Homework 2

Rolando Santos

2023-09-28

Question 1: Consider a simple linear regression model $y = B_0 + B_1 x + \epsilon$. We fit this this model based on a dataset with test score (y) and training hours (x). The fitted model is $y = 10 + 0.56x$.

a. What is the fitted value of the response variable corresponding to $x = 7$?

$$\hat{y} = 10 + 0.56x$$
$$\hat{y} = 10 + 0.56 * 7$$
$$\hat{y} = 10 + 3.92$$
$$\hat{y} = 13.92$$

b. What is the residual corresponding to the data point with $x = 7$ and $y = 17$

$$e_i = y_i - \hat{y}_i$$
$$e_i = 17 - 13.92$$
$$e_i = 3.08$$

c. If the number of training hours is increased by 1, how is the expected test score affected?

The slope of $\hat{y}$ changes if x is increased by 1. $\hat{y}$ will increase by 0.56 (1 * 0.56).

d. Consider the data point in part b. An additional test score is to be obtained for a new observation at $x = 7$. Would the test score for the new observation necessarily be 17? Explain.

It could be, but it also can not be. x is a random variable from a normal distribution, the fitted value of x in our model is 13.92, so it is unlikely we would see 17 as the response.

Question 2: In this question, we will use the teengamb dataset. It concers a study of teenage gambling in Britain. Each row is one teenager's records. Download this dataset from Sakai and read it into R.

a. Fit a regression model with the expenditure on gambling as the response and sex, status, income and verbal score as predictors. Save the model output to a "model" object. Use the summary function to show the model output.

```r
teengamb <- read.csv("teengamb.csv")
head(teengamb)
```

```
##   sex status income verbal gamble
## 1   1     51   2.00      8    0.0
## 2   1     28   2.50      8    0.0
## 3   1     37   2.00      6    0.0
## 4   1     28   7.00      4    7.3
## 5   1     65   2.00      8   19.6
## 6   1     61   3.47      6    0.1
```

```r
teengamb$sex <- factor(teengamb$sex)
```

```r
model <- lm(gamble ~ sex + status + income + verbal, data = teengamb)
summary(model)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex1        -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

b. What percentage of variation in the response is explained by these predictors?

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

```
## [1] 0.5267234
```

$$R^2 = 0.5267234$$

According to the R^2 (from the summary(model)), ~53% of the variatio in the response is explained by these predictors.

c. Use model$residuals to show the residuals. Which observation has the largest (positive) residual?

```
print(paste("Observation with largest residual is:", which.max(model$residuals)))
```

```
## [1] "Observation with largest residual is: 24"
```

```
max(model$residuals)
```

```
## [1] 94.25222
```

```
model$residuals
```

```
##          1          2          3          4          5          6
##  10.6507430   9.3711318   5.4630298 -17.4957487  29.5194692  -2.9846919
##          7          8          9         10         11         12
##  -7.0242994 -12.3060734   6.8496267 -10.3329505   1.5934936  -3.0958161
##         13         14         15         16         17         18
##   0.1172839   9.5331344   2.8488167  17.2107726 -25.2627227 -27.7998544
##         19         20         21         22         23         24
##  13.1446553 -15.9510624 -16.0041386  -9.5801478 -27.2711657  94.2522174
##         25         26         27         28         29         30
##   0.6993361  -9.1670510 -25.8747696  -8.7455549  -6.8803097 -19.8090866
##         31         32         33         34         35         36
##  10.8793766  15.0599340  11.7462296  -3.5932770 -14.4016736  45.6051264
##         37         38         39         40         41         42
##  20.5472529  11.2429290 -51.0824078   8.8669438  -1.4513921  -3.8361619
##         43         44         45         46         47
##  -4.3831786 -14.8940753   5.4506347   1.4092321   7.1662399
```

d. Use model$fitted.values to show the fitted response. Compute the correlation of the residuals with the fitted response.

```
cor(model$residuals, model$fitted.values)
```

```
## [1] -6.215823e-17
```

e. Compute the correlation of the residuals with the income.

```
cor(model$residuals, teengamb$income)
```

```
## [1] 3.247058e-17
```

f. If all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?

From the summary we saw that the coefficient for sex is -22.11833. The dataset describes females = 1 and males = 0, so on average females spent 22.11833 less on gambling than males did.

Question 3: The dataset prostate comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. The description of each variable can be found at https://rafalab.github.io/pages/649/prostate.html. Download and import this dataset from Sakai, answer following questions.

```
prostate <- read.csv("prostate.csv")
head(prostate)
```

```
##       lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa
## 1 -0.5798185  2.7695  50 -1.386294   0 -1.38629       6     0 -0.43078
## 2 -0.9942523  3.3196  58 -1.386294   0 -1.38629       6     0 -0.16252
## 3 -0.5108256  2.6912  74 -1.386294   0 -1.38629       7    20 -0.16252
## 4 -1.2039728  3.2828  58 -1.386294   0 -1.38629       6     0 -0.16252
## 5  0.7514161  3.4324  62 -1.386294   0 -1.38629       6     0  0.37156
## 6 -1.0498221  3.2288  50 -1.386294   0 -1.38629       6     0  0.76547
```

a. Fit a regression model with lpsa as the response and lcavol as the predictor.Show the residual sum of square RSS and the R^2 of this model (hint: check deviance function for RSS).

```
model <- lm(lpsa ~ lcavol, data = prostate)
summary(model)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67625 -0.41648  0.09859  0.50709  1.89673
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.50730    0.12194   12.36   <2e-16 ***
## lcavol       0.71932    0.06819   10.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7875 on 95 degrees of freedom
## Multiple R-squared:  0.5394, Adjusted R-squared:  0.5346
## F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16
```

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

```
## [1] 0.5394319
```

$$R^2 = 0.5394319$$

```r
deviance(model)
```

```
## [1] 58.91476
```

$$RSS = 58.91476$$

b. Add lweight, svi, lbph, age, lcp, pgg45 and gleason as predictors to theregression model. Show the residual sum of square (RSS) and the R^2 of this model

```r
model <- lm(lpsa ~ lcavol + lweight + svi + lbph + age + lcp + pgg45 + gleason, data = prostate)
summary(model)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age + lcp +
##     pgg45 + gleason, data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## svi          0.766157   0.244309   3.136  0.00233 **
## lbph         0.107054   0.058449   1.832  0.07040 .
## age         -0.019637   0.011173  -1.758  0.08229 .
## lcp         -0.105474   0.091013  -1.159  0.24964
## pgg45        0.004525   0.004421   1.024  0.30886
## gleason      0.045142   0.157465   0.287  0.77503
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

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

```
## [1] 0.6547541
```

$$R^2 = 0.6547541$$

```r
deviance(model)
```

```
## [1] 44.16302
```

$$RSS = 44.16302$$

c. Compare the RSS and R^2 of these two models. Explain why you observe such a comparison result.

From the R^2 increasing we can see that we achieved a model with a better fit by adding more predictors. A lower RSS also shows that we start observing a model with a better fit. By adding more predictors we can better explain our response (until we start seeing penalties from the adjusted R^2).

d. Use the method introduced in lecture slides to manually fit the model in b. First construct a design matrix X, then a response vector y, and finally use the formula of parameter estimation. Compare the manually estimated parameters with the result from the lm function.

```
y <- prostate$lpsa
X <- model.matrix(~lcavol + lweight + svi + lbph + age + lcp + pgg45 + gleason, data = prostate)

XtXi <- solve(t(X)%*%X)
XtXi%*%t(X)%*%y
```

```
##                     [,1]
## (Intercept)  0.669336698
## lcavol       0.587021826
## lweight      0.454467424
## svi          0.766157326
## lbph         0.107054031
## age         -0.019637176
## lcp         -0.105474263
## pgg45        0.004525231
## gleason      0.045141598
```

```
model$coefficients
```

```
##  (Intercept)       lcavol      lweight          svi         lbph          age
##  0.669336698  0.587021826  0.454467424  0.766157326  0.107054031 -0.019637176
##          lcp        pgg45      gleason
## -0.105474263  0.004525231  0.045141598
```

We can see that the coefficients created manually are the same as the coefficients done using the lm() function.

Question 4: Thirty samples of cheddar cheese were analyzed for their content of acetic acid, hydrogen sulfide and lactic acid. Each sample was tasted and scored by a panel of judges and the average taste score produced. Use the cheddar dataset from Sakai to answer the following questions.

```
cheddar <- read.csv("cheddar.csv")
head(cheddar)
```

```
##   taste Acetic   H2S Lactic
## 1  12.3  4.543 3.135   0.86
## 2  20.9  5.159 5.043   1.53
```

```
## 3  39.0  5.366 5.438   1.57
## 4  47.9  5.759 7.496   1.81
## 5   5.6  4.663 3.807   0.99
## 6  25.9  5.697 7.601   1.09
```

a. Fit a regression model with taste as the response and the three chemical contents as predictors. Report the values of the regression coefficients.

```
model <- lm(taste ~ Acetic + H2S + Lactic, data = cheddar)
summary(model)
```

```
##
## Call:
## lm(formula = taste ~ Acetic + H2S + Lactic, data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## Acetic        0.3277     4.4598   0.073  0.94198
## H2S           3.9118     1.2484   3.133  0.00425 **
## Lactic       19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

```
model$coefficients
```

```
## (Intercept)      Acetic         H2S      Lactic
## -28.8767696   0.3277413   3.9118411  19.6705434
```

b. Compute the correlation between the fitted values and the true response. What information can you learn from this correlation?

```
cor(cheddar$taste, model$fitted.values)
```

```
## [1] 0.8073256
```

We can observe that there is a high positive correlation between the true response and the fitted values from the model.

c. How do you interpret the value of intercept in this model? Does this value make sense in this setting (tasting cheese)?

```
cheddar$taste
```

```
##  [1] 12.3 20.9 39.0 47.9  5.6 25.9 37.3 21.9 18.1 21.0 34.9 57.2  0.7 25.9 54.9
## [16] 40.9 15.9  6.4 18.0 38.9 14.0 15.2 32.0 56.7 16.8 11.6 26.5  0.7 13.4  5.5
```

The value of the intercept can be viewed as the base taste value if all predictors are 0. This might not make sense in the context of the dataset because the taste of cheese having a negative score if all predictors are 0 (or close to 0) wouldn't make sense considering all taste values in the dataset are positive numbers.

Question 5: Run the following R code:

a. Explain what the code does. Use ?function_name() or Google if you do not know the meaning of any function.

```
#Setting the seed means that whenever we rerun any random sampling
#or generate random data, the same random data will appear
#when rerunning the code.
set.seed(1234)

#Generates 100 random uniform variables from a normal distribution
#between and including 0 and 10, the max and min values.
x <- runif(100,0,10)

#Response (y) is created using x and rnorm() generated 100
#random uniform variables from a normal distribution with 0 as
#the mean and 1 as the standard deviation.
y <- 3+x+x^2+rnorm(100,0,1)
```

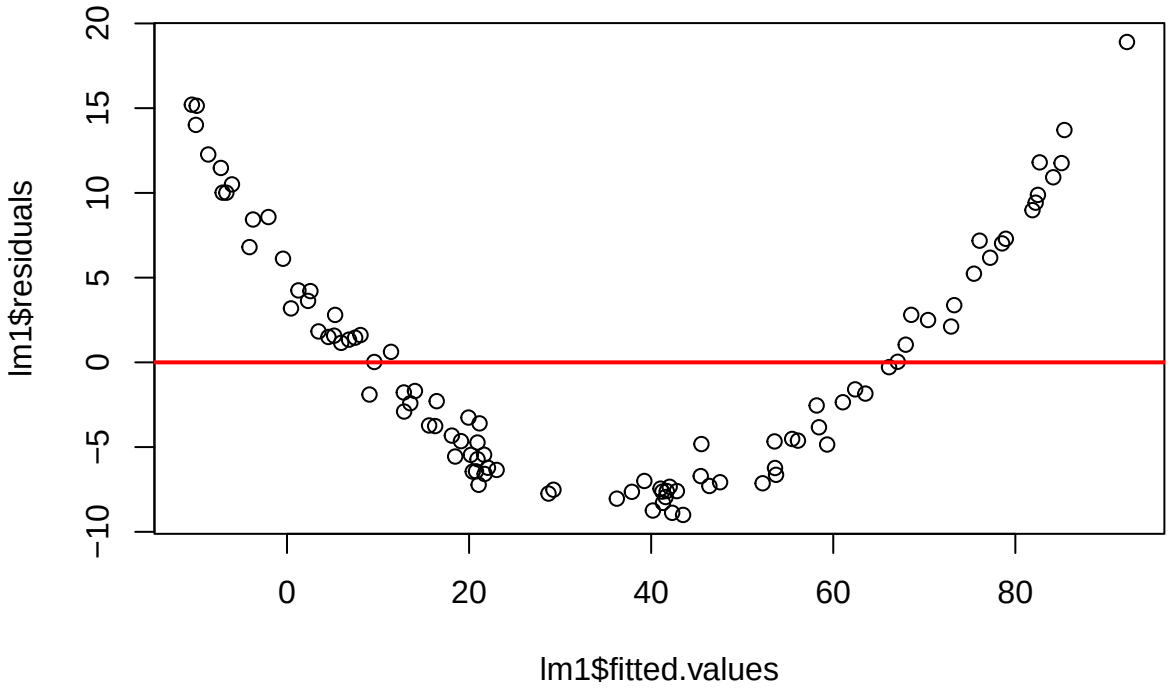
Once you have generated x and y, fit the following two linear models:

```
#Create a single linear regression model with y as the
#response and x as the predictor.
lm1 <- lm(y~x)

#Creates a multiple linear regression model with y as the
#response and x and x^2 as the predictors. I() is used to
#protect data type of x^2 as to not convert it into
#factors or other unwanted data types.
lm2 <- lm(y~x+I(x^2))
```

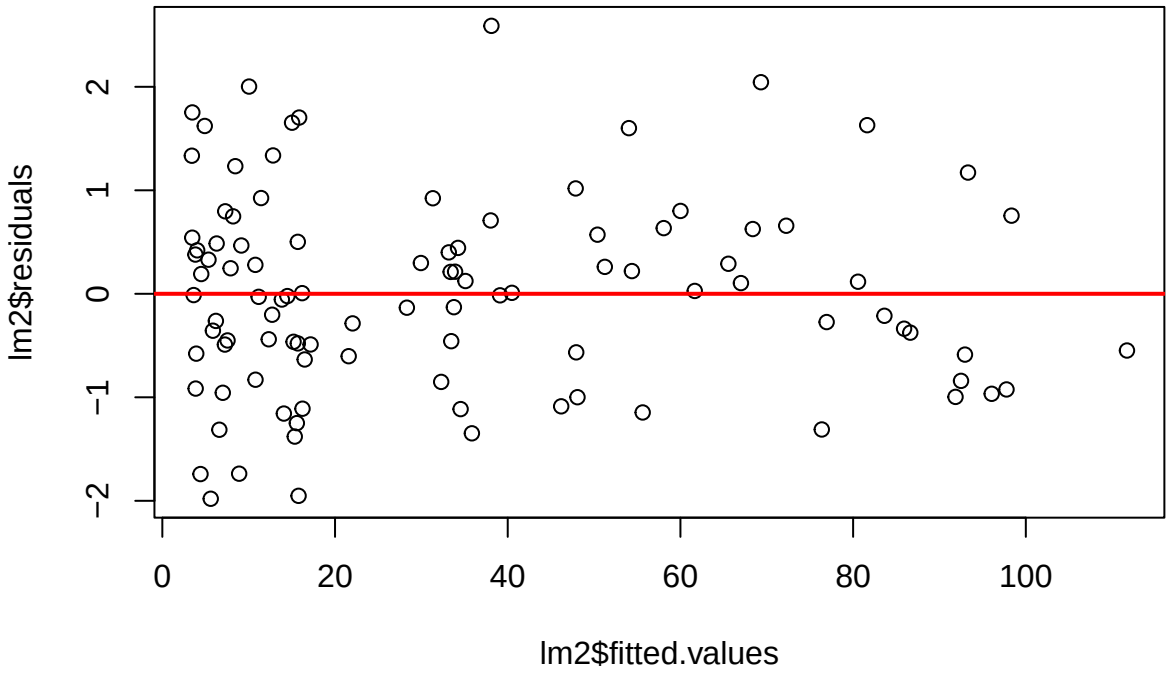
b. For both models, plot the residual versus the fitted response. Describe the pattern you observed in the plots.

```
plot(lm1$fitted.values, lm1$residuals)
abline(h=0 ,col='red', lwd=2)
```



This plot appears to suggest a non-linear relationship between the residuals and fitted values.

```
plot(lm2$fitted.values, lm2$residuals)
abline(h=0 ,col='red', lwd=2)
```



This model suggests a more linear (in comparison to the first plot) relationship between the fitted values residuals.

c. Which model is better? Give your reason.

The second model for $y = 3 + x + x^2 + \epsilon$ with predictors x and x^2 appears as the better model. The first plot suggests that the predictor(s) provided did not have a proper fit for the response (suggested by the quadratic appearance of the plot points). The points in the second plot appear more random, which makes sense considering we use both x and x^2 as predictors, and we know x and x^2 are actual predictors that contribute to calculating the response y.