Applied Statistics With R

Review of Linear Models

John Fox

WU Wien May/June 2006

Outline

General Form of the Linear Model

- The general linear model is given by the equation

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

- Collecting the regressors into a row vector, appending a 1 for the constant, and placing the corresponding parameters in a column vector,

$$\begin{aligned} Y_i &= [1, x_{i1}, x_{i2}, \ldots, x_{ik}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \epsilon_i \\ &= \underset{(1\times k+1)}{\mathbf{x}_i'} \underset{(k+1\times 1)}{\boldsymbol{\beta}} + \epsilon_i \end{aligned}$$

General Form of the Linear Model

Model in Matrix Form

- For a sample of n observations,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times k+1)}{\mathbf{X}} \underset{(k+1\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\epsilon}}$$

- With suitable specification of the contents of $\mathbf{X}$, called the *model matrix*, this equation serves not only for multiple regression, but for linear models generally.

General Form of the Linear Model

Assumptions of the Linear Model

- The errors are assumed to be independent and normally distributed with zero expectation and common variance: $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$.
- The distribution of $\mathbf{y}$ follows:

$$\begin{aligned} \text{Expectation: } \boldsymbol{\mu} &= E(\mathbf{y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\epsilon}) \\ &= \mathbf{X}\boldsymbol{\beta} \\ \text{Covariance Matrix: } V(\mathbf{y}) &= E[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'] \\ &= E[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'] = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') \\ &= \sigma_\epsilon^2 \mathbf{I}_n \end{aligned}$$

  - Because it is simply a translation of $\boldsymbol{\epsilon}$ to a different expectation, $\mathbf{y}$ is also normally distributed: $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma_\epsilon^2 \mathbf{I}_n)$.

Estimating the Linear Model

Least-Squares Fit

- The fitted linear model is

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

- The least-squares estimates of the coefficients are the solution to the normal equations

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$$

- If the model matrix $\mathbf{X}$ is of full rank, then

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Estimating the Linear Model

Distribution of the Least-Squares Coefficients

- Under the assumptions of the linear model,

$$\mathbf{b} \sim N_{k+1}[\boldsymbol{\beta},\ \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}]$$

- Estimating the error variance as $S_E^2 = \mathbf{e}'\mathbf{e}/(n-k-1)$, the estimated covariance matrix of the coefficients is

$$\widehat{V}(\mathbf{b}) = S_E^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{\mathbf{e}'\mathbf{e}}{n-k-1}(\mathbf{X}'\mathbf{X})^{-1}$$

- The standard error $\mathrm{SE}(B_j)$ of the coefficient B_j is the square root of the jth diagonal entry of $\widehat{V}(\mathbf{b})$.

Inference for Regression Coefficients

Individual Coefficients

- To test the hypothesis H_0: $\beta_j = \beta_j^{(0)}$, calculate the test statistic

$$t_0 = \frac{B_j - \beta_j^{(0)}}{\mathrm{SE}(B_j)}$$

  comparing the obtained value of t_0 with the quantiles of t_{n-k-1}.
- Likewise, a $100(1-a)\%$ confidence interval for β_j is given by

$$\beta_j = B_j \pm t_{a/2,\,n-k-1}\mathrm{SE}(B_j)$$

  where $t_{a/2,\,n-k-1}$ is the critical value of t_{n-k-1} with a probability of $a/2$ to the right.

Inference for Regression Coefficients

Several Coefficients

- To test H_0: $\beta_1 = \cdots = \beta_q = 0$ for $q \leq k$, fit
  1. the *full model*
  $$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$
  with residual sum of squares RSS and
  2. the *null model*
  $$\begin{aligned} Y &= \beta_0 + 0x_1 + \cdots + 0x_q + \beta_{q+1} x_{q+1} + \cdots + \beta_k x_k + \epsilon \\ &= \beta_0 + \beta_{q+1} x_{q+1} + \cdots + \beta_k x_k + \epsilon \end{aligned}$$
  with residual sum of squares RSS_0.
- Then, under H_0, the *incremental F-statistic*
$$F_0 = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-k-1)}$$
is distributed as F with q and $n-k-1$ df.

Inference for Regression Coefficients

Several Coefficients

- This test statistic can also be written
$$F_0 = \mathbf{b}_1' \mathbf{V}_{11}^{-1} \mathbf{b}_1 / q S_E^2$$
where $\mathbf{V}_{11}$ represents the square submatrix consisting of the entries in the q rows and q columns of $(\mathbf{X}'\mathbf{X})^{-1}$ that pertain to the coefficients in $\mathbf{b}_1 = [B_1, \ldots, B_q]'$.

Inference for Regression Coefficients

General Linear Hypothesis

- More generally, to test
$$H_0: \underset{(q \times k+1)}{\mathbf{L}} \quad \underset{(k+1 \times 1)}{\boldsymbol{\beta}} \quad = \quad \underset{(q \times 1)}{\mathbf{c}}$$
where $\mathbf{L}$ and $\mathbf{c}$ contain pre-specified constants, and the *hypothesis matrix* $\mathbf{L}$ is of full row rank $q \leq k+1$:
$$F_0 = \frac{(\mathbf{L}\mathbf{b} - \mathbf{c})' \left[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\right]^{-1} (\mathbf{L}\mathbf{b} - \mathbf{c})}{q S_E^2}$$
with q and $n-k-1$ df.
- For example, to test the *omnibus null hypothesis* H_0: $\beta_1 = \beta_2 = 0$ in the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, take
$$\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

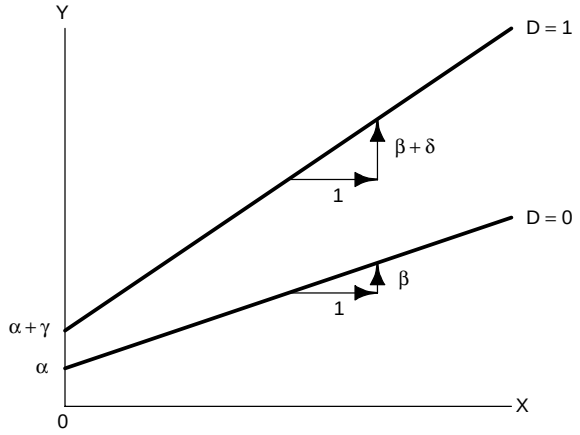
Dummy Regression

Dummy Regression Model

- The model matrices for *dummy-regression* and *analysis-of-variance* models are strongly patterned.
- A dummy-regression model for a *dichotomous* factor:
$$Y_i = \alpha + \beta x_i + \gamma d_i + \delta(x_i d_i) + \epsilon_i$$
where, e.g., Y is income, x is years of education, and the dummy regressor d is coded 1 for men and 0 for women.
- Then:
$$\begin{aligned} \text{Women: } Y_i &= \alpha + \beta X_i + \gamma(0) + \delta(X_i \cdot 0) + \varepsilon_i \\ &= \alpha + \beta X_i + \varepsilon_i \\ \text{Men: } Y_i &= \alpha + \beta X_i + \gamma(1) + \delta(X_i \cdot 1) + \varepsilon_i \\ &= (\alpha + \gamma) + (\beta + \delta) X_i + \varepsilon_i \end{aligned}$$

Dummy Regression

Geometry of the Dummy Regression Model



Dummy Regression

Dummy Regression in Matrix Form

- In matrix form:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{n_1} \\ Y_{n_1+1} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_1} & 0 & 0 \\ 1 & x_{n_1+1} & 1 & x_{n_1+1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{n_1} \\ \epsilon_{n_1+1} \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

  - To emphasize the pattern of the model matrix, the n_1 observations for women (for whom d and hence xd are 0) precede the $n - n_1$ observations for men.

Dummy Regression

Principle of Marginality

- Following Nelder (1977), we say that the separate partial effects, or *main effects*, of education and gender are *marginal* to the education-by-gender interaction.
- We neither test nor interpret the main effects of explanatory variables that interact.
- If we can rule out interaction either on theoretical or on empirical grounds, then we can proceed to test, estimate, and interpret the main effects.
- It is generally not sensible to fit a model that includes an interaction but omits a main effect marginal to the interaction.

Dummy Regression

Hypothesis Tests

- The *principle of marginality* leads to "Type-II" F-tests:
  - To test the null hypothesis H_0: No gender main effect, compare the model that includes both the gender and education main effects to the model that includes only the education main effect. The interaction is absent from both models.
  - To test the null hypothesis H_0: No education main effect, compare the model that includes both the gender and education main effects to the model that includes only the gender main effect.
  - To test the null hypothesis H_0: No gender $\times$ education interaction, compare the full model with the model that includes only the gender and education main effects (but not the interaction).
  - The estimated error variance in the denominator of all F-statistics is based on the largest model fit to the data (here, the model with both main effects and interactions).

Dummy Regression

Polytomous Factors

- A *polytomous* factor is represented by a *set* of dummy regressors, one fewer than the number of *levels* (categories) of the factor.
  - One level, say the first level of an m-level factor, is arbitrarily selected as the "baseline" level, to which others are compared:

Category	D_1	D_2	$\cdots$	D_{m-1}
1	0	0	$\cdots$	0
2	1	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$
m	0	0	$\cdots$	1

  - For example, for a three-level factor:

Category	D_1	D_2
Blue-collar	0	0
White-collar	1	0
Professional and Managerial	0	1

One-Way Analysis of Variance

- The *over-parametrized one-way analysis-of-variance model*:

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij} \quad \text{for groups } j = 1, \ldots, m$$

  - Thus the population mean for group j is

$$\mu_j = E(Y_{ij}) = \mu + \alpha_j$$

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{n_1,1} \\ \hline Y_{12} \\ \vdots \\ Y_{n_2,2} \\ \hline \vdots \\ \hline Y_{1,m-1} \\ \vdots \\ Y_{n_{m-1},m-1} \\ \hline Y_{1m} \\ \vdots \\ Y_{n_m,m} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ \hline 1 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 1 & \cdots & 0 & 0 \\ \hline \vdots & \vdots & \vdots & & \vdots & \vdots \\ \hline 1 & 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 & 0 \\ \hline 1 & 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \\ \alpha_m \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{n_1,1} \\ \hline \epsilon_{12} \\ \vdots \\ \epsilon_{n_2,2} \\ \hline \vdots \\ \hline \epsilon_{1,m-1} \\ \vdots \\ \epsilon_{n_{m-1},m-1} \\ \hline \epsilon_{1m} \\ \vdots \\ \epsilon_{n_m,m} \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

One-Way Analysis of Variance

Producing a Full-Rank Model Matrix by Dummy Coding

- The model matrix is of rank m, one less than the number of columns, because the first column of $\mathbf{X}$ is the sum of the others.
- One solution is to delete a column, implicitly setting the corresponding parameter to 0.
  - Deleting the last column of the model matrix, for example, sets $\alpha_m = 0$, establishing the last category as the baseline for a dummy-coding scheme.

One-Way Analysis of Variance

Producing a Full-Rank Model Matrix by Deviation Coding

- Alternatively, imposing the *sigma constraint* $\sum_{j=1}^{m} \alpha_j = 0$ on the parameters leads to the following *full-rank* model matrix $\mathbf{X}_F$, composed of *deviation-coded regressors*:

$$\underset{(n\times m)}{\mathbf{X}_F} = \left[\begin{array}{ccccc} (\mu) & (\alpha_1) & (\alpha_2) & \cdots & (\alpha_{m-1}) \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ \hline 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \hline \vdots & \vdots & \vdots & & \vdots \\ \hline 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ \hline 1 & -1 & -1 & \cdots & -1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & -1 & -1 & \cdots & -1 \end{array}\right]$$

One-Way Analysis of Variance

Parametric Equation

- The relationship between the group means $\boldsymbol{\mu} = \{\mu_j\}$ and the parameters of the constrained model:

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_{m-1} \\ \mu_m \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & -1 & -1 & \cdots & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \end{bmatrix}$$

$$\underset{(m\times 1)}{\boldsymbol{\mu}} = \underset{(m\times m)}{\mathbf{X}_B}\,\underset{(m\times 1)}{\boldsymbol{\beta}_F}$$

  - $\mathbf{X}_B$ is the *row basis* of the full-rank model matrix, consisting of the m unique rows of $\mathbf{X}_F$, one for each group
  - $\boldsymbol{\beta}_F$ is the parameter vector associated with the full-rank model matrix.

One-Way Analysis of Variance

Parameters as a Function of Group Means

- By construction, the $m \times m$ matrix $\mathbf{X}_B$ is of full column rank and hence non-singular, allowing us to invert $\mathbf{X}_B$ and solve uniquely for the constrained parameters in terms of the group means: $\boldsymbol{\beta}_F = \mathbf{X}_B^{-1}\boldsymbol{\mu}$:

$$\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \end{bmatrix} = \begin{bmatrix} \mu_{\cdot} \\ \mu_1 - \mu_{\cdot} \\ \mu_2 - \mu_{\cdot} \\ \vdots \\ \mu_{m-1} - \mu_{\cdot} \end{bmatrix}$$

where

$$\mu_{\cdot} = \frac{\sum_{j=1}^{m} \mu_j}{m}$$

Two-Way Analysis of Variance

- The *two-way Anova model*:

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk} \\ \mu_{jk} &= E(Y_{ijk}) = \mu + \alpha_j + \beta_k + \gamma_{jk} \end{aligned}$$

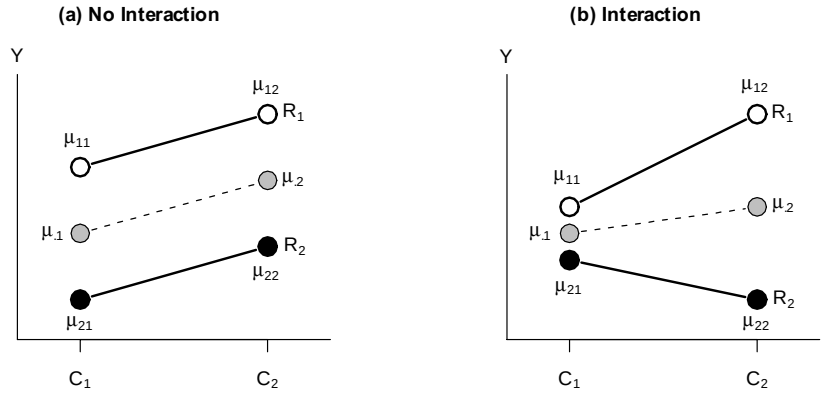
where

  - α_j for $j = 1, \ldots, r$ are the main effects of the *row factor*
  - β_k for $k = 1, \ldots, c$ are the main effects of the *column factor*
  - γ_{jk} are interaction effects
  - μ_{jk} are population cell means

Two-Way Analysis of Variance

Absence and Presence of Interaction

- For the simple case of $r = c = 2$:



Here, e.g., $\mu_{.1} = (\mu_{11} + \mu_{21})/2$ is the *marginal mean* in column 1.

Two-Way Analysis of Variance

Sigma Constraints

- Each set of parameters is constrained to sum to zero over (each of) its coordinates:

$$\begin{aligned} \sum_{j=1}^{r} \alpha_j &= 0 \\ \sum_{k=1}^{c} \beta_k &= 0 \\ \sum_{j=1}^{r} \gamma_{jk} &= 0 \text{ for all } k = 1, \ldots, c \\ \sum_{k=1}^{c} \gamma_{jk} &= 0 \text{ for all } j = 1, \ldots, r \end{aligned}$$

Two-Way Analysis of Variance

Parametric Equation with Sigma Constraints

- The two-way ANOVA model with interactions, for $r = 2$ categories of the row factor and $c = 3$ categories of the column factor:

$$\begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \end{bmatrix}$$

$$\underset{(6\times 1)}{\boldsymbol{\mu}} = \underset{(6\times 6)}{\mathbf{X}_B} \underset{(6\times 1)}{\boldsymbol{\beta}_F}$$

Two-Way Analysis of Variance

Parameters as a Function of Cell Means

- The row basis of the full-rank model matrix is non-singular by construction, yielding the following solution for the parameters in terms of the cell means:

$$\begin{bmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \end{bmatrix} = \begin{bmatrix} \mu_{\cdot\cdot} \\ \mu_{1\cdot} - \mu_{\cdot\cdot} \\ \mu_{\cdot 1} - \mu_{\cdot\cdot} \\ \mu_{\cdot 2} - \mu_{\cdot\cdot} \\ \mu_{11} - \mu_{1\cdot} - \mu_{\cdot 1} + \mu_{\cdot\cdot} \\ \mu_{12} - \mu_{1\cdot} - \mu_{\cdot 2} + \mu_{\cdot\cdot} \end{bmatrix}$$

where, e.g.,

$$\mu_{\cdot\cdot} = \sum_{j=1}^{r} \sum_{k=1}^{c} \mu_{jk}$$

Writing Model Formulas in R

- R implements a version of the Rogers and Wilkinson notation for linear-models formulas.
  - Model formulas, specified as the first argument to `lm()`, are of the form

    left-hand-side $\sim$ right-hand-side

  - On the right-hand side of a model formula the arithmetic operators have special meaning:

Expression	*Interpretation*	*Example*
`A + B`	include both `A` and `B`	`income + education`
`A - B`	exclude `B` from `A`	`a*b*d - a:b:d`
`A:B`	all interactions of `A` and `B`	`type:education`
`A*B`	`A + B + A:B`	`type*education`
`B %in% A`	`B` nested within `A`	`education %in% type`
`A/B`	`A + B %in% A`	`type/education`
`A^k`	all effects crossed up to order `k`	`(a + b + d)^2`

Writing Model Formulas in R

- To perform arithmetic on the right-hand side of a model formula, "protect" the expression via a call to the identity function, `I()`: e.g., `I(a + b)`.
  - Expression involving a function call, such as `log(a + b)` do not require protection.
  - On the left-hand side of a formula, the arithmetic operators have their conventional meaning: e.g., `a + b` $\sim$ `c` adds `a` and `b` and regresses their sum on `c`.
- Model formulas are used for many other kinds of statistical models in R, such as generalized linear models fit by `glm()`.