POLITECNICO
MILANO 1863

Machine Learning

- Regression -

Matteo Matteucci, PhD (matteo.matteucci@polimi.it)
Artificial Intelligence and Robotics Laboratory
Politecnico di Milano

AIRLAB
ARTIFICIAL INTELLIGENCE AND ROBOTICS LAB

Example: Increasing Sales by Advertising

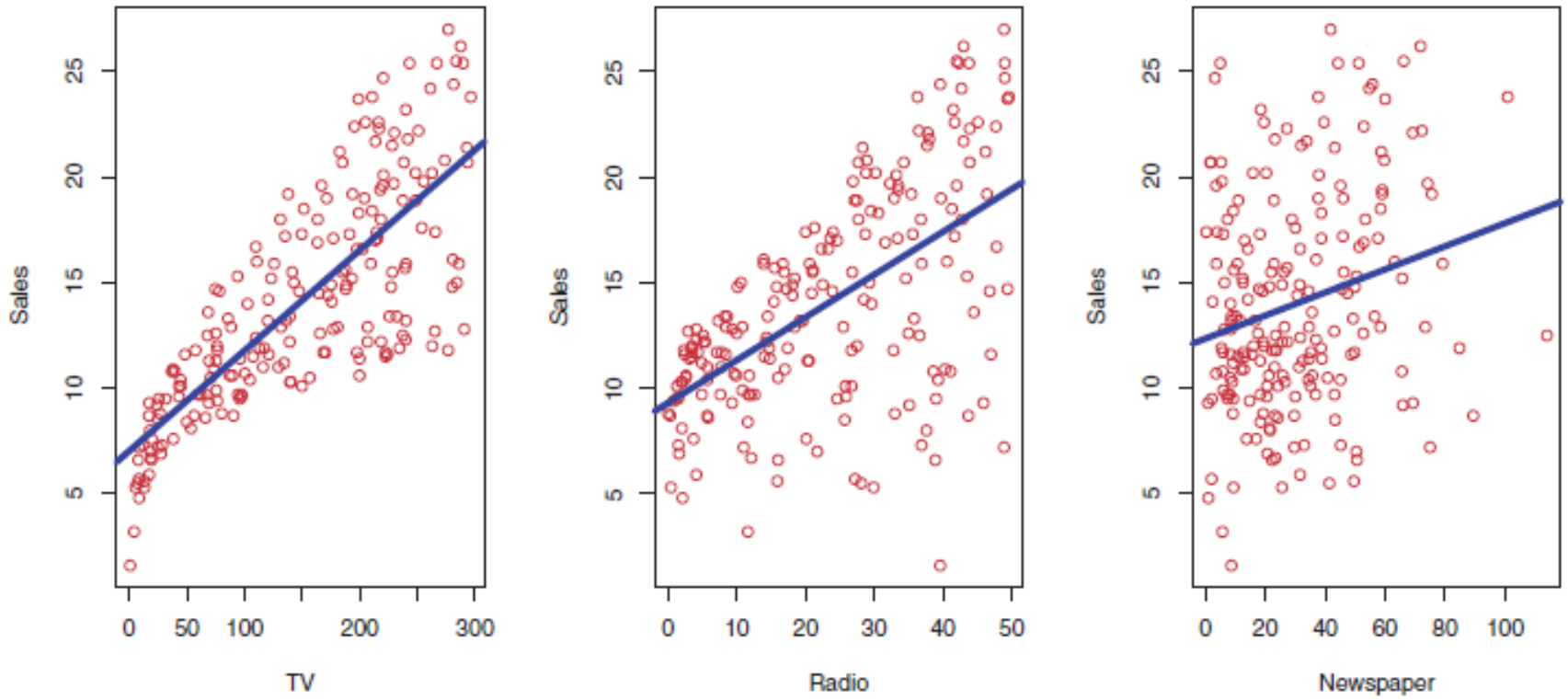
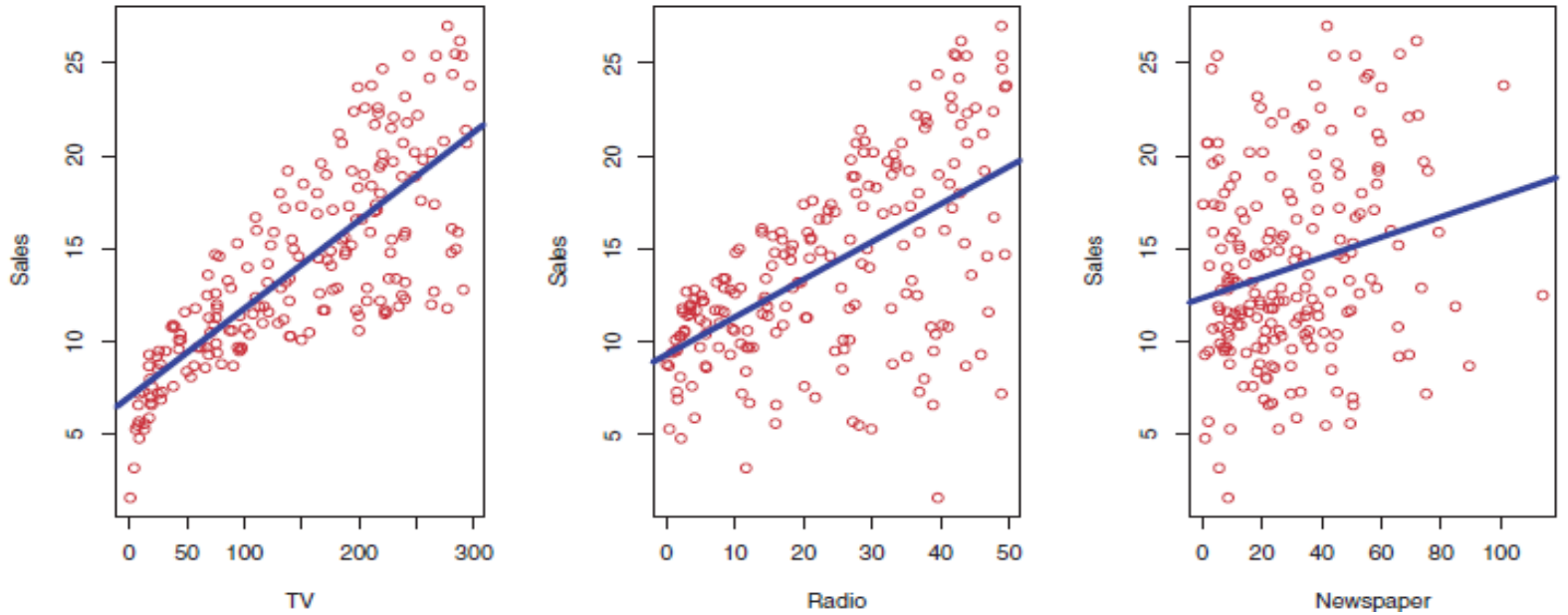


FIGURE 2.1. *The* `Advertising` *data set. The plot displays* `sales`*, in thousands of units, as a function of* `TV`*,* `radio`*, and* `newspaper` *budgets, in thousands of dollars, for* 200 *different markets. In each plot we show the simple least squares fit of* `sales` *to that variable, as described in Chapter 3. In other words, each blue line represents a simple model that can be used to predict* `sales` *using* `TV`*,* `radio`*, and* `newspaper`*, respectively.*

What can we ask to the data?



- *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 estimate the effect of each medium on sales?*
- *Is the relationship linear?*
- *Is there synergy among the advertising media?*

Simple linear regression

Le assume that a linear relationship exists between Y and X

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}$$

We say sales regress on TV through some parameters

- Model coefficients β_0 and β_1
- After training, a new data point can be predicted as

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

Given a datasets the coefficient above can be estimated by using least squares to minimize the Residual Sum of Squares

$$e_i = y_i - \hat{y}_i$$
$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

Example: TV Advertising vs Sales

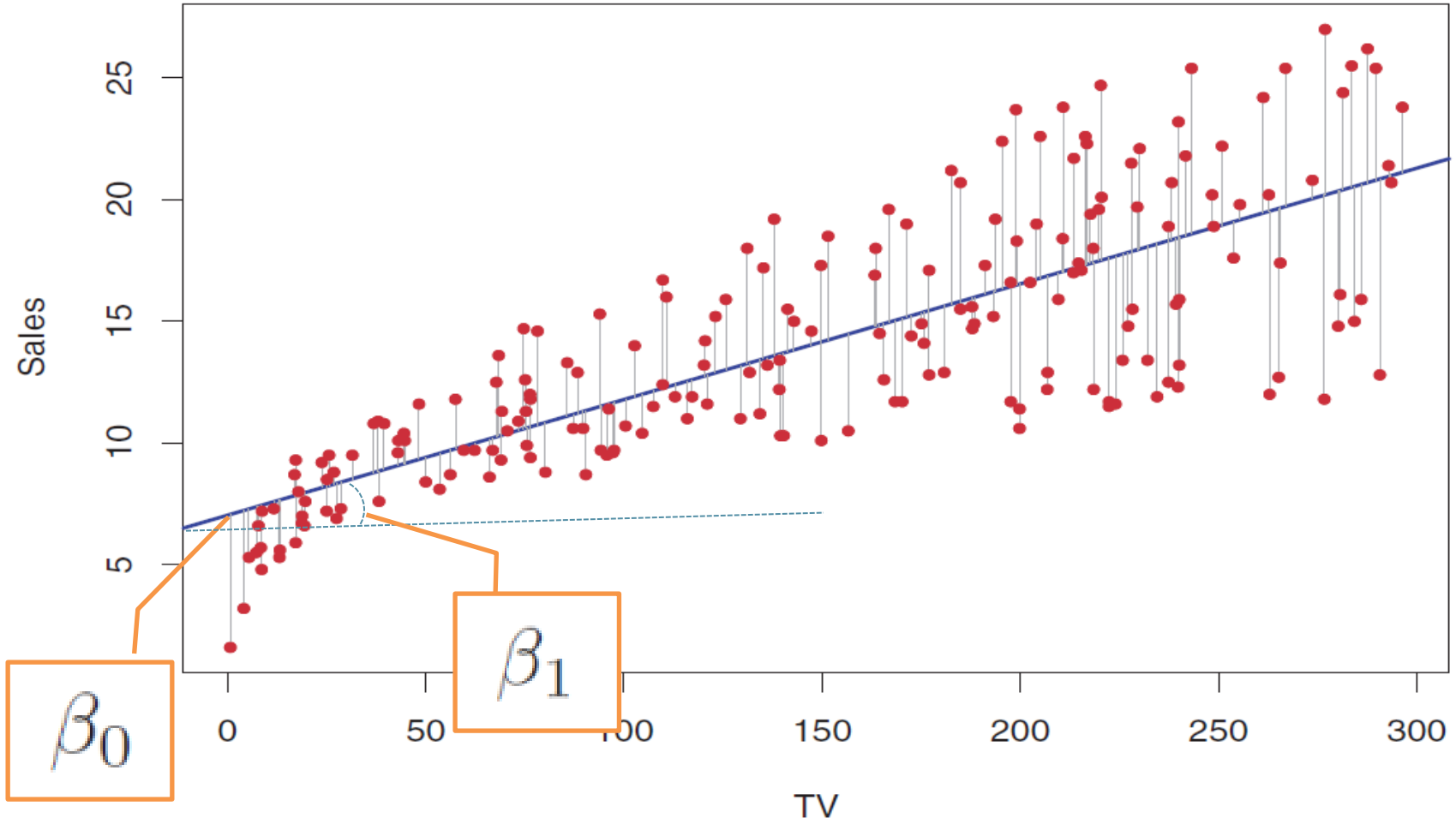


FIGURE 3.1. *For the* `Advertising` *data, the least squares fit for the regression of* `sales` *onto* `TV` *is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of*

Least squares fitting

Least square fitting minimizes RSS (Residual Sum of Squares)

$$e_i = y_i - \hat{y}_i$$

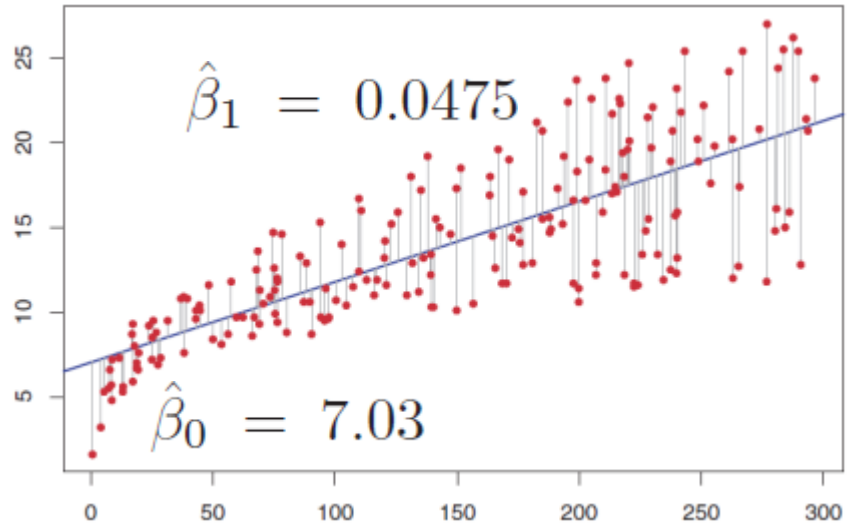
$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

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

Obtaining the following estimates

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$



Least square solution …

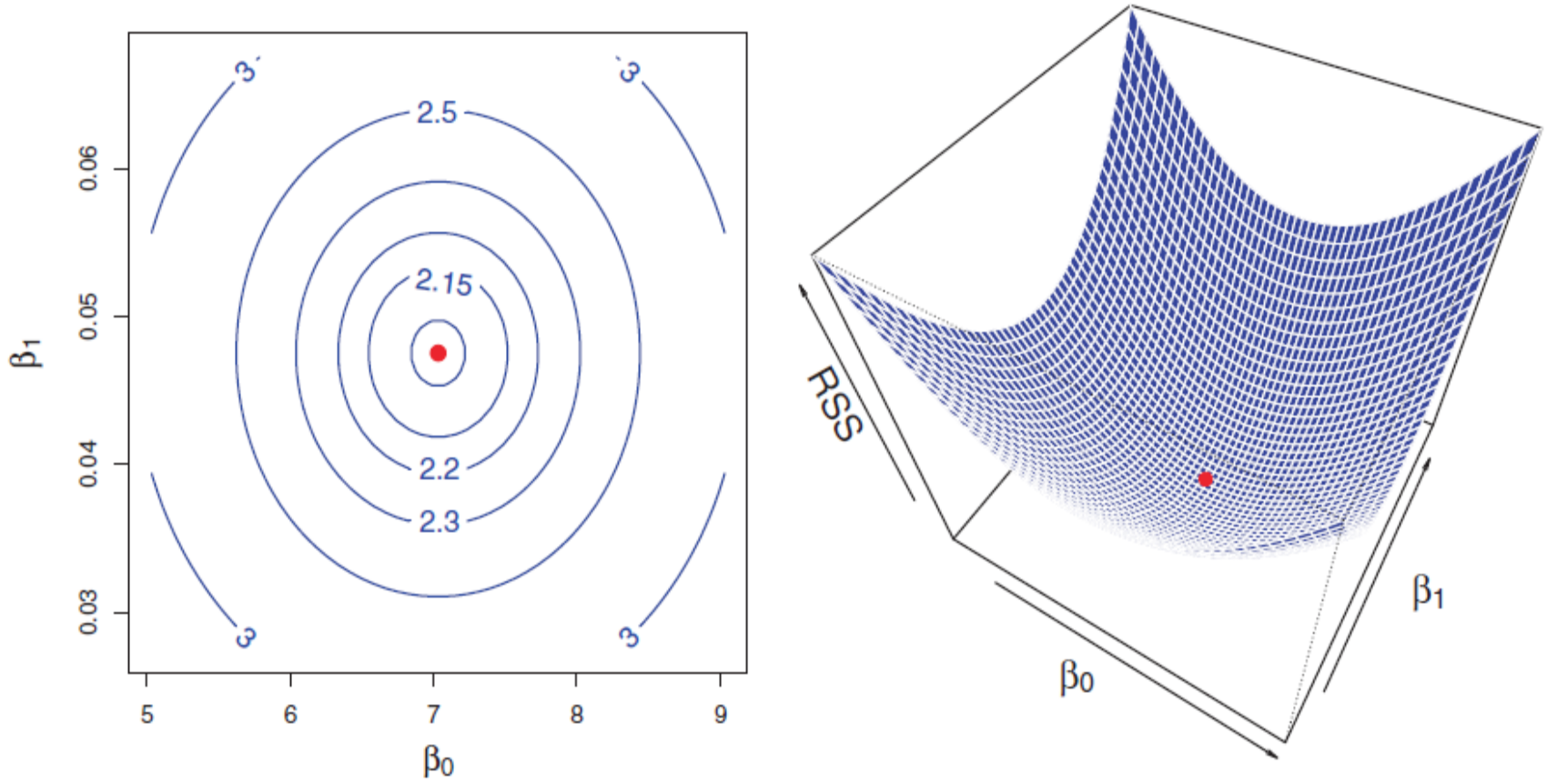


FIGURE 3.2. *Contour and three-dimensional plots of the RSS on the* `Advertising` *data, using* `sales` *as the response and* `TV` *as the predictor. The red dots correspond to the least squares estimates* $\hat{\beta}_0$ *and* $\hat{\beta}_1$*, given by (3.4).*

Least squares fitting

Least square fitting minimizes RSS (Residual Sum of Squares)

$$e_i = y_i - \hat{y}_i$$

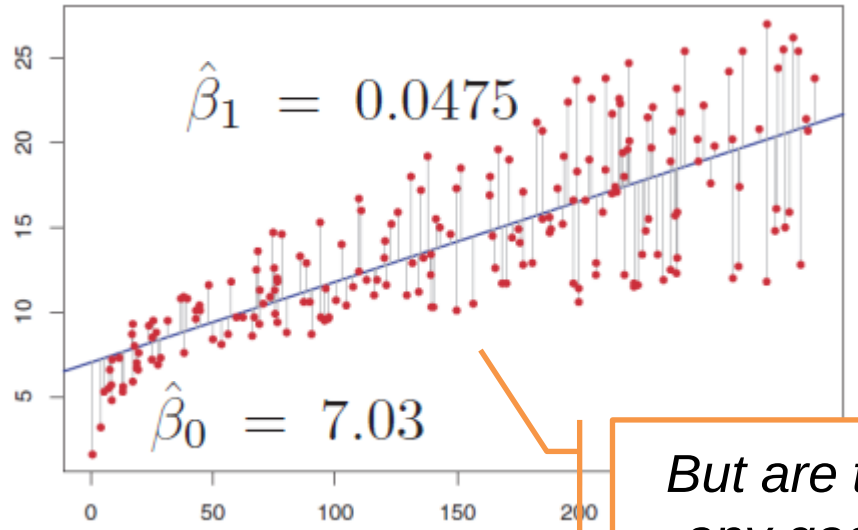
$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

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

Obtaining the following estimates

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$



But are they any good?

Population regression line

Recall from Statistical Learning theory the underlying hypothesis

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

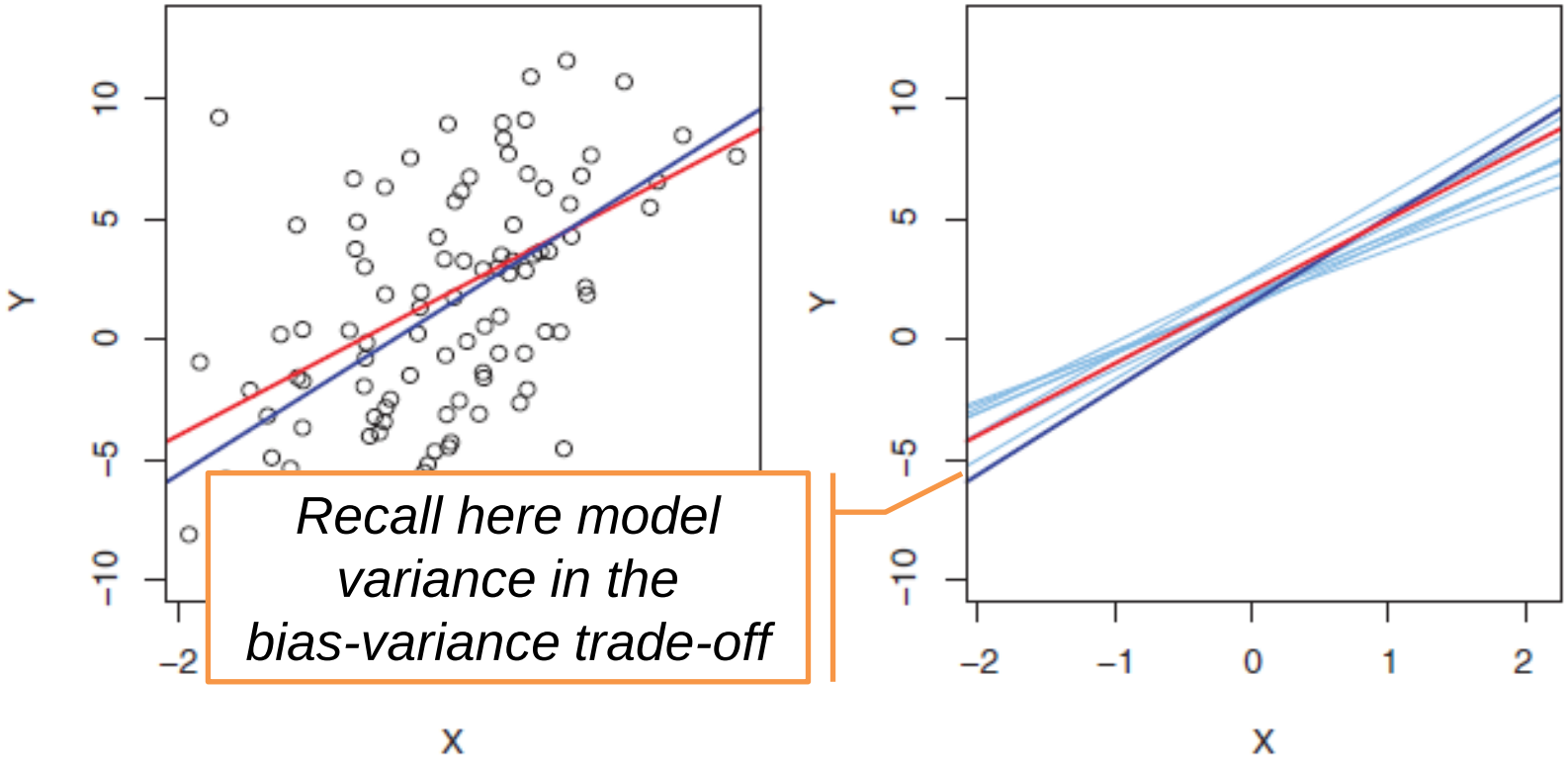
- β_0 the Y value when X = 0
- β_1 the average increase in Y due to unitary increase in X
- the error term captures all the rest …

This model is known as "population regression line"

- the best linear approximation of the true model
- it might differ from the least sq[illegible]

The population regression line stays to the mean of a distribution as the least squares regression line stays to the sample mean …

Example: population regression line



Recall here model variance in the bias-variance trade-off

FIGURE 3.3. *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. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.*

Standard error & linear regression

The (squared) standard error for the *mean estimator* represents the average distance of the sample mean from the real mean

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

We have formulae for standard errors of linear regression coefficients

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

The higher the spread of x the better the estimate

These formulae assume

- uncorrelated errors …
- … having the same (unknown) variance $\sigma^2 = \mathrm{Var}(\epsilon)$

Parameters confidence intervals

In general, errors variance is not known, but it can be estimated from residuals (if the model fits properly)

$$\mathrm{RSE} = \sqrt{\mathrm{RSS}/(n-2)}$$

From standard errors we can compute confidence intervals for the linear regression parameters, e.g., the 95% confidence intervals for the parameters are

$$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0) \qquad \hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

the true slope is, with 95% probability, in the range

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

This should be the 97.5 quantile of a t-distribution

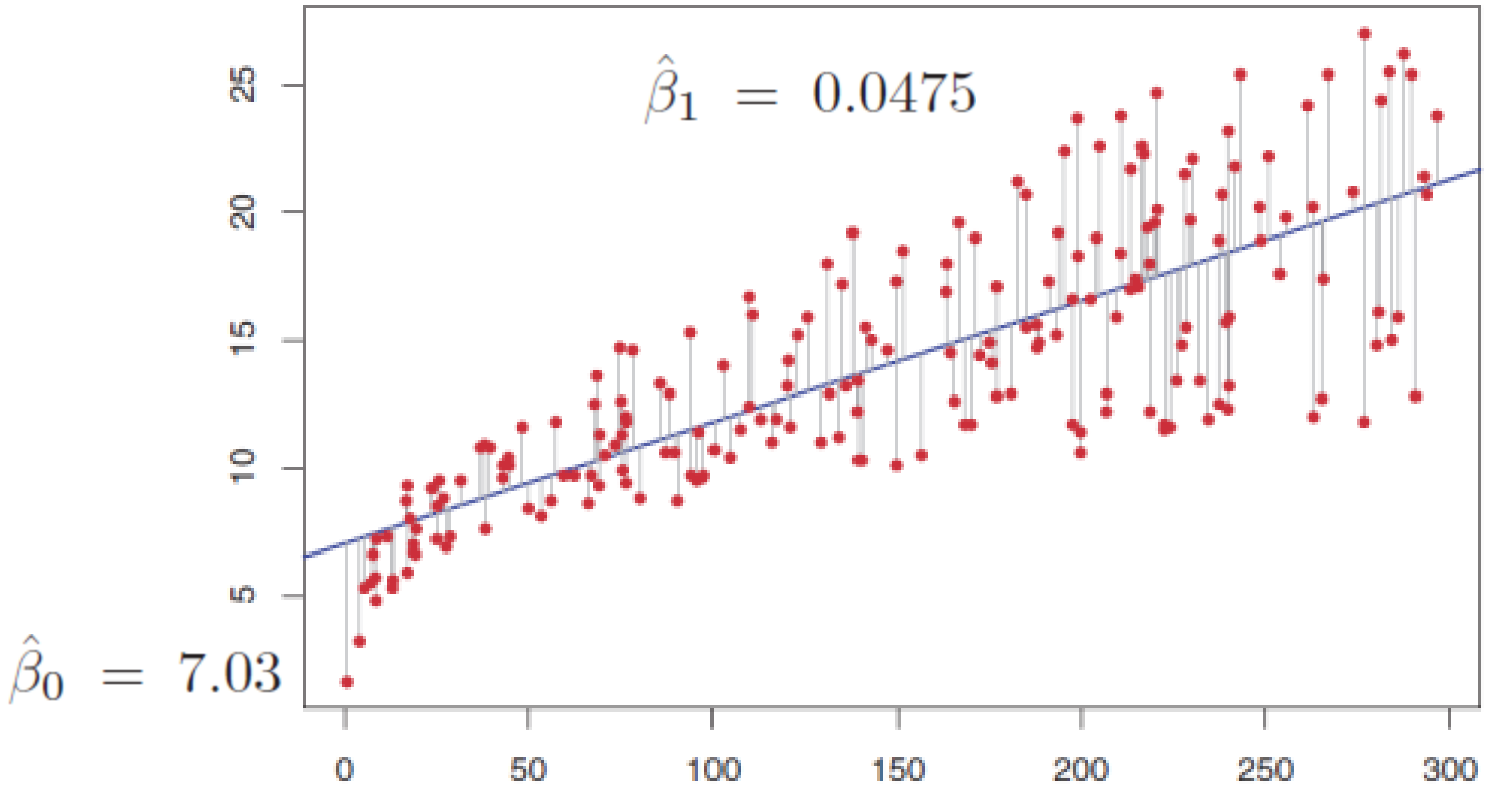
Example: TV Advertising data

If we consider the 95% confidence intervals

- for the intercept we have $[6.130, 7.935]$
- for the slope we have $[0.042, 0.053]$

Sales without any advertising

Average impact of TV advertising



Parameters hypothesis testing

Standard errors can be used for hypothesis testing such as:

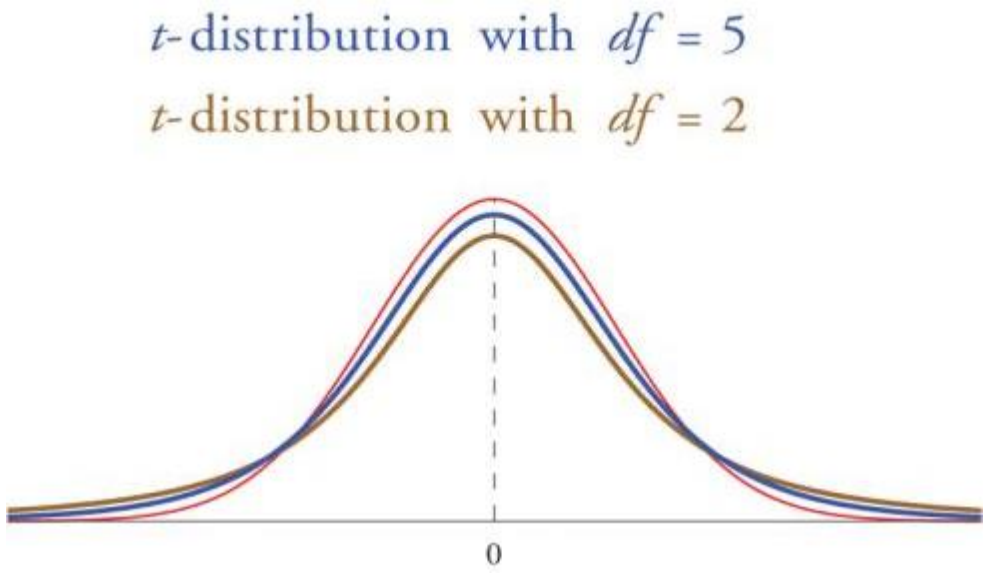
- H_0: there is no relationship between Y and X
- H_a: there is some relationship between Y and X

This translates on parameters hypothesis testing for

$$H_0 : \beta_1 = 0 \quad \text{against} \quad H_a : \beta_1 \neq 0$$

We do not know true parameters so we can use estimates and perform a statistical test using

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



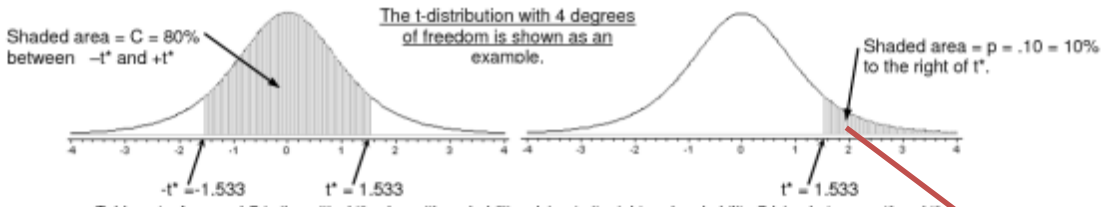


Table entry for p and C is the critical t* value with probability p lying to its right and probability C lying between –t* and t*.

Upper Tail Probability p →	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
Degrees of freedom ↓												
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.894	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.689
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.660
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.974	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
Confidence level C = 1- 2p →	**50%**	**60%**	**70%**	**80%**	**90%**	**95%**	**96%**	**98%**	**99%**	**99.5%**	**99.8%**	**99.9%**

This value is know as p-value

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

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

~1.984

N = 200 data points

p-value = 2.5%

t ~ 1.984

Here you find the confidence interval

Example: TV advertising hypothesis test

We reject the null hypothesis H_0 if the p-value is small

- p-value is the probability of making a wrong choice
- Usually small is as low as 5% or 1%, these percentages, with N>30 correspond to t~2 and t~2.75 respectively
- In other fields, p-values might be significantly different, e.g., in bioinformatics p-values of 10^{-6} are quite common to avoid false discoveries …

	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

TABLE 3.1. *For the* Advertising *data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of* \$1,000 *in the TV advertising budget is associated with an increase in sales by around 50 units (Recall that the* sales *variable is in thousands of units, and the* TV *variable is in thousands of dollars).*

Accuracy of a model: RSE

The classical measure of fit is mean squared error, in linear regression we use the Residual Standard Error

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

How far the model is from least square line on average

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- It estimates the standard deviation of the errors, i.e., the irreducible error.

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

Compared to the average sales
$3{,}260/14{,}000 = 23\,\%$

TABLE 3.2. *For the* `Advertising` *data, more information about the least squares model for the regression of number of units sold on TV advertising budget.*

Accuracy of a model: R^2

We might be interested in computing how much of the data variance is explained by the model ("relative accuracy")

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

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

$$\text{TSS} = \sum(y_i - \bar{y})^2$$

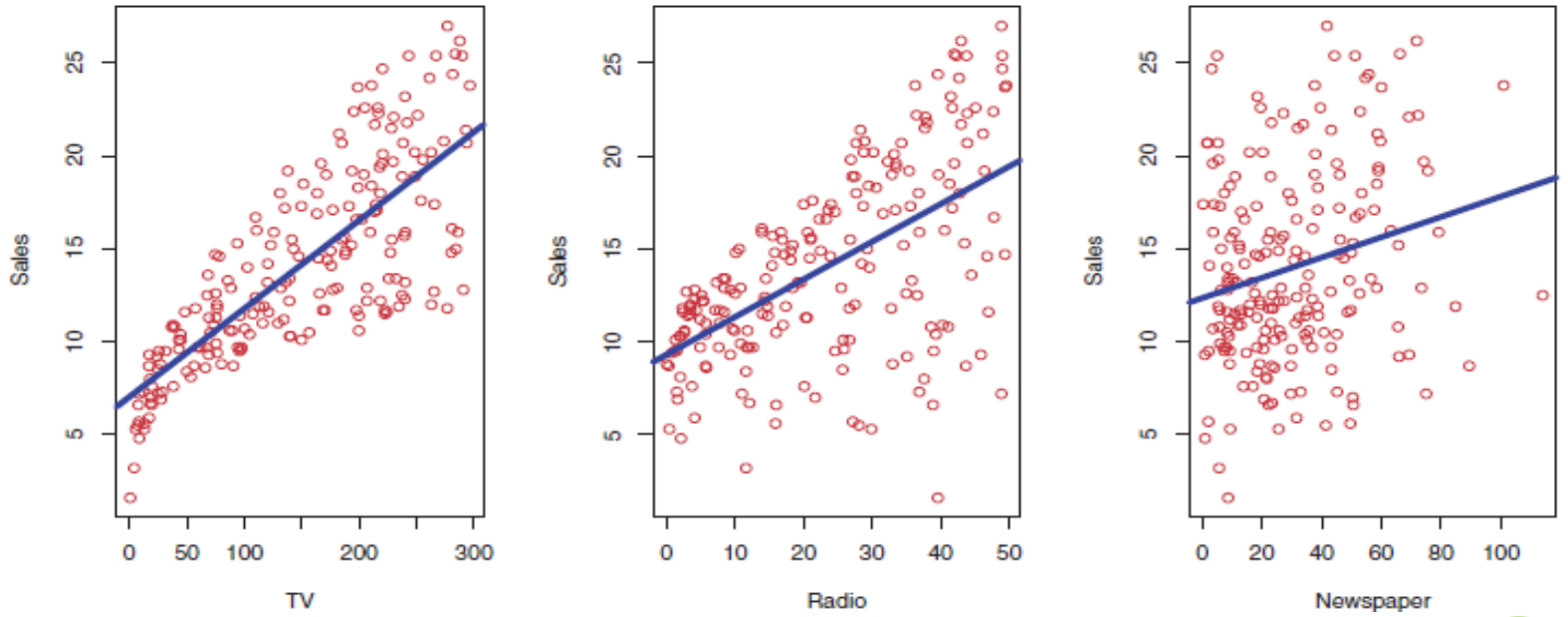
An R^2 close to 1 means the data are almost perfectly explained by our simple linear model, in our case it is just 0.612 …

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

Is this due to the error noise or to the fact that data is not linear?

TABLE 3.2. *For the* `Advertising` *data, more information about the least squares model for the regression of number of units sold on TV advertising budget.*

What can we ask to the data?



- *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 estimate the effect of each medium on sales?* ✗
- *Is the relationship linear?* ✗
- *Is there synergy among the advertising media?* ✗

Multiple linear regression … the easy (wrong) way

Simple regression of `sales` on `radio`

	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

Simple regression of `sales` on `newspaper`

	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

TABLE 3.3. *More simple linear regression models for the* `Advertising` *data. Coefficients of the simple linear regression model for number of units sold on* Top: *radio advertising budget and* Bottom: *newspaper advertising budget. A* \$1,000 *increase in spending on radio advertising is associated with an average increase in sales by around 203 units, while the same increase in spending on newspaper advertising is associated with an average increase in sales by around 55 units (Note that the* `sales` *variable is in thousands of units, and the* `radio` *and* `newspaper` *variables are in thousands of dollars).*

Multiple linear regression … the right way

Treating variables as they were independent

- Does not tell how a sales increase is obtained by changing all input variables
- Coefficients of each input did not take into account the others in the estimate
- If input are highly correlated, using independent estimates can be misleading

Extend linear regression to consider multiple predictors

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

More formally we have the multivariate regression

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

Example: a two dimensional dataset

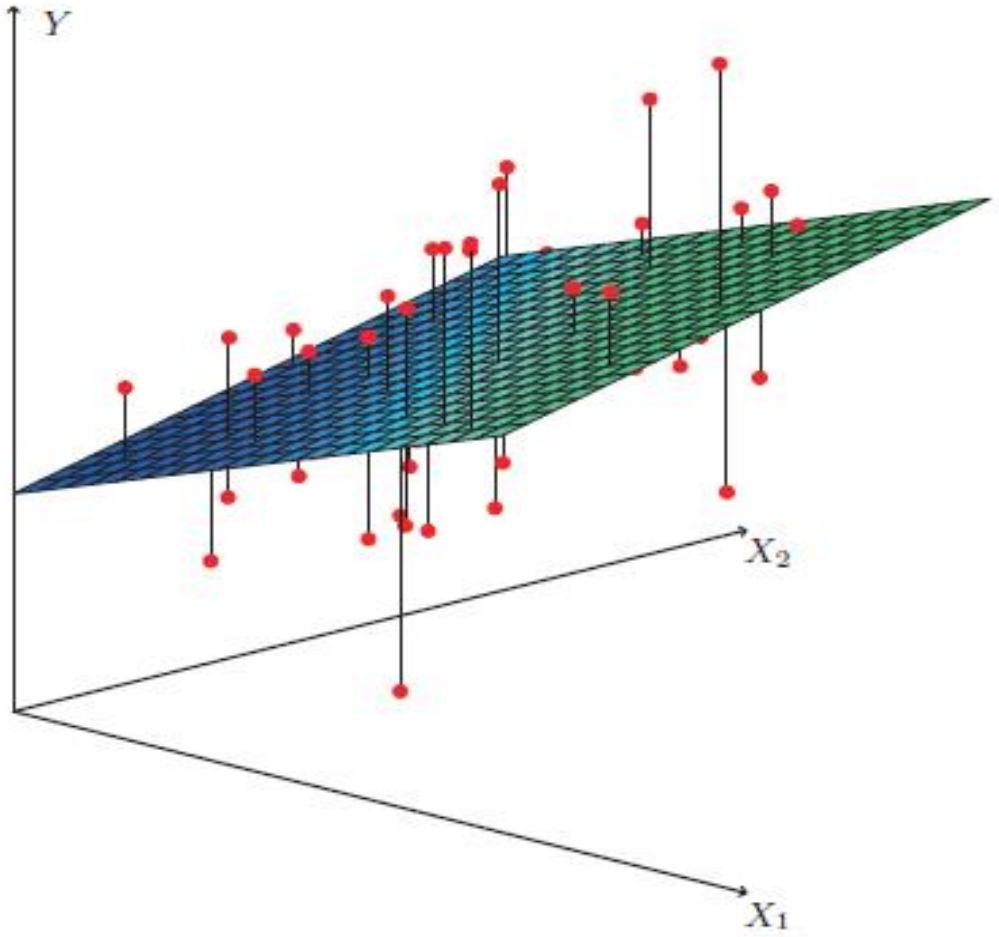


FIGURE 3.4. *In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.*

Linear regression

Linear regression parametric model, i.e., the population line

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

- Each parameters describes the average influence of the associated input keeping all the others fixed

The regression coefficient can be estimated by least squares fit

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

To obtain the least squares predictor

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Example: Advertising dataset

	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	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

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

Example: Advertising dataset

	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	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

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

Example: Advertising dataset

	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

	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

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

Correlation between attributes

	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

TABLE 3.5. *Correlation matrix for* TV, radio, newspaper, *and* sales *for the* Advertising *data.*

Let consider correlations between input and output variables

- If we increase *radio* then *sales* increase
- *Radio* and *newspaper* are highly correla
- If we increase *radio* then *newspaper* increa

Sharks attacks are correlated to ice cream sales at the beach …

The increase on sales is correlate to the increase of newspaper is due to radio, not to the fact that newspaper increases sales

Computing linear regression coefficients (1)

Computing the least regression fit can be done easily using linear algebra

Recall here

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

By taking into account that

- $\mathbf{X}$ is an N x (p+1) data matrix
- $\mathbf{y}$ is N x 1 vector of desired output
- $\boldsymbol{\beta}$ is a (p+1) x 1 vector of model coefficients

We can rewrite the Residuals Sums of Squares as

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

Computing linear regression coefficients (2)

We want to minimize $\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$

Let's compute the RSS derivatives with respect to β

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

Assuming $\mathbf{X}$ has full rank and $\mathbf{X}^T\mathbf{X} > 0$ we have just to compare the first derivative to zero

$$\mathrm{Var}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

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

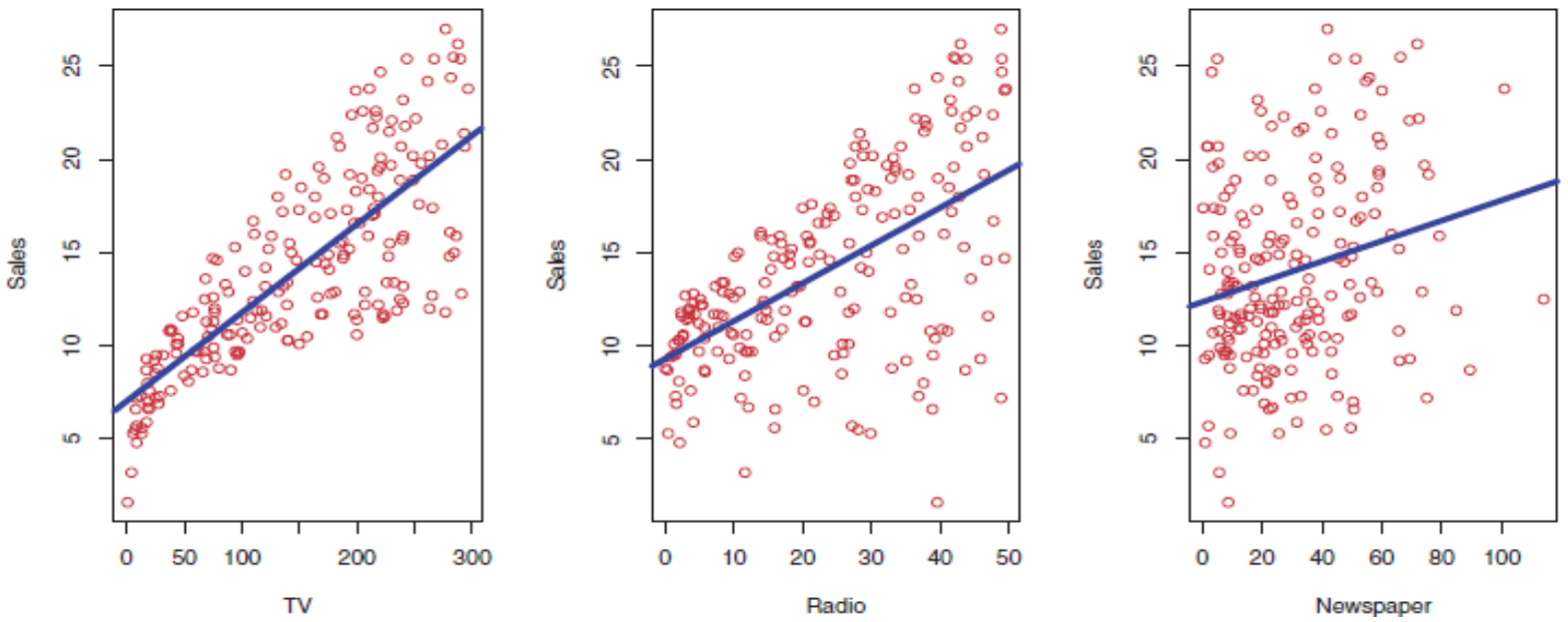
$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Pseudo Inverse

In matrix algebra terms the prediction becomes

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

Let's rephrase the questions in a multivariate setting



- *Is at least one of the predictors X_1, ..., X_p useful in predicting the response?*
- *Do all the predictors help to explain Y, or it is only a subset of the predictors?*
- *How well does the model fit the data?*
- *Given a set of predictor values, what response value should we predict, and accurate is our prediction?*

Hypothesis testing on multiple parameters

Is there any relationship between response and predictors?

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad \text{against} \quad H_a : \text{ at least one } \beta_j \text{ is non-zero}$$

This test is performed using the F-statistics

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

If the linear model assumptions are valid

$$E\{\text{RSS}/(n-p-1)\} = \sigma^2$$

- when H_0 is true $E\{(\text{TSS} - \text{RSS})/p\} = \sigma^2$ then $F \sim 1$
- when H_a is true $E\{(\text{TSS} - \text{RSS})/p\} > \sigma^2$ then $F > 1$

Example: Advertising dataset

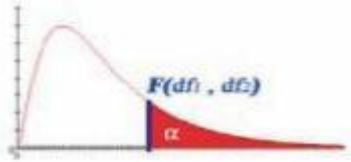
Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

F>1

TABLE 3.6. *More information about the least squares model for the regression of number of units sold on TV, newspaper, and radio advertising budgets in the* `Advertising` *data. Other information about this model was displayed in Table 3.4.*

F is well above 1, a relationship exists!

How much should be F to tell this relationship exists?



F Table for α = .05

df_2=1	df_1=1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28				8.89	8.85	8.81	8.79
4	7.71	6.94	6.59				6.09	6.04	6.00	5.96
5	6.61	5.79	5.41				4.88	4.82	4.77	4.74
6	5.99	5.14	4.76				4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96				33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98			20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89			11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59		2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45

p

n

Testing for subsets of variables

We can test also a subset of the variables

$$H_0: \quad \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0$$

The novel F-statistics for the model fitted on q variables is

RSS of the model without the q variables

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/q}{\mathrm{RSS}/(n-p-1)}$$

RSS of the model with all the variables

If we leave out one variable at the time ($q=1$) we obtain an equivalent formulation of the t-statistics for single parameters

- F-statistics is more accurate than t-statistics computed for each parameter since it corrects for other parameters
- It tells you the partial effect of adding that specific variable to the model

Spurious correlations

If the number of factors p is big, p-values might be tricky

- With p=100 and H_0 true, ~5% of the p=values (by chance) will be lower than 0.05 and we might see 5 predictors associated (by chance) to the response
- F statistic is not affected by the number of factors p in the model

	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

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

Example: Credit Predictors

In the Credit dataset

- 7 quantitative
- 4 qualitative

Qualitative ones

- Gender
- Student (Binary status)
- Status (Binary marital)
- Ethnicity (Caucasian, African American, Asian)

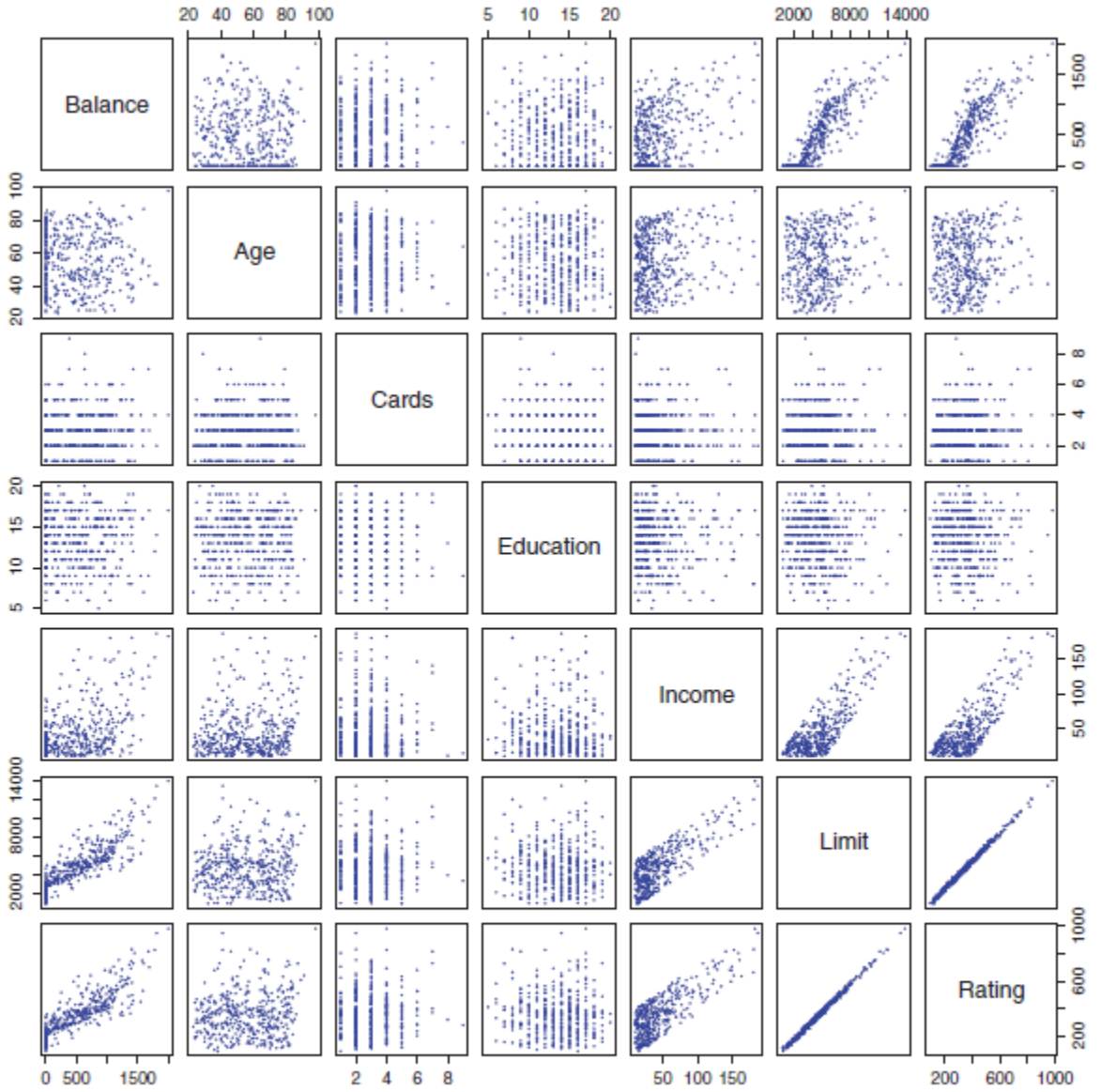


FIGURE 3.6. *The* `Credit` *data set contains information about* `balance`, `age`, `cards`, `education`, `income`, `limit`, *and* `rating` *for a number of potential customers.*

Qualitative Predictors (Two Levels)

Two Levels qualitative predictors can be coded using *Dummy Variables*

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

This results in a "double" model for regression

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

Average difference of balance between males and females

Average balance among males

	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

TABLE 3.7. *Least squares coefficient estimates associated with the regression of* balance *onto* gender *in the* Credit *data set. The linear model is given in (3.27). That is, gender is encoded as a dummy variable, as in (3.26).*

Other Coding for Two Levels

Other possible coding can be devised for Dummy Variables

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

In this case the model becomes

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

Average balance

Amount of balance females are above the average and males are below …

No significant impact on the regression output, but on the interpretation of the coefficients …

Qualitative Predictors (More Levels)

More than 2 levels are handled by using L-1 dummy labels

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

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

This again results in a "multiple output"

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

Average difference between African Americans and Asian

Average balance for African American

Average difference between African Americans and Caucasians

Qualitative Predictors (More Levels)

	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

TABLE 3.8. *Least squares coefficient estimates associated with the regression of* balance *onto* ethnicity *in the* Credit *data set. The linear model is given in (3.30). That is, ethnicity is encoded via two dummy variables (3.28) and (3.29).*

This again "output" model

The non coded level is defined *baseline*

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

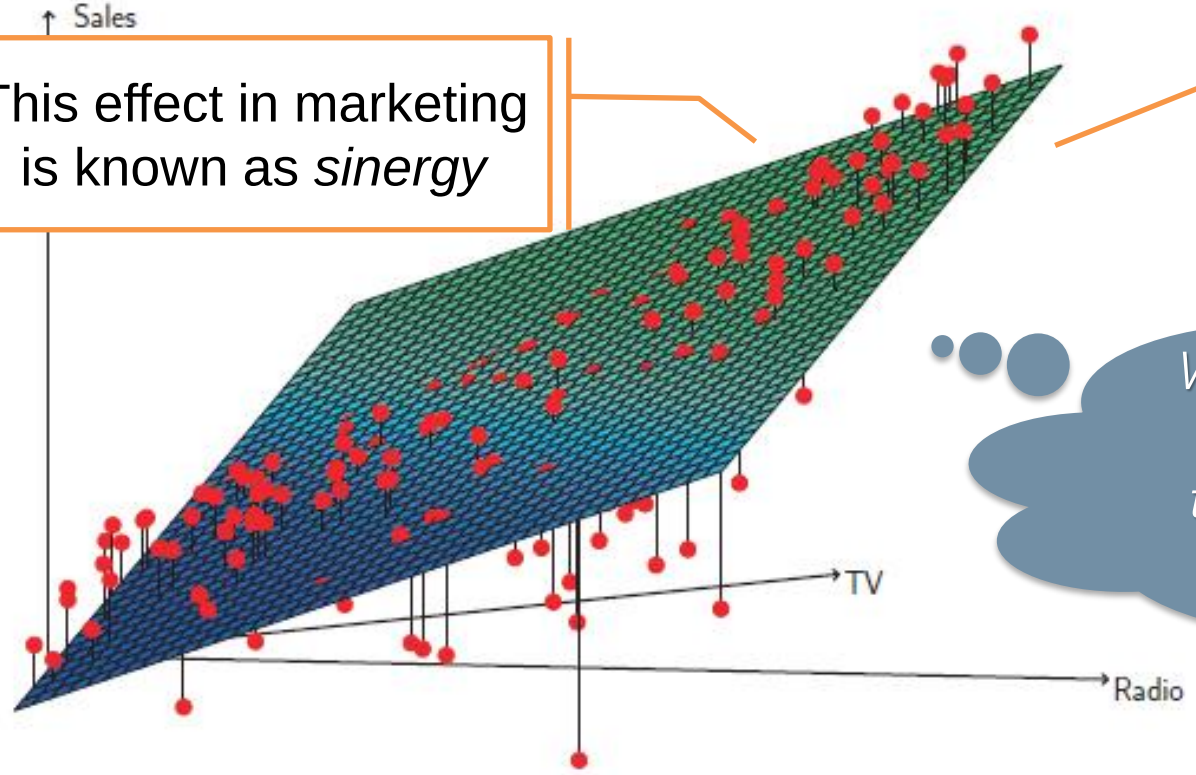
$$H_0 : \beta_1 = \beta_2 = 0$$

F-statistics
p-value 0.96

Variables Interactions (or Sinergies)

So far the linear regression model has assumed

- Linear relationship between predictor and response
- Additive relationship between predictor and response



This effect in marketing is known as *sinergy*

The "slope" of TV increases because of an increase in radio budget

What if allocating half the budget to TV and Radio would increase the sales more than putting it all on one of the two?

Variable Interactions (continued)

Let consider the classical Linear Regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- An increase in X_1 of 1 unit increases Y on average by β_1 units
- Presence or absence of other variables does not affect this

We can extend the previous model with an interaction term

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

This translates in a "linear model"

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

One variable affects other variables influence

Example: Interaction between TV and Radio

We can imagine some interaction in TV and Radio Advertising

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

- Interaction term is the increase of effectiveness of TV for one unit of Radio
- p-value suggests this interaction to be significant
- R^2 increases from 89.7% to 96.8% (69% of missing variance)

	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

TABLE 3.9. *For the* Advertising *data, least squares coefficient estimates associated with the regression of* sales *onto* TV *and* radio*, with an interaction term, as in (3.33).*

Example: MpG Polynomial Regression

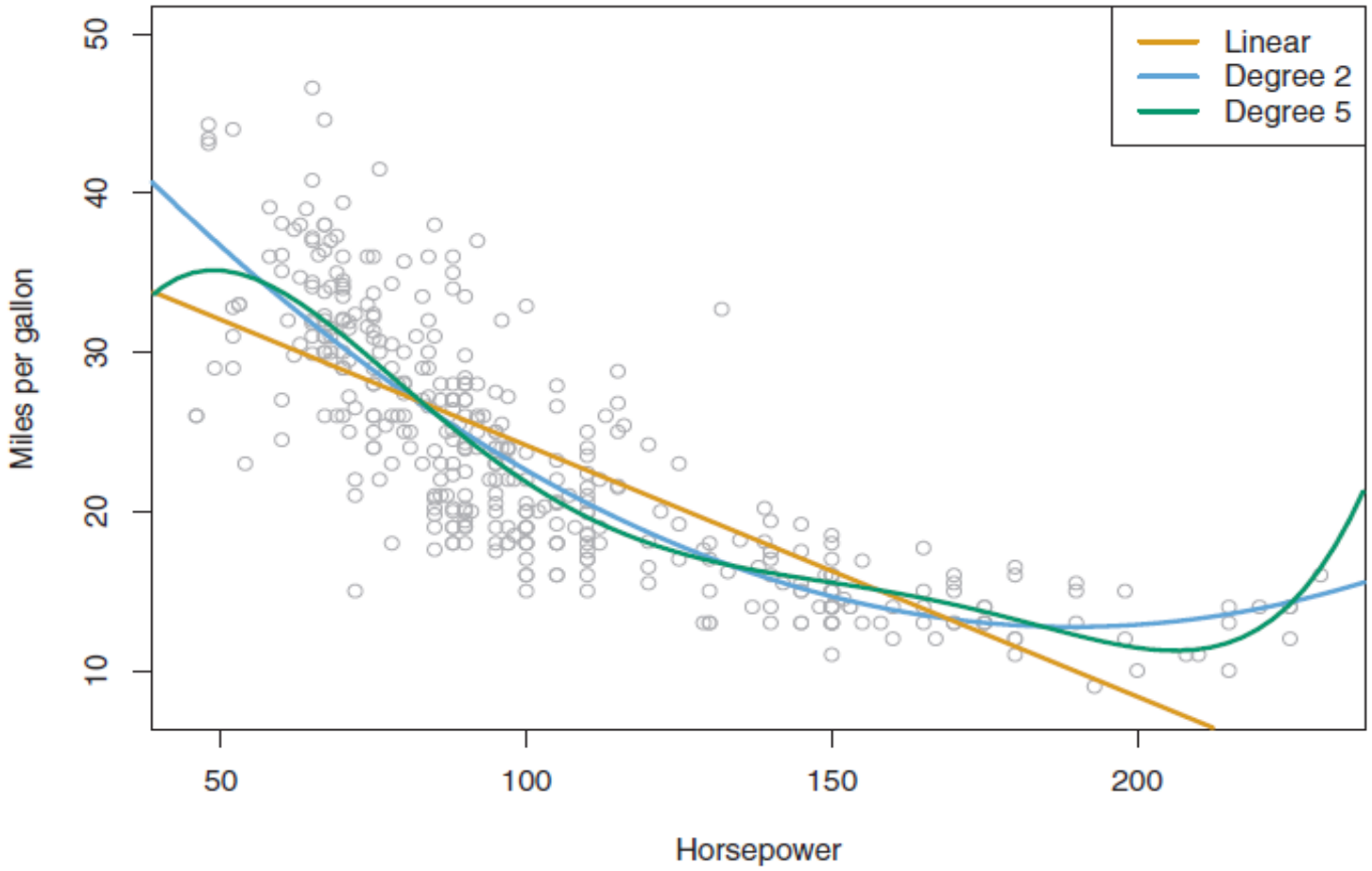


FIGURE 3.8. *The* `Auto` *data set. For a number of cars,* `mpg` *and* `horsepower` *are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes* $\texttt{horsepower}^2$ *is shown as a blue curve. The linear regression fit for a model that includes all polynomials of* `horsepower` *up to fifth-degree is shown in green.*

Non Linear Fitting via Generalized Linear Models

We can use Polynomial Regression to accomodate non linearity

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

- It is still a linear fitting problem !!!!
- A 5th grade polynome is too much, but quadratic term is statistically significant

	Coefficient	Std. error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	−0.4662	0.0311	−15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

TABLE 3.10. *For the* Auto *data set, least squares coefficient estimates associated with the regression of* mpg *onto* horsepower *and* horsepower^2.

Potential Problems in Linear Regression

A number of possible problems might be encountered when fitting the linear regression model.

- Non-linearity of the data
- Dependence of the error terms
- Non-constant variance of error terms
- Outliers
- High leverage points
- Collinearity

> In practice, identifying and overcoming these problems is as much an art as a science. Many pages in countless books have been written on this topic. Since the linear regression model is not our primary focus here, we will provide only a brief summary of some key points.

Non Linearity of the Data

If the linearity assumption does not hold, conclusions might be inaccurate

Check residual plot!

Try to use non linear transformations of the predictor (log X, X^2, sqrt(X), …)

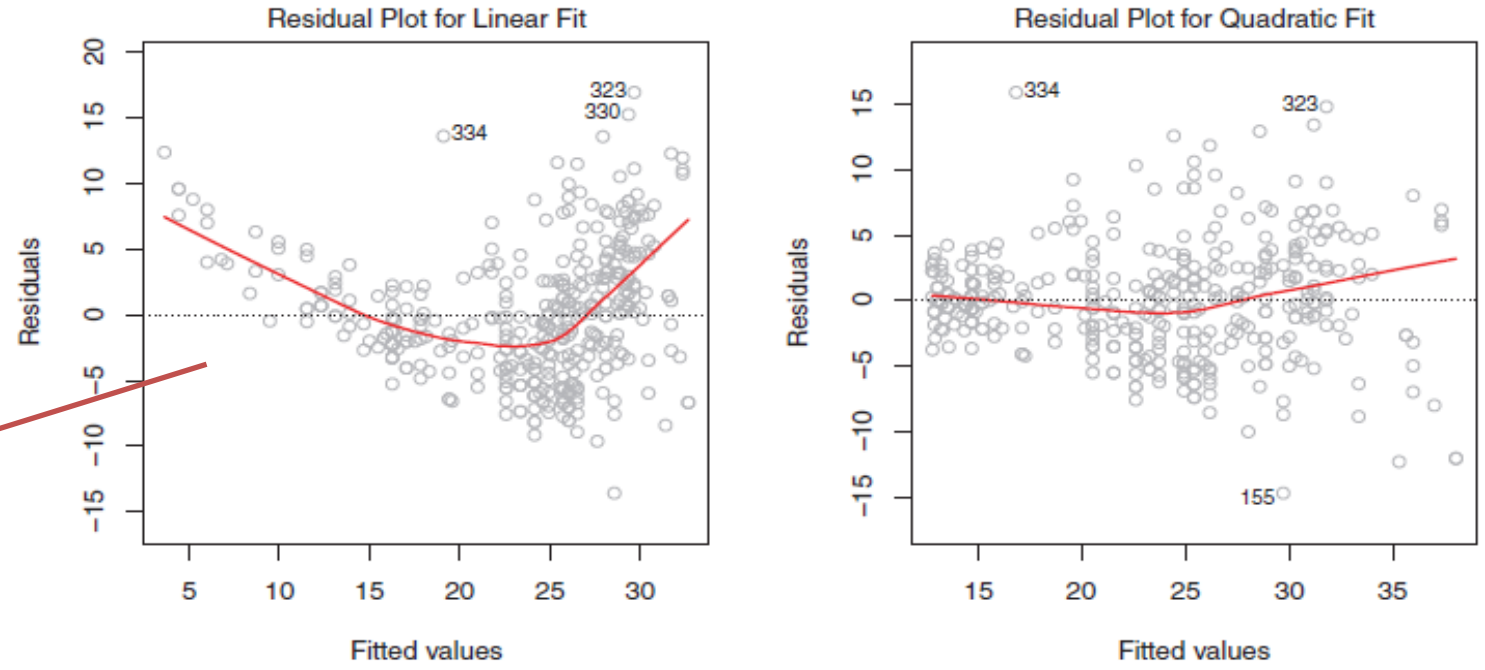


FIGURE 3.9. *Plots of residuals versus predicted (or fitted) values for the* `Auto` *data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend.* Left: *A linear regression of* `mpg` *on* `horsepower`*. A strong pattern in the residuals indicates non-linearity in the data.* Right: *A linear regression of* `mpg` *on* `horsepower` *and* $\texttt{horsepower}^2$*. There is little pattern in the residuals.*

Non Constant Variance of Error Term

Linear Regression assumes no heteroscedasticity in the noise

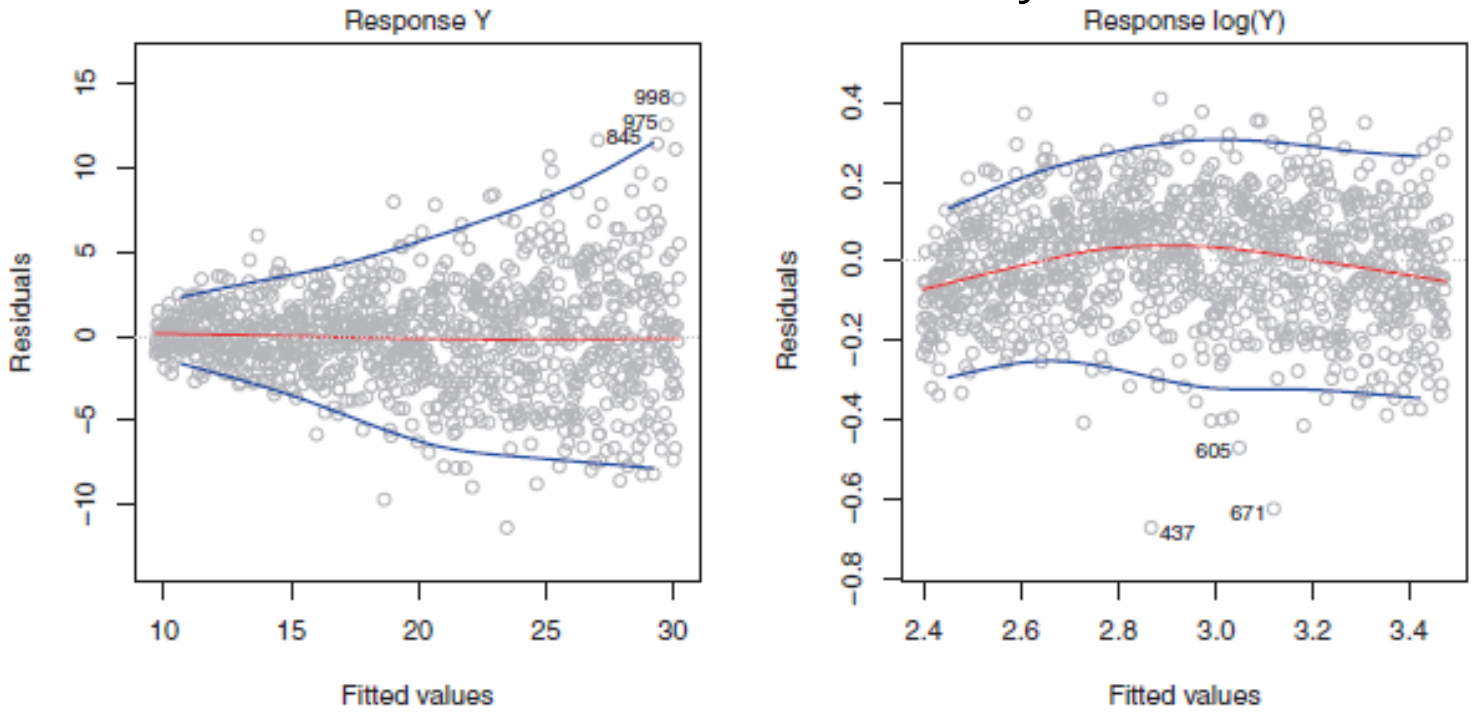


FIGURE 3.11. *Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns.* Left: *The funnel shape indicates heteroscedasticity.* Right: *The predictor has been log-transformed, and there is now no evidence of heteroscedasticity.*

Presence of outliers

An outlier is a point that is too far from its prediction

- Might be due to error in data collection (just remove it)
- Might be due to some missing predictors (revise the model)

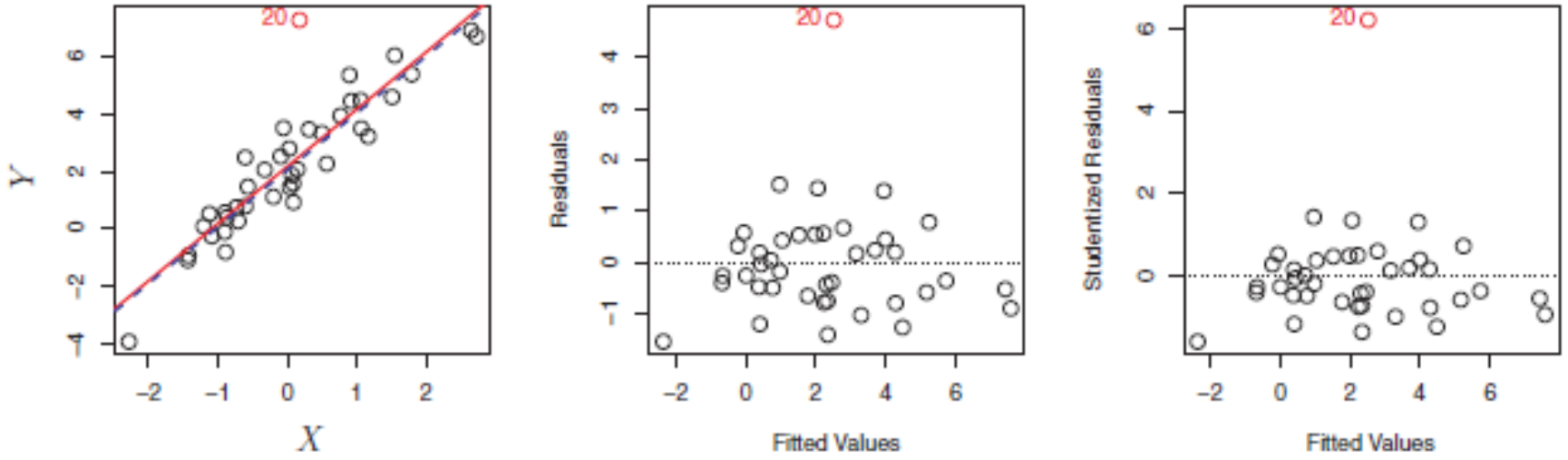


FIGURE 3.12. Left: *The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue.* Center: *The residual plot clearly identifies the outlier.* Right: *The outlier has a studentized residual of 6; typically we expect values between −3 and 3.*

High Leverage Points

High leverage points have unexpected values for a predictor

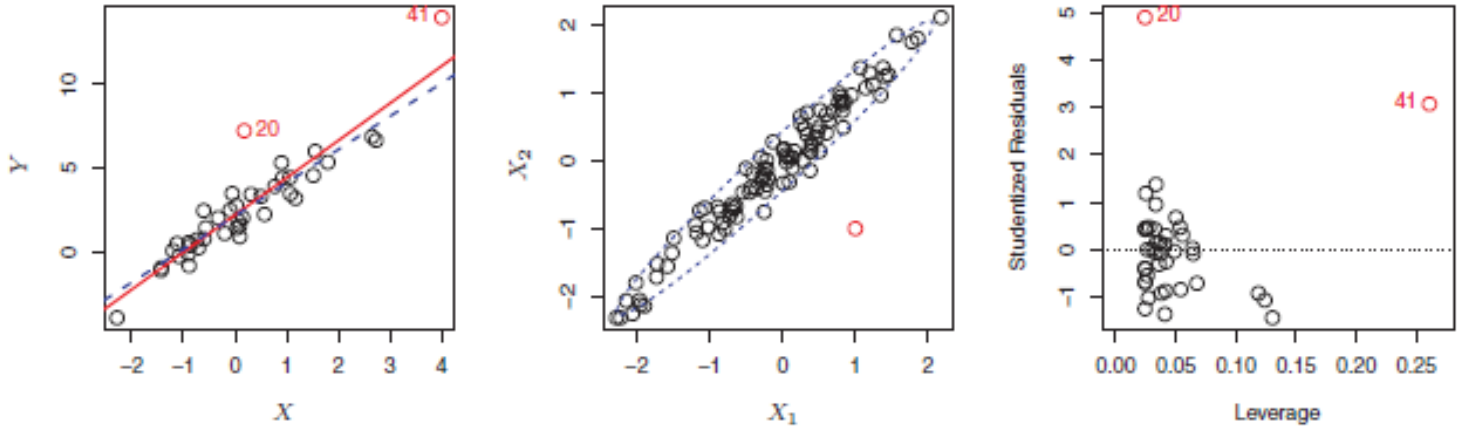


FIGURE 3.13. Left: *Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed.* Center: *The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage.* Right: *Observation* 41 *has a high leverage and a high residual.*

Leverage statistics (between 1/n and 1, average (p+1)/n)

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

Dependence of the Error Term

Errors are supposed to be uncorrelated otherwise standard errors would underestimate true errors …

Tracking phenomenon

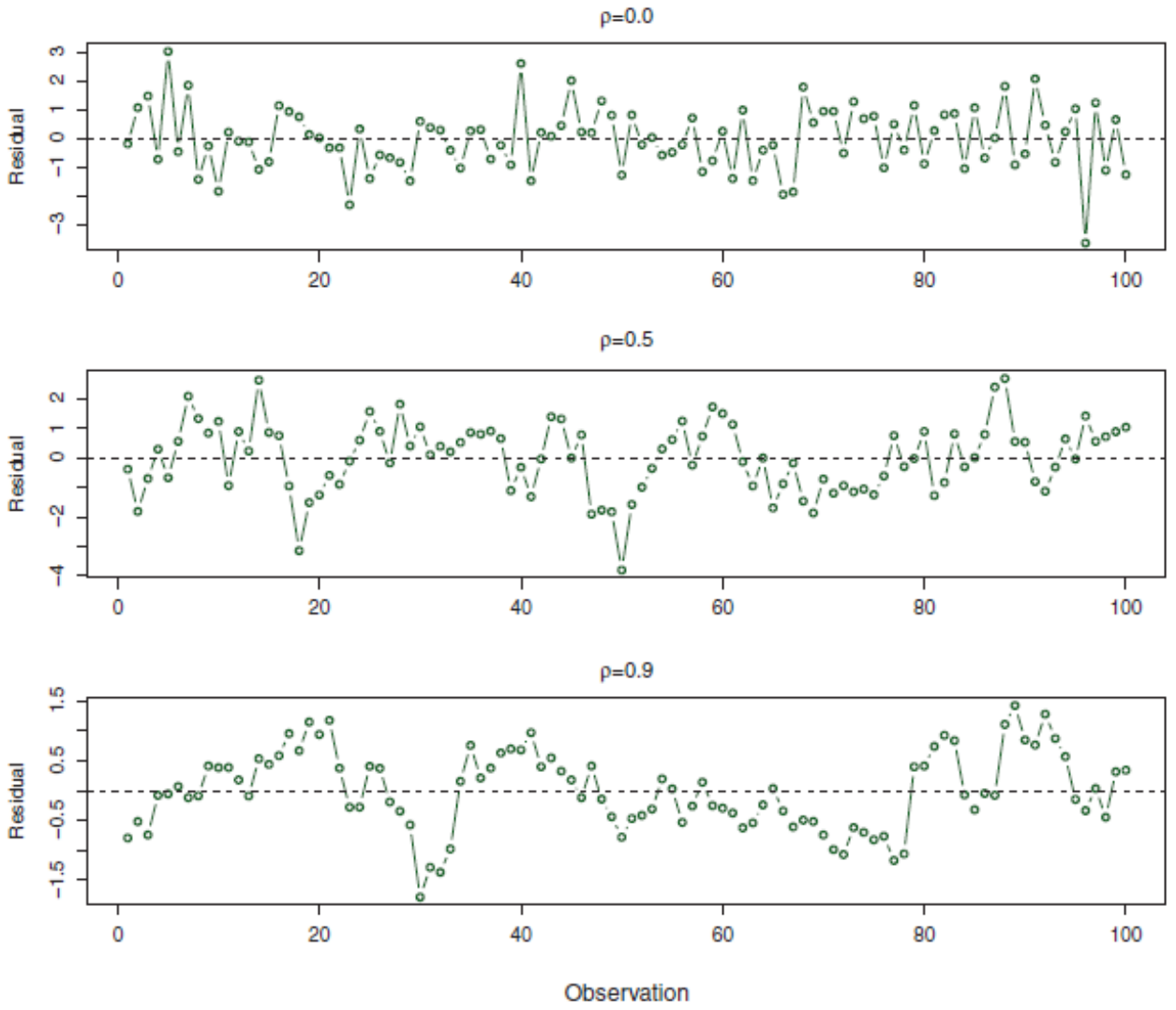


FIGURE 3.10. *Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.*

Colinearity

Factors might be highly related, becomes difficult to separate their effects

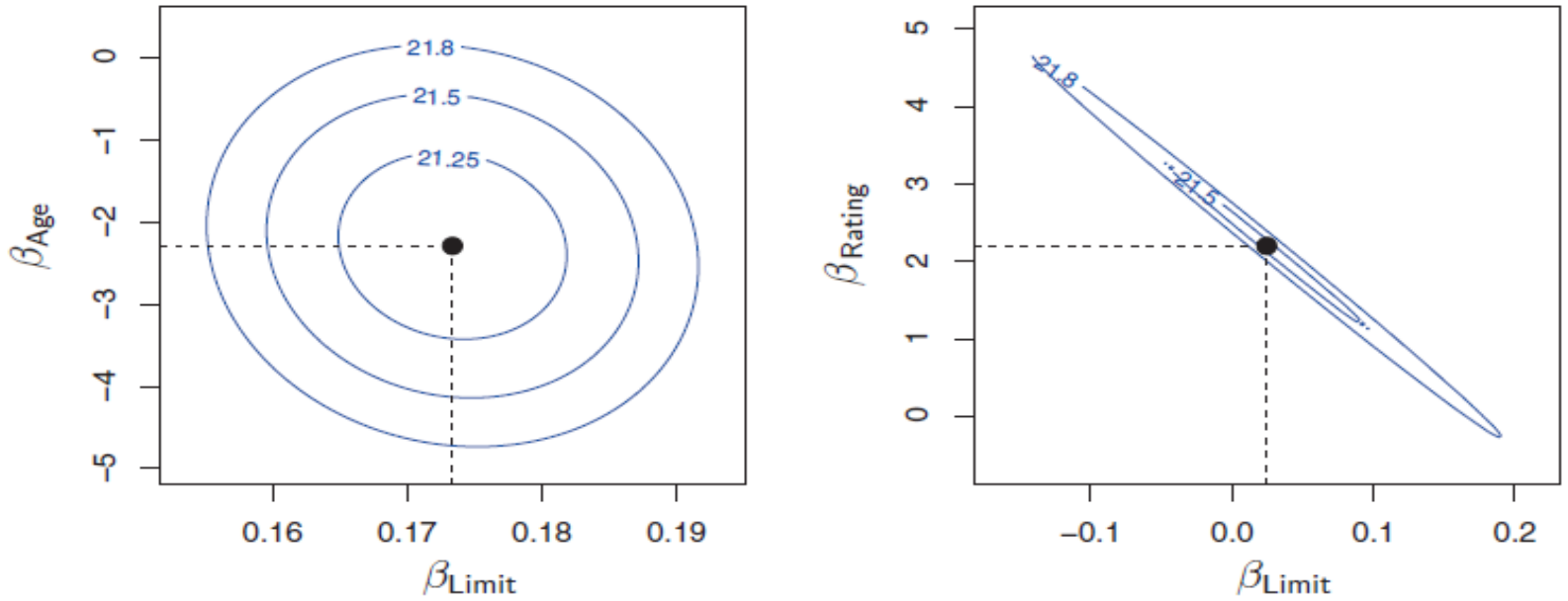
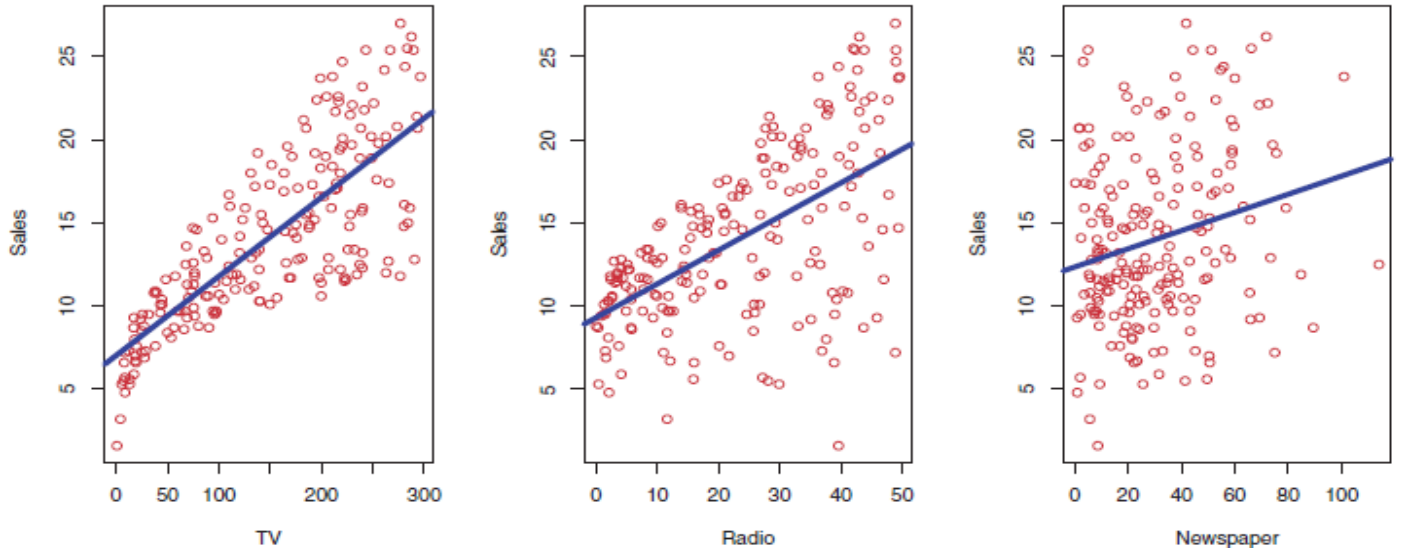


FIGURE 3.15. *Contour plots for the RSS values as a function of the parameters β for various regressions involving the* `Credit` *data set. In each plot, the black dots represent the coefficient values corresponding to the minimum RSS.* Left: *A contour plot of RSS for the regression of* `balance` *onto* `age` *and* `limit`*. The minimum value is well defined.* Right: *A contour plot of RSS for the regression of* `balance` *onto* `rating` *and* `limit`*. Because of the collinearity, there are many pairs $(\beta_{Limit}, \beta_{Rating})$ with a similar value for RSS.*

What can we ask to the data?



- *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 estimate the effect of each medium on sales?*
- *How accurately can we predict future sales?*
- *Is the relationship linear?*
- *Is there synergy among the advertising media?*

POLITECNICO
MILANO 1863

Model (and Features) Selection

Matteo Matteucci, PhD (matteo.matteucci@polimi.it)

Artificial Intelligence and Robotics Laboratory
Politecnico di Milano

AIRLAB
ARTIFICIAL INTELLIGENCE AND ROBOTICS LAB

Improved Linear Regression

We can devise alternative procedures to least squares

- Improve prediction accuracy: if number of data is limited (or p is big) we might have “low bias” but too “high variance” (overfitting) and a poor prediction
- Improve model interpretability: irrelevant variables, beside impacting on accuracy, make models unnecessary complex and difficult to interpret

Several alternatives to remove unnecessary features (predictors)

- Subset Selection: selection of the input variables
- Shrinkage (or regularization): reduction of model variance
- Dimension reduction: projection on an input subspace

Variable selection

Select the variables which are really associated to the prediction

- Exhaustive exploration of model space (2^P)
- Forward selection
- Backward selection
- Mixed selection

If p=30 the number of possible models is 1.073.741.824

Exhaustive exploration is unfeasible because of exponential complexity

- $Y = \beta_0$
- $Y = \beta_0 + \beta_1 * X_1$
- $Y = \beta_0 + \beta_2 * X_2$
- $Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$

Different possible metrics, e.g., C_P, AIC, BIC, adjusted R^2

Recall the Bias-Variance trade-off

For a Linear Model

$$\mathrm{Err}(x_0) = \mathrm{E}[(Y - \hat{f}_\lambda)^2 | X = x_0]$$

$$\sigma^2 + \left[f(x_0) - \mathrm{E}\hat{f}(x_i) \right]^2 + \|\mathbf{h}(x_0)\|^2 \sigma^2$$

$$\frac{1}{N} \sum_{i=1}^{N} \mathrm{Err}(\mathrm{x_i}) = \sigma^2 + \frac{1}{N} \sum_{i=1}^{N} [f(x_i) - \mathrm{E}\hat{f}(x_i)]^2 + \frac{p}{N} \sigma^2$$

Best Subset Selection

Fit a least squares regression for any possible input combination

- A total of 2^P need to be compared
- Best Subset Selection introduces a procedure to evaluate them systematically

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 RS[illegible]st R^2.
3. Select a single best model from amon[illegible] cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Cross-validated prediction error!

Example: Credit data best subset selection

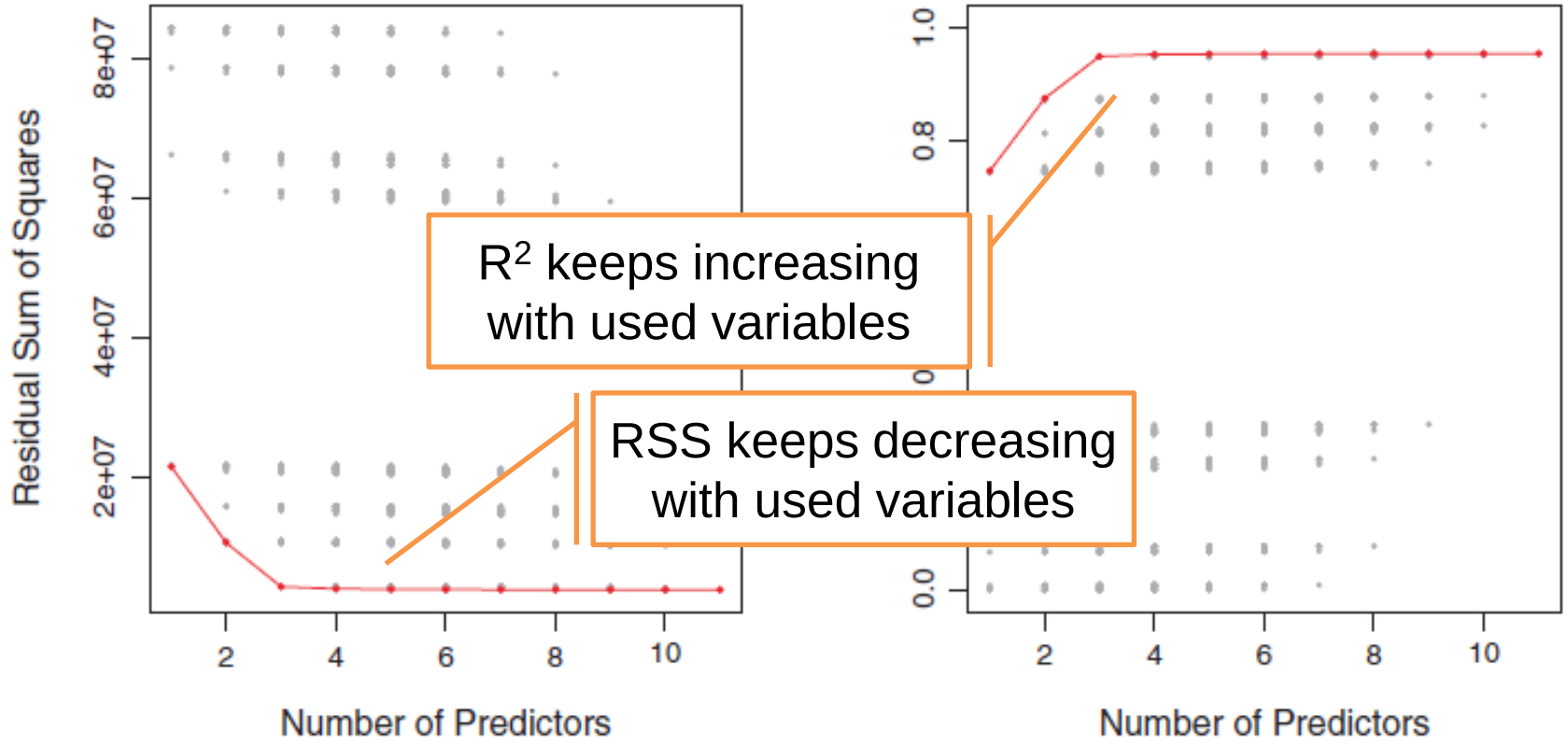


FIGURE 6.1. *For each possible model containing a subset of the ten predictors in the* `Credit` *data set, the RSS and* R^2 *are displayed. The red frontier tracks the best model for a given number of predictors, according to RSS and* R^2*. Though the data set contains only ten predictors, the x-axis ranges from* 1 *to* 11*, since one of the variables is categorical and takes on three values, leading to the creation of two dummy variables.*

Forward Stepwise Selection

Forward stepwise selection is a computationally efficient alternative

- Starts from an empty model with no predictors
- Adds one predictor at the time until "all are in"
- At each stage adds "most improving" variable

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.

Forward Stepwise vs Best Subset Selection

Forward Stepwise is a greedy approach

- Needs to fit 1+p(p+1)/2 models instead of 2^P
- It can be used also when n < p (it will stop with k<n variables)
- It does not "reconsider" its choices and might result in a suboptimal subset

# Variables	Best subset	Forward stepwise
One	rating	rating
Two	rating, income	rating, income
Three	rating, income, student	rating, income, student
Four	cards, income	rating, income,
	student, limit	student, limit

TABLE 6.1. *The first four selected models for best subset selection and forward stepwise selection on the* `Credit` *data set. The first three models are identical but the fourth models differ.*

Backward Stepwise Selection

Backward stepwise selection is yet computationally efficient

- Starts from the model having all predictors
- At each stage removes the "least useful" variable

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.

- Greedy as Forward Stepwise, but cannot be used when n < p

Choosing the Optimal Model (theory)

Feature subset selection algorithms "optimize" the number of features according to RSS and R^2, but what about the test set?

Several approaches estimate test error correcting the training error

Strong statistical background

- Mallows CP

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

d = number of predictors

$\hat{\sigma}^2$ = estimate of the variance associated to the complete model

- Akaike Information Criterion

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

Some constants omitted, but proportional to C_P

- Bayesian Information Criterion

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

Some constants omitted, more stringent than C_P

- Adjusted R^2

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

Equivalent to $\frac{\mathrm{RSS}}{n-d-1}$

Example: Credit data feature selection

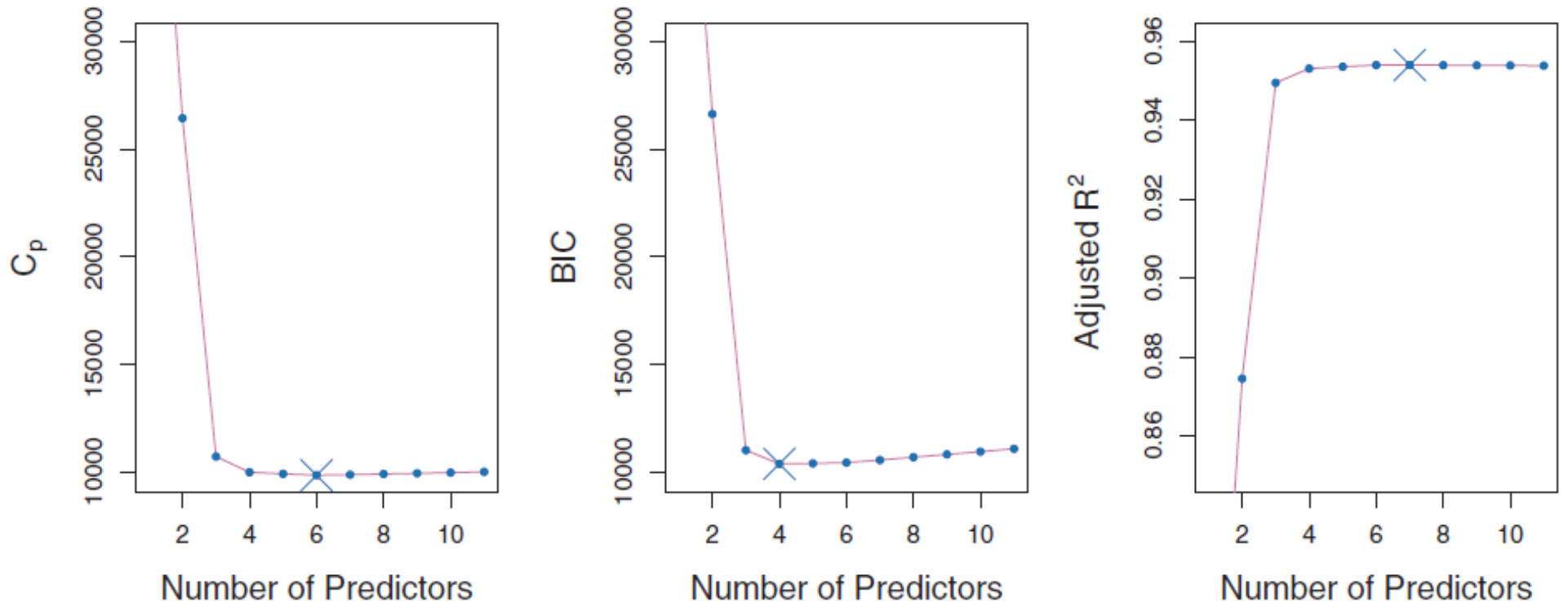


FIGURE 6.2. *C_p, BIC, and adjusted R^2 are shown for the best models of each size for the* `Credit` *data set (the lower frontier in Figure 6.1). C_p and BIC are estimates of test MSE. In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected.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

Choosing the Optimal Model (practice)

We can use data itself to estimate the error on new data

- We can use an hold out set and perform validation

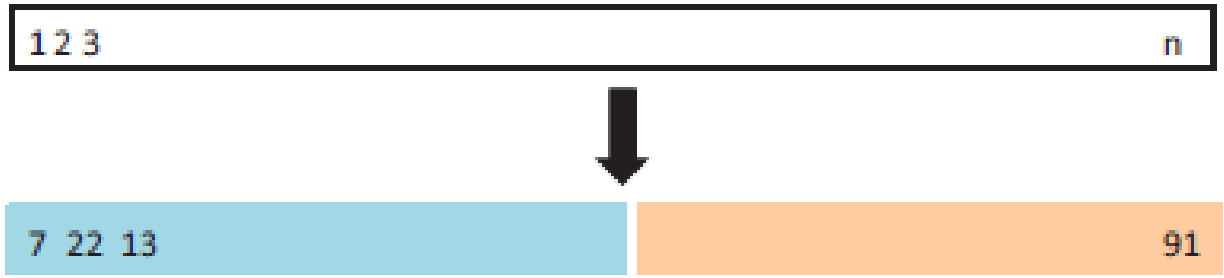


FIGURE 5.1. *A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.*

Choosing the Optimal Model (practice)

We can use data itself to estimate the error on new data

- We can use an hold out set and perform validation
- We can use k-fold cross-validation

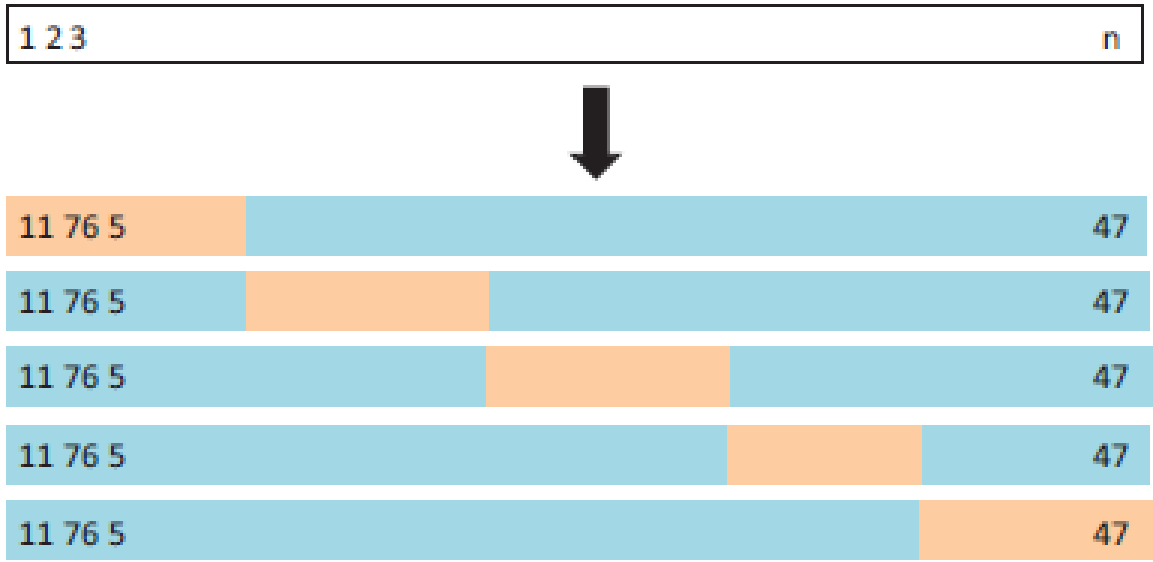


FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

Example: Credit data feature selection

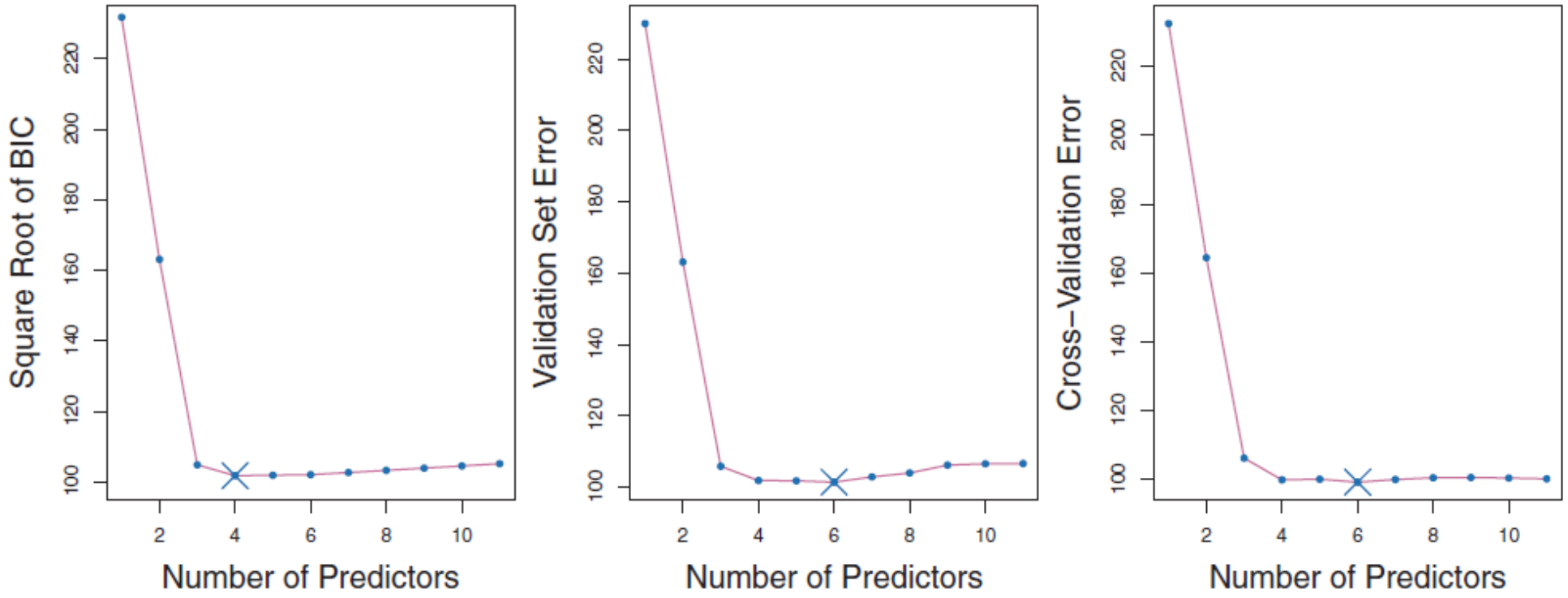


FIGURE 6.3. *For the* `Credit` *data set, three quantities are displayed for the best model containing d predictors, for d ranging from* 1 *to* 11. *The overall* best *model, based on each of these quantities, is shown as a blue cross.* Left: *Square root of BIC.* Center: *Validation set errors.* Right: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

Improved Linear Regression

We can devise alternative procedures to least squares

- Improve prediction accuracy: if number of data is limited (or p is big) we might have "low bias" but too "high variance" (overfitting) and a poor prediction
- Improve model interpretability: irrelevant variables, beside impacting on accuracy, make models unnecessary complex and difficult to interpret

Several alternatives to remove unnecessary features (predictors)

- Subset Selection: selection of the input variables
- Shrinkage (or regularization): reduction of model variance
- Dimension reduction: projection on an input subspace

Shrinkage Methods: Ridge Regression

Ordinary Least Squares (OLS) minimizes

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Ridge Regression minimizes a slightly different function

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- $\lambda \geq 0$ is a tuning parameter to be estimated experimentally
- Shrinkage does not apply to intercept, with centered variables

$$\hat{\beta}_0 = \bar{y} = \sum_{i=1}^{n} y_i / n$$

- $\lambda \sum_j \beta_j^2$ is called shrinkage penalty
- as $\lambda \to \infty$ parameters shrink to zero

Example: Ridge Regression on Credit data

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}}$$

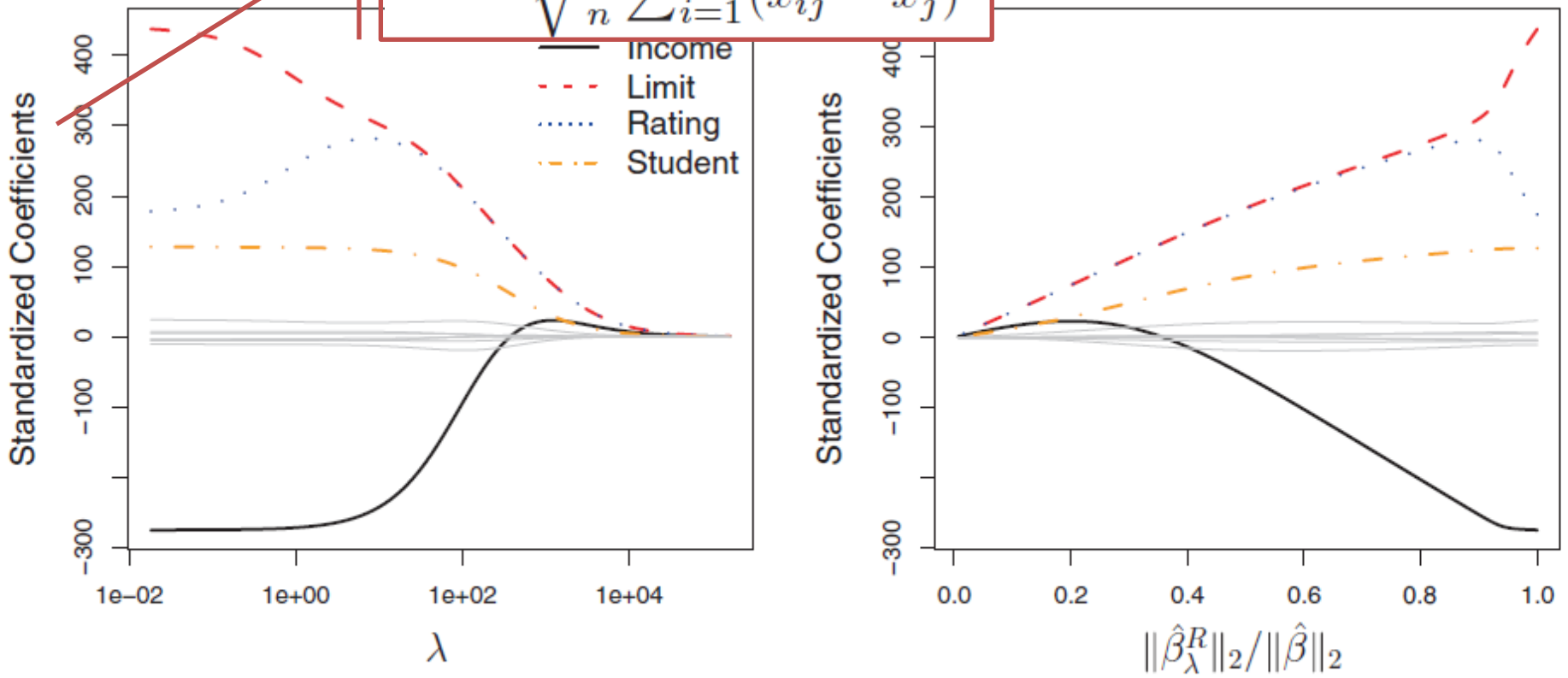


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$.

$$\|\beta\|_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$$

Ridge Regression vs. Ordinary Least Squares

Ridge regression improves OLS because of a reduced model variance (i.e., a better bias-variance trade-off)

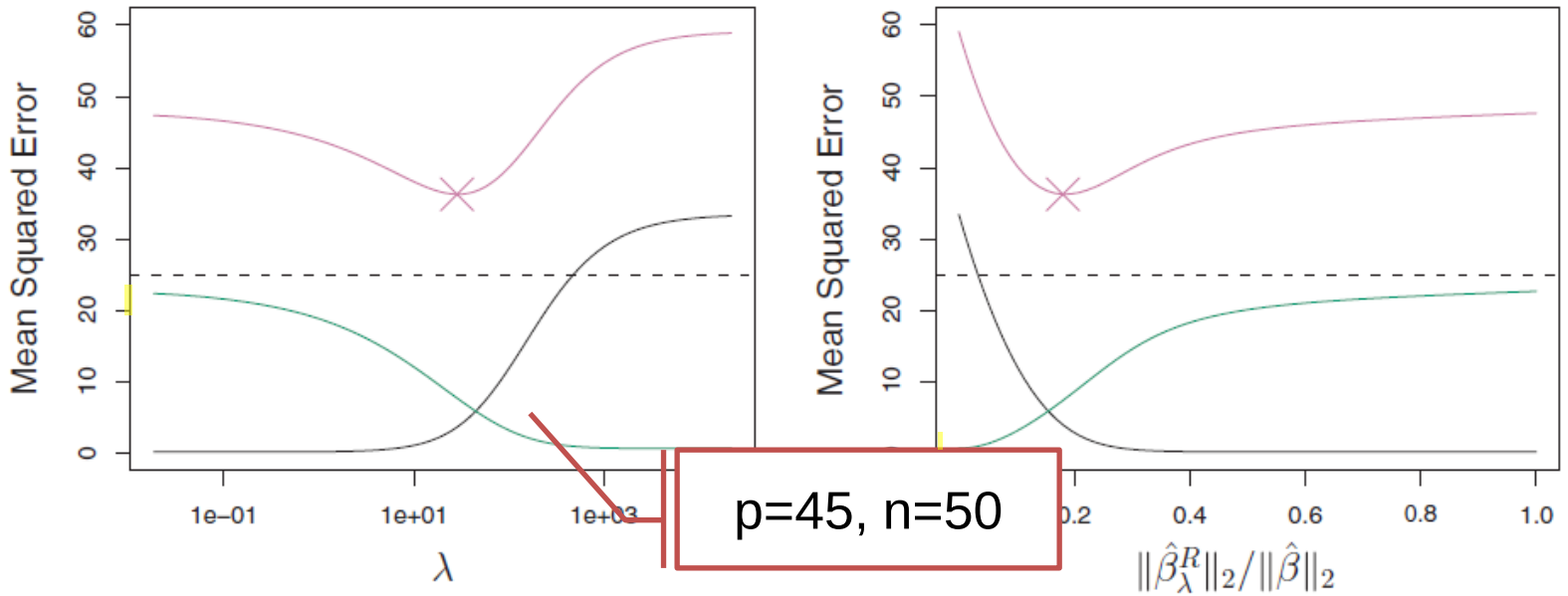


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$. *The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

Shrinkage Methods: The Lasso

Ridge regression more efficient than subset selection, uses all the p input

The Lasso is an alternative to shrink regression coefficients

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|$$

- The $\|\beta\|_1=\sum|\beta_j|$ forces coefficients to be exactly zero
- The Lasso performs *variable section*
- Models are simpler, sparse, and easy to interpret

Example: The Lasso and the Credit data

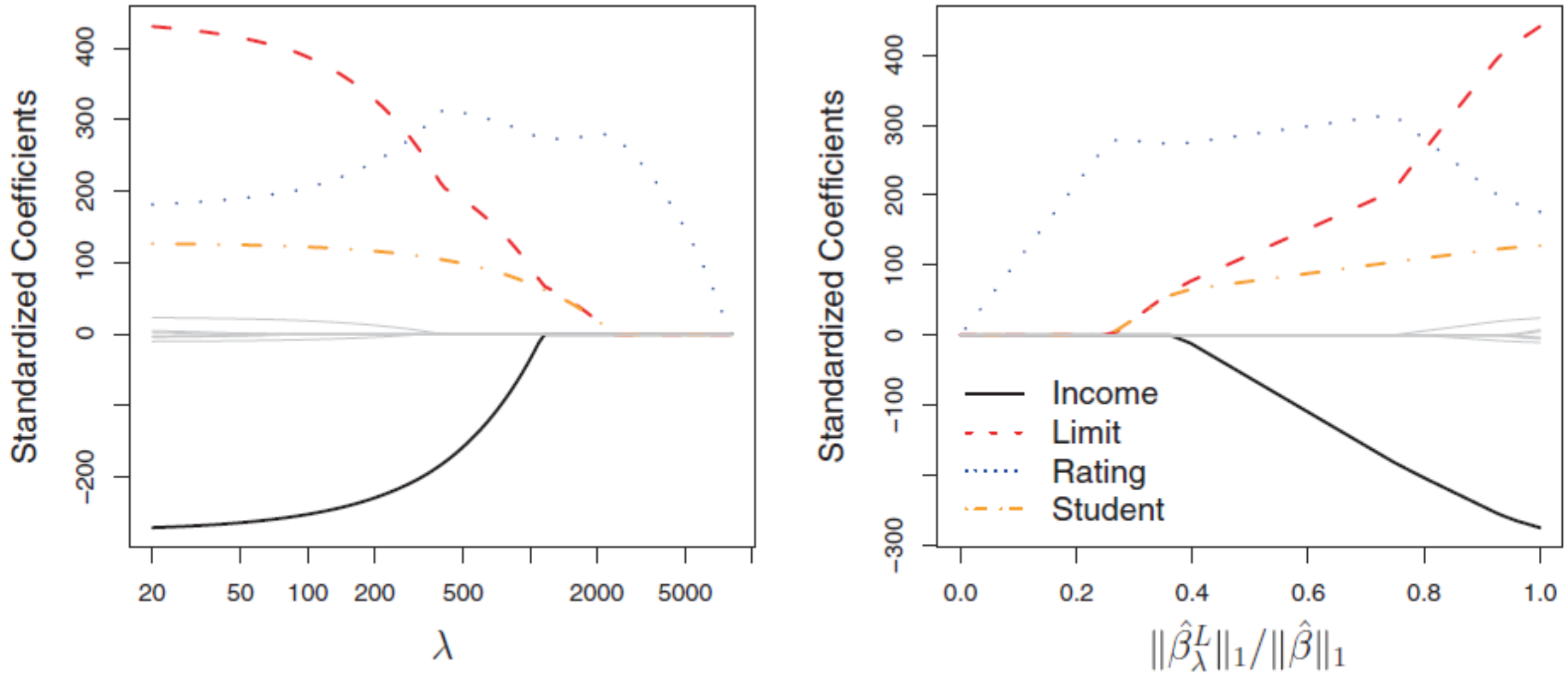


FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

Another interpretation of shrinkage

We can show that Ridge regression solves the problem

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s$$

While The Lasso solves the problem

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

They approximate the Best Subset Selection

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s$$

How can Lasso select variables?

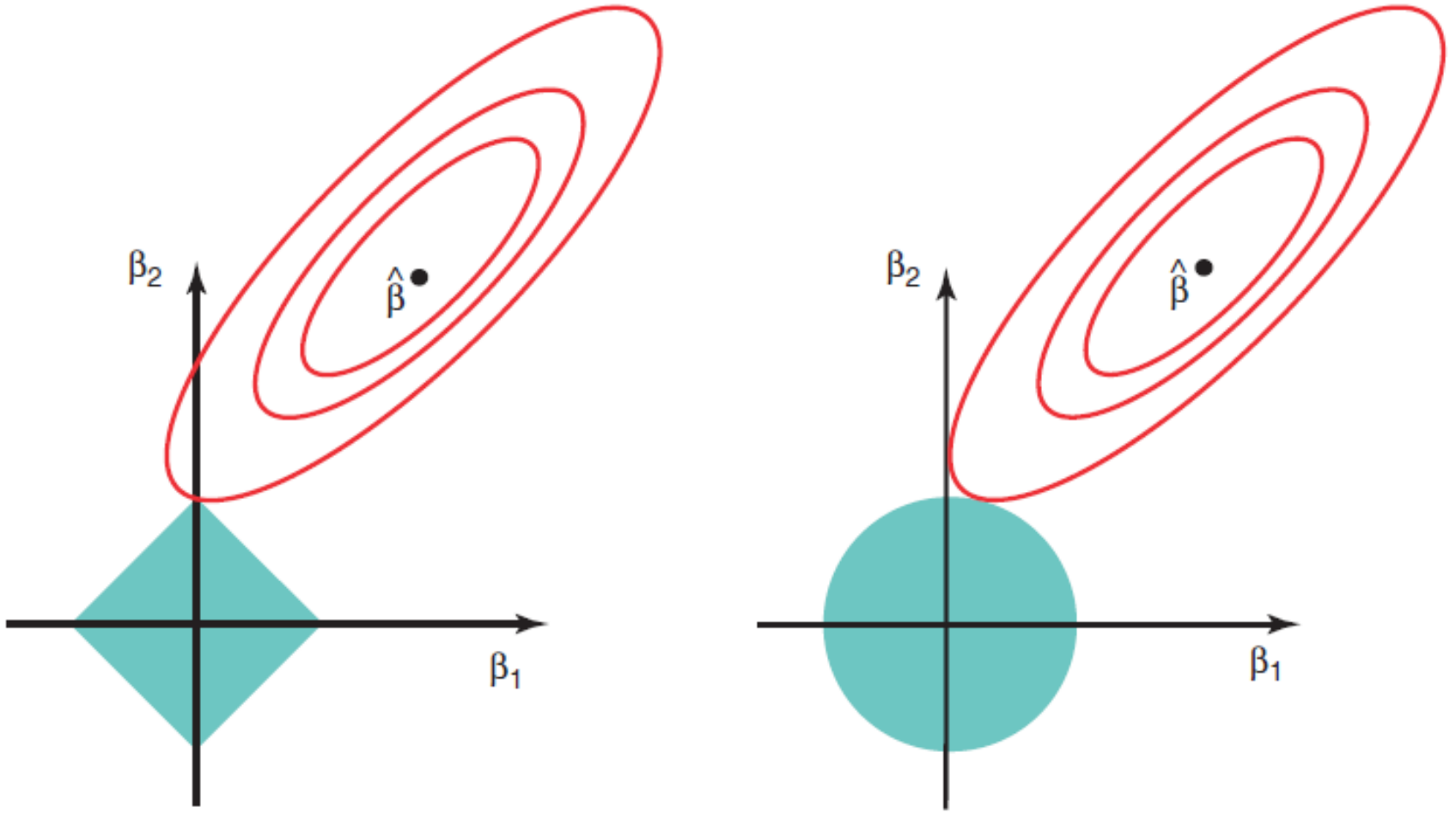


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Lasso vs. Ridge Regression (p=45 all useful)

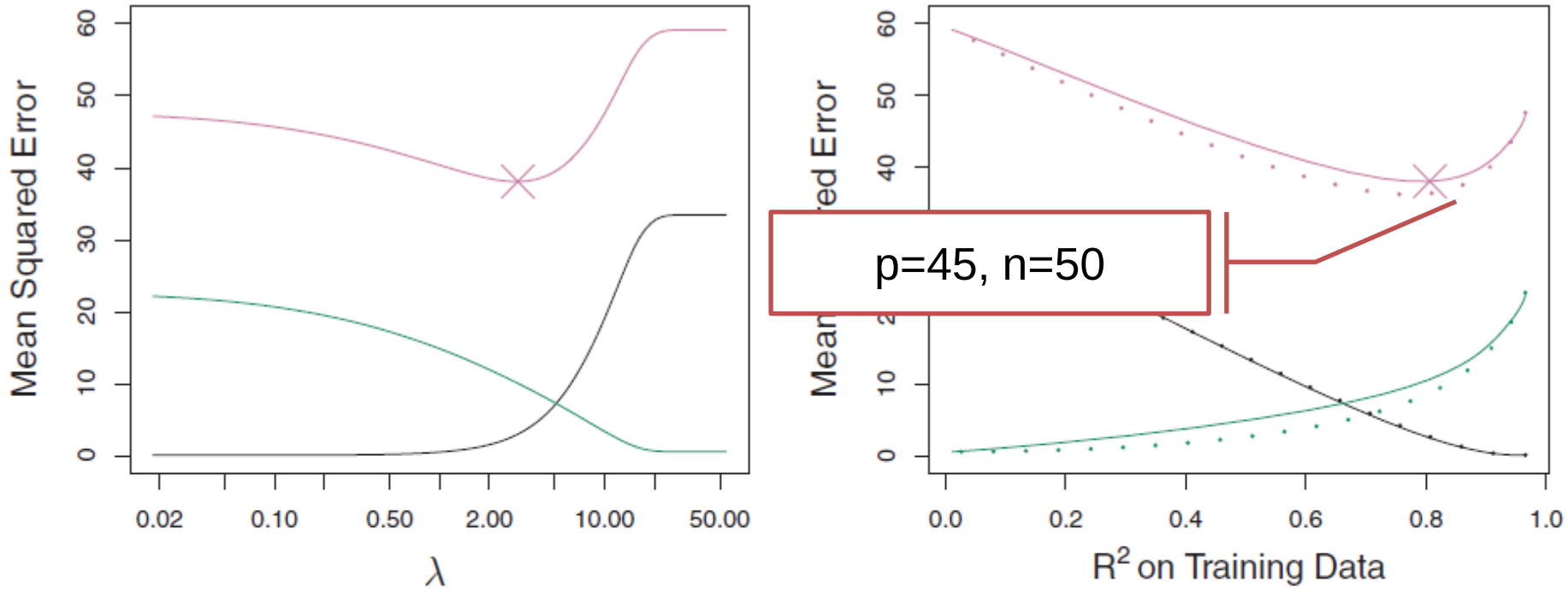


FIGURE 6.8. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on a simulated data set.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dashed). Both are plotted against their* R^2 *on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

Lasso vs. Ridge Regression (p=2 only useful)

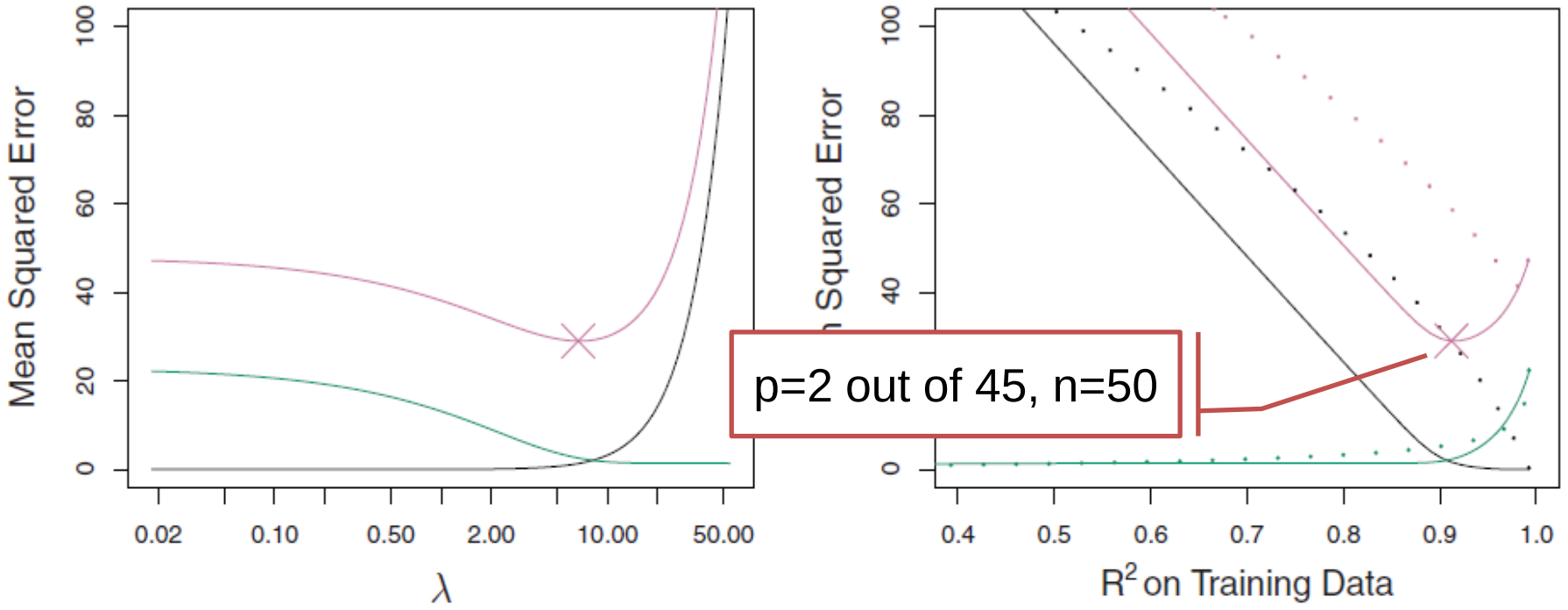


FIGURE 6.9. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso. The simulated data is similar to that in Figure 6.8, except that now only two predictors are related to the response.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dashed). Both are plotted against their* R^2 *on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

Selection of the tuning parameter (Ridge)

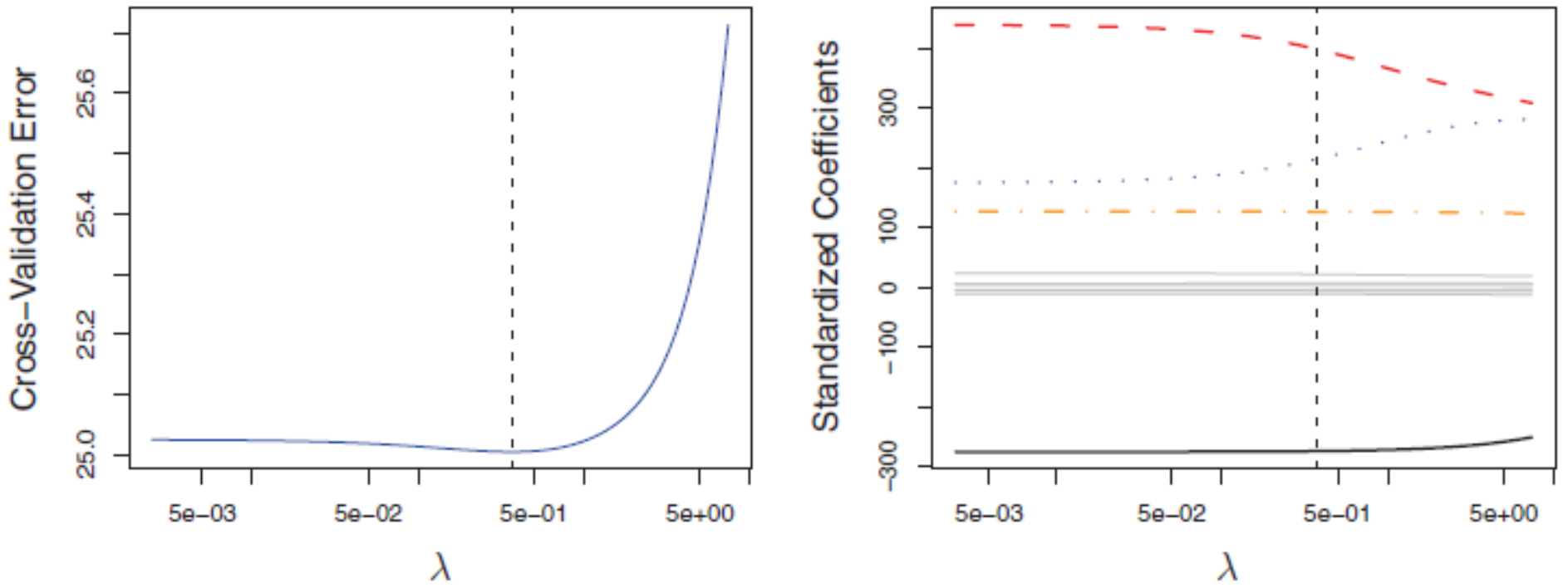


FIGURE 6.12. Left: *Cross-validation errors that result from applying ridge regression to the* `Credit` *data set with various value of* λ. Right: *The coefficient estimates as a function of* λ. *The vertical dashed lines indicate the value of* λ *selected by cross-validation.*

Selection of the tuning parameter (Lasso)

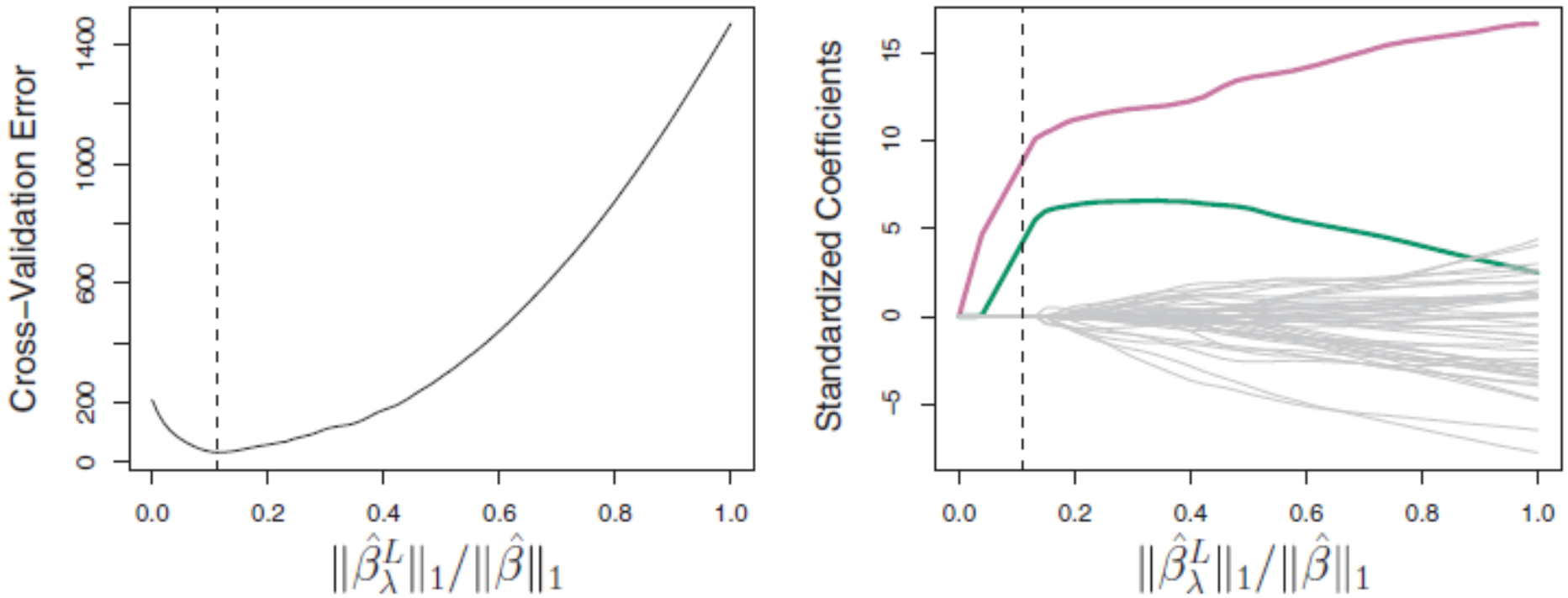


FIGURE 6.13. Left: *Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set from Figure 6.9.* Right: *The corresponding lasso coefficient estimates are displayed. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.*

Bayesian interpretation

The posterior for the coefficients can be written as

$$p(\beta|X,Y) \propto f(Y|X,\beta)p(\beta|X) = f(Y|X,\beta)p(\beta)$$

Assuming the usual linear model $Y = \beta_0 + X_1\beta_1 + \ldots + X_p\beta_p + \epsilon$

- Having independent errors drawn from a normal distribution

If we assume $p(\beta) = \prod_{j=1}^{p} g(\beta_j)$

- ***Ridge regression***: we assume a Gaussian prior with zero mean and variance being a function of lambda
- ***Lasso***: we assume a double-exponential (Laplace) with zero mean and scale parameter a function of lambda

Ridge and Lasso priors …

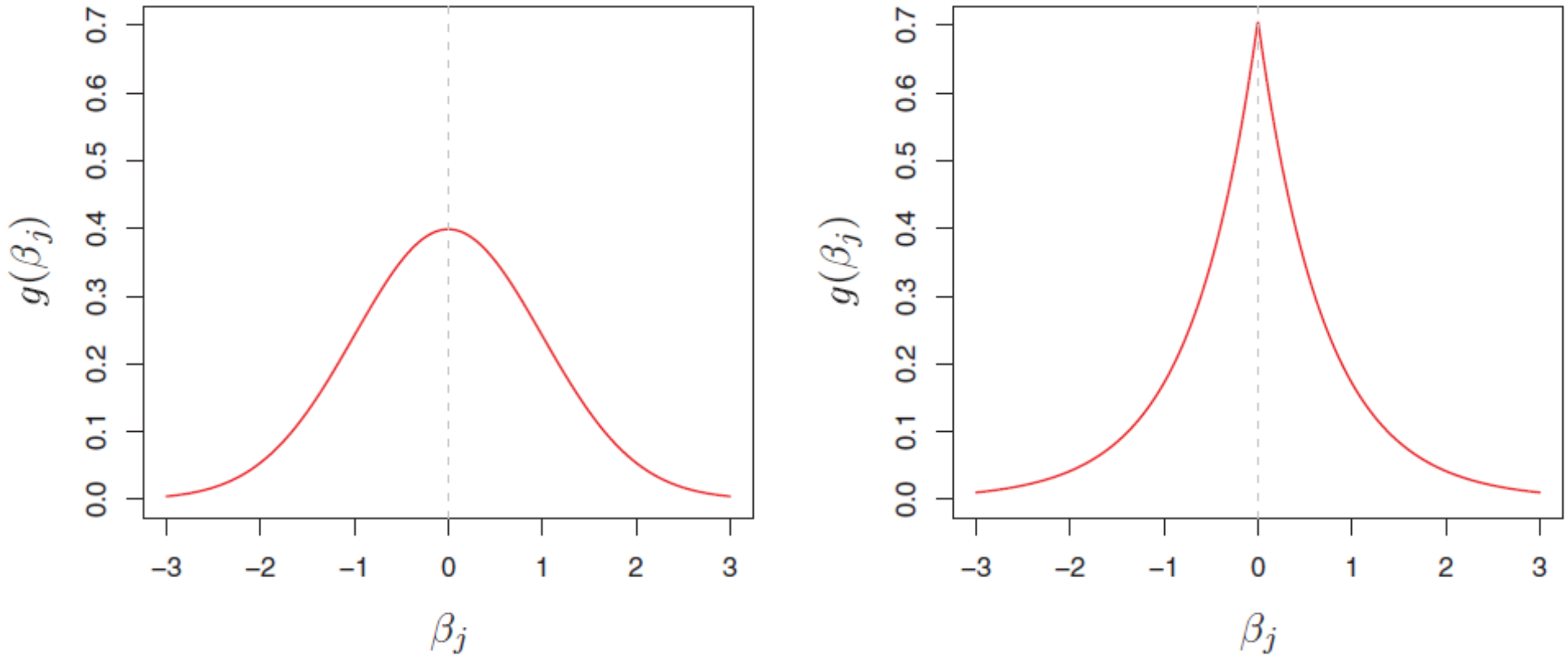


FIGURE 6.11. Left: *Ridge regression is the posterior mode for β under a Gaussian prior.* Right: *The lasso is the posterior mode for β under a double-exponential prior.*