PM592: Regression Analysis for Data Science

Name:
Flemming
Wu

HW3
LINE assumptions, categorical predictors

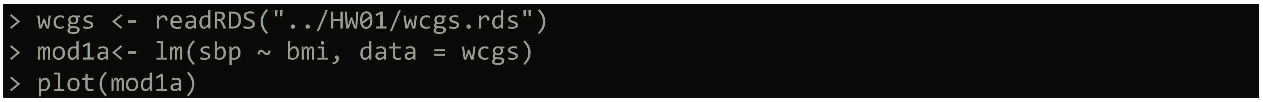
Instructions

- Answer questions directly within this document.
- Upload to Blackboard by the due date & time.
- Clearly indicate your answers to all questions.
- If a question requires analysis, attach all relevant output to this document in the appropriate area. Do not attach superfluous output.
- There are 3 questions and 30 points possible.

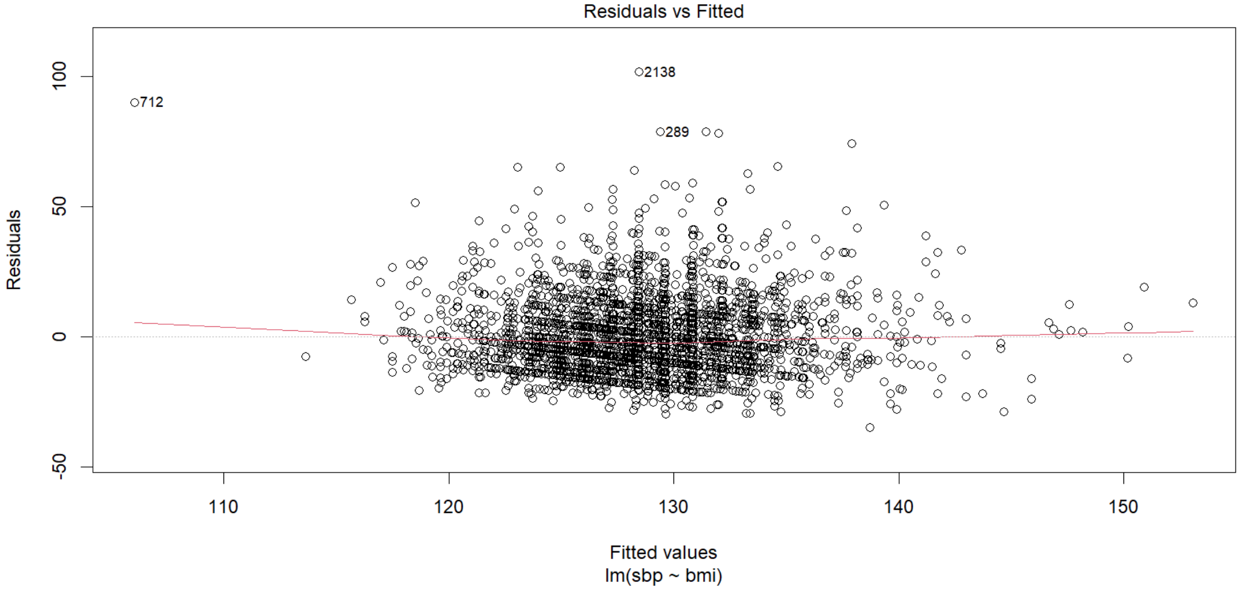
Question 1 [11 points]

Revisit your model from Homework 2, Question 1. Recall this was a regression between SBP (Y) and BMI (X).

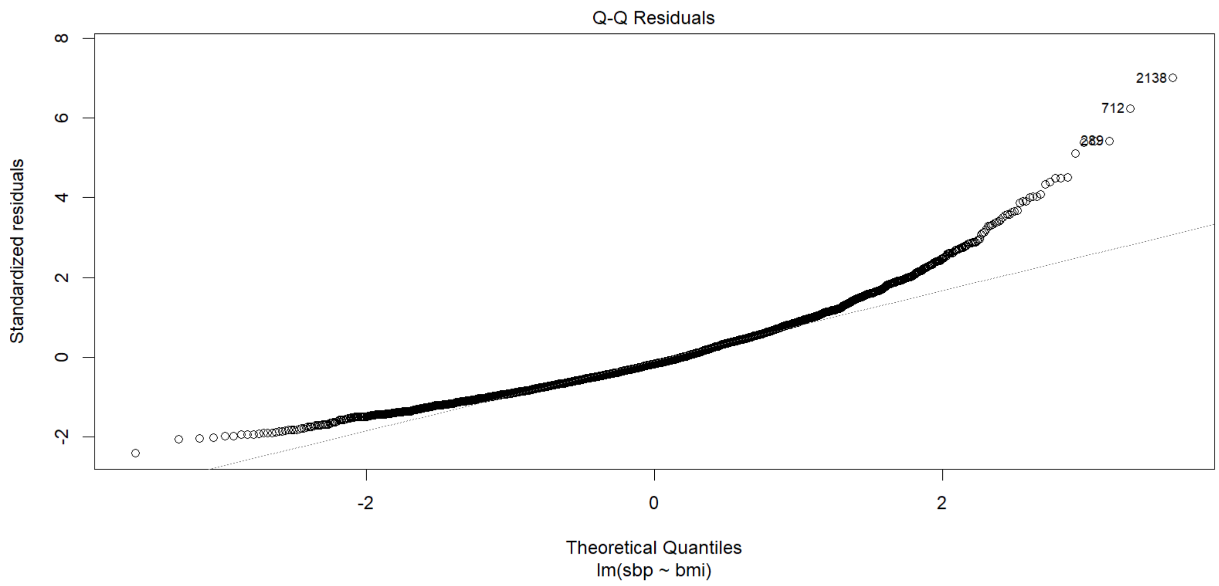
1a. [3 points]. Assess each of the assumptions of linear regression, providing relevant plots and statistics to support your conclusions.



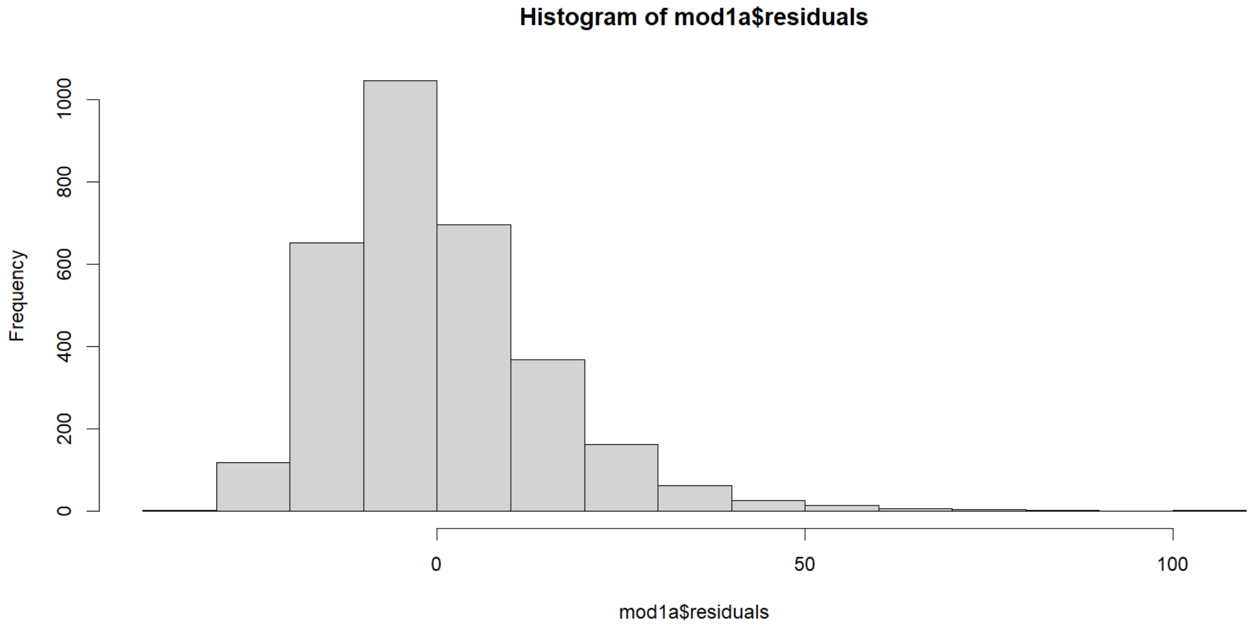
```
> wcgs <- readRDS("../HW01/wcgs.rds")
> mod1a<- lm(sbp ~ bmi, data = wcgs)
> plot(mod1a)
```



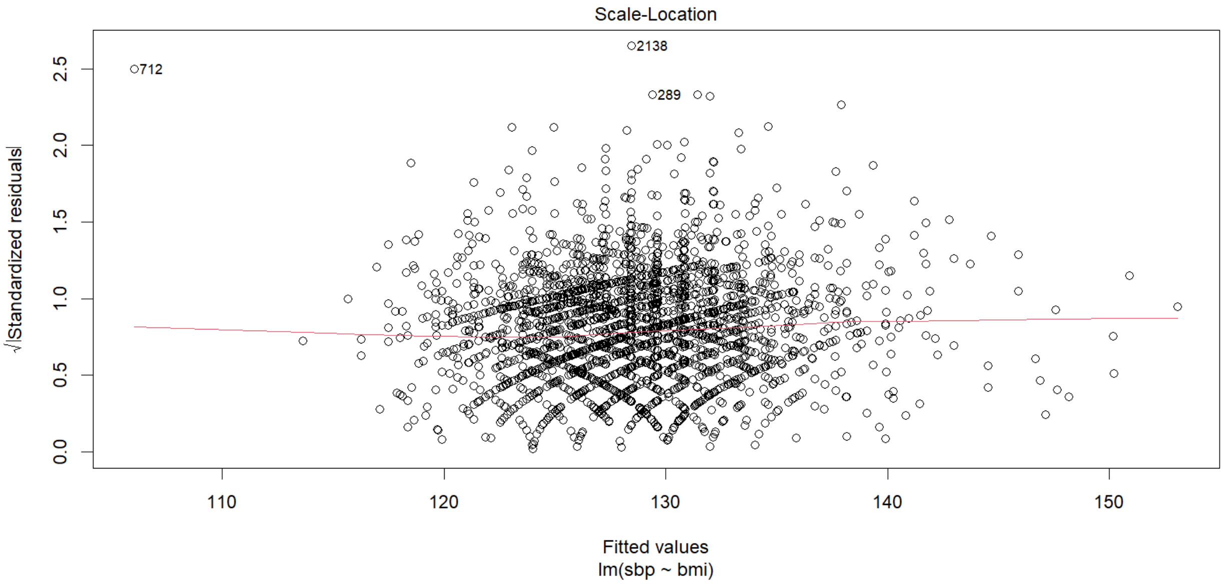
The linearity assumption can be assessed with the above plot of the fitted values versus the residuals. If the relationship between SBP and BMI is linear, there should be a random scatter of points (no pattern) around a horizontal line at 0. For the most part, it looks like the plot does fit these criteria. So I will say that the linearity assumption is met. I do notice a point (712) on the top left corner on the plot which drives the mean line up on the left.



The Normal QQ plot checks to see if the residuals follow a normal distribution. If the residuals do follow a normal distribution the points on the QQ plot should roughly follow a straight line. From the plot above it seems that the assumption of normality is violated, as there is a lot of deviation from the straight line, especially on the right side.

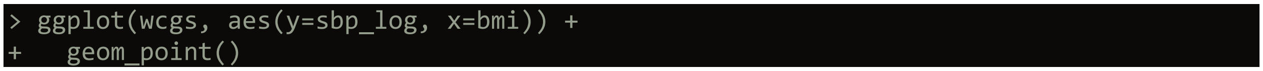


Checking this histogram of the residuals, it also shows a distribution that is slightly skewed right. However, since the number of observations in the data set are in the thousands, I can still make the normality assumption based on the Central Limit Theorem.

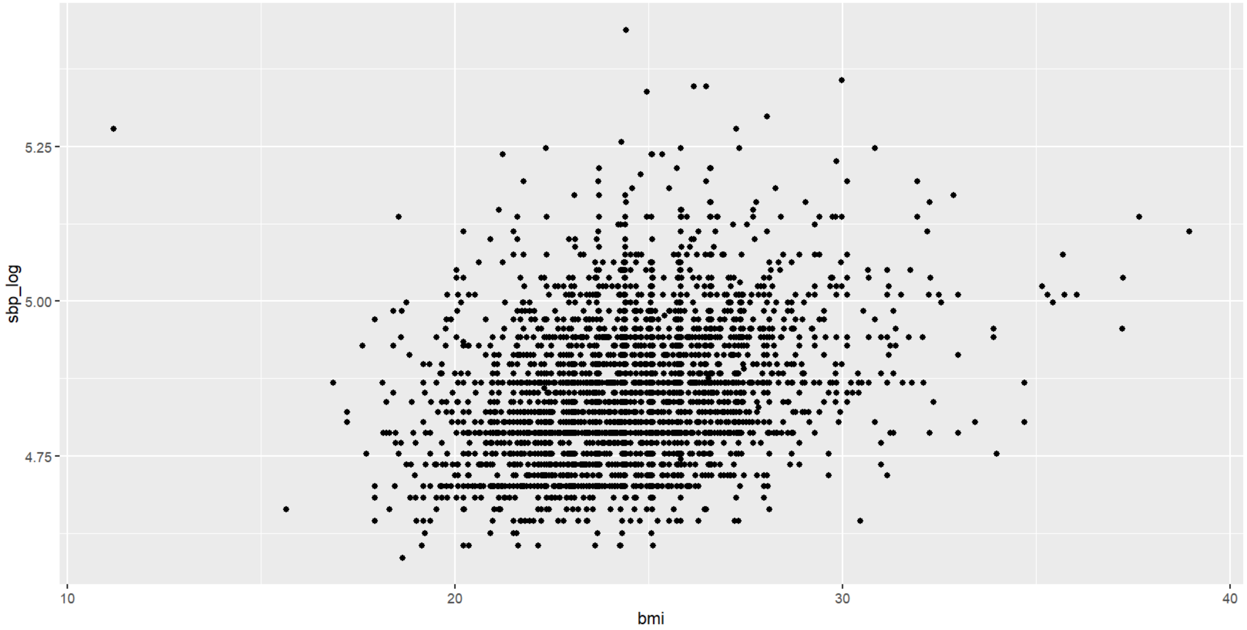


The scale-location plot shows the relationship between the square root of the standardized residuals and the fitted values, and can be used to verify the assumption of homoscedasticity (equal variances). Like the residuals vs fitted plot, there should be a random scatter of points, and indeed I do see that there is no clear pattern in the plot which signals that the assumption of homoscedasticity is met.

1b. [1 point] Provide a scatter plot of the relationship between ln(SBP) and BMI. What are your impressions?



```
> ggplot(wcgs, aes(y=sbp_log, x=bmi)) +
+   geom_point()
```



From the scatterplot there does seem to be a very slight linear relationship between BMI and the natural log of SBP. There is an outlier on the left with a very low BMI of about 11.

1c. [1 point] Provide the theoretical regression equation for the relationship of ln(SBP) regressed on mean-centered BMI. What is the null and alternative hypothesis that would test whether ln(SBP) is associated with BMI?

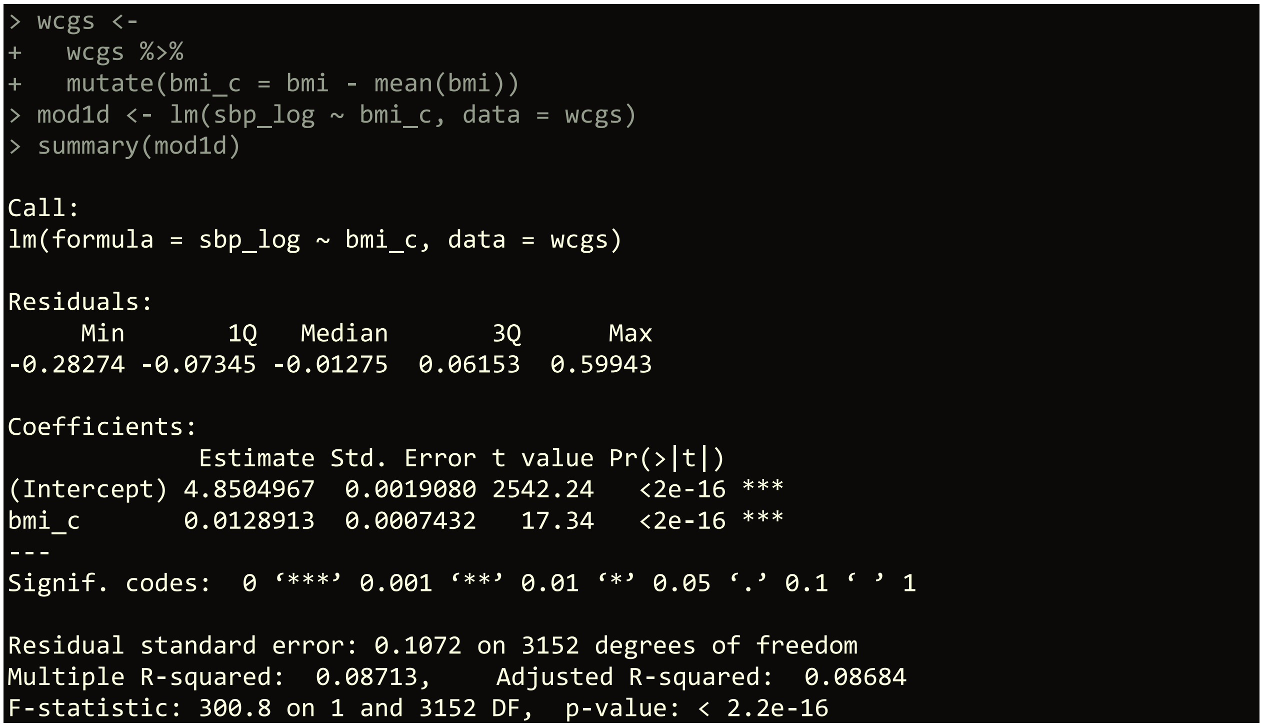
The theoretical regression equation for the relationship of ln(SBP) regressed on mean-centered BMI is:

$$\ln(\hat{Y}) = \beta_0 + \beta_{\mathrm{BMI\_c}}\mathrm{X}$$

Null hypothesis: The natural log of SBP is not associated with BMI $(\beta_{\mathrm{BMI\_c}} = 0)$

Alternative hypothesis: The natural log of SBP is associated with BMI $(\beta_{\mathrm{BMI\_c}} \neq 0)$

1d. [1 point] Provide the best-fit linear regression equation for the relationship of ln(SBP) regressed on mean-centered BMI.



```
> wcgs <-
+   wcgs %>%
+   mutate(bmi_c = bmi - mean(bmi))
> mod1d <- lm(sbp_log ~ bmi_c, data = wcgs)
> summary(mod1d)

Call:
lm(formula = sbp_log ~ bmi_c, data = wcgs)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.28274 -0.07345 -0.01275  0.06153  0.59943 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.8504967  0.0019080 2542.24   <2e-16 ***
bmi_c       0.0128913  0.0007432   17.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1072 on 3152 degrees of freedom
Multiple R-squared:  0.08713,	Adjusted R-squared:  0.08684 
F-statistic: 300.8 on 1 and 3152 DF,  p-value: < 2.2e-16
```

The best-fit linear regression equation for the relationship of ln(SBP) regressed on mean-centered BMI is:

$$\ln(\hat{Y}) = 4.85 + 0.013X$$

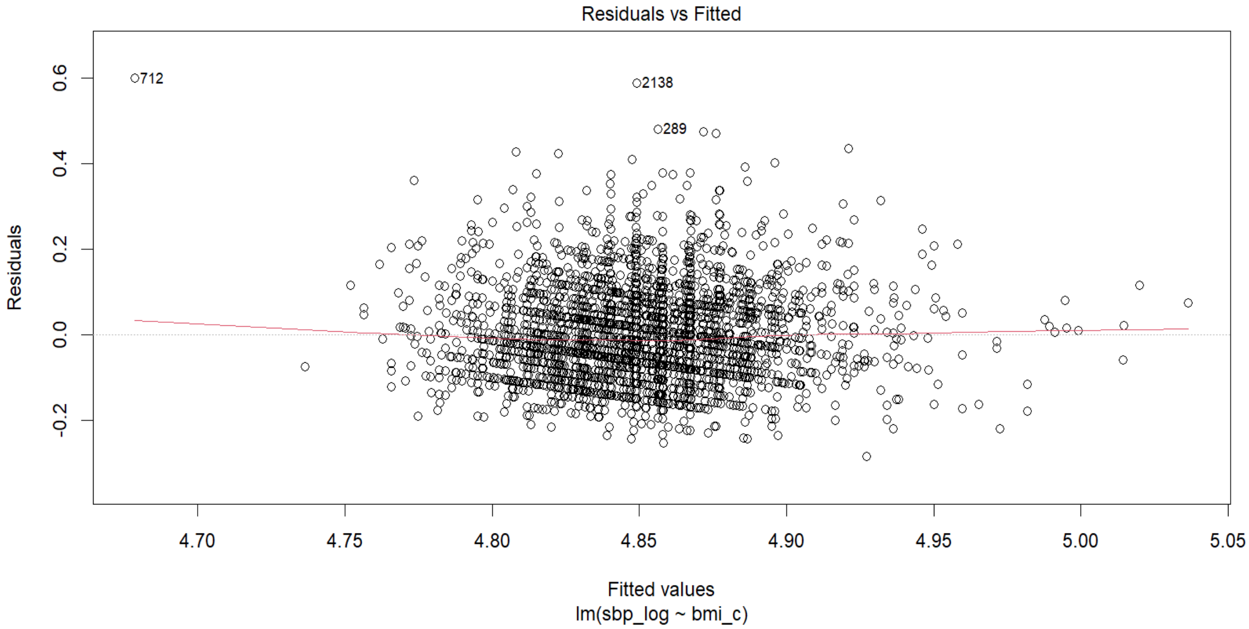
1e. [1 point] Provide the test statistic and p-value that test your hypotheses in (1c).

The test statistic and p-value for testing my hypothesis are 17.34 and p < 2e-16, respectively. These statistics indicate that there is a relationship (p << 0.001) between BMI and the natural log of SBP.

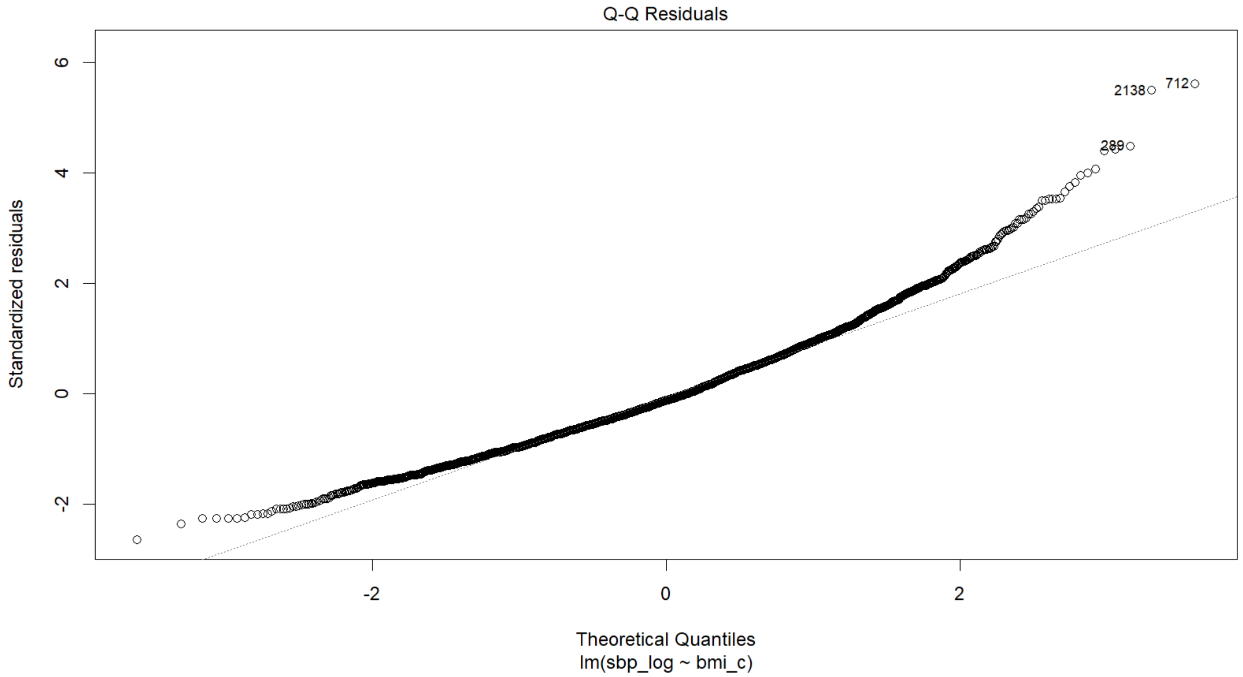
1f. [1 point] What is the interpretation of β_{BMI} in your equation?

A β_{BMI} of 0.0129 in the above equation says: a 1-unit increase in BMI is associated with a $100\left(e^{\beta} - 1\right) = 100(e^{0.13} - 1) = 100(1.013 - 1) = \mathbf{1.29}\%$ increase in SBP.

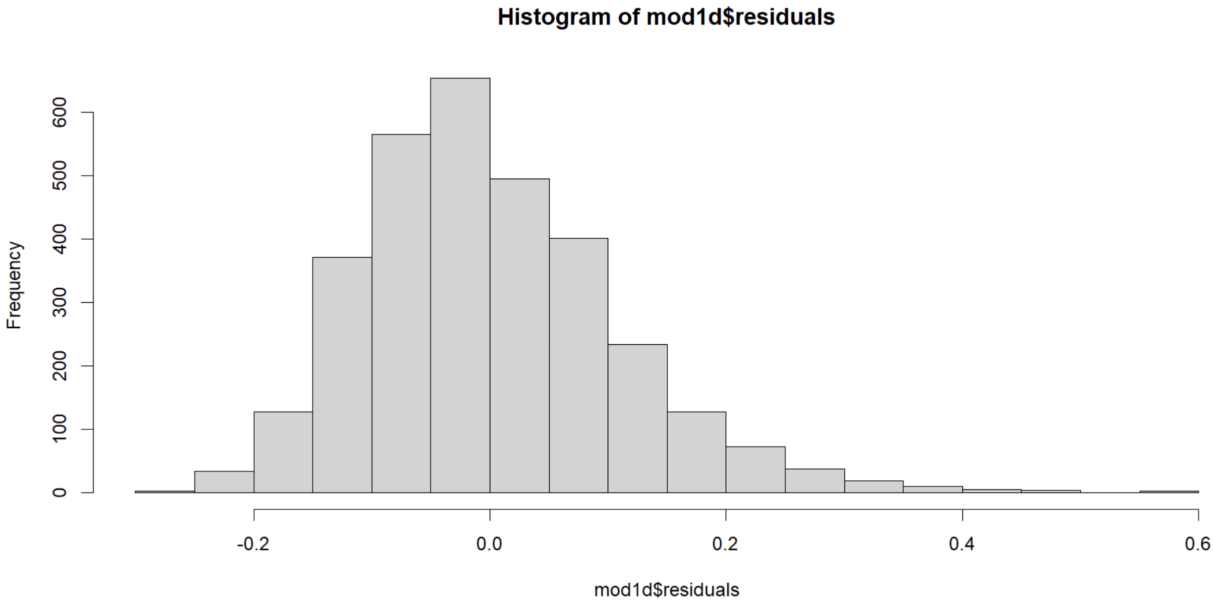
1g. [3 points] Assess each of the assumptions of this linear regression, providing relevant plots and statistics to support your conclusions.



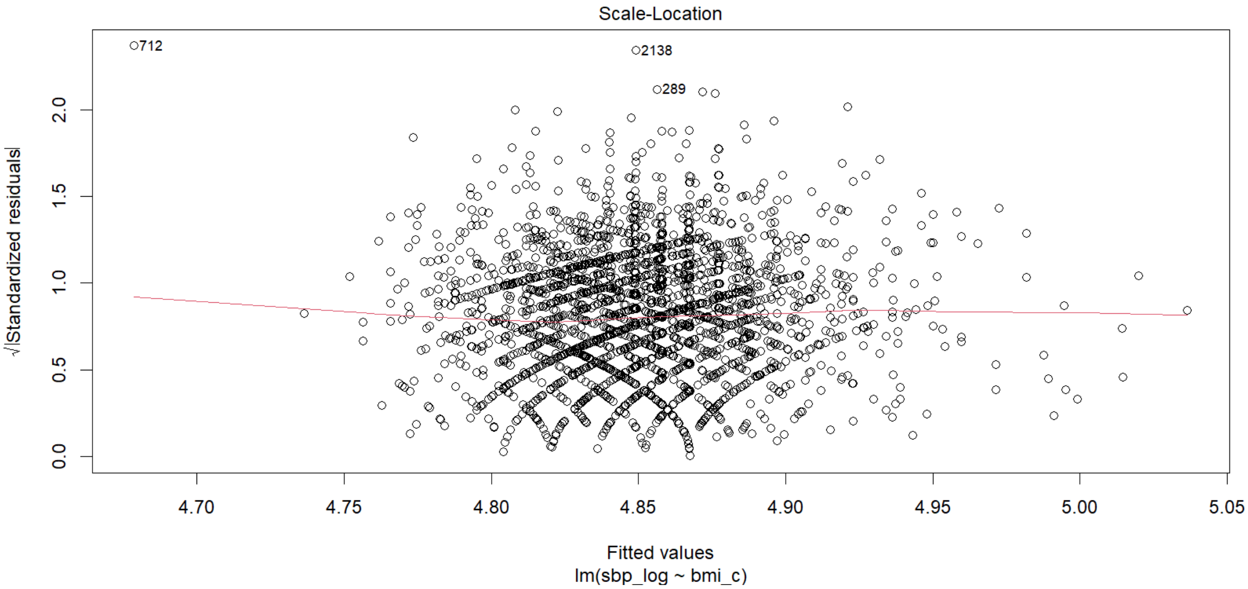
The residual plot shows a relatively flat scatter with a mean of 0, and there doesn't appear to be a pattern, so the assumption of linearity is met.



The QQ plot of the residuals shows deviations from the line at the ends, indicating some deviation from normality.



Judging from the histogram of the residuals of the regression model, it appears that the distribution is slightly skewed right. However, since the number of observations is in the thousands, the Central Limit Theorem will make the regression robust to non-normality of the residuals. Therefore, the assumption of normality can be assumed to be met.

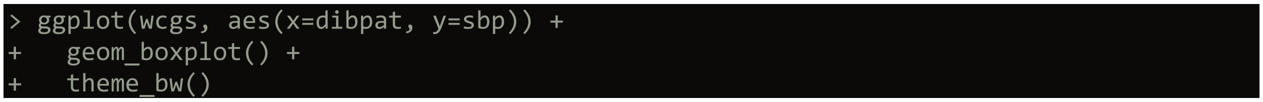


Lastly, the variance of residuals across all fitted values look more or less consistent with no clear pattern in the scale-location plot, meeting the assumption of homoscedasticity.

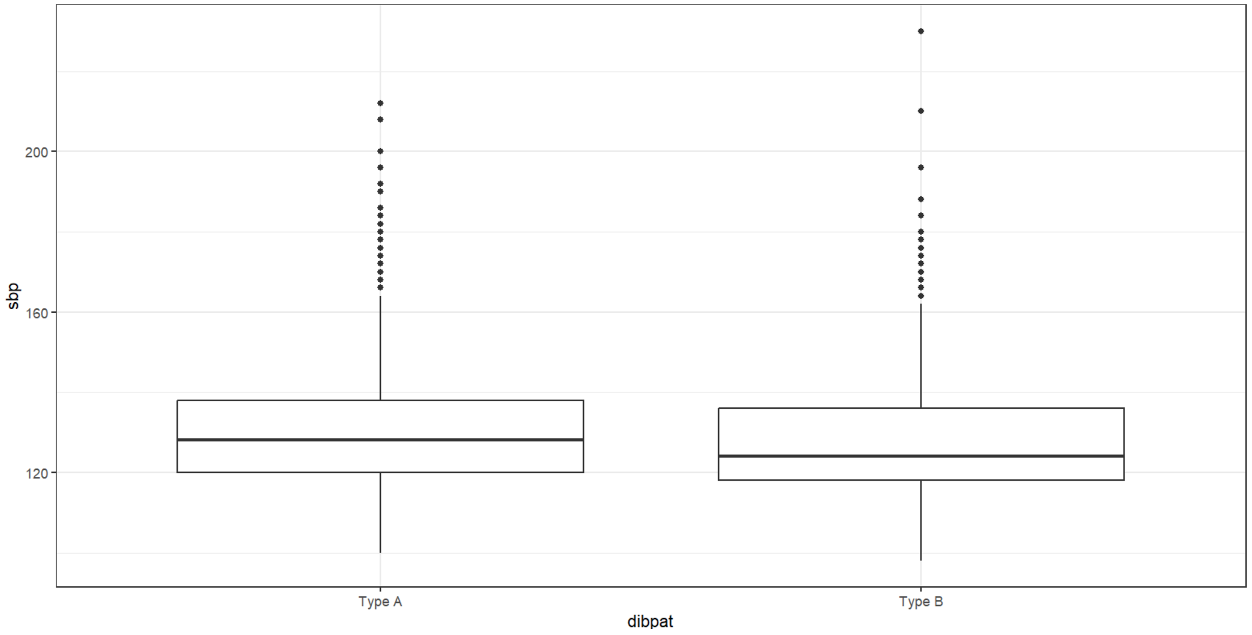
Question 2 [9 points]

Continue to use your WCGS file.

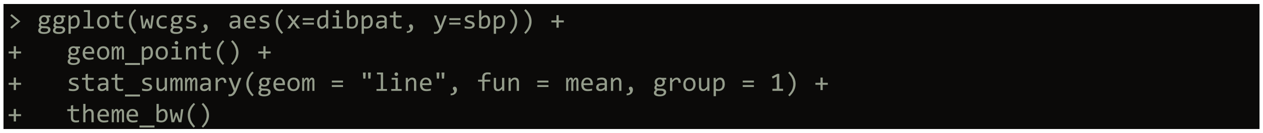
2a. [1 point] Create a boxplot of SBP vs. personality type (dibpat).



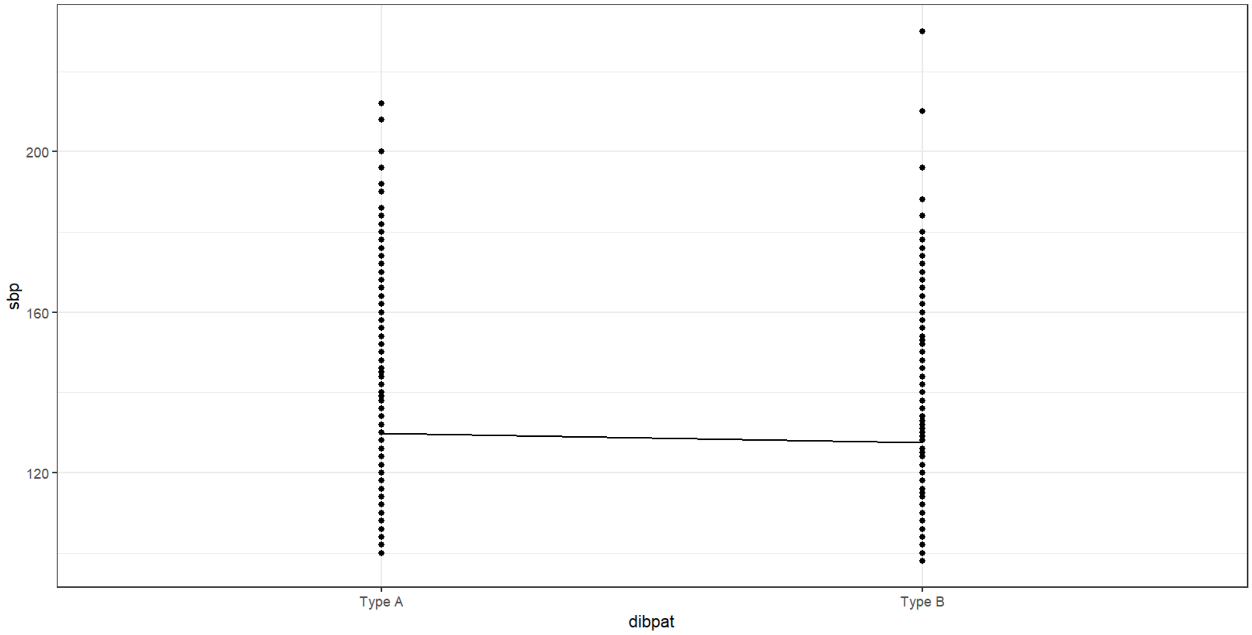
```
> ggplot(wcgs, aes(x=dibpat, y=sbp)) +
+   geom_boxplot() +
+   theme_bw()
```



2b. [1 point] Create a scatterplot of SBP vs. personality type, with a linear regression line.



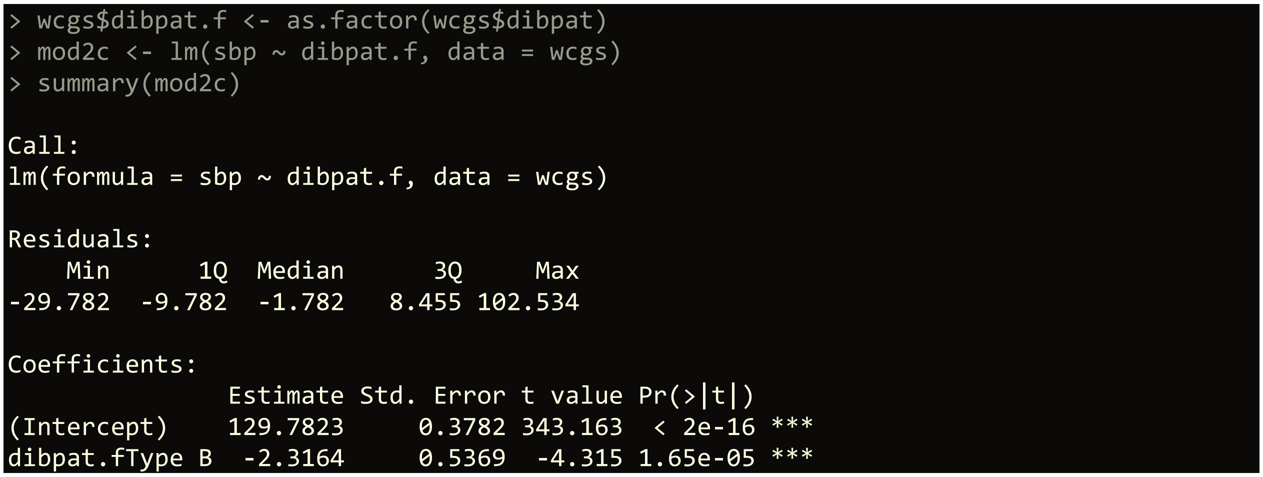
```
> ggplot(wcgs, aes(x=dibpat, y=sbp)) +
+   geom_point() +
+   stat_summary(geom = "line", fun = mean, group = 1) +
+   theme_bw()
```



2c. [1 point] Do you prefer using the boxplot or scatterplot to display this relationship? List at least one pro and con of each.

I personally prefer the boxplots to display the relationship. The boxplots give a better idea of where most points in each dibpat category actually lie, since points of the same SBP value are plotted on top of each other in the scatterplot. However, in the scatterplot it is easier to tell the difference in means between the two groups, whereas in the boxplot it is harder to tell if the mean lines are different between the two groups.

2d. [2 points] Regress SBP on personality type. Provide the best-fit regression equation and interpret the values of the intercept and slope in this equation.



```
> wcgs$dibpat.f <- as.factor(wcgs$dibpat)
> mod2c <- lm(sbp ~ dibpat.f, data = wcgs)
> summary(mod2c)

Call:
lm(formula = sbp ~ dibpat.f, data = wcgs)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.782  -9.782  -1.782   8.455 102.534 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    129.7823     0.3782 343.163  < 2e-16 ***
dibpat.fType B  -2.3164     0.5369  -4.315 1.65e-05 ***
```

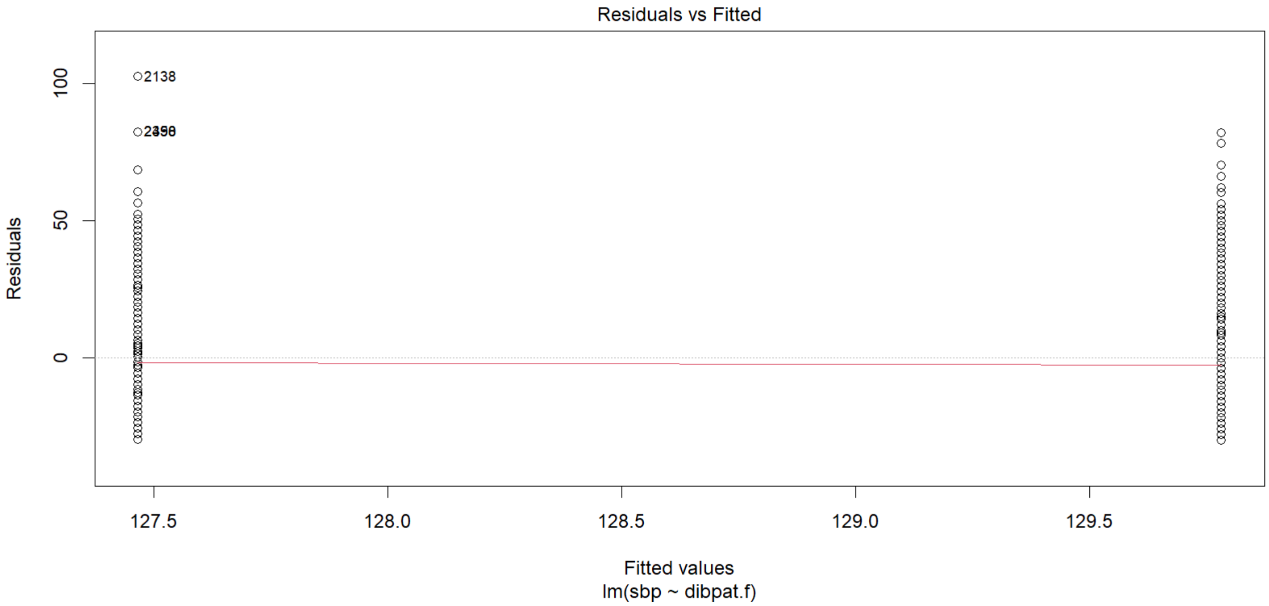
```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.08 on 3152 degrees of freedom
Multiple R-squared:  0.005871,    Adjusted R-squared:  0.005556
F-statistic: 18.62 on 1 and 3152 DF,  p-value: 1.649e-05
```

The best-fit linear regression equation for the relationship of SBP regressed on personality type is:

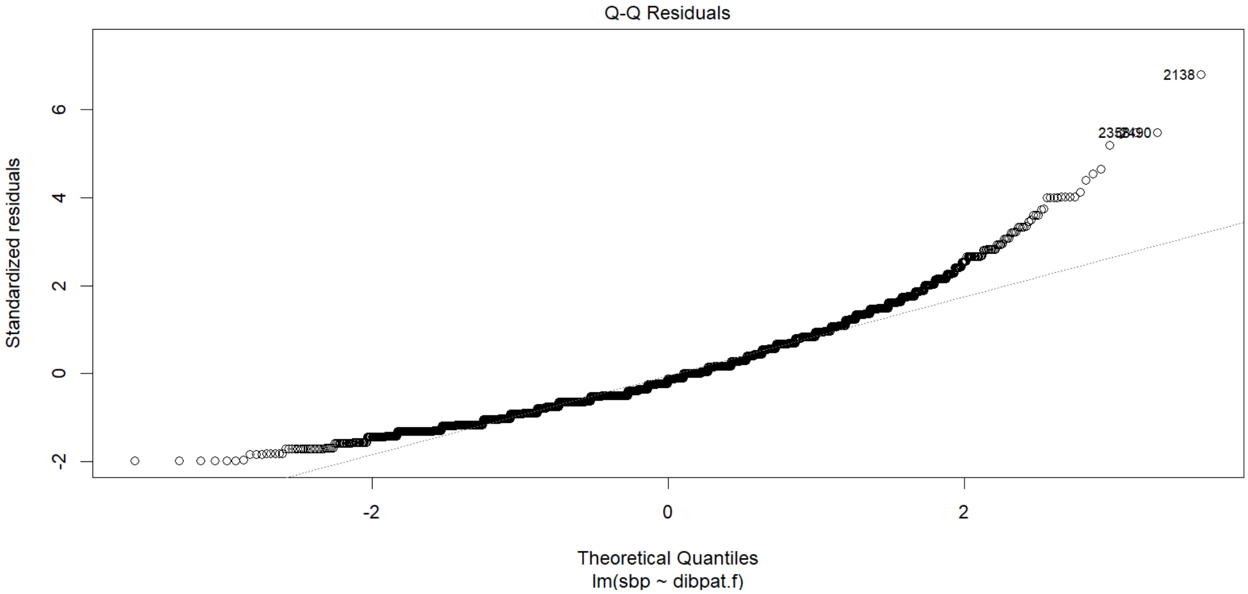
$$\hat{Y} = 129.78 + (-2.32X)$$

The intercept of 129.78 is the mean SBP value for those of Type A personality. The slope indicates that people with Type B personality have a mean SBP that is -2.32 units lower than the mean of those with Type A personality (127.46).

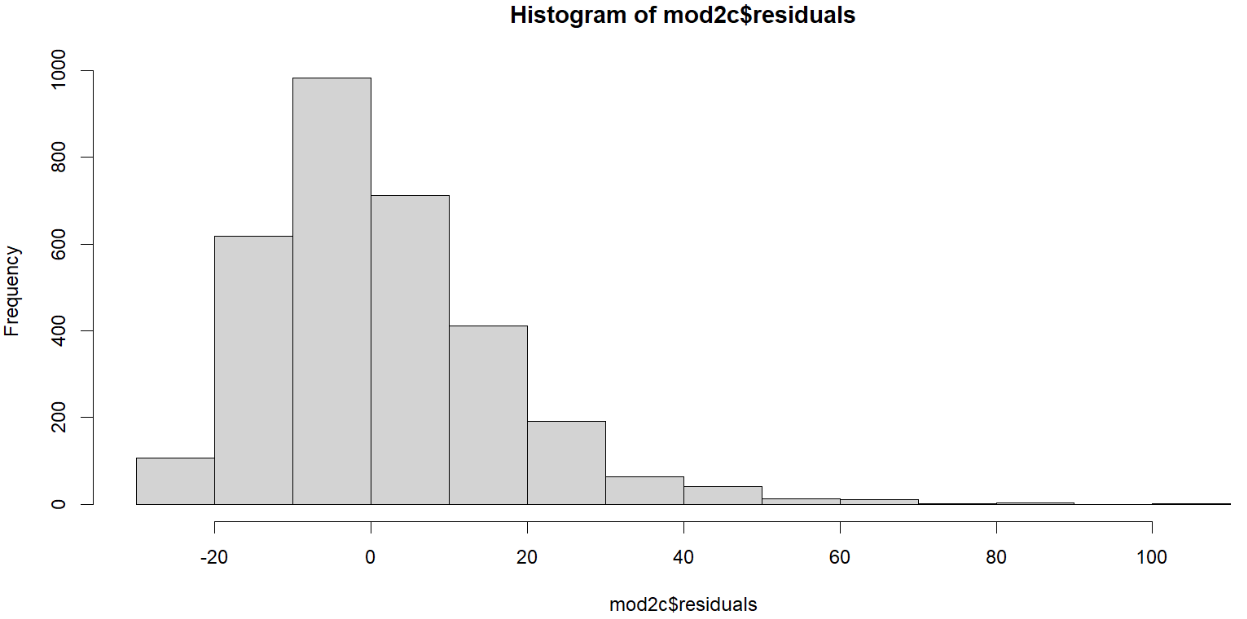
2e. [2 points] Assess each of the assumptions of this linear regression, providing relevant plots and statistics to support your conclusions.



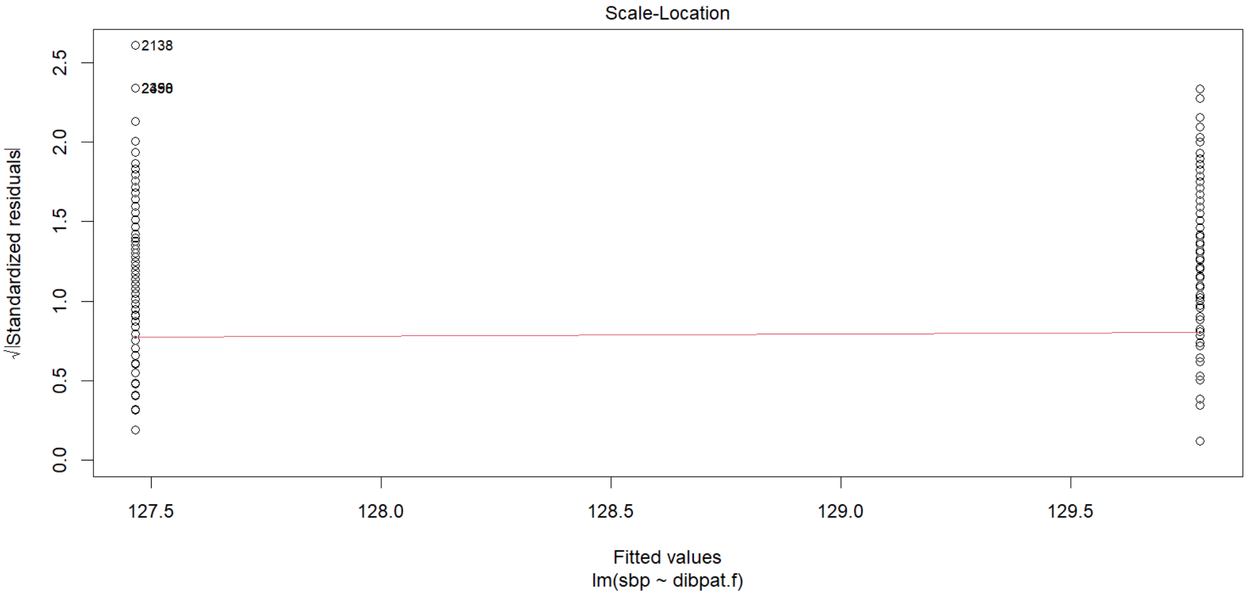
The residual plot has a midline that goes through 0 for and the middle of the residual distribution for both personality types, so the linearity assumption is met.



In the QQ plot of the residuals, there seems to be a deviation from the straight line, especially on the right.



The histogram of the residuals shows a right skew. However, since the number of observations is in the thousands, the Central Limit Theorem will make the regression robust to non-normality of the residuals. Therefore, the assumption of normality can be assumed to be met.

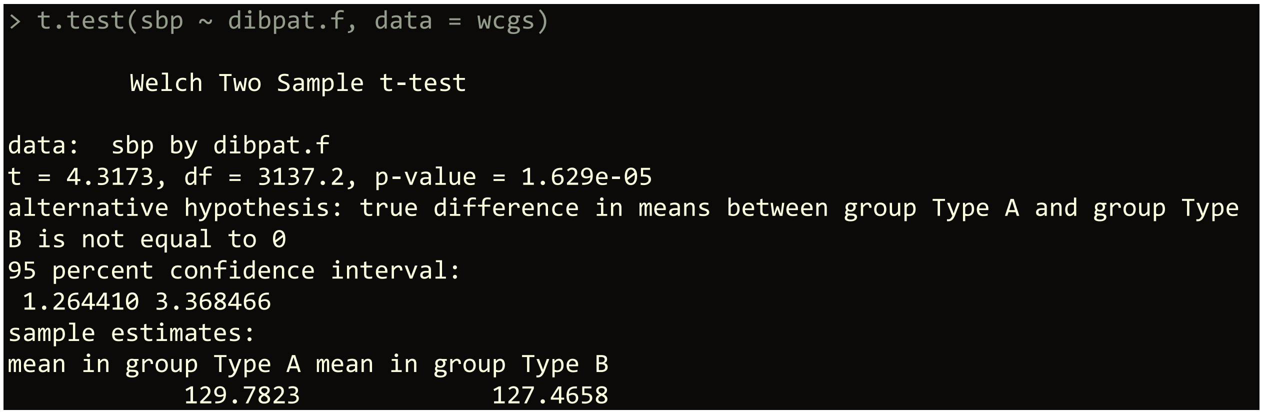


Lastly, the variance of residuals appears to be the same for both personality types, the assumption of homoscedasticity is met.

2f. [2 points] Is this regression method equivalent to 1) a two-sample t-test assuming equal but unknown variance or 2) a two-sample t-test assuming unequal but unknown variance? Explain why, and provide your output comparing the regression results to the appropriate t-test.

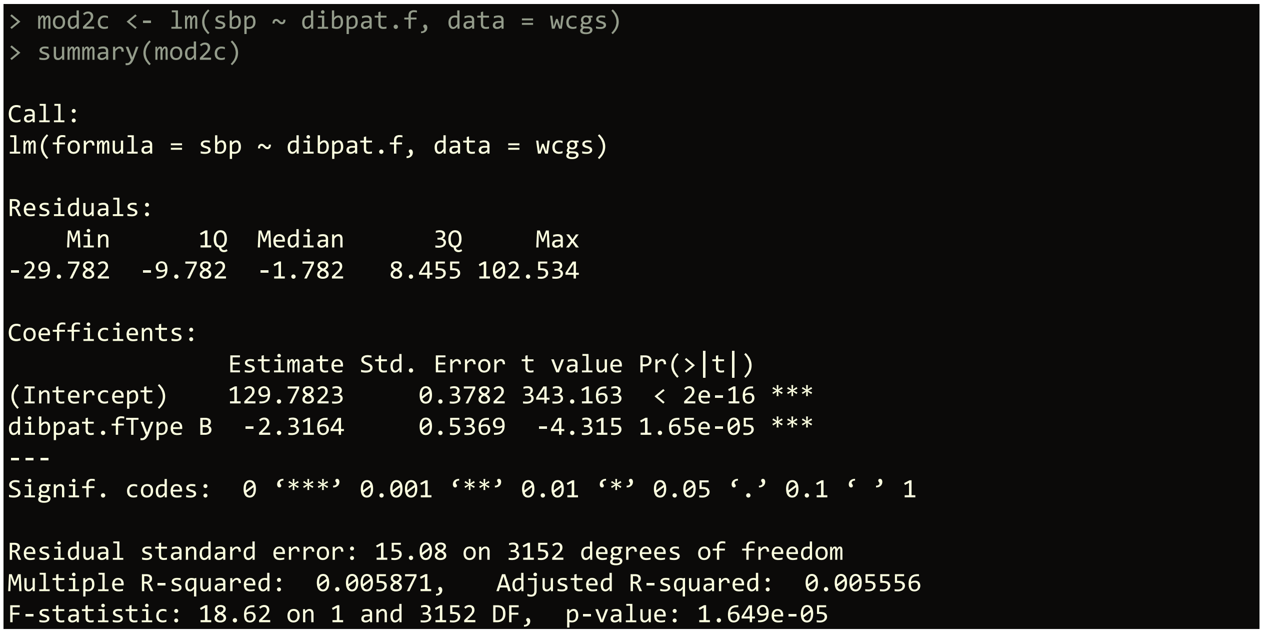
This regression method is equivalent to a two-sample t-test assuming equal but unknown variance because in order for the regression to be valid, it has to meet the assumption of homoscedasticity, which means that the variance of Y should be consistent for all (or in this case both) values of X.

Two-sample t-test assuming equal but unknown variance:



```
> t.test(sbp ~ dibpat.f, data = wcgs)

	Welch Two Sample t-test

data:  sbp by dibpat.f
t = 4.3173, df = 3137.2, p-value = 1.629e-05
alternative hypothesis: true difference in means between group Type A and group Type
B is not equal to 0
95 percent confidence interval:
 1.264410 3.368466
sample estimates:
mean in group Type A mean in group Type B
            129.7823             127.4658
```

Linear regression



```
> mod2c <- lm(sbp ~ dibpat.f, data = wcgs)
> summary(mod2c)

Call:
lm(formula = sbp ~ dibpat.f, data = wcgs)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.782  -9.782  -1.782   8.455 102.534 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    129.7823     0.3782 343.163  < 2e-16 ***
dibpat.fType B  -2.3164     0.5369  -4.315 1.65e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.08 on 3152 degrees of freedom
Multiple R-squared:  0.005871,	Adjusted R-squared:  0.005556 
F-statistic: 18.62 on 1 and 3152 DF,  p-value: 1.649e-05
```

The p-values are the same in both tests: 1.65e-05

The mean for Type A is the same as the regression intercept: 129.78

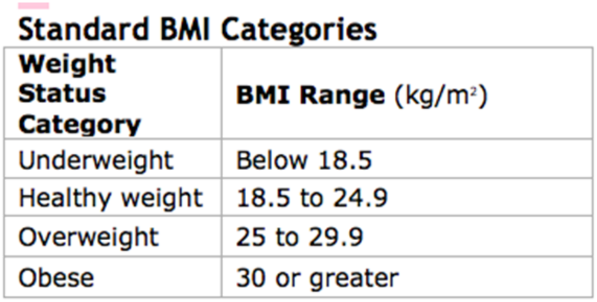
The absolute values of the test statistic are the same in both tests: 4.315

Question 3 [10 points]

Continue to use your WCGS file. In this question we will look at a different way that we can examine the relationship between BMI and SBP.

Whenever we categorize a continuous variable, some information always gets lost. However, in some cases categorizing a continuous variable can help satisfy some regression assumptions. In other cases, a variable’s categorical value may be the only information recorded in the data set.

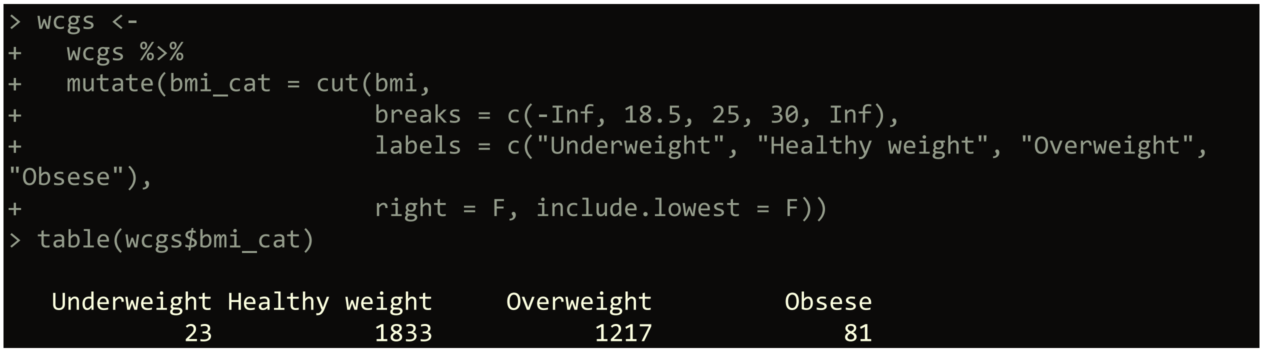
In this question we will gain practice examining a categorical predictor.



Standard BMI Categories

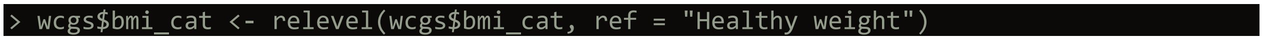
Weight Status Category	**BMI Range** (kg/m^2)
Underweight	Below 18.5
Healthy weight	18.5 to 24.9
Overweight	25 to 29.9
Obese	30 or greater

3a. [1 point] Categorize BMI using the cut point criteria defined above, and provide the number of individuals that fall into each category.



```
> wcgs <-
+   wcgs %>%
+   mutate(bmi_cat = cut(bmi,
+                        breaks = c(-Inf, 18.5, 25, 30, Inf),
+                        labels = c("Underweight", "Healthy weight", "Overweight",
"Obsese"),
+                        right = F, include.lowest = F))
> table(wcgs$bmi_cat)

   Underweight Healthy weight     Overweight         Obsese 
            23           1833           1217             81 
```

3b. [2 points] Define a set of dummy variables that are sufficient to describe these four categories. Make healthy weight the reference group.



```
> wcgs$bmi_cat <- relevel(wcgs$bmi_cat, ref = "Healthy weight")
```

3c. [2 points] Using notation and the variables you defined in 3b, write the theoretical regression equation for the regression of SBP on BMI. State the null and alternative hypotheses that would test if BMI category is related to SBP.

$$\hat{Y} = \beta_0 + \beta_{underweight}X_{underweight} + \beta_{overweight}X_{overweight} + \beta_{obese}X_{obese}$$

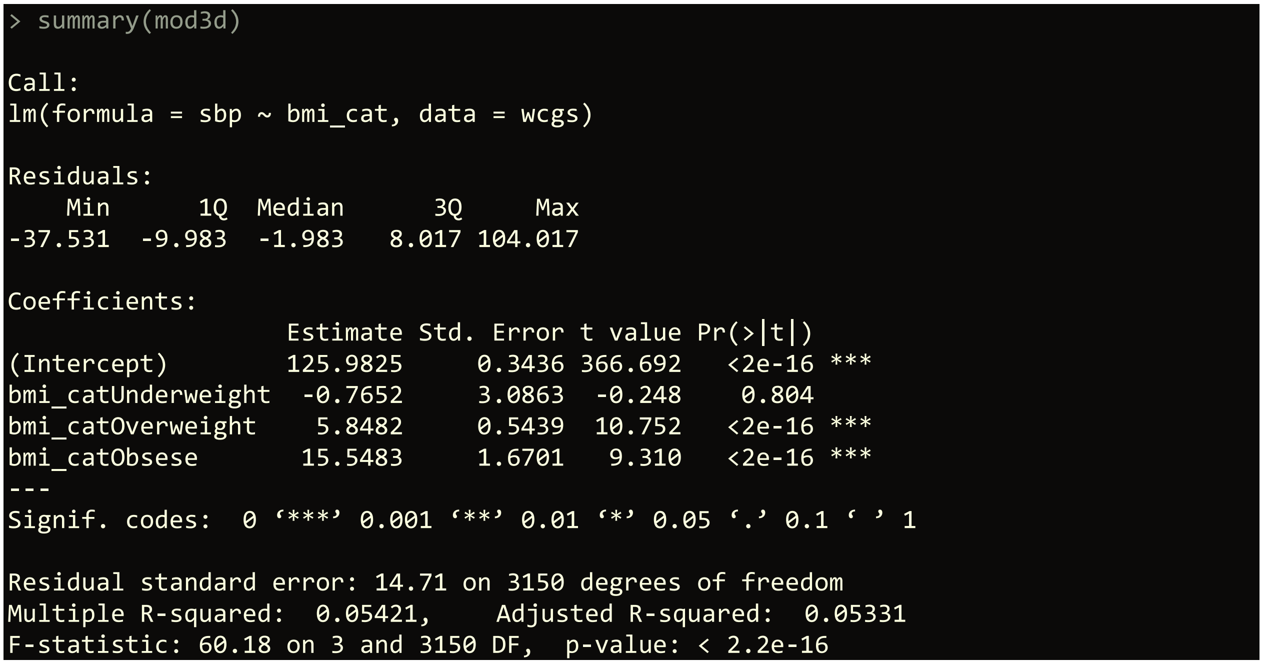
H_0: BMI category is not related to SBP $\beta_{underweight} = 0, \beta_{overweight} = 0, \beta_{obese} = 0$
H_A : BMI category is related to SBP (at least 1 beta value is not equal to 0)

3d. [2 points] Write your best-fit regression equation and interpret the intercept and slope values.

$$\hat{Y} = 125.98 - 0.77X_{underweight} + 5.85X_{overweight} + 15.55X_{obese}$$

The intercept of 125.98 is the mean SBP for individuals that are of healthy weight.
The slope for the underweight category (-0.77) can be interpreted to say that the mean SBP for underweight individuals is -0.77 units lower than the mean SBP for healthy weight individuals.
The slope for the overweight category (5.85) can be interpreted to say that the mean SBP for overweight individuals is 5.85 units higher than the mean SBP for healthy weight individuals.
The slope for the obese category (15.55) can be interpreted to say that the mean SBP for obese individuals is 15.55 units higher than the mean SBP for healthy weight individuals.

3e. [2 points] What are your decisions for the hypotheses in 3c? Provide the test statistic and p-value.



```
> summary(mod3d)

Call:
lm(formula = sbp ~ bmi_cat, data = wcgs)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.531  -9.983  -1.983   8.017 104.017 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        125.9825     0.3436 366.692   <2e-16 ***
bmi_catUnderweight  -0.7652     3.0863  -0.248    0.804    
bmi_catOverweight    5.8482     0.5439  10.752   <2e-16 ***
bmi_catObsese       15.5483     1.6701   9.310   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.71 on 3150 degrees of freedom
Multiple R-squared:  0.05421,	Adjusted R-squared:  0.05331 
F-statistic: 60.18 on 3 and 3150 DF,  p-value: < 2.2e-16
```

The F-test tests whether BMI overall is associated with SBP. The test-statistic and p-value are 60.18 (on 3 and 3150 degrees of freedom) and 2.2e-16, respectively. The low p-value (<< 0.001) suggests there is enough evidence to reject the null hypothesis and accept the alternative hypothesis that BMI is associated with SBP.

3f. [1 point] Which BMI categories have mean SBP values that are significantly different from the “healthy weight” mean SBP value?

It appears that the BMI categories “Overweight” and “Obese” have mean SBP values that are significantly different from the “healthy weight” mean SBP value, both with p-values < 2e-16. On the other hand, the “underweight” category has a mean SBP value that is not significantly different from the “healthy weight” mean SBP value (p = 0.8).