1. Com base no banco de dados budget.xls (o mesmo usado em todas as aulas do curso), gerar o modelo final que resolva o problema da multicolinearidade do modelo “model2” usado nas aulas.
2. O modelo final é significativo?

R - Ainda não.

No modelo apresentado acima precisa sofrer ajustes, pois temos duas variáveis que se pretende prever o gasto com saúde, essas duas variáveis independentes (X) tem uma correlação forte positiva com a variável dependente (Y), a priori, essas variáveis devem ser independentes, mas como elas tem uma forte correlação entre si, do qual “fere a regra de ser independente”, uma pode estar influenciando no comportamento da outra, que por fim, na previsora, deixando o modelo não muito bom. Vide também o p-value do intercepto ou Beta0 com valor igual a 0,134.

1. Como você interpreta os parâmetros deste modelo final (tem intercepto? O que significam as estimativas dos parâmetros?).

**4.** Qual o poder de ajuste deste modelo?

**5.** O modelo tem pontos extremos (outliers)?

**6.** Com base neste novo modelo final, o que as medidas Distancia de Cook, DFBETA, DFFIT, Alavanca e CVR sugerem sobre os outliers? Todos estes outliers podem ser retirados?

**7.** O modelo satisfaz a suposição sobre os Tipos das Variáveis envolvidas?

**8.** O modelo satisfaz a suposição de Multicolinearidade?

**9.** O modelo satisfaz a suposição Homocedasticidade?

**10.** O modelo satisfaz a suposição de que os resíduos devem ser normalmente distribuídos e com média zero?

**11.** O modelo satisfaz a suposição de que os resíduos devem ser aleatórios?

**12.** O modelo satisfaz a suposição de Independência na variável Y?

**13.** O modelo satisfaz a suposição de Linearidade entre as variáveis X e Y?

**14.** Qual a sua sugestão para a melhoria da modelagem para estes dados?

Por fim, é necessário retirar o intercepto ajustando o modelo para -1. Outro ponto, mesmo com p-value baixo na variáveis X, elas têm uma correlação muito forte entre si gerando uma multicolinearidade no modelo.

O modelo foi ajustado sem o intercepto, já que seu p-value estava com 0,134, portanto foi gerado um novo modelo passando o valor -1:

model <- lm(Health.Spending ~ -1 + Leisure.Spending + Spending.on.education, data = data)

summary(model)

Call:

lm(formula = Health.Spending ~ -1 + Leisure.Spending + Spending.on.education,

data = data)

Residuals:

Min 1Q Median 3Q Max

-94.492 -21.594 0.676 30.622 97.348

Coefficients:

Estimate Std. Error t value Pr(>|t|)

Leisure.Spending 0.45634 0.05814 7.849 2.12e-12 \*\*\*

Spending.on.education 0.46374 0.07811 5.937 2.98e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42.82 on 118 degrees of freedom

Multiple R-squared: 0.9609, Adjusted R-squared: 0.9602

F-statistic: 1450 on 2 and 118 DF, p-value: < 2.2e-16

Temos estimativas com p-value abaixo de 0,05%, um R² alto de 96%, que de acordo com a análise estatística é um forte indício de multicolinearidade.