Multiple Linear Regression
and
an Application on Language Attrition

Gulsen Yilmaz

Rijksuniversiteit Groningen,
April 2009
Outline

- Introduction to multiple linear regression
  - Method of least squares
  - Methods of regression
  - Outliers/residuals
  - Assumptions
- How to run and interpret regression analysis
- The study
Introduction to multiple linear regression

- Investigates relationships between variables using several independent variables and predicts numerical variable
- Effect of each variable can be estimated separately
- Used in econometrics, policy making and also ‘linguistics’
- Difference from correlation: predictive power
- Example: income dependent on education, experience, school performance,…
Simple and multiple linear regression

**Simple regression**
- **D**: use of L2 Dutch (IV)
- **A**: attrition in L1 Turkish (DV)

\[ A = a + bD + e \]

- \( a \): constant (attrition with no Dutch use)
- \( b \): ‘coefficient’ of \( D \)
  - effect of an additional unit of Dutch use on attrition
- \( e \): other factors that influence attrition (error, deviation)

→ mean of the outcome depends on one variable
**Multiple regression**

- **P**: positive attitude towards Dutch culture

\[ A = a + bD + yP + e \]

- **b**: estimated effect of additional use of Dutch on attrition, holding positive attitude constant
- **y**: estimated effect of positive attitude on attrition, holding Dutch use constant

→ mean of the outcome depends on two variables
Multiple linear regression model

- Outcome\(_i\) = Model\(_i\) + error\(_i\)

- \(y_i = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n + \epsilon_i\)

- \(y_i\): outcome
- \(b_1\): coefficient of the first predictor \(x\)
- \(b_2\): coefficient of the second predictor \(x\) and so on
- \(\epsilon_i\): deviations, independent and normally distributed
Method of least squares

- Deviation = \( \sum (\text{observed} - \text{model})^2 \)
- Line of best fit: the line that best describes the data
- The best fit if we have more variables
- Multiple regression: selects a plane so that the sum of squared errors is at a minimum
Scatterplot of the relationship between reaction time in L1, Dutch L2 use and positive attitude towards L2 culture
(hypothetical values)
Where does $R^2$ come from

- $SS_T$, Total Sum of Squares: observed data – mean of outcome
- $SS_R$, Residual Sum of Squares: observed data – regression line
- $SS_M$, Model Sum of Squares: mean of outcome – regression line

$R^2 = \frac{SS_M}{SS_T}$

example: $R^2 = 0.81$

81% of the variability in the outcome is captured by the predictors in the equation
19% residual

Smaller the residual, the better the quality of the model
Methods of regression

Stepwise methods for complex models:
- **Enter**: all predictors at once, builds the complex model all at once
- **Forward**: one predictor at a time, the best predictor, then the second best predictor
- **Backward**: builds the complex model, drops the least good predictor, then the second least good one
Which method to choose

- Not too many predictors
  - i.e. principal component analysis
  - correct children
  - regret if they forget L1 importance of L1 for children
  - saturday classes
  - etc.

- Past research

- Supression: Supressor effects occur when a predictor has a significant effect only when another variable is held constant.

- Forward selection → type 2 error due to supressor effects
Outliers and residuals (regression diagnostics)

- Outlier: very different from the rest of the data
- Influential: case with a large influence on our model
- See both outliers and influentials to assess your model
- But, no justification for data removal to have significant results
Some tips for regression diagnostics

Case summaries on the output:

- **Standardized:**
  - no more than 5% of cases > above 2
  - no more than 1% > above 2.5
  - any case > 3 could be an outlier

- **Cook’s distance:** any value above 1, concern

- **Leverage:** values 0-1, big values concern

- **Mahalanobis distance:** values above 25 (N=500, 5 predictors), and values above 15 (N=100, 3 predictors), concern

- **DFBeta:** greater than 1, concern

- **CVR (covariance ratio):** if close to 1, ok
Assumptions of multiple regression

- **Variables:**
  - Predictor: quantitative or categorical (with two categories)
  - Outcome: quantitative, continuous, unbounded

- **Nonzero variance:** Predictors should have some variation in value

- Predictors should be uncorrelated with external variables
Assumptions cont.

- **No perfect collinearity**: no perfect linear relationship between two or more of the predictors

  → otherwise multicollinearity:
    1. weak explanatory power
    2. difficult to assess the importance of individual factors
    3. unstable predictor equations

  → check: **VIF** (variance inflation factor)
    **tolerance** statistic (1/VIF)
    - largest VIF > 10, concern
    - average VIF > 1, regression maybe biased
    - tolerance < 0.1, serious problem
    - tolerance < 0.2, a potential problem
Assumptions cont.

- **Homoscedasticity**: Residuals at each level of the predictors should have the same variance.
  → check by visual inspection of the residual scatter plot

- **Independent errors**: Errors should be uncorrelated
  → check **Durbin-Watson** test
  - If 2: residuals are uncorrelated, fine
  - Concern: values <1 and values >3
Assumptions cont.

- **Normally distributed errors**: Residuals should be normally distributed with a mean of zero.

- **Independence**: Each value of outcome variable should come from a separate entity.

- **Linearity**: The mean values of the outcome variable for each increment of the predictors lie along a straight line.
How to do multiple regression?

How to interpret the output?
This is how the data looks like on SPSS:

<table>
<thead>
<tr>
<th>group</th>
<th>name</th>
<th>L1</th>
<th>ATTCON</th>
<th>Cutoff</th>
<th>HFAv</th>
<th>MFAv</th>
<th>LFAv</th>
<th>RT</th>
<th>age</th>
<th>ageem</th>
<th>emislen</th>
<th>L1fam</th>
<th>prefcul</th>
<th>social</th>
<th>Igforch</th>
<th>Igprofess</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Steps to run a multiple linear regression:
What the statistics options mean:

**Regression Coefficient**
- Estimates
- Confidence intervals
- Covariance matrix
- Model fit
- R squared change
- Descriptives
- Part and partial correlations
- Collinearity diagnostics

**Residuals**
- Durbin-Watson
- Casewise diagnostics
  - Outliers outside: 2 standard deviations
  - All cases

[Buttons: Continue, Cancel, Help]
Regression plot is a good way to check the assumptions of random errors and homoscedasticity

*ZRESID (standardized residuals, errors)

*ZPRED (standardized predicted values of DV based on the model)
Plot of *ZRESID against *ZPRED
- assumptions of linearity and homoscedasticity met?
- yes, because points are random, widely dispersed, no sign of trend
Histogram of residuals
- assumption of normal distribution of errors met?
- yes, a bell shaped curve means normal distribution

Dependent Variable: HFAv

Mean = -0.04
Std. Dev. = 1.034
N = 69
Plot of residuals
- assumption met?
- yes, straight line represents normal distribution
How to interpret multiple regression

### Anova

#### Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
<th>Change Statistics</th>
<th>Durbin-Watson</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.308&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.095&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.083</td>
<td>142.81332</td>
<td>.095</td>
<td>7.657&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup> Predictors: (Constant), age

<sup>b</sup> Dependent Variable: HFAv

<sup>c</sup> Durbin-Watson

#### ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>1</td>
<td>156173.656</td>
<td>7.657</td>
<td>.007&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>73</td>
<td>20395.644</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1645055.643</td>
<td>74</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Predictors: (Constant), age

<sup>b</sup> Dependent Variable: HFAv
Interpretation

- Look at **F-ratio** and **significance** and $R^2$
- For this data F ratio is 7.657 and significant at $p<.01$
- Regression model predicts the outcome well
- $R^2 = 0.095$
  - age accounts for about 10% variation in the reaction time
- Durbin-Watson is close to 2, so fine
How to interpret multiple regression
Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>95% Confidence Interval for B</th>
<th>Correlations</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td>t</td>
<td>Sig.</td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>800.573</td>
<td>79.913</td>
<td>10.018</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>age</td>
<td>4.738</td>
<td>1.712</td>
<td>0.308</td>
<td>2.767</td>
</tr>
</tbody>
</table>

a. Dependent Variable: HFAv
Interpretation

- Look at **t-ratio** and **significance**

- **t- statistics**: If a variable significantly predicts the outcome, it should have a coefficient significantly different from zero

- For this data **t- ratio** is 10.018, significant at p<.001

- Age is a good predictor
Introduction to the study

- **Aim:** Investigate L1 *attrition* among Turks and Moroccans in the Netherlands

- **Attrition:** “a linguistic system in disuse will be vying for memory space with the other linguistic system(s) occupying the same brain, [...] not being kept ‘fresh’ and ‘strong’ through constant use will somehow weaken it, and [...] it will therefore suffer in some way.” (Schmid, 2006:74)
L1 proficiency in a migrant context

- Limited exposure to L1 and less opportunities to use it
- Attitudes towards L1/ L2 and L1/L2 culture
- Factors that enhance L1 maintenance:
  a large community size, symbolic value of language, cultural and linguistic dissimilarity
- Yet, stability of the native language cannot be guaranteed
Activation Threshold Hypothesis (ATH): an account for attrition

ATH: Language disuse $\rightarrow$ higher thresholds $\rightarrow$ attrition

- First affects lexical items
  - Word finding/retrieval problems
  - Decreased lexical diversity
  - Disfluency in speech
- Word retrieval: 2-5 words/second
  - Conceptualization $\rightarrow$ Formulation $\rightarrow$ Articulation
- Bilingual disadvantage
Predictions of the study

- Lexical access problems: Slower Reaction Times (RTs)
- Despite
  - dominant L1 use
  - strong attachment to L1 and L1 culture
The Study

- Informants: first generation Moroccans (n = 35) and Turks (n = 54)
- Degree of bilingualism: various
- Age at arrival: 14 – 42 (mean: 22.00)
- Age: 28 – 65 (mean: 44.73)
- Length of residence: 10 – 43 years (mean: 22.37)
- Control groups: collected, matched
  (age: 25-62, mean: 43.45)
Research Design

1. Picture Naming Task
   - 78 pictures (26 high, 26 mid, 26 low fam.)
   - no cognates, no ambiguous pictures
   - timed: 3000 ms
   - accuracy and reaction time measured
   - E-prime software

2. Sociolinguistic questionnaire
   - L1 and L2 use, social networks, linguistic/cultural affiliation, attitudes towards language learning

3. Free speech
Variables in multiple regression

Predictors (Independent Variables)
- L1 use in the family
- L1 social use
- Preferred culture
- Importance of L1 for children
- L1 professional use

Outcome (Dependent Variable)
- RT on the PNT task
Outliers

- If half or more than half of the participants couldn’t name an object, item excluded
- If the response was below 250 ms, response excluded
- Cutoff point: those subjects with more than 25% invalid responses get a 0, those with less get a 1
Recode between 0 and 1

Example: Do you consider yourself a bilingual?
1= NL better, 2=bilingual, 3=TR better

original 1=NL better, recoded as 0
original 2=bilingual, recoded as 0.5
original 3= TR better, recoded as 1
Check reliability of subscales

Example:

L1 use in family: nationality of partner, language with partner, with children, with grandchildren

Reliability goes up when grandchildren are omitted
Compute mean for predictors and reaction time

Example: Preferred culture is L1 or L2 culture
COMPUTE prefcul = MEAN(mosque,culture,L1friend,L1club,L1media)

- RT measured in milliseconds
  Total RT (78 items)
  High Fam RT (26 items)
  Medium Fam RT (26 items)
  Low Fam RT (26 items)
Picture Naming Task: Reaction Time

Results

![Bar chart showing mean response times for experimental and control groups with different conditions: HF, MF, and LF. The chart indicates a statistically significant difference (p<.05).]
Results

- Slower RTs in the experimental group compared to controls
- LF significant, HF and MF approaching significance
- So, sign of lexical retrieval difficulties
Predicting performance on the PNT on the basis of L1 use/attitudes

Multiple linear regression
- Attrition not related to variables in question except age

T-tests
- Attrition in only MA group
- TR: maintainers
- MA controls faster than TR controls
Discussion: Why Moroccans differ from Turks

Group level differences:
- MA: early multilinguals (Berber and/or French)
- Turks: no other languages before coming to NL
- Moroccans more open to Dutch language and culture

Individual level factors/predictors:
- Total languages, attitudes not related to attrition
- L2 proficiency may be a potential factor
Discussion: Multiple linear regression

- Why the other predictors turned out to be weak?
- Possible correlation between the predictors?
  i.e. if they prefer L1 culture they would automatically use L1 more
- Enough number of participants?
- Small range of variation in reaction time?
  i.e. only 80 ms yields to significant difference
- What other potential predictors can account for the outcome? i.e. Dutch proficiency, language specific factors in TR and MA
Future of the Study

- Data collection in L2 from the same speakers
- Analysis of spoken data in L1 and L2
- Effects of multicompetence on lexical access in L1 and L2
- Signs of lexical attrition in free speech
- Effects of attrition in other domains
THANK YOU!