Homework_3

Tania Ommer

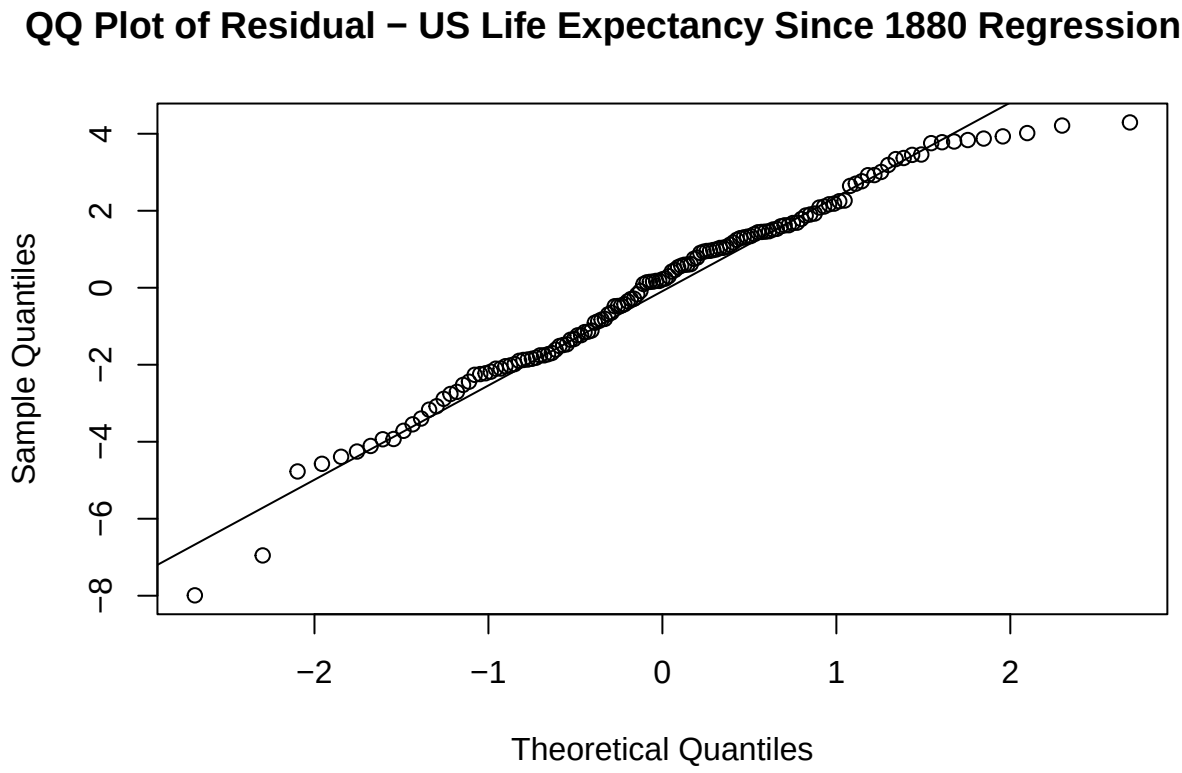
2023-10-03

Revisit the regression model of US life expectancy on number of years since 1880 in homework 2

1. Get the QQ plot of the residual

```
regression <- lm(life_expectancy ~ year, data = US)
residuals <- residuals(regression)

qqnorm(residuals, main="QQ Plot of Residual - US Life Expectancy Since 1880 Regression Model")
qqline(residuals)
```



2. Together with histogram of residual and scatter plot of residual vs. x, check the four assumptions in regression (LINE property).

```
residual_plot <- ggplot(data = data.frame(Year = US$year, Residual = residuals), aes(x = Year, y = Resi
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(
    title = "Scatter Plot of Residuals",
    x = "Number of Years Since 1880",
```

```
    y = "Residuals"
  )

histogram_plot <- ggplot(data = data.frame(Residual = residuals), aes(x = Residual)) +
  geom_histogram(binwidth = 1, color = "white") +
  labs(
    title = "Histogram of Residuals",
    x = "Residuals",
    y = "Frequency"
  )

grid.arrange(residual_plot, histogram_plot, ncol = 2)
```

The four assumptions of regression are linearity, independence, normality, and equal variance (LINE).

Linearity: The distribution of points in the scatterplot above can let us determine if the linearity assumption is met. The points appear to follow an inverted U-shaped pattern and are not randomly scattered around zero. Therefore, the linearity assumption is not met, and the regression model used is not the best fit for the life expectancy data.

Independece: The independence assumption for the plots above focus on the residuals and whether or not they are independent of each other. The residual data does not meet the independence assumption, as there is a systematic pattern in the scatter plot of points, meaning there is some sort of correlation and dependence between the observed and predicted values.

Normality: The histogram of residuals does not follow a normal distribution, as it is slightly skewed to the left and not symmetric. This, paired with the lack of randomness of points in the scatter plot of residuals, allows us to conclude that the normality assumption is not met.

Equal Variance: For the equal variance assumption to be met, the points in the residual scatter plot must be dispersed randomly throughout the graph, not following any distinct pattern. However, the points in scatter plot of residuals are not evenly spread throughout the graph, displaying unequal variance and not meeting the assumption.

Galton's height data

```
height$Height[height$Gender == "F"] <- height$Height[height$Gender == "F"] * 1.08
height$Mother <- height$Mother * 1.08
height$Midparent <- (height$Father + height$Mother)/2
```

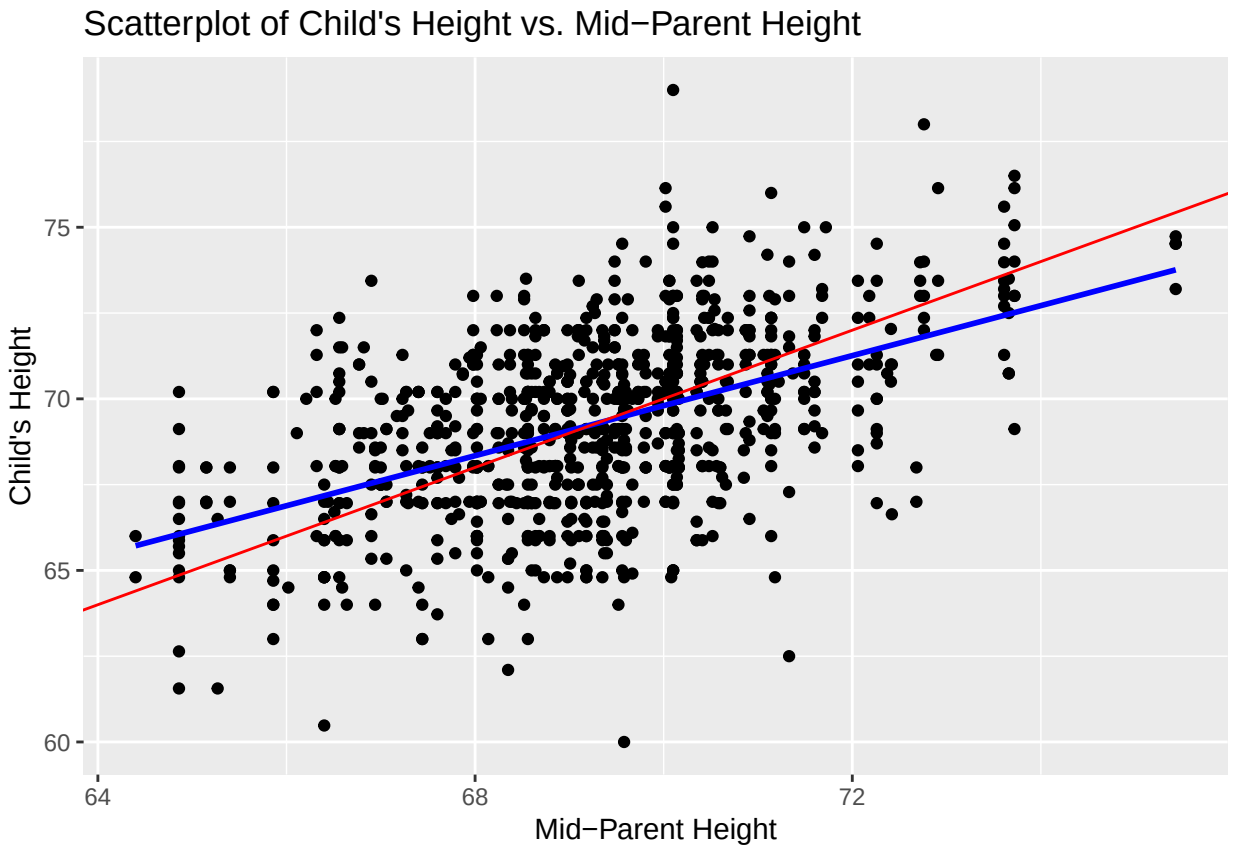
Regression of child's height (gender adjusted) on mid-height of parent

3. Have a scatterplot of y vs. x. On top of scatterplot, add the regression line and the diagonal line y=x, with different colors

```
scatterplot <- ggplot(data = height, aes(x = Midparent, y = Height)) +
  geom_point() + geom_smooth(method = "lm", color = "blue", se = FALSE) +
  geom_abline(intercept = 0, slope = 1, color = "red") +
  labs(
    title = "Scatterplot of Child's Height vs. Mid-Parent Height",
    x = "Mid-Parent Height",
    y = "Child's Height"
  )

scatterplot
```

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



In the scatterplot above, the blue line is the regression line and the red line is the diagonal line y=x.

4. What is the average children's height in the data? What is the average mid-height of the parent?

```
avg_child_height <- mean(height$Height)
avg_midheight_parent <- mean(height$Midparent)

cat("Average Children's Height:", avg_child_height, "inches\n")
```

```
## Average Children's Height: 69.23371 inches
```

```
cat("Average Mid-Height of Parent:", avg_midheight_parent, "inches")
```

```
## Average Mid-Height of Parent: 69.22201 inches
```

5. Among parents whose mid-height between 72 and 73 inches, what is the average height of their children?

```
new_height <- height[height$Midparent > 72 & height$Midparent < 73, ]
new_avg_child <- mean(new_height$Height)
cat("Average Height of Children with Parents' Mid-Height between 72-73 inches:", new_avg_child, "inches
```

```
## Average Height of Children with Parents' Mid-Height between 72-73 inches: 71.3178 inches
```

6. Run regression, is the model significant?

```
reg_model <- lm(Height ~ Midparent, data = height)
summary(reg_model)
```

```
##
## Call:
## lm(formula = Height ~ Midparent, data = height)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4947 -1.4779  0.0995  1.5175  9.1262
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.76698    2.84062   6.607 6.74e-11 ***
## Midparent    0.72906    0.04102  17.772  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.233 on 896 degrees of freedom
## Multiple R-squared:  0.2606, Adjusted R-squared:  0.2598
## F-statistic: 315.9 on 1 and 896 DF,  p-value: < 2.2e-16
```

```
get_regression_table(reg_model)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   18.8       2.84       6.61       0   13.2      24.3
## 2 Midparent    0.729     0.041     17.8        0    0.649     0.81
```

To determine if the regression model is significant, we can look at the p-value for Midparent (the mid-height of both parents) in the summary and regression table. The summary displays p-value: $< 2.2e\text{-}16$ and in the regression table the p-value is listed as 0. Since the p-vale for Midparent is extremely small and less than the default significance level of 0.05, we can determine that the model is significant and a good fit for the data.

7. If the parents' mid-height increases by 1 inch, what is the expected increase in child's height? Is the expected increase larger or smaller than 1 inch?

```
coef_midparent <- coef(reg_model)["Midparent"]
cat("Expected increase in child's height for 1-inch increase in parents' mid-height:", coef_midparent,
```

```
## Expected increase in child's height for 1-inch increase in parents' mid-height: 0.7290562 inches
```

The expected increase in child's height for a 1-inch increase in parents' mid-height can be found in the regression model summary, under the Coefficients section. The coefficient value listed for Midparent is the expected increase in child's height for every one inch increase in parents' mid-height, which is 0.73 inches. Therefore, the expected increase is smaller than 1 inch.

8. Estimate the child's height if the mid-height of parent is 64, 68, 70, 72, 76 respectively, and check their "closeness" to the mean height of all children

```
new_mid_height <- c(64, 68, 70, 72, 76)
new_height <- data.frame(Midparent = new_mid_height)

estimate <- predict(reg_model, newdata = new_height)
mean_height_child <- mean(height$Height)
difference <- abs(estimate - mean_height_child)

results <- data.frame(
  Mid_Height_Parent = new_mid_height,
  Estimated_Height = estimate,
  Closeness_to_Mean = difference
)

results
```

```
##   Mid_Height_Parent Estimated_Height Closeness_to_Mean
## 1                64         65.42658         3.8071363
## 2                68         68.34280         0.8909115
## 3                70         69.80092         0.5672008
## 4                72         71.25903         2.0253132
## 5                76         74.17525         4.9415380
```

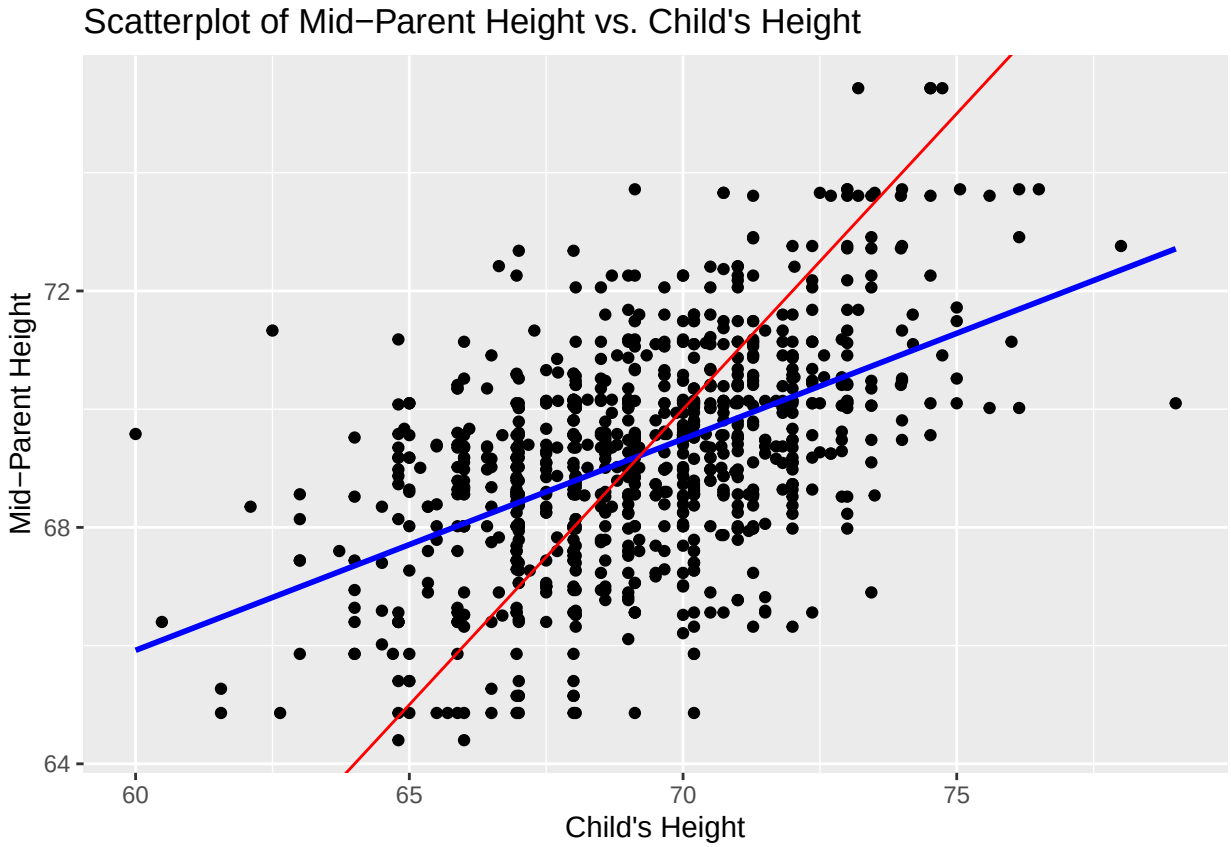
Regression of mid-height of parent on child's height (gender adjusted)

9. Have a scatterplot of y vs. x. On top of scatterplot, add the regression line and the diagonal line y=x, with different colors

```
scatterplot <- ggplot(data = height, aes(x = Height, y = Midparent)) +
  geom_point() + geom_smooth(method = "lm", color = "blue", se = FALSE) +
  geom_abline(intercept = 0, slope = 1, color = "red") +
  labs(
    title = "Scatterplot of Mid-Parent Height vs. Child's Height",
    x = "Child's Height",
    y = "Mid-Parent Height"
  )

scatterplot
```

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



In the scatterplot above, the blue line is the regression line and the red line is the diagonal line y=x.

10. Among all children with height between 72 and 73 inches, what is the mean mid-height of their parents?

```r
new_height_2 <- height[height$Height >= 72 & height$Height <= 73, ]
new_mean_parent <- mean(new_height_2$Midparent)
cat("Mean Mid-Height of Parents with Children's Height between 72-73 inches:", new_mean_parent, "inches
```

```
## Mean Mid-Height of Parents with Children's Height between 72-73 inches: 70.29489 inches
```

11. Run regression, is the model significant?

```r
reg_model_2 <- lm(Midparent ~ Height, data = height)
summary(reg_model_2)
```

```
##
## Call:
## lm(formula = Midparent ~ Height, data = height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7075 -1.0556  0.0547  1.0163  4.7901
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 44.47115    1.39365   31.91   <2e-16 ***
## Height       0.35750    0.02012   17.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.564 on 896 degrees of freedom
## Multiple R-squared:  0.2606, Adjusted R-squared:  0.2598
## F-statistic: 315.9 on 1 and 896 DF,  p-value: < 2.2e-16
```

```r
get_regression_table(reg_model_2)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   44.5        1.39      31.9       0   41.7     47.2
## 2 Height       0.357      0.02      17.8       0    0.318    0.397
```

To determine if the regression model is significant, we can look at the p-value for Midparent (the mid-height of both parents) in the summary and regression table. The summary displays p-value: $< 2.2e\text{-}16$ and in the regression table the p-value is listed as 0. Since the p-vale for Midparent is extremely small and less than the default significance level of 0.05, we can determine that the model is significant and a good fit for the data.

12. If the child's height increases by 1 inch, what is the expected increase in parent's mid-height? Is the expected increase larger or smaller than 1 inch?

```r
coef_child <- coef(reg_model_2)["Height"]
cat("Expected increase in Parents' Mid-Height for 1-inch increase in child's height:", coef_child, "inch
```

```
## Expected increase in Parents' Mid-Height for 1-inch increase in child's height: 0.3574971 inches
```

The expected increase in Parents' mid-height for a 1-inch increase in child's height can be found in the regression model summary, under the Coefficients section. The coefficient value listed for Height is the expected increase in parents' mid-height for every one inch increase in child's height, which is about 0.36 inches. Therefore, the expected increase is smaller than 1 inch.

13. Estimate the parent's mid-height if the child's height is 64, 68, 70, 72, 76 respectively, and check their "closeness" to the mean mid-height of all parents

```r
new_child_height <- c(64, 68, 70, 72, 76)
new_height_2 <- data.frame(Height = new_child_height)

estimate_2 <- predict(reg_model_2, newdata = new_height_2)
mean_mid_height <- mean(height$Midparent)
difference_2 <- abs(estimate_2 - mean_mid_height)

results_2 <- data.frame(
  Child_Height = new_child_height,
  Estimated_Height = estimate_2,
  Closeness_to_Mean = difference_2
)

results_2
```

```
##   Child_Height Estimated_Height Closeness_to_Mean
## 1           64         67.35097         1.8710379
## 2           68         68.78096         0.4410495
## 3           70         69.49595         0.2739447
## 4           72         70.21095         0.9889389
## 5           76         71.64093         2.4189273
```

14. Use the above results to explain regression to the mean.

Regression to the mean is a tendancy for extreme values to become less extreme when followed by additional measurements, taken after the initial observation. In the data set of parents and children's heights, some parents may have mid-heights that are relatively larger or smaller than the mean mid-height (outlier mid-heights that are very short or tall). These outlier mid-heights can be explained by random variation, genetics, or errors in measurement. If you were to take an initial measurement of children's heights and compared it to a second measurement of children's heights taken at a later time, you would observe that the the children's heights in the second measurement are closer to the overall average height of children (population mean) compared to the initial measurement. This occurs because very tall parents are more likely to have children who are shorter than they are, while very short parents are more likely to have children who are taller than they are, with the resulting children's height being closer to the overall population mean.