STAC67: Regression Analysis

Lecture 19

Sohee Kang

Mar. 24, 2021

Full Model (5 Predictors, 6 Parameters, n=158)

- Consider model with Predictors: Age, Tonnage, Passdens, Cabins, Length (Passenger Dropped)

```
fit0 <- lm(crew ~ age + tonnage + length + cabins + passdens)
summary(fit0)
```

```
##
## Call:
## lm(formula = crew ~ age + tonnage + length + cabins + passdens)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1306 -0.5411 -0.0952  0.4797  7.0633
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.968295   0.979282  -2.010 0.046207 *
## age         -0.005458   0.014423  -0.378 0.705611
## tonnage     -0.006110   0.010474  -0.583 0.560525
## length       0.419138   0.117648   3.563 0.000491 ***
## cabins       0.652583   0.077798   8.388 3.15e-14 ***
## passdens     0.027906   0.013319   2.095 0.037802 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.01 on 152 degrees of freedom
## Multiple R-squared:  0.9195, Adjusted R-squared:  0.9169
## F-statistic: 347.3 on 5 and 152 DF,  p-value: < 2.2e-16
```

Full Model

```r
anova(fit0)
```

```
## Analysis of Variance Table
##
## Response: crew
##            Df  Sum Sq Mean Sq   F value    Pr(>F)
## age         1  542.66  542.66  531.7490 < 2.2e-16 ***
## tonnage     1 1118.50 1118.50 1096.0157 < 2.2e-16 ***
## length      1   19.71   19.71   19.3130 2.072e-05 ***
## cabins      1   86.61   86.61   84.8700 2.430e-16 ***
## passdens    1    4.48    4.48    4.3903    0.0378 *
## Residuals 152  155.12    1.02
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
drop1(fit0,test="F")
```

```
## Single term deletions
##
## Model:
## crew ~ age + tonnage + length + cabins + passdens
##          Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                155.12  9.092
## age       1     0.146 155.26  7.241  0.1432 0.7056111
## tonnage   1     0.347 155.47  7.446  0.3403 0.5605252
## length    1    12.953 168.07 19.764 12.6924 0.0004907 ***
## cabins    1    71.806 226.93 67.200 70.3621 3.146e-14 ***
## passdens  1     4.480 159.60 11.591  4.3903 0.0378024 *
## ---
```

Pval for testing single variable significance

Backward Elimination - Model Based AIC (minimize)

```r
library(MASS)
fit1 <- lm(crew ~ age + tonnage + length + cabins + passdens)
stepAIC(fit1,direction="backward")
```

```
## Start:  AIC=9.09
## crew ~ age + tonnage + length + cabins + passdens
##
##            Df Sum of Sq    RSS    AIC
## - age       1     0.146 155.26  7.241
## - tonnage   1     0.347 155.47  7.446
## <none>                  155.12  9.092
## - passdens  1     4.480 159.60 11.591
## - length    1    12.953 168.07 19.764
## - cabins    1    71.806 226.93 67.200
##
## Step:  AIC=7.24
## crew ~ tonnage + length + cabins + passdens
##
##            Df Sum of Sq    RSS    AIC
## - tonnage   1     0.276 155.54  5.521
## <none>                  155.26  7.241
## - passdens  1     5.397 160.66 10.640
## - length    1    12.864 168.13 17.817
## - cabins    1    71.803 227.07 65.299
##
## Step:  AIC=5.52
## crew ~ length + cabins + passdens
##
```

Forward Selection - Model Based AIC (minimize)

```r
fit2 <- lm(crew ~ 1)
stepAIC(fit2,direction="forward",scope=list(upper=fit1,lower=fit2))
```

```
## Start:  AIC=397.18
## crew ~ 1
##
##            Df Sum of Sq     RSS    AIC
## + cabins    1   1742.21  184.88  28.82
## + tonnage   1   1658.03  269.05  88.10
## + length    1   1546.60  380.49 142.86
## + age       1    542.66 1384.42 346.93
## + passdens  1     46.60 1880.48 395.32
## <none>                  1927.08 397.18
##
## Step:  AIC=28.82
## crew ~ cabins
##
##            Df Sum of Sq    RSS     AIC
## + length    1   22.9636 161.91  9.8661
## + passdens  1   14.9541 169.92 17.4948
## + tonnage   1   12.5135 172.36 19.7480
## + age       1    5.4442 179.43 26.0989
## <none>                  184.88 28.8215
##
## Step:  AIC=9.87
## crew ~ cabins + length
##
##            Df Sum of Sq    RSS     AIC
## + passdens  1    6.3732 155.54  5.5212
## <none>                  161.91  9.8661
```

Forward Selection - Model Based AIC (minimize)

```
\begin{verbatim}
Step:  AIC=5.52
crew ~ cabins + length + passdens

          Df Sum of Sq    RSS    AIC
<none>                 155.54 5.5212
+ tonnage  1  0.275559 155.26 7.2410
+ age      1  0.074462 155.47 7.4455

Call:
lm(formula = crew ~ cabins + length + passdens)
\end{vernbatim}
```

Stepwise Regression (AIC Based)

```r
stepAIC(fit2,direction="both",scope=list(upper=fit1,lower=fit2))
```

```
## Start:  AIC=397.18
## crew ~ 1
##
##            Df Sum of Sq     RSS    AIC
## + cabins    1   1742.21  184.88  28.82
## + tonnage   1   1658.03  269.05  88.10
## + length    1   1546.60  380.49 142.86
## + age       1    542.66 1384.42 346.93
## + passdens  1     46.60 1880.48 395.32
## <none>                  1927.08 397.18
##
## Step:  AIC=28.82
## crew ~ cabins
##
##            Df Sum of Sq     RSS    AIC
## + length    1     22.96  161.91   9.87
## + passdens  1     14.95  169.92  17.49
## + tonnage   1     12.51  172.36  19.75
## + age       1      5.44  179.43  26.10
## <none>                   184.88  28.82
## - cabins    1   1742.21 1927.08 397.18
##
## Step:  AIC=9.87
## crew ~ cabins + length
##
##            Df Sum of Sq    RSS    AIC
## + passdens  1     6.373 155.54  5.521
## <none>                  161.91  9.866
```

Stepwise Regression (AIC Based)

```
\begin{verbatim}
Step:  AIC=5.52
crew ~ cabins + length + passdens

           Df Sum of Sq    RSS     AIC
<none>                  155.54   5.521
+ tonnage   1     0.276 155.26   7.241
+ age       1     0.074 155.47   7.446
- passdens  1     6.373 161.91   9.866
- length    1    14.383 169.92  17.495
- cabins    1   214.177 369.72 140.323

Call:
lm(formula = crew ~ cabins + length + passdens)

Coefficients:
(Intercept)       cabins       length     passdens
   -1.83730      0.61878      0.38835      0.02532
\end{verbatim}
```

Summary of Automated Model

- Backward Elimination
  - Drop Age (AIC drops from 9.09 to 7.24)
  - Drop Tonnage (AIC drops from 7.24 to 5.52)
  - Stop: Keep Passdens, Length, Cabins
- Forward Selection
  - Add Cabins (AIC drops from 397.18 to 28.82)
  - Add Length (AIC drops from 28.82 to 9.87)
  - Add Passdens (AIC drops from 9.87 to 5.52)
  - Stop: Keep Passdens, Length, Cabins
- Stepwise - Same as Forward Selection

All Possible (Subset) Regressions

```r
library(leaps)
allcruise <- regsubsets(crew ~ age + tonnage +  length + cabins + passdens,
 nbest=4,data=cruise)
aprout <- summary(allcruise)
n <- length(cruise$crew)
p <- apply(aprout$which, 1, sum)
aprout$aic <- aprout$bic - log(n) * p + 2 * p
with(aprout,round(cbind(which,rsq,adjr2,cp,bic,aic),3))
```

```
##   (Intercept) age tonnage length cabins passdens   rsq adjr2       cp      bic
## 1           1   0       0      0      1        0 0.904 0.903   27.160 -360.238
## 1           1   0       1      0      0        0 0.860 0.859  109.642 -300.954
## 1           1   0       0      1      0        0 0.803 0.801  218.835 -246.201
## 1           1   1       0      0      0        0 0.282 0.277 1202.589  -42.129
## 2           1   0       0      1      1        0 0.916 0.915    6.658 -376.131
## 2           1   0       0      0      1        1 0.912 0.911   14.507 -368.502
## 2           1   0       1      0      1        0 0.911 0.909   16.898 -366.249
## 2           1   1       0      0      1        0 0.907 0.906   23.826 -359.898
## 3           1   0       0      1      1        1 0.919 0.918    2.413 -377.413   Best Model!
## 3           1   1       0      1      1        0 0.917 0.915    6.728 -373.002
## 3           1   0       1      1      1        0 0.917 0.915    7.432 -372.294
## 3           1   0       1      0      1        1 0.913 0.911   14.749 -365.117
## 4           1   0       1      1      1        1 0.919 0.917    4.143 -372.631
## 4           1   1       0      1      1        1 0.919 0.917    4.340 -372.426
## 4           1   1       1      1      1        0 0.917 0.915    8.390 -368.280
## 4           1   1       1      0      1        1 0.913 0.911   16.692 -360.108
## 5           1   1       1      1      1        1 0.920 0.917    6.000 -367.717
##        aic
## 1 -366.363
## 1 -307.080
```

9.6 Model Validation

- The general idea of model validation
- Confirmation that the model is sound and effective for the purpose for which it was intended.
- It requires to assess the effectiveness of the model against an independent set of data, called **validation data set**, and not against the data from which the model was built/fitted, called **model-building data set**.

Mean squared prediction error

MSE F

- The **mean squared prediction error (MSPR)** is the average squared difference between independent observations and predictions from the fitted model

$$MSPR = \sum_{i=1}^{n^*} \frac{(Y_i - \hat{Y}_i)^2}{n^*}$$

where,
1) Y_i is the value of the response variable for ith observation in the validation data set.
2) $\hat{Y}_i$ is the predicted value for the ith observation in the validation data set based on the Model fitted with the model building data set.
3) n^* is the number of cases in the validation data set

Criterion: MSPR $\approx$ MSE — from model building data set

Obtaining an independent data set

- Often impractical to obtain an adequate independent data set, via collection of new data, for instance.
- If the existing data set is sufficiently large, one approach consists of dividing the data set into two representative halves:

70% for training

15% for validation

15% for testing

Excise: Surgical unit example: Model validation

- We used the first 54 out of the 108 patients as model-building data set. The last 54 observations will be used as validation data set. We consider the model with X_1, X_2, and X_3 as predictor variables. We have

$$\underbrace{\sum_{i=1}^{54}(Y_i - \hat{Y}_i)^2 = 4.3877}_{\text{training}} \quad \text{and} \quad \underbrace{\sum_{i=1}^{54}(Y_i - \hat{Y}_i)^2 = 3.1085}_{\text{validation}}$$

for the validation data set, and for the model building data set, respectively. Compute the MSPR and MSE. Can we validate the model?

$$MSPR = \frac{4.3877}{54} \qquad MSE = \frac{3.1085}{54-4} = 0.062$$

$MSPR \approx MSE$, validation quiet good

Cross-Validataion

- Hold-out Sample (Training Sample = 100, Validation = 58)
  - Fit Model on Training Sample, and obtain Regression Estimates
  - Apply Regression Estimates from Training Sample to Validation Sample X levels for Predicted
  - Fit Model on Validation Sample and Compare regression coefficients with model for Training Sample

```
##### Cross-validation with a hold-out sample
##### Randomly sample 100 ships, fit model, obtain predictions for remaining 58 ships
#####    by applying their X-levels to regression coefficients from model

##### Obtain "training" and "validation" sets
set.seed(12345)
cruise.cv.samp <- sample(1:length(crew),100,replace=FALSE)
cruise.cv.in <- cruise[cruise.cv.samp,]
cruise.cv.out <- cruise[-cruise.cv.samp,]

##### Fit model for training set
fit.cv.in <- lm(crew ~ length + cabins + passdens,
   data=cruise.cv.in)
anova(fit.cv.in)
```

```
## Analysis of Variance Table
##
## Response: crew
```

Cross Validation

```
\begin{verbatim}
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.69471    0.85371  -1.985   0.0500 *
length       0.36720    0.15200   2.416   0.0176 *
cabins       0.63501    0.06392   9.934   <2e-16 ***
passdens     0.02487    0.01514   1.643   0.1037
---
\end{verbatim}
```

```
##### Obtain Predicted values and prediction errors for validation sample
##### Regression is based on same 3 predictors as fit3 (columns 6:8 of cruise)
##### Compute MSPR
pred.cv.out <- predict(fit.cv.in,cruise.cv.out[,6:8])
delta.cv.out <- crew[-cruise.cv.samp]-pred.cv.out
n.star = dim(cruise.cv.out)[1]
MSPR <- sum((delta.cv.out)^2)/n.star
MSPR
```

```
## [1] 1.350547
```

$\approx$ MSE = 0.8 ..

PRESS Statistic

$\sum_{i=1}^{n} (Y_i - \hat{Y}_{i(i)})^2 = \text{Press}$

Criterion $\approx$ SSE

```r
library(MPV)
fit.best = lm(crew ~ cabins + length + passdens)
anova(fit.best)
```

```
## Analysis of Variance Table
##
## Response: crew
##            Df  Sum Sq Mean Sq   F value    Pr(>F)
## cabins      1 1742.21 1742.21 1724.9508 < 2.2e-16 ***
## length      1   22.96   22.96   22.7362 4.272e-06 ***
## passdens    1    6.37    6.37    6.3101   0.01304 *
## Residuals 154  155.54    1.01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
PRESS(fit.best)
```

```
## [1] 162.4069
```