

A Comparison of Linear & Logistic Regression

	Linear Regression (<i>regress</i>)	Logistic Regression (<i>logit & logistic</i>)
Dependent Variable/Outcome	Y = a continuous variable	P = a dichotomous variable
Equation	$Y = b_0 + b_1X_1 + b_2 X_2 \dots b_k X_k + E$	$\ln[p/(1-p)] = b_0 + b_1X_1 + b_2 X_2 \dots b_k X_k + E$ (p = probability of outcome event)
Regression Coefficient	Amount of expected (average) change in the DepVar (Y) per unit increase in the IndVar (X's).	Coefficient allows you to calculate the odds ratio for a one point increase in the IndVar (X's).
Key Assumptions	<ol style="list-style-type: none"> 1. Measures of Y are independently and randomly sampled. 2. Linear association between Y and the X's. 3. Normally distributed residuals with homoskedastic variance. 4. Y is measured without error. 5. All potentially relevant X's are in the model and all X's in the model are relevant. 6. The X's are not linear combinations of each other (i.e., no multicollinearity). 	<ol style="list-style-type: none"> 1. Measures of Y are independently and randomly sampled 2. Probabilities of Y is a logit function of the X's. 3. Y is measured without error 4. All potentially relevant X's are in the model and all X's in the model are relevant. 5. The X's are not linear combinations of each other (i.e., no multicollinearity).
Discrimination / Predictiveness Statistic	R^2 (<i>regress</i>)	C Statistic (AUROC curve)(<i>lroc</i>)
Calibration Assessment	Residual Plots <pre>. regress depvar x1 x2 x3 . capture rypplot, saving(graph1, replace) . predict yhat . drop if yhat == . . sort yhat . gen cat5 = int(5*_n/(_N + 1)) . bysort cat5: sum depvar yhat</pre>	Hosmer-Lemeshow Tables <pre>. logit depvar x1 x2 x3 . predict phat . drop if phat == . . sort phat . gen cat10 = int(10*_n/(_N + 1)) . graph bar (mean) phat depvar, by(cat10, rows(1))</pre>

Checklists for Linear & Logistic Regression

	Linear Regression	Logistic Regression
Reading Regression Output	<ol style="list-style-type: none"> 1. Did I lose sample size? (Compare # observations in regression model to number of non-missing values for Y) 2. Is my overall model significant? (Check the F test) 3. How predictive (discriminate) is my model? (Check the R² and examine <i>rvf</i> plots) 4. Which X's are significantly associated with Y? 5. What is the magnitude of these associations? (Check the regression coefficients & examine adjusted values for \hat{Y}) 6. Which X's contribute the most is predicting Y? (Check the standardized "beta" coefficient) 	<ol style="list-style-type: none"> 1. Did I lose sample size? (Compare # observations in regression model to # of non-missing values for Y) 2. Is my overall model significant? (Check the LR χ^2 test) 3. How predictive (discriminate) is my model? (Check the C-statistic [AUROC curve] & examine Hosmer-Lemeshow tables) 4. Which X's are significantly associated with Y? 5. What is the magnitude of these associations? (Check the odds ratios & examine adjusted values for <i>phat</i>)
Questions to ask About Your Multivariable Model	<ol style="list-style-type: none"> a) Have I included all available confounders in the model? b) Which variables may be pathways or mediators, rather than confounders? c) How does the addition or subtraction of individual variables or hierarchical domains (ie., sociodemographics, health status measures, attitudes and behaviors, etc..) affect the model and associations? d) What does it all mean? e) Are there ways to test my interpretation of the data by conducting additional analyses? f) Are there interaction terms that should be evaluated (if you have adequate power to look at more variables)? 	
Regression Diagnostics	<ol style="list-style-type: none"> 1. Visual inspection of the <i>rvfplot</i>!!! (provides info on discrimination, linearity, heteroskedasticity, outliers and calibrations) 2. Check for multicollinearity (check <i>vif</i>) 	<ol style="list-style-type: none"> 1. Visual inspection of the Hosmer-Lemeshow tables/graphs!!! (Provides info on both discrimination and calibrations) 2. Check for multicollinearity (check <i>vif</i>)
Problems & approaches	<ol style="list-style-type: none"> 1. For loss of sample size – Impute 2. For multicollinearity – exclude a duplicative variable or combine correlated variables into a scale. 3. For non-linear fit of residuals – transformation of Y, evaluate complex forms of influential X's (e.g., quadratics), interactions, omitted variables. Last resort in severe instances is to dichotomize Y and do logistic regression 4. For outliers – error check values, exclude the observation only if it is felt the result is due to measurement error extreme or unusual value of the variable (the absolute value for an X or Y) 	<ol style="list-style-type: none"> 1. For loss of sample size – Impute 2. For multicollinearity – exclude a duplicative variable or combine correlated variables into a scale. 3. For poor calibration –evaluate complex forms of influential X's (e.g., quadratics), interactions, omitted variables. A poorly calibrated model should not be used for prediction of adjustment. 4. For outliers – error check values, exclude the observation only if it is felt the result is due to measurement error extreme or unusual value of X