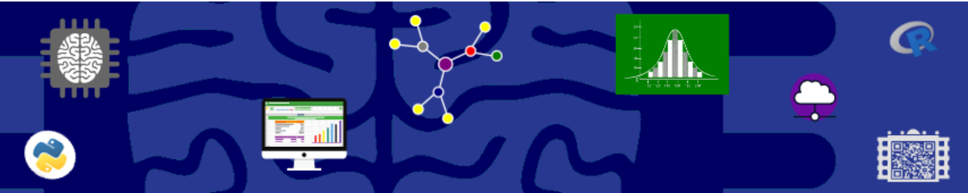
Introduction to Data Science

Winter Semester 2023/24

Oliver Ernst

TU Chemnitz, Fakultät für Mathematik, Professur Numerische Mathematik

Lecture Slides



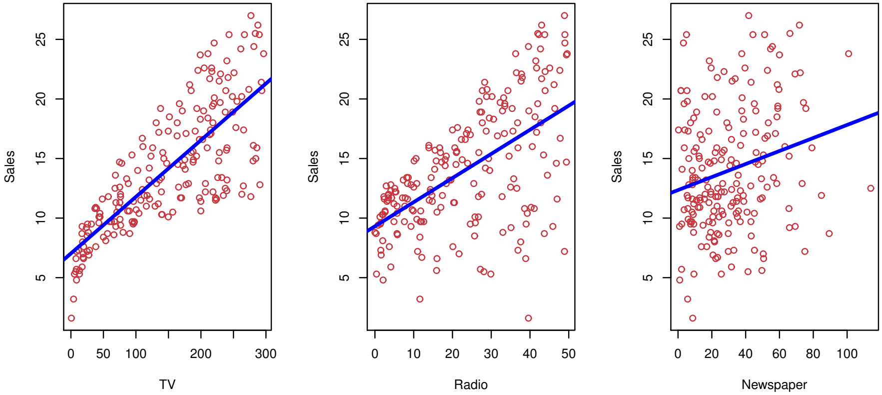
Contents I

Contents II

Contents III

Contents

Recall advertising data set from Slide 30:



We will use the simple and well-established statistical learning technique known as **linear regression** to answer the following questions:

1. Is there a relationship between advertising budget and sales?
   Otherwise, why bother?

Linear Regression

Questions about advertising data set

1. Is there a relationship between advertising budget and sales?
   Otherwise, why bother?
2. How strong is this relationship between advertising budget and sales?
   Prediction possibly better than random guess?

1. Is there a relationship between advertising budget and sales?
   Otherwise, why bother?
2. How strong is this relationship between advertising budget and sales?
   Prediction possibly better than random guess?
3. Which media contribute to sales?
   Separate individual contributions

1. Is there a relationship between advertising budget and sales?
   Otherwise, why bother?
2. How strong is this relationship between advertising budget and sales?
   Prediction possibly better than random guess?
3. Which media contribute to sales?
   Separate individual contributions
4. How accurately can we estimate the effect of each medium on sales?
   Euro by Euro?

Linear Regression

Questions about advertising data set

1. Is there a relationship between advertising budget and sales?
   Otherwise, why bother?
2. How strong is this relationship between advertising budget and sales?
   Prediction possibly better than random guess?
3. Which media contribute to sales?
   Separate individual contributions
4. How accurately can we estimate the effect of each medium on sales?
   Euro by Euro?
5. How accurately can we predict future sales?
   Precise prediction for each medium?

Linear Regression

Questions about advertising data set

1. Is there a relationship between advertising budget and sales?
   Otherwise, why bother?
2. How strong is this relationship between advertising budget and sales?
   Prediction possibly better than random guess?
3. Which media contribute to sales?
   Separate individual contributions
4. How accurately can we estimate the effect of each medium on sales?
   Euro by Euro?
5. How accurately can we predict future sales?
   Precise prediction for each medium?
6. Is the relationship linear?
   If yes, linear regression appropriate (possibly after transforming data)

1. Is there a relationship between advertising budget and sales?
   Otherwise, why bother?
2. How strong is this relationship between advertising budget and sales?
   Prediction possibly better than random guess?
3. Which media contribute to sales?
   Separate individual contributions
4. How accurately can we estimate the effect of each medium on sales?
   Euro by Euro?
5. How accurately can we predict future sales?
   Precise prediction for each medium?
6. Is the relationship linear?
   If yes, linear regression appropriate (possibly after transforming data)
7. Is there synergy among the advertising media?
   Called ***interaction effect*** *in statistics.*

Contents

Linear model for quantitative response Y of single predictor X:

$$Y \approx \beta_0 + \beta_1 X. \qquad (3.1)$$

Statistician: "We are regressing Y onto X."

Simple Linear Regression

Definition, terminology, notation

Linear model for quantitative response Y of single predictor X:

$$Y \approx \beta_0 + \beta_1 X. \tag{3.1}$$

Statistician: "We are regressing Y onto X."

E.g., with predictor `TV` advertising and response `sales`,

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

Simple Linear Regression

Definition, terminology, notation

Linear model for quantitative response Y of single predictor X:

$$Y \approx \beta_0 + \beta_1 X. \tag{3.1}$$

Statistician: "We are regressing Y onto X."

E.g., with predictor `TV` advertising and response `sales`,

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

The values of **coefficients** or **parameters** β_0, β_1 obtained from fitting to the training data are denoted by $\hat{\beta}_0, \hat{\beta}_1$, leading to the prediction values

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{3.2}$$

when $X = x$, where the hat on $\hat{y}$ denotes the predicted value of the reponse.

The **intercept** $\hat{\beta}_0$ and **slope** $\hat{\beta}_1$ in (3.1) are determined choosing these parameters such that the **residuals** or **data misfits**

$$r_i := y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \qquad i = 1, \dots, n,$$

are minimized.

The **intercept** $\hat{\beta}_0$ and **slope** $\hat{\beta}_1$ in (3.1) are determined choosing these parameters such that the **residuals** or **data misfits**

$$r_i := y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \qquad i = 1, \dots, n,$$

are minimized.

There are many options for defining smallness here, in **least squares estimation** this is measured by the **residual sum of squares (RSS)**

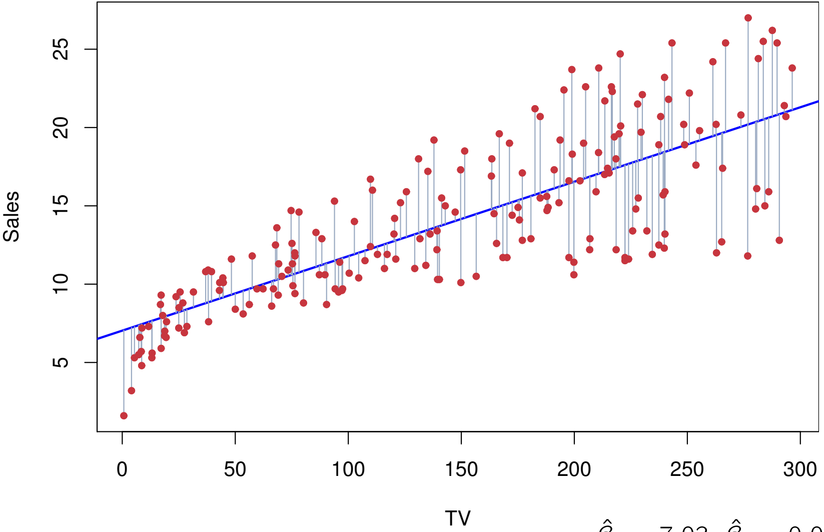
$$\text{RSS} := r_1^2 + \cdots + r_n^2 = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2. \tag{3.3}$$

An easy calculation reveals

$$\begin{aligned} \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x}, & \overline{x} &:= \frac{1}{n}\sum_{i=1}^n x_i, \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}, & \overline{y} &:= \frac{1}{n}\sum_{i=1}^n y_i. \end{aligned} \tag{3.4}$$

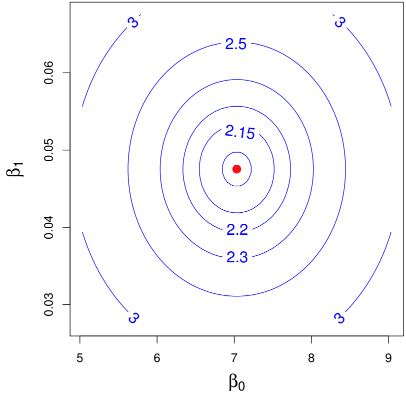
Simple Linear Regression

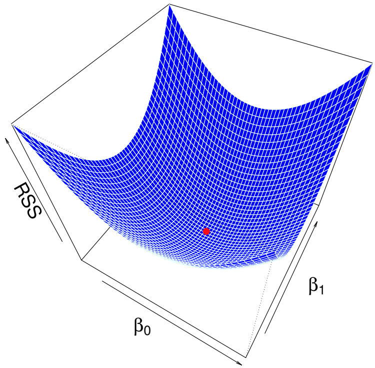
Example: LS fit for advertising data



$\hat{\beta}_0 = 7.03,\ \hat{\beta}_1 = 0.0475$

LS fit of sales vs. TV budget: RSS as a function of (β_0, β_1)





Left: Level curves. Right: Surface plot.

Linear regression yields a linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{3.5}$$

where β_0 : intercept

β_1 : slope

ε : model error, modeled as centered random variable, independent of X.

Linear regression yields a linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{3.5}$$

where β_0 : intercept

β_1 : slope

ε : model error, modeled as centered random variable, independent of X.

Model (3.5) defines the **population regression line**, the best linear approximation to the true (generally unknown) relationship between X and Y.

Linear regression yields a linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{3.5}$$

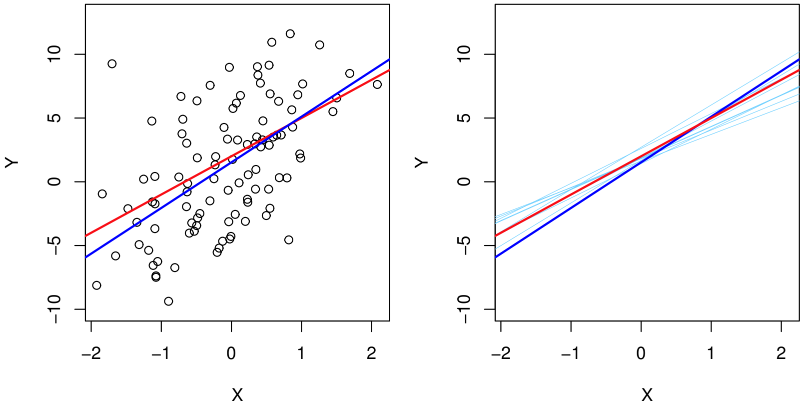
where
β_0 : intercept
β_1 : slope
ε : model error, modeled as centered random variable, independent of X.

Model (3.5) defines the **population regression line**, the best linear approximation to the true (generally unknown) relationship between X and Y.

The linear relation (3.2) containing the coefficients $\hat{\beta}_0, \hat{\beta}_1$ estimated from a given data set is called the **least squares line**.

Simple Linear Regression

Example: population regression line, least squares line



- Left: Simulated data set ($n = 100$) from model $f(X) = 2 + 3X$.
  Red line: population regression line (true model).
  Blue line: least squares line from data (black dots).
- Right: Additionally ten (light blue) least squares lines obtained from ten separate randomly generated data sets from same model; seen to average to the red line.

- Standard statistical approach: use information contained in a sample to estimate characteristics of a large (possibly infinite) population.

[4] Standard deviation of the sample distribution, i.e., average amount $\hat{\mu}$ differs from μ.

Simple Linear Regression

Analogy: estimation of mean

- Standard statistical approach: use information contained in a sample to estimate characteristics of a large (possibly infinite) population.
- Example: approximate **population mean** μ (expectation, expected value) of random variable Y from observations $y_1, \dots, y_n$ by **sample mean** $\hat{\mu} := \overline{y} := \frac{1}{n} \sum_{i=1}^{n} y_i$.

[4] Standard deviation of the sample distribution, i.e., average amount $\hat{\mu}$ differs from μ.

Simple Linear Regression

Analogy: estimation of mean

- Standard statistical approach: use information contained in a sample to estimate characteristics of a large (possibly infinite) population.
- Example: approximate **population mean** μ (expectation, expected value) of random variable Y from observations $y_1, \ldots, y_n$ by **sample mean** $\hat{\mu} := \overline{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$.
- Just as $\hat{\mu} \approx \mu$ but, in general, $\hat{\mu} \neq \mu$, the coefficients $\hat{\beta}_0, \hat{\beta}_1$ defining the least squares line are estimates of the true values β_0, β_1 of the model.

[4] Standard deviation of the sample distribution, i.e., average amount $\hat{\mu}$ differs from μ.

- Standard statistical approach: use information contained in a sample to estimate characteristics of a large (possibly infinite) population.
- Example: approximate **population mean** μ (expectation, expected value) of random variable Y from observations $y_1, \ldots, y_n$ by **sample mean** $\hat{\mu} := \overline{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$.
- Just as $\hat{\mu} \approx \mu$ but, in general, $\hat{\mu} \neq \mu$, the coefficients $\hat{\beta}_0, \hat{\beta}_1$ defining the least squares line are estimates of the true values β_0, β_1 of the model.
- Sample mean $\hat{\mu}$ is an **unbiased estimator** of μ, i.e., it does not *systematically* over- or underestimate the true value μ.
  Same holds for estimators $\hat{\beta}_0, \hat{\beta}_1$.

[4]Standard deviation of the sample distribution, i.e., average amount $\hat{\mu}$ differs from μ.

- Standard statistical approach: use information contained in a sample to estimate characteristics of a large (possibly infinite) population.
- Example: approximate **population mean** μ (expectation, expected value) of random variable Y from observations $y_1, \dots, y_n$ by **sample mean** $\hat{\mu} := \overline{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$.
- Just as $\hat{\mu} \approx \mu$ but, in general, $\hat{\mu} \neq \mu$, the coefficients $\hat{\beta}_0, \hat{\beta}_1$ defining the least squares line are estimates of the true values β_0, β_1 of the model.
- Sample mean $\hat{\mu}$ is an **unbiased estimator** of μ, i.e., it does not *systematically* over- or underestimate the true value μ.
  Same holds for estimators $\hat{\beta}_0, \hat{\beta}_1$.
- How accurate is $\hat{\mu} \approx \mu$?

[4]Standard deviation of the sample distribution, i.e., average amount $\hat{\mu}$ differs from μ.

- Standard statistical approach: use information contained in a sample to estimate characteristics of a large (possibly infinite) population.
- Example: approximate **population mean** μ (expectation, expected value) of random variable Y from observations $y_1, \dots, y_n$ by **sample mean** $\hat{\mu} := \overline{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$.
- Just as $\hat{\mu} \approx \mu$ but, in general, $\hat{\mu} \neq \mu$, the coefficients $\hat{\beta}_0, \hat{\beta}_1$ defining the least squares line are estimates of the true values β_0, β_1 of the model.
- Sample mean $\hat{\mu}$ is an **unbiased estimator** of μ, i.e., it does not *systematically* over- or underestimate the true value μ.
  Same holds for estimators $\hat{\beta}_0, \hat{\beta}_1$.
- How accurate is $\hat{\mu} \approx \mu$?
  Standard error[4] of $\hat{\mu}$, denoted $\mathrm{SE}(\hat{\mu})$, satisfies

$$\mathbf{Var}\,\hat{\mu} = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}, \qquad \text{where } \sigma^2 = \mathbf{Var}\,Y. \tag{3.6}$$

[4] Standard deviation of the sample distribution, i.e., average amount $\hat{\mu}$ differs from μ.

For the regression coefficients (assuming uncorrelated observation errors)

$$
\begin{aligned}
\mathrm{SE}(\hat{\beta}_0)^2 &= \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right], \\
\mathrm{SE}(\hat{\beta}_1)^2 &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}, \qquad\qquad \sigma^2 = \mathbf{Var}\, \varepsilon.
\end{aligned}
\tag{3.7}
$$

For the regression coefficients (assuming uncorrelated observation errors)

$$\begin{aligned} \mathrm{SE}(\hat{\beta}_0)^2 &= \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right], \\ \mathrm{SE}(\hat{\beta}_1)^2 &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}, \qquad \sigma^2 = \mathbf{Var}\, \varepsilon. \end{aligned} \tag{3.7}$$

- $\mathrm{SE}(\hat{\beta}_1)$ smaller when x_i more spread out (provides more leverage to estimate slope).
- $\mathrm{SE}(\hat{\beta}_0) = \mathrm{SE}(\hat{\mu})$ if $\overline{x} = 0$. (Then $\hat{\beta}_0 = \overline{y}$.)

Simple Linear Regression

Standard error of regression coefficients

For the regression coefficients (assuming uncorrelated observation errors)

$$
\begin{aligned}
\mathrm{SE}(\hat{\beta}_0)^2 &= \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right], \\
\mathrm{SE}(\hat{\beta}_1)^2 &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}, \qquad\qquad \sigma^2 = \mathbf{Var}\, \varepsilon.
\end{aligned}
\tag{3.7}
$$

- $\mathrm{SE}(\hat{\beta}_1)$ smaller when x_i more spread out (provides more leverage to estimate slope).
- $\mathrm{SE}(\hat{\beta}_0) = \mathrm{SE}(\hat{\mu})$ if $\overline{x} = 0$. (Then $\hat{\beta}_0 = \overline{y}$.)
- σ generally unknown, can be estimated from the data by **residual standard error**

$$
\mathrm{RSE} := \sqrt{\frac{\mathrm{RSS}}{n-2}}.
$$

When RSE used in place of σ, should write $\widehat{\mathrm{SE}}(\hat{\beta}_1)$.

- 95% **confidence interval**: range of values containing true unknown value of parameter with probability 95%.

- 95% **confidence interval**: range of values containing true unknown value of parameter with probability 95%.
- For linear regression: 95% CI for β_1 approximately

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1), \tag{3.8}$$

i.e., with probability 95%,

$$\beta_1 \in [\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)]. \tag{3.9}$$

- 95% **confidence interval**: range of values containing true unknown value of parameter with probability 95%.
- For linear regression: 95% CI for β_1 approximately

$$\hat{\beta}_1 \pm 2 \cdot \mathsf{SE}(\hat{\beta}_1), \tag{3.8}$$

  i.e., with probability 95%,

$$\beta_1 \in [\hat{\beta}_1 - 2 \cdot \mathsf{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \mathsf{SE}(\hat{\beta}_1)]. \tag{3.9}$$

- Similarly, for β_0, 95% CI approximately given by

$$\hat{\beta}_0 \pm 2 \cdot \mathsf{SE}(\hat{\beta}_0). \tag{3.10}$$

- 95% **confidence interval**: range of values containing true unknown value of parameter with probability 95%.
- For linear regression: 95% CI for β_1 approximately

$$\hat{\beta}_1 \pm 2 \cdot \mathsf{SE}(\hat{\beta}_1), \tag{3.8}$$

  i.e., with probability 95%,

$$\beta_1 \in [\hat{\beta}_1 - 2 \cdot \mathsf{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \mathsf{SE}(\hat{\beta}_1)]. \tag{3.9}$$

- Similarly, for β_0, 95% CI approximately given by

$$\hat{\beta}_0 \pm 2 \cdot \mathsf{SE}(\hat{\beta}_0). \tag{3.10}$$

- For advertising example: with 95% probability

$$\beta_0 \in [6.130, 7.935], \qquad \beta_1 \in [0.042, 0.053].$$

Use SE to test **null hypothesis**

$$H_0 : \text{no relationship between } X \text{ and } Y \tag{3.11}$$

and **alternative hypothesis**

$$H_a : \text{some relationship between } X \text{ and } Y \tag{3.12}$$

or, mathematically,

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

- Reject H_0 if $\hat{\beta}_1$ sufficiently far from 0 relative to $\mathrm{SE}(\hat{\beta}_1)$.
- **t-statistic**

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)} \tag{3.13}$$

measures distance of $\hat{\beta}_1$ from 0 in # standard deviations.

- $\beta_1 = 0$ implies t follows t-distribution with $n - 2$ degrees of freedom.
- We compute probability of observing $|t|$ or larger under assumption $\beta_1 = 0$, its p-**value**.

- $\beta_1 = 0$ implies t follows t-distribution with $n - 2$ degrees of freedom.
- We compute probability of observing $|t|$ or larger under assumption $\beta_1 = 0$, its ***p*-value**.
- Small p-value: unlikely to observe substantial relation between X and Y due to purely random variation, unless the two actually are related.
- In this case we reject H_0.

- $\beta_1 = 0$ implies t follows t-distribution with $n-2$ degrees of freedom.
- We compute probability of observing $|t|$ or larger under assumption $\beta_1 = 0$, its ***p*-value**.
- Small p-value: unlikely to observe substantial relation between X and Y due to purely random variation, unless the two actually are related.
- In this case we reject H_0.
- Typical cutoffs for p-value: 1%, 5%; for $n = 30$ corresponds to t-statistic (3.13) values 2 and 2.75. respectively.

- $\beta_1 = 0$ implies t follows t-distribution with $n - 2$ degrees of freedom.
- We compute probability of observing $|t|$ or larger under assumption $\beta_1 = 0$, its ***p*-value**.
- Small p-value: unlikely to observe substantial relation between X and Y due to purely random variation, unless the two actually are related.
- In this case we reject H_0.
- Typical cutoffs for p-value: 1%, 5%; for $n = 30$ corresponds to t-statistic (3.13) values 2 and 2.75. respectively.

For `TV` sales data in advertising data set:

	Estimate	SE	t-statistic	p-value
β_0	7.0325	0.4578	15.36	< 0.0001
β_1	0.0475	0.0027	17.67	< 0.0001

- Given $X_1, \cdots, X_n$ i.i.d. $\sim N(\mu, \sigma^2)$
- Sample mean:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- (Bessel corrected) sample variance:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

- RV

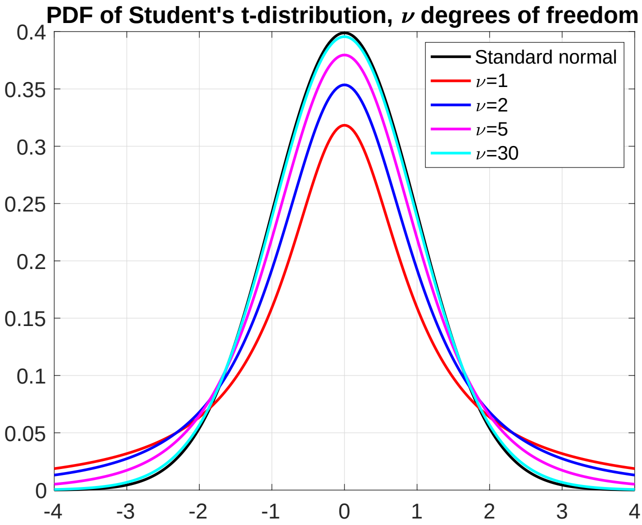
$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

distributed according to $N(0, 1)$.

- RV

$$\frac{\overline{X} - \mu}{S / \sqrt{n}}$$

distributed according to Student's t-distribution with $n - 1$ DoF.



PDF of Student's t-distribution, ν degrees of freedom
Standard normal
ν=1
ν=2
ν=5
ν=30
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4

- **Residual standard error**: estimate of standard deviation of ε (model error)

$$\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{n-2}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{3.14}$$

- For `TV` data $\mathrm{RSS} = 3.26$, i.e., deviation of sales from true regression line on average by 3,260 units (even if exact β_0, β_1 known).
  Corresponds to $3,260/14,000 = 23\%$ error relative to mean value of all sales.

- **Residual standard error**: estimate of standard deviation of ε (model error)

$$\mathrm{RSE} = \sqrt{\frac{\mathrm{RSS}}{n-2}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \qquad (3.14)$$

- For `TV` data $\mathrm{RSS} = 3.26$, i.e., deviation of sales from true regression line on average by 3,260 units (even if exact β_0, β_1 known).
  Corresponds to $3,260/14,000 = 23\%$ error relative to mean value of all sales.
- RSE measures *lack of model fit*.

- R^2 **statistic**: alternative measure of fit: proportion of variance explained.
- $\in [0, 1]$, independent of scale of Y.

- R^2 **statistic**: alternative measure of fit: proportion of variance explained.
- $\in [0, 1]$, independent of scale of Y.
- Defined in terms of **total sum of squares (TSS)** as

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}, \qquad \text{TSS} = \sum_{i=1}^{n} (y_i - \overline{y})^2. \tag{3.15}$$

- TSS : total variance in response Y,
  RSS : amount of variability left unexplained after regression,
  TSS − RSS : response variability explained by regression model,
  R^2 : proportion of variability in Y explained using X.
- $R^2 \approx 0$: linear model wrong, high model error variance.
- For `TV` data $R^2 = 0.61$: 2/3 of `sales` variability explained by (linear regression on) TV budget.
- $R^2 \in [0, 1]$, but sufficient value problem dependent.

- Measure of linear relationship between X and Y: **(sample) correlation**:

$$\mathrm{Cor}(X,Y) = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}}. \tag{3.16}$$

- In simple linear regression: $\mathrm{Cor}(X,Y)^2 = R^2$.
- Correlation expresses association between single pair of variables; R^2 between larger number of variables in multivariate linear regression.

Contents

Multiple Linear Regression

Justification

- $p > 1$ predictor variables
  (as in `advertising` data set: `TV`, `newspaper`, `radio`)
- Easiest option: simple linear regression for each

- $p > 1$ predictor variables
  (as in `advertising` data set: `TV`, `newspaper`, `radio`)
- Easiest option: simple linear regression for each

For `radio` sales data in `advertising` data set:

	Estimate	SE	t-statistic	p-value
β_0	9.312	0.563	16.54	< 0.0001
β_1	0.203	0.020	9.92	< 0.0001

For `newspaper` sales data in `advertising` data set:

	Estimate	SE	t-statistic	p-value
β_0	12.351	0.621	19.88	< 0.0001
β_1	0.055	0.017	3.30	< 0.00115

Multiple Linear Regression
Justification

- How to predict total sales given 3 budgets?
- For given values of the 3 budgets, each simple regression model will give different sales prediction.
- Each separate regression equation ignores the other 2 media.
- For correlated media budgets this can lead to misleading estimates of individual media effects.

- How to predict total sales given 3 budgets?
- For given values of the 3 budgets, each simple regression model will give different sales prediction.
- Each separate regression equation ignores the other 2 media.
- For correlated media budgets this can lead to misleading estimates of individual media effects.

Multiple linear regression model for p predictor variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{3.17}$$

β_j : average effect on Y of 1-unit increase in X_j holding other predictors fixed.

- How to predict total sales given 3 budgets?
- For given values of the 3 budgets, each simple regression model will give different sales prediction.
- Each separate regression equation ignores the other 2 media.
- For correlated media budgets this can lead to misleading estimates of individual media effects.

Multiple linear regression model for p predictor variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{3.17}$$

β_j : average effect on Y of 1-unit increase in X_j holding other predictors fixed.

In advertising example:

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} \tag{3.18}$$

- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, obtain prediction formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p. \tag{3.19}$$

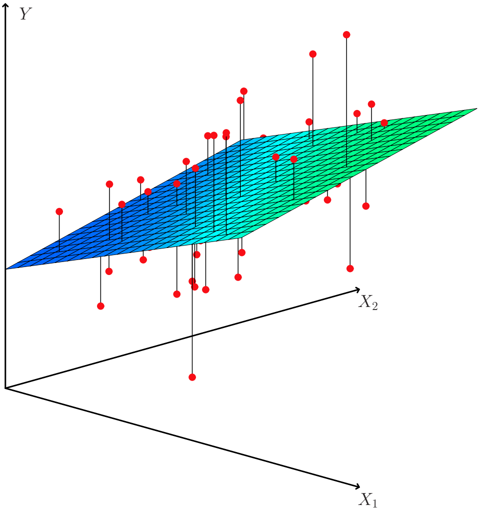
- Same fitting approach: choose $\{\hat{\beta}_j\}_{j=0}^p$ to minimize

$$\mathrm{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \cdots - \hat{\beta}_p x_{i,p})^2, \tag{3.20}$$

yielding the **multiple least squares regression coefficients**

Multiple Linear Regression

Example: multiple linear regression, 2 predictors, 1 response



Multiple Linear Regression

Advertising data

	Estimate	SE	t-statistic	p-value
β_0	2.939	0.3119	9.42	< 0.0001
β_1 (TV)	0.046	0.0014	32.81	< 0.0001
β_2 (radio)	0.189	0.0086	21.89	< 0.0001
β_3 (newspaper)	-0.001	0.0059	-0.18	0.8599

Multiple Linear Regression

Advertising data

	Estimate	SE	t-statistic	p-value
β_0	2.939	0.3119	9.42	< 0.0001
β_1 (TV)	0.046	0.0014	32.81	< 0.0001
β_2 (radio)	0.189	0.0086	21.89	< 0.0001
β_3 (newspaper)	-0.001	0.0059	-0.18	0.8599

- Newspaper slope differs from simple regression: much smaller estimate, p-value no longer significant.

Multiple Linear Regression

Advertising data

	Estimate	SE	t-statistic	p-value
β_0	2.939	0.3119	9.42	< 0.0001
β_1 (TV)	0.046	0.0014	32.81	< 0.0001
β_2 (radio)	0.189	0.0086	21.89	< 0.0001
β_3 (newspaper)	-0.001	0.0059	-0.18	0.8599

- Newspaper slope differs from simple regression: much smaller estimate, p-value no longer significant.
- Now apparently no relation between sales and newspaper budget. Contradiction?

Correlation matrix:

	TV	radio	newapaper	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

Multiple Linear Regression

Correlation matrix:

	TV	radio	newapaper	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

- Correlation between newspaper and radio: ≈ 0.35: tend to spend more on radio ads whenever more is spent on newspaper ads.

Correlation matrix:

	TV	radio	newapaper	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

- Correlation between newspaper and radio: ≈ 0.35: tend to spend more on radio ads whenever more is spent on newspaper ads.
- If correct, i.e., $\beta_{\text{newspaper}} \approx 0$, $\beta_{\text{radio}} > 0$, radio ads increase sales, and where radio budget high, newpaper budget also tends to be high.

Correlation matrix:

	TV	radio	newapaper	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

- Correlation between newspaper and radio: ≈ 0.35: tend to spend more on radio ads whenever more is spent on newspaper ads.
- If correct, i.e., $\beta_{\text{newspaper}} \approx 0$, $\beta_{\text{radio}} > 0$, radio ads increase sales, and where radio budget high, newpaper budget also tends to be high.
- Simple linear regression: indicates newspaper associated with higher sales. Multiple regression indicates absence of such affect.
- Newspaper receives credit for radio's affect on sales.
  Sales due to newspaper advertising is a **surrogate** for sales due to radio advertising.

- Counterintuitive but not uncommon. Consider following (absurd) example.

Multiple Linear Regression

Absurd example, same effect

- Counterintuitive but not uncommon. Consider following (absurd) example.
- Data on `shark attacks` versus `ice cream sales` at beach community would show similar positive relationship as `newpaper` and `radio` ads.
- Should one ban ice cream sales to reduce risk of shark attacks?

Multiple Linear Regression

Absurd example, same effect

- Counterintuitive but not uncommon. Consider following (absurd) example.
- Data on `shark attacks` versus `ice cream sales` at beach community would show similar positive relationship as `newpaper` and `radio` ads.
- Should one ban ice cream sales to reduce risk of shark attacks?
- Answer: High temperatures cause both (more people at beach for shark encounters, more ice cream customers).
- Multiple regression reveals icre cream sales not a predictor for shark attacks after adjusting for temperature.

Multiple Linear Regression

Questions to consider

1. Is *at least one* of the predictors X_1, X_2, ..., X_p useful in predicting the response?
2. Do *all* predictors help to explain Y, or is *only a subset* of the predictors useful?
3. *How well* does the model fit the data?
4. Given a set of predictor values, what response value should we *predict*, and how accurate is our prediction?

- As for simple regression, perform statistical hypothesis test: null hpothesis

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

versus alternative

$$H_a : \quad \text{at least one } \beta_j (j = 1, \ldots, p) \text{ is nonzero.}$$

Multiple Linear Regression

(1) Is there a relationship between response and predictors?

- As for simple regression, perform statistical hypothesis test: null hpothesis

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

versus alternative

$$H_a : \qquad \text{at least one } \beta_j (j = 1, \dots, p) \text{ is nonzero.}$$

- Such a test can be based on the ***F*-statistic** (F ratio)

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}\,/(n - p - 1)} \sim F_{p,n-p-1} \tag{3.21}$$

where, as before,

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

Multiple Linear Regression

(1) Is there a relationship between response and predictors?

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

- Under linear model assumption, can show

$$\mathbf{E}\left[\frac{\mathrm{RSS}}{n-p-1}\right] = \sigma^2$$

(σ^2 again the common noise variance).

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

- Under linear model assumption, can show

$$\mathbf{E}\left[\frac{\mathrm{RSS}}{n-p-1}\right] = \sigma^2$$

(σ^2 again the common noise variance).

- If also H_0 is trueß, can show

$$\mathbf{E}\left[\frac{\mathrm{TSS} - \mathrm{RSS}}{p}\right] = \sigma^2.$$

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

- Under linear model assumption, can show

$$\mathbf{E}\left[\frac{\mathrm{RSS}}{n-p-1}\right] = \sigma^2$$

  (σ^2 again the common noise variance).
- If also H_0 is trueß, can show

$$\mathbf{E}\left[\frac{\mathrm{TSS} - \mathrm{RSS}}{p}\right] = \sigma^2.$$

- Hence $F \approx 1$ if no relationship between response and predictors. Alternatively, if H_a true, $\mathbf{E}\left[(\mathrm{TSS} - \mathrm{RSS})/p\right] > \sigma^2$, hence $F > 1$.

Statistics for multiple regression of `sales` onto `radio`, `TV` and `newspaper` in the advertising data set:

Quantity	Value
RSE	1.69
R^2	0.897
F	570

- $F \gg 1$ strong evidence against H_0.
- Proper threshold value for F depends on n, p.
  Larger F needed to reject H_0 for small n.
- H_0 true, ε_i Gaussian, then F follows **F-distribution**; calculate p-value using statistical software.
- Here, p-value ≈ 0 for $F = 570$ in this example, hence we safely reject H_0.

- To test whether *subset* of last $q < p$ coefficients relevant, use null hypothesis

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

- Fit model using all variables *except* last q, obtaining residual sum of squares RSS_0.
- Appropriate F-statistic now

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

- For multiple regression, t-statistic and p values for each variable indicate whether each predictor related to response after adjusting for the remaining variables.
  Equivalent to F-test omitting single variable ($q = 1$).
  Reports partial effect of adding each variable.

What does F statistic tell us that individual p-values don't?

What does F statistic tell us that individual p-values don't?

- Does single small p-value indicate at least one variable relevant? No.
- Example: $p = 100$, $H_0 : \beta_1 = \cdots = \beta_p = 0$ true.
  Then by chance, 5% of p-values below 0.05.
  Almost guaranteed that $p < 0.05$ for at least one variable by chance.

What does F statistic tell us that individual p-values don't?

- Does single small p-value indicate at least one variable relevant? No.
- Example: $p = 100$, $H_0 : \beta_1 = \cdots = \beta_p = 0$ true.
  Then by chance, 5% of p-values below 0.05.
  Almost guaranteed that $p < 0.05$ for at least one variable by chance.
- Thus, for large p, looking only at p-values of individual t-statistics tends to discover spurious relationships.

What does F statistic tell us that individual p-values don't?

- Does single small p-value indicate at least one variable relevant? No.
- Example: $p = 100$, $H_0 : \beta_1 = \cdots = \beta_p = 0$ true.
  Then by chance, 5% of p-values below 0.05.
  Almost guaranteed that $p < 0.05$ for at least one variable by chance.
- Thus, for large p, looking only at p-values of individual t-statistics tends to discover spurious relationships.
- For F-statistic, if H_0 true, only 5% chance of p-value < 0.05 independently of n, p.

Note: F-statistic approach works for $p < n$.

Multiple Linear Regression

(2) Deciding on important variables

- Typically, not all predictors related to response (**variable selection** problem).
- One approach: try all possible models, select best one. Criteria? **Mallow's C_p**, **Akaike information criterion (AIC)**, **cross validation (CV)**, **Bayesian information criterion (BIC)**, **adjusted R^2** (later)

Multiple Linear Regression

(2) Deciding on important variables

- Typically, not all predictors related to response (**variable selection** problem).
- One approach: try all possible models, select best one. Criteria? **Mallow's C_p**, **Akaike information criterion (AIC)**, **cross validation (CV)**, **Bayesian information criterion (BIC)**, **adjusted R^2** (later)
- For p large, trying 2^p models with subsets of variables impractical.
- *Forward selection:* Start with null model (only β_0), fit p simple regressions, add variable leading to lowest RSS, then add variable leading to two-variable model with lowest RSS, continue until stopping criterion met.

- Typically, not all predictors related to response (**variable selection** problem).
- One approach: try all possible models, select best one. Criteria? **Mallow's C_p**, **Akaike information criterion (AIC)**, **cross validation (CV)**, **Bayesian information criterion (BIC)**, **adjusted R^2** (later)
- For p large, trying 2^p models with subsets of variables impractical.
- *Forward selection:* Start with null model (only β_0), fit p simple regressions, add variable leading to lowest RSS, then add variable leading to two-variable model with lowest RSS, continue until stopping criterion met.
- *Backward selection:* Start with full model, remove variable with largest p-value, fit new $(p-1)$-variable model, keep removing least significant variable, until stopping criterion met.

Multiple Linear Regression

- Typically, not all predictors related to response (**variable selection** problem).
- One approach: try all possible models, select best one. Criteria? **Mallow's C_p**, **Akaike information criterion (AIC)**, **cross validation (CV)**, **Bayesian information criterion (BIC)**, **adjusted R^2** (later)
- For p large, trying 2^p models with subsets of variables impractical.
- *Forward selection:* Start with null model (only β_0), fit p simple regressions, add variable leading to lowest RSS, then add variable leading to two-variable model with lowest RSS, continue until stopping criterion met.
- *Backward selection:* Start with full model, remove variable with largest p-value, fit new $(p-1)$-variable model, keep removing least significant variable, until stopping criterion met.
- *Mixed selection:* Start with null model, adding variables with best fit one-by-one, remove variables whenever their p-value rises above threshold, until model contains only variables with low p-values and excludes those with high p-value.

RSE, R^2 computed and interpreted as in simple linear regression.

- $R^2 = \mathrm{Cor}(X, Y)^2$ for simple linear regression.
- $R^2 = \mathrm{Cor}(\hat{Y}, Y)^2$ for multiple linear regression, maximized by fitted model.
- $R^2 \approx 1$: model explains large portion of response variance.

Multiple Linear Regression

(3) Model fit

RSE, R^2 computed and interpreted as in simple linear regression.

- $R^2 = \mathrm{Cor}(X, Y)^2$ for simple linear regression.
- $R^2 = \mathrm{Cor}(\hat{Y}, Y)^2$ for multiple linear regression, maximized by fitted model.
- $R^2 \approx 1$: model explains large portion of response variance.
- *Advertising example:*

  $$\{\texttt{TV}, \texttt{radio}, \texttt{newspaper}\} \qquad R^2 = 0.8972$$
  $$\{\texttt{TV}, \texttt{radio}\} \qquad R^2 = 0.89719$$

  Small *increase* on including `newspaper` (even though `newspaper` not significant)
- Note: R^2 always increases when variables are added.
- Tiny increase in R^2 on including `newspaper` more evidence this variable can be dropped.
- Including redundant variables promotes overfitting.

Multiple Linear Regression

(3) Model fit

- *Advertising example:*

$$\{\texttt{TV}\} \qquad R^2 = 0.61$$
$$\{\texttt{TV}, \texttt{radio}\} \qquad R^2 = 0.89719$$

Substantial improvement on adding `radio`.
(Could also look at p-value of `radio`'s coefficient in last model.)

Multiple Linear Regression

(3) Model fit

- *Advertising example:*

$$\{\texttt{TV}\} \qquad R^2 = 0.61$$
$$\{\texttt{TV}, \texttt{radio}\} \qquad R^2 = 0.89719$$

  Substantial improvement on adding `radio`.
  (Could also look at p-value of `radio`'s coefficient in last model.)

- *Advertising example:*

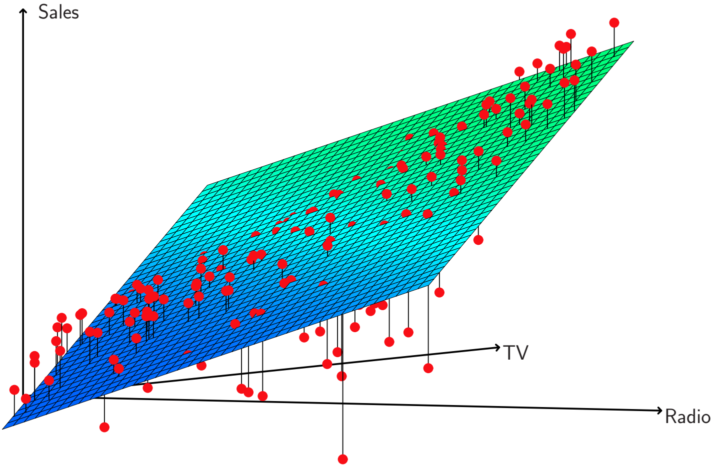
$$\{\texttt{TV}, \texttt{radio}, \ \texttt{newspaper}\} \qquad \text{RSE} = 1.686$$
$$\{\texttt{TV}, \texttt{radio}\} \qquad \text{RSE} = 1.681$$
$$\{\texttt{TV}\} \qquad \text{RSE} = 3.26$$

- Note: for multiple linear regression RSE defined as

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n - p - 1}}.$$

Multiple Linear Regression
(3) Model fit

$\{$`TV`, `radio`$\}$



Previous figure:

- Some observations above, some below least squares regression plane.
- Linear model overestimates sales where most of budget spent either exclusively on `TV` or `radio`.
- Underestimation where budget split between two media.
- Such *nonlinear pattern* not reflected by linear model; suggests *synergy* effect between these two media.

We note three sources of prediction uncertainty:

1. Reducible error: $\hat{Y} \approx f(X)$ since $\hat{\beta}_j \approx \beta_j$.
   Can construct confidence intervals to ascertain closeness $\hat{Y}$ to $f(X)$.
2. *Model bias*: linear model can only yield best *linear* approximation.
3. Irreducible error: $Y = f(X) + \varepsilon$.
   Assess prediction error with **prediction intervals**: incorporate both reducible and irreducible errors.

We note three sources of prediction uncertainty:

1. Reducible error: $\hat{Y} \approx f(X)$ since $\hat{\beta}_j \approx \beta_j$.
   Can construct confidence intervals to ascertain closeness $\hat{Y}$ to $f(X)$.
2. *Model bias*: linear model can only yield best *linear* approximation.
3. Irreducible error: $Y = f(X) + \varepsilon$.
   Assess prediction error with **prediction intervals**: incorporate both reducible and irreducible errors.

Example: Prediction using $\{\texttt{TV}, \texttt{radio}\}$ model.
$X_{\text{TV}} = 100\,000$ \$, $X_{\text{radio}} = 20\,000$ \$.

$$\begin{aligned} &\text{Confidence interval on } \texttt{sales}: && 95\% \text{ confidence interval} : [10.985, 11.528]. \\ &\text{Prediction interval on } \texttt{sales}: && 95\% \text{ prediction interval} : [7.930, 14.580]. \end{aligned}$$

Increased uncertainty about sales for given city in contrast with average sales over many locations.

Contents

- Determining the coefficients $\{\hat{\beta}_j\}_{j=0}^p$ to minimize the RSS in (3.20) is equivalent to minimizing $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2$, where we have introduced the notation

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{bmatrix}, \quad \widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

for the vector $\boldsymbol{y} \in \mathbb{R}^n$ of response observations, the matrix $\boldsymbol{X} \in \mathbb{R}^{n\times(p+1)}$ of predictor observations and vector $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}$ of coefficient estimates.

Computational Solution of Least Squares Problems

Numerical methods for least squares fitting

- Determining the coefficients $\{\hat{\beta}_j\}_{j=0}^p$ to minimize the RSS in (3.20) is equivalent to minimizing $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2$, where we have introduced the notation

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{bmatrix}, \quad \widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

  for the vector $\boldsymbol{y} \in \mathbb{R}^n$ of response observations, the matrix $\boldsymbol{X} \in \mathbb{R}^{n\times(p+1)}$ of predictor observations and vector $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}$ of coefficient estimates.
- The problem of finding a vector $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{b} \approx \boldsymbol{A}\boldsymbol{x}$ for given $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$ is called a **linear regression problem**.

- Determining the coefficients $\{\hat{\beta}_j\}_{j=0}^p$ to minimize the RSS in (3.20) is equivalent to minimizing $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2$, where we have introduced the notation

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{bmatrix}, \quad \widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

  for the vector $\boldsymbol{y} \in \mathbb{R}^n$ of response observations, the matrix $\boldsymbol{X} \in \mathbb{R}^{n\times(p+1)}$ of predictor observations and vector $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}$ of coefficient estimates.
- The problem of finding a vector $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{b} \approx \boldsymbol{A}\boldsymbol{x}$ for given $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$ is called a **linear regression problem**.
- One (of many) possible approaches for achieving this is choosing $\boldsymbol{x}$ to minimize $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2$, which is a **linear least squares problem**.

- A somewhat more general fitting approach using a model

$$y \approx \beta_0 + \beta_1 f_1(\boldsymbol{x}) + \cdots + \beta_p f_p(\boldsymbol{x})$$

with fixed **regression functions** $\{f_j\}_{j=1}^p$ also leads to a linear regression problem, where now $[\boldsymbol{X}]_{i,j} = f_j(x_i)$.

Computational Solution of Least Squares Problems

Numerical methods for least squares fitting

- A somewhat more general fitting approach using a model

$$y \approx \beta_0 + \beta_1 f_1(\boldsymbol{x}) + \cdots + \beta_p f_p(\boldsymbol{x})$$

  with fixed **regression functions** $\{f_j\}_{j=1}^p$ also leads to a linear regression problem, where now $[\boldsymbol{X}]_{i,j} = f_j(x_i)$.
- A linear least squares problem $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2 \to \min$ with $m \geq n$ has a unique solution if the columns of $\boldsymbol{A}$ are linearly independent, i.e., when $\boldsymbol{A}$ has full rank, given by $\boldsymbol{x} = (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \boldsymbol{b}$.
  In this case the solution can be computed using a Cholesky decomposition.

Computational Solution of Least Squares Problems

Numerical methods for least squares fitting

- A somewhat more general fitting approach using a model

$$y \approx \beta_0 + \beta_1 f_1(\boldsymbol{x}) + \cdots + \beta_p f_p(\boldsymbol{x})$$

  with fixed **regression functions** $\{f_j\}_{j=1}^p$ also leads to a linear regression problem, where now $[\boldsymbol{X}]_{i,j} = f_j(x_i)$.
- A linear least squares problem $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2 \to \min$ with $m \geq n$ has a unique solution if the columns of $\boldsymbol{A}$ are linearly independent, i.e., when $\boldsymbol{A}$ has full rank, given by $\boldsymbol{x} = (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{b}$.
  In this case the solution can be computed using a Cholesky decomposition.
- In the (nearly) rank-deficient case, more sophisticated techniques of numerical linear algebra like the QR decomposition or the SVD are required to obtain a (stable) solution.

Computational Solution of Least Squares Problems

Numerical methods for least squares fitting

- A somewhat more general fitting approach using a model

$$y \approx \beta_0 + \beta_1 f_1(\boldsymbol{x}) + \cdots + \beta_p f_p(\boldsymbol{x})$$

  with fixed **regression functions** $\{f_j\}_{j=1}^p$ also leads to a linear regression problem, where now $[\boldsymbol{X}]_{i,j} = f_j(x_i)$.
- A linear least squares problem $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2 \to \min$ with $m \geq n$ has a unique solution if the columns of $\boldsymbol{A}$ are linearly independent, i.e., when $\boldsymbol{A}$ has full rank, given by $\boldsymbol{x} = (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{b}$.
  In this case the solution can be computed using a Cholesky decomposition.
- In the (nearly) rank-deficient case, more sophisticated techniques of numerical linear algebra like the QR decomposition or the SVD are required to obtain a (stable) solution.
- When $\boldsymbol{A}$ is large and sparse or structured, iterative methods such as CGLS or LSQR can be employed which require only matrix-vector products in place of manipulations of matrix entries.

Theorem 3.1 (QR decomposition)

For any matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ of rank p with $n \geq p$ there exists an orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ and an upper triangular matrix $\boldsymbol{R} \in \mathbb{R}^{p\times p}$ such that

$$\boldsymbol{A} = \boldsymbol{Q}\begin{bmatrix}\boldsymbol{R}\\ \boldsymbol{O}\end{bmatrix} \qquad \text{with a zero matrix } \boldsymbol{O} \in \mathbb{R}^{(n-p)\times p}.$$

The decomposition can be made unique by requiring the diagonal entries of $\boldsymbol{R}$ to be positive.

The solution of the least squares problem $\|\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}\|_2 \to \min_{\boldsymbol{x}\in\mathbb{R}^p}$ is given by

$$\hat{\boldsymbol{x}} = \boldsymbol{R}^{-1}\boldsymbol{z}_1, \qquad \|\boldsymbol{b} - \boldsymbol{A}\hat{\boldsymbol{x}}\|_2 = \|\boldsymbol{z}_2\|_2,$$

where

$$\boldsymbol{z}_1 := [\boldsymbol{Q}^\top \boldsymbol{b}]_{1,\dots,p}, \quad \boldsymbol{z}_2 := [\boldsymbol{Q}^\top \boldsymbol{b}]_{p+1,\dots,n}.$$

Theorem 3.2 (Singular value decomposition)

For any matrix $\boldsymbol{A} \in \mathbb{R}^{n \times p}$ of rank $r (\leq \min\{n, p\})$ there exist orthogonal matrices $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{V} \in \mathbb{R}^{p \times p}$ as well as a "diagonal" matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_r & O \\ O & O \end{bmatrix} \in \mathbb{R}^{n \times p} \qquad \text{where} \qquad \boldsymbol{\Sigma}_r = \operatorname{diag}(\sigma_1, \sigma_2, \dots, \sigma_r) \in \mathbb{R}^{r \times r}$$

and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, such that

$$\boldsymbol{A} = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^\top. \tag{SVD}$$

- The positive numbers $\sigma_1, \dots, \sigma_r$ are called the **singular values** of $\boldsymbol{A}$.
- The columns of $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_n]$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_p]$ are the **left and right singular vectors**, respectively.
- (SVD) is called the **singular value decomposition** of $\boldsymbol{A}$.

Singular Value Decomposition

Properties

1. Representation of $\boldsymbol{A}$ as sum of rank-1 matrices:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = [\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_r]\, \boldsymbol{\Sigma}_r\, [\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_r]^\top = \sum_{k=1}^{r} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^\top$$

2. Singular vector mapping properties:

$$\boldsymbol{A}\boldsymbol{v}_k = \begin{cases} \sigma_k \boldsymbol{u}_k & k = 1, 2, \dots, r, \\ \boldsymbol{0} & k = r+1, \dots, p, \end{cases}$$

and

$$\boldsymbol{A}^\top \boldsymbol{u}_k = \begin{cases} \sigma_k \boldsymbol{v}_k & k = 1, 2, \dots, r, \\ \boldsymbol{0} & k = r+1, \dots, n. \end{cases}$$

3.

$\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_r\}$	is an ON-basis of	$\mathscr{R}(\boldsymbol{A})$.
$\{\boldsymbol{u}_{r+1}, \dots, \boldsymbol{u}_n\}$	is an ON-basis of	$\mathscr{N}(\boldsymbol{A}^\top) = \mathscr{R}(\boldsymbol{A})^\perp$.
$\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_r\}$	is an ON-basis of	$\mathscr{R}(\boldsymbol{A}^\top) = \mathscr{N}(\boldsymbol{A})^\perp$.
$\{\boldsymbol{v}_{r+1}, \dots, \boldsymbol{v}_p\}$	is an ON-basis of	$\mathscr{N}(\boldsymbol{A})$.

4. Eigenspaces of $\boldsymbol{A}\boldsymbol{A}^\top$ and $\boldsymbol{A}^\top\boldsymbol{A}$:
   - $\sigma_1^2, \dots, \sigma_r^2$ are the non-zero eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$, respectively:

$$\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{V}\begin{bmatrix} \boldsymbol{\Sigma}_r^2 & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}\boldsymbol{V}^\top,$$

$$\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top = \boldsymbol{U}\begin{bmatrix} \boldsymbol{\Sigma}_r^2 & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}\boldsymbol{U}^\top.$$

   - In particular, the singular values $\sigma_1, \dots, \sigma_r$ are uniquely determined by $\boldsymbol{A}$.
   - The right singular vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_p$ form an ON-basis of $\mathbb{R}^p$ of eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$:

$$\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{v}_k = \begin{cases} \sigma_k^2\boldsymbol{v}_k & k = 1, 2, \dots, r, \\ \boldsymbol{0} & k = r+1, \dots, p. \end{cases}$$

   The left singular vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$ form an ON-basis of $\mathbb{R}^n$ of eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\top$:

$$\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{u}_k = \begin{cases} \sigma_k^2\boldsymbol{u}_k & k = 1, 2, \dots, r, \\ \boldsymbol{0} & k = r+1, \dots, n. \end{cases}$$

5. If $\boldsymbol{A} = \boldsymbol{A}^\top \in \mathbb{R}^{n \times n}$ with non-zero eigenvalues

$$\lambda_1, \dots, \lambda_r, \qquad |\lambda_1| \geq \cdots \geq |\lambda_r| > 0,$$

   then the singular values of $\boldsymbol{A}$ are given by $\sigma_k = |\lambda_k|$.

6. The (p-dimensional) unit sphere is mapped by $\boldsymbol{A}$ to an ellipsoid (in $\mathbb{R}^n$) with center $\mathbf{0}$ and semi-axes $\sigma_k \boldsymbol{u}_k$ ($\sigma_k := 0$ für $k > r$).
7. For $\boldsymbol{A} \in \mathbb{R}^{n \times p}$ there holds $\|\boldsymbol{A}\|_2 = \sigma_1$ and $\|\boldsymbol{A}\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$.
   For $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ invertible, there holds in addition that $\|\boldsymbol{A}^{-1}\|_2 = \sigma_n^{-1}$.
8. Anlalogous statements hold for complex-valued matrices $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H$ ($\boldsymbol{U}, \boldsymbol{V}$ unitary). (In (5) replace $\boldsymbol{A} = \boldsymbol{A}^\top$ by '$\boldsymbol{A}$ normal'.)

Singular Value Decomposition

Best rank-k approximation

- E. Schmidt. Zur Theorie der linearen und nichtlinearen Integralgleichungen. 1. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener. Math. Ann., 63 (1907), pp. 433–476
- C. Eckart, G. Young. The approximation of one matrix by another of lower rank. Psychometrika, 1 (1936), pp. 211–218
- L. Mirsky. Symmetric gauge functions and unitarily invariant norms. Quart. J. Math. Oxford, 11 (1960), pp. 50–59

Theorem 3.3 (Best approximation by matrices of lower rank)

For a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ of rank r with SVD $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ the best approximation problem

$$\min\{\|\boldsymbol{A} - \boldsymbol{B}\|_2 \,:\, \boldsymbol{B} \in \mathbb{R}^{n\times p} \text{ and } \operatorname{rank}(B) \le k\}$$

for $k < r$ is solved by

$$\boldsymbol{A}_k := \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top \qquad \text{with} \qquad \|\boldsymbol{A} - \boldsymbol{A}_k\|_2 = \sigma_{k+1}.$$

- $\boldsymbol{A}_k$ as above is also the closest rank-k matrix to $\boldsymbol{A}$ in the Frobenius-norm $\|\cdot\|_F$, with distance $\|\boldsymbol{A} - \boldsymbol{A}_k\|_F = \sqrt{\sigma_{k+1}^2 + \cdots + \sigma_r^2}$.

Optimality of LS Estimate

The Gauss-Markov[5] theorem

Theorem 3.4 (Gauss-Markov theorem)

Given observations $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \in \mathbb{R}^n$, for which the uncorrelated random noise variables $\{\varepsilon_i\}_{i=1}^n$ have mean zero and constant variance $\sigma^2 > 0$, and assuming that the observation vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_p \in \mathbb{R}^n$ are linearly independent, then the least squares estimate

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \quad \boldsymbol{X} = [\boldsymbol{x}_1 | \cdots | \boldsymbol{x}_p] \in \mathbb{R}^{n \times p}, \quad \boldsymbol{y} \in \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

has minimal variance among all linear unbiased estimators of $\boldsymbol{\beta}$.

[5] C.F. Gauss, 1777–1855, A.A. Markov, 1856–1922

- No assumption is made on the distribution of the errors, only on their first two moments.
- The theorem also holds if $\mathbf{Var}\,\boldsymbol{\varepsilon} = \boldsymbol{C}$ is a (nonsingular) covariance matrix. In this case the best linear unbiased estimator solves the **weighted or generalized least squares** problem

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_{\boldsymbol{C}^{-1}} \to \min_{\boldsymbol{\beta}}, \qquad \|\boldsymbol{x}\|_{\boldsymbol{C}^{-1}}^2 = \boldsymbol{x}^\top \boldsymbol{C}^{-1} \boldsymbol{x}.$$

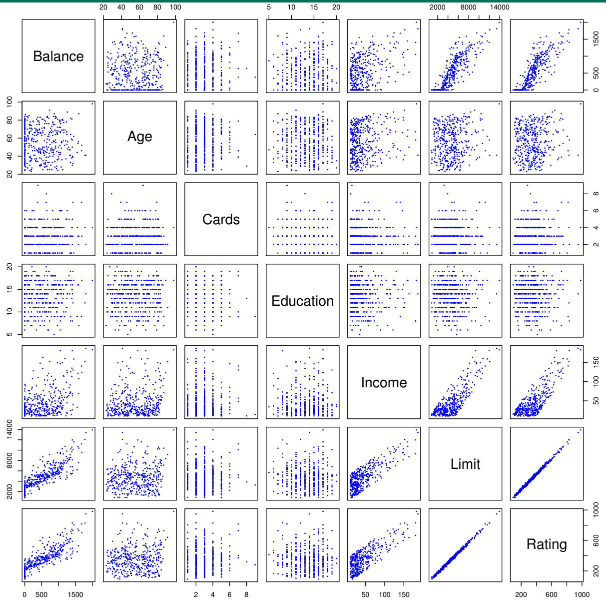
- The theorem does not say there are no better estimators than LS, only that any better estimators are either nonlinear or biased.
- Examples of biased estimators are ridge regression and the lasso.
- ESL:
  "Most models are distortions of the truth, and hence are biased; picking the right model amounts to creating the right balance between bias and variance.

Contents

For a first regression model involving also **qualitative predictor variables**, consider the `Credit` data set:

- Quantitative predictors:
  - `balance`: average credit card debt for a number of individuals
  - `age`
  - `cards` (# credit cards)
  - `education` (years of education)
  - `income` (in thousands of dollars)
  - `limit` (credit limit)
  - `rating` (credit rating)
- Qualitative predictors:
  - `gender`
  - `student` (student status)
  - `status` (marital status)
  - `ethnicity` (Caucasian, African American or Asian)

Other Considerations in the Regression Model
Qualitative predictors



- Goal: investigate differences in credit card balance between males/females.
- `Gender` (qualitative variable, factor) represented with **indicator** (dummy) variable

$$x_i = \begin{cases} 1 & \text{if } i\text{-th person female,} \\ 0 & \text{if } i\text{-th person male.} \end{cases} \tag{3.22}$$

- Using x_i in regression equation results in model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{-th person female,} \\ \beta_0 + \varepsilon_i & \text{if } i\text{-th person male.} \end{cases} \tag{3.23}$$

- Interpretation

$$\begin{aligned} \beta_0 &: \text{ average credit card balance among males,} \\ \beta_0 + \beta_1 &: \text{ average credit card balance among females,} \\ \beta_1 &: \text{average difference in credit card balance male/female.} \end{aligned}$$

Other Considerations in the Regression Model

Two-valued predictors

	Coefficient	Standard error	t-statistic	p-value
β_0	509.80	33.13	15.389	< 0.0001
β_1	19.73	46.05	0.429	0.6690

- Average credit card debt males: \$509.80.
- Average additional credit card debt females: \$19.73.
- Total average female credit card debt: \$529.53.
- High p value for dummy variable. Conclusion?

	Coefficient	Standard error	t-statistic	p-value
β_0	509.80	33.13	15.389	< 0.0001
β_1	19.73	46.05	0.429	0.6690

- Average credit card debt males: \$509.80.
- Average additional credit card debt females: \$19.73.
- Total average female credit card debt: \$529.53.
- High p value for dummy variable. Conclusion?
  Gender not a statistically significant factor for credit card debt.

Other Considerations in the Regression Model

Two-valued predictors

	Coefficient	Standard error	t-statistic	p-value
β_0	509.80	33.13	15.389	< 0.0001
β_1	19.73	46.05	0.429	0.6690

- Average credit card debt males: \$509.80.
- Average additional credit card debt females: \$19.73.
- Total average female credit card debt: \$529.53.
- High p value for dummy variable. Conclusion?
  Gender not a statistically significant factor for credit card debt.
- Switching male/female coding yields estimates

$$\hat{\beta}_0 = \$529.53, \qquad \hat{\beta}_1 = \$ - 19.73, \qquad \hat{\beta}_0 + \hat{\beta}_1 = \$509.80.$$

Another alternative coding of two-valued `gender` predictor:

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

Results in model

$$y_i = \beta_0 + \beta_1 x + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{-th person female,} \\ \beta_0 - \beta_1 + \varepsilon_i & \text{if } i\text{-th person male,} \end{cases}$$

with interpretation now

$$\begin{aligned} \beta_0 &: \text{ average credit card balance (ignoring gender),} \\ \beta_1 &: \text{ amount females are above/males below this average,} \end{aligned}$$

giving estimates

$$\begin{aligned} \hat{\beta}_0 &= \$519.665 && \text{(half way between male and female averages)} \\ \hat{\beta}_1 &= \$\quad 9.865 && \text{(half of \$19.63, average male/female difference).} \end{aligned}$$

Other Considerations in the Regression Model

Multi-valued qualitative predictors

To encode `ethnicity` $\in$ {Caucasian, African American, Asian}, use multiple dummy variables (# values -1)

$$x_{i,1} = \begin{cases} 1 & \text{if } i\text{-th person Asian,} \\ 0 & \text{if } i\text{-th person not Asian,} \end{cases} \tag{3.24}$$

$$x_{i,2} = \begin{cases} 1 & \text{if } i\text{-th person Caucasian,} \\ 0 & \text{if } i\text{-th person not Caucasian,} \end{cases} \tag{3.25}$$

resulting in model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i.2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{-th person Asian} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if } i\text{-th person Caucasian} \\ \beta_0 + \varepsilon_i & \text{if } i\text{-th person African American} \end{cases} \tag{3.26}$$

Interpretation: β_0 : average credit card balance for African Americans (**baseline**),

β_1 : difference between Asian and African Americans,

β_2 : difference between Caucasian and African Americans

	Coefficient	Standard error	t-statistic	p-value
β_0	531.00	46.32	11.464	< 0.0001
β_1 (Asian)	-18.69	65.02	-0.287	0.7740
β_2 (Caucasian)	-12.50	56.68	-0.221	0.8260

- Estimated balance for African Americans (baseline): $531.00.

	Coefficient	Standard error	t-statistic	p-value
β_0	531.00	46.32	11.464	< 0.0001
β_1 (Asian)	-18.69	65.02	-0.287	0.7740
β_2 (Caucasian)	-12.50	56.68	-0.221	0.8260

- Estimated balance for African Americans (baseline): $531.00.
- Asians estimated to have $18.69 less debt than African Americans.
- Caucasians estimated to have $12.50 less debt than African Americans.

Other Considerations in the Regression Model

Multi-valued qualitative predictors

	Coefficient	Standard error	t-statistic	p-value
β_0	531.00	46.32	11.464	< 0.0001
β_1 (Asian)	-18.69	65.02	-0.287	0.7740
β_2 (Caucasian)	-12.50	56.68	-0.221	0.8260

- Estimated balance for African Americans (baseline): \$531.00.
- Asians estimated to have \$18.69 less debt than African Americans.
- Caucasians estimated to have \$12.50 less debt than African Americans.
- $\hat{\beta}_1, \hat{\beta}_2$ have high p-values, indicating no statistical significance for ethnicity as factor in credit card balance.

Other Considerations in the Regression Model

Multi-valued qualitative predictors

	Coefficient	Standard error	t-statistic	p-value
β_0	531.00	46.32	11.464	< 0.0001
β_1 (Asian)	-18.69	65.02	-0.287	0.7740
β_2 (Caucasian)	-12.50	56.68	-0.221	0.8260

- Estimated balance for African Americans (baseline): \$531.00.
- Asians estimated to have \$18.69 less debt than African Americans.
- Caucasians estimated to have \$12.50 less debt than African Americans.
- $\hat{\beta}_1, \hat{\beta}_2$ have high p-values, indicating no statistical significance for ethnicity as factor in credit card balance.
- Coefficients and p-values depend on coding, result does not.
  F-test to reject $H_0 : \beta_1 = \beta_2 = 0$ has p-value 0.96 (cannot reject).

Other Considerations in the Regression Model

Multi-valued qualitative predictors

	Coefficient	Standard error	t-statistic	p-value
β_0	531.00	46.32	11.464	< 0.0001
β_1 (Asian)	-18.69	65.02	-0.287	0.7740
β_2 (Caucasian)	-12.50	56.68	-0.221	0.8260

- Estimated balance for African Americans (baseline): \$531.00.
- Asians estimated to have \$18.69 less debt than African Americans.
- Caucasians estimated to have \$12.50 less debt than African Americans.
- $\hat{\beta}_1, \hat{\beta}_2$ have high p-values, indicating no statistical significance for ethnicity as factor in credit card balance.
- Coefficients and p-values depend on coding, result does not.
  F-test to reject $H_0 : \beta_1 = \beta_2 = 0$ has p-value 0.96 (cannot reject).
- Dummy variable approach works for combining qualitative and quantitative predictors.
  (Other coding schemes for qualitative variables possible.)

Other Considerations in the Regression Model

Extending the linear model

- Restrictive assumptions in linear model: **linearity**, **additivity**.
- Additivity: effect on Y of changing X_j independent of remaining variables.
- Linearity: rate of change in Y with respect to X_j constant in X_j.

- Restrictive assumptions in linear model: **linearity**, **additivity**.
- Additivity: effect on Y of changing X_j independent of remaining variables.
- Linearity: rate of change in Y with respect to X_j constant in X_j.
- Recall advertising data set: indication that higher `radio` budget made effect of `TV` spending stronger (interaction effect, synergy).

Other Considerations in the Regression Model

Extending the linear model

- Restrictive assumptions in linear model: **linearity**, **additivity**.
- Additivity: effect on Y of changing X_j independent of remaining variables.
- Linearity: rate of change in Y with respect to X_j constant in X_j.
- Recall advertising data set: indication that higher `radio` budget made effect of `TV` spending stronger (interaction effect, synergy).
- Add **interaction term** to two-predictor model:

.

- Restrictive assumptions in linear model: **linearity**, **additivity**.
- Additivity: effect on Y of changing X_j independent of remaining variables.
- Linearity: rate of change in Y with respect to X_j constant in X_j.
- Recall advertising data set: indication that higher `radio` budget made effect of `TV` spending stronger (interaction effect, synergy).
- Add **interaction term** to two-predictor model:

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

.

- Restrictive assumptions in linear model: **linearity**, **additivity**.
- Additivity: effect on Y of changing X_j independent of remaining variables.
- Linearity: rate of change in Y with respect to X_j constant in X_j.
- Recall advertising data set: indication that higher `radio` budget made effect of `TV` spending stronger (interaction effect, synergy).
- Add **interaction term** to two-predictor model:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \varepsilon \end{aligned}$$

.

- Restrictive assumptions in linear model: **linearity**, **additivity**.
- Additivity: effect on Y of changing X_j independent of remaining variables.
- Linearity: rate of change in Y with respect to X_j constant in X_j.
- Recall advertising data set: indication that higher `radio` budget made effect of `TV` spending stronger (interaction effect, synergy).
- Add **interaction term** to two-predictor model:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \varepsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \varepsilon, \qquad \tilde{\beta}_1 := \beta_1 + \beta_3 X_2. \end{aligned}$$

- Restrictive assumptions in linear model: **linearity**, **additivity**.
- Additivity: effect on Y of changing X_j independent of remaining variables.
- Linearity: rate of change in Y with respect to X_j constant in X_j.
- Recall advertising data set: indication that higher `radio` budget made effect of `TV` spending stronger (interaction effect, synergy).
- Add **interaction term** to two-predictor model:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \varepsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \varepsilon, \qquad\qquad \tilde{\beta}_1 := \beta_1 + \beta_3 X_2. \end{aligned}$$

$\tilde{\beta}_1$ changes with X_2, hence effect of X_1 on Y changes with X_2.

Example: factory productivity.

- Predict # produced `units` based on # production `lines` and # `workers`.

Other Considerations in the Regression Model

Extending the linear model: factory example

Example: factory productivity.

- Predict # produced `units` based on # production `lines` and # `workers`.
- Expected: increase in # production `lines` will depend on # `workers`.
- In linear model of `units`, include interaction term between `lines` and `workers`. A regression fit may yield, e.g.,

$$
\begin{aligned}
\texttt{units} &\approx 1.2 + 3.4 \times \texttt{lines} + 0.22 \times \texttt{workers} + 1.4 \times (\texttt{lines} \times \texttt{workers}) \\
&= 1.2 + (3.4 + 1.4 \times \texttt{workers}) \times \texttt{lines} + 0.22 \times \texttt{workers}
\end{aligned}
$$

- Adding additional line will increase # produced units by $3.4+1.4\times\texttt{workers}$. The more `workers`, the stronger the effect of adding a `line`.

Linear model for `sales` predicted by interacting `TV`, `radio` terms:

$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \varepsilon\end{aligned} \tag{3.27}$$

Linear model for `sales` predicted by interacting `TV`, `radio` terms:

$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \varepsilon\end{aligned} \tag{3.27}$$

Interpretation of β_3 : increase in effectiveness of TV advertising for one-unit increase in radio advertising.

Linear model for `sales` predicted by interacting `TV`, `radio` terms:

$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \varepsilon\end{aligned} \tag{3.27}$$

Interpretation of β_3 : increase in effectiveness of TV advertising for one-unit increase in radio advertising.

	Coefficient	Standard error	t-statistic	p-value
β_0	6.7502	0.248	27.23	< 0.0001
β_1	0.0191	0.002	12.70	< 0.0001
β_2	0.0289	0.009	3.24	0.0014
β_3	0.0011	0.000	20.73	< 0.0001

Linear model for `sales` predicted by interacting `TV`, `radio` terms:

$$\begin{aligned} \texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \varepsilon \end{aligned} \tag{3.27}$$

Interpretation of β_3 : increase in effectiveness of TV advertising for one-unit increase in radio advertising.

	Coefficient	Standard error	t-statistic	p-value
β_0	6.7502	0.248	27.23	< 0.0001
β_1	0.0191	0.002	12.70	< 0.0001
β_2	0.0289	0.009	3.24	0.0014
β_3	0.0011	0.000	20.73	< 0.0001

- Model with interaction term superior to that including only **main effects**.
- Low p-value of interaction term strong evidence for rejecting $H_0 : \beta_3 = 0$.

Other Considerations in the Regression Model

Extending the linear model: advertising example

- Model (3.27) has $R^2 = 96.8\%$
  (vs. $R^2 = 89.7\%$ for model without interaction term).

- Model (3.27) has $R^2 = 96.8\%$
(vs. $R^2 = 89.7\%$ for model without interaction term).
- Interpretation: of the variability remaining after fitting the model without interaction term,
$$\frac{96.8\% - 89.7\%}{100\% - 89.7\%} = 69\%$$
is explained by model (3.27) which includes the interaction term.

Other Considerations in the Regression Model

Extending the linear model: advertising example

- Model (3.27) has $R^2 = 96.8\%$
  (vs. $R^2 = 89.7\%$ for model without interaction term).
- Interpretation: of the variability remaining after fitting the model without interaction term,
  $$\frac{96.8\% - 89.7\%}{100\% - 89.7\%} = 69\%$$
  is explained by model (3.27) which includes the interaction term.
- \$1000 increase in TV budget associated with sales increase of
  $(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units.
  \$1000 increase in radio budget associated with sales increase of
  $(\hat{\beta}_2 + \hat{\beta}_3 \times \texttt{TV}) \times 1000 = 29 + 1.1 \times \texttt{TV}$ units.

Other Considerations in the Regression Model

Extending the linear model: advertising example

- Model (3.27) has $R^2 = 96.8\%$
  (vs. $R^2 = 89.7\%$ for model without interaction term).
- Interpretation: of the variability remaining after fitting the model without interaction term,
  $$\frac{96.8\% - 89.7\%}{100\% - 89.7\%} = 69\%$$
  is explained by model (3.27) which includes the interaction term.
- \$1000 increase in TV budget associated with sales increase of
  $(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units.
  \$1000 increase in radio budget associated with sales increase of
  $(\hat{\beta}_2 + \hat{\beta}_3 \times \texttt{TV}) \times 1000 = 29 + 1.1 \times \texttt{TV}$ units.
- **Hierarchical principle**: for every interaction term, include all associated main effects, even if the p values of their coefficients not significant.
  `Rationale:` If $X_1 X_2$ related to response, vanishing coefficients for X_1, X_2 unimportant. $X_1 X_2$ typically correlated with X_1, X_2; leaving these out alters meaning of interaction.

`Credit` data set: predict `balance` using `income` (quantitative) and `student` (qualitative). Without interaction term:

$$
\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{-th person student} \\ 0 & \text{otherwise} \end{cases} \\
&= \beta_1 \times \texttt{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{-th person student} \\ \beta_0 & \text{otherwise.} \end{cases}
\end{aligned}
\tag{3.28}
$$

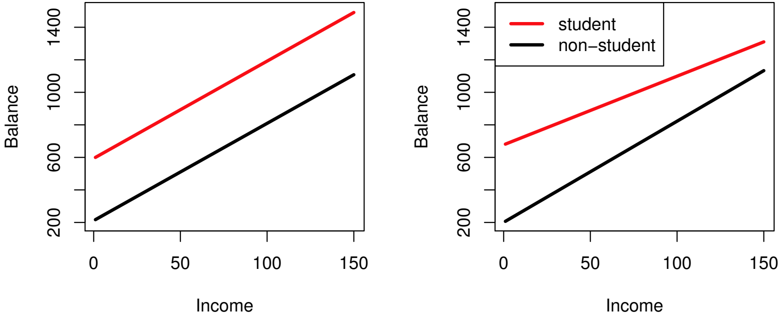
`Credit` data set: predict `balance` using `income` (quantitative) and `student` (qualitative). Without interaction term:

$$\begin{aligned}\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases}\beta_2 & \text{if } i\text{-th person student}\\ 0 & \text{otherwise}\end{cases}\\ &= \beta_1 \times \texttt{income}_i + \begin{cases}\beta_0+\beta_2 & \text{if } i\text{-th person student}\\ \beta_0 & \text{otherwise.}\end{cases}\end{aligned} \tag{3.28}$$

- Results in fitting two parallel lines to data (one each for students and non-students).
- Parallel implies: average affect on balance of one-unit increase in `income` independent of `Student` status.
- Reflects model shortcoming: change in `income` may have very different effect on credit card balance for students and non-students.

Other Considerations in the Regression Model

Extending the linear model: credit example



With interaction term: multiply `income` with dummy variable for `student`

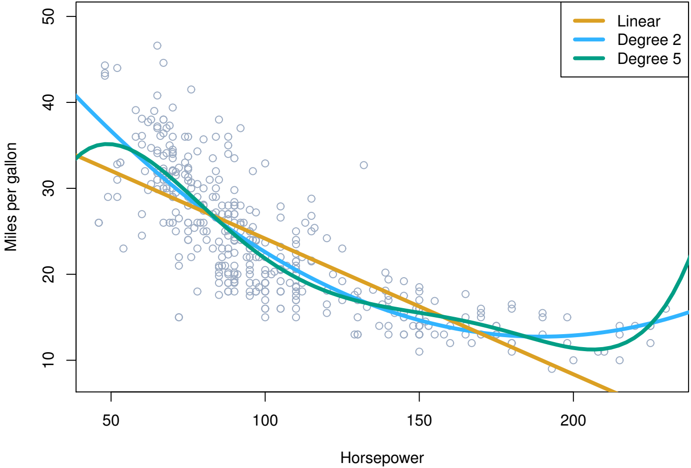
$$
\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if } i\text{-th person student} \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if } i\text{-th person student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{otherwise.} \end{cases}
\end{aligned} \tag{3.29}
$$

- Now the two lines have different intercepts and different slopes.
- Slope for students lower, indicates increases in income associated with smaller increase in credit card balance than for non-students.

Other Considerations in the Regression Model

Extending the linear model: nonlinear relationships

Polynomial regression vs. linear regression:



`Auto` data set showing `mpg` (miles per gallon) versus `horsepower` for different cars.

Other Considerations in the Regression Model

Extending the linear model: nonlinear relationships

Since the data seem to suggest curved relationship, add quadratic term:

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \varepsilon. \tag{3.30}$$

	Coefficient	Standard error	t-statistic	p-value
β_0	56.9001	1.8004	31.6	< 0.0001
β_1	-0.4662	0.0311	-15.0	< 0.0001
β_2	0.0012	0.0001	10.1	< 0.0001

- Linear fit has $R^2 = 0.606$, quadratic fit has $R^2 = 0.688$.
- p-value for quadratic term highly significant.
- Degree 5 fit more oscillatory, doesn't appear to explain data any better than quadratic.

Other Considerations in the Regression Model

Potential problems

Most common problems when fitting a linear regression model to a data set:
(identification and solution as much an art as a science)

1. Nonlinear dependence of response on predictors
2. Correlated error terms
3. Non-constant variance of error terms
4. Outliers
5. High-leverage points
6. Collinearity

Other Considerations in the Regression Model

Potential problems: (1) Nonlinear dependence

Inference and prediction from linear regression model are suspect when true model nonlinear.

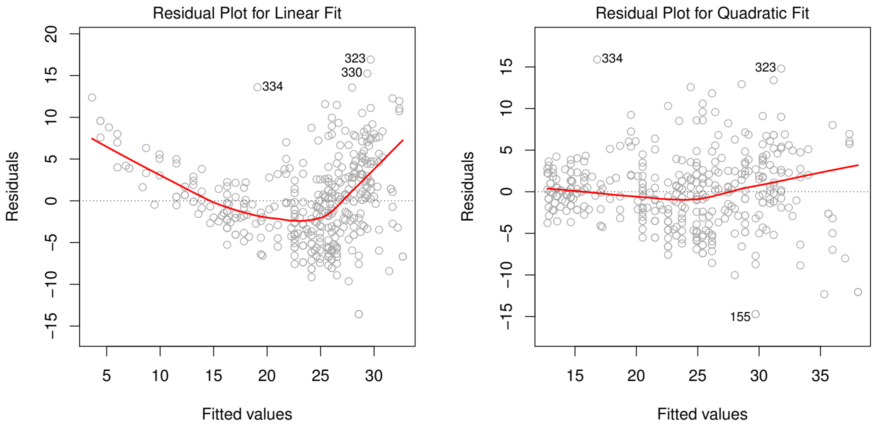
- Identifying nonlinearity aided by **residual plots**

$$r_i = y_i - \hat{y}_i \quad \text{against predictors } x_i.$$

- For multiple regression models, plot residuals against predicted (fitted) values $\hat{y}_i$.
- Ideal picture: no discernible pattern.
- Presence of pattern indicates possible problem with model.
- When nonlinearity is suggested, introduce nonlinear functions of predictors as **regression functions** into the regression model.

Other Considerations in the Regression Model

Potential problems: (1) Nonlinear dependence



Residuals versus predicted values for `Auto` data set.

Red line is smooth fit to residuals to aid in identifying trends.

Left: linear regression of `mpg` on `horsepower` (strong pattern).

Right: linear regression of `mpg` on `horsepower` and `horsepower`^{2} (little pattern).

Other Considerations in the Regression Model

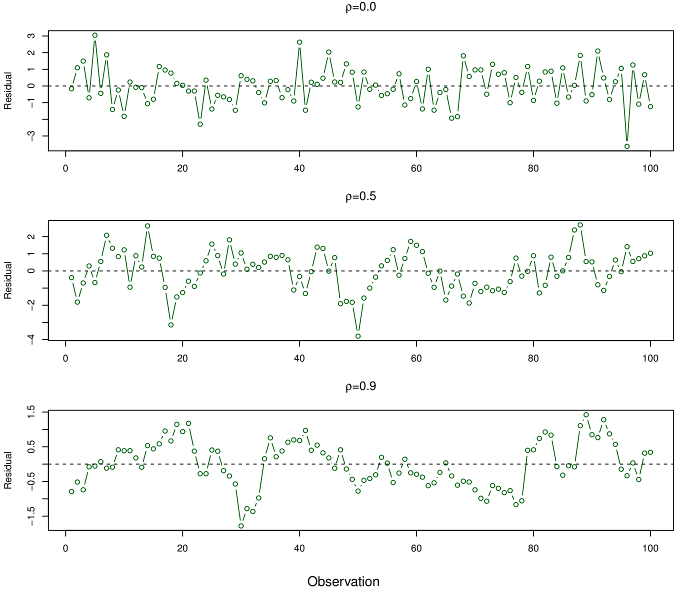
Potential problems: (2) Correlated error terms

- Linear regression assumes uncorrelated errors ε_i.
- Computation of SE for coefficient estimates, fitted values, based on this assumption. Otherwise estimated SE tend to underestimate true SE, confidence and prediction intervals too optimistic (narrow), p-values lower than they should be.

- Linear regression assumes uncorrelated errors ε_i.
- Computation of SE for coefficient estimates, fitted values, based on this assumption. Otherwise estimated SE tend to underestimate true SE, confidence and prediction intervals too optimistic (narrow), p-values lower than they should be.
- Extreme example: double data (observations, error terms identical in pairs). SE calculations use sample size $2n$ in place of n, hence CI narrower by factor of $\sqrt{2}$.
- Detection for **time series**: plot residuals as function of time. No correlations implies no visible pattern; correlations lead to **tracking** of residuals.
- Example (next slide): time series with error correlation $\rho = 0, 0.5, 0.9$
- Example: study of persons' heights predicted from their weights. Uncorrelatedness assumption violated if, e.g., individuals related, same diet or environmantal factors.

Other Considerations in the Regression Model

Potential problems: (2) Correlated error terms



Other Considerations in the Regression Model

Potential problems: (3) Non-constant variance of error terms

- SE, CI, hypothesis tests associated with linear model rely on assumption $\mathbf{Var}\, \varepsilon_i = \sigma^2 \ (\forall i)$.

Other Considerations in the Regression Model

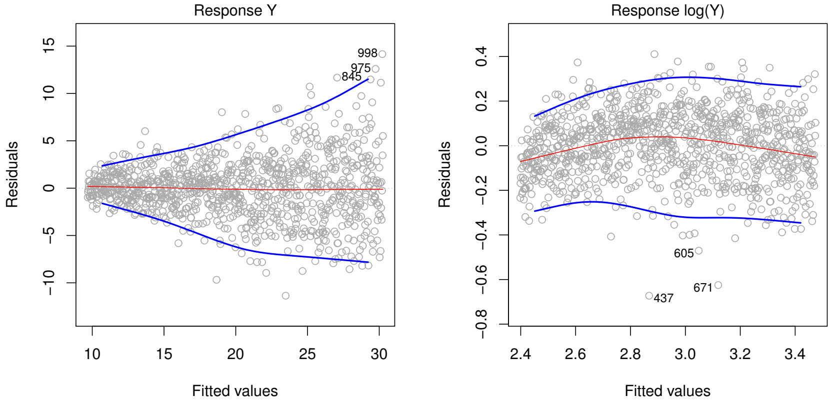
Potential problems: (3) Non-constant variance of error terms

- SE, CI, hypothesis tests associated with linear model rely on assumption $\mathbf{Var}\,\varepsilon_i = \sigma^2$ $(\forall i)$.
- Non-constant error variance (**heteroscedasticity**), e.g. increase with response value, leads to *funnel-shaped* residual plot.
- Possible solution: transform response Y using concave function such as $\log Y$ or $\sqrt{Y}$, leads to damping of larger responses, reducing heteroscedasticity.

Other Considerations in the Regression Model

Potential problems: (3) Non-constant variance of error terms

- SE, CI, hypothesis tests associated with linear model rely on assumption $\mathbf{Var}\,\varepsilon_i = \sigma^2\ (\forall i)$.
- Non-constant error variance (**heteroscedasticity**), e.g. increase with response value, leads to *funnel-shaped* residual plot.
- Possible solution: transform response Y using concave function such as $\log Y$ or $\sqrt{Y}$, leads to damping of larger responses, reducing heteroscedasticity.
- When variation of response variance known, e.g., i-th response average of n_i observations which are uncorrelated with variance σ^2, then average has variance $\sigma_i^2 = \sigma^2/n_i$. Remedy: **weighted least squares** with weights proportional to inverse variances.



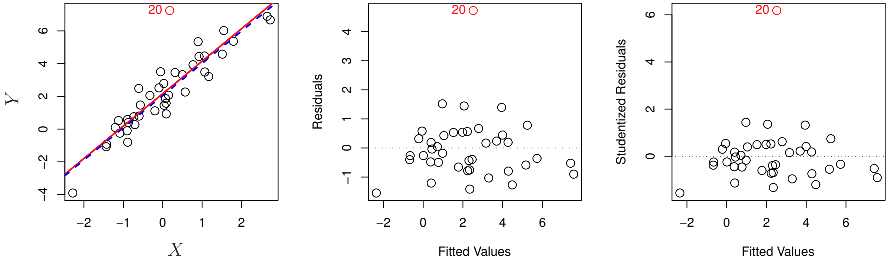
Residual plots. Red: smooth fit of residuals. Blue: track outer quantiles of residuals.
Left: funnel shape indicating heteroscedasticity.

Right: After log-transforming response, heteroscedasticity removed.

Other Considerations in the Regression Model

Potential problems: (4) Outliers

- **Outlier**: point where y_i far from value predicted by model.
- Possible causes: observation errors.

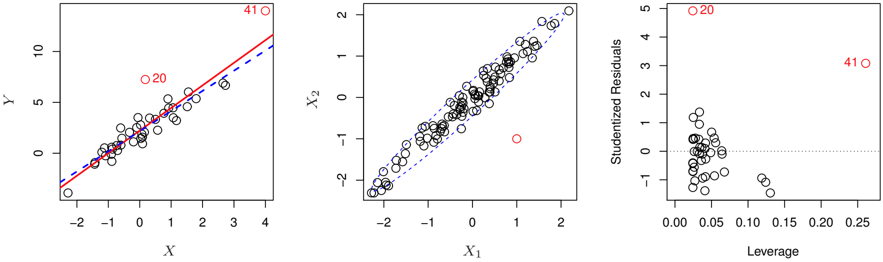


Left: red solid line: least squares line with outlier, blue: without.
Center: Residual plot identifies outlier.
Right: Outlier seen to have studentized residual (divide e_i by its estimated standard error) of 6 (between -3 and 3 expected).
R^2 declines from 0.892 to 0.805 on including outlier.

- Outliers: observations where y_i is unusual given x_i.
- Observations with **high leverage** have unusual value for x_i.
- If least squares line strongly affected by certain points, problems with these may invalidate entire fit, hence important to identify such observations.
- Simple linear regression: extremal x-values;
  multiple linear regression: in range of all other observation coordinates, but unusual (difficult to detect for more than two predictors).
- Large value of **leverage statistic** indicates high leverage.
  For simple linear regression:

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2} \in \left(\frac{1}{n}, 1\right). \tag{3.31}$$

Average value always $\frac{p+1}{n}$, deviation from average indicates high leverage.

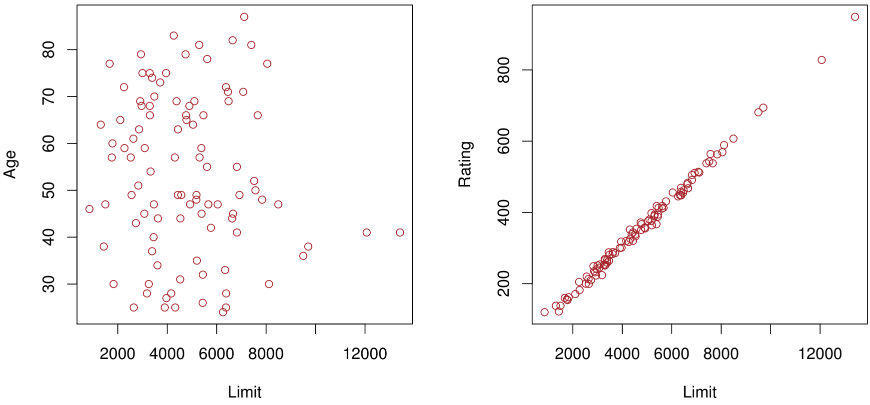


Left: Same data as previous figure, with added observation 41 (red) of high leverage. Red solid line is least squares fit with, blue dashed without observation 41.

Center: two predictor variables, most observations within blue dashed ellipse, red observation distinctly outside, but either coordinate well within normal range.

Right: same data as in left panel, studentized residuals vs. leverage statistic. Observation 41 has high leverage and high residual, i.e., outlier *and* high-leverage point. Outlier observation 20 has low leverage.

Collinearity: two or more predictor variables closely related.



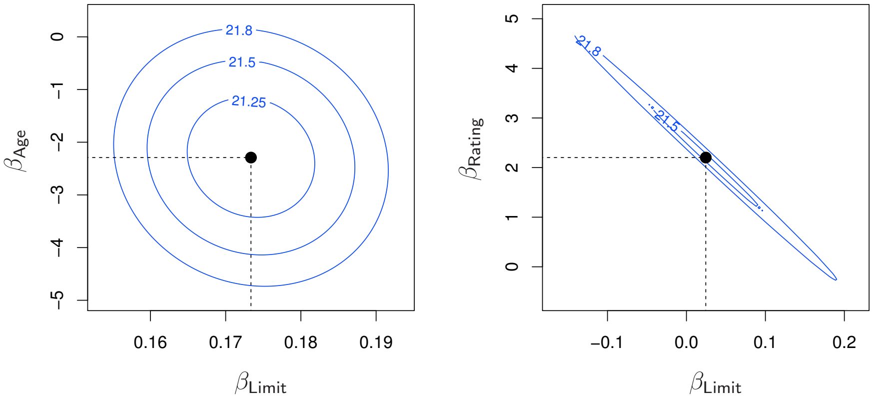
From `Credit` data set.

Left: `limit` vs. `age`. Right: `limit` vs. `rating` (strongly collinear).

Other Considerations in the Regression Model

Potential problems: (6) Collinearity

Difficult to separate individual effects of collinear variables on response.



Contour plots of **RSS** for different coefficient estimates for `Credit` data set. Axes scaled to include 4 **SE** on either side of optimum. Regression of `balance`

Left: on `limit` and `age`. Right: on `limit` and `rating`.

- Collinearity increases SE, hence reduces t-statistic, and we will more likely fail to reject $H_0 : \beta_j = 0$. This reduces the **power** of the hypothesis test, i.e., the probability of correctly detecting a nonzero coefficient.

	Coefficient	Standard error	t-statistic	p-value
	Model 1			
β_0	-173.411	43.828	-3.957	< 0.0001
β_1 (`age`)	-2.292	0.672	-3.407	0.0007
β_2 (`limit`)	0.173	0.005	34.496	< 0.0001
	Model 2			
β_0	-377.537	45.254	-8.343	< 0.0001
β_1 (`rating`)	2.202	0.952	2.312	0.0213
β_2 (`limit`)	0.025	0.064	0.384	0.7012

- Model 1: `age`, `limit` both highly significant.
  Model 2: collinearity between `rating` and `limit` increases SE for `limit` coefficient by factor 12, p-value increases to 0.701. Collinearity masks importance of `limit` variable.

- Important to detect collinearity when fitting a model.
- Correlation matrix may give indication.
- **Multicollinearity**: collinearity between 3 or more variables which each have low pairwise correlation.
- **Variance inflation factor (VIF)**: ratio of variance of $\hat{\beta}_j$ when fitting the full model and variance of $\hat{\beta}_j$ when fitted separately.
- VIF ≥ 1, minimum at complete absence of collinearity.
  Problematic if VIF exceeds 5 or 10.
- $$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$
  $R^2_{X_j|X_{-j}}$: R^2 from regression of X_j onto all other predictors.
- In `Credit` data example: predictors have VIF values of 1.01, 160.67, 160.59.
- Remedies: drop problematic variables, combine collinear variables into single predictor.

Contents

Revisiting the Marketing Data Questions

Recall the seven questions relating to the `Advertising` data set we set out to answer on Slide 75:

1. Is there a relationship between advertising budget and sales?
2. How strong is this relationship between advertising budget and sales?
3. Which media contribute to sales?
4. How accurately can we estimate the effect of each medium on sales?
5. How accurately can we predict future sales?
6. Is the relationship linear?
7. Is there synergy among the advertising media?

We revisit each in turn.

Revisiting the Marketing Data Questions

1. Is there a relationship between advertising budget and sales?

Revisiting the Marketing Data Questions

1. Is there a relationship between advertising budget and sales?

- Fit multiple regression model of `sales` onto `TV`, `radio` and `newspaper`.
- Test hypothesis $H_0 : \beta_{\mathrm{TV}} = \beta_{\mathrm{radio}} = \beta_{\mathrm{newspaper}} = 0.$
- Rejection/non-rejection based on F-statistic (Slide 102).
- For advertising data: low p-value of F-statistic (table on Slide 104) strong evidence for rejecting H_0.

Revisiting the Marketing Data Questions

2. How strong is this relationship between advertising budget and sales?

Revisiting the Marketing Data Questions

2. How strong is this relationship between advertising budget and sales?

- Measure of model error: RSE (see Slide 84), estimates standard deviation of response from (true) population regression line.
- Advertising data:
  For multiple regression model of `sales` on `TV` and `radio`, RSE $= 1,681$ units (Slide 109).
  Relative to response sample mean of $14,022$ units, this is an error of 12%.

Revisiting the Marketing Data Questions

2. How strong is this relationship between advertising budget and sales?

- Measure of model error: RSE (see Slide 84), estimates standard deviation of response from (true) population regression line.
- Advertising data:
  For multiple regression model of `sales` on `TV` and `radio`, RSE $= 1,681$ units (Slide 109).
  Relative to response sample mean of $14,022$ units, this is an error of 12%.
- Measure of model error: R^2 (Slide 91), measures proportion of response variability explained by model.
- Advertising data:
  For multiple regression model of `sales` on `TV`, `radio` and `newspaper`, $R^2 = 0.897$, i.e., $\approx 90\%$ of `sales` variability explained by multiple linear regression model (Slide 104).

Revisiting the Marketing Data Questions

3. Which media contribute to sales?

Revisiting the Marketing Data Questions

3. Which media contribute to sales?

- p-values of t-statistic in multiple regression model of `sales` on `TV`, `radio` and `newspaper`: small for `TV` and `radio`, large for `newspaper`.
- Suggest only `TV` and `radio` budgets related to `sales`.

Revisiting the Marketing Data Questions

4. How accurately can we estimate the effect of each medium on sales?

Revisiting the Marketing Data Questions

4. How accurately can we estimate the effect of each medium on sales?

- Confidence intervals for β_j constructed from SE of $\hat{\beta}_j$.
- Advertising data: 95%-confidence intervals for multiple regression coefficients are

`TV`	$(0.043, 0.049)$
`radio`	$(0.172, 0.206)$
`newspaper`	$(-0.013, 0.011)$

Revisiting the Marketing Data Questions

4. How accurately can we estimate the effect of each medium on sales?

- Confidence intervals for β_j constructed from SE of $\hat{\beta}_j$.
- Advertising data: 95%-confidence intervals for multiple regression coefficients are

`TV`	$(0.043, 0.049)$
`radio`	$(0.172, 0.206)$
`newspaper`	$(-0.013, 0.011)$

- Wide CI due to collinearity? (Slide 151).
  VIF scores for `TV`, `radio` and `newspaper` are 1.005, 1.145, 1.145, so not likely.

Revisiting the Marketing Data Questions

4. How accurately can we estimate the effect of each medium on sales?

- Confidence intervals for β_j constructed from SE of $\hat{\beta}_j$.
- Advertising data: 95%-confidence intervals for multiple regression coefficients are

`TV`	$(0.043, 0.049)$
`radio`	$(0.172, 0.206)$
`newspaper`	$(-0.013, 0.011)$

- Wide CI due to collinearity? (Slide 151).
  VIF scores for `TV`, `radio` and `newspaper` are 1.005, 1.145, 1.145, so not likely.
- Separate simple regressions of `sales` on `TV`, `radio` and `newspaper` show strong association of `TV` and `radio` with `sales`, mild association of `newspaper` with `sales`, when remaining two predictors ignored.

Revisiting the Marketing Data Questions

5. How accurately can we predict future sales?

Revisiting the Marketing Data Questions

5. How accurately can we predict future sales?

- Can use (3.19) for prediction.
- Prediction intervals assess accuracy of predicting individual responses $Y = f(X) + \varepsilon$.
- Confidence intervals assess accuracy of predicting average responses $Y = f(X)$.
- Former always wider due to accounting for additional variability due to irreducible error ε.

Revisiting the Marketing Data Questions

6. Is the relationship linear?

Revisiting the Marketing Data Questions

6. Is the relationship linear?

- Identify nonlinearity using residual plots of linear model (Slide 142).
- Advertising data:
  Nonlinear effects visible in figure on Slide 110.
- Discussed regression functions which are nonlinear in the predictor variables.

Revisiting the Marketing Data Questions

7. Is there synergy among the advertising media?

Revisiting the Marketing Data Questions

7. Is there synergy among the advertising media?

- Non-additive relationships modeled by interaction term in model (Slide 132).
- Presence of interaction (synergy) confirmed by small p-value of interaction term.
- Advertising data:
  Including interaction term increased R^2 from $\approx 90\%$ to $\approx 97\%$.

Contents

Linear Regression vs. K-Nearest Neighbors

Non-parametric approach

- Linear regression is a parametric method.
- Non-parametric methods make no (strong) a priori assumptions on functional form of model $Y \approx f(X)$, more flexibility in adapting to data.

Linear Regression vs. K-Nearest Neighbors

Non-parametric approach

- Linear regression is a parametric method.
- Non-parametric methods make no (strong) a priori assumptions on functional form of model $Y \approx f(X)$, more flexibility in adapting to data.
- Here: **K-nearest neighbors (KNN) regression** (Cf. KNN classifier in Chapter 2).

Linear Regression vs. K-Nearest Neighbors

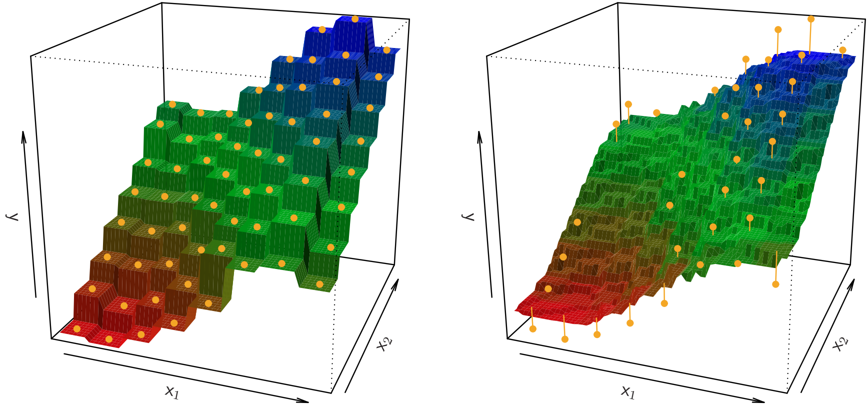
Non-parametric approach

- Linear regression is a parametric method.
- Non-parametric methods make no (strong) a priori assumptions on functional form of model $Y \approx f(X)$, more flexibility in adapting to data.
- Here: **K-nearest neighbors (KNN) regression** (Cf. KNN classifier in Chapter 2).
- Given prediction point x_0, first determine set $\mathscr{N}_0$ consisting of K ($K \in \mathbb{N}$) training observations closest to x_0.
- Predict $\hat{y}_0$ to be average training response in $\mathscr{N}_0$, i.e.,

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathscr{N}_0} y_i.$$

Linear Regression vs. K-Nearest Neighbors

Non-parametric approach



Two KNN fits on a data set with 64 observations using $p = 2$ predictors.
Left: $K = 1$. Interpolation, rough step-like function.
Right: $K = 9$. Not interpolatory, smoother.

- Flexibility of model controlled by K: less flexible. smoother fit, for large K.
- Bias-variance tradeoff.
- Flexible model: low bias, high variance
  (prediction depends on only one nearby observation).
  Unflexible model: high bias, low variance (changing one observation has smaller effect, averaging introduces bias).
- Optimal value of K? (later)

Q: In what setting will a parametric approach outperform a non-parametric approach?

Linear Regression vs. K-Nearest Neighbors

Parametric vs. non-parametric

Q: In what setting will a parametric approach outperform a non-parametric approach?

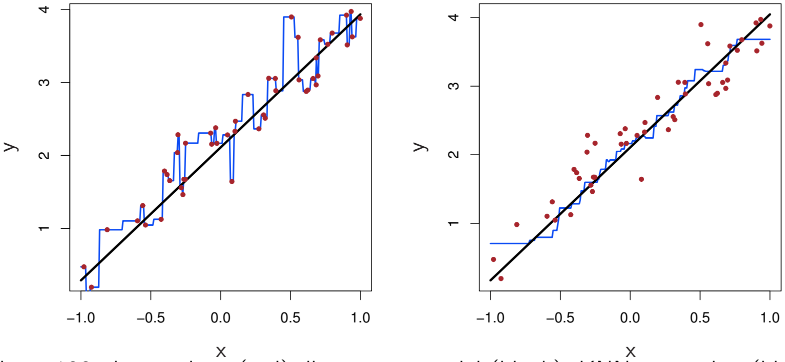
A: Depends on how closely assumed form of f matches true form.

Linear Regression vs. K-Nearest Neighbors

Parametric vs. non-parametric

Q: In what setting will a parametric approach outperform a non-parametric approach?

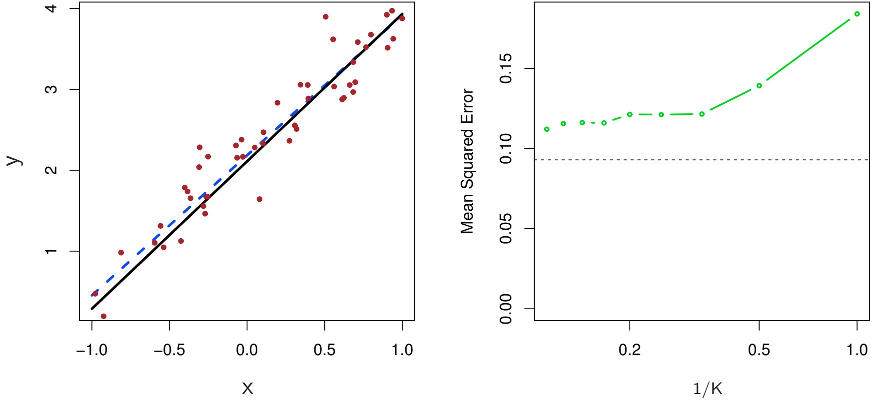
A: Depends on how closely assumed form of f matches true form.



1D data, 100 observations (red), linear true model (black), KNN regression (blue).
Left: $K = 1$, right: $K = 9$.

Linear Regression vs. K-Nearest Neighbors

Parametric vs. non-parametric

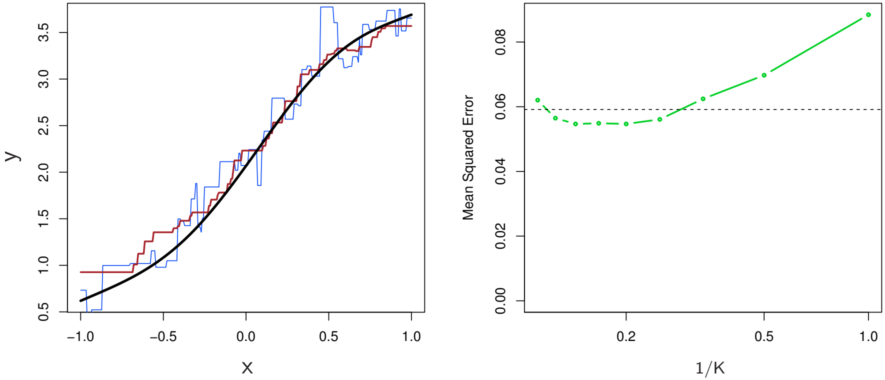


Left: same data, linear regression fit.

Right: test set MSE for linear regression (dotted line) and KNN for different values of K (plotted against $1/K$).

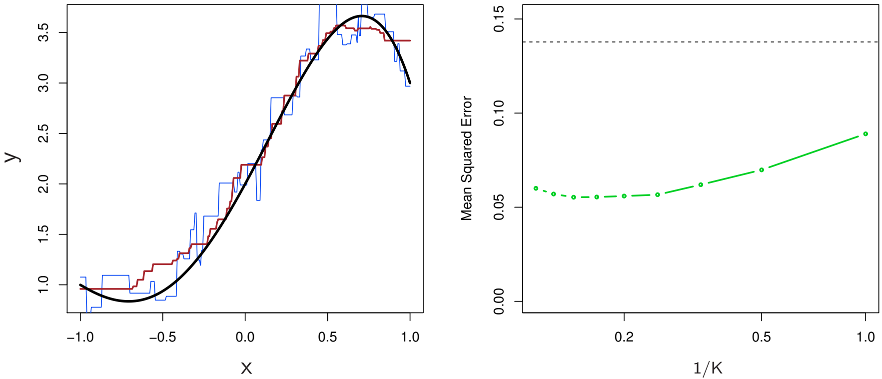
Linear Regression vs. K-Nearest Neighbors

Parametric vs. non-parametric



Left: slightly nonlinear data, true model (black), KNN regression with $K = 1$ (blue) and $K = 9$ (red).

Right: test set MSE for linear regression (dotted line) and KNN (against $1/K$). KNN wins for $K \geq 4$.

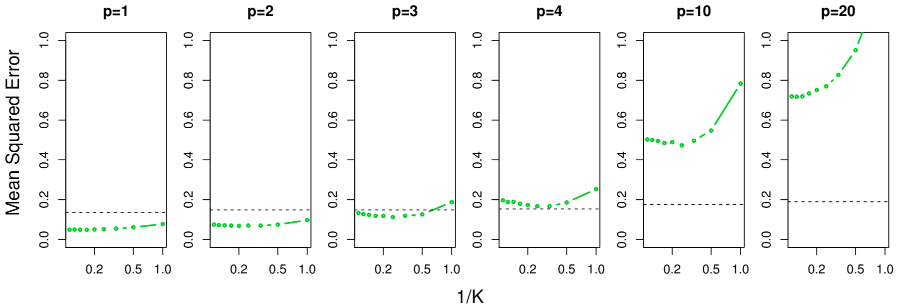


Left: stronly nonlinear data, true model (black), KNN regression with $K = 1$ (blue) and $K = 9$ (red).

Right: test set MSE for linear regression (dotted line) and KNN (against $1/K$). KNN wins for all K displayed.

Linear Regression vs. K-Nearest Neighbors

Parametric vs. non-parametric



Strongly nonlinear case, added noise predictors not associated with response. Linear regression MSE deteriorates only slightly as p rises, KNN regression MSE much more sensitive.

- For $p = 1$ KNN seems at most slightly worse than linear regression. For $p > 1$ this is no longer true.
- **Curse of dimensionality**: for $p = 20$, many of the 100 observations have no nearby observations.

Linear Regression vs. K-Nearest Neighbors

Parametric vs. non-parametric

- General rule: parametric methods tend to outperform non-parametric methods when there is a small number of observations per predictor.
- Even for small p, parametric methods offer the added advantage of better interpretability.