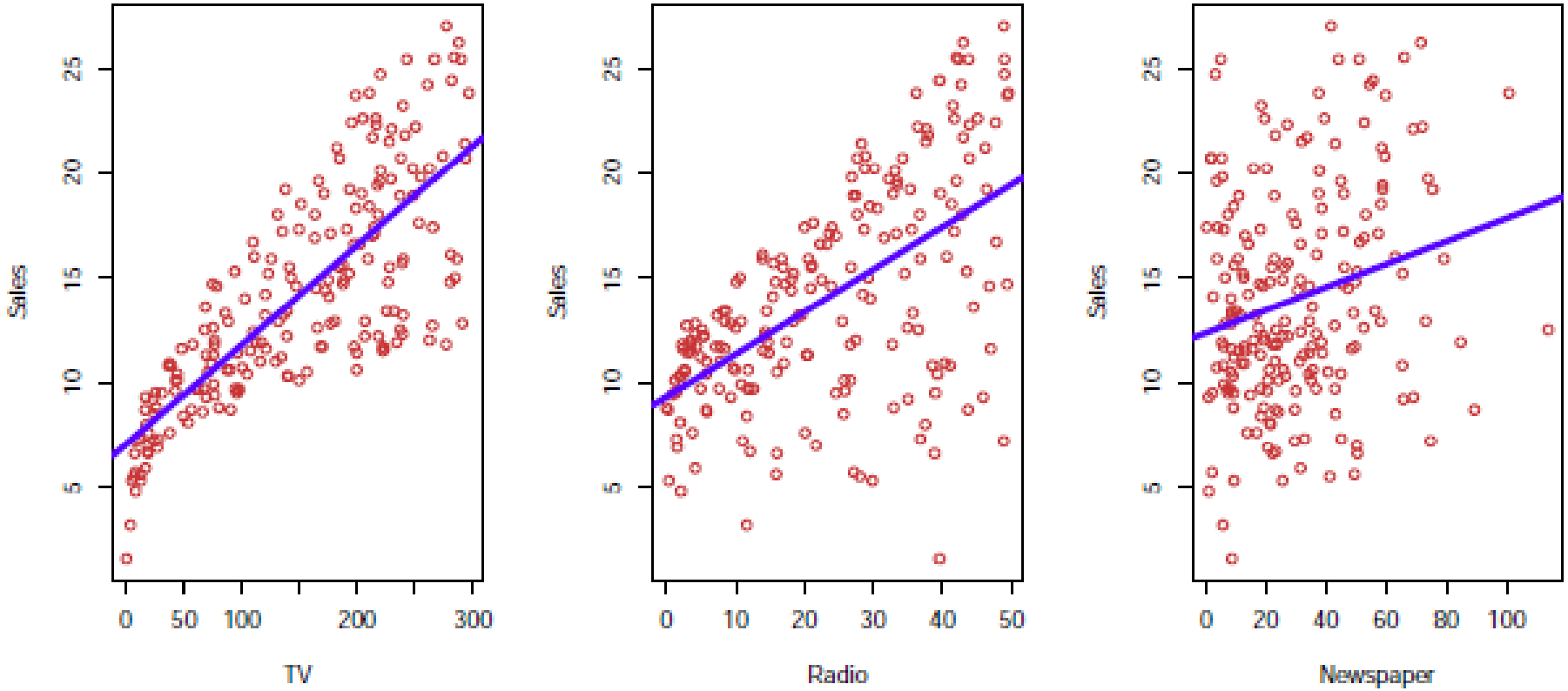
Linear Regression

CS-309

Sourav Kumar Dandapat

Statistical Learning

- Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product.
- The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.



Shown are `Sales` vs `TV`, `Radio` and `Newspaper`, with a blue linear-regression line fit separately to each.

Can we predict `Sales` using these three?

Perhaps we can do better using a model

$$\texttt{Sales} \approx f(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper})$$

Here `Sales` is a *response* or *target* that we wish to predict. We generically refer to the response as Y.
`TV` is a *feature*, or *input*, or *predictor*; we name it X_1.
Likewise name `Radio` as X_2, and so on.
We can refer to the *input vector* collectively as

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

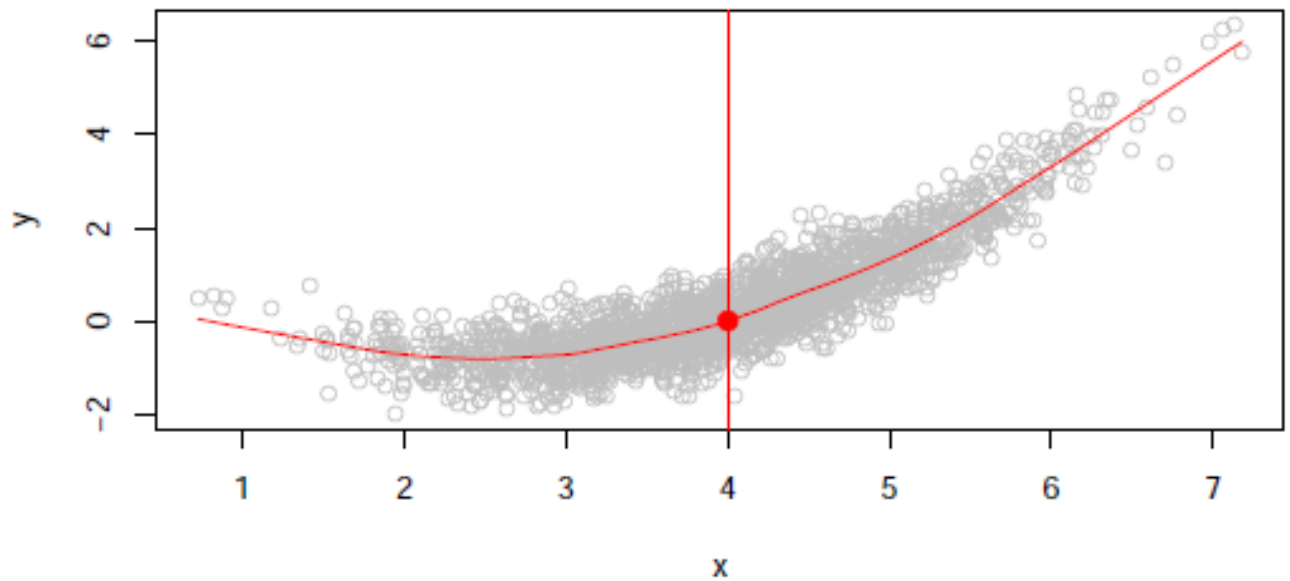
Now we write our model as

$$Y = f(X) + \epsilon$$

where ϵ captures measurement errors and other discrepancies.

What is $f(X)$ good for?

- With a good f we can make predictions of Y at new points $X = x$.
- We can understand which components of $X = (X_1, X_2, \ldots, X_p)$ are important in explaining Y, and which are irrelevant. e.g. `Seniority` and `Years of Education` have a big impact on `Income`, but `Marital Status` typically does not.
- Depending on the complexity of f, we may be able to understand how each component X_j of X affects Y.



Is there an ideal $f(X)$? In particular, what is a good value for $f(X)$ at any selected value of X, say $X = 4$? There can be many Y values at $X = 4$. A good value is

$$f(4) = E(Y|X = 4)$$

$E(Y|X = 4)$ means *expected value* (average) of Y given $X = 4$.

This ideal $f(x) = E(Y|X = x)$ is called the *regression function*.

The regression function $f(x)$

- Is also defined for vector X; e.g.
  $f(x) = f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$
- Is the *ideal* or *optimal* predictor of Y with regard to mean-squared prediction error: $f(x) = E(Y|X = x)$ is the function that minimizes $E[(Y - g(X))^2|X = x]$ over all functions g at all points $X = x$.
- $\epsilon = Y - f(x)$ is the *irreducible* error — i.e. even if we knew $f(x)$, we would still make errors in prediction, since at each $X = x$ there is typically a distribution of possible Y values.
- For any estimate $\hat{f}(x)$ of $f(x)$, we have

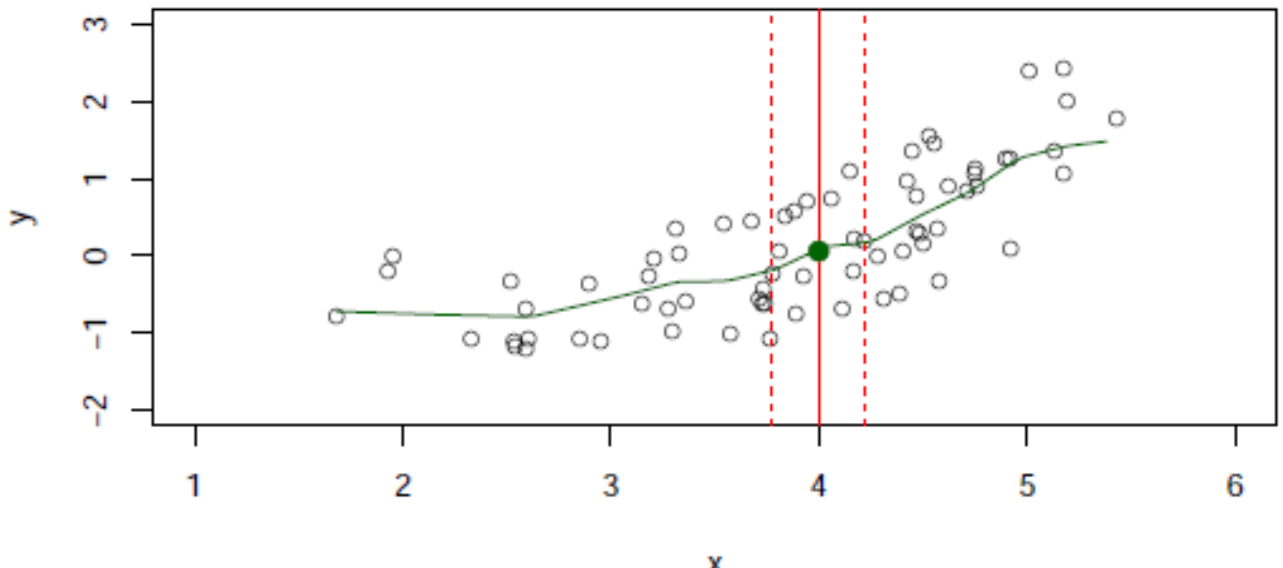
$$E[(Y - \hat{f}(X))^2|X = x] = \underbrace{[f(x) - \hat{f}(x)]^2}_{\textit{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\textit{Irreducible}}$$

How to estimate f

- Typically we have few if any data points with $X = 4$ exactly.
- So we cannot compute $E(Y|X = x)$!
- Relax the definition and let

$$\hat{f}(x) = \text{Ave}(Y|X \in \mathcal{N}(x))$$

where $\mathcal{N}(x)$ is some *neighborhood* of x.

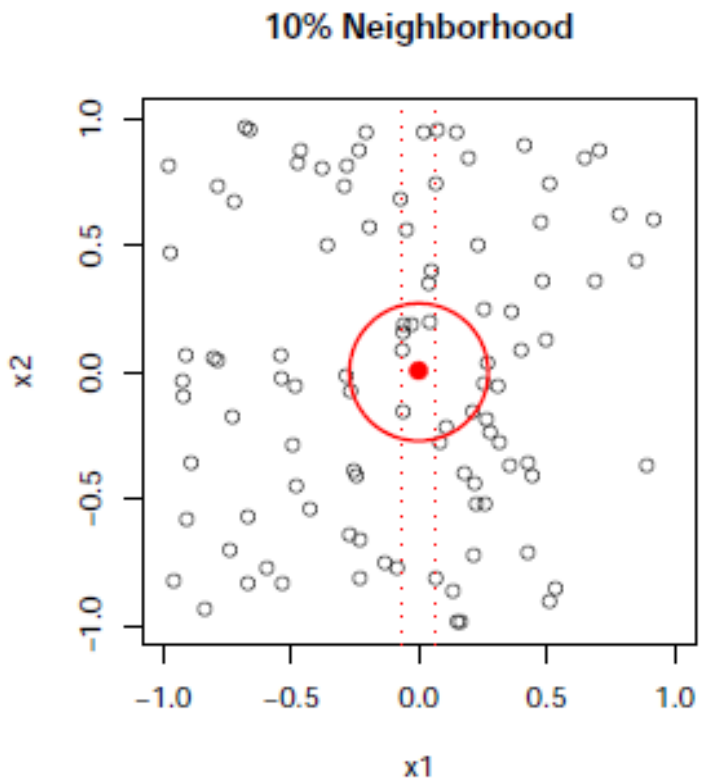


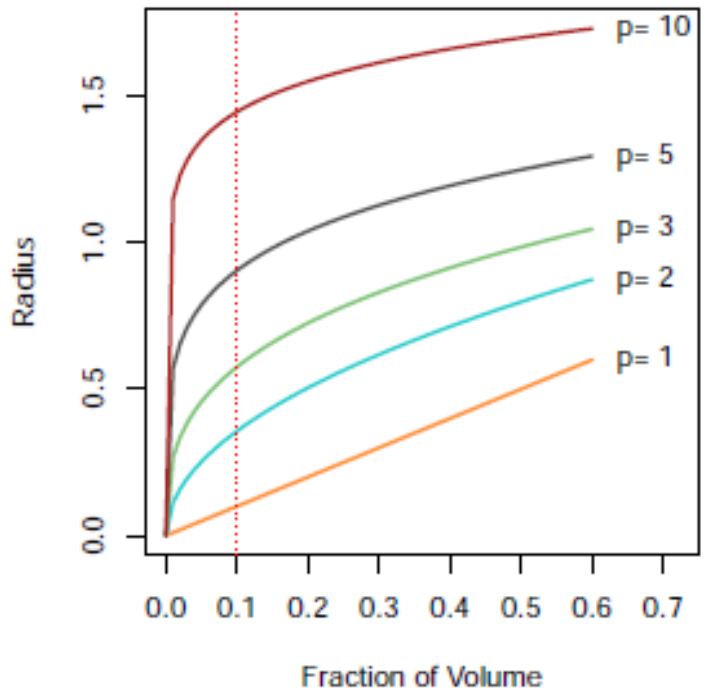
- Nearest neighbor averaging can be pretty good for small p — i.e. $p \leq 4$ and large-ish N.

- Nearest neighbor methods can be *lousy* when p is large. Reason: the *curse of dimensionality*. Nearest neighbors tend to be far away in high dimensions.

    - We need to get a reasonable fraction of the N values of y_i to average to bring the variance down—e.g. 10%.
    - A 10% neighborhood in high dimensions need no longer be local, so we lose the spirit of estimating $E(Y|X = x)$ by local averaging.

The curse of dimensionality





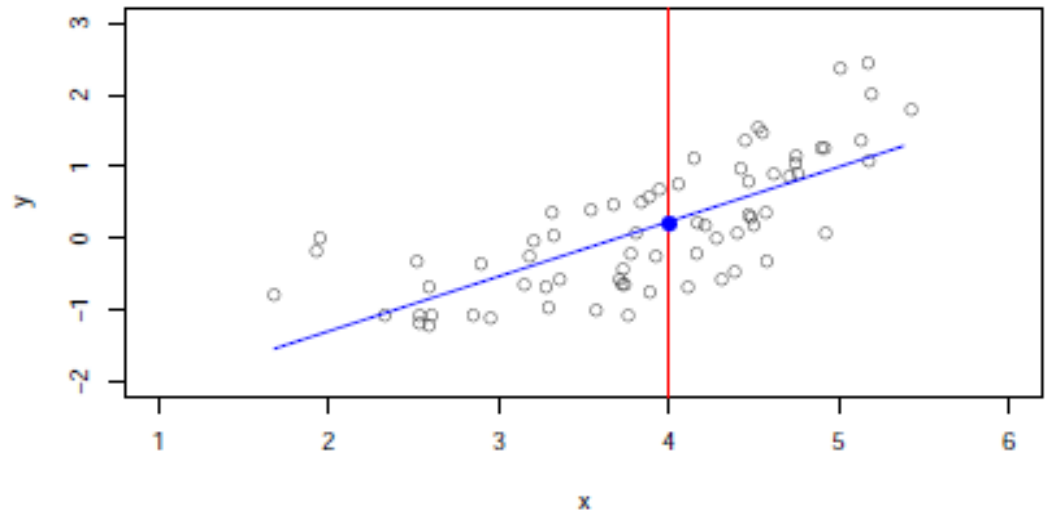
Parametric and structured models

The *linear* model is an important example of a parametric model:

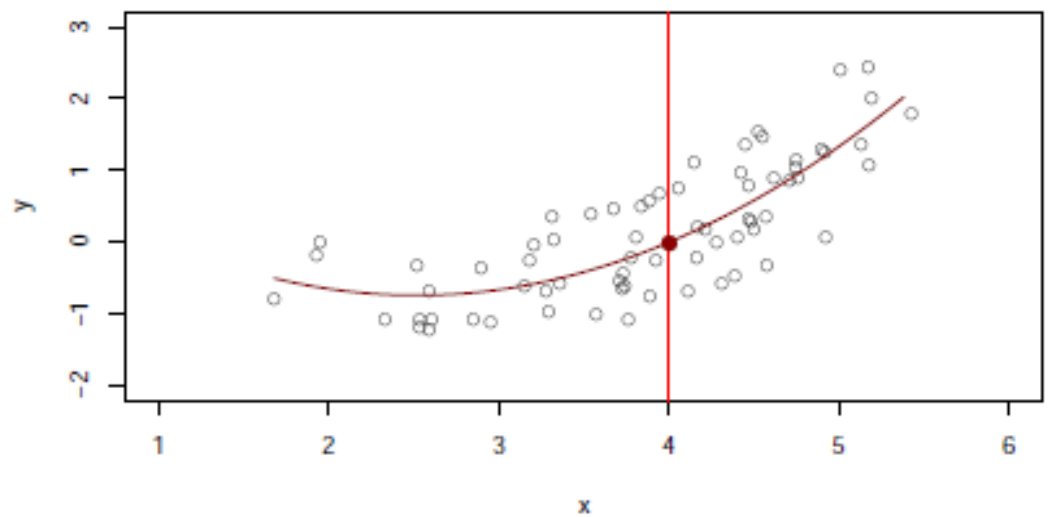
$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p.$$

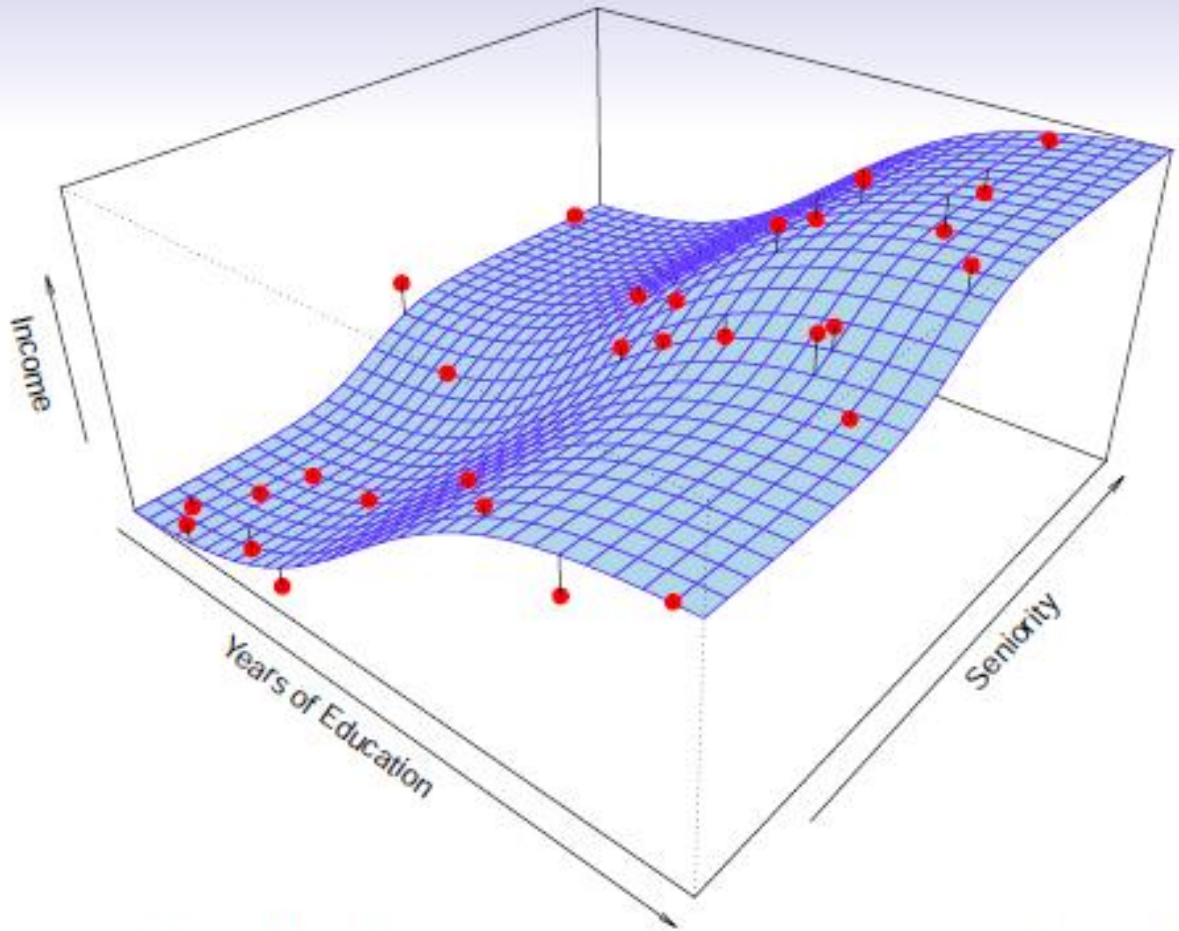
- A linear model is specified in terms of $p + 1$ parameters $\beta_0, \beta_1, \ldots, \beta_p$.
- We estimate the parameters by fitting the model to training data.
- Although it is *almost never correct*, a linear model often serves as a good and interpretable approximation to the unknown true function $f(X)$.

A linear model $\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ gives a reasonable fit here



A quadratic model $\hat{f}_Q(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$ fits slightly better.

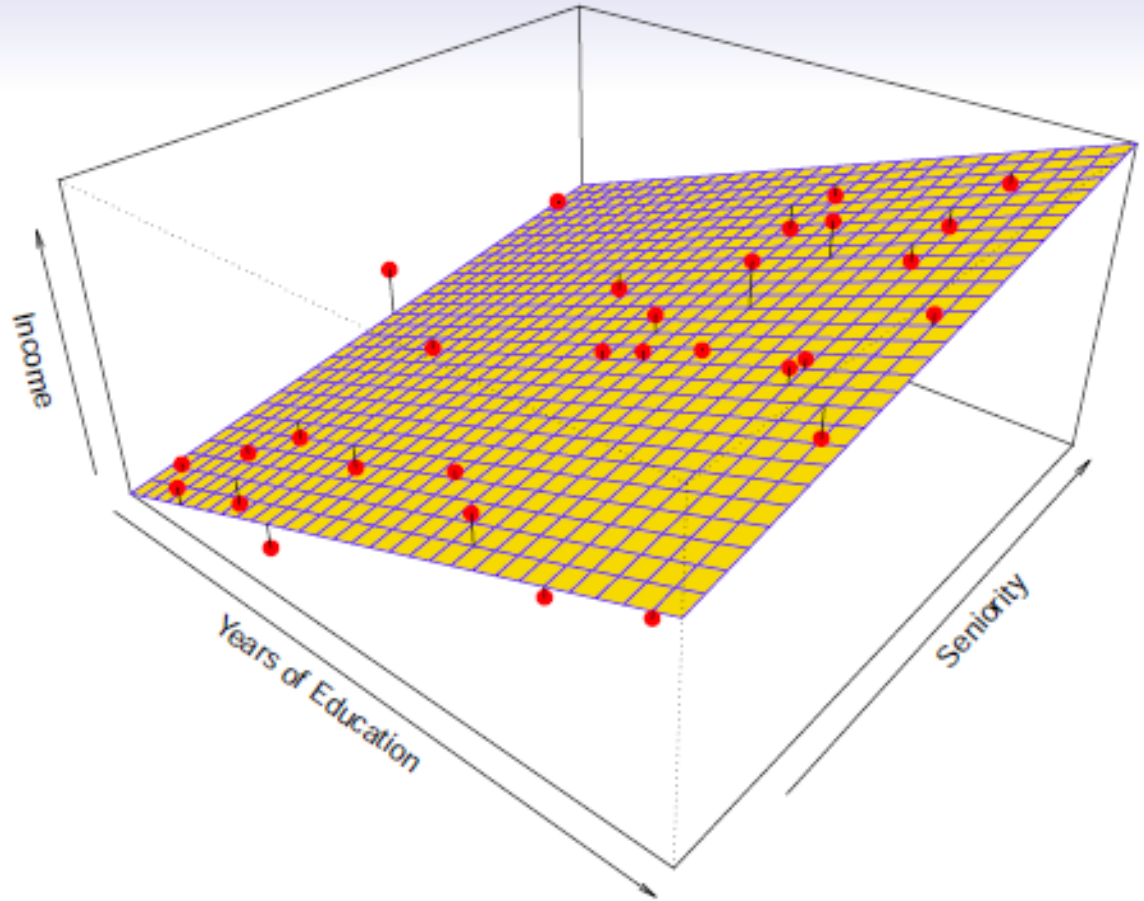




Simulated example. Red points are simulated values for `income` from the model

$$\texttt{income} = f(\texttt{education}, \texttt{seniority}) + \epsilon$$

f is the blue surface.

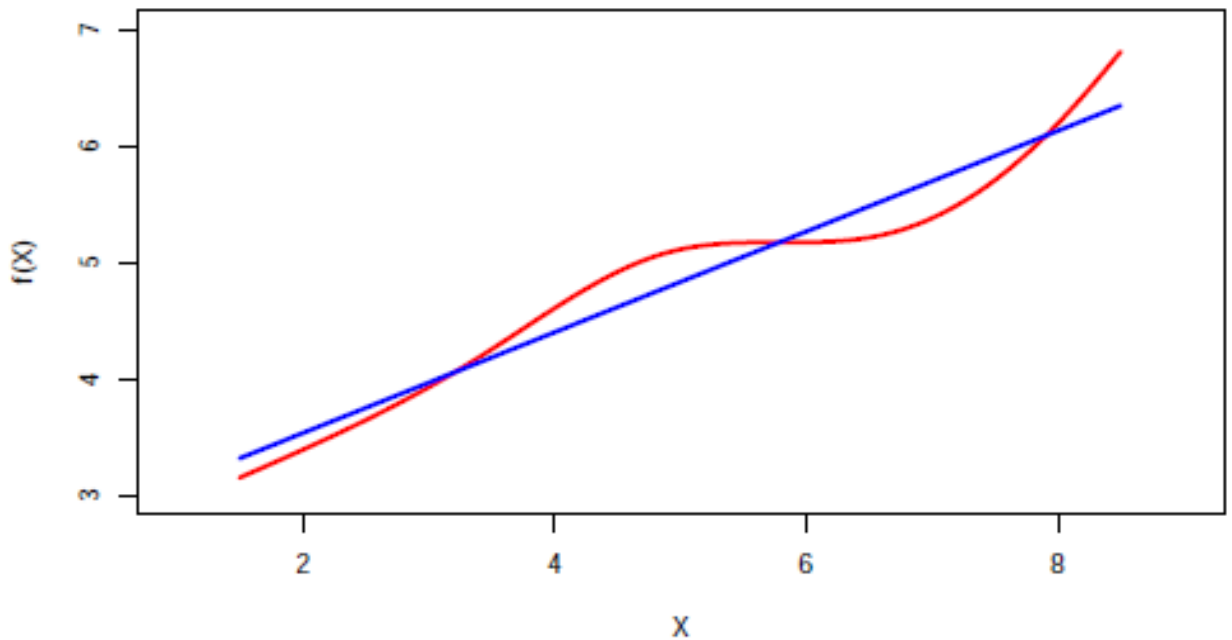


Linear regression model fit to the simulated data.

$$\hat{f}_L(\texttt{education}, \texttt{seniority}) = \hat{\beta}_0 + \hat{\beta}_1 \times \texttt{education} + \hat{\beta}_2 \times \texttt{seniority}$$

Linear regression

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $X_1, X_2, \ldots X_p$ is linear.
- True regression functions are never linear!



- although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

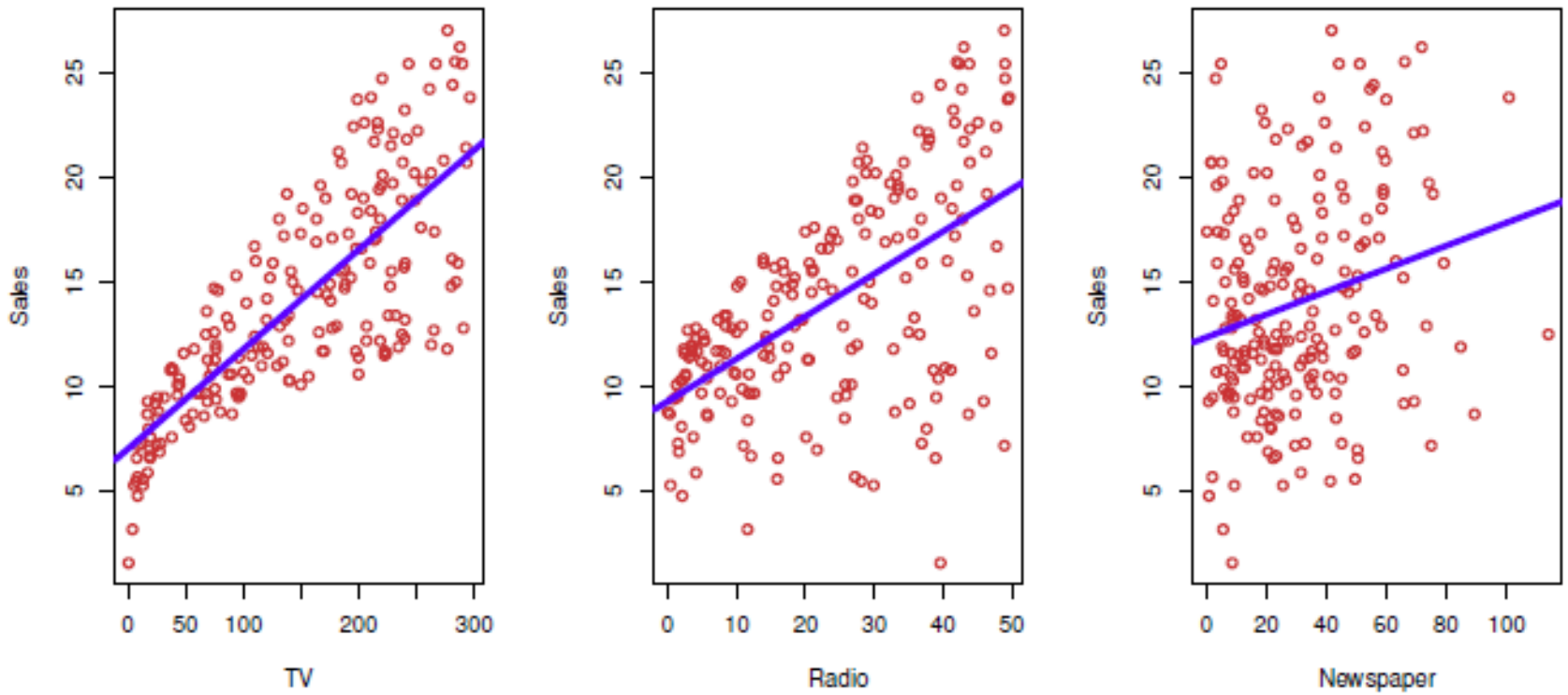
Linear regression for the advertising data

Consider the advertising data shown on the next slide.

Questions we might ask:

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Advertising data



Simple linear regression using a single predictor X.

- We assume a model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where β_0 and β_1 are two unknown constants that represent the *intercept* and *slope*, also known as *coefficients* or *parameters*, and ϵ is the error term.

- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future sales using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$. The *hat* symbol denotes an estimated value.

- For example, *X* may represent TV advertising and *Y* may represent sales. Then we can regress sales onto TV by fitting the model

$$\text{sales} \approx \beta 0 + \beta 1 \times \text{TV}$$

Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future sales using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

- In practice, β0 and β1 are unknown. So before we can use the last equation to make predictions, we must use data to estimate the coefficients. Let

  $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ represent n observation pairs, each of which consists of a measurement of X and a measurement of Y .

- In the Advertising example, this data set consists of the TV advertising budget and product sales in n = 200 different markets.

Estimation of the parameters by least squares

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith *residual*
- We define the *residual sum of squares* (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

Derivation of Parameters

- Least Squares (L-S):

  Minimize squared error

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2$$

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_0} = \frac{\partial \sum \left(y_i - \beta_0 - \beta_1 x_i \right)^2}{\partial \beta_0}$$

$$= -2\left(n\bar{y} - n\beta_0 - n\beta_1 \bar{x} \right)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Derivation of Parameters

- Least Squares (L-S):

  Minimize squared error

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_1} = \frac{\partial \sum (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_1}$$

$$= -2 \sum x_i (y_i - \beta_0 - \beta_1 x_i)$$

$$= -2 \sum x_i (y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i)$$

$$\beta_1 \sum x_i (x_i - \bar{x}) = \sum x_i (y_i - \bar{y})$$

$$\beta_1 \sum (x_i - \bar{x})(x_i - \bar{x}) = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

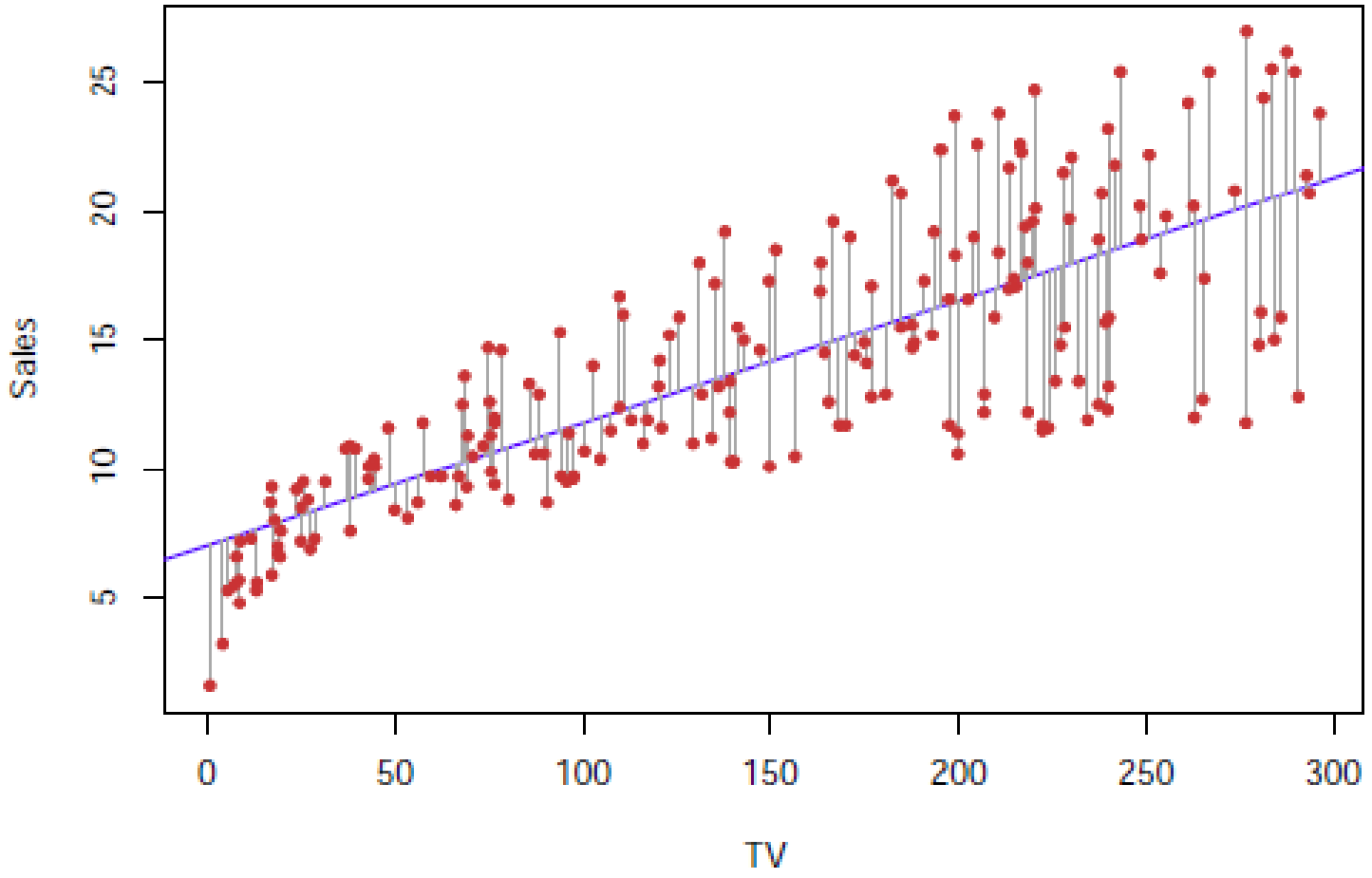
The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

Example: advertising data



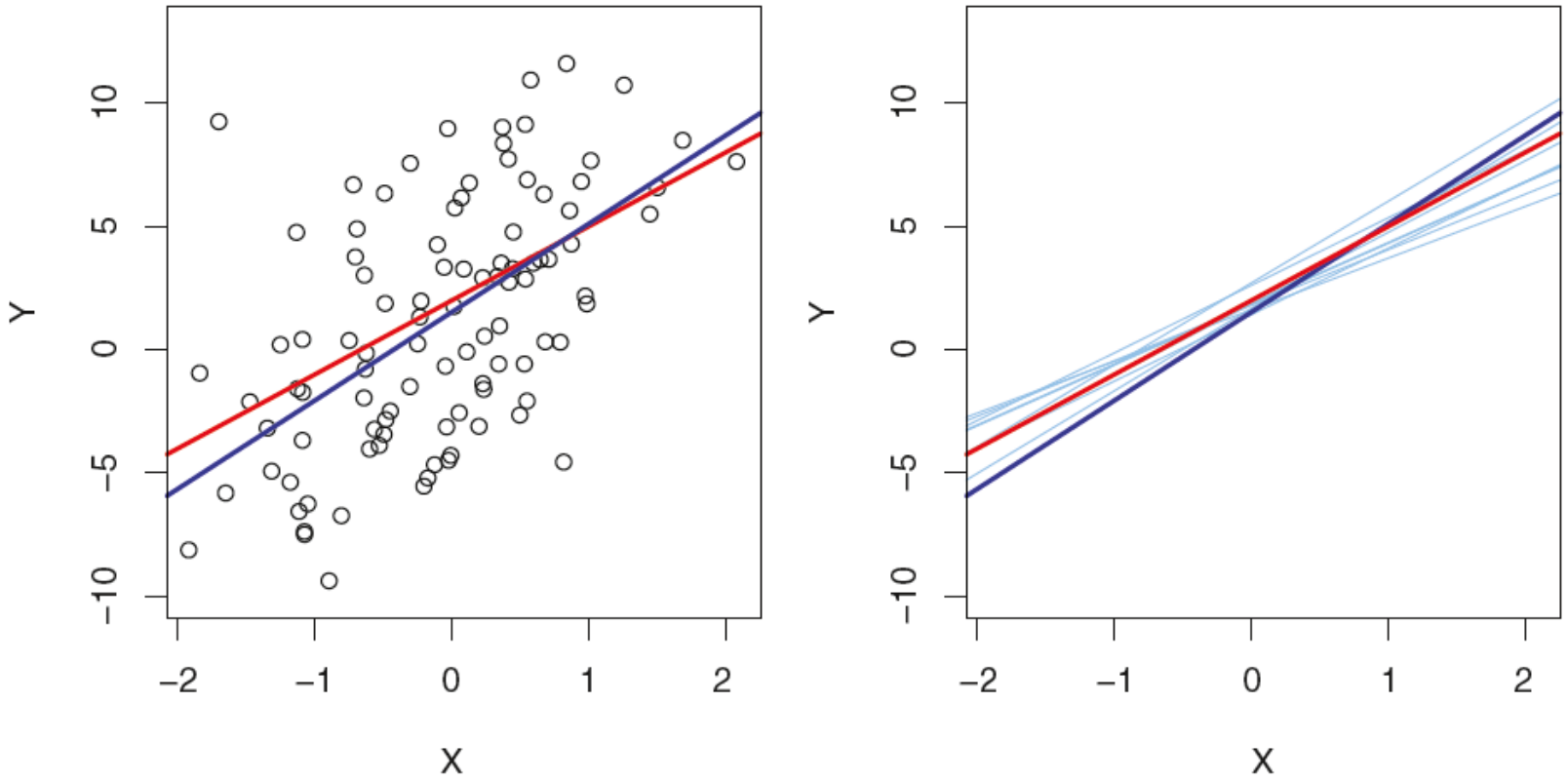
The least squares fit for the regression of `sales` onto `TV`.
In this case a linear fit captures the essence of the relationship

Assessing the Accuracy of the Coefficient Estimates

- we assume that the *true* relationship between *X* and *Y* takes the form $Y = f(X) + \xi$ for some unknown function *f*, where ξ is a normally distributed mean-zero random error term. If *f* is to be approximated by a linear function, then we can write this relationship as

$$Y = \beta_0 + \beta_1 X + \xi$$

- We created 100 random *X*s, and generated 100 corresponding *Y* s from the model $Y = 2 + 3X + \xi$, where ξ was generated from a normal distribution with mean zero.



A simulated data set. Left: *The red line represents the true relationship, f(X) = 2+3X, which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for f(X) based on the observed data, shown in black.* Right: *The population regression line is again shown in red, and the least squares line in dark blue. In light blue, ten least squares lines are shown, each computed on the basis of a separate random set of observations.*

- Let’s say Y is a random variable with population mean μ. Let us say, we try to estimate mean of Y using n number of observation. A reasonable estimate of μ is = ȳ, where ȳ = $\sum_{i=1}^{n} y_i$ is the sample mean.

Assessing the Accuracy of the Coefficient Estimates

- The standard error of an estimator reflects how it varies under repeated sampling. We have

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

  where $\sigma^2 = \mathrm{Var}(\epsilon)$

- These standard errors can be used to compute *confidence intervals*. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. It has the form

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1).$$

Confidence intervals — continued

That is, there is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

will contain the true value of β_1 (under a scenario where we got repeated samples like the present sample)

For the advertising data, the 95% confidence interval for β_1 is $[0.042, 0.053]$

Hypothesis testing

- Standard errors can also be used to perform *hypothesis tests* on the coefficients. The most common hypothesis test involves testing the *null hypothesis* of

  H_0 : There is no relationship between X and Y

  versus the *alternative hypothesis*

  H_A : There is some relationship between X and Y.

- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

Hypothesis testing — continued

- To test the null hypothesis, we compute a *t-statistic*, given by

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)},$$

- This will have a t-distribution with $n - 2$ degrees of freedom, assuming $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the *p-value*.

Results for the advertising data

	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	< 0.0001

Multiple Linear Regression

- Here our model is

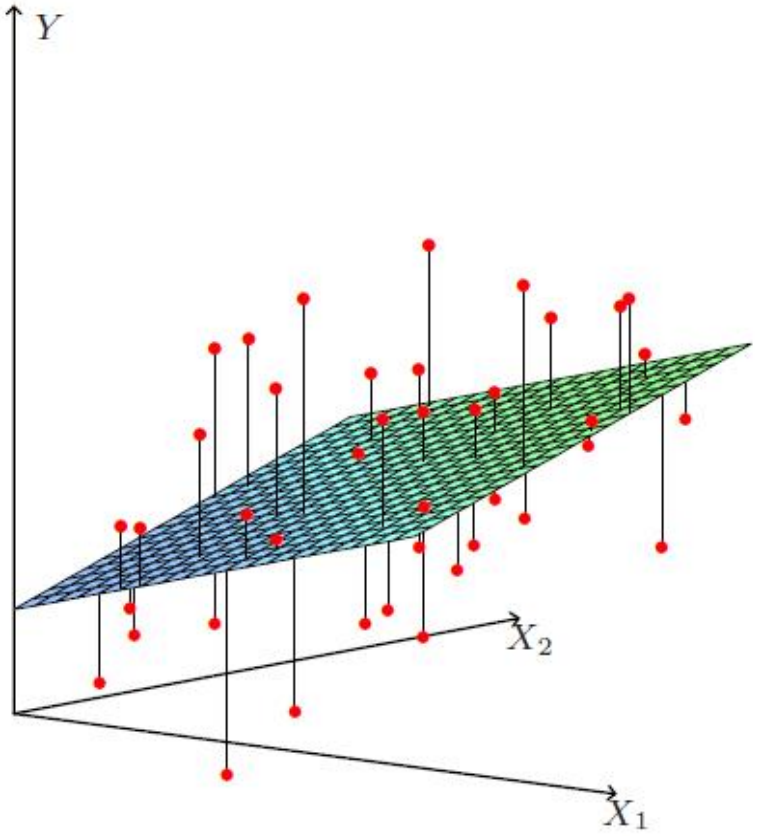
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

- We interpret β_j as the *average* effect on Y of a one unit increase in X_j, *holding all other predictors fixed*. In the advertising example, the model becomes

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

Interpreting regression coefficients

- The ideal scenario is when the predictors are uncorrelated — a *balanced design*:
  - Each coefficient can be estimated and tested separately.
  - Interpretations such as *"a unit change in X_j is associated with a β_j change in Y, while all the other variables stay fixed"*, are possible.
- Correlations amongst predictors cause problems:
  - The variance of all coefficients tends to increase, sometimes dramatically
  - Interpretations become hazardous — when X_j changes, everything else changes.
- *Claims of causality* should be avoided for observational data.



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 .

$$\begin{aligned}\mathrm{RSS}(\beta) &= \sum_{i=1}^{N}(y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N}\Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2\end{aligned}$$

$$\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta).$$

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

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- $Y=\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}$ $X=\begin{bmatrix} 1 & x_{11} & . & . & x_{1k} \\ 1 & x_{21} & . & . & x_{2k} \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{n1} & . & . & x_{nk} \end{bmatrix}$ $\beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{bmatrix}$
- Dimension of Y is (nX1), X is (nX(k+1))=(nXp), β is (pX1)
- $RSS = (Y-X\beta)^T(Y-X\beta)$

  $=(YT- \beta^T X^T)(Y-X\beta)$

  $=Y^TY-Y^TX\beta- \beta^TX^TY- \beta^TX^TX\beta$ ---(i)

- $Y=\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}$ $X=\begin{bmatrix} 1 & x_{11} & . & . & x_{1k} \\ 1 & x_{21} & . & . & x_{2k} \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{n1} & . & . & x_{nk} \end{bmatrix}$ $\beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{bmatrix}$

- $X\beta=\begin{bmatrix} \beta_0 & +\beta_1 x_{11} & . & . & +\beta_k x_{1k} \\ \beta_0 & +\beta_1 x_{21} & . & . & +\beta_k x_{2k} \\ . & . & . & . & . \\ . & . & . & . & . \\ \beta_0 & +\beta_1 x_{n1} & . & . & +\beta_k x_{nk} \end{bmatrix}$

- $Y^T X\beta=[y_1\ y_2 \quad y_n] \begin{bmatrix} \beta_0 & +\beta_1 x_{11} & . & . & +\beta_k x_{1k} \\ \beta_0 & +\beta_1 x_{21} & . & . & +\beta_k x_{2k} \\ . & . & . & . & . \\ . & . & . & . & . \\ \beta_0 & +\beta_1 x_{n1} & . & . & +\beta_k x_{nk} \end{bmatrix}$

$$Y^T X\beta = [y_1\ y_2 \quad y_n] \begin{bmatrix} \beta_0 & +\beta_1 x_{11} & . & . & +\beta_k x_{1k} \\ \beta_0 & +\beta_1 x_{21} & . & . & +\beta_k x_{2k} \\ . & . & . & . & . \\ . & . & . & . & . \\ \beta_0 & +\beta_1 x_{n1} & . & . & +\beta_k x_{nk} \end{bmatrix}$$

$$= \beta_0(y_1 + y_2 \quad \ldots + y_n) + \beta_1(x_{11}y_1 + x_{21}y_2 + \ldots + x_{n1}y_n) + \ldots$$

$$+\beta_k\,(x_{1k}y_1 + x_{2k}y_2 + \ldots + x_{nk}y_n)$$

$$= [\beta_0\ \beta_1 \ldots\ \beta_k] \begin{bmatrix} y_1 & +y_2 & . & . & +y_n \\ x_{11}y_1 & +x_{21}y_2 & . & . & +x_{n1}y_n \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{1k}y_1 & +x_{2k}y_2 & . & . & +x_{nk}y_n \end{bmatrix}$$

$$= \beta^T X^T Y \quad \text{-----(ii)}$$

- $\frac{\partial(\beta^T X^T Y)}{\partial \beta} = \frac{\partial(\beta^T M)}{\partial \beta}$ [assuming $X^T Y = M$ (pX1)]
- $\beta^T M = \beta_0 m_0 + \beta_1 m_1 + \ldots + \beta_k m_k$
- $\frac{\partial(\beta^T M)}{\partial \beta} = \begin{bmatrix} \frac{\partial(\beta^T M)}{\partial \beta_0} \\ \frac{\partial(\beta^T M)}{\partial \beta_1} \\ . \\ . \\ \frac{\partial(\beta^T M)}{\partial \beta_k} \end{bmatrix} = \begin{bmatrix} m_0 \\ m_1 \\ . \\ . \\ m_k \end{bmatrix} = M = X^T Y$ ---(iii)

- X^TX is a symmetric matrix $(X^TX)^T = X^TX = A$ (pXp)
- $\beta^T A\beta = [\beta_1\ \beta_2]\begin{bmatrix} a_{11} & a \\ a & a_{22} \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = [\beta_1\ \beta_2]\begin{bmatrix} a_{11}\beta_1 + a\beta_2 \\ a\beta_1 + a_{22}\beta_2 \end{bmatrix} = [a_{11}\beta_1^2 + 2a\,\beta_1\beta_2 + a_{22}\beta_2^2]$
- $\frac{\partial(\beta^T A\beta)}{\partial\beta} = \begin{bmatrix} \frac{\partial(a_{11}\beta_1^2 + 2a\,\beta_1\beta_2 + a_{22}\beta_2^2)}{\partial\beta_1} \\ \frac{\partial(a_{11}\beta_1^2 + 2a\,\beta_1\beta_2 + a_{22}\beta_2^2)}{\partial\beta_2} \end{bmatrix} = \begin{bmatrix} 2a_{11}\beta_1 + 2a\beta_2 \\ 2a\beta_1 + 2a_{22}\beta_2 \end{bmatrix} = 2A\beta$ –(iv)

- From eqn (i)
- $RSS = Y^TY - Y^TX\beta - \beta^TX^TY - \beta^TX^TX\beta$
- $= Y^TY - 2\beta^TX^TY - \beta^TX^TX\beta$ [using eqn (ii), $Y^TX\beta = \beta^TX^TY$]
- $\frac{\partial(RSS)}{\partial\beta} = 2X^TY - 2X^TX\beta = 0$
- $X^TX\beta = X^TY$
- $\beta = (X^TX)^{-1}X^TY$

Compute β with given X and Y

$$Y=\begin{bmatrix}10\\20\\30\\40\\50\end{bmatrix},\quad X=\begin{bmatrix}1 & 5\\1 & 7\\1 & 10\\1 & 12\\1 & 20\end{bmatrix}$$

$$Y=\begin{bmatrix}10\\20\\30\\40\\50\end{bmatrix},\quad X=\begin{bmatrix}1 & 5\\1 & 7\\1 & 10\\1 & 12\\1 & 20\end{bmatrix}$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\mathrm{X}^T\mathrm{X} = \begin{bmatrix}1 & 1 & 1 & 1 & 1\\5 & 7 & 10 & 12 & 20\end{bmatrix}\begin{bmatrix}1 & 5\\1 & 7\\1 & 10\\1 & 12\\1 & 20\end{bmatrix}$$

$$= \begin{bmatrix}5 & 54\\54 & 718\end{bmatrix}$$

$$(\mathrm{X}^T\mathrm{X})^{-1} = \frac{1}{|X^TX|}adj(X^TX)$$

$$= \frac{1}{674}\begin{bmatrix}718 & -54\\-54 & 5\end{bmatrix}$$

$$= \begin{bmatrix}1.07 & -0.08\\-0.08 & 0.007\end{bmatrix}$$

$$\mathrm{X^TY} = \begin{bmatrix}1 & 1 & 1 & 1 & 1\\5 & 7 & 10 & 12 & 20\end{bmatrix}\begin{bmatrix}10\\20\\30\\40\\50\end{bmatrix}$$

$$= \begin{bmatrix}150\\1970\end{bmatrix}$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{bmatrix}1.07 & -0.08\\-0.08 & 0.007\end{bmatrix}\begin{bmatrix}150\\1970\end{bmatrix}$$

$$= \begin{bmatrix}2.7\\1.775\end{bmatrix}$$

Results for advertising data

	Coefficient	Std. Error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Correlations:

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

- while the newspaper regression coefficient estimate in simple linear regression was significantly non-zero, the coefficient estimate for newspaper in the multiple regression model is close to zero, and the corresponding p-value is no longer significant, with a value around 0.86
- This illustrates that the simple and multiple regression coefficients can be quite different.
- This difference stems from the fact that in the simple regression case, the slope term represents the average effect of a $1,000 increase in newspaper advertising, ignoring other predictors such as TV and radio.

- In contrast, in the multiple regression setting, the coefficient for newspaper represents the average effect of increasing newspaper spending by $1,000 while holding TV and radio fixed.
- Note that the correlation between radio and newspaper is 0.35.
- This reveals a tendency to spend more on newspaper advertising in markets where more is spent on radio advertising.
- in a simple linear regression which only examines sales versus newspaper, we will observe that higher values of newspaper tend to be associated with higher values of sales, even though newspaper advertising does not actually affect sales.

Is at least one predictor useful?

For the first question, we can use the F-statistic

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)} \sim F_{p,n-p-1}$$

$$\text{TSS} = \sum(y_i - \bar{y})^2 \qquad \text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

Deciding on the important variables

- The most direct approach is called *all subsets* or *best subsets* regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.
- However we often can't examine all possible models, since they are 2^p of them; for example when $p = 40$ there are over a billion models!
  Instead we need an automated approach that searches through a subset of them. We discuss two commonly use approaches next.

Forward selection

- Begin with the *null model* — a model that contains an intercept but no predictors.
- Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
- Add to that model the variable that results in the lowest RSS amongst all two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

Backward selection

- Start with all variables in the model.
- Remove the variable with the largest p-value — that is, the variable that is the least statistically significant.
- The new $(p-1)$-variable model is fit, and the variable with the largest p-value is removed.
- Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold.

Model Fit

The quality of a linear regression fit is typically assessed using two related quantities: the *residual standard error* (RSE) and the R^2 statistic.

- *R-squared* or fraction of variance explained is

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *total sum of squares*.

- It can be shown that in this simple linear regression setting that $R^2 = r^2$, where r is the correlation between X and Y:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

- The RSE provides an absolute measure of lack of fit of the model to the data. But since it is measured in the units of Y , it is not always clear what constitutes a good RSE.
- The R^2 statistic provides an alternative measure of fit. It takes the form of a *proportion*—the proportion of variance explained—and so it always takes on a value between 0 and 1, and is independent of the scale of Y .
- TSS measures the total variance in the response Y , and can be squares thought of as the amount of variability inherent in the response before the regression is performed.

- In contrast, RSS measures the amount of variability that is left unexplained after performing the regression.
- Hence, TSS−RSS measures the amount of variability in the response that is explained (or removed) by performing the regression, and R^2 measures the *proportion of variability in Y that can be explained using X*.
- An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression.

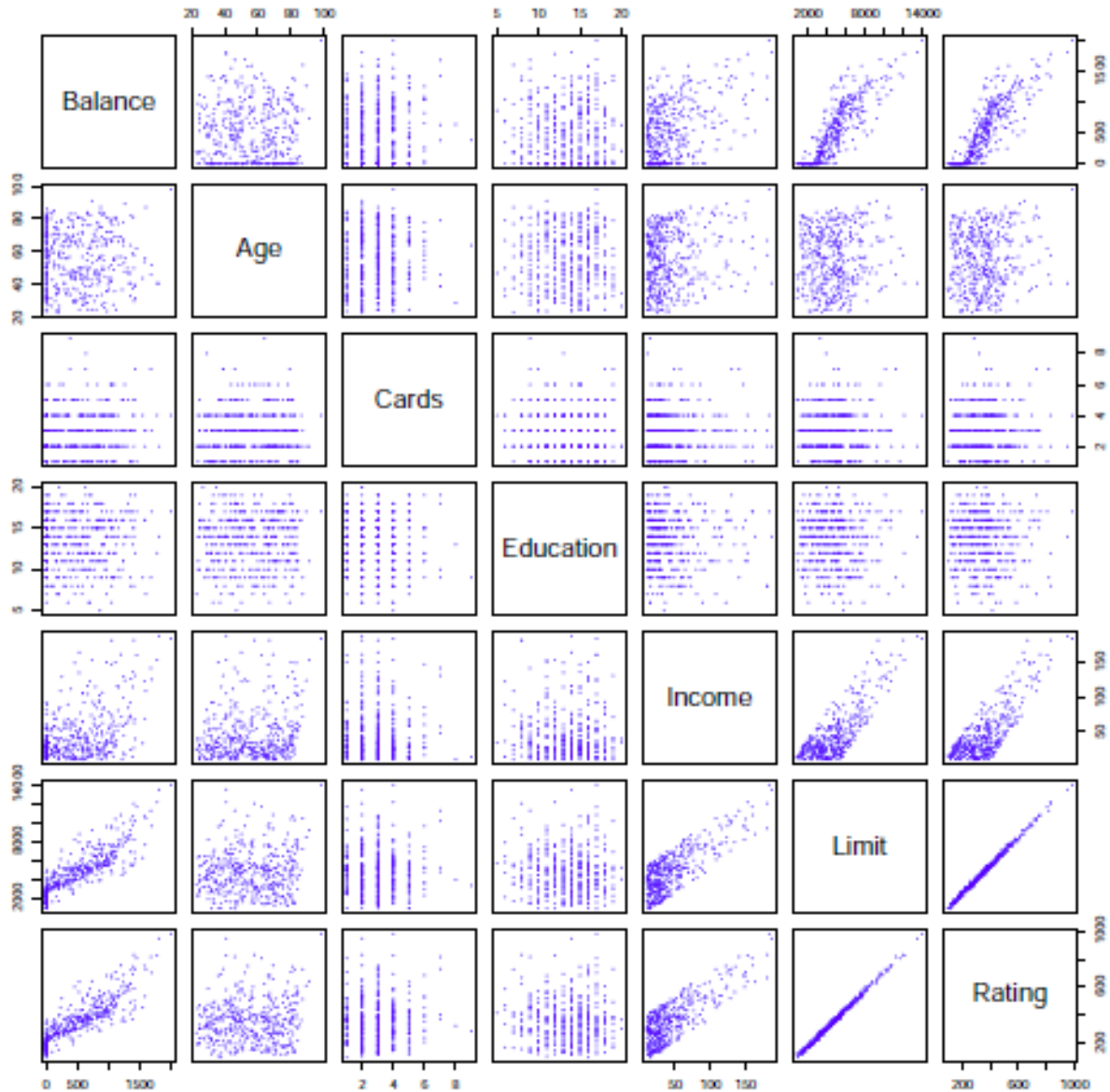
Other Considerations in the Regression Model

Qualitative Predictors

- Some predictors are not *quantitative* but are *qualitative*, taking a discrete set of values.
- These are also called *categorical* predictors or *factor variables*.
- See for example the scatterplot matrix of the credit card data in the next slide.

  In addition to the 7 quantitative variables shown, there are four qualitative variables: `gender`, `student` (student status), `status` (marital status), and `ethnicity` (Caucasian, African American (AA) or Asian).

Credit Card Data



Qualitative Predictors — continued

Example: investigate differences in credit card balance between males and females, ignoring the other variables. We create a new variable

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Resulting model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Intrepretation?

Credit card data — continued

Results for gender model:

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

- Alternatively, instead of a 0/1 coding scheme, we could create a dummy variable which will take value 1/-1 and use this in regression

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ -1 & \text{if } i\text{th person is male} \end{cases}$$

- This results in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 - \beta_1 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Qualitative predictors with more than two levels

- With more than two levels, we create additional dummy variables. For example, for the `ethnicity` variable we create two dummy variables. The first could be

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian,} \end{cases}$$

and the second could be

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian.} \end{cases}$$

Qualitative predictors with more than two levels — continued.

- Then both of these variables can be used in the regression equation, in order to obtain the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

- There will always be one fewer dummy variable than the number of levels. The level with no dummy variable — African American in this example — is known as the *baseline*.

Results for ethnicity

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`ethnicity[Asian]`	-18.69	65.02	-0.287	0.7740
`ethnicity[Caucasian]`	-12.50	56.68	-0.221	0.8260

Lasso Regression

- Lasso regression is a regularization technique.
- This model uses shrinkage.
- LASSO regression introduces an additional penalty term based on the absolute values of the coefficients.
- The L1 regularization term is the sum of the absolute values of the coefficients multiplied by a tuning parameter λ
- $L1 = \lambda*(|\beta_1|+|\beta_2|+\ldots+|\beta_p|)$
- Objective function is to minimize RSS +L1

Ridge Regression

- Ridge regression shrinks the regression coefficients by imposing a penalty on their size.

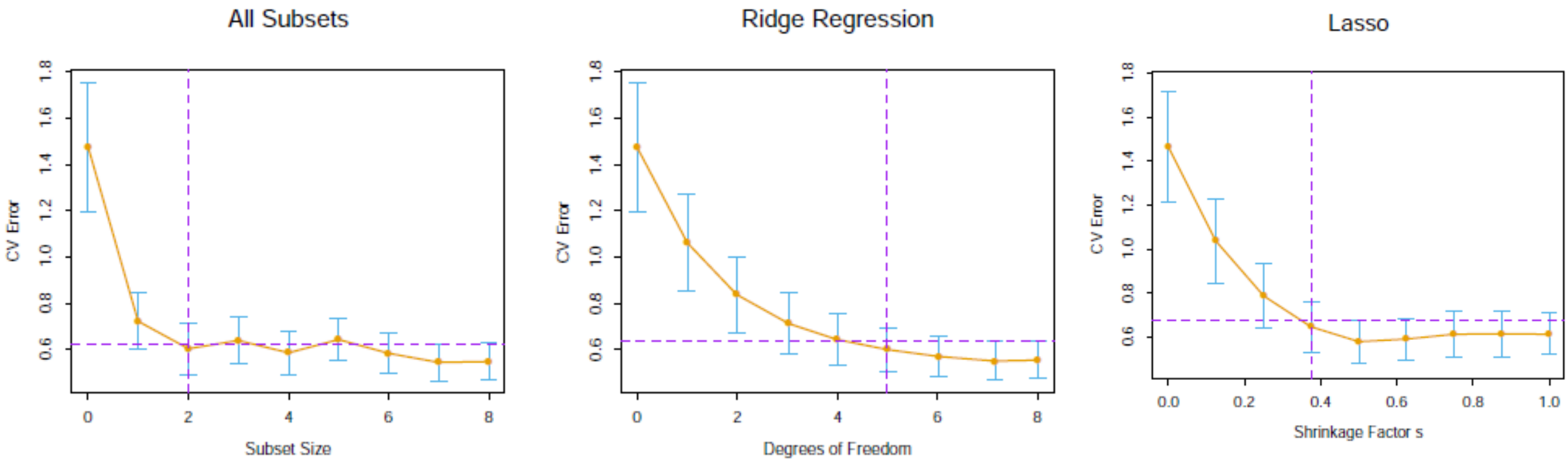
$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

- Here λ>= 0 is a complexity parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage.
- The coefficients are shrunk toward zero.

- An equivalent way to write the ridge problem is

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2$$

$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 \leq t,$$


All Subsets
CV Error
0.6
0.8
1.0
1.2
1.4
1.6
1.8
0
2
4
6
8
Subset Size
Ridge Regression
CV Error
0.6
0.8
1.0
1.2
1.4
1.6
1.8
0
2
4
6
8
Degrees of Freedom
Lasso
CV Error
0.6
0.8
1.0
1.2
1.4
1.6
1.8
0.0
0.2
0.4
0.6
0.8
1.0
Shrinkage Factor s

Extensions of the Linear Model

Removing the additive assumption: *interactions* and *nonlinearity*

Interactions:

- In our previous analysis of the `Advertising` data, we assumed that the effect on `sales` of increasing one advertising medium is independent of the amount spent on the other media.
- For example, the linear model

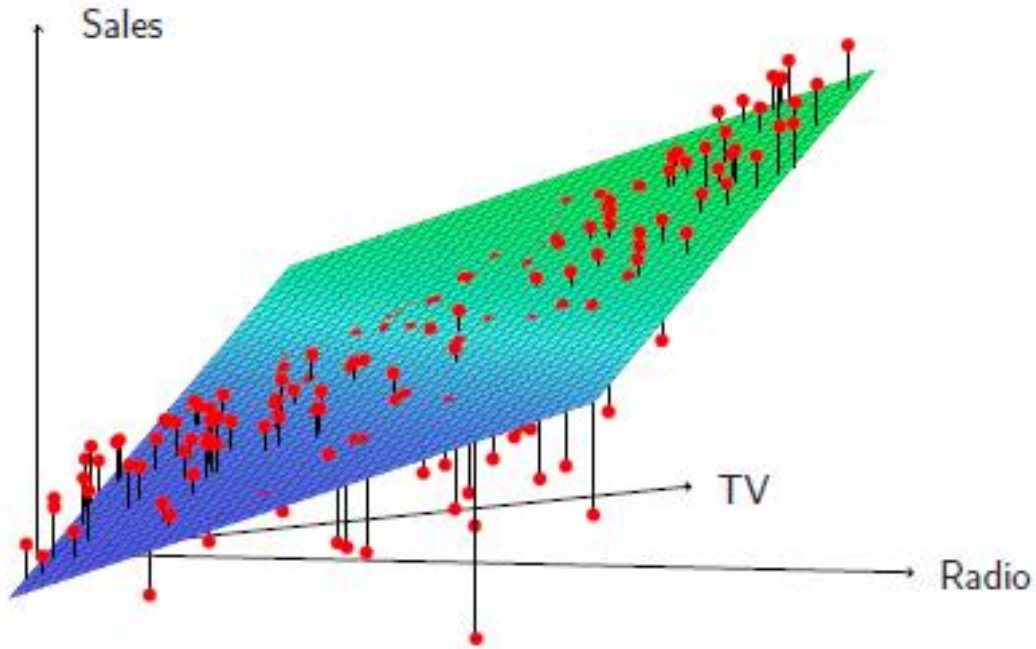
$$\widehat{\texttt{sales}} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper}$$

states that the average effect on `sales` of a one-unit increase in `TV` is always β_1, regardless of the amount spent on `radio`.

Interactions — continued

- But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for `TV` should increase as `radio` increases.
- In this situation, given a fixed budget of $\$100,000$, spending half on `radio` and half on `TV` may increase `sales` more than allocating the entire amount to either `TV` or to `radio`.
- In marketing, this is known as a *synergy* effect, and in statistics it is referred to as an *interaction* effect.

Interaction in the Advertising data?



When levels of either `TV` or `radio` are low, then the true `sales` are lower than predicted by the linear model.
But when advertising is split between the two media, then the model tends to underestimate `sales`.

Modelling interactions — Advertising data

Model takes the form

$$\begin{aligned} \texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon. \end{aligned}$$

Results:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Interpretation

- The results in this table suggests that interactions are important.
- The p-value for the interaction term `TV`×`radio` is extremely low, indicating that there is strong evidence for $H_A : \beta_3 \neq 0$.
- The R^2 for the interaction model is 96.8%, compared to only 89.7% for the model that predicts `sales` using `TV` and `radio` without an interaction term.

Interpretation — continued

- This means that $(96.8 - 89.7)/(100 - 89.7) = 69\%$ of the variability in `sales` that remains after fitting the additive model has been explained by the interaction term.
- The coefficient estimates in the table suggest that an increase in TV advertising of $\$1,000$ is associated with increased sales of
  $(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units.
- An increase in radio advertising of $\$1,000$ will be associated with an increase in sales of
  $(\hat{\beta}_2 + \hat{\beta}_3 \times \texttt{TV}) \times 1000 = 29 + 1.1 \times \texttt{TV}$ units.

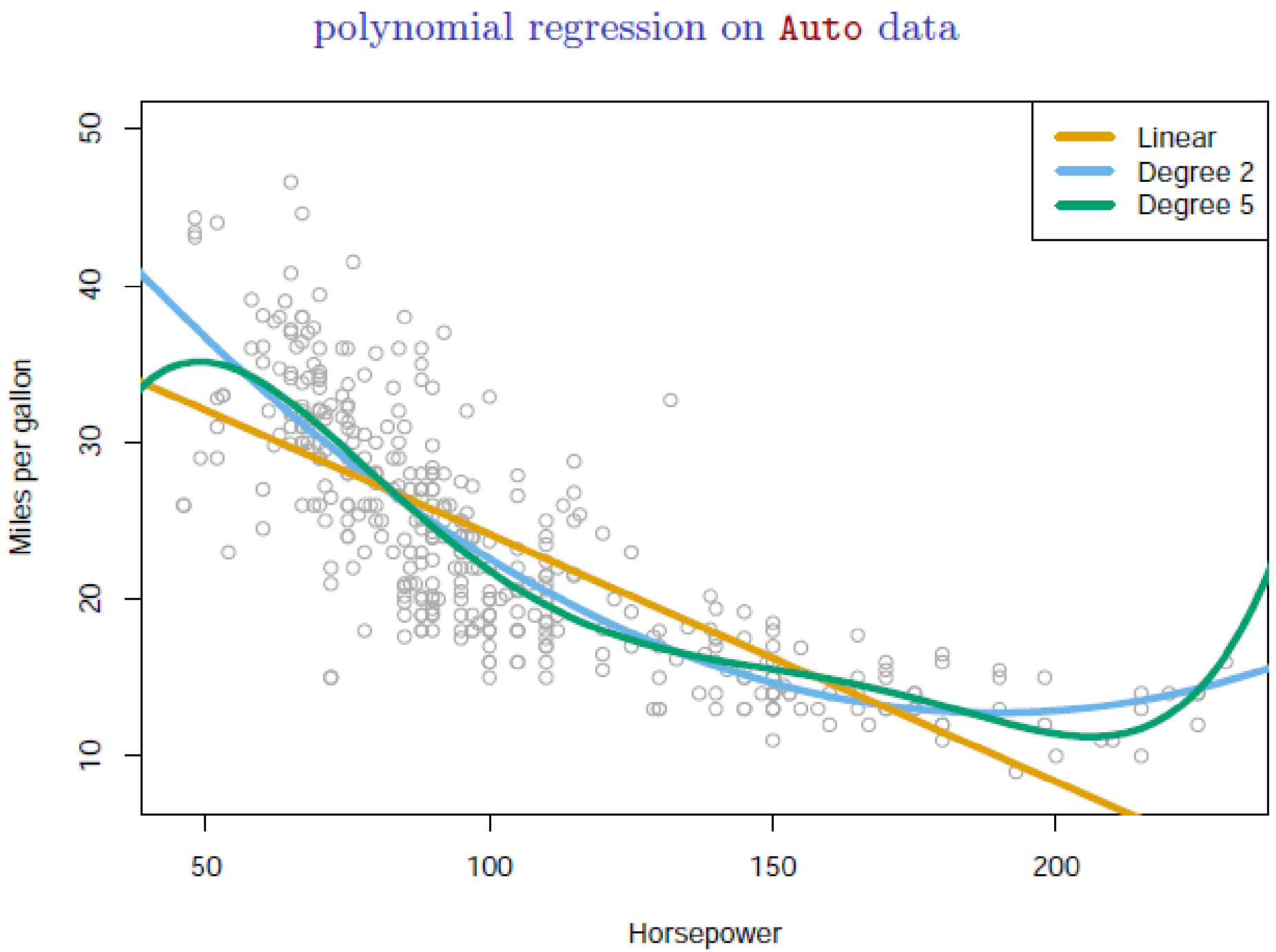
Hierarchy

- Sometimes it is the case that an interaction term has a very small p-value, but the associated main effects (in this case, `TV` and `radio`) do not.
- The *hierarchy principle*:

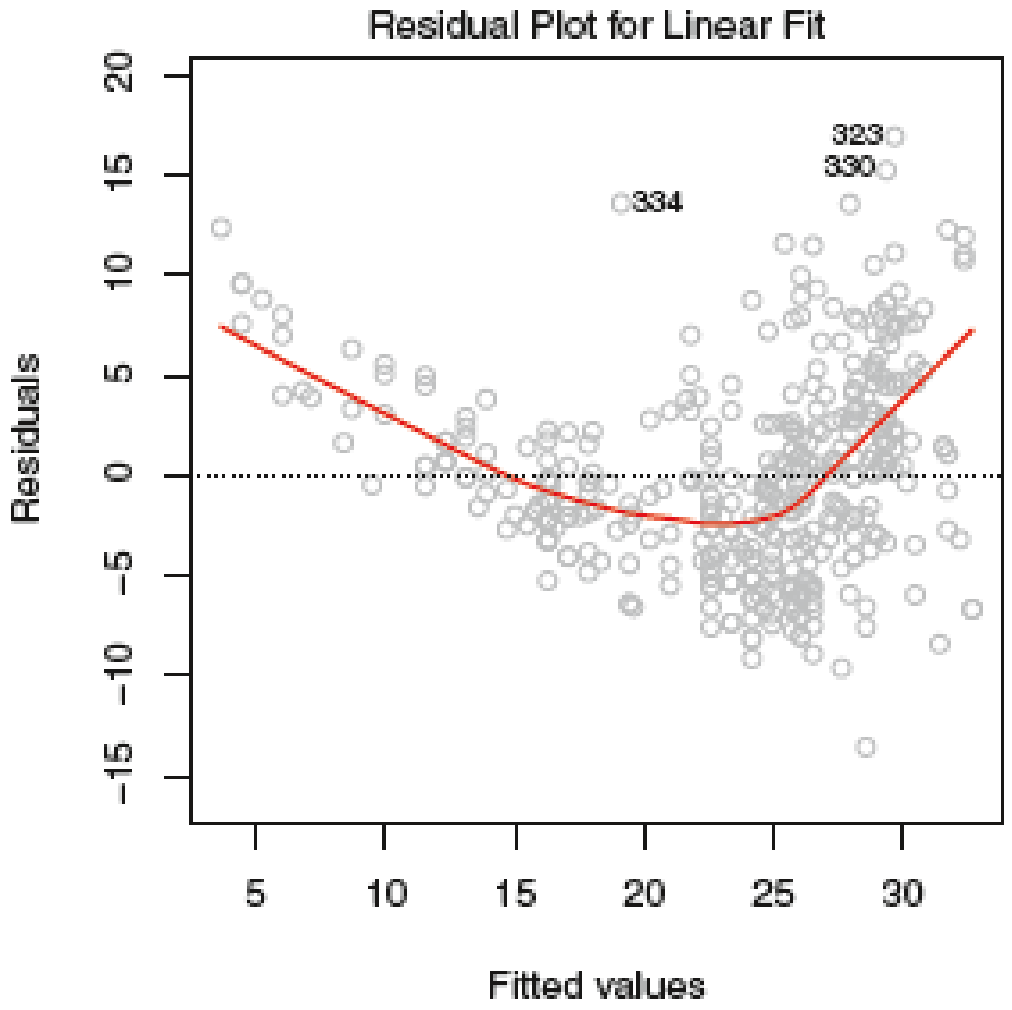
  If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

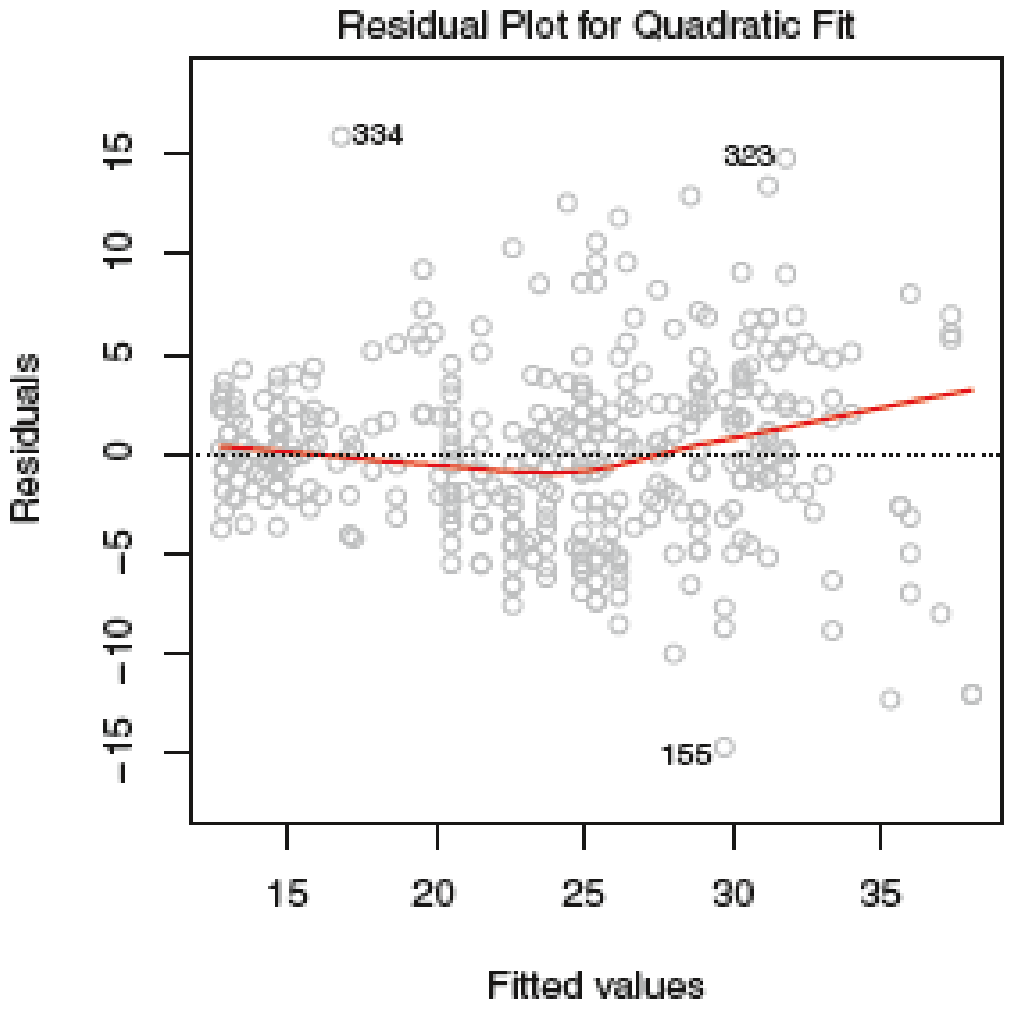
Non-Linear Relationship

- The linear regression model assumes a linear relationship between the response and predictors.
- But in some cases, the true relationship between the response and the predictors may be nonlinear.
- Here we present a very simple way to directly extend the linear model to accommodate non-linear relationships, using *polynomial regression*.


polynomial regression on Auto data
Linear
Degree 2
Degree 5
Miles per gallon
10
20
30
40
50
50
100
150
200
Horsepower

- There is a pronounced relationship between mpg and horsepower, but it seems clear that this relationship is in fact non-linear: the data suggest a curved relationship.
- A simple approach for incorporating non-linear associations in a linear model is to include transformed versions of the predictors in the model.
- $\text{mpg} = \beta_0 + \beta_1 \text{ x horsepower} + \beta_2 \text{ x horsepower}^2 + \in$
- Above equation involves predicting mpg using a non-linear function of horsepower. *But it is still a linear model!* That is, equation is simply a multiple linear regression model with X_1 = horsepower and X_2 = horsepower^2.
- The R^2 of the quadratic fit is 0.688, compared to 0.606 for the linear fit

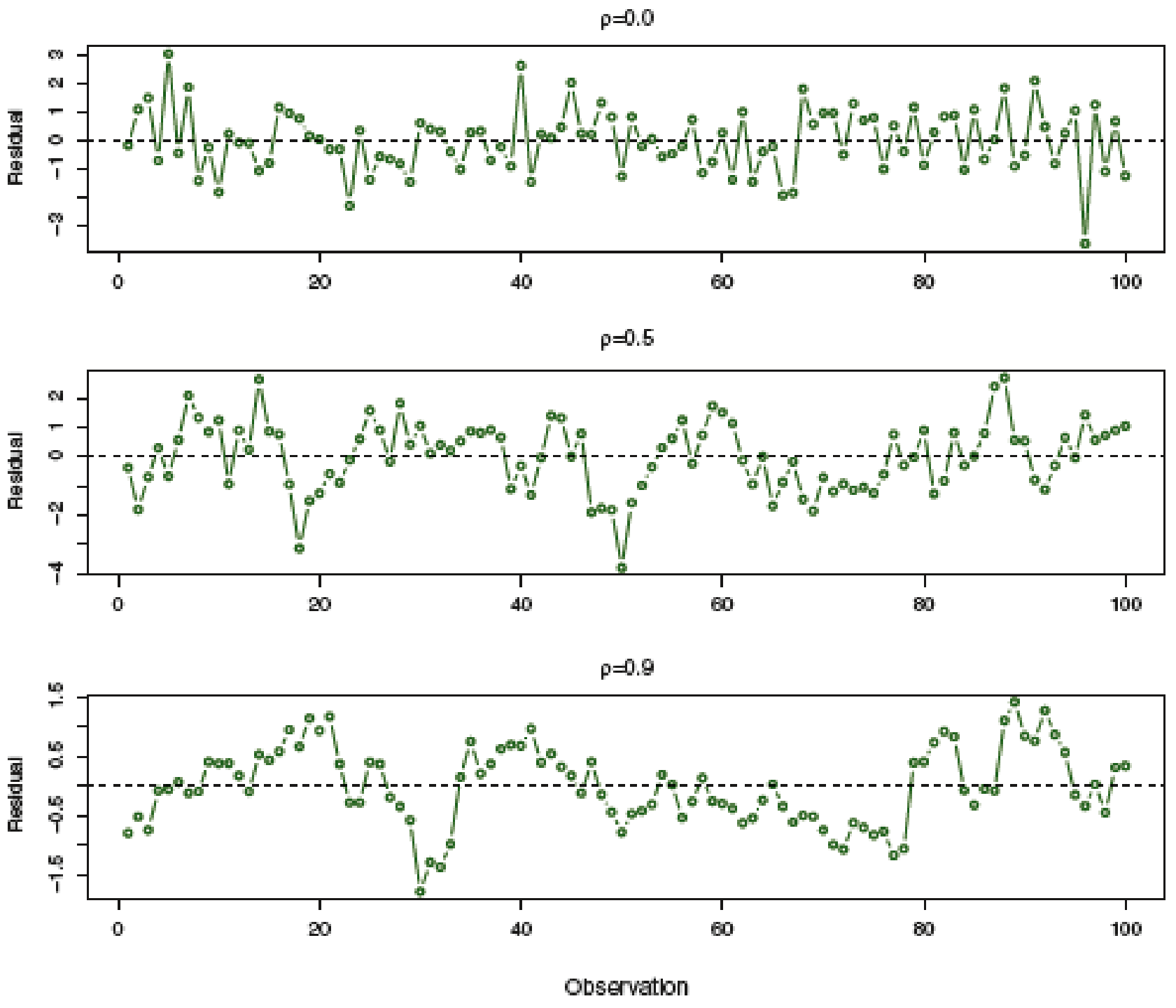




If the residual plot indicates that there are non-linear associations in the data, then a simple approach is to use non-linear transformations of the predictors, such as log$X$, $\sqrt{X}$, and X^2, in the regression model.

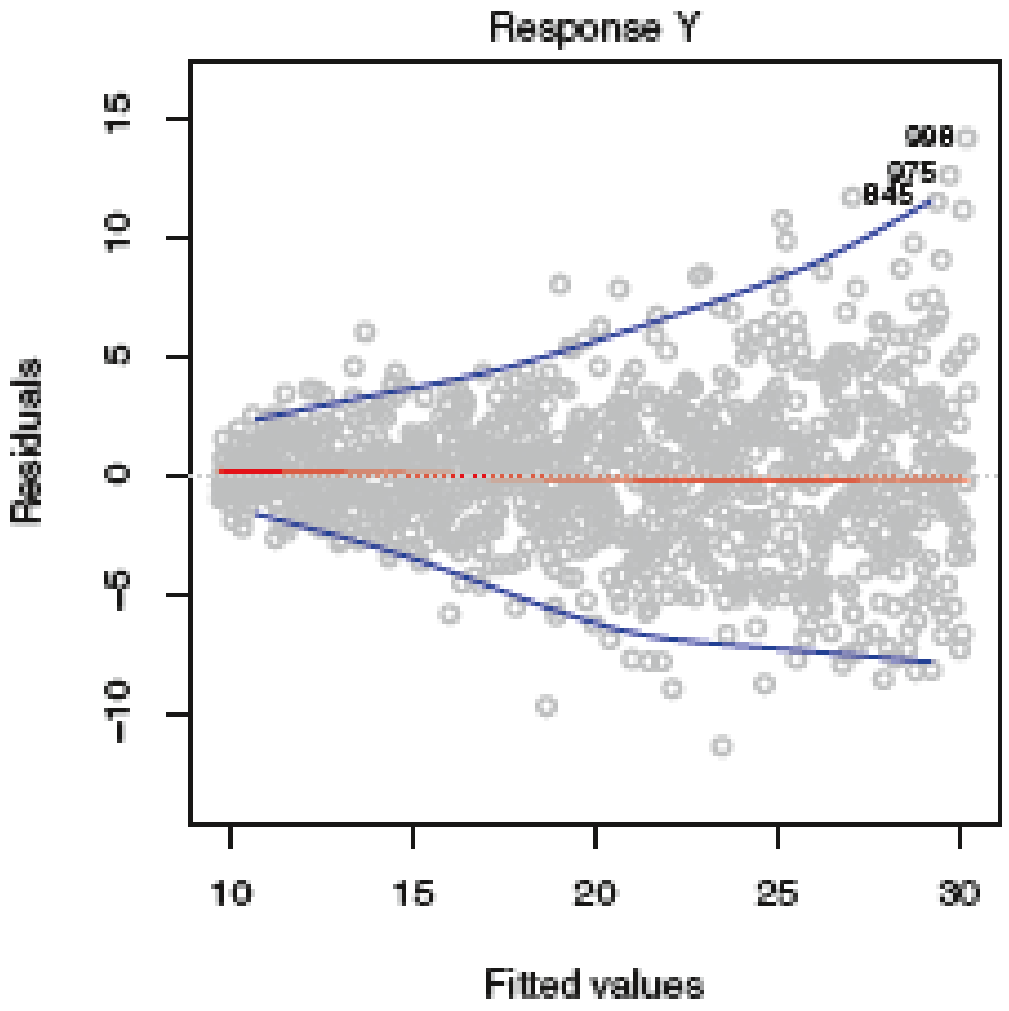
Correlation of Error Terms

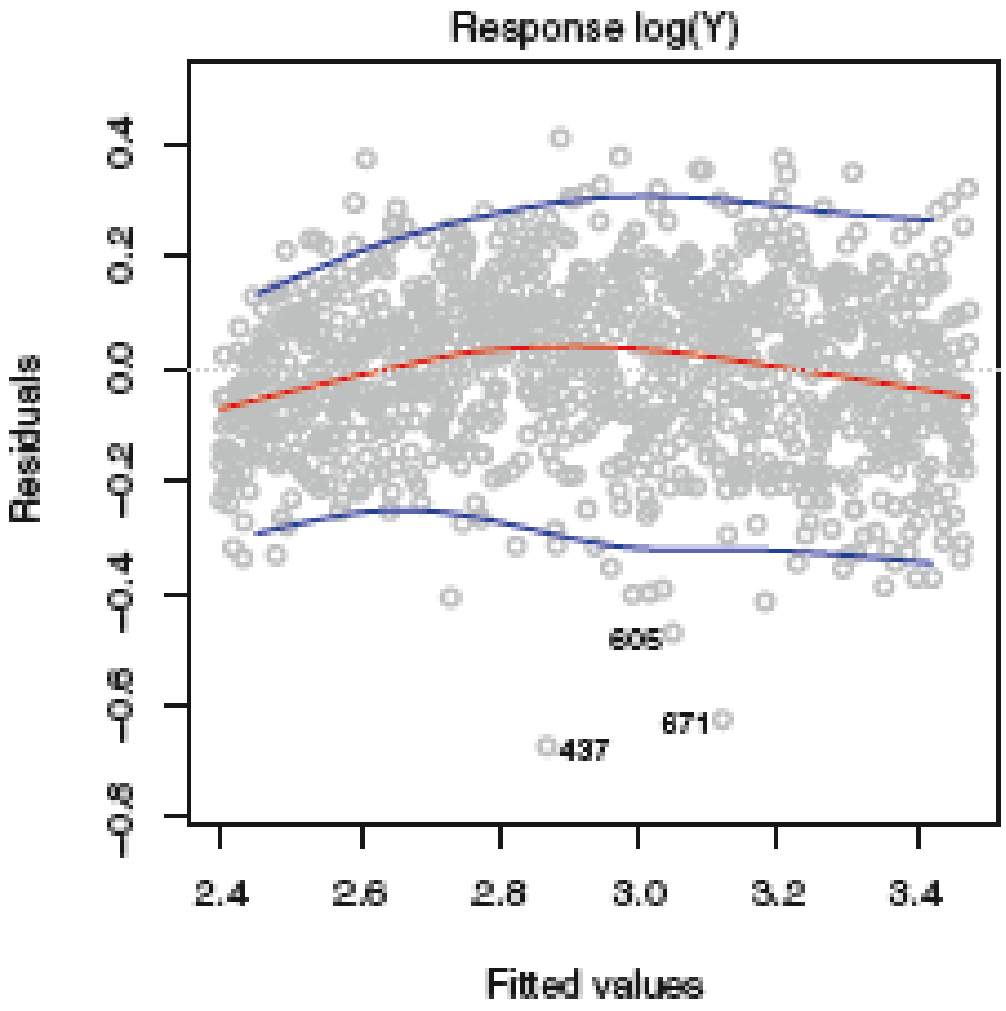
- An important assumption of the linear regression model is that the error terms, $1, 2, \ldots, n$, are uncorrelated.
- For instance, if the errors are uncorrelated, then the fact that $\in_i$ is positive provides little or no information about the sign of $\in_{i+1}$.
- If in fact there is correlation among the error terms, confidence and prediction intervals will be narrower than they should be.
- For example, a 95% confidence interval may in reality have a much lower probability than 0.95 of containing the true value of the parameter.


ρ=0.0
Residual
3
2
1
0
−1
−3
0
20
40
60
80
100
ρ=0.5
Residual
2
1
0
−2
−4
0
20
40
60
80
100
ρ=0.9
Residual
1.5
0.5
−0.5
−1.5
0
20
40
60
80
100
Observation

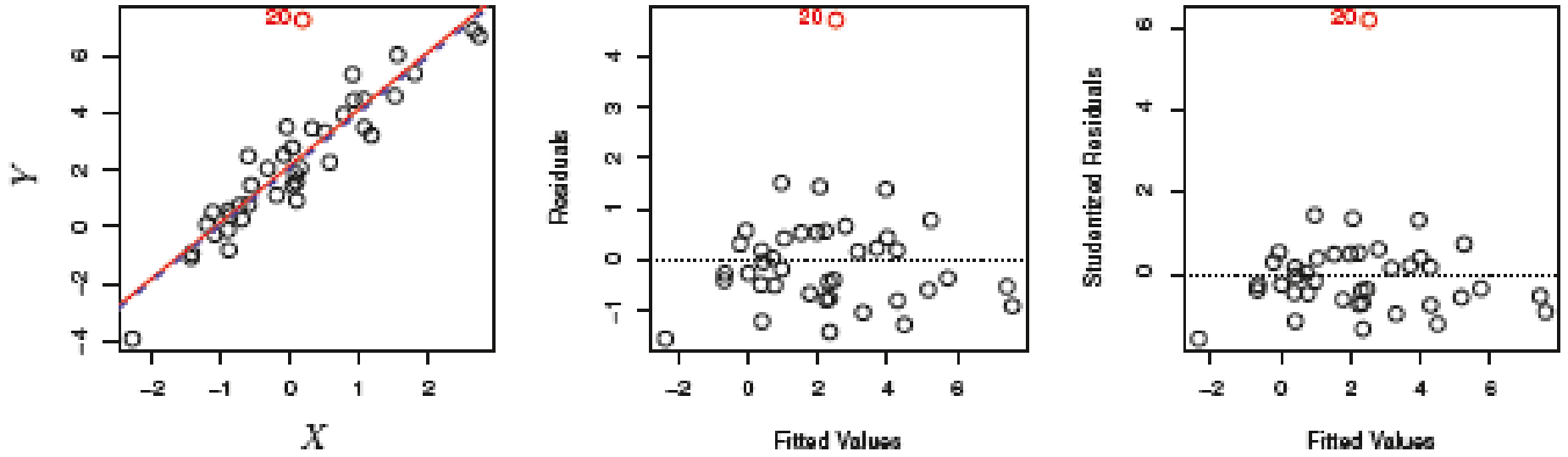
Non-constant Variance of Error Terms

- Another important assumption of the linear regression model is that the error terms have a constant variance, $\text{Var}(\in_i) = \sigma^2$.
- Unfortunately, it is often the case that the variances of the error terms are non-constant.
- One can identify non-constant variances in the errors, or *heteroscedasticity*, from the presence of a *funnel shape* in residual plot.
- When faced with this problem, one possible solution is to transform the response Y using a concave function such as $\log Y$ or $\sqrt{Y}$.


Response Y
Residuals
Fitted values
15
10
5
0
−5
−10
10
15
20
25
30
908
875
845


Response log(Y)
Residuals
Fitted values
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
2.4
2.6
2.8
3.0
3.2
3.4
605
871
437

Outliers



- An *outlier* is a point for which y_i is far from the value predicted by the model.
- The red solid line is the least squares regression fit, while the blue dashed line is the least squares fit after removal of the outlier.
- However, even if an outlier does not have much effect on the least squares fit, it can cause other problems.
- Like inclusion of the outlier causes the R^2 to decline from 0.892 to 0.805.
- we can plot the studentized residuals, computed by dividing each residual e_i by its estimated standard error. Observations whose studentized residuals are greater than 3 in absolute value are possible outliers.

High Leverage Points

- observations with *high leverage* have an unusual value for x_i.