Hw05ST430Yu

Haozhe (Jerry) Yu

2023-10-27

Question 1

```
software <- as_tibble(read_table("Datasets/software.txt",
            col_names = TRUE,
            ))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Rep = col_double(),
##   Software = col_double(),
##   SalesLastQuarter = col_double(),
##   SalesThisQuarter = col_double()
## )
```

1A. Fit a model in which sales last quarter is ignored. We want to know whether software package has any effect on sales.

```
softwareslr <- lm(SalesThisQuarter ~ as.factor(Software), software)
```

i. Write E(y|x).

E(y|x) = 81.5833333 + -2software2 + -7.6666667software3.

ii. What proportion of the variation in sales this quarter is explained by software package?

The proportion of the variation in sales this quarter that can be explained w the software package is 0.0817601.

iii. What is the null hypothesis for testing whether software package has any effect on sales?

H0: The type of software package has no effect on sales this quarter. That is $\beta_{software2} = \beta_{software3} = 0$.

iv. Give the test statistic.

The test statistic is 1.4691611.

v. Give the p-value.

The p value is 0.2447802.

vi. Do you reject H0 at = 0.05?

No, 0.2447802 > 0.05.

vii. Are the results statistically significant at the 0.05 level?

No, the results are not significant at the 0.05 level because out p value is higher than 0.05. We fail to reject the null hypothesis and conclude there is not enough evidence to support the claim that the type of software package has an effect on the sales for this quarter, if we ignore the sales form last quarter.

1B. Fit a model with software package and sales last quarter as the explanatory variables, and sales this quarter as the response variable.

```
softwaremlr <-
  lm(SalesThisQuarter ~ as.factor(Software) + SalesLastQuarter,
     software)
```

i. Write E(y|x).

E(y|x) = -36.442295 + 0.7535141software2 + -1.2835203software3 + 1.5019168Sales Last Quarter.

ii. What is the null hypothesis for testing whether software package has any effect on sales this quarter once you control for sales last quarter?

H0 = $\beta_{software2} = \beta_{software3} = 0$.

iii. Give the test statistic.

```
softf <-
  ftest(softwaremlr, matrix(c(0, 1, 0, 0, 0, 0, 1, 0), nrow = 2, byrow =
                              TRUE))
```

the F statistic for the general linear test is 0.2422432.

iv. Give the p-value.

The p value for the general linear test is 0.7862915.

v. Do you reject H0 at = 0.05?

No, 0.7862915 > 0.05.

vi. Are the results statistically significant at the 0.05 level?

No, 0.7862915 > 0.05, so we fail to reject the null hypothesis and conclude that there is not evidence for the claim that the software package has an effect on sales this quarter once you control for sales last quarter.

vii. What proportion of the remaining variation in sales this quarter is explained by software package once you allow for sales last quarter?

```
softwaremlrr <-
  lm(SalesThisQuarter ~ SalesLastQuarter,
     software)
```

The proportion of the remaining variation in sales this quarter explained by the software package controlling for last quarter's sales is 0.0149144. (Calculated using `rsp.partial`)

1C. Fit a full model (with interaction) in which the slopes and intercepts of the regression lines relating sales last quarter to sales this quarter might depend on the kind of software the sales representatives are using.

```
softwaremlri <- lm(SalesThisQuarter~as.factor(Software)*SalesLastQuarter,software)
```

i. Write E(y/x).

E(y|x) = -92.2593254 + 144.6515746software2 + 48.1096684software3 + 2.2122077Sales Last Quarter + -1.8579265software2*Sales Last Quarter + -0.6238716software3*Sales Last Quarter

ii. What is the null hypothesis for testing whether the three slopes are equal?

H0 = $\beta_{software2*SalesLastQuarter} = \beta_{software3*SalesLastQuarter} = 0$.

iii. What is the null hypothesis for testing whether the effect of software program on sales this quarter depends on sales last quarter?

H0 = $\beta_{software2*SalesLastQuarter} = \beta_{software3*SalesLastQuarter} = 0$.

iv. Carry out an F-test to determine whether the effect of software type on sales depends on the representative's performance last quarter.

```
tImatrix <- matrix(c(0,0,0,0,1,0,
                     0,0,0,0,0,1),
                   nrow = 2,
                   byrow = TRUE
                   )
```

A. Give the test statistic.

The test statistic (F) calculated by `ftest()` is 10.3049569.

B. Give the p-value.

The p value (p) calculated by `ftest()` is 3.919989×10^{-4}.

C. Do you reject H0 at = 0.05?

As $3.919989 \times 10^{-4} < 0.05$, we reject H0.

D. Are the results statistically significant at the 0.05 level?

As we reject the null hypothesis, we can conclude that our results are statistically significant at $\alpha = 0.05$ and that there is evidence to support the claim that the effect of software type on sales depends on the representative's performance last quarter.

v. Estimate the slopes and intercepts of the three regression lines.

- E(SalesThisQuarter|Software = 1) = -92.2593254 + 2.2122077*SalesLastQuarter
- E(SalesThisQuarter|Software = 2) = 52.3922493 + 0.3542812*SalesLastQuarter
- E(SalesThisQuarter|Software = 2) = -44.1496569 + 1.5883362*SalesLastQuarter

vi. Test whether the slope is different from zero for software package two.

A. State the null hypothesis.

H0: $\beta_{software2*SalesLastQuarter} + \beta_{SalesLastQuarter} = 0$

B. Give the test statistic.

```
softs42 <- matrix(c(0,0,0,1,1,0),
                  nrow=1,
                  byrow = TRUE)
summary(softwaremlri)
```

```
##
## Call:
## lm(formula = SalesThisQuarter ~ as.factor(Software) * SalesLastQuarter,
##     data = software)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6222 -3.9426  0.8822  2.8198 11.7895
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           -92.2593    20.6075  -4.477 0.000102 ***
## as.factor(Software)2                  144.6516    31.9049   4.534 8.66e-05 ***
## as.factor(Software)3                   48.1097    29.6808   1.621 0.115504
## SalesLastQuarter                        2.2122     0.2614   8.462 1.92e-09 ***
## as.factor(Software)2:SalesLastQuarter  -1.8579     0.4106  -4.525 8.88e-05 ***
## as.factor(Software)3:SalesLastQuarter  -0.6239     0.3879  -1.609 0.118200
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.649 on 30 degrees of freedom
## Multiple R-squared:  0.7938, Adjusted R-squared:  0.7594
## F-statistic:  23.1 on 5 and 30 DF,  p-value: 1.849e-09
```

The F value is 1.2518795.

C. Give the p-value.

The p value is 0.2720729.

D. Do you reject H0 at = 0.05?

As 0.2720729 > 0.05, we fail to reject H0.

E. Are the results statistically significant at the 0.05 level?

As we fail to reject the null hypothesis, we conclude that the results are not statistically significant and there is not evidence to support the claim that the slope is different from zero for software package two.

1D. Test the hypothesis that you would test in order to answer this question: Controlling for sales last quarter, is average expected sales this quarter for software 1 and 3 different from expected sales this quarter for package 2?

H0: $\frac{\beta_{\text{software2*Sales Last Quarter}}}{2} + \frac{\beta_{\text{software2}}}{2} = \beta_{\text{software2*Sales Last Quarter}} + \beta_{\text{software2}}$

```
btest <- matrix(c(0,1,-1/2,0,1,-1/2),
                  nrow=1,
                  byrow = TRUE)
ftest(softwaremlri,btest)
```

```
##            F          df1          df2      p-value
## 17.883932568  1.000000000 30.000000000  0.000202823
```

As $\alpha < 0.05$, we reject the null hypothesis and conclude that there is evidence suggesting that controlling for sales last quarter, average expected sales this quarter for software 1 and 3 is different from expected sales this quarter for package 2.

Question 2

2. Pigs are routinely given large doses of antibiotics even when they show no signs of illness, to protect their health under unsanitary conditions. Pigs were randomly assigned to one of three antibiotic drugs. Dressed weight (weight of the pig after slaughter and removal of head, intestines and skin) was the dependent variable. Independent variables are Drug type, Mother's live adult weight and Father's live adult weight.

```
library(readr)
pigs <- as_tibble(read_table("Datasets/pig.txt",
                             col_names = TRUE))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Drug = col_double(),
##   Momweight = col_double(),
##   Dadweight = col_double(),
##   Pigweight = col_double()
## )
```

```
pigs <- add_column(pigs,
                   c2 = ifelse(pigs$Drug == 2, 1, 0),
                   c3 = ifelse(pigs$Drug == 3, 1, 0))
```

a. Write the regression equation for the full model, including error term.

```
pigsm <- lm(Pigweight~Momweight + Dadweight + c2 +c3,pigs)
```

The regression equation is

E(Pigweight|x) = 7.4816313 + 0.2636323*Momweight + 0.1744219*Dadweight + -1.6055653*Drug2+ -0.7048*Drug3 + ϵ

b. Make a table with one row for every drug, with columns showing how the dummy variables were defined. Make another column giving E(y/x) for each drug controlling other predictor variables.

```
ptable <- pigs %>% select(Drug, c2, c3) %>% distinct() %>%

  add_column("E(Y|x)" = c(
    paste0(
      round(summary(pigsm)$coefficients[1, 1], 3),
      " + ",
      round(summary(pigsm)$coefficients[2, 1], 3),
      "*Momweight + ",
      round(summary(pigsm)$coefficients[3, 1], 3),
      "*Dadweight"
    ),
    paste0(
      round(
        summary(pigsm)$coefficients[1, 1] + summary(pigsm)$coefficients[4, 1],
        3
      ),
      " + ",
      round(summary(pigsm)$coefficients[2, 1], 3),
      "*Momweight + ",
      round(summary(pigsm)$coefficients[3, 1], 3),
      "*Dadweight"
```

```
    ),
    paste0(
      round(
        summary(pigsm)$coefficients[1, 1] + summary(pigsm)$coefficients[5, 1],
        3
      ),
      " + ",
      round(summary(pigsm)$coefficients[2, 1], 3),
      "*Momweight + ",
      round(summary(pigsm)$coefficients[3, 1], 3),
      "*Dadweight"
    )
  ))

ptable
```

```
## # A tibble: 3 x 4
##    Drug    c2    c3 `E(Y|x)`
##   <dbl> <dbl> <dbl> <chr>
## 1     1     0     0 7.482 + 0.264*Momweight + 0.174*Dadweight
## 2     2     1     0 5.876 + 0.264*Momweight + 0.174*Dadweight
## 3     3     0     1 6.777 + 0.264*Momweight + 0.174*Dadweight
```

c. Predict the dressed weight of a pig getting Drug 2, whose mother weighed 140 pounds, and whose father weighed 185 pounds.

```
pigpred <- tibble(
  c2=1,
  c3=0,
  Momweight = 140,
  Dadweight = 185
)

predict(pigsm,pigpred,interval = "prediction",level=0.95)
```

```
##        fit      lwr      upr
## 1 75.05263 71.15725 78.94801
```

```
predict(pigsm,pigpred,interval = "prediction",level=0.95)[1]
```

```
## [1] 75.05263
```

The dressed weight of a pig getting Drug 2, whose mother weighed 140 pounds, and whose father weight 185 pounds is 75.0526279, with a prediction interval of (71.157246,78.9480098)

d. This parallel plane regression model (no interaction model) specifies that the differences in expected weight for the different drug treatments are the same for every possible combination of mother's weight and father's weight. Give a 95% confidence interval for the difference in expected weight between drug treatments 1 and 2. Show your calculations.

Null hypothesis: $\beta_{drug2} = 0$.

So we can use the t value directly from the `summary()` output. Thus, using `confint()` with t = -3.0415545 and df = 70, we derive a 95% confidence interval of (-2.6583819, -0.5527487)

e. In symbols, give the null and alternate hypotheses you would test to answer the following questions. Your answers are statements involving the /*beta* values from your regression equation. Give the value of the t or F statistic (a number from the printout), and indicate whether or not you reject the null hypothesis.

i. Controlling for mother's weight and father's weight, does type of drug have an effect on the expected weight of a pig?

```
pigshi <- matrix(c(0,0,0,1,0,
                   0,0,0,0,1),
                   nrow=2,
                   byrow = TRUE)
ftest(pigsm,pigshi)
```

```
##           F         df1         df2     p-value
##  4.63559465  2.00000000 70.00000000  0.01286451
```

H0: $\beta_{drug2} = \beta_{drug3} = 0$ HA: At least β_{drug2} or $\beta_{drug3} \neq 0$.

F = 4.6355947

As p< 0.05, we reject the null hypothesis and conclude that there is enough evidence to support the claim that type of drug does have an effect on the expected weight of a pig, controlling for mother's weight and father's weight, at α -= 0.05.

ii. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 1 or Drug 2?

```
summary(pigsm)$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  7.4816313 9.14916679  0.8177391 4.162810e-01
## Momweight    0.2636323 0.04726554  5.5776841 4.282202e-07
## Dadweight    0.1744219 0.03464907  5.0339547 3.580316e-06
## c2          -1.6055653 0.52787655 -3.0415545 3.311102e-03
## c3          -0.7048000 0.52871229 -1.3330501 1.868378e-01
```

H0: $\beta_{drug2} = 0$ HA: $\beta_{drug2} = 0 \neq 0$.

t = -3.0415545

As p< 0.05, we reject the null hypothesis and conclude that there is enough evidence to support the claim that Drug 1 helps the average pig gain more weight than Drug 2, controlling for mother's weight and father's weight, at α -= 0.05.

iii. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 1 or Drug 3?

```
summary(pigsm)$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  7.4816313 9.14916679  0.8177391 4.162810e-01
## Momweight    0.2636323 0.04726554  5.5776841 4.282202e-07
## Dadweight    0.1744219 0.03464907  5.0339547 3.580316e-06
## c2          -1.6055653 0.52787655 -3.0415545 3.311102e-03
## c3          -0.7048000 0.52871229 -1.3330501 1.868378e-01
```

H0: $\beta_{drug3} = 0$ HA: $\beta_{drug3} = 0 \neq 0$.

t = -1.3330501

As p> 0.05, we fail to reject the null hypothesis and conclude that there is not enough evidence to support the claim that either Drug 1 or Drug 3 helps the average pig gain more weight, controlling for mother's weight and father's weight, at α -= 0.05.

iv. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 2 or Drug 3?

```
pigshi <- matrix(c(0,0,0,1,-1),
                 nrow=1,
                 byrow = TRUE)
ftest(pigsm,pigshi)
```

```
##           F         df1         df2     p-value
##  2.79120433  1.00000000 70.00000000  0.09924863
```

```
ftest(pigsm,pigshi)[1]
```

```
##        F
## 2.791204
```

```
summary(pigsm)
```

```
##
## Call:
## lm(formula = Pigweight ~ Momweight + Dadweight + c2 + c3, data = pigs)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.905 -1.174  0.187  1.351  3.657
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.48163    9.14917   0.818  0.41628
## Momweight    0.26363    0.04727   5.578 4.28e-07 ***
## Dadweight    0.17442    0.03465   5.034 3.58e-06 ***
## c2          -1.60557    0.52788  -3.042  0.00331 **
## c3          -0.70480    0.52871  -1.333  0.18684
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.855 on 70 degrees of freedom
## Multiple R-squared:  0.4561, Adjusted R-squared:  0.425
## F-statistic: 14.67 on 4 and 70 DF,  p-value: 9.393e-09
```

H0: $\beta_{drug2} = \beta_{drug3}$ HA: $\beta_{drug2} \neq \beta_{drug3}$.

F = 2.7912043

As p> 0.05, we fail to reject the null hypothesis and conclude that there is not enough evidence to support the claim that either Drug 2 or Drug 3 helps the average pig gain more weight, controlling for mother's weight and father's weight, at α -= 0.05.