Charles Hwang
Dr. Xi
STAT 408-001
27 September 2022

Homework 2

1A) $y = 10 + 0.56x$

$y = 10 + 0.56(7)$

$y = 10 + 3.92$

$\boxed{\hat{y} = 13.92}$

1B) $y - \hat{y} =$

$17 - 13.92 =$

$\boxed{4.08}$

1C) We can see from the coefficient that the expected test score increases by 10 for every additional training hour.

1D) No, the test score for a new observation at $x=7$ training hours would not necessarily be $y=17$. It is very possible for the new observation's test score to be different. Residuals show that there can be very different values of y for observations with the same x.

Homework 2

Charles Hwang

9/27/2022

Problem 2

```
rm(list=ls())
g<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analysis/teengamb.csv")
```

Problem 2a

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

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = g)
##
## 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
```

Problem 2b

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

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

We can see from the output that $r^2 = 0.5267234$. Approximately 52.6723413 percent of variation in the response variable is explained by these four predictor variables.

Problem 2c

```
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
```

```
sort(model$residuals)[length(model$residuals)]
```

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

We can see from sorting the residuals that observation 24 has the largest positive residual (94.2522174).

Problem 2d

```
model$fitted.values
```

```
##           1           2           3           4           5           6
## -10.6507430  -9.3711318  -5.4630298  24.7957487  -9.9194692   3.0846919
##           7           8           9          10          11          12
##   8.4742994  18.9060734  -5.1496267  10.4329505  -1.4934936   8.4958161
##          13          14          15          16          17          18
##   1.0827161  -5.9331344  -0.4488167 -13.8107726  25.3627227  36.1998544
##          19          20          21          22          23          24
##  -1.1446553  15.9510624  17.0041386  10.7801478  27.3711657  61.7477826
##          25          26          27          28          29          30
##  37.8006639  11.2670510  40.3747696  11.7455549   7.4803097  29.4090866
##          31          32          33          34          35          36
##  77.1206234  38.1400660  78.2537704   6.5932770  28.5016736  24.3948736
##          37          38          39          40          41          42
##  17.9527471  45.9570710  57.0824078  16.1330562   8.3513921  73.5361619
##          43          44          45          46          47
##  17.6831786  15.4940753  32.5493653  12.9907679  12.0337601
```

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

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

We can see the correlation between the residuals and the fitted response is $r = -1.0706588 \times 10^{-16} \approx 0$, which is expected.

Problem 2e

```r
cor(model$residuals,g$income)
```

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

We can see the correlation between the residuals and the income variable is $r = -7.2423817 \times 10^{-17} \approx 0$, which is expected.

Problem 2f

```r
summary(model)$coefficients["sex","Estimate"] # 0 = Male; 1 = Female
```

```
## [1] -22.11833
```

We can see the predicted annual expenditure on gambling for a male is approximately £22.1183301 greater than the predicted annual expenditure on gambling for a female, holding all other predictors constant.

Problem 3

```r
p<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analysis/prostate.csv")
```

Problem 3a

```r
summary(lm(lpsa~lcavol,data=p))
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol, data = p)
##
## 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
```

```r
summary(lm(lpsa~lcavol,data=p))$sigma
```

```
## [1] 0.7874994
```

```r
summary(lm(lpsa~lcavol,data=p))$r.squared
```

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

We can see that the residual sum of squares is 0.7874994 and $r^2 = 0.5394319$.

Problem 3b

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

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age + lcp +
##     pgg45 + gleason, data = p)
##
## 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(lm(lpsa~lcavol+lweight+svi+lbph+age+lcp+pgg45+gleason,data=p))$sigma
```

```
## [1] 0.7084155
```

```r
summary(lm(lpsa~lcavol+lweight+svi+lbph+age+lcp+pgg45+gleason,data=p))$r.squared
```

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

We can see that the residual sum of squares is 0.7084155 and $r^2 = 0.6547541$.

Problem 3c

```r
summary(lm(lpsa~lcavol,data=p))$sigma
```

```
## [1] 0.7874994
```

```r
summary(lm(lpsa~lcavol+lweight+svi+lbph+age+lcp+pgg45+gleason,data=p))$sigma
```

```
## [1] 0.7084155
```

```r
summary(lm(lpsa~lcavol,data=p))$r.squared
```

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

```r
summary(lm(lpsa~lcavol+lweight+svi+lbph+age+lcp+pgg45+gleason,data=p))$r.squared
```

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

We observe different values of RSS and r^2 because the two models are different: the model from problem 3a is more basic with only one predictor variable, while the model from problem 3b is more complex with eight predictor variables. We can see the model with more predictor variables has a higher r^2 and a lower RSS, which is generally expected.

Problem 3d

```
X<-model.matrix(~lcavol+lweight+svi+lbph+age+lcp+pgg45+gleason,data=p)
Y<-matrix(p[,"lpsa"])
solve(t(X)%*%X)%*%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
```

```
summary(lm(lpsa~lcavol+lweight+svi+lbph+age+lcp+pgg45+gleason,data=p))$coefficients[,"Estimate"]
```

```
##  (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 the manually estimated parameters from using the design matrix are the same as the parameters calculated with the linear model, which makes sense intuitively as they are both calculating the same thing.

Problem 4

```
c<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analysis/cheddar.csv")
```

Problem 4a

```
summary(lm(taste~Acetic+H2S+Lactic,data=c))
```

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

```
summary(lm(taste~Acetic+H2S+Lactic,data=c))$coefficients[,"Estimate"]
```

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

Problem 4b

```
cor(lm(taste~Acetic+H2S+Lactic,data=c)$fitted.values,c$taste)
```

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

We can see there is a strong to very strong correlation ($r = 0.8073256$) between the fitted values from the model and the true values from the response variable. This suggests the model predicts the response variable well.

Problem 4c

```
summary(lm(taste~Acetic+H2S+Lactic,data=c))
```

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

```
summary(lm(taste~Acetic+H2S+Lactic,data=c))$coefficients["(Intercept)","Estimate"]
```

```
## [1] -28.87677
```

For a cheese with no acetic acid, hydrogen sulfide, or lactic acid content, the linear regression model predicts that the average taste score produced by a panel of judges would be -28.8767696. This does not make sense in the context of this problem because a cheese with no acetic acid, hydrogen sulfide, or lactic acid content would mean that it has not aged at all and likely would not be considered by judges. The negative score also likely does not make sense.

Problem 5

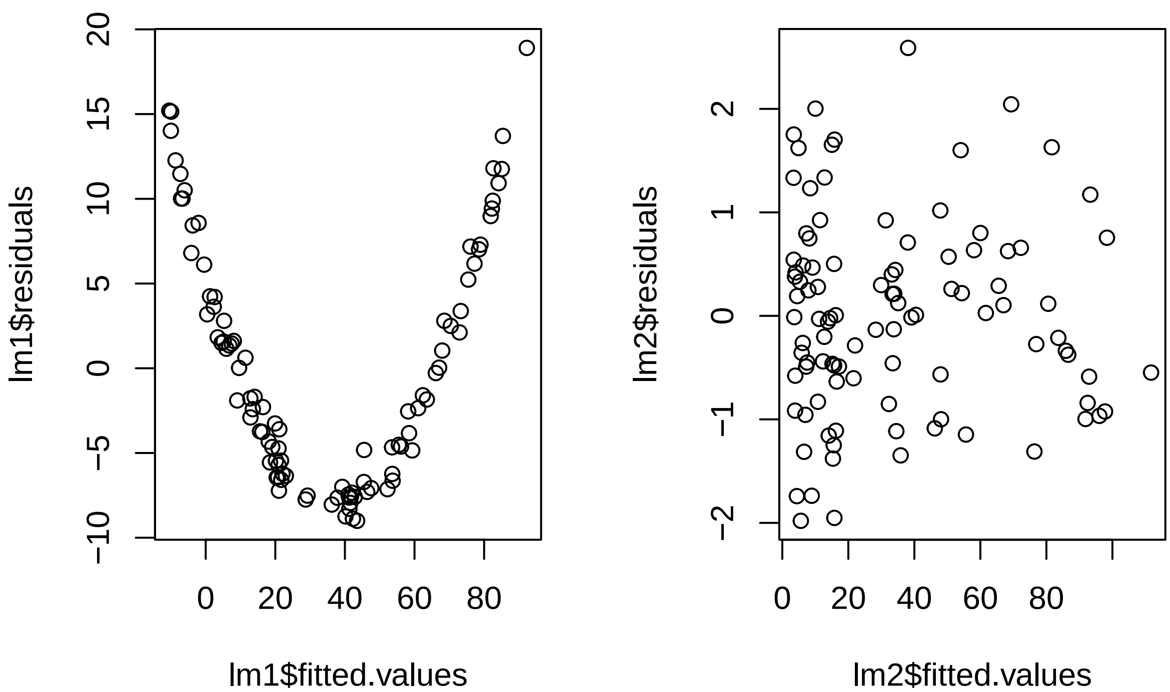
```
set.seed(1234)
x<-runif(100,0,10)
y<-3+x+x^2+rnorm(100,0,1)
lm1<-lm(y~x)
lm2<-lm(y~x+I(x^2))
```

Problem 5a

The code sets the seed number for random number generation (RNG) to `1234`; randomly samples `100` values from a uniform distribution $\in [0, 10]$ and stores them as x; randomly samples `100` values from a standard normal distribution $N(0, 1)$ and adds `3` to them, storing them as part of a quadratic equation $y = 3 + x + x^2 + N(0, 1)$; and fits two linear models: one of y on x, and the other of y on x and its quadratic term x^2. We can see that $N(0, 1)$ acts as the random error term ϵ in a linear model.

Problem 5b

```
par(mfrow=c(1,2))
plot(lm1$residuals~lm1$fitted.values)
plot(lm2$residuals~lm2$fitted.values)
```



We can see a clear positive quadratic pattern in the first plot. There is no clear pattern in the second plot, but there is a reverse megaphone effect and some clustering near the y-axis.

Problem 5c

The second model is better. Patterns in residual plots indicate an inappropriate model choice, and we can see there is clearly a pattern in the first plot with the linear model, as the response variable y is quadratic. A linear model is not an appropriate choice for these data.