STAT 331: Applied Linear Models (Fall 2019)

Lecture 23: A Case Study

Instructor: Leilei Zeng

Department of Statistics and Actuarial Science
University of Waterloo

December 3, 2019

A Case Study of SENIC Data

The US Department of Health is concerned about the infections in hospitals. A sample of 113 hospitals were randomly selected.

The data file "`senic.dat`" contains the observations of following variables:

Response :	`Risk`	(average infection risk in percentage)
Covariates :	`Stay`	(average length of hospital stay in days)
	`Age`	(average age of patient in years)
	`Culture`	(number of bacteriological tests per non-infected patient $\times 100$)
	`Xray`	(number of X-ray per non-infected patient $\times 100$)
	`Beds`	(number of beds)
	`School`	(medical school affiliation 1 =yes, 0 = no)
	`Region`	(geographic region 1=NE, 2=NC, 3=S, 4=W)
	`Pat`	(average number of patients a day)
	`Nurse`	(average number of full-time trained nurses)
	`Facility`	(percentage of available facilities out of a total list of 35)

The objective is to find a good statistical model that describes the infection risk and possible causes.

Data Exploration

As a first step, we produce some summary statistics for each of the variables

```
> senic=read.table("senic.dat", col.names=c("ID", "Stay", "Age",
    "Risk", "Culture", "Xray", "Beds", "School", "Region", "Pat",
    "Nurse", "Facility"))
> senic=data.frame(senic)
> summary(senic)
       ID              Stay             Age              Risk
 Min.   :  1   Min.   : 6.700   Min.   :  38.8   Min.   :  1.30
 1st Qu.: 29   1st Qu.: 8.340   1st Qu.:  50.9   1st Qu.:  3.70
 Median : 57   Median : 9.420   Median :  53.2   Median :  4.40
 Mean   : 57   Mean   : 9.648   Mean   : 141.2   Mean   : 13.15
 3rd Qu.: 85   3rd Qu.:10.470   3rd Qu.:  56.2   3rd Qu.:  5.20
 Max.   :113   Max.   :19.560   Max.   :9999.0   Max.   :999.00
    Culture             Xray              Beds           School
 Min.   :   1.6   Min.   :  39.6   Min.   : 29.0   Min.   :1.00
 1st Qu.:   8.4   1st Qu.:  69.5   1st Qu.:106.0   1st Qu.:2.00
 Median :  14.0   Median :  82.5   Median :186.0   Median :2.00
 Mean   : 104.0   Mean   : 169.4   Mean   :252.2   Mean   :1.85
 3rd Qu.:  20.3   3rd Qu.:  95.9   3rd Qu.:312.0   3rd Qu.:2.00
 Max.   :9999.0   Max.   :9999.0   Max.   :835.0   Max.   :2.00
```

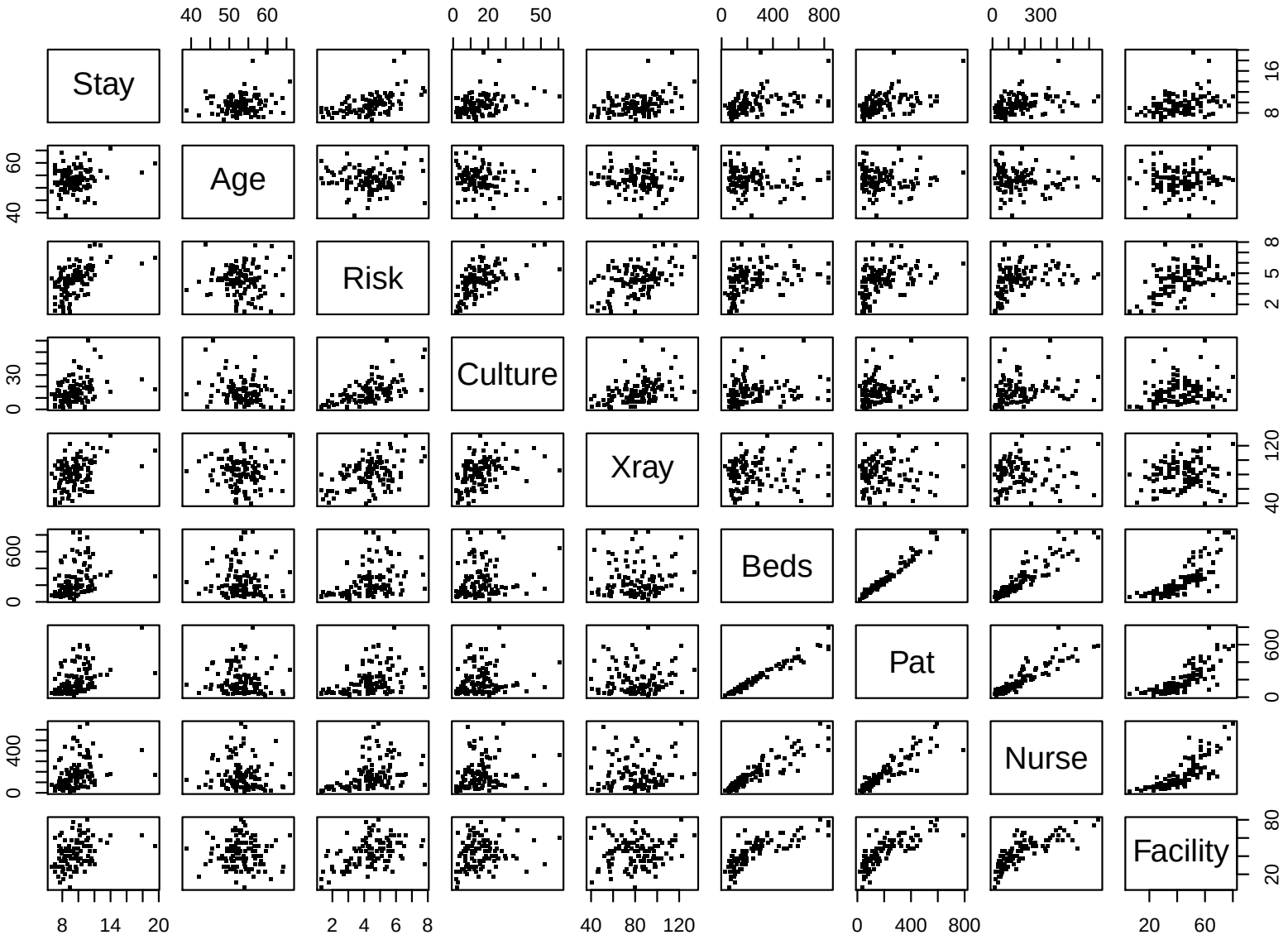
```
     Region            Pat              Nurse          Facility
 Min.   :1.000   Min.   : 20.0   Min.   : 14.0   Min.   : 5.70
 1st Qu.:1.000   1st Qu.: 68.0   1st Qu.: 66.0   1st Qu.:31.40
 Median :2.000   Median :143.0   Median :132.0   Median :42.90
 Mean   :2.345   Mean   :191.4   Mean   :173.2   Mean   :43.15
 3rd Qu.:3.000   3rd Qu.:252.0   3rd Qu.:218.0   3rd Qu.:54.30
 Max.   :4.000   Max.   :791.0   Max.   :656.0   Max.   :80.00
```

Missing Data

- Variables `Age`, `Risk`, `Culture`, and `Xray` have unreasonably large values of "9999" or "999" which are used to represent missing data
- One shall use the default symbol "`NA`" for missing values in R so that missing observations can be omitted in the analysis

```
> senic$Age[senic$Age>500]=NA
> senic$Risk[senic$Risk>500]=NA
> senic$Xray[senic$Xray>500]=NA
> senic$Culture[senic$Culture>500]=NA
```

Pairwise Scatter Plots



Think about how the variables should influence `Risk`

- `Risk` should increase with length of `Stay`, there is some indication of such a relationship in the scatter plot
- `Age` expected to be positively correlated with risk of infection-presumably, older patients are more susceptible, however, the scatter plot does not indicate clearly whether or not this is the case
- `Culture` and `Xray` are measures of how hard the hospital looks for otherwise unsuspected infection, if you look hard you should find more so that the coefficients of these variables would be expected to be positive
- `Beds`, `Pat`, and `Nurses` all measure the size of the hospital, they may be highly correlated which will make it hard to tell which of them should be used in the model
- `Facilities` measures the sophistication of medical treatment at the hospital, it seems positively related to Risk which might suggest more exotic diseases among the patients

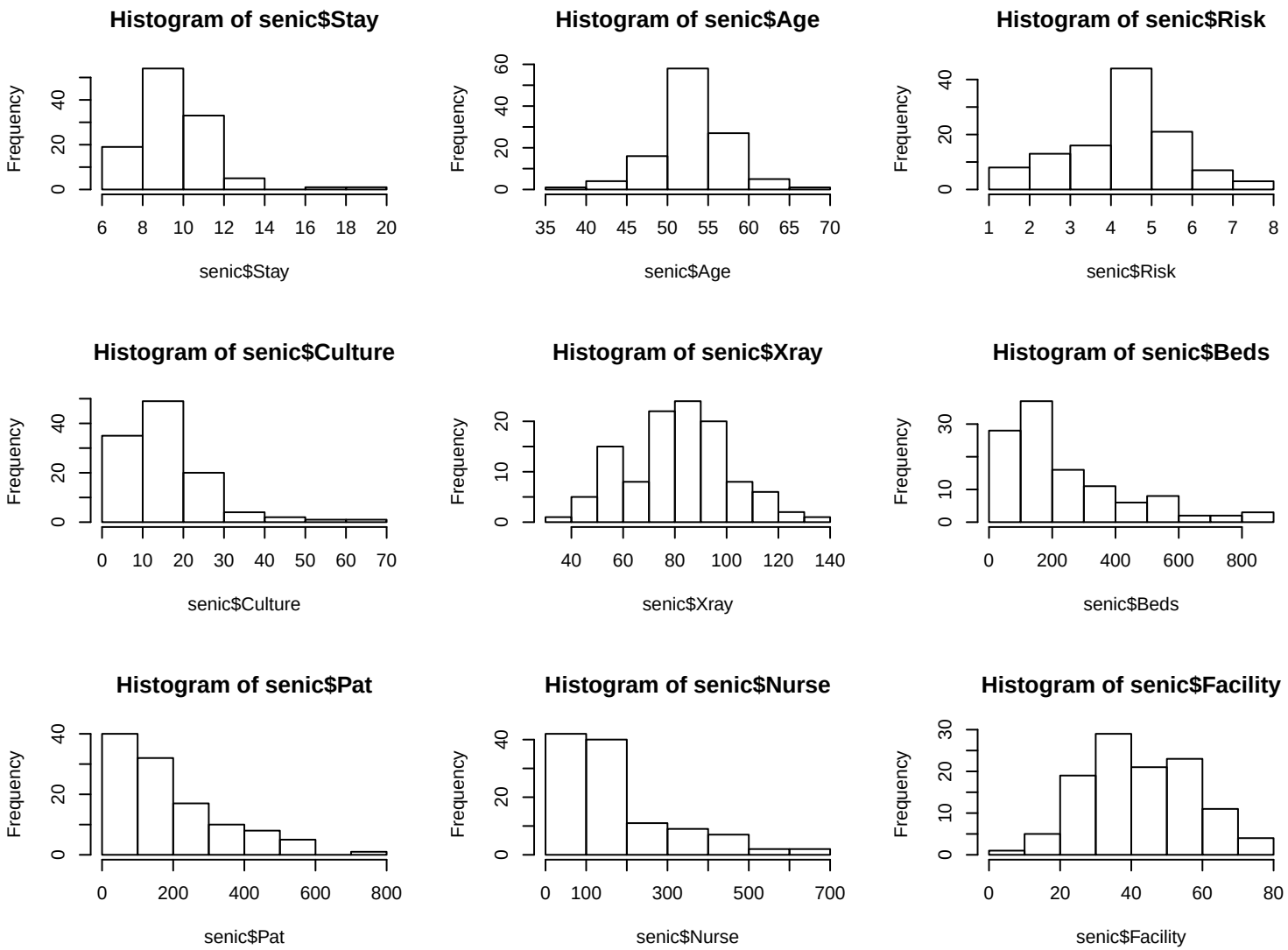
These hypotheses can be checked formally by multiple regression.

Pairwise Correlation Matrix

```
> round(cor(senic[, -c(1, 8, 9)], use="pairwise.complete.obs"), 2)
          Stay   Age Risk Culture  Xray  Beds   Pat Nurse Facility
Stay      1.00  0.19 0.53    0.33  0.38  0.41  0.47  0.34     0.36
Age       0.19  1.00 0.00   -0.23 -0.01 -0.05 -0.05 -0.08    -0.03
Risk      0.53  0.00 1.00    0.57  0.45  0.37  0.39  0.40     0.42
Culture   0.33 -0.23 0.57    1.00  0.43  0.15  0.15  0.21     0.18
Xray      0.38 -0.01 0.45    0.43  1.00  0.05  0.06  0.08     0.11
Beds      0.41 -0.05 0.37    0.15  0.05  1.00  0.98  0.92     0.79
Pat       0.47 -0.05 0.39    0.15  0.06  0.98  1.00  0.91     0.78
Nurse     0.34 -0.08 0.40    0.21  0.08  0.92  0.91  1.00     0.78
Facility  0.36 -0.03 0.42    0.18  0.11  0.79  0.78  0.78     1.00
```

We explore the relationship between any two continuous variables via pairwise scatter plots and correlation coefficients, as expected, several of the variables are quite highly correlated, think about collinearity.

Histograms of Variables



`Stay`, `Culture`, `Beds`, `Pat` and `Nurse` are skewed, may consider transformation

Initial Model Fitting

We begin by fitting a model using all the explanatory variables

```
> senic$School.f = factor(senic$School)
> senic$Region.f = factor(senic$Region)
> names(senic)
 [1] "ID"       "Stay"     "Age"      "Risk"     "Culture"
 [6] "Xray"     "Beds"     "School"   "Region"   "Pat"
[11] "Nurse"    "Facility" "School.f" "Region.f"
>
>
> fitfull =lm(Risk~., data=senic[, -c(1, 8, 9)])
> summary(fitfull)
Call:
lm(formula = Risk ~ ., data = senic[, -c(1, 8, 9)])

Residuals:
     Min       1Q   Median       3Q      Max
-1.54227 -0.59967 -0.08366  0.59201  2.57579
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.216133   1.323901  -1.674  0.09736 .
Stay         0.216548   0.070635   3.066  0.00281 **
Age          0.016070   0.022226   0.723  0.47140
Culture      0.056734   0.010864   5.222 1.01e-06 ***
Xray         0.010682   0.005297   2.016  0.04653 *
Beds        -0.003169   0.002693  -1.177  0.24216
Pat          0.003868   0.003471   1.114  0.26795
Nurse        0.001491   0.001704   0.875  0.38367
Facility     0.019428   0.010205   1.904  0.05990 .
School.f2    0.654834   0.324148   2.020  0.04612 *
Region.f2    0.285863   0.261019   1.095  0.27615
Region.f3    0.158685   0.273825   0.580  0.56359
Region.f4    0.982185   0.336859   2.916  0.00441 **
---
Residual standard error: 0.9184 on 97 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.5842,Adjusted R-squared:  0.5328
F-statistic: 11.36 on 12 and 97 DF,  p-value: 7.245e-14
```

The output suggests that `Stay`, `Culture`, `Xray` and `School` are important predictors

For the categorical predictor `Region`, the coefficient for one of it indicator variables is significant, which implies that it might be an important factor but F-test is required to make a formal conclusion

None of the three variables related to the hospital size: `Beds`, `Pat`, and `Nurses` is important, however, exploratory analysis indicates that there might exist very strong collinearity between them

Checking multicollinearity

```
> library(car)
> vif(fitfull)
              GVIF Df GVIF^(1/(2*Df))
Stay      2.396134  1        1.547945
Age       1.278664  1        1.130780
Culture   1.567189  1        1.251874
Xray      1.393640  1        1.180525
Beds     35.537963  1        5.961373
Pat      37.580492  1        6.130293
Nurse     7.416404  1        2.723307
Facility  3.125182  1        1.767818
School.f  1.790363  1        1.338044
Region.f  1.745751  3        1.097313
```

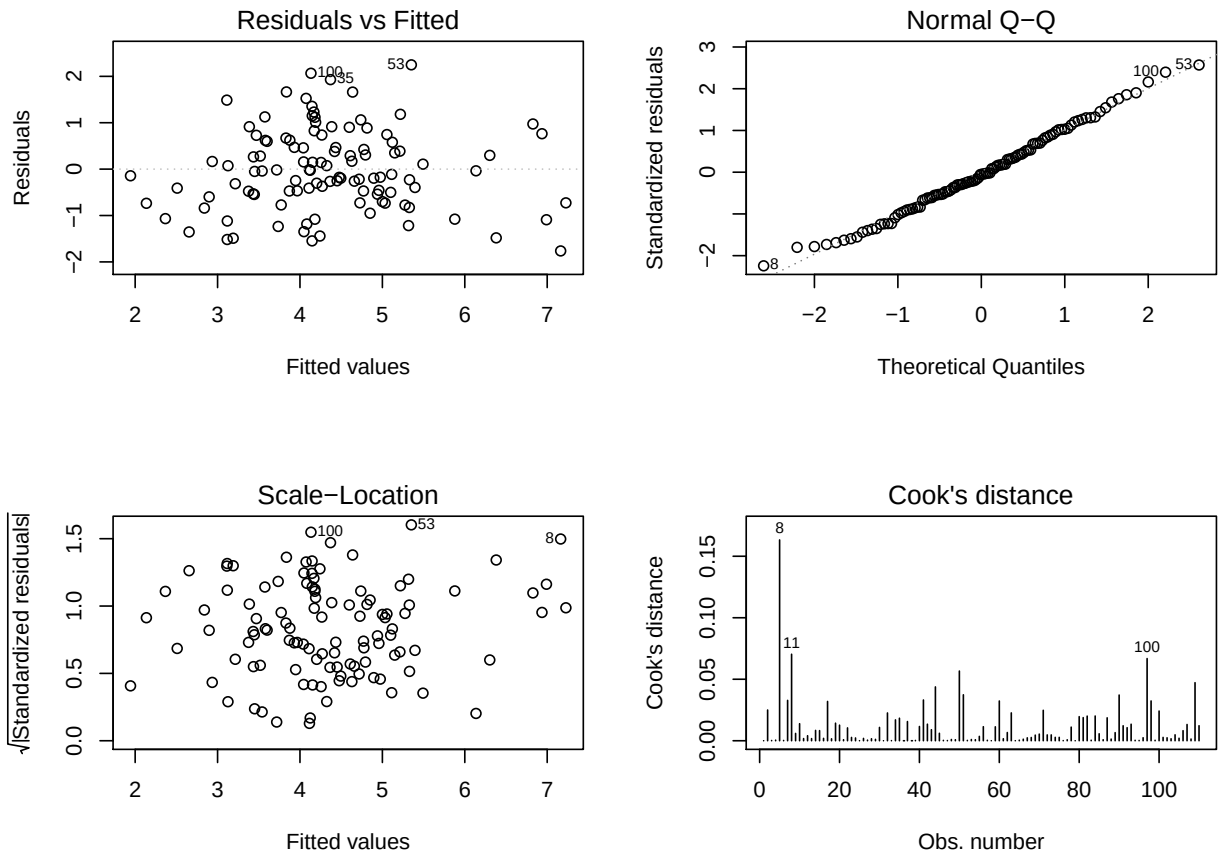
- If all predictors in a linear model have 1 DF (only one coefficient for a variable), then the usual *VIF*s are calculated (i.e. $VIF = GVIF$)
- For any categorical predictors with more than 1 DF (more than one coefficients for a variable), GVIF is calculated
- Measures of hospital size: `Beds`, `Pat`, and `Nurses` are multi-collinear, we drop `Beds` and `Pat` which has $VIF >> 10$, and keep `Nurse` in the model.

- Update the full model by dropping Bed and Pat

```
> fitfull.new =update(fitfull, .~.-Beds-Pat)
> summary(fitfull.new)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.079885   1.312931  -1.584 0.116348
Stay         0.238149   0.061446   3.876 0.000191 ***
Age          0.013241   0.021939   0.604 0.547520
Culture      0.054974   0.010543   5.214 1.01e-06 ***
Xray         0.010637   0.005243   2.029 0.045156 *
Nurse        0.001356   0.001086   1.249 0.214464
Facility     0.017078   0.009593   1.780 0.078099 .
School.f2    0.589380   0.311541   1.892 0.061437 .
Region.f2    0.224129   0.252950   0.886 0.377732
Region.f3    0.113517   0.263121   0.431 0.667098
Region.f4    0.911632   0.330528   2.758 0.006925 **
---

Residual standard error: 0.9157 on 99 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.5781,Adjusted R-squared:  0.5355
F-statistic: 13.57 on 10 and 99 DF,  p-value: 1.009e-14
```

Diagnostic Plots

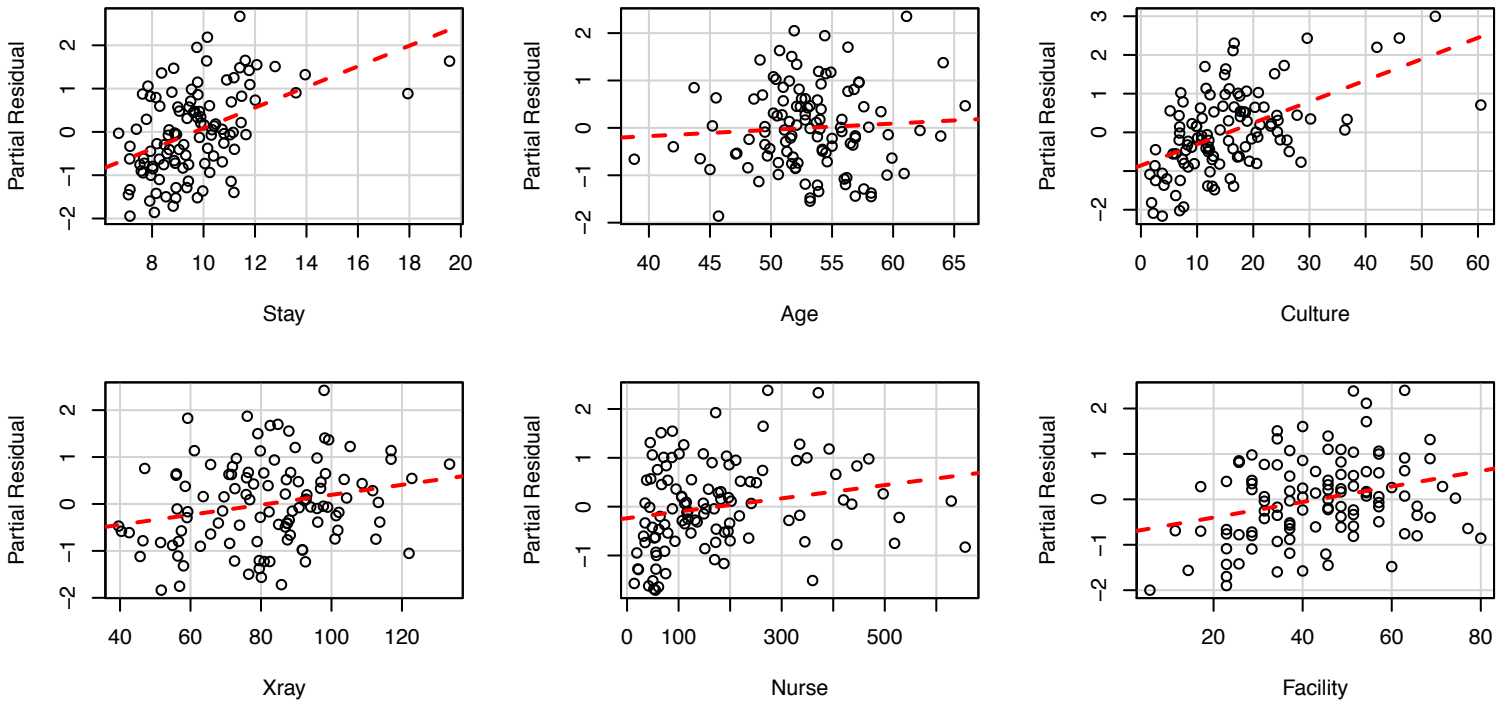
```
> plot(fitfull.new, which=1:4)
```



- Plot r vs $\widehat{y}$: some deviation from a random scatter, think about non-linearity

Partial Residual Plots

```
> crPlots(fitfull.new, main="", ylab="Partial Residual", smooth=F)
```



- Stay, Culture and Nurse are skewed, and their partial residuals plots also show some curvatures, a *log* transformation for these variables might improve the fit of the model

Data Transformation

We create a new dataset that contains

- response variable: `Risk`
- explanatory variables:

ln(`Stay`), ln(`Culture`), ln(`Nurse`),
`Age`, `Xray`, `Facility`, `School`, `Region`

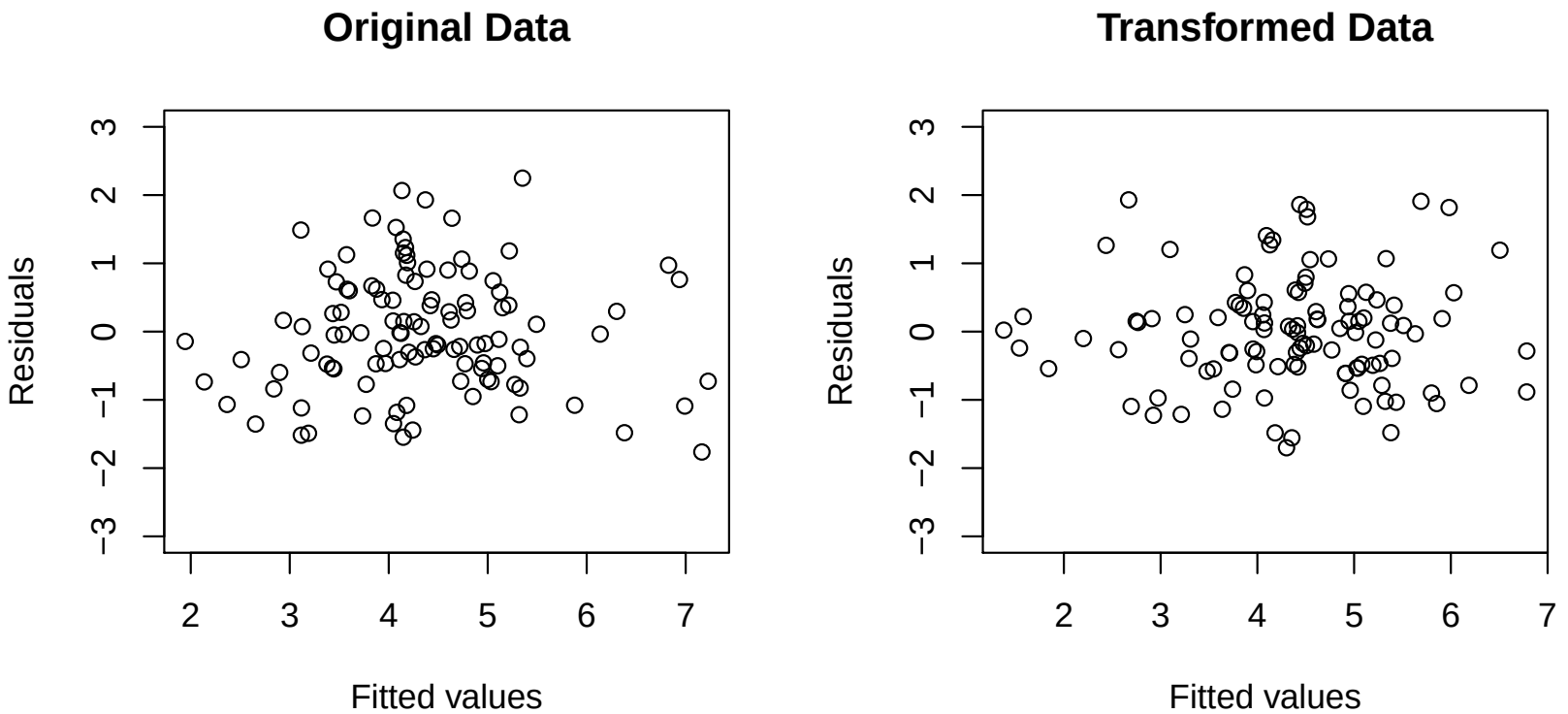
```
> senic.new=data.frame(cbind(lnStay=log(senic$Stay), Age=senic$Age,
    Risk=senic$Risk, lnCulture=log(senic$Culture), Xray=senic$Xray,
    lnNurse=log(senic$Nurse), Facility=senic$Facility,
    School.f=factor(senic$School), Region.f=factor(senic$Region))
```

Fit a full model include all the explanatory variables in this new dataset

```
> fitfull.tr <-lm(Risk~., data=na.omit(senic.new))
```

```
> summary(fitfull.tr)

Call:
lm(formula = Risk ~ ., data = na.omit(senic.new))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.686599   1.641912  -5.291 7.32e-07 ***
lnStay       2.607928   0.642075   4.062 9.76e-05 ***
Age          0.028828   0.020619   1.398  0.16520
lnCulture    0.863767   0.152855   5.651 1.54e-07 ***
Xray         0.005830   0.005018   1.162  0.24813
lnNurse      0.592580   0.201961   2.934  0.00416 **
Facility    -0.009573   0.010677  -0.897  0.37212
School.f2    0.311137   0.271383   1.146  0.25436
Region.f2    0.302774   0.235145   1.288  0.20089
Region.f3    0.126371   0.243808   0.518  0.60539
Region.f4    0.910764   0.313265   2.907  0.00450 **
---

Residual standard error: 0.8409 on 99 degrees of freedom
Multiple R-squared:  0.6442,Adjusted R-squared:  0.6083
F-statistic: 17.93 on 10 and 99 DF,  p-value: < 2.2e-16
```

Plot residuals vs fitted values



Model Selection - Backward Elimination

```
> drop1(fitfull.tr, test="F")
Single term deletions

Model:
Risk ~ lnStay + Age + lnCulture + Xray + lnNurse + Facility +
    School.f + Region.f
           Df Sum of Sq    RSS      AIC F value    Pr(>F)
<none>                  70.004 -27.7127
lnStay      1   11.6655 81.669 -12.7584 16.4975 9.758e-05 ***
Age         1    1.3822 71.386 -27.5619  1.9548  0.165199
lnCulture   1   22.5796 92.583   1.0391 31.9324 1.538e-07 ***
Xray        1    0.9543 70.958 -28.2232  1.3496  0.248134
lnNurse     1    6.0876 76.091 -20.5402  8.6091  0.004157 **
Facility    1    0.5684 70.572 -28.8231  0.8039  0.372118
School.f    1    0.9294 70.933 -28.2618  1.3144  0.254357
Region.f    3    7.3632 77.367 -22.7114  3.4710  0.018992 *
---
```

```
> drop1(update(fitfull.tr, .~.-Facility), test="F")
Single term deletions

          Df Sum of Sq    RSS      AIC F value    Pr(>F)
<none>                  70.572 -28.8231
lnStay     1   12.4337 83.006 -12.9728 17.6184 5.859e-05 ***
Age        1    1.0985 71.670 -29.1241  1.5565 0.2150920
lnCulture  1   22.8271 93.399   0.0042 32.3459 1.285e-07 ***
Xray       1    1.0210 71.593 -29.2430  1.4468 0.2318828
lnNurse    1    9.2215 79.794 -17.3142 13.0668 0.0004727 ***
School.f   1    1.3508 71.923 -28.7375  1.9141 0.1695923
Region.f   3    7.7914 78.363 -23.3035  3.6801 0.0145985 *
>
> drop1(update(fitfull.tr, .~.-Facility-Xray), test="F")
Single term deletions

          Df Sum of Sq     RSS      AIC F value    Pr(>F)
<none>                  71.593 -29.2430
lnStay     1    15.365  86.958  -9.8557 21.6764 9.841e-06 ***
Age        1     1.163  72.756 -29.4707  1.6405  0.203193
lnCulture  1    32.922 104.515  10.3739 46.4451 6.970e-10 ***
lnNurse    1     8.626  80.219 -18.7285 12.1697  0.000721 ***
School.f   1     1.470  73.063 -29.0068  2.0743  0.152892
Region.f   3     8.237  79.830 -23.2633  3.8736  0.011441 *
```

```
> drop1(update(fitfull.tr, .~.-Facility-Xray-Age), test="F")
Single term deletions

          Df Sum of Sq     RSS      AIC F value    Pr(>F)
<none>                  72.756 -29.4707
lnStay     1    19.990  92.746  -4.7673 28.0256 6.900e-07 ***
lnCulture  1    32.165 104.921   8.7996 45.0932 1.086e-09 ***
lnNurse    1     8.174  80.930 -19.7590 11.4591  0.001011 **
School.f   1     1.853  74.609 -28.7036  2.5984  0.110064
Region.f   3     8.656  81.412 -23.1050  4.0453  0.009214 **
>
> drop1(update(fitfull.tr,.~.-Facility-Xray-Age-School.f),test="F")
Single term deletions

Model:
Risk ~ lnStay + lnCulture + lnNurse + Region.f
          Df Sum of Sq     RSS      AIC F value    Pr(>F)
<none>                  74.609 -28.7036
lnStay     1    18.729  93.338  -6.0673 25.8561 1.655e-06 ***
lnCulture  1    32.694 107.303   9.2694 45.1344 1.037e-09 ***
lnNurse    1     6.362  80.971 -21.7023  8.7830  0.003777 **
Region.f   3     8.010  82.619 -23.4862  3.6859  0.014399 *
```

The sequence of variables removed: Facility, Xray, Age and School

The resulting model is

$$\begin{aligned}\texttt{Risk} \quad = \quad & \beta_0 + \beta_1 \ln(\texttt{Stay}) + \beta_2 \ln(\texttt{Culture}) + \beta_3 \ln(\texttt{Nurse}) \\ & + \beta_4 I(\texttt{Region} = 2) + \beta_5 I(\texttt{Region} = 3) + \beta_6 I(\texttt{Region} = 4) + \epsilon\end{aligned}$$

```
> final.b=lm(Risk~lnStay+lnCulture+lnNurse+Region.f,
     data=senic.new)
> summary(final.b)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.5421     1.2829  -5.099 1.56e-06 ***
lnStay        2.9891     0.5878   5.085 1.65e-06 ***
lnCulture     0.9000     0.1340   6.718 1.04e-09 ***
lnNurse       0.3373     0.1138   2.964  0.00378 **
Region.f2     0.2500     0.2335   1.071  0.28664
Region.f3     0.1480     0.2422   0.611  0.54246
Region.f4     0.9538     0.3125   3.052  0.00289 **
---

Residual standard error: 0.8511 on 103 degrees of freedom
Multiple R-squared:  0.6208,Adjusted R-squared:  0.5987
F-statistic: 28.11 on 6 and 103 DF,  p-value: < 2.2e-16
```

Question: Can we combine region 1, 2 and 3?

- We wish to test hypothesis

$$H_0 : \beta_4 = \beta_5 = 0$$

  The model under the null is the one with a binary indictor for region 4 only

- F-test contrasts this null model against the one with the four level categorical predictor `Region.f` gives a p-value of 0.5648, so we do not have enough evidence to reject H_0.

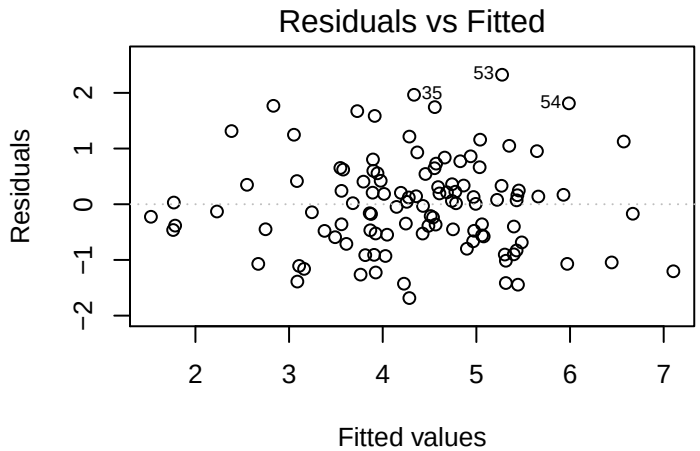
```
> senic.new$Region4=ifelse(senic.new$Region.f==4, 1, 0)
> final.b0=lm(Risk~lnStay+lnCulture+lnNurse+Region4,
      data=senic.new)
> anova(final.b, final.b0)
Analysis of Variance Table

Model 1: Risk ~ lnStay + lnCulture + lnNurse + Region.f
Model 2: Risk ~ lnStay + lnCulture + lnNurse + Region4
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    103 74.609
2    105 75.442 -2   -0.8322 0.5744 0.5648
```

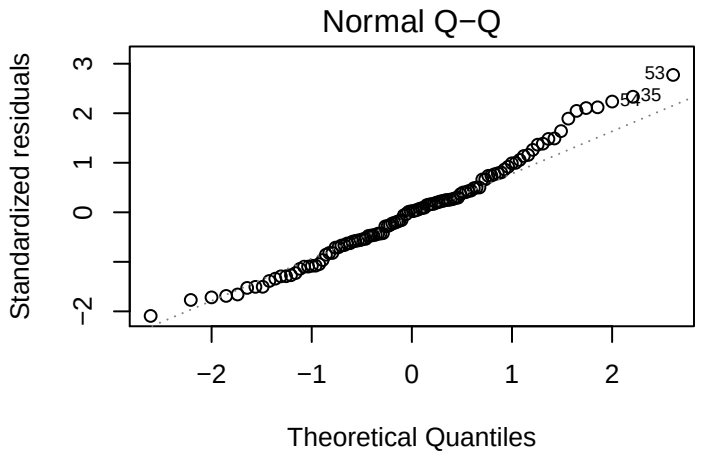
Thus we arrive at a final model

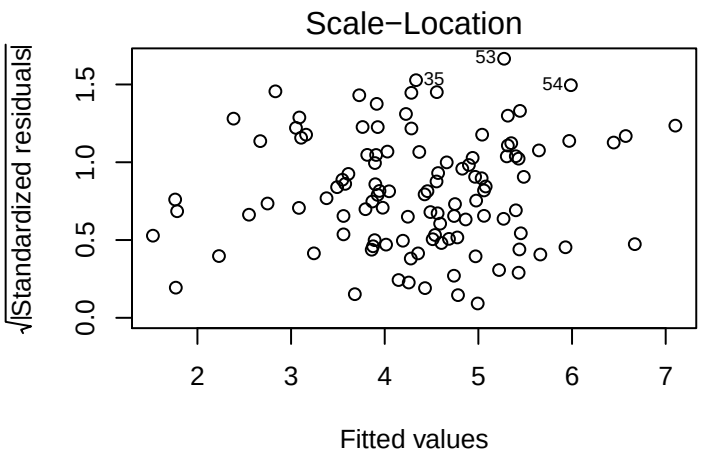
$$\begin{aligned} \texttt{Risk} &= \beta_0 + \beta_1 \ln(\texttt{Stay}) + \beta_2 \ln(\texttt{Culture}) + \beta_3 \ln(\texttt{Nurse}) \\ &\quad + \beta_4 I(\texttt{Region} = 4) + \epsilon \end{aligned}$$

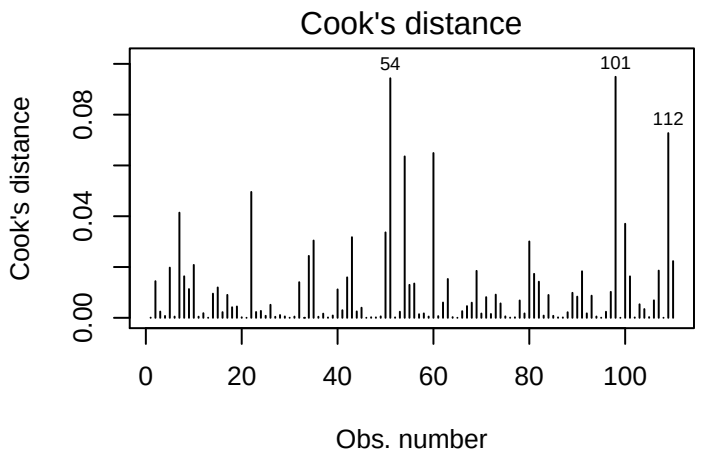
```
 > summary(final.b0)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.0598     1.1184  -5.418 3.86e-07 ***
lnStay        2.8343     0.5517   5.138 1.29e-06 ***
lnCulture     0.8742     0.1296   6.746 8.56e-10 ***
lnNurse       0.3526     0.1124   3.136  0.00222 **
Region4       0.7851     0.2484   3.161  0.00206 **
---

Residual standard error: 0.8476 on 105 degrees of freedom
Multiple R-squared:  0.6166,Adjusted R-squared:  0.602
F-statistic: 42.22 on 4 and 105 DF,  p-value: < 2.2e-16
```

Diagnostic Plots







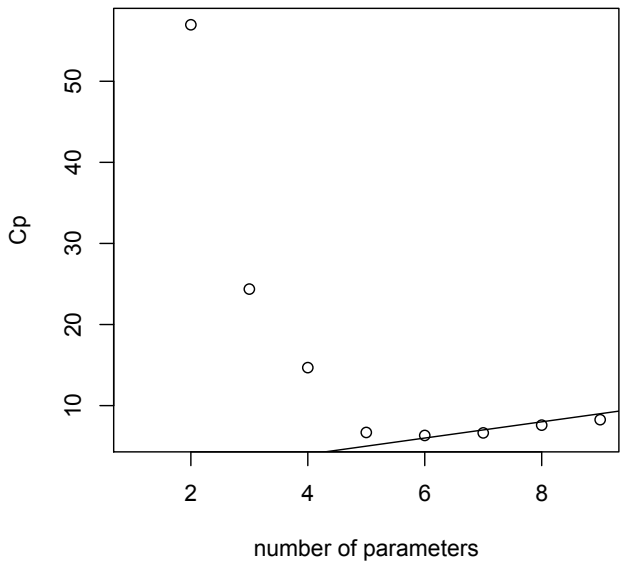


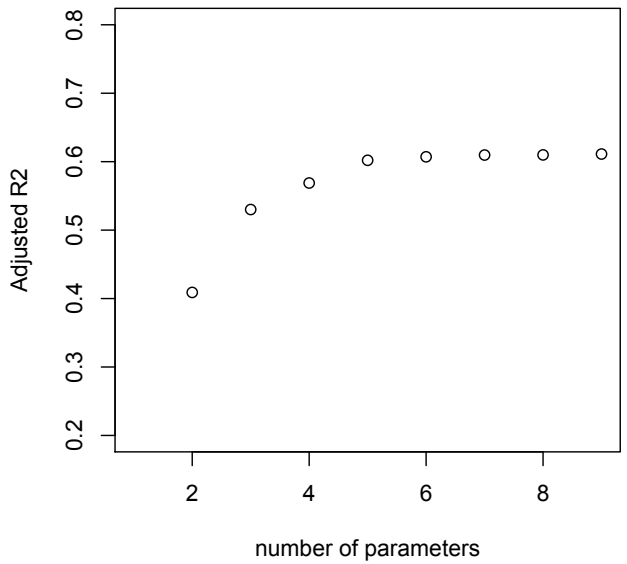
Model Selection - All Subsets

```
> best.subset=regsubsets(Risk~., data=senic.new, nbest=1)
> summary(best.subset)
1 subsets of each size up to 8
Selection Algorithm: exhaustive
   lnStay Age lnCulture Xray lnNurse Facility School.f2
1  " "    " " "*"       " "  " "     " "      " "
2  "*"    " " "*"       " "  " "     " "      " "
3  "*"    " " "*"       " "  " "     " "      " "
4  "*"    " " "*"       " "  "*"     " "      " "
5  "*"    " " "*"       " "  "*"     " "      "*"
6  "*"    " " "*"       "*"  "*"     " "      "*"
7  "*"    "*" "*"       "*"  "*"     " "      "*"
8  "*"    "*" "*"       "*"  "*"     " "      "*"
   Region.f2 Region.f3 Region.f4
1  " "       " "       " "
2  " "       " "       " "
3  " "       " "       "*"
4  " "       " "       "*"
5  " "       " "       "*"
6  " "       " "       "*"
7  " "       " "       "*"
8  "*"       " "       "*"
```

```
> summary(best.subset)$cp
[1] 56.964810 24.369942 14.684773  6.690327  6.321414  6.635915
[7]  7.591801  8.263792
> summary(best.subset)$adjr2
[1] 0.4089568 0.5300748 0.5688211 0.6019983 0.6070934 0.6096885
[7] 0.6098715 0.6111591
> summary(best.subset)$bic
[1] -49.45814 -71.00565 -76.80359 -81.95312 -79.72256 -76.81384
[7] -73.23812 -69.98504
```

- When using Mallow's C_p, the models with C_p values that fall near or under the line is good, and generally small values of C_p are desirable.
- When using adjusted R^2, select the model with the largest value, here note that the 4-, 5-, 6-, 7- and 8-predictor models are all quite compatible (differ by less than 1%)
- When using BIC (Bayesian information criterion), which is the Bayesian extension of AIC, we select the model with the lowest value of the criterion.





- When the models are very compatible according to selection criterion, we generally prefer the simpler model, so we settle with the 4-predictor model as the final model

$$\begin{aligned} \texttt{Risk} \quad = \quad & \beta_0 + \beta_1 \ln(\texttt{Stay}) + \beta_2 \ln(\texttt{Culture}) + \beta_3 \ln(\texttt{Nurse}) \\ & + \beta_4 I(\texttt{Region} = 4) + \epsilon \end{aligned}$$

  which is the same as the one selected by backward elimination