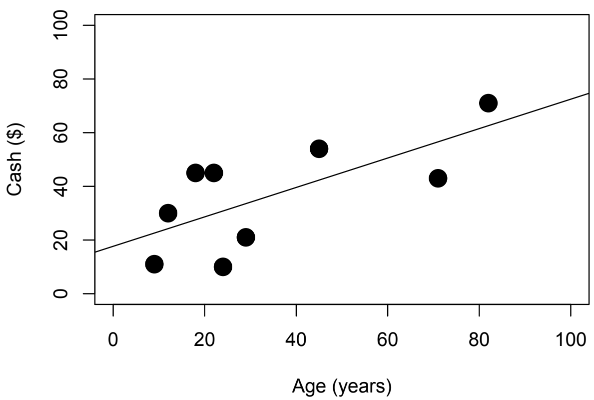
Stat 306:
Finding Relationships in Data.
Lecture 24
Review of Regression

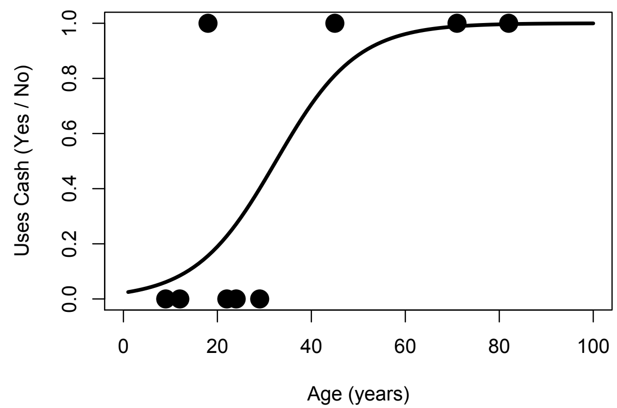
Stat 306:
Finding Relationships in Data.

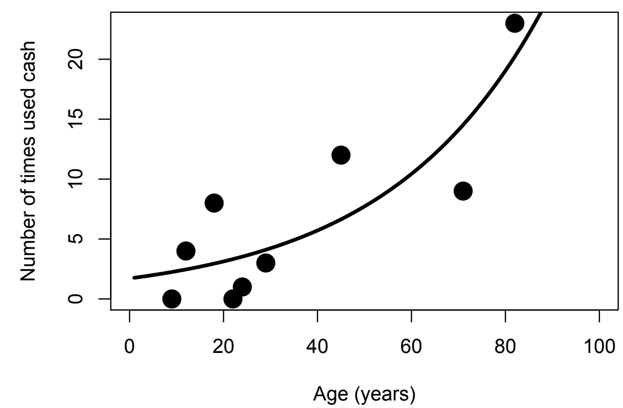
The main topic of this course is **regression**, which means fitting prediction equations.

Regression is a common statistical method in scientific research.

LINEAR REGRESSION	LOGISTIC REGRESSION	POISSON REGRESSION







LINEAR REGRESSION

$$Y_i \sim Normal(\mu(X_i), \sigma^2) ,$$

$$\text{where: } \mu(X_i) = X_i \beta$$

Minimize Least Squares

A one unit increase in x is associated with a β increase in Y.

Var(β) = $\sigma^2 (X^T X)^{-1}$

Using properties of Normal Distribution:

$$se(\hat{\mu}_Y(x)) \quad = \quad \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{(n-1)s_x^2}} .$$

```
> lm(y~x)
```

LOGISTIC REGRESSION

$$Y_i \sim Bernoulli(\pi(X_i)) ,$$

$$\text{where: } \pi(X_i) = \frac{exp(X_i \beta)}{1+exp(X_i \beta)}$$

Maximum Likelihood

A one unit increase in x is associated with a β increase in the log odds Y.

Var(β) = $[\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \pi_i (1 - \pi_i)]^{-1}$

Using delta method:

$$se(\hat{\pi}_Y(x)) = ...$$

```
> glm(y~x, family="binomial")
```

POISSON REGRESSION

$$Y_i \sim Poisson(\lambda(X_i)) ,$$

$$\text{where: } \lambda(X_i) = exp(X_i \beta)$$

Maximum Likelihood

A one unit increase in x is associated with increasing Y by a factor of a e^{β}.

Var(β) = $[\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \exp(\mathbf{x}_i^T \beta)]^{-1}$

Using delta method:

$$se(\hat{\lambda}_Y(x)) = ...$$

```
> glm(y~x, family="poisson")
```

Simple Linear Regression

Three important things to know about a normal random variable

Thing 1:

Linear combinations of independent normal random variables also have normal distributions! Remember…

$$\operatorname{Var}(aX \pm bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) \pm 2ab \operatorname{Cov}(X, Y)$$

Thing 2:

A normal random variable can be converted to a standard normal random variable.

Thing 3:

If the variance is unknown, we must use the t distribution.

The Sum of Squared Residuals:

The goal is to minimize $S(b_0, b_1) = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$.

$$
\begin{aligned}
(2.18) \qquad \hat{b}_0 &= \overline{y} - \hat{b}_1 \overline{x}, \\
(2.19) \qquad 0 &= \sum_{i=1}^{n} x_i y_i - [\overline{y} - \hat{b}_1 \overline{x}] n \overline{x} - \hat{b}_1 \sum_{i=1}^{n} x_i^2, \\
(2.20) \qquad 0 &= \sum_{i=1}^{n} x_i y_i - n \overline{x}\overline{y} + n \hat{b}_1 \overline{x}^2 - \hat{b}_1 \sum_{i=1}^{n} x_i^2, \\
(2.21) \qquad \hat{b}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n \overline{x}\overline{y}}{\sum_{i=1}^{n} x_i^2 - n \overline{x}^2} \\
(2.22) \qquad &= \frac{(n-1) s_{xy}}{(n-1) s_x^2} \\
(2.23) \qquad &= \frac{r_{xy} s_x s_y}{s_x^2} = \frac{r_{xy} s_y}{s_x}.
\end{aligned}
$$

The solution is therefore:

$$\hat{b}_0 = \overline{y} - \hat{b}_1 \overline{x}$$

$$\hat{b}_1 = r_{xy} s_y / s_x$$

Step 0: From θ, define estimator, $\hat{\theta}$

Step 1: Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Step 2: Determine $E[\hat{\Theta}]$ (to confirm it's unbiased) $Var[\hat{\Theta}]$ (to calculate se)

Step 3: Define se($\hat{\theta}$) = estimate of $\sqrt{\text{Var}(\hat{\Theta})}$

Step 4: Define (1-α)% C.I. = $\hat{\theta} \pm c \times se(\hat{\theta})$

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**	**Expected Value of the estimator**	**Variance of the estimator**	**Standard Error of estimator**	**Confidence Interval**
β_0	b_0	B_0	$E[B_0]$	$Var[B_0]$	$se(b_0)$	C.I. for β_0
β_1	b_1	B_1	$E[B_1]$	$Var[B_1]$	$se(b_1)$	C.I. for β_1
σ^2	s^2	S^2	$E[S^2]$	$Var[S^2]$	$se(s^2)$	C.I. for σ^2
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$(\hat{\mu}_Y(x))$	$E(\hat{\mu}_Y(x))$	$Var(\hat{\mu}_Y(x))$	$se(\hat{\mu}_Y(x))$	C.I. for $\mu_Y(x)$

2.5.2 Derivations

Steps to get 95% C.I. for b_1

1. Consider the sample statistic
   b_1 as the random variable B_1

2. Determine $Var[B_1]$

3. Define $se(b_1)$ as an estimate of $sqrt(Var(B_1))$

4. 95% C.I. = $[\ b_1 - c^*se(b_1)\ ,\ b_1 + c^*se(b_1)\ ]$

2.5.2 Derivations

The standard errors come from the variances when the estimators are considered as random variables. $\hat{\beta}_1$ as a random variable $\hat{B}_1$ is:

(2.50)
$$\hat{B}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})Y_i}{(n-1)s_x^2} = \sum_{i=1}^{n} a_i Y_i,$$

where

(2.51)
$$a_i = (x_i - \overline{x})/[(n-1)s_x^2].$$

2.5.2 Derivations

$$b_1 = r_{xy}\frac{s_y}{s_x}$$

$$= \frac{\sum_{i=1}^{n}(y_i)(x_i - \bar{x})}{(n-1)s_x^2}$$

$$= \sum_{i=1}^{n}\frac{(x_i - \bar{x})}{(n-1)s_x^2}(y_i)$$

Step 1. Consider the sample statistic b_1 as the random variable B_1:

$$B_1 = \sum_{i=1}^{n}\frac{(x_i - \bar{x})}{(n-1)s_x^2}(Y_i)$$

$$= \sum_{i=1}^{n} a_i Y_i$$, where: $$a_i = \frac{(x_i - \bar{x})}{(n-1)s_x^2}$$

2.5.2 Derivations

Step 1. Consider the sample statistic b_1 as the random variable B_1 :

$$B_1 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{(n-1)s_x^2}(Y_i)$$

$$= \sum_{i=1}^{n} a_i Y_i$$

, where: $$a_i = \frac{(x_i - \bar{x})}{(n-1)s_x^2}$$

Step 2. Determine Var[B_1]

First, recall that for random variable Y_i, we have:

(2.33) $$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

(2.58) $$\text{Var}\,(\hat{B}_1) = \sum_{i=1}^{n} a_i^2 \text{Var}\,(Y_i) = \sigma^2 \sum_{i=1}^{n} a_i^2 = \sigma^2 \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{[(n-1)s_x^2]^2}$$

(2.59) $$= \frac{\sigma^2}{(n-1)s_x^2}.$$

2.5.2 Derivations

Steps to get 95% C.I. for b_1

1. Consider the sample statistic
   b_1 as the random variable B_1

2. **Determine $Var[B_1]$** $= \frac{\sigma^2}{(n-1)s_x^2}.$

3. Define $se(b_1)$ as an estimate of $sqrt(Var(B_1))$

4. 95% C.I. = $[\ b_1 - c*se(b_1)\ ,\ b_1 + c*se(b_1)\ ]$

2.5.2 Derivations

Steps to get 95% C.I. for b_1

1. Consider the sample statistic
   b_1 as the random variable B_1

2. Determine $Var[B_1]$ $= \frac{\sigma^2}{(n-1)s_x^2}$.

3. Define se(b_1) as an estimate of sqrt(Var(B_1))

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n-1}\, s_x}$$

where: $\hat{\sigma} = \text{residual SD} = \left\{(n-2)^{-1}\sum_{i=1}^{n} e_i^2\right\}^{1/2}$

2.5.2 Derivations

<u>Steps to get 95% C.I. for b_1</u>

1. Consider the sample statistic
   b_1 as the random variable B_1

2. Determine Var[B_1] $= \frac{\sigma^2}{(n-1)s_x^2}.$

3. Define se(b_1) as an estimate of sqrt(Var(B_1)) : $se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n-1}\, s_x}$

4. **95% C.I. = [b_1 - c*se(b_1) , b_1 + c*se(b_1)]**

2.5.2 Derivations

Steps to get 95% C.I. for b_1

1. Consider the sample statistic
   b_1 as the random variable B_1

2. Determine Var[B_1] $= \frac{\sigma^2}{(n-1)s_x^2}$.

3. Define se(b_1) as an estimate of sqrt(Var(B_1)) : $se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n-1}\, s_x}$

4. **95% C.I. = [b_1 - c*se(b_1) , b_1 + c*se(b_1)]**

we take $c = t_{n-2,0.975}$

2.5.2 Derivations

Steps to get 95% C.I. for b_1

1. Consider the sample statistic
   b_1 as the random variable B_1
2. Determine $Var[B_1]$ $= \frac{\sigma^2}{(n-1)s_x^2}.$
3. Define $se(b_1)$ as an estimate of $sqrt(Var(B_1))$: $se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n-1}\, s_x}$
4. **95% C.I. = [b_1 - $c*se(b_1)$, b_1 + $c*se(b_1)$]**

we take $c = t_{n-2,0.975}$

Then we have :

95% C.I. for β_1 : $[b_1 - t_{n-2,0.975}\frac{\hat{\sigma}}{\sqrt{n-1}s_x}, \quad b_1 + t_{n-2,0.975}\frac{\hat{\sigma}}{\sqrt{n-1}s_x}]$

2.5.2 Derivations

Steps to get 95% C.I. for b_1

1. Consider the sample statistic b_1 as the random variable B_1
2. Determine Var[B_1] $= \dfrac{\sigma^2}{(n-1)s_x^2}.$
3. Define se(b_1) as an estimate of sqrt(Var(B_1)) : $se(\hat{\beta}_1) = \dfrac{\hat{\sigma}}{\sqrt{n-1}\, s_x}$
4. **95% C.I. = [b_1 - c*se(b_1) , b_1 + c*se(b_1)]**

we take c = $t_{n-2,0.975}$

Then we have :

95% C.I. for β_1 : $[b_1 - t_{n-2,0.975}\frac{\hat{\sigma}}{\sqrt{n-1}s_x}, \quad b_1 + t_{n-2,0.975}\frac{\hat{\sigma}}{\sqrt{n-1}s_x}]$

where: $\hat{\sigma} = \text{residual SD} = \left\{(n-2)^{-1}\sum_{i=1}^{n} e_i^2\right\}^{1/2}$ (also known as "s")

Step 0: From θ, define estimator, $\hat{\theta}$

Step 1: Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Step 2: Determine $E[\hat{\Theta}]$ (to confirm it's unbiased) $Var[\hat{\Theta}]$ (to calculate se)

Step 3: Define se($\hat{\theta}$) = estimate of $\sqrt{\text{Var}(\hat{\Theta})}$

Step 4: Define (1-α)% C.I. = $\hat{\theta} \pm c \times se(\hat{\theta})$

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**	**Expected Value of the estimator**	**Variance of the estimator**	**Standard Error of estimator**	**Confidence Interval**
β_0	b_0	B_0	$E[B_0]$	$Var[B_0]$	$se(b_0)$	C.I. for β_0
β_1	b_1	B_1	$E[B_1]$	$Var[B_1]$	$se(b_1)$	C.I. for β_1
σ^2	s^2	S^2	$E[S^2]$	$Var[S^2]$	$se(s^2)$	C.I. for σ^2
$\mu_Y(x)$	$(\hat{\mu}_Y(x))$	$(\hat{\mu}_Y(x))$	$E(\hat{\mu}_Y(x))$	$Var(\hat{\mu}_Y(x))$	$se(\hat{\mu}_Y(x))$	C.I. for $\mu_Y(x)$

se(subpopulation mean) VS. se(prediction error)

Subpopulation mean:

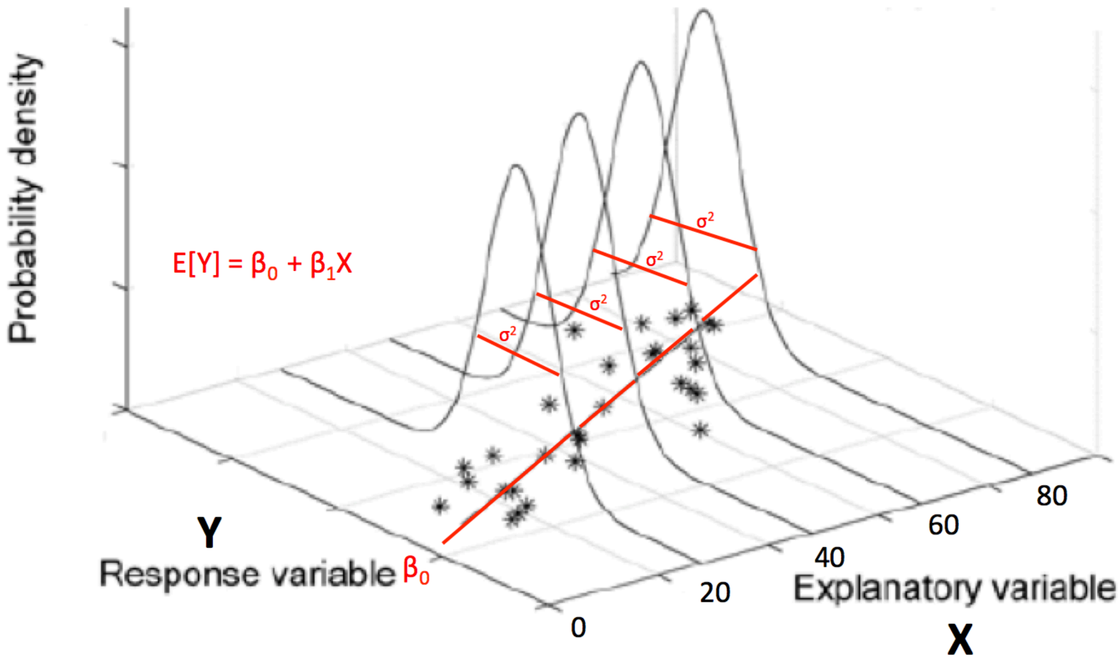
$$se(\hat{\mu}_Y(x)) \quad = \quad \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{(n-1)s_x^2}}$$

Whereas, the (estimated) SE of the prediction error is:

(2.69) $$\hat{\sigma} \times \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{(n-1)s_x^2}},$$

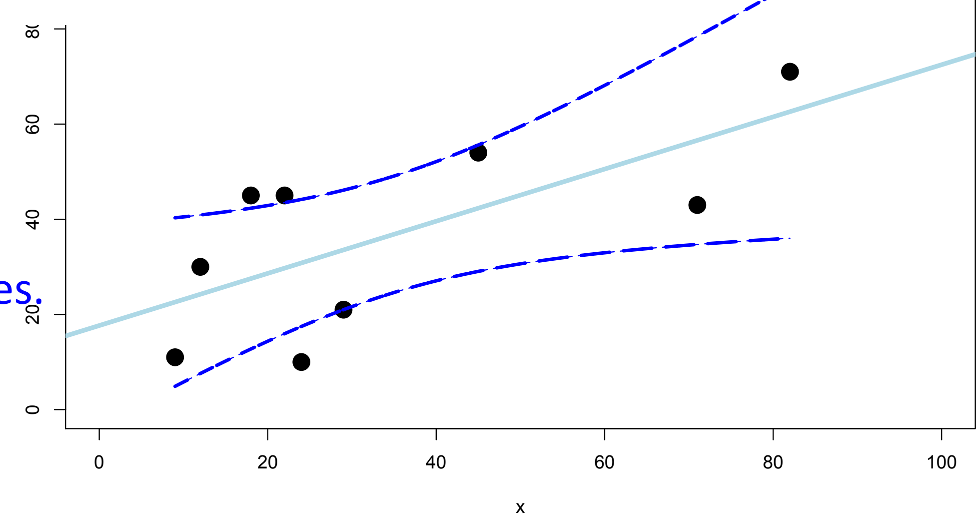
and this does not decrease to 0 as $n \to \infty$.

- **Confused about homogeneity vs. non-consistent width of confidence intervals?**



σ^2 is the variance of Y; constant regardless of the value of x.

The blue dashed line is the confidence interval for the subpopulation mean.
In other words, it represents the variability in our estimate of the mean of Y as x changes.



Multiple Linear Regression

design matrix or data matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \hat{\mathbf{b}} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_p \end{pmatrix}. \tag{3.18}$$

Least Squares for multiple Regression:

$SS(b) = (y - Xb)^T(y - Xb)$

$$\frac{\delta SS(\mathbf{b})}{\delta \mathbf{b}} = 0$$

$$\Rightarrow \quad (\mathbf{X}^T\mathbf{X})\hat{\mathbf{b}} = \mathbf{X}^T\mathbf{y}$$

$$\text{or} \quad \hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

The *system of normal equations*

3.6 Interval estimates and standard errors

Step 0: From θ, define estimator, $\hat{\theta}$

Step 1: Consider the sample statistic, $\hat{\theta}$, as a random variable $\hat{\Theta}$

Step 2: Determine E[$\hat{\Theta}$] (to confirm it's unbiased) Var[$\hat{\Theta}$] (to calculate se)

Step 3: Define se($\hat{\theta}$) = estimate of $\sqrt{\text{Var}(\hat{\Theta})}$

Step 4: Define (1-α)% C.I. = $\hat{\theta} \pm c \times se(\hat{\theta})$

Population parameter or "something we would like to estimate"	**Sample statistic ("estimator")**	**Estimator as a Random Variable**	**Expected Value of the estimator**	**Variance of the estimator**	**Standard Error of estimator**	**Confidence Interval**
β	**b** 1. $= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	**B ~** 2. **N(β,** $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$**)**	E[**b**] = **β** 3.	Var[**B**] 4. $= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$	se(**b**) 5. $= \hat{\sigma}\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}$	C.I. for **β** 6.
σ^2	**s^2 or MS(Res)** 1.	S^2 2.	$E[S^2]$ 3.	$Var[S^2]$	$se(s^2)$	C.I. for σ^2
$\mu_Y(\mathbf{x})$	$(\hat{\mu}_Y(x))$ 1.	$(\hat{\mu}_Y(x))$ 2.	E $(\hat{\mu}_Y(x))$ 3.	Var$(\hat{\mu}_Y(x))$ 4.	$se(\hat{\mu}_Y(x))$ 5.	C.I. for $\mu_Y(x)$ 6.

3.6 Interval estimates and standard errors

From (3.11), with $\hat{\mathbf{B}} = \hat{\boldsymbol{\beta}}$ as a random vector, and $k = p + 1$ as the dimension of $\hat{\boldsymbol{\beta}}$,

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{A}\mathbf{Y}, \tag{3.66}$$

$$\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \begin{pmatrix} \mathbf{a}_0^T \\ \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_p^T \end{pmatrix}, \tag{3.67}$$

$$(k \times n) \qquad (k \times k) \quad (k \times n) \tag{3.68}$$

- **Thing 1:**
  - Linear combinations of independent normal random variables also have normal distributions! (see Appendix B)

3.6 Interval estimates and standard errors

where $\mathbf{a}_j^T$ is a $1 \times n$ row vector. The covariance matrix of $\mathbf{Y}$ is $\boldsymbol{\Sigma}_\mathbf{Y} = \sigma^2 \mathbf{I}_n$ ($n \times n$ identity matrix because the ϵ_i are independent and identically distributed $N(0, \sigma^2)$ random variables). From the Appendix A for linear combinations,

$$\text{Var}(\hat{B}_1) = \text{Var}(\mathbf{a}_1^T \mathbf{Y}) = \mathbf{a}_1^T \boldsymbol{\Sigma}_\mathbf{Y} \mathbf{a}_1 = \mathbf{a}_1^T (\sigma^2 \mathbf{I}_n) \mathbf{a}_1 = \sigma^2 \mathbf{a}_1^T \mathbf{a}_1 \tag{3.69}$$

$$\text{Var}(\hat{B}_2) = \text{Var}(\mathbf{a}_2^T \mathbf{Y}) = \mathbf{a}_2^T \boldsymbol{\Sigma}_\mathbf{Y} \mathbf{a}_2 = \sigma^2 \mathbf{a}_2^T \mathbf{a}_2 \tag{3.70}$$

$$\vdots = \vdots$$

$$\text{Var}(\hat{B}_p) = \text{Var}(\mathbf{a}_p^T \mathbf{Y}) = \mathbf{a}_p^T \boldsymbol{\Sigma}_\mathbf{Y} \mathbf{a}_p = \sigma^2 \mathbf{a}_p^T \mathbf{a}_p \tag{3.71}$$

$$\text{Cov}(\hat{B}_1, \hat{B}_2) = \text{Cov}(\mathbf{a}_1^T \mathbf{Y}, \mathbf{a}_2^T \mathbf{Y}) = \mathbf{a}_1^T \boldsymbol{\Sigma}_\mathbf{Y} \mathbf{a}_2 = \sigma^2 \mathbf{a}_1^T \mathbf{a}_2 \tag{3.72}$$

$$\vdots = \vdots$$

$$\text{Var}(\ B\ ) = \text{Var}(AY)$$
$$\text{Var}(B) = A\,\text{Var}(Y)\,A^T$$

https://en.wikipedia.org/wiki/Covariance_matrix#Generalization_of_the_variance

3.6 Interval estimates and standard errors

From (3.11), with $\hat{\mathbf{B}} = \hat{\boldsymbol{\beta}}$ as a random vector, and $k = p + 1$ as the dimension of $\hat{\boldsymbol{\beta}}$,

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{A}\mathbf{Y}, \tag{3.66}$$

$$\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \begin{pmatrix} \mathbf{a}_0^T \\ \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_p^T \end{pmatrix}, \tag{3.67}$$

$$(k \times n) \qquad (k \times k) \quad (k \times n) \tag{3.68}$$

$$\mathbf{Y} \sim Normal(\mu, \sigma^2 I_n)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ .. \\ y_n \end{bmatrix} \sim Normal \begin{bmatrix} \mu_1 & , & \sigma^2 & 0 \ldots & 0 \\ \mu_2 & , & 0 & \sigma^2 \ldots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ \mu_n & & 0 & \ldots 0 & \sigma^2 \end{bmatrix}$$

3.6 Interval estimates and standard errors

From (3.11), with $\hat{\mathbf{B}} = \hat{\boldsymbol{\beta}}$ as a random vector, and $k = p + 1$ as the dimension of $\hat{\boldsymbol{\beta}}$,

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{A}\mathbf{Y}, \tag{3.66}$$

$$\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \begin{pmatrix} \mathbf{a}_0^T \\ \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_p^T \end{pmatrix}, \tag{3.67}$$

$$(k \times n) \qquad (k \times k) \quad (k \times n) \tag{3.68}$$

Var(B) = Var(AY)

Var (B) = A Var(Y) A^T

$$\mathbf{Y} \sim Normal(\mu, \sigma^2 I_n)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ .. \\ y_n \end{bmatrix} \sim Normal \begin{bmatrix} \mu_1 & \sigma^2 & 0 \dots & 0 \\ \mu_2 \quad , & 0 & \sigma^2 \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mu_n & 0 & \dots 0 & \sigma^2 \end{bmatrix}$$

3.6 Interval estimates and standard errors

From (3.11), with $\hat{\mathbf{B}} = \hat{\boldsymbol{\beta}}$ as a random vector, and $k = p + 1$ as the dimension of $\hat{\boldsymbol{\beta}}$,

(3.66) $$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{A}\mathbf{Y},$$

(3.67) $$\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \begin{pmatrix} \mathbf{a}_0^T \\ \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_p^T \end{pmatrix},$$

(3.68) $$(k \times n) \qquad (k \times k) \quad (k \times n)$$

Variance – Covariance Matrix of **Y**

$$\mathbf{Y} \sim Normal(\mu, \sigma^2 I_n)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ .. \\ y_n \end{bmatrix} \sim Normal \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}, \begin{pmatrix} \sigma^2 & 0 \ldots & 0 \\ 0 & \sigma^2 \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots 0 & \sigma^2 \end{pmatrix}$$

Var(B) = Var(AY)

Var (B) = A Var(Y) A^T

3.6 Interval estimates and standard errors

Putting everything together, one gets:

$$
\begin{pmatrix}
\mathrm{Var}(\hat{B}_0) & \mathrm{Cov}(\hat{B}_0,\hat{B}_1) & \cdots & \mathrm{Cov}(\hat{B}_0,\hat{B}_p) \\
\mathrm{Cov}(\hat{B}_1,\hat{B}_0) & \mathrm{Var}(\hat{B}_1) & \cdots & \mathrm{Cov}(\hat{B}_1,\hat{B}_p) \\
\vdots & \vdots & \ddots & \vdots \\
\mathrm{Cov}(\hat{B}_p,\hat{B}_0) & \mathrm{Cov}(\hat{B}_p,\hat{B}_1) & \cdots & \mathrm{Var}(\hat{B}_p)
\end{pmatrix}
= \sigma^2
\begin{pmatrix}
\mathbf{a}_0^T\mathbf{a}_0 & \mathbf{a}_0^T\mathbf{a}_1 & \cdots & \mathbf{a}_0^T\mathbf{a}_p \\
\mathbf{a}_1^T\mathbf{a}_0 & \mathbf{a}_1^T\mathbf{a}_1 & \cdots & \mathbf{a}_1^T\mathbf{a}_p \\
\vdots & \vdots & \ddots & \vdots \\
\mathbf{a}_p^T\mathbf{a}_0 & \mathbf{a}_p^T\mathbf{a}_1 & \cdots & \mathbf{a}_p^T\mathbf{a}_p
\end{pmatrix}
\tag{3.73}
$$

$$
= \sigma^2
\begin{pmatrix}
\mathbf{a}_0^T \\ \mathbf{a}_1^T \\ \cdots \\ \mathbf{a}_p^T
\end{pmatrix}
\begin{pmatrix}
\mathbf{a}_0 & \cdots & \mathbf{a}_p
\end{pmatrix}
= \sigma^2 \mathbf{A}\mathbf{A}^T
\tag{3.74}
$$

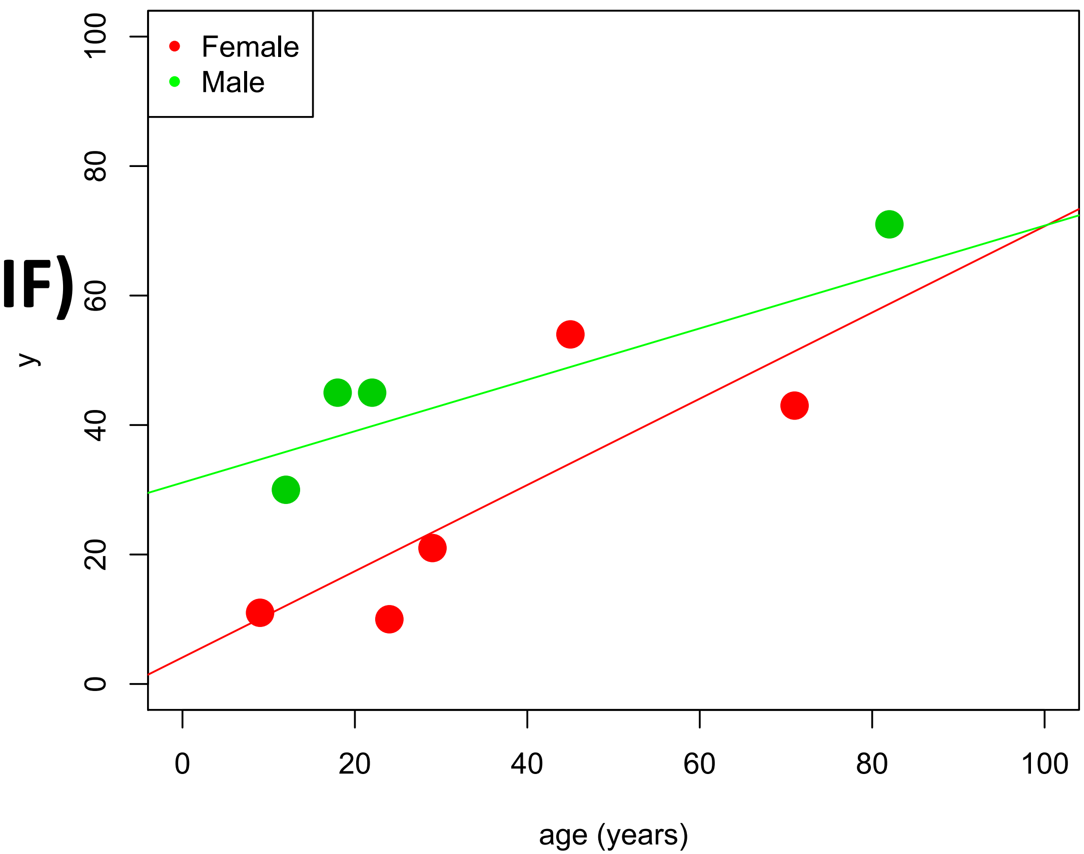
$$
= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}
\tag{3.75}
$$

$$
= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \stackrel{\text{def}}{=} \mathbf{\Sigma}_{\hat{\boldsymbol{\beta}}}.
\tag{3.76}
$$

$$
\mathrm{Var}(\beta) = \sigma^2(X^TX)^{-1}
$$

Multiple Linear Regression

- **Categorical covariates**
- **Interaction effects**
- **Multi-collinearity (VIF)**

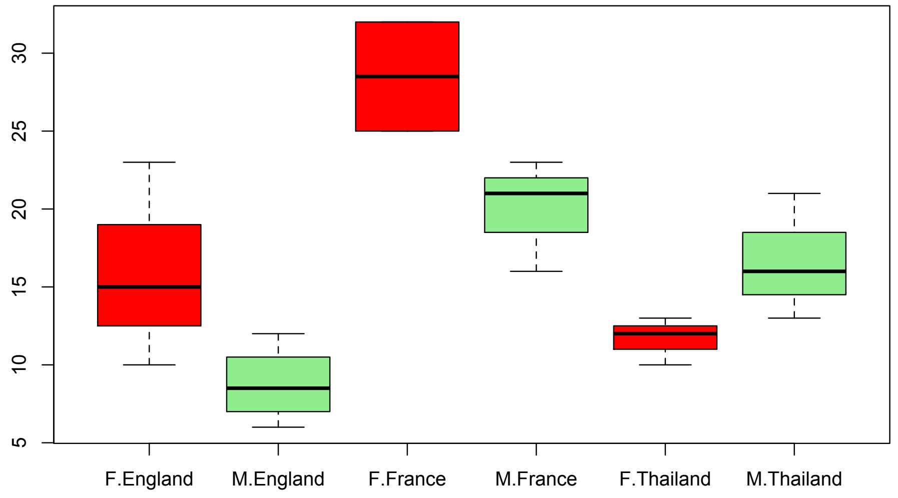


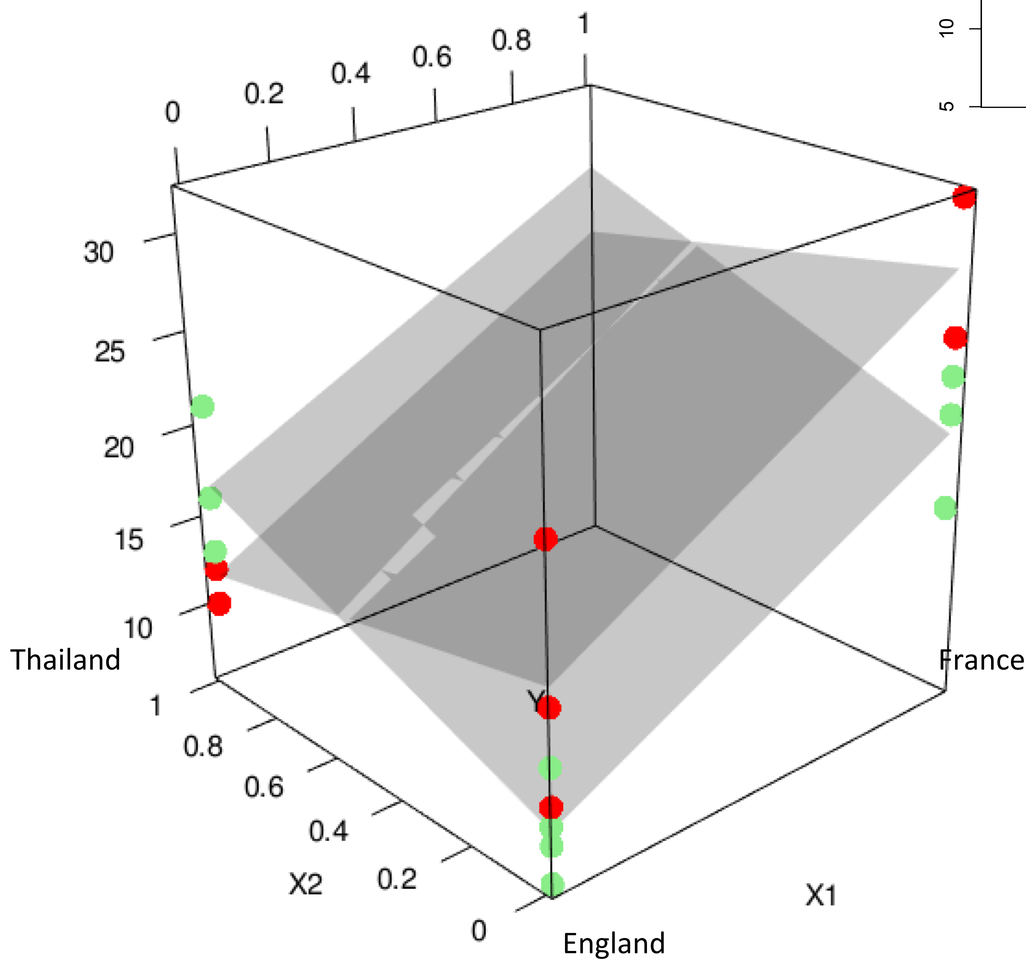
The **model**:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

The **model** with an interaction effect:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 (X_3 X_1) + \beta_5 (X_3 X_2)$$





What hypotheses can we test?

The art of linear regression

- Categorical predictors
- Quadratic (polynomial) relationships
- Outliers (Leverage, Influence)
- How to fix heterogeneity
- Regression to the mean
- Simpsons Paradox
- Unobserved Confounding

Four categories of scientific study

	Observational	Experimental
Goal is Explanation	1.	2.
Goal is Prediction	3.	4.

Goal is **Explanation**

1. What questions do you want to ask ?

2. Define an appropriate model.

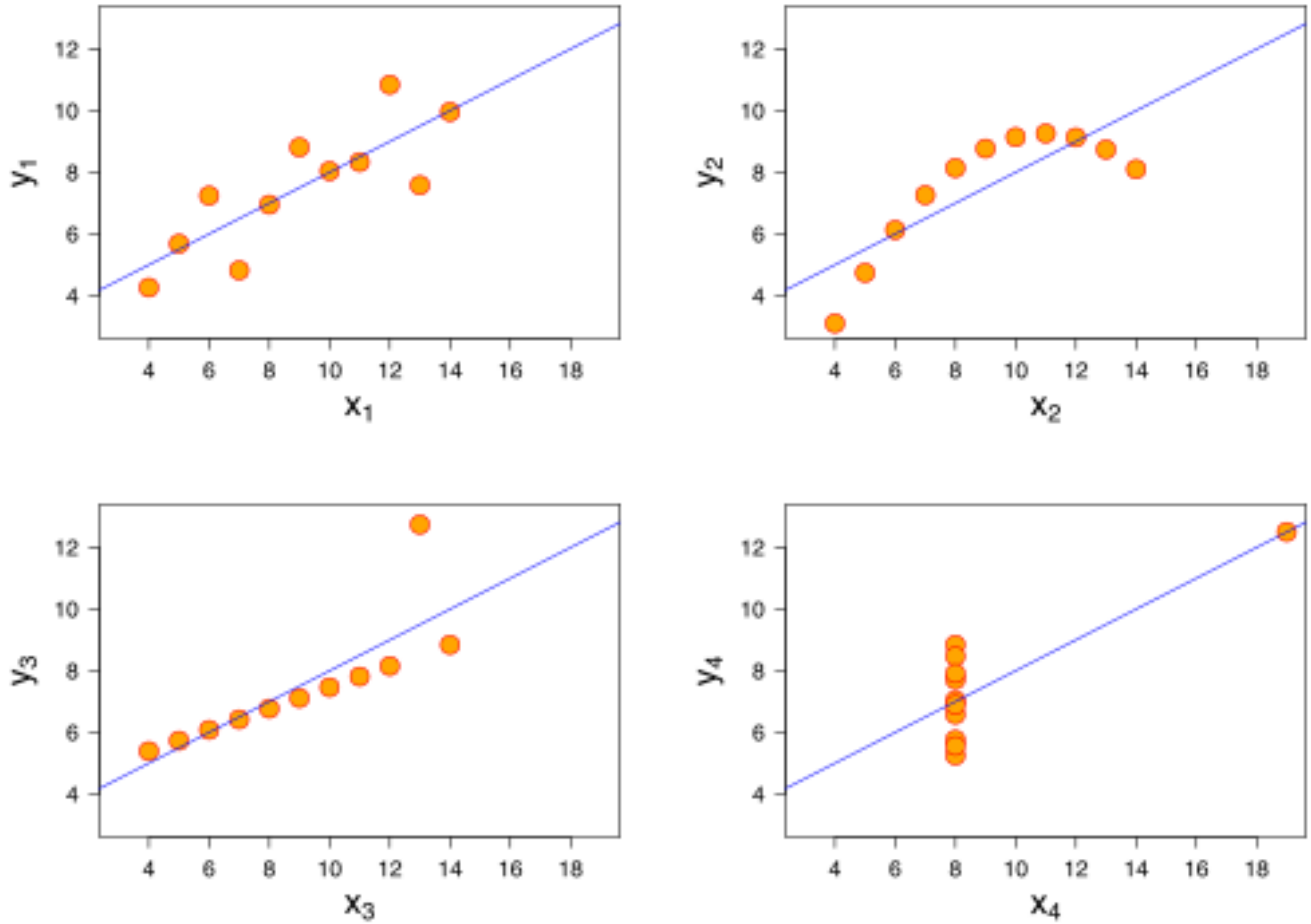
3. Define the hypotheses that correspond to the questions of interest.

4. Collect the data.

5. Fit the model as defined earlier.

6. Answer your questions with uncertainty quantification
   (i.e. with p-values, Confidence Intervals).

Classic example: Anscombe's quartet



Goal is
Prediction

1. What do you want to predict?

2. Define an appropriate metric for evaluating quality of predictions (e.g. RMSE, absolute prediction error, ROC curve).

3. Collect the data.

4. Separate your data into “train” and “holdout” subsets.

5. Fit many different models to the “train” subset of the data.

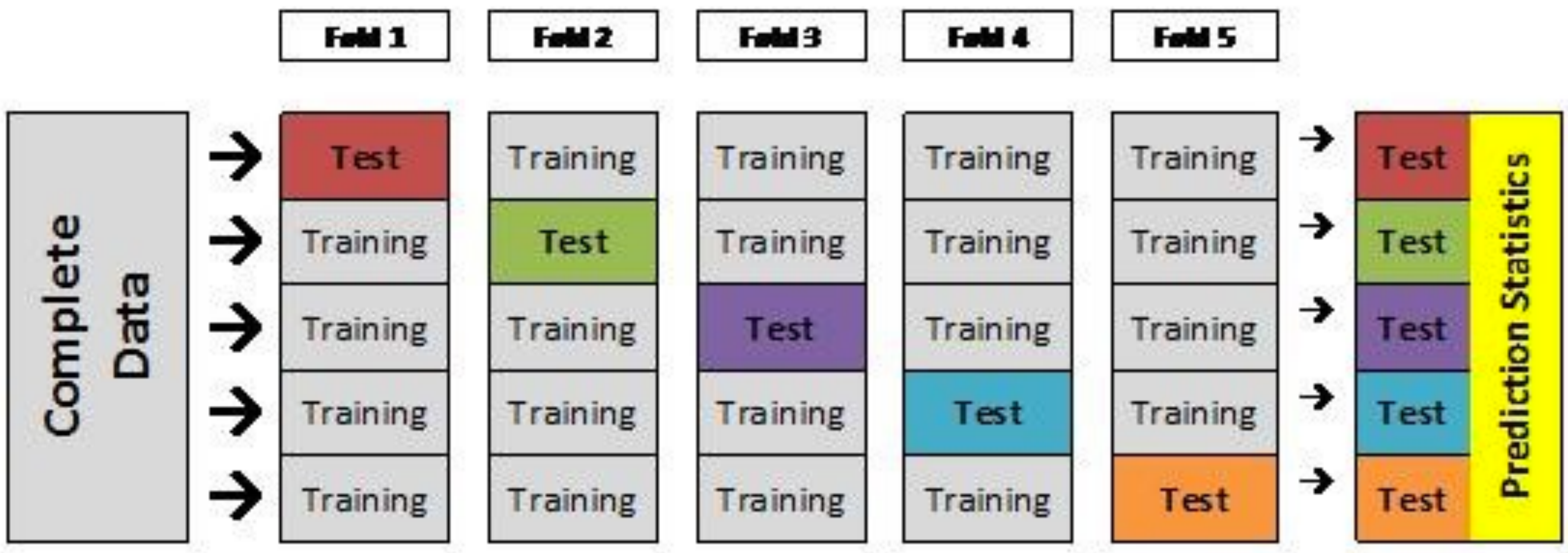
6. Pick the model that is “best” (according to your chosen outcome) for making predictions on the “holdout” subset of the data.

7. Note that p-values and Confidence intervals are not valid.

For each model, we do 5-fold CV:

Metric:

Mean Absolute Prediction Error:



12

8

6

9

5

K-averaged metric = 40/5 = 8

Source: http://blog.goldenhelix.com/goldenadmin/cross-validation-for-genomic-prediction-in-svs/

Logistic Regression

- Maybe there are better measure to describe the effect?
- Since OR is so difficult to interpret, perhaps we should use RR?

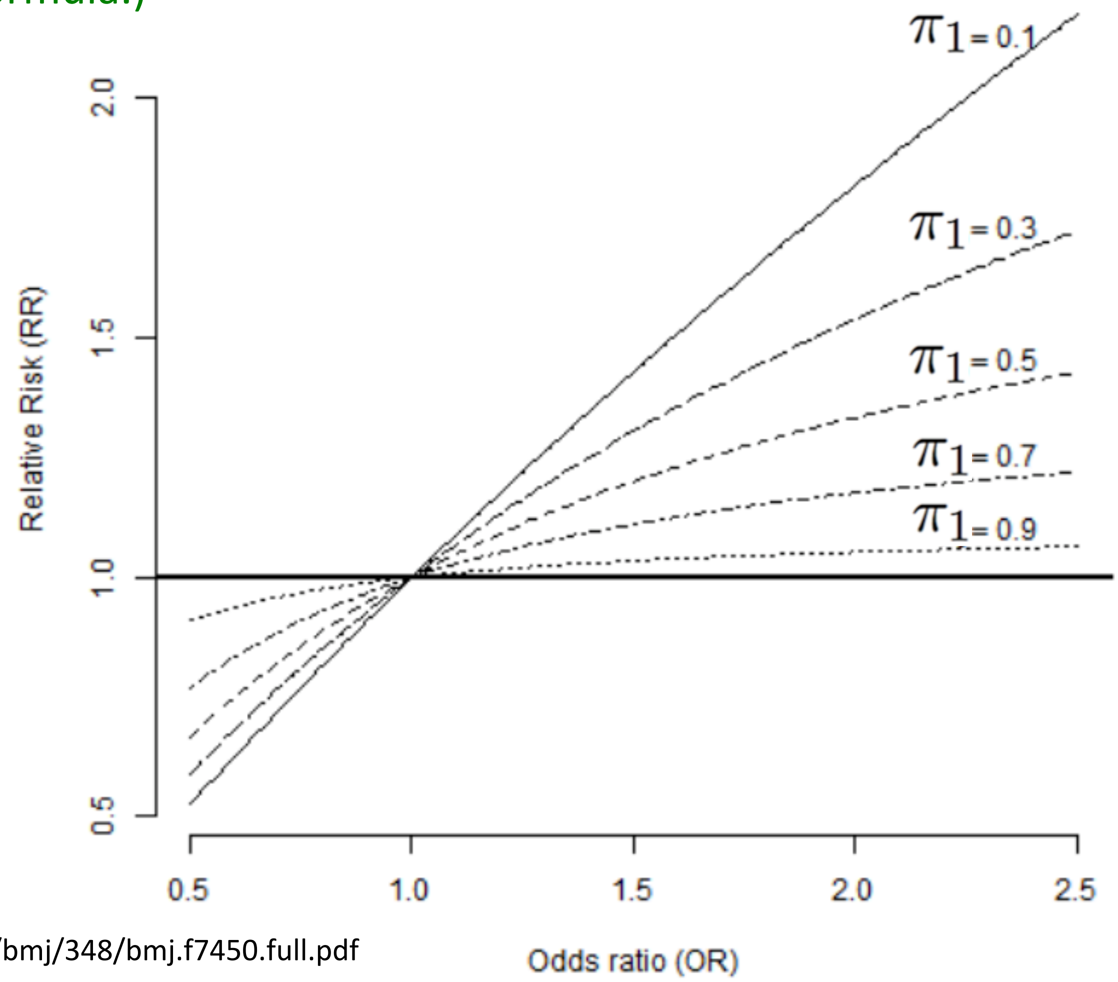
Type θ	*Expression*	*Domain*	*Null Value*
Risk difference (RD)	$\pi_1 - \pi_2$	$[-1, 1]$	0
Relative risk (RR)	π_1/π_2	$(0, \infty)$	1
log RR	$\log(\pi_1) - \log(\pi_2)$	$(-\infty, \infty)$	0
Odds ratio (OR)	$\dfrac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$	$(0, \infty)$	1
log OR	$\log \dfrac{\pi_1}{1-\pi_1} - \log \dfrac{\pi_2}{1-\pi_2}$	$(-\infty, \infty)$	0

Biostatistical Methods: The Assessment of Relative Risks
By John M. Lachin

To convert an Odds Ratio to a Relative Risk, you need to know π_1, which in our example is Pr(Y=1|X=0). Here is the formula:

$$RR = OR/(1 - \pi_1 + (\pi_1 \cdot OR))$$

(Exercise : Derive the formula.)



More info: http://www.bmj.com/content/bmj/348/bmj.f7450.full.pdf

Misclassification and the ROC curve

Note that: The misclassification rate among the true 0s is $n_{01}/[n_{00}+n_{01}]$ and this decreases as τ increases. The misclassification rate among the true 1s is $n_{10}/[n_{10}+n_{11}]$ and this increases as τ increases.

Sensitivity: True Positive rate (= n11/(n11+n10))

Specificity: True Negative rate (= n00/(n00 + n01))

	$\hat{\pi} \leq \tau$	$\hat{\pi} > \tau$	count	misclass rate
$y=0$	n_{00}	n_{01}	$n_{00}+n_{01}$	$n_{01}/[n_{00}+n_{01}]$
$y=1$	n_{10}	n_{11}	$n_{10}+n_{11}$	$n_{10}/[n_{10}+n_{11}]$
all	$n_{00}+n_{10}$	$n_{01}+n_{11}$	n	$(n_{01}+n_{10})/n$

True Negatives

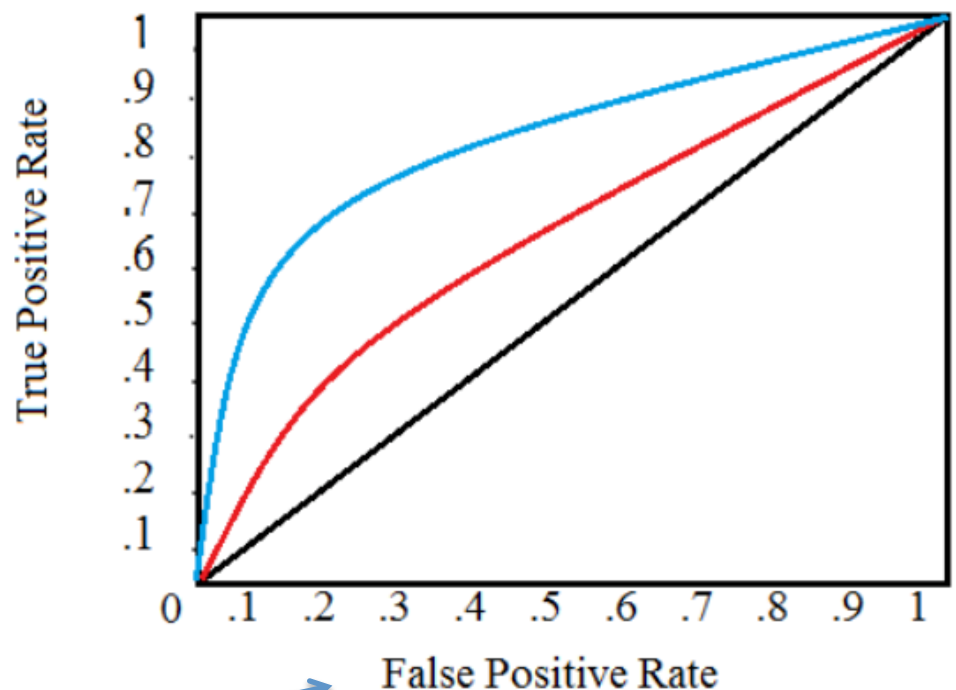
False Negatives

True Positives

False Positives

The Receiver Operating Characteristic curve (ROC curve)

The ROC curve is a plot that show how **Sensitivity** and **Specificity** change with different values for the threshold:



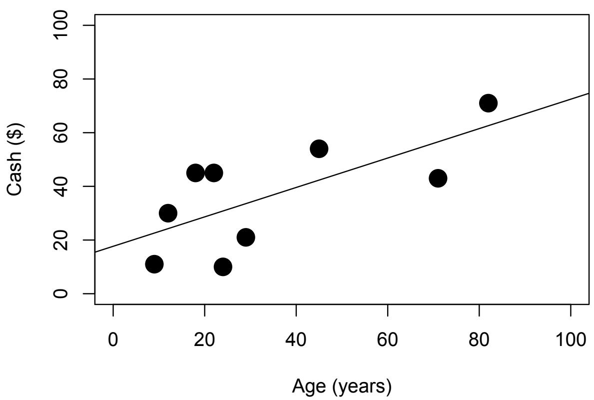
Sensitivity

1- Specificity

A ROC curve showing two tests. The red test is closer to the diagonal and is therefore less accurate than the green test.

http://www.statisticshowto.com/receiver-operating-characteristic-roc-curve/

LINEAR REGRESSION



$$Y_i \sim Normal(\mu(X_i), \sigma^2) \ ,$$

$$\text{where: } \mu(X_i) = X_i\beta$$

Minimize Least Squares

A one unit increase in x is associated with a β increase in Y.

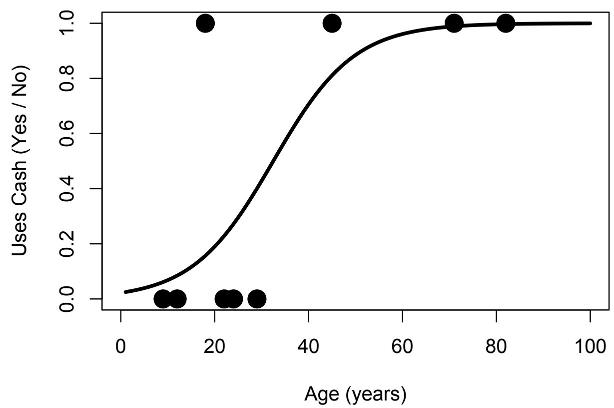
$Var(\beta) = \ \sigma^2(X^T X)^{-1}$

Using properties of Normal Distribution:

$$se(\hat{\mu}_Y(x)) \quad = \quad \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)s_x^2}} \ .$$

```
> lm(y~x)
```

LOGISTIC REGRESSION



$$Y_i \sim Bernoulli(\pi(X_i)) \ ,$$

$$\text{where: } \pi(X_i) = \frac{exp(X_i\beta)}{1+exp(X_i\beta)}$$

Maximum Likelihood

A one unit increase in x is associated with a β increase in the log odds Y.

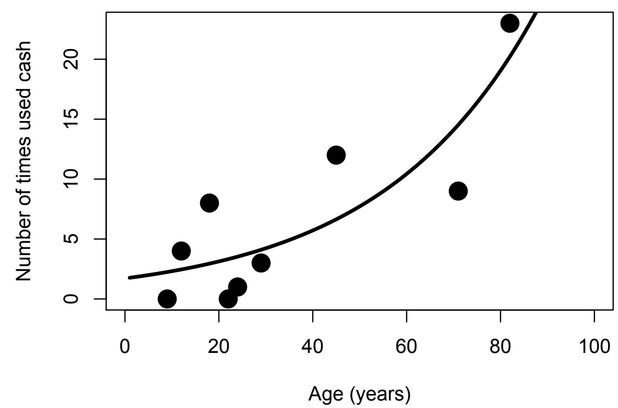
$Var(\beta) = [\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T \pi_i(1-\pi_i)]^{-1}$

Using delta method:

$$se(\hat{\pi}_Y(x)) = ...$$

```
> glm(y~x, family="binomial")
```

POISSON REGRESSION



$$Y_i \sim Poisson(\lambda(X_i)) \ ,$$

$$\text{where: } \lambda(X_i) = exp(X_i\beta)$$

Maximum Likelihood

A one unit increase in x is associated with increasing Y by a factor of a e^{β} .

$Var(\beta) = [\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T \exp(\mathbf{x}_i^T\beta)]^{-1}$

Using delta method:

$$se(\hat{\lambda}_Y(x)) = ...$$

```
> glm(y~x, family="poisson")
```

What's next for regression…

- Other distributions for Y
  - Survival times or Time-to-event data
  - Semicontinuous data
  - Mixture models

- Penalized Regression
  - Lasso
  - Ridge Regression

- Observations are not independent
  - Random effects models
  - Methods for clustered data
  - Time series models or longitudinal models
  - Spatial models

- Bayesian Methods
  - Incorporating Prior knowledge about the parameters
  - Updating your likelihood (posterior distribution) as you collect more data.

"An approximate answer to the right problem is
worth a good deal more than an exact answer
to an approximate problem." -- John Tukey

"All models are wrong but some are useful".
– George Box

“Absence of evidence is not evidence of absence.”

“The most important is to know what questions to ask of the data.”