# Coherence relations and DRD identification: theory and analysis

Ted Sanders (Utrecht University)
Wilbert Spooren (Radboud University Nijmegen)

Radboud University

# Regression and Logistic Regression

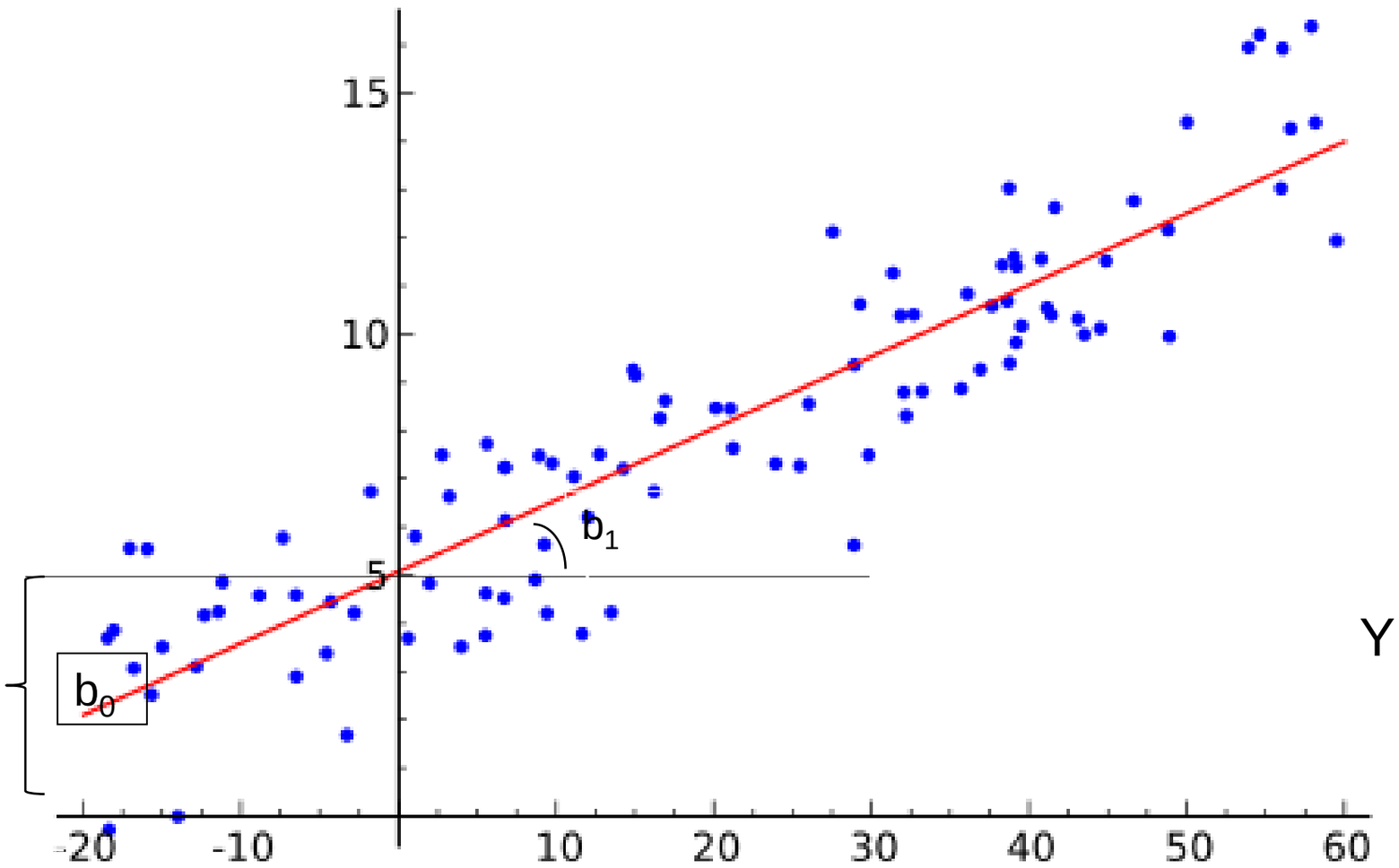
# Statistical modeling

- We male a model of our data in accordance with our hypotheses
- The fit of the model is an indication how valuable the model is

- Statistical models often have the from of a regression equation
  - $Y = b_0 + b_1 * X_1$
  - Number of records sold = constant + b1 * advertising budget

- Two crucial questions:
  - How good is our model? (what is the fit?)
  - Do the independent variables contribute to the prediction?

# Linear regression




$Y = b_0 + b_1 * X_1$

# How well does the model fit?

- Basic structure of any statistical model:
  - true score = predicted score + Error

- Model predicts well if the error component is small
  - little difference between predicted score and true score

- How to measure the amount of error?
  - By analyzing the deviations from the predicted scores: squared deviations (sums of squares)
  - $SS_{error}$ = Σ(true score –predicted score)$^2$
  - $SS_{total}$ =Σ(true score –mean score)$^2$
  - $SS_{predicted}$ = Σ(predicted score – mean score)$^2$

# Example from Field (2005)

| | Adverts | Sales | Airplay | Attract |
|---|---|---|---|---|
| 1 | Model: | | | 10 |
| 2 | Sales = 134,14 + 0,096*Adverts | | | 7 |
| 3 | | | | 7 |
| 4 | Adverts | Sales | Sales_Pre | Error | Error_Sq | 7 |
| 5 | 10,26 | 330 | 135,12 | 194,88 | 37976,28 | 5 |
| 6 | 985,69 | 120 | 228,77 | -108,77 | 11830,09 | 5 |
| 7 | 1445,56 | 360 | 272,91 | 87,09 | 7584,01 | 1 |
| 8 | 1188,19 | 270 | 248,21 | 21,79 | 474,97 | 9 |
| 9 | 574,51 | 220 | 189,29 | 30,71 | 942,92 | 7 |
| 10 | 568,95 | 170 | 188,76 | -18,76 | 351,91 | 7 |
| 11 | 471,81 | 70 | 179,43 | -109,43 | 11975,75 | 7 |
| 12 | 537,35 | 210 | 185,73 | 24,27 | 589,25 | 2 |
| | 514,07 | 200 | 183,49 | 16,51 | 272,56 | |
| | 174,09 | 300 | 150,85 | 149,15 | 22244,93 | |

| Model | | Sum of Squares | icients |
|---|---|---|---|
| 1 | Regression | 433687,833 | Error |
| | Residual | 862264,167 | 7,537 |
| | Total | 1295952,000 | ,010 |

a. Dependent Variable: sales

| Model | R | R Square |
|---|---|---|
| 1 | ,578[a] | ,335 |

# Does the independent variable make a significant contribution to the model?

- In a regression formula:
  - Sales = 134,14 + 0,096*Adverts

- Does the coëfficient of the predictor differ signifantly from 0?
  - Is “0,096” significant different from 0?

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95,0% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | 134,140 | 7,537 | | 17,799 | ,000 | 119,278 | 149,002 |
| | adverts | ,096 | ,010 | ,578 | 9,979 | ,000 | ,077 | ,115 |

a. Dependent Variable: sales

# Multiple regression

- More than one predictor

**Model Summary**

| Mode | | | | | Change Statistics | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | Adjusted R | Std. Error of | R Square | | | |
| 2 | | | | | | | | | |

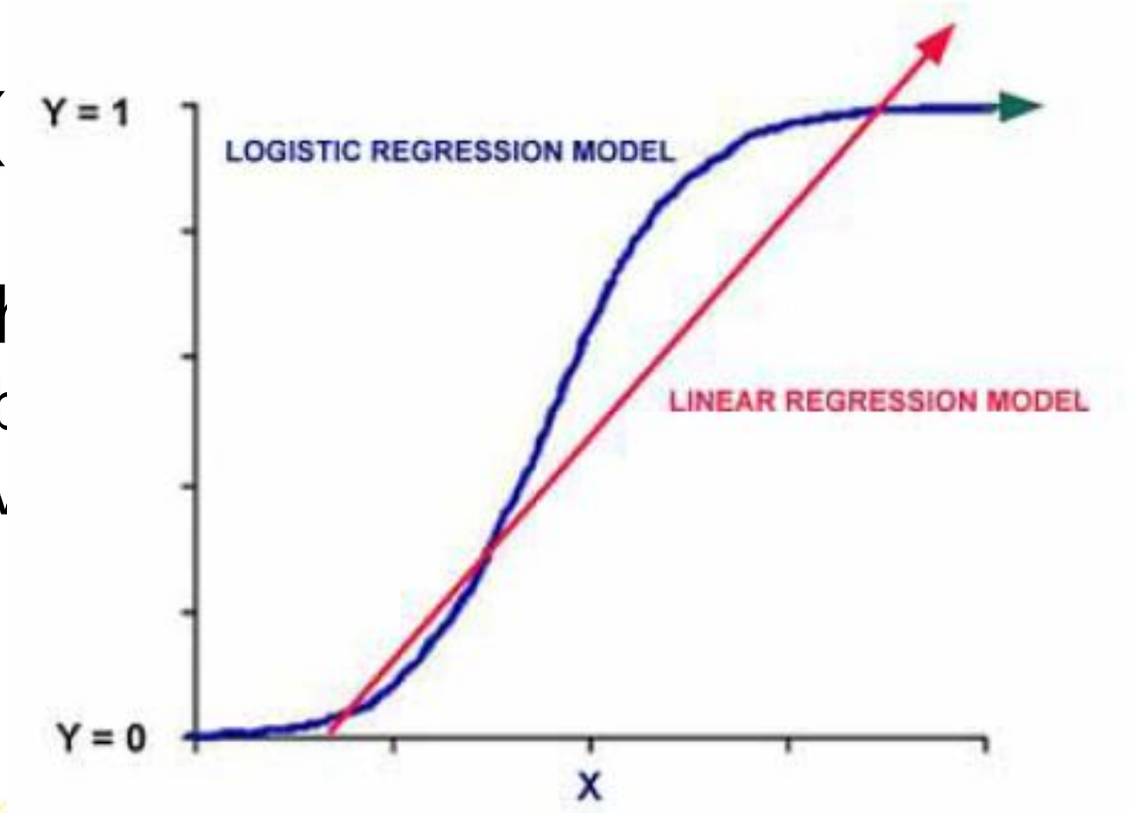
**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95,0% Confidence Interval for B | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | | Lower Bound | Upper Bound |
| 1 | (Constant) | 134,140 | 7,537 | | 17,799 | ,000 | 119,278 | 149,002 |
| | Advertsing Budget (Thousands of Pounds) | ,096 | ,010 | ,578 | 9,979 | ,000 | ,077 | ,115 |
| 2 | (Constant) | 41,124 | 9,331 | | 4,407 | ,000 | 22,722 | 59,525 |
| | Advertsing Budget (Thousands of Pounds) | ,087 | ,007 | ,523 | 11,991 | ,000 | ,073 | ,101 |
| | No. of plays on Radio | 3,589 | ,287 | ,546 | 12,513 | ,000 | 3,023 | 4,154 |

a. Dependent Variable: Album Sales (Thousands)

# Logistic regression

- The predicted variable is binary (eg. 1=correct; 0=false)
- The aim is to predict the probability that the variable has the value 1 (P(y=1))
- For this you use natural logarithms

- ln( 

- Wh
  - b
    v

in the case of binary

# Logistic regression formula

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1 X$$

# An example

- Can we predict whether the DRD is omdat or want on the basis of Relation (subjective, objective) and Genre (spoken, chat, written)?

| | DRD | Relation | Genre | Freq | va |
|---|---|---|---|---|---|
| 1 | Omdat | objective | spoken | 4021 | |
| 2 | Omdat | objective | chat | 1277 | |
| 3 | Omdat | objective | written | 4733 | |
| 4 | Omdat | subjective | spoken | 52 | |
| 5 | Omdat | subjective | chat | 26 | |
| 6 | Omdat | subjective | written | 25 | |
| 7 | Want | objective | spoken | 818 | |
| 8 | Want | objective | chat | 637 | |
| 9 | Want | objective | written | 773 | |
| 10 | Want | subjective | spoken | 1745 | |
| 11 | Want | subjective | chat | 2797 | |
| 12 | Want | subjective | written | 2545 | |
| 13 | | | | | |

# An example

- Step 0: how good is the model without predictors?

**Iteration History**[a,b,c]

| Iteration | | -2 Log likelihood | Coefficients: Constant |
|---|---|---|---|
| Step 0 | 1 | 26927,541 | -,084 |
| | 2 | 26927,541 | -,084 |

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 26927,541

c. Estimation terminated at iteration number 2 because parameter estimates changed by less than ,001.

**Classification Table**[a,b]

| Observed | | | Predicted: DRD: Omdat | Predicted: DRD: Want | Predicted: Percentage Correct |
|---|---|---|---|---|---|
| Step 0 | DRD | Omdat | 10134 | 0 | 100,0 |
| | | Want | 9315 | 0 | ,0 |
| | Overall Percentage | | | | 52,1 |

a. Constant is included in the model.

b. The cut value is ,500

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 0 | Constant | -,084 | ,014 | 34,468 | 1 | ,000 | ,919 |

1. The measure for the quality of the model is -2LL (a type of Chi2)

2. The analysis stops after two steps because it does not improve anymore

3. The model only contains a constant; it simply predicts that all DRDs are “omdat”. This is correct in 52.1% of the cases.

4. The analysis produces a regression coefficient and a test statistic (Wald), comparable to Chi2. We also see the significance test

# An example (cont'd)

- Step 1: how good is a model with two predictors added?

| Iteration | | -2 Log likelihood | Coefficients | | | |
|---|---|---|---|---|---|---|
| | | | Constant | Relation(1) | Genre(1) | Genre(2) |
| Step 1 | 1 | 13889,199 | 1,778 | -3,132 | ,050 | ,387 |
| | 2 | 12675,163 | 2,645 | -4,317 | ,103 | ,807 |
| | 3 | 12417,864 | 3,341 | -5,099 | ,139 | 1,027 |
| | 4 | 12381,844 | 3,761 | -5,533 | ,148 | 1,061 |
| | 5 | 12380,249 | 3,877 | -5,649 | ,149 | 1,062 |
| | 6 | 12380,244 | 3,884 | -5,656 | ,149 | 1,062 |
| | 7 | 12380,244 | 3,884 | -5,656 | ,149 | 1,062 |

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 26927,541

d. Estimation terminated at iteration number 7 because parameter estimates changed by less than ,001.

1. The -2LL is now 12380

2. The analysis stops after seven iterations because the result does not improve anymore

**Omnibus Tests of Model Coefficients**

| | | Chi-square | df | Sig. |
|---|---|---|---|---|
| Step 1 | Step | 14547,296 | 3 | ,000 |
| | Block | 14547,296 | 3 | ,000 |
| | Model | 14547,296 | 3 | ,000 |

3. The change of the fit compared to the previous model is chi2=14547 ($p<.001$)

**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 12380,244[a] | ,527 | ,703 |

4. The fit of the model (+ translation to a type of $R^2$)

# An example (cont'd)

- Step 1: how good is a model with two predictors added?

The classification has improved much (88.0 % in stead of 52.1 %)

**Classification Table[a]**

| Observed | | | Predicted | | |
|---|---|---|---|---|---|
| | | | DRD | | Percentage Correct |
| | | | Omdat | Want | |
| Step 1 | DRD | Omdat | 10031 | 103 | 99,0 |
| | | Want | 2228 | 7087 | 76,1 |
| | Overall Percentage | | | | 88,0 |

a. The cut value is ,500

# An example (cont'd)

- Step 1: how important are the predictors?

Both Relation and Genre give a significant contribution to the prediction

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 1[a] | Relation(1) | -5,656 | ,102 | 3056,246 | 1 | ,000 | ,003 |
| | Genre | | | 338,953 | 2 | ,000 | |
| | Genre(1) | ,149 | ,053 | 7,927 | 1 | ,005 | 1,161 |
| | Genre(2) | 1,062 | ,060 | 311,254 | 1 | ,000 | 2,893 |
| | Constant | 3,884 | ,102 | 1445,937 | 1 | ,000 | 48,623 |

a. Variable(s) entered on step 1: Relation, Genre.

De odds ratio (Exp(B)): when Relation changes from 0 to 1 (when Relation is Objective) the odds that the DRD is "want" become 3/1000 times smaller

# Similarities and differences between linear and logistic regression

- In both types of analyses: a regression formula, coefficients, goodness-of-fit, etcetera
- What is predicted:
  - linear regression: dependent variale; log. regression: probability that dependent variable = 1
- Quality of the model:
  - linear regression: $R^2$; logistic regression: loglikelihood
- Evaluation improvement of the model by adding new factors
  - linear regression: use of F-test, logistic regression: $\chi^2$
- Significance of predictors
  - linear regression: t-test; logistic regression: Wald (type of $\chi^2$)