10/17/2024 (Week 8)

Jingjing Yang, PhD

Associate Professor of Human Genetics

Jingjing.yang@emory.edu

Outline

1

Pearson's Correlation Test

2

Linear Regression

- Single variant regression
- Multivariate regression

3

Generalized Linear Regression

- Logistic regression

Study relationship between two variables (X, Y)

- Hypothesis testing : e.g., t-test
- Pearson's correlation coefficient r
  - Unit free
  - Dose not depend on number of samples

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

Pearson's Correlation Test

- $H_0: r = 0; \quad H_a: r \neq 0$

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

Pearson's Correlation Test

- **Under $\boldsymbol{H_0: r = 0}$**, with sample size n, the standard error of the correlation coefficient r is given by

$$\sigma_r = \frac{1 - r^2}{\sqrt{n-2}}$$

- **Under $\boldsymbol{H_0}$**: Test statistic t follows a **Student's t-distribution** with degrees of freedom $n - 2$

$$t = \frac{r}{\sigma_r} = r\sqrt{\frac{n-2}{1-r^2}}$$

Q1: Why is the Pearson’s Correlation Test follows a Student’s t-distribution with degrees of freedom of n-2?

Answer

Estimating the population standard deviation has N–1 degrees of freedom because the standard deviation requires an estimate of the mean, so the univariate t-test has N–1 df.

In contrast, in the case of the correlation coefficient, we have to estimate two population standard deviations (one for X and one for Y), plus the covariance.

To do this, we need estimates of both means, so the degrees of freedom for a correlation is N–2.

Q2: What is p-value?

what is p-value

A p-value is a statistical measure that helps you determine the significance of your results in hypothesis testing. Specifically, it represents the probability of obtaining results at least as extreme as those observed in your sample data, assuming that the null hypothesis is true.

Here's a breakdown:

- **Null Hypothesis (H0)**: This is the default assumption that there is no effect or no difference.
- **Alternative Hypothesis (H1)**: This is what you want to prove, indicating that there is an effect or a difference.
- **Low p-value (typically ≤ 0.05)**: Suggests that the observed data is unlikely under the null hypothesis, leading you to reject H0.
- **High p-value (> 0.05)**: Indicates that the observed data is consistent with the null hypothesis, and you fail to reject H0.

It's important to remember that a p-value does not measure the size of an effect or the importance of a result; it merely indicates whether the observed data are statistically significant.

Pearson's Correlation Test: cor.test()

```r
cor.test(~ age + wholeWeight, data = abalone, alternative = "two.sided",
         method = "pearson")
```
```

```

```
	Pearson's product-moment correlation

data:  age and wholeWeight
t = 41.498, df = 4175, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5185606 0.5615148
sample estimates:
      cor
0.5403897
```

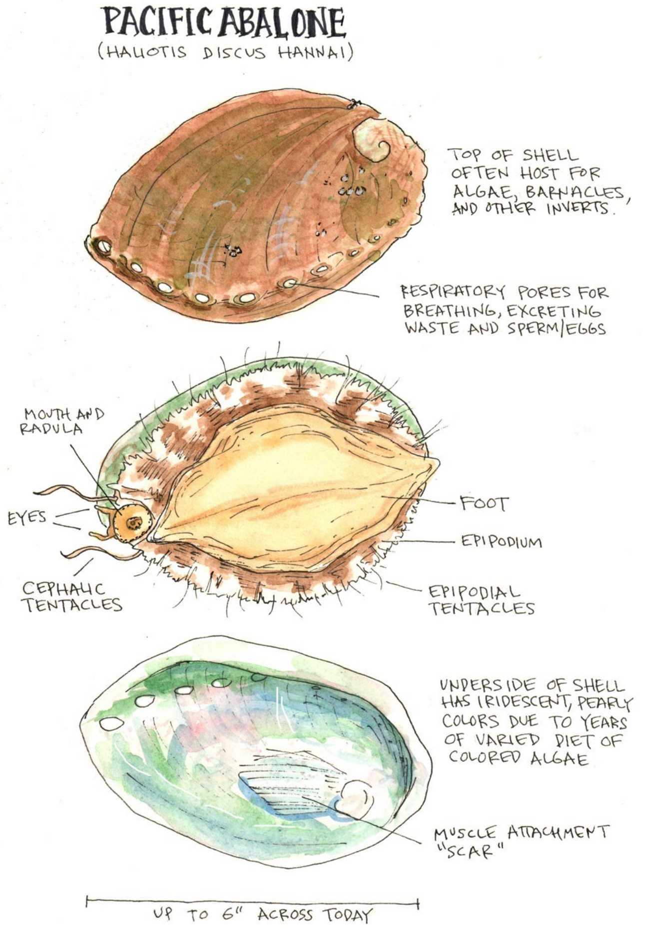
Beyond Simple Hypothesis Testing

- Quantify correlation between two variables
- Quantify correlation between one outcome variable and multiple predictor variables
- Account for confounding factors in the test
- Predict one outcome variable by using one or multiple predictor variables

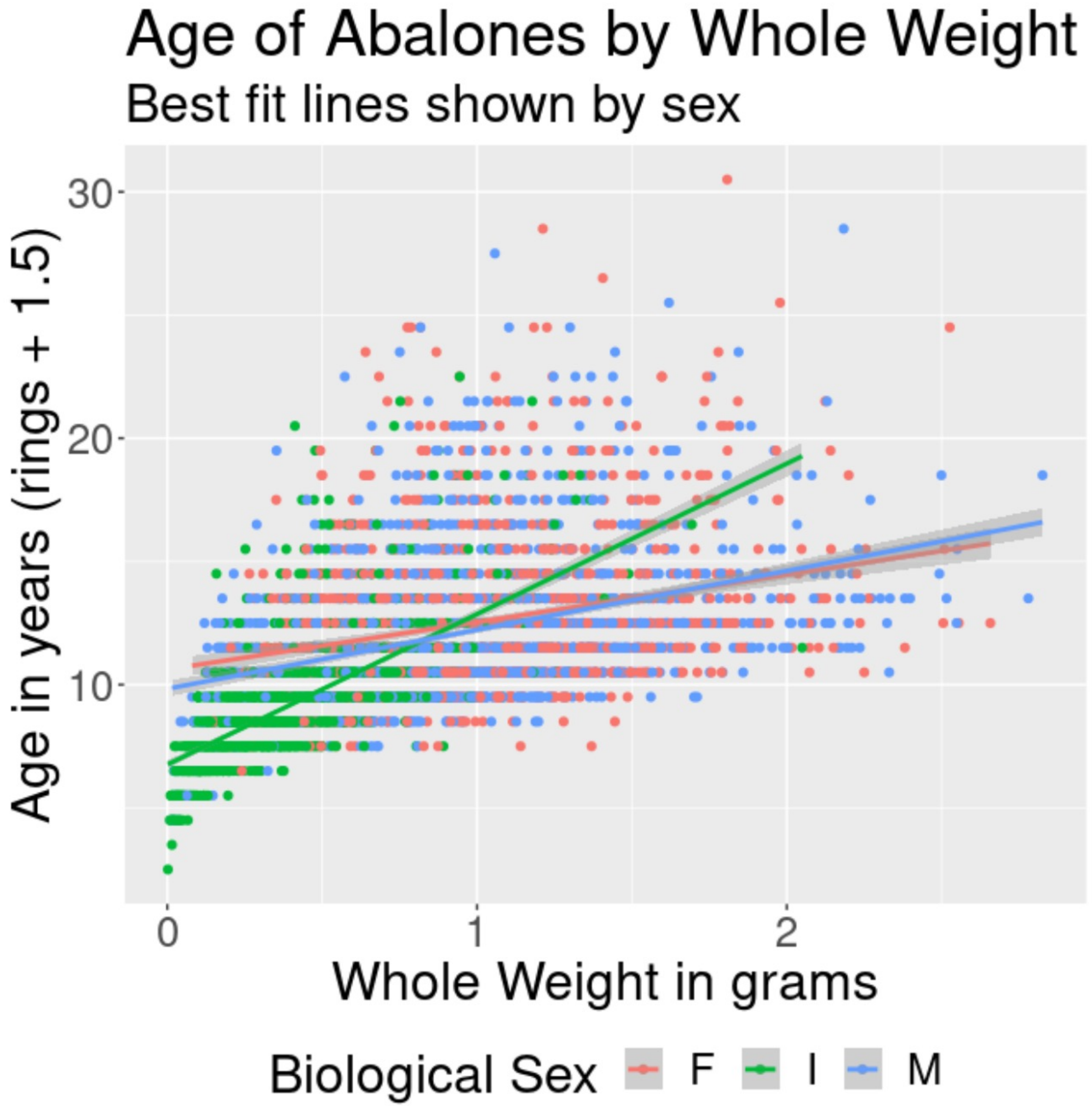
Relationship between one response variable and multiple predictor variables?

Abalones Dataset

Name	Data Type	Measurement Unit	Description
Sex	nominal	–	M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer	–	+1.5 gives the age in years



Relationship between Abalone age/rings and Whole Weight



Regression

- Technique used for the modeling and analysis of numerical data
- Exploits the relationship between two or more variables so that we can gain information about one of them through knowing values of the other
- Regression can be used for prediction, estimation, hypothesis testing, and modeling causal relationships

Linear Regression

Single variant linear regression model

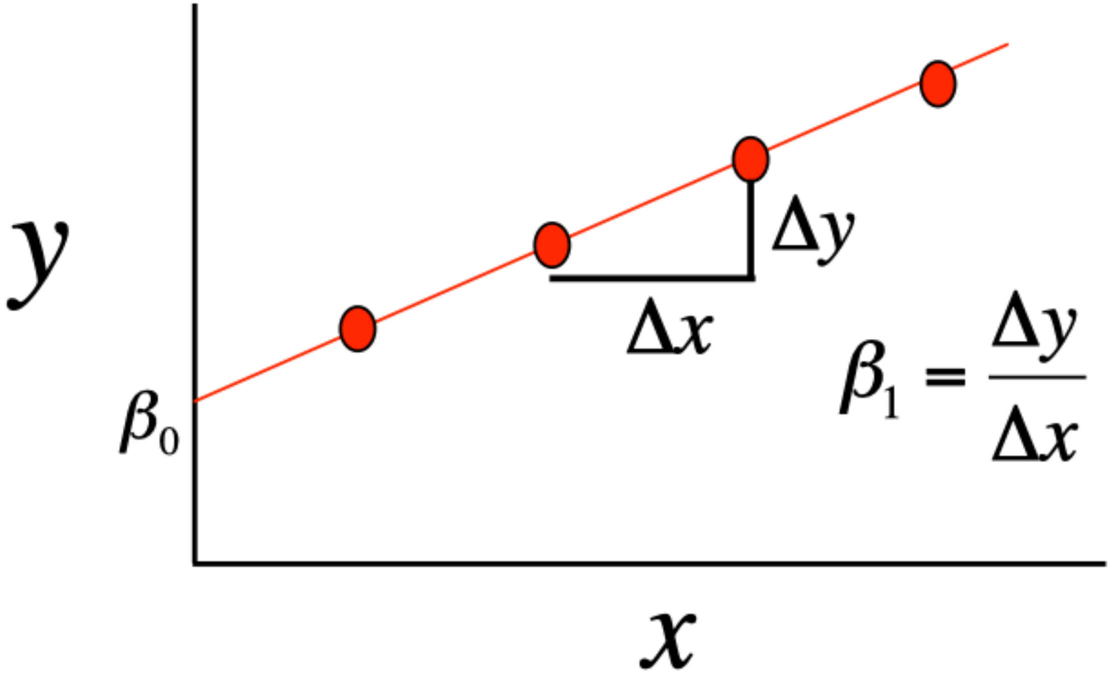
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{i=1,…, n}$$

- x_i : Independent (explanatory, predictor, covariate) Variable value for sample i
- y_i : Dependent (response, outcome) Variable value for sample i
- β_0 : Intercept of the fitted linear line
- β_1 : Slope of the fitted linear line, coefficient of X
- $\varepsilon_i \sim N(0, \sigma^2)$: Residual value for sample i

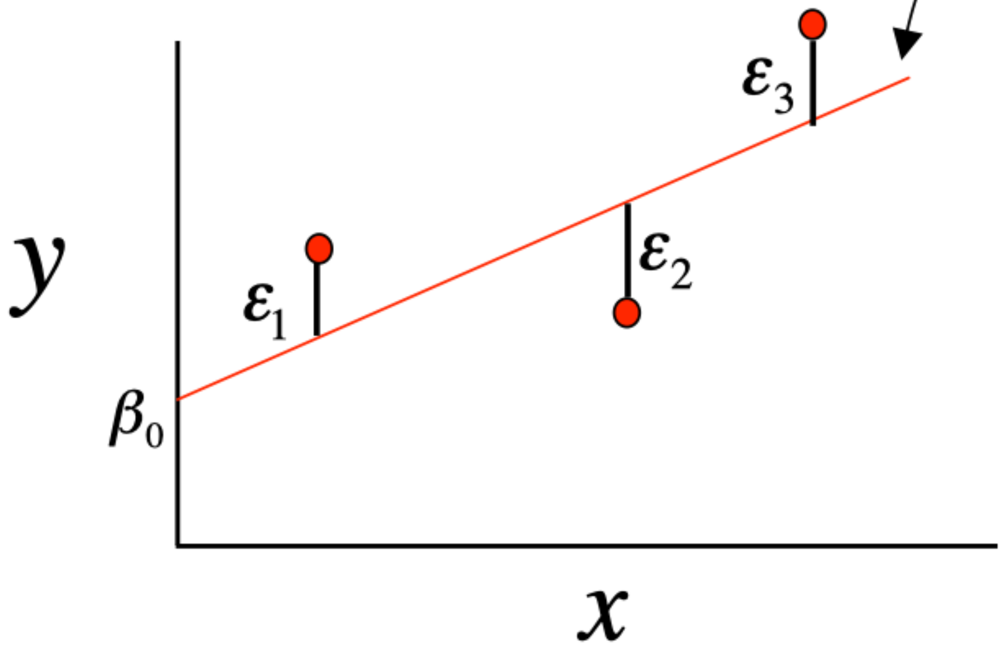
How to fit the model?

- How to find the linear line by estimating the intercept β_0 and slope β_1 ?

$$y = \beta_0 + \beta_1 x$$



Residuals in the linear regression model

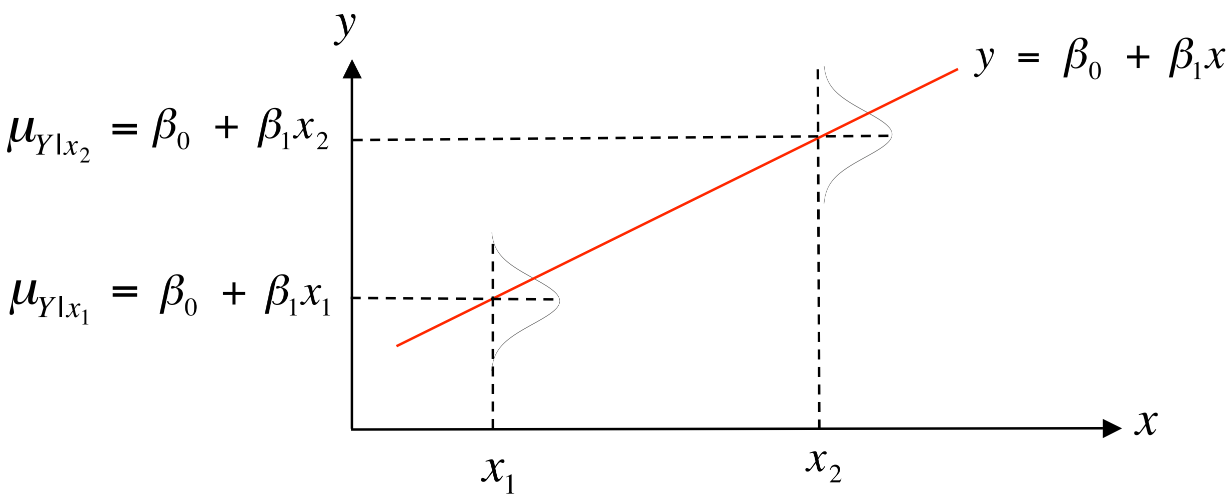


True Regression Line

$$y = \beta_0 + \beta_1 x$$

The expected value of the outcome variable Y is a linear function of the predictor X

Graphical Interpretation

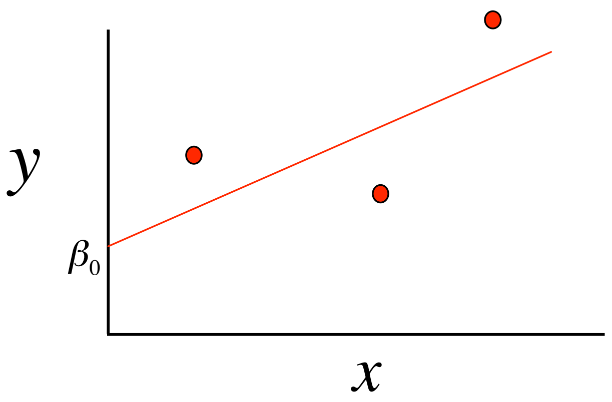


- For example, if x = height and y = weight then $\mu_{Y|x=60}$ is the average weight for all individuals 60 inches tall in the population

Ordinary Least Square Estimates

- Point estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained by the principle of least squares

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$



Calculate the Intercept (β_0): Once you have β_1, you can calculate the intercept using:

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Where:

- $\bar{Y}$ is the mean of the Y values
- $\bar{X}$ is the mean of the X values

Calculate the Slope (β_1): The slope can be calculated using the formula:

$$\beta_1 = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$$

Where:

- n is the number of observations
- $\sum XY$ is the sum of the product of X and Y
- $\sum X$ is the sum of X values
- $\sum Y$ is the sum of Y values
- $\sum X^2$ is the sum of squared X values

Predicted and Residual Values

- ***Predicted***, or fitted, values are values of y predicted by the least-squares regression line obtained by plugging in $x_1, x_2, \ldots, x_n$ into the estimated regression line

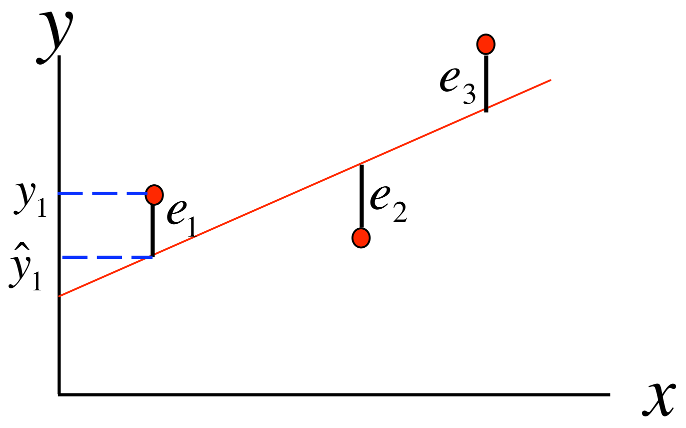
$$\hat{y}_1 = \hat{\beta}_0 - \hat{\beta}_1 x_1$$

$$\hat{y}_2 = \hat{\beta}_0 - \hat{\beta}_1 x_2$$

- ***Residuals*** are the deviations of observed and predicted values

$$e_1 = y_1 - \hat{y}_1$$

$$e_2 = y_2 - \hat{y}_2$$



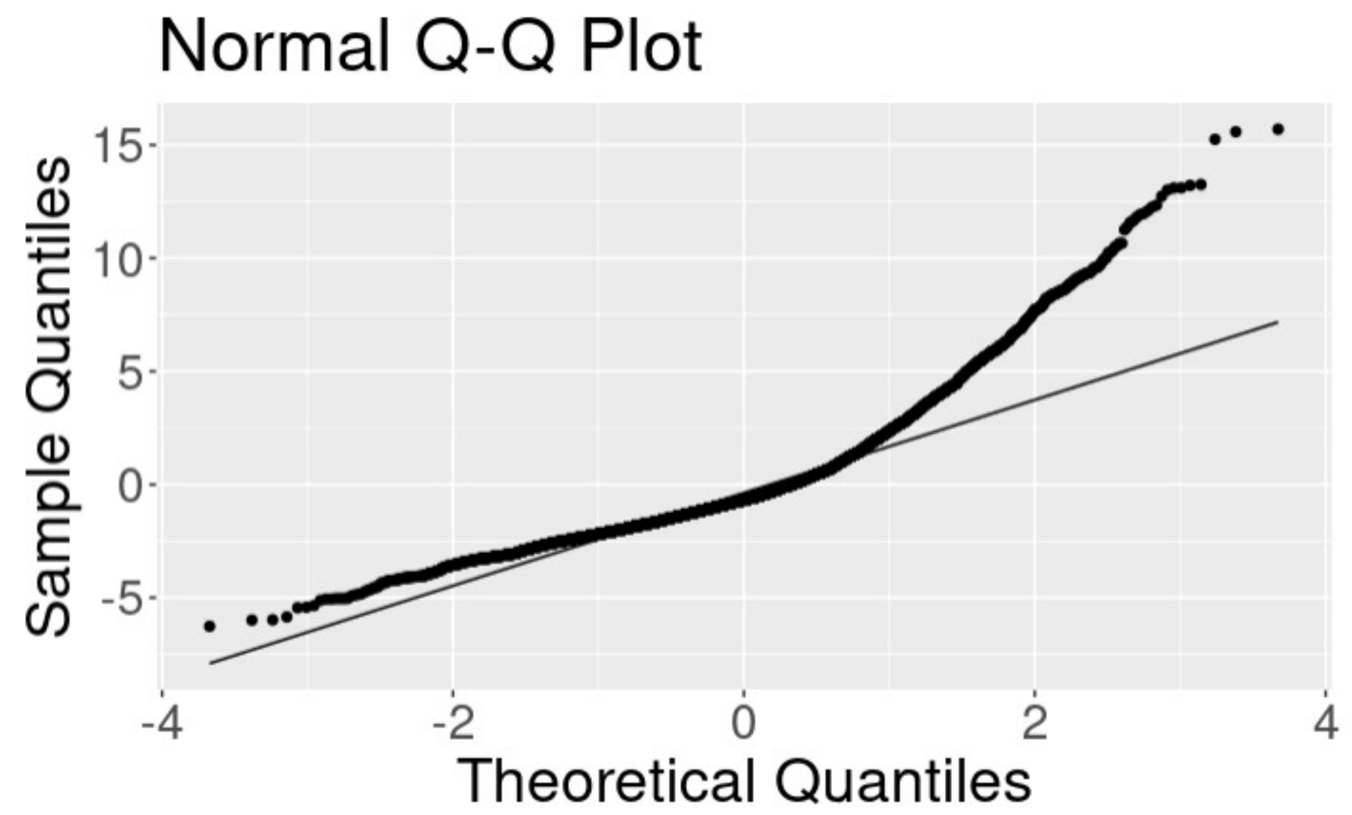
Linear Regression in R by lm()

````
```{r}
fit1 <- lm(age ~ wholeWeight, data = abalone)
summary(fit1)
```
````

```
Call:
lm(formula = age ~ wholeWeight, data = abalone)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.2693 -1.7518 -0.6874  1.0177 15.7029 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.48924    0.08244   103.0   <2e-16 ***
wholeWeight  3.55291    0.08562    41.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.713 on 4175 degrees of freedom
Multiple R-squared:  0.292,	Adjusted R-squared:  0.2919 
F-statistic:  1722 on 1 and 4175 DF,  p-value: < 2.2e-16
```

Check residuals distribution

```{r}
residuals.df <- data.frame(residuals = fit1$residuals)
ggplot(residuals.df, aes(sample = residuals)) +
  stat_qq() + stat_qq_line() +
  labs(x = "Theoretical Quantiles", y = "Sample Quantiles", title = "Normal Q-Q Plot")
```



1. Relationship between rings/age and whole weight while accounting for Sex?
2. Predict Abalone age/rings by multiple measurements?

Abalones Dataset

Name	Data Type	Measurement Unit	Description
Sex	nominal	–	M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer	–	+1.5 gives the age in years

Multivariate Linear Regression

- Extension of the simple linear regression model to two or more independent/predictor variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Exercise: fit the following multivariate linear regression model with the Abalone data.
  - $Age \sim Sex + length\ +\ diameter\ +\ height\ +\ wholeWeight\ +\ shuckedWeight\ +\ wisceraWeight\ +\ shellWeight$

How to quantify categorical independent variable?

Binary variable: coded as 0/1

The sex variable in the abalone dataset has three levels: F, I, M ?

How to quantify categorical independent variable?

- The sex variable in the abalone dataset has three levels: F, M, I?
- Code through (k-1) dummy variables for k levels:

Sex	X1	X2
F	1	0
M	0	1
I	0	0

Fit a multivariate linear regression model with sex and wholeWeight

```{r}
fit2 <- lm(age ~ factor(sex) + wholeWeight, data = abalone)
summary(fit2)
```

```
Call:
lm(formula = age ~ factor(sex) + wholeWeight, data = abalone)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.0404 -1.7442 -0.5449  0.9935 15.7240 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    9.6770     0.1290  74.987  < 2e-16 ***
factor(sex)I  -1.5034     0.1207 -12.454  < 2e-16 ***
factor(sex)M  -0.2684     0.1004  -2.674  0.00753 ** 
wholeWeight    2.8210     0.1013  27.849  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.661 on 4173 degrees of freedom
Multiple R-squared:  0.3195,	Adjusted R-squared:  0.319 
F-statistic: 653.2 on 3 and 4173 DF,  p-value: < 2.2e-16
```

In-Class Exercise : lm()

Question Need to Answer for In-class participation credit.

- What is Regression R-square?
- What dose it mean if you get increased Regression R-square by adding additional predictor variables? The same question is included in Task 5 in Exercise 1. Rmd.

Generalized Linear Regression

What's the difference between general and generalized linear models?

General

$$E[Y] = \beta_0 + \beta_1 X_1$$

$Y \sim N(\mu, \sigma^2)$

Generalized

$$E[g(Y)] = \beta_0 + \beta_1 X_1$$

$$Y \sim \begin{cases} Bernoulli, Binomial \\ Poisson \\ Negative\ binomial \\ etc \end{cases}$$

g ~ "link" function to transform Y

$g(Y) \sim N(\mu, \sigma^2)$

Why generalized?

Apply linear regression to outcome variables that are clearly not normally distributed

- Binary : case/control, yes/no, 0/1

$$Y \sim Bernoulli\ (p), \qquad 0 \le p \le 1$$

- Poisson distributed counts

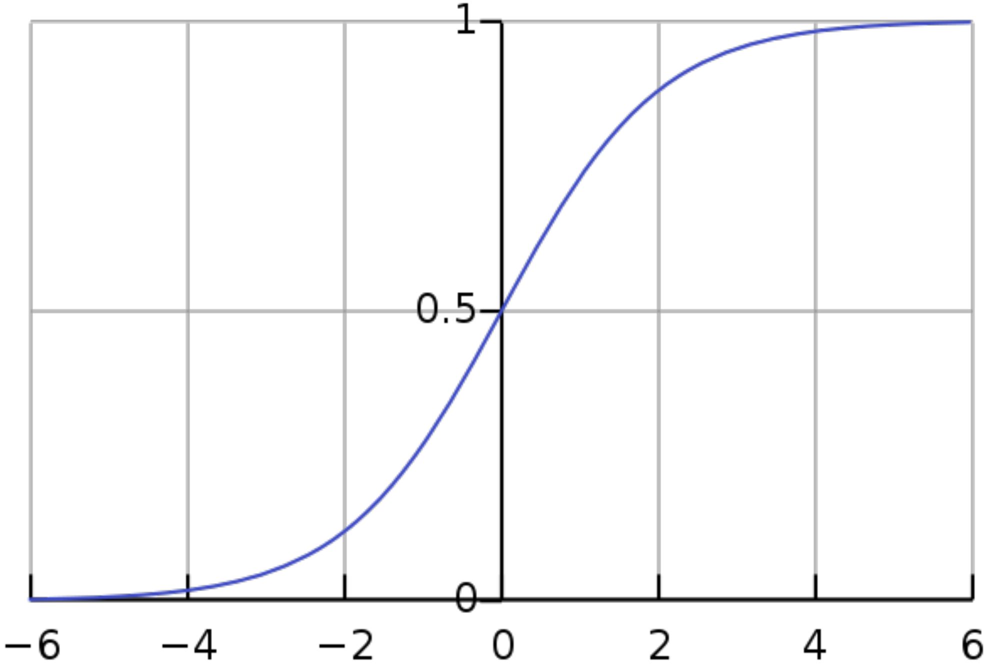
$$Y \sim Poisson\ (\lambda), \qquad \lambda > 0$$

Generalized linear regression model

- The mean/expectation function of Y can usually be expressed as a function of the distribution parameters
  - Binary outcome: $\mathrm{E}[Y]= p$
  - Poisson outcome: $\mathrm{E}[Y]= \lambda$
- Model a linear relation ship between $E[g(Y)]$ and explanatory/independent/predictor variables X

Logistic Regression: $Y \sim Bernoulli\ (p)$

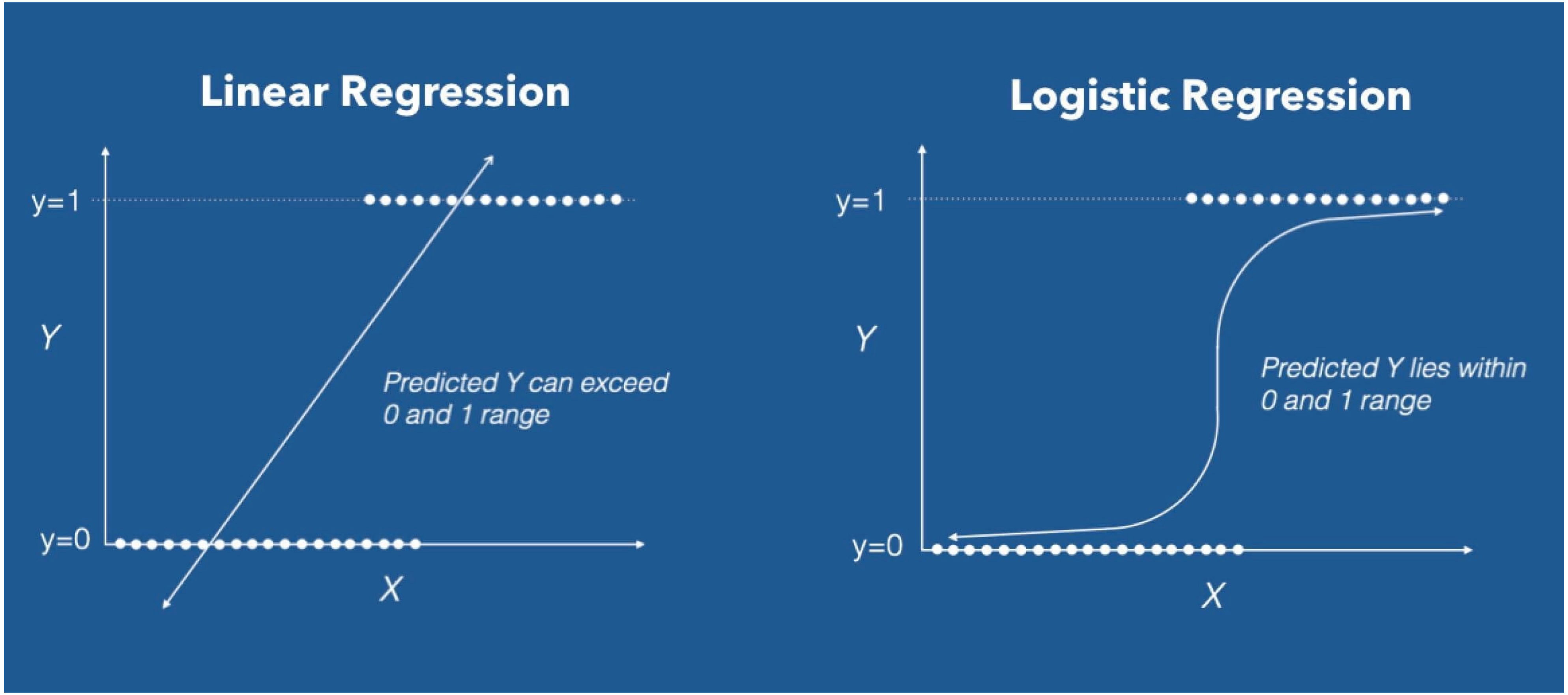
- $l_{\text{LogOdds}} = \log\left(\frac{p}{1-p}\right) = \beta X;\qquad p = Prob(Y = 1)$
- $p = \frac{1}{1+e^{-X\beta}} = \sigma(X\beta),$ Sigmoid function of $X\beta$



$g(E[Y])$ is the log odds of success probability or logit

Model will be fitted by maximizing the likelihood function

Logistic Regression: $Y \sim Bernoulli\ (p)$



Logit link function

Generalized linear model: $log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1$

- A one unit change in X_1 leads to a β_1 change in the log odds
- In terms of odds: $odds(Y=1) = e^{b_0 + b_1 X}$
- In terms of probability or proportion: $\Pr(Y=1) = \frac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}}$

Logit, odds, and probability are different ways of expressing the same thing

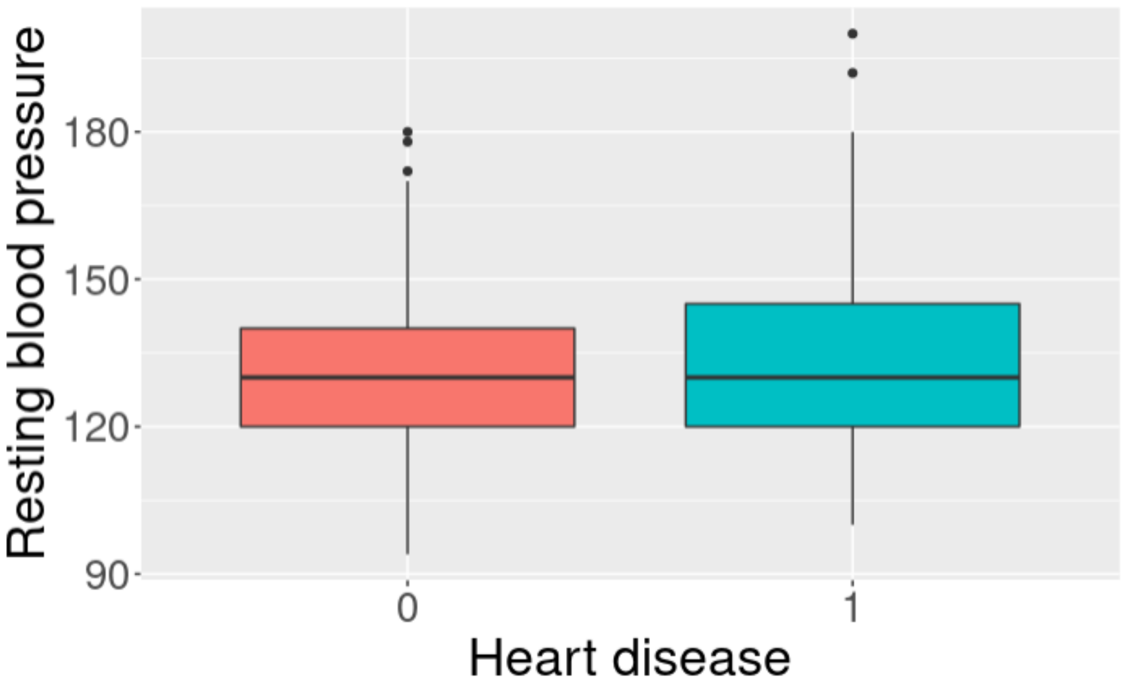
Logit link function

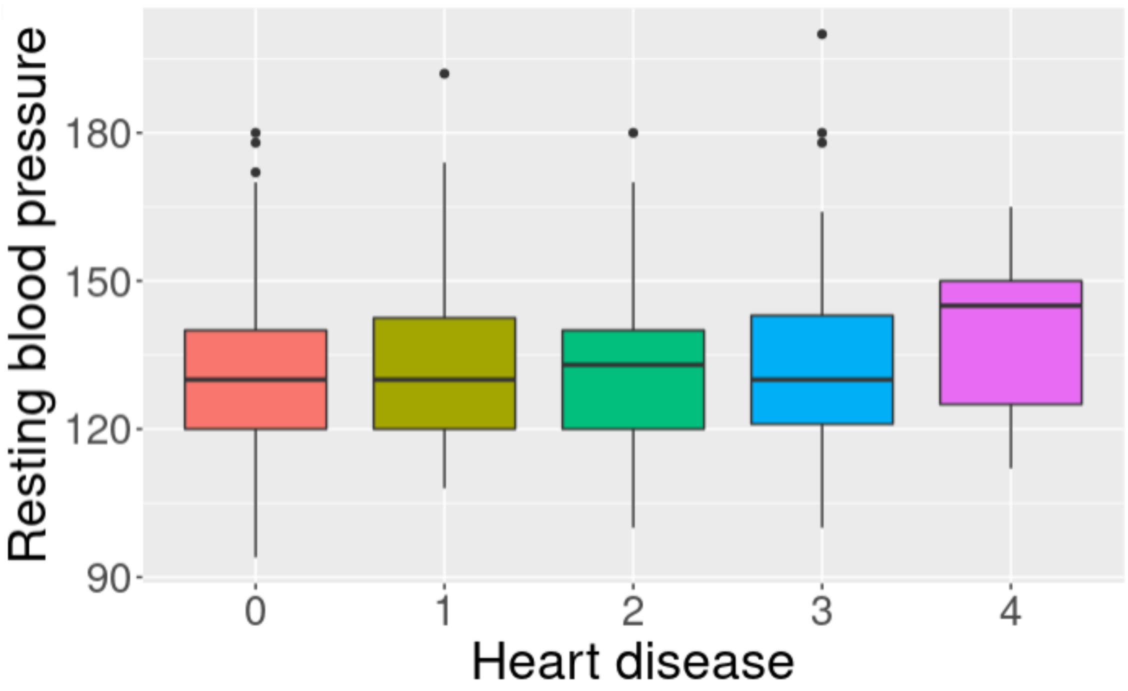
- Logit
  - Natural log (e) of an odds
  - Often called a *log odds*
    - *The logit scale linearizes odds!*
- Logits are continuous and are centered on zero (think of as the z-score for the binomial world!)
  - p = 0.50, odds = 1, then logit = 0
  - p = 0.70, odds = 2.33, then logit = 0.85
  - p = 0.30, odds = .43, then logit = -0.85

Example dataset : Cleveland heart disease

Name	Data Type	Description
age	continuous	age in years
sex	binary	1 = male; 0 = female
cp	categorical	chest pain type – 1: typical angina; 2: atypical angina; 3: non-anginal pain; 4: asymptomatic
trestbps	continuous	resting blood pressure (in mm Hg on admission to the hospital)
chol	continuous	serum cholestoral in mg/dl
fbs	continuous	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg	continuous	resting electrocardiographic results – 0: normal; 1: having ST-T wave abnormality; 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
thalach	continuous	maximum heart rate achieved
exang	binary	exercise induced angina (1 = yes; 0 = no)
oldpeak	continuous	ST depression induced by exercise relative to rest
slope	categorical	the slope of the peak exercise ST segment– 1: upsloping; 2: flat; 3: downsloping
ca	continuous	number of major vessels (0-3) colored by flourosopy
thal	categorical	3 = normal; 6 = fixed defect; 7 = reversable defect
disease	categorical	absence (0) vs. presence (1, 2, 3, 4)

Study the relationship between resting blood pressure would affect heart disease presence





Study the relationship between resting blood pressure would affect heart disease presence

```
        Pearson's product-moment correlation

data:  HD and trestbps
t = 2.647, df = 301, p-value = 0.008548
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.03880692 0.25910016
sample estimates:
      cor
0.1508254
```

Study the relationship between resting blood pressure would affect heart disease presence

```
        Welch Two Sample t-test

data:  trestbps by HD
t = -2.6152, df = 274.64, p-value = 0.009409
alternative hypothesis: true difference in means between group 0 and group 1 is not equal
to 0
95 percent confidence interval:
 -9.321775 -1.314915
sample estimates:
mean in group 0 mean in group 1
       129.2500        134.5683
```

Logistic Regression: HD ~ trestbps

```{r}
fit3 <- glm(HD ~ trestbps, data = cleveland, family = "binomial")
summary(fit3)
```

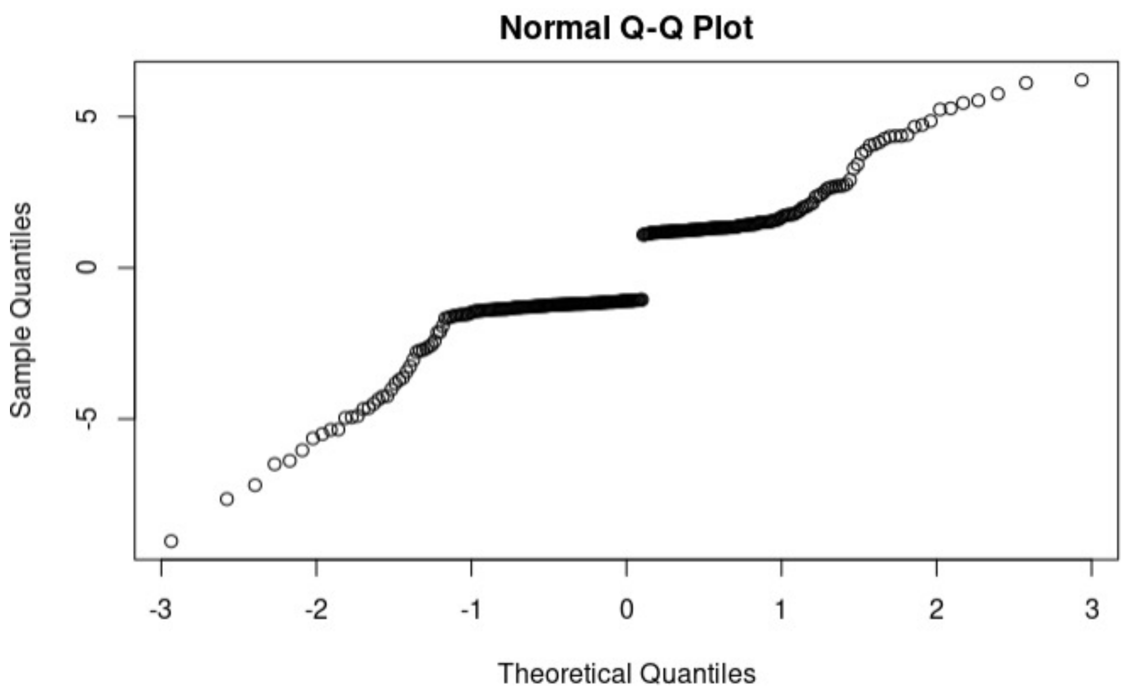
```
Call:
glm(formula = HD ~ trestbps, family = "binomial", data = cleveland)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4773  -1.0948  -0.9414   1.2394   1.4966

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.483687   0.903634  -2.749  0.00599 **
trestbps     0.017587   0.006796   2.588  0.00966 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.98  on 302  degrees of freedom
Residual deviance: 411.03  on 301  degrees of freedom
AIC: 415.03

Number of Fisher Scoring iterations: 4
```

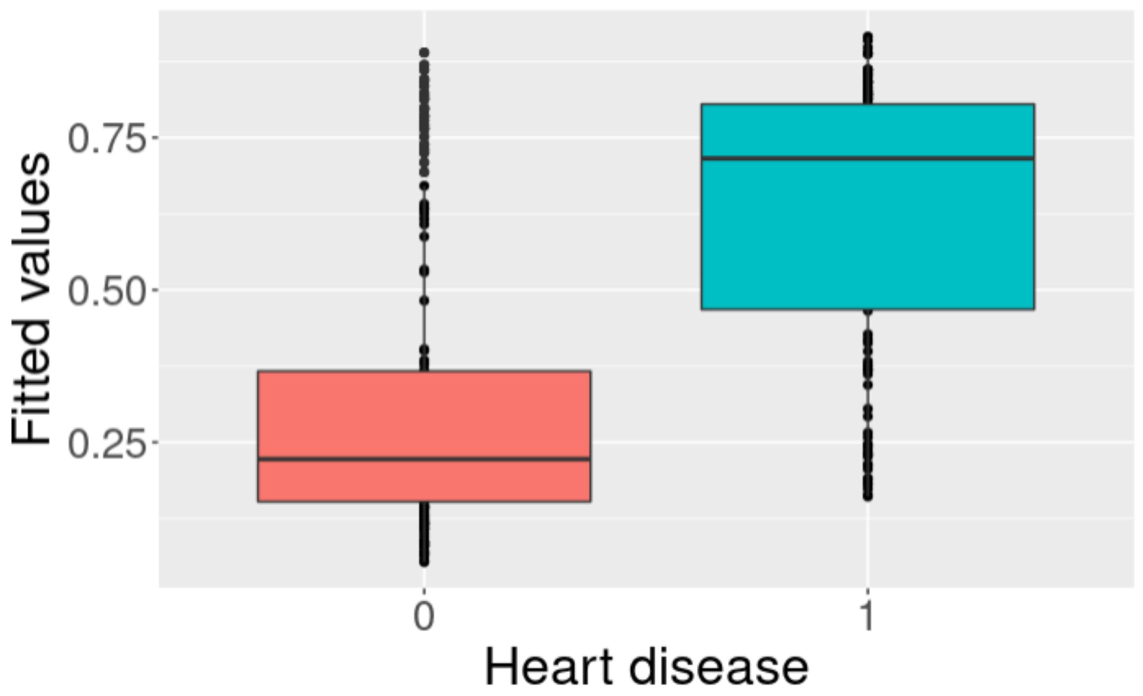
Account for age, sex, and thal

```{r}
fit4 <- glm(HD ~ age + sex + trestbps + factor(thal), data = cleveland, family = "binomial")
summary(fit4)
```

```
Call:
glm(formula = HD ~ age + sex + trestbps + factor(thal), family = "binomial",
    data = cleveland)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0986  -0.7282  -0.4232   0.7656   1.9112

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.735162   1.360779  -4.215 2.50e-05 ***
age            0.052540   0.016683   3.149  0.00164 **
sex            0.773658   0.339110   2.281  0.02252 *
trestbps       0.009081   0.008436   1.076  0.28175
factor(thal)6  1.511252   0.561693   2.691  0.00713 **
factor(thal)7  2.140144   0.306639   6.979 2.97e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 415.20  on 300  degrees of freedom
Residual deviance: 311.38  on 295  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 323.38

Number of Fisher Scoring iterations: 4
```

Logistic regression results





Generalized linear model families

Normal outcome
- Gaussian

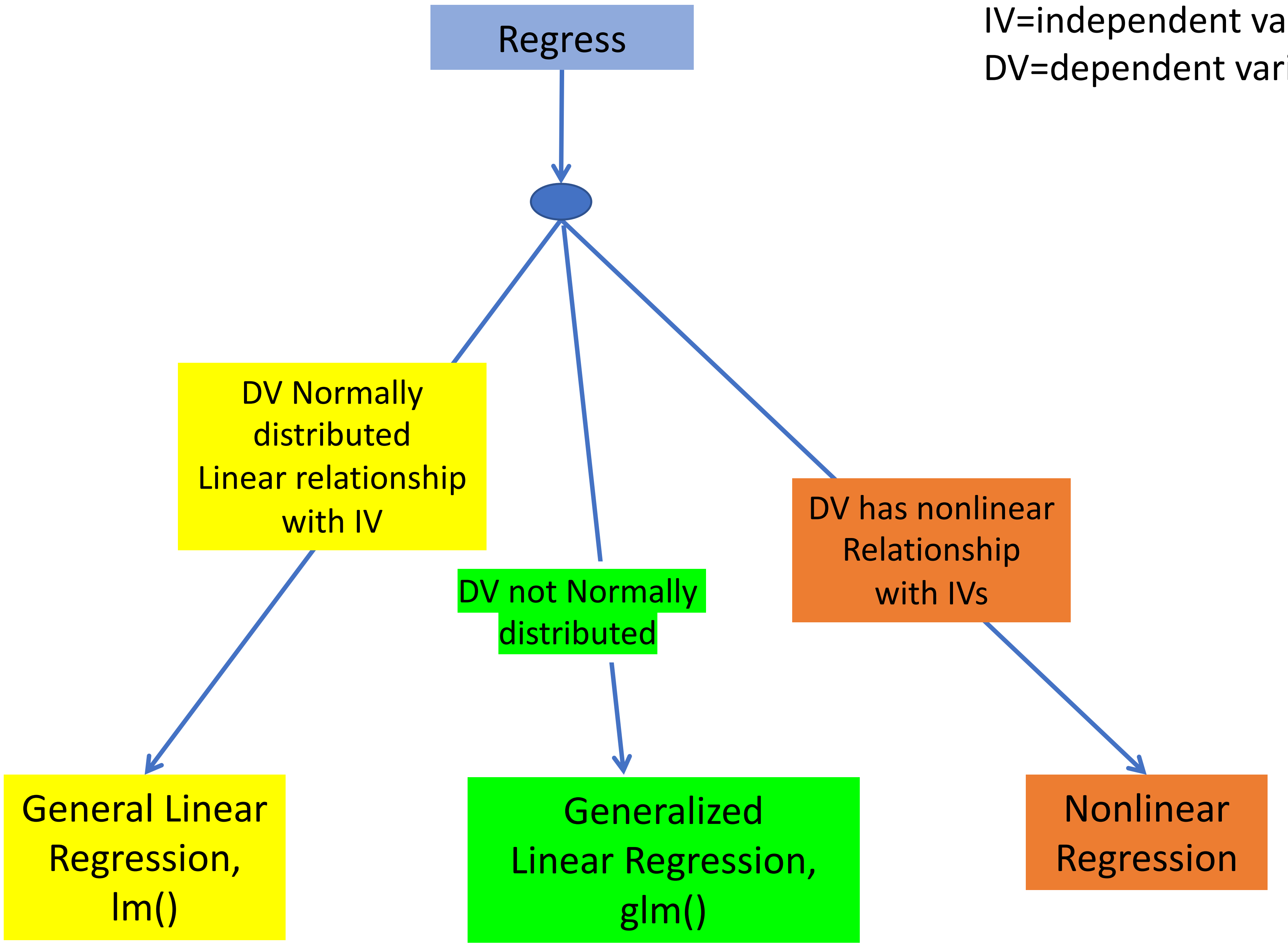
Binary outcome
- Binomial

Count outcome
- Poisson
- Negative binomial

Continuous positive outcome
- Gamma
- Inverse Gaussian

Common link functions: identity, logit, log, square-root, inverse, etc.

IV=independent variable
DV=dependent variable



Checking Assumptions

- Critically important to examine data and check assumptions underlying the regression model
  - Outliers
  - Normality
  - Constant variance
  - Independence among residuals

- Standard diagnostic plots include:
  - scatter plots of y versus x_i (outliers)
  - qq plot of residuals (normality)
  - residuals versus fitted values (independence, constant variance)
  - residuals versus x_i (outliers, constant variance)

Summary

- Regression offers a single cohesive approach to inference and estimating effect sizes

Response ~ Predictors

- Only reason to stick with t-tests/ANOVA are
  - Mostly just care about “statistical significance”
  - No other confounding covariates
  - Cultural (engrained in biomedical community)

Regression or ANOVA/t-tests?

- ANOVA/t-tests thinking emphasize “statistical significance” after experiment

- Regression thinking emphasizes overall weight of an independent variable predictively

- Regression is easy-peasy for “completely randomized” samples
  - lm() –for general linear model
  - glm() –for generalized linear model

In-Class Exercise 2 : glm()