Multiple Regression

(adapted in part from Dr. Andy Field)

Goals

- When To Use Multiple Regression.
- The multiple regression equation and what the betas represent.
- Different Methods of Regression Model Building
- How to Interpret multiple regression.
- Assumptions of Multiple Regression and how to test them.

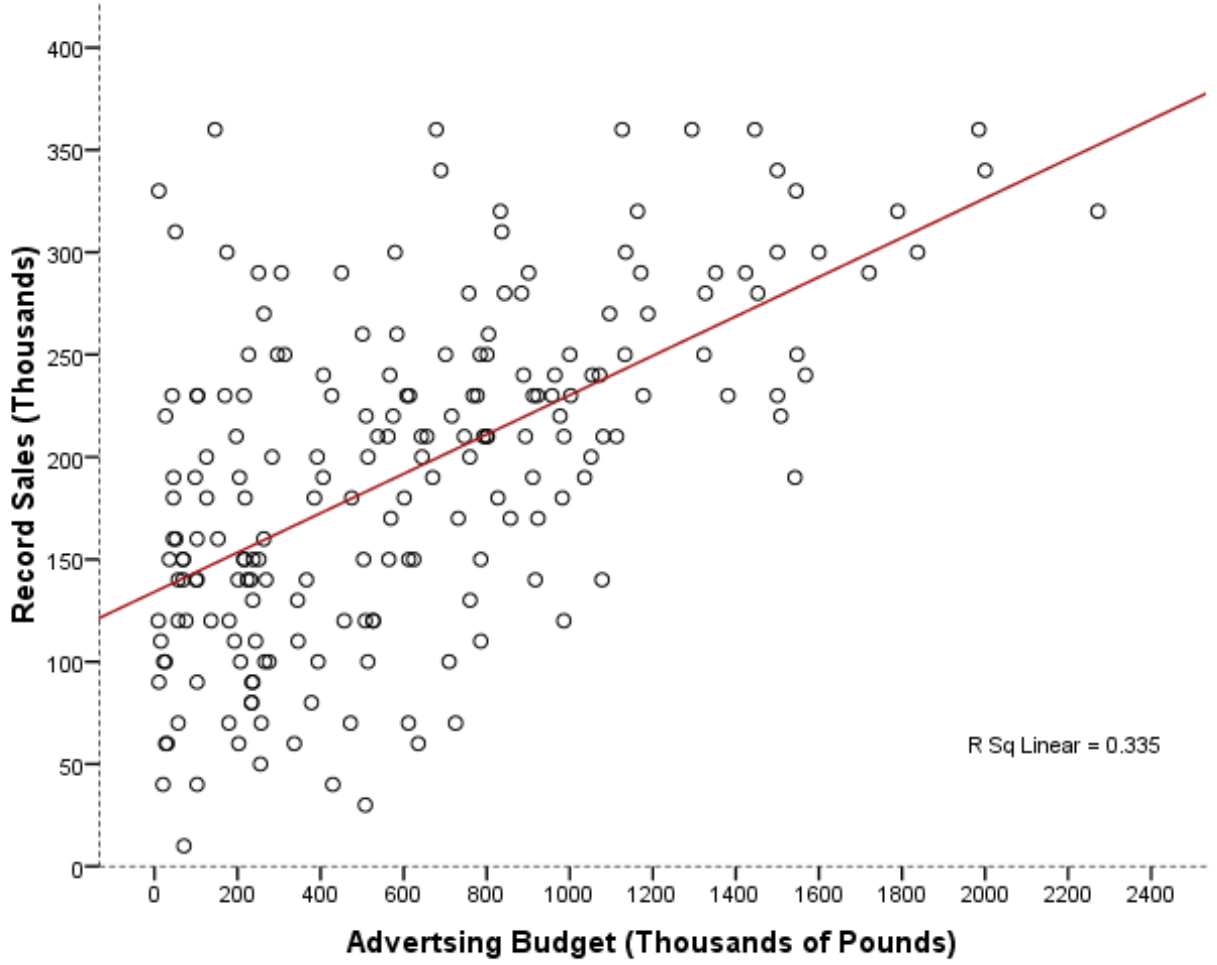
What is Multiple Regression?

- Linear Regression is a model to predict the value of one variable from another.
- Multiple Regression is a natural extension of this model:
  - We use it to predict values of an outcome from *several* predictors.
  - It is a hypothetical model of the relationship between several variables.

Regression: An Example

- A record company boss was interested in predicting record sales from advertising.
- Data
  - 200 different album releases
- Outcome variable:
  - Sales (CDs and Downloads) in the week after release
- Predictor variables
  - The amount (in thousands of £s) spent promoting the record before release
  - Number of plays on the radio (Radio 1)
  - Attractiveness of band (scale)

The Model with One Predictor



Multiple Regression as an Equation

- With multiple regression the relationship is described using a variation of the equation of a straight line.

$$y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n + \varepsilon_i$$

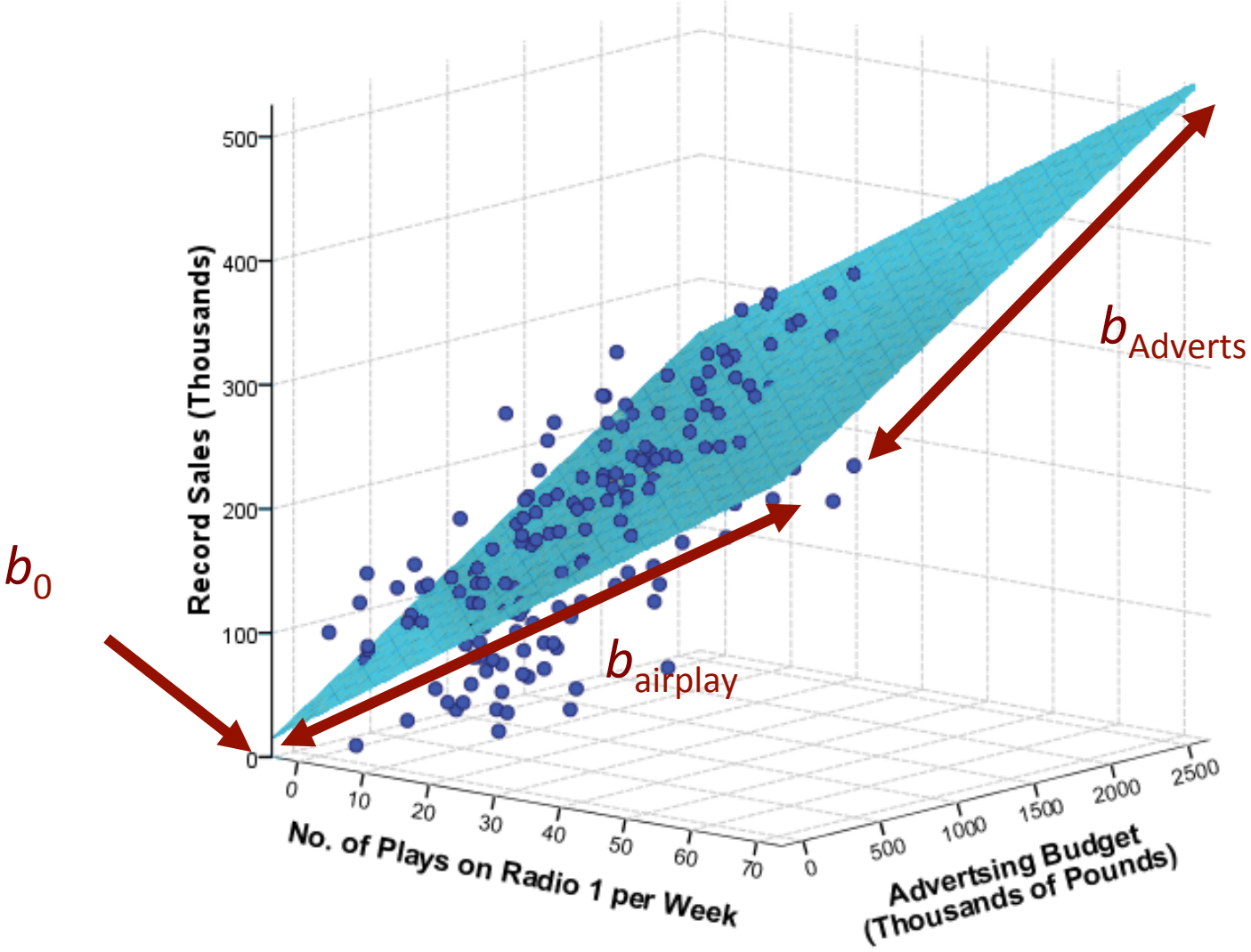
b_0

- b_0 is the intercept.
- The intercept is the value of the Y variable when all Xs = 0.
- This is the point at which the regression plane crosses the Y-axis (vertical).

Beta Values

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

The Model with Two Predictors



Methods of Regression

- Stepwise:
  - Predictors are selected using their semi-partial correlation with the outcome.
- Forced Entry:
  - All predictors are entered simultaneously.
- Hierarchical:
  - Experimenter decides the order in which variables are entered into the model.


Which method of regression should I use?

Stepwise Regression

- Variables are entered into the model based on mathematical criteria.
- Computer selects variables in steps.

Problems with Stepwise Methods

- Rely on a mathematical algorithm.
- Should be used only for exploration

Forced Entry Regression

- 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.

Hierarchical Regression

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

Hierarchical Regression

- 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:
  - Relies on the experimenter knowing what they're doing!

RESEARCH QUESTION:

Can advertising budget, number of plays on the radio, and attractiveness of band predict record sales?

Output: Model Summary

```
Call:
lm(formula = SALES ~ ADVERTS + AIRPLAY + ATTRACT, data =
advert)

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
```

R and R^2

```
Multiple R-squared:  0.6647,     Adjusted R-squared:  0.6595
```

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

```
F-statistic: 129.5 on 3 and 196 DF,  p-value: < 2.2e-16
```

- The F-test
  - 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.

Output: betas

```
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
```

How to Interpret Beta Values

- Beta values:
  - the change in the outcome associated with a unit change in the predictor.

Beta Values

- b_1= 0.085.
  - So, as advertising increases by £1,000, record sales increase by 0.085 units (thousands of records).

```
Estimate Std. Error t value Pr(>|t|)
ADVERTS       0.084885   0.006923  12.261  < 2e-16 ***
```

- b_2= 3.367.
  - So, each time (per week) a song is played on radio 1 its sales increase by 3.367 units (thousands of records).

```
Estimate Std. Error t value Pr(>|t|)
AIRPLAY       3.367425   0.277771  12.123  < 2e-16 ***
```

Beta Values

- $b_3 = 11.086$.
  - So, as attractiveness of band increases by 1 point on the scale, record sales increase by 11.086 units (thousands of records).

```
Estimate Std. Error t value Pr(>|t|)
ATTRACT      11.086335   2.437849   4.548 9.49e-06 ***
```

Predicting Outcomes

$$y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$$

```
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 ***
```

- What is the predicted outcome if:
  - £11 million advertising budget
  - 15 plays on Radio One per week
  - Band attractiveness is a “2”

Reporting the Model

TABLE 7.2 How to report multiple regression

	B	*SE B*	β
Constant	−26.61	17.35	
Advertising Budget	0.09	0.01	.51*
Plays on BBC Radio 1	3.37	0.28	.51*
Attractiveness	11.09	2.44	.19*

Predictors as "Controls"

- Very often, but not always, you are looking at the relationship between a key independent variable and a dependent variable.

Predictors as “Controls”

- You may need to include other independent variables to ‘control for’ other known correlates with the outcome
- Thus, some independent variables are called ‘controls’

Generalization

- When we run regression, 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.

Straightforward Assumptions

- Variable Type / Measurement:
  - 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.

Sample Size

- N should be greater than: 10 * k (k= the number of independent variables

The More Tricky Assumptions

- Models should be **parsimonious**:
  - or, as 'lean' as possible.
  - Don't add variables that have no theoretical basis.
- No Multicollinearity:
  - Predictors must not be highly correlated (Pearson's R of 0.7 or higher).
- No circularity:
  - Or, no variables that predict Y because they are in some way a representation of Y (Pearson's R of 0.8 or higher).

The gravest Regression error

- Omitted variable bias
  - Theory is most important here
- Remedies
  - More data collection
  - Careful survey design
  - A proxy for the omitted variable

What will be on the final exam?

- **Bivariate Hypothesis Tests**
  - Two-Sample t-tests
  - ANOVA
  - Chi-Square
  - Simple Regression
- **Multivariate Hypothesis Test**
  - Multiple Regression
- **Measures of Association/Correlation**
- **Key concepts of statistics**

Format of the final exam

- Minimal calculations/formulas
- Interpretation of R output
- Cases: research ‘situations’ presented, and the answer will be an essay on the approach to take, which variables to use, which hypothesis test, and what the results would tell you
- Definition of key concepts of statistics