Lecture 4

Tahani Coolen-Maturi

Contents

1 Simple Linear Regression: Assumptions

Recall that the simple linear regression model for Y on x is

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where

Y : the dependent or response variable

x : the independent or predictor variable, assumed known

β_0, β_1 : the regression parameters, the intercept and slope of the regression line

ϵ : the random regression error around the line.

and the regression equation for a set of n data points is $\hat{y} = b_0 + b_1\ x$, where

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

and

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

where b_0 is called the **y-intercept** and b_1 is called the **slope**.

The **residual standard error** s_e can be defined as

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

s_e indicates how much, on average, the observed values of the response variable differ from the predicated values of the response variable.

1.1 Simple Linear Regression Assumptions (SLR)

We have a collection of n pairs of observations $\{(x_i, y_i)\}$, and the idea is to use them to estimate the unknown parameters β_0 and β_1.

$$\epsilon_i = Y_i - (\beta_0 + \beta_1\ x_i)\ ,\quad i = 1, 2, \ldots, n$$

We need to make the following key assumptions on the errors:

A. $E(\epsilon_i) = 0$ (errors have mean zero and do not depend on x)

B. $Var(\epsilon_i) = \sigma^2$ (errors have a constant variance, homoscedastic, and do not depend on x)

C. $\epsilon_1, \epsilon_2, \ldots \epsilon_n$ are independent.

D. ϵ_i are all i.i.d. $N(0,\ \sigma^2)$, meaning that the errors are independent and identically distributed as Normal with mean zero and constant variance σ^2.

The above assumptions, and conditioning on β_0 and β_1, imply:

a. Linearity: $E(Y_i|X_i) = \beta_0 + \beta_1\ x_i$

b. Homogenity or homoscedasticity: $Var(Y_i|X_i) = \sigma^2$

c. Independence: $Y_1, Y_2, \ldots, Y_n$ are all independent given X_i.

d. Normality: $Y_i|X_i \sim N(\beta_0 + \beta_1 x_i,\ \sigma^2)$

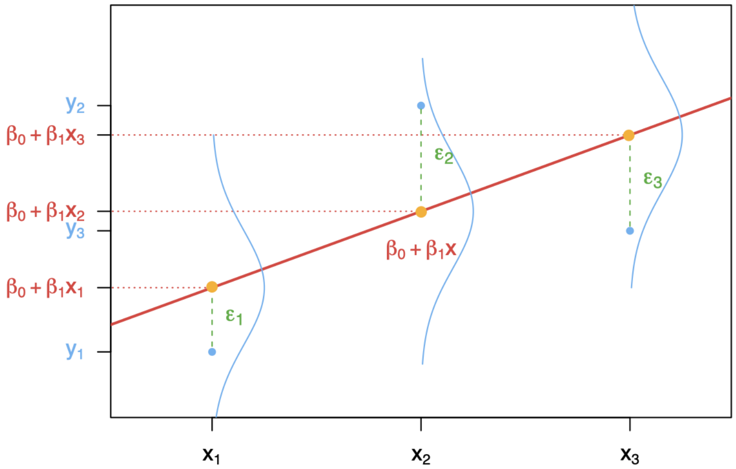
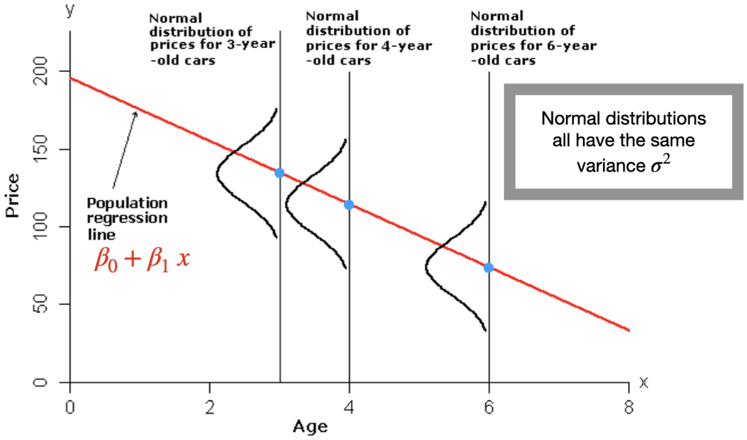


Image Credit: Jonathan Cumming

1.2 Used cars example

The table below displays data on Age (in years) and Price (in hundreds of dollars) for a sample of cars of a particular make and model.(Weiss, 2012)

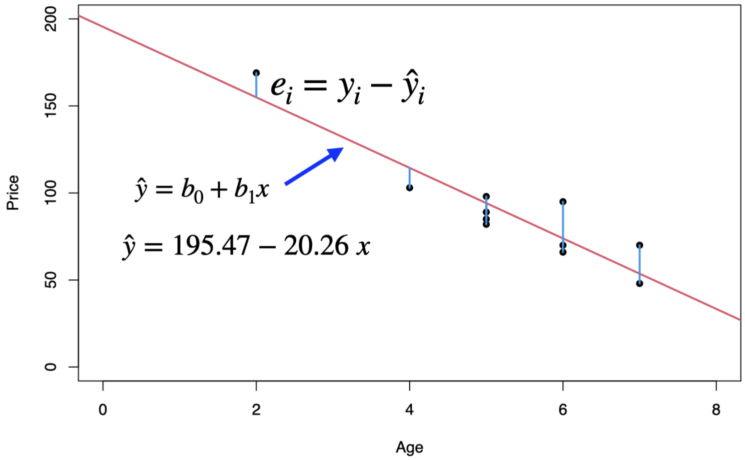
Price (y)	Age (x)
85	5
103	4
70	6
82	5
89	5
98	5
66	6
95	6
169	2
70	7
48	7


y
Normal distribution of prices for 3-year -old cars
Normal distribution of prices for 4-year -old cars
Normal distribution of prices for 6-year -old cars
Normal distributions all have the same variance σ^2
Population regression line
$\beta_0+\beta_1\,x$
Price
0
50
100
150
200
0
2
4
6
8
x
Age

We can see that for each age, the mean price of all cars of that age lies on the regression line $E(Y|x) = \beta_0+\beta_1x$. And, the prices of all cars of that age are assumed to be normally distributed with mean equal to $\beta_0+\beta_1x$

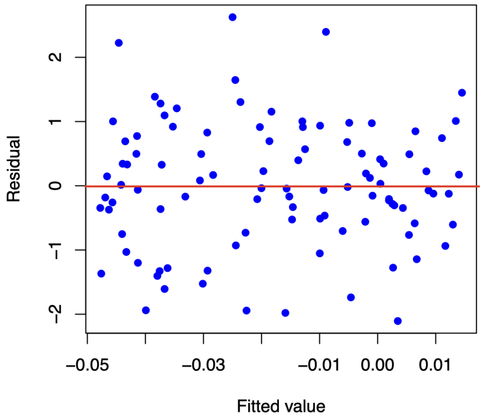
and variance σ^2. For example, the prices of all 4-year-old cars must be normally distributed with mean $\beta_0 + \beta_1(4)$ and variance σ^2.

We used the least square method to find the best fit for this data set, and residuals can be obtained as $e_i = y_i - \hat{y}_i = y_i - (195.47 - 20.26x_i)$.



1.3 Residual Analysis

The easiest way to check the simple linear regression assumptions is by constructing a scatterplot of the residuals ($e_i = y_i - \hat{y}_i$) against the fitted values ($\hat{y}_i$) or against x. If the model is a good fit, then the **residual plot** should show an even and random scatter of the residuals.



1.3.1 Linearity

The regression needs to be linear in the parameters.

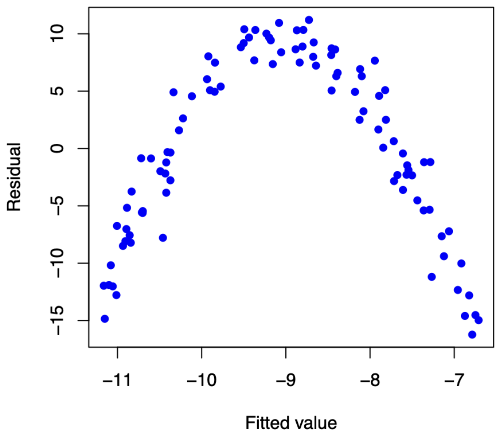
$$Y = \beta_0 + \beta_1\ x + \epsilon$$
$$E(Y_i|X_i) = \beta_0 + \beta_1\ x_i \equiv E(\epsilon_i|X_i) = E(\epsilon_i) = 0$$

✘ $\beta_0 + \beta_1^2 x_i$

✓ $\beta_0 + \beta_1 \log(x_i)$

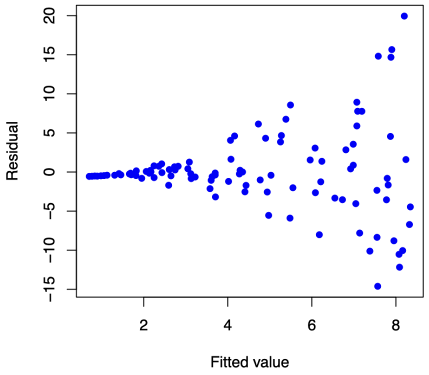
✓ $\beta_0 + \beta_1\ x_i^2$

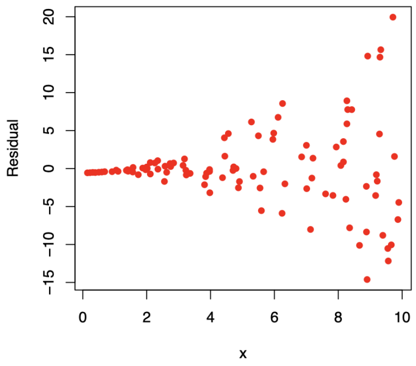
The residual plot below shows that a linear regression model is not appropriate for this data.



1.3.2 Constant error variance (homoscedasticity)

The plot shows the spread of the residuals is increasing as the fitted values (or x) increases, which indicates that we have Heteroskedasticity (non-constant variance). The standard errors are biased when heteroskedasticity is present, but the regression coefficients will still be unbiased.





How to detect?

- Residuals plot (fitted vs residuals)
- Goldfeld–Quandt test
- Breusch-Pagan test

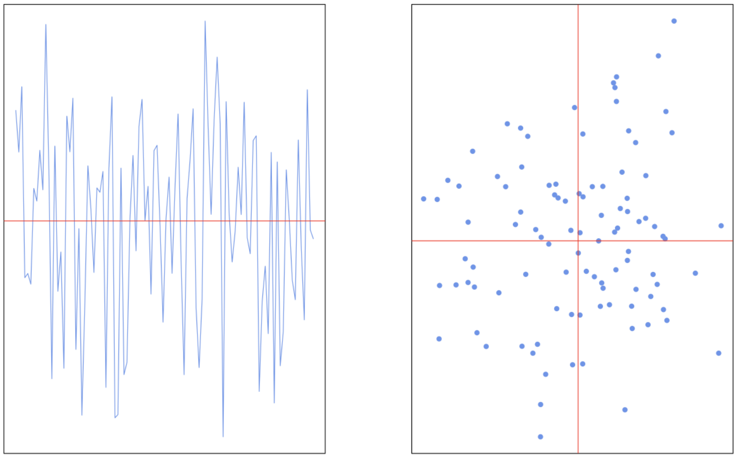
How to fix?

- White's standard errors
- Weighted least squares model
- Taking the log

1.3.3 Independent errors terms (no autocorrelation)

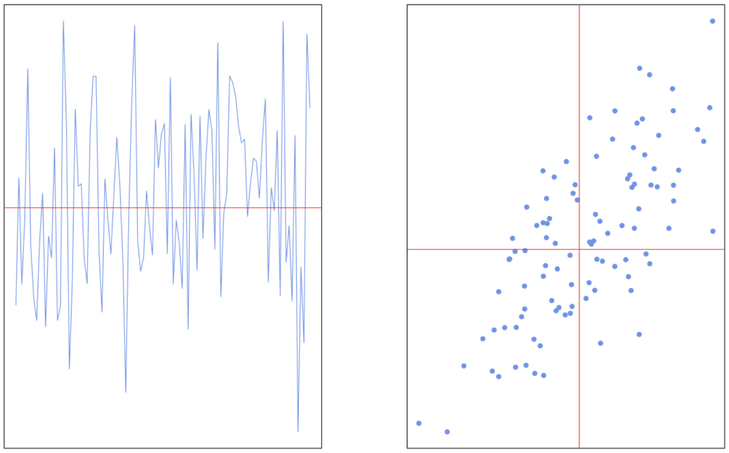
The problem of autocorrelation is most likely to occur in time series data, however, it can also occur in cross-sectional data, e.g. if the model is incorrectly specified. When autocorrelation is present, the regression coefficients will still be unbiased, however, the standard errors and test statitics are no longer valid.

An example of no autocorrelation



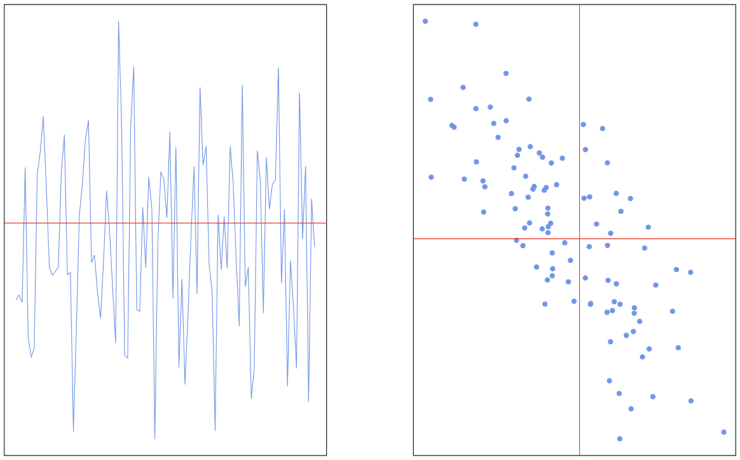
$time\ vs\ e_i$ $e_i\ vs\ e_{i-1}$

An example of positive autocorrelation



$time\ vs\ e_i$ $e_i\ vs\ e_{i-1}$

An example of negative autocorrelation



$time\ vs\ e_i$ $e_i\ vs\ e_{i-1}$

How to detect?

- Residuals plot
- Durbin-Watson test
- Breusch-Godfrey test

How to fix?

- Investigate omitted variables (e.g. trend, business cycle)
- Use advanced models (e.g. AR model)

1.3.4 Normality of the errors

We need the errors to be normally distributed. Normality is only required for the sampling distributions, hypothesis testing and confidence intervals.

How to detect?

- Histogram of residuals
- Q-Q plot of residuals
- Kolmogorov–Smirnov test
- Shapiro–Wilk test

How to fix?

- Change the functional form (e.g. taking the log)
- Larger sample if possible

1.4 Example: Infant mortality and GDP

Let us investigate the relationship between infant mortality and the wealth of a country. We will use data on 207 countries of the world gathered by the UN in 1998 (the ‘UN’ data set is available from the R package ‘car’). The data set contains two variables: the infant mortality rate in deaths per 1000 live births, and the GDP per capita in US dollars. There are some missing data values for some countries, so we will remove the missing data before we fit our model.

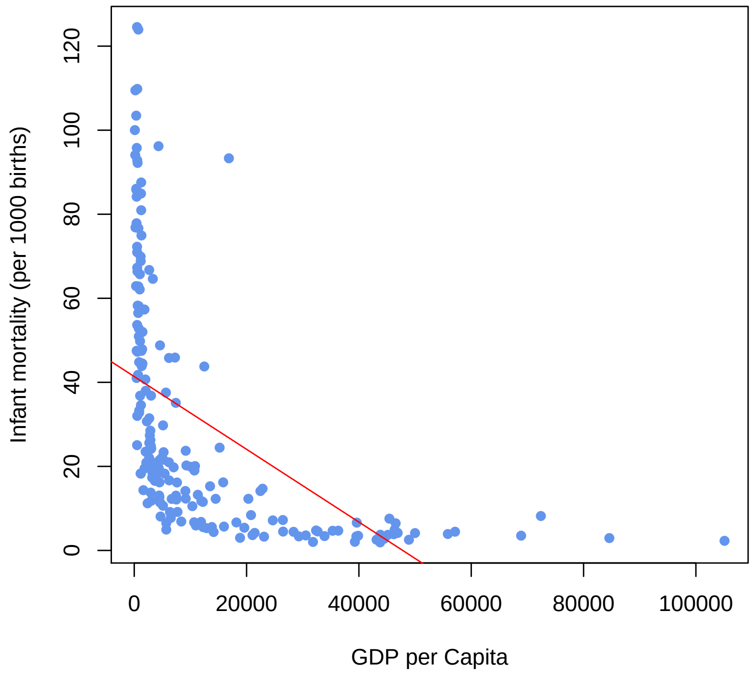
```r
# install.packages("carData")
library(carData)
data(UN)
options(scipen=999)
# Remove missing data
newUN<-na.omit(UN)
str(newUN)
```

```
## 'data.frame':    193 obs. of  7 variables:
##  $ region         : Factor w/ 8 levels "Africa","Asia",..: 2 4 1 1 5 2 3 8 4 2 ...
##  $ group          : Factor w/ 3 levels "oecd","other",..: 2 2 3 3 2 2 2 1 1 2 ...
##  $ fertility      : num  5.97 1.52 2.14 5.13 2.17 ...
##  $ ppgdp          : num  499 3677 4473 4322 9162 ...
##  $ lifeExpF       : num  49.5 80.4 75 53.2 79.9 ...
##  $ pctUrban       : num  23 53 67 59 93 64 47 89 68 52 ...
##  $ infantMortality: num  124.5 16.6 21.5 96.2 12.3 ...
##  - attr(*, "na.action")= 'omit' Named int [1:20] 4 6 21 35 38 54 67 75 77 78 ...
##   ..- attr(*, "names")= chr [1:20] "American Samoa" "Anguilla" "Bermuda" "Cayman Islands" ...
```

```r
fit<- lm(infantMortality ~ ppgdp, data=newUN)
summary(fit)
```

```
##
## Call:
## lm(formula = infantMortality ~ ppgdp, data = newUN)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31.48 -18.65  -8.59  10.86  83.59
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 41.3780016  2.2157454  18.675 < 0.0000000000000002 ***
## ppgdp       -0.0008656  0.0001041  -8.312   0.0000000000000173 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.13 on 191 degrees of freedom
## Multiple R-squared:  0.2656, Adjusted R-squared:  0.2618
## F-statistic: 69.08 on 1 and 191 DF,  p-value: 0.0000000000000173
```

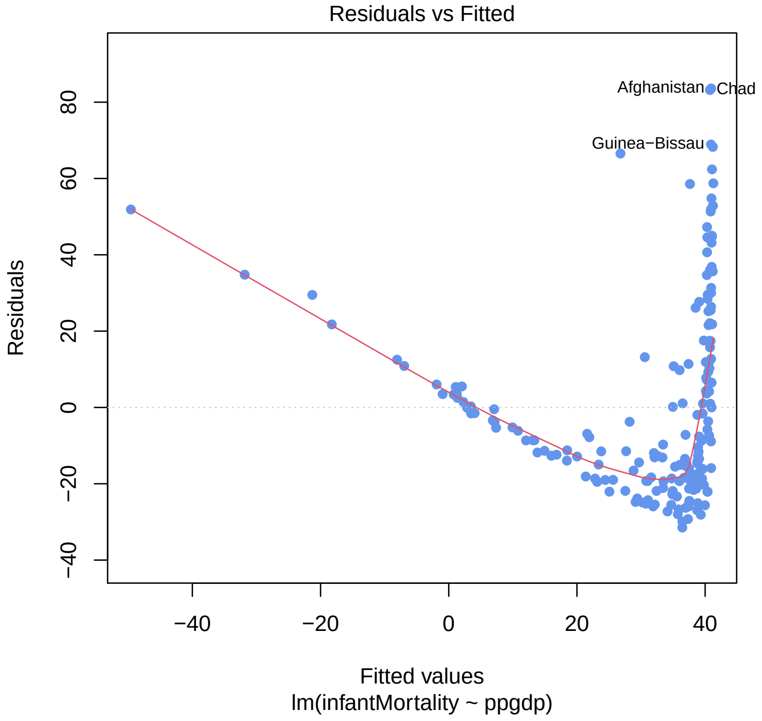
```r
plot(newUN$infantMortality ~ newUN$ppgdp, xlab="GDP per Capita", ylab="Infant mortality (per 1000 births
abline(fit,col="red")
```



We can see from the scatterplot that the relationship between the two variables is not linear. There is a concentration of data points at small values of GDP (many poor countries) and a concentration of data points at small values of infant mortality (many countries with very low mortality). This suggests a skewness to both variables which would not conform to the normality assumption. Indeed, the regression line (the red line) we construct is a poor fit and only has an R^2 of 0.266.

From the residual plot below we can see a clear evidence of structure to the residuals suggesting the linear relationship is a poor description of the data, and substantial changes in spread suggesting the assumption of homogeneous variance is not appropriate.

```r
# diagnostic plots
plot(fit,which=1,pch=16,col="cornflowerblue")
```

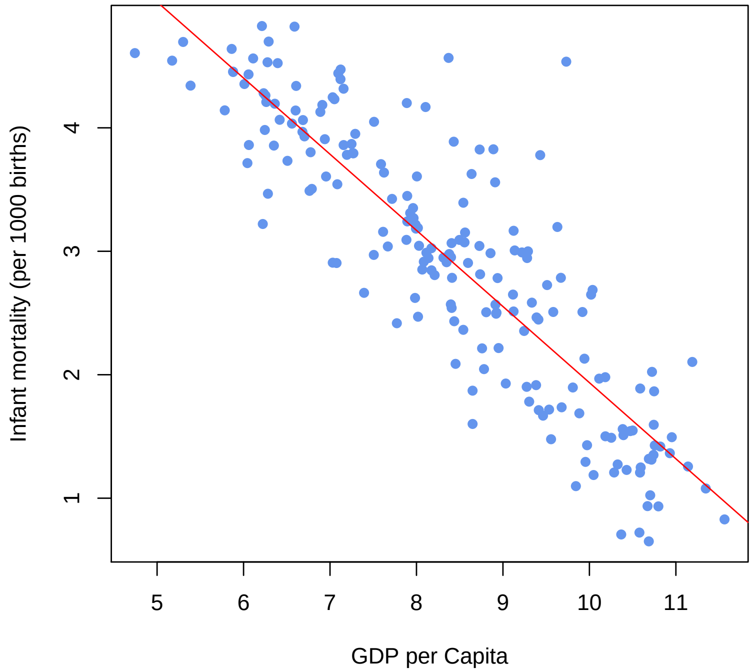


So we can apply a transformation to one or both variables, e.g. taking the log or adding a quadratic form. Notice that this will not affect (violet) the linearity assumption as the regression will still be linear in the parameters. So if we take the logs of both variables gives us the scatterplot of the transformed data set, below, which appears to show a more promising linear structure. The quality of the regression is now improved, with an R^2 value of 0.766, which is still a little weak due to the rather large spread in the data.

```
fit1<- lm(log(infantMortality) ~ log(ppgdp), data=newUN)
summary(fit1)
```

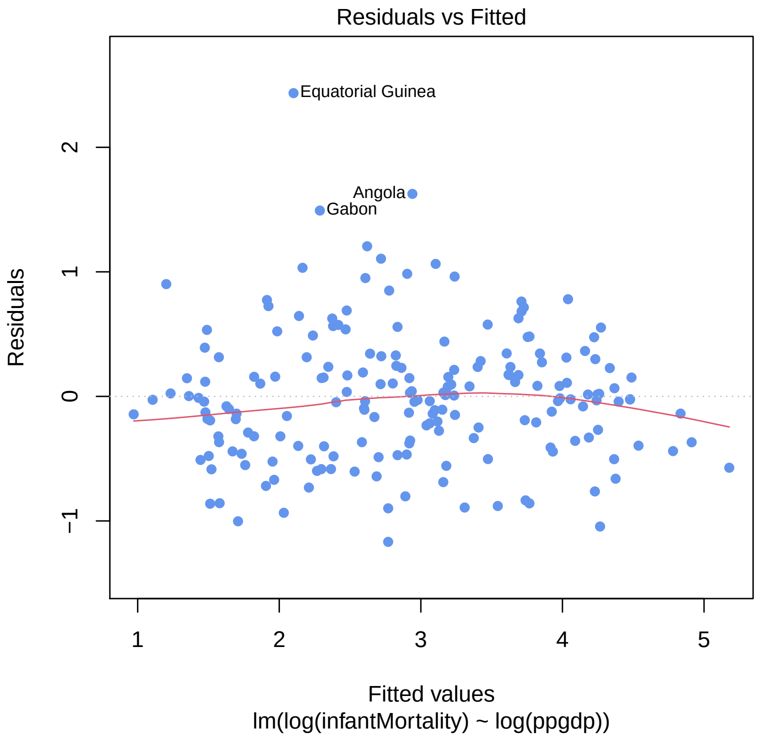
```
##
## Call:
## lm(formula = log(infantMortality) ~ log(ppgdp), data = newUN)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16789 -0.36738 -0.02351  0.24544  2.43503
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  8.10377    0.21087   38.43 <0.0000000000000002 ***
## log(ppgdp)  -0.61680    0.02465  -25.02 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5281 on 191 degrees of freedom
## Multiple R-squared:  0.7662, Adjusted R-squared:  0.765
## F-statistic: 625.9 on 1 and 191 DF,  p-value: < 0.00000000000000022
```

```
plot(log(newUN$infantMortality) ~ log(newUN$ppgdp), xlab="GDP per Capita", ylab="Infant mortality (per 
abline(fit1,col="red")
```



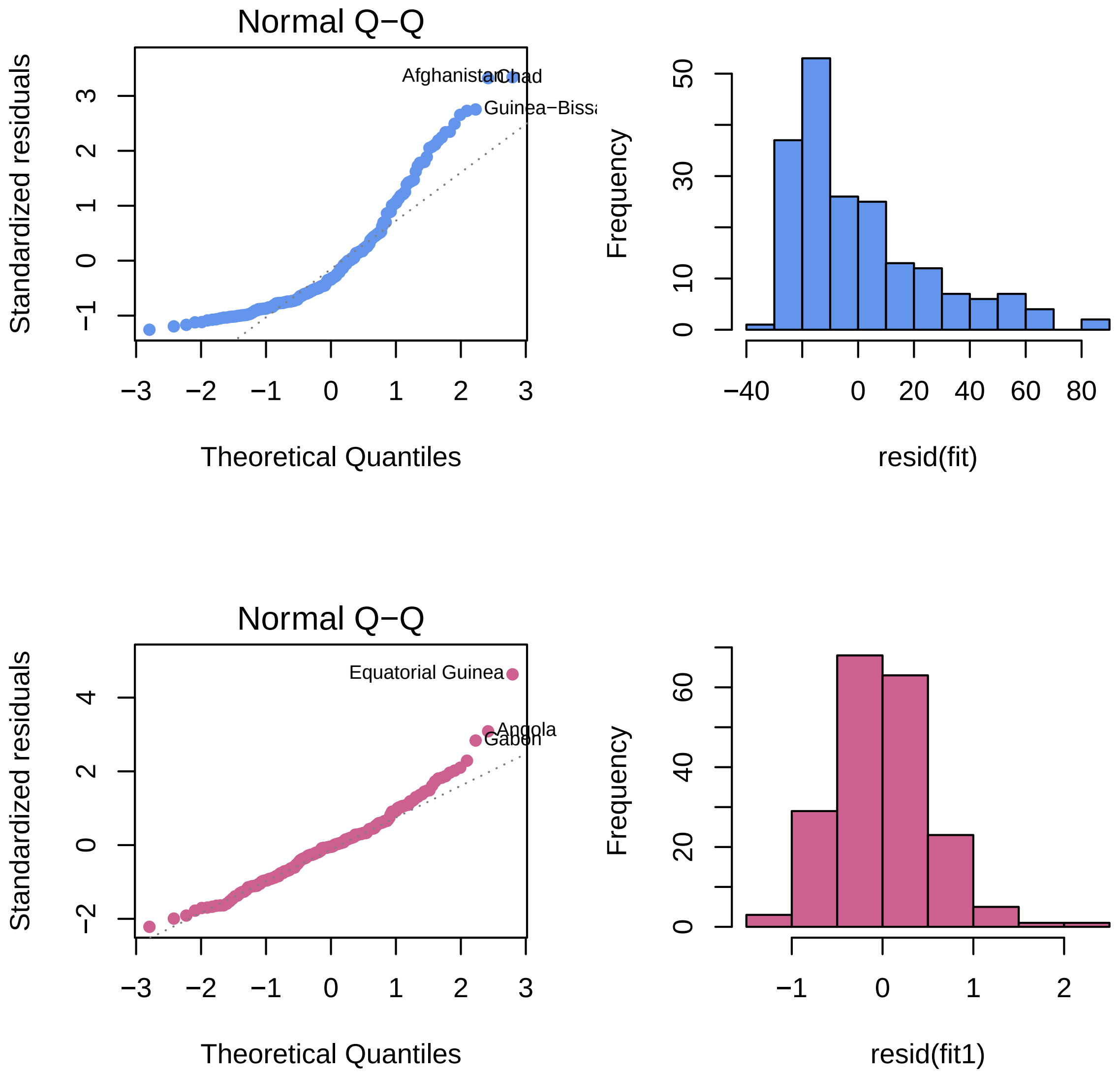
So we check the residuals again, as we can see from the residuals plot below that the log transformation has corrected many of the problems with with residual plot and the residuals now much closer to the expected random scatter.

```
# diagnostic plots
plot(fit1,which=1,pch=16,col="cornflowerblue")
```



Now let us check the Normality of the errors by creating a histogram and normal QQ plot for the residuals, before and after the log transformation. The normal quantile (QQ) plot shows the sample quantiles of the residuals against the theoretical quantiles that we would expect if these values were drawn from a Normal distribution. If the Normal assumption holds, then we would see an approximate straight-line relationship on the Normal quantile plot.

```r
par(mfrow=c(2,2))
# before the log  transformation.
plot(fit, which = 2,pch=16, col="cornflowerblue")
hist(resid(fit),col="cornflowerblue",main="")
# after the log  transformation.
plot(fit1, which = 2, pch=16, col="hotpink3")
hist(resid(fit1),col="hotpink3",main="")
```



The normal quantile plot and the histogram of residuals (before the log transformation) shows strong departure from the expectation of an approximate straight line, with curvature in the tails which reflects the skewness of the data. Finally, the normal quantile plot and the histogram of residuals suggest that residuals are much closer to Normality after the transformation, with some minor deviations in the tails.

2 Simple Linear Regression: Inference

2.1 Simple Linear Regression Assumptions

- Linearity of the relationship between the dependent and independent variables
- Independence of the errors (no autocorrelation)
- Constant variance of the errors (homoscedasticity)
- Normality of the error distribution.

2.2 Simple Linear Regression

The simple linear regression model for Y on x is

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where

Y : the dependent or response variable

x : the independent or predictor variable, assumed known

β_0, β_1 : the regression parameters, the intercept and slope of the regression line

ϵ : the random regression error around the line.

2.3 The simple linear regression equation

- The regression equation for a set of n data points is $\hat{y} = b_0 + b_1\ x$, where

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

  and

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

- y is the dependent variable (or response variable) and x is the independent variable (predictor variable or explanatory variable).
- b_0 is called the **y-intercept** and b_1 is called the **slope**.

2.4 Residual standard error, s_e

The residual standard error, s_e, is defined by

$$s_e = \sqrt{\frac{SSE}{n-2}}$$

where SSE is the error sum of squares (also known as the residual sum of squares, RSS) which can be defined as

$$SSE = \sum e_i^2 = \sum(y_i - \hat{y}_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

s_e indicates how much, on average, the observed values of the response variable differ from the predicated values of the response variable. Under the simple linear regression assumptions, s_e is an unbiased estimate for the error standard deviation σ.

2.5 Properties of Regression Coefficients

Under the simple linear regression assumptions, the least square estimates b_0 and b_1 are unbiased for the β_0 and β_1, respectively, i.e.

$E[b_0] = \beta_0$ and $E[b_1] = \beta_1$.

The variances of the least squares estimators in simple linear regression are:

$$Var[b_0] = \sigma^2_{b_0} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

$$Var[b_1] = \sigma^2_{b_1} = \frac{\sigma^2}{S_{xx}}$$

$$Cov[b_0, b_1] = \sigma_{b_0,b_1} = -\sigma^2 \frac{\bar{x}}{S_{xx}}$$

We use s_e to estimate the error standard deviation σ:

$$s^2_{b_0} = s^2_e \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

$$s^2_{b_1} = \frac{s^2_e}{S_{xx}}$$

$$s_{b_0,b_1} = -s^2_e \frac{\bar{x}}{S_{xx}}$$

2.6 Sampling distribution of the least square estimators

For the Normal error simple linear regression model:

$$b_0 \sim N(\beta_0, \sigma^2_{b_0}) \rightarrow \frac{b_0 - \beta_0}{\sigma_{b_0}} \sim N(0,1)$$

and

$$b_1 \sim N(\beta_1, \sigma^2_{b_1}) \rightarrow \frac{b_1 - \beta_1}{\sigma_{b_1}} \sim N(0,1)$$

We use s_e to estimate the error standard deviation σ:

$$\frac{b_0 - \beta_0}{s_{b_0}} \sim t_{n-2}$$

and

$$\frac{b_1 - \beta_1}{s_{b_1}} \sim t_{n-2}$$

2.7 Degrees of Freedom

- In statistics, degrees of freedom are the number of independent pieces of information that go into the estimate of a particular parameter.
- Typically, the degrees of freedom of an estimate of a parameter are equal to the number of independent observations that go into the estimate, minus the number of parameters that are estimated as intermediate steps in the estimation of the parameter itself.

- The sample variance has $n-1$ degrees of freedom, since it is computed from n pieces of data, minus the 1 parameter estimated as intermediate step, the sample mean. Similarly, having estimated the sample mean we only have $n-1$ independent pieces of data left, as if we are given the sample mean and any $n-1$ of the observations then we can determine the value of remaining observation exactly.

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

- In linear regression, the degrees of freedom of the residuals is $df = n - k^*$, where k^* is the numbers of estimated parameters (including the intercept). So for the simple linear regression, we are estimating β_0 and β_1, thus $df = n - 2$.

2.8 Inference for the intercept β_0

- Hypotheses:

$$H_0 : \beta_0 = 0 \text{ against } H_1 : \beta_0 \neq 0$$

- Test statistic:

$$t = \frac{b_0}{s_{b_0}}$$

has a t-distribution with $df = n - 2$, where s_{b_0} is the standard error of b_0, and given by

$$s_{b_0} = s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

and

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

We reject H_0 at level α if $|t| > t_{\alpha/2}$ with $df = n - 2$.

- 100(1-α)% confidence interval for β_0,

$$b_0 \pm t_{\alpha/2}.\, s_{b_0}$$

where $t_{\alpha/2}$ is critical value obtained from the t-distribution table with $df = n - 2$.

2.9 Inference for the slope β_1

- Hypotheses:

$$H_0 : \beta_1 = 0 \text{ against } H_1 : \beta_1 \neq 0$$

- Test statistic:

$$t = \frac{b_1}{s_{b_1}}$$

has a t-distribution with $df = n - 2$, where s_{b_1} is the standard error of b_1,and given by

$$s_{b_1} = \frac{s_e}{\sqrt{S_{xx}}}$$

and

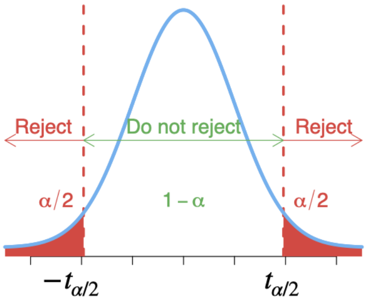
$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

We reject H_0 at level α if $|t| > t_{\alpha/2}$ with $df = n - 2$.

- 100(1-α)% confidence interval for β_1,

$$b_1 \pm t_{\alpha/2}\, s_{b_1}$$

where $t_{\alpha/2}$ is critical value obtained from the t-distribution table with $df = n - 2$.



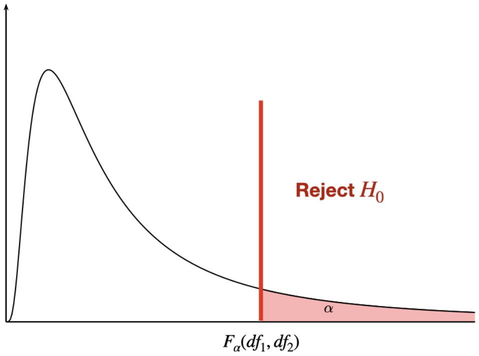
2.10 How useful is the regression model?

Goodness of fit test

- We test the null hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$, the F-statistic

$$F = \frac{MSR}{MSE} = \frac{SSR}{SSE/(n-2)}$$

  has F-distribution with degrees of freedom $df_1 = 1$ and $df_2 = n - 2$.
- We reject H_0, at level α, if $F > F_\alpha(df_1, df_2)$.
- For a simple linear regression ONLY, F-test is equivalent to t-test for β_1.



2.11 Regression in R (Used cars example)

```
Price<-c(85, 103,  70,  82,  89,  98,  66,  95, 169,  70,  48)
Age<- c(5, 4, 6, 5, 5, 5, 6, 6, 2, 7, 7)
carSales<-data.frame(Price,Age)
str(carSales)
```

```
## 'data.frame':    11 obs. of  2 variables:
##  $ Price: num  85 103 70 82 89 98 66 95 169 70 ...
##  $ Age  : num  5 4 6 5 5 5 6 6 2 7 ...
```

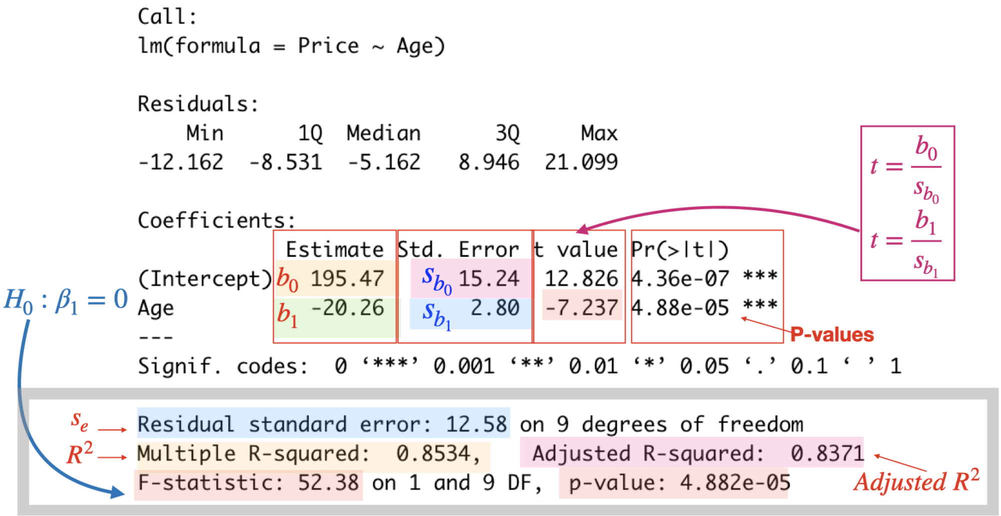
```r
# simple linear regression
reg<-lm(Price~Age)
summary(reg)
```

```
##
## Call:
## lm(formula = Price ~ Age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.162  -8.531  -5.162   8.946  21.099
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept)   195.47      15.24  12.826 0.000000436 ***
## Age           -20.26       2.80  -7.237 0.000048819 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.58 on 9 degrees of freedom
## Multiple R-squared:  0.8534, Adjusted R-squared:  0.8371
## F-statistic: 52.38 on 1 and 9 DF,  p-value: 0.00004882
```

```r
# To obtain the confidence intervals
confint(reg, level=0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) 160.99243 229.94451
## Age         -26.59419 -13.92833
```

2.12 R output



```
Call:
lm(formula = Price ~ Age)

Residuals:
    Min      1Q  Median      3Q     Max
-12.162  -8.531  -5.162   8.946  21.099

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   195.47      15.24  12.826 4.36e-07 ***
Age           -20.26       2.80  -7.237 4.88e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.58 on 9 degrees of freedom
Multiple R-squared:  0.8534,   Adjusted R-squared:  0.8371
F-statistic: 52.38 on 1 and 9 DF,  p-value: 4.882e-05
```

3 Simple Linear Regression: Confidence and Prediction intervals (Workshop 5)

Earlier we have introduced the simple linear regression as a basic statistical model for the relationship between two random variables. We used the least square method for estimating the regression parameters.

Recall that the simple linear regression model for Y on x is

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where

Y : the dependent or response variable

x : the independent or predictor variable, assumed known

β_0, β_1 : the regression parameters, the intercept and slope of the regression line

ϵ : the random regression error around the line.

and the regression equation for a set of n data points is $\hat{y} = b_0 + b_1\ x$, where

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

and

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

where b_0 is called the **y-intercept** and b_1 is called the **slope**.

Under the simple linear regression assumptions, the residual standard error s_e is an unbiased estimate for the error standard deviation σ, where

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$

s_e indicates how much, on average, the observed values of the response variable differ from the predicated values of the response variable.

Below we will see how we can use these least square estimates for prediction. First, we will consider the inference for the conditional mean of the response variable y given a particular value of the independent variable x, let us call this particular value x^*. Next we will see how to predicting the value of the response variable Y for a given value of the independent variable x^*. These confidence and predictive intervals, to be valid, the usual four simple regression assumptions must hold.

3.1 Inference for the regression line $E\left[Y|x^*\right]$

Suppose we are interested in the value of the regression line at a new point x^*. Let's denote the unknown true value of the regression line at $x = x^*$ as μ^*. From the form of the regression line equation we have

$$\mu^* = \mu_{Y|x^*} = E\left[Y|x^*\right] = \beta_0 + \beta_1 x^*$$

but β_0 and β_1 are unknown. We can use the least square regression equation to estimate the unknown true value of the regression line, so we have

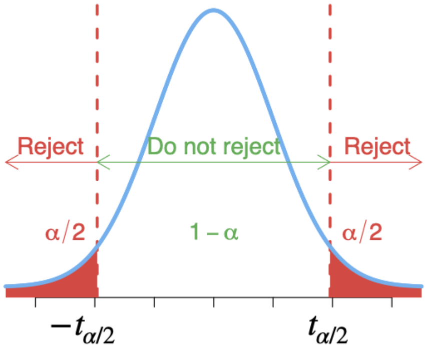
$$\hat{\mu}^* = b_0 + b_1 x^* = \hat{y}^*$$

This is simply a point estimate for the regression line. However, in statistics, point estimate is often not enough, and we need to express our uncertainty about this point estimate, and one way to do so is via confidence interval.

A $100(1-\alpha)\%$ confidence interval for the conditional mean μ^* is

$$\hat{y}^* \pm t_{\alpha/2} \ \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, and $t_{\alpha/2}$ is the $\alpha/2$ critical value from the t-distribution with $df = n - 2$.



3.2 Inference for the response variable Y for a given $x = x^*$

Suppose now we are interested in predicting the value of Y^* if we have a new observation at x^*.

At $x = x^*$, the value of Y^* is unknown and given by

$$Y^* = \beta_0 + \beta_1 x^* + \epsilon$$

where but β_0, β_1 and ϵ are unknown. We will use $\hat{y}^* = b_0 + b_1 \, x^*$ as a basis for our prediction.

A $100(1-\alpha)\%$ prediction interval for Y^* at $x = x^*$ is

$$\hat{y}^* \pm t_{\alpha/2} \ \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

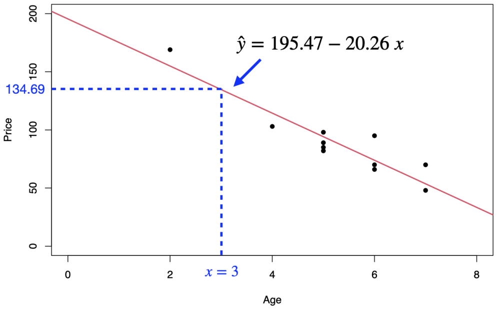
The extra '1' under the square root sign, we have here to account for the extra variability of a single observation about the mean.

Note: we construct a confidence interval for a parameter of the population, which is the conditional mean in this case, while we construct a prediction interval for a single value.

3.3 Used cars example

Estimate the mean price of all 3-year-old cars, $E[Y|x = 3]$**:**

$$\hat{\mu}^* = 195.47 - 20.26(3) = 134.69 = \hat{y}^*$$



A 95% confidence interval for the mean price of all 3-year-old cars is

$$\hat{y}^* \pm t_{\alpha/2} \ \times se\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

$$[195.47 - 20.26(3)] \pm 2.262 \times 12.58\sqrt{\frac{1}{11} + \frac{(3 - 5.273)^2}{(11 - 1) \times 2.018}}$$

$$134.69 \pm 16.76$$

that is

$$117.93 < \mu^* < 151.45$$

Predict the price of a 3-year-old car, $Y|x = 3$:

$$\hat{y}^* = 195.47 - 20.26(3) = 134.69$$

A 95% predictive interval for the price of a 3-year-old car is

$$\hat{y}^* \pm t_{\alpha/2} \ \times se\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

$$[195.47 - 20.26(3)] \pm 2.262 \times 12.58\sqrt{1 + \frac{1}{11} + \frac{(3 - 5.273)^2}{(11 - 1) * \times 2.018}}$$

$$134.69 \pm 33.025$$

that is

$$101.67 < Y^* < 167.72$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = (n-1)Var(x)$.

3.4 Regression in R

```
# Build linear model
Price<-c(85, 103,  70,  82,  89,  98,  66,  95, 169,  70,  48)
Age<- c(5, 4, 6, 5, 5, 5, 6, 6, 2, 7, 7)
carSales<-data.frame(Price=Price,Age=Age)

reg <- lm(Price~Age,data=carSales)
summary(reg)
```

```
##
## Call:
## lm(formula = Price ~ Age, data = carSales)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.162  -8.531  -5.162   8.946  21.099
##
## Coefficients:
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept)   195.47      15.24  12.826 0.000000436 ***
## Age           -20.26       2.80  -7.237 0.000048819 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.58 on 9 degrees of freedom
## Multiple R-squared:  0.8534, Adjusted R-squared:  0.8371
## F-statistic: 52.38 on 1 and 9 DF,  p-value: 0.00004882
```

```
mean(Age)
```

```
## [1] 5.272727
```

```
var(Age)
```

```
## [1] 2.018182
```

```
qt(0.975,9)
```

```
## [1] 2.262157
```

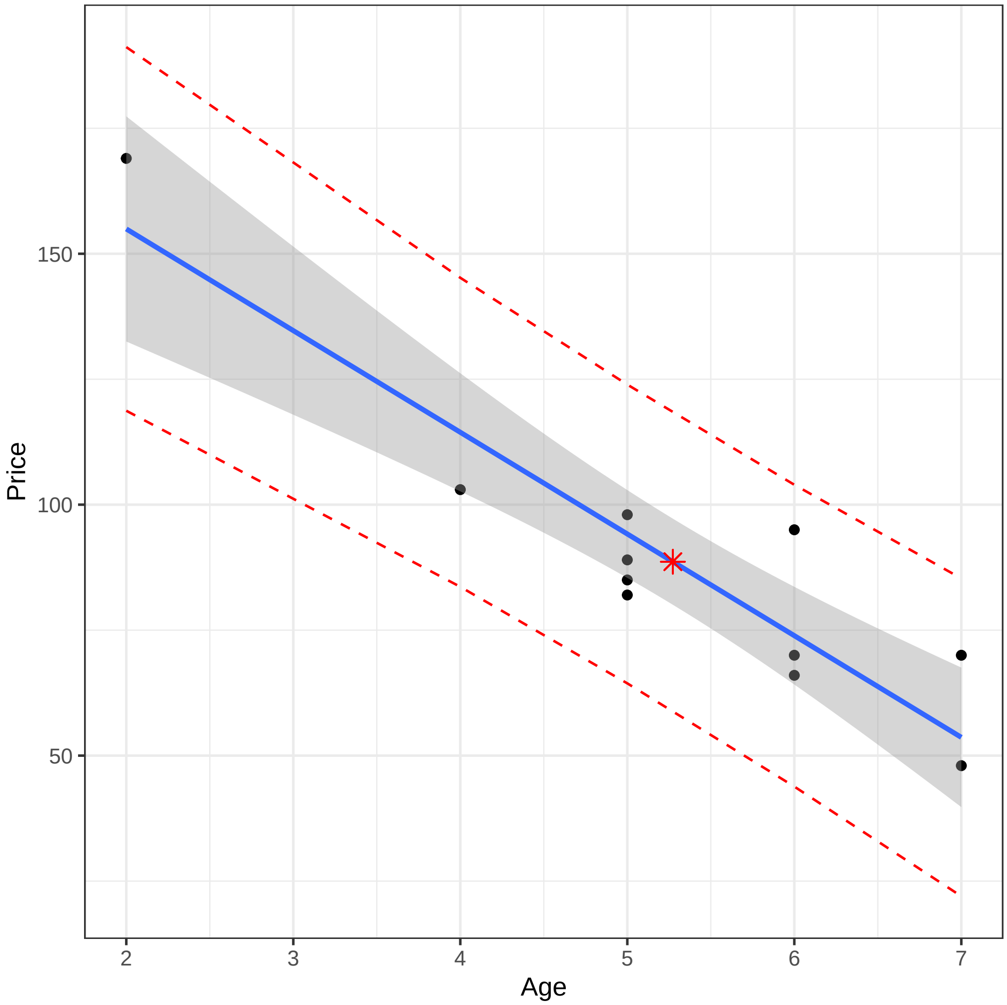
```
newage<- data.frame(Age = 3)
predict(reg, newdata = newage, interval = "confidence")
```

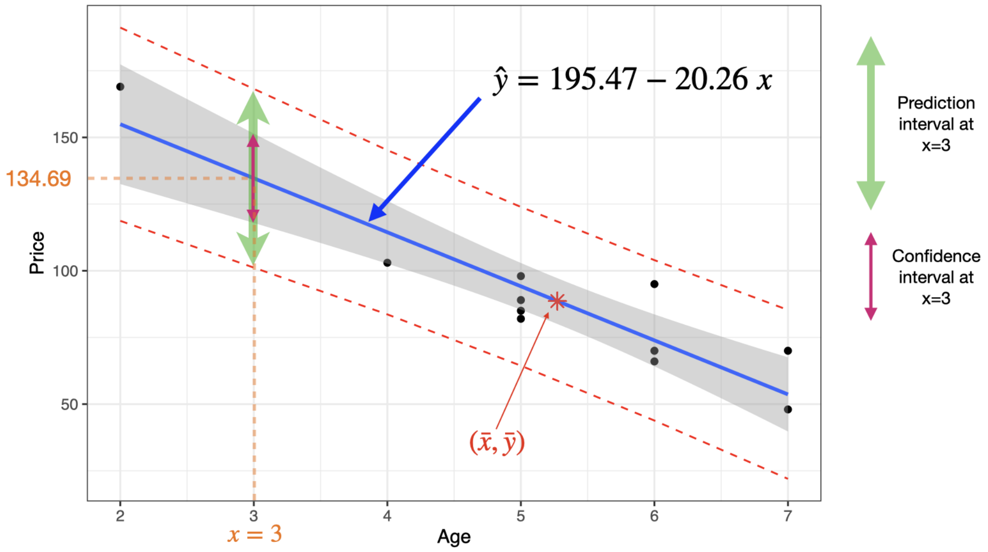
```
##        fit      lwr      upr
## 1 134.6847 117.9293 151.4401
```

```
predict(reg, newdata = newage, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 134.6847 101.6672 167.7022
```

We can plot the confidence and prediction intervals as follows:


Price
150
100
50
2
3
4
5
6
7
Age


$\hat{y} = 195.47 - 20.26\, x$
150
134.69
100
50
Price
$(\bar{x}, \bar{y})$
2
3
4
5
6
7
$x = 3$
Age
Prediction interval at x=3
Confidence interval at x=3

4 Multiple Linear Regression: Introduction

4.1 Multiple linear regression model

In simple linear regression, we have one dependent variable (y) and one independent variable (x). In multiple linear regression, we have one dependent variable (y) and several independent variables ($x_1, x_2, \ldots, x_k$).

- The multiple linear regression model, for the **population**, can be expressed as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

  where ϵ is the error term.

- The corresponding least square estimate, from the **sample**, of this multiple linear regression model is given by

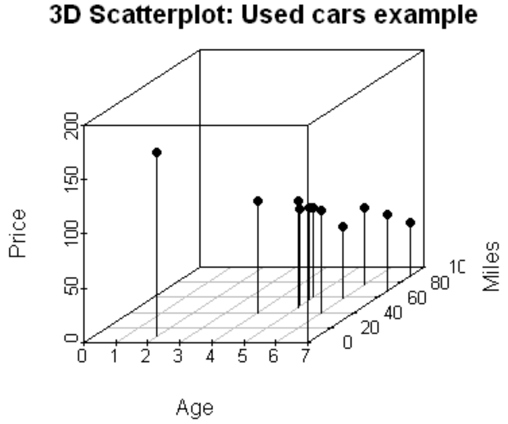
$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

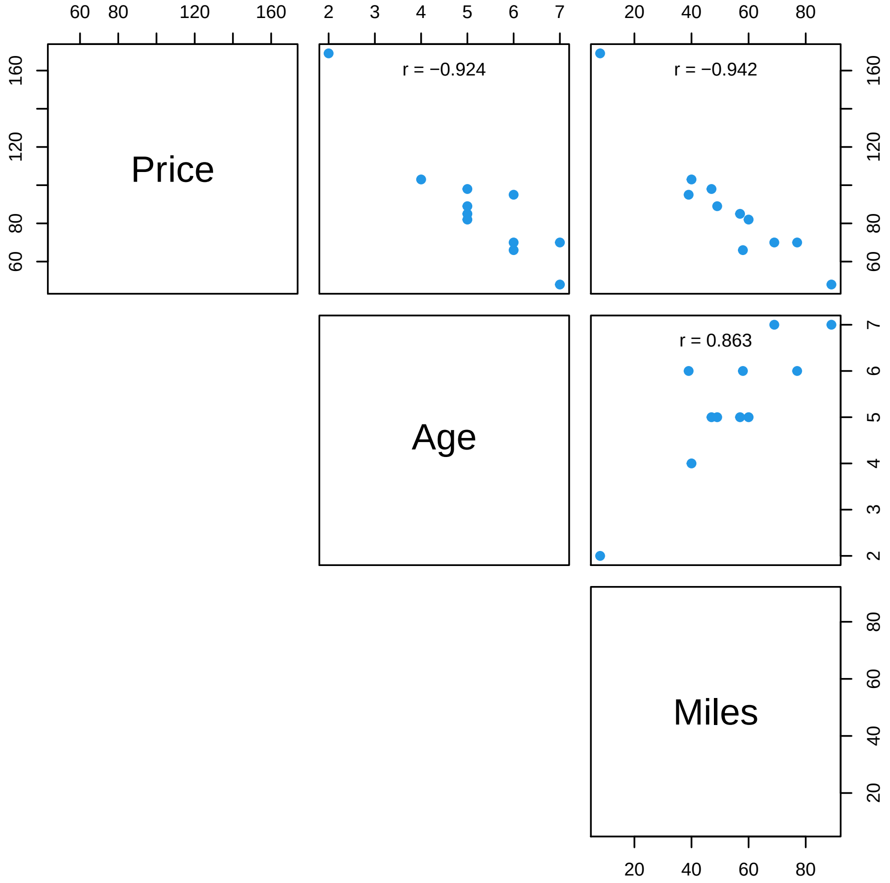
- The coefficient b_0 (or β_0) represents the y-intercept, that is, the value of y when $x_1 = x_2 = \ldots = x_k = 0$. The coefficient b_i (or β_i) ($i = 1, \ldots, k$) is the partial slope of x_i, holding all other x's fixed. So b_i (or β_i) tells us the change in y for a unit increase in x_i, holding all other x's fixed.

4.2 Used cars example

The table below displays data on Age, Miles and Price for a sample of cars of a particular make and model.

Price (y)	Age (x_1)	Miles (x_2)
85	5	57
103	4	40
70	6	77
82	5	60
89	5	49
98	5	47
66	6	58
95	6	39
169	2	8
70	7	69
48	7	89





The scatterplot and the correlation matrix show a fairly negative relationship between the price of the car and both independent variables (age and miles). It is desirable to have a relationship between each independent variable and the dependent variable. However, the scatterplot also shows a positive relationship between the age and the miles, which isundesirable as it will cause the issue of Multicollinearity.

4.3 Coefficient of determination, R^2 and adjusted R^2

- Recall that, R^2 is a measure of the proportion of the total variation in the observed values of the response variable that is explained by the multiple linear regression in the k predictor variables $x_1, x_2, \ldots, x_k$.
- R^2 will increase when an additional predictor variable is added to the model. One should not simply select a model with many predictor variables because it has the highest R^2 value, it is often good to have a model with high R^2 value but only few x's included.
- Adjusted R^2 is a modification of R^2 that takes into account the number of predictor variables.

$$\text{Adjusted-}R^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

4.4 The residual standard error, s_e

- Recall that,

$$\text{Residual} = \text{Observed value} - \text{Predicted value}.$$

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

- In a multiple linear regression with k predictors, the standard error of the estimate, s_e, is defined by

$$s_e = \sqrt{\frac{SSE}{n-(k+1)}} \quad \text{where } SSE = \sum (y_i - \hat{y}_i)^2$$

- The standard error of the estimate, s_e, indicates how much, on average, the observed values of the response variable differ from the predicated values of the response variable. The s_e is the estimate of the common standard deviation σ.

4.5 Inferences about a particular predictor variable

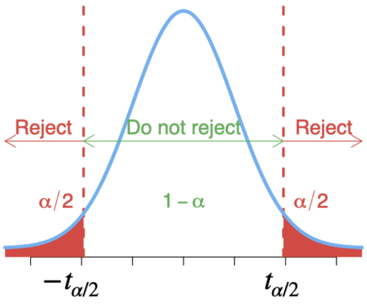
- To test whether a particular predictor variable, say x_i, is useful for predicting y we test the null hypothesis $H_0 : \beta_i = 0$ against $H_1 : \beta_i \neq 0$.
- The test statistic

$$t = \frac{b_i}{s_{b_i}}$$

  has a t-distribution with degrees of freedom $df = n - (k+1)$. So we reject H_0, at level α, if $|t| > t_{\alpha/2}$.
- Rejection of the null hypothesis indicates that x_i is useful as a predictor for y. However, failing to reject the null hypothesis suggests that x_i may not be useful as a predictor of y, so we may want to consider removing this variable from the regression analysis.
- 100(1-α)% confidence interval for β_i is

$$b_i \pm t_{\alpha/2}.s_{b_i}$$

  where s_{b_i} is the standard error of b_i.



4.6 How useful is the multiple regression model?

Goodness of fit test

To test how useful is this model, we test the null hypothesis

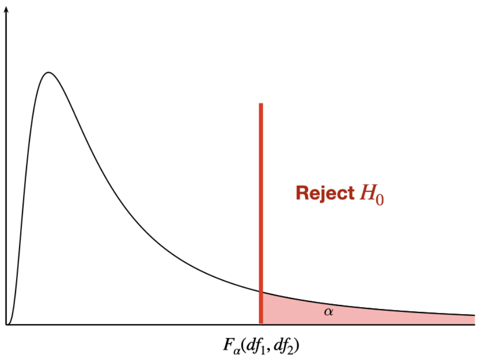
$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$, against

H_1 : at least one of the β_i's is not zero. - The F-statistic

$$F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)}$$

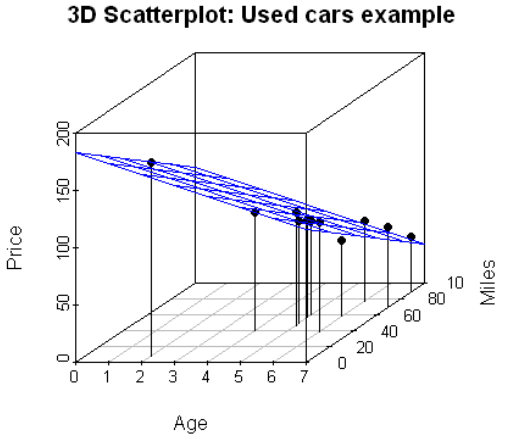
with degrees of freedom $df_1 = k$ and $df_2 = n - (k+1)$.

We reject H_0, at level α, if $F > F_\alpha(df_1, df_2)$.



4.7 Used cars example continued

Multiple regression equation: $\hat{y} = 183.04 - 9.50x_1 - 0.82x_2$



The predicted price for a 4-year-old car that has driven 45 thousands miles is

$$\hat{y} = 183.04 - 9.50(4) - 0.82(45) = 108.14$$

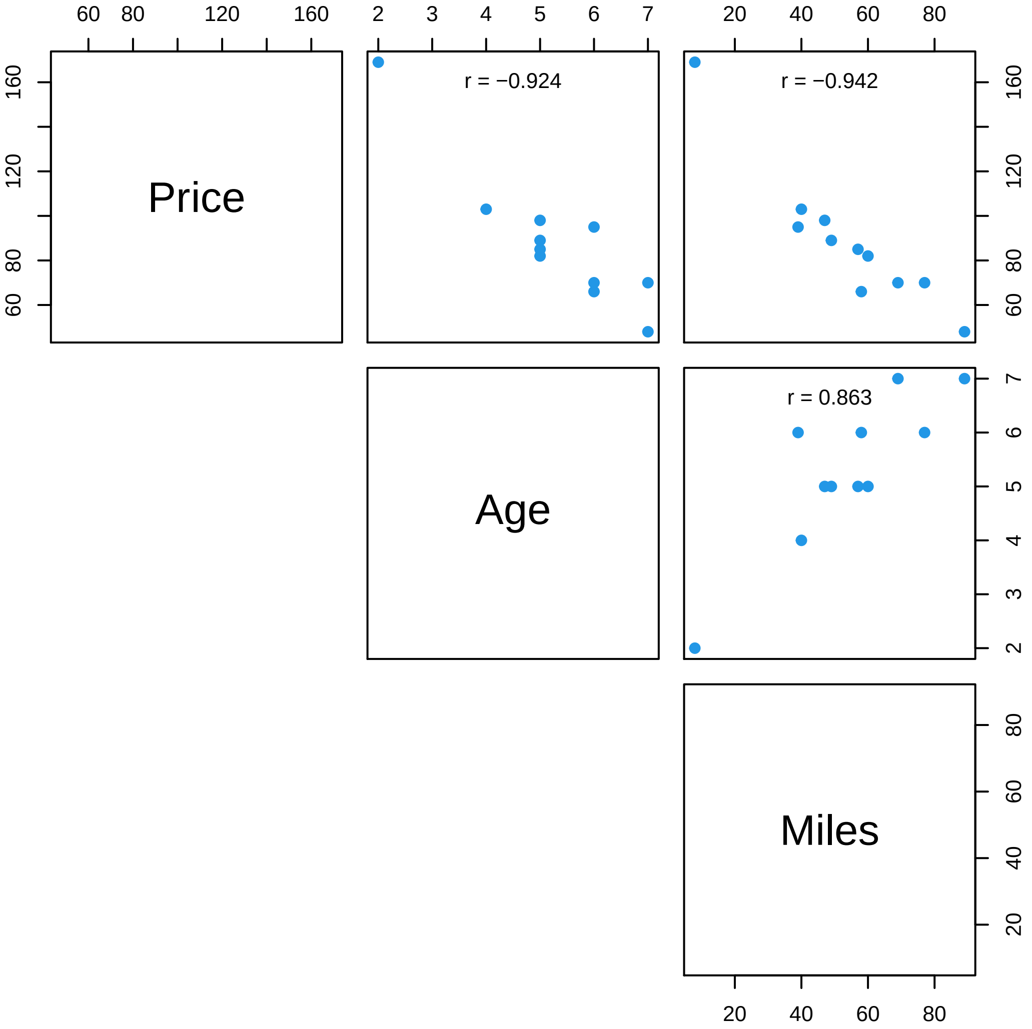
(as units of \$100 were used, this means \$10814)

Extrapolation: we need to look at the region (all combined values) not only the range of the observed values of each predictor variable separately.

4.8 Regression in R

```r
Price<-c(85, 103,  70,  82,  89,  98,  66,  95, 169,  70,  48)
Age<- c(5, 4, 6, 5, 5, 5, 6, 6, 2, 7, 7)
Miles<-c(57,40,77,60,49,47,58,39,8,69,89)
carSales<-data.frame(Price=Price,Age=Age,Miles=Miles)


# Scatterplot matrix
# Customize upper panel
upper.panel<-function(x, y){
  points(x,y, pch=19, col=4)
  r <- round(cor(x, y), digits=3)
  txt <- paste0("r = ", r)
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  text(0.5, 0.9, txt)
}
pairs(carSales, lower.panel = NULL,
      upper.panel = upper.panel)
```



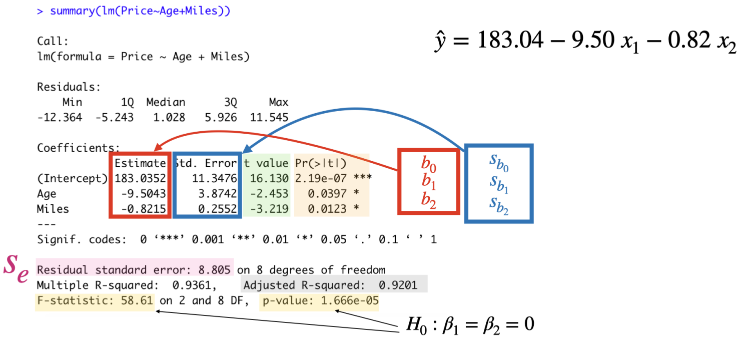
```
reg <- lm(Price~Age+Miles,data=carSales)
summary(reg)
```

```
##
## Call:
## lm(formula = Price ~ Age + Miles, data = carSales)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.364  -5.243   1.028   5.926  11.545
##
## Coefficients:
##               Estimate Std. Error t value    Pr(>|t|)
## (Intercept) 183.0352    11.3476  16.130 0.000000219 ***
## Age          -9.5043     3.8742  -2.453      0.0397 *
## Miles        -0.8215     0.2552  -3.219     0.0123 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.805 on 8 degrees of freedom
## Multiple R-squared:  0.9361, Adjusted R-squared:  0.9201
## F-statistic: 58.61 on 2 and 8 DF,  p-value: 0.00001666
```

```r
confint(reg, level=0.95)
```

```
##                  2.5 %      97.5 %
## (Intercept) 156.867552 209.2028630
## Age         -18.438166  -0.5703751
## Miles        -1.409991  -0.2329757
```

4.8.1 Summary



```
> summary(lm(Price~Age+Miles))

Call:
lm(formula = Price ~ Age + Miles)

Residuals:
    Min      1Q  Median      3Q     Max 
-12.364  -5.243   1.028   5.926  11.545 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 183.0352    11.3476  16.130 2.19e-07 ***
Age          -9.5043     3.8742  -2.453   0.0397 *  
Miles        -0.8215     0.2552  -3.219   0.0123 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.805 on 8 degrees of freedom
Multiple R-squared:  0.9361,	Adjusted R-squared:  0.9201 
F-statistic: 58.61 on 2 and 8 DF,  p-value: 1.666e-05
```

$\hat{y} = 183.04 - 9.50\, x_1 - 0.82\, x_2$
b_0 b_1 b_2
s_{b_0} s_{b_1} s_{b_2}
s_e
$H_0 : \beta_1 = \beta_2 = 0$

	β_0	β_1	β_2
H_0 H_1	$H_0 : \beta_0 = 0$ $H_1 : \beta_0 \neq 0$	$H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$	$H_0 : \beta_2 = 0$ $H_1 : \beta_2 \neq 0$
Estimate of β_i	$b_0 = 183.04$	$b_1 = 9.50$	$b_2 = 0.82$
$t = \frac{b_i}{s_{b_i}}$	16.130	-2.453	-3.219
P-value	0	0.040	0.012
Decision*	reject H_0	reject H_0	reject H_0
95% CI for β_i	(156.868,209.203)	(-18.438,-0.570)	(-1.410,-0.233)

$*$ at $\alpha = 0.05$.

4.9 Multiple Linear Regression Assumptions

- **Linearity**: For each set of values, $x_1, x_2, \ldots, x_k$, of the predictor variables, the conditional mean of the response variable y is $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$.
- **Equal variance (homoscedasticity)**: The conditional variance of the response variable are the same (equal to σ^2) for all sets of values, $x_1, x_2, \ldots, x_k$, of the predictor variables.
- **Independent observations**: The observations of the response variable are independent of one another.
- **Normally**: For each set values, $x_1, x_2, \ldots, x_k$, of the predictor variables, the conditional distribution of the response variable is a normal distribution.

- **No Multicollinearity**: Multicollinearity exists when two or more of the predictor variables are highly correlated.

4.9.1 Multicollinearity

- Multicollinearity refers to a situation when two or more predictor variables in our multiple regression model are highly (linearly) correlated.
- The least square estimates will remain unbiased, but unstable.
- The standard errors (of the affected variables) are likely to be high.
- Overall model fit (e.g. R-square, F, prediction) is not affected.

4.9.2 Multicollinearity: Detect

- Scatterplot Matrix
- **Variance Inflation Factors**: the Variance Inflation Factors (VIF) for the i^{th} predictor is

  $$VIF_i = \frac{1}{1 - R_i^2}$$

  where R_i^2 is the R-square value obtained by regressing the i^{th} predictor on the other predictor variables.
- $VIF = 1$ indicates that there is no correlation between i^{th} predictor variable and the other predictor variables.
- As rule of thumb if $VIF > 10$ then multicollinearity could be a problem.

4.9.3 Multicollinearity: How to fix?

Ignore: if the model is going to be used for prediction only.

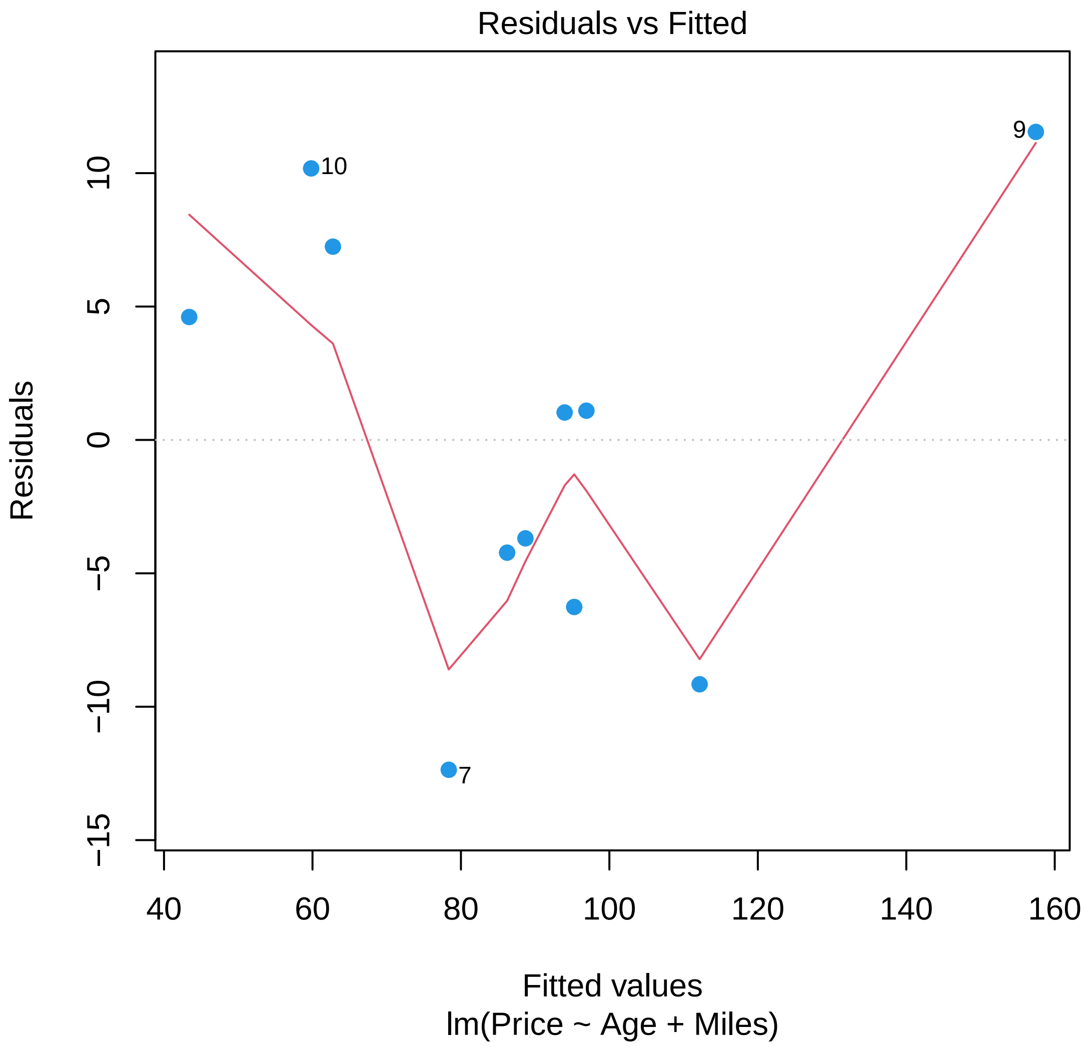
Remove: e.g. see if the variables are providing the same information.

Combine: combining highly correlated variables.

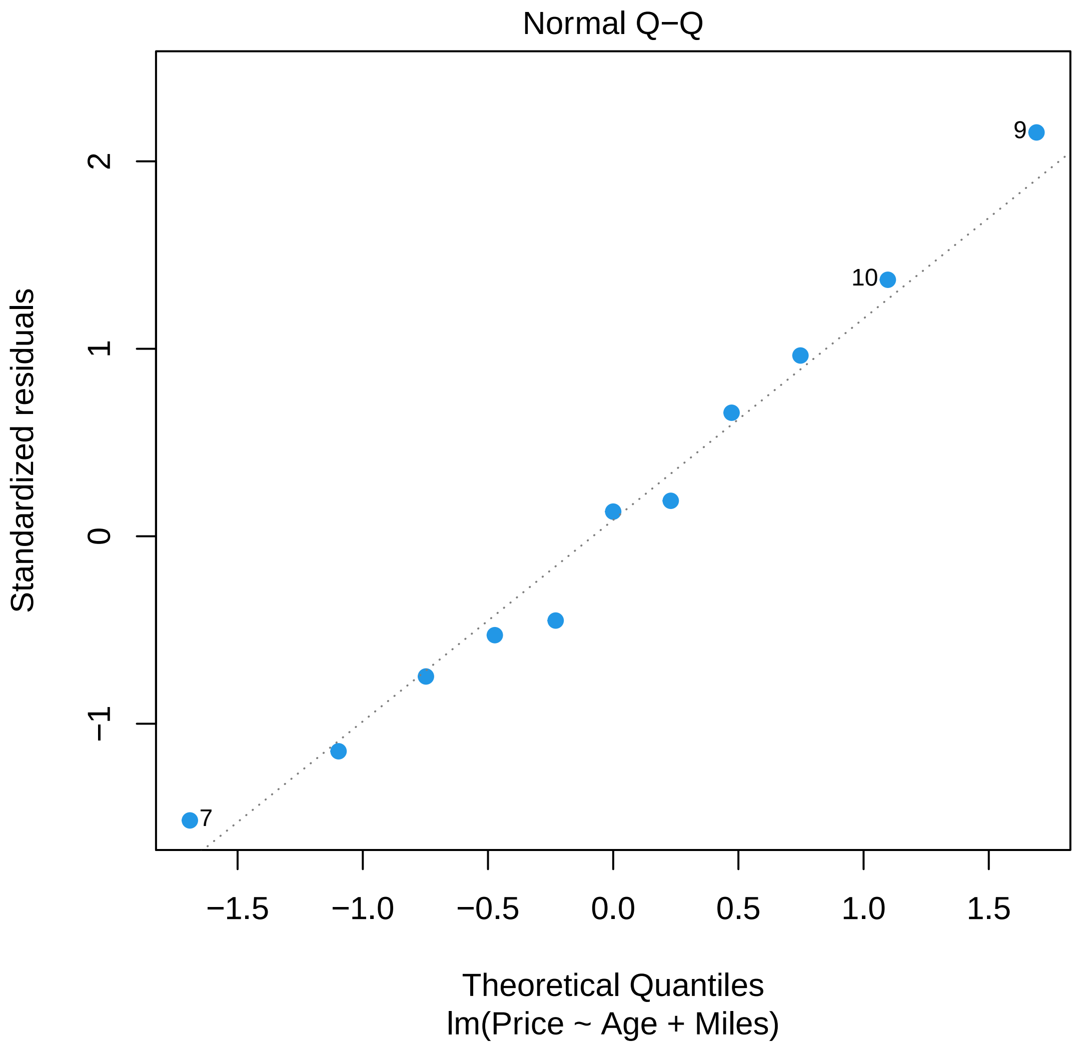
Advanced: e.g. Principal Components Analysis, Partial Least Squares.

4.10 Regression in R (regression assumptions)

```r
plot(reg, which=1, pch=19, col=4)
```



```
plot(reg, which=2, pch=19, col=4)
```



```
# install.packages("car")
library(car)
vif(reg)
```

```
##      Age    Miles
## 3.907129 3.907129
```

The value of $VIF = 3.91$ indicates a moderate correlation between the age and the miles in the model, but this is not a major concern.