Regression analysis_Homework Assignment 4

心理所碩二 **R08227112** 林子堯

2020/11/02

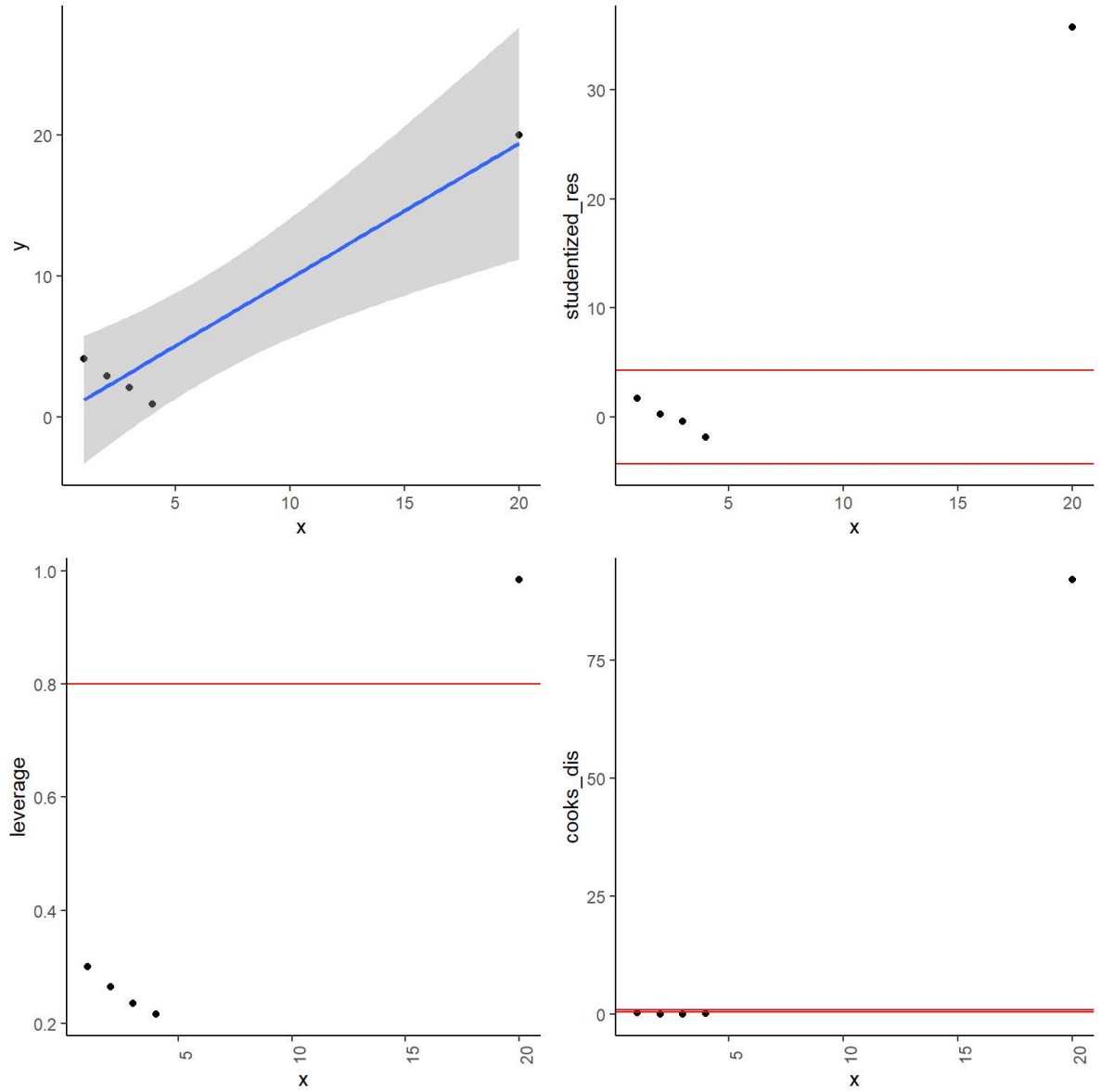
1. Fit a line by least squares to the following points: $(4, 0.9), (3, 2.1), (2, 2.9), (1, 4.1)$ and $(20, 20)$. Obtain Studentized residuals and also plot the points and the estimated line. Does the point $(20, 20)$ appear as an outlier? Using a suitable indicator variable, numerically demonstrate the indicator variable interpretation of RSTUDENT's. Also demonstrate that DFFIT and the DFBETA's do indeed measure what has been claimed for them.

```
library(tidyverse)
library(grid)
library(gridExtra)
library(ggrepel)
```

```
data1 ← data.frame(y = c(0.9, 2.1, 2.9, 4.1, 20),
                   x = c(4, 3, 2, 1, 20))

# define my custom function for part 1.
getLM1 ← function(data){
  data_lm ← lm(y ~ 1 + x, data)
  data ← add_column(data,
                    studentized_res = rstudent(data_lm),
                    leverage = hatvalues(data_lm),
                    cooks_dis = cooks.distance(data_lm))

  g1 ← ggplot(data, aes(x = x, y = y)) +
    geom_point() + geom_smooth(method = "lm") + theme_classic()
  g2 ← ggplot(data, aes(x = x, y = studentized_res)) +
    geom_point() +
    geom_hline(yintercept = qt(0.05/2, data_lm$df.residual - 1), color = "red") +
    geom_hline(yintercept = qt((1-0.05/2), data_lm$df.residual - 1), color = "red") +
    theme_classic()
  g3 ← ggplot(data, aes(x = x, y = leverage)) +
    geom_point() +
    geom_hline(yintercept = 2*length(data_lm$coef)/nrow(data), color = "red") + # 2p/n
    theme_classic() +
    theme(axis.text.x = element_text(angle = 90))
  g4 ← ggplot(data, aes(x = x, y = cooks_dis)) +
    geom_point() +
    geom_hline(yintercept = c(0.5, 1), color = "red") +
    theme_classic() +
    theme(axis.text.x = element_text(angle = 90))

  plot = arrangeGrob(g1, g2, g3, g4, nrow = 2)

  return(list(data = data, lm = data_lm, plot = plot))
}

data1 ← getLM1(data1)
summary(data1$lm)
```

```
Call:
lm(formula = y ~ 1 + x, data = data)

Residuals:
      1       2       3       4       5
-3.1816 -1.0224  0.7368  2.8960  0.5712

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2448     1.5303   0.160   0.8831
x             0.9592     0.1650   5.813   0.0101 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.609 on 3 degrees of freedom
Multiple R-squared:  0.9184,    Adjusted R-squared:  0.8913
F-statistic: 33.79 on 1 and 3 DF,  p-value: 0.01014
```

```
grid.draw(data1$plot)
```



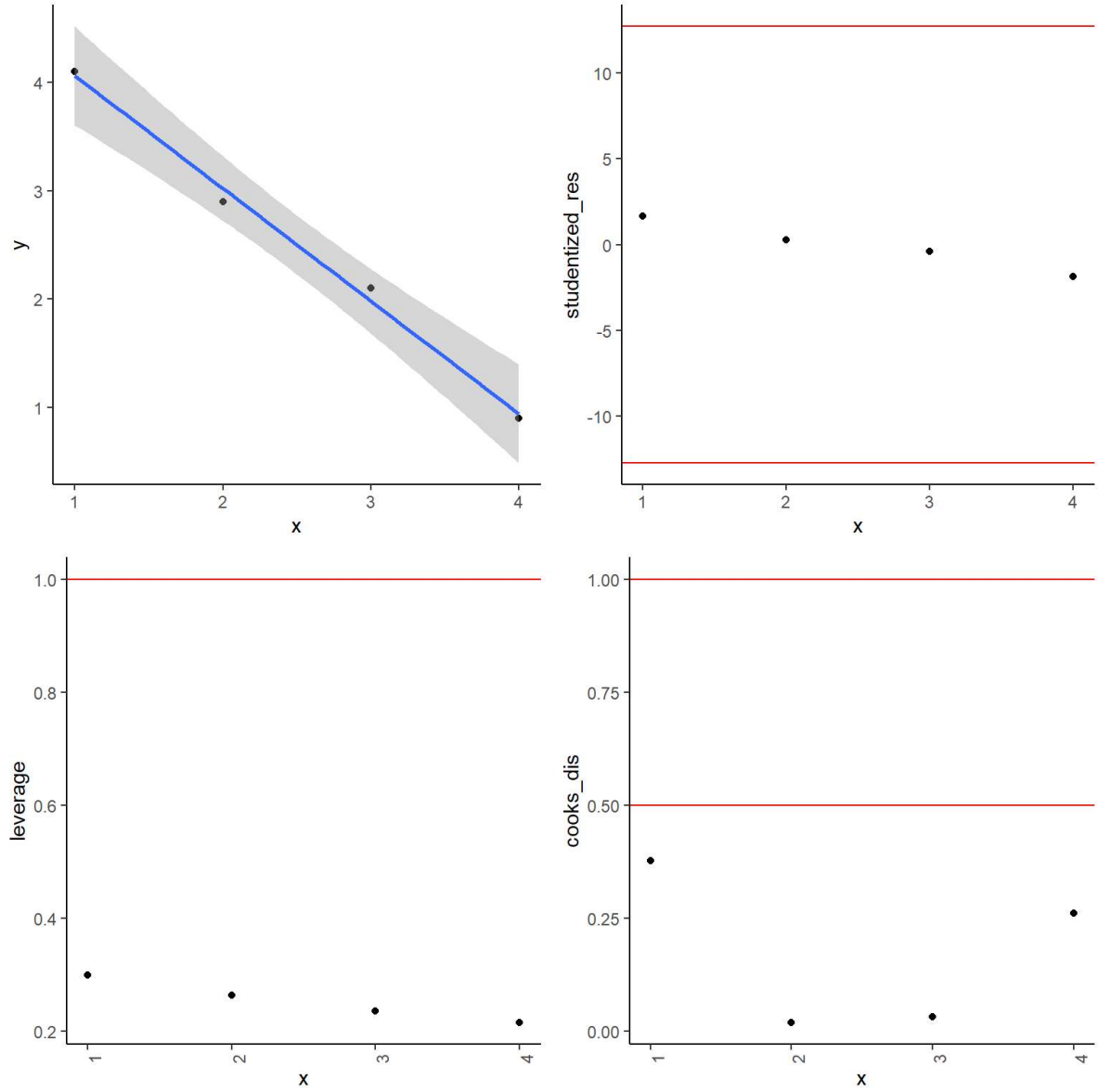
The pattern of first foue point is decreasing, however, the last point is on the opposite direction. From the studentized residual plot, we can exactly observe that the point $(20, 20)$ is an outlier. Furthermore, the leverage and Cook's distance of $(20, 20)$ is too large. It shows that the last point impacts the regression line largely.

If we remove that point, the result is

```
data1_rmOutlier ← data1$data %>%
  filter(studentized_res > qt(0.05/2, data1$lm$df.residual - 1),
         studentized_res < qt((1-0.05/2), data1$lm$df.residual - 1))

data1_rmOutlier ← getLM1(data1_rmOutlier)
summary(data1_rmOutlier$lm)
```

```
Call:
lm(formula = y ~ 1 + x, data = data)

Residuals:
    1     2     3     4
-0.04  0.12 -0.12  0.04

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.10000    0.15492   32.92 0.000921 ***
x           -1.04000    0.05657  -18.39 0.002946 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1265 on 2 degrees of freedom
Multiple R-squared:  0.9941,    Adjusted R-squared:  0.9912
F-statistic:   338 on 1 and 2 DF,  p-value: 0.002946
```

```
grid.draw(data1_rmOutlier$plot)
```



We can find the relation of x and y are negative correlation and the regression line is fitted perfectly. And the DFFIT and the DFBETA's are as follows

```
DFFIT ← predict(data1$lm, data.frame(x = c(1:4))) - predict(data1_rmOutlier$lm, data.fr
ame(x = c(1:4)))
DFFIT
```

```
      1       2       3       4
-2.8560 -0.8568  1.1424  3.1416
```

```
DFBETA ← coef(data1$lm) - coef(data1_rmOutlier$lm)
DFBETA
```

```
(Intercept)           x
    -4.8552      1.9992
```

There are large different in DFFIT and DFBETA's. So we can convince that it is not appropriate to use simple linear regression to fit all five points. Maybe one can choose quadratic line to fit this pattern or remove the problem point then use linear regression.

2. Consider the data in Exhibit 8.12 on male deaths per million in 1950 for lung cancer (y) and per capita cigarette consumption in 1930 (x).

(A discussion of part 1 of this example is contained in Thfte, 1974, p. 78 et seq. The data was used in some of the earlier reports on Smoking and Health by the Advisory Committee to the U.S. Surgeon General. See reference to Doll, 1955.)

a. Estimate a model expressing y as a linear function of x. Do any of the points look particularly influential? Delete the United States, rerun the model and check if the influences of Great Britain and Finland have been substantially altered. Now, put the U.S. back and delete Great Britain and examine the influence of points for the resultant model.

Our simple linear model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

```
data2 ← readxl::read_xlsx("Exhibit_8-12.xlsx")

getLM2 ← function(data){
  data_lm ← lm(y ~ 1 + x, data)

  data ← mutate(data,
                studentized_res = rstudent(data_lm),
                leverage = hatvalues(data_lm),
                cooks_dis = cooks.distance(data_lm))

  g1 ← ggplot(data, aes(x = x, y = y)) +
    geom_point() +
    geom_smooth(method = "lm") +
    geom_label_repel(aes(label = Country), box.padding = 0.5, size = 2) +
    theme_classic()
  g2 ← ggplot(data, aes(x = Country, y = leverage)) +
    geom_point() +
    geom_hline(yintercept = 2*length(data_lm$coef)/nrow(data), color = "red") + # 2p/
n
    theme_classic() +
    theme(axis.text.x = element_text(angle = 90))
  g3 ← ggplot(data, aes(x = Country, y = cooks_dis)) +
    geom_point() +
    geom_hline(yintercept = c(0.5, 1), color = "red") +
    theme_classic() +
    theme(axis.text.x = element_text(angle = 90))
  plot = arrangeGrob(g1, g2, g3, nrow = 1)

  return(list(data = data, lm = data_lm, plot = plot))
}

data2 ← getLM2(data2)
summary(data2$lm)
```

```

Call:
lm(formula = y ~ 1 + x, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-169.016  -32.813    0.004   45.804  136.914

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.74886   48.95871   1.343  0.21217
x            0.22912    0.06921   3.310  0.00908 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 84.13 on 9 degrees of freedom
Multiple R-squared:  0.549, Adjusted R-squared:  0.4989
F-statistic: 10.96 on 1 and 9 DF,  p-value: 0.009081
```

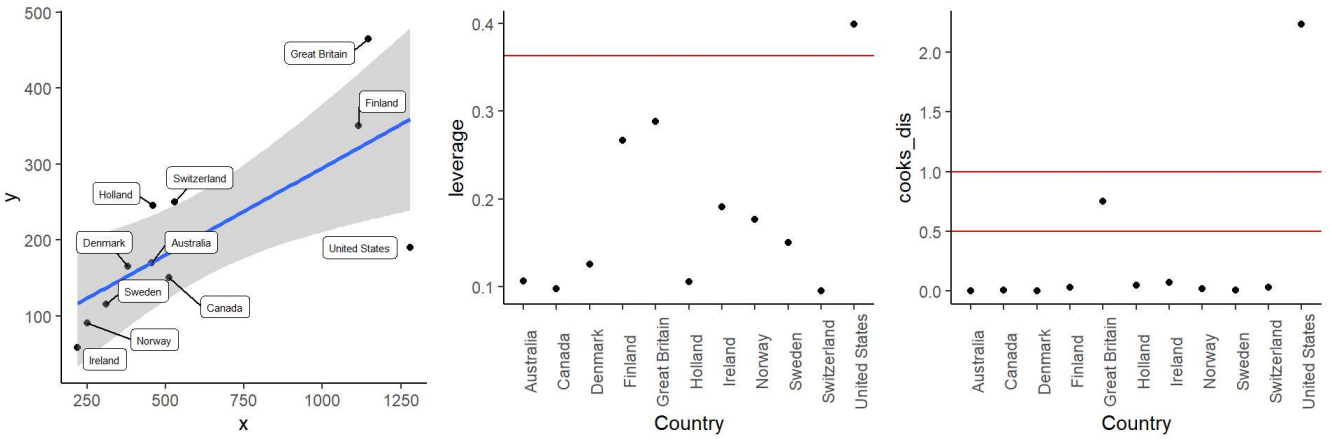
and the fitted result is

$$\hat{y}_i = 65.75 + 0.23x_i$$

We can find the male death rate is linear dependent with the per capita cigarette consumption. When consumption add 1 unit, the male death rate enhance 0.23 unit.

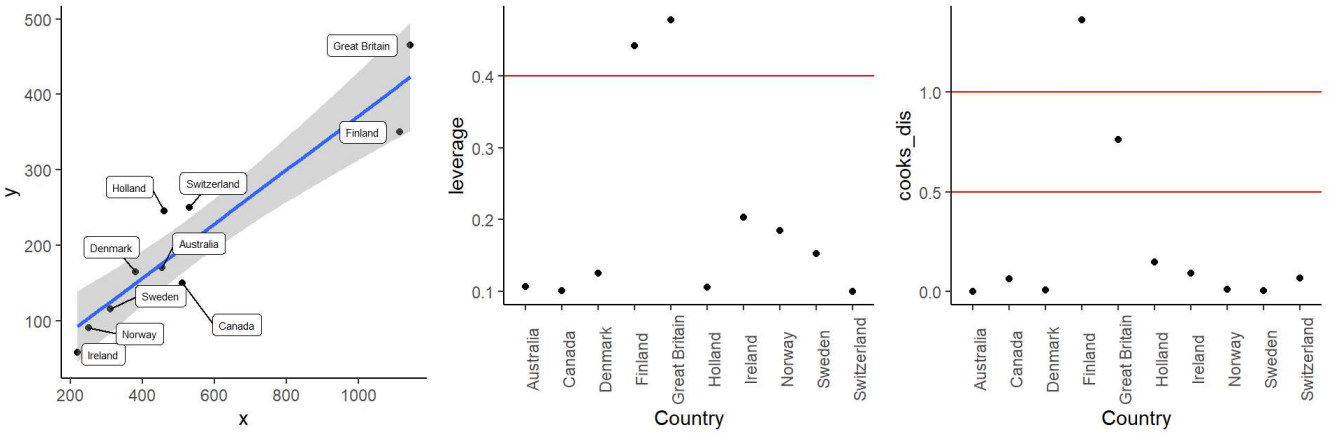
If we check is there any influential point by using "leverage" and "Cook's distance" in this data

```
grid.draw(data2$plot)
```



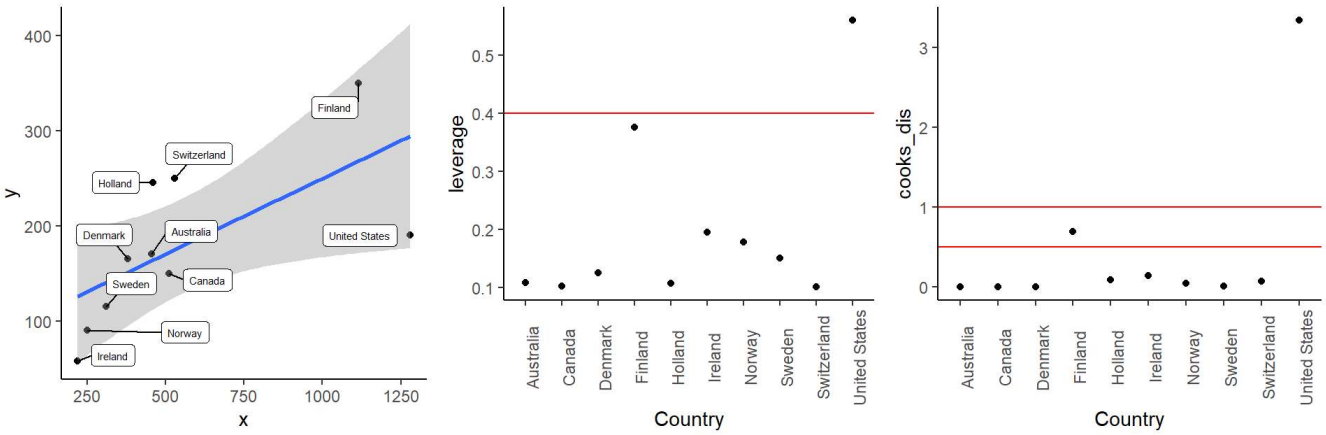
We can find (see middle and right plot) that the United States may be a highly influential observations since its leverage and Cook's distance are extremely high. If we remove the United States

```
data2_rmUS ← data2$data %>% filter(Country ≠ "United States")
data2_rmUS ← getLM2(data2_rmUS)
grid.draw(data2_rmUS$plot)
```



The influences of Great Britain and Finland have been substantially altered but are not as extremely high as the influences of United States. And, if we put the U.S. back again and delete Great Britain

```
data2_rmGB ← data2$data %>% filter(Country ≠ "Great Britain")
data2_rmGB ← getLM2(data2_rmGB)
grid.draw(data2_rmGB$plot)
```



We can find the influence of United Stats is still very high, but the influence of Finland is decreasing. In conclusion, the influence of point would effect to each others and the United States is the most influence point. Otherwise, this point may be a outlier, since it has the largest deviation form the fitted regression line.

b. Try an appropriate broken line regression and examine the residuals. If you notice heteroscedasticity, run an appropriately weighted regression. Examine the data points for outliers or undue influence.

The broken line regression model is

$$y_i = \alpha + \beta_1 \min(x_i - \theta, 0) + \beta_2 \max(x_i - \theta, 0) + \varepsilon_i$$

I use the `lm.br` package in R to automatically fit this data. Then the result is

```
library(lm.br)
data2_lmbr ← lm.br(y ~ 1 + x, data = data2$data)
data2_lmbr
```

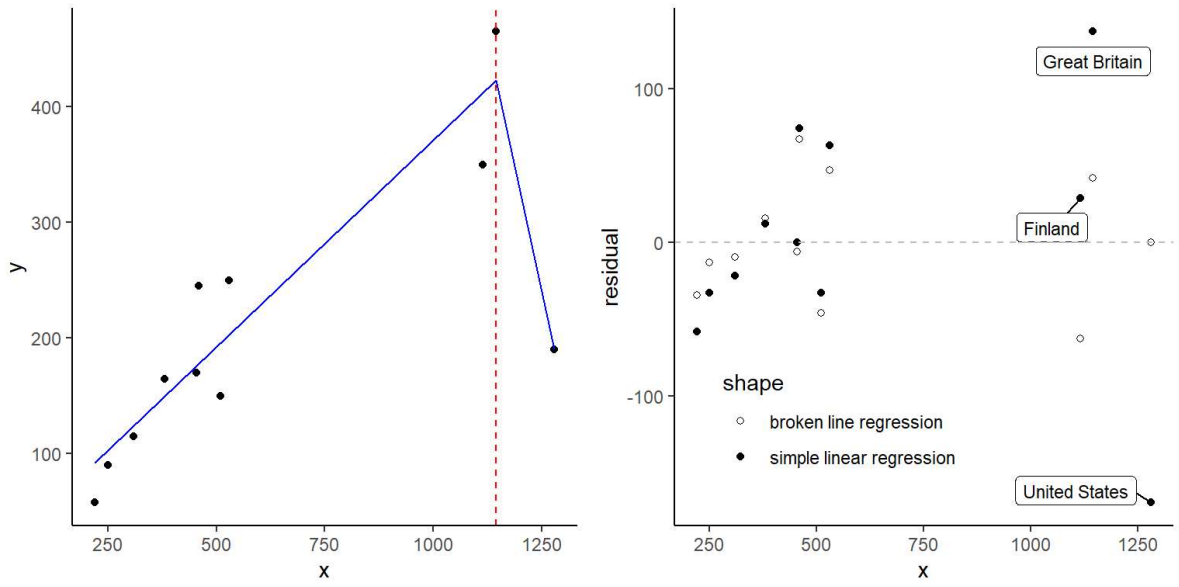
```
Call:
lm.br(formula = y ~ 1 + x, type = "LL", data = data2$data)

Changepoint and coefficients:
      theta       alpha   x < theta   x > theta
 1145.00000   423.08333     0.35767    -1.72654

Significance Level of H0:"no changepoint" vs H1:"one changepoint"
    SL= 0.0093749  for theta0 = 61  by method CLR

95-percent confidence interval for changepoint 'theta' by CLR
    [ 861.797, 1280 ]
```

$$\hat{y}_i = 423.08 + 0.38 \min(x_i - 1145, 0) - 1.73 \max(x_i -, 0)$$

The following are the broken line regression plot (left) and residual plot (right), which compares the (raw) residuals in the original simple linear regression and in the broken line regression.

Note: I only presented the raw residual plot, since the function `hatvalues()` *(calculating the leverages) and* `rstudent()` *(calculating studentized residual) can't use on the object which class is "lm.br" (the output result from the* `lm.br()`*). I don't know other methods to calculate the leverages and studentized residual from the broken line regression. QQ*

```
theta = data2_lmbr$coef["theta"]
alpha = data2_lmbr$coef["alpha"]
beta1 = data2_lmbr$coef["  x < theta"]
beta2 = data2_lmbr$coef["  x > theta"]

g1 ← ggplot(data2$data, aes(x = x, y = y)) +
  geom_point() +
  geom_vline(xintercept = theta, color = "red", linetype = "dashed") +
  geom_segment(aes(x = min(x), y = alpha+beta1*(min(x)-theta),
                   xend = theta, yend = alpha),
               color = "blue") +
  geom_segment(aes(x = theta, y = alpha,
                   xend = max(x), yend = alpha+beta2*(max(x)-theta)),
               color = "blue") +
  theme_classic()

data2$data ← data2$data %>%
  add_column(residual = residuals(data2$lm),
             residual_brokenLine = residuals(data2_lmbr))
potentialOutlier ← c("United States", "Great Britain", "Finland")
g2 ← ggplot(data2$data, aes(x = x)) +
  geom_point(aes(y = residual, shape = "simple linear regression")) +
  geom_point(aes(y = residual_brokenLine, shape = "broken line regression")) +
  geom_label_repel(aes(y = residual,
                       label = ifelse(Country %in% potentialOutlier, Country, "")),
                   box.padding = 0.5, size = 3) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "gray") +
  scale_shape_manual(values = c(1, 16)) +
  theme_classic() +
  theme(legend.position = c(0.33, 0.2))
grid.arrange(g1, g2, nrow = 1)
```



It seems like that the broken line regression fitted very well. The residual becomes smaller (hollow circles crossing to 0) than the orginial residual from simple linear regression (solid circles). It showed be noted that the data of the United Stats may be an outlier for the simple linear regression. But in the broken line regression, there is no obvious outlier or influence point. Furthermore, the variance of residual seems homoscedastic, so I propose that we don't need use weighted least squared method on the broken line regression.

c. Do you think a plausible reason for using broken line regression is that the number of women who smoke might be much higher in countries with high per capita cigarette consumption?

From the (a) and (b) results, we can find that when the cigarette consumption increases, the male death rate increases, but except for the male death rate in the United States is decreasing. The problem is that the independent variable x, per capita cigarette consumption, which contains the "male" and "female" buyer, but the dependent variable y is only

considered the male! It is possible that the more cigarette consumption, the more people died. However, the women who smoke in the United States is more than in other countries. So, the broken line regression may be suitable for this data. Beyond doubt, it is better for analysis if we can have each gender cigarette consumption and death rate data.

d. Write a report discussing your various efforts and your final conclusion.

We found data in the United States may be an influence point and outlier from the part (a) and (b), and the probable reason for casing this result from the part (c). Since the property of that data point, I suggest we can:

1. remote the outlier (the United States then fitted by simple linear regression;
2. use the broken line regression directly like part (b);
3. consider the missing variable (like each gender's death rate) for controlling gender effect; or
4. use the more proper variable (like per "male" cigarette consumption) as an independent variable in the original analysis.

3. Consider the model $y_i = \beta x_i + \varepsilon_i$, where ε_i's are independent and distributed as $N(0, \sigma^2 x_i^2)$. Find the weighted least squares estimator for β and its variance. Give reasons why you would not wish to use ordinary least squares in this case.

Let $Cov(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{W}^{-1}$ where $\boldsymbol{W^{-1}} = diag(x_1^2, \ldots, x_n^2)$.

We can transform the original model to the classical linear model

$$\begin{aligned} \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{y} &= \beta \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{x} + \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{\varepsilon} \\ \Rightarrow \quad \boldsymbol{y^*} &= \beta \boldsymbol{x^*} + \boldsymbol{\varepsilon^*} \end{aligned}$$

For the weighted least squares estimator we get

$$\begin{aligned} \hat{\beta} &= (\boldsymbol{x^*}^\top \boldsymbol{x^*})^{-1} \boldsymbol{x^*}^\top \boldsymbol{y^*} \\ &= (\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{y} \\ &= \frac{\sum_{i=1}^n w_i x_i y_i}{\sum_{i=1}^n w_i x_i^2} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i} \end{aligned}$$

and it's expectation and variance respectively are

$$\begin{aligned} E(\hat{\beta}) &= (\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{E}(\boldsymbol{y}) = (\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \beta = \beta \\ Var(\hat{\beta}) &= [(\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{W}] \boldsymbol{Cov}(\boldsymbol{y}) [(\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{W}]^\top \\ &= [(\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{W}] [\sigma^2 \boldsymbol{W^{-1}}] [[\boldsymbol{W}^\top \boldsymbol{x} (\boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x})^{-1}] \\ &= \sigma^2 (x^\top W x)^{-1} \\ &= \frac{\sigma^2}{\sum_{i=1}^n w_i x_i^2} = \frac{\sigma^2}{n} \end{aligned}$$

We can find the weighted least squares estimator $\hat{\beta}$ is unbiased and its variance is a constant. On the other hand, the ordinary least squares for β is

$$\begin{aligned} \tilde{\beta} &= (\boldsymbol{x}^\top \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{y} \\ E(\tilde{\beta}) &= (\boldsymbol{x}^\top \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{x} \beta = \beta \\ Var(\tilde{\beta}) &= \sigma^2 (\boldsymbol{x}^\top \boldsymbol{x})^{-1} \boldsymbol{x}^\top \boldsymbol{W^{-1}} \boldsymbol{x} (\boldsymbol{x}^\top \boldsymbol{x})^{-1} \\ &= \sigma^2 \frac{\sum_{i=1}^n x_i^2 / w_i}{(\sum_{i=1}^n x_i^2)(\sum_{i=1}^n x_i^2)} = \sigma^2 \frac{\sum_{i=1}^n x_i^4}{(\sum_{i=1}^n x_i^2)^2} \end{aligned}$$

In this case, hence, the ordinary least squares developed for the classical linear model is still unbiased for β within the general linear model setting. However, the variance does not correspond with the one found for the classical model, $\sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$. Thus, all derivations that are based on the covariance matrix of $\tilde{\beta}$ are wrong. In particular, we obtain incorrect variances and standard errors for the estimated regression coefficients, and thus incorrect tests and confidence intervals. And it can be checked the $Var(\hat{\beta}) < Var(\tilde{\beta})$, so the weighted least squared estimator $\hat{\beta}$ is more efficient tha the ordinary least squared estimator $\tilde{\beta}$.

4. Consider the model $y_i = x_i^\top \beta + \varepsilon_i$, where $i = 1, \ldots, n$, x_i's are $k+1$-vectors, the ε_i's are independently distributed with means zero and variances $w_i^{-1}\sigma^2$ and the w_i's are known positive integers. Show that the weighted least squares estimate of β can be obtained using ordinary least squares in the following way: Construct a data set in which each of the cases (y_i, x_i) is repeated w_i times. Show that the ordinary least squares estimate of β obtained from this data set

is $(X^\top WX)^{-1}X^\top Wy$, and an unbiased estimate of σ^2 is $\sum_{i=1}^n w_i(y_i-\hat{y}_i)^2/(n-k-1)$, where $X^\top=(x_1,\ldots,x_n)$ and $W=diag(w_1,\ldots,w_n)$. (These estimates are therefore the same as the corresponding weighted least squares estimates using the w_i's as weights.)

[Hint: Let $\mathbf{1_i}=(1,\ldots,1)^\top$, be a vector of 1's of dimension w_i. To obtain the OLS estimator, we are using the model $Dy=DX\beta+D\varepsilon$, where

$$\boldsymbol{D}=\begin{pmatrix} 1_1 & 0 & \ldots & 0 \\ 0 & 1_2 & \ldots & \ldots \\ \ldots & \ldots & \ldots & 0 \\ 0 & \ldots & 0 & 1_n \end{pmatrix}$$

with dimension $(\sum_{i=1}^n w_i)\times n$.]

By using the setting from the hint, the ordinary least square estimator is

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= ((\boldsymbol{DX})^\top \boldsymbol{DX})^{-1}(\boldsymbol{DX})^\top \boldsymbol{Dy} \\ &= (\boldsymbol{X}^\top \boldsymbol{D}^\top \boldsymbol{DX})^{-1}\boldsymbol{X}^\top \boldsymbol{D}^\top \boldsymbol{Dy} \\ &= (\boldsymbol{X}^\top \boldsymbol{WX})^{-1}\boldsymbol{X}^\top \boldsymbol{Wy}\end{aligned}$$

since $\boldsymbol{D}^\top \boldsymbol{D}=diag(w_1,\ldots,w_n)=\boldsymbol{W}$. Therefore, the ordinary least squares estimator $\hat{\boldsymbol{\beta}}$ from this data font is identical to the weighted least squares estimator from the original data font.

Furthermore, the estimator $\sum_{i=1}^n w_i(y_i-\hat{y}_i)^2/(n-k-1)$ can rewrite as

$$\begin{aligned}\frac{\sum_{i=1}^n w_i(y_i-\hat{y}_i)^2}{n-k-1} &= \frac{1}{n-k-1}(\boldsymbol{y}-\hat{\boldsymbol{y}})^\top \boldsymbol{W}(\boldsymbol{y}-\hat{\boldsymbol{y}}) \\ &= \frac{1}{n-k-1}\boldsymbol{y}^\top(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{WX})^{-1}\boldsymbol{X}^\top \boldsymbol{W})^\top \boldsymbol{W}(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{WX})^{-1}\boldsymbol{X}^\top \boldsymbol{W})\boldsymbol{y} \\ &= \frac{1}{n-k-1}\boldsymbol{y}^\top(\boldsymbol{I}-\boldsymbol{H})^\top \boldsymbol{W}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}\end{aligned}$$

where the matrix $\boldsymbol{H}=\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{WX})^{-1}\boldsymbol{X}^\top \boldsymbol{W}$ and $\boldsymbol{I}-\boldsymbol{H}$ are idempotent but not symmetric. One can check that $\boldsymbol{HX}=\boldsymbol{0}$ and $trace(\boldsymbol{I}-\boldsymbol{H})=rank(\boldsymbol{I}-\boldsymbol{H})=n-(k+1)$. Therefore

$$\begin{aligned}E\left(\frac{\sum_{i=1}^n w_i(y_i-\hat{y}_i)^2}{n-k-1}\right) &= E\left(\frac{1}{n-k-1}\boldsymbol{y}^\top(\boldsymbol{I}-\boldsymbol{H})^\top \boldsymbol{W}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}\right) \\ &= \frac{1}{n-k-1}(\boldsymbol{E(y)}^\top(\boldsymbol{I}-\boldsymbol{H})^\top \boldsymbol{W}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{E(y)}+\boldsymbol{trace}((\boldsymbol{I}-\boldsymbol{H})^\top \boldsymbol{W}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Cov(y)})) \\ &= \frac{1}{n-k-1}(\boldsymbol{0}+\boldsymbol{trace}((\boldsymbol{I}-\boldsymbol{H})^\top \boldsymbol{W}(\boldsymbol{I}-\boldsymbol{H})\sigma^2\boldsymbol{W^{-1}})) \\ &= \frac{(n-k-1)\sigma^2}{n-k-1}=\sigma^2\end{aligned}$$

So $\sum_{i=1}^n w_i(y_i-\hat{y}_i)^2/(n-k-1)$ is an unbiased estimator for σ^2.

5. This example compares in-field ultrasonic measurements of the depths of defects in the Alaska oil pipeline to measurements of the same defects in a laboratory. The lab measurements were done in six different batches. The goal is to decide if the field measurement can be used to predict the more accurate lab measurement. In this analysis, the field measurement is the response variable and the laboratory measurement is the predictor variable. The data, in the file W_pipeline.csv, were given at https://www.itl.nist.gov/div898/handbook/pmd/section6/pmd621.htm (https://www.itl.nist.gov/div898/handbook/pmd/section6/pmd621.htm). The three variables are called *Field*, the in-field measurement, *Lab*, the more accurate in-lab measurement, and *Batch*, the batch number.

a. Draw the scatterplot of *Lab* versus *Field*, and comment on the applicability of the simple linear regression model.

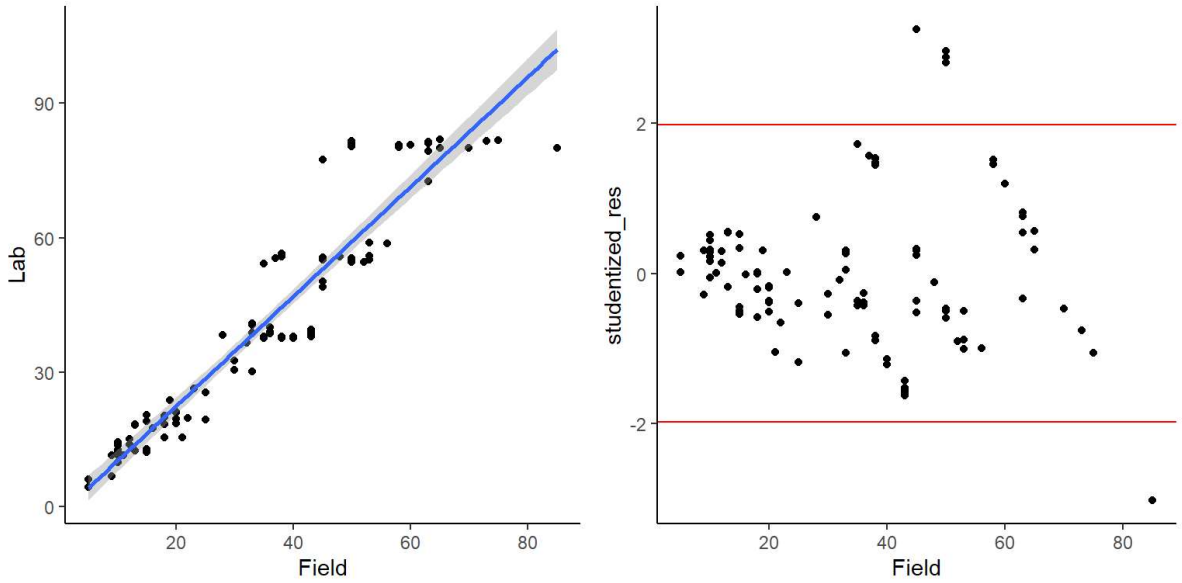
```
data5 ← read_csv("W_pipeline.csv")
data5_lm ← lm(Lab ~ 1 + Field, data5)
summary(data5_lm)
```

```
Call:
lm(formula = Lab ~ 1 + Field, data = data5)

Residuals:
    Min      1Q  Median      3Q     Max
-21.985  -4.072  -1.431   2.504  24.334

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.96750    1.57479  -1.249    0.214
Field        1.22297    0.04107  29.778   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.865 on 105 degrees of freedom
Multiple R-squared:  0.8941,    Adjusted R-squared:  0.8931
F-statistic: 886.7 on 1 and 105 DF,  p-value: < 2.2e-16
```

The fitted simple linear regression is

$$\hat{Lab}_i = -1.97 + 1.22 Field_i$$

```
data5 ← data5 %>%
  add_column(residual = residuals(data5_lm),
             studentized_res = rstudent(data5_lm))
g1 ← ggplot(data5, aes(x = Field, y = Lab)) +
    geom_point() + geom_smooth(method = "lm") + theme_classic()

g2 ← ggplot(data5, aes(x = Field, y = studentized_res)) +
  geom_point() +
  geom_hline(yintercept = qt(0.05/2, data5_lm$df.residual - 1), color = "red") +
  geom_hline(yintercept = qt((1-0.05/2), data5_lm$df.residual - 1), color = "red") +
  theme_classic()
grid.arrange(g1, g2, nrow = 1)
```



The scatter plot (left) with the simple linear regression line and studentized residual plot (right) are above. We can find that the *Lab* is increasing as the *Field* increases. In general, the model appears to be a good fit. However, we can clearly see the "right opening megaphone" pattern, which indicates non-constant variance, in the studentized residual plot. So I use the Breusch–Pagan test to ensure this data has or not has homoscedastic errors.

```
library(lmtest)
bptest(data5_lm, studentize = TRUE)
```

```
    studentized Breusch-Pagan test

data:  data5_lm
BP = 16.045, df = 1, p-value = 6.185e-05
```

The $p-value < 0.05$, we reject the null hypothesis of homoscedastic variance. Maybe we should use the weighted least squared method to estimate this simple linear model.

b. Perform a two-stage least squares approach that models the conditional variance of *Lab* as a function of *Field*. State your findings and conclusions.

From the above observation, I propose that the relation of the conditional variance of *Lab* σ_i and *Field_i_* are

$$\sigma_i^2 = e^{\alpha_0+\alpha_1 Feild_i+\nu_i}$$

where ν_i's are the deviations of the squared errors from their expectations. In this model, $\boldsymbol{\alpha}$ can be estimated with the help of the regression

$$log(\hat{\varepsilon}_i^2) = \alpha_0 + \alpha_1 Fidld_i + \nu_i$$

and the weights for the regression between *Lab* and *Field* are then given by

$$\hat{w}_i = \frac{1}{\hat{\sigma}_0 + \hat{\sigma}_1 Field_i}$$

Therefore, the two-stage estimator for this model by R is as follows

```
res_lm ← lm(I(log(residual^2)) ~ 1 + Field, data5)
weight ← 1 / exp(fitted(res_lm))
data5_wlm ← lm(Lab ~ 1 + Field, weights = weight, data5)
summary(data5_wlm)
```

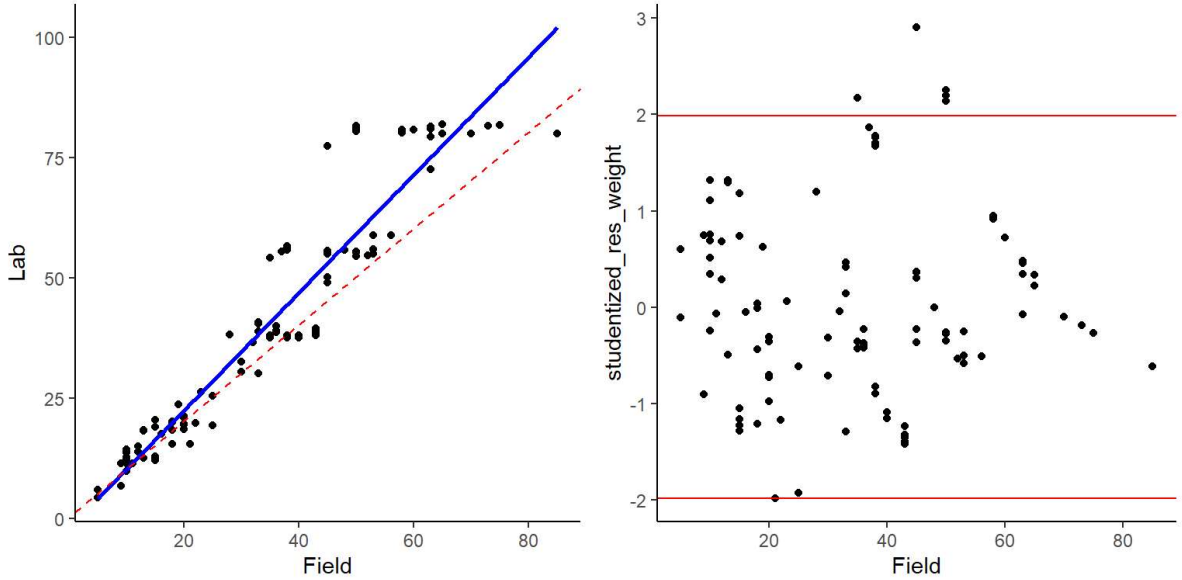
```
Call:
lm(formula = Lab ~ 1 + Field, data = data5, weights = weight)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-3.5363 -1.1928 -0.3392  1.0976  5.0620

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4145     0.8575   -1.65    0.102
Field         1.1925     0.0394   30.27   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.82 on 105 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8962
F-statistic: 916.1 on 1 and 105 DF,  p-value: < 2.2e-16
```

The fitted simple linear regression by the weighted least square method is

$$\hat{Lab}_i = -1.41 + 1.19 Field_i$$

It is different from the OLS. The intercept is increased and the slope is decreased.

```
data5 ← data5 %>%
  add_column(studentized_res_weight = rstudent(data5_wlm))
g3 ← ggplot(data5, aes(x = Field, y = Lab)) +
    geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  geom_abline(intercept = res_lm$coefficients[1],
              slop = res_lm$coefficients[2],
              color = "red", linetype = "dashed") +
  theme_classic()
g4 ← ggplot(data5, aes(x = Field, y = studentized_res_weight)) +
  geom_point() +
  geom_hline(yintercept = qt(0.05/2, data5_lm$df.residual - 1), color = "red") +
  geom_hline(yintercept = qt((1-0.05/2), data5_lm$df.residual - 1), color = "red") +
  theme_classic()
grid.arrange(g3, g4, nrow = 1)
```



In the left figure, the blue solid line is the original regression line by the OLS and the red dashed line is by the WLS. The right figure is the studentized residual plot. The variance error is more approach homoscedastic, but the *Field* underestimates the *Lab*, especially after $Field > 40$. In conclusion, I thought the WLS does not do better than the orginial performance in part (a), though, it was solved the problem of the heteroscedastic variance.

c. In addition to the variable *Field*, consider the variable *Batch*, treating *Batch* as a class (or categorical) variable, in the regression model. Does the conditional variance vary with *Batch*? State your findings and conclusions.

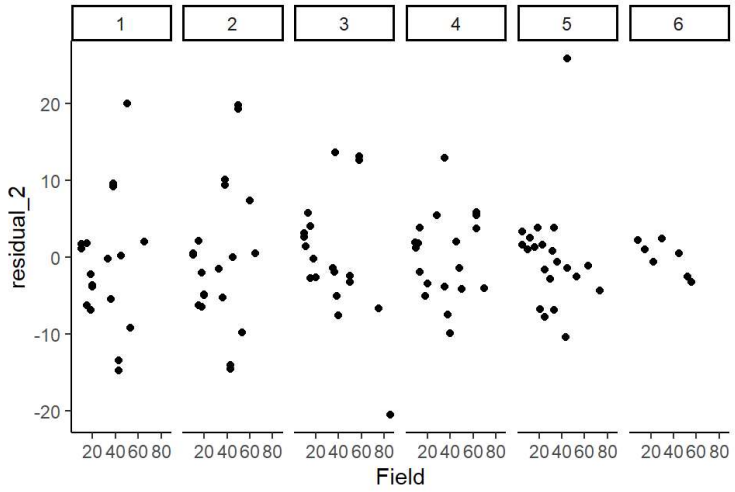
According the content of this item, our model becomes to

$$Lab_i = \beta_0 + \beta_1 Field_i + \beta_2 Batch_i + \varepsilon_i$$

```
data5 ← data5 %>%
  mutate(Batch = factor(Batch))

data5_lm2 ← lm(Lab ~ 1 + Field + Batch, data5)
summary(data5_lm2)
```

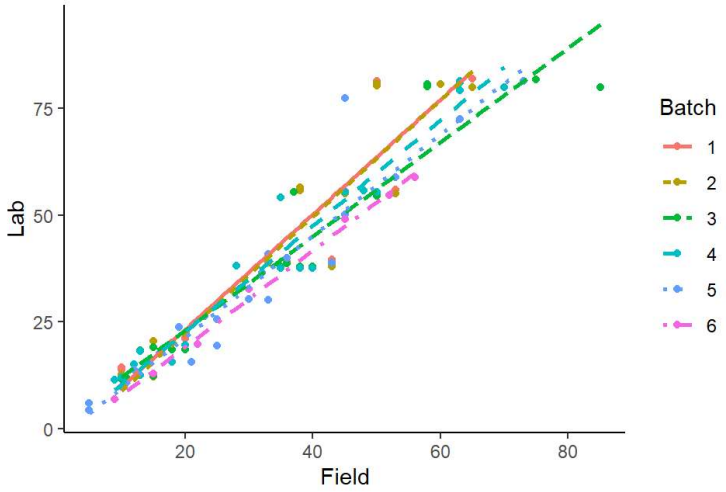
```
Call:
lm(formula = Lab ~ 1 + Field + Batch, data = data5)

Residuals:
    Min      1Q  Median      3Q     Max
-20.521  -4.087  -0.578   2.488  25.821

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.37570    2.23259   0.168   0.8667
Field        1.22236    0.04092  29.871   <2e-16 ***
Batch2      -0.39586    2.49917  -0.158   0.8745
Batch3      -3.75504    2.50281  -1.500   0.1367
Batch4      -1.88039    2.49991  -0.752   0.4537
Batch5      -3.80262    2.47052  -1.539   0.1269
Batch6      -6.86426    3.44832  -1.991   0.0493 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.799 on 100 degrees of freedom
Multiple R-squared:  0.9008,    Adjusted R-squared:  0.8949
F-statistic: 151.4 on 6 and 100 DF,  p-value: < 2.2e-16
```

```
data5_2 ← data5 %>% add_column(residual_2 = residuals(data5_lm2))
ggplot(data5_2, aes(x = Field, y = residual_2)) +
  geom_point() +
  facet_wrap(~ Batch, nrow = 1) +
  theme_classic()
```



The variance of residuals is significantly difference in different Batch level. The variance is lower when the batch level increases. It is not appropriate aggregate all data with fitting in a single linear regression. In my opinion, a suitable approach to this data is fitting the regression lines separately in each Batch level (like below plot), where each model may obey the homoscedastic variance assumption in each separate line.

```
ggplot(data5, aes(x = Field, y = Lab, color = Batch, linetype = Batch)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_classic()
```


Lab
75
50
25
0
20
40
60
80
Field
Batch
1
2
3
4
5
6