Problem Set 3

Ian Sims

February 19, 2019

Load packages:

1. (5 points.) Using the data set BGSgirls in package alr4:

(a) Fit a linear regression model to predict girls' weight at age 18 (variable WT18) using weight at age 2 (WT2) and weight at age 9 (WT9) as the regressors. Display the resulting model in a format that someone who has not used R can understand.

```
m.wghts = lm(WT18 ~ WT2 + WT9, data = BGSgirls)
tidy(m.wghts)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic       p.value
##   <chr>          <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept)   34.3       6.55       5.23 0.00000180
## 2 WT2           -0.974     0.701     -1.39 0.169
## 3 WT9            1.20      0.179      6.70 0.00000000509
```

This linear regression produces the model: WT18 = 34.26 - .97WT2 + 1.2WT9

(b) The model with both WT2 and WT9 as the regressors has a negative coefficient for WT2. A friend sees this and says, "The negative sign means that girls who are heavier than average at age 2 will usually be heavier than average at age 18." Patiently explain why your friend is mistaken, and give a correct interpretation of the negative sign.

Actually a negative coefficient implies a negative relationship to the response variable. In this case if a person has a higher weight at age 2 the model predicts they will have a lower weight at age 18. Though this shouldn't be relied on too much from this model since this model contains two predictor variables. If we were to run a linear model just using the age 2 weight we might get a different coefficient if there is not independence between WT2 and WT9.

(c) The model with both WT2 and WT9 as the regressors has a coefficient of 1.2 for WT9. A friend sees this and says, "If two girls have a one pound difference in weight at age 9, the model predicts they'll have a 1.2 difference in weight at age 18." Is your friend correct? Why or why not?

This would be true based on this model, assuming that both the girls had the same weight at age 2. Again this shouldn't be relied on too much considering there are two predictor

variables that are likely not independent. If WT9 was modeled independenlty, the coeeficient could change.

2. (10 points.) The data set MinnLand in package alr4 contains data on "nearly every farm sale" in six economic regions in Minnesota from 2002 to 2011. Suppose we wish to model how sale price per acre (acrePrice) depends on year. Since sales price per acre is strongly right-skewed, we'll take log(acrePrice) as the response in our regressions.

(a) Fit a linear regression model to predict log(acrePrice) from year alone, taking year as a continuous variable. Write down the regression equation you obtain.

```
m.price = lm(log(acrePrice) ~ year, data = MinnLand)
tidy(m.price)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept) -194.      3.98        -48.7       0
## 2 year           0.100   0.00199      50.6       0
```

The regresion equation is: log(acrePrice) = -193.88 + 0.10*year

(b) Fit a regression model to predict log(acrePrice) from year alone, taking year as a factor. State the coefficient for the year 2008, and explain what this coefficient means.

```
mnlnd = MinnLand
mnlnd$year.fctr = as.factor(MinnLand$year)
#head(model.matrix(acrePrice ~ year.fctr, mnlnd), 10)
m.price.fctr = lm(log(acrePrice) ~ year.fctr, data = mnlnd)
summary(m.price.fctr)
```

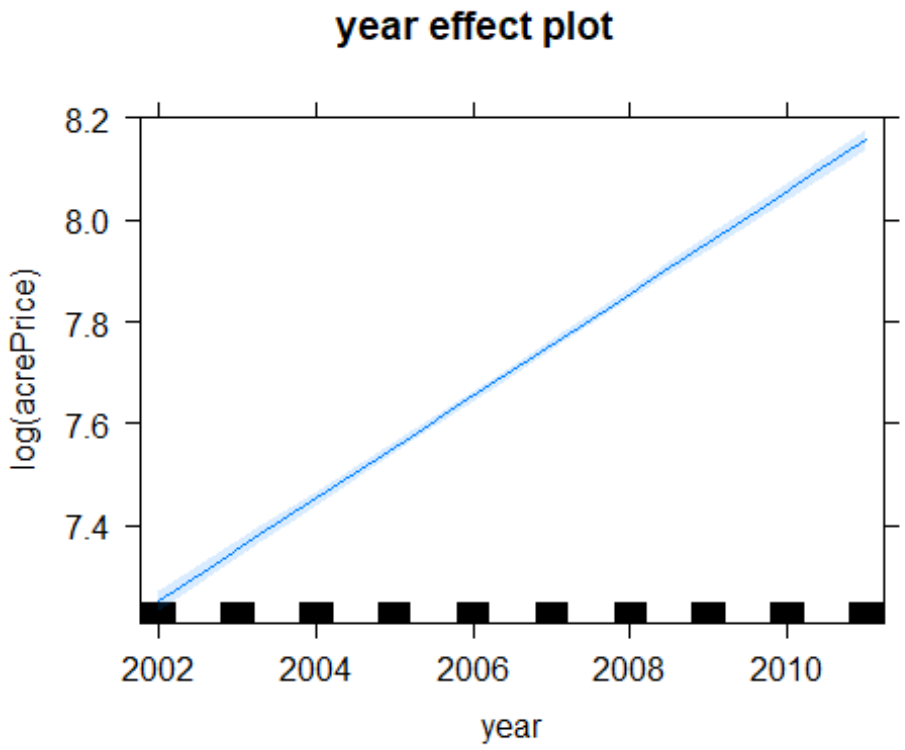
```
##
## Call:
## lm(formula = log(acrePrice) ~ year.fctr, data = mnlnd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9499 -0.3785  0.1301  0.4354  2.3456
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.27175    0.02848 255.345  < 2e-16 ***
## year.fctr2003 -0.00155    0.03207  -0.048    0.961
## year.fctr2004  0.14794    0.03155   4.689 2.76e-06 ***
## year.fctr2005  0.36026    0.03176  11.343  < 2e-16 ***
```

```
## year.fctr2006  0.39392    0.03195  12.329  < 2e-16 ***
## year.fctr2007  0.47682    0.03186  14.965  < 2e-16 ***
## year.fctr2008  0.68364    0.03162  21.620  < 2e-16 ***
## year.fctr2009  0.71407    0.03355  21.284  < 2e-16 ***
## year.fctr2010  0.75733    0.03260  23.231  < 2e-16 ***
## year.fctr2011  0.72071    0.03526  20.437  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6775 on 18690 degrees of freedom
## Multiple R-squared:  0.1293, Adjusted R-squared:  0.1289
## F-statistic: 308.5 on 9 and 18690 DF,  p-value: < 2.2e-16
```

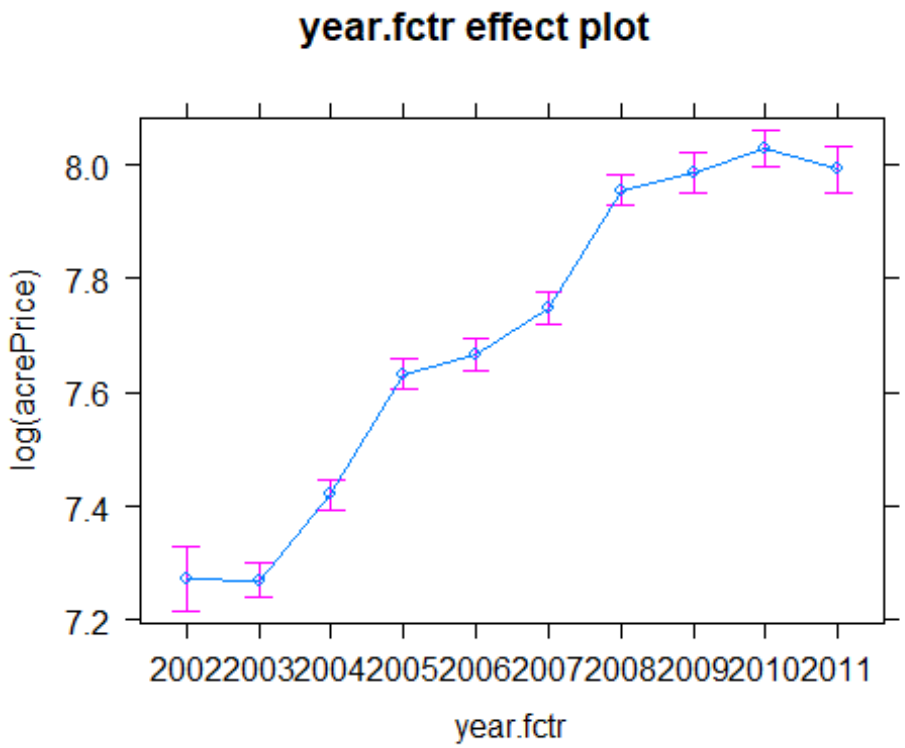
The coefficient for the log(acrePrice) for 2008 is 0.68. This means that the estimated average log(acrePrice) for 2008 is 7.27 + .68 or 7.95. Which means the estimated average acrePrice for 2008 is: $2835.58.

(c) Each of these two models can be used to (retrospectively) predict the expected log of sale price per acre from 2002 to 2011. Plot these predictions for the two models, and describe the differences.

```
plot(Effect("year", m.price))
```



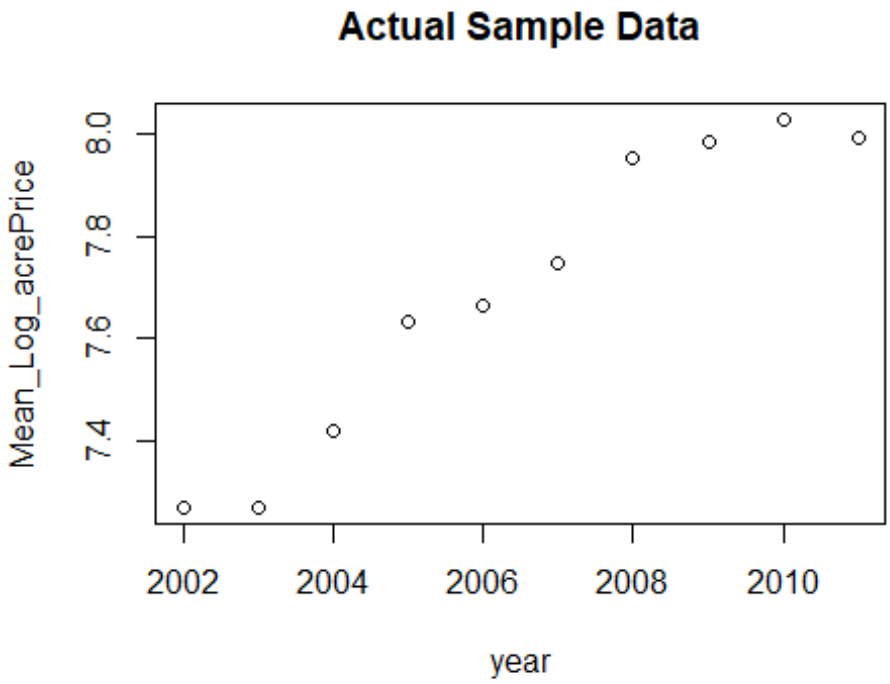

```
plot(Effect("year.fctr", m.price.fctr))
```



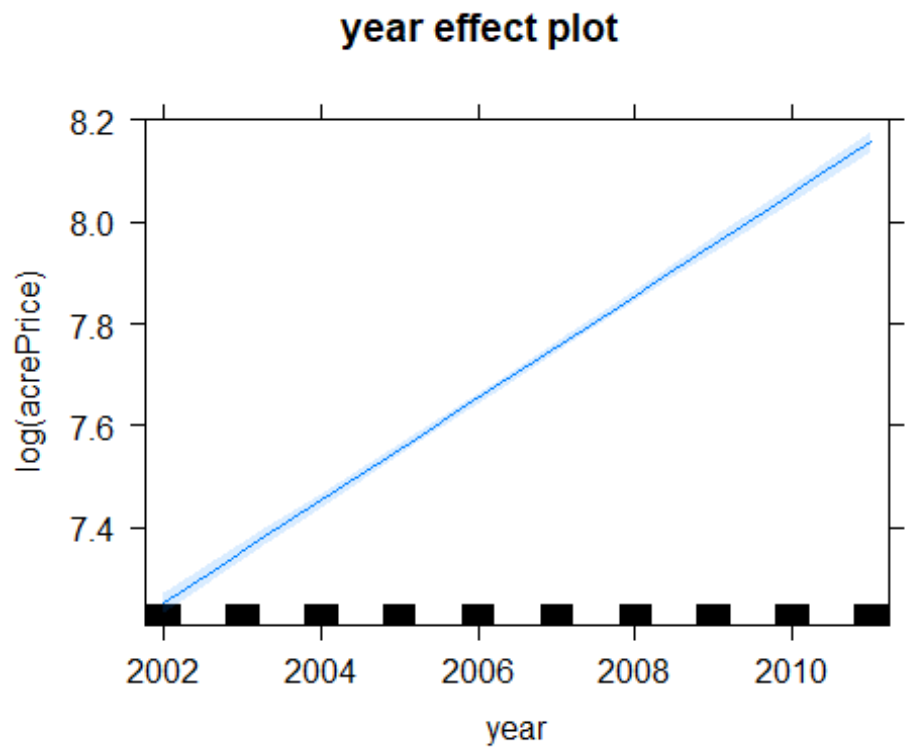
The linear model that uses year as a continuos variable predicts a perfectly linear increase over time. The linear model based on year as a factor shows more variation over time. For example the latest four years of data show considerable flattening.

(d) Which of these two models fits the data better? Support your answer using graphs or otherwise.

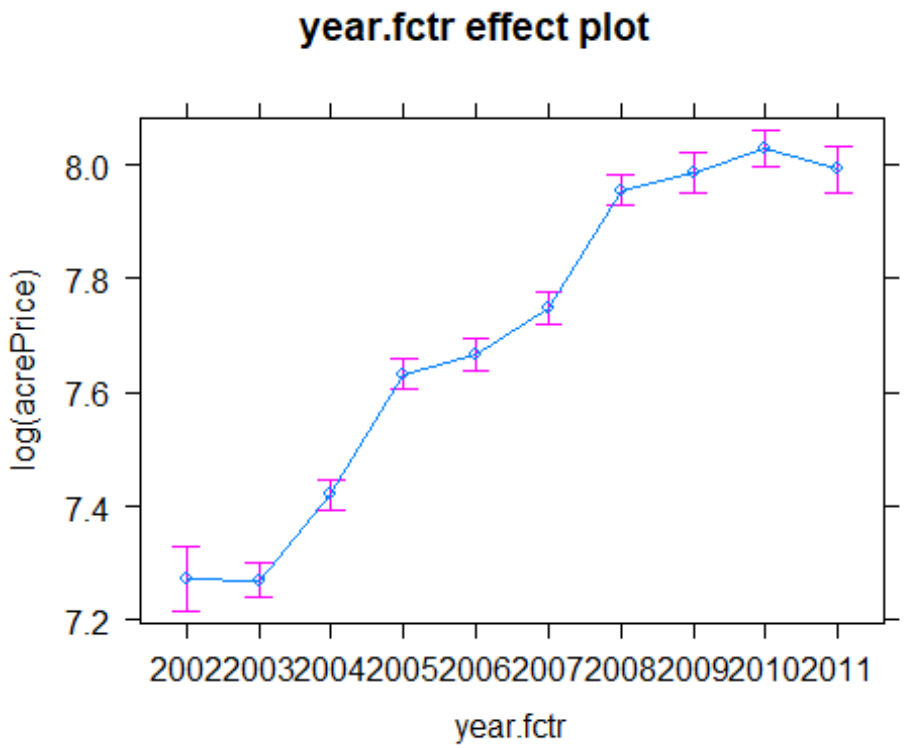
```
mnlnd$logAcrePrice = log(mnlnd$acrePrice)
mnlnd_grp = mnlnd %>%
              group_by(year) %>%
              dplyr::summarize(Mean_Log_acrePrice = mean(logAcrePrice,
na.rm=TRUE))
plot(Mean_Log_acrePrice ~ year, data = mnlnd_grp)
title(main = "Actual Sample Data")
```




```
plot(Effect("year", m.price))
```




```
plot(Effect("year.fctr", m.price.fctr))
```



The first graph plots the average log(acrePrice) by year from the sample data. The second graph show the linear model with year as a continuous variable. The third plot shows the linear model with year as a factor. The factored model definitely seems to more closely resemble the actual sample experience.

Also given each year is very likely not independent, considering the linear model with year as continuos variable is problematic.

3. (10 points.) The data set Moore in the package carData contains data from an experiment to see how conformity with someone else's opinion was related to the other person's status. Subjects were paired with a partner of either high or low status; the partners were secretly collaborators of the investigators. On 40 key questions, the partners were told to disagree with the subjects. The experimenters counted the number of times each subject "conformed" by changing their opinion to agree with their partner. Each subject was also (presumably before the experiment) given a questionnaire to measure their authoritarianism, as authoritarianism could potentially affect how the subject reacted to disagreement.

The variables in the data set are:**

. conformity: number of conforming responses-could potentially be 0 to 40; observed values ranged from 4 to 24

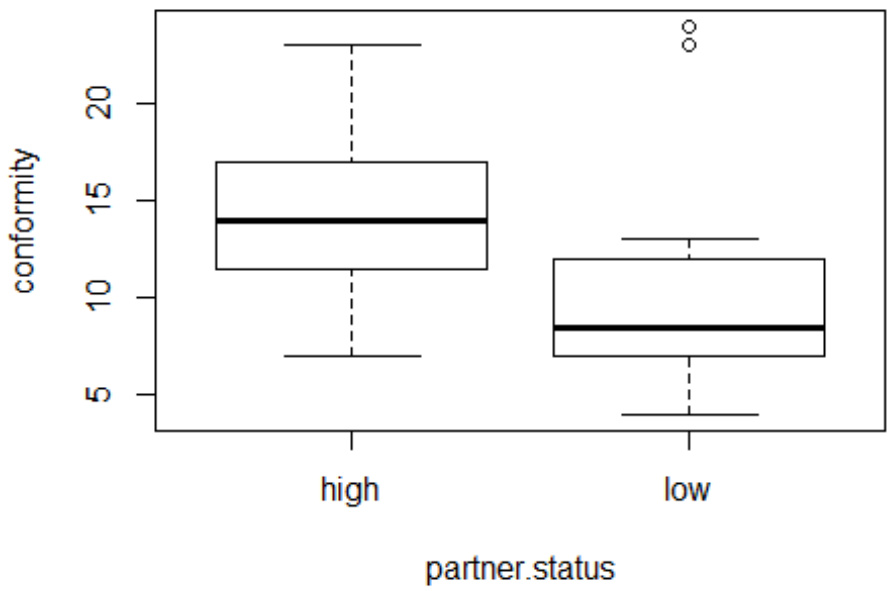
. partner.status: a factor: high or low

. fscore: authoritarianism score-observed values ranged from 15 to 68

The data frame also includes fcategory, a categorized version of fscore; ignore this.

(a) Show that there's evidence that partner.status affects conformity. (This might not require any regression. . .)

```
plot(conformity ~ partner.status, data = Moore)
```



Just looking at a simple plot of partner.status to conformity shows there are distinct differences both in the average conformity and the spread of the individual partner.status values over conformity. It looks like there is evidence to suspect that partner.status affects conformity.

(b) Does the effect of partner.status differ for people with different fscores? One way to look at this is to fit a linear model with conformity as the response and fscore, partner.status, and their interaction as regressors. Fit this model, give the P-value, and explain what the P-value means and what it tells you about whether the effect of partner.status differs for people with different fscores.

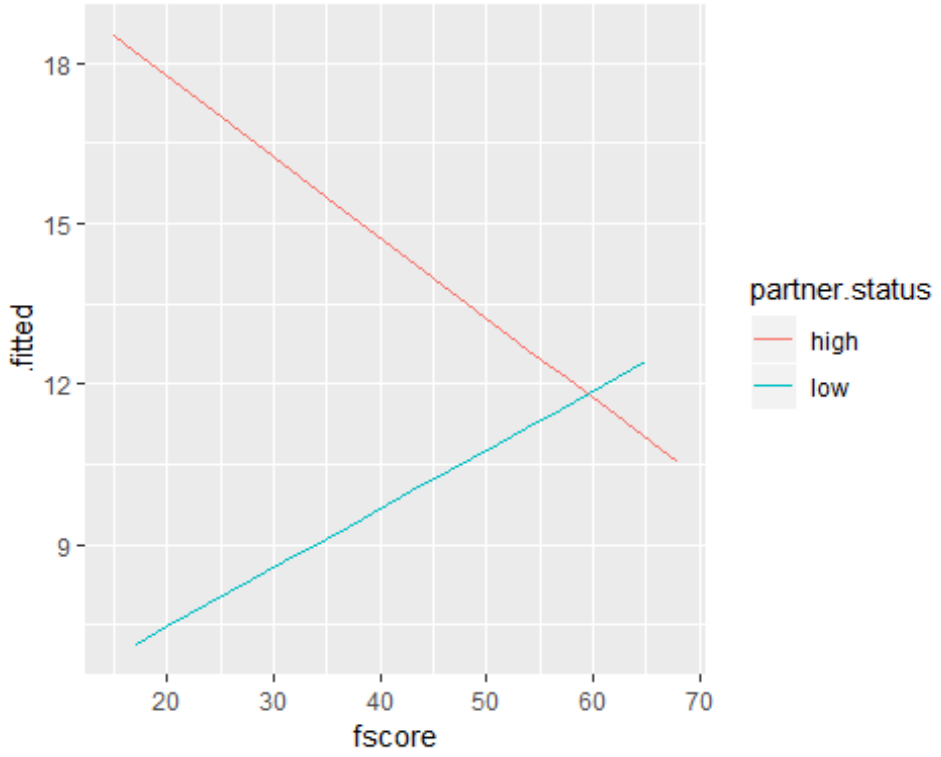
```
m.conformity = lm(conformity ~ partner.status + fscore +
partner.status:fscore, data = Moore)
get_regression_table(m.conformity)
```

```
## # A tibble: 4 x 7
##   term               estimate std_error statistic p_value lower_ci
upper_ci
##   <chr>                 <dbl>     <dbl>     <dbl>   <dbl>    <dbl>
<dbl>
## 1 intercept             20.8       3.26       6.37   0       14.2     27.4
## 2 partner.statuslow    -15.5       4.4       -3.53   0.001  -24.4     -6.65
## 3 fscore                -0.151     0.072     -2.11   0.041   -0.296   -
0.006
## 4 partner.statuslow~     0.261     0.097      2.69   0.01     0.065
0.457
```

The P-Value for the interaction coefficient for fscore and partner.status is 0.01. In essence, the P-Values tell you the effect of adding additional terms to the model. This is a low P-Value which suggest that adding this interaction term does seem significant. This suggests that the effect of partner.status does differ for people with different fscores.

(c) A broader question is how the effect of partner.status differs for people with different fscore. This is perhaps easiest to study graphically. Using your model in (b), make predictions for conformity for people with fscores ranging from 15 to 68, for both the high status and low status treatments. Plot these predictions on the same graph, clearly distinguishing between the lines for the high and low status groups (e.g. by color.) Assuming your model is close to right, what does this graph tell you about how the effect of partner.status differs for people with different fscores?

```
m.conformity.df = augment(m.conformity)
m.conformity.df.filt = m.conformity.df[which(m.conformity.df$fscore >= 15 &
m.conformity.df$fscore <= 68), ]
ggplot(m.conformity.df.filt, aes(x = fscore, y = .fitted, color =
partner.status)) + geom_line()
```

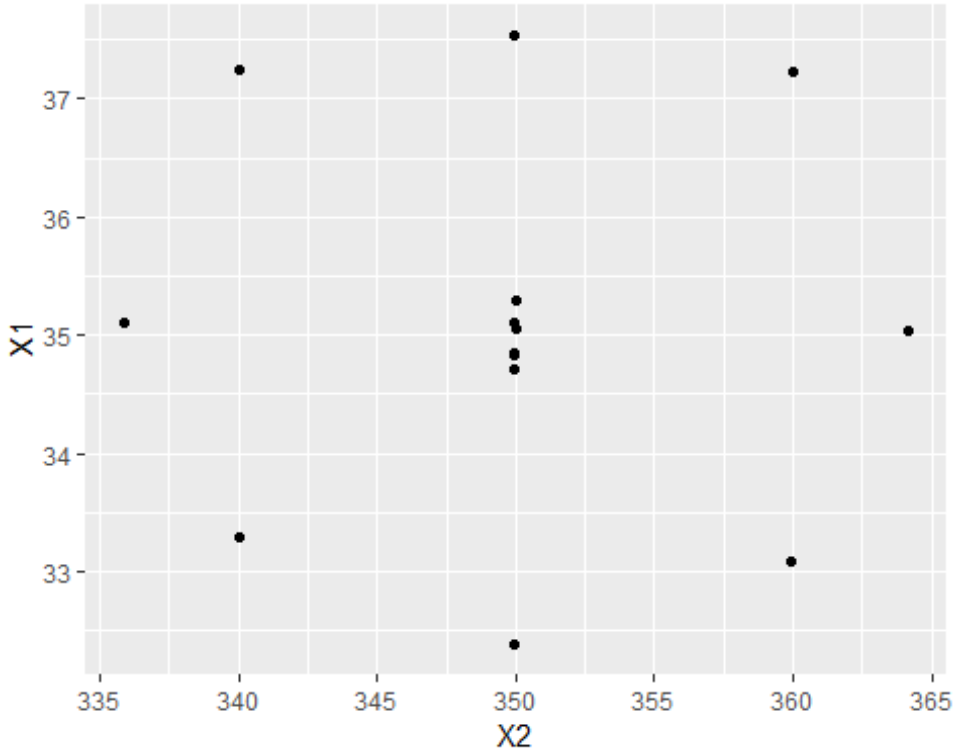


There is a clear difference in how the effect of partner.status differs for people with different fscores. We see that when partner.status is high the predicted conformity decreases as the fscore increases. When partner.status is low the predicted conformity actually increased with higher fscores.

4. (10 points.) The data set cakes contains data from a baking experiment using packaged cake mix. The response, Y, is a "palatability score" (higher is tastier.) The explanatory variables are X1, baking time in minutes, and X2, baking temperature in degrees Fahrenheit. (Ignore the block variable.)

(a) Show graphically that it is not appropriate to model expected palatability score as a linear function of X1 and X2. Explain why we should have known this even before we looked at the data.

```
ggplot(cakes, aes(X2, X1)) +
    geom_jitter(position = position_jitter(width = .1))
```



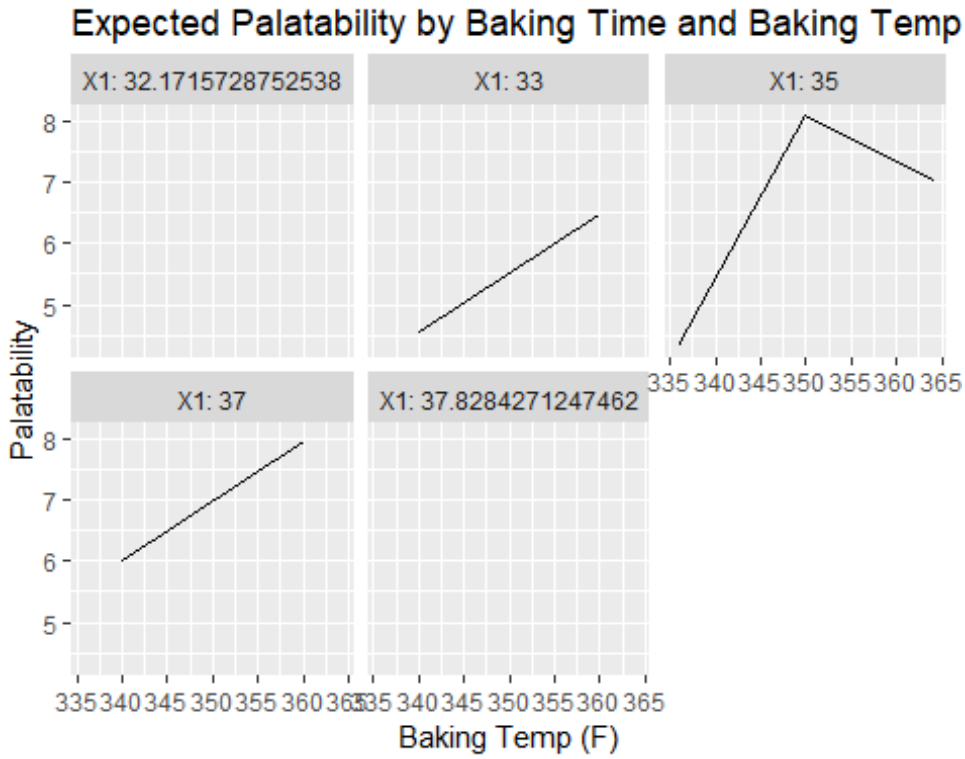
When plotting the data points with X1 and X2 on the axis, we can see a clear pattern. These two variables have seem to have a correlation and therefore it is not appropriate to model expected palatability as a linear function of both variables.

It is fairly intuitive to assume there is a relationship between cooking time and oven tempature.

(b) Fit a model to predict palatability score as the sum of quadratic functions of baking time and baking temperature. (For simplicity, we recommend you do not fit any interaction.) Display the fitted model graphically, e.g. through colored or faceted plots.

```
m.palatability = lm(Y ~ X1 + X2 + I(X1^2) + I(X2^2), data = cakes)
m.palatability.df = augment(m.palatability)
ggplot(m.palatability.df, aes(x = X2, y = .fitted)) + geom_line() +
facet_wrap(~ X1, labeller = label_both) + ylab("Palatability") + xlab("Baking
Temp (F)") + ggtitle("Expected Palatability by Baking Time and Baking Temp")

## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```



(c) For how long and at what temperature should you bake a cake using this mix to maximize the predicted palatability? (Hint: Recall from Calc I that a quadratic ax2 + bx + c is maximized at ???b/(2a) if a is negative.)

```
m.palatability = lm(Y ~ X1 + X2 + I(X1^2) + I(X2^2), data = cakes)
rslts = tidy(m.palatability)
rslts

## # A tibble: 5 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) -1695.     325.          -5.22 0.000550
## 2 X1             11.3      4.42         2.57 0.0304
## 3 X2              8.46     1.77         4.78 0.000996
## 4 I(X1^2)        -0.157    0.0632      -2.48 0.0348
## 5 I(X2^2)        -0.0120   0.00253     -4.73 0.00107
```

```
max_X1 = ( - rslts[2, 2]) / (2 * rslts[4, 2])
max_X1
```

```
##   estimate
## 1  36.1715
```

```
max_X2 = ( - rslts[3, 2]) / (2 * rslts[5, 2])
max_X2
```

```
##   estimate
## 1 354.0332
```

The baking time and tempature for the predicted maximum palatablility are 36.2 minutes and 354 degrees F, repsectively.

5. (10 points.) Returning to the MinnLand data set, one subject the data was collected to answer was the relationship between sale price per acre and crpPct, the percentage of the land enrolled in the Conservation Reserve Program. However, there are many potential confounding variables associated with crpPct that could affect sale prices. For example, land in the Conservation Reserve Program is disproportionately in northwest Minnesota, and sale prices in northwest Minnesota tend to be lower than in the rest of the state for reasons that may have less to do with the program than with negative temperatures in the winter. One way to study this would be to fit models that include both crpPct and region as predictors. However, it is not clear a priori whether an interaction between crpPct and region will help.

(a) Fit a linear regression model to predict log(acrePrice) from crpPct and region, with no interaction. State the coefficient of crpPct in this model, and explain what this coefficient tells you about the relationship between crpPct and log(acrePrice).

```
m.acrePrice = lm(log(acrePrice) ~ crpPct + region, data = MinnLand)
summary(m.acrePrice)
```

```
##
## Call:
## lm(formula = log(acrePrice) ~ crpPct + region, data = MinnLand)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1365 -0.3363  0.0076  0.3577  2.5741
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.8831275  0.0092553  743.70   <2e-16 ***
## crpPct              -0.0048681  0.0002388  -20.38   <2e-16 ***
## regionWest Central   0.7565716  0.0130780   57.85   <2e-16 ***
## regionCentral        1.0451863  0.0124530   83.93   <2e-16 ***
## regionSouth West     1.0333409  0.0140783   73.40   <2e-16 ***
## regionSouth Central  1.2577899  0.0137854   91.24   <2e-16 ***
## regionSouth East     1.2639919  0.0153082   82.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5464 on 18693 degrees of freedom
## Multiple R-squared:  0.4335, Adjusted R-squared:  0.4334
## F-statistic:  2384 on 6 and 18693 DF,  p-value: < 2.2e-16
```

Based on this model the coefficient for crpPct is near zero, which suggest that there isn't a relationship between crpPct and acrePrice.

(b) Fit a regression model to predict log(acrePrice) from crpPct and region with an interaction. Explain what this model tells you about the relationship between crpPct and log(acrePrice).

```
m.acrePrice.intct = lm(log(acrePrice) ~ crpPct + region + crpPct:region, data
= MinnLand)
summary(m.acrePrice.intct)
```

```
##
## Call:
## lm(formula = log(acrePrice) ~ crpPct + region + crpPct:region,
##     data = MinnLand)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.13296 -0.33644  0.00808  0.35798  2.57467
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 6.8825323  0.0095174 723.150  < 2e-16 ***
## crpPct                     -0.0048146  0.0003118 -15.439  < 2e-16 ***
## regionWest Central          0.7530229  0.0137536  54.751  < 2e-16 ***
## regionCentral               1.0463803  0.0127785  81.886  < 2e-16 ***
## regionSouth West            1.0416076  0.0145350  71.662  < 2e-16 ***
## regionSouth Central         1.2585863  0.0140545  89.550  < 2e-16 ***
## regionSouth East            1.2610612  0.0155927  80.875  < 2e-16 ***
## crpPct:regionWest Central   0.0007133  0.0006115   1.167  0.24340
## crpPct:regionCentral       -0.0004686  0.0009388  -0.499  0.61766
## crpPct:regionSouth West    -0.0028393  0.0008824  -3.218  0.00129 **
## crpPct:regionSouth Central -0.0002849  0.0014823  -0.192  0.84760
## crpPct:regionSouth East     0.0032408  0.0015620   2.075  0.03802 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5462 on 18688 degrees of freedom
## Multiple R-squared:  0.4341, Adjusted R-squared:  0.4338
## F-statistic:  1303 on 11 and 18688 DF,  p-value: < 2.2e-16
```

Including the interaction term again shows that all of the coefficients for interactions by region are close to zero. This, again suggests that there isn't a relationship between crpPct and acrePrice.

(c) Perform an ANOVA to compare your models from parts (a) and (b). State the P-value that you get, and explain what, if anything, this P-value tells you.

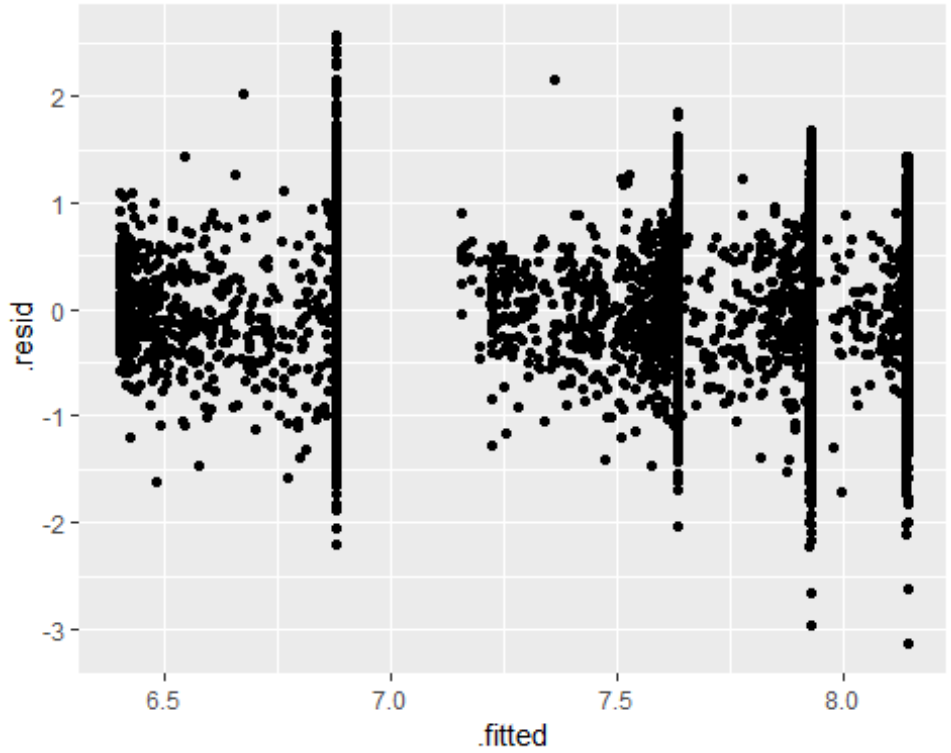
```
anova(m.acrePrice, m.acrePrice.intct)

## Analysis of Variance Table
##
## Model 1: log(acrePrice) ~ crpPct + region
## Model 2: log(acrePrice) ~ crpPct + region + crpPct:region
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1  18693 5581.7
## 2  18688 5576.2  5    5.4969 3.6844 0.002469 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The P-Value is 0.002, which is very low. This suggest that there is evidence that the addition of the interaction term does positively contribute to the model.

(d) Your ANOVA in part (c) made certain assumptions. Check the residuals of your model from (b) to see if these assumptions are close to satisfied.

```
m.acrePrice.intct = lm(log(acrePrice) ~ crpPct + region + crpPct:region, data
= MinnLand)
m.acrePrice.intct.df = augment(m.acrePrice.intct)
ggplot(m.acrePrice.intct.df, aes(x = .fitted, y = .resid)) + geom_point()
```

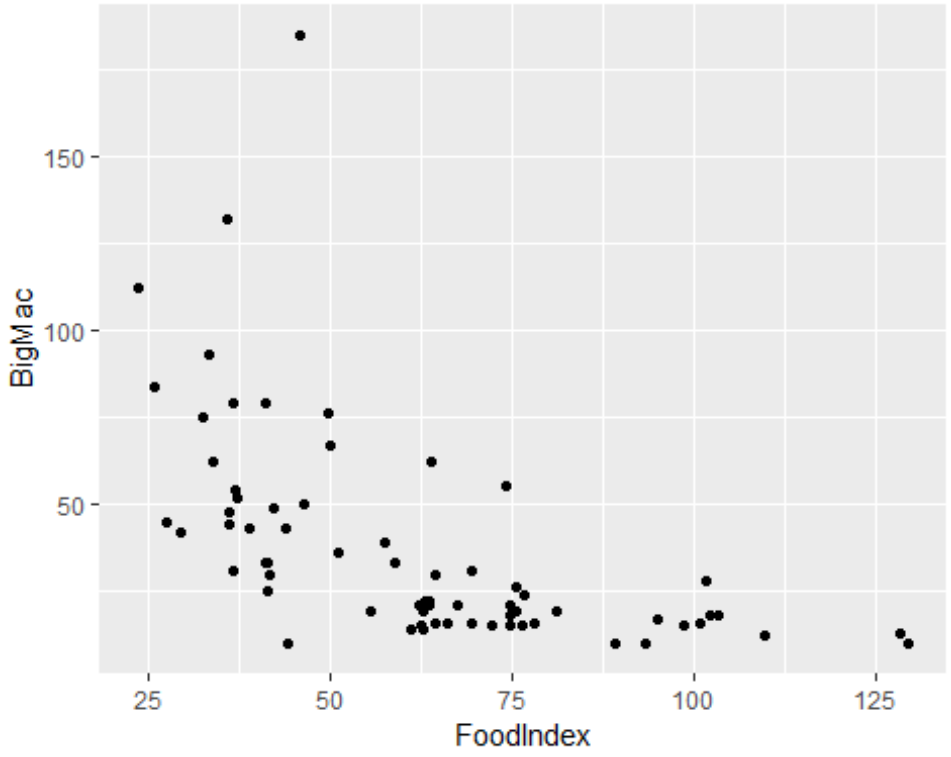


ANOVA assumes heteroskedacticity of the residuals. Plotting the fitted values against the residual there are definite differences particularly at higher fitted values. This is an indication that the assumption does not hold.

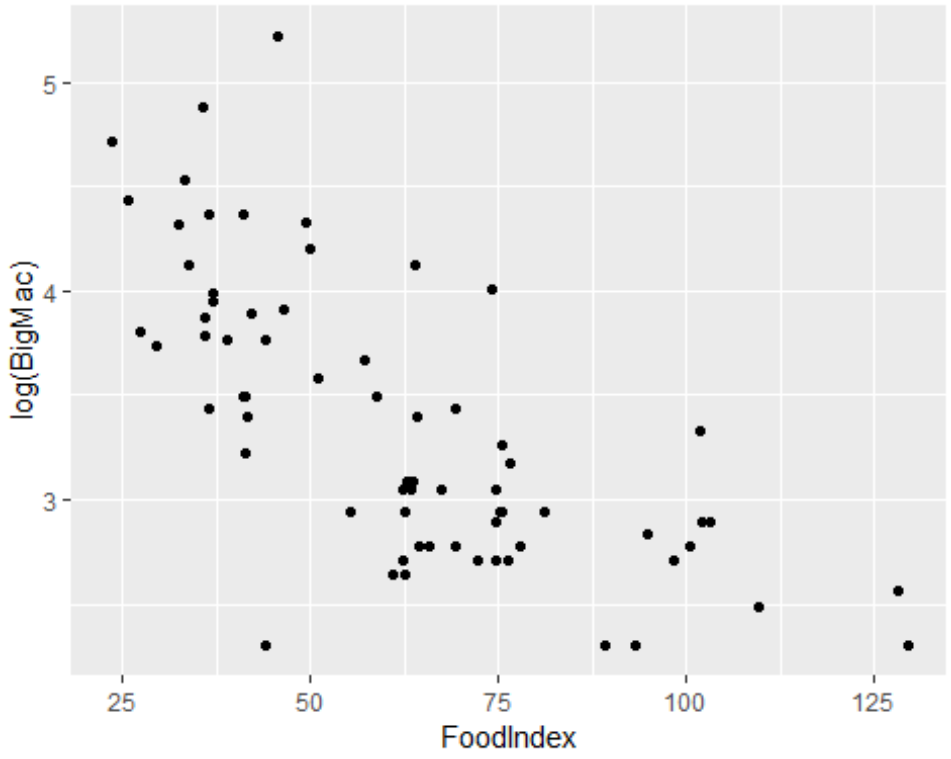
6. (10 points.) The data set BigMac2003 in alr4 gives the price of a Big Mac in 2003 (BigMac), measured in minutes of labor required to buy one, in 69 cities. Of the many potential explanatory variables in the data set, FoodIndex, a measure of food prices (relative to a baseline where Zurich is 100), both logically makes sense as a predictor and has a fairly strong correlation with BigMac. We thus wish to first try a model to predict BigMac from FoodIndex, but these variables may require transformation.

(a) Choose interpretable transformations to apply to FoodIndex and BigMac, such that the relationship between the transformed variables is approximately linear. (Note that you may choose "no transformation" for either variable.) Justify your choice using graphs or otherwise.

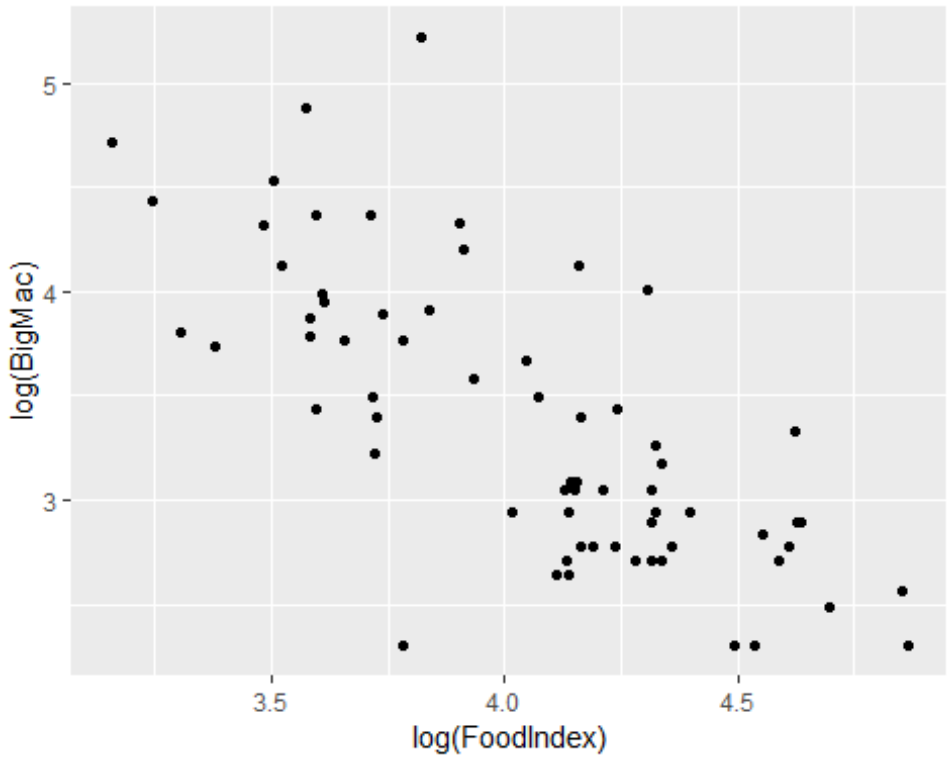
```
ggplot(BigMac2003, aes(x = FoodIndex, y = BigMac)) + geom_point()
```



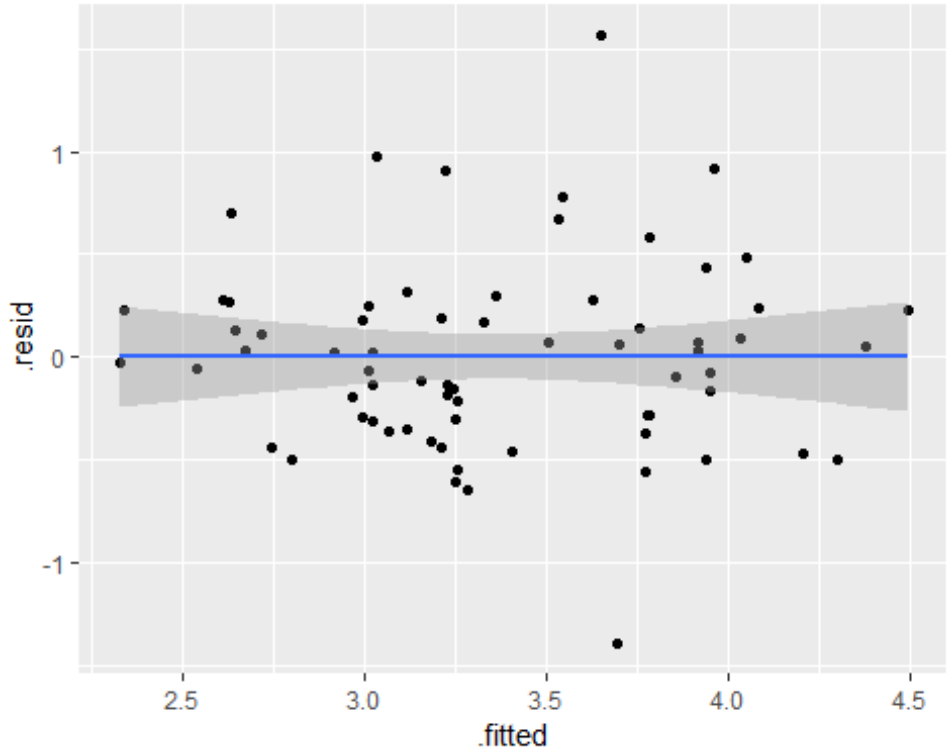

```
ggplot(BigMac2003, aes(x = FoodIndex, y = log(BigMac))) + geom_point()
```




```
ggplot(BigMac2003, aes(x = log(FoodIndex), y = log(BigMac))) + geom_point()
```




```
m.bigmac = lm(log(BigMac) ~ log(FoodIndex), data = BigMac2003)
m.bigmac.df = augment(m.bigmac)
ggplot(m.bigmac.df, aes(x = .fitted, y = .resid)) + geom_point() +
geom_smooth(method = "gam", formula = y ~ s(x))
```



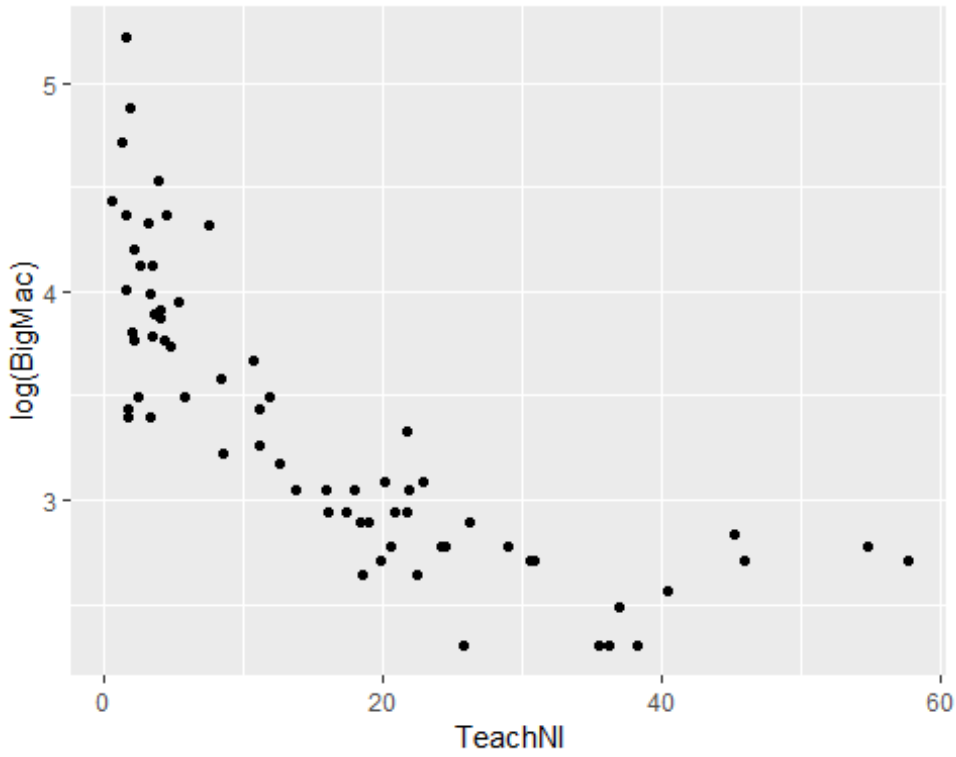
The first graphs plots the actual values of FoodIndex to BigMac. This shows a defintie right-skewdness. This suggest a ptoential log transformation. A log transformation is also am easily interpretable transformation. Apply a log transformation to the BigMac response variable (graph 2) helps to remove some of the skewness. However there does still seem to be a nonlinear trend in this data. The third graph includes a log transformation of both BigMac and Food index. This definitely seems like a more linear trend.

The final graph is a plot of the residuals based on a linear model applied to the log transformation of both variables. The residula plot confirms that these transformations do produce a nice linerarity.

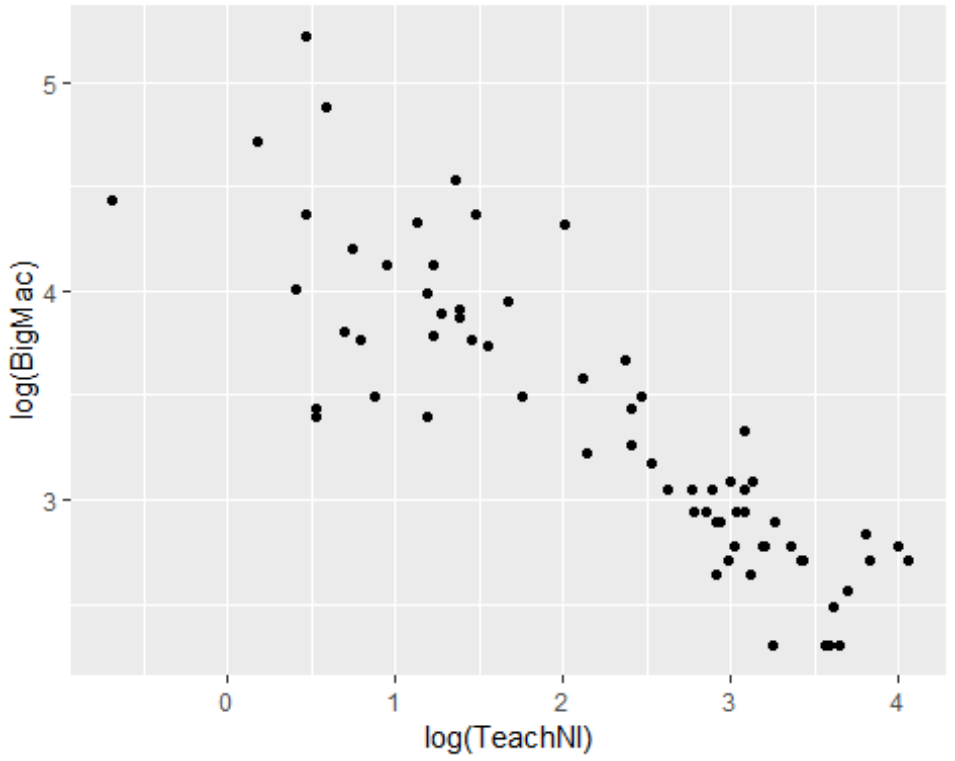
(b) The model can straightforwardly be improved by adding another predictor. Fit a better model that predicts the transformed BigMac variable from FoodIndex and one other variable. Convince the grader that your model is an improvement. (Your model may include complex regressors such as interactions if you wish.)

After exploring the data associated with BigMac2003, it was determined that a reasonable addition to this model would be the TeachNI ("Primary teacher's net income, 1000s of US dollars"). This seems reasonble as the BigMac price is likely a function of overall food prices and the cost of labor.

```
ggplot(BigMac2003, aes(x = TeachNI, y = log(BigMac))) + geom_point()
```



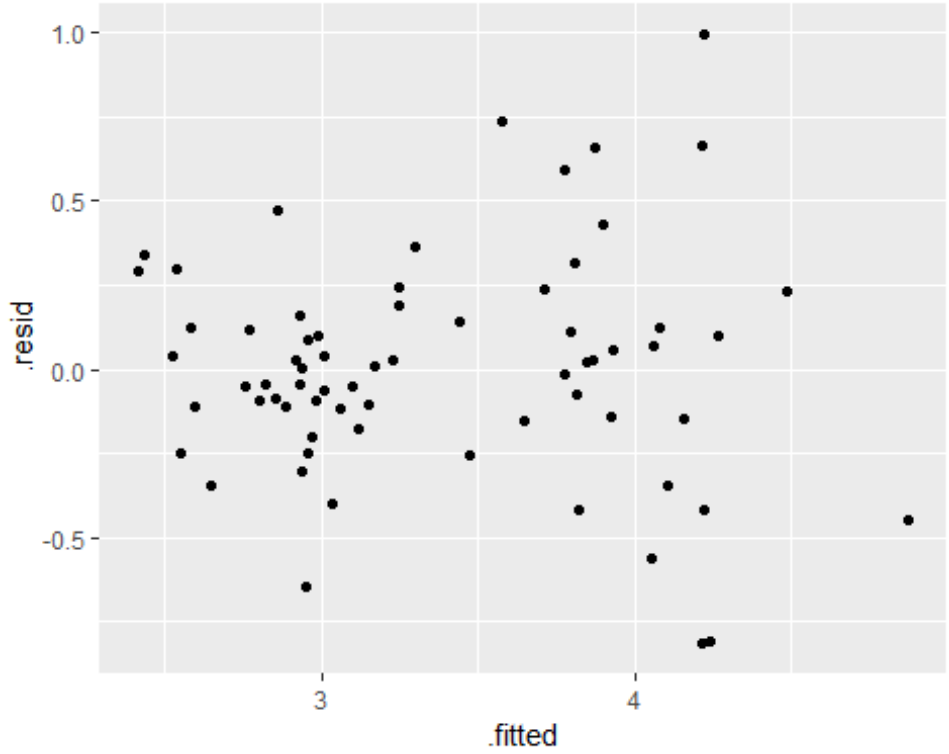

```
ggplot(BigMac2003, aes(x = log(TeachNI), y = log(BigMac))) + geom_point()
```



```r
m.bigmac = lm(log(BigMac) ~ log(FoodIndex), data = BigMac2003)
m.bigmac.plus = lm(log(BigMac) ~ log(FoodIndex) + log(TeachNI), data =
BigMac2003)
anova(m.bigmac, m.bigmac.plus)
```

```
## Analysis of Variance Table
##
## Model 1: log(BigMac) ~ log(FoodIndex)
## Model 2: log(BigMac) ~ log(FoodIndex) + log(TeachNI)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     67 14.3348
## 2     66  7.6639  1    6.6709 57.448 1.498e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
m.bigmac.plus.df  = augment(m.bigmac.plus)
ggplot(m.bigmac.plus.df, aes(x = .fitted, y = .resid)) + geom_point()
```




```r
m.bigmac.plus2 = lm(log(BigMac) ~ log(FoodIndex) + log(TeachNI) +
log(FoodIndex):log(TeachNI), data = BigMac2003)
anova(m.bigmac.plus, m.bigmac.plus2)
```

```
## Analysis of Variance Table
##
## Model 1: log(BigMac) ~ log(FoodIndex) + log(TeachNI)
## Model 2: log(BigMac) ~ log(FoodIndex) + log(TeachNI) +
log(FoodIndex):log(TeachNI)
```

```
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     66 7.6639
## 2     65 7.6495  1  0.014378 0.1222 0.7278
```

The first graph considers the relationship between TeachNI and the log of BigMac. Again we see some left-skewdness. Applying an additional log transfomation to TeachNI (second graph), produces a more linear relationship.

Fitting a model that includes the log transformations for FoodIndex and TeachNI and performing an anova test prduces a P-Value very near zero which does suggest the additional variable improves the model. Plotting the residuals for the added filed model (graph 3) also demonstrates that the ANOVA assumption of heteroskedacticity is fairly reasonable.

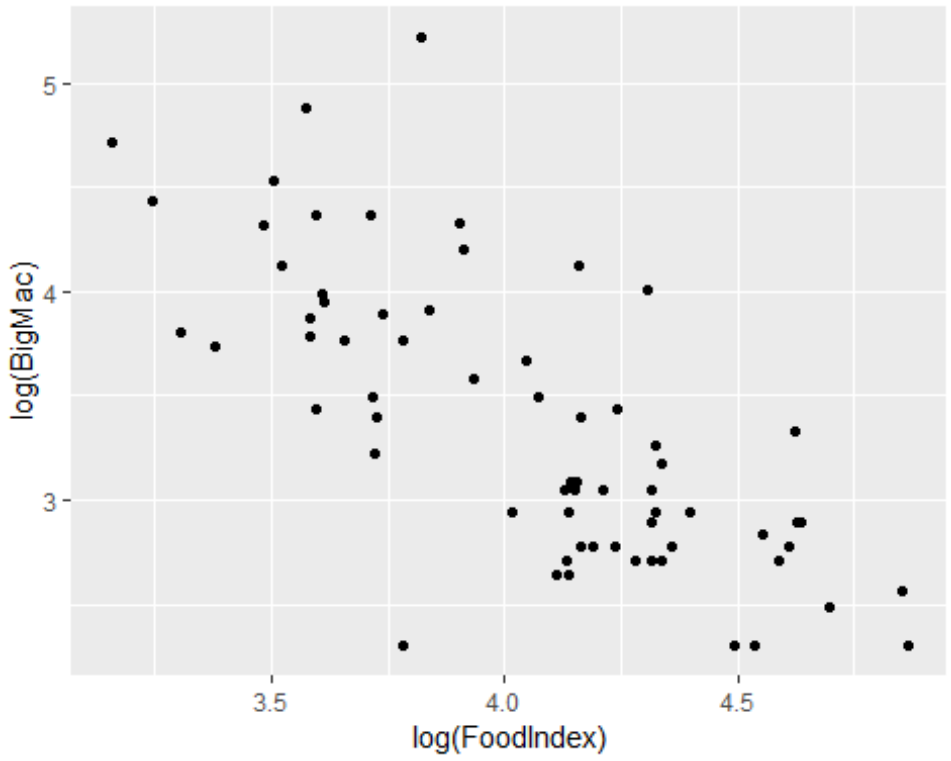
Finally, adding an interaction term for FoodIndex and TeachNI and running an ANOVA against the model with the non-interaction prodices a P-Value of .73, which suggests that addding the interaction term does not improve the model.

(c) Using numbers, graphs, and words, explain what your improved model tells you about how the price of Big Macs relates to the Food Index and your other explanatory variable.

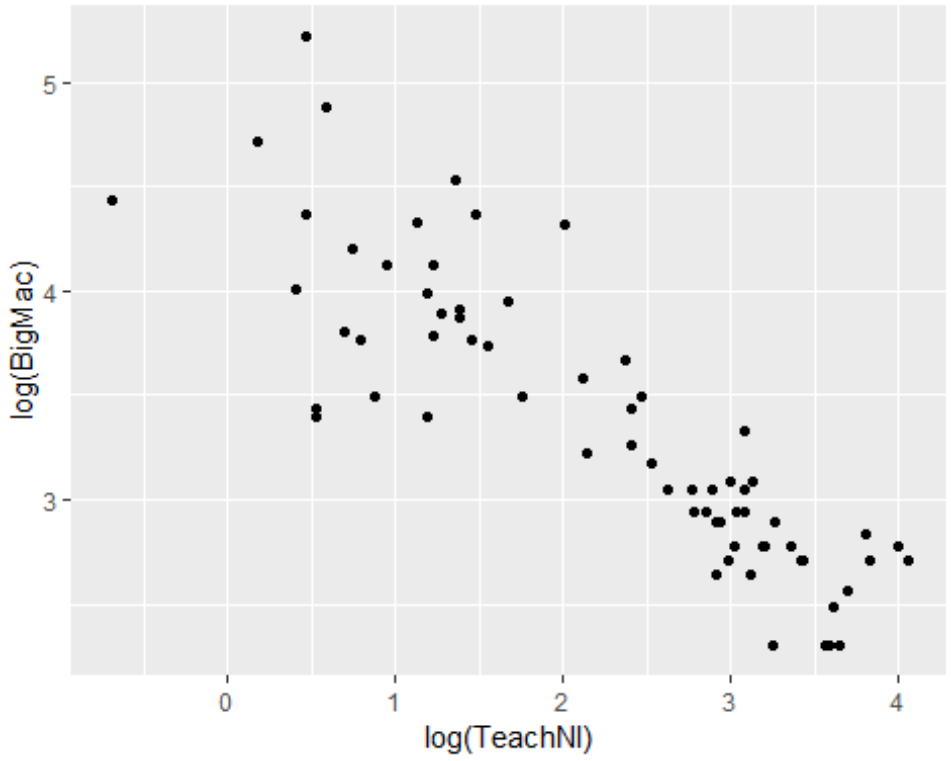
```
m.bigmac.plus = lm(log(BigMac) ~ log(FoodIndex) + log(TeachNI), data =
BigMac2003)
summary(m.bigmac.plus)

##
## Call:
## lm(formula = log(BigMac) ~ log(FoodIndex) + log(TeachNI), data =
BigMac2003)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81326 -0.15321  0.00609  0.14247  0.99675
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.20490    0.60251   8.639 1.91e-12 ***
## log(FoodIndex) -0.20007    0.17441  -1.147    0.255
## log(TeachNI)   -0.46175    0.06092  -7.579 1.50e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3408 on 66 degrees of freedom
## Multiple R-squared:  0.7618, Adjusted R-squared:  0.7546
## F-statistic: 105.5 on 2 and 66 DF,  p-value: < 2.2e-16

ggplot(BigMac2003, aes(x = log(FoodIndex), y = log(BigMac))) + geom_point()
```




```
ggplot(BigMac2003, aes(x = log(TeachNI), y = log(BigMac))) + geom_point()
```



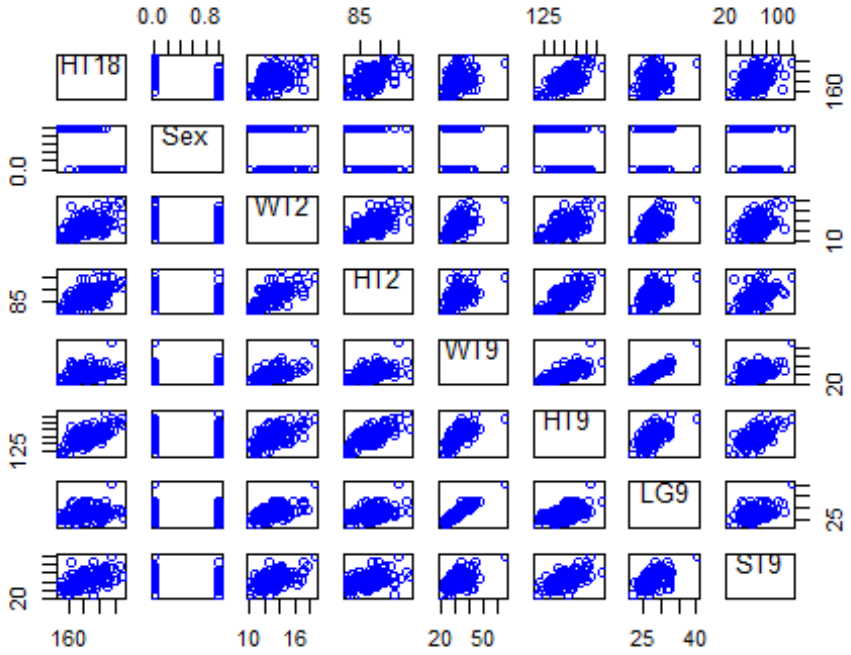
Looking at the coefficients produced by this enhanced model, the relationship between both FoodIndex and TechNI is negatively correlated with BigMac price. This makes sense when looking at the scatterplots of the actual amaple data. We see that as FoodIndex and TeachNI increase the BigMac price decreases.

7. (5 points.) The data set BGSall in alr4 gives measurements on all subjects of the Berkeley Guidance Study, both male and female. Our goal is to find the model that best predicts height at age 18 (HT18 gives this height in cm) using measurements available at age 9: Sex, WT2, HT2, WT9, HT9, LG9, and ST9. See ?BGSall for definitions of all these variables. Find the best predictive model you can. You may not transform HT18 but you may transform any of the predictors. You may also consider interactions and higher-order terms. AIC isn't the be-all and end-all, but I managed to get a model with an AIC of 708.1 without looking too hard, so your model's AIC should get close to that. In addition, you should give some measure of how large you would expect the prediction errors if your model was applied to individuals similar to those in the data set (children born in California in 1928-29.)

```
var1 = c("HT18", "Sex", "WT2", "HT2", "WT9", "HT9", "LG9", "ST9")
summary(BGSall[,var1])
```

```
##       HT18            Sex              WT2             HT2       
##  Min.   :153.6   Min.   :0.0000   Min.   :10.10   Min.   :80.90  
##  1st Qu.:166.2   1st Qu.:0.0000   1st Qu.:12.20   1st Qu.:85.97  
##  Median :172.5   Median :1.0000   Median :13.20   Median :87.70  
##  Mean   :172.6   Mean   :0.5147   Mean   :13.21   Mean   :87.80  
##  3rd Qu.:179.2   3rd Qu.:1.0000   3rd Qu.:14.10   3rd Qu.:89.70  
##  Max.   :195.1   Max.   :1.0000   Max.   :18.60   Max.   :98.20  
##       WT9             HT9             LG9             ST9       
##  Min.   :19.90   Min.   :121.4   Min.   :21.80   Min.   : 22.00  
##  1st Qu.:27.88   1st Qu.:132.5   1st Qu.:26.30   1st Qu.: 57.00  
##  Median :30.90   Median :135.7   Median :27.30   Median : 64.00  
##  Mean   :31.63   Mean   :135.5   Mean   :27.68   Mean   : 64.57  
##  3rd Qu.:34.12   3rd Qu.:139.4   3rd Qu.:28.65   3rd Qu.: 73.00  
##  Max.   :66.80   Max.   :152.5   Max.   :40.40   Max.   :121.00
```

```
scatterplotMatrix(BGSall[,var1],smooth = F, regLine = F, diagonal = F)
```




```
dd1 = powerTransform(cbind(WT2, HT2, WT9, HT9, LG9, ST9) ~ 1, BGSall)
summary(dd1)

## bcPower Transformations to Multinormality
##     Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## WT2   -1.4683          -1      -2.5276      -0.4091
## HT2   -2.7857           0      -5.9128       0.3415
## WT9   -1.0069          -1      -1.4805      -0.5334
## HT9   -1.1725           1      -3.6436       1.2987
## LG9   -1.4653          -1      -2.4819      -0.4486
## ST9    0.6799           1       0.2422       1.1175
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                       LRT df       pval
## LR test, lambda = (0 0 0 0 0 0) 36.01753  6 2.7351e-06
##
## Likelihood ratio test that no transformations are needed
##                                       LRT df       pval
## LR test, lambda = (1 1 1 1 1 1) 96.34636  6 < 2.22e-16

m.ht18 = lm(HT18 ~ WT9 + HT9, data = BGSall)
m.ht18.intct = lm(HT18 ~ WT9 + HT9 + WT9:HT9, data = BGSall)
m.ht18.plus = lm(HT18 ~ WT9 + HT9 + LG9 + ST9, data = BGSall)
m.ht18.plus2 = lm(HT18 ~ WT9 + HT9 + LG9 + ST9 + WT9:LG9, data = BGSall)

anova(m.ht18, m.ht18.intct)
```

```
## Analysis of Variance Table
##
## Model 1: HT18 ~ WT9 + HT9
## Model 2: HT18 ~ WT9 + HT9 + WT9:HT9
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    133 5861.9
## 2    132 5861.6  1   0.30507 0.0069 0.9341
```

```
anova(m.ht18, m.ht18.plus)
```

```
## Analysis of Variance Table
##
## Model 1: HT18 ~ WT9 + HT9
## Model 2: HT18 ~ WT9 + HT9 + LG9 + ST9
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    133 5861.9
## 2    131 5380.8  2    481.13 5.8568 0.003662 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m.ht18.plus, m.ht18.plus2)
```

```
## Analysis of Variance Table
##
## Model 1: HT18 ~ WT9 + HT9 + LG9 + ST9
## Model 2: HT18 ~ WT9 + HT9 + LG9 + ST9 + WT9:LG9
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    131 5380.8
## 2    130 5364.6  1    16.177 0.392 0.5323
```

```
summary(m.ht18.plus)
```

```
##
## Call:
## lm(formula = HT18 ~ WT9 + HT9 + LG9 + ST9, data = BGSall)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.7750  -4.8799   0.2831   5.3425  11.5410
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 47.22095   21.42016   2.205  0.02923 *
## WT9          0.06539    0.27256   0.240  0.81078
## HT9          1.09588    0.15095   7.260 3.07e-11 ***
## LG9         -1.21209    0.58947  -2.056  0.04174 *
## ST9          0.12932    0.04396   2.942  0.00386 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.409 on 131 degrees of freedom
```

```
## Multiple R-squared:  0.4905, Adjusted R-squared:  0.4749
## F-statistic: 31.52 on 4 and 131 DF,  p-value: < 2.2e-16
```

```
AIC(m.ht18.plus2)
```

```
## [1] 899.7403
```

The scatterplots of each of these variables shows a fair amount of linearity with transformation. This is confimred by performing the power transformation test which suggests that the only term that could benefit from transfromation is the HT2. However, due to the probable correlation between the age 2 variables and the age 9 variables, we chose to exclude the age 2 variables from the model. It also makes mroe intuitive sense that age 9 variables will be more powerful when predicting a factor related to age 18.

Starting with a simple model that includes only WT9 and HT9 we compare this to more complex models using ANOVA. Adding an interaction term between WT9 and HT9 produced an ANOVA P-Value of 0.93, suggesting the addition of the interaction term did not improve the model.

Next we compared the simple model to a model that includes all variables except the age 2 variables. Comparing these models using ANOVA produced a P-VALU close to zero, suggesting these additional variables do improve the model.

Finally, we added an interaction term between WT9 and LG9 (maybe an intuitive relationship) and compared this using ANOVA to previous model which produce a P-Value of 0.53. This suggests this interaction term didn't improve the model.

The best model we produces seemed to be a simple linear model using the valriables: WT9, HT9, LG9 and ST9. with no transformations, as follows:

```
m.ht18.plus = lm(HT18 ~ WT9 + HT9 + LG9 + ST9, data = BGSall)
```

```
summary(m.ht18.plus)
```

```
##
## Call:
## lm(formula = HT18 ~ WT9 + HT9 + LG9 + ST9, data = BGSall)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.7750  -4.8799   0.2831   5.3425  11.5410
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 47.22095   21.42016   2.205  0.02923 *
## WT9          0.06539    0.27256   0.240  0.81078
## HT9          1.09588    0.15095   7.260 3.07e-11 ***
## LG9         -1.21209    0.58947  -2.056  0.04174 *
## ST9          0.12932    0.04396   2.942  0.00386 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.409 on 131 degrees of freedom
## Multiple R-squared:  0.4905, Adjusted R-squared:  0.4749
## F-statistic: 31.52 on 4 and 131 DF,  p-value: < 2.2e-16
```

Looking at the predicted residuals we see:

```
m.ht18.plus.df = augment(m.ht18.plus)
ggplot(m.ht18.plus.df, aes(x = .fitted, y = .resid)) + geom_point() +
geom_smooth(method = "gam", formula = y ~ s(x))
```