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: 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:

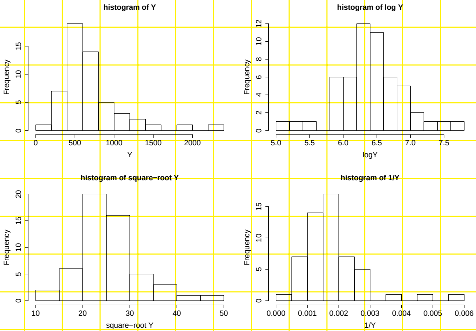


Figure: Distribution of survival times (Y)

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 predictor variables:

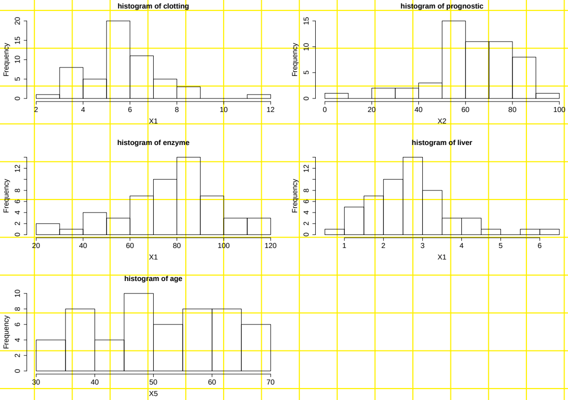
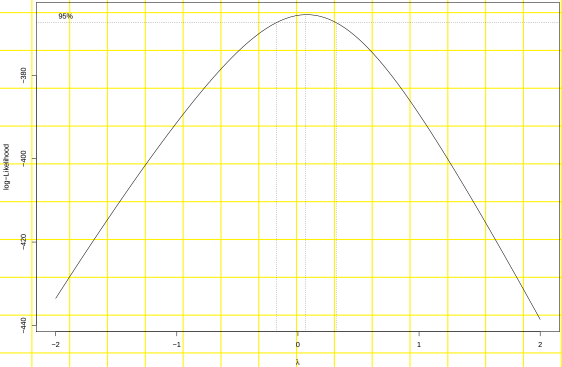


Figure: Distributions of quantitative predictor variables

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

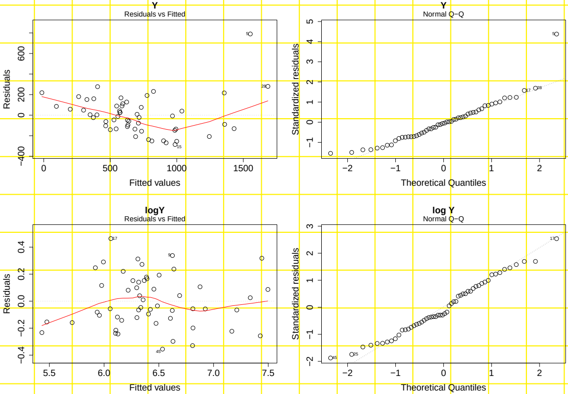
```
fit1=lm(Y˜., data=data.o) ##fit a first-order model with all X variables
plot(fit1, which=1)  ## residuals vs. fitted shows nonlinearity and nonconstant variance
plot(fit1, which=2) ##residuals Q-Q shows heavy right tail
library(MASS)
boxcox(fit1)  ### boxcox procedure suggests logarithm transformation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



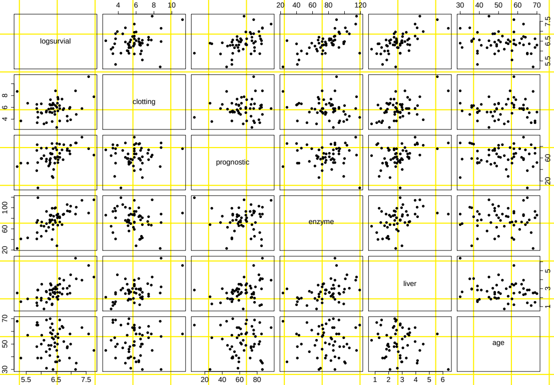
Boxcox procedure suggests a transformation ($\lambda =$) for the response variable (consistent with the exploratory data analysis).

Figure: 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

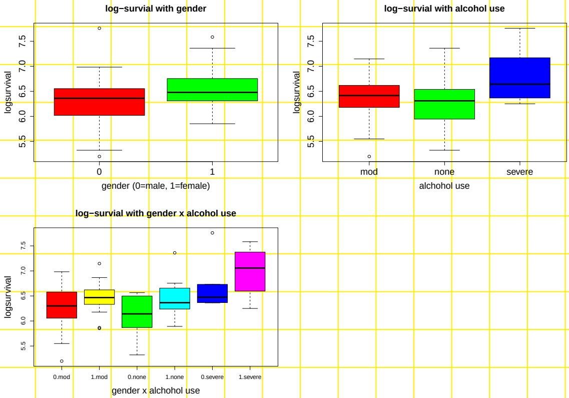


No obvious nonlinearity in regression relations. Log-survival is positively correlated with `clotting`, `prognostic`, `enzyme`, `liver`, and is weakly negatively correlated with `age`.

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.

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 $\Longrightarrow$ **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

Correct Models vs. Good Models

- Correct models are those that contain all important X variables.
- Consequently, correct models have model
bias.
- However, a correct model is a good
model because it may include
which lead to
- A good model should contain (
), and at the same time it should
( ).
- 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., $\{X_1, x_2\}, \{X_1, X_2, X_1X_2\}$,v$\{X_1, X_2, X_1^2, X_2^2\}$,$\{X_1, X_2, X_3, X_4, X_5\}$.
    - These models lead to of the mean response and error variance.
    - However, some of them may have model variance such that the estimates behave erratically with even very small perturbation of the data. Such models, although correct, are .
- On the other hand, the models $\{X_1\}$ or $\{X_2\}$ both have an important X variable being omitted and thus they lead to

Model Variance

Assume the response 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 model variance:

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

Model Bias

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

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

- The biases depend on

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

Mean-Squared-Estiamtion-Errors of a Model

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

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

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

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 of the model to the **observed data Y**.
- E(SSE) is affected by three factors:

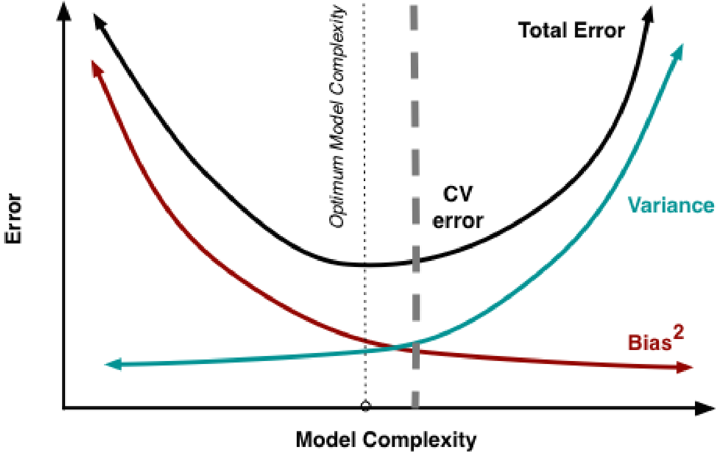
$$E(SSE) \quad =$$

- 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., $\mu = E(\mathbf{Y}) \quad \langle \mathbf{X} \rangle$, then $E(SSE)$ and $E(MSE)$.

Summary: Model Varinace and Model Bias

- Larger models always have a 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 .
- For two correct models, the larger model always has a $E(SSE)$, so they tend to the observed data. On the other hand, the larger models have overall variance and thus they have overall 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.

Bias-Variance Trade-off



Key Components for Model Selection

- **Criterion to compare models**:
  - 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.

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 σ^2. E.g.,

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

  - $\hat{\sigma}^2$ 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 as an Estimator of msee

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

So C_p can be viewed as an estimator of the

.

How to Use C_p?

- If a model has no (in-sample) bias, i.e., $bias_{in}(\mathbb{M}) = \mathbf{0}$, then $E(C_p)$. Otherwise $E(C_p)$ tends to be than p.
- When C_p is plotted against p, then models with will tend to fall near the diagonal line $C_p = p$.
- On the other hand, models with 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 bias and number of X variables (thus 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.

AIC_p and BIC_p Criteria

- *Akaike's information criterion (AIC)*:

$$AIC_p = n \log \frac{SSE_p}{n} + 2p.$$

- *Bayesian information criterion (BIC)*:

$$BIC_p = n \log \frac{SSE_p}{n} + (\log n)p.$$

- **We should look for models with small AIC (BIC).**
    - Surgical unit. The model with X_1, X_2, X_3 has the smallest AIC and BIC among the models being considered.

- The first term: $n \log \frac{SSE_p}{n}$ reflects the of the model to the observed data.
  - It by adding more X variables into the model.
- The second term, $2p$ for AIC and $(\log n)p$ for BIC, reflects
  .
  - It by adding more X variables into the model.
  - If $n \geq 8$, then $\log n > 2$ and BIC puts penalty on model complexity and tends to choose models than AIC.

- Overly simplified models have model complexity (p), but they tend to have SSE (underfitting; high).
- Overly complicated models may have a SSE, but they have model complexity (overfitting, high).
- By minimizing AIC (or BIC), we are trying to find a model that between model complexity and the goodness-of-fit.

$Press_p$ Criterion

Predicted residual sum of squares ($Press_p$):

$$Press_p = \sum_{i=1}^{n} (Y_i - \widehat{Y}_{i(i)})^2.$$

- Y_i is the observed response of the *ith* case.
- $\widehat{Y}_{i(i)}$ is the predicted value for the ith case obtained by fitting the model only using $n - 1$ cases excluding case i.
- $Press_p$ is also known as *leave-one-out-cross-validation (LOOCV)*.
- Models with small $Press_p$ are considered good in terms of predictive ability.
  - Surgical unit: the model with X_1, X_2, X_3 has $Press_p = 3.914$ which is the smallest among all models being considered here.

Calculate $Press_p$

$Press_p$ can be calculated without actually performing n regressions.

- This is because the *deleted residual* for the *ith* case:

$$d_i := Y_i - \widehat{Y}_{i(i)} = \qquad\qquad\qquad\qquad , \quad i = 1, \cdots, n.$$

  where $e_i = Y_i - \widehat{Y}_i$ is the residual of the *ith* case and h_{ii} is the *ith* diagonal element of the hat matrix **H**, both from the regression fit using .

- So

Derive the Deleted Residuals

Optional Reading.

- Define $\tilde{\mathbf{Y}}$ by replacing the *i*th element of the response vector $\mathbf{Y}$ with the leave-*i*-out predicted value $\hat{Y}_{i(i)}$ of the *i*th case:

$$\tilde{\mathbf{Y}} = (Y_1, \cdots, Y_{i-1}, \hat{Y}_{i(i)}, Y_{i+1}, \cdots, Y_n)^T.$$

- Let $\hat{\beta}_{(i)}$ be the leave-i-out LS fitted regression coefficients. Then $\hat{\beta}_{(i)}$ is also the LS fitted regression coefficients by using $\tilde{\mathbf{Y}}$ as the response vector, i.e. $\hat{\beta}_{(i)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\tilde{\mathbf{Y}}$. *Why?*
- The leave-i-out fitted values are:

$$\hat{\mathbf{Y}}_{(i)} = \mathbf{X}\hat{\beta}_{(i)} = H\tilde{\mathbf{Y}} = H(\mathbf{d}_{(i)}+\mathbf{Y}), \quad \mathbf{d}_{(i)} = \tilde{\mathbf{Y}}-\mathbf{Y} = (0, \cdots, -d_i, \cdots, 0)^T.$$

- Subtracting the *i*th element from Y_i on both sides gives:

$$d_i = h_{ii}d_i + e_i \Longrightarrow d_i = \frac{e_i}{1 - h_{ii}}.$$

Surgical Unit: Full Model X_1, X_2, X_3, X_4

```
> fit.f =lm(log(Y)˜X1+X2+X3+X4, data=data.o)
> summary(fit.f)
Call:
lm(formula = log(Y) ˜ X1 + X2 + X3 + X4, data = data.o)
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.851933   0.266263  14.467  < 2e-16 ***
X1          0.083739   0.028834   2.904  0.00551 **
X2          0.012671   0.002315   5.474 1.50e-06 ***
X3          0.015627   0.002100   7.440 1.38e-09 ***
X4          0.032056   0.051466   0.623  0.53627
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1
Residual standard error: 0.2509 on 49 degrees of freedom
Multiple R-squared: 0.7591,     Adjusted R-squared: 0.7395
F-statistic: 38.61 on 4 and 49 DF,  p-value: 1.398e-14
> anova(fit.f)
Analysis of Variance Table

Response: log(Y)
Df Sum Sq Mean Sq  F value    Pr(>F)
X1         1 0.7770  0.7770  12.3443 0.0009618 ***
X2         1 2.5904  2.5904  41.1565 5.341e-08 ***
X3         1 6.3286  6.3286 100.5490 1.838e-13 ***
X4         1 0.0244  0.0244   0.3879 0.5362698
Residuals 49 3.0841  0.0629
```

Surgical Unit: Full Model

- Full model has $P = 5$ and

  $$SSE = 3.0841, \;\; MSE = 0.0629, \;\; R^2 = 0.7591, \;\; R_a^2 = 0.7395.$$

- By definition, for the full model, $C_P = P = 5$.
- Sample size $n = 54$, so for the full model:
  $AIC_P = 54\log(3.0841/54) + 2 \times 5 = -144.5871$ and
  $BIC_P = 54\log(3.0841/54) + \log(54) \times 5 = -134.6422$.
- $Press_p = 4.069$.

```
> e.f=fit.f$residuals  ## residuals
> h.f=influence(fit.f)$hat  ## diagonals of hat matrix
> press.f= sum(e.f^2/(1-h.f)^2)  ## calculate press
```