Homework One

Chiayu Tu (Louis Tu)

2022-10-09

Question One

1. The dataset teengamb concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as the response and the sex, status, income and verbal score as predictors. Present the output.

(a) What percentage of variation in the response is explained by these predictors?
(b) Which observation has the largest (positive) residual? Give the case number.
(c) Compute the mean and median of the residuals.
(d) Compute the correlation of the residuals with the fitted values.
(e) Compute the correlation of the residuals with the income.
(f) For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?

```
#Question One a
#Summary of teengamb
t_model <- lm(gamble~.,data = teengamb)
summary(t_model)
```

```
##
## Call:
## lm(formula = gamble ~ ., 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
## sex         -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
```

```
#Question One b
#Find the largest residual
largest_residual <- max(t_model$residuals)
largest_residual_num <- t_model$residuals[t_model$residuals == largest_residual]
largest_residual_num
```

```
##       24
## 94.25222
```

```
print("The largest residual is 94, and the case number is 24")
```

```
## [1] "The largest residual is 94, and the case number is 24"
```

```
#Question One c
#Mean
residual_mean <- mean(t_model$residuals)
print(paste("The residual mean   is: ", residual_mean))
```

```
## [1] "The residual mean   is:  -3.06529329106115e-17"
```

```
#Medium
residual_medium <- median(t_model$residuals)
print(paste("The residual medium is: ", residual_medium))
```

```
## [1] "The residual medium is:  -1.45139206896952"
```

```
#Question One d
cor_fit.value <- zapsmall(cor(fitted(t_model), resid(t_model)))
print("The correlation of the residuals with the fitted values is: ")
```

```
## [1] "The correlation of the residuals with the fitted values is: "
```

```
print(cor_fit.value)
```

```
## [1] -1.070659e-16
```

```
#Question One e
print("Compute the correlation of the residuals with the income is: ")
```

```
## [1] "Compute the correlation of the residuals with the income is: "
```

```
cor(teengamb$income, resid(t_model))
```

```
## [1] -7.242382e-17
```

Question Two

2. The dataset uswages is drawn as a sample from the Current Population Survey in 1988. Fit a model with weekly wages as the response and years of education and experience as predictors. Report and give a simple interpretation to the regression coefficient for years of education. Now fit the same model but with logged weekly wages. Give an interpretation to the regression coefficient for years of education. Which interpretation is more natural?

```
view(uswages)
E_model <- lm(formula = wage ~ educ + exper, data = uswages)
summary(E_model)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper, data = uswages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1018.2  -237.9   -50.9   149.9  7228.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -242.7994    50.6816  -4.791 1.78e-06 ***
## educ          51.1753     3.3419  15.313  < 2e-16 ***
## exper          9.7748     0.7506  13.023  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 427.9 on 1997 degrees of freedom
## Multiple R-squared:  0.1351, Adjusted R-squared:  0.1343
## F-statistic:   156 on 2 and 1997 DF,  p-value: < 2.2e-16
```

```
E_model_1 <- lm(formula = log(wage) ~ educ + exper - 1, data = uswages)
summary(E_model_1)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper - 1, data = uswages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7681 -0.4896  0.2144  0.8670  7.5580
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## educ  0.380897   0.002758  138.09   <2e-16 ***
## exper 0.054867   0.001630   33.65   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.099 on 1998 degrees of freedom
## Multiple R-squared:  0.9687, Adjusted R-squared:  0.9687
## F-statistic: 3.093e+04 on 2 and 1998 DF,  p-value: < 2.2e-16
```

Question Three

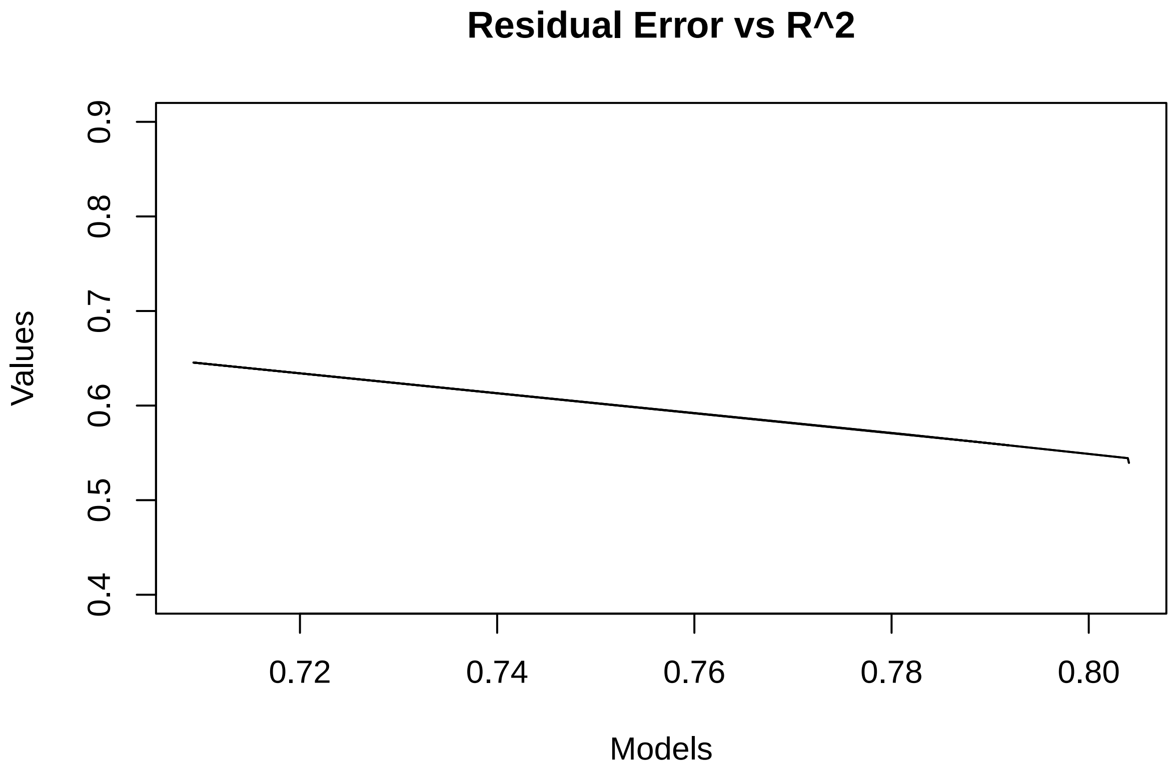
4. The dataset prostate comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Fit a model with lpsa as the response and lcavol as the predictor. Record the residual standard error and the R^2. Now add lweight, svi, lbph, age, lcp, pgg45 and gleason to the model one at a time. For each model record the residual standard error and the R^2. Plot the trends in these two statistics.

```
#view(prostate)
P_model <- lm(lcavol ~ lpsa, data = prostate)
summary(P_model)
```

```
##
## Call:
## lm(formula = lcavol ~ lpsa, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15948 -0.59383  0.05034  0.50826  1.67751
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.50858    0.19419  -2.619   0.0103 *
## lpsa         0.74992    0.07109  10.548   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8041 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
```

```
model1 <- lm(lcavol ~ lpsa, data = prostate)
v <- summary(model1)
R2<- c()
Sgma <- c()
R2 = v$r.squared
Sgma = v$sigma
for(i in 1:7){
  model.temp = lm(prostate$lcavol ~ prostate$lpsa + prostate[,i+1], data = prostate)
  v = summary(model.temp)
  R2[i+1] = v$r.squared
  Sgma[i+1] = v$sigma
}


plot(Sgma, R2, type = "l", main = "Residual Error vs R^2", xlab = "Models", ylab = "Values" , ylim = c(
points(Sgma, R2, type = "l" )
```



Qurstion Four

6. 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 data to answer the following:

(a) Fit a regression model with taste as the response and the three chemical contents as predictors. Report the values of the regression coefficients.
(b) Compute the correlation between the fitted values and the response. Square it. Identify where this value appears in the regression output.
(c) Fit the same regression model but without an intercept term. What is the value of R^2 reported in the output? Compute a more reasonable measure of the goodness of fit for this example.

```r
#view(cheddar)
C_model <- lm(taste ~ Acetic + H2S + Lactic, data = cheddar)
summary(C_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
```

```
#b

(cor(cheddar$taste,cheddar$Acetic))^2
```

```
## [1] 0.3019934
```

```
(cor(cheddar$taste,cheddar$H2S))^2
```

```
## [1] 0.5711615
```

```
(cor(cheddar$taste,cheddar$Lactic))^2
```

```
## [1] 0.4959486
```

```
(cor(C_model$fitted.values,cheddar$Acetic))^2
```

```
## [1] 0.4633402
```

```
(cor(C_model$fitted.values,cheddar$H2S))^2
```

```
## [1] 0.8763174
```

```
(cor(C_model$fitted.values,cheddar$Lactic))^2
```

```
## [1] 0.7609203
```

```
#c
#without an intercept term
C_model_1 <- lm(taste ~ 0 + Acetic + H2S + Lactic, data = cheddar)
summary(C_model_1)
```

```
##
## Call:
## lm(formula = taste ~ 0 + Acetic + H2S + Lactic, data = cheddar)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4521  -6.5262  -0.6388   4.6811  28.4744
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## Acetic   -5.454      2.111  -2.583  0.01553 *
## H2S       4.576      1.187   3.854  0.00065 ***
## Lactic   19.127      8.801   2.173  0.03871 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 10.34 on 27 degrees of freedom
## Multiple R-squared:  0.8877, Adjusted R-squared:  0.8752
## F-statistic: 71.15 on 3 and 27 DF,  p-value: 6.099e-13
```

```
(cor(C_model_1$fitted.values,cheddar$Acetic))^2
```

```
## [1] 0.2705027
```

```
(cor(C_model_1$fitted.values,cheddar$H2S))^2
```

```
## [1] 0.8796605
```

```
(cor(C_model_1$fitted.values,cheddar$Lactic))^2
```

```
## [1] 0.6859282
```