Lab Assignment 3 BIS 505b

Wenxin Xu

3/21/2021

Contents

Instructions

This Lab Assignment uses the data from the study conducted to investigate the impacts of herbicide exposure on maternal health described in **Lab Assignment 0**, `hgb.csv`. Instead of comparing the hemoglobin change between groups of pregnant women exposed and unexposed to herbicides through tap water, we would like to investigate the relationship between hemoglobin change (g/dL) [`change`] and amount of tap water consumed (L) [`water`] in women who consumed some or all tap water during their pregnancies (groups 1 and 3). In this assignment, report any p-values that are less than 0.0001 as **<0.0001** and round values reported in your narrative text to **3** decimal places. **Perform all plotting using `ggplot()`.**

Assignment

1. [5 points] Import the CSV file `hgb.csv` in the third code chunk above. Name your data frame `hgb` and re-create the variable `change` that you created in **Lab Assignment 0**. Since 1-liter changes in water consumed are very small changes, let's instead consider 100-liter changes. To do this, we will re-scale `water` and create the variable `water100` that is equal to `water/100` for use in this Lab Assignment. `water100` measures the amount of tap water consumed in 100s of liters. Finally, create a subset of `hgb` called `hgb13` that includes women who consumed some or all tap water during their pregnancies (groups 1 and 3). You will work with this data frame in this Lab Assignment. [Note: No written response is required for this question. Display the code chunk(s) that perform the requested data management steps for this question.]

Variable Name	Definition
`change`	Change in hemoglobin between Week 9 and Week 36 (g/dL) (negative if hemoglobin decreases)
`water100`	Amount of tap water consumed (in 100s of liters)

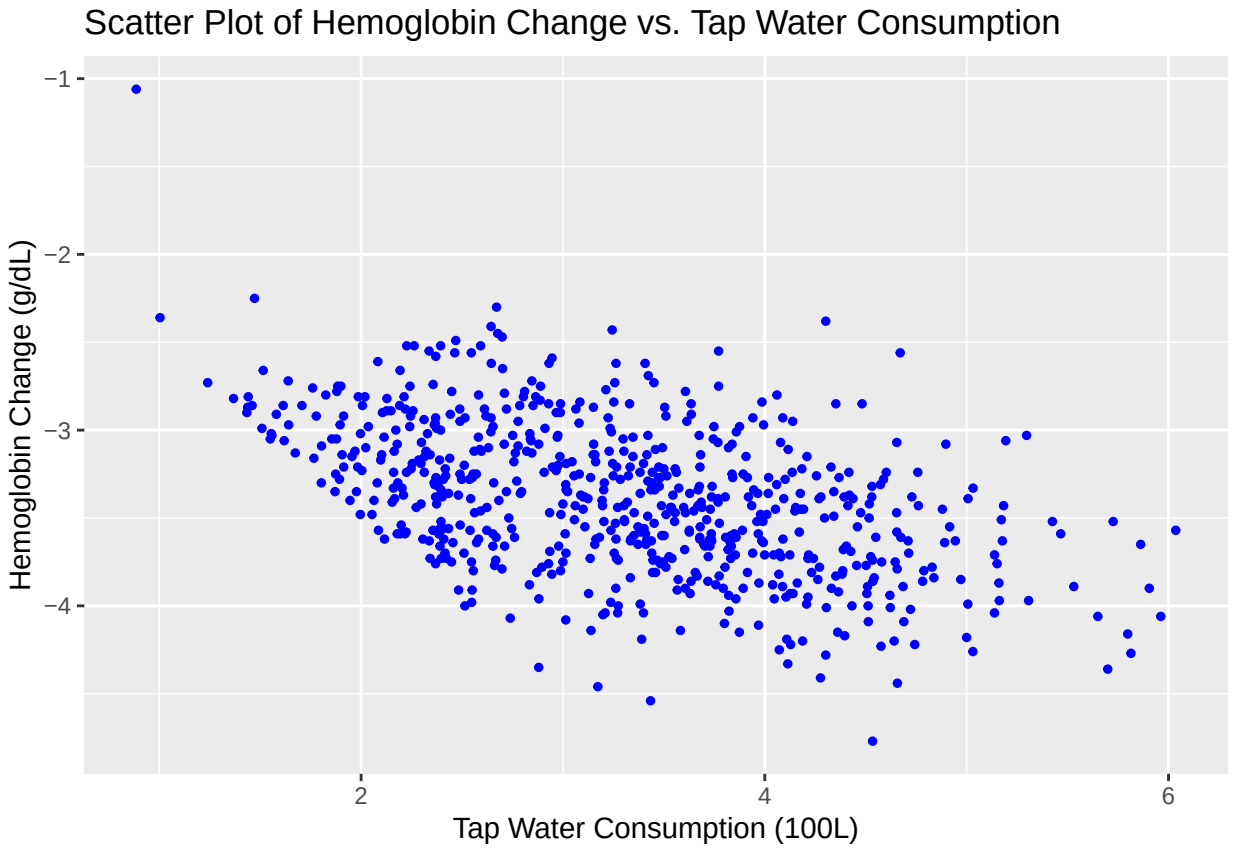
```
hgb$change <- hgb$hgb36 - hgb$hgb9

hgb$water100 <- hgb$water/100
```

```
hgb13 <- subset(hgb, group==1|group==3)
```

2. The **research question** is: Is exposure to herbicides in drinking water (measured by tap water consumption) associated with hemoglobin change during pregnancy?

a. [10 points] Create a scatter plot of the association between these two quantitative variables. Include axis labels and a title. Is it clear which variable is the independent (x) variable and which is the dependent (y) variable? If so, create your scatter plot to reflect this. In words, describe the relationship that you see between these two variables in the context of the problem. Do you see any "outlying values" (i.e., values that are not clustered with the rest of the points)?

```
ggplot(data=hgb13, aes(x=water100, y=change ))+
  geom_point(size=1,shape=19,col="blue")+
  labs(title="Scatter Plot of Hemoglobin Change vs. Tap Water Consumption",
       x="Tap Water Consumption (100L)", y="Hemoglobin Change (g/dL)")
```



Answer: Tap Water Consumption is the independent variable and Hemoglobin Change is the dependent variable. The scatter plot shows a modest negative linear relationship between them. I see 1 outlying values on the top left corner.

b. [10 points] Fit a simple linear regression model with hemoglobin change as the dependent variable (y) and tap water consumption as the independent variable (x). From this model, report the fitted least squares linear regression line and interpret the estimated slope. Provide a 95% confidence interval for the slope parameter.

```r
# Simple linear regression model: modeling hemoglobin change using tap water consumption
reg <- lm(change ~ water100, data=hgb13)

# output results of fitted model
summary(reg)
```

```
##
## Call:
## lm(formula = change ~ water100, data = hgb13)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12293 -0.24692 -0.02443  0.23843  1.78627
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -2.64779    0.05251  -50.43 <0.0000000000000002
## water100    -0.22402    0.01526  -14.68 <0.0000000000000002
##
## Residual standard error: 0.3702 on 662 degrees of freedom
## Multiple R-squared:  0.2457, Adjusted R-squared:  0.2445
## F-statistic: 215.6 on 1 and 662 DF,  p-value: < 0.00000000000000022
```

```r
# 95% CI for model params
confint(reg, level=0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept) -2.7508880 -2.5446909
## water100    -0.2539752 -0.1940624
```

1) The fitted least squares linear regression line is $\hat{y}$ = -2.648 -0.224 x.

2) The estimated slope is equal to -0.224 [95% CI (-0.254, -0.194)] indicates that a 1-unit increase in tap water consumption is associated with a 0.224-unit average decrease in hemoglobin change.

c. [10 points] Based on your regression model output from part **(b)**, perform the hypothesis test at the $\alpha = 0.05$-level to determine if there is a linear relationship between hemoglobin change and amount of tap water consumed. (i) State the null and alternative hypotheses of this test. (ii) From your **R** output, report the value of the test statistic and p-value. (iii) State your statistical conclusion and your conclusion in the context of the problem.

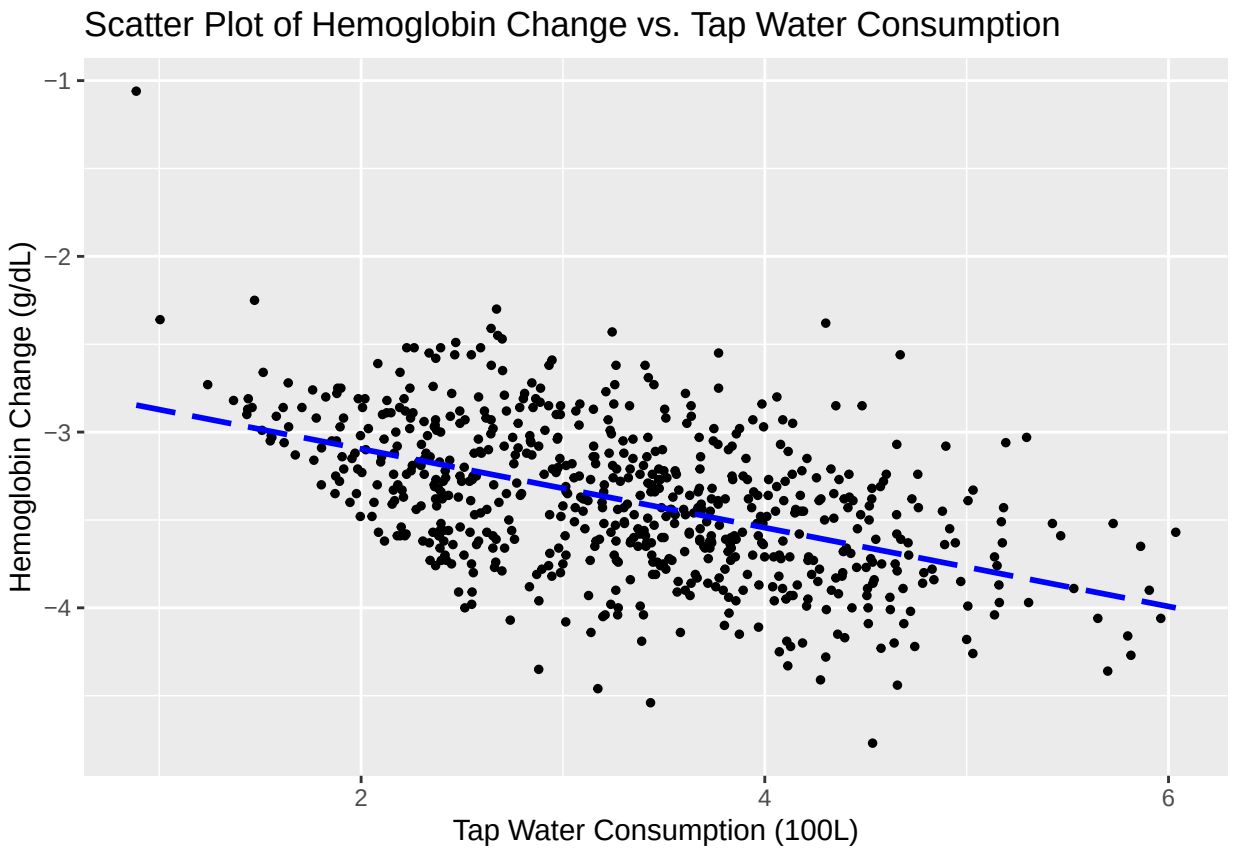
(i) $H_0 : \beta = 0$ $vs.H_1 : \beta \neq 0$

(ii) The t-statistic is t =-14.684, p-value is <.0001.

(iii) Because p-value is less than 0.0001, we have evidence to reject H_0 and conclude there is a significant linear relationship between hemoglobin change (g/dL) and amount of tap water consumed (L) in women who consumed some or all tap water during their pregnancies (groups 1 and 3) at the (α) = 0.05-level of significance.

d. [7 points] Re-create the scatter plot from part **(a)**, adding the fitted regression line. Comment on how well this line appears to fit the data. Is the outlying value estimated well by this regression line? Will this observation have a positive or negative residual?

```
ggplot(data=hgb13, aes(x=water100, y=change ))+
  geom_point(size=1,shape=19)+
  geom_smooth(method = "lm", formula = y ~ x,
se = FALSE, col = "blue", linetype = "longdash") + # add regression line (no CI)
  labs(title="Scatter Plot of Hemoglobin Change vs. Tap Water Consumption",
       x="Tap Water Consumption (100L)", y="Hemoglobin Change (g/dL)")
```



This line fits the data well. The outlying value is not estimated well by this regression line. This observation has a positive residual.

e. [8 points] Report and interpret the Coefficient of Determination for the fitted model. Report the residual standard error of the model; in words, what does this value represent?

The Coefficient of Determination is rather low at 0.246, indicating that tap water consumption only explains 24.6% of the total variability in hemoglobin change. The residual standard error $s_{y|x}$ is 0.37, this value represent the estimated variability in Y ($\sigma_{y|x}$) about the regression line is expected to hold for all values of X (i.e., the constant variance assumption of linear regression).

f. [10 points] Use **R** to report the expected change in hemoglobin and the 95% confidence interval for the mean change in hemoglobin when 200 L of tap water are consumed. Do the same assuming 400 L of tap water are consumed.

```
# value of x (tap water consumption) used to estimate y (hemoglobin change)
x.star <- data.frame(water100=c(200,400))     c(2, 4)

# fitted value and lower and upper CI for mean at values of x in x.star
predict(reg, newdata=x.star, interval="confidence", level=0.95)
```

these estimates are a bit wrong
*because one unit of water consumption here is *100 L*.*
therefore for our new data frame we should be making c(2,4) not c(200,400)

```
##         fit        lwr       upr
## 1 -47.45155  -53.34373 -41.55936
## 2 -92.25531 -104.13873 -80.37188
```

The fitted line estimates a mean hemoglobin change of -47.452 in those with tap water consumption = 200 [95% CI (-53.344, -41.559)]. The estimated mean hemoglobin change in those with tap water consumption = 400 is equal to -92.255 [95% CI (-104.139, -80.372)].

g. [10 points] Create a scatter plot of residuals vs. fitted values. Does the constant variance assumption appear to hold? Why or why not? Does the outlying value identified in part **(a)** stand out in this scatter plot? Report the subject ID number (`id`) and residual e_i for this outlying value.

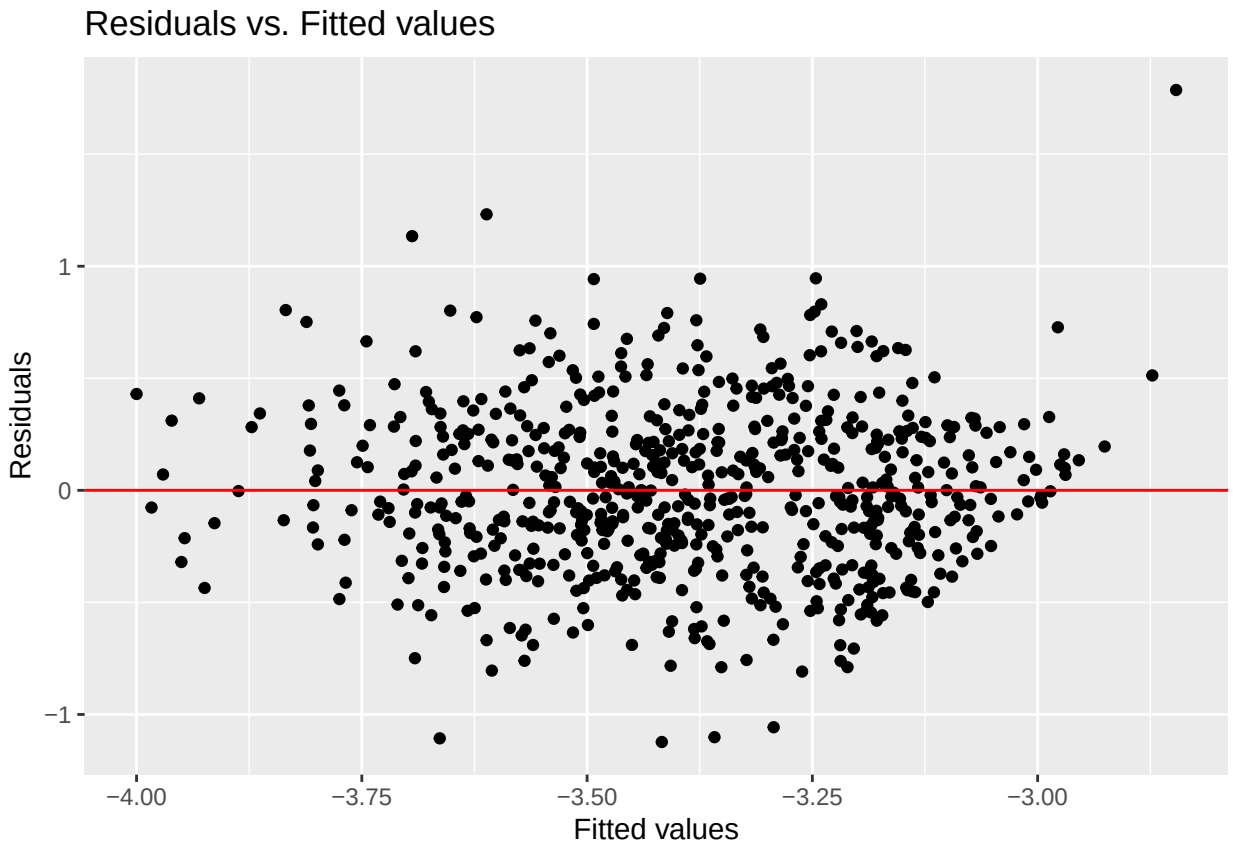
```
# append column of predicted values to dataset
hgb13$predicted <- predict(reg)

# append column of residuals to dataset
hgb13$residuals <- resid(reg)

# summary statistics of residuals
summary(hgb13$residuals)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.12293 -0.24692 -0.02443  0.00000  0.23843  1.78627
```

```
# sort data frame by largest |residuals|
hgb13sorted <- hgb13[order(-abs(hgb13$residuals)),]

head(hgb13sorted)
```

```
##      id group age edyrs income   wt0   wt1 parity prenatal psmoke  hgb9 hgb36
## 979 979     3  25    12     NA 133.5 178.8      1        1      0 10.54  9.48
## 161 161     1  24    12  3.661 212.4 250.8      1        1      0  9.65  7.27
## 104 104     1  21     8  3.083 161.6 192.6      1        0      0 11.67  9.11
## 688 688     3  26    11  2.699 193.5 237.9      3        0      1 11.90  7.36
## 83   83     1  22    10  2.840 170.9 205.4      3        0      1 11.35  6.58
## 685 685     3  25    10  2.684 128.4 174.9      2        0      1 10.84  6.38
##     water change water100 predicted residuals
## 979  88.6  -1.06    0.886 -2.846270  1.786270
## 161 430.2  -2.38    4.302 -3.611518  1.231518
## 104 467.1  -2.56    4.671 -3.694181  1.134181
## 688 343.4  -4.54    3.434 -3.417070 -1.122930
## 83  453.4  -4.77    4.534 -3.663491 -1.106509
## 685 317.3  -4.46    3.173 -3.358601 -1.101399
```

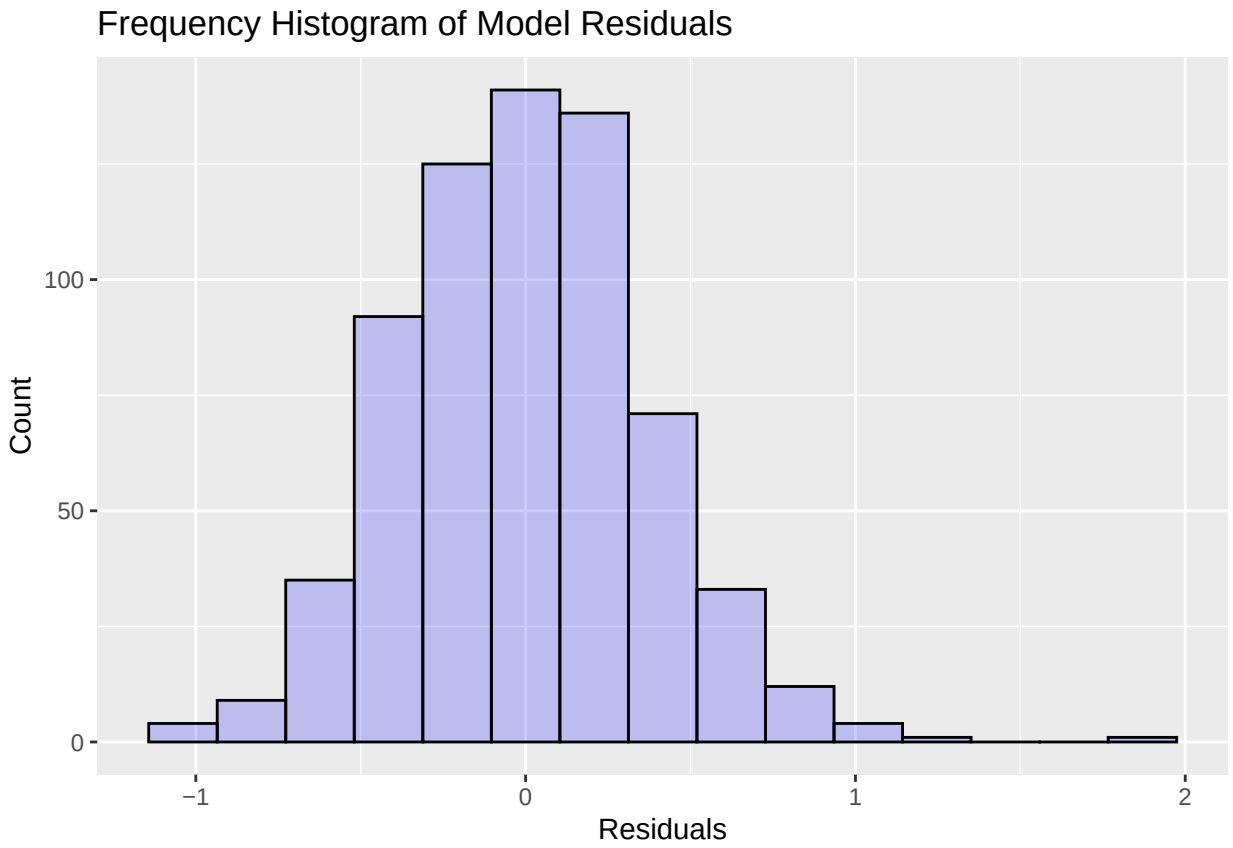
```
# scatter plot of residuals vs. fitted values
ggplot(data=hgb13,aes(x=predicted,y=residuals))+
  geom_point()+
  geom_hline(yintercept=0,col="red")+
  labs(title="Residuals vs. Fitted values",
       x="Fitted values",y="Residuals")
```



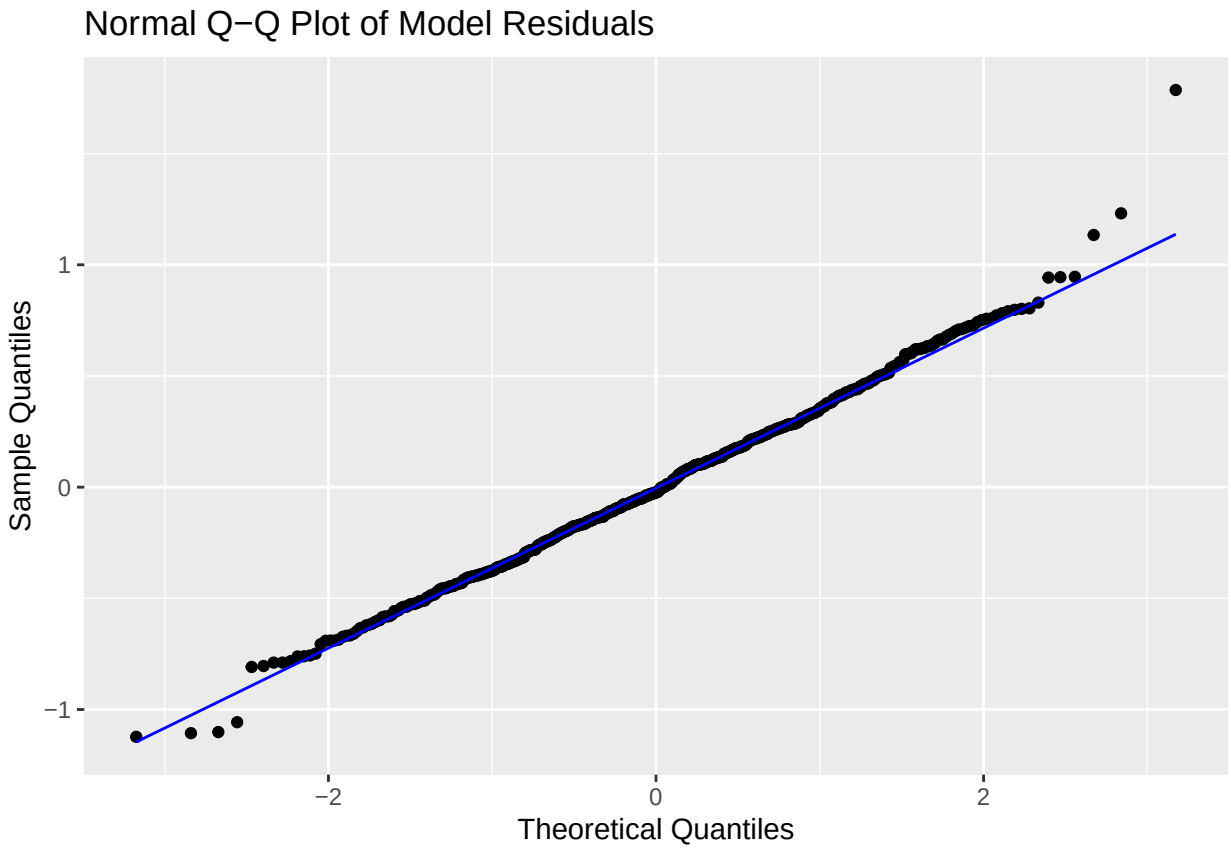
The constant variance assumption appears to hold because the plot of residuals vs. fitted values above does appear random. The outlying value identified in part **(a)** stand out in this scatter plot since the residual is far from the main cluster of residuals.The outlying value's subject ID number is 979 and residual e is 1.786.

h. [10 points] Create a histogram and Normal Q-Q plot of residuals. What linear model assumption is being evaluated by these plots? Does this assumption appear to be violated? Does the outlying value identified in part **(a)** stand out in these plots?

```
# histogram of residuals
ggplot(data=hgb13,aes(x=residuals))+
  geom_histogram(bins=15,
                 col="black",
                 fill="blue",
                 alpha=0.2,
                 closed="left",
                 na.rm=TRUE)+
  labs(title="Frequency Histogram of Model Residuals",
       x="Residuals",y="Count")
```




```
# Normal Q-Q plot of residuals
ggplot(data=hgb13,aes(sample=residuals))+
  geom_qq()+
  geom_qq_line(col="blue")+
  labs(title="Normal Q-Q Plot of Model Residuals",
       x="Theoretical Quantiles",y="Sample Quantiles")
```



Normality of residuals is being evaluated by these plots. These plots do indicate right skewed residuals. The assumption this appear to be violated. Subject 979's large residual does stand out in these plots.

i. [15 points] (i) Remove the observation identified in part **(g)** from the `hgb13` data frame and re-fit the model. Report the re-fitted regression line. Are the intercept and slope close to those from the original model fit in part **(b)**? (ii) Next, highlight the outlying value that was removed from the analysis in a scatter plot of hemoglobin change vs. tap water consumption. Include the original fitted regression line from **(b)** as a blue line and the re-fitted regression line as a red line in that scatter plot. Has the fitted line changed greatly after removing the outlying value (that is, would you consider the outlier to be an "influential" point)? (iii) Finally, create a scatter plot of residuals vs. fitted values, histogram of residuals and Normal Q-Q plot of residuals for the re-fit model and comment on whether these plots have changed/improved in any way after removing the outlying observation.

(i)

```
# filter hgb13 to remove id==979
sensitivity_hgb13 <- subset(hgb13, id !=hgb13sorted$id[1])

# sensitivity analysis
reg.sens <- lm(change ~ water100, data=sensitivity_hgb13)

summary(reg.sens)
```

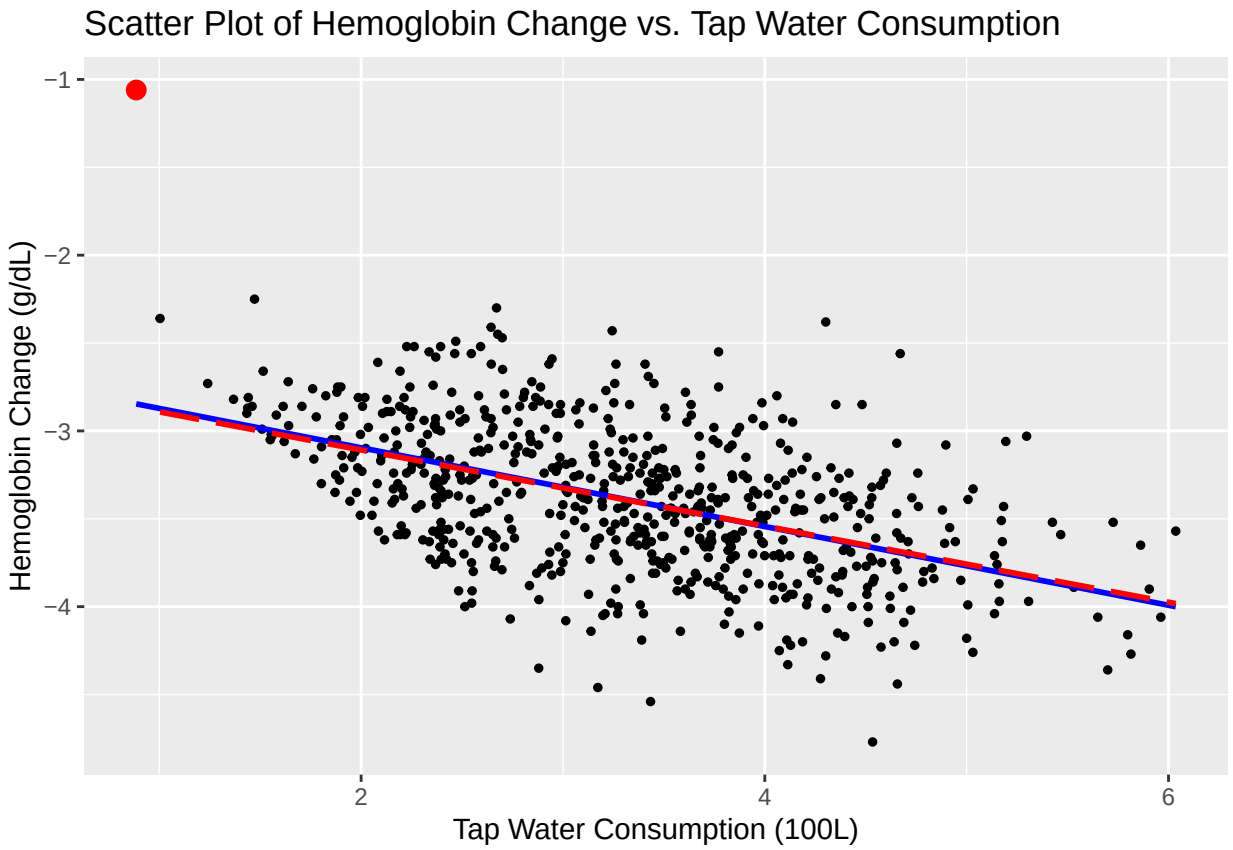
```
##
## Call:
## lm(formula = change ~ water100, data = sensitivity_hgb13)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12113 -0.24227 -0.01738  0.24280  1.22686
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -2.67514    0.05190  -51.55 <0.0000000000000002
## water100    -0.21658    0.01507  -14.37 <0.0000000000000002
##
## Residual standard error: 0.3638 on 661 degrees of freedom
## Multiple R-squared:  0.2381, Adjusted R-squared:  0.2369
## F-statistic: 206.6 on 1 and 661 DF,  p-value: < 0.00000000000000022
```

The refitted regression line is $\hat{y}$ = -2.675 -0.217 x. The intercept (-2.675 [95% CI (-2.777, -2.573)]) and slope (-0.217 [95% CI (-0.246, -0.187)]) are close to intercept (-2.648 [95% CI (-2.751, -2.545)]) and slope (-0.224 [95% CI (-0.254, -0.194)]) from the original model fit in part **(b)**.

(ii)

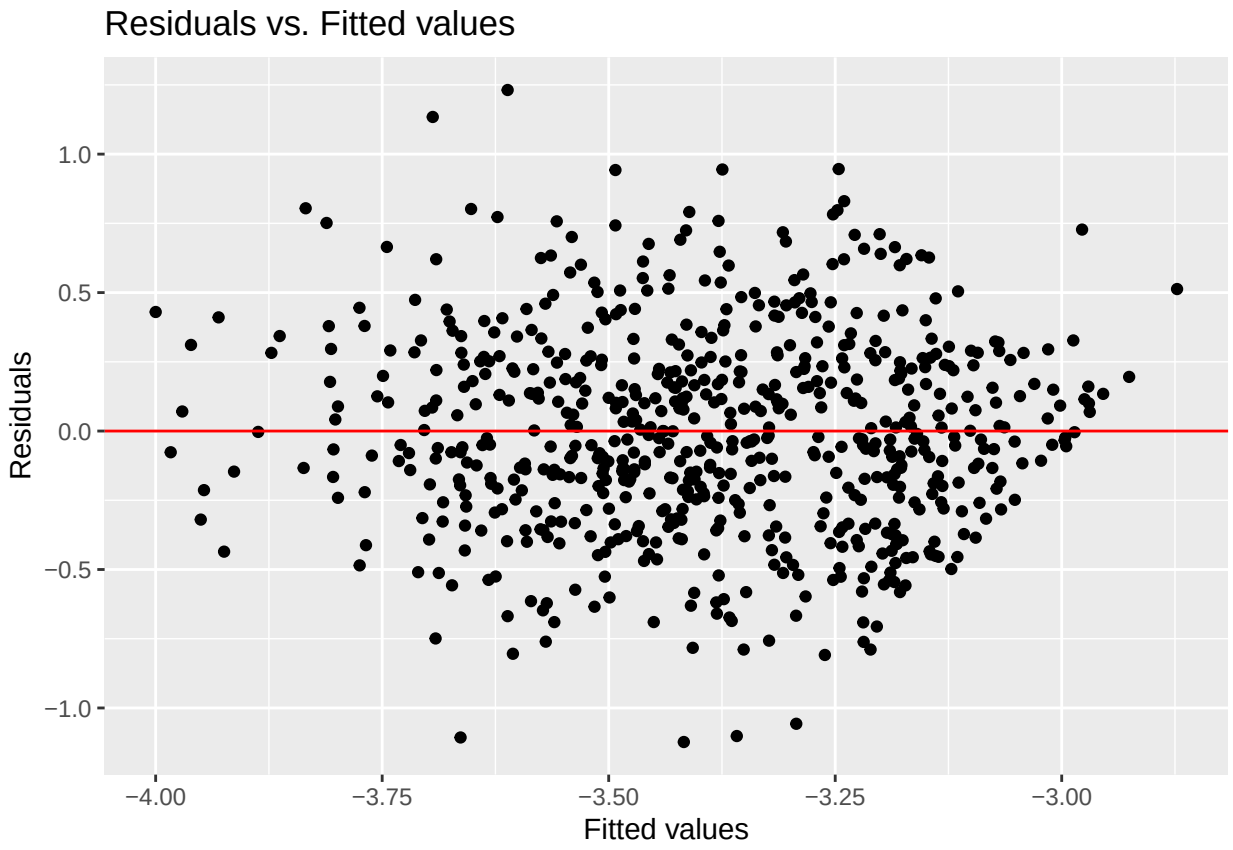
```
highlight_hgb13 <- subset(hgb13, id==hgb13sorted$id[1])

ggplot()+
  geom_point(data=hgb13, aes(x=water100, y=change ), size=1,shape=19)+
  geom_smooth(data=hgb13, aes(x=water100, y=change ),
              method = "lm", formula = y ~ x, se = FALSE, col = "blue") + # add regression line as blue
    geom_smooth(data=sensitivity_hgb13, aes(x=water100, y=change ),
              method = "lm", formula = y ~ x, se = FALSE, col = "red", linetype = "longdash") + # add r
  labs(title="Scatter Plot of Hemoglobin Change vs. Tap Water Consumption",
       x="Tap Water Consumption (100L)", y="Hemoglobin Change (g/dL)")+
  geom_point(data=highlight_hgb13, aes(x=water100, y=change),
             col="red",size=3)
```



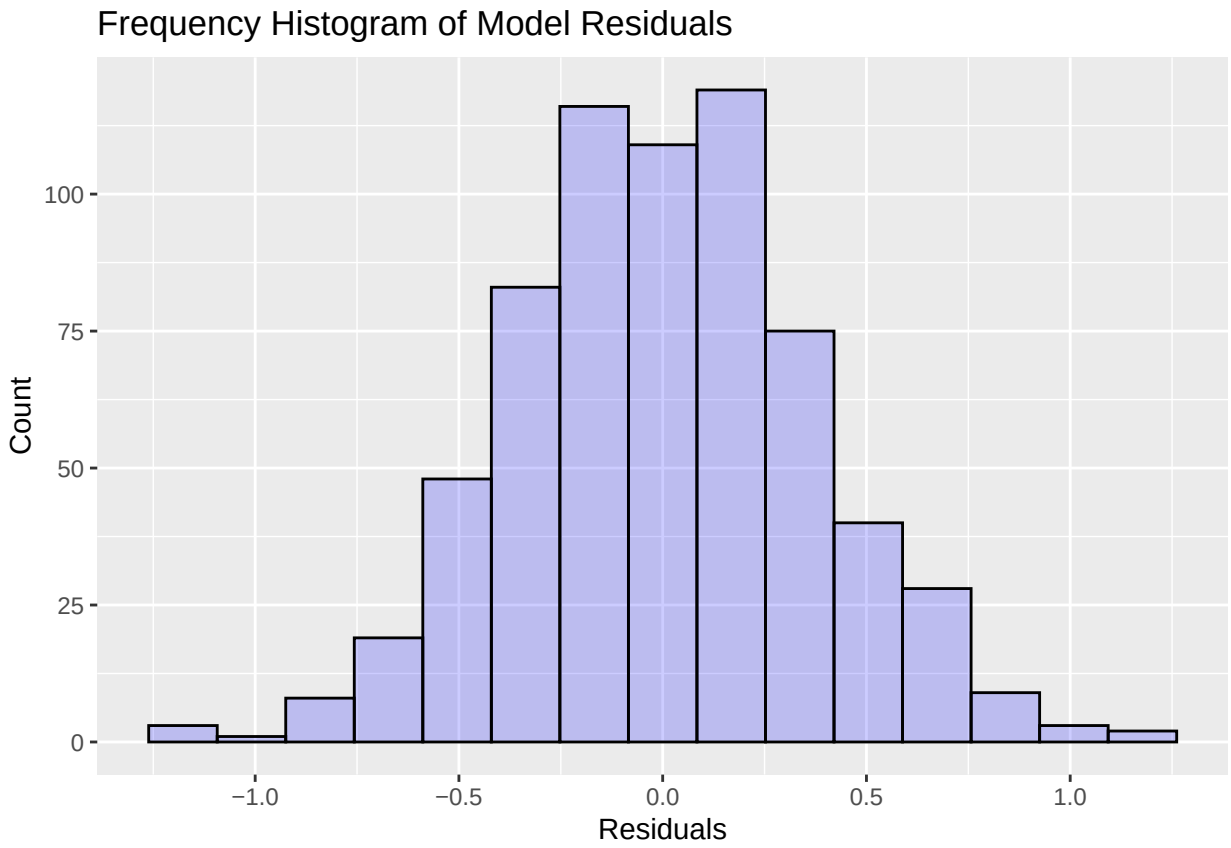
The fitted line have not changed greatly after removing the outlying value , thus, I would not consider the outlier to be an "influential" point.

(iii)

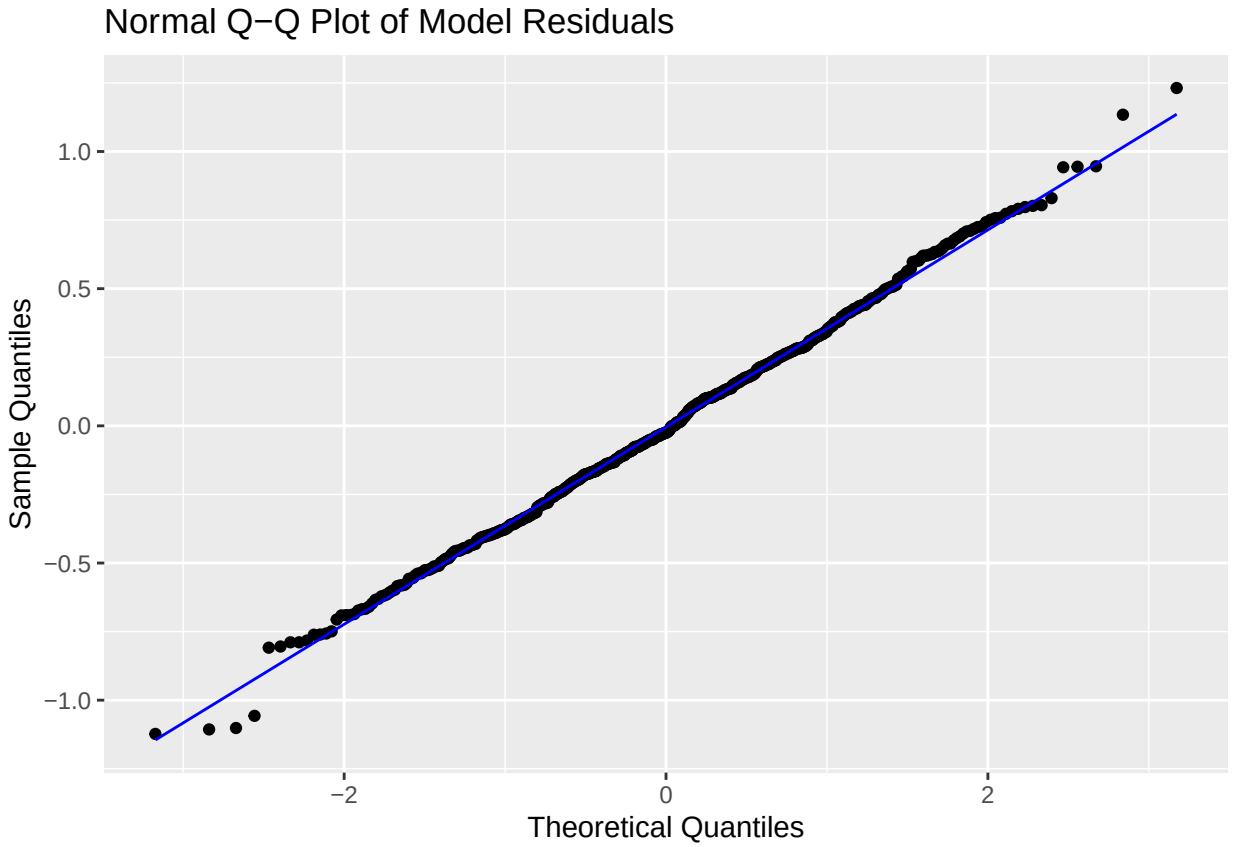
```
# scatter plot of residuals vs. fitted values for the re-fit model
ggplot(data=sensitivity_hgb13,aes(x=predicted,y=residuals))+
  geom_point()+
  geom_hline(yintercept=0,col="red")+
  labs(title="Residuals vs. Fitted values",
       x="Fitted values",y="Residuals")
```




```
# histogram of residuals for the re-fit model
ggplot(data=sensitivity_hgb13,aes(x=residuals))+
  geom_histogram(bins=15,
                 col="black",
                 fill="blue",
                 alpha=0.2,
                 closed="left",
                 na.rm=TRUE)+
  labs(title="Frequency Histogram of Model Residuals",
       x="Residuals",y="Count")
```




```
# Normal Q-Q plot of residuals for the re-fit model
ggplot(data=sensitivity_hgb13,aes(sample=residuals))+
  geom_qq()+
  geom_qq_line(col="blue")+
  labs(title="Normal Q-Q Plot of Model Residuals",
       x="Theoretical Quantiles",y="Sample Quantiles")
```



After removing the outlier, all the figures are improved. The scatter plot does appear random which indicates constant variance. The histogram becomes bell-shape, and the Q-Q plot becomes more aligned along the diagonal reference line, both 2 plots indicate the distribution of residuals is normal.

j. [5 points] Report the residual standard error from the re-fit model. Comment on how this value has changed from the original model and why the direction of the change makes sense.

The residual standard error $s_{y|x}$ is 0.364, which is smaller than the residual standard error of the original model (0.37). Since we remove one outlier value, the variability in Y ($\sigma_{y|x}$) about the regression line is expected to hold for all values of X should be smaller, thus, the new residual standard error should be smaller.