# Linear Regression Models

Basics

Source: Fields (2018)

# Aims

- Understand the linear model and its assumptions
- Understand how we assess the fit of the model
- Understand how we interpret model parameters
- Understand how to assess the generalizability of the model
- Fit and interpret linear models using Python

Source: Fields (2018)

# What is the linear model?

Part 1

Source: Fields (2018)

# What is the linear model?

- A way of predicting the value of one variable from another.
  - It is a hypothetical model of the relationship between two variables.
  - The model used is a linear one.
  - Therefore, we describe the relationship using the equation of a straight line.

Source: Fields (2018)

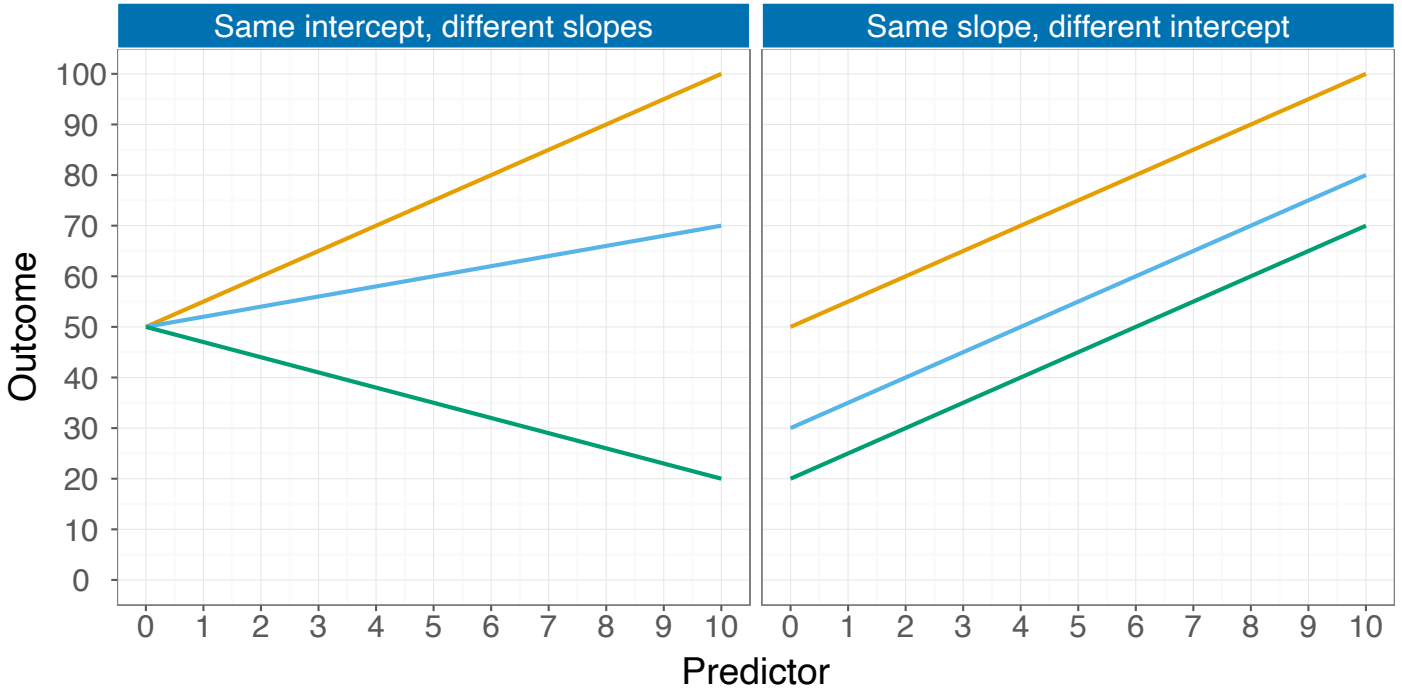
# Describing a straight line

$$y_i = (b_0 + b_1 X_i) + \varepsilon_i$$

- $b_i$
  - Coefficient for the predictor
  - Gradient (slope) of the line
  - Direction/strength of relationship
- $b_0$
  - Intercept (value of *Y* when *X* = 0)
  - Point at which the regression line crosses the *Y*-axis (ordinate)

Source: Fields (2018)

# Intercepts and gradients




Source: Fields (2018)

# A musical example

- A record company boss was interested in predicting album sales from advertising.
- Data
  - 200 different album releases
- Outcome variable:
  - Sales (CDs and Downloads) in the week after release
- Predictor variables
  - The amount (in £s) spent promoting the album before release
  - Number of plays on the radio
  - Image of the band

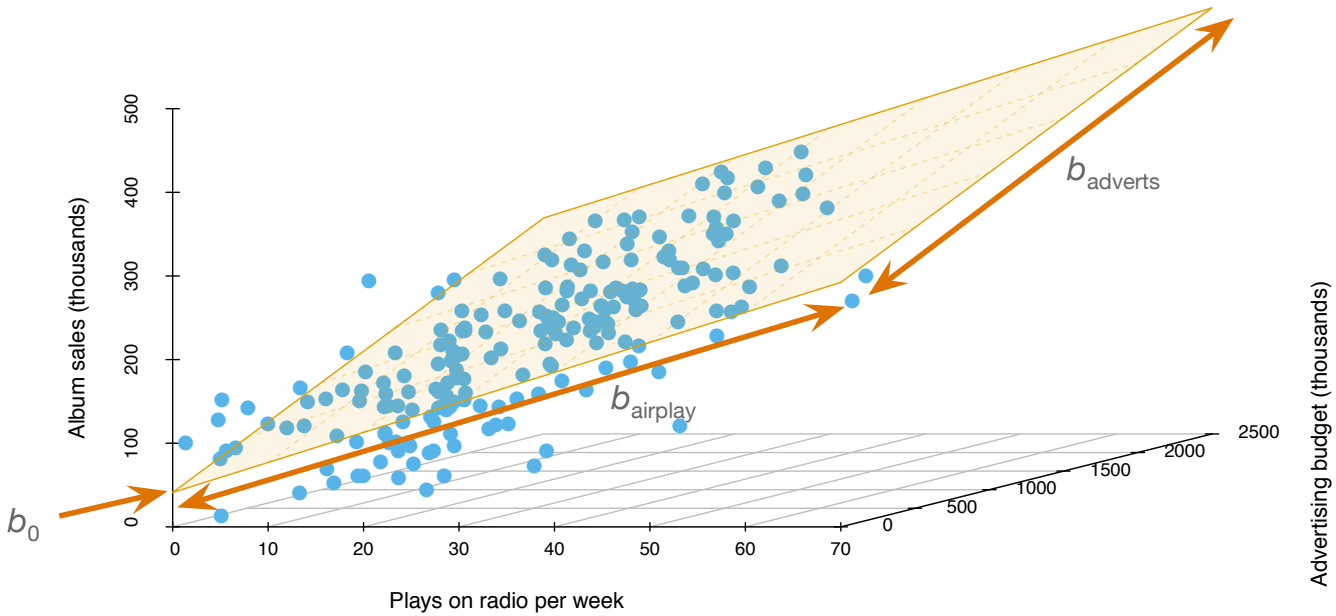
Source: Fields (2018)

# Linear model with one predictor


**FIGURE 8.12** Scatterplot showing the relationship between album sales and the amount spent promoting the album

Source: Fields (2018)

# Linear model with two predictors




Source: Fields (2018)

# The model as an equation

- With several predictors the model is described using a variation of the equation of a straight line.

$$Y_i = (b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_n X_{ni}) + \varepsilon_i$$

Source: Fields (2018)

# $b_0$

- $b_0$ is the intercept.
- The intercept is the value of the *Y* variable when all *X*s = 0.
- This is the point at which the model plane crosses the *Y*-axis (vertical).

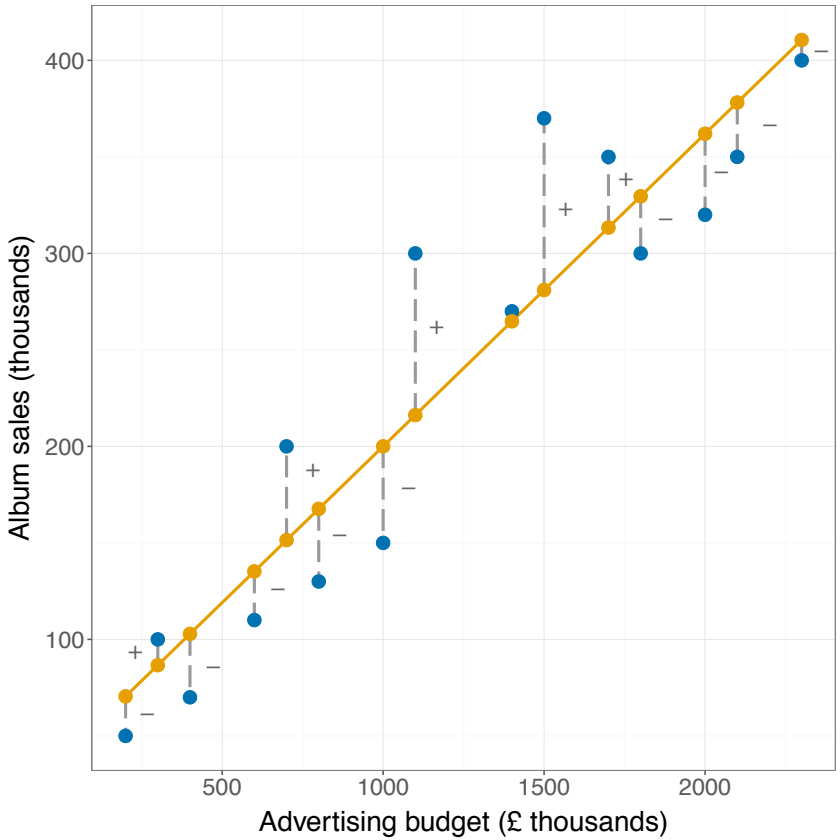
Source: Fields (2018)

# Beta values

- $b_1$ is the coefficient for variable 1.
- $b_2$ is the coefficient for variable 2.
- $b_n$ is the coefficient for $n^{th}$ variable.

Source: Fields (2018)

# Estimation: the method of least squares




$$\text{Total error} = \sum_{i=1}^{n} (\text{observed}_i - \text{model}_i)^2$$

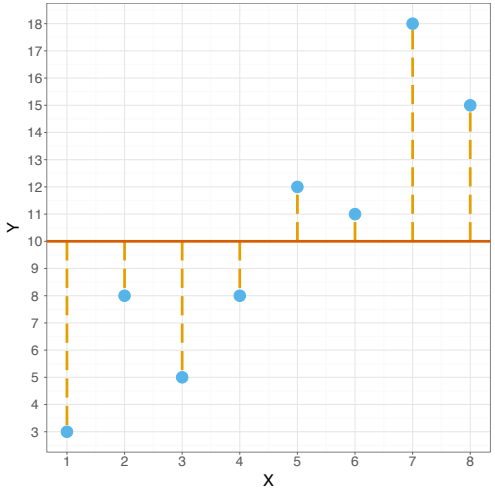
Source: Fields (2018)

# The fit of the model?

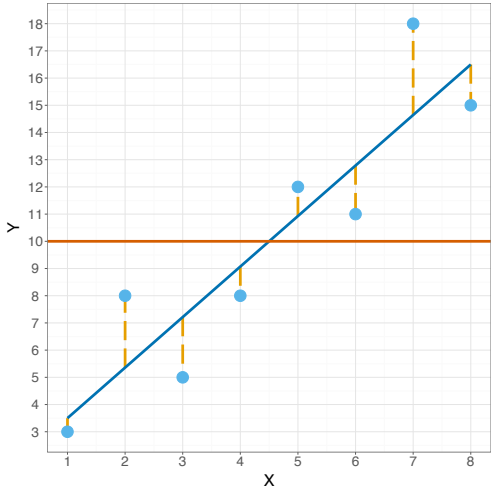
- The model based on the data might not reflect reality.
  - We need some way of testing how well the model fits the observed data.
  - How?
    - $F$
    - $R^2$

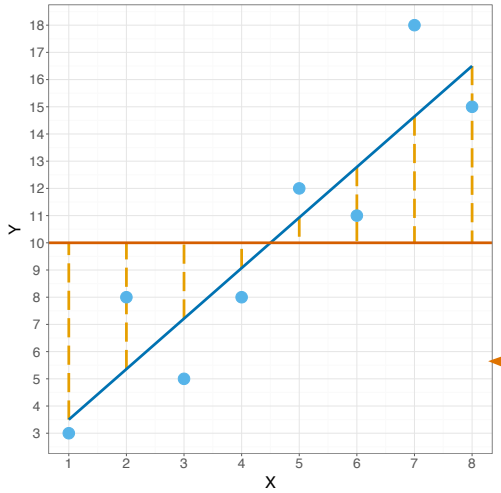
Source: Fields (2018)

$SS_T$ uses the differences between the observed data and the mean value of *Y*




$SS_R$ uses the differences between the observed data and the model




$SS_M$ uses the differences between the mean value of *Y* and the model

# Sums of squares

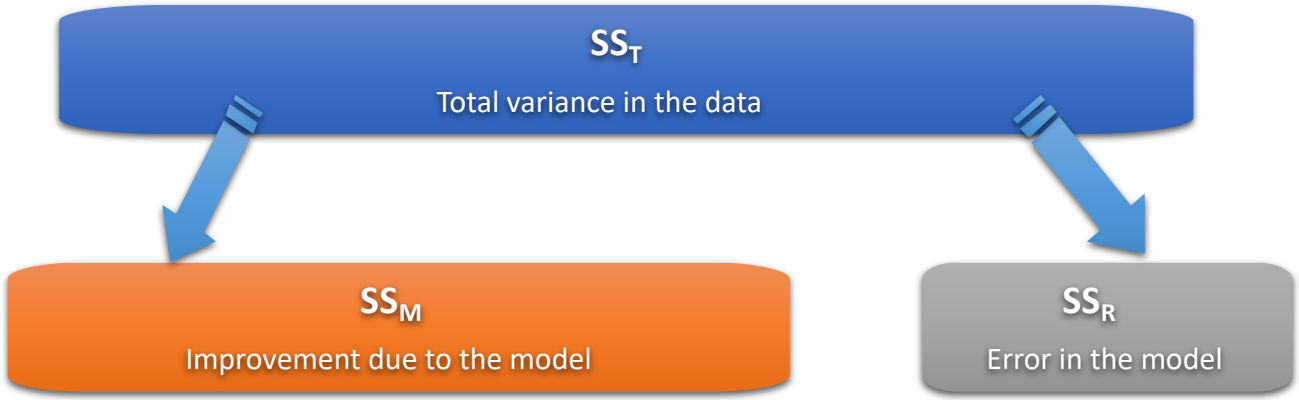
Source: Fields (2018)

# Summary

- $SS_T$
  - Total variability (variability between scores and the mean).
- $SS_R$
  - Residual/error variability (variability between the model and the actual data).
- $SS_M$
  - Model variability (difference in variability between the model and the mean).

Source: Fields (2018)

# Testing the fit: *F*-statistic




- If the model results in better prediction than using the mean, then we expect $SS_M$ to be much greater than $SS_R$

Source: Fields (2018)

# Testing the fit: *F*-statistic

- Mean Squared Error
  - Sums of Squares are total values.
  - They can be expressed as averages.
  - These are called Mean Squares, MS

$$F = \frac{MS_M}{MS_R}$$

**$SS_M$**
Improvement due to the model

**$SS_R$**
Error in the model

Source: Fields (2018)

# *F*-statistic

- Looks at whether the variance explained by the model ($SS_M$) is significantly greater than the error within the model ($SS_R$).
- It tells us whether using the regression model is significantly better at predicting values of the outcome than using the mean.

Source: Fields (2018)

# Testing the fit: $R^2$

- $R^2$
  - The proportion of variance accounted for by the regression model.
  - The Pearson correlation coefficient squared

$$R^2 = \frac{SS_M}{SS_T}$$

Source: Fields (2018)

# $R$ and $R^2$

- $R$
  - The correlation between the observed values of the outcome, and the values predicted by the model.
- $R^2$
  - The proportion of variance accounted for by the model.
- Adj. $R^2$
  - An estimate of $R^2$ in the population (*shrinkage*).

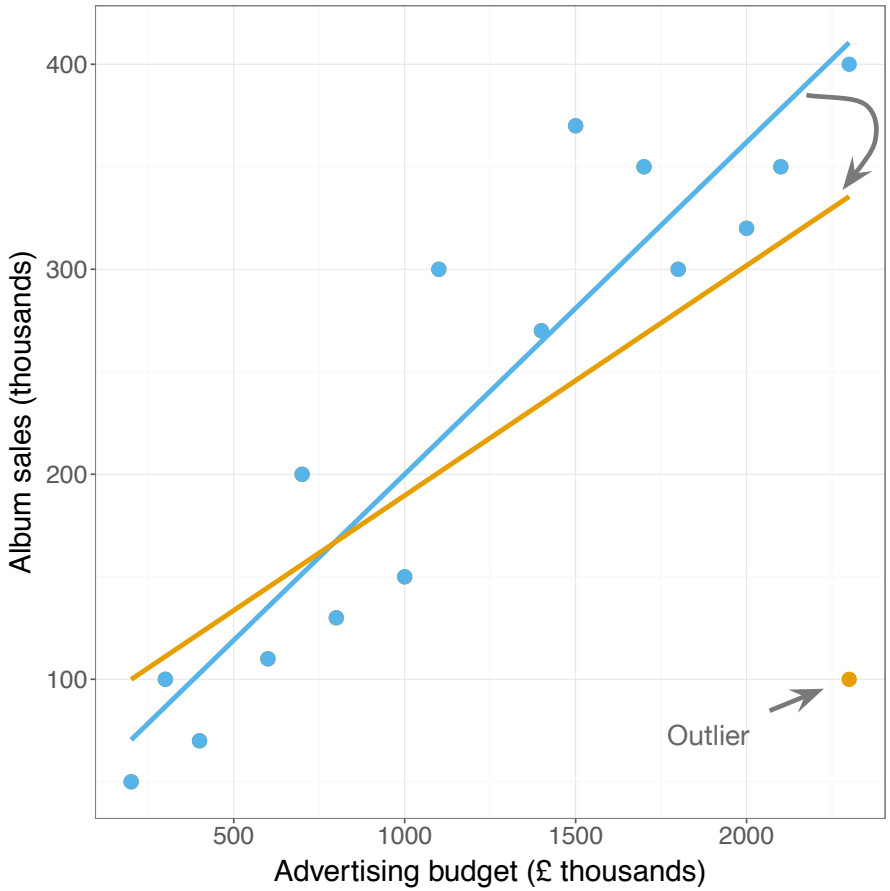
Source: Fields (2018)

# Bias and generalization

Part 3

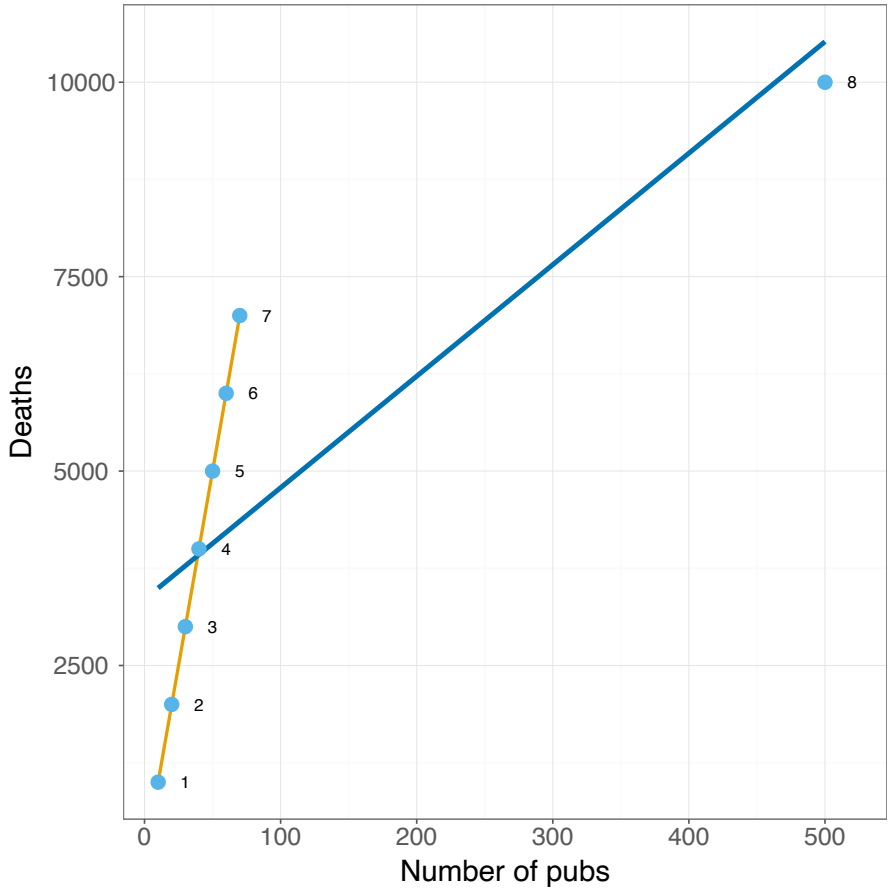
Source: Fields (2018)

# Bias: outliers




Source: Fields (2018)

# Bias: influential cases




Source: Fields (2018)

# How well does the model fit the data?

- There are two ways to assess the accuracy of the model in the sample:
- Residual statistics
  - Standardized residuals
- Influential cases
  - Cook's distance

Source: Fields (2018)

# Standardized residuals

- In an average sample, 95% of standardized residuals should lie between ± 2.
- 99% of standardized residuals should lie between ± 2.5.
- Outliers
  - Any case for which the absolute value of the standardized residual is 3 or more, is likely to be an outlier.

Source: Fields (2018)

# Cook's distance

- Measures the influence of a single case on the model as a whole.
- Weisberg (1982):
  - **Absolute values greater than 1 may be cause for concern.**

Source: Fields (2018)

# Generalizing the model

- We hope to be able to generalize the sample model to the entire population.
- To do this, several assumptions must be met
- Violating these assumptions stops us generalizing conclusions to our target population.

Source: Fields (2018)

# Straightforward assumptions

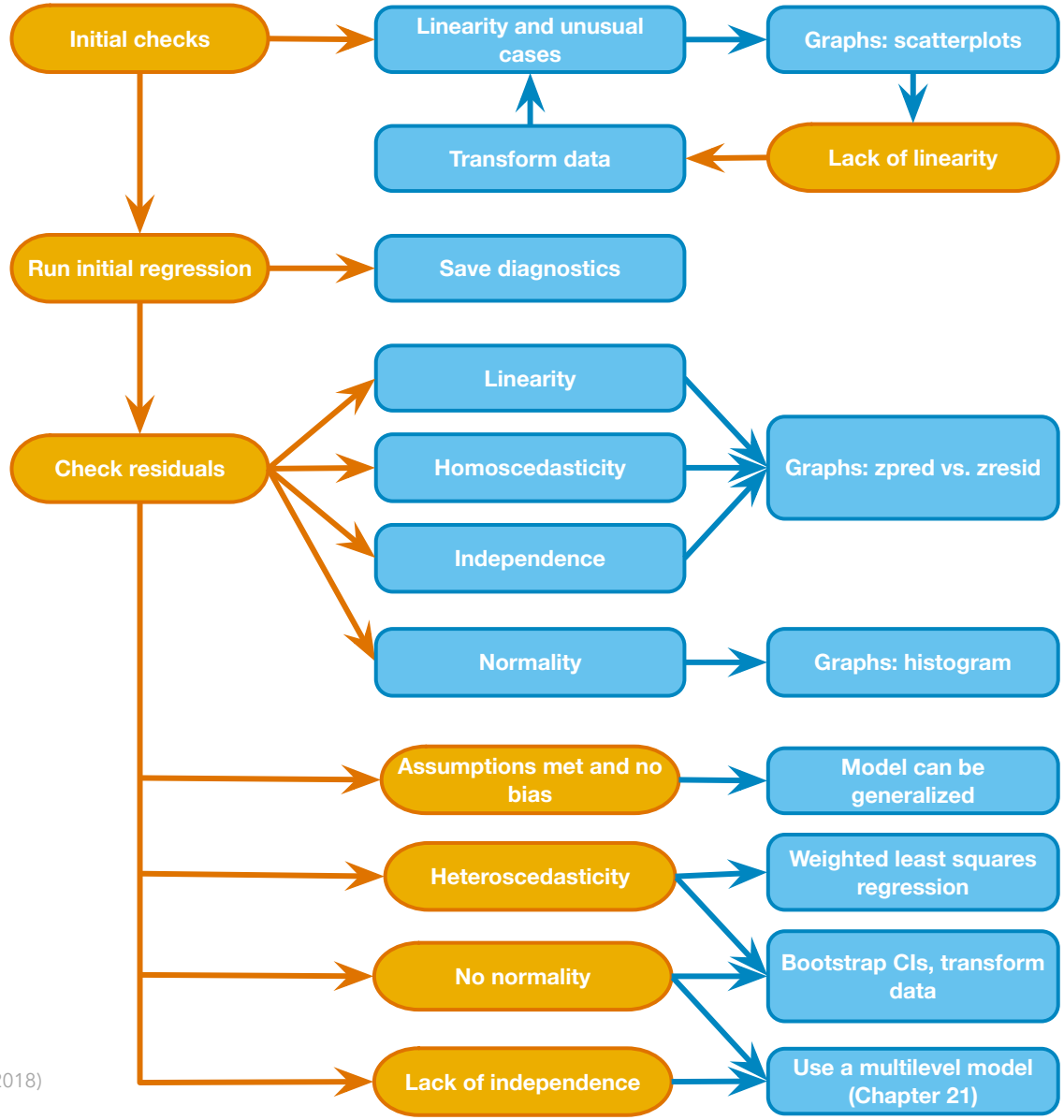
- Variable type:
  - Outcome must be continuous
  - Predictors can be continuous or dichotomous.
- Non-zero variance:
  - Predictors must not have zero variance.
- Linearity:
  - The relationship we model is, in reality, linear.
- Independence:
  - All values of the outcome should come from a different person.

Source: Fields (2018)

# Assumptions that matter

- Non-linearity of the response-predictor relationships
- Normally distributed errors (outliers).
- Correlation of error terms.
- Non-constant variance of error terms (heteroskedasticity).
- High-leverage points.
- Multicollinearity.

Source: Fields (2018)




Source: Fields (2018)

# The record sales example (recap)

- A record company boss was interested in predicting album sales from advertising.
- Data
  - 200 different album releases
- Outcome variable:
  - Sales (CDs and Downloads) in the week after release
- Predictor variables
  - The amount (in £s) spent promoting the album before release
  - Number of plays on the radio
  - Image of the band

Source: Fields (2018)

# model parameters

- *b*-values:
  - The change in the outcome associated with a unit change in the predictor.
- Standardised *b*-values :
  - Tell us the same but expressed as standard deviations.

Source: Fields (2018)

# Using the model

$$\text{album sales}_i = b_0 + b_1 \text{advertising budget}_i$$

$$= 134.14 + \left(0.096 \times \text{ advertising budget}_i\right)$$

$$\text{album sales}_i = 134.14 + \left(0.096 \times \text{ advertising budget}_i\right)$$

$$= 134.14 + \left(0.096 \times 100\right)$$

$$= 143.74$$

Source: Fields (2018)

# *b*-values

- **Advertising budget**: $b = 0.085$
  - As advertising budget increases by one unit, album sales increase by 0.085 units. Both variables were measured in thousands; therefore, for every £1000 more spent on advertising, an extra 0.085 thousand albums (85 albums) are sold. This interpretation is true only if the effects of band image and airplay are held constant.
- **Airplay**: $b = 3.367$
  - As the number of plays on radio in the week before release increases by one, album sales increase by 3.367 units. Every additional play of a song on radio (in the week before release) is associated with an extra 3.367 thousand albums (3367 albums) being sold. This interpretation is true only if the effects of the band's image and advertising budget are held constant.
- **Image**: $b = 11.086$
  - If a band can increase their image rating by 1 unit they can expect additional album sales of 11.086 units. Every unit increase in the band's image rating is associated with an extra 11.086 thousand albums (11,086 albums) being sold. This interpretation is true only if the effects of airplay and advertising are held constant.

Source: Fields (2018)