An Introduction to Statical Learning 笔记

第三章 线性回归

3.3.3 线性回归中遇到的问题

主要问题：

1. 因变量与自变量之间是非线性关系。
2. 误差项的相关。
3. 误差项的非尝试方差。
4. 离群值(outliers)。
5. 高杠杆数据
6. 共线性。

4. 分类器

4.2 为何不用线性回归

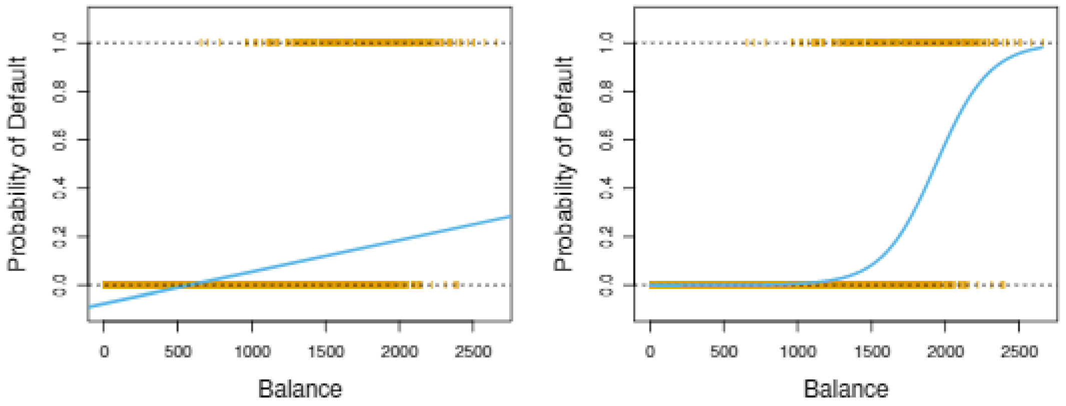


FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default`*(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

4.3 Logistic回归

4.4 其他分类方法

4.4.1 LDA（线性判别式）当 p = 1

三种方法比较：

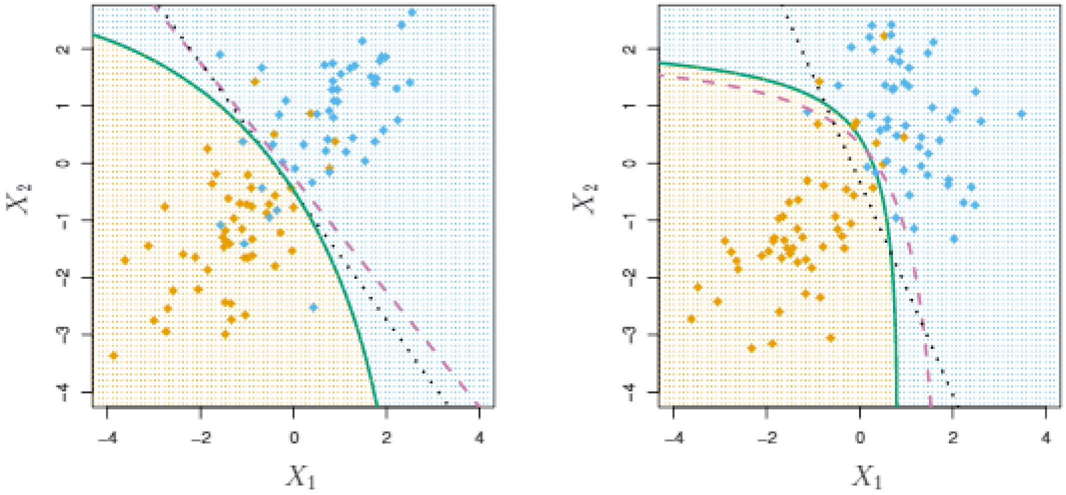


FIGURE 4.9. Left: *The Bayes (purple dashed), LDA (black dotted), and QDA (green solid) decision boundaries for a two-class problem with* $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. *The shading indicates the QDA decision rule. Since the Bayes decision boundary is linear, it is more accurately approximated by LDA than by QDA.* Right: *Details are given in the left-hand panel, except that* $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. *Since the Bayes decision boundary is non-linear, it is more accurately approximated by QDA than by LDA.*

4.5 不同模型的比较

4.5.2 实证比较

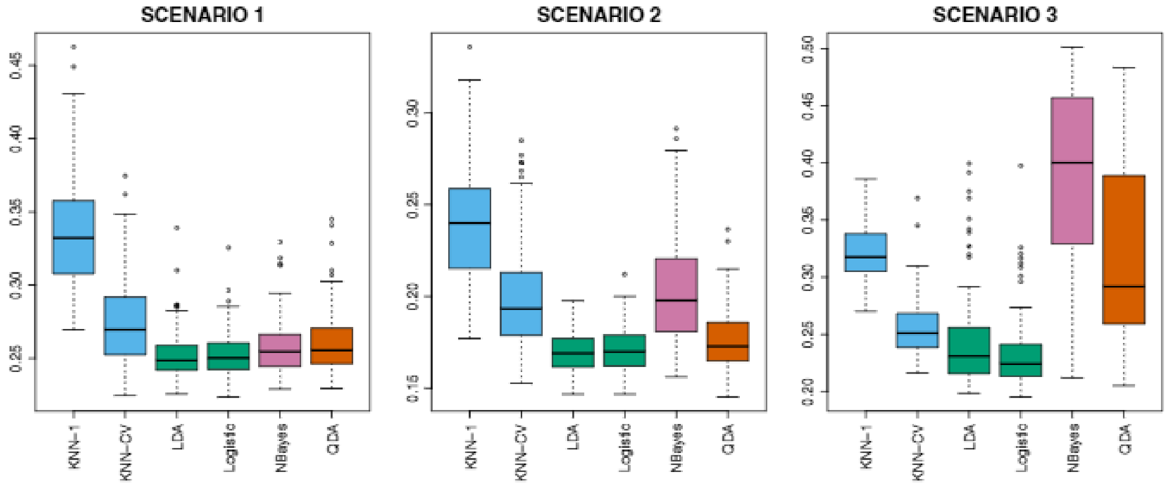


FIGURE 4.11. *Boxplots of the test error rates for each of the linear scenarios described in the main text.*

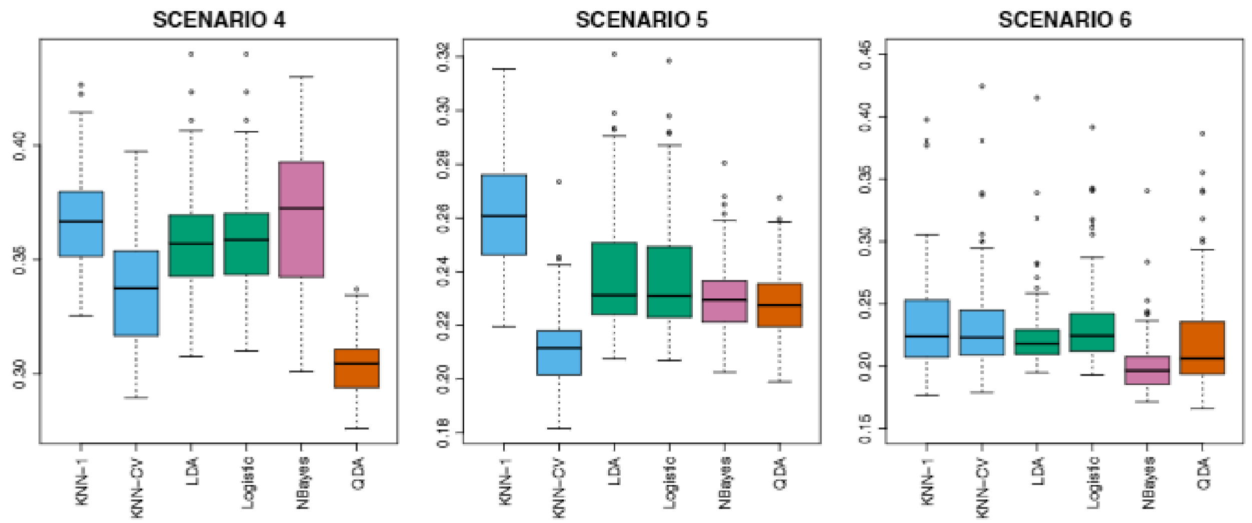


FIGURE 4.12. *Boxplots of the test error rates for each of the non-linear scenarios described in the main text.*

5. 重抽样方法

5.1 交叉验证(Cross-Validation)

5.1.1 验证集方法

使用一个比例的验证集(*Validation Set*)

5.1.2 Leave-One-Out Cross-Validation

每次取出一个做验证，然后用平均值作为结果：

$$
CV_{(n)} = \frac 1 n \sum_{i=1}^n {MSE}_i
$$

5.1.3 k-Fold Cross-Valisdation

$$
CV_{(k)} = \frac 1 k \sum^k_{i=1} {MSE}_i
$$

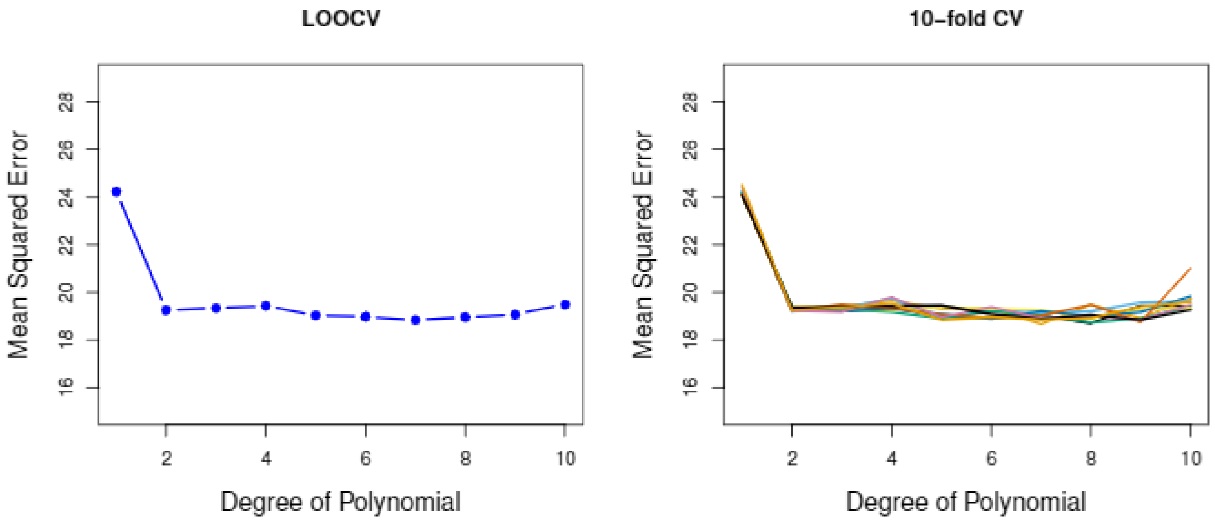


FIGURE 5.4. *Cross-validation was used on the* Auto *data set in order to estimate the test error that results from predicting* mpg *using polynomial functions of* horsepower. Left: *The LOOCV error curve.* Right: *10-fold CV was run nine separate times, each with a different random split of the data into ten parts. The figure shows the nine slightly different CV error curves.*

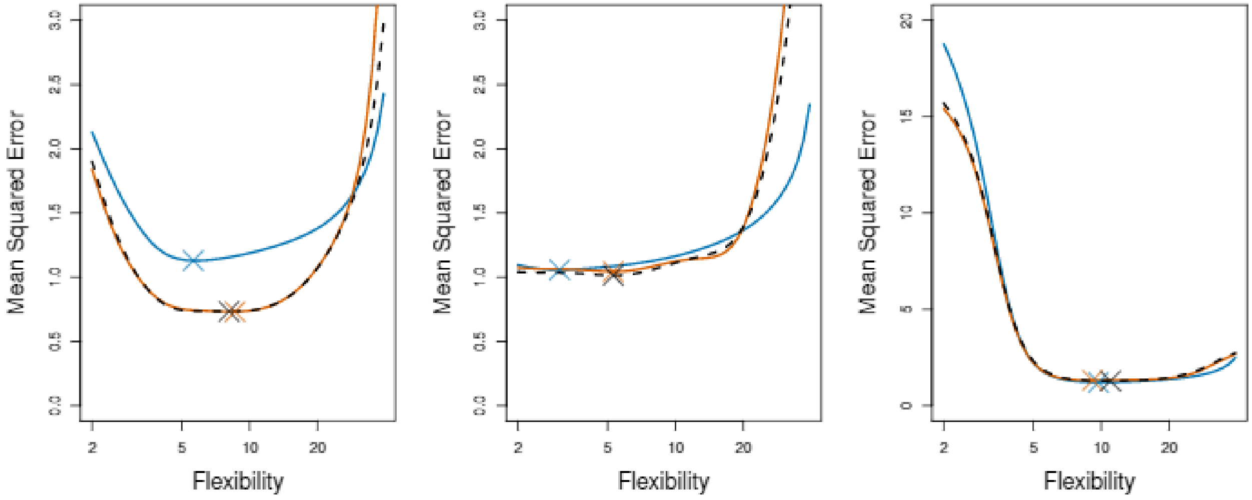


FIGURE 5.6. *True and estimated test MSE for the simulated data sets in Figures 2.9 (*left*), 2.10 (*center*), and 2.11 (*right*). The true test MSE is shown in blue, the LOOCV estimate is shown as a black dashed line, and the 10-fold CV estimate is shown in orange. The crosses indicate the minimum of each of the MSE curves.*

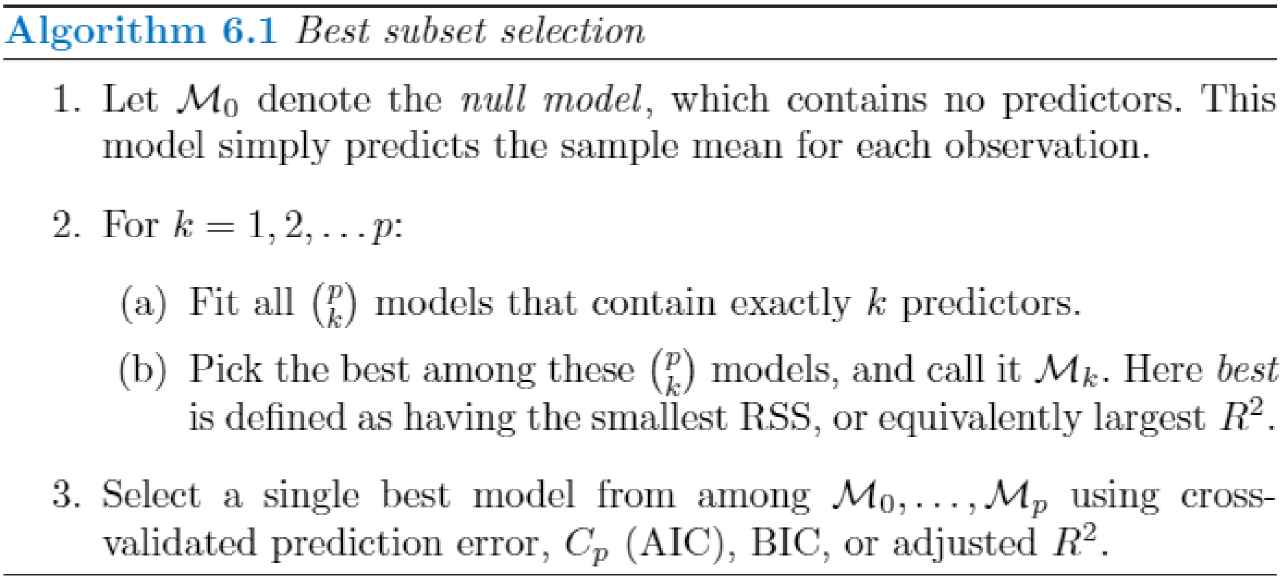
5.2 The Bootstrap

第六章 线性模型选择和正则化

拟合过程的修改基于两个原因：**预测精确度**和**模型可解释性**

6.1 子集选择(Subset Selection)

6.1.1 最优子集选择

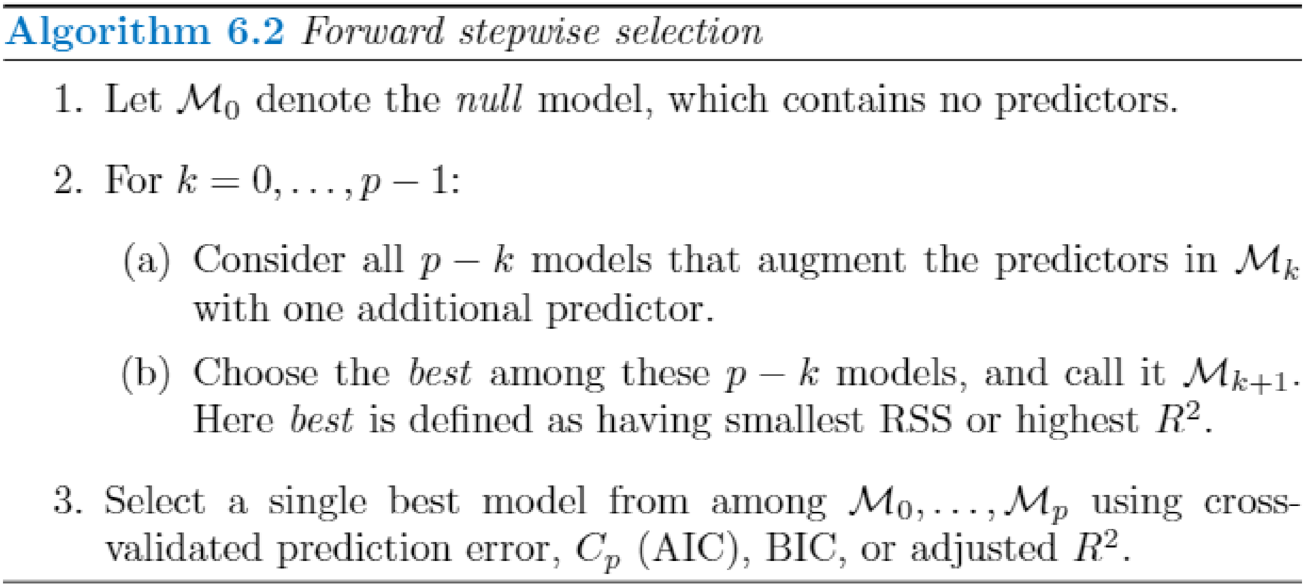


Algorithm 6.1 *Best subset selection*

```
1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This
   model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here best
       is defined as having the smallest RSS, or equivalently largest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-
   validated prediction error, $C_p$ (AIC), BIC, or adjusted $R^2$.
```

6.1.2 分步选择(Stepwise Selection)

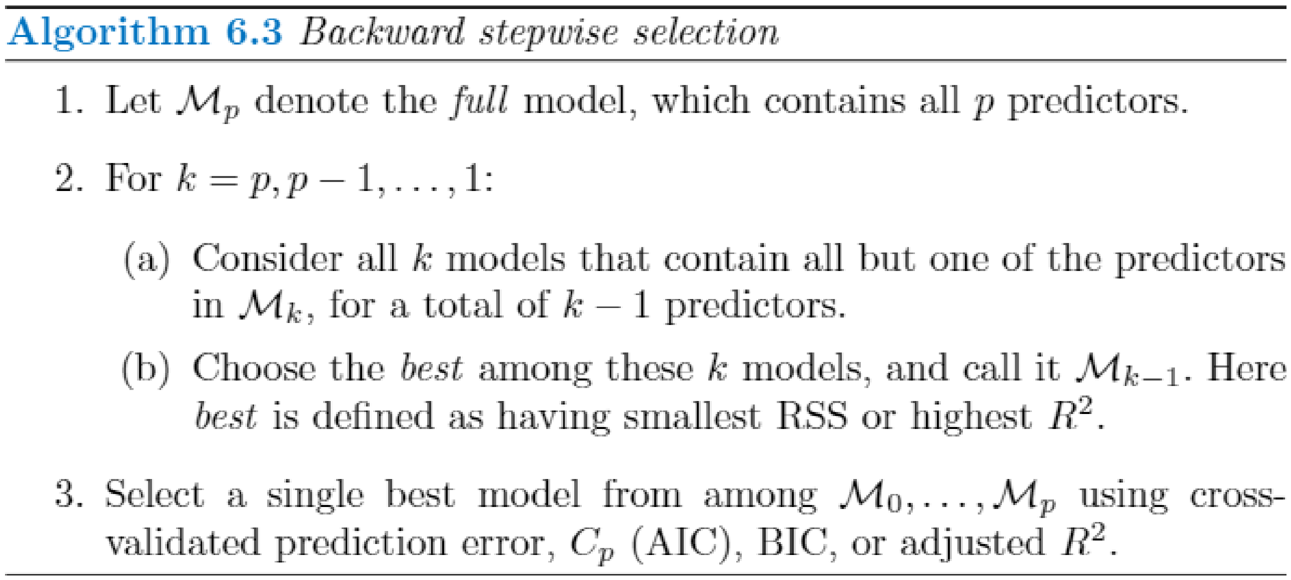
前向分步选择



Algorithm 6.2 *Forward stepwise selection*

```
1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors.
2. For $k = 0, \ldots, p - 1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$
       with one additional predictor.
   (b) Choose the best among these $p - k$ models, and call it $\mathcal{M}_{k+1}$.
       Here best is defined as having smallest RSS or highest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-
   validated prediction error, $C_p$ (AIC), BIC, or adjusted $R^2$.
```

后向分步选择



Algorithm 6.3 *Backward stepwise selection*

```
1. Let $\mathcal{M}_p$ denote the *full* model, which contains all $p$ predictors.
2. For $k = p, p-1, \ldots, 1$:
    (a) Consider all $k$ models that contain all but one of the predictors
        in $\mathcal{M}_k$, for a total of $k-1$ predictors.
    (b) Choose the *best* among these $k$ models, and call it $\mathcal{M}_{k-1}$. Here
        *best* is defined as having smallest RSS or highest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-
   validated prediction error, $C_p$ (AIC), BIC, or adjusted $R^2$.
```

6.1.3 选择最佳模型

如何选择

1. 非直接地使用训练错误
2. 直接使用CV方法。

C_p、AIC, BIC和调整的R^2

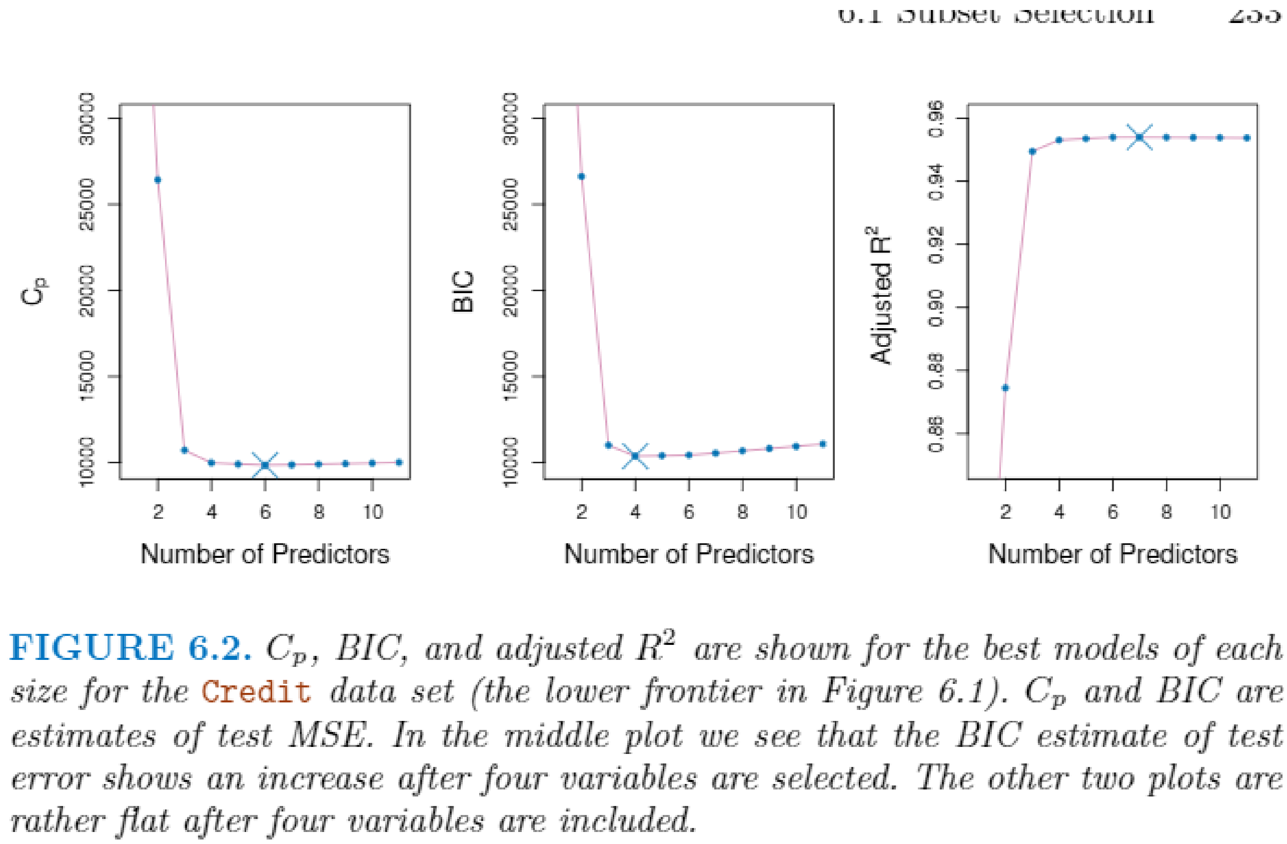


FIGURE 6.2. *C_p, BIC, and adjusted R^2 are shown for the best models of each size for the* Credit *data set (the lower frontier in Figure 6.1). C_p and BIC are estimates of test MSE. In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected. The other two plots are rather flat after four variables are included.*

$$
C_p = \frac 1 n (RSS+ 2d\hat{\sigma ^ 2})
$$

$$
AIC = \frac 1 n (RSS + 2d\hat{\sigma ^ 2})
$$

$$
BIC = \frac 1 n (RSS + log(n)d\hat{\sigma ^ 2})
$$

$$
\text{Adusted } {R}^2 = 1 - frac {RSS/(n - d - 1)} {TSS / (n - 1)}
$$

Validation和CV

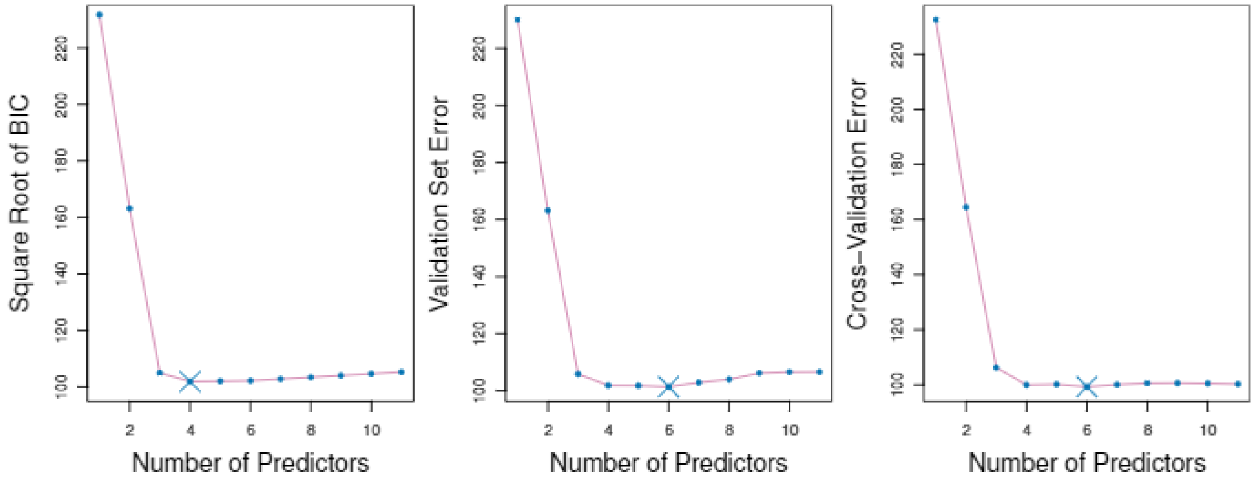


FIGURE 6.3. *For the* Credit *data set, three quantities are displayed for the best model containing d predictors, for d ranging from* 1 *to* 11*. The overall* best *model, based on each of these quantities, is shown as a blue cross.* Left: *Square root of BIC.* Center: *Validation set errors.* Right: *Cross-validation errors.*

6.2 收缩方法

6.2.1 岭回归

$$
RSS = \sum^n_{i = 1} (y_i - \beta_0 - \sum^p_{j=1} \beta_j x_{ij}) ^ 2
$$

岭回归的优化函数

$$
RSS + \lambda \sum^p_{j=1} \beta ^ 2 _ j
$$

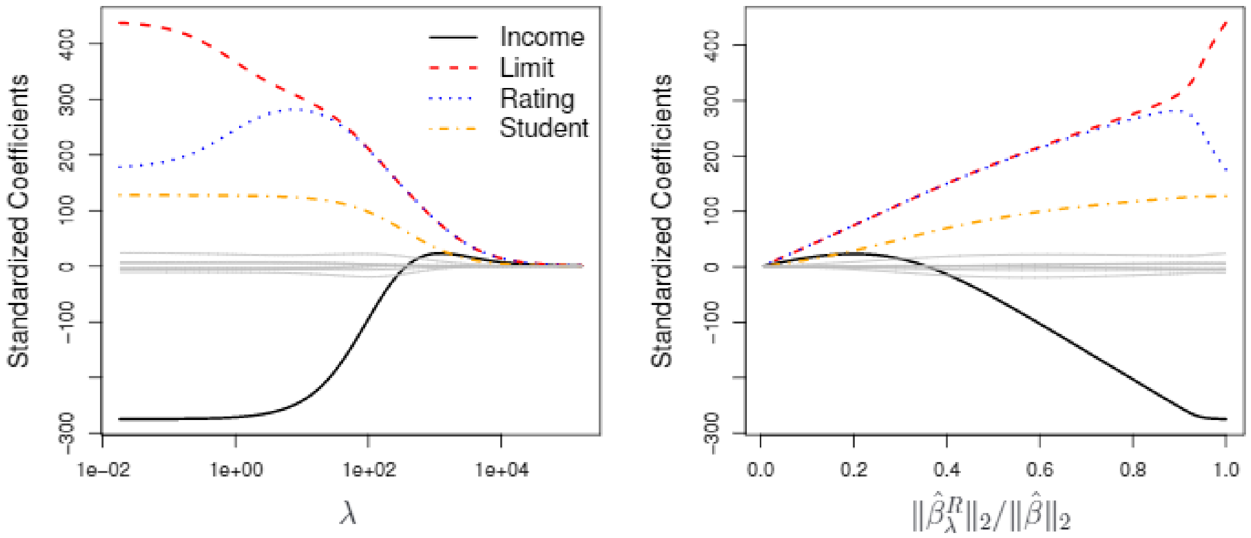


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.

6.2.2 The Lasso

$$
RSS + \lambda \sum^p_{j=1} |\beta_j|
$$

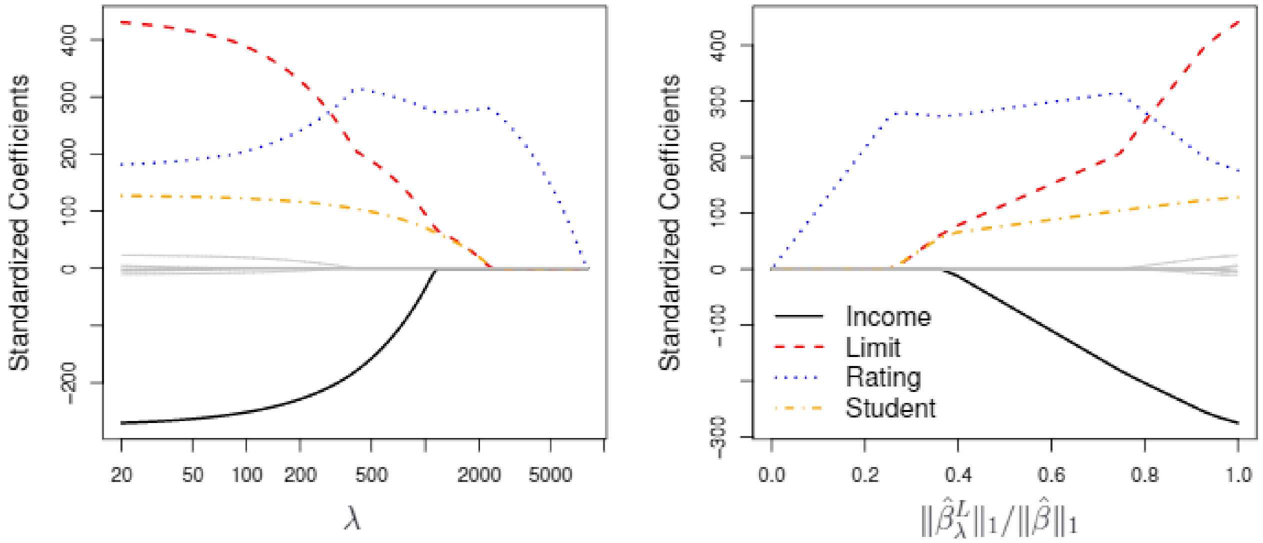


FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

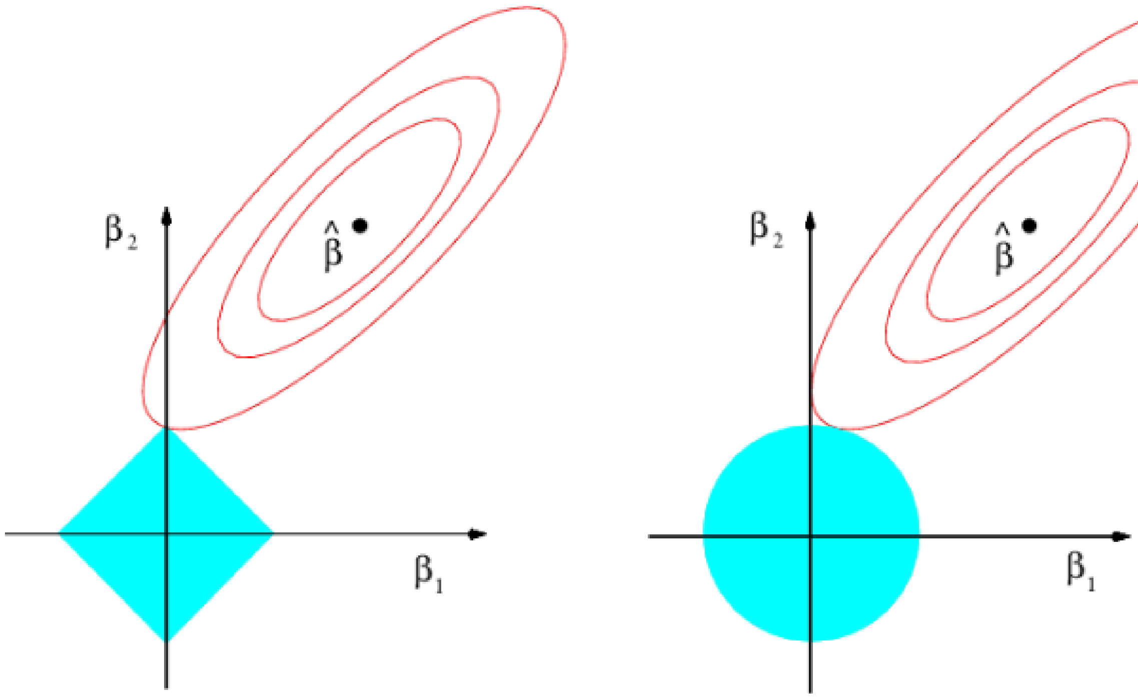


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, while the red ellipses are the contours of the RSS.*

6.3 降维方法

6.3.1 PCA

6.3.2 偏最小二乘回归(Partial Least Squares)

6.4 高维度数据

第八章 树模型

8.1.4 树模型的优缺点

优点：

- 非常容易向人解释
- 有些人认为人类就是类似这么做决策的
- 可以用图的方式解释，甚至给一个非技术人员
- 不需要使用哑变量

缺点：

- 树模型没有很强的预测准确度
- 树模型不够稳健。

8.2 Bagging, RF, Boosting, Bayesian Addictive Regression Trees

第十一章 生存分析和截尾数据

11.1 生存和截尾时间

一般我们能观察到两个时间，一个是截尾时间$C$和一个生存时间$T$。即我们可以看到下面的随机变量：

$$
Y = min(T, C)
$$

此外我们可以设定一个$\delta$来记录是否在真的截尾了。

$$
\delta = \begin{cases}
1 && \text{if } T \le C \
0 && \text{if } T > C
\end{cases}
$$

这样可以得到下面的图片：

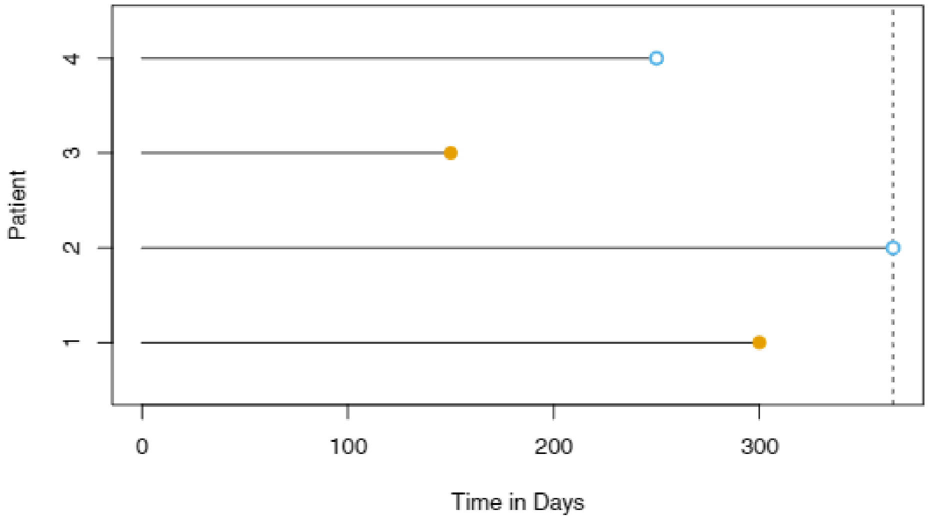


FIGURE 11.1. *Illustration of censored survival data. For patients 1 and 3, the event was observed. Patient 2 was alive when the study ended. Patient 4 dropped out of the study.*

11.3 Kaplan-Meier生存曲线

生存曲线，生存函数定义为

$$
S(t) = Pr(T > t)
$$

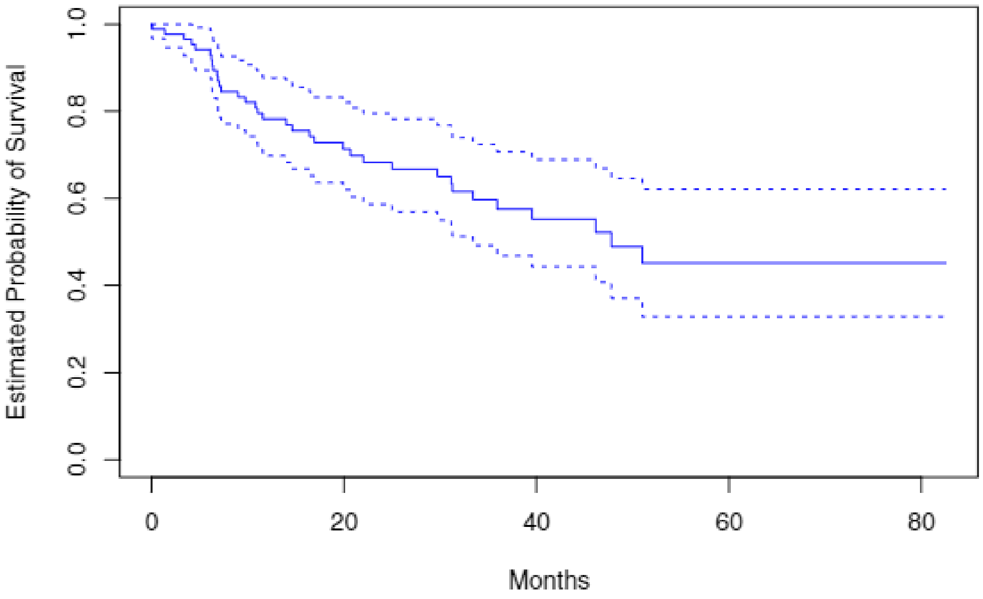


FIGURE 11.2. *For the* BrainCancer *data, we display the Kaplan-Meier survival curve (solid curve), along with standard error bands (dashed curves).*

11.4 Log-Rank测试

Log-Rank测试是用来验证两条KM曲线是否相同。

	Group 1	Group 2	Total
Died	q_{1k}	q_{2k}	q_k
Survived	$r_{1k} - q_{1k}$	$r_{2k} - q_{2k}$	$r_k - q_k$
Total	r_{1k}	r_{2k}	r_{k}

设定W为

$$
W = \frac {X - E(X)} {\sqrt{Var(X)}}
$$

$$
E(q_{1k}) = \frac {r_{1k}} {r_k} q_k
$$

$$
Var(q_{1k}) = \frac
{q_k(r_{1k}/r_k)(1 - r_{1k}/r_k)(r_k - q_k)}
{r_k - 1}
$$

$$
Var(\sum^K_{k = 1} q_{1k}) \approx \sum^K_{k=1} Var(q_{1k})
= \sum_{k=1}^K \frac

{q_k(r_{1k}/r_{k})(1 - r_{1k}/ r_k)(r_k - q_k)}
{r_k - 1}
$$

从而

$$
W = \frac
{\sum_{k=1}^K(q_{1k} - E(q_{1k}))}
{\sqrt{\sum^K_{k=1}Var(q_{1k})}}
= \frac
{\sum{k=1}^K(q_{1k} - \frac{q_k}{r_k}r_{1k})}
{\sqrt{\sum^K_{k=1}
\frac
{q_k(r_{1k}/r_k)(1 - r_{1k}/r_{k})(r_k - q_k)}
{r_k - 1}}}
$$

11.5 回归模型

11.5.1 Hazard函数

$$
h(t) = \lim_{\Delta t \to 0} \frac
{Pr(t < T \le t + \Delta t | T > t)}
{\Delta t}
\approx \frac
{Pr(t < T \le t + \Delta t | T > t)}
{\Delta t}
= \frac {f(t)} {S(t)}
$$

其第i的likelihood为

$$
\begin{align}
L_i && =
\begin{cases}
f(y_i) && \text{if the }i\text{th obsevation is not censored}\
S(y_i) && \text{if the }i\text{th obsevation is censored}
\end{cases}
\
= && f(y_i)^{\delta_i}S(y_i)^{1 - \delta_i}
\end{align}
$$

11.5.2 比例Hazards

$$
h(t|x_i) = h_0(t) exp(\sum^p_{j=1} x_{ij}\beta_j)

$$

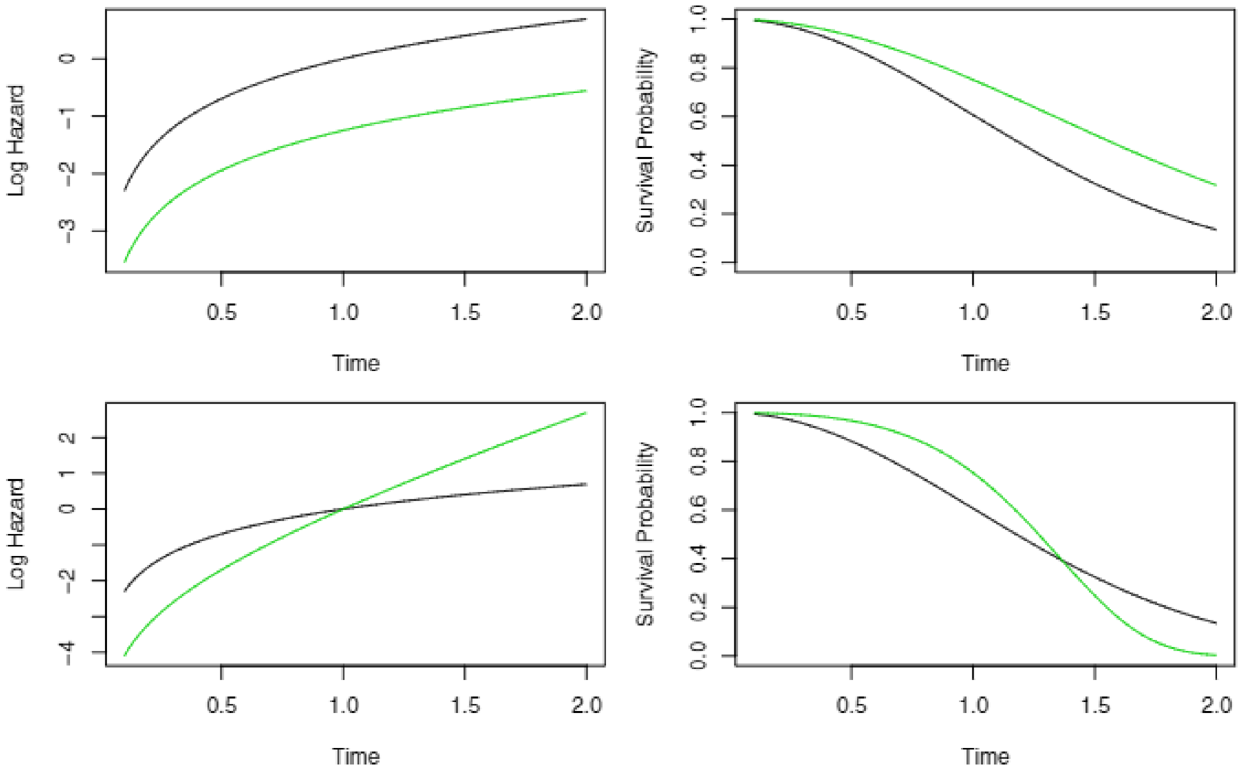


FIGURE 11.4. Top: *In a simple example with* $p = 1$ *and a binary covariate* $x_i \in \{0, 1\}$, *the log hazard and the survival function under the model* (11.14) *are shown (green for* $x_i = 0$ *and black for* $x_i = 1$*). Because of the proportional hazards assumption* (11.14)*, the log hazard functions differ by a constant, and the survival functions do not cross.* Bottom: *Again we have a single binary covariate* $x_i \in \{0, 1\}$*. However, the proportional hazards assumption* (11.14) *does not hold. The log hazard functions cross, as do the survival functions.*