Biostatistics for Health Care Providers: A Short Course

Multiple Linear Regression and Logistic Regression

Patrick Monahan, Ph.D.
Division of Biostatistics
Department of Medicine

Objectives

- Review Simple Linear Regression
- Describe Assumptions, Interpretations, and Model Checking for Multiple Linear Regression
- Describe Assumptions, Interpretations, and Model Checking for Logistic Regression
- Discuss Model Selection Procedures

Review Simple Linear Regression

- 1 Independent variable (IV).
- 1 Dependent variable (DV).
- 1 IV used to predict 1 DV.

- IV also called predictor or explanatory variable, or Covariate.
- IV is categorical or continuous.

- DV also called outcome or response variable.
- DV is a continuous numerical variable.
- Formula:

$$y = \alpha + \beta_1 x_1 + \varepsilon$$

Multiple Linear Regression

- More than 1 IV.
- 1 DV.
- More than 1 IV used to predict 1 DV.

- IVs are categorical or continuous.
- DV is a continuous numerical variable.

- Formula:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

Assumptions

- Assumptions of linear regression pertain to distribution of the error terms, or equivalently, the distribution of the dependent variable (y).

<u>Assumptions of multiple linear regression</u>

- The errors (or y's) are independent of one another.

- For any fixed combination of the x's, the errors (or y's) are normally distributed with variance σ^{2}.

- The variance of errors (or y's) is the same for any fixed combination of the x's (i.e., constant variance). (Homogeneity of variance, or homoscedasticity).

Assumptions: A subtle but important point

- Assumptions apply to population.
- If we knew population data satisfied assumptions, we could validly perform tests even if our sample data appeared to violate assumptions.
- However, we usually don't know distribution of population, so we use our sample data to assess whether population data satisfies assumptions.

Data

- Sample: n = 1039 women non-adherent to breast cancer screening guidelines.

- Intervention study: women randomly assigned
  - intervention promoting mammography or
  - “treatment as usual”.

- This example: we attempt to predict women’s perceived barriers to obtain mammography from their age, education and race.

Data: Dependent variable (outcome)

Perceived barriers to mammography screening

- Measured 2-months post-intervention.
- 16-item scale.
- Total score: sum of 16 items, with range 16-80.

- Example Items:
- Having a mammogram is embarrassing for you.
- You don't have the time to get a mammogram.
- Cost would keep you from having a mammogram.
- (1) = strongly disagree, (5) = strongly agree

- Higher total score = worse = greater perceived barriers.

Data: Independent variables

- Independent variables (predictors or covariates):
  - Age, in years.
  - Education (years of school completed).
  - Race
    - 1 = Caucasian,
    - 0 = other, mostly African American.

Selected SAS output

- Overall model: $F = 16.58$, $df = 3$ and 1035, $p < .0001$
- $R^2 = .046$ Adjusted $R^2 = .043$
- Least Squares Estimates of Parameters

Variable	df	Estimated Regression Coefficient	Estimated Standard Error	t Value	Pr > |t|
Intercept	1	26.630	2.277	11.69	<.0001
Age	1	0.087	0.026	3.28	.0024
Education	1	-0.265	0.088	-3.00	.0011
Race	1	3.218	0.559	5.76	<.0001

Interpretation: Statistical Significance of Overall model

- The linear combination of age, education, and race is highly statistically significant in predicting perceived barriers. How do we know this?
- Examine F test for significance of overall model:
- $F = 16.58$, $df = 3 \& 1035$, $p < .0001$).
- However, in large samples, statistical tests have high power to detect small, clinically unimportant, magnitudes of relationships.
- Thus, it is important to report effect sizes along with statistical tests.
- Effect size = index that provides the magnitude or strength of a relationship (i.e., practical importance).

Interpretation: Practical or clinical importance of Overall Model

- R^2 is an effect size, indicating the strength of overall model, commonly used in multiple linear regression.
- R^2 measures the strength of the *combined* effect of age, education and race in predicting barriers.
- R^2 = Coefficient of multiple determination.
- R^2 = proportion of the total variation in y that is explained by the linear combination of x1, x2, x3.
- R^2 = .046 = 4.6% is small.
- However, interpretation of R^2 as small, medium or large depends on context.

Interpretation: Adjusted R^2

- Adjusted R^2 = Coefficient of multiple determination adjusted down for number of predictors in the model.
- Use Adjusted R^2 to compare models with different number of predictors.
- Adjusted R^2 = .043 = 4.3%.

Interpretation: Statistical Significance of Individual Covariates

- The two-sided partial t test is a test of significance of each predictor after adjusting for the effect of other predictors in model:
  - $H_0 : \beta_{age} = 0$, adjusted for education and race.
  - $H_A : \beta_{age} \neq 0$, adjusted for education and race.
- "Adjusted for" can also be phrased as ...
- "partialling out".
- "Controlling for".
- "Holding constant".

Interpretation: p-value from partial t test

- For partial t test, $p < .05$ for age, education and race.
- Age, education and race each add significant prediction of barriers after adjusting for each other.

- However, the effect size is small.
- How do we know?
- What is the effect size that measures the magnitude of *partial association*?
- E.g., what is the association between age and barriers after adjusting for education and race?

Interpretation: Practical or clinical importance of Individual Covariates

- Semi-partial correlation coefficient squared (r^2).
- Semi-partial r^2 = effect size to accompany partial t test.
- Proportion of total variation in y that is explained by the predictor after adjusting for other predictors in model.
- SCORR2 option in SAS REG procedure.

	Semi-partial r^2
Age	.0099 (i.e., 1.0%)
Education	.0083 (i.e., 0.8%)
Race	.0306 (i.e., 3.1%)

- E.g., Age accounts for 1% of the variation in perceived barriers after adjusting for education and race.

Interpretation: Estimated Regression Coefficients

- Age (.087): The model predicts that for every one-unit increase in age (i.e., 1 year) the perceived barriers score increases by .087, adjusted for education and race.
- Education (-.265): For every one-unit increase in Education (i.e., 1 year of school completed) the perceived barriers score decreases by .265, adjusted for age and race.
- Race (3.218): Caucasians are estimated to have a higher (worse) perceived barriers score than African Americans by 3.218, when age and education are held constant.
- Race: In other words, based on this sample, Caucasians are predicted to have a slightly higher barriers score on average than African Americans of comparable age and education, in the population.

Model Checking

- I prefer to check model fit graphically rather than with statistical tests, because ...
- Large samples yield high power for statistical tests to detect minor deviations of model from the data, i.e. minor misfit of the model.
- No model fits data perfectly.
- If you increase sample size enough, all statistical tests will eventually reject Ho of good fit of the model.
- Question is: how great is the misfit and what are the consequences?

Model Checking: graphical

- Very popular, effective graphical method:
- Examine scatter plot of fitted (i.e., predicted) values of the DV (y-hat) on the x-axis and standardized (mean = 0, SD = 1) residuals on the y-axis.
- If assumptions reasonably satisfied, should see a random scatter of points with no patterns.
- Non-normality: residuals not normally distributed.
- Non-linearity: non-random *trend* where most residuals above the zero line (Y = 0) at some predicted values and below the zero line at other predicted values.
- Non-constant variance: residuals fan out to left or right or in middle (i.e., greater spread of residuals for some fitted values than other fitted values).

Model Checking: Additional steps

- Check for outliers
- Diagnostics
- Are standard errors extremely large compared to regression coefficients?

Model Checking: Assessing Collinearity

- If two or more predictors are highly correlated with each other, could cause a problem:
- F test and R^2 for overall model are valid.
- But difficult to determine partial effects of predictors.
- Examine correlations among predictors.
- Use COLLIN option in SAS REG.
- Condition index > 30 indicates collinearity problem.

Logistic Regression (LR)

- Use instead of linear regression when outcome (y) is binary.
- Why use LR instead of linear regression?
- Practical limitation:

  Linear regression allows y-hat > 1 and y-hat < 0, which is not possible for probability of binary outcome.
- Theoretical limitation:

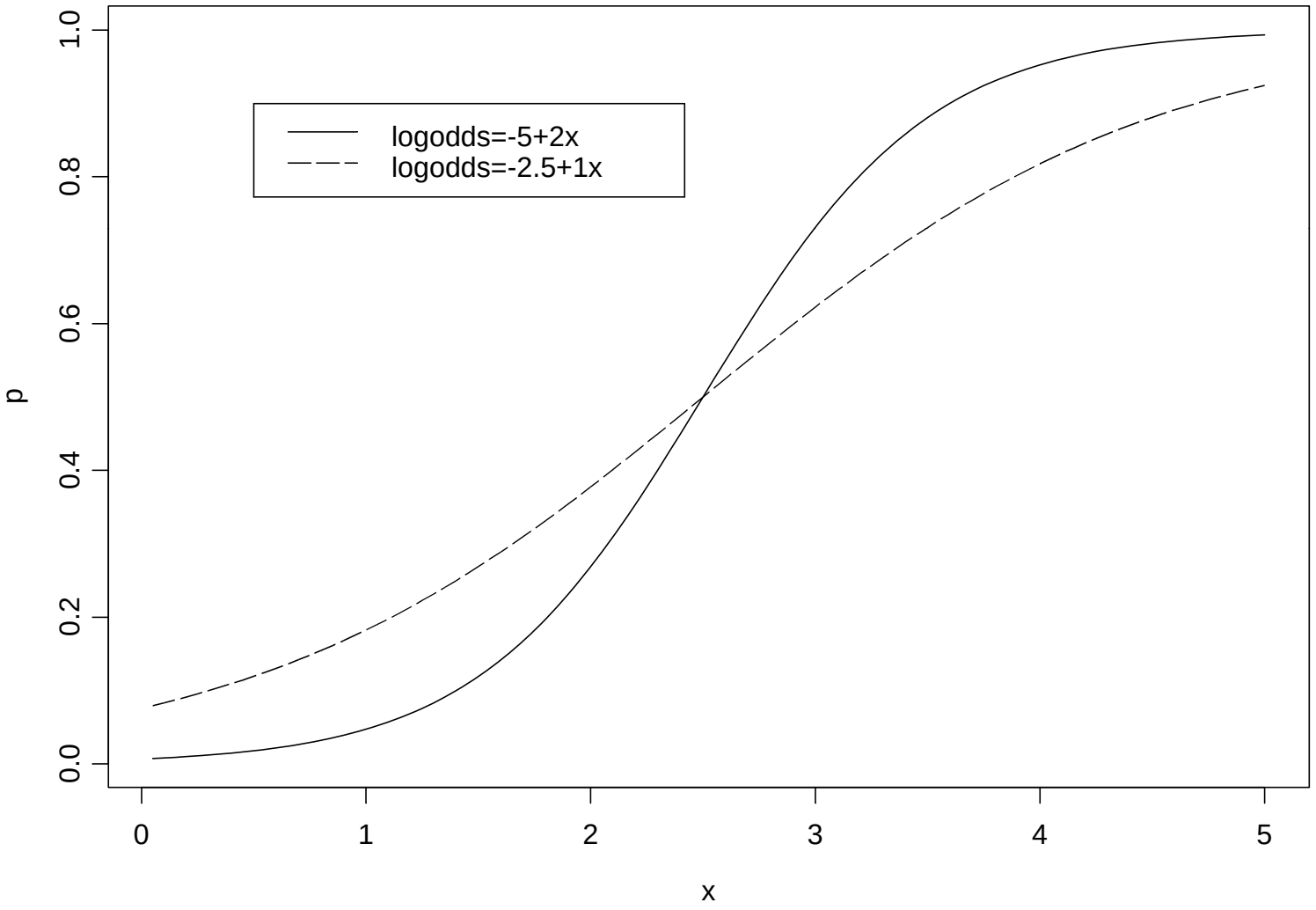
  Errors for binary outcome tend to be binomially distributed, whereas linear regression assumes errors are normally distributed.
- Simple LR (single predictor).
- Multiple LR (more than 1 predictor).

Logistic Regression Model

- Coding of outcome or DV:
  - 1 = event of interest
  - 0 = not the event of interest
- Logistic regression provides a model of the probability of the event of interest as a function of 1 or more IVs.
- Logistic function is an S-shaped curve.
- Next slide: graph of logistic function (1 continuous IV).
- Formula, using 1 IV:

$$p = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Logistic Regression: Graph of Logistic Function



Logistic regression: Fitting the model

- We linearize S-shaped logistic function by modeling the logit (i.e., log odds) of outcome:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- Log = natural log.

Data

- Same mammography data, but n = 1238.
- Now, predict binary outcome.
- Dependent variable:
  - Adherence to mammography screening (assessed at 4-months post intervention).
    - 1 = yes
    - 0 = No

- Independent variables:
  - Age, in years.
  - Education (years of formal education).
  - Race
    - 1 = Caucasian,
    - 0 = other, mostly African American.

SAS coding

- By default, SAS Logistic procedure models lowest category as event of interest.
- Therefore, if the outcome is coded (as I suggest) as 0 or 1, where 1 represents the event of interest, SAS will by default model 0 as event of interest, when in fact 1 is your event of interest.
- Recommendation:

  If your event of interest is coded 1 (as I suggest), then in order to model category 1 as event of interest ...

  (1) use DESCENDING option in SAS LOGISTIC procedure.

  or

  (2) use the EVENT='x' option (e.g., event = '1').
- Method (1) is popular, but I prefer method (2) because it is more explicit and less likely to allow mistakes.

Selected SAS output

- Overall model:
  (i.e., testing Global Null Hypothesis: $\beta_1 = \beta_2 = \beta_3$)
  Likelihood ratio chi-square = 18.442, df = 3, p = .0004

- R^2 = .015 max-rescaled R^2 = .021

- Maximum likelihood estimates of parameters

Variable	df	Estimated Regression Coefficient	Estimated Standard Error	Wald chi-square	Pr > ChiSq
Intercept	1	-0.769	0.566	1.85	.174
Age	1	-0.010	0.006	2.52	.113
Education	1	0.055	0.021	6.57	.010
Race	1	-0.162	0.132	1.52	.218

Interpretation: Statistical Significance of Overall Model

- The linear combination of age, education, and race is significant in predicting whether women obtain a mammogram by 4 months after intervention:

  Likelihood ratio chi-square = 18.442,

  df = 3,

  p = .0004.

Interpretation: Practical or Clinical Importance of Overall Model

- However, the magnitude of relationship (i.e., effect size) (here, R^2) is small (.015).
- R^2: generalized R^2 for logistic regression.

  Not exactly the same thing as the coefficient of multiple determination (R^2) in linear regression but similar interpretation as proportion of variance explained.
- Approximately 1.5% of variation in mammography screening is accounted for by the combination of age, education, and race.

Interpretation: maximum-rescaled R^2

- The maximum possible value for generalized R^2 is less than 1.0, therefore ...
- max-rescaled R^2 converts R^2 to a scale that has 1.0 as a maximum.

- Max-rescaled R^2 = .021 = 2.1% is small.
- Recommendation: I generally use the max-rescaled R^2 instead of the R^2 for logistic regression, because I know what it means when it can range from 0 to 1.

Interpretation: Statistical Significance of Individual Covariates

- The two-sided partial Wald test is a test of significance of each predictor after adjusting for the effect of other predictors in model:
  - $H_0 : \beta_{age} = 0$, adjusted for education and race
  - $H_A : \beta_{age} \neq 0$, adjusted for education and race
- "Adjusted for" can also be phrased as ...
- "partialling out".
- "Controlling for".
- "Holding constant".

Interpretation, cont.

- For partial Wald test, $p < .05$ only for education.
- Only education provides significant prediction of barriers after adjusting for the other two predictors.

Interpretation: Practical or clinical importance of Individual Covariates

- What is our effect size or measure of association for partial effects?
- Adjusted odds ratio.
- Exponential of each regression coefficient equals odds ratio of event for a 1-unit increase in predictor, adjusted for other predictors in model.

- Exp (β_{age})
- Exp ($\beta_{education}$)
- Exp (β_{race})

- Exponential = e^{x} or exp on calculator.

Interpretation: Adjusted odds ratios

Variable	df	b	Odds ratio exp(b)
Intercept	1	-0.769	
Age	1	-0.010	0.990
Education	1	0.055	1.056
Race	1	-0.162	0.850

H_0 (no relationship):

$b_{age} = 0$, or

Odds Ratio = 1.

Interpretation, Adjusted odds ratios cont.

- Age: Because age was not significant, we should not interpret the odds ratio for age since its magnitude could be due to random sampling error (we have no reason to believe the age coefficient is different from zero in the population).
- Race: Race was also not significant, so we should not interpret race coefficient. However, we do so here to demonstrate convenience of odds ratio interpretation for categorical independent variables.
- The odds ratio for obtaining a mammogram was estimated to be 0.85 times as great for Caucasians compared to African Americans of comparable age and education.

Interpretation: Practical Importance of Education

- Coefficient: The model predicts that for every one-unit increase in education (i.e., 1 year of school) the *log odds* of obtaining a mammogram within 4 months of intervention increases by .055, adjusted for age and race.
- Odds ratio: Log odds does not have much meaning to us, so we will focus on interpreting the odds ratio.
- The model predicts that for every one-unit increase in education (i.e., 1 year of school) the *odds* of obtaining a mammogram within 4 months of intervention increases by a factor of 1.056, adjusted for age and race.
- In other words, the estimated *odds* for obtaining a mammogram was 1.056 times greater for women with 1 more year of schooling compared to other women, controlling for age and race.

Interpretation: Customized Odds ratios

- For continuous covariates, odds ratios per *1-unit* increase in predictor may not be interesting.
- It may make more sense to report odds ratios for a *c-unit* increase, where *c* is 5 or 10 for example.
- Exp ($c \times$ b).
- For 5-unit increase in education:
  Exp (5 $\times$ .055) = Exp (.275) = 1.317.
- In other words, the estimated odds for obtaining a mammogram was 1.317 times greater for women who are 5 years more educated than other women, controlling for age and race.
- UNITS option in SAS LOGISTIC procedure.

Interpretation: Wald tests versus Likelihood ratio tests for statistical significance of individual covariates

- Wald chi-square for partial tests are reported by default with SAS Logistic procedure.
- Wald chi-square tests are adequate but are generally slightly conservative, even in large sample sizes.
- Statisticians prefer the likelihood ratio chi-square tests.
- In SAS LOGISTIC procedure, you must run the model with and without predictor of interest, and subtract the likelihood ratio chi-squares by hand, to obtain the likelihood ratio chi-square partial test.
- In SAS GENMOD, you can obtain likelihood ratio chi-square tests with an option.
- To perform LR with SAS GENMOD, specify logit link and binomial error distribution.

Model Checking

- As in linear regression, create scatter plot; however, ...
- because outcome binary, assess linearity of logit by
  - grouping continuous covariate.
  - Smoothing.
- Check for outliers.
- Diagnostics.
- Are standard errors extremely large compared to regression coefficients?

Model Checking: Test of Goodness of Fit of the Overall Model

- Hosmer and Lemeshow test
  - Appropriate when one or more covariates are continuous.
  - To be valid, this test needs to create more than 6 groups of predicted probabilities.
  - LACKFIT option in SAS LOGSISTIC.

- Pearson or Deviance chi-square summary statistics.
  - Appropriate when only a few categorical covariates, that is, when there are a small number of unique covariate patterns (compared to total number of observations).

Model Selection Procedures: Explanatory versus Prediction

- Explanatory example: intervention study where predict outcome from intervention group while adjusting for appropriate potentially confounding covariates.

- Prediction example: Use a set of many predictors to predict outcome.

Model Selection Procedures: Use Judgment

- Part art, part science.
- Prediction scenario:
  - Combine clinical knowledge, theory, and literature with results of statistical procedures to build a “best-predicting” model.
  - Statistical procedures must be guided wisely with clinical knowledge.

Model Selection Procedures

- Statistical procedures (linear or logistic regression)
- Forward: Start with no variables in the model and add significant terms
- Backward: Start with all variables in the model and remove non-significant terms
- Stepwise: Hybrid of forward and backward.
- All possible (i.e., “best subsets”).
  - Run all possible combinations
  - Preferred over forward, backward or stepwise.
  - By listing several top models, guards against idolizing 1 model as best.

Recommended text for logistic regression

- Hosmer, David. W., & Lemeshow, Stanley. (2000). *Applied logistic regression* (Second ed.). New York: Wiley.

  - Classic text.
  - Accessible to practitioners.

SAS coding for our two examples

- Multiple linear regression
  - proc reg;
  - model barriers = age educ race / scorr2 collin;

- Multiple logistic regression
  - proc logistic;
  - class race (param=ref);
  - model screen (event='1') = age educ race;
  - units educ = **5**;