Model Selection and Validation

Zhenisbek Assylbekov

Department of Mathematics

Regression Analysis

Salary example

https://github.com/zh3nis/MATH440/tree/main/chp09/salary.R

Model annual salary (in $1000) as function of

Salary example

https://github.com/zh3nis/MATH440/tree/main/chp09/salary.R

Model annual salary (in $1000) as function of

- age (in years),

Salary example

https://github.com/zh3nis/MATH440/tree/main/chp09/salary.R

Model annual salary (in $1000) as function of

- `age` (in years),
- `education` (years of post-high-school education), and

Salary example

https://github.com/zh3nis/MATH440/tree/main/chp09/salary.R

Model annual salary (in $1000) as function of

- `age` (in years),
- `education` (years of post-high-school education), and
- `political affiliation` (`pol` = D for Democrat, `pol` = R for Republican, and `pol` = O for other).

Salary example

https://github.com/zh3nis/MATH440/tree/main/chp09/salary.R

Model annual salary (in $1000) as function of

- age (in years),
- education (years of post-high-school education), and
- political affiliation (pol = D for Democrat, pol = R for Republican, and pol = O for other).

```
> salary_data = read.table("path/to/salary.txt", header=FALSE)
> colnames(salary_data) = c('salary', 'age', 'educ', 'pol')
> head(salary_data)
  salary age educ pol
1     38  25    4   D
2     45  27    4   R
3     28  26    4   O
4     55  39    4   D
5     74  42    4   R
6     43  41    4   O
```

Salary example

https://github.com/zh3nis/MATH440/tree/main/chp09/salary.R

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.0313     7.3459   2.318  0.03735 *
age           0.8983     0.1968   4.565  0.00053 ***
educ          1.5039     1.1841   1.270  0.22632
polO        -16.5404     4.8807  -3.389  0.00484 **
polR          9.1587     4.8482   1.889  0.08139 .
---

Residual standard error: 8.209 on 13 degrees of freedom
Multiple R-squared:  0.8374,Adjusted R-squared:  0.7873
```

- We can also test quadratic effects and interactions.
- From the initial fit, `educ` is not needed with `age` and `pol` in the model. Let's refit:

Droc educ?

```
> m2 = update(m1, . ~ . - educ)
> anova(m1, m2)
Analysis of Variance Table

Model 1: salary ~ age + educ + pol
Model 2: salary ~ age + pol
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     13 876.03
2     14 984.72 -1    -108.7 1.6131 0.2263

> summary(m2)
lm(formula = salary ~ age + pol, data = salary_data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.5172     6.5806   3.270  0.00559 **
age           1.0345     0.1686   6.136 2.58e-05 ***
polO        -16.7414     4.9838  -3.359  0.00468 **
polR          8.6379     4.9354   1.750  0.10196
Residual standard error: 8.387 on 14 degrees of freedom
Multiple R-squared:  0.8172,Adjusted R-squared:  0.778
```

Add 2nd order terms?

```
> salary_data$age2 = salary_data$age^2
> m3 = update(m2, . ~ . + age2 + age*pol)
> summary(m3)

Call:
lm(formula = salary ~ age + pol + age2 + age:pol, data = salary_data)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.94355   20.34047  -1.030  0.32528
age           3.17751    0.93793   3.388  0.00606 **
polO        -16.90846   22.83050  -0.741  0.47444
polR         -1.18699   21.44129  -0.055  0.95684
age2         -0.02514    0.01255  -2.004  0.07037 .
age:polO      0.05101    0.65536   0.078  0.93936
age:polR      0.28944    0.61956   0.467  0.64950
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.434 on 11 degrees of freedom
Multiple R-squared:  0.8871,Adjusted R-squared:  0.8256
```

Drop 2nd order terms?

```
> anova(m3, m2)
Analysis of Variance Table

Model 1: salary ~ age + pol + age2 + age:pol
Model 2: salary ~ age + pol
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     11 607.88
2     14 984.72 -3   -376.84 2.2731 0.1369
```

Drop 2nd order terms?

```
> anova(m3, m2)
Analysis of Variance Table

Model 1: salary ~ age + pol + age2 + age:pol
Model 2: salary ~ age + pol
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     11 607.88
2     14 984.72 -3   -376.84 2.2731 0.1369
```

We don't need *all* the 2nd order terms ($p = 0.137$), although theres some indication in the table of regression effects that age^2 might be needed.

Drop 2nd order terms?

```
> anova(m3, m2)
Analysis of Variance Table

Model 1: salary ~ age + pol + age2 + age:pol
Model 2: salary ~ age + pol
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     11 607.88
2     14 984.72 -3   -376.84 2.2731 0.1369
```

We don't need *all* the 2nd order terms ($p = 0.137$), although theres some indication in the table of regression effects that age^2 might be needed.

```
> anova(m4, m2)
Analysis of Variance Table

Model 1: salary ~ age + pol + age2
Model 2: salary ~ age + pol
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     13 645.09
2     14 984.72 -1   -339.64 6.8444 0.02134 *
```

Final model

Our final model is

$$\text{salary}_i = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \mathbb{I}[\text{pol} = \text{O}] + \beta_3 \cdot \mathbb{I}[\text{pol} = \text{R}] + \beta_4 \cdot \text{age}^2 + \epsilon_i$$

```
> summary(m4)

Call:
lm(formula = salary ~ age + pol + age2, data = salary_data)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.751745  18.529243  -1.336  0.20452
age           3.272388   0.867049   3.774  0.00232 **
polO        -15.891696   4.198662  -3.785  0.00227 **
polR          9.260234   4.152253   2.230  0.04399 *
age2         -0.024576   0.009394  -2.616  0.02134 *
---

Residual standard error: 7.044 on 13 degrees of freedom
Multiple R-squared:  0.8802,Adjusted R-squared:  0.8434
```

Scatterplots

- Scatterplots show the *marginal* relationship between Y and each of the $x_1, \ldots, x_k$.

Scatterplots

- Scatterplots show the *marginal* relationship between Y and each of the $x_1, \ldots, x_k$. They *cannot* show you anything about the joint relationship among the $Y, x_1, \ldots, x_k$.

Scatterplots

- Scatterplots show the *marginal* relationship between Y and each of the $x_1, \dots, x_k$. They *cannot* show you anything about the joint relationship among the $Y, x_1, \dots, x_k$.
- Nonlinear relationship between Y and x_j $(j = 1, \dots, k)$ *marginally* may or may not be present in the joint relationship.

Scatterplots

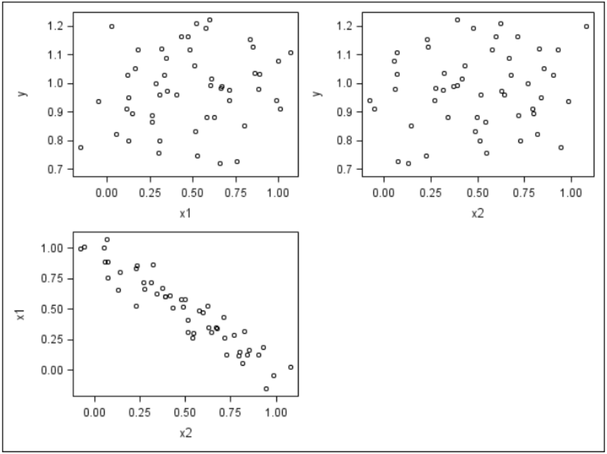
- Scatterplots show the *marginal* relationship between Y and each of the $x_1, \ldots, x_k$. They *cannot* show you anything about the joint relationship among the $Y, x_1, \ldots, x_k$.
- Nonlinear relationship between Y and x_j ($j = 1, \ldots, k$) *marginally* may or may not be present in the joint relationship.
- Actually, any strong relationship between Y and x_j marginally doesn't mean that x_j will be needed in the presence of other variables.

Scatterplots

- Scatterplots show the *marginal* relationship between Y and each of the $x_1, \ldots, x_k$. They *cannot* show you anything about the joint relationship among the $Y, x_1, \ldots, x_k$.
- Nonlinear relationship between Y and x_j ($j = 1, \ldots, k$) *marginally* may or may not be present in the joint relationship.
- Actually, any strong relationship between Y and x_j marginally doesn't mean that x_j will be needed in the presence of other variables.
- Seeing no marginal relationship between Y and x_j does not mean that x_j is not needed in a model including other predictors.

No relationship?

Here Y vs. x_1 and Y vs. x_2 shows nothing. There seems to be some multicollinearity though.



proc reg output

x_1 important marginally? $Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.94576	0.03602	26.26	<.0001
x1	1	0.06974	0.06311	1.11	0.2745

x_2 important marginally? $Y_i = \beta_0 + \beta_2 x_{i2} + \epsilon_i$

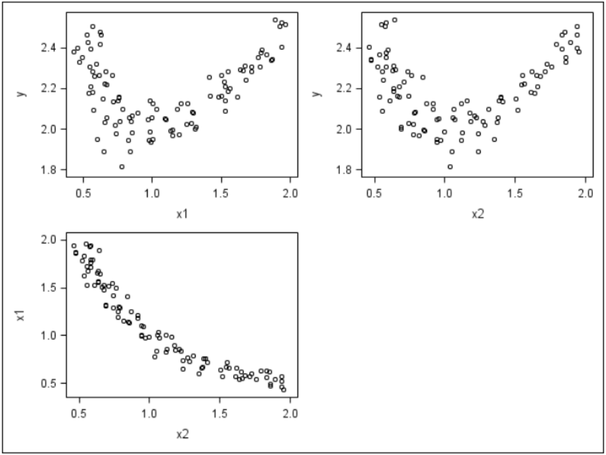
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.95180	0.03730	25.52	<.0001
x2	1	0.05603	0.06458	0.87	0.3898

x_1, x_2 important jointly? $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.08151	0.10876	-0.75	0.4572
x1	1	1.07327	0.11065	9.70	<.0001
x2	1	1.08548	0.11271	9.63	<.0001

Nonlinear relationship?

Marginally, x_1 and x_2 have highly nonlinear relationships with Y. Should we transform?



Regression output

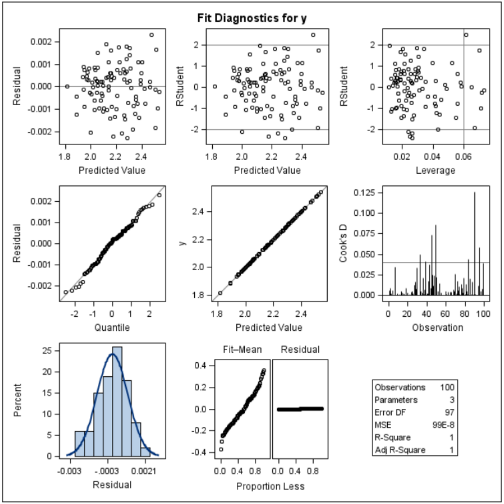
Let's try fitting a simple main effects model without any transformation.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.00036626	0.00130	-0.28	0.7791
x1	1	1.00022	0.00059936	1668.80	<.0001
x2	1	1.00009	0.00060998	1639.54	<.0001

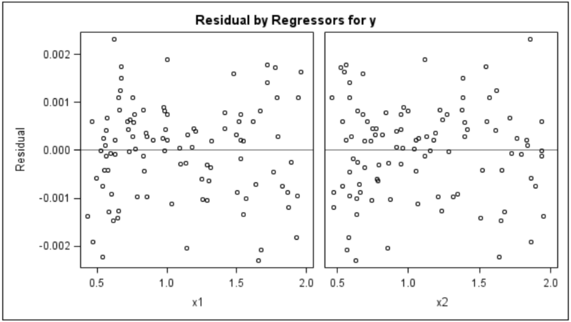
both x_1 and x_2 are important, but does the model fit okay?

Model fit is okay



Look at Y_i vs. $\hat{Y}_i$ and R^2!

No pattern here, either



9.1 Model building overview (pp. 343–349)

Book outlines four steps in data analysis:

1. Data collection and preparation

9.1 Model building overview (pp. 343–349)

Book outlines four steps in data analysis:

1. Data collection and preparation
2. Reduction of explanatory variables. Mass screening for "good" predictors.

9.1 Model building overview (pp. 343–349)

Book outlines four steps in data analysis:

1. Data collection and preparation
2. Reduction of explanatory variables. Mass screening for "good" predictors.
3. Model refinement and selection.

9.1 Model building overview (pp. 343–349)

Book outlines four steps in data analysis:

1. Data collection and preparation
2. Reduction of explanatory variables. Mass screening for "good" predictors.
3. Model refinement and selection.
4. Model validation.

9.1 Model building overview (pp. 343–349)

Book outlines four steps in data analysis:

1. Data collection and preparation
2. Reduction of explanatory variables. Mass screening for "good" predictors.
3. Model refinement and selection.
4. Model validation.

The way the data was collected may affect steps 2 & 3.

9.1 Model building overview (pp. 343–349)

Book outlines four steps in data analysis:

1. Data collection and preparation
2. Reduction of explanatory variables. Mass screening for "good" predictors.
3. Model refinement and selection.
4. Model validation.

The way the data was collected may affect steps 2 & 3.

Model validation should not be confused with model diagnostics (residual analysis).

Data collection strategies

- **Controlled experiments**: subjects (experimental units) assigned to x-levels by experimenter

Data collection strategies

- **Controlled experiments**: subjects (experimental units) assigned to x-levels by experimenter
  - **Purely controlled experiments**: researcher uses only predictors that were assigned to units

Data collection strategies

- **Controlled experiments**: subjects (experimental units) assigned to x-levels by experimenter
  - **Purely controlled experiments**: researcher uses only predictors that were assigned to units
  - **Controlled experiments with covariates** (uncontrolled variables): researcher has additional predictors associated with units

Data collection strategies

- **Controlled experiments**: subjects (experimental units) assigned to x-levels by experimenter
  - **Purely controlled experiments**: researcher uses only predictors that were assigned to units
  - **Controlled experiments with covariates** (uncontrolled variables): researcher has additional predictors associated with units
- **Observational studies**: subjects have x-levels associated with them (not assigned by researcher)

Data collection strategies

- **Controlled experiments**: subjects (experimental units) assigned to x-levels by experimenter
  - **Purely controlled experiments**: researcher uses only predictors that were assigned to units
  - **Controlled experiments with covariates** (uncontrolled variables): researcher has additional predictors associated with units
- **Observational studies**: subjects have x-levels associated with them (not assigned by researcher)
  - **Confirmatory studies**: it is *hypothesized* that new (primary) predictors are associated with Y; while there are predictors, *known* to be associated with Y (called risk factors).

Data collection strategies

- **Controlled experiments**: subjects (experimental units) assigned to x-levels by experimenter
  - **Purely controlled experiments**: researcher uses only predictors that were assigned to units
  - **Controlled experiments with covariates** (uncontrolled variables): researcher has additional predictors associated with units
- **Observational studies**: subjects have x-levels associated with them (not assigned by researcher)
  - **Confirmatory studies**: it is *hypothesized* that new (primary) predictors are associated with Y; while there are predictors, *known* to be associated with Y (called risk factors).
  - **Exploratory studies**: it is hypothesized that some or all of potential predictors are associated with Y.

Reduction of Explanatory Variables

- Controlled experiments

Reduction of Explanatory Variables

- Controlled experiments
  - Purely controlled experiments: rarely any need or desire to reduce number of predictors

Reduction of Explanatory Variables

- Controlled experiments
  - Purely controlled experiments: rarely any need or desire to reduce number of predictors
  - Controlled experiments with covariates: remove any covariates that do not reduce the error variance

Reduction of Explanatory Variables

- Controlled experiments
  - Purely controlled experiments: rarely any need or desire to reduce number of predictors
  - Controlled experiments with covariates: remove any covariates that do not reduce the error variance
- Observational Studies

Reduction of Explanatory Variables

- Controlled experiments
  - Purely controlled experiments: rarely any need or desire to reduce number of predictors
  - Controlled experiments with covariates: remove any covariates that do not reduce the error variance
- Observational Studies
  - Confirmatory Studies: Must keep in all risk factors to compare with previous research, should keep all primary variables as well

Reduction of Explanatory Variables

- Controlled experiments
  - Purely controlled experiments: rarely any need or desire to reduce number of predictors
  - Controlled experiments with covariates: remove any covariates that do not reduce the error variance
- Observational Studies
  - Confirmatory Studies: Must keep in all risk factors to compare with previous research, should keep all primary variables as well
  - Exploratory Studies: Often have many potential predictors (and polynomials and interactions).

Reduction of Explanatory Variables

- Controlled experiments
  - Purely controlled experiments: rarely any need or desire to reduce number of predictors
  - Controlled experiments with covariates: remove any covariates that do not reduce the error variance
- Observational Studies
  - Confirmatory Studies: Must keep in all risk factors to compare with previous research, should keep all primary variables as well
  - Exploratory Studies: Often have many potential predictors (and polynomials and interactions). Want to fit parsimonious model that explains much of the variation in Y, while keeping model as basic as possible.

Reduction of Explanatory Variables

- Controlled experiments
  - Purely controlled experiments: rarely any need or desire to reduce number of predictors
  - Controlled experiments with covariates: remove any covariates that do not reduce the error variance
- Observational Studies
  - Confirmatory Studies: Must keep in all risk factors to compare with previous research, should keep all primary variables as well
  - Exploratory Studies: Often have many potential predictors (and polynomials and interactions). Want to fit parsimonious model that explains much of the variation in Y, while keeping model as basic as possible. Caution: one shall not make decisions based on single variable t-tests.

9.2 Surgical unit example

- First steps often involve plots:
  - Plots to indicate correct functional form of predictors and/or response.
  - Plots to indicate possible interaction.
  - Exploration of correlation among predictors (maybe).
  - Often a first-order model is a good starting point.
- Once a reasonable set of potential predictors is identified, formal model selection begins.
- If the number of predictors is large, say $k \geq 10$, we can use (automated) stepwise procedures to reduce the number of variables (and models) under consideration.

Surgical unit example: predictors

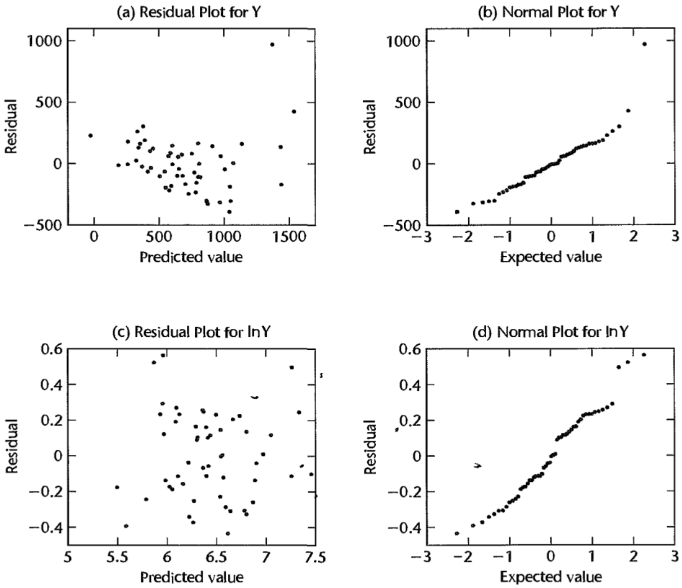
A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random selection of 108 patients was available for analysis.

X_1	blood clotting score
X_2	prognostic index
X_3	enzyme function test score
X_4	liver function test score

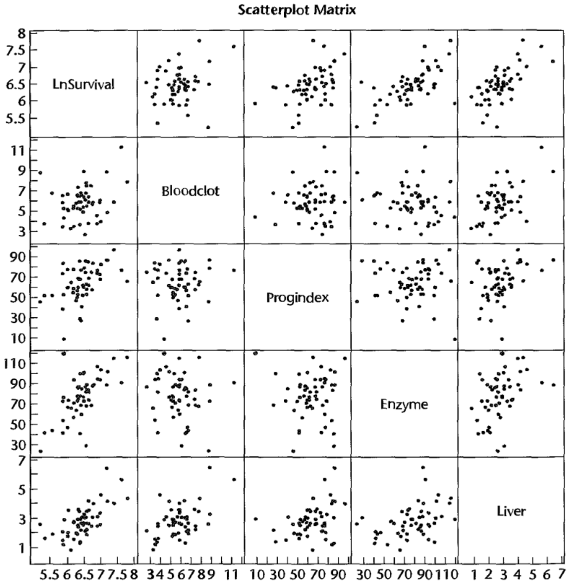
...

Case Number i	Blood-Clotting Score X_{i1}	Prognostic Index X_{i2}	Enzyme Test X_{i3}	Liver Test X_{i4}	Age X_{i5}	Gender X_{i6}	Alc. Use: Mod. X_{i7}	Alc. Use: Heavy X_{i8}	Survival Time Y_i	$Y_i' = \ln Y_i$
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
...	...	...	...	...	...	...	...	...	...	...
52	6.4	85	40	1.21	58	0	0	1	579	6.361
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

Surgical unit example: residual plots



Surgical unit example: scatterplot matrix



9.3: Model selection

Once we reduce the set of potential predictors to a reasonable number, we can examine all possible models and choose the "best" according to some criterion.

9.3: Model selection

Once we reduce the set of potential predictors to a reasonable number, we can examine all possible models and choose the "best" according to some criterion.

Say we have k predictors $x_1, \ldots, x_k$ and we want to find a good subset of predictors that predict the data well. There are several useful criteria to help choose a subset of predictors.

Adjusted-R^2, R_a^2

The *adjusted* R^2, denoted R_a^2 “fixes” R^2 to provide a measure of how good the model will predict data not used to build the model. For a candidate model with $p-1$ predictors

$$R_a^2 = 1 - \frac{\text{SSE}/(n-p)}{\text{SSTO}/(n-1)}$$

Adjusted-R^2, R_a^2

The *adjusted* R^2, denoted R_a^2 "fixes" R^2 to provide a measure of how good the model will predict data not used to build the model. For a candidate model with $p-1$ predictors

$$R_a^2 = 1 - \frac{\text{SSE}/(n-p)}{\text{SSTO}/(n-1)} \quad \left(= 1 - \frac{\text{MSE}}{S_Y^2}\right).$$

- Equivalent to choosing the model with the *smallest* MSE.

Adjusted-R^2, R_a^2

The *adjusted* R^2, denoted R_a^2 "fixes" R^2 to provide a measure of how good the model will predict data not used to build the model. For a candidate model with $p-1$ predictors

$$R_a^2 = 1 - \frac{\text{SSE}/(n-p)}{\text{SSTO}/(n-1)} \quad \left(= 1 - \frac{\text{MSE}}{S_Y^2} \right).$$

- Equivalent to choosing the model with the *smallest* MSE.
- If irrelevant variables are added, R_a^2 may decrease unlike "regular" R^2 (R_a^2 can be negative!).

Adjusted-R^2, R_a^2

The *adjusted* R^2, denoted R_a^2 "fixes" R^2 to provide a measure of how good the model will predict data not used to build the model. For a candidate model with $p-1$ predictors

$$R_a^2 = 1 - \frac{\mathrm{SSE}/(n-p)}{\mathrm{SSTO}/(n-1)} \quad \left(= 1 - \frac{\mathrm{MSE}}{S_Y^2}\right).$$

- Equivalent to choosing the model with the *smallest* MSE.
- If irrelevant variables are added, R_a^2 may decrease unlike "regular" R^2 (R_a^2 can be negative!).
- R_a^2 penalizes model for being too complex.

Adjusted-R^2, R_a^2

The *adjusted* R^2, denoted R_a^2 "fixes" R^2 to provide a measure of how good the model will predict data not used to build the model. For a candidate model with $p-1$ predictors

$$R_a^2 = 1 - \frac{\mathrm{SSE}/(n-p)}{\mathrm{SSTO}/(n-1)} \quad \left(= 1 - \frac{\mathrm{MSE}}{S_Y^2}\right).$$

- Equivalent to choosing the model with the *smallest* MSE.
- If irrelevant variables are added, R_a^2 may decrease unlike "regular" R^2 (R_a^2 can be negative!).
- R_a^2 penalizes model for being too complex.
- Problem: R_a^2 is greater for a "bigger" model whenever the F-statistic for comparing bigger to smaller is greater than 1 (**show this**).

Adjusted-R^2, R_a^2

The *adjusted* R^2, denoted R_a^2 "fixes" R^2 to provide a measure of how good the model will predict data not used to build the model. For a candidate model with $p-1$ predictors

$$R_a^2 = 1 - \frac{\mathrm{SSE}/(n-p)}{\mathrm{SSTO}/(n-1)} \quad \left(= 1 - \frac{\mathrm{MSE}}{S_Y^2}\right).$$

- Equivalent to choosing the model with the *smallest* MSE.
- If irrelevant variables are added, R_a^2 may decrease unlike "regular" R^2 (R_a^2 can be negative!).
- R_a^2 penalizes model for being too complex.
- Problem: R_a^2 is greater for a "bigger" model whenever the F-statistic for comparing bigger to smaller is greater than 1 (**show this**). We usually want F-statistic to be a *lot* bigger than 1 before adding in new predictors

Adjusted-R^2, R_a^2

The *adjusted* R^2, denoted R_a^2 "fixes" R^2 to provide a measure of how good the model will predict data not used to build the model. For a candidate model with $p-1$ predictors

$$R_a^2 = 1 - \frac{\text{SSE}/(n-p)}{\text{SSTO}/(n-1)} \quad \left(= 1 - \frac{\text{MSE}}{S_Y^2}\right).$$

- Equivalent to choosing the model with the *smallest* MSE.
- If irrelevant variables are added, R_a^2 may decrease unlike "regular" R^2 (R_a^2 can be negative!).
- R_a^2 penalizes model for being too complex.
- Problem: R_a^2 is greater for a "bigger" model whenever the F-statistic for comparing bigger to smaller is greater than 1 (**show this**). We usually want F-statistic to be a *lot* bigger than 1 before adding in new predictors $\Rightarrow$ *too liberal*.

Akaike Information Criterion

In general, **Akaike Information Criterion (AIC)** is

$$\mathrm{AIC} = -2 \ln L(\hat{\boldsymbol{\theta}}) + 2p$$

for a model with parameters $\boldsymbol{\theta} \in \mathbb{R}^p$ and likelihood function L.

Akaike Information Criterion

In general, **Akaike Information Criterion (AIC)** is

$$\mathrm{AIC} = -2 \ln L(\hat{\boldsymbol{\theta}}) + 2p$$

for a model with parameters $\boldsymbol{\theta} \in \mathbb{R}^p$ and likelihood function L.

Exercise. Show that for the MLR with normal errors,

$$\mathrm{AIC} = n \ln(\mathrm{SSE}) + 2p + C$$

where C is a constant that does not depend on SSE and p.

Akaike Information Criterion

In general, **Akaike Information Criterion (AIC)** is

$$\mathrm{AIC} = -2 \ln L(\hat{\boldsymbol{\theta}}) + 2p$$

for a model with parameters $\boldsymbol{\theta} \in \mathbb{R}^p$ and likelihood function L.

Exercise. Show that for the MLR with normal errors,

$$\mathrm{AIC} = n \ln(\mathrm{SSE}) + 2p + C$$

where C is a constant that does not depend on SSE and p.

- $2p$ is "penalty" term for adding predictors.

Akaike Information Criterion

In general, **Akaike Information Criterion (AIC)** is

$$\text{AIC} = -2 \ln L(\hat{\boldsymbol{\theta}}) + 2p$$

for a model with parameters $\boldsymbol{\theta} \in \mathbb{R}^p$ and likelihood function L.

Exercise. Show that for the MLR with normal errors,

$$\text{AIC} = n \ln(\text{SSE}) + 2p + C$$

where C is a constant that does not depend on SSE and p.

- $2p$ is "penalty" term for adding predictors.
- Like R_a^2, AIC favors models with small SSE, but penalizes models with too many variables p.

Akaike Information Criterion

In general, **Akaike Information Criterion (AIC)** is

$$\mathrm{AIC} = -2 \ln L(\hat{\boldsymbol{\theta}}) + 2p$$

for a model with parameters $\boldsymbol{\theta} \in \mathbb{R}^p$ and likelihood function L.

Exercise. Show that for the MLR with normal errors,

$$\mathrm{AIC} = n \ln(\mathrm{SSE}) + 2p + C$$

where C is a constant that does not depend on SSE and p.

- $2p$ is "penalty" term for adding predictors.
- Like R_a^2, AIC favors models with small SSE, but penalizes models with too many variables p.
- $\Rightarrow$ Between two models, we prefer the one with lower AIC.

Bayesian Information Criterion

In general, **Bayesian Information Criterion (BIC)** is

$$\mathrm{BIC} = -2 \ln L(\hat{\boldsymbol{\theta}}) + p \ln(n)$$

Bayesian Information Criterion

In general, **Bayesian Information Criterion (BIC)** is

$$\text{BIC} = -2 \ln L(\hat{\boldsymbol{\theta}}) + p \ln(n)$$

Exercise. Show that for the MLR with normal errors,

$$\text{BIC} = n \ln(\text{SSE}) + p \log(n) + C$$

Bayesian Information Criterion

In general, **Bayesian Information Criterion (BIC)** is

$$\mathrm{BIC} = -2\ln L(\hat{\boldsymbol{\theta}}) + p\ln(n)$$

Exercise. Show that for the MLR with normal errors,

$$\mathrm{BIC} = n\ln(\mathrm{SSE}) + p\log(n) + C$$

- BIC is similar to AIC, but for $n \geq 8$, the BIC "penalty term" is more severe.

Mallow's C_p

Full model with k predictors and Reduced model with $p - 1$ predictors.

Mallow's C_p

Full model with k predictors and Reduced model with $p - 1$ predictors. **Mallow's** C_p is

$$C_p = \frac{\text{SSE(Reduced)}}{\text{MSE(Full)}} - n + 2p$$

Mallow's C_p

Full model with k predictors and Reduced model with $p-1$ predictors. **Mallow's** C_p is

$$C_p = \frac{\mathrm{SSE(Reduced)}}{\mathrm{MSE(Full)}} - n + 2p$$

- Estimates $\frac{\mathrm{E}[\hat{Y}_i - \mathrm{E}[Y_i]]^2}{\sigma^2}$, where $\hat{Y}_i$ is from the Reduced model (pp. 357–359).

Mallow's C_p

Full model with k predictors and Reduced model with $p-1$ predictors. **Mallow's** C_p is

$$C_p = \frac{\text{SSE(Reduced)}}{\text{MSE(Full)}} - n + 2p$$

- Estimates $\frac{\mathrm{E}[\hat{Y}_i - \mathrm{E}[Y_i]]^2}{\sigma^2}$, where $\hat{Y}_i$ is from the Reduced model (pp. 357–359).
- Recall that $\mathrm{E}[\text{MSE(Full)}] = \sigma^2$

Mallow's C_p

Full model with k predictors and Reduced model with $p-1$ predictors. **Mallow's** C_p is

$$C_p = \frac{\text{SSE(Reduced)}}{\text{MSE(Full)}} - n + 2p$$

- Estimates $\frac{\mathrm{E}[\hat{Y}_i - \mathrm{E}[Y_i]]^2}{\sigma^2}$, where $\hat{Y}_i$ is from the Reduced model (pp. 357–359).
- Recall that $\mathrm{E}[\text{MSE(Full)}] = \sigma^2$
- If the Reduced model is unbiased, i.e. $\mathrm{E}[\text{MSE(Reduced)}] = \sigma^2$,

Mallow's C_p

Full model with k predictors and Reduced model with $p-1$ predictors. **Mallow's** C_p is

$$C_p = \frac{\text{SSE(Reduced)}}{\text{MSE(Full)}} - n + 2p$$

- Estimates $\frac{\mathrm{E}[\hat{Y}_i - \mathrm{E}[Y_i]]^2}{\sigma^2}$, where $\hat{Y}_i$ is from the Reduced model (pp. 357–359).
- Recall that $\mathrm{E}[\text{MSE(Full)}] = \sigma^2$
- If the Reduced model is unbiased, i.e. $\mathrm{E}[\text{MSE(Reduced)}] = \sigma^2$, then $\mathrm{E}[\text{SSE(Reduced)}] = (n-p)\sigma^2$,

Mallow's C_p

Full model with k predictors and Reduced model with $p-1$ predictors. **Mallow's** C_p is

$$C_p = \frac{\text{SSE(Reduced)}}{\text{MSE(Full)}} - n + 2p$$

- Estimates $\frac{\mathrm{E}[\hat{Y}_i - \mathrm{E}[Y_i]]^2}{\sigma^2}$, where $\hat{Y}_i$ is from the Reduced model (pp. 357–359).
- Recall that $\mathrm{E}[\text{MSE(Full)}] = \sigma^2$
- If the Reduced model is unbiased, i.e. $\mathrm{E}[\text{MSE(Reduced)}] = \sigma^2$, then $\mathrm{E}[\text{SSE(Reduced)}] = (n-p)\sigma^2$, and

$$C_p \approx \frac{(n-p)\sigma^2}{\sigma^2} - n + 2p = p$$

Mallow's C_p

Full model with k predictors and Reduced model with $p-1$ predictors. **Mallow's** C_p is

$$C_p = \frac{\text{SSE(Reduced)}}{\text{MSE(Full)}} - n + 2p$$

- Estimates $\frac{\mathrm{E}[\hat{Y}_i - \mathrm{E}[Y_i]]^2}{\sigma^2}$, where $\hat{Y}_i$ is from the Reduced model (pp. 357–359).
- Recall that $\mathrm{E}[\text{MSE(Full)}] = \sigma^2$
- If the Reduced model is unbiased, i.e. $\mathrm{E}[\text{MSE(Reduced)}] = \sigma^2$, then $\mathrm{E}[\text{SSE(Reduced)}] = (n-p)\sigma^2$, and

$$C_p \approx \frac{(n-p)\sigma^2}{\sigma^2} - n + 2p = p$$

- The Full model always has $C_{k+1} = k + 1$.

Mallow's C_p

If $C_p \approx p$ then the reduced model predicts as well as the full model. If $C_p < p$ then the reduced model is estimated to be less *biased* than the full model.

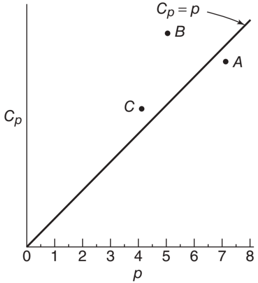


Figure: A C_p plot

Mallow's C_p

If $C_p \approx p$ then the reduced model predicts as well as the full model. If $C_p < p$ then the reduced model is estimated to be less *biased* than the full model.

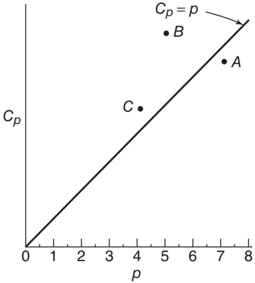


Figure: A C_p plot

In practice, just choose model with smallest C_p.

Which criteria to use?

R_a^2, AIC, BIC, and C_p may give different "best" models, or they may agree. Ultimate goal is to find model that balances:

Which criteria to use?

R_a^2, AIC, BIC, and C_p may give different "best" models, or they may agree. Ultimate goal is to find model that balances:

- A good fit to the data.

Which criteria to use?

R_a^2, AIC, BIC, and C_p may give different "best" models, or they may agree. Ultimate goal is to find model that balances:

- A good fit to the data.
- Low bias.

Which criteria to use?

R_a^2, AIC, BIC, and C_p may give different "best" models, or they may agree. Ultimate goal is to find model that balances:

- A good fit to the data.
- Low bias.
- Parsimony.

All else being equal, the simpler model is often easier to interpret and work with. Christensen (1996) recommends C_p and notes the similarity between C_p and AIC.

Methods for "automatically" picking variables

- Any regression textbook will caution against not thinking about the data at all and simply using automated procedures.

Methods for “automatically” picking variables

- Any regression textbook will caution against not thinking about the data at all and simply using automated procedures.
- Automated procedures *cannot* assess a good functional form for a predictor, *cannot* think about which interactions might be important, etc.

Methods for "automatically" picking variables

- Any regression textbook will caution against not thinking about the data at all and simply using automated procedures.
- Automated procedures *cannot* assess a good functional form for a predictor, *cannot* think about which interactions might be important, etc.
- Anyway, automated procedures are widely used and *can* produce good models.

Methods for "automatically" picking variables

- Any regression textbook will caution against not thinking about the data at all and simply using automated procedures.
- Automated procedures *cannot* assess a good functional form for a predictor, *cannot* think about which interactions might be important, etc.
- Anyway, automated procedures are widely used and *can* produce good models. But they can also produce models that are *substantially inferior* to other models built from the same predictors using scientific input and common sense.

Example: cruise ships

https://github.com/zh3nis/MATH440/blob/main/chp09/cruise.R

A cruise ship company wishes to model the `crew size` needed for a ship using predictors such as: `age`, `tonnage`, `passengers`, `length`, `cabins` and passenger density (`passdens`).

Example: cruise ships

https://github.com/zh3nis/MATH440/blob/main/chp09/cruise.R

A cruise ship company wishes to model the `crew size` needed for a ship using predictors such as: `age`, `tonnage`, `passengers`, `length`, `cabins` and passenger density (`passdens`).

```
> cruise <- read.fwf("http://www.stat.ufl.edu/~winner/data/cruise_ship.dat", ...)
> head(cruise)
         ship      cline age tonnage passengers length cabins passdens  crew
1 Journey        Azamara   6  30.277       6.94   5.94   3.55    42.64  3.55
2 Quest          Azamara   6  30.277       6.94   5.94   3.55    42.64  3.55
3 Celebration   Carnival  26  47.262      14.86   7.22   7.43    31.80  6.70
4 Conquest      Carnival  11 110.000      29.74   9.53  14.88    36.99 19.10
5 Destiny       Carnival  17 101.353      26.42   8.92  13.21    38.36 10.00
6 Ecstasy       Carnival  22  70.367      20.52   8.55  10.20    34.29  9.20
```

Example: cruise ships

https://github.com/zh3nis/MATH440/blob/main/chp09/cruise.R

A cruise ship company wishes to model the `crew size` needed for a ship using predictors such as: `age`, `tonnage`, `passengers`, `length`, `cabins` and passenger density (`passdens`).

```
> cruise <- read.fwf("http://www.stat.ufl.edu/~winner/data/cruise_ship.dat", ...)
> head(cruise)
         ship      cline age tonnage passengers length cabins passdens  crew
1 Journey        Azamara   6  30.277       6.94   5.94   3.55    42.64  3.55
2 Quest          Azamara   6  30.277       6.94   5.94   3.55    42.64  3.55
3 Celebration   Carnival  26  47.262      14.86   7.22   7.43    31.80  6.70
4 Conquest      Carnival  11 110.000      29.74   9.53  14.88    36.99 19.10
5 Destiny       Carnival  17 101.353      26.42   8.92  13.21    38.36 10.00
6 Ecstasy       Carnival  22  70.367      20.52   8.55  10.20    34.29  9.20
```

Without concerning ourselves with potential interactions we will look at simple additive models.

Cruise ships — Full model

```
> fit0 = lm(crew ~ age + tonnage + passengers + length + cabins + passdens, data=cruise)
> summary(fit0)

Call:
lm(formula = crew ~ age + tonnage + passengers + length + cabins +
    passdens, data = cruise)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7700 -0.4881 -0.0938  0.4454  7.0077

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.5213400  1.0570350  -0.493  0.62258
age         -0.0125449  0.0141975  -0.884  0.37832
tonnage      0.0132410  0.0118928   1.113  0.26732
passengers  -0.1497640  0.0475886  -3.147  0.00199 **
length       0.4034785  0.1144548   3.525  0.00056 ***
cabins       0.8016337  0.0892227   8.985 9.84e-16 ***
passdens    -0.0006577  0.0158098  -0.042  0.96687
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9819 on 151 degrees of freedom
Multiple R-squared:  0.9245,Adjusted R-squared:  0.9215
F-statistic:   308 on 6 and 151 DF,  p-value: < 2.2e-16
```

Best subsets

```
> library(leaps)
> allcruise <- regsubsets(crew ~ age + tonnage + passengers + length + cabins + passdens,
+                         nbest=4, data=cruise)
> all_output <- summary(allcruise)
> with(all_output, round(cbind(which, rsq, adjr2, cp, bic), 3))
  (Intercept) age tonnage passengers length cabins passdens   rsq adjr2      cp      bic
1           1   0       0          0      0      1        0 0.904 0.903  37.772 -360.238
1           1   0       1          0      0      0        0 0.860 0.859 125.086 -300.954
1           1   0       0          1      0      0        0 0.838 0.837 170.523 -277.122
1           1   0       0          0      1      0        0 0.803 0.801 240.675 -246.201
2           1   0       0          0      1      1        0 0.916 0.915  15.952 -376.131
2           1   0       0          0      0      1        1 0.912 0.911  24.261 -368.502
2           1   0       1          0      0      1        0 0.911 0.909  26.792 -366.249
2           1   0       0          1      0      1        0 0.908 0.907  32.443 -361.332
3           1   0       0          1      1      1        0 0.922 0.921   5.857 -382.878
3           1   0       0          0      1      1        1 0.919 0.918  11.341 -377.413
3           1   0       1          1      0      1        0 0.918 0.916  14.023 -374.808
3           1   1       0          0      1      1        0 0.917 0.915  15.909 -373.002
4           1   0       1          1      1      1        0 0.924 0.922   3.847 -381.933
4           1   1       0          1      1      1        0 0.923 0.921   5.084 -380.652
4           1   0       0          1      1      1        1 0.923 0.921   5.197 -380.534
4           1   0       1          0      1      1        1 0.919 0.917  13.056 -372.631
5           1   1       1          1      1      1        0 0.924 0.922   5.002 -377.752
5           1   0       1          1      1      1        1 0.924 0.922   5.781 -376.939
5           1   1       0          1      1      1        1 0.924 0.921   6.240 -376.462
5           1   1       1          0      1      1        1 0.920 0.917  14.904 -367.717
6           1   1       1          1      1      1        1 0.924 0.921   7.000 -372.692
```

Best subsets

```
> library(leaps)
> allcruise <- regsubsets(crew ~ age + tonnage + passengers + length + cabins + passdens,
+                         nbest=4, data=cruise)
> all_output <- summary(allcruise)
> with(all_output, round(cbind(which, rsq, adjr2, cp, bic), 3))
  (Intercept) age tonnage passengers length cabins passdens   rsq adjr2      cp      bic
1           1   0       0          0      0      1        0 0.904 0.903  37.772 -360.238
1           1   0       1          0      0      0        0 0.860 0.859 125.086 -300.954
1           1   0       0          1      0      0        0 0.838 0.837 170.523 -277.122
1           1   0       0          0      1      0        0 0.803 0.801 240.675 -246.201
2           1   0       0          0      1      1        0 0.916 0.915  15.952 -376.131
2           1   0       0          0      0      1        1 0.912 0.911  24.261 -368.502
2           1   0       1          0      0      1        0 0.911 0.909  26.792 -366.249
2           1   0       0          1      0      1        0 0.908 0.907  32.443 -361.332
3           1   0       0          1      1      1        0 0.922 0.921   5.857 -382.878
3           1   0       0          0      1      1        1 0.919 0.918  11.341 -377.413
3           1   0       1          1      0      1        0 0.918 0.916  14.023 -374.808
3           1   1       0          0      1      1        0 0.917 0.915  15.909 -373.002
4           1   0       1          1      1      1        0 0.924 0.922   3.847 -381.933
4           1   1       0          1      1      1        0 0.923 0.921   5.084 -380.652
4           1   0       0          1      1      1        1 0.923 0.921   5.197 -380.534
4           1   0       1          0      1      1        1 0.919 0.917  13.056 -372.631
5           1   1       1          1      1      1        0 0.924 0.922   5.002 -377.752
5           1   0       1          1      1      1        1 0.924 0.922   5.781 -376.939
5           1   1       0          1      1      1        1 0.924 0.921   6.240 -376.462
5           1   1       1          0      1      1        1 0.920 0.917  14.904 -367.717
6           1   1       1          1      1      1        1 0.924 0.921   7.000 -372.692
```

A good model choice might be the model with 4 predictors: tonnage, passengers, length, and cabins, whose $R_a^2 = 0.922$, $C_p = 3.847$, and BIC $= 381.933$.

AIC full vs reduced

```
> fit3 <- update(fit0, . ~ . - age - passdens)
> AIC(fit3)
[1] 448.3229
> AIC(fit0)
[1] 451.4394
```

9.4 Automated variable search

As discussed, it is possible to have a large set of predictor variables (including interactions). The goal is to fit a "parsimoneous" model that explains as much variation in the response as possible with a relatively small set of predictors.

9.4 Automated variable search

As discussed, it is possible to have a large set of predictor variables (including interactions). The goal is to fit a "parsimoneous" model that explains as much variation in the response as possible with a relatively small set of predictors.

There are 3 automated procedures

9.4 Automated variable search

As discussed, it is possible to have a large set of predictor variables (including interactions). The goal is to fit a "parsimoneous" model that explains as much variation in the response as possible with a relatively small set of predictors.

There are 3 automated procedures

- Backward Elimination (Top down approach)

9.4 Automated variable search

As discussed, it is possible to have a large set of predictor variables (including interactions). The goal is to fit a "parsimoneous" model that explains as much variation in the response as possible with a relatively small set of predictors.

There are 3 automated procedures

- Backward Elimination (Top down approach)
- Forward Selection (Bottom up approach)

9.4 Automated variable search

As discussed, it is possible to have a large set of predictor variables (including interactions). The goal is to fit a "parsimoneous" model that explains as much variation in the response as possible with a relatively small set of predictors.

There are 3 automated procedures

- Backward Elimination (Top down approach)
- Forward Selection (Bottom up approach)
- Stepwise Regression (Combines Forward/Backward)

9.4 Automated variable search

As discussed, it is possible to have a large set of predictor variables (including interactions). The goal is to fit a "parsimoneous" model that explains as much variation in the response as possible with a relatively small set of predictors.

There are 3 automated procedures

- Backward Elimination (Top down approach)
- Forward Selection (Bottom up approach)
- Stepwise Regression (Combines Forward/Backward)

We will explore these procedures using two different elimination/selection criteria.

9.4 Automated variable search

As discussed, it is possible to have a large set of predictor variables (including interactions). The goal is to fit a "parsimoneous" model that explains as much variation in the response as possible with a relatively small set of predictors.

There are 3 automated procedures

- Backward Elimination (Top down approach)
- Forward Selection (Bottom up approach)
- Stepwise Regression (Combines Forward/Backward)

We will explore these procedures using two different elimination/selection criteria. One that uses t-test and p-value and another that uses the AIC value.

Backward elimination

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

Backward elimination

1. Select a significance level to *stay* in the model (e.g. $\alpha_s = 0.20$, generally .05 is too low, causing too many variables to be removed).
2. Fit the full model with all possible predictors.

Backward elimination

1. Select a significance level to *stay* in the model (e.g. $\alpha_s = 0.20$, generally .05 is too low, causing too many variables to be removed).
2. Fit the full model with all possible predictors.
3. Consider the predictor with lowest t-statistic (highest p-value).

Backward elimination

1. Select a significance level to *stay* in the model (e.g. $\alpha_s = 0.20$, generally .05 is too low, causing too many variables to be removed).
2. Fit the full model with all possible predictors.
3. Consider the predictor with lowest t-statistic (highest p-value).
   - If p-value $> \alpha_s$, remove the predictor and fit model without this variable (must re-fit model here because partial regression coefficients change).

Backward elimination

1. Select a significance level to *stay* in the model (e.g. $\alpha_s = 0.20$, generally .05 is too low, causing too many variables to be removed).
2. Fit the full model with all possible predictors.
3. Consider the predictor with lowest t-statistic (highest p-value).
   - If p-value $> \alpha_s$, remove the predictor and fit model without this variable (must re-fit model here because partial regression coefficients change).
   - If p-value $\leq \alpha_s$, stop and keep current model.

Backward elimination

1. Select a significance level to *stay* in the model (e.g. $\alpha_s = 0.20$, generally .05 is too low, causing too many variables to be removed).
2. Fit the full model with all possible predictors.
3. Consider the predictor with lowest t-statistic (highest p-value).
   - If p-value $> \alpha_s$, remove the predictor and fit model without this variable (must re-fit model here because partial regression coefficients change).
   - If p-value $\leq \alpha_s$, stop and keep current model.
4. Continue until all predictors have p-values $\leq \alpha_s$.

Forward selection

1. Select a significance level to *enter* the model (e.g. $\alpha_e = 0.20$, generally .05 is too low, causing too few variables to be entered).

Forward selection

1. Select a significance level to *enter* the model (e.g. $\alpha_e = 0.20$, generally .05 is too low, causing too few variables to be entered).
2. Fit all simple regression models.

Forward selection

1. Select a significance level to *enter* the model (e.g. $\alpha_e = 0.20$, generally .05 is too low, causing too few variables to be entered).
2. Fit all simple regression models.
3. Consider the predictor with the highest t-statistic (lowest p-value).

Forward selection

1. Select a significance level to *enter* the model (e.g. $\alpha_e = 0.20$, generally .05 is too low, causing too few variables to be entered).
2. Fit all simple regression models.
3. Consider the predictor with the highest t-statistic (lowest p-value).
   - If p-value $\leq \alpha_e$, keep this variable and fit all two variable models that include this predictor.

Forward selection

1. Select a significance level to *enter* the model (e.g. $\alpha_e = 0.20$, generally .05 is too low, causing too few variables to be entered).
2. Fit all simple regression models.
3. Consider the predictor with the highest t-statistic (lowest p-value).
    - If p-value $\leq \alpha_e$, keep this variable and fit all two variable models that include this predictor.
    - If p-value $> \alpha_e$, stop and keep previous model.

Forward selection

1. Select a significance level to *enter* the model (e.g. $\alpha_e = 0.20$, generally .05 is too low, causing too few variables to be entered).
2. Fit all simple regression models.
3. Consider the predictor with the highest t-statistic (lowest p-value).
   - If p-value $\leq \alpha_e$, keep this variable and fit all two variable models that include this predictor.
   - If p-value $> \alpha_e$, stop and keep previous model.
4. Continue until no new predictors have p-values $\leq \alpha_e$.

Stepwise regression

1. Select α_s and α_e, $(\alpha_e < \alpha_s)$.

Stepwise regression

1. Select α_s and α_e, $(\alpha_e < \alpha_s)$.
2. Start like Forward Selection (bottom up process) where new variables must have p-value $\leq \alpha_e$ to enter.

Stepwise regression

1. Select α_s and α_e, $(\alpha_e < \alpha_s)$.
2. Start like Forward Selection (bottom up process) where new variables must have p-value $\leq \alpha_e$ to enter.
3. Re-test all "old variables" that have already been entered, must have p-value $\leq \alpha_s$ to stay in model.

Stepwise regression

1. Select α_s and α_e, ($\alpha_e < \alpha_s$).
2. Start like Forward Selection (bottom up process) where new variables must have p-value $\leq \alpha_e$ to enter.
3. Re-test all "old variables" that have already been entered, must have p-value $\leq \alpha_s$ to stay in model.
4. Continue until no new variables can be entered and no old variables need to be removed.

Backward and Forward for cruise ships

```
> library(olsrr)
> ols_step_backward_p(fit0)

                              Elimination Summary
------------------------------------------------------------------------
        Variable                  Adj.
Step    Removed     R-Square    R-Square     C(p)        AIC        RMSE
------------------------------------------------------------------------
   1    passdens      0.9245       0.922    5.0017    449.4412    0.9786
   2    age            0.924      0.9221    3.8468    448.3229    0.9782
------------------------------------------------------------------------

> ols_step_forward_p(fit0)

                              Selection Summary
---------------------------------------------------------------------------
        Variable                  Adj.
Step     Entered      R-Square    R-Square     C(p)         AIC        RMSE
---------------------------------------------------------------------------
   1    cabins          0.9041      0.9034    37.7724    479.2060    1.0886
   2    length          0.9160      0.9149    15.9524    460.2507    1.0221
   3    passengers      0.9220      0.9205     5.8566    450.4411    0.9878
   4    tonnage         0.9240      0.9221     3.8468    448.3229    0.9782
---------------------------------------------------------------------------
```

Stepwise procedure for cruise ships

```
> ols_step_both_p(fit0)
```

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	cabins	addition	0.904	0.903	37.7720	479.2060	1.0886
2	length	addition	0.916	0.915	15.9520	460.2507	1.0221
3	passengers	addition	0.922	0.921	5.8570	450.4411	0.9878
4	tonnage	addition	0.924	0.922	3.8470	448.3229	0.9782

PRESS_p criterion

$$\text{PRESS}_p = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(i)})^2 \quad \left(= \sum_{i=1}^{n} \left[\frac{e_i}{1 - h_{ii}} \right]^2 \right),$$

where $\hat{Y}_{i(i)}$ is the fitted value at $\mathbf{x}_i$ with the $(\mathbf{x}_i, Y_i)$ omitted.

PRESS_p criterion

$$\text{PRESS}_p = \sum_{i=1}^{n}(Y_i - \hat{Y}_{i(i)})^2 \quad \left(= \sum_{i=1}^{n}\left[\frac{e_i}{1-h_{ii}}\right]^2 \right),$$

where $\hat{Y}_{i(i)}$ is the fitted value at $\mathbf{x}_i$ with the $(\mathbf{x}_i, Y_i)$ omitted.

- This is leave-one-out prediction error. The smaller, the better.

PRESS_p criterion

$$\text{PRESS}_p = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(i)})^2 \quad \left(= \sum_{i=1}^{n} \left[\frac{e_i}{1 - h_{ii}} \right]^2 \right),$$

where $\hat{Y}_{i(i)}$ is the fitted value at $\mathbf{x}_i$ with the $(\mathbf{x}_i, Y_i)$ omitted.

- This is leave-one-out prediction error. The smaller, the better.
- Having $\text{PRESS}_p \approx \text{SSE}_p$ supports the *validity* of the model with p predictors (p. 374).

Caveats for automated procedures

- There is no “best” way to search for good models.

Caveats for automated procedures

- There is no "best" way to search for good models.
- There may be *several* "good" models.

Caveats for automated procedures

- There is no "best" way to search for good models.
- There may be *several* "good" models.
- If you use the same data to *estimate* the model and *choose* the model, the regression effects are *biased*! $This$ leads to the idea of data splitting; one portion of the data is the *training data* and the other portion is the *validation set* (Section 9.6, p. 372). PRESS_p can also be used.