What are the drivers of phishing attack exposure?

Oleh Stupak

December 6, 2017

1 Introduction

This report presents the results of work that examines casual relationships between enterprises' exposure to a concrete cyberattack and basic indices of enterprises' activity in the information environment. The research is of quantitative nature. It uses regression analysis method to answer the question:

What are the drivers of phishing attack exposure?

The paper should be considered as one written for study purpose only. Its main objective is to show decent statistical knowledge and skills. Therefore, the paper focuses on technical aspects of the research and in-depth model analyses. It reveals the research's methodological aspects and describes the inquiry step-by-step.

The structure of the paper is as follows: this section introduces and does general research overview. Hypotheses presented in Section 2. Section 3 provides information on data and variables. Section 4 reveals the empirical strategy and base model. Verification and adjustment are presented in Section 5. Section 6 presents final model and results. Section 7 concludes.

2 Hypotheses

Before paper goes into in-depth analyses, it is necessary to present the hypotheses that serve as a base for the model. The broad hypothesis is:

The enterprises' activity in the information environment has a direct influence on the enterprises' phishing attack exposure.

For the sake of the feasibility, transparency, and simplicity the presented hypothesis should be narrowed and detailed. Therefore, the paper purposes the next set of hypotheses:

1. *Enterprises' computer usage environment has a direct influence on the enterprises' phishing attack exposure;*

2. *Enterprises' total turnover from e-commerce has a direct influence on the enterprises' phishing attack exposure;*

3. *Enterprises' internet usage for banking and financial purposes has a direct influence on the enterprises' phishing attack exposure;*

4. *Enterprises' open-source software usage has a direct influence on the enterprises' phishing attack exposure;*

5. *Enterprises' usage of advanced e-signatures has a direct influence on the enterprises' phishing attack exposure.*

The hypotheses have been built relying on available data and author's local knowledge.

3 Data

3.1 Dataset

The research uses Eurostat working database with the results of the surveys on the usage of information and communication technologies in enterprises and households for modelling purpose.

The dataset is built leaning on questionnaires from approximately 148 000 companies from all over the European Union. It aggregates the enterprises by country and economic activities that correspond to the classification NACE Revision 2. You can see countries' table of frequencies in the appendix (Appendix 1).

To avoid double counting aggregated observations have been dropped (EA16, EU15, EU25, EU27, EU28).

The all-European survey was composed by numerous statistical agencies. The survey was conducted on the online basis. The database includes 1129 variables revealing all aspects of enterprises' life in the information environment.

3.2 Dependent variable

As mentioned before, the primary objective of the paper is to model enterprises' exposure to phishing attacks. Therefore, the dependent variable is constructed as a share of enterprises in the industry experienced ICT related security incidents that resulted in the disclosure of confidential data due to intrusion, pharming or phishing attacks. Unfortunately, the variable

was collected only for the 2010 year. That imposes limitations on further research, e.g., impossibility to investigate the dynamics of the variable. This restriction forces to construct the cross-sectional dataset and ignore the time dimension. You can find variable's content below (Figure 1).

Figure 1: Dependent variable characteristics

```
E_SECICNFA                Share of enterprises that experienced Intrusion, pharming, phishing

             type:  numeric (double)

            range:  [0,.391667]                  units:  1.000e-06
    unique values:  1273                     missing .:  12418/14139

             mean:    .01434
         std. dev:   .023615

      percentiles:        10%       25%       50%       75%       90%
                            0   .000638   .007534   .017929      .036
```

It is evident that another possible limitation is a large share of missing values. There are two possible explanations:

- As mentioned before, observations are collected only for the 2010 year;
- Limited ICT awareness among some industries (e.g., industries that are not operating in information environment at all).

3.3 Independent variables

The set of independent variables is represented by:

1. Employees internet activity. Constructed as a portion of persons employed using computers with access to World Wide Web in the overall quantity of labour employed;
2. Turnover from electronic commerce. Constructed as a percentage of trading in products or services using computer networks (such as the Internet) in the overall turnover;
3. The usage of electronic finance and banking systems. Constructed as a share of enterprises using the World Wide Web for the financial and banking purpose;

4. Open-source software usage. Constructed as a proportion of enterprises that use third-party open-source software (free, editable software available to everyone for any purpose);

5. Electronic signature usage. Constructed as a percentage of enterprises that use e-signature for identity verification or online contract submitting.

Each variable corresponds to one of hypothesis mentioned above. Variables' contents could be found in the appendix (Appendix 2). The independent variables summative table could be found below (Figure 2).

Figure 2: Independent variables' summative table

Variable	Obs	Mean	Std. Dev.	Min	Max
P_CUSE	11262	.5598495	.2370545	0	1
E_ETURN	8758	.1170855	.1123799	0	.886114
E_IBK	3276	.8199685	.1562884	0	1
E_OSOPEN	3787	.2149106	.1677576	0	1
E_DIGSIGN	3831	.3199061	.2316285	0	1

For the sake of dependent variable, the research considers only 2010 year's observations.

4 Empirical strategy

As an estimation method, the paper suggests the use of ordinary least squares (OLS). In order to test hypotheses, following specification estimated:

$$SECICNFA = \beta_0 + \beta_1 * CUSE + \beta_2 * ETURN + \beta_3 * IBK +$$

$$+\beta_4 * OSOPEN + \beta_5 * DIGSIGN + \varepsilon$$

Where,

- $SECICNFA$ - dependent variable; share of enterprises that experienced intrusion, pharming, phishing;
- $CUSE$ - explanatory variable; share of persons employed using computers;
- $ETURN$ - explanatory variable; Share of enterprises' total turnover from e-commerce;

- IBK - explanatory variable; share of enterprises that use the Internet for banking and financial services;
- $OSOPEN$ - explanatory variable; share of enterprises that use open source operating systems;
- $DIGSIGN$ - explanatory variable; share of enterprises that use advanced e-signatures;
- β_{0-5} - regression coefficients;
- ε - error term.

Regression table for the base model could be found below (Figure 3).

Figure 3: Base model regression table

Source	SS	df	MS		
Model	.011920482	5	.002384096	Prob > F	0.0000
Residual	.285675746	883	.000323529	R-squared	0.0401
Total	.297596228	888	.000335131	Root MSE	.01799

Number of obs	=	889				
F(5, 883)	=	7.37				
Prob > F	=	0.0000				
R-squared	=	0.0401				
Adj R-squared	=	0.0346				
Root MSE	=	.01799				

E_SECICNFA	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
P_CUSE	-.0012596	.0030698	-0.41	0.682	-.0072847	.0047654
E_ETURN	-.0087033	.005946	-1.46	0.144	-.0203734	.0029667
E_IBK	.013546	.0054752	2.47	0.014	.0028001	.024292
E_OSOPEN	.0195445	.004631	4.22	0.000	.0104554	.0286337
E_DIGSIGN	-.003691	.0030747	-1.20	0.230	-.0097256	.0023436
_cons	.0010351	.0037139	0.28	0.781	-.0062539	.0083241

We can observe that model's p-value is equal to 0. It means that relationships between dependent and independent variables are statistically significant. However, the goodness of fit (R^2) is quite low. Moreover, only one variable ($OSOPEN$) is statistically significant in explaining phishing attack exposure.

The model requires verification and adjustment procedures that could be found in the next section.

5 Verification and adjustment

5.1 Heteroskedasticity

The convenient way to test for heteroskedasticity is the Breusch-Pagan test (Cook, Weisberg, 1983). It tests whether the variance of the errors from a regression is dependent on the values of the independent variables. In that case, heteroskedasticity is present. You can find test results below (Figure 4).

Figure 4: Breusch-Pagan test on heteroskedasticity

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of E_SECICNFA

         chi2(1)      =   244.43
         Prob > chi2  =   0.0000
```

Regarding test results it is obvious that the heteroskedasticity problem is present in the model ($p - value$ is less than 0.05 threshold); $H0$ rejected, and heteroskedasticity assumed.

There is no need for extra evidence to support the presence of heteroskedasticity. However, original variables' scatter plots could be found in the appendix (Appendix 3) to support the statement.

Presented scatter plots also contain locally weighted scatterplot smoothing (LOWESS) line that could be used to check the relationships between dependent and independent variables and foresee trends. The method is also used to test for nonlinearities (Cleveland, 1981). However, LOWEES line will be plotted again after an appropriate data transformation.

Weighted east squares (WLS) could be used as convenient method for dealing with heteroscedasticity (Bjorck, 1996). However, after performing a few WLS estimations, the problem has not been fixed, and results remained similar. You can find example WLS regression table and weighted residuals graph in the appendix (Appendix 4).

Therefore, the paper states that another approach to heteroskedasticity problem should be used. The possible source of heteroskedasticity could be the data format. Variables presented in the model are all from 0 to 1 range and could be considered as proportions. Arcsine transformation could be used to treat the problem. This consists of taking the arcsine of the square

root of a variable. The operation is performed for every variable used in the model. The new variable' summative table could be found below (Figure 5).

Figure 5: Transformed variables summative table

Variable	Obs	Mean	Std. Dev.	Min	Max
secicnfa_t	1721	.0918026	.078688	0	.6761991
cuse_t	11262	.868131	.2867924	0	1.570796
eturn_t	8758	.3133477	.1735747	0	1.226568
ibk_t	3276	1.165979	.2097196	0	1.570796
osopen_t	3787	.4564513	.2066642	0	1.570796
digsign_t	3831	.5838852	.2831199	0	1.570796

The OLS regression now should be re-estimated with transformed variables and again tested for heteroskedasticity. Re-estimated regression table could be found below (Figure 6).

Figure 6: Base regression with transformed variables

Source	SS	df	MS
Model	.165956985	5	.033191397
Residual	4.45387122	883	.005044022
Total	4.61982821	888	.005202509

Number of obs	= 889
F(5, 883)	= 6.58
Prob > F	= 0.0000
R-squared	= 0.0359
Adj R-squared	= 0.0305
Root MSE	= .07102

secicnfa_t	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cuse_t	-.0396665	.0099795	-3.97	0.000	-.0592528	-.0200802
eturn_t	.0196345	.0149486	1.31	0.189	-.0097044	.0489734
ibk_t	.0260986	.0171731	1.52	0.129	-.0076064	.0598035
osopen_t	.0610149	.0152588	4.00	0.000	.0310672	.0909626
digsign_t	-.0101515	.0101199	-1.00	0.316	-.0300134	.0097104
_cons	.0679385	.0149534	4.54	0.000	.0385902	.0972868

We can observe that even the situation has not changed dramatically, another coefficient now became statistically significant in explaining phishing attacks ($cuse_t$). Breusch-Pagan test has shown different results as well (Figure 7).

Figure 7: Breusch-Pagan test on heteroskedasticity with transformed variables

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of secicnfa_t

         chi2(1)      =     6.26
         Prob > chi2  =   0.0123
```

Even the heteroskedasticity is still present, it is significantly reduced comparing to the base model. It means that robust standard errors method could be used to control it (White, 1980). It relaxes assumptions that errors are both independent and identically distributed. As heteroskedasticity is clearly present, robust standard errors tend to be more trustworthy. The finial model should be estimated using *robust* option.

5.2 Nonlinearity

The OLS regression assumes linearity in parameters. As mentioned before, to test this assumption on transformed variables and visualise relationships between dependent and independent variables, scatter plots with fitted LOWESS lines should be built. Five scatter plots representing this relationship could be found below (Figures 8-12).

Figure 8: Transformed phishing exposure and computer usage relationship

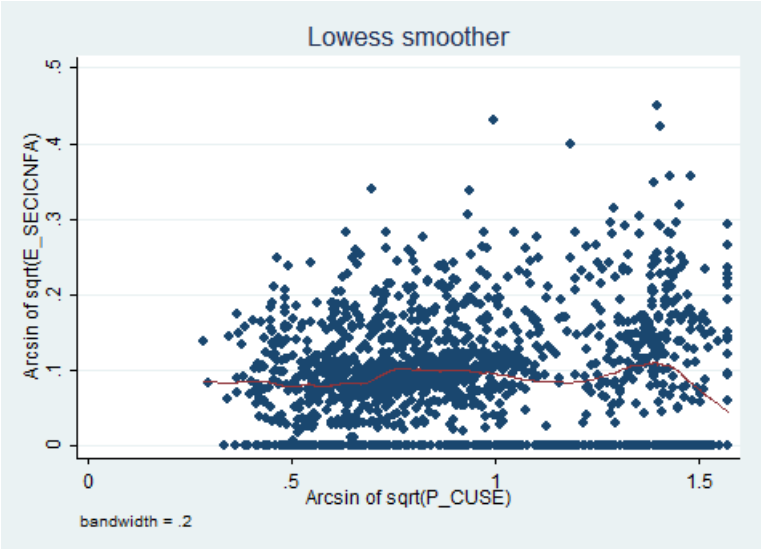


Figure 9: Transformed phishing exposure and e-commerce turnover relationship

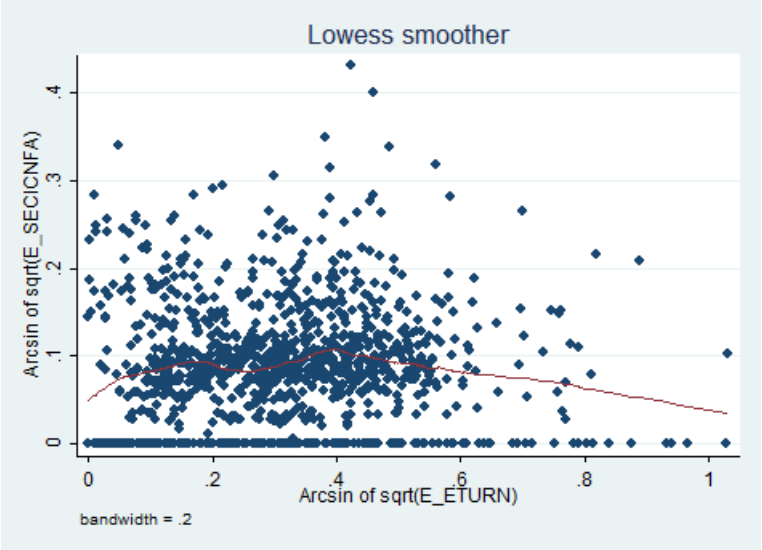


Figure 10: Transformed phishing exposure and banking/financial usage relationship

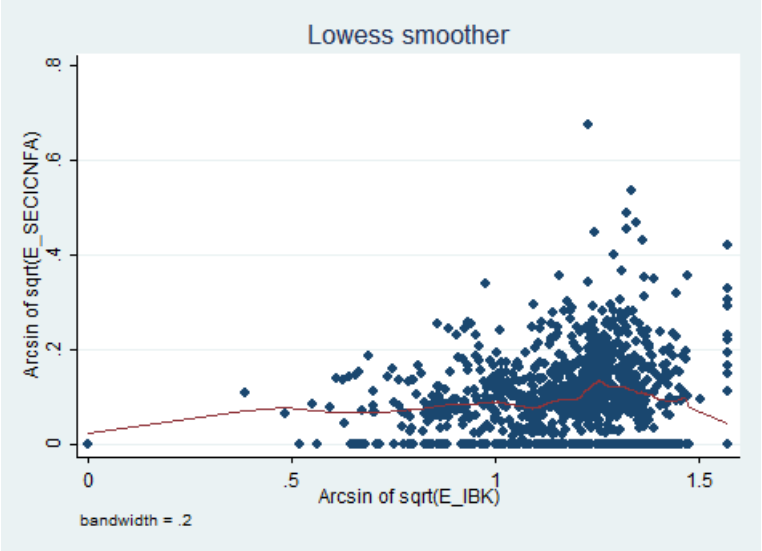


Figure 11: Transformed phishing exposure and open-source software usage relationship

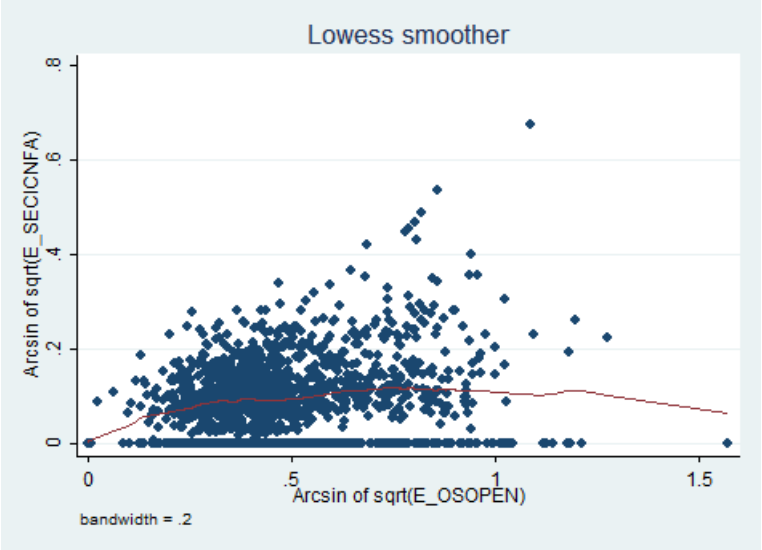
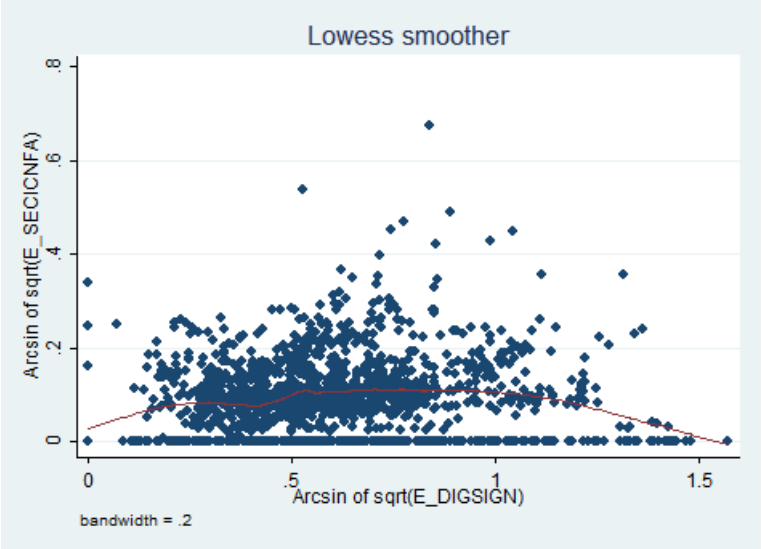


Figure 12: Transformed phishing exposure and advanced digital signature usage relationship



Considering graphs it is possible to assume the linearity in parameters (or the absence of strong nonlinearity). It is also evident that transformation done in the previous subsection significantly reduced heteroskedasticity.

5.3 Collinearity

To test for collinearity (multicollinearity), the paper proposes the procedure that described in 'Regression collinearity diagnostic' (Belsley, Kuh, Welsch, 1980). The procedure examines the "conditioning" of the matrix of independent variables. The matrix built for the presented model could be found below (Figure 13).

Figure 13: Condition indices and variance-decomposition matrix

Condition number using scaled variables = 13.34

Condition Indexes and Variance-Decomposition Proportions

	condition index	cuse_t	eturn_t	ibk_t	osopen_t	digsign_t
1	1.00	0.00	0.01	0.00	0.00	0.01
2	4.78	0.00	0.68	0.00	0.02	0.22
3	5.23	0.16	0.24	0.00	0.04	0.31
4	7.74	0.17	0.03	0.04	0.92	0.08
5	13.34	0.67	0.04	0.96	0.01	0.38

Authors suggest that collinearity could be present if the condition number is higher than 15. In our case, the number is 13.34. Therefore, it is possible to conclude that there is no collinearity in the model.

5.4 Outliers

Another problem that could significantly influence the outcome is highly influential observations (or outliers). To detect them the paper suggests the use of studentized Cook's residuals method. It is commonly used to estimate the influence of a data point when performing a least-squares regression analysis. (Cook, 1977)

The summary of estimated Cook's residuals could be found below (Figure 14).

Figure 14: Cook's residuals summary

stdresid2 Cook's D

type:	numeric (float)		
range:	[9.904e-12,.02799248]	units:	1.000e-19
unique values:	887	missing .:	13250/14139
mean:	.001285		
std. dev:	.00243		

percentiles:	10%	25%	50%	75%	90%
	5.1e-06	.000031	.000394	.001497	.003742

Cook suggested that the point could be considered of a high influence if the residual's values are higher than 1. In our case, the maximum value that Cook's residual achieve is around 0.03. Therefore, the paper concludes that there is no outliers' problem.

5.5 Control variables

Another necessary adjustment to consider is the country effect influence. The research is not particularly interested in location influences on the dependent variable. Therefore, effects could be removed by introducing the set of scientific constants that represent countries (Freedman, Pisani, Purves, 1998). Thirty-four dummy variables were generated for this purpose and included in the final model.

6 Final model and results

The final model is estimated using OLS regression method. Variables were transformed to control for heteroskedasticity. Robust standard error method is also used to obtain more trustworthy standard errors. The model also included 34 dummy variables to control for country effect.

The final specification takes following form:

$$secicnfa_t = \beta_0 + \beta_1 * cuse_t + \beta_2 * eturn_t + \beta_3 * ibk_t +$$

$$+\beta_4 * osopen_t + \beta_5 * digsign_t + \sum_{n=6}^{39} \beta_n * country + \varepsilon$$

Where,

- $secicnfa_t$ - dependent variable; transformed share of enterprises that experienced intrusion, pharming, phishing;

- $cuse_t$ - explanatory variable; transformed share of persons employed using computers;
- $eturn_t$ - explanatory variable; transformed share of enterprises' total turnover from e-commerce;
- ibk_t - explanatory variable; transformed share of enterprises that use the Internet for banking and financial services;
- $osopen_t$ - explanatory variable; transformed share of enterprises that use open-source operating systems;
- $digsign_t$ - explanatory variable; transformed share of enterprises that use advanced e-signatures;
- $country$ - country control dummy variables;
- β_{0-39} - regression coefficients;
- ε - error term.

The final regression table could be found below (Figure 15).

Specification could be rewritten:

$$secicnfa = .061 + -.017 * cuse + .025 * eturn + -.047 * ibk +$$

$$+.094 * osopen + .006 * digsign + \sum_{n=6}^{39} \beta_n * country$$

The final model has a significantly increased goodness of fit(0.5854). The effect could be explained by the presence of country control variables that, probably, have a high influence on the dependent variable. The paper discusses each coefficient independently below.

The rise of arcsine of the square root of a share of persons employed using computers by 1 unit causes the 0.017 decrease in phishing attacks exposure. The coefficient for $cuse_t$ is not statistically significant because its p-value of 0.163 is higher than 0.05. The result is counter-intuitive as the volume of employees using computers should, in theory, increase the risk of cyberattack exposure. However, the coefficient's insignificance could explain this fact.

The arc sinus of the square root of a share of enterprises' total turnover from e-commerce is another statistically insignificant regressor (0.05 is less than 0.117). According to the model, the rise of the regressor by 1 unit

Figure 15: Final regression table

Linear regression

Number of obs	= 889
F(26, 862)	= 75.77
Prob > F	= 0.0000
R-squared	= 0.5854
Root MSE	= .04714

secicnfa_t	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
cuse_t	-.017044	.0122085	-1.40	0.163	-.0410058	.0069178
eturn_t	.0254131	.0161823	1.57	0.117	-.0063483	.0571745
ibk_t	-.0473605	.0177683	-2.67	0.008	-.0822347	-.0124864
osopen_t	.0940013	.0201174	4.67	0.000	.0545165	.133486
digsign_t	.0065411	.0195459	0.33	0.738	-.0318221	.0449043
country_en						
CY	-.0030205	.0164896	-0.18	0.855	-.035385	.0293441
CZ	.0194929	.0113461	1.72	0.086	-.0027763	.0417622
EA	.0596081	.0113708	5.24	0.000	.0372904	.0819258
EL	-.0529238	.0224098	-2.36	0.018	-.0969079	-.0089396
ES	.0552839	.0098448	5.62	0.000	.0359612	.0746065
FR	.0436741	.0124831	3.50	0.000	.0191733	.068175
HR	.0090172	.0128685	0.70	0.484	-.0162401	.0342746
HU	-.028008	.0134007	-2.09	0.037	-.0543099	-.0017061
IE	.0466344	.015627	2.98	0.003	.0159631	.0773058
IS	-.0009507	.019565	-0.05	0.961	-.0393513	.03745
IT	.0384508	.0123804	3.11	0.002	.0141516	.0627501
LT	.1043725	.0135816	7.68	0.000	.0777157	.1310294
MK	-.0066454	.0227438	-0.29	0.770	-.0512852	.0379943
MT	.0602133	.0364259	1.65	0.099	-.0112806	.1317073
NL	.1532085	.0138288	11.08	0.000	.1260665	.1803505
NO	.0545584	.0165955	3.29	0.001	.0219861	.0871307
PL	.0395325	.0099718	3.96	0.000	.0199605	.0591044
PT	.0884603	.0126232	7.01	0.000	.0636845	.1132361
RO	.0392332	.0134605	2.91	0.004	.0128141	.0656524
SI	-.029299	.0159564	-1.84	0.067	-.060617	.002019
SK	.1589951	.0135742	11.71	0.000	.1323528	.1856374
_cons	.0619277	.0162846	3.80	0.000	.0299656	.0938898

causes the 0.025 increase in arcsine of the square root of phishing attacks exposure. This result is quite logical. The more enterprise trade online, the more profitable could be the attack for the foe as the enterprise has more assets in the information environment.

The first statistically significant regressor is arcsine of the square root of a share of enterprises that use the Internet for banking and financial services (0.05 is higher than 0.008). The rise of the transformed regressor by 1 unit causes 0.047 decrease in arcsine of the square root of phishing attacks exposure. It could be explained by the fact that enterprises that use ICT

for financial and banking purposes are more aware of cyberattack mitigation and use more sophisticated security systems.

The second statistically significant regressor is arcsine of the square root of a share of enterprises that use open-source operating systems ($0.05 > 0.000$). The rise of this regressor by 1 unit causes 0.094 increase in arcsine of the square root of phishing attacks exposure. The possible explanation is that open-source software could be developed with no security in mind, and, therefore, is more exposed to cyberattacks. The paper finds that result quite logical.

The last analysed regressor is the arcsine of the square root of a transformed share of enterprises that use advanced e-signatures. The estimator is not statistically significant (0.05 is significantly smaller than 0.738). The rise of the regressor by 1 unit causes 0.0065 increase in arcsine of the square root of phishing attacks exposure. Again, the result is counterintuitive as advanced security measures should result in a better security state. However, it could be explained by coefficient's statistical insignificance.

Interestingly, the most influential regressor is the usage of open-sourced software. In could be valuable to perform comparative analyses of closed and open software usage from security perspective in the future research.

Another point to discuss is postestimation analyses. For this purpose, it is convenient to use *predict* command to obtain predictions and residuals. Summative table for prediction could be found below.

Figure 16: Predictions summative table

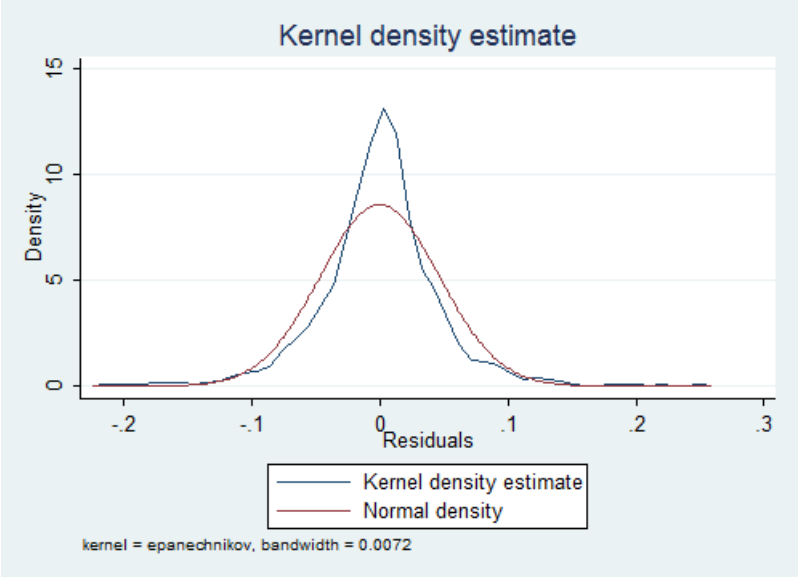
seciconfa_pr — **Fitted values**

type:	numeric (float)		
range:	[-.02583442,.24456605]	units:	1.000e-11
unique values:	1941	missing .:	12197/14139
mean:	.09052		
std. dev:	.05271		

percentiles:	10%	25%	50%	75%	90%
	.030728	.052414	.083898	.108515	.186492

The normality of estimations could be assessed by plotting kernel density of residuals over the normal distribution (Figure 17).

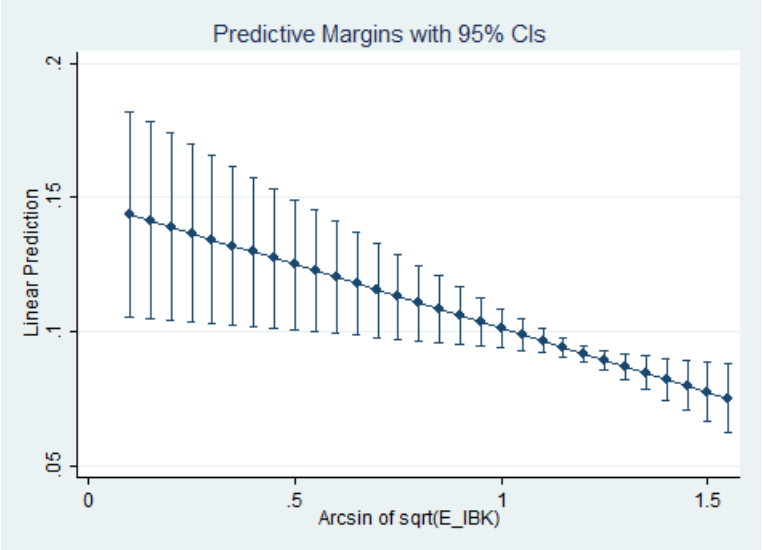
Figure 17: Kernel density plot of residuals over normal distribution



It is evident that there are some deviations in the centre part of the graph. However, the paper assumes that it is insignificant for now.

The next step is the estimation of margins of predicted variable. Margins table estimated on the base of arcsine of the square root of a share of enterprises that use the Internet for banking and financial services with 0.05 step could be found in the appendix (Appendix 5). Margins plot is presented below (Figure 18). However, the paper suggests relying on textual interpretation above.

Figure 18: Predicted margins plot



7 Conclusion

The paper attempted to answer the question:

What are the drivers of phishing attack exposure?

Using regression analyses methods. It introduced and tested five hypotheses and rejected three of them. The research confirmed that:

1. Enterprises' internet usage for banking and financial purposes has a direct influence on the enterprises' phishing attack exposure;
2. Enterprises' open-source software usage has a direct influence on the enterprises' phishing attack exposure.

The paper also presented detailed methodology and verification description. It should be considered as one done for study purposes.

References

1. Belsley, D. A., Kuh, E., Welsch, R. E. (1980). Regression Diagnostics. Hoboken, NJ, USA: John Wiley Sons, Inc. https://doi.org/10.1002/0471725153

2. Bjorck, A. (1996). Numerical methods for least squares problems. Philadelphia: SIAM.

3. Cleveland, W. S. (1981). LOWESS: A Program for Smoothing Scatterplots by Robust Locally Weighted Regression. The American Statistician, 35(1), 54. https://doi.org/10.2307/2683591

4. Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. Technometrics, 19(1), 15. https://doi.org/10.2307/1268249

5. COOK, R. D., WEISBERG, S. (1983). Diagnostics for heteroscedasticity in regression. Biometrika, 70(1), 110. https://doi.org/10.1093/biomet/70.1.1

6. Freedman, D., Pisani, R., Purves, R. (1998). Statistics. W.W. Norton

7. White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica, 48(4), 817. https://doi.org/10.2307/1912934

Appendix

Appendix 1. Countries' table of frequencies

```
. tab country
```

ExpCountry	Freq.	Percent	Cum.
AT	403	2.60	2.60
BE	414	2.67	5.27
BG	397	2.56	7.82
CY	409	2.64	10.46
CZ	429	2.76	13.23
DE	671	4.32	17.55
DK	391	2.52	20.07
EA	396	2.55	22.62
EA16	52	0.34	22.96
EE	444	2.86	25.82
EL	437	2.82	28.63
ES	599	3.86	32.50
EU15	393	2.53	35.03
EU25	140	0.90	35.93
EU27	396	2.55	38.48
EU28	396	2.55	41.04
FI	331	2.13	43.17
FR	395	2.55	45.71
HR	373	2.40	48.12
HU	467	3.01	51.13
IE	397	2.56	53.69
IS	241	1.55	55.24
IT	393	2.53	57.77
LT	400	2.58	60.35
LU	391	2.52	62.87
LV	391	2.52	65.39
MK	374	2.41	67.80
MT	461	2.97	70.77
NL	442	2.85	73.62
NO	456	2.94	76.56
PL	491	3.16	79.72
PT	656	4.23	83.95
RO	421	2.71	86.67
RS	48	0.31	86.97
SE	411	2.65	89.62
SI	453	2.92	92.54
SK	678	4.37	96.91
TR	69	0.44	97.36
UK	410	2.64	100.00
Total	15,516	100.00	

Appendix 2. Dependent variables' content

P_CUSE — Share of persons employed using computers

type:	numeric (double)		
range:	[0,1]	units:	1.000e-06
unique values:	10956	missing .:	2877/14139
mean:	.55985		
std. dev:	.237055		

percentiles:	10%	25%	50%	75%	90%
	.273747	.37917	.521162	.72048	.946637

E_ETURN — Share of enterprises' total turnover from e-commerce

type:	numeric (double)		
range:	[0,.886114]	units:	1.000e-06
unique values:	8465	missing .:	5381/14139
mean:	.117085		
std. dev:	.11238		

percentiles:	10%	25%	50%	75%	90%
	.009497	.035041	.090343	.164615	.250339

E_IBK — Share of enterprises that use the Internet for banking and financial services

type:	numeric (double)		
range:	[0,1]	units:	1.000e-06
unique values:	3059	missing .:	10863/14139
mean:	.819968		
std. dev:	.156288		

percentiles:	10%	25%	50%	75%	90%
	.59195	.768599	.864001	.926986	.96668

E_OSOPEN — Share of enterprises that use open source operating systems

type:	numeric (double)		
range:	[0,1]	units:	1.000e-06
unique values:	3641	missing .:	10352/14139
mean:	.214911		
std. dev:	.167758		

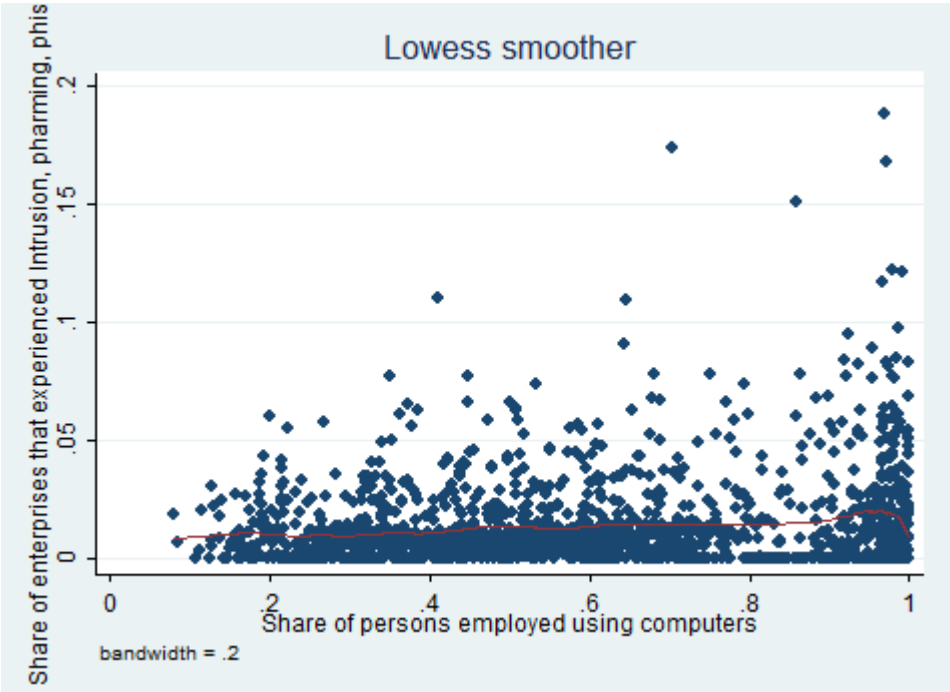
percentiles:	10%	25%	50%	75%	90%
	.061194	.101235	.159572	.274597	.487394

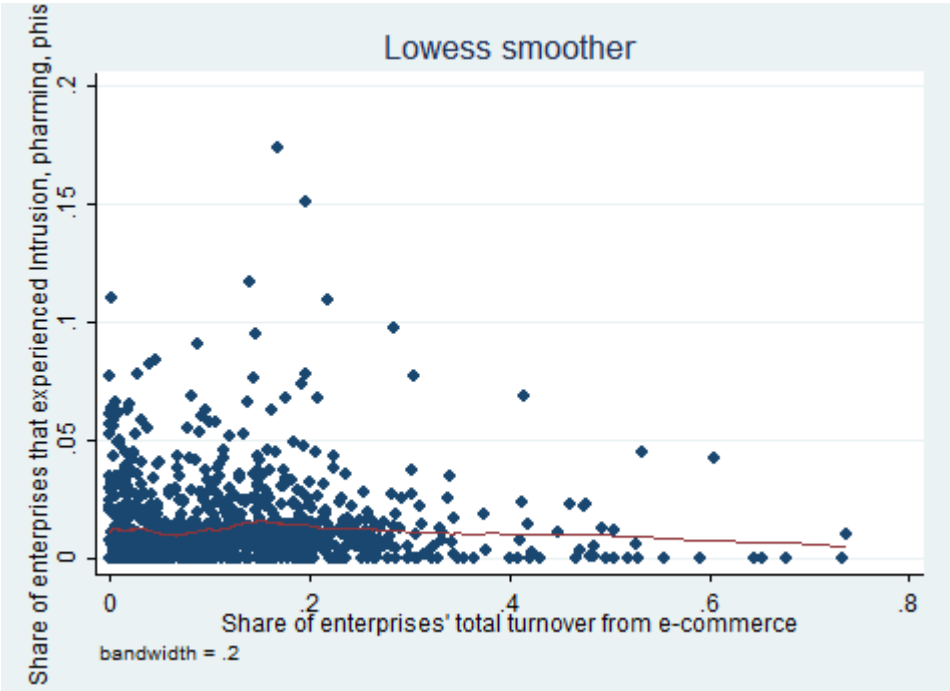
E_DIGSIGN — Share of enterprises that use advanced e-signatures in relations with suppliers

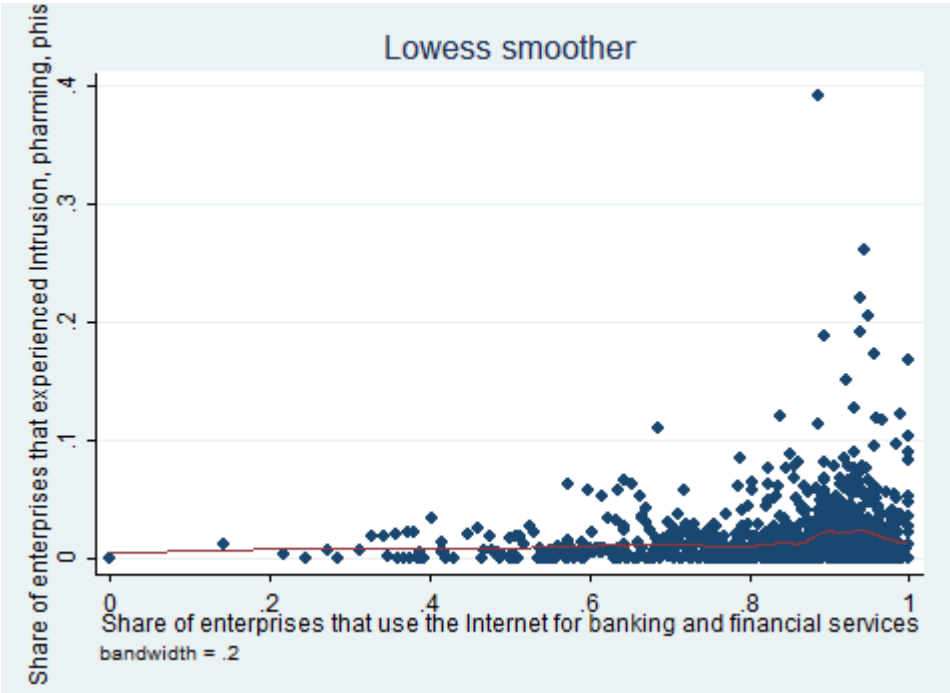
type:	numeric (double)		
range:	[0,1]	units:	1.000e-06
unique values:	3675	missing .:	10308/14139
mean:	.319906		
std. dev:	.231628		

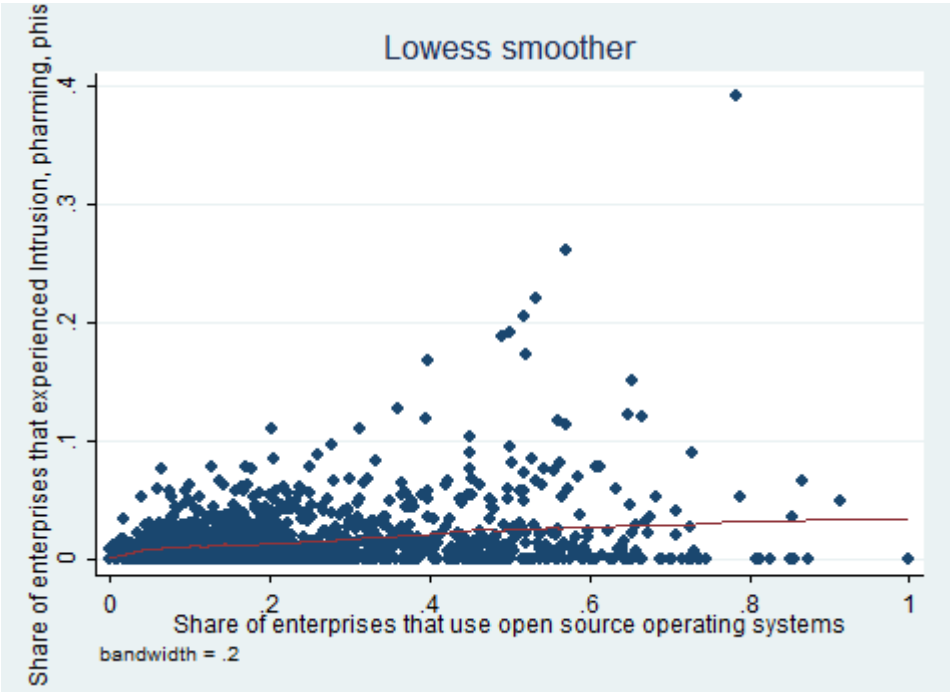
percentiles:	10%	25%	50%	75%	90%
	.082796	.149234	.26122	.430039	.681818

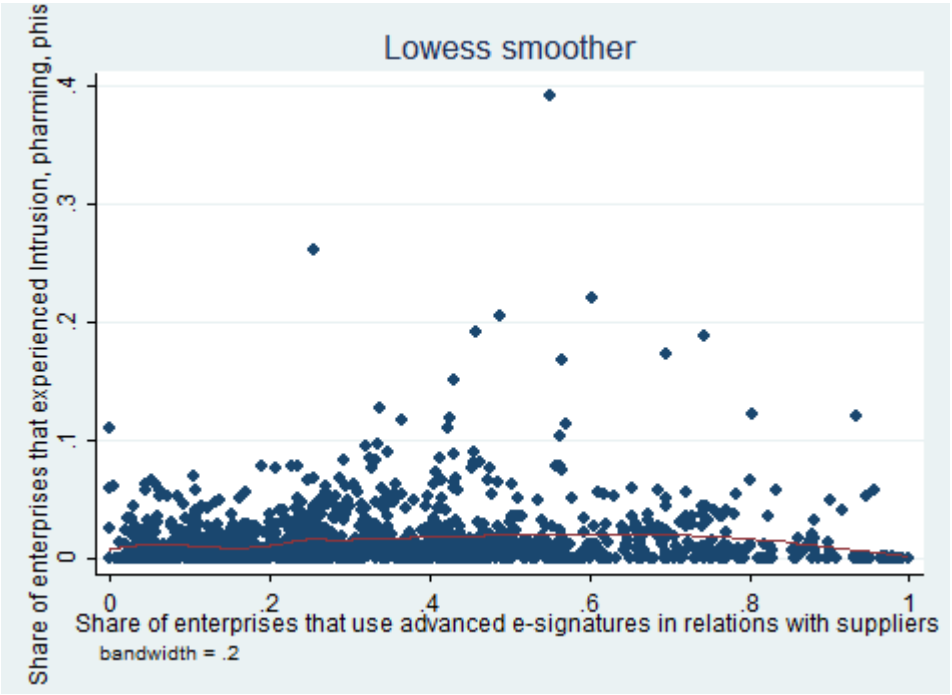
Appendix 3. Original variables scatter plots (by LOWESS smoothing)






Lowess smoother
Share of enterprises that experienced Intrusion, pharming, phis
0
.1
.2
.3
.4
0
.2
.4
.6
.8
1
Share of enterprises that use the Internet for banking and financial services
bandwidth = .2


Lowess smoother
Share of enterprises that experienced Intrusion, pharming, phis
0
.1
.2
.3
.4
0
.2
.4
.6
.8
1
Share of enterprises that use open source operating systems
bandwidth = .2


Lowess smoother
Share of enterprises that experienced Intrusion, pharming, phis
0
.1
.2
.3
.4
0
.2
.4
.6
.8
1
Share of enterprises that use advanced e-signatures in relations with suppliers
bandwidth = .2

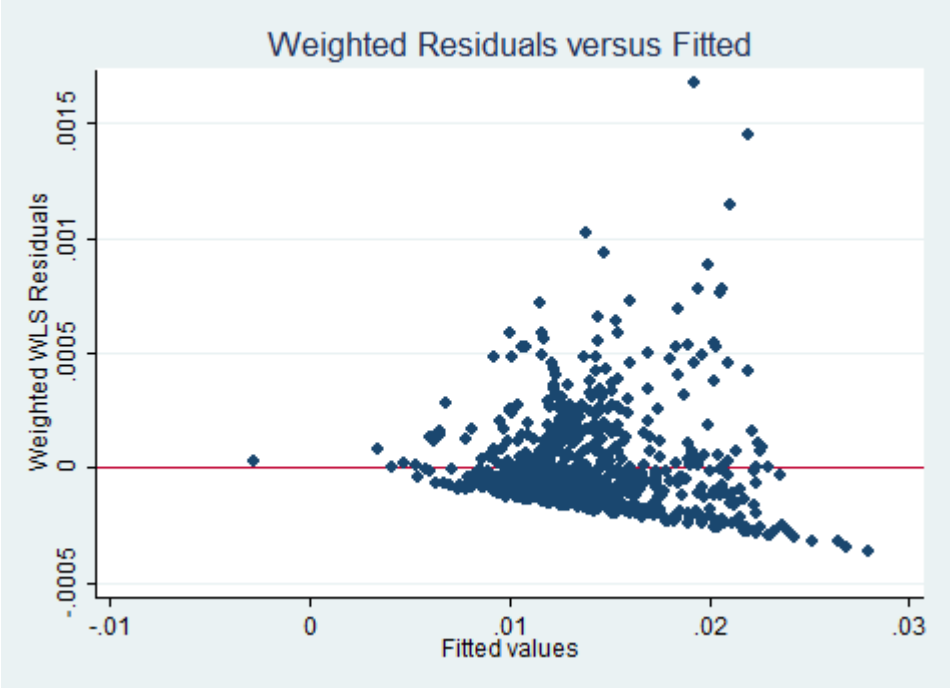
Appendix 4. WLS regression table and weighted residuals plot

WLS regression - type: proportional to log(e^2)

(sum of wgt is 9.3940e+00)

Source	SS	df	MS		
Model	.011422126	5	.002284425	Number of obs	= 889
Residual	.297107569	883	.000336475	F(5, 883)	= 6.79
				Prob > F	= 0.0000
Total	.308529695	888	.000347443	R-squared	= 0.0370
				Adj R-squared	= 0.0316
				Root MSE	= .01834

E_SECICNFA	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
P_CUSE	-.002107	.0030799	-0.68	0.494	-.0081518	.0039377
E_ETURN	-.0060576	.0062156	-0.97	0.330	-.0182566	.0061414
E_IBK	.0119251	.0055961	2.13	0.033	.0009419	.0229083
E_OSOPEN	.0200547	.0046275	4.33	0.000	.0109725	.0291369
E_DIGSIGN	-.003673	.0031328	-1.17	0.241	-.0098216	.0024756
_cons	.0024474	.0038216	0.64	0.522	-.0050529	.0099478



Appendix 5. Margins table

	Delta-method Margin	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_at						
1	.1438807	.0195084	7.38	0.000	.1055912	.1821703
2	.1415127	.018623	7.60	0.000	.104961	.1780644
3	.1391447	.0177379	7.84	0.000	.1043302	.1739591
4	.1367767	.0168531	8.12	0.000	.1036988	.1698545
5	.1344086	.0159687	8.42	0.000	.1030666	.1657507
6	.1320406	.0150848	8.75	0.000	.1024334	.1616478
7	.1296726	.0142014	9.13	0.000	.1017992	.157546
8	.1273045	.0133187	9.56	0.000	.1011636	.1534455
9	.1249365	.0124369	10.05	0.000	.1005264	.1493466
10	.1225685	.011556	10.61	0.000	.0998872	.1452497
11	.1202005	.0106764	11.26	0.000	.0992456	.1411553
12	.1178324	.0097984	12.03	0.000	.0986009	.137064
13	.1154644	.0089225	12.94	0.000	.0979521	.1329767
14	.1130964	.0080492	14.05	0.000	.097298	.1288948
15	.1107283	.0071798	15.42	0.000	.0966365	.1248202
16	.1083603	.0063155	17.16	0.000	.0959647	.1207559
17	.1059923	.0054591	19.42	0.000	.0952776	.1167069
18	.1036243	.0046147	22.46	0.000	.0945668	.1126817
19	.1012562	.0037906	26.71	0.000	.0938164	.1086961
20	.0988882	.0030033	32.93	0.000	.0929936	.1047829
21	.0965202	.0022912	42.13	0.000	.0920231	.1010172
22	.0941522	.0017488	53.84	0.000	.0907198	.0975845
23	.0917841	.0015637	58.70	0.000	.088715	.0948533
24	.0894161	.0018468	48.42	0.000	.0857913	.0930409
25	.0870481	.0024403	35.67	0.000	.0822585	.0918376
26	.08468	.0031745	26.67	0.000	.0784494	.0909107
27	.082312	.0039723	20.72	0.000	.0745156	.0901084
28	.079944	.0048019	16.65	0.000	.0705191	.0893689
29	.077576	.0056495	13.73	0.000	.0664876	.0886644
30	.0752079	.006508	11.56	0.000	.0624346	.0879813