Does Homework Predict the Final Grade?

Eric W. Overholser

July 22, 2010

Contents

1 Introduction

A student can usually expect homework, quizzes, labs, and exams to determine their final grade in a course. Also, they can expect to hear from their professor, "If you do well on your homework, you will do well in this course!" Using data obtained from four classes taught by the same professor, we will determine the most likely factors behind the final grade in the course.

First, we will regress the final course grade on the exlanatory variables gender, student class, homework average, and concept test. The homework average and concept test affect the final course grade equally for this course. Second, we will determine if any of the variables can be removed to arrive at a simpler linear model. This will be done at the 90% confidence level considering the t-statistic along with other characteristics such as multi-collinearity between the remaining variables. Last, we will make a comparison of the adjusted R-squared for each regression to determine the most favorable regression model.

2 The (2,1) Interactive Model

Consider the regression model

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \gamma_1 D_{1j} + \delta_{11} X_{1j} D_{1j} + \delta_{12} X_{2j} D_{1_j} + \epsilon_j$$

where

Variable	Description
Y_j	Logit of Final Course Grade
X_1	Logit of Homework Average
X_2	Logit of Concept Test Average
D_1	Gender

The linear estimates for the coefficients will be A, B_1, B_2 for α, β_1, β_2, respectively. We transformed each of the variables X_1, X_2, and Y using the logit function since each of the variables are bounded in $(0, 100)$. The coefficients and other statistical information for the (2,1) Interactive Model is shown in Figure 2.1.

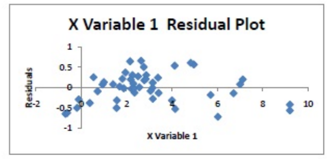
SUMMARY OUTPUT

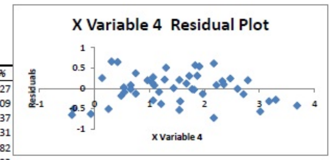
Regression Statistics	
Multiple R	0.902273001
R Square	0.814096568
Adjusted R Square	0.791965207
Standard Error	0.382560842
Observations	48

ANOVA

	df	SS	MS	F	Significance F
Regression	5	26.91775494	5.383551	36.7847494	2.66123E-14
Residual	42	6.146817504	0.146353		
Total	47	33.06457244			

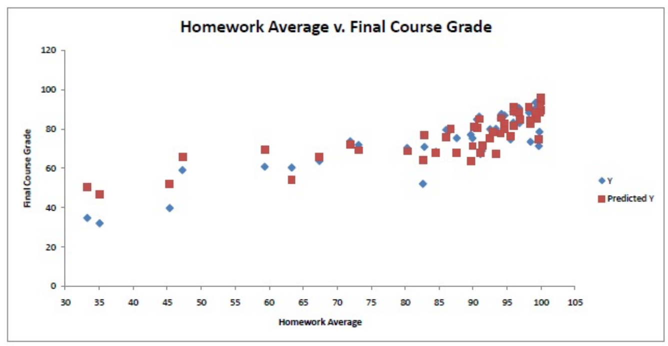
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.299550945	0.135377118	2.212715	0.032406973	0.026348862	0.572753027	0.026348862	0.572753027
Logit(HW Avg)	0.10027646	0.054241288	1.848711	0.071547356	-0.00918689	0.209739809	-0.00918689	0.209739809
Gender	-0.298698832	0.210306587	-1.420302	0.162901081	-0.723114702	0.125717037	-0.723114702	0.125717037
Logit(HW Avg)xGender	0.036825672	0.072140023	0.510475	0.612391446	-0.108758788	0.182410131	-0.108758788	0.182410131
Logit(Concept)	0.5131557	0.114076394	4.498351	5.32847E-05	0.282940219	0.743371182	0.282940219	0.743371182
Logit(Concept)xGender	0.118371979	0.171806433	0.688985	0.494621756	-0.228347435	0.465091393	-0.228347435	0.465091393





RESIDUAL OUTPUT

Observation	Predicted Logit(Y)	Residuals	Fit Y
1	1.138152415	0.217520638	75.73403
2	0.087075405	-0.501270405	52.17551
3	0.922338677	-0.127240306	71.55184
4	2.076770647	-0.00906999	88.86248
5	0.165674833	0.255010457	54.13242
6	2.7413998	0.201753374	93.94258
7	1.381391865	-0.018132687	79.92144
8	1.546821448	-0.529103845	82.44542
9	0.660929022	-0.29419551	65.9469
10	0.021601595	-0.650729393	50.54002
11	0.780827719	0.081994694	68.58585
12	2.142768481	0.081443446	89.49911
13	0.725571992	0.661139088	67.38328
14	2.331983818	-0.315714503	91.14915
15	0.652801404	-0.091918432	65.76415
16	1.772967162	0.179613625	85.48263
17	1.576313118	0.310512641	82.86817
18	2.026959383	0.244333587	88.35987
19	1.36809682	0.209539089	79.70725
20	0.588895112	-0.504207766	64.31116
21	1.475627668	0.121421458	81.39113
22	2.310326975	-0.277695684	90.97287
23	1.738712479	0.098054619	85.05235
24	1.412061098	0.307111432	80.40908
25	0.564624991	0.649341379	63.7522
26	1.100141314	0.277494984	75.02866
27	1.714279233	-0.134291521	84.73905
28	0.95463302	0.074481505	72.2046
29	1.451421283	-0.029582517	81.02171
30	2.019177428	-0.722336847	88.27959
31	1.164101961	-0.084527826	76.20773
32	0.745321808	0.025326463	67.81585
33	1.203259188	-0.31486667	76.91041
34	0.738120992	-0.019016232	67.65848
35	3.16887333	-0.566071756	95.9646
36	-0.122965779	-0.628190091	46.92972
37	0.744799952	0.372881161	67.80446
38	0.911175358	0.199675085	71.32406
39	2.365922421	-0.140428168	91.41915
40	1.092940839	-0.181641818	74.89351
41	0.806134821	0.134303221	69.12852
42	1.297182606	0.006508624	78.53604
43	1.738917088	0.570620808	85.05495
44	1.245999503	0.509339964	77.66066
45	0.8197329	-0.381682239	69.41796
46	3.114923173	-0.422177516	95.75041
47	2.035627982	0.613022708	88.44873
48	1.693779869	0.541647671	84.47206



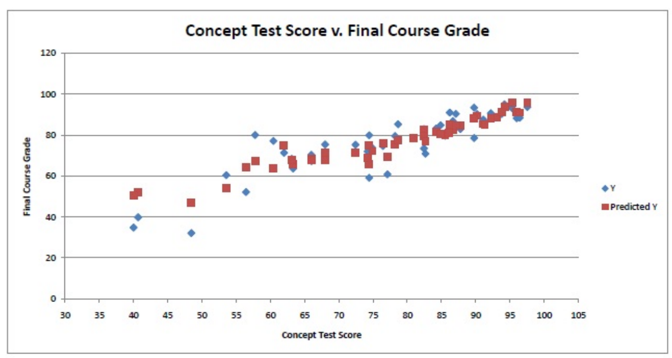


Figure 2.1: The (2,1) Interactive Model

Figure 2.1 shows that the concept tests are the most significant explanatory variable predicting the final course grade. This model proposes that for every point you earn on the concept test predicts that you will earn a higher final course grade compared to every point from the homework average. A possible explaination for this is the dedication given to a homework assignment compared to a concept test. There were relatively few concept tests given as compared to many homework assignments assigned throughout the semester. With 81.4% of the data explaned by these variables (79.2% using the adjusted R-square) this model seems to be a fairly reliable predictor of the final course grade.

The interactive coefficients δ_{11} and δ_{12} have high P-values. These interactive terms lean towards not being significant. So our next model will remove these interactive terms in an attempt to improve the model.

3 The (2,1) Model

Consider the regression model

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \gamma_1 D_{1j} + \epsilon_j$$

using the same variable definitions as with the (2,1) Interactive Model. Figure 3.1 shows the regression statistics for the model. Again, the concept tests are a powerful predictor of the final course grade. How does this model compare to the prior model. The R-squared has slightly dropped, and the adjusted R-squared has also dropped. This indicates that we have a little explanatory power, but it is preferred to the prior model if one wants to insure that more explanatory variables are significant at the 90% level.

The gender dummy variable is the only variable with a coefficient that is not at the 90% confidence level. This indicates that gender was not a factor in the final course grade. We will remove this dummy variable for our next model.

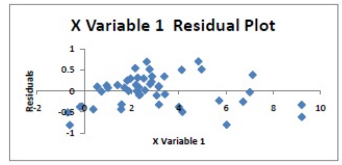
SUMMARY OUTPUT

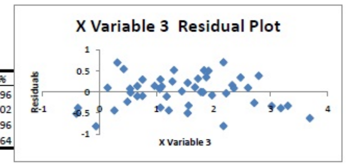
Regression Statistics	
Multiple R	0.896073603
R Square	0.802947903
Adjusted R Square	0.789512533
Standard Error	0.384809379
Observations	48

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	26.5491291	8.84971	59.7637346	1.46685E-15
Residual	44	6.515443344	0.148078		
Total	47	33.06457244			

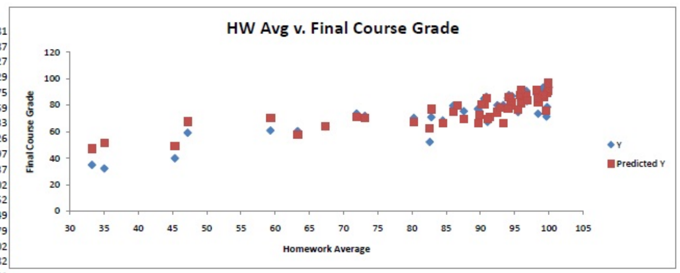
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.190820985	0.117441611	1.624816	0.111344706	-0.045867027	0.427508996	-0.045867027	0.427508996
Logit(HW Avg)	0.123254428	0.035941868	3.429272	0.001325509	0.050818353	0.195690502	0.050818353	0.195690502
Gender	-0.01789922	0.111289088	-0.160835	0.872959271	-0.242187635	0.206389196	-0.242187635	0.206389196
Logit(Concept)	0.539115652	0.083794201	6.433806	7.76299E-08	0.37023954	0.707991764	0.37023954	0.707991764





RESIDUAL OUTPUT

Observation	*Predicted Y*	*Residuals*	*Fit Y*
1	1.104423745	0.251249309	75.10881
2	-0.035728997	-0.378466003	49.10687
3	0.887251759	-0.092153388	70.83227
4	1.964628021	0.103072636	87.70329
5	0.317018796	0.103666495	57.85975
6	2.556753133	0.38640004	92.80259
7	1.363101659	0.000157518	79.62633
8	1.51846652	-0.500748918	82.03126
9	0.735722225	-0.368988713	67.60597
10	-0.113667554	-0.515460244	47.16137
11	0.72160436	0.141218054	67.29602
12	2.252839163	-0.028627236	90.48952
13	0.685723402	0.700987678	66.50149
14	2.397780272	-0.381510957	91.66579
15	0.574938886	-0.014055914	63.9902
16	1.788621759	0.163959028	85.67582
17	1.535674644	0.351151115	82.28351
18	1.923596496	0.347696474	87.2539
19	1.357793586	0.219842323	79.54009
20	0.522715562	-0.438028217	62.77825
21	1.482927936	0.11412119	81.50144
22	2.359945432	-0.327314141	91.37215
23	1.743974723	0.092792376	85.11912
24	1.400117132	0.319055398	80.22025
25	0.667614056	0.546352314	66.09687
26	1.076895371	0.300740927	74.5906
27	1.656850342	-0.076862629	83.98147
28	0.895891457	0.133223067	71.01045
29	1.425014954	-0.003176187	80.61234
30	2.105407037	-0.808566456	89.14276
31	1.18498155	-0.105407416	76.58423
32	0.689535697	0.081112574	66.58636
33	1.209163301	-0.320770783	77.01509
34	0.816736976	-0.097632217	69.35433
35	2.934601271	-0.331799696	94.95306
36	0.063391577	-0.814547446	51.58426
37	0.820151336	0.297529776	69.42685
38	0.963629495	0.147220948	72.38479
39	2.48212585	-0.256631596	92.28792
40	1.140692657	-0.229393636	75.78068
41	0.86659525	0.073842792	70.40367
42	1.28485506	0.018836171	78.32751
43	1.791978975	0.517558922	85.71697
44	1.234575144	0.520764324	77.46183
45	0.874324324	-0.436273663	70.56447
46	3.313529887	-0.62078423	96.48901
47	1.939981818	0.708668872	87.43501
48	1.729448176	0.505979364	84.93418



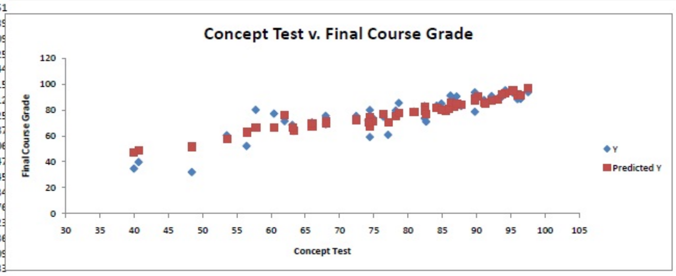


Figure 3.1: The (2,1) Model

4 The (2,0) Model

Consider the regression model

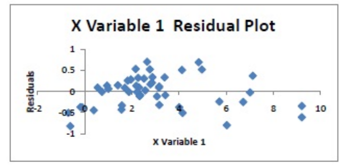
$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \epsilon_j$$

using the same variable definitions as with the (2,1) Interactive Model. Figure 4.1 shows the regression statistics for this regression model. Just as with the previous two models, the concept tests are a powerful predictor of the final course grade. Comparing the coefficients of the transformed variables, the transformed concept test coefficient is five times the transformed homework average coefficient. The R-squared and adjusted R-squared have only slightly dropped (by a smaller amount than we saw from the (2,1) Interactive Model to the (2,1) Model). Although, we still have above 80% explanation of our data. Notice that the coefficients are significant at the 90% significance level.

Since the concept test coefficient is much more significant than the homework average cofficient, we will take a look at one last regression model with only the concept test average as the explanatory variable.

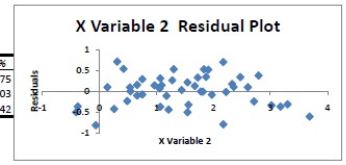
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.896008959
R Square	0.802832054
Adjusted R Square	0.794069034
Standard Error	0.380621534
Observations	48



ANOVA

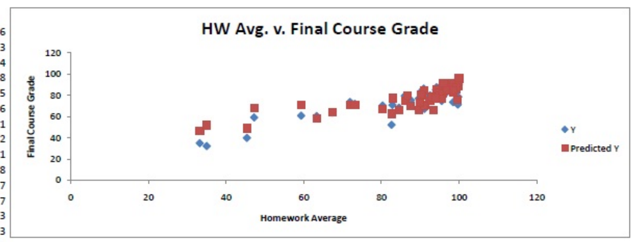
	df	SS	MS	F	Significance F
Regression	2	26.54529861	13.27265	91.61591213	1.36087E-16
Residual	45	6.519273834	0.144873		
Total	47	33.06457244			



	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.181701017	0.101726337	1.786175	0.080809837	-0.023186341	0.386588375	-0.023186341	0.386588375
X Variable 1	0.123412442	0.035537432	3.472745	0.001149938	0.05183638	0.194988503	0.05183638	0.194988503
X Variable 2	0.539169189	0.082881622	6.505292	5.53627E-08	0.372237035	0.706101342	0.372237035	0.706101342

RESIDUAL OUTPUT

Observation	Predicted Y	Residuals	Fit Y
1	1.09565973	0.260013323	74.9446
2	-0.044898537	-0.369296463	48.87773
3	0.878544007	-0.083445636	70.65204
4	1.974043984	0.093656673	87.80448
5	0.325891945	0.094793345	58.07595
6	2.566808494	0.376344679	92.86946
7	1.372272509	-0.009013332	79.77471
8	1.527984542	-0.510266939	82.17112
9	0.744541255	-0.377807744	67.79881
10	-0.122919349	-0.506208449	46.93088
11	0.712741936	0.150080478	67.10067
12	2.244946162	-0.020734236	90.42137
13	0.677037912	0.709673168	66.30773
14	2.389463274	-0.373193959	91.60203
15	0.565962877	-0.005079905	63.78311
16	1.780066281	0.172514506	85.57051
17	1.545007471	0.341818289	82.41915
18	1.933043354	0.338249616	87.35859
19	1.367105623	0.210530286	79.69121
20	0.513855805	-0.429168459	62.57099
21	1.47439655	0.122652576	81.37247
22	2.351501884	-0.318870593	91.30535
23	1.735344311	0.101422787	85.00947
24	1.391447026	0.327725503	80.08232
25	0.676757149	0.537209221	66.30145
26	1.068229624	0.309406675	74.42601
27	1.666274915	-0.086287202	84.10785
28	0.886978678	0.142135846	70.82663
29	1.434241947	-0.01240318	80.75614
30	2.097356197	-0.800515616	89.06459
31	1.194307672	-0.114733538	76.75106
32	0.680712697	0.089935574	66.38977
33	1.218274953	-0.329882435	77.17598
34	0.825918681	-0.106813921	69.54913
35	2.944997384	-0.34219581	95.00265
36	0.072070096	-0.823225966	51.80097
37	0.829279728	0.288401384	69.62026
38	0.972806114	0.138044329	72.56785
39	2.474213638	-0.248719385	92.23142
40	1.15040046	-0.239101439	75.95841
41	0.875589489	0.064848553	70.59074
42	1.2762177	0.02747353	78.18053
43	1.783742406	0.525795491	85.61584
44	1.225961225	0.529378242	77.31109
45	0.883228337	-0.445177675	70.74908
46	3.306062632	-0.613316976	96.46362
47	1.949641068	0.699009622	87.54075
48	1.721082379	0.514345162	84.82682



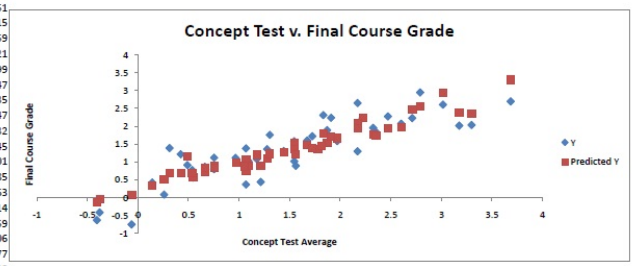


Figure 4.1: The (2,0) Model

5 The (1,0) Model

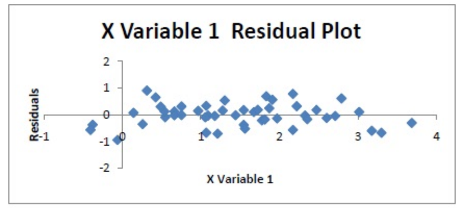
Consider the regression model

$$Y_j = \alpha + \beta_1 X_{1j} + \epsilon_j$$

using the same variable definitions as with the (2,1) Interactive Model. Figure 5.1 gives the regression statistics determined for this model. This model shows that only 75% of the data is explained as compared to over 80% with the other models. But all of the coefficients are significant according to their respective t-values.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.866020339
R Square	0.749991228
Adjusted R Square	0.744556255
Standard Error	0.423916346
Observations	48



ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	24.79813929	24.79814	137.9935441	1.90843E-15
Residual	46	8.266433149	0.179705		
Total	47	33.06457244			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.237730331	0.111863417	2.125184	0.038973882	0.012560954	0.462899708	0.012560954	0.462899708
X Variable 1	0.747644002	0.063645173	11.74707	1.90843E-15	0.619532916	0.875755088	0.619532916	0.875755088

RESIDUAL OUTPUT

Observation	*Logit(Y)*	*Residuals*	*Fit Y*
1	1.193715574	0.161957479	76.74049
2	-0.04470154	-0.36949346	48.88265
3	0.801283298	-0.006184927	69.02489
4	2.184508372	-0.116807715	89.88497
5	0.3442417	0.07644359	58.52205
6	2.324880927	0.618272246	91.09168
7	1.567681398	-0.20442222	82.74528
8	1.394151838	-0.376434236	80.12542
9	1.037110104	-0.670376593	73.8292
10	-0.065413225	-0.563714573	48.36525
11	0.733638274	0.12918414	67.56032
12	1.898740971	0.325470956	86.9749
13	0.472234523	0.914476556	61.59125
14	2.613783215	-0.5975139	93.17434
15	0.646350455	-0.085467483	65.61876
16	1.977705548	-0.025124761	87.84364
17	1.637172001	0.249653758	83.71498
18	2.086601955	0.184691015	88.95941
19	1.394151838	0.183484071	80.12542
20	0.431534316	-0.34684697	60.624
21	1.489926619	0.107122507	81.60673
22	2.705296321	-0.67266503	93.73384
23	1.998455557	-0.161688458	88.06348
24	1.528098761	0.191073769	82.1728
25	0.554745231	0.659221139	63.52358
26	1.037110104	0.340526194	73.8292
27	1.711749792	-0.131762079	84.70631
28	1.054677277	-0.025562753	74.16721
29	1.594920986	-0.173082219	83.13073
30	1.862191392	-0.565350811	86.55522
31	1.117858766	-0.038284632	75.35913
32	0.639205029	0.131443242	65.45737
33	1.405688597	-0.517296079	80.3085
34	0.733638274	-0.014533515	67.56032
35	2.493323299	0.109478275	92.36724
36	0.191195242	-0.942351111	54.76537
37	0.801283298	0.316397815	69.02489
38	0.960414848	0.150435595	72.32049
39	2.265939581	-0.040445327	90.60166
40	0.603738125	0.307560896	64.65111
41	1.028401845	-0.087963803	73.66059
42	1.31643601	-0.01274478	78.85881
43	1.60881295	0.700724946	83.32465
44	1.213367416	0.541972051	77.08942
45	1.145815589	-0.707764928	75.87458
46	2.993997805	-0.301252149	95.23022
47	1.862191392	0.786459298	86.55522
48	1.666342567	0.569084973	84.10876

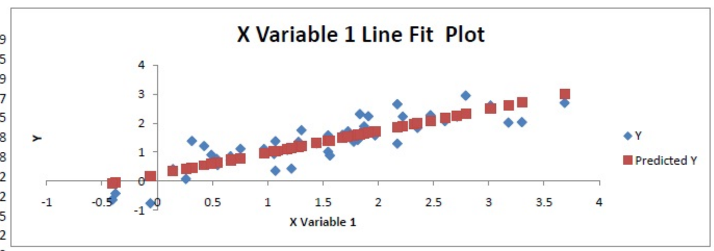


Figure 5.1: The (1,0) Model

6 The Selected Regression Model

Based on the analysis of this study, we have selected the (2,0) Model

$$\text{Logit}(Y_j) = 0.181701 + 0.1234124\text{Logit}(X_1) + 0.5391692\text{Logit}(X_2)$$

Although this model does not have the highest explanatory power (i.e. the highest R-square), it does have some benefits. One benefit is that all of the coefficients are significant at the 90% level. Further, the model explains just over 80% of the data. This slightly lower than the more complicated models, and significantly higher than the single variable model.

7 Conclusion

The perfered model that predicts the final course grade is given by

$$\text{Logit}(Y_j) = 0.181701 + 0.1234124\text{Logit}(X_1) + 0.5391692\text{Logit}(X_2)$$

where Y_j is the predicted final course grade, X_1 is the homework average, and X_2 is the concept test average. This model proposes that an increase in concept test average or homework average will increase your final course grade, but the concept test average more than the homework average. Using the Logit function can rewrite the above equation as

$$\left(\frac{Y_j}{1-Y_j}\right) = e^{0.182} \cdot \left(\frac{X_j}{1-X_j}\right)^{0.123} \cdot \left(\frac{X_2}{1-X_2}\right)^{0.539}.$$