# PSTAT 126 - Regression Analysis – Fall 2017

Kendall Brown ,Tuesday 6pm

Homework #5 – Due in Section Nov 21 - 23

Note: These course materials are the sole property of Dr. Todd M. Gross. They are strictly for use by students enrolled in a course taught by Dr. Gross. They may not be altered, excerpted, distributed, or posted to any website or other document-sharing service without express permission.

*Notes: Use R to perform data analyses, and provide annotated code and output.*

1. Explain in what sense the regression sum of squares SSR(X1) in simple linear regression is an extra sum of squares?  
     
   1. As more predictor variables are added to a linear model the SSR and SSE increase and decrease respectively, whilst still summing to the same value SSTO. In the case of SSR(X1), we consider the model Y(i)=b(0)+e(i). In this situation the SSR=0 and the SSE=SSTO. Adding the X1 parameter

we take Y(i)=b(0)+b(1)X(1)+e(i). By doing this we include an extra sum of squares into our model. Subtracting away from the SSE and adding back into SSR we achieve SSTO=SSR+SSE. Where SSR(X1) is an extra sum of squares added to the model with the goal of decreasing error.

1. State the number of degrees of freedom that are associated with each of the following extra sum of squares:
   1. SSR(X1|X2)=1
   2. SSR(X2|X1,X3)=1
   3. SSR(X1,X2|X3,X4)=2
   4. SSR(X1,X2,X3|X4, X5)=3
2. Complete the following problem:

The model *Y = β0 + β1X1 + β2X2 + ε* is ﬁtted to a data set with 10 observations.

(a) Supply values for the missing entries in the following ANOVA table.

Table 1: ANOVA Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | SS | df | MS | F |
| Regression | 20 | 2 | 10 | 5 df1=2,df2=7 |
| Error | 14 | 7 | 2 |  |
| Total | 34 | 9 |  |  |

(b) Is the overall regression signiﬁcant (Hint: Use R to obtain the critical value of F)?  
From the F-statistic we observe a p-value of .0447. At a 5% alpha we reject significance, at a 1% alpha we fail to reject significance.

(c) What proportion of the variance of Y is explained by the regression model?  
  
R2=SSR/SSTo=20/34=.5882. Approximately 58.82% of the variance can be explained by the regression model.

(d) Another model *Y = β0 +β1X1 +ε* is ﬁtted to the same data set. Supply values for the missing entries in the following ANOVA table.

Table 2: ANOVA Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | SS | df | MS | F |
| Regression | 10 | 1 | 10 | 3.33 df1=1,df2=8 |
| Error | 24 | 8 | 3 |  |
| Total | 34 |  |  |  |

(e) Use the extra sum of squares principle to test if variable X2 is important. State your hypothesis.  
  
X2 has MS of 10 and a corresponding F-statistic of 5 with df1=1 and df2=7. From this we compute a p-value of .0604 and conclude there is not enough evidence to reject X2’s significance at a 5% alpha.

1. A hospital administrator was interested in the relationship between patient satisfaction (Y) and patient age (X1), severity of illness (X2) and anxiety level (X3). The data for this problem is posted on Gauchospace in **patsat.txt**. Higher scores on Y, X2, and X3 indicate more satisfaction, increased severity, and higher anxiety, respectively.
   1. Generate a scatterplot matrix for these data.
   2. Fit the full regression model, *Y = β0 + β1X1 + β2X2 + β3X3 + ε* to these data. Interpret the value of each of the regression coefficients, and interpret R2.
   3. Test the significance of the full regression model. State the null and alternative hypotheses, the value of F, the p-value, and your statistical conclusion.
   4. Obtain the analysis of variance table that decomposes the regression sum of squares into extra sum of squares associate with X2; with X1 given X2; and with X3 given X2 and X1.
   5. Test whether X3 can be dropped from the regression model given that X1 and X2 are retained. State the null and alternative hypotheses, the value of F, the p-value, and your statistical conclusion.
   6. Test whether both X2 and X3 can be dropped from the regression model given that X1 is retained. State the null and alternative hypotheses, the value of F, the p-value, and your statistical conclusion.
2. Critical values of F-statistics

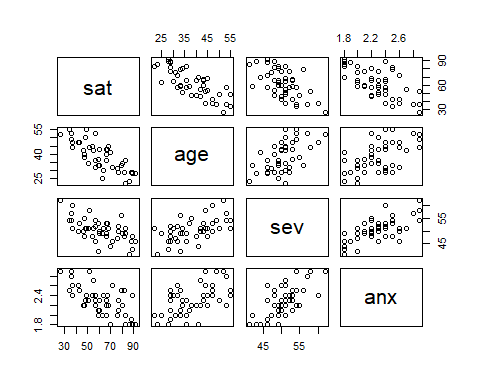
1-pf(5,2,7)#p-value of regression model

## [1] 0.04479937

1-pf(5,1,7)#p-vlaue of X2

## [1] 0.06042742

patsat=read.table("C:/Users/kebro/Desktop/Pstat 126/patsat.txt",header=T)  
plot(patsat)



attach(patsat)  
fitpat=lm(sat~age+sev+anx)  
summary(fitpat)

##   
## Call:  
## lm(formula = sat ~ age + sev + anx)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.3524 -6.4230 0.5196 8.3715 17.1601   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 158.4913 18.1259 8.744 5.26e-11 \*\*\*  
## age -1.1416 0.2148 -5.315 3.81e-06 \*\*\*  
## sev -0.4420 0.4920 -0.898 0.3741   
## anx -13.4702 7.0997 -1.897 0.0647 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.06 on 42 degrees of freedom  
## Multiple R-squared: 0.6822, Adjusted R-squared: 0.6595   
## F-statistic: 30.05 on 3 and 42 DF, p-value: 1.542e-10

1. From this model, we can expect a patient to have a satisfaction level of 158.4913 decreasing by 1.1416, .4420, and 13.4702 per unit increase of age, severity, and anxiety level respectively. From R2 we estimate that 68.22% of satisfactions observed variation can be accounted for by age, severity, and anxiety level.
2. We calculate a F-statistic of 30.05 on 3 and 42 degrees of freedom with a corresponding p-value of approvimately 0. We reject the null hypothesis of the model bing insignificant and conclude there is not enough evidence to reject significance.

anova(lm(sat~sev+age+anx))

## Analysis of Variance Table  
##   
## Response: sat  
## Df Sum Sq Mean Sq F value Pr(>F)   
## sev 1 4860.3 4860.3 48.0439 1.822e-08 \*\*\*  
## age 1 3896.0 3896.0 38.5126 2.008e-07 \*\*\*  
## anx 1 364.2 364.2 3.5997 0.06468 .   
## Residuals 42 4248.8 101.2   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

fitpat2=lm(sat~age+sev)  
fitpat3=lm(sat~age+sev+anx)  
anova(fitpat2,fitpat3)

## Analysis of Variance Table  
##   
## Model 1: sat ~ age + sev  
## Model 2: sat ~ age + sev + anx  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 43 4613.0   
## 2 42 4248.8 1 364.16 3.5997 0.06468 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We test the null hypothesis, the slope correlating to anxiety level is 0, against the alt hypothesis, the slope of anxiety level is not 0. we obtain an F-statistic of 3.5997 with a correlating p-value of .06468. We fail to reject the null at a 5% alpha level and conclude that the anxiety predictor is insignificant to a model given age and severity level.

fitpat4=lm(sat~age)  
anova(fitpat4,fitpat3)

## Analysis of Variance Table  
##   
## Model 1: sat ~ age  
## Model 2: sat ~ age + sev + anx  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 44 5093.9   
## 2 42 4248.8 2 845.07 4.1768 0.02216 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We test the null hypothesis, the slopes correlating to severity and anxiety level are 0, against the alt hypothesis, the slopes of severity and anxiety level are not 0. we obtain an F-statistic of 4.1768 with a correlating p-value of .02216. We reject the null at a 5% alpha level and conclude that the severity and anxiety predictors are significant to a model given age.