ON SCREEN AND ON AIR TALENT

AN ASSESSMENT OF THE BBC'S APPROACH AND IMPACT

A REPORT FOR THE BBC TRUST

APPENDIX VIII –
THE REGRESSION ANALYSIS

BY OLIVER & OHLBAUM ASSOCIATES

APRIL 2008

APPENDIX VIII – THE REGRESSION ANALYSIS

INTRODUCTION

O&O performed regression analysis to determine the importance of talent on television viewing audiences. The analysis uses two sets of data: a large dataset consisting of strands from the five main networks (BBC1, BBC2, ITV1, Channel 4 and Five) and within sixteen sub-genres[1] and a smaller dataset containing BBC-only data which also included talent costs.

The regression analysis implemented OLS (ordinary least squares) regression on both sets of data, with average audience as the dependent variable (the left-hand side of the equation). The independent variables (right-hand side variables) mainly comprised of dummy variables[2] to denote strand characteristics, such as the broadcast channel, sub-genre, and stratification of the main talent.

REGRESSION ANALYSIS WITH LARGE DATASET (BBC, ITV, C4 and FIVE)

The equation of the first regression is in log-linear form as denoted below:

$$\begin{aligned}\ln AUD_i = {} & C + \alpha_1 year06 + \alpha_2 year07 + \beta_1 BBC2 + \beta_2 ITV + \beta_3 CH4 + \beta_4 FIVE + \beta_5 chat + \beta_6 family + \beta_7 othercomedy + \\ & + \beta_8 quiz + \beta_9 sitcom + \beta_{10} science + \beta_{11} naturalhist + \beta_{12} serials + \beta_{13} series + \beta_{14} plays + \beta_{15} factent + \beta_{16} human + \\ & + \beta_{17} factdrama + \beta_{18} lifestyle + \beta_{19} cook + \beta_{20} prepeak + \beta_{21} earlypeak + \beta_{22} midpeak + \beta_{23} latepeak + \beta_{24} afterpeak + \\ & + \beta_{25} SUPER + \beta_{26} TOP + \beta_{27} KNOWN + \varepsilon\end{aligned}$$

Where the variables are defined as the following:

VARIABLE	DEFINITION
C	Constant
lnAUD	Natural log of viewing audience
year06	Dummy variable takes the value of 1 if strand occurred in 2006
year07	Dummy variable takes the value of 1 if strand occurred in 2007
BBC2	Dummy variable takes the value of 1 if strand occurred on BBC2
ITV	Dummy variable takes the value of 1 if strand occurred on ITV1
CH4	Dummy variable takes the value of 1 if strand occurred on Channel 4
FIVE	Dummy variable takes the value of 1 if strand occurred on Channel 5
chat	Dummy variable takes the value of 1 if strand fell into the chat sub-genre
family	Dummy variable takes the value of 1 if strand fell into the family show sub-genre
othercomedy	Dummy variable takes the value of 1 if strand fell into the other comedy sub-genre
quiz	Dummy variable takes the value of 1 if strand fell into the quiz/panel game sub-genre
sitcom	Dummy variable takes the value of 1 if strand fell into the sitcom sub-genre

[1] The sixteen sub-genres are: Chat shows, family shows, quiz/panel games, other comedy, sitcoms, drama series, drama serials, drama single plays, factual entertainment, human interest, factual drama, lifestyle, and cookery.

[2] Dummy variables are binary and take the value of "1" when the status is "true" or "0" when the status is "not true." For example, the dummy variable *ITV* would take the value of "1" if the strand was broadcast on ITV1.

science	Dummy variable takes the value of 1 if strand fell into the science/medical sub-genre
naturalhist	Dummy variable takes the value of 1 if strand fell into the natural history sub-genre
serials	Dummy variable takes the value of 1 if strand fell into the drama-serials sub-genre
series	Dummy variable takes the value of 1 if strand fell into the drama-series sub-genre
plays	Dummy variable takes the value of 1 if strand fell into the drama-single plays sub-genre
factent	Dummy variable takes the value of 1 if strand fell into the factual entertainment sub-genre
human	Dummy variable takes the value of 1 if strand fell into the human interest sub-genre
factdrama	Dummy variable takes the value of 1 if strand fell into the factual drama sub-genre
lifestyle	Dummy variable takes the value of 1 if strand fell into the lifestyles and makeover sub-genre
cook	Dummy variable takes the value of 1 if strand fell into the cookery show sub-genre
prepeak	Dummy variable takes the value of 1 if strand occurred between 4pm-6pm
earlypeak	Dummy variable takes the value of 1 if strand occurred between 6pm-8pm
midpeak	Dummy variable takes the value of 1 if strand occurred between 8pm-10pm
latepeak	Dummy variable takes the value of 1 if strand occurred between 10pm-11pm
afterpeak	Dummy variable takes the value of 1 if strand occurred between 11pm-6am
SUPER	Dummy variable takes the value of 1 if strand had "super talent"
TOP	Dummy variable takes the value of 1 if strand had "top talent"
KNOWN	Dummy variable takes the value of 1 if strand had "known talent"

The variable coefficients (the ß's) measure the influence that variable has on the dependent variable. However, when a regression equation is in log-linear form, a calculation is required to derive the true influence of the dummy variables.

If the general equation takes the form $\ln Y = \beta_1 + \beta_2 X + \beta_3 D_1$ where D is a dummy variable, the percentage increase in Y for the dummy taking a value of "1" compared to the base case is $(e^{\beta_3} - 1) * 100$

REGRESSION RESULTS

The results from the first regression are presented below.

Dependent Variable: LN_AUD	COEFFICIENTS	STD ERROR	T-STAT	SIGNIFICANCE	DUMMY INFLUENCE (%)
Constant	5.45	0.16	33.48	0.00	
BBC2	-0.63	0.33	-1.91	-0.06	***-46.7***
ITV1	**-0.18**	**0.07**	**-2.41**	**0.02**	***-16.2***
CH4	0.18	0.17	1.06	0.29	***19.4***
FIVE	**-0.28**	**0.12**	**-2.27**	**0.02**	***-24.2***
YEAR 2006	**-0.16**	**0.06**	**-2.63**	**0.01**	***-14.6***
YEAR 2007	**-0.27**	**0.06**	**-4.22**	**0.00**	***-23.9***
CHAT SHOWS	0.18	0.24	0.73	0.46	***19.2***
FAMILY SHOWS	**1.16**	**0.19**	**6.27**	**0.00**	***220.6***
OTHER COMEDY	0.31	0.21	1.48	0.14	***36.7***
QUIZ/PANEL SHOWS	**0.64**	**0.20**	**3.19**	**0.00**	***89.7***
SITCOMS	0.23	0.21	1.09	0.28	***25.6***
SCIENCE/MEDICAL	0.40	0.25	1.59	0.11	***48.6***
NATURAL HISTORY	**0.50**	**0.20**	**2.50**	**0.01**	***65.5***
SERIALS	**1.06**	**0.22**	**4.79**	**0.00**	***188.8***
SERIES	**1.12**	**0.19**	**5.96**	**0.00**	***206.9***
SINGLE PLAYS	**1.06**	**0.20**	**5.43**	**0.00**	***188.6***
FACTUAL ENTERTAINMENT	**1.12**	**0.17**	**6.55**	**0.00**	***207.9***
HUMAN INTEREST	**0.42**	**0.15**	**2.79**	**0.01**	***52.1***
FACTUAL DRAMA	**1.36**	**0.42**	**3.26**	**0.00**	***289.2***
LIFESTYLE	0.03	0.16	0.15	0.88	***2.6***
COOKERY	-0.24	0.21	-1.18	0.24	***-21.7***
PRE PEAK	0.26	0.15	1.68	0.09	***29.2***
EARLY PEAK	**0.56**	**0.10**	**5.71**	**0.00**	***75.2***
MID PEAK	**0.92**	**0.10**	**9.45**	**0.00**	***151.4***
LATE PEAK	0.17	0.12	1.37	0.17	***18.2***
AFTER PEAK	**-0.60**	**0.13**	**-4.67**	**0.00**	***-45.2***
SUPER TALENT	**1.33**	**0.10**	**12.89**	**0.00**	***278.9***
TOP TALENT	**1.23**	**0.09**	**13.16**	**0.00**	***243.5***
KNOWN TALENT	**0.98**	**0.09**	**10.98**	**0.00**	***167.3***
R^2=66%					

Variables in bold indicate that they are 95% significant which means that there is only a 5% chance that their influence would occur randomly.

Firstly, since the equation is in log linear form, the correct interpretation of the dummy variable coefficients is derived from the calculation stated in the equation above and denoted in the last column of the table. Secondly, dummy variables can only be interpreted relative to the base case. In regression equations the base case is implicit. For example in the above equation, the base sub-genre is 'history.' This means that all

sub-genres have a representative dummy variable in the equation *except* 'history' since 'history' would be the implied sub-genre if all sub-genre dummy variables took the value of "0." In the regression above, the base time slot is 'daytime', the base channel is 'BBC1', the base year is '2005' and the base level of talent is 'unknown.' So for example, with all else being equal, a strand with 'super talent' on average has 278& higher audiences relative to shows without any talent.

MODEL LIMITATIONS:

- The large dataset is still a very small dataset when compared to the entire population of programmes on the five networks. Not all genres are represented and of the genres that are in the model, only a subset of strands within those subgenres is included. Including only a small proportion of the entire population of data possibly introduces a selection bias, although it is unknown if this has actually occurred in this case.

- One-off programmes and repeats are not removed from this dataset, and there is a high likelihood of substantial programme repeats contained within the dataset. Additionally, the BARB average audience measure will also contain repeats, in other words the average audience of a strand will be influenced by audiences viewing its repeat showing. Programmes with a higher repeat rate will be more susceptible to this type of data bias. This bias may especially present a problem where shows are repeated in the daytime or during late nights, but where the original show was broadcast during peak time. Thus one could attribute some degree of measurement error to the audience viewing data.

- Where appropriate, the talent value survey provided influence on the coding of talent. However, the majority of talent coding had to be done on a subjective basis. There is a potential the data to be miscoded for some strands.

- There has been no testing done with respect to the run of causality, i.e. do large audiences watch programmes and therefore create talent (bigger audiences lead to higher talent costs because channels can afford it) or does talent drive larger audiences. This type of testing (Granger causality testing) has not been performed in this analysis.

- A potentially more robust regression analysis would involve not only the entire population of strands but also data over time. A time series approach is more insightful because it facilitates the use of differenced equations which would capture the *changes* in talent over time. Secondly, a differenced equation would eliminate unobserved effects not captured by the measured variables. There exists the possibility that some effects are unobservable either because there exists no tangible data in which to measure the effect or simply because there is no access to the data which would capture it. If it is the case that these 'un-captured' effects are correlated or enveloped within another variable that is included in the regression, then the coefficient of the variable will not distinguish between two potentially distinct effects. Thus the significance and measure of the coefficient for the variable will be incorrect. Lastly, a time series approach would also enable the representation of 'strand time-specific' effects, e.g. legacy, loyalty, cult status, etc.

There exists the potential for multi-collinearity: when the variables on the right-hand side of the equation are correlated. For example, some genres might only be shown in peak times so the timeslot and genre might correlate with each other. This will also have an impact of the variable coefficients and may make them seem significant when in actuality they are not.

REGRESSION ANALYSIS WITH BBC DATASET (SMALL BBC SET)

A second regression analysis used a much smaller data set containing limited BBC-only information. The regression equation is also log-linear and takes the following form:

$$\ln AUD_i = C + \alpha_1 year06 + \alpha_2 year07 + \beta_1 BBC2 + \beta_2 BBC3 + \beta_3 earlypeak + \beta_4 midpeak + \beta_5 serials + \beta_6 plays + \\ + \beta_7 panel + \beta_8 family + \beta_9 sitcom + \beta_{10} othercom + \beta_{11} lifestyle + \beta_{12} natural + \beta_{13} history + \beta_{14} human + \beta_{15} \ln talent\cos t + \varepsilon$$

Where the variables are defined as the following:

VARIABLE	DEFINITION
lnAUD	Natural log of viewing audience
year06	Dummy variable takes the value of 1 if strand occurred in 2006
year07	Dummy variable takes the value of 1 if strand occurred in 2007
BBC2	Dummy variable takes the value of 1 if strand occurred on BBC2
BBC3	Dummy variable takes the value of 1 if strand occurred on BBC3
earlypeak	Dummy variable takes the value of 1 if strand occurred between 6pm-8pm
midpeak	Dummy variable takes the value of 1 if strand occurred between 8pm-10pm
serials	Dummy variable takes the value of 1 if strand belongs to the drama: serials sub-genre
plays	Dummy variable takes the value of 1 if strand belongs to the drama: plays sub-genre
panel	Dummy variable takes the value of 1 if strand belongs to the quiz/panel game sub-genre
family	Dummy variable takes the value of 1 if strand belongs to the family show sub-genre
sitcom	Dummy variable takes the value of 1 if strand belongs to the sitcom sub-genre
othercom	Dummy variable takes the value of 1 if strand belongs to the other comedy sub-genre
lifestyle	Dummy variable takes the value of 1 if strand belongs to the lifestyle sub-genre
natural	Dummy variable takes the value of 1 if strand belongs to the natural history sub-genre
history	Dummy variable takes the value of 1 if strand belongs to the history sub-genre
human	Dummy variable takes the value of 1 if strand belongs to the human interest sub-genre
ln talentcosts	The natural log of talent costs for the strand that year

This regression equation is also log linear but in this case the talent measure used is not a dummy variable, but rather the actual talent costs for the strand. Thus one can interpret the coefficient on the talent variable without manipulation. So if the talent cost coefficient is 0.2, then a 1% increase in talent costs increases the audience by 0.2%. The

other variables are dummy variables and their coefficients should be interpreted as in the first regression.

Dependent Variable: LN AUD	COEFFICIENT	STD. ERROR	T-STAT	SIGNIFICANCE	DUMMY SIGNIFICANCE (%)
(Constant)	4.84	1.27	3.81	0.00	
YEAR 06	-0.13	0.12	-1.09	0.28	**-11.8**
YEAR 07	-0.14	0.12	-1.17	0.25	**-12.8**
EARLY_PEAK	**0.68**	**0.29**	**2.35**	**0.02**	**96.5**
MID_PEAK	**0.48**	**0.17**	**2.83**	**0.01**	**62.2**
BBC2	**-0.68**	**0.16**	**-4.16**	**0.00**	**-49.4**
BBC3	**-1.88**	**0.39**	**-4.85**	**0.00**	**-84.7**
SERIALS	0.07	0.21	0.32	0.75	**6.8**
PLAYS	0.23	0.34	0.66	0.51	**25.3**
PANEL GAMES	-0.23	0.47	-0.48	0.63	**-20.2**
FAMILY SHOWS	**0.49**	**0.16**	**3.04**	**0.00**	**63.6**
SITCOM	-0.38	0.27	-1.41	0.16	**-31.8**
OTHER COMEDY	-0.01	0.31	-0.02	0.98	**-0.7**
NATURAL HISTORY	**0.94**	**0.38**	**2.49**	**0.02**	**157.0**
LIFESTYLE	0.22	0.26	0.85	0.40	**24.4**
HISTORY	0.66	0.40	1.65	0.10	**94.4**
HUMAN INTEREST	0.50	0.47	1.05	0.30	**64.5**
LN_COST	**0.21**	**0.09**	**2.45**	**0.02**	**23.9**
R^2 =91%					

In this regression the base variables were 'BBC1' for the channel, 'off-peak' for the time-slot, 'drama series' for sub-genre, and '2005' for year. Again, variables that are 95% significant are denoted in bold type.

MODEL LIMITATIONS:

- The BBC dataset was a very small dataset (less than 100 observations and less than 50 strands) which limits meaningful extrapolations. Because the sample is small, there may not be sufficient variety represented by the dataset and there is a very strong possibility of selection bias in this sample. For example, one cookery strand would not be indicative of the entire cookery sub-genre, but using only one strand within a regression analysis would give that strand undue influence.

- The talent cost data represents more than just the cost of actors and presenters. This would mask or over-emphasise the true influence of pure talent. If this measurement error is combined with the limitations of a small dataset, the bias is magnified.