Chapter 5: Answers

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 on the CD-ROM 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 3 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 generalises 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 generalise well.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	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:

$$\begin{aligned}\text{Salary} &= \beta_0 + \beta_1\text{Age}_i + \beta_2\text{Experience}_i + \beta_3\text{Attractiveness}_i \\ &= -60.89 + \left(6.23\text{Age}_i\right) - \left(5.56\text{Experience}_i\right) - \left(0.02\text{Attractiveness}_i\right)\end{aligned}$$

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

Collinearity Diagnostics[a]

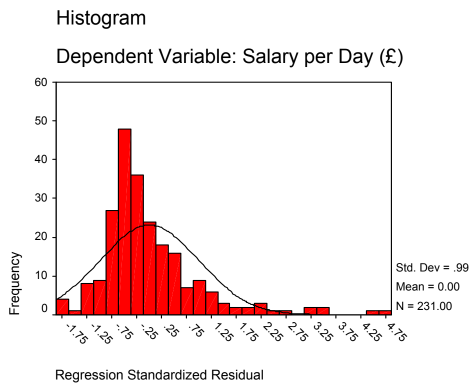
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	Age (Years)	Number of Years as a Model	Attractiveness (%)
1	1	3.925	1.000	.00	.00	.00	.00
	2	.070	7.479	.01	.00	.08	.02
	3	.004	30.758	.30	.02	.01	.94
	4	.001	63.344	.69	.98	.91	.04

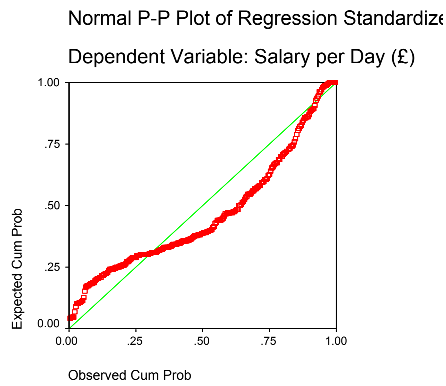
a. Dependent Variable: Salary per Day (£)

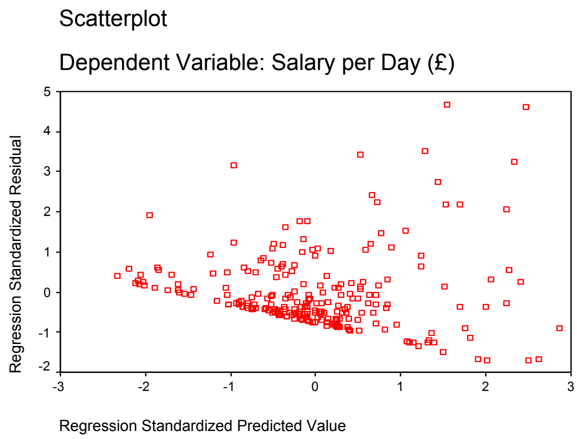
Casewise Diagnostics[a]

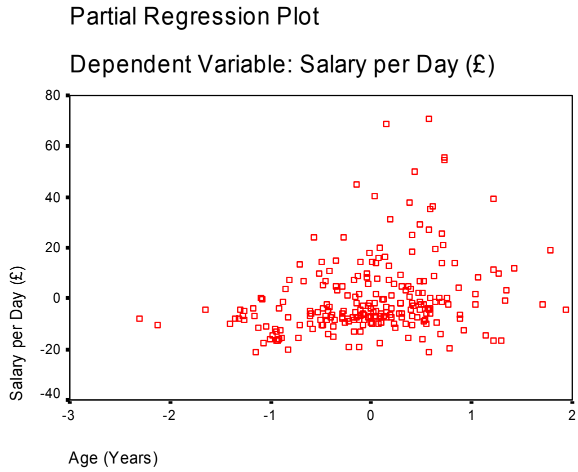
Case Number	Std. Residual	Salary per Day (£)	Predicted Value	Residual
2	2.186	53.72	21.8716	31.8532
5	4.603	95.34	28.2647	67.0734
24	2.232	48.87	16.3444	32.5232
41	2.411	51.03	15.8861	35.1390
91	2.062	56.83	26.7856	30.0459
116	3.422	64.79	14.9259	49.8654
127	2.753	61.32	21.2059	40.1129
135	4.672	89.98	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

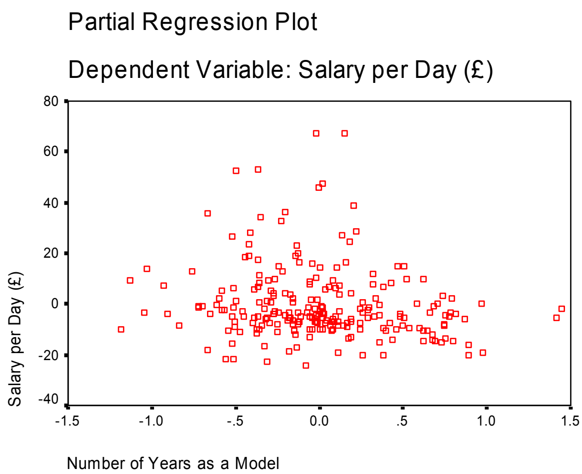
a. Dependent Variable: Salary per Day (£)

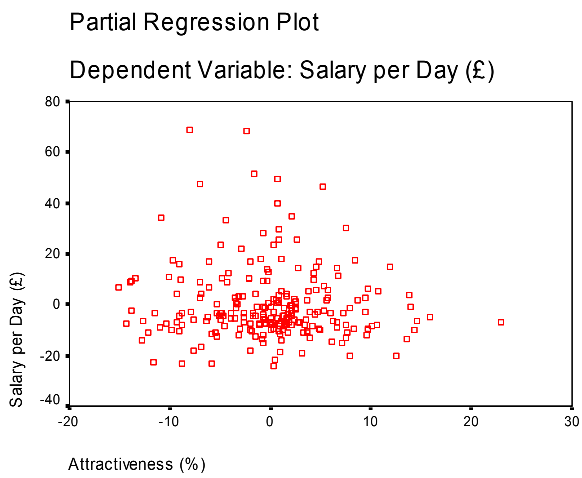












- *Residuals*: there 6 cases that has a standardized residual greater than 3, and two of these are fairly substantial (case 5 and 135). We have 5.19% of cases with standardized residuals above 2, so that's as we expect, but 3% of cases with residuals above 2.5 (we'd expect only 1%), which indicates possible 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 dotted 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]

					Change Statistics					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	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		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	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]

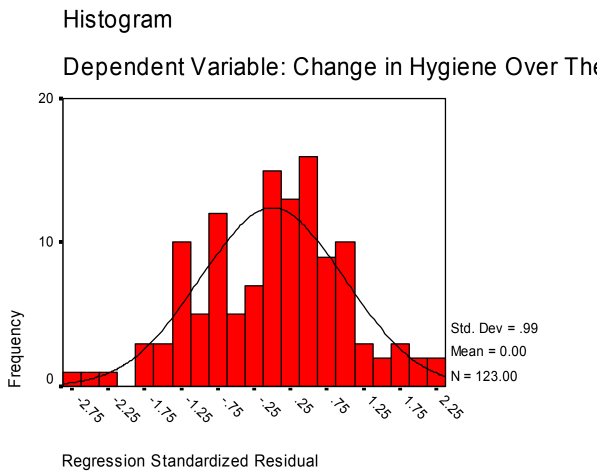
				Variance Proportions			
Model	Dimension	Eigenvalue	Condition Index	(Constant)	No Affiliation vs. Crusty	No Affiliation vs. Metaller	No Affiliation vs. Indie Kid
1	1	1.727	1.000	.14	.08	.08	.05
	2	1.000	1.314	.00	.37	.32	.00
	3	1.000	1.314	.00	.07	.08	.63
	4	.273	2.515	.86	.48	.52	.32

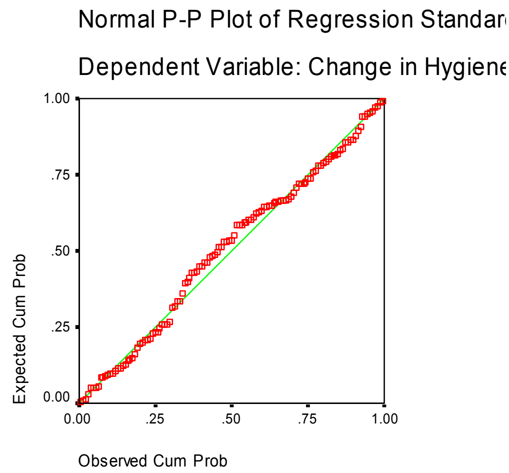
a. Dependent Variable: Change in Hygiene Over The Festival

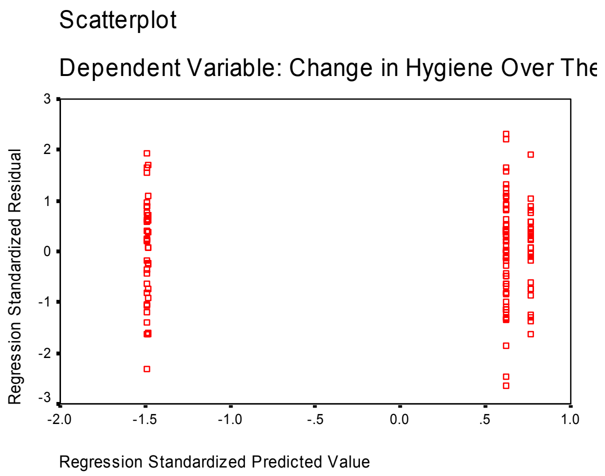
Casewise Diagnostics[a]

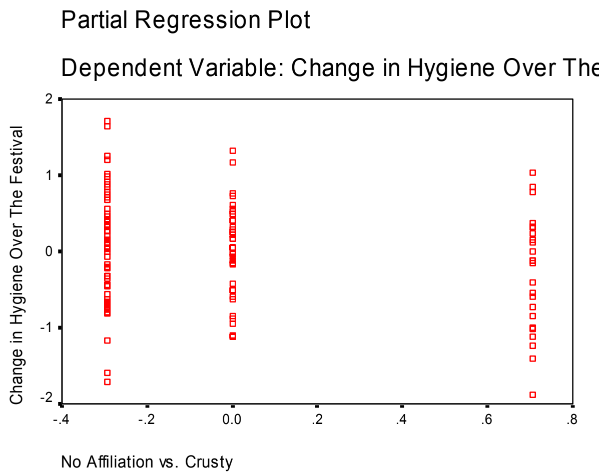
Case Number	Std. Residual	Change in Hygiene Over The Festival	Predicted Value	Residual
31	-2.302	-2.55	-.9658	-1.5842
153	2.317	1.04	-.5543	1.5943
202	-2.653	-2.38	-.5543	-1.8257
346	-2.479	-2.26	-.5543	-1.7057
479	2.215	.97	-.5543	1.5243

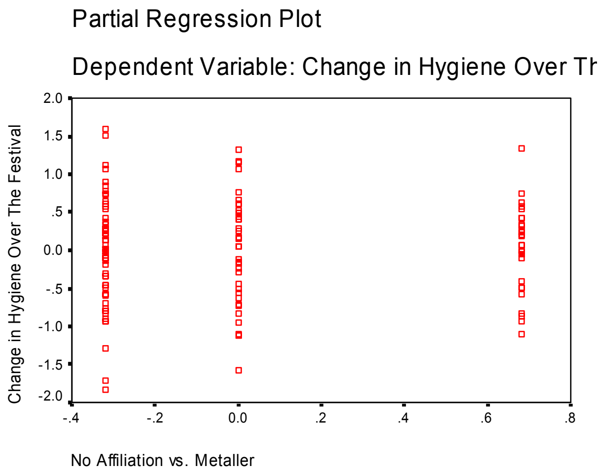
a. Dependent Variable: Change in Hygiene Over The Festival

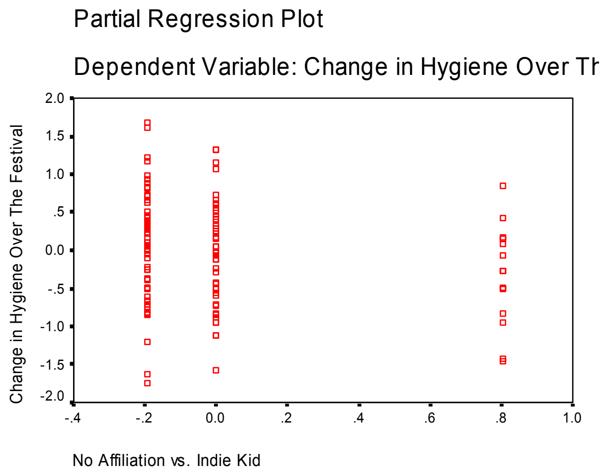












- *Residuals*: there are no cases that have a standardized residual greater than 3. We have 4.07% of cases with standardized residuals above 2, so that's as we expect, and .81% of cases with residuals above 2.5 (and we'd expect 1%), which indicates the data are consistent with what we'd expect.
- *Normality 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 dotted line doesn't deviates 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*: 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!).