

# Models

Merlise Clyde

January 18, 2017

# Problem Setting

- ▶ Data: Observe for each case  $i$  ( $Y_i, X_i$ )
- ▶ Response or dependent variable  $Y_i$
- ▶ Predictor(s) or **independent** variable  $X_i$

Goals:

- ▶ Exploring distribution of  $p(y|X = x)$  as a function of  $x$
- ▶ Understanding the mean in  $Y$  as a function of  $x$  :  
$$E(Y | X = x) = f(x)$$

## Special cases:

- ▶ regression (normal  $Y$ ) or additive error model
- ▶ classification (binary or Bernoulli  $Y$  where probability  $p(Y = 1 | X)$  depends on  $x$ )
- ▶ exponential family models
  - ▶ Poisson regression (counts)
  - ▶ Gamma regression (continuous, positive)
- ▶ Survival models

# Additive Error Model

- ▶ Assume  $E[\epsilon_i] = 0$  for  $i = 1, \dots, n$ ,

$$Y_i = f(X_i) + \epsilon_i$$

- ▶ Regression function  $E(Y | x) = f(x)$
- ▶ ideal or optimal predictor of  $Y$  given  $X = x$
- ▶ minimizes  $E[(Y - g(x))^2 | X = x]$  over all functions  $g(x)$  at all points  $X = x$
- ▶ for prediction  $\epsilon = Y - f(x)$  is *irreducible error* as even if we know  $f(x)$  there are still errors in predicting  $Y$
- ▶ for any estimator  $\hat{f}(x)$  we have

$$E[(Y - \hat{f}(x))^2 | X = x] = \underbrace{(f(x) - \hat{f}(x))^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}$$

# Linear Regression

- ▶ Taylor's series expansion of regression function about point  $x_0$

$$f(x_i) = f(x_0) + f'(x_0)(x_i - x_0) + \text{Remainder}$$

leads to locally linear approximation

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

- ▶  $\varepsilon_i$  : errors (sampling/measurement errors  $\epsilon$ , lack of fit)

# Regression in Matrix Notation

Simple Linear Regression:

$$Y_i = \beta_0 + x_i\beta_1 + \epsilon_i \text{ for } i = 1, \dots, n$$

Rewrite in vectors:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \beta_1 + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

## Big Picture:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \stackrel{iid}{\sim} N(0, \sigma^2)$$

- ▶ Estimate parameters  $(\beta, \sigma)$
- ▶ interpretation of parameters:  $\beta, \sigma$
- ▶ assess model fit — adequate? good? if inadequate, how?
- ▶ move to more complicated model?
- ▶ predict new (“future”) responses at new  $x_{n+1}, \dots$
- ▶ how much variability does  $x$  explain?
- ▶ how important is  $X$  is predicting  $Y$

# Body Fat Data

- ▶ For a group of 252 male subjects, various body measurements were obtained
- ▶ An accurate measurement of the percentage of body fat is recorded for each
- ▶ Goal is to use the other body measurements as a proxy for predicting body fat
- ▶ Understand how changing one measurement may lead to changes in Bodyfat



# Data

```
library(BAS)
data(bodyfat)  #from BAS  help(bodyfat)
dim(bodyfat)
```

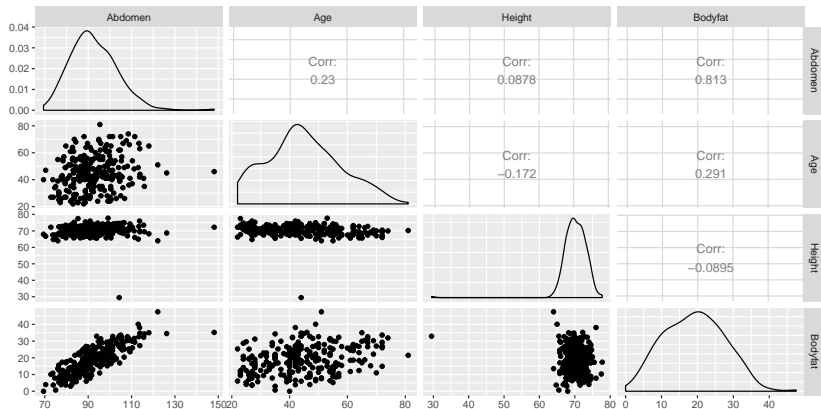
```
## [1] 252  15
```

```
summary(bodyfat)  # anything strange ?
```

##	Density	Bodyfat	Age	Weight
##	Min. :0.995	Min. : 0.00	Min. :22.00	Min. :112.50
##	1st Qu.:1.041	1st Qu.:12.47	1st Qu.:35.75	1st Qu.:152.50
##	Median :1.055	Median :19.20	Median :43.00	Median :168.00
##	Mean :1.056	Mean :19.15	Mean :44.88	Mean :171.28
##	3rd Qu.:1.070	3rd Qu.:25.30	3rd Qu.:54.00	3rd Qu.:183.00
##	Max. :1.109	Max. :47.50	Max. :81.00	Max. :279.00
##	Height	Neck	Chest	Abdomen
##	Min. :29.50	Min. :31.10	Min. : 79.30	Min. :94.50
##	1st Qu.:68.25	1st Qu.:36.40	1st Qu.: 94.35	1st Qu.:104.00

# Pairs Plots

```
library(GGally)
ggpairs(bodyfat, columns=c(8,3,5,2))
```



# Ordinary Least Squares

- ▶ OLS estimates of parameters  $\beta_0$  and  $\beta$  minimize sum of squared errors

$$\sum_{i=1}^n (Y_i - (\beta_0 + X_i \beta_1))^2$$

$$L(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta)$$

- ▶ OLS estimate of  $\beta$

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

- ▶ Ad Hoc
- ▶ Equivalent to Maximum Likelihood Estimates with assumption that errors are iid Normal (Model based)

# Summarizing Model Fit

- ▶ Fitted values

$$\hat{Y}_i = x_i^T \hat{\beta}$$

- ▶ Residuals (estimates of errors)

$$e_i = Y_i - \hat{Y}_i = \hat{e}_i$$

- ▶ Sum of Squared Errors

$$SSE = \sum e_i^2$$

- ▶ measures remaining residual variation in response
- ▶  $MSE = SSE/(n - 2)$  (or more generally  $n - p$ ) is an estimate of  $\sigma^2$
- ▶ degrees of freedom  $n - p$

## Fitting Models in R

```
bodyfat.lm = lm(Bodyfat ~ Abdomen, data=bodyfat)
summary(bodyfat.lm)    #summary of regression output
```

```
##
```

```
## Call:
```

```
## lm(formula = Bodyfat ~ Abdomen, data = bodyfat)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -19.0160  -3.7557   0.0554   3.4215  12.9007
```

```
##
```

```
## Coefficients:
```

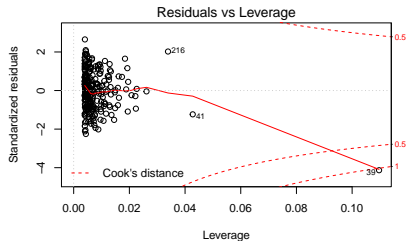
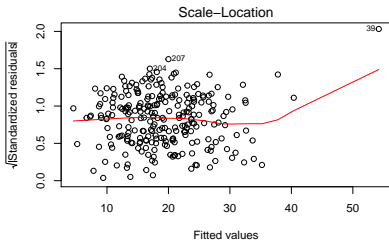
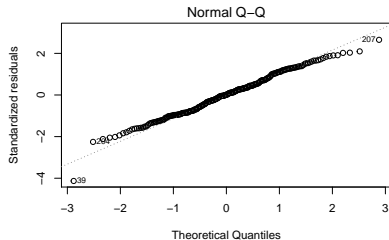
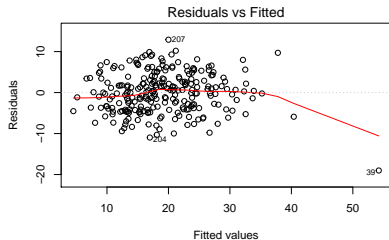
```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -39.28018    2.66034  -14.77  <2e-16 ***
## Abdomen      0.63130     0.02855   22.11  <2e-16 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

```
##
```

# Residual Plots



# Diagnostic Plots

- ▶ Residuals versus fitted values
- ▶ Normal Quantile: check normality of residuals or look for heavier tails than normal where observed quantiles are larger than expected under normality
- ▶ Scale-Location plot:  
Detect if the spread of the residuals is constant over the range of fitted values. (Constant variance with mean)
- ▶ standardized residuals versus leverage with contours of Cook's distance: shows influential points where points greater than 1 or  $4/n$  are considered influential

Case 39 appears to be influential!

# Hat Matrix

- ▶ predictions

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1}X^T Y$$

$$H = X(X^T X)^{-1}X^T$$

- ▶ Hat Matrix or Projection Matrix

- ▶ idempotent  $HH = H$
- ▶ symmetric
- ▶ leverage values are the diagonal elements  $h_{ii}$

$$\hat{Y}_i = h_{ii}Y_i + \sum_{i \neq j} h_{ij}Y_j$$

$$0 \leq h_{ii} \leq 1$$

- ▶ leverage values near 1 imply  $\hat{Y}_i = Y_i$
- ▶ potentially influential
- ▶ measure of how far  $x_i$  is from center of data



# Residual Analysis

- ▶ residuals

$$e = Y - \hat{Y} = (I - H)Y$$

$$\text{var}(e_i) = \hat{\sigma}^2(1 - h_{ii})$$

- ▶ Standardize:

$$r_i = e_i / \sqrt{\text{var}(e_i)}$$

- ▶ if leverage is near 1 then residual is near 0 and variance is near 0 and  $r_i$  is approximately 0 (may not be helpful)

## Cook's Distance

Measure of influence of case  $i$  on predictions

$$D_i = \frac{\|Y - \hat{Y}_{(i)}\|^2}{\hat{\sigma}^2 p}$$

after removing the  $i$ th case

Easier way to calculate

$$D_i = \frac{e_i^2}{\hat{\sigma}^2 p} \left[ \frac{h_{ii}}{(1 - h_{ii})^2} \right],$$

$$D_i = \frac{r_{ii}}{p} \frac{h_{ii}}{1 - h_{ii}}$$

# Model Assessment

- ▶ Always look at residual plots!
- ▶ Check constant variance, outliers, influence, normality assumption
- ▶ Treat  $e_i$  as 'new data' -- look at structure, other predictors & plots
- ▶ Case 39 looks influential!

How should we proceed?