11 Simple Linear Regression

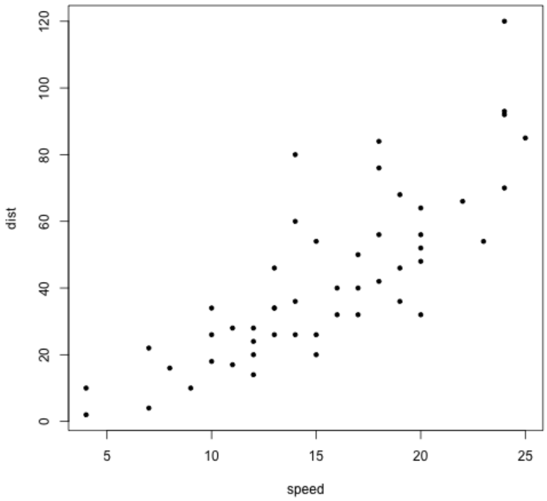
Chapter Preview

In Chapter 10, we considered ANOVA which examines the relationship between a qualitative variable (three or more populations/treatments) and a quantitative variable. Linear regression considers the relationship between two quantitative variables.

11.1: The Simple Linear Regression Model

Often, we wish to quantify and analyze the relationship between two or more quantitative variables.

Example: The following data give the speed of cars (mph) and the distances (feet) taken to stop. Note that the data were recorded in the 1920s.



Variables:

- X = independent variable / predictor variable / explanatory variable
- Y = dependent variable / response variable

Simple linear regression considers one predictor variable and one response variable.

Multiple linear regression considers more than one predictor variable and one response variable.

Goal: Use a simple random sample to build a model for the relationship between X and Y (both numerical values) for the entire population.

Given some (X,Y) data set, we want to find the line of best fit which we call the least squares regression line.

SLR Model

Estimated Regression Equation:

Population Model:

Assumptions and Notes about SLR:

1. Errors, ϵ_i, are independent (hence uncorrelated) and identically distributed normal random variables with mean 0 and variance σ^2. One goal of SLR is to estimate σ^2.

2. For a given value of X, the mean value of Y given by the line of best fit is: $E[Y|X] = \beta_0 + \beta_1 x$.

3. Since the data we observe does not fall exactly on the same line, we estimate σ by quantifying how much spread or variability the data has around the line of best fit.

Correlation

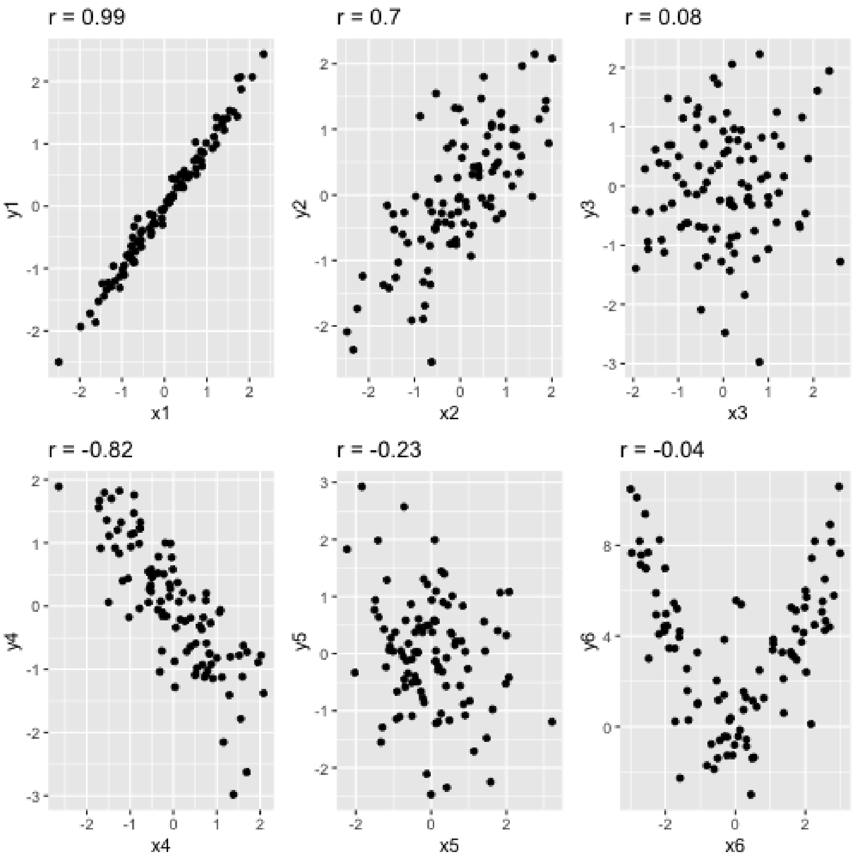
Recall: *Correlation* is a measure of linear relationship between two quantitative variables. The population correlation coefficient is represented by ρ and the sample correlation coefficient is represented by r.

Correlation is related to covariance, $\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$, but it is independent of units and is bounded by ± 1, i.e., $-1 \leq \rho \leq 1$.

If two variables are uncorrelated, then they are not necessarily independent. This could be the case if the variables have a nonlinear relationship.

$$\boxed{r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}, \text{where } S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}$$

Example: Some examples of some scatterplots and their associated sample correlation coefficients.



Note: Correlation does not imply causation!

11.2: Estimating Model Parameters

Criterion for determining the line of best fit

We want to minimize the sum of squared errors (SSE), that is, we want to find b_0 and b_1 that minimize $SSE = \sum_{i=1}^{n}[y_i - (b_0 + b_1 x)]^2$.

Illustration:

Normal Equations:

SLR estimates

$$\boxed{\text{Estimated Regression Equation: } \hat{y} = b_0 + b_1 x}$$

$$\boxed{b_0 = \bar{y} - b_1 \bar{x}}$$

$$\boxed{b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}}$$

$$\boxed{S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}$$

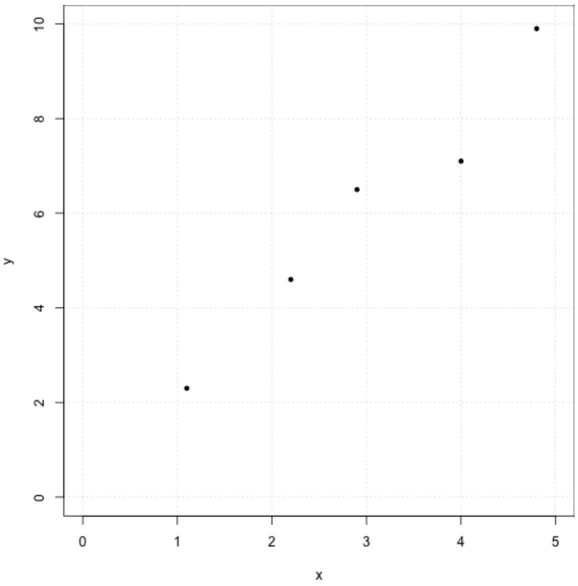
Interpreting b_1 and b_0

The estimated slope, b_1, can be interpreted as the estimated change in y for a unit increase in x.

The estimated intercept, b_0, can be interpreted as the estimated value of y when x is 0.

Example: Find and plot the estimated regression equation for the following data.

x_i	1.1	2.2	2.9	4.0	4.8
y_i	2.3	4.6	6.5	7.1	9.9



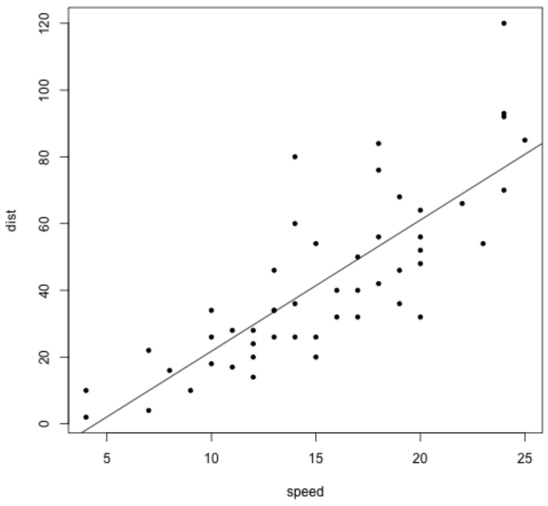
x_i	1.1	2.2	2.9	4.0	4.8	
$x_i - \bar{x}$						
$(x_i - \bar{x})^2$						
y_i	2.3	4.6	6.5	7.1	9.9	
$(y_i - \bar{y})$						
$(x_i - \bar{x})(y_i - \bar{y})$						

Example: Here's how to use R to do the calculations.

```
# enter the data
x = c(1.1,2.2,2.9,4,4.8)
y = c(2.3,4.6,6.5,7.1,9.9)

# run the linear regression and see the results
model = lm(y~x)
summary(model)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##        1        2        3        4        5
## -0.15435  0.04659  0.61082 -0.88824  0.38518
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3553     0.7488   0.474  0.66757
## x             1.9082     0.2289   8.335  0.00362 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6674 on 3 degrees of freedom
## Multiple R-squared:  0.9586,Adjusted R-squared:  0.9448
## F-statistic: 69.48 on 1 and 3 DF,  p-value: 0.003619
```

Example: Here is the car stopping distance example given at the beginning of the chapter. Note that we will load the “cars” dataset that is provided in R.

```
# load the data
data(cars)

# run the regression and summarize the results
model = lm(dist~speed,data=cars)
summary(model)

##
## 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
```



11.3: Assessing Model Fit and Inferences for the Slope Parameter, β_1

We can, of course, find an estimated regression equation for any set of pairs of data, however, we may be interested in analyzing if the linear model is appropriate.

There are multiple ways to assess model fit.

1. Coefficient of determination, R^2
2. Linear Regression Hypothesis Test for b_1
3. Model diagnostics (more on this later)

ANOVA for Regression

Often, we wish to analyze the quality of the estimated regression equation line using an *analysis of variance approach*. In this setup, the total variation of the dependent variable is partitioned into subcomponents.

$$\boxed{SST = \sum_{i=1}^{n}(y_i - \bar{y})^2} = \text{total sum of squares} = \text{“total variation”}$$

$\boxed{df_T = n - 1}$ (n = sample size)

$$\boxed{SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} = \text{regression sum of squares} = \text{“variation explained by the predictor x”}$$

$\boxed{df_R = p}$ (p = # of predictors, p=1 for SLR)

$$\boxed{SSE = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2} = \text{error sum of squares} = \text{“variation not explained by the predictor x”}$$

$\boxed{df_E = n - p - 1}$

Note: We can estimate σ^2 with $s^2 = MSE$.

Example: Here's how to get the ANOVA output for the car stopping example.

```
anova(model)

## Analysis of Variance Table
##
## Response: dist
##           Df Sum Sq Mean Sq F value   Pr(>F)
## speed      1  21186 21185.5  89.567 1.49e-12 ***
## Residuals 48  11354   236.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Coefficient of Determination

$$\boxed{R^2 = \frac{SSR}{SST}}$$

This is the proportion of the variation in Y that is explained by X.

Higher R^2 is typically preferred and suggests a better “fit” (*careful!*)

$R^2 = r_{xy}^2$, where r_{xy} is the correlation coefficient.

Linear Regression Hypothesis Test

Is there a linear relationship between the predictor(s) and the response? $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

$$\boxed{F_{test} = \frac{MSR}{MSE}} \; (df_1 = p,\, df_2 = n - p - 1)$$

For SLR, this is equivalent to a t-test.

If F_{test} is large, then there is evidence that there is a significant linear relationship between the predictor(s) and the response.

Example: Let’s look again at regression output for the cars dataset to see the results of the hypothesis test.

```
summary(model)

##
## 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
```

Inferences About the Slope Parameter β_1

Assuming we have met the assumptions of linear regression (i.e., IID normal errors), then we can create a confidence interval for β_1 or run a hypothesis test of linear relationship ($H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$.

1. $E[b_1] = \beta_1$ (b_1 is an unbiased estimator of β_1)

2. $Var(b_1) = \sigma^2_{b_1} = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$

3. $s_{b_1} = \frac{s}{\sqrt{S_{xx}}} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$

4. The estimator b_1 has a normal distribution.

Theorem: $t = \frac{b_1 - \beta_1}{s_{b_1}}$ has a t-distribution with $n - p - 1$ degrees of freedom.

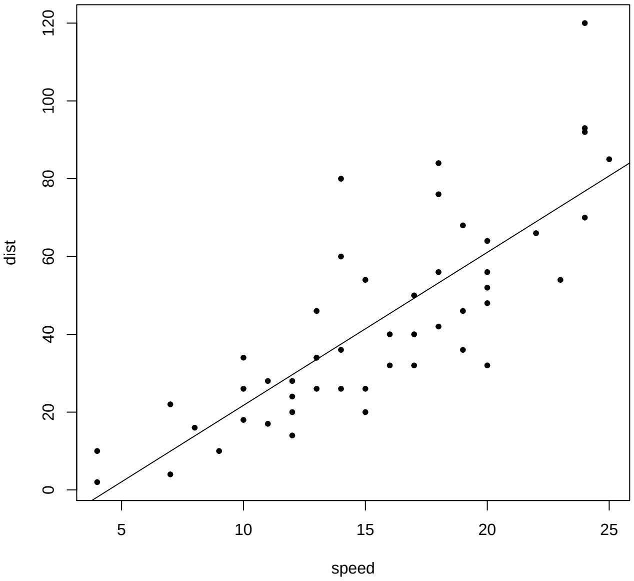
Example: Create a 95% confidence interval for β_1 using the “cars” dataset and determine if there is evidence of a linear relationship between *speed* and *stopping distance.*

11.4: The Prediction of Future Y Values

Given a specific value of X (denoted x_p), we can use the estimated regression equation to calculate a point estimate or fitted value of Y, denoted $\hat{y}_p$, by simply plugging x_p into the estimated regression equation. $\hat{y}_p = b_0 + b_1 x_p$

The residual of the i^{th} observation is $r_i = y_i - \hat{y}_i =$ actual y value $-$ predicted y value.

Example: Suppose we want to estimate the stopping distance for a car going 20.5mph. We can get this estimate by simply plugging this value into our estimated regression equation. Recall that the estimated regression equation is: $\hat{y} = -17.5791 + 3.9324x$.



In the above example, we computed a point estimate for the stopping distance of a car going 20.5mph. As with other topics we've explored so far in statistics, we prefer to have a measure of uncertainty attached with all point estimates.

For regression, this leads us to two kinds of intervals:

1. Confidence interval for mean response
2. Prediction interval for a new data value

Confidence intervals for mean response tell us how well we have determined the mean function, $E[Y|X]$, i.e., the line of best fit.

Prediction intervals for a new data value tell us how well we can estimate the dependent value for a given new predictor value.

These are fundamentally two different kinds of intervals.

Confidence Interval for Mean Response

A confidence interval for mean response creates an uncertainty estimate for the line of best fit, i.e., the mean function, $E[Y|X]$.

In other words, a confidence interval for mean response creates an interval of plausible values for the average response value for a given predictor value.

Theorem: Let x_p be the value of the predictor variable and let $\hat{y}_p = b_0 + b_1 x_p$ be the fitted value. Then:

1. $E[\hat{y}_p] = \beta_0 + \beta_1 x_p$
2. $Var(\hat{y}_p) = MSE \cdot \left(\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)$
3. $SE(\hat{y}_p) = \sqrt{MSE \cdot \left(\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}$
4. A confidence interval for the mean response is given by:

$$\hat{y}_p \pm t_{\alpha/2, n-2} \cdot \sqrt{MSE \cdot \left(\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}$$

The theorem above allows for us to create confidence intervals for $E[y_p]$, i.e., the mean response for a given x-value.

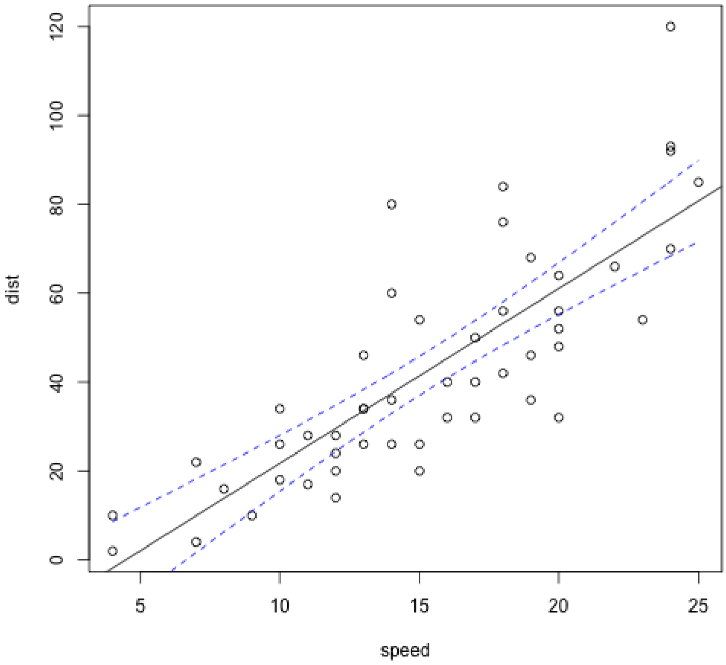
Example: Create a confidence interval for the mean response stopping distance for a car moving at 20.5 miles per hour. Note: $n = 50$, $\bar{x} = 15.4$, $\sum (x_i - \bar{x})^2 = 1370$, $\sqrt{MSE} = 15.38$, $t_{c.v.} = 2.01$.

```
# create confidence interval for mean response at x_p = 20.5
predict(model, newdata = data.frame(speed=20.5), interval = "confidence")

##        fit      lwr      upr
## 1 63.03528 56.92968 69.14089
```

We can also create a simultaneous confidence band for mean response for the entire regression line. Note that these band are parabolic.

Example: Create a 95% confidence band for the mean response of the cars dataset.

```
# plot the data and the line of best fit
plot(cars)
abline(model)

# create a grid of speed values and put into a data frame
speed.grid = seq(min(cars$speed), max(cars$speed), by = 0.1)
speed.df = data.frame(speed = speed.grid)

# calculate the confidence interval at each grid point and plot
conf_interval = predict(model, newdata = speed.df, interval = "confidence", level = 0.95)
lines(speed.grid, conf_interval[, 2], col = "blue", lty = 2)
lines(speed.grid, conf_interval[, 3], col = "blue", lty = 2)
```



Prediction Interval for a New Data Value

A prediction interval for a new data value creates an uncertainty estimate for a new data value. Note that this uncertainty is a combination of uncertainty from the line of best fit as well as the uncertainty from a single new observation.

In other words, a prediction interval for a new data value creates an interval of plausible values for the response value for a given new predictor value.

Theorem: Let x_p be the value of the predictor variable and let $\hat{y}_p = b_0 + b_1 x_p$ be the fitted value. Then:

1. $E[\hat{y}_p] = \beta_0 + \beta_1 x_p$
2. $Var(\hat{y}_p) = MSE \cdot \left(1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)$
3. $SE(\hat{y}_p) = \sqrt{MSE \cdot \left(1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$
4. A prediction interval for a new observation is given by:

$$\hat{y}_p \pm t_{\alpha/2,n-2} \cdot \sqrt{MSE \cdot \left(1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)}$$

The theorem above allows for us to create prediction intervals for y_p, i.e., the response for a new x-value..

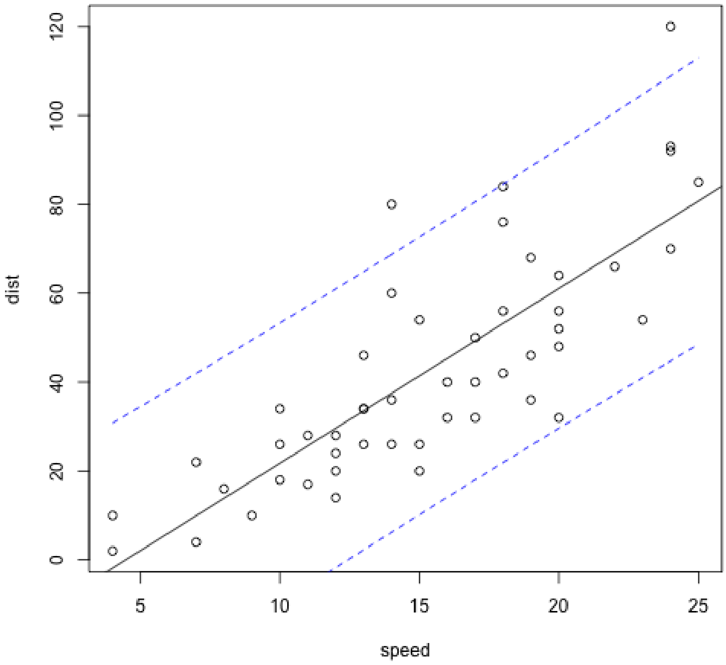
Example: Create a prediction interval for a car moving at 20.5 miles per hour. Note: $n = 50$, $\bar{x} = 15.4$, $\sum(x_i - \bar{x})^2 = 1370$, $\sqrt{MSE} = 15.38$, $t_{c.v.} = 2.01$.

```
# create confidence interval for mean response at x_p = 20.5
predict(model, newdata = data.frame(speed=20.5), interval = "prediction")

##        fit      lwr      upr
## 1 63.03528 31.51555 94.55502
```

We can also create a simultaneous prediction band for a new data value for the entire regression line. Note that these band are linear and parallel.

Example: Create a 95% prediction band for a new response of the cars dataset.

```
# plot the data and the line of best fit
plot(cars)
abline(model)

# create a grid of speed values and put into a data frame
speed.grid = seq(min(cars$speed), max(cars$speed), by = 0.1)
speed.df = data.frame(speed = speed.grid)

# calculate the confidence interval at each grid point and plot
conf_interval = predict(model, newdata = speed.df, interval = "prediction", level = 0.95)
lines(speed.grid, conf_interval[, 2], col = "blue", lty = 2)
lines(speed.grid, conf_interval[, 3], col = "blue", lty = 2)
```



11.5: Model Diagnostics

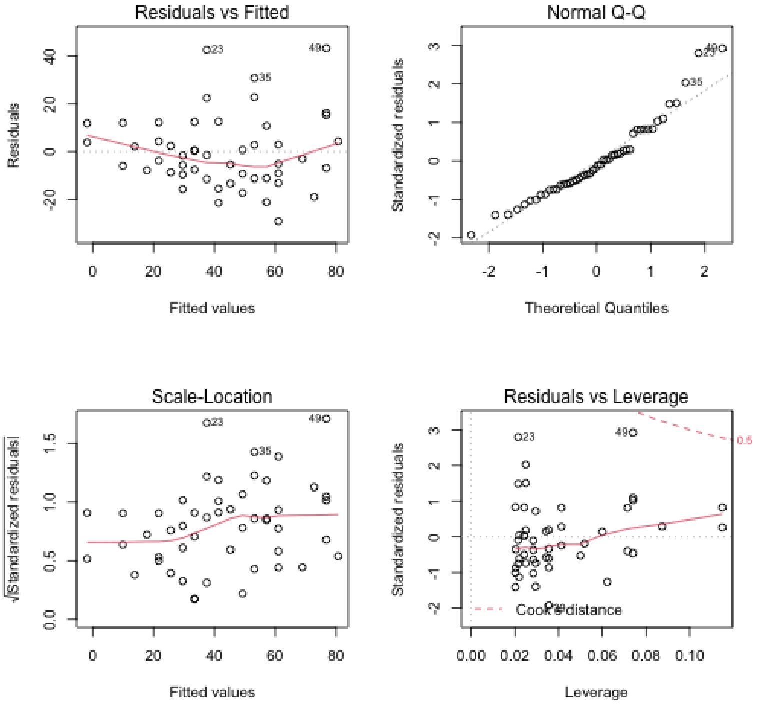
Recall: Simple linear regression makes a number of assumptions which should be checked to make sure that such a model is appropriate.

Assumptions:
1. **Linearity:** the mean response of the response variable is a linear function of the predictor variable. In multiple regression, the response variable is a linear combination of the predictor variables.
2. **Constant Variance:** different values of the predictor variables have the same variance in their errors. This is also called **homoscedasticity**.
3. **Normal Errors:** all errors are normally distributed (with constant variance).
4. **Independent Errors:** all errors are independent.

Examples of how to diagnose assumptions (1)-(3) will be given with toy datasets that either meet or violate the assumptions.

Here's how to obtain diagnostic plots in R.

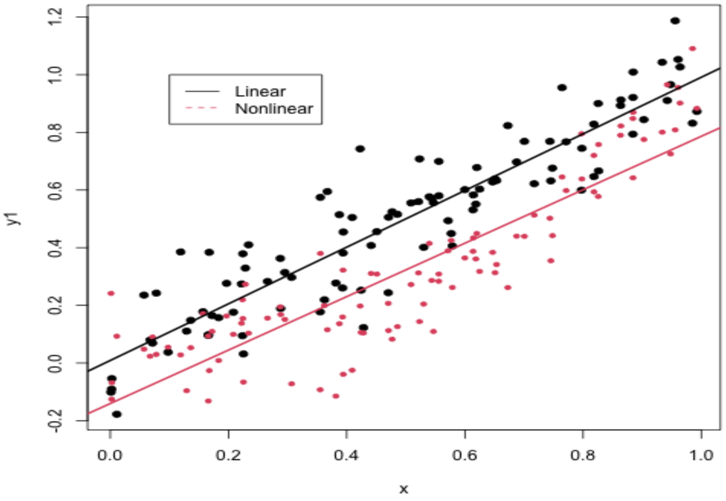
```
model = lm(dist~speed,data=cars)
plot(model)
```

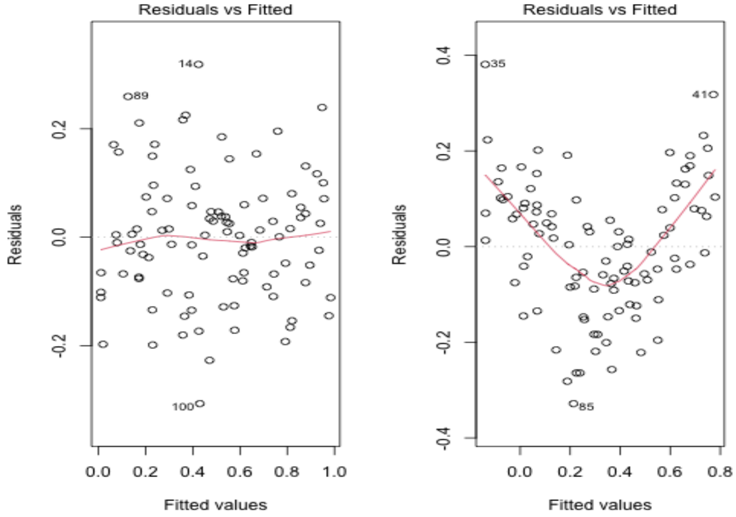


Linearity

Two toy datasets will be created, one linear and one nonlinear. Residual analysis will be used identify violations to the linearity assumption.

```r
set.seed(2020)
n=100
x = runif(n) # some random x-values
y1 = x + rnorm(n,sd=0.1) # linear
y2 = x^2 + rnorm(n,sd=0.1) # nonlinear
```

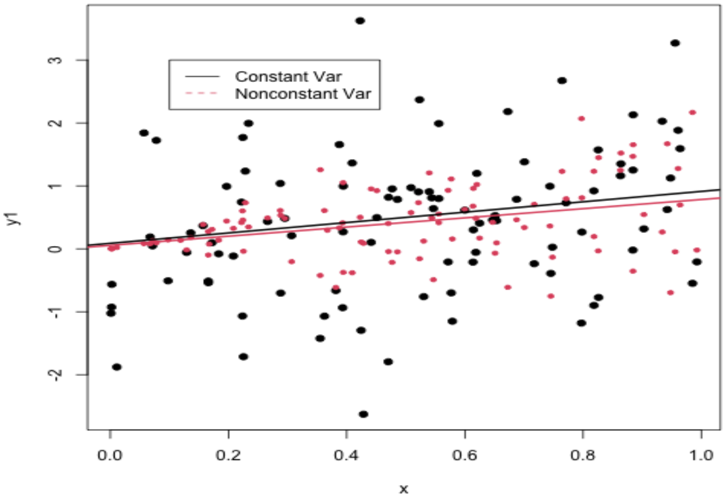


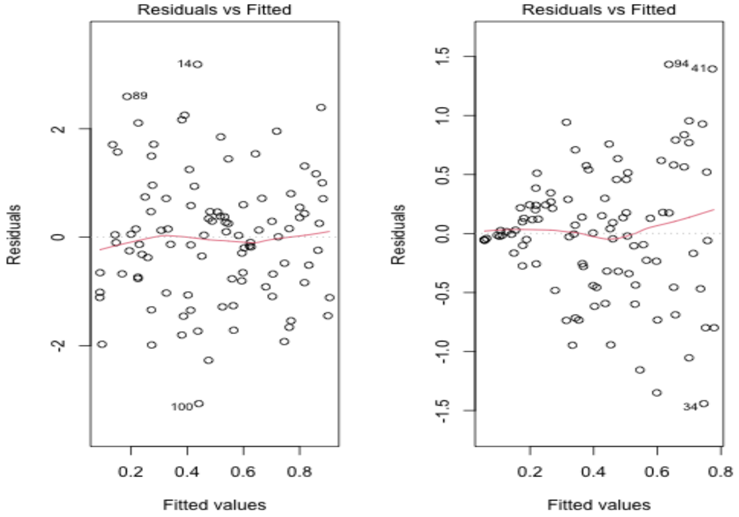


Constant Variance

Two toy datasets will be created, one with constant variance and one with nonconstant. Residual analysis will be used identify violations to the constant variance assumption.

```r
set.seed(2020)
n=100
x = runif(n) # some random x-values
y1 = x + rnorm(n,sd=1) # constant variance
y2 = x + rnorm(n,sd=x) # nonconstant variance
```

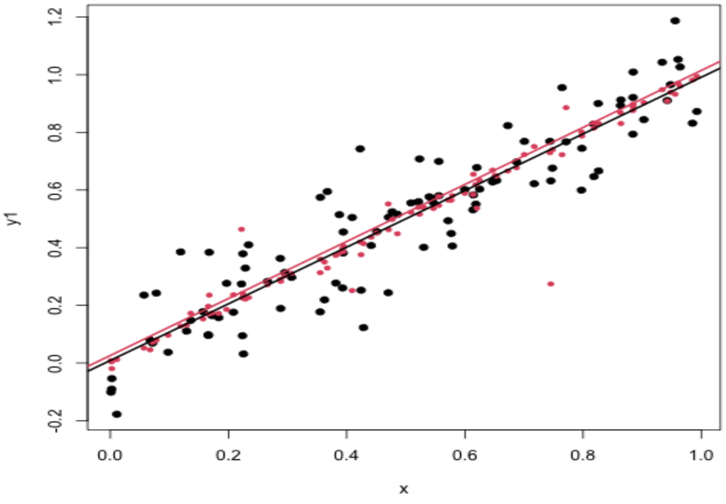


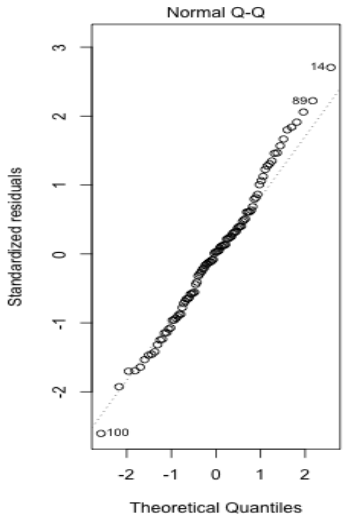


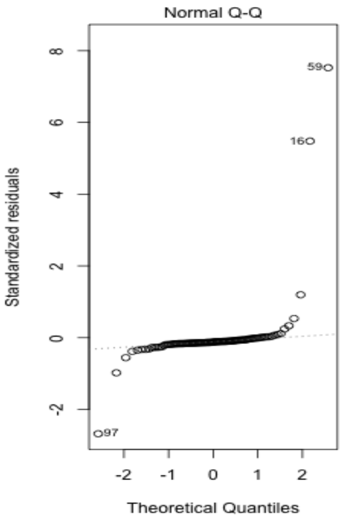
Normal Errors

Two toy datasets will be created, one with normal errors and one with Cauchy errors. A QQ-plot will be used identify violations to the normal errors assumption.

```r
set.seed(2020)
n=100
x = runif(n) # some random x-values
y1 = x + rnorm(n,sd=0.1) # normal errors
y2 = x + rcauchy(n,scale = 0.01) # laplace errors
```


y1
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.2
0.4
0.6
0.8
1.0
x


Normal Q-Q
Standardized residuals
Theoretical Quantiles


Normal Q-Q
Standardized residuals
Theoretical Quantiles