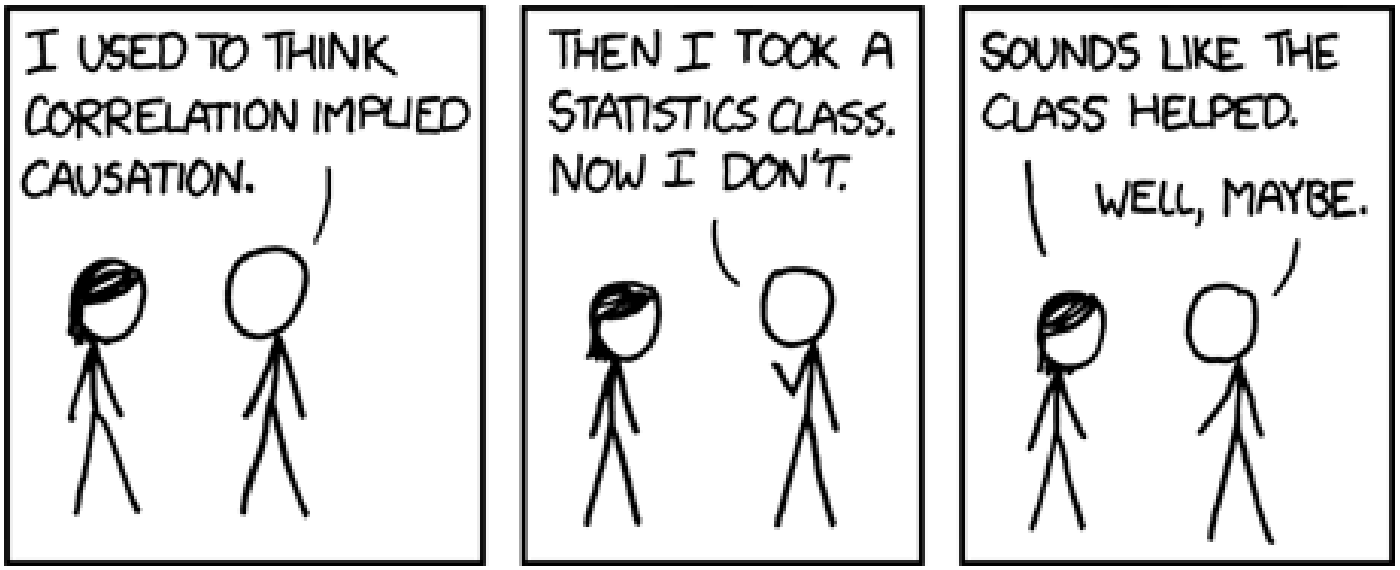
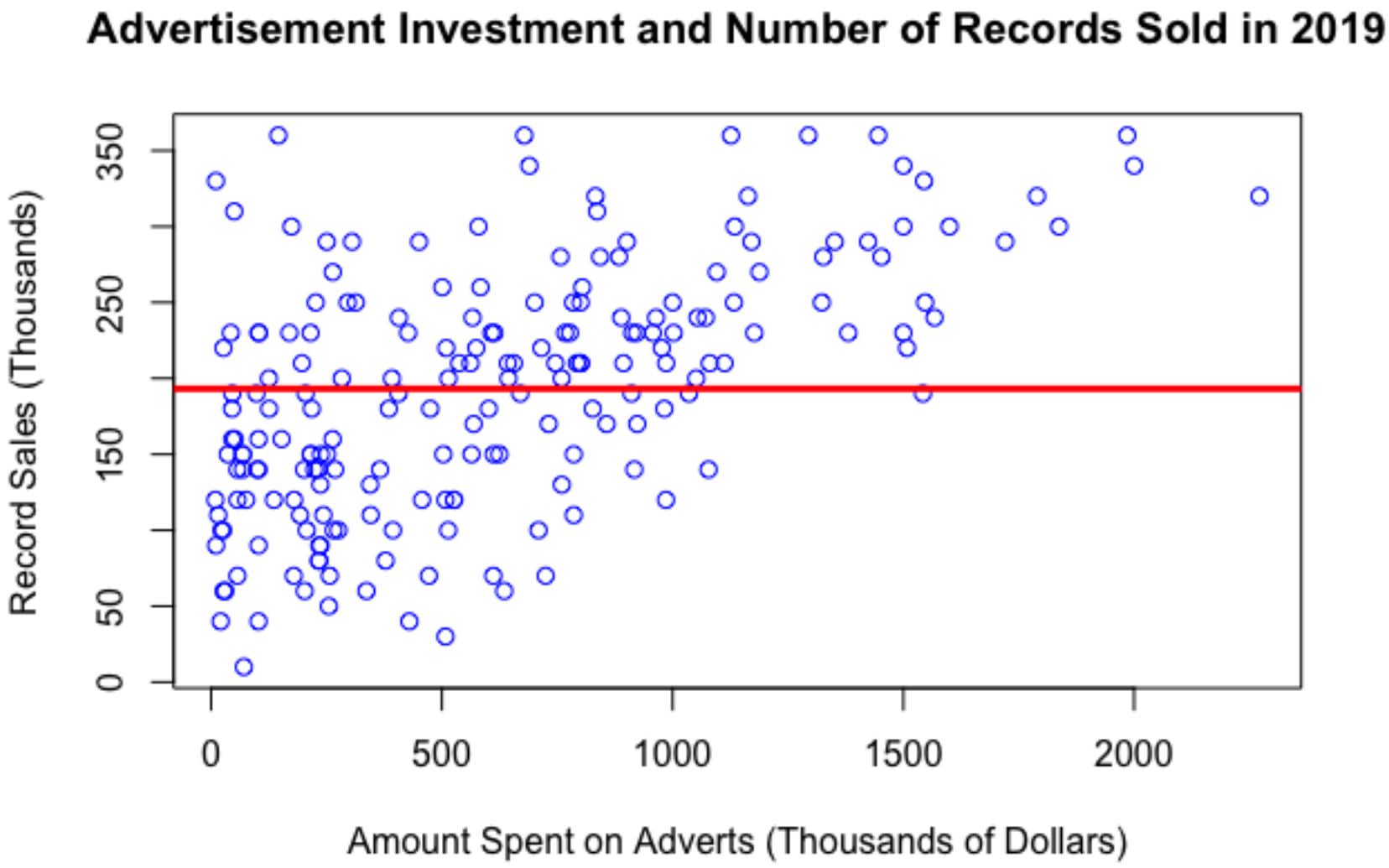
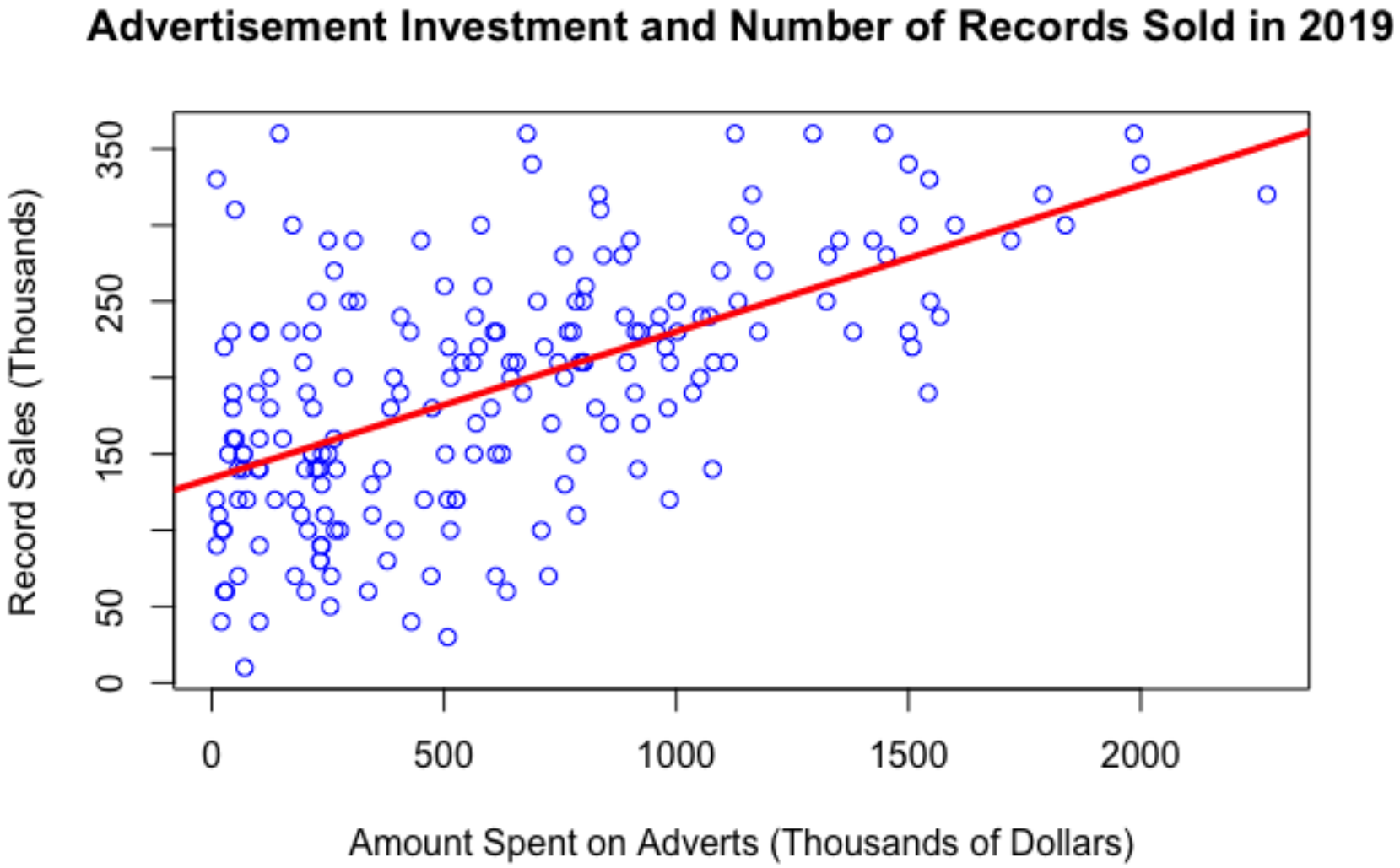
# Linear Models – Multiple Regression



# The mean:
# A very simple statistical model



# The method of least squares



# Constructing Simple Regressions in R

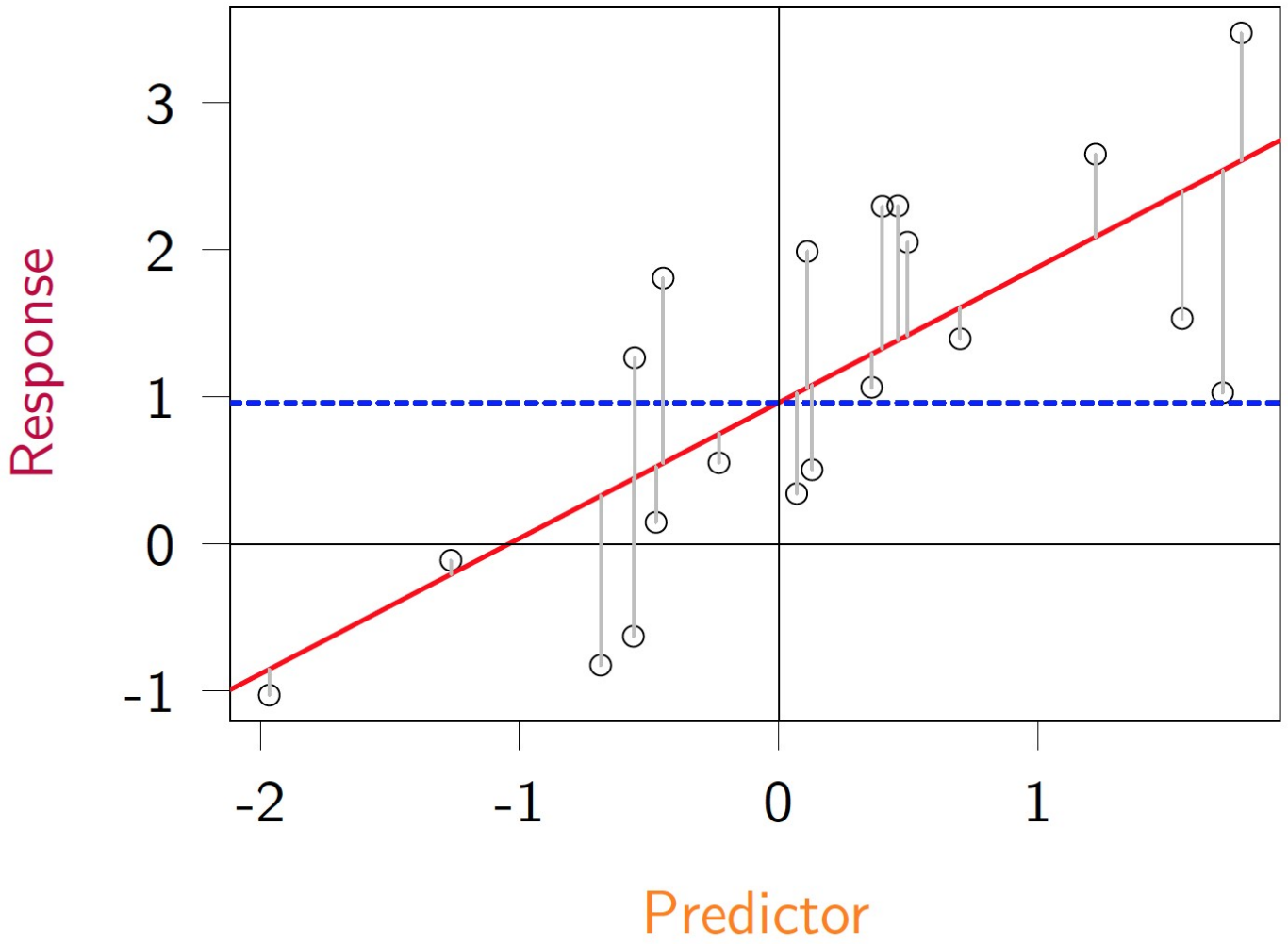
```
album_lm_1 <- lm(sales ~ 1 + adverts, data = album_data)
```

```
summary(album_lm_1)
```

```
Call:
lm(formula = sales ~ 1 + adverts, data = album_data)

Residuals:
     Min       1Q   Median       3Q      Max
-152.949  -43.796   -0.393   37.040  211.866

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.341e+02  7.537e+00  17.799   <2e-16 ***
adverts     9.612e-02  9.632e-03   9.979   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 65.99 on 198 degrees of freedom
Multiple R-squared:  0.3346,    Adjusted R-squared:  0.3313
F-statistic: 99.59 on 1 and 198 DF,  p-value: < 2.2e-16
```

# Linear Regression

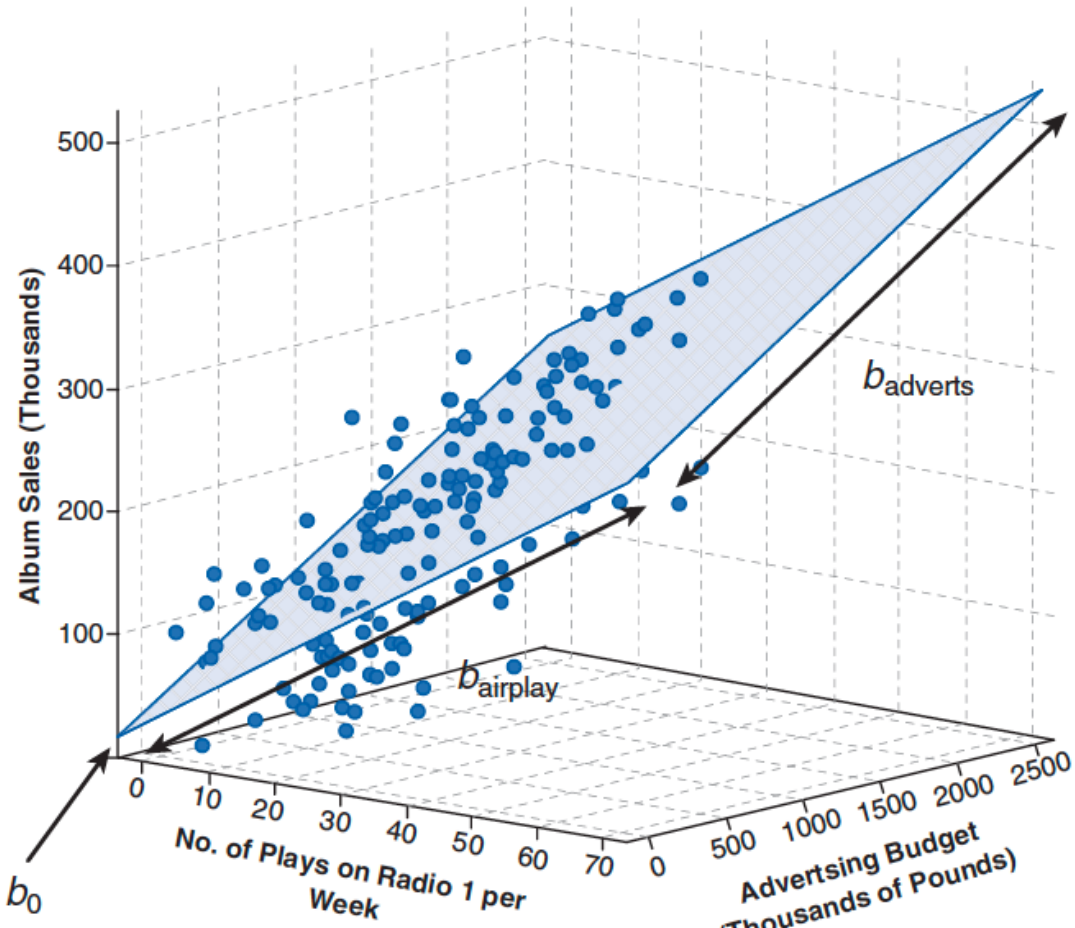
**Response** = **Intercept** + **Slope** × **Predictor** + **Error**




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

# Multiple Regression

**Response** = **Intercept** + **Slope1** × **Predictor1** + **Slope2** × **Predictor2** + **Error**




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

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

# Selecting Predictors

- The values of the regression coefficients depend upon the variables in the model.
- Therefore, the predictors included and the way in which they are entered into the model can have a great impact.
- Predictors should be based on past research (i.e. theoretical importance).

# Methods of regression

Theoretically based (Theory testing):

- Hierarchical
- Predictors are selected based on past research.
- Order to enter predictors is determined by the experimenter.

- Forced Entry
- Predictors are selected on past research.
- All predictors are forced into the model simultaneously.

Mathematically based (Exploratory analysis):

- Stepwise methods
- Decisions about the order in which predictors are entered into the model are based on purely mathematical criterion.
- The predictor variable that is found to contribute most to the dependent variable is inserted first

- All-subsets methods
- Tries every combination of variables to see which gives the best fit.
- Number of combinations increases with number of predictors.

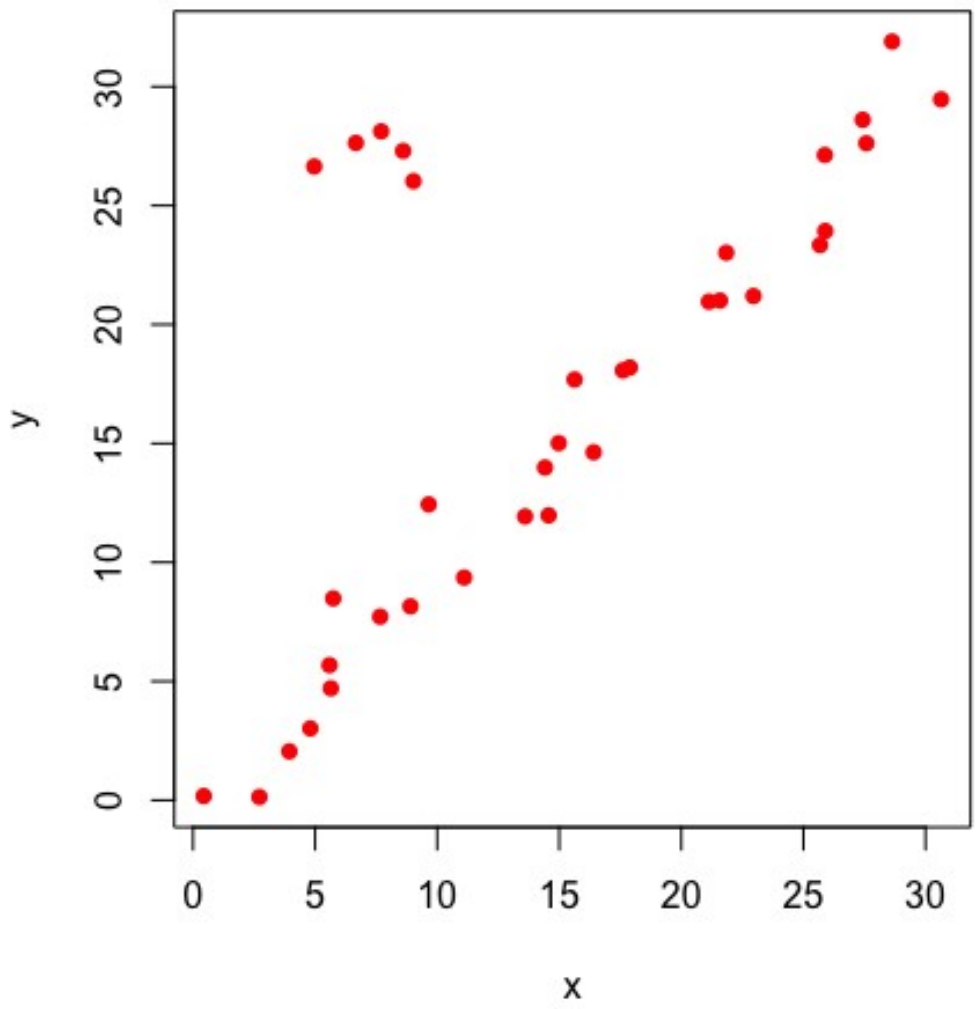
# Assessing/Interpreting the regression model

1) Does the model fit the observed data well, or is it influenced by a small number of cases? (**Part 1: Diagnostics**)
2) How do I interpret the output? (**Part 2: Interpretation)**
3) Can my model be generalised to other samples? (**Part 3: Assumptions**)

# Model Fit Diagnostics: Hierarchical Regression

- Does $R^2$ significantly improve? Test this using an F-ratio (using the anova() function). Note: when comparing hierarchical models, the second model must contain everything from the first model plus something new. The third model must contain everything from the second model plus something new (and so on…)
- Akaike information criterion (AIC) which is a measure of fit which penalizes the model for having more variables (like the adjusted $R^2$). Larger AIC values indicate worse fit. Note: You can only compare the AIC between models using the same data. Also the AIC means nothing on its own – therefore, a good value is defined relative to other AIC values.

# Model Fit Diagnostics: Outliers (and Residuals)



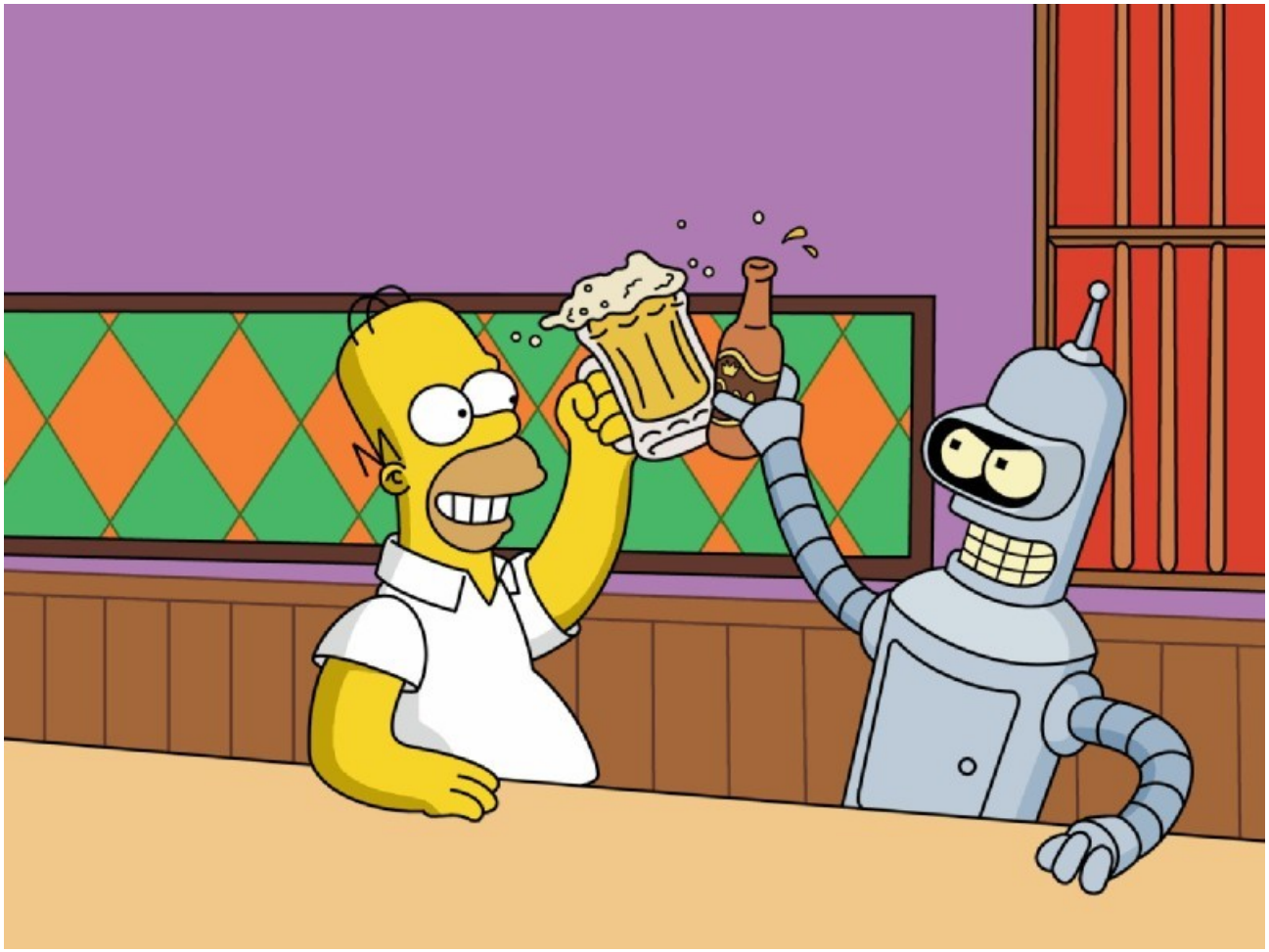
# Model Fit Diagnostics: Outliers and Residuals

- Outliers can cause your model to be biased because they affect the values of the estimated regression coefficients.
- We can assess potential outliers by checking observations with large residuals.
- We can **standardised residuals** to obtain a definition of what constitutes a large residual using z scores.
- In a normally distributed sample:
- 95% of scores are between -1.96 and +1.96
- 99% of scores are between -2.58 and +2.58
- 99.9% of scores are between -3.29 and + 3.29

# Model Fit Diagnostics: Standardised Residuals

1) Standardised residuals > absolute value of 3.29 are a cause for concern.
2) If more than 1% of our sample have standardised residuals > absolute value of 2.58, there is evidence that the level of error in our model is unacceptable.
3) If more than 5% of our sample have standardised residuals > absolute value of 1.96, there is evidence that the level of error in our model is unacceptable.

# Exercise 1: Are residuals enough?



# Model Fit Diagnostics: Influential Cases

- The **adjusted predicted value** that uses leave one out analyses (i.e. what effect does removing a case have on the regression coefficients?)
- **Studentised residual =** the residual based on the adjusted predicted value (i.e. difference between the adjusted predicted value and the original observed value, divided by the standard error)
- Studentised residuals are good for assessing the influence of a single case on the ability of the model to predict that single case but not on how a single case influences the model's ability to predict ALL cases.

# Model Fit Diagnostics: Influential Cases

- **Cook's distance** (considers the effect of a single case on the model as a whole). Values >1 are a possible cause for concern.
- **Leverage (or hat) values** which measure how far away the predictor variable (x) values of an observation are, relative to the observations (of x) as a whole.
- Leverage values lie between 0 (no influence) and 1 (a case has compete influence over a prediction). Most cases should be close to the average leverage value. The average leverage is (k + 1)/n where k is the number of predictors and n is the number of participants.

# Model Fit Diagnostics: Influential Cases

*Cook’s distance = leverage + influence*

*Conceptually yes, mathematically, not exactly...

# Model Fit Diagnostics: Influential Cases

- DFBeta is the standardised difference between a parameter estimated using all cases and estimated when one case is excluded.
- DFFit is the standardised difference between the predicted value for a case when the model is calculated including that case and when the model is calculated excluding that case.

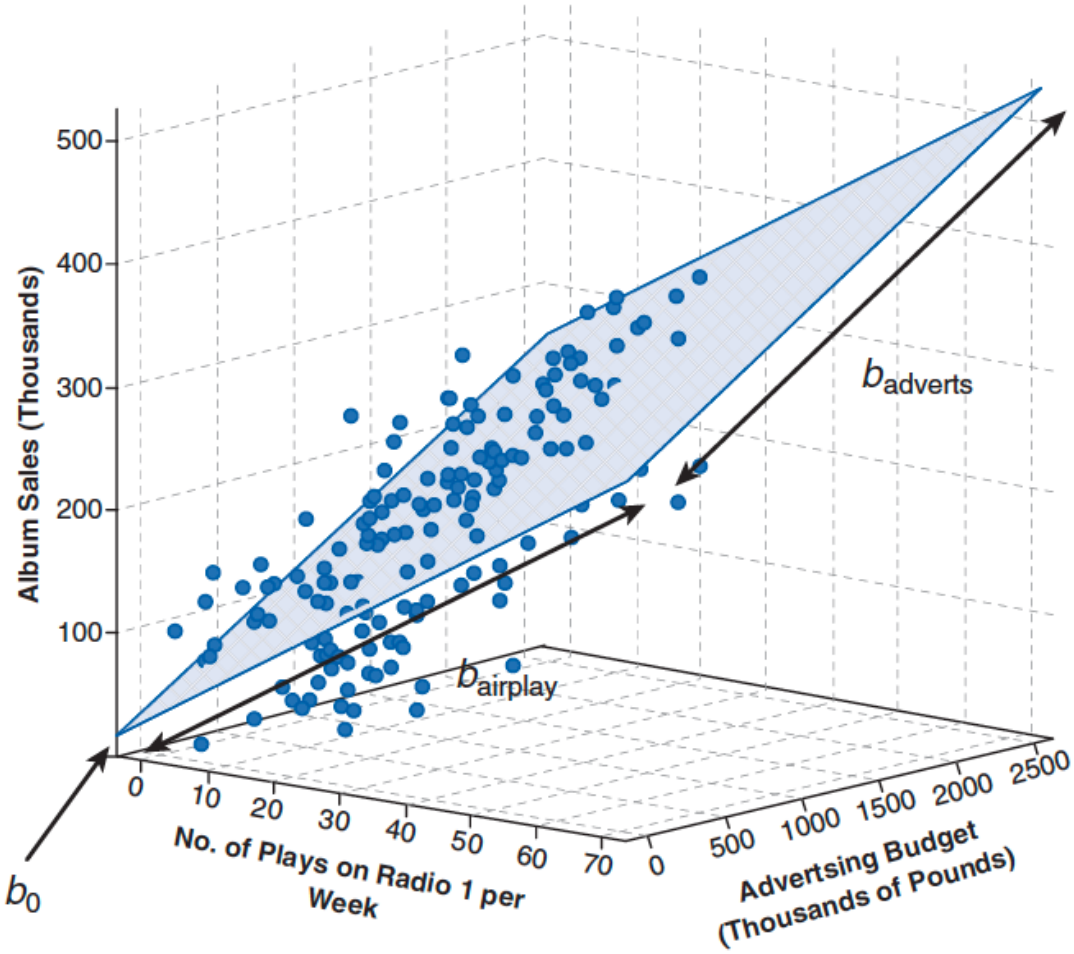
# Model Fit Diagnostics: Influential Cases

- Covariance Ratio (CVR) is a measure of whether a case influences the variance of the regression parameters. A ratio close to 1 means the case has little influence. Note – if CVR > 1 + 3[(k+1)/n] then deleting the case will damage the precision of model parameters. If CVR < 1 – 3[(k+1)/n], then deleting the case will improve the precision of some model parameters. (Remember: 1 plus or minus 3 times the leverage)

# Remember!

- Diagnostic tools help you see how good or bad your model is in terms of fitting the sampled data. They are a way of assessing your model. They are NOT, however, a way of justifying the removal of data points to effect some desirable change in the regression parameters.
- If you deem the removal of outliers necessary (due to theoretically sound reasons), then consider reporting results with and without outliers, if appropriate.

# Exercise 2: Multiple Regression



# Choosing your model

$$F = \frac{(N-k-1)R^2}{k(1-R^2)}$$

$$AIC = n\ln\left(\frac{\text{SSE}}{n}\right) + 2k$$

```
> # Compare R^2 between models
> anova(album_lm_0, album_lm_1, album_lm_2)
Analysis of Variance Table

Model 1: sales ~ 1
Model 2: sales ~ 1 + adverts
Model 3: sales ~ 1 + adverts + airplay + attract
  Res.Df     RSS Df Sum of Sq       F    Pr(>F)
1    199 1295952
2    198  862264  1    433688 195.600 < 2.2e-16 ***
3    196  434575  2    427690  96.447 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # Compare AIC between models
> AIC(album_lm_0, album_lm_1, album_lm_2)
             df      AIC
album_lm_0  2 2326.863
album_lm_1  3 2247.375
album_lm_2  5 2114.337
```

# Interpreting your output

```
Call:
lm(formula = sales ~ 1 + adverts + airplay + attract, data = album_data)

Residuals:
     Min       1Q   Median       3Q      Max
-121.324  -28.336   -0.451   28.967  144.132

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.612958  17.350001  -1.534    0.127
adverts       0.084885   0.006923  12.261  < 2e-16 ***
airplay       3.367425   0.277771  12.123  < 2e-16 ***
attract      11.086335   2.437849   4.548 9.49e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.09 on 196 degrees of freedom
Multiple R-squared:  0.6647,    Adjusted R-squared:  0.6595
F-statistic: 129.5 on 3 and 196 DF,  p-value: < 2.2e-16
```

$$\begin{aligned}\text{sales}_i &= b_0 + b_1\text{advertising}_i + b_2\text{airplay}_i + b_3\text{attractiveness}_i \\ &= -26.61 + (0.08\ \text{advertising}_i) + (3.37\ \text{airplay}_i) + (11.09\ \text{attractiveness}_i)\end{aligned}$$

# Interpreting your output

- The beta values tells us the relationship between album sales and each predictor as well as the degree to which each predictor affects the outcome if the effects of all other predictors are held constant.

# Interpreting your output

**Advertising budget (b = 0.085)**: This value indicates that 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 attractiveness of the band and airplay are held constant.

# Interpreting your output

**Airplay (b = 3.367)**: This value indicates that as the number of plays on radio in the week before release increases by one, album sales increase by 3.367 units. Therefore, 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 attractiveness of the band and advertising are held constant.

# Interpreting your output

**Attractiveness (b = 11.086)**: This value indicates that a band rated one unit higher on the attractiveness scale can expect additional album sales of 11.086 units. Therefore, every unit increase in the attractiveness of the band is associated with an extra 11.086 thousand albums (11,086 albums) being sold. This interpretation is true only if the effects of radio airplay and advertising are held constant.

# Interpreting your output

- Remember that the magnitude of the t-value can indicate which predictor has the biggest impact on the response.
- Alternatively, we can use standardised beta estimates to compare the magnitude of effect across predictors.

```
> # Standardised beta-values
> lm.beta(album_lm_2)
  adverts   airplay   attract
0.5108462 0.5119881 0.1916834
```

# Interpreting your output

**Advertising budget (standardised b = 0.511)**: This value indicates that as advertising budget increases by one standard deviation ($485,655), album sales increase by 0.511 standard deviations. The standard deviation for album sales is 80,699 and so this constitutes a change of 41,240 sales (0.511 x 80,699). Therefore, for every $485,655 more spent on advertising, an extra 41,240 albums are sold. This interpretation is true only if the effects of attractiveness of the band and airplay are held constant.

```
> sd(album_data$adverts)
[1] 485.6552
> sd(album_data$sales)
[1] 80.69896
```

# Interpreting your output

**Airplay (standardised b = 0.512)**: This value indicates that as the number of plays on radio in the week before release increases by 1 standard deviation (12.27), album sales increase by 0.512 standard deviations. This constitutes a change of 41,320 sales (0.512 x 80,699). Therefore if the radio station plays the song an extra 12.27 times in the week before release, 41,320 extra album sales can be expected. This interpretation is true only if the effects of attractiveness of the band and advertising are held constant.

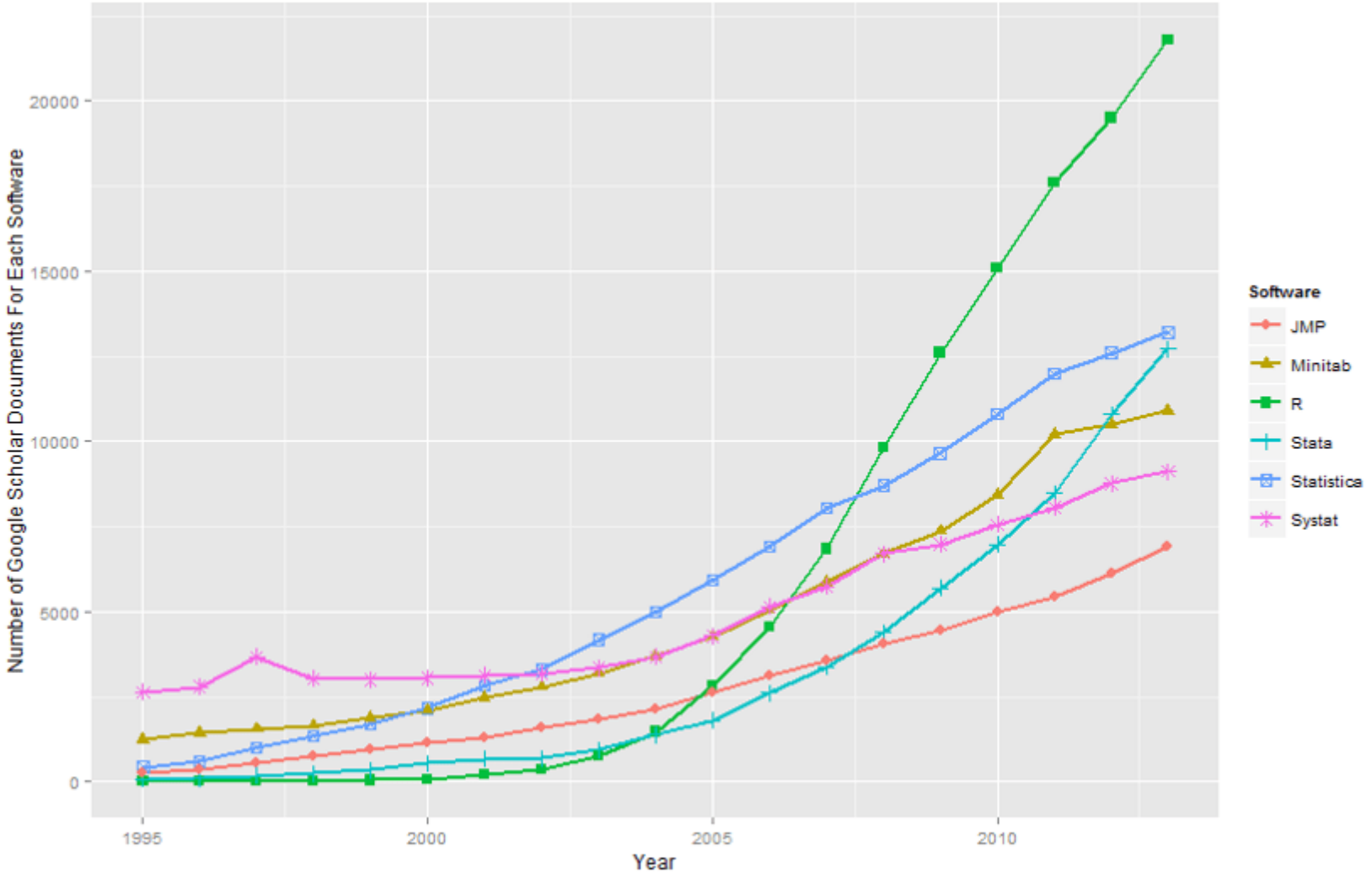
```
> sd(album_data$airplay)
[1] 12.26958
> sd(album_data$sales)
[1] 80.69896
```

# Interpreting your output

**Attractiveness (standardised b = 0.192)**: This value indicates that a band rated one standard deviation (1.40 units) higher on the attractiveness scale can expect additional album sales of 0.192 standard deviations. This constitutes a change of 15,490 sales (0.192 x 80,699). Therefore, a band with an attractiveness rating 1.40 higher than another band can expect 15,490 additional sales. This interpretation is true only if the effects of airplay and advertising are held constant.

```
> sd(album_data$attract)
[1] 1.39529
> sd(album_data$sales)
[1] 80.69896
```

# Exercise 3:
# Outliers and Influential Cases



# Model Generalisation: Assumptions

1) Variable types: All predictor variables must be quantitative or categorical (with two categories) and the outcome variable must be quantitative, continuous and unbounded (i.e. no constraints on the variability of the outcome). If you have a categorical variable with more than 2 categories you will need to dummy code the variable – Note: R does this for you if you specify that the categorical variable is a factor.

2) Non-zero variance: The predictors should have some variation in value (i.e. they do not have variances of 0) – variables need to…vary

# Model Generalisation: Assumptions

3) No **perfect** multicollinearity: There should be no perfect linear relationship between two or more of the predictors i.e the predictor variables should not correlate too highly. (VIF < 10 for each predictor, average VIF should not be substantially greater than 1 and the tolerance statistics (1/VIF) should not be less than 0.1 or 0.2). Note: these thresholds are arguable.

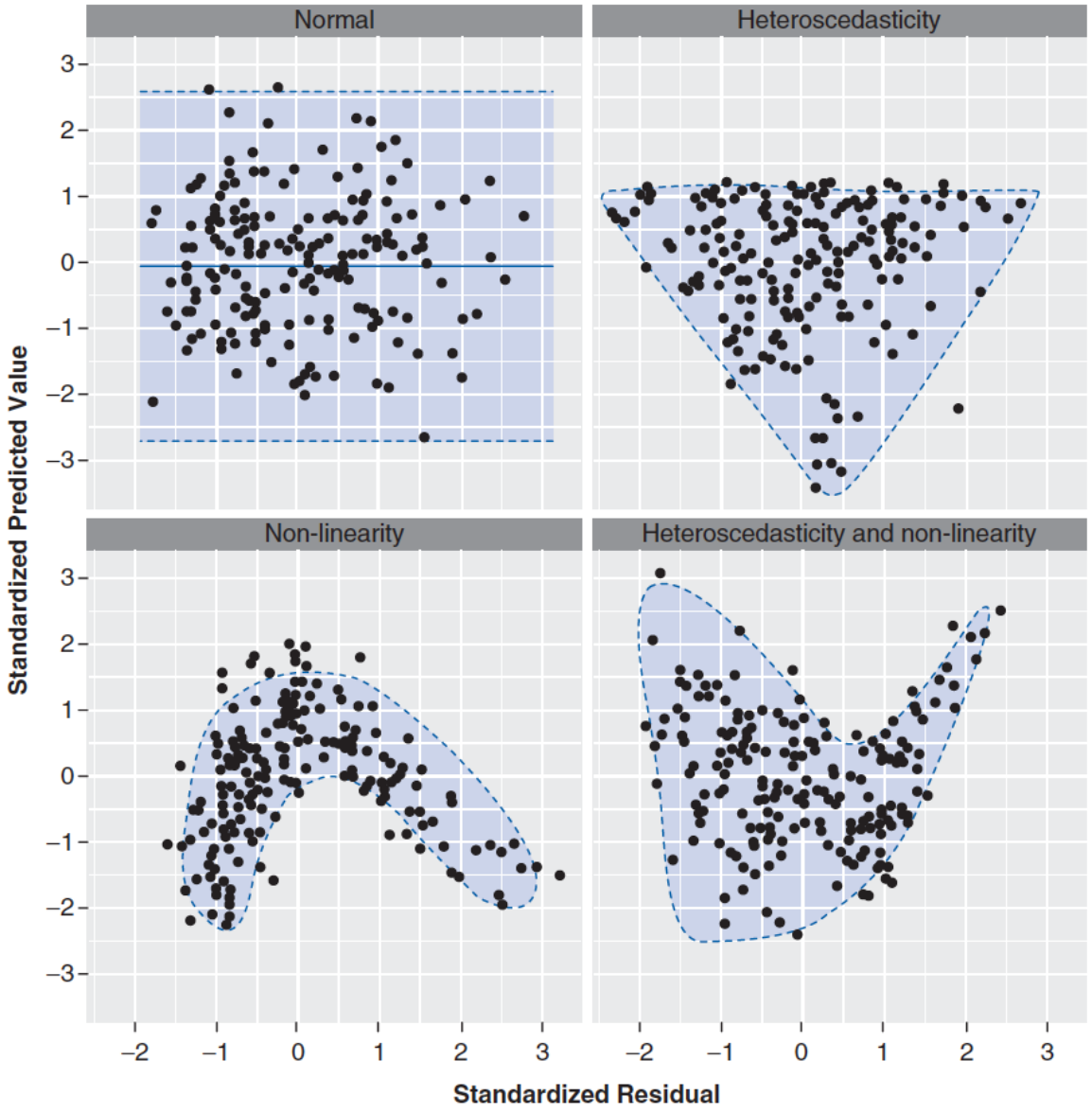
4)Predictors are uncorrelated with “external variables”: External variables are variables that haven’t been included in the regression model which influence the outcome variable i.e. “third variable problem”.

# Model Generalisation: Assumptions

5) Homoscedasticity: At each level of the predictor variable(s) the variance of the residual terms should be constant.

6) Linearity: The relationship that we are modeling is a linear one.

# Model Generalisation: Assumptions



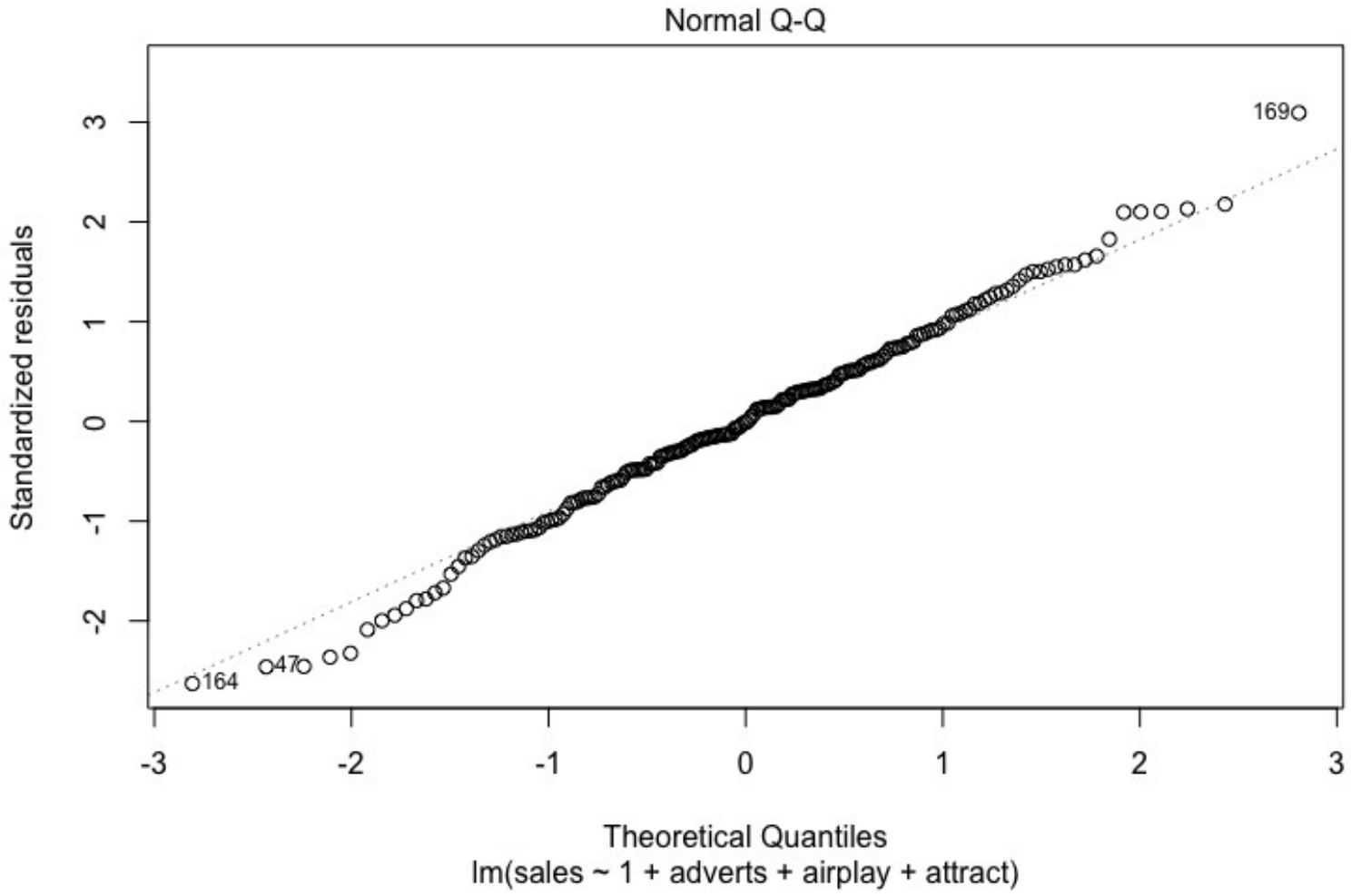
# Model Generalisation: Assumptions

7) Independent errors: For any two observations the residual terms should be uncorrelated (or independent). Use Durbin–Watson (DW) test for this. The test statistic varies between 0 and 4 with a value of 2 meaning no correlation between residuals. Values greater than 2 indicate a negative correlation between adjacent residuals and a value less than 2 indicate positive correlations. Note, values greater than 3 and less than 1 should raise alarm bells. Also, reordering your data will change the value of the DW test. Because DW uses bootstrapping to obtain p-value, this value will change slightly each time you run the analysis.

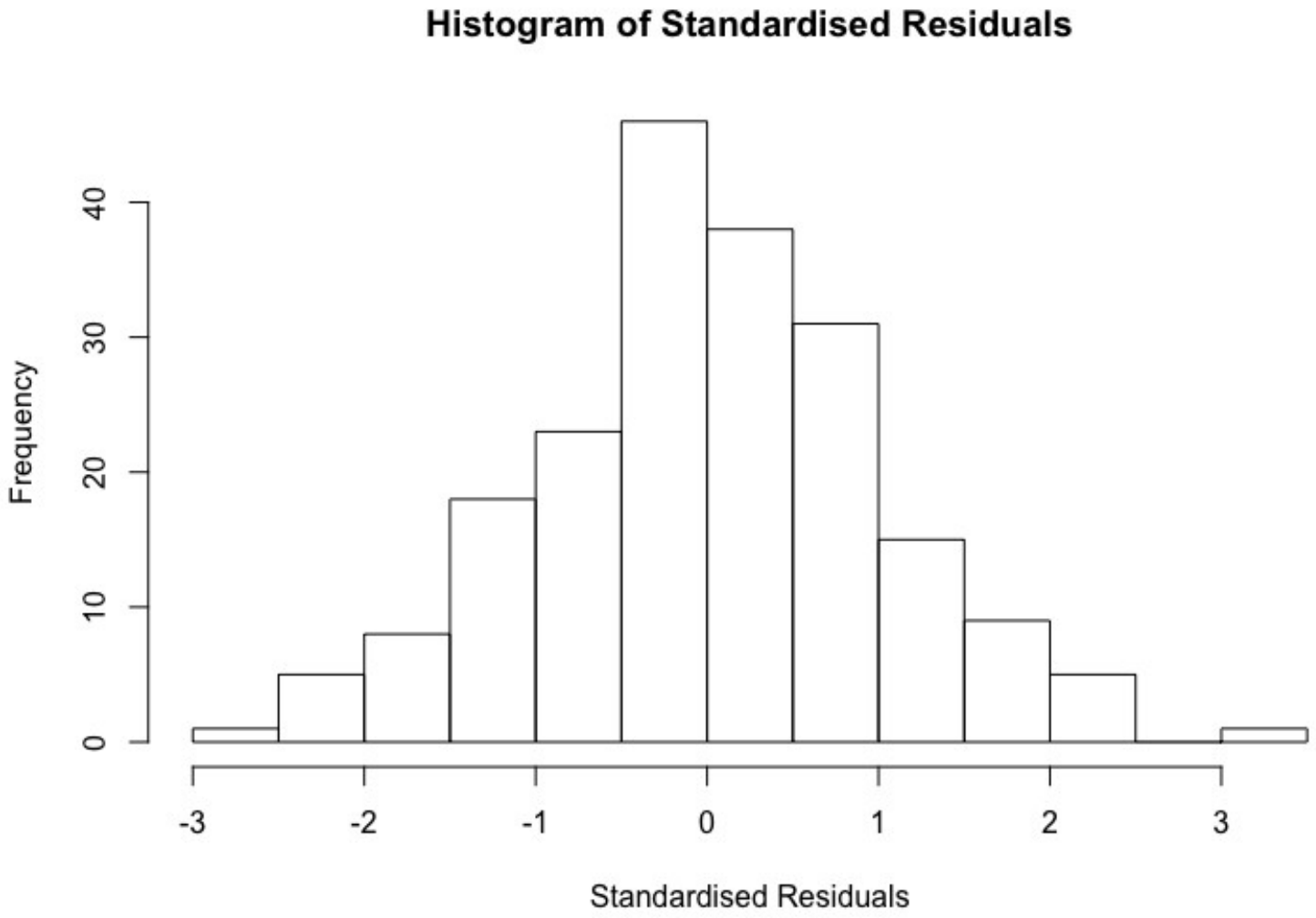
# Model Generalisation: Assumptions

8) Normally distributed errors. The residuals in the model should be random and normally distributed with a mean of 0. Note – predictors do not need to be normally distributed.

# Model Generalisation: Assumptions



# Model Generalisation: Assumptions



# Model Generalisation: Assumptions

9) Independence: It is assumed that all of the values of the outcome variable are independent (i.e. each value of the outcome variable comes from a separate entity).

# What if I violate an assumption?

- If assumptions are met we can generalise our model to the population.

- If assumptions have not been met, the model cannot be generalised to the population of interest. Instead, inferences need to be restricted to the sample only.

# What if I violate an assumption?

- If residuals show problems with heteroscedasticity or non-normality, you could try transforming the raw data (but this won't necessarily affect the residuals).
- Violation of linearity may be addressed with logistic regression, if the outcome can be categorical or non-linear models.
- Robust regressions (i.e. bootstrapping) is also possible.

# Statistics are important!

*To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.*

*- Sir Ronald Fisher*

# But always be skeptical...

*All models are wrong, but some are useful.*

*- George Box*

# Thank you!