STAT 331 – Lecture 11 (Data analysis)

We study a dataset taken from *Applied Linear Statistical Models* (5th edition) by Kutner et al. (2005). The primary objective of the study was to determine whether infection surveillance and control programs have reduced the risk of nosocomial (hospital-acquired) infection in United States hospitals. The data set consists of a random sample of 113 hospital selected from an original of 338 hospitals surveyed. The following variables were observed:

- `ID`: Hospital identification number (1–113) → not explanatory var.
- `Stay`: Average length of stay of all patients in hospital (in days)
- `Age`: Average patient age (in years)
- `InfctRsk` = Risk of acquiring infection in hospital
- `Culture` = Number of cultures performed / number of patients without signs or symptoms of hospital-acquired infection, times 100
- `Xray` = Number of xrays/Number of patients without signs or symptoms of pneumonia, times 100
- `Beds` = Average number of beds (in use) during study period
- `MedSchool` = Med school affiliation (1=Yes, 0=No)
- `Region` = Geographic region, where: 1 =NE, 2=NC, 3=S, 4=W → convert into indicator variables.
- `Census` = Average number of patients in hospital per day during study period

1 Read and view data from csv file

```
> library(data.table); library(dplyr)
> hospital_dat_tmp = fread("hospital_dat.csv", header=T)
> hospital_dat = hospital_dat_tmp %>% select(Stay,Age,InfctRsk,Xray,Region)

> ## View data
> head(hospital_dat)
    Stay  Age InfctRsk  Xray Region
1:  7.13 55.7       4.1  39.6      4
2:  8.82 58.2       1.6  51.7      2
3:  8.34 56.9       2.7  74.0      3
4:  8.95 53.7       5.6 122.8      4
5: 11.20 56.5       5.7  88.9      1
6:  9.76 50.9       5.1  97.0      2
```

→ requires recoding.

2 Fit a multiple linear regression model

First define our model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \epsilon_i, \quad i = 1, \ldots, 113, \quad \epsilon_i \sim N(0, \sigma^2) \text{ iid}$$

where x_1 denotes `Stay`, x_2 denotes `Age`, x_3 denotes `Xray`, x_4 denotes an indicator variable for `NC` region, x_5 denotes an indicator variable for `S` region and x_6 denotes an indicator variable for `W` region. We first code the `Region` variable to reflect the conversion of this variable into indicator variables given by x_4 to x_6.

```
## create some indicators variables for region
hospital_dat$regionNC = ifelse(hospital_dat$Region==2, 1, 0)
hospital_dat$regionS = ifelse(hospital_dat$Region==3, 1, 0)
hospital_dat$regionW = ifelse(hospital_dat$Region==4, 1, 0)
hospital_dat$Region = NULL
```

Next, we fit the multiple linear regression in `R`:

```
> myfit = lm(InfctRsk ~ Stay + Age + Xray + regionNC + regionS + regionW,
           data = hospital_dat)
> summary(myfit) ## Shows a summary of our fitted model

Call:
lm(formula = InfctRsk ~ Stay + Age + Xray + regionNC + regionS +
    regionW, data = hospital_dat)

Residuals:
     Min       1Q   Median       3Q      Max
-2.69155 -0.67937  0.00912  0.70820  2.62489

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.49901    1.41391   0.353  0.72485
Stay         0.36394    0.06516   5.586 1.81e-07 ***
Age         -0.02601    0.02363  -1.101  0.27349
Xray         0.01908    0.00578   3.302  0.00131 **
regionNC     0.13120    0.29301   0.448  0.65523
regionS      0.04471    0.29677   0.151  0.88052
regionW      0.84512    0.38150   2.215  0.02888 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.067 on 106 degrees of freedom
Multiple R-squared:  0.4007,    Adjusted R-squared:  0.3667
F-statistic: 11.81 on 6 and 106 DF,  p-value: 4.147e-10
```

3 Obtain the ANOVA table

To obtain the ANOVA table in R, we use the `anova()` function. R will display an ANOVA table where each explanatory variable is given its own separate row. In order to get SS(reg), we will need to add up the sums of squares for each explanatory variable. (need to sum up □)

```
> anova(myfit)
Analysis of Variance Table

Response: InfctRsk
           Df  Sum Sq Mean Sq F value    Pr(>F)
Stay        1  57.305  57.305 50.3286 1.526e-10 ***
Age         1   2.075   2.075  1.8224  0.179899
Xray        1  13.719  13.719 12.0486  0.000751 ***
regionNC    1   0.057   0.057  0.0500  0.823467
regionS     1   1.943   1.943  1.7063  0.194297
regionW     1   5.587   5.587  4.9072  0.028883 *
Residuals 106 120.694   1.139
---
Signif. codes:  0 '***' 0.001 '*' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Find SSRes:
> SSRes = anova(myfit)$'Sum Sq'[7]
> SSRes
[1] 120.6936

## Find SSReg:
> SSReg = sum(anova(myfit)$'Sum Sq'[1:6])
> SSReg
[1] 80.6862

## Find R^2:
> R2 = SSReg/(SSRes + SSReg);
> R2
[1] 0.4006668

## F-statistic (from ANOVA table)
> dfRes = anova(myfit)$'Df'[7]; dfRes;
[1] 106
> dfReg = sum(anova(myfit)$'Df'[1:6]); dfReg;
[1] 6
> MSRes = SSRes/dfRes
> MSReg = SSReg/dfReg
> F_statistic = MSReg/MSRes; F_statistic
[1] 11.81054

## F-test
> pval <- 1 - pf(F_statistic, df1 = dfReg, df2 = dfRes); pval
[1] 4.147185e-10

## or compare F-statistic with the following:
> qf(0.95, df1=dfReg, df2=dfRes)
[1] 2.185293
```

SS(tot)

MS(reg) / MS(res)

or n-p-1 = 106

or # of covariates

CDF of a R.V. $Y \sim F_{df1,df2}$ at F (or $P(Y \le F)$)

F

↳ equals $c = F_{0.95, df1, df2}$.

4 Fitting reduced models

4.1 Fit a reduced multiple linear regression without Stay

Consider now a new regression model where we omit the variable Stay.

```
> myfit_reduced = lm(InfctRsk ~ Age + Xray + regionNC + regionS + regionW,
                     data = hospital_dat)
> summary(myfit_reduced)

Call:
lm(formula = InfctRsk ~ Age + Xray + regionNC + regionS + regionW,
    data = hospital_dat)

Residuals:
     Min       1Q   Median       3Q      Max
-2.98981 -0.82817 -0.04748  0.84127  3.00838

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.1846408  1.5641677   1.397    0.165
Age          0.0009519  0.0261959   0.036    0.971
Xray         0.0288228  0.0062402   4.619 1.08e-05 ***
regionNC    -0.2375510  0.3232605  -0.735    0.464
regionS     -0.4729324  0.3192500  -1.481    0.141
regionW     -0.0782325  0.3893377  -0.201    0.841
---
Signif. codes:  0 '***' 0.001 '*' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.208 on 107 degrees of freedom
Multiple R-squared:  0.2243,    Adjusted R-squared:  0.188
F-statistic: 6.187 on 5 and 107 DF,  p-value: 4.472e-05

> anova(myfit_reduced) ——→ To find SS_A(res)
Analysis of Variance Table

Response: InfctRsk
           Df  Sum Sq Mean Sq F value    Pr(>F)
Age         1   0.000   0.000  0.0002    0.9898
Xray        1  41.415  41.415 28.3667 5.595e-07 ***
regionNC    1   0.001   0.001  0.0010    0.9749
regionS     1   3.686   3.686  2.5247    0.1150
regionW     1   0.059   0.059  0.0404    0.8411
Residuals 107 156.218   1.460
---
Signif. codes:  0 '***' 0.001 '*' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In order to test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$, we can use the F-test based on an F-statistic that we calculate below: → $A = [0\ 1\ 0\ 0\ 0\ 0\ 0]_{1\times 7}$

$$F = \frac{[SS_A(res) - SSCres)]/l}{SSCres)/n-p-1}$$

```
> ## Extract SS(res)
> SSRes_reduced = anova(myfit_reduced)$'Sum Sq'[6]
> l = 1 #we are testing H_0: beta_1=0
> n = nrow(hospital_dat)
> p = length(myfit$coefficients)-1

## F_statistic
> F_statistic = ((SSRes_reduced-SSRes)/l) / (SSRes/(n-p-1));
```

```
[1] 31.1997
> pval <- 1 - pf(F_statistic, df1 = l, df2 = n-p-1); pval
[1] 1.813912e-07
```

Note that this is the same p-value that we can obtain via a t-test from before (see also the results from `summary(myfit)`).

4.2 Fit a reduced multiple linear regression without Region

Now, we consider a new regression model where we omit the variables associated with `Region`:

```
> myfit_reduced = lm(InfctRsk ~ Stay + Age + Xray, data = hospital_dat)
> summary(myfit_reduced) ## Shows a summary of our fitted model

Call:
lm(formula = InfctRsk ~ Stay + Age + Xray, data = hospital_dat)

Residuals:
     Min       1Q   Median       3Q      Max
-2.77320 -0.73779 -0.03345  0.73308  2.56331

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.001162   1.314724   0.761 0.448003
Stay         0.308181   0.059396   5.189 9.88e-07 ***
Age         -0.023005   0.023516  -0.978 0.330098
Xray         0.019661   0.005759   3.414 0.000899 ***
---
Signif. codes:  0 '***' 0.001 '*' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.085 on 109 degrees of freedom
Multiple R-squared:  0.363,     Adjusted R-squared:  0.3455
F-statistic:  20.7 on 3 and 109 DF,  p-value: 1.087e-10

> anova(myfit_reduced)
Analysis of Variance Table

Response: InfctRsk
           Df  Sum Sq Mean Sq F value    Pr(>F)
Stay        1  57.305  57.305 48.6920 2.444e-10 ***
Age         1   2.075   2.075  1.7632 0.1870031
Xray        1  13.719  13.719 11.6568 0.0008992 ***
Residuals 109 128.281   1.177
```

We wish to test $H_0 : \beta_4 = \beta_5 = \beta_6 = 0$ vs. H_a : at least one of the βs is not zero.

```
> ## Extract SS(res)
> SSRes_reduced = anova(myfit_reduced)$'Sum Sq'[4]
> l = 3 #we are testing H_0: beta_3=beta_4=beta_5=0
> n = nrow(hospital_dat)
> p = length(myfit$coefficients)-1

> ## F-statistic
> F_statistic = ((SSRes_reduced-SSRes)/l) / (SSRes/(n-p-1));
[1] 2.22118
> pval <- 1 - pf(F_statistic, df1 = l, df2 = n-p-1); pval
[1] 0.0899514
```