Linear Methods for Regression

Hastie – Chap - 3; Part A

Introduction

- A linear regression model assumes that the regression function $E(Y\ |X)$ is linear in the inputs $X_1,\ldots,X_p$.

- They are simple and often provide an adequate and interpretable description of how the inputs affect the output. For prediction purposes they can sometimes outperform fancier nonlinear models, especially in situations with small numbers of training cases, low signal-to-noise ratio or sparse data.

Linear Regression Models and Least Squares

- Purpose: - to predict a real-valued output Y. The linear regression model has the form.

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \,. \qquad (3.1)$$

- The linear model either assumes that the regression function $E(Y|X)$ is linear, or that the linear model is a reasonable approximation. Here the β_j's are unknown parameters or coefficients, and the variables X_j can come from different sources:

- We have a set of training data $(x_1, y_1) \ldots (x_N, y_N)$ from which to estimate the parameters . Each $x_i = (x_{i1}\ x_{i2}\ \ldots\ x_{ip})^T$ is a vector of feature measurements for the i^{th} case. The most popular estimation method is $least\ squares$, in which we pick the coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ to minimize the residual sum of squares

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

$$= \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2. \qquad (3.2)$$

- From a statistical point of view, this criterion is reasonable if the training observations (x_i, y_i) represent independent random draws from their population. Even if the x_i's were not drawn randomly, the criterion is still valid if the y_i's are conditionally independent given the inputs x_i.

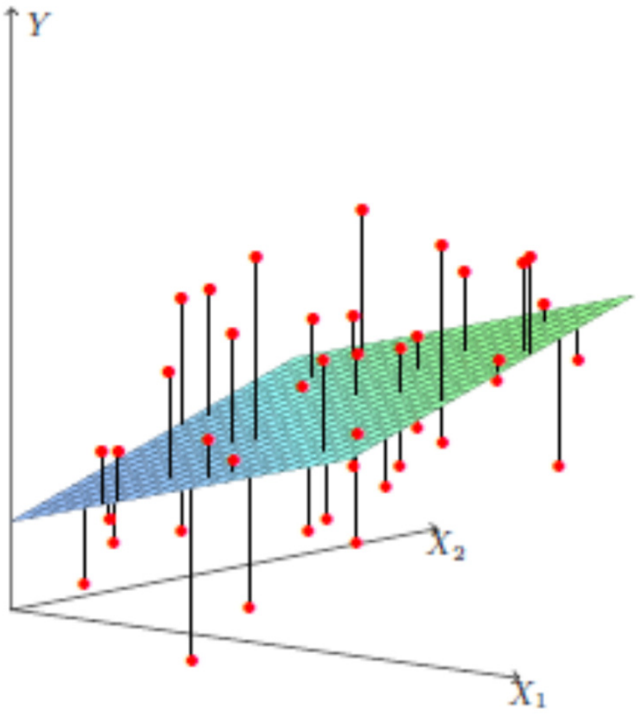


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

Figure 3.1 illustrates the geometry of least-squares fitting in the (p+1)-dimensional space occupied by the pairs (X, Y).

- Figure 3.1 illustrates the geometry of least-squares fitting in the $\mathbb{R}^{p+1}$ −dimensional space occupied by the pairs (X, Y). Note that (3.2) makes no assumptions about the validity of model (3.1); it simply finds the best linear fit to the data. Least squares fitting is intuitively satisfying no matter how the data arise; the criterion measures the average lack of fit.

$$\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

- How do we minimize (3.2)?

Denote by $\mathbf{X}$ the $N \times (p + 1)$ matrix with each row an input vector (with a 1 in the first position), and similarly let $\boldsymbol{y}$ be the N-vector of outputs in the training set. Then we can write the residual sum-of-squares as

$$RSS(\beta) = (\boldsymbol{y} - \boldsymbol{X}\beta)^T (\boldsymbol{y} - \boldsymbol{X}\beta). \qquad (3.3)$$

- This is a quadratic function in the $p + 1$ parameters. Differentiating with respect to we obtain

$$\frac{\partial RSS}{\partial \beta} = -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta)$$

$$\frac{\partial^2 RSS}{\partial \beta \partial \beta^T} = \qquad (3.4)$$

- Assuming (for the moment) that $\boldsymbol{X}$ has full column rank, and hence $\boldsymbol{X}^T\boldsymbol{X}$ is *positive* definite, we set the first derivative to zero

$$\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta) = 0 \qquad (3.5)$$

- To obtain the unique solution

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}. \qquad (3.6)$$

Figure 3.2 shows a different geometrical representation of the least squares estimate, this time in $\mathbb{R}^N$. We denote the column vectors of $\mathbf{X}$ by $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_p$,

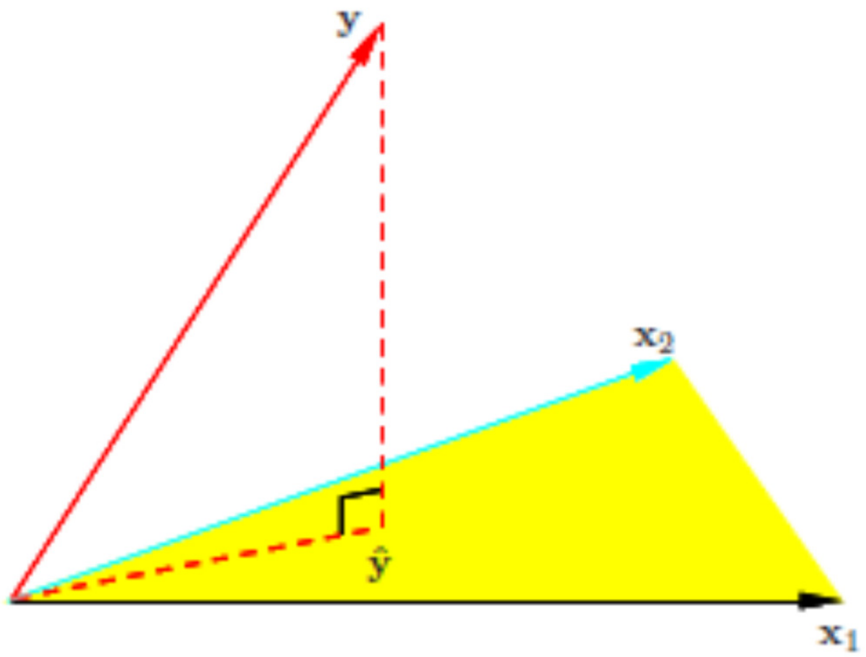


FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector y is orthogonally projected onto the hyperplane spanned by the input vectors x_1 and x_2. The projection $\hat{y}$ represents the vector of the least squares predictions; x_0 = [**1**];*

What is the dimension of y, in fig ?

- The predicted values at an input vector x_0 are given by $\hat{f}(x_0) = (1 : x_0)^T \hat{\beta}$;the fitted values at the training inputs are

$$\hat{y} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y, \qquad (3.7)$$

where $\hat{y}_i = \hat{f}(x_i)$ .The matrix $\mathbf{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ appearing in equation (3.7) is sometimes called the "hat" matrix because it puts the hat on y.

- The hat matrix $\mathbf{H}$ computes the orthogonal projection, and hence it is also known as a projection matrix. It might happen that the columns of $\boldsymbol{X}$ are not linearly independent, so that $\mathbf{X}$ is not of full rank. for example, if two of the inputs were perfectly correlated, $(e.g., x_2 = 3\mathrm{x}_1)$.

- Then $\boldsymbol{X}^T\boldsymbol{X}$ is singular and the least squares coefficients $\hat{\beta}$ are not uniquely defined. However, the fitted values $\hat{y} = \boldsymbol{X}\hat{\beta}$ are still the projection of y onto the columns pace of $\boldsymbol{X}$; The non-full-rank case occurs most often when one or more qualitative inputs are coded in a redundant fashion.

- There is usually a natural way to resolve the non-unique representation, by recoding and/or dropping redundant columns in $\boldsymbol{X}$.

- Rank deficiencies can also occur in signal and image analysis, where the number of inputs p can exceed the number of training cases N. In this case, the features are typically reduced by filtering or else the fitting is controlled by regularization

- Assume that the observations y_i are uncorrelated and have constant variance σ^2, and that the x_i are fixed (non random). The variance–covariance matrix of the least squares parameter estimates is easily derived from (3.6)
  - $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$ (3.6)

and is given by

$$Var(\hat{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2. \quad (3.8)$$

- Typically one estimates the variance σ^2 by.

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i{}^2). \qquad (3.8)$$

- The $N - p - 1$ rather than N in the denominator makes $\hat{\sigma}^2$ an unbiased estimate of σ^2: $E(\hat{\boldsymbol{\sigma}}^{\mathbf{2}}) = \boldsymbol{\sigma}^{\mathbf{2}}$.

- The conditional expectation of Y is linear in $X_1, \ldots, X_p$. We also assume that the deviations of Y around its expectation are additive and Gaussian. Hence

$$Y = E(Y \mid X_1, \ldots, X_p) + \varepsilon$$

$$= \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon, \qquad (3.9)$$

where the error ε is a Gaussian random variable with expectation zero and variance σ^2, written $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ .Under (3.9), it is easy to show that

$$\hat{\beta} \sim N(\beta, (\boldsymbol{X}^T \boldsymbol{X})^{-1} \sigma^2). \qquad (3.10)$$

- This is a multivariate normal distribution with mean vector and variance–covariance matrix as shown.

The Gauss–Markov Theorem

- One of the most famous results in statistics asserts that the **least squares estimates of the parameters β have the smallest variance among all linear unbiased estimates**.

- This observation will lead us to consider biased estimates such as ridge regression later. We focus on estimation of any linear combination of the parameters $\theta = a^T\beta$; for example, predictions $f(x_0) = x_0^T\beta$ are of this form.

- The least squares estimate of $a^T\beta$ is

$$\hat{\theta} = a^T\hat{\beta} = a^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \qquad (3.17)$$

- Considering $\boldsymbol{X}$ to be fixed, this is a linear function $\boldsymbol{c}_0^T\boldsymbol{y}$ of the response vector $\boldsymbol{y}$. If we assume that the linear model is correct, $\alpha^T\hat{\beta}$ is unbiased since

$$E(a^T\hat{\beta}) = E(a^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y})$$

$$= a^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\beta$$

$$= a^T\beta. \qquad (3.18)$$

- The Gauss–Markov theorem states that if we have any other linear estimator $\bar{\theta} = \boldsymbol{c}^T\boldsymbol{y}$ that is unbiased for $a^T\beta$, that is, $E(\boldsymbol{c}^T\boldsymbol{y}) = a^T\beta$, then

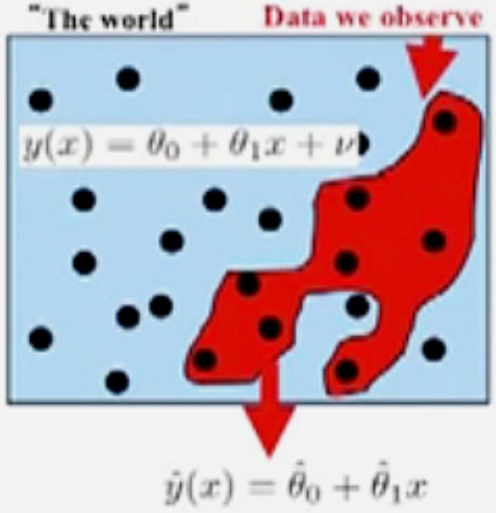
$$Var(a^T\hat{\beta}) \leq Var(\boldsymbol{c}^T\boldsymbol{y}). \qquad (3.19)$$

- Consider the mean squared error of an estimator $\bar{\theta}$ in estimating θ:

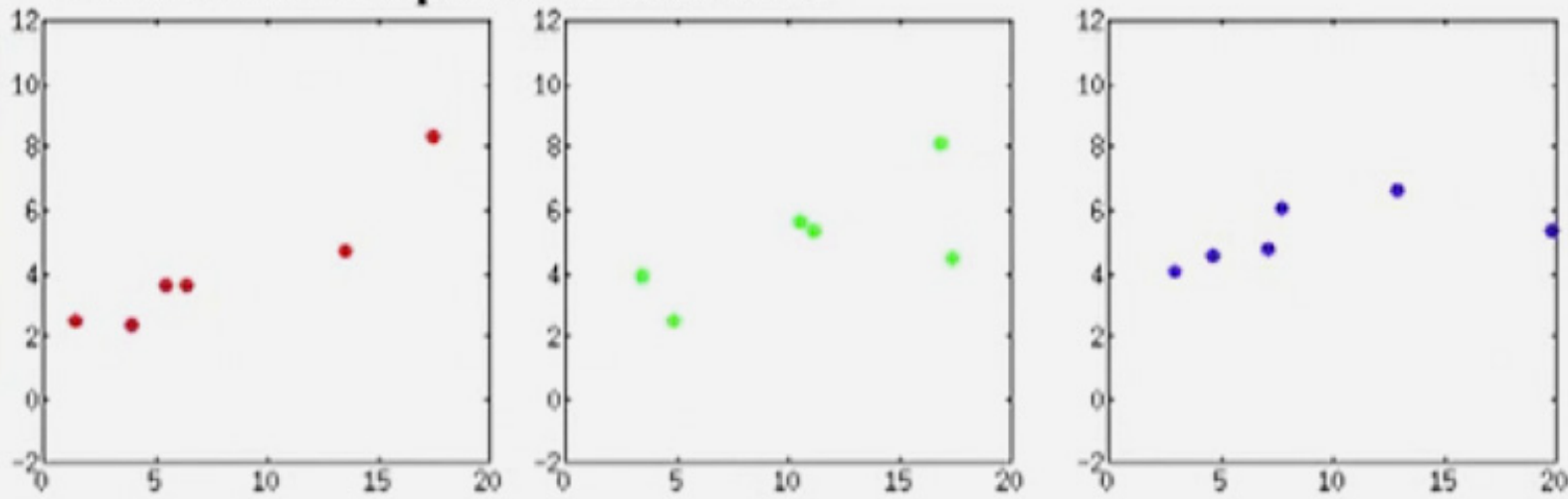
$$MSE(\bar{\theta}) = E(\bar{\theta} - \theta)^2$$
$$= Var(\bar{\theta}) + [E(\bar{\theta}) - \theta]^2. \qquad (3.20)$$

- The first term is the variance, while the second term is the squared bias. The Gauss-Markov theorem implies that the least squares estimator has the smallest mean squared error of all linear estimators with no bias.

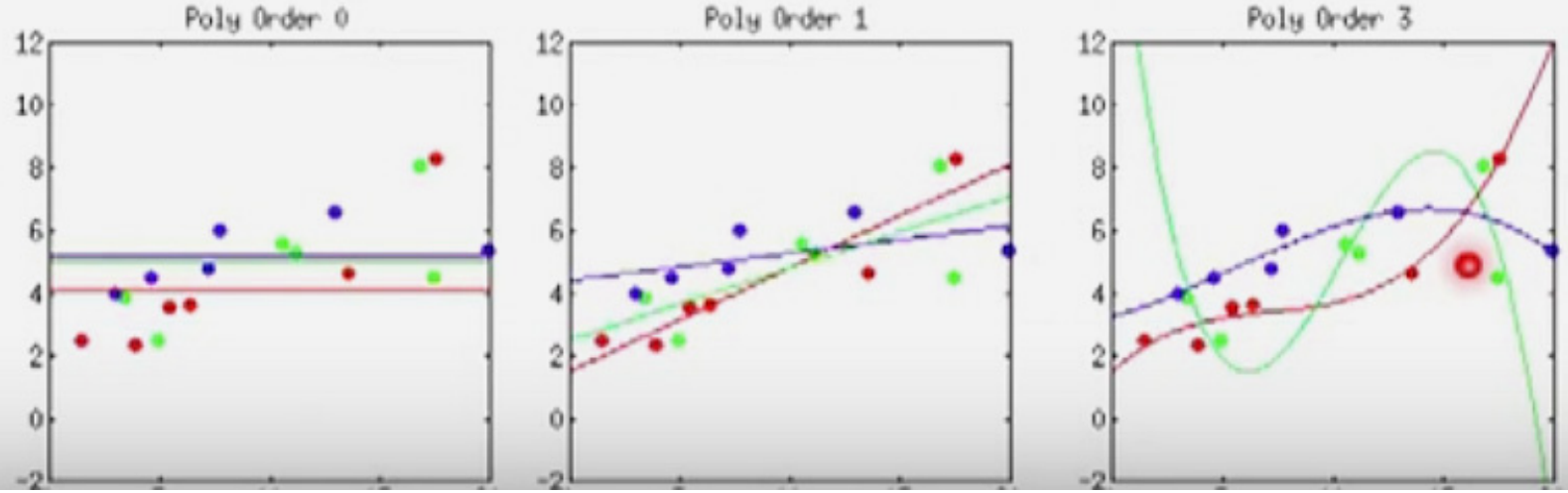
Bias & variance



Three different possible data sets:



Each would give different predictors for any polynomial degree:



Subset Selection

- There are two reasons why we are often not satisfied with the least squares estimates (3.6).
  - $\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$ (3.6)

❖ The first is $prediction\ accuracy$: the least squares estimates *(not just linear) <u>often</u>* have low bias but large variance. Prediction accuracy can sometimes be improved by shrinking or setting some coefficients to zero. By doing so we sacrifice a little bit of bias to reduce the variance of the predicted values, and hence may improve the overall prediction accuracy.

❖ The second reason is $interpretation$. With a large number of predictors, we often would like to determine a smaller subset that exhibit the strongest effects. In order to get the "big picture," we are willing to sacrifice some of the small details.

- We describe a number of approaches to variable subset selection with linear regression. In later sections we discuss shrinkage and hybrid approaches for controlling variance, as well as other dimension-reduction strategies. These all fall under the general heading $model\ selection$.

- With subset selection we retain only a subset of the variables, and eliminate the rest from the model. Least squares regression is used to estimate the coefficients of the inputs that are retained. There are a number of different strategies for choosing the subset.

$\boldsymbol{Best - Subset\ Selection}$

- Best subset regression finds for each $k \in \{0, 1, 2, \ldots, p\}$ the subset of size k that gives smallest residual sum of squares (3.2). An efficient algorithm— the leaps and bounds procedure (Furnivall and Wilson, 1974)—makes this feasible for p as large as 30 or 40.

- The lower boundary represents the models that are eligible for selection by the best-subsets approach. The best-subset curve (blue lower boundary in Figure 3.5) is necessarily decreasing, so cannot be used to select the subset size k.

- There are a number of criteria that one may use; typically we choose the smallest model that minimizes an estimate of the expected prediction error. E.g. cross-validation to estimate prediction error and select k; the AIC criterion is a popular alternative.

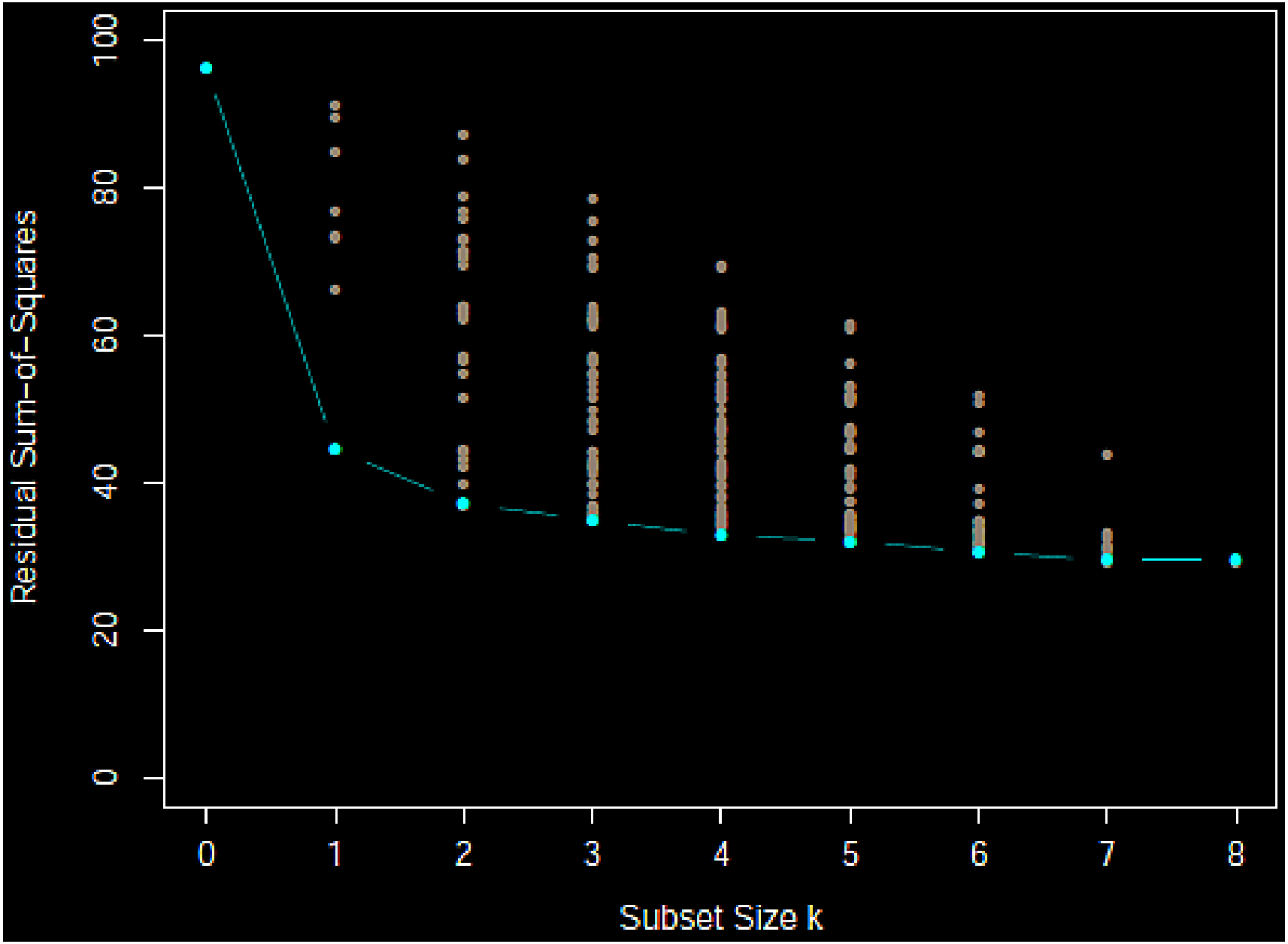


FIGURE 3.5. All possible subset models for an (the prostate cancer) example. At each subset size is shown the residual sum-of-squares for each model of that size.

Contd. - in part B;