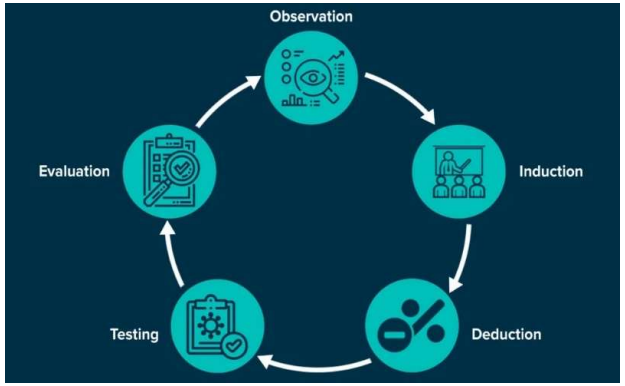
Empirical research in management and economics
Exercise

Thorsten Pachur, Linus Hof, Rebecca West, Sebastian Hellmann, Nuno Busch

Technical University of Munich
School of Management
Chair of Behavioral Research Methods



Please remember to take part in the brief mid-term survey!

https://forms.gle/bt71nwC1Q2r247hE6

(Please participate at the survey irrespective of whether you are attending in person or watching the recording.)

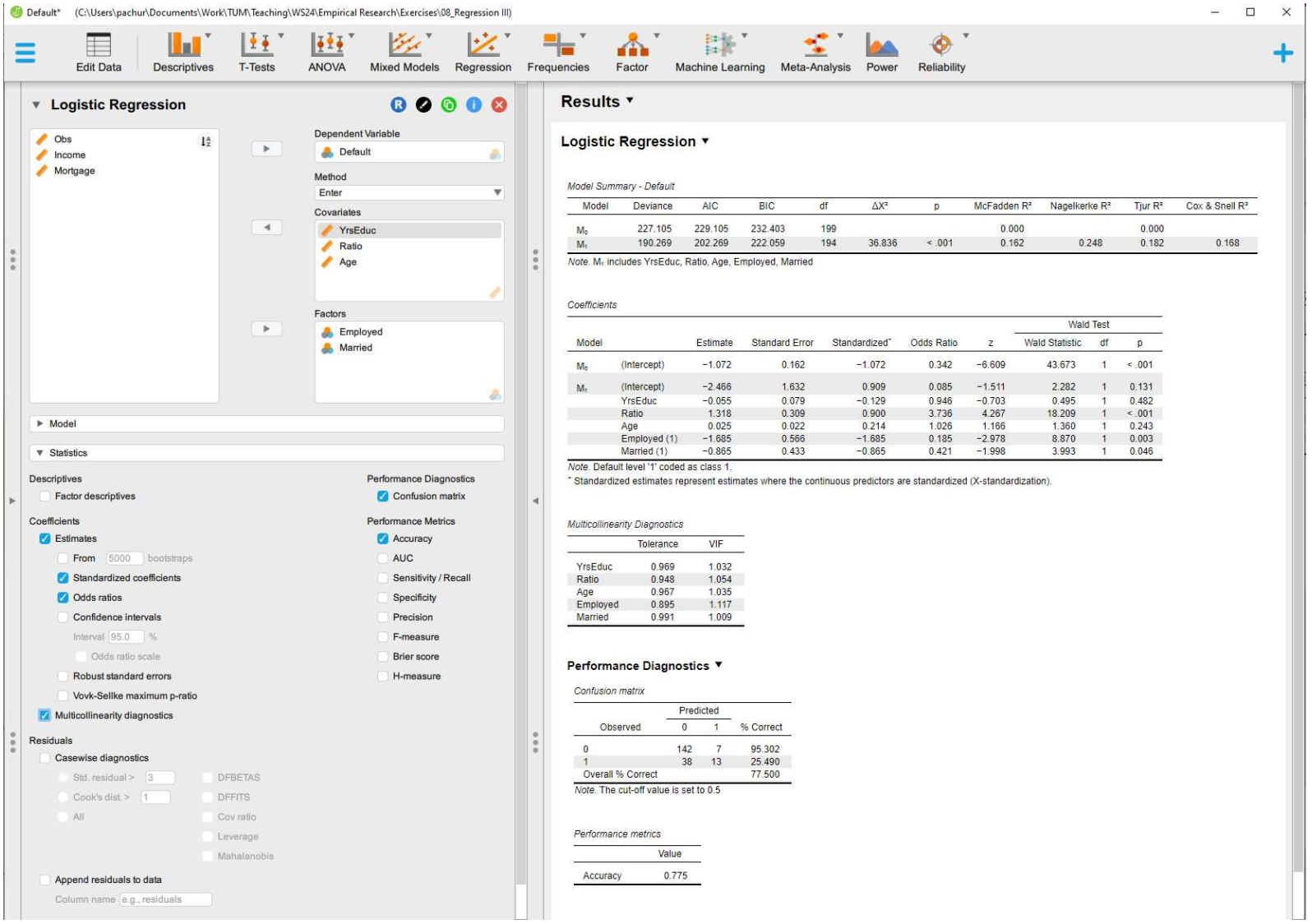
Exercise: Logistic regression I

- Open the dataset “Default.csv“ in JASP

→Sample of mortgage loans from 200 borrowers. For each loan, it is given whether the loan defaulted or not and some information on the borrower:

  - Age
  - Years of education
  - Employed (yes, no)
  - Married (yes, no)
  - Ratio of mortgage to income

- Conduct a logistic regression analysis
  - What factors predict occurrence of a default—and how? Also interpret the odds ratios in the JASP output.
  - Calculate the predicted probability of a default of a $235,500 loan by a 48 year-old, employed, and married borrower, who has an annual income of $70,090 and 16 years of education.
  - Do the same for a 47 year-old borrower of a $ 367,300 loan, who is employed and unmarried, and has an annual income of $115,880 and 12 years of education.


Default* (C:\Users\pachur\Documents\Work\TUM\Teaching\WS24\Empirical Research\Exercises\08_Regression III)
Edit Data
Descriptives
T-Tests
ANOVA
Mixed Models
Regression
Frequencies
Factor
Machine Learning
Meta-Analysis
Power
Reliability
Logistic Regression
Obs
Income
Mortgage
Dependent Variable
Default
Method
Enter
Covariates
YrsEduc
Ratio
Age
Factors
Employed
Married
Model
Statistics
Descriptives
Factor descriptives
Coefficients
Estimates
From 5000 bootstraps
Standardized coefficients
Odds ratios
Confidence intervals
Interval 95.0 %
Odds ratio scale
Robust standard errors
Vovk-Sellke maximum p-ratio
Multicollinearity diagnostics
Residuals
Casewise diagnostics
Std. residual > 3
Cook's dist. > 1
All
DFBETAS
DFFITS
Cov ratio
Leverage
Mahalanobis
Append residuals to data
Column name e.g., residuals
Performance Diagnostics
Confusion matrix
Performance Metrics
Accuracy
AUC
Sensitivity / Recall
Specificity
Precision
F-measure
Brier score
H-measure
Results
Logistic Regression
Model Summary - Default
Model Deviance AIC BIC df ΔX² p McFadden R² Nagelkerke R² Tjur R² Cox & Snell R²
M0 227.105 229.105 232.403 199 0.000 0.000
M1 190.269 202.269 222.059 194 36.836 < .001 0.162 0.248 0.182 0.168
Note. M1 includes YrsEduc, Ratio, Age, Employed, Married
Coefficients
Model Estimate Standard Error Standardized* Odds Ratio z Wald Test Wald Statistic df p
M0 (Intercept) −1.072 0.162 −1.072 0.342 −6.609 43.673 1 < .001
M1 (Intercept) −2.466 1.632 0.909 0.085 −1.511 2.282 1 0.131
YrsEduc −0.055 0.079 −0.129 0.946 −0.703 0.495 1 0.482
Ratio 1.318 0.309 0.900 3.736 4.267 18.209 1 < .001
Age 0.025 0.022 0.214 1.026 1.166 1.360 1 0.243
Employed (1) −1.685 0.566 −1.685 0.185 −2.978 8.870 1 0.003
Married (1) −0.865 0.433 −0.865 0.421 −1.998 3.993 1 0.046
Note. Default level '1' coded as class 1.
* Standardized estimates represent estimates where the continuous predictors are standardized (X-standardization).
Multicollinearity Diagnostics
Tolerance VIF
YrsEduc 0.969 1.032
Ratio 0.948 1.054
Age 0.967 1.035
Employed 0.895 1.117
Married 0.991 1.009
Performance Diagnostics
Confusion matrix
Observed Predicted 0 Predicted 1 % Correct
0 142 7 95.302
1 38 13 25.490
Overall % Correct 77.500
Note. The cut-off value is set to 0.5
Performance metrics
Value
Accuracy 0.775

Predicted probability of default

Regression equation (predicted log odds)

$$\log\left(\frac{\widehat{p(default)}}{p(no\ default)}\right) = -2.466 - 0.055 \times YrsEduc + 1.318 \times Ratio + 0.025 \times Age - 1.685 \times Employed - 0.865 \times Married$$

YrsEduc = 16 Ratio = 235,500/70,090 = 3.36 Age = 48 Employed = 1 Married = 1

Exponentiated regression equation (predicted odds)

$$\frac{\widehat{p(default)}}{p(no\ default)} = e^{-2.466-0.055\times16+1.318\times3.36+0.025\times48-1.685\times1-0.865\times1} = e^{-0.268} = .765$$

Predicted probability from predicted odds

$$\widehat{p(default)} = \frac{.765}{1+.765} = .43$$

The person has a 43% chance of not being able to pay back the loan.

Predicted probability of default

Regression equation (predicted log odds)

$$\log\left(\frac{\widehat{p(default)}}{p(no\ default)}\right) = -2.466 - 0.055 \times YrsEduc + 1.318 \times Ratio + 0.025 \times Age - 1.685 \times Employed - 0.865 \times Married$$

YrsEduc = 12 Ratio = 367,300/115,880 = 3.17 Age = 47 Employed = 1 Married = 0

Exponentiated regression equation (predicted odds)

$$\frac{\widehat{p(default)}}{p(no\ default)} = e^{-2.466-0.055\times12+1.318\times3.17+0.025\times47-1.685\times1-0.865\times0} = e^{0.542} = 1.720$$

Predicted probability from predicted odds

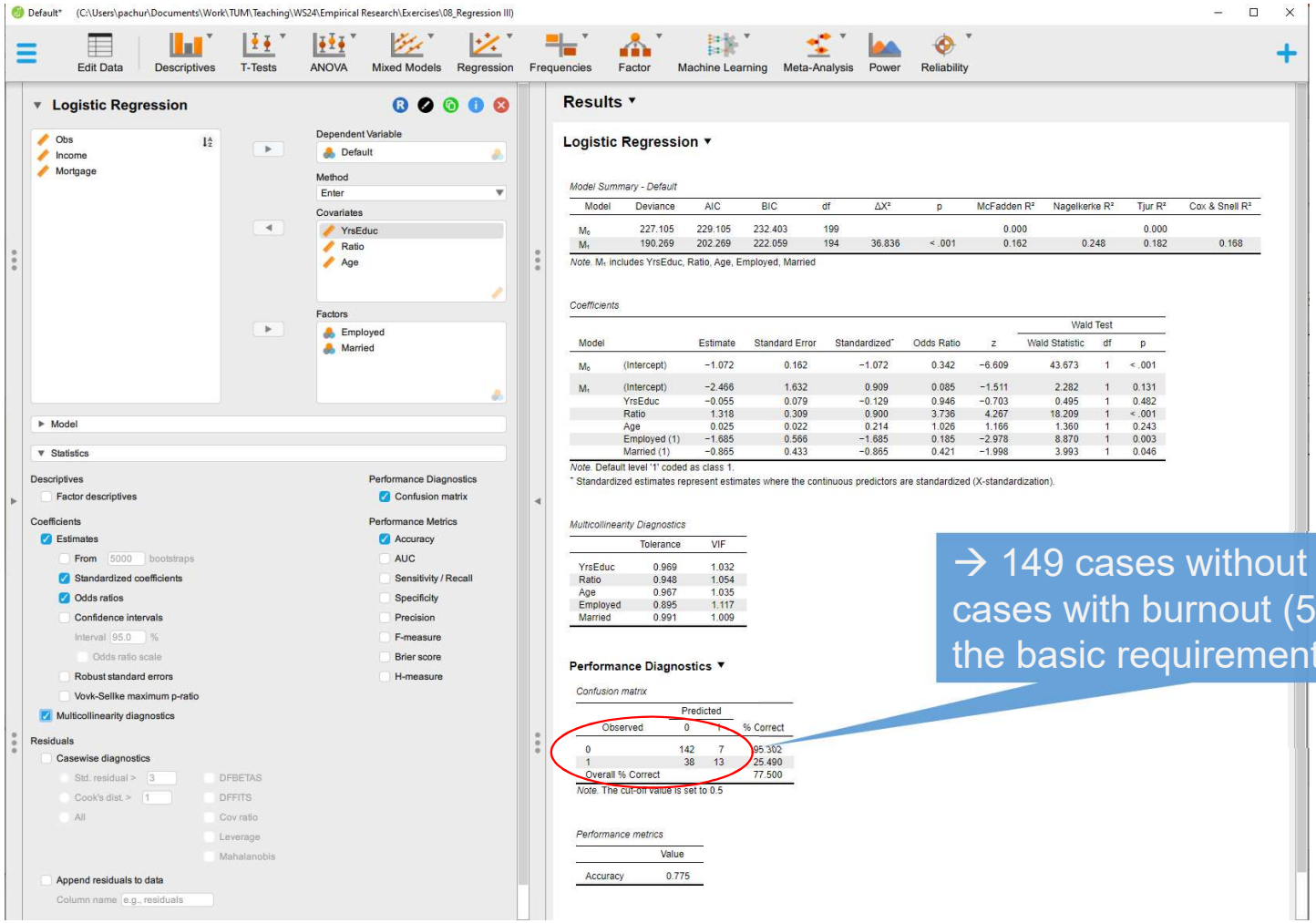
$$\widehat{p(default)} = \frac{1.720}{1+1.720} = .63$$

The person has a 63% chance of not being able to pay back the loan.

Exercise: Sample-size considerations

Assess whether in the dataset “Default.csv“ the sample-size considerations discussed in the lecture are fullfilled

- How many cases are there for the different levels of the dependent variable and which level occurs less frequently?
- What is the minimum number of cases given the number of predictors in the dataset?


Default* (C:\Users\pachur\Documents\Work\TUM\Teaching\WS24\Empirical Research\Exercises\08_Regression III)
Edit Data
Descriptives
T-Tests
ANOVA
Mixed Models
Regression
Frequencies
Factor
Machine Learning
Meta-Analysis
Power
Reliability
Logistic Regression
Obs
Income
Mortgage
Dependent Variable
Default
Method
Enter
Covariates
YrsEduc
Ratio
Age
Factors
Employed
Married
Model
Statistics
Descriptives
Factor descriptives
Coefficients
Estimates
From 5000 bootstraps
Standardized coefficients
Odds ratios
Confidence intervals
Interval 95.0 %
Odds ratio scale
Robust standard errors
Vovk-Sellke maximum p-ratio
Multicollinearity diagnostics
Residuals
Casewise diagnostics
Std. residual > 3
Cook's dist. > 1
All
DFBETAS
DFFITS
Cov ratio
Leverage
Mahalanobis
Append residuals to data
Column name e.g., residuals
Performance Diagnostics
Confusion matrix
Performance Metrics
Accuracy
AUC
Sensitivity / Recall
Specificity
Precision
F-measure
Brier score
H-measure
Results
Logistic Regression
Model Summary - Default
Model Deviance AIC BIC df ΔX² p McFadden R² Nagelkerke R² Tjur R² Cox & Snell R²
M0 227.105 229.105 232.403 199 0.000 0.000
M1 190.269 202.269 222.059 194 36.836 < .001 0.162 0.248 0.182 0.168
Note. M1 includes YrsEduc, Ratio, Age, Employed, Married
Coefficients
Model Estimate Standard Error Standardized* Odds Ratio z Wald Test: Wald Statistic df p
M0 (Intercept) −1.072 0.162 −1.072 0.342 −6.609 43.673 1 < .001
M1 (Intercept) −2.466 1.632 0.909 0.085 −1.511 2.282 1 0.131
YrsEduc −0.055 0.079 −0.129 0.946 −0.703 0.495 1 0.482
Ratio 1.318 0.309 0.900 3.736 4.267 18.209 1 < .001
Age 0.025 0.022 0.214 1.026 1.166 1.360 1 0.243
Employed (1) −1.685 0.566 −1.685 0.185 −2.978 8.870 1 0.003
Married (1) −0.865 0.433 −0.865 0.421 −1.998 3.993 1 0.046
Note. Default level '1' coded as class 1.
* Standardized estimates represent estimates where the continuous predictors are standardized (X-standardization).
Multicollinearity Diagnostics
Tolerance VIF
YrsEduc 0.969 1.032
Ratio 0.948 1.054
Age 0.967 1.035
Employed 0.895 1.117
Married 0.991 1.009
Performance Diagnostics
Confusion matrix
Predicted
Observed 0 1 % Correct
0 142 7 95.302
1 38 13 25.490
Overall % Correct 77.500
Note. The cut-off value is set to 0.5
Performance metrics
Value
Accuracy 0.775
→ 149 cases without burnout, 51 cases with burnout (51 > 5*10, so the basic requirement is fullfilled)

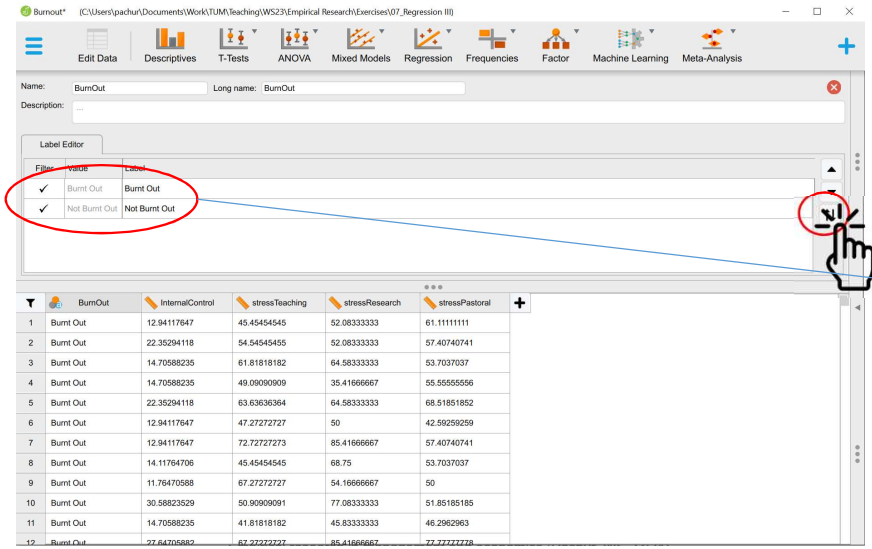
Exercise: Logistic regression II

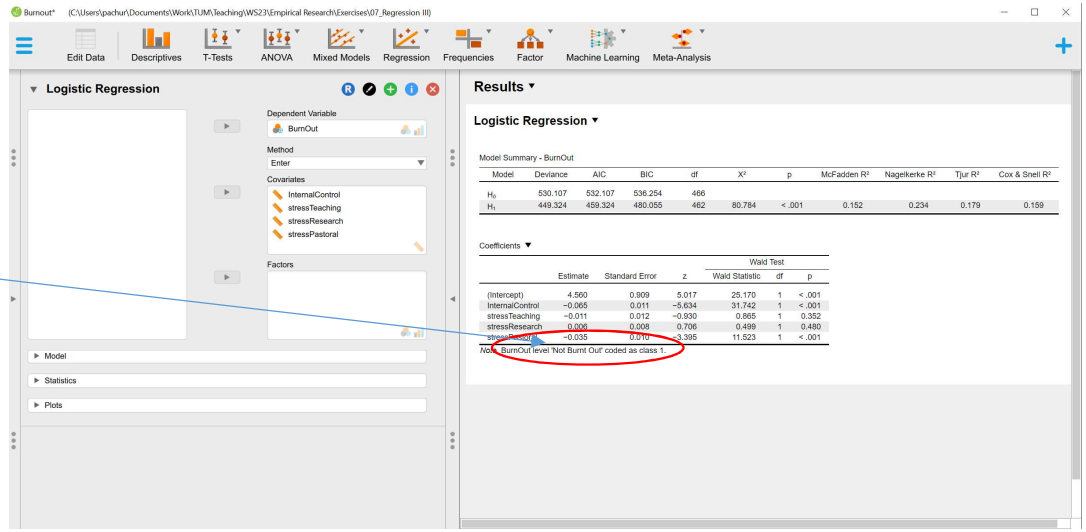
- Open the dataset “Burnout.csv“ in JASP

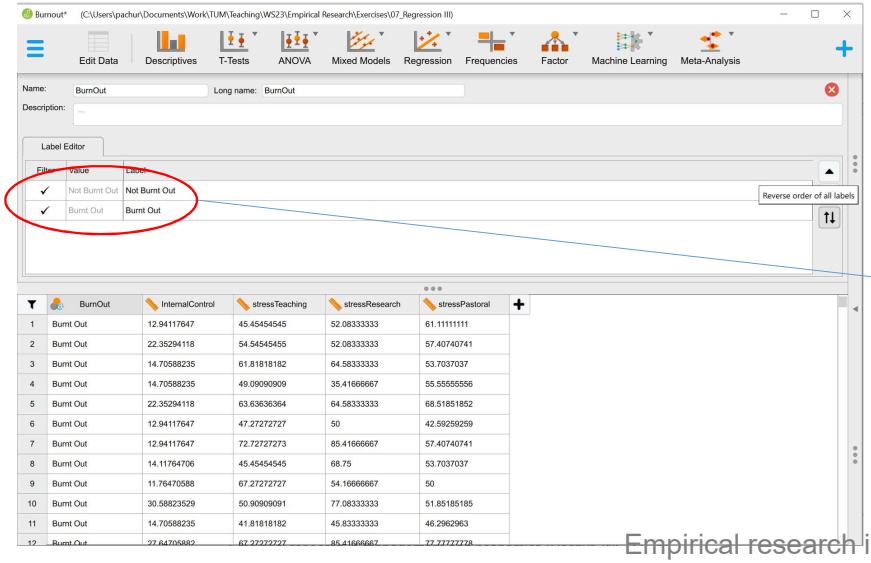
→ Burnout among academics predicted by

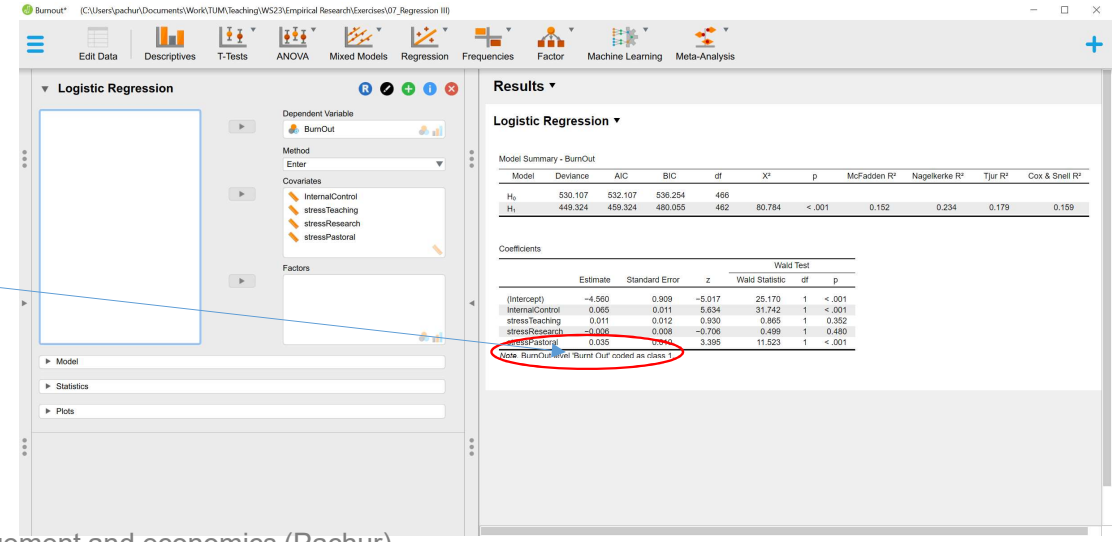
perceived internal locus of control (high value=low control!), perceived stress from teaching activities, perceived stress from research activities, perceived stress from pastoral care

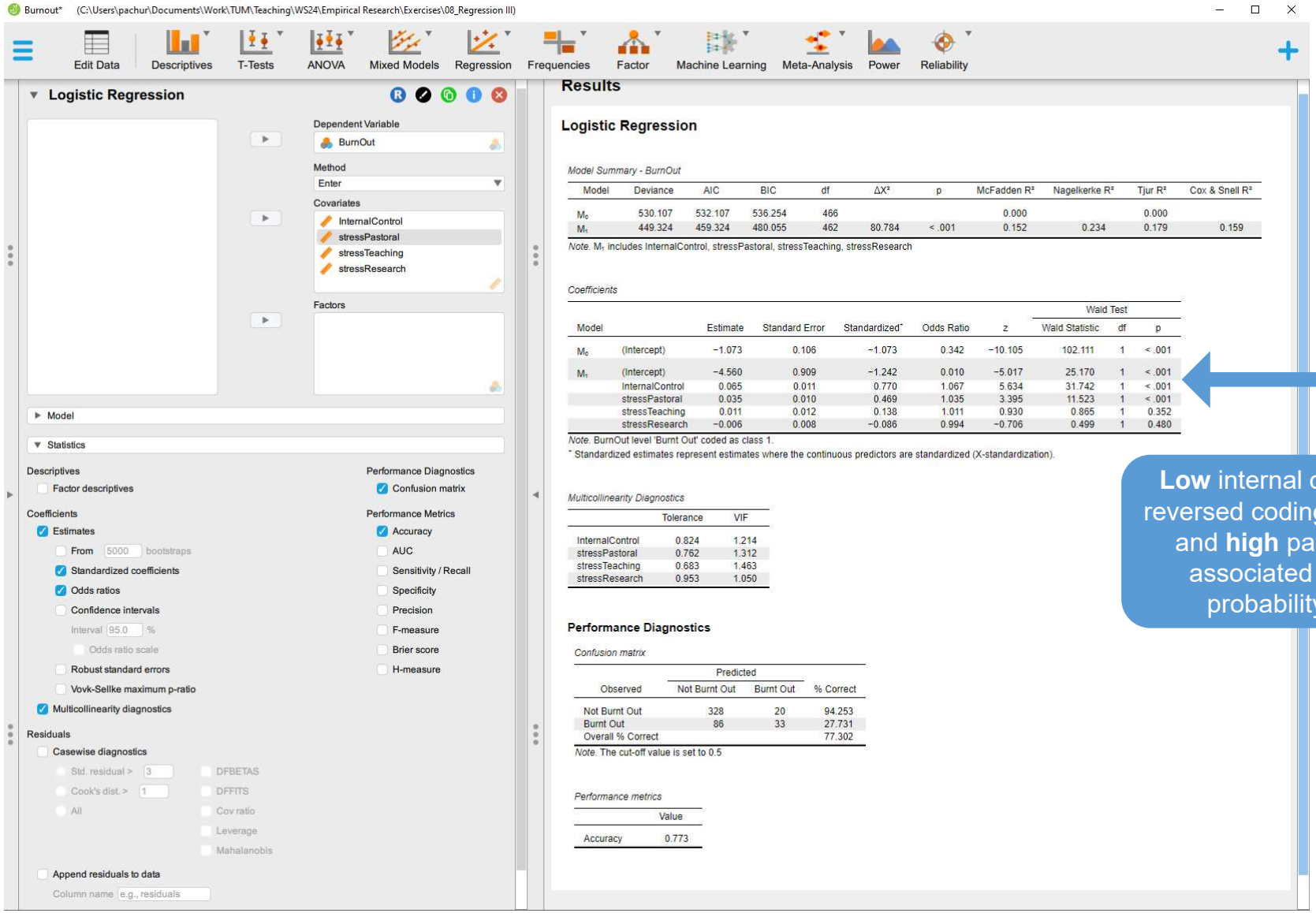
- Conduct a logistic regression analysis
  - What factors predict occurrence of burnout—and how? Also interpret the odds ratios in the JASP output.
  - Compare models with different numbers of predictors in terms of AIC and BIC to observe the trade-off between model fit and model complexity.
  - Use the best-performing model (in terms of AIC) to calculate the predicted probability of burnout for a person with scores on perceived internal control of 18, stress from teaching of 60, stress from research of 52, and stress from pastoral care of 55.


Edit Data
Descriptives
T-Tests
ANOVA
Mixed Models
Regression
Frequencies
Factor
Machine Learning
Meta-Analysis
Name: BurnOut
Long name: BurnOut
Description:
Label Editor
Burnt Out
Not Burnt Out

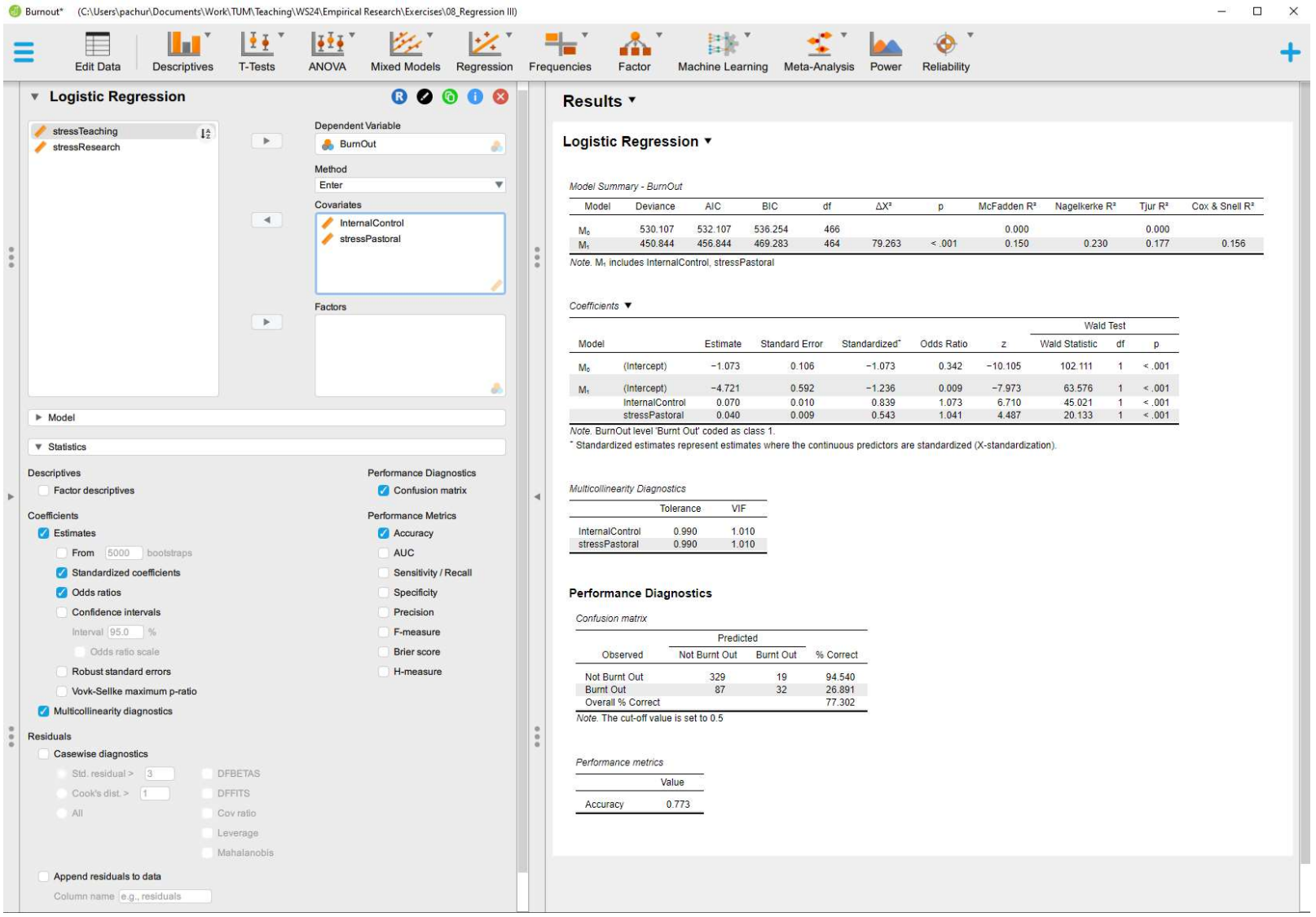

Logistic Regression
Dependent Variable
BurnOut
Method
Enter
Covariates
InternalControl
stressTeaching
stressResearch
stressPastoral
Factors
Model
Statistics
Plots
Results
Logistic Regression
Model Summary - BurnOut
Coefficients

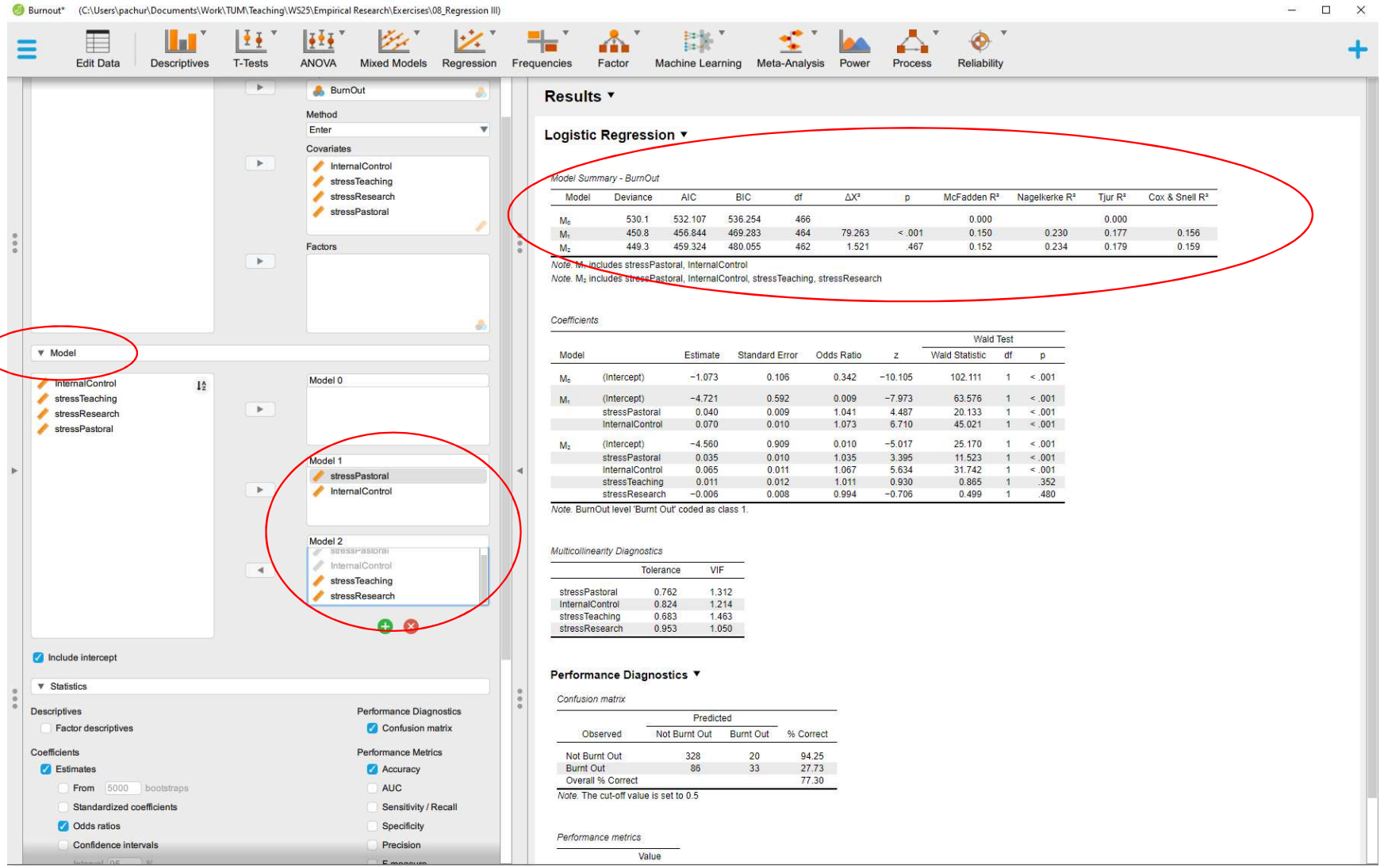

Edit Data
Descriptives
T-Tests
ANOVA
Mixed Models
Regression
Frequencies
Factor
Machine Learning
Meta-Analysis
Name: BurnOut
Long name: BurnOut
Description:
Label Editor
Not Burnt Out
Burnt Out
Reverse order of all labels


Logistic Regression
Dependent Variable
BurnOut
Method
Enter
Covariates
InternalControl
stressTeaching
stressResearch
stressPastoral
Factors
Model
Statistics
Plots
Results
Logistic Regression
Model Summary - BurnOut
Coefficients



Low internal control (note the reversed coding of the variable!) and **high** pastoral care are associated with a higher probability of burnout


Burnout* (C:\Users\pachur\Documents\Work\TUM\Teaching\WS24\Empirical Research\Exercises\08_Regression III)
Edit Data
Descriptives
T-Tests
ANOVA
Mixed Models
Regression
Frequencies
Factor
Machine Learning
Meta-Analysis
Power
Reliability
Logistic Regression
stressTeaching
stressResearch
Dependent Variable
BurnOut
Method
Enter
Covariates
InternalControl
stressPastoral
Factors
Model
Statistics
Descriptives
Factor descriptives
Coefficients
Estimates
From 5000 bootstraps
Standardized coefficients
Odds ratios
Confidence intervals
Interval 95.0 %
Odds ratio scale
Robust standard errors
Vovk-Sellke maximum p-ratio
Multicollinearity diagnostics
Residuals
Casewise diagnostics
Std. residual > 3
Cook's dist. > 1
All
DFBETAS
DFFITS
Cov ratio
Leverage
Mahalanobis
Append residuals to data
Column name e.g., residuals
Performance Diagnostics
Confusion matrix
Performance Metrics
Accuracy
AUC
Sensitivity / Recall
Specificity
Precision
F-measure
Brier score
H-measure
Results
Logistic Regression
Model Summary - BurnOut
Model Deviance AIC BIC df ΔX² p McFadden R² Nagelkerke R² Tjur R² Cox & Snell R²
M0 530.107 532.107 536.254 466 0.000 0.000
M1 450.844 456.844 469.283 464 79.263 < .001 0.150 0.230 0.177 0.156
Note. M1 includes InternalControl, stressPastoral
Coefficients
Model Estimate Standard Error Standardized* Odds Ratio z Wald Test: Wald Statistic df p
M0 (Intercept) −1.073 0.106 −1.073 0.342 −10.105 102.111 1 < .001
M1 (Intercept) −4.721 0.592 −1.236 0.009 −7.973 63.576 1 < .001
InternalControl 0.070 0.010 0.839 1.073 6.710 45.021 1 < .001
stressPastoral 0.040 0.009 0.543 1.041 4.487 20.133 1 < .001
Note. BurnOut level 'Burnt Out' coded as class 1.
* Standardized estimates represent estimates where the continuous predictors are standardized (X-standardization).
Multicollinearity Diagnostics
Tolerance VIF
InternalControl 0.990 1.010
stressPastoral 0.990 1.010
Performance Diagnostics
Confusion matrix
Observed Predicted: Not Burnt Out Predicted: Burnt Out % Correct
Not Burnt Out 329 19 94.540
Burnt Out 87 32 26.891
Overall % Correct 77.302
Note. The cut-off value is set to 0.5
Performance metrics
Value
Accuracy 0.773


Burnout* (C:\Users\pachur\Documents\Work\TUM\Teaching\WS25\Empirical Research\Exercises\08_Regression III)
Edit Data
Descriptives
T-Tests
ANOVA
Mixed Models
Regression
Frequencies
Factor
Machine Learning
Meta-Analysis
Power
Process
Reliability
BurnOut
Method
Enter
Covariates
InternalControl
stressTeaching
stressResearch
stressPastoral
Factors
Model
InternalControl
stressTeaching
stressResearch
stressPastoral
Model 0
Model 1
stressPastoral
InternalControl
Model 2
InternalControl
stressTeaching
stressResearch
Include intercept
Statistics
Descriptives
Factor descriptives
Coefficients
Estimates
From 5000 bootstraps
Standardized coefficients
Odds ratios
Confidence intervals
Performance Diagnostics
Confusion matrix
Performance Metrics
Accuracy
AUC
Sensitivity / Recall
Specificity
Precision
Results
Logistic Regression
Model Summary - BurnOut
Model Deviance AIC BIC df ΔX² p McFadden R² Nagelkerke R² Tjur R² Cox & Snell R²
M0 530.1 532.107 536.254 466 0.000 0.000
M1 450.8 456.844 469.283 464 79.263 < .001 0.150 0.230 0.177 0.156
M2 449.3 459.324 480.055 462 1.521 .467 0.152 0.234 0.179 0.159
Note. M1 includes stressPastoral, InternalControl
Note. M2 includes stressPastoral, InternalControl, stressTeaching, stressResearch
Coefficients
Model Estimate Standard Error Odds Ratio z Wald Test
Wald Statistic df p
M0 (Intercept) −1.073 0.106 0.342 −10.105 102.111 1 < .001
M1 (Intercept) −4.721 0.592 0.009 −7.973 63.576 1 < .001
stressPastoral 0.040 0.009 1.041 4.487 20.133 1 < .001
InternalControl 0.070 0.010 1.073 6.710 45.021 1 < .001
M2 (Intercept) −4.560 0.909 0.010 −5.017 25.170 1 < .001
stressPastoral 0.035 0.010 1.035 3.395 11.523 1 < .001
InternalControl 0.065 0.011 1.067 5.634 31.742 1 < .001
stressTeaching 0.011 0.012 1.011 0.930 0.865 1 .352
stressResearch −0.006 0.008 0.994 −0.706 0.499 1 .480
Note. BurnOut level 'Burnt Out' coded as class 1.
Multicollinearity Diagnostics
Tolerance VIF
stressPastoral 0.762 1.312
InternalControl 0.824 1.214
stressTeaching 0.683 1.463
stressResearch 0.953 1.050
Performance Diagnostics
Confusion matrix
Observed Predicted Not Burnt Out Burnt Out % Correct
Not Burnt Out 328 20 94.25
Burnt Out 86 33 27.73
Overall % Correct 77.30
Note. The cut-off value is set to 0.5
Performance metrics
Value

Predicted probability of burnout

Regression equation (predicted log odds)

$$\log\left(\frac{p(\widehat{burnout})}{p(no\ burnout)}\right) = -4.721 + 0.04 \times stressPastoral + 0.07 \times internalControl$$

stressPastoral = 55 internalControl = 18

Exponentiated regression equation (predicted odds)

$$\frac{p(\widehat{burnout})}{p(no\ burnout)} = e^{-4.721+\ .04\times55+0.07\times18} = e^{-1.261} = .283$$

Predicted probability from predicted odds

$$p(\widehat{burnout}) = \frac{.283}{1+.283} = .22$$

A person with a perceived internal control score of 18 and a stress from pastoral care score of 55 has a chance to experience burnout of 22%