Stat 615/415 (Regression) HW #4 – Solutions

Analysis of Variance (sec. 2.6-2.9).

1. (**2.17**) An analyst fitted normal error regression model and conducted an F test of $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. The P-value of the test was 0.033, and the analyst concluded that $\beta_1 \neq 0$. 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?

   SOLUTION. *With the P-value of 0.033, H_0 is rejected at any level of significance $\alpha > 0.033$. In particular, for $\alpha = 0.01$, H_0 is not rejected, and the slope is found not significant.*

2. (**2.18**) For conducting statistical tests concerning the parameter β_1, why is the t-test more versatile than the F-test?

   SOLUTION. *The t-test has a one-sided option while the F-test can only distinguish $\beta_1 = 0$ from $\beta_1 \neq 0$.*

   In other words, the t-test can be used for

$$H_0 : \beta = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$
$$H_0 : \beta = 0 \quad \text{vs.} \quad H_1 : \beta_1 < 0$$
$$H_0 : \beta = 0 \quad \text{vs.} \quad H_1 : \beta_1 > 0$$

   whereas the F-test can only be used for

$$H_0 : \beta = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

3. (**2.19**) When testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$, why is the F-test a one-sided test even though H_1 includes both cases $\beta_1 < 0$ and $\beta_1 > 0$?

   SOLUTION. *Even in a two-sided test of $H_1 : \beta_1 \neq 0$, the null hypothesis is rejected when the F-statistic is large. Indeed, $F = MSReg/MSErr$ is large when $MSReg$ is large which suggests that $\mathbf{E}(MSReg) = \sigma^2 + b_1^2 S_{xx}$ is large comparing with $\mathbf{E}(MSErr) = \sigma^2$, leading to $\beta_1 \neq 0$.*

4. (Continued from HW-2,3) The time it takes to transmit a file always depends on the file size. Suppose you transmitted 30 files, with the average size of 126 Kbytes and the standard deviation of 35 Kbytes. The average transmittance time was 0.04 seconds with the standard deviation of 0.01 seconds. The correlation coefficient between the time and the size was 0.86.

   In the previous homework, we fit a regression model that predicted the time it will take to transmit a 400 Kbyte file.

   (a) Compute the total, regression, and error sums of squares.

   (b) Complete the ANOVA table. Include degrees of freedom, mean squares, and the F-statistic.

   (c) Use this F-statistic to test significance of our regression model that relates transmission time to the size of the file. State H_0 and H_1, calculate the p-value, and make a conclusion. How does your test statistic and the p-value relate to your results in problem #3c of Homework-3?

   (d) Calculate R^2 and explain what it means.

   SOLUTION. *We are given: $n = 30$, $s_x = 35$, $s_y = 0.01$, and $r = 0.86$.*

(a)

$$\begin{aligned} SSTot &= (n-1)s_y^2 = (29)(0.01^2) = \quad 0.0029 \\ SSReg &= r^2 SSTot = (0.86)^2(0.0029) = \quad 0.00214 \\ SSErr &= SSTot - SSReg = 0.0029 - 0.00214 = \quad 0.00076 \end{aligned}$$

(b)

Sums of Squares	*DF*	*Mean Squares*	*F*
$SSReg = 0.00214$	1	$MSReg = SSReg/1 = 0.00214$	$F = MSReg/MSErr = 79.3$
$SSErr = 0.00076$	$n-2 = 28$	$MSErr = SSErr/28 = 0.000027$	
$SSTot = 0.0029$	$n-1 = 29$		

(c) *Test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.*
We obtained $F = 79.3$.
The p-value is $P = \boldsymbol{P}\{F > 79.3\} < 0.0001$, from F-distribution with 1 and 28 d.f.
Conclusion: reject H_0. The slope is significant. There is evidence of a linear relation between X and Y, the file size and the transmission time.

Comparing with the previous homework, the p-value is also very close to 0, and we have $F = t^2$ because $79.3 = 8.91^2$.

(d) $R^2 = \dfrac{SSReg}{SSTot} = \dfrac{0.00214}{0.0029} =$ *0.738 or 73.8%. It means that 73.8% of the total variation of transmission times is explained solely by the file sizes.*

5. (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.

	Sample mean	Standard deviation
Mileage	24,598	14,634
Miles per gallon	23.8	3.4

The sample correlation coefficient is $r = -0.17$. In the previous homework, we fit a regression model that described how the number of miles per gallon depends on the mileage.

(a) Complete the ANOVA table. Include sums of squares, degrees of freedom, mean squares, and the F-statistic.

(b) What statement can be tested using this F-statistic? Calculate the p-value and state a conclusion for this ANOVA F-test.

(c) From your ANOVA table, estimate the standard deviation of responses, $\sigma = \text{Std}(Y_i)$.

(d) Calculate R^2 and comment on the goodness of fit in this regression problem.

SOLUTION. *In this problem, $n = 180$, $s_x = 14,634$, $s_y = 3.4$, and $r = -0.17$.*

(a)
$$\begin{aligned} SSTot &= (n-1)s_y^2 = (179)(3.4^2) = \quad 2069.2 \\ SSReg &= r^2 SSTot = (-0.17)^2(2069.2) = \quad 59.8 \\ SSErr &= SSTot - SSReg = 2069.2 - 59.8 = \quad 2009.4 \end{aligned}$$
The ANOVA table is below.

Sums of Squares	*DF*	*Mean Squares*	*F*
$SSReg = 59.8$	1	$MSReg = SSReg/1 = 59.8$	$F = MSReg/MSErr = 5.29$
$SSErr = 2009.4$	$n-2 = 178$	$MSErr = SSErr/178 = 11.3$	
$SSTot = 2069.2$	$n-1 = 179$		

(b) *This ANOVA F-statistic is for testing* $H_0 : \beta_1 = 0$ *versus* $H_1 : \beta_1 \neq 0$.
We obtained $F = 5.29$.
The p-value is $P = \boldsymbol{P}\{F > 5.29\} = 0.023$, *from F-distribution with 1 and 178 d.f.*
This is a borderline case when the P-value is between 0.01 and 0.1. Thus, there is no obvious conclusion. For all α *levels above 0.023, including* $\alpha = 0.05$, H_0 *is rejected in favor of* H_1, *and the slope is found significant. Notice that* R^2 *is only 2.89% here, but nevertheless, the slope is (marginally) significant.*

(c) *The standard deviation of responses* $\sigma = \text{Std}(Y_i)$ *is estimated by* $s = \sqrt{MSErr} = \sqrt{11.3} = 3.36$ *(and that's the value we used in the previous h/w).*

(d) $R^2 = r^2 = (-0.17)^2 = 0.0289$ *or 2.89%. It means that only 2.89% of the total variation of responses (miles per gallon) is explained by X, the mileage on a car. This is not a good fit. Either there is a lot of variation, or pure noise, in our responses, or more likely, there are other factors which can explain more of the total variation.*

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

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

(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 $\beta_1 = 0$. Control the α level at 0.01. State the alternative and your conclusion.

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

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

(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.

Any problem reading the data? Since the textbook data are also available on the web, you can read them by one command,

```
A = read.table(url("http://statweb.lsu.edu/EXSTWeb/StatLab/DataSets/NKNWData/CH01PR19.txt"))
```

and then you can do `attach(A); names(A); head(A);`

etc.

SOLUTION. *Results are based on the R output below.*

(a) *The ANOVA table is*

Sums of Squares	*DF*	*Mean Squares*	*F*
$SSReg = 3.588$	1	$MSReg = 3.588$	9.2402
$SSErr = 45.818$	118	$MSErr = 0.388$	
$SSTot = 45.818 + 3.588 = 49.406$	119		

(b) *As we derived in class, MSR is an unbiased estimator of* $\sigma^2 + \beta_1^2 S_{xx}$, *and MSE*$= s^2$ *is an unbiased estimator of* σ^2. *That is,*

$$\mathbf{E}(MSReg) = \sigma^2 + \beta_1^2 S_{xx} \quad \text{and} \quad \mathbf{E}(MSErr) = \sigma^2.$$

They estimate the same quantity if $\beta_1 = 0$, which means the population slope is 0, X and Y are uncorrelated, and the linear regression of Y on X cannot be used to predict Y.

(c) *Test $H_0 : \beta_1 = 0$ vs $H_1 \beta_1 \neq 0$. The F-statistic is $F = 9.2402$, and the P-value is 0.002917. Since $P < 0.01$, the null hypothesis is rejected at the level $\alpha = 0.01$.*

The data present significant evidence that the regression slope is different from zero.

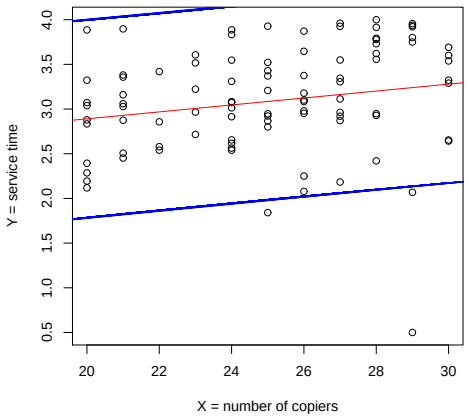
(d) *Variation of Y is reduced by $SSTot - SSErr = SSReg = 3.588$.*

The relative reduction is $SSReg/SSTot = R^2 =$ 0.07262 or 7.262%.

(e) *$r = \sqrt{R^2} =$ 0.269 . The sign is positive because the slope $b_1 = 0.03883$ is positive.*

Alternatively, you can compute the correlation coefficient with the R command `cor(X,Y)`.

(f)



R code

```
 # Reading data from the text file, renaming variables, fitting regression
> GPA = read.table("C:\\Data\\Book data\\Chapter  1 Data Sets\\CH01PR19.txt")
> attach(GPA); X = V2; Y = V1;
> reg = lm(Y~X)

> anova(reg)
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq F value   Pr(>F)
X           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

> summary(reg)

Call: lm(formula = Y ~ X)

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 ***
X            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

    # Confidence band for the entire population regression line
> n = length(X)
> W = sqrt(qf(0.95,2,n-2))
> Sxx = (n-1)*var(X)
> e = reg$residuals
> s = sqrt( sum(e^2)/(n-2) )
> margin = W*s*sqrt(1 + 1/n + (X - mean(X))^2/Sxx)
> upper.band = predict(reg) + margin
> lower.band = predict(reg) - margin

    # Plots
> plot(X,Y,xlab="X = number of copiers",ylab="Y = service time",xlim=c(20,30))
> abline(reg,col="red")
> lines(X,upper.band,col="blue")
> lines(X,lower.band,col="blue")
```