Module 4 Assessment

Include your name

3/11/2022

Instructions

Please submit the markdown file and compiled pdf to canvas before Mar 21 at 11:59pm. For this assignment, you can discuss with classmates, but please at least attempt to go through it individually first so that you can see what you understand or don't understand. Ultimately, the final product you turn in should be your own work. So you can discuss questions with classmates, but your answers should be written in your own words.

Intro

In this module assessment, you'll be considering data from "Soil nutrients influence growth response tree species to drought" by Levesque, Walthert, and Weber (Journal of Ecology, 2016). The authors consider how certain properties of soil and climate are associated with tree growth. The data we will be using today can be accessed at this link: https://datadryad.org/stash/dataset/doi:10.5061%2Fdryad.sd40d, but can also be taken directly from my github using the code chunk below.

In particular, the authors consider Basal area increment (BAI) (i.e., a measure of tree cross section growth) as the dependent variable and measure 538 trees across 52 sites in Switzerland and northwestern Italy. Measurements were taken on the same trees annually from 1957-2006. Take a few minutes to read skim the study (at least the abstract) to get a sense for the scientific question of interest.

```
## load libraries which we will need later
library("sandwich")
library("lmtest")
library("lme4")
library("lmboot")


# Load in the data
tree_growth <- read.csv("https://raw.githubusercontent.com/ysamwang/btry6020_sp22/main/lectureData/Tree
# Remove data points with missing data
tree_growth <- na.omit(tree_growth)

## Dimensions of data set
dim(tree_growth)
```

```
## [1] 26627    17
```

```
# variables in data set
names(tree_growth)
```

```
##  [1] "year"    "site"    "siteid"  "species" "treeid"  "trw"     "bai"
##  [8] "radius"  "age"     "awc100"  "pH"      "BS"      "C.N"     "tmean"
## [15] "prec"    "dri"     "co2"
```

Question 1 (2 pts)

The authors are interested in the dependent variable Basal area increment which is in the column named `bai`. Specifically, they take the log transform (and add 1 so that they avoid taking the log of 0) and ultimately use `log(bai + 1)`. Fit a linear model which examines the association between `log(bai +1)` and the covariates: radius (radius of the tree), age, prec (which is precipitation in mm) and co2 (atmospheric CO_2). Don't transform the covariates. Write a sentence about the interpretation of the estimated coefficient corresponding to `prec`. You can interpret the model as if we aren't adding the additional 1 to `bai` inside the log.

```
### Fit the linear model specified above here
reg_model <- lm(log(bai + 1)~ radius + age + prec + co2, data = tree_growth)
summary(reg_model)
```

```
##
## Call:
## lm(formula = log(bai + 1) ~ radius + age + prec + co2, data = tree_growth)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.59982 -0.30271  0.03646  0.34771  1.81438
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  1.954e+00  5.961e-02   32.781  < 2e-16 ***
## radius       1.054e-01  5.506e-04  191.349  < 2e-16 ***
## age         -1.062e-02  9.025e-05 -117.667  < 2e-16 ***
## prec         3.711e-04  2.029e-05   18.289  < 2e-16 ***
## co2         -5.140e-04  1.822e-04   -2.821  0.00479 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5156 on 26622 degrees of freedom
## Multiple R-squared:  0.6201, Adjusted R-squared:  0.6201
## F-statistic: 1.087e+04 on 4 and 26622 DF,  p-value: < 2.2e-16
```

Answer to Question 1: interpretation of estimated coefficient

Since we are taking the log transform of BAI, the estimated coefficient corresponds to $100(\exp .0003711 - 1) = 0.037$ percentage change in BAI. Thus, two trees who differ by 1 mm of precipitation but have the same age, radius and co2 levels, have an expected 0.037% difference in BAI.

(1 point) Students should not interpret the estimate in a causal manner, but as an association (1 point) Students should use the % change interpretation and calculate the value correctly.

Question 2 (1 pt)

Using the model above, form a 95% confidence interval for the coefficient corresponding to the coefficient corresponding to `prec`.

```
### Calculate the lower and upper part of the CI here
coefci(reg_model)
```

```
##                      2.5 %        97.5 %
## (Intercept)  1.8370947378  2.0707544478
## radius       0.1042868916  0.1064454975
## age         -0.0107958976 -0.0104421231
## prec         0.0003312927  0.0004108256
```

```
## co2         -0.0008711858 -0.0001568904
```

```
# or to form the CI "by hand"
# number of observations
n <- nrow(tree_growth)
# lower bound
summary(reg_model)$coef[4, 1] - summary(reg_model)$coef[4, 2] * qt(.975, df = n - 5)
```

```
## [1] 0.0003312927
```

```
# upper bound
summary(reg_model)$coef[4, 1] + summary(reg_model)$coef[4, 2] * qt(.975, df = n - 5)
```

```
## [1] 0.0004108256
```

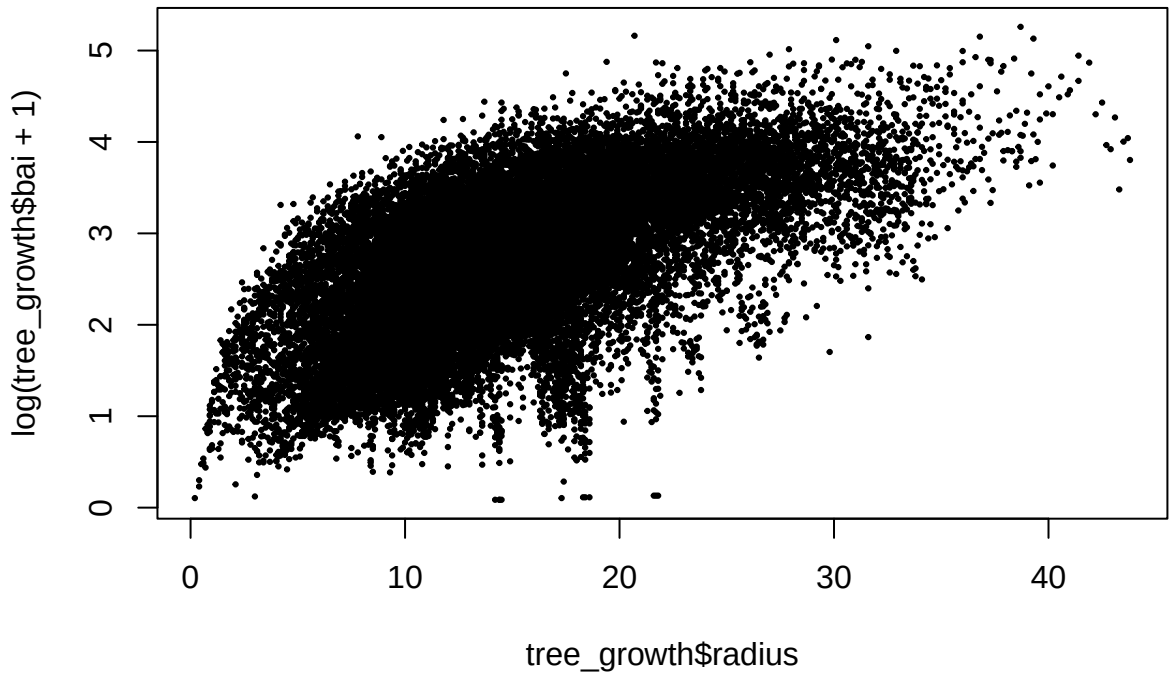
(1 point) Students should either use the coefci command, or if forming by hand should calculate the df correctly

Question 3 (5 pts)

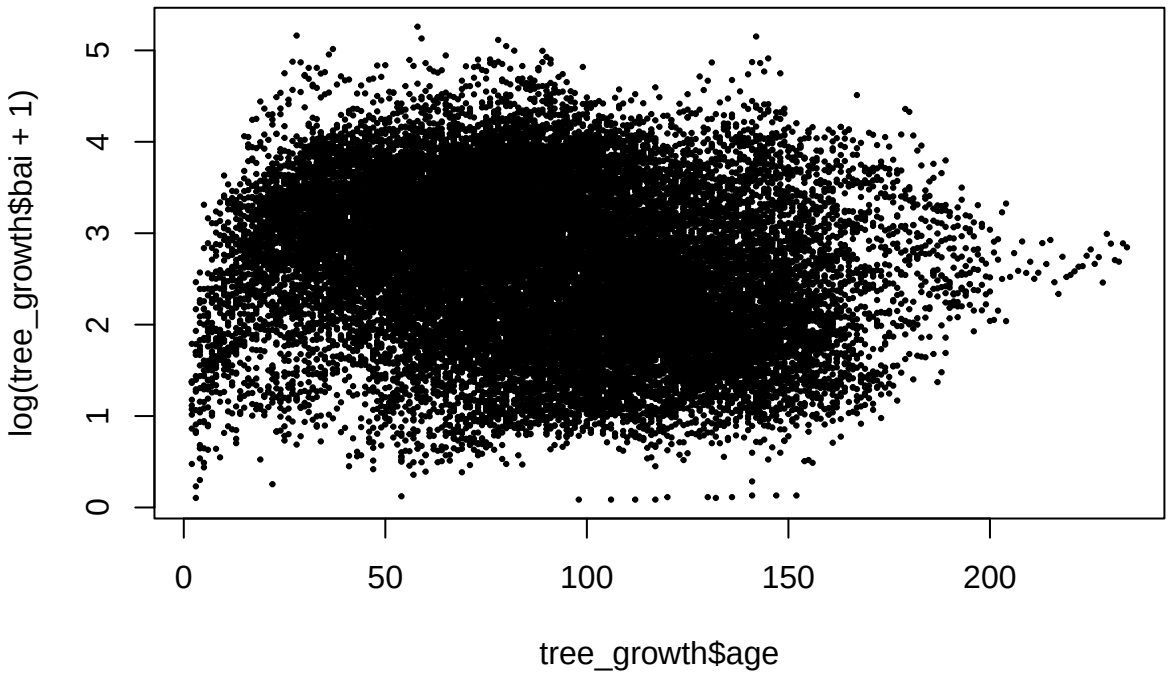
Are the modeling assumptions we typically make when creating confidence intervals and hypothesis tests satisfied? Consider at least two of the required assumptions. Explain why or why not. If helpful, you may provide plots to support your claim.

Answer for Question 3 **Linear model** Examining the univariate plots of the dependent variable against the covariates and the plots of the residuals against the covariates, the dependent variable is probably not exactly linear in the covariates, but it doesn't look like a terrible assumption either. We can also examine the fitted values plotted against the actual values. Again, it doesn't look like it's exactly linear, but it's close. In summary, linearity is a on the border and a reasonable case can be made in either direction, but I would lean towards being okay.

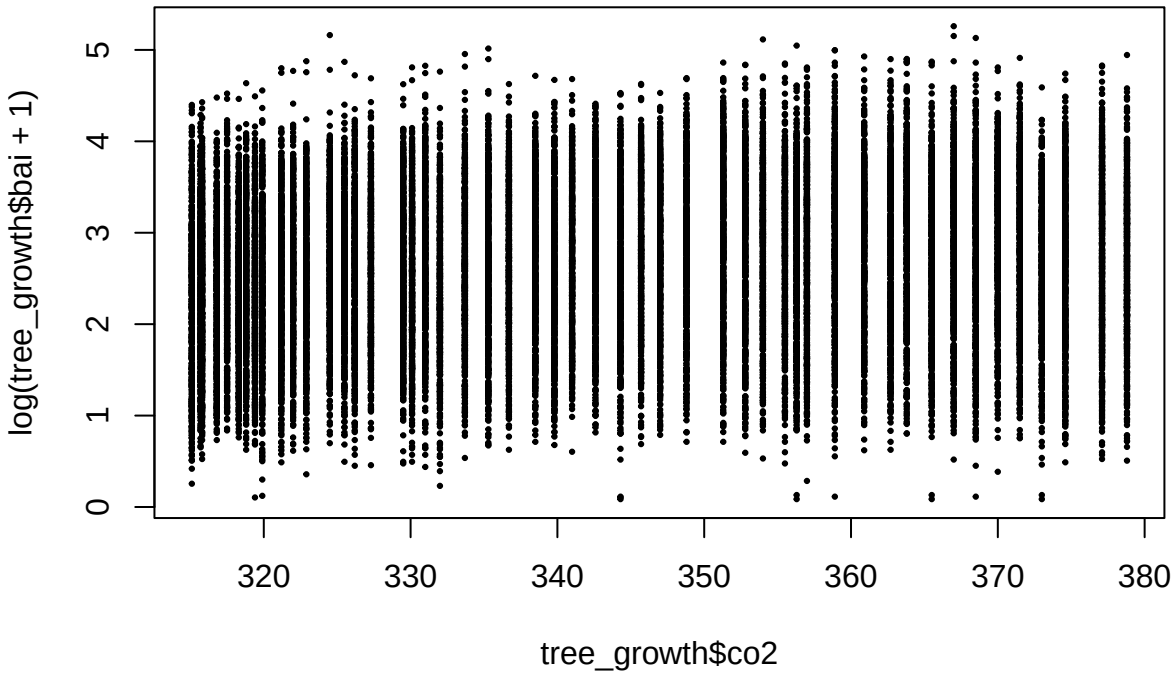
```
plot(tree_growth$radius, log(tree_growth$bai + 1), pch = 19, cex = .3)
```



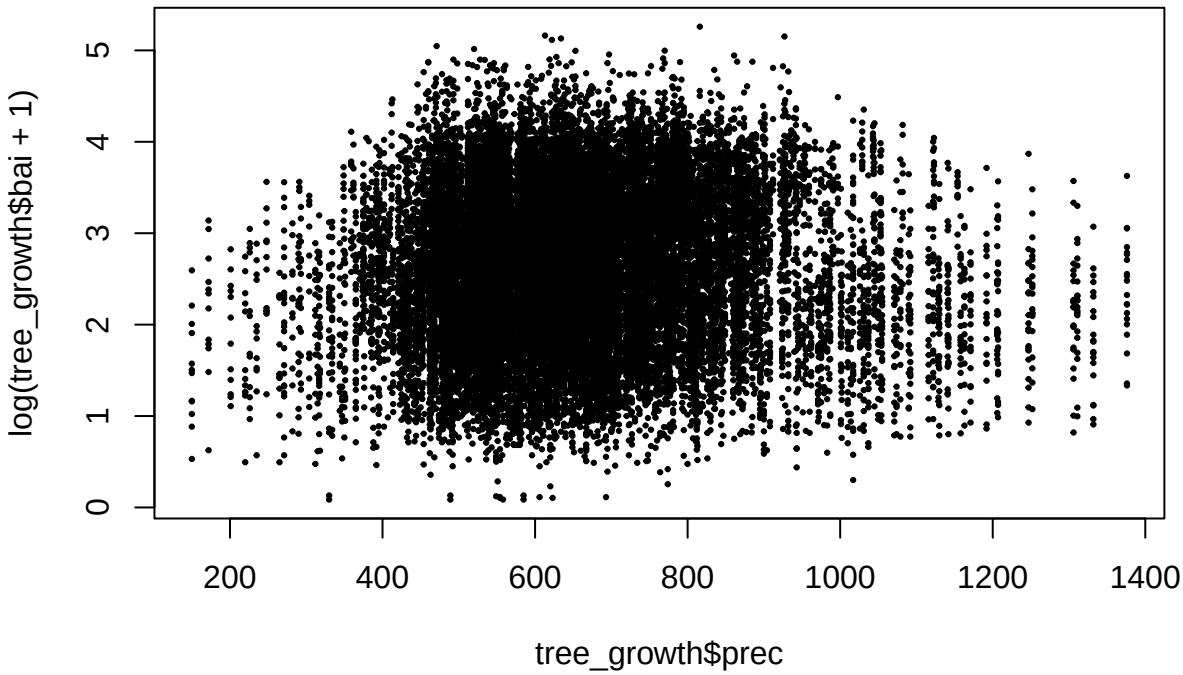

```
plot(tree_growth$age, log(tree_growth$bai + 1), pch = 19, cex = .3)
```



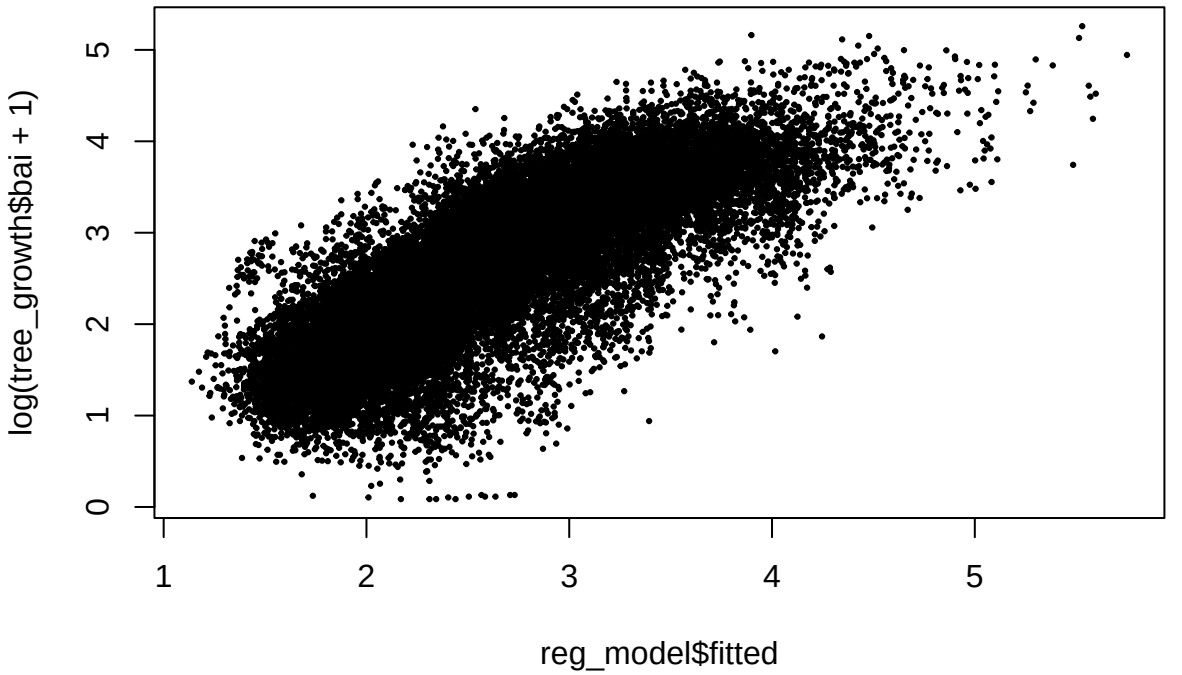

```
plot(tree_growth$co2, log(tree_growth$bai + 1), pch = 19, cex = .3)
```




```
plot(tree_growth$prec, log(tree_growth$bai + 1), pch = 19, cex = .3)
```

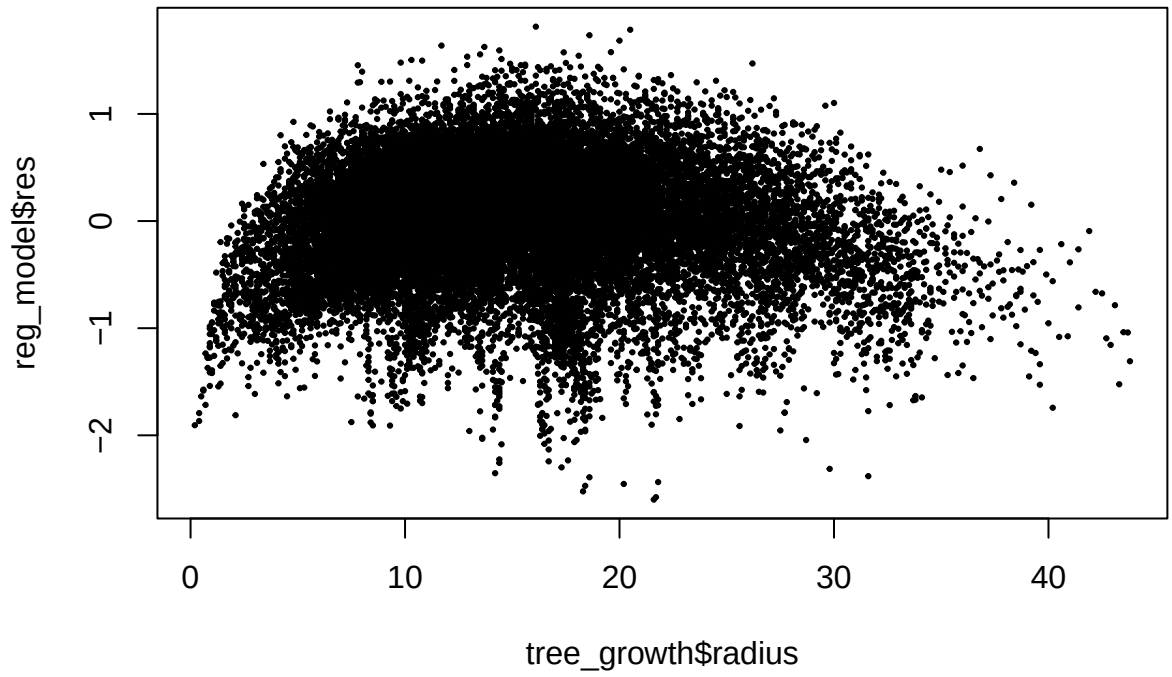



```
plot(reg_model$fitted, log(tree_growth$bai + 1), pch = 19, cex = .3)
```

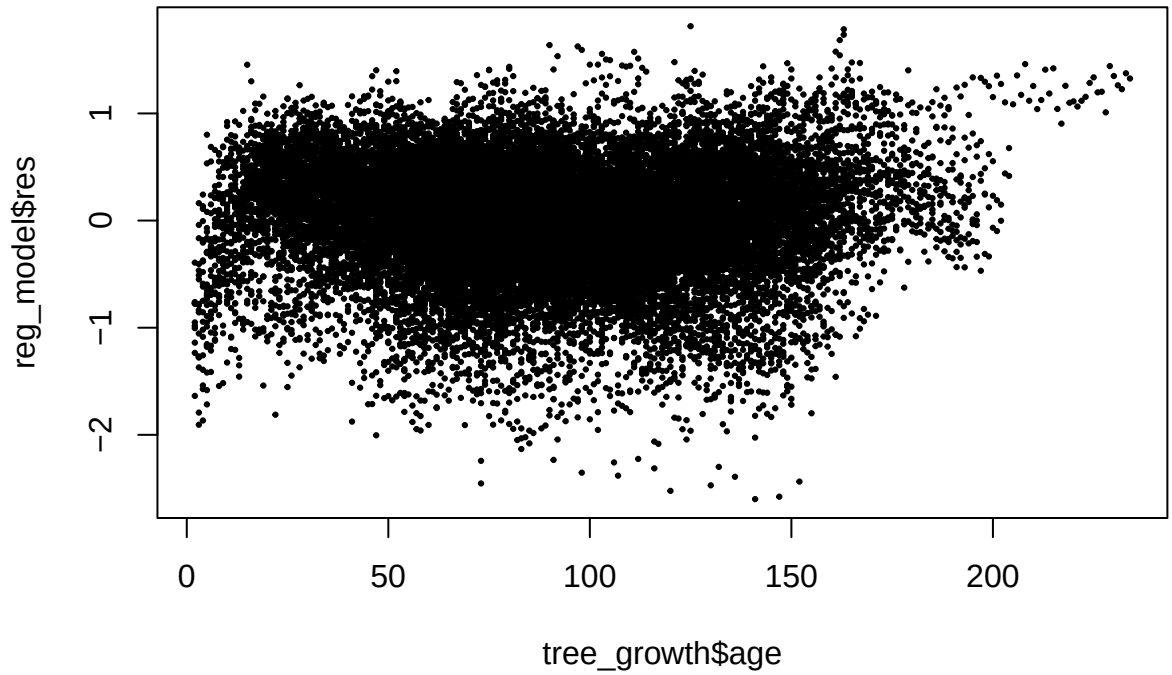


Homoscedastic Examining the plots of the residuals against the covariates, it seems that the variance of the residuals is smaller at the higher values of radius, precipitation and age. To check, we can also use a Breusch-Pagan test which indicates strong evidence against the homoscedastic assumption.

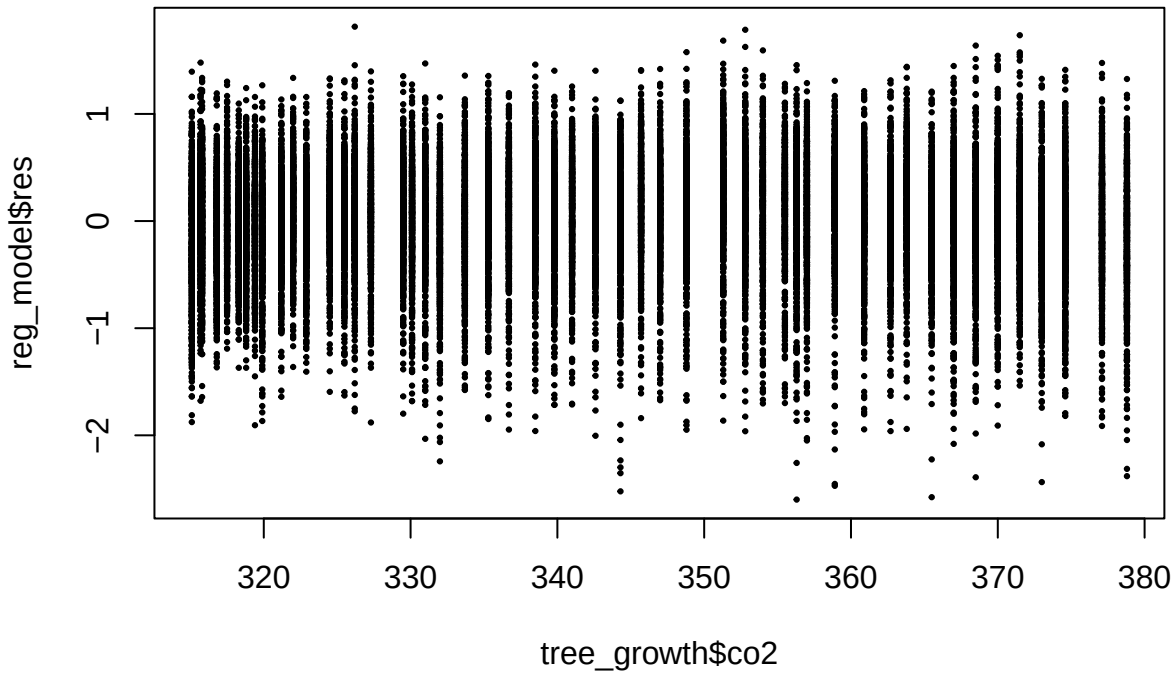
```
plot(tree_growth$radius, reg_model$res, pch = 19, cex = .3)
```



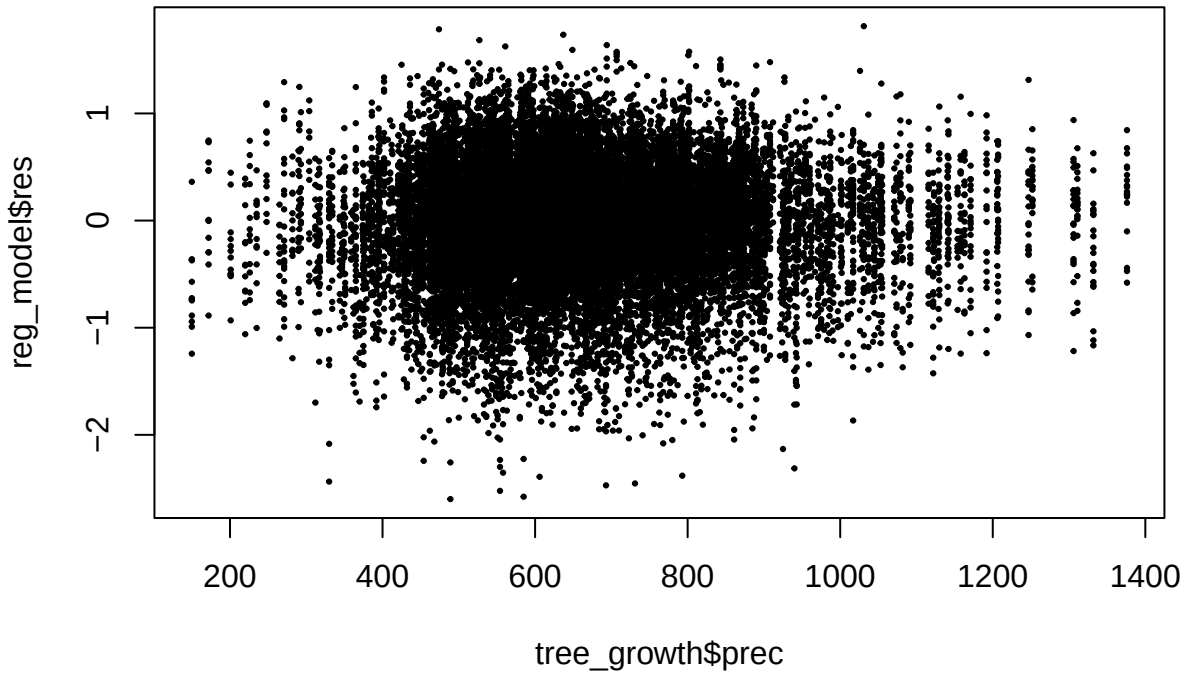

```
plot(tree_growth$age, reg_model$res, pch = 19, cex = .3)
```




```
plot(tree_growth$co2, reg_model$res, pch = 19, cex = .3)
```




```
plot(tree_growth$prec, reg_model$res, pch = 19, cex = .3)
```




```
bptest(reg_model)
```

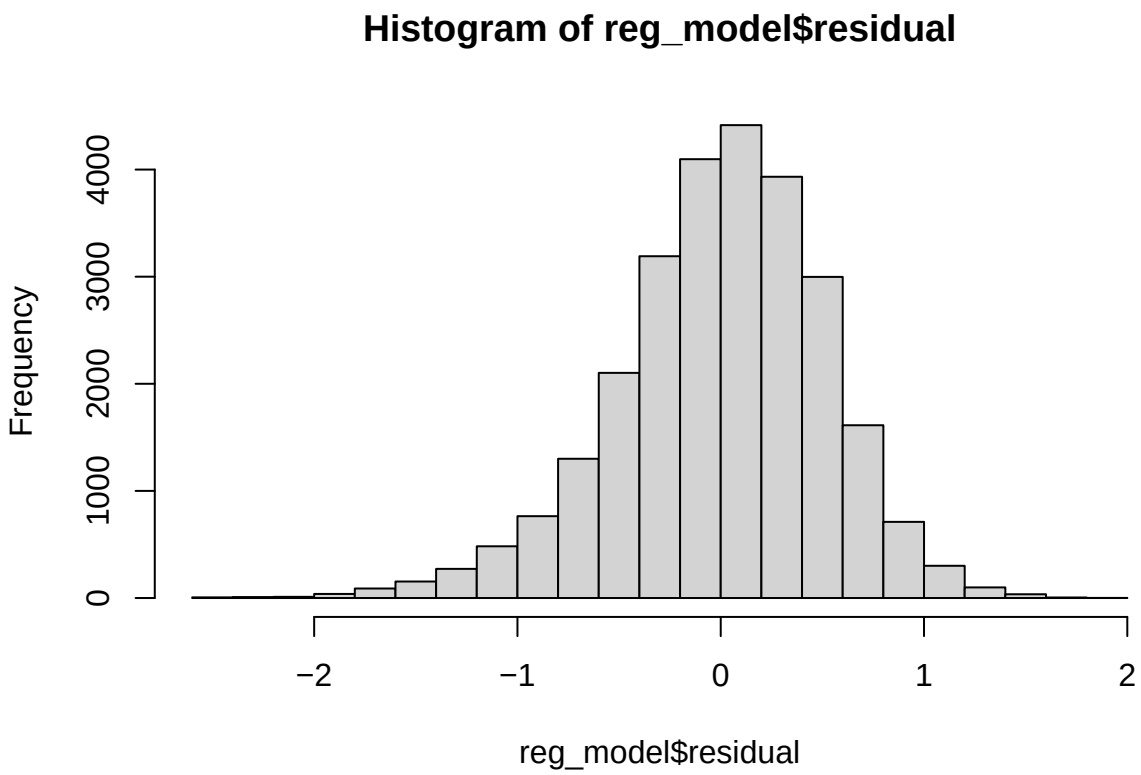
```
##
##  studentized Breusch-Pagan test
##
## data:  reg_model
## BP = 183.1, df = 4, p-value < 2.2e-16
```

Independent error terms Because we include multiple measurements of each tree, it is very likely that the observations are not independent. Measurements from the same tree or from trees at the same site are

likely dependent.

Gaussian errors We can see from the plot below that the residuals look close to normal, but it potentially could be skewed on the left hand side.

```
hist(reg_model$residual)
```



(1 point) Students should correctly identify at least two assumptions we've discussed required for linear regression (2 points) For each assumption discussed, students should give reasonable support for why they think the assumption is violated or not

Question 4 (1 pt)

Using the model above, form a 95% confidence interval for the coefficient corresponding to the coefficient corresponding to `prec` using the robust standard errors (i.e., sandwich standard errors).

```
### Calculate the lower and upper part of the CI using robust standard errors
coefci(reg_model, vcov. = sandwich::vcovHC(reg_model, type = "HC3"))
```

```
##                    2.5 %        97.5 %
## (Intercept)  1.8388755264  2.0689736592
## radius       0.1042636206  0.1064687685
## age         -0.0108017388 -0.0104362818
## prec         0.0003308964  0.0004112219
## co2         -0.0008668466 -0.0001612296
```

```
# or to form the CI "by hand"
# number of observations
n <- nrow(tree_growth)

# robust covariance
robustVar <- sandwich::vcovHC(reg_model, type = "HC3")
```

```r
# square root of value on diagonal for prec
robustSe <- sqrt(robustVar[4,4])

# lower bound
summary(reg_model)$coef[4, 1] - robustSe * qt(.975, df = n - 5)
```

```
## [1] 0.0003308964
```

```r
# upper bound
summary(reg_model)$coef[4, 1] + robustSe * qt(.975, df = n - 5)
```

```
## [1] 0.0004112219
```

(1) Students should either use the coefci command with the sandwich standard error, or form the CI by hand

Question 5 (1 pt)

Do you think it's better to use the robust standard errors or the model based standard errors for this setting? Explain.

Answer for Question 5 The homoscedastic assumption seems to be violated, so we ought to use the robust standard errors. In addition, the sample size is gigantic so concerns about loss of power to detect when the null hypothesis is false aren't a large concern. In general, it is typically safer to use the sandwich standard errors unless we have a very strong reason to believe that the errors are actually homoscedastic.

(1 point) Students should indicate that robust standard errors are appropriate because the homoscedastic assumption is violated

Question 6 (1 pt)

Using the pairs bootstrap with the percentile method, create a 95% confidence interval for the estimated coefficient corresponding to `prec`.

```r
### Calculate the lower and upper part of the CI using the bootstrap here
library("lmboot")
paired.output <- paired.boot(log(bai + 1) ~ age + radius + prec + co2, data = tree_growth, B = 5000)
```

```
## Warning in paired.boot(log(bai + 1) ~ age + radius + prec + co2, data = tree_growth, : Number of boo
```

```r
paired.pct <- apply(paired.output$bootEstParam,
                    MAR = 2, FUN = quantile, prob = c(.025, .975))

paired.pct
```

```
##       (Intercept)         age    radius         prec           co2
## 2.5%     1.837312 -0.01080083 0.1042488 0.0003309968 -0.0008675036
## 97.5%    2.071013 -0.01043684 0.1064571 0.0004104903 -0.0001586142
```

(1 point) The students should use the `paired.boot` command and then the code from lab to calculate the quantiles of the bootstrap estimates.

Question 7 (Bonus: 1 pt)

Is the pairs bootstrap appropriate for this setting? Why or why not?

Answer for Question 7 (1 point) This was a tough question because we only briefly mentioned this in class (which is why it's a bonus). The pairs bootstrap may not be appropriate for this setting, because when we resample pairs, we assume that each observation is independent of the other observations, which as discussed earlier is probably not true. It would be more appropriate to use a cluster bootstrap, where instead of resampling each individual observation, we might resample trees.

If students correctly answer this question, they may make up for a point lost elsewhere, but the total score cannot exceed the total maximum score.

Question 8 (2 pts)

The authors use a random effects model to model the data and include a random effect for the specific tree and the site. Explain why the authors might be motivated to do that.

Answer for Question 8 The data set includes multiple measurements of each tree, and the set of observations for each tree are likely dependent. In addition, since trees in the same site are located close to one another, there may be unobserved environment variables which influence BAI measurements for all trees in the same site. Including a random effect for tree and site would model that dependence.

(1 point) students should mention that the observations are dependent because they involve the same tree (1 point) Students should mention that random effects allow for modeling that dependence

Question 9 (1 pt)

Fit a mixed effects model which uses `log(bai +1)` as the dependent variable and includes fixed effects for age, radius, prec and co2. Include a random effect for the tree (`treeid`) and site (`siteid`).

```
reg_model_rand <- lme4::lmer(log(bai + 1) ~ age + radius + prec + co2 + (1| treeid) + (1|siteid), data =
summary(reg_model_rand)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log(bai + 1) ~ age + radius + prec + co2 + (1 | treeid) + (1 |
##     siteid)
##    Data: tree_growth
##
## REML criterion at convergence: 27652.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.4941 -0.5577  0.0566  0.6164  3.9349
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  treeid   (Intercept) 0.08948  0.2991
##  siteid   (Intercept) 0.07438  0.2727
##  Residual             0.15433  0.3928
## Number of obs: 26627, groups:  treeid, 503; siteid, 30
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept)  2.048e+00  9.366e-02   21.86
## age         -5.643e-03  4.025e-04  -14.02
## radius       8.063e-02  1.075e-03   75.00
## prec         2.772e-04  2.217e-05   12.50
## co2         -8.036e-04  3.079e-04   -2.61
##
```

```
## Correlation of Fixed Effects:
##        (Intr) age    radius prec
## age     0.573
## radius  0.141 -0.339
## prec   -0.133 -0.018  0.006
## co2    -0.796 -0.790 -0.166 -0.016
```

(1 point) Students should use the `lmer` function disucssed in lab and correctly specify the random effects for treeid and siteid.

Question 10 (1 pt)

Fit the same model as above, but this time use a fixed effect for each tree and site.

```
reg_model_fixed <- lm(log(bai + 1) ~ age + radius + prec + co2 + treeid + siteid, data = tree_growth)
summary(reg_model_fixed)$coef[1:5, ]
```

```
##                  Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept)  2.4181372387 1.123001e-01 21.532819 5.914607e-102
## age         -0.0043727918 5.143262e-04 -8.501982  1.961895e-17
## radius       0.0785056242 1.125657e-03 69.742047  0.000000e+00
## prec         0.0002756903 2.219742e-05 12.419926  2.563840e-35
## co2         -0.0014374148 3.765418e-04 -3.817411  1.351705e-04
```

(1 point) Students should use the `lm` function and include treeid and siteid as categorical variables

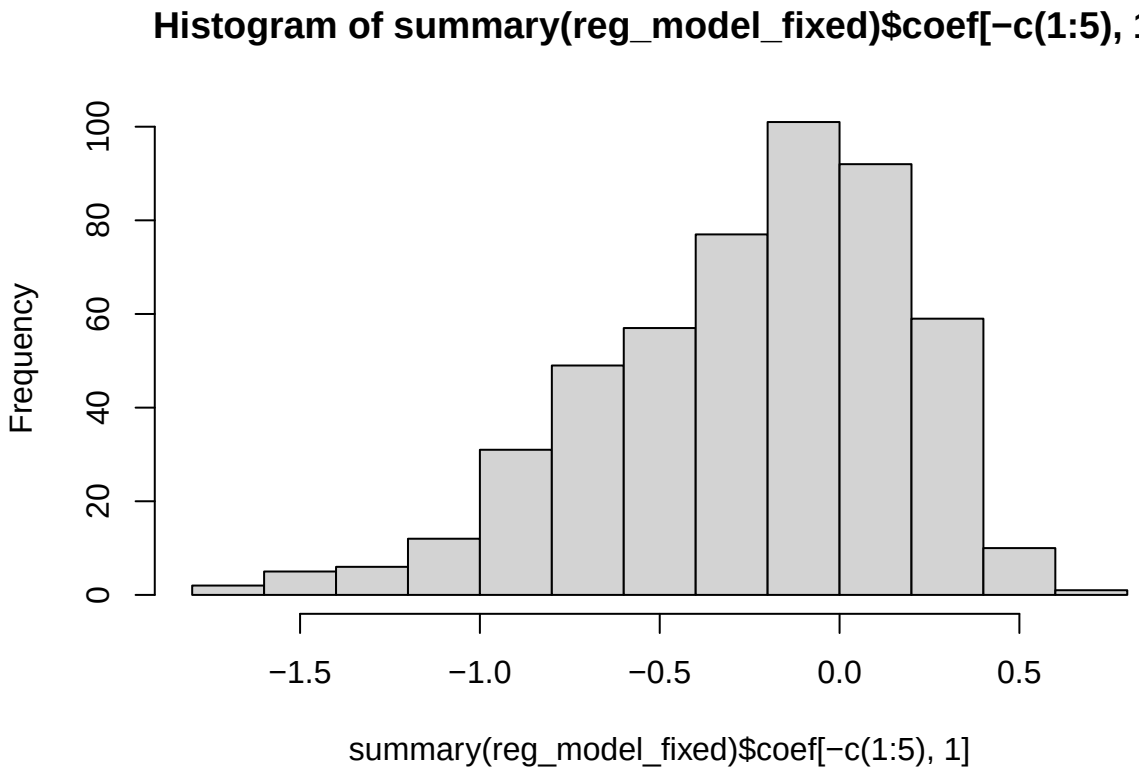
Question 11 (2 pts)

The two procedures give similar estimates, and neither model is clearly better than the other. But, explain some reasons why you might pick one over the other.

Answer for Question 11 In this case, we might have a lots of measurements for each tree, the gain in precision from using the mixed effects model might not be very large when estimating the coefficients we care about. In addition, when examining the estimated fixed effects, the normality assumption may not be true. However, its also likely the case that the measured covariates are not independent of random effects.

On the other hand, if you wanted to think about how large the individual level effect of a new tree might be, using a random effects model would allow you to do that.

```
hist(summary(reg_model_fixed)$coef[-c(1:5), 1])
```



(1 point) students should discuss the potential benefit of the procedure picked and why it might outweight the costs (1 point) students should discuss the potential costs of the procedure picked and why it might be outweighed by the benefit