# hw_06.Rmd

Vinay K L

2023-11-16

### Q1

Fit a linear model to the given data

```r
# Given data
x = c(110.5, 105.4, 118.1, 104.5, 93.6, 84.1, 77.8, 75.6)
y = c(5.755, 5.939, 6.010, 6.545, 6.730, 6.750, 6.899, 7.862)
```

```r
# Given equation corresponds to basic linear regression model.
# To fit the model
linear_model <- lm(y ~ x)

# Checking the summary of the model results
summary(linear_model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34626 -0.27605 -0.09448  0.27023  0.53495
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.137455   0.842265  12.036    2e-05 ***
## x           -0.037175   0.008653  -4.296  0.00512 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3624 on 6 degrees of freedom
## Multiple R-squared:  0.7547, Adjusted R-squared:  0.7138
## F-statistic: 18.46 on 1 and 6 DF,  p-value: 0.005116
```

---

## a) Least squares estimates of the slope.

```r
# Extracting the coefficient for the slope
slope_estimate <- coef(linear_model)[2]

slope_estimate
```

```
##           x
## -0.03717469
```

Interpretation : For each additional unit increase in plant height, the estimated change in grain yield is approximately equal to ($\hat{\beta_1}$) units. The sign of ($\hat{\beta_1}$) indicates the direction of the relationship. If ($\hat{\beta_1}$) is positive, it suggests a positive correlation, meaning higher plant heights are associated with higher grain yields. If ($\hat{\beta_1}$) is negative, it suggests a negative correlation.

---

# b) Perform F-test and then T-test.

```r
# First let us do a F-test. F-test can be conducted by using the idea of ANOVA
anova(linear_model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## x          1 2.42357 2.42357  18.455 0.005116 **
## Residuals  6 0.78794 0.13132
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Now we can check for T-test, Summary of the model-fitting will have the p-values which can be used to
summary(linear_model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34626 -0.27605 -0.09448  0.27023  0.53495
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.137455   0.842265  12.036    2e-05 ***
## x           -0.037175   0.008653  -4.296  0.00512 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3624 on 6 degrees of freedom
## Multiple R-squared:  0.7547, Adjusted R-squared:  0.7138
## F-statistic: 18.46 on 1 and 6 DF,  p-value: 0.005116
```

In both the cases, F-test and T-test, p value is less than 0.05 which provides evidence to reject the null hypothesis of $H_0 : \beta_1 = 0$ . This suggests that there is a significant linear relationship between the predictor variable (plant height) and the response variable (grain yield).

---

## c) Construct a 95% CI by hand and compare to what R gives.

```r
# ALpha value is given
alpha <- 0.05
n <- length(x)

# Extracting the values from the summary of model fitting

intercept_estimate <- coef(linear_model)[1]
SE_intercept <- summary(linear_model)$coefficients[1, "Std. Error"]

# Getting critical t value

t_critical <- qt(alpha/2, n-2)

# Calculting the intereval as upper and lower boundry
lower_bound <- intercept_estimate - t_critical * SE_intercept
upper_bound <- intercept_estimate + t_critical * SE_intercept


# Displaying the results
lower_bound # Calculated by hand
```

```
## (Intercept)
##     12.1984
```

```r
upper_bound # Calcualted by hand
```

```
## (Intercept)
##    8.076507
```

```r
confint(linear_model) # Calculated by R
```

```
##                   2.5 %      97.5 %
## (Intercept)  8.07650745 12.19840320
## x           -0.05834895 -0.01600043
```

---

# d) Raw residuals.

---

# e) Estimate of the error variance.
"'r # The estimate of the error variance is obtained as the mean squared residual from the regression model.

| # d) Raw residuals. |
|---|
| # We can obrain the same in R with following code # Calculate the estimate of the error variance error_variance <- sum(residuals^2) / (length(x) - 2)<br># Display the result error_variance "'<br>`## [1] 0.1313228` |

## f) Expected yield of the rice variety

```
# Given values
x_0 <- 100
alpha <- 0.05

# Calculate the expected yield
expected_yield <- coef(linear_model)[1] + coef(linear_model)[2] * x_0

# Calculate the standard error of the predicted values
SE_expected_yield <- sqrt(error_variance * (1/n + (x_0 - mean(x))^2 / sum((x - mean(x))^2)))

# Calculate the critical t-value
t_critical <- qt(alpha/2, length(x) - 2)

# Calculate the confidence interval
lower_bound <- expected_yield - t_critical * SE_expected_yield
upper_bound <- expected_yield + t_critical * SE_expected_yield

# Display the results
expected_yield
```

```
## (Intercept)
##    6.419986
```

```
lower_bound
```

```
## (Intercept)
##    6.743651
```

```
upper_bound
```

```
## (Intercept)
##    6.096321
```

---

## g) Prediction of the yield of new rice variety

```r
# Calculate the standard error of the prediction
SE_prediction <- sqrt(error_variance * (1 + 1/n + (x_0 - mean(x))^2 / sum((x - mean(x))^2)))

# Calculate the prediction interval
lower_bound_prediction <- expected_yield - t_critical * SE_prediction
upper_bound_prediction <- expected_yield + t_critical * SE_prediction

# Display the results
lower_bound_prediction
```

```
## (Intercept)
##    7.363934
```

```r
upper_bound_prediction
```

```
## (Intercept)
##    5.476038
```

Comparing the results from f, new variety of rice has a wider 95% prediction interval.

---

## h) Compute R2 and interpret the results

```r
R_squared <- summary(linear_model)$r.squared

R_squared
```

```
## [1] 0.7546518
```

Interpretation : A higher r square suggests that the linear regression model does a good job of explaining the variability in grain yield based on plant height.

==================================================================

### Q2

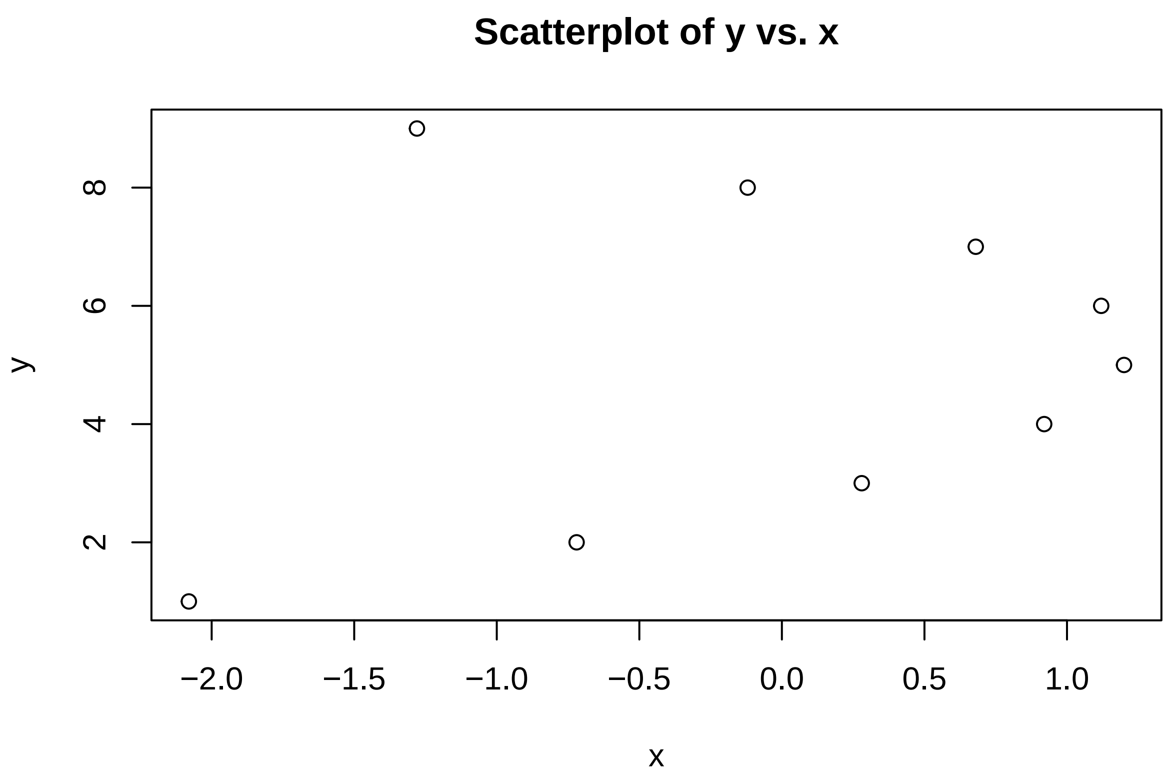
### Answers

```r
# Given artificial data
x <- c(1, 2, 3, 4, 5, 6, 7, 8, 9)
y <- c(-2.08, -0.72, 0.28, 0.92, 1.20, 1.12, 0.68, -0.12, -1.28)

# Fitting a linear model
linear_model <- lm(y ~ x)

summary(linear_model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -1.68  -0.42   0.48   1.02   1.20
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5000     0.8674  -0.576    0.582
## x             0.1000     0.1541   0.649    0.537
##
## Residual standard error: 1.194 on 7 degrees of freedom
## Multiple R-squared:  0.05672,    Adjusted R-squared:  -0.07804
## F-statistic: 0.4209 on 1 and 7 DF,  p-value: 0.5372
```

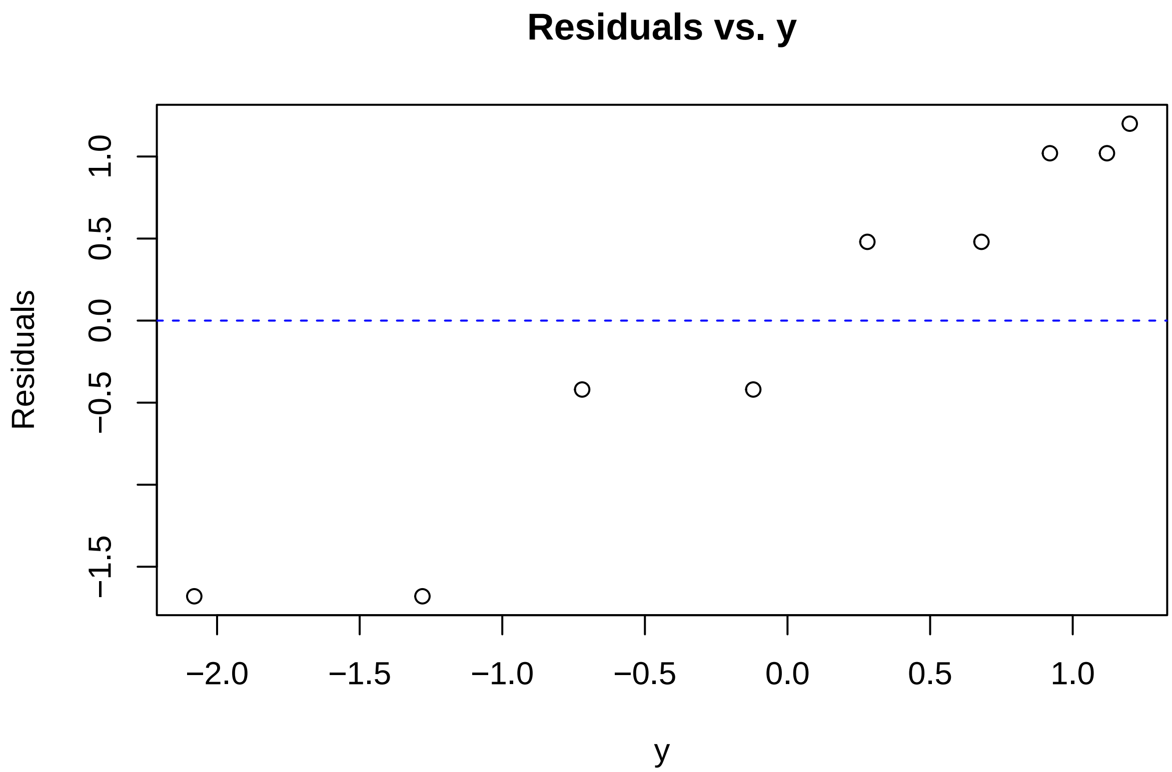
## a) Plot y vs x

```
plot(y, x, main = "Scatterplot of y vs. x", xlab = "x", ylab = "y")
```



## b) Plot the raw residuals vs. y

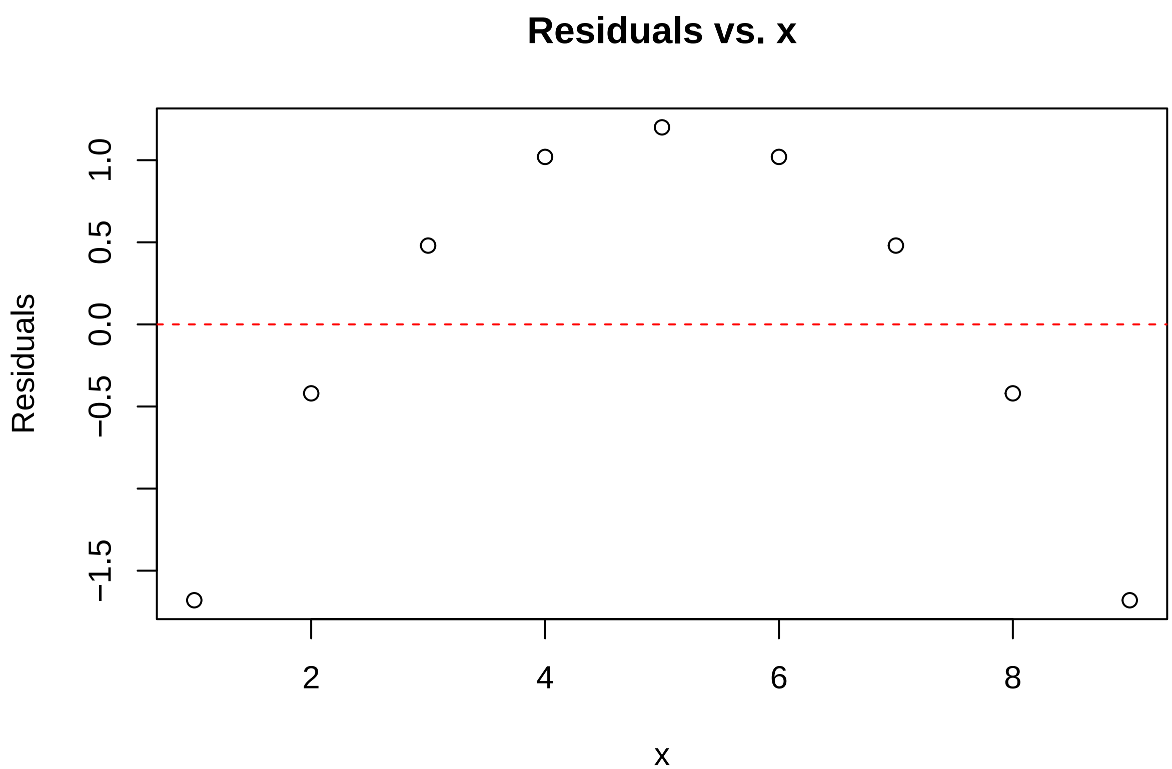
```r
raw_residuals <- residuals(linear_model)

plot(y, raw_residuals, main = "Residuals vs. y", xlab = "y", ylab = "Residuals")
abline(h = 0, col = "blue", lty = 2)
```




---

## c) Plot raw residuals vs x

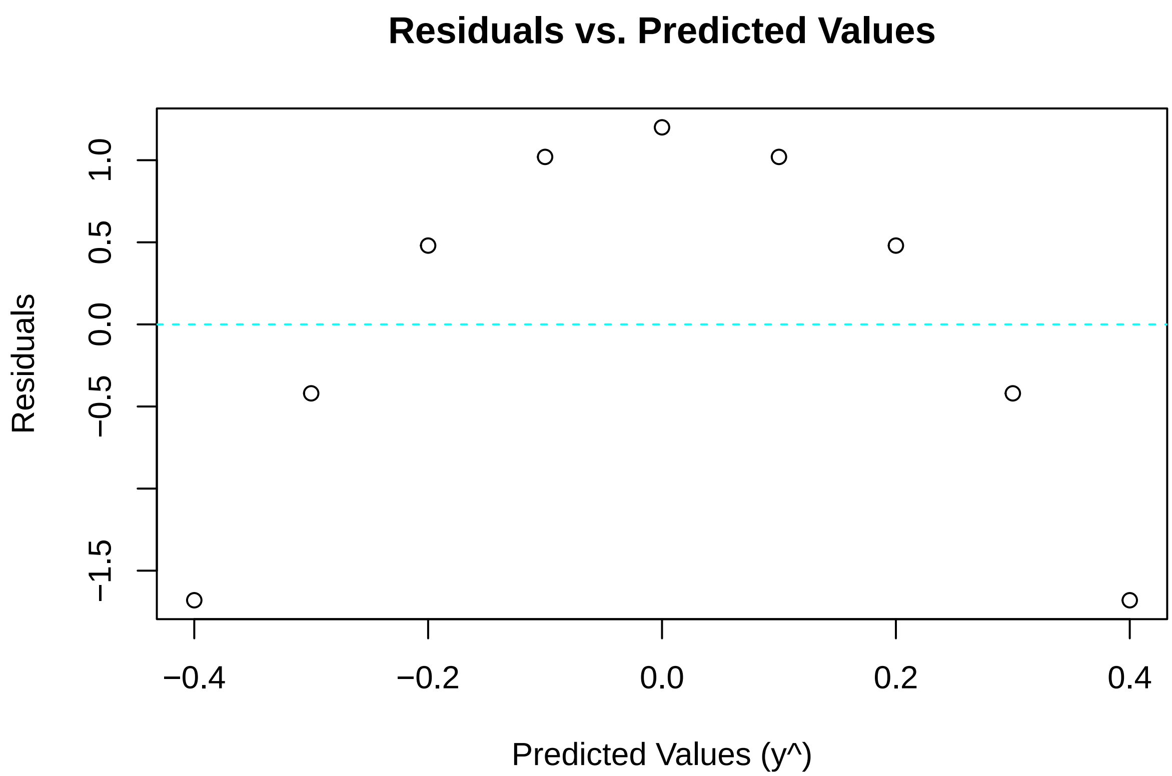
```r
# Plot residuals against x
plot(x, raw_residuals, main = "Residuals vs. x", xlab = "x", ylab = "Residuals")
abline(h = 0, col = "red", lty = 2)
```




---

## d) plot raw residuals vs y^

```
predicted_values <- predict(linear_model)

plot(predicted_values, raw_residuals, main = "Residuals vs. Predicted Values", xlab = "Predicted Values
abline(h = 0, col = "cyan", lty = 2)
```




---

## e) Which explains the better model fit?

While (b) and (c) provide valuable information, (d) Residuals vs. yˆ gives a better indication of the lack of fit as it directly assesses the performance of the model in predicting the response variable `y`. If there is a pattern or trend in (d), it suggests that the linear model might not be appropriate for capturing the underlying relationship in the data.