Probability and statistics for data analysis

2nd Assignment

December 23, 2023

Professors: I. Vrontos

Athens University of Economics and Business

MSc in Data Science

Contents

1 ANOVA Testing

1.1 One-way ANOVA Tests

The relationship between W and $Y, X1, X2, X3, X4$ can be found in Figure 1. Specifically:

- There are significant differences between X1 and W on a 90% confidence level $p = 0.0915$. The normality assumptions hold on a 95% confidence level (S-W $p = 0.268$, K-S $p = 0.1506$) and so does the homogeneity assumption (Lev $p = 0.3367$).
- There are no statistically significant differences between X2 and W on a 90% confidence level $p = 0.128$. The normality assumptions hold on a 95% confidence level (S-W $p = 0.8049$, K-S $p = 0.2343$) and so does the homogeneity assumption (Lev $p = 0.3412$).
- There are no statistically significant differences between X3 and W on a 90% confidence level $p = 0.876$. The normality assumptions hold on a 95% confidence level (S-W $p = 0.2555$, K-S $p = 0.1112$). We reject the homogeneity assumption on a 90% confidence level (Lev $p = 0.0007$) and as such our results may not be accurate.
- There are statistically significant differences between X2 and W on a 95% confidence level $p = 0.0168$. The normality assumptions hold on a 95% confidence level (S-W $p = 0.4243$, K-S $p = 0.4261$) and so does the homogeneity assumption (Lev $p = 0.4261$).

1.2 Graphical Representation of X_i effect on Y depending on W

Figure 2 shows the relation of Y with the other variables $X_i, i = 1, 2, 3, 4$ depending on W.

1.3 Simple Regression

Table 1 shows the basic model where $Y = \beta_1 X4 + \beta_0 + \varepsilon, \varepsilon \sim N(0, \sigma^2)$.

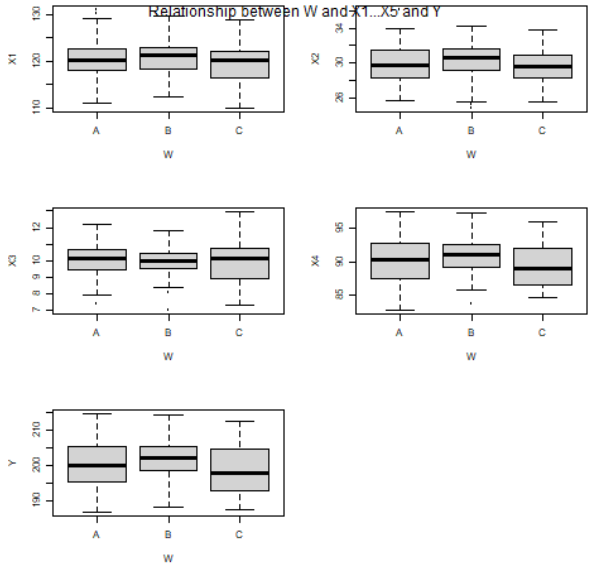

Figure 1: Boxplots of W in relation to the other variables.

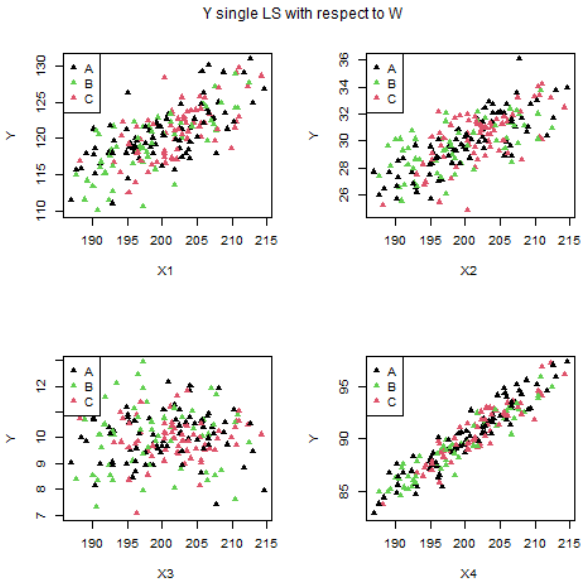

Figure 2: Scatter-plot of Y and $X_i i = 1, 2, 3, 4$ depending on W.

Table 1: Linear Regression of Y depending on X4

	Dependent variable:
	Y
X4	1.935^{***}
	p = 0.000
Constant	26.197^{***}
	p = 0.00000
Observations	200
R^2	0.886
Adjusted R^2	0.886
Residual Std. Error	2.129 (df = 198)
F Statistic	$1{,}540.022^{***}$ (df = 1; 198)
Note:	*p<0.1; **p<0.05; ***p<0.01

1.4 Multiple Regression

Table 2 shows the full model with base effects and interactions.

1.5 Checking the Full Model Assumptions

The model suffers from multi-collinearity ($GVIF > 10$) on most variables, meaning the model cannot be interpreted. We remove the variables with the biggest VIF scores and thus produce the model found in Table 3.

We do not reject the normality assumption on the new model on a 95% confidence level (S-W $p = 0.2253$, K-S $p = 0.5005$) nor the homogeneity assumption (Lev $p = 0.2985$, Bart $p = 0.4587$). Therefore the LS regression assumptions hold.

1.6 Stepwise Model Selection

We use the stepwise selection procedure with the full model being the valid model presented in Table 3.

The resulting model can be found in Table 4. Note that the dimensionality of the model has indeed been significantly reduced and that almost all terms are statistically significant on a 95% confidence level.

Table 2: Mutiple Linear Regression of Y with main effects and interaction

	Dependent variable:
	Y
X1	1.168^{***}
	p = 0.00001
WB	−8.239
	p = 0.481
WC	-24.413^{**}
	p = 0.025
X2	2.701^{***}
	p = 0.00000
X3	0.322
	p = 0.166
X4	−0.586
	p = 0.245
X1:WB	−0.212
	p = 0.538
X1:WC	−0.439
	p = 0.227
WB:X2	−0.923
	p = 0.201
WC:X2	-1.356^{*}
	p = 0.068
WB:X3	0.284
	p = 0.450
WC:X3	−0.309
	p = 0.317
WB:X4	0.657
	p = 0.335
WC:X4	1.348^{*}
	p = 0.057
Constant	28.361^{***}
	p = 0.0002
Observations	200
R^2	0.917
Adjusted R^2	0.911
Residual Std. Error	1.879 (df = 185)
F Statistic	146.212^{***} (df = 14; 185)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3: Mutiple Linear Regression with no multicolinearity issues

	Dependent variable:
	Y
X1	0.968***
	p = 0.000
X2	1.939***
	p = 0.000
X3	0.245*
	p = 0.077
X4	0.058
	p = 0.839
WB	0.654**
	p = 0.045
WC	0.340
	p = 0.319
Constant	17.944***
	p = 0.0002
Observations	200
R^2	0.909
Adjusted R^2	0.907
Residual Std. Error	1.924 (df = 193)
F Statistic	322.679*** (df = 6; 193)
Note:	*p<0.1; **p<0.05; ***p<0.01

The model presents no multi-colinearity issues. We do not reject the normality assumption on the new model on a 95% confidence level (S-W $p = 0.2006$, K-S $p = 0.6342$) nor the homogeneity assumption (Lev $p = 0.4434$, Bart $p = 0.4434$). Therefore the LS regression assumptions hold.

Table 4: Stepwise Mutiple Linear Regression on Y

	Dependent variable:
	Y
X1	0.997^{***}
	p = 0.000
X2	1.999^{***}
	p = 0.000
X3	0.251^{*}
	p = 0.064
WB	0.655^{**}
	p = 0.045
WC	0.337
	p = 0.322
Constant	17.878^{***}
	p = 0.0002
Observations	200
R^2	0.909
Adjusted R^2	0.907
Residual Std. Error	1.919 (df = 194)
F Statistic	389.128^{***} (df = 5; 194)
Note:	*p<0.1; **p<0.05; ***p<0.01

1.7 Estimating Y

When $X1 = 120, X2 = 30, X3 = 10, X4 = 90, W = B$ the stepwise model presented in Table 4 predicts a value of $Y = 200.6143$ with a 95% confidence interval of $(200.1456 \leq Y \leq 201.083)$.

Z	*W*			***Total***
	A	B	C	
(82.9,87.7]	20 40.8 %	9 18.4 %	20 40.8 %	49 100 %
(87.7,89.9]	14 28 %	19 38 %	17 34 %	50 100 %
(89.9,92.5]	21 42 %	21 42 %	8 16 %	50 100 %
(92.5,97.4]	20 40 %	18 36 %	12 24 %	50 100 %
Total	75 37.7 %	67 33.7 %	57 28.6 %	199 100 %

χ^2=12.690 · df=6 · Cramer's V=0.179 · p=0.048

Figure 3: Contingency table between W and Z (quantiles of X4)

1.8 Categorizing Continuous Variable

The contingency table can be found in 3.

Executing the two-way ANOVA test between Y and W*Z, we determine that there are statistically significant differences in the means of Y depending on the two variables on a 95% confidence interval (W: $p = 5.12e-09$, Z: $p = 0$) but not their interaction (W:Z $p = 0.305$). Therefore while the means deviate, the slope of Y to Z and Y to W does not change.

We do not reject the normality assumption on the ANOVA test on a 95% confidence level (S-W $p = 0.8318$, K-S $p = 0.9147$) nor the homogeneity assumption (Lev $p = 0.6634$). Therefore the ANOVA assumptions hold.