Stat21 HW3

Dohyun Lee

Problem 1

In class, we said that if the standard regression model assumptions are satisfied, then the least-squares line passes through the conditional means. Sketch a picture of a dataset in which not all of the standard regression model assumptions are satisfied, and the least-squares line does not pass through all the conditional means of the Y-values. Explain briefly or indicate on your picture where the LS line misses the conditional mean.

Problem 2

In a survey of 988 men aged 18–24, the regression equation for predicting height from weight was height in inches = 62.4 + (0.047)(weight in pounds) Is the following statement a correct interpretation of the regression line: "If someone puts on 10 pounds, he will get taller by (0.047)(10) = 0.47 inches"? If not, explain what the slope means.

If someone puts on weight then that doesn't mean that they would grow in height by whatever amount. The correct interpretation would be that for every inch taller that a man is, you can expect them to weigh 0.047 lbs more – 0.047 is the slope and this is the amount that the response variable would change for every 1 unit change in the explanatory variable.

Problem 3

In the dataset above, suppose the conditional SD of Y is s = 2.2 inches. What percentage of all 200-pound men are taller than 74 inches? (Assume the regression model assumptions are met.)

```
meanHeight <- 62.4 + (0.047*200) #mean height predicted from regression model
meanHeight
```

```
## [1] 71.8
```

```
condSD <- 2.2
height <- 74

heightDiff <- height - meanHeight
heightDiff
```

```
## [1] 2.2
```

```
testStat <- heightDiff/condSD
testStat
```

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

```
p_value <- 1 - pnorm(testStat)
p_value
```

```
## [1] 0.1586553
```

15.9% of all 200-lbs men are above 74 inches.

Problem 4

Suppose we have two datasets each having n observations: dataset 1 with variables X1 and Y1, and dataset 2 with variables X2 and Y2. (X1 and X2 are the explanatory variables, and Y1 and Y2 are the response variables.) There is a positive linear relationship between X1 and Y1, and a positive linear relationship between X2 and Y2. Suppose Y1 has marginal standard deviation 10 and conditional standard deviation 8, and Y2 has marginal standard deviation 10 and conditional standard deviation 2. For each of Y1 and Y2, assume the conditional standard deviation is constant. Which one of the following statements must be true? Explain briefly or draw a picture.

(a) The correlation between X1 and Y1 equals the correlation between X2 and Y2.

(b) The correlation between X1 and Y1 is greater than the correlation between X2 and Y2.

(c) The correlation between X1 and Y1 is less than the correlation between X2 and Y2.

(d) None of the above statements can be determined from the information given.

c is the only correct option because the conditional SD for dataset 1 is bigger than dataset 2. Their marginal SDs are the same, so when we look at the conditional SDs, this means that at any given x value, the spread is bigger for dataset 1 than 2, meaning on a fitted line, the points are farther from the line, which means that there is a lower correlation between the X1 and Y1 variables.

Problem 5

Is the line drawn in the scatterplot at right the regression line? Why or why not? Explain briefly. You may assume that the standard regression model assumptions are satisfied.

The line in the scatterplot is not at the right position to be a regression line because at every given x value, the residuals aren't uniform and the line doesn't pass through the mean of the points at each x. The regression line seems to overestimate every low value at the low end and underestimate every high value of the high end.

Skyscrapers (18 points)

Problem 6

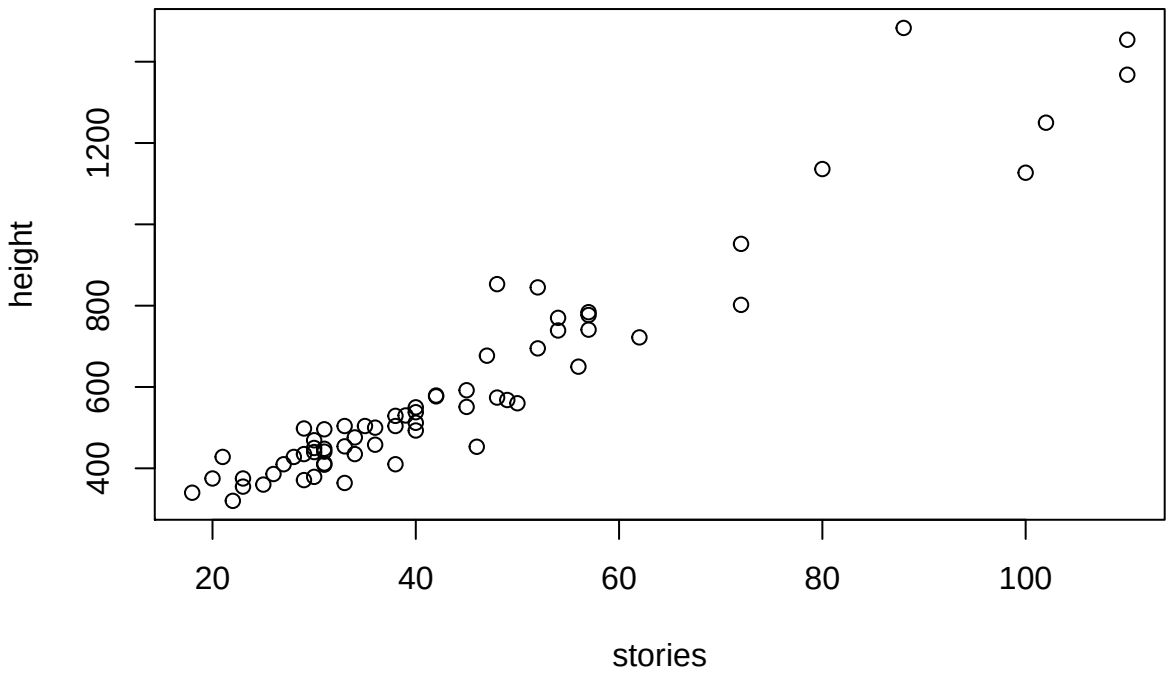
How does the height (Y) of a skyscraper depend on the number of stories it has (x)? Sixty-five buildings were selected at random from a table of US tall buildings given in the World Almanac, and their heights and number of stories were recorded. A few other selected skyscrapers were added to the dataset. The data are on Moodle. You can read in the data using the following commands:

```
# read in dataset
data <- read.csv("/Users/Dohyun/Desktop/school stuff/year2/spring20/stat21/skyscrapers.csv")

# define variables
height <- data[,"height"]
stories <- data[,"stories"]
year <- data[,"year"]
building <- data[,"building"]
```

(a) How does the height of a tall building depend on the number of stories it has? To explore this question, make a scatterplot of height vs stories using the plot command. (Height should be on the Y axis.) Briefly describe the relationship between height and stories: Does it appear to be linear? Is the relationship a strong one? Copy the scatterplot and hand it in (paste it into your write-up).

```
plot(stories, height)
```



There appears to be a strong, linear relationship between stories and height.

(b) Calculate the correlation of height and stories using the cor command. What is the value of the correlation coefficient. Would you say this is a strong correlation, medium, or low?

```
corr <- cor(stories, height)
corr
```

```
## [1] 0.9556949
```

The correlation comes out to be a strong 0.96.

(c) Calculate the regression of height and stories using the following commands:

```
fit1 <- lm(height ~ stories)
summary(fit1)
```

```
##
## Call:
## lm(formula = height ~ stories)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -168.66  -41.48    0.81   24.10  353.23
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65.1574    23.0423   2.828  0.00628 **
## stories      12.0979     0.4695  25.770  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 80.05 on 63 degrees of freedom
## Multiple R-squared:  0.9134, Adjusted R-squared:  0.912
## F-statistic: 664.1 on 1 and 63 DF,  p-value: < 2.2e-16
```

Copy the output and hand it in. What are the equation of the regression line, the value of the conditional SD of height, and the value of R-squared?

The regression line is: height = 65.16 + 12.1(stories). The conditional SD is:

```
80/(63-2)
```

```
## [1] 1.311475
```

and the R-squared is 0.9134

(d) Calculate a 95% confidence interval for ß1. How would you explain the meaning of this confidence interval, in the specific context of this dataset, to an architect who has not taken statistics?

```
t <- qt(0.025, 61)
12.1 + c(t,-t)*0.4695
```

```
## [1] 11.16118 13.03882
```

We are 95% confident that the slope of the regression ranges from 11.16 to 13.04. This means that for every 1 unit increase in building stories, the building height will increase by any number ranging from 11.16 to 13.04.

(e) Using the output from (c), test the hypothesis that ß1 = 0. State your null and alternative hypotheses and report the test statistic and p-value; circle or highlight the relevant part of the output from (c). Is your conclusion surprising?

$H_0 : \beta_1 = 0$ $H_A : \beta_1 \neq 0$

```
testStat <- 12.0979/0.4695
testStat #this is our test statistic
```

```
## [1] 25.76763
```

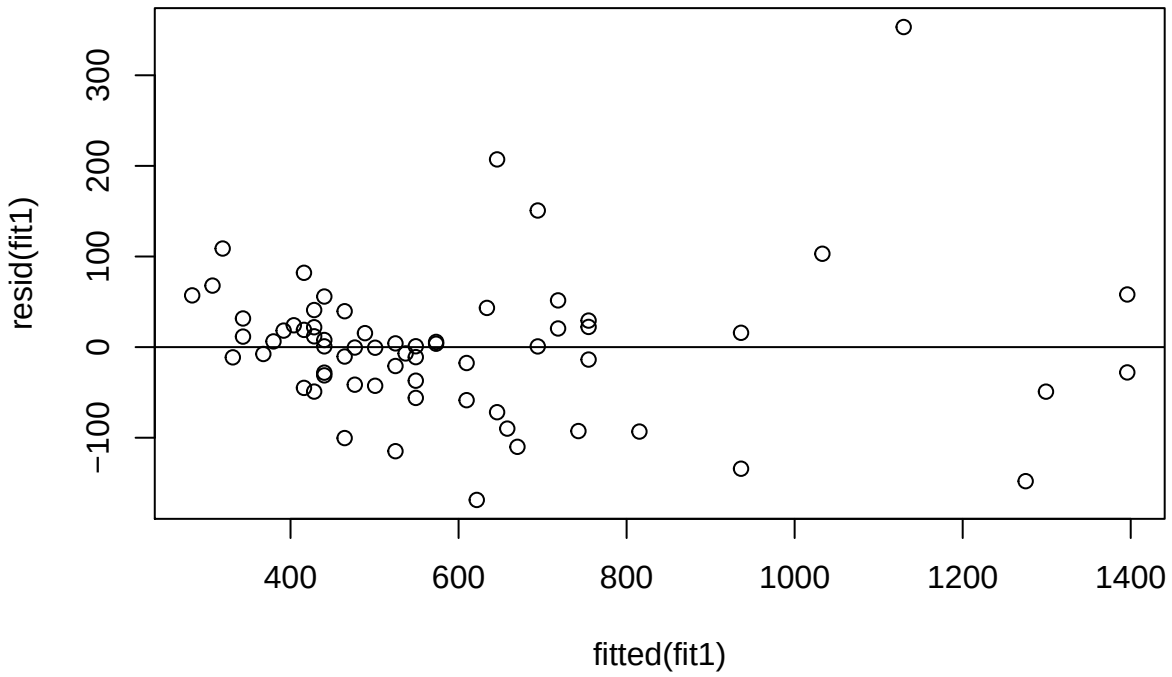
```
p_value <- 2*pt(-testStat, 61)
p_value #this is our p-value
```

```
## [1] 1.732765e-34
```

Our p-value is less than 0.05, which means we reject H_0 – the slope is not zero. This conclusion makes sense because our confidence interval does not include zero.

(f) Make a residual plot for the regression of height and stories using the following commands:

```
plot(fitted(fit1), resid(fit1))
abline(h=0)
```

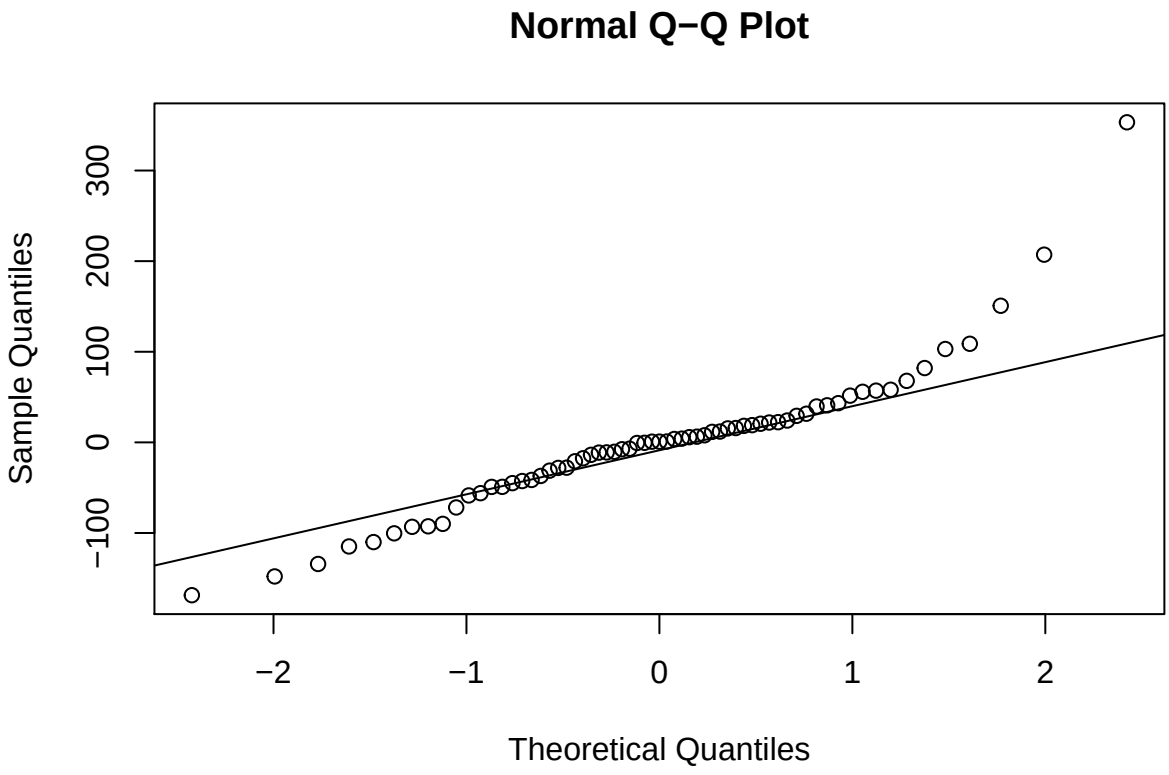


Copy the residual plot and hand it in. Are there any apparent violations of the regression model assumptions? Explain briefly.

There doesn't seem to be any clear violations of the model assumptions. The spread of the plot doesn't appear to take on any specific shape and seems to be randomly spread out.

(g) Make a Normal probability plot of the residuals using the following commands:

```
qqnorm(resid(fit1))
qqline(resid(fit1))
```



Copy the NPP and hand it in. Are there any apparent violations of the regression model assumptions? Explain briefly.

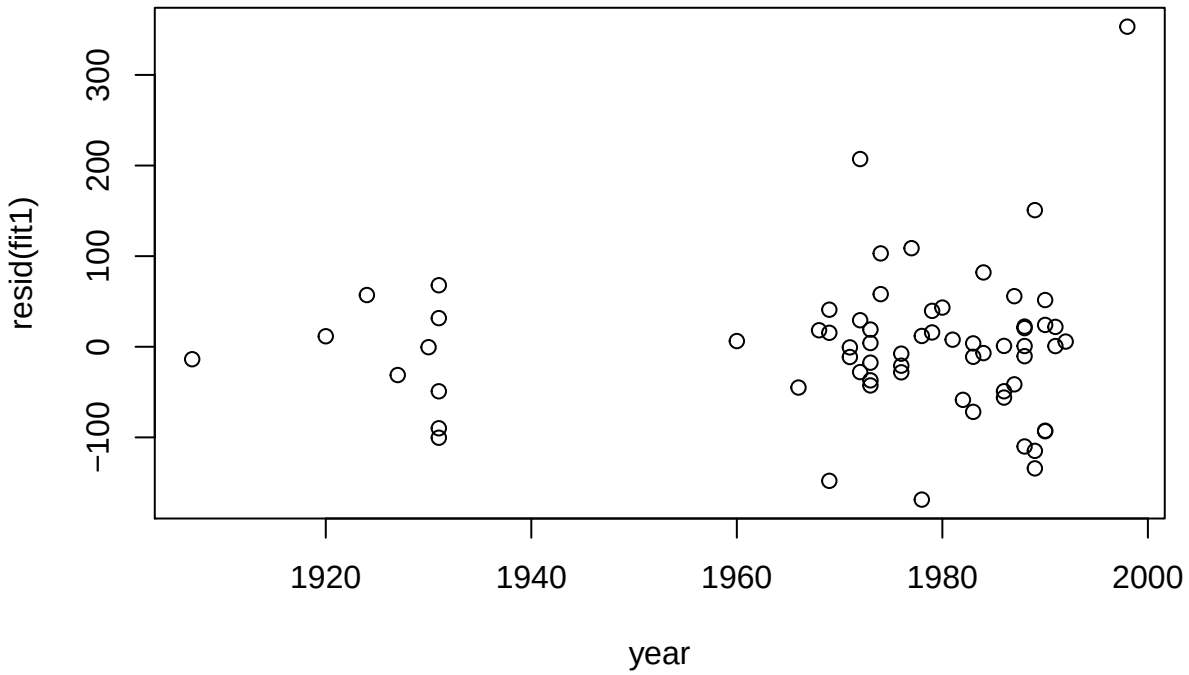
There is a violation of the regression model assumption. The points on the high end are above the line and the points on the low end are below the line, diverging from the line. This means that the conditional standard deviation for Y at any given X is not normal, which violates the assumption of having normal residuals.

(h) Have there been any trends in the height/story relationship over time? Make a scatterplot of the residuals vs year and hand it in. Are there any patterns over time or other notable features?

```
plot(year, stories)
```


```
plot(year, building)
```

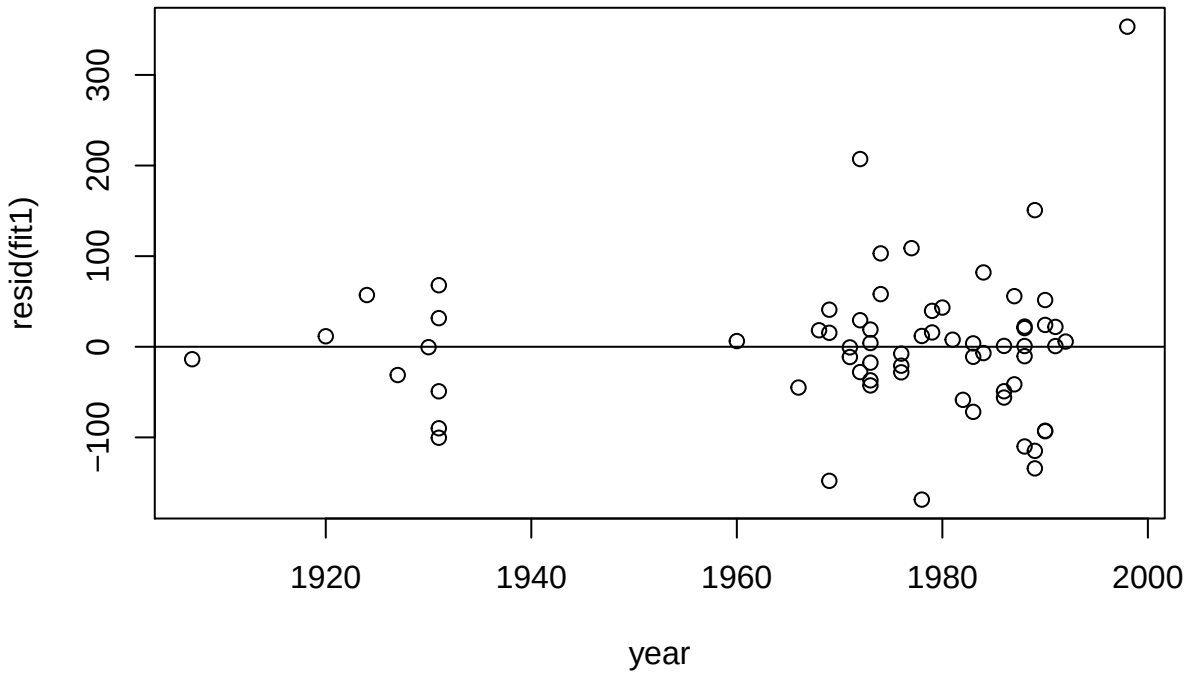

```
plot(year, resid(fit1))
```



There is a noticeable gap between what seems to be the late 1930s and 1960. This could tell us that during World War II, steel and concrete were mainly distributed to the military for warfare, rather than infrastructure. Then from 1960 onward, there is a huge boom in the number of buildings built. This makes sense as urban sprawl and the rise of corporations could have inspired the need for more and higher buildings. Variance between height and story seems to also be evenly spread out, which tells us that over time architecture has been diversified for practical or stylistic reasons.

(i) Are there any buildings that seem to be unusually high for their number of stories? If so, which building(s)? You can identify points on the residual plot using the following commands:

```
plot(year, resid(fit1))
abline(h=0)
identify(year, resid(fit1), labels=building, cex=.6)
```




```
## integer(0)
```

Liberty Place, the Transmaerica Pyramid, and the Petronas Towers have unusually high heights given their number of stories. This doesn't take antenna height into account and both these buildings have tapered tops, which add to their height.