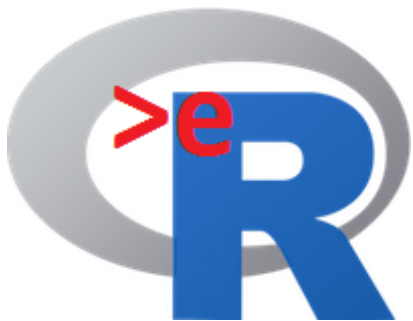




This course was developed as a part of the VLIR-UOS Cross-Cutting project s:

- Statistics: 2011-2012, 2013,2014-2015.
- Statistics: 2016/2017.
- Statistics for development : 2018-2020.



The >eR-Biostat initiative
Making R based education materials in
statistics accessible for all

Linear regression in R

Slides developed by Ziv Shkedy (Hasselt University, Belgium, July 2017)

based on an online course developed by
Marc Lavielle

Inria Saclay (Xpop) & Ecole Polytechnique (CMAP)
March, 2017



ER-BioStat



<https://github.com/eR-Biostat>

Email: erbiostat@gmail.com



@erbiostat

Contents

- Fitting polynomial models
 - Fitting a polynomial of degree 0
 - Fitting a polynomial of degree 1
 - Numerical results
 - Some diagnostic plots
 - The predictive performance of the model
 - Confidence interval and prediction interval
 - Fitting a polynomial of degree 2
 - Fitting a polynomial without intercept
 - Using orthogonal polynomials
- Model comparison
 - t-test
 - Analysis-of-variance (anova)
 - Likelihood ratio test (LRT)
 - Information criteria
- Data transformation
 - log based transformations
 - Diagnostic plots
 - Confidence interval and prediction interval
 - Model comparison

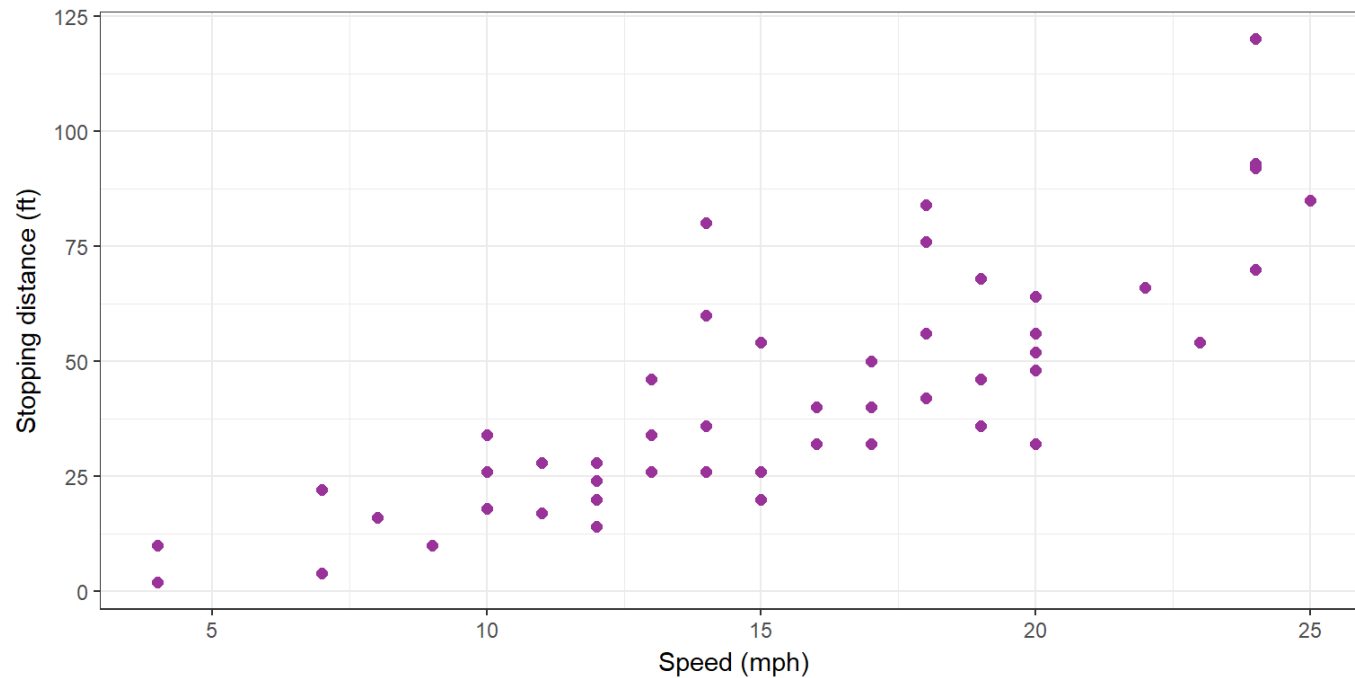
YouTube tutorials

- YouTube tutorials are available for:
 - Linear regression in R (host by Christoph Scherber): <https://www.youtube.com/watch?v=Xh6Rex3ARjc&t=838s>
 - Linear regression in R (host by Ani Katachova): <https://www.youtube.com/watch?v=fInEw5LTvxM>
 - Simple Linear regression in R (host by Mike Marin): https://www.youtube.com/watch?v=66z_MRwtFJM&list=PLqzoL9-eJTnBJrvFcN-ohc5G13E7Big0e
 - Checking Linear Regression Assumptions in R (host by Mike Marin): <https://www.youtube.com/watch?v=eTZ4VUZHxw>

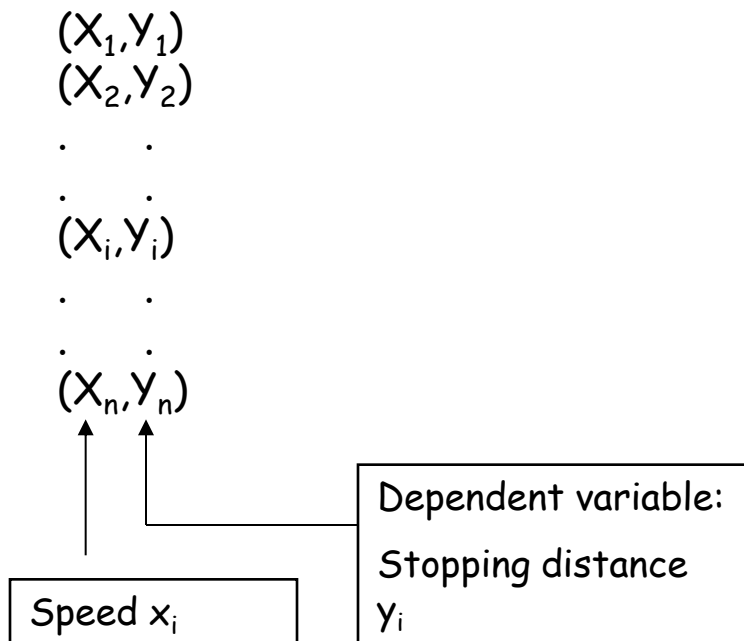
The Car data

- Cars' speed and stopping distance.
- Cars from the 1920s

The data



Data Structure



The cars data in R

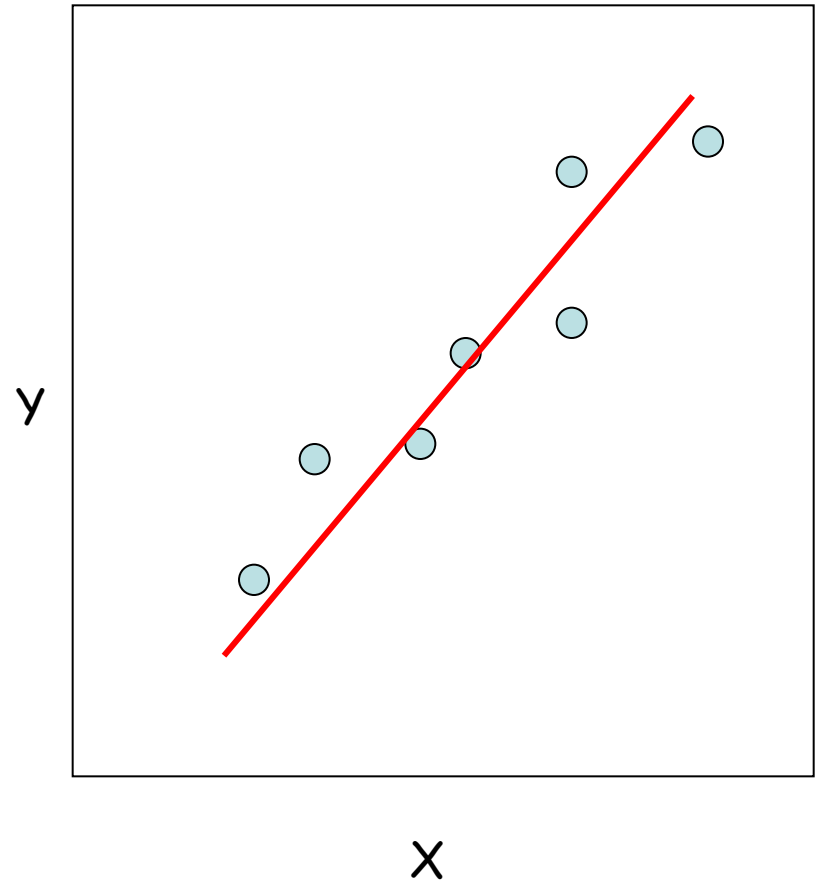
```
> data(cars)
> head(cars)
  speed  dist
1     4     2
2     4    10
3     7     4
4     7    22
5     8    16
6     9    10
```

Speed x_i

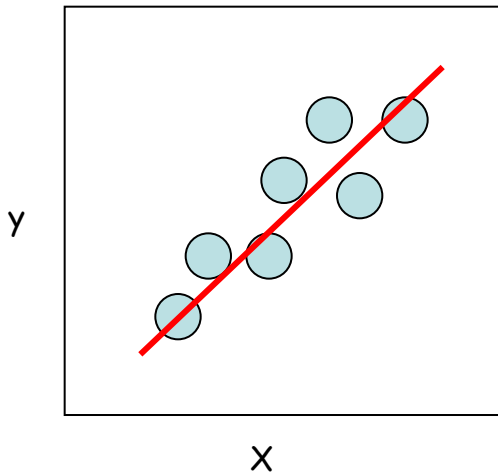
Stopping Distance y_i

What is a Simple Linear Regression Model ?

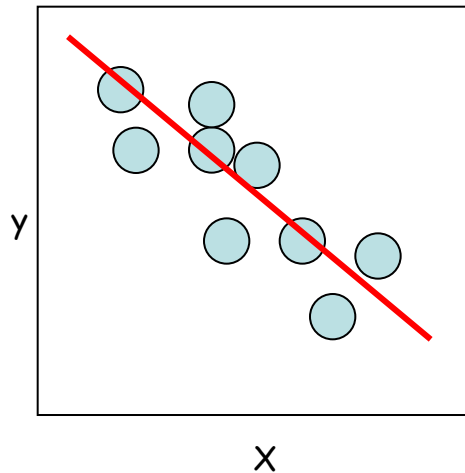
- A regression model is a statistical model which aims to describe the relationship between a predictor (the dose level) and the dependent variable (test score) with a **straight line**.



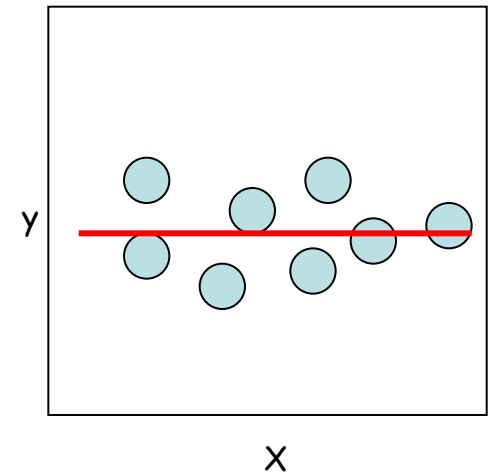
Properties of Simple Linear Regression Models : Trends



Upward trend

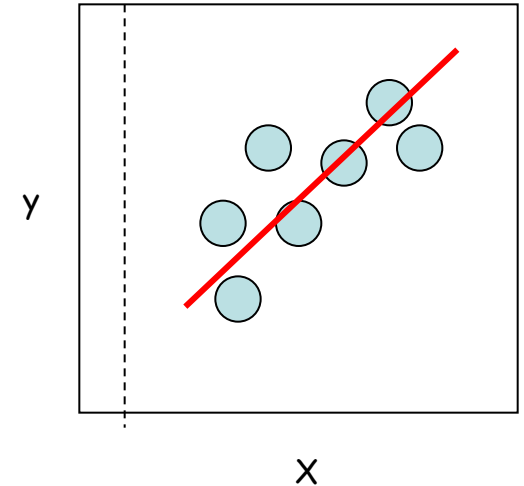
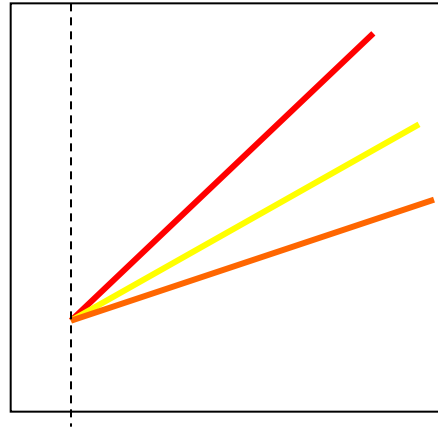


Downward trend

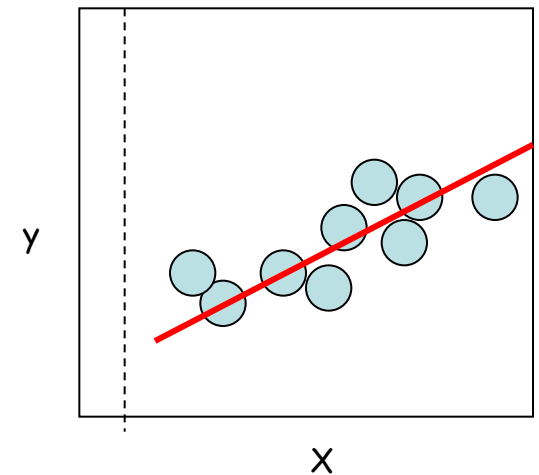


Y does not depend on
X

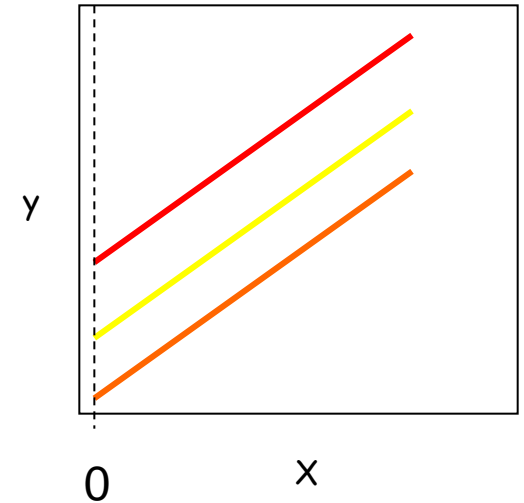
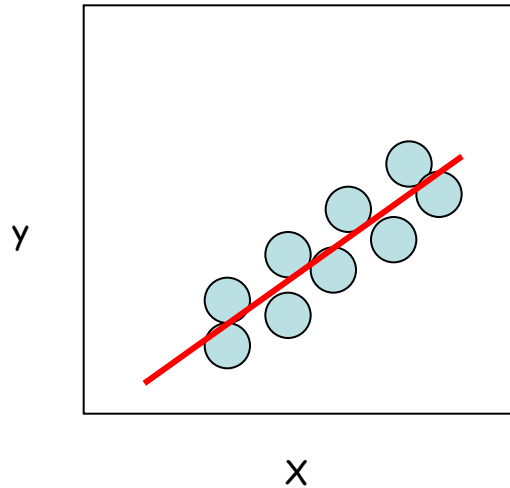
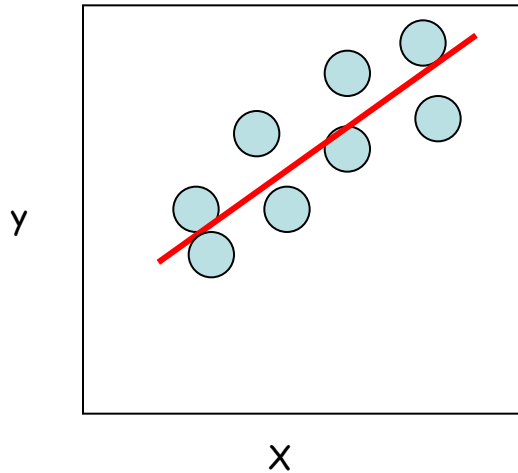
Properties of Simple Linear Regression Models : Slope



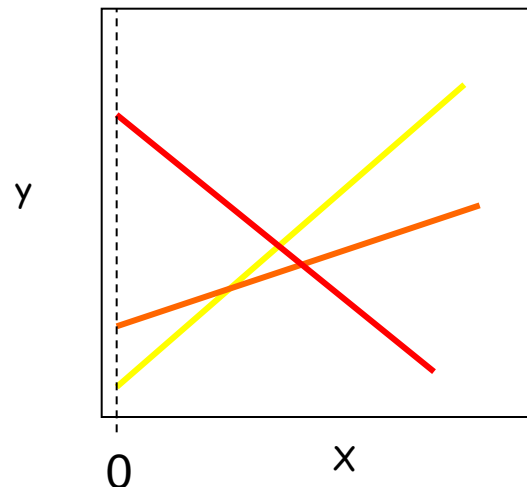
The slope is the change in the mean of Y for a unit change in X.



Properties of Simple Linear Regression Models : Intercept

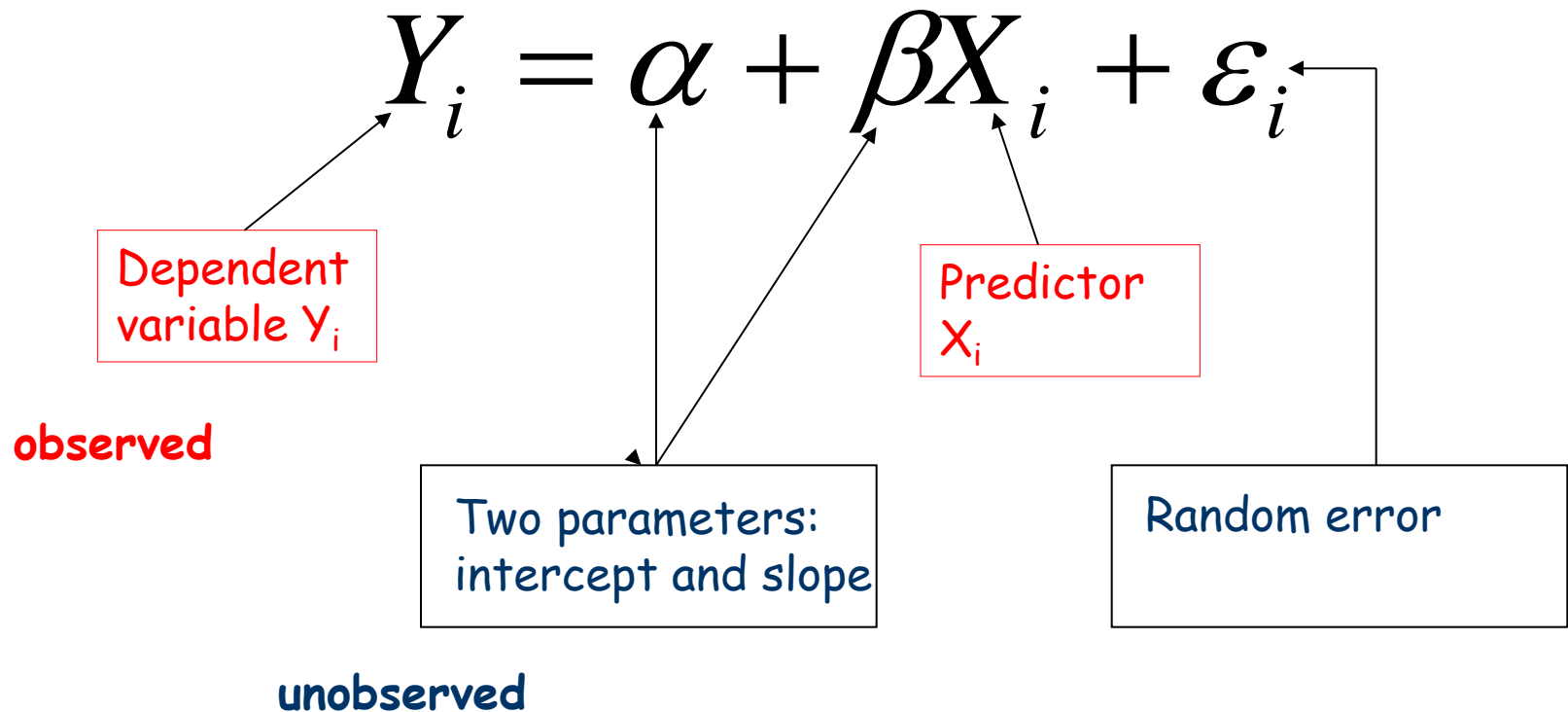


The intercept is the height of the regression line when $x=0$.



A Simple Linear Regression Model

- We assume that the relationship between the predictor and the response can be describe with the model:



Estimation (I)

- We need to estimate the unobserved parameters of the model:
- The estimator for the random error:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$
$$\hat{Y}_i = a + bX_i$$
$$e_i = \hat{Y}_i - Y_i$$

The diagram illustrates the estimation process. It shows the true model $Y_i = \alpha + \beta X_i + \varepsilon_i$ and the estimated model $\hat{Y}_i = a + bX_i$. The parameters α and β are represented by pink and orange vertical bars, respectively, while their estimators a and b are represented by pink and orange text. A blue box represents the random error ε_i , and its estimator, the residual e_i , is shown in a blue box. A box on the right defines the residual $e_i = \hat{Y}_i - Y_i$. Arrows indicate that a estimates α and b estimates β , and that the residual e_i is the difference between the predicted value and the observed value.

predicted value for
the test score

(the estimator for
stopping distance)

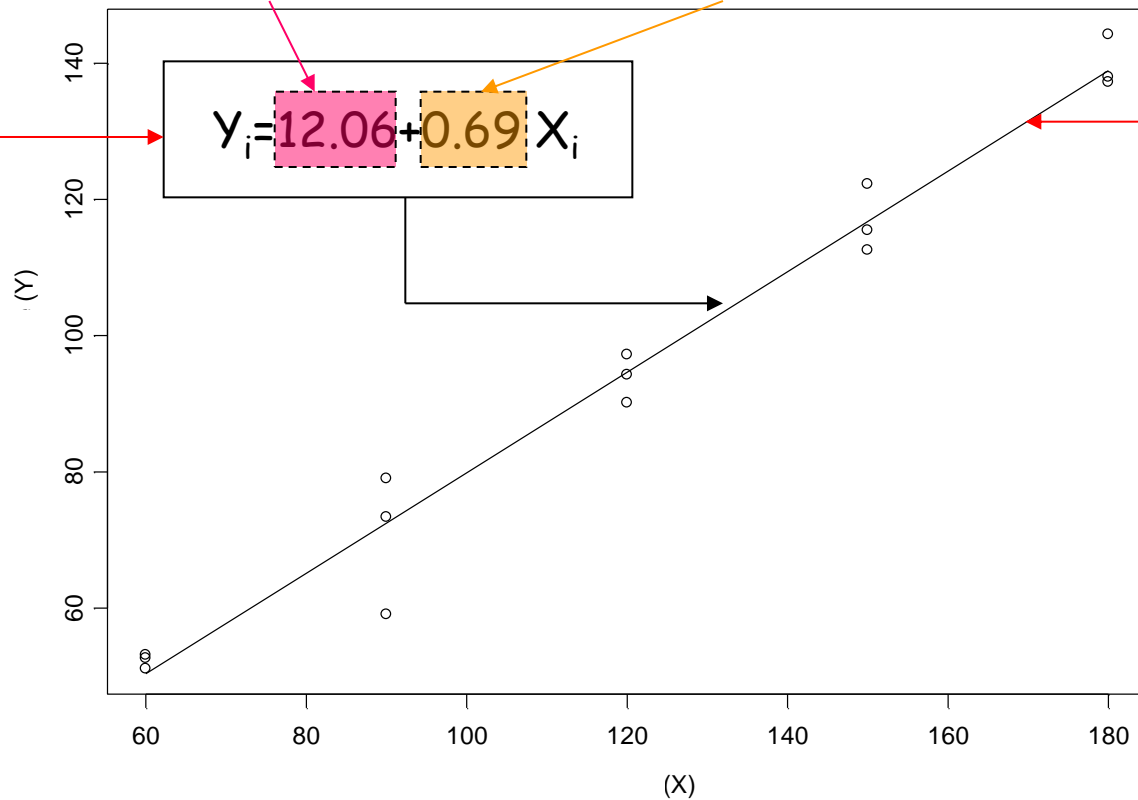
- a and b are the estimators for alpha and beta
- e_i (the residual) is the estimator for the random error

Regression Model and Data

a: the estimate
for alpha

b: the estimate
for beta

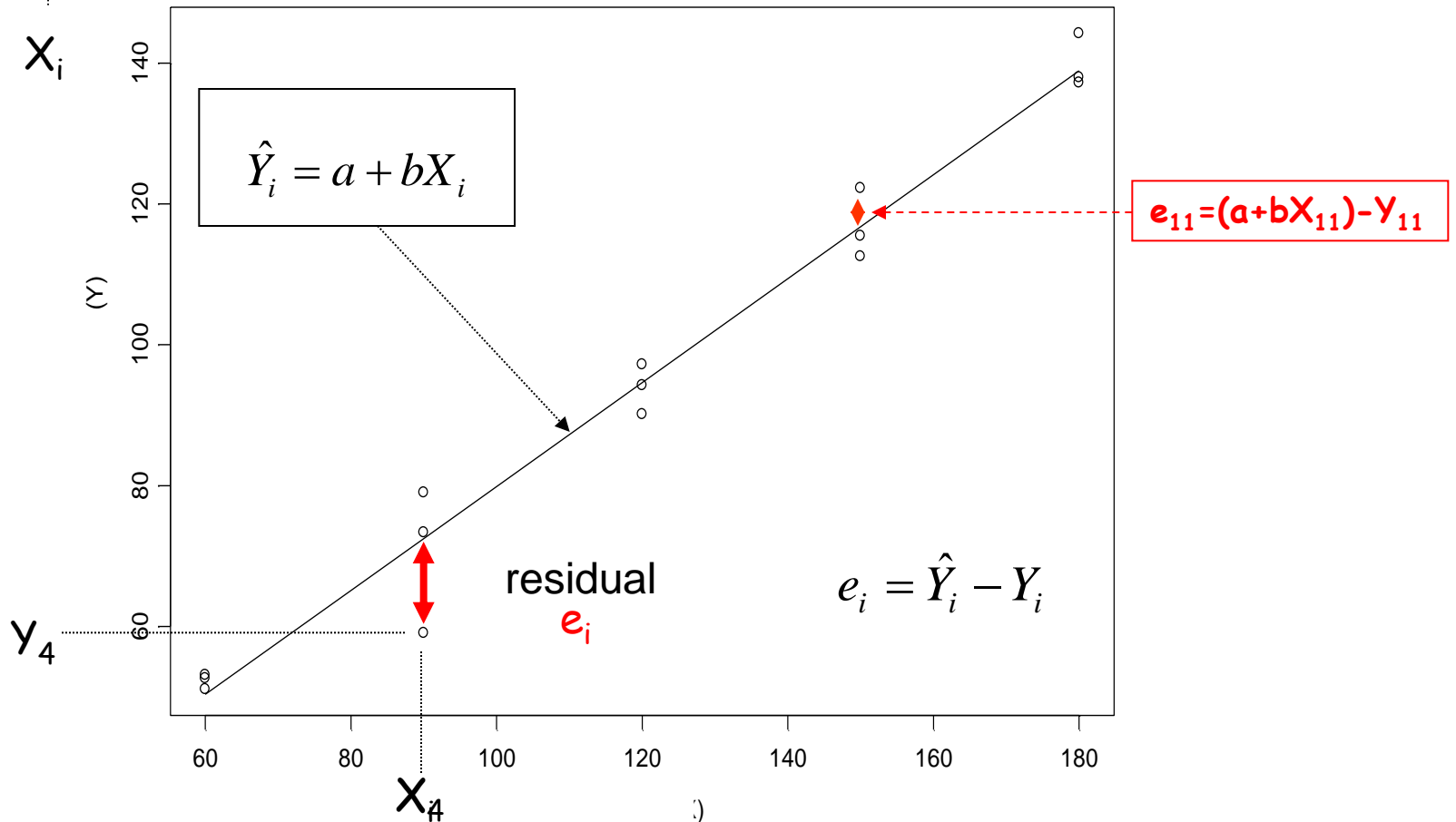
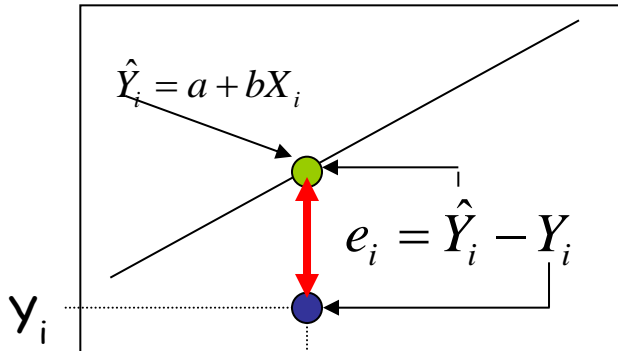
The
estimated
(fitted)
model



The regression
line

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad \text{The model}$$

The Residuals



Estimation (II): The Least Squares Criterion

- How to estimate the intercept and slope?
- We want that the fitted model (the line which describes the relationship between Y and X) will be “close” to the data.
- The residual sum of squares = $\text{sum}(\text{residual})^2$.
- The least squares criterion: choose intercept and slope which minimize the residual sum of squares

$$RSS = \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$$

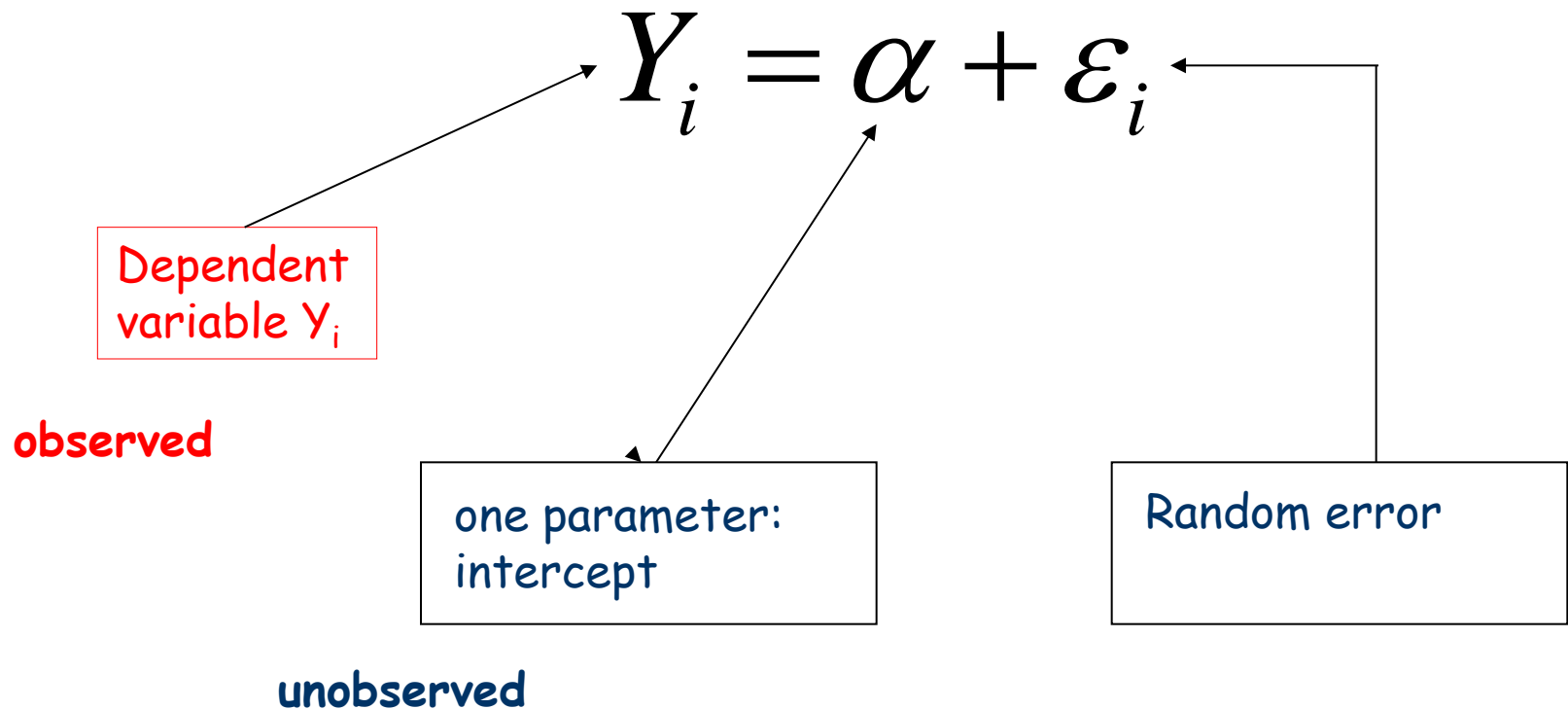


contents

- Fitting polynomial models
 - **Fitting a polynomial of degree 0**
 - Fitting a polynomial of degree 1
 - Numerical results
 - Some diagnostic plots
 - The predictive performance of the model
 - Confidence interval and prediction interval
 - Fitting a polynomial of degree 2
 - Fitting a polynomial without intercept
 - Using orthogonal polynomials
- Model comparison
 - t-test
 - Analysis-of-variance (anova)
 - Likelihood ratio test (LRT)
 - Information criteria
- Data transformation
 - log based transformations
 - Diagnostic plots
 - Confidence interval and prediction interval
 - Model comparison

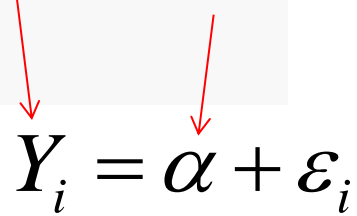
A Simple Linear Regression Model

- We assume that the relationship between the predictor and the response is constant and can be describe with the model:



Polynomial of degree 0 in R

```
attach(cars)  
lm0 <- lm(dist ~ 1)
```


$$Y_i = \alpha + \varepsilon_i$$

R output for the estimated model

```
> summary(lm0)
```

Call:

```
lm(formula = dist ~ 1)
```

Residuals:

Min	1Q	Median	3Q	Max
-40.98	-16.98	-6.98	13.02	77.02

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	42.980	3.644	11.79	6.38e-16 ***

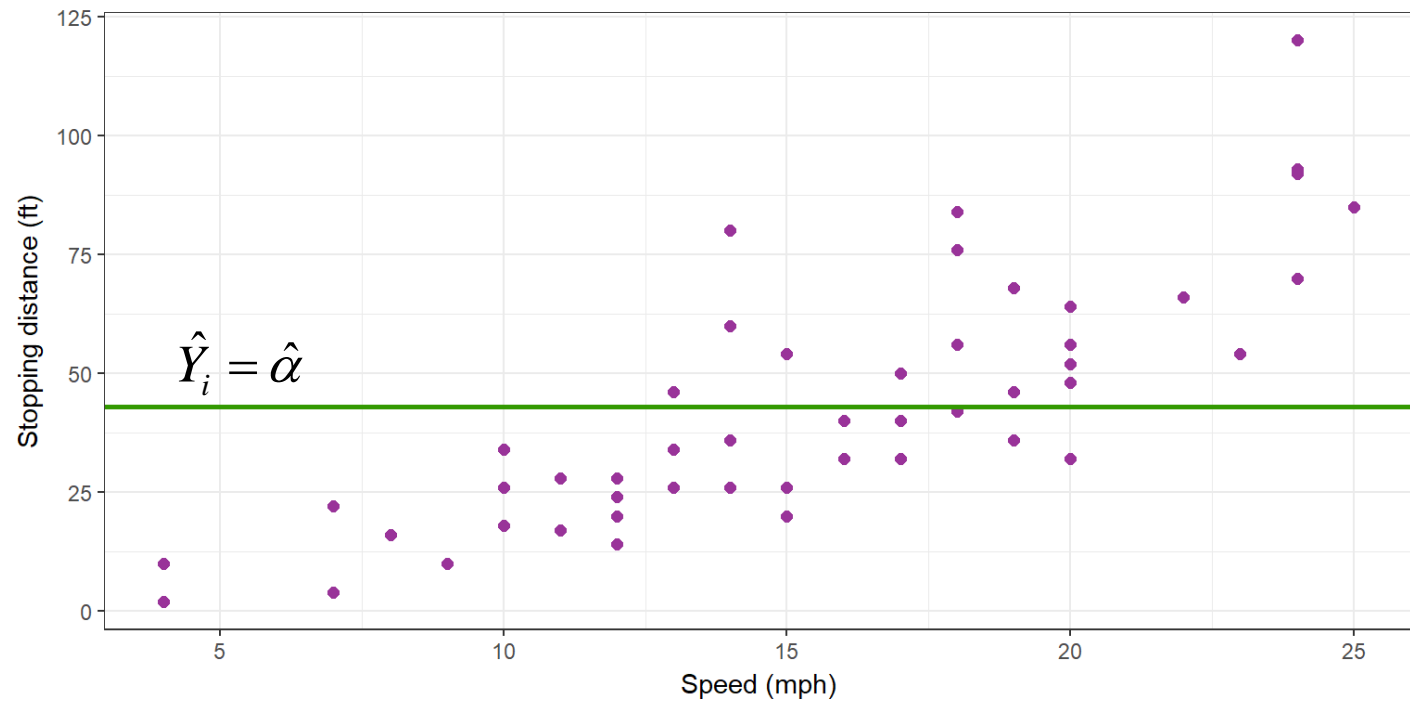
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.77 on 49 degrees of freedom

$\hat{\alpha}$



Data and estimated model



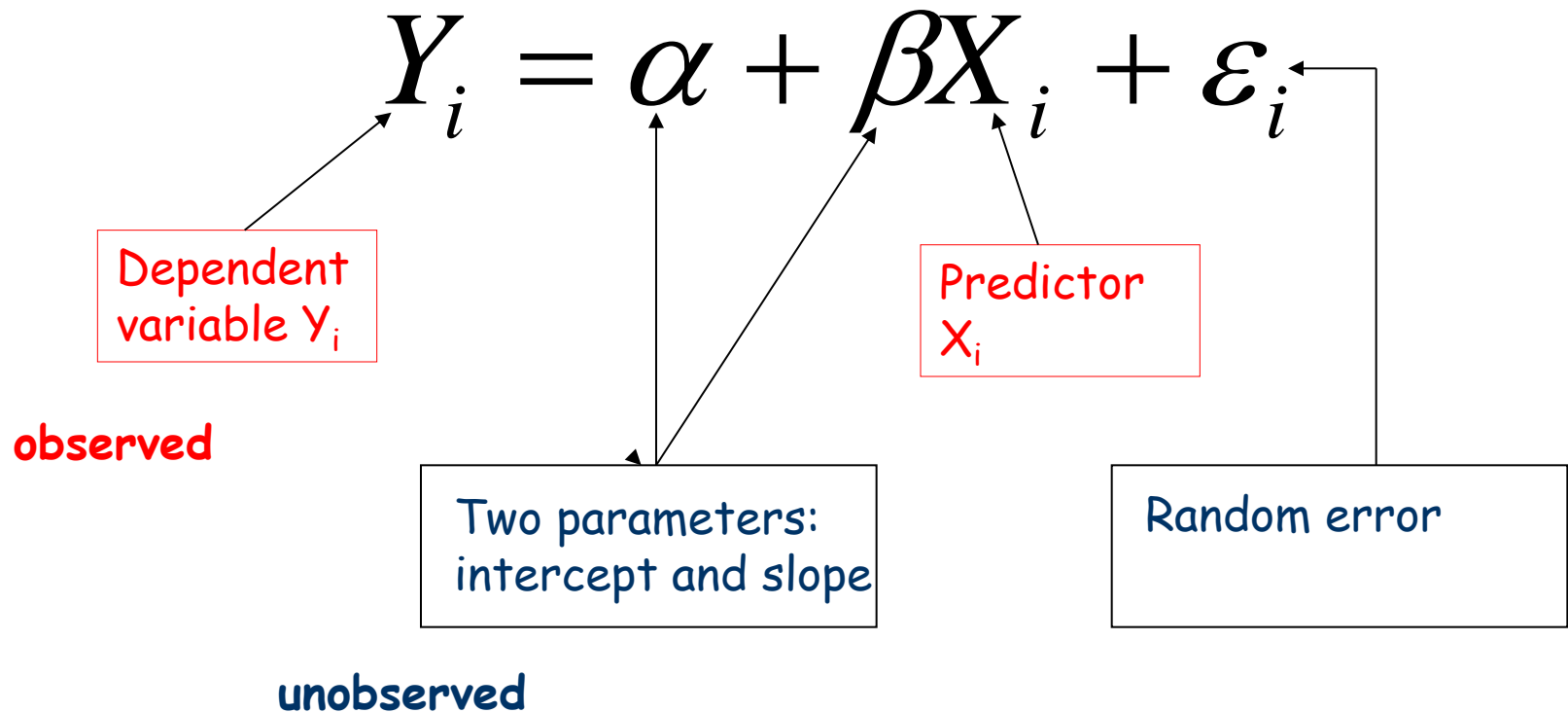


contents

- Fitting polynomial models
 - Fitting a polynomial of degree 0
 - **Fitting a polynomial of degree 1**
 - **Numerical results**
 - Some diagnostic plots
 - The predictive performance of the model
 - Confidence interval and prediction interval
 - Fitting a polynomial of degree 2
 - Fitting a polynomial without intercept
 - Using orthogonal polynomials
- Model comparison
 - t-test
 - Analysis-of-variance (anova)
 - Likelihood ratio test (LRT)
 - Information criteria
- Data transformation
 - log based transformations
 - Diagnostic plots
 - Confidence interval and prediction interval
 - Model comparison

A Simple Linear Regression Model

- We assume that the relationship between the predictor and the response can be describe with the model:



A simple linear regression in R

```
lm1 <- lm(dist ~ speed, data=cars)
```

```
coef(lm1)
```

```
## (Intercept)
```

```
## -17.579095
```

```
speed
```

```
3.932409
```

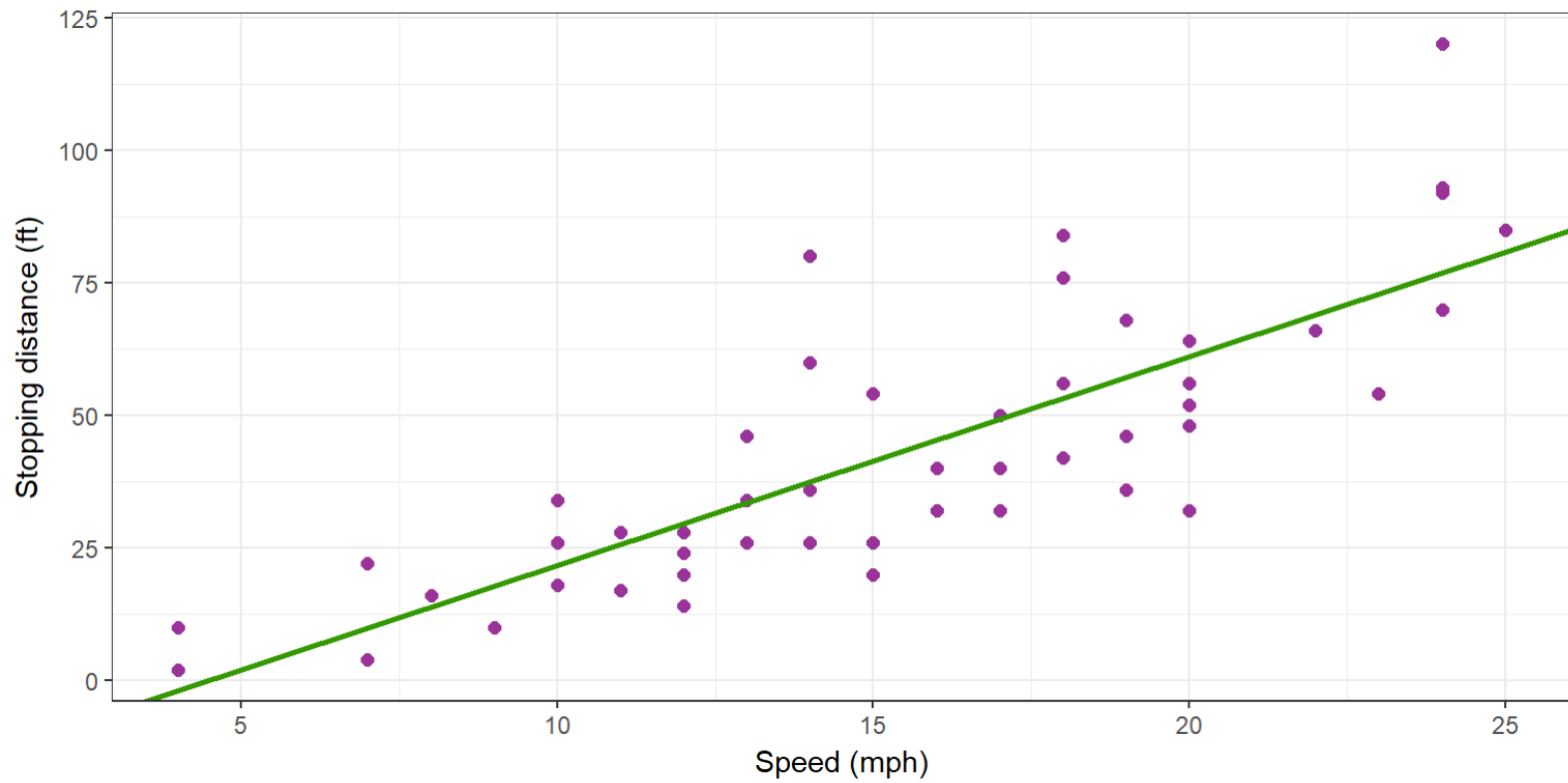


$\hat{\alpha}$



$\hat{\beta}$

Data and estimated model



Estimated model in R

```
> summary(lm1)
```

Call:

```
lm(formula = dist ~ speed, data = cars)
```

Residuals:

Min	1Q	Median	3Q	Max
-29.069	-9.525	-2.272	9.215	43.201

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-17.5791	6.7584	-2.601	0.0123	*
speed	3.9324	0.4155	9.464	1.49e-12	***

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

Residual standard error: 15.38 on 48 degrees of freedom

Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438

F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12

Confidence interval for the parameters

```
> confint(lm1)
```

	2.5 %	97.5 %
(Intercept)	-31.167850	-3.990340
speed	3.096964	4.767853

Summary

Technical Details (Estimation)

- A simple linear regression model has the form:

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

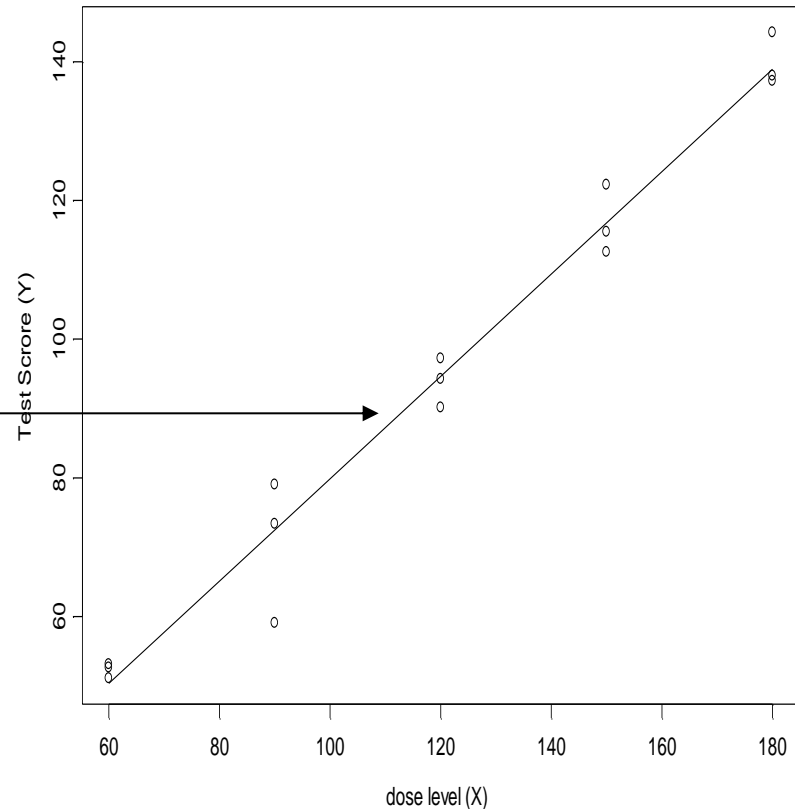
- α and β are the parameters in the model and ε is the random error.
- We can estimate α and β by minimizing the residual sum of squares

$$RSS = \sum_{i=1}^n (\alpha - \beta X_i - Y_i)^2$$

Technical Details (Estimation)

- The estimated model

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



- The residual

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

Technical Details (Estimation)

We assume that the relationship between Y_i and X_i can be described with a statistical model $Y_i = \alpha + \beta X_i + \varepsilon_i$

We assume that the random error \mathcal{E} is normally distributed.	$\varepsilon \sim N(0, \sigma^2)$
The mean of \mathcal{E} is equal to zero	$E(\varepsilon_i) = 0$
The conditional mean of Y_i (given the value of X_i)	$E(Y_i X_i) = \alpha + \beta X_i$
The estimator for the conditional mean of Y_i (the fitted model=the regression line)	$\hat{E}(Y_i X_i) = a + bX_i = \hat{Y}_i$
The residual: the estimator for \mathcal{E}	$e_i = \hat{Y}_i - Y_i$
Least square criterion: choose a and b that minimize the residuals sum of squares	$RSS = \sum_{i=1}^n (\alpha - \beta X_i - Y_i)^2$



contents

- Fitting polynomial models
 - Fitting a polynomial of degree 0
 - Fitting a polynomial of degree 1
 - Numerical results
 - **Some diagnostic plots**
 - The predictive performance of the model
 - Confidence interval and prediction interval
 - Fitting a polynomial of degree 2
 - Fitting a polynomial without intercept
 - Using orthogonal polynomials
- Model comparison
 - t-test
 - Analysis-of-variance (anova)
 - Likelihood ratio test (LRT)
 - Information criteria
- Data transformation
 - log based transformations
 - Diagnostic plots
 - Confidence interval and prediction interval
 - Model comparison

Simple regression model and its assumptions

We focus on model diagnostic. We consider the following **linear** regression model

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

The random error is **assumed** to be normal distributed:

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

We also assume that the variance is constant, i.e., the variance of $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ is equal.

How to check the model assumptions? (1)

- The random error, ε_i , is unknown but we can estimate ε_i with the residuals
- The residuals can be used in order to check the model assumptions.

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

Observed Predicted

- We focus on two things:
 - 1) the distribution of e_i
 - 2) the variability of e_i

How to check the model assumptions? (2)

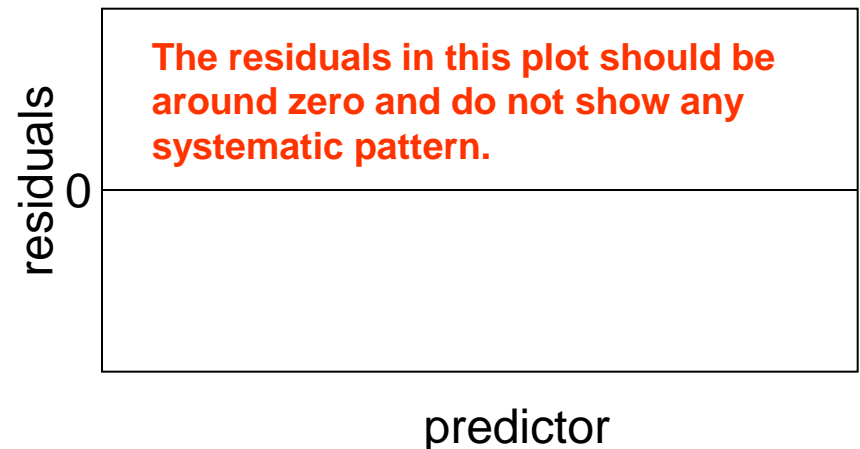
- We assume that the mean of Y_i is linear with respect to X :

$$E(Y)_i = \alpha + \beta \times X_i$$

- This is true only if

$$E(\varepsilon_i) = 0$$

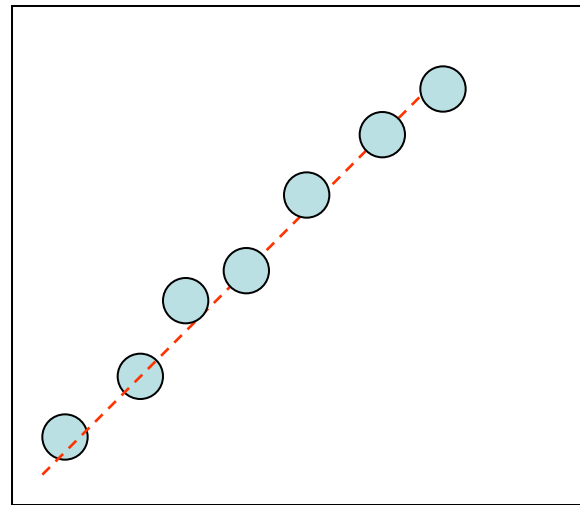
- Once again, the residuals can be used in order to check the linearity assumption .



Assumption 1: The distribution of e_i

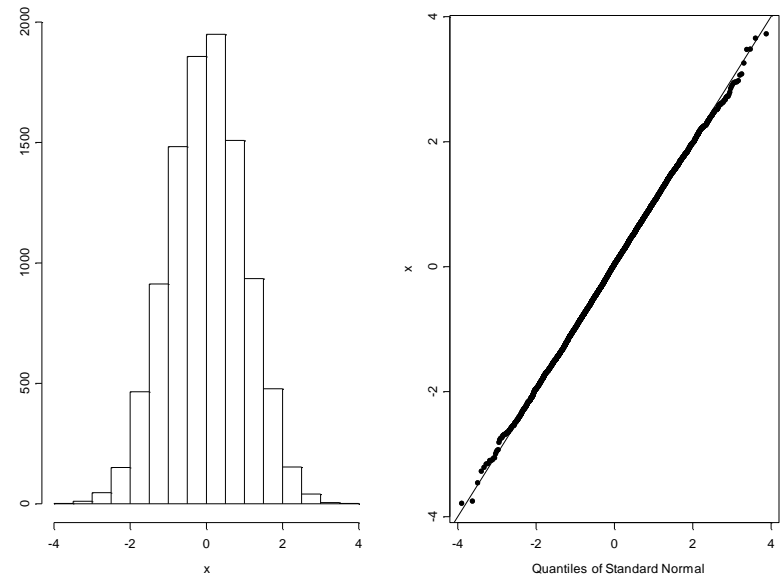
- The distribution of e_i is expected to be normal with mean zero and variance σ^2
- qq-normal plot (or normal probability plot) is a graphical tool that can be used in order to assess the normality assumption.

If the normality assumption holds we expect qq-normal plot will be a straight line.



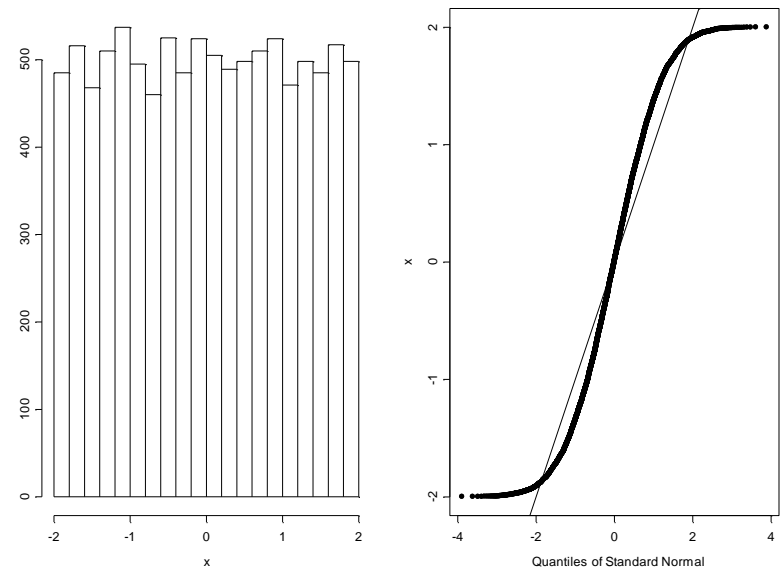
Example of qqnormal plot form $N(0,1)$

- Sample of 10000 observations from $N(0,1)$
- The qqnormal plot is a stright line.
- If the random error ε_i is normal distributed, the qqnormal plot of the residuals should be a stright line.



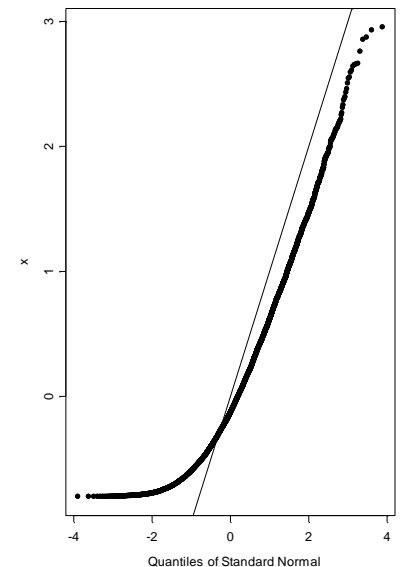
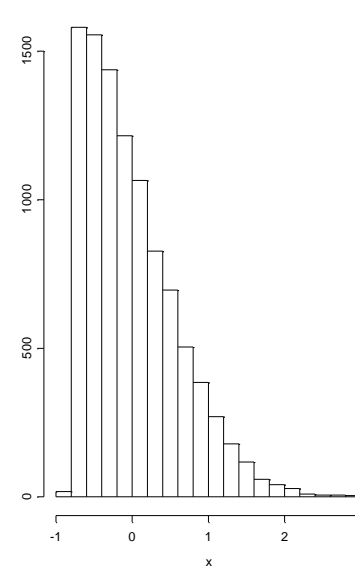
Example of qqnormal plot from heavy tailed distribution

- Sample of 10000 observations from $U(-2,2)$.
- S shape of the qqnormal plot.
- This is an example of a symmetric distribution with more observations (relatively to the normal distribution) at the tails.



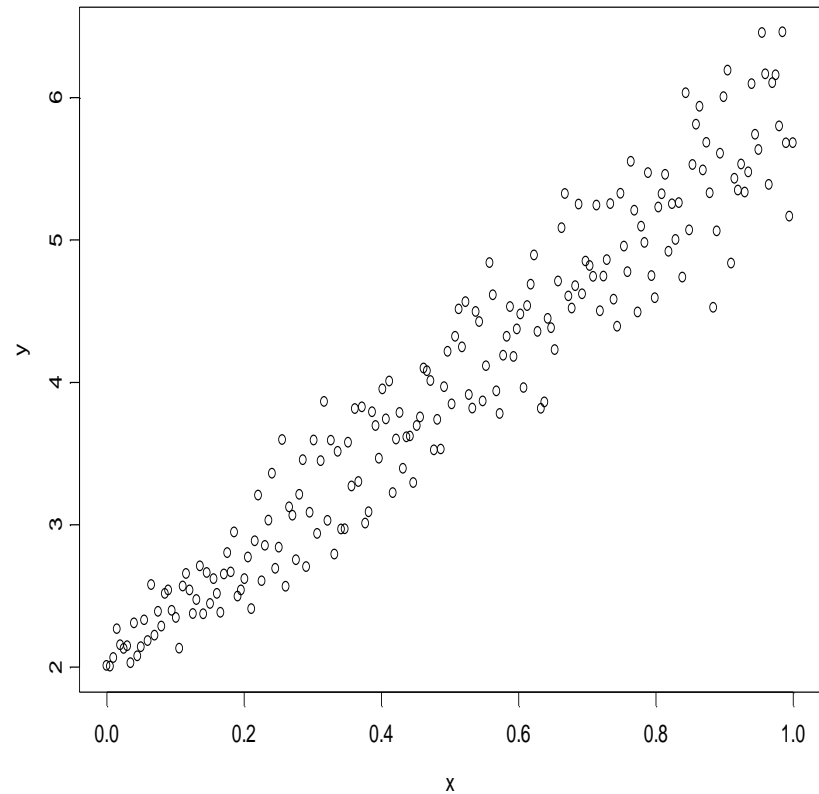
Example of qqnormal plot from skewed distribution

- Sample of 10000 observations from a skewed distribution.
- The distribution is skewed to the right and the points in the qqplot are not follow the straight line.



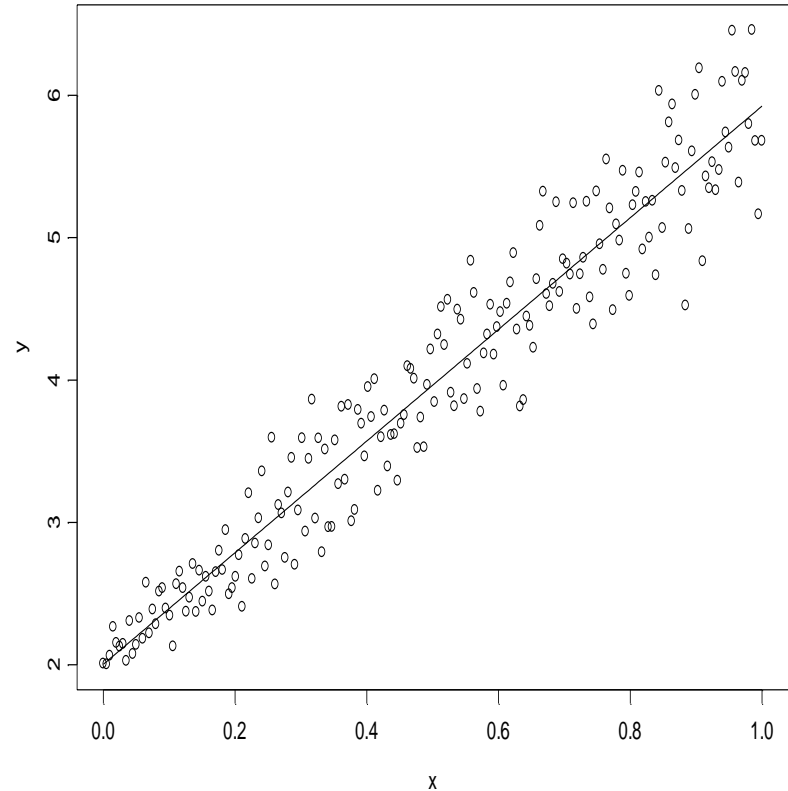
Assumption 2: Constant variance

- This is an example of a dataset in which the variance is not constant.
- The variance increases when the value of X increases.
- However, there is a linear relationship between the predictor and the response.



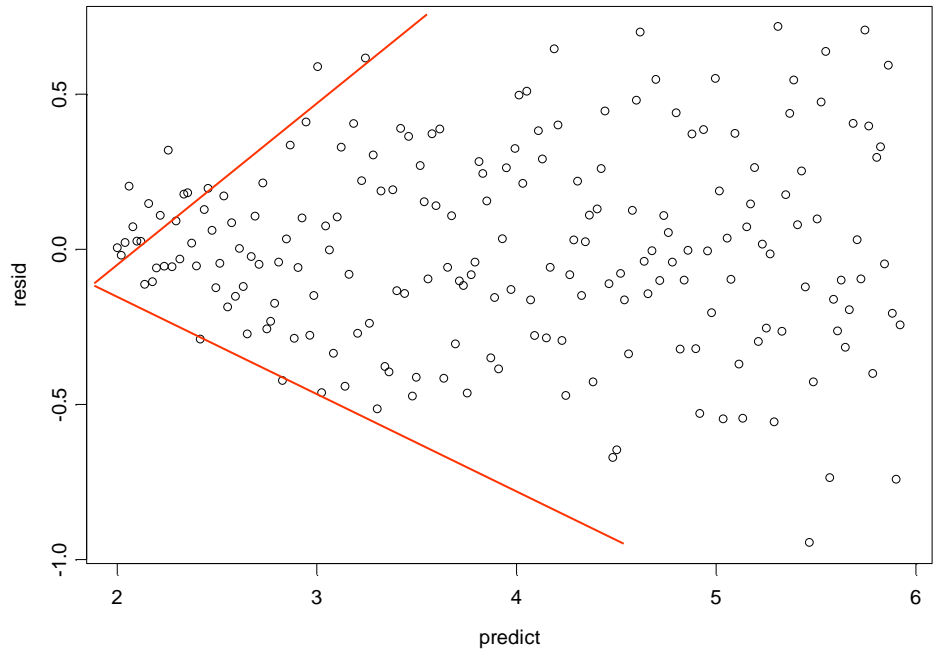
The data and the resrsson line

- The model seems to fit the data well in the sense that it captures that structure of the mean.

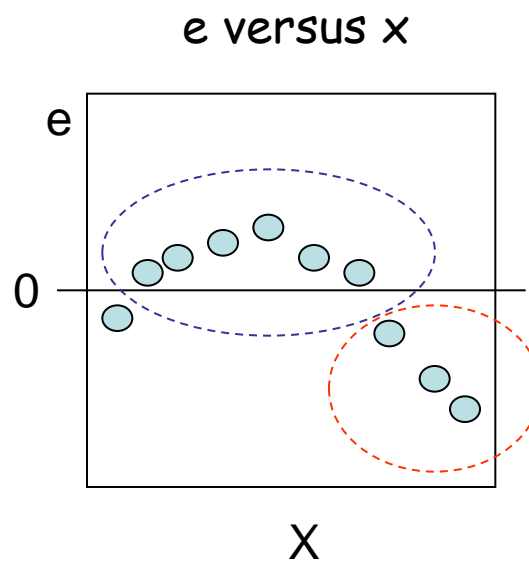
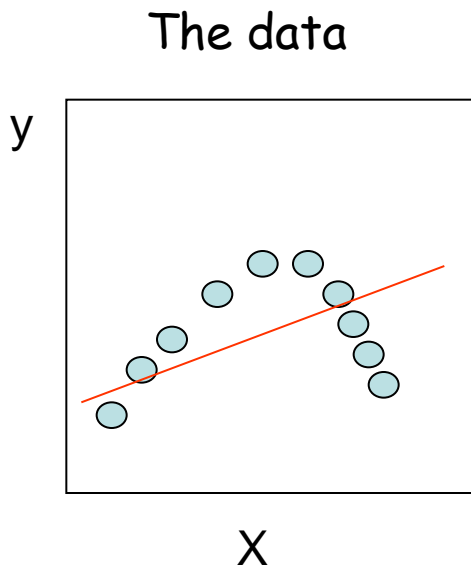


Residuals plot: Residuals versus the predicted values

- In this plot we can see clearly a pattern. As the predicted values increase the variability among the residuals increase (a “megaphone” shape).



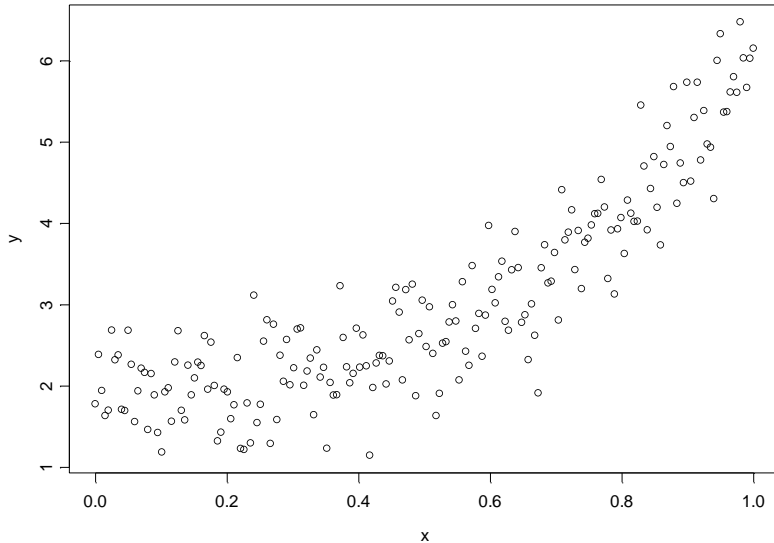
Assumptio 3: Linearity



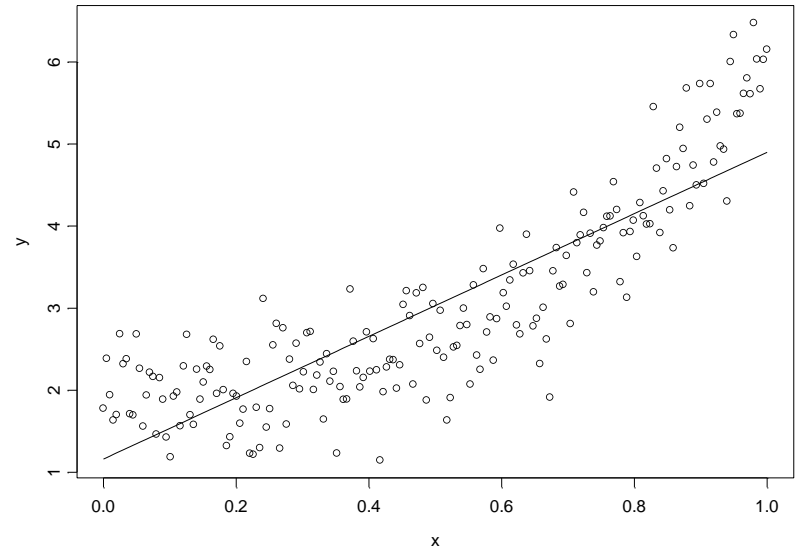
Example of
systematic
pattern in the
residual plot.

The scatterplot of the data reveals that the association between the response and the predictor is not linear. The residuals plot (in the right) reveals a clear pattern among the residuals which depends on the value of X.

Systematic patterns



data

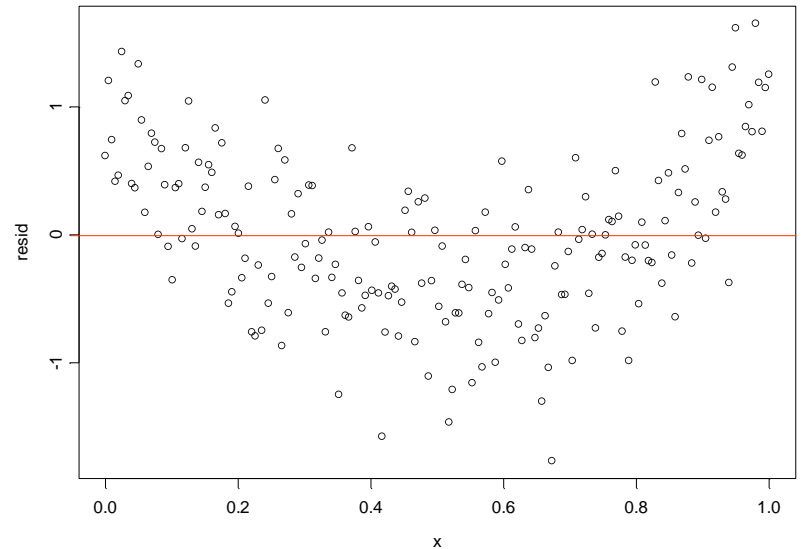


data and fitted model

The model underestimates the value of Y when the value of X is relatively small or large.

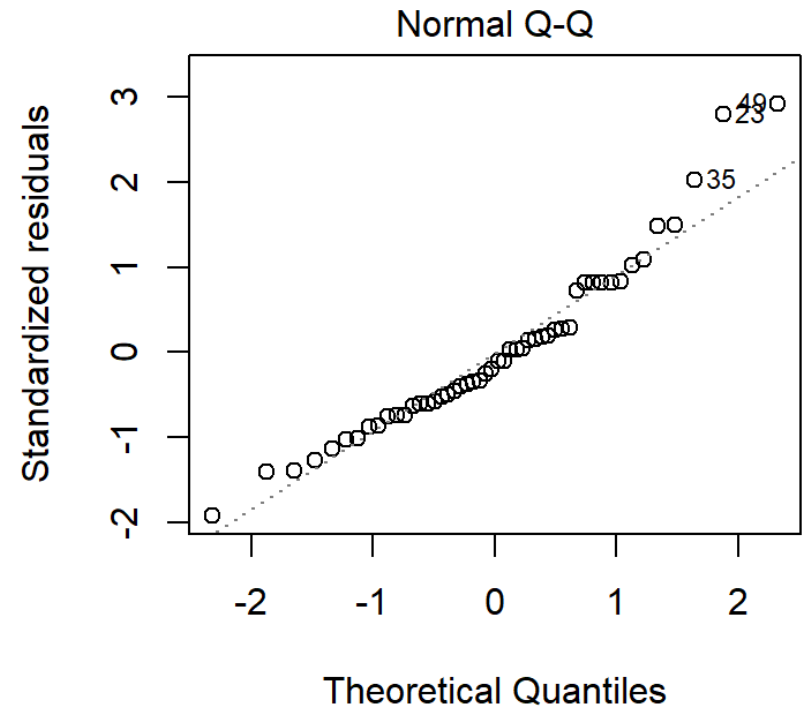
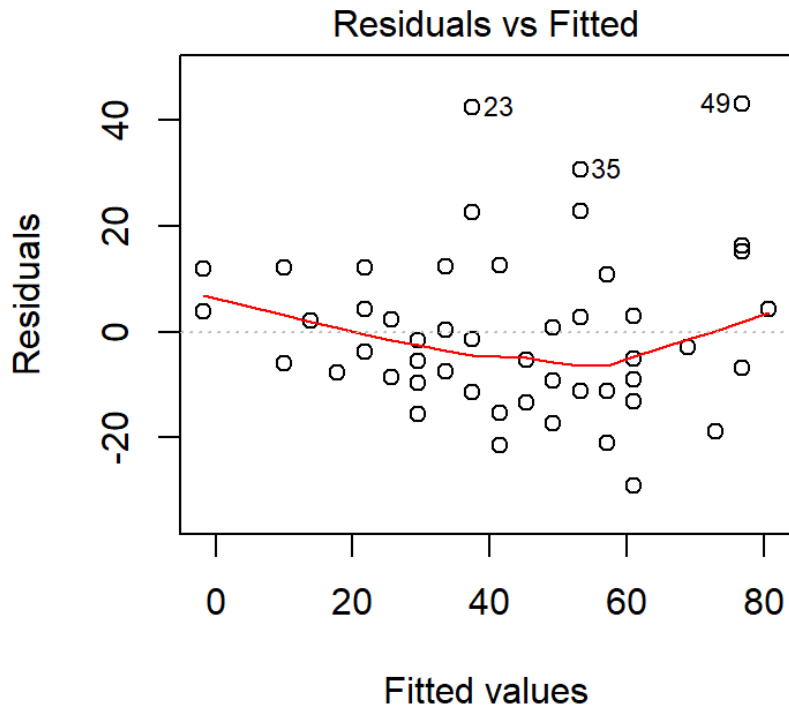
Linearity: Residuals plot

This plot reveals clear systematic pattern among the residuals: the residuals are positive for small and large value of X and negative in the middle. This means that there is structure in the data that the linear regression model did not capture.



Diagnostic plots in R

```
lm1 <- lm(dist ~ speed, data=cars)  
par(mfrow = c(1, 2))  
plot(lm1, which=c(1,2))
```





contents

- Fitting polynomial models
 - Fitting a polynomial of degree 0
 - Fitting a polynomial of degree 1
 - Numerical results
 - Some diagnostic plots
 - **The predictive performance of the model**
 - Confidence interval and prediction interval
 - Fitting a polynomial of degree 2
 - Fitting a polynomial without intercept
 - Using orthogonal polynomials
- Model comparison
 - t-test
 - Analysis-of-variance (anova)
 - Likelihood ratio test (LRT)
 - Information criteria
- Data transformation
 - log based transformations
 - Diagnostic plots
 - Confidence interval and prediction interval
 - Model comparison

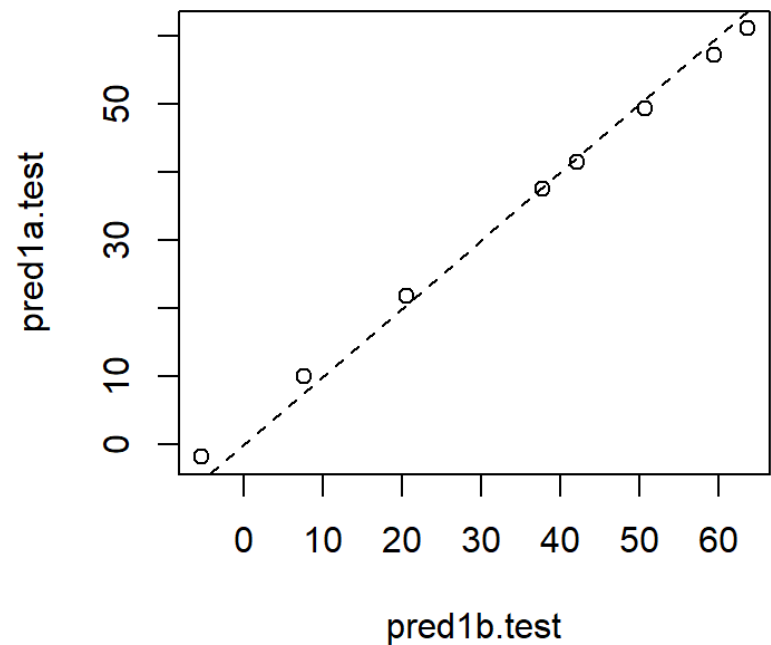
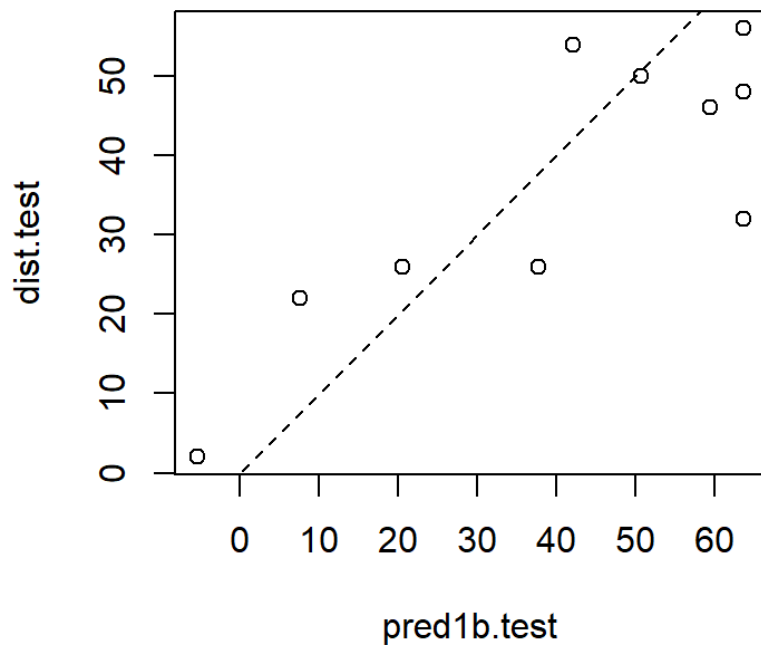
Cross-validation for prediction

- A common practice is to split the dataset into a 80:20 sample (training:test), then, build the model on the 80% sample and then use the model thus built to predict the response variable on test data.

Predicted values for the test data

```
> set.seed(100)
> n <- nrow(cars)
> i.training <- sort(sample(n,round(n*0.8)))
> cars.training <- cars[i.training,]
> cars.test <- cars[-i.training,]
> pred1a.test <- predict(lm1, newdata=cars.test)
>
>
> lm1.training <- lm(dist ~ speed, data=cars.training)
> pred1b.test <- predict(lm1.training, newdata=cars.test)
> data.frame(cars.test, pred1a.test, pred1b.test)
  speed dist pred1a.test pred1b.test
1     4    2  -1.849460  -5.392776
4     7   22   9.947766   7.555787
8    10   26  21.744993  20.504349
20    14   26  37.474628  37.769100
26    15   54  41.407036  42.085287
31    17   50  49.271854  50.717663
37    19   46  57.136672  59.350038
39    20   32  61.069080  63.666225
40    20   48  61.069080  63.666225
42    20   56  61.069080  63.666225
```


Predicted values for two test dataset and the observed data



contents

- Fitting polynomial models
 - Fitting a polynomial of degree 0
 - Fitting a polynomial of degree 1
 - Numerical results
 - Some diagnostic plots
 - The predictive performance of the model
 - **Confidence interval and prediction interval**
 - Fitting a polynomial of degree 2
 - Fitting a polynomial without intercept
 - Using orthogonal polynomials
- Model comparison
 - t-test
 - Analysis-of-variance (anova)
 - Likelihood ratio test (LRT)
 - Information criteria
- Data transformation
 - log based transformations
 - Diagnostic plots
 - Confidence interval and prediction interval
 - Model comparison

Confidence intervals for regression line

```
> alpha <- 0.05
> df.new <- data.frame(speed=(6:23))
> conf.dist <- predict(lm1, newdata = df.new, interval="confidence",
                      level=1-alpha)
> head(conf.dist)
```

	fit	lwr	upr
1	6.015358	-2.973341	15.00406
2	9.947766	1.678977	18.21656
3	13.880175	6.307527	21.45282
4	17.812584	10.905120	24.72005
5	21.744993	15.461917	28.02807
6	25.677401	19.964525	31.39028

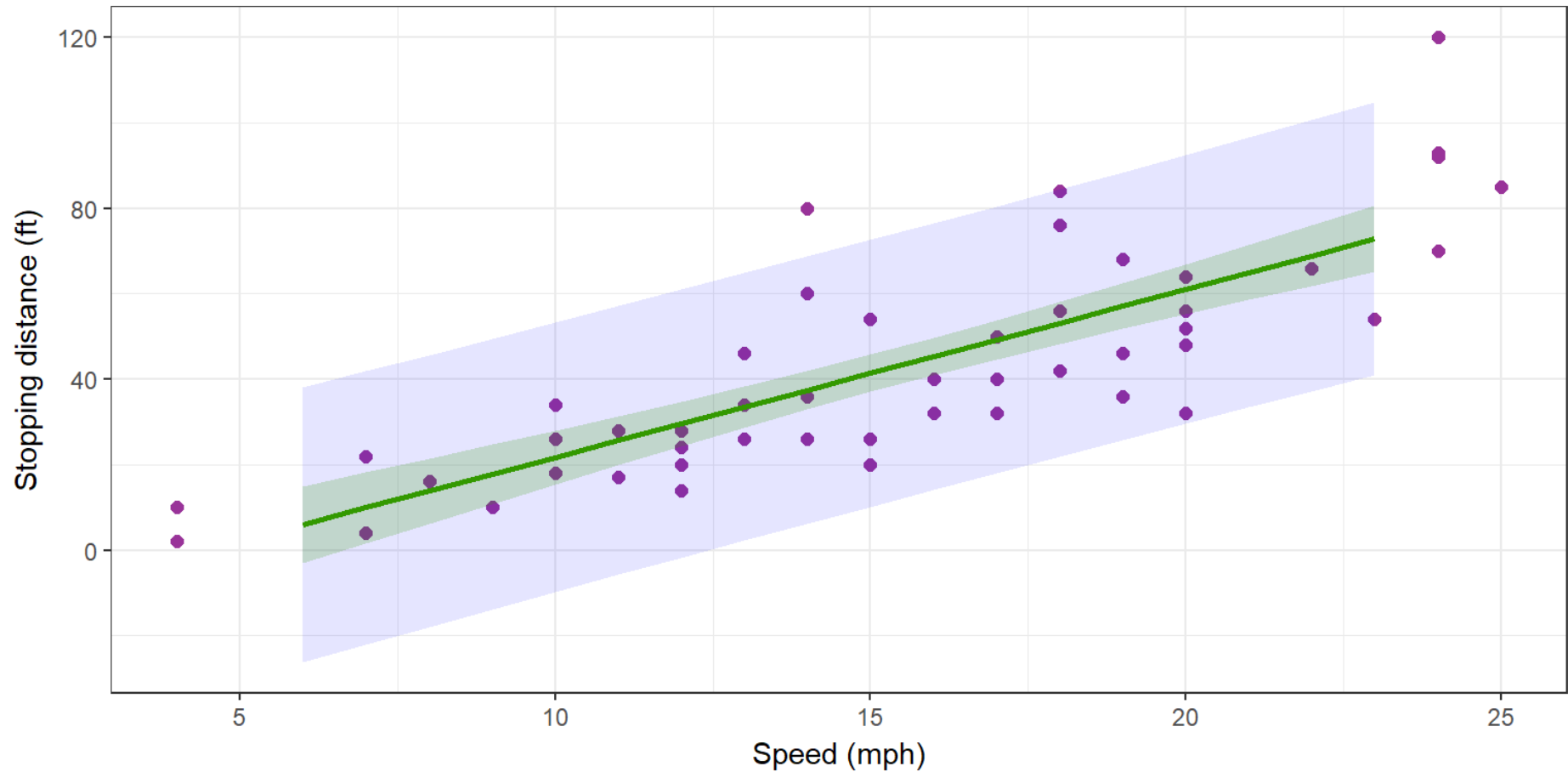
A prediction interval for a new measured distance

```
> pred.dist <- predict(lm1, newdata = df.new, interval="prediction",  
                        level=1-alpha)
```

```
> head(pred.dist)
```

	fit	lwr	upr
1	6.015358	-26.187314	38.21803
2	9.947766	-22.061423	41.95696
3	13.880175	-17.956287	45.71664
4	17.812584	-13.872245	49.49741
5	21.744993	-9.809601	53.29959
6	25.677401	-5.768620	57.12342

Graphical display for the two confidence intervals





contents

- Fitting polynomial models
 - Fitting a polynomial of degree 0
 - Fitting a polynomial of degree 1
 - Numerical results
 - Some diagnostic plots
 - The predictive performance of the model
 - Confidence interval and prediction interval
 - **Fitting a polynomial of degree 2**
 - Fitting a polynomial without intercept
 - Using orthogonal polynomials
- Model comparison
 - t-test
 - Analysis-of-variance (anova)
 - Likelihood ratio test (LRT)
 - Information criteria
- Data transformation
 - log based transformations
 - Diagnostic plots
 - Confidence interval and prediction interval
 - Model comparison

Polynomial of degree 2

```
> lm2 <- lm(dist ~ speed + I(speed^2))
```

```
> summary(lm2)
```

Call:

```
lm(formula = dist ~ speed + I(speed^2))
```

Residuals:

Min	1Q	Median	3Q	Max
-28.720	-9.184	-3.188	4.628	45.152

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.47014	14.81716	0.167	0.868
speed	0.91329	2.03422	0.449	0.656
I(speed^2)	0.09996	0.06597	1.515	0.136

Residual standard error: 15.18 on 47 degrees of freedom

Multiple R-squared: 0.6673, Adjusted R-squared: 0.6532

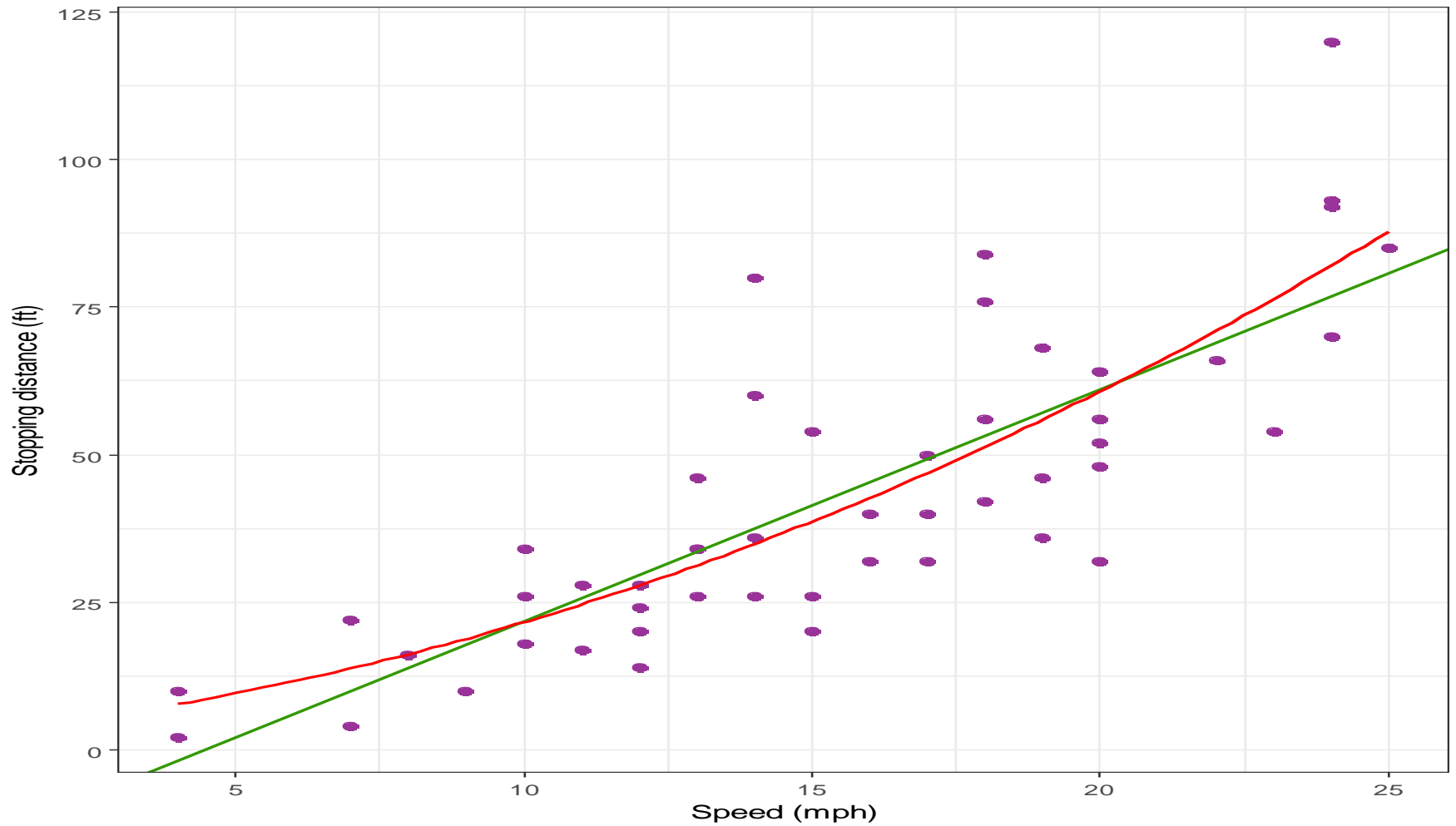
F-statistic: 47.14 on 2 and 47 DF, p-value: 5.852e-12

Covariance matrix of the vector of estimates

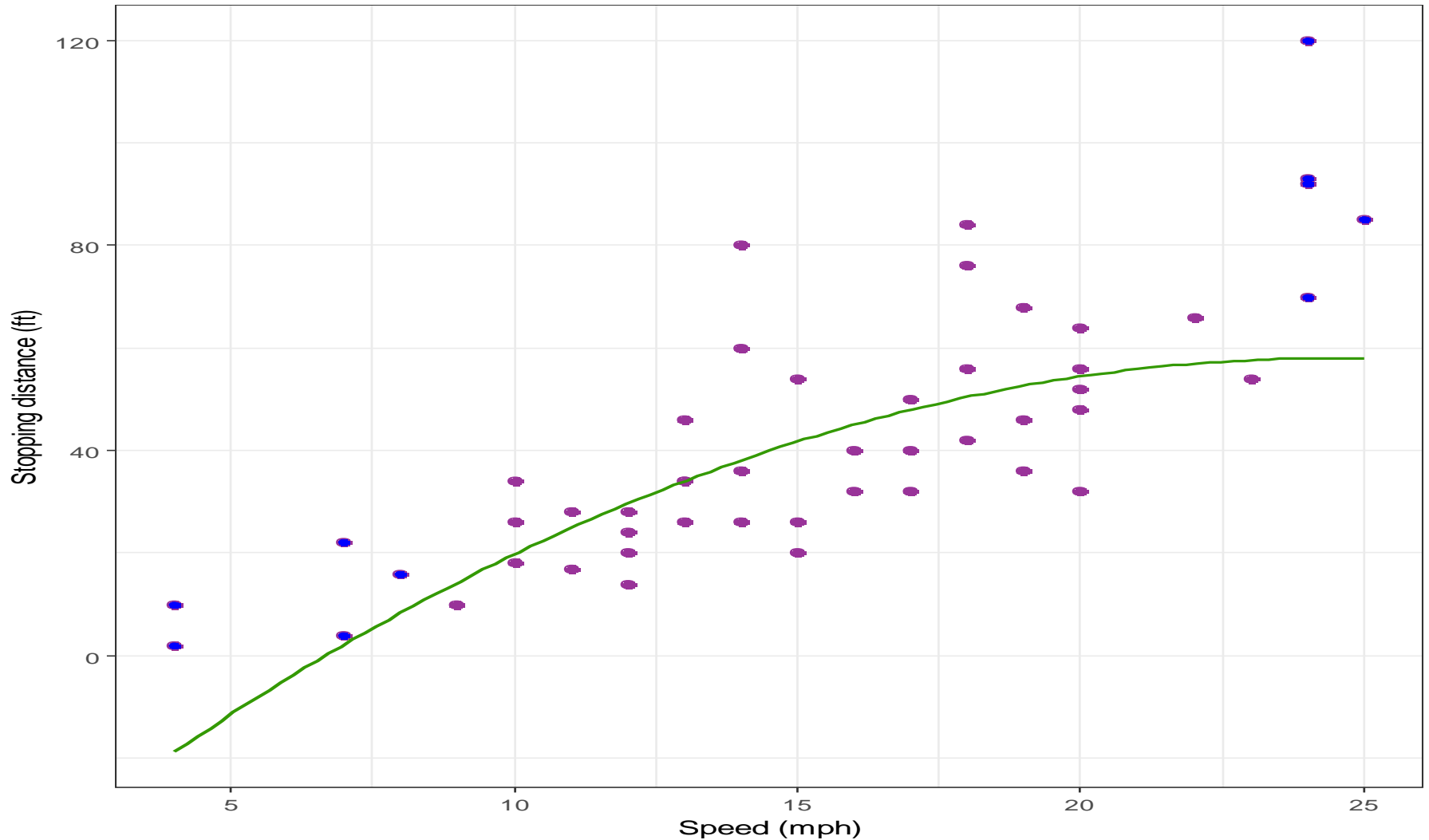
```
> X <- model.matrix(lm2)
> S <- solve(t(X)%*%X)
> d <- sqrt(diag(S))
> R <- S/(d%*%t(d))
> R
```

	(Intercept)	speed	I(speed^2)
(Intercept)	1.0000000	-0.9605503	0.8929849
speed	-0.9605503	1.0000000	-0.9794765
I(speed^2)	0.8929849	-0.9794765	1.0000000

Data and estimated model



Predicted model with the 80% most central data points

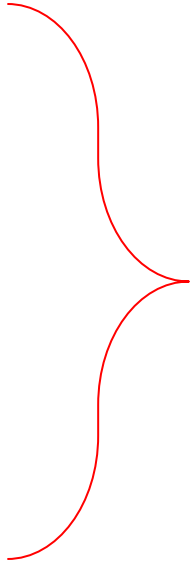


contents

- Fitting polynomial models
 - Fitting a polynomial of degree 0
 - Fitting a polynomial of degree 1
 - Numerical results
 - Some diagnostic plots
 - The predictive performance of the model
 - Confidence interval and prediction interval
 - Fitting a polynomial of degree 2
 - **Fitting a polynomial without intercept**
 - Using orthogonal polynomials
- Model comparison
 - t-test
 - Analysis-of-variance (anova)
 - Likelihood ratio test (LRT)
 - Information criteria
- Data transformation
 - log based transformations
 - Diagnostic plots
 - Confidence interval and prediction interval
 - Model comparison

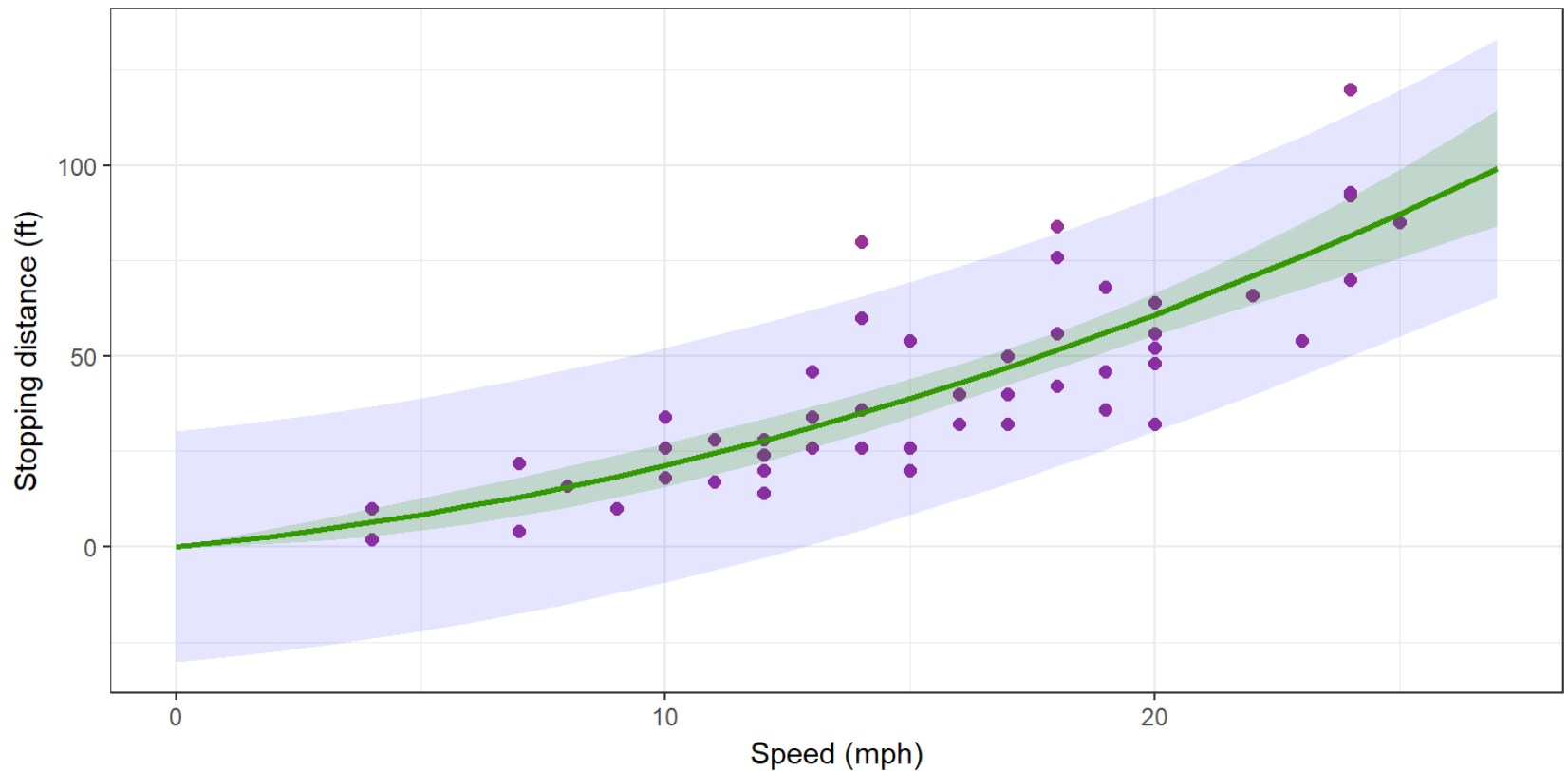
Fitting a polynomial without intercept

```
> lm2.noint <- lm(dist ~ -1 + speed + I(speed^2))  
  
> coef(lm2.noint)  
      speed I(speed^2)  
1.23902996 0.09013877  
  
> X <- model.matrix(lm2.noint)  
> head(X)  
  speed I(speed^2)  
1      4      16  
2      4      16  
3      7      49  
4      7      49  
5      8      64  
6      9      81
```



Design matrix

Data, predicted model, prediction intervals and confidence intervals





contents

- Fitting polynomial models
 - Fitting a polynomial of degree 0
 - Fitting a polynomial of degree 1
 - Numerical results
 - Some diagnostic plots
 - The predictive performance of the model
 - Confidence interval and prediction interval
 - Fitting a polynomial of degree 2
 - Fitting a polynomial without intercept
 - **Using orthogonal polynomials**
- Model comparison
 - t-test
 - Analysis-of-variance (anova)
 - Likelihood ratio test (LRT)
 - Information criteria
- Data transformation
 - log based transformations
 - Diagnostic plots
 - Confidence interval and prediction interval
 - Model comparison

Orthogonal polynomials

```
> lm2.poly <- lm(dist ~ poly(speed, degree=2, raw=T))
```

```
> M <- model.matrix(lm2.poly)
```

```
> head(M)
```

	(Intercept)	poly(speed, degree = 2, raw = T)1	
1	1	4	Design matrix
2	1	4	
3	1	7	
4	1	7	
5	1	8	
6	1	9	
	poly(speed, degree = 2, raw = T)2		Design matrix
1	16		
2	16		
3	49		
4	49		
5	64		
6	81		

contents

- Fitting polynomial models
 - Fitting a polynomial of degree 0
 - Fitting a polynomial of degree 1
 - Numerical results
 - Some diagnostic plots
 - The predictive performance of the model
 - Confidence interval and prediction interval
 - Fitting a polynomial of degree 2
 - Fitting a polynomial without intercept
 - Using orthogonal polynomials
- Model comparison
 - t-test
 - Analysis-of-variance (anova)
 - Likelihood ratio test (LRT)
 - Information criteria
- Data transformation
 - log based transformations
 - Diagnostic plots
 - Confidence interval and prediction interval
 - Model comparison

T-test

```
> summary(lm1.ortho)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	42.9800	2.175002	19.76090	1.061332e-24
poly(speed, degree = 1)	145.5523	15.379587	9.46399	1.489836e-12



contents

- Fitting polynomial models
 - Fitting a polynomial of degree 0
 - Fitting a polynomial of degree 1
 - Numerical results
 - Some diagnostic plots
 - The predictive performance of the model
 - Confidence interval and prediction interval
 - Fitting a polynomial of degree 2
 - Fitting a polynomial without intercept
 - Using orthogonal polynomials
- Model comparison
 - t-test
 - **Analysis-of-variance (anova)**
 - Likelihood ratio test (LRT)
 - Information criteria
- Data transformation
 - log based transformations
 - Diagnostic plots
 - Confidence interval and prediction interval
 - Model comparison

anova(model1,model2)

```
> anova(lm0, lm1)
```

Analysis of Variance Table

Model 1: dist ~ 1 $Y_i = \alpha + \varepsilon_i$

Model 2: dist ~ speed $Y_i = \alpha + \beta \times X_i + \varepsilon_i$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	49	32539				
2	48	11354	1	21186	89.567	1.49e-12 ***

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

anova(model1,model2)

```
> anova(lm1, lm2)
```

Analysis of Variance Table

Model 1: dist ~ speed

Model 2: dist ~ speed + I(speed^2)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	48	11354				
2	47	10825	1	528.81	2.296	0.1364



contents

- Fitting polynomial models
 - Fitting a polynomial of degree 0
 - Fitting a polynomial of degree 1
 - Numerical results
 - Some diagnostic plots
 - The predictive performance of the model
 - Confidence interval and prediction interval
 - Fitting a polynomial of degree 2
 - Fitting a polynomial without intercept
 - Using orthogonal polynomials
- Model comparison
 - t-test
 - Analysis-of-variance (anova)
 - **Likelihood ratio test (LRT)**
 - Information criteria
- Data transformation
 - log based transformations
 - Diagnostic plots
 - Confidence interval and prediction interval
 - Model comparison

Likelihood ratio test (LRT)

```
> logLik(lm0)
'log Lik.' -232.9012 (df=2)
> logLik(lm1)
'log Lik.' -206.5784 (df=3)
> logLik(lm2)
'log Lik.' -205.386 (df=4)
>
```

```
> dl <- 2*as.numeric(logLik(lm1) - logLik(lm0))
> 1-pchisq(dl,1)
[1] 3.995693e-13
>
```

Model 0 versus model 1

```
> dl <- 2*as.numeric(logLik(lm2) - logLik(lm1))
> 1-pchisq(dl,1)
[1] 0.122521
```

Model 1 versus model 2

contents

- Fitting polynomial models
 - Fitting a polynomial of degree 0
 - Fitting a polynomial of degree 1
 - Numerical results
 - Some diagnostic plots
 - The predictive performance of the model
 - Confidence interval and prediction interval
 - Fitting a polynomial of degree 2
 - Fitting a polynomial without intercept
 - Using orthogonal polynomials
- Model comparison
 - t-test
 - Analysis-of-variance (anova)
 - Likelihood ratio test (LRT)
 - **Information criteria**
- Data transformation
 - log based transformations
 - Diagnostic plots
 - Confidence interval and prediction interval
 - Model comparison

Information criteria

```
> AIC(lm0,lm1,lm2)
```

	df	AIC
lm0	2	469.8024
lm1	3	419.1569
lm2	4	418.7721

```
> BIC(lm0,lm1,lm2)
```

	df	BIC
lm0	2	473.6265
lm1	3	424.8929
lm2	4	426.4202

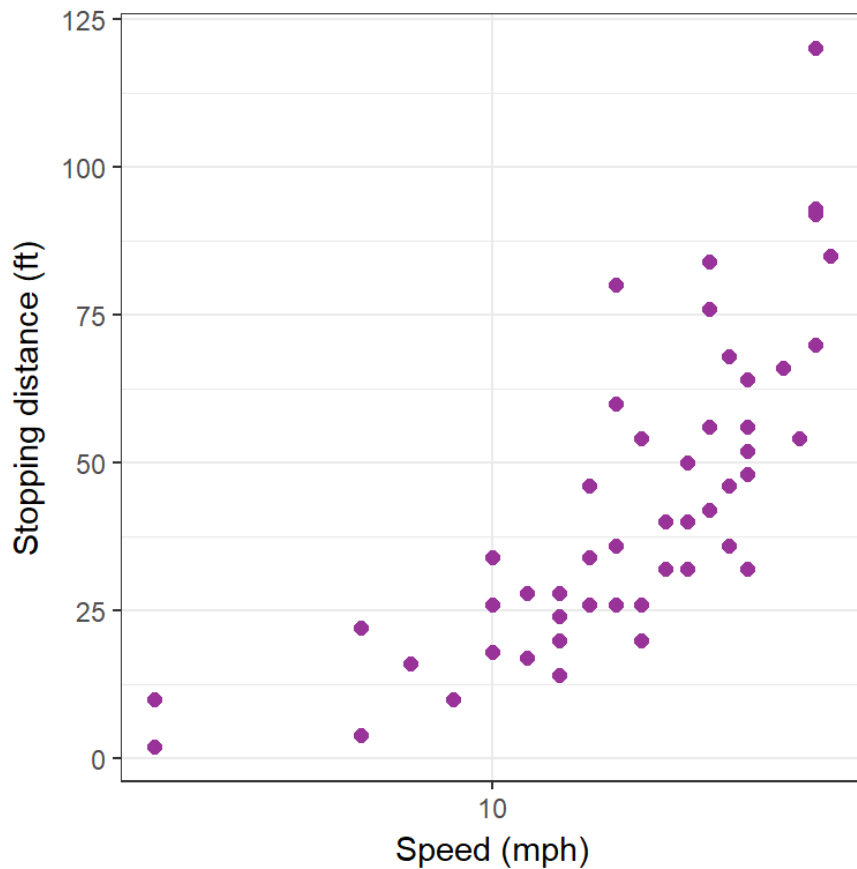


contents

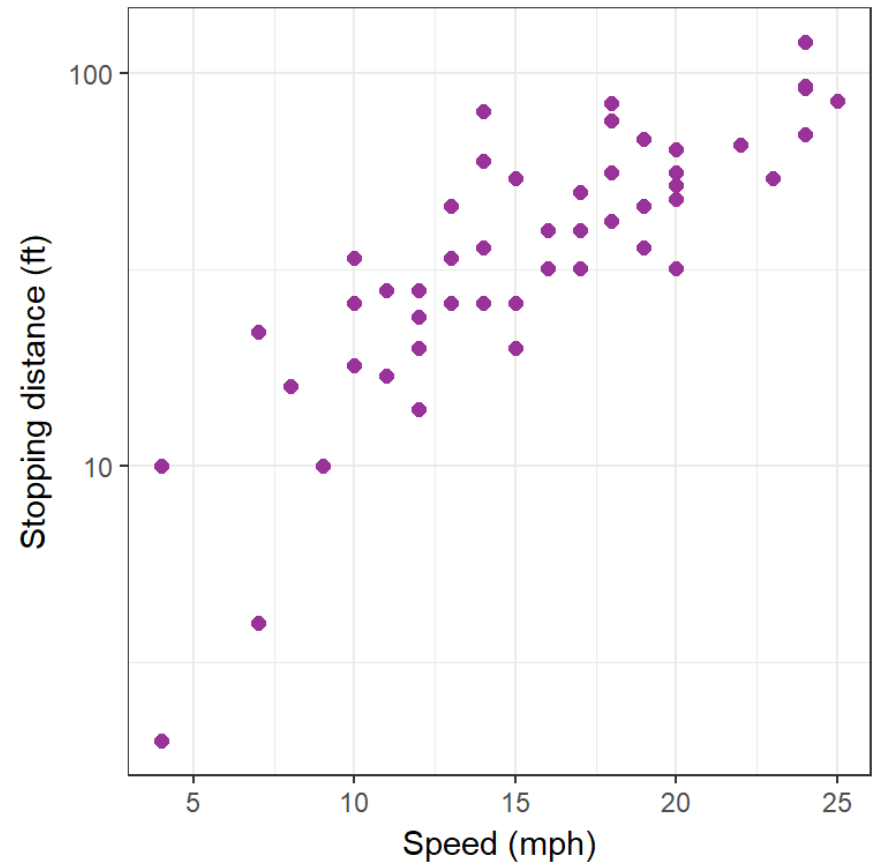
- Fitting polynomial models
 - Fitting a polynomial of degree 0
 - Fitting a polynomial of degree 1
 - Numerical results
 - Some diagnostic plots
 - The predictive performance of the model
 - Confidence interval and prediction interval
 - Fitting a polynomial of degree 2
 - Fitting a polynomial without intercept
 - Using orthogonal polynomials
- Model comparison
 - t-test
 - Analysis-of-variance (anova)
 - Likelihood ratio test (LRT)
 - Information criteria
- Data transformation
 - **log based transformations**
 - Diagnostic plots
 - Confidence interval and prediction interval
 - Model comparison

Logarithmic scale

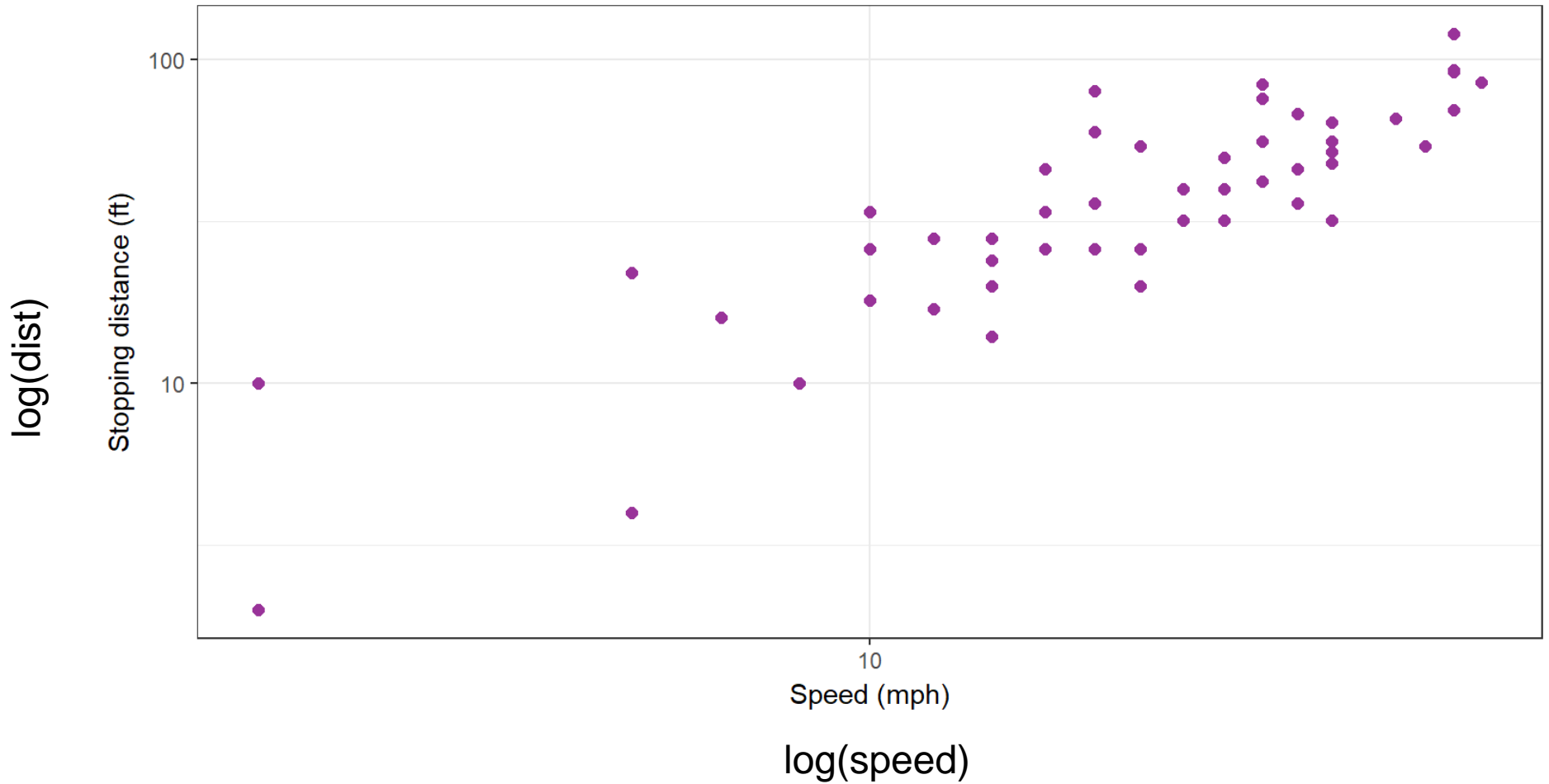
logarithmic scale for the x-axis



logarithmic scale for the y-axis



The data using log-log transformation.



Regression model for log(dist) and log(speed)

```
> lm1.log <- lm(log(dist) ~ log(speed))     $\log(Y_i) = \alpha + \beta \times \log(X_i) + \varepsilon_i$   
> summary(lm1.log)
```

Call:

```
lm(formula = log(dist) ~ log(speed))
```

Residuals:

Min	1Q	Median	3Q	Max
-1.00215	-0.24578	-0.02898	0.20717	0.88289

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-0.7297	0.3758	-1.941	0.0581	.
log(speed)	1.6024	0.1395	11.484	2.26e-15	***

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

Residual standard error: 0.4053 on 48 degrees of freedom

Multiple R-squared: 0.7331, Adjusted R-squared: 0.7276

F-statistic: 131.9 on 1 and 48 DF, p-value: 2.259e-15

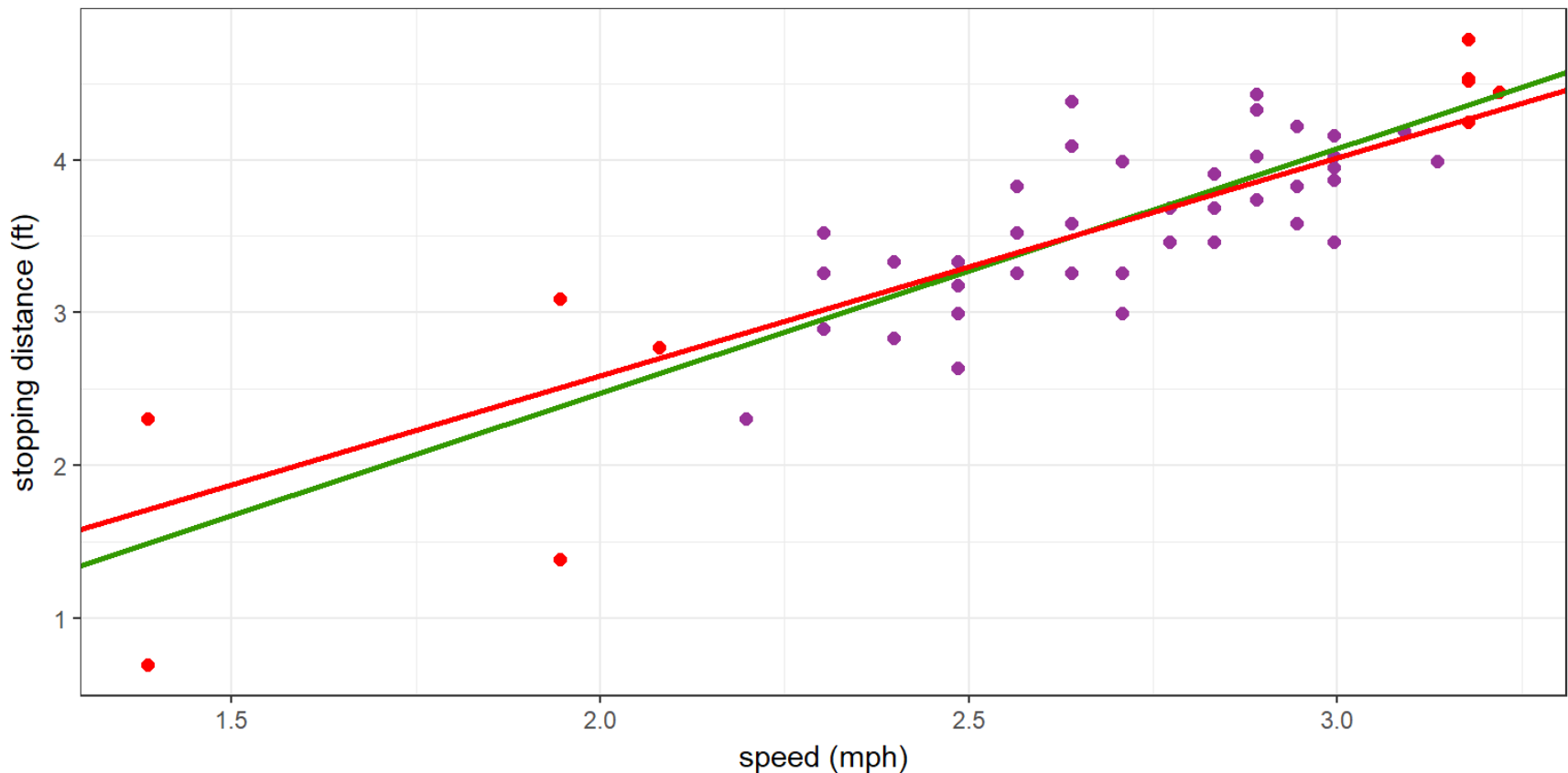


contents

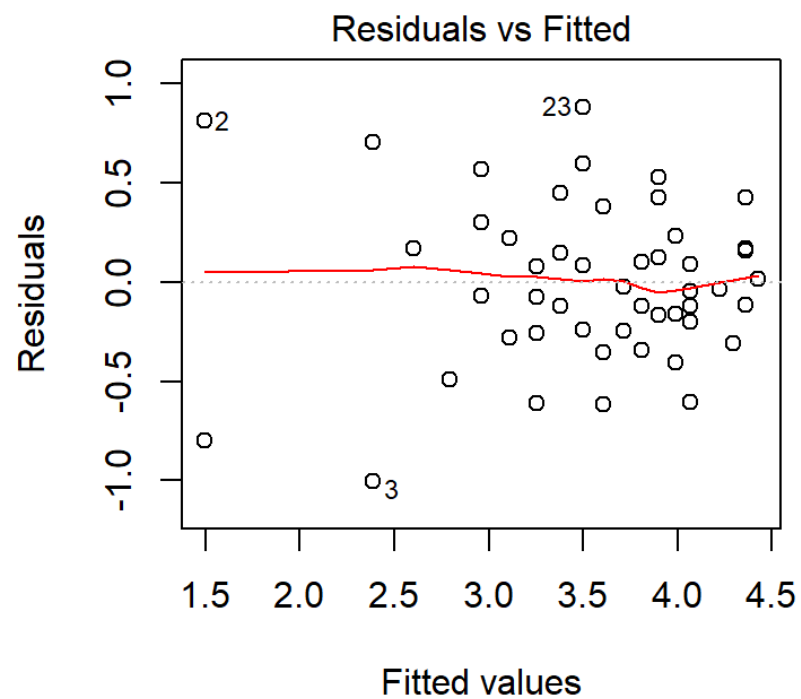
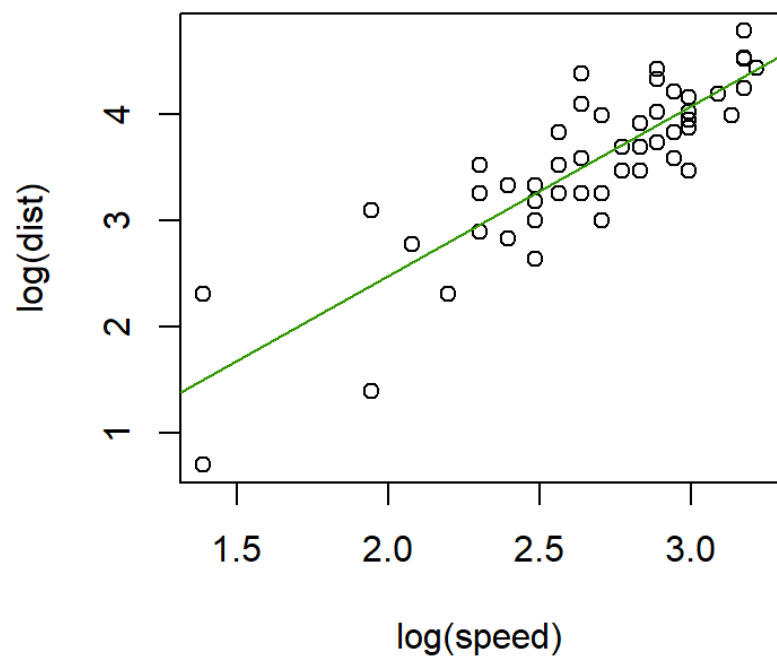
- Fitting polynomial models
 - Fitting a polynomial of degree 0
 - Fitting a polynomial of degree 1
 - Numerical results
 - Some diagnostic plots
 - The predictive performance of the model
 - Confidence interval and prediction interval
 - Fitting a polynomial of degree 2
 - Fitting a polynomial without intercept
 - Using orthogonal polynomials
- Model comparison
 - t-test
 - Analysis-of-variance (anova)
 - Likelihood ratio test (LRT)
 - Information criteria
- Data transformation
 - log based transformations
 - **Diagnostic plots**
 - Confidence interval and prediction interval
 - Model comparison

Data and estimated models

A model fitted using the complete data (green) and a model fitted when only the training sample is used (red).



Diagnostic plots



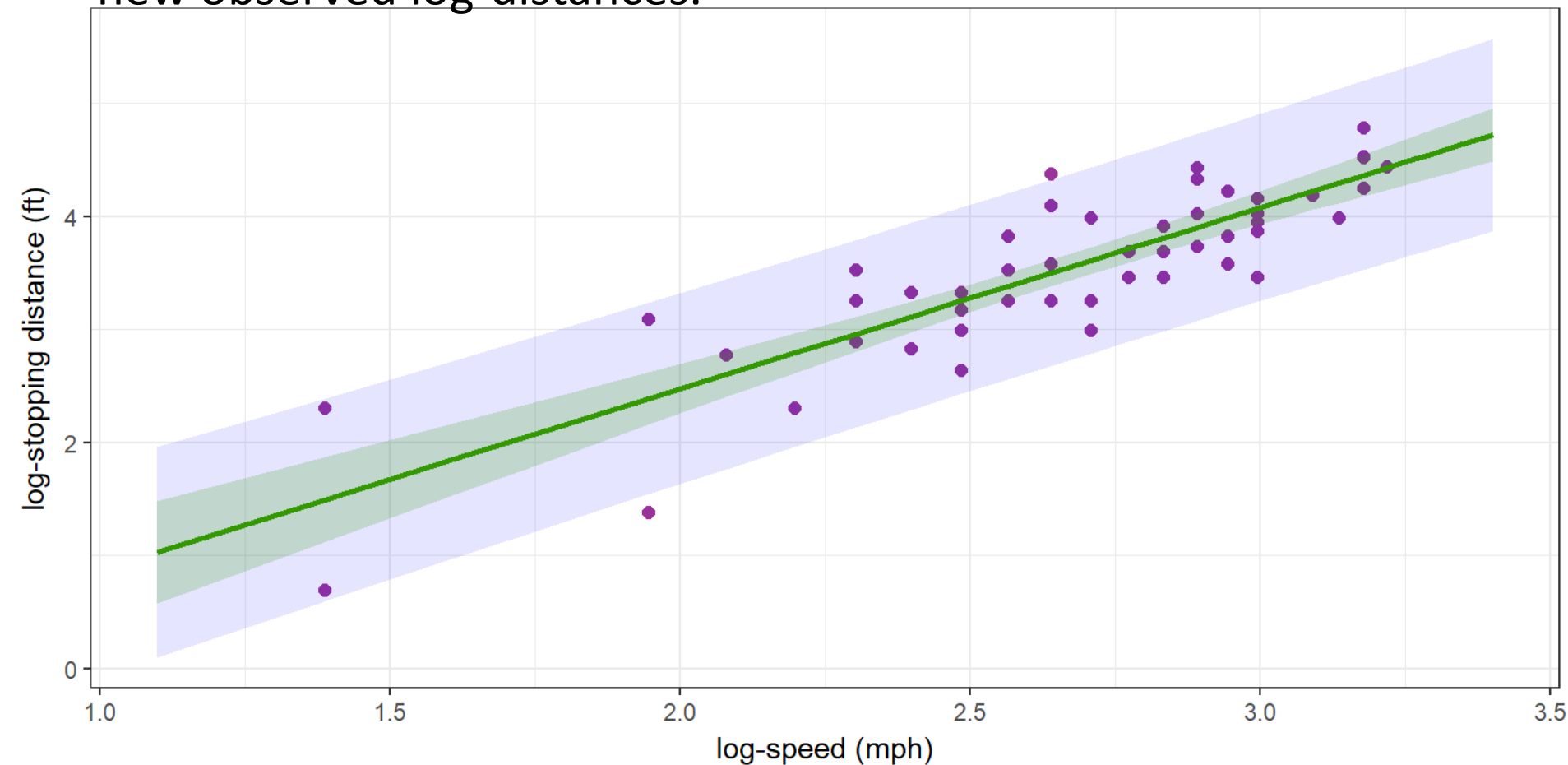


contents

- Fitting polynomial models
 - Fitting a polynomial of degree 0
 - Fitting a polynomial of degree 1
 - Numerical results
 - Some diagnostic plots
 - The predictive performance of the model
 - Confidence interval and prediction interval
 - Fitting a polynomial of degree 2
 - Fitting a polynomial without intercept
 - Using orthogonal polynomials
- Model comparison
 - t-test
 - Analysis-of-variance (anova)
 - Likelihood ratio test (LRT)
 - Information criteria
- Data transformation
 - log based transformations
 - Diagnostic plots
 - **Confidence interval and prediction interval**
 - Model comparison

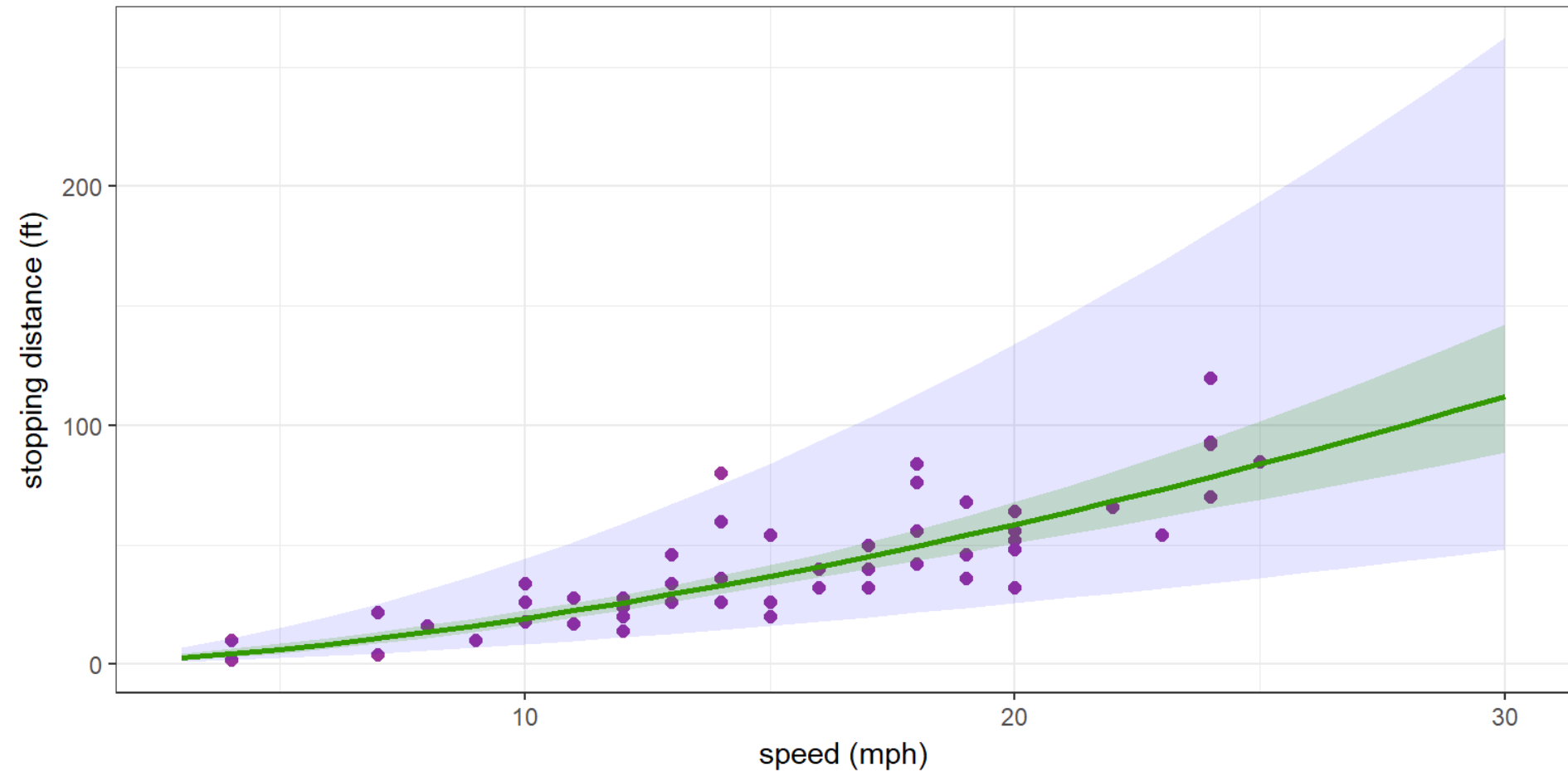
Confidence and prediction intervals

Confidence interval for the regression line (of the model $\text{lm}(\log(\text{dist}) \sim \log(\text{speed}))$) and the prediction interval for new observed log-distances.



Confidence and prediction intervals

Original scale.





contents

- Fitting polynomial models
 - Fitting a polynomial of degree 0
 - Fitting a polynomial of degree 1
 - Numerical results
 - Some diagnostic plots
 - The predictive performance of the model
 - Confidence interval and prediction interval
 - Fitting a polynomial of degree 2
 - Fitting a polynomial without intercept
 - Using orthogonal polynomials
- Model comparison
 - t-test
 - Analysis-of-variance (anova)
 - Likelihood ratio test (LRT)
 - Information criteria
- Data transformation
 - log based transformations
 - Diagnostic plots
 - Confidence interval and prediction interval
 - **Model comparison**

Model comparison

```
> logLik(lm1.log)
'log Lik.' -24.76592 (df=3)
> logLik(lm1.log) - sum(log(dist))
'log Lik.' -201.5613 (df=3)
> AIC(lm1.log) - 2*sum(log(1/dist))
[1] 409.1226
> BIC(lm1.log) - 2*sum(log(1/dist))
[1] 414.8587
```