

# **REGRESIÓN APLICADA USANDO R**

**Edgar Acuña Fernandez**

**Departamento de Ciencias Matemáticas**

**Universidad de Puerto Rico**

**Recinto Universitario de Mayaguez**

**Enero 15, 2015**

**©2015, Derechos reservados por Edgar Acuña. Prohibida su reproducción sin  
permiso del autor**

## PREFACIO

La razón principal de escribir este libro es la carencia de un texto completo de regresión que cubra las diversas técnicas de regresión, especialmente aquellas que han tomado auge en la última década. Un par de buenos libros de regresión son el “Classical and Modern Regression with applications” de Myers (2000) y el “Applied linear Regression” de Weisberg (2005), pero ambos cubren muy poco material u omiten temas importantes en regresión tales como: selección de variables, regresión logística, regresión robusta y la muy importante área de regresión no paramétrica. Existen por otro lado buenos textos cubriendo solamente Regresión Robusta como el “Robust Regression and Outlier Detection” de Rousseeuw y Leroy (2005) y otros que tratan exclusivamente Regresión no paramétrica como el “Applied Nonparametric Regression” de Härdle (1994). El objetivo de este texto es cubrir la parte más transcendental de los libros antes mencionados.

En el transcurso de los quince años que he venido desarrollando el texto he usado varios programas estadísticos tales como: MINITAB, SAS, MATLAB, S-PLUS y últimamente R. La meta final es desarrollar todo el texto usando el programa gratuito R, el cual está disponible en [www.r-project.org](http://www.r-project.org). Aún quedan en el texto algunas salidas de MINITAB. Las salidas de SAS, MATLAB y S-Plus han sido prácticamente eliminadas.

Aunque el texto es en regresión aplicada también se ha tratado de probar varias identidades y propiedades de estimadores que aparecen en regresión. Sin embargo no es nuestra intención llenar el texto con demostraciones teóricas. Dos buenos textos donde se ve el lado teórico de Regresión son “Linear Regression Analysis” de Seber (2003) y “Linear Statistical Inference and its Applications” de Rao (2008).

El texto está organizado en 9 capítulos. El primer capítulo se enfoca en regresión lineal simple y el segundo en regresión lineal múltiple. En el tercer capítulo se discute los diversos métodos de diagnosticar si las suposiciones del modelo de regresión se están cumpliendo o no. En el capítulo 4 se estudian diferentes transformaciones que se pueden hacer de las variables predictoras y de la variable de respuesta con la finalidad de mejorar el modelo de regresión para que haga un mejor ajuste de los datos. En el capítulo 5 se discute modelos de regresión considerando la presencia de variables categóricas. Aquí se incluye el estudio de la regresión logística. El capítulo 6 está dedicada al importante problema de selección de variables en regresión y en el problema 7 se discute la forma de detectar y resolver el problema de multicolinealidad entre las variables predictoras. Los capítulos 8 y 9 están dedicados a regresión robusta y regresión no paramétrica respectivamente.

Los conjuntos de datos que aparecen en este texto pueden ser obtenidos en el siguiente sitio de la internet en <http://academic.uprm.edu/eacuna/class6205.html>.

Finalmente, deseo agradecer la ayuda de mis pasados asistentes de investigación por colaborar conmigo en la depuración de errores presentes en el texto, así como en la edición de algunos capítulos y en la preparación de las transparencias del texto.

Por favor para reportar cualquier sugerencia o error mandarme un e-mail a [edgar.acuna@upr.edu](mailto:edgar.acuna@upr.edu).

Mayagüez, Febrero 14, 2018

# CONTENIDO

1	Regresión lineal simple.	1
1.1	Introducción.	1
1.1.1.	Usos del Análisis de Regresión.	5
1.2	El modelo de Regresión Lineal Simple.	5
1.2.1	Estimación de la línea de regresión usando mínimos cuadrados.	6
1.2.2	Interpretación de los coeficientes de regresión estimados.	9
1.2.3	Propiedades de los estimadores minimos cuadraticos de regression.	9
1.2.4	Distribución de los estimadores minimos cuadraticos.	11
1.2.5	Propiedades de los residuales.	11
1.2.6	Estimación de la varianza del error	12
1.2.7	Descomposición de la suma de cuadrados.	13
1.2.8	El coeficiente de Determinación $R^2$	16
1.3	Inferencia en Regresion Lineal Simple.	16
1.3.1	Inferencia acerca de la pendiente y el intercepto usando la prueba t.	17
1.3.2	El análisis de Varianza para regresión lineal simple.	20
1.3.3	Intervalo de predicción e intervalo de confianza para el valor medio de la variable de respuesta.	21
1.4	El coeficiente de Correlación	24
1.5	Análisis de Residuales.	27
1.5.1	Cotejando Normalidad en los errores y detectando outliers.	28
1.5.2	Cotejando que la varianza sea constante.	30
1.5.3	Cotejando si los errores están correlacionados.	32
2	Regresión Lineal Multiple.	41
2.1	Introducción.	41
2.2	El Modelo de Regresión lineal multiple.	46
2.2.1	Estimación de B por minimos cuadrados.	47
2.2.2	Propiedades del estimador $\hat{\beta}$	48
2.2.3	Estimación de la varianza $\sigma^2$	49
2.3	Inferencia en regresión lineal múltiple.	51
2.3.1	Prueba de hipotesis acerca de un coeficiente de regresión individual	51
2.3.2	Prueba de Hipótesis de que todos los coeficientes de regresión sean ceros.	52
2.3.3	Prueba de hipótesis para un subconjunto de coeficientes de regresión.	54
2.3.4	Intervalo de Confianza y de Predicción en Regresión Lineal Múltiple.	56
2.3.5	La prueba de Falta de Ajuste.	57
3	Anomalías en regresión y medidas remediales.	64
3.1	“Outliers”, puntos de leverage alto y valores influenciales.	64
3.2	Residuales y detección de outliers”	67
3.2.1	Media y Varianza del vector de residuales.	67
3.2.2	Residuales Estudentizados internamente.	68
3.2.3	Residuales Estudentizados externamente.	70
3.3	Diagnósticos para detectar “outliers” y puntos de leverage alto.	75
3.4	Plot de Residuales para detectar el efecto de variables y casos influenciales	79

3.5	Plot de Residuales para detectar Normalidad.....	82
3.6	Detectando varianza no constante.....	85
3.7	Errores correlacionados en regresión.....	84
4	Transformaciones en Regresión.....	92
4.1	Transformaciones para linealizar modelos.....	92
4.2	Transformaciones para estabilizar la varianza .....	95
4.3	Transformaciones de las variables predictoras en regresión multiple .....	98
4.4	Transformaciones para mejorar la normalidad de la variable de respuesta.....	104
4.5	Mínimos cuadrados ponderados.....	108
4.6	Mínimos cuadrados generalizados .....	113
5	Regresión con variables cualitativas.....	117
5.1	Regresión con variables predictoras cualitativas.....	117
5.1.1	Regresión con una sola variable cualitativa.....	117
5.1.2	Comparando las líneas de regresión de mas de dos grupos.....	121
5.2	Regresión Logística.....	121
5.2.1	Estimación del modelo logístico.....	126
5.2.2	Medidas de confiabilidad del modelo.....	127
5.2.3	Medidas influenciales para regresión logística.....	128
5.2.4	Uso de regresión logística en clasificación.....	132
6	Selección de variables en Regresión.....	136
6.1	Métodos “Stepwise”.....	136
6.1.1	“Backward Elimination” (Eliminación hacia atrás).....	136
6.1.2	“Forward Selection” (Selección hacia adelante).....	137
6.1.3	“Stepwise Selección” (Selección Paso a Paso).....	138
6.2	Método de los mejores subconjuntos.....	142
6.3	Criterios para elegir el mejor modelo.....	142
6.3.1	El coeficiente de Determinación $R^2$ .....	142
6.3.2	El $R^2$ ajustado.....	143
6.3.3	La varianza estimada del error ( $s^2$ ).....	143
6.3.4	$C_p$ de Mallows.....	143
6.3.5	PRESS. Suma de cuadrados de Predicción.....	146
6.3.6	Validación Cruzada .....	148
6.3.7	AIC .....	149
6.3.8	BIC .....	154
6.3.9	Validación cruzada Generalizada.....	156
6.3.10	Otros Criterios.....	157
6.3.11	Recomendación para elegir el mejor modelo.....	157
6.4	Otros métodos de selección de variable.....	158
6.4.1	Métodos basados en remuestreo.....	158
6.4.2	Métodos basados en regresión penalizada.....	158
6.4.3	Métodos Bayesianos.....	159
6.4.4	Algoritmos Genéticos.....	159
7.	Multicolinealidad .....	161
7.1	Multicolinealidad.....	161
7.1.1	Efectos de Multicolinealidad.....	161
7.1.2	Diagnósticos de Multicolinealidad.....	163

7.1.3	Medidas remediales al problema de multicolinealidad.....	166
7.2	Regresión Ridge.....	166
7.2.1	Aplicación de Regresión Ridge a Selección de variables.....	174
7.3	Componentes principales para Regresión.....	176
8	Regresión Robusta.....	186
8.1	Introducción.....	186
8.2	Regresión L1.....	187
8.3	Regresión M.....	190
8.3.1	Cálculo de los estimadores M de regresión.....	198
8.4	Regresión GM o Regresión de Influencia acotada.....	201
8.5	Regresión de Medianas de Cuadrados Mínima.....	202
9	Regresión Noparamétrica.....	206
9.1	Introducción.....	206
9.2	Suavización bivariada o Suavizadores de diagramas de puntos .....	207
9.2.1	El regresograma.....	207
9.2.2	“Running Means” y “Running Lines”.....	208
9.2.3	Suavizador por los k vecinos más cercanos.....	210
9.2.4	Suavización por kernels.....	210
9.2.5	Regresión local ponderada, LOWESS.....	212
9.2.6	Regresión Polinomial .....	213
9.2.7	Regresión por Splines.....	215
9.2.8	Suavización por Splines.....	217
9.3	Suavización multidimensional.....	220
9.3.1	Modelos Aditivos generalizados, GAM.....	220
9.3.2	Regresión usando árboles de decisión (CART).....	222
	Apéndice A: Revisión de Matrices.....	231
	Apéndice B: Laboratorios en R para Regresión Aplicada.....	240
	Referencias.....	265