Homework #4

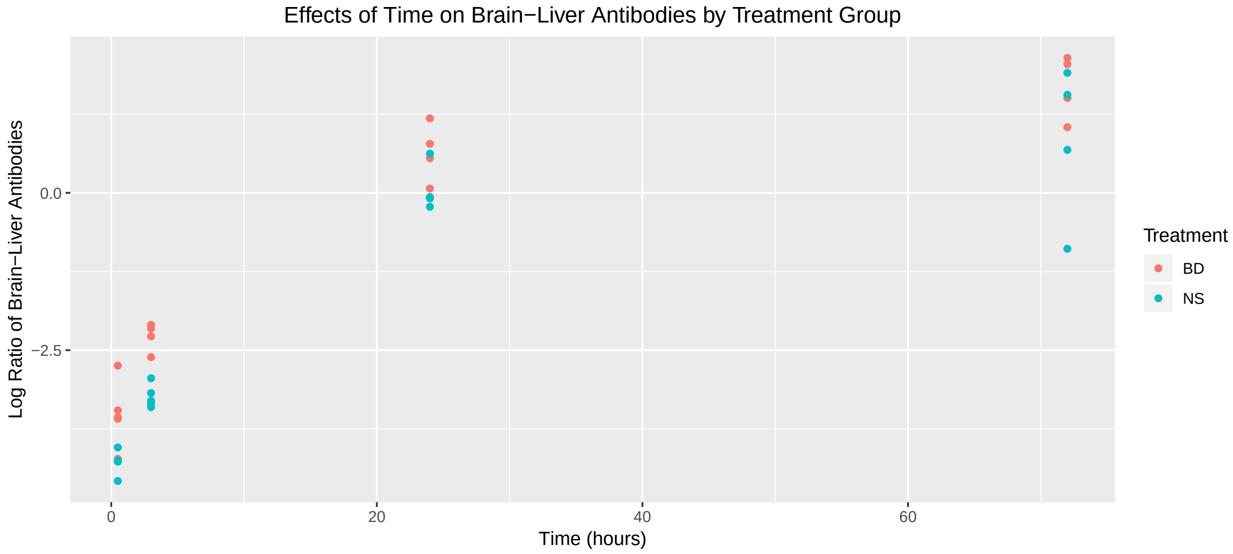
Dustin Leatherman

February 16, 2019

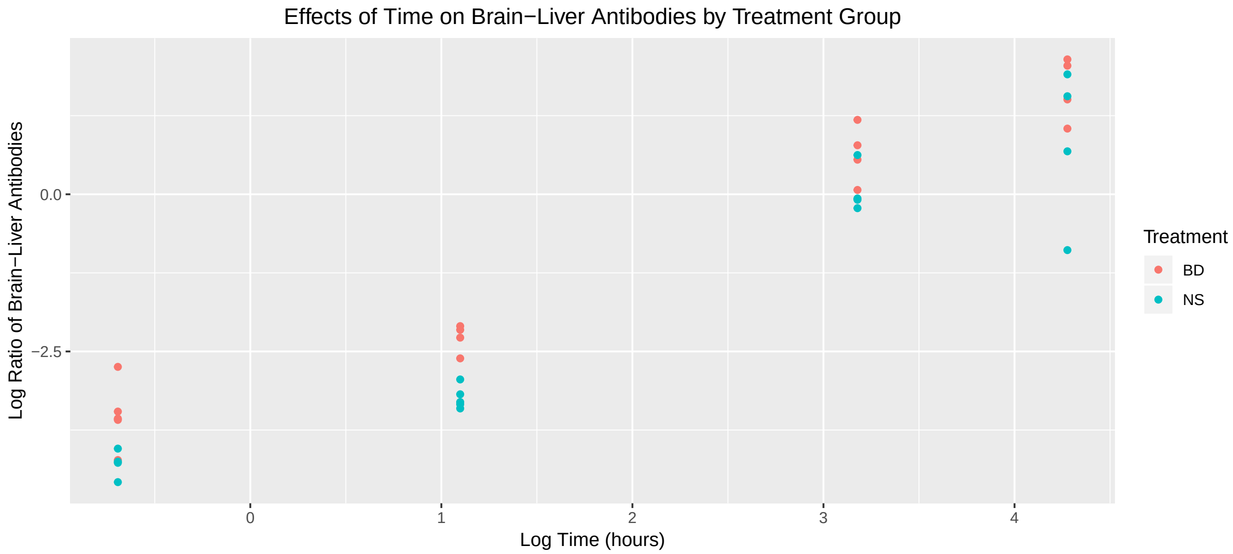
1

Create a plot of the logarithm of the ratio of brain to liver antibodies against the sacrifice time, using different colors to code the treatment categories. Comment on the relationship between the response and the design variables.

```
case1102$ratio <- with(case1102, Brain / Liver)

qplot(Time, log(ratio), color = Treatment, data = case1102) +
  xlab("Time (hours)") +
  ylab("Log Ratio of Brain-Liver Antibodies") +
  ggtitle("Effects of Time on Brain-Liver Antibodies by Treatment Group") +
    theme(plot.title = element_text(hjust = 0.5))
```



```
qplot(log(Time), log(ratio), color = Treatment, data = case1102) +
  xlab("Log Time (hours)") +
  ylab("Log Ratio of Brain-Liver Antibodies") +
  ggtitle("Effects of Time on Brain-Liver Antibodies by Treatment Group") +
    theme(plot.title = element_text(hjust = 0.5))
```

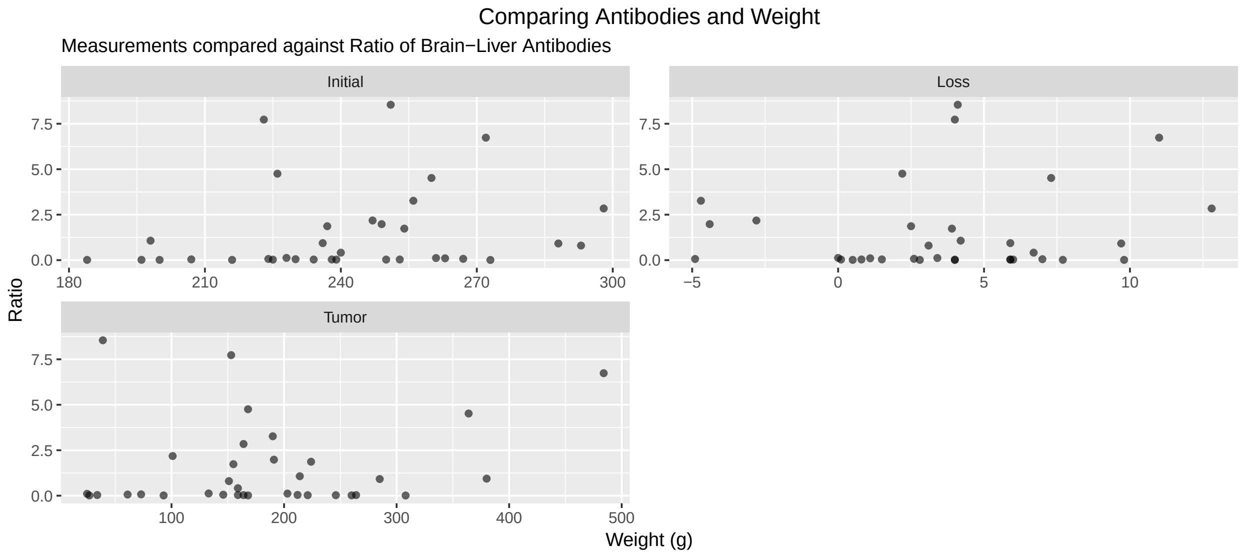


As time increases, there is a non-linear increase in the log ratio of Brain-Liver antibodies. By using Log(Time) as a design variable, there is a more noticeable linear increase in antibodies than using Time without any transformation.

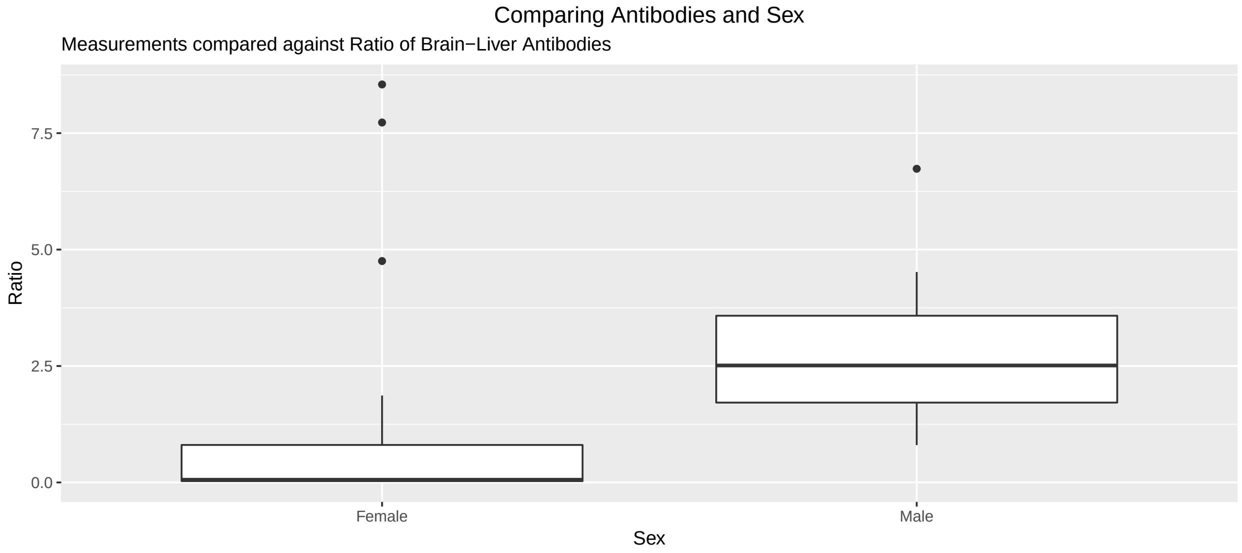
2

For each covariate (sex, weight, loss, and tumor) create a plot to examine the relationship with the response (ratio of brain to liver antibodies). Do there seem to be any relationships?

```r
case1102 %>%
  select(
    `Initial` = Weight,
    Loss,
    Tumor,
    Ratio = ratio
    ) %>%
  gather(-c(Ratio), key = "variable", value = "value") %>%
  ggplot(aes(x = value, y = Ratio)) +
    geom_point(alpha = 0.6) +
    facet_wrap(~ variable , ncol = 2, scales = "free") +
      labs(subtitle = "Measurements compared against Ratio of Brain-Liver Antibodies") +
      xlab("Weight (g)") +
      ylab("Ratio") +
      ggtitle("Comparing Antibodies and Weight") +
        theme(plot.title = element_text(hjust = 0.5))
```



```r
case1102 %>%
  select(
    Sex,
    Ratio = ratio
    ) %>%
  ggplot(aes(x = Sex, y = Ratio)) +
    geom_boxplot() +
      labs(subtitle = "Measurements compared against Ratio of Brain-Liver Antibodies") +
      xlab("Sex") +
      ylab("Ratio") +
      ggtitle("Comparing Antibodies and Sex") +
        theme(plot.title = element_text(hjust = 0.5))
```

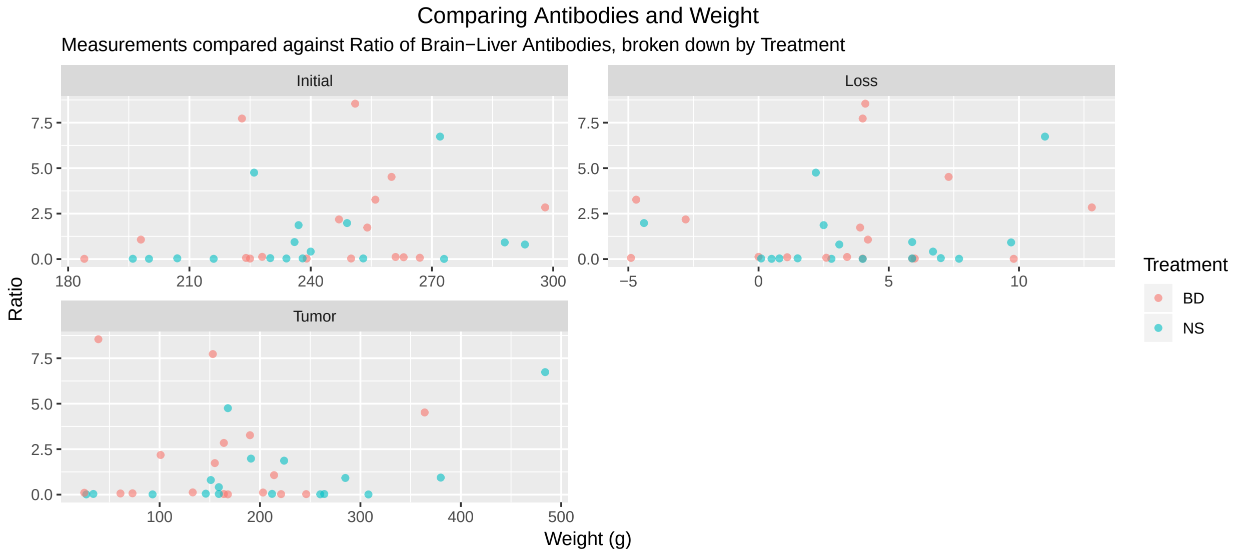


Females tend to have a lower Ratio of Brain-Liver Antibodies but there are 3 female values that appear to be outliers. Males have a lower variance but are less concentrated around low Ratio values. There doesn't appear to be any relationship that particularly stands out. There are a number of responses that have Ratios around 0 which are interesting. It is worth seeing if that is the control Treatment.

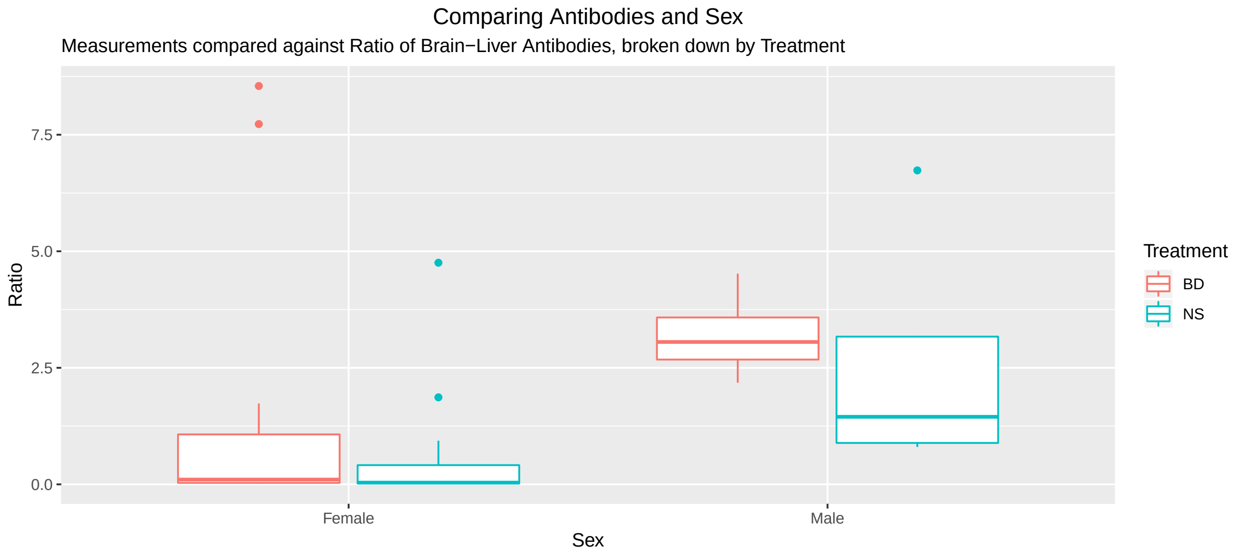
3

For each covariate use plots or summary statistics to explore if thereare relationships between the covariate and treatment assignment.

```r
# Plots except for the dotplot of gender
case1102 %>%
  select(
    `Initial` = Weight,
    Loss,
    Tumor,
    Treatment,
    Ratio = ratio
    ) %>%
  gather(-c(Ratio, Treatment), key = "variable", value = "value") %>%
  ggplot(aes(x = value, y = Ratio, color = Treatment)) +
    geom_point(alpha = 0.6) +
    facet_wrap(~ variable , ncol = 2, scales = "free") +
      labs(subtitle = "Measurements compared against Ratio of Brain-Liver Antibodies, broken down by Tre
      xlab("Weight (g)") +
      ylab("Ratio") +
      ggtitle("Comparing Antibodies and Weight") +
        theme(plot.title = element_text(hjust = 0.5))
```



```r
# Boxplot of Treatment and Gender breakdown for antibodies
case1102 %>%
  select(
    Sex,
    Ratio = ratio,
    Treatment
    ) %>%
  ggplot(aes(x = Sex, y = Ratio, color = Treatment)) +
    geom_boxplot() +
      labs(subtitle = "Measurements compared against Ratio of Brain-Liver Antibodies, broken down by Tre
      xlab("Sex") +
      ylab("Ratio") +
      ggtitle("Comparing Antibodies and Sex") +
        theme(plot.title = element_text(hjust = 0.5))
```



There does not appear to be a clear relationship between Initial, Tumor, or Loss in Weight and Brain-Liver Antibody Ratio. There appears to be a relationship between Sex and Brain-Liver Antibody Ratio though. Males have a higher Antibody count overall than females for both Treatment groups but the Blood Barrier Treatment noticeably increases the Antibody count for males whereas it minorly increases it for females.

4

Fit the tentative model: $\mu\{\log(\text{antibody ratio})|\ \text{TIME, TREAT, DAYS, FEM, weight, loss, tumor}\} = \beta_0 + \beta_1\text{weight} + \beta_2\text{loss} + \beta_3\text{tumor} + \beta_4\text{TIME} + \beta_5\text{TREAT} + \beta_6(\text{TIME} \times \text{TREAT})$

```
# relevel gender so we can get Female as the variable of interest
case1102$SexRelevel <- relevel(case1102$Sex, ref = "Male")

model.full <- lm(log(ratio) ~ Weight + Loss + Tumor + factor(Days) + factor(Time) + Treatment + SexRele

# Hack to replace and surround ':'  with a space so multiple lines work when generating pdf and it does
tidy(model.full) %>%
  mutate(term = gsub(":", " X ", term)) %>%
  kable(
    align = c('l'),
    digits = 4,
    #col.names = c("Waist Girth (cm)","Gender","Fit", "Lower", "Upper", "Fit", "Lower", "Upper"),
    caption = "Full Model for log(antibody ratio)"
  ) %>%
    kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position") %>%
    row_spec(0, bold = T) %>%
    row_spec(which(summary(model.full)$coefficients[,4] < 0.05), color = "black", bold = T) %>%
    column_spec(1, width = "5em")
```

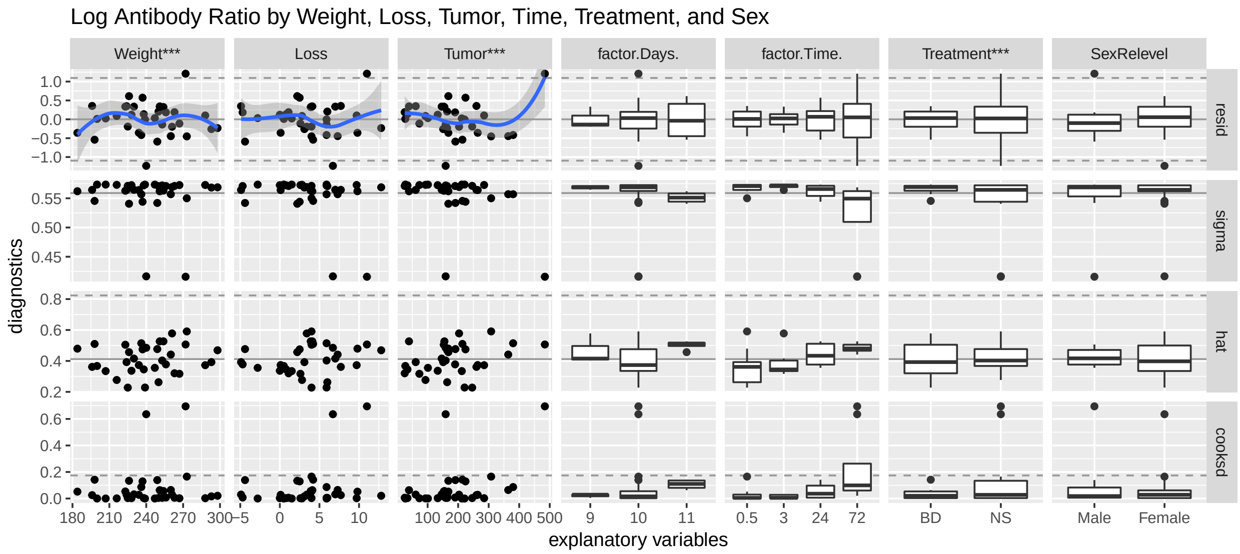
Table 1: Full Model for log(antibody ratio)

term	**estimate**	**std.error**	**statistic**	**p.value**
(Intercept)	-2.8335	1.5527	-1.8249	0.0830
Weight	-0.0005	0.0053	-0.0965	0.9241
Loss	**-0.0651**	**0.0305**	**-2.1341**	**0.0454**
Tumor	0.0015	0.0012	1.2308	0.2327
factor(Days)10	-0.3926	0.4197	-0.9353	0.3608
factor(Days)11	-0.1065	0.5864	-0.1817	0.8577
factor(Time)3	**1.0882**	**0.4296**	**2.5331**	**0.0198**
factor(Time)24	**3.7662**	**0.4584**	**8.2161**	**0.0000**
factor(Time)72	**5.2430**	**0.4700**	**11.1560**	**0.0000**
TreatmentNS	**-0.7981**	**0.3754**	**-2.1259**	**0.0462**
SexRelevelFemale	-0.1629	0.3943	-0.4132	0.6838
factor(Time)3 X TreatmentNS	-0.2198	0.5478	-0.4013	0.6925
factor(Time)24 X TreatmentNS	0.4207	0.5807	0.7245	0.4771
factor(Time)72 X TreatmentNS	-0.3902	0.5640	-0.6918	0.4970

5

Examine the residuals versus fitted values plot for evidence of violations of the assumptions/conditions, and identify two unusual observations.

```r
ggnostic(model.full, title = "Log Antibody Ratio by Weight, Loss, Tumor, Time, Treatment, and Sex")
```



The top row contains residual plots for this model. The residual plot for Tumor Weight violates of the linearity assumption as seen by its trailing upwards. The other assumptions: constant spread, normality, and independence are met. There are two observations which the magnitude of the residuals are greater than 1. One observation has a largest tumor weight, largest loss weight, and towards the upper end of initial weight. The other observation has values that are more central to the other data points. It is intersting to note that both of these values are a part of the control group.

6

```r
augment(model.full) %>%
  select(Weight, Loss, Tumor, Treatment, Fit = .fitted, Residual = .resid, CooksD = .cooksd, Hat = .hat
  filter(abs(Residual) > 1) %>%
  kable(
    align = c('l'),
    digits = 4,
    caption = "Unusual Observations based on high residual magnitude"
  ) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position") %>%
    row_spec(0, bold = T)
```

Table 2: Unusual Observations based on high residual magnitude

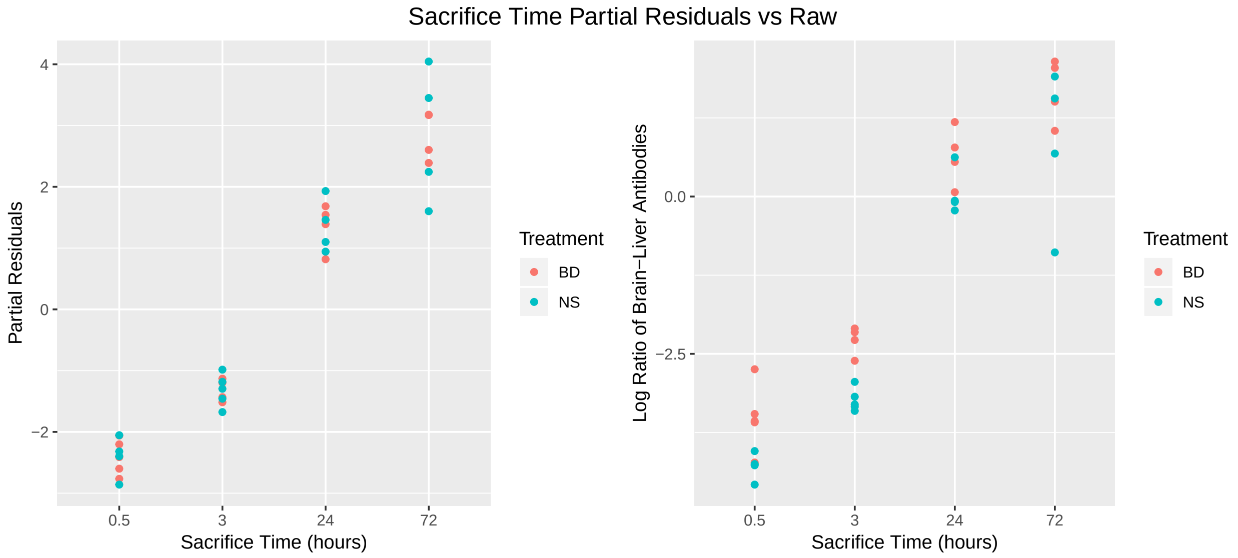
Weight	Loss	Tumor	Treatment	Fit	Residual	CooksD	Hat
272	11.0	484	NS	0.6980	1.2093	0.6950	0.5064
240	6.7	159	NS	0.3451	-1.2331	0.6356	0.4850

The both observations appear to be influential. Both observations have a large Cook's Distance. The Leverage plots look alright. The Studentized Residuals stand out but they are within a magnitude of 2 so it is questionable. The Residual plots show the best-fit line being pulled in the direction of these observations which further indicates its influentiality.

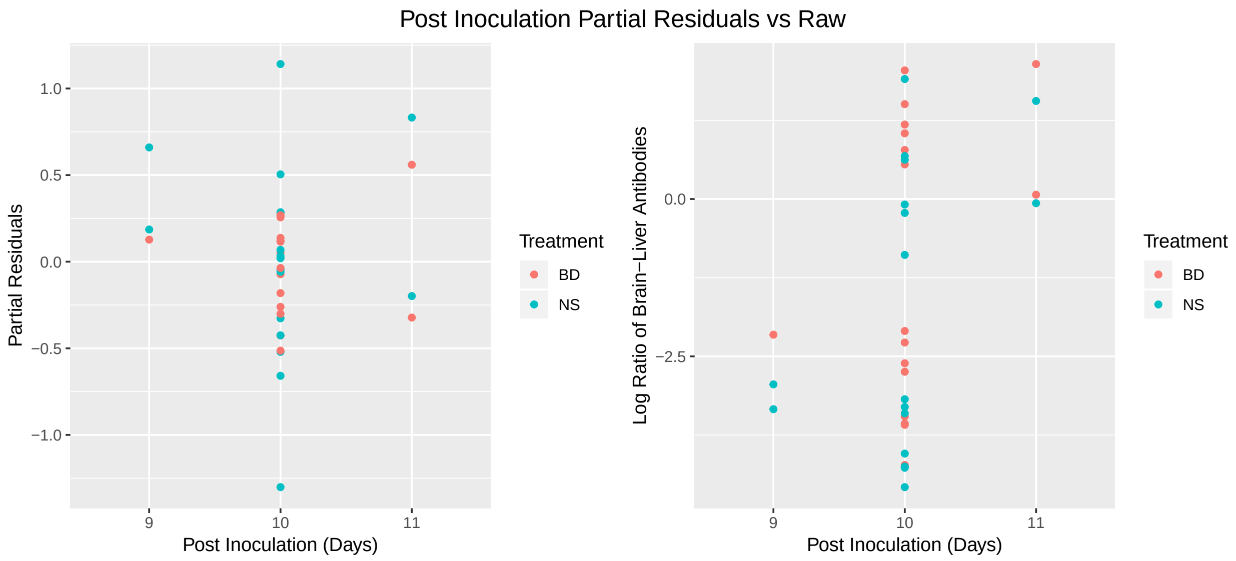
7

```r
# get partial residuals and append to augmented model column
resids <- model.full %>%
  residuals(type = "partial") %>%
  as_tibble %>%
  select(pres.Treatment = Treatment, pres.Time = `factor(Time)`, pres.Sex = SexRelevel, pres.Days = `fa
  bind_cols(augment(model.full)) %>%
  select(Time = `factor.Time.`, Treatment, Sex = SexRelevel, Days = `factor.Days.`, pres.Treatment, pres

time.part <-
  qplot(Time, pres.Time, color = Treatment, data = resids) +
    xlab("Sacrifice Time (hours)") +
    ylab("Partial Residuals")

time.raw <-
  qplot(Time, LogRatio, color = Treatment, data = resids) +
    xlab("Sacrifice Time (hours)") +
    ylab("Log Ratio of Brain-Liver Antibodies")

days.part <-
  qplot(Days, pres.Days, color = Treatment, data = resids) +
    xlab("Post Inoculation (Days)") +
    ylab("Partial Residuals")

days.raw <-
```

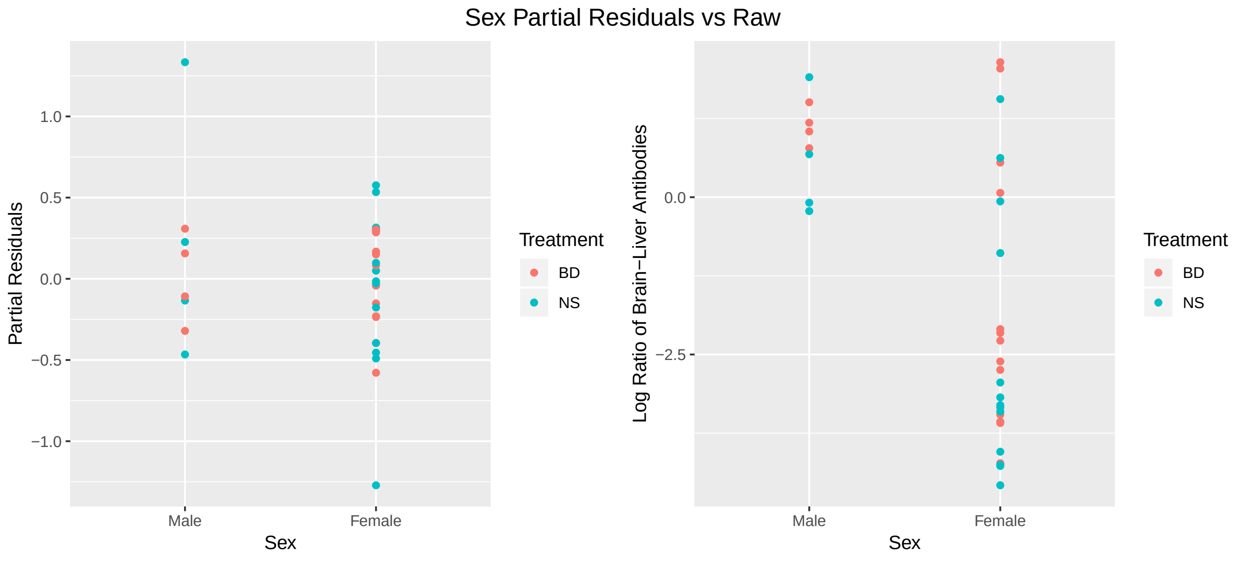
```
  qplot(Days, LogRatio, color = Treatment, data = resids) +
    xlab("Post Inoculation (Days)") +
    ylab("Log Ratio of Brain-Liver Antibodies")

sex.part <-
  qplot(Sex, pres.Sex, color = Treatment, data = resids) +
    xlab("Sex") +
    ylab("Partial Residuals")

sex.raw <-
  qplot(Sex, LogRatio, color = Treatment, data = resids) +
    xlab("Sex") +
    ylab("Log Ratio of Brain-Liver Antibodies")


grid.arrange(time.part, time.raw,
             ncol = 2,
             widths = c(1, 1),
             top = textGrob("Sacrifice Time Partial Residuals vs Raw",
                            gp=gpar(fontsize=14,font=1),just=c("center")))
```



```
grid.arrange(days.part, days.raw,
             ncol = 2,
             widths = c(1, 1),
             top = textGrob("Post Inoculation Partial Residuals vs Raw",
                            gp=gpar(fontsize=14,font=1),just=c("center")))
```



```r
grid.arrange(sex.part, sex.raw,
             ncol = 2,
             widths = c(1, 1),
             top = textGrob("Sex Partial Residuals vs Raw",
                            gp=gpar(fontsize=14,font=1),just=c("center")))
```



Sacrifice Time

Sacrifice Times doesn't look much different than the raw plot though the spread has been reduced. This is saying that depending on the Sacrifice Time in hours, the coefficient representing Sacrifice Time can be fairly large and affect the overall model. The p-value for all Time factors are considered significant and thus affect the relationship between covariates and the response after accounting for treatment.

Days

The relationship with Post Inoculation Days becomes less linear after taking into account the other variables. There is not a clear line between the three factors so I would imagine this coefficient to be less impactful.

The p-values for all Day factors are not considered significant so this shows on here. This does not appear to affect the response much after accounting for treatment.

Sex

The relationship with Sex is less linear after taking into account other variables. Since this is centered around 0, Sex does not appear to be impactful after accounting for Treatment.

8

Model 1 = $\mu\{\log(\text{antibody ratio})|\ \text{TIME, TREAT, DAYS, FEM, weight, loss, tumor}\} = \beta_0 + \beta_1\text{TIME} + \beta_2\text{DAYS} + \beta_3\text{FEM} + \beta_4\text{TREAT} + \beta_5(\text{TIME} \times \text{TREAT})$

Model 2 = $\mu\{\log(\text{antibody ratio})|\ \text{TIME, TREAT, DAYS, FEM, weight, loss, tumor}\} = \beta_0 + \beta_1\text{TIME} + \beta_2\text{TREAT} + \beta_3(\text{TIME} \times \text{TREAT})$

Model 3 = $\mu\{\log(\text{antibody ratio})|\ \text{TIME, TREAT, DAYS, FEM, weight, loss, tumor}\} = \beta_0 + \beta_1\text{DAYS} + \beta_2\text{FEM}$

```r
model1 <- lm(log(ratio) ~ factor(Days) + factor(Time) + Treatment + SexRelevel + factor(Time):Treatment
model2 <- lm(log(ratio) ~ factor(Time) * Treatment, data = case1102)
model3 <- lm(log(ratio) ~ factor(Days) + SexRelevel, data = case1102)

# comparing model 2 vs model 1
tidy(anova(model2, model1)) %>%
  kable(
    align = c('l'),
    digits = 4
  ) %>%
  row_spec(0, bold = T) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position")
```

res.df	rss	df	sumsq	statistic	p.value
26	8.0768	NA	NA	NA	NA
23	7.7528	3	0.324	0.3204	0.8105

```r
# comparing model 3 vs model 1
tidy(anova(model3, model1)) %>%
  kable(
    align = c('l'),
    digits = 4
  ) %>%
  row_spec(0, bold = T) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position")
```

res.df	rss	df	sumsq	statistic	p.value
30	71.5940	NA	NA	NA	NA
23	7.7528	7	63.8412	27.0567	0

```r
# show the estimated values for Model 1 and Model 2 and their absolute difference
tidy(model1) %>%
  inner_join(tidy(model2), by = "term", suffix = c(".model1", ".model2")) %>%
  mutate(estimate.abs.diff = abs(estimate.model1 - estimate.model2), term = gsub(":", " X ", term)) %>%
  select(term, estimate.model1, estimate.model2, estimate.abs.diff) %>%
  kable(
    align = c('l'),
    digits = 4,
    col.names = c("Term", "Model 1", "Model 2", "Abs. Diff"),
    caption = "Estimated values for Design variables with Covariates (Model 1) vs Design Variables witho
  ) %>%
  row_spec(0, bold = T) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position") %>%
  column_spec(1, width = "5em")
```

Table 3: Estimated values for Design variables with Covariates (Model 1) vs Design Variables without Covariates (Model 2)

Term	**Model 1**	**Model 2**	**Abs. Diff**
(Intercept)	-3.0878	-3.5169	0.4291
factor(Time)3	1.1906	1.2311	0.0405
factor(Time)24	3.9428	4.1624	0.2195
factor(Time)72	4.9831	5.2027	0.2195
TreatmentNS	-0.7698	-0.7698	0.0000
factor(Time)3 X TreatmentNS	-0.2043	-0.1800	0.0243
factor(Time)24 X TreatmentNS	0.1867	0.1867	0.0000
factor(Time)72 X TreatmentNS	-0.1006	-0.1006	0.0000

There is not enough evidence to suggest that Model 1 is a better fit than Model 2 (Sum of Squares F-Test. p-value = 0.8105). Model 2 being the simpler model, this means that Days and Sex are not significant when the design variables Time, Treatment, and their interaction are included.

There is convincing evidence that Model 1 is a better fit than Model 3 (Sum of Squares F-Test. p-value = 1.17e-9). This means that design variables are considered significant compared to the covariates.

The absolute difference between the estimates is minute between the models meaning that the covariates don't affect the conclusion.

9

```
tidy(model2) %>%
  mutate(term = gsub(":", " X ", term)) %>%
  kable(
    align = c('l'),
    digits = 4,
    caption = "Summary of Model 2"
  ) %>%
  row_spec(0, bold = T) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position") %>%
  column_spec(1, width = "5em")
```

Table 4: Summary of Model 2

term	estimate	std.error	statistic	p.value
(Intercept)	-3.5169	0.2493	-14.1094	0.0000
factor(Time)3	1.2311	0.3739	3.2928	0.0029
factor(Time)24	4.1624	0.3739	11.1328	0.0000
factor(Time)72	5.2027	0.3739	13.9151	0.0000
TreatmentNS	-0.7698	0.3739	-2.0590	0.0496
factor(Time)3 X TreatmentNS	-0.1800	0.5288	-0.3404	0.7363
factor(Time)24 X TreatmentNS	0.1867	0.5432	0.3437	0.7338
factor(Time)72 X TreatmentNS	-0.1006	0.5432	-0.1852	0.8545

```r
model2.nointer <- lm(log(ratio) ~ factor(Time) + Treatment, data = case1102)

tidy(model2.nointer) %>%
  mutate(term = gsub(":", " X ", term)) %>%
  kable(
    align = c('l'),
    digits = 4,
    caption = "Summary of Model 2 with No interaction Terms"
  ) %>%
  row_spec(0, bold = T) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position") %>%
  column_spec(1, width = "5em")
```

Table 5: Summary of Model 2 with No interaction Terms

term	estimate	std.error	statistic	p.value
(Intercept)	-3.5049	0.1954	-17.9365	0e+00
factor(Time)3	1.1341	0.2520	4.5007	1e-04
factor(Time)24	4.2573	0.2591	16.4310	0e+00
factor(Time)72	5.1539	0.2591	19.8916	0e+00
TreatmentNS	-0.7968	0.1834	-4.3457	2e-04

```r
tidy(anova(model2.nointer, model2)) %>%
  kable(
    align = c('l'),
    digits = 4
  ) %>%
  row_spec(0, bold = T) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position")
```

res.df	rss	df	sumsq	statistic	p.value
29	8.2326	NA	NA	NA	NA
26	8.0768	3	0.1558	0.1672	0.9175

Since the interaction terms in model 2 are not significant, I expect that an F-test will confirm that fact. A Sum of Squares F-Test shows that there is not enough evidence to suggest that the interaction term has a significant effect on the model (p-value = 0.9175).

10

```r
# Readjust parameters to interpret BD treatment affect.
case1102$TreatmentRelevel <- relevel(case1102$Treatment, ref = "NS")
model2.nointer <- lm(log(ratio) ~ factor(Time) + TreatmentRelevel, data = case1102)

# get last row which contains the Treatment group.
confint_tidy(model2.nointer)[5,] %>%
  kable(
    align = c('l'),
    digits = 4
  ) %>%
  row_spec(0, bold = T) %>%
  kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position")
```

conf.low	**conf.high**
0.4218	1.1718

With 95% confidence, the mean log ratio of Brain-Liver antibodies is between 0.4218 and 1.1718 greater for the BD Treatment group than the Control Group.