CSC 423 Homework 3

Akhil Kumar Ramasagaram

April 16, 2016

Deep space survey of quasars

a) Hypothesize a first-order model for equivalent width, y, as a function of the first four variables.
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

b) Give the least squares prediction equation.

```
load("rdata/QUASAR.Rdata")
lm_model <- lm(RFEWIDTH ~ REDSHIFT + LINEFLUX + LUMINOSITY + AB1450, data = QUASAR)
summary(lm_model)
```

```
##
## Call:
## lm(formula = RFEWIDTH ~ REDSHIFT + LINEFLUX + LUMINOSITY + AB1450,
##     data = QUASAR)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.757  -9.039  -2.250   1.756  48.628
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21087.951  18553.161   1.137   0.2691
## REDSHIFT      108.451     88.740   1.222   0.2359
## LINEFLUX      557.910    315.990   1.766   0.0927 .
## LUMINOSITY   -340.166    320.763  -1.060   0.3016
## AB1450         85.681      6.273  13.658 1.34e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.42 on 20 degrees of freedom
## Multiple R-squared:  0.9118, Adjusted R-squared:  0.8942
## F-statistic: 51.72 on 4 and 20 DF,  p-value: 2.867e-10
```

```
lm_coeff <- summary(lm_model)$coefficients[,1]
lm_coeff
```

```
## (Intercept)    REDSHIFT    LINEFLUX  LUMINOSITY      AB1450
## 21087.95124   108.45084   557.90980  -340.16553    85.68102
```

The least square prediction equation is $RFEWIDTH = 21087.95 + 108.45 * REDSHIFT + 557.90 * LINEFLUX - 340.165 * LUMINOSITY + 85.68 * AB1450$

c) Interpret the β estimates in the model.
β_0 is the intercept on y axis which is the value when all our predictors are zero. β_1 is our first predictor

which is REDSHIFT, the same can be said for $\beta_2, \beta_3, \beta_4$ for LINEFLUX, LUMINOSITY & AB1450. For every 1 unit increase in REDSHIFT the RFEWIDTHincreases by 108.45 value. The same can be interpreted for other β values. A positive/negative represents increase of decrease for unit increase.

d) Test to determine whether redshift (x1) is a useful linear predictor of equivalent width (y), using $\alpha = .05$.

At $\alpha = 0.05$, the p-value of redsgift is 0.23, which is more than α, we can conclude that this is not a useful linear predictor.

e) Locate R^2 and R_a^2 from the output. Interpret these values. Which statistic is the preferred measure of model fit? Explain?

The R^2 and R_a^2 from the text book output is 0.912 & 0.894 which can be located in the second table under model summary. The R-Square is more related to individual features performance and adjusted R-square is the performance of all the variable combined. Since we user multiplt variable in our model we should use adjusted R-square here.

f) Locate the global F-value?

we can locate the F-value in the third table(ANOVA) which is 51.720.

Removing oil from a water/oil mix.

```
load("rdata/WATEROIL.Rdata")
lm_model <- lm(VOLTAGE ~  VOLUME + SALINITY + SURFAC, data = WATEROIL)
predict(lm_model, newdata = data.frame('VOLUME' = 80,'SALINITY' = 1, 'SURFAC' =2), interval = 'predicti
```

```
##           fit       lwr     upr
## 1 -0.09795082 -1.233442 1.03754
```

the above prediction interval is an estimate of an interval in which our future observations will fall, with a 95% probability, given what has already been observed.

Arsenic in groundwater.

```
load("rdata/ASWELLS.Rdata")
lm_model <- lm(ARSENIC  ~ (LATITUDE + LONGITUDE) * DEPTHFT, data = ASWELLS)
summary(lm_model)
```

```
##
## Call:
## lm(formula = ARSENIC ~ (LATITUDE + LONGITUDE) * DEPTHFT, data = ASWELLS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -175.75  -65.04  -23.02   29.82  480.01
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       10845.07   67720.06   0.160   0.8729
## LATITUDE          -1279.76    1053.11  -1.215   0.2252
```

```
## LONGITUDE           217.40     814.50   0.267   0.7897
## DEPTHFT           -1549.22     985.58  -1.572   0.1170
## LATITUDE:DEPTHFT    -11.00      11.86  -0.927   0.3547
## LONGITUDE:DEPTHFT    19.98      11.20   1.783   0.0755 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103.1 on 321 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1372, Adjusted R-squared:  0.1238
## F-statistic: 10.21 on 5 and 321 DF,  p-value: 4.306e-09
```

The least square equation for the above model can be written as
$ARSEINC = 10845.07 - 1279.76 * LATITUDE + 217.40 * LONGITUDE - 1549.22 * DEPTH - 11 * (LATITUDE : DEPTHFT) + 19.98 * (LONGITUDE : DEPTHFT)$

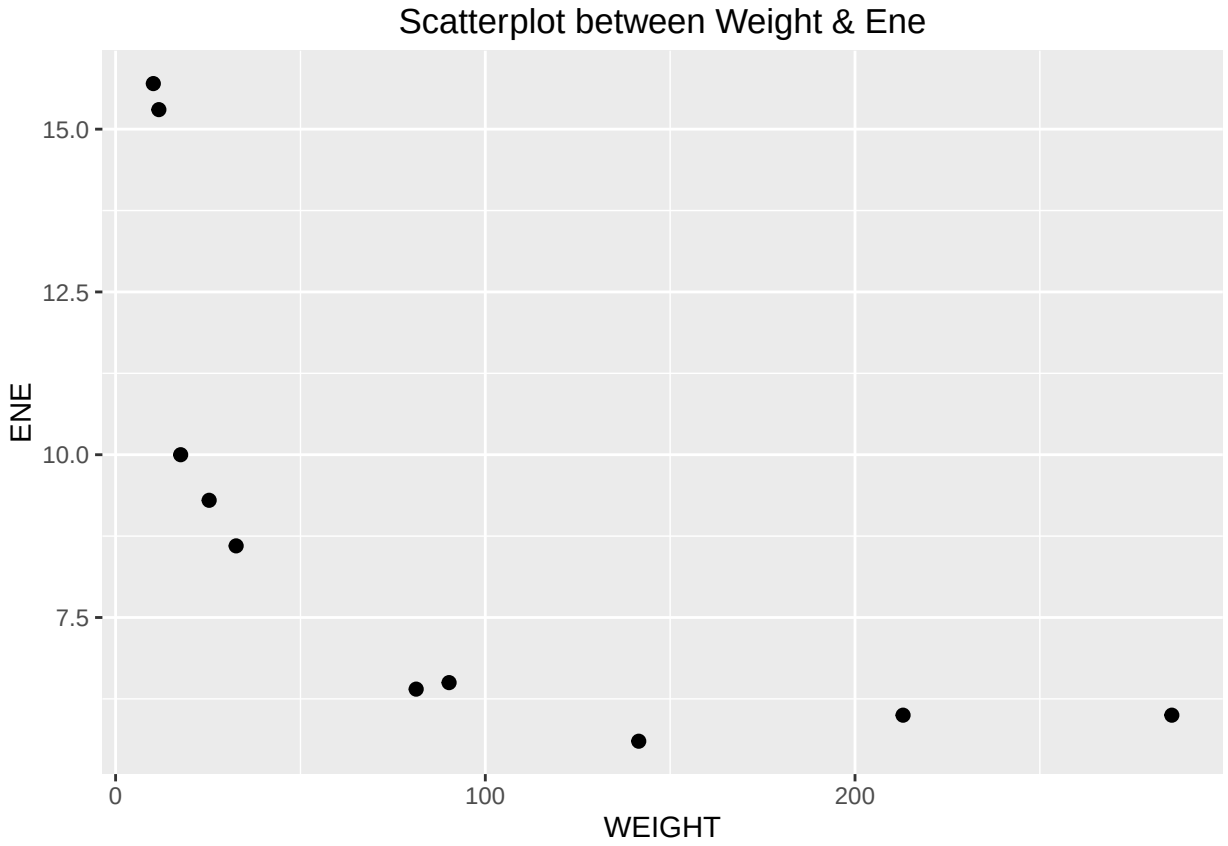
With $alpha = 0.05$ the intereaction term does not have an effect on the arsenic level. The p-value for the LATITUDE:DEPTHFT term is 0.35 with t = -0.93 and for LONGITUDE:DEPTHFT term the p-value is 0.0755 with t = 1.78.

Carp diet study.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```

```
load("rdata/CARP.Rdata")
ggplot(CARP, aes(x = WEIGHT, y = ENE)) + geom_point(size = 2) +
  ggtitle("Scatterplot between Weight & Ene")
```



Yes, there is a clear pattern in the above scatterplot. From the output, we can see that the p-value for the β_2 is 0.031. Our $H_o : \beta_2 = 0$ is rejected.

Homework assistance for accounting students.

```
library(stringr)
library(nnet)
load("rdata/ACCHW.Rdata")
ACCHW$ASSIST <- as.character(ACCHW$ASSIST)
ACCHW$ASSIST <- str_trim(ACCHW$ASSIST)
ACCHW <- cbind(ACCHW, data.frame(class.ind(ACCHW$ASSIST)))
lm_model <- lm(IMPROVE ~ FULL + CHECK, ACCHW)
summary(lm_model)
```

```
##
## Call:
## lm(formula = IMPROVE ~ FULL + CHECK, data = ACCHW)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.433 -2.433  0.050  1.567  6.567
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.4333     0.4941   4.925  5.2e-06 ***
## FULL         -0.4833     0.7813  -0.619    0.538
```

```
## CHECK         0.2867     0.7329   0.391    0.697
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.706 on 72 degrees of freedom
## Multiple R-squared:  0.01244,    Adjusted R-squared:  -0.01499
## F-statistic: 0.4535 on 2 and 72 DF,  p-value: 0.6372
```

We can use categorical features in a regression by converting into dummy variable. In the above case the class.ind function from the neural net package creates a dummy variable. The equation can be writtern as $Improve = 2.433 - 0.48 * Full + 0.28 * Check$. Note that the model fails the overall F-test (p-value = 0.6372).