Switching Regression as Robust Estimation against Misclassification in Machine Learning Classification

Aleksandr Michuda

UC Davis Agricultural and Resource Economics

Introduction

- How does urban labor supply respond to agricultural income shocks?
  - Rural-Urban linkages
  - Driving Uber as insurance/income diversification
- Can machine learning help when location data is unavailable?
  - Requires less data
  - But introduces misclassification
  - **Can we develop an estimator that is robust to that misclassification?**

Data

- Uber Driver Data
  - Hours online
  - Earnings
- Weather Shocks
  - Drought Indices (SPI, NDVIA)
- Predicted Rural Place Origin
  - SAP Region (4)
  - FAO Agro-ecological Zones (10)
    - Distribution of probabilities

Prediction Table

Name	Central	East	North	West
Ahimbisibwe	0.144	0.003	0.001	0.925
Amin	0.149	0.057	0.651	0.140
Auma	0.040	0.267	0.674	0.017
Kadaga	0.148	0.797	0	0.054
Makubuya	0.964	0.018	0	0.017
Museveni	0.164	0.022	0.042	0.769
Oculi	0.015	0.263	0.717	0.003

Misclassification is a problem

- Predictions might contain misclassification error
  - For ex: we are classifying "Amin" into North, but it might be that they are actually more connected to the Center.
  - Not from any systematic bias during machine learning process
  - We can "imperfectly" break drivers into groups

What if I use OLS?

- We can estimate with OLS
- But response to drought will be attenuated.
- Is there a better way to estimate it?
  - We can model the misclassification directly
  - "What's the probability that I categorize Amin into the North regime, given that they're truly from the Center?, etc."

Objectives Today

- Presenting a Maximum Likelihood Estimator
  - inspired by switching regression literature
- How does this estimator perform under varying levels of misclassification compared to OLS?
- Explore through Monte Carlo Simulations

The Hours Function

- Each regime $i \in I$ can be expressed as a linear function of SPI^i and *Hours*:

$$Hours = \beta_0^i + \beta_1^i SPI^i + \varepsilon^i$$

- The goal is to recover β_1^0 and β_1^1

Without Misclassification

- Suppose there's a true membership indicator, I.
- The conditional expectation function is then:

$$E(Hours|I, SPI^i) = 1\{I = 0\}(\beta_0^0 + \beta_1^0 SPI^0) + 1\{I = 1\}(\beta_0^1 + \beta_1^1 SPI^1)$$

- Without misclassification and with a separation indicator, we can recover β_1^0 and β_1^1 without bias, by using I as a variable in an OLS regression.

Misclassification in Regimes

- In our case we do not observe I, but we do observe r.
- r gives us a measure of I with *measurement error*.
- We can express the measurement error in terms of a matrix of conditional probabilities with $p_i^j = Pr(r = i | I = j)$.

-	$r = 0$	$r = 1$
$I = 0$	p_0^0	p_1^0
$I = 1$	p_0^1	p_1^1

- If $p_0^1 = p_1^0 = 0$, then there is no misclassification.

Conditional Expectation with Misclassification

- In the case of misclassification, the conditional expectation function is then:

$$
\begin{aligned}
&E(Hours|r) = \\
&\overbrace{(\beta_0^0 + \beta_1^1 SPI^0)}^{E(Hours|I=0)} \cdot (1-r) \cdot \overbrace{(1-\lambda)p_0^0}^{Pr(r=0,I=0)} + \\
&\overbrace{(\beta_0^1 + \beta_1^1 SPI^1)}^{E(Hours|I=1)} \cdot (1-r) \cdot \overbrace{\lambda p_0^1}^{Pr(r=0,I=1)} + \\
&\overbrace{(\beta_0^0 + \beta_1^0 SPI^0)}^{E(Hours|I=0)} \cdot r \cdot \overbrace{(1-\lambda)p_1^0}^{Pr(r=1,I=0)} + \\
&\overbrace{(\beta_0^1 + \beta_1^1 SPI^1)}^{E(Hours|I=1)} \cdot r \cdot \overbrace{\lambda p_1^1}^{Pr(r=1,I=1)}
\end{aligned}
$$

- $Pr(I=1) = \lambda$

Using OLS with Misclassification

- If we use the same OLS strategy as before:

$$Hours = 1\{r=0\} + 1\{r=1\} + \beta_1^0 SPI^0 \cdot 1\{r=0\} + \beta_1^1 SPI^1 \cdot 1\{r=1\} + \varepsilon$$

- Leads to biased estimate, proportionate to extent of misclassification
  - $ABias(\beta^0) = \frac{(1-\lambda)p_0^1}{p_0^0+p_0^1} \cdot (\Sigma_{00}^{-1}\Sigma_{01}\beta^1 - \beta^0)$
    - $\beta^r = [\beta_0^r \ \beta_1^r]$, $\Sigma_{jk} = E(x_j' x_k)$
  - $ABias(\beta^1) = \frac{\lambda p_1^0}{p_1^0+p_1^1} \cdot (\Sigma_{11}^{-1}\Sigma_{10}\beta^0 - \beta^1)$
    - $\beta^r = [\beta_0^r \ \beta_1^r]$, $\Sigma_{jk} = E(x_j' x_k)$
    - $x_r = [1 \ SPI^r]$

ML Approach

- Generalizing Lee and Porter (1985) to more than two regimes
  - Switching Regression with imperfect sample separation
- Flatten probabilities to a categorical
  - Take maximum of probabilities as truth, r

original_name	Central	East	North	West	Region Indicator (r)
Ahimbisibwe	0.144	0.003	0.001	0.925	West
Amin	0.149	0.057	0.651	0.140	North
Auma	0.040	0.267	0.674	0.017	North
Kadaga	0.148	0.797	0	0.054	East
Makubuya	0.964	0.018	0	0.017	Central
Museveni	0.164	0.022	0.042	0.769	West
Oculi	0.015	0.263	0.717	0.003	North

A Maximum Likelihood Alternative

- Each regime is normally distributed with mean $Hours - \beta_0^r - \beta_1^r SPI^r$ and standard deviation σ_r, with density f_r.
- We can then write the joint density of ε_r and r as:

$$f(\varepsilon_r, r) = f_0(\varepsilon_0)\left[r\lambda p_0^0 + (1-r)\lambda(1-p_0^0)\right] +$$
$$f_1(\varepsilon_1)\left[r(1-\lambda)(1-p_1^1) + (1-r)(1-\lambda)p_1^1\right]$$

The Likelihood Function

- The likelihood function of the estimator is then:

$$
\begin{aligned}
&L(\beta, \sigma, p, \lambda) = \\
&[f_0(\varepsilon_{i1t})\lambda p_{11} + f_1(\varepsilon_{i1t})(1-\lambda)p_{10}]^r \\
&\cdot [f_0(\varepsilon_{i0t})\lambda(1-p_{11}) + f_1(\varepsilon_{i0t})(1-\lambda)(1-p_{10})]^{1-r}
\end{aligned}
$$

- We can maximize the log-likelihood to find optimal parameters for each of the parameters above.
- We can run Monte Carlo simulations of the MLE and an OLS analogue to compare the performance of the estimator.

Baseline Values for Simulation

- Data is modelled as crossection
- Actual data is panel

Parameter
Simulations in each =200
Drivers = 275
Time periods = 10
Regimes=2
$\sigma_0 = \sigma_1 = 1$
$E(SPI^0) = E(SPI^1) = 0$
$Var(SPI^0) = Var(SPI^1) = 1$
$Cov(SPI^0, SPI^1) = 0$
$\beta_0^0 = 20$, $\beta_0^1 = 35$
$\beta_1^0 = -1$. $\beta_1^1 = -2$

How is misclassification created?

Misclassification Plots $R = 2$

- Increase severity of misclassification

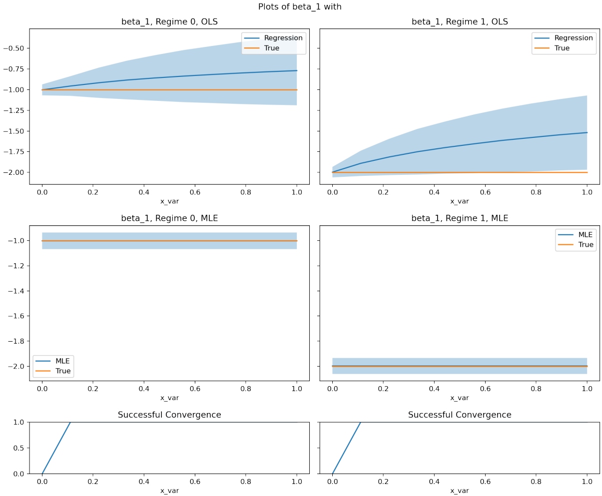


Figure 1: Increasing Misclassification $R = 2$

Generalizing to $R > 2$

- For $R > 2$, r becomes a categorical variable and we now use the mutually exclusive and exhaustive indicator functions for each regime, $G_i \equiv 1\{r = i\}$
- There are now $R - 1$, λ parameters
- The probability matrix will be an RxR matrix
- Consider R=3:

	$G_0 = 1$	$G_1 = 1$	$G_2 = 1$
$I = 0$	p_0^0	p_1^0	p_2^0
$I = 1$	p_0^1	p_1^1	p_2^1
$I = 2$	p_0^2	p_1^2	p_2^2

Generalizing to $R > 2$

- The likelihood function now becomes:

$$L(\beta, \sigma, p, \lambda) =$$

$$\left[f_0(\varepsilon_{i0t})\lambda_0 p_0^0 + f_1(\varepsilon_{i0t})\lambda_1 p_0^1 + f_2(\varepsilon_{i0t})(1-\lambda_0-\lambda_1)p_0^2\right]^{G_0}$$

$$\cdot \left[f_0(\varepsilon_{i1t})\lambda_0 p_1^0 + f_1(\varepsilon_{i1t})\lambda_1 p_1^1 + f_2(\varepsilon_{i1t})(1-\lambda_0-\lambda_1)p_1^2\right]^{G_1}$$

$$\cdot \left[f_0(\varepsilon_{i2t})\lambda_0 p_2^0 + f_1(\varepsilon_{i2t})\lambda_1 p_2^1 + f_2(\varepsilon_{i2t})(1-\lambda_0-\lambda_1)p_2^2\right]^{G_2}$$

Baseline Values for Simulation ($R = 3$)

- Unless otherwise stated the values of each parameter in question will be as follows:

Parameter
Simulations in each=200
Drivers =275
Time Periods =10
Regimes =3
$\sigma_0 = \sigma_1 = \sigma_2 = 1$
$E(SPI_0) = E(SPI_1) = E(SPI_2) = 0$
$Var(SPI_0) = Var(SPI_1) = Var(SPI_2) = 1$
$Cov(SPI_j, SPI_k) = 0$
$\beta_0^0 = 10$, $\beta_0^1 = 20$, $\beta_0^2 = 35$
$\beta_1^0 = -1$,$\beta_1^1 = -2$, $\beta_1^2 = -3$

Misclassification Plot $R = 3$

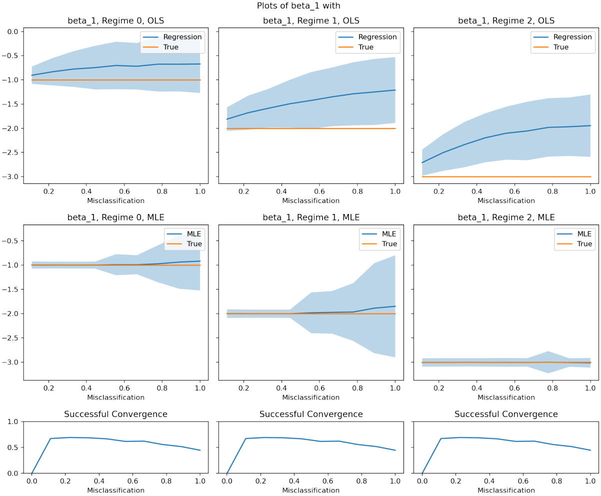


Figure 2: Increasing Misclassification $R = 3$

Conclusion

- MLE method is robust to misclassification
  - but converges less often with more regimes
  - better ways to specify function or calculate standard errors?
- How best to sell results?
- Regressions using real data require many regimes
  - OLS regressions suggest promising results

Misclassification 2 Beta 0

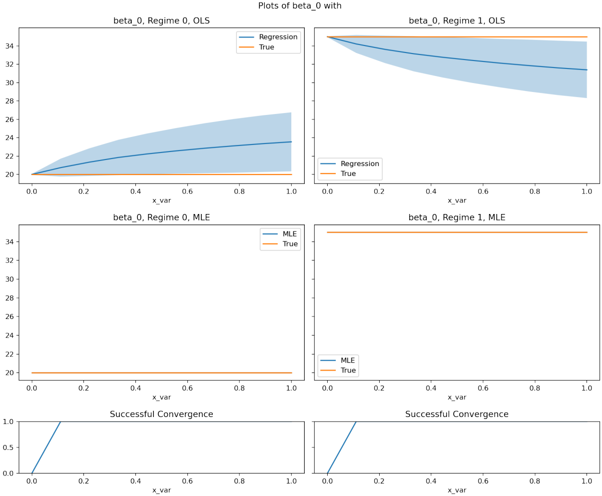


Figure 3: Changing STN of Hours

Misclassification 2 Sigma

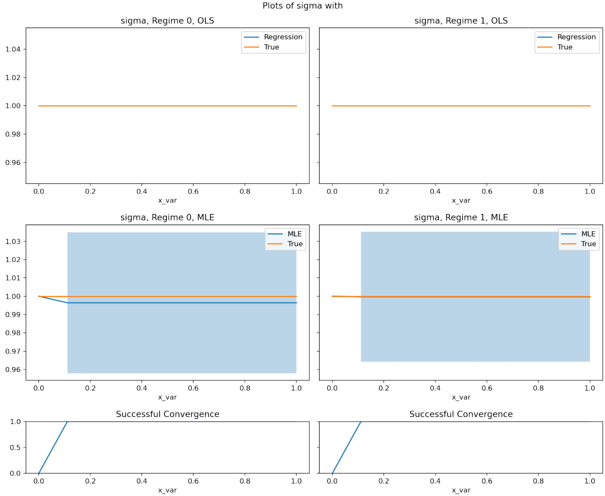


Figure 4: Changing STN of Hours

Misclassification 3 Beta 0

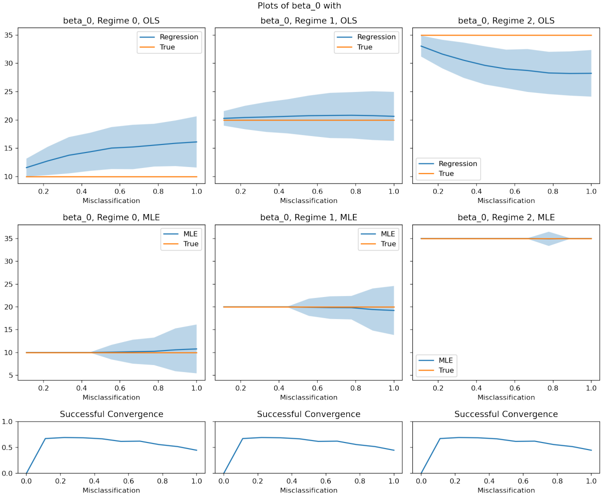


Figure 5: Changing STN of Hours

Misclassification 3 Sigma

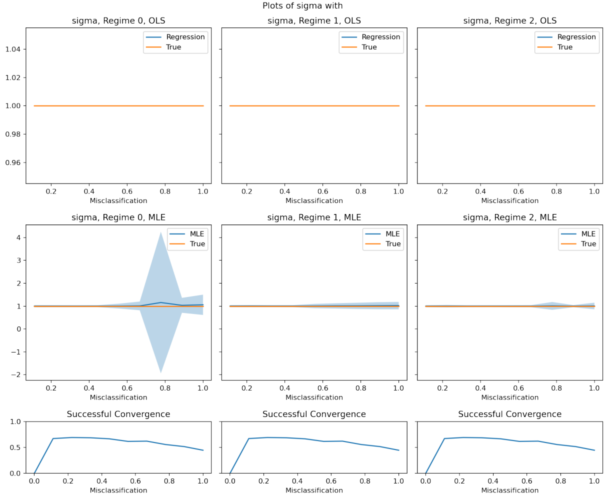
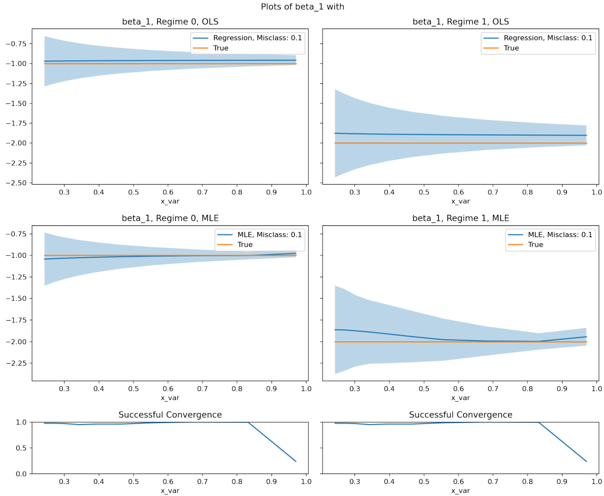


Figure 6: Changing STN of Hours

Noise to Signal Ratio of Hours

- Focus on increasing the signal to noise ratio symmetrically across the two regimes
- $STN = \frac{E(y_r)}{\sigma_r}$



Correlation of SPI Shocks

- Increase correlation between SPI variables
  - Increase from 0 to 0.9

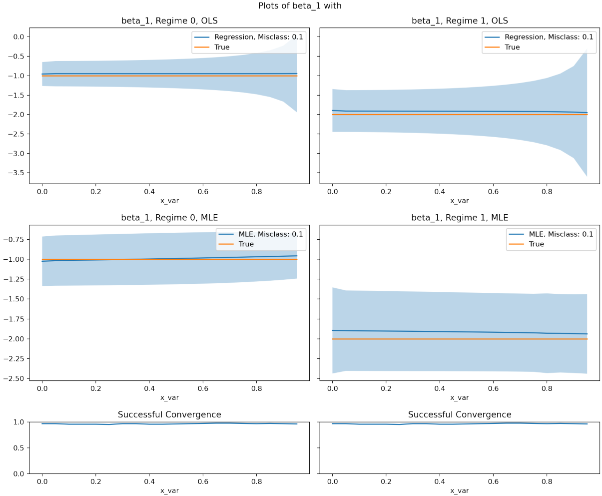


Figure 8: Changing Correlation of Drought Shocks

Difference across Regime Responses

- $\beta_1^0 = 0$
- β_1^1 ranges from 0 to 2

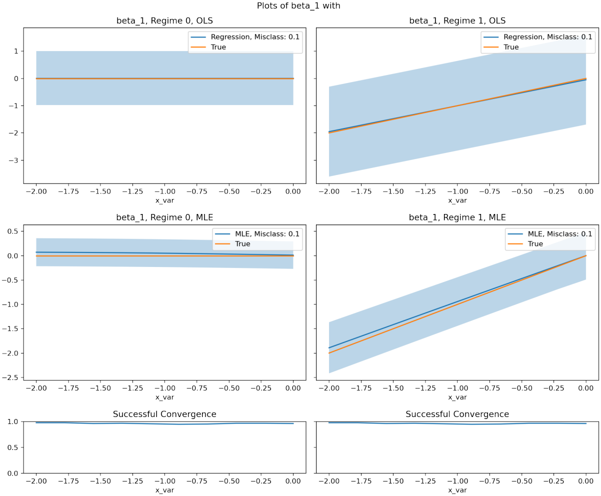


Figure 9: Regime Response Heterogeneity

Noise to Signal Ratio of Hours ($R = 3$)

- Same idea as before, but three regimes now

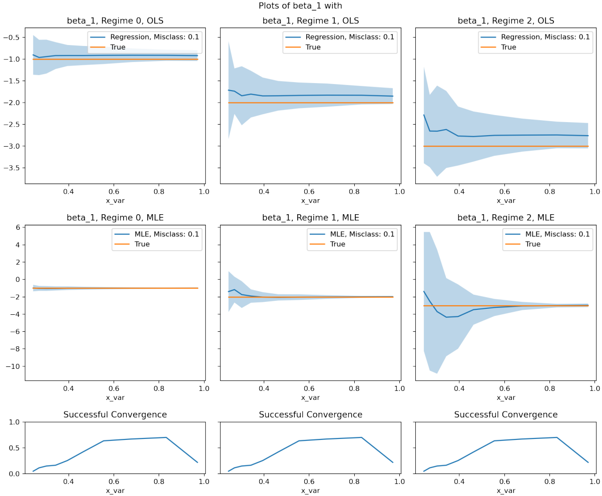


Figure 10: Changing STN of Hours

Correlation of Drought Shocks ($R = 3$)

- Increase correlation between drought variables
  - Increase from 0 to 0.9

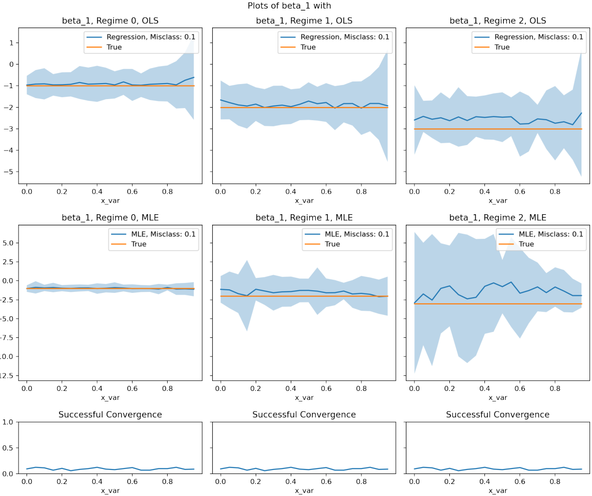


Figure 11: Changing Correlation of Drought Shocks, $R = 3$

Difference across Regime Responses ($R = 3$)

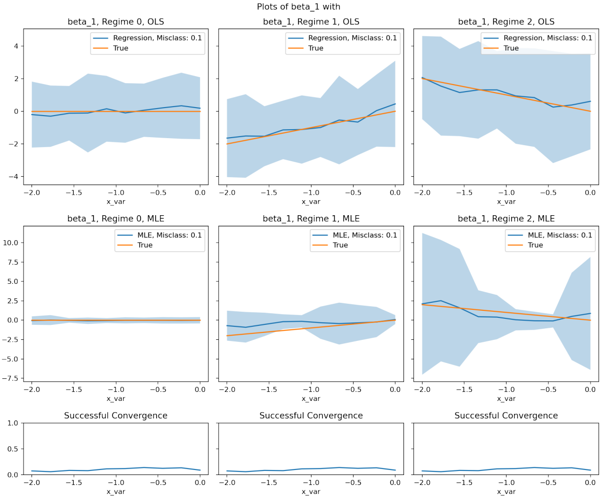
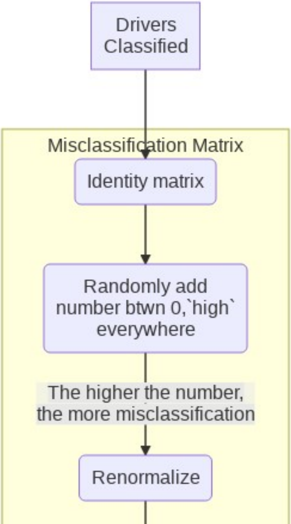


Figure 12: Changing Regime Response Heterogeneity, $R = 3$

β_0 σ

Misclassification Procedure

- The misclassification matrix is a “jittered” matrix that introduces misclassification to the drivers after their memberships have already been chosen.



Two Regimes STN Sigma

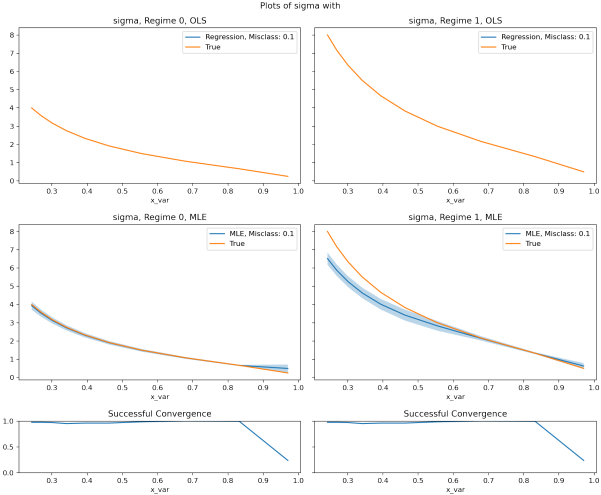


Figure 14: Two Regimes STN Sigma

Back

Two Regimes STN Beta 0

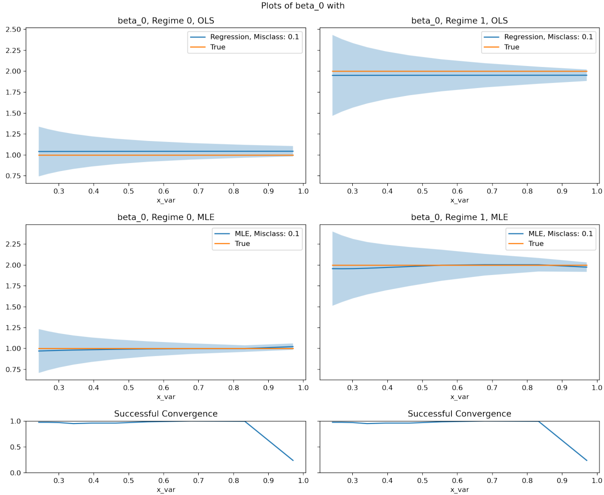


Figure 15: Two Regimes STN β_0

Back

Two Regimes Drought Correlation Sigma

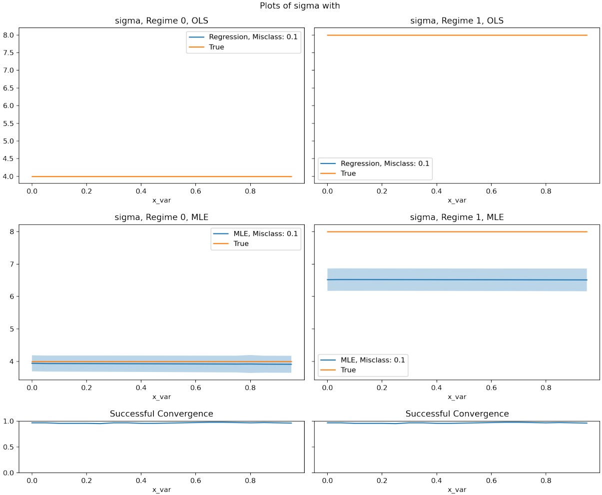


Figure 16: Two Regimes Drought Correlation Sigma

Back

Two Regimes Drought Correlation Beta 0

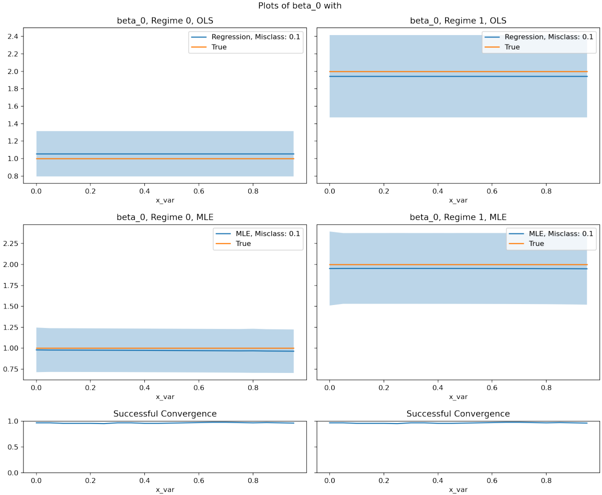


Figure 17: Two Regimes Drought Correlation Beta 0

Back

Two Regimes Response Sigma

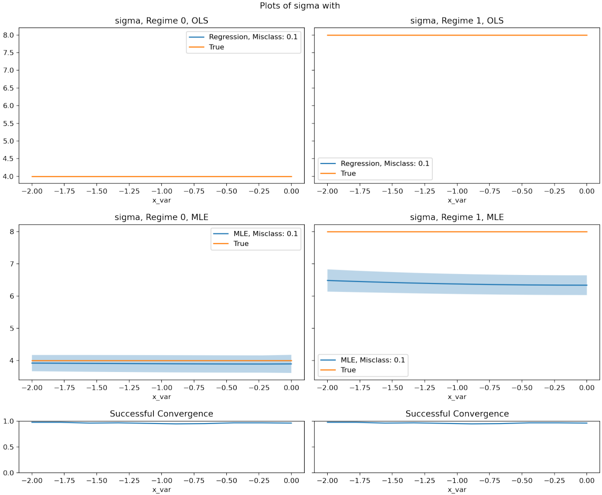


Figure 18: Two Regimes Response Sigma

Back

Two Regimes Response Beta 0

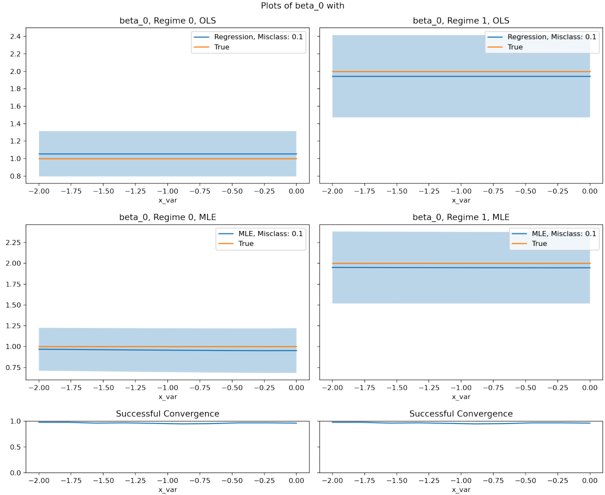


Figure 19: Two Regimes Response Sigma

Back

Three Regimes STN Sigma

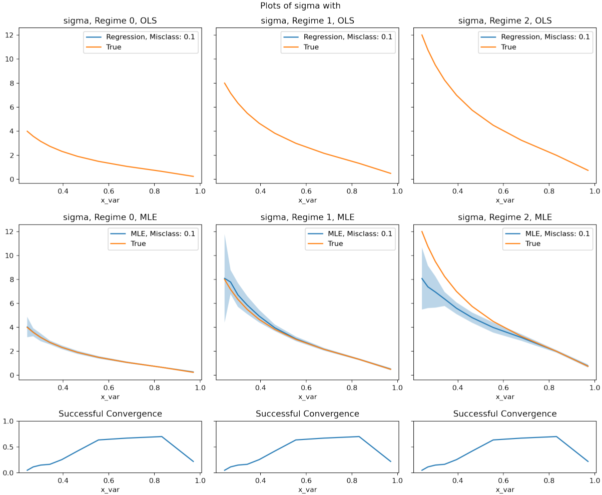


Figure 20: Three Regimes STN Sigma

Back

Three Regimes STN Beta 0

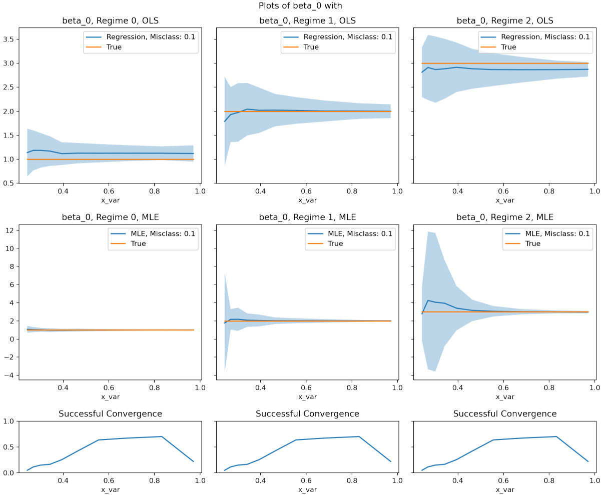
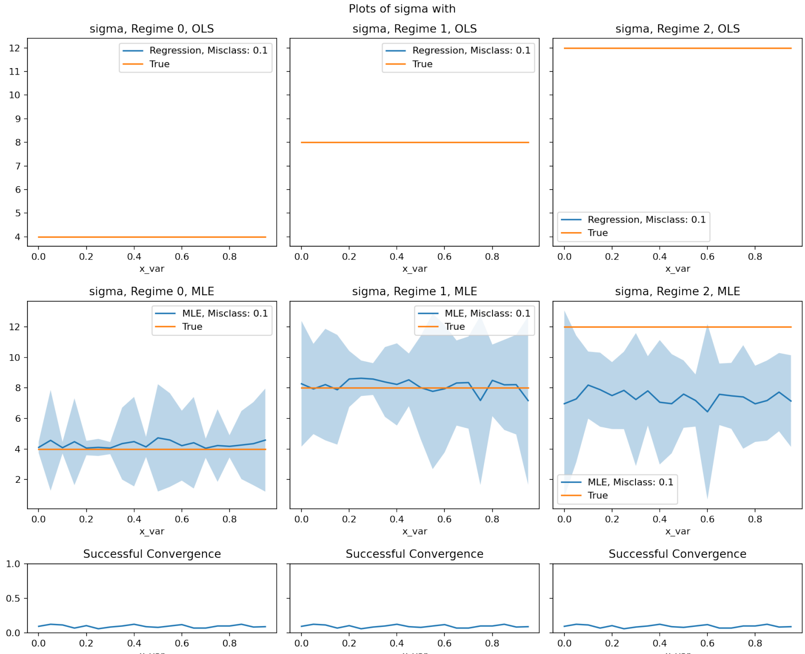


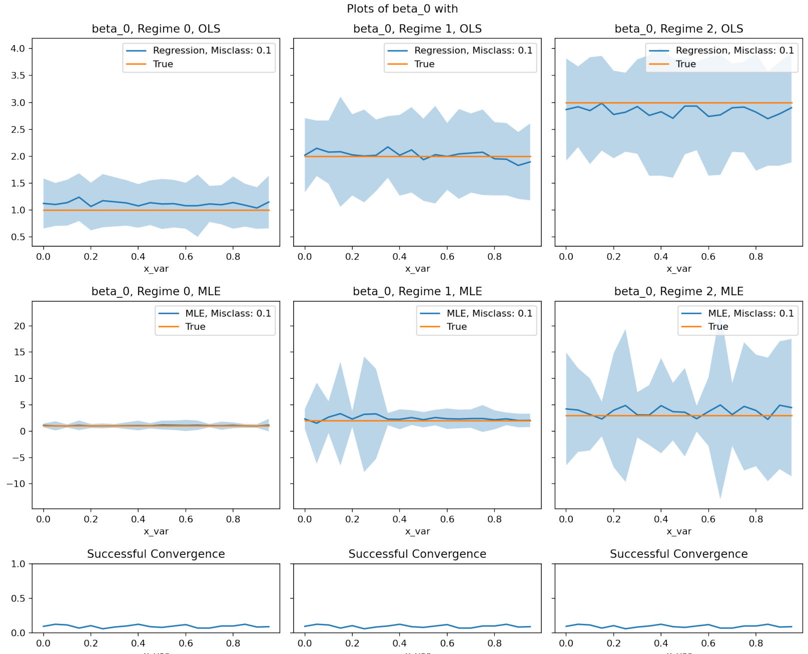
Figure 21: Three Regimes STN β_0

Back

Three Regimes Drought Correlation Sigma



Three Regimes Drought Correlation Beta 0



Three Regimes Response Sigma

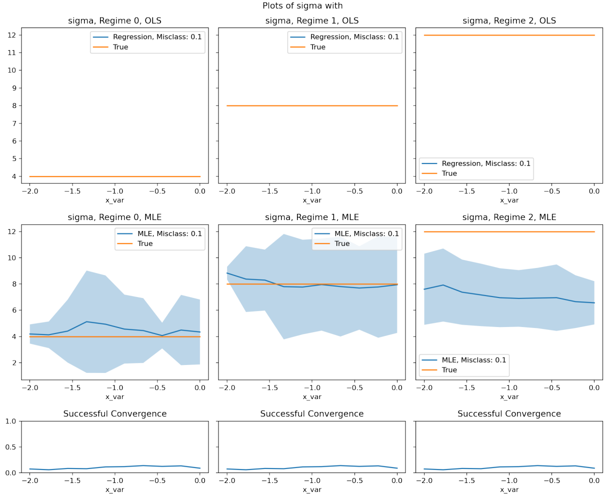


Figure 24: Three Regimes Response Sigma

Back

Three Regimes Response Beta 0

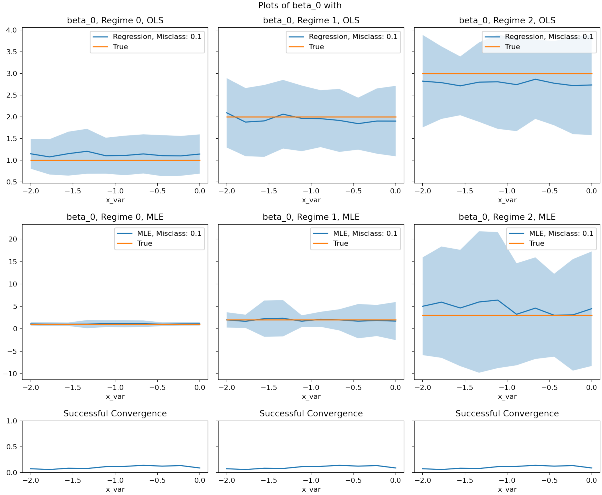


Figure 25: Three Regimes Response Sigma

Back

OLS Regressions on Region

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
region_class_Central	74.024	0.7525	98.372	0.0000	72.549	75.499
region_class_East	69.941	0.8790	79.570	0.0000	68.219	71.664
region_class_North	63.988	1.1560	55.354	0.0000	61.723	66.254
region_class_West	74.381	0.7921	93.902	0.0000	72.828	75.933
region_class_Central:lagged_Central	0.1891	0.0584	3.2357	0.0012	0.0746	0.3037
region_class_East:lagged_East	0.1619	0.0529	3.0634	0.0022	0.0583	0.2655
region_class_North:lagged_North	0.1621	0.0927	1.7485	0.0804	-0.0196	0.3437
region_class_West:lagged_West	0.1224	0.0551	2.2221	0.0263	0.0144	0.2303

Figure 26: Regression Results using SAP Region

MLE Estimates on Region

- Unavailable as MLE does not converge.