Data Mining Laboratory

Ewa Figielska

Regression

- Regression modeling represents a powerful and elegant method for estimating the value of a continuous target variable.

- Simple linear regression

  A straight line is used to approximate the relationship between a single continuous predictor variable and a single continuous response variable.

- Multiple regression

  Several predictor variables are used to estimate a single response.

Regression equation

- The regression line is written in the following form

$$\textbf{(1)} \qquad \hat{\boldsymbol{y}} = \boldsymbol{b_0} + \boldsymbol{b_1 x}$$

where:

$\hat{y}$ is the estimated value of the response variable

x is the predictor variable

b_0 is the *y-intercept* of the regression line

b_1 is the *slope* of the regression line

b_0 and b_1, together, are called the **regression coefficients**

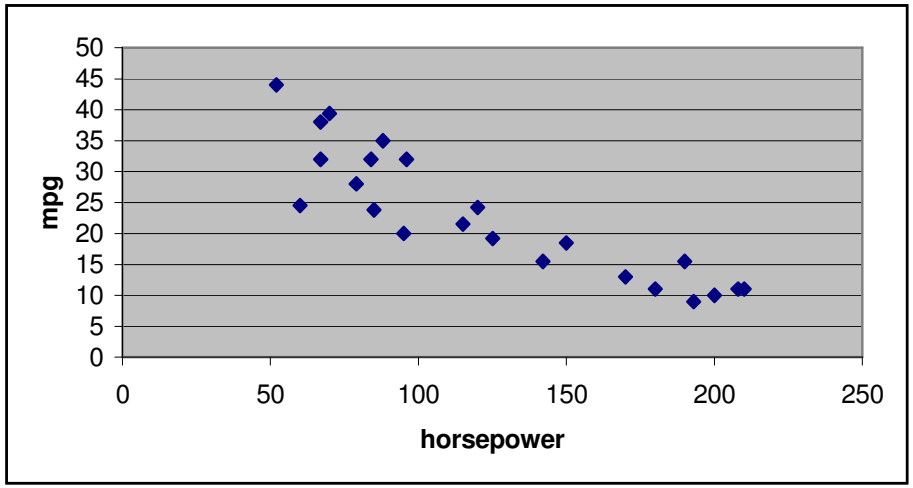
- Eq. (1) is called the **regression equation** or the **estimated regression equation (ERE)**

- The regression line is used as a **linear approximation** of the relationship between the x (predictor) and y (response) variables.
- We can use the regression line to make estimates or predictions.

Example

cylinder	displacement	horsepower	weight	acceleration	mpg
8	307	200	4376	15	10
8	318	210	4382	13.5	11
8	304	193	4732	18.5	9
8	429	208	4633	11	11
8	360	170	4654	13	13
8	350	180	3664	11	11
6	198	95	3102	16.5	20
4	98	60	2164	22.1	24.5
8	400	190	4325	12.2	15.5
4	85	70	2070	18.6	39.4
4	151	85	2855	17.6	23.8
6	231	115	3245	15.4	21.5
8	351	142	4054	14.3	15.5
8	267	125	3605	15	19.2
8	360	150	3940	13	18.5
4	122	88	2500	15.1	35
6	146	120	2930	13.8	24.2
4	91	67	1965	15.7	32
4	91	67	1995	16.2	38
4	144	96	2665	13.9	32
4	97	52	2130	24.6	44
4	135	84	2295	11.6	32
4	120	79	2625	18.6	28

- Given the values of horspower estimate the values of mpg



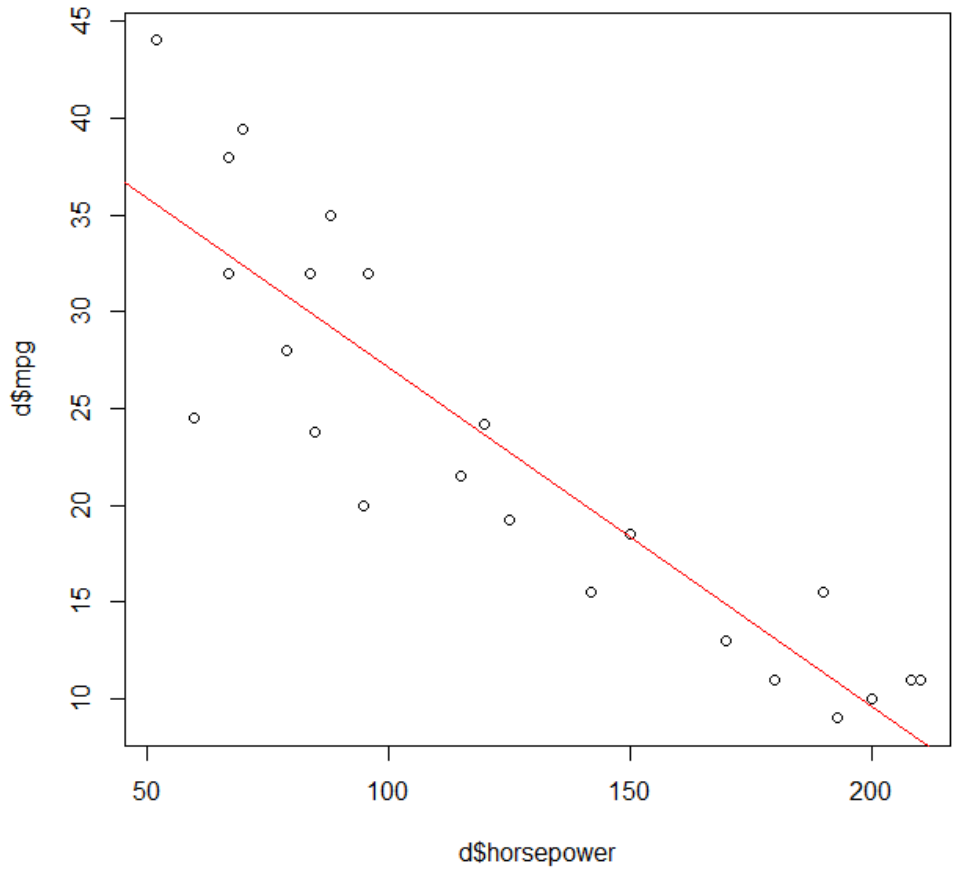
Example

- The ERE is given as $\hat{y} = 44.6422 - 0.1752x$
- "The estimated mpg equals 44.6422 minus 0.1752 times the horsepower"

```
d<-read.csv(file="autompg.csv"
model<-lm(d$mpg ~ d$horsepower) //lm() creates a linear model
plot(d$horsepower,d$mpg) //creates a scatter plot
abline(model,col='red') //adds the model line to the plot
```

The result of calling the `model:`

Call:
lm(formula = d$mpg ~ d$horsepower)
Coefficients:
(Intercept) d$horsepower
44.6422 -0.1752

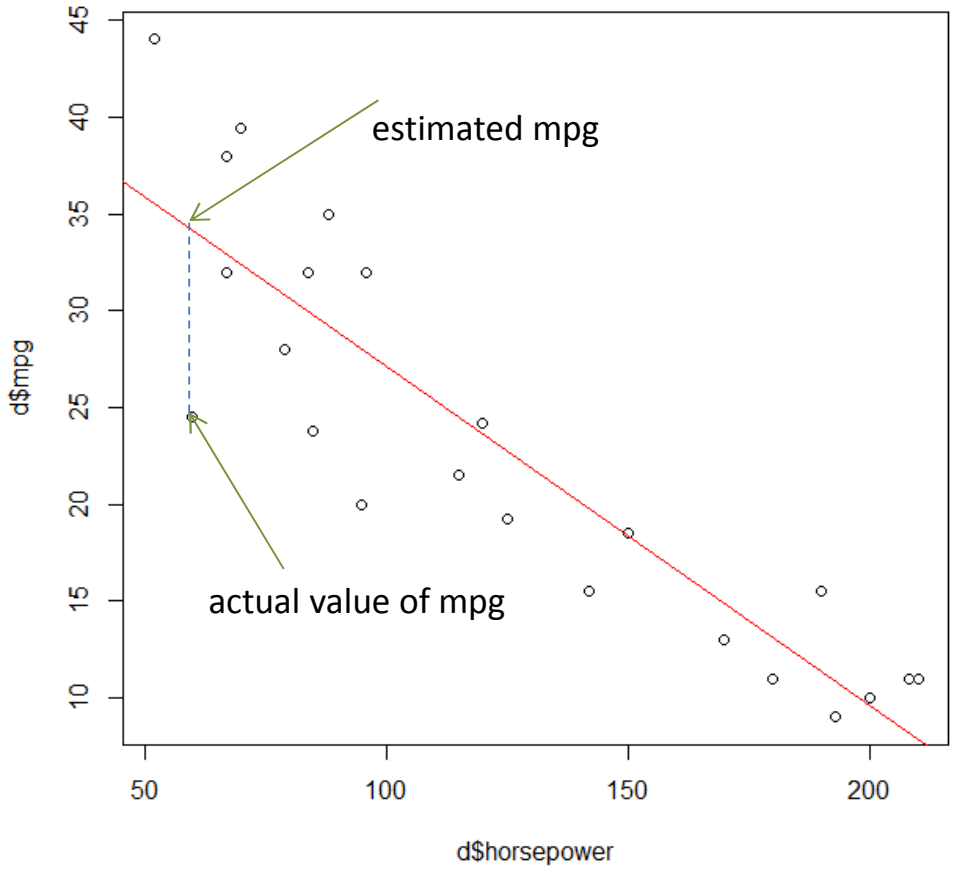


Estimation/prediction error

- Suppose that we are interested in estimating the mpg for a new car (not in the original data) with horsepower = 60.
- Using the ERE, we find the estimated mpg, $\hat{y} = 44.642 - 0.175 \cdot 60 = 34.129.$

- The prediction is too high by $34.129 - 24.5 = 9.629$, which represents the vertical distance from the data point to the regression line.

- The vertical distance $(\boldsymbol{y} - \hat{\boldsymbol{y}})$, is known as the **prediction error**, **estimation error**, **or residual**.



Least-squares estimates

- The most common method for determining the regression line: **least squares regression**
- Least squares regression works by choosing the unique regression line that minimizes the sum of squared residuals over all the data points.
- The true linear relationship is given by:

(2) $\boldsymbol{y = \beta_0 + \beta_1 x + \varepsilon,}$ where ε is an error term

Expression (2) is called **true population regression equation.**

- We seek to minimize the overall size of the prediction errors.

Least-squares estimates

- Suppose that we have *n* observations:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ for } i = 1, \ldots n$$

- The population **sum of squared errors**:

$$(3) \quad SSE_p = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- The least-squares line minimizes the population sum of squared errors, SSE_p .

- We may find the values of $\beta_0$and β_1 which minimize SSE_p by differentiating (3) with respect to β_0 and β_1 and setting the results equal to zero.
- The partial derivatives:

$$\frac{dSSE_p}{d\beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{dSSE_p}{d\beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

- The estimates b_0 and b_1, are calculated from the following equations:

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0$$

- Finally, we obtain:

$$(4) \quad b_1 = \frac{\sum x_i y_i - [(\sum x_i)(\sum y_i)]/n}{\sum x_i^2 - (\sum x_i)^2/n}$$

$$(5) \quad b_0 = \bar{y} - b_1 \bar{x}$$

Measuring the quality of a regression model

- A least-squares regression line can be found to approximate the relationship between any two continuous variables.
- This does not guarantee that the regression will be useful.

- How we can determine whether a particular estimated regression equation is useful for making predictions?

Does the linear relationship exists? Hypothesis testing.

- The population regression equation: $y = \beta_0 + \beta_1 x + \varepsilon$, where ε represents a random variable for modeling the errors.
- When $\beta_1 = 0$, the population regression equation becomes $y = \beta_0 + \varepsilon$, i.e. there is no relationship between x and y. When $\beta_1 \neq 0$, there is a linear relationship between x and y

- To test for the existence of a linear relationship between x and y, the following hypothesis test can be performed:

  H_0: β_1= 0. No relationship between x and y.

  H_a: $\beta_1 \neq 0$. Linear relationship between x and y.

- To decide whether to support or reject the null hypothesis H_0 the **test statistic (t-statistic)** is used.
- The test statistic for this hypothesis test is:

  $\boldsymbol{t = \frac{b_1}{s_{b_1}}}$, where s_{b_1} is the **standard error** of coefficient b_1

  s_{b_1} is a measure of the variability in the slope of the regression line observed among the samples. Large values of s_{b_1} tend to reduce the size of t-statistic.

  Standard error of regression slope,s_{b_1}, is given by: $s_{b_1} = \sqrt{\sum (y_i - \hat{y}_i)^2/(n-2)} / \sqrt{\sum (x_i - \bar{x})^2}$

- **p-value** indicates the probability of observing the t-value (given by $\boldsymbol{t = b_1/s_{b_1}}$), if there is no relationship between x and y.
- If p-value (usually < 0.05) is small, the hypothesis H_0 is rejected. This indicates that $\beta_1 \neq 0$, i.e. that there exists a linear relationship between x and y.

Example. Hypothesis testing.

#	x (horsep.)	y (mpg)	$\hat{y}_i = b_1 x_i + b_0$ (Score predicted)	$y_i - \hat{y}_i$ (Residual)	$(y_i - \hat{y}_i)^2$ (Squared error, SE)	$(x_i - \bar{x})^2$
1	200	10	9.60	0.40	0.16	5815.72
2	210	11	7.85	3.15	9.92	7440.94
3	193	9	10.83	-1.83	3.34	4797.07
4	208	11	8.20	2.80	7.84	7099.89
5	170	13	14.86	-1.86	3.45	2140.07
6	180	11	13.11	-2.11	4.44	3165.29
7	95	20	28.00	-8.00	63.97	825.94
8	60	24.5	34.13	-9.63	92.74	4062.68
9	190	15.5	11.35	4.15	17.19	4390.50
10	70	39.4	32.38	7.02	49.31	2887.89
11	85	23.8	29.75	-5.95	35.40	1500.72
12	115	21.5	24.49	-2.99	8.97	76.37
13	142	15.5	19.76	-4.26	18.18	333.46
14	125	19.2	22.74	-3.54	12.55	1.59
15	150	18.5	18.36	0.14	0.02	689.63
16	88	35	29.22	5.78	33.36	1277.29
17	120	24.2	23.62	0.58	0.34	13.98
18	67	32	32.90	-0.90	0.82	3219.33
19	67	38	32.90	5.10	25.97	3219.33
20	96	32	27.82	4.18	17.45	769.46
21	52	44	35.53	8.47	71.71	5146.50
22	84	32	29.93	2.07	4.30	1579.20
23	79	28	30.80	-2.80	7.85	2001.59
Sum =	2846.00	528.10	528.15	-0.05	489.26	62454.43
Mean =	123.74	22.96	22.96	-0.002 (median = 0.14)	21.27	2715.41

SSE (sum of squared errors (residuals))

small value due to the least squares regression

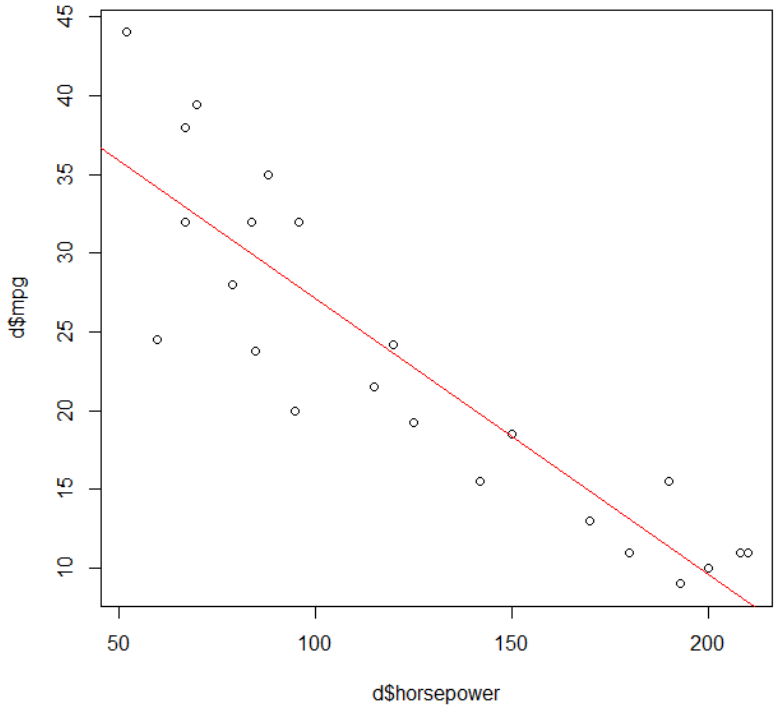
$s_{b_1} = \sqrt{\sum (y_i - \hat{y}_i)^2/(n-2)} / \sqrt{\sum (x_i - \bar{x})^2} = 0.0193$

t-statistic $= -9.071$, p-value < 0.00001 (significance level $= 0.05$)

Residual standard error (i.e. residual standard deviation) $= \sqrt{\sum (y - \hat{y})^2/(n-2)} = 4.827$ (here (n – 2)= 21 degree of freedom)

Example. Hypothesis testing.

```
model<-lm(d$mpg ~ d$horsepower)
summary(model)
```




```
Call:
lm(formula = d$mpg ~ d$horsepower)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6291 -2.8960  0.1405  3.6514  8.4692

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.64218    2.59321  17.215 7.35e-14 ***
d$horsepower -0.17522    0.01931  -9.072 1.04e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.827 on 21 degrees of freedom
```

Coeficient of determination, R^2

- **Coefficient of determination, R^2** (r-squared), measures how closely linear regression fits the data.

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

- SSR is the **sum of squares regression** (regression sum of squares), $SSR = \sum_i (\hat{y}_i - \bar{y})^2$
- SSR represents the differences between the prediction for each observation and the population mean (the improvement in estimation from using the regression model as compared to just using $\bar{y}$ to estimate y).

- SST is the **sum of squares total** (total sum of squares), $SST = \sum_i (y_i - \bar{y})^2$
- SST represents the variability in the y variable.

- The values of R^2 close to 1 indicate a perfect fit.

Example. R^2

#	x (horsep.)	y (mpg)	$\hat{y}_i = b_1 x_i + b_0$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$ (Square regression, SR)	$y_i - \bar{y}$	$(y_i - \bar{y})^2$ (Square total, ST)
1	200	10	9.60	-13.36	178.45	-12.96	167.98
2	210	11	7.85	-15.11	228.33	-11.96	143.06
3	193	9	10.83	-12.13	147.19	-13.96	194.91
4	208	11	8.20	-14.76	217.87	-11.96	143.06
5	170	13	14.86	-8.10	65.65	-9.96	99.22
6	180	11	13.11	-9.85	97.11	-11.96	143.06
7	95	20	28.00	5.04	25.37	-2.96	8.77
8	60	24.5	34.13	11.17	124.75	1.54	2.37
9	190	15.5	11.35	-11.61	134.71	-7.46	55.66
10	70	39.4	32.38	9.42	88.69	16.44	270.25
11	85	23.8	29.75	6.79	46.10	0.84	0.70
12	115	21.5	24.49	1.53	2.35	-1.46	2.13
13	142	15.5	19.76	-3.20	10.22	-7.46	55.66
14	125	19.2	22.74	-0.22	0.05	-3.76	14.14
15	150	18.5	18.36	-4.60	21.15	-4.46	19.90
16	88	35	29.22	6.26	39.23	12.04	144.94
17	120	24.2	23.62	0.66	0.43	1.24	1.54
18	67	32	32.90	9.94	98.86	9.04	81.71
19	67	38	32.90	9.94	98.86	15.04	226.18
20	96	32	27.82	4.86	23.64	9.04	81.71
21	52	44	35.53	12.57	158.03	21.04	442.65
22	84	32	29.93	6.96	48.50	9.04	81.71
23	79	28	30.80	7.84	61.47	5.04	25.39
Sum =	2846.00	528.10	528.15	0.05	1917.04	0.00	2406.69
Determination = 1917.04 /2406.69 = 0.797							

Example. R^2

```
model<-lm(d$mpg ~ d$horsepower)
summary(model)
help(summary.lm) # provides help on summary
```

```
Call:
lm(formula = d$mpg ~ d$horsepower)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6291 -2.8960  0.1405  3.6514  8.4692

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.64218    2.59321  17.215 7.35e-14 ***
d$horsepower -0.17522    0.01931  -9.072 1.04e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.827 on 21 degrees of freedom
Multiple R-squared:  0.7967,    Adjusted R-squared:  0.787
F-statistic:  82.3 on 1 and 21 DF,  p-value: 1.037e-08
```

Making predictions and determining confidence intervals

- The predicted values of mpg for 3 new values of the horsepower parameter, 91, 205 and 133.

$$x = 91, \quad \hat{y} = 44.6422 - 0.1752 * 91 = 28.699$$
$$x = 205, \quad \hat{y} = 44.6422 - 0.1752 * 205 = 8.726$$
$$x = 133, \quad \hat{y} = 44.6422 - 0.1752 * 133 = 21.341$$

- Confidence intervals
  - **Confidence intervals are used to estimate the mean of all values of y for a given x,**
  - Confidence interval gives an estimated range of values which is likely (with probability given by the **confidence level**) to include an unknown population parameter (here mean of all values of y for a given x), the estimated range being calculated from a given set of sample data.

```
d<-read.csv(file="autompg.csv")
hrsp<-d$horsepower
mpg<-d$mpg
model<-lm(mpg ~ hrsp)
new <- data.frame(hrsp = c(91, 205, 133)) # new values
pred.conf <- predict(model, new, interval = "confidence",level=0.95)
```

```
> pred.conf
       fit       lwr      upr
1 28.697351 26.225487 31.16922
2  8.722511  4.845111 12.59991
3 21.338199 19.212352 23.46405
```

fit = the predictied values
lwr = lower bound of the confidence interval
upr = upper bound of the confidence interval

Prediction intervals

- **Prediction intervals are used to estimate the value of a randomly chosen value of y for a given x.**

- For example, it is not unusual for a randomly selected student score to exceed 95%, but is quite unusual for the class average to be so high.

- Smaller variability is associated with the mean of a variable than with a randomly selected value of that variable.
- The prediction interval is always wider than the analogous confidence interval.

```
pred.conf <- predict(model, new, interval = "confidence",level=0.95)
pred.pred <- predict(model, new, interval = "prediction",level=0.95)
```

```
> pred.conf <- predict(model, new, interval = "confidence",l
evel=0.95)
> pred.conf
        fit       lwr      upr
1 28.697351 26.225487 31.16922
2  8.722511  4.845111 12.59991
3 21.338199 19.212352 23.46405
> pred.pred <- predict(model, new, interval = "prediction",l
evel=0.95)
> pred.pred
        fit       lwr      upr
1 28.697351 18.359560 39.03514
2  8.722511 -2.038253 19.48327
3 21.338199 11.077642 31.59876
```

Evaluation of the model using a test set of data. $Pseudo - R^2$

- $Pseudo - R^2 = 1 - \frac{SSE_{test}}{SST_{train}}$

- Assume that we are given the test set containing the following 3 recods:

#	horsepower	mpg
1	91	38
2	205	12
3	133	15

- $SSE_{test} = \sum_i (y_i - \hat{y}_i)^2 = (38 - 28.697)^2 + (12 - 8.772)^2 + (15 - 21.338)^2 = 52.288$
- $SST_{train} = \sum_i (y_i - \bar{y})^2 = 2406.69$
- $Pseudo - R^2 = 1 - 52.288 / 2406.69 = 0.94$

- The best value of the $pseudo - R^2 = 1$;
- If the regression model is worse than the model of estimation by mean, to $pseudo - R^2 < 0$;
- If the regression model is better than the model of estimation by mean, to $pseudo - R^2 > 0$;

Multiple regression

- Estimated regression equation:

$$\hat{y} = b_0 + b_1 x_1 + \ldots + b_m x_m$$

```
d<-read.csv(file="autompg.csv")
hrsp<-d$horsepower
mpg<-d$mpg
accl<-d$acceleration
model<-lm(mpg ~ hrsp+accl)
summary(model)
```

```
Call:
lm(formula = mpg ~ hrsp + accl)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9534 -2.8161 -0.0995  3.5575  9.4991

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 47.90096    8.27930   5.786 1.16e-05 ***
hrsp        -0.18136    0.02463  -7.362 4.11e-07 ***
accl        -0.16136    0.38845  -0.415    0.682
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.925 on 20 degrees of freedom
Multiple R-squared:  0.7984,    Adjusted R-squared:  0.7783
F-statistic: 39.61 on 2 and 20 DF,  p-value: 1.106e-07
```

$$\hat{y} = 47.90 - 0.18x_1 - 0.16x_2,$$

where

x_1 = horsepower

x_2 = acceleration

$\hat{y}$ = estimated value of mpg

***F*-statistic** tells whether the regression as a whole is performing 'better than random'. It is used for testing whether the model outperforms 'noise' as a predictor. F-test can assess multiple coefficients simultaneously.

Hypothesis for F-test:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_m = 0$.

H_a: At least one of the β_i values is different than 0.

H_0 : there is no linear relationship between y and set $\{x_1, x_2, \ldots x_m\}$. If p-value is small, H_0 is rejected.

Dictionary

- Prediction error – błąd oszacowania, wartość resztowa
- Estimated regression equation – oszacowane równanie regresji