ST517-HW4

Nora Quick

1. Recall the Lab 5 data and models:

   Model 1: Both birds and non-echolocating bats have possibly different energy costs in flight to echolocating bats, after accounting for a linear relationship between log energy and log mass.

   Model 2: The energy costs for non-echolocating bats and echolocating bats is the same, but possibly different to birds, after accounting for a linear relationship between log energy and log mass.

   (a) Use these two models to demonstrate that the Extra Sum of Squares F-test comparing models that only differ by one parameter is equivalent to a t-test of that parameter, and that the F-statistic is the t-statistic squared.

```
mod1 <- lm(log(Energy) ~ log(Mass) + Type, data = case1002)
mod2 <- lm(log(Energy) ~ log(Mass) + I(Type == "non-echolocating birds"), data = case1002)


rss1 <- deviance(mod1)
rss2 <- deviance(mod2)

df1 <- df.residual(mod1)
df2 <- df.residual(mod2)

fstat <- ((rss2 -rss1) / (df2 - df1)) / (rss1 / df1)
fstat
```

```
## [1] 0.1506364
```

```
summary(mod1)
```

```
##
## Call:
## lm(formula = log(Energy) ~ log(Mass) + Type, data = case1002)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23224 -0.12199 -0.03637  0.12574  0.34457
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -1.49770    0.14987  -9.993 2.77e-08 ***
## log(Mass)                     0.81496    0.04454  18.297 3.76e-12 ***
## Typenon-echolocating bats    -0.07866    0.20268  -0.388    0.703
## Typenon-echolocating birds    0.02360    0.15760   0.150    0.883
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Residual standard error: 0.186 on 16 degrees of freedom
## Multiple R-squared:  0.9815, Adjusted R-squared:  0.9781
## F-statistic: 283.6 on 3 and 16 DF,  p-value: 4.464e-14
```

```
sq_t <- (-0.388)^2
sq_t
```

```
## [1] 0.150544
```

(b) Consider these two model specified in R's lm() notation:

lm(log(Energy) ~ log(Mass), data = case1002) lm(log(Energy) ~ log(Mass) + Type, data = case1002)

Describe in non-technical terms, (i.e. to someone who doesn't use R), why these two models, that look like they only differ by one parameter, cannot be compared with a single t-test.

In this situation we are testing the effects of different variables on Energy. We need to test the same variables for their effect on the variable (if they effect the variable or not).

The way we want to do this is to check the same variables but in different ways. From the model we see if (a) we are checking the Type in different ways to check if it has a significant effect.

If we don't check the same vairables we will be checking different outcomes overall and when comparing the models for assiciation between independent variables and the outcomes of those variables.

2. Consider the data `ex1014` in the `Sleuth3` package. The data describes an experiment in which researchers randomly allocated 25 beakers containing minnow larvae to treatment combinations of zinc and copper. After four days, the minnows in each beaker were measured for their protein levels.

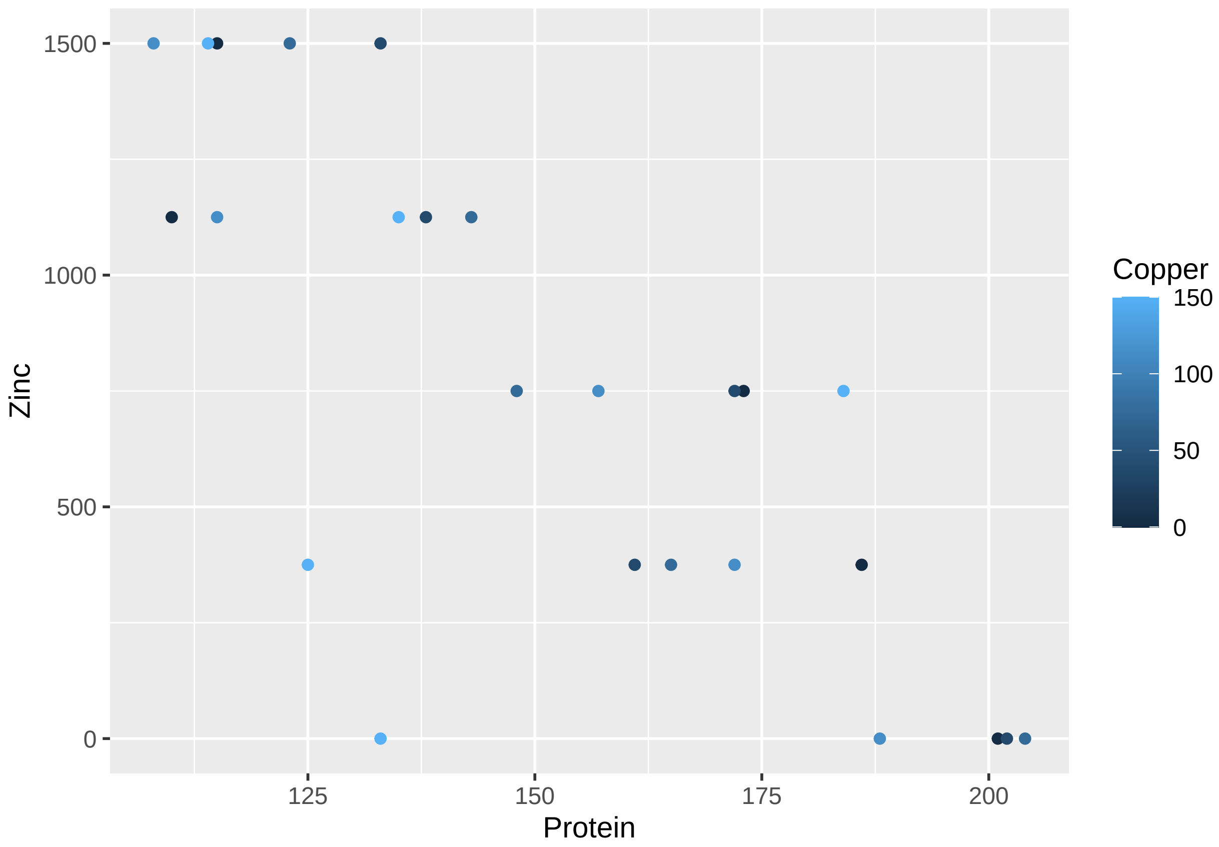
```
head(ex1014)
```

```
##   Copper Zinc Protein
## 1      0    0     201
## 2      0  375     186
## 3      0  750     173
## 4      0 1125     110
## 5      0 1500     115
## 6     38    0     202
```

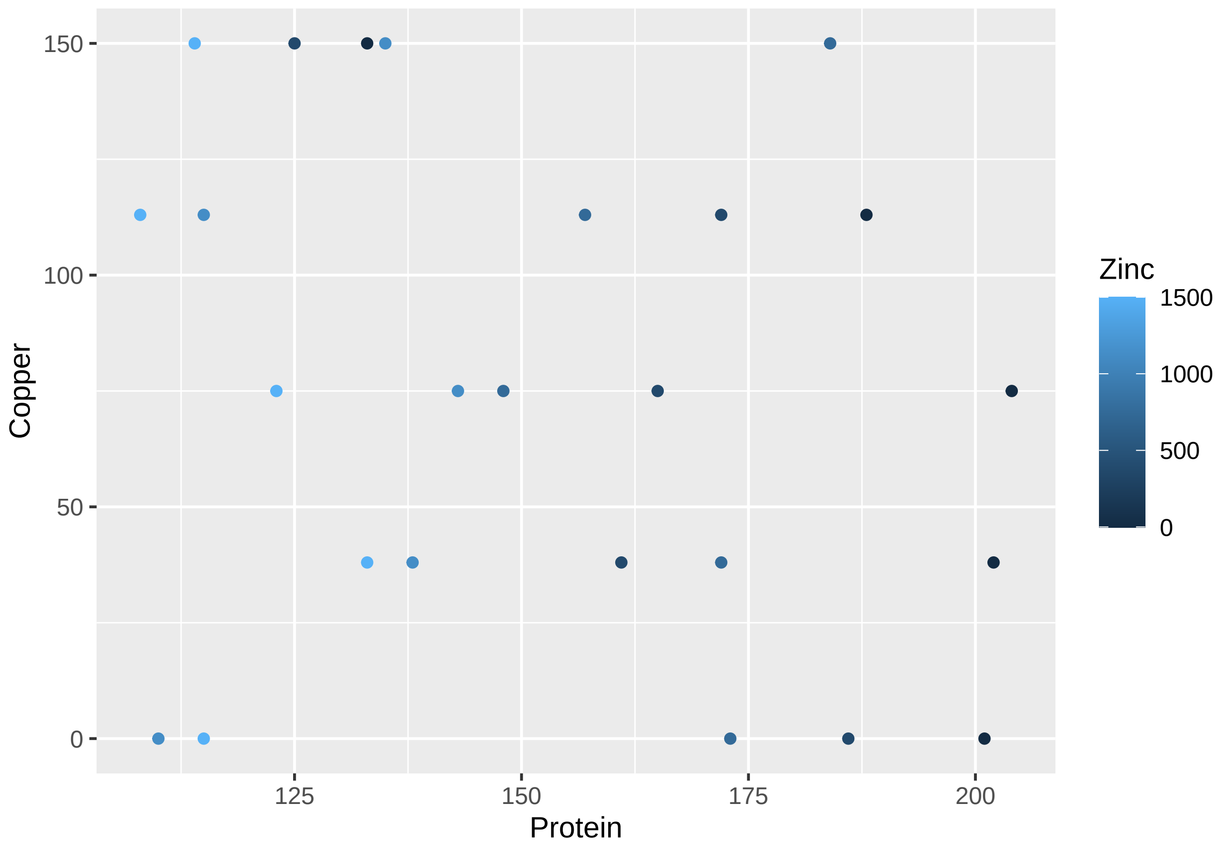
```
Copper <- ex1014$Copper
Zinc <- ex1014$Zinc
Protein <- ex1014$Protein
```

(a) Create a plot of protein against zinc, with points colored by the level of copper, and a plot of protein against copper, with points colored by the level of zinc. Describe any relationships you see.

```
qplot(Protein, Zinc, data = ex1014, color = Copper)
```



```
qplot(Protein, Copper, data = ex1014, color = Zinc)
```



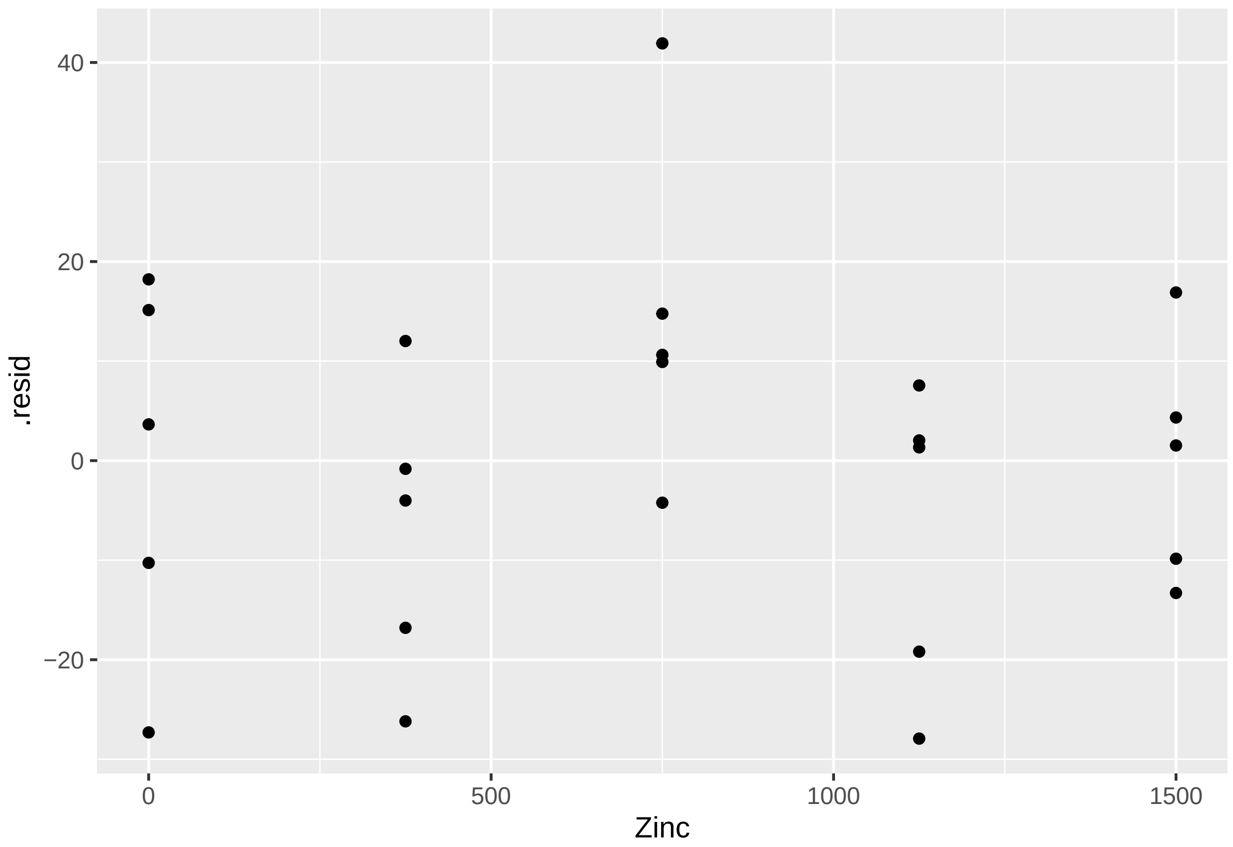
When comparing Protein and Zinc we can see that as there is an increase in Protein and a decrease in Zinc there is less Copper. When comparing Protein and Copper we can see that as there is an increase in Protein there is less Copper no matter what the Copper level is.

(b) Fit a model for protein that includes both main and interaction terms for Zinc and Copper. Examine the residual plots and comment on the validity of the assumptions.

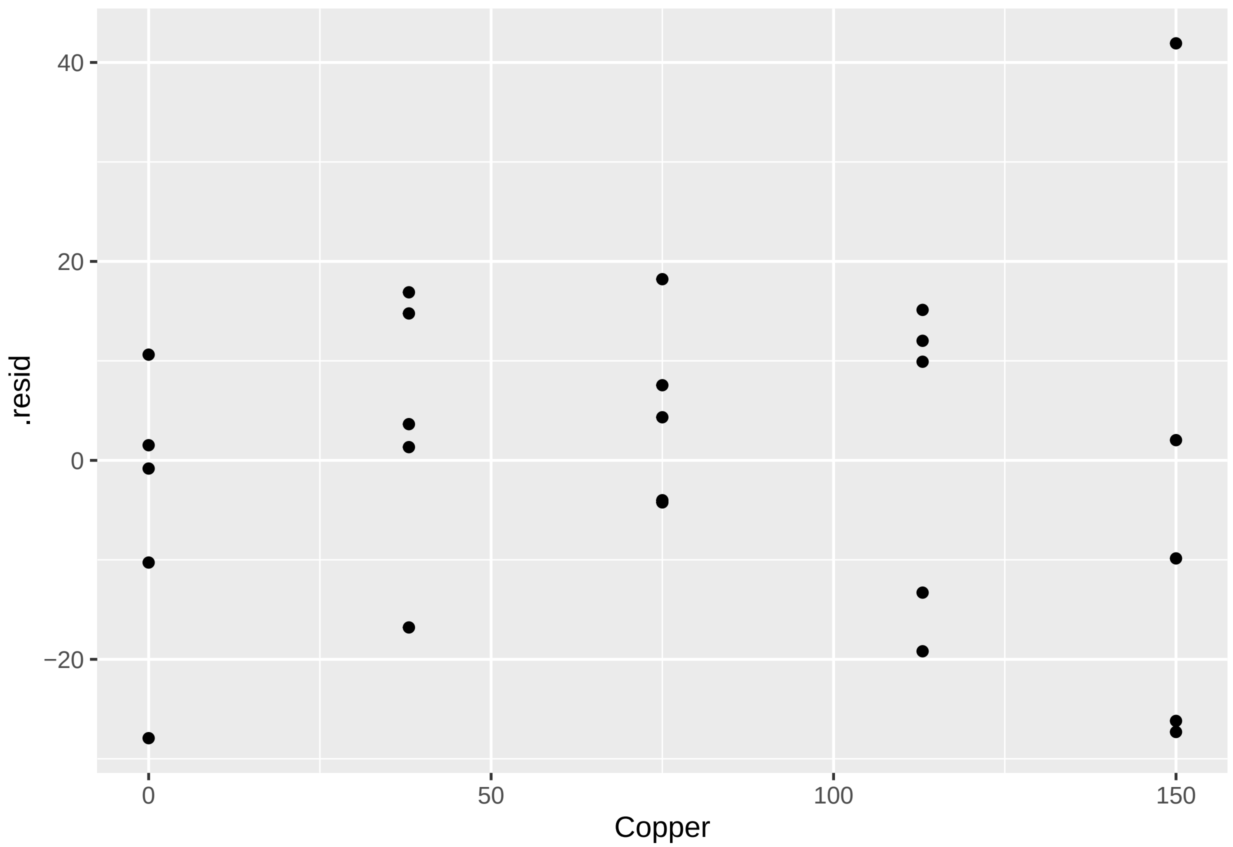
```
modOne <- lm(Protein ~ Zinc + Copper + Zinc:Copper, data = ex1014)
plotmodOne <- augment(modOne, ex1014)

head(plotmodOne)
```

```
## # A tibble: 6 x 9
##   Copper  Zinc Protein .fitted  .resid  .hat .sigma  .cooksd .std.resid
##    <int> <int>   <int>   <dbl>   <dbl> <dbl>  <dbl>    <dbl>      <dbl>
## 1      0     0     201    211. -10.3   0.361   17.9 0.0749      -0.728
## 2      0   375     186    187.  -0.823 0.181   18.1 0.000146    -0.0515
## 3      0   750     173    162.  10.6   0.120   17.9 0.0141       0.642
## 4      0  1125     110    138. -27.9   0.181   16.7 0.168       -1.75
## 5      0  1500     115    113.   1.52  0.361   18.1 0.00165      0.108
## 6     38     0     202    198.   3.64  0.179   18.1 0.00282      0.227
```

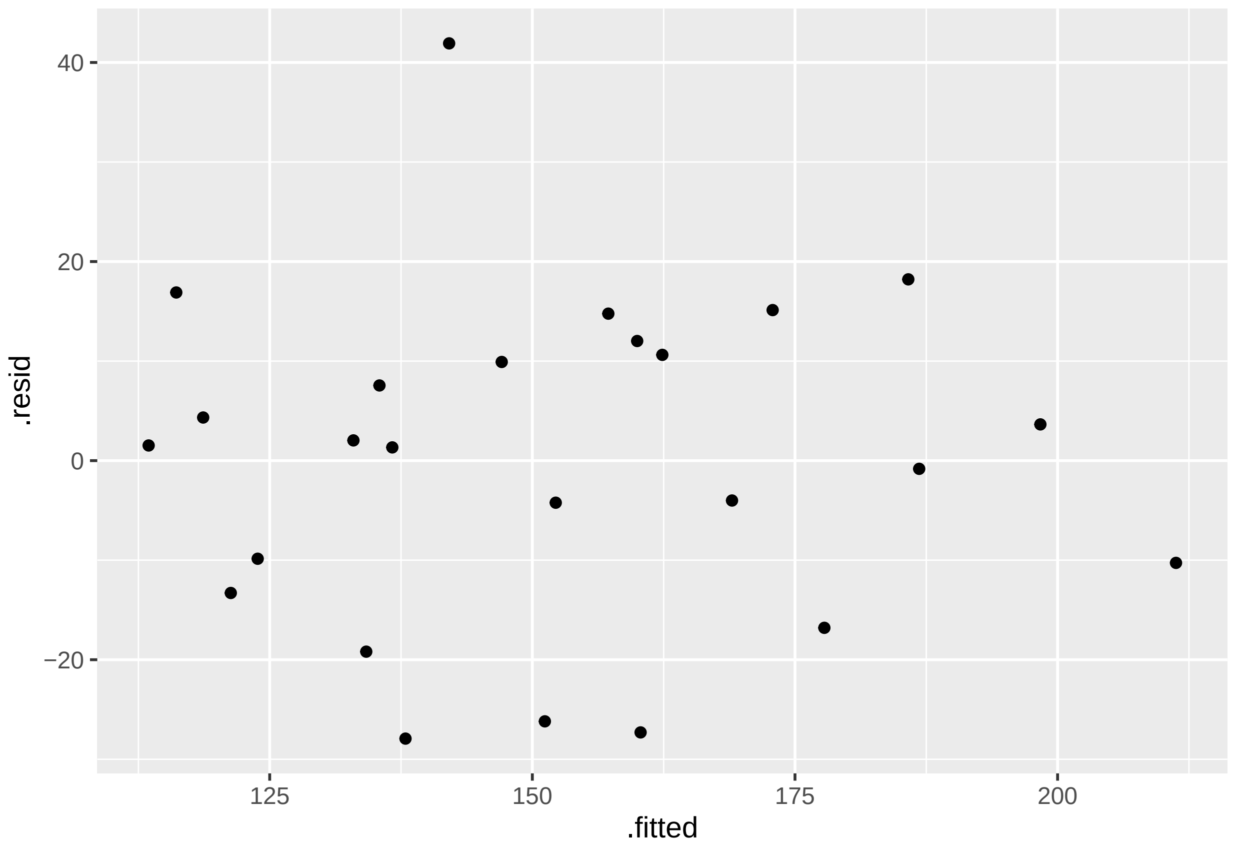
```
qplot(Zinc, .resid, data = plotmodOne)
```



```
qplot(Copper, .resid, data = plotmodOne)
```



```
qplot(.fitted, .resid, data = plotmodOne)
```



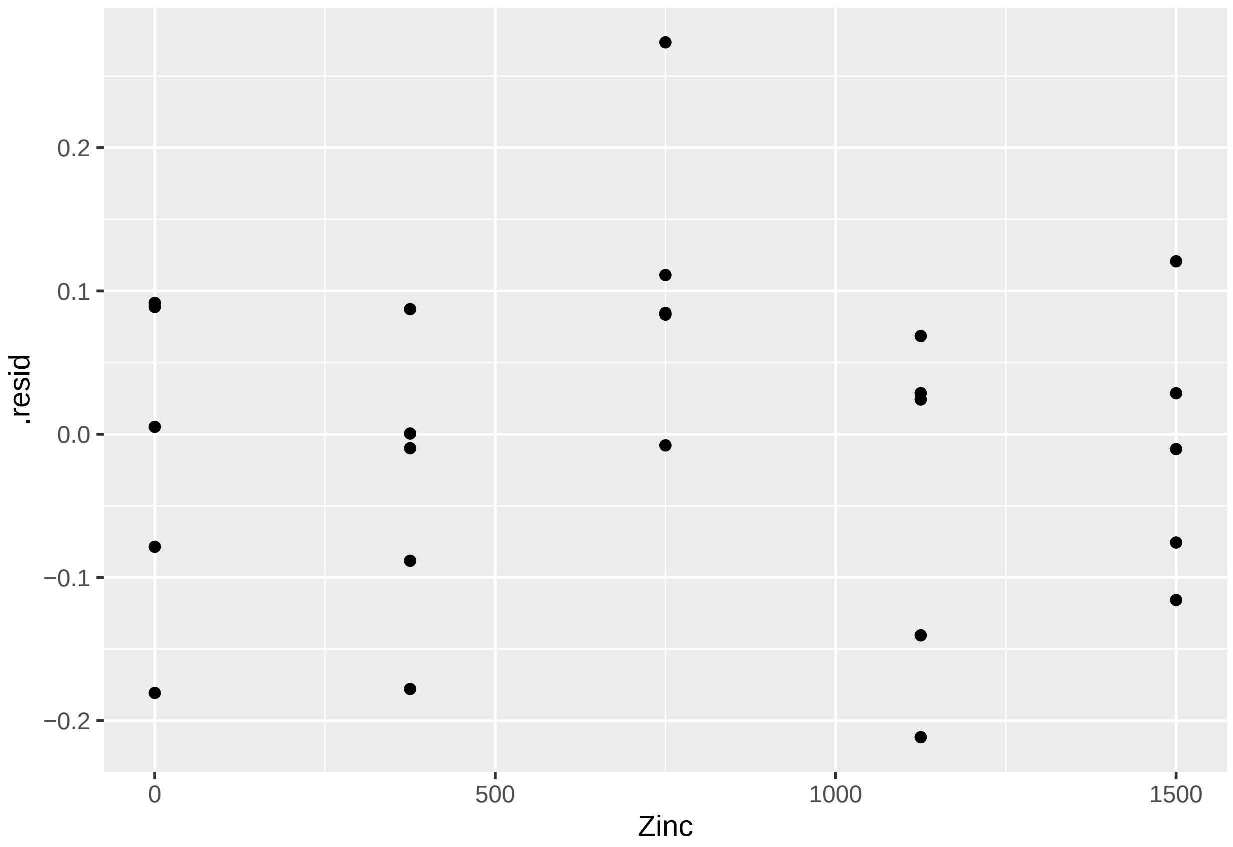
From what we can see from the table and the plots both Zinc and Copper effect the minnow Protein levels. When there is no Copper or Zinc the minnow Protein level is almost at it's highest. When theres no Copper but Zinc the minnow Protein levels are at their lowerst and when theres no Zinc but Copper theres the highest level of minnow Protein (about on par with no Copper or Zinc). We can conclude that they both effect minnow Protein but Zinc in a negative and Copper in a positive way and it is likely that the minnow Protein levels are find without Copper or Zinc.

(c) Fit a model for log protein that includes both main and interaction terms for Zinc and Copper. Examine the residual plots and comment on the validity of the assumptions. Is there evidence the model on the log scale better satisfies the assumptions?

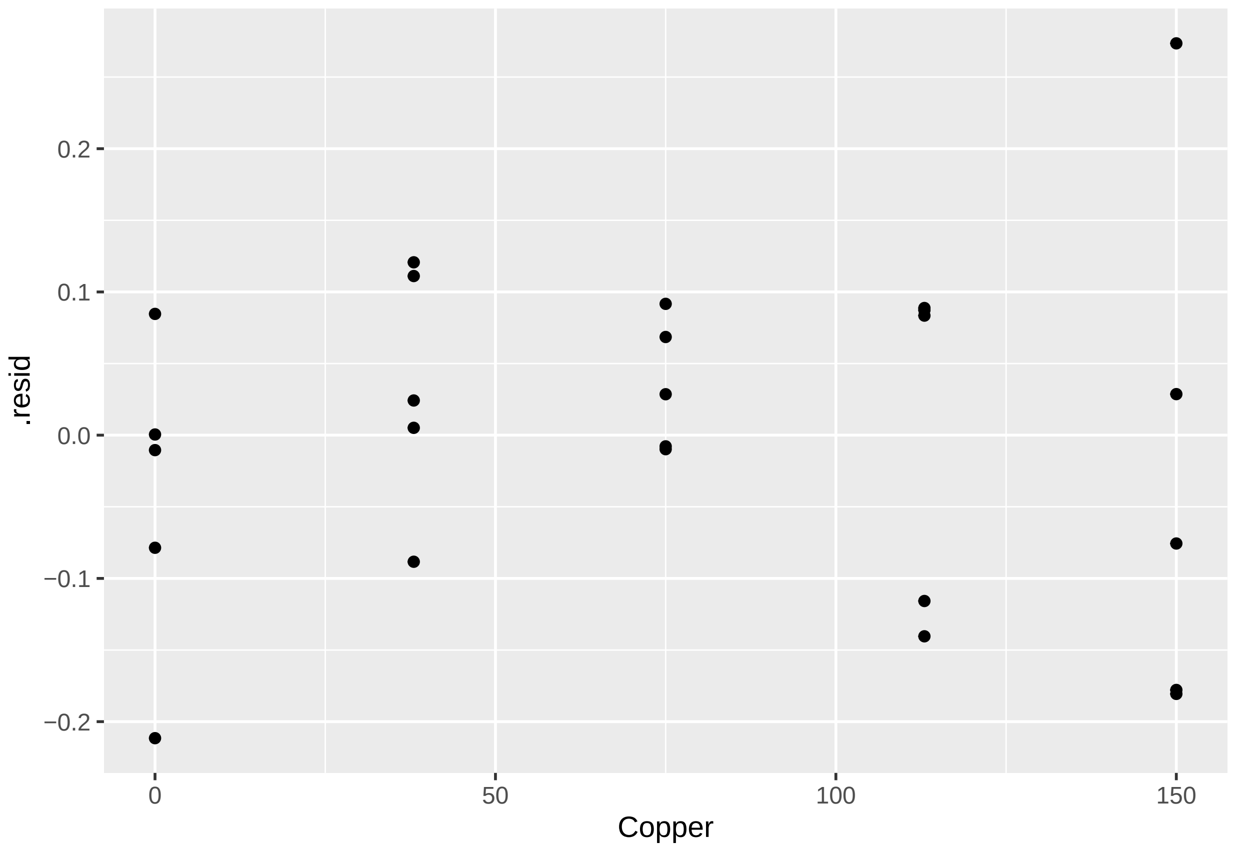
```
modTwo <- lm(log(Protein) ~ Zinc + Copper + Zinc:Copper, data=ex1014)
plotmodTwo <- augment(modTwo, ex1014)

head(plotmodTwo)
```

```
## # A tibble: 6 x 9
##   Copper  Zinc Protein .fitted    .resid  .hat .sigma    .cooksd .std.resid
##    <int> <int>   <int>   <dbl>     <dbl> <dbl>  <dbl>      <dbl>      <dbl>
## 1      0     0     201    5.38 -0.0786   0.361  0.122 0.0930       -0.811
## 2      0   375     186    5.23  0.000472 0.181  0.124 0.00000102    0.00430
## 3      0   750     173    5.07  0.0847   0.120  0.123 0.0190        0.744
## 4      0  1125     110    4.91 -0.212    0.181  0.113 0.205        -1.93
## 5      0  1500     115    4.76 -0.0104   0.361  0.124 0.00164      -0.108
## 6     38     0     202    5.30  0.00513  0.179  0.124 0.000119      0.0467
```

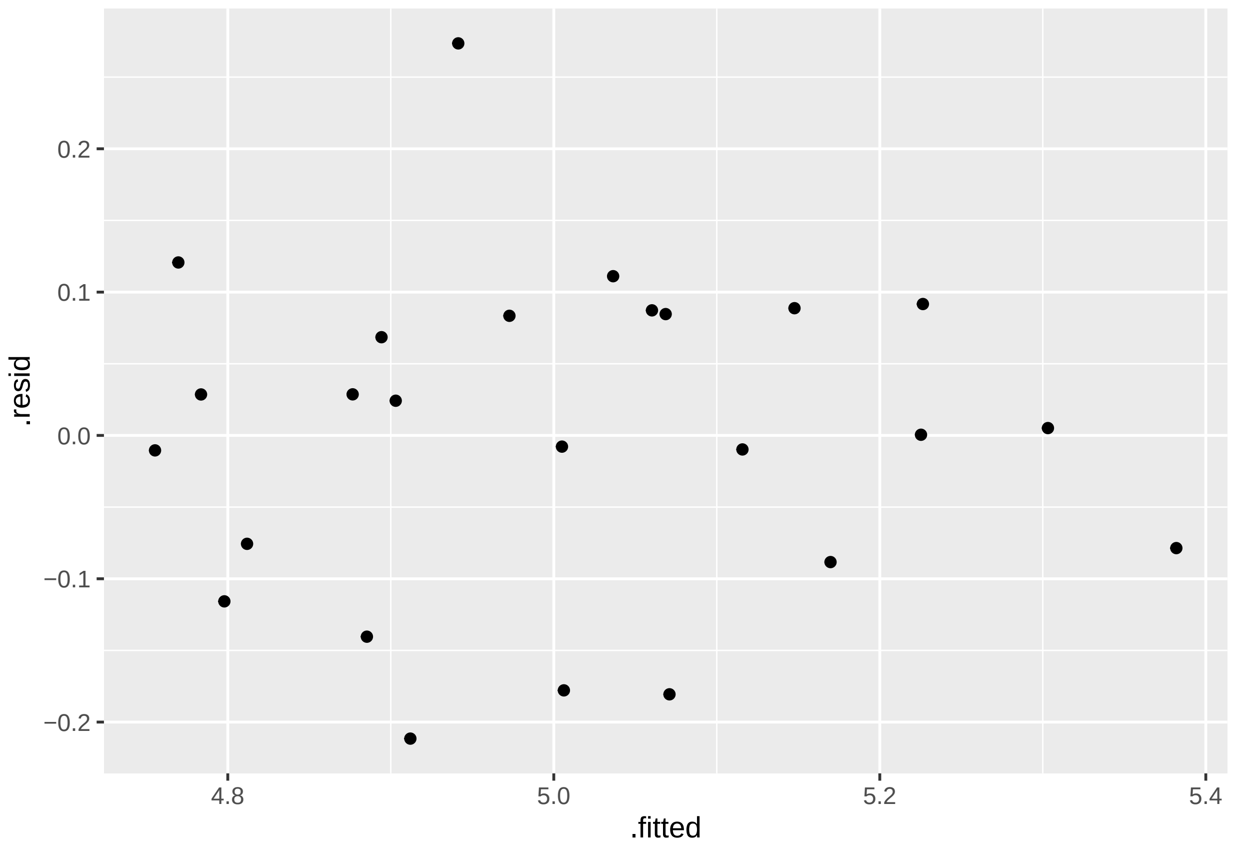
```
qplot(Zinc, .resid, data = plotmodTwo)
```



```
qplot(Copper, .resid, data = plotmodTwo)
```



```
qplot(.fitted, .resid, data = plotmodTwo)
```



The log scale slightly better satisfies the assumption that neither are truly effective. From .resid and .sigma from the table we can see that there aren't any truly sigificant values. From the graphs we can also see that there is almost no significant values.

(d) Conduct a test for the interaction term in the model in (b). Make sure you include completely specify your model, hypotheses, test statistic and p-value. Write a short non-technical summary based on your result in the context of the study.

```
modThree <- lm(Protein ~ Zinc + Copper, ex1014)

rss1 <- deviance(modOne)
rss3 <- deviance(modThree)

df1 <- df.residual(modOne)
df3 <- df.residual(modThree)

fstat <- ((rss3 - rss1)/(df3 - df1))/(rss1/df1)
pstat <- 1 - pf(fstat, df3 - df1, df3)

# fstat
# pstat

anova(modThree, modOne)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: Protein ~ Zinc + Copper
## Model 2: Protein ~ Zinc + Copper + Zinc:Copper
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     22 8020.1
## 2     21 6549.9  1    1470.2 4.7135 0.04153 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The models we are testing are the model with the interaction between Zinc and Copper on Protein and the model with both just affecting Protein. Our hypothesis for this test is that the interaction between the two does not have enough veriability to sufficiently justify adding the interaction to the model. For this an ANOVA test for Sum of Squares F-test was used and we found an f-stat of 4.7 and a p-value of 0.041

From the p-value we can conclude that there is enough variability with the interaction term between Zinc and Copper to justify it being in model one.

(e) Produce mean protein levels, along with confidence intervals, for all combinations of Zinc and Copper based on the model in (b). Describe the effect of the interaction between Zinc and Copper on mean protein. (Hint: making a plot of these predictions might help).

```
mean_modOne <- mean(predict(modOne, interval = "confidence"))
ci_modOne <- predict(modOne, interval = "confidence")

mean_modOne
```

```
## [1] 152.2
```

```
ci_modOne
```

```
##         fit       lwr      upr
## 1  211.2718 189.19625 233.3473
## 2  186.8229 171.21316 202.4327
## 3  162.3740 149.62873 175.1194
## 4  137.9252 122.31541 153.5349
## 5  113.4763  91.40077 135.5518
## 6  198.3600 182.81932 213.9006
## 7  177.7964 166.80754 188.7853
## 8  157.2329 148.26050 166.2053
## 9  136.6694 125.68046 147.6583
## 10 116.1058 100.56517 131.6465
## 11 185.7880 173.06509 198.5108
## 12 169.0075 160.01108 178.0039
## 13 152.2271 144.88151 159.5726
## 14 135.4466 126.45018 144.4430
## 15 118.6662 105.94329 131.3890
## 16 172.8761 157.25239 188.4999
## 17 159.9810 148.93337 171.0287
## 18 147.0859 138.06554 156.1063
## 19 134.1908 123.14314 145.2385
## 20 121.2957 105.67193 136.9195
## 21 160.3041 138.30695 182.3013
## 22 151.1921 135.63774 166.7465
## 23 142.0801 129.38000 154.7802
## 24 132.9680 117.41369 148.5224
## 25 123.8560 101.85884 145.8532
```

The interation of Zinc and Copper on Protein reduces the mean level of Protein the minnows have when there is no Zinv and Copper.

(f) Try fitting a model where the levels of both Zinc and Copper are treated as categories, and include an interaction between the now categorical Zinc and Copper. Examine the model. Describe the problem with this model that prevents inference from proceeding.

```
modFour <- lm(Protein ~ factor(Zinc) * factor(Copper), data =  ex1014)

summary(modFour)
```

```
##
## Call:
## lm(formula = Protein ~ factor(Zinc) * factor(Copper), data = ex1014)
##
## Residuals:
## ALL 25 residuals are 0: no residual degrees of freedom!
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             201         NA      NA       NA
## factor(Zinc)375                         -15         NA      NA       NA
## factor(Zinc)750                         -28         NA      NA       NA
## factor(Zinc)1125                        -91         NA      NA       NA
## factor(Zinc)1500                        -86         NA      NA       NA
## factor(Copper)38                          1         NA      NA       NA
## factor(Copper)75                          3         NA      NA       NA
## factor(Copper)113                       -13         NA      NA       NA
## factor(Copper)150                       -68         NA      NA       NA
## factor(Zinc)375:factor(Copper)38        -26         NA      NA       NA
## factor(Zinc)750:factor(Copper)38         -2         NA      NA       NA
## factor(Zinc)1125:factor(Copper)38        27         NA      NA       NA
## factor(Zinc)1500:factor(Copper)38        17         NA      NA       NA
## factor(Zinc)375:factor(Copper)75        -24         NA      NA       NA
## factor(Zinc)750:factor(Copper)75        -28         NA      NA       NA
## factor(Zinc)1125:factor(Copper)75        30         NA      NA       NA
## factor(Zinc)1500:factor(Copper)75         5         NA      NA       NA
## factor(Zinc)375:factor(Copper)113        -1         NA      NA       NA
## factor(Zinc)750:factor(Copper)113        -3         NA      NA       NA
## factor(Zinc)1125:factor(Copper)113       18         NA      NA       NA
## factor(Zinc)1500:factor(Copper)113        6         NA      NA       NA
## factor(Zinc)375:factor(Copper)150         7         NA      NA       NA
## factor(Zinc)750:factor(Copper)150        79         NA      NA       NA
## factor(Zinc)1125:factor(Copper)150       93         NA      NA       NA
## factor(Zinc)1500:factor(Copper)150       67         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 24 and 0 DF,  p-value: NA
```

The issue with this format is that there is almost no significant data provided in the model. There are no Degrees of Freedom, no standard error, p-values, or t-values. From this model we cannot accurratly see which Zinc and Copper levels are significant or not. The esitimated std. does not give enough data to make inferences from the model.