Table of Contents

List of Figures

Question 1

Estimate some regression equations from the Social Stratification in Eastern Europe after 1989 data for Czechoslovakia (csk.dta). The data was collected in 1993. Turn year of birth (birth) in age. Recode respondent's own education and spouse's education to have the same and a manageable number of categories (3 or 4). For each regression estimate, exclude all cases with missing data on any variable. Make sure for all variables that all values that are missing values, are recognized by Stata as missing values (i.e. have the value ".").

a) *Regress household income on whether the respondent was ever a member of the communist party and her or his age. Write a paragraph interpreting your results. Consider the coefficients, their statistical significance (i.e. mention the null hypothesis, the test statistic, whether or not the null hypothesis was rejected, and what you substantively learned from that test.), and the amount of variance explained by the equation.*

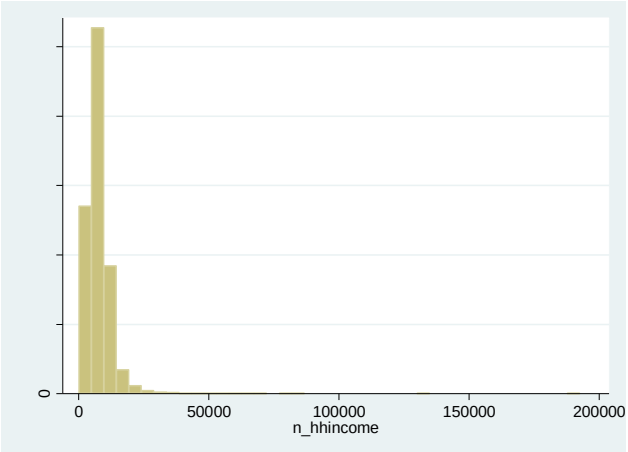

Fig. 1 – Household Income Distribution

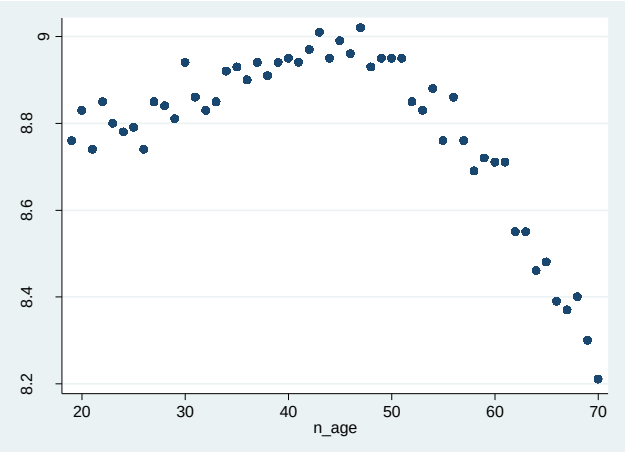

Fig. 2 –Mean of Household Income per Age

Additional preparation

- Due to the distribution of household income, it gets logarithmized (Fig. 1).
- To accept the linear assumption underlying a linear regression-model, we make a spline of age at the age of 48 years (Fig. 2).

Source	SS	df	MS	Number of obs	=	10,233
				F(3, 10229)	=	440.19
Model	357.385296	3	119.128432	Prob > F	=	0.0000
Residual	2768.2588	10,229	.270628487	R-squared	=	0.1143
				Adj R-squared	=	0.1141
Total	3125.64409	10,232	.305477335	Root MSE	=	.52022

n_hhincome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
n_party						
Yes	.142755	.0146979	9.71	0.000	.1139441	.1715658
age1	.0075873	.0007189	10.55	0.000	.0061781	.0089965
age2	-.0312087	.0009247	-33.75	0.000	-.0330213	-.029396
_cons	8.981745	.0103783	865.44	0.000	8.961402	9.002089

Fig. 3 – Regression-Model of Household Income, Party-Membership & Age

The regression-model (Fig. 3) contains 10.233 observations. Since 'Prob > F' is below 0.05, the model is appropriate for interpretations and has some explanatory power. The regression-equation explains 11.43% of the variance, which can be considered okay in social sciences (Adj. r^2= 11.41%). Since we logarithmized the household income, the constant is complex to interpret, but we can observe a geometric mean of 7956.5 monthly income if the respondent was never part of the CP and 48 years old.[1]

If the respondent ever was in the CP, the geometric mean of household income raises about 15%. This coefficient is statistically significant ($p < 0.000$), which means, that the null-hypothesis, that the effect equals zero and thus no difference in the household income between respondents, that were ever in the CP or not, is observable, can be rejected. The 95%-conf.-interval indicates that the coefficient is located between 12.1% and 18.7% in 95 of 100 samples.[2] Therefore, we can assume, that the party-membership has a positive effect on the household income.

If the respondent is between 19-48 years old, the geometric mean of household income raises about 0.8% with every additional year. This coefficient is statistically significant ($p < 0.000$), which means, that the null-hypothesis, that the effect equals zero, can be rejected. Therefore, we can assume, that every additional year of age, until 48 years, has a minor positive effect on the household income. In comparison to that, the household income decreases by 3% with every

[1] Interpretations are based on https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqhow-do-i-interpret-a-regression-model-when-some-variables-are-log-transformed/. E.g. in this case the coefficients are exponentiated (see do-file "an_fwilhelm_task2").

[2] Since the confidence-interval was not mentioned in the question, I omit to calculate it for every coefficient, because this would take a lot of time and I am not quite sure, if I am obligated to do so. I hope this sentence is enough to show you that I understand the concept of confidence intervals and can differentiate it from e.g. credibility intervals (Bayesian statistics).

additional year after being 48 years old. This effect is also statistically significant ($p < 0.000$), which means, that the null-hypothesis, that the effect equals zero, can be rejected. Therefore, we can assume, that being older than 48 years has a negative effect on the household income.

b) Diagnose the model, and if necessary fix it. Do your substantive results change?

Diagnosing the model with a residual-versus-fitted plot ('rvfplot'), a test concerning the centered variance inflation factors for the independent variables ('estat vif') and a heteroskedasticity-test ('estat hettest') show no noteworthy problems.

However, Fig. 4 indicates some leverage and influential observations, that could skew our regression-model. Therefore, we exclude some of these observations, run a second regression model and compare both models.

Fig. 5 shows the comparison of both models. We see that the differences concerning the second model without influential observations are minor and thus the substantive results do not change.

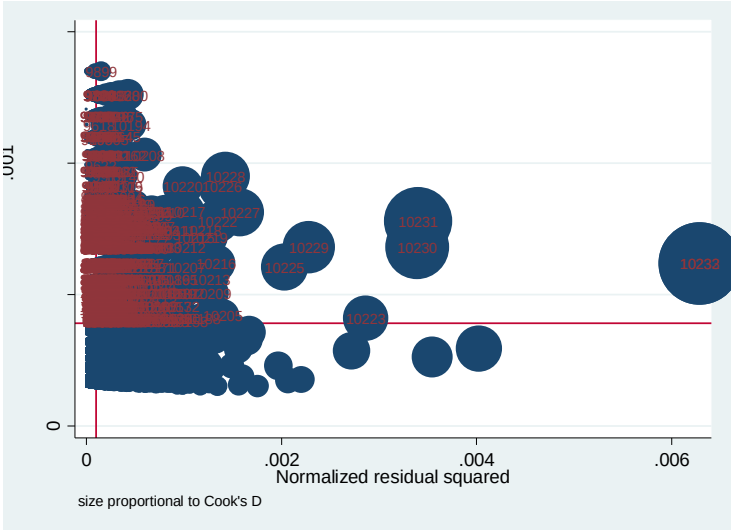

Fig. 4 – 'lvr2plot2'

	(1) ln_hhincome	(2) ln_hhincome
0bn.n_party	.	.
	.	.
1.n_party	0.143*	0.141*
	(0.000)	(0.000)
age1	0.00759*	0.00724*
	(0.000)	(0.000)
age2	-0.0312*	-0.0311*
	(0.000)	(0.000)
_cons	8.982*	8.980*
	(0.000)	(0.000)
N	10233	10228
r2	0.114	0.116

p-values in parentheses
* p<0.05

Fig. 5 – Regression models with (1) and without (2) influential observations.

c) Add the education of the respondent and the spouse to your model. Write a paragraph interpreting your results. In particular, comment on what the introduction of education does to the coefficients for communist party membership. Is education an intervening or a confounding variable in this context?

This regression-model (Fig. 6) contains 7.030 observations and since 'Prob > F' remains below 0.05, the model is appropriate for interpretations and has some explanatory power. The regression-equation explains 15.76% of the variance, which is a little bit better than the previous model (Adj. r^2= 15.65%). We can see a geometric mean of 7424.4 monthly income if the respondent was never part of the CP, is 48 years old, and the respondent and their spouse have not completed primary education. The reduced constant indicates already, that the education-variables could be a confounding variables.

If the respondent ever was in the CP, the household income now only increases by 4.5%, but remains statistically significant ($p < 0.001$). Therefore, we can assume, that the party-membership has a positive effect on the household income. In the context of introducing both education variables in this model, we can observe that education seems to explain the monthly income better than the party-membership. The effect reduces drastically but remains statistically significant. In this case, education of the respondent and the spouse are confounding variables.

	(1) ln_hhincome	(2) ln_hhincome
0bn.n_party	.	.
	.	.
1.n_party	0.143*	0.0443*
	(0.000)	(0.001)
age1	0.00759*	0.0134*
	(0.000)	(0.000)
age2	-0.0312*	-0.0237*
	(0.000)	(0.000)
0bn.n_educ		.
		.
1.n_educ		0.0555
		(0.062)
2.n_educ		0.159*
		(0.000)
3.n_educ		0.291*
		(0.000)
0bn.n_speduc		.
		.
1.n_speduc		0.0673
		(0.082)
2.n_speduc		0.141*
		(0.000)
3.n_speduc		0.202*
		(0.000)
_cons	8.982*	8.913*
	(0.000)	(0.000)
N	10233	7030
r2	0.114	0.158

p-values in parentheses
* p<0.05

Fig. 6 – Regression-Model of Household Income, Party-Membership, Age, respondent's Education & spouse's Education

If the respondent is between 19-48 years old, the household income raises now about 1.4% with every additional year and remains statistically significant ($p < 0.000$). Therefore, we can assume, that with every additional year of age, until being 48 years old, has a positive effect on the household income. In comparison to that, the household income decreases by 2% with every additional year after being 48 years old. This effect remains also statistically significant (p <

0.000). Therefore, we can assume, that being older than 48 years has a negative effect on the household income.

In comparison to a respondent without primary education, the monthly income increases by 5.7%, if the respondent completed only primary education, but the effect is not statistically significant on a 5%-significance-level ($p < 0.062$). Thus, the null-hypothesis, that there is no difference between a respondent without primary education and with only primary education, cannot be rejected. Furthermore, for respondents, that completed secondary education, the monthly income increases by 17.2% – for respondents, that completed tertiary education, it increases even by 33.8% – in comparison to respondents without primary education. These effects are both statistically significant ($p<0.000$) and thus different to the effect of the reference-category. These results show that education of the respondents got a relatively large impact on the monthly income.

The education of the respondent's spouse indicates a similar appearance. In this case the monthly income increases by 7%, if the respondent's spouse got just primary education, in comparison to a respondent's spouse without primary education. Also, this effect is not statistically significant on a 5%-significance-level ($p < 0.082$) and therefore the null-hypothesis, that these effects are the same, can not be rejected. Moreover, for the respondent's spouse, that completed secondary education, the monthly income increases by 15% – for the respondent's spouse, that completed tertiary education, it increases even by 22% – in comparison to the respondent's spouse without primary education. These effects are both statistically significant ($p<0.000$) and thus different to the effect of the reference-category. These results show that the education of the respondent's spouse got a relatively big impact on the monthly household income.

d) Create two new variables: education of the male member of the household and education of the female member of the household (assume only heterosexual relationships). Add those two variables instead of the previous education variables. Write a paragraph interpreting your results. In particular, answer the question whether male and female education are equally important for household income or not.

This regression-model (Fig. 7) contains 7.030 observations and since 'Prob > F' remains below 0.05, the model is appropriate for interpretations and has some explanatory power. The regression-equation explains 15.77% of the variance (Adj. r^2= 15.66%). We can see a geometric mean of 7510.4 monthly income, if the respondent was never part of the CP, is 48 years old, and neither the male or female household-member completed primary education.

If the respondent ever was in the CP, the household income increases by 5% and the coefficient is statistically significant ($p < 0.000$). Therefore, we can assume, that the positive effect of the party-membership on the household income remains.

If the respondent is between 19-48 years old, the household income raises about 1.3% with every additional year and remains statistically significant ($p < 0.000$). Therefore, we can assume, that every additional year until 48 years has a positive effect on the household income. In comparison to that, the household income decreases by 2% with every additional year after being 48 years old and remains statistically significant ($p < 0.000$). Therefore, we can assume, that being older than 48 years has a negative effect on the household income.

Source	SS	df	MS	Number of obs	=	7,030
				F(9, 7020)	=	146.04
Model	228.449577	9	25.3832864	Prob > F	=	0.0000
Residual	1220.16066	7,020	.17381206	R-squared	=	0.1577
				Adj R-squared	=	0.1566
Total	1448.61024	7,029	.206090516	Root MSE	=	.41691

ln_hhincome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
n_party						
No	0	(base)				
Yes	.0488889	.0135457	3.61	0.000	.0223352	.0754425
age1	.0133179	.0007624	17.47	0.000	.0118233	.0148124
age2	-.0236797	.0009627	-24.60	0.000	-.0255669	-.0217925
m_educ						
Primary Incomplete	0	(base)				
Primary Complete	.0255751	.0366442	0.70	0.485	-.0462585	.0974088
Secondary Complete	.0999664	.0378528	2.64	0.008	.0257635	.1741694
Tertiary Complete	.2362969	.0397122	5.95	0.000	.158449	.3141449
f_educ						
Primary Incomplete	0	(base)				
Primary Complete	.084169	.0306002	2.75	0.006	.0241833	.1441547
Secondary Complete	.182205	.0319253	5.71	0.000	.1196219	.2447882
Tertiary Complete	.2280759	.0367749	6.20	0.000	.155986	.3001659
_cons	8.924047	.0438594	203.47	0.000	8.83807	9.010025

Fig. 7 – Regression-Model of Household Income, Party-Membership, Age and male's & female's Education

In comparison to a male household-member with incomplete primary education, the monthly income increases by 2.6%, if the male household-member completed just primary education. This coefficient is in comparison to a male household-member with incomplete primary education not statistically significant ($p < 0.485$), if we consider a significance-level of e.g. 5%. Therefore, we cannot reject the null-hypothesis, that there is no difference between male household-member with incomplete and with just primary education. But we can observe an increase of 10.5% ($p < 0.008$) of the monthly income, if the male household-member obtained secondary education, and an increase of 26.7% ($p < 0.000$), if he obtained tertiary education, in comparison to the male household-member with incomplete primary education. These coefficients are both statistically significant and for this reason the null-hypothesis of indifference towards the reference-category can be rejected. Accordingly, we can observe a positive effect of male's education on the monthly income.

In comparison to a female household-member with incomplete primary education, the monthly income increases by 9% ($p < 0.006$), if the female household-member completed only primary education, 20% ($p < 0.000$), if the female household-member completed secondary education, and 25.6% ($p < 0.000$), if the female household-member completed tertiary education. These coefficients are statistically significant, which means that we can reject the null-

hypothesis, that these coefficients do not differ from the reference-category, an incomplete primary education. Therefore, even a completed primary education has a positive effect on the monthly income in comparison to an incomplete primary education of a female household-member.

Comparing the coefficients of the education of male and female household-members, we can observe some differences. E.g. considering the education of female household-members there is a difference between incomplete primary education and just primary education, which is unique in this task. Therefore, we can assume that the level of female education is a bit more important concerning these two categories. Regarding the effect of a complete secondary education, we also see that the effect is bigger among female household-members, than among male household-members. Here, it is necessary to note, that we cannot just compare the coefficients of male and female household-members, since these coefficients only refer to their reference-category. Furthermore, we can acknowledge only a minor difference of a completed tertiary education between male and female household-members. It seems, that at this education-level the gender does not have a noteworthy effect on the household-income. Recapitulating, we can answer, that male and female education are not equally important for household income, concerning incomplete primary, primary and secondary education-levels.

Question 2

Estimate a logistic regression from the NORC 2004 General Social Survey data (gss_2004.dta). Recode divorce such that it is 0 when never divorced and 1 when divorced. Make sure for all variables that all values that are missing values, are recognized by Stata as missing values (i.e. have the value ".").

Additional preparation

- This task is a little bit tricky, since the variable 'divorce' only contains observation about respondents, that were currently married or widowed.[3] Accordingly, the variable miss information about respondents that are currently divorced.

  Furthermore, the variable 'Marital', that also should be used to investigate this topic adequately, contains the category 'separated'. Since someone, who was not 'Never married' and not 'Divorced, but 'Separated' they are still married. Unfortunately, there is no detailed information about 'Separated'. It could be that 'Separated' means 'Legally Separated' which would be the same than 'Divorced', but since this information is lacking, I treat these observations as missings. Additionally, it is possible that these respondents e.g. live separately and only misinterpreted the question. Due to these uncertainties I am willing to 'loose' these 95 observations.

  Conclusively, a new variable is being generated, that contains information of the variables 'Marital' and 'Divorce'.
- Theoretically, I would not use the variable 'educ' as education-variable, since years of schooling do not provide adequate interpretations about the education of a respondent. In such cases, the highest degree of the respondent is a better measurement for the education ('degree'). But regarding Task 3 b), where we should use an interaction effect, the variable 'educ' seems more reasonable, because the logit-model gets a bit confusing, when adding an interaction-effect to a categorial-variable with several characteristics.
- Since age seems to have an u-shaped relationship with divorce, we make a spline of age at the age of 55 years (Fig. 8).

[3] www.thearda.com/Archive/Files/Codebooks/GSS2004_CB.asp

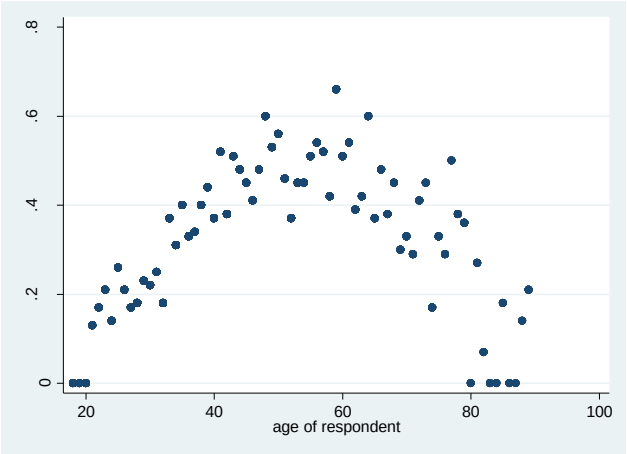

Fig. 8 – Proportion of divorce per age

a) *Estimate a logistic regression explaining whether a person was divorced with their education, age, and how often they attend a religious service. Interpret the model in terms of odds ratios and marginal effects.*

The model contains 2.078 observations. The chi2-test shows that none of the coefficients, in this case odds ratio's, are equal zero. Also, we can detect a pseudo-r2 of ~7%, which should be interpreted carefully. Among respondents, that completed 12 years of education, are 55 years old and never attend religious service, we can expect 2.63 respondents, that were at least divorced one time, for every respondent, that never was divorced.

```
Logistic regression                               Number of obs     =      2,078
                                                  LR chi2(7)        =     183.31
                                                  Prob > chi2       =     0.0000
Log likelihood = -1298.2639                       Pseudo R2         =     0.0659
```

n_div	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
c_educ	.9403603	.015535	-3.72	0.000	.9103998	.9713067
age_0	1.055097	.0060546	9.35	0.000	1.043297	1.067031
age_1	.9372366	.0071052	-8.55	0.000	.9234137	.9512665
n_attend						
Never	1	(base)				
(lt) once a year	.7622544	.1191419	-1.74	0.082	.5611196	1.035486
st a year/monthly	.6064884	.0956821	-3.17	0.002	.4451777	.8262501
nrly every week	.5044863	.0846994	-4.08	0.000	.3630271	.7010672
weekly and more	.3804018	.0560931	-6.55	0.000	.2849224	.5078771
_cons	2.626372	.3703546	6.85	0.000	1.992165	3.462481

Fig. 9 – Logit-Model of Divorce, Years of Education, Age & Church Attendance

This odds decreases by 6% with every additional year of education and is statistically significant, since p<0.000, which means, that the null-hypothesis, that no effect is observable, can be rejected. Therefore, we can assume, that higher education reduces the odds of being divorced a little bit.

The odds of being divorced increases by a factor 1.06 with every additional year, if the respondent is younger than 55 years. Alternatively, this odds decreases by a factor .94 with every additional year, if the respondent is older than 55 years. Both effects are statistically significant. Consequently, we can argue, that younger people rather get divorced, than older people, but the odds remain pretty small.

In comparison to a respondent, that never attends religious services, the odds of being divorced decreases by 24%, if one attends these services less than or once a year; by 39%, if one attends these services several times a year or monthly; by 49%, if one attends these services nearly every week; and by 62%, if one attends these services weekly or even more. Concerning the null-hypothesis, that the effect of attending religious services does not differ between respondents, that never or less than/once a year attend these services, the p-values does not allow us to reject this null-hypothesis on 5%-significance-level. In comparison to that, the increased attendance of religious services is statistically significant in comparison to the reference-category. Therefore, we can assume, that attending religious services at least several times a years has an impact on the odds of getting divorced.

```
Average marginal effects                        Number of obs     =      2,078
Model VCE    : OIM

Expression   : Pr(n_div), predict()
dy/dx w.r.t. : c_educ age_0 age_1 1.n_attend 2.n_attend 3.n_attend 4.n_attend
```

	dy/dx	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
c_educ	-.0133856	.0035522	-3.77	0.000	-.0203477	-.0064234
age_0	.0116747	.0011504	10.15	0.000	.00942	.0139294
age_1	-.0141099	.0015443	-9.14	0.000	-.0171366	-.0110831
n_attend						
(lt) once a year	-.063362	.0364229	-1.74	0.082	-.1347496	.0080255
st a year/monthly	-.1157245	.0363028	-3.19	0.001	-.1868768	-.0445723
nrly every week	-.1564656	.0378613	-4.13	0.000	-.2306724	-.0822589
weekly and more	-.2152495	.0326954	-6.58	0.000	-.2793313	-.1511676

```
Note: dy/dx for factor levels is the discrete change from the base level.
```

Fig. 10 – Average marginal effects of previous Logit-Model

Since the marginal effects show us the chance in probability for a unit change in x, we can observe, that with every additional year of completed schooling the chance of getting divorced decreases by 1.3 percent-points. Being younger than 55 years, an additional year increases the chance by 1.2 percent-points, while being older than 55 years, an additional year decreases the chance by 1.4 percent-points. If one attends religious services less than once a year or once a year the chance getting divorced decreases by 6.3 percent-points; 11.6 percent-points, if one attends it several times a year or monthly; 15.7 percent-points, if one attends it nearly every week and 21.5 percent-points, if one attends it weekly or even more, in comparison to someone, who does not attend any religious service.

The results remain the same. Education, being older than 55 years and attending religious services reduces the chance getting divorced, while being younger than 55 years slightly increases the chance getting divorced.

b) Use an interaction effect to investigate whether the effect of education is the same for men and women. Look substantively at the size of the difference and whether or not that difference is statistically significant (again, mention the null hypothesis, the test statistic used, whether or not you reject that null hypothesis, and what you substantively learned from that test.).

Logistic regression — Number of obs = 2,078; LR chi2(9) = 185.78; Prob > chi2 = 0.0000; Log likelihood = -1297.0271; Pseudo R2 = 0.0668

n_div	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
c_educ	.9523788	.0220937	-2.10	0.035	.9100457	.9966812
n_sex						
Female	1	(base)				
Male	.9089752	.1007693	-0.86	0.389	.7314548	1.129579
n_sex#c.c_educ						
Male	.9773358	.0318377	-0.70	0.482	.9168854	1.041772
age_0	1.055641	.0060746	9.41	0.000	1.043802	1.067615
age_1	.9370503	.0071305	-8.54	0.000	.9231786	.9511306
n_attend						
Never	1	(base)				
(lt) once a year	.7580091	.1186253	-1.77	0.077	.5577825	1.030111
st a year/monthly	.596848	.0944681	-3.26	0.001	.4376602	.8139364
nrly every week	.4963042	.0836005	-4.16	0.000	.356752	.6904457
weekly and more	.3724768	.0553007	-6.65	0.000	.2784347	.4982818
_cons	2.782446	.4277288	6.66	0.000	2.058622	3.76077

Fig. 11 – Logit-Model of Divorce, Years of Education, Gender, Interaction Effect of Education and Gender, Age & Church Attendance

This model contains 2.078 observations. The chi2-test shows that none of the coefficients, in this case odds ratio's, are equal zero. Also, pseudo-r2 remains at ~7%, which should be interpreted carefully.

Among female respondents, that completed 12 years of education, are 55 years old and never attend religious service, we can expect 2.78 respondents, that were at least divorced one time, for every respondent, that never was divorced.

This odds decreases by 9%, if the respondent is male. This observation could be interesting, if we suggest homosexual marriages, but in the year of 2004, I am not quite sure, if homosexual marriages were allowed in the USA. Therefore, this could be an interesting observation, but I will not follow up this topic. Besides, we cannot reject the null-hypothesis, that there is no difference between the genders ($p< 0.389$) and therefore, we cannot observe an actual effect of being male on the odds.

The effect of education is nearly the same for men and women. We can observe an additional minor decrease of education about 2% on the odds, if the respondent is male. But that difference is not statistically significant, since $p<0.482$, which means that we cannot reject the null-hypothesis of no difference of the effect of education between men and women, if we consider a 5%-significance-level, which is reasonable in this case. Conclusively, we can answer the question, that the effect of education on the odds of getting divorced remains the same for men and women.