Smart Alex's Answers

Chapter 7

Task 1

- A fashion student was interested in factors that predicted the salaries of catwalk models. She collected data from 231 models. For each model she asked them their salary per day on days when they were working (**salary**), their age (**age**), how many years they had worked as a model (**years**), and then got a panel of experts from modelling agencies to rate the attractiveness of each model as a percentage with 100% being perfectly attractive (**beauty**). The data are in the file **Supermodel.sav**. Unfortunately, this fashion student bought some substandard statistics text and so doesn't know how to analyse her data.☺ Can you help her out by conducting a multiple regression to see which factor predict a model's salary? How valid is the regression model?

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.429[a]	.184	.173	14.57213	.184	17.066	3	227	.000	2.057

a. Predictors: (Constant), Attractiveness (%), Number of Years as a Model, Age (Years)

b. Dependent Variable: Salary per Day (£)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10871.964	3	3623.988	17.066	.000[a]
	Residual	48202.790	227	212.347		
	Total	59074.754	230			

a. Predictors: (Constant), Attractiveness (%), Number of Years as a Model, Age (Years)

b. Dependent Variable: Salary per Day (£)

To begin with, a sample size of 231 with three predictors seems reasonable because this would easily detect medium to large effects (see the diagram in the chapter).

Overall, the model accounts for 18.4% of the variance in salaries and is a significant fit of the data (F(3, 227) = 17.07, p < .001). The adjusted R^2 (.17) shows some shrinkage from the unadjusted value (.184) indicating that the model may not generalize well. We can also use Stein's formula:

$$\begin{aligned}\text{adjusted } R^2 &= 1-\left[\left(\frac{231-1}{231-3-1}\right)\left(\frac{231-2}{231-3-2}\right)\left(\frac{231+1}{231}\right)\right](1-0.184)\\ &= 1-[1.031](0.816)\\ &= 1-0.841\\ &= 0.159\end{aligned}$$

This also shows that the model may not cross-generalize well.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-60.890	16.497		-3.691	.000	-93.396	-28.384		
	Age (Years)	6.234	1.411	.942	4.418	.000	3.454	9.015	.079	12.653
	Number of Years as a Model	-5.561	2.122	-.548	-2.621	.009	-9.743	-1.380	.082	12.157
	Attractiveness (%)	-.196	.152	-.083	-1.289	.199	-.497	.104	.867	1.153

a. Dependent Variable: Salary per Day (£)

In terms of the individual predictors we could report:

	B	SE B	β
Constant	–60.89	16.50	
Age	6.23	1.41	.94**
Years as a model	–5.56	2.12	–.55*
Attractiveness	–0.20	0.15	–.08

Note: R^2 = .18 (p < .001). * p < .01, ** p < .001.

It seems as though salaries are significantly predicted by the age of the model. This is a positive relationship (look at the sign of the beta), indicating that as age increases, salaries increase too. The number of years spent as a model also seems to significantly predict salaries, but this is a negative relationship indicating that the more years you've spent as a model, the lower your salary. This finding seems very counter-intuitive, but we'll come back to it later. Finally, the attractiveness of the model doesn't seem to predict salaries.

If we wanted to write the regression model, we could write it as:

$$\begin{aligned}\text{Salary} &= \beta_0 + \beta_1\text{Age}_i + \beta_2\text{Experience}_i + \beta_2\text{Attractiveness}_i \\ &= -60.89 + (6.23\text{Age}_i) - (5.56\text{Experience}_i)(0.02\text{Attractiveness}_i)\end{aligned}$$

The next part of the question asks whether this model is valid.

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	Age (Years)	Number of Years as a Model	Attractiveness (%)
1	1	3.925	1.000	.00	.00	.00	.00
	2	.070	7.479	.01	.00	.08	.02
	3	.004	30.758	.30	.02	.01	.94
	4	.001	63.344	.69	.98	.91	.04

a. Dependent Variable: Salary per Day (£)

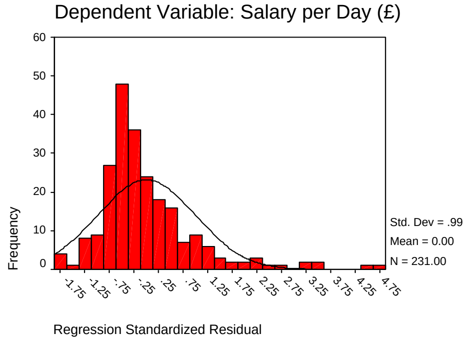
Casewise Diagnostics[a]

Case Number	Std. Residual	Salary per Day (£)	Predicted Value	Residual
2	2.186	53.72	21.8716	31.8532
5	4.603	95.34	28.2647	67.0734
24	2.232	48.87	16.3444	32.5232
41	2.411	51.03	15.8861	35.1390
91	2.062	56.83	26.7856	30.0459
116	3.422	64.79	14.9259	49.8654
127	2.753	61.32	21.2059	40.1129
135	4.672	89.98	21.8946	68.0854
155	3.257	74.86	27.4025	47.4582
170	2.170	54.57	22.9401	31.6254
191	3.153	50.66	4.7164	45.9394
198	3.510	71.32	20.1729	51.1478

a. Dependent Variable: Salary per Day (£)

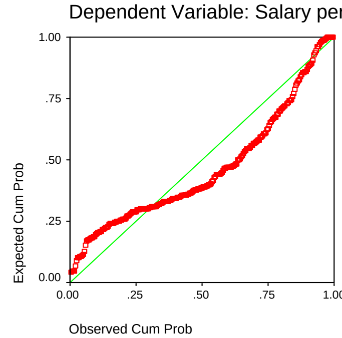
Histogram

Dependent Variable: Salary per Day (£)



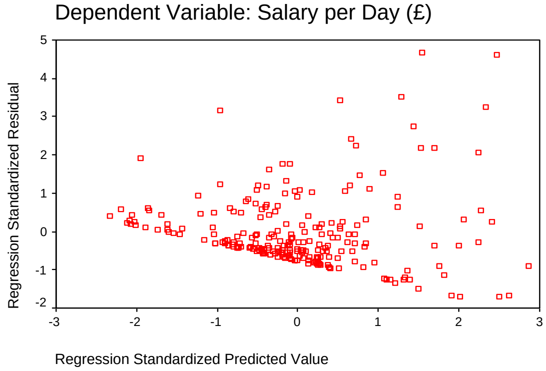
Normal P-P Plot of Regression Standardiz

Dependent Variable: Salary per Day (£)



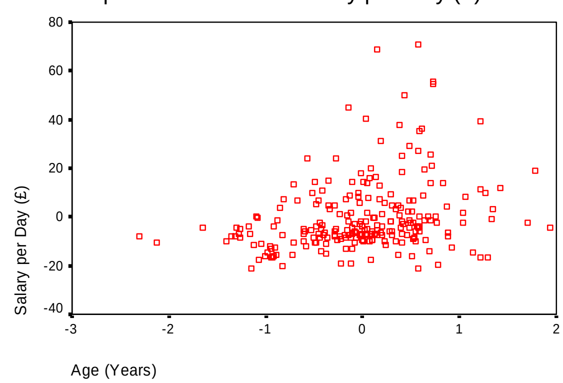
Scatterplot

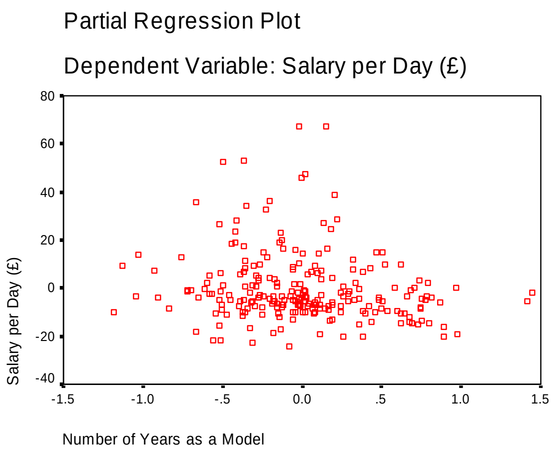
Dependent Variable: Salary per Day (£)

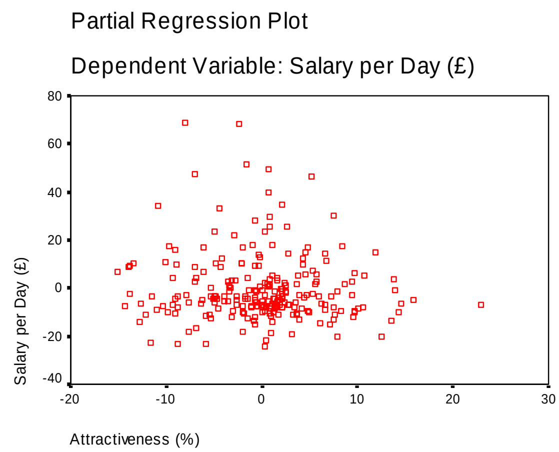


Partial Regression Plot

Dependent Variable: Salary per Day (£)







- ✓ *Residuals*: There six cases that have a standardized residual greater than 3, and two of these are fairly substantial (case 5 and 135). We have 5.19% of cases with standardized residuals above 2, so that's as we expect, but 3% of cases with residuals above 2.5 (we'd expect only 1%), which indicates possible outliers.

- ✓ *Normality of errors*: The histogram reveals a skewed distribution indicating that the normality of errors assumption has been broken. The normal P–P plot verifies this because the dashed line deviates considerably from the straight line (which indicates what you'd get from normally distributed errors).

- ✓ *Homoscedasticity and independence of errors*: The scatterplot of ZPRED vs. ZRESID does not show a random pattern. There is a distinct funnelling indicating heteroscedasticity. However, the Durbin–Watson statistic does fall within Field's recommended boundaries of 1–3, which suggests that errors are reasonably independent.

- ✓ *Multicollinearity*: For the age and experience variables in the model, VIF values are above 10 (or alternatively, tolerance values are all well below 0.2) indicating multicollinearity in the data. In fact, if you look at the correlation between these two variables it is around .9! So, these two variables are measuring very similar things. Of course, this makes perfect sense because the older a model is, the more years she would've spent modelling! So, it was fairly stupid to measure both of these things! This also explains the weird result that the number of years spent modelling negatively predicted salary (i.e. more experience = less salary!): in fact if you do a simple regression with experience as the only predictor of salary you'll find it has the expected positive relationship. This hopefully demonstrates why multicollinearity can bias the regression model.

All in all, several assumptions have not been met and so this model is probably fairly unreliable.

Task 2

- Using the Glastonbury data from this chapter (with the dummy coding in **GlastonburyDummy.sav**), which you should've already analysed, comment on whether you think the model is reliable and generalizable.

This question asks whether this model is valid.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.276[a]	.076	.053	.68818	.076	3.270	3	119	.024	1.893

a. Predictors: (Constant), No Affiliation vs. Indie Kid, No Affiliation vs. Crusty, No Affiliation vs. Metaller

b. Dependent Variable: Change in Hygiene Over The Festival

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	-.554	.090		-6.134	.000	-.733	-.375		
	No Affiliation vs. Crusty	-.412	.167	-.232	-2.464	.015	-.742	-.081	.879	1.138
	No Affiliation vs. Metaller	.028	.160	.017	.177	.860	-.289	.346	.874	1.144
	No Affiliation vs. Indie Kid	-.410	.205	-.185	-2.001	.048	-.816	-.004	.909	1.100

a. Dependent Variable: Change in Hygiene Over The Festival

Collinearity Diagnostics[a]

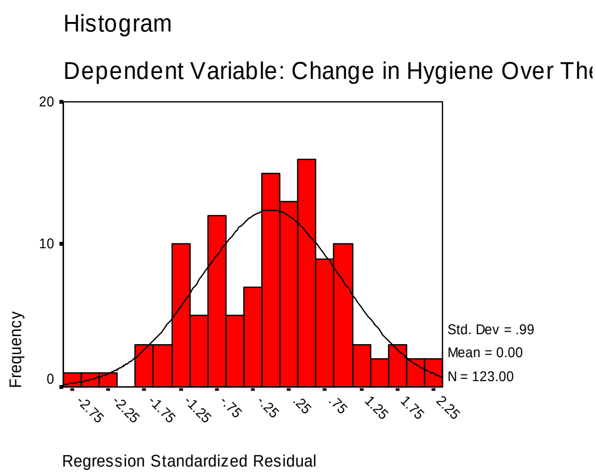
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	No Affiliation vs. Crusty	No Affiliation vs. Metaller	No Affiliation vs. Indie Kid
1	1	1.727	1.000	.14	.08	.08	.05
	2	1.000	1.314	.00	.37	.32	.00
	3	1.000	1.314	.00	.07	.08	.63
	4	.273	2.515	.86	.48	.52	.32

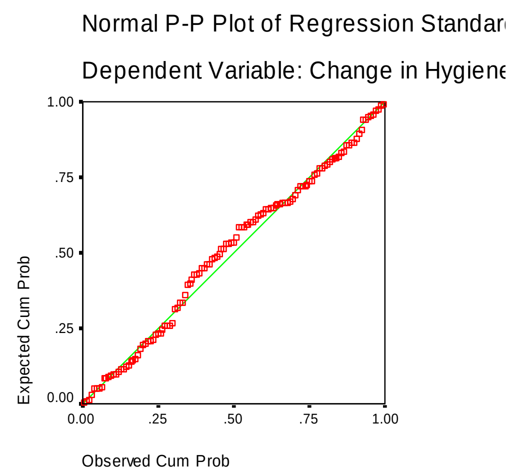
a. Dependent Variable: Change in Hygiene Over The Festival

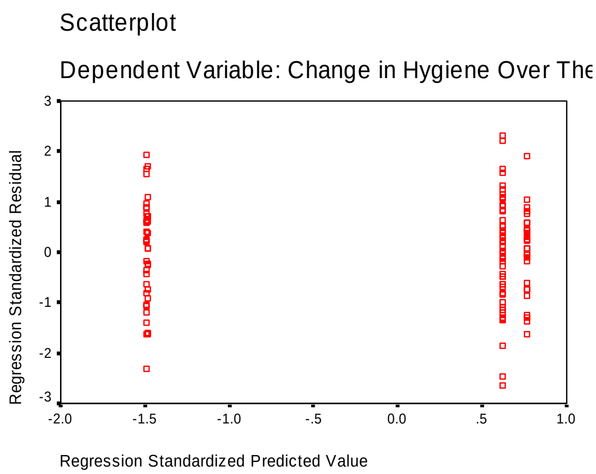
Casewise Diagnostics[a]

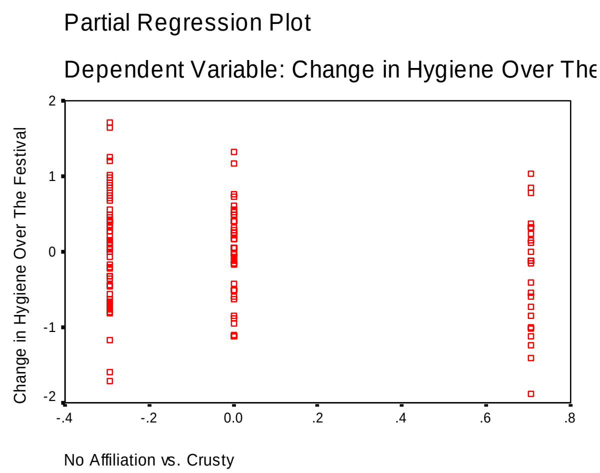
Case Number	Std. Residual	Change in Hygiene Over The Festival	Predicted Value	Residual
31	-2.302	-2.55	-.9658	-1.5842
153	2.317	1.04	-.5543	1.5943
202	-2.653	-2.38	-.5543	-1.8257
346	-2.479	-2.26	-.5543	-1.7057
479	2.215	.97	-.5543	1.5243

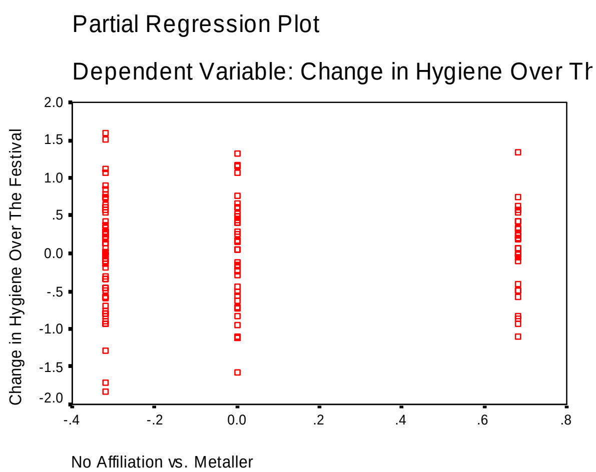
a. Dependent Variable: Change in Hygiene Over The Festival

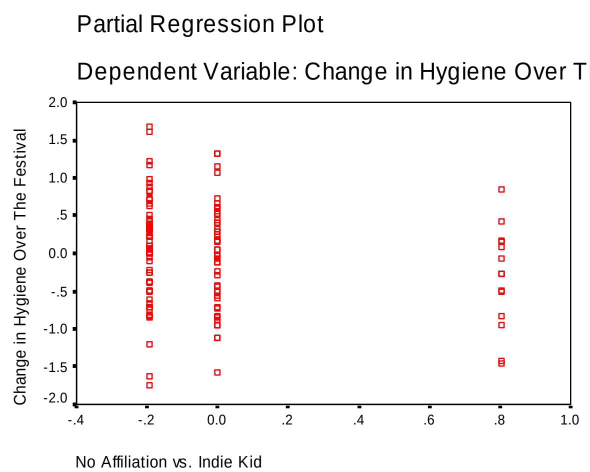












- ✓ *Residuals*: There are no cases that have a standardized residual greater than 3. We have 4.07% of cases with standardized residuals above 2, so that's as we expect, and .81% of cases with residuals above 2.5 (and we'd expect 1%), which indicates the data are consistent with what we'd expect.

- ✓ *Normality of errors*: The histogram looks reasonably normally distributed indicating that the normality of errors assumption has probably been met. The normal P–P plot verifies this because the dashed line doesn't deviate much from the straight line (which indicates what you'd get from normally distributed errors).

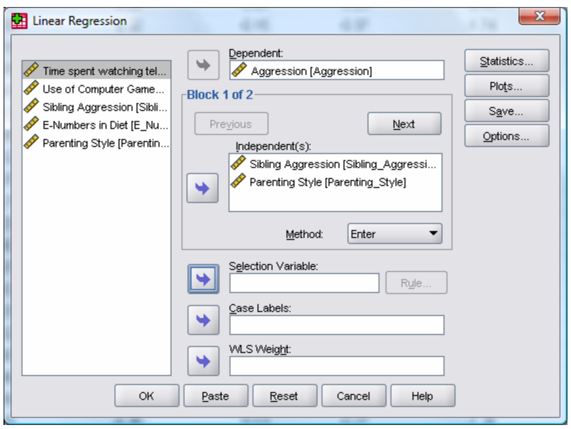
- ✓ *Homoscedasticity and independence of errors*: The scatterplot of ZPRED vs. ZRESID does look a bit odd with categorical predictors, but essentially we're looking for the height of the lines to be about the same (indicating the variability at each of the three levels is the same). This is true indicating homoscedasticity. The Durbin–Watson statistic also falls within Field's recommended boundaries of 1–3, which suggests that errors are reasonably independent.
- ✓ *Multicollinearity*: For all variables in the model, VIF values are below 10 (or alternatively, tolerance values are all well above 0.2) indicating no multicollinearity in the data.

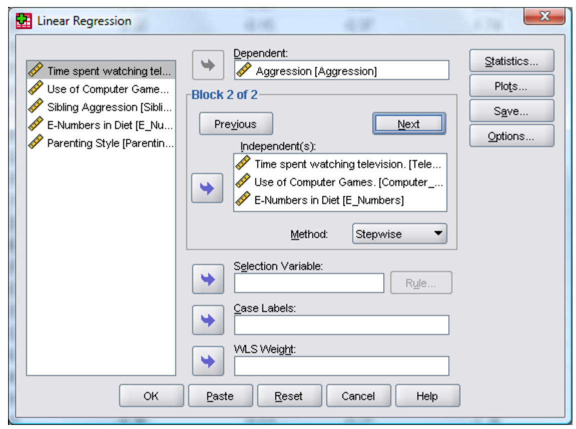
All in all, the model looks fairly reliable (but you should check for influential cases!).

Task 3

- A study was carried out to explore the relationship between aggression and several potential predicting factors in 666 children who had an older sibling. Variables measured were **Parenting_Style** (high score = bad parenting practices), **Computer_Games** (high score = more time spent playing computer games), **Television** (high score = more time spent watching television), **Diet** (high score = the child has a good diet low in E-numbers), and **Sibling_Aggression** (high score = more aggression seen in their older sibling). Past research indicated that parenting style and sibling aggression were good predictors of the level of aggression in the younger child. All other variables were treated in an exploratory fashion. The data are in the file **Child Aggression.sav**. Analyse them with multiple regression.

We need to conduct this analysis hierarchically entering parenting style and sibling aggression in the first step (forced entry) and the remaining variables in a second step (stepwise):





Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.231[a]	.053	.050	.31125	.053	18.644	2	663	.000	
2	.264[b]	.070	.066	.30875	.017	11.787	1	662	.001	
3	.286[c]	.082	.076	.30697	.012	8.682	1	661	.003	1.911

a. Predictors: (Constant), Parenting Style, Sibling Aggression

b. Predictors: (Constant), Parenting Style, Sibling Aggression, Use of Computer Games.

c. Predictors: (Constant), Parenting Style, Sibling Aggression, Use of Computer Games., Good Diet

d. Dependent Variable: Aggression

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound	Correlations Zero-order	Correlations Partial	Correlations Part	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	-.006	.012		-.479	.632	-.029	.018					
	Sibling Aggression	.093	.038	.096	2.491	.013	.020	.167	.129	.096	.094	.970	1.031
	Parenting Style	.062	.012	.194	5.057	.000	.038	.086	.211	.193	.191	.970	1.031
2	(Constant)	-.007	.012		-.574	.566	-.030	.017					
	Sibling Aggression	.068	.038	.070	1.793	.073	-.006	.142	.129	.070	.067	.933	1.072
	Parenting Style	.054	.012	.170	4.385	.000	.030	.079	.211	.168	.164	.937	1.067
	Use of Computer Games.	.126	.037	.134	3.433	.001	.054	.197	.186	.132	.129	.918	1.090
3	(Constant)	-.006	.012		-.497	.619	-.029	.017					
	Sibling Aggression	.086	.038	.088	2.258	.024	.011	.161	.129	.087	.084	.908	1.101
	Parenting Style	.062	.013	.194	4.925	.000	.037	.087	.211	.188	.184	.897	1.115
	Use of Computer Games.	.143	.037	.153	3.891	.000	.071	.216	.186	.150	.145	.893	1.120
	Good Diet	-.112	.038	-.118	-2.947	.003	-.186	-.037	-.009	-.114	-.110	.870	1.150

a. Dependent Variable: Aggression

Excluded Variables[d]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance	Collinearity Statistics VIF	Collinearity Statistics Minimum Tolerance
1	Time spent watching television.	.049[a]	1.091	.276	.042	.704	1.421	.704
	Use of Computer Games.	.134[a]	3.433	.001	.132	.918	1.090	.918
	Good Diet	-.092[a]	-2.313	.021	-.090	.894	1.119	.894
2	Time spent watching television.	.044[b]	.986	.324	.038	.703	1.423	.703
	Good Diet	-.118[b]	-2.947	.003	-.114	.870	1.150	.870
3	Time spent watching television.	.032[c]	.715	.475	.028	.697	1.436	.669

a. Predictors in the Model: (Constant), Parenting Style, Sibling Aggression

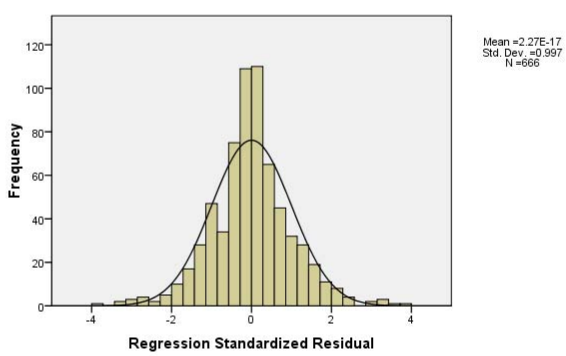
b. Predictors in the Model: (Constant), Parenting Style, Sibling Aggression, Use of Computer Games.

c. Predictors in the Model: (Constant), Parenting Style, Sibling Aggression, Use of Computer Games., Good Diet

d. Dependent Variable: Aggression

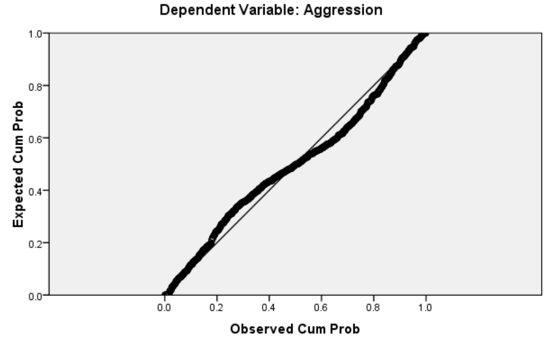
Histogram

Dependent Variable: Aggression



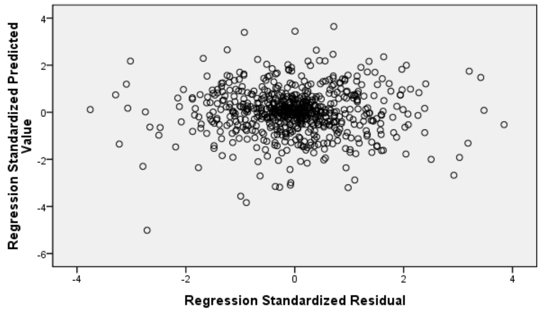
Normal P-P Plot of Regression Standardized Residual

Dependent Variable: Aggression



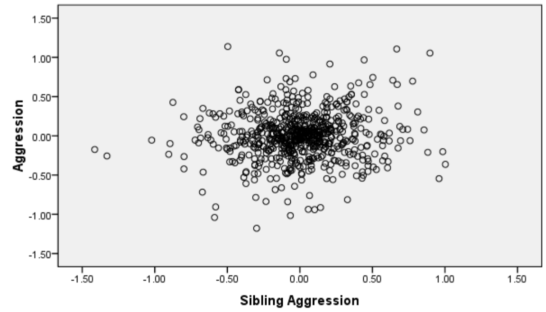
Scatterplot

Dependent Variable: Aggression



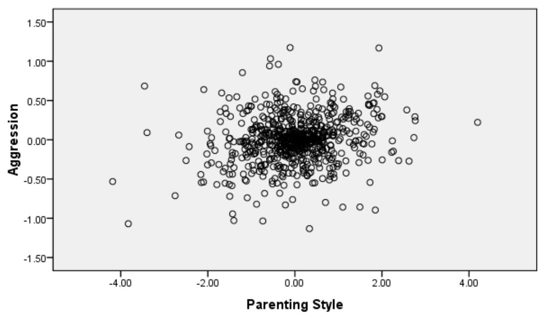
Partial Regression Plot

Dependent Variable: Aggression



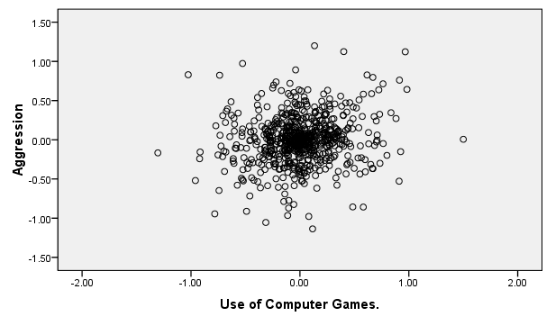
Partial Regression Plot

Dependent Variable: Aggression



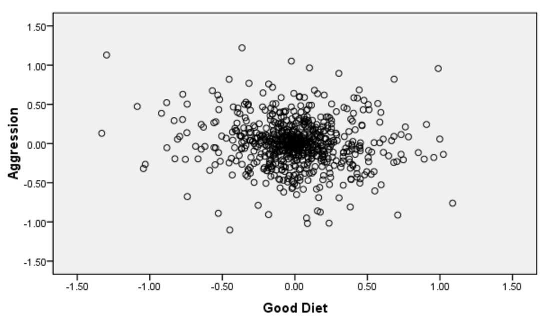
Partial Regression Plot

Dependent Variable: Aggression



Partial Regression Plot

Dependent Variable: Aggression



Casewise Diagnostics[a]

Case Number	Std. Residual	Aggression	Predicted Value	Residual
2	2.281	.77	.0710	.70014
45	-3.067	-.93	.0106	-.94162
47	2.405	.84	.1053	.73842
71	-2.496	-.86	-.0942	-.76622
75	2.126	.74	.0849	.65261
157	3.845	1.13	-.0529	1.18037
163	-2.084	-.68	-.0423	-.63962
169	3.182	.85	-.1251	.97673
182	2.051	.81	.1775	.62946
199	2.505	.58	-.1879	.76897
200	3.026	.75	-.1805	.92899
204	2.080	.63	-.0120	.63837
217	-2.712	-1.30	-.4630	-.83263
221	3.205	1.14	.1543	.98372
266	2.085	.59	-.0533	.64012
270	-3.018	-.73	.1936	-.92649
351	2.386	.74	.0101	.73259
374	2.923	.65	-.2495	.89716
375	2.263	.68	-.0170	.69483
379	-2.789	-1.07	-.2150	-.85618
386	2.388	.65	-.0841	.73290
407	-2.148	-.61	.0502	-.65934
411	-2.188	-.81	-.1394	-.67154
421	-2.045	-.54	.0833	-.62772
431	-2.472	-.82	-.0643	-.75895
439	-3.092	-.85	.1041	-.94922
440	-3.290	-.95	.0624	-1.00982
463	-3.756	-1.15	.0055	-1.15286
482	3.476	1.07	.0025	1.06707
505	-3.223	-1.12	-.1284	-.98938
539	3.416	1.18	.1300	1.04877
589	2.042	.46	-.1671	.62679
630	-2.119	-.63	.0169	-.65047
635	-2.661	-.88	-.0625	-.81672
639	-2.743	-.85	-.0037	-.84210
640	2.024	.56	-.0629	.62135

a. Dependent Variable: Aggression

Based on the final model (which is actually all we're interested in) the following variables predict aggression:

- Parenting style (b = 0.062, β = 0.194, t = 4.93, p < .001) significantly predicted aggression. The beta value indicates that as parenting increases (i.e. as bad practices increase), aggression increases also.
- Sibling aggression (b = 0.086, β = 0.088, t = 2.26, p < .05) significantly predicted aggression. The beta value indicates that as sibling aggression increases (became more aggressive), aggression increases also.
- Computer games (b = 0.143, β = 0.037, t = 3.89, p < .001) significantly predicted aggression. The beta value indicates that as the time spent playing computer games increases, aggression increases also.
- E-numbers (b = -.112, β =-0.118, t = -2.95, p < .01) significantly predicted aggression. The beta value indicates that as the diet improved, aggression decreased.

The only factor not to predict aggression was:

- ✓ Television (b if entered = .032, t = 0.72, p > .05) did not significantly predict aggression.

Based on the standardized beta values, the most substantive predictor of aggression was actually parenting style, followed by computer games, diet and then sibling aggression.

R^2 is the squared correlation between the observed values of aggression and the values of aggression predicted by the model. The values in this output tell us that sibling aggression and parenting style in combination explain 5.3% of the variance in aggression. When computer game use is factored in as well, 7% of variance in aggression is explained (i.e. an additional 1.7%). Finally, when diet is added to the model, 8.2% of the variance in aggression is explained (an additional 1.2%). With all four of these predictors in the model still less than of the variance in aggression can be explained.

The Durbin–Watson statistic tests the assumption of ‘independence of errors’, which means that for any two observations (cases) in the regression, their residuals should be uncorrelated (or independent). In this output the Durbin–Watson statistic falls within the recommended boundaries of 1–3, which suggests that errors are reasonably independent.

The scatterplot helps us to assess both *homoscedasticity* and *independence of errors*. The scatterplot of ZPRED vs. ZRESID does show a random pattern and so indicates no violation of the independence of errors assumption. Also, the errors on the scatterplot do not funnel out, indicating homoscedascitity of errors, thus no violations of these assumptions.