生物统计（2）回归模型
Chapter 18

俞章盛

教授　生命学院生物信息与生物统计系
客座教授　耶鲁大学生物统计系

Outline

- Regression concept
- Simple linear regression model
    - Formulation of the model
    - Estimation of regression coefficients
    - Inference of regression coefficients
    - Inference for predicted values
    - Model evaluation
- Extension

Linear regression concept

- Correlation coefficient tells us the magnitude at which two random variables are linearly associated with each other; it does not tell us how change in one variable impact the value of the other one
- Linear regression seeks to identify the linear functional form (intercept and slope) between the mean of one variable (***response variable, dependent variable*** or ***outcome variable***) and any fixed value of the other variable (***explanatory variable***, ***independent variable***, ***covariate, predictor*** or ***regressor***)
- The ultimate objective is
  - Assess how change in the predictor impact value of the response.
  - Estimate or predict the response that is associated with a fixed value of the predictor.

Representation of a line

A line can be represented as

$$y = \alpha + \beta x$$

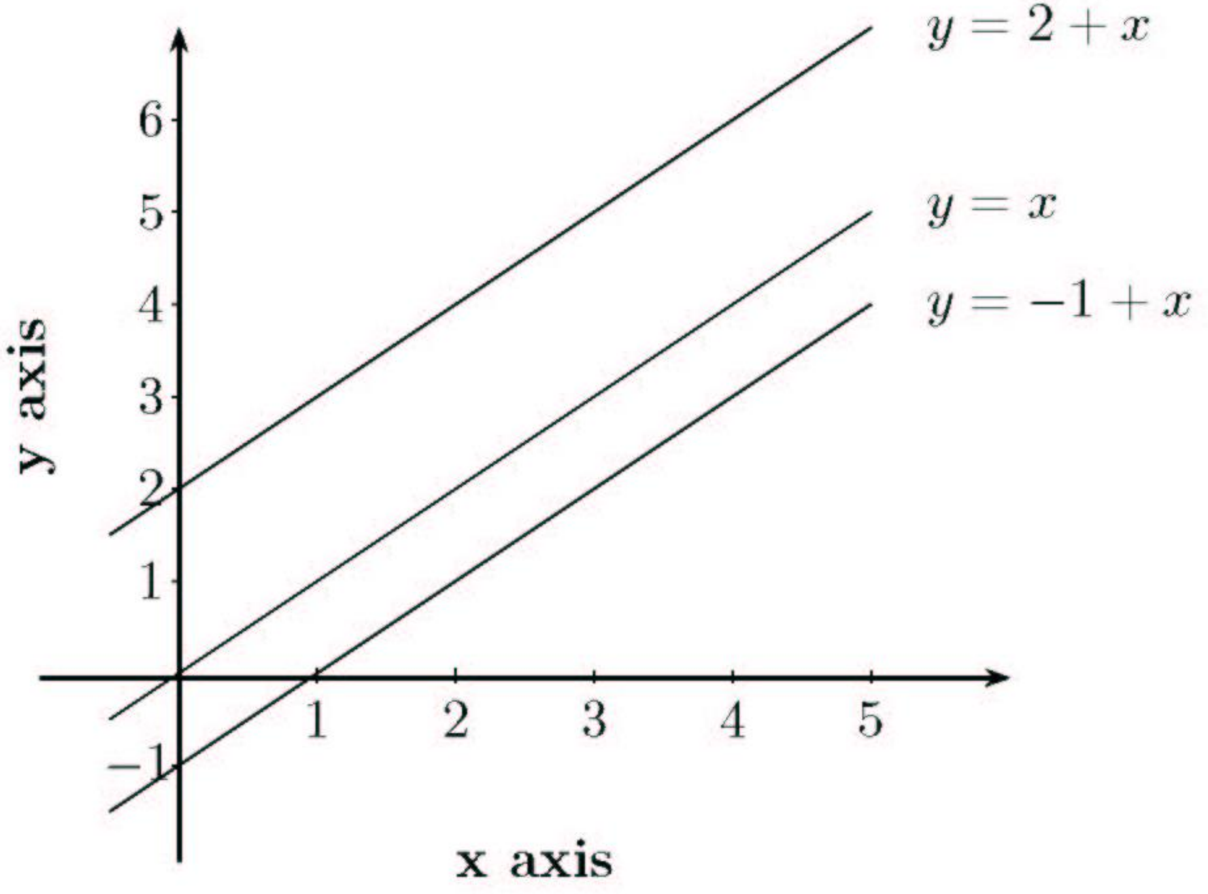
- x and y correspond to the coordinates on X axis and Y axis, respectively
- α is called the **intercept**; α is the value of y when $x = 0$
- β is called the **slope**:

  * When $\beta > 0$ or positive slope, y increases as x increases. For every unit increase in x, the increase in y is β.

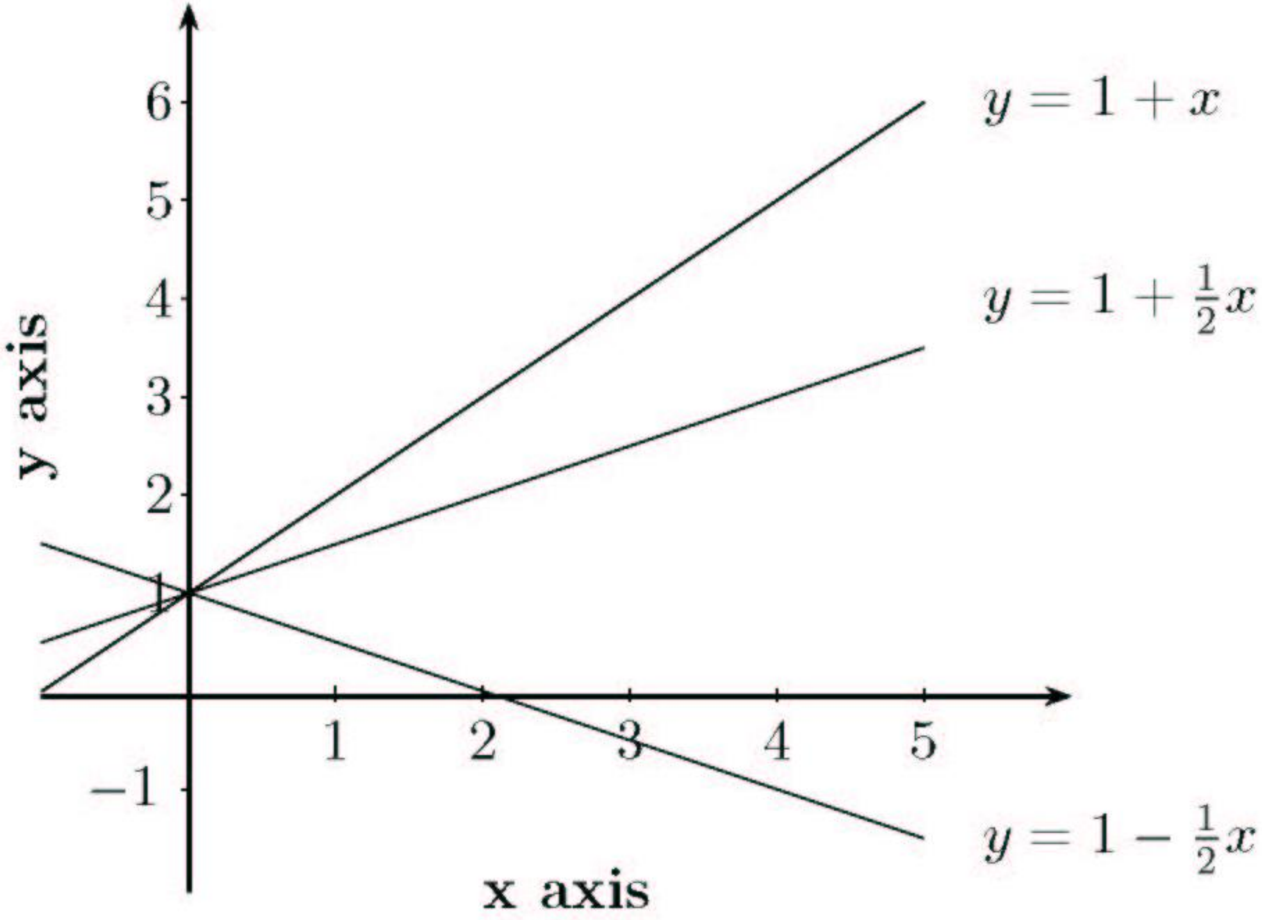
  * When $\beta < 0$ or negative slope, y decreases as x increases. For every unit increase in x, the decrease in y is -β.

  * When $\beta = 0 \Rightarrow$ a line parallel to X axis and y is a constant that does not change as x ranges over all possible values

Lines with different intercept and same slope



Lines with same intercept and different s



The model

Consider dependent variable Y and independent variable X.

We assume that

- Given any fixed value x of X, Y follows a normal distribution with mean $\mu_{y|x}$ and variance $\sigma_{y|x}^2$,
- $\mu_{y|x} = \alpha + \beta x \Rightarrow$ the mean of Y is linearly associated the values of X (α and β are called **regression coefficients**)
- $\sigma_{y|x}^2 = \sigma^2$ is a constant $\Rightarrow$ the variance of Y given any fixed value x is the same (**homoscedasicity**)

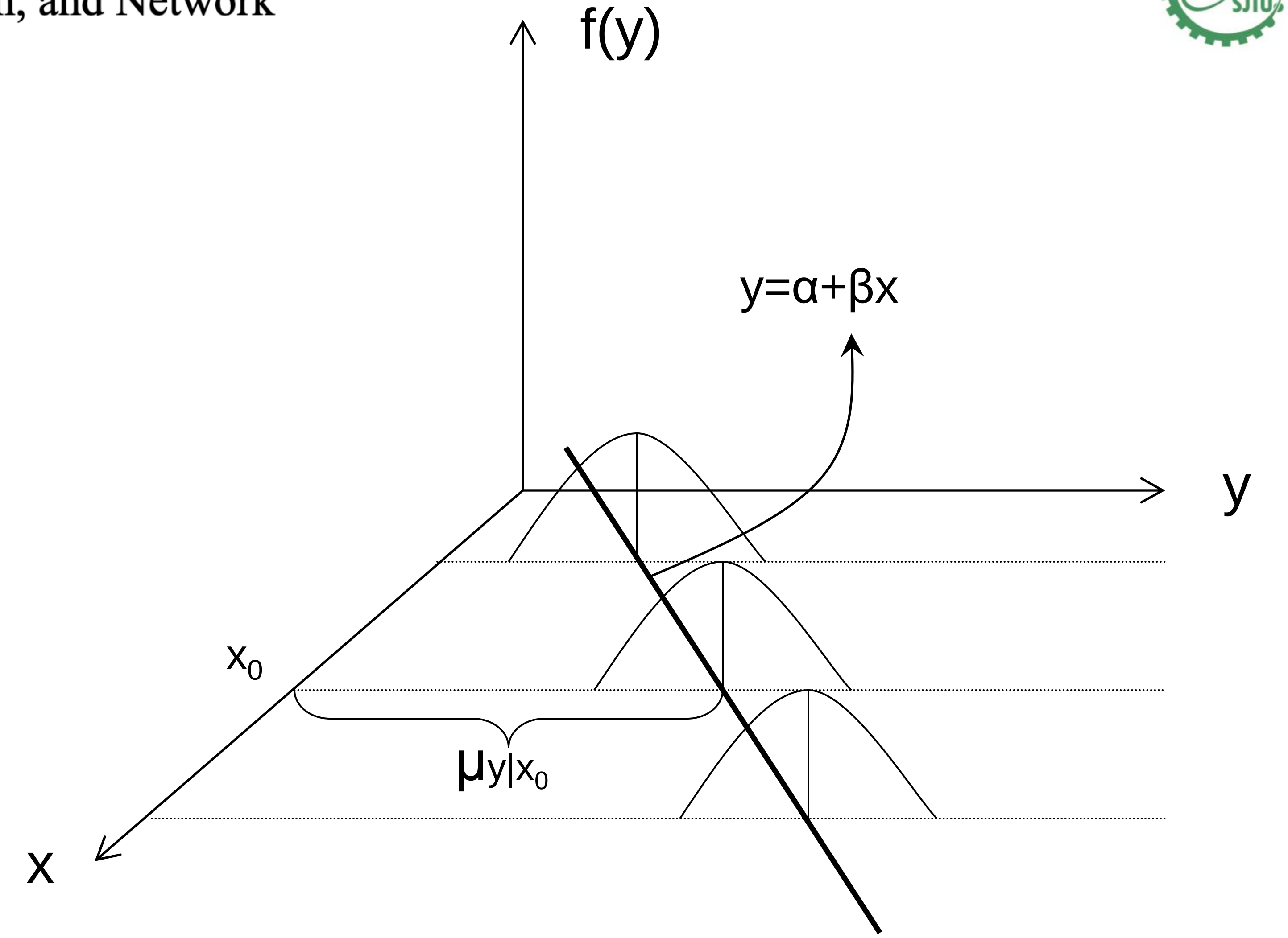
The model cont.

Now, let $(x_i, y_i),\ i = 1, 2, \ldots, n$, be a random sample. Based on the assumption just proposed, we have

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

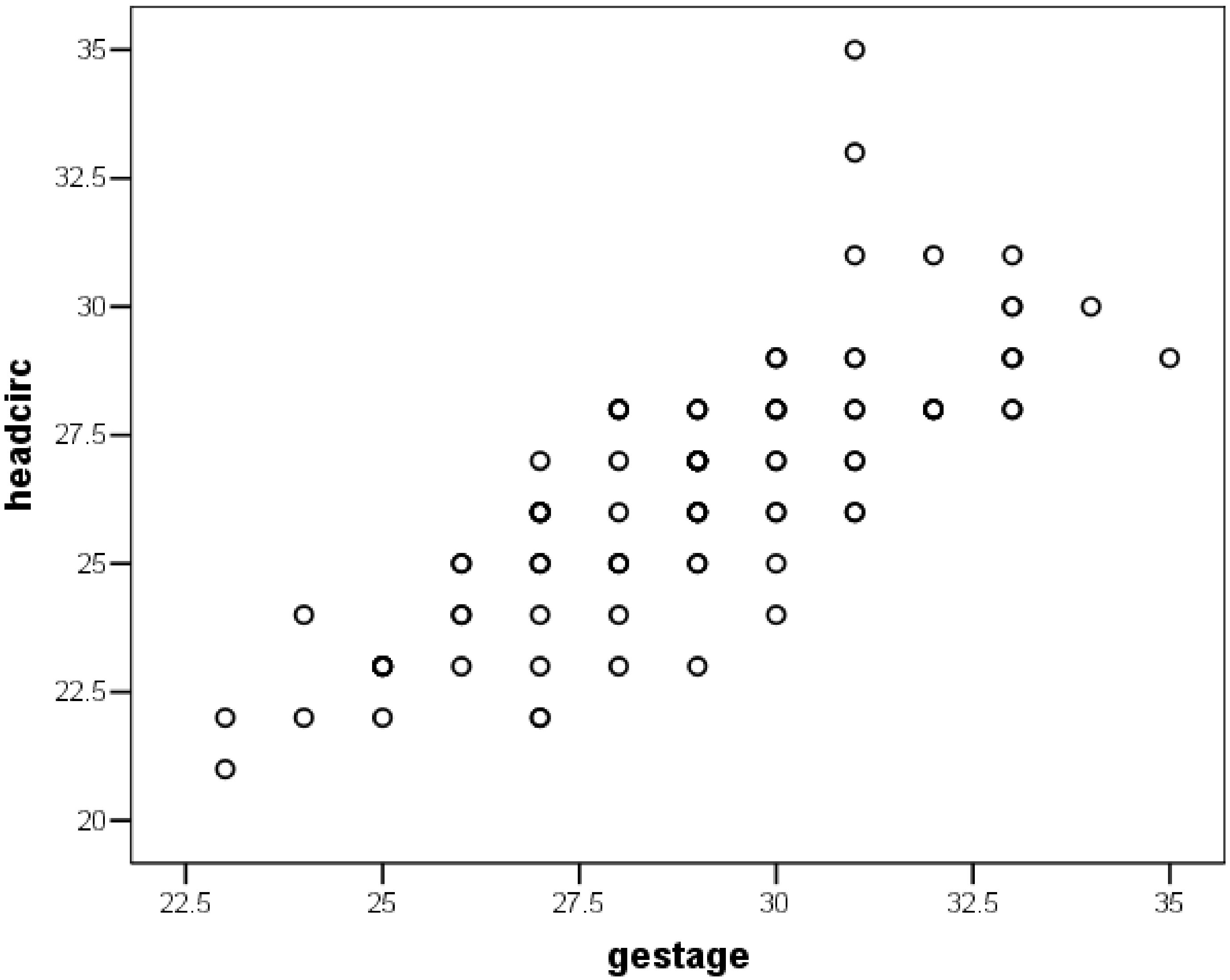
where ε_i is an error term that accounts for variability from what is expected from a line. In accordance with previous assumptions, the error terms is assumed to have a normal distribution with mean 0 and variance $\sigma_{y|x}^2 = \sigma^2$.

We also have to assume y_i are independent of y_j for different i and j.

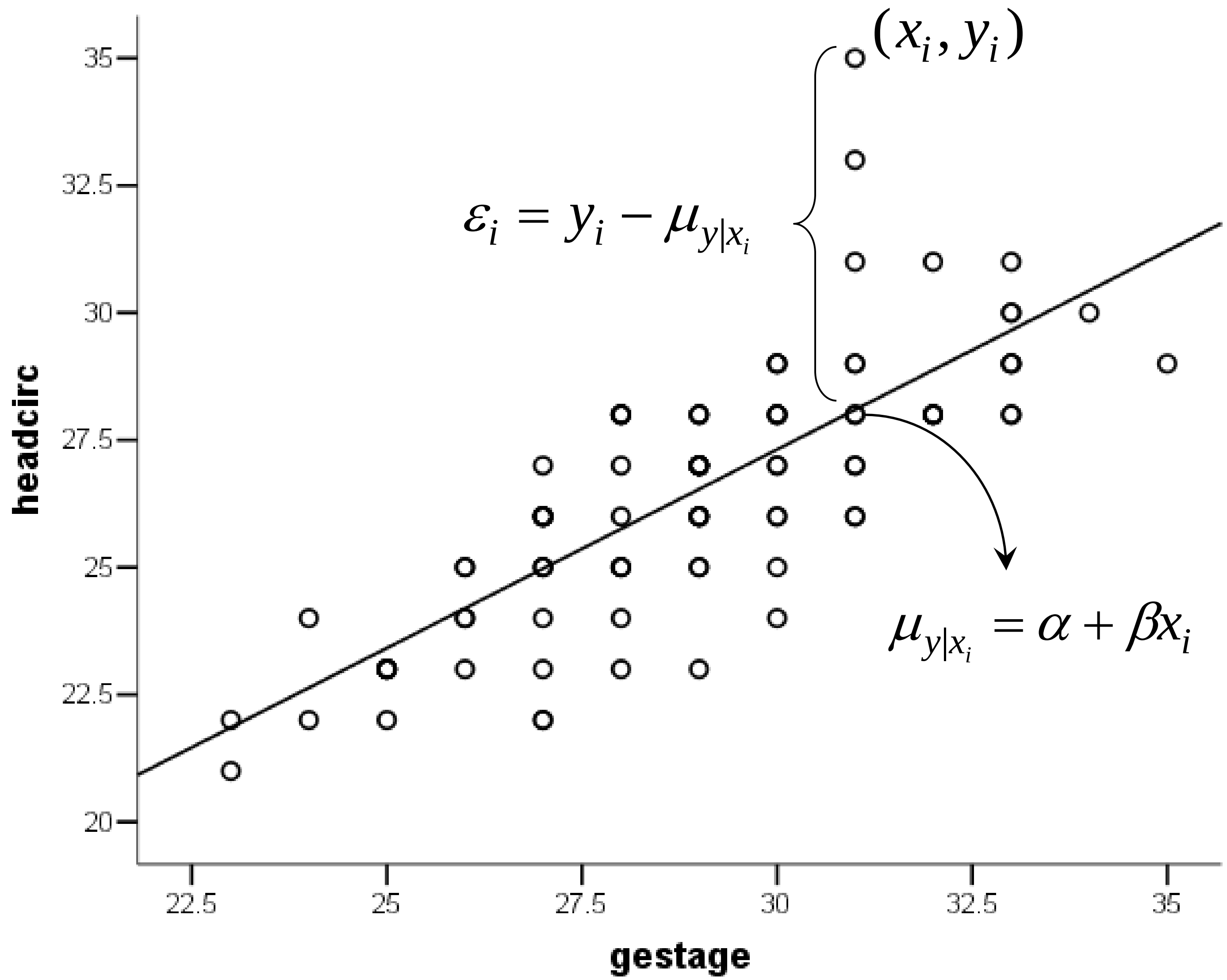

f(y)
y=α+βx
y
x₀
μy|x₀
x

Head circumference example

- Among children of both sexes, head circumference appears to increase linearly **with gestational age**.
- Head circumference is the outcome variable and **gestational** age is the independent variable
- **An understanding of their relationship helps parents and pediatricians to monitor growth and detect possible cases of macrocephaly and microcephaly**
- A sample of 100 low birth weight infants born in Boston is available for analysis
- Mean(head circumference)=$\alpha+\beta\times$gestational age

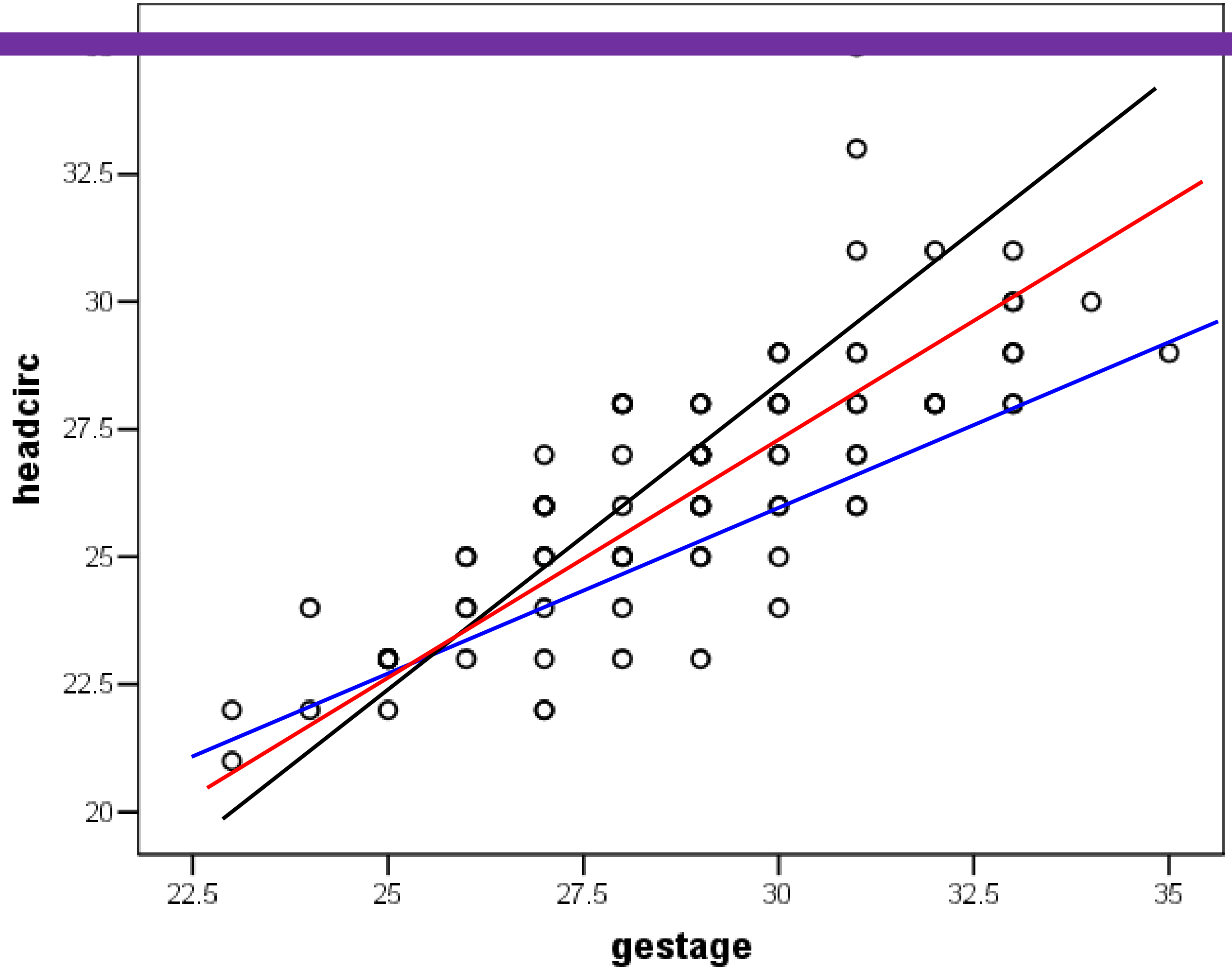


headcirc
35
32.5
30
27.5
25
22.5
20
22.5
25
27.5
30
32.5
35
gestage



(x_i, y_i)
$\varepsilon_i = y_i - \mu_{y|x_i}$
$\mu_{y|x_i} = \alpha + \beta x_i$
headcirc
35
32.5
30
27.5
25
22.5
20
22.5
25
27.5
30
32.5
35
gestage

Which line is better?



Least Square Estimate

To fit a line to data $(x_i, y_i),\ i = 1,\ 2,\ \ldots n,$ we would like that the points are as close to the line as possible. Obviously, it is impossible to find a line that passes every point. Therefore, we need to have some criteria on what is the best line in terms of the distance between the data points and a fitted line. One criteria is the least squares criteria and the associated method in finding the line is called **method of least squares**.

Let $\hat{y}_i = \alpha + \beta x_i$ be the value on the fitted line corresponding to x_i, then define the **residual** e_i as

$$e_i = y_i - \hat{y}_i.$$

Least Square Estimate

Intuitively, we want to fit a line that makes the residuals as small as possible. The method of least squares seeks a line that minimizes the sum of the squares of the residuals, or the **error sum of squares** (SSE).

Specifically, we try to find $(\hat{\alpha}, \hat{\beta})$ such that

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

is minimized.

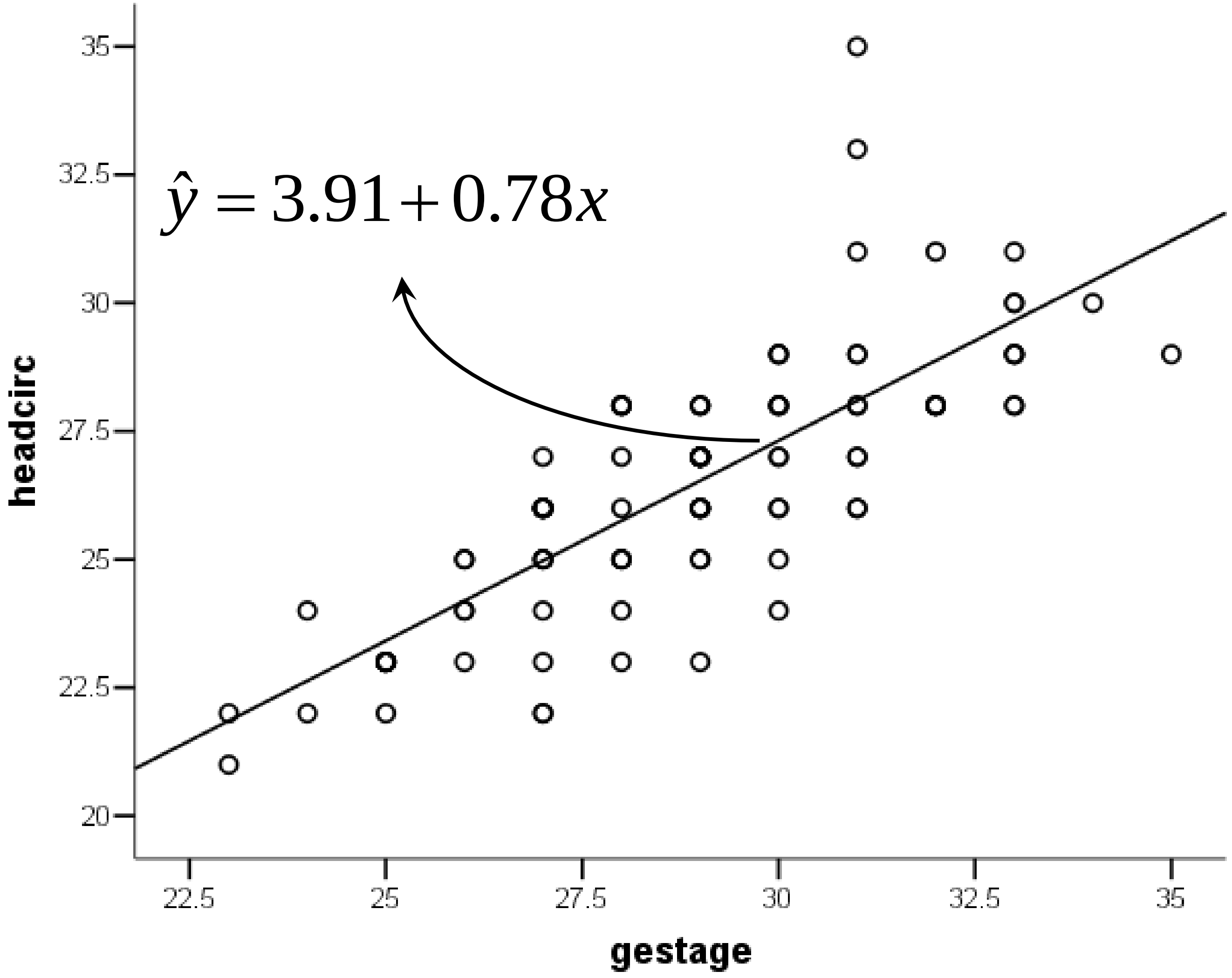
Methods of least squares cont.

The estimates of α and β based on method of least squares are

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

with

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ and } \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$


$\hat{y} = 3.91 + 0.78x$
headcirc
gestage
35
32.5
30
27.5
25
22.5
20
22.5
25
27.5
30
32.5
35

Inference for regression coefficients cont.

It can be shown mathematically that

$$\mathrm{se}(\hat{\beta}) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \quad \text{and} \quad \mathrm{se}(\hat{\alpha}) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Usually, σ, the standard deviation of the error term, is not known. Hence, we need to estimate it by the standard deviation of the residuals:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{n-2}}$$

Therefore,

$$\mathrm{se}^*(\hat{\beta}) = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \quad \text{and} \quad \mathrm{se}^*(\hat{\alpha}) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Inference for regression coefficients cont.

Inference of β

- The slope is usually the more important coefficient in that it quantifies the average change in y that correspond each one-unit change in x. In particular, $\beta = 0$ implies that there is no linear relationship between x and y; the mean value of y is the same regardless of the value of x.
- Hypothesis testing:

$$H_0 : \beta = \beta_0 \text{ versus } H_A : \beta \neq \beta_0$$

Under the null hypothesis, $T = \dfrac{\hat{\beta} - \beta_0}{\mathrm{se}^*(\hat{\beta})}$ has a t distribution with n-2 degrees of freedom. Hence, reject the null hypothesis when $|T| > t_{n-2,\alpha/2}$.

- $(1-\alpha)\times 100\%$ CI for β:

$$(\hat{\beta} - t_{n-2,\alpha/2}\mathrm{se}^*(\hat{\beta}),\ \hat{\beta} + t_{n-2,\alpha/2}\mathrm{se}^*(\hat{\beta}))$$

Matrix Derivation of LSE

$$\underline{Y} = (y_1, y_2, \cdots, y_n)^T \qquad X^T = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & & x_{np} \end{pmatrix} \qquad \underline{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$$

$$\underline{Y} = X^T \underline{\beta} + \underline{\varepsilon} \qquad \underline{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$$

$$\text{LSE } \underline{\hat{\beta}} = \underset{\underline{\beta}}{\arg\min} \underbrace{(Y - X^T\beta)^T (Y - X^T\beta)}_{LSE}$$

$$\frac{\partial LSE(\beta)}{\partial \underline{\beta}} = 2X(Y - X^T\beta)^T = 2X\underline{Y} - (XX^T)\underline{\beta} = 0$$

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

$$\text{If } X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T Y$$

$$\mathrm{Cov}(\hat{\beta}) \stackrel{?}{=}$$

Matrix Derivation of LSE

Vector derivative

$$f(x)=x^TB \xrightarrow{df(x)/dx} = B$$

$$f(x)=x^Tb \longrightarrow = b$$

$$f(x)=x^Tx \longrightarrow 2x$$

$$f(x)=x^TBx \longrightarrow 2Bx$$

$$f(x)=f(b(x)) \longrightarrow f'(b(x))\,\cancel{b}\,\frac{db(x)}{dx}$$

$$= b(x)^Tb(x) \longrightarrow 2b(x)\frac{db(x)}{dx}$$

Inference for regression coefficients cont.

Inference of α

- Hypothesis testing :

$$H_0: \alpha = \alpha_0 \ \text{ versus } H_A: \alpha \neq \alpha_0$$

Under the null hypothesis, $T = \dfrac{\hat{\alpha} - \alpha_0}{\text{se}^*(\hat{\alpha})}$ has a t distribution with n-2 degrees of freedom. Hence,

reject the null hypothesis when $|T| > t_{n-2,\alpha/2}$.

- $(1 - \alpha) \times 100\%$ CI for α :

$$(\hat{\alpha} - t_{n-2,\alpha/2}\text{se}^*(\hat{\alpha}), \hat{\alpha} + t_{n-2,\alpha/2}\text{se}^*(\hat{\alpha}))$$

Remarks

- Recall that α is the mean value of y corresponding to $x = 0$. The x_i's are far away from 0 under many situations and x is not even allowed to be 0 sometimes. Therefore, the interpretation of such an coefficient is not very meaningful under such circumstances. In practice, we usually use a centered version of x_i in the regression analysis. That is, create

$$x_i^c = x_i - \bar{x},$$

and fit a regression line $y = \alpha^c + \beta x$ for (x_i^c, y_i). Here, β has exactly the same interpretation as before and α^c is the mean value of y associated with mean value $\bar{x}$.

- A test of $H_0 : \beta = 0$ is equivalent to the test of $H_0 : \rho = 0$, where ρ is the population correlation coefficient between x and y.

Head circumference example cont.

$$s = 1.59 \qquad s_x = 2.534 \quad \beta = 0.78$$

$$H_0 : \beta = 0 \quad \text{vs. Ha: } \beta \neq 0$$

$$\text{se}^*(\hat{\beta}) = s / \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} = 1.59 / \sqrt{99 \times 2.534^2} = 0.063$$

$$T = \frac{0.78}{0.063} = 12.36$$

For a $t_{98}(0.025)$=1.98. Since T>1.98, we reject the null hypotheses *at* 5% significance level and conclude that with each unit increase of gestational age, there is a significant change (increase) of the mean head circumference.

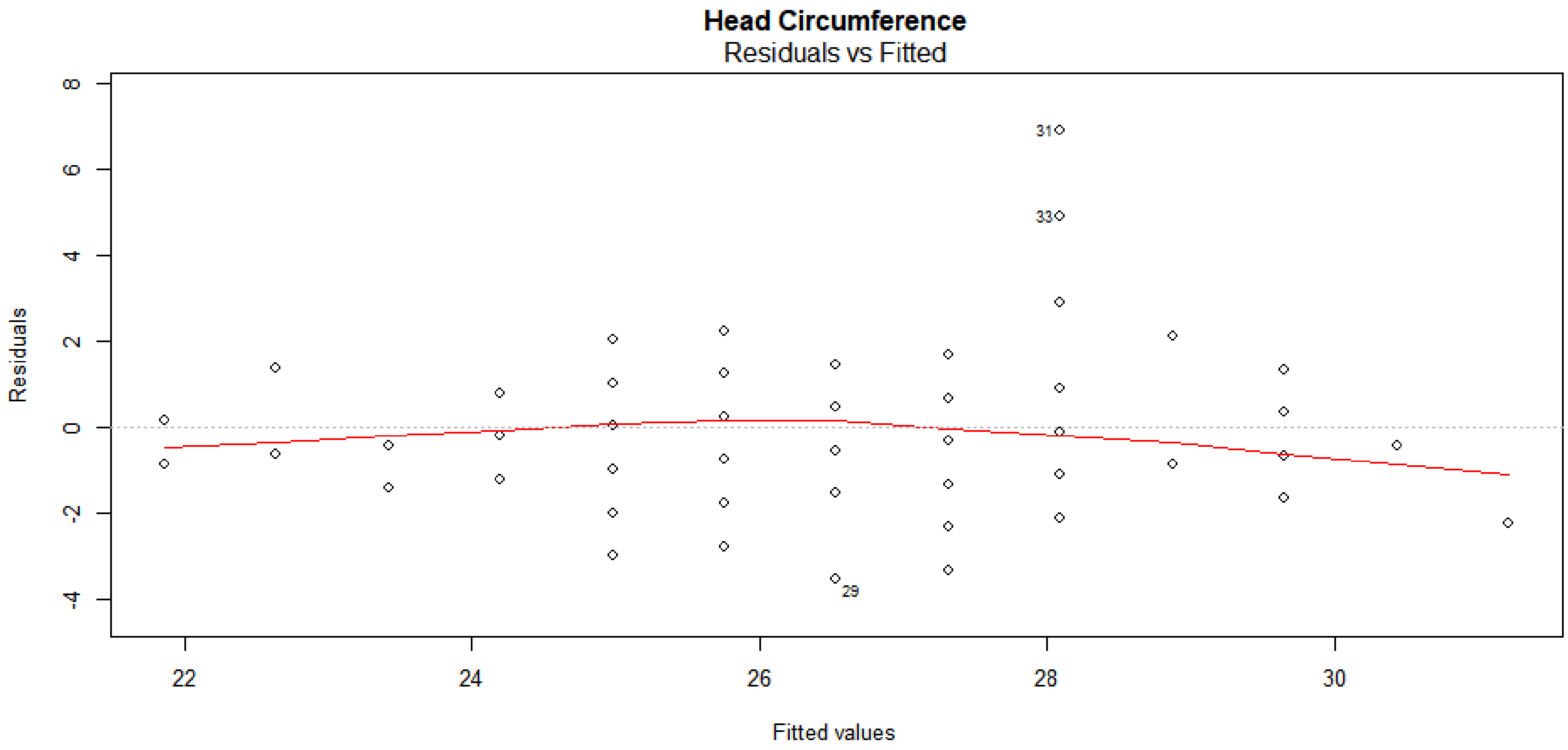
95% CI: $(0.78\text{-}1.98(0.063), 0.78 + 1.98(0.063)) = (0.656, 0.904)$

What is we want to test H0: beta=1?

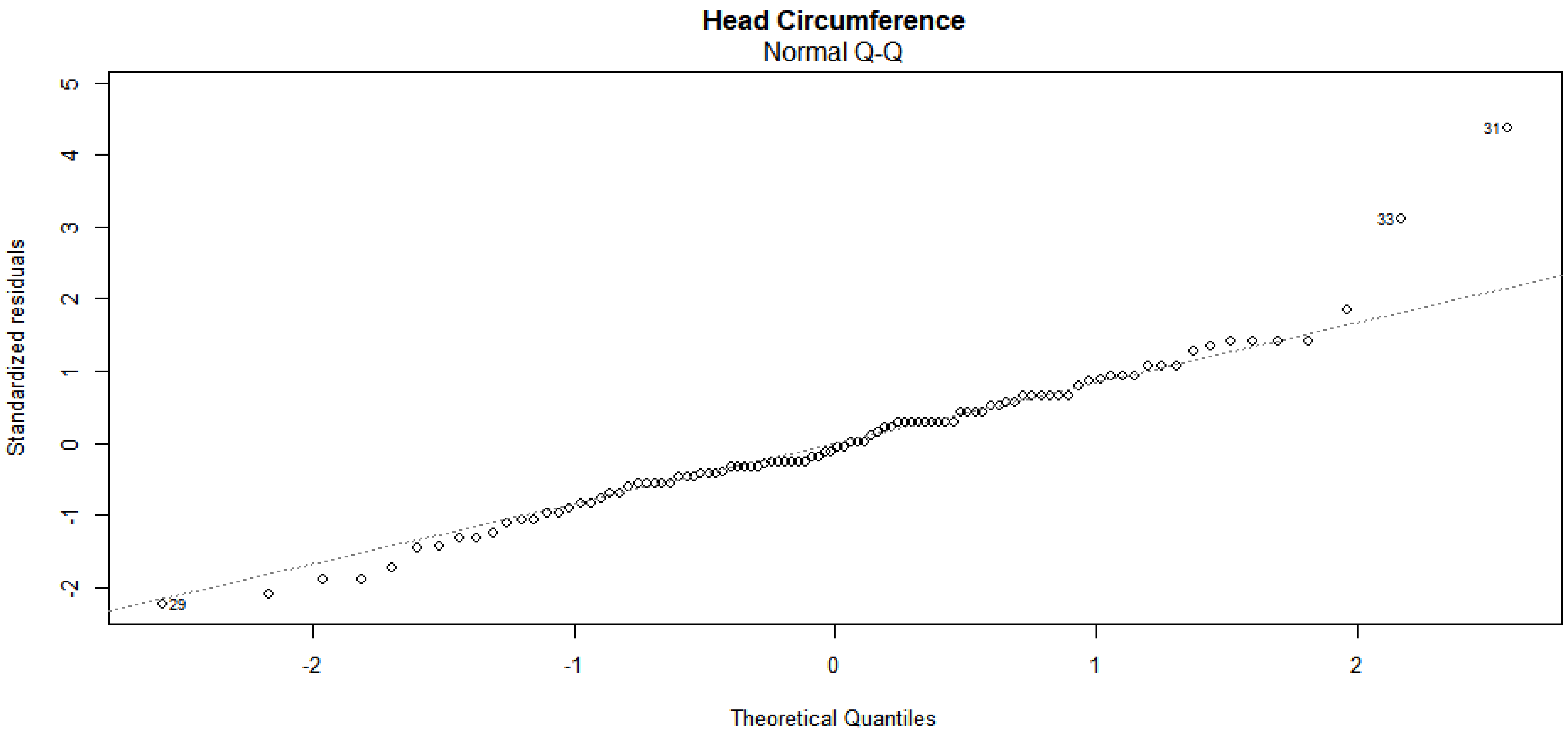
Model Diagnosis—Goodness of fit

```
> fit_lbwi<-lm(headcirc~gestage,data=data_lbwi)
> summary(fit_lbwi)

Call:
lm(formula = headcirc ~ gestage, data = data_lbwi)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5358 -0.8760 -0.1458  0.9041  6.9041 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.91426    1.82915    2.14   0.0348 *  
gestage      0.78005    0.06307   12.37   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.59 on 98 degrees of freedom
Multiple R-squared:  0.6095,	Adjusted R-squared:  0.6055 
F-statistic: 152.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

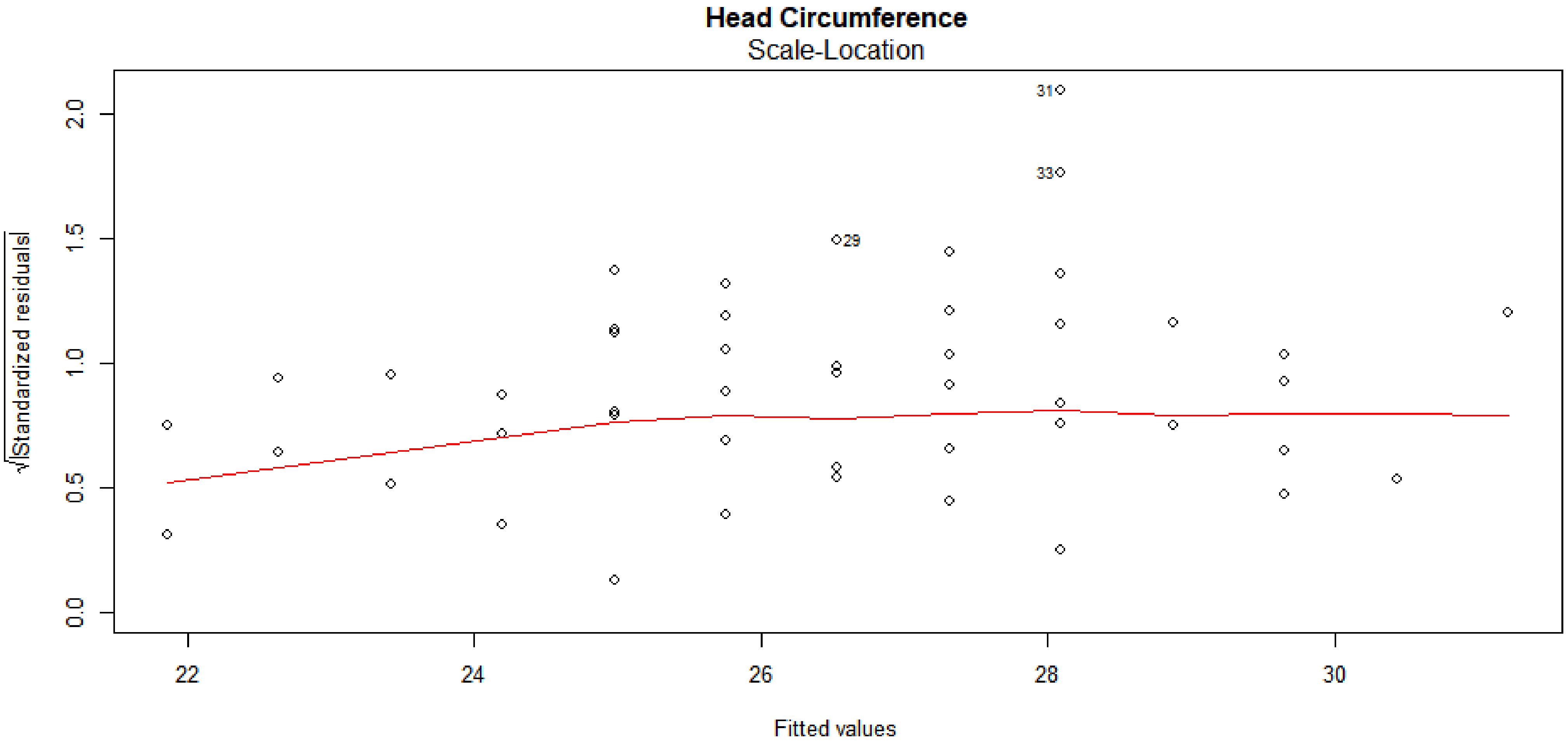
Model Diagnosis—Residual plots



Model Diagnosis—Normality



Model Diagnosis—Homogeneity



Connection to ANOVA

Simple linear regression is closely relates to the concept of ANOVA.

- For each fixed value of x, the mean value of y is $\mu_{y|x} = \alpha + \beta x$. Therefore, the null hypothesis that $\beta = 0$ is equivalent to saying that these infinite number of populations have the same mean.
- The within group variation in the simple linear regression setting is measured by mean squares of error

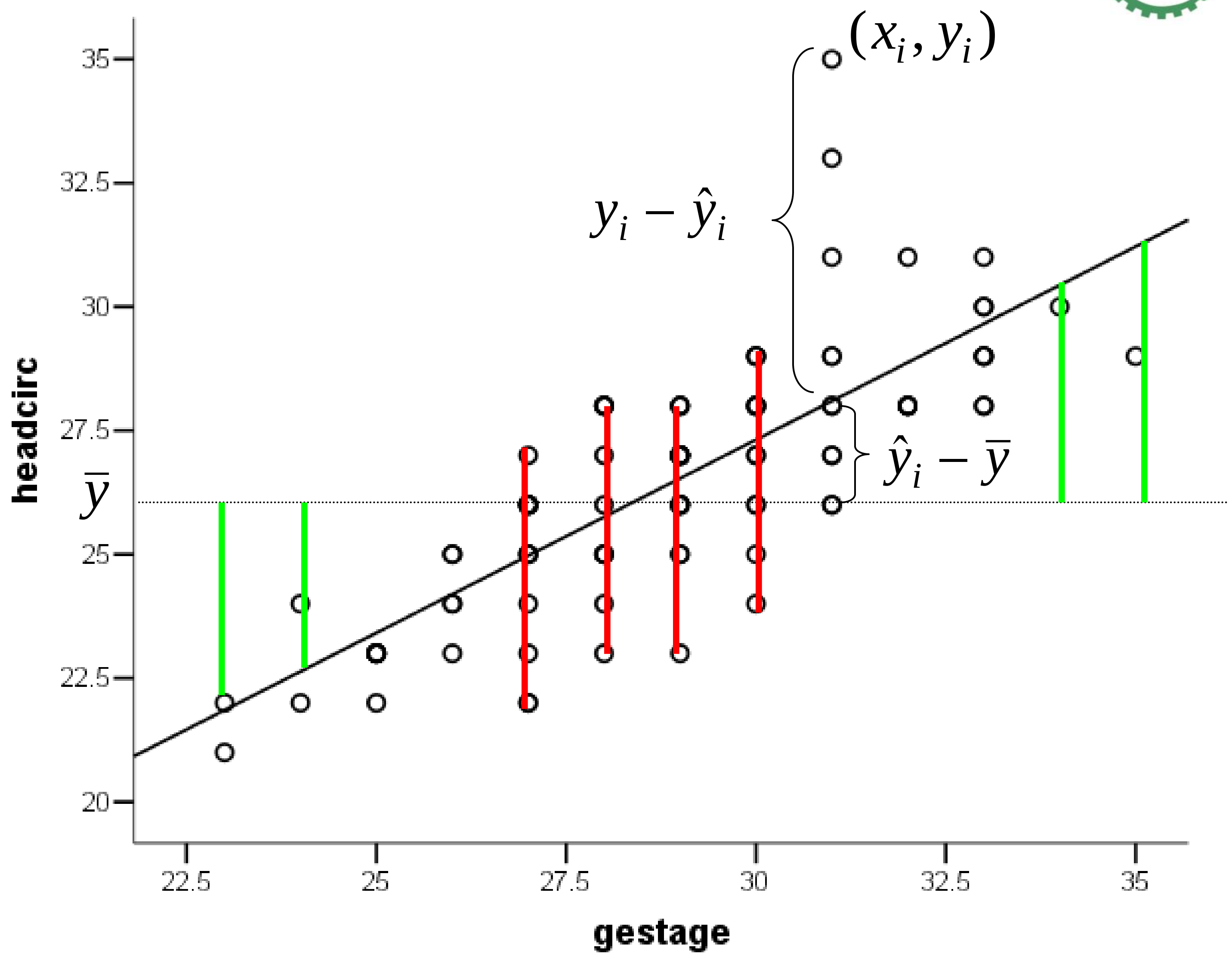
$$\text{MSE} = \frac{\text{SSE}}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = s^2.$$

Connection to ANOVA

- The between group variation in the simple linear regression setting is measured by mean squares due to regression

$$\text{MSR} = \frac{\text{sum of squares due to regression}}{1} = \frac{\text{SSR}}{1} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

- The statistic $F = \dfrac{\text{MSR}}{\text{MSE}}$ follows an F distribution with 1 and n-2 degrees of freedom under the null hypothesis that the infinite number of populations have the same mean. The F test is equivalent to the t test of $\beta = 0$.


(x_i, y_i)
$y_i - \hat{y}_i$
$\hat{y}_i - \bar{y}$
$\bar{y}$
headcirc
gestage
35
32.5
30
27.5
25
22.5
20
22.5
25
27.5
30
32.5
35

ANOVA table

Source of variation	Sum of Squares (SS)	*df*	Mean Squares (MS)	F Statistic	P-value
Regression	$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	1	$MSR=\frac{SSR}{1}$	$F=\frac{MSR}{MSE}$	$P(F>\text{calculated } F)$
Residuals (Errors)	$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	*n*-2	$MSE=\frac{SSE}{n-2}$	$H_0: \beta=0$	
Total	$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2$	*n*-1	s^2		

Multiple Linear Regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i .$$

with the same distribution assumption for the noise term.

- Remember $\hat{\beta} = (XX^T)^{-1}(X^T Y)$
- Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related.
- If the XX^T is not of full rank(collinear), the $\hat{\beta}$ is not computable.
- Even if XX^T is invertible, a high correlation among Xs will affect the standard error estimators.
- Using variance inflation factor: VIFj=$\frac{1}{1-R_j^2}$ where R_j^2 is the coefficient of determination of Xj versus other Xs.

Coefficient of Determination

- R_j^2 is obtained by regress Xj vs X1, X2, ⋯, Xp without Xj. The R^2 of the regression model is R_j^2

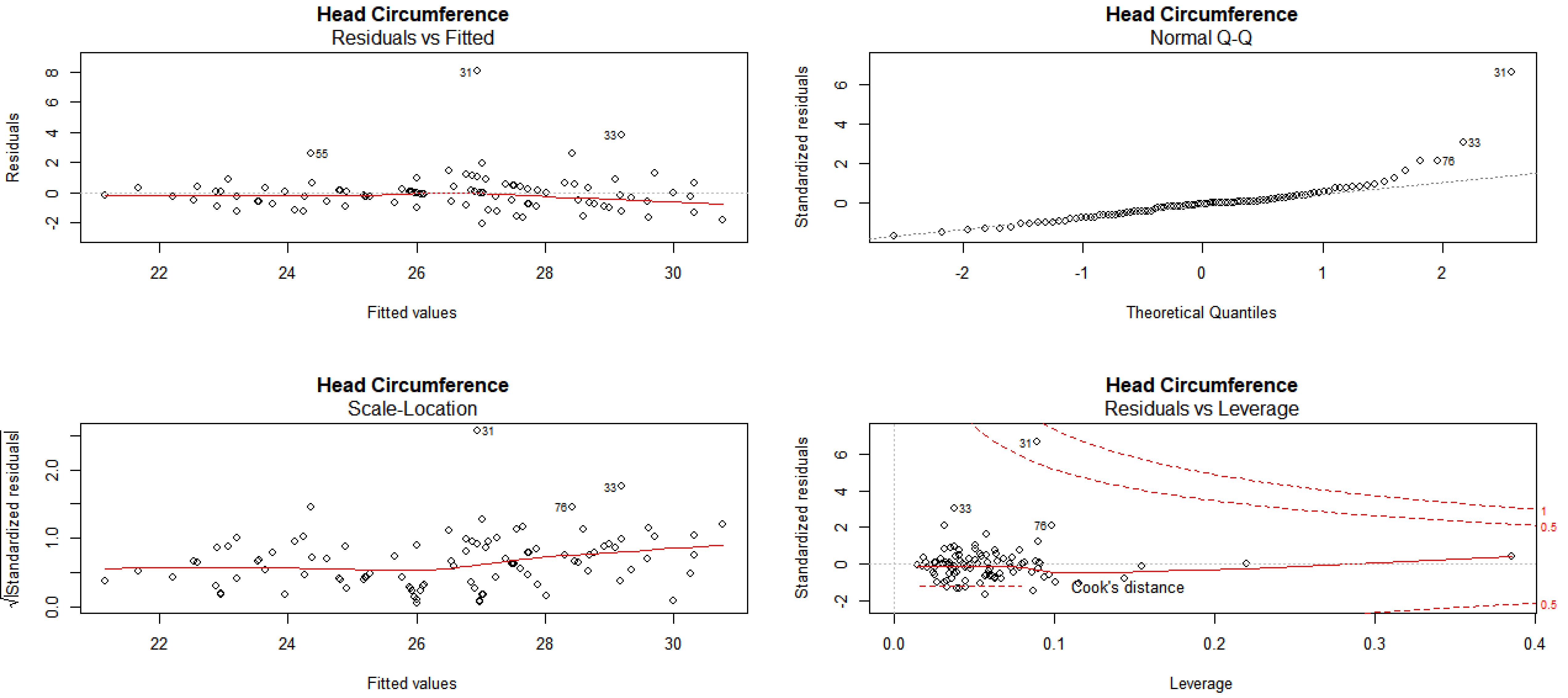
```
> summary(lm(gestage~+length+birthwt+momage+toxemia ,data=data_lbwi)) ##calculate the R-square of gestage
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.798835   2.050346   7.705 1.25e-11 ***
length       0.193789   0.077733   2.493 0.014397 *
birthwt      0.003918   0.001012   3.872 0.000198 ***
momage       0.042371   0.026949   1.572 0.119222
toxemia      2.264834   0.389897   5.809 8.34e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.558 on 95 degrees of freedom
Multiple R-squared:  0.6372,    Adjusted R-squared:  0.6219
F-statistic: 41.71 on 4 and 95 DF,  p-value: < 2.2e-16
```

Multiple Regression Model

```
> vif(data_lbwi[,2:6]) #calcualte the VIF.
  Variables      VIF
1    length 3.348036
2   gestage 2.756302
3   birthwt 3.523658
4    momage 1.087551
5   toxemia 1.407603
```

If VIF0>0, there is a problem with variance estimation of the coefficients

```
> summary(mfit_lbwi)

Call:
lm(formula = headcirc ~ gestage + length + gestage + birthwt +
    momage + toxemia, data = data_lbwi)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0190 -0.6712 -0.0364  0.3334  8.0421

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.2097216  2.1285705   3.387  0.00103 **
gestage      0.5261922  0.0835553   6.298 9.62e-09 ***
length       0.0082711  0.0653434   0.127  0.89954
birthwt      0.0042555  0.0008867   4.799 5.99e-06 ***
momage      -0.0300651  0.0222312  -1.352  0.17950
toxemia     -0.5160581  0.3696445  -1.396  0.16597
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.269 on 94 degrees of freedom
Multiple R-squared:  0.7615,    Adjusted R-squared:  0.74
```


Head Circumference
Residuals vs Fitted
Residuals
Fitted values
Head Circumference
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Head Circumference
Scale-Location
√|Standardized residuals|
Fitted values
Head Circumference
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

Model selection- stepwise regression

- Build regression model from a set of candidate predictor variables by entering and removing predictors based on p values, in a stepwise manner until there is no variable left to enter or remove any more.
- At each step, each variables will be added into the model at a time. The variable with the smallest p-value (and lower than the prespecified threshold p-value for inclusion) will be included.
- At the same time, in the new model, the p-value of all variable will be examined. If there is any variable with updated p-value larger than the threshold p-value for removal), it will be removed.

Model selection- stepwise regression based on p-value

- install.package(olsrr)
- library(olsrr)
- ols_step_both_p(mfit_lbwi,pent = 0.2, prem = 0.15, progress = TRUE, details = FALSE)
- stsel=ols_step_both_p(mfit_lbwi,pent = 0.2, prem = 0.15, progress = TRUE, details = FALSE)
- plot(stsel)

All possible models, best subset

- ##All possible models
- ols_step_all_possible(mfit_lbwi,pent = 0.2, prem = 0.15, progress = TRUE, details = FALSE)
- allpos=ols_step_all_possible(mfit_lbwi,pent = 0.2, prem = 0.15, progress = TRUE, details = FALSE)
- plot(allpos)

- ##Best subset
- ols_step_best_subset(mfit_lbwi,pent = 0.2, prem = 0.15, progress = TRUE, details = FALSE)
- bestsubset=ols_step_best_subset(mfit_lbwi,pent = 0.2, prem = 0.15, progress = TRUE, details = FALSE)
- plot(bestsubset)

Model selection- LASSO

- Bias-variance tradeoff
- Ridge regression
- LASSO(Least Absolute Shrinkage and Selection Operator)

Model selection- LASSO

For the OLS, we obtain $\hat{\beta}$ by minimizing $\sum (Y_i - x_i^T\beta)^2$

Prediction error (PE): If we have a new observation,

The estimated $\hat{Y}_i$ should be close to new Y_i

$$PE(x_0) = E_{Y|x=x_0}[(Y-\hat{f}(x))^2 \mid X=x_0]$$

$$= E_{Y|x=x_0}[f(x_0)+\varepsilon - E(\hat{f}(x_0)) + E[\hat{f}(x_0)] - \hat{f}(x_0)]^2$$

$$= E_{Y|x=x_0}(\varepsilon^2) + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2 + E[f(x_0) - E\hat{f}(x_0)]^2$$

$$= \sigma_\varepsilon^2 + Var(\hat{f}(x_0)) + [bias(\hat{f}(x_0))]^2$$

① Introduce a little bias may decrease the variance.

Model selection- LASSO

If # of predictor is large, we sometime may have difficulty obtaining $(X^TX)^{-1}$.

One way is to introduce "penalty

$$\sum_{i=1}^{n}(y_i - \underbrace{\cancel{f(x)}}_{x^T\beta})^2 - \lambda\sum_{j}^{p}\beta_j^2$$

The solution $\hat{\beta}_{PLS} = (X^TX)^{-1}(X^TY)$

$$\hat{\beta}_{PLS} = (X^TX + \lambda I_p)^{-1}X^TY$$

This is call ridge regression.

λ is called the tuning parameter

Is $\hat{\beta}_{PLS}$ biased?

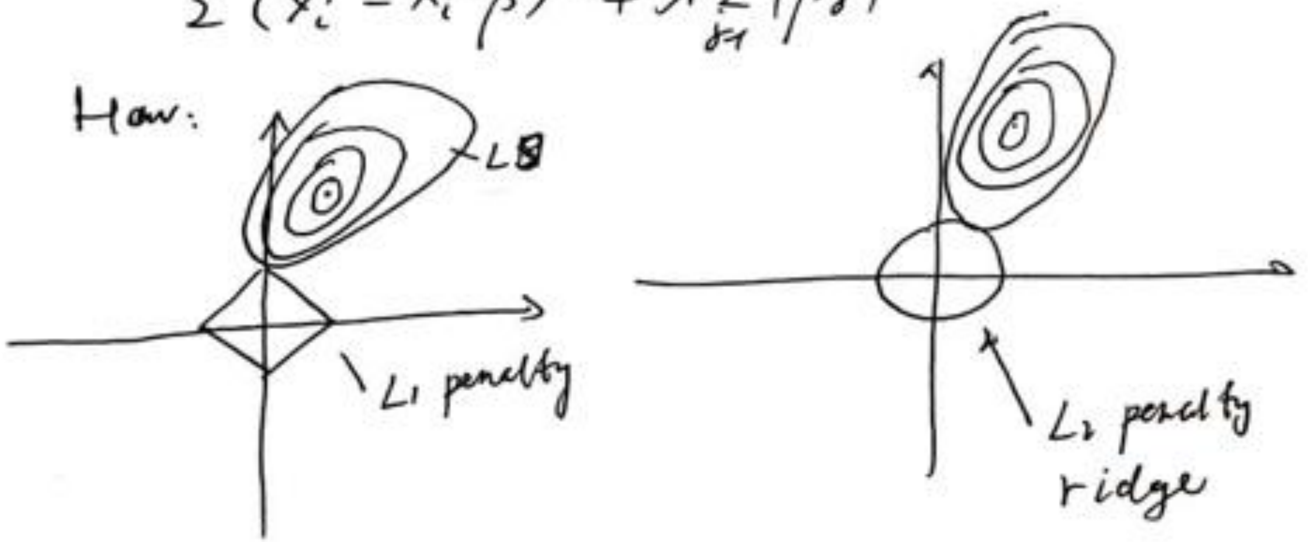
Model selection- LASSO

LASSO: (Least absolute shrinkage and selection operator)

When p is large, and there are some X_j has no effect on Y, we would like to kick out those by

$$\sum (Y_i - x_i^T\beta)^2 + \lambda \sum_{j=1}^{k} |\beta_j|$$

How:




```
library(glmnet)
las=glmnet(data_lbwi[,2:6],data_lbwi[,1], family = c("gaussian"))
```

Interactions

$$y_i = \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon_i$$

$$y_i = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_1 X_2\beta_3 + \varepsilon_i$$

If β_3 is significant, then there is an interaction effect between X_1, X_2.

If $X_1 \in \{0, 1\}$ $X_2 \in \{0, 1\}$

$$\beta_3 = \underbrace{E[Y|X_1=1, X_2=1] - E[Y|X_1=1, X_2=0]}_{\beta_2+\beta_3}$$

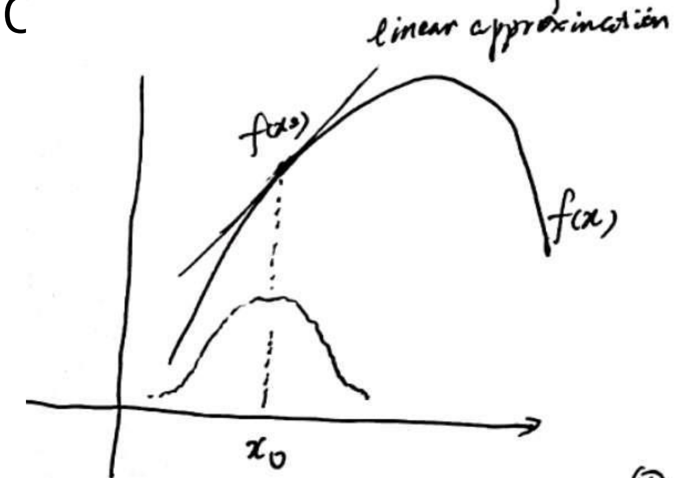
$$- \underbrace{E[Y|X_1=0, X_2=1] - E[Y|X_1=0, X_2=0]}_{\beta_2}$$

$$= \beta_3$$

Non-linear Covariate Effect

- Local Kernel Method

$y_i = f(x_i) + \varepsilon_i$ f(x) is a smooth function with certain order of derivative without a prespecified form.



① For a fixed target point x_0, we want to estimate $f(x_0)$

② We approximate $f(x)$ around x_0 by

$$\tilde{f}(x) = \beta_0 + \beta_1 (x - x_0) + \cdots + \beta_p (x - x_0)^p$$

$$\beta_j = f^{(j)}(x_0) \quad \text{Taylor expansion.}$$

④ We approximate the likelihood

$$\sum_{i=1}^{n} (y_i - f(x_i))^2$$

by

$$-LLH = \sum [y_i - \tilde{f}(x_i)]^2 K_h(x_i - x_0)$$

$$= \sum [y_i - \beta_0 - \beta_1 (x_i - x_0) - \cdots - \beta_p (x - x_p)^p]^2 K_h(x_i -$$

$\hat{\beta}_0, \hat{\beta}_1 \cdots \hat{\beta}_p$ are the minizer of local likelihood (LLH)

Note: $\hat{\beta}_j = \hat{f}^{(j)}(x_0)$, $\hat{\beta}_0 = f(x_0)$

Non-linear Covariate Effect

- Local Kernel Method

The $K_h(x_i - x_0) = \frac{1}{h} k\left(\frac{x_i - x_0}{h}\right)$ which is the kernel weight.

$k(x)$ can be a symmetric function.

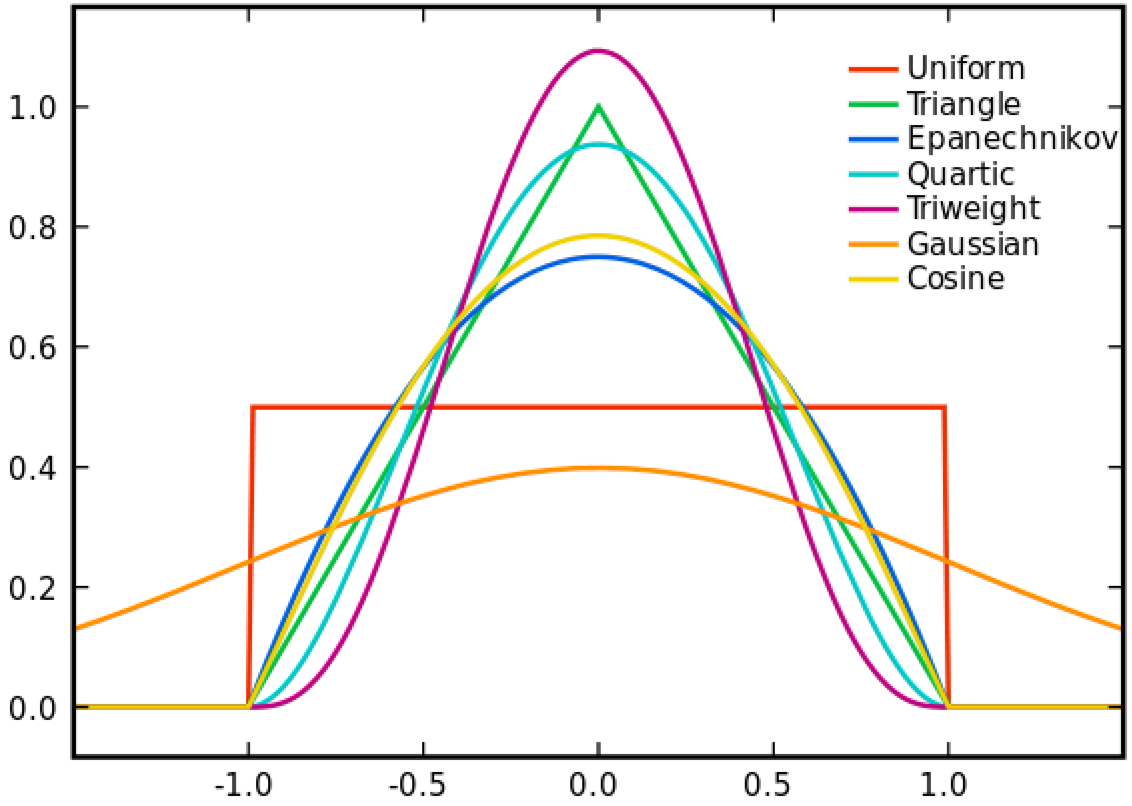
e.g. $\cancel{k(x) =}$

h is the bandwidth control the smoothness of the estimated function

Non-linear Covariate Effect

Epanechnikov $$K(u) = \frac{3}{4}(1 - u^2)$$

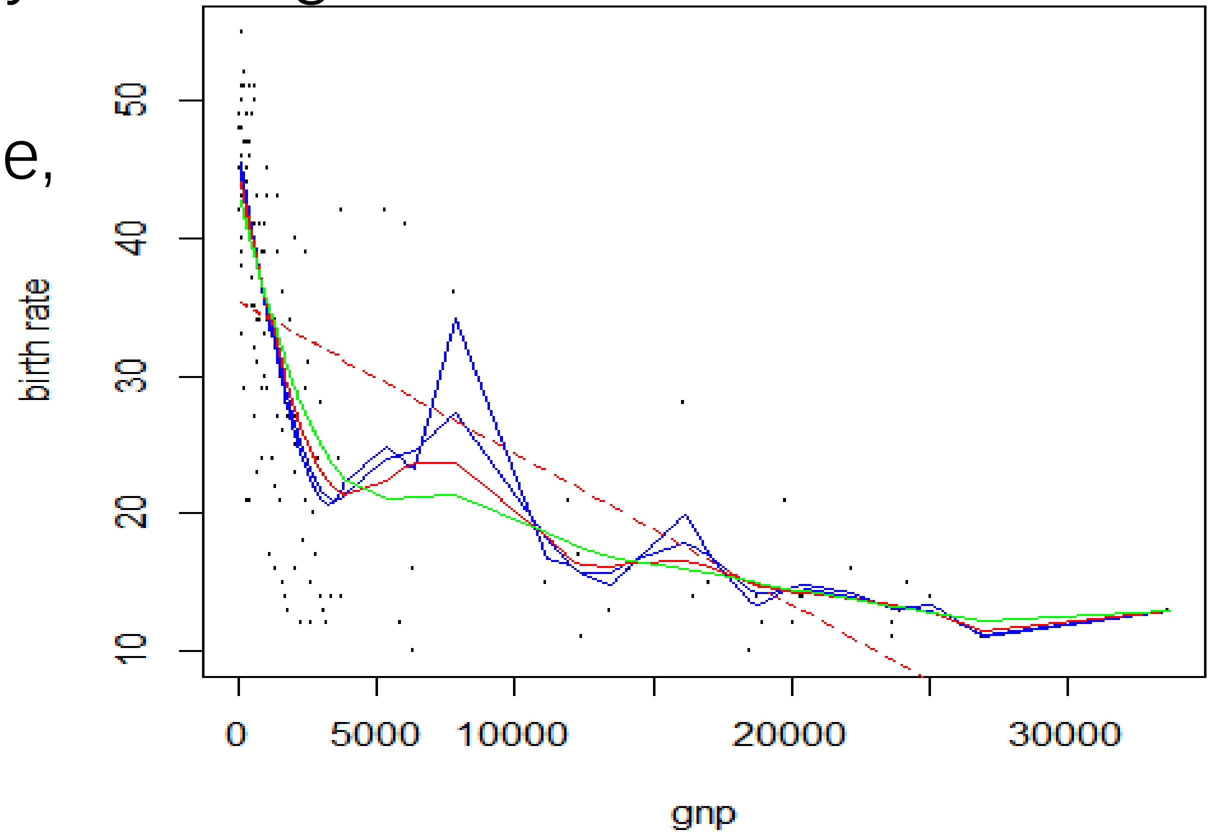
Gaussian $$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$$



Bandwidth

- As h->0, the bias introduced by the weight ->0, the variance increases.
- As h increases, the bias increase,

the variance decreases.



Bandwidth

```
library(np)
ord=order(data_lbwi$length)
data_lbwi=data_lbwi[ord,]
model.np <-
npreg(data_lbwi$headcir~data_lbwi$length ,regtype =
"ll",bwmethod = "cv.aic",gradients = TRUE)
summary(model.np)
npsigtest(model.np)

model.par <- lm(data_lbwi$headcir~data_lbwi$length)
plot(data_lbwi$length, data_lbwi$headcir, xlab = "length", ylab
= "head circumference", cex=.1)
lines(data_lbwi$length, fitted(model.np), lty=1,  col = "blue")
lines(data_lbwi$length, fitted(model.par), lty = 2, col = " red")
```

Smoothing Spline

$Y = f(x_i) + \varepsilon_i \quad i = 1 \cdots n$

LSE: $\Sigma (Y_i - f(x_i))^2$ or $[Y - f(x)]^T [Y - f(x)]$

$Y = (y_1, \cdots, y_n)^T \quad X = (x_1, x_2, \cdots, x_n)^T$

minimizer of LSE is $\hat{f}(x_i) = Y_i$ (meaningless)

$PLH = \Sigma (Y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx$

(penalized the second derivative)

☞ The minimizer will be a cubic spline with all x_i as the knots. $\hat{f}(x_i) = \sum_{j=1}^{K} \beta_j B_j(x_i)$ for some β_j

$B_j(x)$ is the cubic spline basis.

rewrite $\hat{f}(x_i) = \underline{\beta}^T \underline{B(x_i)}$

$\underline{\beta} = (\beta_1 \cdots \beta_p)^T \quad \underline{B(x)} = (B_1(x), \cdots B_p(x))^T$

Smoothing Spline

If we accept the above conclusion

$$\int f^{(2)}(x)^2 dx = \int \left[\sum_j \beta_j B_j^{(2)}(x)\right]^2 dx = \int \underline{\beta}^T \underline{B^{(2)}(x)}\, \underline{B^{(2)}(x)}^T \underline{\beta}\, dx$$

$$= \underline{\beta}^T \int B^{(2)}(x)\, B^{(2)}(x)^T dx\, \underline{\beta} = \underline{\beta}^T \underline{\underline{P}}\, \underline{\beta}$$

Therefore, the $PLH = [\underline{Y} - \underline{\underline{B(X)}}\underline{\beta}]^T [\underline{Y} - \underline{\underline{B(X)}}\underline{\beta}] + \lambda \underline{\beta}^T \underline{\underline{P}}\, \underline{\beta}$

where $\underline{\underline{B(X)}} = (B(x_1), \cdots, B(x_n))^T$

$$= \begin{pmatrix} B_1(x_1) & \cdots & B_p(x_1) \\ B_1(x_2) & \cdots & B_p(x_2) \\ \vdots & & \\ B_1(x_n) & \cdots & B_p(x_n) \end{pmatrix}$$

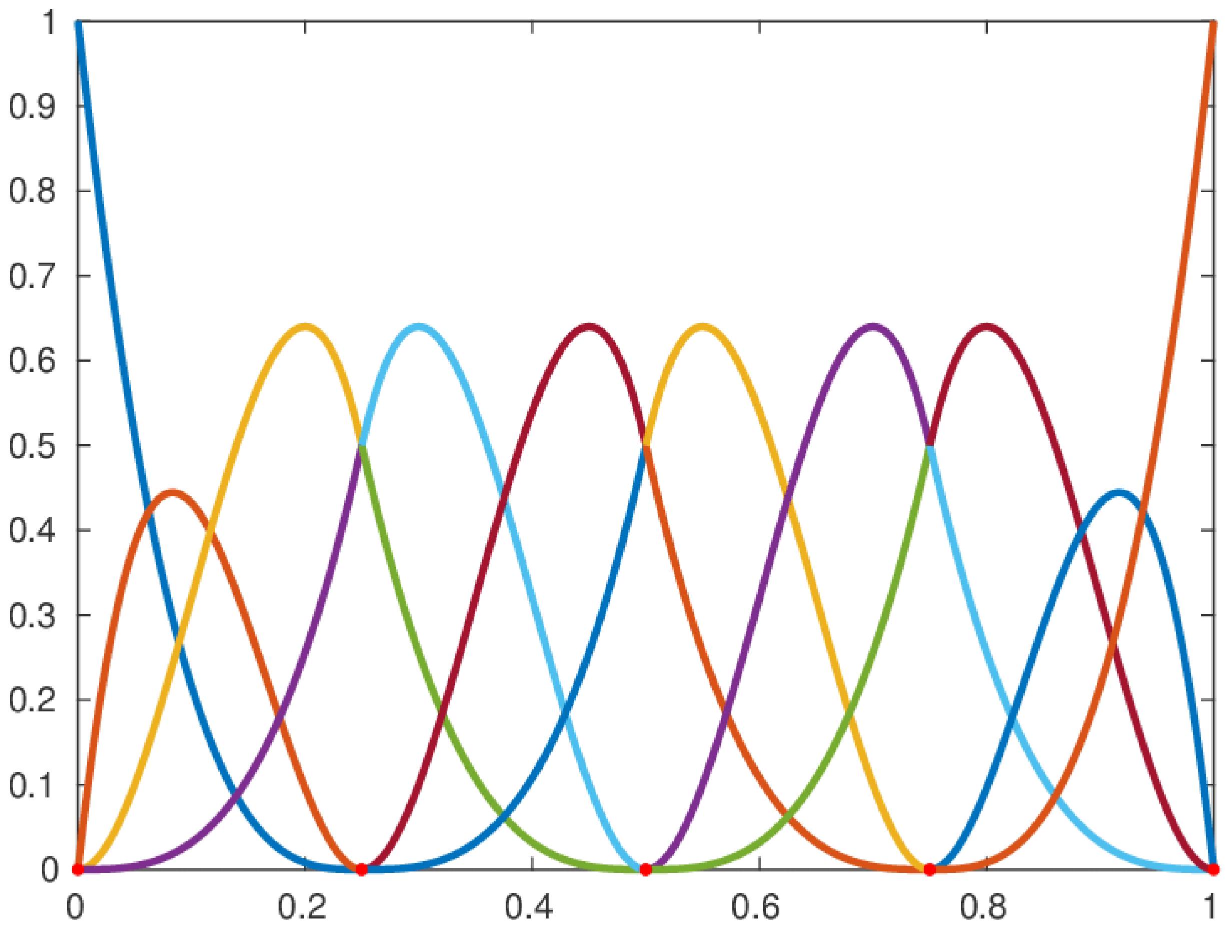
$$\frac{\partial PLH}{\partial \underline{\beta}} = \overset{-2\underline{\underline{B(X)}}^T(\underline{Y} - \underline{\underline{B(x)}}\beta)}{\cancel{-2(Y - B(x)\beta)}} + 2\lambda \underline{\underline{P}}\,\underline{\beta} = 0$$

$$+2\underline{\underline{B(x)}}^T \underline{Y} \cancel{+2\lambda P\beta} = 2\underline{\underline{B(x)}}^T \underline{\underline{B(x)}}\, \beta + 2\lambda \underline{\underline{P}}\,\underline{\beta}$$

$$\underline{\underline{B(x)}}^T \underline{Y} = (\underline{\underline{B(x)}}^T \underline{\underline{B(x)}} + \lambda \underline{\underline{P}})\, \underline{\beta}$$

$$\beta = (\underline{\underline{B(x)}}^T \underline{\underline{B(x)}} + \lambda \underline{\underline{P}})\, \underline{\underline{B(x)}}^T Y$$

Cubic B-spline basis functions



Fit Spline using MGCV

```
library(mgcv)

b <- gam(birthrt~gnp,family=gaussian(),data=data_gnp)
b <- gam(birthrt~s(gnp),family=gaussian(),bs=cr,
data=data_gnp)
plot(b)
```