Linear Regression

Please download from "Example Datasets":
- ***Record2.sav***
- ***ExamAnxiety.sav***

2 Oct 2009
CPSY501
Dr. Sean Ho
Trinity Western University

Outline for today

- Correlation and Partial Correlation
- Linear Regression (Ordinary Least Squares)
  - Using Regression in Data Analysis
  - Requirements: Variables
  - Requirements: Sample Size
- Assignments & Projects

Causation from correlation?

- Ancient proverb: "correlation ≠ causation"
- *But*: sometimes correlation can suggest causality, in the right situations!
  - Need >2 variables, and it's
  - Not the "ultimate" (only, direct) cause, but
  - Degree of influence one var has on another
    - ("causal relationship", "causality")
- Is a significant fraction of the variability in one variable explained by the other variable(s)?

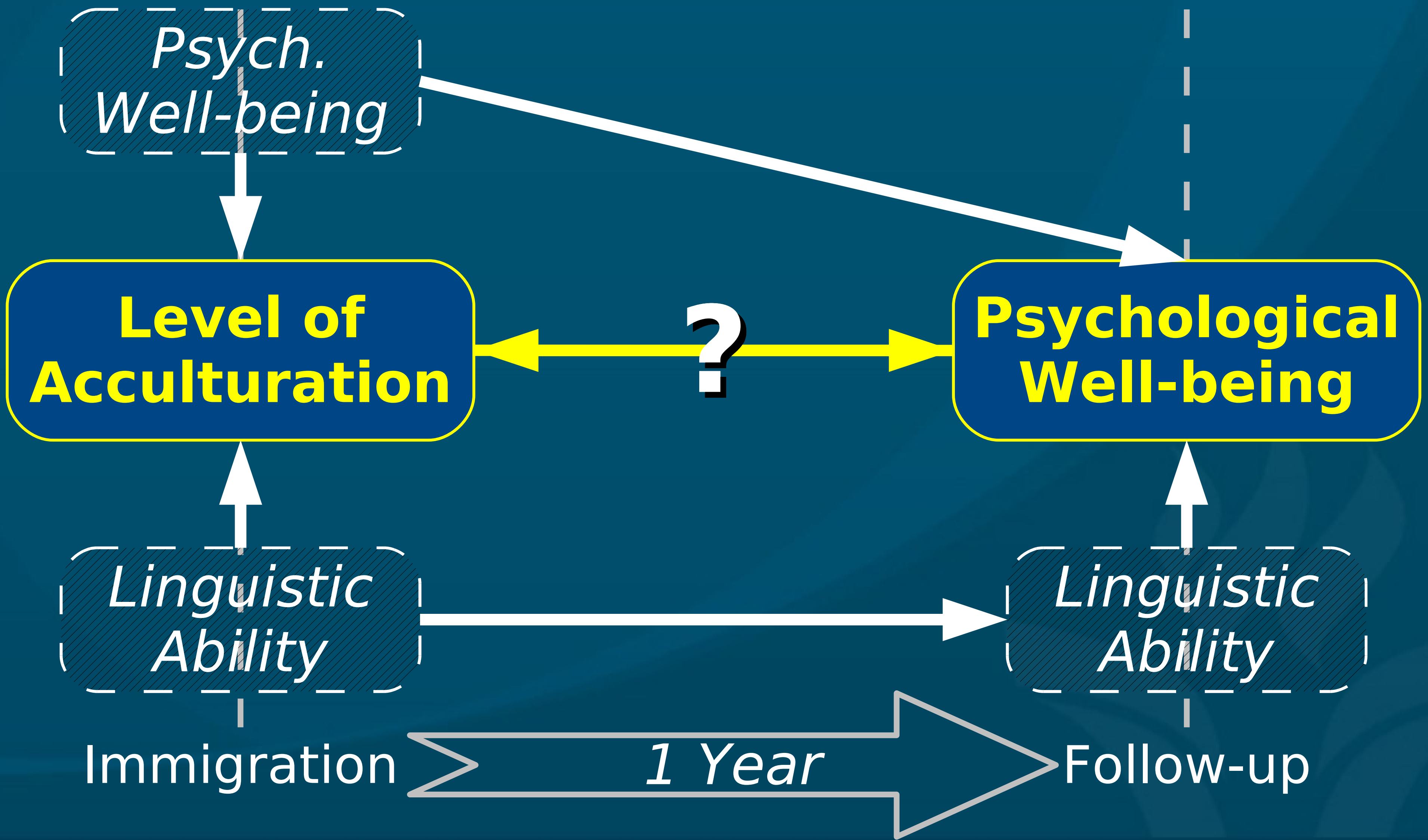
How to get causality?

- Use theory and prior empirical evidence
    - e.g., temporal sequencing: can't be T2 → T1
  - Order of causation? *(gender & career aspir.)*
- Best is a controlled experiment:
    - Randomly assign participants to groups
    - Change the level of the IV for each group
    - Observe impact on DV
- But this is not always feasible/ethical!
- Still need to “rule out” other potential factors

Potential hidden factors

- What other vars might explain these correlatns?
  - Height and hair length
    - e.g., both are related to gender
  - Increased time in bed and suicidal thoughts
    - e.g., depression can cause both
  - # of pastors and # of prostitutes (!)
    - e.g., city size
- To find causality, must rule out hidden factors
    - e.g., cigarettes and lung cancer

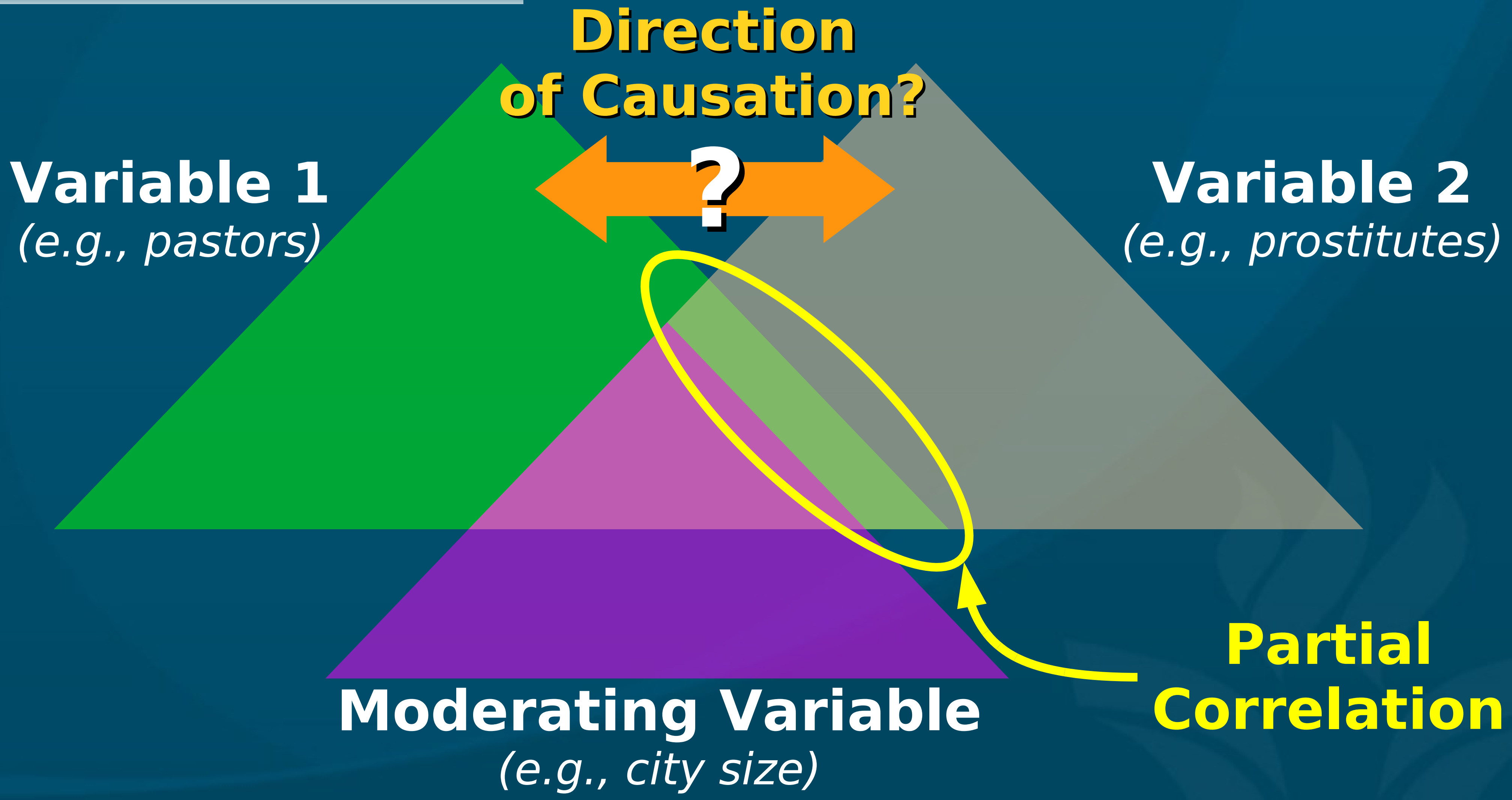
Example: immigration study



Partial Correlation

- *Purpose*: to measure the unique relationship between two variables – after the effects of other variables are "controlled for".
- Two variables may have a large correlation, but
  - It may be largely due to the effects of other moderating variables
  - So we must factor out their influence, and
  - See what correlation remains (partial)
- SPSS algorithm assumes parametricity
  - There exist non-parametric methods, though

Visualizing Partial Correlation

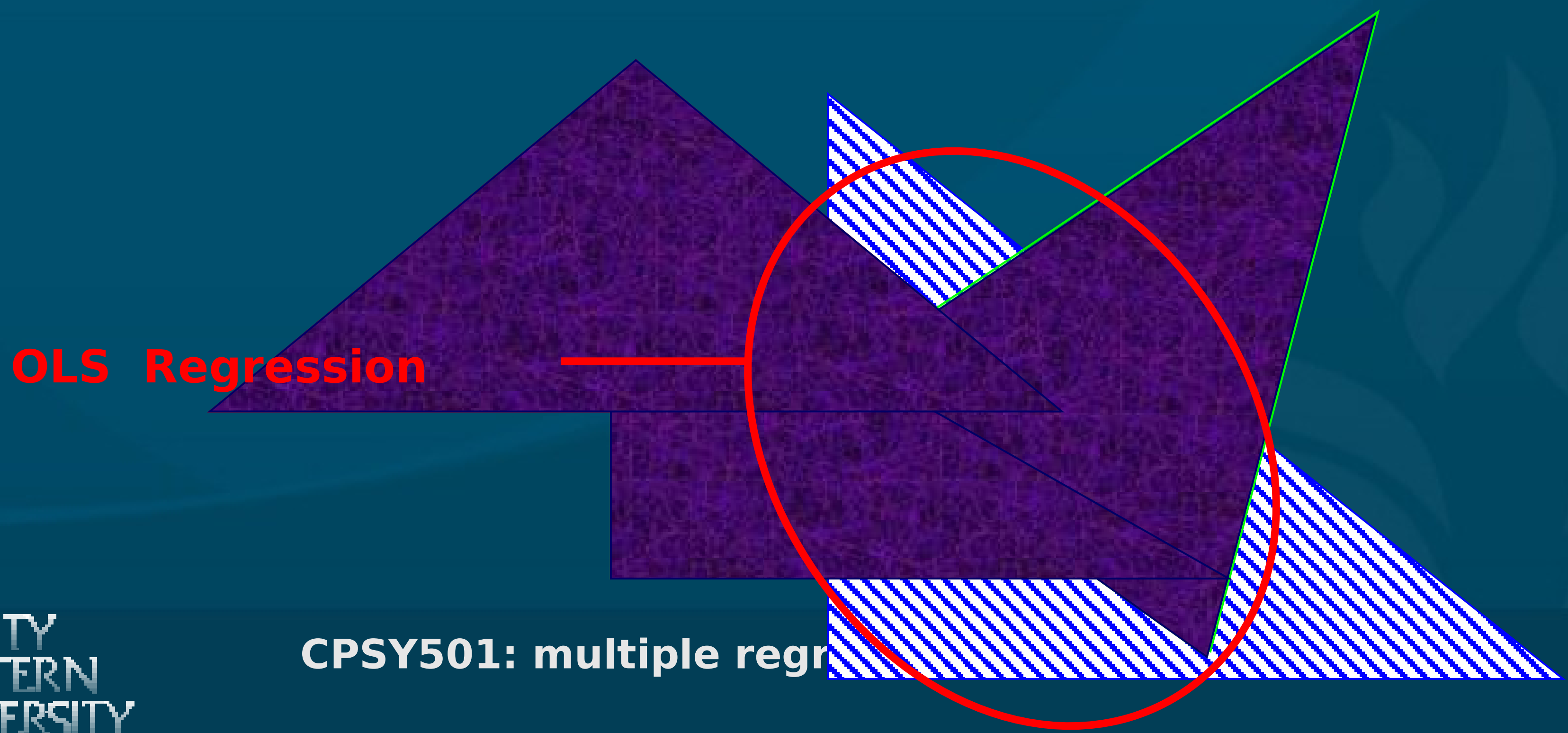


Practise: Partial Correlation

- Dataset: Record2.sav
- Analyze → Correlate → Partial,
  - Variables: adverts, sales
  - Controlling for: airplay, attract
- Or: Analyze → Regression → Linear:
  - Dependent: sales; Independents: all other var
  - Statistics → “Part and partial correlations”
    - Uncheck everything else
  - All partial *r*, plus regular (0-order) *r*

Linear Regression

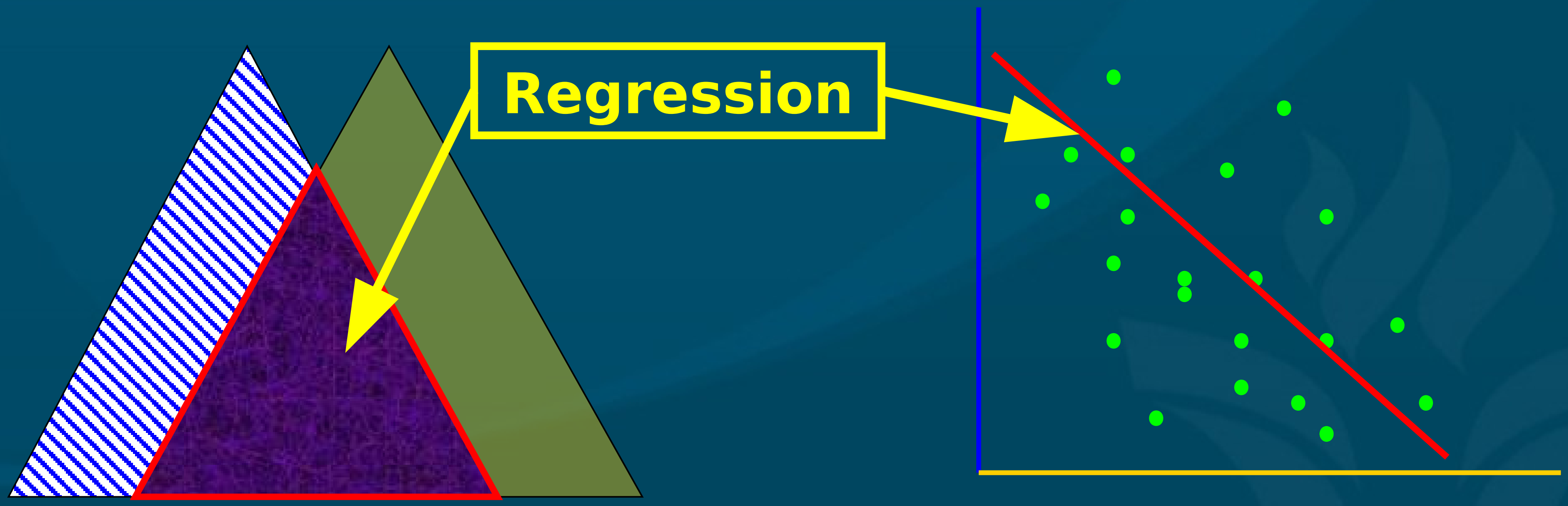
- When we use correlation to try to infer direction of causation, we call it regression:
- Combining the influence of a number of variables (predictors) to determine their total effect on another variable (outcome).



Simple Regression: 1 predictor

Simple regression is predicting scores on an outcome variable from a single predictor variable

- Mathematically equiv. to bivariate correlation



OLS Regression Model

- In Ordinary Least-Squares (OLS) regression, the "best" model is defined as the line that minimizes the error between model and data (sum of squared differences).
- Conceptual description of regression line (General Linear Model):

$$Y = b_0 + b_1 X_1 + (B_2 X_2 \ldots + B_n X_n) + \varepsilon$$

Outcome

Intercept

Gradient

Predictor

Error

Assessing Fit of a Model

- R^2: proportion of variance in outcome accounted for by predictors: SS_{model} / SS_{tot}
  - Generalization of r^2 from correlation
- F ratio: variance attributable to the model divided by variance unexplained by model
  - F ratio converts to a *p*-value, which shows whether the "fit" is good
- If fit is poor, predictors may not be linearly related to the outcome variable

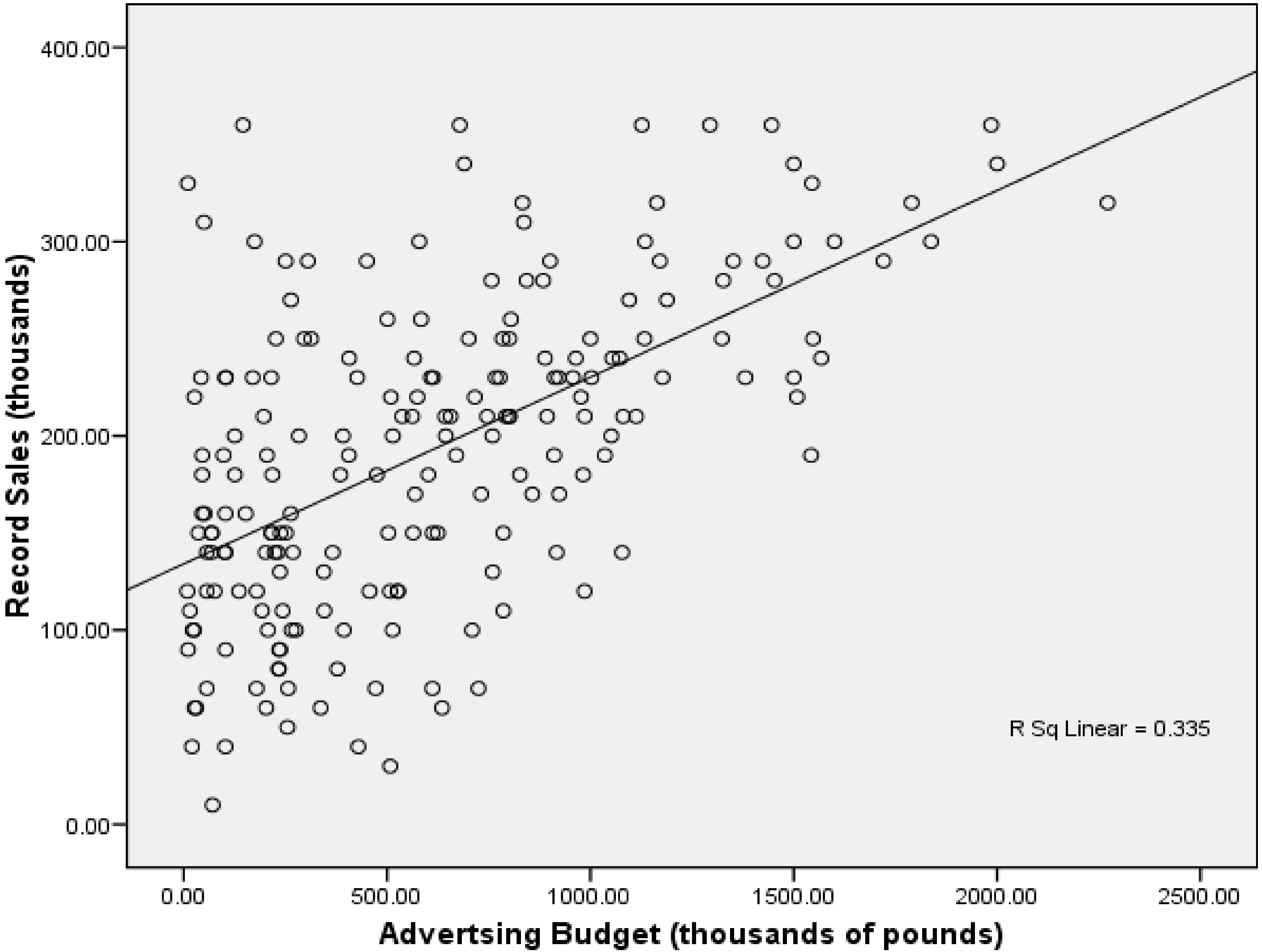
Example: *Record Sales*

- Dataset: Record2.sav
- Outcome variable: Record sales
- Predictor: Advertising Budget

- Analyze → Regression → Linear
  - Dependent: sales; Independent: adverts
  - Statistics → Estimates, Model fit, Part and partial correlations

Example: Interpreting Results

- Model Summary: R Square = .335
  - adverts explains 33.5% of variability in sales
- ANOVA: F = 99.587 and Sig. = .000
  - There is a significant effect
- Report: $R^2 = .335$, $F(1, 198) = 99.587$, $p < .001$
- Coefficients: (Constant) B = 134.140 (Y-intcpt)
  - Unstandardized advert B = .096 (slope)
  - Standardized advert Beta = .578
    - (with all variables converted to z-scores)
  - Linear model: $\hat{Y} = .578 * AB_z + 134$



400.00
300.00
200.00
100.00
0.00
Record Sales (thousands)
0.00
500.00
1000.00
1500.00
2000.00
2500.00
Advertising Budget (thousands of pounds)
R Sq Linear = 0.335

General Linear Model

- Actually, nearly all parametric methods can be expressed as generalizations of regression!
- Multiple Regression: several scale predictors
- Logistic Regression: categorical outcome
- ANOVA: categorical IV
  - *t*-test: dichotomous IV
  - ANCOVA: mix of categorical + scale IVs
  - Factorial ANOVA: several categorical IVs
- Log-linear: categorical IVs and DV
- We'll talk about each of these in detail later

Regression Modelling Process

(1) Develop RQ: IVs/DVs, metrics
- Calc required sample size & collect data

(2) Clean: data entry errors, missing data, outliers

(3) Explore: assess requirements, xform if needed

(4) Build model: try several models, see what fits

(5) Test model: “diagnostic” issues:
- Multivariate outliers, overly influential cases

(6) Test model: “generalizability” issues:
- Regression assumptions: rebuild model

Selecting Variables

- According to your model or theory, what variables might relate to your outcomes?
  - Does the literature suggest important vars?
- Do the variables meet all the requirements for an OLS multiple regression?
  - (more on this next week)
- Record sales example:
  - DV: what is a possible outcome variable?
  - IV: what are possible predictors, and why?

Choosing Good Predictors

- It's tempting to just throw in hundreds of predictors and see which ones contribute most
  - Don't do this! There are requirements on how the predictors interact with each other!
  - Also, more predictors → less power
- Have a theoretical/prior justification for them
- Example: what's a DV you are interested in?
  - Come up with as many possible good IVs as you can – have a justification!
    - Background, internal personal, current external environment

Using Derived Variables

You may want to use derived variables in regression, for example:

- Transformed variables (to satisfy assumptions)
- Interaction terms: ("moderating" variables)
  - e.g., Airplay * Advertising Budget
- Dummy variables:
  - e.g., coding for categorical predictors
- Curvilinear variables (non-linear regression)
  - e.g., looking at X^2 instead of X

Required Sample Size

- Depends on effect size and # of predictors
  - Use G*Power to find exact sample size
  - Rough estimates on pp. 172-174 of Field
- Consequences of insufficient sample size:
  - Regression model may be overly influenced by individual participants (not generalizable)
  - Can't detect “real” effects of moderate size
- Solutions:
  - Collect more data from more participants!
  - Reduce number of predictors in the model

Requirements on DV

- Must be interval/continuous:
  - If not: mathematics simply will not work
  - Solutions:
    - Categorical DV: use Logistic Regression
    - Ordinal DV: use Ordinal Regression,
      or possibly convert into categorical form
- Independence of scores (research design)
  - If not: invalid conclusions
  - Solutions: redesign data set, or
    - Multi-level modelling instead of regression

Requirements on DV, cont.

- Normal (use normality tests):
  - If not: significance tests may be misleading
  - Solutions: Check for outliers, transform data, use caution in interpreting significance
- Unbounded distribution (obtained range of responses versus possible range of responses)
  - If not: artificially deflated R^2
  - Solutions:
    - Collect more data from missing range
    - Use a more sensitive instrument

Requirements on IV

- Scale-level
  - Can be categorical, too (see next page)
  - If ordinal, either threshold into categorical or treat as if scale (only if have sufficient number of ranks)
- Full-range of values
  - If an IV only covers 1-3 on a scale of 1-10, then the regression model will predict poorly for values beyond 3
- More requirements on the behaviour of IVs with each other and the DV (covered next week)

Categorical Predictors

- Regression can work for categorical predictors:
- If dichotomous: code as 0 and 1
  - e.g., 1 dichotomous predictor, 1 scale DV:
  - Regression is equivalent to *t*-test!
  - And the slope B_1 is the difference of means
- If *n* categories: use “dummy” variables
  - Choose a base category
  - Create *n-1* dichotomous variables
  - e.g., {BC, AB, SK}: dummys are isAB, isSK