

# Stat 206: Linear Models

## Lecture 14

Nov. 18, 2019

# Overview of Model-Building

- Data collection and processing
- Exploratory data analysis
- Preliminary model investigation
- Model selection
- Model diagnostic and validation

# Exploratory Data Analysis

- Type of each variable: quantitative or qualitative?
- Distribution of each variable: symmetric or skewed? outliers?
  - Quantitative: histogram, boxplot, summary statistics, etc.
  - Qualitative: pie chart, frequency table, etc.
- Relationships among variables.
  - scatter plot matrix, correlation matrix,
  - nonlinear pattern? clusters? outliers?

# Preliminary Model Fitting

- Residual plots based on initial fits:
  - nonlinearity? departure from Normality? nonconstant error variance?
  - transformations needed?
  - omission of important predictors/interaction terms/high-order power terms?
- The goal is to decide on:
  - Functional forms in which variables should enter the regression model.
  - Potential pool of predictors, interactions and higher-order powers to be considered in subsequent analysis.
- This process should be aided by prior knowledge and domain expertise if possible.

# Surgical Unit

A hospital surgical unit was interested in predicting survival times of patients ( $Y$ , in days, ascertained in a follow-up study) undergoing a particular type of liver operation. 108 such patients were randomly selected for this study. The following variables were measured for each patient: bleeding clotting score ( $X_1$ ), prognostic index ( $X_2$ ), enzyme function test score ( $X_3$ ), liver function test score ( $X_4$ ), age ( $X_5$ , in years), gender (male or female) and history of alcohol use (none, moderate or severe). The two qualitative variables are quantified by the following indicator variables:

$$X_6 = \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}$$

$$X_7 = \begin{cases} 1 & \text{if moderate use} \\ 0 & \text{if otherwise} \end{cases} \quad X_8 = \begin{cases} 1 & \text{if severe use} \\ 0 & \text{if otherwise} \end{cases}$$

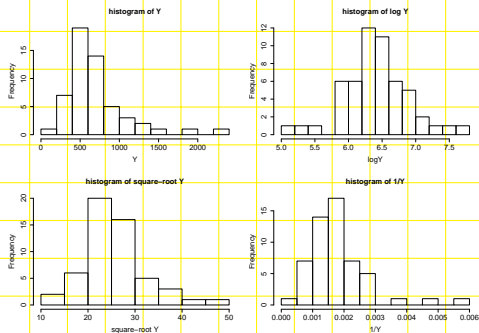
These constitute the pool of potential  $X$  variables.

We use half of the data to build the model (**training data**) and use the other half to perform model validation (**validation data**) later.

case	clotting	prognostic	enzyme	liver	age	gender	alcohol_moderate	alcohol_severe	survival
X1	X2	X3	X4	X5	X6	X7	X8	Y	
1	6.7	62	81	2.59	50	0	1	0	695
2	5.1	59	66	1.70	39	0	0	0	403
3	7.4	57	83	2.16	55	0	0	0	710
...	...	...	...	...	...	...	...	...	...
53	6.4	59	85	2.33	63	0	1	0	550
54	8.8	78	72	3.20	56	0	0	0	651

Explore the response variable:

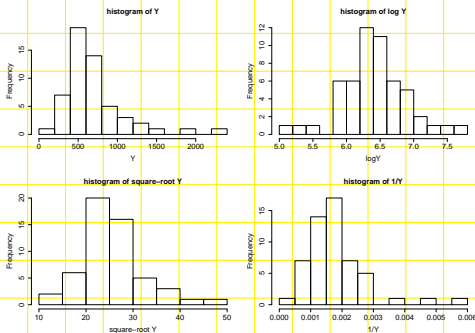
Figure: Distribution of survival times (Y)



Distribution of survival time is                     , so we may want to consider a transformation to make it more normal like. The transformation seems to work the best in this case.

Explore the response variable:

Figure: Distribution of survival times (Y)

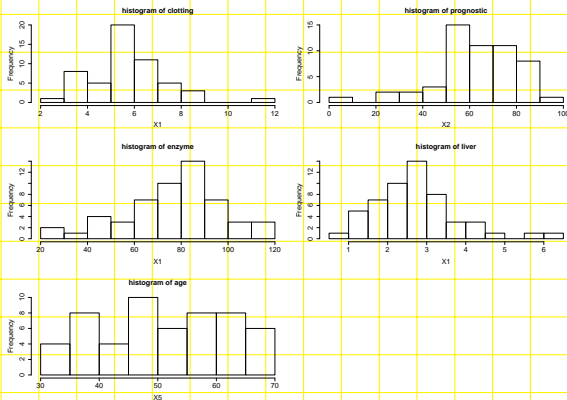


Distribution of survival time is right-skewed, so we may want to consider a transformation to make it more normal like. The logarithm transformation seems to work the best in this case.



Explore the predictor variables:

**Figure:** Distributions of quantitative predictor variables



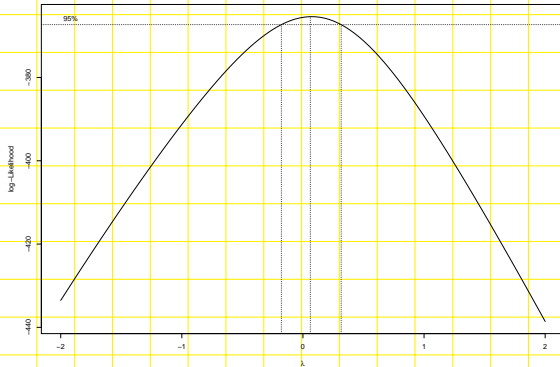
## Preliminary investigation: fit a first-order model with all variables to explore whether transformations, etc., are needed.

```
fit1=lm(Y~., data=data.o) ##fit a first-order model with all X variables
plot(fit1, which=1) ## residuals vs. fitted shows nonlinearity and nonconstant variance
plot(fit1, which=2) ##residuals Q-Q shows heavy right tail
library(MASS)
boxcox(fit1) ### boxcox procedure suggests logarithm transformation of the response variable.

fit2=lm(log(Y)~., data=data.o) ##fit a first-order model with all X variables and log Y as response
plot(fit2, which=1) ## no obvious nonlinearity and nonconstant variance
plot(fit2, which=2) ## no obvious departure from normality

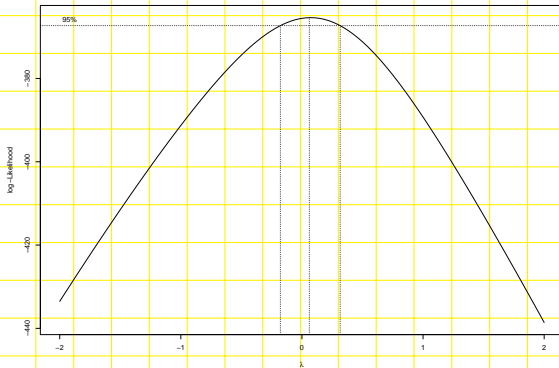
> summary(fit2)
Call:
lm(formula = log(Y) ~ ., data = data.o)
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.050949    0.251741  16.092 < 2e-16 ***
X1           0.068551    0.025420   2.697  0.00982 **
X2           0.013459    0.001947   6.913  1.37e-08 ***
X3           0.014948    0.001809   8.261  1.44e-10 ***
X4           0.007931    0.046706   0.170  0.86592
X5          -0.003567    0.002751  -1.296  0.20145
X6           0.084151    0.060746   1.385  0.17279
X7           0.057313    0.067480   0.849  0.40019
X8           0.388190    0.088374   4.393  6.73e-05 ***
---
```

Figure: Plot for boxcox procedure



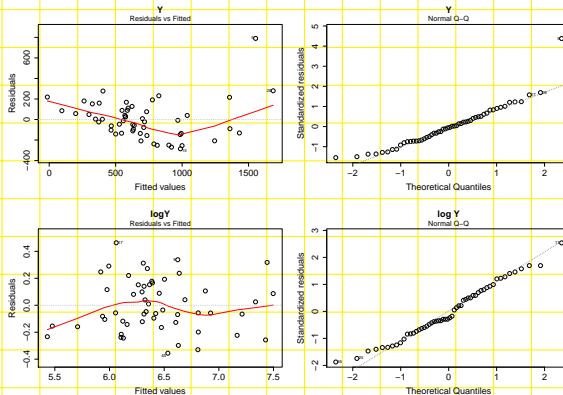
Boxcox procedure suggests a transformation  
( $\lambda =$  ) for the response variable (consistent with the exploratory  
data analysis).

Figure: Plot for boxcox procedure



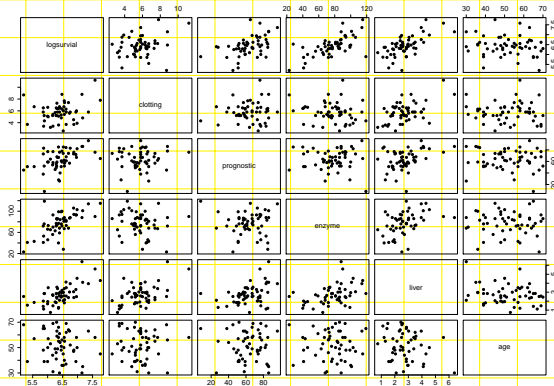
Boxcox procedure suggests a logarithm transformation ( $\lambda = 0$ ) for the response variable (consistent with the exploratory data analysis).

**Figure:** Residual plots of models using survival time and log-survival as response variable, respectively.



Logarithm transformation is able to remedy departures from model assumptions.

Figure: Pairwise scatter plots among quantitative variables

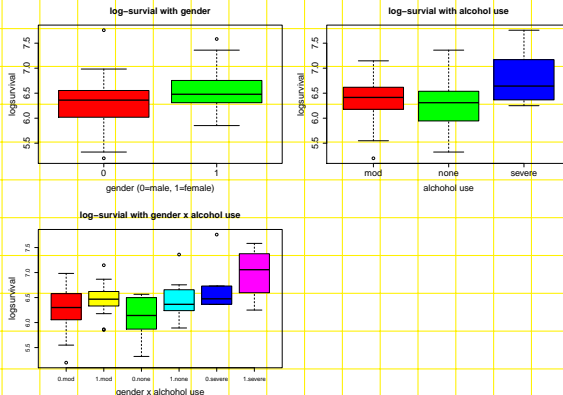


No obvious nonlinearity in regression relations. Log-survival is positively correlated with clotting, prognostic, enzyme, liver, and is weakly negatively correlated with age.

## Pairwise correlation matrix.

```
> temp=cor(cbind(log(data.o$Y),data.o$X1, data.o$X2, data.o$X3, data.o$X4, data.o$X5))
> round(temp,2)
logsurvial clotting prognostic enzyme liver age
logsurvial      1.00      0.25      0.47  0.65  0.65 -0.15
clotting         0.25      1.00      0.09 -0.15  0.50 -0.02
prognostic       0.47      0.09      1.00 -0.02  0.37 -0.05
enzyme          0.65     -0.15     -0.02  1.00  0.42 -0.01
liver           0.65      0.50      0.37  0.42  1.00 -0.21
age            -0.15     -0.02     -0.05 -0.01 -0.21  1.00
```

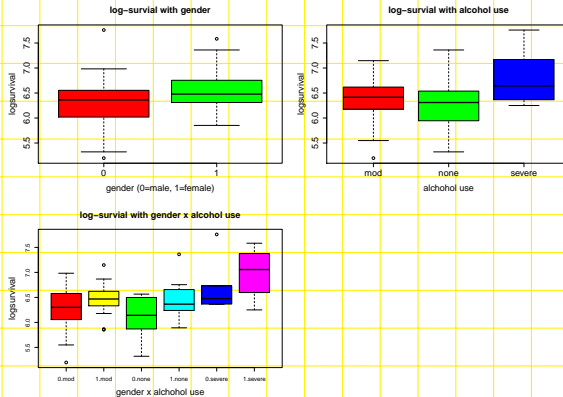
**Figure:** Distribution of log-survival within each class of gender and alcohol use.



Women tend to have longer survival and so do severe alcohol users. There is also an interaction between gender and alcohol use.



**Figure:** Distribution of log-survival within each class of gender and alcohol use.



Women tend to have longer survival and so do severe alcohol users. There is also mild interaction between gender and alcohol use.

Based on these preliminary investigations, we decide to:

- use log-survival as the response variable
- not include any interaction terms: this can be further examined by plotting residuals against various interaction terms.

Next, we should examine whether all predictors are needed or a subset of them is adequate in explaining log-survival  $\implies$  **model selection**.

# Model Selection

- Why is there a need for model selection?
  - Models with many  $X$  variables tend to have sampling variability. They are also hard to maintain and interpret.
  - On the other hand, omission of key  $X$  variables leads to fitted regression functions and predictions.
- The goal of model selection is to choose a subset of  $X$  variables which balances between  
, i.e., achieves

# Model Selection

- Why is there a need for model selection?
  - Models with many  $X$  variables tend to have large sampling variability. They are also hard to maintain and interpret.
  - On the other hand, omission of key  $X$  variables leads to biased fitted regression functions and predictions.
- The goal of model selection is to choose a subset of  $X$  variables which balances between model variance and bias, i.e., achieves **bias-variance trade-off**.

# Correct Models vs. Good Models

- Correct models are those that contain all important  $X$  variables.
- Consequently, correct models have  $0$  model bias.
- However, a correct model is not necessarily a good model because it may include irrelevant variables which lead to overfitting.
- A good model should contain only relevant variables (no bias), and at the same time it should have low variance (no overfitting).
- In summary, a good model achieves *bias-variance trade-off*.

## Correct Models vs. Good Models

- Correct models are those that contain all important  $X$  variables.
- Consequently, correct models have little model bias.
- However, a correct model is not necessarily a good model because it may include too many nuisance variables which lead to large sampling variability and overfitting.
- A good model should contain all important  $X$  variables (correct: little bias), and at the same time it should have few nuisance variables (simple: small variability).
- In summary, a good model achieves *bias-variance trade-off*.

*Example.* The response variable  $Y$  is generated by:

$$Y_i = 1 + 2X_1 + 3X_2 + \epsilon_i, \quad \epsilon_i \sim i.i.d. (0, \sigma^2).$$

- Any model contains  $(X_1, X_2)$  is a correct model, e.g.,  $\{X_1, X_2\}, \{X_1, X_2, X_1X_2\}, \{X_1, X_2, X_1^2, X_2^2\}, \{X_1, X_2, X_3, X_4, X_5\}$ .
  - These models lead to  $\infty$  of the mean response and error variance.
  - However, some of them may have  $\infty$  model variance such that the estimates behave erratically with even very small perturbation of the data. Such models, although correct, are  $\infty$ .
- On the other hand, the models  $\{X_1\}$  or  $\{X_2\}$  both have an important  $X$  variable being omitted and thus they lead to  $\infty$ .

*Example.* The response variable  $Y$  is generated by:

$$Y_i = 1 + 2X_1 + 3X_2 + \epsilon_i, \quad \epsilon_i \sim_{i.i.d.} (0, \sigma^2).$$

- Any model contains  $(X_1, X_2)$  is a correct model, e.g.,  $\{X_1, X_2\}, \{X_1, X_2, X_1X_2\}, \{X_1, X_2, X_1^2, X_2^2\}, \{X_1, X_2, X_3, X_4, X_5\}$ .
  - These models lead to unbiased estimates of the mean response and error variance.
  - However, some of them may have very large model variance such that the estimates behave erratically with even very small perturbation of the data. Such models, although correct, are not good or useful.
- On the other hand, the models  $\{X_1\}$  or  $\{X_2\}$  both have an important  $X$  variable being omitted and thus they lead to substantial model bias and underfitting.



## Model Variance

Assume the response vector  $\mathbf{Y}$  has  $\text{Var}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$ . Let  $\mathbb{M} = \mathbb{M}(X_1, \dots, X_{p-1})$  be an arbitrary model (**not necessarily** a correct model) with design matrix  $\mathbf{X}$  on these  $n$  cases.

- The (in-sample) variances of  $\mathbb{M}$  are the variances of its fitted values  $\hat{\mathbf{Y}}$ , where

$$\hat{\mathbf{Y}} = H(\mathbf{X})\mathbf{Y}, \quad \text{Var}(\hat{\mathbf{Y}}) = \sigma^2 H(\mathbf{X}), \quad H(\mathbf{X}) = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$$

- The overall model variance:
- Therefore, larger models always have overall  
(in-sample) variance, whether they are correct or not.

## Model Variance

Assume the response vector  $\mathbf{Y}$  has  $\text{Var}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$ . Let  $\mathbb{M} = \mathbb{M}(X_1, \dots, X_{p-1})$  be an arbitrary model (**not necessarily** a correct model) with design matrix  $\mathbf{X}$  on these  $n$  cases.

- The (in-sample) variances of  $\mathbb{M}$  are the variances of its fitted values  $\hat{\mathbf{Y}}$ , where

$$\hat{\mathbf{Y}} = H(\mathbf{X})\mathbf{Y}, \quad \text{Var}(\hat{\mathbf{Y}}) = \sigma^2 H(\mathbf{X}), \quad H(\mathbf{X}) = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$$

- The overall model variance:

$$\text{Var}_{in}(\mathbb{M}) := \sum_{i=1}^n \text{Var}(\hat{Y}_i) = \text{Tr}(\text{Var}(\hat{\mathbf{Y}})) = \sigma^2 \text{Tr}(H(\mathbf{X})) = p\sigma^2.$$

- Therefore, larger models always have larger overall (in-sample) variance, whether they are correct or not.

# Model Bias

- The (in-sample) biases of a model  $\mathbb{M} = \mathbb{M}(X_1, \dots, X_{p-1})$  are the biases of the fitted values:

$$\text{bias}_{in}(\mathbb{M}) = E(\hat{\mathbf{Y}}) - E(\mathbf{Y}) = (H(\mathbf{X}) - \mathbf{I})\mu, \quad \hat{\mathbf{Y}} = H(\mathbf{X})\mathbf{Y}, \quad \mu = E(\mathbf{Y}).$$

- The biases depend on

- If  $\mathbb{M}$  is a correct model, then  $\text{bias}_{in}(\mathbb{M}) =$  .

# Model Bias

- The (in-sample) biases of a model  $\mathbb{M} = \mathbb{M}(X_1, \dots, X_{p-1})$  are the biases of the fitted values:

$$\text{bias}_{in}(\mathbb{M}) = E(\hat{\mathbf{Y}}) - E(\mathbf{Y}) = (H(\mathbf{X}) - \mathbf{I})\boldsymbol{\mu}, \quad \hat{\mathbf{Y}} = H(\mathbf{X})\mathbf{Y}, \quad \boldsymbol{\mu} = E(\mathbf{Y}).$$

- The biases depend on how well the space  $\langle \mathbf{X} \rangle$  approximates the mean response vector:

$$\boldsymbol{\mu} = E(\mathbf{Y}) = \boldsymbol{\mu}_X + \boldsymbol{\mu}_{X^\perp}, \quad \boldsymbol{\mu}_X \in \langle \mathbf{X} \rangle, \quad \boldsymbol{\mu}_{X^\perp} \in \langle \mathbf{X} \rangle^\perp.$$

$$\text{bias}_{in}(\mathbb{M}) = -\boldsymbol{\mu}_{X^\perp}, \quad \|\text{bias}_{in}(\mathbb{M})\|_2^2 = \boldsymbol{\mu}^T (\mathbf{I} - H(\mathbf{X})) \boldsymbol{\mu} = \|\boldsymbol{\mu}_{X^\perp}\|_2^2 \leq \|\boldsymbol{\mu}\|_2^2.$$

- If  $\mathbb{M}$  is a correct model, then  $\text{bias}_{in}(\mathbb{M}) = \mathbf{0}$  because:

$$\boldsymbol{\mu} = E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \in \langle \mathbf{X} \rangle, \quad \text{so,} \quad \boldsymbol{\mu}_{X^\perp} = \mathbf{0}.$$

# Mean-Squared-Estimation-Errors of a Model

- **The msee equals to variance plus squared bias, i.e.,**

$$msee_h(\mathbb{M}) = Var_h(\mathbb{M}) + bias_h^2(\mathbb{M}).$$

$$msee_h(\mathbb{M}) =$$

# Mean-Squared-Estimation-Errors of a Model

- **The msee equals to variance plus squared bias, i.e.,**

$$msee_h(\mathbb{M}) = \text{Var}_h(\mathbb{M}) + \text{bias}_h^2(\mathbb{M}).$$

$$\begin{aligned} msee_h(\mathbb{M}) &= E((\hat{y}_h - \mu_h)^2) = E((\hat{y}_h - E(\hat{y}_h)) + (E(\hat{y}_h) - \mu_h))^2 \\ &= \text{Var}_h(\hat{y}_h) + (E(\hat{y}_h) - \mu_h)^2 \\ &= \text{Var}_h(\mathbb{M}) + \text{bias}_h^2(\mathbb{M}). \end{aligned}$$

## E(SSE) of a Model

- $SSE = \mathbf{e}^T \mathbf{e} = \mathbf{Y}^T (\mathbf{I} - H(\mathbf{X})) \mathbf{Y}$ , is a measure of the model to the **observed data**  $\mathbf{Y}$ .
- $E(SSE)$  is affected by three factors:

$$E(SSE) =$$

- If  $\mathbb{M}$  is a correct model, then  $bias_{in}(\mathbb{M}) =$  and thus  $E(SSE) =$  and  $E(MSE) =$ .
- If  $\mathbb{M}$  is an **underfitted model**, i.e.,  $\mu = E(\mathbf{Y}) \neq \langle \mathbf{X} \rangle$ , then  $E(SSE)$  and  $E(MSE)$ .

## E(SSE) of a Model

- $SSE = \mathbf{e}^T \mathbf{e} = \mathbf{Y}^T (\mathbf{I} - H(\mathbf{X})) \mathbf{Y}$ , is a measure of *goodness-of-fit* of the model to the **observed data**  $\mathbf{Y}$ .
- $E(SSE)$  is affected by three factors: (i) model complexity  $p$ ; (ii) error variance  $\sigma^2$ ; (iii) and model bias  $bias_{in}$ .

$$\begin{aligned} E(SSE) &= E(\text{Tr}((\mathbf{I} - H(\mathbf{X})) \mathbf{Y} \mathbf{Y}^T)) = \text{Tr}((\mathbf{I} - H(\mathbf{X})) E(\mathbf{Y} \mathbf{Y}^T)) \\ &= \text{Tr}((\mathbf{I} - H(\mathbf{X})) (\sigma^2 \mathbf{I} + \boldsymbol{\mu} \boldsymbol{\mu}^T)) \\ &= (n - p) \sigma^2 + \boldsymbol{\mu}^T (\mathbf{I} - H(\mathbf{X})) \boldsymbol{\mu} \\ &= (n - p) \sigma^2 + \|bias_{in}\|_2^2 \geq (n - p) \sigma^2. \end{aligned}$$

- If  $\mathbb{M}$  is a correct model, then  $bias_{in}(\mathbb{M}) = \mathbf{0}$  and thus  $E(SSE) = (n - p) \sigma^2$  and  $E(MSE) = \sigma^2$ .
- If  $\mathbb{M}$  is an **underfitted model**, i.e.,  $\boldsymbol{\mu} = E(\mathbf{Y}) \notin \langle \mathbf{X} \rangle$ , then  $E(SSE) > (n - p) \sigma^2$  and  $E(MSE) > \sigma^2$ .



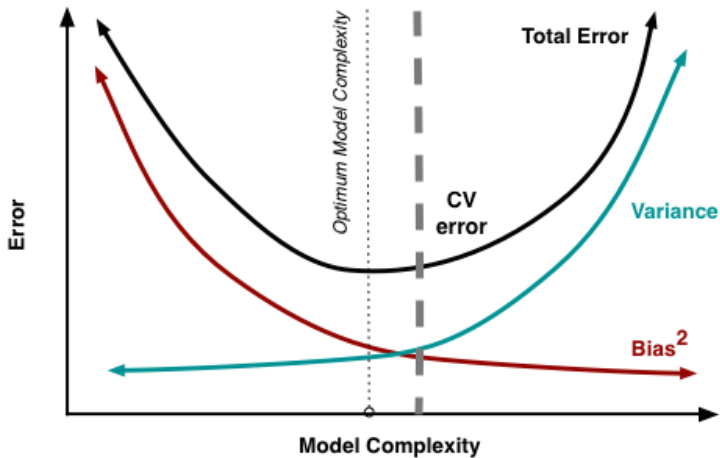
## Summary: Model Variance and Model Bias

- Larger models always have a  $E(SSE)$  overall variance.
- The overall bias of a model depends on how well the column space of its design matrix approximates the mean response vector. Correct models are unbiased.
- For two correct models, the larger model always has a  $E(SSE)$ , so they tend to overfit the observed data. On the other hand, the larger models have higher overall variance and thus they have higher overall mean-squared-estimation-error.
- Under-fit models have higher  $E(SSE)$  than correct models of the same size. So they tend to underfit the observed data. Their MSE is higher than the error variance.

## Summary: Model Variance and Model Bias

- Larger models always have a larger overall variance.
- The overall bias of a model depends on how well the column space of its design matrix approximates the mean response vector. Correct models are unbiased.
- For two correct models, the larger model always has a smaller  $E(SSE)$ , so they tend to *overfit* the observed data. On the other hand, the larger models have larger overall variance and thus they have larger overall mean-squared-estimation-error.
- Under-fit models have larger  $E(SSE)$  than correct models of the same size. So they tend to underfit the observed data. Their MSE overestimates the error variance.

## Bias-Variance Trade-off



# Key Components for Model Selection

- **Criterion to compare models:**
  - $R_a^2$ ,  $C_p$ ,  $AIC_p$ ,  $BIC_p$ ,  $Press_p$ , etc.
- **Procedure to search for good model(s):**
  - *Best subset selection*: Exhaustive search; When the number of potential  $X$  variables is not too big
  - *Stepwise regression*: Greedy search; The number of potential  $X$  variables can be large.

## Full Model vs. Candidate Model

- *Full model*: The model that contains all  $P$  potential  $X$  variables in the pool.
  - Assume the full model is a correct model.
  - It is often used to provide an unbiased estimate for the error variance.
- *Candidate model*: A model that contains a subset of  $p - 1$   $X$  variables with  $1 \leq p \leq P$ .
- The goal is to choose good model(s) (subset(s) of  $X$  variables) that balances bias and variance.

## Surgical Unit: Model Selection Criteria

Consider  $X_1, X_2, X_3, X_4$  (clotting, prognostic, enzyme, liver) as the potential pool of  $X$  variables. There are 16 sub-models.

p	intercept	X1	X2	X3	X4	sse	R <sup>2</sup>	R <sup>2</sup> _a	Cp	aic	bic	press
1	1	0	0	0	0	12.805	0.000	0.000	151.569	-75.716	-73.727	13.292
2	1	0	0	1	0	7.334	0.427	0.416	66.518	-103.811	-99.833	8.329
2	1	0	0	0	1	7.408	0.421	0.410	67.696	-103.268	-99.290	8.024
2	1	0	1	0	0	9.974	0.221	0.206	108.469	-87.205	-83.227	10.738
2	1	1	0	0	0	12.028	0.061	0.043	141.093	-77.096	-73.118	13.508
3	1	0	1	1	0	4.313	0.663	0.650	20.523	-130.479	-124.512	5.066
3	1	0	0	1	1	5.132	0.599	0.583	33.536	-121.089	-115.122	6.123
3	1	1	0	1	0	5.783	0.548	0.531	43.873	-114.644	-108.677	6.989
3	1	0	1	0	1	6.620	0.483	0.463	57.175	-107.342	-101.375	7.474
3	1	1	0	0	1	7.299	0.430	0.408	67.961	-102.070	-96.103	8.472
3	1	1	1	0	0	9.437	0.263	0.234	101.937	-88.194	-82.227	11.055
4	1	1	1	1	0	3.109	0.757	0.743*	3.388*	-146.161*	-138.205*	3.914*
4	1	0	1	1	1	3.615	0.718	0.701	11.434	-138.011	-130.055	4.598
4	1	1	0	1	1	4.970	0.612	0.589	32.960	-120.823	-112.867	6.209
4	1	1	1	0	1	6.568	0.487	0.456	58.358	-105.763	-97.807	7.902
5	1	1	1	1	1	3.084	0.759*	0.739	5.000	-144.587	-134.642	4.069

Within each subset size, models are sorted in ascending  $SSE$ .

Consequently, within each subset size,  $R_p^2, R_{a,p}^2$  are from the largest to the smallest and  $C_p, BIC_p, AIC_p$  are from the smallest to the largest.  $Press_p$  may not be monotone with  $SSE$ .

## Mallows' $C_p$ Criterion

Mallows'  $C_p$  for a model with  $p$  regression coefficients:

$$C_p := \frac{SSE_p}{\hat{\sigma}^2} - (n - 2p).$$

- $n$ : sample size (constant across models).
- $SSE_p$ : error sum of squares of the candidate model.
- $\hat{\sigma}^2$ : an estimator of the error variance  $\sigma^2$ . E.g.,

$$\hat{\sigma}^2 = MSE_{\text{full model}} = MSE(X_1, \dots, X_{P-1}).$$

- $\hat{\sigma}^2$  is unbiased due to the assumption that the full model contains  $X$  variables so that
- $C_p$  of the full model is always

## Mallows' $C_p$ Criterion

Mallows'  $C_p$  for a model with  $p$  regression coefficients:

$$C_p := \frac{SSE_p}{\hat{\sigma}^2} - (n - 2p).$$

- $n$ : sample size (constant across models).
- $SSE_p$ : error sum of squares of the candidate model.
- $\hat{\sigma}^2$ : an unbiased estimator of the error variance  $\sigma^2$ . E.g.,

$$\hat{\sigma}^2 = MSE_{\text{full model}} = MSE(X_1, \dots, X_{P-1}).$$

- $\hat{\sigma}^2$  is unbiased due to the assumption that the full model contains all important  $X$  variables so that it has no model bias.
- $C_p$  of the full model is always  $P$ .



## Mallows' $C_p$ as an Estimator of msee

Let  $\mathbb{M} = \mathbb{M}(X_1, \dots, X_{p-1})$  be a model. Then:

So  $C_p$  can be viewed as an estimator of the

.

## Mallows' $C_p$ as an Estimator of msee

Let  $\mathbb{M} = \mathbb{M}(X_1, \dots, X_{p-1})$  be a model. Then:

$$\begin{aligned} E(C_p(\mathbb{M})) &\approx \frac{E(\text{SSE}(\mathbb{M}))}{\sigma^2} - (n - 2p) \\ &= \frac{(n - p)\sigma^2 + \|\text{bias}_{in}(\mathbb{M})\|_2^2}{\sigma^2} - (n - 2p) \\ &= \frac{p\sigma^2 + \|\text{bias}_{in}(\mathbb{M})\|_2^2}{\sigma^2} \\ &= \frac{\text{Var}_{in}(\mathbb{M}) + \|\text{bias}_{in}(\mathbb{M})\|_2^2}{\sigma^2} = \frac{\text{msee}_{in}(\mathbb{M})}{\sigma^2}. \end{aligned}$$

So  $C_p$  can be viewed as an estimator of the overall (in-sample) mean-squared-estimation-error divided by the error variance.

## How to Use $C_p$ ?

- If a model has no (in-sample) bias, i.e.,  $\text{bias}_{in}(\mathbb{M}) = \mathbf{0}$ , then  $E(C_p)$  tends to be close to  $p$ . Otherwise  $E(C_p)$  tends to be greater than  $p$ .
- When  $C_p$  is plotted against  $p$ , then models with low bias will tend to fall near the diagonal line  $C_p = p$ .
- On the other hand, models with high bias will tend to fall considerably above this line.
- **We should look for models with (i) the  $C_p$  value not far above  $p$  and (ii) small  $C_p$  value.** Such models have low bias and low number of  $X$  variables (thus low model variance).
  - Surgical unit. The model with  $X_1, X_2, X_3$  has  $C_p = 3.38 < p = 4$ , indicating little or no bias. Its  $C_p$  value is also the smallest among all models being considered.

## How to Use $C_p$ ?

- If a model has no (in-sample) bias, i.e.,  $bias_{in}(\mathbb{M}) = \mathbf{0}$ , then  $E(C_p) \approx p$ . Otherwise  $E(C_p)$  tends to be larger than  $p$ .
- When  $C_p$  is plotted against  $p$ , then models with little bias will tend to fall near the diagonal line  $C_p = p$ .
- On the other hand, models with substantial bias will tend to fall considerably above this line.
- **We should look for models with (i) the  $C_p$  value not far above  $p$  and (ii) small  $C_p$  value.** Such models have small bias and small number of  $X$  variables (thus small model variance).
  - Surgical unit. The model with  $X_1, X_2, X_3$  has  $C_p = 3.38 < p = 4$ , indicating little or no bias. Its  $C_p$  value is also the smallest among all models being considered.