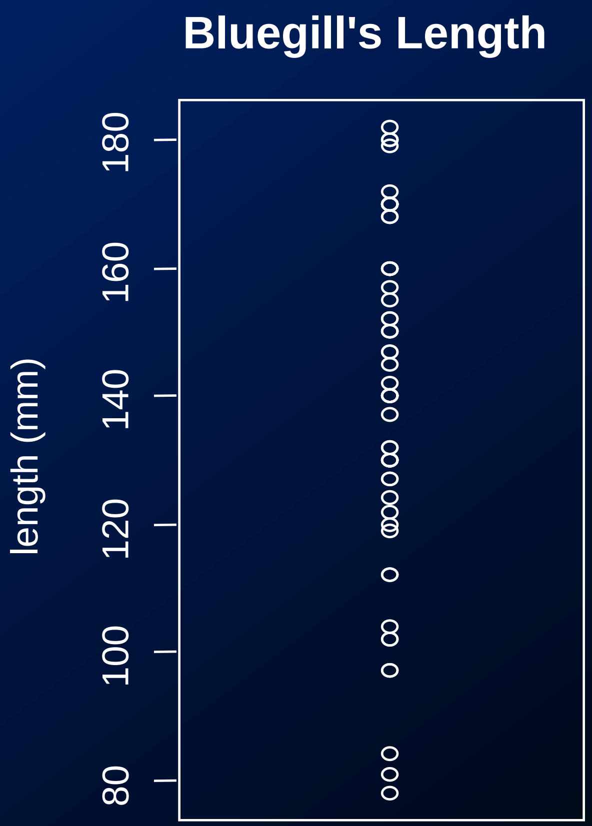
Regression

To reproduce the output install Library s20x

Example: Camp Lake Bluegills

- 66 bluegills were captured from Camp Lake, Minnesota.
- Variables:
  - length (mm).
  - age of the fish (years)
  - radius of a key scale (mm/100)



- We wish to build a model to **predict or explain** the length of bluegills.

Introduction to Regression I

- In the previous section, we looked at linear models of the form:

$$y_i = \mu + \varepsilon_i$$

- We can explain the behaviour of, or predict, *Y* in terms of centre (mean) and spread (variance).

- Camp Lake bluegills:

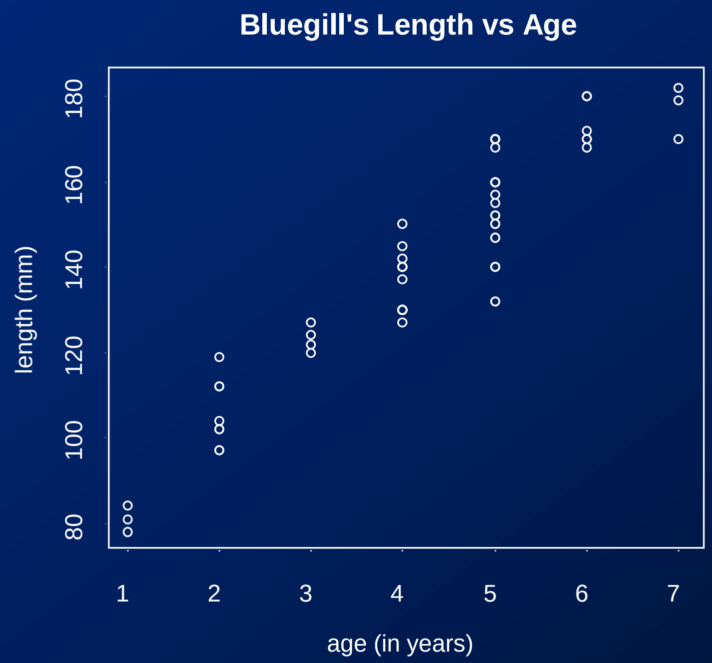
```
Mean value:              142.20
Variance:                679.91
Standard deviation:       26.08
```

Introduction to Regression II

- In predicting *Y*, we are constrained by the **limited information** we have at hand.
  - Our best prediction is our estimate of the population mean, μ, i.e. $\bar{y} = 142.2\ \text{mm}$
  - Everything else is assumed to be due to **random variation** ($\widehat{\text{Var}}(Y) = 679.91\ \text{mm}^2$) about our estimate of the population mean.
- Difference between individuals is just thought of as random variation.

Introduction to Regression III

- We also measured the age of the bluegills.



- We can **make much more accurate predictions** of a fish's expected length, **given that we know its age**.

Introduction to Regression IV

```
> summaryStats(Length~Age,data=camplake.df)
  Sample size     Mean Median   Std Dev Midspread
1           3  81.0000   81.0  3.000000      3.00
2           8 105.6250  103.0  7.909082     11.25
3           4 123.2500  123.0  2.986079      3.25
4          16 135.6875  133.5  6.877197     10.00
5          25 155.2800  157.0 12.204917     21.00
6           7 174.2857  172.0  5.468525     10.00
7           3 177.0000  179.0  6.244998      6.00
```

Introduction to Regression V

- Expected length of the fish (*Y*) **given** its age (*X*):

$$\hat{E}(Y \mid X = 4) = 135.7$$

- Estimate the variance in the lengths of 4-year-old fish:

$$\widehat{Var}(Y \mid X = 4) = 6.877197^2 = 47.3$$

Introduction to Regression VI

- Now our estimates of the mean and variance, given age, are far **more precise**.

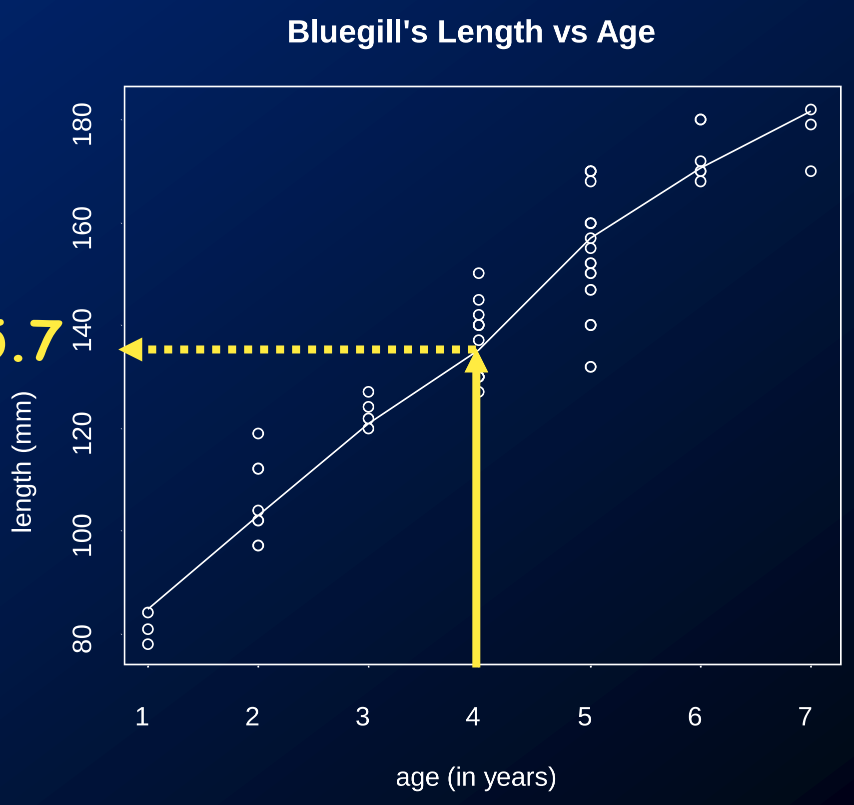
- **Not surprisingly, the more relevant information we have, the more accurate our predictions of bluegill length are likely to be.**

Regression Analysis I

- In regression analysis there are two concepts of interest:
  - $E(Y \mid X) = \mu_{Y|X}$
    and
  - $Var(Y \mid X) = \sigma^2_{Y|X}$

For 4-year-old fish ($X = 4$)

$\hat{E}(Y|X=4)=135.7$

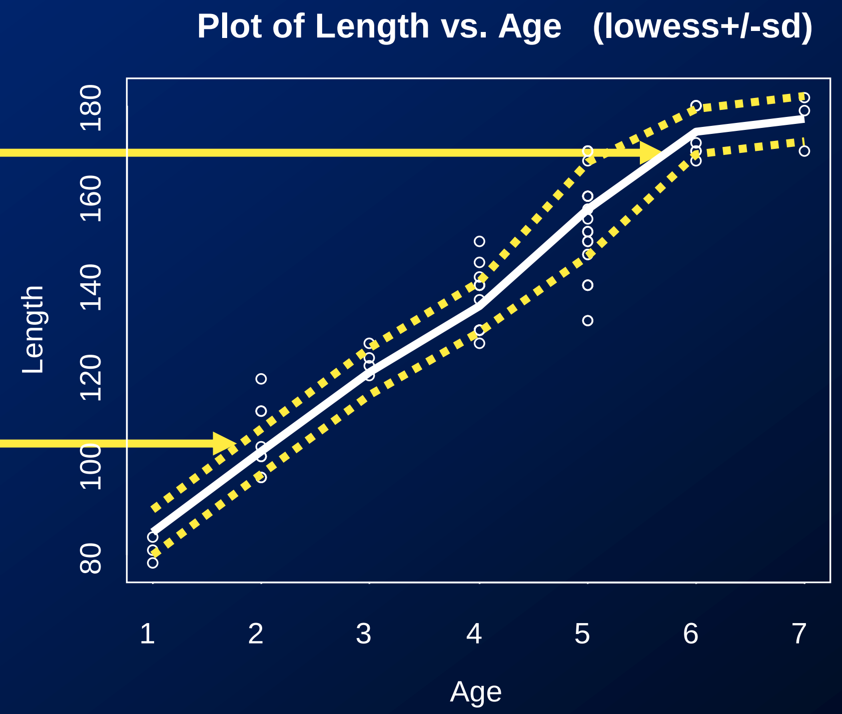


Regression Analysis II

- Var(Y | X), the scatter or variability of our observations, given X.

$\hat{E}(Y|X)$

1 sd bands around $\hat{E}(Y|X)$

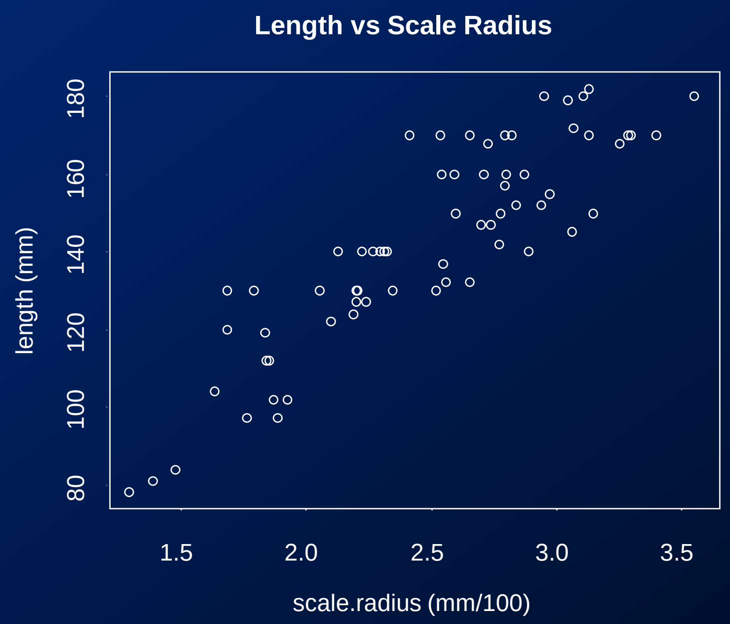


- The plot now gives a visual impression of the scatter, Var(Y | X) as well as a visual impression of the expected value, E(Y | X).

Regression Analysis III

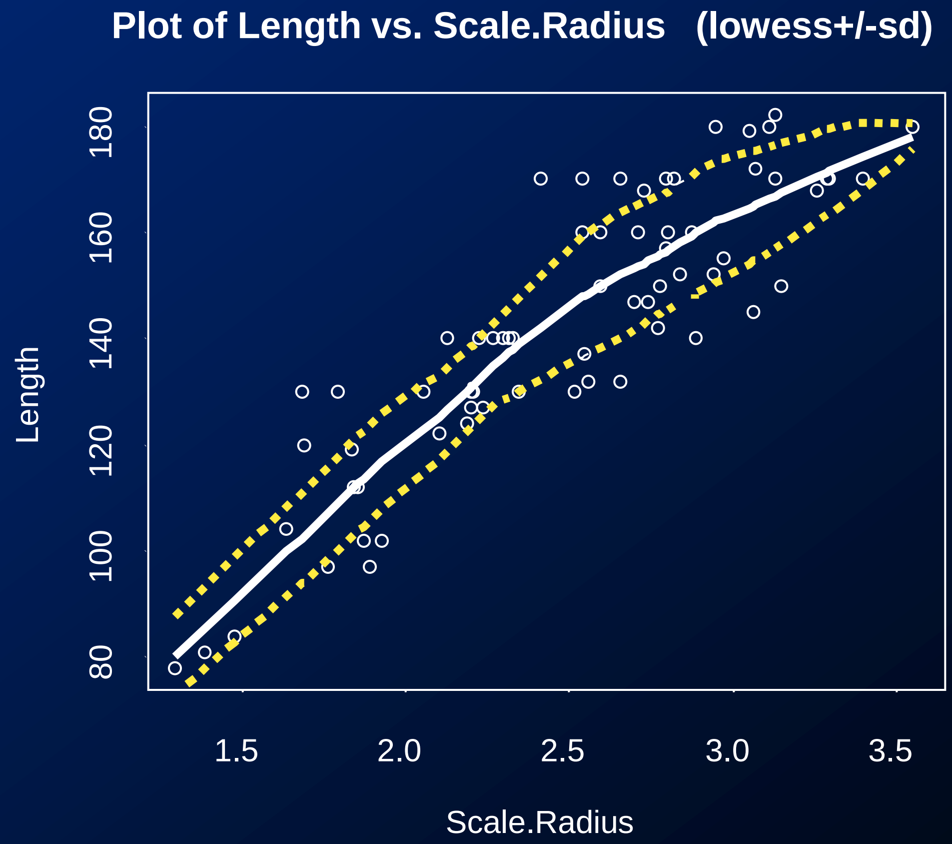
- Let's use **a continuous *X*-variable**, key scale radius, (more accurate measure of age).

- Age: count of rings on the key scale



Regression Analysis IV

- We still focus on the same two concepts:
  - $E(Y \mid X) = \mu_{Y|X}$
    and
  - $Var(Y \mid X) = \sigma^2_{Y|X}$



Regression Analysis V

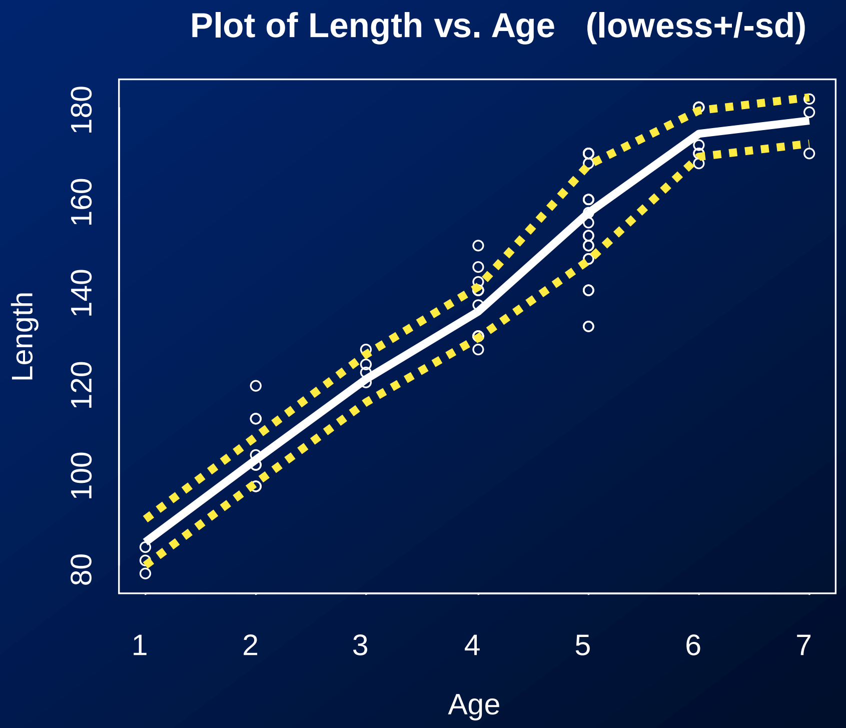
- The Camp Lake data:
  - A (near) linear relationship between age (or scale radius) and length.
  - Variance was (relatively) constant.

- This type of **linear relationship with constant scatter** was discussed in the lab **Simple Linear Regression**.

- However, not all data are reasonably well behaved!!

Trends I

- We can visualise or estimate E(Y | X):
  - By eye.
  - By using smoothers.
  - By fitting parametric curves.



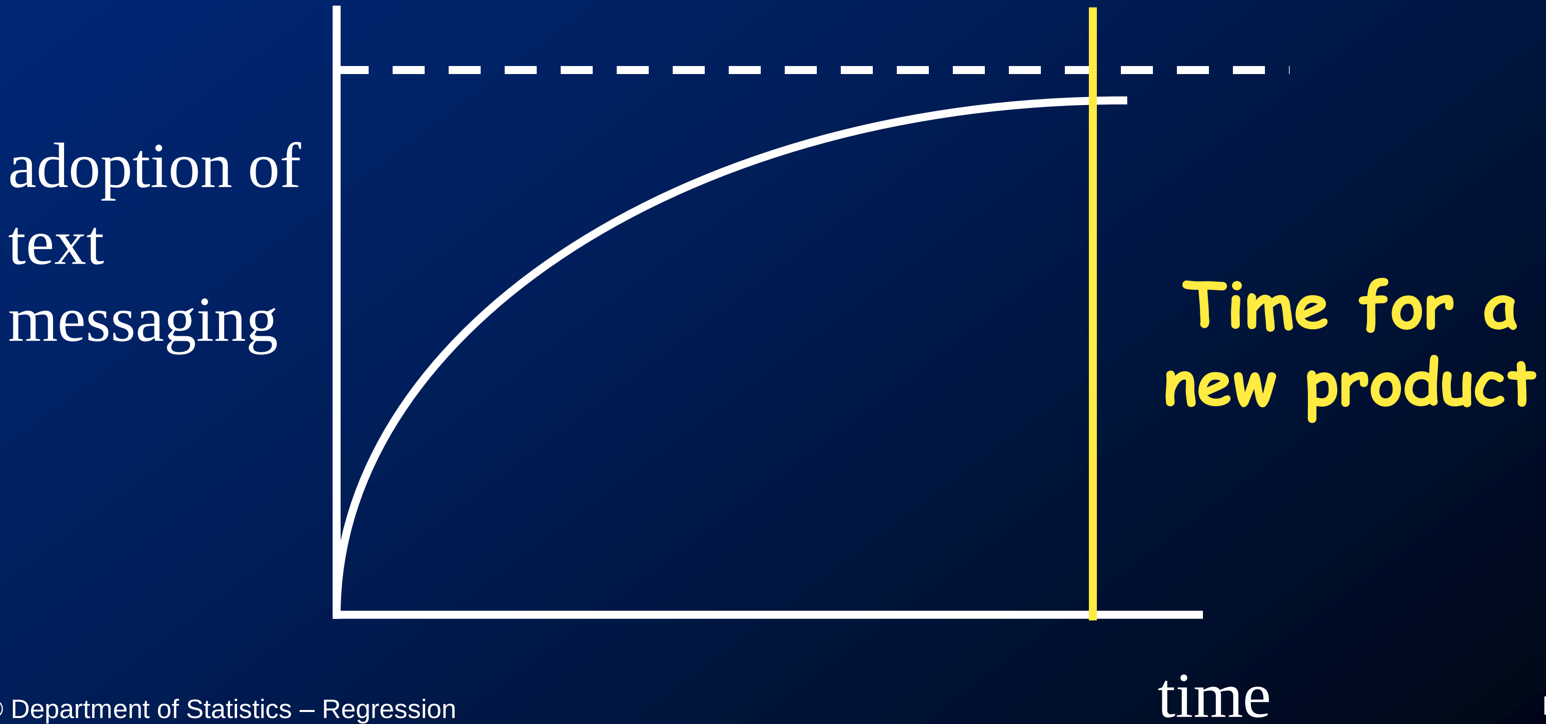
Trends II

- The **trend summarises the main pattern** we see in the scatter plot.
  - Trend enables us to make predictions.
  - The **trend is the expected, or average, value of the response variable, *Y*, given *X*,** E(Y | X).

- Not all trends are linear. Non-linear trends can still summarise data and be used for prediction.
  - However, **interpreting a non-linear trend model may be difficult**.

Trends III

- The shape of a non-linear trend can suggest questions.
- Marketers may plot the adoption of text messaging against time:



Trends IV

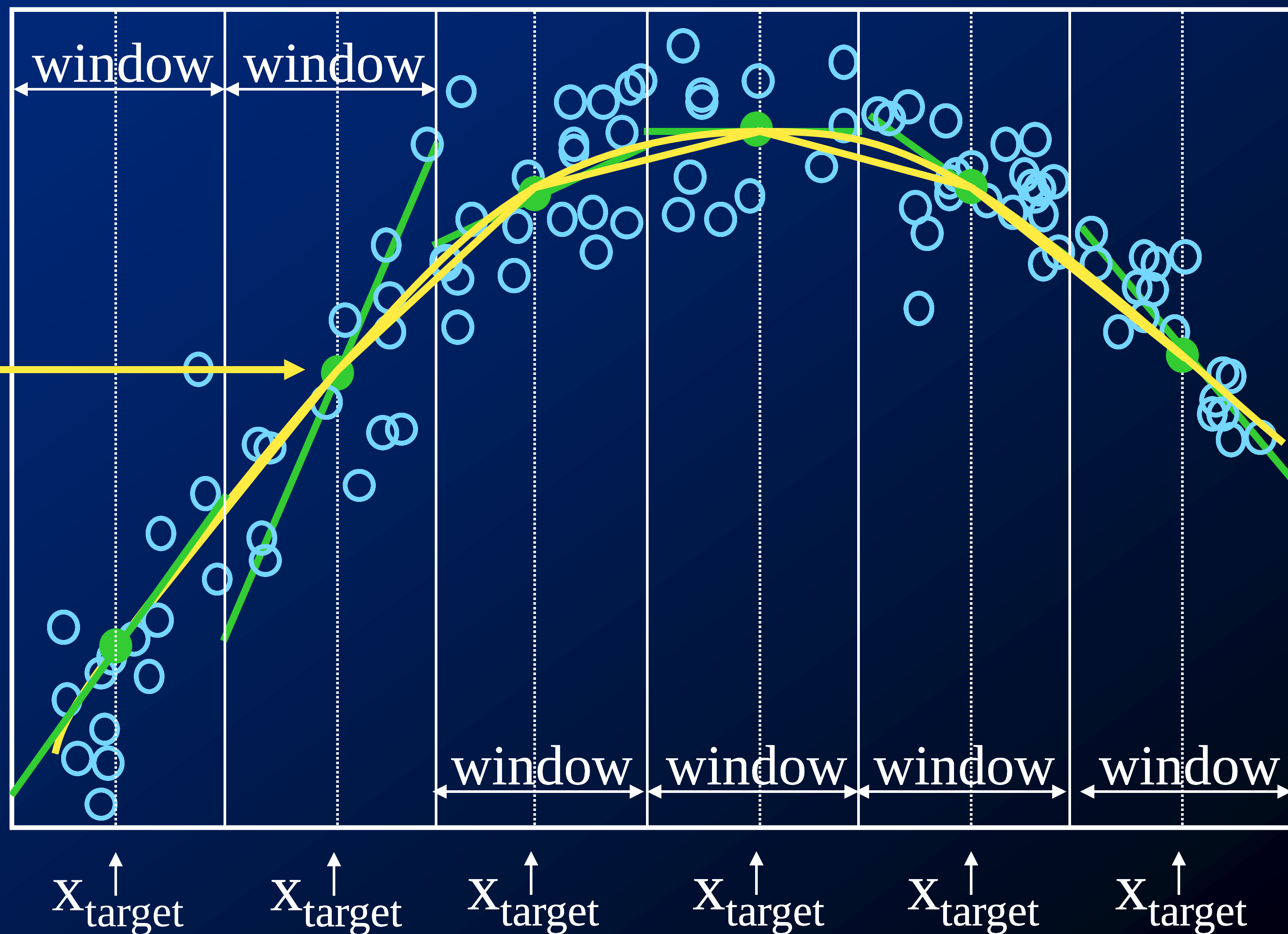
- Assessing trends by eye may suggest a suitable ***parametric family*** to fit to the data (e.g. a quadratic, a cubic, etc).

- Or we can use **a smoother** to get an indication of which ***parametric family*** may be appropriate.
  - A **smoother traces out the main pattern** in a scatter plot or in a residual plot.
  - We will use a **lowess smoother**.
  - We need reasonably **large sample ($n \geq 50$)**.

Smoothers I

- Tracing out a lowess smoother:

Point on the fitted line at the target X-value



Smoothers II

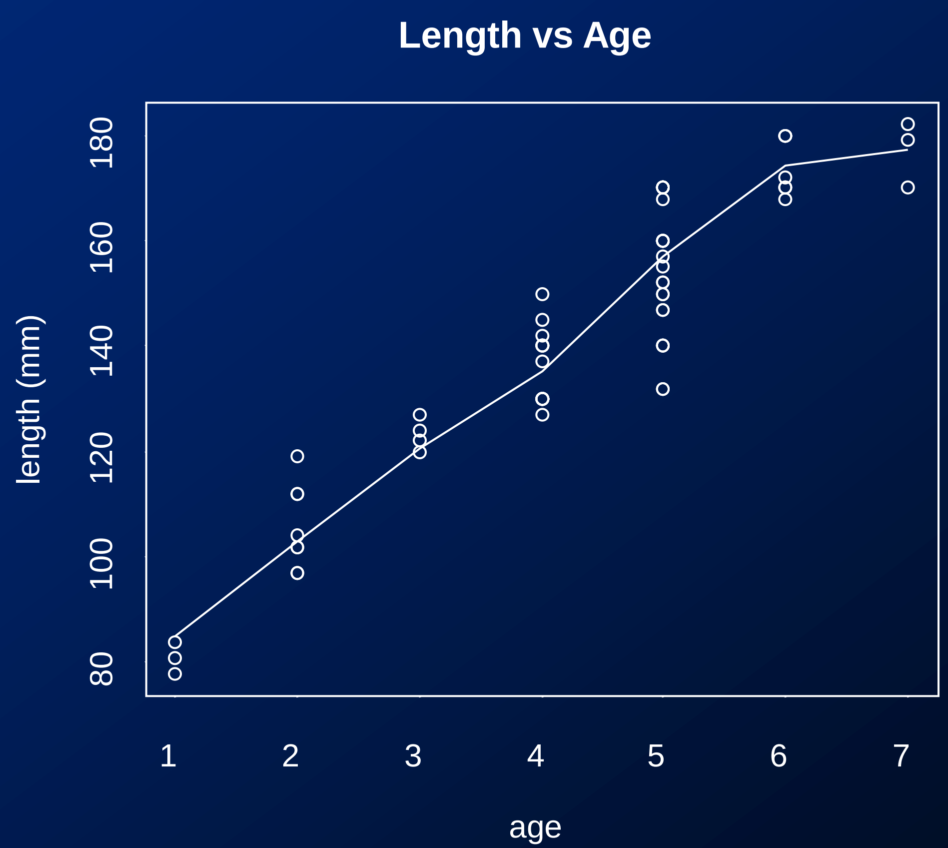
- We can put a lowess smoother on a scatter plot (or residual plot) using the following commands in **R**:

```
plot(y~x,data=data.df)
lines(lowess(data.df$x,data.df$y,f=2/3))
```

  - f is the lowess smoothing constant (0 < f < 1).
  - The larger f, the smoother the curve.
  - We select f by **trial and error**.

Smoothers III

```
> plot(Length~Age,main="Length vs Age",xlab="age",
   ylab="length (mm)",data=camplake.df)
> lines(lowess(camplake.df$Age,camplake.df$Length,f=0.5))
```

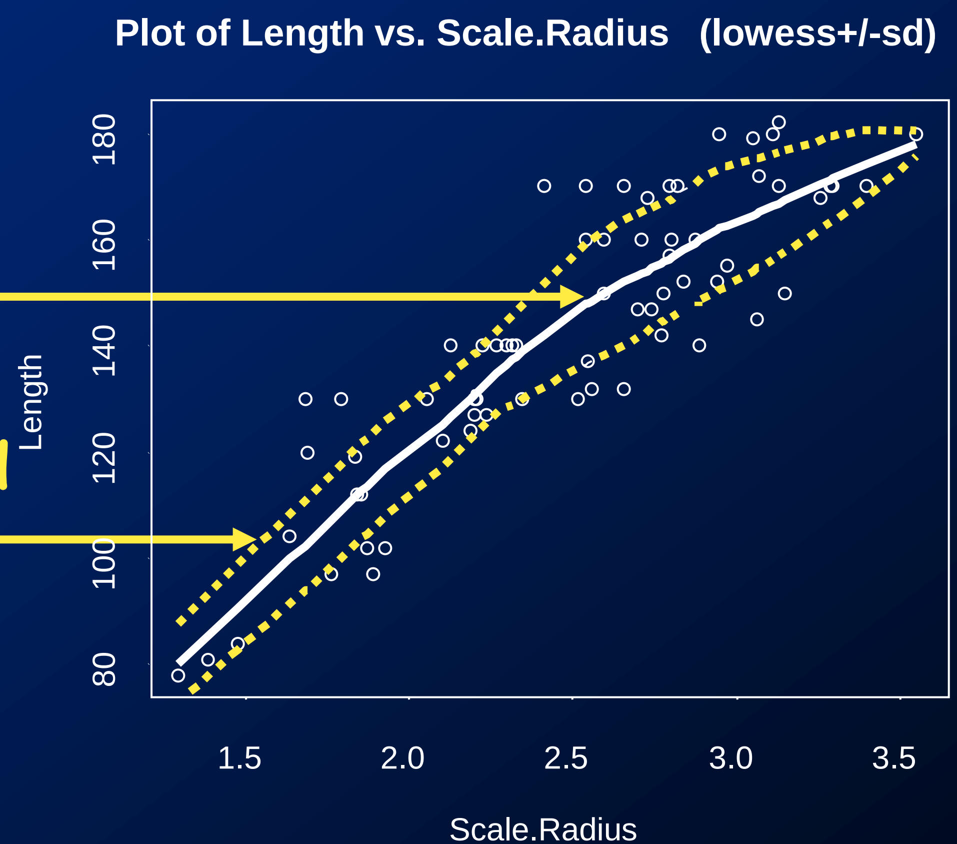


Smoothers IV

```
> trendscatter(Length~Scale.Radius,f=0.5,data=camplake.df)
```

$\hat{E}(Y|X)$

1 sd bands around $\hat{E}(Y|X)$



Smoothers V

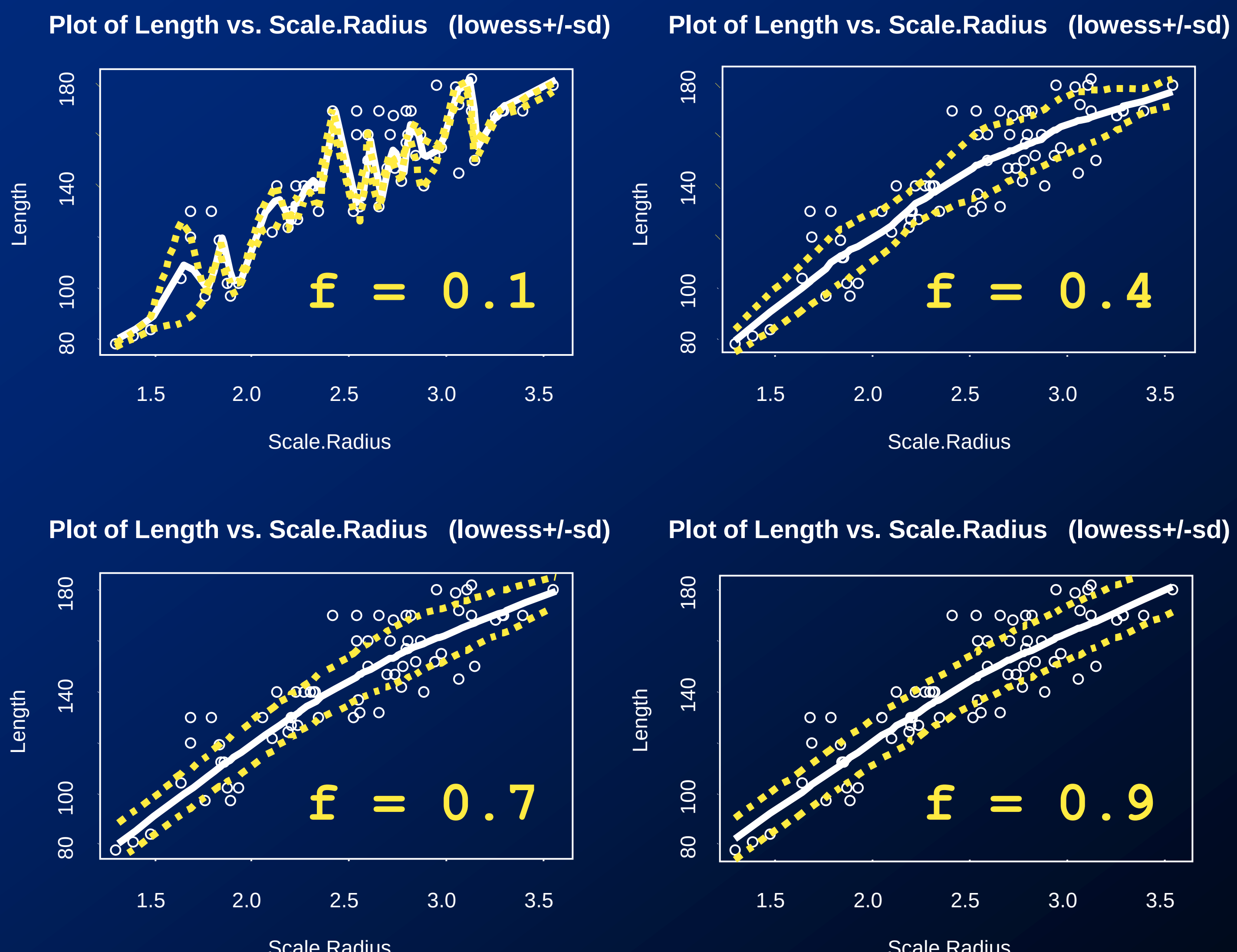
- **In R:** `trendscatter(y~x,f=0.5,data=`***`data.df`***`)`

```
> trendscatter(Length~Scale.Radius,f=0.1,data=camplake.df)

> trendscatter(Length~Scale.Radius,f=0.4,data=camplake.df)

> trendscatter(Length~Scale.Radius,f=0.7,data=camplake.df)

> trendscatter(Length~Scale.Radius,f=0.9,data=camplake.df)
```

Smoothers VI

Not smooth

Better



Looks good

Too much?

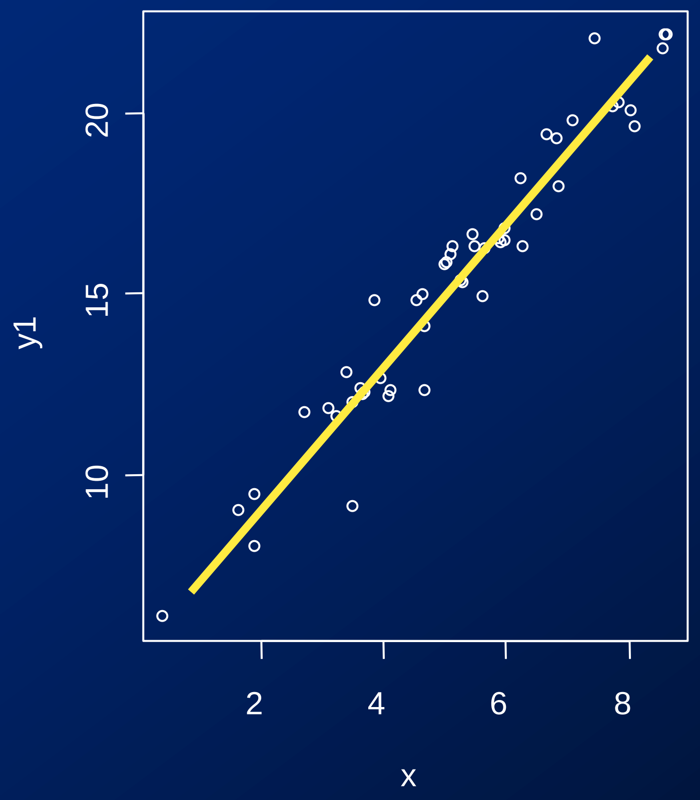
Smoothers VII

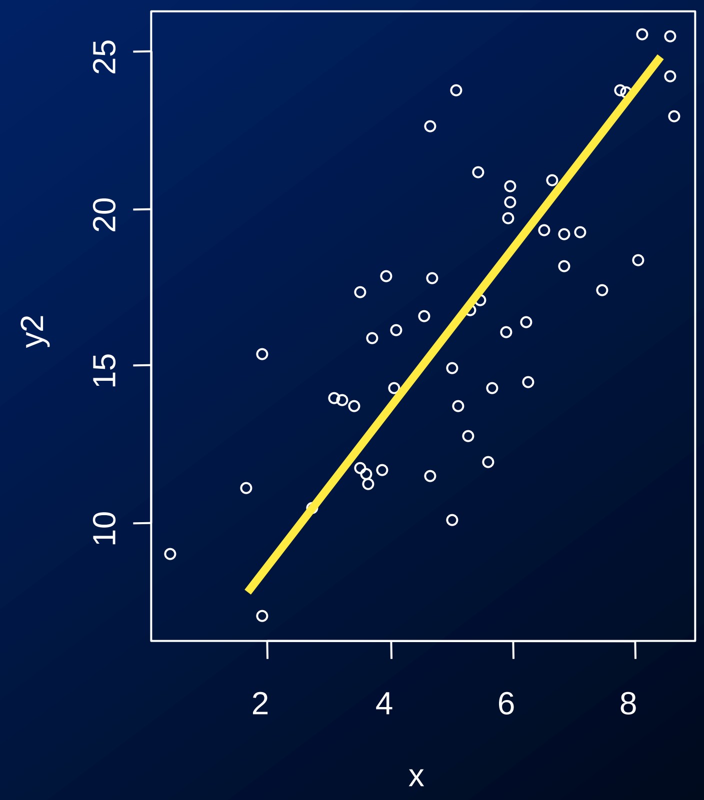
- **Note: We want to use as small a value of the smoothing parameter, f as possible**

- but **we also want to trace out the pattern(s) in the data.**

Scatter I

- The amount of **scatter about the trend** determines **how well the chosen model fits the data**.



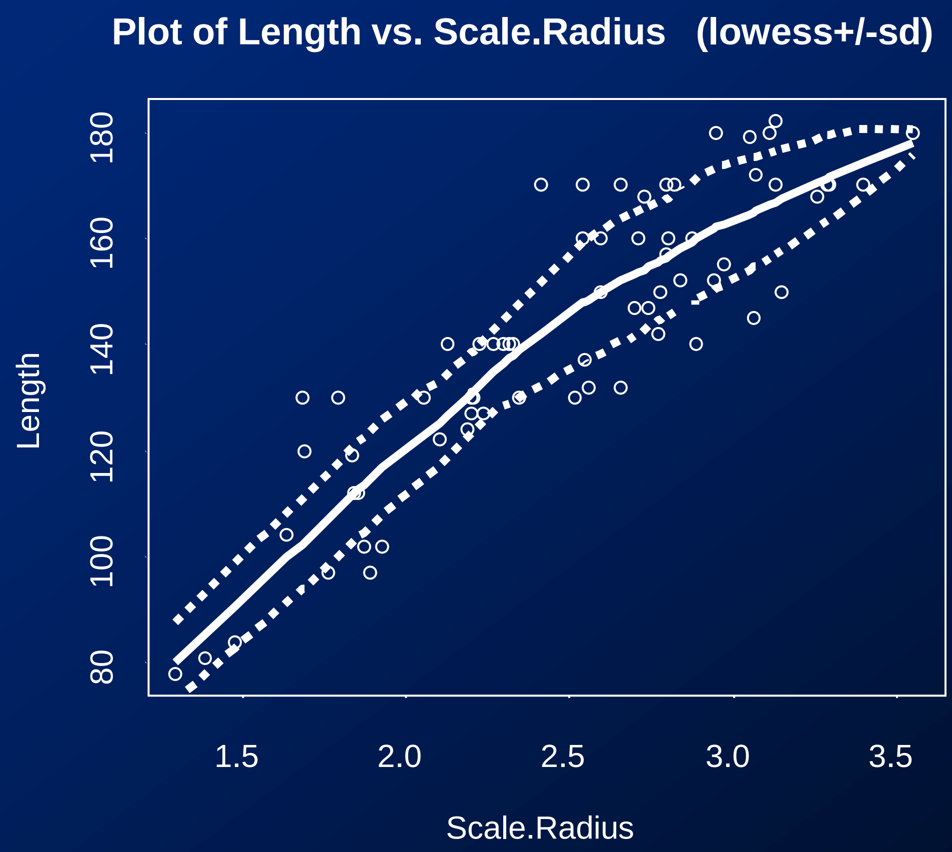


Scatter II

- We require that the scatter is constant throughout the trend.

- When the scatter is constant, our estimation technique (least squares) is reliable in the sense that all the points have the same influence in estimating the trend, $E(Y \mid X)$.

Camp Lake Bluegills I

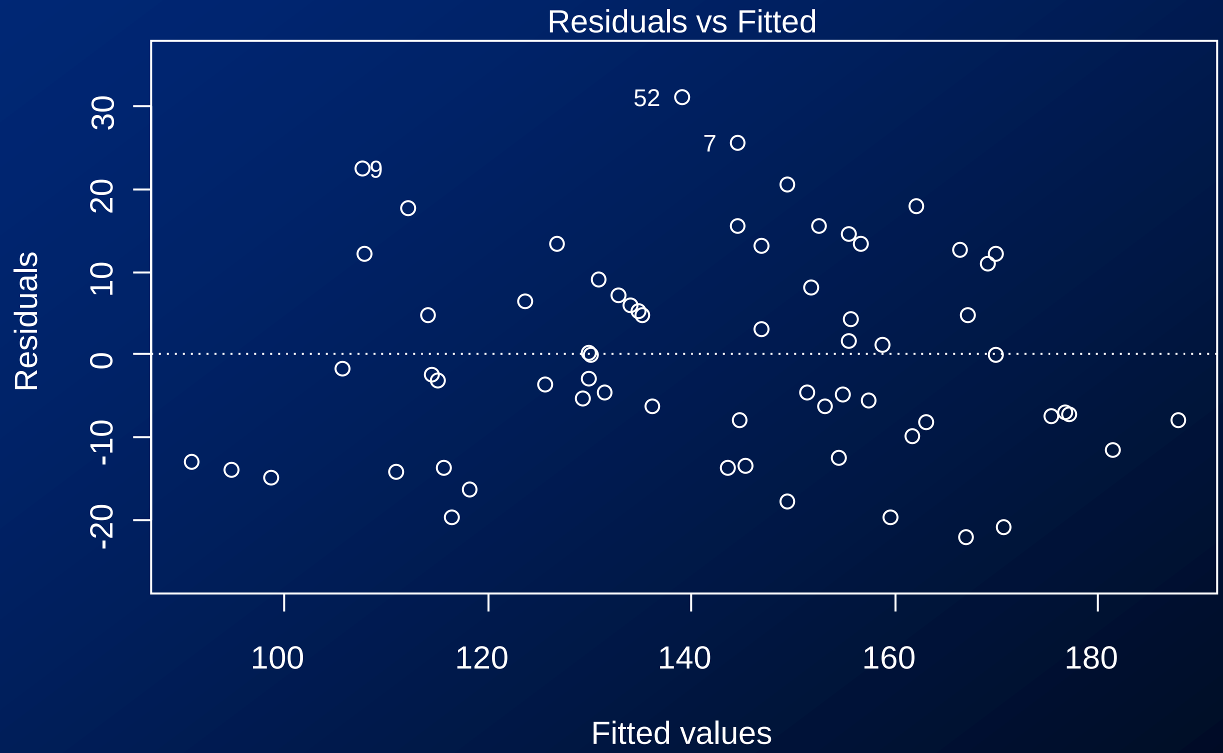


Camp Lake Bluegills II

```
> bluegill.fit<-lm(Length~Scale.Radius,data=camplake.df)
```

response ~ explanatory

```
> eovcheck(bluegill.fit)
```



Camp Lake Bluegills III

5 number summary of the residuals

```
> summary(bluegill.fit)
Call:
lm(formula = Length ~ Scale.Radius, data = camplake.df)
Residuals:
    Min       1Q   Median       3Q      Max
-21.928  -8.041  -1.941    8.946   31.028
Coefficients:
```

$\hat{\beta}_j$ $\quad$ $\text{se}(\hat{\beta}_j)$ $\quad$ $H_0 : \beta_j = 0$

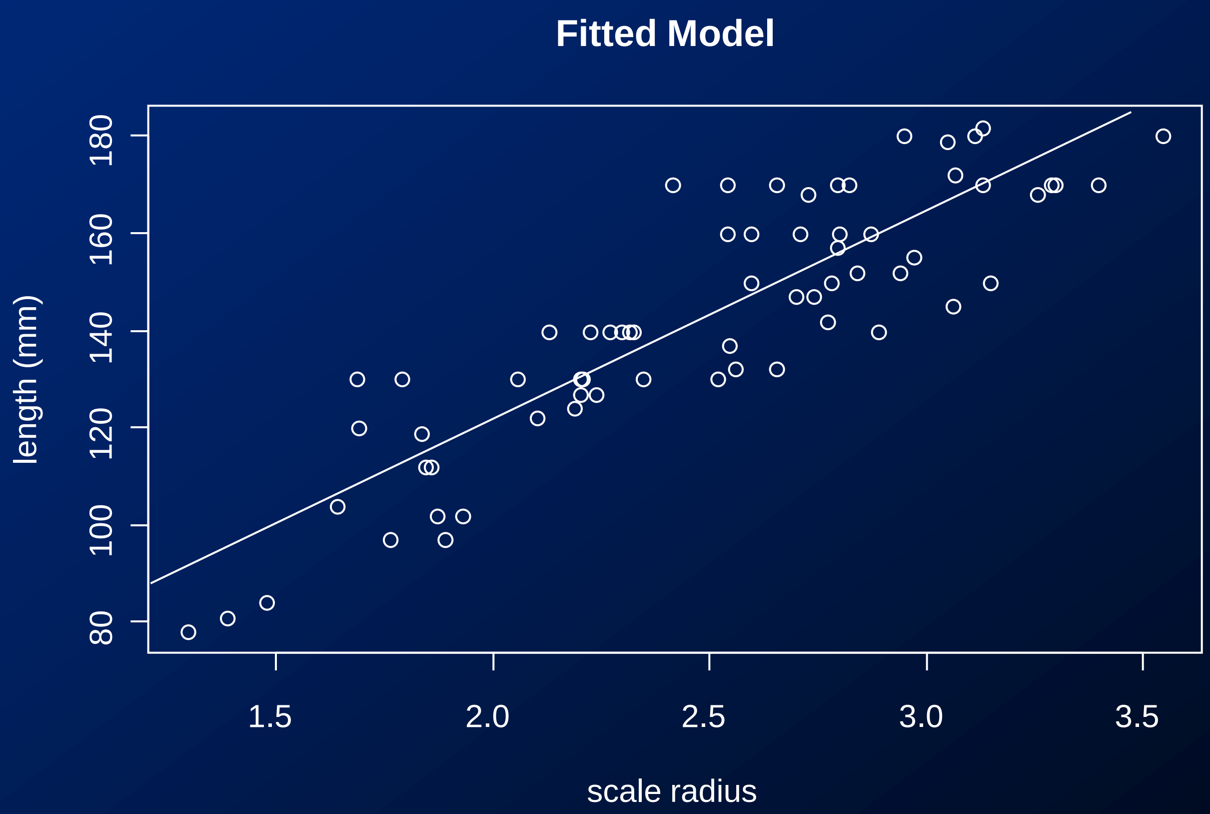
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    34.920      7.356   4.747 1.20e-05 ***
Scale.Radius   43.126      2.893  14.908  < 2e-16 ***
---
Signif. codes: 0'***'0.001'**'0.01'*'0.05'.'0.1' '1
```

$\hat{\sigma}$ $\quad$ $n-2$

```
Residual standard error: 12.43 on 64 degrees of freedom
Multiple R-squared: 0.7764,   Adjusted R-squared: 0.7729
F-statistic: 222.3 on 1 and 64 DF,  p-value: < 2.2e-16
```

Camp Lake Bluegills IV

```
> plot(Length~Scale.Radius,main="Fitted Model",
  xlab="scale radius",ylab="length (mm)",
  data=bluegill.df)
> abline(bluegill.fit)
```



Does the model fit very well?

A) Yes

B) No

C) I don't know

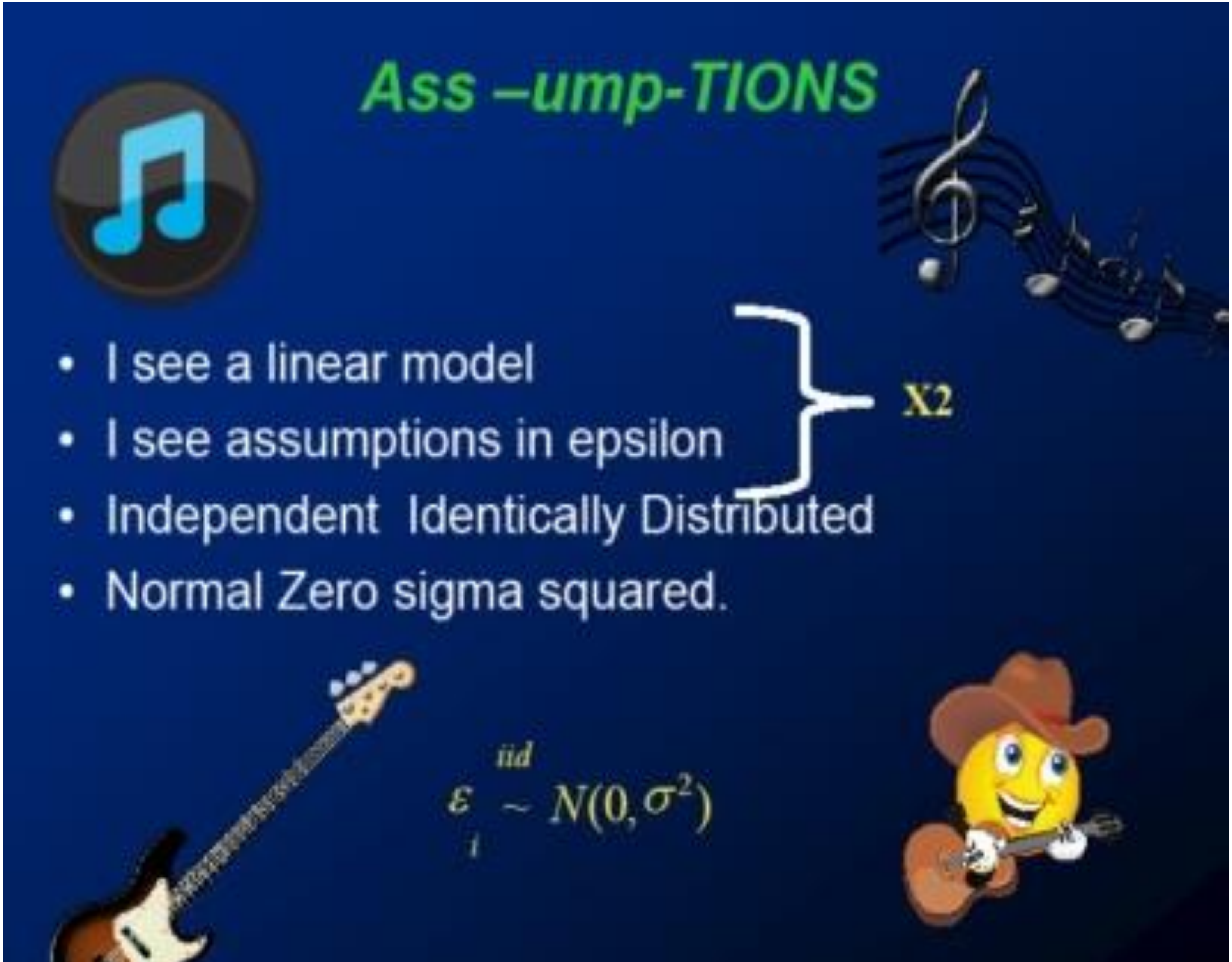
D) Yes, reasonably well

ASS-ump-TIONS

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

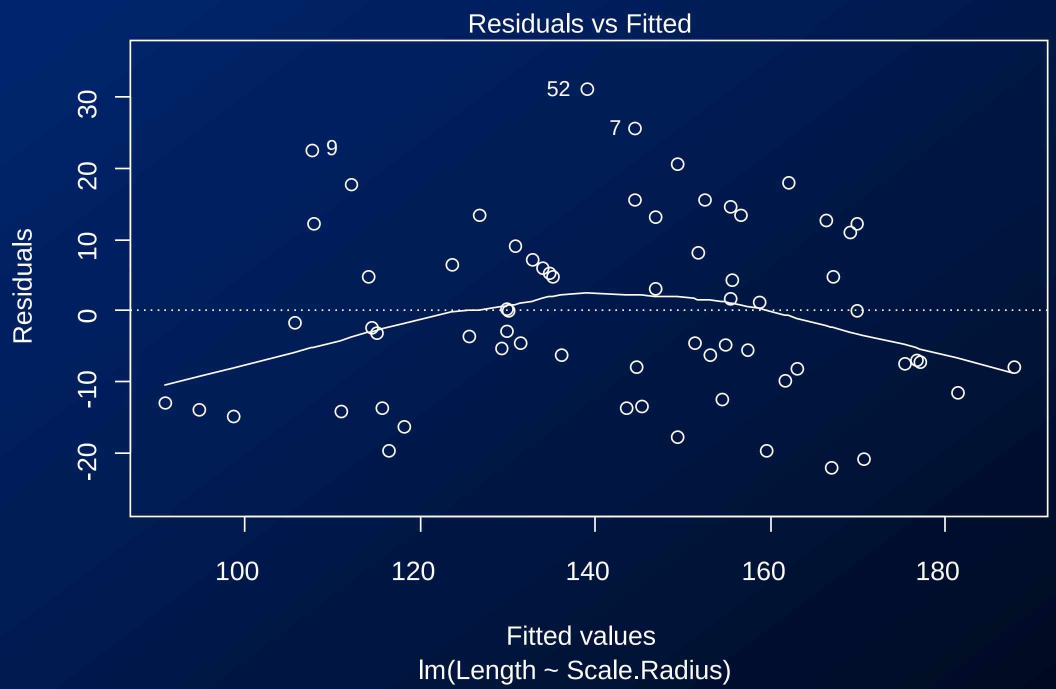
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\mu_i = \beta_0 + \beta_1 x_i$$


Ass –ump-TIONS
• I see a linear model
• I see assumptions in epsilon
X2
• Independent Identically Distributed
• Normal Zero sigma squared.
$\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

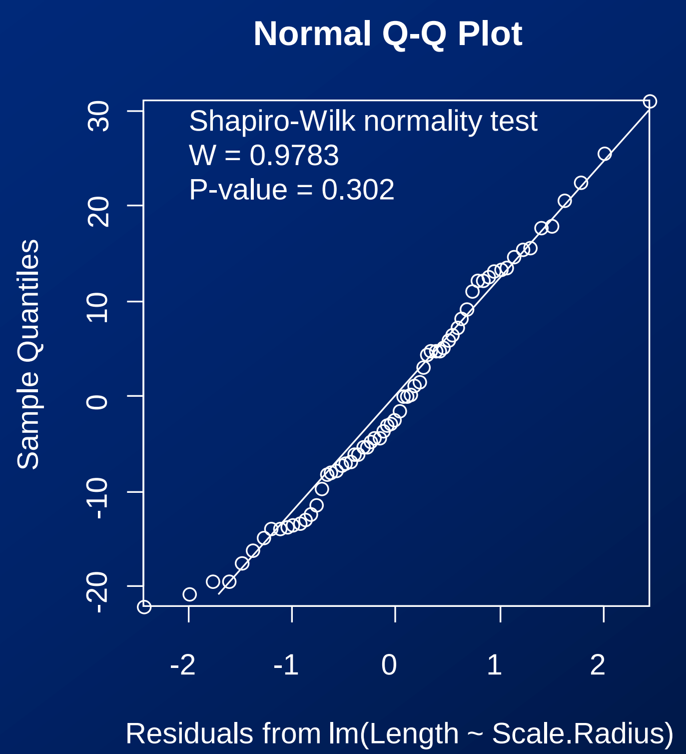
Camp Lake Bluegills V

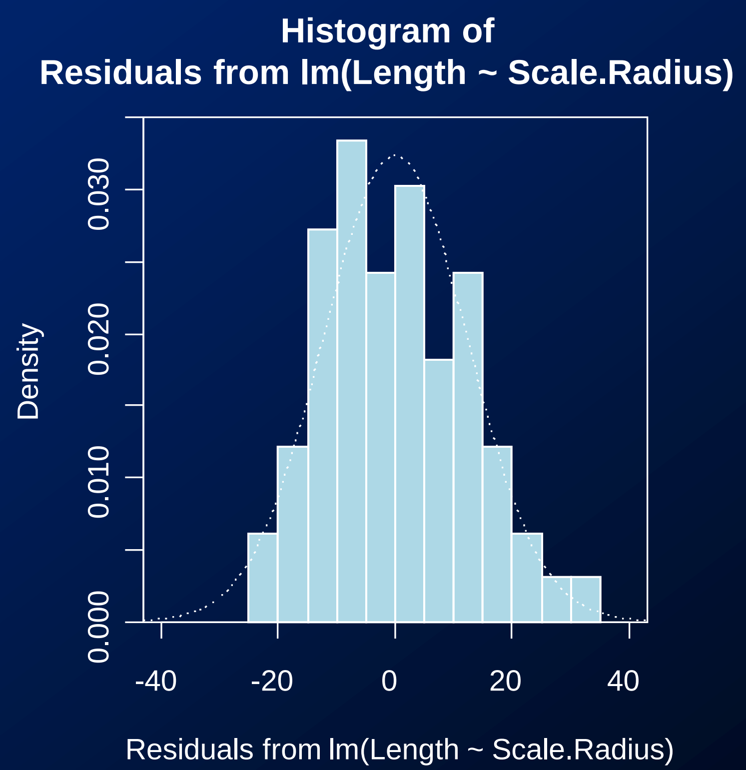
```
> ciReg(bluegill.fit)
               95 % C.I.lower    95 % C.I.upper
(Intercept)          20.22412          49.61655
Scale.Radius         37.34749          48.90540

> plot(bluegill.fit,which=1)
```



Camp Lake Bluegills VI

```
> normcheck(bluegill.fit)
```





Any problem with normality?

A) No

B) Yes

C) Maybe

Simple Linear Regression I

- Regression uses information in ***explanatory variables*** to predict/explain a ***response variable***.

- Response (dependent, endogeneous) variable :
  - Usually denoted Y, is random and is (usually) continuous.

- Explanatory (predictor, independent, exogenous, carrier, covariate) variable:
  - Usually denoted X, and may be continuous or discrete or a factor (See: Blocks 7 & 8).

Simple Linear Regression II

- Regression models:
  - **Identify and model all the patterns (or structure)**.
  - Everything that is left over, we model as being completely random.

Simple Linear Regression III

- Thus our modelling framework is:

trend and scatter:

$$\underbrace{y_i}_{\text{Observed value}} = \underbrace{E(y_i \mid x_i)}_{\text{Fitted or Predicted value}} + \underbrace{\varepsilon_i}_{\text{Error}}$$

What we observe

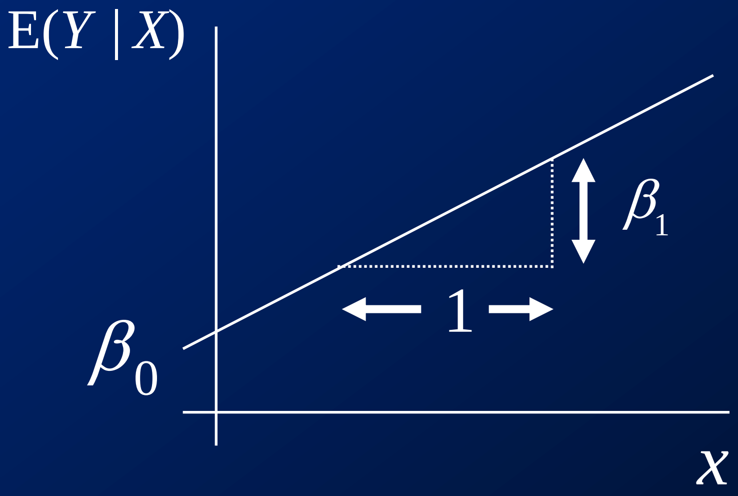
What we expect to see

Difference (or deviation) between what we observe and what we expect to see

Simple Linear Regression IV

- A straight line relationship for the trend is:

$$E(Y \mid X) = \beta_0 + \beta_1 x$$



Simple Linear Regression V

- Our ***Simple Linear Regression*** model is:

$$y_i = \underbrace{E(y_i|x_i)} + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Data estimates $\quad y_i = \underbrace{\hat{E}(y_i|x_i)} + r_i$

$$y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_i} + r_i$$

$$\underbrace{y_i} \quad = \quad \underbrace{\hat{y}_i} \quad + \quad \underbrace{r_i}$$

What we observe

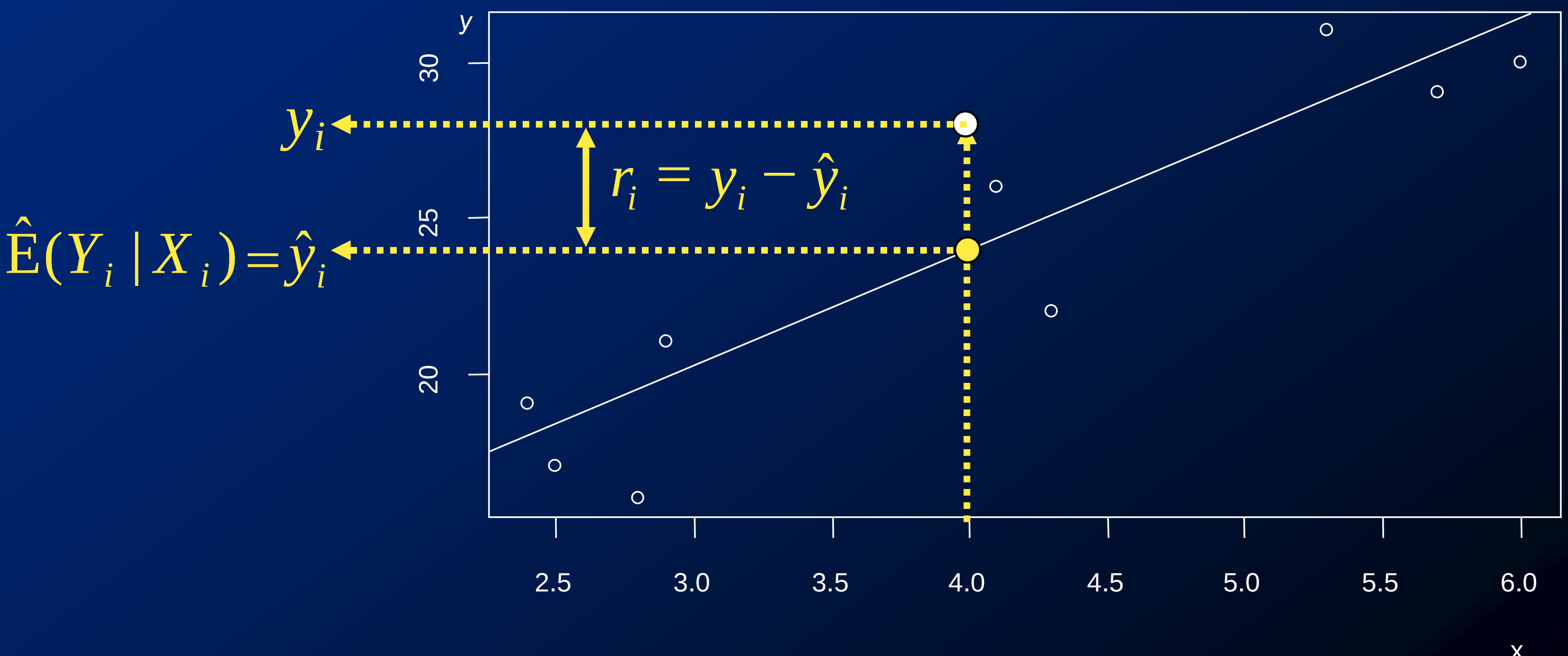
What we expect to see

Difference (or deviation) between what we observe and what we expect to see

Simple Linear Regression VI

- The residual, r_i:

$$r_i = y_i - \hat{y}_i$$



Simple Linear Regression VII

- We find estimates for β_0 and β_1 such that **the sum of the squared residuals (RSS) is minimised**.

- We minimise: $$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- The equation above can be written as:

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- The estimates found in this way are called the ***Least Squares* estimates**.

Regression Model Assumptions I

- The **trend** we have used for our model **is correct**.

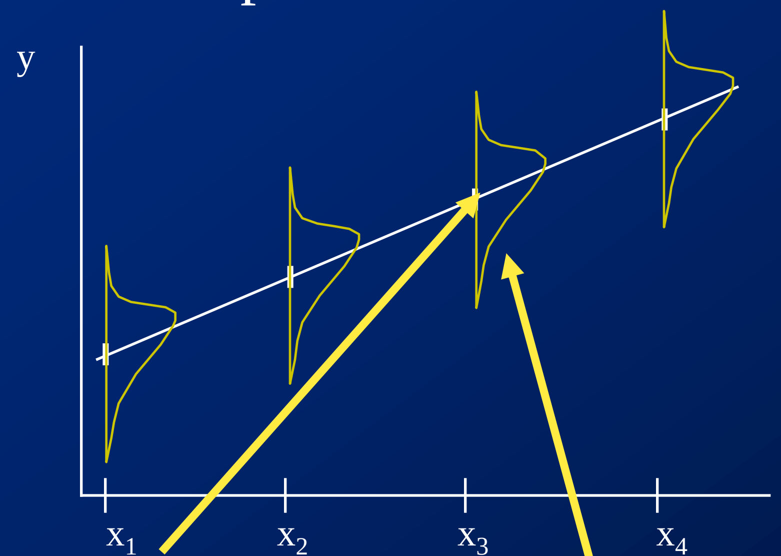
- The other assumptions of the regression model concern the **random scatter** about the trend.

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

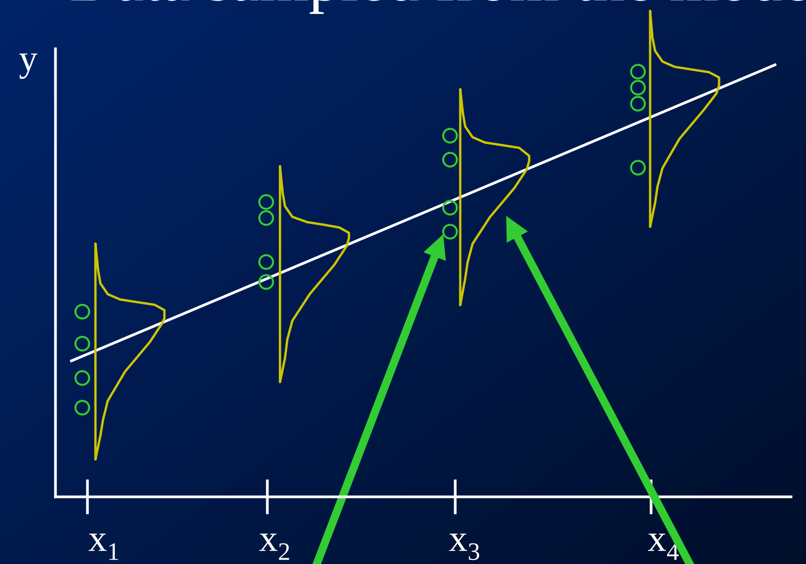
Regression Model Assumptions II

- Diagrammatic views of the Model and Sample.

Simple Linear Model



Data sampled from the model



Expected value of y given $X = x_3$ $E(Y|X=x_3)$

Normal curves perpendicular to the page with means lying on the trend

Observed y – values at x_3

y_i come from this distribution

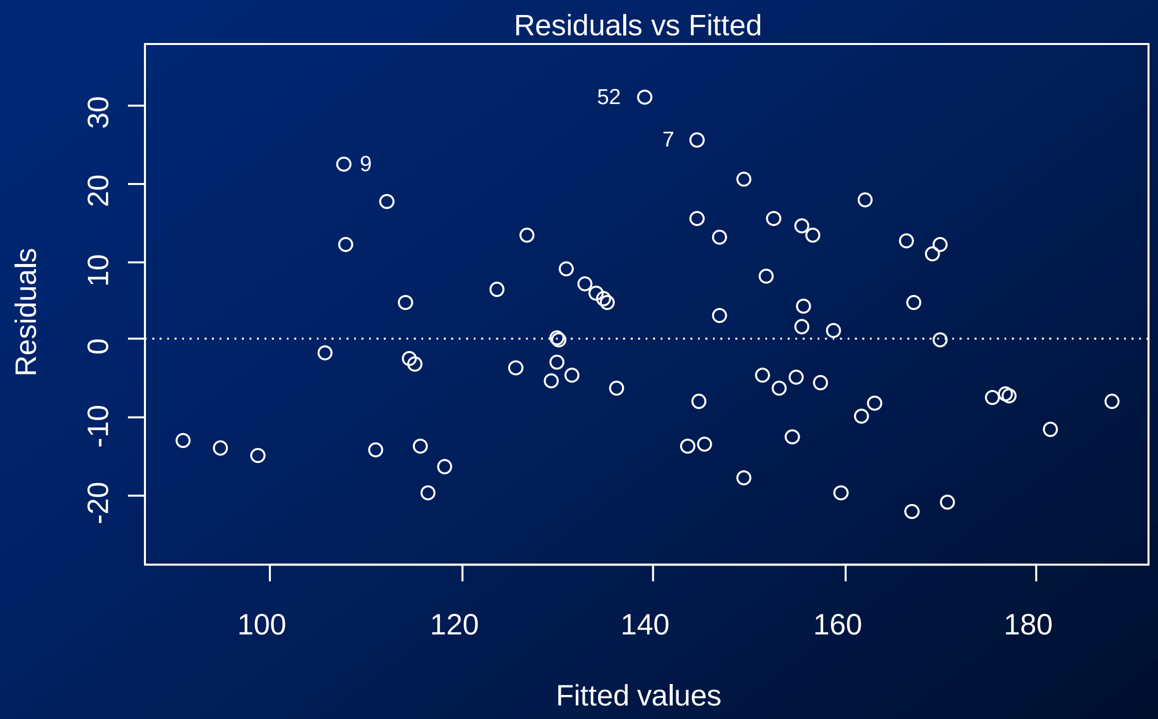
Diagnostics I

- The independence assumption is, as usual, assessed by a careful examination of how the data were collected, or by examining the design of the experiment that produced the data.

- The next most important assumption is that we have **constant variance** (scatter) of the residuals about the fitted model.

Diagnostics II

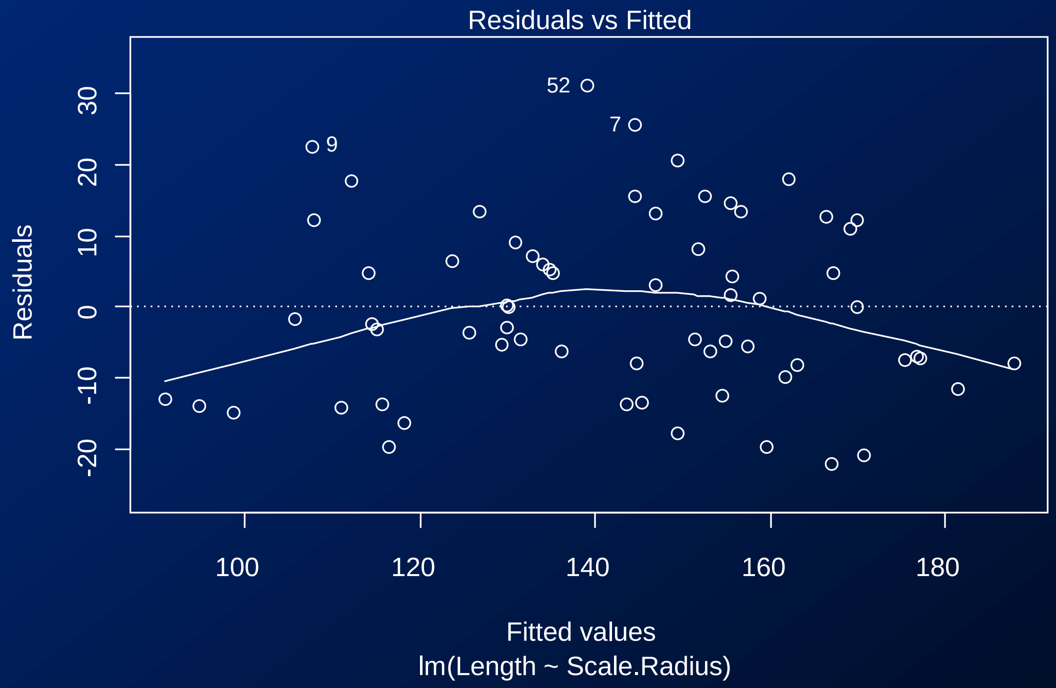
```
> eovcheck(bluegill.fit)
```



Diagnostics III

- Next, we check that the relationship is linear.

```
> plot(bluegill.fit,which=1)
```

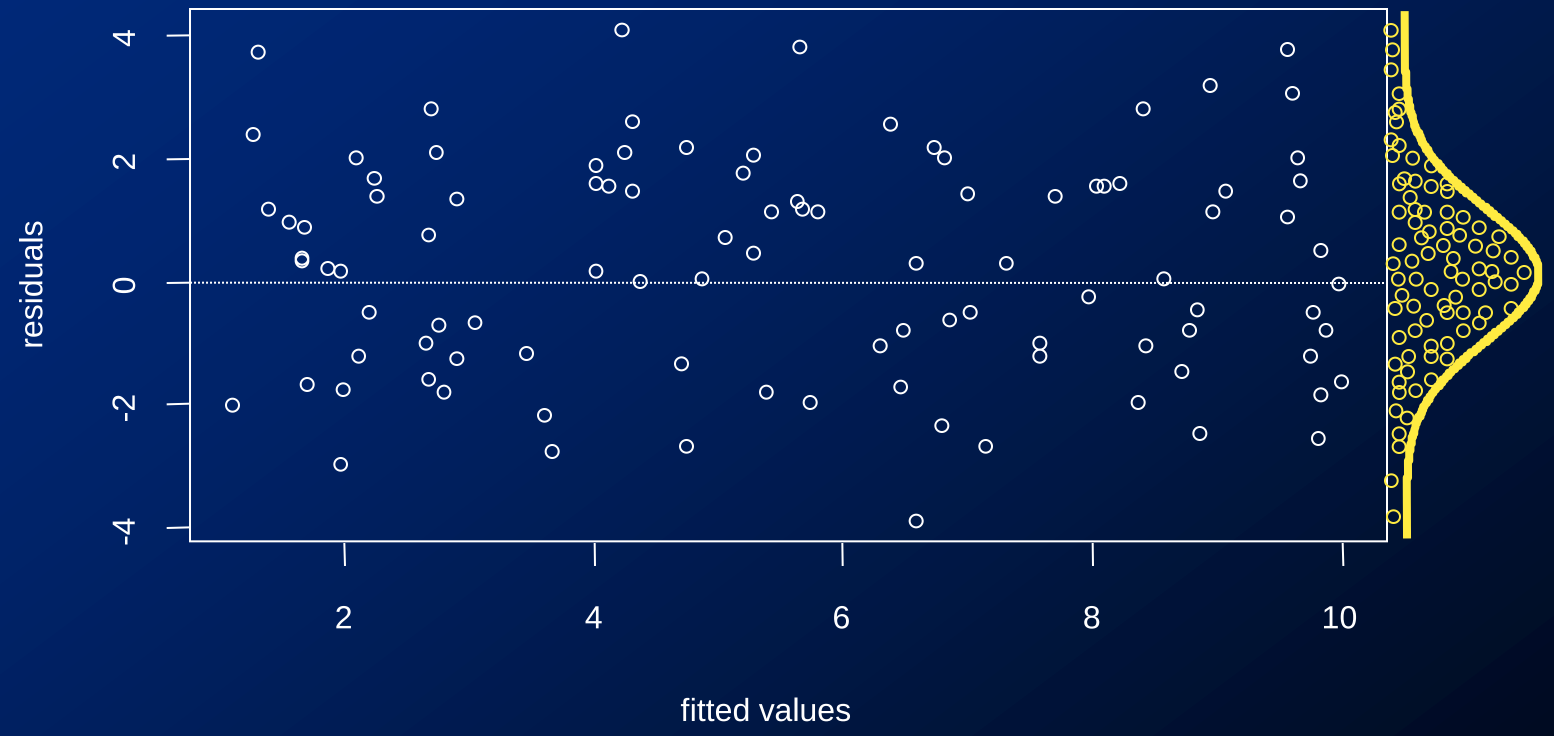


Diagnostics IV

- The **ideal residual plot is a *patternless horizontal band* of points**, centred at 0.
  - Why?
  - If we have successfully modelled all of the patterns in the data, there will be no pattern left in the residuals.

Diagnostics V

- An ideal residual plot:



Diagnostics VI

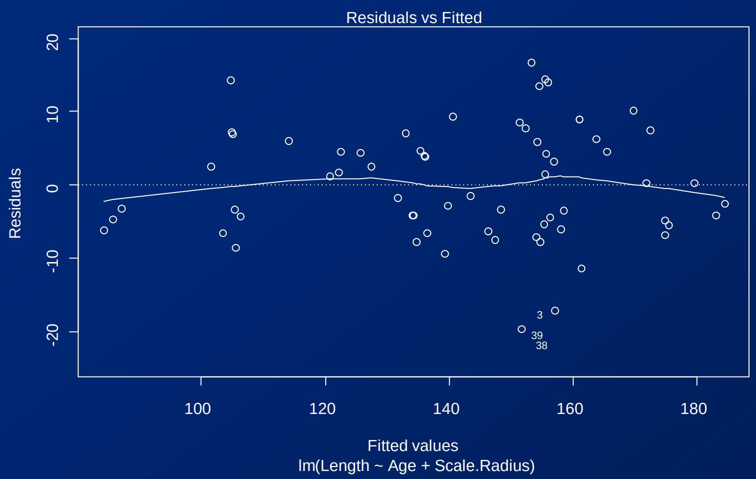
- **If we detect patterns in the residual plot, it means there are further patterns in the data that have not been captured in the model**.

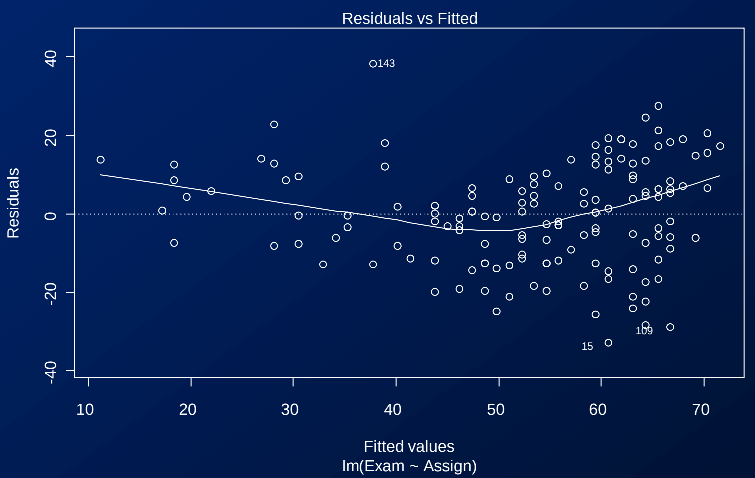
- If problems occur, we need to go beyond the Simple Linear Regression model.

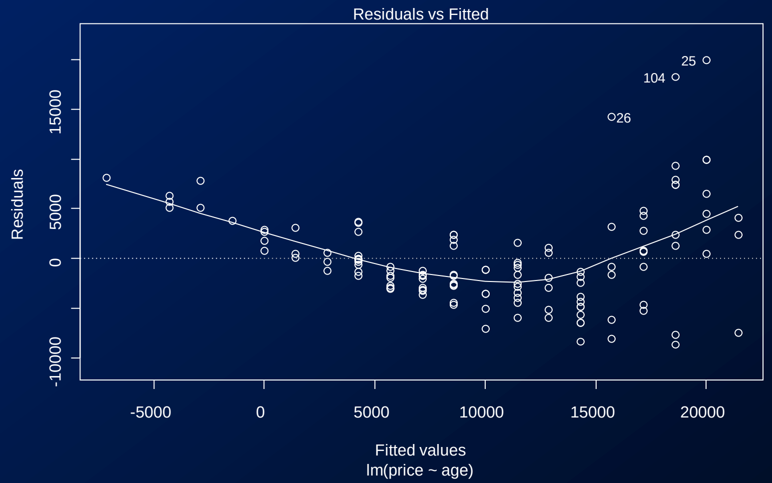
  (See: Block 5)

Diagnostics VII

- Examples of "**unsatisfactory**" residual plots:



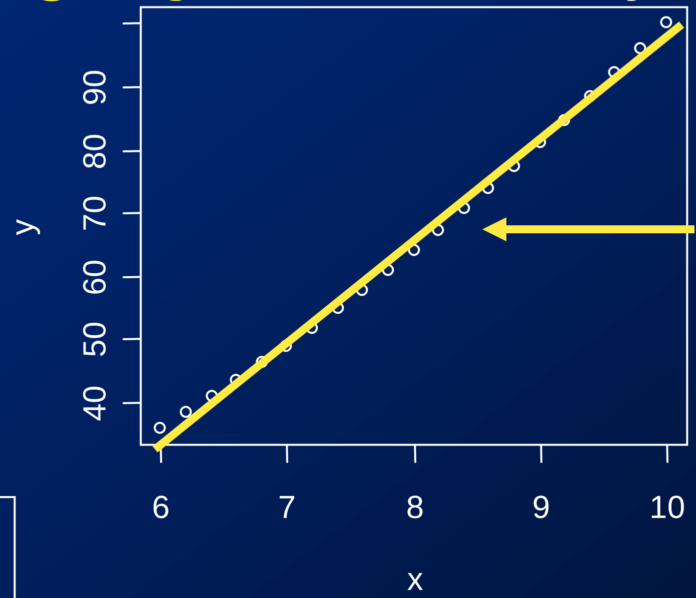




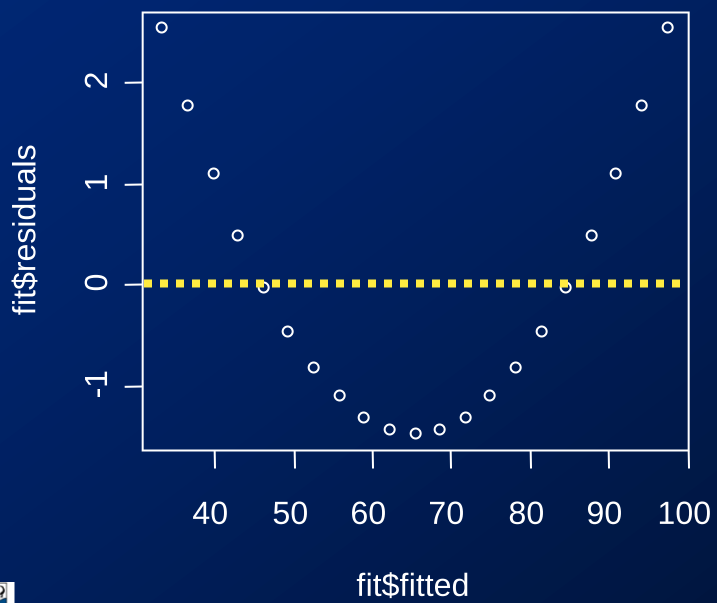
Diagnostics VIII

- Residual plots "**magnify**" non-linearity:

The trend looks linear in the scatter plot of y vs x



Scatter plot is of $y = x^2$



The residual plot from a linear fit shows the non-linearity much more clearly

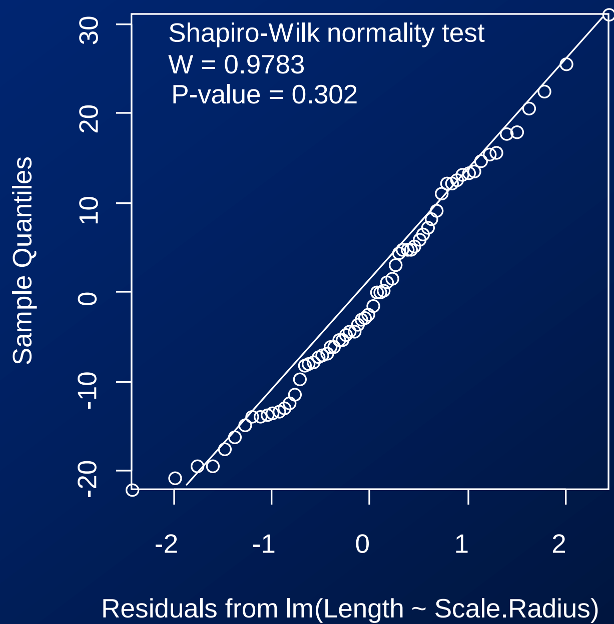
Diagnostics IX

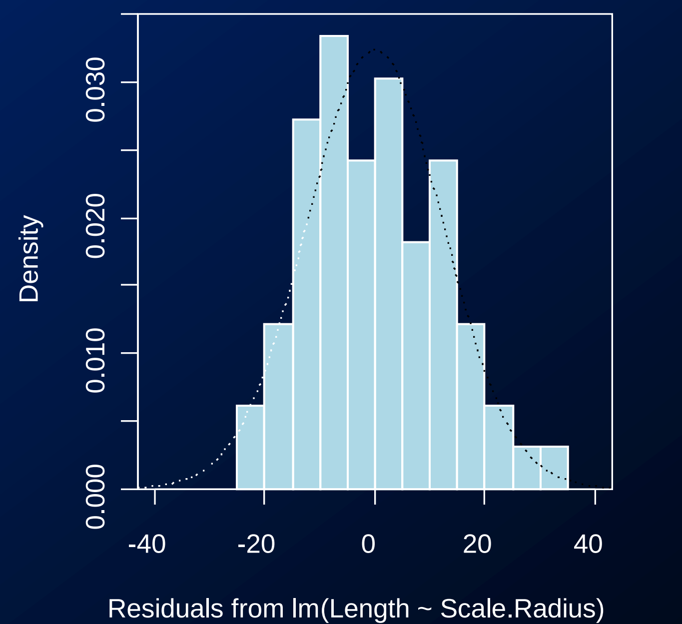
- **Always do a residual plot to assess the true extent of the scatter.**
- Another thing we need to be careful with is assessing scatter about a non-linear trend.
  - Scatter often looks smaller than it really is when the curve is steep.
  - Scatter often looks larger than it really is when the curve is flat.
- The key to assessing scatter with any trend is to make sure you **assess the scatter vertically**, not horizontally!!

Diagnostics X

- Lastly, we check whether the residuals could have come from a **normal distribution** as a check on the normality of the population errors.

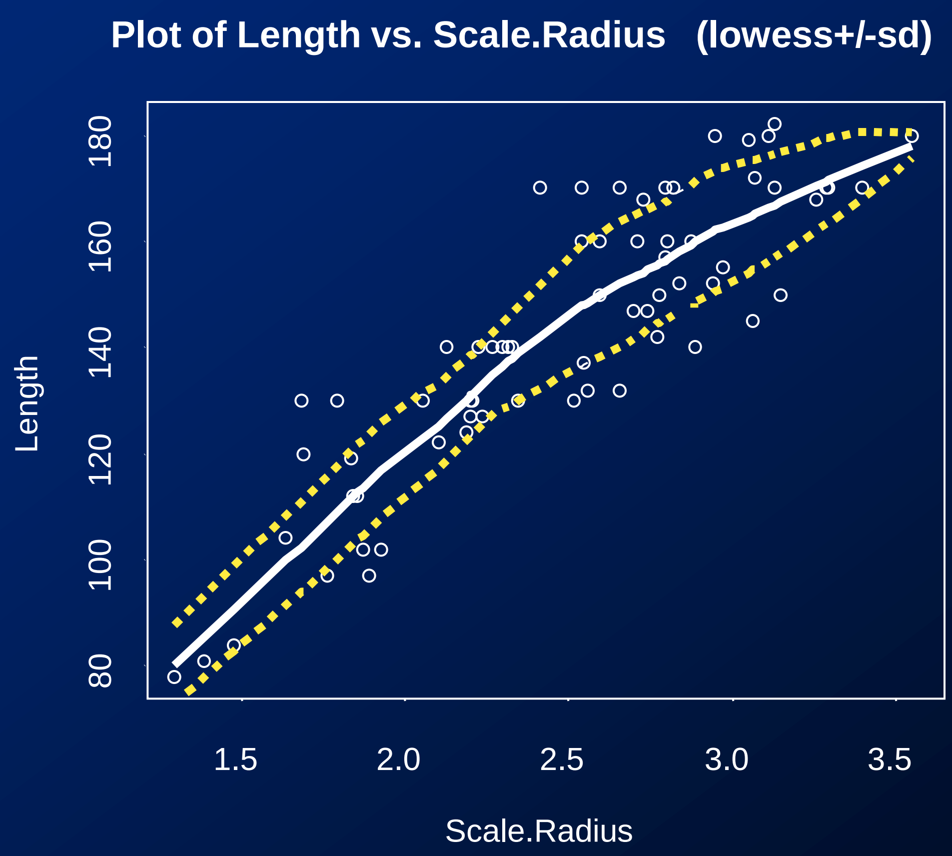
```
> normcheck(bluegill.fit)
```





Diagnostics XI

- We only worry about normality after we are satisfied the relationship is linear, with constant scatter.



Diagnostics XII

```
Multiple R-Squared: 0.7764
```

- The Multiple R-Squared ($0 \le R^2 \le 1$) tells us, when expressed as a percentage ($0\% \le R^2 \le 100\%$), the percentage of the variation in *Y* that our regression model can explain, using the variation in *X*.

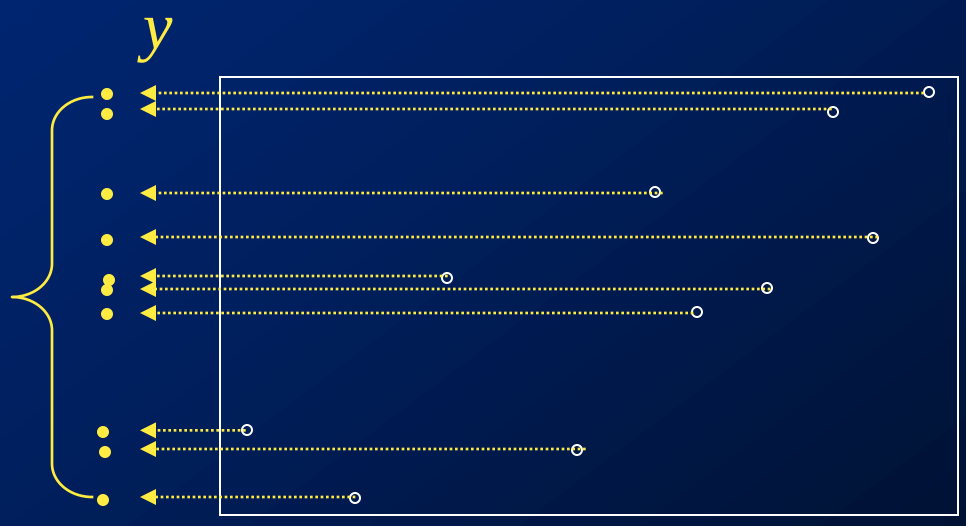
- It is a measure of **how well our model fits the data, <u>assuming</u> that we have correctly identified and modelled the trend**.
- However, it ***tells us nothing*** about how well the error assumptions are satisfied.

Diagnostics XIII

- R^2 is best explained diagrammatically:

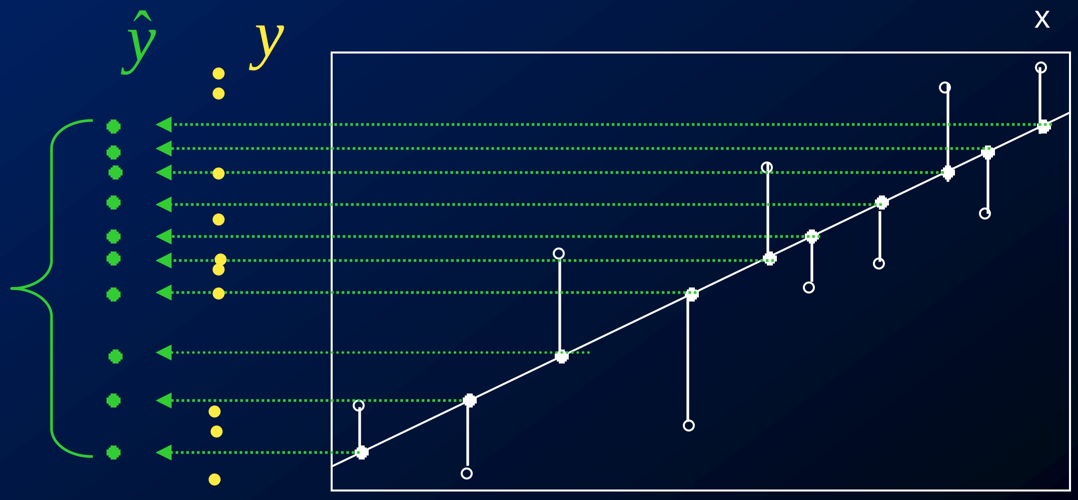
Dotplot of the *y*'s

Shows the variation in the *y*'s



Dotplot of the $\hat{y}$'s

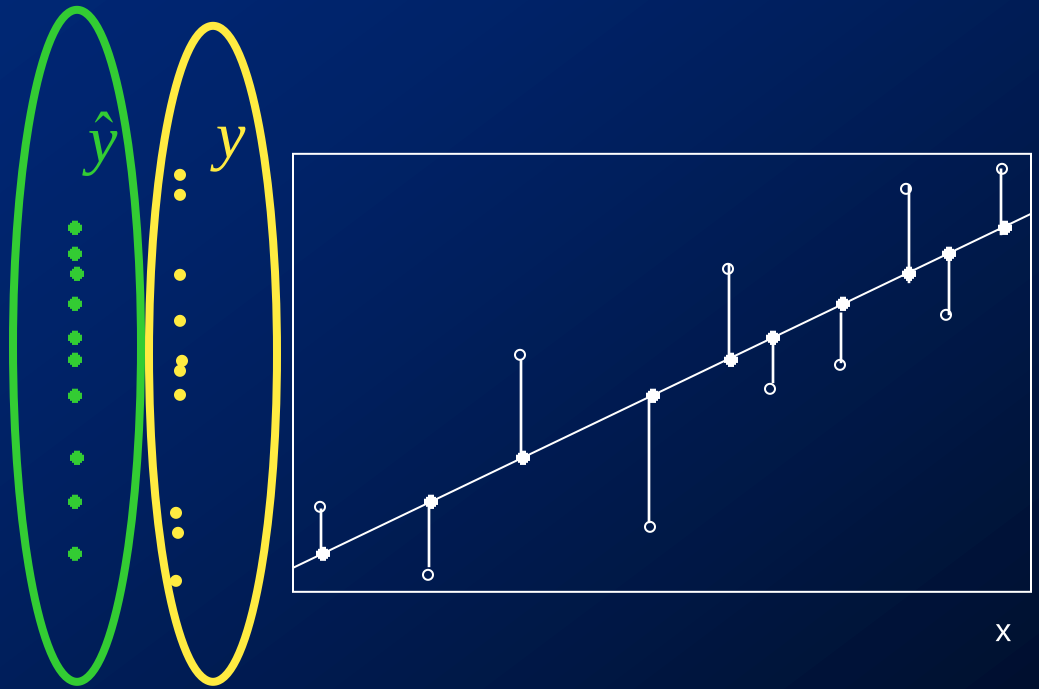
Shows the variation in the $\hat{y}$'s



Diagnostics XIV

Variation in the $\hat{y}$'s

This amount of variation can be ***explained*** as transmitted from the *x*'s



We see some additional variation in the *y*'s here.

The ***excess*** (residual variation) is ***not explained*** by the model.

Diagnostics XV

$$R^2 = \frac{\text{Variation in the } \hat{y}'s}{\text{Variation in the } y's} = \frac{\text{Reg SS}}{\text{Total SS}} = 1 - \frac{\text{Res SS}}{\text{Total SS}}$$

- Assuming the trend is correctly modelled:
  - R^2 near 1 (100%) shows the model fits well (residuals are small).
  - R^2 near 0 shows the model does not fit the data well (residuals are large). A weak relationship.

$$R^2 = \mathrm{cor}(y, \hat{y})^2$$

$$R^2 = r^2 = \mathrm{cor}(\mathrm{x}, \mathrm{y})^2$$

Diagnostics XVI

```
Adjusted R-squared: 0.7729
```

- Adjusted R^2 is of interest in a Multiple Regression setting. (See: Block 9)

- We will also use Adjusted R^2 for Model Building. (See: Block 11)

Diagnostics XVII

- The model we have fitted to the Camp Lake data has an R^2 of 0.7764 or 78%.

- We are reasonably satisfied that our model satisfies the assumptions of Simple Linear Regression.
  - Therefore we can say that **78% of the variation in bluegill's length is explained by the variation in scale radius** (i.e by our model).

Diagnostics XVIII

```
F-statistic: 222.3 on 1 and 64 DF,  p-value: < 2.2e-16
```

- The null hypothesis for this *F*-test is:

  H_0: **None of the explanatory variables are related to the response**

- The Regression ANOVA Identity is:

  TSS = RegSS + ResSS

Diagnostics XIX

- The Regression ANOVA Table:

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	*F*- Ratio	*P-Value*
Regression	k	*Reg SS*	*Reg MS*	$\dfrac{Reg\ MS}{Res\ MS}$	$\Pr(F \geq F_0)$
Residual	$n-k-1$	*Res SS*	*Res MS*		
Total	$n-1$	*Tot SS*			

- *k* is the number of explanatory variables used in the model.

Diagnostics XX

- Or, with the formulae for the Sums of Squares:

Source of Variation	**Degrees of Freedom**	**Sum of Squares**	**Mean Square**	***F*- Ratio**	***P*-Value**
Regression	k	$\sum(\hat{y}_i - \bar{y})^2$	*Reg MS*	$\dfrac{Reg\ MS}{Res\ MS}$	$\Pr(F \geq F_0)$
Residual	$n-k-1$	$\sum(y_i - \hat{y}_i)^2$	*Res MS*		
Total	$n-1$	$\sum(y_i - \bar{y})^2$			

- *k* is the number of explanatory variables used in the model.

Diagnostics XXI

```
> anova(bluegill.fit)
Analysis of Variance Table
Response: Length
             Df Sum Sq Mean Sq F value    Pr(>F)
Scale.Radius  1  34314   34314  222.26 < 2.2e-16 ***
Residuals    64   9881     154
---
Signif. codes: 0`***'0.001`**'0.01`*'0.05`.'0.1` '1
```

Statistical Inference I

```
> bluegill.fit<-lm(Length~Scale.Radius,data=camplake.df)
> summary(bluegill.fit)
Call:
lm(formula = Length ~ Scale.Radius, data = camplake.df)
Residuals:
    Min       1Q   Median       3Q      Max
-21.928  -8.041  -1.941   8.946  31.028
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    34.920      7.356   4.747 1.20e-05 ***
Scale.Radius   43.126      2.893  14.908  < 2e-16 ***
---
Signif. codes: 0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1
Residual standard error: 12.43 on 64 degrees of freedom
Multiple R-squared: 0.7764,   Adjusted R-squared: 0.7729
F-statistic: 222.3 on 1 and 64 DF,  p-value: < 2.2e-16
```

Is the intercept of interest?

A) Yes

B) No

C) Maybe

Statistical Inference II

- Once we are satisfied with our model, we can then use it to make statistical inferences.
  - Estimating the unknown population parameters (i.e. the β_j's) and the error standard deviation, σ
- We estimate σ using:

$$\hat{\sigma} = s = \sqrt{\text{ResMS}} = \text{Residual Standard Error}$$

```
Residual standard error: 12.43 on 64 degrees of freedom
```

Statistical Inference III

- The estimated coefficients, estimate the unknown parameters, β_j

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     34.920      7.356   4.747 1.20e-05 ***
Scale.Radius    43.126      2.893  14.908  < 2e-16 ***
```

- The *P-value* from the test of H_0: $\beta_1 = 0$ tells us the strength of the relationship between the explanatory variable and the response variable.

Statistical Inference IV

- We can do tests of hypotheses concerning the true values of the parameters, β_0 and β_1: $H_0 : \beta_j = \beta_j^{hyp}$
- The test statistic is:

$$t_0 = \frac{\hat{\beta}_j - \beta_j^{hyp}}{\mathrm{se}(\hat{\beta}_j)}$$

$$P-value = 2 \times \Pr(T \geq |t_0|) \quad \text{where}\, T \sim t_{df}$$

- Most statistical packages (including **R**) routinely print the *t*-statistics and *P-values* for the tests:

$$H_0 : \beta_j = 0$$

Statistical Inference V

- **Testing H_0: $\beta_1 = 0$ is important**.
- A **large *P-value*** means we cannot reject the hypothesis that there is **no relationship** between the *X* and *Y* variables.
  - Knowing the value of *X* does not give us any useful information about the value of *Y*.

- We are ***not usually*** interested in testing H_0: $\beta_0 = 0$, **unless the intercept ($x = 0$) is a value of interest**, **and** **we have *x*-values close to 0** in our data.

Statistical Inference VI

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     34.920      7.356   4.747 1.20e-05 ***
Scale.Radius    43.126      2.893  14.908  < 2e-16 ***
```

Statistical Inference VII

- We also want confidence intervals for the true values of the β_j's.
  - A confidence interval for β_j is given by:

$$\text{estimate} \pm t_{df,\alpha/2} \times \text{se(estimate)}$$

$$\hat{\beta}_j \pm t_{df,\alpha/2}\,\text{se}(\hat{\beta}_j)$$

- $df = n - 2$ where n is the sample size

```
> ciReg(bluegill.fit)
             95 % C.I.lower    95 % C.I.upper
(Intercept)        20.22412          49.61655
Scale.Radius       37.34749          48.90540
```

Statistical Inference VIII

- Our interpretation is that if X increases by 1 unit, Y changes by β_1 units.
  - We can also interpret the estimate and the confidence interval for a ***w-unit change*** in X.
  - Our interpretation would then be that if **X increases by w units, Y changes by $\beta_1 \times w$ units**.

- We estimate that a **0.1-unit** increase in scale radius is associated with an increase in a bluegill's length of between **3.7 mm and 4.9 mm**, on average.

Prediction I

- Suppose we wish to predict the length of a bluegill, given its scale radius is 3.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	34.920	7.356	4.747	1.20e-05	***
Scale.Radius	43.126	2.893	14.908	< 2e-16	***

- Our estimated model is:

$$\hat{y} = \hat{\mathrm{E}}(Y \mid X) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$= 34.920 + 43.126 \times \text{Scale.Radius}$$

- Our point prediction, $\hat{\mathrm{E}}(Y \mid X{=}3)$, for the expected length of a bluegill with a scale radius of 3 is:

$$34.920 + 43.126 \times 3 = 164.298 \text{ mm}$$

Prediction II

- This gives us the **point on the fitted regression line** when x (scale radius) = 3
- If we wish to build a confidence interval for the ***average (mean) value of the response***, $E(Y \mid X)$, for a given value of the explanatory variable, we have to take account of the uncertainty in the estimation of the trend.
  - This, gives us a Confidence Interval for the true trend value, $E(Y \mid X)$ which is often called a ***Confidence Interval for the mean*** (think of the interval as being confidence bands for the regression line).

Prediction III

- We want a confidence interval for:

$$E(Y \mid X) = \beta_0 + \beta_1 x$$

based on our sample estimates:

$$\hat{E}(Y \mid X) = \hat{\beta}_0 + \hat{\beta}_1 x$$

- When we use a regression model for **prediction**, we have **two sources of variation** that need to be taken into account when we build prediction intervals:
  - **uncertainty in the estimation of the trend**, $E(Y \mid X)$, as above, and
  - **uncertainty due to the residual scatter** about the trend, $\mathrm{Var}(Y \mid X)$.

Prediction IV

- If we want to build a prediction interval for a ***new observation*** given a value of the explanatory variable, we need to take account of the uncertainty in the estimation of the trend ***AND*** the uncertainty due to the residual scatter about the trend.

- We want a prediction interval for:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

based on our sample estimates:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + r_i$$

Prediction V

- A Prediction Interval for a new observation will be ***wider*** than a Confidence Interval for the mean.

- In **R** we do predictions using the function:

  `predict20x(`***`fit,prediction.data.frame`***`)`

- This function takes two arguments:
  - The object in which we stored the output from our model.
  - A data frame of the values of the explanatory variable for the prediction(s).

Prediction VI

```
> bluegill.to.predict<-data.frame(3)

> predict20x(bluegill.fit,bluegill.to.predict)
  Predicted  Conf.lower Conf.upper  Pred.lower Pred.upper
1     164.3     160.044    168.555     139.115    189.484
```

Predicted value, or point on the estimated trend $\hat{y} = \hat{\mathrm{E}}(Y \mid X)$

Confidence Interval for the true trend value $\mathrm{E}(Y \mid X)$ (***CI for the mean***)

Prediction Interval for a new observation

Prediction VII

- Predictions can only be precise if the scatter is small when compared to the range of the trend.
- **Provided the model assumptions hold**:
  - $R^2 > 90\%$ - accurate predictions (narrow prediction intervals).
  - $80\% < R^2 < 90\%$ - reasonable predictions (reasonable prediction intervals).

- **NOTE**: If the interval comprises a substantial proportion of the range of the *Y*-values, then the predictions are of little practical use.

Prediction VIII

- When R^2 is less than 80%, our model will give us fairly wide Prediction Intervals.
  - However, if we have evidence of a relationship **we can still use the model to explain the behaviour of *Y*, in terms of the explanatory variable, *X*.**

- **NOTE**: **Prediction outside the range of the data is unreliable**.

Which Assumptions Matter?

- **The assumption about the trend always matters**

- Which of the error assumptions matter?
  - Tests and Confidence Intervals for coefficients are:
    - Sensitive to the independence assumption.
    - Sensitive to the constant scatter assumption.
    - Robust against non-normality (especially in moderate to large samples).
  - **Prediction Intervals are:**
    - **Sensitive to all 3 error assumptions.**