**KEY TERMS FOR SIMPLE LINEAR REGRESSION**

***Response***

The variable we are trying to predict.

Synonyms

dependent variable, Y-variable, target, outcome

***Independent variable***

The variable used to predict the response.

Synonyms

X-variable, feature, attribute

***Record***

The vector of predictor and outcome values for a specific individual or case.

Synonyms

row, case, instance, example

***Intercept***

The intercept of the regression line—that is, the predicted value when X = 0 .

Synonyms

b 0 , β 0

***Regression coefficient***

The slope of the regression line.

Synonyms

slope, b 1 , β 1 , parameter estimates, weights

***Fitted values***

The estimates Y ^ i obtained from the regression line.

Synonyms

predicted values

***Residuals***

The difference between the observed values and the fitted values.

Synonyms

errors

***Least squares***

The method of fitting a regression by minimizing the sum of squared residuals.

Synonyms

ordinary least squares

**KEY TERMS FOR MULTIPLE LINEAR REGRESSION**

***Root mean squared error***

The square root of the average squared error of the regression (this is the most widely used metric to compare regression models).

Synonyms

RMSE

***Residual standard error***

The same as the root mean squared error, but adjusted for degrees of freedom.

Synonyms

RSE

***R-squared***

The proportion of variance explained by the model, from 0 to 1.

Synonyms

coefficient of determination, R 2

***t-statistic***

The coefficient for a predictor, divided by the standard error of the coefficient, giving a metric to compare the importance of variables in the model. See [“t-Tests”](https://www.oreilly.com/library/view/practical-statistics-for/9781491952955/ch03.html#tTest).

***Weighted regression***

Regression with the records having different weights.

##### KEY IDEAS

* Multiple linear regression models the relationship between a response variable Y and multiple predictor variables X 1 , ... , X p .
* The most important metrics to evaluate a model are root mean squared error (RMSE) and R-squared (R2).
* The standard error of the coefficients can be used to measure the reliability of a variable’s contribution to a model.
* Stepwise regression is a way to automatically determine which variables should be included in the model.
* Weighted regression is used to give certain records more or less weight in fitting the equation.

##### KEY TERMS FOR FACTOR VARIABLES

**Dummy variables**

Binary 0–1 variables derived by recoding factor data for use in regression and other models.

**Reference coding**

The most common type of coding used by statisticians, in which one level of a factor is used as a reference and other factors are compared to that level.

Synonyms

treatment coding

**One hot encoder**

A common type of coding used in the machine learning community in which all factors levels are retained. While useful for certain machine learning algorithms, this approach is not appropriate for multiple linear regression.

**Deviation coding**

A type of coding that compares each level against the overall mean as opposed to the reference level.

Synonyms

sum contrasts

##### KEY TERMS FOR INTERPRETING THE REGRESSION EQUATION

**Correlated variables**

When the predictor variables are highly correlated, it is difficult to interpret the individual coefficients.

**Multicollinearity**

When the predictor variables have perfect, or near-perfect, correlation, the regression can be unstable or impossible to compute.

Synonyms

collinearity

**Confounding variables**

An important predictor that, when omitted, leads to spurious relationships in a regression equation.

**Main effects**

The relationship between a predictor and the outcome variable, independent from other variables.

**Interactions**

An interdependent relationship between two or more predictors and the response.

##### KEY TERMS FOR REGRESSION DIAGNOSTICS

**Standardized residuals**

Residuals divided by the standard error of the residuals.

**Outliers**

Records (or outcome values) that are distant from the rest of the data (or the predicted outcome).

**Influential value**

A value or record whose presence or absence makes a big difference in the regression equation.

**Leverage**

The degree of influence that a single record has on a regression equation.

Synonyms

hat-value

**Non-normal residuals**

Non-normally distributed residuals can invalidate some technical requirements of regression, but are usually not a concern in data science.

**Heteroskedasticity**

When some ranges of the outcome experience residuals with higher variance (may indicate a predictor missing from the equation).

**Partial residual plots**

A diagnostic plot to illuminate the relationship between the outcome variable and a single predictor.

Synonyms

added variables plot