Machine Learning for Business

Module 5: Linear regression
Day 3, 9.00 – 12.00

Asst. Prof. Dr. Santitham Prom-on

Department of Computer Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi

Module 5 Overview

- Linear correlation: selecting right variables
- Linear regression concept
- Model building
- Prediction and evaluation

** Hands-on workshop: demand forecast

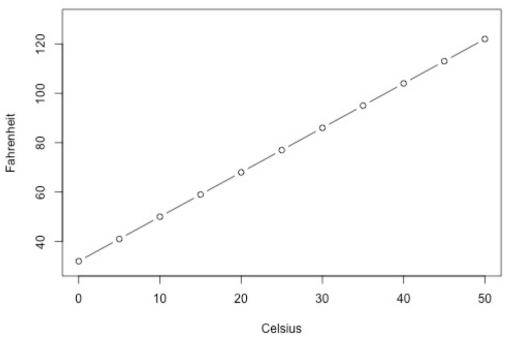
Simple linear regression

Simple linear regression allows us to summarize and study relationships between two continuous (quantitative) variables:

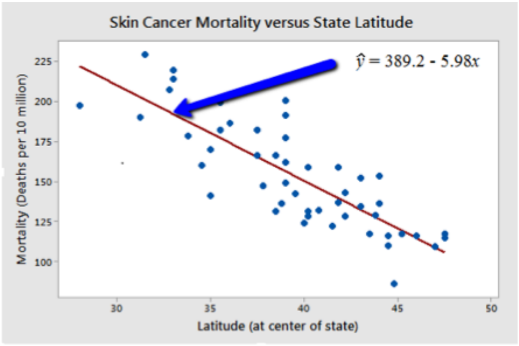
- One variable, denoted x, is regarded as the **predictor**, **explanatory**, or **independent** variable.
- The other variable, denoted y, is regarded as the **response**, **outcome**, or **dependent** variable.

Types of relationship

Deterministic



Statistical



Example of statistical relationships

- Height and weight
- Alcohol consumed and blood alcohol content
- Spending amount and number of spending
- Number of exercises per week and loosing weight
- Credit-line and average credit usage

Assessing statistical linear relationship

Graphical

- Scatter plot

Statistics

- Correlation

Correlation coefficient

- The correlation coefficient computed from the sample data measures **the strength and direction of a linear relationship** between two variables.
- The symbol for the sample correlation coefficient is *r*.
- The symbol for the population correlation coefficient is ρ.
- There are several types of correlation coefficients. The one explained in this section is called the Pearson product moment correlation coefficient (PPMC)

Pearson product moment correlation

- Given *n* pairs of observations (x_1,y_1), (x_2,y_2), $\ldots,(x_n,y_n)$
  - It is natural to speak of *x* and *y* having a positive relationship if large *x*'s are paired with large *y*'s and small *x*'s with small *y*'s
  - On the contrary, if large *x*'s are paired with small *y*'s and small *x*'s with large *y*'s, then a negative relationship between the variable is implied

Pearson product moment correlation

- Consider the quantity

$$s_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- Then, if the relationship is strongly positive, an x_{i} above the mean will tend to be paired with a y_{i} above the mean, so that
  and this product will also be positive whenever both x_{i} and y_{i} are below their means

$$(x_i - \bar{x})(y_i - \bar{y}) > 0$$

Pearson product moment correlation

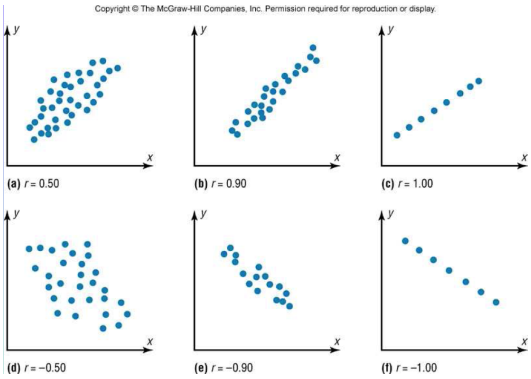
- To make this measure dimensionless, we divide as follow

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

- A more convenient form for this equation is

$$r = \frac{n\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}\sqrt{n\left(\sum y_i^2\right) - \left(\sum y_i\right)^2}}$$

Correlation coefficient and scatter plot



(a) $r = 0.50$ (b) $r = 0.90$ (c) $r = 1.00$

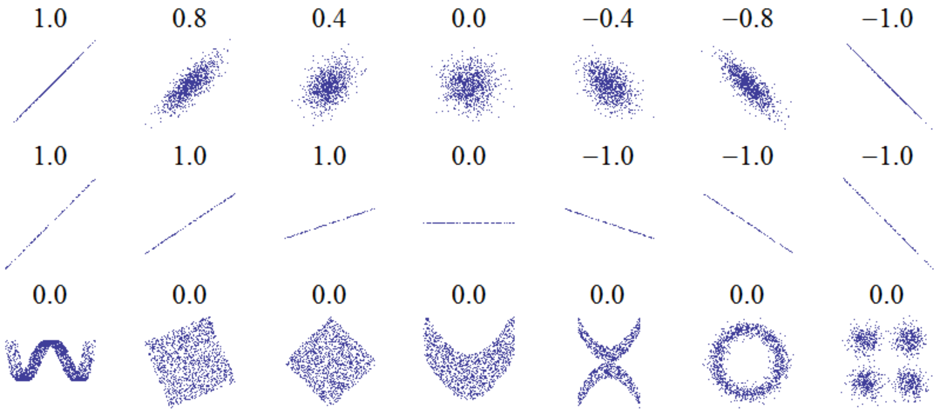
(d) $r = -0.50$ (e) $r = -0.90$ (f) $r = -1.00$

- The range of the correlation coefficient is from -1 to 1.
- If there is a strong positive linear relationship between the variables, the value of *r* will be close to 1.
- If there is a strong negative linear relationship, the value of *r* will be close to -1.
- When there is no linear relationship between the variables or only a weak one, the value of *r* will be close to 0

Strong correlation?

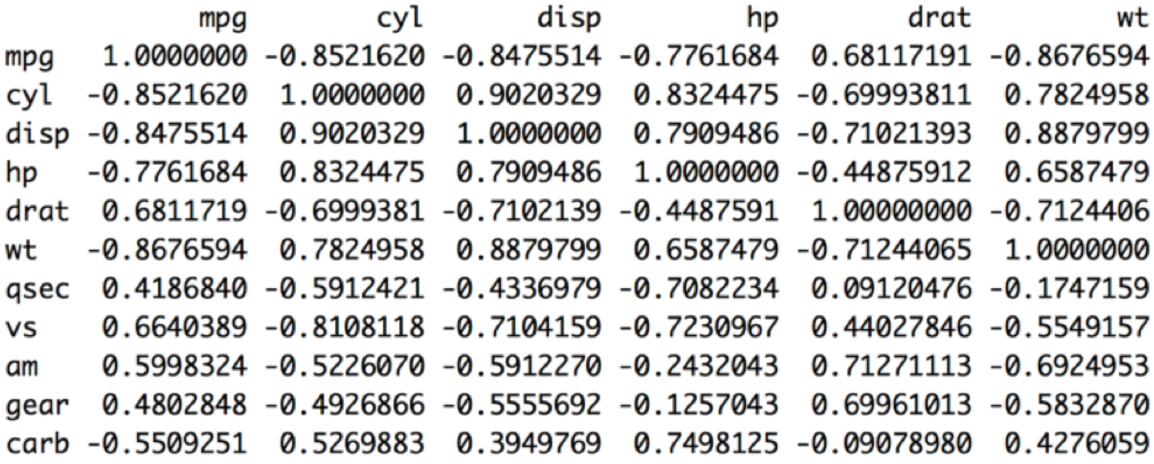
- A frequently asked question is: “what can it be said that there is a strong correlation between variables, and when is the correlation weak?”
- A reasonable rule of thumb is to say that the correlation is
  - weak if 0≤|r| ≤0.5
  - strong 0.8≤|r|≤1
  - moderate otherwise

Correlation and shape



R: correlation
Assessing for every pair

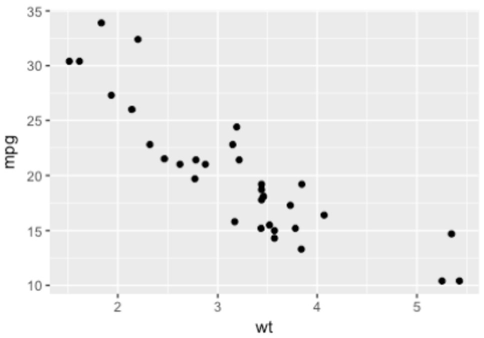
```
cor(mtcars)
```



```
            mpg        cyl       disp         hp        drat         wt
mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
```

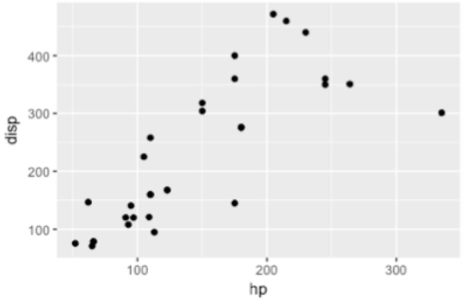
R: correlation and scatter plot
Negative relationship, mpg vs wt

```
ggplot(data = mtcars,
       mapping  = aes(x = wt,
                      y = mpg)) +
  geom_point()

cor(mtcars$mpg, mtcars$wt)
[1] -0.8676594
```



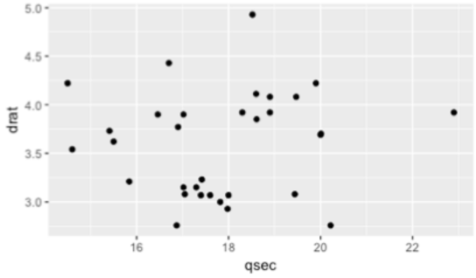
R: correlation and scatter plot
Positive relationship, hp vs disp

```
ggplot(data = mtcars,
       mapping  = aes(x = hp,
                      y = disp)) +
  geom_point()

cor(mtcars$hp, mtcars$disp)
[1] 0.7909486
```



R: correlation and scatter plot
No relationship, qsec vs drat

```
ggplot(data = mtcars,
       mapping  = aes(x = qsec,
                      y = drat)) +
  geom_point()

cor(mtcars$qsec, mtcars$drat)
[1] 0.09120476
```



Scatter and trend

- Since we are interested in summarizing the trend between two quantitative variables, the natural question arises — "what is the best fitting line?"
- Scatter plot shows points distribution and potentially the trend in the data
- Look at the figure in next pages, which lines do you think best summarizes the trend between height and weight?

Line equation

- y_i denotes the observed response for experimental unit i
- x_i denotes the predictor value for experimental unit i
- $\hat{y}_i$ is the predicted response (or fitted value) for experimental unit i
- The equation for best fitting line is:

$$\hat{y}_i = b_0 + b_1 x_i$$

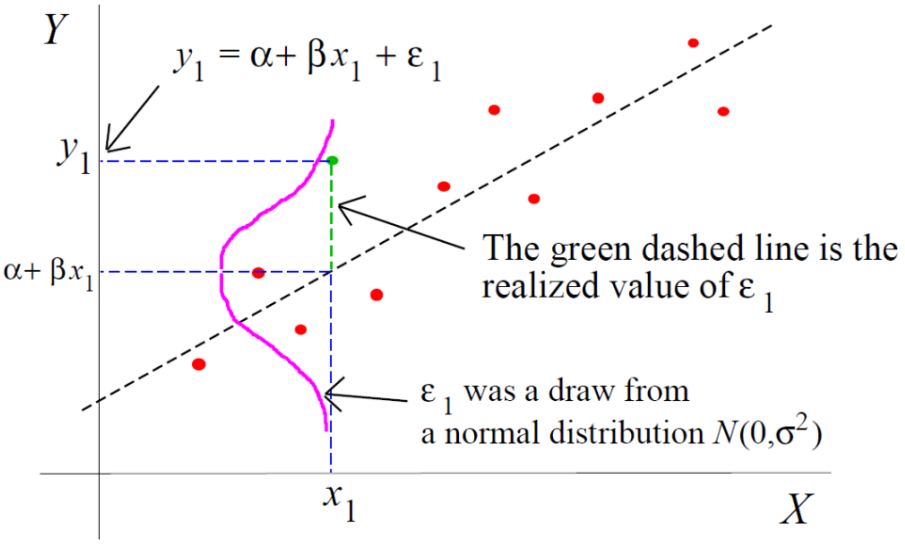
Regression parameters

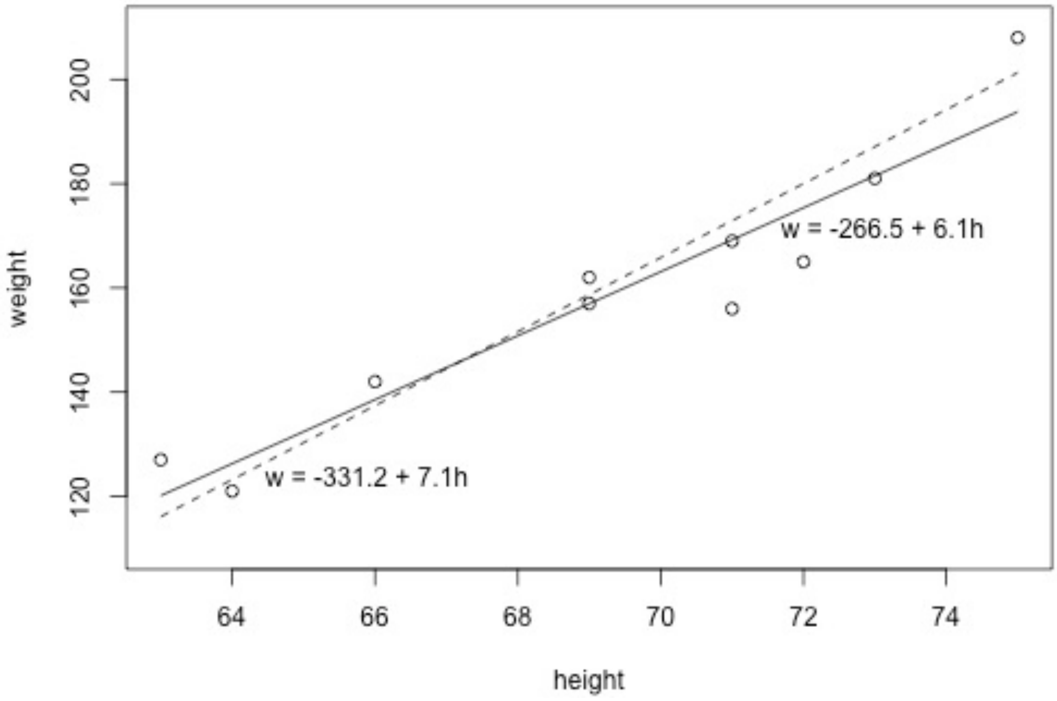
- Given a value x_i, how do we interpret b_0 and b_1.

b_1 tells us: if the value we saw for x was one unit bigger, how much would our prediction for y changes?

b_0 tells us: what would we predict for y if $x = 0$?

Residual must be normally distributed with zero mean

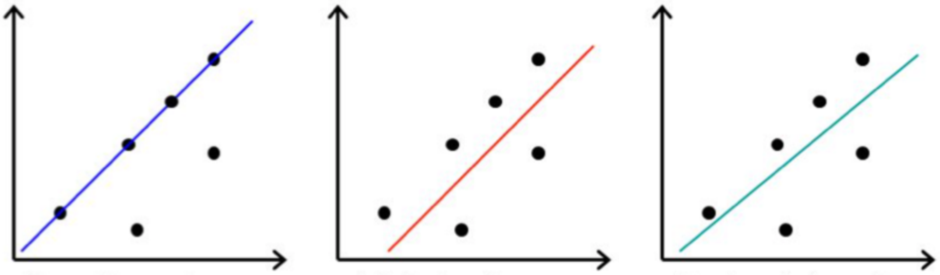

Y
$y_1 = \alpha + \beta x_1 + \varepsilon_1$
y_1
$\alpha + \beta x_1$
The green dashed line is the realized value of ε_1
ε_1 was a draw from a normal distribution $N(0,\sigma^2)$
x_1
X



Best line

Objective: Want to fit the "best" line to the data points (that exhibit linear relation).

- How do we define "best"?



Pass through as many points as possible

Minimize the maximum residual of each point

Each point carries the same weight

Error

- In general, when we use $\hat{y}_i = b_0 + b_1 x_i$ to predict the actual response y_i, we make a prediction error (or residual error) of size:

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

- A line that fits the data "**best**" will be one for which the ***n*** **prediction errors** — one for each observed data point — **are as small as possible in some overall sense**.

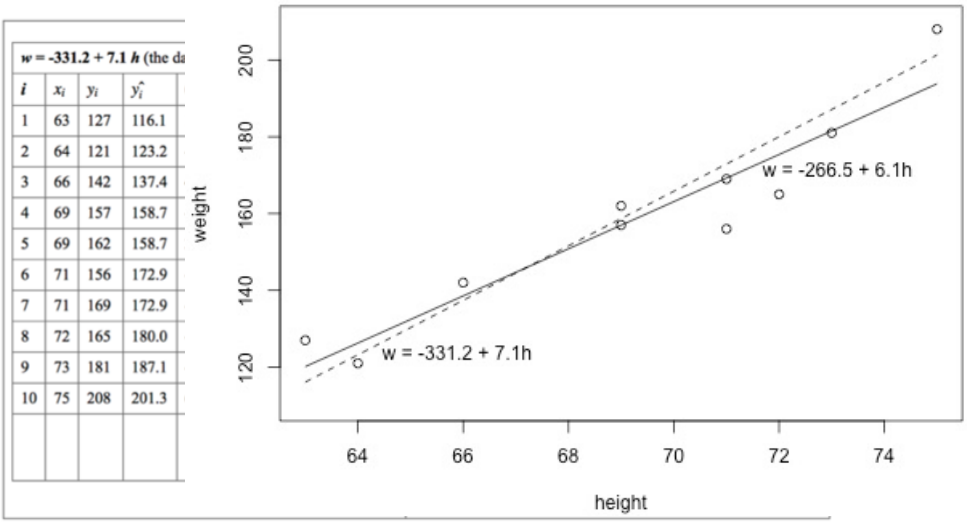
Method of least squares

- Choose the b's so that the sum of the squares of the errors, e_i, are minimized
- The error function is

$$S = \sum_{i=1}^{n} e_i$$

$$= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

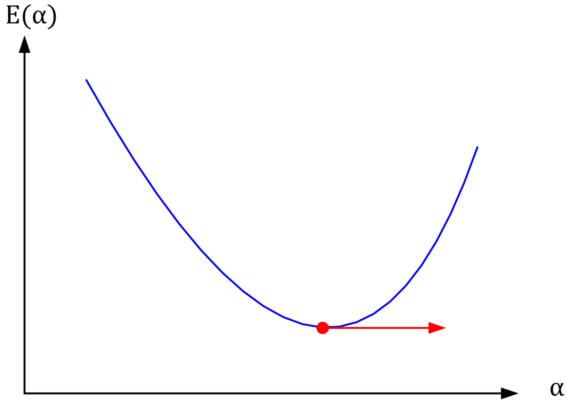
Error comparison



w = -331.2 + 7.1 h (the da			
i	x_i	y_i	$\hat{y_i}$
1	63	127	116.1
2	64	121	123.2
3	66	142	137.4
4	69	157	158.7
5	69	162	158.7
6	71	156	172.9
7	71	169	172.9
8	72	165	180.0
9	73	181	187.1
10	75	208	201.3

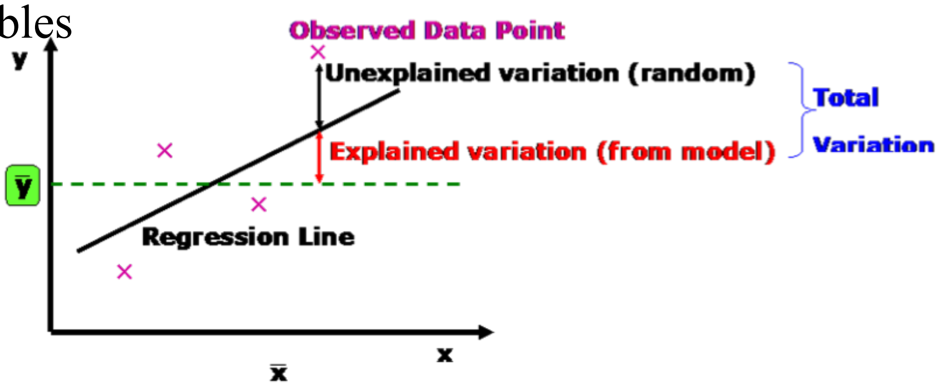
Ordinary least square solution

Minimum of a function is the point where the slope is zero



Coefficient of determination (r^2)

The coefficient of determination is a number that indicates the proportion of the variance in the dependent variable that is predictable from the independent variables



Coefficient of determination

- The coefficient of determination R^2 (or sometimes r^2) is another measure of how well the least squares equation

$$Y = \alpha + \beta X$$

perform as a predictor of y

- R^2 is computed as:

$$R^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{SS_{yy}}{SS_{yy}} - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

- R^2 measures the relative sizes of SS_{yy} and SSE.
- The smaller SSE, the more reliable the predictions obtained from the model.

Coefficient of determination

- SS_{yy} measures the deviation of the observations from their mean:

$$SS_{yy} = \sum_i \left(y_i - \bar{y}\right)^2$$

- SSE measures the deviation of observations from their predicted values

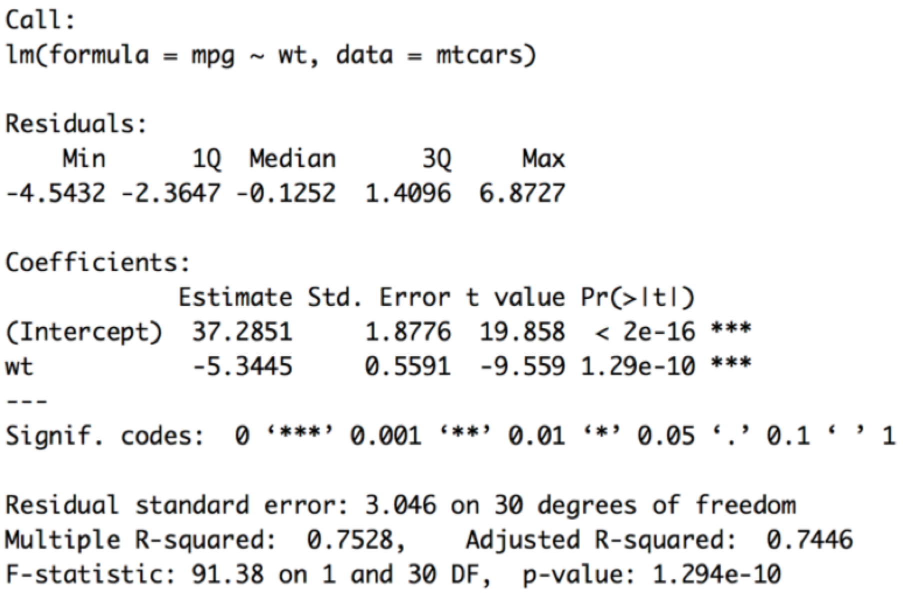
$$SSE = \sum_i \left(y_i - Y_i\right)^2$$

Coefficient of determination

- The higher the R^2, the more useful the model
- R^2 takes on values between 0 and 1
- Essentially, R^2 tells us how much better we can do in predicting y by using the model and computing Y than by just using the mean of y as a predictor.
- Note that when we use the model and compute Y the prediction depends on X because $Y = \alpha + \beta X$.
- Thus, we act as if x contains information about y.
- If we just use the mean of y to predict y, then we are saying that x does not contribute information about y and thus our predictions of y do not depend on x.

R: simple linear regression

```
model <- lm(mpg ~ wt, mtcars)
summary(model)
```



```
Call:
lm(formula = mpg ~ wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5432 -2.3647 -0.1252  1.4096  6.8727 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,	Adjusted R-squared:  0.7446 
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Coefficients

- **Estimates:** values of the coefficients
- **Standard errors:** This measures the average amount that the coefficient estimates vary from the actual average value of our response variable
- **t-value:** The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0.
- ***Pr(>t)*****:** This indicates whether the probability that the impact of the parameter is due to chance.

Residual and R^2

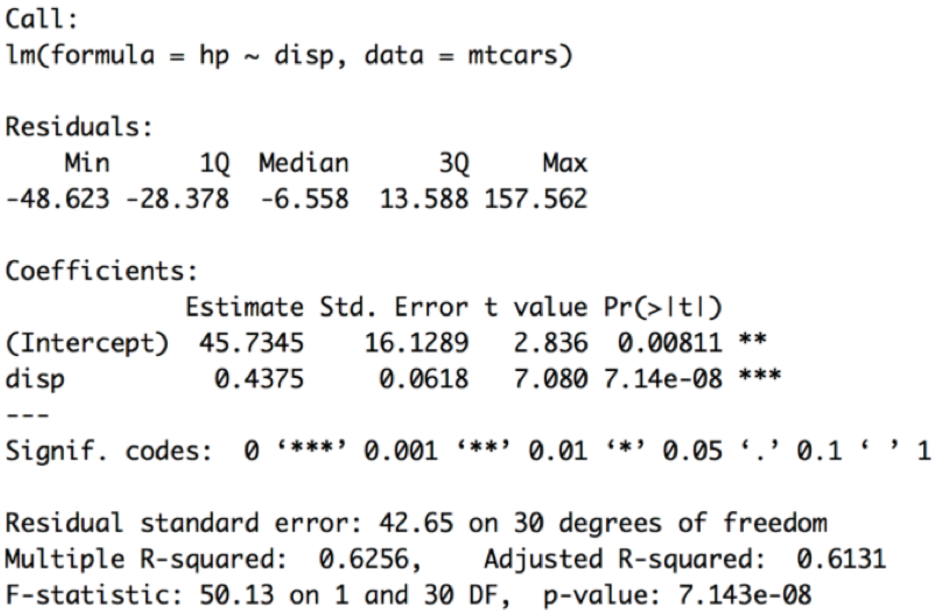
- The Residual Standard Error is the average amount that the response (dist) will deviate from the true regression line.
- In multiple regression settings, the R^2 will always increase as more variables are included in the model.
- That’s why the adjusted R^2 is the preferred measure as it adjusts for the number of variables considered.

F-statistics

- F-statistic is a good indicator of whether there is a relationship between our predictor and the response variables.
- The further the F-statistic is from 1 the better it is.

R: simple linear regression

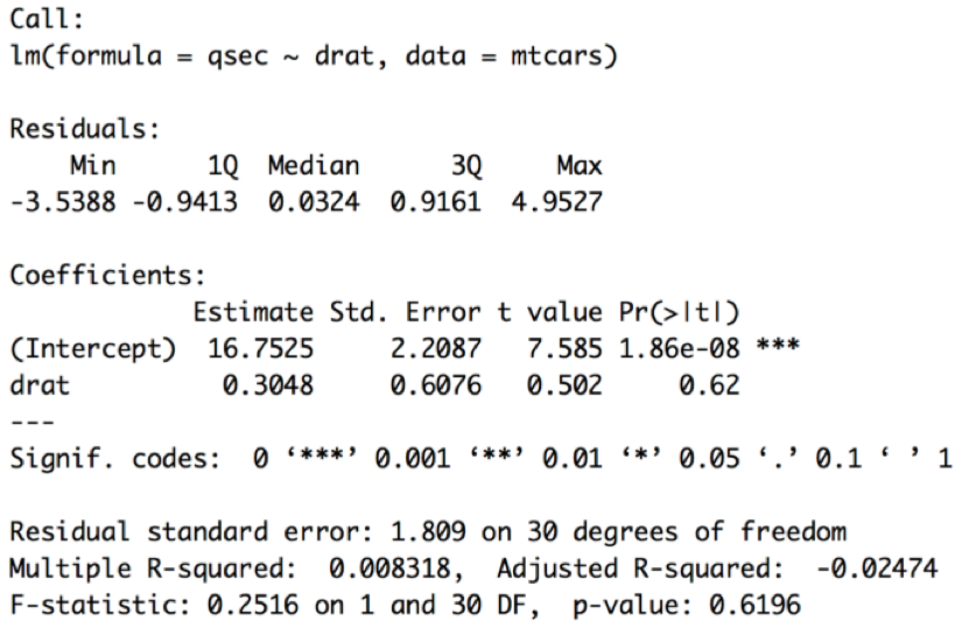
```
model1 <- lm(hp ~ disp, mtcars)
summary(model1)
```



```
Call:
lm(formula = hp ~ disp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.623 -28.378  -6.558  13.588 157.562 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  45.7345    16.1289   2.836  0.00811 ** 
disp          0.4375     0.0618   7.080 7.14e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42.65 on 30 degrees of freedom
Multiple R-squared:  0.6256,	Adjusted R-squared:  0.6131 
F-statistic: 50.13 on 1 and 30 DF,  p-value: 7.143e-08
```

R: simple linear regression

```
model2 <- lm(qsec ~ drat, mtcars)
summary(model2)
```



```
Call:
lm(formula = qsec ~ drat, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5388 -0.9413  0.0324  0.9161  4.9527 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  16.7525     2.2087   7.585 1.86e-08 ***
drat          0.3048     0.6076   0.502     0.62    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.809 on 30 degrees of freedom
Multiple R-squared:  0.008318,	Adjusted R-squared:  -0.02474 
F-statistic: 0.2516 on 1 and 30 DF,  p-value: 0.6196
```

Multiple linear regression

- We move from the simple linear regression model with one predictor to the multiple linear regression model with two or more predictors.
- That is, we use the adjective "simple" to denote that our model has only predictor, and we use the adjective "multiple" to indicate that our model has at least two predictors

Multiple regression
Additive model, no interaction

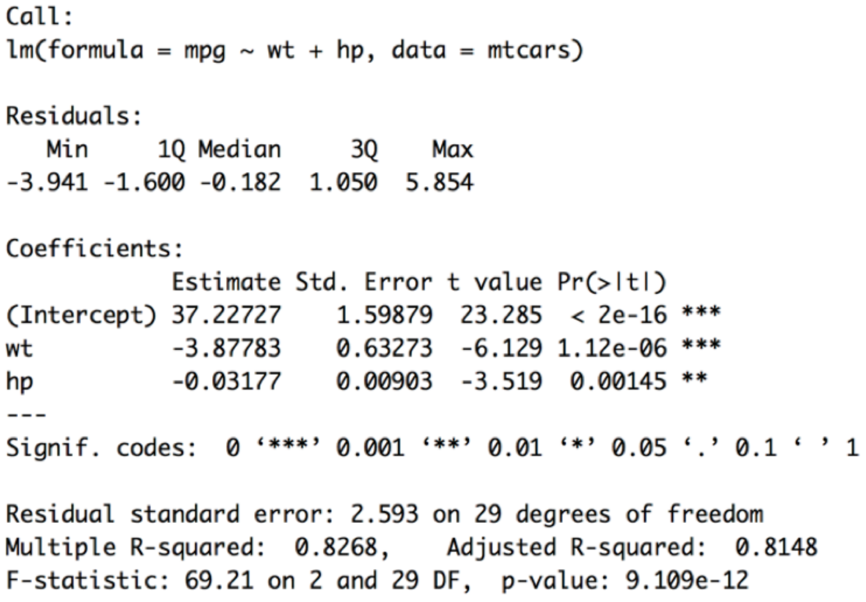
- Multiple linear regression model structure is exactly the same as the linear regression

$$f\left(\mathbf{x}\right) = w_0 + w_1 x_1 + w_2 x_2 + \cdots$$

- Mathematically, parameters are obtained by least square method

R: multiple regression additive model

```
model3 <- lm(mpg ~ wt + hp, mtcars)
summary(model3)
```

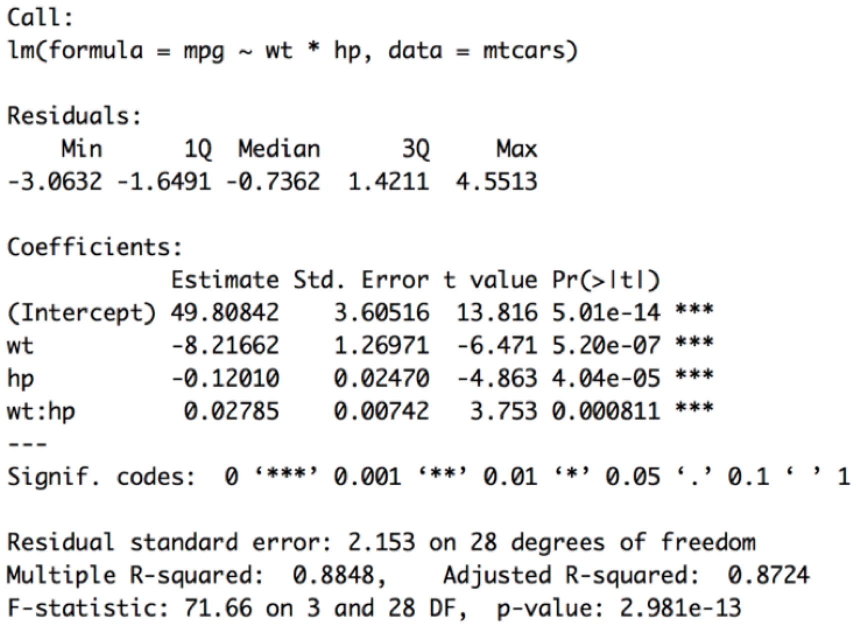


```
Call:
lm(formula = mpg ~ wt + hp, data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max 
-3.941 -1.600 -0.182  1.050  5.854 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
hp          -0.03177    0.00903  -3.519  0.00145 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,	Adjusted R-squared:  0.8148 
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

Interaction in multiple regression

- Adding interaction terms to a regression model can greatly expand understanding of the relationships among the variables in the model
- This occurs when two or more variables depend on one another for the outcome
- For example, drug and alcohol may interact and creates addition (adverse) affect
- Another example, credit card types and number of active month may interact (depend) and have different spending results

```
model4 <- lm(mpg ~ wt * hp, mtcars)
summary(model4)
```



```
Call:
lm(formula = mpg ~ wt * hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0632 -1.6491 -0.7362  1.4211  4.5513 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 49.80842    3.60516  13.816 5.01e-14 ***
wt          -8.21662    1.26971  -6.471 5.20e-07 ***
hp          -0.12010    0.02470  -4.863 4.04e-05 ***
wt:hp        0.02785    0.00742   3.753 0.000811 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.153 on 28 degrees of freedom
Multiple R-squared:  0.8848,	Adjusted R-squared:  0.8724 
F-statistic: 71.66 on 3 and 28 DF,  p-value: 2.981e-13
```

R: multiple regression multiplicative model (interaction)

Activity
Daily demand forecast

- Given the dataset of daily demand forecast
  - The dataset was collected during 60 days, this is a real database of a Brazilian logistics company.
  - The dataset has twelve predictors and a target that is the total of orders for daily treatment.
- Experiment and create the best regression model for predicting daily
- Why is sometime adding predictors do not help prediction?

Thank you

Question?