Homework #4

Yunting Chiu

2021-02-19

1. **(2.17)** An analyst fitted normal error regression model and conducted an F test of $H0 : \beta 1 = 0$ versus $H1 : \beta 1 = 0$. The P-value of the test was 0.033, and the analyst concluded that $\beta 1 \; 6= 0$. Was the α level used by the analyst greater than or smaller than 0.033? If the α level had been 0.01, what would have been the appropriate conclusion?

The hypothesis is $H0 : \beta 1 = 0$ v.s. $H1 : \beta 1 = 0$. With the small p-value 0.033, we have evidence to reject the null, meaning that $\alpha > 0.033$. If the α level had been 0.01 ???

2. **(2.18)** For conducting statistical tests concerning the parameter $\beta 1$, why is the t-test more versatile than the F-test?

Because t-test have one-sided test(left tail & right tail), and two-sided test for $\beta 1$. Conversely, F-test (most notably in ANOVA) can only detect $H0 : \beta 1 = 0$ v.s. $H1 : \beta 1 \; != 0$.

3. **(2.19)** When testing $H0 : \beta 1 = 0$ versus $H1 : \beta 1 \; 6= 0$, why is the F-test a one-sided test even though H1 includes both cases $\beta 1 < 0$ and $\beta 1 > 0$?

4. **(Continued from HW-2,3)** At a gas station, 180 drivers were asked to record the mileage of their cars and the number of miles per gallon. The results are summarized in the table.

5. Computer project (2.23, 2.67).
Grade point average (this data set was already used in Homework-2,3).

```
# read the data
gpa <- read.table("./data/CH01PR19.txt")

reg <- lm(V1 ~ V2, data = gpa)
# call the regression model summary table
summary(reg)
```

```
##
## Call:
## lm(formula = V1 ~ V2, data = gpa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74004 -0.33827  0.04062  0.44064  1.22737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
## V2           0.03883    0.01277   3.040  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6231 on 118 degrees of freedom
## Multiple R-squared:  0.07262,    Adjusted R-squared:  0.06476
```

```
## F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

(a) Set up the ANOVA table. Use it to answer questions (b-e).

```r
anova(reg)
```

```
## Analysis of Variance Table
##
## Response: V1
##            Df Sum Sq Mean Sq F value   Pr(>F)
## V2          1  3.588  3.5878  9.2402 0.002917 **
## Residuals 118 45.818  0.3883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b) (Stat-615 only) What is estimated by MSR in your ANOVA table? by MSE? Under what conditions do MSR and MSE estimate the same quantity?

(c) Conduct an F-test of whether or not $\beta1 = 0$. Control the α level at 0.01. State the alternative and your conclusion

The F-test is 9.2402, and the p-value falls into significant level between 0.001 to 0.01. We can conclude the null hypothesis can be rejected at level 0.01 in favor of the alternative hypothesis.

(d) How much does the variation of Y reduce when X is introduced into the regression model? What is the relative reduction?

SST = 49.406, SSE = 45.818, SSR = 3.588. The coefficient of determination is 7 %. It means that 7 % of total variation of GPA score is explained by the ACT score.

6d:

$SST = SSE + SSR$

$49.406 = 45.818 + 3.588$

$9.24 = F = \dfrac{\frac{3.588}{1} \; SSR}{\frac{45.818}{118} \; SSE} = \dfrac{3.588 \; MSR}{0.39 \; MSE} \doteqdot 9.24$ (recheck F-stat)

relative reduction:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{3.588}{49.406} = 0.07 = 7\%$$

(e) Obtain the sample correlation coefficient and attach the appropriate sign to it, positive or negative.

Firstly, $\beta1$ is 0.03883, which is positive slope so the correlation coefficient is a positive number. Thus, the sample correlation coefficient is 0.26.

be:

R = sample correlation coefficient.

$R^2 = 0.07 \Rightarrow R = \sqrt{0.07} = 0.26$

(d) Coefficient of Determination $\Rightarrow$ is the square of the correlation coeff.

N4 $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ $\quad -1 \le R \le 1$

The coefficient of determination is interpreted as the proportion of observed variation in Y that can be explained by the simple linear regression model.

(f) (leftover from the last homework) On the same graph, plot

• the data • the least squares regression line for ACT scores • the 95 percent confidence band for the true regression line for ACT scores between 20 and 30.

```
attach(gpa)
n = length(V2) #sample sizes
e = reg$residuals # residuals
s = sqrt(sum(e^2)/(n-2)) # estimated standard deviation = root MSE
s
```

```
## [1] 0.623125
```

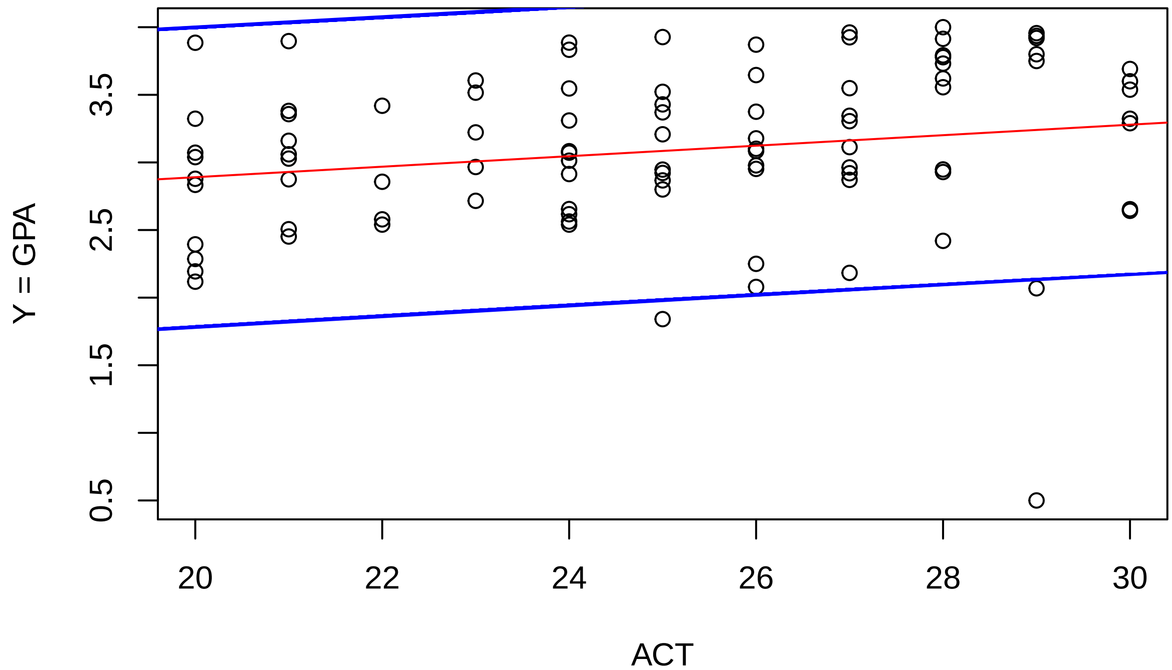
```
W = sqrt(qf(0.95,2,n-2))  # quantity of F-distribution
W
```

```
## [1] 1.753023
```

```
Yhat = fitted.values(reg) # Yhat = b0 + b1x = predict(reg)
Sxx = (n-1)*var(V2)

margin = W*s*sqrt(1/n + (V2 - mean(V2))^2/Sxx)
upper.band = Yhat + W*s*sqrt(1 + 1/n + (V2 - mean(V2))^2/Sxx) # 95% upper
lower.band = Yhat - W*s*sqrt(1 + 1/n + (V2 - mean(V2))^2/Sxx) # 95% lower

plot(V2, V1, xlab = "ACT", ylab = "Y = GPA", xlim = c(20,30))
abline(reg,col="red")
lines(V2 ,upper.band,col="blue")
lines(V2 ,lower.band,col="blue")
```


Y = GPA
0.5
1.5
2.5
3.5
20
22
24
26
28
30
ACT