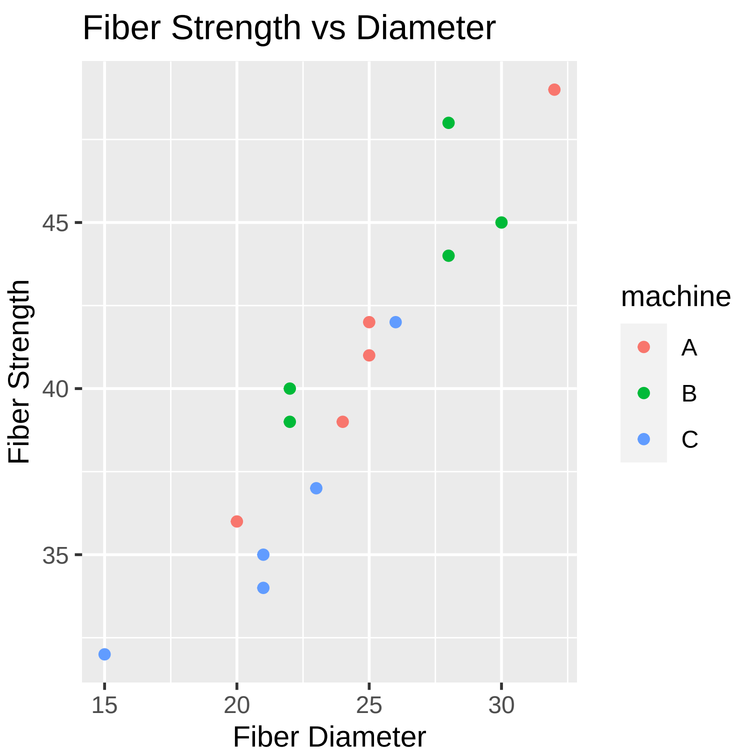
Homework Four

Josh Virene

Problem 1

```
## # A tibble: 3 x 6
##   machine mean_strength sd_strength mean_diameter sd_diameter     n
##   <chr>           <dbl>       <dbl>         <dbl>       <dbl> <int>
## 1 A                41.4        4.83          25.2        4.32     5
## 2 B                43.2        3.70          26          3.74     5
## 3 C                36          3.81          21.2        4.02     5
```



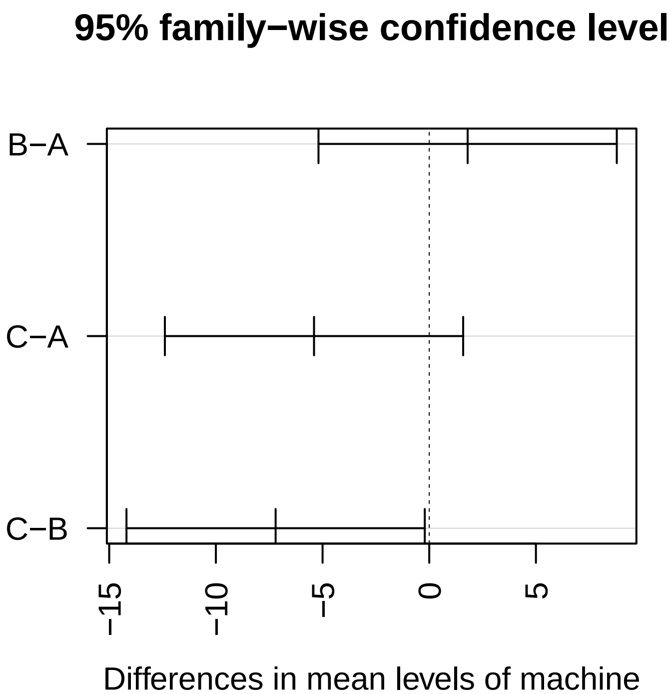
Problem 2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
machine	2	140.4	70.20000	4.08932	0.0442319
Residuals	12	206.0	17.16667	NA	NA

Based on the results of this ANOVA table, **there is evidence of different population mean breaking strengths for the different machines**. This is because the table returns a p-value that is < 0.05. The ANOVA model tests against the null hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ Because the p value is less than 0.05, this suggests that we will reject the null hypothesis that population mean breaking strengths are equal.

Problem 3

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = strength ~ machine, data = fiber)
##
## $machine
##     diff        lwr        upr     p adj
## B-A  1.8  -5.190957  8.7909567 0.7753939
## C-A -5.4 -12.390957  1.5909567 0.1402027
## C-B -7.2 -14.190957 -0.2090433 0.0434397
```




```
##  contrast estimate   SE df t.ratio p.value
##  A - B        -1.8 2.62 12  -0.687  0.7754
##  A - C         5.4 2.62 12   2.061  0.1402
##  B - C         7.2 2.62 12   2.748  0.0434
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

Based on the Tukey multiple comparisons of means analysis and the graph above, the only statistically signficant different mean breaking strength pairwise comparisons was between machines B and C. The other two pairwise comparisons (between machines A vs. B and A vs. C) had confidence intervals that contained zero, as well as p values that are greater than 0.05 and therefore were not statistically significant. From this we conclude that the only significant difference is between machines B and C, and that the standard error of each estimated difference is 2.62, and this is the explanation for the differences between the machines.

Problem 4

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
machine	2	140.4	70.20000	4.08932	0.0442319
Residuals	12	206.0	17.16667	NA	NA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
machine	2	140.40000	70.200000	27.59248	5.17e-05
diameter	1	178.01411	178.014110	69.96938	4.30e-06
Residuals	11	27.98589	2.544172	NA	NA

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
12	206.00000	NA	NA	NA	NA
11	27.98589	1	178.0141	69.96938	4.3e-06

These two models are different in that the reduced model is a one way ANOVA model that does not include diameter as a covariate it has strength as the independent / response variable and only machines as the dependent variable; it is considered a partial F test. In the ANCOVA model, this includes diameter as a co variate and machine as a factor variable and still strength as a response variable. Looking at the results, the important statistics to consider are the F tests, MS_{error}, and $MS_{treatment}$. The f statistic is the "signal to noise ratio", which indicates the difference in means for each treatment group. In the ANOVA model, the F statistic for machine is 4.08, which is very low compared to the F statistic in the ANCOVA model. In the ANCOVA model, the F statistic for machine is 27.593, meaning that the signal to noise ratio has increased (MS_{error} decreased between these models). Based on these differences, it is clear that adding in diameter has reduced the noise (MS_{error}) in the model and the ANCOVA model is a better fit.

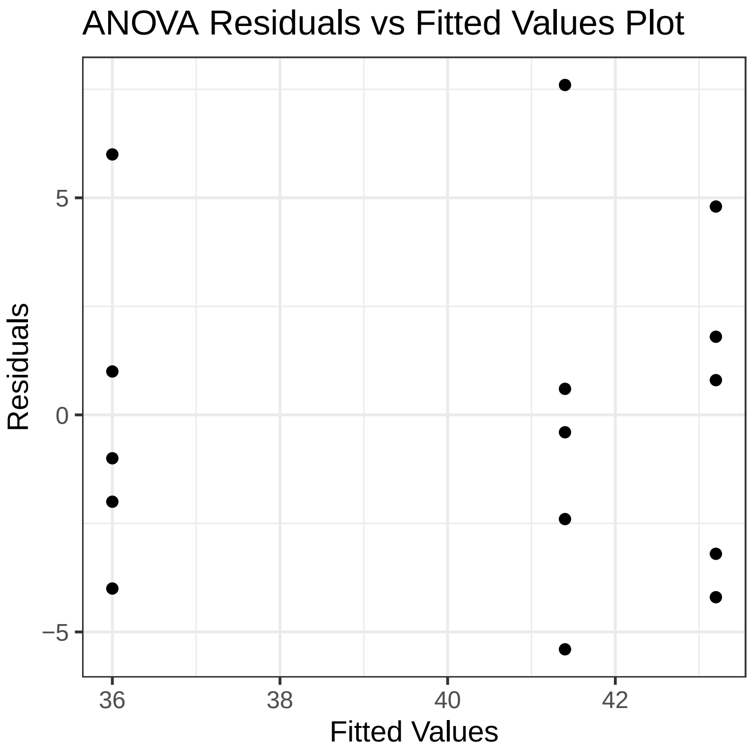
Problem 5

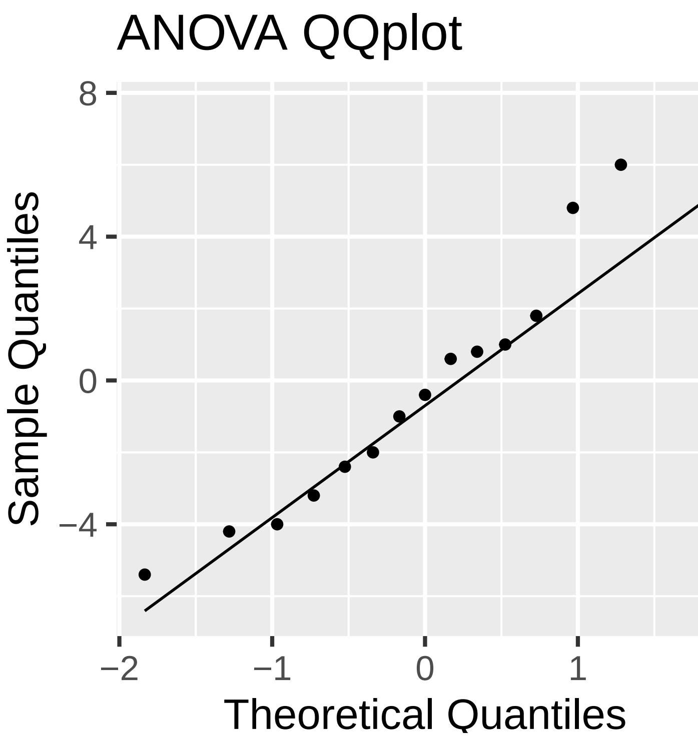
```
##
## Call:
## lm(formula = strength ~ machine + diameter, data = fiber)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0160 -0.9586 -0.3841  0.9518  2.8920
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.360      2.961   5.862 0.000109 ***
## machineB       1.037      1.013   1.024 0.328012
## machineC      -1.584      1.107  -1.431 0.180292
## diameter       0.954      0.114   8.365 4.26e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.595 on 11 degrees of freedom
## Multiple R-squared:  0.9192, Adjusted R-squared:  0.8972
## F-statistic: 41.72 on 3 and 11 DF,  p-value: 2.665e-06
```

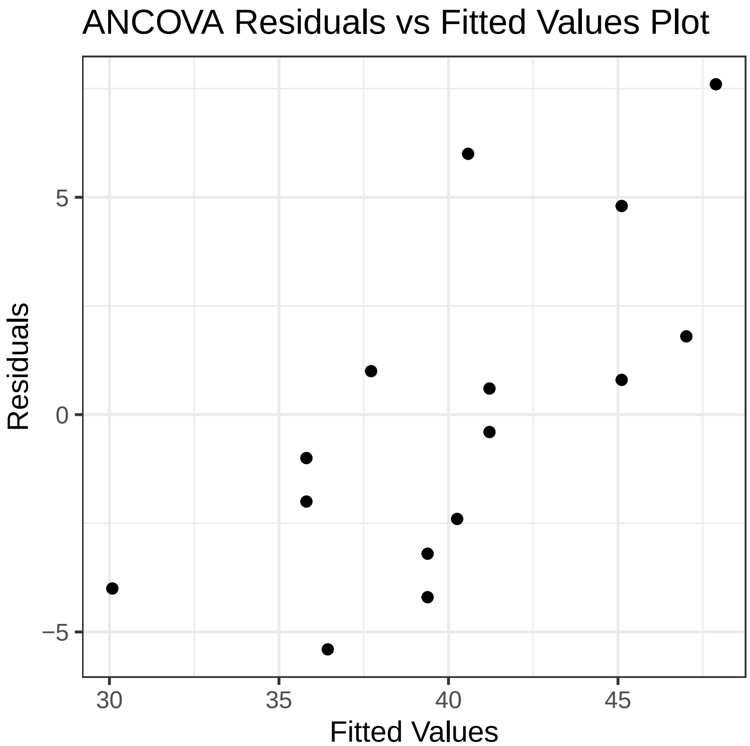
```
## diameter = 24.1:
##  contrast estimate   SE df t.ratio p.value
##  A - B       -1.04 1.01 11  -1.024  0.5781
##  A - C        1.58 1.11 11   1.431  0.3596
##  B - C        2.62 1.15 11   2.283  0.1005
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

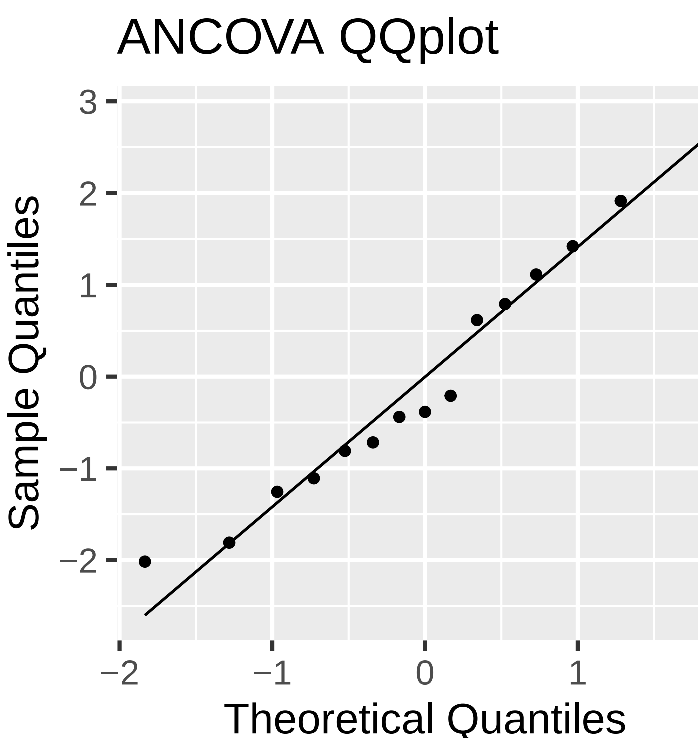
a. **Estimated mean breaking strength for each machine, and their standard errors:** Machine A has a breaking strength of 17.360 units, and a standard error of 2.961. Machine B has a breaking strength of 17.360 + 1.037 = 18.397 units, and a standard error of 1.013. Machine C has a breaking strength of 17.360 - 1.584 = 15.776 units, and a standard error of 1.107.
b. **The estimated difference in mean breaking strength for each pair of machines and the standard error of these differences:** For the comparison A vs. B, the difference in means is -1.04 units, and the standard error is 1.01. For the comparison A vs. C, the difference in means is 1.58 units, and the standard error is 1.11. For the comparison B vs. C, the difference in means is 2.62 units, and the standard error is 1.15.
c. **Test statistics and p values for each pairwise comparison** For the comparison A vs. B, the test statistic is -1.024, and the p value is 0.5781. For the comparison A vs.C, the test statistic is 1.431, and the p value is 0.3596. For the comparison B vs. C, the test statistic is 2.283, and the p value is 0.1005. **Note important differences** None of these comparisons are significant at the 95% family-wise confidence level. This is different than the comparison in question three because there was a statistically signficant pairwise comparison between machines B and C in that model. A reason for the lack of statistical significance for each of these pairwise comparisons is that R is estimating an average value for thread diameter, so the interpretation in this model is comparing the difference in means across machines, given an average value for thread diameter. **The model here assumes that thread diameter is equal across machine types. Interpretation- how the ANOVA and ANCOVA interpretations differ** The ANOVA model only considers the affect of machine type on the strength of the fiber, whereas the ANCOVA model considers both the impact of the machine type, as well as the diameter of the fiber on the strength of the fiber. In this model, it is important to recognize that there is no interaction between machine type and fiber diameter (i.e., fiber strength does not depend on how machine type in conjunction with fiber diameter change together.)

Problem 6









a. ANOVA plots: The assumption of constant variance is not violated, the residuals versus fitted values plot shows that the residuals are fairly constant across fitted values. The assumption of normality is violated in this plot, there is a right tail on the QQ plot. Due to the central limit theorem however, this may not be a huge concern because the coefficients are normally distributed.

b. ANCOVA plots: The assumption of constant variance in the ANCOVA residuals versus fitted values plot is violated because the residuals are not constant across fitted values; their distribution is higher in the middle of the plot than on the left and right ends. The assumption of normality is met, the points all fall very close to the 1-1 line.

Problem 7

In the ANOVA residuals versus fitted values plot, the residuals are stacked in columns, while for the ANCOVA residuals versus fitted values plot, the residuals are spread horizontally. The reason for this is that in the ANOVA model, the linear model that ANOVA draws on is strength ~ machines, and machines is a categorical variable with discrete values "A", "B", "C", so the residuals can only fall into these three categories. In the ANCOVA model, this draws on a linear model with predictor variables machines and diameter. Machines is still a categorical discrete variable, but diameter is continuous, so the residuals can now fall within a range of values instead of the three defined categories.

Problem 8

The residuals take on larger values for the ANOVA model compared to the ANCOVA model because there is more error and uncertainty in the ANOVA model. The residuals account for the difference between the values of the observed data points and the predicted values from the regression model, so these are larger in the ANOVA model because the absence of the continuous covariate diameter means the model doesn't make as good of a prediction. When this co variate (diameter) is introduced in the ANCOVA model, the model becomes a better fit of the data thereby decreasing the residual values.

Problem 9

```
##
## Call:
## lm(formula = strength ~ machine + diameter + machine * diameter,
##     data = fiber)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8272 -0.8707 -0.1791  0.5816  3.0857
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        13.5722     4.9375   2.749 0.022520 *
## machineB            7.3421     7.6684   0.957 0.363355
## machineC            4.1068     6.6631   0.616 0.552932
## diameter            1.1043     0.1937   5.702 0.000294 ***
## machineB:diameter  -0.2471     0.2960  -0.835 0.425337
## machineC:diameter  -0.2401     0.2843  -0.845 0.420215
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.675 on 9 degrees of freedom
## Multiple R-squared:  0.9271, Adjusted R-squared:  0.8866
## F-statistic:  22.9 on 5 and 9 DF,  p-value: 7.191e-05
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
machine	2	140.400000	70.200000	25.0230585	0.0002107
diameter	1	178.014110	178.014110	63.4538104	0.0000229
machine:diameter	2	2.737177	1.368589	0.4878387	0.6292895
Residuals	9	25.248712	2.805413	NA	NA

One notable result from the ANCOVA model with interaction is that the signal to noise error is very low, resulting in a low value for the F test and a lack of statistical significance for the ANCOVA F test on the interaction variable. This happens because $MS_{treatment}$ - the signal, is very low, and MS_{error} - is high in comparison. The interpretation of this is that the interaction between diameter and machine type does not explain the independent variable = strength. Based on these results, there is not strong evidence of an interaction.

Appendix

```
library(knitr)
#install the tidyverse library (do this once)
#install.packages("tidyverse")
library(tidyverse)
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, fig.width = 4, fig.height = 4, tid
#code for prob1

#load the dataset
library(readr)
fiber <- read_csv("fiber.csv")

fiber %>% group_by(machine) %>%
summarise("mean_strength" = mean(strength), "sd_strength" = sd(strength),
"mean_diameter" = mean(diameter), "sd_diameter" = sd(diameter),
"n" = length(strength))

g_fiber <- ggplot(fiber, aes(x = diameter, y = strength, col = machine)) +
                             xlab("Fiber Diameter") +
                             ylab("Fiber Strength") +
                              ggtitle("Fiber Strength vs Diameter") +
geom_point()
g_fiber
#code for prob2
model1 <- lm(strength ~ machine, data = fiber)
kable(anova(model1))
#code for prob3
library(emmeans)

fiber.Tukey <- TukeyHSD(aov(strength~machine,data = fiber))
fiber.Tukey
plot(fiber.Tukey,las = 2)

lsFoneway <- emmeans(model1, pairwise ~ machine)
lsFoneway$contrasts
#code for prob4
model2 <- lm(strength ~ machine + diameter, data = fiber)
```

```
kable(anova(model1))
kable(anova(model2))
kable(anova(model1, model2))
#code for prob5a
summary(model2)
emmeansF <- emmeans(model2, pairwise ~ machine|diameter)
emmeansF$contrasts
#code for prob6

#ANOVA Plots
attach(fiber)
df1 <- data.frame(predict(model1),model1$residuals,strength)

strength %>%
  data(df1)
  ggplot() +
  theme_bw() +
  geom_point(aes(x = predict(model1),
                 y = model1$residuals)) +
xlab("Fitted Values") +
ylab("Residuals") +
  ggtitle("ANOVA Residuals vs Fitted Values Plot") +
  theme(legend.position = "none")

ggplot(model1,aes(sample = model1$residuals))+
  stat_qq()+
  stat_qq_line()+
  xlab("Theoretical Quantiles")+
  ylab("Sample Quantiles")+
  ggtitle("ANOVA QQplot")+
  theme_gray(base_size = 16)

#ANCOVA Plots
df2 <- data.frame(predict(model2),model2$residuals,strength)

strength %>%
  data(df2)
  ggplot() +
  theme_bw() +
  geom_point(aes(x = predict(model2),
                 y = model1$residuals)) +
xlab("Fitted Values") +
ylab("Residuals") +
  ggtitle("ANCOVA Residuals vs Fitted Values Plot") +
  theme(legend.position = "none")

ggplot(model2,aes(sample = model2$residuals))+
  stat_qq()+
  stat_qq_line()+
  xlab("Theoretical Quantiles")+
  ylab("Sample Quantiles")+
  ggtitle("ANCOVA QQplot")+
  theme_gray(base_size = 16)
```

```r
#code for prob7

#code for prob8
#code for prob9

#first, we'll need to convert machine from a factor to a vector
attach(fiber)
model3 <- lm(strength ~ machine + diameter + machine * diameter, data = fiber)
summary(model3)
kable(anova(model3))
```