Making Predictions: Regression

Vanessa LoBue

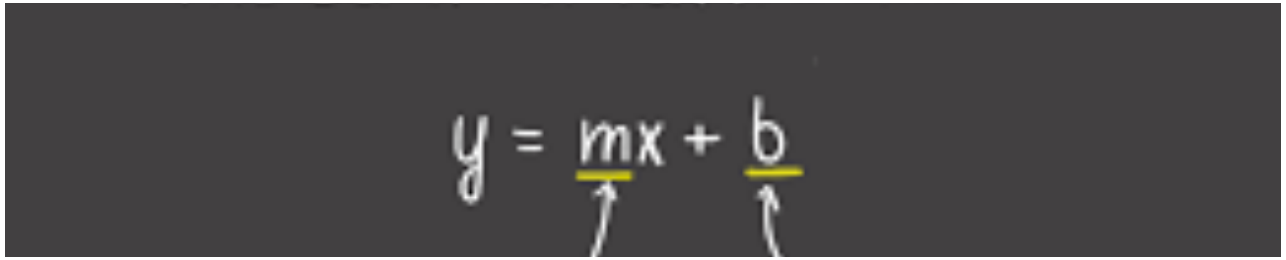
Jamil Bhanji

with a little help from Andy Field

Aims

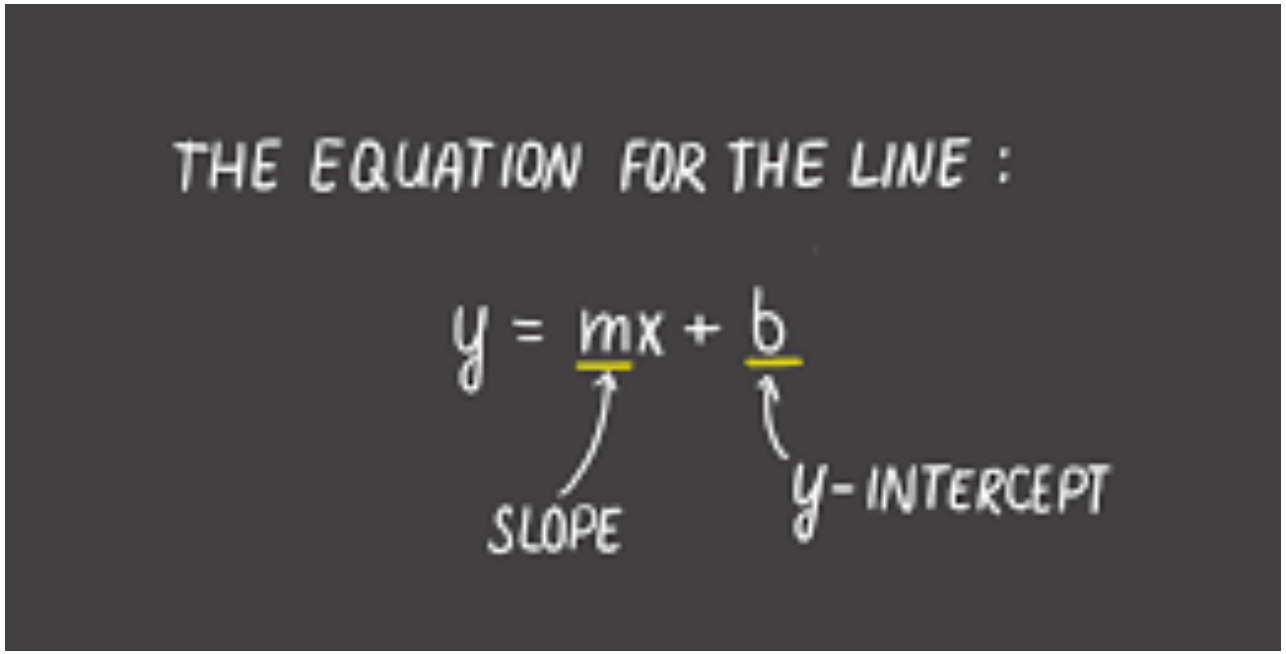
- Understand the linear model and its assumptions
- Understand how we assess the fit of the model
- Understand how we interpret model parameters
- Understand how to assess the generalizability of the model

Do you remember…?



Thank your math teachers!

Do you remember…?



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.

Describing a Straight Line

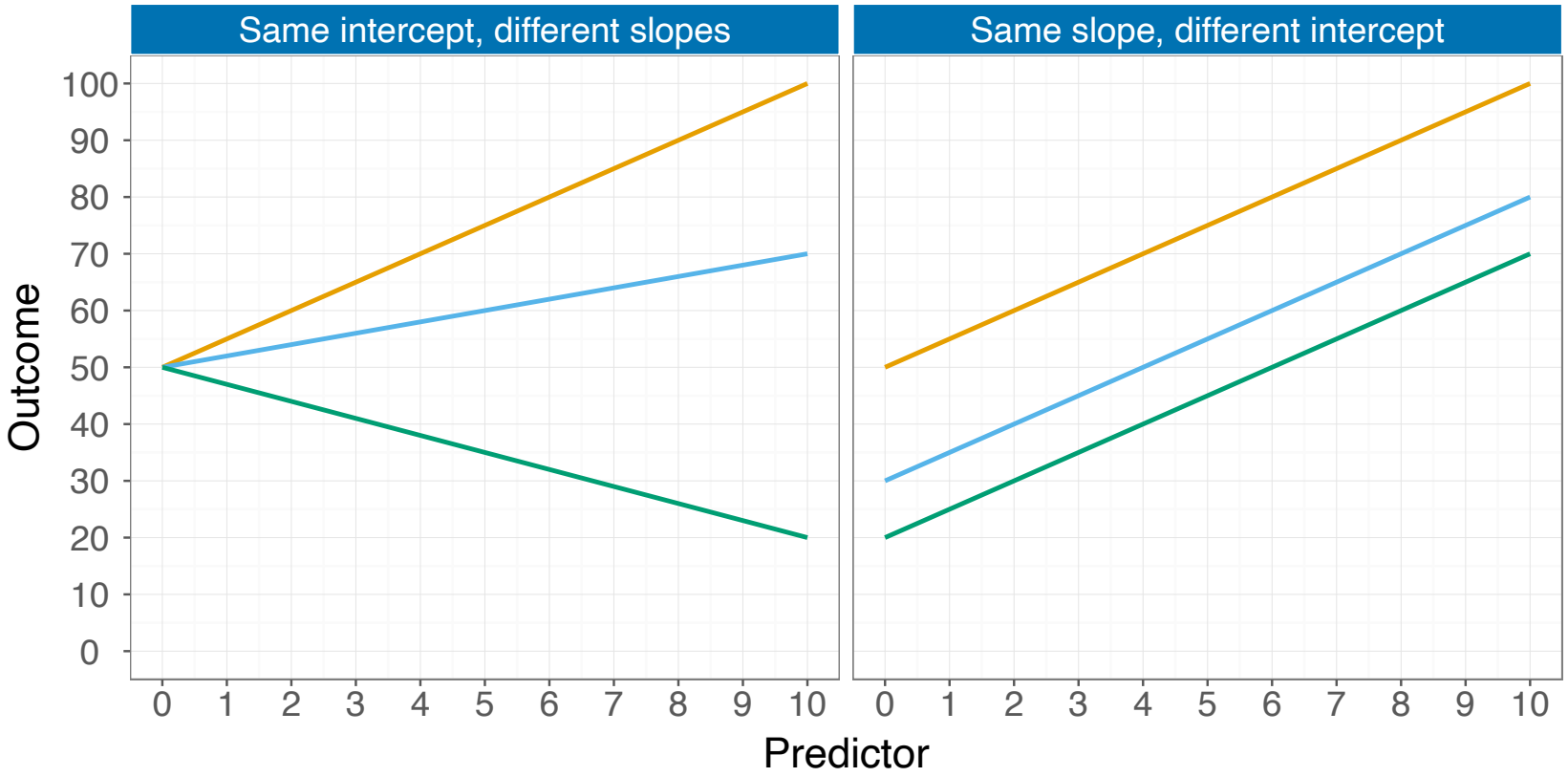
slope

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

y-intercept error

- 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)

Intercepts and Gradients



Linear Model with One Predictor

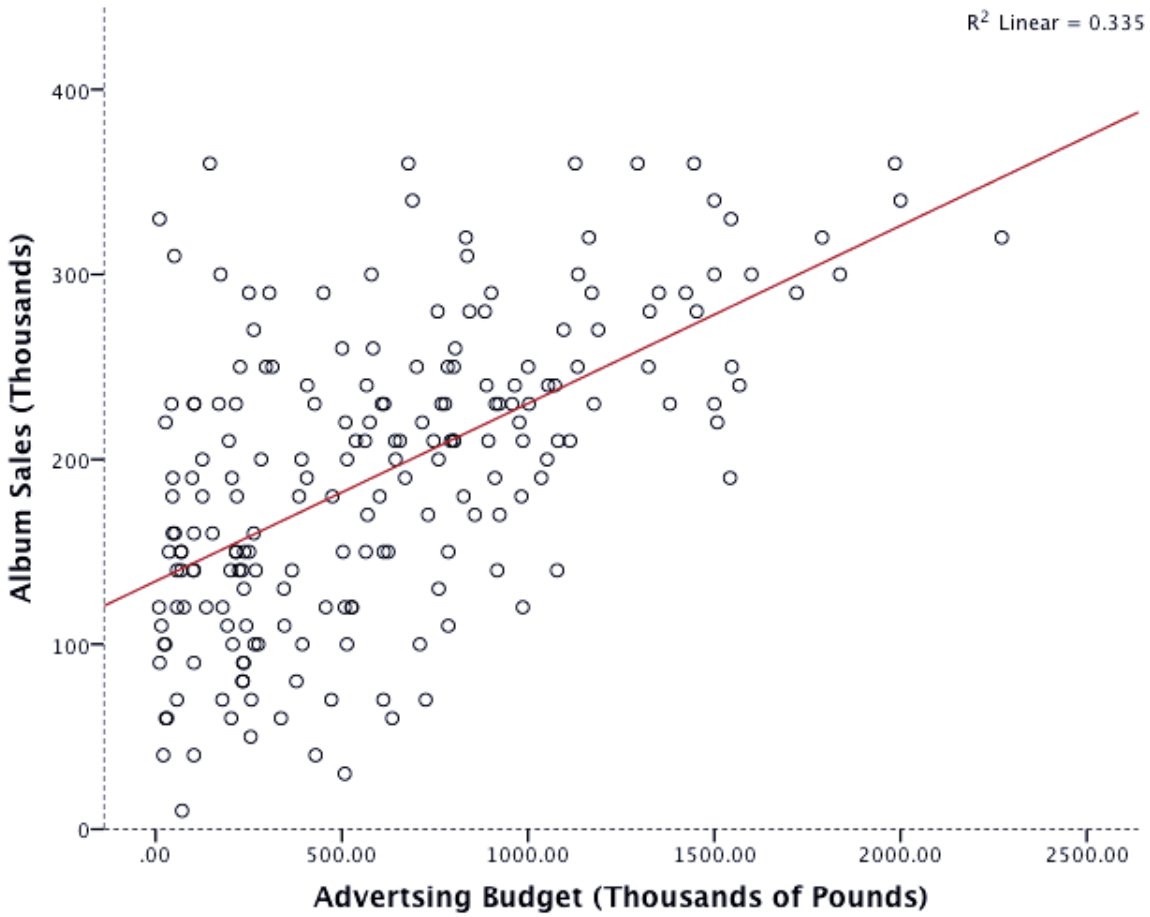


FIGURE 8.12 Scatterplot showing the relationship between album sales and the amount spent promoting the album

The Model as an Equation

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

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

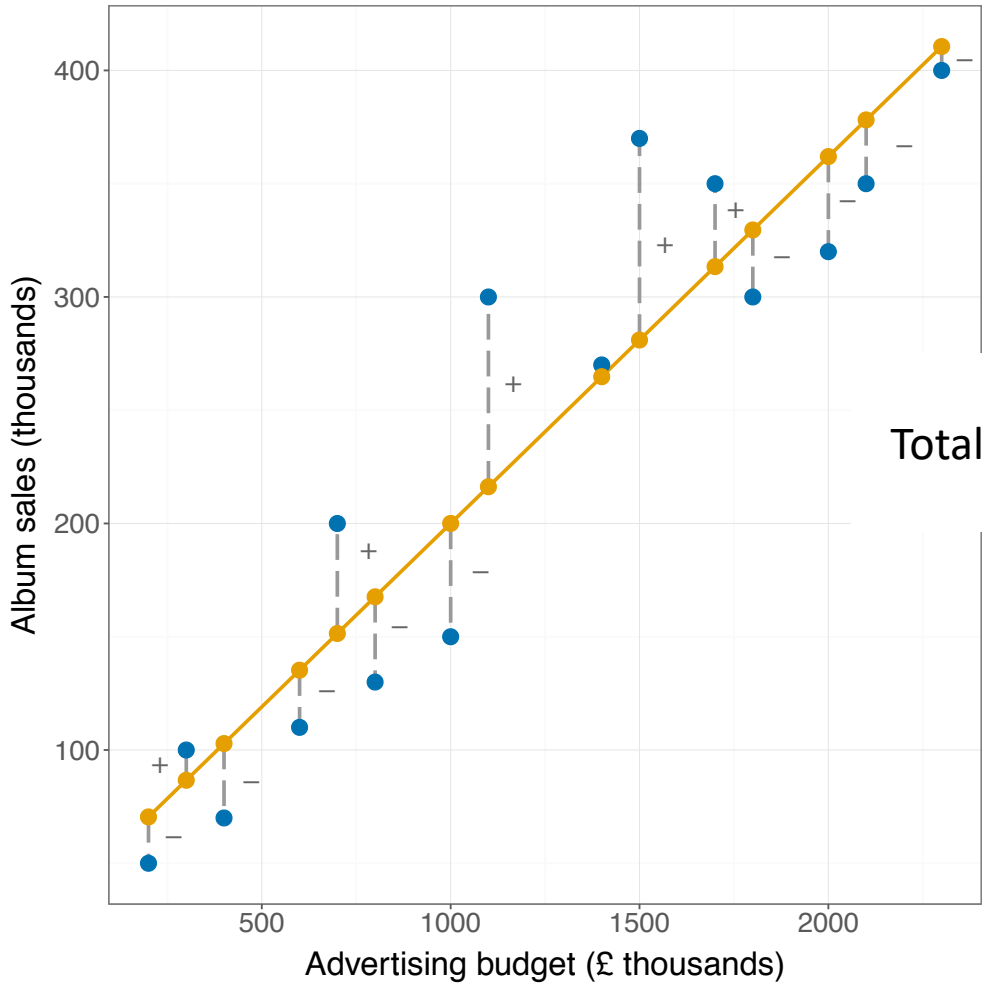
b_0

- 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).

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.

Error



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

The Fit of the Model?

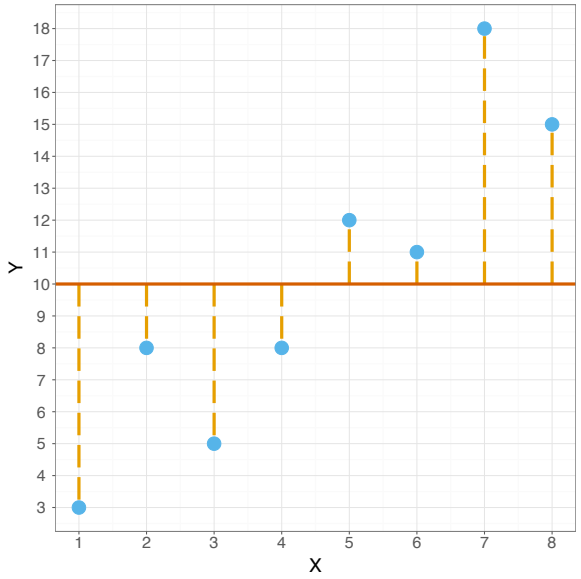
- The model based on the data might not reflect reality.
  - We need some way of testing how well the model fits the observed data.

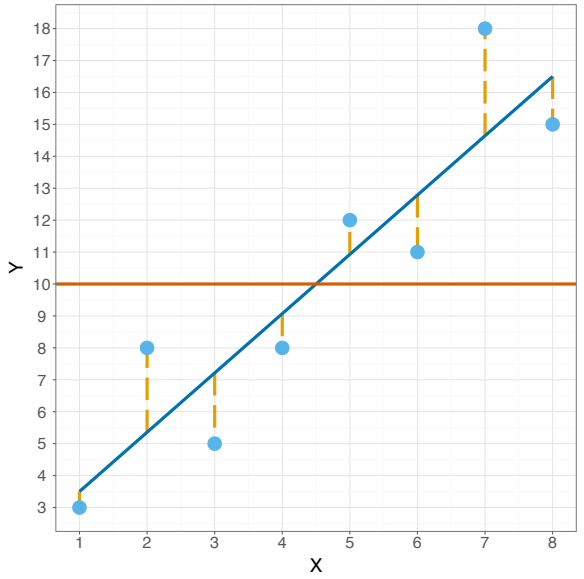
  - How?
    - F
    - R^2

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

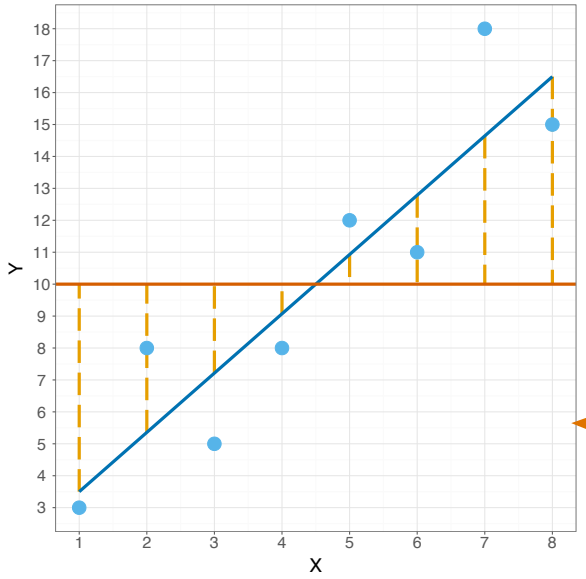
SS_T is in the ”Total” row in SPSS output





SS_R is in the ”Residual” row in SPSS output

SS_R is in the ”Regression” row in SPSS output



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

Sums of squares

Summary

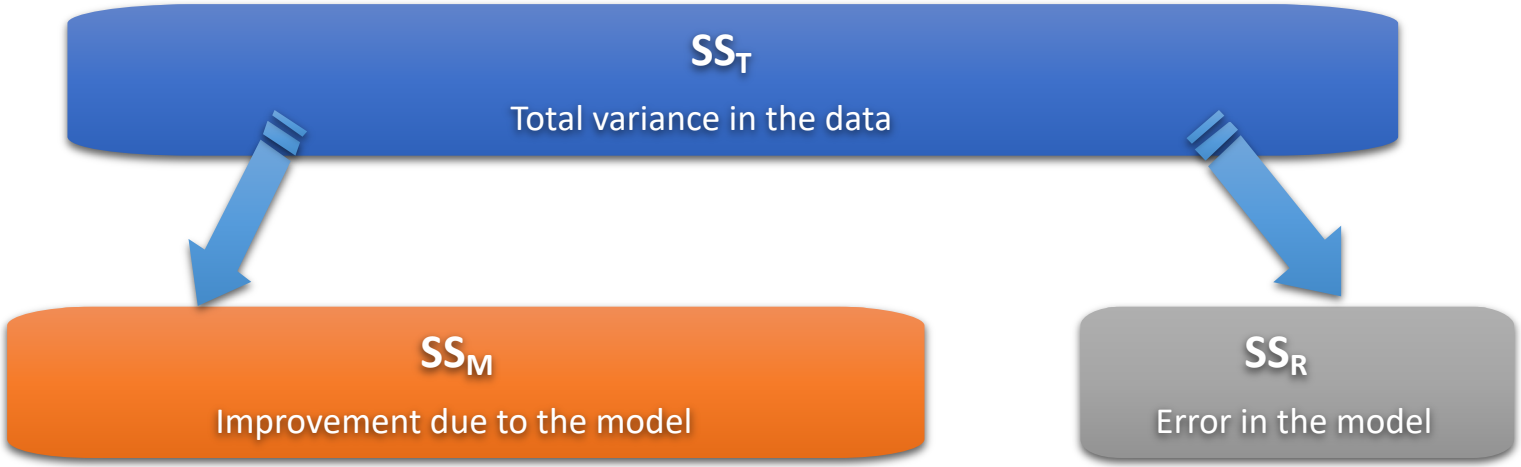
- SS_T
  - Total variability (variability between scores and the mean).
- SS_R
  - Residual/error variability (variability between the model and the actual data).
- SS_M
  - Model variability (difference in variability between the model and the mean).

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

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

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

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

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.

Testing the Fit: R^2

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

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

Individual Predictors

$$t = \frac{b_{\text{observed}} - b_{\text{expected}}}{SE_b} = \frac{b_{\text{observed}}}{SE_b}$$

Building the Model: Hierarchical Entry

- Known predictors (based on past research) are entered into the regression model first
- New predictors are then entered in a separate step/block
- Researcher makes the decisions

Hierarchical Entry

- It is the best method:
  - Based on theory testing.
  - You can see the unique predictive influence of a new variable on the outcome because known predictors are held constant in the model.

- Bad point:
  - It relies on the researcher knowing what they're doing!

Forced Entry

- All variables are entered into the model simultaneously.

- The results obtained depend on the variables entered into the model.
  - It is important, therefore, to have good theoretical reasons for including a particular variable.

Stepwise entry

- Variables are entered into the model based on mathematical criteria.

- Computer selects variables in steps.

Stepwise Entry

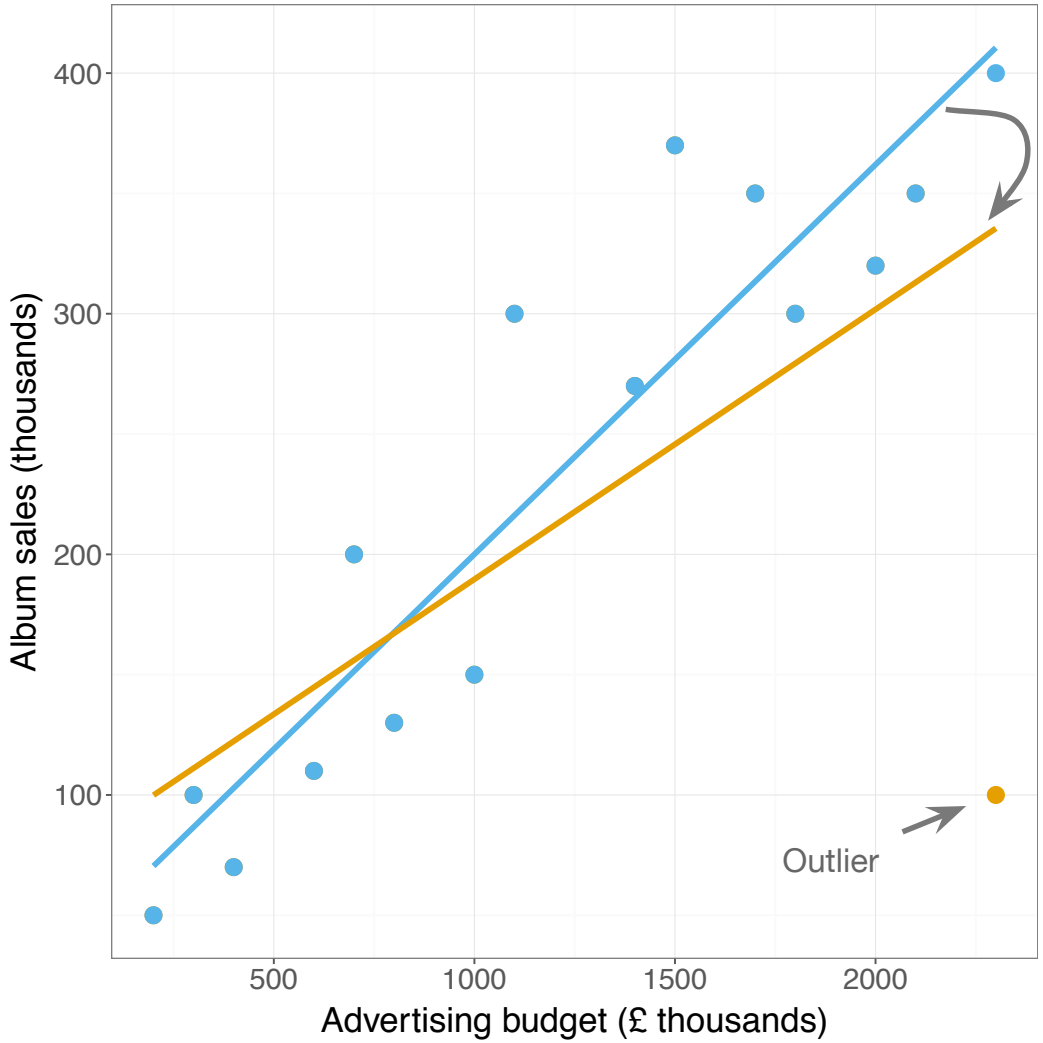
- Step 1
  - SPSS looks for the predictor that can explain the most variance in the outcome variable.

- Step 2:
  - Having selected the 1st predictor, a second one is chosen from the remaining predictors.
  - The *semi-partial correlation* (think back to the last lecture) is used as a criterion for selection.

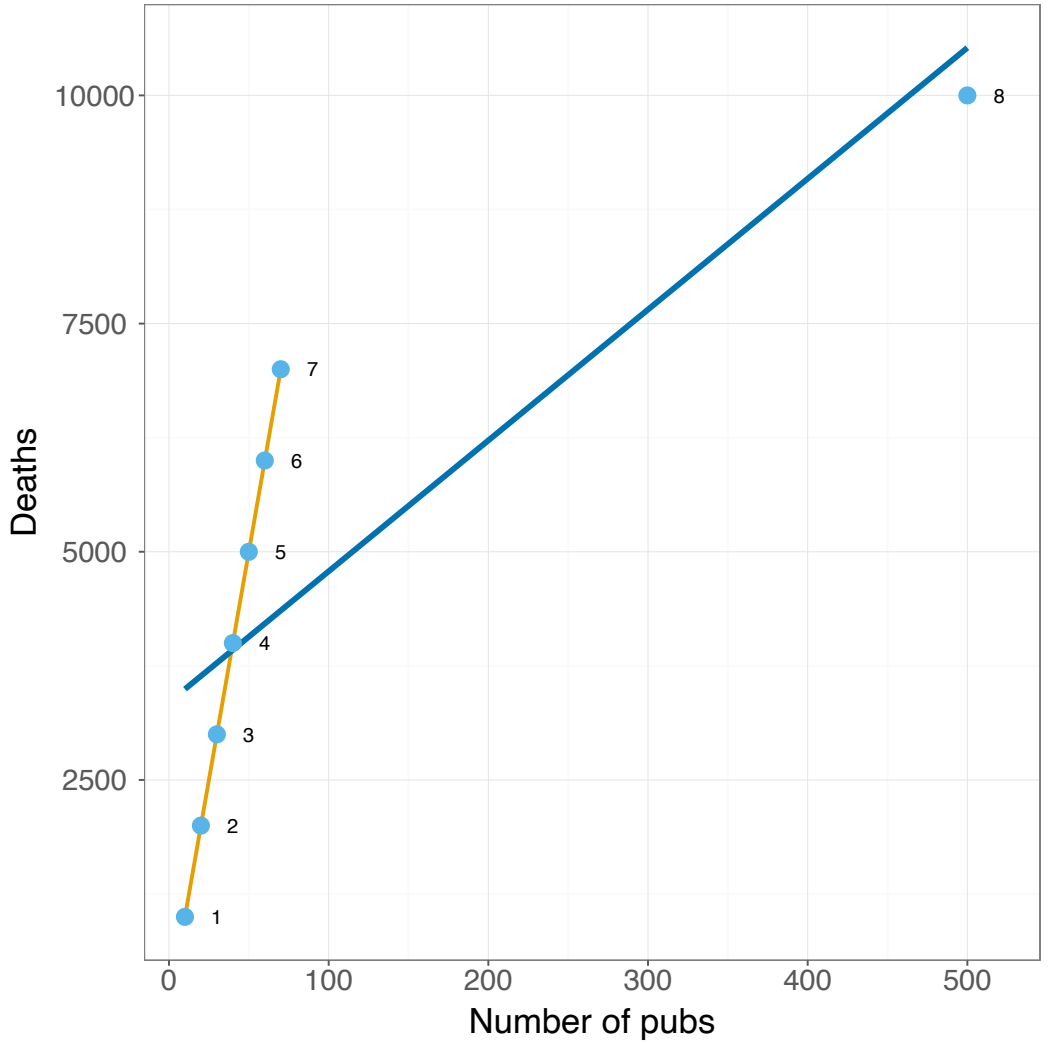
Problems with Stepwise Entry

- Reliance on a mathematical criterion
  - Variable selection may depend upon only slight differences in the semi-partial correlation
  - These slight numerical differences can lead to major theoretical differences
- Should be used only for exploration

Bias: Outliers



Bias: Outliers & Influential Cases



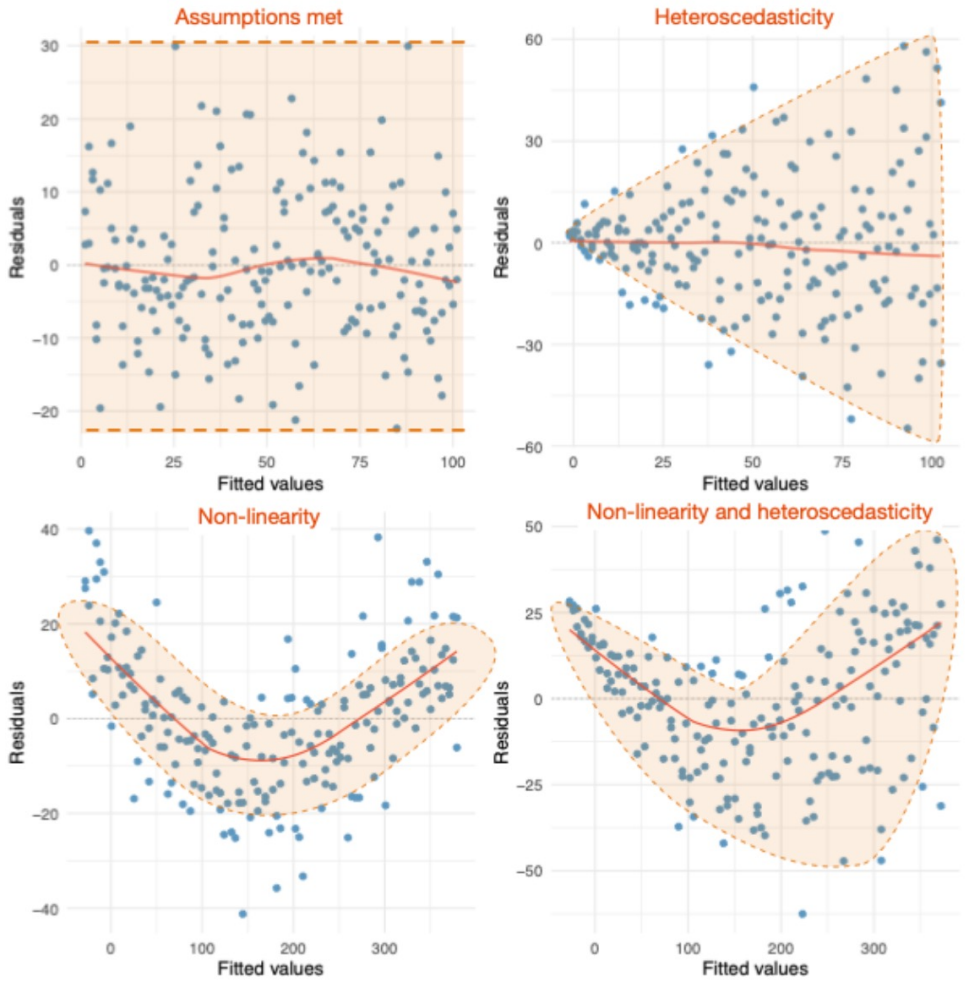
Generalizing the model

- 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.

Assumptions (a reminder…)

- Additivity and linearity
- Spherical residuals
  - Homoscedasticity: For each value of the predictors the variance of the error term should be constant
  - Independent residuals: for any pair of observations, the error terms should be uncorrelated.

Assumptions (a reminder…)



Assumptions (a reminder…)

- Additivity and linearity
- Spherical residuals
  - Homoscedasticity: For each value of the predictors the variance of the error term should be constant
  - Independent residuals: for any pair of observations, the error terms should be uncorrelated.
- Normally-distributed errors
- No Multicollinearity:
  - Predictors must not be highly correlated.

Multicollinearity

- Multicollinearity exists if predictors are highly correlated.
- This assumption can be checked with collinearity diagnostics.

Interpreting Model Parameters

- *b*-values:
  - The expected change in the outcome associated with a unit change in the predictor.

- Standardised *b*-values:
  - Tell us the same but expressed as standard deviations.

Logistic Regression

- To predict an outcome variable that is categorical from one or more categorical or continuous predictor variables
- Used because having a categorical outcome variable violates the assumption of linearity in normal regression
- Example: Explain whether a person chooses/refuses a tutoring session, using their GPA and beliefs about intelligence (e.g., growth/fixed)

Interpreting Logistic Regresssion Model Parameters

- *b*-values:
  - The expected change in **the log(odds) of the outcome** associated with a unit change in the predictor.
- exponentiate the coefficient (e^b) to calculate the odds ratio.
  - e.g., odds ratio of 1.2 indicates that a one unit increase in the predictor means the odds of the outcome are 1.2 times the odds of the outcome without the increase in the predictor

Multinomial Logistic Regression

- Logistic regression to predict membership of more than two categories.
- It (basically) works in the same way as binary logistic regression.
- The analysis breaks the outcome variable down into a series of comparisons between two categories.
  - E.g., if you have three outcome categories (A, B and C), then the analysis will consist of two comparisons that you choose:
    - Compare everything against your first category (e.g. A vs. B and A vs. C),
    - Or your last category (e.g. A vs. C and B vs. C),
    - Or a custom category (e.g. B vs. A and B vs. C).