Supervised Learning: Linear & Logistic Regression

Lauren Vanderlinden, PhD, MS
T15 Postdoctoral Fellow Computational Biology
Division of Rheumatology & Department of Biomedical Informatics
School of Medicine, University of Colorado Anschutz Medical Campus

CPBS 7602 - December 2, 2024

Outline

- Analyses and Measurements of Error
- Simple Linear Regression
- Multiple Linear Regression
- Interaction Terms
- Logistic Regression

Why Would We Use Analyses?

Simple Statistical Model

All statistical models can be represented by the equation:

$$\text{response}_i = \text{model} + \text{error}_i$$

This means that every response can be predicted from the model we choose to fit to the data plus come error.

Simple Statistical Model (Cont.)

- The model is typically specified in terms of parameters, which are unknown constants that describe a population characteristic
  - Parameters are usually represented by Greek letters such as: μ, σ, and β
- A statistic is an estimate of a parameter calculated from observed data
  - Statistics are typically represented by phonetic letters (or modified Greek letters) such as: $\bar{x}$, s, and $\hat{\beta}$

Example: the mean as a statistical model

Suppose we would like to summarize the number of friends statistics instructors have using the model:

$$friends_i = \mu + error_i$$

where μ is the mean number of friends for all statistics instructors.

Suppose we randomly selected 5 statistics instructors and measure the number of friends that they have. We observe 1, 2, 3, 3, 4. The mean number of friends is $\bar{x} = 2.6$

Use $\bar{x}$ serves as an estimate for μ

Assessing model fit

The fit of a model is often characterized using the error and sum of squares

The **error** for observation i is the difference between the response for observation *i* and the estimated model.

- Using a formula, the deviance is

$$error_i = response_i - model$$

- Negative error indicates that our model overestimated the response
- Positive error indicates that our model underestimated the response
- Note: Deviance is another term often used in model fit language. In this simple statistical model (and linear regression), deviance and error are the same.

Example: error

- For our example, the errors are given by the formula:

$$error_i = x_i - \bar{x}$$

where x_i is the response for observation i.

- Friends/statistics instructors where $\bar{x} = 2.6$

Instructor 1: $error_1 = 1 - 2.6 = -1.6$

Instructor 2: $error_2 = 2 - 2.6 = -0.6$

Instructor 3: $error_3 = 3 - 2.6 = 0.4$

Instructor 4: $error_3 = 3 - 2.6 = 0.4$

Instructor 5: $error_3 = 4 - 2.6 = 1.4$

Sum of squared errors

The sum of squared errors (SS) is the sum of the squared errors and is a way to measure the fit of a model.

$$SS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- The larger the SS is, the more poorly our model fits the data.

- Note: the sample variance is $s^2 = ss/(n-1)$
- Note: the sample standard deviation is $s = \sqrt{s^2}$

Comparison of variability measures

Measure	What it measures	Formula
Variance (σ^2 or s^2)	Measures the spread of data points around the mean (how much data varies).	Population Variance (σ^2): $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$ Sample Variance (s^2): $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standard Deviation (σ or s)	Measures the spread of data points around the mean (how much data varies). It's the square root of the variance.	Population SD (σ): $\sigma = \sqrt{\sigma^2}$ Sample SD (s): $s = \sqrt{s^2}$
Standard Error (SE)	Measures the precision of the sample mean as an estimate of the population mean.	Standard Error of the Mean (SEM): $SE = \frac{s}{\sqrt{n}}$

Regression Analysis

What is regression?

- Regression analysis provides a way of describing the distribution of a response (outcome/dependent) variable as a function of one or more explanatory (predictor/independent) variables.

- The **regression of the response variable on the explanatory variable** described the mathematical relationship between the response variable and the explanatory variable.

Simple Linear Regression

Regression Analyses

Simple Linear Regression Model

The simple linear regression model postulates that the mean of the response variable Y, as a function of a single explanatory variable X denoted by $\mu_{Y|X}$, is given by the linear relationship

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

This is called a **regression line**.

Relationship between Y_i and X_i

The relationship between the response of the *i*th observation (Y_i) and the value of the explanatory variable of the *i*th observation (X_i) is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Notice that

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \mu_{Y|X} + \varepsilon_i$$

$$Y_i = model + \varepsilon_i$$

Thus, the value of the *i*th observation (Y_i) is the mean of Y when $X = X_i$ plus some random error

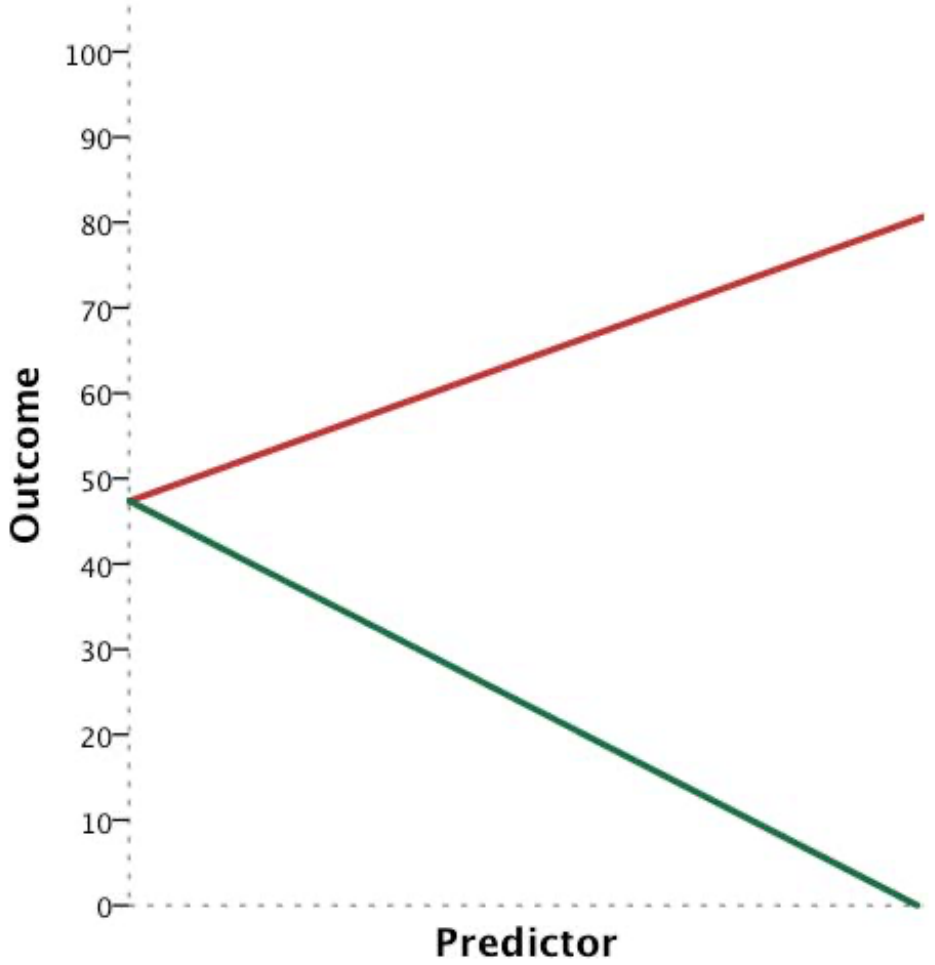
Terms in a simple linear regression model

- β_0 is the *intercept* of the model
  - It is the mean value for Y when $X = 0$.
  - It is where the regression line crosses the y-axis.

- β_1 is the *slope* of the model
  - It is the mean change in Y when the explanatory variable X increases by 1 unit.
  - Describes how the response variable is affected by the explanatory variable.

β_0 and β_1 are called **regression coefficients**.

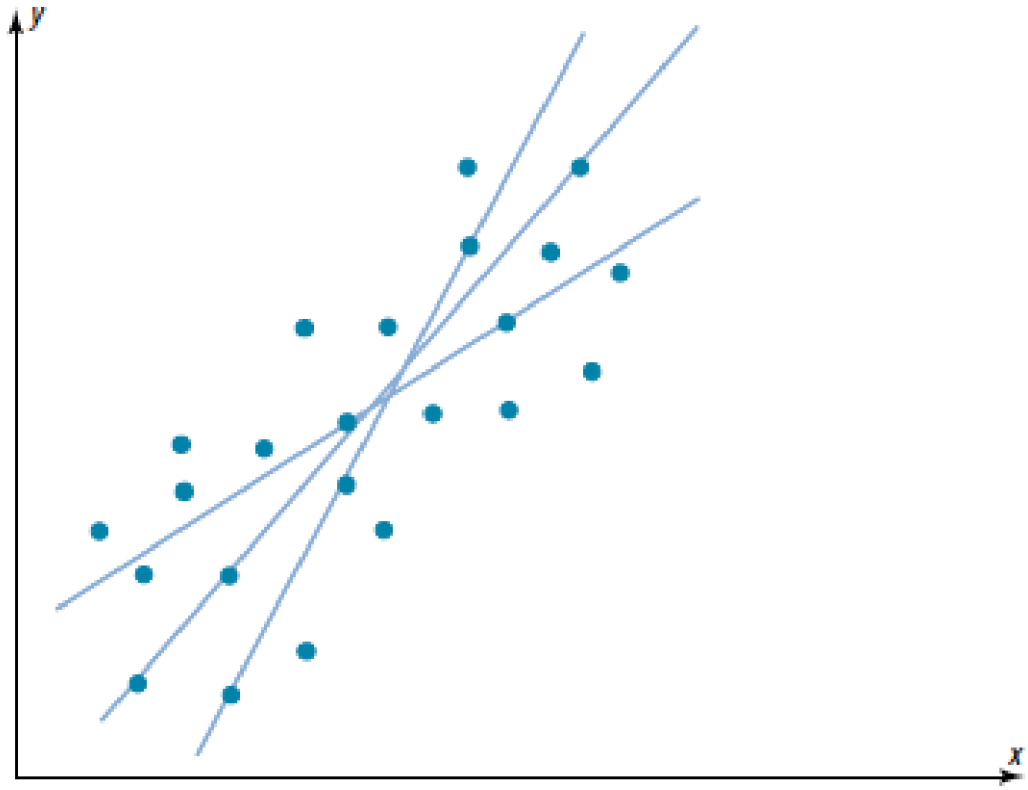
Relationship between intercepts and slopes



Best fit line for the data

- We want to describe the relationship between the response variable and the explanatory variable with a regression line.

- How do we decide which model is the “best” model since there are many possible straight lines?



Method of least squares

Method of least squares – a regression procedure that estimates the regression coefficients with the values that minimize the sum of the squared deviations (residuals).

The estimated mean response for Y_i is $$\hat{Y}_i = \hat{\mu}_{Y|X} = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated values of β_0 and β_1

$\hat{Y}_i$ is known as the *i*th fitted value.

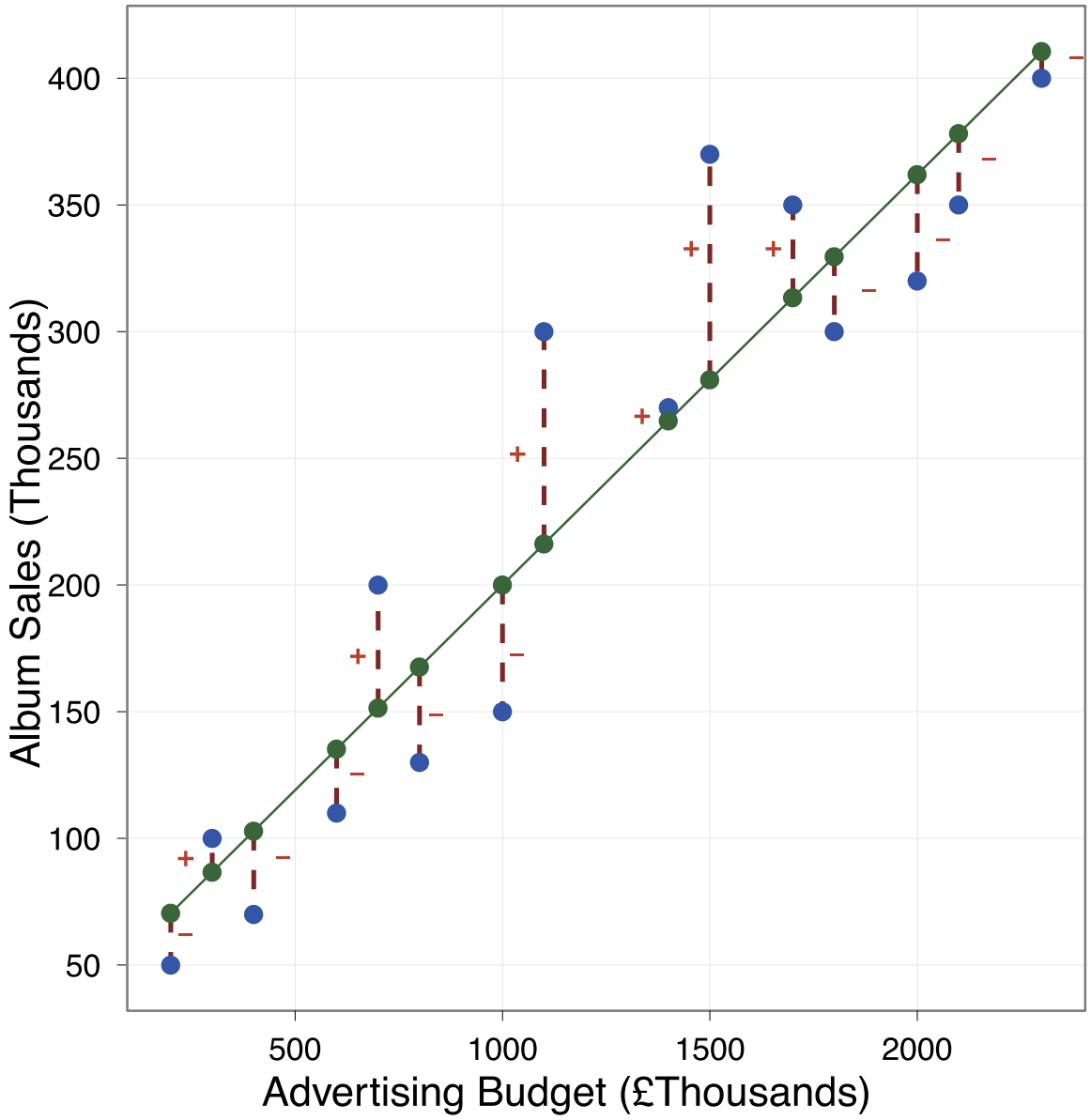
The *i*th residual (deviation) is $e_i = Y_i - \hat{Y}_i$

- This is the difference between the observed response and the estimated response for the *i*th observation

The least squares method finds the $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

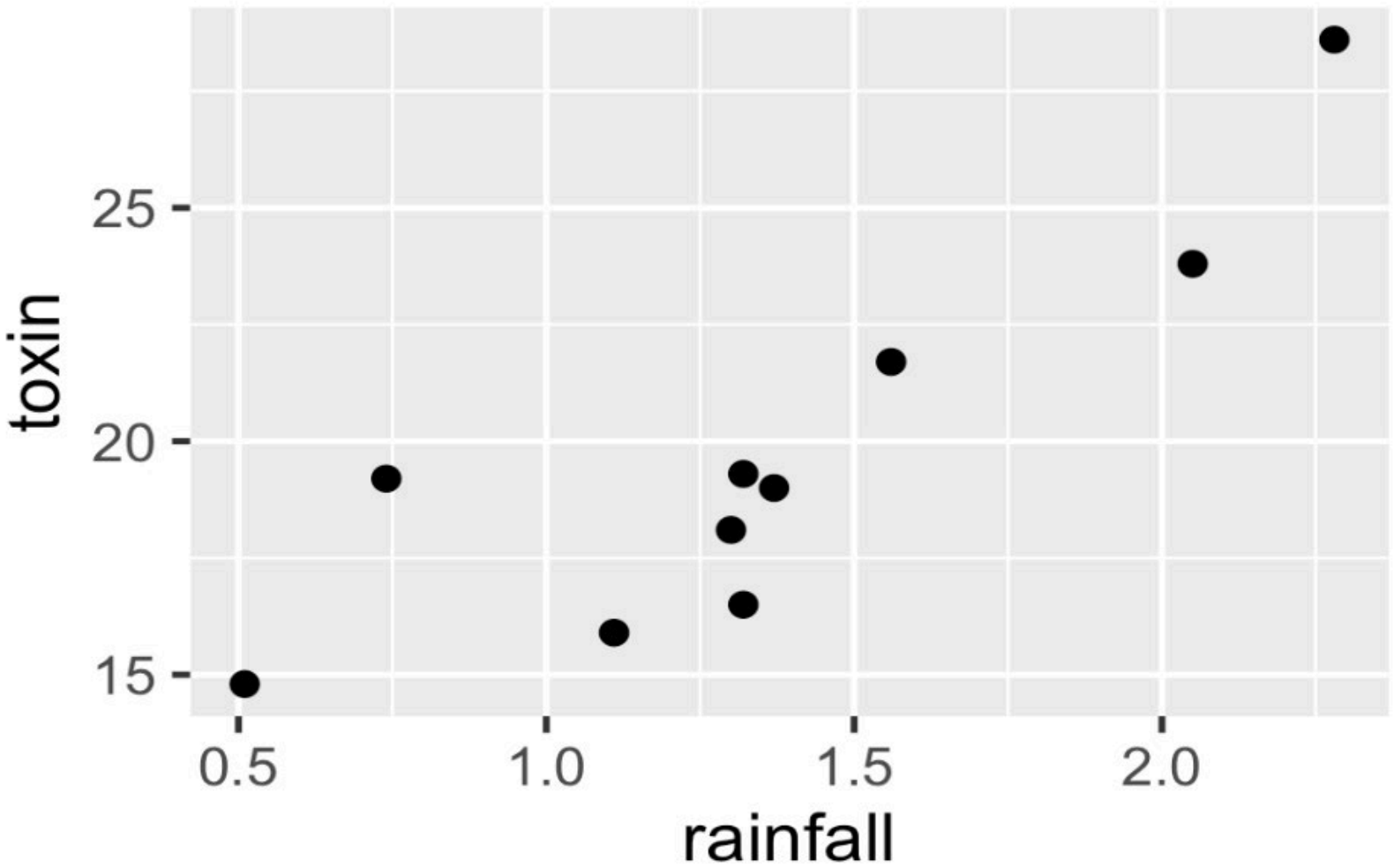
$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} e_i^2$$

Goal: minimize those vertical dashed lines



Simple Linear Regression Example

A drug precursor molecule is extracted from a type of nut. The nuts are commonly contaminated by a fungal toxin that is difficult to remove during the purification process. We suspect that the amount of fungus (and hence toxin) depends on the rainfall at the growing site. We would like to predict toxin concentration from rainfall in order to judge whether it would be worth paying additional rental charges for relatively drier sites. We analyze the toxin content in a series of batches of nuts and we know the rainfall at the growing sites during the four months when the nuts are forming.



Toxin Content (microgram per 100 grams) By Amount of Rainfall (cm per week)

rainfall	toxin
1.30	18.1
2.28	28.6
1.11	15.9
0.74	19.2
1.32	19.3
0.51	14.8
1.56	21.7
1.32	16.5
2.05	23.8
1.37	19.0

Code/Results in R

```r
simple_regression <- lm(toxin ~ rainfall, data=toxin)
summary(simple_regression)
```

```
Call:
lm(formula = toxin ~ rainfall, data = toxin)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9479 -1.1061 -0.3528  0.7596  3.6531

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.570      1.961   5.390 0.000654 ***
rainfall       6.726      1.356   4.961 0.001105 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.16 on 8 degrees of freedom
Multiple R-squared:  0.7547,    Adjusted R-squared:  0.724
F-statistic: 24.61 on 1 and 8 DF,  p-value: 0.001105
```

```
Call:
lm(formula = toxin ~ rainfall, data = toxin)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9479 -1.1061 -0.3528  0.7596  3.6531

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.570      1.961   5.390 0.000654 ***
rainfall       6.726      1.356   4.961 0.001105 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.16 on 8 degrees of freedom
Multiple R-squared:  0.7547,    Adjusted R-squared:  0.724
F-statistic: 24.61 on 1 and 8 DF,  p-value: 0.001105
```

Intercept $(\hat{\beta}_0)$ = 10.57 micrograms / 100 grams

Slope $(\hat{\beta}_1)$ = 6.73 micrograms per 100 grams / cm per week

Regression Model:

$$10.57 + 6.73 * rainfall$$

Results from example

An increase in rainfall is significantly associated with an increase in nut toxin concentration (slope = 6.73, standard error = 1.36, p-value = 0.0011). The full regression model is toxin concentration = 10.57 + 6.73 * rainfall.

Making predictions using the regression equation

Having obtained the regression equation, we might now have the chance to rent two agricultural locations where we could grow a crop of nuts. We show that the weekly rainfall at Sites A and B during the fruiting season are 2.05 and 1.25 cm/week respectively. Therefore, we can predict the level of toxin in nuts grown at these two sites.

Site A:

$$Toxin = 10.6 + 6.73 * Rainfall\left(\frac{cm}{week}\right)$$
$$= 10.6 + 6.73*2.05$$
$$= 10.6 + 13.8$$
$$= 24.4$$

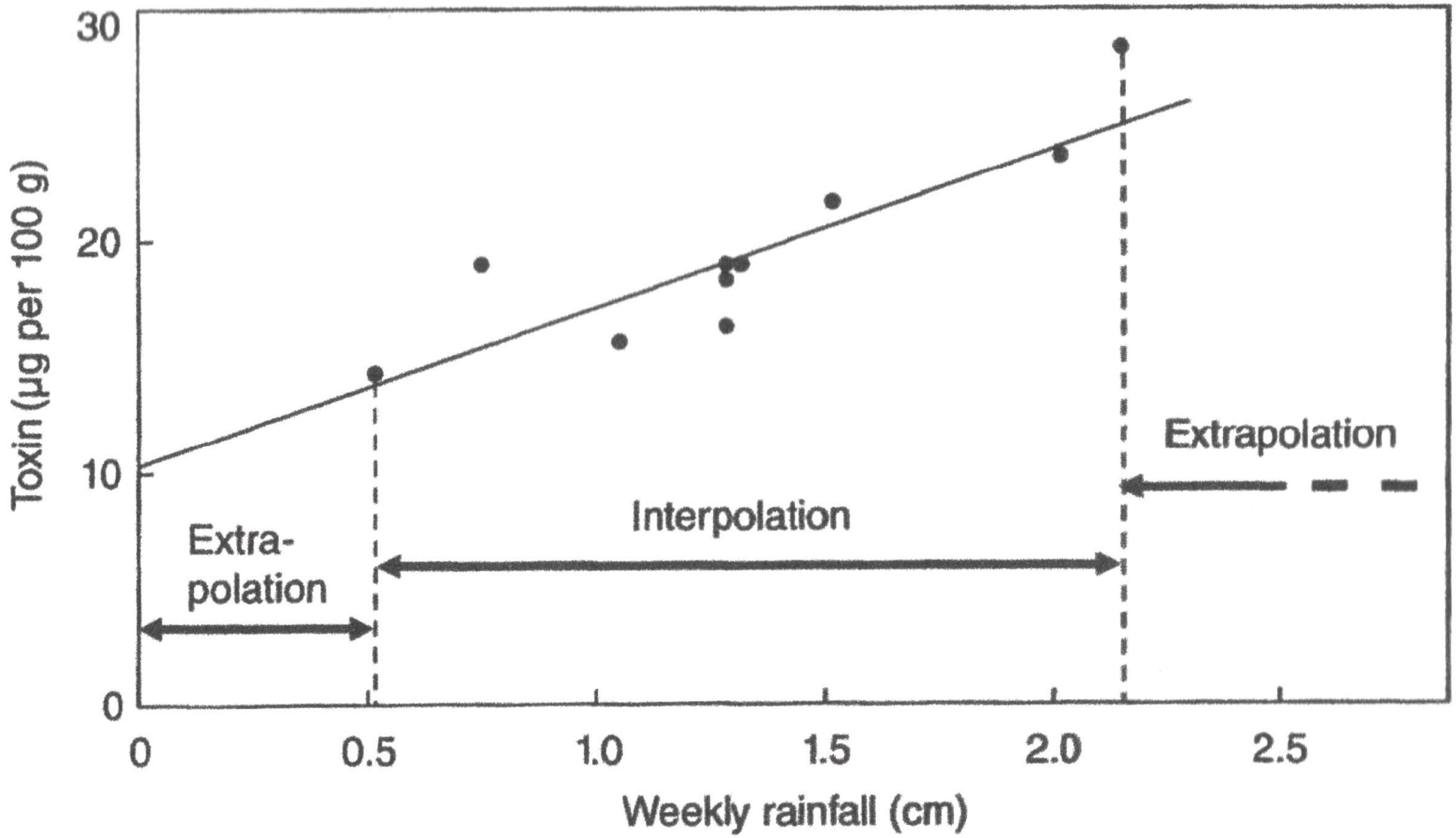
Site B:

$$Toxin = 10.6 + 6.73 * Rainfall\left(\frac{cm}{week}\right)$$
$$= 10.6 + 6.73*1.25$$
$$= 10.6 + 8.4$$
$$= 19.0$$

Interpolation and Extrapolation

- **Interpolation**: A prediction using a value of the independent variable that is within the observed range - uncontroversial.

- **Extrapolation**: A prediction using a value of the independent variable that lies outside the observed range. Extrapolation should be avoided unless there is sound reason to believe that the linear relationship extends beyond the observed range.


30
20
10
0
Toxin (µg per 100 g)
0
0.5
1.0
1.5
2.0
2.5
Weekly rainfall (cm)
Extra-
polation
Interpolation
Extrapolation

Summary of simple linear regression

- A simple statistical model indicates that the response (dependent variable) equals the statistical model plus a sample-specific error.
- Simple linear regression includes only one predictor variable.
- The ‘intercept’ from a linear regression model is the average *y* value when *x* is zero. The ‘slope’ from a linear regression model is the average change in *y* when *x* is increased by one unit.

Multiple Linear Regression

Regression Analyses

Simple vs. Multiple Linear Regression

Simple linear regression - regression assuming that the relationship between a single explanatory variable and the response variable is linear.

Multiple linear regression - regression using multiple explanatory variables that assumes the relationship can be written as a linear equation.

Multiple Linear Regression

- Multiple regression is a linear regression model with one response variable and multiple explanatory variables.
- Our model is similar to before, but we have extra regression coefficients and explanatory variables.
- Multiple regression is often used when we want to ‘control’ for a variable (i.e. confounders, precision covariates…)

Multiple Regression

The relationship between the response for the *i*th observation (Y_i) and the values for the explanatory variables for the *i*th observation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

$$Y_i = \mu_{Y|X_1,\ldots,X_{p-1}} + \epsilon_i$$

where X_{i1} is the value of the first explanatory variable for observation *i*, X_{i2} is the value of the second explanatory variable for the observation *i*, etc., and β_j is the regression coefficient for the *j*th explanatory variable.

Interpretation of regression coefficients

- β_0 - the mean value of Y when ALL the explanatory variables are equal to zero.
- β_j - the mean change in Y when the *j*th explanatory variable increases by 1 unit and the other explanatory variables are held constant.
- $\hat{\beta}_0$ - the estimated mean value of Y when ALL the explanatory variables are equal to zero.
- $\hat{\beta}_j$ - the estimated mean change in Y when the *j*th explanatory variable increases by 1 unit and the other explanatory variables are held constant.

Estimating Regression Coefficients

We estimate the regression coefficients for multiple regression using the method of least squares, i.e., estimating β_j with $\hat{\beta}_j$ that minimize

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 = \sum_{i=1}^{n} e_i^2$$

where

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta_{p-1}} X_{p-1}$$

Multiple Linear Regression Example

We have 4 potential predictors of concentration of fungal toxin in nuts (µg per 100g).

1. **rainfall** - average amount of rainfall in cm per week
2. **noon_temp** - average temperature (degrees Celsius) at noon
3. **sunshine** - average number of hours per day of sunshine
4. **wind_speed** - average wind speed in km per hour

Multiple Predictors of Toxin Content (microgram per 100 grams)

rain	noon_temp	sunshine	wind_speed	toxin
1.30	20.9	6.23	13.3	18.1
2.28	25.4	8.13	10.8	28.6
1.11	28.2	10.21	10.9	15.9
0.74	23.7	6.96	8.2	19.2
1.32	26.5	9.04	9.8	19.3
0.51	23.9	7.84	12.3	14.8
1.56	26.7	6.69	10.0	21.7
1.32	30.0	8.30	12.2	16.5
2.05	24.9	9.22	10.7	23.8
1.37	22.0	8.37	15.0	19.0

Code/Results in R

```r
multiple_regression <- lm(toxin ~ rain + noon_temp + sunshine + wind_speed, data=toxin)
summary(multiple_regression)
```

```
Call:
lm(formula = toxin ~ rain + noon_temp + sunshine + wind_speed,
    data = toxin)

Residuals:
      1       2       3       4       5       6       7       8       9      10
-1.8818  2.0498 -0.6314  0.4787 -0.5805  1.2508 -0.1921 -0.1813 -1.1552  0.8429

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.6084     7.1051   4.449  0.00671 **
rain          7.0676     1.0031   7.046  0.00089 ***
noon_temp    -0.4201     0.2413  -1.741  0.14215
sunshine     -0.2375     0.5086  -0.467  0.66018
wind_speed   -0.7936     0.2977  -2.666  0.04458 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.574 on 5 degrees of freedom
Multiple R-squared:  0.9186,    Adjusted R-squared:  0.8535
F-statistic: 14.11 on 4 and 5 DF,  p-value: 0.006232
```

```
Call:
lm(formula = toxin ~ rain + noon_temp + sunshine + wind_speed,
    data = toxin)

Residuals:
      1       2       3       4       5       6       7       8       9      10
-1.8818  2.0498 -0.6314  0.4787 -0.5805  1.2508 -0.1921 -0.1813 -1.1552  0.8429

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.6084     7.1051   4.449  0.00671 **
rain          7.0676     1.0031   7.046  0.00089 ***
noon_temp    -0.4201     0.2413  -1.741  0.14215
sunshine     -0.2375     0.5086  -0.467  0.66018
wind_speed   -0.7936     0.2977  -2.666  0.04458 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.574 on 5 degrees of freedom
Multiple R-squared:  0.9186,    Adjusted R-squared:  0.8535
F-statistic: 14.11 on 4 and 5 DF,  p-value: 0.006232
```

Regression model:

$$\hat{\beta}_0 + \hat{\beta}_{rain}X_{rain} + \hat{\beta}_{noon\_temp}X_{noon\_temp} + \hat{\beta}_{sunshine}X_{sunshine} + \hat{\beta}_{wind\_speed}X_{wind\_speed}$$

$$31.61 + 7.07X_{rain} - 0.42X_{noon_{temp}} - 0.24X_{sunshine} - 0.79X_{wind\_speed}$$

Interpretation of regression coefficients

- $\hat{\beta}_0$ - the mean fungus content is 31.6 micrograms per 100 grams of nut when the area has:
  - Average rainfall of 0 cm per week
  - Average noon temperature of 0℃
  - Average of 0 hours of sunshine per day
  - Average windspeed of 0 km per hour

- $\hat{\beta}_{rainfall}$ - A 1 cm/week increase in rainfall is associated with an increase of 7.1 micrograms of fungus per 100 grams of nut when temperature, amount of sunshine, and windspeed do not change.

Interaction Effects

- For effect modifiers, you want to model an interaction term
  - Example: variable Z is an effect modifier on the relationship between X -> Y
- **Hierarchy principal**: When a product term is included in a model, you must retain the variables included in the interaction as predictors in the model.

$$Y = \beta_0 + \beta_1 * X + \beta_2 * Z + \beta_3 * X * Z$$

Interaction Example

We are interested in the effect of rainfall (potential predictor) of concentration of fungal toxin in nuts (µg per 100g). However, there is previous studies that have shown this relationship also depends on how hot the temperature is.

1. **rainfall** - average amount of rainfall in cm per week
2. **noon_temp** – average temperature (degrees Celsius) at noon

For ease of interpretation in modeling, we are going to stratify our noon_temp as being low or high (depending on if it's below or above the median temperature value).

Code/Results in R

```r
toxin$temp_binary <- ifelse(toxin$noon_temp>median(toxin$noon_temp), 1, 0)
interaction_regression <- lm(toxin ~ rain*temp_binary, data=toxin)
summary(interaction_regression)
```

```
Call:
lm(formula = toxin ~ rain * temp_binary, data = toxin)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7151 -0.9304 -0.1032  0.8287  2.3493

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       13.380      1.693   7.905 0.000217 ***
rain               4.690      1.292   3.629 0.010978 *
temp_binary       -9.731      3.176  -3.064 0.022123 *
rain:temp_binary   6.345      2.144   2.960 0.025287 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.556 on 6 degrees of freedom
Multiple R-squared:  0.9046,    Adjusted R-squared:  0.8569
F-statistic: 18.96 on 3 and 6 DF,  p-value: 0.001831
```

CAUTION: in R using an * automatically includes all lower order effects (i.e. main effects)

In SAS this is NOT the case and need to hard code this.

Interpretation of Interaction Coefficients

- We cannot interpret the interaction effect beta coefficient directly
- $\hat{\beta}_0$ - the mean fungus content is 13.4 micrograms per 100 grams of nut when the area has:
  - Average rainfall of 0 cm per week
  - The noon temperature is low
- $\hat{\beta}_{rain}$ - A 1 cm/week increase in rainfall is associated with an increase of 4.7 micrograms of fungus per 100 grams of nut **when** average noon temperature is low
- $\hat{\beta}_{rain} + \hat{\beta}_{rain*noon\_temp}$ - A 1 cm/week increase in rainfall is associated with an increase of 11 (4.7 + 6.3) micrograms of fungus per 100 grams of nut **when** average noon temperature is high

Hypothesis tests for individual regression coefficients

$$H_0 : \beta_i = 0$$

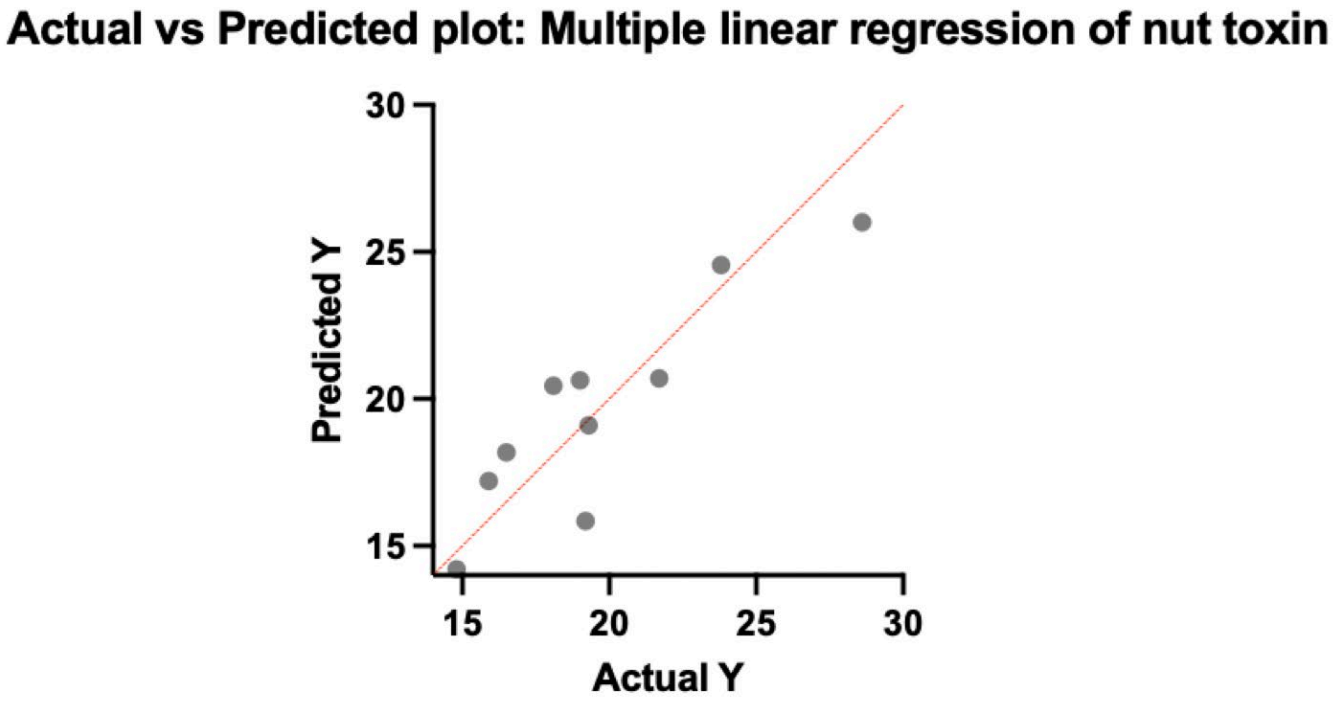
$$H_a : \beta_i \neq 0$$

Null Hypothesis: There is no linear association between the explanatory variable i and the outcome when the other explanatory variables are held constant

Alternative Hypothesis: There is an association between the explanatory variable i and the outcome when the the other explanatory variables are held constant.

Linear Regression Assumptions

- **Linearity assumption**: The relationship between the mean response and the explanatory variables is linear in the explanatory variables.



Linear Regression Assumptions

- **Assumptions about errors:**
  1. The errors are assumed to follow a normal distribution.
  2. The errors have a mean of zero.
  3. The errors have constant variance (homoscedasticity or homogeneity of variance).
  4. The errors are independent of one another.

Linear Regression Assumptions

- **Assumptions about explanatory variables**
  1. The explanatory variables are nonrandom (fixed).
  2. The explanatory variables are not highly correlated with one another.

Linear Regression Assumptions

- **Assumption about observations:** All observations are equally reliable and have approximately equal role in determining the regression results in influencing conclusions.

Model Selection

- If we have several explanatory variables, there are many models we may potentially fit. How do we choose which one is best?

- There is no best model. In general, we want the simplest model that adequately explains the data, e.g., the most parsimonious model.

Backward selection

- Start with all variables and drop one variable at a time.
- The variable with the largest p-value is considered for removal. The variable is removed if it is not significant, and a new model (excluding that variable) is fit.
- Variables are removed in this way until all variables in the model are significant.

Forward selection

- Start with only an intercept in the model.
- Look for the variable most associated with the response variable. If this regression coefficient is significantly different from zero a search for a second variable is made.
- The next variable selected is the most highly correlated with the residuals from the model with the first explanatory variable. If this regression coefficient is significantly different from zero, a search for a third regression variable is made.
- This process continues until there are no additional variables that are significantly different from zero.

Evaluate potential confounding

- Instead of solely relying on p-values, look at the effect size f your primary explanatory variable (i.e. β estimate) change between including and excluding the potential confounder
- If percent change ≥ 10%, include the variable

$$Percent\ Change = \frac{\beta_{with\ potential\ confounder} - \beta_{without\ potential\ confounder}}{\beta_{without\ potential\ confounder}} * 100$$

Choosing a model

- “When there is sound theoretical literature available, you base your model upon what past research tells you.”
  - e.g., in human genetics studies age, gender, and ethnicity are often always included in the regression model regardless of whether or not they significantly contribute.
- Use caution when using an automated selection method (i.e., forward or backward selection)
  - These methods rely on a mathematical criterion instead of human intelligence.
  - Variable selection may depend upon only slight differences in significance. Slight numerical differences can lead to major differences in the model.
  - Backward selection is preferred (not applicable if the number of predictors is larger than the number of observations).

Model Fit: Coefficient of Determination

Coefficient of determination, R^2 - measures the proportion of total response variation explained by the linear regression model with *p-1* explanatory variables compared to the regression model with only an intercept.

- R^2 measures the relative reduction in the residual sum of squares for our regression model in comparison to the model having only an intercept.
- When R^2 is close to 1, the regression model explains a high proportion of the variation in the observed data compared to the simplest model.

Adjusted Coefficient of Determination

- Since the coefficient of determination, R^2, measures the proportion of additional variance explained by the regression model in comparison to the simple model (just an intercept), it may seem like a reasonable statistic to choose between models.
- However, R^2 will never decrease, and almost always will increase, as you add explanatory variables to the model.
  - If R^2 is used to choose between models, we will always choose the model with the most explanatory variables.
  - This is **NOT** a good thing. We may make our model too complex. This is called overfitting.

Summary of Multiple Linear Regression

- Regression can be extended to multiple regression, allowing several factors to participate in the prediction of the dependent variable.
- Regression coefficients are interpreted in the context that the other explanatory variables are held constant.
- The most parsimonious model can be found using either backward (preferred when the number of explanatory variables is smaller than the number of observations) or forward selection.

Logistic Regression

Regression Analyses

What is Logistic Regression?

Logistic regression is a type of regression used when the responses are binary.

Why not use linear regression?

$$\pi = P(Y = 1|X) = \mu_{Y|X} = \beta_0 + \beta_1 X$$

1. $0 \leq \pi \leq 1$, but the linear regression model doesn't make this constraint.
2. $\mathrm{Var}(Y) = \pi(1\text{-}\pi)$, where π is dependent on x. Therefore, the variance of Y is not constant for all values of the explanatory variable, i.e., homogeneity of variance isn't satisfied.
3. If $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, then the errors ϵ_i are not normally distributed.

Generalized linear model

- Logistic regression is a type of *generalized* linear model.
- A generalized linear model is a model in which the mean relationship between the response variable and the explanatory variables is linear in the regression coefficients **AFTER** some transformation of the mean function.
- In this case, the logit transformation does the work for us.

$$logit\left(\pi\right) = ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

Probability of an event vs. logit of the probability

Scenario	Odds = P/(1-P)	Logit = ln(Odds)
No probability of an event (P=0)	0/(1-0) = 0	ln(0) = -∞
Neutral point (P=0.5)	0.5/(1-0.5) = 1	ln(1) = 0
Certainty of about event (P=1)	1/(1-1) = ∞	ln(∞) = +∞

Logistic Regression Example

- A questionnaire has been used to gather demographics data from some pharmacists, the data include their age and gender.
- It also asked ‘Would you personally be happy to provide training in the use of the App? (Yes/No)’
- We are concerned with the influence of age on pharmacists’ willingness to train patients in use of a phone App.

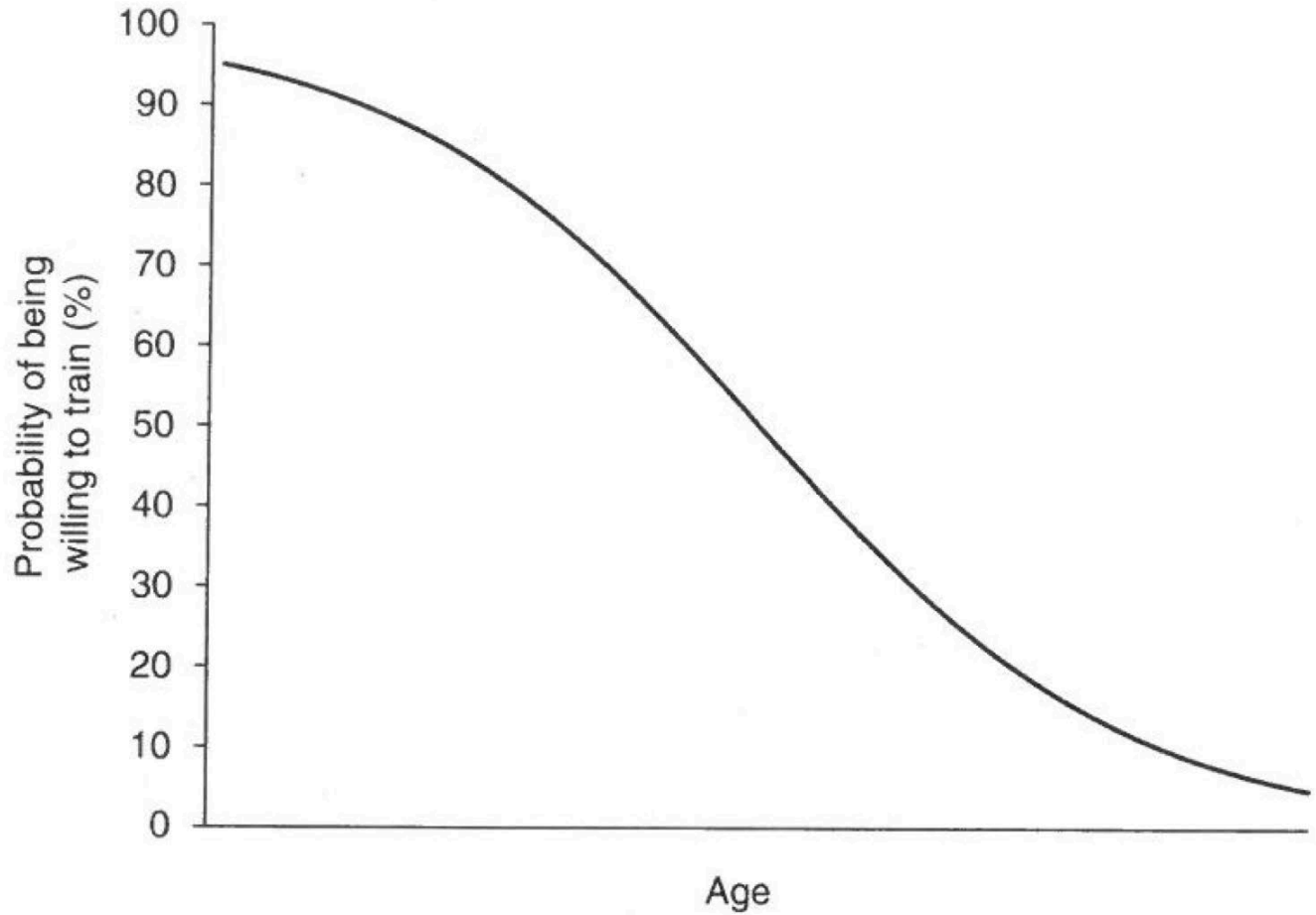


Figure 20.3 A more realistic, sigmoidal relationship between willingness to provide training and age

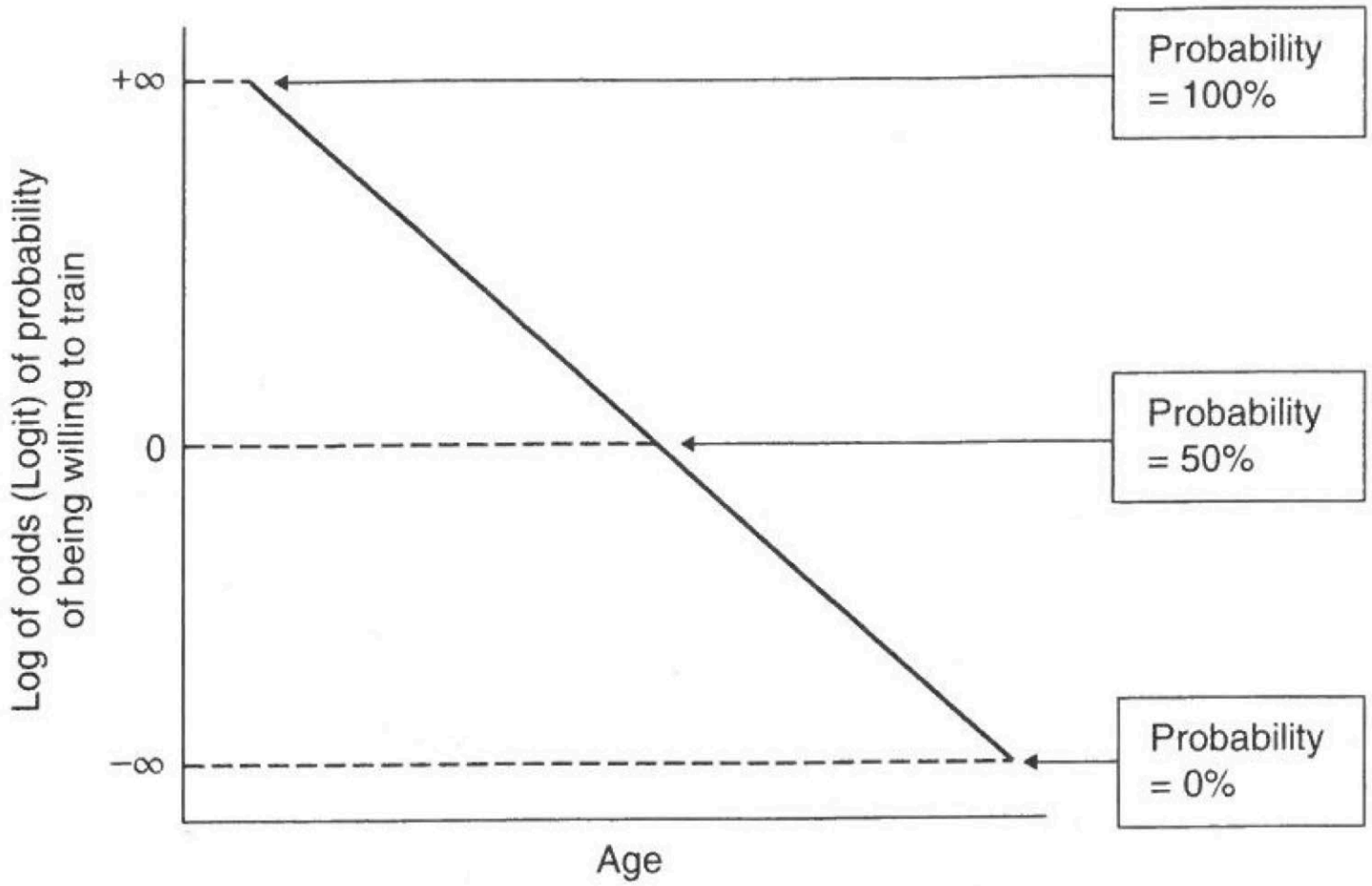
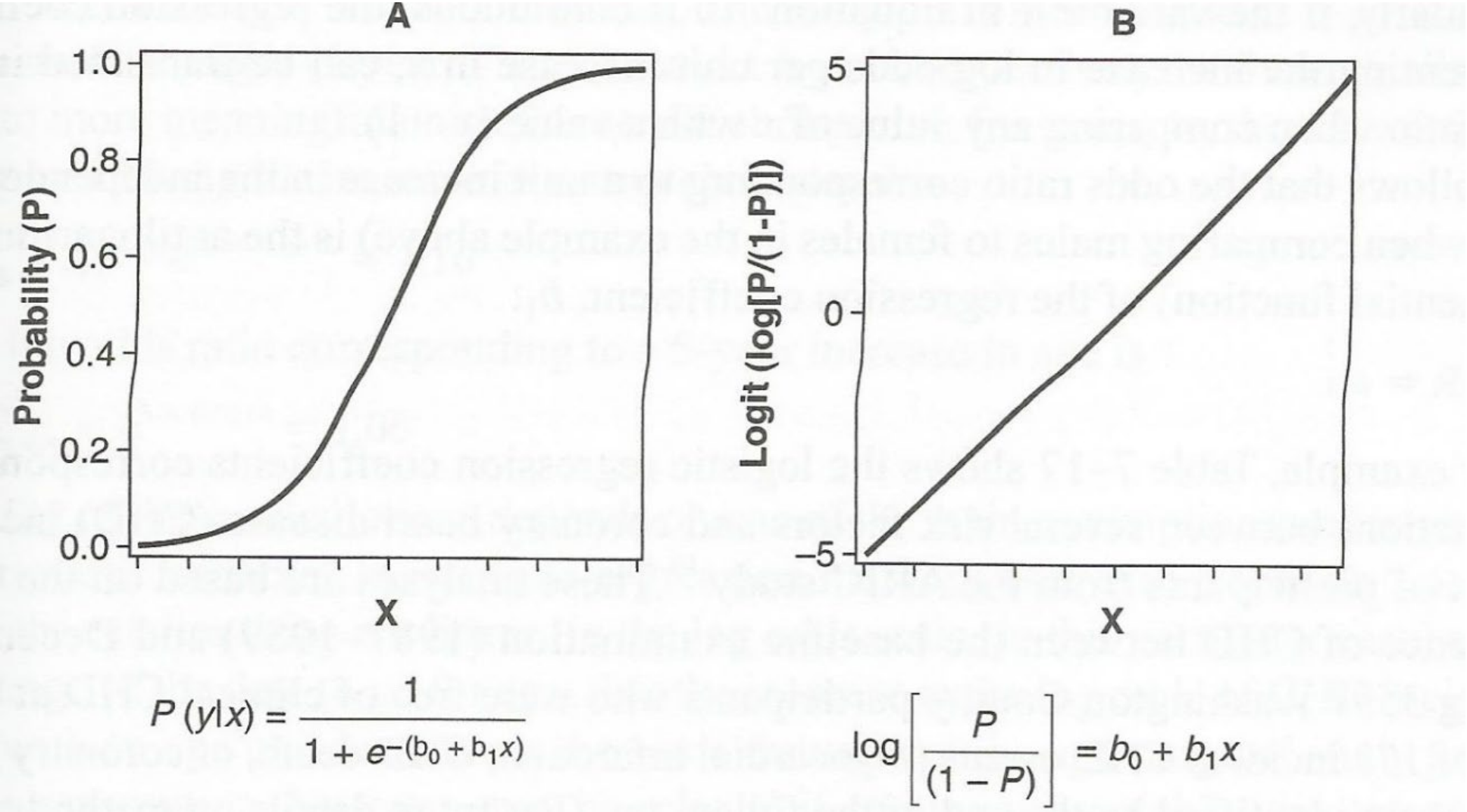


Figure 20.4 The logit of the probability of being willing to train can take values between plus infinity (Corresponds to 100% probability) and minus infinity (0% probability). A logit of zero corresponds to 50% probability



Szklo & Nieto, Beyond the Basics, 2nd Ed, Ch.7

Estimation of Logistic Regression Coefficients

- Logistic regression coefficients are estimated using maximum likelihood fitting.
- This means we estimate β_0 and β_1 with the values that are most likely to have produced the data.
- We will skip the horrifying mathematical details!

Interpretation of Logistic Regression Parameters

Recall: $\pi = P(Y = 1|X)$

Thus, $logit(\pi) = ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$ is the natural logarithm of the odds in favor of Y=1.

Then the odds in favor of Y=1 is

$$\exp(\beta_0 + \beta_1 * X)$$

Let's compare the ratio of the odds when the explanatory variable is X+1 relative to when it is X, i.e., X increases by 1.

The odds ratio in this case is

$$\frac{odds(Y=1|X+1)}{odds(Y=1|X)} = \frac{exp\left(\beta_0+\beta_1(X+1)\right)}{exp\left(\beta_0+\beta_1(X)\right)}$$
$$= exp(\beta_1)$$

Thus, the odds change by a multiplicative factor of $\exp(\beta_1)$ when X increases by 1 (i.e. effect of 1 unit of X has an odds ratio of $\exp(\beta_1)$)

Coefficients and Odds Ratios (ORs)

- **Regression coefficients** can range from $-\infty$ to $+\infty$. Values less than 0 indicate a reduction in the probability of an event. Values greater than 0 indicate an increase in the probability of an event.

- **Odds Ratios** can range from 0 to $+\infty$. Values less than 1 indicate a reduction in the probability of an event. Values greater than 1 indicate an increase in the probability of an event.

Odds Ratio from Coefficients

$$logit\,(\text{Probability of willingness to train}) = 4.605 - 0.0921 \times Age$$

$$OR = exp(-0.0921) = 0.912$$

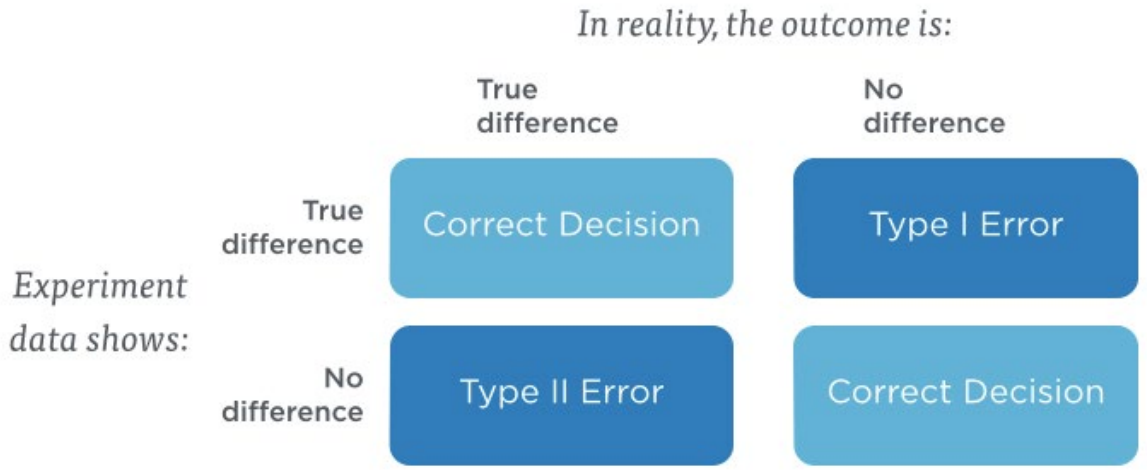
If one pharmacist is one year older than another, the odds that the older pharmacist will be willing to train will be 0.912 times those for the younger individual.

Assessing the effectiveness of the model

Can visualize with a 2x2 contingency table

Classify each participant (or sample) as a:

- **true positive** - predicted to have event and had event
- **true negative** - predicted to not have an event and did not have an even
- **false positive**- predicted to have event, but did NOT have event
- **false negative** - predicted to NOT have event, but had event



Classifying participants in example

	Observed to be willing	Observed to be not willing
Predicted as willing	True positive = 116	False positive = 45
Predicted as not willing	False positive = 29	True negative = 46

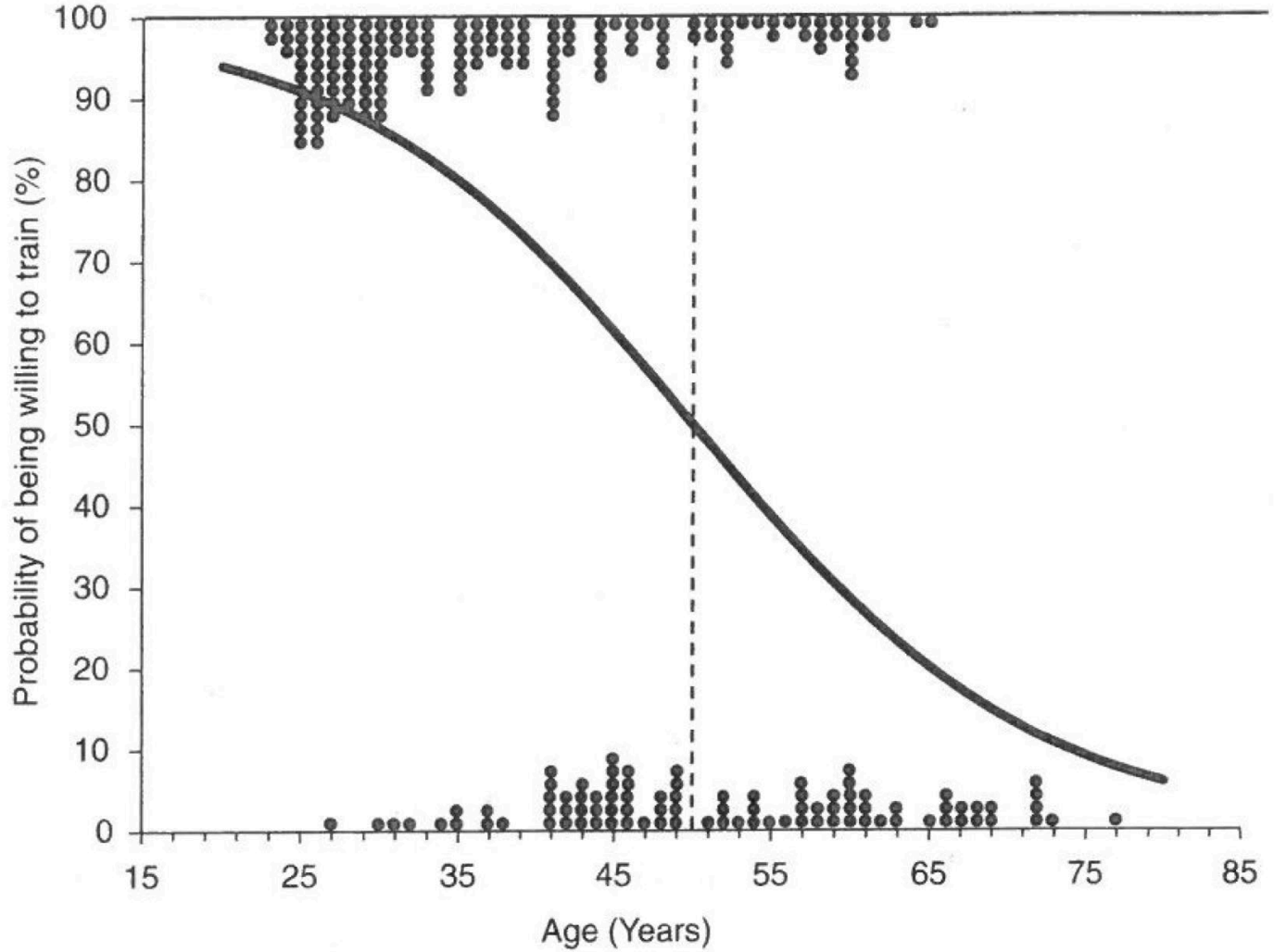


Figure 20.6 True and false classifications of willingness to train, achieved by logistic regression based on age

Code/Results in R

```r
log_mod <- glm(Would_train ~ Age + Gender, data=willing, family = "binomial")
summary(log_mod)
```

```
Call:
glm(formula = Would_train ~ Age + Gender, family = "binomial",
    data = willing)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1138  -0.8938   0.4759   0.7868   1.7942

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.60567    0.62828   7.331 2.29e-13 ***
Age         -0.09203    0.01372  -6.706 2.00e-11 ***
GenderM     -0.01010    0.32196  -0.031    0.975
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 314.70  on 235  degrees of freedom
Residual deviance: 249.74  on 233  degrees of freedom
AIC: 255.74

Number of Fisher Scoring iterations: 4
```

10 example observations from study about pharmacists willingness to train patients on device

Age	Gender	Would_train
44	F	1
25	F	1
30	M	1
27	F	1
26	F	1
46	M	0
33	F	1
43	F	0
30	F	1
59	M	1

Code/Results for ORs

```r
# get odds ratios and their confidence intervals
exp(cbind(coef(log_mod),confint(log_mod)))
```

```
Waiting for profiling to be done...
                                  2.5 %       97.5 %
(Intercept) 100.0503976 30.9718057 366.9704989
Age           0.9120775  0.8866970   0.9358826
GenderM       0.9899492  0.5289607   1.8762006
```

In the multiple logistic regression model, age was negatively associated with the probability of the pharmacist being willing to train patients (odds ratio = 0.91; 95% CI for OR = 0.89 to 0.94; p-value<0.0001). There was no association between gender and the probability of a pharmacist being willing to train patients (p-value = 0.98).

Summary of Logistic Regression

- Logistic regression allows us to determine the influences nominal and/or interval factors have on a dichotomous nominal outcome.
- With a logistic regression equation, we can predict the likelihood that a given individual will fall into a particular category by taking account of one or more factors.
- The Odds Ratio of a factor describes the multiplicative effect on the odds of an event when the factor is increases in the value by 1.0.

Conclusions

What Did We Learn?

- Regression analysis is often used when the goal is to predict the outcome based on one or more variables or when the goal is to understand the relationship between two variables when ”controlling” for other variables, i.e., when other variables don’t change.
- There are several types of regression analyses that depend on the nature of the outcome variable, e.g., continuous, binary.
- Effect size between a predictor and the outcome is measured as a ‘regression coefficient’.