



6

Baseline model

- Linear regression: psa_00 school attendance: yes vs. target
- $y = 100 * x - 5$
- Error for Baseline model: +/- 6.6 %
- High school attendance leads to lower present income!

7

Complex model and predictions

- For simplicity we chose **Linear Regression**
- Complex model has slightly better results (target = +/- 4.0 %)

Features	#
Type of dwelling	13
School attendance	5
Satellite TV	2
Car	2
Landline	2
Language	13
Population group	6
Electricity for lighting	1
Piped water access	7
Ward code	1
Coordinates	2
Highlights for the area	1

8

Simpler model and predictions

- Reduced number of features: 11
- Features **easy to access** by stakeholder
- Simple model: target = +/- 4.4 %

Features	#
Type of dwelling	13
School attendance	2
Satellite TV	1
Car	1
Landline	1
Language	2
Population group	2
Electricity for lighting	1
Piped water access	2
Ward code	1
Coordinates	2
Highlights for the area	1

9

Feature Importance

Features containing informations about school attendance and owning a car are most important to the models

10

Clustering the data

In comparison to cluster 0 and 2, cluster 1 has higher percentage of women headed households with income below R 19.6 K

11

Conclusions and Recommendations

- Conclusions
- Recommendations for the stakeholder:
- Outlook:

	Error	No of features
Baseline model	6.6	1
Complex model	4.0	51
Simple model	4.4	11

Conclusions and Recommendations

- Conclusions

	Error	No of features
Baseline model	6.6	1
Complex model	4.0	51
Simple model	4.4	11

- Recommendations for the stakeholder:

- Focus on **essential features** (simpler model)
- Provide financial support for families with high **school attendance**
- Improve **public transportation** might have an impact
- Population group/spoken language/landline ownership** still have some impact

- Outlook:

- include **geographical data** into model
- a higher accuracy might be obtained using even a **higher complex model** i.e. Neural Network
- try data from **other years** on the model

