

# Regression

Bret Hanlon and Bret Larget

Department of Statistics  
University of Wisconsin—Madison

December 7–14, 2010

# Example

## Example

- The proportion of blackness in a male lion's nose increases as the lion ages.
- This proportion can be used to predict the age of a lion with unknown age.
- To find a predictive equation, researchers determined the proportion of blackness in 32 male lions of known age.
- The data is displayed in a scatter plot, and a good-fitting line is found for the data.

$$(\text{age in years}) = a + b \times (\text{proportion of blackness in the nose})$$

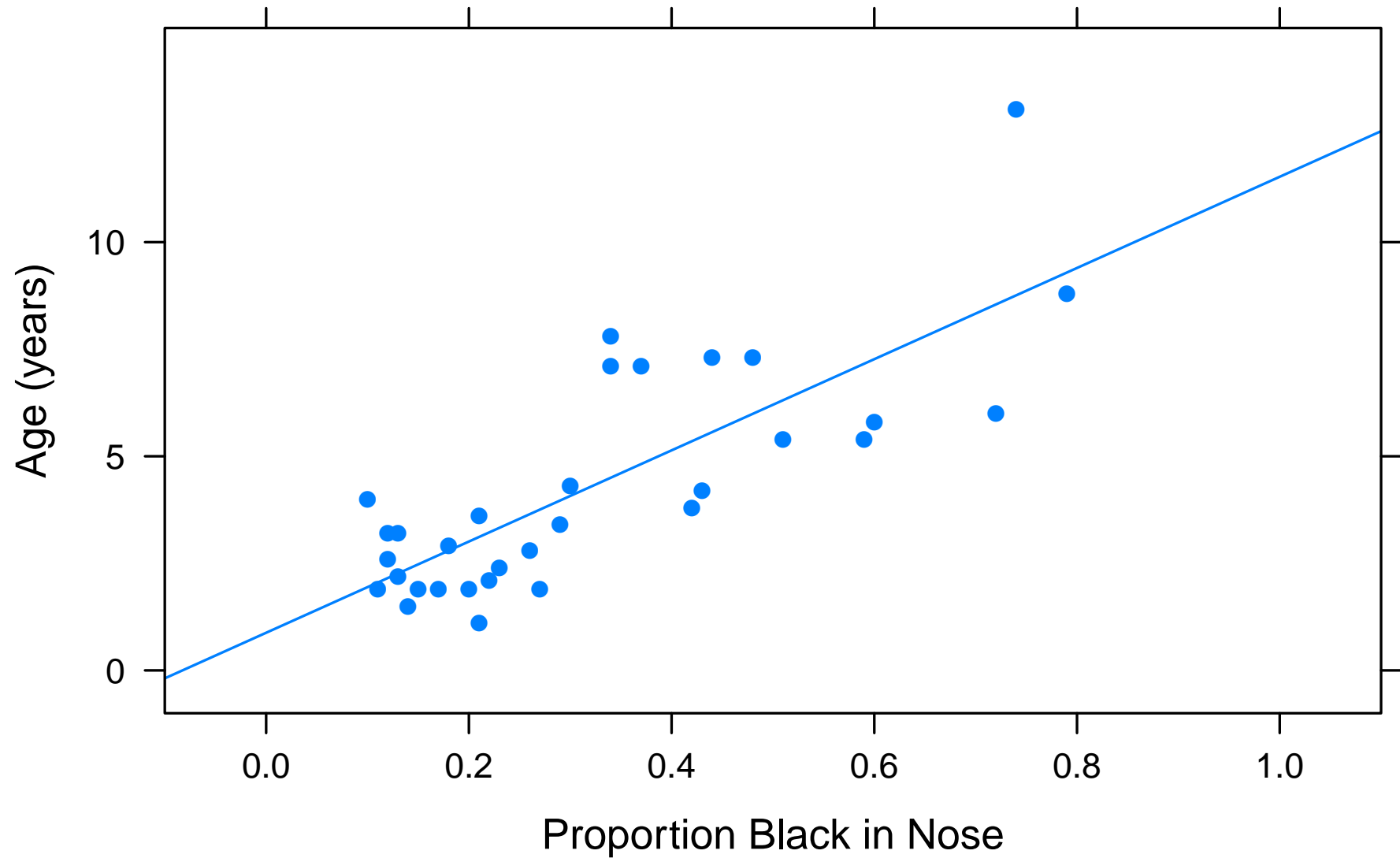
- The line may be interpreted as a *conditional expectation*: the expected age of a lion given a proportion of blackness in its nose.

# The Data

- age is the age of a male lion in years;
- `proportion.black` is the proportion of a lion's nose that is black.

age	proportion.black
1.1	0.21
1.5	0.14
1.9	0.11
2.2	0.13
2.6	0.12
3.2	0.13
3.2	0.12
...	

# Lion Data Scatter Plot



# Observations

- We see that *age and blackness in the nose are positively associated*.
- The points do not fall exactly on a line.
- How do we find a good-fitting line?
- How do we decide if a line is a sufficient summary of the relationship between the variables?

# Simple Linear Regression

## Definition

*Simple linear regression* is the statistical procedure for describing the relationship between an quantitative explanatory variable  $X$  and a quantitative response variable  $Y$  with a straight line;

$$Y = a + bX$$

- The value  $a$  is the  $Y$ -intercept of the estimated line.
- It is the location where the line crosses the  $Y$ -axis, and may be interpreted as an estimate of  $E(Y | X = 0)$ , the expected value of  $Y$  given  $X$  is zero, which may or may not be sensible in the context.
- The slope  $b$  is the estimated change in  $Y$  per unit change in  $X$ .

# A Model

- The *simple linear regression model* for the data is

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where

- ▶  $i$  varies from 1 to  $n$ , the sample size;
- ▶  $\alpha$  and  $\beta$  are fixed population parameters;
- ▶  $\varepsilon_i$  is the random vertical deviation between the line and the  $i$ th observed data point; and
- ▶ the deviations are assumed to be independent and normally distributed with standard deviation  $\sigma$ .

$$\varepsilon_i \sim \text{i.i.d } N(0, \sigma^2)$$

- Notice that in this model, there is a *common variance* for all observations.
- This means that we should expect the size of a typical deviation from the line to be the same size at all locations.

# Fitted Values and Residuals

- The estimated regression line takes the form

$$Y = a + bX$$

where  $a$  is an estimate of  $\alpha$  and  $b$  is an estimate of  $\beta$ .

- The height of the point on the line at  $X = X_i$  is called the  $i$ th *fitted value* or *predicted value*.

$$\hat{Y}_i = a + bX_i$$

- The difference between the  $i$ th data point  $Y_i$  and the  $i$ th predicted value is called the  $i$ th *residual*,  $Y_i - \hat{Y}_i$ .

# Estimation

- The parameters of the model may be estimated either by the criteria of *least squares*, which *minimizes the sum of squared residuals*.
- The parameters may also be estimated by *maximum likelihood*, which makes the probability density of the observed data as large as possible.
- In simple linear regression, the least squares and maximum likelihood estimates of  $\alpha$  and  $\beta$  are identical.
- The maximum likelihood estimate of  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{(\text{sum of squared residuals})}{n}$$

is slightly different than the conventional unbiased estimate.

$$s^2 = \frac{(\text{sum of squared residuals})}{n - 2}$$

- Note that there are 2 parameters used to describe all means ( $\alpha$  and  $\beta$ ) as two points determine a line, and so there are  $n - 2$  remaining pieces of information remaining to estimate variation around the line.

# Formulas for Estimation

- It is an exercise in calculus (or inspired algebra) to show that

$$b = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

is the least squares (and maximum likelihood) estimate of the slope.

- There is a simple formula for the estimated intercept given the estimated slope.

$$a = \bar{Y} - b\bar{X}$$

- The *residual sum of squares* is

$$\text{RSS} = \sum_{i=1}^n (Y_i - (a + bX_i))^2$$

is used to estimate  $\sigma^2$  by dividing by either  $n - 2$  (for an unbiased estimate) or  $n$  (for the maximum likelihood estimate).

## An alternative formulation

- The estimated parameters may also be described in an alternative manner based on the means and standard deviations of  $X$  and  $Y$  and the correlation between them.
- The formulas are based on this idea:

*When  $X$  is  $z = \frac{X - \bar{X}}{s_x}$  standard deviations above the mean, ( $z < 0$  when  $X$  is less than  $\bar{X}$ ), the predicted  $Y$  is  $rz$  standard deviations above its mean, or*

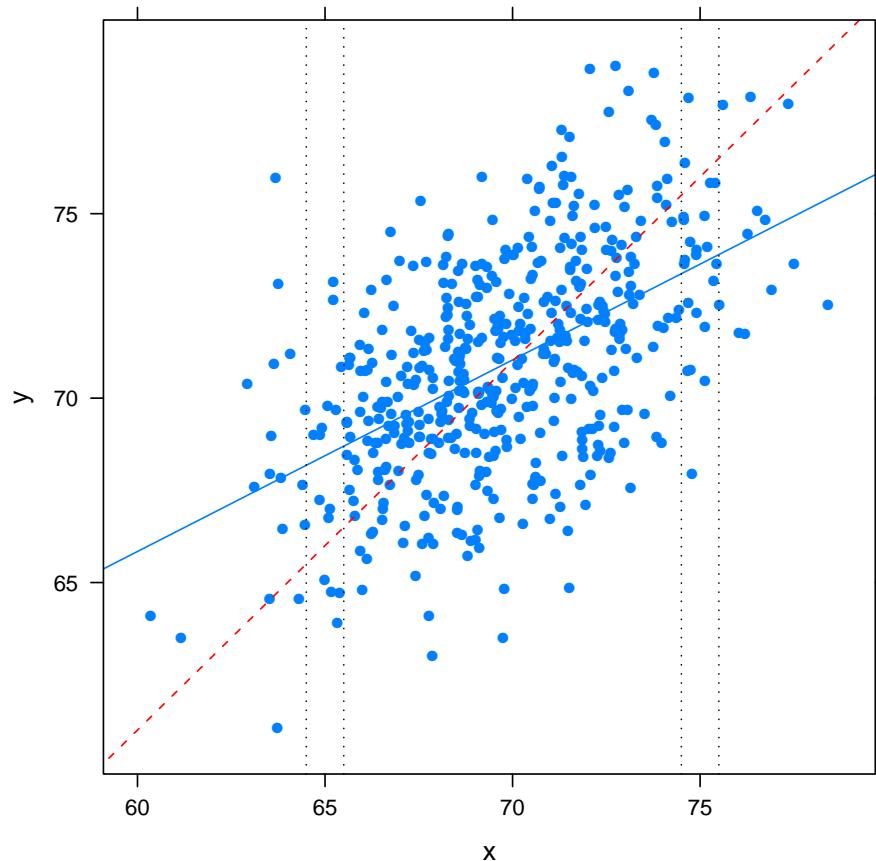
$$\hat{Y} = \bar{Y} + r \left( \frac{X - \bar{X}}{s_x} \right) s_y = \left( \bar{Y} - \left( r \frac{s_y}{s_x} \right) \bar{X} \right) + \left( r \frac{s_y}{s_x} \right) X$$

- The slope is  $b = rs_y/s_x$ .
- When  $X = \bar{X}$ , then  $z = 0$  and  $\hat{Y} = \bar{Y}$ , so the regression line goes through the point  $(\bar{X}, \bar{Y})$ .
- When  $X = \bar{X} + s_x$  is one standard deviation above the mean, the predicted value is  $\hat{Y} = \bar{Y} + rs_y$   $r$  standard deviations above the mean.

# Regression Toward the Mean

- Heights of fathers and sons in human populations often have a correlation close to  $r = 0.5$ .
- If one uses the height of the father to predict the height of the son, the average heights of all sons of fathers whose height is one standard deviation above the mean is only about one half of a standard deviation above the mean.
- Similarly, the heights of sons of fathers that are one standard deviation below the mean are expected to be only half a standard deviation below the mean.
- This general phenomenon is called *regression toward the mean*.

# Scatter Plot to Illustrate Regression Toward the Mean



- The solid blue line is the regression line with slope  $rs_Y/s_X$ .
- The dashed line has slope  $s_Y/s_X$  and passes through the principal axis of the data.
- Notice that the average  $Y$  value of points in the narrow bands is much closer to the regression line than the other.
- The regression line is flatter than the red dashed line.
- If we exchanged  $X$  and  $Y$ , the red dashed line would just flip, *but the regression line would be different.*

# Reconsider the Lion Data

- Reconsider the lion data
- Use the function `lm()` to fit the regression model and `summary()` to display the results.
- Here is R code to read the data and fit the model.

```
> lions = read.csv("lions.csv")
```

```
> str(lions)
```

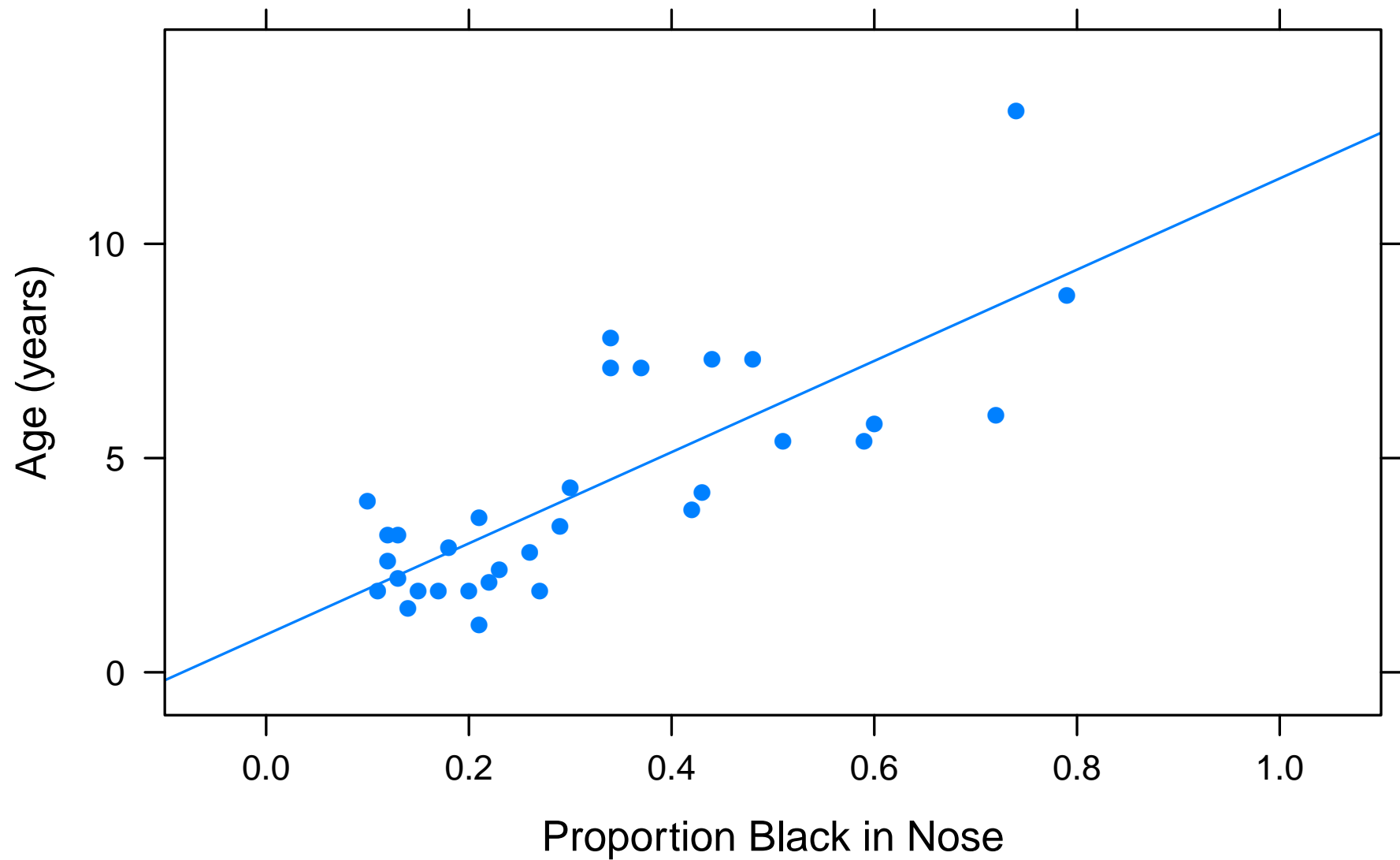
```
'data.frame': 32 obs. of 2 variables:
```

```
$ age          : num  1.1 1.5 1.9 2.2 2.6 3.2 3.2 2.9 2.4 2.1 ...
```

```
$ proportion.black: num  0.21 0.14 0.11 0.13 0.12 0.13 0.12 0.18 0.23 0.22
```

```
> lions.lm = lm(age ~ proportion.black, data = lions)
```

# Lion Data Scatter Plot



# Summary

```
> summary(lions.lm)
```

Call:

```
lm(formula = age ~ proportion.black, data = lions)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.5449	-1.1117	-0.5285	0.9635	4.3421

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.8790	0.5688	1.545	0.133
proportion.black	10.6471	1.5095	7.053	7.68e-08 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.669 on 30 degrees of freedom

Multiple R-squared: 0.6238, Adjusted R-squared: 0.6113

F-statistic: 49.75 on 1 and 30 DF, p-value: 7.677e-08

# Interpretation

- The estimated intercept is  $a = 0.879$  years.
- This means that a lion with no black its nose is expected to be about 10 or 11 months old. (This interpretation may not be very reliable as it *extrapolates* beyond the range of the data).
- The slope is  $b = 10.65$  years per proportion black. This means (using more practical numbers) that if the proportion of black increases by 0.1, the age is expected to increase by about  $10.65/10 \doteq 1.07$  years.
- The standard errors may be used to find confidence intervals.
- The hypothesis test  $\beta = 0$  has very strong evidence against it (two-sided  $t$ -test,  $p < 10^{-7}$ ,  $t = 7.05$ ,  $df = 30$ ).

# Verifying Textbook Formulas

```
> x = with(lions, proportion.black)
> y = with(lions, age)
> c(mean(x), sd(x), mean(y), sd(y), cor(x, y))

[1] 0.3221875 0.1985550 4.3093750 2.6765842 0.7898272

> b.text = sum((x - mean(x)) * (y - mean(y)))/sum((x - mean(x))^2)
> b.text

[1] 10.64712

> b.cor = cor(x, y) * sd(y)/sd(x)
> b.cor

[1] 10.64712

> a = mean(y) - b.text * mean(x)
> a

[1] 0.8790062

> sigma.hat = sqrt(sum(residuals(lions.lm)^2)/(nrow(lions) - 2))
> sigma.hat

[1] 1.668764
```

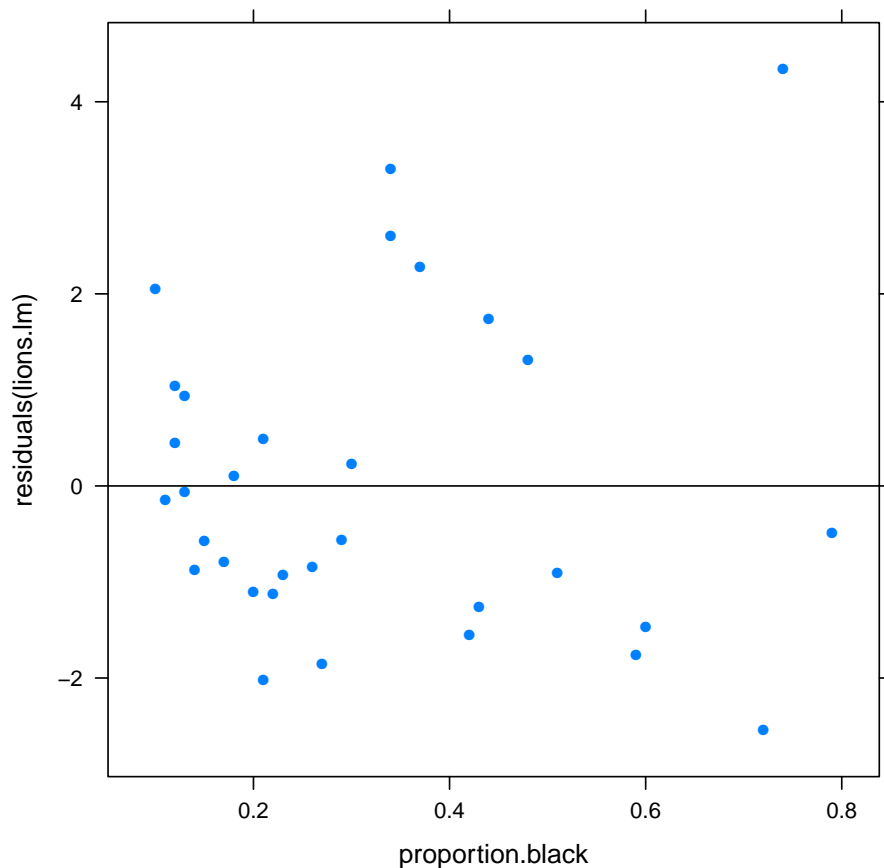
# Assumptions

- In addition to random sampling from a population of interest, simple linear regression makes these assumptions:
  - ① **Linearity:** there is a linear relationship between  $X$  and  $E(Y | X)$ .
  - ② **Independence:** observations are independent of each other.
  - ③ **Constant Variance:** the random deviations of the observed values from the true regression line have the same standard deviation for all observations.
  - ④ **Normality:** the random deviations of the observed values from the true regression line are normally distributed.
- It is good statistical practice to check these assumptions.
- Regression is robust to modest deviations from normality and constant variance.
- Violations of independence should be dealt with by using more complex models.

# Residual Plots

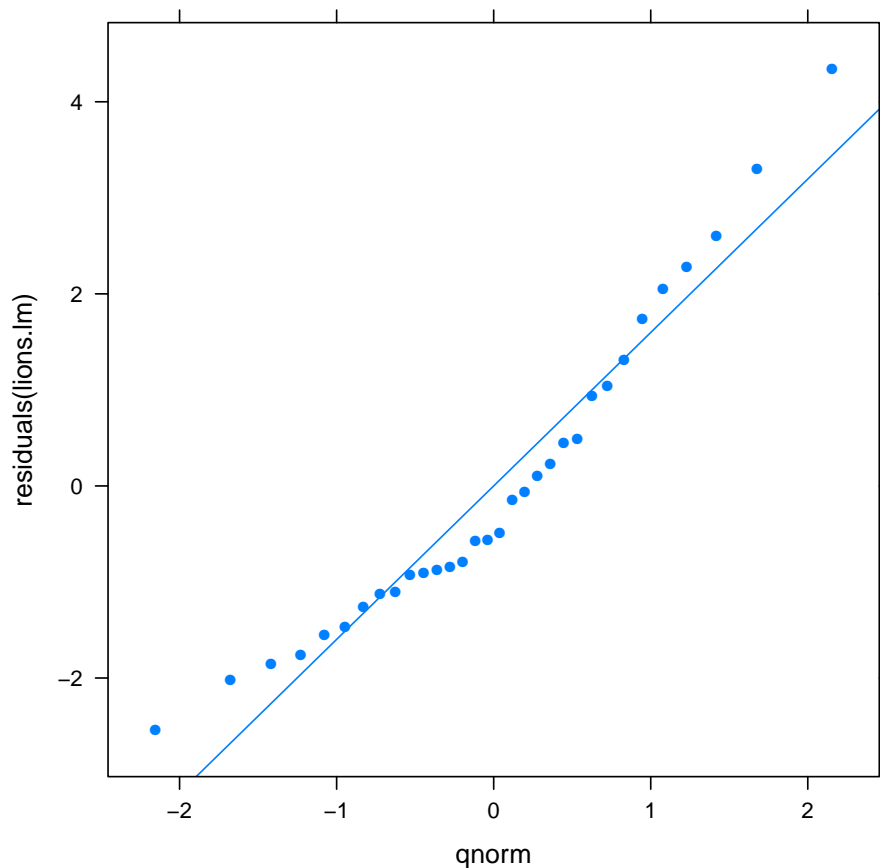
- A *residual plot* might plot residuals versus  $X$  or versus fitted values.
- (Plots versus  $X$  are easy to interpret in simple linear regression models with one explanatory variable; in models with more than one explanatory variable, plots versus fitted values are feasible.)
- Examine the residual plot for *deviations from linearity*: Is there an up-down pattern to the residuals?
- Examine the residual plot for *deviations from constant variance*: Does the size of residuals change depending on  $X$  (or fitted values)?
- Examine the residual plot for *deviations from normality*: Are there obvious signs of skewness or outliers?

# Lion Data Residual Plot



- There is no strong apparent nonlinearity.
- Residual size might increase slightly as  $X$  increases, but not too much to be worried about.
- There are no extreme outliers or strong skewness.

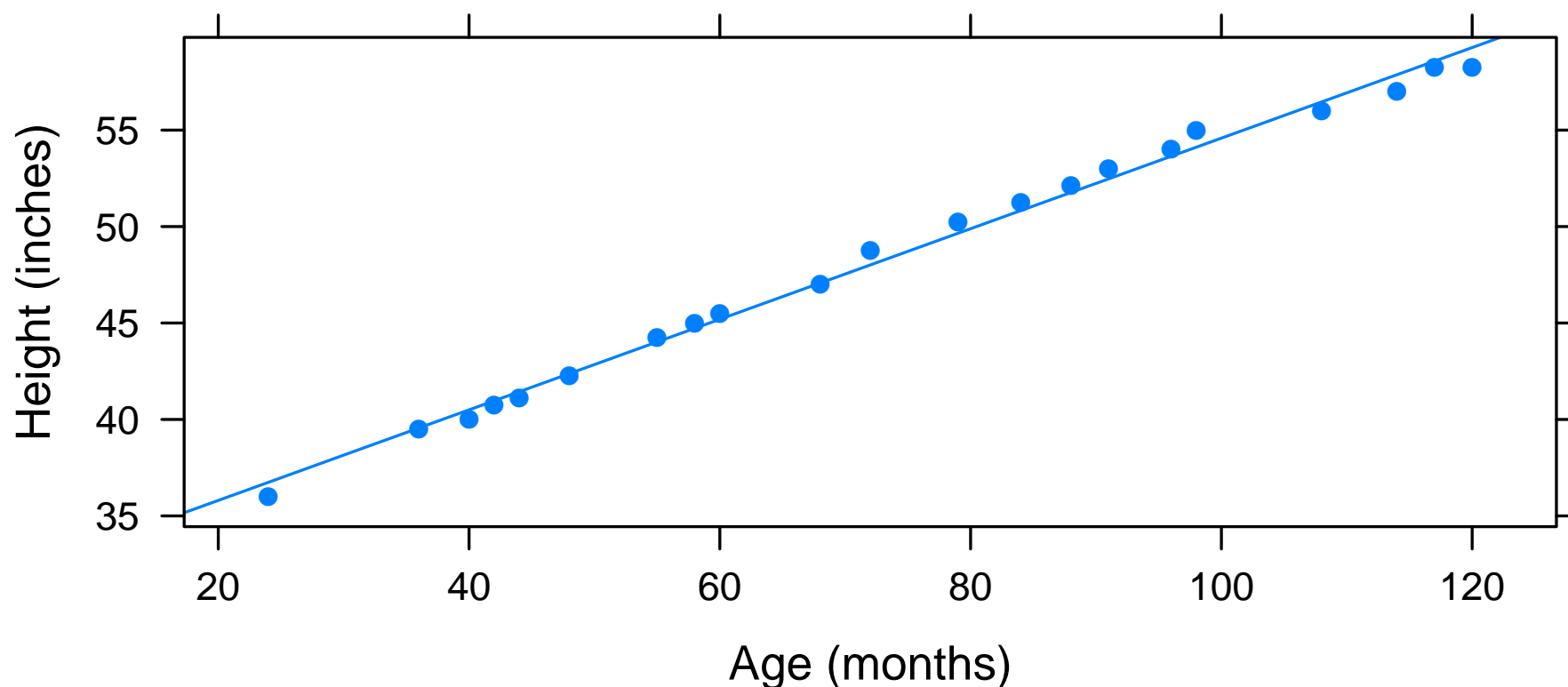
# Normal Quantile Plot of Residuals



- The line is curved a up bit, which indicates residual distribution is skewed slightly to the right.
- The data is not perfectly normal, but nonnormality is not too severe, and inferences based on normality are probably okay.

## Riley Larget

- Riley Larget is Professor Larget's oldest child. He will turn 18 soon.
- As I am sure almost almost all of your parents have done for you and any siblings, there is a special place in the Larget household where children height is tracked with carefully measured pencil marks.
- Below is a plot of his height versus his age, from ages 2 to 10 years.



# Linear Fit

- The previous scatter plot shows a pretty good fit to a linear model for height versus age in this age range.
- Examine a the data and a fitted linear regression model.

```
age  height
...
24   36
36   39.5
40   40
42   40.75
44   41.125
48   42.25
...
```

```
> riley = read.table("riley.txt", header = T)
> age.2.10 = with(riley, (age > 23) & (age < 121))
> riley.lm = lm(height ~ age, data = riley, subset = age.2.10)
```

# Summary

```
> summary(riley.lm)
```

Call:

```
lm(formula = height ~ age, data = riley, subset = age.2.10)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.0290	-0.3247	-0.0580	0.3588	0.8860

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	31.106153	0.332111	93.66	<2e-16 ***
age	0.234774	0.004218	55.66	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.549 on 19 degrees of freedom

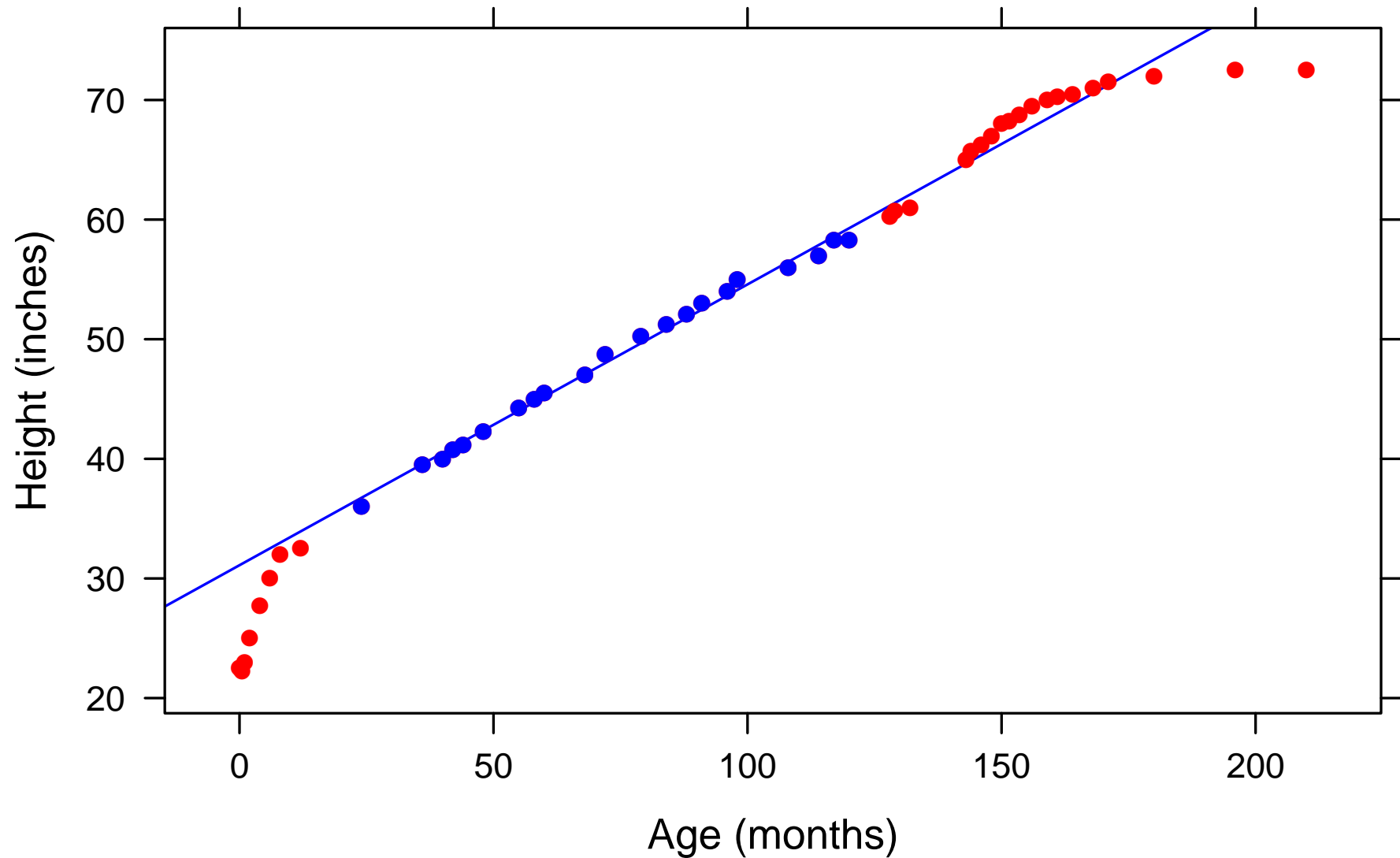
Multiple R-squared: 0.9939, Adjusted R-squared: 0.9936

F-statistic: 3097 on 1 and 19 DF, p-value: < 2.2e-16

# Interpretation

- The intercept of 31.11 inches can be interpreted as Riley's height at age 0 (or birth). Is this a reasonable estimate?
- The slope of 0.235 inches per month can be interpreted as Riley's rate of growth.
- This is just under 3 inches per year. Is this a reasonable estimate?

## Plot of Data from Birth to 17.5 years



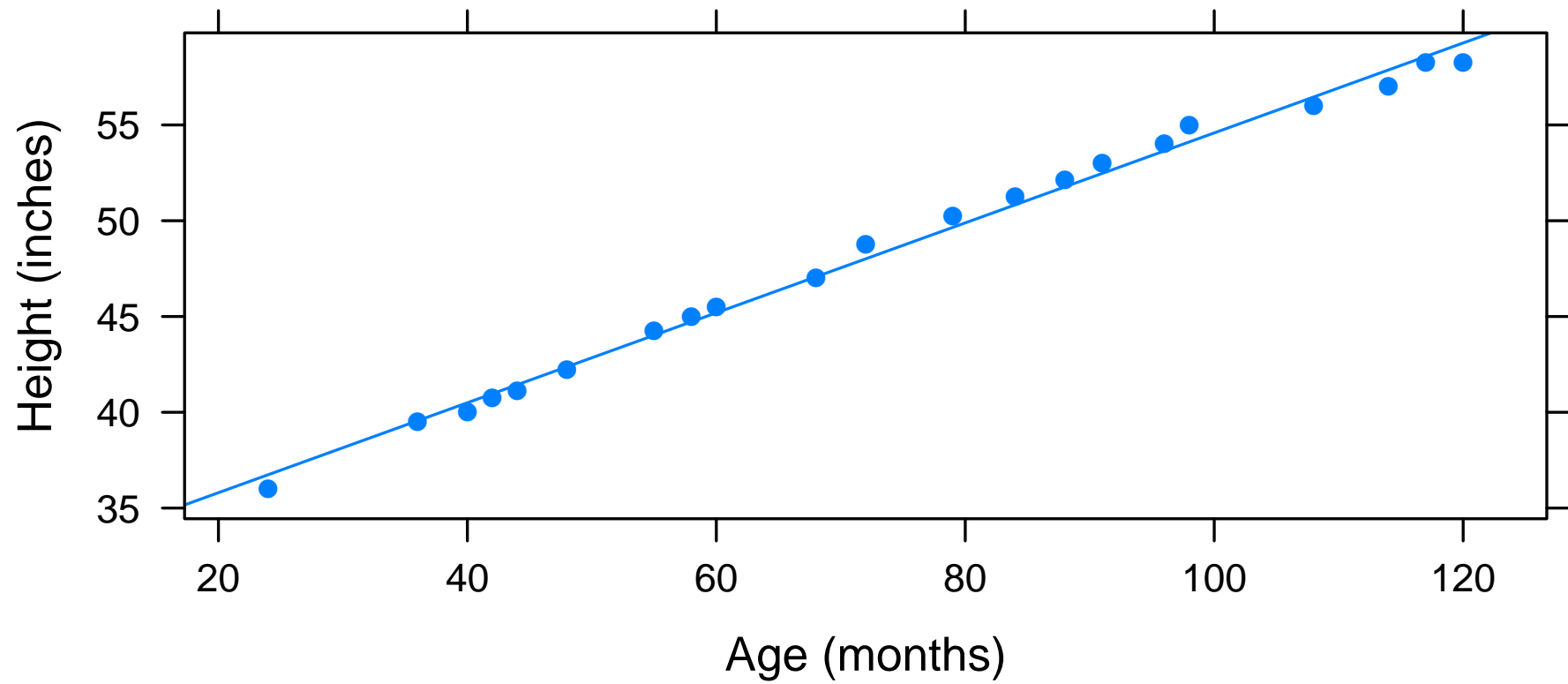
# Extrapolation

- The previous plot shows Riley's height from age birth to 17.5 years.
- Blue points were used to fit the first regression model.
- Red points were not used for the model.
- We see that the estimated intercept of 31.1 inches is much larger than the actual value of 22.5 inches.
- This estimate is *an extrapolation*, or the use of a regression model to make predictions outside the range of data for which the model was estimated.
- Extrapolations are very susceptible to non-linearity.
- A linear model may be reasonable within the range of some of the data, but it may not necessarily extend past the range of observed data.
- Similarly, while the linear model captures well the growth up to age about 12 (144 months), it misses a growth spurt in the early teens and misses Riley's eventual plateau at his current height of 72.5 inches.

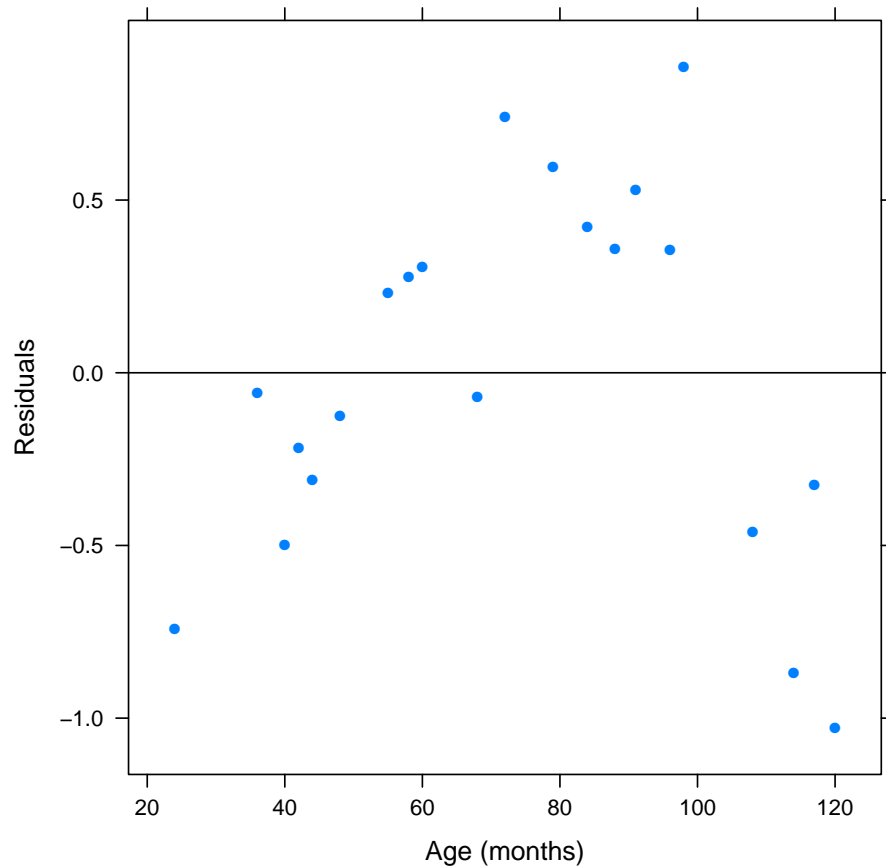
# Fitting a Curve

- *Multivariate regression* is an extension of simple linear regression in which there are more than one explanatory variables.
- These explanatory variables need not be independent.
- In fact, if we use  $X$  and  $X^2$  as two explanatory variables for  $Y$ , we fit a curve (specifically, a parabola or degree 2 polynomial) to data.
- Let's reexamine the Riley's height data to see if a curve fits much better than a straight line.

## Riley's Height, age 2-10 years



# Residual Plot



- Notice the pattern in the residual plot where residuals tend to be negative, then positive, and then negative again.
- This is an indication that a curve may fit better than a straight line.
- Note that the correlation between height and age (for ages from 2–10 years) is 0.997, but despite this very high value, a curve may still be significantly better.

# Fitting a Quadratic Model

- The quadratic model is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

where  $\varepsilon_i \sim N(0, \sigma^2)$  and the parameters again are estimated by least squares or maximum likelihood (which again is identical except for  $\sigma^2$ ).

- Even though we do not fit a straight line, this model is still in the class of *linear models* because the model for the mean is a sum of terms, each of which is a parameter times a predictor variable.
- When specifying the formula in a linear model in R, the symbols \* and ^ mean something other than multiplication and exponentiation.
- The command I() is necessary in the formula below so that R interprets the symbol ^ as exponentiation.

```
> riley.lm2 = lm(height ~ age + I(age^2), data = riley, subset = age.2.10)
```

# Summary

```
> summary(riley.lm2)
```

Call:

```
lm(formula = height ~ age + I(age^2), data = riley, subset = age.2.10)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.53103	-0.24210	-0.02807	0.21074	0.73646

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.827e+01	4.821e-01	58.644	< 2e-16 ***
age	3.241e-01	1.417e-02	22.866	9.43e-15 ***
I(age^2)	-6.002e-04	9.391e-05	-6.392	5.10e-06 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

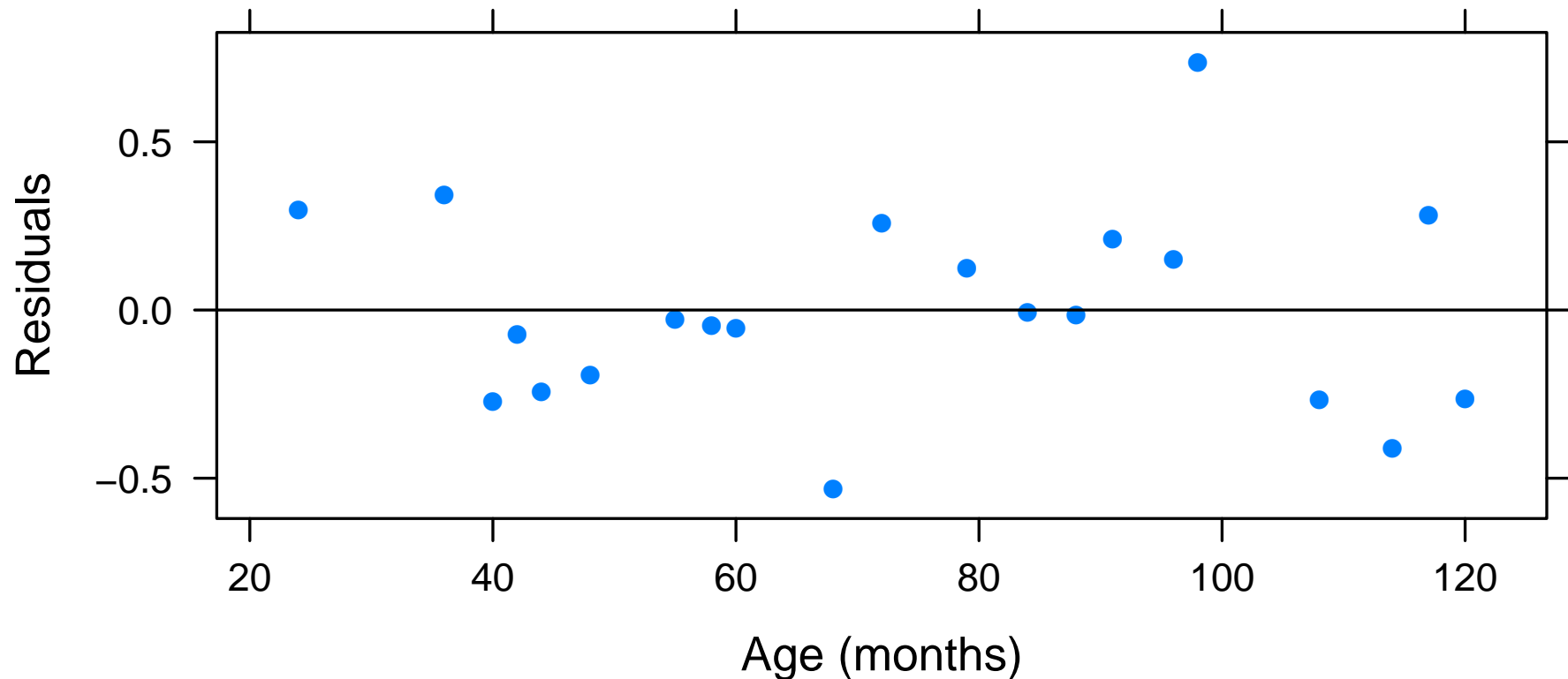
Residual standard error: 0.3119 on 18 degrees of freedom

Multiple R-squared: 0.9981, Adjusted R-squared: 0.9979

F-statistic: 4818 on 2 and 18 DF, p-value: < 2.2e-16

# Observations

- The summary indicates that the parameter for the quadratic term improves the fit significantly.
- The residual plot now shows an up/down/up/down pattern.



# Fitting a Cubic Model

- The cubic model is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \varepsilon_i$$

where  $\varepsilon_i \sim N(0, \sigma^2)$ .

```
> riley.lm3 = lm(height ~ age + I(age^2) + I(age^3), data = riley,  
+ subset = age.2.10)
```

# Summary

```
> summary(riley.lm3)
```

Call:

```
lm(formula = height ~ age + I(age^2) + I(age^3), data = riley,  
    subset = age.2.10)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.48304	-0.12357	-0.01843	0.08316	0.56923

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	3.043e+01	9.978e-01	30.500	2.77e-16	***
age	2.145e-01	4.733e-02	4.533	0.000294	***
I(age^2)	1.037e-03	6.870e-04	1.509	0.149579	
I(age^3)	-7.423e-06	3.092e-06	-2.401	0.028080	*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

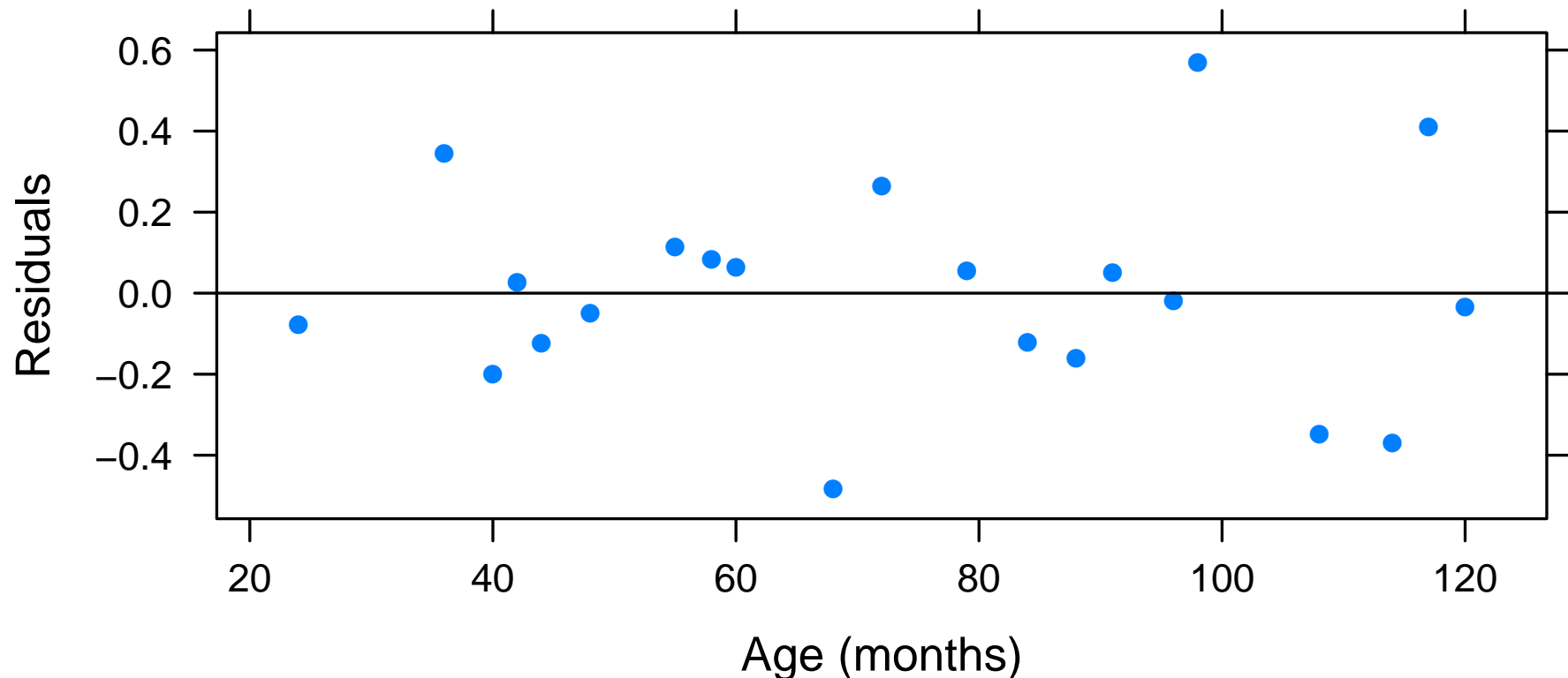
Residual standard error: 0.2774 on 17 degrees of freedom

Multiple R-squared: 0.9986, Adjusted R-squared: 0.9984

F-statistic: 4064 on 3 and 17 DF, p-value: < 2.2e-16

# Observations

- The summary indicates that the parameter for the cubic term also improves the fit significantly.
- When leaving in a term for a higher power, you should also leave in terms for all lesser powers.
- The residual plot does not show a pattern now.



# Fitting a Degree 4 Polynomial

- Even though the residual plot does not show a pattern, we can see if adding another term improves the fit.
- The degree four model is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \beta_4 X_i^4 + \varepsilon_i$$

where  $\varepsilon_i \sim N(0, \sigma^2)$ .

```
> riley.lm4 = lm(height ~ age + I(age^2) + I(age^3) + I(age^4),  
+ data = riley, subset = age.2.10)
```

# Summary

```
> summary(riley.lm4)
```

Call:

```
lm(formula = height ~ age + I(age^2) + I(age^3) + I(age^4), data = riley,  
    subset = age.2.10)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.47258	-0.13139	-0.01963	0.08789	0.55926

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.015e+01	2.339e+00	12.889	7.26e-10 ***
age	2.355e-01	1.618e-01	1.456	0.165
I(age^2)	5.179e-04	3.884e-03	0.133	0.896
I(age^3)	-2.211e-06	3.847e-05	-0.057	0.955
I(age^4)	-1.821e-08	1.340e-07	-0.136	0.894

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2857 on 16 degrees of freedom

Multiple R-squared: 0.9986, Adjusted R-squared: 0.9983

# Observations

- The summary indicates that the parameter for the degree 4 term is not significant.
- The `anova()` function in R can also be used to summarize a nested sequence of models, from simpler to more complex.
- The table shows residual degrees of freedom, RSS, changes in RSS, and an inference on the significance of the change.
- Each F statistic is the change in RSS from the previous model divided by the estimate of  $\sigma^2$  from the last model (1.3063/16) and degrees of freedom are change in df and df for the last model.

```
> anova(riley.lm, riley.lm2, riley.lm3, riley.lm4)
```

Analysis of Variance Table

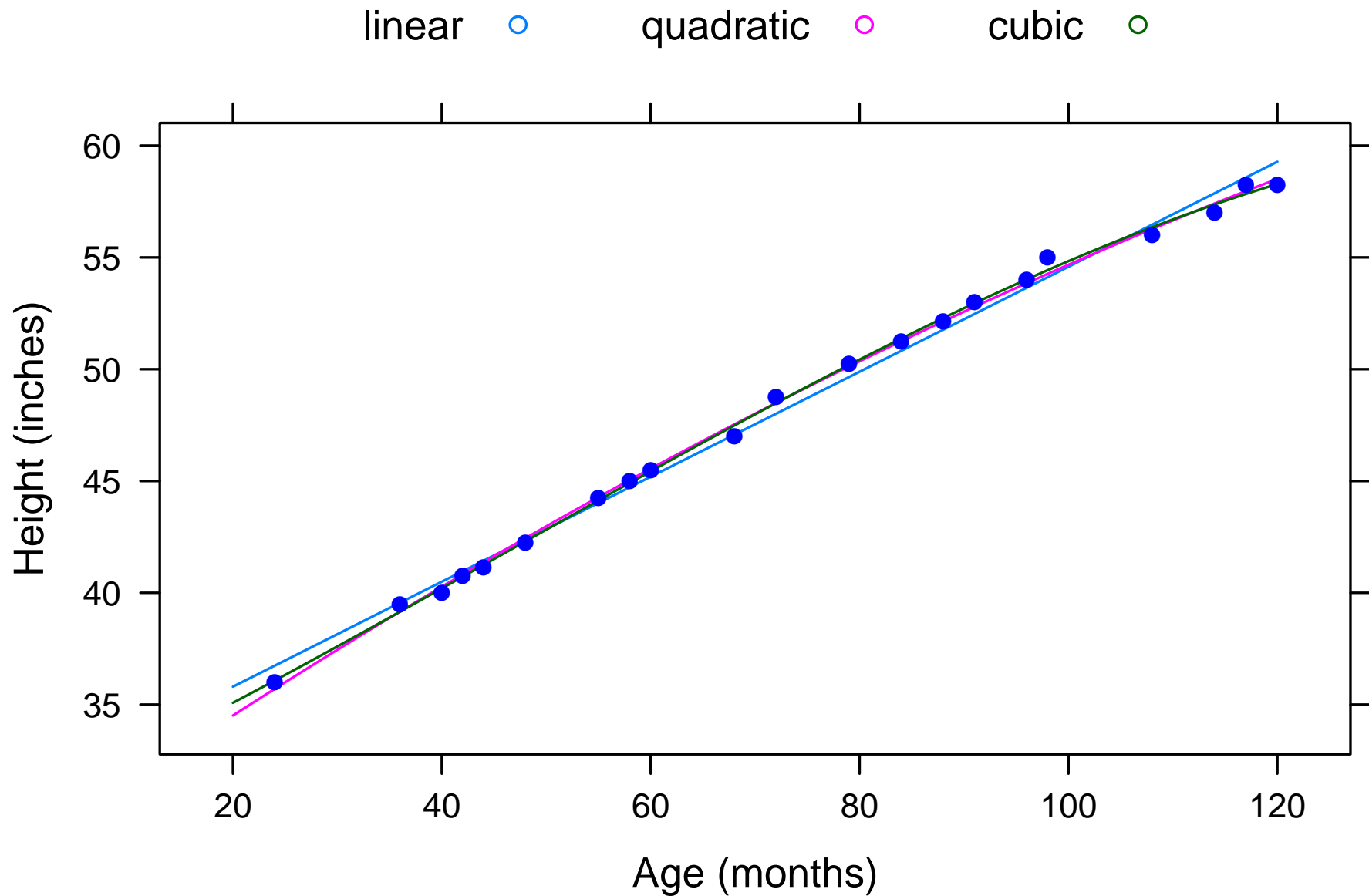
```
Model 1: height ~ age
Model 2: height ~ age + I(age^2)
Model 3: height ~ age + I(age^2) + I(age^3)
Model 4: height ~ age + I(age^2) + I(age^3) + I(age^4)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	19	5.7265				
2	18	1.7513	1	3.9752	48.6887	3.115e-06 ***
3	17	1.3078	1	0.4434	5.4312	0.03319 *
4	16	1.3063	1	0.0015	0.0185	0.89357

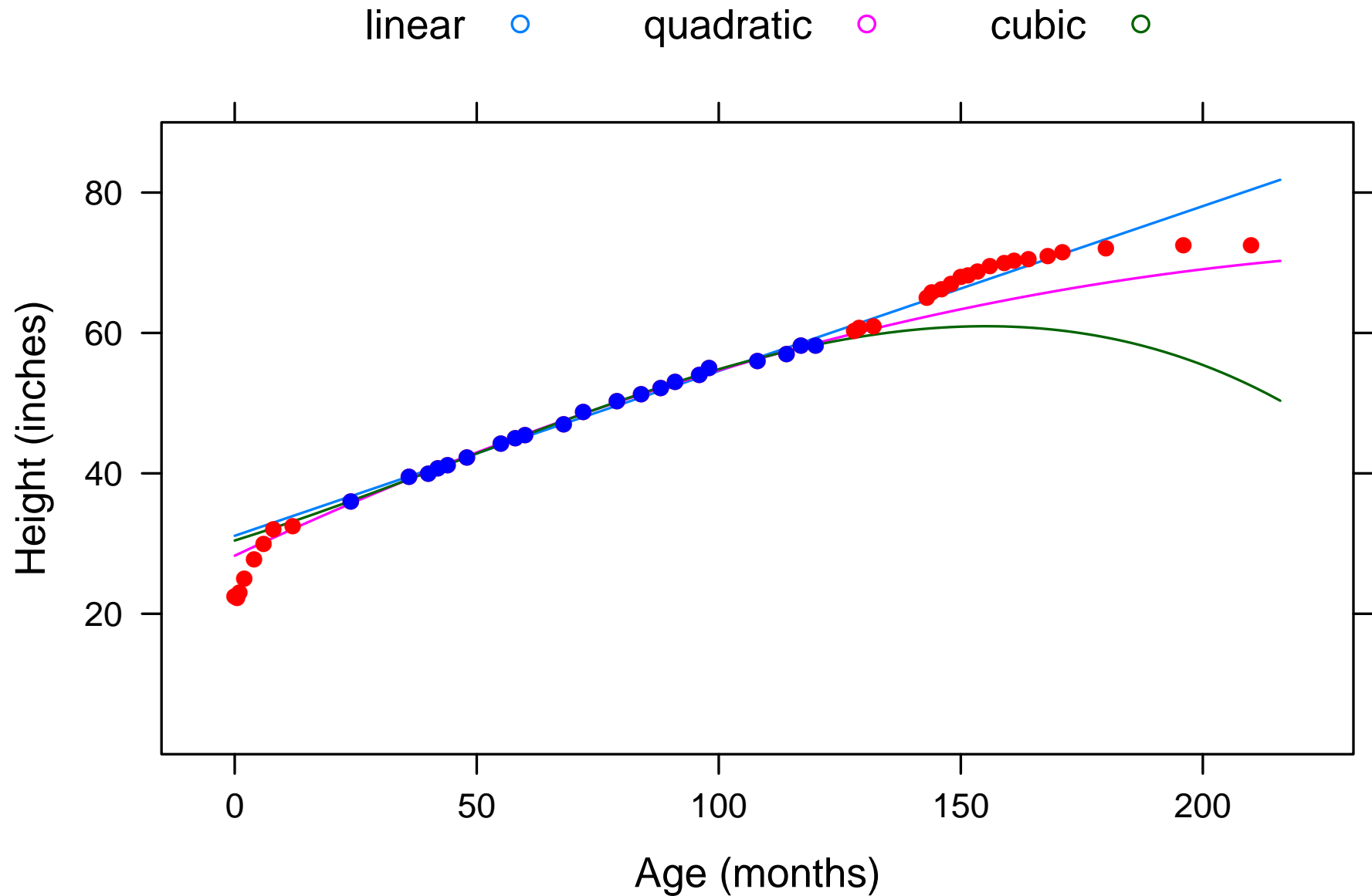
---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Plots of Three Fits



# Plots of Three Fits Over Full Range



# Observations

- While the quadratic and cubic models are significantly better than the linear model (in the region from 2 to 10 years), the actual differences in predicted heights within this range are not so different.
- None of the models do well when extrapolated beyond the range of the data.
- To find a good model for the entire data range (fit using all of the data), some sort of *nonlinear regression* would be necessary.
- In particular, we would want a model where the function eventually flattens to a horizontal asymptote.

# Confidence and Prediction Intervals

- The summaries of regression models include *inferences about model parameters*.
- We may also wish to make inferences about *conditional means* and predictions about *new individual observations*.
- For example, how can we find a 95% confidence interval for the average age of a male lion whose nose is 20 percent black?
- In addition, how can we find a 95% prediction interval for the age of a single male lion whose nose is 20 percent black?
- The first type of interval will be much smaller than the second, because the first is only about uncertainty about the location of the regression line while the second incorporates both uncertainty about the line and individual variation.

# Confidence Intervals

- In simple linear regression, the estimate of  $E(Y | X)$  has standard error

$$SE = \sqrt{\frac{RSS}{\text{residual df}} \left( \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)}$$

- Notice that this interval gets wider as  $X$  moves away from  $\bar{X}$ .
- Also note that as  $n$  becomes very large, both fractions under the square root get small and the SE approaches zero.
- A confidence interval for  $E(Y | X)$  has the form

$$\hat{Y} - t^*SE < E(Y | X) < \hat{Y} + t^*SE$$

where  $t^*$  is a critical value from a  $t$  distribution with  $n - 2$  degrees of freedom.

## Example

- While possible to use summary data to find these intervals, it is more convenient to use the R function `predict()`.
- The first argument to `predict()` is a fitted model.
- The second argument is a data frame with the desired explanatory variables.
- Here are 95% confidence intervals for the average age of lions with either 0.2, 0.3, and 0.4 proportion black noses.
- In the data set, the mean proportion is 0.32, so the confidence interval is most narrow at 0.3.

```
> predict(lions.lm, data.frame(proportion.black = c(0.2, 0.3, 0.4)),  
+       level = 0.95, interval = "confidence")
```

	fit	lwr	upr
1	3.008430	2.297898	3.718962
2	4.073142	3.466804	4.679480
3	5.137854	4.489386	5.786322

# Prediction Intervals

- In simple linear regression, the prediction of a single  $Y$  given  $X$  has standard error

$$SE = \sqrt{\frac{RSS}{\text{residual df}} \left( 1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)}$$

- Notice that this interval gets wider as  $X$  moves away from  $\bar{X}$ .
- Also note that as  $n$  becomes very large, both fractions under the square root get small, but 1 does not, so the SE approaches  $\sigma$ .
- A prediction interval for a single observation  $Y$  given  $X$  has the form

$$\hat{Y} - t^*SE < Y < \hat{Y} + t^*SE$$

where  $t^*$  is a critical value from a  $t$  distribution with  $n - 2$  degrees of freedom.

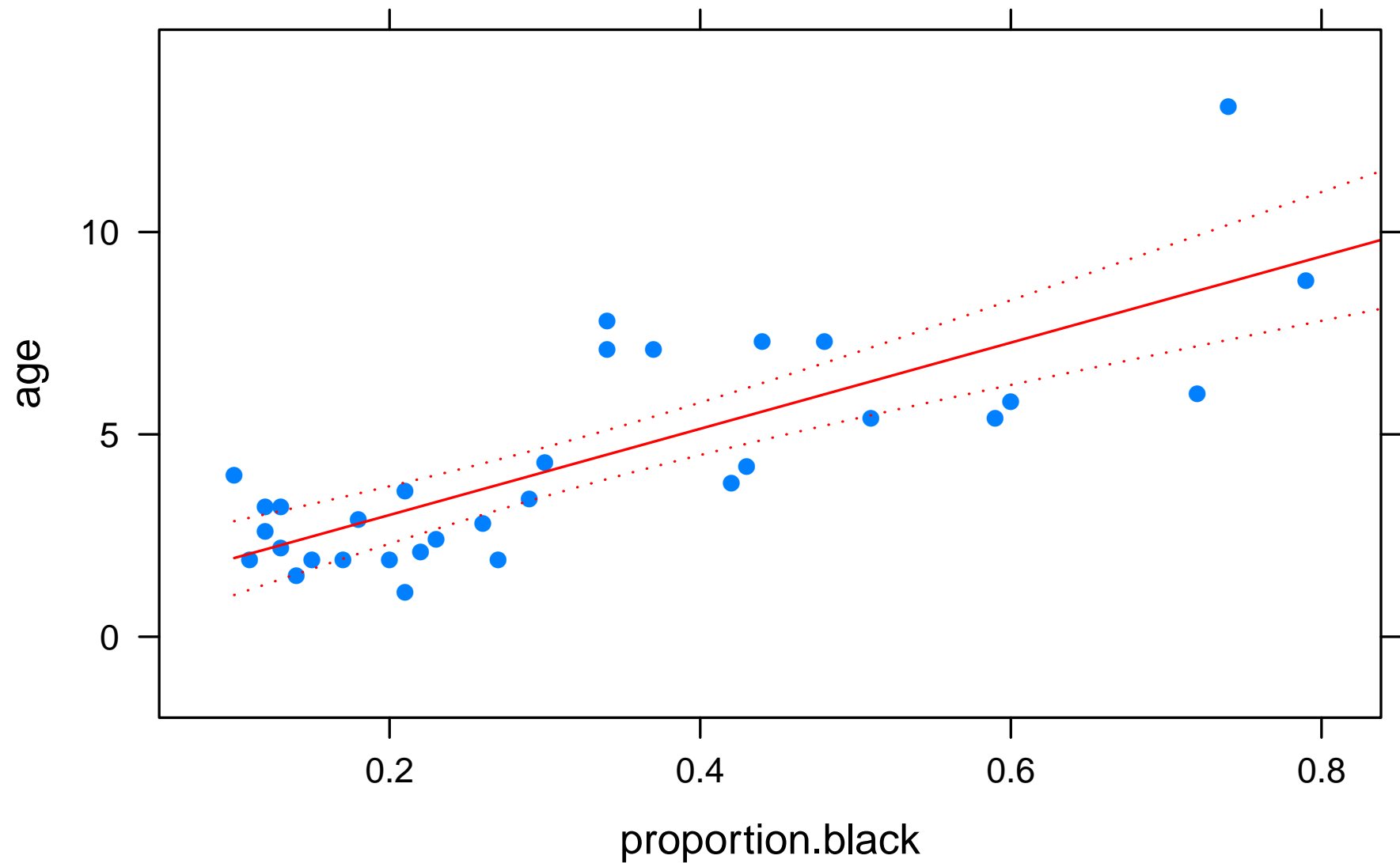
## Example

- The R function `predict()` also can be used for prediction intervals.
- The first argument to `predict()` is a fitted model.
- The second argument is a data frame with the desired explanatory variables.
- The argument `interval="prediction"` specifies a prediction interval.
- Here are 95% intervals for the age of individual lions with either 0.2, 0.3, and 0.4 proportion black noses.

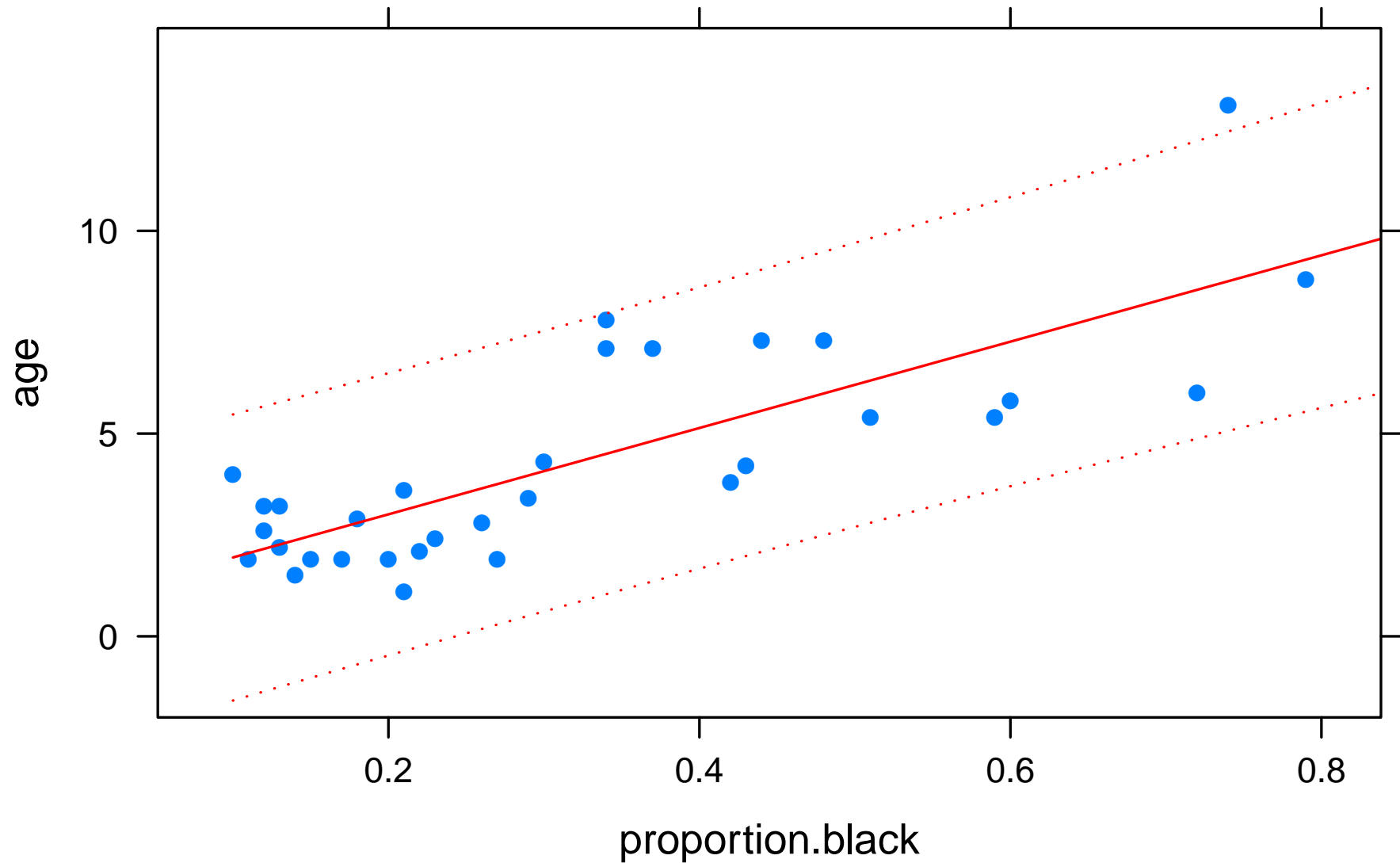
```
> predict(lions.lm, data.frame(proportion.black = c(0.2, 0.3, 0.4)),  
+       level = 0.95, interval = "prediction")
```

	fit	lwr	upr
1	3.008430	-0.4729207	6.489781
2	4.073142	0.6115538	7.534730
3	5.137854	1.6686382	8.607070

# Plot of Confidence Intervals



# Plot of Prediction Intervals



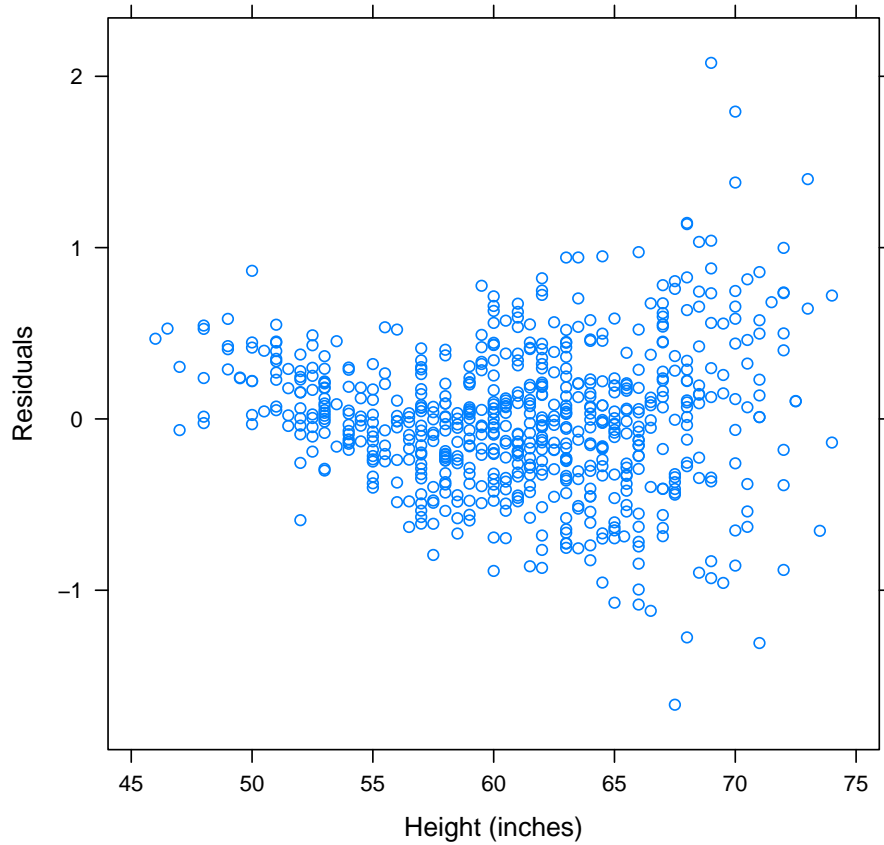
# A New Example

## Example

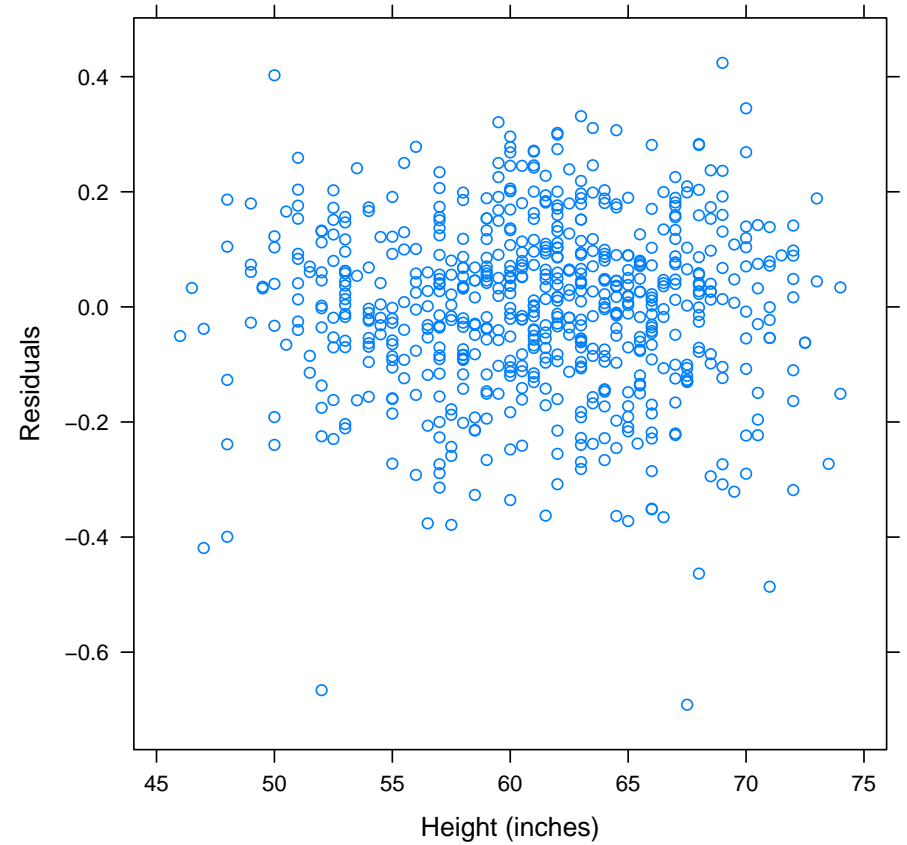
- FEV (forced expiratory volume) is a measure of lung capacity and strength.
- FEV increases as children grow larger.
- The next page has two separate residual plots, one of the residuals versus height from a fit of FEV versus height and one of the residuals versus height from a fit of  $\log(\text{FEV})$  versus height.
- Notice that the residuals in the first plot have a fan shape but those in the second are more evenly spread out. Also, the first plot shows a nonlinear trend that is not present in the second plot.
- When a residual plot shows a fan-shaped pattern with larger residuals for larger values, a log transformation of the response variable often improves the fit of the data to assumptions of a linear model.

# Two Residual Plots

**Original Data**



**Log-Transformed Data**



# Cautions and Concerns

- Inferences to larger populations assume random sampling; be cautious in interpretation when samples are not random.
- Plotting both raw data and residuals after fitting models is essential.
- Many response variables depend on multiple explanatory variables; be aware of the potential of confounding if the model does not include some important explanatory variables.
- Extrapolating beyond the range of the data can be dangerous.
- Do not confuse confidence and prediction intervals.
- Sometimes transformations are useful to better align data with regression assumptions.

# Extensions

- Multiple regression includes multiple explanatory variables;
- Nonlinear regression allows fitting more general curves;
- Spline methods fit curves where the basic form of the curve can change for different values of explanatory variables;
- Linear models allow both quantitative and categorical explanatory models;
- Logistic regression is used for binomial or binary response variables;
- Poisson regression is another example of a generalized linear model in which the response variable is not modeled as normally distributed.

# Summary

- Regression and ANOVA are among the most useful and widely used applications of statistical methods in biology.
- If you need to use these methods in your own research, it is advisable to take at least one additional course that examines these models in much greater depth over an entire semester.
- Statistics 572 with Jun Zhu in Spring 2011 may be the right next course for you.

# Final Exam Information

- The final examination is *Sunday, December 19, at 7:45am* in a room to be determined.
- Both lecture sections take the exam in the same location at the same time.
- There are two hours for the exam.
- The exam will be of similar length and style as the midterm exams.
- Each of the five questions will have multiple parts.
- There will be one question each on ANOVA and regression.
- The other three questions may be form any topics of the course.
- The final exam determines 40% of your semester grade.

Good luck!