Section 4E. Hypothesis Testing in Logistic Regression

Statistics for Data Science

Victor M. Preciado, PhD MIT EECS
Dept of Electrical & Systems Engineering
University of Pennsylvania
preciado@seas.upenn.edu

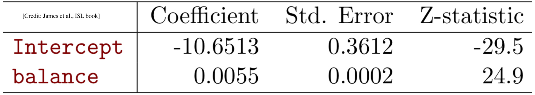
Hypothesis Testing in Logistic Regression (single variable)

How do we decide if the input influences the output? We can use the hypothesis testing framework...

1. If $\beta_1 = 0$ (i.e., the input does not influence the output), then $\widehat{\beta}_1 \sim \mathcal{N}\left(0, \mathrm{SD}(\widehat{\beta}_1)^2\right)$
2. Define the variable $z_1 = \widehat{\beta}_1 / \mathrm{SD}(\widehat{\beta}_1)$. Hence, we have that (see Hypothesis Testing section)

$$\Pr\left(|z_1| < 2\right) = 0.95$$

3. Therefore, if $|z_1| > 2$, we have statistical evidence to reject the hypothesis that the input does not influence the output.



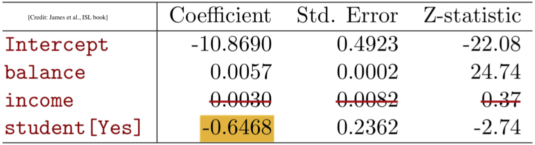
[Credit: James et al., ISL book]	Coefficient	Std. Error	Z-statistic
Intercept	-10.6513	0.3612	-29.5
balance	0.0055	0.0002	24.9

Hypothesis Testing with Multiple Variables

When $p > 1$ we can use a *multivariate logistic function*, defined below:

$$p_1\left(\mathbf{x};\beta_0,\beta_1,\ldots,\beta_p\right)=\frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}$$

- We can find estimates $\widehat{\beta}_0,\widehat{\beta}_1,\ldots,\widehat{\beta}_p$ using the Maximum Likelihood Criterion
- Predictions are made using the function $\widehat{p}_1\left(\mathbf{x}\right)=p_1\left(\mathbf{x};\widehat{\beta}_0,\widehat{\beta}_1,\ldots,\widehat{\beta}_p\right)$
- We can decide if a particular input is not relevant looking at the Z-values



[Credit: James et al., ISL book]	Coefficient	Std. Error	Z-statistic
Intercept	-10.8690	0.4923	-22.08
balance	0.0057	0.0002	24.74
income	~~0.0030~~	~~0.0082~~	~~0.37~~
student[Yes]	-0.6468	0.2362	-2.74

Penn Engineering