Regression

Today's goals

- introduce linear regression models
- learn to interpret the parameters of a regression model
- Identify the assumptions of linear regression
- Share some useful protips
- introduce logistic regression (if we have time)

Definition linear regression

. . . a method that summarizes how the average values of a numerical *outcome* variable vary over subpopulations defined by linear functions of *predictors*. . . . Regression can be used to predict an outcome given a lienar function of these predictors, and regression coefficients can be thought of as comparisons across predicted values or as comparisons among averages in the data.

(Gelman and Hill, 2007, p.31)

What is linear regression?

- simple statistical model
- model of mean and variance of normally (Gaussian) distributed variance
- mean as *additive* combination of *weighted* variables
- constant variance

Linear models

- models of normally distributed data
  - t-test, single regression, multiple regression, ANOVA, ANCOVA, MANOVA, MANCOVA
  - all just special cases of linear regression
- focus on the strategy, not the magic words

One predictor

y is a continuous variable defined for $-\infty$ to ∞ (e.g. test score)

x is a vector of 0s and 1s (e.g. completed high school)

$$y = \alpha + \beta x + \epsilon$$
$$\epsilon \sim \text{Normal}(0, \sigma)$$

Parts of this model

$$y = \alpha + \beta x + \epsilon$$
$$\epsilon \sim \text{Normal}(0, \sigma)$$

- α
    - "the intercept"
    - expected value of y when all x are 0
- β
    - regression coefficient
    - expected change in y per unit change in x
- ϵ
    - "error" term
    - describes the residuals – the dispersion around the linear element
- σ
    - standard deviation of dispersion around linear element

Another way to write out linear model

$$y = \alpha + \beta x + \epsilon$$
$$\epsilon \sim \text{Normal}(0, \sigma)$$

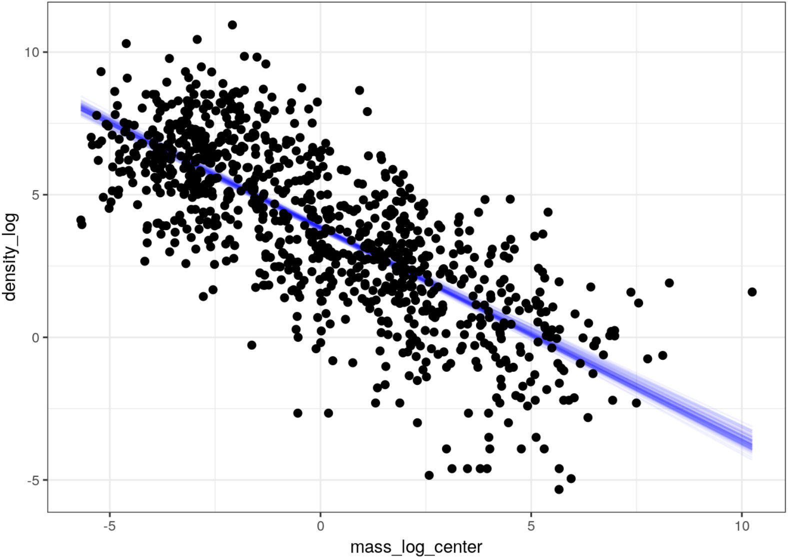
is the same as

$$y \sim \text{Normal}(\alpha + \beta x, \sigma)$$

Regression coefficients

The expected difference in y between two units that differ by 1 in a single predictor.

One predictor

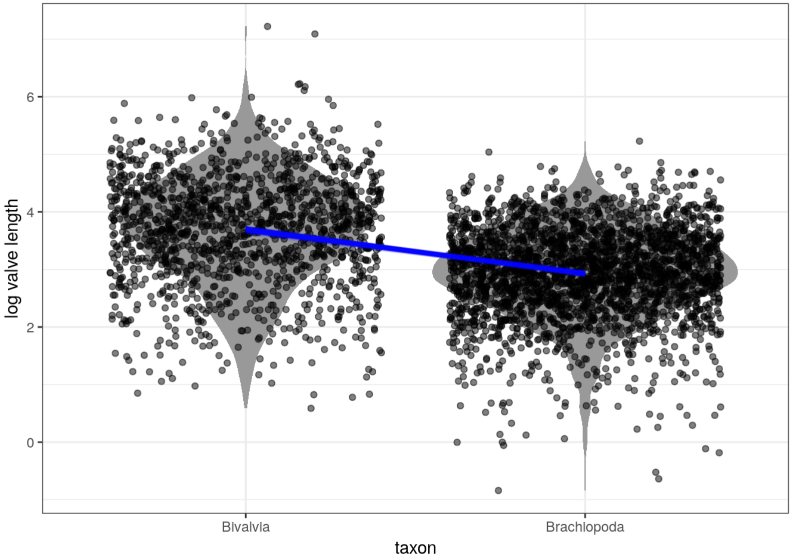


One *binary* predictor

For a binary predictor, the regression coefficient is the difference between the averages of the two groups.

(Gelman and Hill, 2007, p.31)

One *binary* predictor



Multiple predictors

Typical advice is to interpret each coefficient "with all the other predictors held constant."

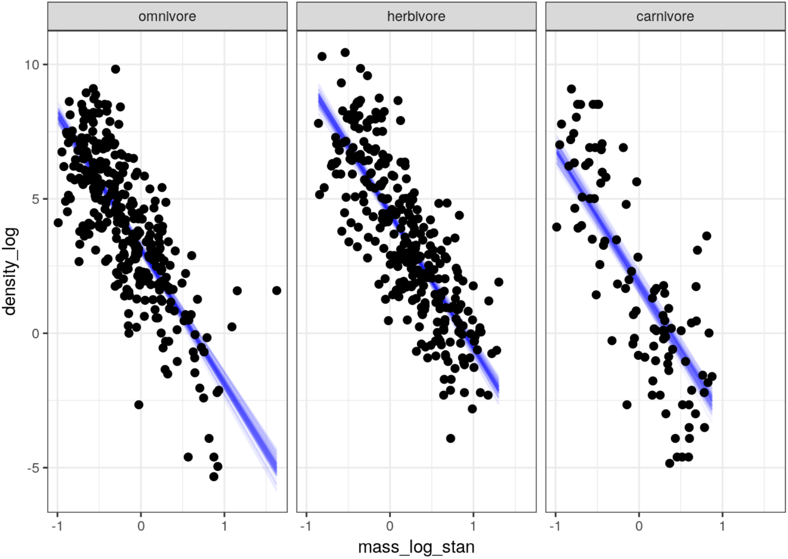
(Gelman and Hill, 2007, p.31)

Multiple predictors. . . what?

$$y \sim \text{Normal}(\alpha + \beta_1 x_1 + \beta_2 x_2, \sigma)$$

β_1 only describes change in y based on x_1. changing x_2 is assumed indendent.

Multiple predictor



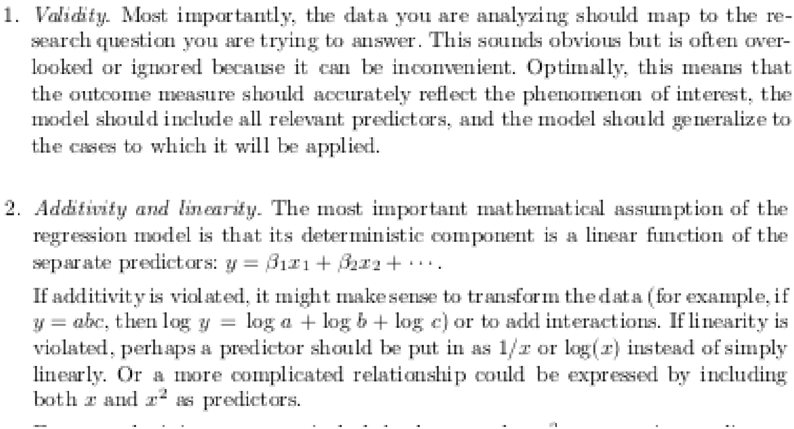
Vector-matrix notation

$$\alpha + \beta_1 x_1 + \beta_2 x_2 = \alpha + X\beta$$

or, if first column of X is all 1s

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = X\beta$$

Assumptions

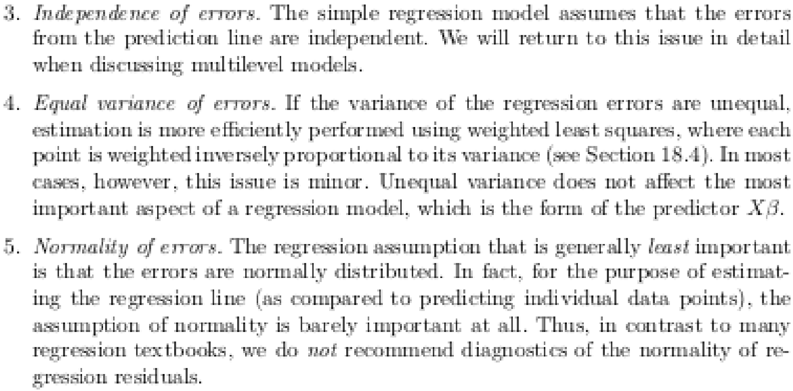


1. *Validity.* Most importantly, the data you are analyzing should map to the research question you are trying to answer. This sounds obvious but is often overlooked or ignored because it can be inconvenient. Optimally, this means that the outcome measure should accurately reflect the phenomenon of interest, the model should include all relevant predictors, and the model should generalize to the cases to which it will be applied.

2. *Additivity and linearity.* The most important mathematical assumption of the regression model is that its deterministic component is a linear function of the separate predictors: $y = \beta_1 x_1 + \beta_2 x_2 + \cdots$.

   If additivity is violated, it might make sense to transform the data (for example, if $y = abc$, then $\log y = \log a + \log b + \log c$) or to add interactions. If linearity is violated, perhaps a predictor should be put in as $1/x$ or $\log(x)$ instead of simply linearly. Or a more complicated relationship could be expressed by including both x and x^2 as predictors.

Assumptions

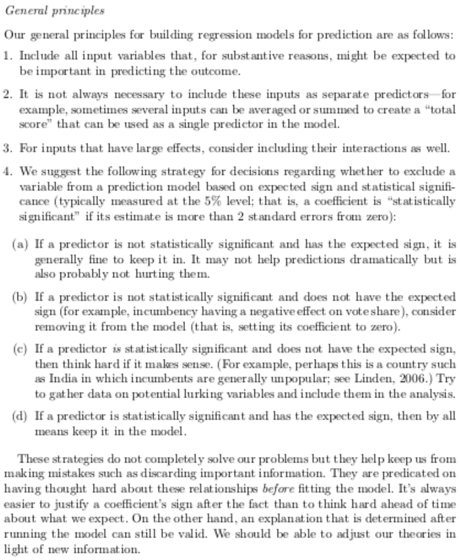


3. *Independence of errors.* The simple regression model assumes that the errors from the prediction line are independent. We will return to this issue in detail when discussing multilevel models.

4. *Equal variance of errors.* If the variance of the regression errors are unequal, estimation is more efficiently performed using weighted least squares, where each point is weighted inversely proportional to its variance (see Section 18.4). In most cases, however, this issue is minor. Unequal variance does not affect the most important aspect of a regression model, which is the form of the predictor $X\beta$.

5. *Normality of errors.* The regression assumption that is generally *least* important is that the errors are normally distributed. In fact, for the purpose of estimating the regression line (as compared to predicting individual data points), the assumption of normality is barely important at all. Thus, in contrast to many regression textbooks, we do *not* recommend diagnostics of the normality of regression residuals.

Advice



General principles

Our general principles for building regression models for prediction are as follows:

1. Include all input variables that, for substantive reasons, might be expected to be important in predicting the outcome.
2. It is not always necessary to include these inputs as separate predictors—for example, sometimes several inputs can be averaged or summed to create a "total score" that can be used as a single predictor in the model.
3. For inputs that have large effects, consider including their interactions as well.
4. We suggest the following strategy for decisions regarding whether to exclude a variable from a prediction model based on expected sign and statistical significance (typically measured at the 5% level; that is, a coefficient is "statistically significant" if its estimate is more than 2 standard errors from zero):

   (a) If a predictor is not statistically significant and has the expected sign, it is generally fine to keep it in. It may not help predictions dramatically but is also probably not hurting them.

   (b) If a predictor is not statistically significant and does not have the expected sign (for example, incumbency having a negative effect on vote share), consider removing it from the model (that is, setting its coefficient to zero).

   (c) If a predictor *is* statistically significant and does not have the expected sign, then think hard if it makes sense. (For example, perhaps this is a country such as India in which incumbents are generally unpopular; see Linden, 2006.) Try to gather data on potential lurking variables and include them in the analysis.

   (d) If a predictor is statistically significant and has the expected sign, then by all means keep it in the model.

These strategies do not completely solve our problems but they help keep us from making mistakes such as discarding important information. They are predicated on having thought hard about these relationships *before* fitting the model. It's always easier to justify a coefficient's sign after the fact than to think hard ahead of time about what we expect. On the other hand, an explanation that is determined after running the model can still be valid. We should be able to adjust our theories in light of new information.

Logistic regression

The standard way to model binary outomes.

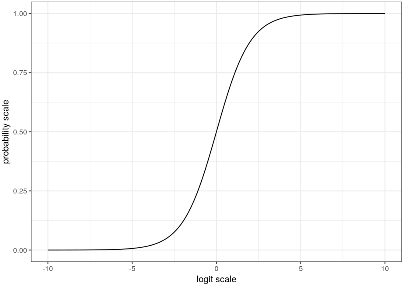
$$\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta)$$

Logit, or log-odds, function

$$\text{logit}(p) = \log \frac{p}{1-p}$$

maps continuous space from $-\infty$ to ∞ to (0, 1).

Logit, or log-odds, function



When x has a magnitude of 3 , y is either close to its maxima or minima. The approximate slope of the logistic function for x between -3 and 3 is much larger than the slope of the logistic function for values with a magnitude of 3+. Any change on the logit scale is compressed at the ends of the probability scale.

Alternative notation

$$\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta)$$

is the same as

$$y_i \sim \text{Bernoulli}(\theta_i)$$
$$\theta_i = \text{logit}^{-1}(X_i\beta)$$

Interpreting logistic regression coefs

The definition of a regression coefficient is that it describes the expected change in the response per unit change in its predictor. However, the logit (or inverse logit) function introduced into our model creates a nonlinearity which complicates the simplicity of this interpretation.

$$
\begin{aligned}
y &\sim \text{Bernoulli}(\theta) \\
\theta &= \text{logit}^{-1}(-1 + -0.5x)
\end{aligned}
$$

Interpreting logistic regression coefs: intercept

The intercept of a logistic can only be interpreted assuming zero values for all the other predictors. This point is where on the probability axis the logistic function crosses $x = 0$.

With mean-centered data, the central point describes the baseline probability of the result being a 1.

Interpreting logistic regression coefs: probability

For the coefficient -0.5, we can calculate this difference in probability when $\mathrm{x} = 0$ and when $\mathrm{x} = 1$.

An increase of x by 1 would decrease the probability of $y = 1$ by $\mathrm{logit}^{-1}(-1 + -0.5 \cdot 1) - \mathrm{logit}^{-1} -1 + -0.5 \cdot 0 \approx -0.09$.

This is the absolute effect of the predictor on y.

Interpreting logistic regression coefs: odds-ratios

Two outcomes have probabilities $(p, 1 - p)$, then $p/(1 - p)$ is the odds. Odds of 1 is equivalent to a probability of 0.5 $(1 = 0.5/(1 - 0.5))$. Odds of 0.5 or 2 represent probabilities of $1/3$ and $2/3$, respectively.

The ratio of two odds (e.g. $(p_1/(1 - p_1))/(p_2/(1 - p_2))$) is called an odds ratio. An odds ratio of 2 corresponds to a change from $p = 0.33$ to $p = 0.5$, or a change from $p = 0.5$ to $p = 0.67$.

Interpreting logistic regression coefs: odds-ratios

Exponentiated logistic regression coefficients can be interpreted as odds ratios.

$$\log \left(\frac{\Pr(y = 1|x)}{\Pr(y = 0|x)} \right) = \alpha + \beta x$$

Adding 1 to x in this equation as the effect of adding β to both sides of the equation.

Exponentiating both sides means that the odds are then multiplied by e^{β}.

So if $\beta = 0.3$, then a unit change in x corresponds to multiplicative change of $e^{0.3} \approx 1.35$ in the odds.

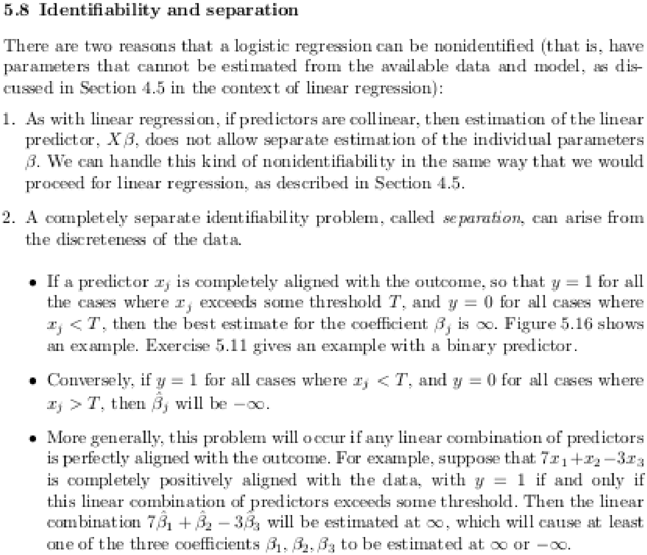
Interpreting logistic regression coefs: aside

Most researchers don't understand odds, so using odds ratios only increases confusion.

By working with probabilities your interpretations are on the same scale as the data, not a transform of the data.

However, most people are taught the odds-ratio interpretation, so it is worth getting used to even if you don't use it.

Identifiability and separation



5.8 Identifiability and separation

There are two reasons that a logistic regression can be nonidentified (that is, have parameters that cannot be estimated from the available data and model, as discussed in Section 4.5 in the context of linear regression):

1. As with linear regression, if predictors are collinear, then estimation of the linear predictor, $X\beta$, does not allow separate estimation of the individual parameters β. We can handle this kind of nonidentifiability in the same way that we would proceed for linear regression, as described in Section 4.5.

2. A completely separate identifiability problem, called *separation*, can arise from the discreteness of the data.

   - If a predictor x_j is completely aligned with the outcome, so that $y = 1$ for all the cases where x_j exceeds some threshold T, and $y = 0$ for all cases where $x_j < T$, then the best estimate for the coefficient β_j is ∞. Figure 5.16 shows an example. Exercise 5.11 gives an example with a binary predictor.

   - Conversely, if $y = 1$ for all cases where $x_j < T$, and $y = 0$ for all cases where $x_j > T$, then $\hat{\beta}_j$ will be $-\infty$.

   - More generally, this problem will occur if any linear combination of predictors is perfectly aligned with the outcome. For example, suppose that $7x_1 + x_2 - 3x_3$ is completely positively aligned with the data, with $y = 1$ if and only if this linear combination of predictors exceeds some threshold. Then the linear combination $7\hat{\beta}_1 + \hat{\beta}_2 - 3\hat{\beta}_3$ will be estimated at ∞, which will cause at least one of the three coefficients $\beta_1, \beta_2, \beta_3$ to be estimated at ∞ or $-\infty$.