Estatística Aplicada às Ciências e Engenharia - 2014/2015
Regressão linear múltipla

1. Considere de novo os dados das galáxias da última folha.
   (a) Encontre um modelo linear para prever a velocidade à custa das outras variáveis.
   (b) Teste se alguma das variáveis pode ser removida do modelo.

2. Os resultados de uma análise de regressão linear de Y em $x_1, x_2, \ldots, x_p$, podem ser colocados numa tabela ANOVA do tipo:

Source	Sum of Squares	d.f.	Mean Square	F-test
Regression	SSR	p	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$
Residuals	SSE	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$	

Considere as seguintes tabelas ANOVA para um problema de regressão com seis variáveis explicativas (1a tabela) e com apenas duas destas seis, x_1 e x_3, na 2a tabela. O tamanho da maostra é $n = 30$.

Source	Sum of Squares	d.f.	Mean Square	F-test
Regression	3147.97	6	524.661	10.5
Residuals	1149.00	23	49.9565	

Source	Sum of Squares	d.f.	Mean Square	F-test
Regression	3042.32	2	1521.1600	32.7
Residuals	1254.65	27	46.4685	

   (a) Teste a hipótese dos coeficientes $\beta_1, ..., \beta_6$ serem todos zero e escreva os modelos correspondentes às hipóteses.
   (b) Teste a hipótese dos coeficientes $\beta_2, \beta_4, \beta_5, \beta_6$ serem todos zero e escreva os modelos correspondentes às hipóteses.
   (c) Podemos concluir que a remoção das variáveis x_2, x_4, x_5, x_6 não afeta significativamente o poder explicativo do modelo?
   (d) Calcule os coeficientes de correlação múltipla, de determinação e de determinação ajustado para ambos os modelos. Diga que modelo prefere, com base no coeficientes de determinação ajustado.

(e) * Diga como se poderia testar a hipóteses $H_0 : \beta_1 = \beta_3$ e $\beta_2 = \beta_4 = \beta_5 = \beta_6 = 0$

3. Considere de novo os dados das galáxias. Teste se a variável com o menor valor, em módulo, da estatística t pode ser removida do modelo linear. Verifique que o quadrado do valor dessa estatística coincide com o valor dado pela tabela anova para testar essa mesma hipótese.

4. Considere um dos exercícios da folha anterior sobre regressão linear simples e adeque agora um modelo polinomial de grau 3 aos dados. De seguida desenhe esse modelo assim como as bandas de confiança e de predição. Finalmente teste se o modelo linear é adequado.

5. A tabela ANOVA seguinte mostra o resultado de uma regressão linear simples de Y em x para 20 observações, com alguns números em falta. Complete os números que faltam na tabela e depois calcule $Var(Y)$ e $Var(x)$

Source	Sum of Squares	d.f.	Mean Square	F-test
Regression	1848.76	-	-	-
Residuals	-	-	-	

Variable	Coefficient	s.e.	t-test	p-value
Constant	-23.4325	12.74	-	0.0824
x	-	0.1528	8.32	< 0.0001

$n =$ - $\quad R^2 = -$ $\quad R_a^2 = -$ $\quad \hat{\sigma} = -$

6. * Prove que o estimador MMM de β satisfaz $Var(\hat{\beta}) = (X^T X)^{-1}\sigma^2$

7. * Na regressão Ridge, β_0 é um caso à parte, pois não faz sentido penalizar β_0. Deve-se começar por centrar os dados, isto é, $x_{ij} \leftarrow x_{ij} - \overline{x}_j$, e estima-se β_0 por $\overline{y}$. De seguida consideram-se apenas p variáveis e aplica-se regressão Ridge. Mostre que os estimadores de Regressão Ridge podem ser obtidos pelo MMQ usual, com a matriz de dados centrada e aumentada pelas linhas $\sqrt{\lambda}$I e y aumentado de 'p zeros.

8. Escolha, por exemplo em http://archive.ics.uci.edu/ml/datasets.html , um conjunto de dados de regressão e aplique os vários métodos de regressão linear múltipla que aprendeu e compare-os.