Lab 8 Regression

Chen Wang, Junke Wang, Zhuozhi Xiong

due 11/27/19

Together as a class:

We are first going to install and load packages that we will need.

1. Install and load both the `modelr` package and the `openintro` package. Load the `tidyverse` package.

```
install.packages('modelr')
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.6'
## (as 'lib' is unspecified)
```

```
install.packages('openintro')
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.6'
## (as 'lib' is unspecified)
```

```
install.packages("tidyverse")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.6'
## (as 'lib' is unspecified)
```

```
library(modelr)
library(openintro)
```

```
## Please visit openintro.org for free statistics materials

##
## Attaching package: 'openintro'

## The following objects are masked from 'package:datasets':
##
##     cars, trees
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

For this problem we will be using the data set called `gpa` from the `openintro` package. We are interest in association between the number of hours of sleep a student gets and their gpas.

```
install.packages("modelr")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.6'
## (as 'lib' is unspecified)
```

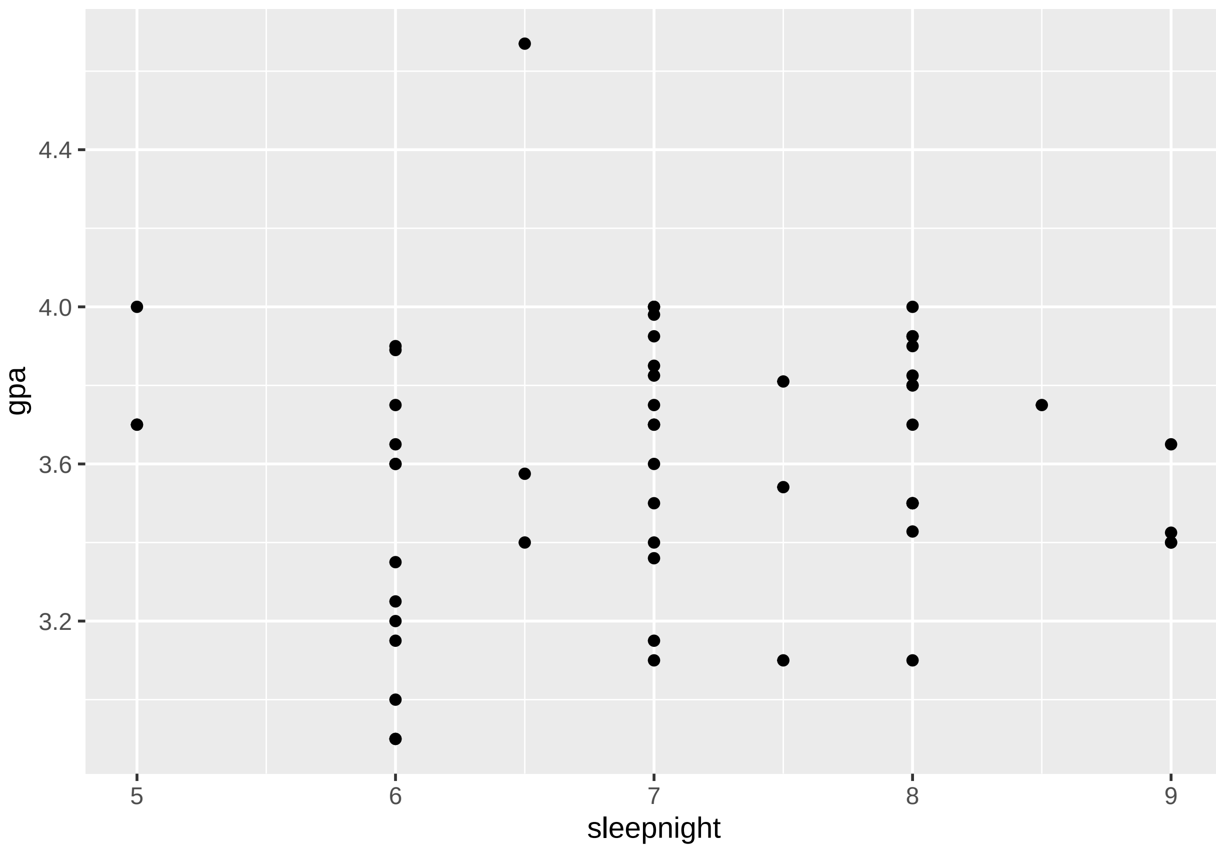
```r
install.packages("openintro")
```

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/3.6'
## (as 'lib' is unspecified)
```

```r
library(modelr)
library(openintro)
library(tidyverse)
```

2. Make a scatter plot and describe the association that you see.

```r
gpa %>%
  ggplot(aes(x=sleepnight, y=gpa)) +
  geom_point()
```



It doesn't seem that there is a relationship between hours of sleep and GPA. We can't tell if there is a positive or a negative trend.

3. Create the linear regression model for this relationship between hours of sleep and GPA. We are regressing `gpa` onto `hours of sleep` Call this model `gpa_model`.

```r
gpa_model <- lm(gpa ~ sleepnight, data=gpa)

gpa_model %>%
  summary()
```

```
##
## Call:
## lm(formula = gpa ~ sleepnight, data = gpa)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.67898 -0.22123  0.02102  0.21627  1.08110 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  3.46000    0.31819  10.874 4.14e-15 ***
## sleepnight   0.01983    0.04458   0.445    0.658    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3381 on 53 degrees of freedom
## Multiple R-squared:  0.003719,   Adjusted R-squared:  -0.01508 
## F-statistic: 0.1978 on 1 and 53 DF,  p-value: 0.6583
```

4. Find the correlation of this regression line and give the estimated β values.

```
correlation <- sqrt(0.003719)#add sign of intercept
correlation
```

```
## [1] 0.0609836
```

```
cor(gpa$gpa, gpa$sleepnight)
```

```
## [1] 0.06098308
```

the correlation is 0.061. The estimated intercept is 3.46. The estimated slope is 0.01983.

5. Interpret the $\hat{\beta}$ values. There are 2.

GPA = intercept + slope(hours of sleep)

We can intercept the intercept as us having an estimated GPA of 3.46 when we have 0 hours of sleep.
We can interpret the slope as for every 1 extra hour of sleep we get, we will increase our GPA by 0.01983 points.

6. Add the predicted values of GPA and the Residuals to the data frame `gpa` using the `add_predictions()` and `add_residuals()` functions from the `modelr` package.

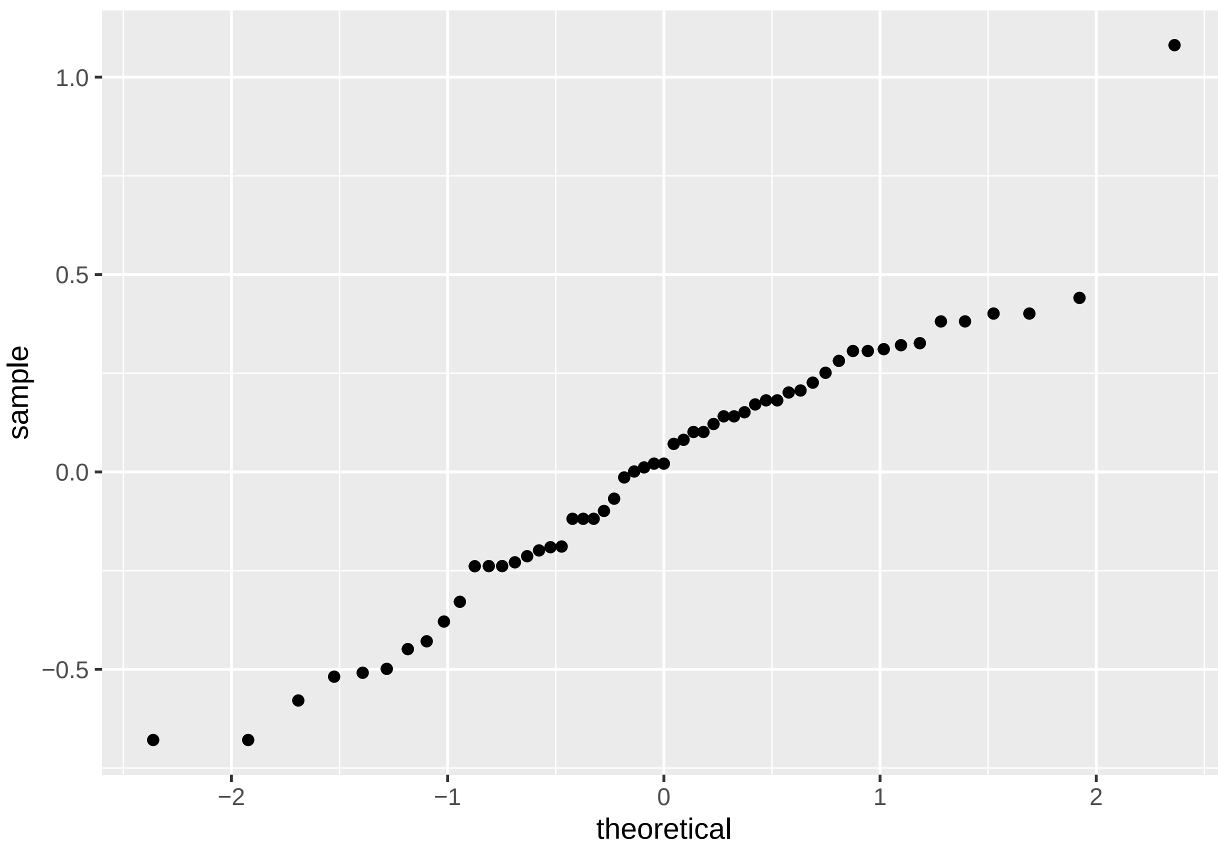
```
gpa<-gpa %>%
  add_predictions(gpa_model) %>%
  add_residuals(gpa_model)
```

```
gpa<-gpa %>%
  add_predictions(gpa_model) %>%
  add_residuals(gpa_model)
```

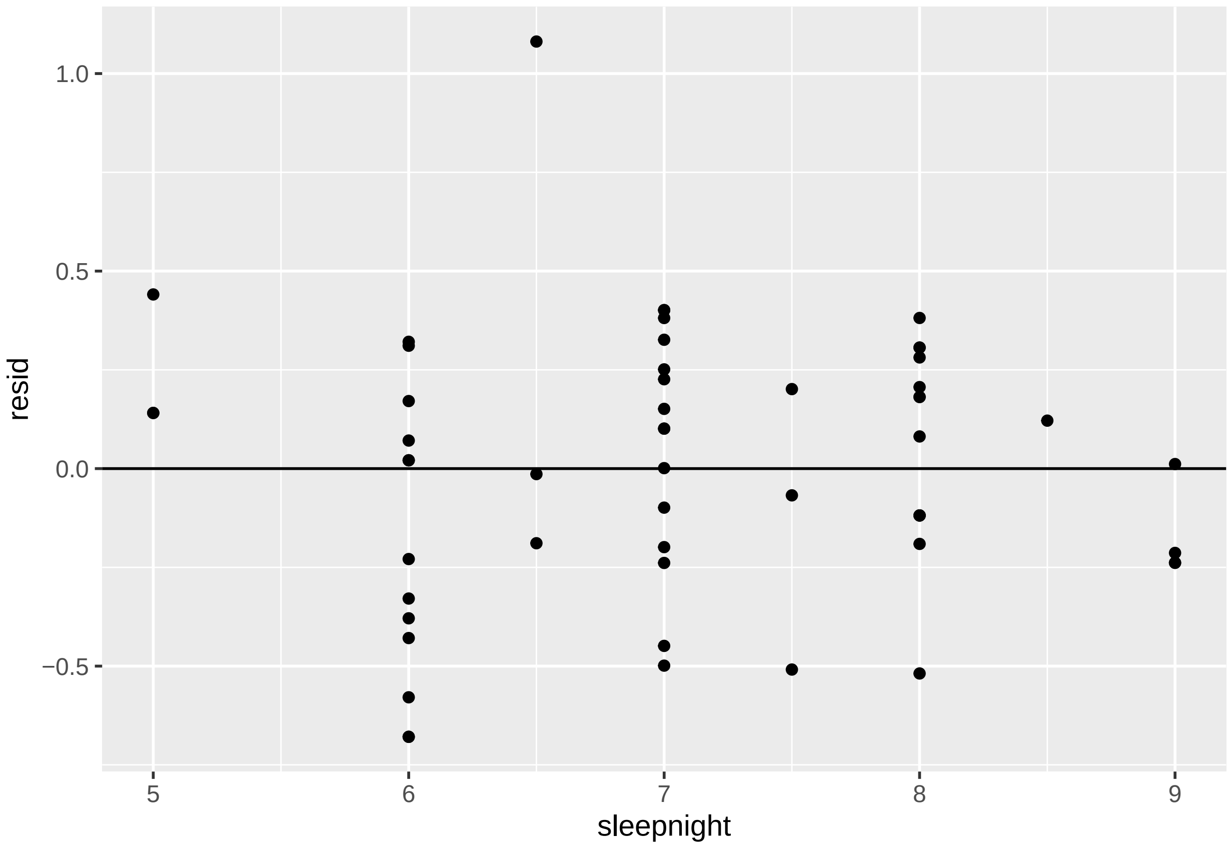
7. Now we want to check the conditions needed to use the least squares regression line. Create a qqplot and a residual plots in order to check the conditions. Are the conditions meet? Are there any outliers?
   1 - independence of data
   2 - linear relationship of GPA and hours of sleep
   3 - normality of residuals
   4 - constant variability

<<<<<<< HEAD

```
gpa %>%
  ggplot(aes(sample=resid))+
  geom_qq()
```




```
gpa %>%
  ggplot(aes(x=sleepnight,y=resid))+
  geom_point()+
  geom_hline(yintercept=0)
```



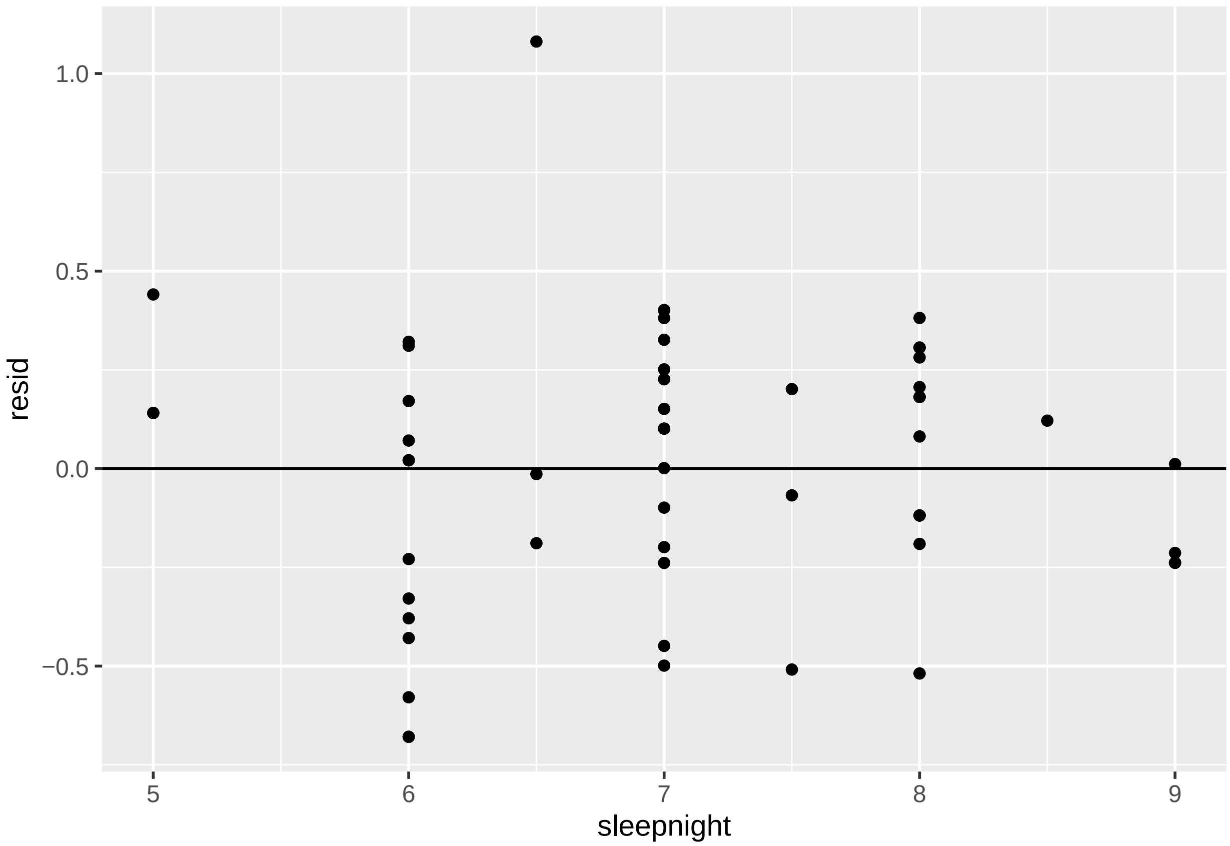
All of our assumptions check out. We can use linear regression.

1 - independence of data

2 - linear relationship of GPA and hours of sleep

4 - constant variability >>>>>>> ba9d74e8d944f5180aec53926d33f133f71d592c

```
gpa %>%
  ggplot(aes(x=sleepnight,y=resid))+
  geom_point()+
  geom_hline(yintercept=0)
```



All of our assumptions check out. We can use linear regression.

8. Conduct a hypothesis test to see if there is an association between how much sleep a student gets and their GPA. What can we conclude?

$H_0 : \beta_1 = 0$
(We assume the slope is 0. Hours of sleep and GPA have no linear relationship.)
$H_0 : \beta_1 \neq 0$
(We want to prove that the slope is not 0. That hours of sleep and GPA have a linear relationship of some kind.)

```
gpa_model %>%
  summary()
```

```
##
## Call:
## lm(formula = gpa ~ sleepnight, data = gpa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67898 -0.22123  0.02102  0.21627  1.08110
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.46000    0.31819  10.874 4.14e-15 ***
## sleepnight   0.01983    0.04458   0.445    0.658
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3381 on 53 degrees of freedom
## Multiple R-squared:  0.003719,   Adjusted R-squared:  -0.01508
## F-statistic: 0.1978 on 1 and 53 DF,  p-value: 0.6583
```

Our t-statistics is 0.445, our p-value is 0.658. Because our p-value is greater than 0.05, we fail to reject the null hypothesis, we cannot say there is a linear relationship between hours of sleep and GPA.

$$H_0 : \beta_1 = 0$$

(The population slope is 0. There is no relationship between hours of sleep and GPA.)

$$H_A : \beta_1 \neq 0$$

(The population slope is not 0. There is a relationship between hours of sleep and GPA.)


```
gpa_model %>%
  summary()
```

```
##
## Call:
## lm(formula = gpa ~ sleepnight, data = gpa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67898 -0.22123  0.02102  0.21627  1.08110
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.46000    0.31819  10.874 4.14e-15 ***
## sleepnight   0.01983    0.04458   0.445    0.658
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3381 on 53 degrees of freedom
## Multiple R-squared:  0.003719,   Adjusted R-squared:  -0.01508
## F-statistic: 0.1978 on 1 and 53 DF,  p-value: 0.6583
```

Our test statistic T = 0.0445. Our p-value is 0.658. Because our p-value>0.05, we fail to reject the null hyposis. We cannot say if there is a relationship between hours of sleep and GPA.

On your own:

In this problem, we will be using the `babies` data set from the `openintro` package. information about the data set can be found by running the following code in an `r` chunk: `?babies`. We are interested in finding out how the length of gestation(`gestation`), the age of the mother(`age`), the height of the mother(`height`), the mother's weight(`weight`), and whether or not this was the mother's first pregnacy (`parity`) are related to how much the baby will weight at birth (`bwt`).

1.Load the data and save it as `baby`.

```
baby <- babies
```

2. Create the linear regression model that regresses `bwt` the variables mentioned above. Print out the summary.

```
baby_model <- lm(bwt ~ gestation + age + height + weight + parity, data= baby)

baby_model %>%
  summary()
```

```
##
## Call:
## lm(formula = bwt ~ gestation + age + height + weight + parity,
##     data = baby)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.569 -10.506   0.453  10.063  54.285
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -86.59435   14.75733  -5.868 5.73e-09 ***
## gestation     0.46579    0.02992  15.568  < 2e-16 ***
## age           0.05233    0.08820   0.593   0.5531
## height        1.04614    0.21064   4.967 7.82e-07 ***
## weight        0.06564    0.02584   2.540   0.0112 *
## parity       -2.96394    1.16403  -2.546   0.0110 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.35 on 1178 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.2106, Adjusted R-squared:  0.2072
## F-statistic: 62.84 on 5 and 1178 DF,  p-value: < 2.2e-16
```

3. Interpret the coefficients for `gestation` and `parity`.
   The coefficient for gestination is 0.46579, which means when other explanatory variables don't change, and only gestination increases by 1 unit, the weight of baby at birth will increase by 0.46579 unit.
   The coefficient for parity is -2.96394, which means when other explanatory variables don't change, and only parity increases by 1 unit, the weight of baby at birth will decrease by 2.96394 units.

4. Does the intercept value have any meaning in context?
   Theoretically, the intercept value means when gestation, age, height, weight and parity are all 0, then the weight of the baby will -86,59435. But since these factors can not be 0, and baby's weight can not be negative, so the intercept value doesn't have meaning in context.

5. What is the Multiple R^2 value for this model.

```r
baby_model %>%
  summary()
```

```
##
## Call:
## lm(formula = bwt ~ gestation + age + height + weight + parity,
##     data = baby)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.569 -10.506   0.453  10.063  54.285
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -86.59435   14.75733  -5.868 5.73e-09 ***
## gestation     0.46579    0.02992  15.568  < 2e-16 ***
## age           0.05233    0.08820   0.593   0.5531
## height        1.04614    0.21064   4.967 7.82e-07 ***
## weight        0.06564    0.02584   2.540   0.0112 *
## parity       -2.96394    1.16403  -2.546   0.0110 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.35 on 1178 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.2106, Adjusted R-squared:  0.2072
## F-statistic: 62.84 on 5 and 1178 DF,  p-value: < 2.2e-16
```

We can see that multiple R^2 for this model is 0.2106.

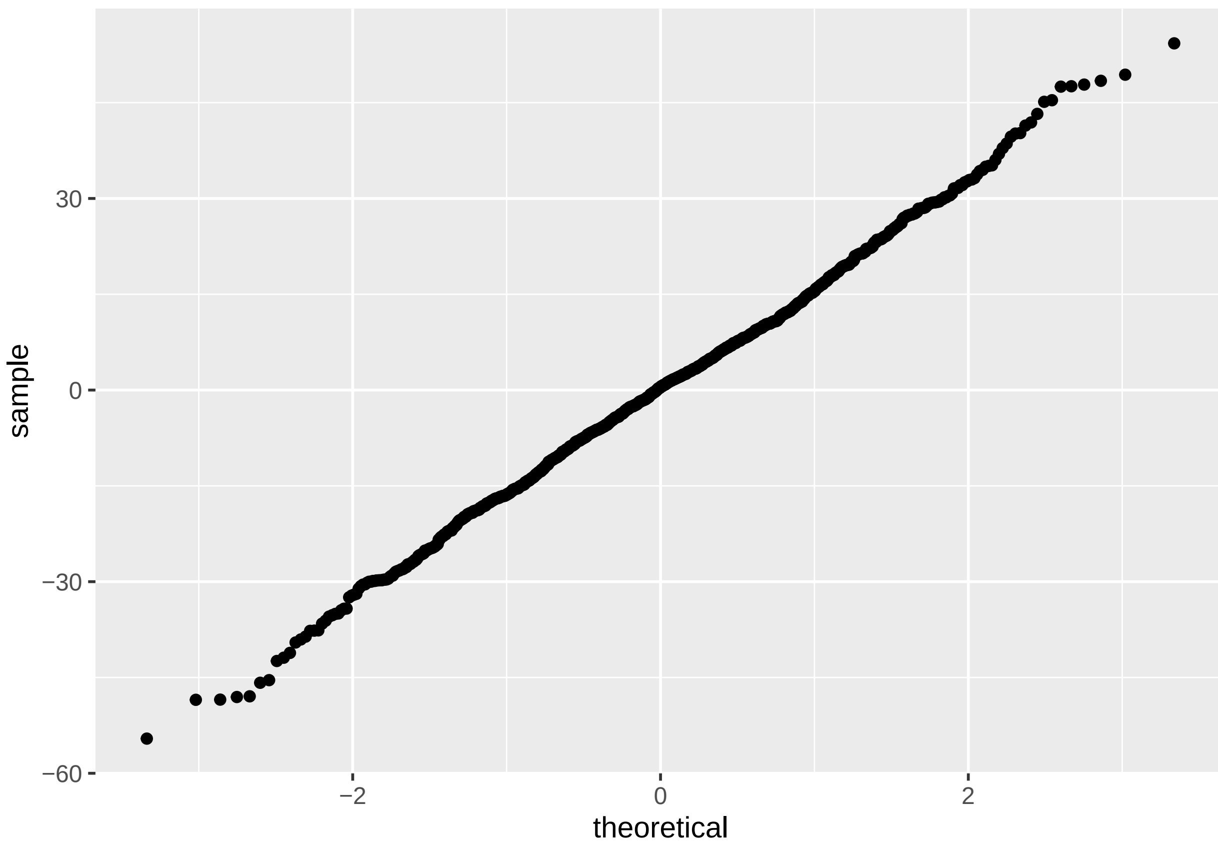
6. Add the residuals and predicted values to the data frame.

```
baby <- baby %>%
  add_predictions(baby_model) %>%
  add_residuals(baby_model)
```

7. Create a residual plot and a qqplot. Comment on whether or not the conditions are met to use the model you found in part 2.
   The conditions include:
   1 - independence of data
   2 - linear relationship
   3 - normality of residuals
   4 - constant variability

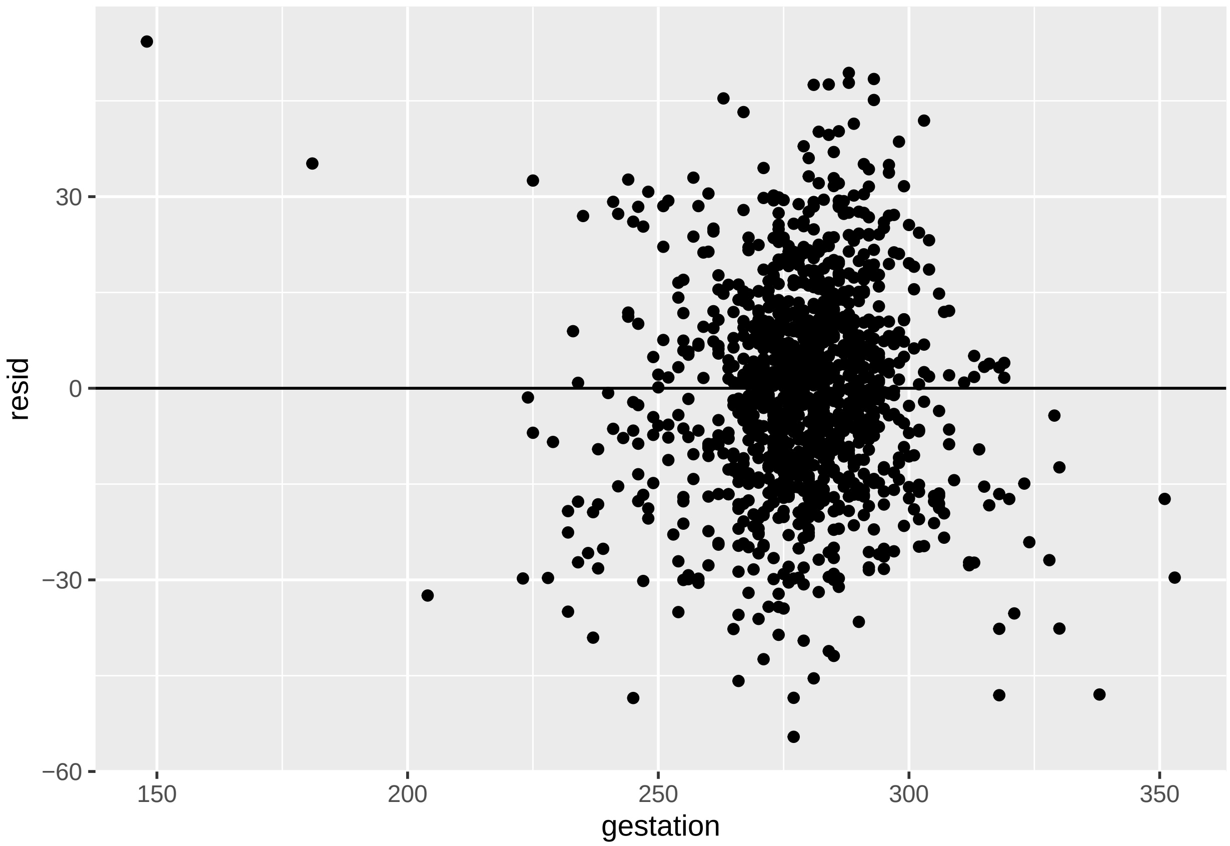
```
baby %>%
  ggplot(aes(sample=resid))+
  geom_qq()
```

```
## Warning: Removed 52 rows containing non-finite values (stat_qq).
```



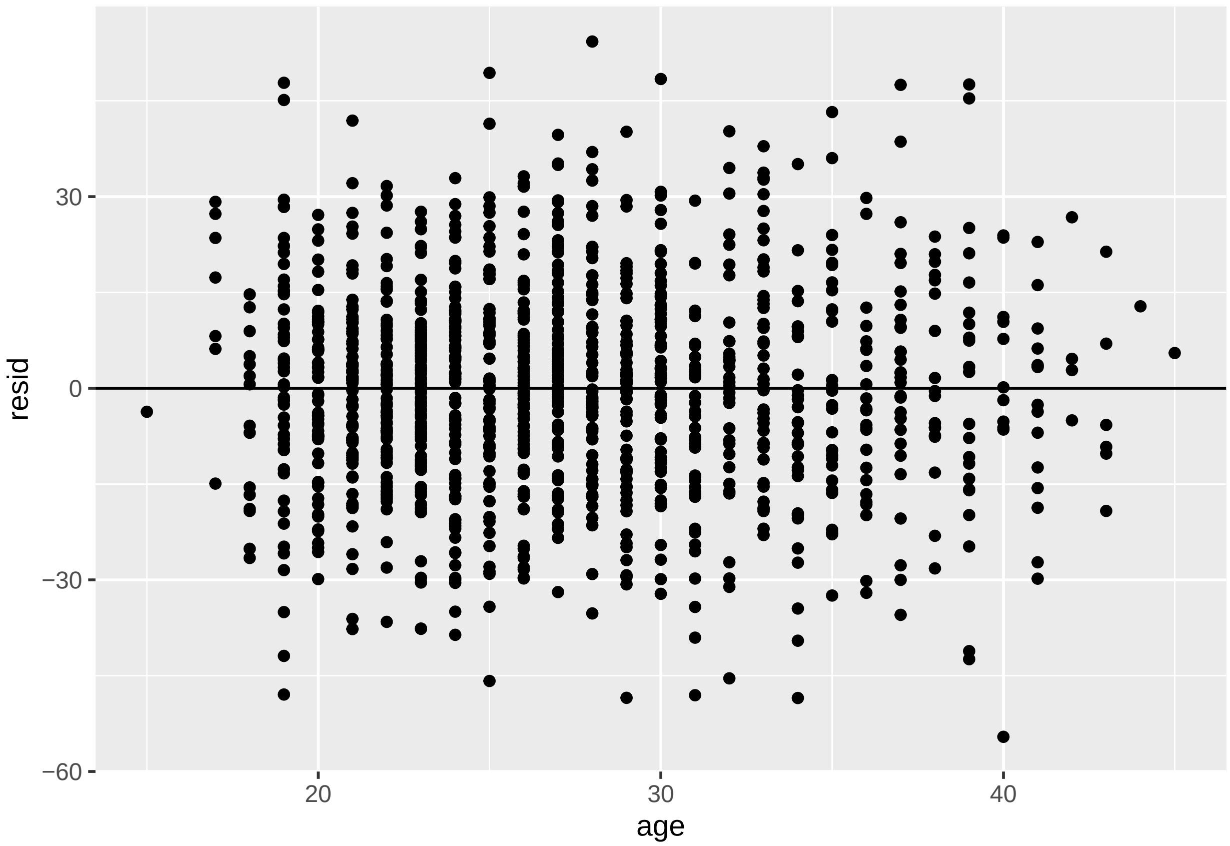

```
baby %>%
  ggplot(aes(x=gestation,y=resid))+
  geom_point()+
  geom_hline(yintercept=0)
```

```
## Warning: Removed 52 rows containing missing values (geom_point).
```



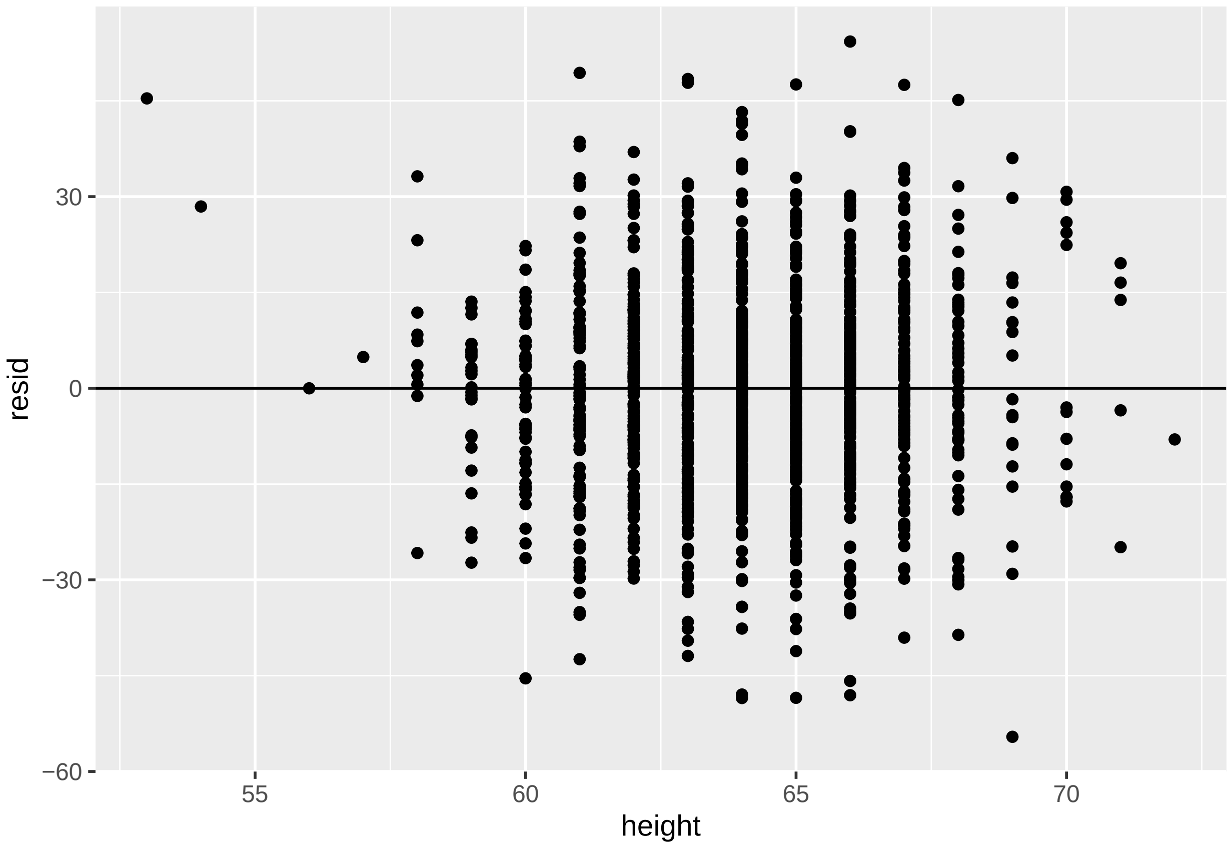

```
baby %>%
  ggplot(aes(x=age,y=resid))+
  geom_point()+
  geom_hline(yintercept=0)
```

```
## Warning: Removed 52 rows containing missing values (geom_point).
```



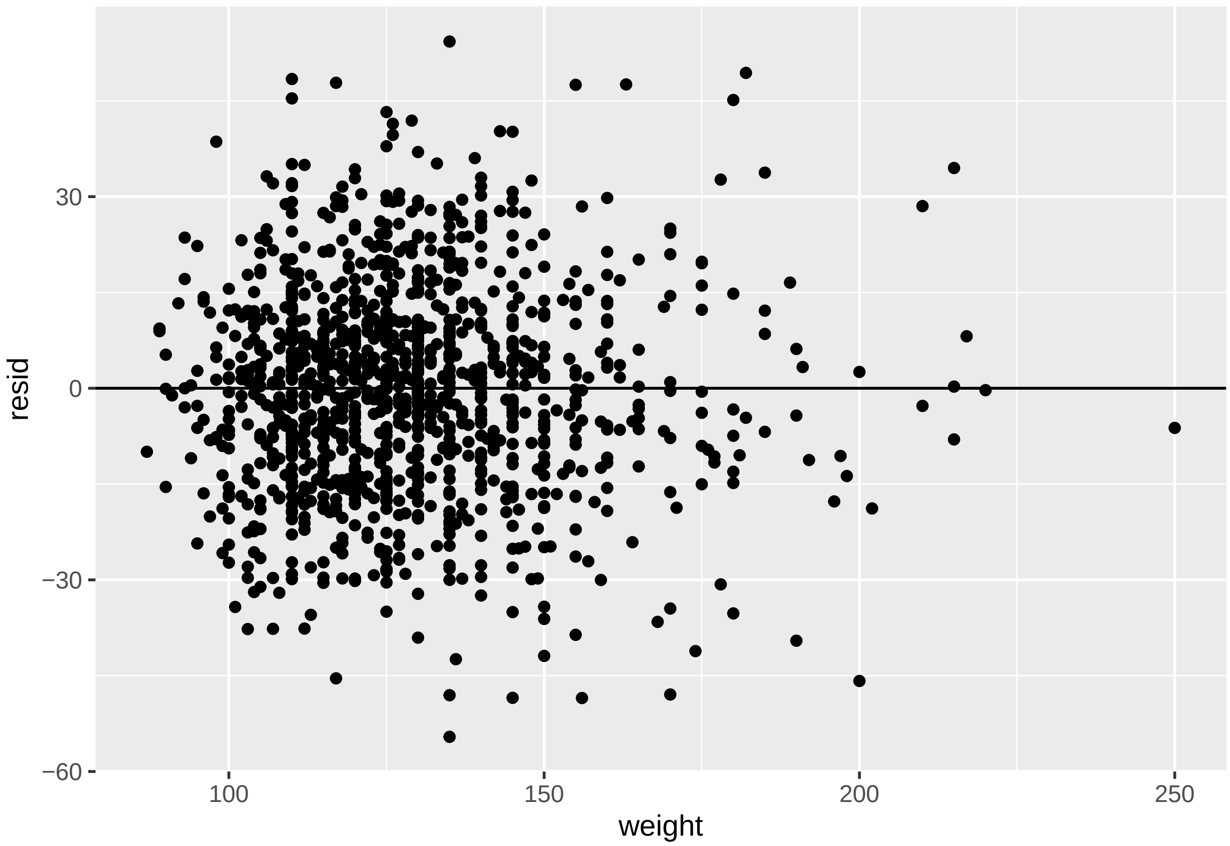

```
baby %>%
  ggplot(aes(x=height,y=resid))+
  geom_point()+
  geom_hline(yintercept=0)
```

```
## Warning: Removed 52 rows containing missing values (geom_point).
```



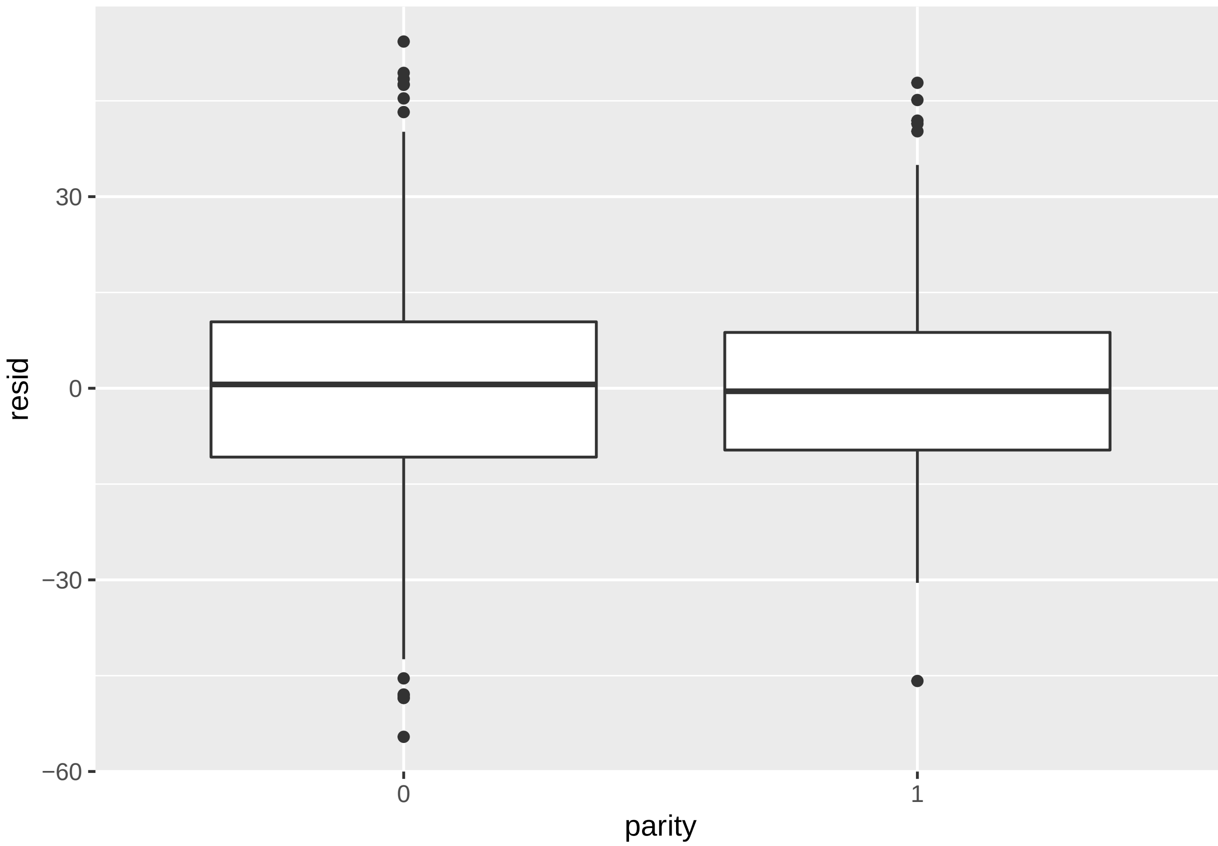

```
baby %>%
  ggplot(aes(x=weight,y=resid))+
  geom_point()+
  geom_hline(yintercept=0)
```

```
## Warning: Removed 52 rows containing missing values (geom_point).
```




```
baby %>%
  ggplot(aes(x=factor(parity),y=resid))+
  geom_boxplot()+
  xlab("parity")
```

```
## Warning: Removed 52 rows containing non-finite values (stat_boxplot).
```



Therefore, from the diagrams above, we can see that all the conditions are met: from the qqplot we can see the residual nearly follows normality and from residual plot we can see the variability is constant, so we can use a linear regression.

8. Obtain a 95% confidence interval for the coefficient on `gestation` and `age`. Interpret both confidence intervals.

```
baby_model %>%
  summary()
```

```
##
## Call:
## lm(formula = bwt ~ gestation + age + height + weight + parity,
##     data = baby)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.569 -10.506   0.453  10.063  54.285
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -86.59435   14.75733  -5.868 5.73e-09 ***
## gestation     0.46579    0.02992  15.568  < 2e-16 ***
## age           0.05233    0.08820   0.593   0.5531
## height        1.04614    0.21064   4.967 7.82e-07 ***
## weight        0.06564    0.02584   2.540   0.0112 *
## parity       -2.96394    1.16403  -2.546   0.0110 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.35 on 1178 degrees of freedom
##   (52 observations deleted due to missingness)
## Multiple R-squared:  0.2106, Adjusted R-squared:  0.2072
## F-statistic: 62.84 on 5 and 1178 DF,  p-value: < 2.2e-16
```

```
z_cri <- qnorm(0.025)
l1 <- 0.46579 + z_cri*0.02992
r1 <- 0.46579 - z_cri*0.02992
l1
```

```
## [1] 0.4071479
```

```
r1
```

```
## [1] 0.5244321
```

```
l2 <- 0.05233 + z_cri*0.08820
r2 <- 0.05233 - z_cri*0.08820
l2
```

```
## [1] -0.1205388
```

```
r2
```

```
## [1] 0.2251988
```

The 95% confidence interval for coefficient on gestation is (0.4071479, 0.5244321), which means we are 95% confident that the coefficient on gestation is between 0.4071479 and 0.5244321.
The 95% confidence interval for coefficient on age is (-0.1205388, 0.2251988), which means we are 95% confident that the coefficient on age is between -0.1205388 and 0.2251988.