Stat 206: Linear Models

Lecture 12

Nov. 9, 2015

Recall: LS Fitted Regression Coefficients as Partial Coefficients

The LS fitted regression coefficients $\hat{\beta}$ are indeed partial coefficients.

- Consider $p-1$ X variables in the model. Let $\hat{\beta}_j$ be the LS fitted regression coefficient for X_j.
- Then $\hat{\beta}_j$ equals to the LS fitted regression coefficient when regressing the residuals $e(Y|X_{-(j)}) = Y - \hat{Y}(X_{-(j)})$ to the residuals $e(X_j|X_{-(j)}) = X_j - \hat{X}_j(X_{-(j)})$, where $X_{-(j)} = \{X_l : 1 \leq l \neq j \leq p\}$.
- The coefficient of partial determination $R^2_{Y,j|1,\cdots,j-1,j+1,\cdots,p-1}$ is the coefficient of simple determination by regressing $Y - \hat{Y}(X_{-(j)})$ to the residuals $X_j - \hat{X}_j(X_{-(j)})$.

Added-Variable Plots

Added-variable plot of X_j: Check for X_j.

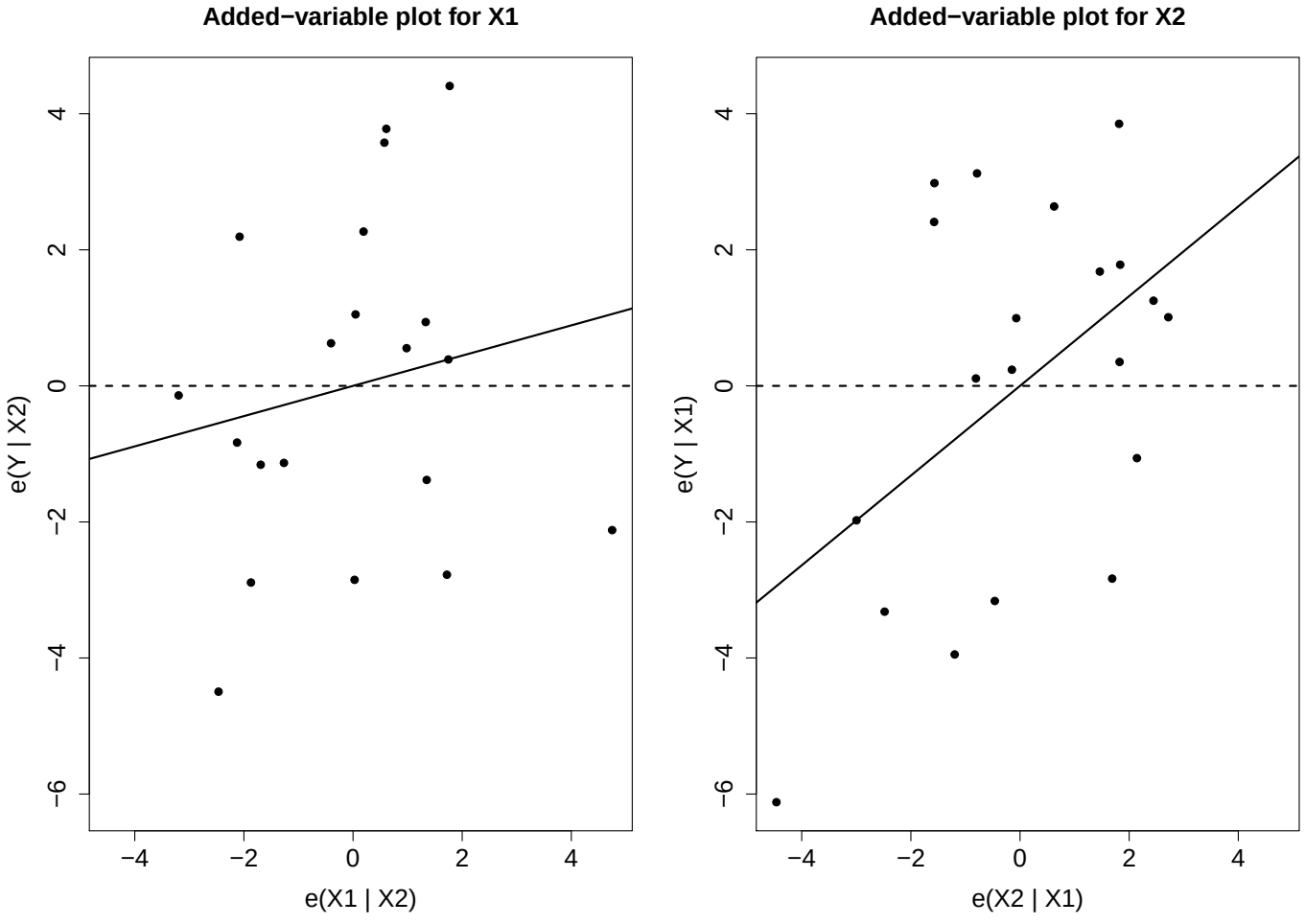
- Both the response variable Y and X_j are regressed onto the rest of the X variables, denoted by $X_{-(j)}$, in the model.
- The residuals reflect the part of Y (X_j) that with the rest of the X variables.
- The plot of these two sets of residuals against each other:
  - shows the of X_j in reducing the residual variability in Y after accounting for .
  - provides information about the nature of the marginal effect of X_j on Y, e.g., linear or curvilinear.

Added-Variable Plots

Added-variable plot of X_j: Check model adequacy for X_j.

- Both the response variable Y and X_j are regressed onto the rest of the X variables, denoted by $X_{-(j)}$, in the model.
- The residuals reflect the part of Y (X_j) that is not linearly associated with the rest of the X variables.
- The plot of these two sets of residuals against each other:
  - shows the marginal importance of X_j in reducing the residual variability in Y after accounting for the linear effects in the rest of the X variables.
  - provides information about the nature of the marginal effect of X_j on Y, e.g., linear or curvilinear.

Figure: Body Fat $Y \sim X_1, X_2$: Added-variable plots



Which plot corresponds to a larger coefficient of determination?

Model 3: $Y \sim X_1, X_2$.

- Added-variable plot for X_1 implies that X_1 is of
in explaining Y when X_2 is already in the model. *Why?* This is consistent with $R^2_{Y1|2} = 3.1\%$.
- Added-variable plot for X_2 shows that X_2 is of
in explaining Y when X_1 is already in the model. From previous slides, $R^2_{Y2|1} = 23.2\%$. It also shows that a term of X_2 in the model is adequate.

Model 3: $Y \sim X_1, X_2$.

- Added-variable plot for X_1 implies that X_1 is of little additional help in explaining Y when X_2 is already in the model. *Why?* This is consistent with $R^2_{Y1|2} = 3.1\%$.
- Added-variable plot for X_2 shows that X_2 is of some help in explaining Y when X_1 is already in the model. From previous slides, $R^2_{Y2|1} = 23.2\%$. It also shows that a linear term of X_2 in the model is adequate.

Body Fat: Model 3

```
> summary(fit3)

Call:
lm(formula = Y ~ X1 + X2, data = fat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.1742     8.3606  -2.293   0.0348 *
X1            0.2224     0.3034   0.733   0.4737
X2            0.6594     0.2912   2.265   0.0369 *
---

Residual standard error: 2.543 on 17 degrees of freedom
Multiple R-squared: 0.7781,     Adjusted R-squared: 0.7519
F-statistic:  29.8 on 2 and 17 DF,  p-value: 2.774e-06

> anova(fit3)
Analysis of Variance Table

Response: Y
Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 352.27  352.27 54.4661 1.075e-06 ***
X2         1  33.17   33.17  5.1284    0.0369 *
Residuals 17 109.95    6.47
```

back

Overview of Model-Building

- Data collection and processing.
- Exploratory data analysis and preliminary investigation.
- Model selection.
- Model diagnostic and validation.

Data Collection

- Observational studies.
  - Exploratory: Search for predictors that might be related to the response variable. Often a large number of predictors are considered.
  - Confirmatory: Confirm or disproof existing hypotheses.
- Controlled experiments.
  - Experimenters set the levels of the predictors and randomly assign a treatment to each experimental unit and then record the response.
- Quality control: Are there gross errors in the data?

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 investigation.

- 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 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: blooding 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:

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

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

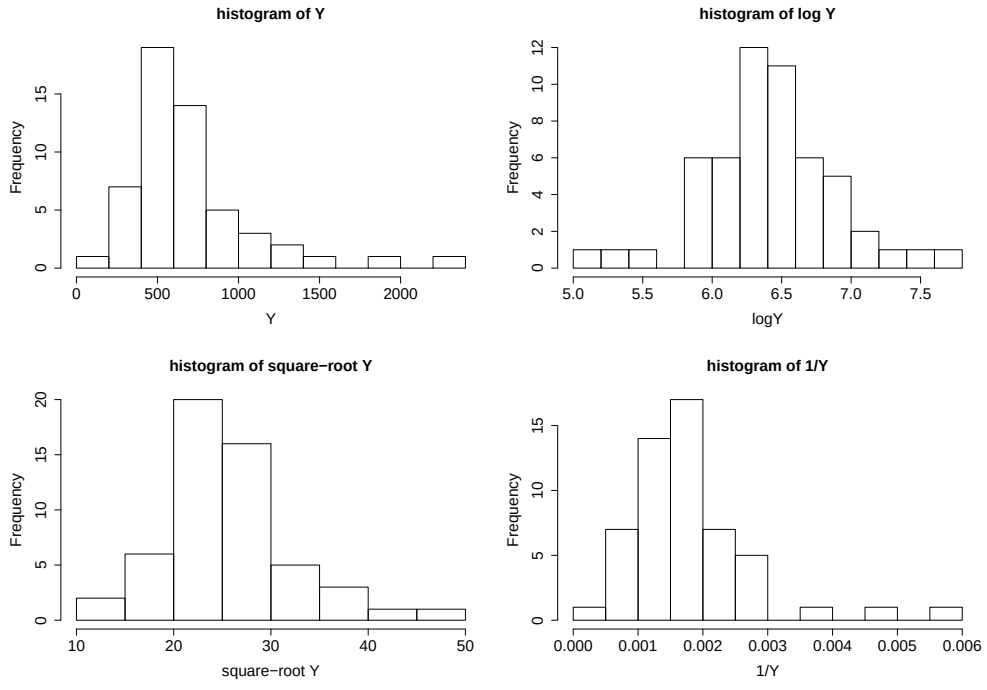
These constitute the pool of potential X variables.

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.

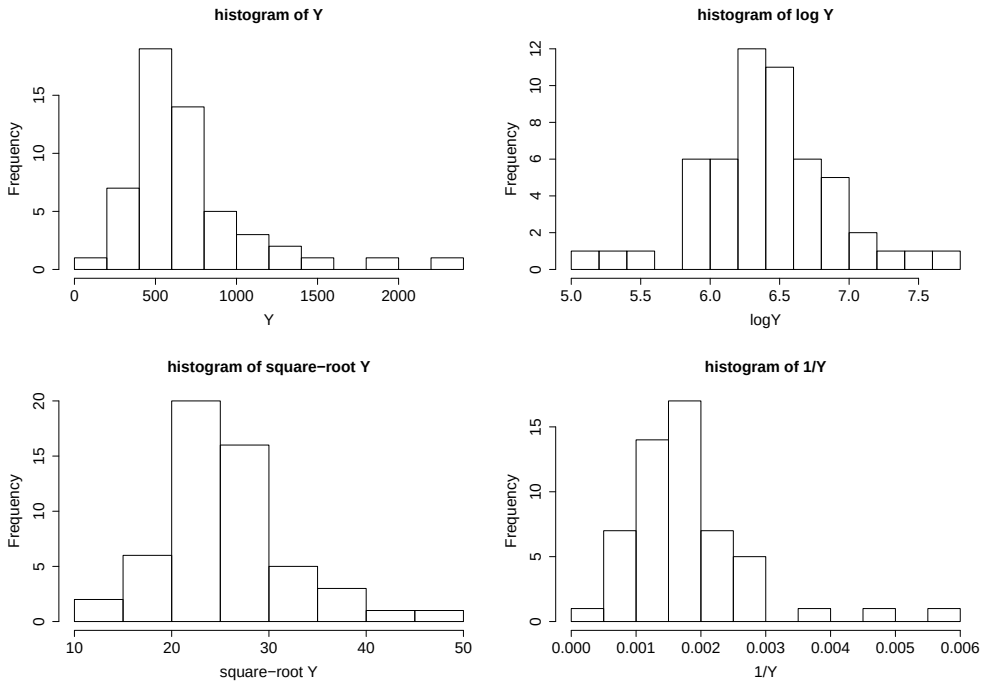
Figure: Distribution of survival times (Y)



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.

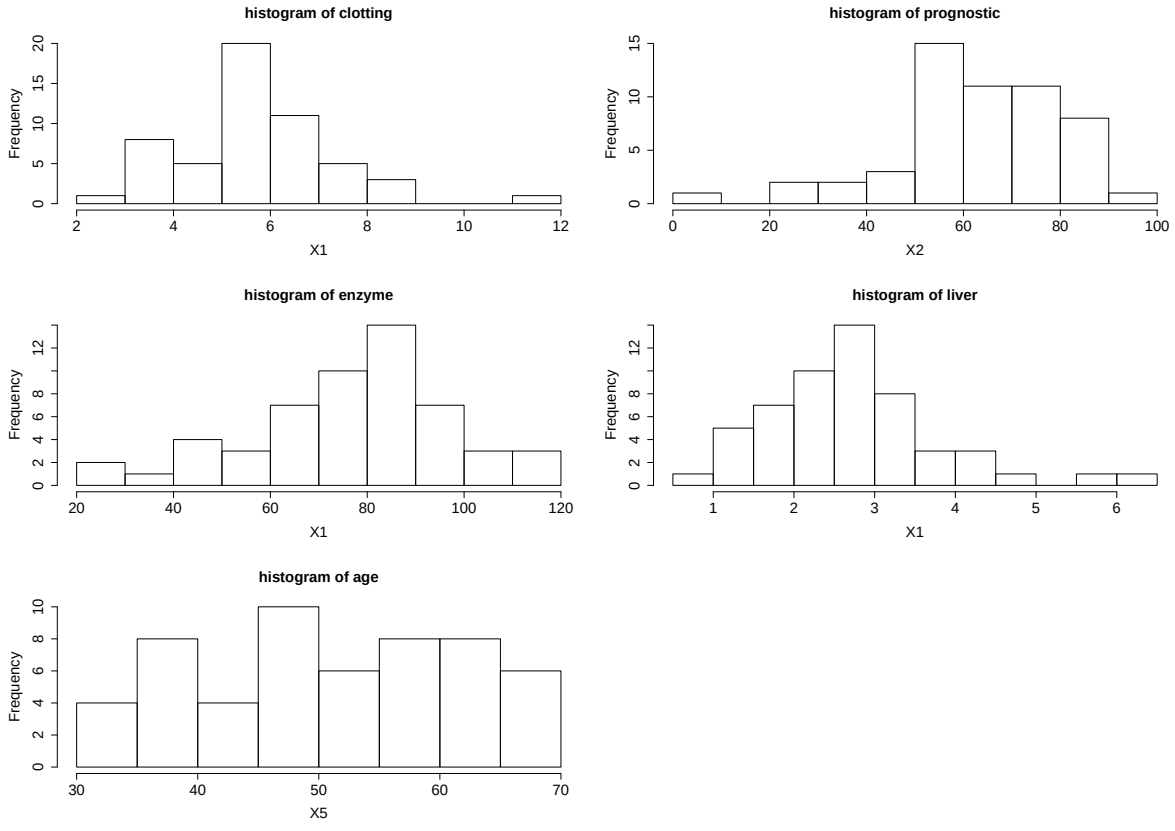
Figure: Distribution of survival times (Y)



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.

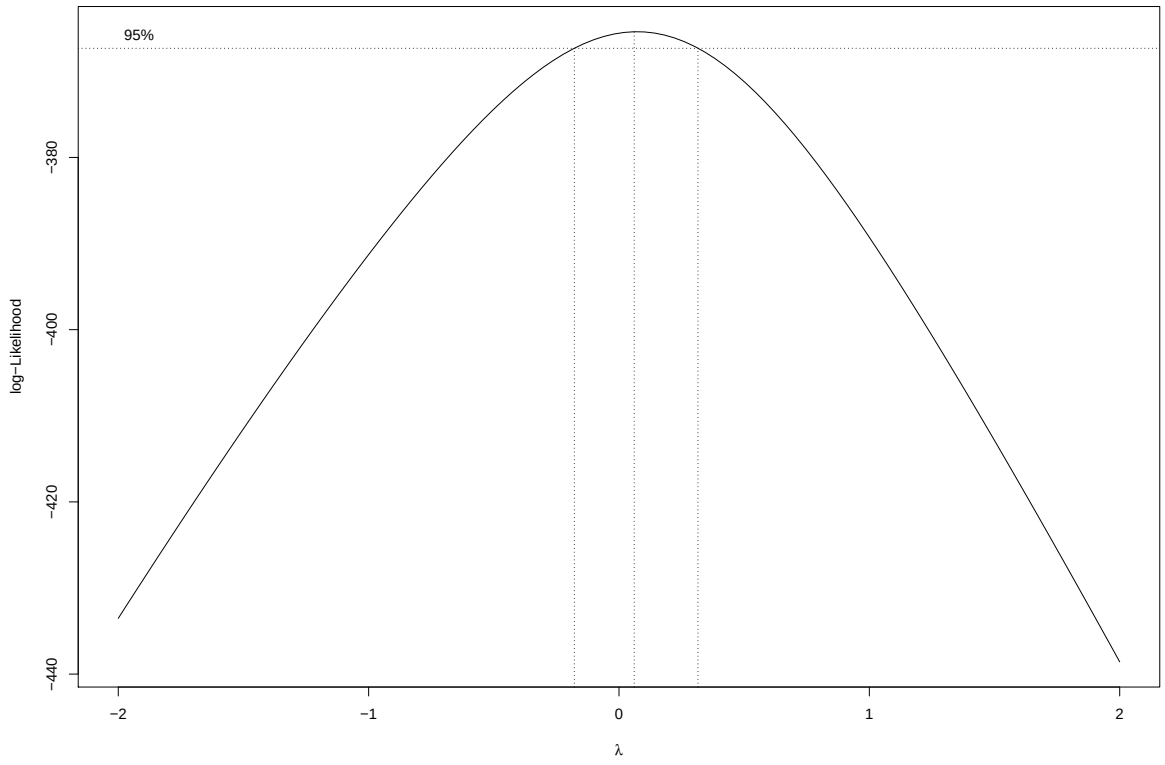
Figure: Distributions of quantitative predictor variables



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

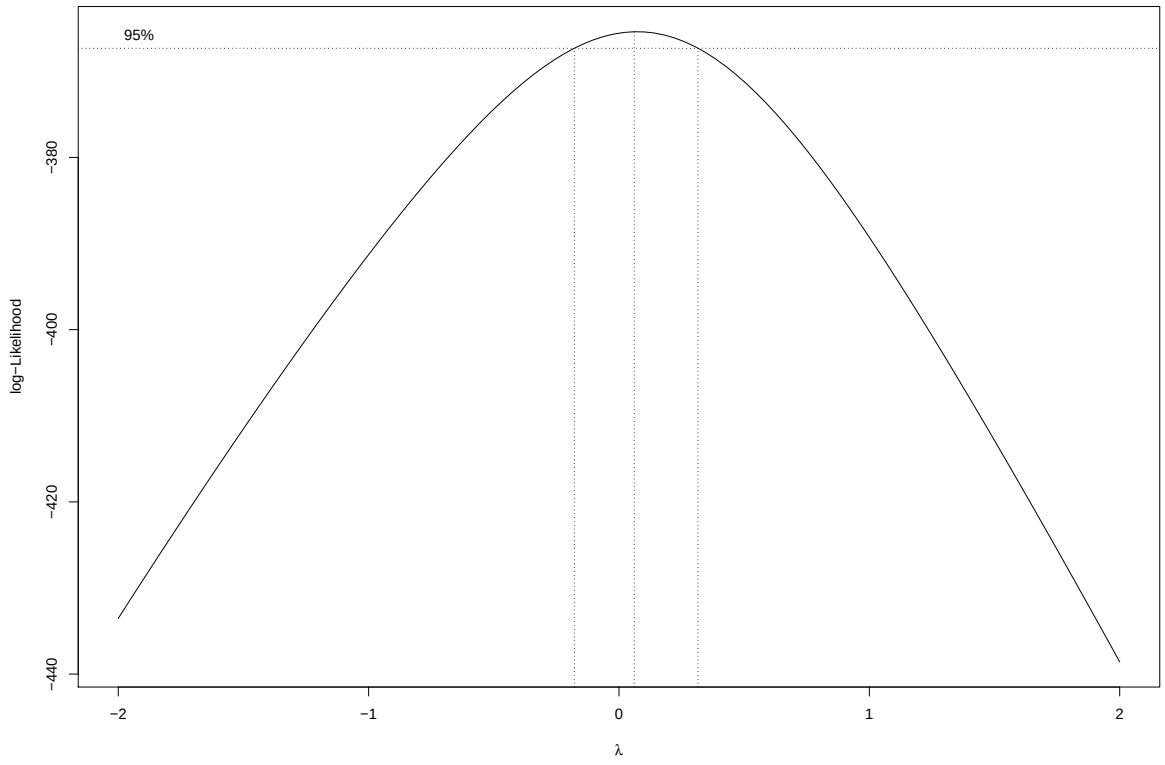
```
fit1=lm(Y~., data=data.o) ##fit a first-order model with all X variables
plot(fit1, which=1)  ## residuals vs. fitted shows nonlinearity and nonconstant variance
plot(fit1, which=2) ##residuals Q-Q shows heavy right tail
library(MASS)
boxcox(fit1)  ### boxcox procedure suggests logarithm transformation 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 ***
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1
Residual standard error: 0.2093 on 45 degrees of freedom
Multiple R-squared: 0.8461,     Adjusted R-squared: 0.8187
F-statistic: 30.93 on 8 and 45 DF,  p-value: 7.823e-16
```

Figure: Plot for boxcox procedure



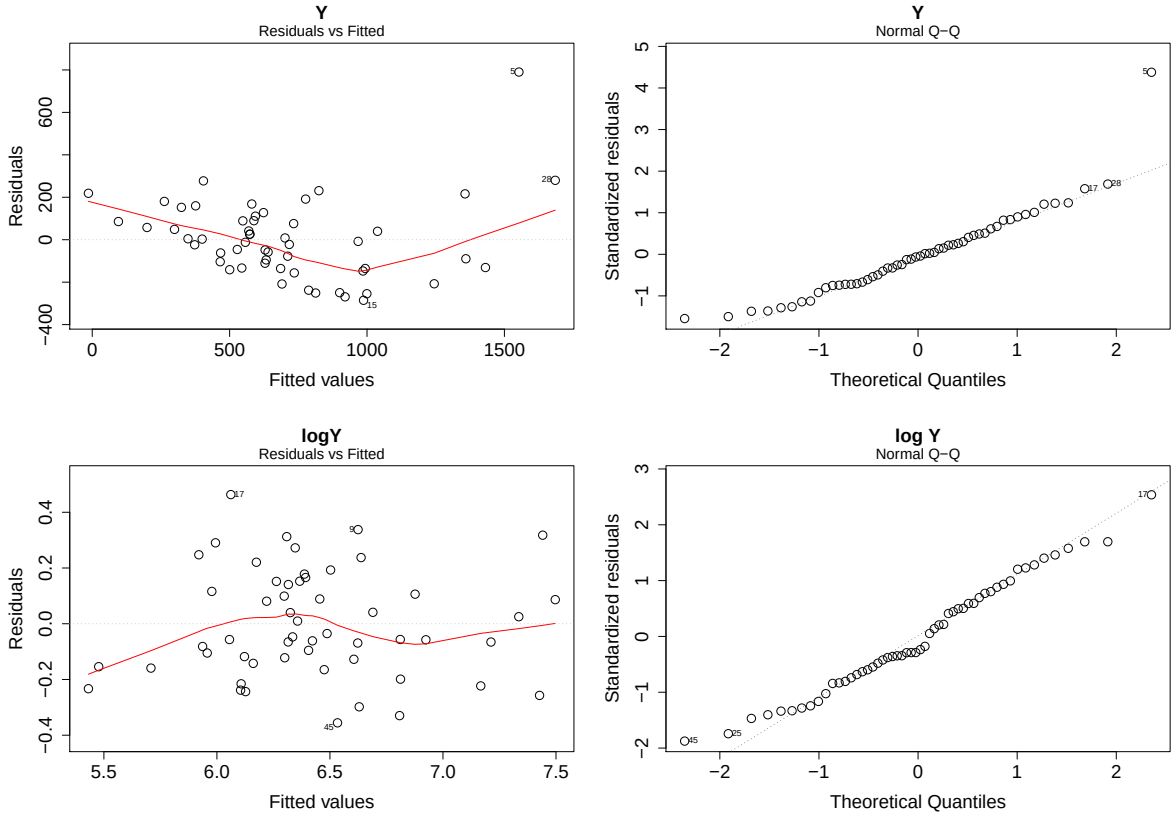
Boxcox procedure suggests a transformation ($\lambda =$) for the response variable (consistent with the exploratory data analysis).

Figure: Plot for boxcox procedure



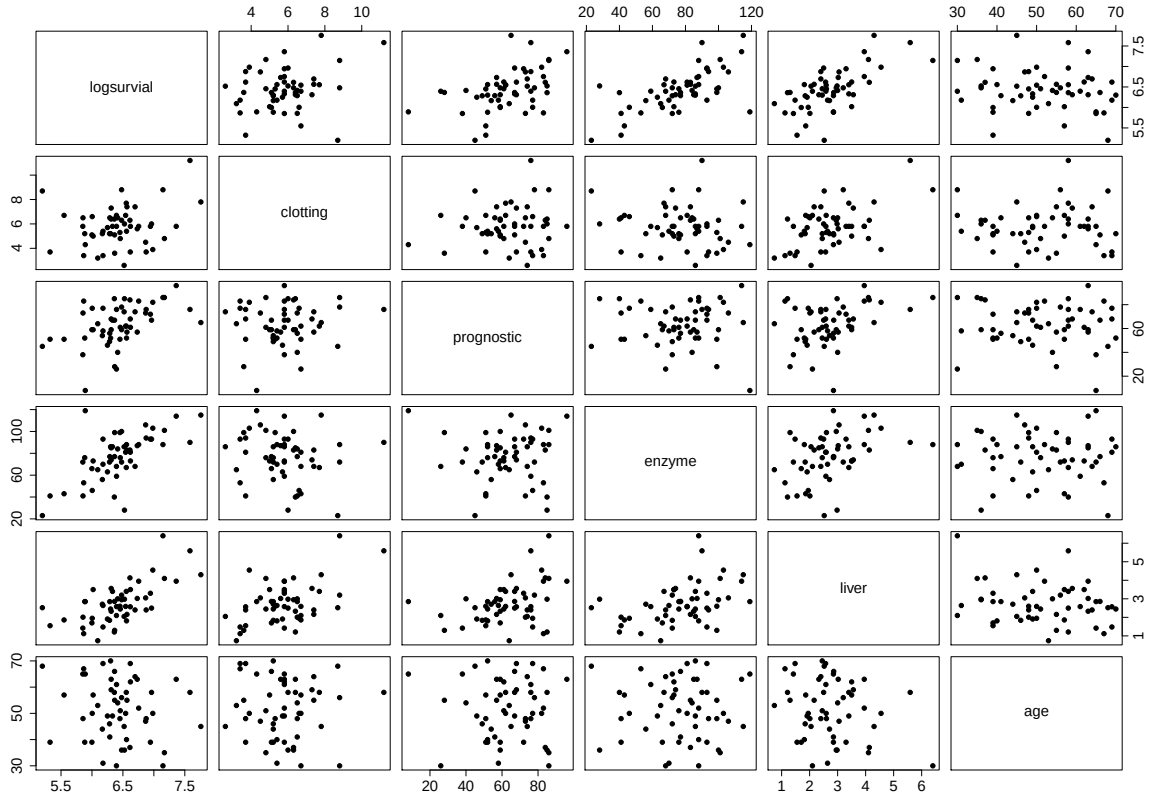
Boxcox procedure suggests a logarithm transformation ($\lambda = 0$) for the response variable (consistent with the exploratory data analysis).

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

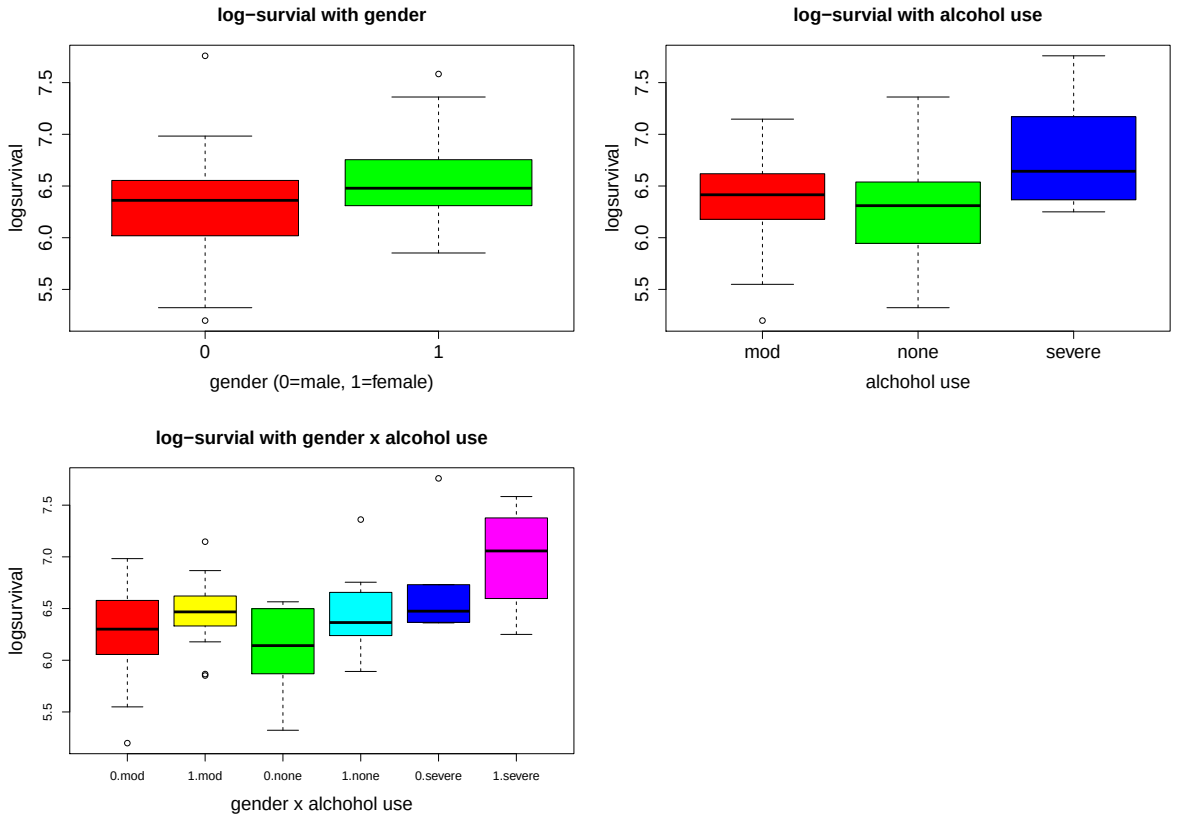


No obvious nonlinearity in regression relations. Log-survival is positively correlated with `clotting`, `prognostic`, `enzyme`, `liver`, and is weakly negatively correlated with `age`.

Pairwise correlation matrix.

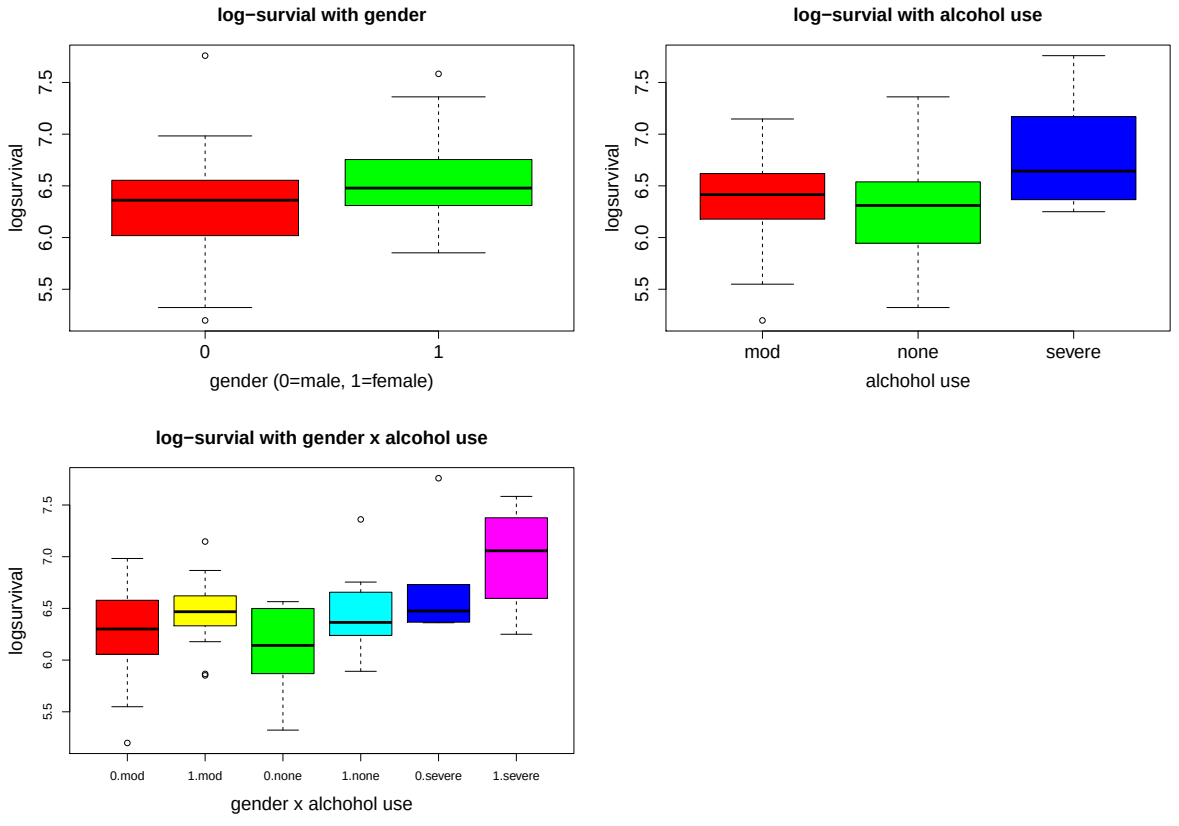
```
> temp=cor(cbind(log(data.o$Y),data.o$X1, data.o$X2, data.o$X3, data.o$X4, data.o$X5))
> round(temp,2)
logsurvial clotting prognostic enzyme liver   age
logsurvial       1.00     0.25       0.47   0.65  0.65 -0.15
clotting         0.25     1.00       0.09  -0.15  0.50 -0.02
prognostic       0.47     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 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 to 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 $\Longrightarrow$ model selection.

Model Selection

- Why is there a need for model selection?
  - Models with many X variables tend to have sampling variability and predictive abilities due to . 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 *bias-variance trade-off*.

Model Selection

- Why is there a need for model selection?
  - Models with many X variables tend to have increased sampling variability and worsened predictive abilities due to *overfitting*. 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 variance and bias, i.e., achieves *bias-variance trade-off*.

Model Diagnostic and Validation

After one or several candidate models are chosen, we should employ diagnostic tools to check:

- do model assumptions hold?
- are there residual nonlinearity or interactions?
- are there any influential outlying observations?
- is there severe multicollinearity?

Remedial measures such as transformations may be employed based on diagnostic results.

Finally, for model(s) passing the diagnostic checks, model validation should be employed to check their validity.

- *Model validity* includes stability of the fitted regression coefficients, plausibility of the fitted regression function, and the *generalizability of the model (i.e., the ability to predict)*.
- Model validation is usually done through checking the model using new data or through data splitting (*cross-validation*).

Correct Models vs. Good Models

- In regression analysis, correct models are those that contain all important X variables in terms of explaining the response variable.
- Correct models have model bias.
- A correct model is **not necessarily** a good model because it may include too many nuisance variables which could lead to

- A good model should contain all important X variables ( ), and at the same time it should have few nuisance variables ( ).
- **In practice, often none of the models under consideration is correct. Then a good model is one that balances model bias and model variance such that the overall error is small. This is called *bias-variance trade-off*.**

Correct Models vs. Good Models

- In regression analysis, correct models are those that contain all important X variables in terms of explaining the response variable.
- Correct models have little model bias.
- A correct model is not necessarily a good model because it may include too many nuisance variables which could lead to large sampling variability and overfitting.
- A good model should contain all important X variables (correct: little bias, i.e. accurate), and at the same time it should have few nuisance variables (simple: small model variance, i.e. precise).
- **In practice, often none of the models under consideration is correct. Then a good model is one that balances model bias and model variance such that the overall error is small. This is called *bias-variance trade-off*.**

Example. Suppose the response variable Y is generated according to:

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

- Then any model contains the two variables X_1, X_2 would be a correct model, e.g., $\{X_1, x_2\}, \{X_1, X_2, X_1X_2\}$ or $\{X_1, X_2, X_1^2, X_2^2\}$ or $\{X_1, X_2, X_3, X_4, X_5\}$, etc.
  - These models are correct since they give 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 model $\{X_1\}$ or the model $\{X_2\}$ both have an important X variable being omitted and therefore they lead to

Example. Suppose the response variable Y is generated according to:

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

- Then any model contains the two variables X_1, X_2 would be a correct model, e.g., $\{X_1, x_2\}, \{X_1, X_2, X_1X_2\}$ or $\{X_1, X_2, X_1^2, X_2^2\}$ or $\{X_1, X_2, X_3, X_4, X_5\}$, etc.
  - These models are correct since they give 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 model $\{X_1\}$ or the model $\{X_2\}$ both have an important X variable being omitted and therefore they lead to substantial model bias and tend to underfit the data.

Model Variance and Model Bias

Definitions and notations.

- A model $\mathbb{M} = \mathbb{M}(X_1, \cdots, X_{p-1})$ refers to a set of X variables to be used to fit a response variable Y. p is referred to as the *size or the complexity* of $\mathbb{M}$.
- A model $\mathbb{M} = \mathbb{M}(X_1, \cdots, X_{p-1})$ is called to be *nested within (or a sub-model of)* another model $\mathbb{M}^* = \mathbb{M}(X_1^*, \cdots, X_{q-1}^*)$, denoted by $\mathbb{M} \subseteq \mathbb{M}^*$, if $\{X_1, \cdots, X_{p-1}\}$ is a subset of $\{X_1^*, \cdots, X_{q-1}^*\}$.
- A model $\mathbb{M} = \mathbb{M}(X_1, \cdots, X_{p-1})$ is called a *correct model*, if there exists a vector $\boldsymbol{\beta} \in \mathbb{R}^p$ such that for any subject, indexed by h,

$$y_h = x_h^T \boldsymbol{\beta} + \epsilon_h, \quad x_h^T = (1, X_{h1}, \cdots, X_{h,p-1}),$$

where ϵ_h is a random variable with $E(\epsilon_h) = 0$ and $Var(\epsilon_h) = \sigma^2$.

Model Variance

Suppose the observations vector $\mathbf{Y}$ has $Var(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$. Let $\mathbb{M} = \mathbb{M}(X_1, \cdots, X_{p-1})$ 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

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

- The overall (in-sample) model variance:

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

- Therefore, larger models (i.e., models with more X variables) always have overall (in-sample) variance.

Model Variance

Suppose the observations vector $\mathbf{Y}$ has $Var(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$. Let $\mathbb{M} = \mathbb{M}(X_1, \cdots, X_{p-1})$ 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

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

- The overall (in-sample) model variance:

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

- Therefore, larger models (i.e., models with more X variables) always have larger overall (in-sample) variance.

Model Bias

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

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

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

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

  Then

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

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

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

Mean-Squared-Estiamtion-Errors of a Model

- The *mean-squared-estimation-error (msee)* of a model $\mathbb{M} = \mathbb{M}(X_1, \cdots, X_{p-1})$ at a case h is defined as:

  $$msee_h(\mathbb{M}) = E((\hat{y}_h - \mu_h)^2), \ \ \hat{y}_h = x_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \ \mu_h := E(y_h),$$

  and $\hat{y}_h - \mu_h$ is referred to as the *estimation error* of model $\mathbb{M}$ at case h.
- The mean-squared-estimation-error equals to variance plus squared bias, i.e.,

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

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

E(SSE) of a Model

Consider an arbitrary model $\mathbb{M} = \mathbb{M}(X_1, \cdots, X_{p-1})$.

- $SSE = \mathbf{e}^T\mathbf{e} = \mathbf{Y}^T(\mathbf{I} - H(\mathbf{X}))\mathbf{Y}$, is a measure of of the model to the **observed data Y**, where the residuals vector $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - H(\mathbf{X}))\mathbf{Y}$.
- E(SSE):

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

- If $\mathbb{M}$ is a correct model, then $bias_{in}(\mathbb{M}) = \mathbf{0}$ and thus $E(SSE) = (n-p)\sigma^2$ and $E(MSE) = \sigma^2$, where $MSE = SSE/(n-p)$.
- 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 thus $E(MSE) > \sigma^2$.

E(SSE) of a Model

Consider an arbitrary model $\mathbb{M} = \mathbb{M}(X_1, \cdots, X_{p-1})$.

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

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

- If $\mathbb{M}$ is a correct model, then $bias_{in}(\mathbb{M}) = \mathbf{0}$ and thus $E(SSE) = (n-p)\sigma^2$ and $E(MSE) = \sigma^2$, where $MSE = SSE/(n-p)$.
- 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 thus $E(MSE) > \sigma^2$.

Summary: Model Varinace and Model Bias

- Larger models always have a overall (in-sample) variance.
- The overall (in-sample) bias of a model depends on how well the column space of its design matrix approximates the mean response vector.
- Correct models are .
- For two correct models, the larger model always has a $E(SSE)$, but a (overall) in-sample variance. Thus the larger model tends to the observed data (i.e.,).
- For two correct models, the larger model always has a overall (in-sample) mean-squared-estimation-error.
- Under-fit models have $E(SSE)$ (than correct models of the same size). So they tend to the observed data. Their MSE the error variance.

Summary: Model Varinace and Model Bias

- Larger models always have a larger overall (in-sample) variance.
- The overall (in-sample) 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 $E(SSE)$, but a larger (overall) in-sample variance. Thus the larger model tends to *overfit* the observed data (i.e., smaller SSE and larger variability).
- For two correct models, the larger model always has a larger overall (in-sample) 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.

Simulation Study

Simulation setting.

- Consider one predictor variable X and the true regression function (i.e., true mean response) being :

$$f(x) = 1 - 6x + x^2 + x^3.$$

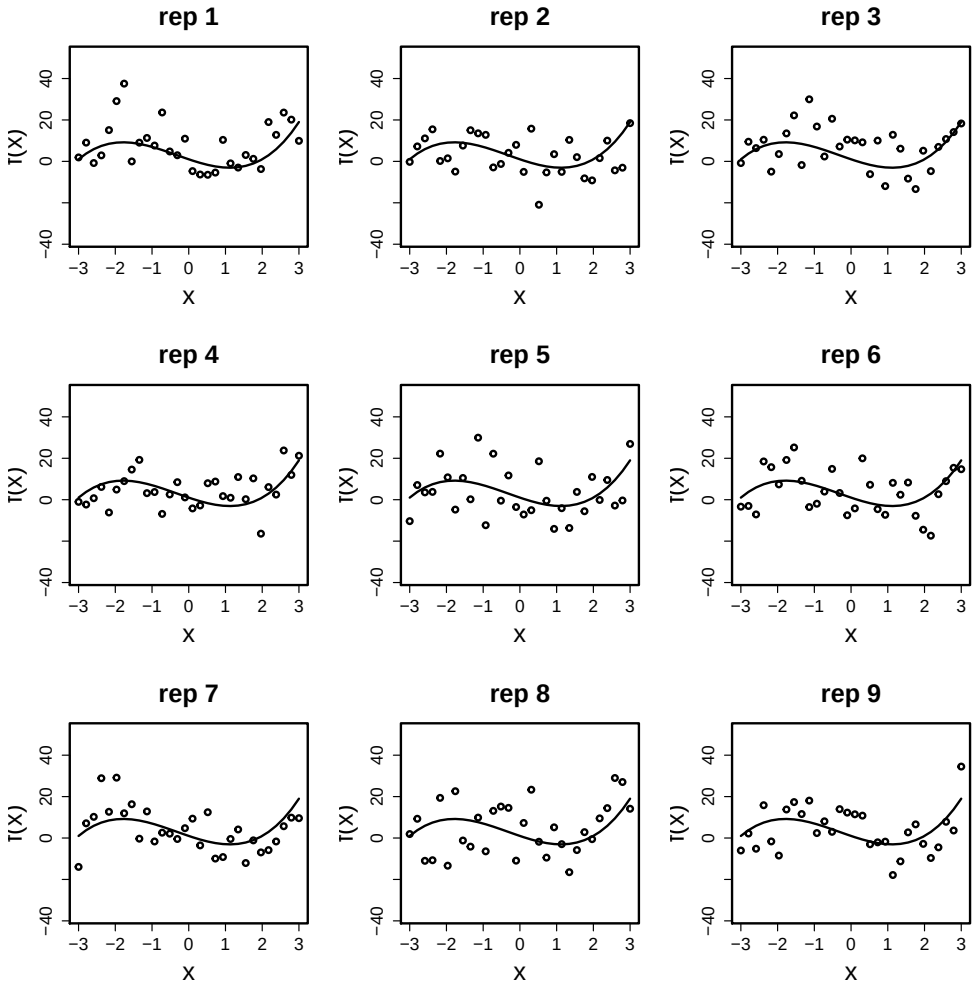
- The X values (i.e., design points) are n equally spaced points on $[-3, 3]$. These are fixed throughout the simulation.
- The observations are generated according to :

$$Y_i = f(X_i) + \epsilon_i, \quad i = 1, \cdots, n,$$

where ϵ_i are i.i.d. $N(0, \sigma^2)$, where $n = 30, \sigma = 10$ are fixed throughout the simulation.

- Repeat the generation of $\mathbf{Y}$ 1000 times. Thus get 1000 independent data sets (i.e., replicates).

Figure: Observations for nine replicates



The true mean response curve (the black curve) is the same across replicates, but the observations **Y** are changing due to *sampling variability.*

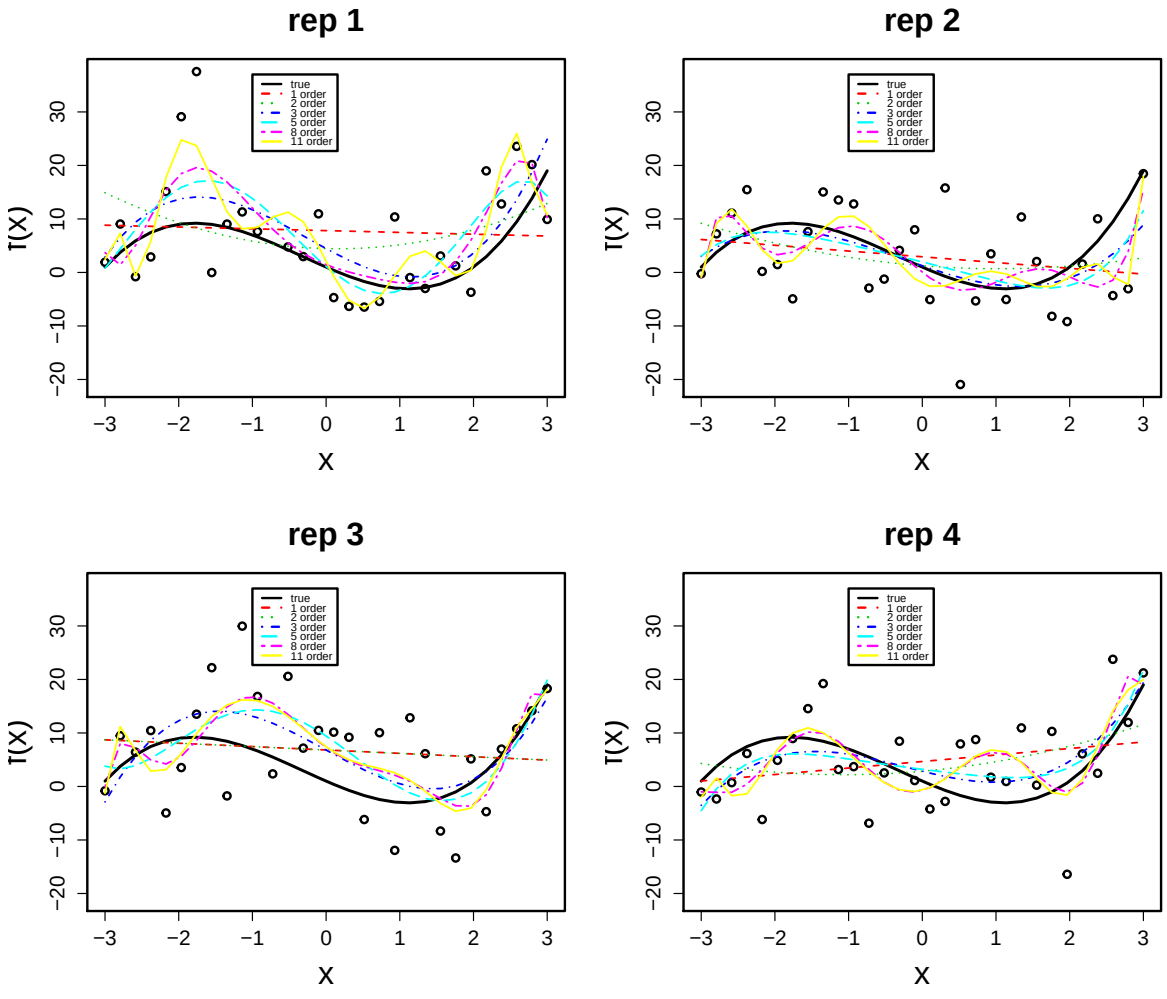
Models to be fit.

- Denote a polynomial regression model with oder l by $\mathbb{M}_l$.
- Consider the following orders: $l = 1$ (linear), $l = 2$ (quadratic), $l = 3$ (cubic), $l = 5$, $l = 8$, and $l = 11$.
- $\mathbb{M}_1$ and $\mathbb{M}_2$ are models.
- $\mathbb{M}_3, \mathbb{M}_5, \mathbb{M}_8, \mathbb{M}_{11}$ are models.
- $\mathbb{M}_3$ is the data generating model.

Models to be fit.

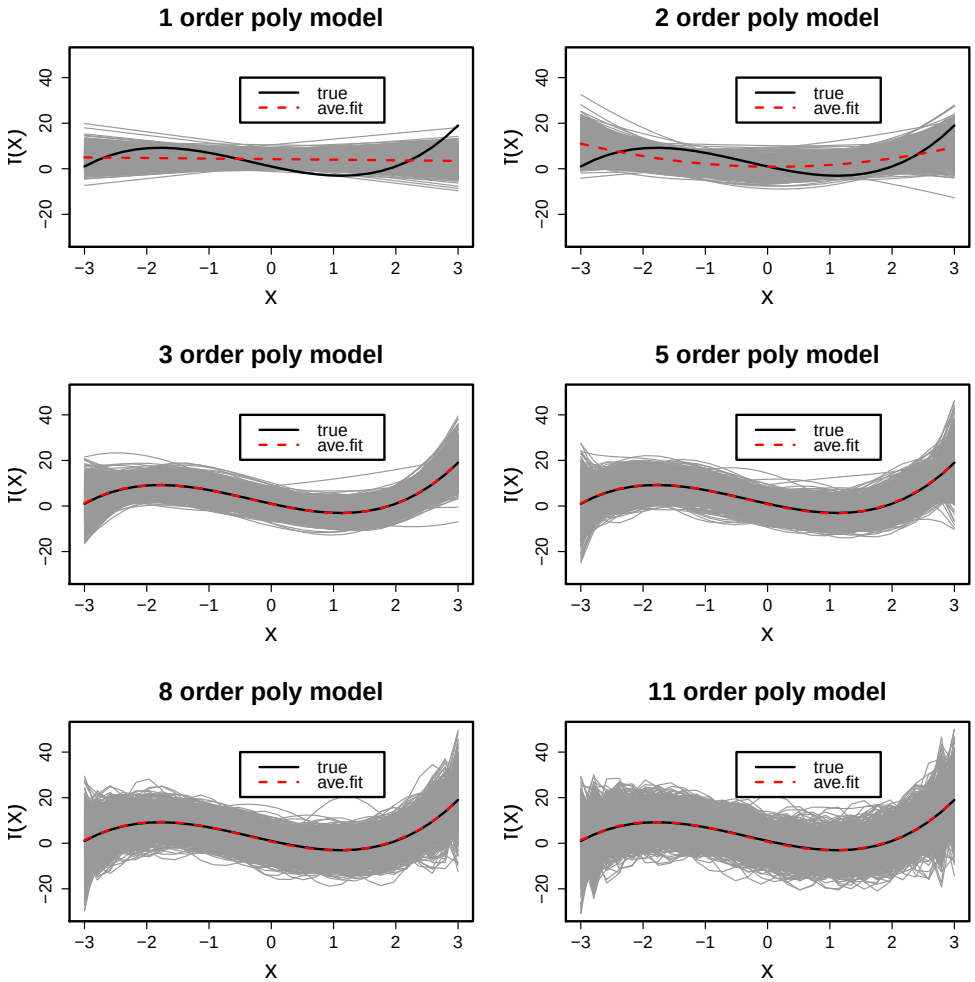
- Denote a polynomial regression model with oder l by $\mathbb{M}_l$.
- Consider the following orders: $l = 1$ (linear), $l = 2$ (quadratic), $l = 3$ (cubic), $l = 5$, $l = 8$, and $l = 11$.
- $\mathbb{M}_1$ and $\mathbb{M}_2$ are underfitted models.
- $\mathbb{M}_3, \mathbb{M}_5, \mathbb{M}_8, \mathbb{M}_{11}$ are correct models.
- $\mathbb{M}_3$ is the data generating model.

Figure: Fitted response curves for four replicates



Fitted response curves are changing across replicates due to . $\mathbb{M}_8, \mathbb{M}_{11}$ tend to the observed data, while $\mathbb{M}_1, \mathbb{M}_2$ tend to the observed data.

Figure: Fitted response curves across 1000 replicates

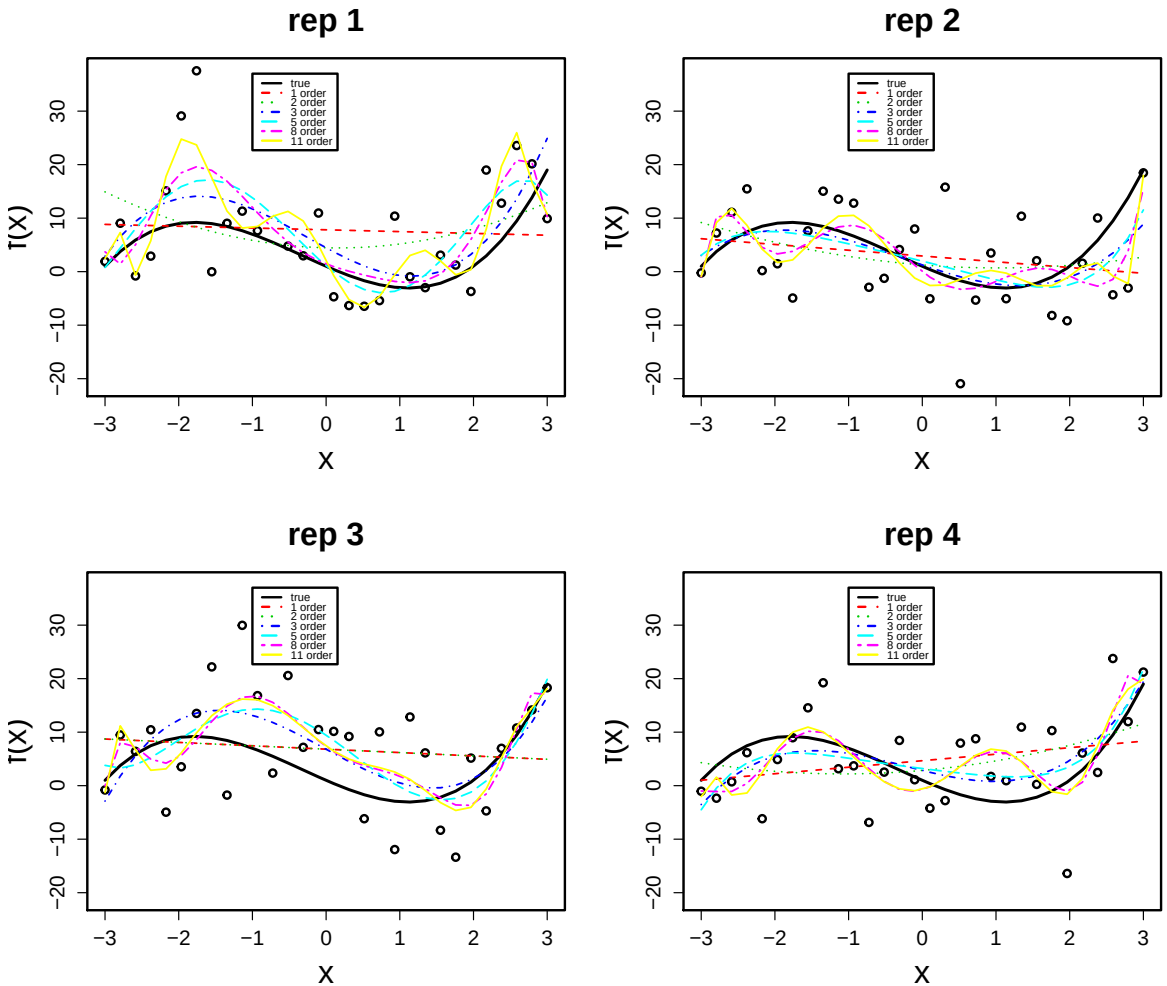


$\mathbb{M}_1, \mathbb{M}_2$ have

with the order l.

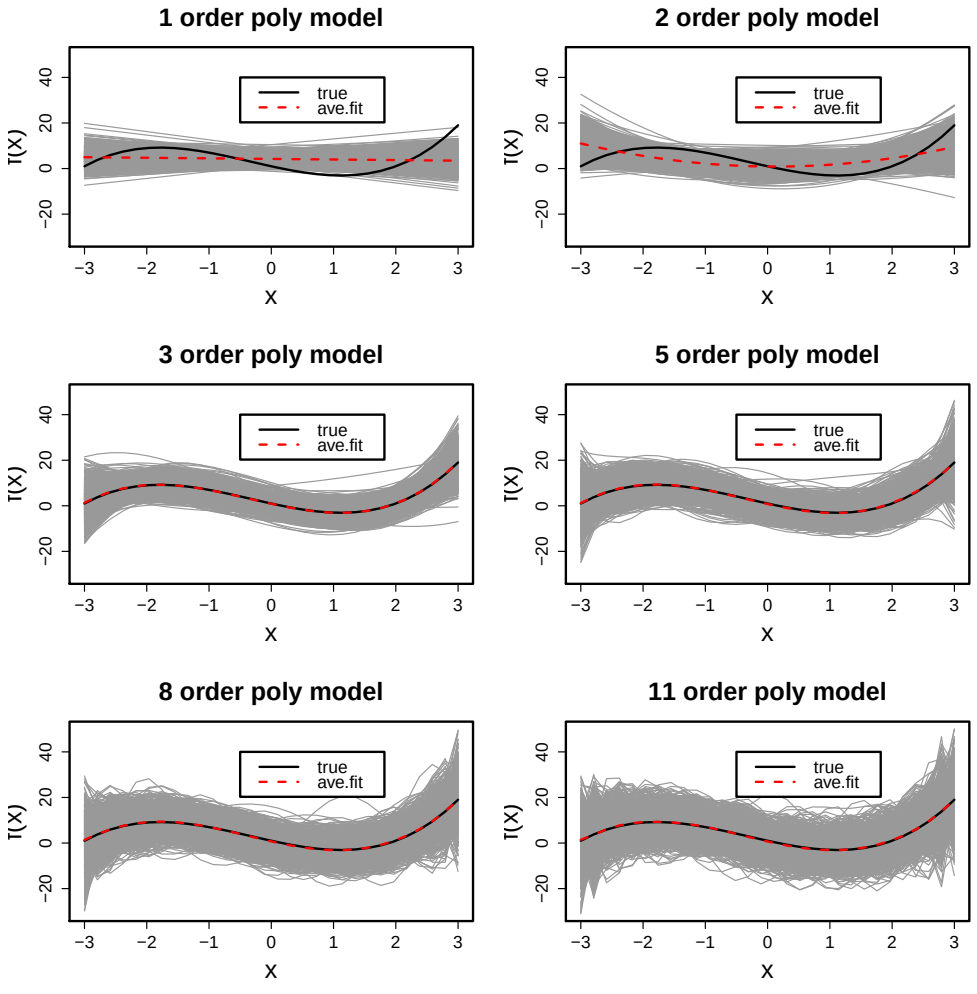
model bias, while model variance

Figure: Fitted response curves for four replicates



Fitted response curves are changing across replicates due to sampling variability. $\mathbb{M}_8, \mathbb{M}_{11}$ tend to overfit the observed data, while $\mathbb{M}_1, \mathbb{M}_2$ tend to underfit the observed data.

Figure: Fitted response curves across 1000 replicates



$\mathbb{M}_1, \mathbb{M}_2$ have large model bias, while model variance increases with the order l.

- Derive the empirical versions of $E(SSE)$, $bias_{in}$, Var_{in} and $msee_{in}$ for each model by averaging across the 1000 replicates.
- Suppose $SSE_k(\mathbb{M}_l)$ is the SSE of the kth replicate under model $\mathbb{M}_l$, then

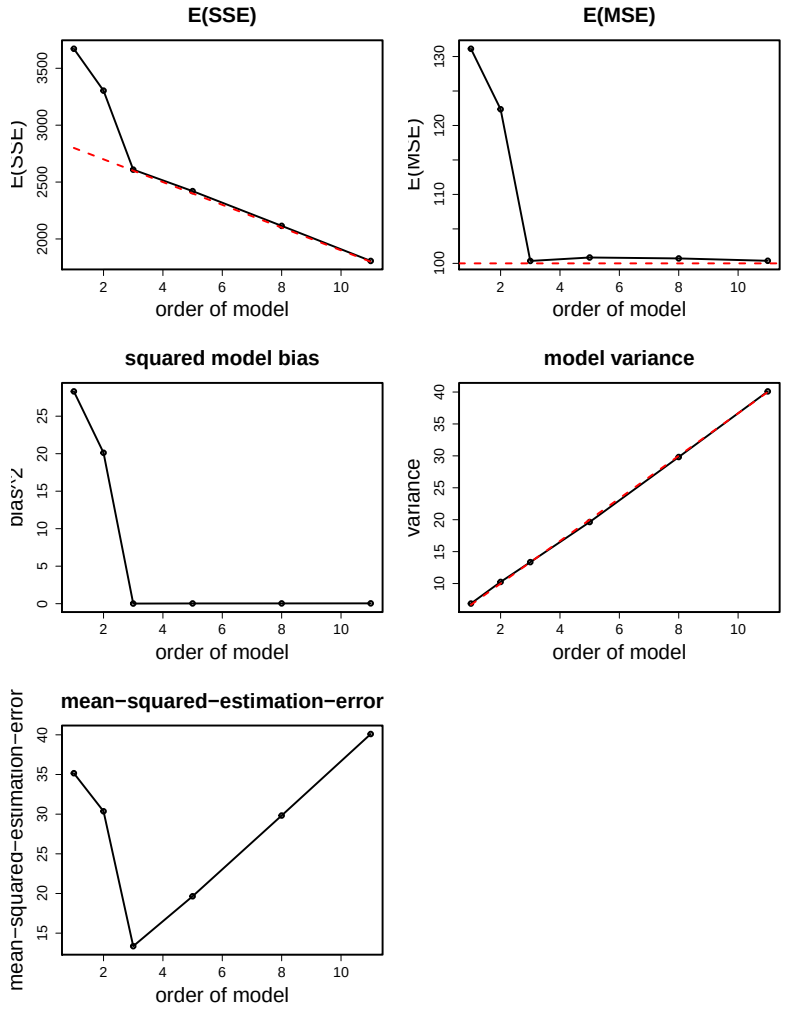
$$E(SSE(\mathbb{M}_l)) \approx \frac{1}{1000} \sum_{k=1}^{1000} SSE_k(\mathbb{M}_l).$$

- Suppose $\hat{Y}^k(\mathbb{M}_l)$ is the fitted values vector of the kth replicate under model $\mathbb{M}_l$, then

$$bias_{in}(\mathbb{M}_l) \approx \frac{1}{1000} \sum_{k=1}^{1000} \hat{Y}^k(\mathbb{M}_l) - \boldsymbol{\mu},$$

  where $\boldsymbol{\mu} = (f(X_1), \cdots, f(X_n))^T$ is the true mean response vector. *How to approximate Var_{in} and $msee_{in}$?*
- The more replicates, the closer the empirical versions tend to be to their population counterparts. So it is important to use sufficient number of replicates in a simulation study.

Figure: $E(SSE)$, $E(MSE)$, $bias_{in}$, Var_{in} and $msee_{in}$



$bias_{in}^2$(Var_{in}, $msee_{in}$) are averaged across the n cases to produce one overall number for each model. *What are the two best models according to these plots?*