

R Console Output: Regression Analyses

The following is a copy of the statistical report for simple linear and multiple regression analyses of the Spring 2009 data. It includes a line-by-line description of the report output provided in the R console. This will be a useful reference for application of regression analyses to new data as the semester progresses. All of the highlighted values MUST be included in any report that you submit in which you utilize regression analysis to examine relationships between predictor and response variables.

These output lines correspond to the code that R uses to import data into the workspace.

```
> #Ponderosa Pine Lab R Script
>
> #Initialize variables
> #These following steps allow us to create objects in R that correspond to our data variables.
Recall that these are case sensitive and must be entered EXACTLY as we define them during
analysis. Data MUST be in a comma separated file format (.csv)
>
> pine <-read.csv(file.choose(), header=TRUE)
>
> attach(pine) # Here we tell R to attach our headers from the .csv file.
```

The following object(s) are masked from pine (position 3) :

```
canyon dbh density elevation height n_dbh n_height rate rings semester
```

```
>
```

These are the data as read by R software. You should ALWAYS check your data to ensure that they are inputted correctly. I haven't included all the data lines in this report, which is the reason it jumps from data line 8 to line 452 (the last line in the data set).

```
> pine # look at all the data to make sure it is loaded in R
canyon elevation dbh height n_dbh n_height rings rate
```

```

1    BC 2020.00 40.50 10.80  NA   NA 50.00 0.100
2    BC 2020.00 40.00 10.10  NA   NA 51.00 0.098
3    SC 2001.00 48.00  7.30  NA   NA 26.00 0.192
4    SL 2407.00 16.00 13.00  3.00 0.30 37.00 0.135
5    SL 2407.00 26.00  6.60  3.50 0.35 43.00 0.116
6    CC 1952.24 45.50  7.80  4.50 0.51 19.00 0.263
7    SC 1921.76 94.50 13.25 15.70 0.60 28.00 0.179
8    BC 2321.00  0.33 17.50  0.01 0.61 12.00 0.417

```

...

```
452  LH 2358.00 130.00 17.30  NA   NA 32.00 0.156
```

```
density semester
```

```

1    0.00 Spring 2008
2    0.00 Spring 2008
3    4.00  Fall 2008
4    0.00  Fall 2008
5   11.00  Fall 2008
6    4.00 spring 2007
7    6.00  Fall 2006
8    1.00 Spring 2008

```

...

```
452  4.00  Spring 09
```

>

> # Call up a few variables to see that you can see data for each. Use the same name as those used in the original excel file or you can scroll up in the R Console to see the heading for each column. Here's one to get you started:

>

```
> elevation
```

```
[1] 2020.00 2020.00 2001.00 2407.00 2407.00 1952.24 1921.76 2321.00
```

...

```
[449] 2363.00 1897.00 1893.00 2358.00
```

>

You should always report the mean, variance, and standard deviation of the variables included in a regression analysis.

> #Descriptive Statistics

> #Substitute x with the variable of interest

> `mean(elevation, na.rm= TRUE)` # Here the "na.rm=TRUE" tells R to remove the records that are missing a value. For example, some groups may not have had a non-ponderosa tree growing within sight of the target ponderosa tree and thus did not record a value. Because there was a blank cell in the excel file that we turned into a .csv file and loaded into R, R has placed "NA" in this place. Here we must tell R to ignore or remove these "NAs".

[1] 2154.005

> `var(elevation, na.rm= TRUE)` #The variance of a sample is a non-negative number which gives an idea of how widely spread the values of sample are likely to be; the larger the variance, the more scattered the observations around the mean.

[1] 91885.39

> `sd(elevation, na.rm= TRUE)` #68% of the data lies within one standard deviation on either side of the mean, if our data is normally distributed. It's another measure of data dispersion. The steps are: (1) Compute the mean for the data set. (2)Compute the deviation by subtracting the mean from each value. (3)Square each individual deviation.(4) Add up the squared deviations.(5)Divide by one less than the sample size.(6)Take the square root.

[1] 303.126

> `se <- function(x) {sd(x,na.rm=TRUE)/sqrt(length(x))}` #leave x as is in this line of code. Note that within the brackets you have the equation for standard error (standard deviation divided by the square root of the sample size).

> # This is a function that you just made!! You now can find the standard error of the mean for any variable in the dataset with just a few characters of code.

> `se(elevation)` # substitute the name of the variable of interest.

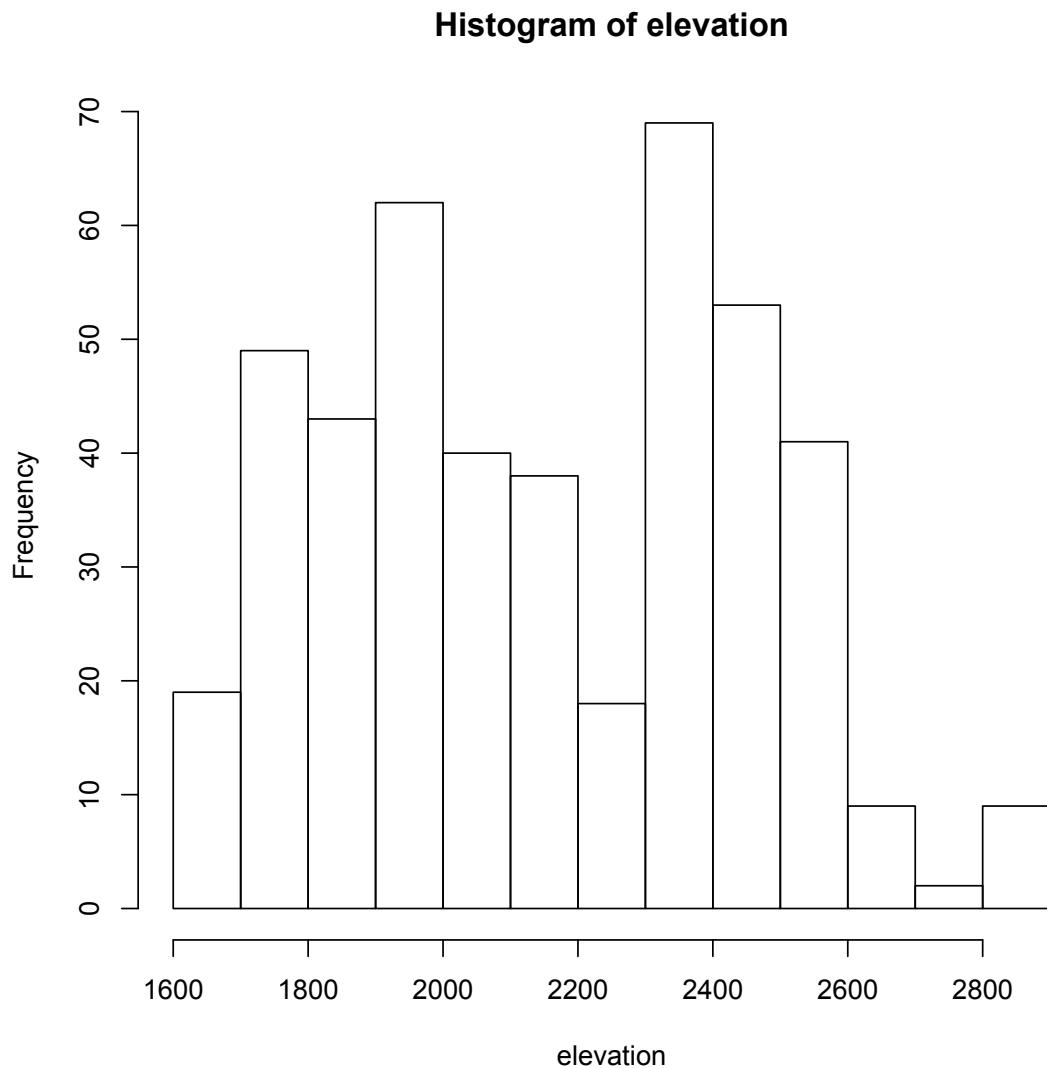
[1] 14.25785

You should visualize the data before proceeding with any data analysis. Histograms allow you to examine (but not statistically assess) if the data are normally distributed and to identify any outliers. Clearly, the data on elevation do not appear to be normally distributed, a fact that we will ignore in subsequent regression analyses. Again, methods are available that are better suited to analyzing data that aren't normally distributed, but we won't go into them in this class.

It is also essential that you examine data for outliers before proceeding with any analysis. Outliers are observations that are numerically distant from other data points. The statistical methods we use in this course are extremely sensitive to the effects of outliers

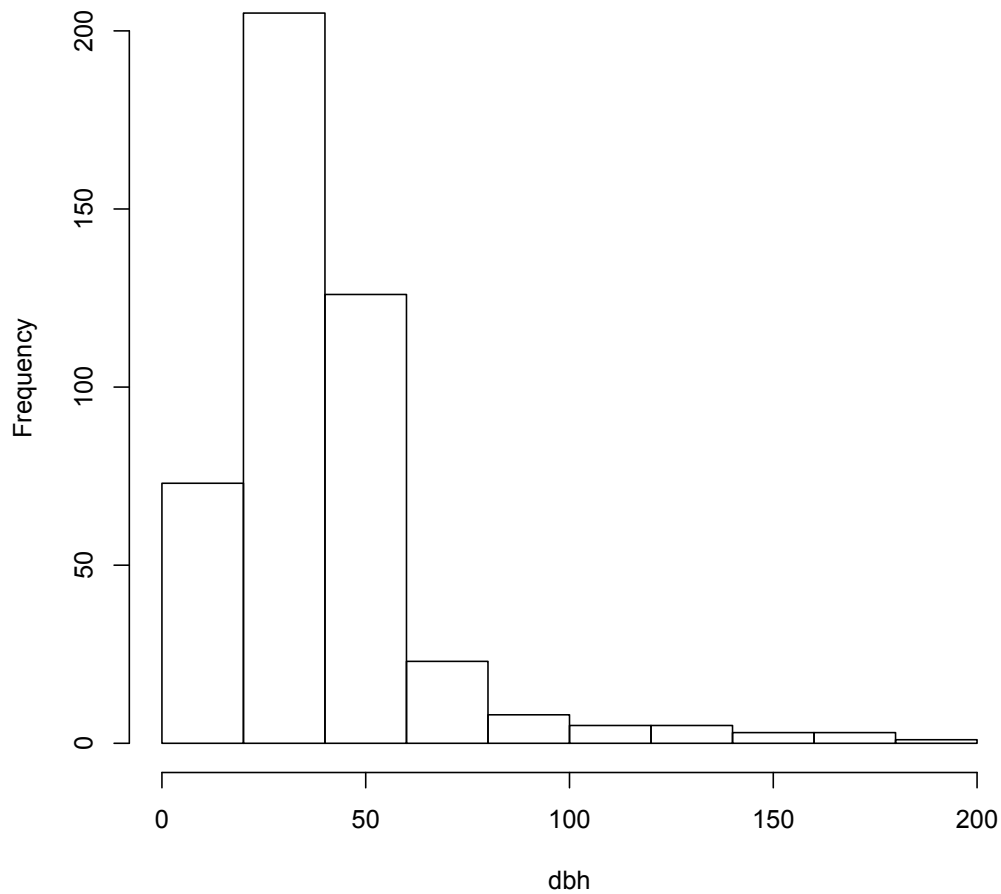
and consequently we must consider whether to exclude the outliers from the analysis or consider an alternative method to analyze data that include outliers. You should only exclude outliers from a data set if you have reason to suspect the veracity of the data collected. For example, in the Ponderosa pine data set, some of the data points for density of surrounding Ponderosa pine were unusually high. This could reflect that an area has an unusually high Ponderosa pine density, in which case, although the data point is an outlier, it is valid and cannot be excluded. It could, however, be the consequence of sampling too large an area, resulting in an overly large number of trees and an inaccurate estimate of density. Generally speaking, sampling error of this nature will be obvious because all the values for a given individual/group will be outliers. This is also one of the many reasons to take good field notes that allow you to reconstruct conditions in the field. It is generally acceptable to exclude outlying data points you suspect are due to sampling error, provided you can document the source of the error.

> hist(elevation) # visualize the spread in the data for variables of interest.



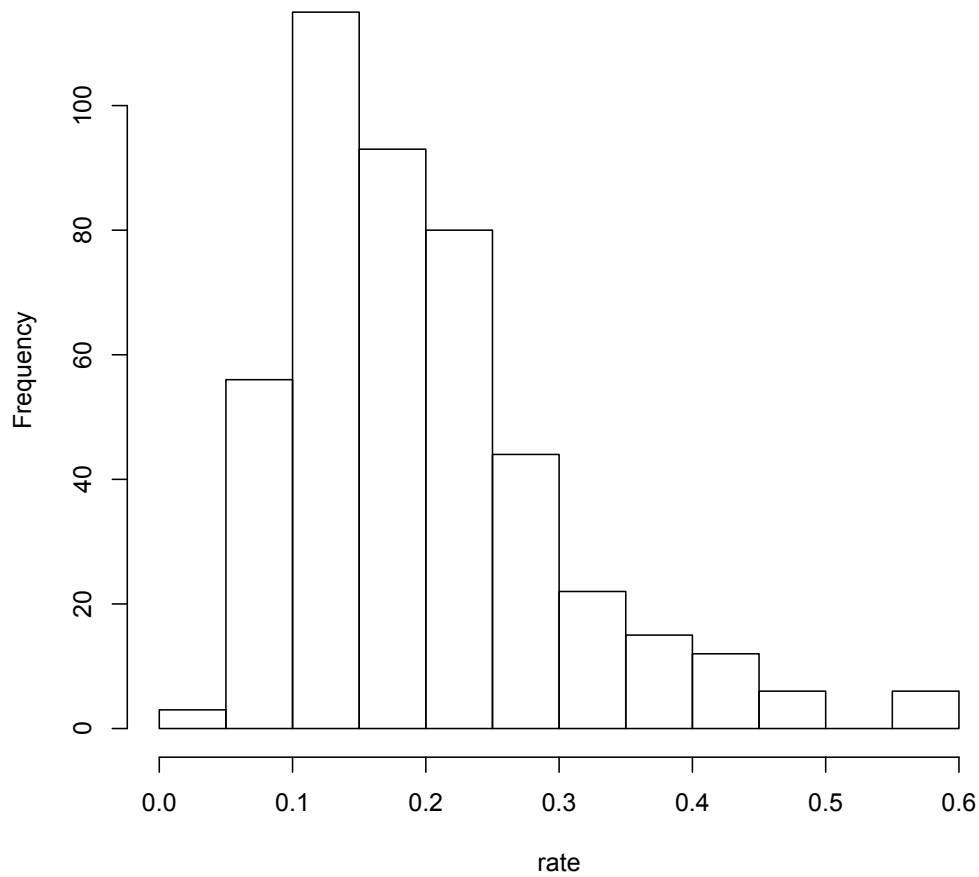
```
>
> mean(dbh, na.rm= TRUE)
[1] 39.79493
> var(dbh, na.rm= TRUE)
[1] 667.9714
> sd(dbh, na.rm= TRUE)
[1] 25.84514
> se <- function(x) {sd(x,na.rm=TRUE)/sqrt(length(x))} >
> se(dbh)
[1] 1.215653
> hist(dbh)
```

Histogram of dbh



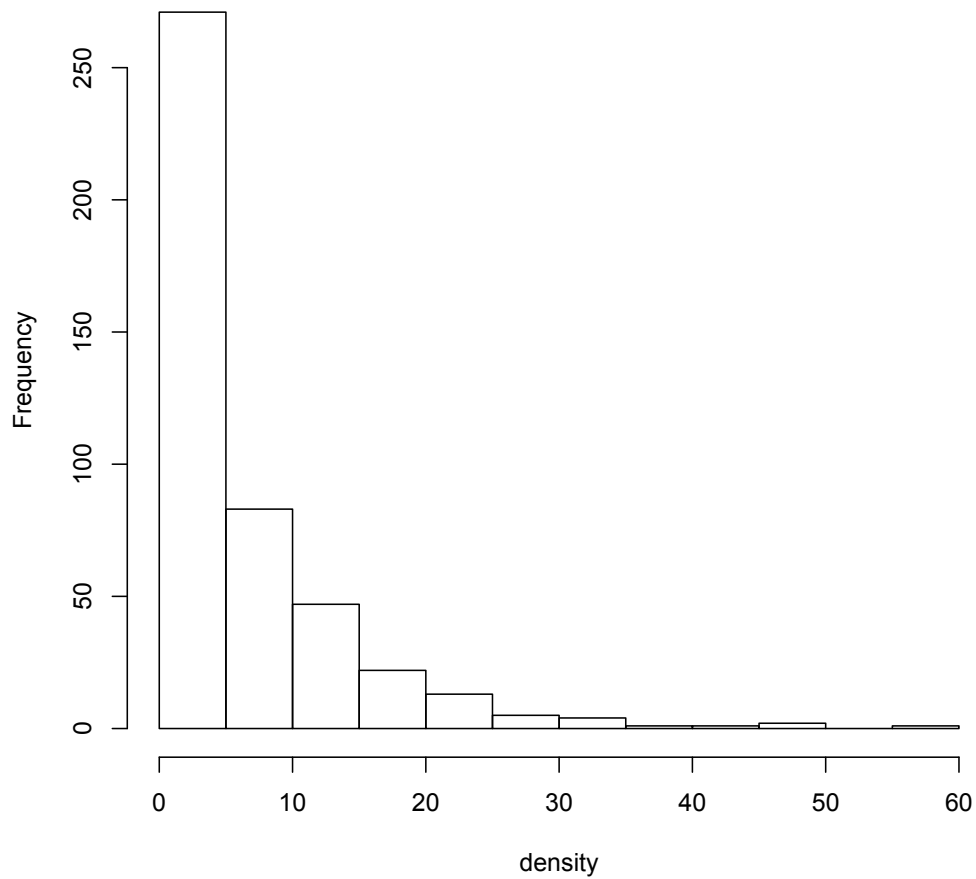
```
>
> mean(rate, na.rm= TRUE)
[1] 0.1995288
> var(rate, na.rm= TRUE)
[1] 0.009941957
> sd(rate, na.rm= TRUE)
[1] 0.09970936
> se <- function(x) {sd(x,na.rm=TRUE)/sqrt(length(x))}
> se(rate)
[1] 0.004689934
> hist(rate)
```

Histogram of rate



```
>
> mean(density, na.rm= TRUE)
[1] 6.941467
> var(density, na.rm= TRUE)
[1] 61.05647
> sd(density, na.rm= TRUE)
[1] 7.813864
> se <- function(x) {sd(x,na.rm=TRUE)/sqrt(length(x))}
> se(density)
[1] 0.3675333
> hist(density)
```

Histogram of density

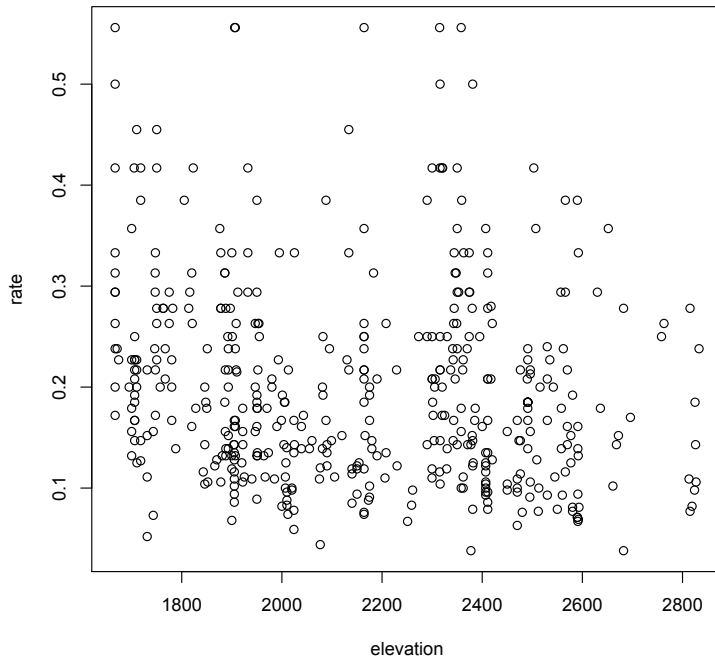


>

In addition to constructing histograms of the individual variables, you should always construct scatter plots for any data you subject to regression analyses. This may help you identify previously overlooked outliers.

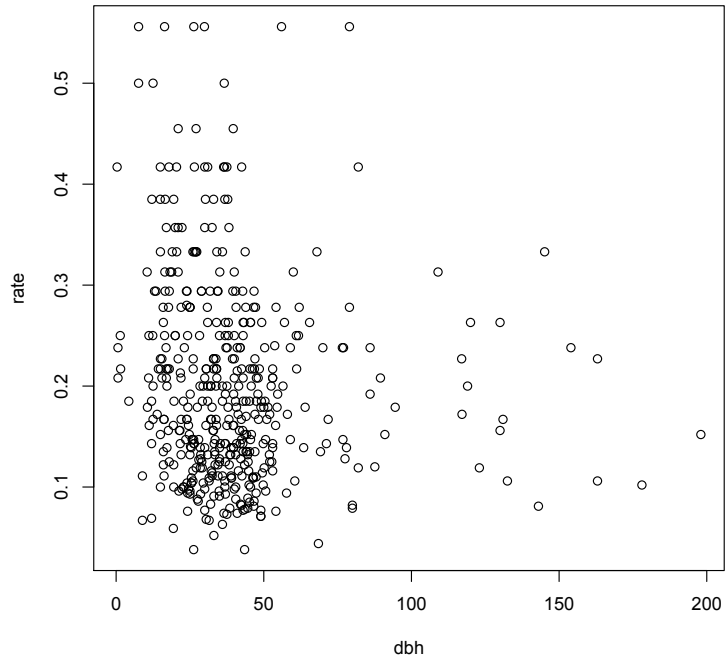
```
> #Scatter Plot
```

```
> plot(elevation,rate) #This will show a plot of x vs y (predictor vs. response variable)
```



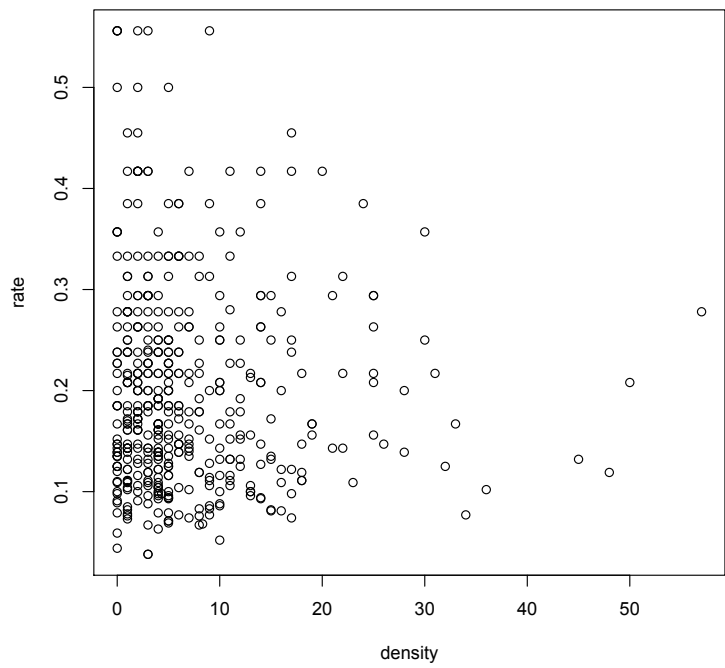
```
>
```

```
> plot(dbh,rate)
```



v

> plot(density,rate)



v

Having examined our data using descriptive statistics and multiple graphing methods, we can now proceed to the regression analyses. You should always start with more simple models (simple linear regression) before proceeding to more complex models (multiple regression) in any data analysis.

```
> #Simple Linear Regression: growth rate against elevation
```

```
> model1 <- lm (rate~elevation) #In this line of code we create the model
```

```
> model1 #This runs the model
```

Call:

```
lm(formula = rate ~ elevation)
```

Coefficients:

```
(Intercept)  elevation
```

```
3.147e-01 -5.348e-05
```

```
> summary (model1) #This creates our data output
```

Call:

```
lm(formula = rate ~ elevation)
```

Residuals:

```
Min      1Q  Median      3Q      Max
```

```
-0.17015 -0.07459 -0.02038  0.05101  0.36738
```

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 3.147e-01 3.328e-02 9.457 < 2e-16 ***
```

```
elevation -5.348e-05 1.530e-05 -3.495 0.00052 ***
```

```
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09849 on 450 degrees of freedom

Multiple R-squared: 0.02643, Adjusted R-squared: 0.02427

F-statistic: 12.22 on 1 and 450 DF, p-value: 0.0005204

In your report, you should interpret the results of the simple linear regression analysis as follows:

If $p > 0.05$ – We fail to reject the null hypothesis that there is no linear relationship between growth rate and elevation, $F_{(\text{num}, \text{den})} = \text{value reported}$, $p = \text{value reported}$.

If $p \leq 0.05$ – We reject the null hypothesis that there is no linear relationship between growth rate and elevation, $F_{(1,450)} = 12.22$, $p = 0.0005204$. We find evidence for a statistically significant linear relationship between growth rate and elevation: $\text{rate} = 3.147e-01 + (-5.348e-05)\text{elevation}$, $R^2 = 0.02643$.

***Report the multiple R-squared value for a simple linear regression.

>

```
> plot(elevation,rate,xlab="Elevation (m)", ylab="Growth rate (cm/yr)") # make sure you change the x and y labels to something that makes sense.
```

```
> abline(reg=model1, col ="blue") #try any other color you prefer.
```

```
> title( sub= "Figure 1: Plot of Ponderosa pine growth rate against elevation in the Boulder foothills.")
```

>

```
> #Simple Linear Regression: growth rate against DBH
```

```
> model2 <- lm (rate~dbh)
```

```
> model2
```

Call:

```
lm(formula = rate ~ dbh)
```

Coefficients:

```
(Intercept)      dbh
```

```
0.2175240 -0.0004522
```

```
> summary (model2)
```

Call:

```
lm(formula = rate ~ dbh)
```

Residuals:

```
    Min     1Q  Median     3Q     Max
-0.16765 -0.07159 -0.01890  0.05228  0.37420
```

Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.2175240  0.0085674  25.390 <2e-16 ***
dbh          -0.0004522  0.0001806  -2.504  0.0126 *
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09913 on 450 degrees of freedom

Multiple R-squared: 0.01374, Adjusted R-squared: 0.01155

F-statistic: 6.269 on 1 and 450 DF, p-value: 0.01264

In your report, you should interpret the results of the simple linear regression analysis as follows:

If $p > 0.05$ – We fail to reject the null hypothesis that there is no linear relationship between growth rate and DBH, $F_{(num, den)} = \text{value reported}$, $p = \text{value reported}$.

If $p \leq 0.05$ – We reject the null hypothesis that there is no linear relationship between growth rate and DBH, $F_{(1,450)} = 6.269$, $p = 0.01264$. We find evidence for a statistically significant linear relationship between growth rate and DBH: $\text{rate} = 2.175e-01 + (-4.522e-04)\text{DBH}$, $R^2 = 0.01374$.

```
>
```

```
> plot(dbh,rate,xlab="DBH (cm)", ylab="Growth rate (cm/yr)")
```

```
> abline(reg=model2, col ="red")
```

```
> title( sub= "Figure 2: Plot of Ponderosa pine growth rate against DBH in the Boulder foothills.")
```

```

>
> #Simple Linear Regression: growth rate against density
> model3 <- lm (rate~density)
> summary (model3)

```

Call:

```
lm(formula = rate ~ density)
```

Residuals:

```

      Min       1Q   Median       3Q      Max
-0.16274 -0.07266 -0.02137  0.04999  0.35746

```

Coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.2018383  0.0063038  32.018 <2e-16 ***
density      -0.0003666  0.0006035  -0.607  0.544

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09992 on 448 degrees of freedom

(2 observations deleted due to missingness)

Multiple R-squared: 0.0008231, Adjusted R-squared: -0.001407

F-statistic: 0.3691 on 1 and 448 DF, p-value: 0.5438

In your report, you should interpret the results of the simple linear regression analysis as follows:

If $p > 0.05$ – We fail to reject the null hypothesis that there is no linear relationship between growth rate and density, $F_{(1, 448)} = 0.3691$, $p = 0.5438$. We fail to find evidence for a statistically significant linear relationship between growth rate and density. Because we find no evidence of a linear relationship, it is inappropriate to provide the regression equation because we do not have statistical evidence to support the best-fit line for the data.

If $p \leq 0.05$ – We reject the null hypothesis that there is no linear relationship between growth rate and density, $F_{(\text{num}, \text{den})} = \text{value reported}$, $p = \text{value reported}$.

>

```
> plot(density, rate, xlab="Ponderosa pine density (# trees/100 m^2)", ylab="Growth rate (cm/yr)")
```

```
> abline(reg=model3, col="green")
```

```
> title(sub="Figure 3: Plot of Ponderosa pine growth rate against Ponderosa pine density in the Boulder foothills.")
```

>

```
> #Multiple Regression
```

```
> model4 <- lm(rate ~ elevation + dbh + density) #Include here predictor variables of interest as a function of a response variable
```

```
> model4
```

Call:

```
lm(formula = rate ~ elevation + dbh + density)
```

Coefficients:

```
(Intercept)  elevation      dbh      density
 3.356e-01  -5.341e-05  -4.414e-04  -5.232e-04
```

```
> summary(model4)
```

Call:

```
lm(formula = rate ~ elevation + dbh + density)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-0.17128 -0.07387 -0.02004  0.04926  0.37245
```

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
```

(Intercept) 3.356e-01 3.458e-02 9.703 < 2e-16 ***

elevation -5.341e-05 1.534e-05 -3.480 0.00055 ***

dbh -4.414e-04 1.790e-04 -2.465 0.01407 *

density -5.232e-04 5.955e-04 -0.879 0.38011

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09812 on 446 degrees of freedom

(2 observations deleted due to missingness)

Multiple R-squared: 0.04081, Adjusted R-squared: 0.03436

F-statistic: 6.325 on 3 and 446 DF, p-value: 0.0003302

In your report, you should interpret the results of the multiple regression analysis as follows:

Full Model (F-value)

If $p > 0.05$ – We fail to reject the null hypothesis that there is no relationship between growth rate and elevation, DBH and density as a group, $F_{(num, den)} = \text{value reported}$, $p = \text{value reported}$.

If $p \leq 0.05$ – We reject the null hypothesis that there is no relationship between growth rate and elevation, DBH and density as a group, $F_{(3,446)} = 6.325$, $p = 0.0003302$. We find evidence for a statistically significant relationship between growth rate and elevation, DBH and density as a group: $\text{rate} = 3.356e-01 + (-5.341e-05)\text{elevation} + (-4.414e-04)\text{DBH} + (-5.232e-04)\text{density}$, $R^2 = 0.03436$.

***Report the adjusted R-squared value for a multiple regression, which has been adjusted to account for multiple predictor variables.

Residual Model–Elevation (t-value)

If $p > 0.05$ – We fail to reject the null hypothesis that there is no relationship between growth rate and elevation, $t_{(num, den)} = \text{value reported}$, $p = \text{value reported}$.

If $p \leq 0.05$ – We reject the null hypothesis that there is no relationship between growth rate and elevation, $t_{(1,448)} = -3.480$, $p = 0.00055$. We find evidence for a statistically significant relationship between growth rate and elevation while controlling for variation due to DBH and Ponderosa pine density.

Residual Model–DBH (t-value)

If $p > 0.05$ – We fail to reject the null hypothesis that there is no relationship between growth rate and DBH, $t_{(\text{num}, \text{den})} = \text{value reported}$, $p = \text{value reported}$.

If $p \leq 0.05$ – We reject the null hypothesis that there is no relationship between growth rate and DBH, $t_{(1,448)} = -2.465$, $p = 0.01407$. We find evidence for a statistically significant relationship between growth rate and DBH while controlling for variation due to elevation and Ponderosa pine density.

Residual Model–Ponderosa pine density (t-value)

If $p > 0.05$ – We fail to reject the null hypothesis that there is no relationship between growth rate and density, $t_{(1,448)} = -0.879$, $p = 0.38011$. We fail to find evidence for a statistically significant linear relationship between growth rate and density while controlling for variation due to elevation and DBH.

If $p \leq 0.05$ – We reject the null hypothesis that there is no relationship between growth rate and density, $F_{(\text{num}, \text{den})} = \text{value reported}$, $p = \text{value reported}$.

Sample Report:

Do growing season length, precipitation, tree age, and/or congeneric competition affect the growth of ponderosa pines? We address several competing hypotheses regarding the cause(s) of variation in ponderosa pine growth rate along an elevational gradient in the Boulder foothills.

Simple Linear Regression: Elevation and ponderosa pine growth rates

Both precipitation and the length of the growing season vary along an elevational gradient in the Boulder foothills and may affect ponderosa pine growth rate. As elevation increases, so does precipitation and increased precipitation at higher elevations may produce a faster growth rate. If the amount of precipitation limits growth of ponderosa pines, then we expect to find a significant positive linear relationship between growth rate and elevation. However, as elevation increases, the growing season also decreases and a shorter growing season at higher elevations may produce a slower growth rate. If the length of the growing season limits growth of ponderosa pines, then we expect to find a significant negative linear relationship between growth rate and elevation. To test these hypotheses, we conducted a simple linear regression of growth rate on elevation. We reject the null hypothesis that there is no linear relationship between growth rate and elevation, $F_{(1,450)} = 12.22$, $p = 0.0005204$. We find evidence for a significant negative linear relationship between growth rate (cm/yr) and elevation (m): $\text{rate} = 3.147e-01 + (-5.348e-05)\text{elevation}$, $R^2 = 0.02643$ (Figure 1). As elevation increases, growth rate decreases, an observation that supports the hypothesis that growing season limits growth rate in ponderosa pine. However, only 2.6% of the variation in growth rate is due to changes in elevation and we conclude that elevation is not a very good predictor of growth rate.

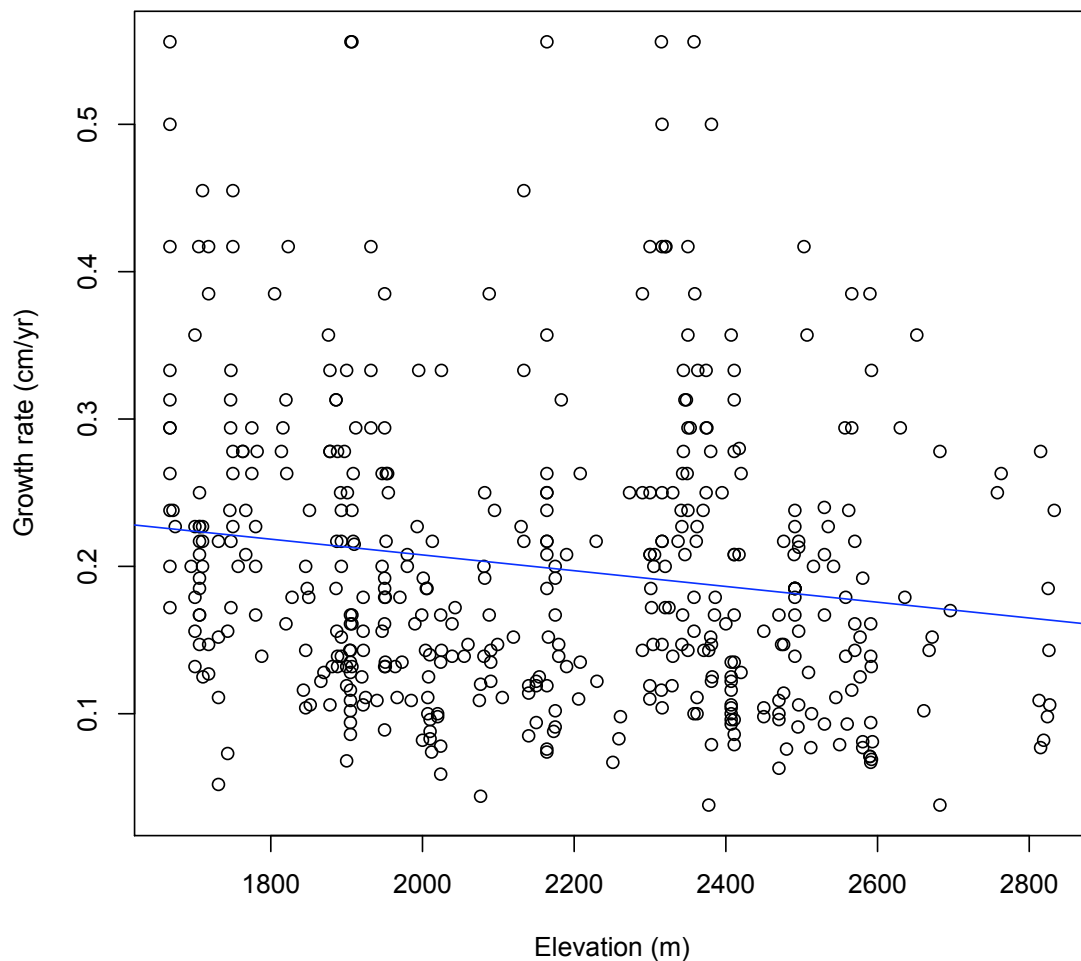


Figure 1: Plot of Ponderosa pine growth rate against elevation in the Boulder foothills

Simple Linear Regression: DBH and ponderosa pine growth rates

Older trees often grow more slowly than younger trees. In addition, as trees grow, they add wood to their girth. Consequently, diameter at breast height (DBH) can be used as a proxy for age in most tree species, i.e., older trees have a larger diameter than younger trees. If tree age limits ponderosa pine growth rates, then we expect to find a significant negative linear relationship between growth rate and DBH. To test this hypothesis, we conducted a simple linear regression of growth rate on DBH. We reject the null hypothesis that there is no linear relationship between growth rate and DBH, $F_{(1,450)} = 6.269$, $p = 0.01264$. We find evidence for a significant negative linear relationship between growth rate (cm/yr) and DBH (cm): $\text{rate} = 2.175e-01 + (-4.522e-04)\text{DBH}$, $R^2 = 0.01374$ (Figure 2). As DBH increases, growth rate decreases, an observation that supports the hypothesis that age limits growth rate in ponderosa pine. However, only 1.4% of the variation in growth rate is due to differences in DBH and we conclude that DBH is not a very good predictor of growth rate.

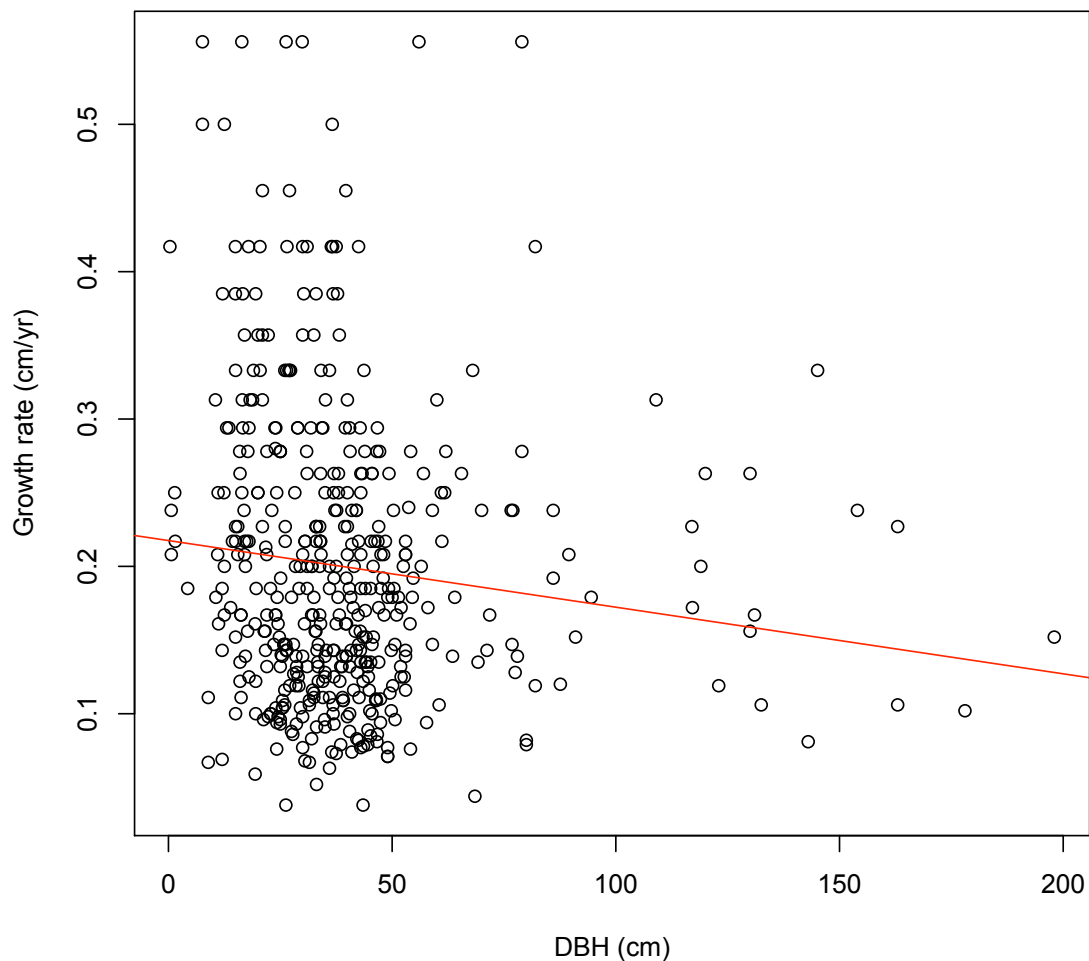


Figure 2: Plot of Ponderosa pine growth rate against DBH in the Boulder foothills.

Simple Linear Regression: Ponderosa pine density and growth rates

Competition with other ponderosa pines may limit the resources available for ponderosa pine growth. If congeneric competition limits ponderosa pine growth rates, then we expect to find a significant negative linear relationship between growth rate and surrounding ponderosa pine density. To test this hypothesis, we conducted a simple linear regression of growth rate on density. We fail to reject the null hypothesis that there is no linear relationship between growth rate (cm/yr) and density (# trees/100 m²), $F_{(1, 448)} = 0.3691$, $p = 0.5438$. We fail to find evidence for a significant negative linear relationship between growth rate and density (Figure 3). We reject the hypothesis that

congeneric competition limits ponderosa pine growth rates.

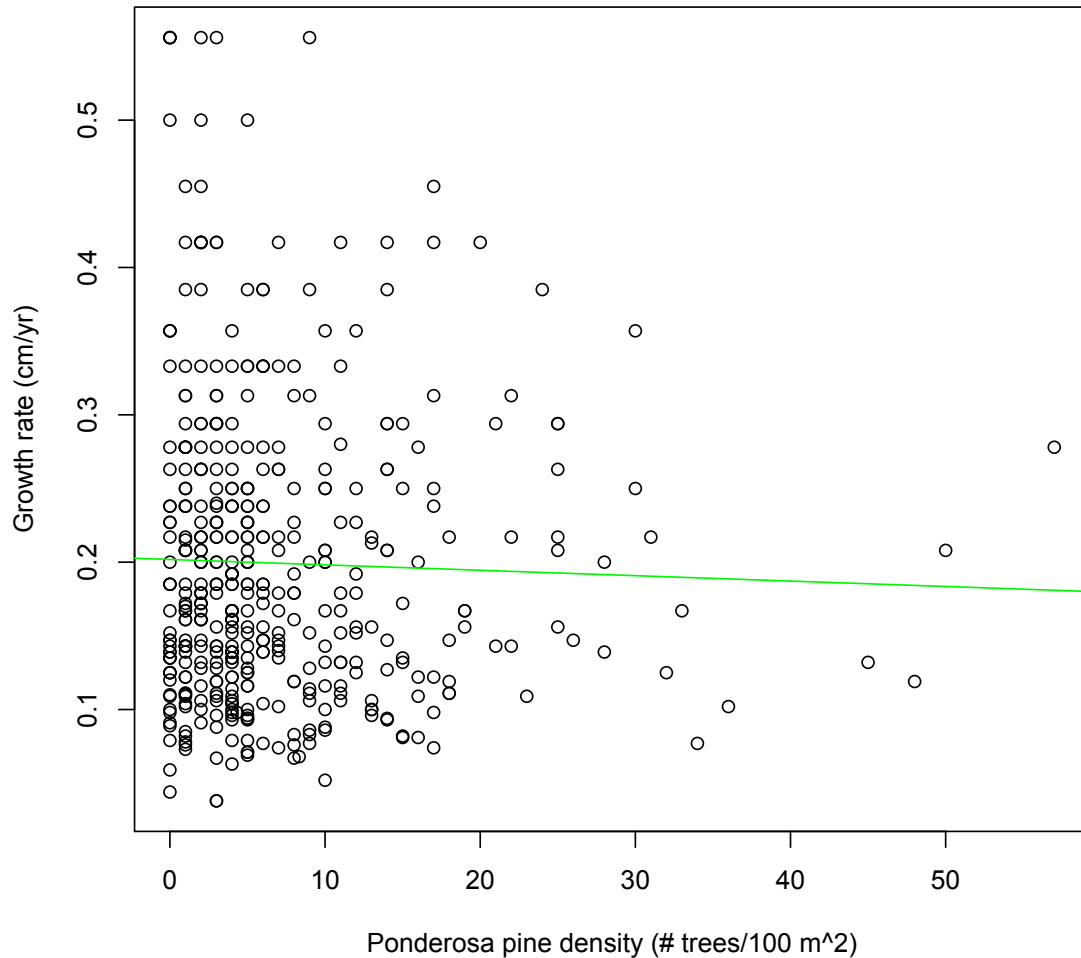


Figure 3: Plot of Ponderosa pine growth rate against Ponderosa pine density in the Boulder foothills

Multiple Regression: Elevation, DBH, density, and ponderosa pine growth rate

Elevation, tree age, and ponderosa pine density may, as a group, influence ponderosa pine growth rate. In order to determine if all three of these variables have a simultaneous effect on growth, we conducted a multiple linear regression analysis of growth rate on elevation, DBH, and ponderosa pine neighbor density. When considered as a group, we reject the null hypothesis that there is no relationship between growth rate and elevation, DBH and density as a group, $F_{(3,446)} = 6.325$, $p = 0.0003302$. We find evidence for a statistically significant relationship between growth rate and elevation, DBH, and density as a group: $\text{rate} = 3.356e-01 + (-5.341e-05)\text{elevation} + (-4.414e-04)\text{DBH} + (-5.232e-04)\text{density}$, $R^2 = 0.03436$. However, when taken together, these variables only explain 3.4% of the observed variation in growth rate and consequently are not good predictors of growth rate.

Controlling for the effects of DBH and ponderosa pine density, we found that growth rate significantly decreases as elevation increases, $t_{(1,448)} = -3.480$, $p = 0.00055$. Therefore, we conclude that growing season length has a greater effect on growth rate than precipitation levels, regardless of tree age and competition. We also found that growth rate significantly decreases as DBH increases, $t_{(1,448)} = -2.465$, $p = 0.01407$, while controlling for the effects of elevation and ponderosa pine density. Older trees grow more slowly than younger trees, regardless of growing season length and competition. We fail to find evidence for a statistically significant relationship between growth rate and ponderosa pine density, $t_{(1,448)} = -0.879$, $p = 0.38011$, while controlling for the effects of growing season length and tree age. Based on these results, we find support for the hypotheses that growing season length and tree age limit growth of Ponderosa pine in the Boulder foothills, but not for the hypothesis that congeneric competition limits growth rate.

***CAUTION: If you test a multiple linear regression model and determine that it is not a significant predictor of the dependent variable, stop right there! You cannot continue to interpret the individual effects of the variables included in the model, but rather should continue testing other biologically sound models until you find one that fits the data better. It is also important to remember that with each additional variable added to a model, the power to detect an effect is decreased. In other words, you are less likely to find evidence of a significant effect, even if an effect really does exist.