Lecture 18: Linear Regression: OLS, Ridge, LASSO
Setup and Practical Considerations

Foundations of Data Science: Algorithms and Mathematical Foundations

Mihai Cucuringu
mihai.cucuringu@stats.ox.ac.uk

CDT in Mathematics of Random System
University of Oxford

28 September, 2023

Advertising data set

- sales of a product in 200 different markets
- + budgets for the product in each of those markets for three different media: TV, radio, and newspaper

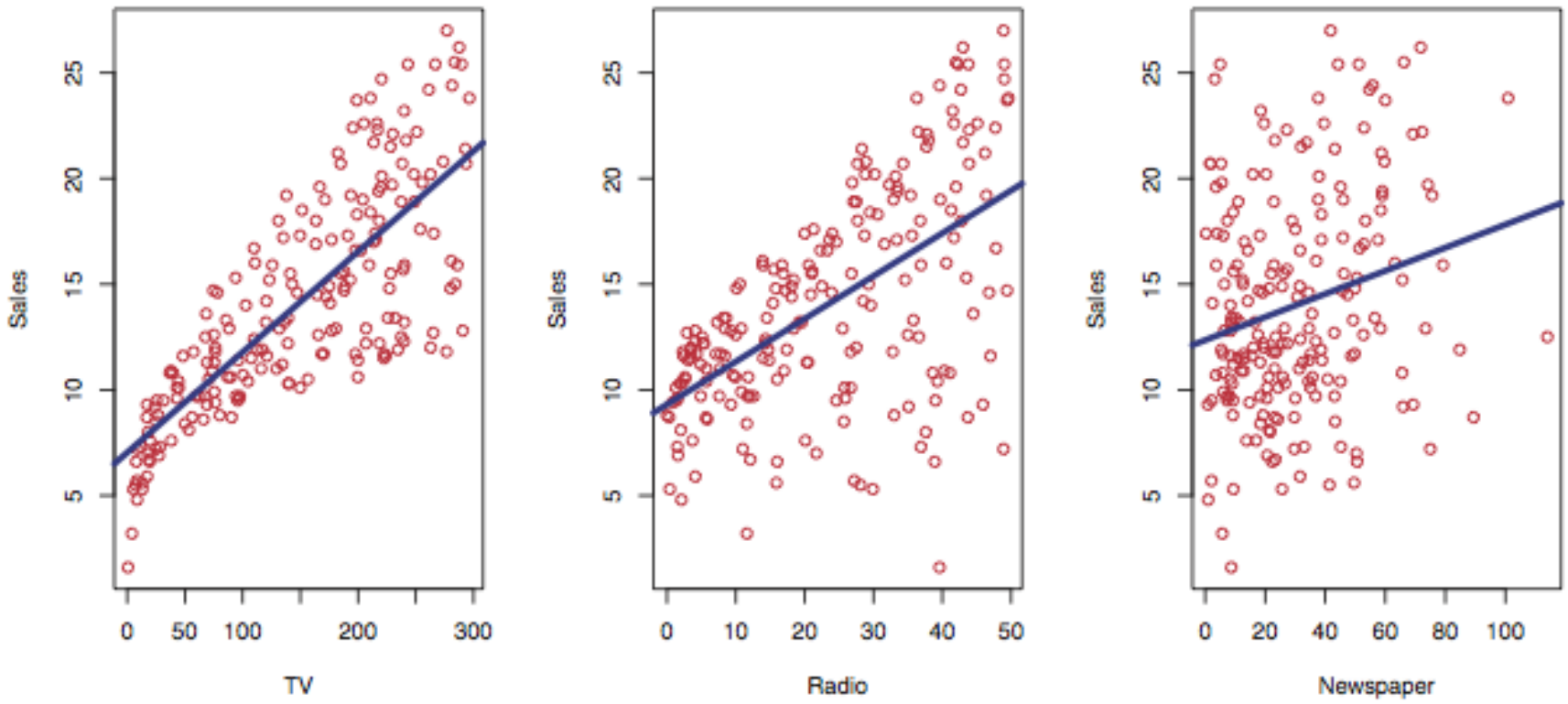
- goal: predict sales given the three media budgets

- input variables (denoted by $X_1, X_2, \ldots$)
  - X_1 TV budget
  - X_2 radio budget
  - X_3 newspaper budget

- inputs known as such as *predictors, independent variables, features, variables, covariates...*

- the output variable (sales) is the response or dependent variable (denoted by Y)

Advertising data set



Linear Regression

- 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? (50k on TV + 50k on radio > 100k on either one) (interaction effect)

Errors

Model: $Y \approx \beta_0 + \beta_1 X$

Example: sales $\approx \beta_0 + \beta_1 \times$ radio

Define the residual sum of squares (RSS):

$$\text{RSS} = \sum_{i=1}^{n} \epsilon_i^2 \tag{1}$$

$$\epsilon_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad \forall i = 1, \ldots, n \tag{2}$$

Least squares fit

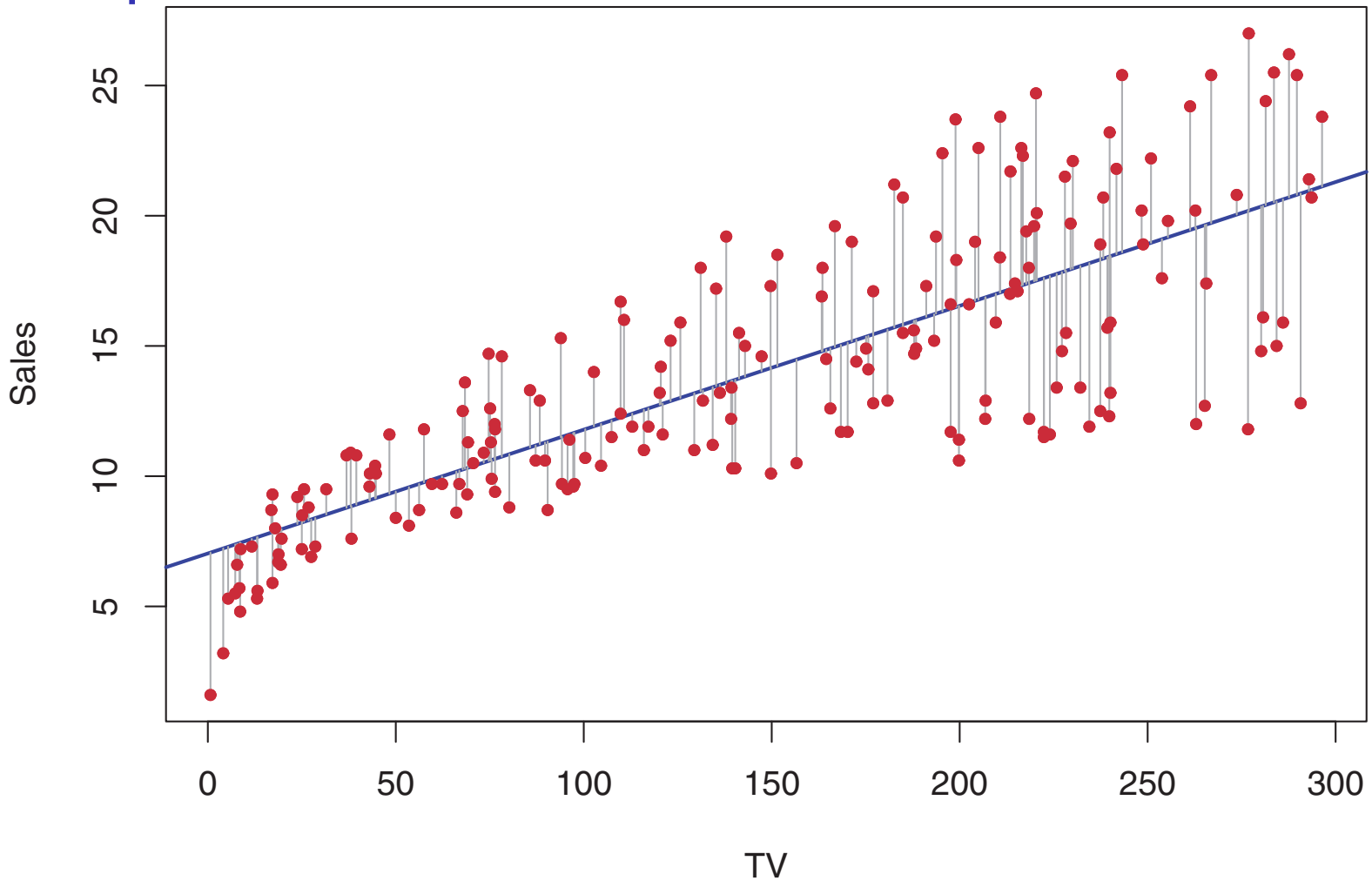


Figure: The least squares fit for the regression of sales onto TV. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case, a linear fit captures the essence of the relationship, although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

Least squares fit

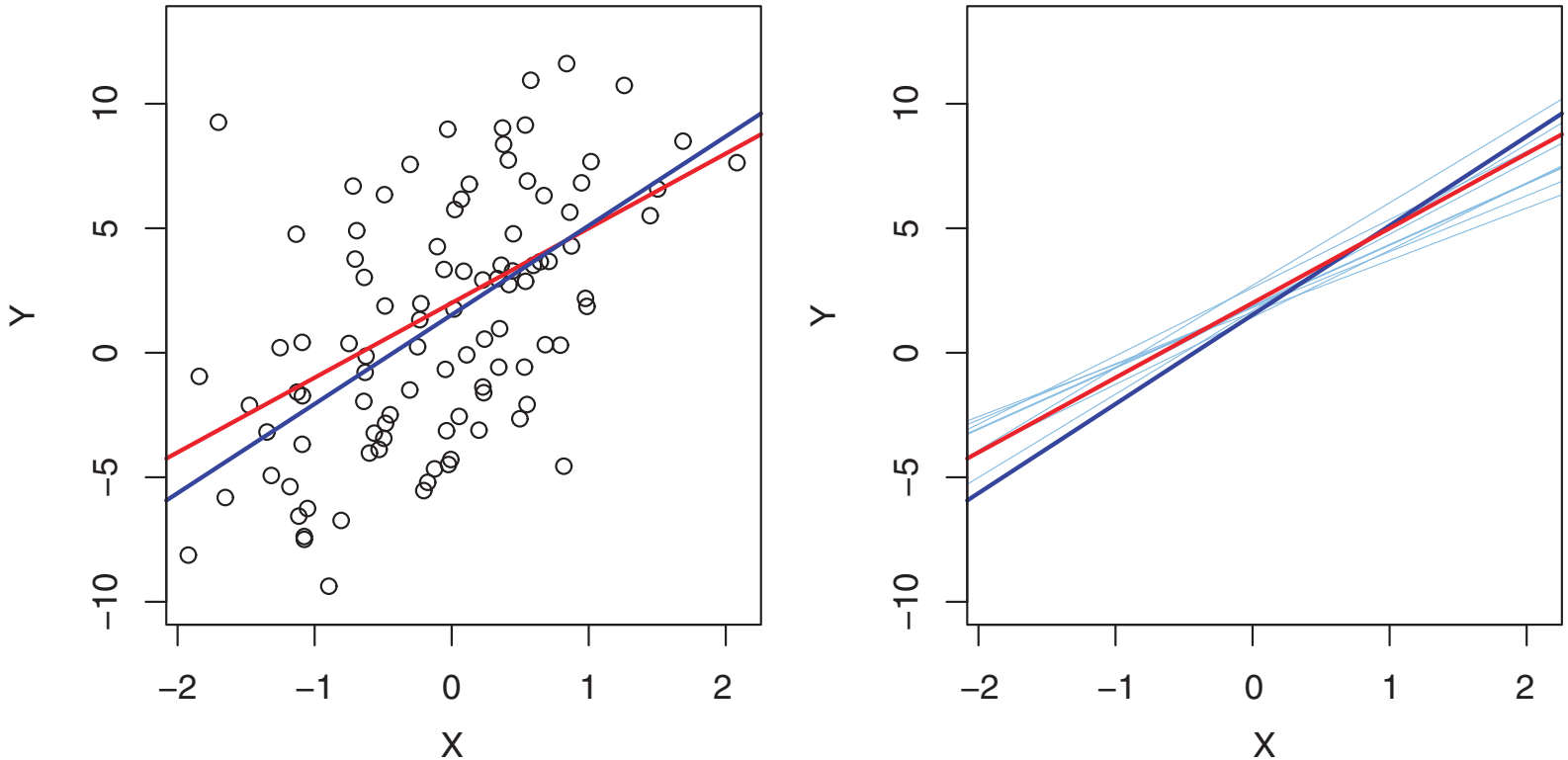


Figure: A simulated data set. **Left**: The red line represents the true relationship, $f(X) = 2 + 3X$, which is known as the population regression line. The blue line is the least squares line: it is the least squares estimate for $f(X)$ based on the observed data, shown in black. **Right**: The population 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.

Recall the OLS estimators

The least squares coefficient estimates for simple linear regression

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{4}$$

where $\bar{y} = \frac{1}{n}\sum_i^n y_i$ and $\bar{x} = \frac{1}{n}\sum_i^n x_i$ denote the sample means.

The corresponding standard errors are given by

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \tag{5}$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6}$$

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

Confidence intervals

- 95 % confidence interval for β_1

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

- i.e., 95 % prob. the β_1 lies in

$$[\hat{\beta}_1 - 2 \cdot \text{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \text{SE}(\hat{\beta}_1)$$

- similarly for β_0

Advertising data, the 95% confidence interval

- $\beta_0 \in [6.130, 7.935]$: without any advertising $\Rightarrow$ sales will situate around 6,130 and 7,940 units.

- $\beta_1 \in [0.042, 0.053]$: each \$1,000 increase in TV advertising $\Rightarrow$ average increase in sales by between 42 and 53 units.

Hypothesis testing: the null hypothesis

H_0 : There is no relationship between X and Y

$$\beta_1 = 0$$

H_1 : There is some relationship between X and Y

$$\beta_1 \neq 0$$

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

- compute the t-statistic given by

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

  i.e., the number of standard deviations $\hat{\beta}_1$ is away from 0
- if no relationship between X and Y, t ~ t-distribution with n-2 degrees of freedom
- for $n > 30$, t-distribution is similar to the Gaussian

Hypothesis testing: the null hypothesis

- p-value: probability of observing any value equal to $|t|$ or larger, assuming $\beta_1 = 0$
- Small p-value: unlikely to observe such a substantial association between *X* and *Y* due to chance, (if X and Y were truly unrelated)
- Typical p-values for rejecting the null hypothesis: 5% or 1%

	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).*

Quality metrics

Recall:

- TSS= $\sum(y_i - \bar{y})^2$, the total variance in the response Y
- RSS= $\sum(y_i - \hat{y}_i)^2$, the amount of variability that is left unexplained after the regression

Quality metrics

- RSE: measures lack of fit of the model to the data

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

- R^2: measures the proportion of variance explained

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

- for simple linear regression: $R^2 = \rho^2$, where ρ is the usual Pearson correlation

From Simple to Multiple Linear Regression

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

Figure: A \$1,000 increase in radio spending $\Rightarrow$ an average increase in sales by 203 units. A \$1,000 increase in newspaper spending $\Rightarrow$ an average increase in sales by around 55 units.

Multiple Linear Regression

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

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

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \cdots + \hat{\beta}_p x_{i,p} \quad \forall i = 1, \ldots, n$$

$$\textit{sales} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{radio} + \beta_3 \text{newspaper} + \epsilon$$

Errors being minimized

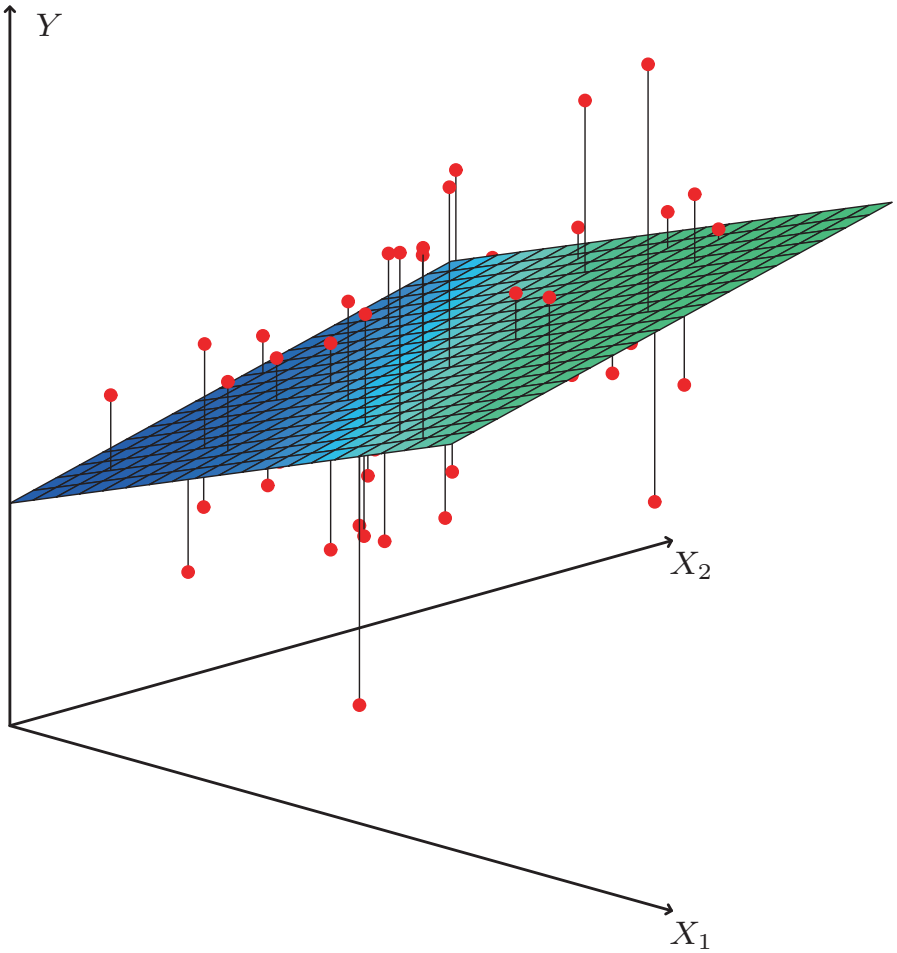


Figure: In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.

Multiple Linear Regression

	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.*

Fixing TV and newspaper advertising, spending an additional \$1,000 on radio $\Rightarrow$ sales increase 189 units

Note $\beta_{newspaper}$ is now very close to zero, with a small t-statistic and p-value.

	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.*

- corr(radio,newspaper) = 0.35
- newspaper gets ”credit” for the effect of radio on sales
- shark attacks vs ice cream sales at a given beach shows a positive relationship
- higher temperatures $\Rightarrow$ more people visit the beach $\Rightarrow$ more ice cream sales and more shark attacks
- ice cream no longer significant after adjusting for temperature

Variable selection

Which predictors are associated with the response? (in order to fit a single model involving only those d predictors)

- Note: R^2 always increase as you add more variables to the model
- adjusted R^2: $1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$
- Mallow's: $C_p = \frac{1}{n}(\text{RSS} + 2p\hat{\sigma}^2)$
- Akaike Information criterion AIC $= \frac{1}{n\hat{\sigma}^2}(\text{RSS} + 2p\hat{\sigma}^2)$

Cannot consider all 2^p models...

- Best Subset Selection: fit a separate least squares regression for each possible k-combination of the p predictors, and select the best one
- Forward selection: start with the null model and keep adding predictors one by one
- Backward selection: start with all variables in the model, and remove the variable with the largest p-value

Other considerations (see the textbook)

- prediction intervals
- extensions of the linear model

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

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

• R^2 for this model 96.8% vs 89.7% for the model that uses TV and radio without an interaction term.
• The hierarchical principle: if we include $X \times Y$, you should also include the main effects X and Y (even if their p-values are not significant)

- Non-linear Relationships

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

Potential Problems with Linear Regression

- Non-linearity of the response-predictor relationships
- Correlation of error terms
- Non-constant variance of error terms
- Outliers
- High-leverage points
- Collinearity

(1) Non-linearity of the Data

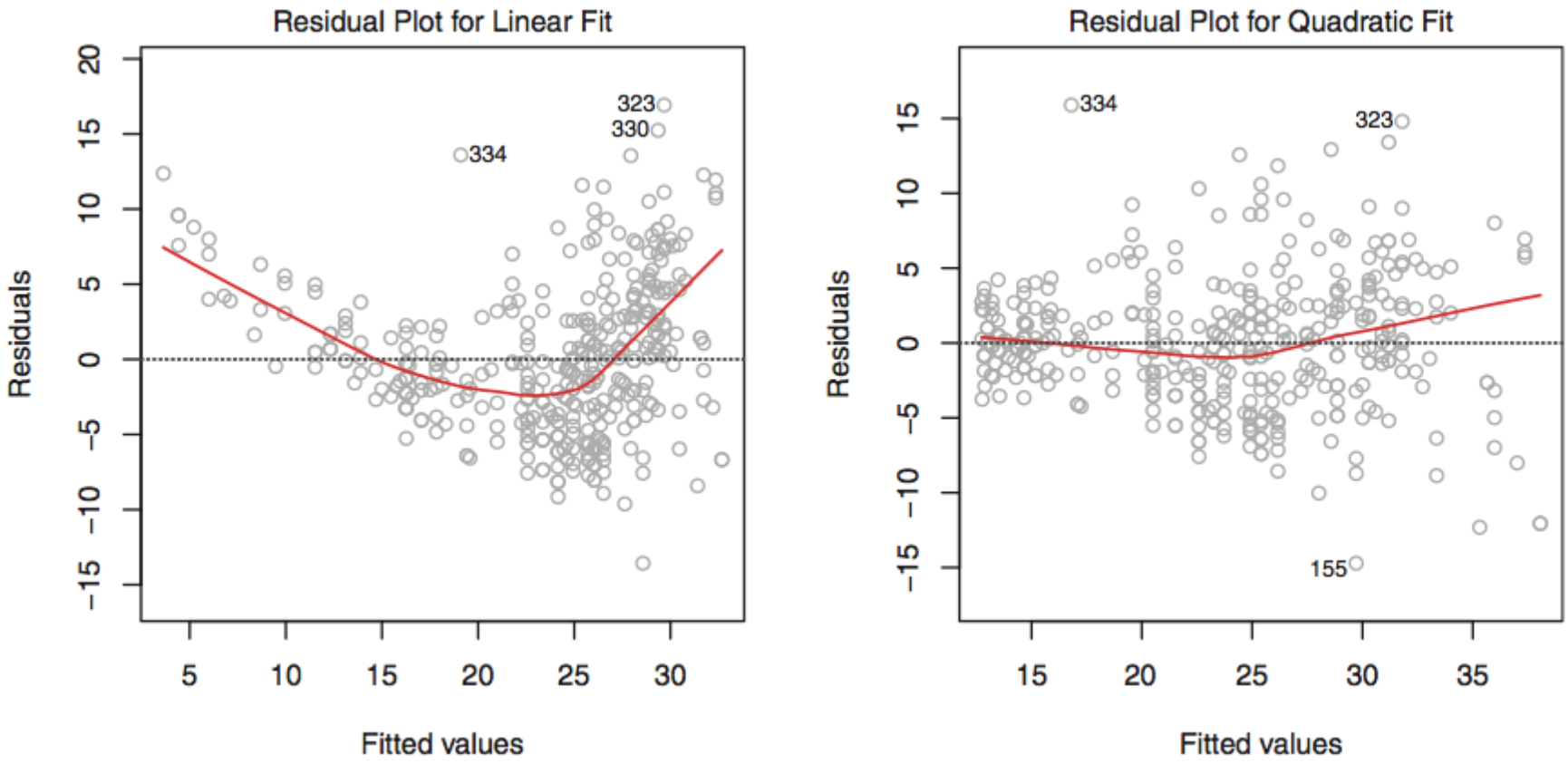


Figure: Residuals vs. predicted (or fitted) values for the Auto data set. In each plot, the red line is a smooth fit to the residuals. Left: $Y \sim X$, Right: $Y \sim X^2$.

(2) Time series of residuals - (Correlation of error terms)

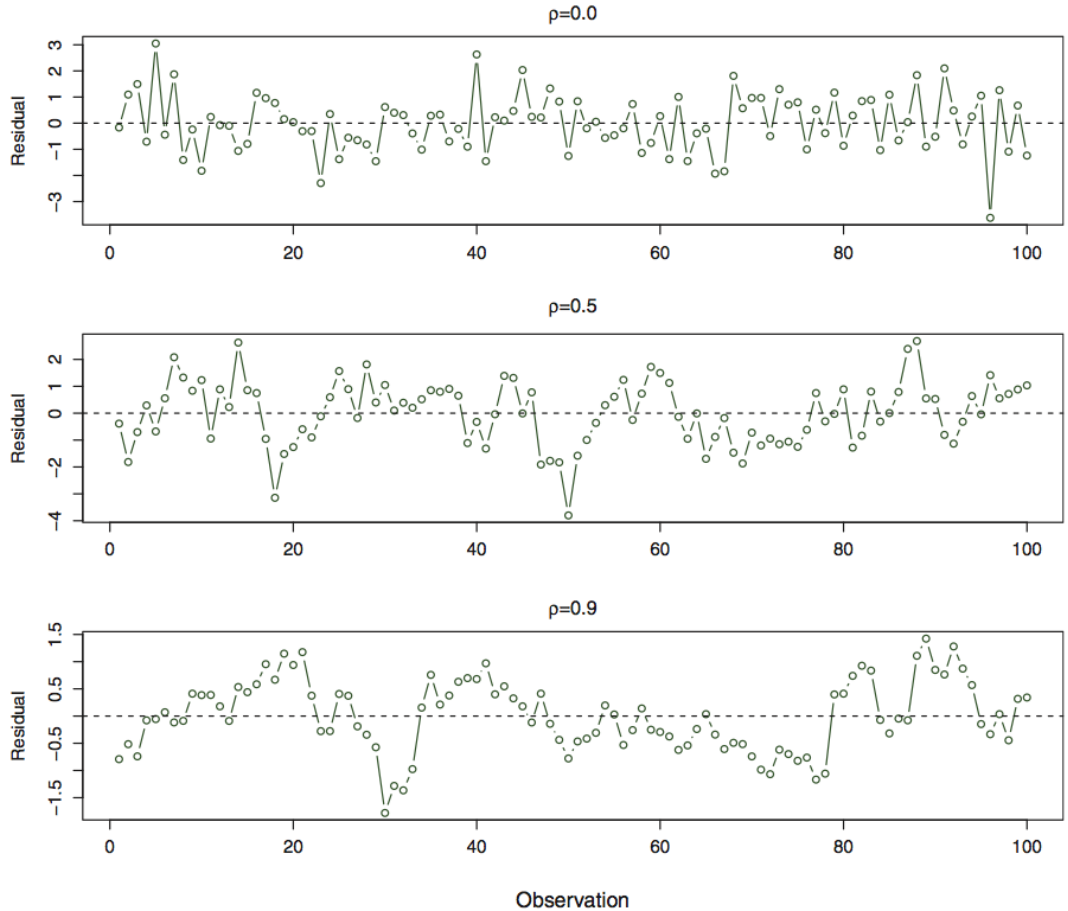


FIGURE 3.10. *Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.*

Figure: Plots of residuals from simulated time series data sets generated with differing levels of correlation $\rho = \{0$ (top), 0.5 (middle), 0.9 (bottom) between error terms for adjacent time points.

- See the Newey–West estimator, for handling autocorrelation (serial correlation), and heteroskedasticity in the error terms.

(3) Residual plots - (Non-constant variance of error terms)

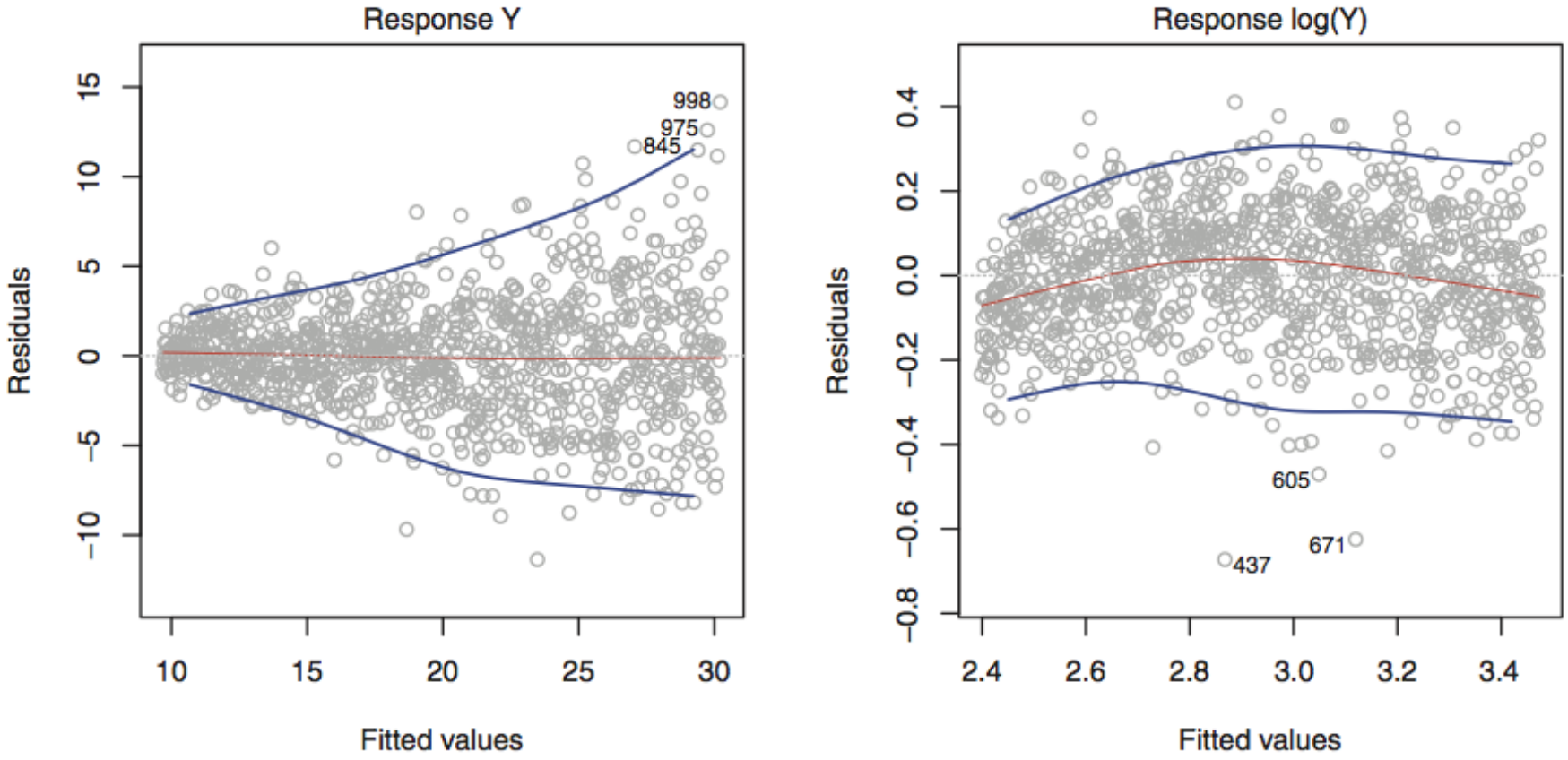


Figure: Red line: smooth fit to the residuals. Blue lines: track the outer quantiles of the residuals. Left: The funnel shape indicates heteroscedasticity (variance of the errors is not constant). Right: The predictor has been log-transformed $\Rightarrow$ no evidence of heteroscedasticity.

Read the entire Chapter 3 in ISLR.