Introduction to Statistical Modeling

Simple Linear Regression

Joris Vankerschaver

Introduction

Regression

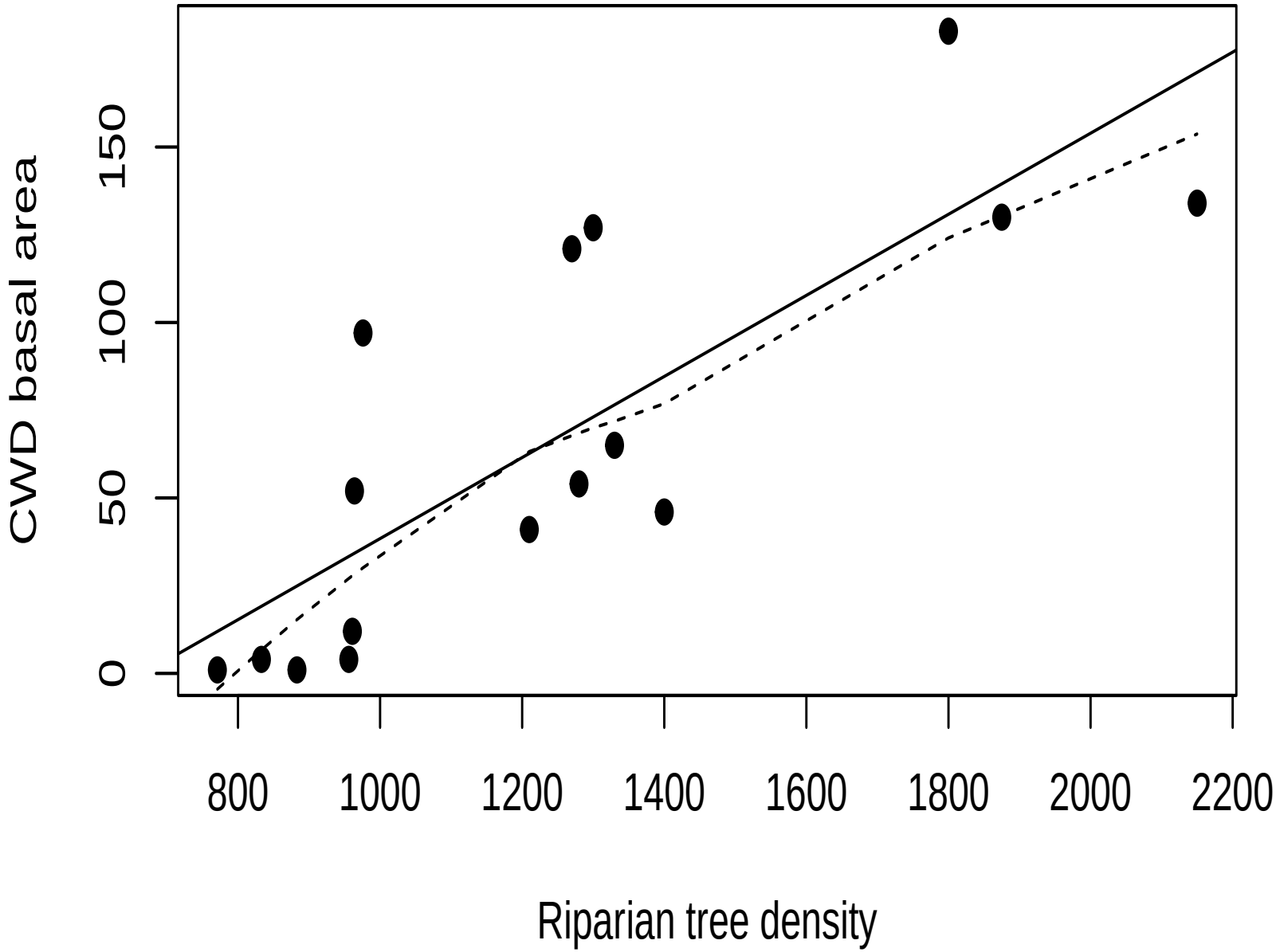
Goal: describe the relationship between 2 series of observations (X_i, Y_i), obtained for individual subjects $i = 1, ..., n$

Example:

Basal area of coarse woody debris (CWD) versus tree density along 16 North American lakes

- **Dependent variable, outcome, response** Y: CWD basal area
- **Independent variable, explanatory variable, predictor** X: tree density (in number per km)

CWD versus tree density



Regression

For fixed X, Y will be some function of X plus random noise:
observation = signal + noise

Mathematical modelling of observation:

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

where $f(x)$ is the expected outcome for subjects with $X_i = x$

$$E(Y_i|X_i = x) = f(x)$$

and ϵ_i is on average 0 for subjects with same X_i:

$$E(\epsilon_i|X_i) = 0.$$

Linear regression

- To obtain accurate and interpretable results, $f(X)$ is often chosen as linear function of unknown parameters
- Use **linear regression model**

$$E(Y|X = x) = \alpha + \beta x$$

  with unknown **intercept** α and **slope** β.
- Linear regression model makes assumption on distribution of X and Y, so can be incorrect.

Use of linear regression

- **Prediction**: when Y unknown, but X known, we can predict Y based on X:

$$E(Y|X = x) = \alpha + \beta x.$$

- **Association**: describe biological relation between variable X and continuous measurement Y
  - Slope β: difference in mean outcome between subjects that differ 1 unit in the value of X:

$$\begin{aligned} E(Y|X = x + \delta) - E(Y|X = x) &= \alpha + \beta(x + \delta) - \alpha - \beta x \\ &= \beta\delta. \end{aligned}$$

Least squares estimates

- Least squares (regression) line: line that 'best' fits data.
- Found by choosing values for α and β that minimize sum of squares of **residuals**:

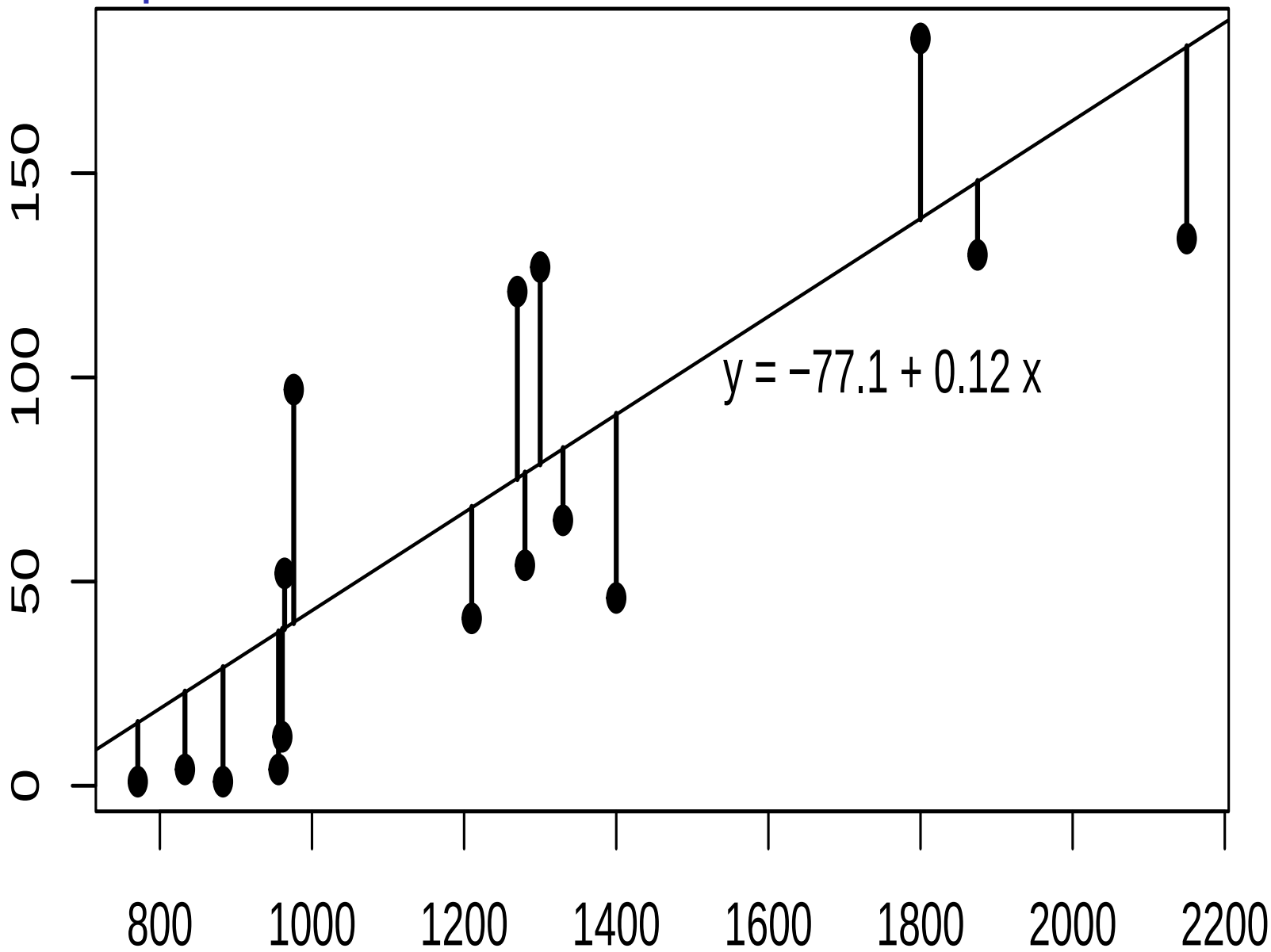
$$\sum_{i=1}^{n} (\underbrace{Y_i - \alpha - \beta X_i}_{\text{Residual}})^2$$

- Estimates for β and α:

$$\hat{\beta} = \text{Cor}(x, y) \frac{S_y}{S_x} \quad \text{and} \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}.$$

with $\text{Cor}(x, y)$ the sample correlation between x and y and S_x, S_y the sample standard deviation.

Residuals plot



See also: residuals animation.

Output linear regression (coefficients only)

```
model <- lm(CWD.BASA ~ RIP.DENS)
summary(model)$coefficients
```

```
                Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) -77.0990778 30.60800907 -2.518918 0.0245520121
RIP.DENS      0.1155161  0.02343233  4.929772 0.0002216405
```

Regression line:

$$E(Y|X = x) = -77.10 + 0.12x$$

Output linear regression (full)

```
summary(model)
```

```
Call:
lm(formula = CWD.BASA ~ RIP.DENS)

Residuals:
   Min     1Q Median     3Q    Max
-38.62 -22.41 -13.33  26.16  61.35

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -77.09908   30.60801  -2.519 0.024552 *
RIP.DENS      0.11552    0.02343   4.930 0.000222 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36.32 on 14 degrees of freedom
Multiple R-squared:  0.6345,    Adjusted R-squared:  0.6084
F-statistic:  24.3 on 1 and 14 DF,  p-value: 0.0002216
```

Interpreting linear regression

- Model: $E(Y|X = x) = -77.10 + 0.12x$
- Expected CWD basal area is 1.2 m^2 larger alongside lakes with 10 more trees per km
- Expected CWD basal area alongside lakes with 1,600 trees per km shoreline:

$$-77.10 + 0.12 \times 1600 = 108 \text{ m}^2$$

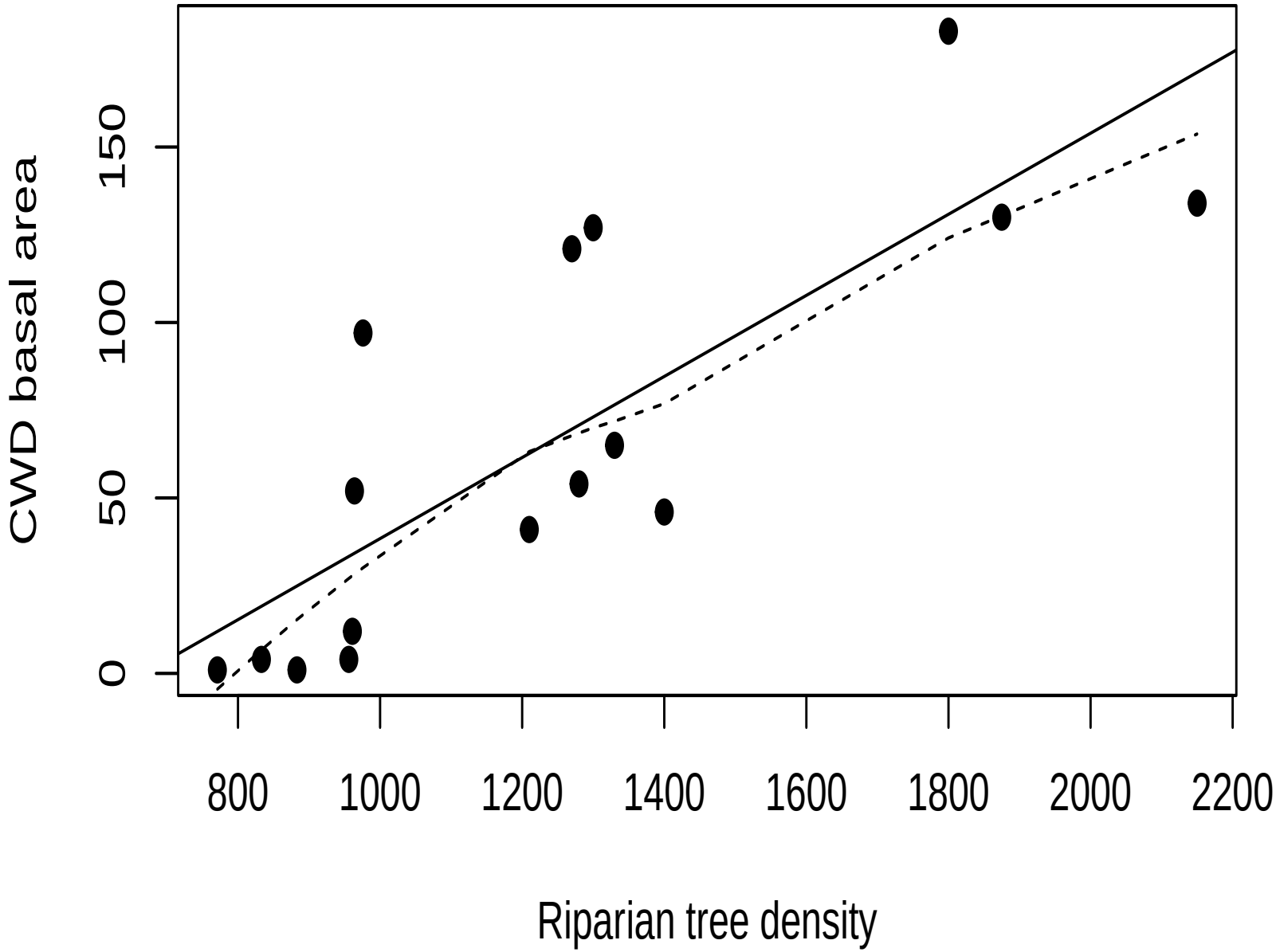
- Expected CWD basal area alongside lakes with 500 trees per km shoreline:

$$-77.10 + 0.12 \times 500 = -17 \text{ m}^2$$

- **Be careful with extrapolation!** (linearity assumption can only be verified within range of data)

Assumptions for linear regression

Verifying linearity assumption



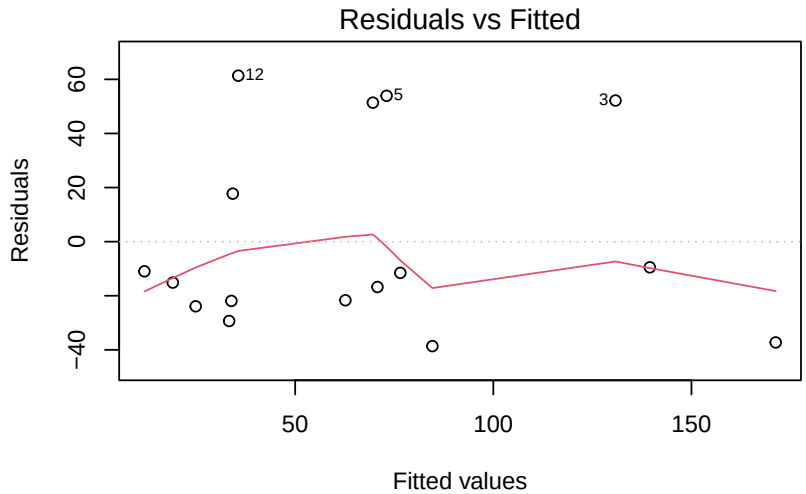
Verifying linearity assumption

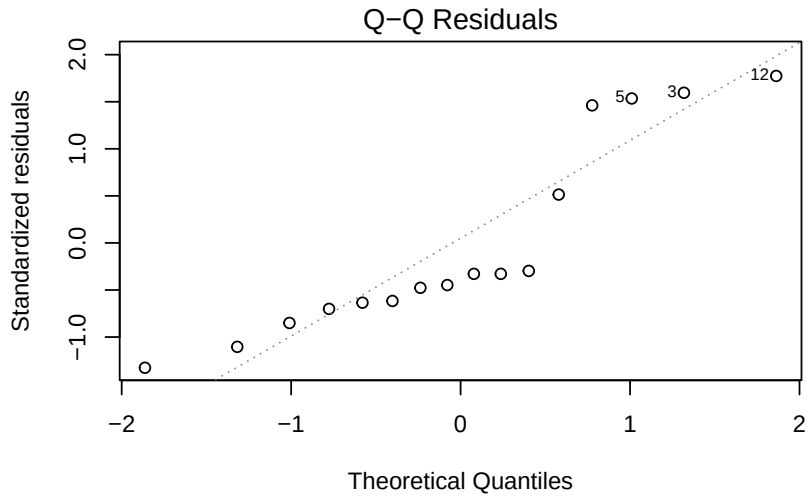
- An alternative (more convenient when there are multiple predictors) is a **residual plot**.
- Note: residuals are prediction errors:

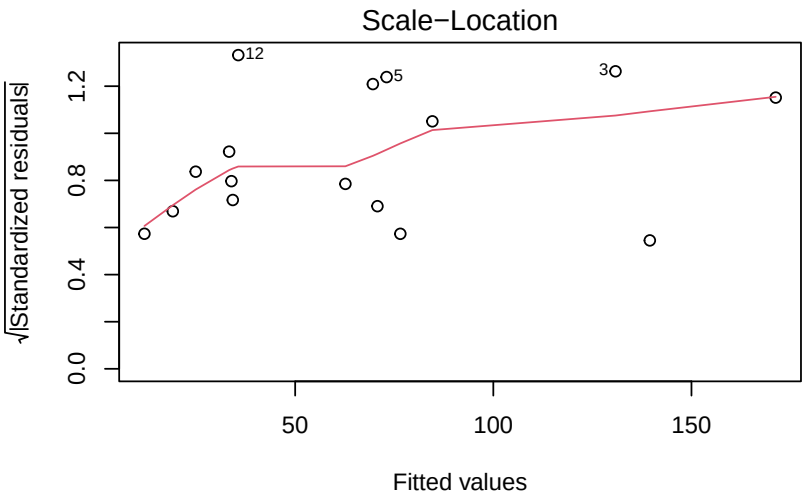
$$e_i = y_i - \hat{\alpha} - \hat{\beta} x_i$$

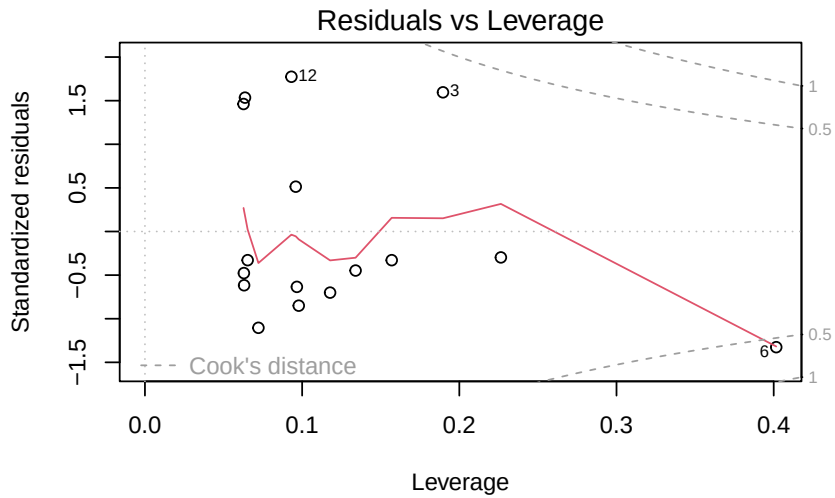
- If linear model correct, then scatterplot of e_i versus x_i or predictions $\hat{\alpha} + \hat{\beta} x_i$ shows no pattern

Verifying linearity assumption









Inference for simple linear regression

To be able to draw conclusions about the linear regression model

$$E(Y|X) = \alpha + \beta X$$

we need extra assumptions:

- **Homoscedasticity**: for fixed X, Y has constant variance

$$\text{Var}(Y|X) = \sigma^2,$$

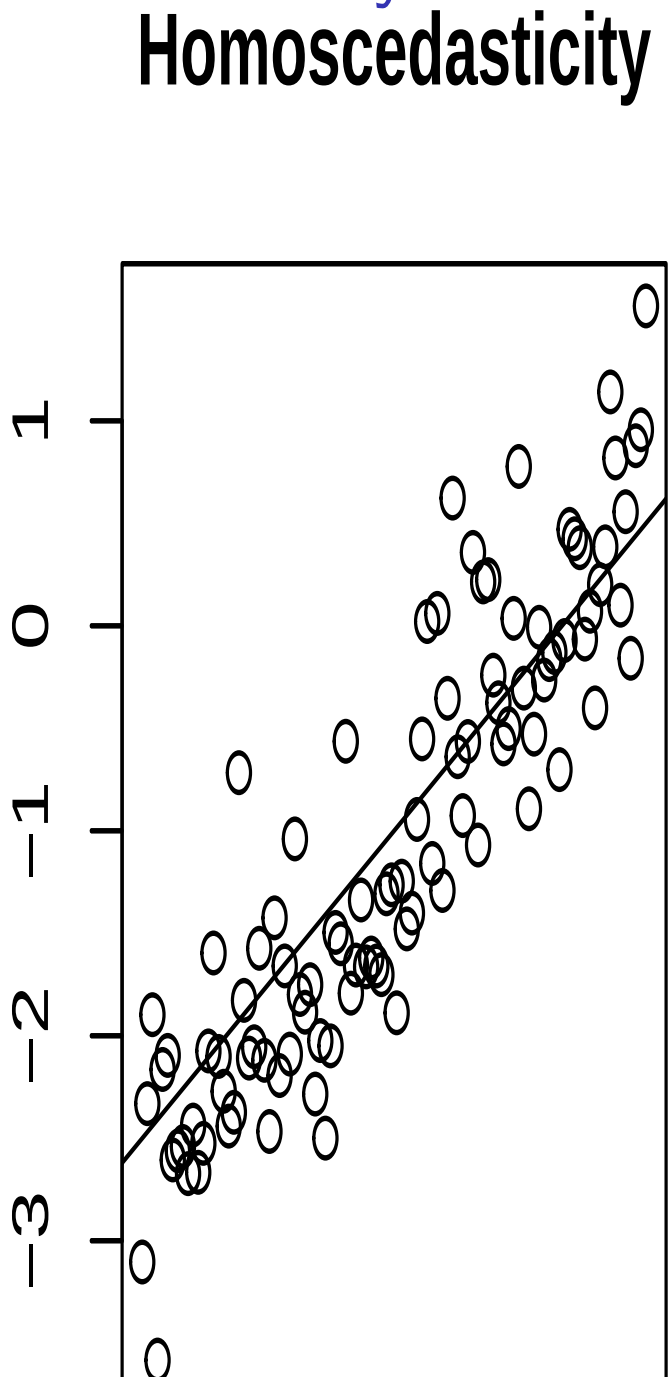
estimated by the residual mean square error:

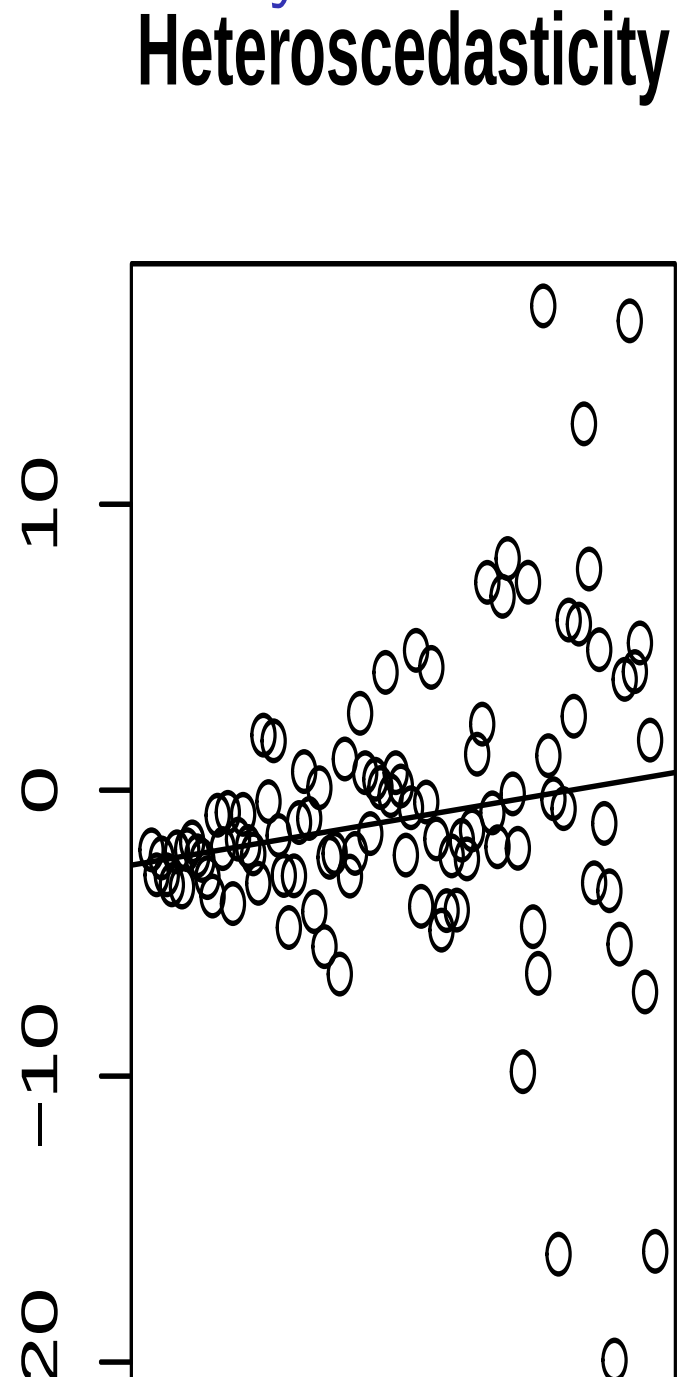
$$\text{MSE} = \sum_{i=1}^{n} e_i^2/(n-2)$$

- **Normality**: for fixed X, Y is normally distributed

$$Y|X \sim N(\alpha + \beta X, \sigma^2)$$

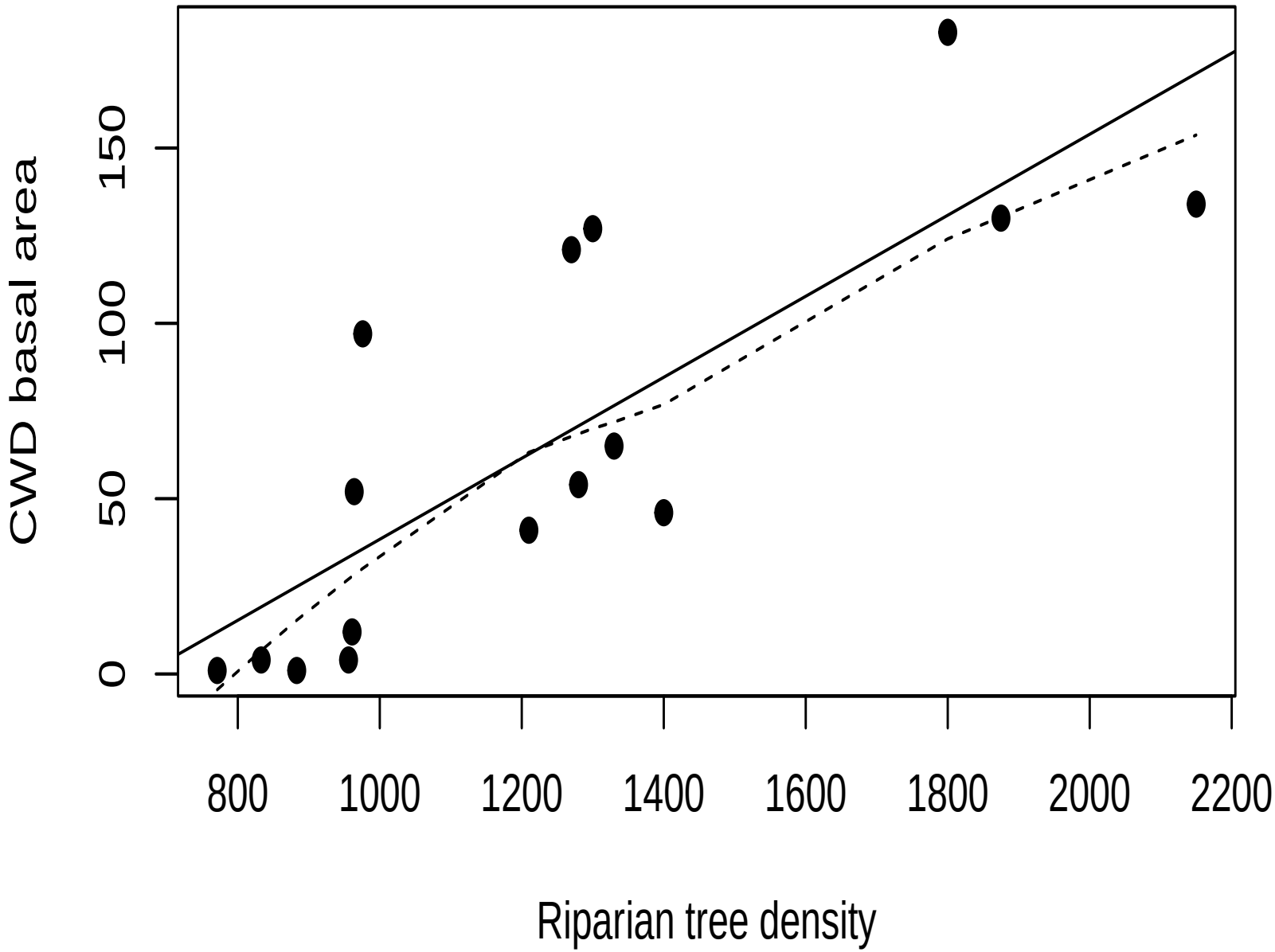
Homoscedasticity versus heteroscedasticity





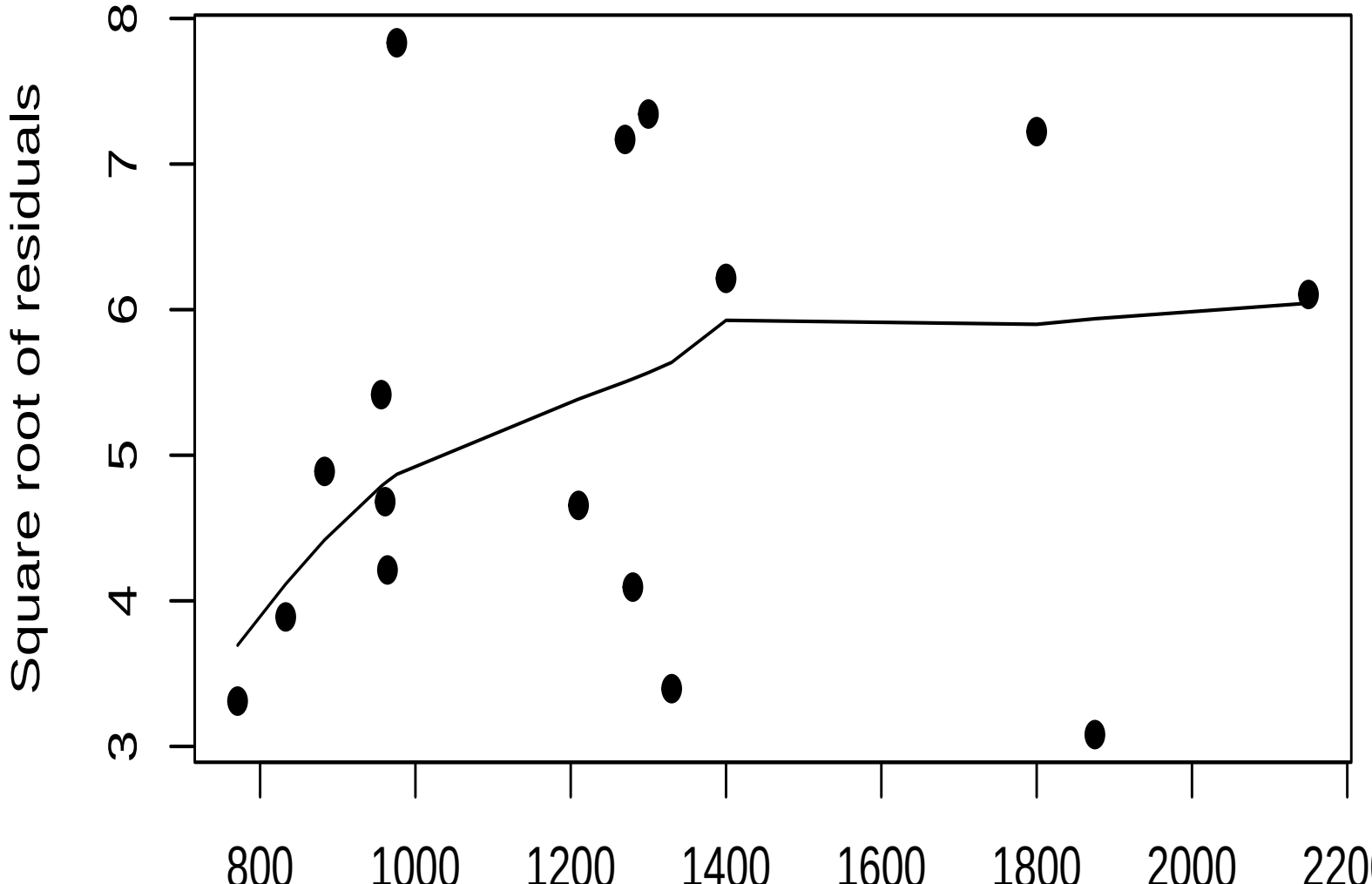
Homoscedasticity?

Hard to check on regression plot directly!

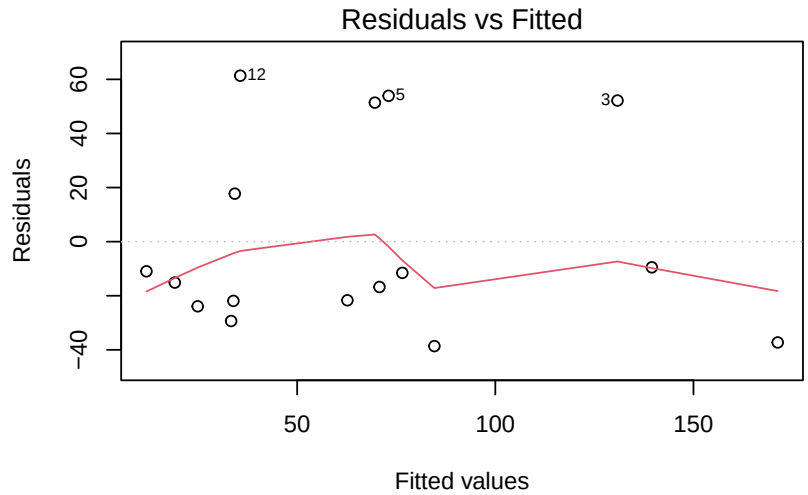


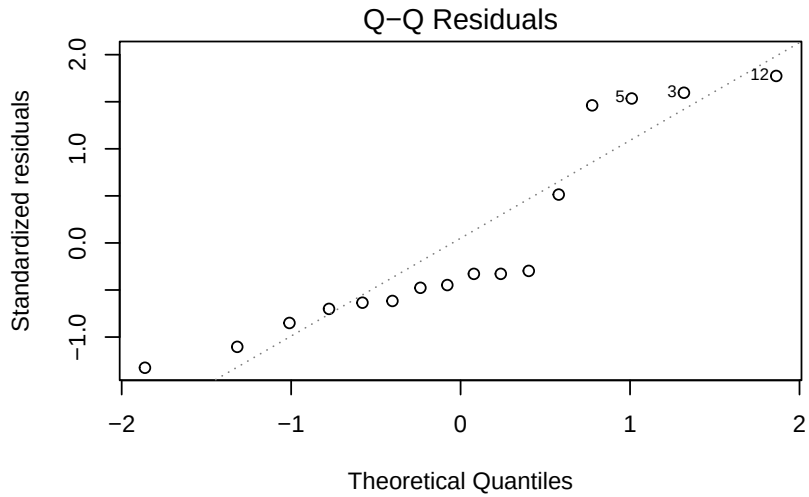
Assumption of homoscedasticity

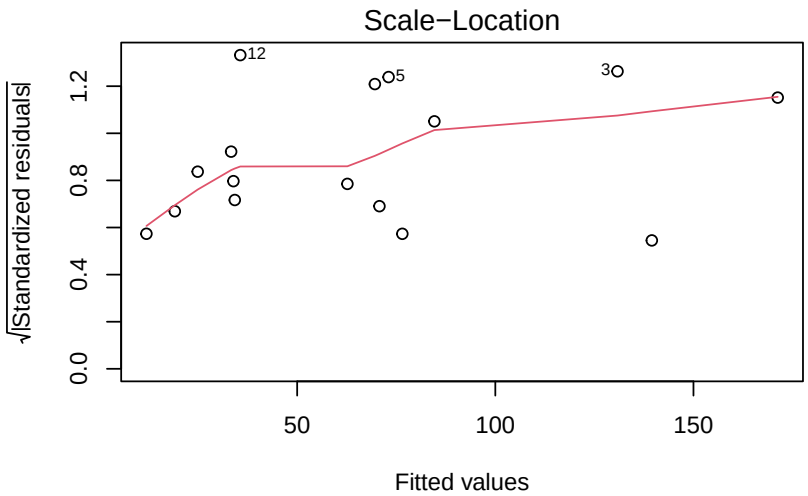
- Squared residuals carry information on residual variability.
- If these are associated with explanatory variable, then indication of **heteroscedasticity**.
- Scatterplot of e_i^2 or $\sqrt{|e_i|}$ versus x_i or predictions.

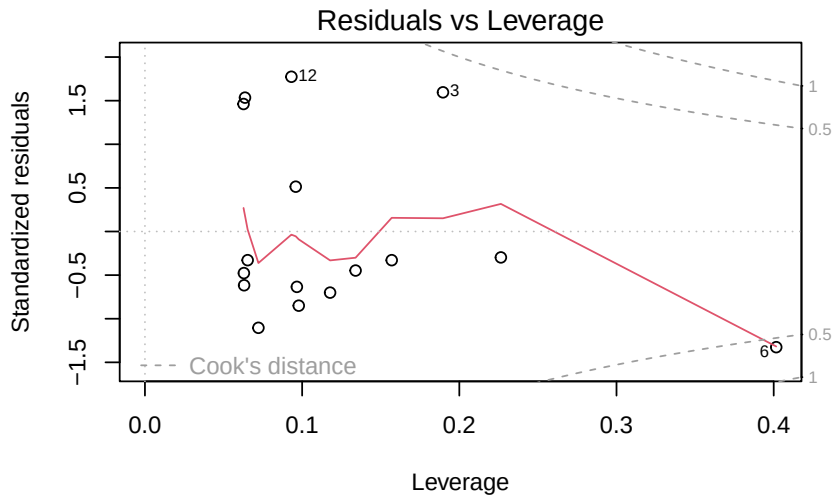


Assumption of homoscedasticity









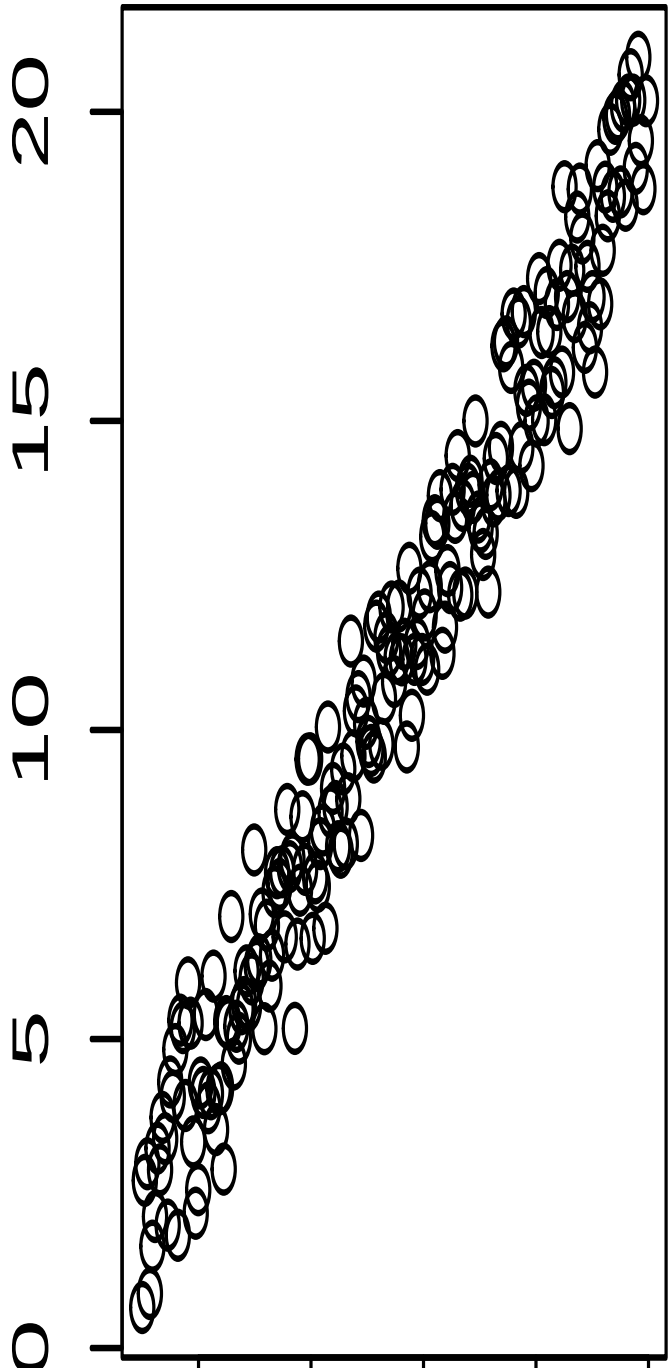
Normality assumption

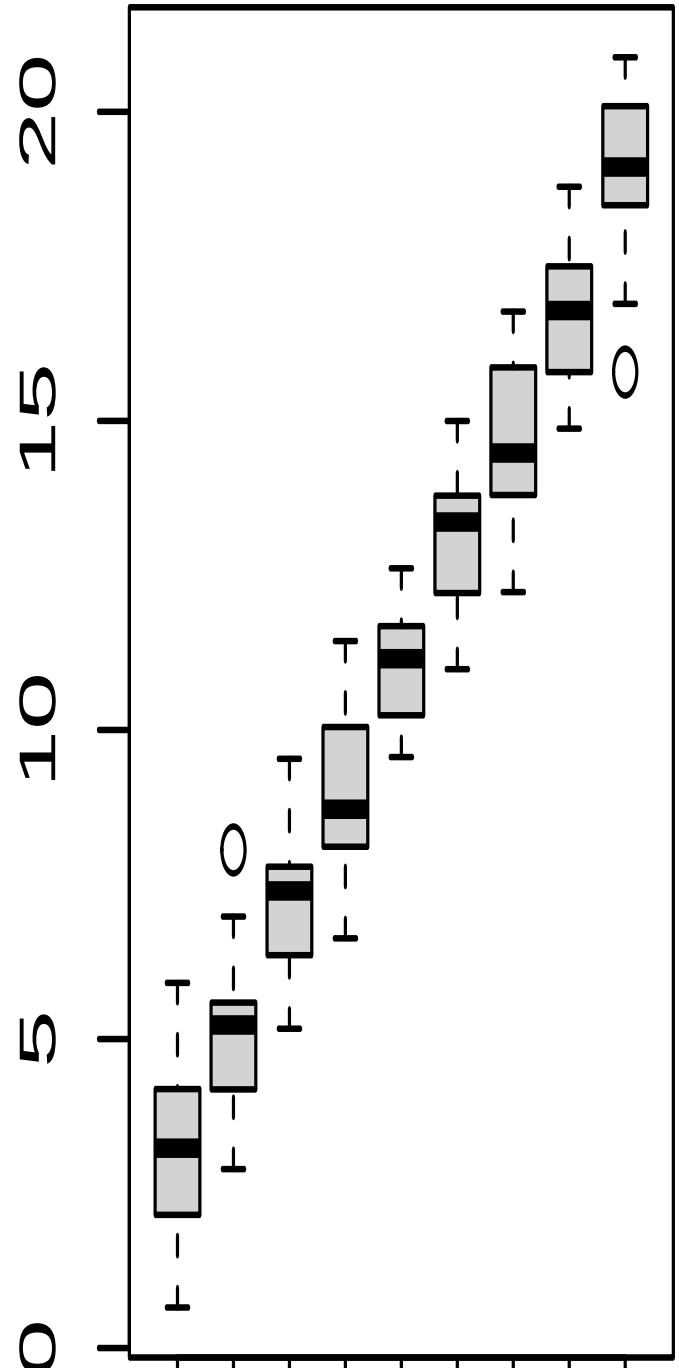
- Assumption: outcomes normally distributed **for fixed values of explanatory variable**:

$$Y|X \sim N(a + bX, \sigma^2).$$

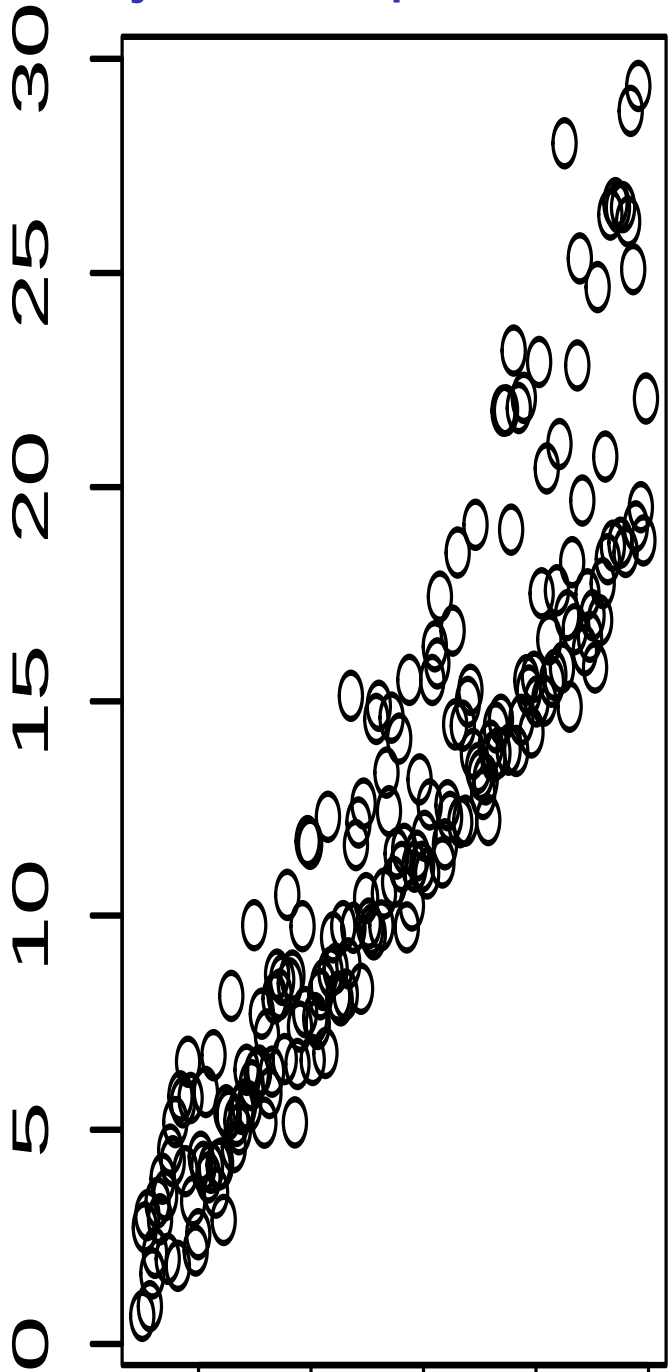
- Can be checked using QQ-plot of the residuals.

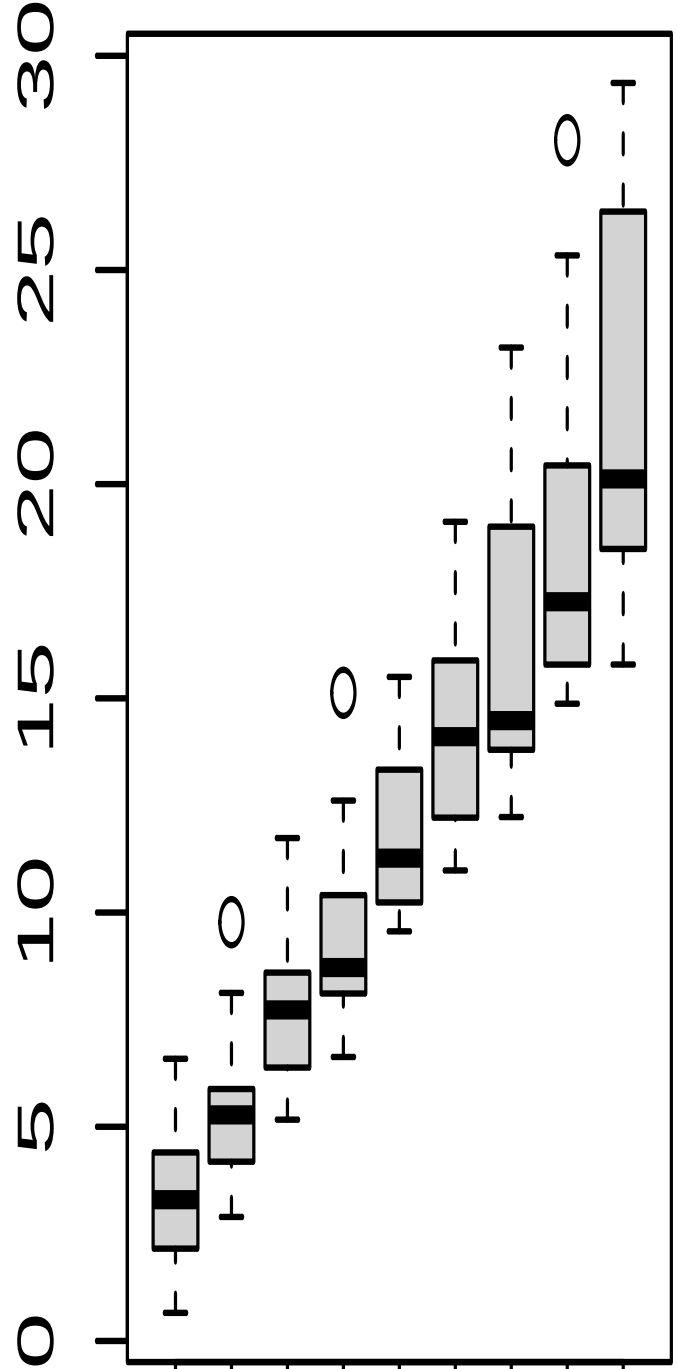
Normality assumption valid



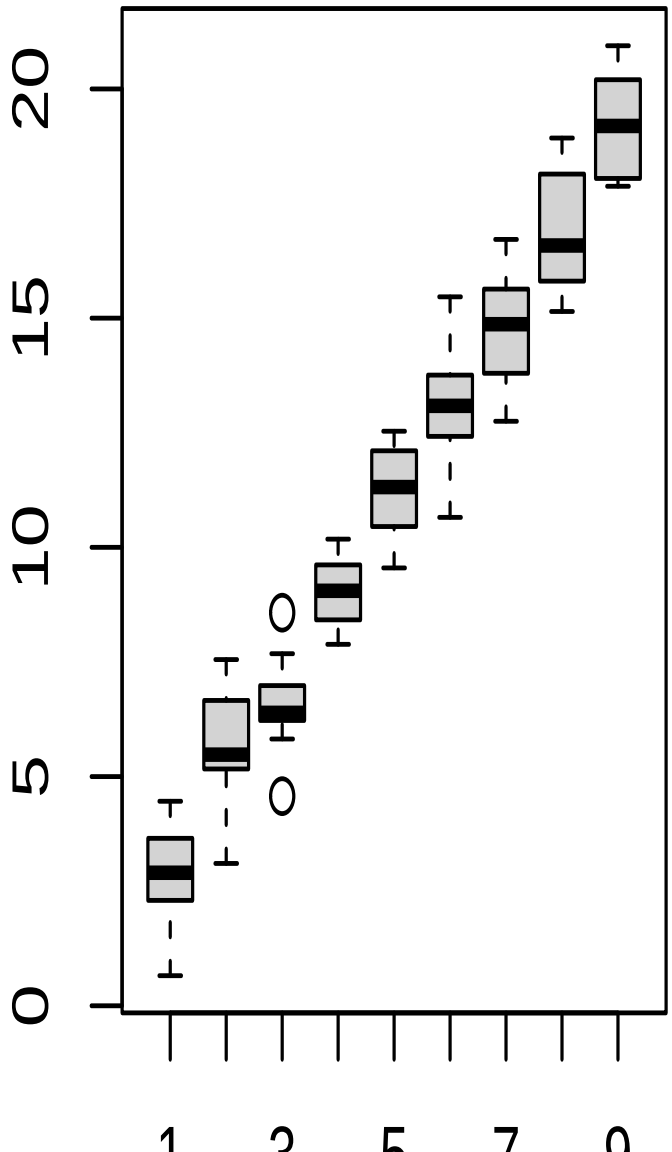


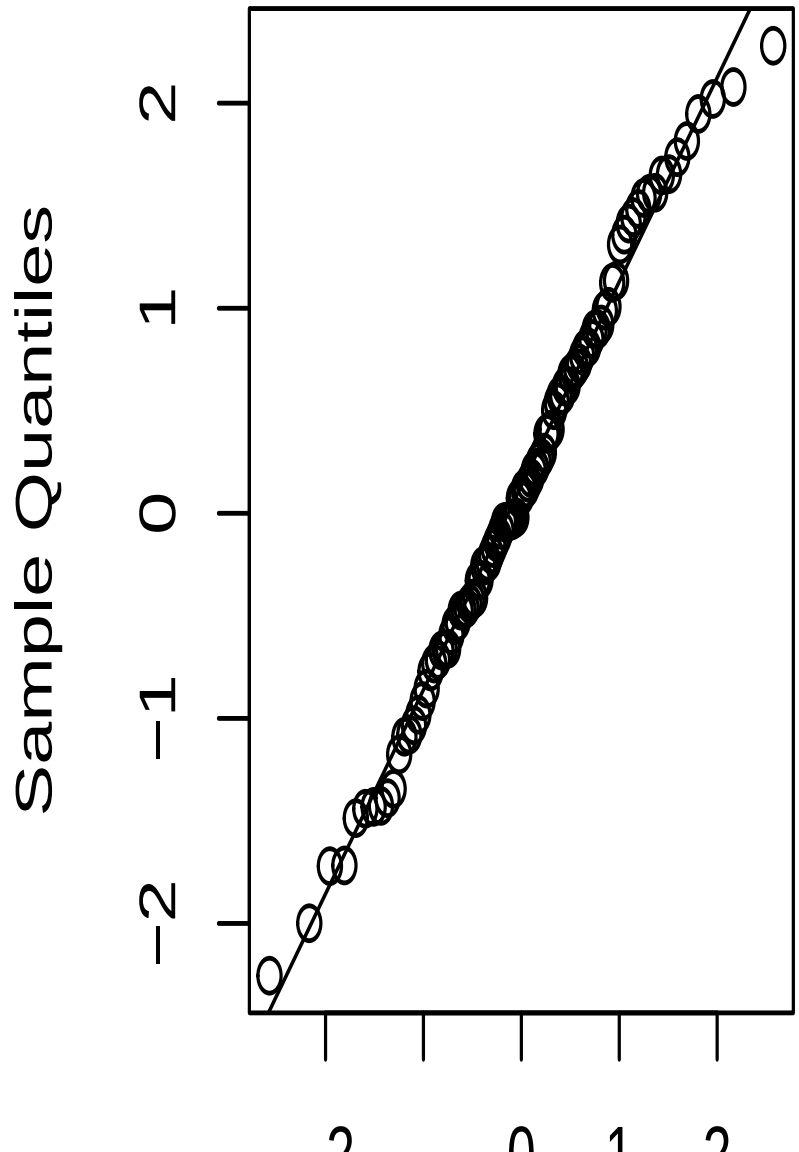
Normality assumption not valid



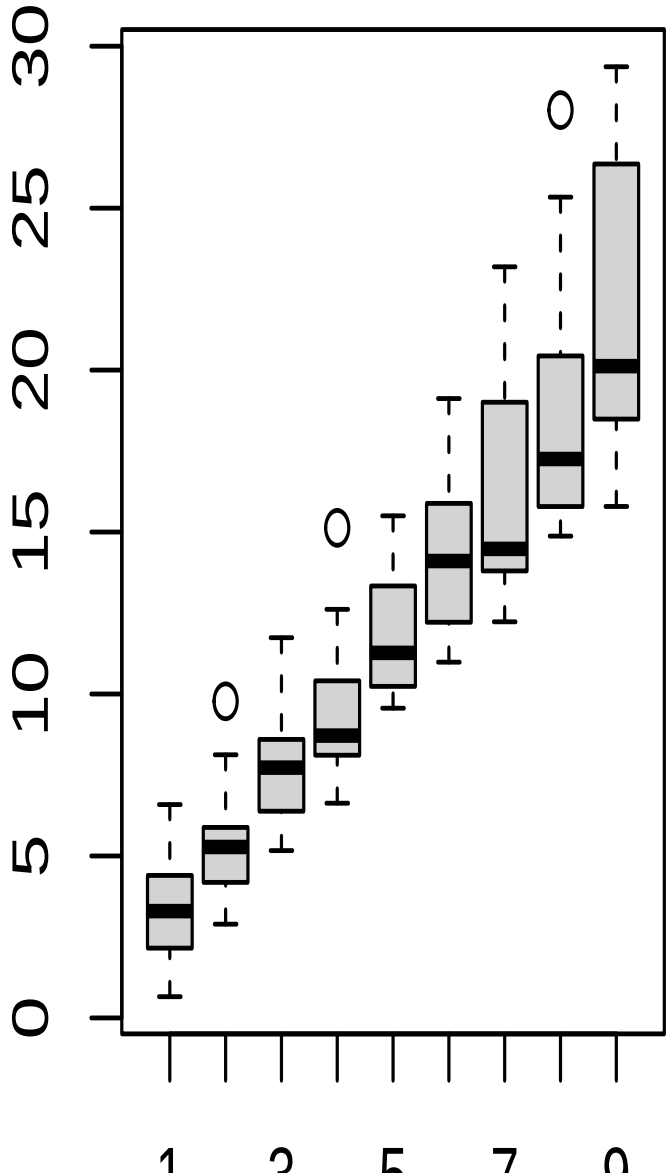


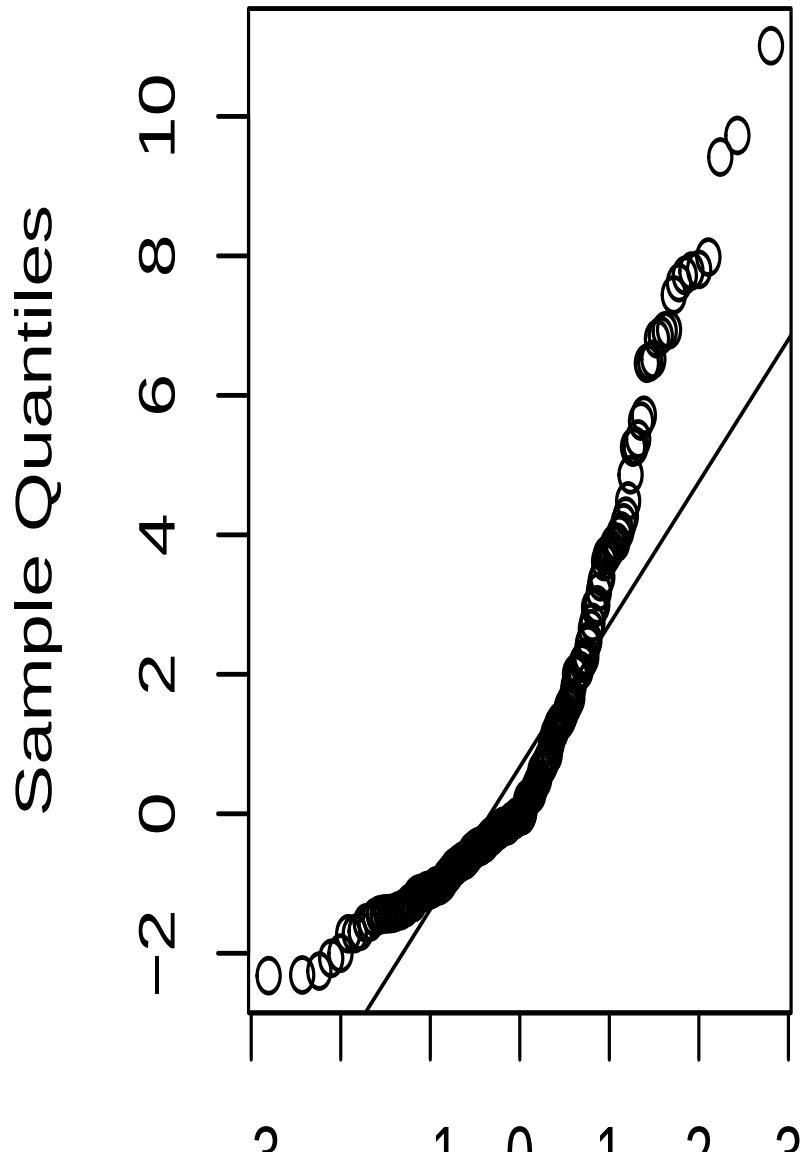
QQ plot of residuals (Y|X normal)



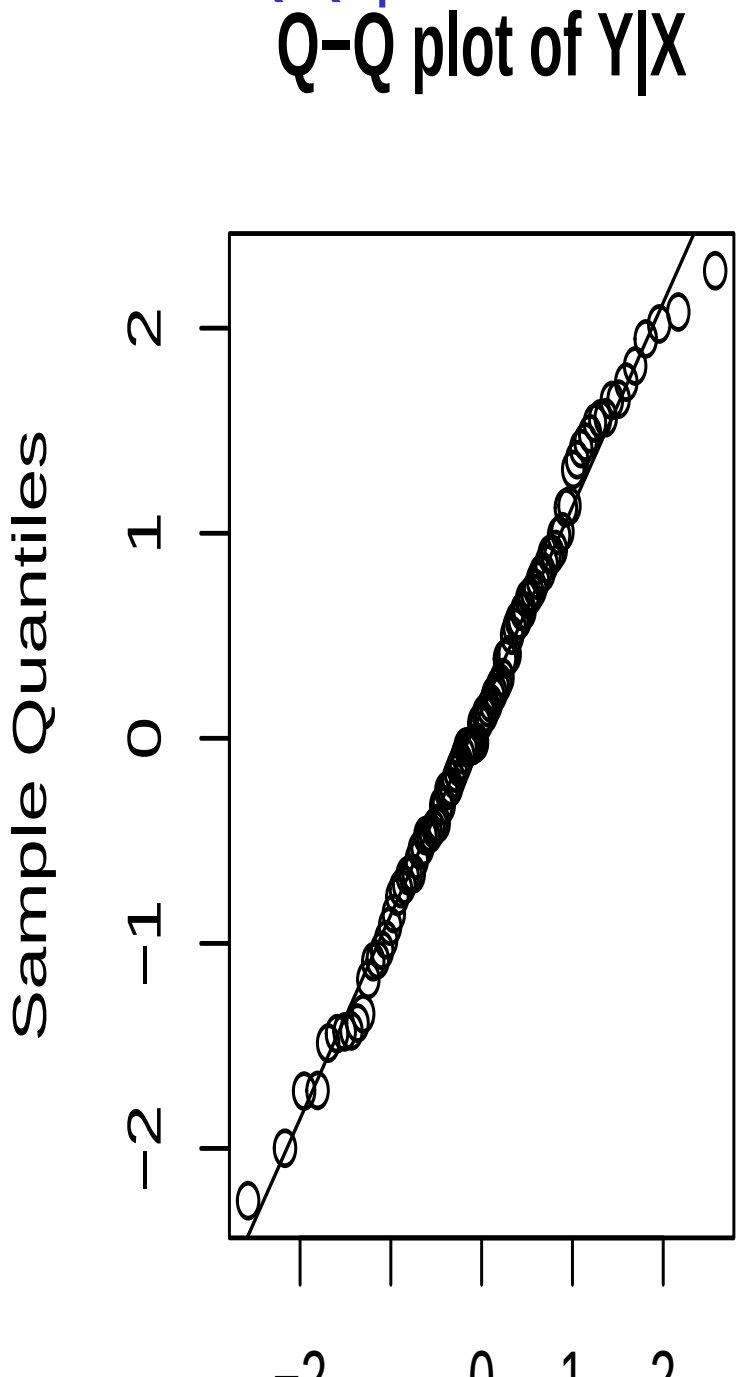


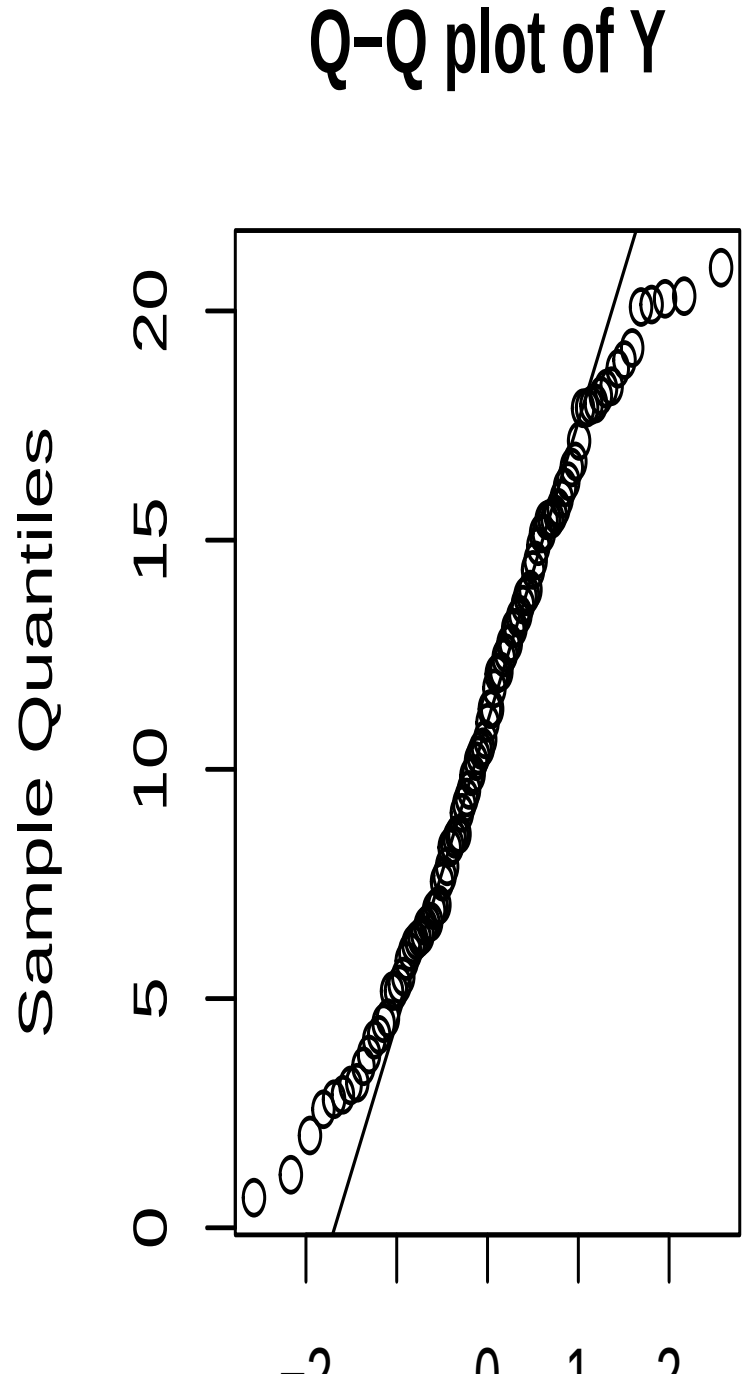
QQ plot of residuals (Y|X not normal)



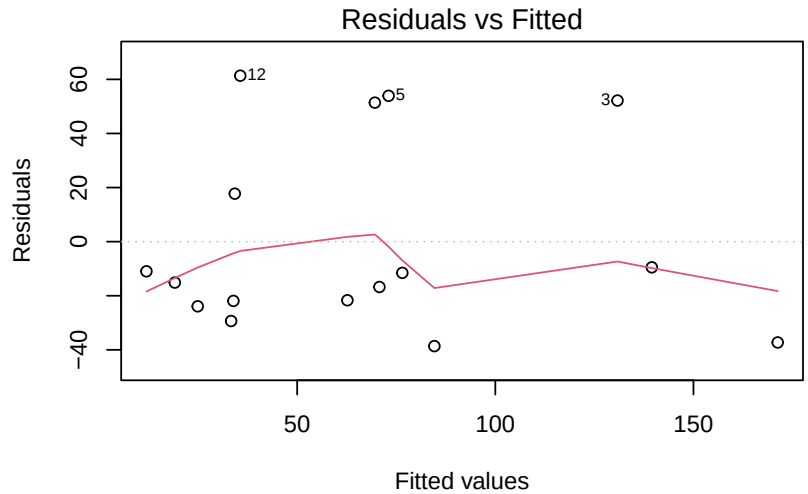


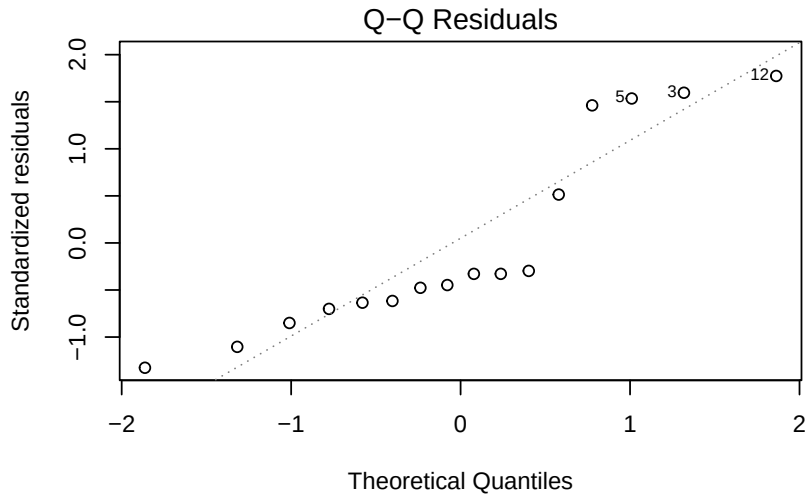
Do not use QQ-plot of Y!

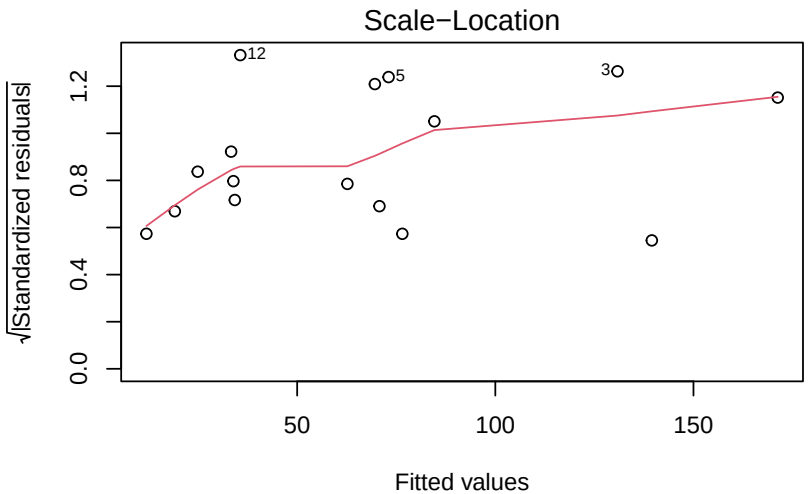


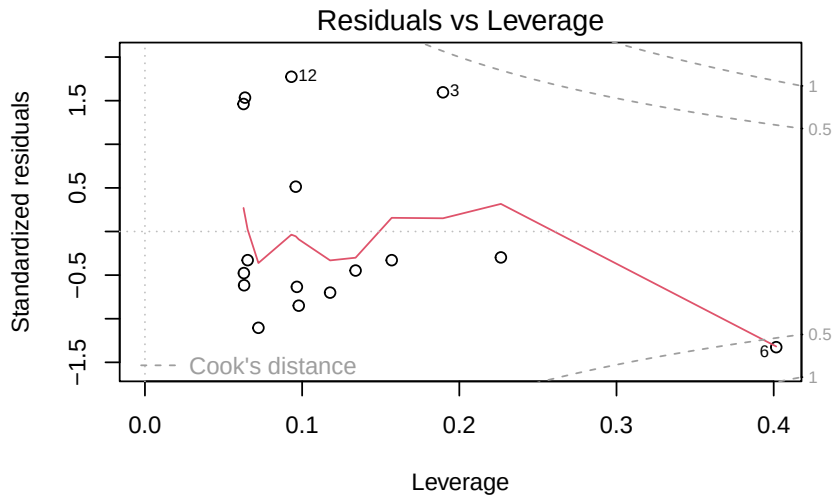


Checking for normality with R diagnostic plots









What if homoscedasticity or normality false?

- Transformation of **dependent variable** can help to obtain normality and homoscedasticity.
- Example transformations:
  $\sqrt{Y}, Y^2, 1/Y, \exp Y, \exp(-Y), \ln Y.$
- Transformation of **independent variable** does not change distribution of Y for given X:
  - does not help in obtaining normality or homoscedasticity.
  - does help to obtain linearity if normality and homoscedasticity are ok.

What if homoscedasticity or normality false?

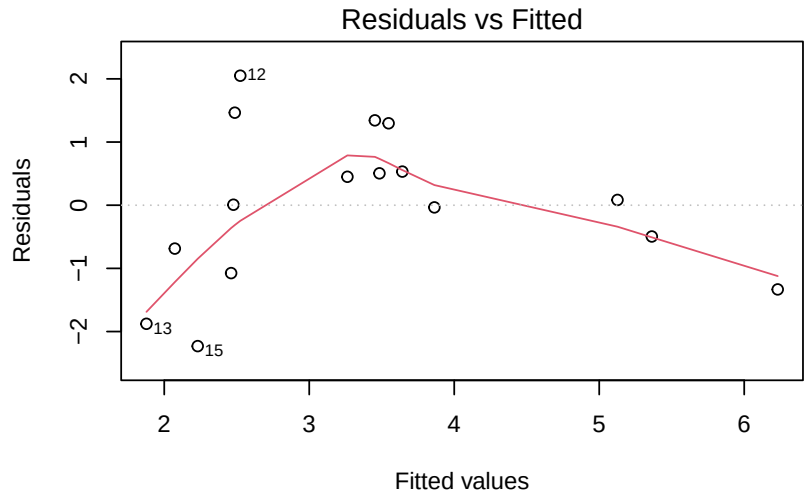
- Often because outcome can only take on values in certain interval (e.g. $[0, 1]$, positive numbers, ...)
- **Solution**: transform outcome such that it can take on all real values
- Example: `CWD.BASA` is always positive: take $\ln$ to make outcome real-valued:

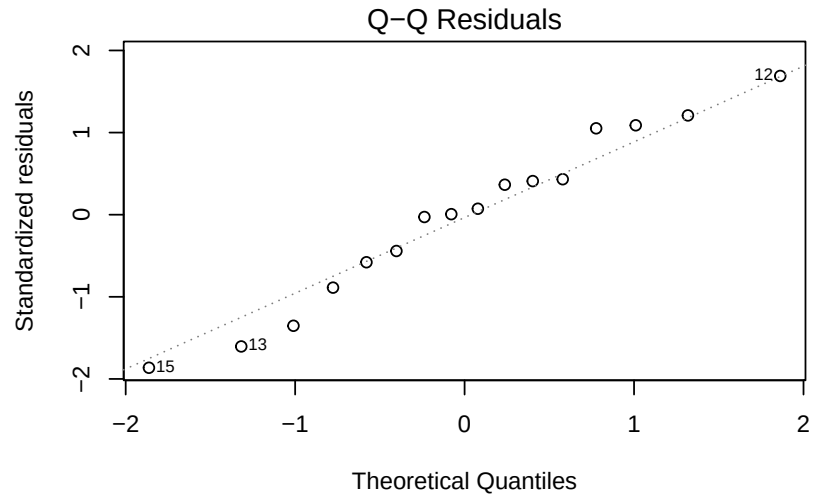
Transforming the outcome

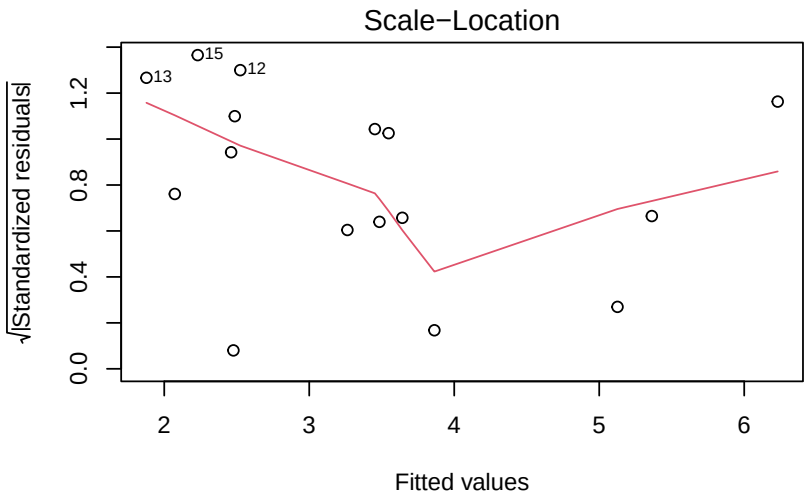
```
model2 <- lm(log(CWD.BASA) ~ RIP.DENS)
summary(model2)
```

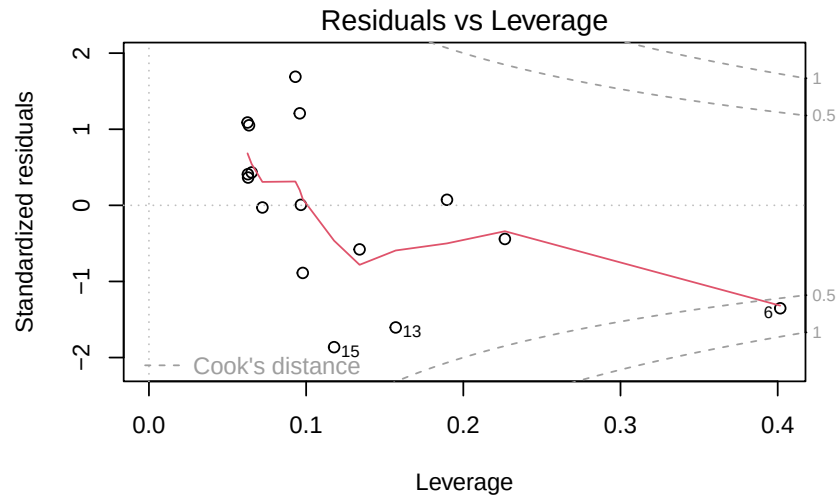
```
Call:
lm(formula = log(CWD.BASA) ~ RIP.DENS)

Residuals:
     Min       1Q   Median       3Q      Max
-2.23086 -0.78379  0.04559  0.72335  2.05022

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.5570100  1.0739690  -0.519   0.6121
RIP.DENS     0.0031573  0.0008222   3.840   0.0018 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.274 on 14 degrees of freedom
Multiple R-squared:  0.513, Adjusted R-squared:  0.4782
F-statistic: 14.75 on 1 and 14 DF,  p-value: 0.001802
```

Residual plots









Higher-order regression

What if linearity assumption is false?

- Transformation of dependent variable
- Transformation of independent variable
- If residuals reveal **quadratic association**, such that

$$e_i \approx \delta_0 + \delta_1 x_i + \delta_2 x_i^2$$

then

$$y_i = \hat{\alpha} + \hat{\beta} x_i + e_i \approx (\hat{\alpha} + \delta_0) + (\hat{\beta} + \delta_1) x_i + \delta_2 x_i^2$$

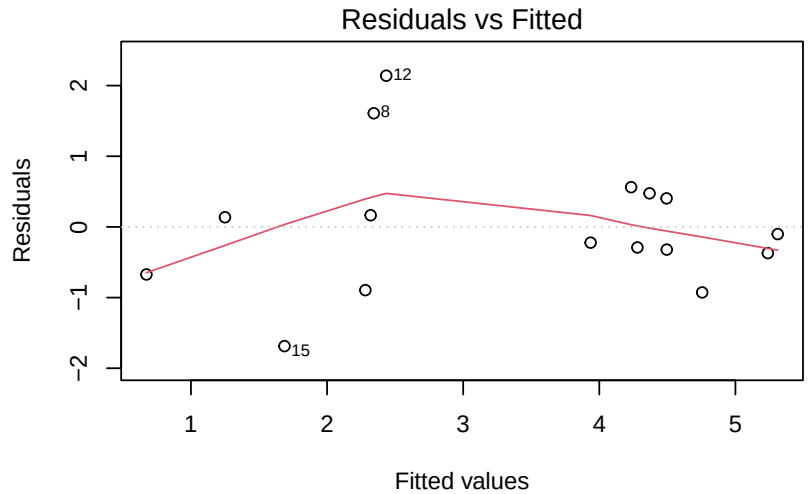
Quadratic regression

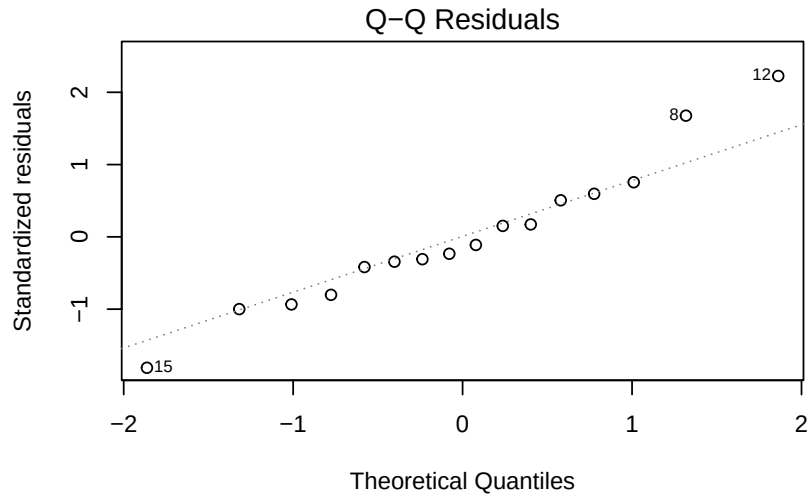
- We assume
$$E(Y|X) = \alpha + \beta X + \gamma X^2$$
- Unknown parameters estimated by **least squares method**: minimize
$$\sum_{i=1}^{n} (Y_i - \alpha - \beta X_i - \gamma X_i^2)^2$$

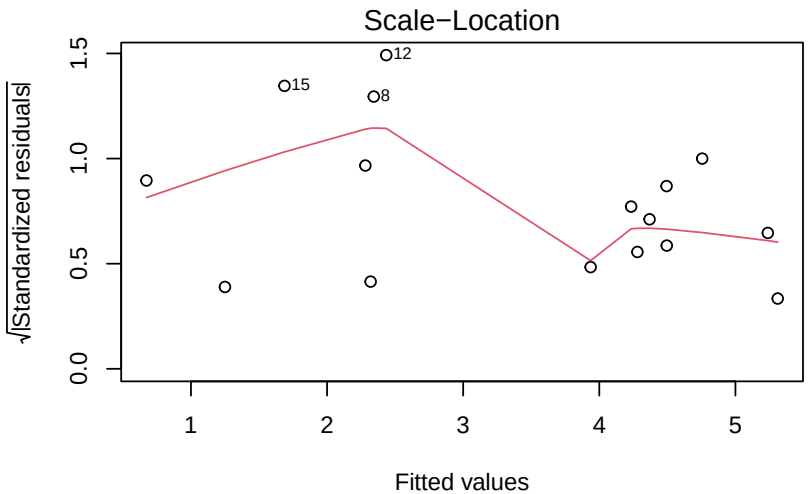
Quadratic regression

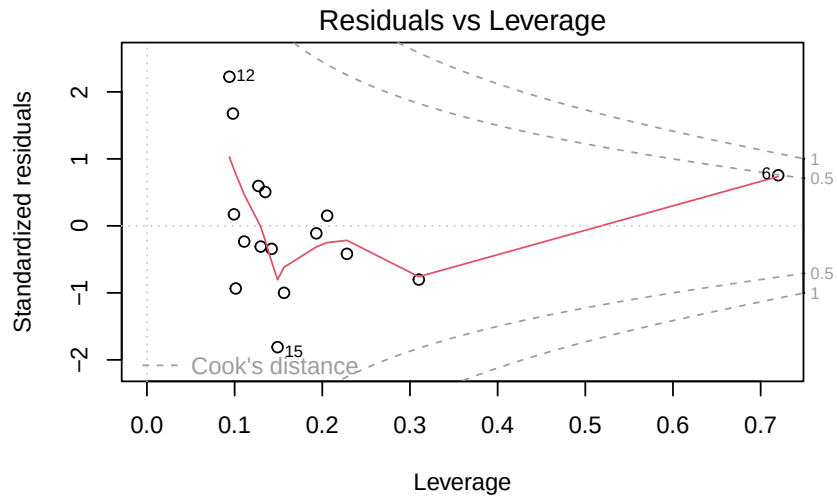
```
Call:
lm(formula = log(CWD.BASA) ~ RIP.DENS + I(RIP.DENS^2))

Residuals:
    Min      1Q  Median      3Q     Max
-1.6872 -0.4462 -0.1621  0.4214  2.1399

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.686e+00  3.114e+00  -3.110  0.00828 **
RIP.DENS       1.726e-02  4.673e-03   3.693  0.00270 **
I(RIP.DENS^2) -4.960e-06  1.628e-06  -3.047  0.00935 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.01 on 13 degrees of freedom
Multiple R-squared:  0.7159,    Adjusted R-squared:  0.6722
F-statistic: 16.38 on 2 and 13 DF,  p-value: 0.0002801
```

Residual plots









Building model proceeds hierarchically

```
                    Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)   -9.686141e+00 3.114228e+00 -3.110287 0.008281158
RIP.DENS       1.726034e-02 4.673452e-03  3.693275 0.002704501
I(RIP.DENS^2) -4.960374e-06 1.627703e-06 -3.047469 0.009345380
```

- Add terms to model and keep those as long as they are significantly associated with outcome
- Example: adding third order term is not significant contribution (p-value 0.26)
- Adding proceeds **hierarchically**: lower order terms are kept as long as higher order terms are in model

Results

- We conclude

$$E\{\ln(Y)|X\} = -9.69 + 0.017X - 4.96\ 10^{-6}X^2$$

  or equivalently that geometric mean CWD basal area for given tree density X is equal to

$$\exp(-9.69 + 0.017X - 4.96\ 10^{-6}X^2)$$

- For $X = 500$ we now find 0.086 m^2 (previously: -17 m^2)
- **How precise is this?**

Interpreting the results of a regression model

Inference for mean outcome

Given an input $X = x_h$, what do we expect the outcome Y to be on average?

Use the regression equation:

- $\hat{y}_h = \hat{\alpha} + \hat{\beta} x_h$ is unbiased estimator of $E(Y|X = x_h) = \alpha + \beta x_h$.

What is the uncertainty of this estimator?

For this we need the standard error of $\hat{Y}_h$.

Inference for mean outcome: uncertainty

- Standard error of $\hat{Y}_h$ is

$$SE(\hat{Y}_h) = \sqrt{MSE\left\{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right\}}.$$

- Tests and CI for $E(Y|X_h)$ based on

$$\frac{\hat{Y}_h - E(Y|X_h)}{SE(\hat{Y}_h)} \sim t_{n-p}$$

with p number of unknown parameters in model.

Inference for mean outcome: intuition

- Highest precision for predictions in $X_h = \bar{X}$: relative confidence in predictions for X **close to mean**.
- Lower precision as predictions have X further **away from the mean**.

Prediction in R

```
model3 <- lm(I(log(CWD.BASA)) ~ RIP.DENS + I(RIP.DENS^2))
p <- predict(model3,
             newdata = data.frame(RIP.DENS=c(1000, 1500)),
             interval = "confidence")
p
```

```
       fit      lwr      upr
1 2.613829 1.966541 3.261118
2 5.043534 4.149293 5.937775
```

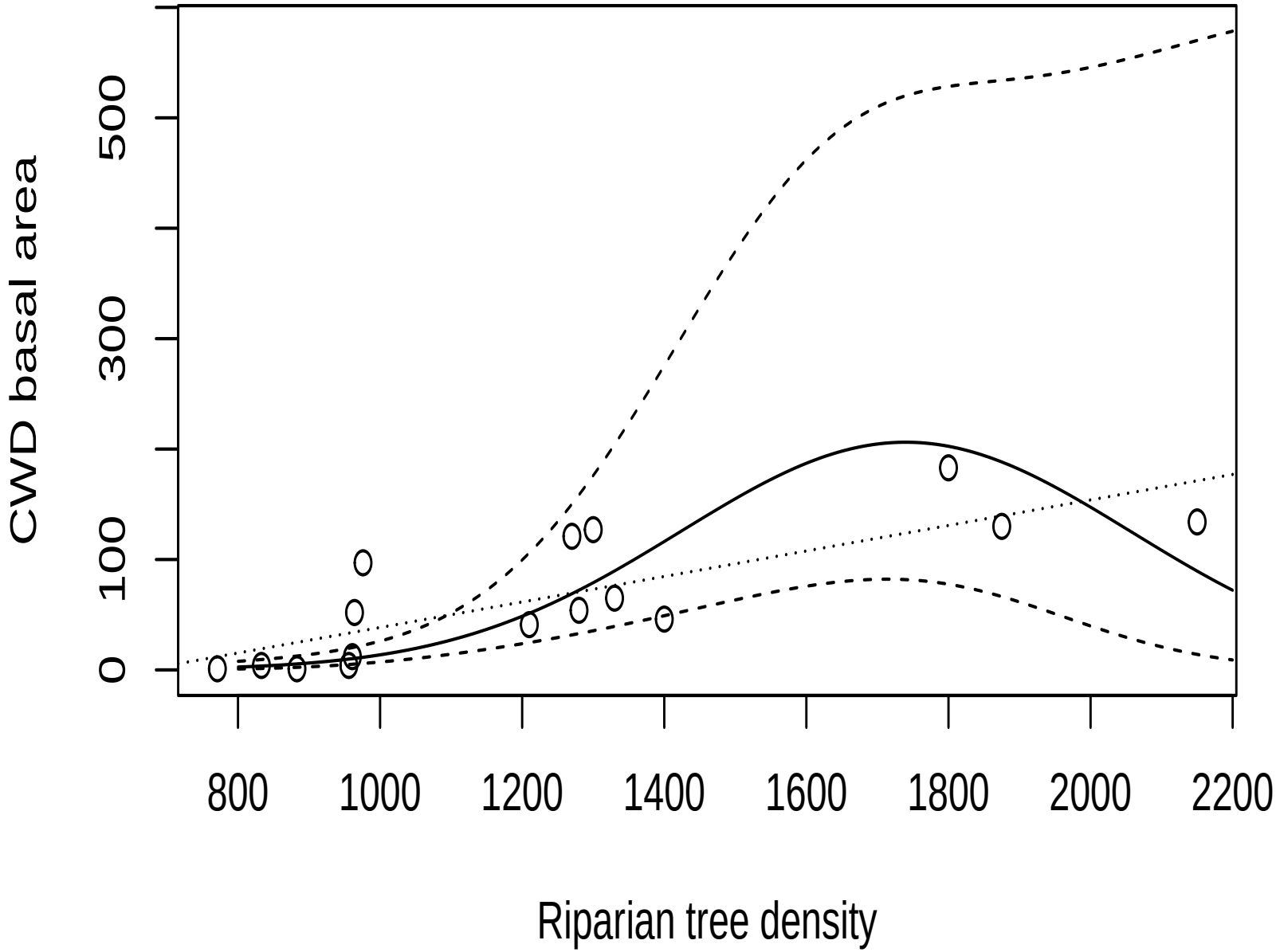
Prediction in R

Predictions and lower/upper bound of the CI are for `log(CWD.BASA)` and need to be transformed back:

```
exp(p)
```

```
         fit        lwr        upr
1   13.65123   7.145917   26.07867
2 155.01687 63.389137 379.09068
```

Expected outcome with 95% CI



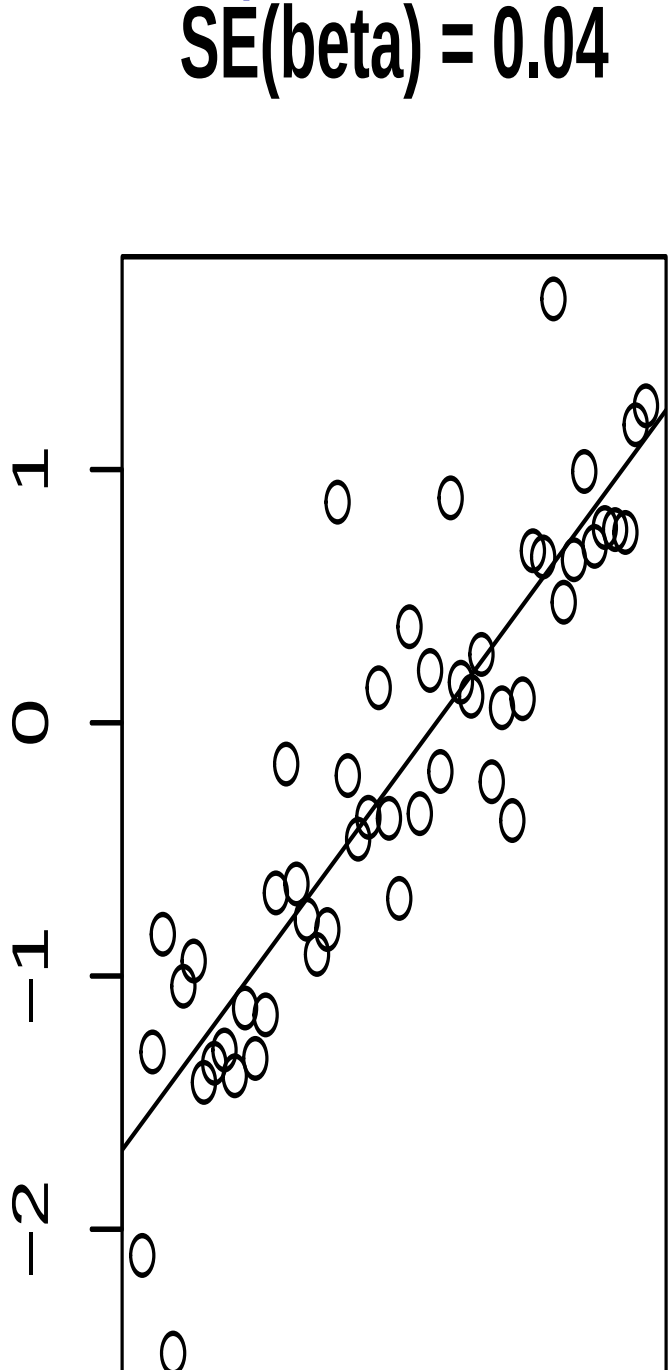
Inference for slope β

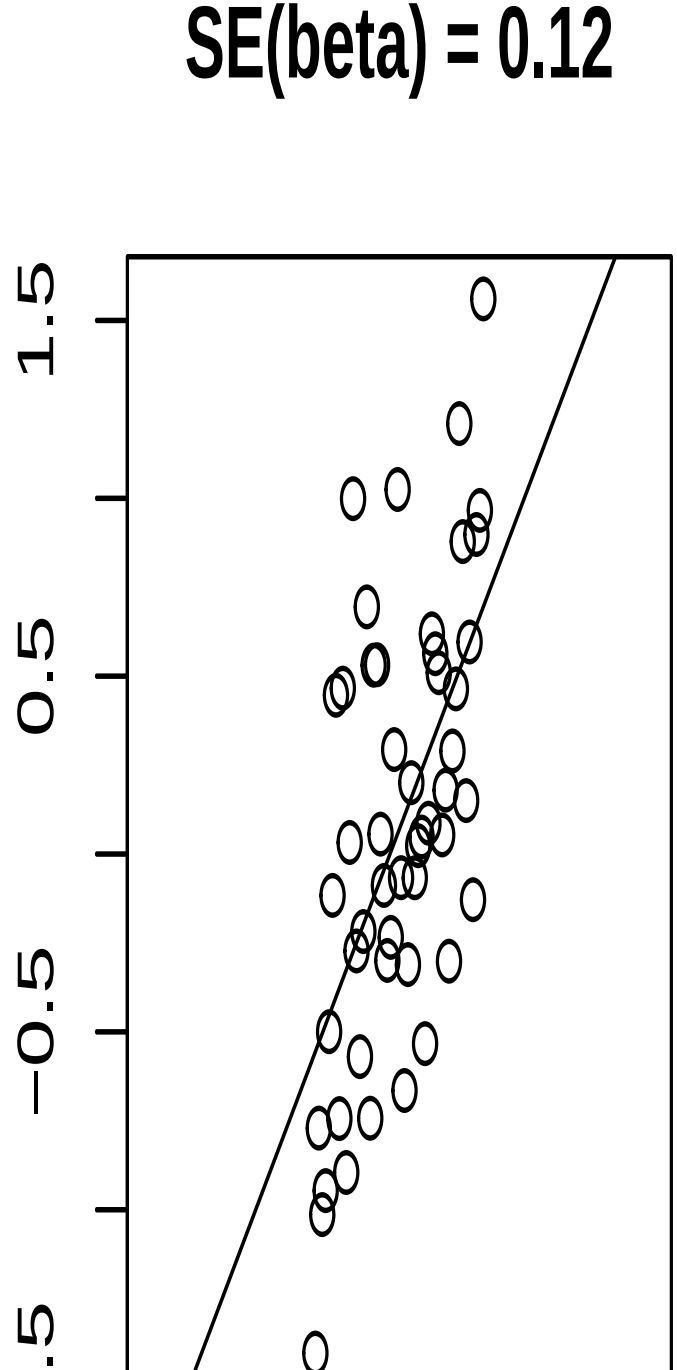
- The regression coefficient $\hat{\beta}$ is an (unbiased) estimator of β, the population regression coefficient.
- It comes with a measure of uncertainty: standard error of $\hat{\beta}$:

$$SE(\hat{\beta}) = \sqrt{\frac{MSE}{\sum_i (X_i - \bar{X})^2}}.$$

  with $MSE = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$
- Large spread on X improves precision.

Spread and precision





Association tree density vs. CWD

Tests and confidence intervals for β are based on

$$\frac{\hat{\beta} - \beta}{SE(\hat{\beta})} \sim t_{n-2}$$

```
summary(model)$coefficients
```

```
                Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) -77.0990778 30.60800907 -2.518918 0.0245520121
RIP.DENS      0.1155161  0.02343233  4.929772 0.0002216405
```

Association tree density vs. CWD

- 95% CI for β needs $t_{14,0.975} = 2.14$
- CI is given by

$$[0.116 - 2.14 \times 0.0234, 0.116 + 2.14 \times 0.0234] = [0.066, 0.166]$$

```
confint(model)
```

```
                    2.5 %      97.5 %
(Intercept) -142.74672817 -11.4514274
RIP.DENS       0.06525871   0.1657734
```