# Homework 3

Solution

## Question 1. Reality TV and cosmetic surgery (Data set: BDYIMG))

(a) (10 pts) Fit the first-order model, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$, to the data in the file. Give the least squares prediction equation

```
source("anova_alt.R")
### Read the data
setwd("/cloud/project/STAT 3113 Data Sets")
bdyimg = read.csv("BDYIMG.csv", header = TRUE, sep = ",", dec = ".")
attach(bdyimg)

### Fit the MLR model
fit_bdyimg = lm(DESIRE ~ GENDER + SELFESTM + BODYSAT + IMPREAL)


anova_alt(fit_bdyimg)
```

```
## Analysis of Variance Table
##
##          Df      SS      MS      F           P
## Source    4  827.83 206.958 40.849 9.1886e-24
## Error   165  835.95   5.066
## Total   169 1663.79   9.845
```

```
summary(fit_bdyimg)
```

```
##
## Call:
## lm(formula = DESIRE ~ GENDER + SELFESTM + BODYSAT + IMPREAL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6628 -1.6688 -0.0767  1.6087  6.1345
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.01066    0.77534  18.070  < 2e-16 ***
## GENDER      -2.18649    0.67663  -3.231 0.001487 **
## SELFESTM    -0.04794    0.03669  -1.307 0.193157
## BODYSAT     -0.32233    0.14348  -2.247 0.025998 *
## IMPREAL      0.49310    0.12739   3.871 0.000156 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.251 on 165 degrees of freedom
## Multiple R-squared:  0.4976, Adjusted R-squared:  0.4854
```

```
## F-statistic: 40.85 on 4 and 165 DF,  p-value: < 2.2e-16
```

Answer: The least squares prediction equation is

$$\hat{y} = 14.011 - 2.186x_1 - 0.0479x_2 - 0.322x_3 + 0.493x_4$$

(b) (20 pts) Interpret the $\beta$ estimates in the words of the problem. (Yes, you need to find the values of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$ and interpret them one by one.)

Answer:

- $\hat{\beta}_0 = 14.01$. This has no meaning other than the y-intercept.
- $\hat{\beta}_1 = -2.186$. The mean value of desire to have cosmetic surgery is estimated to be 2.186 units lower for males than females, holding all other variables constant.
- $\hat{\beta}_2 = -0.0479$. For each unit increase in self-esteem, the mean value of desire to have cosmetic surgery is estimated to decrease by 0.0479 units, holding all other variables constant.
- $\hat{\beta}_3 = -0.322$. For each unit increase in body satisfaction, then mean value of desire to have cosmetic surgery is estimated to decrease by 0.322 units, holding all other variables constant.
- $\hat{\beta}_4 = 0.493$. For each unit increase in impression of reality TV, the mean value of desire to have cosmetic surgery is estimated to increase by 0.493 units, holding all other variables constant.

(c) (15 pts) Is the overall model statistically useful for predicting desire to have cosmetic surgery? Test using $\alpha = .01$. (When performing the hypothesis testing, do not forget to write down the hypotheses first!)

Answer:

To determine if the overall model is useful for predicting desire to have cosmetic surgery, we test:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_a : \text{At least one } \beta_i \neq 0$$

From the printout, the test statistic is $F = 40.85$ and the $p$-value is 0.000. Since the $p$-value is less than $\alpha$ $(0.000 < 0.01)$, $H_0$ is rejected. There is sufficient evidence to indicate the overall model is useful for predicting desire to have cosmetic surgery at $\alpha = 0.01$.

## Question 2. Arsenic in groundwater (Data set: ASWELLS)

```
### Import data and fit the MLR model
setwd("/cloud/project/STAT 3113 Data Sets")
aswells = read.csv("ASWELLS.csv", header = TRUE, sep = ",", dec = ".")

### There is a missing value in the DEPTH.FT column

aswells$DEPTH.FT = as.numeric(aswells$DEPTH.FT)
attach(aswells)

options(scipen = 1)

fit_aswells = lm(ARSENIC ~ LATITUDE + LONGITUDE + DEPTH.FT)
anova_alt(fit_aswells)
```

```
## Analysis of Variance Table
##
##          Df      SS     MS      F           P
## Source   3  505770 168590 15.799 1.3078e-09
```

```
## Error  323 3446791   10671
## Total  326 3952562   12124
```

```
summary(fit_aswells)
```

```
##
## Call:
## lm(formula = ARSENIC ~ LATITUDE + LONGITUDE + DEPTH.FT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -134.41  -65.51  -26.85   27.05  469.32
##
## Coefficients:
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -86867.9174  31224.2677  -2.782   0.00572 **
## LATITUDE     -2218.7568    526.8165  -4.212 0.0000329 ***
## LONGITUDE     1542.1627    373.0721   4.134 0.0000455 ***
## DEPTH.FT        -0.3496      0.1566  -2.232   0.02628 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103.3 on 323 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.128,  Adjusted R-squared:  0.1199
## F-statistic:  15.8 on 3 and 323 DF,  p-value: 1.308e-09
```

(a) (10 pts) Write a first-order model for arsenic level ($y$) as a function of latitude, longitude, and depth.

Answer:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

where $x_1$ is latitude, $x_2$ is longitude, and $x_3$ is depth.

(b) (10 pts) Fit the model to the data using the method of least squares.

Answer:

From the R output, the fitted regression equation is

$$\hat{y} = -86,868 - 2,219x_1 + 1,542x_2 - 0.350x_3.$$

(c) (10 pts) Find the value of $\hat{\beta}_2$ and give a practical interpretation of $\hat{\beta}_2$.

Answer:

- $\hat{\beta}_2 = 1,542$. We estimate that the mean arsenic level will increase by 1,542 $\mu$g/liter for each additional degree increase in longitude, holding all other variables constant.

(d) (10 pts) Find the model standard deviation, $s$, and interpret its value.

Answer:

$s = 103.3$. We could expect that most observed values of arsenic levels to fall within $2s = 2 * 103.3 = 206.6$ units of their predicted values.

(f) (15 pts) Conduct a test of overall model utility at $\alpha = .05$. (When performing the hypothesis testing, do not forget to write down the hypotheses first!)

Answer:

To determine if the overall model is adequate, we test:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_a : \text{At least one } \beta_i \neq 0$$

The test statistic is $F = 15.80$ and the $p$-value is 0.000. Since the $p$-value is less than $\alpha$ $(0.000 < 0.05)$, $H_0$ is rejected. There is sufficient evidence to indicate the overall model is adequate for predicting arsenic levels at $\alpha = 0.05$.