

Multiple Regression

7 Oct 2011
CPSY501
Dr. Sean Ho
Trinity Western University

*Please download from
“Example Datasets”:*

- ***Record2.sav***
- ***Domene.sav***

Outline: Multiple Regression

- Regression Modelling Process
- Building a Regression Model
 - Shared vs. Unique Variance
 - Strategies for Entering IVs
 - Interpreting Output
- Diagnostic Tests:
 - Residuals, Outliers, and Influential Cases
- Checking Assumptions:
 - Non-multicollinearity, independence, normality, homoscedasticity, linearity

Encouragement on Research

- Undergrad students: “is this on the test?”
 - “What do I need to do to pass?”
 - Doing the bare **minimum**: 1 DV, 2 IVs, 1 test
- Graduate students / prep for **research**:
 - “What structure/effects are in the data?”
 - Do **whatever it takes** to understand the data
- You may need **several RQs**
- Your RQs may **change** as you progress
- Have a **theme**/goal and aim to **tell a story** about the effects in the dataset

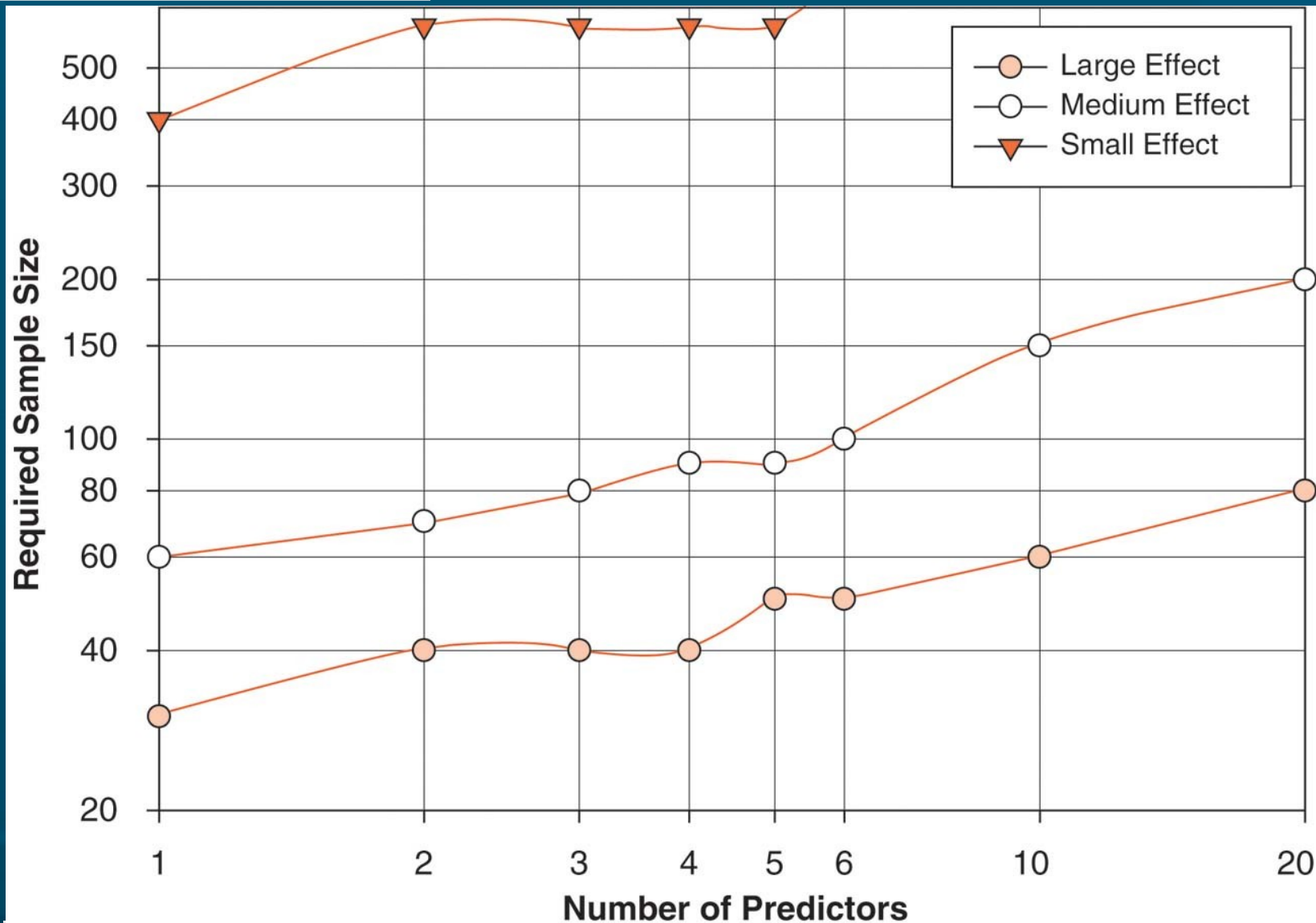
Regression Modelling Process

- (1) RQ: IVs/DVs, metrics, sample size, collect data
- (2) Clean: data entry errors, missing data, outliers
- (3) Explore: assess requirements, xform if needed
- (4) Build model: order & method of entry from RQ
- (5) Test model: “diagnostic” issues:
 - Multivariate outliers, overly influential cases
- (6) Test model: “generalizability” issues:
 - Multicollinearity, linearity of residuals
- (7) Run final model and interpret results

Required Sample Size

- Depends on **effect size** and # of **predictors**
 - Use **G*Power** to find exact sample size
 - Rough **estimates** on pp.222-223 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

Sample Size Estimates (Field)



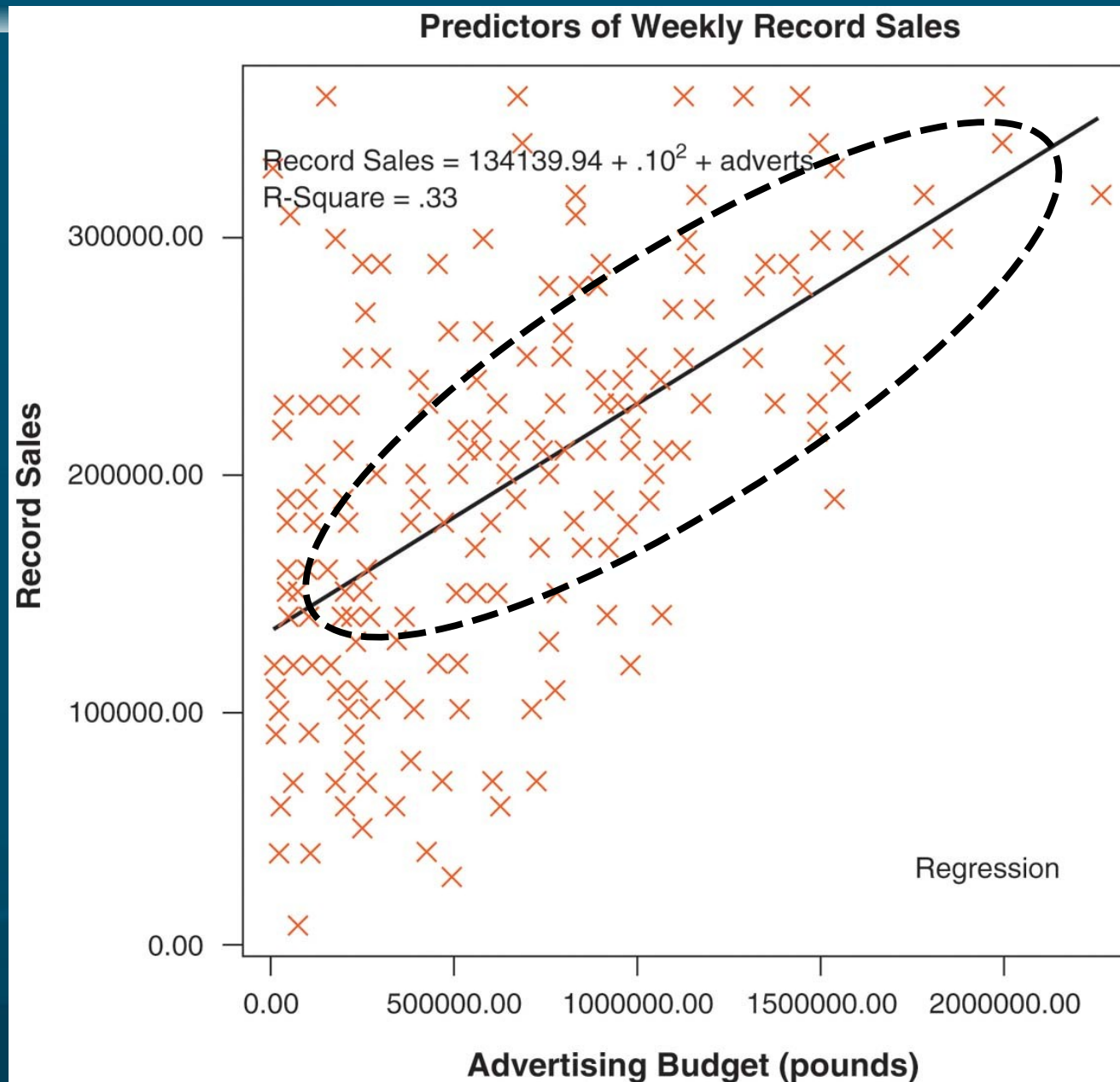
Outline: Multiple Regression

- Regression Modelling Process
- Building a Regression Model
 - Shared vs. Unique Variance
 - Strategies for Entering IVs
 - Interpreting Output
- Diagnostic Tests:
 - Residuals, Outliers, and Influential Cases
- Checking Assumptions:
 - Non-multicollinearity, independence, normality, homoscedasticity, linearity

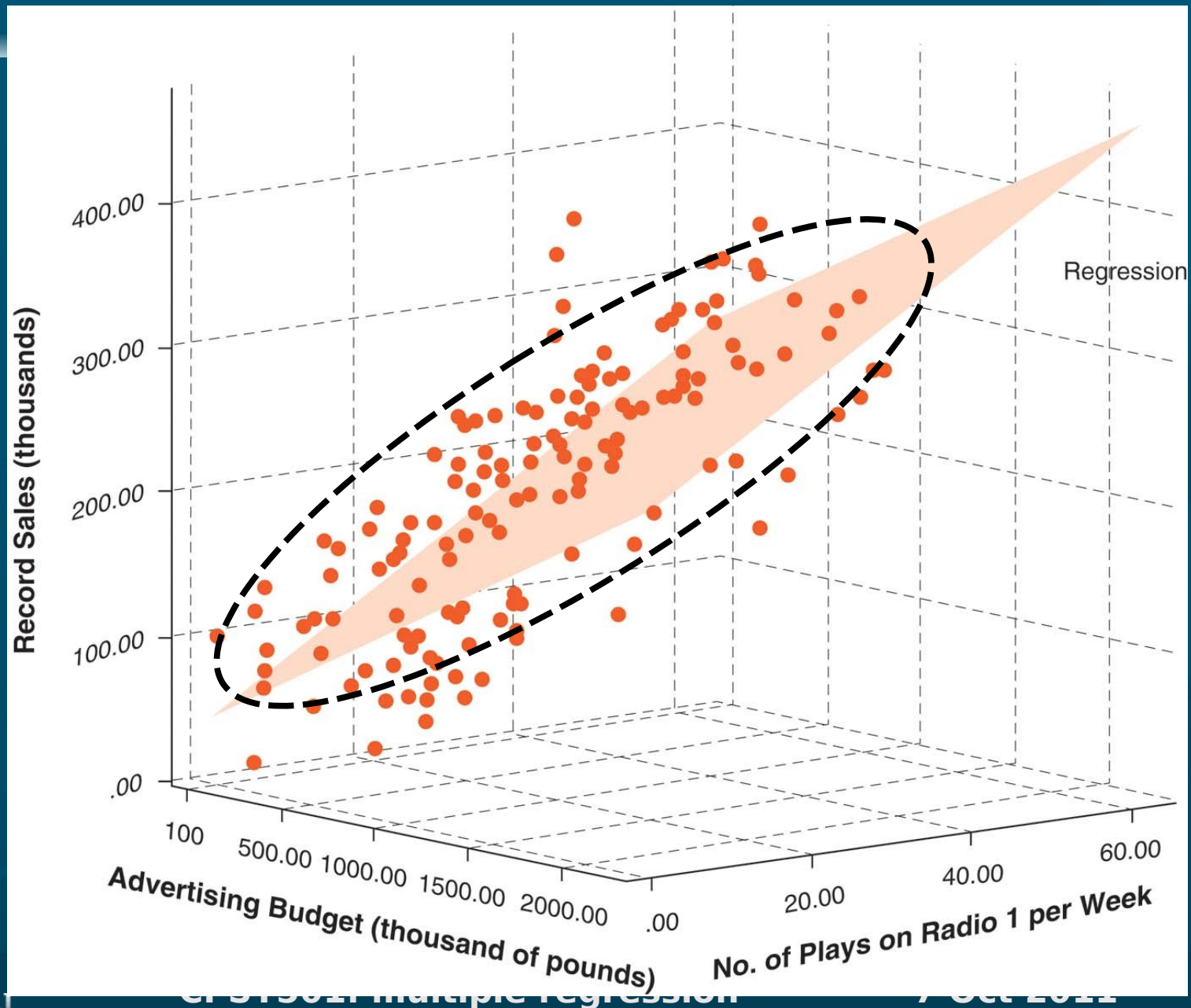
Example: Record Sales data

- Dataset: Record2.sav
- Outcome (“criterion”): record sales (RS)
- Predictors: advertising budget (AB), airtime (AT)
 - Both have good ‘variability’, and n=200
- Research Question: Do AB and AT both show unique effects in explaining Record Sales?
- Research design:
Cross-sectional, correlational study (same year?)
with 2 quantitative IVs & 1 quantitative DV
- Analysis strategy: Multiple regression (MR)

Regression Model with 1 IV



Regression Model with 2 IVs



Asking Precise RQs

- What does **literature** say about **AB** and **AT** in relation to **record sales**?
 - Previous lit may be **theoretical** or **empirical**
 - May focus on these **variables** or others
 - May be **consistent** or **conflicting** results
- Contrast these two seemingly similar RQs:
 - Is AB or AT more important for Sales?
 - Do AB and AT both show unique effects in accounting for the variance of Sales?

Example: Record Sales

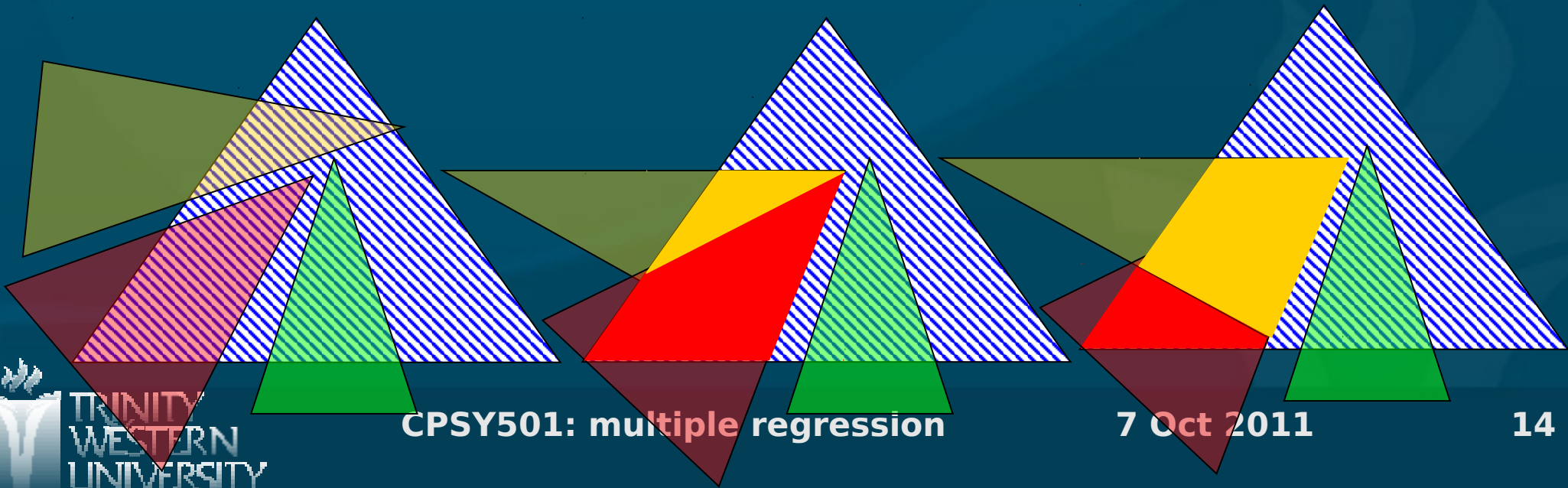
- Dataset: Record2.sav
- Analyze → Regression → Linear
- Dependent: Record Sales (RS)
- Independent: Advertising (AB) & Airtime (AT)
 - This is a “simultaneous” regression
- Statistics: check R^2 change and partial correl.
- Review output: t -test for each β coefficient: significance of unique effects for each predictor

Outline: Multiple Regression

- Regression Modelling Process
- Building a Regression Model
 - Shared vs. Unique Variance
 - Strategies for Entering IVs
 - Interpreting Output
- Diagnostic Tests:
 - Residuals, Outliers, and Influential Cases
- Checking Assumptions:
 - Non-multicollinearity, independence, normality, homoscedasticity, linearity

Shared vs. Unique Variance

- When predictors are **correlated**, they account for **overlapping** portions of variance in outcome
 - Redundant IVs, **mediation**, shared **background** effects, etc.
- **Order of entry** will help distinguish **shared** and **unique** contributions



Order of Entry

- Predictors in same **block** are entered into model at the **same** time
- Subsequent blocks only look at **remaining** variance after **previous** blocks have been **factored out**
- To find a predictor's **unique** contribution, put it **last** after other predictors are factored out
- Try **several** runs with **different** orderings to get **each** predictor's unique effect
- Order for your **final run** should reflect **theory** about relative importance of predictors

Options for Variable Selection

- Within each **block**, not all IVs need to be used:
 - **Manual** method: “Enter” (forced entry)
 - ◆ **All** specified IVs will be included
 - “Stepwise” **automatic** methods:
 - ◆ Forward: **add** significant IVs one-at-a-time
 - ◆ Backward: **eliminate** non-significant IVs
- Best to use “Enter”: **manual control**
 - You decide order according to **theory**/lit
- Automatic methods might not show **shared** effects, **interaction** effects

Record Sales Example

- Analyze → Regression → Linear
- Dependent: Record Sales
- Statistics: check R^2 change
- Run 1: “simultaneous” regression
 - Both AB and AT in Block 1
- Run 2: AB in Block 1, and AT in Block 2
- Run 3: AT in Block 1, and AB in Block 2

Calculating Shared Variance

- Output from Run 1: Total effect size from both predictors together is 63%
- Run 2: Airtime's unique effect size is 30%
 - Look at last ΔR^2 : when airtime is added
- Run 3: Advertising's unique effect size is 27%
- Shared variance:
 - = Total minus all unique effects
 - = 63% - 30% - 27% \approx 6%

Steps for Entering IVs

- First, create a conceptual **outline** of all IVs and their **connections** & **order** of entry.
 - Run a **simultaneous** regression: look at beta weights & *t*-tests for all **unique effects**
- Second, create “**blocks**” of IVs (in order) for any variables that **must** be in the model
 - Use “**Enter**” method to force vars into model
 - **Covariates** may go in these blocks
 - **Interaction** and **curvilinear** terms go in last of these blocks

Steps for Entering IVs (cont.)

- Any **remaining** variables go in a separate block: try all possible **combinations** to sort out **shared** & **unique** variance portions.
 - See record sales example above (no interaction terms were used)
- Summarize the final **sequence of entry** that clearly presents the predictors & their respective unique and shared effects.
- **Interpret** the relative sizes of the unique & shared effects for the Research Question

Entering IVs: SPSS tips

- Plan out your order and method on paper
- Each set of variables that should be entered in at the same time should be in a single block.
 - Other vars & interactions go in later blocks
- Usually choose “Enter” method (default)
 - Try automatic (“Backward”) only if needed
- Confirm correct order & method of entry in your SPSS output
 - Usually only need a few blocks of IVs

Outline: Multiple Regression

- Regression Modelling Process
- Building a Regression Model
 - Shared vs. Unique Variance
 - Strategies for Entering IVs
 - **Interpreting Output**
- Diagnostic Tests:
 - Residuals, Outliers, and Influential Cases
- Checking Assumptions:
 - Non-multicollinearity, independence, normality, homoscedasticity, linearity

Output: “Model Summary”

- R^2 : the variance in the outcome accounted for by the model (i.e., **combined effect** of all IVs)
 - Interpretation is similar to r^2 in correlation
 - Multiply by 100 to convert into a **percentage**
 - **Adjusted R^2** : unbiased estimate of the model, always smaller than R^2
- **R^2 Change (ΔR^2)**: **Increase** in effect size from one **block** of predictors to the next.
 - **F-test** checks whether this “improvement” is **significant**.

Output: “ANOVA” Table

- Summarizes results for the model as a **whole**: Is the “**simultaneous**” regression a better predictor than simply using the **mean score** of the outcome?
- Proper **APA format** for reporting F statistics (see also pp. 136-139 of APA publication manual):

● $F(3, 379) = 126.43, p < .001$

df-regression

df-residual

F-ratio

statistical
significance

Output: “Coefficients” Table

- **Individual** contribution of each predictor, and whether its contribution is **significant**
- **B** (b-weight, slope, gradient): Change in **outcome**, for every unit change of the **predictor**
- **beta** (β): **Standardized** b-weight. Compares the **relative strength** of the different predictors.
- **t-test** (*p*-value): Tests whether a particular variable contributes a **significant** unique effect in the outcome variable for that equation.

Non-significant Predictors

What if the *t*-test shows a predictor's unique effect is **non-significant**?

- In general, the ΔR^2 will be **small**. If not, then you have **low power** for that test & must report that.
- **Remove** the IV unless there is a **theoretical** reason for retaining it in the model (e.g., **low power**, help for interpreting **shared effects**)
- **Re-run** the regression after any variables have been removed

Outline: Multiple Regression

- Regression Modelling Process
- Building a Regression Model
 - Shared vs. Unique Variance
 - Strategies for Entering IVs
 - Interpreting Output
- Diagnostic Tests:
 - Residuals, Outliers, and Influential Cases
- Checking Assumptions:
 - Non-multicollinearity, independence, normality, homoscedasticity, linearity

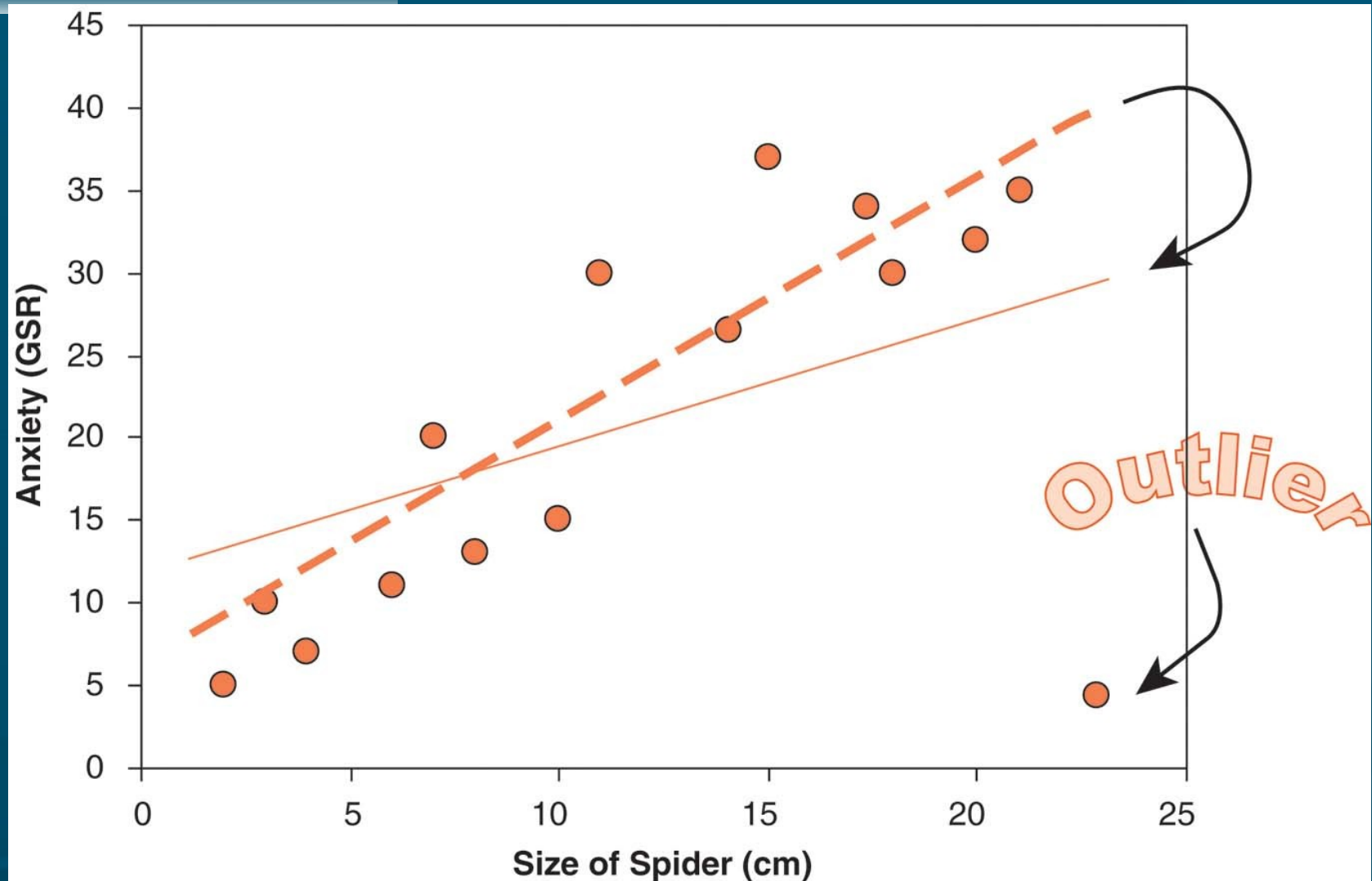
Residuals in Regression

- A **residual** is the difference between the **actual** score and the score **predicted** by the model
 - I.e., the amount of **error** for each case
- Examine the residuals in a trial run
 - Include **all** IVs: simultaneous regression
 - **Save** the residuals in a new variable:
- Analyze → Regression → Linear → Save:
“**standardized**” and/or “**unstandardized**”

Multivariate Outliers

- **Definition:** Cases from a **different** population than what we want to study
 - Combination of scores across predictors is substantially different from rest of sample
- **Consequence:** **distortion** of regression line, reduced **generalizability**
- **Screening:** Standardized **residual** $\geq \pm 3$, and **Cook's distance** > 1 (*these are rules of thumb*)
- **Solution:** **remove** outliers from from sample (if they exert too much influence on the model)

Effect of Multivariate Outliers



Overly-Influential Cases

- **Definition:** A case that has a substantially greater **effect** on the regression model than the majority of other cases in the sample
- **Consequence:** reduced **generalizability**
- **Screening & Solution** (rules of thumb):
 - if **leverage** > 0.50 then remove the case;
 - if $0.20 \leq \text{leverage} \leq 0.50$ and **Cook's distance** > 1 , then remove the case

Outliers & Influential cases

- Outliers and influential cases should be examined and removed together
 - Unlike other aspects of MR, screen only once
 - Why shouldn't you repeat this screening?
- SPSS: Analyze → Regression → Linear:
 - Save: Standardized Resid, Cook's, Leverage
 - Will be saved as additional vars in dataset
- Examine the Residual Statistics table
- Examine the saved scores in the data set
 - Try sorting: Data → Sort

Outline: Multiple Regression

- Regression Modelling Process
- Building a Regression Model
 - Shared vs. Unique Variance
 - Strategies for Entering IVs
 - Interpreting Output
- Diagnostic Tests:
 - Residuals, Outliers, and Influential Cases
- Checking Assumptions:
 - Non-multicollinearity, independence, normality, homoscedasticity, linearity

Multicollinearity

- **Definition:** Predictors **covary** too highly; i.e., too much **overlap** of shared variance
- **Consequences:** deflated R^2 ; may interfere with evaluation of β (depending on RQ & design)
- In “Statistics”: check “Collinearity Diagnostics”
- **Indicators** of possible problems: any of:
 - ◆ **Any VIF** (Variance Inflation Factor) score > 10
 - ◆ **Average VIF** is NOT approximately = 1
 - ◆ **Tolerance** < 0.2
- **Solution:** **delete**, **combine**, or **transform** some of the multicollinear variables

Independence of Residuals

- **Definition:** Residuals for different cases should not be systematically related
- **Consequence:** Can interfere with α and power, although effect size is unaffected
- **Screening:** Durbin-Watson scores that are relatively far away from 2 (on possible range of 0 to 4) indicate a problem with independence.
 - D-W sensitive to case ordering, so ensure cases aren't inherently ordered in dataset
- **Solution:** Re-evaluate sampling technique, or try multi-level modelling.

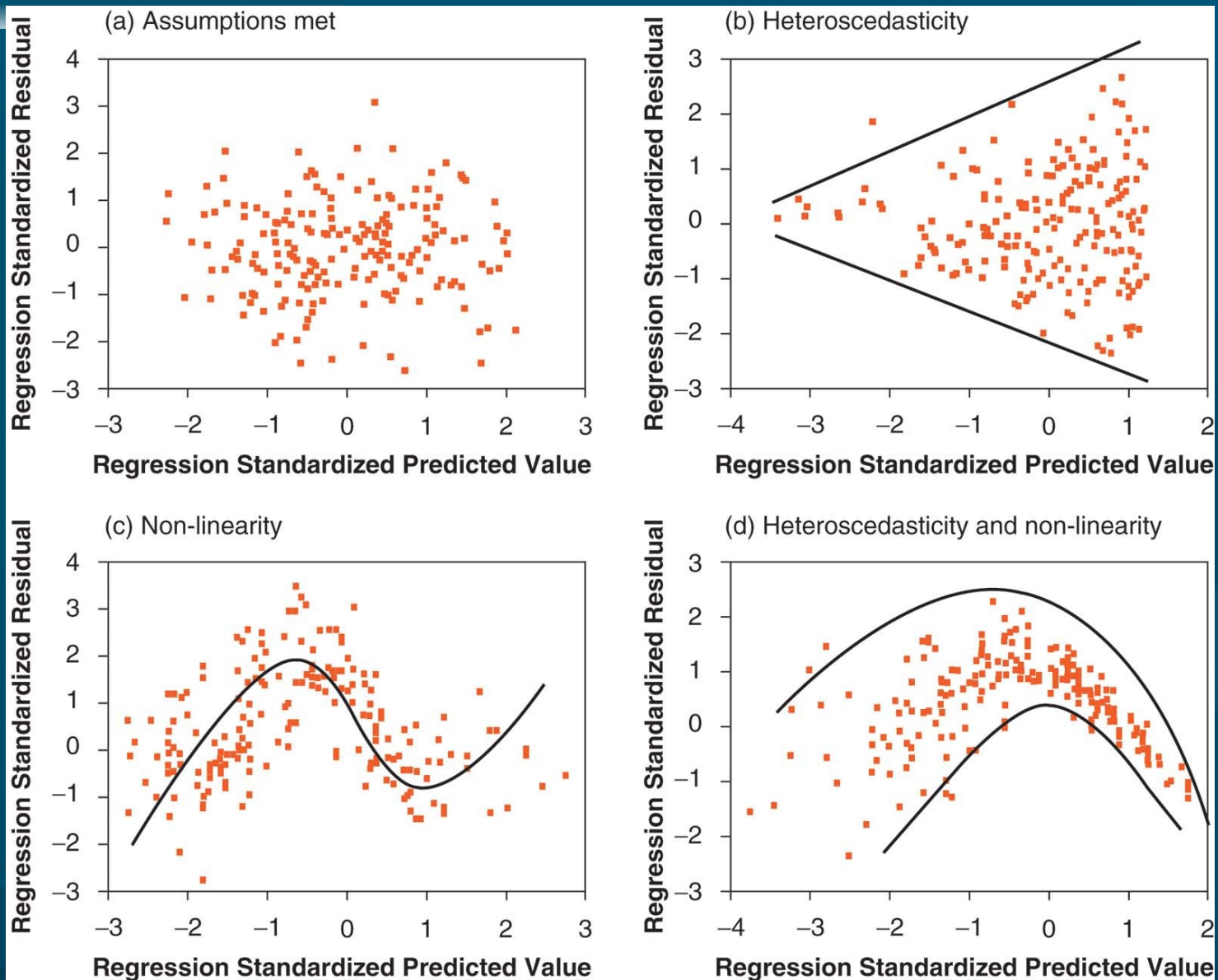
Normally Distributed Residuals

- **Definition:** Residuals normally distributed
 - Predictors don't have to be normal!
- **Consequence:** reduced generalizability (predictive value of the model is distorted)
- **Screening:** normality tests/plots on residuals
 - save standardized residuals
 - Analyze → Descriptives → Explore → “Normality tests with plots”
- **Solution:** check if predictors are non-normal or ordinal; look for non-linearity

Homoscedastic Residuals

- **Definition:** Residuals should have similar variances at every point on the regression line
 - Generalisation of homogeneity of variance
- **Consequence:** the model is less accurate for some people than others
- **Screening:** fan-shaped residual scatterplots:
 - Analyze → Regression → Linear → Plots:
X: "ZPRED" Y: "ZRESID"
- **Solution:** identify moderators and include, try weighted regression, or accept it and acknowledge the drop in accuracy

Heteroscedasticity



Non-linear Relationships

- **Definition:** Relationship between **predictor** and **outcome** is not **linear** (i.e., a straight line).
- **Consequences:** sub-optimal **fit** for the model (R^2 is **lower** than it could be)
- **Screening:** examine residual **scatterplots**
OR try **curve estimation**:
 - Analyze → Regression → Curve estimation
- **Solutions:** **Model** the non-linear relationship by entering a **polynomial** term into the regression equation (e.g., X^2 , X^3)
 - Or just accept the poorer fit

Exercise: Regression with SPSS

- Dataset: Domene.sav
- You try it! Build a regression model with:
 - DV: “educational attainment”
 - IV: Block 1: “academic performance”
 - IV: Block 2: “educational aspirations” and “occupational aspirations”
 - Use “Enter” method (force entry)
- Ask SPSS for ΔR^2 and partial correlation scores