STATISTICAID Interpretationguide

Note: The sample-PDF's purpose is to provide an impression of the interpretation guide you get. Even if the guide might sound sophisticated - it won't after my personal tutoring.

No worries: The R commands are only blackened out in this sample-PDF. Thus, all the results can be replicated with the PDF you will receive - if desired.

Introduction: The following interpretation guide was programmed with "R". R is a statistic software package free of charge. In order to replicate the results mentioned below please download R from: "https://www.r-project.org/".

Guide structure: Each sections begins with the "Start" and finishes with the "End" sign encompassing the R command, the R output and my comments (=unframed section). Texts within frames are R commands (=framed section with grey background) and their outputs (= framed section with white background) while text without frames are my comments. These comments' purpose is a better understanding having no impact on variables and/or calculations.

Start

```
[redacted]
```

Comment: This command inserts and loads your data file. **Change the directory to the location where you have stored the data on your device, otherwise data is not loaded!**

Start

```
[redacted]
```

Comment: Transform a variable to be classified as a numerical variable.

End

Start

```
[redacted]
```

Comment: This command replaces a numercial value of a variable with the one we specifiy.

End

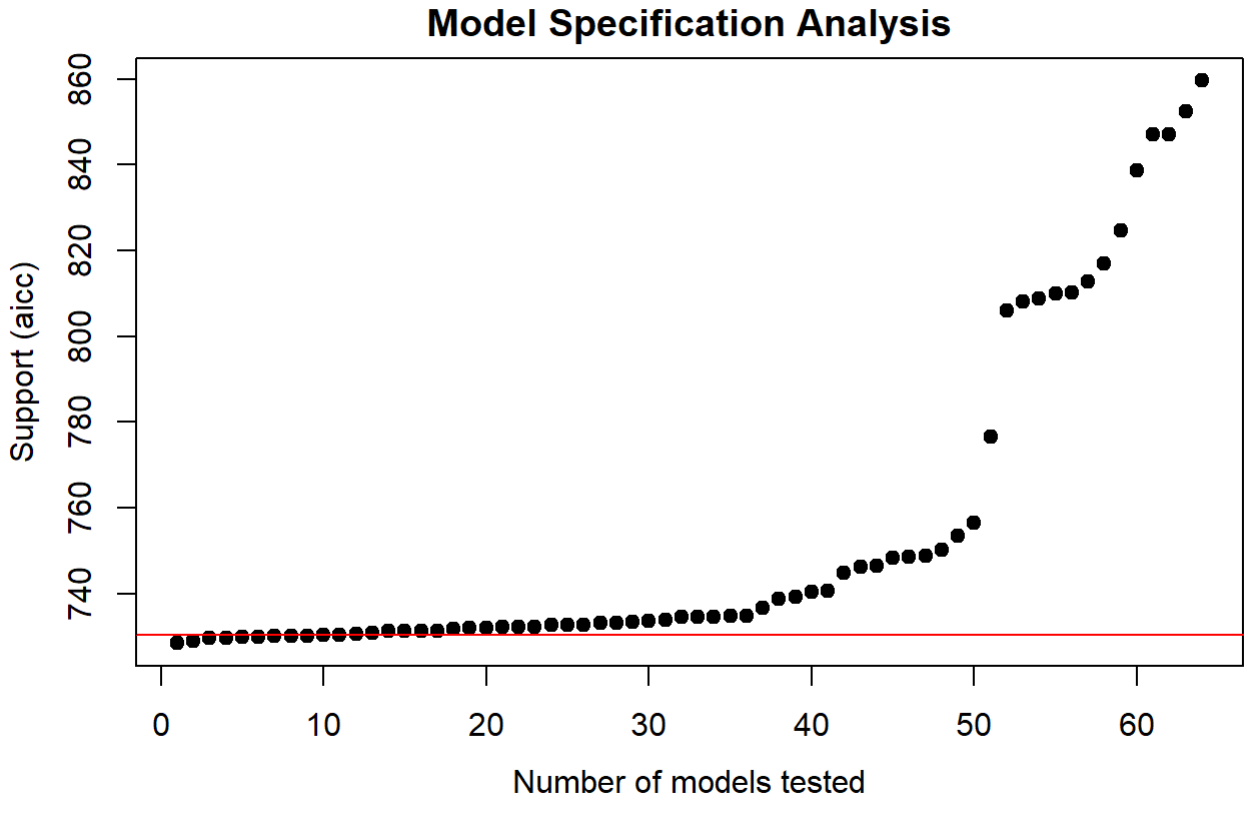
End

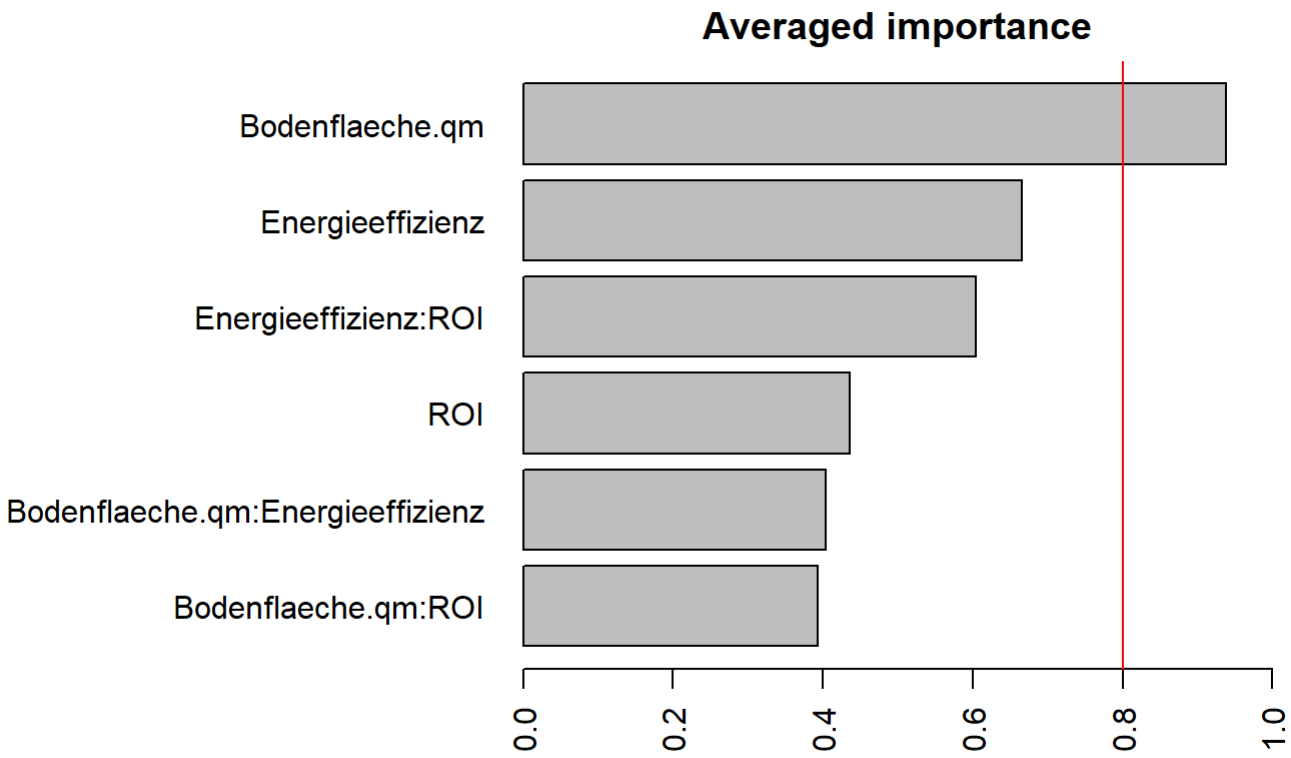
Start

[redacted]

```
## Initialization...
## TASK: Exhaustive screening of candidate set.
## Fitting...
##
## After 50 models:
## Best model: Wohnung.vs.Haus~1+Energieeffizienz+Bodenflaeche.qm+ROI:Energieeffizienz
## Crit= 728.493855159826
## Mean crit= 761.587840427385
## Completed.
```

[redacted]





```
## [1] 1
## Wohnung.vs.Haus ~ 1 + Energieeffizienz + Bodenflaeche.qm + ROI:Energieeffizienz
## <environment: 0x00000000122fb9d0>
## [1] 2
## Wohnung.vs.Haus ~ 1 + Energieeffizienz + Bodenflaeche.qm + ROI
## <environment: 0x00000000122fb9d0>
## [1] 3
## Wohnung.vs.Haus ~ 1 + Bodenflaeche.qm + ROI:Energieeffizienz +
##     ROI:Bodenflaeche.qm
## <environment: 0x00000000122fb9d0>
## [1] 4
## Wohnung.vs.Haus ~ 1 + Bodenflaeche.qm + ROI + ROI:Energieeffizienz
## <environment: 0x00000000122fb9d0>
## [1] 5
## Wohnung.vs.Haus ~ 1 + Energieeffizienz + Bodenflaeche.qm + ROI:Bodenflaeche.qm
## <environment: 0x00000000122fb9d0>
## [1] 6
## Wohnung.vs.Haus ~ 1 + Bodenflaeche.qm + Bodenflaeche.qm:Energieeffizienz +
```

```
##     ROI:Energieeffizienz
## <environment: 0x00000000122fb9d0>
## [1] 7
## Wohnung.vs.Haus ~ 1 + Energieeffizienz + Bodenflaeche.qm
## <environment: 0x00000000122fb9d0>
## [1] 8
## Wohnung.vs.Haus ~ 1 + Energieeffizienz + Bodenflaeche.qm + ROI:Energieeffizienz +
##     ROI:Bodenflaeche.qm
## <environment: 0x00000000122fb9d0>
## [1] 9
## Wohnung.vs.Haus ~ 1 + Energieeffizienz + Bodenflaeche.qm + Bodenflaeche.qm:Energieeffizi
enz +
##     ROI:Energieeffizienz
## <environment: 0x00000000122fb9d0>
## [1] 10
## Wohnung.vs.Haus ~ 1 + Energieeffizienz + Bodenflaeche.qm + ROI +
##     ROI:Energieeffizienz
## <environment: 0x00000000122fb9d0>
```

First part (glmulti command): This command provides support for identifiying a suitable model specification. Put differently, it analyses all possible combination of independent variables/their interactions for a linear multivariate regression and provides an oversight which specific variables should be and should not be included in the model. For priorising the the tested models the so called “Akaike information criterion” is used. The Akaike information criterion is a measure of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Hence, AIC provides a means for model selection. AIC is founded on information theory: it offers a relative estimate of the information lost when a given model is used to represent the process that generates the data. In doing so, it deals with the trade-off between the goodness of fit of the model and the complexity of the model. AIC does not provide a test of a model in the sense of testing a null hypothesis, so it can tell nothing about the quality of the model in an absolute sense. If all the candidate models fit poorly, AIC will not give any warning of that.

The first plot: plots the IC profile (the IC values form the best to the worst model). A horizontal line delineates models that are less than 2 IC units away from the best model. The models represented by the the points below the line are possible canditates for further analysis. The smaller the Akaike IC the better.

The second plot: plots the relative importance of model terms (variables and interactions), i.e. the overall support for each variable across all models tested. A vertical line is drawn at 80

Choosing the model: The following section provides a list of the the ten best model specifications identified to choose and test from.

[redacted]

```
## $Models
##

## Model: "glm, Wohnung.vs.Haus ~ ROI + Energieeffizienz, binomial(link = \"logit\"), myda
ta"
```

```
## Null:  "glm, Wohnung.vs.Haus ~ 1, binomial(link = \"logit\"), mydata"

##
## $Pseudo.R.squared.for.model.vs.null
##                              Pseudo.R.squared
## McFadden                           0.0637542
## Cox and Snell (ML)                 0.0772568
## Nagelkerke (Cragg and Uhler)       0.1077990
##
## $Likelihood.ratio.test
##  Df.diff LogLik.diff  Chisq    p.value
##       -2     -27.337 54.675 1.3412e-12
##
## $Number.of.observations
##
## Model: 680
## Null:  680
##
## $Messages
## [1] "Note: For models fit with REML, these statistics are based on refitting with ML"
##
## $Warnings
## [1] "None"
```

[redacted]

```
##
## Call:
## glm(formula = Wohnung.vs.Haus ~ ROI + Energieeffizienz, family = binomial(link = "logit
"),
##     data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2993  -0.9108  -0.6921   1.1894   2.2770
##
## Coefficients:
##                   Estimate  Std. Error z value Pr(>|z|)
## (Intercept)      -2.892181   0.337608  -8.567  < 2e-16 ***
## ROI               0.154233   0.049192   3.135  0.00172 **
## Energieeffizienz  0.030095   0.005613   5.361 8.27e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 857.59  on 679  degrees of freedom
## Residual deviance: 802.92  on 677  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 808.92
##
## Number of Fisher Scoring iterations: 4
```

[redacted]

```
##                  Odds_Ratio       2.5 %    97.5 %
## (Intercept)      0.05545511  0.02800748 0.1053685
## ROI              1.16676327  1.06029540 1.2861922
## Energieeffizienz 1.03055231  1.01949488 1.0422056
```

```
##              ROI  Energieeffizienz
##        0.1567606         0.2923028
```

The binomial logistic regression:

Assumptions: A binomial Logistic Regression isn't applicable for inference beyond the sample if the following assumptions are not met:

1. Binary logistic regression requires the dependent variable to be binary.
2. Secondly, since logistic regression assumes that P(Y=1) is the probability of the event occurring, it is necessary that the dependent variable is coded accordingly. That is, for a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
3. Thirdly, the model should be fitted correctly. Neither over fitting nor under fitting should occur. That is only the meaningful variables should be included, but also all meaningful variables should be included. A good approach to ensure this is to use a stepwise method to estimate the logistic regression.
4. Fourthly, the error terms need to be independent. Logistic regression requires each observation to be independent. That is that the data -points should not be from any dependent samples design, e.g., before-after measurements, or matched pairings. Also the model should have little or no multicollinearity. That is that the independent variables should be independent from each other. However, there is the option to include interaction effects of categorical variables in the analysis and the model. If multicollinearity is present centering the variables might resolve the issue, i.e. deducting the mean of each variable.If this does not lower the multicollinearity, a factor analysis with orthogonally rotated factors should be done before the logistic regression is estimated.
5. Fifthly, logistic regression assumes linearity of independent variables and log odds. Whilst it does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds. Otherwise the test under -estimates the strength of the relationship and rejects the relationship too easily, that is being not significant (not rejecting the null hypothesis) where it should be significant. A solution to this problem is the categorization of the independent variables. That is transforming metric variables to ordinal level and then including them in the model. Another approach would be to use discriminant analysis, if the assumptions of homoscedasticity, multivariate normality, and absence of multicollinearity are met.
6. Lastly, it requires quite large sample sizes. Because maximum likelihood estimates are less powerful than ordinary least squares (e.g., simple linear regression, multiple linear regression); whilst OLS needs 5 cases per independent variable in the analysis, ML needs at least 10 cases per independent variable, some statisticians recommend at least 30 cases for each parameter to be estimated.

Highly important: Probability vs. odds

Since a logistic regressen relies on odds their understanding is required to interprete the regression table. In common parlance, probability and odds are used interchangeably. However, in statistics, probability and odds are not the same. (!!!) The odds of an event happening is defined as the probability that the event occurs divided by the

probability that the event does not occur. To continue with our coin-tossing example, the probability of getting heads is 0.5 and the probability of not getting heads (i.e., getting tails) is also 0.5. Hence, the odds are .5/.5 = 1.Note that the probability of an event happening and its compliment, the probability of the event not happening, must sum to 1. Now let's pretend that we alter the coin so that the probability of getting heads is .6. The probability of not getting heads is then .4. The odds of getting heads is .6/.4 = 1.5. If we had altered the coin so that the probability of getting heads was .8, then the odds of getting heads would have been .8/.2 = 4. As you can see, when the odds equal one, the probability of the event happening is equal to the probability of the event not happening. When the odds are greater than one, the probability of the event happening is higher than the probability of the event not happening, and when the odds are less than one, the probability of the event happening is less than the probability of the event not happening. Also note that odds can be converted back into a probability: probability = odds / (1+odds). Now let's consider an odds ratio. As the name suggests, it is the ratio of two odds. Let's say we have males and females who want to join a team. Let's say that 75% of the women and 60% of men make the team. So the odds for women are .75/.25 = 3, and for men the odds are .6/.4 = 1.5. The odds ratio would be 3/1.5 = 2, meaning that the odds are 2 to 1 that a woman will make the team compared to men(!!!).

In summary: The number of times the event occurs divided by the number of times the event could occur (possible values range from 0 to 1). **odds:** the probability that an event will occur divided by the probability that the event will not occur: probability(success) / probability(failure). **Odds ratio:** the ratio of the odds of success for one group divided by the odds of success for the other group: (probability(success)A/probability(failure)A) / (probability(success)B/probability(failure)B). **log odds:** the natural log of the odds

Interpretation of the regression output:

Pseudo.R.squared:

In the linear regression model, the coefficient of determination, R^2, summarizes the proportion of variance in the dependent variable associated with the predictor (independent) variables, with larger R^2 values indicating that more of the variation is explained by the model, to a maximum of 1. For logistic regression models with a categorical dependent variable, it is not possible to compute a single R^2 statistic that has all of the characteristics of R^2. So these Pseudo-R^2-approximations are computed instead. The following methods are used to estimate the pseudo-coefficient of determination.

McFadden's R^2 is based on the log-likelihood kernels for the intercept-only model and the full estimated model. Usually, but not always McFadden's is referred to.

Cox and Snell's R^2 is based on the log likelihood for the model compared to the log likelihood for a baseline model. However, with categorical outcomes, it has a theoretical maximum value of less than 1, even for a "perfect" model.

Nagelkerke's R^2 is an adjusted version of the Cox & Snell R-square that adjusts the scale of the statistic to cover the full range from 0 to 1.
What constitutes a "good" R^2 value varies between different areas of application. While these statistics can be suggestive on their own, they are most useful when comparing competing models for the same data. The model with the largest R^2 statistic is "best" according to this measure.

Likelihood.ratio.test

In statistics, a likelihood ratio test is a statistical test used for comparing the goodness of fit of two models, one of which (the null model) is a special case of the other (the alternative model). The test is based on the likelihood ratio's logarithm, which expresses how many times more likely the data are under one model than the other. Simply speaking, if the p.value of the (Chisq) likelihood ratio test is statistically significant - i.e. smaller than 0.05 - it means it is better to have this model than none. So if the model is significant, a further interpretation of the regression is meaningful.

Call:

In the section, the first thing we see is the call, this is R reminding us what the model we ran was, what options we specified, etc.

Deviance Residuals:

Next we see the deviance residuals, which are a measure of model fit. This part of output shows the distribution of the deviance residuals for individual cases used in the model. Residuals are essentially the difference between the actual observed response values and the response values that the model predicted. The R residuals section of the model output breaks it down into 5 summary points.

Coefficients

Estimate - The coefficient estimate contains two rows; the first one is the intercept (= the constant) and second one are the coefficients. The intercept can be interpreteted as the average of log odds accounted for the effects of the independent variables. Often, it does not have a special meaning. The slope term in the model is represented by the coefficients only. These are the values for the logistic regression equation for predicting the dependent variable from the independent variables.
Note: In case of statistically significance you can interprete the coefficients as follows: If X (=independent variable) increases by one unit, the log-odds of Y (= dependent variable) increases by k unit (= coefficient), given the other variables in the model are held constant.
Note: A comparison between coefficients (such as which coefficient has bigger/smaller effect on the dependent variable) is not allowed. For doing so, see paragraph “Standardized coefficients”.
Note: If a coefficient’s z-statistic is not significant, don’t interpret it at all! (For significance see columns Pr(>|z|)). Remember: If a non statistically coefficient is at hand (i.e. the coefficient’s P(>|z|) value is above 0.05) you should state so. But do not speculate about any relationship. You could say something like the following: “No statistically significant dependence of Y on x was detected.” Nevertheless, these coefficients can be interpreted more easily and intuitively if the are visualized as odd ratios. And thus, odd ratios should be preferred. (See paragraph “Odds ratio”)

Std. Error - The coefficient standard error measures the average amount that the coefficient estimates vary from the actual average value of our response variable. We’d ideally want a lower number relative to its coefficients. It is also to be used to compute confidence intervals.

z value - The coefficient z-value is the test value. We want it to be far away from zero as this would indicate we could reject the null hypothesis - that is, we could declare that a relationship between dependent and independent variable exist.

Pr(>|z|)- The Pr(>|z|) acronym found in the model output relates to the probability of observing any value equal or larger than |z|. A small p-value indicates that it is unlikely we will observe a relationship between the predictor and response variables due to chance. Typically, a p-value of 5% or less is a good cut-off point (= significance level). In other words: If the p-value of a coefficient / constant yields a value below 0.05 the coefficient is statistically significantly different from 0. If it yields a p-value above 0,05 the coefficient / constant is not statistically significant.
Note the ‘signif. Codes’ associated to each estimate. Three stars (or asterisks) represent a highly significant p-value. Consequently, a small p-value for the intercept and the slope indicates that we can reject the null hypothesis which allows us to conclude that there is a relationship between the dependent and independent variable.

Odds Ratio:

Odds Ratios - These are the original coefficients of the logistic regression (see “Coefficients”) displayed in odds ratios (!!!) in order to make an interpretation easier. Note: You can interprete the odd ratios as follows: For one-unit increase/decrease in the respective independent variable an k (= coefficient) an odds increase/decrease in the dependent variable is predicted that a certain event occurs, holding all other variables constant. Put differently, if the odds ratio is greater than 1, odds are rising that the event coded 1 within the dependent variable compared to the odds that the event is not occurring (= coded 0 within the dependent variable). If the odds ratio is smaller than 1, odds are decreasing that the event coded 1 within the dependent variable compared to the odds that the event is not occurring (= coded 0 within the dependent variable). It is easier to understand in case of a dichothomous independent variable => For instance: The dependent variable represents admission to a university (= coded 1 = event) and non admission (= coded 0). If the independent variable is gender (coded 1 for male and 0 for female) with a odds ratio of 5.00 at hand, this means that odds for admission of male applicants is five times that of females.
Note: A comparison between the odd ratios (such as which odd ratio has bigger/ smaller effect on the dependent variable) is not allowed. For doing so, see under command “Standardized Coefficients”.
Note: If the odds ratios original coefficient’s z-statistic is not significant, don’t interpret it at all (see *Pr(>|z|)*.

Confidence intervals - this part shows the lower and upper confidence intervals of the individual odds ratio. We are

95% confident that the true odds ratio in the model which generated this data falls within those values.

Standardized coefficients

The standardized coefficients are the coefficicents standardized to standard deviations. Because the standardized coefficients are all measured in standard deviations, instead of the units of the variables, they can be compared to one another. Put differently, only the standardized coefficients allow you to determine which independent variable can be assumed to have the biggest/smallest effect on the dependent variabel und thus, enables a comparison between the coefficients calculated.

Note: The scale of measurement are standard deviations leading to the following conclusion: For instance, a standardized coefficient value of 2.5 explains one standard deviation increase in independent variable on average - i.e. a 2.5 standard deviation increase in the log odds of dependent variable.

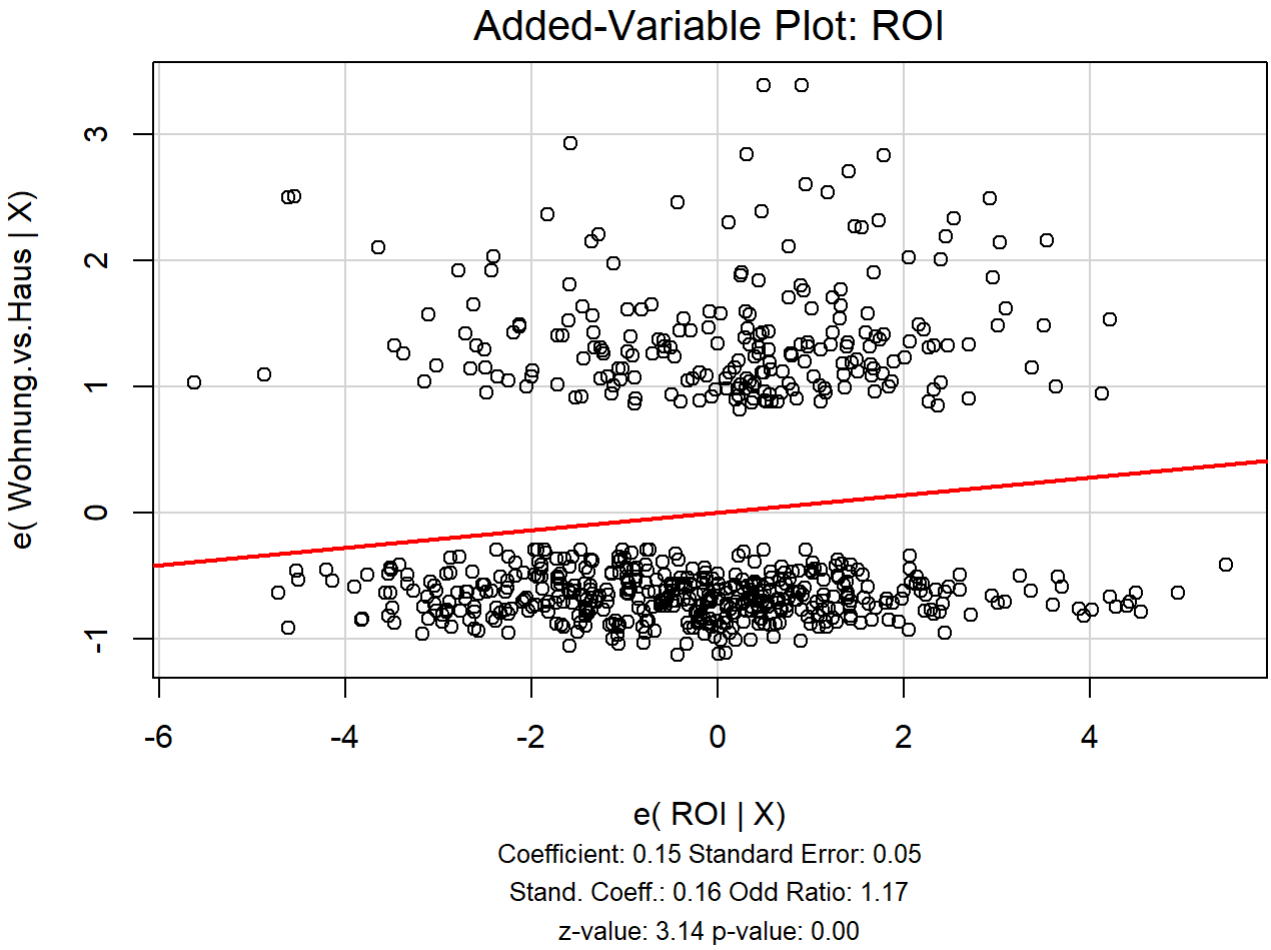
Note: The calculation of the standardized coefficients in logistic regression is not as straight forward as in OLS Regression and differs among statistical packages. Thus, be cautious to interpret the results.

Note: If the odds ratios original coefficient's z-statistic is not significant, don't interpret it at all (see *Pr(>|z|)*.

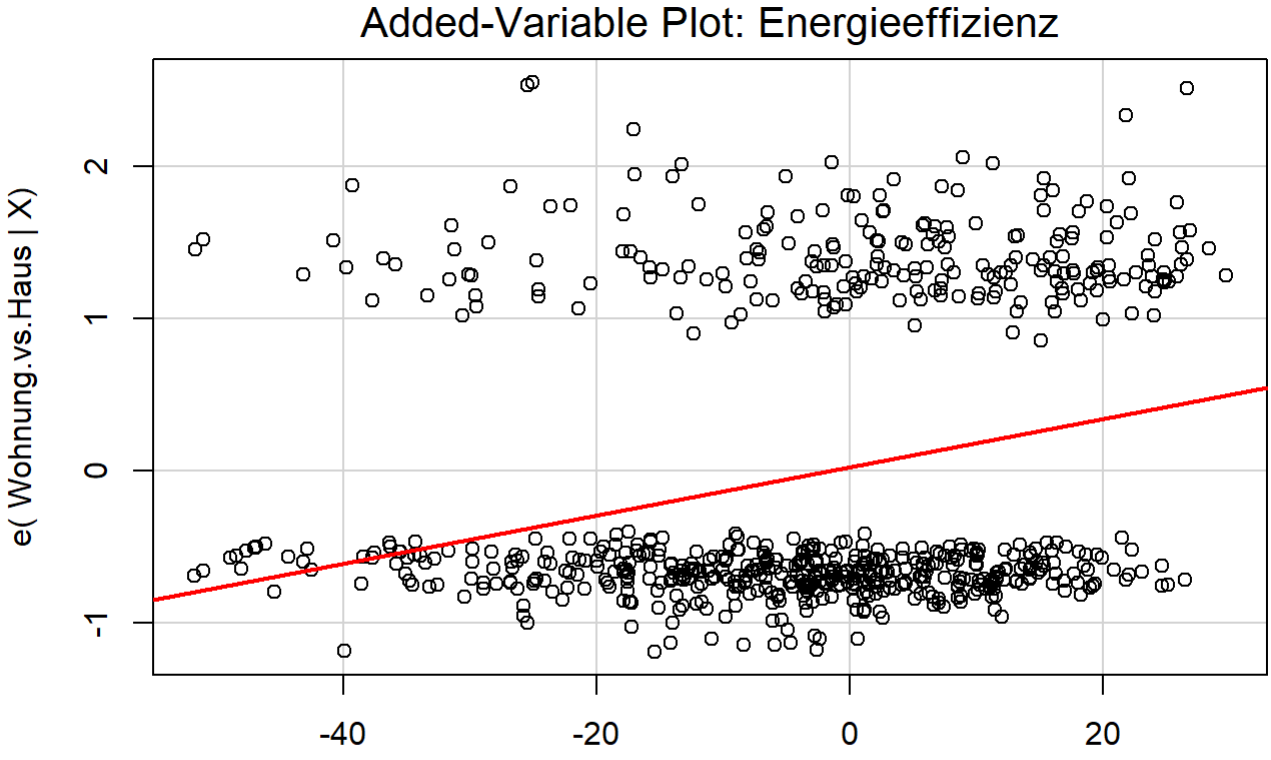
[redacted]

```
## [1] "ROI"              "Energieeffizienz"
```

[redacted]



Coefficient: 0.15 Standard Error: 0.05
Stand. Coeff.: 0.16 Odd Ratio: 1.17
z-value: 3.14 p-value: 0.00



Coefficient: 0.03 Standard Error: 0.01
Stand. Coeff.: 0.29 Odd Ratio: 1.03
z-value: 5.36 p-value: 0.00

[redacted]

Comment: The added variable plot basically plots the residuals from the regression of the response on a subset of the regressors versus the residuals from the regression of the new regressor on the same subset of regressors. Simply speaking, it shows the influence of a single independent variable onto the dependent variable by correcting for the influences of the other independent variables. That's why it is also called "partial residual plot".

etc….