Homework 4

Solution

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

4.12(d) (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. Then identify R^2 and R_a^2 from the R output. Which statistic is the preferred measure of model fit? Practically interpret the value of this statistic.

```r
### Fit the MLR model and find the statistics
bdyimg = read.csv("STAT 3113 Data Sets/BDYIMG.csv")

fit_bdyimg = lm(DESIRE ~ GENDER + SELFESTM + BODYSAT + IMPREAL, data=bdyimg)
summary(fit_bdyimg)
```

```
##
## Call:
## lm(formula = DESIRE ~ GENDER + SELFESTM + BODYSAT + IMPREAL,
##     data = bdyimg)
##
## 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:

- $R^2 = 0.4976$
- $R_a^2 = 0.4854$
- Which statistic, R^2 or R_a^2 is preferred measure?

  R_a^2 is preferred measure of model fit.

- Practically interpret R_a^2.

  48.54% of the total sample variation in desire values is explained by the model containing gender, self-esteem, body satisfaction and impression of reality TV, adjusting for the sample size and the number of variables in the model.

4.12(e) (10 pts) Conduct a test to determine whether desire to have cosmetic surgery decreases linearly as level of body satisfaction increases. Use $\alpha = .05$.

Answer:

$H_0 : \beta_3 = 0$

$H_a : \beta_3 < 0$

From the R output, the p-value for the left-tailed test is 0.025998/2= 0.013 $< \alpha = .05$, we reject H_0. There is sufficient evidence to indicate the desire to have cosmetic surgery decreases linearly as level of body satisfaction increases, holding all other variables constant at $\alpha = 0.05$.

4.12(f) (10 pts) Find a 95% confidence interval for β_4. Practically interpret the result.

```
confint(fit_bdyimg, level=0.95)
```

```
##                  2.5 %      97.5 %
## (Intercept) 12.4797837 15.54153425
## GENDER      -3.5224532 -0.85051691
## SELFESTM    -0.1203848  0.02450268
## BODYSAT     -0.6056281 -0.03903592
## IMPREAL      0.2415720  0.74463423
```

Answer:

From the R printout, the 95% confidence interval for β_4 is (0.242, 0.745).

– Interpret the confidence interval for β_4.

We are 95% confident that the increase in mean desire for cosmetic surgery is between 0.242 and 0.745 for each unit increase in impression of reality TV, holding all other variables constant.

4.22(b) (10 pts) Find the confidence interval in R and interpret the confidence interval for E(y) for student 4.

```
new.data = data.frame(SELFESTM=22, BODYSAT=9, IMPREAL=4, GENDER=1)

CI = predict(fit_bdyimg, newdata=new.data, se.fit=TRUE,
             interval="confidence", level=.95)

CI$fit
```

```
##        fit      lwr      upr
## 1 9.840895 8.790836 10.89096
```

Answer:

The confidence interval is (8.79, 10.89). We are 95% confident that the mean desire to have cosmetic surgery is between 8.79 and 10.89 for males with a self-esteem of 22, body satisfaction of 9, and impression of reality TV of 4.

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

```
### Import data and fit the MLR model
aswells = read.csv("STAT 3113 Data Sets/ASWELLS.csv")

aswells$DEPTH.FT=as.numeric(aswells$DEPTH.FT)
options(scipen=1)
```

```r
fit_aswells = lm(ARSENIC ~ LATITUDE + LONGITUDE + DEPTH.FT,
                 data=aswells)

source("anova_alt.R")
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
```

```r
summary(fit_aswells)
```

```
##
## Call:
## lm(formula = ARSENIC ~ LATITUDE + LONGITUDE + DEPTH.FT, data = aswells)
##
## 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
```

(e) (10 pts) Interpret the values of R^2 and R_a^2.

Answer:

- $R^2 = 0.1280$. Interpretation: 12.8% of the total variation in the arsenic levels is explained by the regression model containing latitude, longitude, and depth.
- $R_a^2 = 0.1199$. Interpretation: 11.99% of the total variation in the arsenic levels is explained by the regression model containing latitude, longitude, and depth, adjusted for the number of independent variables in the model and the sample size.

(g) (5 pts) Based on the results you got in HW 3 and HW 4 about this question, would you recommend using model to predict arsenic level? Explain.

Answer: Based on the results we got, this model is questionable. Even though the arsenic level is significantly related to the predictors, the R^2 level is quite low. Only 12.8% of the variation in the arsenic level is explained by the model. In addition, the standard deviation is 103.3, which is quite large.

Question 3. Cooling Method for Gas Turbines (Data set: GASTURBINE)

4.15(a) (5 pts) Write a first-order model for heat rate (y) as a function of speed, inlet temperature, exhaust temperature, cycle pressure ratio, and air flow rate.

Answer:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon,$$

where x_1 is speed, x_2 is inlet temperature, x_3 is exhaust temperature, x_4 is cycle pressure ratio, and x_5 is air flow rate.

4.15(b) (5 pts) Fit the model to the data using the method of least squares.

```r
### Import data and fit the MLR model
gasturbine = read.csv("STAT 3113 Data Sets/GASTURBINE.csv")
names(gasturbine)
```

```
## [1] "ENGINE"     "SHAFTS"     "RPM"        "CPRATIO"    "INLET.TEMP"
## [6] "EXH.TEMP"   "AIRFLOW"    "POWER"      "HEATRATE"
```

```r
fit_gasturbine = lm(HEATRATE ~ RPM + INLET.TEMP + EXH.TEMP + CPRATIO + AIRFLOW, data=gasturbine)

anova_alt(fit_gasturbine)
```

```
## Analysis of Variance Table
##
##        Df        SS       MS     F          P
## Source  5 155055273 31011055 147.3 1.0671e-32
## Error  61  12841935   210524
## Total  66 167897208  2543897
```

```r
summary(fit_gasturbine)
```

```
##
## Call:
## lm(formula = HEATRATE ~ RPM + INLET.TEMP + EXH.TEMP + CPRATIO +
##     AIRFLOW, data = gasturbine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1007.0  -290.9  -105.8   240.8  1414.0
##
## Coefficients:
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept) 13614.46078   870.01294  15.649  < 2e-16 ***
## RPM             0.08879     0.01391   6.382 2.64e-08 ***
## INLET.TEMP     -9.20087     1.49920  -6.137 6.86e-08 ***
## EXH.TEMP       14.39385     3.46095   4.159 0.000102 ***
## CPRATIO         0.35190    29.55568   0.012 0.990539
## AIRFLOW        -0.84796     0.44211  -1.918 0.059800 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 458.8 on 61 degrees of freedom
## Multiple R-squared:  0.9235, Adjusted R-squared:  0.9172
## F-statistic: 147.3 on 5 and 61 DF,  p-value: < 2.2e-16
```

Answer:

$\hat{y} = 13614 + 0.0888x_1 - 9.2x_2 + 14.39x_3 + 0.4x_4 - 0.848x_5$

4.15(e) (10 pts) Find the adjusted-R^2 value and interpret it.

Answer: $R_a^2 = 0.9172$.

Interpretation: 91.72% of the total variation in heat rates is explained by the model containing the 5 independent variables, adjusted for the number of predictors in the model and the sample size.

4.15(f) (10 pts) Is the overall model statistically useful at predicting head rate (y)? Test using $\alpha = .01$.

Answer:

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_a : At least one $\beta_i \neq 0$

The p-value for the F-statistic is 0.000. Since the p-value is less than α, H_0 is rejected. There is sufficient evidence to indicate that the model is useful in predicting the heat rate at $\alpha = 0.01$.

4.24(a) (5 pts) Find and interpret the 95% prediction interval for y in the words of the problem.

```
new.data = data.frame(RPM=7500, INLET.TEMP=1000, EXH.TEMP=525,
                      CPRATIO=13.5, AIRFLOW=10)

PI = predict(fit_gasturbine, newdata=new.data, se.fit = TRUE, interval="prediction", level = 0.95)

PI$fit
```

```
##        fit      lwr      upr
## 1 12632.53 11599.56 13665.49
```

Answer:

With 95% confidence we can predict that the heat rate level is between 11599.6, 13665.5 for RPM=7500, INLET.TEMP=1000, EXH.TEMP=525, CPRATIO=13.5, and AIRFLOW=10.

4.24(b) (5 pts) Find and interpret the 95% confidence interval for $E(y)$ in the words of the problem.

```
CI = predict(fit_gasturbine, newdata=new.data, se.fit = TRUE, interval="confidence", level = 0.95)

CI$fit
```

```
##        fit      lwr      upr
## 1 12632.53 12157.93 13107.12
```

Answer:

With 95% confidence we can say that the mean heat rate level is between 12157.9 and 13107.11 for RPM=7500, INLET.TEMP=1000, EXH.TEMP=525, CPRATIO=13.5, and AIRFLOW=10.

4.24(c) (5 pts) Will the confidence interval for $E(y)$ always be narrower than the prediction interval for y? Explain.

Answer:

The confidence interval for $E(y)$ will always be narrower than the corresponding prediction interval for a single point.

The variance for a single point includes the variation for locating the mean plus the variation of the y once the mean has been located. The variance for $E(y)$ only includes the variation for locating the mean.