STAC67: Regression Analysis

Lecture 18

Sohee Kang

Mar. 18, 2021

Chapter 9

Building the Regression Model I: Model Selection and Validation

Introduction

- When we have many predictors, we may wish to use an algorithm to determine which variables to include in the model.
- This section answers the question: which variables should be included in the model?
- These variables can be main effects, interactions, and polynomial terms. Note that there are two common approaches.
  - One method involves testing variables based on t-tests, or equivalently F-tests for partial regression coefficients.
  - An alternative method involves comparing models based on model based measures, such as Akaike Information Criterion (AIC), or Schwartz Bayesian Information criterion (BIC or SBC).

Surgical unit example

- A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random selection of 108 patients was available for analysis. For each patient record, the following information was extracted from the preoperation evaluation:
  - X_1: blood clotting score
  - X_2: prognostic index
  - X_3: enzyme function test score
  - X_4: liver function test score
  - X_5: age, in years
  - X_6: indicator variable for gender (0 = male, 1 = female)
  - X_7 and X_8 indicator variables for history of alcohol use:

Alcohol use : {None, Moderate, Severe}
(None: baseline category)

Surgical unit example

- A portion of the data is shown below.

Case number i	Blood clotting score X_{i1}	Prognostic index X_{i2}	Enzyme function test score X_{i3}	Liver function test score X_{i4}	Age X_{i5}	Gender X_{i6}	Alcohol use moderate X_{i7}	Alcohol use severe X_{i8}	Survival time Y_i	ln Y_i
1	6.7	62	81	2.59	50	0	1	0	695	6.544
2	5.1	59	66	1.70	39	0	0	0	403	5.999
3	7.4	57	83	2.16	55	0	0	0	710	6.565
4	6.5	73	41	2.01	48	0	0	0	349	5.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
53	6.4	59	85	2.33	63	0	1	0	550	6.310
54	8.8	78	72	3.20	56	0	0	0	651	6.478

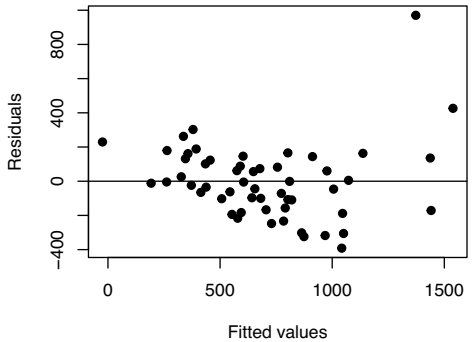
We will use only the first four variables and the first 54 out of the 108 patients (model-building set, see later).

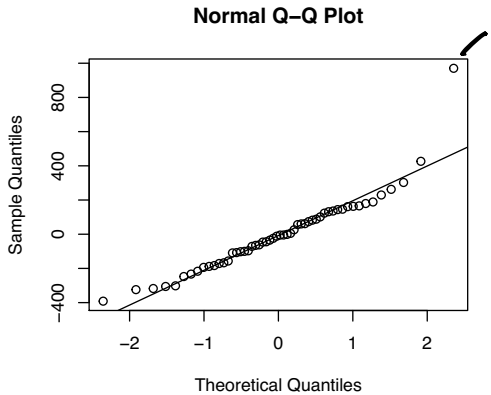
Variable Transformation

- We fit a model with Y as response variable and X_1, X_2, X_3, and X_4 as predictor variables. We obtain the following residual plot and normal qq-plot of the residuals

main effect model.

```
par(mfrow=c(2,2))
Surgic = read.table("Table9-1.txt", header=T)
fit = lm(Y ~ X1 + X2 +X3 +X4, data=Surgic)
resid = fit$residuals
fit.Y = predict(fit)
plot(fit.Y, resid, pch=20, cex=1.5, xlab="Fitted values", ylab="Residuals")
abline(0,0)
qqnorm(resid)
qqline(resid)
```



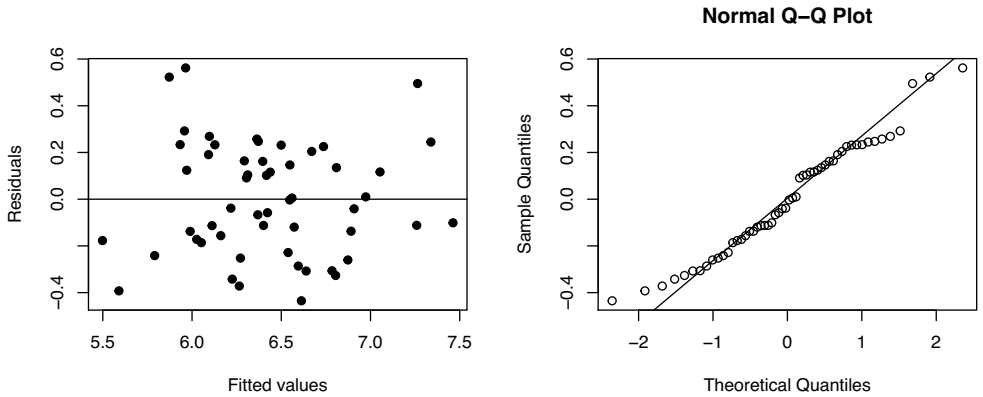


outlied

Variable Transformation

- We decide to transform the response variable. We fit a model with lnY as response variable and X_1, X_2, X_3, and X_4 as predictor variables. We obtain the following residual plot and normal qq-plot of the residuals.

```
par(mfrow=c(2,2))
fit2 = lm(lnY~ X1 + X2 +X3 +X4, data=Surgic)
resid2 = fit2$residuals
fit.Y2 = predict(fit2)
plot(fit.Y2, resid2, pch=20, cex=1.5, xlab="Fitted values", ylab="Residuals")
abline(0,0)
qqnorm(resid2)
qqline(resid2)
```



Two models

```r
Model.1 = lm(lnY~ X1 + X2 +X3 +X4, data=Surgic)
summary(Model.1)
```

```
##
## Call:
## lm(formula = lnY ~ X1 + X2 + X3 + X4, data = Surgic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43500 -0.17591 -0.02091  0.18400  0.56192
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.851948   0.266258  14.467  < 2e-16 ***
## X1          0.083684   0.028833   2.902  0.00554 **
## X2          0.012665   0.002315   5.471 1.51e-06 ***
## X3          0.015632   0.002100   7.443 1.37e-09 ***
## X4          0.032161   0.051465   0.625  0.53493
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2509 on 49 degrees of freedom
## Multiple R-squared:  0.7592, Adjusted R-squared:  0.7396
## F-statistic: 38.62 on 4 and 49 DF,  p-value: 1.388e-14
```

Two models

```r
Model.2 = lm(lnY~ X1 + X3, data=Surgic)
summary(Model.2)
```

```
##
## Call:
## lm(formula = lnY ~ X1 + X3, data = Surgic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06195 -0.21620  0.01228  0.24610  0.87470
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.546228   0.260654  17.442  < 2e-16 ***
## X1          0.107917   0.029178   3.699 0.000531 ***
## X3          0.016342   0.002201   7.426 1.16e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3367 on 51 degrees of freedom
## Multiple R-squared:  0.5486, Adjusted R-squared:  0.5309
## F-statistic:    31 on 2 and 51 DF,  p-value: 1.55e-09
```

Which Model?

- How much emphasis should we place on eliminating variables? It depends on the objective.
- If the objective is to **describe the behavior** of the response variable, eliminating variables in not so important. The best description of the response (in terms of minimum residual sum of squares) is provided by the full model.
- Eliminating the variables is more important when the objective is different, such as **prediction** and **estimation of the mean response**. A model with fewer variables is appealing in terms of simplicity. Moreover, there is are theoretical advantages of eliminating irrelevant variables (discussed later). As far as how many variables we should keep in our model, we want to balance between loss of predictability and theoretical advantages/simplicity.

Model selection and validation

- Model Selection The investigator starts with a usually large set of variables and reduces the number of variables to include in the model. He narrows the number of competing models down to one or a few models.
- Model validation The investigator confirms that the model is sound and effective for the purpose for which it was intended. Model validation requires to assess the effectiveness of the fitted model on an independent set of data (training set).

Model selection procedures

- From any set of p predictor variables, there are
  possible alternative models (each predictor variable can be either included or excluded from the model).
- Some model selection procedures to identify a model (or small group of models) that are good according to a criterion:
  - R_p^2 Criterion
  - $R_{p,adj}^2$ Criterion
  - Mallows' C_p criterion
  - Akaike's Information Criterion (AIC_p) or BIC_p
  - Forward stepwise selection
  - Forward selection
  - Backward elimination

R_p^2 Criterion

- The coefficient of determination:

$$R_p^2 = \frac{SSR_p}{SST} = 1 - \frac{SSE_p}{SST}$$

, where SSR_p is the regression sum of squares from the p predictor variable subset model being considered, is the proportion of the variance in the response variable that is explained by the model.

- As we add predictor variables to the model, the coefficient of determination R_p^2 (increases/decreases/remains the same).
- Therefore, the model that gives the maximum R_p^2 will necessarily be the model that contains (more/less) predictor variables.
- **Criterion**: find the point where adding more predictor variables is not worthwhile because it leads to a very small increase in R_p^2 .

$R^2_{p,adj}$ Criterion

- The adjusted coefficient of determination

$$R^2_{p,adj} = 1 - \frac{(1-R^2_p)(n-1)}{n-p'} = 1 - \frac{n-1}{n-p'}\left(\frac{SSE_p}{SST}\right)$$

gives a quantity that is more comparable than R^2_p over models involving different numbers of parameters.

- unlike R^2_p, $R^2_{p,adj}$ does not always increase as predictor vars are added to the model.

- **Criterion**: The simplist model with R^2_{adj} close to upper limit.

Mallows' C_p criterion

- The C_p statistic is an estimate of the standardized total mean squared error of estimation for the current set of data and is defined as:

$$C_p = \frac{SSE_p}{MSE_F} + 2p' - n$$

- When the model is correct, the residual sum of squares is an unbiased estimate of $(n-p')\sigma^2$ and $C_p \approx p'$.
- When important predictor variables are omitted from model, C_p is expected to be greater than p'
- **Criterion** look for model with small $C_p \approx p'$.

Akaike's Information Criterion (AIC_p)

- Similarly to $R^2_{p,adj}$ and C_p, AIC_p criterion is a model selection criterion that penalizes models having a large number of predictor variables.
- It is defined as

$$AIC_p = n \cdot \ln(SSE_p) - \ln(n) + 2p'$$

- The first term (increases/decreases) as the number of predictor variables p increases. The second term is fixed, and the third term (increases/decreases) as p increases.
- **Criterion**: Select model with smallest AIC_p.

- Bayesian Information Criterion (BIC) Or SBC

$$BIC_p = n \ln(SSE_p) - n \ln(n) + \ln(n) p'$$

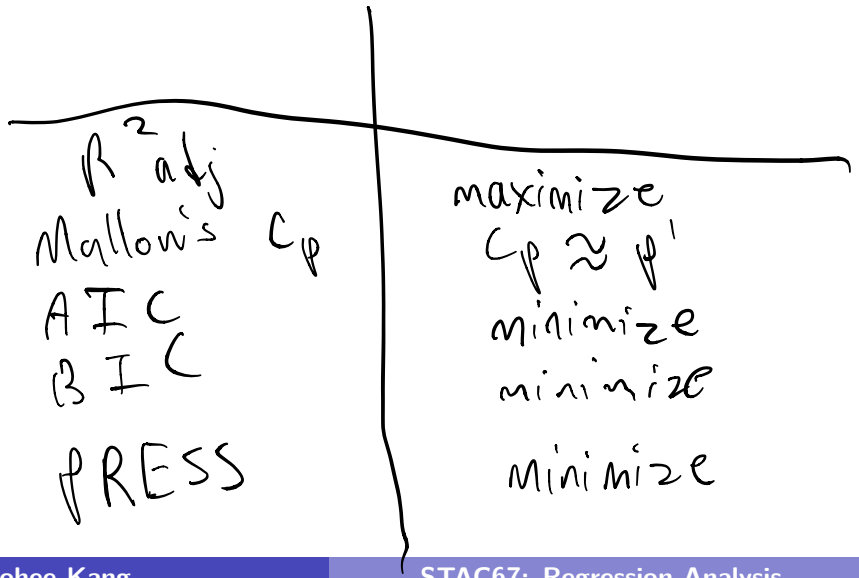
Same criterion as AIC_p

$PRESS_p$ Creterion (Jack Knife)

$$PRESS_p = \sum (Y_i - \hat{Y}_{i(i)})^2$$

$\hat{Y}_{i(i)}$ = Fitted value for ith case is deleted

- **Criterion**:
  Smallest $PRESS_p$ value.



Surgical unit example

Exercise 1 Compute the four criteria R^2, R_a^2 , Mallows' C, and AIC for the following two models

- Full model, i.e. model with the four predictor variables (X_1 to X_4), and
- Reduced model with two auxiliary variables (X_1 and X_3), Which model is preferred for each criterion?

Hint: We have SST $= 12.808$. Moreover, we have SSE $= 3.084$ for the full model and SSE $= 5.781$ for the reduced model.

Surgical unit example: criteria for all the possible models

Indep. var.	$SSres$	R^2	R^2_{adj}	C	AIC
None	12.81	0.00	0.00	151.50	-75.70
X_1	12.03	0.06	0.04	141.16	-77.08
X_2	9.98	0.22	0.21	108.56	-87.18
X_3	7.33	0.43	0.42	66.49	-103.83
X_4	7.41	0.42	0.41	67.71	-103.26
X_1, X_2	9.44	0.26	0.23	102.03	-88.16
X_1, X_3	5.78	0.55	0.53	43.85	-114.66
X_1, X_4	7.30	0.43	0.41	67.97	-102.07
X_2, X_3	4.31	0.66	0.65	20.52	-130.48
X_2, X_4	6.62	0.48	0.46	57.21	-107.32
X_3, X_4	5.13	0.60	0.58	33.50	-121.11
X_1, X_2, X_3	3.11	0.76	0.74	3.39	-146.16
X_1, X_2, X_4	6.57	0.49	0.46	58.39	-105.75
X_1, X_3, X_4	4.97	0.61	0.59	32.93	-120.84
X_2, X_3, X_4	3.61	0.72	0.70	11.42	-138.02
X_1, X_2, X_3, X_4	3.08	0.76	0.74	5.00	-144.59

R codes

- The R package **leaps** will perform all possible regression. It does not include AIC, however it can be computed from BIC.

```
library(leaps)
allreg <- regsubsets(lnY~ X1 + X2 +X3 + X4, nbest=4, data=Surgic)
n=dim(Surgic)[1]
aprout=summary(allreg)
pprime = apply(aprout$which, 1, sum)
aprout$aic <- aprout$bic - log(n) * pprime + 2 * pprime
with(aprout,round(cbind(which,rsq,adjr2,cp,bic, aic),3))
```

```
##   (Intercept) X1 X2 X3 X4   rsq adjr2      cp     bic     aic
## 1           1  0  0  1  0 0.428 0.417  66.489 -22.146 -26.124
## 1           1  0  0  0  1 0.422 0.410  67.715 -21.581 -25.559
## 1           1  0  1  0  0 0.221 0.206 108.556  -5.498  -9.476
## 1           1  1  0  0  0 0.061 0.043 141.164   4.602   0.624
## 2           1  0  1  1  0 0.663 0.650  20.520 -46.814 -52.781
## 2           1  0  0  1  1 0.599 0.584  33.504 -37.443 -43.410
## 2           1  1  0  1  0 0.549 0.531  43.852 -30.989 -36.956
## 2           1  0  1  0  1 0.483 0.463  57.215 -23.654 -29.621
## 3           1  1  1  1  0 0.757 0.743   3.391 -60.502 -68.458
## 3           1  0  1  1  1 0.718 0.701  11.424 -52.365 -60.321
## 3           1  1  0  1  1 0.612 0.589  32.932 -35.186 -43.142
## 3           1  1  1  0  1 0.487 0.456  58.392 -20.089 -28.045
## 4           1  1  1  1  1 0.759 0.740   5.000 -56.942 -66.887
```

Regression Model Building

- Setting: Possibly a large set of predictor variables (including interactions).
- Goal: Fit a parsimonious model that explains variation in Y with a small set of predictors
- Automated Procedures and all possible regressions:
  - Backward Elimination (Top down approach)
  - Forward Selection (Bottom up approach)
  - Stepwise Regression (Combines Forward/Backward)

Backward Elimination (Traditional approach)

This is a "top-down" method, which begins with a "Complete" Model, with all potential predictors.

Step 1: Select a significance level to stay (SLS) in the model (e.g. $\alpha = 0.20$, generally .05 is too low, causing too many variables to be removed)

Step 2: Fit the full model with all possible predictors. Consider the predictor with lowest t-statistic (highest P-value).

- If $P - value > \alpha$, remove the predictor and fit model without this variable (must re-fit model here because partial regression coefficients change)
- If $P - value \leq \alpha$ stop and keep current model

Step 3: Continue until all predictors have P-values below α

Note: R uses model based criteria: AIC instead.

Forward selction (Traditional approach)

- This is a "bottom-up" method, which begins with all "Simple" Models, each with one predictor.

Step 1: Choose a significance level to enter (SLE) the model (e.g. SLE=0.20, generally .05 is too low, causing too few variables to be entered)

Step 2: Fit all simple regression models. Consider the predictor with the highest t-statistic (lowest P-value)

- If $P - value \leq SLE$, keep this variable and fit all two variable models that include this predictor
- If $P - value > SLE$, stop and keep previous model

Step 3: Continue until no new predictors have $P - value \leq SLE$

Note: R uses model based criteria: AIC, SBC instead

Stepwise Regression - Traditional Approach

Step 1: Select SLS and SLE (SLE<SLS)

Step 2: Starts like Forward Selection (Bottom up process)

- New variables must have $P \leq SLE$ to enter
- Re-tests all "old variables" that have already been entered, must have $P \leq SLS$ to stay in model

Step 3: Continues until no new variables can be entered and no old variables need to be removed

Note: R uses model based criteria: AIC, SBC instead

Surgical unit example: stepwise selection

- Function stepAIC( ) from the MASS package performs stepwise selection (forward, backward, both).

```
library(MASS)
fit_select = lm(lnY ~ X1 + X2 + X3 + X4, data = Surgic)
step = stepAIC(fit_select, direction = "both"); step$anova # display results
```

direction : Forward
Backward
Both

```
## Start:  AIC=-144.59
## lnY ~ X1 + X2 + X3 + X4
##
##        Df Sum of Sq    RSS     AIC
## - X4    1    0.0246 3.1085 -146.16
## <none>               3.0840 -144.59
## - X1    1    0.5302 3.6141 -138.02
## - X2    1    1.8839 4.9678 -120.84
## - X3    1    3.4862 6.5702 -105.75
##
## Step:  AIC=-146.16
## lnY ~ X1 + X2 + X3
##
##        Df Sum of Sq    RSS      AIC
## <none>              3.1085 -146.161
## + X4    1    0.0246 3.0840 -144.590
## - X1    1    1.2040 4.3125 -130.483
## - X2    1    2.6724 5.7810 -114.658
## - X3    1    6.3341 9.4427  -88.162
```

Case study: Predicting Number of Crew Members of Cruise Ships

- Data Description: n=158 Cruise Ships
  - Dependent Variable - Crew Size (100s)
  - Potential Predictor Variables
    - Age (2013 - Year Built)
    - Tonnage (1000s of Tons)
    - Passengers (100s)
    - Length (100s of feet)
    - Cabins (100s)
    - Passenger Density (Passengers/Space)

Data - First 20 cases

Ship	Cruise Line	Age	Tonnage	Pssngrs	Length	Cabins	PassDens	Crew
Journey	Azamara	6	30.277	6.94	5.94	3.55	42.64	3.55
Quest	Azamara	6	30.277	6.94	5.94	3.55	42.64	3.55
Celebration	Carnival	26	47.262	14.86	7.22	7.43	31.8	6.7
Conquest	Carnival	11	110	29.74	9.53	14.88	36.99	19.1
Destiny	Carnival	17	101.353	26.42	8.92	13.21	38.36	10
Ecstasy	Carnival	22	70.367	20.52	8.55	10.2	34.29	9.2
Elation	Carnival	15	70.367	20.52	8.55	10.2	34.29	9.2
Fantasy	Carnival	23	70.367	20.56	8.55	10.22	34.23	9.2
Fascination	Carnival	19	70.367	20.52	8.55	10.2	34.29	9.2
Freedom	Carnival	6	110.239	37	9.51	14.87	29.79	11.5
Glory	Carnival	10	110	29.74	9.51	14.87	36.99	11.6
Holiday	Carnival	28	46.052	14.52	7.27	7.26	31.72	6.6
Imagination	Carnival	18	70.367	20.52	8.55	10.2	34.29	9.2
Inspiration	Carnival	17	70.367	20.52	8.55	10.2	34.29	9.2
Legend	Carnival	11	86	21.24	9.63	10.62	40.49	9.3
Liberty*	Carnival	8	110	29.74	9.51	14.87	36.99	11.6
Miracle	Carnival	9	88.5	21.24	9.63	10.62	41.67	10.3
Paradise	Carnival	15	70.367	20.52	8.55	10.2	34.29	9.2
Pride	Carnival	12	88.5	21.24	9.63	11.62	41.67	9.3
Sensation	Carnival	20	70.367	20.52	8.55	10.2	34.29	9.2

R codes:

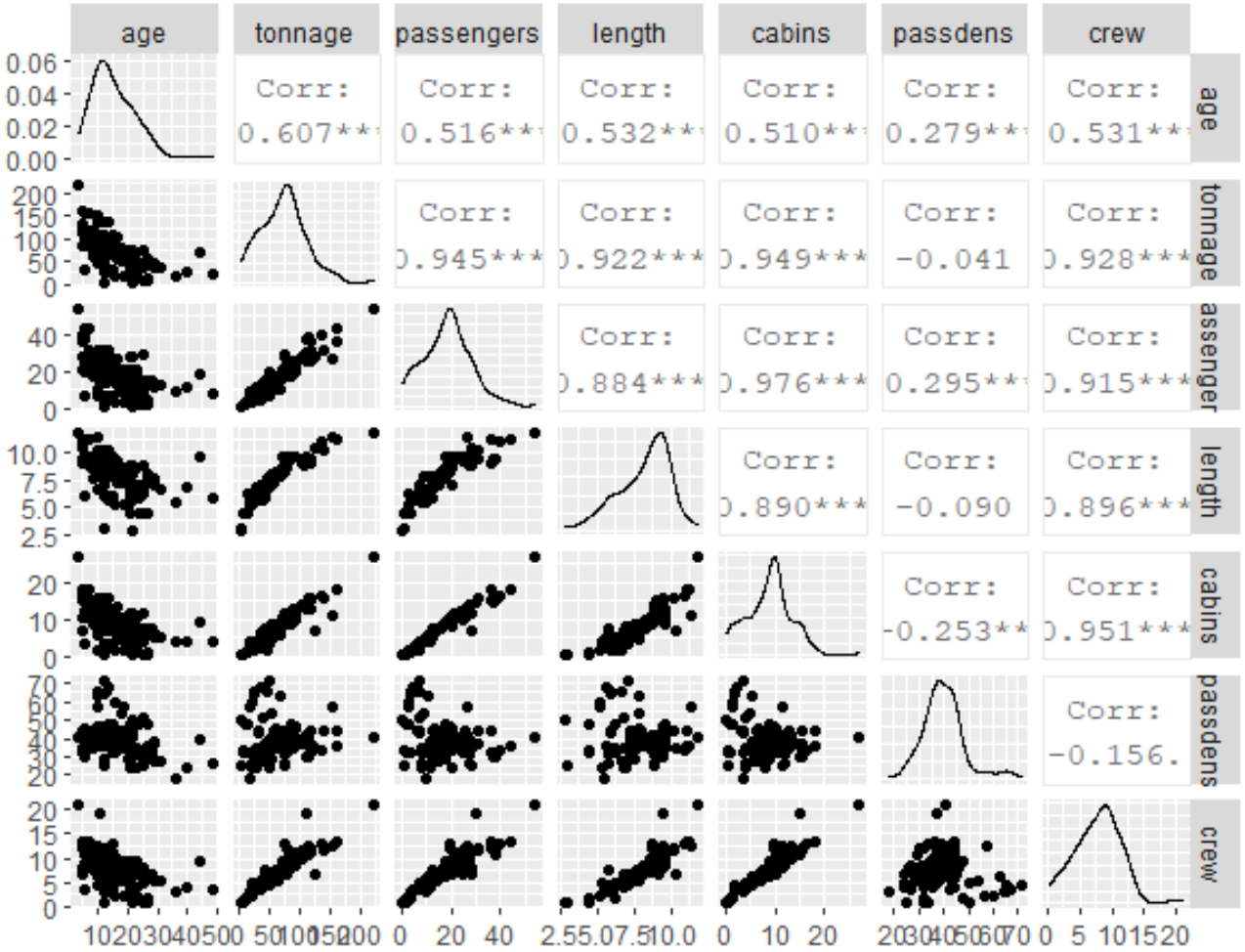
```
cruise = read.table("cruise_ship.txt", col.names=c("ship", "cline", "age", "tonnage", "passengers", "length", "
attach(cruise)

cor(cruise[,-c(1,2)])
```

```
##                  age     tonnage passengers      length     cabins    passdens
## age        1.0000000 -0.60664609 -0.5155423 -0.53228589 -0.5100190 -0.27883020
## tonnage   -0.6066461  1.00000000  0.9450614  0.92236832  0.9487636 -0.04084624
## passengers -0.5155423  0.94506140  1.0000000  0.88353479  0.9763414 -0.29486708
## length    -0.5322859  0.92236832  0.8835348  1.00000000  0.8897982 -0.09048847
## cabins    -0.5100190  0.94876357  0.9763414  0.88979821  1.0000000 -0.25318074
## passdens  -0.2788302 -0.04084624 -0.2948671 -0.09048847 -0.2531807  1.00000000
## crew      -0.5306565  0.92756881  0.9152341  0.89585663  0.9508226 -0.15550928
##                 crew
## age        -0.5306565
## tonnage     0.9275688
## passengers  0.9152341
## length      0.8958566
## cabins      0.9508226
## passdens   -0.1555093
## crew        1.0000000
```

```
#library(GGally)
#library(ggplot2)
#ggpairs(cruise[,-c(1,2)])
```

Correlation Matrix



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

- Consider model with Predictors: Age, Tonnage, Passdens, Cabins, Length

```
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
```