Data Science Homework 6

N.N.

11/14/2021

Question 1a

```
d = data.frame(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))

lm_grain_yield = lm(y ~ x, data = d)
summary(lm_grain_yield)
```

```
##
## Call:
## lm(formula = y ~ x, data = d)
##
## 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
```

```
print(coef(lm_grain_yield)["x"])
```

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

Above is are the summary statistics for a simple linear regression model between variables plant height and grain yield in eight varieties of rice.

The least squares estimate of the slope of this regression is the estimated correlation between plant height and grain yield. In other words, for every 1 unit of measure plant height increases, its yield is estimated to decrease by ~.037 unit of measure.

Question 1b

```
#f-test
anova(lm_grain_yield)
```

```
## 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
```

```
#t-test
summary(lm_grain_yield)
```

```
##
## Call:
## lm(formula = y ~ x, data = d)
##
## 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
```

The result of both an f-test and t-test yield a p-value of ~.005, so we would reject the null hypothesis that the real correlation between plant height and yield is 0.

Question 1c

```
confint(lm_grain_yield)
```

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

```
print(qt(.05/2, 6)) #2.447
```

```
## [1] -2.446912
```

95% CI = Estimated Intercept +/- (t-score)*(std. error) = 10.137455 +/- (2.446912)*0.842265 = [8.07650745, 12.19840320]

Using the equation for the 95% confidence interval of the estimated intercept from lecture, I was able to replicate the results of the confint() function. According to the t-distributions we've used to fit our model, the confidence interval indicates that there is a 95% chance that the real value of the intercept falls within the [8.07650745, 12.19840320] interval.

Question 1d

```
regression = function(x_obs){
  y_pred[i] = 10.137455 - x_obs*0.037175 #regression equation
}
print(resid(lm_grain_yield))
```

```
##          1          2          3          4          5          6          7 
## -0.2746519 -0.2802428  0.2628757  0.2922999  0.0720958 -0.2610638 -0.3462643 
##          8 
##  0.5349514
```

The raw residuals are calculated by subracting the estimated value of y from the observed values.

Question 1e

The mean squared error (estimated error variance) is 0.13132 (From anova().

Question 1f

```
predict(lm_grain_yield, newdata = data.frame(x=100), interval = "confidence")
```

```
##        fit      lwr      upr
## 1 6.419986 6.096321 6.743651
```

The 95% confidence interval for this new data point is [6.096, 6.743]

Question 1g

```
predict(lm_grain_yield, newdata = data.frame(x=100), interval = "prediction")
```

```
##        fit      lwr      upr
## 1 6.419986 5.476038 7.363934
```

The 95% prediction interval for this new data point is [5.476038, 7.363934]. This interval is larger than the confidence interval, since it predicts what range this particular value falls in, which is more uncertain than the range of all predicted values in general. The confidence interval is smaller because it only takes into account the uncertainty of the entire sample.

Question 1h

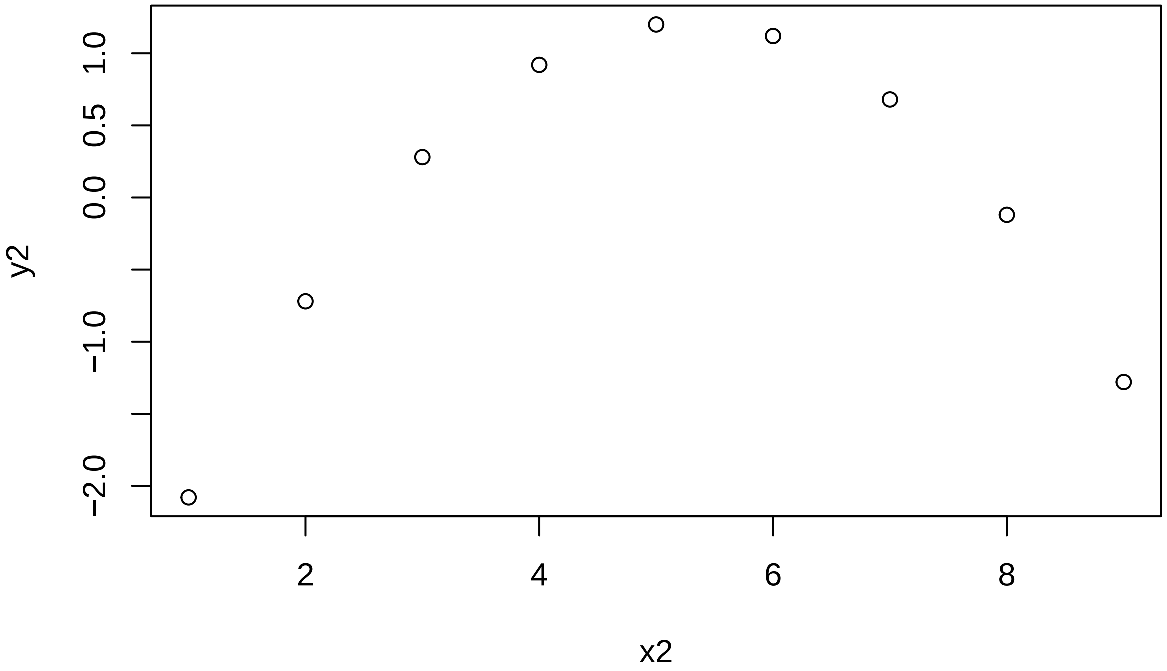
```
summary(lm_grain_yield)$r.squared
```

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

R^2 is .7546518.

Question 2a

```
x2 = c(1, 2, 3, 4, 5, 6, 7, 8, 9)
y2 = c(-2.08, -0.72, 0.28, 0.92, 1.20, 1.12, 0.68, -0.12, -1.28)
plot(x2, y2)
```

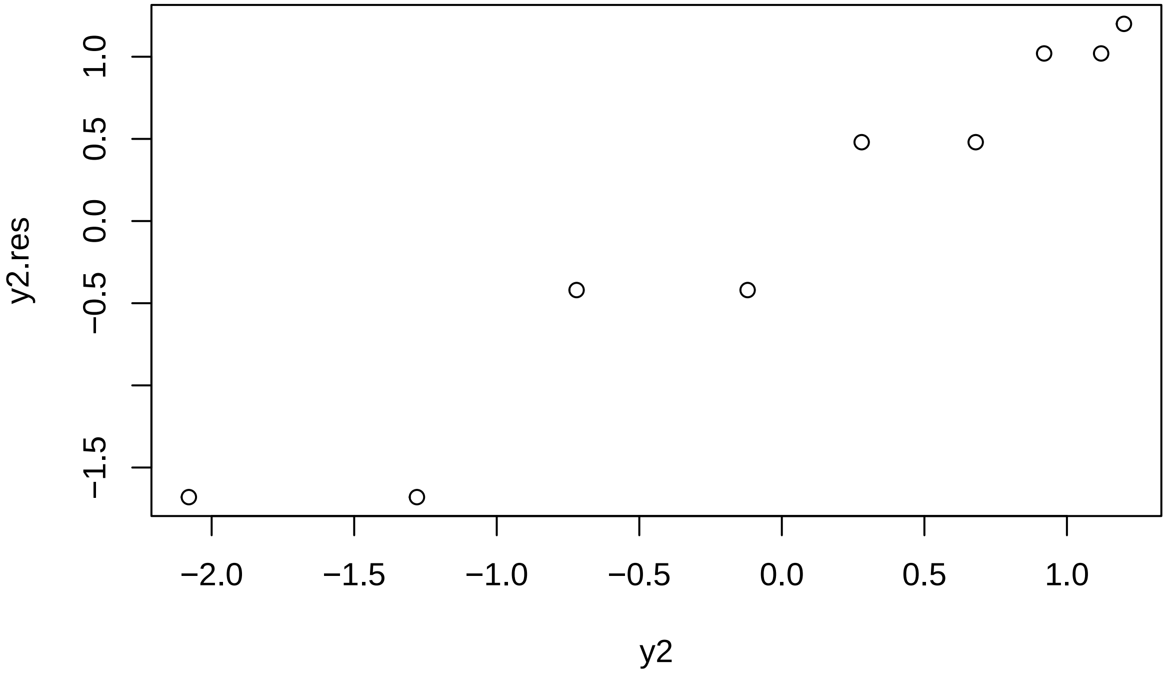


```
data_lm = lm(y2~x2)
summary(data_lm)
```

```
##
## Call:
## lm(formula = y2 ~ x2)
##
## 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
## x2            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
```

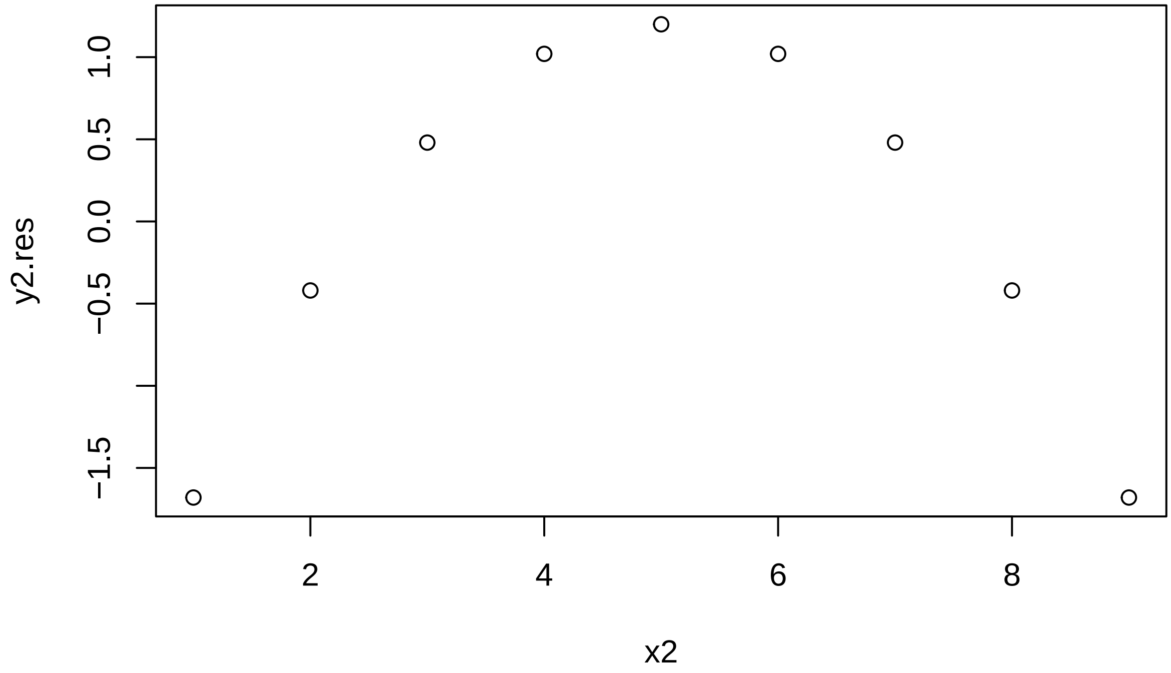
Question 2b

```
y2.res = resid(data_lm)
plot(y2, y2.res)
```



## Question 2c

```
plot(x2, y2.res)
```

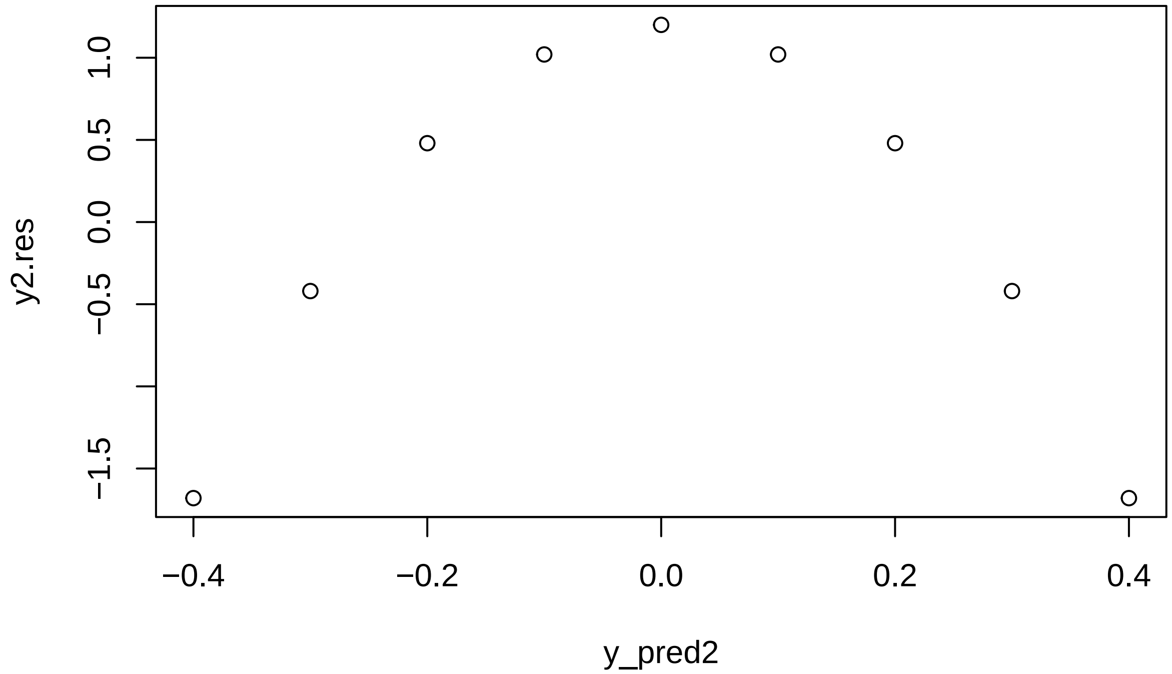


## Question 2d

```
y_pred2 = numeric(9)
for (i in 1:9){
  y_pred2[i] = -.5 + .1*x2[i]
}
print(y_pred2)
```

```
## [1] -0.4 -0.3 -0.2 -0.1  0.0  0.1  0.2  0.3  0.4
```

```
plot(y_pred2, y2.res)
```



## Question 1e Plotting the residuals vs x or estimated y values gives a similarly quadratic plot, since there is no real difference between either comparison. This is because we've fitted our data with a linear regression model, and since that's just a line, flipping between x and estimated y values on the x axis will just flip the plot horizontally. It's hard to see it in the above plot, because the graph happens to be symetrical.

Plot D ends up being more useful than plot B to demonstrate a lack of fit, since plotting the residuals vs fitted values should result in a random scatter plot if we've fit a model correctly. Our plot shows a distinct relationship between residuals and fitted values, which indicates that there is a lack of fit. Since there can be some correlation between Y and the residuals, regardless of the model's fit (or lack thereof), it can't be used as a measure of fit.