Huiwen Chen unique name:huiwenc ID: 02156341 Department of Statistics

**STAT 500 HW3**

1. Using the sat data (see help(sat) for the description of variables) from ”faraway” package:

1. Fit a model with total sat score as the response and takers, ratio and salary as predictors. Comment on the coefficients and the good- ness of fit. Test the hypothesis βsalary = 0. Test the hypothesis βtakers = βratio = βsalary = 0.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1057.8982 44.3287 23.865 <2e-16 \*\*\*

takers -2.9134 0.2282 -12.764 <2e-16 \*\*\*

ratio -4.6394 2.1215 -2.187 0.0339 \*

salary 2.5525 1.0045 2.541 0.0145 \*

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.41 on 46 degrees of freedom

Multiple R-squared: 0.8239,Adjusted R-squared: 0.8124

F-statistic: 71.72 on 3 and 46 DF, p-value: < 2.2e-16

**Comments:** 1) As shown in the table above, all of the Pr(>|t|) of coefficients are < 0.05, which means the linear relationship with total sat score and takers, ratio, salary respectively is significant at 95% level. The predictor “takers” has the most significant relationship with total sat score, since the Pr(>|t|) is much less than 0.05.

1. The Multiple R-squared 0.8239 and adjusted R-squared 0.8124 are close to 1, that means more than 80% of the variation in the response is explained by the predictors. Basically we can say the model indicates good fit.

**i) Test the hypothesis βsalary = 0:**

i) H0: βsalary = 0; HA: not H0

Analysis of Variance Table

Model 1: total ~ takers + ratio

Model 2: total ~ takers + ratio + salary

Res.Df RSS Df Sum of Sq F Pr(>F)

1 47 55097

2 46 48315 1 6781.6 6.4566 0.01449

**Comments:** Since the Pr(>F) =0.01449 <0.05, then we have 95% confidence to reject H0. So βsalary shall not be excluded from the linear regression model.

**ii) Test the hypothesis βtakers = βratio = βsalary = 0:**

H0: βtakers = βratio = βsalary = 0; HA: not H0

Residual standard error: 32.41 on 46 degrees of freedom

Multiple R-squared: 0.8239, Adjusted R-squared: 0.8124

F-statistic: 71.72 on 3 and 46 DF, p-value: < 2.2e-16

**Comments:** From the summary of the model1, the p-value< 2.2e-16 is so small, then the null hypothesis is rejected.

2.Compute 95% and 99% CIs for the parameter associated with salary. Using just these intervals, what can we deduce about the p-value for salary in the regression summary?

**Solution:**

2.5 % 97.5 %

(Intercept) 968.6691802 1147.1271438

takers -3.3727807 -2.4539197

ratio -8.9098147 -0.3690414

salary 0.5304797 4.5744605

0.5 % 99.5 %

(Intercept) 938.786432 1177.009892

takers -3.526644 -2.300057

ratio -10.339965 1.061109

salary -0.146684 5.251624

The 95% CI for βsalary is [0.5304797, 4.5744605], and the 99% CI for βsalary is [-0.146684, 5.251624]. Since in null hypothesis βsalary =0 is within 99%CI but out of 95% CI, the null hypothesis H0 would be rejected with a confidence between 1% and 5%. We can deduce that p-value will be larger than 1%, but smaller than 5%.

3. Compute and display a 95% joint confidence region for the parameters associated with ratio and salary. Add the origin to the plot. The location of the origin on the plot tells us the outcome of a certain hypothesis test. State that test and its outcome.

H0: βsalary=βratio =0 ; Ha: not H0



Since the origin lies outside the ellipse, at the level of 95% we reject the null hypothesis H0: βsalary=βratio =0。

1. Now add expend (current expenditure per pupil) to the model and comment on the coefficients,their significance and the goodness of fit as compared to the model in question 1.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1045.9715 52.8698 19.784 < 2e-16 \*\*\*

takers -2.9045 0.2313 -12.559 2.61e-16 \*\*\*

ratio -3.6242 3.2154 -1.127 0.266

salary 1.6379 2.3872 0.686 0.496

expend 4.4626 10.5465 0.423 0.674

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.7 on 45 degrees of freedom

Multiple R-squared: 0.8246, Adjusted R-squared: 0.809

F-statistic: 52.88 on 4 and 45 DF, p-value: < 2.2e-16

**Comments:1)** Compared with coefficients in model 1, Pr(>|t|) of coefficients in model 2 become larger. In model 2, estimate of βsalary, βratio , βexpend are >0.05, whereas in model 1, βsalary, βratio are <0.05. So these coefficients become less interesting, and can be excluded from the whole model. **2)** P-value of model 2 is still small enough, which means the relationship of all predictors and the response is significant.  **3)** The adjusted R-squared of model 2 is slightly less than that of model1, but it is still close to 1, so the fitness is fairly good.

1. In the model of question 4, test the hypothesis βsalary = βexpend = βratio = 0. Based on your entire analysis, do you feel any of these predictors have an effect on the response?

H0: βsalary = βexpend = βratio = 0; Ha= no H0

Analysis of Variance Table

Model 1: total ~ takers

Model 2: total ~ takers + ratio + salary + expend

Res.Df RSS Df Sum of Sq F Pr(>F)

1 48 58433

2 45 48124 3 10309 3.2133 0.03165 \*

1) Since Pr(>F)= 0.03165 <0.05, so H0 can be rejected at 95% confidence level. Then βsalary, βexpend, βratio should not be excluded from the model at the same time.

2) The parameter associated with takers has a significant effect on the response, since the p-value of βtakers is extremely small.

6. Based on Chapter 3, problem 5 (p. 50). Find a formula relating R2 and the F-test (statistic) for the regression.

Solution:

In a regression setting, F is denoted as:

Where TSS== total sum of squares, RSS==residual sum of squares, n is the number of observations and p+1 is the number of predictors, including the constant.

Recall that, =

It is easy to find the formula that F =

R codes:

##Load the library

library (faraway)

##read in and check out the data

data (sat)

attach (sat)

sat

##problem 1

model1 <- lm(total~ takers+ ratio+ salary, sat)

summary(model1)

##model h0\_1 under H0:βsalary = 0

h0\_1 <- lm(total~ takers+ ratio, sat)

summary(h0)

##model1 is under H0 U HA

anova(h0\_1, model1)

##test H0:βtakers = βratio = βsalary = 0

##anova(model1)

summary(model1)

##problem 2

conf <- confint(model1,level = 0.95)

conf

conf <- confint(model1,level = 0.99)

conf

##problem 3

library(ellipse)

plot(ellipse(model1, c(3,4)), type = "l", xlim=c(-10, 1))

points(0,0)

points(coef(model1)[3], coef(model1)[4], pch=18)

abline (v=confint(model1, level = 0.95)[3,], lty=2)

abline (h=confint(model1, level = 0.95)[4,], lty=2)

##problem 4

model2 <- lm(total~ takers+ ratio+ salary+ expend, sat)

summary(model2)

##problem 5

h0\_2 <- lm(total~ takers, sat)

anova(h0\_2, model2)