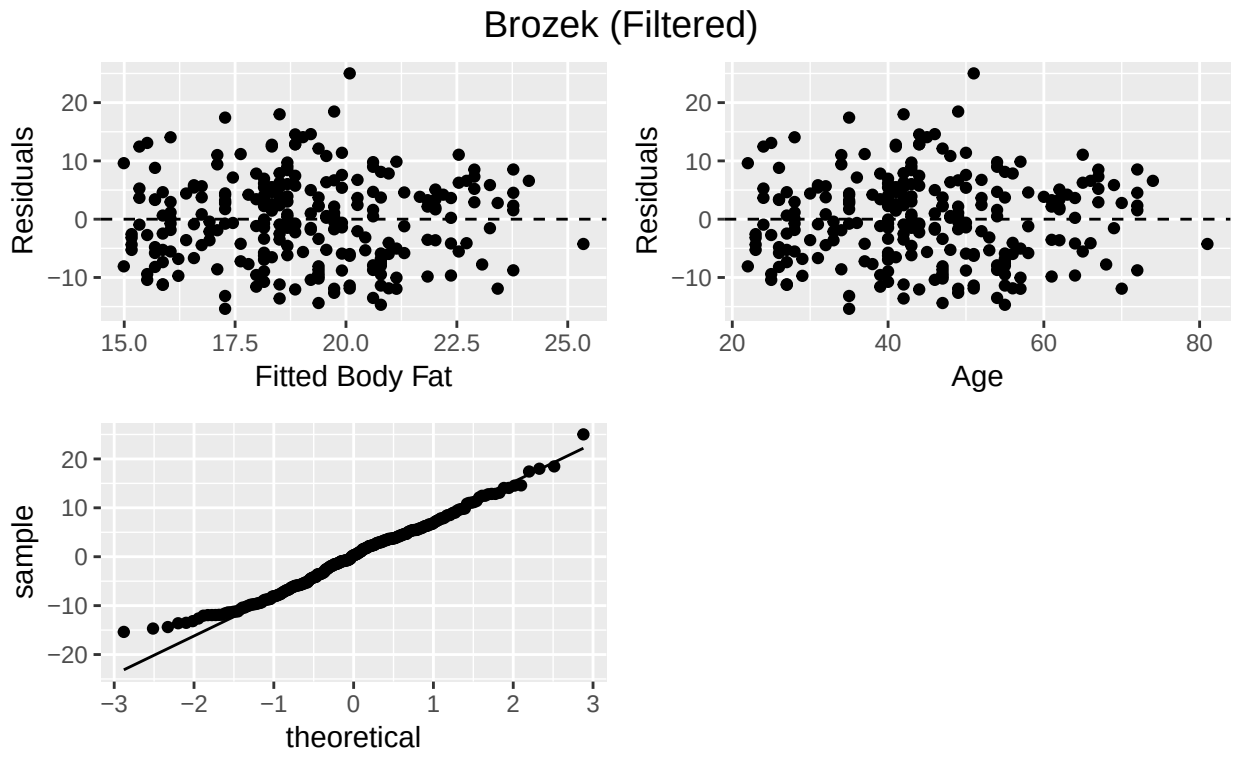
Are body fat and age correlated? For determining correlation between two continuous variables, linear regression is the most appropriate model. The main assumption that needs to be proved is independence between outcomes under experimental conditions. This is done through analyzing residual charts.

Residual Plots

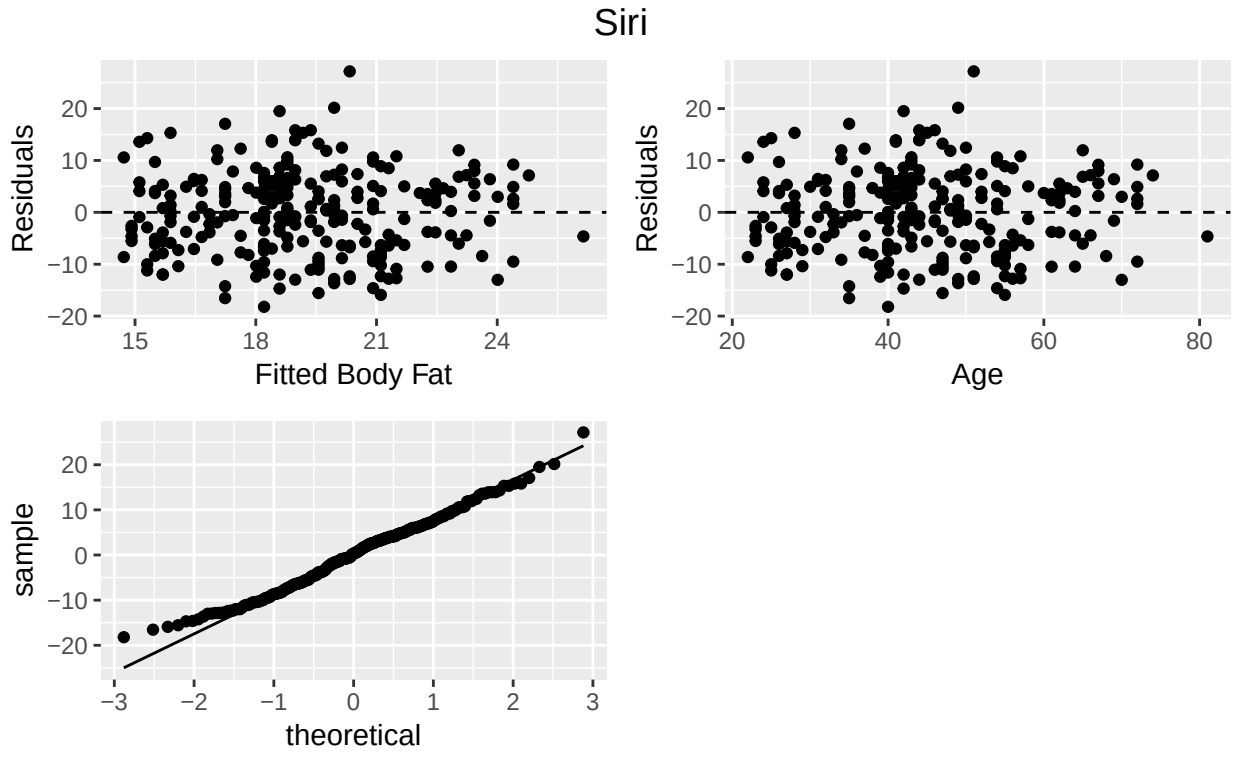
```
#----------------------------------------
# Brozek
#----------------------------------------

model.brozek <- lm(brozek ~ age, data = fat)

# Residual vs fitted values
brozek.resfit <- ggplot(model.brozek) +
  geom_point(aes(x = .fitted, y = .resid)) +
  geom_hline(yintercept = 0, lty = 2) +
  xlab("Fitted Body Fat") +
  ylab("Residuals")

# Residual vs predictor (age)
brozek.predict <- ggplot(model.brozek) +
  geom_point(aes(x = age, y = .resid)) +
  geom_hline(yintercept = 0, lty = 2) +
  xlab("Age") +
  ylab("Residuals")

# Normality plot against residuals
brozek.qq <- ggplot(aes(sample = .resid), data = model.brozek) +
  stat_qq() +
  stat_qq_line()

grid.arrange(brozek.resfit, brozek.predict, brozek.qq,
             widths = c(1,1),
             ncol = 2,
             top = textGrob("Brozek", gp=gpar(fontsize=14,font=1),just=c("center")))
```



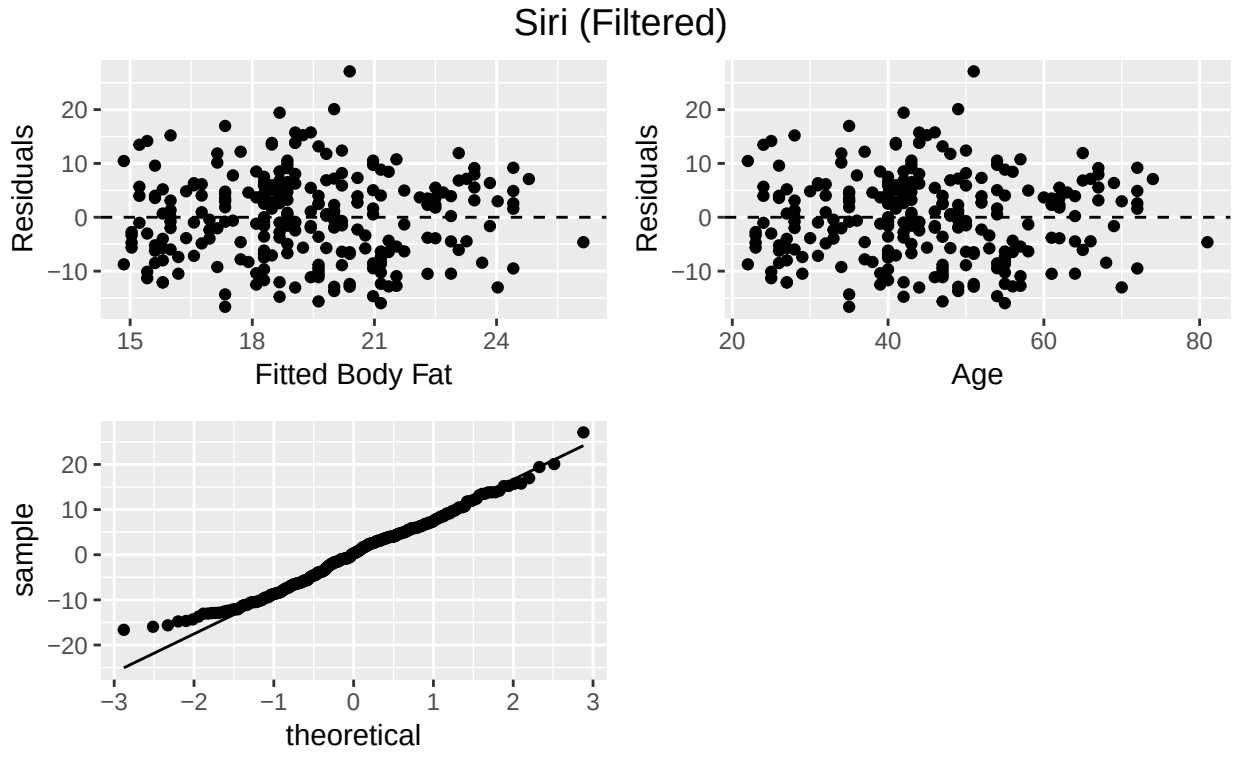

```
#----------------------------------------
# Brozek with body fat % > 0
#----------------------------------------
filterBro <- fat %>% filter(brozek > 0)
model.brozek.no0 <- lm(brozek ~ age, data = filterBro)

# Residual vs fitted values
brozek.no0.resfit <- ggplot(model.brozek.no0) +
  geom_point(aes(x = .fitted, y = .resid)) +
  geom_hline(yintercept = 0, lty = 2) +
  xlab("Fitted Body Fat") +
  ylab("Residuals")

# Residual vs predictor (age)
brozek.no0.predict <- ggplot(data = model.brozek.no0) +
  geom_point(aes(x = age, y = .resid)) +
  geom_hline(yintercept = 0, lty = 2) +
  xlab("Age") +
  ylab("Residuals")

# Normality plot against residuals
brozek.no0.qq <- ggplot(aes(sample = .resid), data = model.brozek.no0) +
  stat_qq() +
  stat_qq_line()

grid.arrange(brozek.no0.resfit, brozek.no0.predict, brozek.no0.qq,
             widths = c(1,1),
             ncol = 2,
             top = textGrob("Brozek (Filtered)", gp=gpar(fontsize=14,font=1),just=c("center")))
```




```
#----------------------------------------
# Siri
#----------------------------------------

model.siri <- lm(siri ~ age, data = fat)

# Residual vs fitted values
siri.resfit <- ggplot(model.siri) +
  geom_point(aes(x = .fitted, y = .resid)) +
  geom_hline(yintercept = 0, lty = 2) +
  xlab("Fitted Body Fat") +
  ylab("Residuals")

# Residual vs predictor (age)
siri.predict <- ggplot(model.siri) +
  geom_point(aes(x = age, y = .resid)) +
  geom_hline(yintercept = 0, lty = 2) +
  xlab("Age") +
  ylab("Residuals")

# Normality plot against residuals
siri.qq <- ggplot(aes(sample = .resid), data = model.siri) +
  stat_qq() +
  stat_qq_line()

grid.arrange(siri.resfit, siri.predict, siri.qq,
             widths = c(1,1),
             ncol = 2,
             top = textGrob("Siri", gp=gpar(fontsize=14,font=1),just=c("center")))
```




```
#----------------------------------------
# Siri with body fat % > 0
#----------------------------------------
filterSiri <- fat %>% filter(siri > 0)
model.siri.no0 <- lm(siri ~ age, data = filterSiri)

# Residual vs fitted values
siri.no0.resfit <- ggplot(model.siri.no0) +
  geom_point(aes(x = .fitted, y = .resid)) +
  geom_hline(yintercept = 0, lty = 2) +
  xlab("Fitted Body Fat") +
  ylab("Residuals")

# Residual vs predictor (age)
siri.no0.predict <- ggplot(data = model.siri.no0) +
  geom_point(aes(x = age, y = .resid)) +
  geom_hline(yintercept = 0, lty = 2) +
  xlab("Age") +
  ylab("Residuals")

# Normality plot against residuals
siri.no0.qq <- ggplot(aes(sample = .resid), data = model.siri.no0) +
  stat_qq() +
  stat_qq_line()

grid.arrange(siri.no0.resfit, siri.no0.predict, siri.no0.qq,
             widths = c(1,1),
             ncol = 2,
             top = textGrob("Siri (Filtered)", gp=gpar(fontsize=14,font=1),just=c("center")))
```



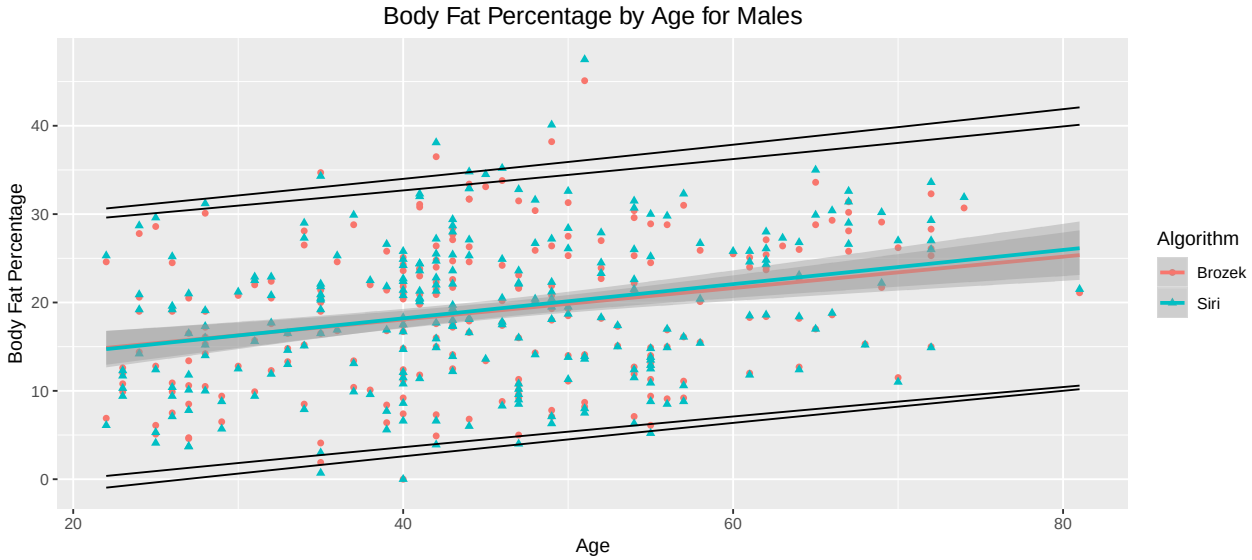
Interpreting the Residuals

Residual vs Fitted plots for each of the 4 categories look similar. There is no discernable pattern within the **Residual vs Fitted Plots** which indicate that there are no obvious issues with using this model. Removing body fat percentages equal to 0 do not appear to have any effect on the residual plots. There are also no discernable patterns between the Residuals vs Age plots. The normality plots indicate that this model is normal except for some deviations in extremely small values. In the models where body fat equal to 0 have been removed, the value for the smallest residual shifts further away from line which indicates that value is less normal. This is expected and is acceptable since it doesn't affect the model in any significantly.

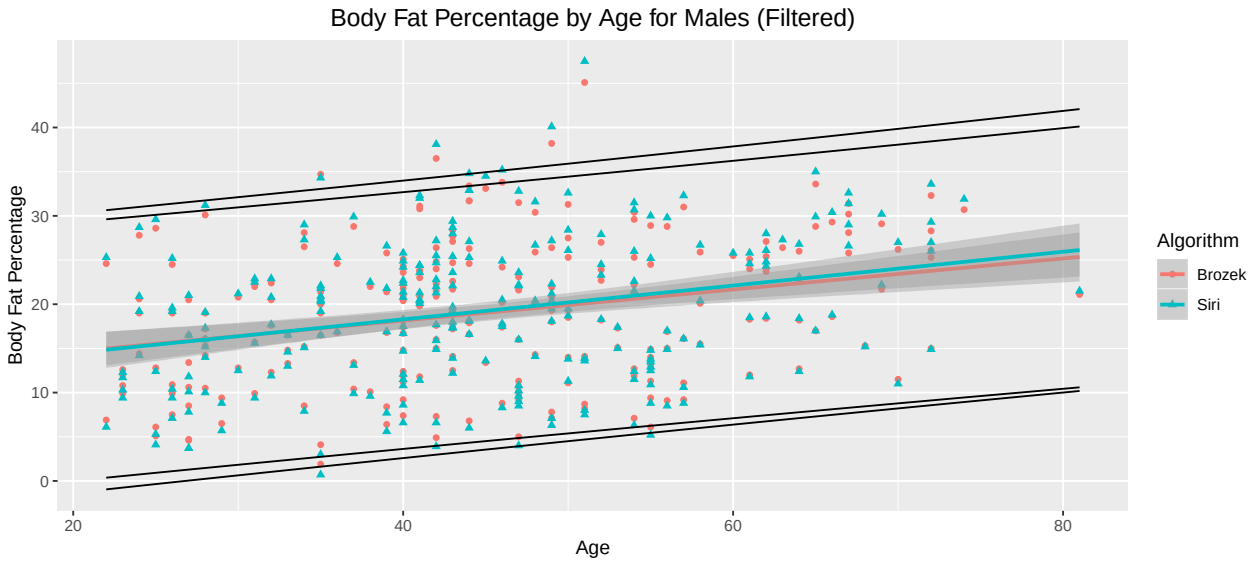
Summary

```
gatherWithPred <- tidypredict_to_column(gatherBodyFat,
                                        model.brozek.no0,
                                        add_interval = T,
                                        vars = c("brozek_fit","brozek_upper","brozek_lower"))

gatherWithPred_siri <- tidypredict_to_column(gatherWithPred,
                                             model.siri.no0,
                                             add_interval = T,
                                             vars = c("siri_fit","siri_upper","siri_lower"))

ggplot(gatherWithPred_siri, aes(x = age, y = BodyFat, shape = Algorithm, color = Algorithm)) +
  geom_point() +
  geom_smooth(method = "lm") +
  ylab("Body Fat Percentage") +
  xlab("Age") +
  ggtitle("Body Fat Percentage by Age for Males") +
  theme(plot.title = element_text(size = 14, hjust = 0.5)) +
```

```
  # for some reason, I was not able to get the prediction interval lines to match
  # changing the color for a given line shifted all the colors 1 and added it to the legend
  # It is not a far leap to determine which prediction interval belongs to which Algorithm
  geom_line(aes(x = age, y = brozek_lower), inherit.aes = FALSE) +
  geom_line(aes(x = age, y = brozek_upper), inherit.aes = FALSE) +
  geom_line(aes(x = age, y = siri_lower), inherit.aes = FALSE) +
  geom_line(aes(x = age, y = siri_upper), inherit.aes = FALSE)
```




```
# Remove values where BodyFat <= 0
ggplot(gatherWithPred_siri %>% filter(BodyFat > 0), aes(x = age, y = BodyFat, shape = Algorithm, color =
  geom_point() +
  geom_smooth(method = "lm") +
  ylab("Body Fat Percentage") +
  xlab("Age") +
  ggtitle("Body Fat Percentage by Age for Males (Filtered)") +
  theme(plot.title = element_text(size = 14, hjust = 0.5)) +
  geom_line(aes(x = age, y = brozek_lower), inherit.aes = FALSE) +
  geom_line(aes(x = age, y = brozek_upper), inherit.aes = FALSE) +
  geom_line(aes(x = age, y = siri_lower), inherit.aes = FALSE) +
  geom_line(aes(x = age, y = siri_upper), inherit.aes = FALSE)
```



Brozek

```
kable(
  tidy(model.brozek) %>% inner_join(model.brozek.no0 %>% tidy, by = c("term")),
  digits = 4,
  align = c('l'),
  col.names = c("","Est", "Std. Err.", "T", "p-value", "Estimate (Filt)", "Std. Err (Filt)", "T (Filt)"
  caption = "Body Fat Percentage (Brozek) by Age"
)
```

Table 3: Body Fat Percentage (Brozek) by Age

	Est	Std. Err.	T	p-value	Estimate (Filt)	Std. Err (Filt)	T (Filt)	p-value (Filt)
(Intercept)	10.9555	1.7358	6.3116	0	11.1273	1.7199	6.4698	0
age	0.1779	0.0372	4.7763	0	0.1756	0.0369	4.7626	0

There is convincing evidence that there is an association between between body fat percentage (Brozek) and age (filtered model p-value = 3.25e-06, unmodified model p-value = 3.04e-6). A newborn male is estimated to have 11.12734 and 10.9555 body fat percentage for the filtered and unfiltered model respectively. The estimated increase in body fat percentage per year is 0.1756 and 0.1779 for the filtered and unfiltered model respectively. With 95% confidence, an increase of 1 year of age is associated with an increase in mean Body Fat Percentage via the Brozek algorithim between 0.1045 and 0.2512 percent for the unfiltered model and 0.103 and 0.2483 percent for the filtered model. Taking the difference in the confidence intervals for the filtered and unfiltered models show that the filtered model has a slightly smaller range and standard error. **While both models are valid, the filtered model is more accurate.**

```
confInt <- predict(model.brozek.no0, data.frame(age=28), interval = "confidence") %>%
  tidy %>%
  mutate(Interval="Confidence")

predInt <- predict(model.brozek.no0, data.frame(age=28), interval = "prediction") %>%
  tidy %>%
  mutate(Interval="Prediction")
```

```
kable(
  confInt %>%
    union(predInt) %>%
    select(Interval,fit,lwr,upr),
  col.names = c("Interval","Estimate (Filt)", "LCL", "UCL"),
  align = c('l'),
  digits = 4,
  caption = "Body Fat % (Brozek) Intervals"
)
```

Table 4: Body Fat % (Brozek) Intervals

Interval	Estimate (Filt)	LCL	UCL
Prediction	16.045	1.4676	30.6224
Confidence	16.045	14.5138	17.5763

With 95% confidence, the average body fat percentage (Brozek) for a 28 year old male is between 14.5138 and 17.5763 (Using filtered model). A 95% prediction interval for the body fat percentage of a 28 year old male is between 1.4676 and 30.6224.

Siri

```
kable(
  tidy(model.siri) %>%
    inner_join(model.siri.no0 %>% tidy, by = c("term")),
  digits = 4,
  align = c('l'),
  col.names = c("","Est", "Std. Err", "T", "p-value", "Estimate % (Filt)", "Std. Err. % (Filt)", "T (Fi
  caption = "Body Fat Percentage (Siri) by Age"
)
```

Table 5: Body Fat Percentage (Siri) by Age

	Est	Std. Err	T	p-value	Estimate % (Filt)	Std. Err. % (Filt)	T (Filt)	p-value (Filt)
(Intercept)	10.4633	1.8728	5.5870	0	10.6364	1.8585	5.7230	0
age	0.1936	0.0402	4.8175	0	0.1913	0.0399	4.8007	0

There is convincing evidence that there is an association between between body fat percentage (Siri) and age (filtered model p-value = 2.73e-6, unmodified model p-value = 2.52e-6). With 95% confidence, a newborn male is estimated to have an average of 10.6364 and 10.4633 percent body fat for the filtered and unfiltered model respectively. The estimated average increase in body fat percentage per year is 0.1913 and 0.1936 for the filtered and unfiltered model respectively. With 95% confidence, an increase of 1 year of age is associated with an increase in mean Body Fat Percentage via the Siri algorithim between 0.113 and 0.27 percent for the filtered model and 0.114 and 0.273 percent for the unfiltered model. Taking the difference in the confidence intervals for the filtered and unfiltered models show that the filtered model has a slightly smaller range and standard error. **While both models are valid, the filtered model is more accurate.**

```
confInt <- predict(model.siri.no0, data.frame(age=28), interval = "confidence") %>%
  tidy %>%
  mutate(Interval="Confidence")
```

```r
predInt <- predict(model.siri.no0, data.frame(age=28), interval = "prediction") %>%
  tidy %>%
  mutate(Interval="Prediction")

kable(
  confInt %>%
    union(predInt) %>%
    select(Interval,fit,lwr,upr),
  col.names = c("Interval","Estimate (Filt)", "LCL", "UCL"),
  align = c('l'),
  digits = 4,
  caption = "Body Fat % (Siri) Intervals"
)
```

Table 6: Body Fat % (Siri) Intervals

Interval	Estimate (Filt)	LCL	UCL
Prediction	15.9931	0.2405	31.7457
Confidence	15.9931	14.3384	17.6478

With 95% confidence, the average body fat percentage (Siri) for a 28 year old male is between 14.3384 and 17.6478 (Using filtered model). A 95% prediction interval for the body fat percentage of a 28 year old male is between 0.2405 and 31.7457.

Final Thoughts

A filtered and unfiltered model have been presented against the Siri and Brozek algorithms for body fat percentage to help the reader understand the impacts of these findings. It is dangerous to interpret the intercept of these models as it is out of the scope. The youngest subject in the sample is 22 and the oldest is 81 so any ages outside of this range are considered extrapolations so interpretation should be avoided.

This linear regression model is not the most appropriate model for body fat percentage. It has a low adjusted r squared value (0.08) which indicates that there is large amount of variability that is not explained by the model. Since there is only one predictor in this model, there may be other significant factors that contribute to body fat percentage. This may also cause the prediction intervals for 28 year old males to be inaccurate. Using multiple parameters, such as neck, chest, and stomach measurements, may provide a better approximation of body fat percentage in males.