Problem

In logistic regression model, B_1 and B_2 are coefficients of two predictors not statistically significant α = 0.05. When you remove one of the predictors, the remaining one becomes statistically significant. How can this happen?

Tips

Question Type: Statistics
Where: Phone Screening

> First, break-down your understanding of how beta coefficients in a logistic regression models are evaluated for statistical significance.
> Consider, a modeling condition that would lead to change in the statistical significance of the model.

Solution

Let's unbox the steps required to evaluate the statistical significance of a predictor in logistic regression.

When P-value is less than or equal to alpha, let's say 0.05, the predictor is statistically significant. Otherwise, there is no statistical significance.

But what determines P-value? It's the value of Wald statistic:

$\omega_i = \frac{\beta_i}{SE[\beta_i]}$
p-value $= pr(W \geq |\omega_i|)$

Here are some patterns on Wald statistic (↑ = increase; ↓ = decrease):

1. β ↑, ω ↑, *p-value* ↓
2. β ↓, ω ↓, *p-value* ↑
3. $SE[\beta]$ ↑, ω ↓, *p-value* ↑
4. $SE[\beta]$ ↓, ω ↑, *p-value* ↓

You just observed how the statistical significance of a predictor is evaluated. Now, can you explain what's causing the change in the statistical significance when you include two predictors versus just one of them?

Here's the hint - it's multicollinearity.

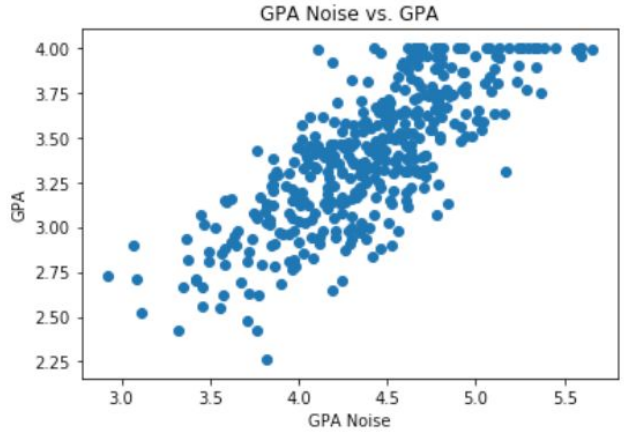
Think about what the standard error of coefficient conveys. It's the error estimating the true value of a coefficient. Intuitively, when you have two predictors highly-correlated with one another, how does the model know which one of the two is driving the effects observed in the binary outcome? It's not possible. That's why two predictors, highly-correlated, are in the model, the standard errors of the coefficients will be much higher than that of the coefficient observed in a model with just one of the predictors, excluding the other.

As noted above, when standard error is high, Wald statistic decreases; thereby, P-value increases, which increases the likelihood that there are no statistical significance in the effects of highly-correlated predictors

Consider this simulation:

A logistic regression model was fitted on a college admission data, containing GRE, GPA, and rank attribute.

A new predictor was generated by adding noise, generated from a standard normal distribution, to the GPA predictor. The two predictors are highly correlated at 0.78 based Pearson correlation.



Two models were fitted - one without the GPA-noise and one with it. Compare the model summaries between the two.

Model #1: $logit[y] = B_1 * GPA + B_2 * GRE + B_3 * Rank + Intercept$

Dep. Variable:	admit	**No. Observations:**	400
Model:	Logit	**Df Residuals:**	396
Method:	MLE	**Df Model:**	3
Date:	Wed, 08 May 2019	**Pseudo R-squ.:**	0.08107
Time:	22:18:51	**Log-Likelihood:**	-229.72
converged:	True	**LL-Null:**	-249.99
		LLR p-value:	8.207e-09

	coef	std err	z	P>|z|	[0.025	0.975]
gre	0.0023	0.001	2.101	0.036	0.000	0.004
gpa	0.7770	0.327	2.373	0.018	0.135	1.419
rank	-0.5600	0.127	-4.405	0.000	-0.809	-0.311
intercept	-3.4495	1.133	-3.045	0.002	-5.670	-1.229

Note that GPA is statistically significant at alpha 0.05.

Model #2: $logit[y] = B_1 * GPA + B_2 * GRE + B_3 * Rank + B_4 * GPA(Noise) + Intercept$

Dep. Variable:	admit	**No. Observations:**	400
Model:	Logit	**Df Residuals:**	395
Method:	MLE	**Df Model:**	4
Date:	Wed, 08 May 2019	**Pseudo R-squ.:**	0.08275
Time:	22:42:20	**Log-Likelihood:**	-229.30
converged:	True	**LL-Null:**	-249.99
		LLR p-value:	2.249e-08

	coef	std err	z	P>|z|	[0.025	0.975]
gre	0.0023	0.001	2.124	0.034	0.000	0.004
gpa	0.4397	0.492	0.894	0.371	-0.524	1.404
rank	-0.5734	0.128	-4.467	0.000	-0.825	-0.322
intercept	-3.7661	1.187	-3.173	0.002	-6.092	-1.440
gpa_noise	0.3374	0.369	0.915	0.360	-0.385	1.060

When GPA-noise is included in the model, standard error of GPA increases from 0.327 to 0.492, and P-value increases from 0.018 to 0.371, which means that there is no statistical significance.