R_Project2_S24_

Zheng Yi Lai

2024-02-06

Question 1. In this question, you will create some **simulated data** and will fit simple linear regression models to it. Make sure to use **set.seed(1)** at the beginning of your R codes to ensure consistent results.

(a) Follow the steps below to create a group of simulated data.

Step 1: [**Generate vector** x] - Using the **rnorm()** function, create a vector x, containing 100 observations drawn from a N(0, 1) distribution

Step 2: [**Generate random error** eps] – Using the **rnorm()** function, create a vector *eps*, containing 100 observations drawn from a N(0, 0.25) distribution - a normal distribution with mean zero and variance 0.25. (Hint: please note that in rnorm(), sd is used, instead of variance.)

Step 3: [**Generate vector** y] – Using x and *eps*, generate a vector y according to the model

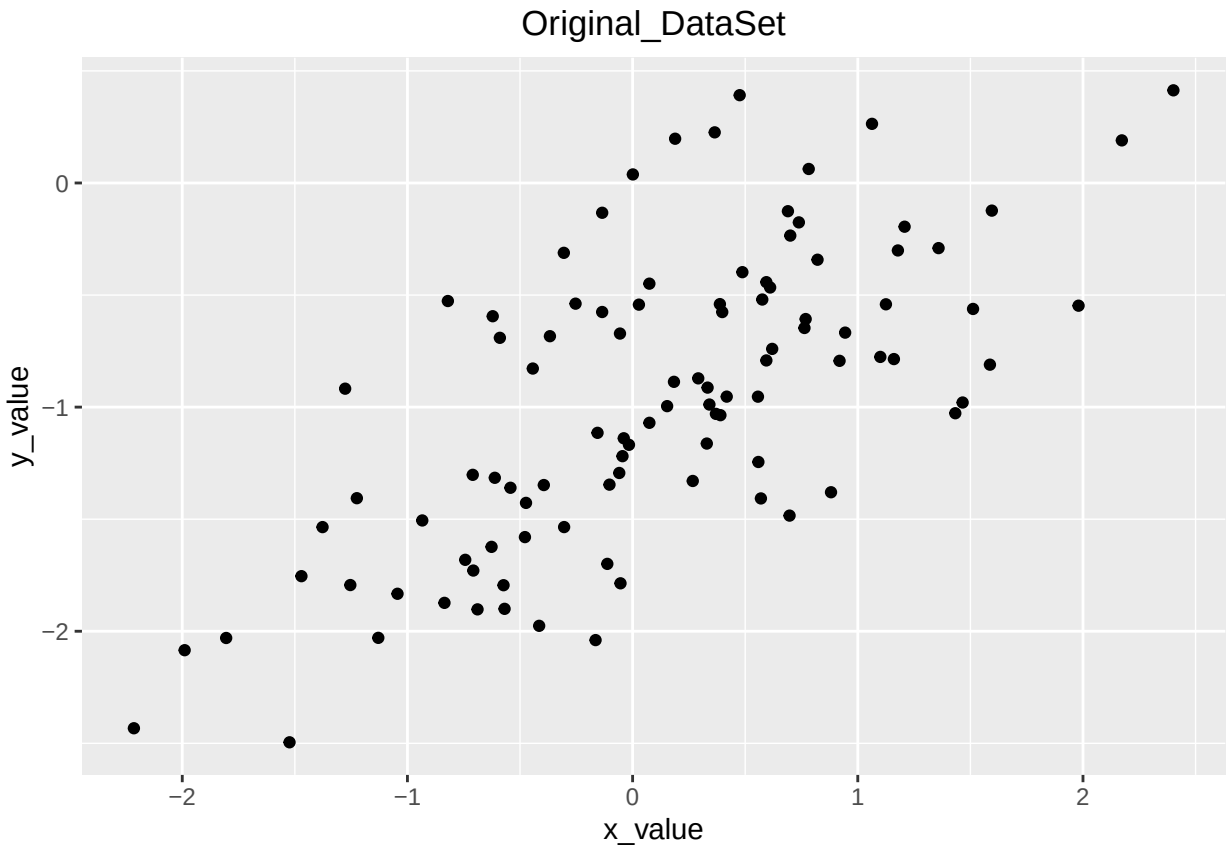
Question: What is the length of the vector y ? What are the values of β_0 and β_1 in this linear model?

```
#Set seed so we have the same randomization
set.seed(1)

#Declare variables
x <- rnorm(100, mean = 0, sd = 1)
eps <- rnorm(100, mean = 0, sd = sqrt(0.25))
y <- -1 + 0.5 * x + eps
df <- data.frame(x,y)

#length of y is 100
beta_0 <- -1
beta_1 <- 0.5
length(y)
```

```
## [1] 100
```

(b) Create a scatterplot displaying the relationship between x and y. Comment on what you observe.

```
library(ggplot2)

ggplot(df, aes(x=x, y=y))+
  ggtitle("Original_DataSet")+
  theme(plot.title =element_text(hjust = 0.5))+
  labs(x = "x_value",
       y = "y_value")+
  geom_point(col="black")
```



I observe a discernible upward trajectory within the scatter plot graph, indicating a positive trend.

(c) Fit a least squares linear model predict y using x. Comment on the model obtained. How do beta_hat_0 and beta_hat_1 compare to β_0 and β_1? How does variance_hat compare to variance?

```
#Fit it into a regression model
linear_model <- lm(y ~ x)
 summary(linear_model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93842 -0.30688 -0.06975  0.26970  1.17309
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.01885    0.04849 -21.010  < 2e-16 ***
## x            0.49947    0.05386   9.273 4.58e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4814 on 98 degrees of freedom
```

```
## Multiple R-squared:  0.4674, Adjusted R-squared:  0.4619
## F-statistic: 85.99 on 1 and 98 DF,  p-value: 4.583e-15
```

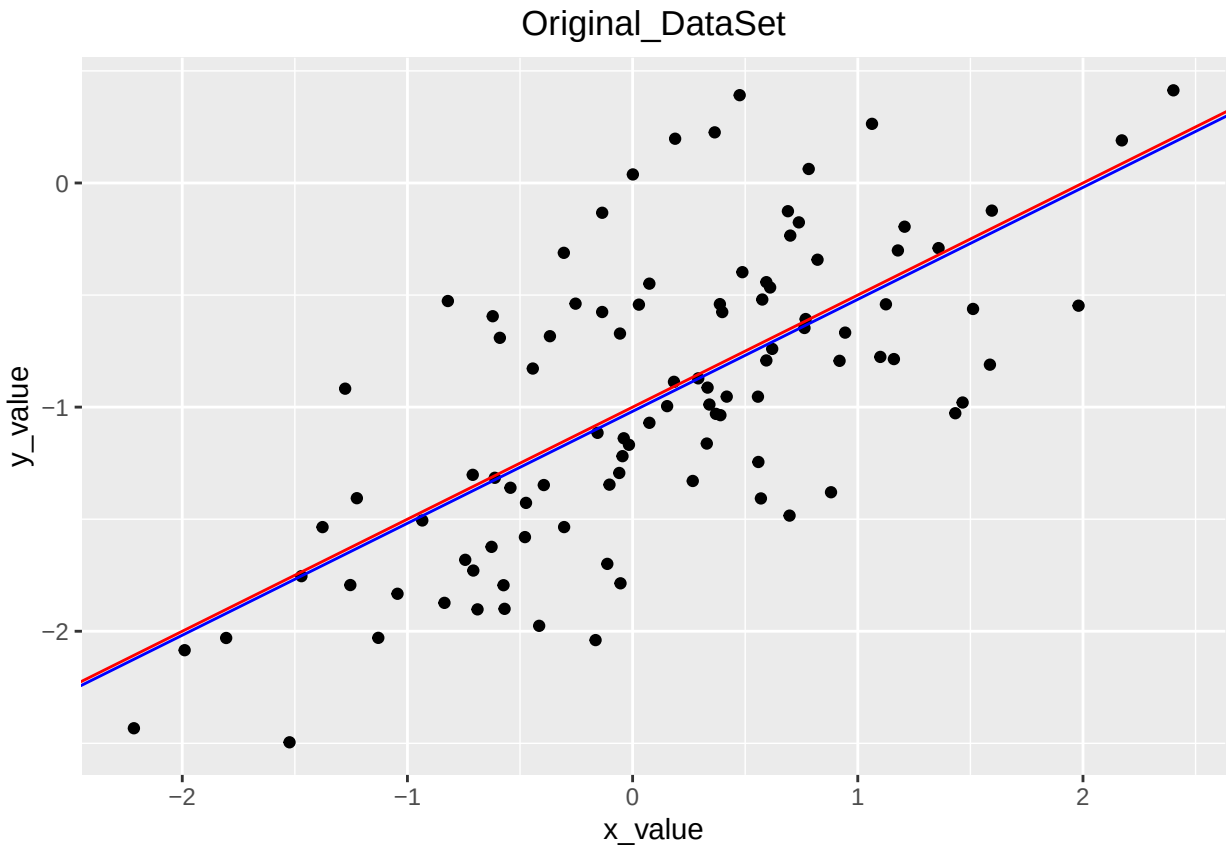
```
# \( \beta_0\) is -1 and beta_hat_0 is -1.01885 from the summary
 beta_0 <- -1
 beta_1 <- 0.5
variance <- var(y)
pop_value <-c(beta_0, beta_1,variance)

# \( \beta_1 \) is 0.5 and beta_hat_1 is 0.49947 from the summary
beta_hat_0 <-  summary(linear_model)$coefficients[1,1]
beta_hat_1 <-  summary(linear_model)$coefficients[2,1]
variance_hat<-  summary(linear_model)$sigma
model_value <-c(beta_hat_0,beta_hat_1,variance_hat)

#Compare them in the data frame they are not apart from each other, but when we
#compare variance and variance hat they are slightly different
variable <- c("beta_0"," \beta_1", "variance")
df1<- data.frame(variable,pop_value,model_value)
df1
```

```
##   variable  pop_value model_value
## 1   beta_0 -1.0000000  -1.0188463
## 2  \beta_1  0.5000000   0.4994698
## 3 variance  0.4306459   0.4813767
```

When comparing beta values in the data frame, they are not distinguishable from each other. However, when we compare the variance and the estimated variance (variance hat), there are slight differences.

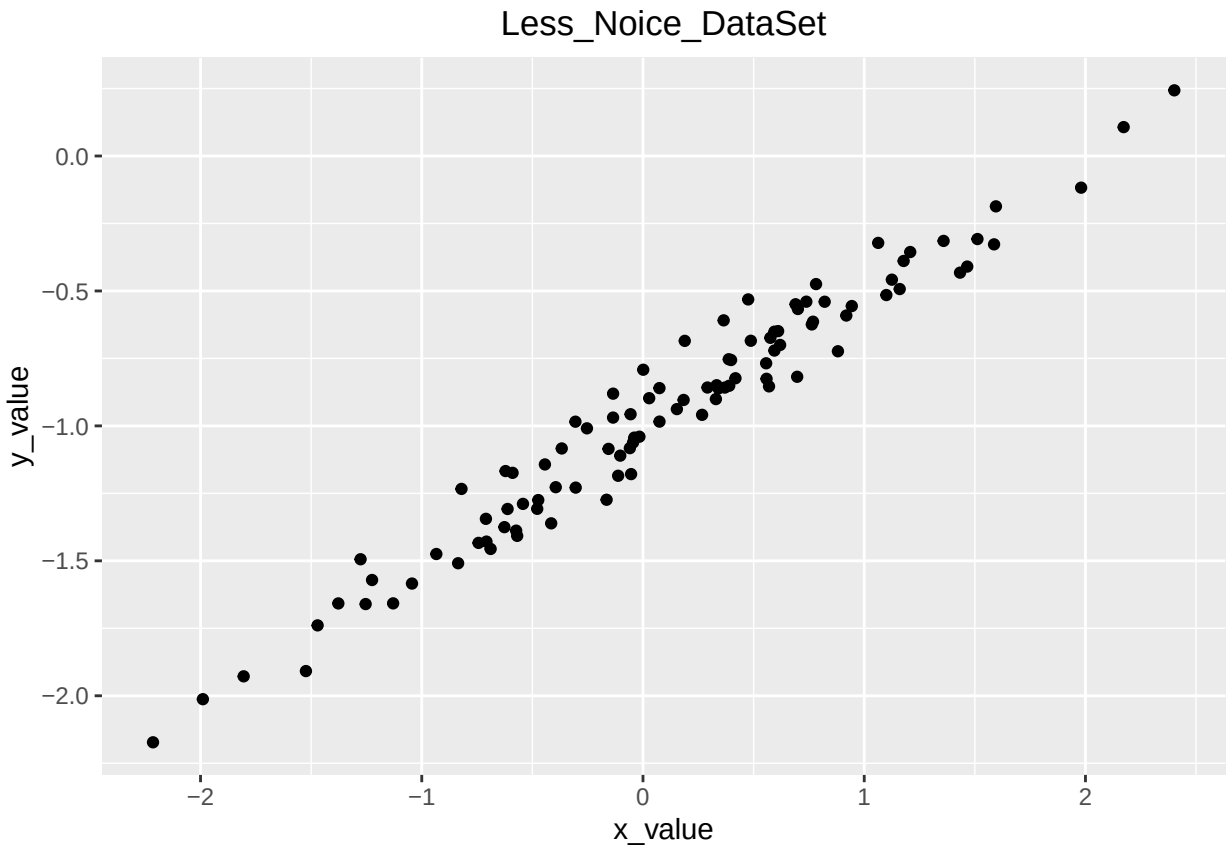
(d) On the scatterplot, display the least squares line (line in color blue) obtained in (c). Add the population regression line (in color red) to the plot.

```
ggplot(df, aes(x=x, y=y))+
  ggtitle("Original_DataSet")+
  theme(plot.title =element_text(hjust = 0.5))+
  labs(x = "x_value",
       y = "y_value")+
  geom_point(col="black")+
  geom_abline(slope=beta_hat_1, intercept=beta_hat_0,color="blue") +
  geom_abline(slope= beta_1, intercept= beta_0, color="red")
```



(e) Repeat (a) - (c) after modifying the data generation process in such a way that there is *less noise* in the data. The model (Model-1) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term e in Step 2, for example, use N(0, 0.01). Describe your results.

```
#Set seed so we have the same randomization
set.seed(1)

#Declare variables
x_1 <- rnorm(100, mean = 0, sd = 1)
eps_1 <- rnorm(100, mean = 0, sd = sqrt(0.01))
y_1 <- -1 + 0.5 * x_1 + eps_1
df2 <- data.frame(x_1,y_1)

#Plot the graph
ggplot(df2, aes(x=x_1, y=y_1))+
  ggtitle("Less_Noice_DataSet")+
  theme(plot.title =element_text(hjust = 0.5))+
  labs(x = "x_value",
       y = "y_value")+
  geom_point(col="black")
```




```
#Fit in the model
linear_model_1 <- lm(y_1 ~ x_1)
summary(linear_model_1)
```

```
##
## Call:
## lm(formula = y_1 ~ x_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18768 -0.06138 -0.01395  0.05394  0.23462
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.003769   0.009699  -103.5   <2e-16 ***
## x_1          0.499894   0.010773    46.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09628 on 98 degrees of freedom
## Multiple R-squared:  0.9565, Adjusted R-squared:  0.956
## F-statistic:  2153 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
# \( \beta_0\) is -1 and beta_hat_0 is  -1.003769  from the summary
beta_0_1 <- -1
```

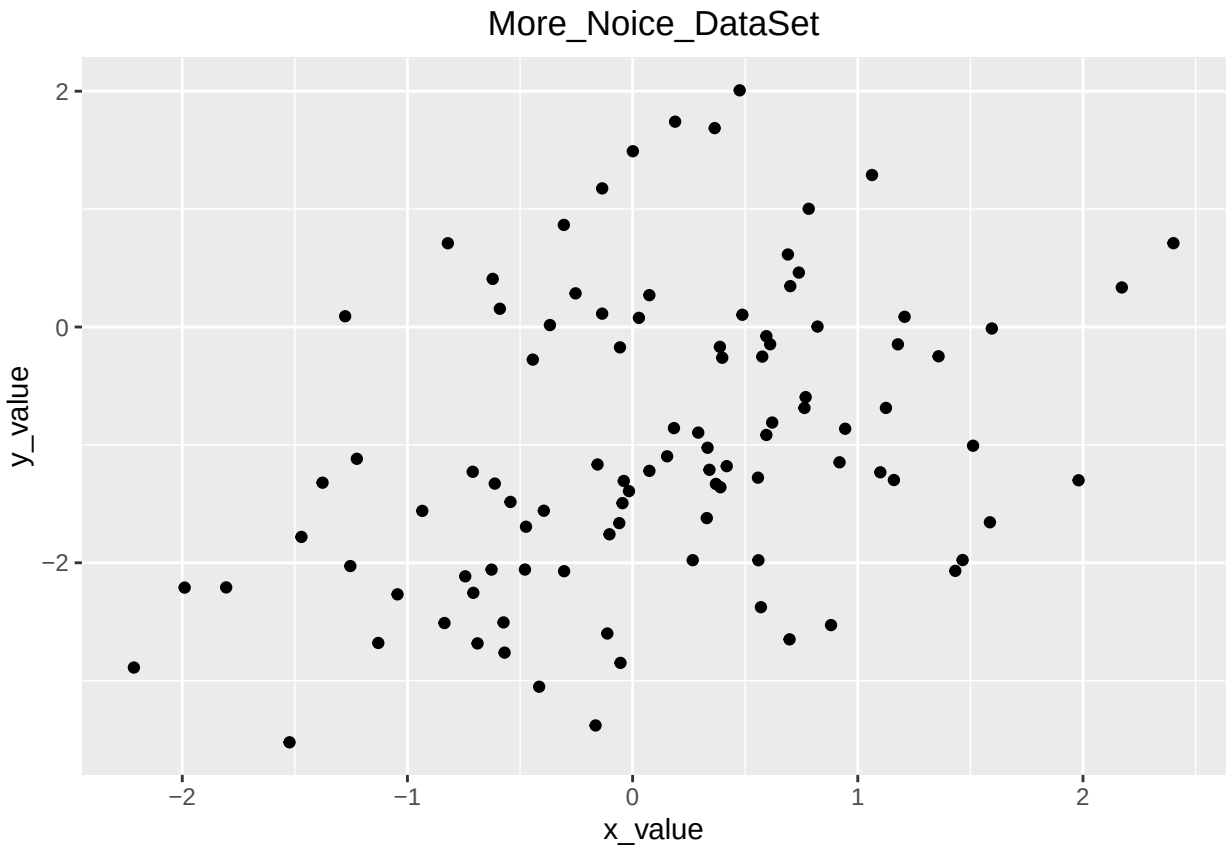
```r
beta_1_1 <- 0.5
variance_1 <- var(y_1)
pop_value_1 <-c( beta_0_1, beta_1_1,variance_1)

# \( \beta_1 \) is 0.5 and beta_hat_1 is 0.499894  from the summary
beta_hat_0_1 <-   summary(linear_model_1)$coefficients[1,1]
beta_hat_1_1 <-   summary(linear_model_1)$coefficients[2,1]
variance_hat_1 <-  summary(linear_model_1)$sigma
model_value_1 <-c(beta_hat_0_1,beta_hat_1_1,variance_hat_1)

#Compare them in the data frame they are not apart from each other, but when we
#compare variance and variance hat they are slightly different
variable <- c("beta_0","beta_1", "variance")
df3<- data.frame(variable,pop_value_1,model_value_1)
df3
```

```
##   variable pop_value_1 model_value_1
## 1   beta_0  -1.0000000   -1.00376926
## 2   beta_1   0.5000000    0.49989396
## 3 variance   0.2107803    0.09627533
```

When examining the beta values in the data frame, they appear indistinguishable from each other. Nonetheless, disparities become evident when comparing the variance and the estimated variance (variance hat). These differences are notable which is variance hat is much smaller than variance, especially as observed in the scatter plot, where data points are closer together.

(f) Repeat (a) - (c) after modifying the data generation process in such a way that there is *more noise* in the data. The model (Model-1) should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term e in Step 2, for example, use N(0, 1.44). Describe your results.

```r
#Set seed so we have the same randomization
set.seed(1)

#Declare variables
x_2 <- rnorm(100, mean = 0, sd = 1)
eps_2 <- rnorm(100, mean = 0, sd = sqrt(1.44))
y_2 <- -1 + 0.5 * x_2 + eps_2
df4 <- data.frame(x_2,y_2)

#Plot the scatter plot
ggplot(df4, aes(x=x_2, y=y_2))+
  ggtitle("More_Noice_DataSet")+
  theme(plot.title =element_text(hjust = 0.5))+
  labs(x = "x_value",
       y = "y_value")+
  geom_point(col="black")
```




```
#Fit into the liner model
linear_model_2 <- lm(y_2 ~ x_2)
summary(linear_model_2)
```

```
##
## Call:
## lm(formula = y_2 ~ x_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2522 -0.7365 -0.1674  0.6473  2.8154
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.0452     0.1164  -8.981 1.97e-14 ***
## x_2           0.4987     0.1293   3.858 0.000205 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.155 on 98 degrees of freedom
## Multiple R-squared:  0.1319, Adjusted R-squared:  0.123
## F-statistic: 14.88 on 1 and 98 DF,  p-value: 0.0002048
```

```
# \( \beta_0\) is -1 and beta_hat_0 is  -1.003769  from the summary
 beta_0_2 <- -1
```

```r
beta_1_2 <- 0.5
variance_2 <- var(y_2)
pop_value_2 <-c( beta_0_2, beta_1_2,variance_2)

# \( \beta_1 \) is 0.5 and beta_hat_1 is 0.499894  from the summary
beta_hat_0_2 <-   summary(linear_model_2)$coefficients[1,1]
beta_hat_1_2 <- summary(linear_model_2)$coefficients[2,1]
variance_hat_2 <-  summary(linear_model_2)$sigma
model_value_2 <-c(beta_hat_0_2,beta_hat_1_2,variance_hat_2)

#Compare them in the data frame they are not apart from each other, but when we
#compare variance and variance hat they are slightly different
variable <- c(" beta_0","beta_1", "variance")
df5<- data.frame(variable,pop_value_2,model_value_2)
df5
```

```
##   variable pop_value_2 model_value_2
## 1   beta_0    -1.00000    -1.0452311
## 2   beta_1     0.50000     0.4987275
## 3 variance     1.52191     1.1553040
```

When examining the beta values in the data frame, they appear indistinguishable from each other. Nonetheless, disparities become evident when comparing the variance and the estimated variance (variance hat). These differences are notable which is variance hat is much bigger than variance, especially as observed in the scatter plot, where data points are more spread out.

(g) What are the coefficient of determinations R^2 and residual standard errors variance hat based on the original data set, the noisier dataset, and the less noisy data set? Comment on your results.

```r
#Create a dataframe to display
variable_1 <- c("original","noisier","less noisy")
Residual_Standard_Error<-c(sigma(linear_model),sigma(linear_model_1),sigma(linear_model_2))
R_Squared <-c(summary(linear_model)$r.squared,summary(linear_model_1)$r.squared,
              summary(linear_model_2)$r.squared)
df6 <- data.frame(variable_1,Residual_Standard_Error,R_Squared)
df6
```

```
##   variable_1 Residual_Standard_Error R_Squared
## 1   original              0.48137666 0.4673515
## 2    noisier              0.09627533 0.9564698
## 3 less noisy              1.15530399 0.1318509
```

In the original data set, there's a slight difference between the values of RSE and R-squared. However, in the noisier data set, the RSE is significantly lower than the R-squared value. Conversely, in the less noisy dataset, the RSE is notably larger than the R-squared value.

(h) What are the 95% confidence intervals for β_0 and β_1 based on the original data set, the noisier dataset, and the less noisy data set? Comment on your results

```r
#Set seed so we have the same randomization
original <- confint(linear_model, level=0.95)
noiser <-confint(linear_model_1, level=0.95)
less_noise <- confint(linear_model_2, level=0.95)

#Display interval
original
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.1150804 -0.9226122
## x            0.3925794  0.6063602
```

```r
noiser
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0230161 -0.9845224
## x_1          0.4785159  0.5212720
```

```r
less_noise
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.2761929 -0.8142694
## x_2          0.2421906  0.7552645
```

Upon examining the original dataset, it's evident that the confidence interval for β_1 (with a range of approximately 0.1) is slightly narrower than that of the less noisy dataset, where it spans from -1.28 to -0.81, resulting in a range of around 0.4. Conversely, the noisier dataset exhibits the smallest range, approximately 0.04. As for β_0, the original dataset presents a range of approximately 0.21. In contrast, the noisier dataset's range is about 0.05, and the less noisy dataset has the widest range, approximately 0.51.

Question 2. The website *www.playbill.com* provides weekly reports on the box office ticket sales for plays on Broadway in New York. We shall consider the data for the week October 11-17, 2004 (referred to below as the current week). The data are in the form of the gross box office results for the current week and the gross box office results for the previous week (i.e., October 3-10, 2004). The data are included in the file *playbill.csv*. Fit the following model to the data: : where Y is the gross box office results for the current week (in $) and x is the gross box office results for the previous week (in $). Complete the following tasks:

```
#Import csv file
data_csv <- readr::read_csv("playbill.csv",show_col_types = FALSE)

#Fit the model
y_data <- data_csv$CurrentWeek
x_data <- data_csv$LastWeek
fit <- lm(y_data ~ x_data ,data=data_csv)
summary(fit)
```

```
##
## Call:
## lm(formula = y_data ~ x_data, data = data_csv)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36926  -7525  -2581   7782  35443
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.805e+03  9.929e+03   0.685    0.503
## x_data      9.821e-01  1.443e-02  68.071   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18010 on 16 degrees of freedom
## Multiple R-squared:  0.9966, Adjusted R-squared:  0.9963
## F-statistic:  4634 on 1 and 16 DF,  p-value: < 2.2e-16
```

(a) Identify and provide an interpretation of the coefficient of determination.

```
c_d <-summary(fit)$r.squared
c_d
```

```
## [1] 0.9965589
```

The coefficient of determination, which is $R^2 = 0.9965589$, is remarkably close to 1. This near-perfect fit indicates that the regression model excellently captures the variability in the dependent variable, suggesting that the independent variable(s) effectively predict the outcome.

(b) Find a 95% confidence interval for the slope of the regression model, β_1,. Is 1 a plausible value for β_1,? Give a reason to support your answer.

```
c_i <- confint(fit, level=0.95)
c_i
```

```
##                      2.5 %       97.5 %
## (Intercept) -1.424433e+04 27854.099443
## x_data       9.514971e-01     1.012666
```

Indeed, the value of 1 falls well within the confidence interval of β_1, which ranges from 0.9515 to 1.0127 (9.514971e-01 to 1.012666). This suggests that 1 is a plausible estimate for β_1 based on the given confidence interval.

(c) Test the null hypothesis $\beta_0 = 10000$ against a two-sided alternative. Interpret your result.

```
#Calculate t statistics using equation
t_statistic <- ( summary(fit)$coefficients[1,1] - 10000) / summary(fit)$coefficients[1,2]

#Finding the p- value using pt function
p_value <- 2 * pt(-abs(t_statistic), 16)
p_value
```

```
## [1] 0.7517807
```

The obtained p-value of approximately 0.7517807 is greater than the significance level of 0.05. This indicates that we do not have sufficient evidence to reject the null hypothesis ($\beta_0 = 10000$). Therefore, based on the results of the t-test, we fail to reject the null hypothesis, suggesting that the coefficient β_0 is not significantly different from 10000 at the 5% level of significance.

(d) Use the fitted regression model to estimate the gross box office results for the current week (in $) for a production with 400,000 in gross box office the previous week. Find a 95% prediction interval for the gross box office results for the current week (in $) for a production with 400,000 in the gross box office the previous week. Is $450,000 a feasible value for the gross box office results in the current week, for a production with $400,000 in gross box office the previous week? Give a reason to support your answer.

```
predict(fit, newdata=data.frame(x_data=400000),se=TRUE,interval="prediction")$fit
```

```
##        fit      lwr      upr
## 1 399637.5 359832.8 439442.2
```

The prediction interval is (359,832.8, 439,442.2). $450,000 is not in the interval so it is not a feasible value for the gross box office results in the current week, for a production with $400,000 in gross box office the previous week.

(e) Some promoters of Broadway plays use the prediction rule that next week's gross box office results will be equal to this week's gross box office results. Comment on the appropriateness of this rule.

```
#Create data frame with current week values as last week
newdata = data.frame(LastWeek=data_csv$CurrentWeek)

#Generate prediction interval
predict_i <- predict(fit, newdata, interval="confidence", level=.95)
predict_i<- data.frame(predict_i)
predict_i$LastWeek <- data_csv$LastWeek
predict_i
```

```
##          fit       lwr       upr LastWeek
## 1   689780.7  680508.2  699053.2   695437
## 2   496833.1  487078.0  506588.2   498969
## 3   594655.3  585628.6  603682.0   598576
## 4   526320.1  516881.7  535758.5   528994
## 5   559681.4  550503.1  568859.7   562964
## 6   284515.9  270778.5  298253.3   282778
## 7   579532.2  570455.7  588608.7   583177
## 8   156899.3  139957.6  173841.1   152833
## 9   110608.9   92429.6  128788.3   105698
## 10  828766.8  817626.5  839907.2   836959
## 11  959610.5  945673.4  973547.6   970190
## 12  646933.5  637890.2  655976.7   651808
## 13  378265.4  366576.7  389954.2   378238
## 14 1100362.4 1082847.7 1117877.1  1113510
## 15  610044.5  601043.5  619045.6   614246
## 16  923894.2  910786.8  937001.5   933822
## 17 1187793.2 1167893.3 1207693.2  1202536
## 18  486673.5  476791.8  496555.1   488624
```

Using the `predict()` function, I utilized the current week's data to predict the gross box office results for next week. From the resulting predictions, it is evident that next week's gross box office results will not be equal to this week's gross box office results. Moreover, the predictions suggest that next week's gross box office results will be even less than the current week's.

Question 3. Regression through the origin.

Occasionally, a mean function in which the intercept is known a priori to be 0 may be fit. The model is $Y_i = \beta x_i + e_i$, i.e., the mean function is $E(Y|X = x_i) = \beta x_i$. The residual sum of squares for this model, assuming the errors are independent with common variance σ^2, is $RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}x_i)^2$.

(c) The data file `snake` in the `alr4` package gives X as the water content of snow on April 1 and Y as the water yield from April to July in inches in the Snake River watershed in Wyoming for $n = 17$ years from 1919 to 1935 (Wilm, 1950). Fit a regression through the origin

```
#Import library and dataset
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
data(snake)

#Fit in the model
lm_snake <- lm(snake$Y ~ snake$X +0)

#From summary we know the beta_1 hat = 0.52039, variance hat is 1.69975
summary(lm_snake)
```

```
##
## Call:
## lm(formula = snake$Y ~ snake$X + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4207 -1.4924 -0.1935  1.6515  3.0771
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## snake$X  0.52039    0.01318   39.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.7 on 16 degrees of freedom
## Multiple R-squared:  0.9898, Adjusted R-squared:  0.9892
## F-statistic:  1559 on 1 and 16 DF,  p-value: < 2.2e-16
```

```
#Value of beta_hat
beta_hat_snake <-summary(lm_snake)$coefficients[1,1]
beta_hat_snake
```

```
## [1] 0.520394
```

```
#Value of variance_hat
variance_hat_snake <-  summary(lm_snake)$sigma
variance_hat_snake
```

```
## [1] 1.69975
```

(d) Based on the fitted regression through the origin, we test for evidence of a linear relationship between X and Y using a t-test at $\alpha = 0.05$. If the absolute value of the t-statistic exceeds the critical t-value, there is evidence of a linear relationship. Otherwise, there is no evidence of a linear relationship.

```
#From summary we know the p-value is less than 2e-16
summary(lm_snake)
```

```
##
## Call:
## lm(formula = snake$Y ~ snake$X + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4207 -1.4924 -0.1935  1.6515  3.0771
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## snake$X  0.52039    0.01318   39.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.7 on 16 degrees of freedom
## Multiple R-squared:  0.9898, Adjusted R-squared:  0.9892
## F-statistic:  1559 on 1 and 16 DF,  p-value: < 2.2e-16
```

```
p_value_snake <- 2e-16
p_value_snake
```

```
## [1] 2e-16
```

In the context of the snake dataset, the p-value obtained from fitting a linear model through the origin is significantly less than 2×10^{-16}. This extremely low p-value leads us to reject the null hypothesis, which states that the slope (β_0) is equal to 0. Rejecting the null hypothesis indicates strong evidence that there is indeed a linear relationship between the independent variable (X) and the dependent variable (Y). Therefore, we can conclude that there is a significant linear relationship between X and Y in the snake dataset.