UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Statistics Workshop: Part II

Linglong Kong

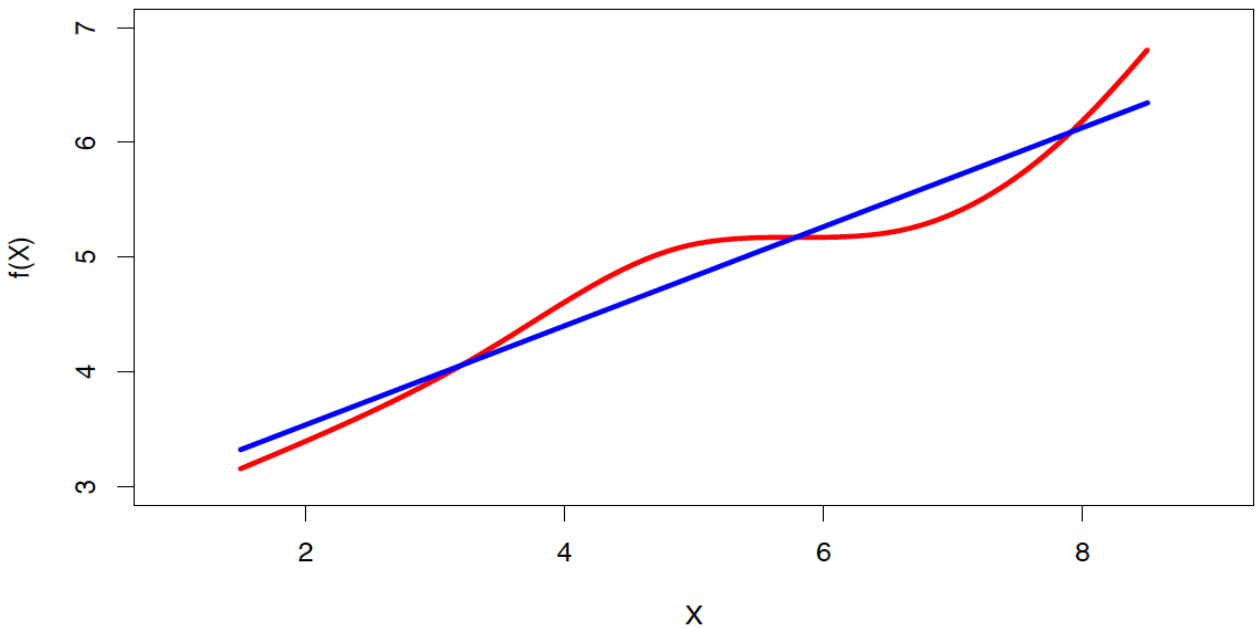
Department of Mathematical and Statistical Sciences
University of Alberta

May 12, 2016

Outline

Simple Linear Regression

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $X_1, X_2, \cdots, X_p$ is linear.
- True regression functions are never linear! although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.



An Introduction to Statistical Learning, by G. James, D. Witten, T. Hastie and R. Tibshirani. 2013, Springer.

Simple Linear Regression

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- Simple Linear Regression Model (SLR) has the form of

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where β_0 and β_1 are two unknown parameters (coefficients), called intercept and *slope*, respectively, and ε is the error term.

- Given the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, the estimated regression line is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x.$$

- For $X = x$, we predict Y by $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, where the hat symbol denotes an estimated value.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Estimate the parameters

- Let (y_i, x_i) be the i-th observation and $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, we call $e_i = y_i - \hat{y}_i$ the ith residual.
- To estimate the parameters, we minimized the residual sums of squares (RSS),

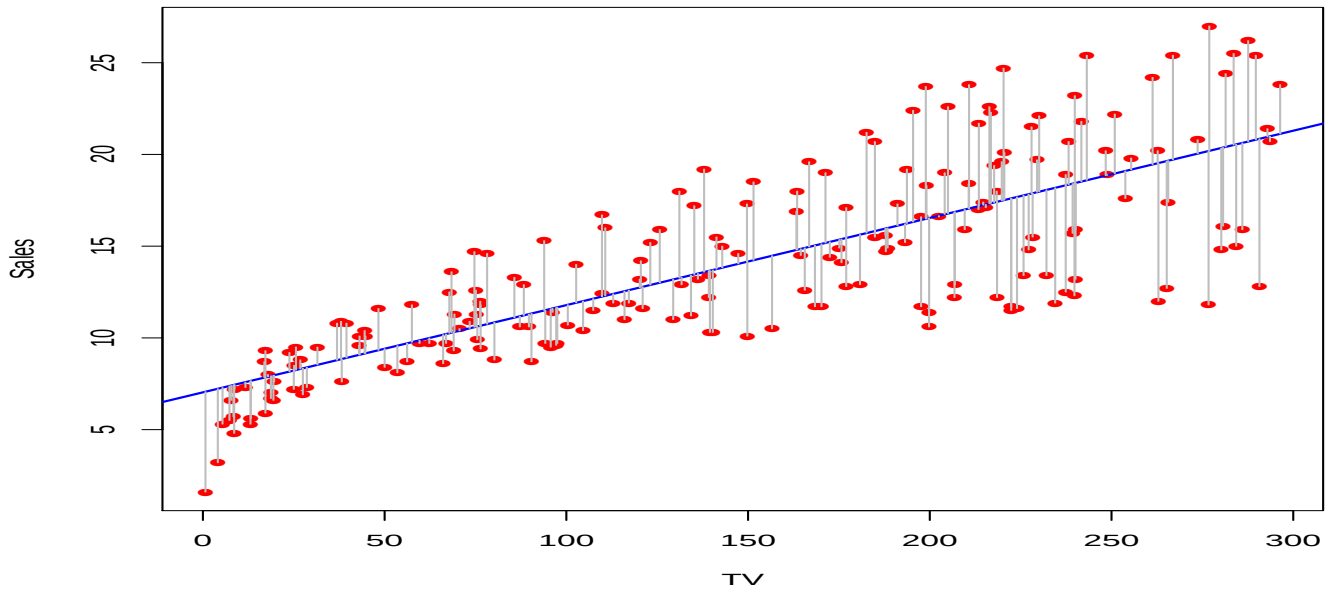
$$\text{RSS} = \sum_i e_i^2 = \sum_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \right)^2.$$

- Denote $\bar{y} = \sum_i y_i / n$ and $\bar{x} = \sum_i x_i / n$. The minimized values are

$$\hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \left(r \frac{\sqrt{\sum_i (y_i - \bar{y})^2}}{\sqrt{\sum_i (x_i - \bar{x})^2}} \right),$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Example



- Advertising data: the least square fit for the regression of `sales` and `TV`.
- Each grey line segment represents an error, and the fit makes a compromise by averaging their squares.
- In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

An Introduction to Statistical Learning, by G. James, D. Witten, T. Hastie and R. Tibshirani. 2013, Springer.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Assess the coefficient estimates

- The standard error of an estimator reflects how it varies under repeated sampling.

$$\text{SE}(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum(x_i - \bar{x})^2}}, \qquad \text{SE}(\hat{\beta}_0) = \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right)},$$

  where $\sigma^2 = \text{Var}(\varepsilon)$.

- A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.
- It has the form

$$\hat{\beta}_1 \pm 2 \cdot \text{SE}(\hat{\beta}_1).$$

- For the advertising data, the 95% confidence interval for β_1 is $[0.042, 0.053]$, which means, there is approximately 95% chance this interval contains the true value of β_1 (under a scenario where we got repeated samples like the present sample).

Hypothesis testing

- Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of

  H_0: There is no relationship between X and Y versus the alternative hypothesis

  H_A: There is some relationship between X and Y.
- Mathematically, we test

$$H_0 : \beta_1 = 0 \text{ versus } H_A : \beta_1 \neq 0,$$

since if $\beta_0 = 0$ then the model reduces to $Y = \beta_0 + \varepsilon$, and X is not associated with Y.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Hypothesis testing

- To test the null hypothesis, we compute a t-statistics,

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}\left(\hat{\beta}_1\right)}.$$

- This statistics follows t_{n-2} under the null hypothesis $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the p-value.
- Results for the advertising data

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Measure of fit

- We compute the Residual Standard Error

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_i (y_i - \hat{y}_i)^2},$$

  where the residual sum-of-squares is $\text{RSS} = \sum_i (y_i - \hat{y}_i)^2$.
- R-squared or fraction of variance explained is

$$\text{R}^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}},$$

  where $\text{TSS} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares.
- It can be shown that in this simple linear regression setting that $\text{R}^2 = r^2$, where r is the correlation between Y and X:

$$r = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_i (y_i - \bar{y})^2}\sqrt{\sum_i (x_i - \bar{x})^2}} = \left(\hat{\beta}_1 \frac{\sqrt{\sum_i (x_i - \bar{x})^2}}{\sqrt{\sum_i (y_i - \bar{y})^2}}\right).$$

R code

```
> TVadData = read.csv('... Advertising.csv')
> attach(TVadData)
> TVadlm = lm(Sales~TV)
> summary(TVadlm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- Multiple Linear Regression has more than one covariates,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

  where usually $\varepsilon \sim N(0, \sigma^2)$.
- We interpret β_j as the average effect on Y of a one unit increase in X_j, while holding all the other covariates fixed.
- In the advertising example, the model becomes

$$\text{Sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{Radio} + \beta_3 \times \text{Newspaper} + \varepsilon.$$

Coefficient Interpretation

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- The ideal scenario is when the predictors are uncorrelated — a balanced design.
  - Each coefficient can be estimated and tested separately.
  - Interpretations such as a unit change in X_j is associated with a β_j change in Y , while all the other variables stay fixed, are possible.
- Correlations amongst predictors cause problems.
  - The variance of all coefficient tends to increase, sometimes dramatically.
  - Interpretations become hazardous — when X_j changes, everything else changes.
- Claims of causality should be avoided for observational data.

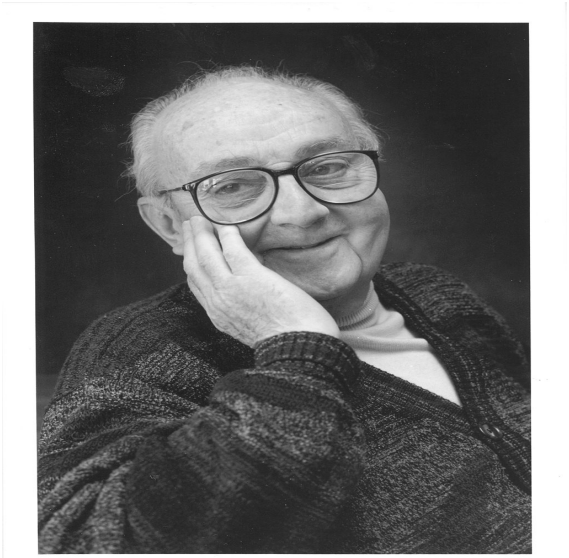
The woes of regression coefficients

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Data Analysis and Regression, Mosteller and Tukey 1977

- A regression coefficient β_j estimates the expected change in Y per unit change in X_j, with all other predictors held fixed. But predictors usually change together!
- Example: Y total amount of change in your pocket; $X_1 = \#$ of coins; $X_2 = \#$ of pennies, nickels and dimes. By itself, regression coefficient of Y on X_2 will be > 0. But how about with X_1 in model?
- $Y =$ number of tackles by a football player in a season; W and H are his weight and height. Fitted regression model is $Y = \beta_0 + 0.50W - 0.10H$. How do we interpret $\hat{\beta}_2 < 0$?

Two quotes by famous Statisticians



1919 - 2013 (aged 93)

- Essentially, all models are wrong, but some are useful.
  George Box
- The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively.
  Fred Mosteller and John Tukey, paraphrasing George Box

Coefficient estimation

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- Given the estimates $\hat{\beta}_0, \hat{\beta}_1, \cdots,$ and $\hat{\beta}_p$, the estimated regression line is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p.$$

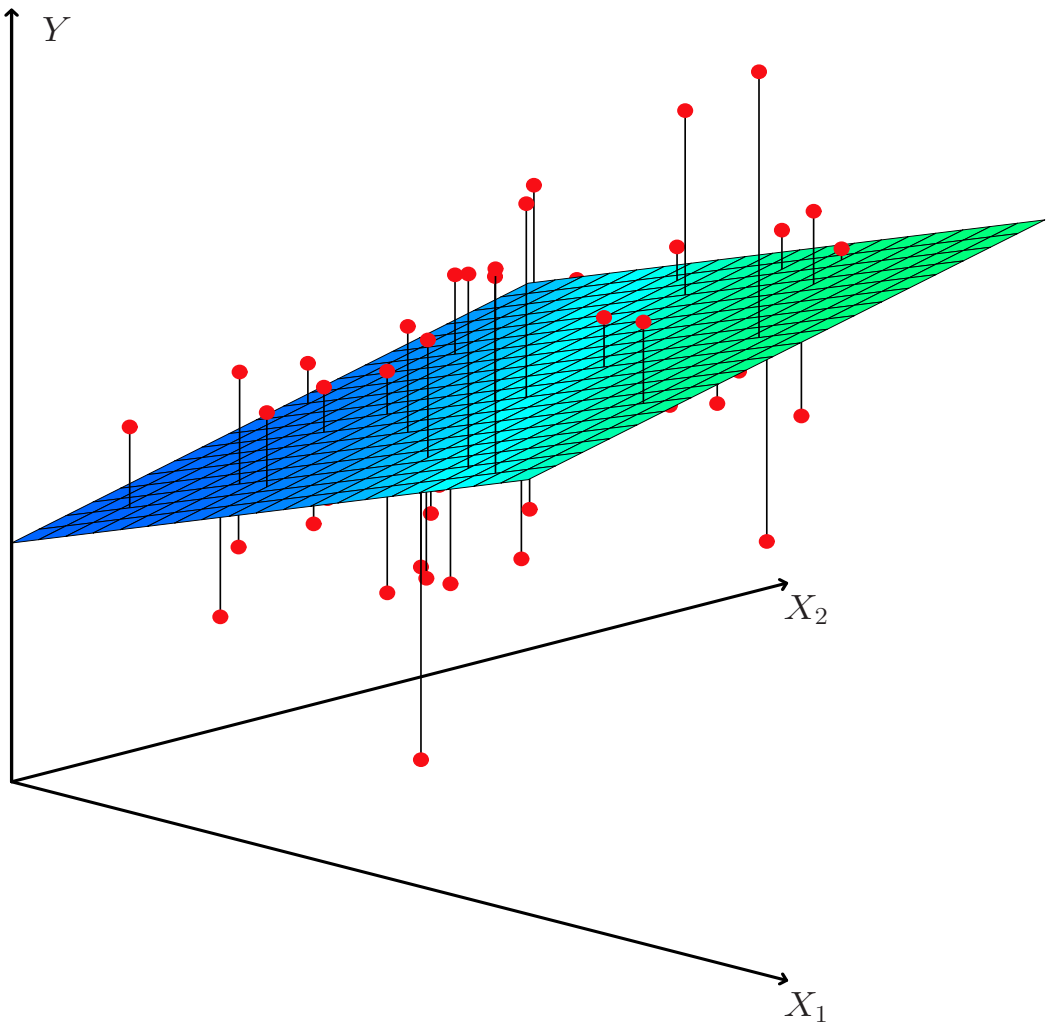
- We estimate all the coefficients $\beta_i, i = 0, 1, \cdots, p$ as the values that minimize the sum of squared residuals

$$\text{RSS} = \sum_i (y_i - \hat{y}_i)^2,$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$ is the predicted values.

- This is done using standard statistical software. The values $\hat{\beta}_0, \hat{\beta}_1, \cdots,$ and $\hat{\beta}_p$ that minimize RSS are the multiple least squares regression coefficient estimates.

Estimation Example



An Introduction to Statistical Learning, by G. James, D. Witten, T. Hastie and R. Tibshirani. 2013, Springer.

Inference

- Is at least one predictor useful?

$$\mathrm{F} = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} \sim F_{p,n-p-1}.$$

- What about an individual coefficient, say if β_i useful?

$$t = \frac{\hat{\beta}_i - 0}{\mathrm{SE}\left(\hat{\beta}_i\right)} \sim t_{n-p-1}.$$

- For given $x_1, \cdots, x_p$, what is the prediction interval (PI) of the corresponding y?
- What about the estimation interval (CI) of y?
- What is the difference — PI, individual and CI, average, PI wider than CI.

Advertising example

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16

> predict(TVadlm, newdata, interval="c", level=0.95)
       fit      lwr      upr
1 20.52397 19.99627 21.05168
> predict(TVadlm, newdata, interval="p", level=0.95)
       fit      lwr      upr
1 20.52397 17.15828 23.88967
```

Indicator Variables

- Some predictors are not quantitative but are qualitative, taking a discrete set of values.
- These are also called categorical predictors or factor variables.
- Example: investigate difference in credit card balance between males and females, ignoring the other variables. We create a new variable,

$$x_i = \begin{cases} 1 & \text{if } i\text{-th person is female,} \\ 0 & \text{if } i\text{-th person is male} \end{cases}.$$

- Resulting model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{-th person is female,} \\ \beta_0 + \varepsilon_i & \text{if } i\text{-th person is male} \end{cases}.$$

- Interpretation and more than two levels (categories)?

Indicator Variables

- In general, if we have k levels, we need $(k-1)$ indicator variables.
- For example, we have 3 levels — $A, B,$ and C for a covariate x,

$$x_A = \begin{cases} 1 & \text{if } x \text{ is A,} \\ 0 & \text{if } x \text{ is not A} \end{cases}; \quad x_B = \begin{cases} 1 & \text{if } x \text{ is B,} \\ 0 & \text{if } x \text{ is not B} \end{cases}.$$

- If x is C, then $x_A = x_B = 0$. We call C as baseline.
- β_A is the contrast between A and C and β_B is the contrast between B and C.

ANOVA by Regression

- Effects model: $y_{ij} = \eta + \tau_i + \epsilon_{ij}, i = 1, \ldots, k; j = 1, \ldots, n_i$,
  where $y_{ij} = j$th observation with treatment i,
  $\epsilon_{ij} =$ error, independent $N(0, \sigma^2)$.
- Constraints:
  - zero-sum is most common: $\sum \tau_i = 0$. With differing group sizes we use instead $\sum n_i \tau_i = 0$, where $\sum n_i = N$.
  - Interpretation: η grand/overall mean, $\tau_i = i$th treatment effect
  - Can be justified as follows: define $\mu_i = E[y_{ij}]$, $\mu = \frac{1}{N} \sum n_i \mu_i = E[\bar{y}_{..}]$; now define $\eta = \mu, \tau_i = \mu_i - \mu$.
  - baseline constraint sometimes used: $\tau_1 = 0$; then $\eta = E[y_{1j}] = \mu_1$ and $\tau_i = \mu_i - \mu_1$.

NOVA by Regression

- Let $\boldsymbol{y} = (\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, \ldots, \boldsymbol{y}_k^T)^T$ and $\boldsymbol{\beta} = (\eta, \tau_2, \ldots, \tau_k)^T$, where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$.
- The design matrixes for the zero-sum and baseline constraints are, respectively,

$$X = \begin{pmatrix} 1 & -n_2/n_1 & \cdots & -n_k/n_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & -n_2/n_1 & \cdots & -n_k/n_1 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 1 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 1 \end{pmatrix}.$$

- Then the regression model is $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$.

Model Fitting

- Under the zero-sum constraint, we have

$$\begin{aligned} y_{ij} &= \hat{\eta} + \hat{\tau}_i + r_{ij} \\ &= \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.}), \end{aligned}$$

where " . " means average over the particular subscript.

- These $(\hat{\eta}, \hat{\tau}_i)$ are the Least Squares Estimates (LSEs) - they minimize

$$S(\eta, \tau_1, ..., \tau_k) = \sum_{i,j} (y_{ij} - \eta - \tau_i)^2 .$$

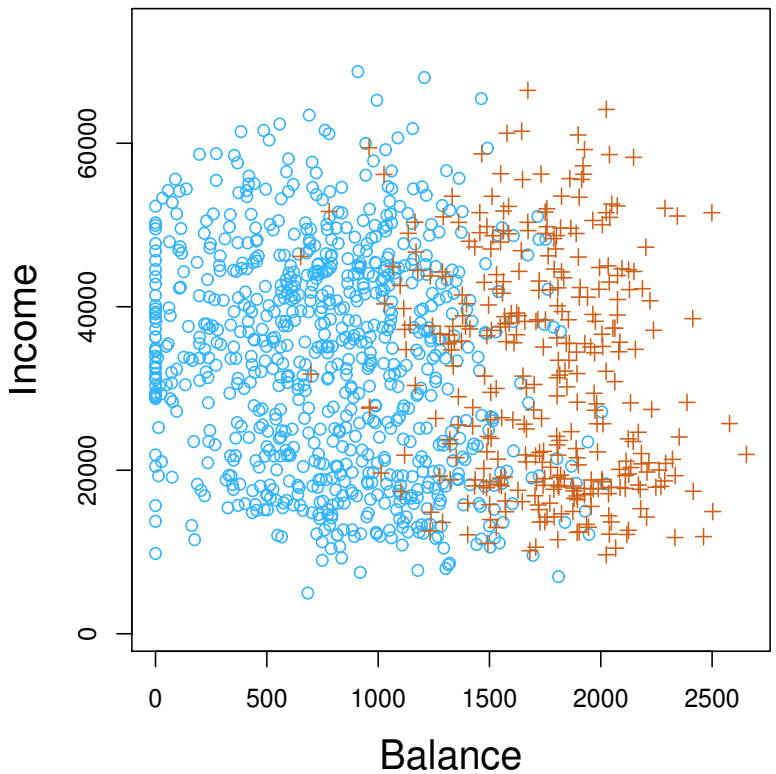
- Proof: Under the zero-sum constraints we have

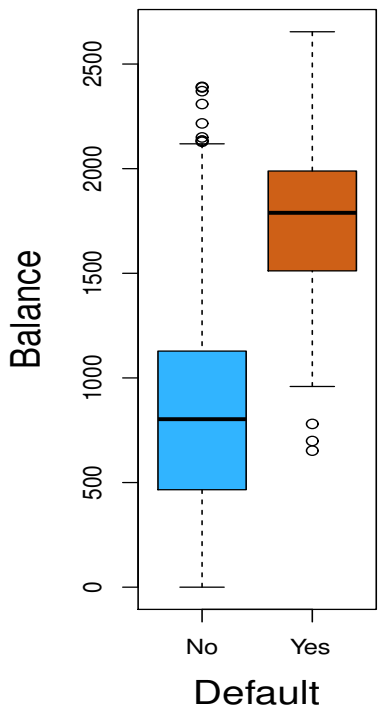
$$\begin{aligned} S &= \sum_{i,j} \{(y_{ij} - \bar{y}_{i.}) + (\hat{\tau}_i - \tau_i) + (\hat{\eta} - \eta)\}^2 \\ &= \sum_{i,j} (y_{ij} - \bar{y}_{i.})^2 + \sum_i n_i (\hat{\tau}_i - \tau_i)^2 + N(\hat{\eta} - \eta)^2 \end{aligned}$$

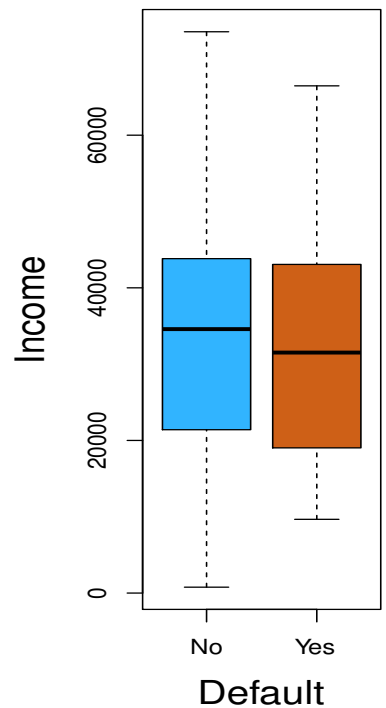
Qualitative Response

- There are many qualitative response taking values in an unordered set $\mathcal{C}$ such as
  `eye color ∈ {brown; blue; green}`.
- Given a feature vector X and a qualitative response Y taking values in the set $\mathcal{C}$, the classification task is to build a function $C(X)$ (learn a rule) that takes as input the feature vector X and predicts its value for Y; i.e. $C(X) \in \mathcal{C}$.
- Often we are more interested in estimating the probabilities that X belongs to each category in $\mathcal{C}$.
- For example, it is more valuable to have an estimate of the probability that an insurance claim is fraudulent, than a classification fraudulent or not.

Cedit Card Defualt







Individuals who defaulted in a given month in orange, and did not in blue.

An Introduction to Statistical Learning, by G. James, D. Witten, T. Hastie and R. Tibshirani. 2013, Springer.

Linear Regression Model

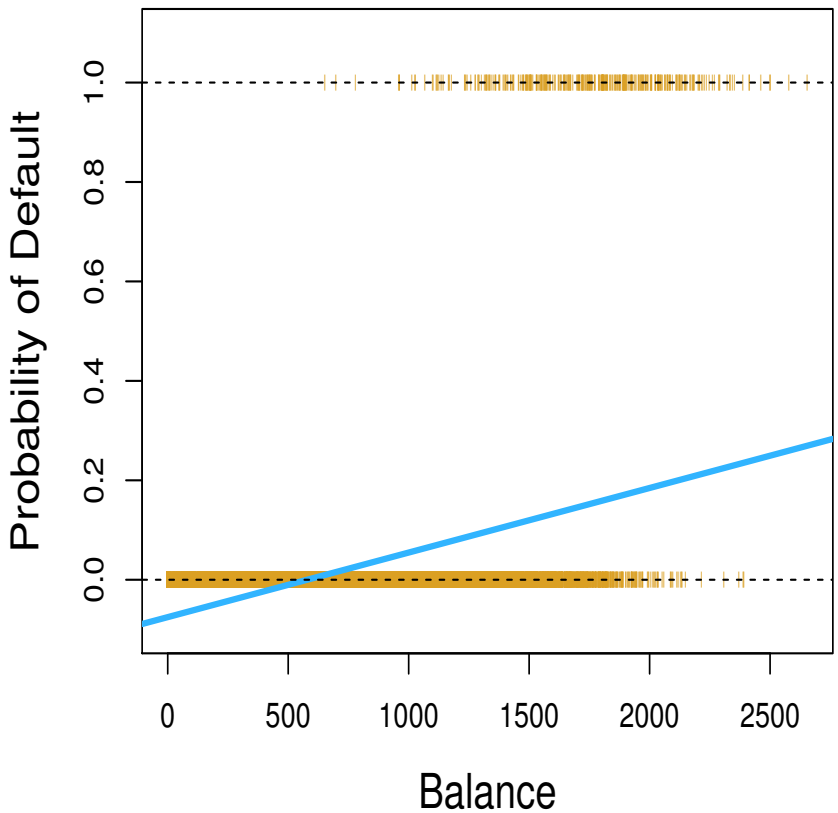
- Suppose for the Default classification task that we code

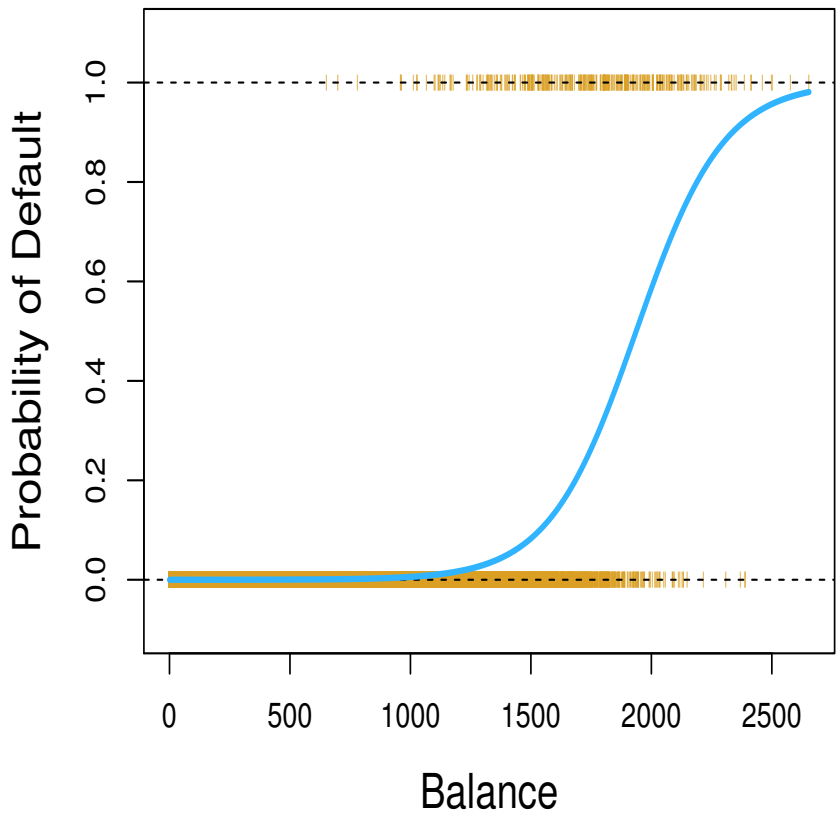
$$Y = \begin{cases} 1 & \text{if Yes} \\ 0 & \text{if No} \end{cases}.$$

  Can we simply perform a linear regression of Y on X and classify as Yes if $\hat{Y} > 0.5$?

- In this case of a binary outcome, linear regression does a good job as a classifier, and is equivalent to linear discriminant analysis which we discuss later.
- Since in the population $\mathrm{E}(Y|X = x) = \Pr(Y = 1|X = x)$, we might think that regression is perfect for this task.
- However, linear regression might produce probabilities less than zero or bigger than one. Logistic regression is more appropriate.

Credit data example





The orange marks indicate the response Y, either 0 or 1. Linear regression does not estimate $\Pr(Y = 1|X)$ well. Logistic regression seems well suited to the task.

An Introduction to Statistical Learning, by G. James, D. Witten, T. Hastie and R. Tibshirani. 2013, Springer.

Logistic Regression

- Denote $p(X) = \Pr(Y = 1|X)$ consider using balance to predict default. Logistic regression uses the form

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

- It is easy to see that no matter what values β_0, β_1 or X take, $p(X)$ will have values between 0 and 1.
- A bit of rearrangement gives

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X.$$

- This monotone transformation is called the log odds or logit transformation of $p(X)$.

Estimation

- We use maximum likelihood to estimate the parameters.

$$l(\beta_0, \beta_1) = \Pi_{i:y_i=1} p(x_i) \Pi_{i:y_i=0} (1 - p(x_i)).$$

- This likelihood gives the probability of the observed zeros and ones in the data. We pick β_0 and β_1 to maximize the likelihood of the observed data.
- Most statistical packages can fit linear logistic regression models by maximum likelihood. In R we use the glm function.

```
> glm.fit=glm(default~balance,data=defaultData,family=binomial)
> summary(glm.fit)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

Interpretation

- Interpreting what β_1 means is not very easy with logistic regression, simply because we are predicting $\Pr(Y = 1|X)$ and not Y.
- If $\beta_1 = 0$, this means that there is no relationship between Y and X.
- If $\beta_1 > 0$, this means that when X gets larger so does the probability that $Y = 1$.
- If $\beta_1 < 0$, this means that when X gets larger, the probability that $Y = 1$ gets smaller.
- But how much bigger or smaller depends on where we are on the slope.

Hypothesis Testing

- We still want to perform a hypothesis test to see whether we can be sure that are β_0 and β_1 significantly different from zero.
- We use a z test instead of a t test, but of course that doesn't change the way we interpret the p-value
- Here the p-value for balance is very small, and β_1 is positive, so we are sure that if the balance increase, then the probability of default will increase as well.

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

Prediction

- What is our estimated probability of `default` for someone with a balance of 1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.5613+0.0055\times 1000}}{1+e^{-10.5613+0.0055\times 1000}} = 0.006.$$

- The predicted probability of default for an individual with a balance of \$1000 is less than 1%.
- For a balance of \$2000, the probability is much higher, and equals to 0.586(58.6%).

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.5613+0.0055\times 2000}}{1+e^{-10.5613+0.0055\times 2000}} = 0.586.$$

```
> predict(glm.fit, list(balance = c(1000,2000)),type="response")
          1           2
0.005752145 0.585769370
```

Logistic Regression with indicator variable

- We can predict if an individual default by checking if she is a student or not. Thus we can use a qualitative variable `Student` coded as (Student $= 1$, Non-student $= 0$).

```
> glm.fit=glm(default~student,data=defaultData,family=binomial)
> summary(glm.fit)
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.50413    0.07071  -49.55  < 2e-16 ***
factor(student)Yes  0.40489    0.11502    3.52 0.000431 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

- β_1 is positive. This indicates students tend to have higher default probabilities than non-students.

```
> predict(glm.fit, list(student = c('Yes','No')),type="response")
         1          2
0.04313859 0.02919501
```

Multiple logistic Regression

- Logistic Regression with several covariates

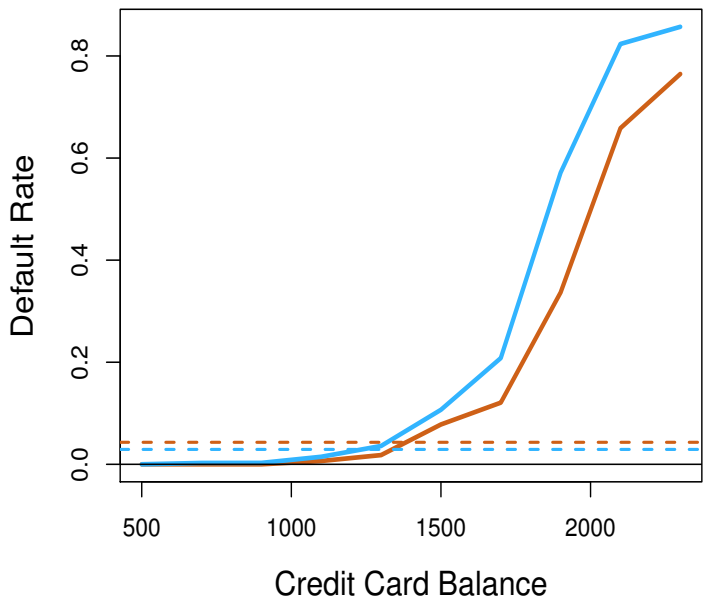
$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

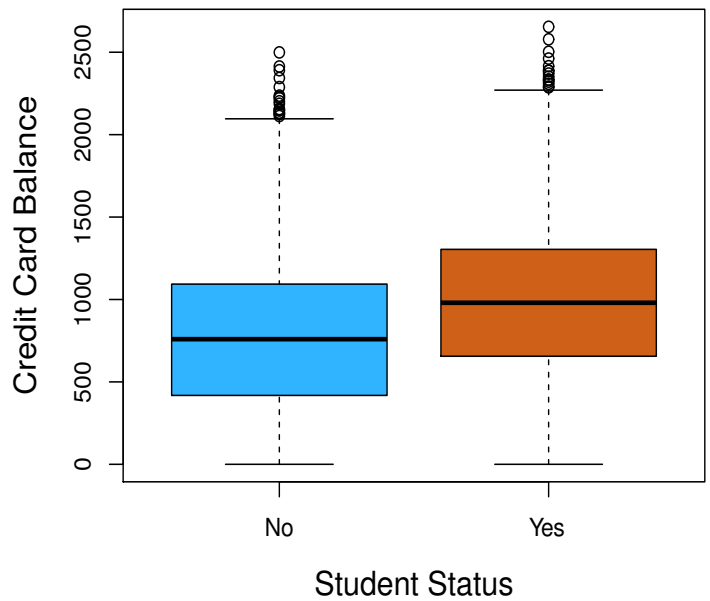
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0+\beta_1 X_1+\cdots+\beta_p X_p.$$

```
> glm.fit=glm(default~balance+income+student,data=defaultData,family=
> summary(glm.fit)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

- Why is coefficient for `student` negative, while it was positive before?

Confounding





- Students tend to have higher balances than non-students, so their marginal default rate is higher than for non-students.
- But for each level of balance, students default less than non-students.
- Multiple logistic regression can tease this out.

An Introduction to Statistical Learning, by G. James, D. Witten, T. Hastie and R. Tibshirani. 2013, Springer.

Linear Mixed Models

- *Linear mixed models (LMM)* are defined as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \quad \boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{\psi}_{\boldsymbol{\theta}}), \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$$

- $\boldsymbol{y}$ is the response vector
- $\boldsymbol{X}$ and $\boldsymbol{Z}$ are fixed design matrix with *fixed and random effects* $\boldsymbol{\beta}$ and $\boldsymbol{b}$, respectively.
- $\boldsymbol{\psi}_{\boldsymbol{\theta}}$ is positive definition variance matrix for random effect $\boldsymbol{b}$ and depends on some parameter $\boldsymbol{\theta}$.
- $\boldsymbol{I}$ is positive definite, usually with simple structure.
- $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ are independent.

Applications

- Clustered Data
  1. response is measured for each subject
  2. each subject belongs to a group of subjects (cluster)
- Ex.:
  1. math scores of student grouped by classrooms (class room forms cluster)
  2. birth weigths of rats grouped by litter (litter forms cluster)
- Longitudinal Data
  1. response is measured at several time points
  2. number of time points is not too large (in contrast to time series)
- Ex.: sales of a product at each month in a year (12 measurements)

Estimation of Fixed Effects

- Let $\boldsymbol{e} = \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$, then we have $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ with $\boldsymbol{e} \sim (\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\sigma^2)$, where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{Z}\psi_{\boldsymbol{\theta}}\boldsymbol{Z}^T/\sigma^2 + \boldsymbol{I}$.
- To estimate $\boldsymbol{\beta}$, we use the weighted LSE, that is,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}.$$

- It is easy to estimate σ^2, say, unbiased estimate $\sigma^2 = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T\Sigma_{\boldsymbol{\theta}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/(n-p)$.
- Now we know how to estimate $\boldsymbol{\beta}$ and σ^2.
- How do we estimate $\boldsymbol{\theta}$? We use *profiling likelihood*.

Prediction of Random Effects

- To predict random effects $\boldsymbol{b}|\boldsymbol{y}$, we need some facts on *multivariate Normal Distribution*.
- The prediction of $\boldsymbol{b}$ given $\boldsymbol{y}$ is $\hat{\boldsymbol{b}} = \psi_{\hat{\boldsymbol{\theta}}} \boldsymbol{Z}^T \boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}^{-1} (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/\hat{\sigma}^2$.
- This is the best linear unbiased prediction of $\boldsymbol{b}$ (BLUP).
- The predicted value of $\boldsymbol{y}$ is

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{b}}.$$

Inference on fixed effects

- For estimated fixed effects $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}$, we have

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\boldsymbol{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{X})^{-1}\sigma^2),$$

  where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \boldsymbol{Z}\boldsymbol{\psi}_{\boldsymbol{\theta}}\boldsymbol{Z}^T/\sigma^2 + \boldsymbol{I}$.
- We can use the estimated $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}^2$ to plug in to obtain $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}}$ and carry on to find confidence intervals.
- It is easy to conduct hypothesis test on single or multiple fixed coefficients, say, t-test.
- For linear combination of fixed coefficient, we can do Wald-type test or likelihood ratio test.
- However, the plug-in may underestimate the variance of $\hat{\boldsymbol{\beta}}$, as it does not consider the *variance* of $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}^2$.

Inference on random effects

- It is not easy to do inference on *random variable* in LMM.
- For example, if we want to test if there exists one random variable, say for covariate z_1, we then need to test if $\psi_{11} = 0$, which can not be easily carried out using *likelihood ratio test*.
- *WHY?* It is on the boundary. All the past MLE results on $\hat{\boldsymbol{\beta}}$ may not be available.
- The test statistics is well defined but the distribution (under some conditions, it is weighted χ^2 mixture)is not easy to find.
- This is still an active research area, especially nowadays *functional data analysis (FDA)* is borrowing strength from LMM.

Linear Mixed Models in R

- In R, the package `nlme` is available for LMM.
- However, by default, the main function `lme` assumes that your data are grouped according to the levels of some factors.
- Meanwhile, the random effects structure are the same for each group and are independent between groups.
- For example, `lme(y~ x1+x2,data,random=~x1|g)` or `lme(y~x1+x2,data,list(g=~x1))`.
- This means a model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + b_1 x_1 + \varepsilon,$$

where β_0, β_1, and β_2 are fixed effects, and b_1 is a random effect.

Example

```
> library(MASS)
> data(oats)
> names(oats) = c('block', 'variety', 'nitrogen', 'yield')
> oats$mainplot = oats$variety
> oats$subplot = oats$nitrogen
>
> summary(oats)
 block           variety     nitrogen      yield              mainplot    subplot
 I  :12   Golden.rain:24   0.0cwt:18   Min.   : 53.0   Golden.rain:24   0.0cwt:18
 II :12   Marvellous :24   0.2cwt:18   1st Qu.: 86.0   Marvellous :24   0.2cwt:18
 III:12   Victory    :24   0.4cwt:18   Median :102.5   Victory    :24   0.4cwt:18
 IV :12                    0.6cwt:18   Mean   :104.0                    0.6cwt:18
 V  :12                                3rd Qu.:121.2
 VI :12                                Max.   :174.0
```

Example

- ```
  > oats
     block     variety nitrogen yield    mainplot subplot
  1      I     Victory   0.0cwt   111     Victory  0.0cwt
  2      I     Victory   0.2cwt   130     Victory  0.2cwt
  3      I     Victory   0.4cwt   157     Victory  0.4cwt
  4      I     Victory   0.6cwt   174     Victory  0.6cwt
  5      I Golden.rain   0.0cwt   117 Golden.rain  0.0cwt
  6      I Golden.rain   0.2cwt   114 Golden.rain  0.2cwt
  7      I Golden.rain   0.4cwt   161 Golden.rain  0.4cwt
  8      I Golden.rain   0.6cwt   141 Golden.rain  0.6cwt
  9      I  Marvellous   0.0cwt   105  Marvellous  0.0cwt
  10     I  Marvellous   0.2cwt   140  Marvellous  0.2cwt
  11     I  Marvellous   0.4cwt   118  Marvellous  0.4cwt
  12     I  Marvellous   0.6cwt   156  Marvellous  0.6cwt
  13    II     Victory   0.0cwt    61     Victory  0.0cwt
  14    II     Victory   0.2cwt    91     Victory  0.2cwt
  15    II     Victory   0.4cwt    97     Victory  0.4cwt
  ...
  ```
- Totally 72 subjects: 6 `block` and each `block` has 12 subjects
- 3 `variety`/`mainplot` nested within each `block`. Within each `block`, each `variety`/`mainplot` has 4 subjects.
- 4 `nitrogen`/`subplot` nested within each `variety`/`mainplot`. Within each `variety`/`mainplot`, each `nitrogen`/`subplot` has 1 subject.
  ```

Example

- Let x_1 denote `block` and x_{1i} be the indicator of the i-th `block`.
- Let x_2 denote `variety/mainplot` and x_{2j} be the indicator of the j-th `variety/mainplot` .
- Let x_3 denote `nitrogen/subplot` and x_{3k} be the indicator of the k-th `nitrogen/subplot`.
- We look at the model

$$y_{ijk} = \beta_0 + \sum_{j=1}^{2} \beta_{2j} x_{2j} + \sum_{k=1}^{3} \beta_{3k} x_{3k} + \sum_{j=1}^{2} \sum_{k=1}^{3} \beta_{23jk} x_{2j} x_{3k} + \sum_{i=1}^{5} x_{1i} b_{1i} + \varepsilon_{ijk},$$

where β's are fixed effects and b's are random effects.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Example

```
> library(nlme)
> m0.nlme = lme(yield ~ variety*nitrogen, random = ~1|block,data = oats)
> summary(m0.nlme)
Linear mixed-effects model fit by REML
 Data: oats
      AIC      BIC    logLik
   564.69 594.0108 -268.345

Random effects:
 Formula: ~1 | block
        (Intercept) Residual
StdDev:    15.60138 15.94425

Fixed effects: yield ~ variety * nitrogen
                                  Value Std.Error DF   t-value p-value
(Intercept)                    80.00000  9.106977 55  8.784473  0.0000
varietyMarvellous               6.66667  9.205419 55  0.724211  0.4720
varietyVictory                 -8.50000  9.205419 55 -0.923369  0.3598
nitrogen0.2cwt                 18.50000  9.205419 55  2.009686  0.0494
nitrogen0.4cwt                 34.66667  9.205419 55  3.765898  0.0004
nitrogen0.6cwt                 44.83333  9.205419 55  4.870320  0.0000
varietyMarvellous:nitrogen0.2cwt  3.33333 13.018428 55  0.256047  0.7989
varietyVictory:nitrogen0.2cwt    -0.33333 13.018428 55 -0.025605  0.9797
varietyMarvellous:nitrogen0.4cwt -4.16667 13.018428 55 -0.320059  0.7501
varietyVictory:nitrogen0.4cwt     4.66667 13.018428 55  0.358466  0.7214
varietyMarvellous:nitrogen0.6cwt -4.66667 13.018428 55 -0.358466  0.7214
varietyVictory:nitrogen0.6cwt     2.16667 13.018428 55  0.166431  0.8684
...

Number of Observations: 72
Number of Groups: 6
```

Example

```
> anova(m0.nlme)
                  numDF denDF   F-value p-value
(Intercept)           1    55 245.14092  <.0001
variety               2    55   3.51343  0.0366
nitrogen              3    55  26.25097  <.0001
variety:nitrogen      6    55   0.21094  0.9719
> m0.nlme$coef
$fixed
                     (Intercept)                varietyMarvellous ...
                      80.0000000                        6.6666667 ...
                   nitrogen0.4cwt                   nitrogen0.6cwt ...
                      34.6666667                       44.8333333 ...
varietyMarvellous:nitrogen0.4cwt    varietyVictory:nitrogen0.4cwt ...
                      -4.1666667                        4.6666667 ...

$random
$random$block
    (Intercept)
I     28.850101
II     3.015334
III   -7.410566
IV    -5.340718
V    -12.010228
VI    -7.103922
```

Example

- ```
  > m1.nlme = lme(yield ~ variety*nitrogen,random = ~ 1|block/mainplot,data = oats)
  >
  > #summary(m1.nlme)
  > #anova(m1.nlme)
  > #m1.nlme$coef
  >
  > anova(m0.nlme,m1.nlme)
          Model df      AIC      BIC    logLik   Test L.Ratio p-value
  m0.nlme     1 14 564.6900 594.0108 -268.3450
  m1.nlme     2 15 559.0285 590.4437 -264.5143 1 vs 2 7.66146  0.0056
  ```
- We can also specify the covariance structure of random effects $\boldsymbol{b}$.
  ```

Introduction

- Principal Component Analysis (PCA) produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have maximal variance, and are mutually uncorrelated.
- Apart from producing derived variables for use in supervised learning problems, say dimension reduction, PCA also serves as a tool for data visualization.
- PCA was invented in 1901 by Karl Pearson, as an analogue of the principal axis theorem in mechanics; it was later independently developed (and named) by Harold Hotelling in the 1930s.
- other names: Kosambi-Karhunen-Loève transform (KLT), spectral decomposition ...

Principal Components Analysis

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- The first principal component of a set of features $X_1, \cdots, X_p$ is the normalized linear combination of the features

$$Z_1 = \phi_{11} X_1 + \cdots + \phi_{p1} X_p$$

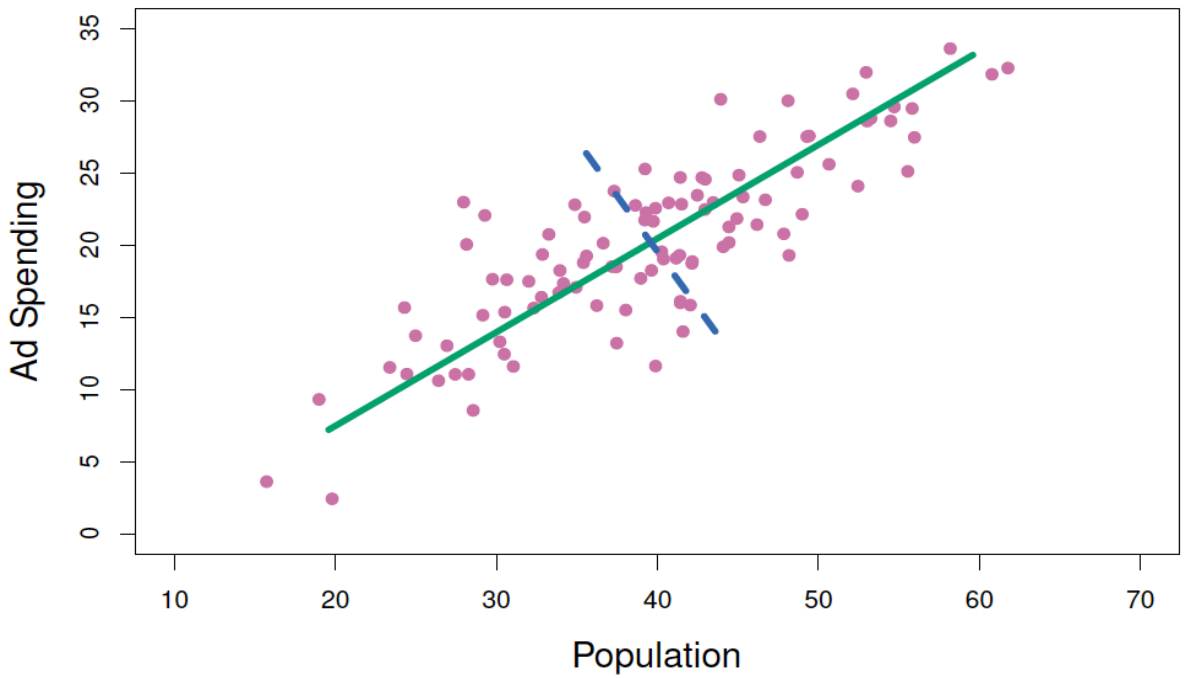
that has the largest variance. By normalized, we mean that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.

- We refer to the elements $\phi_{11}, \cdots, \phi_{p1}$ as the loadings of the first principal component; together, the loadings make up the principal component loading vector, $\phi_1 = (\phi_{11}, \cdots, \phi_{p1})^T$.
- We constrain the loadings so that their sum of squares is equal to one, since otherwise setting these elements to be arbitrarily large in absolute value could result in an arbitrarily large variance.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

An Example

- The population size (pop) and ad spending (ad) for 100 different cities are shown as purple circles.
- The green solid line indicates the first principal component direction, and the blue dashed line indicates the second principal component direction.



Geometry of PCA

- The loading vector ϕ_1 with elements $\phi_{11}, \cdots, \phi_{p1}$ defines a direction in feature space along which the data vary the most.
- If we project the n data points $x_1, \cdots, x_n$ onto this direction, the projected values are the principal component scores $z_{11}, \cdots, z_{n1}$ themselves.

Further principal components

- The second principal component is the linear combination of $X_1, \cdots, X_p$ that has maximal variance among all linear combinations that are uncorrelated with Z_1.
- The second principal component scores $z_{12}, \cdots, z_{n2}$ take the form

$$z_{i2} = \phi_{12} x_{i1} + \cdots + \phi_{p2} x_{ip},$$

where ϕ_2 is is the second principal component loading vector, with elements $\phi_{12}, \cdots, \phi_{p2}$.

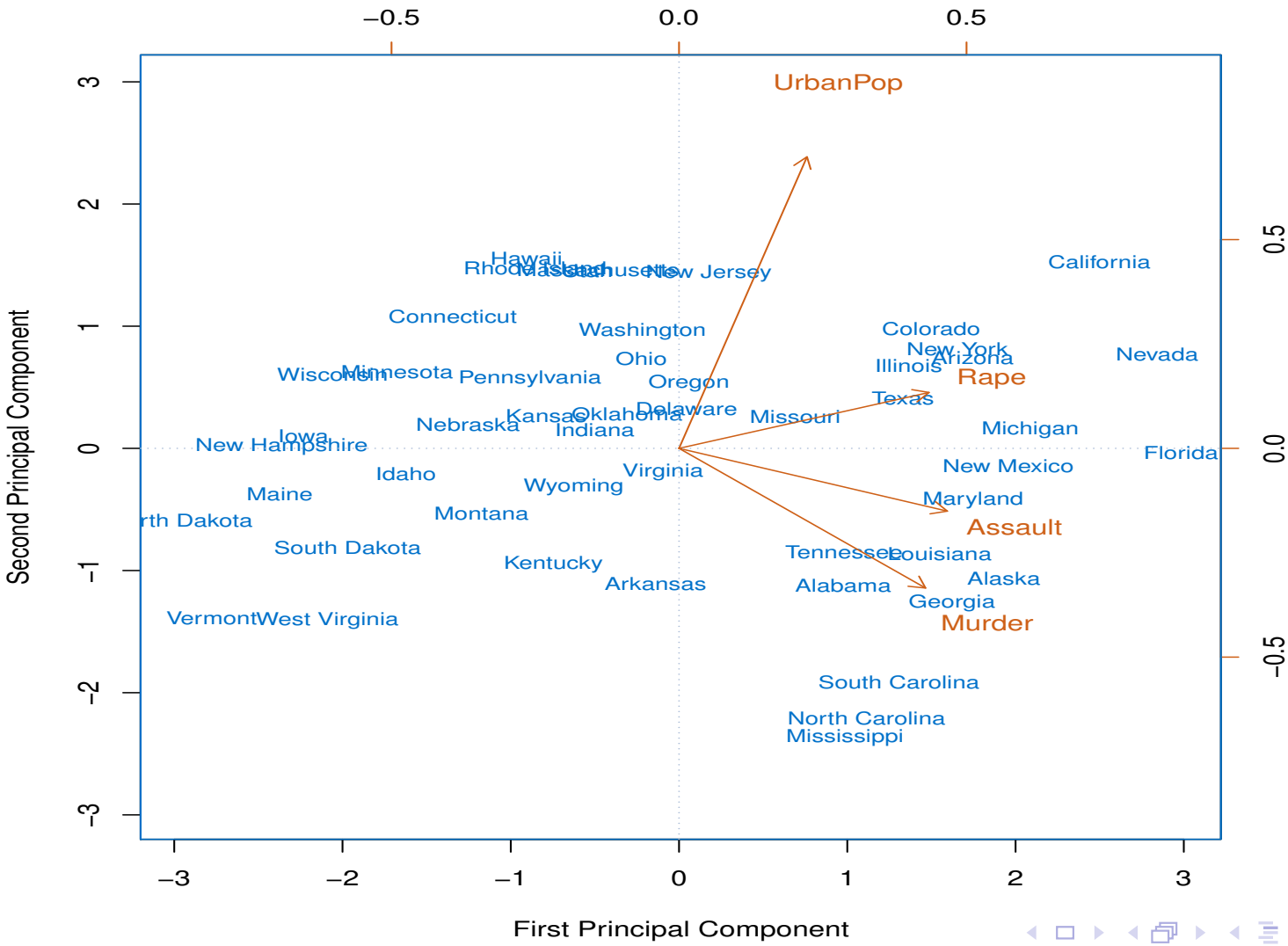
Illustration

- `USAarrests` data: For each of the fifty states in the United States, the data set contains the number of arrests per $100,000$ residents for each of three crimes: `Assault`, `Murder`, and `Rape`. We also record `UrbanPop` (the percent of the population in each state living in urban areas).
- The principal component score vectors have length $n = 50$, and the principal component loading vectors have length $p = 4$.
- PCA was performed after standardizing each variable to have mean zero and standard deviation one.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

PCA plot

The first two principal components for the `USArrests` data.



PCA plot

- The blue state names represent the scores for the first two principal components.
- The orange arrows indicate the first two principal component loading vectors (with axes on the top and right).
- For example, the loading for `Rape` on the first component is 0.54, and its loading on the second principal component 0.17 [the word `Rape` is centered at the point $(0.54, 0.17)$].
- This figure is known as a biplot, because it displays both the principal component scores and the principal component loadings.

PCA loadings

The first two principal component loadings for the `USArrests` data.

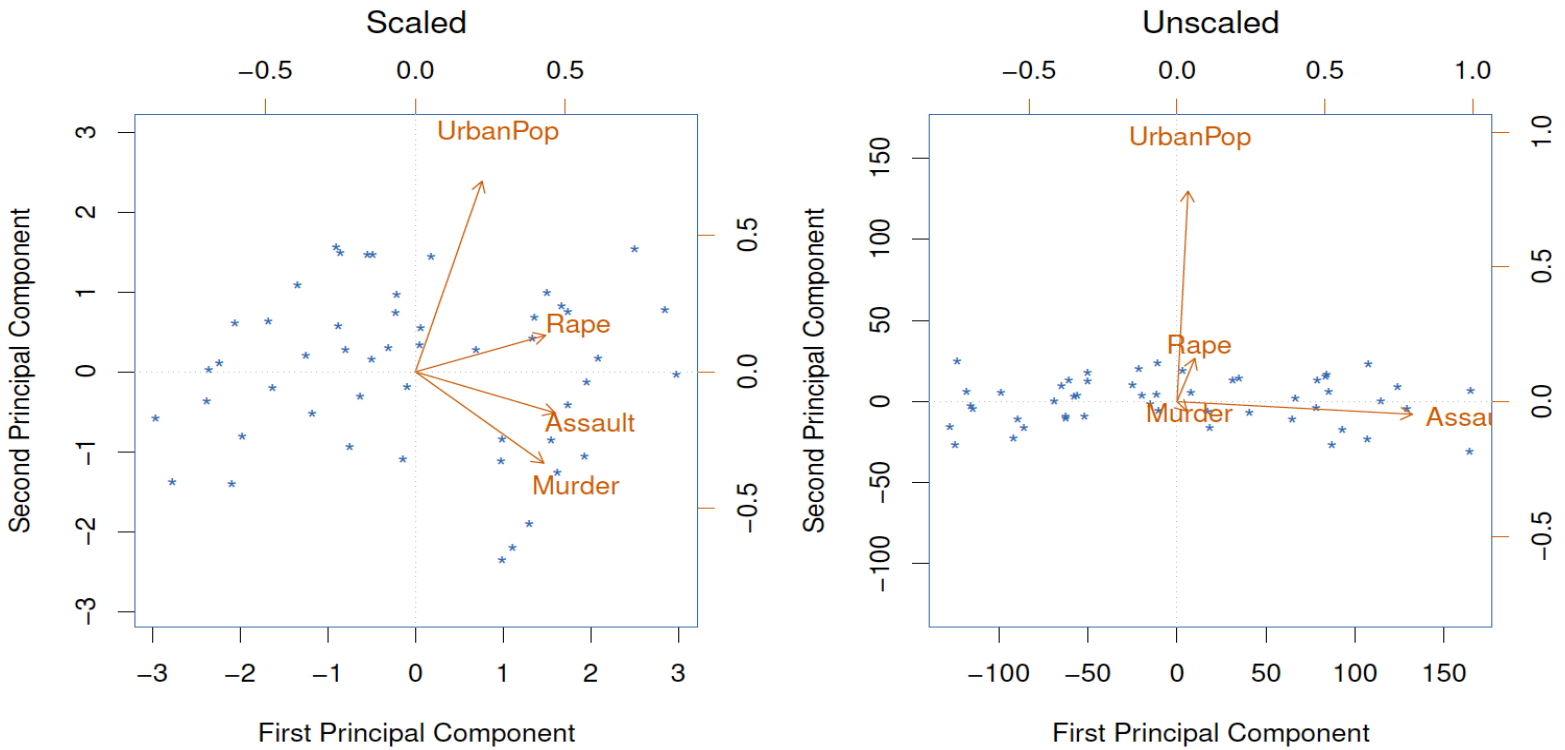
	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

An Introduction to Statistical Learning, by G. James, D. Witten, T. Hastie and R. Tibshirani. 2013, Springer.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Scaling of the variables matters

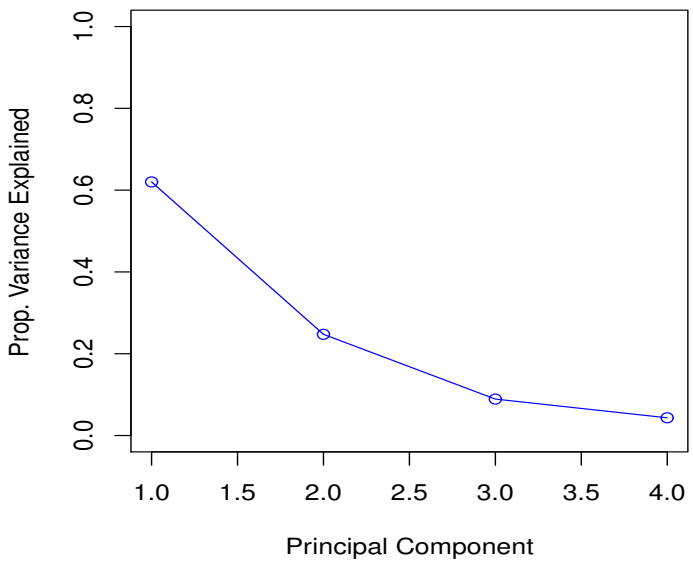
- If the variables are in different units, scaling each to have standard deviation equal to one is recommended.
- If they are in the same units, you might or might not scale the variables.

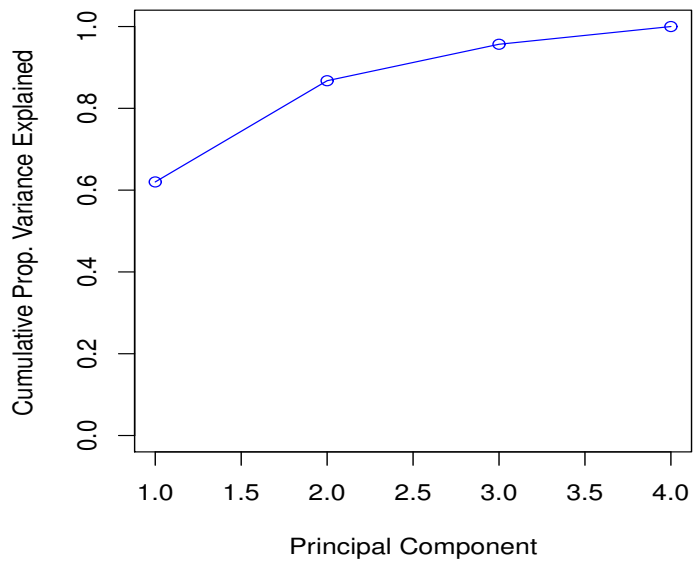


An Introduction to Statistical Learning, by G. James, D. Witten, T. Hastie and R. Tibshirani. 2013, Springer.

Proportion Variance Explained

- To understand the strength of each component, we are interested in knowing the proportion of variance explained (PVE) by each one.
- The PVE of the mth principal component is given by the positive quantity between 0 and 1.
- The PVEs sum to one. We sometimes display the cumulative PVEs.





An Introduction to Statistical Learning, by G. James, D. Witten, T. Hastie and R. Tibshirani. 2013, Springer.

Number of PCA

- If we use principal components as a summary of our data, how many components are sufficient?
- No simple answer to this question, as cross-validation is not available for this purpose. Why not?
- When could we use cross-validation to select the number of components?
- the scree plot on the previous slide can be used as a guide: we look for an elbow.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

p-value

The statement's six principles, many of which address misconceptions and misuse of the *p*-value, are the following:

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

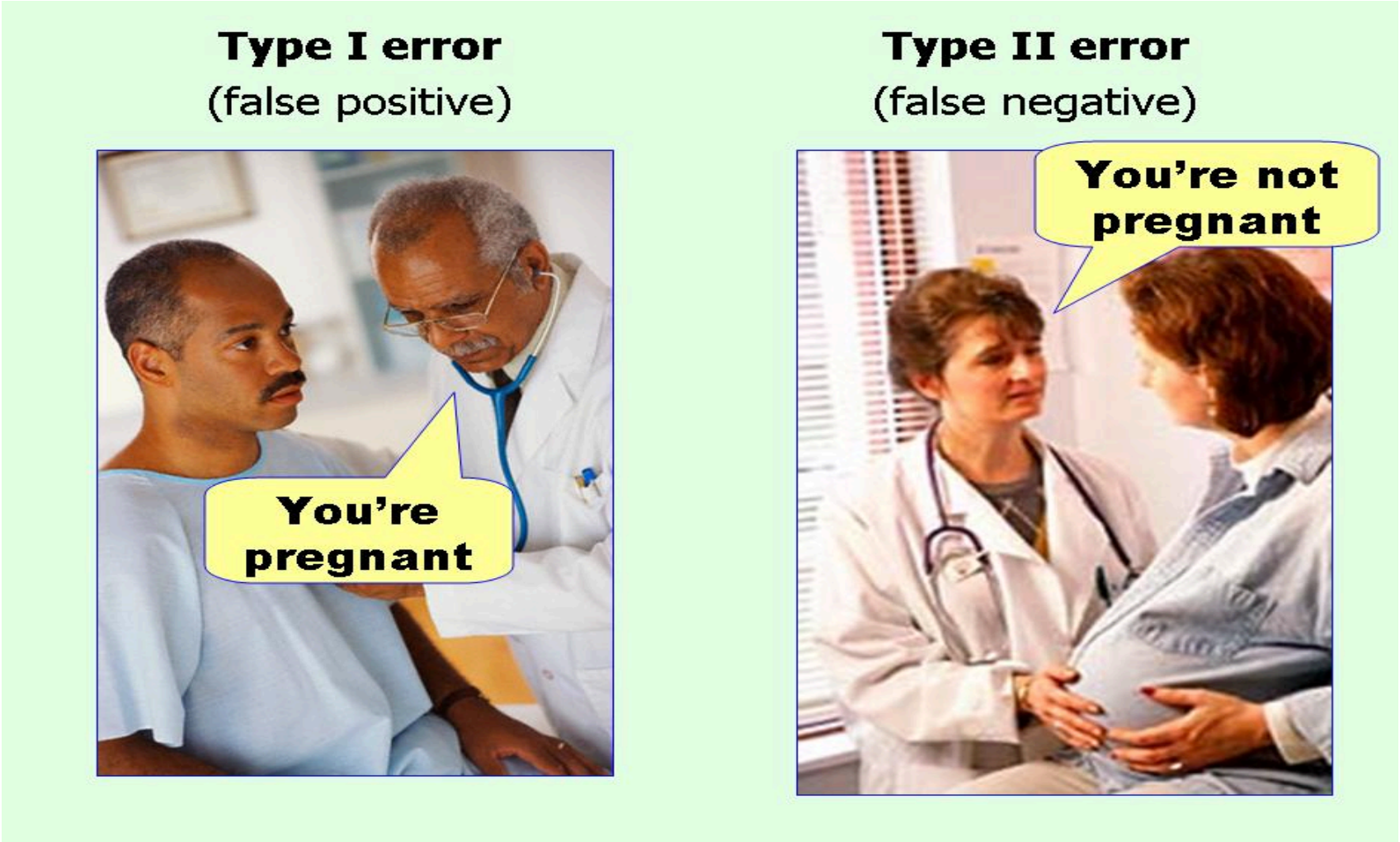
Two errors

Helpful Chart for remembering definitions

Drawing this chart before doing any power presentation will be very helpful!!!

Decision From Sample	Truth About The Population: Ho: is true	Truth About The Population: Ha: is true
Reject Ho:	Type I Error α	**POWER** $1-\beta$
Fail to Reject Ho:	Correct	Type II Error β

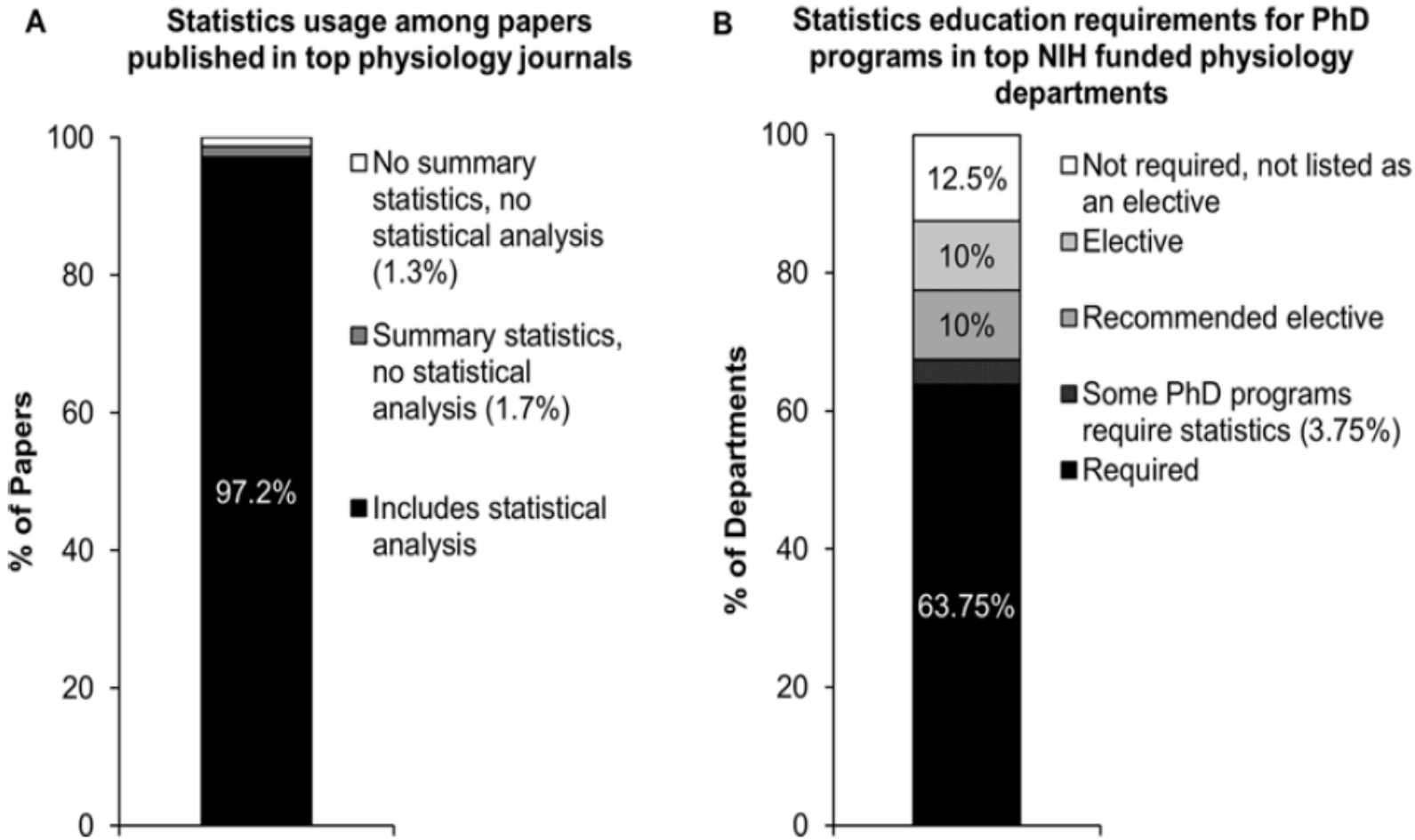
Another look at the two errors



https://effectsizefaq.files.wordpress.com/2010/05/type-i-and-type-ii-errors.jpg

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Reinventing Biostatistics Education for Basic Scientists



Reinventing Biostatistics Education for Basic Scientists

Our recommendations for improving statistics training for basic biomedical scientists include:

- Encouraging departments to require statistics training
- Tailoring coursework to the student?s field of research
- Developing tools and strategies to promote education and dissemination of statistical knowledge

Reinventing Biostatistics Education for Basic Scientists

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

To Code or Not to Code. . .That Is the Question

- The ability to reuse code enhances reproducibility and saves time: Analyses run in coding-based programs are more reproducible.
- Cost and accessibility:Most universities and research centers purchase an institutional license for a particular statistics program with a user interface, then build their courses around that program.
- Ability to run more complex analyses: Multidisciplinary research training is becoming increasingly common, and young investigators may transition among different fields or specialties early in their careers.
- Promoting knowledge retention: Programs that have a user interface often make decisions about what test to use based on the characteristics of the data.

Solution: to use R, a code-based free-of-charge statistical program.

Take courses and consulting statisticians

The courses we offer

- STAT 151252 introduction to applied statistics - too simple?
- STAT 337 Biostatistics
- STAT 441 Applied statistical method in data mining
- STAT 378 Applied Regression Analysis
- STAT 368 Experiment Design

Consider Training Consulting Centre,

`http://www.stat.ualberta.ca/~tcc/#home`

1pm-3pm TWR @ CAB493

Summary

The following are some guidelines you may want to consider in your research.

- Formulate your questions
- Design your experiment
- Collect your data
- Summary statistics and Exploratory plots
- t-test, ANOVA, Linear regression, and so on
- Model assessment - assumptions, normality, equal variance, sample size, residual analysis, outlier detection ...
- Power analysis
- Draw conclusion, statistical and scientific?
- presentation and Limitations