Stats Lab 6

Wynona

2024-02-24

```
tdat<- read.csv("~/ENV 872/EDA_Spring2024/Data/Raw/TreePlots.csv", header=T)
attach(tdat)
```

```
library(GGally)
require(ggplot2)
```

Problems

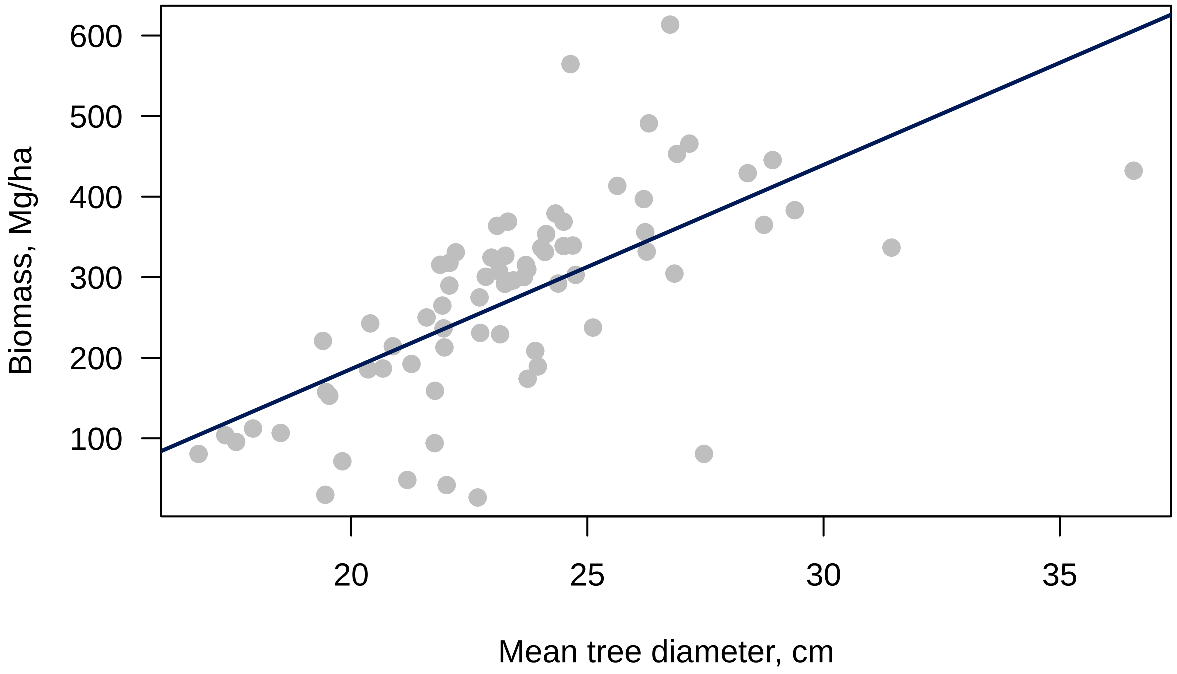
1 & 2. Your assignment is to conduct two different linear regressions on the TreePlots_Lab6.csv data.

Model (1) mean tree diameter (mDBH.cm) versus plot biomass (AGBH.Mg.ha), and (2) mean height (mH.m) versus mean wood density (mWD.g.m3). In the first model, biomass should be your dependent variable. In the second regression, mean height should be your dependent variables.

```
lm_1 <- lm(AGBH.Mg.ha ~ mDBH.cm, data = tdat)
anova(lm_1)
```

```
## Analysis of Variance Table
##
## Response: AGBH.Mg.ha
##           Df Sum Sq Mean Sq F value    Pr(>F)
## mDBH.cm    1 505149  505149  56.902 1.172e-10 ***
## Residuals 71 630303    8878
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
dukeblue <- "#001A57"
plot(mDBH.cm, AGBH.Mg.ha, lwd = 2, las = 1, pch = 19,
col = "grey", cex = 1, xlab = "Mean tree diameter, cm",
ylab = expression(paste("Biomass, Mg/ha")))
abline(lm(AGBH.Mg.ha ~ mDBH.cm), col = dukeblue, lwd = 2)
```



```r
summary(lm_1)
```

```
##
## Call:
## lm(formula = AGBH.Mg.ha ~ mDBH.cm, data = tdat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -294.67  -36.82   19.97   49.99  260.77
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -320.88      79.61  -4.031 0.000138 ***
## mDBH.cm        25.35       3.36   7.543 1.17e-10 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 94.22 on 71 degrees of freedom
## Multiple R-squared:  0.4449, Adjusted R-squared:  0.4371
## F-statistic:  56.9 on 1 and 71 DF,  p-value: 1.172e-10
```

MODEL 1: biomass (AGBH.Mg.ha) vs. mean tree diameter (mDBH.cm)

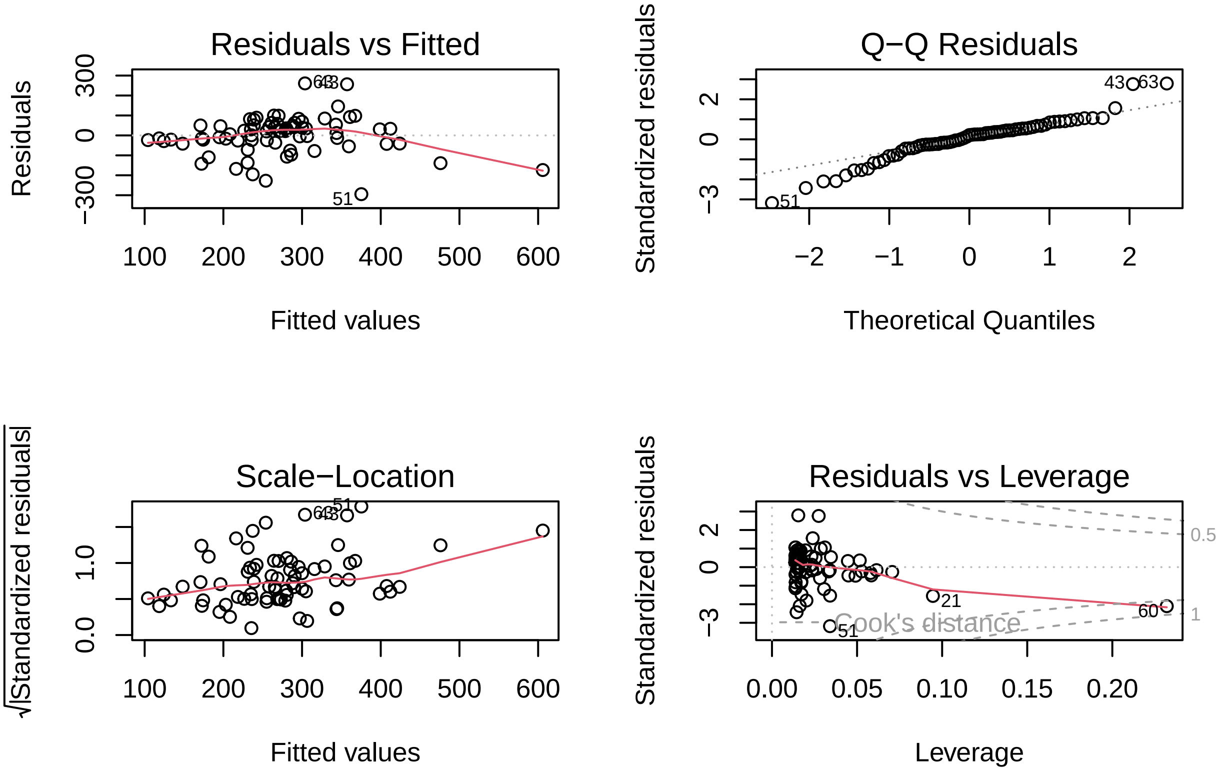
The null hypothesis is that tree diameter has no effect on plot biomass. The alternative hypothesis is that tree diameter is related to plot biomass.

Since the coefficient mDBH.cm is positive, as tree diameter increases, plot biomass increases. For every 1 cm increase in tree diameter, the plot biomass increases by 25.35 Mg ha-1 (R^2 = 0.4449, F-value= 56.902, p < 0.001). 44.49% of variation of plot biomass can be explained by its linear relationship with tree diameter. The small p-value suggests that this is relationship is statistically significant, in other words, the regression model explains a significant part of the variation in the data.

To check the assumptions of my model, I examined diagnostic plots such as Residuals vs. Fitted, Q-Q Residuals, Scale-Location and Residuals vs. Leverage. Datapoints with their values deviating away from the centreline or diagonal (depending on the plot) and are marked with a number are noted as follows.

In the residuals vs. fitted line, the points should be randomly distributed around the centerline. However, there is a deviation where the line forms a banana shape especially due to the values of 43, 51 and 63, suggesting a violation of linearity or homoscedasticity. Similarly, for the Q-Q residuals plot, the points should be aligned with the diagonal dotted line but the same three values are sticking out on the top right corner, suggesting that there are violations from the normality assumption. The scale-location plot shows those three data points scattered away from the center while the residuals-leverage plot shows that one of those values (60) are past the dotted line representing Cook's distance, meaning it is definitely an outlier. These points have the largest effect on parameter estimates.

```r
par(mfrow=c(2,2))
plot(lm_1)
```

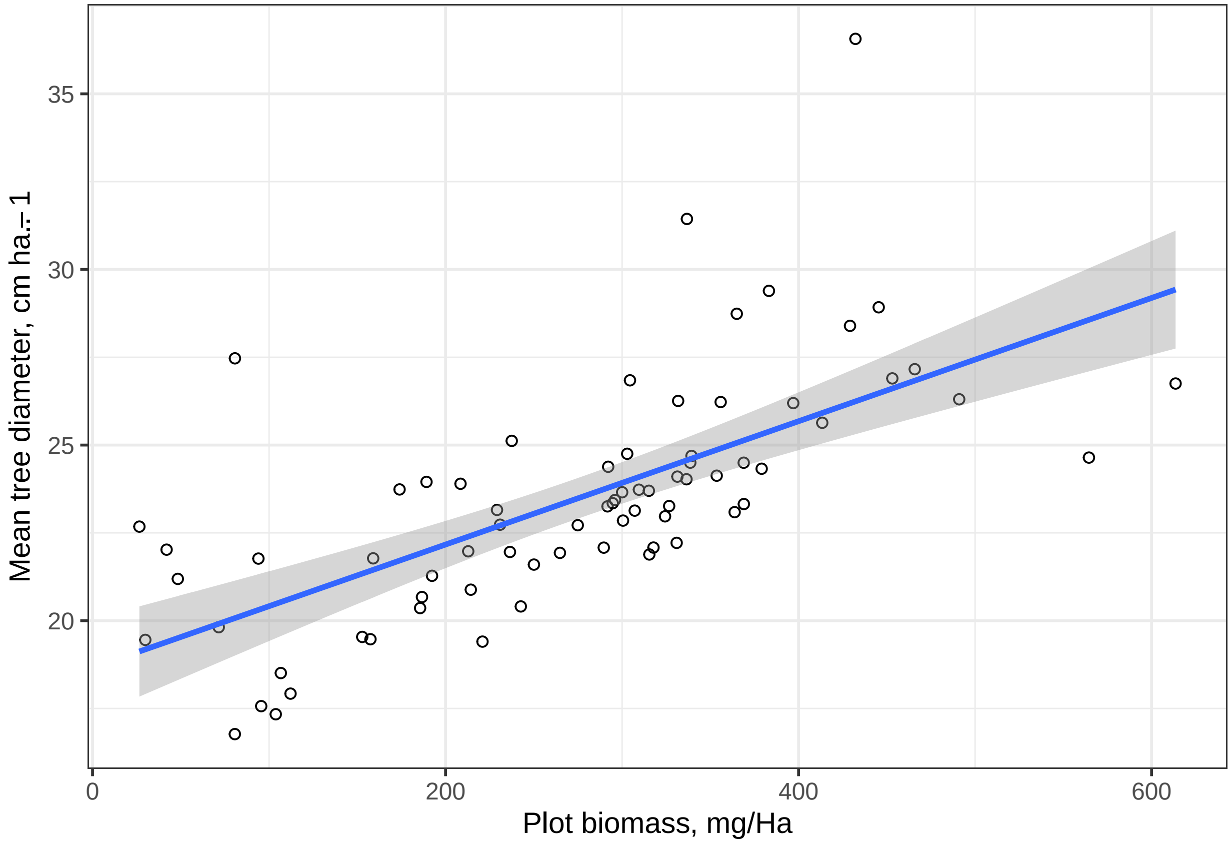


As the values 21, 43 and 63 show up repeatedly as having a high influence on the model, and since 60 is an outlier, I would like to remove these data points to see how the model would change. It is a good sign that the R-squared value increases to 0.5116, which means that the variability of biomass can be explained to a significant degree by changes in diameter at breast height. This model can explain only 51.16% of variability of biomass. Another change is that the diameter at breast height coefficient increased to 29.423, which means that as diameter increases by 1 cm, biomass increases by 29.423 Mg/ha.

```
lm_1_no_outliers <- with(tdat[-c(21,43,60,63), ], lm(AGBH.Mg.ha ~ mDBH.cm))
summary(lm_1_no_outliers)
```

```
##
## Call:
## lm(formula = AGBH.Mg.ha ~ mDBH.cm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -309.55  -19.97   12.39   47.57  135.08
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -418.005     81.695  -5.117 2.81e-06 ***
## mDBH.cm       29.423      3.512   8.377 5.04e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 80.82 on 67 degrees of freedom
## Multiple R-squared:  0.5116, Adjusted R-squared:  0.5043
## F-statistic: 70.17 on 1 and 67 DF,  p-value: 5.037e-12
```

```
require(ggplot2)
tdat_1<-tdat[-c(21,43,60,63)]
ggplot(tdat_1, aes(x = AGBH.Mg.ha, y = mDBH.cm)) +
geom_point(shape = 1) +
geom_smooth(method = lm) +
xlab("Plot biomass, mg/Ha") +
ylab(expression(paste("Mean tree diameter, cm ", ha^-1))) +
theme_bw()
```



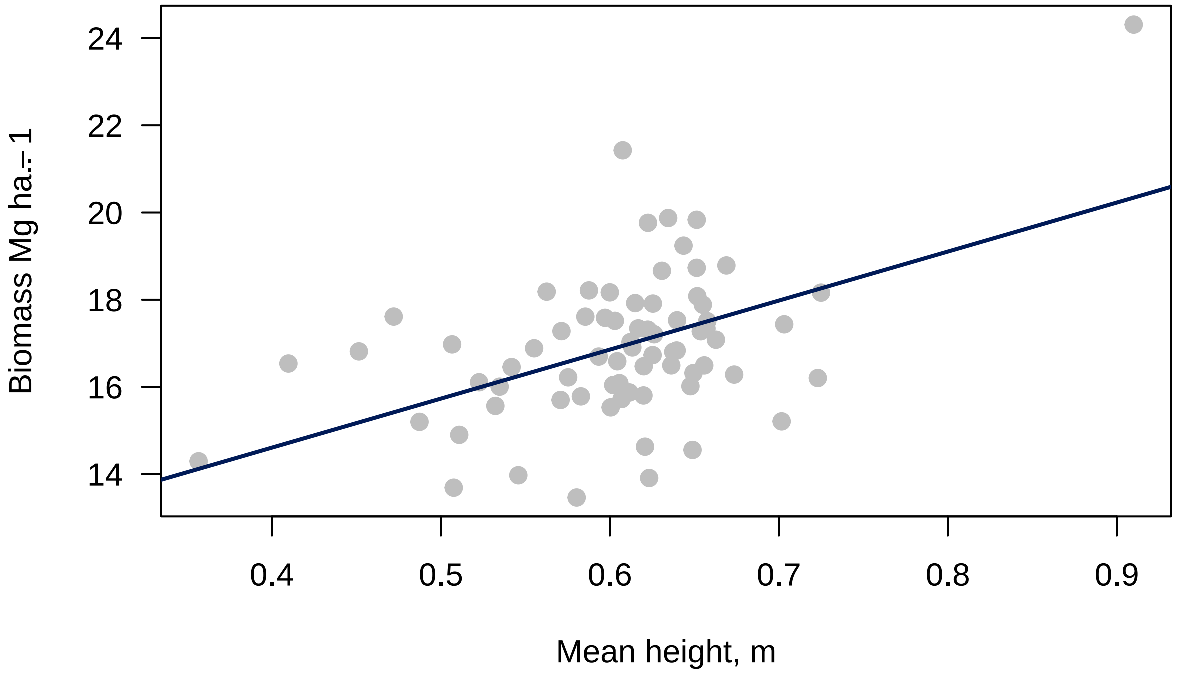
Model (1) mean tree diameter (mDBH.cm) versus plot biomass (AGBH.Mg.ha), and (2) mean height (mH.m) versus mean wood density (mWD.g.m3). In the first model, biomass should be your dependent variable. In the second regression, mean height should be your dependent variable

MODEL 2: mean height (mH.m) versus mean wood density (mWD.g.m3)

```
lm_2 <- lm(mH.m ~ mWD.g.m3, data = tdat)
anova(lm_2)
```

```
## Analysis of Variance Table
##
## Response: mH.m
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## mWD.g.m3   1  51.255  51.255  22.269 1.155e-05 ***
## Residuals 71 163.413   2.302
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
dukeblue <- "#001A57"
plot(mWD.g.m3, mH.m, lwd = 2, las = 1, pch = 19,
col = "grey", cex = 1, xlab = "Mean height, m",
ylab = expression(paste("Biomass Mg ", ha^-1)))
abline(lm(mH.m ~ mWD.g.m3), col = dukeblue, lwd = 2)
```



```
summary(lm_2)
```

```
##
## Call:
## lm(formula = mH.m ~ mWD.g.m3, data = tdat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2065 -0.8346 -0.0980  0.7695  4.4858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.111      1.456   6.945 1.48e-09 ***
## mWD.g.m3      11.242      2.382   4.719 1.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.517 on 71 degrees of freedom
## Multiple R-squared:  0.2388, Adjusted R-squared:  0.228
## F-statistic: 22.27 on 1 and 71 DF,  p-value: 1.155e-05
```

The null hypothesis is that the wood density has no effect on mean height. The alternative hypothesis is that wood density is related to mean height.

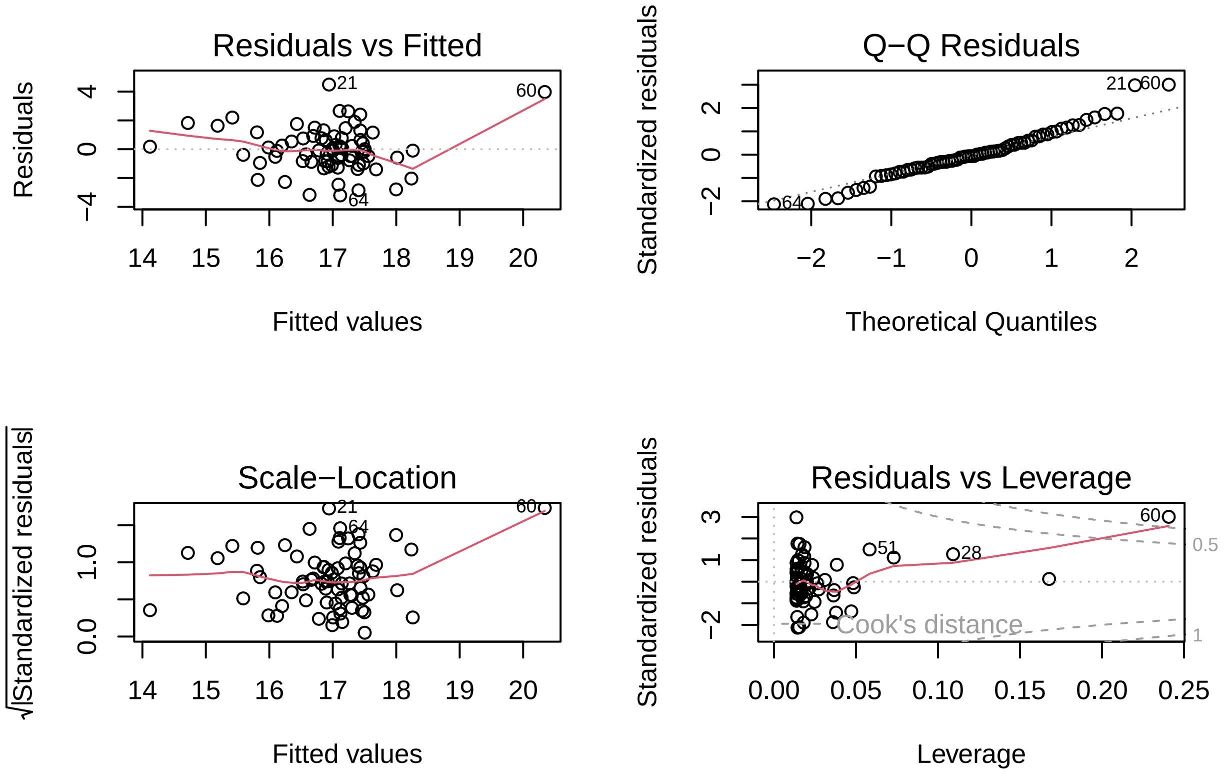
The coefficient of wood density is positive which means that as wood density increases, mean height also increases. As wood density increases by 1 g/mˆ3, mean height increases by 11.242 m. (Rˆ2 = 0.2388,

F-value= 22.269, $p < 0.001$). 23.88% of variation of mean height can be explained by its linear relationship with wood density.

To check the assumptions of my model, I examined diagnostic plots such as Residuals vs. Fitted, Q-Q Residuals, Scale-Location and Residuals vs. Leverage. If there are data points marked for deviating significantly from the centerline or diagonal line (in the case of Q-Q Residuals), they were taken note of.

In the residuals vs. fitted line, the points should be randomly distributed around the centerline, but in this case, there is a clear deviation especially when the fitted value is 21 and 60, suggesting a violation of linearity or homoscedasticity. Similarly, for the Q-Q residuals plot, the points should be aligned with the diagonal dotted line but that is not the case for values 21 and 60, suggesting that there are violations from the normality assumption. The residuals-leverage plot and scale-location plot shows that values 60, 28 and 51 have the largest effect on parameter estimates.

```r
par(mfrow=c(2,2))
plot(lm_2)
```



Since values 60 and 21 show up repeatedly here as potential outliers, I would like to test how the linear model would change if those data points were removed. Interestingly, the R-squared value decreases to 0.119, which means that the variability of height is determined more heavily by other variables other than wood density. This model can explain only 11.9% of variability of height. Another change is that the wood density coefficient decreased to 7.253, which means that as wood density increases by 1g/m3, height increases by 7.253 m.

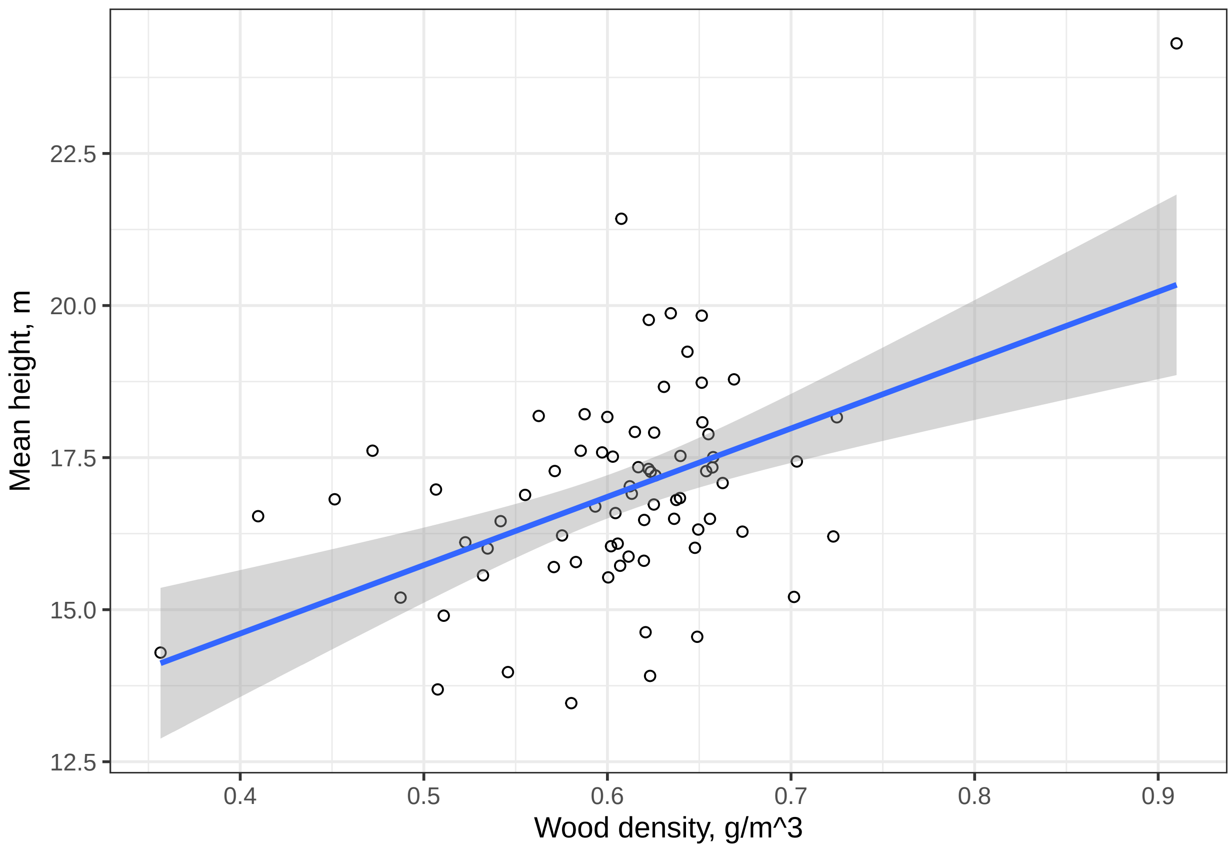
```r
lm_2_no_outliers <- with(tdat[-c(21,60), ], lm(mH.m ~ mWD.g.m3))
summary(lm_2_no_outliers)
```

```
##
```

```
## Call:
## lm(formula = mH.m ~ mWD.g.m3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1396 -0.7586 -0.0022  0.8836  2.8773
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   12.394      1.439   8.611 1.53e-12 ***
## mWD.g.m3       7.253      2.376   3.053  0.00322 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.327 on 69 degrees of freedom
## Multiple R-squared:  0.119,  Adjusted R-squared:  0.1062
## F-statistic: 9.321 on 1 and 69 DF,  p-value: 0.003217
```

- A scatter graph showing the data and the best-fit regression line

```
require(ggplot2)
tdat_2<-tdat[-c(21,60)]
ggplot(tdat_2, aes(x = mWD.g.m3, y = mH.m)) +
geom_point(shape = 1) +
geom_smooth(method = lm) +
xlab("Wood density, g/m^3") +
ylab(expression(paste("Mean height, m"))) +
theme_bw()
```



3. Also model the relationship between brain weight, kg, and body weight, kg, of different animals. Test the alternative hypothesis that brain weight increases with body weight. If the data show large deviations from normality, then use an appropriate transformation. Complete all the bullet points in problem 1 for this problem.

```
animal.data<-read.csv("~/ENV 872/EDA_Spring2024/Data/Raw/x01.csv", header=T)
head(animal.data)
```

```
##   Index Brain.Weight Body.Weight
## 1     1        3.385        44.5
## 2     2        0.480        15.5
## 3     3        1.350         8.1
## 4     4      465.000       423.0
## 5     5       36.330       119.5
## 6     6       27.660       115.0
```

```
colnames(animal.data)
```

```
## [1] "Index"        "Brain.Weight" "Body.Weight"
```

```
lm_3 <- lm(Brain.Weight~Body.Weight, data = animal.data)
anova(lm_3)
```

```
## Analysis of Variance Table
##
## Response: Brain.Weight
##             Df   Sum Sq  Mean Sq F value    Pr(>F)
## Body.Weight  1 43037593 43037593  411.19 < 2.2e-16 ***
## Residuals   60  6279999   104667
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
summary(lm_3)
```

```
##
## Call:
## lm(formula = Brain.Weight ~ Body.Weight, data = animal.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1552.25    -8.00    47.36    55.10  1553.42
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -56.85555   42.97805  -1.323    0.191
## Body.Weight   0.90291    0.04453  20.278   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 323.5 on 60 degrees of freedom
## Multiple R-squared:  0.8727, Adjusted R-squared:  0.8705
## F-statistic: 411.2 on 1 and 60 DF,  p-value: < 2.2e-16
```

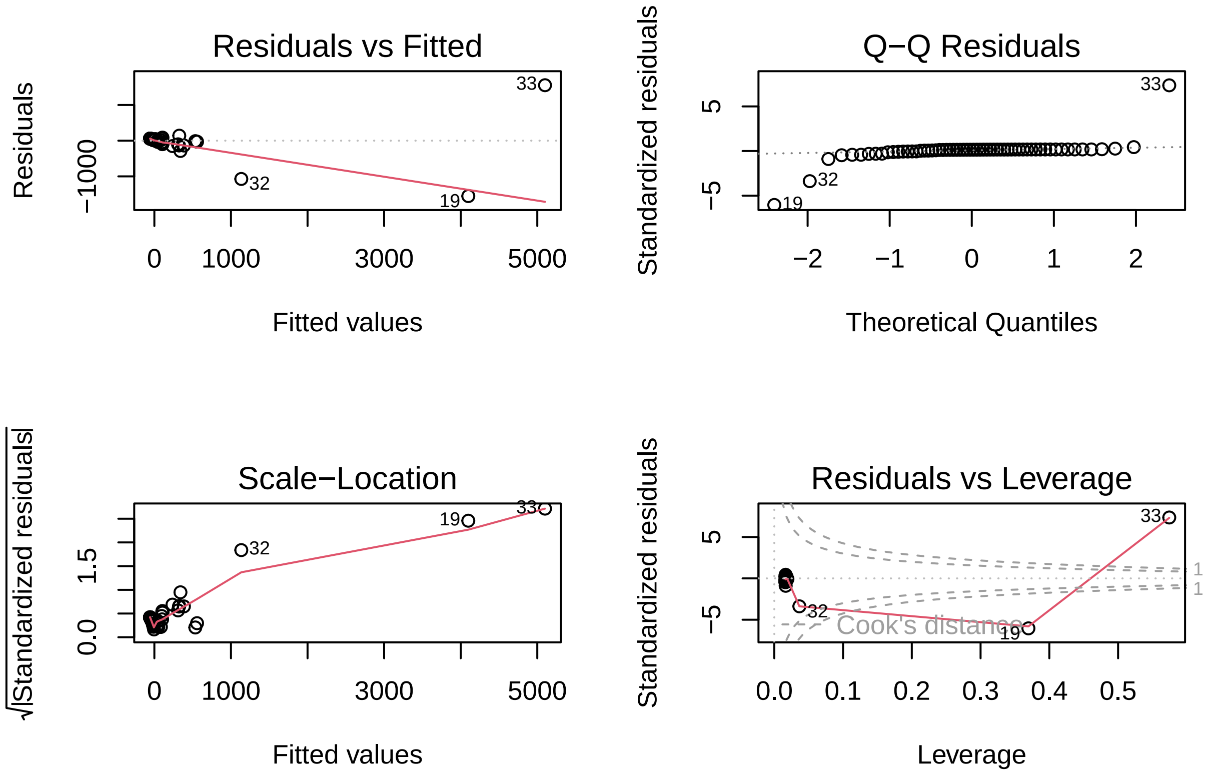
The null hypothesis is that the body weight has no effect on brain weight The alternative hypothesis is that body weight is related to brain weight.

The coefficient of body weight is positive which means that as body weight increases, brain weight also increases. As body weight increases by 1 kg, brain weight increases by 0.90291 m. (Rˆ2 = 0.8727, F-value= 411.19, $p < 0.001$). 87.27% of variation of brain weight can be explained by its linear relationship with body weight.

To check the assumptions of my model, I examined diagnostic plots such as Residuals vs. Fitted, Q-Q Residuals, Scale-Location and Residuals vs. Leverage. If there are data points marked for deviating significantly from the centerline or diagonal line (in the case of Q-Q Residuals), they were taken note of.

In the residuals vs. fitted line, the points should be randomly distributed around the centerline, but in this case, there is a very noticeable deviation for values 19, 32 and 33, suggesting a violation of linearity or homoscedasticity. Similarly, for the Q-Q residuals plot, the points should be aligned with the diagonal dotted line but that is not the case for the same three values, suggesting that there are violations from the normality assumption. The residuals-leverage plot and scale-location plot shows that these three values have the largest effect on parameter estimates.

```
par(mfrow=c(2,2))
plot(lm_3)
```

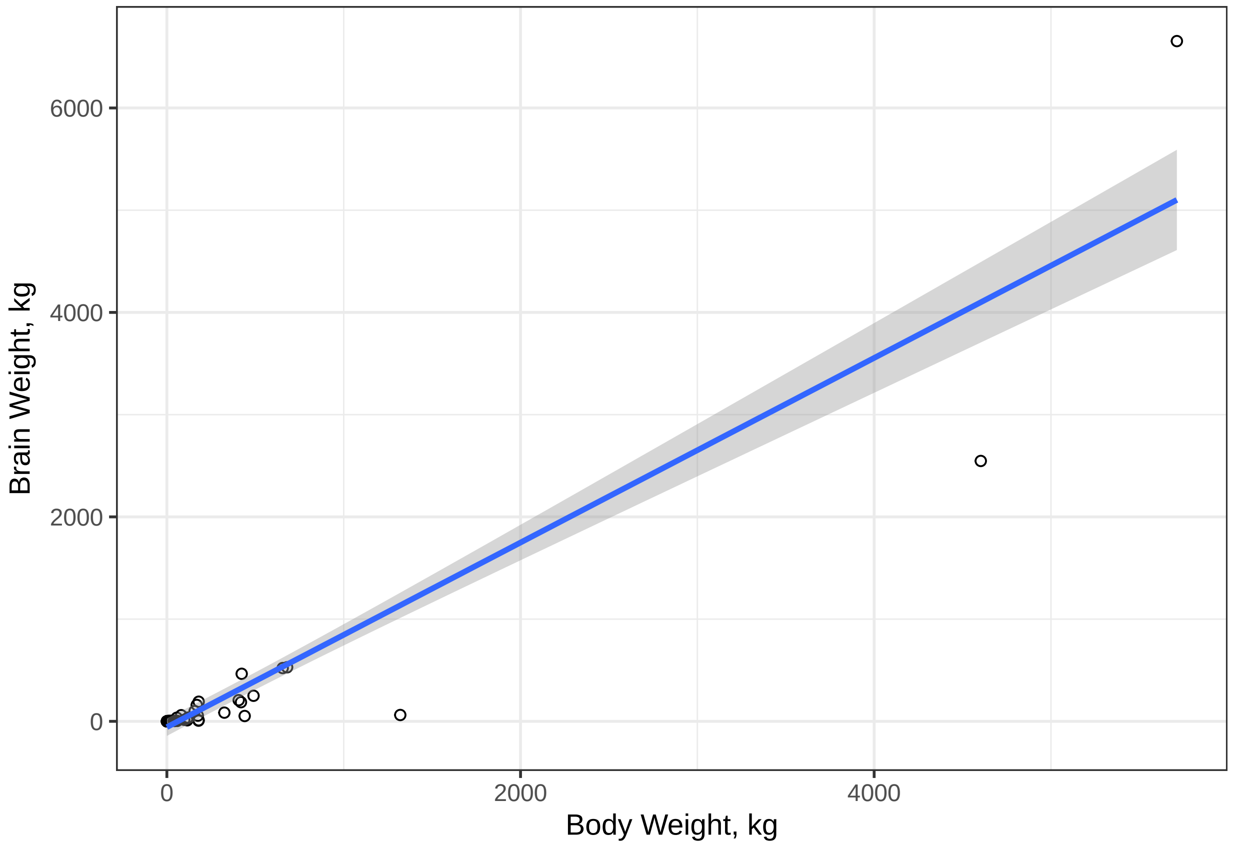


Since values 19, 32 and 33 show up repeatedly here as potential outliers, I would like to test how the linear model would change if those data points were removed. Interestingly, the R-squared value decreases to 0.7893, but it is still high which means that the variability in brain weight can be explained a large extent (78.93%) by its relationship to body weight. Another change is that the body weight coefficient decreased to 0.64248, which means that as body weight increases by 1 kg, brain weight increases by 0.64248 kg.

```r
lm_3_no_outliers <- with(animal.data[-c(19,32,33), ], lm(Brain.Weight~Body.Weight))
summary(lm_3_no_outliers)
```

```
##
## Call:
## lm(formula = Brain.Weight ~ Body.Weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -217.971  -10.491    9.122   11.827  205.791
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.56054    8.49839  -1.478    0.145
## Body.Weight   0.64248    0.04397  14.611   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55.79 on 57 degrees of freedom
## Multiple R-squared:  0.7893, Adjusted R-squared:  0.7856
## F-statistic: 213.5 on 1 and 57 DF,  p-value: < 2.2e-16
```

```r
require(ggplot2)
animal.data_transformed<-animal.data[-c(19,32,33)]
ggplot(animal.data_transformed, aes(x = Body.Weight, y =  Brain.Weight)) +
geom_point(shape = 1) +
geom_smooth(method = lm) +
xlab("Body Weight, kg") +
ylab(expression(paste("Brain Weight, kg"))) +
theme_bw()
```


Brain Weight, kg
Body Weight, kg