Regresión lineal múltiple para estimar esfuerzo en completar proyectos de desarrollo de software

Mitsiu Alejandro Carreño Sarabia
E23S-18014

Introducción

La correcta estimación del esfuerzo requerido para desarrollar un proyecto de software es f**undamental para una planeación y estimación** de costo e inversión de un proyecto.

La investigación se realizó en una empresa multinacional mediana que estaba implementando el modelo Capability Maturity Model Integration (CMMI) en nivel de madurez 2 (informal).

Link de publicación:
https://www.researchgate.net/profile/Leonor-Teixeira/publication/250308495_Software_Effort_Estimation_with_Multiple_Linear_Regression_Review_and_Practical_Application/links/00b4953c7a07f74147000000/Software-Effort-Estimation-with-Multiple-Linear-Regression-Review-and-Practical-Application.pdf

Introducción

Fallar en la estimación de esfuerzo tanto en sobreestimación como subestimación tiene consecuencias graves en el resultado del proyecto. Subestimar provoca **retrasos en las entregas y baja calidad**. Sobreestimar puede provocar la pérdida de clientes, así como una **distribución ineficiente de recursos**.

Introducción

Usualmente la estimación se requiere al inicio de la fase de desarrollo, lo cual vuelve más compleja la estimación en la que se deben tomar en cuenta factores como:

- **Establecer el alcance del proyecto**
- **Establecer la capacidad técnica base**
- **Prever riesgos y flujos de trabajo**

Introducción

Finalmente una buena estimación debe tomar en cuenta factores como:

- Eventos inesperados
- Tareas pasadas por alto
- Cambios solicitados por el cliente
- Problemas por falta de recursos
- Especificación de requisitos superficiales

Metodología

Para la investigación se eligió la regresión lineal múltiple debido a que la empresa analizada tenía r**egistros históricos (2 años) tanto de las estimaciones como del desarrollo verdadero** del proyecto.

Recordemos que la regresión lineal múltiple nos permite **descubrir el grado de dependencia** de una variable (y) respecto a un conjunto de variables independientes (xi).

Metodología

La regresión lineal múltiple se considera útil[1] cuando se cubren los siguientes criterios:

- La cantidad de ejemplos (casos) es significativamente mayor a la cantidad de parámetros a estimar.
- La información tiene un comportamiento estable.
- Existe poca o nula información faltante
- Una pequeña cantidad de variables independientes son suficientes para linealmente predecir valores de salida (independientemente de si es necesario transformarlas).

1. A. R. Gray and S. G. MacDonell, “A comparison of techniques for developing pre- dictive models of software metrics,” Information and Software Technology, Vol. 39, 1997, pp. 425-437.

Contexto

Históricamente la empresa realizaba sus estimaciones en el juicio de un experto, pero al aplicar CMMI es requerido establecer fundamentación en la estimación.

Históricamente la estimación se realizaba tomando en cuenta 4 unidades las cuales se distribuían, prototipado 6%, desarrollo 64%, pruebas 25%, documentación 5%

Para el análisis se desestimó el prototipado y la documentación dado su bajo peso.

Datos

Acrónimo	Descripción de variables
Dev_Eff	Horas efectivas de desarrollo
Dev_Frc	Horas estimadas de desarrollo
QA_Eff	Horas efectivas de testing
QA_Frc	Horas estimadas de testing
Nr_Req	Número de requisitos
Nr_CRs	Número de cambios de requisitos
Nr_Moduls	Número de módulos afectados por Nr_CRs
Prot	Booleano, indica si se debe prototipar
Code_Complex	Variable ordinal que indica la complejidad (1-3)

Estadística Descriptiva

	Equipo desarrollo		Equipo testing	
	Dev_Eff	Dev_Frc	QA_Eff	QA_Frc
Promedio	**67.69**	**57.45**	**21.68**	**20.22**
Desviación estándar	72.49	58.46	21.55	20.02
Mínimo	3.00	3.00	1.00	2.00
Máximo	363.50	280.00	120.00	105.00
Mediana	38.25	35.00	14.00	14.00

Regresión Lineal Múltiple

Modelo	Tam Muestra	R^2	Variable Dependiente (Y)		Variable Independiente (X_n)				
			Dev_ Eff	ln(Dev_Eff)	Prot	Nr_ Req	Code_ Complex	Nr_CRs	Nr_ Moduls
1	106	0.493	X		X		X	X	
2	106	0.382		X		X	X	X	X
3	89	0.508	X		X			X	X
4	**89**	**0.547**		**X**	**X**		**X**	**X**	
5	84	0.537	X		X	X		X	
6	84	0.524	X		X	X	X	X	
7	84	0.524		X	X	X	X		
8	84	0.358		X		X	X	X	
9	76	0.429		X	X	X	X		
10	69	0.092	X					X	

Regresión Lineal Múltiple

La varianza en el tamaño de muestra es causada a la eliminación sucesiva de outliers.

El valor R^2 corresponde a el porcentaje de precisión del modelo.

Implementación en python

Conclusiones personales

Es un **problema interesante**.

Las variables seleccionadas (dev_eff, dev_frc, nm_req, etc) me parecieron una **buena elección de variables**.

Creo que hay factores que sería interesante agregar, como cantidad de programadores, antigüedad/experiencia, aparte **se dejó de lado factores externos** como retraso de insumos, y fase de soporte.

Creo que el hecho de que comenzaran a filtrar variables en sus modelos escaló la **complejidad combinatoria del proyecto**.

Sería interesante analizar en los datos ground truth **cuáles variables tenían una relación lineal con la estimación de esfuerzo desde el principio**.