POLITECNICO DI MILANO

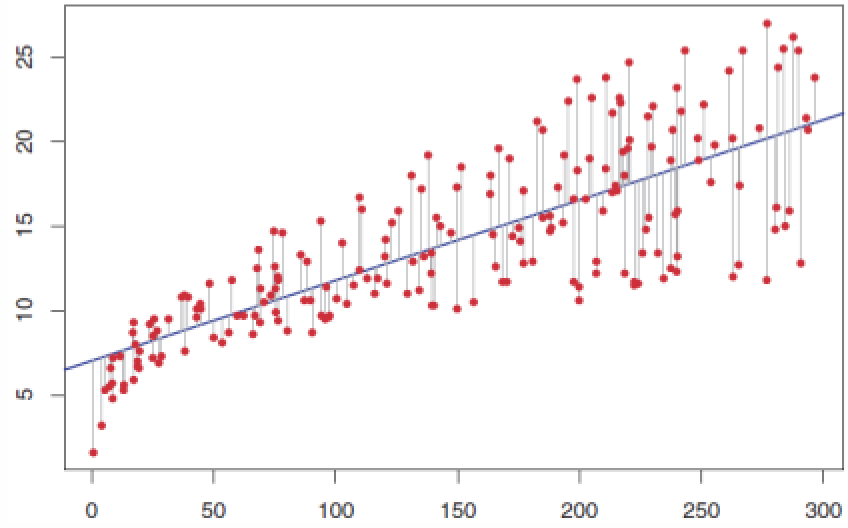
Pattern Analysis and Machine Intelligence

Linear Regression

Prof. Matteo Matteucci

Outline

- Simple Linear Regression Model
  - Least Squares Fit
  - Measures of Fit
  - Inference in Regression
- Multi Variate Regession Model
  - Least Squares Fit
  - Inference in Regression
- Other Considerations in Regression Model
  - Qualitative Predictors
  - Interaction Terms
- Potential Fit Problems
- Linear vs. KNN Regression



Example: Increasing Sales by Advertising

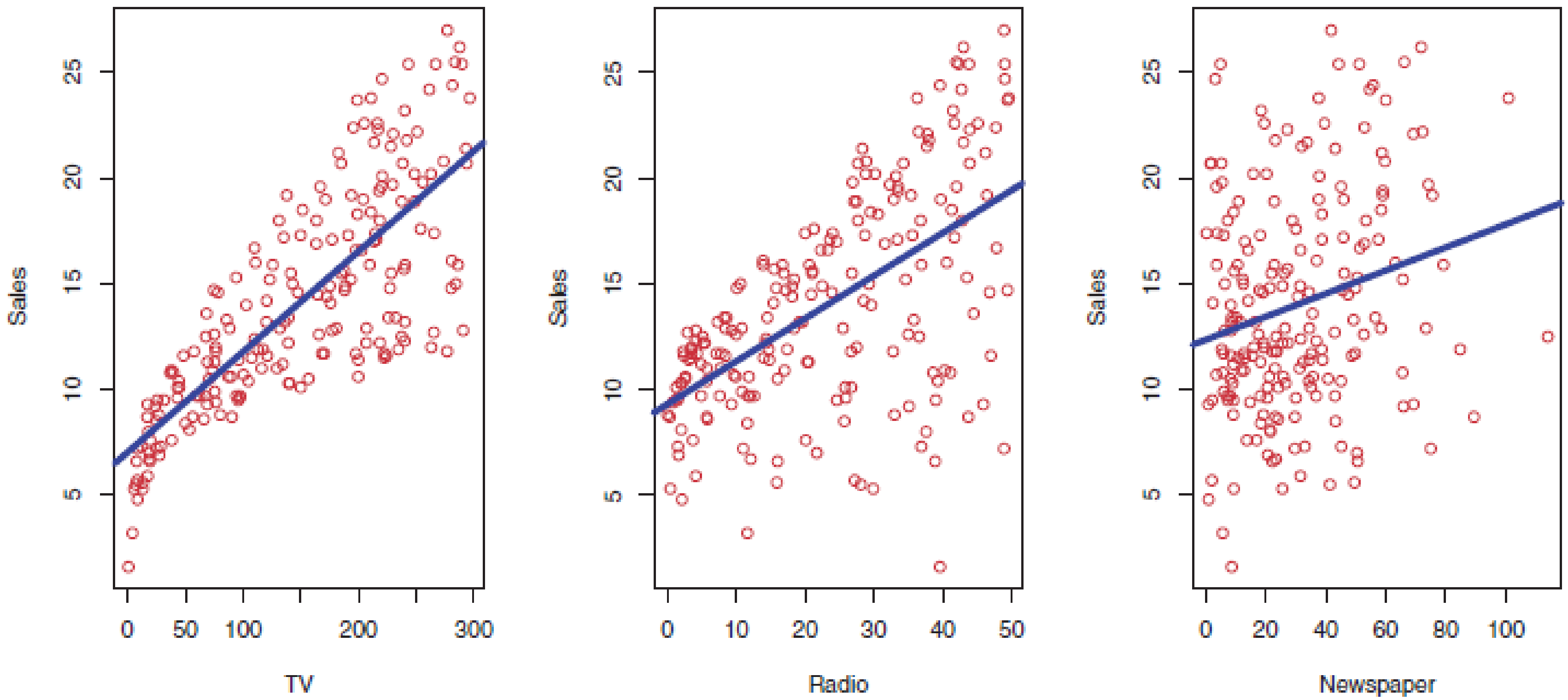
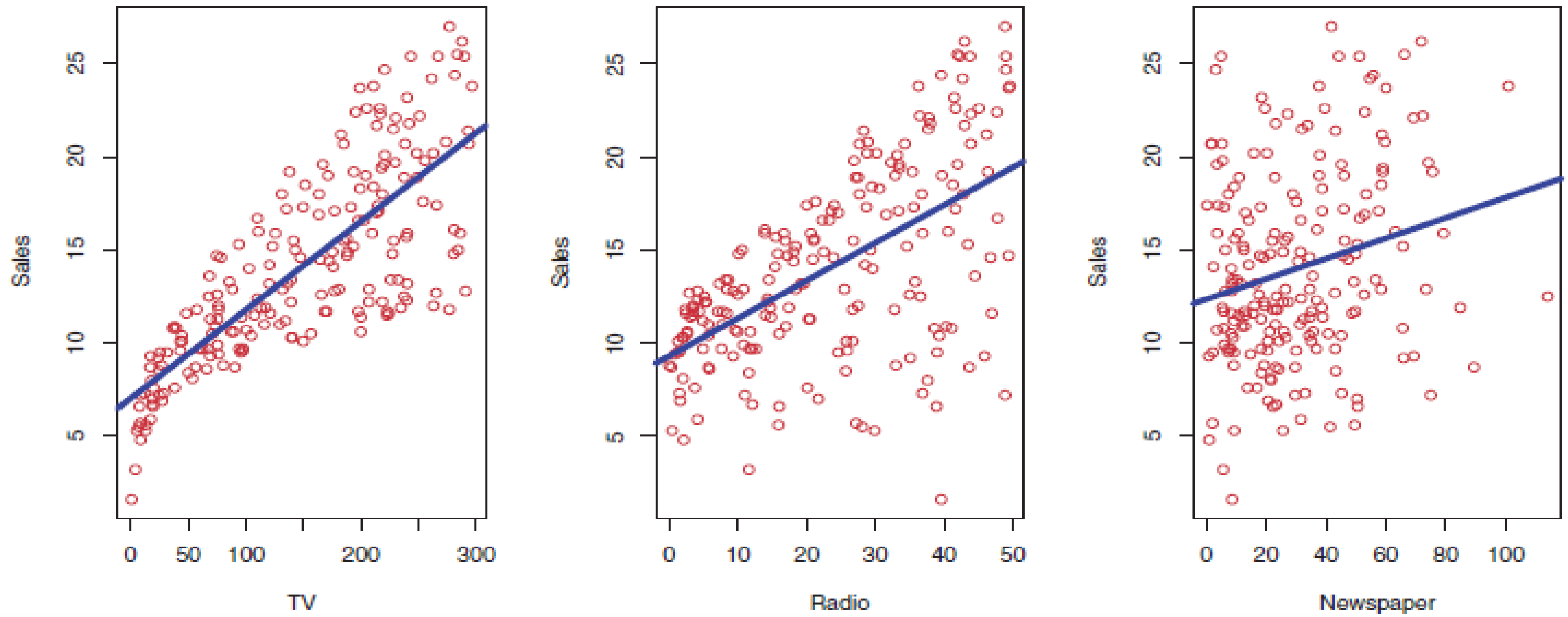


FIGURE 2.1. *The* Advertising *data set. The plot displays* sales*, in thousands of units, as a function of* TV*,* radio*, and* newspaper *budgets, in thousands of dollars, for* 200 *different markets. In each plot we show the simple least squares fit of* sales *to that variable, as described in Chapter 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?*
- *How accurately can we predict future 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

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

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

Example: TV Advertising vs Sales

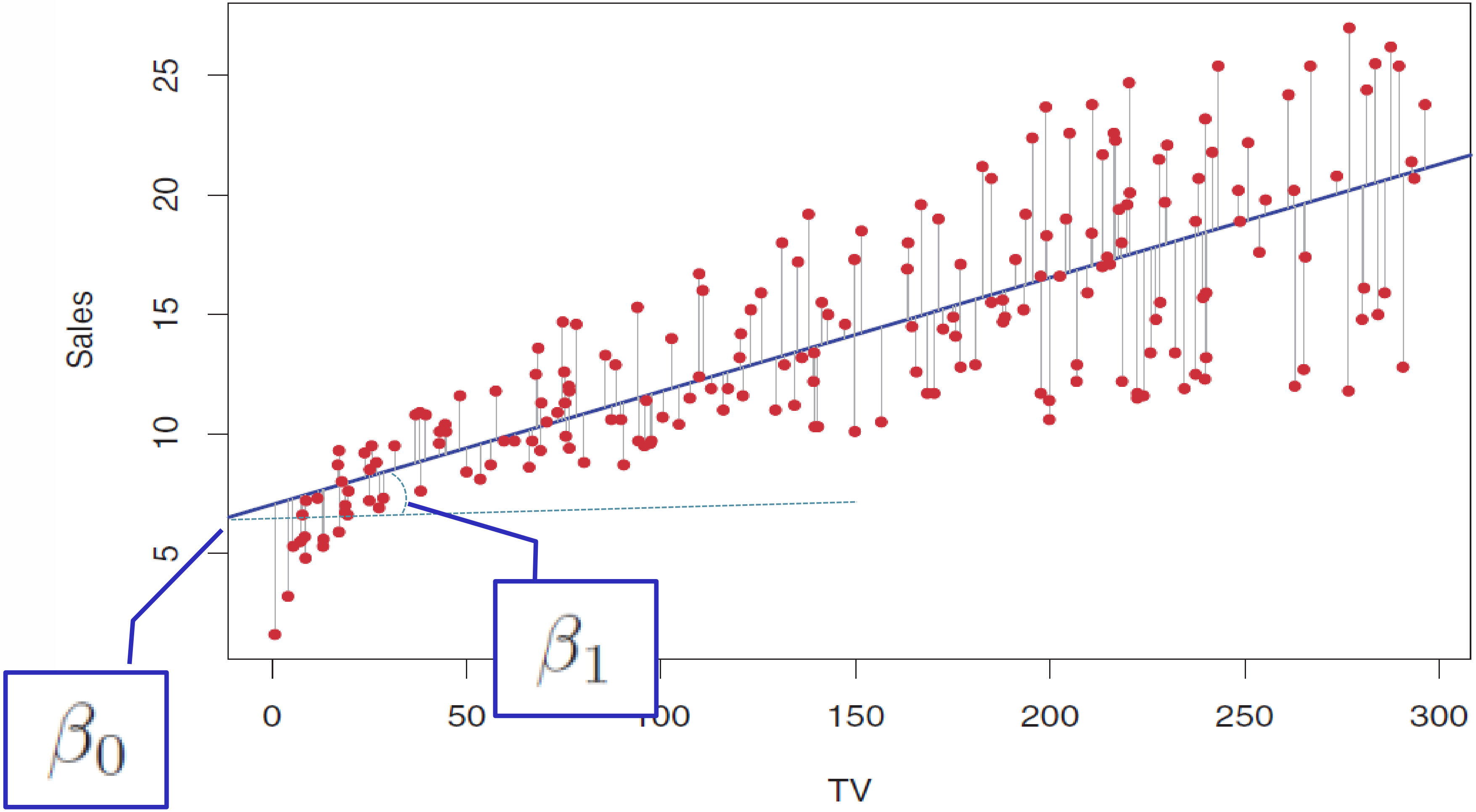


FIGURE 3.1. *For the* Advertising *data, the least squares fit for the regression of* sales *onto* TV *is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures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

Least squares fitting

- Least square fitting minimizes RSS (Residual Sum of Squares)

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

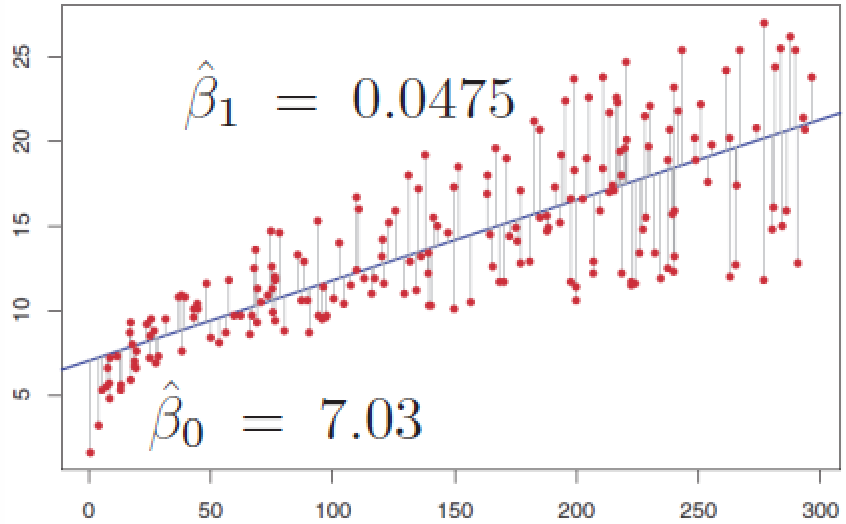
$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

$$\mathrm{RSS} = (y_1 - \hat\beta_0 - \hat\beta_1 x_1)^2 + (y_2 - \hat\beta_0 - \hat\beta_1 x_2)^2 + \ldots + (y_n - \hat\beta_0 - \hat\beta_1 x_n)^2$$

- Obtaining the following estimates

$$\hat\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat\beta_0 = \bar{y} - \hat\beta_1 \bar{x}$$



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

Least square solution …

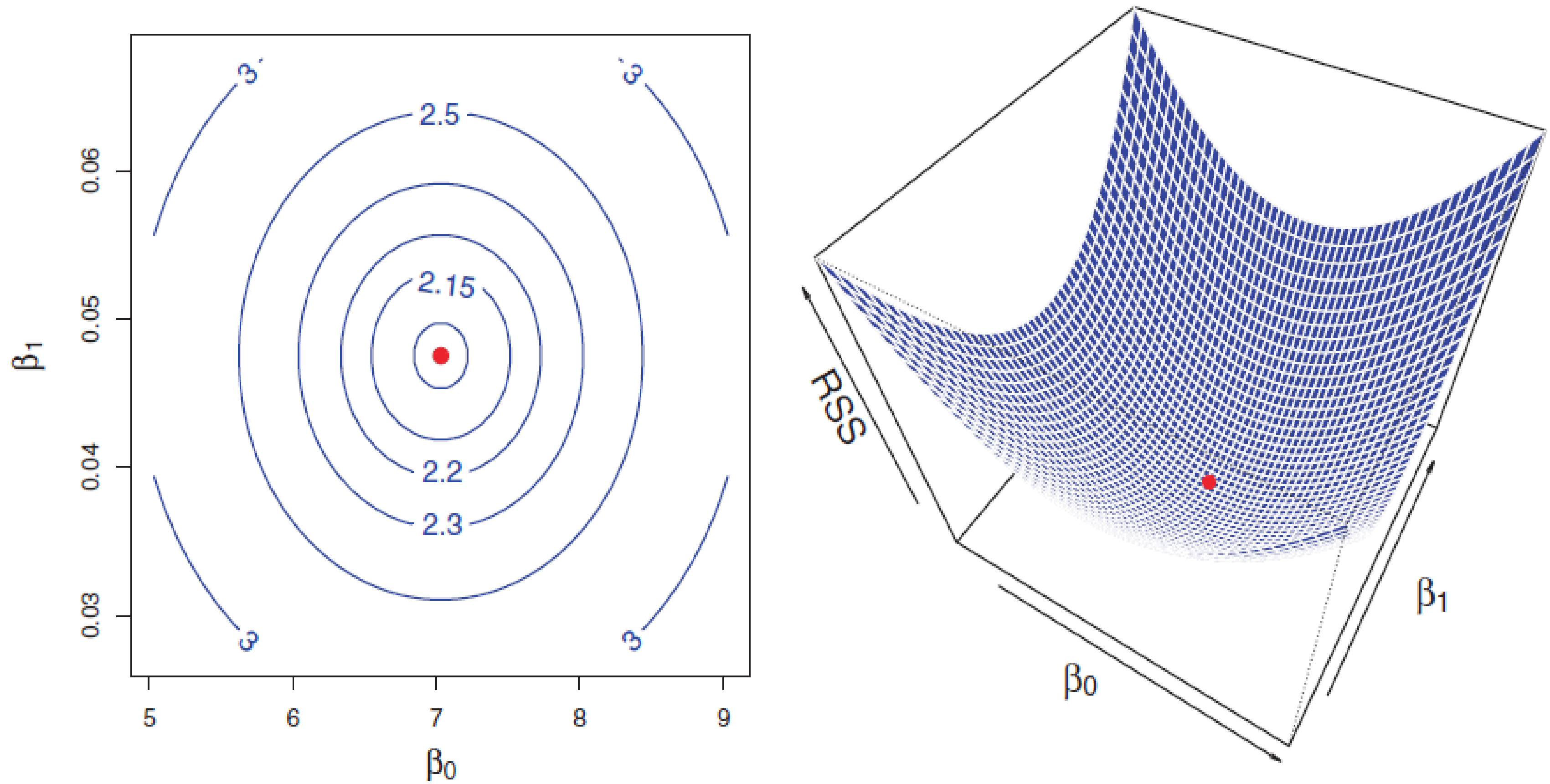


FIGURE 3.2. *Contour and three-dimensional plots of the RSS on the* `Advertising` *data, using* `sales` *as the response and* `TV` *as the predictor. The red dots correspond to the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, given by (3.4).*

Least squares fitting

- Least square fitting minimizes RSS (Residual Sum of Squares)

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

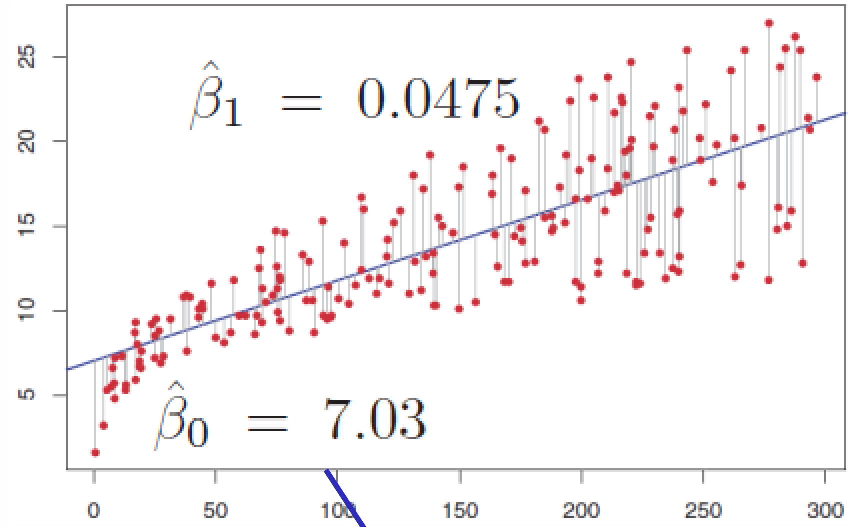
$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

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

- Obtaining the following estimates

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

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



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

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 squares regression line

“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

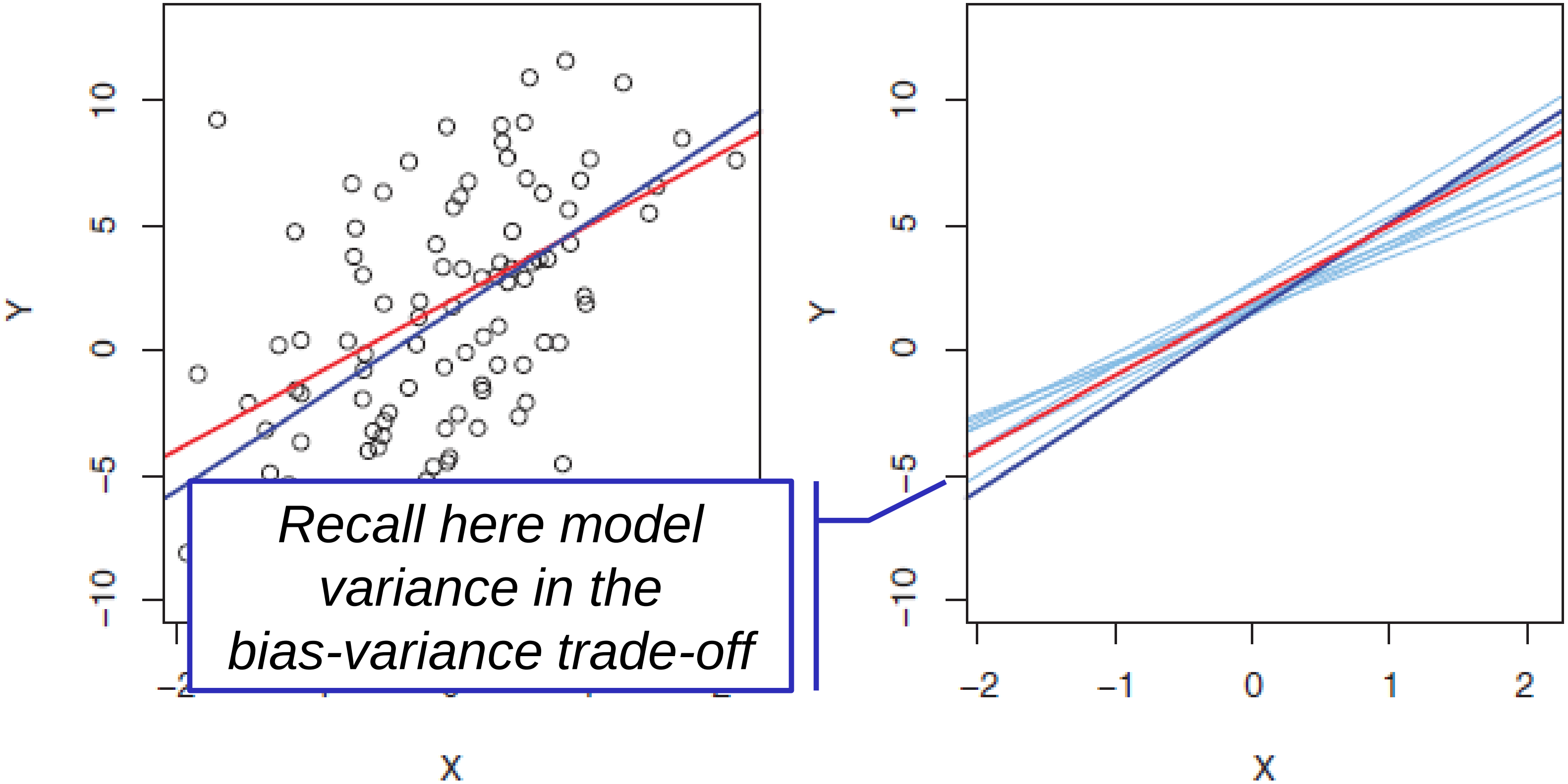


FIGURE 3.3. *A simulated data set.* Left: *The red line represents the true relationship, $f(X) = 2 + 3X$, which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for $f(X)$ based on the observed data, shown in black.* Right: *The population 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 can use similar formulae for the standard errors of the linear regression coefficients …

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

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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

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

This should be the 97.5 quantile of a t-distribution

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

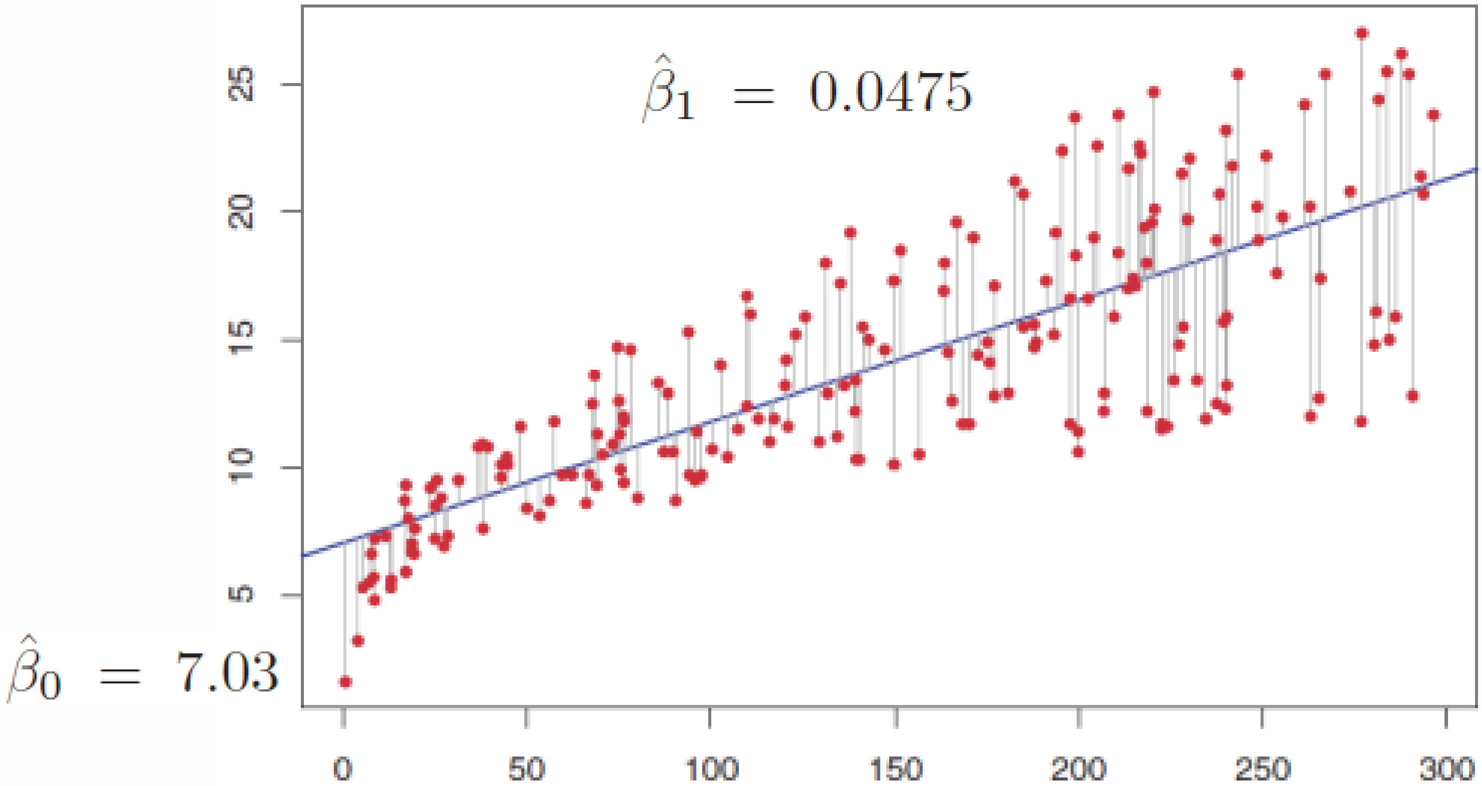
$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

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

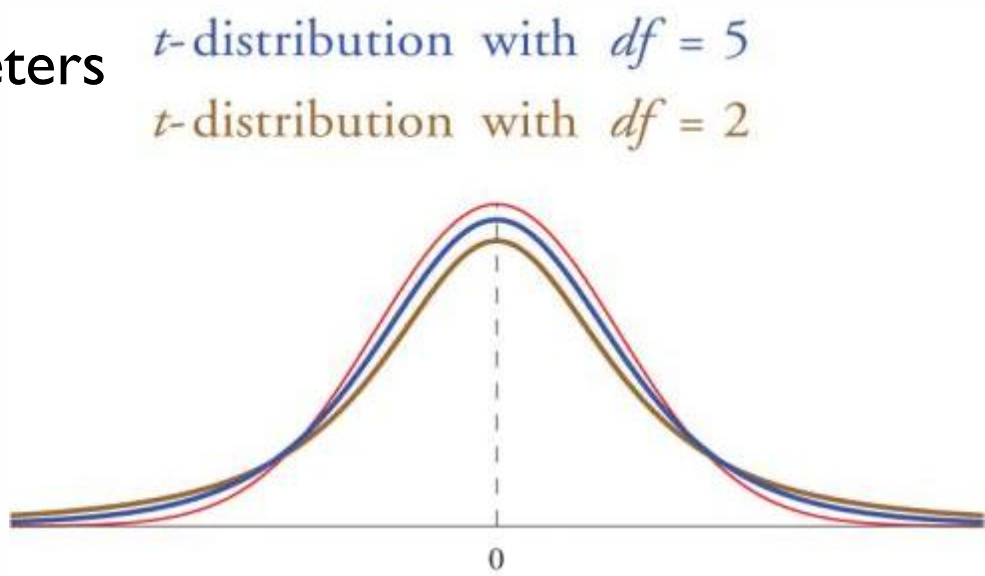
- Standard errors can be used for hypothesis testing such as:
  - H_0: there is no relationship between Y and X
  - H_a: there is some relationship between Y and X

- 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

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



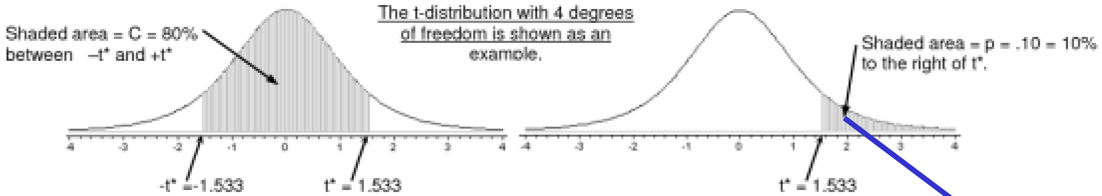


Table entry for p and C is the critical t* value with probability p lying to its right and probability C lying between −t* and t*.

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	$\boxed{1.533}$	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	2.971	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

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

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

  - 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

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

How far the model is from least square line on average

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

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")

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

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

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

- 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.*

R^2 vs Correlation

- Recall the definition of correlation between two variables, it is a measure of the (linear) relationship between them

$$\mathrm{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

- In the simple univariate linear regression setting we have

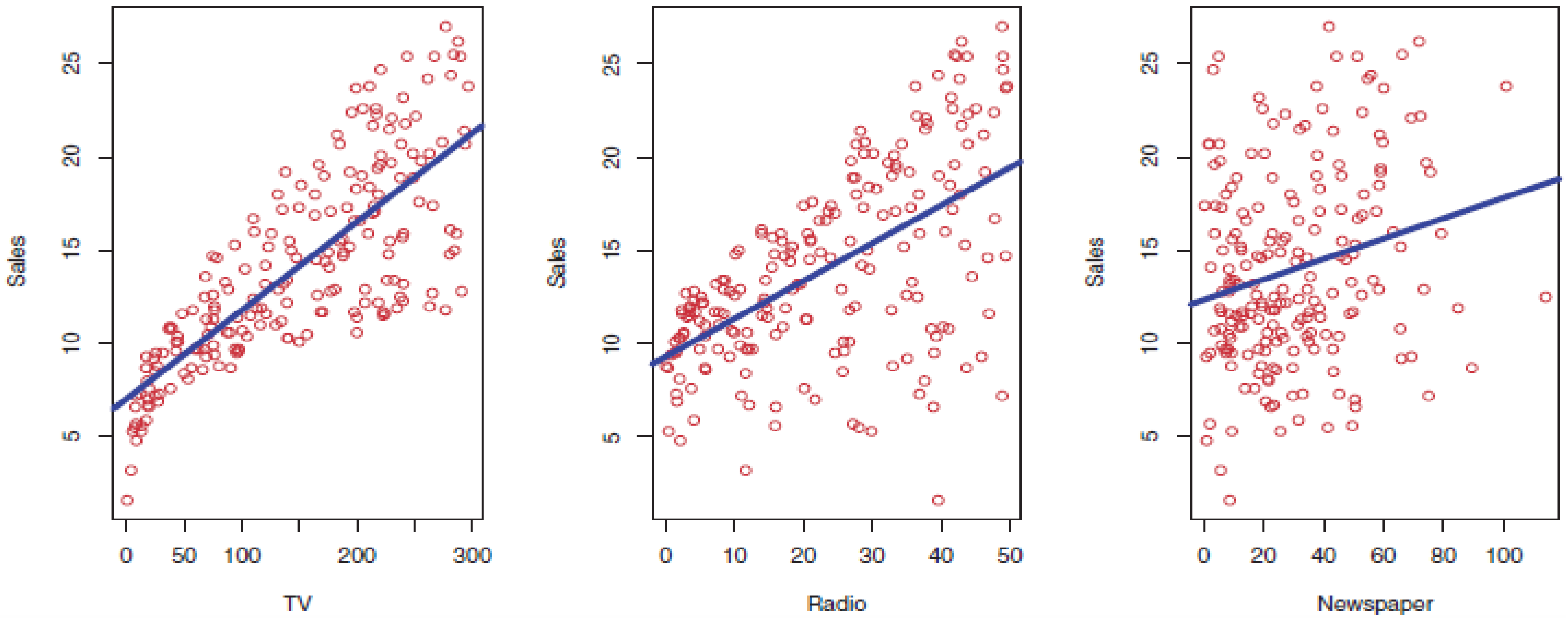
$$r = \mathrm{Cor}(X,Y)$$

$$R^2 = r^2$$

- In a multivariate case this does not holds and R^2 is used to extend the correlation concept to multiple variables

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

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?*

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 an increase in sales is obtained by changing the three input variables
  - the coefficients of each input did not take into account the other input in the estimate process
  - 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

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

Example: a two dimensional dataset

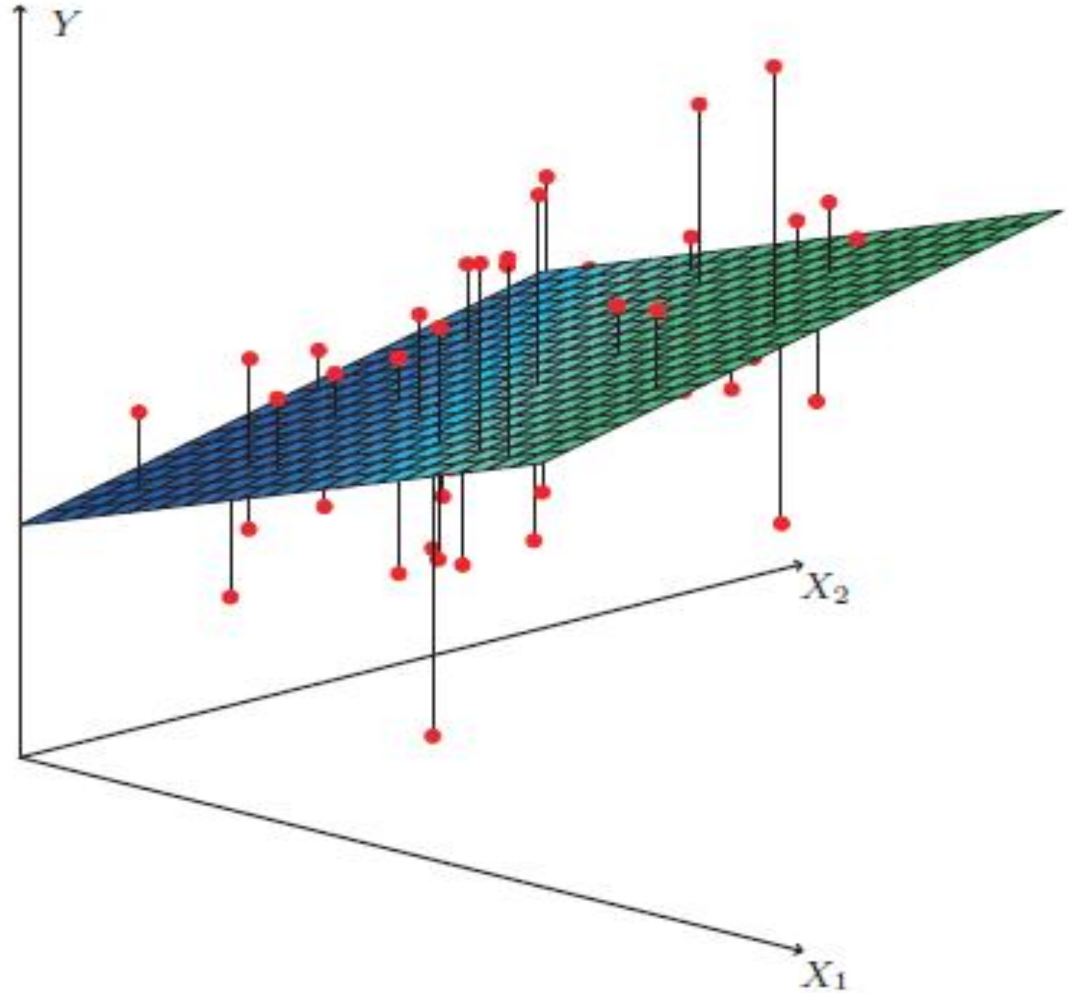


FIGURE 3.4. *In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.*

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

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

- 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	$\boxed{< 0.0001}$

	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	$\boxed{0.8599}$

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 correlated is some markets
  - If we increase *radio* then *newspaper* increases
- The increase on sales is correlate to the increase of newspaper is due to radio, not to the fact that newspaper increases sales
  - E.g., increase in sharks attacks are correlated to ice cream sales at the beach … because of people!

Computing linear regression coefficients (1)

- Computing the least regression fit can be done easily using linear algebra calculus
- Recall here
$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2$$
$$= \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$
- By taking into account that
  - **X** is an N x (p+1) data matrix
  - **y** is N x 1 vector of desired output
  - 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

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

$$\frac{\partial^2 \mathrm{RSS}}{\partial \beta \partial \beta^T} = 2\mathbf{X}^T\mathbf{X}$$

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

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

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

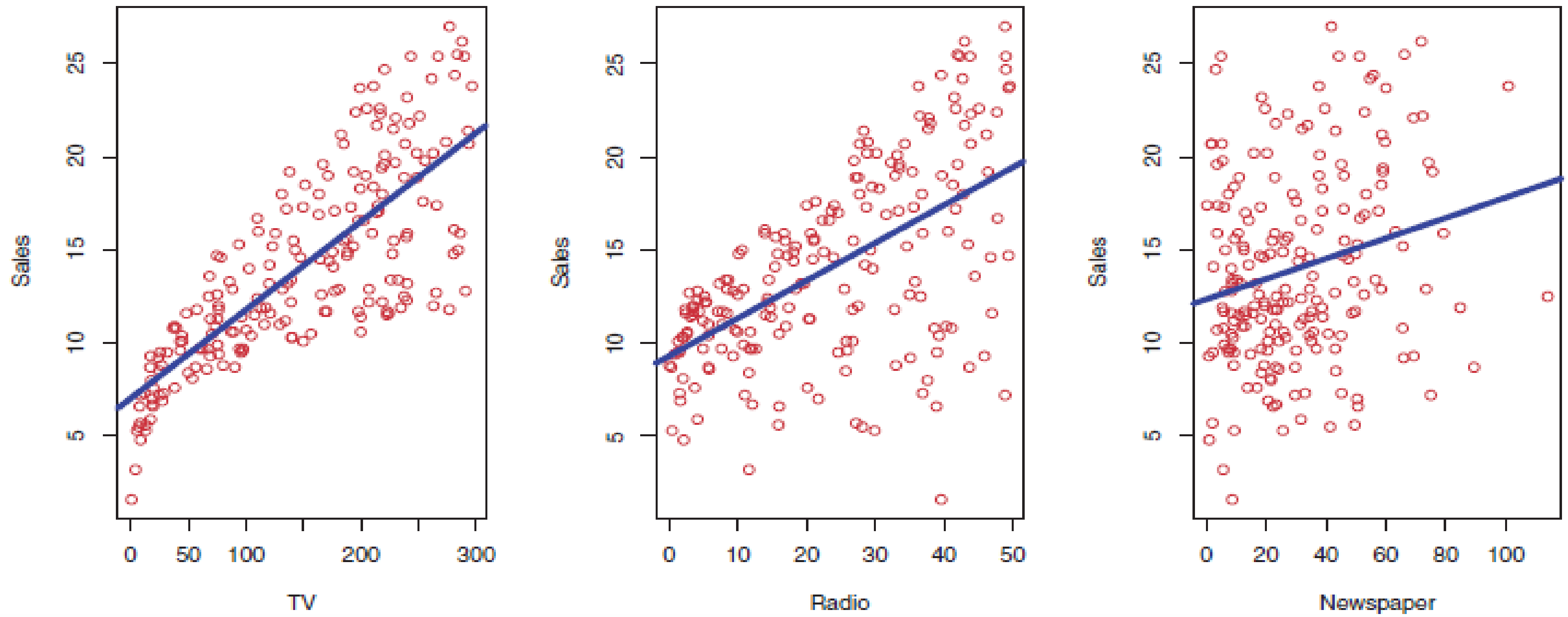
$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Pseudo Inverse

- In matrix algebra terms the prediction becomes

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

Rephrase the questions



- *Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting the response?*
- *Do all the predictors help to explain Y, or it is only a subset of the predictors useful*
- *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?

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

against

$$H_a : \text{ at least one } \beta_j \text{ is non-zero}$$

- This test is performed using the F-statistics

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

- If the linear model assumptions are valid

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

  - when H_0 is true $E\{(\text{TSS} - \text{RSS})/p\} = \sigma^2$ $F \sim 1$
  - 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 $\alpha = .05$

	df_1=1	2	3	4	5	6	7	8	9	10
df_2=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	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	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	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	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	[illegible]	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45

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

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

- 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.*

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 =
  - Y =
  - Y =
  - Y =

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

Practical methods for variable selection

- Forward selection starts with the null model
  - Fit *p* simple linear regressions and select the lower RSS
  - Fit *p-1* bivariate linear models and select the lower RSS
  - …
- Backward selection starts with the full model
  - Remove the variable with the smallest p-value
  - Retrain the (p-1) variables model
  - …
- Mixed selection
  - Alternate forward selection
  - Perform some backward selection from time to time

Model fit in multiple variable selection

- As for the simple linear regression R^2 and *RSE* are used
- In multiple linear regression R^2 equals $\mathrm{Cor}(Y, \hat{Y})^2$
- An R^2 close to 1 means that the model explains a large amount of variance in the data
- The more variables in the model the

Removing newspaper this goes from 0.8972 to 0.89719

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

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.*

- In a multi variable scenario RSE becomes $\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{RSS}}.$

Data visualization

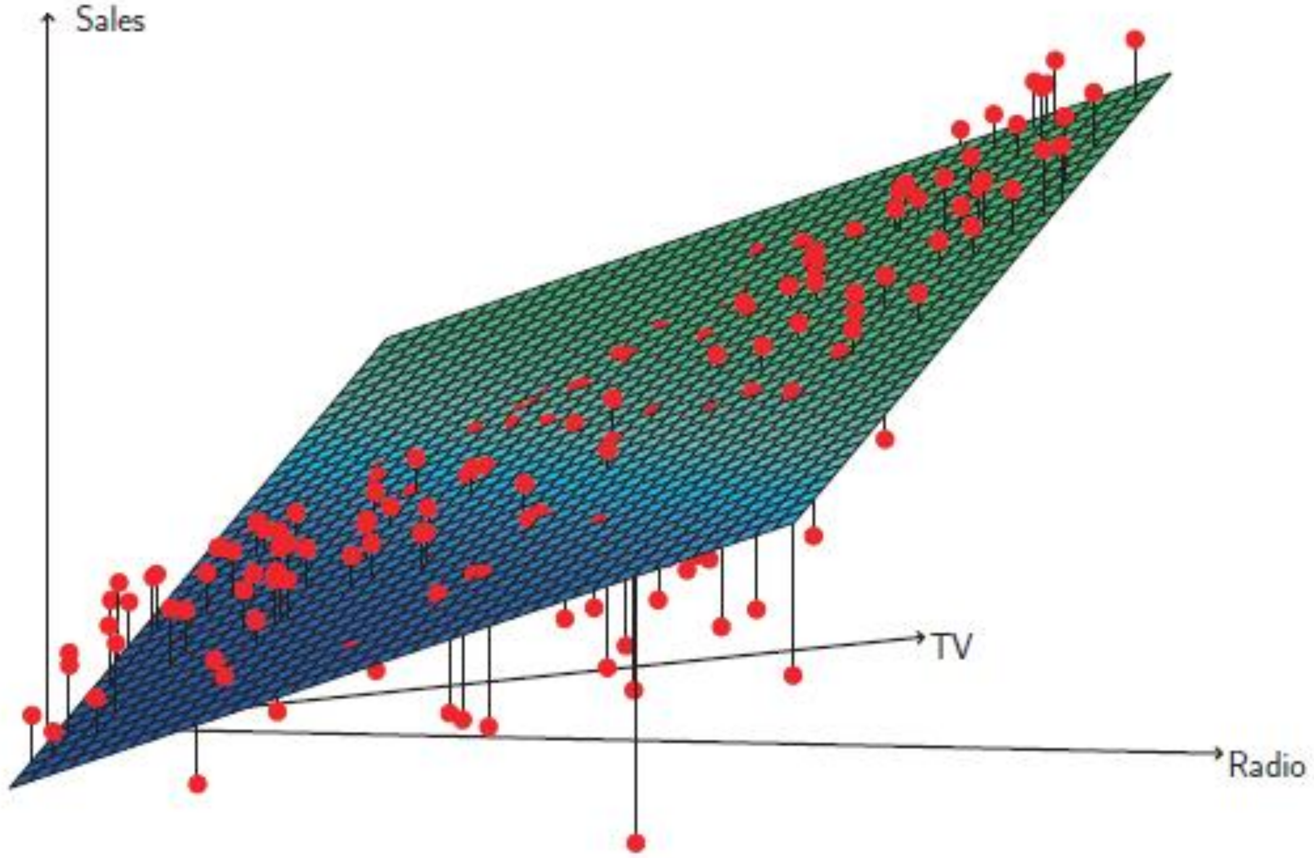


FIGURE 3.5. *For the* Advertising *data, a linear regression fit to* sales *using* TV *and* radio *as predictors. From the pattern of the residuals, we can see that there is a pronounced non-linear relationship in the data. The positive residuals (those visible above the surface), tend to lie along the 45-degree line, where TV and Radio budgets are split evenly. The negative residuals (most not visible), tend to lie away from this line, where budgets are more lopsided.*

Prediction uncertainty

- The least squares regression estimated from the data

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

can differ from the population regression plane (model variance)

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Moreover a linear model might not be enough to capture all the systematic components of the data (model bias)

- Finally the irreducible error is there …

- Also in the case of multivariate linear regression we can compute the confidence intervals for the plane coefficients

Example: Credit Predictors

- In the Credit dataset
  - 7 quantitative
  - 4 qualitative
- Quantitative ones
  - Gender
  - Student (status)
  - Status (marital)
  - Ethnicity
    - Caucasian
    - African American
    - Asian

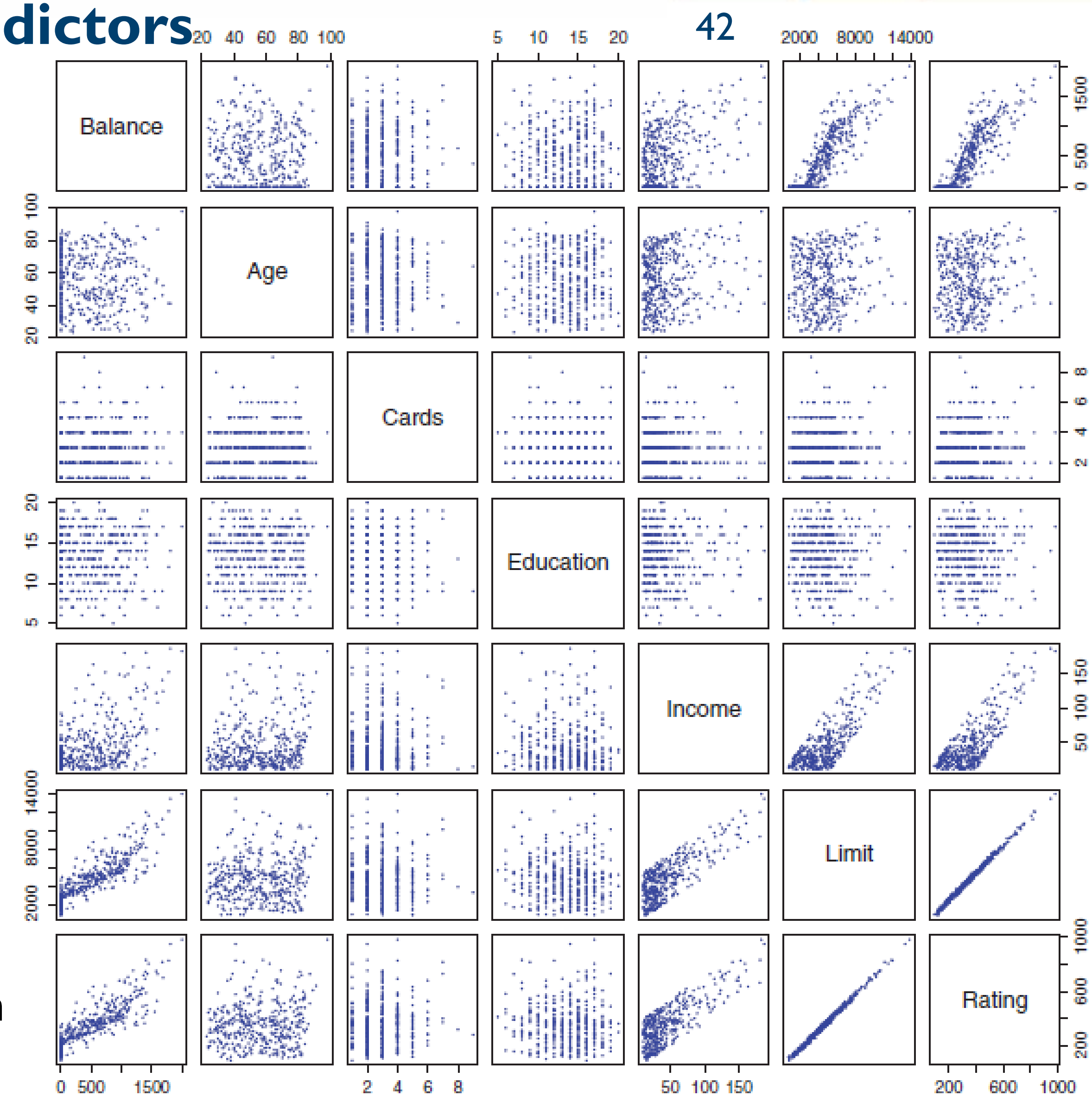


FIGURE 3.6. *The* Credit *data set contains information about* balance, age, cards, education, income, limit, *and* rating *for a number of potential customers.*

- Quantitative predictors with Two Levels 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 balance among males

Average difference of balance between males and females

	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 salary females are above the average and males are below …

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

Quantitative 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" model

Average difference between African Americans and Asian

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

Average balance for African American

Average difference between African Americans and Caucasians

Quantitative 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 results in a "multiple output" model

The non coded level is defined *baseline*

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

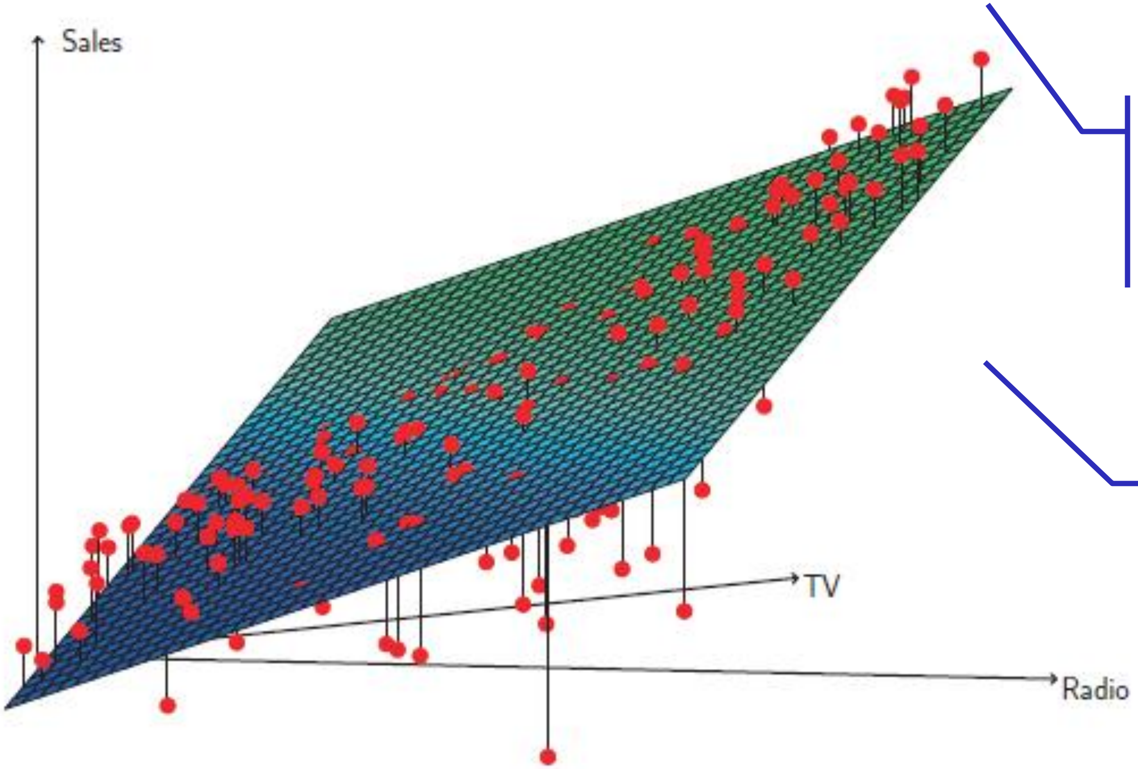
- F-statistics should be used instead of p-values

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

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



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

This effect in marketing is known as *sinergy*

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"

One variable affects other variables influence

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

Example: Interaction between TV and Radio

- We can imaging some interaction in TV and Radio Advertising

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

  - The interaction term is the increase of effectiveness of TV advertising for one unit of Radio advertising
  - 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

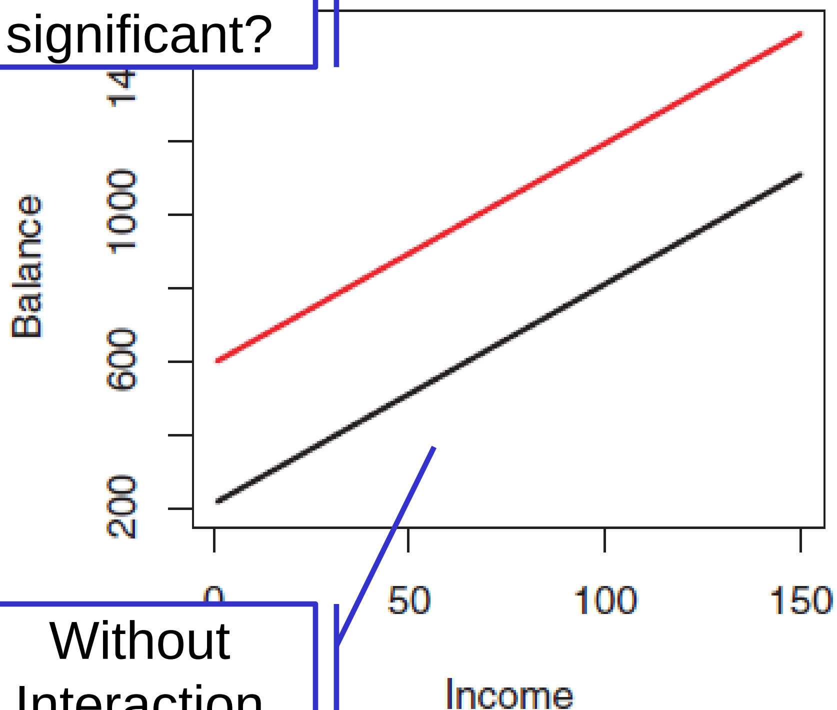
TABLE 3.9. *For the* Advertising *data, least squares coefficient estimates associated with the regression of* sales *onto* TV *and* radio*, with an interaction term, as in (3.33).*

Interaction in Qualitative Predictors

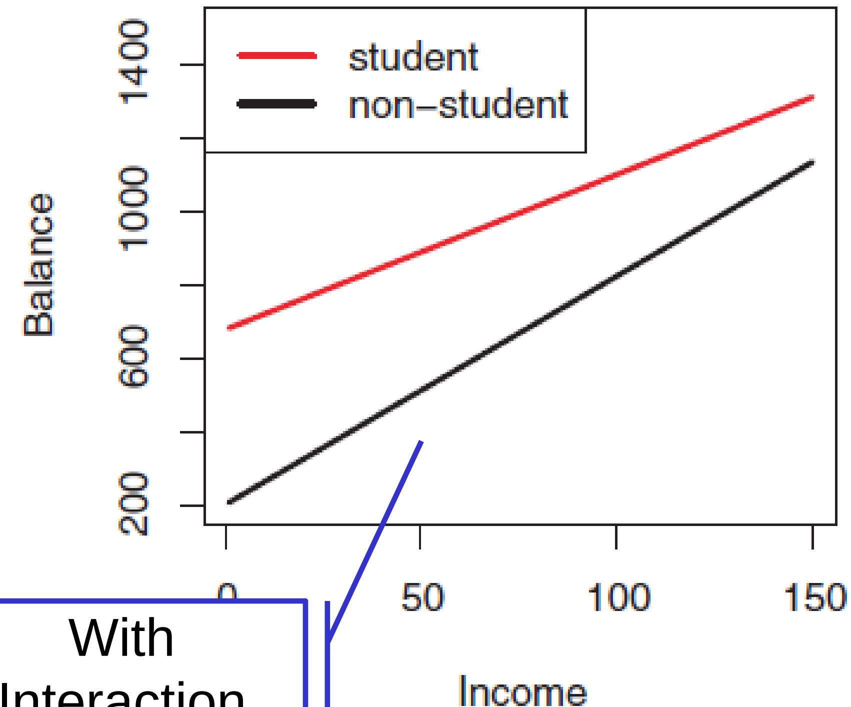
- We can check for interactions with qualitative predictors

$$
\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\
&= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases}
\end{aligned}
$$

Is this term significant?



Without Interaction



With Interaction

Example: MpG Polynomial Regression

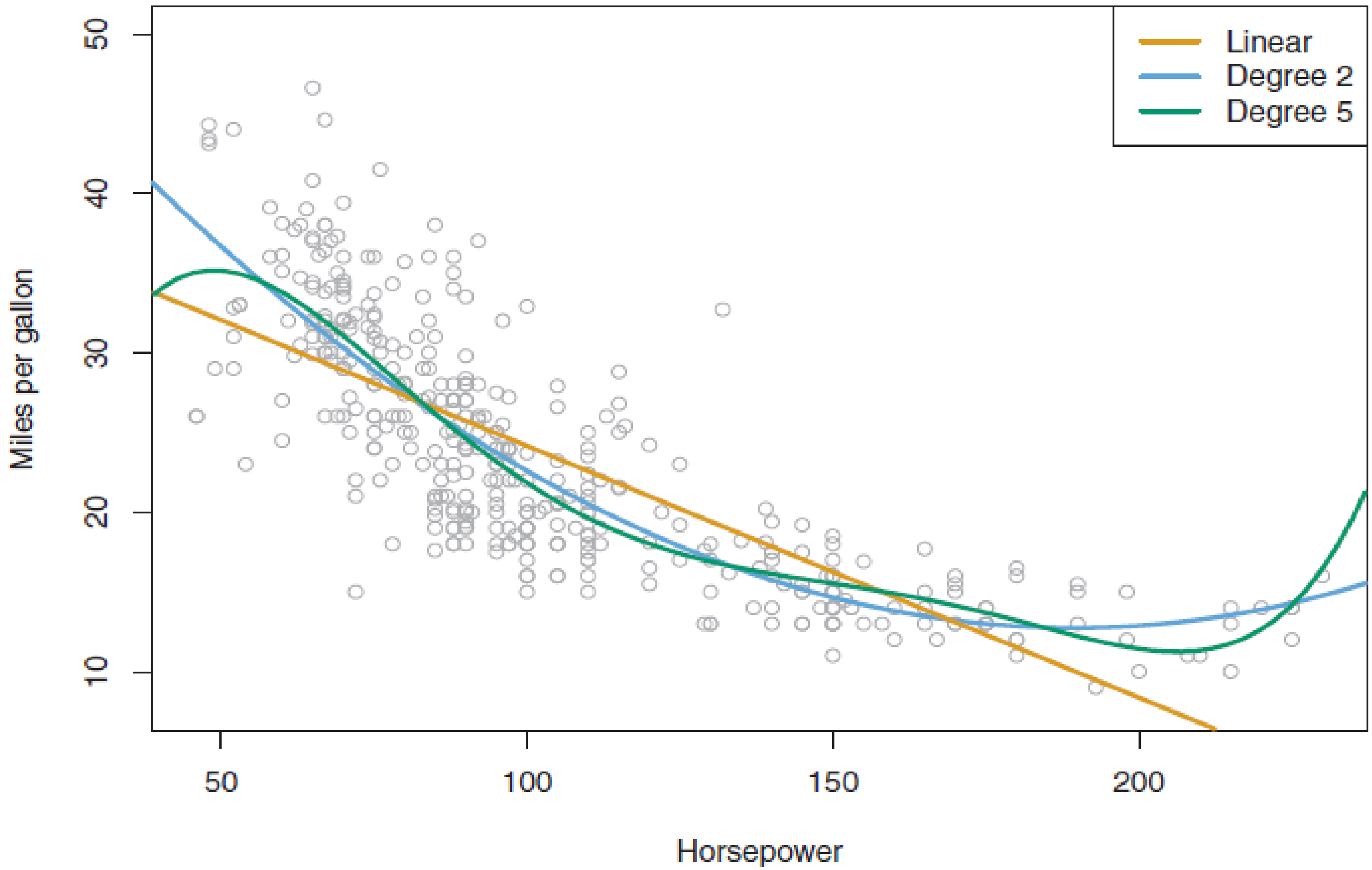


FIGURE 3.8. *The* Auto *data set. For a number of cars,* mpg *and* horsepower *are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes* horsepower^2 *is shown as a blue curve. The linear regression fit for a model that includes all polynomials of*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

- We can use Polynomial Regression to accomodate non linearity

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

  - It is still a linear fitting problem !!!!
  - A 5th grade polynome seems too much, but the 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
$\mathtt{horsepower}^2$	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* $\mathtt{horsepower}^2$.

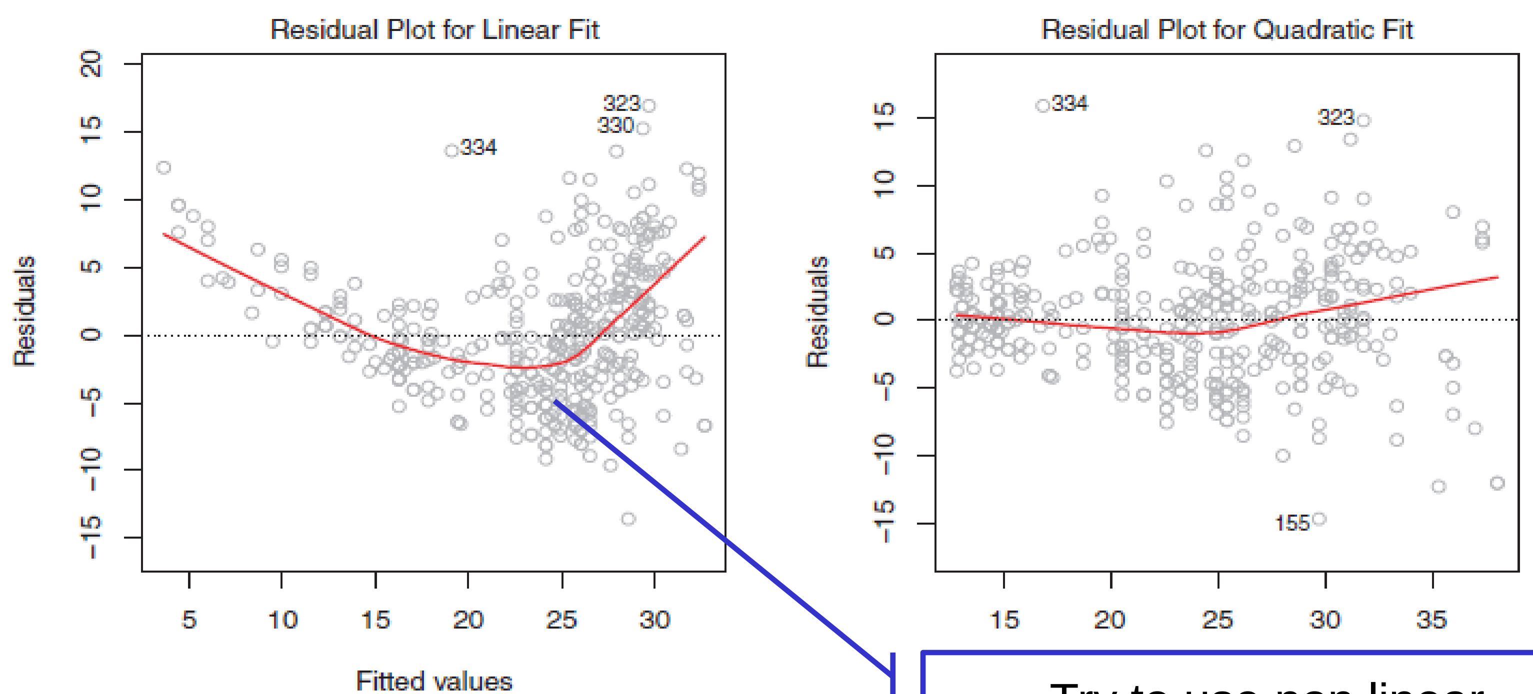
Potential Problems in Linear Regression

- 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, the conclusions drawn might be (at least) inaccurate … check residual plot!



Try to use non linear transformations of the predictor (log X, X2, sqrt(X), …)

FIGURE 3.9. *Plots of residuals versus predict… data set. In each plot, the red line is a smooth fit t… 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* horsepower^2*. There is little pattern in the residuals.*

Dependence of the Error Term

- Errors are supposed to uncorrelated otherwise standard errors would underestimate the true errors … tracking phenomenon

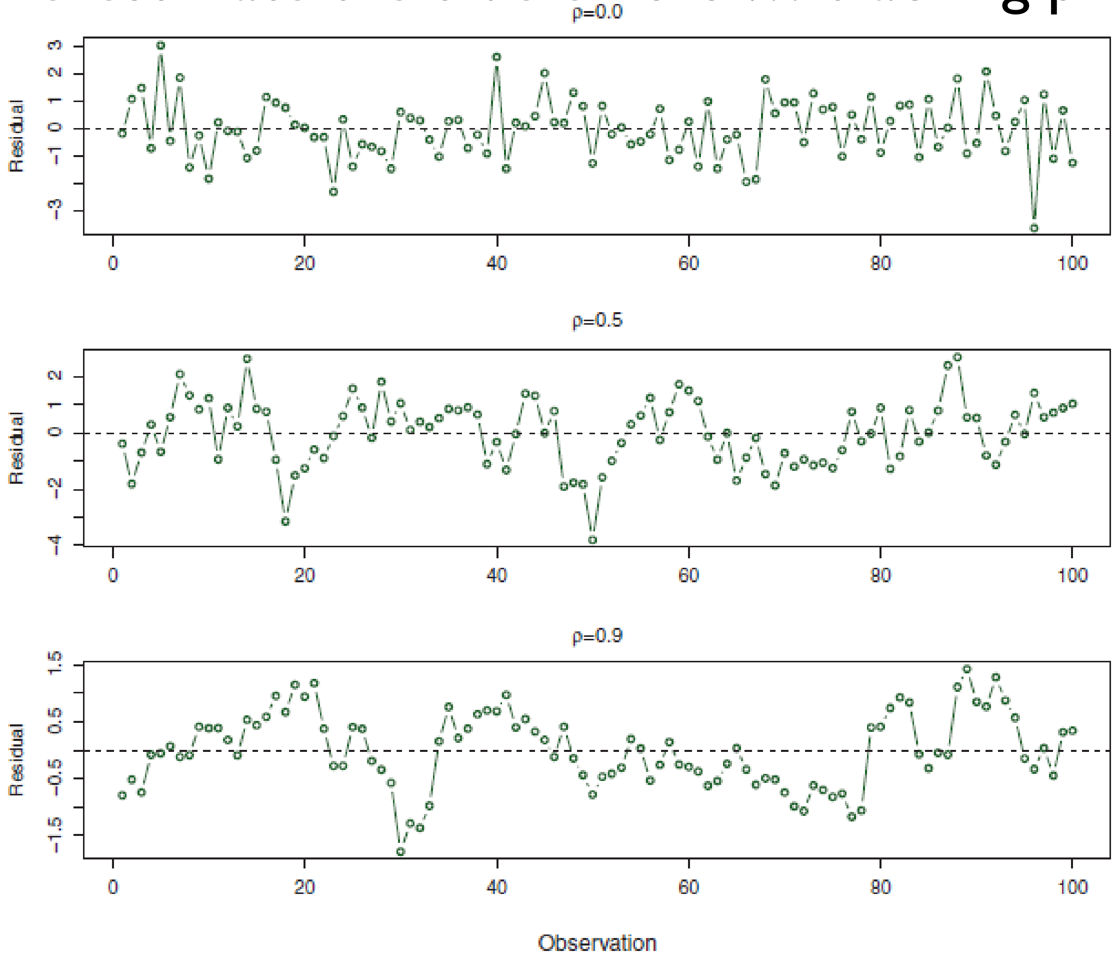


FIGURE 3.10. *Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.*

Non Constant Variance of Error Term

- Linear Regression assumes no heteroscedasticity in the noise

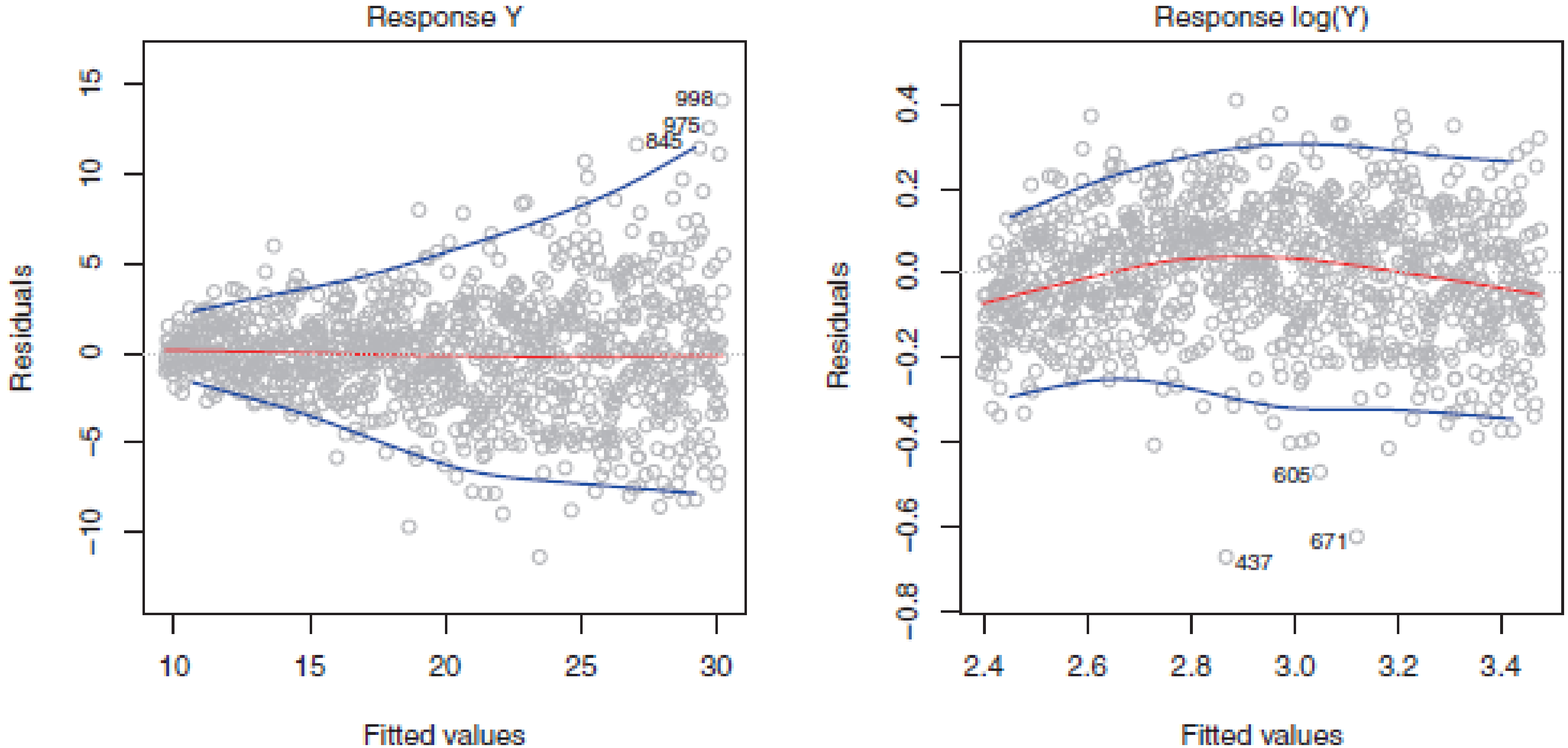


FIGURE 3.11. *Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns. Left: The funnel shape indicates heteroscedasticity. Right: The predictor has been 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)

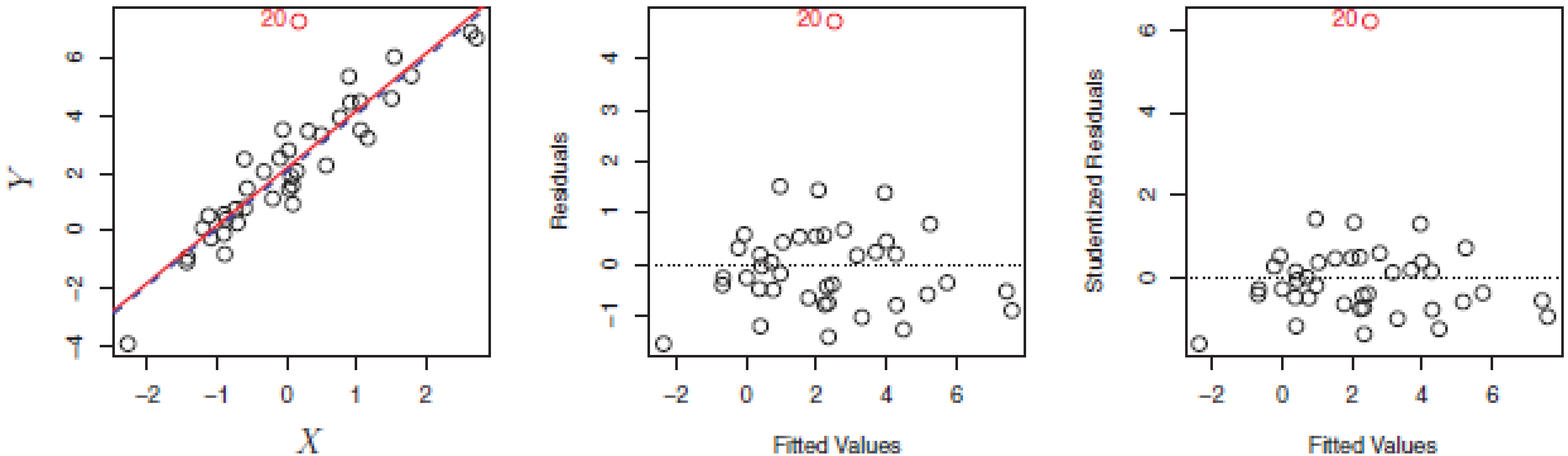


FIGURE 3.12. Left: *The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue.* Center: *The residual plot clearly identifies the outlier.* Right: *The outlier has a studentized residual of 6; typically we expect values between −3 and 3.*

High Leverage Points

- High leverage points have unexpected values for a predictor

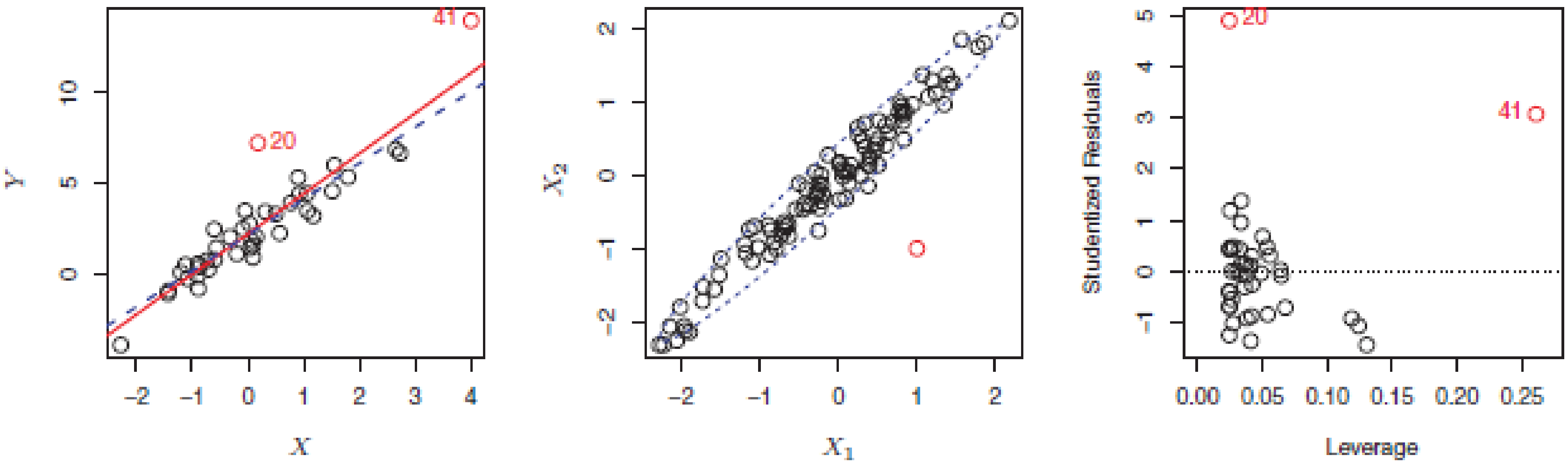


FIGURE 3.13. Left: *Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed.* Center: *The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of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

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

Colinearity

- Some factors might be highly related so it might be difficult to separate their effects

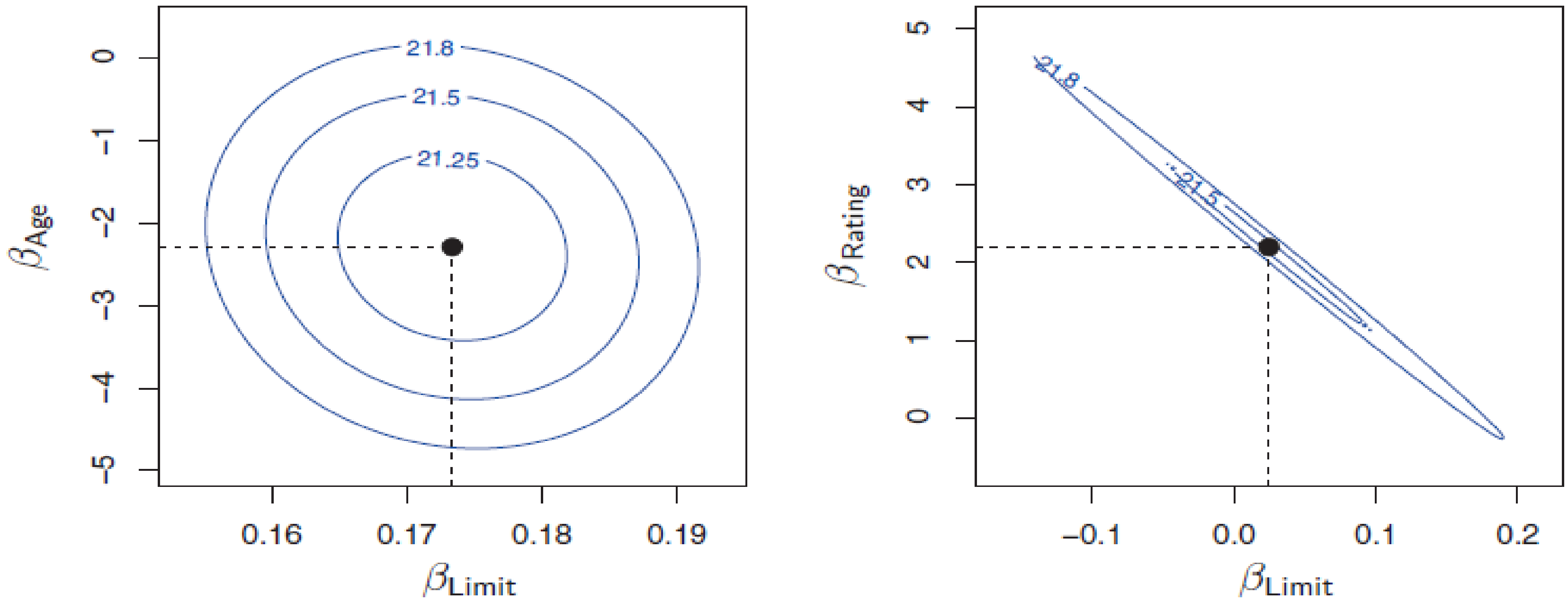
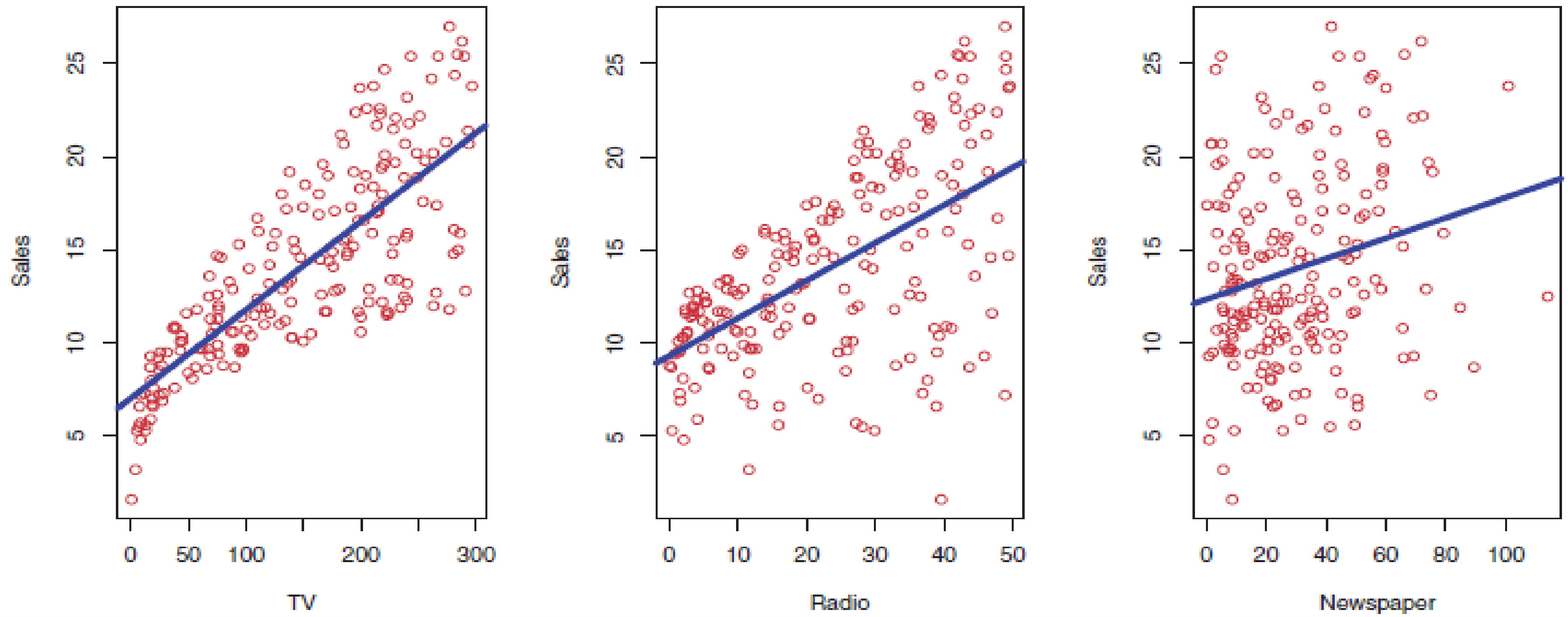


FIGURE 3.15. *Contour plots for the RSS values as a function of the parameters β for various regressions involving the* `Credit` *data set. In each plot, the black dots represent the coefficient values corresponding to the minimum RSS.* Left: *A contour plot of RSS for the regression of* `balance` *onto* `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?*