A study about stress in a demanding west-European society

Introduction

Our west-European society is becoming increasingly demanding in its expectations towards its citizens. More than ever, both men and women are expected to build a successful career. Aside from this, they are expected to have an active part in their family life, in sharp contrast with the classical pattern of the earning father and the stay-at-home housewife, which was predominant in the past decennia. Furthermore, they are expected to invest in at least a single form of self-development, be it through sports, hobbies, an active role in socio-cultural organizations, and so forth. Likewise, they must encourage their children to do the same. We conducted an exploratory observational study to explore the impact of career choices on one's stress level.

Methods

The data for our study were collected through a questionnaire (see protocol) that was randomly sent out via e-mail and social media to family, friends, acquaintances and colleagues of the group members. The survey consists of questions related to work, career, partner, children and stress and was anonymously filled in by the subjects. A total of 158 responses were collected. After reducing the amount of variables, the final dataset consisted of 20 explanatory variables listed below. After the data collection the data were screened for errors as well as extreme outliers (Figure 1, Appendix A). Subjects with partner and/or children that did not fill in all the relevant questions regarding partner and/or children were removed (n=13). Obvious outliers were removed (n=3), these consisted of answers of a salary of 1 Euro and personal quality time of 180 hours a week. After data preparation the final sample size consisted of 140 observations.

Variable name	**Unit/scale**
Age	Years
CareerYears	Years
ChildcareComparison	Likert scale 0-10
CommuteOptions	By foot/bike/public transport/car (inclusive carpooling) or motor bike
CommutingTime	Minutes per day
EmployedToDegree	Likert scale 0-10
FamilyQualityTime	Hours per week
Gender	Male/female
HasChildren	Yes/no
HasPartner	Yes/no
HighestDegree	Primary school/secondary school/Bachelor/Master/PhD
IsPartnerOvertime	Yes/no
Overtime	Hours per day
PartOrFulltime	Fulltime/parttime
PartnerQualityTime	Hours per week
PersonalQualityTime	Hours per week
Salary	Monthly net in euros
SectorChangeCount	Absolute count
StressDegree	Likert scale 0-25

The variables in the dataset were explored to get a better understanding of their distribution and bivariate relationships. Histograms, boxplots, qqplots and scatter plot matrices were used as visual tools for the descriptive statistics and to check the presence of gaps and outliers in the data. A correlation matrix was created to check the relationship between all the variables in the study. A scatter plot was created to explore the relationship between the response variable StressDegree and the predictor variable of interest Salary. To assess whether a linear relationship between the two variables was reasonable, the best fitting curve and lowess smoother were added to the plot. A simple linear regression was then performed of the following form:

$$Y_i = \beta_0 + \beta_1 .\ X_{i1} + \varepsilon_i$$

where:

Y_i = StressDegree, X_{i1} = Salary, where $\varepsilon_i \sim N(0,\sigma^2)$ (i = 1,...,140). The regression coefficients were estimated, a *t* test was conducted to test if the slope is different from zero at the α 5% level of significance. Large values of the *t* test statistic lead to rejection of the null hypothesis in favour of the alternative hypothesis: H_0: $\beta_1 = 0$ versus Ha: $\beta_1 \neq 0$. Diagnostics procedures were performed using different graphical tools, mostly based on residual plots and formal tests to identify lack of fit, outliers, influential observations, multicollinearity etc. and the assumptions underlying the regression model were evaluated. Remedial measures were evaluated, such as a Box-Cox analysis to find out which transformations could possibly improve the linearity of the regression function.
Aanvullen

Missing data were generated in our survey for questions related to partner and children for people without partner and/or children (variables ChildcareComparison, FamilyQualityTime, IsPartnerOvertime, PartnerQualityTime and PersonalQualityTime). To deal with the missing values missingness indicators were introduced as interaction terms between those variables and the binary variables for partner and children (HasPartner and HasChildren), that take on the values 0 and 1. The model then becomes of the following form:

$$Y_i = \beta_0 + \beta_1 .X_{i1} + \beta_2 .X_{i2} + \beta_3 .X_{i2} .X_{i3} + \dots \beta_k X_k + \varepsilon_i$$

where:

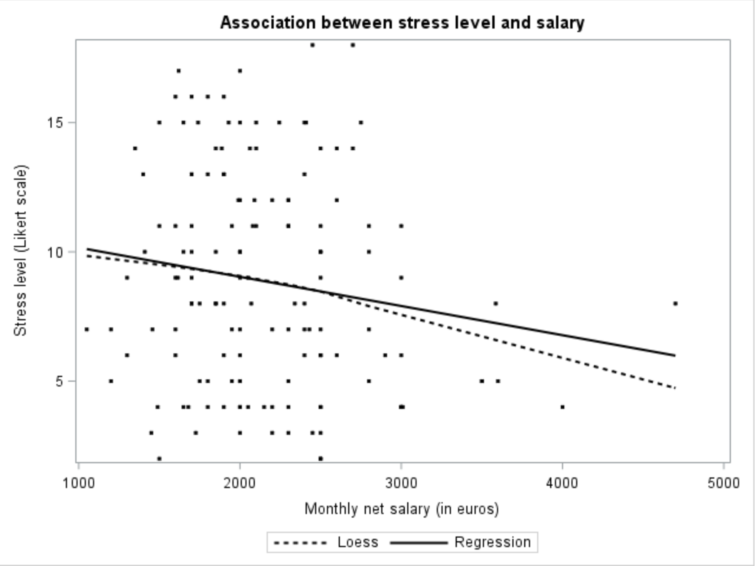
Y_i = StressDegree, X_{i1} = Salary, X_{i2} = HasPartner or HasChildren, X_{i3} = variables related to partner or children (ChildcareComparison, FamilyQualityTime, IsPartnerOvertime, PartnerQualityTime, PersonalQualityTime) where $\varepsilon_i \sim N(0,\sigma^2)$ (i = 1,...,140) and k the number of covariates in the dataset. For people without partner and/or children (X_{i2} =0) the terms $\beta_2 .X_{i2} + \beta_3 .X_{i2} .X_{i3}$ disappear from the model.

The model building started with a first-order regression model fitted with all the predictor variables as a starting point. Partial residual plots were evaluated to detect possible residual trends, interactions etc. Then the forward selection method was used to decide which parameters should be added to the model and makes use of the *F* test. The *F* test tests whether a term $\beta_k X_k$ needs to be included in the model or not at the α 5% level of significance. Large values of the *F* test statistic, that reflect the variable's contribution to the model if it is included, lead to rejection of the null hypothesis in favour of the alternative hypothesis: H_0: $\beta_k = 0$ versus H_a: $\beta_k \neq 0$. The forward selection started with only the variable of interest Salary in the model. For each of the independent variables, the *F* statistics and *p*-values were calculated and compared to the level of significance. Only the variables yield significant results were added to the model. The quality of the model, assumptions and remedial measures were evaluated like was done for the simple linear regression model. During every model building step the model validation criteria were evaluated.
The SAS software (version 9.4) was used for the data analysis.

Results & Discussion

Descriptive analysis Julien

The relationship between the response variable and the predictor variable was explored by creating a scatter plot (Figure X below). From the plot the nature of the regression relationship is not clearly visible. The relationship between the two variables seems rather random. The best fitting curve and smoother were plotted together. Both support a negative linear relationship between the variables. However, the relationship seems rather weak at first sight.



A simple linear regression was fitted next. The estimates for the regression coefficients were obtained from the Linear regression table X. For the intercept an estimate of 11.3 was found, with a standard error of 1.4 and a 95% confidence interval from 8.6 to 14.0. For the slope β_1 an estimate of -0.0011 was found, with a standard error of 0.00062 and a 95% confidence interval from -0,0024 to -0,0000095. The estimated regression function therefore is:

(1) StressDegree = 11.3 - 0.0011*Salary

An increase by 1000 Euros in the month net salary of an employee is assiociated with an expected decrease in the mean stress degree by 1.1 on the Likert scale. However, the null hypothesis of the slope equal to zero cannot be rejected at the α 5% level of significance. A *t* value of -1.82 and a corresponding *p*-value of 0.0704 was found. There is not enough evidence of a linear relationship between stress level and salary. Therefore we cannot use the interpretation above and can only interpet the model as the mean stress level is around 11 on the Likert scale, irrespective of the levels of salary. Also the coefficient of determination, which gives some information about the goodness of fit of the linear model for our data, is very low (R^2 =0.0235). The regression line does not fit the data well. The reason for this could be because the response variable and the explanatory variable tend to be uncorrelated, as seen in the previous figure. The Pearson correlation coefficient between the variables is only -0.15 (Correlation matrix, Appendix A). Since salary is our variable of interest, we keep it in the model anyway.

The appropriateness of the regression model and the assumptions underlying the model were then evaluated. First, it was checked whether there were any outlying values for the predictor variables that could influence the appropriateness of the fitted regression function. This was evaluated by creating a boxplot for the predictor variable of interest salary (Figure X, Appendix A). As described earlier, salary seems to be slightly right skewed distributed, since the median is lower than the mean and not in the middle of the box. Two potential outliers are visible, showed as the points above the top whisker. These need to be explored formally to see if they have an influence on the fitted model (see further). The plot of the residuals of the fitted model versus the predictor variable (figure X) shows that the points are randomly dispersed around the horizontal axis. There is no residual trend visible so we have no direct clue to modify the model to another type of regression function. The variance of the error terms seems constant in the first half of the residual plot. However, there is a decrease for values of salary exceeding 3000. These values could also indicate potential outliers as we saw earlier from the boxplot. This is confirmed by the squared residual plot (Figure X, Appendix A). The Brown-Forsythe test was performed, to see if the errors are similar in 2 equally sized ranges of X-values. The minimum salary is 1050, whilst the maximum is 4700. The splitting point was calculated as follows: 1050 + (4700-1050)/2 = 2875. The first segment of the X-values contained 126 data points, while the second segment only contained 14 data points. The t-test for equal variance yielded t*_bf= 3.14 (p= 0.0021). The critical value for this test is qt(α=0.975, df=140-2) = 1.977304. Because |t*_BF| > 1.98, we conclude that both segments have a different variance and are therefore heteroscedastic. The Breusch-Pagan test was performed next. This test yielded, χ2 $_{\text{BP}}$=0.53383 (P=0.4650). The threshold for this χ2 distribution is qchisq(0.95, 1) = 3.84. Since 0.534 < 3.84, this test confirms that both variances are equal. Since there are more observations in the first segment than the second one, the Brown-Forsythe test appears to be less trustworthy to determine the nonconstancy of error variance, so we opted for the Bruesch-Pagan test result as a more reliable outcome for our data.

Presence of outliers

To check whether there were outlying cases that could have an influence on the fitted regression function, diagnostic procedures were undertaken to identify potential cases, evaluate their influence and consider remedial measurements. Identification of cases with outlying *Y* observations was done by looking at the studentized residuals and studentized deleted residuals. For both the studentized and studentized deleted residuals no absolute values exceeding 2.5 were found (Figures X and X, Appendix A). Identification of outlying *X* observations was done by looking at Cook's distances and the leverage values. For a sample size of 140 observations, a Cook's distance of ≥ 0.029 is considered a potential outlier. Two observations showed a high Cook's distance (0.029 and 0.039) in our dataset (Figure X, Appendix A). Leverages with values higher than 0.5 are considered large and moderate between 0.5 and 0.2. None of the observations showed leverages higher than 0.15 (Figure X, Appendix A) so we do not expect the outliers to cause serious problems.

Nonnormality of error terms

The normality of the error terms was evaluated by looking at the histogram, boxplot and normal quantile plot of the studentized residuals (Figure X, Appendix A). The distribution does not seem perfectly normal, the points do not follow a perfect linear line on the normal quantile plot, but the deviation from normality does not seem to be too dramatic. Our dataset is large enough so we do not expect this to cause serious problems.

Independency of the error terms

Since there is not a particular time-sensitive aspect to fill in a survey, we assume that the errors are independent. The residuals were plotted against the sequence in which the questionnaire was filled in by the respondents (Figure X). There is no clear indication of a trend related to time or sequence.

The Forward selection procedure started with only the variable of interest Salary in the model. One by one the other covariates were added and tested if they were significant to enter the model. Only HasChildren and ChildcareComparison*HasChildren were significant at this first stage. *T* values of 2.26 (*p*=0.0252) and 2.20 (*p*=0.0292) were found respectively. Also Salary now becomes significant in the model, with a *t* value of -2.20 (*p*=0.0295). HasPartner was initially included in the model selection with covariates related to partner. However the covariate did not yield a significant result to be included in the model. Removing HasPartner barely affected the estimates of the other regression parameters. The resulting model with estimates for the regression coefficients then becomes:

(2) StressDegree = 11.7 -0.0014*Salary + 4.1*HasChildren
-0.67*(ChildcareComparison*HasChildren)

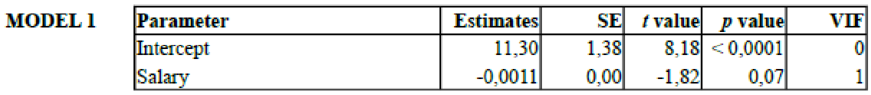
The interpretation of the second model would be an increase of 1000 Euros in the net monthly salary of an employee is associated with an expected decrease in the mean stress level by 1.4 on the Likert scale, while holding the fact of having a child or not and the care of the children constant. When a person has children, the level of stress is associated with an increase of 4.1 on the Likert scale, when monthly net salary and childcare are held constant. Likewise, when a person has children, an increase in the care of the children on the Likert scale by 1 is associated with an expected decrease of the stress level by -0.67 on the Likert scale, when salary is held constant. Then the remaining covariates were tested again to enter this new model. Only Gender gave a significant result to be included, with a *t* value -2.15 (*p*=0.0335). The final model with estimates for the regression coefficients became:

(3) StressDegree = 12 -0.0012*Salary + 5.6*HasChildren
-0.92*(ChildcareComparison*HasChildren) - 1.6*Gender

The interpretation of the third model would be an increase by 1000 euros in the net monthly salary of an employee is associated with an expected decrease in the mean stress level by 1.2 on the Likert scale, while holding the fact of having a child or not and the care of the children constant and the gender of the employee constant. When a person has children, the level of stress is associated with an increase of 5.6 on the Likert scale, when monthly net salary, childcare and gender of the employee are held constant. For male employees, the stress level is expected to decrease by 1.6 on the Likert scale, when salary and childcare are held constant.

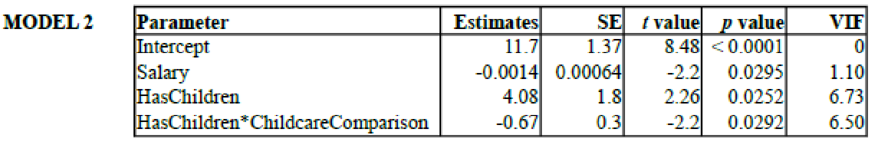
Next, we checked if interaction terms should be added to the model, if there is confounding and multicollinearity. No interaction terms produced significant results to enter the model. During the forward selection process, the parameters did not change dramatically when adding new covariates. Therefore, we assume that there is no confounding to deal with. The variance inflation factors calculated for every estimate in our models indicated no problems concerning multicollinearity. From the partial residuals plots from the built models there is no clear indication that the predictors have a nonlinear relationship with the response variable. An overview of the parameters, their significance and variance inflation factors for all the model build is given in Table X below. Plots of the variable selection criteria can be found in Figure X, Appendix A. Both the R2 and adjusted R2 increase with the number of added parameters in the model. The AIC, PRESS and SBC are more optimal with every added variable (Table X, Appendix A).

Table X Overview of parameter estimates and their significance per model



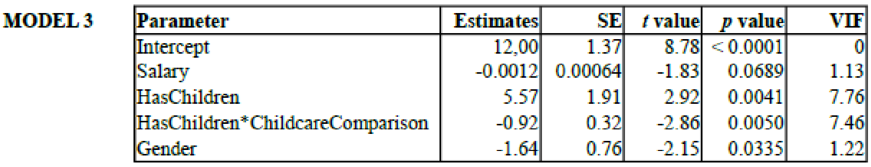
MODEL 1

Parameter	Estimates	SE	*t* value	*p* value	VIF
Intercept	11,30	1,38	8,18	<0,0001	0
Salary	-0,0011	0,00	-1,82	0,07	1



MODEL 2

Parameter	Estimates	SE	*t* value	*p* value	VIF
Intercept	11.7	1.37	8.48	<0.0001	0
Salary	-0.0014	0.00064	-2.2	0.0295	1.10
HasChildren	4.08	1.8	2.26	0.0252	6.73
HasChildren*ChildcareComparison	-0.67	0.3	-2.2	0.0292	6.50



MODEL 3

Parameter	Estimates	SE	*t* value	*p* value	VIF
Intercept	12,00	1.37	8.78	<0.0001	0
Salary	-0.0012	0.00064	-1.83	0.0689	1.13
HasChildren	5.57	1.91	2.92	0.0041	7.76
HasChildren*ChildcareComparison	-0.92	0.32	-2.86	0.0050	7.46
Gender	-1.64	0.76	-2.15	0.0335	1.22

Conclusion

Most explanatory variables seem to be unrelated to the response variable. Three models were built based with the covariates for salary, children, childcare and gender. We don't think the models are perfect to make predictions about stress level, since this is a variable that is influenced by a lot of other factors. In our study we only looked at career choices and the family situation. These are situations were stress may play a great deal, but other factors should also be considered such as health status, emotional problems etc. Our third model that included gender lead to non-significance of our variable of interest Salary. AANVULLEN

Prologue

If a marketeer wants to target an audience with a certain level of stress he or she could make use of our second and final model, depending of the purpose of his study. If the gender of the study population is important, the marketeer could make use of our final model, where this variable is included. However, salary is not significant in this model, so if gender is not relevant for the goal of the study, he should make use of our second model, where salary is significant. However, depending on the nature of his study our model could not be good to make predictions, since the model is only built for aspects of stress experience related to work and family situation. Then the marketeer should have to consider other covariates to build a useful regression model. It is important that a 'good' set of explanatory variables is chosen and that the number is small enough so that the costs are manageable and the analysis is facilitated, but yet large enough so that adequate description, control or prediction is possible for his study.

The goals of this study, as mentioned in the protocol, was to gain insight into the way people deal with stress in our modern, West-European society. As the study didn’t really uncover all that much relevant and significant insights into the matter, we can conclude that this data analysis has not met its objectives.

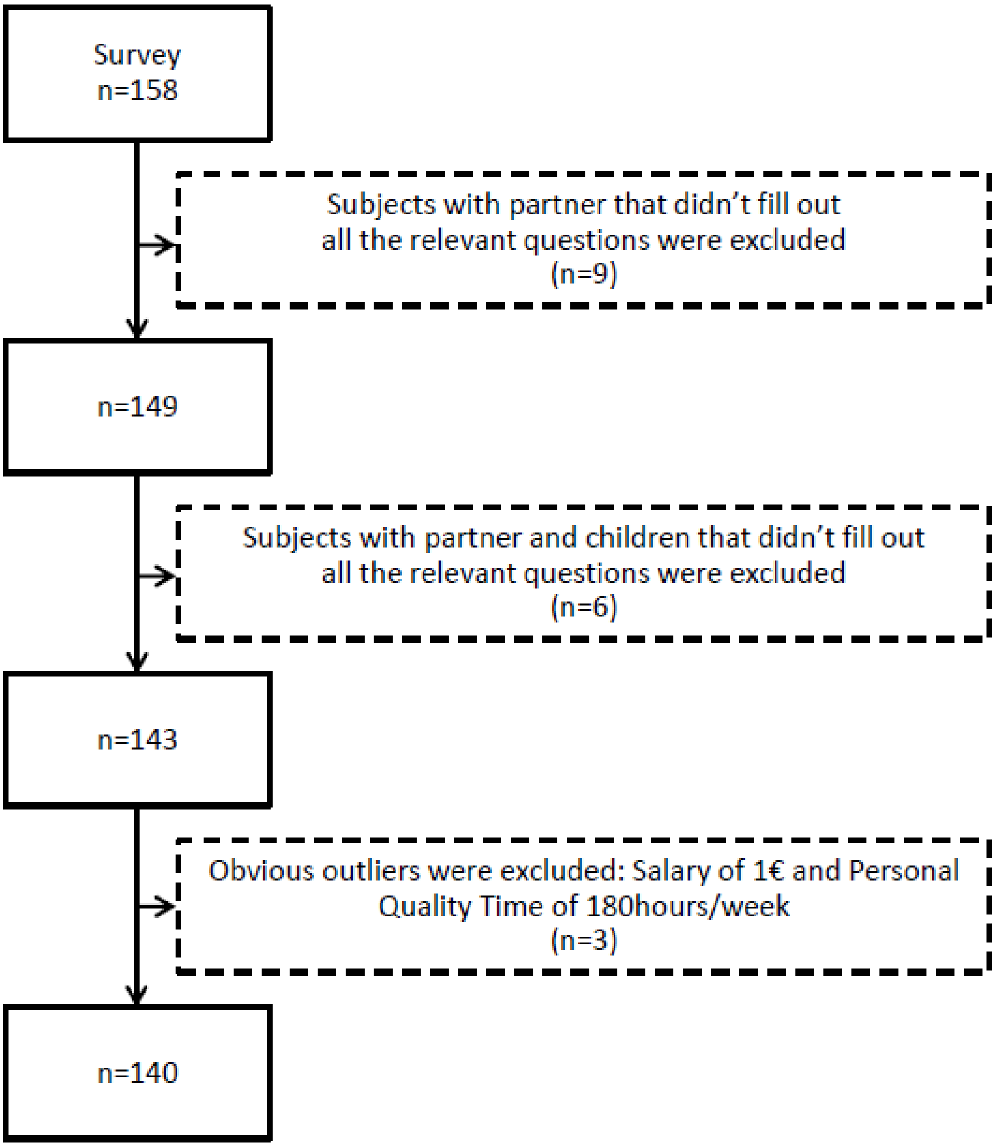
What we would do differently, in a follow-up study:

- **About our sample (diversity):** We would aim at obtaining a more diverse sample of the population. As can be seen, most people in our sample work at companies that are open to part time working regimes, which is perhaps not that common in that many companies nowadays. We can also see that the majority of our respondents work in services, and almost none work in construction or trade. Most of our subjects also have a masters’ degree while none have only obtained their secondary school degree.
- **About our data gathering methodology:** Our calculation of one’s perceived degree of stress is rather basic. There are more profound, in-depth methods of quantifying these rather subjective phenomena in a manner that could prevent registration errors and provide a better distribution on the “stress scale”. For instance, previous studies mention several traumatic life events, such as the death of a loved one, to be potential sources of stress. These studies attach weights to these traumatic events, so that the less influential events have less impact on the overall score. We could

also have included some open questions to the respondents, to gain insight into both the sources of stress and how people cope with them. These open questions could then provide valuable input about potential gaps in our survey.

- **About other covariates:** There are numerous other potential covariates:
  - Do you use a daycare provider for your baby?
  - Do you use a childcare provider for your children?
  - How often do you order food (delivered) because of lack of time for cooking?
  - Do you use external house cleaning services?
  - How often do (did) you require the help of therapeutic counseling?
  - Is your job mostly about responsibilities or more operational?
  - Do you have a management position at your company?
  - How many people depend on your work in your company?

Appendix A: Figures and Tables



Survey
n=158

Subjects with partner that didn't fill out all the relevant questions were excluded (n=9)

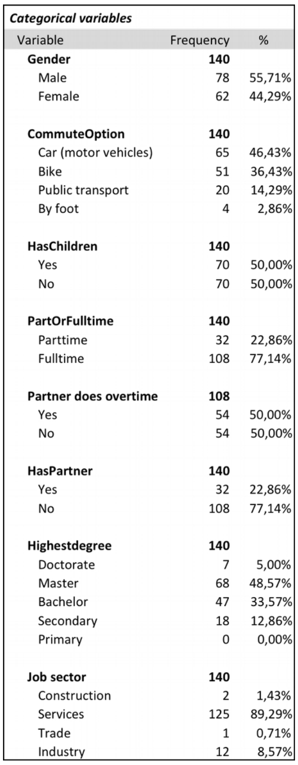
n=149

Subjects with partner and children that didn't fill out all the relevant questions were excluded (n=6)

n=143

Obvious outliers were excluded: Salary of 1€ and Personal Quality Time of 180hours/week (n=3)

n=140

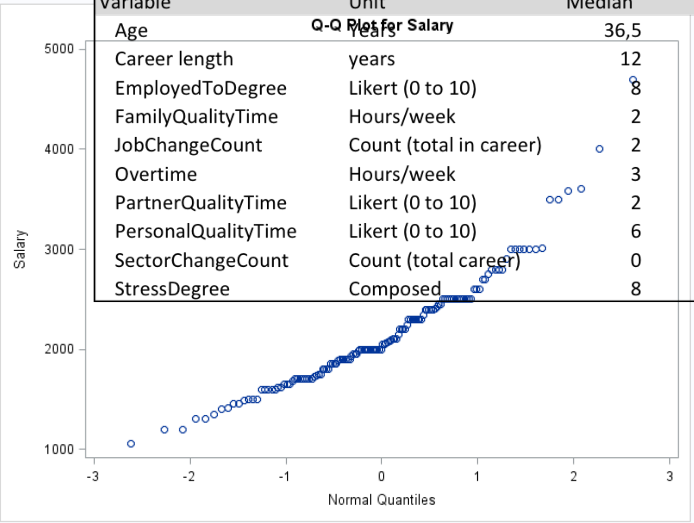


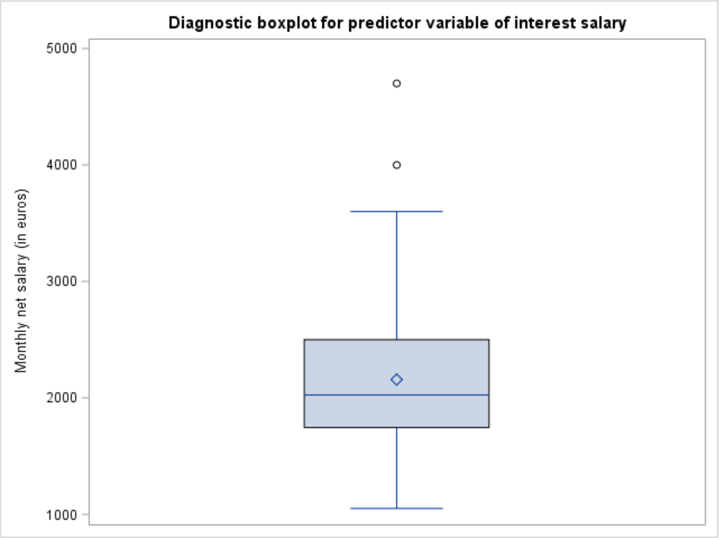
Categorical variables

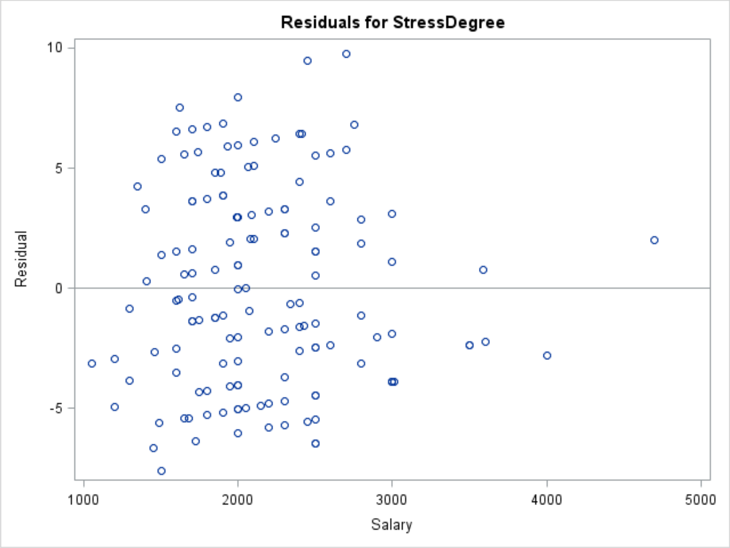
Variable	Frequency	%
Gender	**140**	
Male	78	55,71%
Female	62	44,29%
CommuteOption	**140**	
Car (motor vehicles)	65	46,43%
Bike	51	36,43%
Public transport	20	14,29%
By foot	4	2,86%
HasChildren	**140**	
Yes	70	50,00%
No	70	50,00%
PartOrFulltime	**140**	
Parttime	32	22,86%
Fulltime	108	77,14%
Partner does overtime	**108**	
Yes	54	50,00%
No	54	50,00%
HasPartner	**140**	
Yes	32	22,86%
No	108	77,14%
Highestdegree	**140**	
Doctorate	7	5,00%
Master	68	48,57%
Bachelor	47	33,57%
Secondary	18	12,86%
Primary	0	0,00%
Job sector	**140**	
Construction	2	1,43%
Services	125	89,29%
Trade	1	0,71%
Industry	12	8,57%

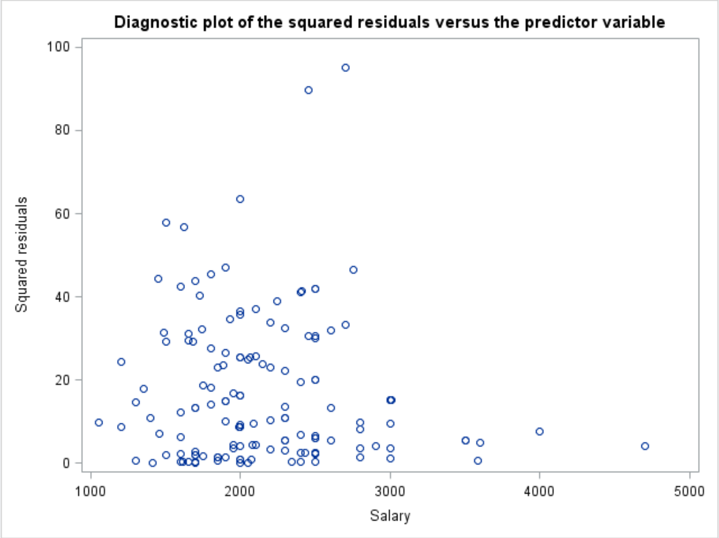
Continuous variables

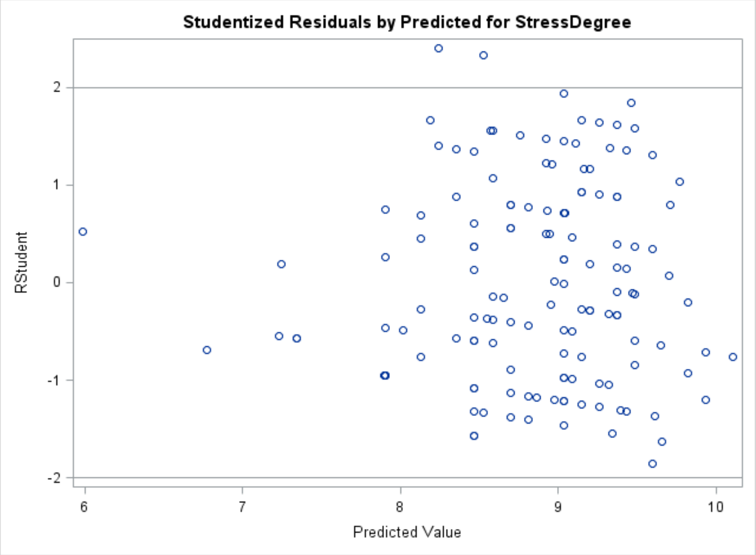
Variable	Unit	Mean	Std. Dev.	N
ChildcareComparison	Likert (0 to 10)	5,36	1,64	70
CommutingTime	Hours/day	51,81	39,65	140
Salary	Euro/month (net)	2157,25	569,89	140
Variable	**Unit**	**Median**	**IQR**	**N**
Age	Years	36,5	14,50	140
Career length	years	12	13,50	140
EmployedToDegree	Likert (0 to 10)	8	4,00	140
FamilyQualityTime	Hours/week	2	4,00	70
JobChangeCount	Count (total in career)	2	3,00	140
Overtime	Hours/week	3	4,00	140
PartnerQualityTime	Likert (0 to 10)	2	3,00	108
PersonalQualityTime	Likert (0 to 10)	6	6,00	140
SectorChangeCount	Count (total career)	0	2,00	140
StressDegree	Composed	8	7,00	140

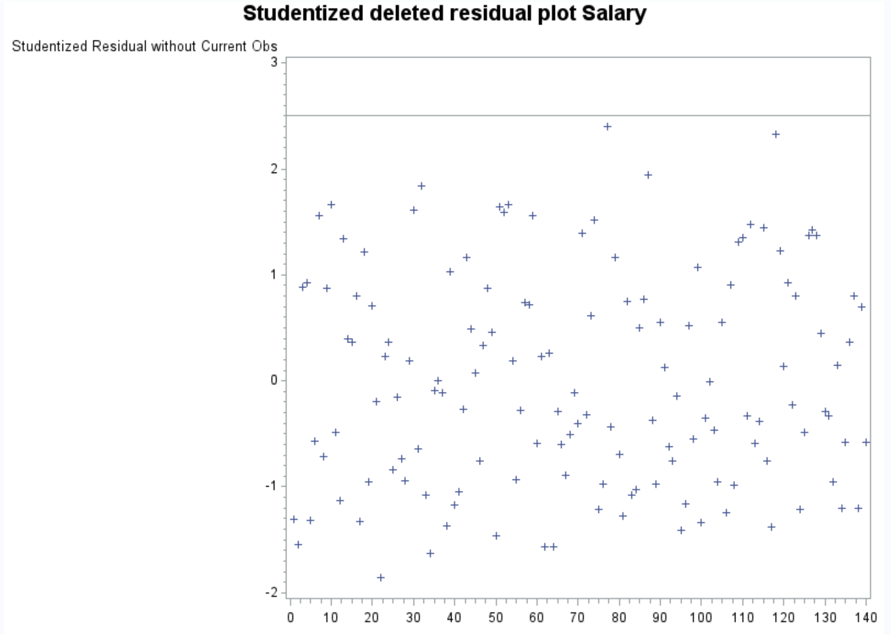


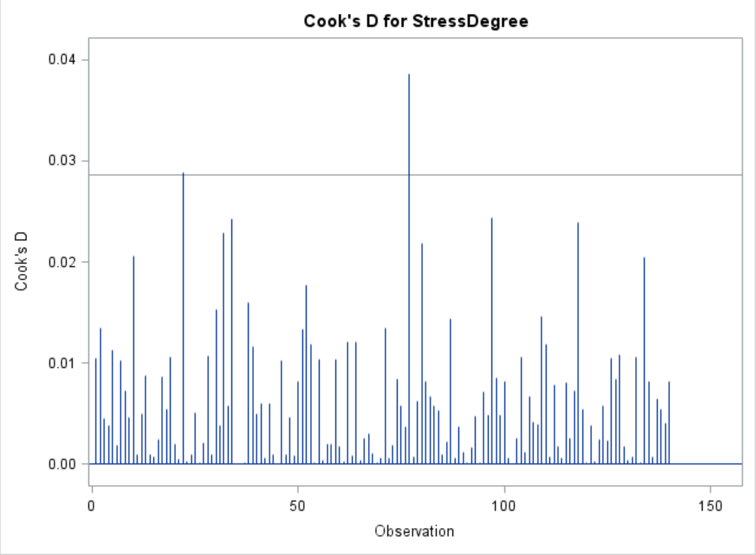

Diagnostic boxplot for predictor variable of interest salary
Monthly net salary (in euros)
5000
4000
3000
2000
1000

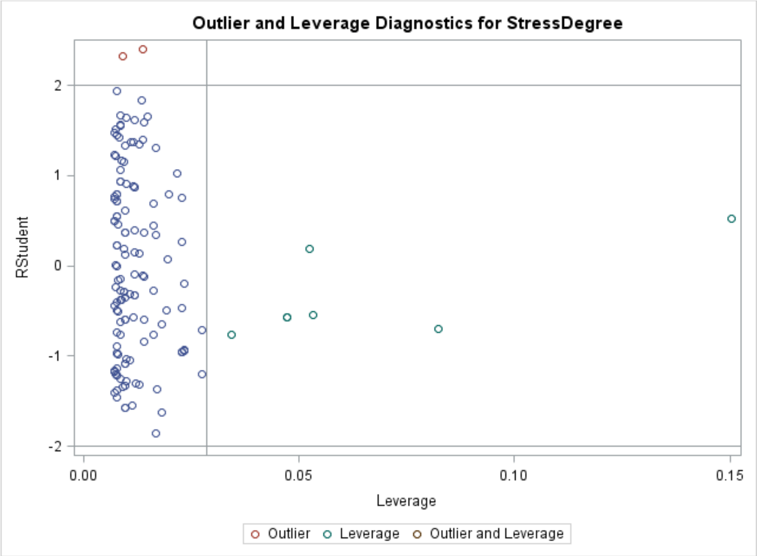

Residuals for StressDegree
Residual
10
5
0
-5
1000
2000
3000
4000
5000
Salary

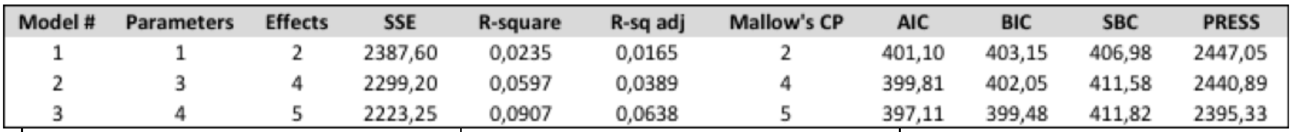

Diagnostic plot of the squared residuals versus the predictor variable
Squared residuals
100
80
60
40
20
0
1000
2000
3000
4000
5000
Salary


Studentized Residuals by Predicted for StressDegree
RStudent
2
1
0
-1
-2
6
7
8
9
10
Predicted Value

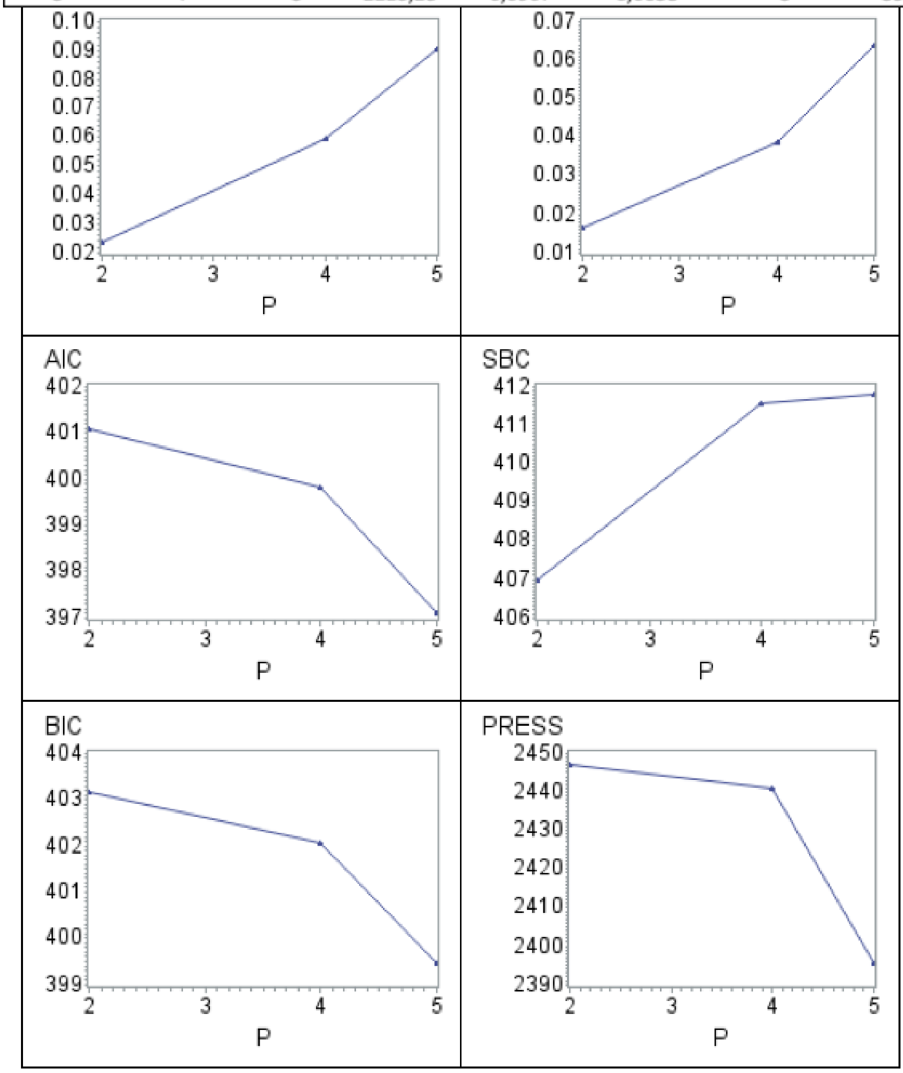

Studentized deleted residual plot Salary
Studentized Residual without Current Obs
3
2
1
0
-1
-2
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140


Cook's D for StressDegree
Cook's D
0.04
0.03
0.02
0.01
0.00
0
50
100
150
Observation





Model #	Parameters	Effects	SSE	R-square	R-sq adj	Mallow's CP	AIC	BIC	SBC	PRESS
1	1	2	2387,60	0,0235	0,0165	2	401,10	403,15	406,98	2447,05
2	3	4	2299,20	0,0597	0,0389	4	399,81	402,05	411,58	2440,89
3	4	5	2223,25	0,0907	0,0638	5	397,11	399,48	411,82	2395,33


P
P
AIC
P
SBC
P
BIC
P
PRESS
P