Homework 1 - Stats 230

Dhyey Mavani

date

Due Monday, Feb 14 at 10:00pm

Dhyey Mavani

PROBLEMS TO TURN IN: #0.08, #0.14, #0.15, #1.35, #1.42, Additional #1, Additional #2

Exercise 0.08 - Predicting NFL wins (Pg. 90 of my book pdf)

part a:

SOLUTION: "Wins" is a response variable in this case. "Wins" is a quantitative variable.

part b:

SOLUTION: "PF" and "PA" are the explanatory variables in this case. They are both quantitative variables.

part c:

SOLUTION: They expect to achieve 0.5 * 3 = 1.5 more wins if they increase their scoring by an average of 3 points per game.

part d:

SOLUTION: They expect to achieve 0.3 * 3 = 0.9 more wins if they decrease their points allowed by an average of 3 points per game.

part e:

SOLUTION: Based on my answers to the previous two parts, I think it makes more sense for the team to focus more on improving it offense compared to improving its defense because the same improvement (increase of PF or decrease of PA) leads to more wins ($1.5 > 0.9$) when the team focuses on improving their offense by increasing PF.

part f:

SOLUTION: The data analysed in this study are observational.

Exercise 0.14 - Predicting NFL wins: Patriots/Chargers (Pg. 91 of my book pdf)

part a:

SOLUTION: PF = 441/16, PA = 250/16, Predicted Wins = 3.6 + 0.5 * (PF) - 0.3 * (PA) = 3.6 + 13.78125 - 4.6875 = 12.69375 = 13 games (aprroximately, rounded to nearest whole number)

part b:

SOLUTION: The residual error would be 14 - 12.69375 = 1.30625. This means that our model's prediction of the number of wins was off by 1.30625 games from the actual number of wins which was realized later in time through observation.

part c:

SOLUTION: The residual of -3.48 means that the our model overestimated the number of wins by 3.48, or the actual number of wins were 3.48 less than the predicted number of wins.

Exercise 0.15 - Roller coasters

part a:

SOLUTION: For wooden roller coster, TopSpeed = 54 + 7.6 * (0) = 54 mph

part b:

SOLUTION: For steel roller coster, TopSpeed = 54 + 7.6 * (1) = 61.6 mph

part c:

SOLUTION: The steel roller coster is 7.6 mph faster than the wooden roller coster. This stems from the coefficient of boolean variable TypeCode (which takes value 0 or 1)

```
data(Pines)
```

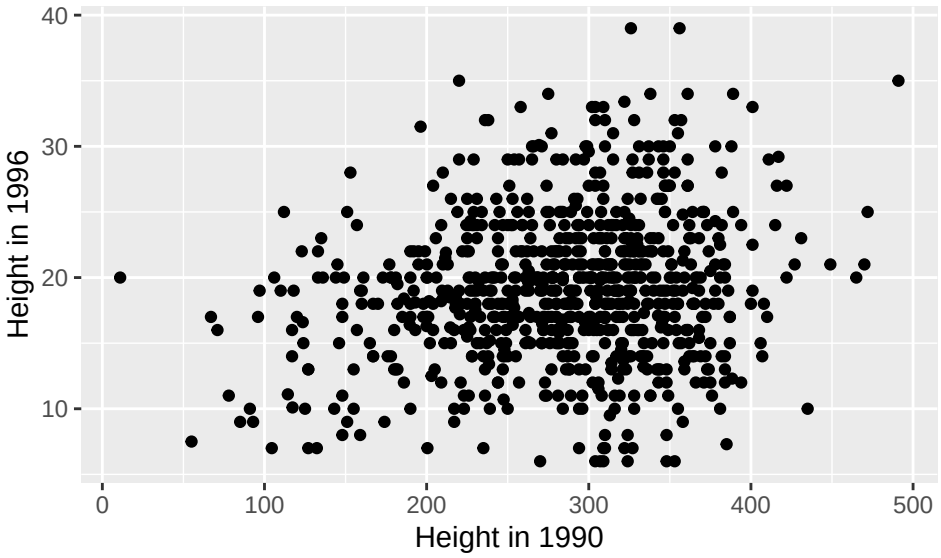
Exercise 1.35 - Pines

part a: Scatterplot and comment on relationship

SOLUTION: Generally we can see that we have a positive or directly proportional relationship between the heights in 1990 and heights in 1996.

```
gf_point(Hgt90 ~ Hgt96, data = Pines) %>%
  gf_labs(x = "Height in 1990", y = "Height in 1996")
```

```
## Warning: Removed 193 rows containing missing values (geom_point).
```



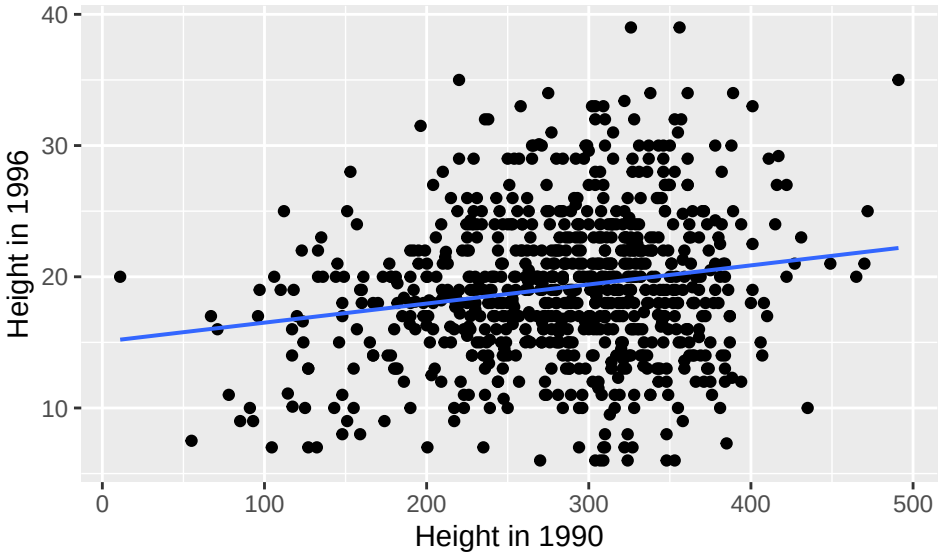
part b: Fit a least squares line and report it

SOLUTION: Least squares line: Hgt96 = 15.04822 + 0.01455 * Hgt90

```
#show appropriate output and report line below
gf_point(Hgt90 ~ Hgt96, data = Pines) %>%
  gf_labs(x = "Height in 1990", y = "Height in 1996") %>%
  gf_lm()
```

```
## Warning: Removed 193 rows containing non-finite values (stat_lm).
```

```
## Warning: Removed 193 rows containing missing values (geom_point).
```



```r
# Fit
fm <- lm(Hgt90 ~ Hgt96, data = Pines)
msummary(fm)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.04822    0.81657   18.43  < 2e-16 ***
## Hgt96        0.01455    0.00279    5.22  2.3e-07 ***
##
## Residual standard error: 5.55 on 805 degrees of freedom
##   (193 observations deleted due to missingness)
## Multiple R-squared:  0.0327, Adjusted R-squared:  0.0315
## F-statistic: 27.3 on 1 and 805 DF,  p-value: 2.28e-07
```

```r
confint(fm, level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) 13.44535604 16.6510797
## Hgt96        0.00907876  0.0200209
```

> part c: Text says - are you satisfied with the fit of this SLR? This means assess and discuss the fit of the model (this does NOT mean the p-values!) Assess = conditions. Fit = some other numbers too!

SOLUTION: Yes, at first pass from a visual standpoint we can proceed. We will need to use the LINE pnuemonic to consider during the ASSESS phase.

Linearity - Assess via a scatterplot of X and Y.

We can see the linearity using the scatterplot above.

Independence - Read the research design/methods to see if it observations were independent (randomly selected)

We cannot say anything about this using R, so we need to read the experiment and documentation.

Normality - Assess the distribution of errors with histograms/qqplots of residuals errors should be centered at zero (ZERO MEAN), no skew or pattern (RANDOM)- necessary for inference.

The histogram is relatively normal and centered at zero (latter condition is always true for least squares technique). There are a few high residuals > 15 that many need to be considered. All other residuals fall between +/- 15.

In the QQPlot, we can see that most points are on the line, there is no real shape and only a few are off at the ends, hence we are good there!

Equal (Constant) Variance - Assess with a residual vs. fitted plot. Looking for no pattern

Next, we can check the condition for constant variance of errors by looking at the residual vs fitted plots.

A few outliers are noted in red but nothing exerting a great deal of leverage (unduly influencing the overall relationship). There is no pattern of errors (looks like a cloud) and is mostly linear. IF the variance in Y was not equal across X we could see a fan shape indicating heteroscedastcity which would need to be resolved with a transformation.

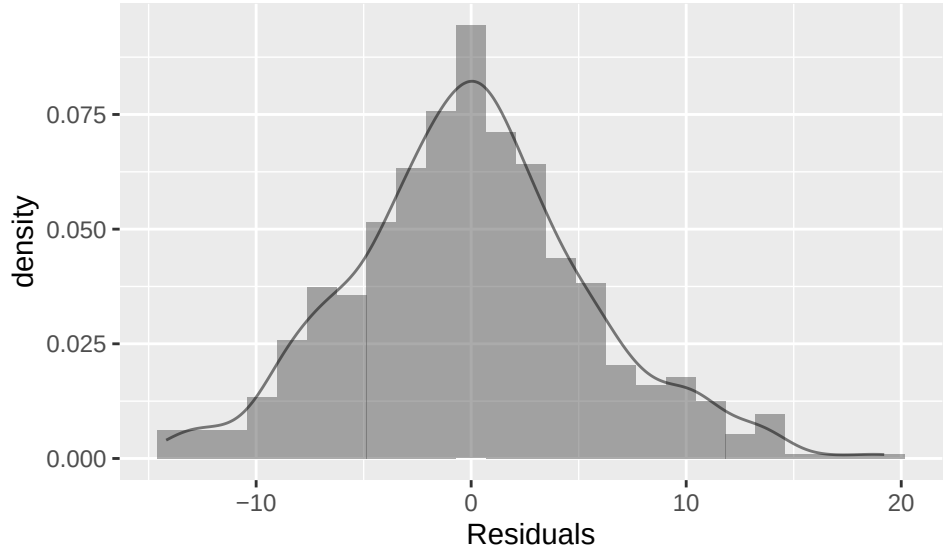
```r
# Fit
fm <- lm(Hgt90 ~ Hgt96, data = Pines)
msummary(fm)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.04822    0.81657   18.43  < 2e-16 ***
## Hgt96        0.01455    0.00279    5.22  2.3e-07 ***
##
## Residual standard error: 5.55 on 805 degrees of freedom
##   (193 observations deleted due to missingness)
## Multiple R-squared:  0.0327, Adjusted R-squared:  0.0315
## F-statistic: 27.3 on 1 and 805 DF,  p-value: 2.28e-07
```

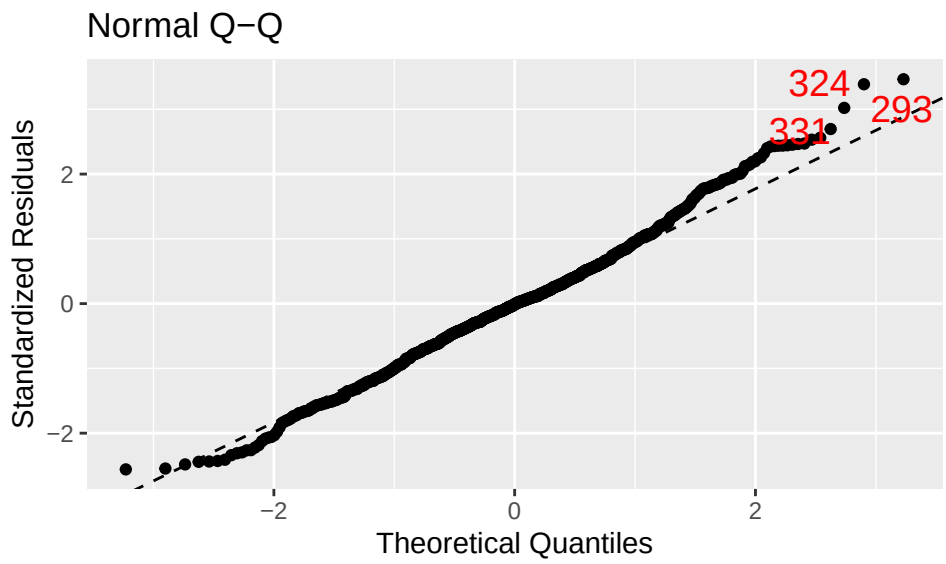
```
confint(fm, level = 0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) 13.44535604 16.6510797
## Hgt96        0.00907876  0.0200209
```

```
# Assess
# generates a histogram with fitted density curve
gf_dhistogram(~ residuals(fm), xlab = "Residuals") %>%
  gf_dens()
```

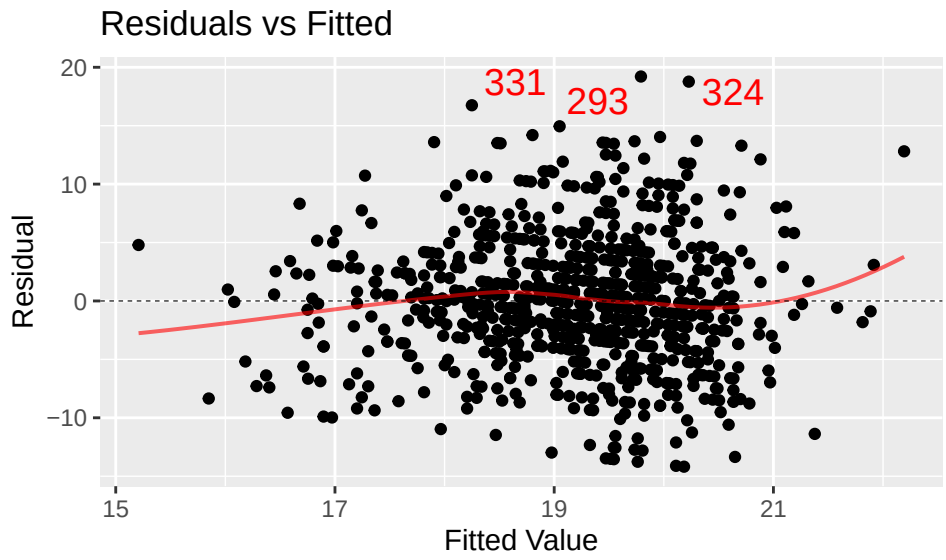



```
# easiest way to get QQplot for residuals
mplot(fm, which = 2)
```




```
# residual vs fitted plot
mplot(fm, which = 1)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
data(Goldenrod) #description is in Exercise 1.41 - Read it!
```

Exercise 1.42 - Goldenrod Galls 2004 A note about understanding Pairwise.Complete.Obs
We have missing some missing data for Goldenrod Gall data. We have to know how to deal with this because certain assumptions are made in R when fitting a linear regression model. R will just ignore these missing values when you produce the scatterplot and the linear model.

The following commands produce different results
COMMAND 1: cor(Gdiam04 ~ Stdiam04, data = Goldenrod) COMMAND 2: cor(Gdiam04 ~ Stdiam04, data = Goldenrod, use = "pairwise.complete.obs")

Command 1 will result in nothing because R doesn't know what to do with the missing data. If you want to get correlations, you need to set the "use" option, shown in Command 2. The "use" option tells R to use all pairwise complete observations. This means that observations are included in the computation as long as they have values for both variables under consideration. It is NOT the same as using only complete cases (values must exist for all variables, even the ones you aren't working with).

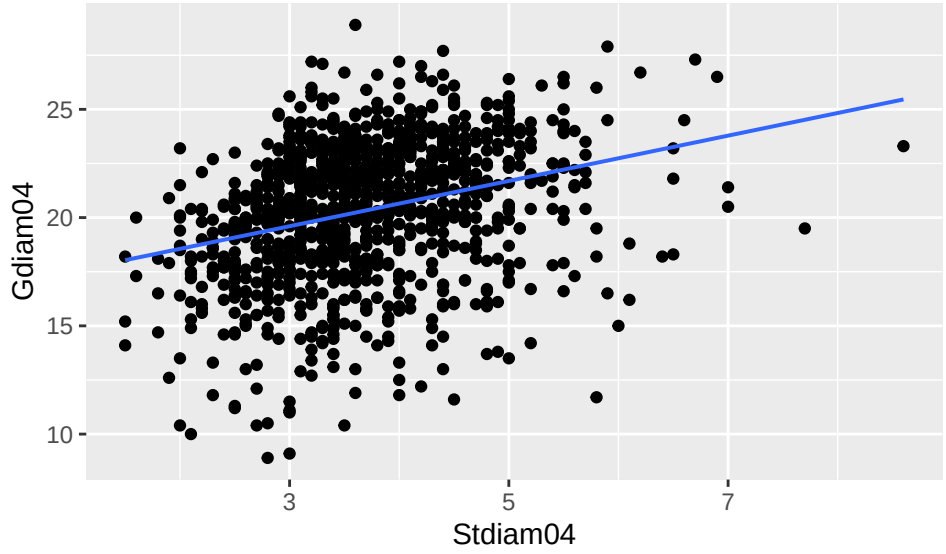
part a: Check for a positive correlation in 2004

SOLUTION: There is positive correlation between Gdiam04 and Stdiam04. Because the correlation is greater than 0 we can say that there is a positive correlation.

```
gf_point(Gdiam04 ~ Stdiam04, data = Goldenrod) %>% gf_lm()
```

```
## Warning: Removed 3 rows containing non-finite values (stat_lm).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```




```
cor(Gdiam04 ~ Stdiam04, data = Goldenrod, use = "pairwise.complete.obs") #necessary because not all are
```

```
## [1] 0.300735
```

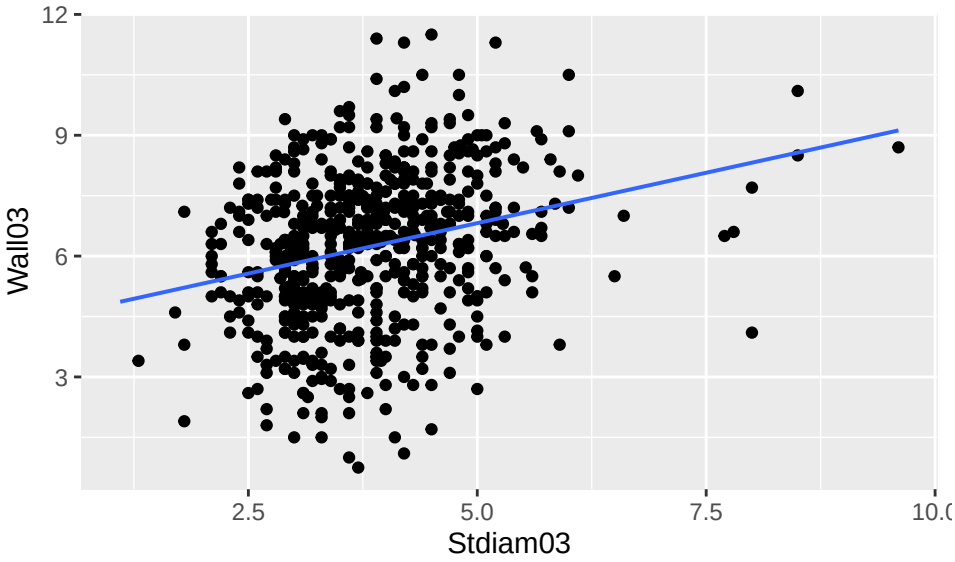
part b: Compare relationships with wall thickness

SOLUTION: We can see that even in the year 2003, both the diameters (gall and stem) are having a positive correlation with wall thickness since correlation coefficients are greater than 1. But, the correlation coefficient is significantly higher between wall thickness and gall diameter.

```
gf_point(Wall03 ~ Stdiam03, data = Goldenrod) %>% gf_lm()
```

```
## Warning: Removed 460 rows containing non-finite values (stat_lm).
```

```
## Warning: Removed 460 rows containing missing values (geom_point).
```



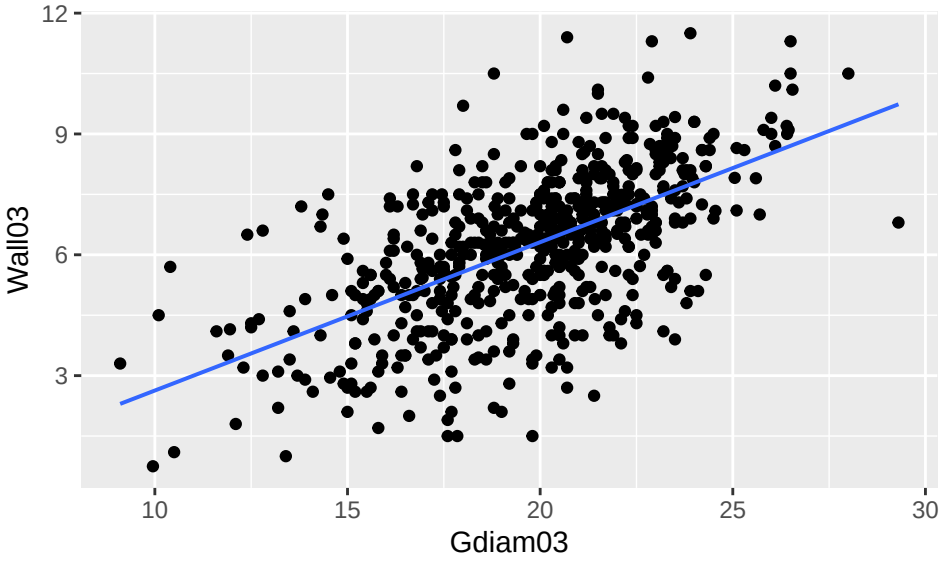

```
cor(Wall03 ~ Stdiam03, data = Goldenrod, use = "pairwise.complete.obs") #necessary because not all are
```

```
## [1] 0.266839
```

```
gf_point(Wall03 ~ Gdiam03, data = Goldenrod) %>% gf_lm()
```

```
## Warning: Removed 460 rows containing non-finite values (stat_lm).
```

```
## Warning: Removed 460 rows containing missing values (geom_point).
```



```
cor(Wall03 ~ Gdiam03, data = Goldenrod, use = "pairwise.complete.obs") #necessary because not all are c
```

```
## [1] 0.602387
```

part c: Fit a least squares line and report it for stronger relationship

SOLUTION: Wall03 = -1.052 + 0.368 * Gdiam03

```
# Fit
fm <- lm(Wall03 ~ Gdiam03, data = Goldenrod)
msummary(fm)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.052      0.401   -2.62   0.0089 **
## Gdiam03        0.368      0.020   18.38   <2e-16 ***
##
## Residual standard error: 1.5 on 593 degrees of freedom
##   (460 observations deleted due to missingness)
## Multiple R-squared:  0.363,  Adjusted R-squared:  0.362
## F-statistic:  338 on 1 and 593 DF,  p-value: <2e-16
```

```
confint(fm, level = 0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) -1.839744 -0.264477
## Gdiam03      0.328864  0.407565
```

part d: Find fitted value and residual for second observation

SOLUTION: We can see that our fitted value for the second observation is 6.93815 and our residual for the second observation is 0.361852502.

```
library(broom)
augmentG <- augment(fm)
show(augmentG)
```

```
## # A tibble: 595 x 9
##    .rownames Wall03 Gdiam03 .fitted .resid    .hat .sigma   .cooksd .std.resid
##    <chr>      <dbl>   <dbl>   <dbl>  <dbl>   <dbl>  <dbl>     <dbl>      <dbl>
##  1 1            6      20.7    6.57 -0.570 0.00183   1.50 0.000133      -0.380
##  2 2            7.3    21.7    6.94  0.362 0.00234   1.50 0.0000682      0.241
##  3 3            7.8    20.5    6.50  1.30  0.00177   1.50 0.000671       0.869
##  4 5            7.1    19.5    6.13  0.972 0.00169   1.50 0.000356       0.648
##  5 6            7.1    18.1    5.61  1.49  0.00218   1.50 0.00108        0.992
##  6 7            8.6    23      7.42  1.18  0.00353   1.50 0.00110        0.790
##  7 8           10.5    26.5    8.71  1.79  0.00973   1.50 0.00709        1.20
##  8 10          10      21.5    6.86  3.14  0.00221   1.50 0.00484        2.09
##  9 11           7.5    20      6.31  1.19  0.00169   1.50 0.000531       0.792
## 10 12           8.1    17.9    5.54  2.56  0.00231   1.50 0.00338        1.71
## # ... with 585 more rows
```

part e: Report value of a typical residual (Hint: this is a value in the output that helps to assess model fit.)

SOLUTION: typical residual is 1.5 based on the summary output.

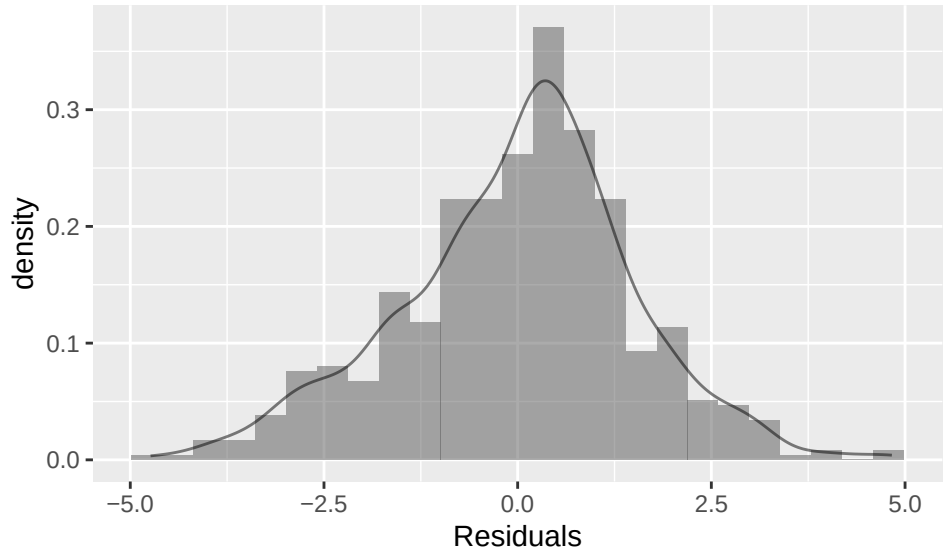
```
# Fit
fm <- lm(Wall03 ~ Gdiam03, data = Goldenrod)
msummary(fm)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.052      0.401   -2.62   0.0089 **
## Gdiam03        0.368      0.020   18.38   <2e-16 ***
##
## Residual standard error: 1.5 on 593 degrees of freedom
##   (460 observations deleted due to missingness)
## Multiple R-squared:  0.363,  Adjusted R-squared:  0.362
## F-statistic:  338 on 1 and 593 DF,  p-value: <2e-16
```

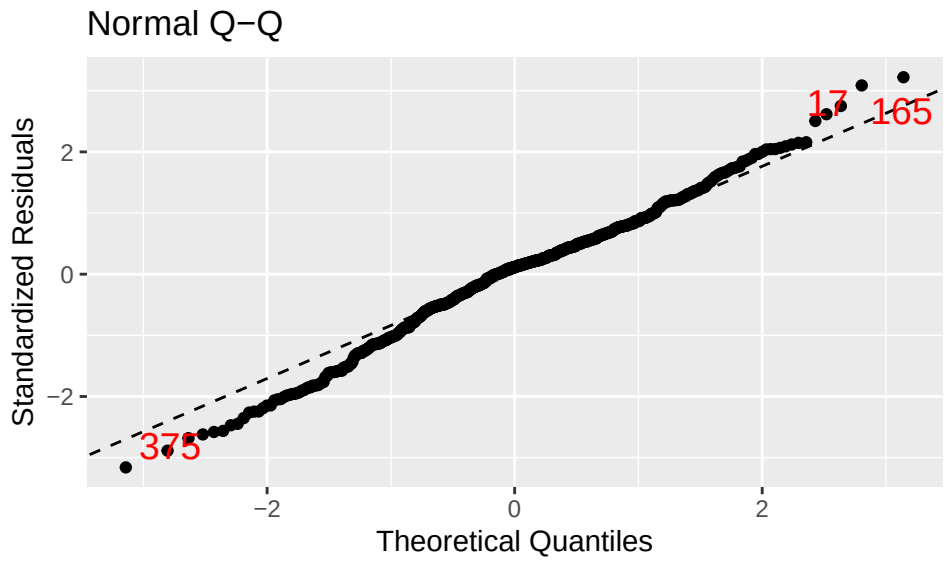
```
confint(fm, level = 0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) -1.839744 -0.264477
## Gdiam03      0.328864  0.407565
```

```
# Assess
# generates a histogram with fitted density curve
gf_dhistogram(~ residuals(fm), xlab = "Residuals") %>%
  gf_dens()
```

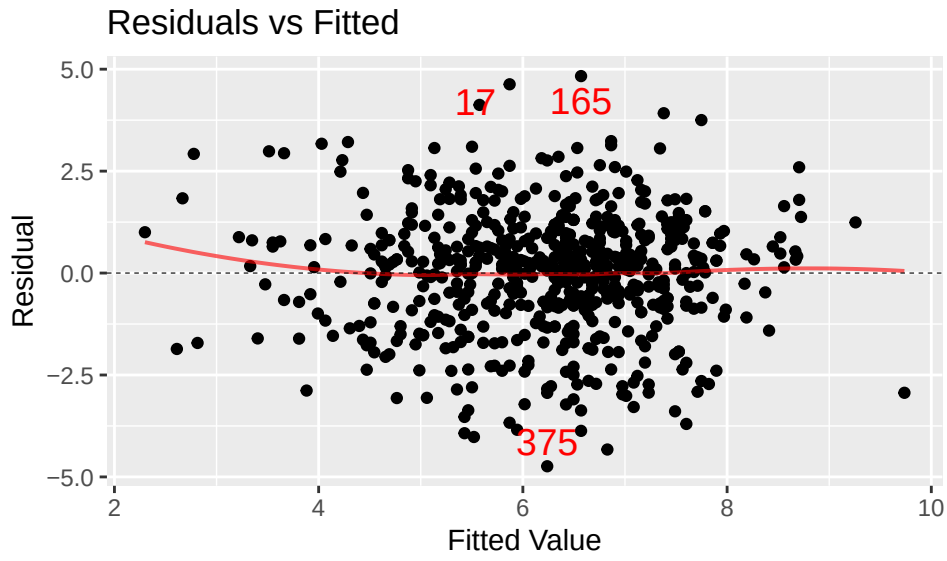



```
# easiest way to get QQplot for residuals
mplot(fm, which = 2)
```




```
# residual vs fitted plot
mplot(fm, which = 1)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Additional #1 (Intro Stats Review) This problem is designed to be review - it does not require fitting any models or using inference - just basic plots and descriptive statistics. It uses the Pines data set, already loaded above for a previous problem.

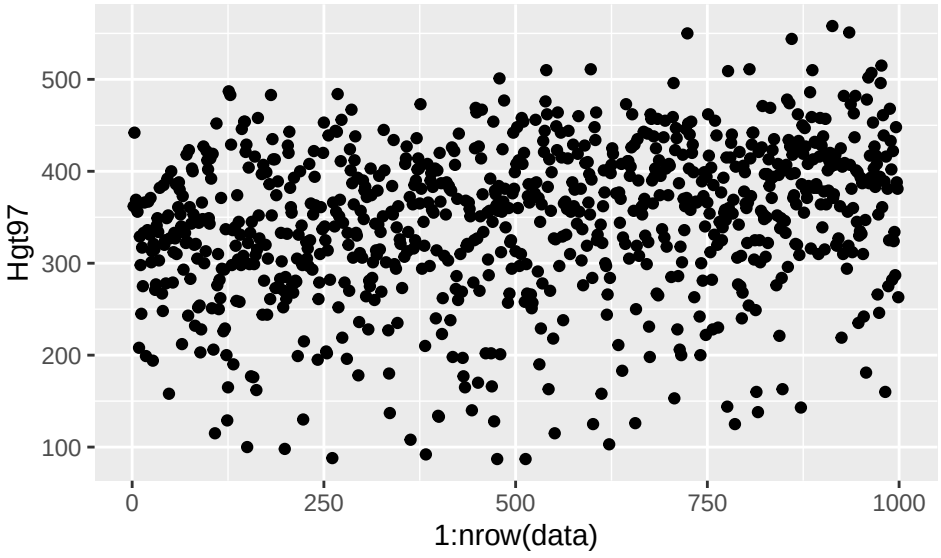
part a: Make an appropriate plot (i.e. at least one) and use appropriate numerical summaries to help you describe the distribution of the variable Hgt97 in a few sentences.

SOLUTION: In the plot we can see frequency of different values of the variable Hgt97 in the dataset. We can say that it is centered at around 350 and is skewed towards left. We can see two unusually high peaks

signifying high frequency at values near 325 adn 410 for Hgt97. Now, we can look at the numerical summaries generated below. We can say that Hgt97 lies between 97 to 558 over the entire dataset. Moreover, our data is centered with 357 as median of Hgt97 values. Our interquartile range for the Hgt97 is from 301 to 406, which is 105 wide.

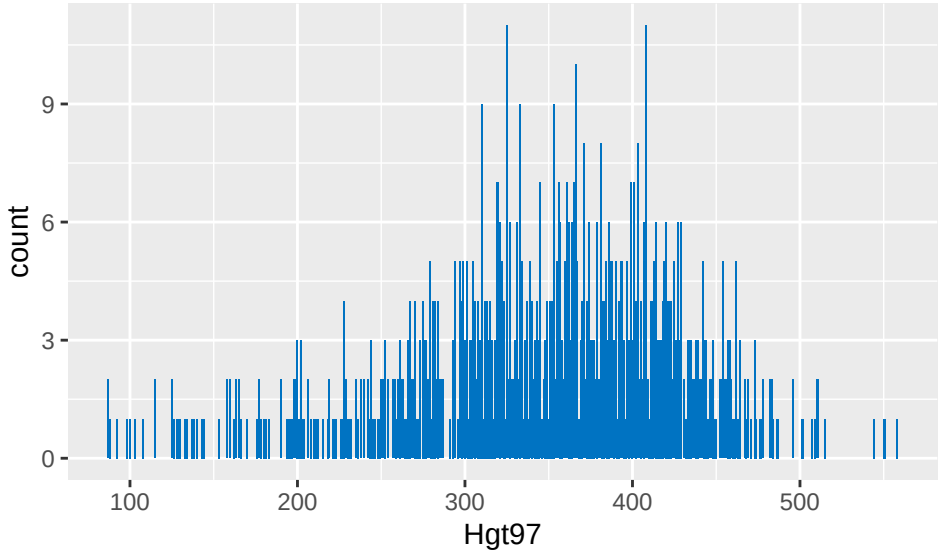
```
data <- Pines
ggplot(data, aes(x = 1:nrow(data), y = Hgt97)) +  # Apply nrow function
  geom_point()
```

```
## Warning: Removed 135 rows containing missing values (geom_point).
```



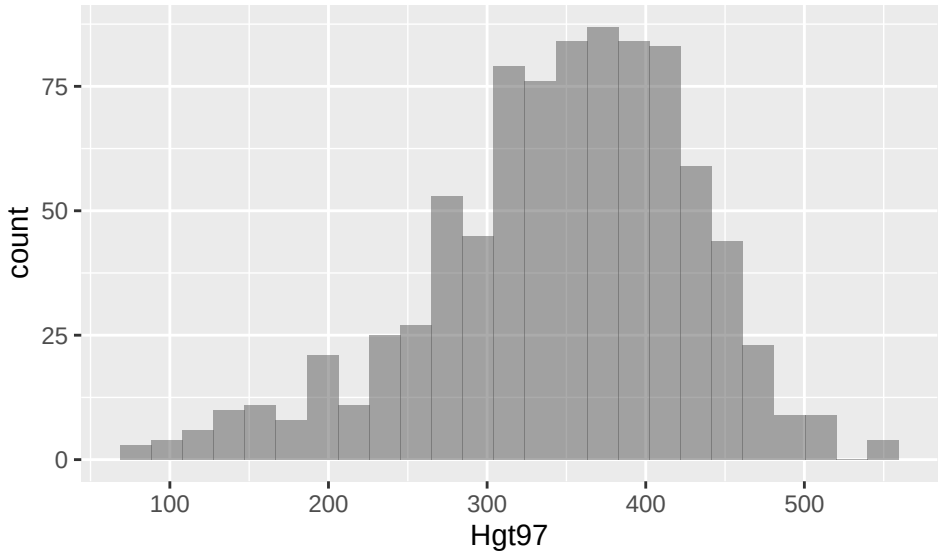

```
ggplot(data, aes(Hgt97)) +
  geom_bar(fill = "#0073C2FF")
```

```
## Warning: Removed 135 rows containing non-finite values (stat_count).
```



```
gf_histogram(~Hgt97, bins = 25, data = Pines)
```

```
## Warning: Removed 135 rows containing non-finite values (stat_bin).
```




```
favstats(~Hgt97, data=data)
```

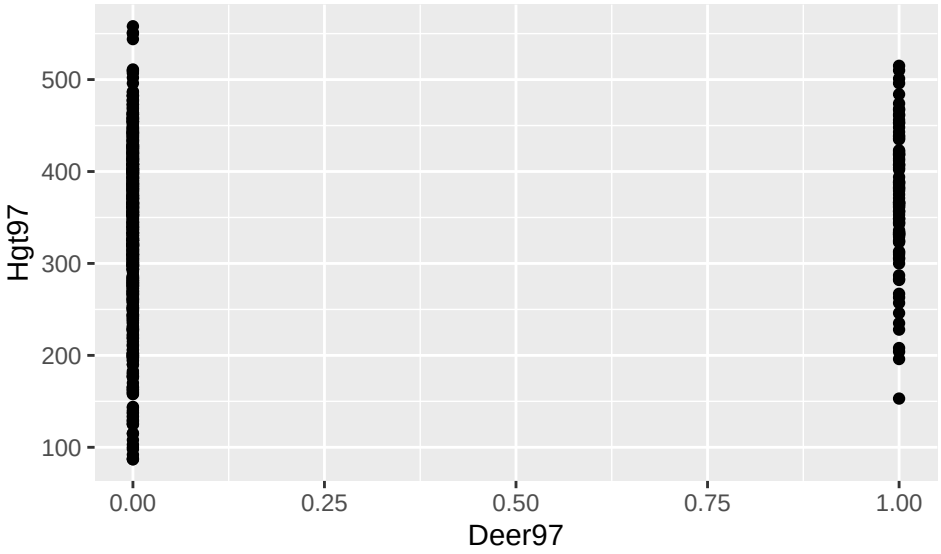
```
##  min  Q1 median  Q3 max    mean      sd   n missing
##   87 301    357 406 558 346.618 82.5964 865     135
```

part b: Make an appropriate plot (i.e. at least one) and use appropriate numerical summaries to help you describe the distribution of Hgt97 by Deer97 in a few sentences. In particular, address the question: does the distribution of Hgt97 appear to be the same for the two values of Deer97?

SOLUTION: From the first graph we can see that the distribution of Hgt97 is different in the case of Deer97 equals to 1 and 0. When the Deer907 is 0, we can see that the data is distributed over the range of 87 to 558, but when the Deer97 is 1, Hgt97 is distributed over a smaller range which is 153 to 515. Also, from the histograms we can say that both distributions are skewed towards the left side (or lower Hgt97 values). When the Deer97 is 0, distribution's median is 355, and the interquartile range is 299 to 403 (which is 104 wide) as we can see in the numerical summary output. On the other hand, when the Deer97 is 1, distribution's median is 366, and the interquartile range is 330.5 to 418.25 (which is 87.75 wide) as we can see in the numerical summary output. Hence, we can say that the distribution of Hgt97 differs in the above-mentioned fashion with different values of Deer97.

```
#no hypothesis test is needed - describe what you see in appropriate output
data <- Pines
ggplot(data, aes(x = Deer97, y = Hgt97)) +
  geom_point()
```

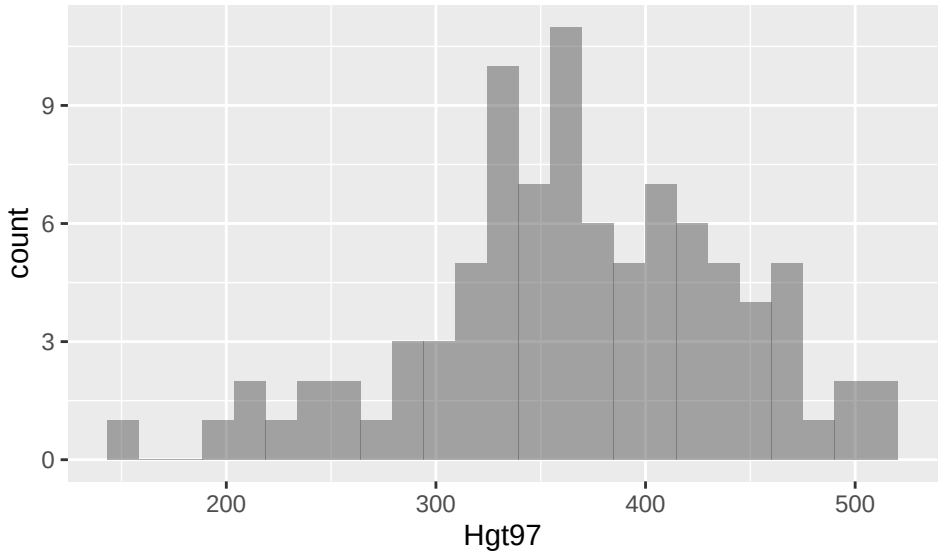
```
## Warning: Removed 137 rows containing missing values (geom_point).
```




```
data_deer97_is_1 <- data[data$Deer97 == 1, ]
#ggplot(data_deer97_is_1, aes(Hgt97)) +
#  geom_bar(fill = "#0073C2FF")

gf_histogram(~Hgt97, bins = 25, data = data_deer97_is_1)
```

```
## Warning: Removed 135 rows containing non-finite values (stat_bin).
```



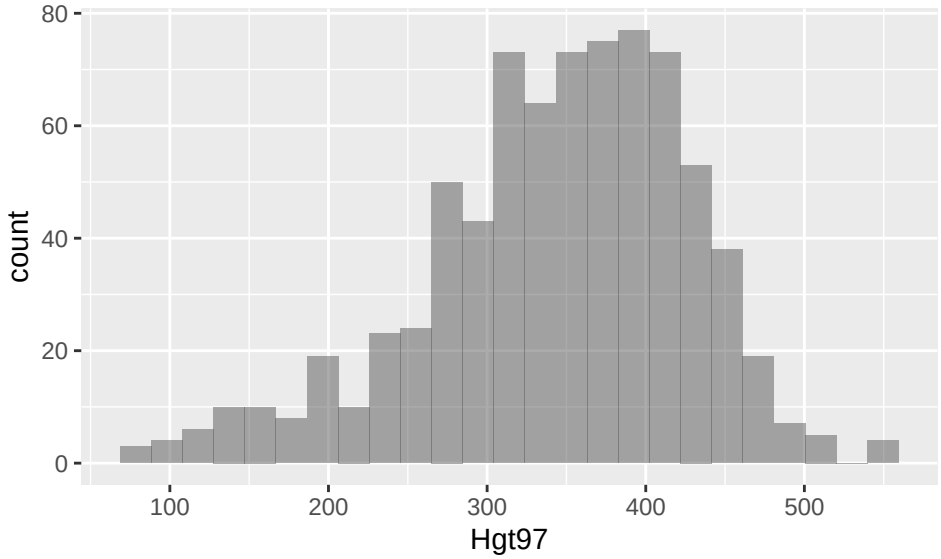

```
favstats(~Hgt97, data=data_deer97_is_1)
```

```
##  min    Q1 median     Q3 max    mean      sd  n missing
##  153 330.5    366 418.25 515 367.011 74.0799 92     135
```

```r
data_deer97_is_0 <- data[data$Deer97 == 0, ]
#ggplot(data_deer97_is_0, aes(Hgt97)) +
#  geom_bar(fill = "#0073C2FF")

gf_histogram(~Hgt97, bins = 25, data = data_deer97_is_0)
```

```
## Warning: Removed 137 rows containing non-finite values (stat_bin).
```




```r
favstats(~Hgt97, data=data_deer97_is_0)
```

```
##  min  Q1 median  Q3 max    mean      sd   n missing
##   87 299    355 403 558 343.933 83.1459 771     137
```