QUANTITATIVE RESEARCH METHODS

DR. MEIKE MORREN

Lecture 5

contents

- Linear regression
  - Deterministic vs Probabilistic
  - Simple regression
  - T-test

- Sums of squares (ANOVA approach)
- Ordinary least squares

- Plotting the line

LINEAR REGRESSION

Simple linear regression

- Straight line
- Y is called the response/dependent variable
- x is called the predictor or independent variable (sometimes explanatory)
- The model is written as:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

R plot mtcars

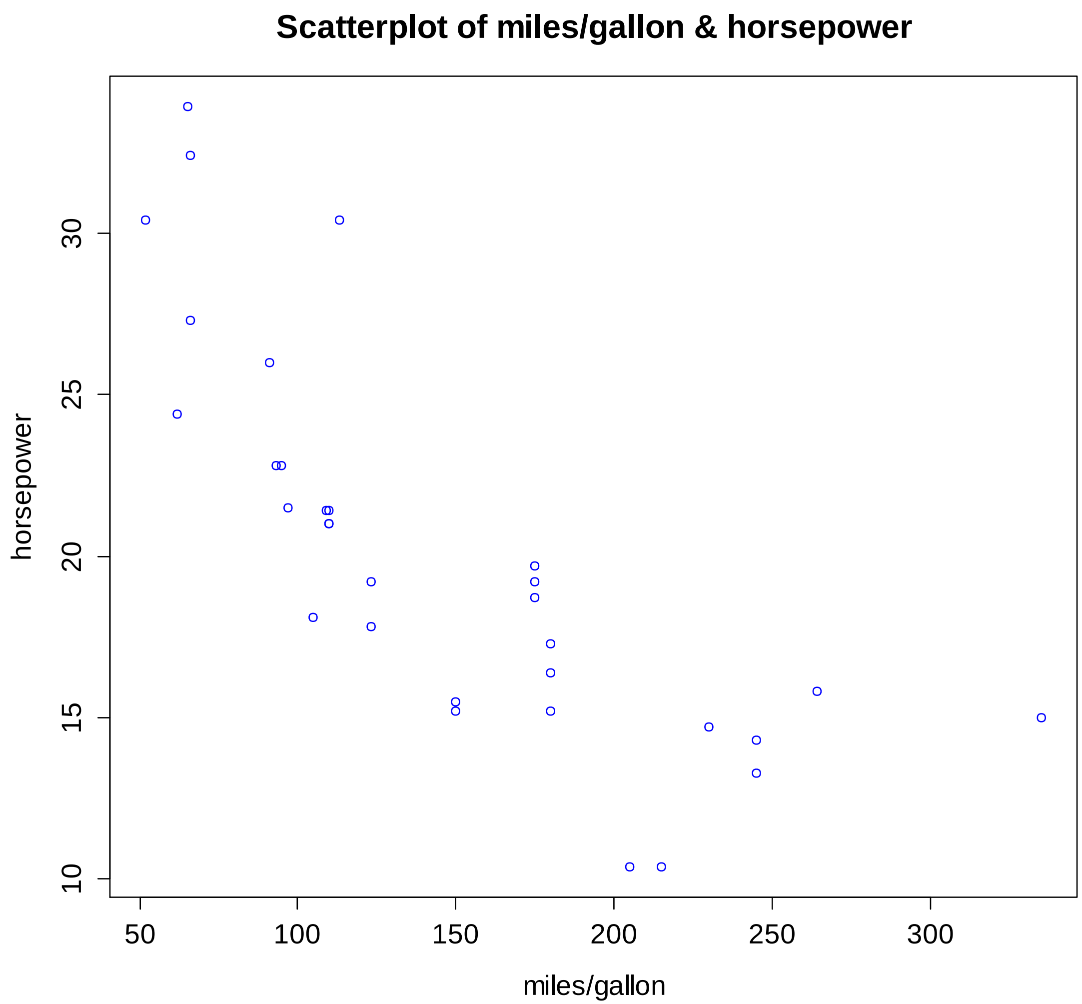
- Y <- mtcars$mpg
- x <- mtcars$hp
- plot(x. Y. col="blue".

  main="Scatterplot of miles/gallon & + horsepower".

  xlab="miles/gallon".

  ylab="horsepower")

Scatterplot



Deterministic vs probabilistic

- Deterministic

$$Y_i = \beta_0 + \beta_1 x_i$$

- Probabilistic

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Estimation of parameters

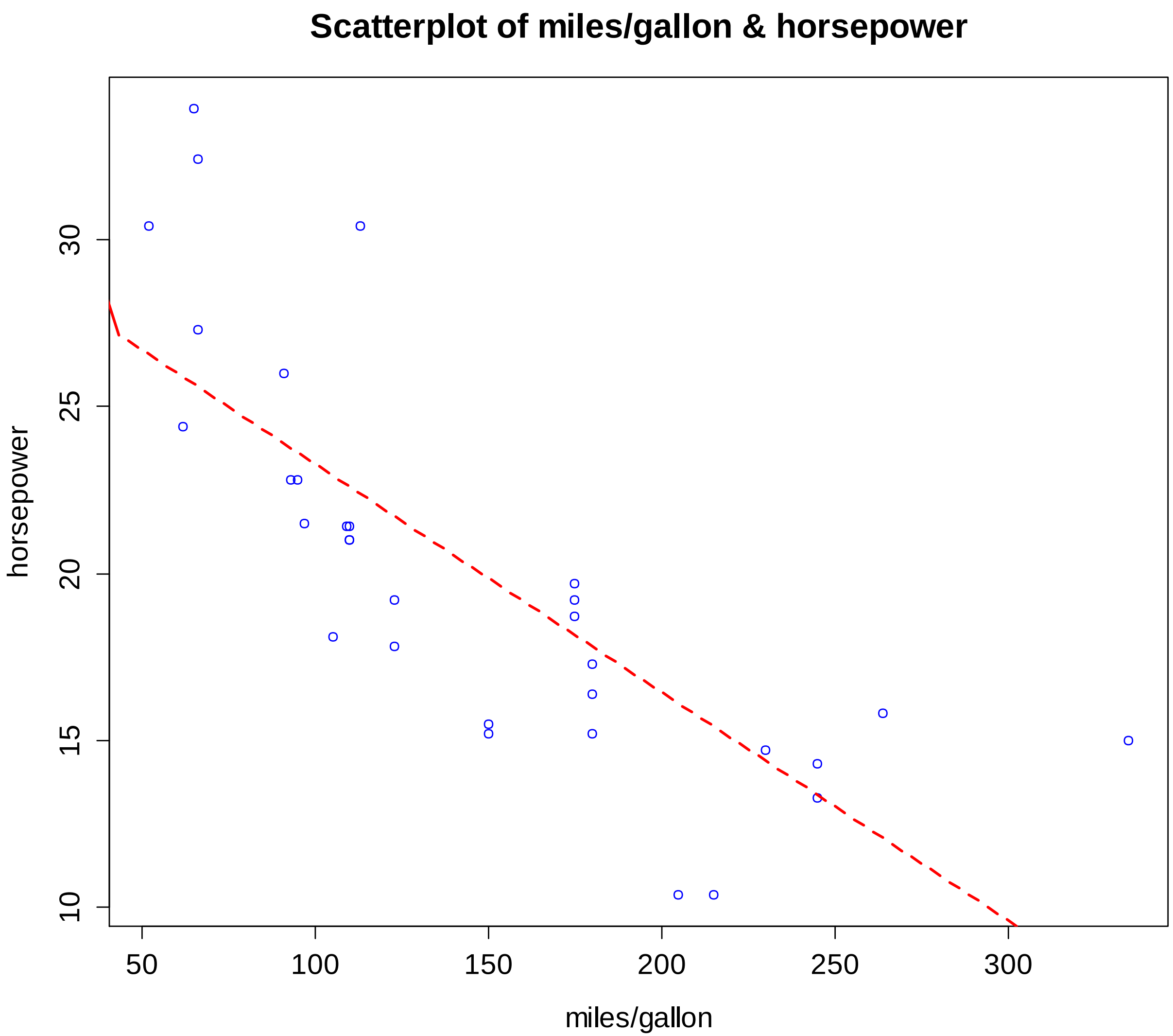
$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Using the expected value which is the mean here and can also be written as:

$$\bar{y} = E(y) = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Ad line (first estimate parameters)



Equation (1)

$$Y_i = b_0 + b_1 x_1 + \varepsilon_i$$

An increase of 1pt in X1 leads to increase of b1 in Y

Assess fit

Y	X	Y predicted	Error	Error squared
21	110	22.59	-1.59	2.54
22.8	110	22.59	-1.59	2.54
21.4	931	23.75	-.95	.91
18.7	110	22.59	-1.19	1.43
18.1	175	18.16	.54	.29
14.3	105	22.93	-4.83	23.38
24.4	245	13.38	.92	.84
22.8	62	25.87	-1.47	2.16
19.2	95	23.62	-.82	.67

Assess fit

- Calculate predicted values using the parameters
- Find the errors (= difference between predicted and actual values)
- Sum all squared errors

Model fit (1)

SSE = sum of squared errors

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

SST = sum of squares (total variation)

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

SSR = sum of squares regression (explained variation)

$$SSR = \sum_{i=1}^{n} (\widehat{Y}_i - \bar{Y})^2$$

Model fit (2)

SST = SSE + SSR

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

R^2 = 1- SSE/SST

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

R^2 adjusted =

1- (SSE/(n-k)) / (SST/(n-1))

$$SSR = \sum_{i=1}^{n} (\widehat{Y}_i - \bar{Y})^2$$

T-TEST, DUMMY VARIABLES

Compare t-test with regression model with dummy variable

- Compare two countries
- No exercise!

Equation

$$Y_i = b_0 + b_1 x_1 + \varepsilon_i$$

The value of Y where the X1 variable is zero

Exercise 5_1.r

Use the WVS dataset

- *Relate happiness to income*

a) Write vectorised function of linear regression using equations of b0 and b1 (see lecture)
b) Calculate the R squared (optional)
c) Check with lm function
d) Estimate a model with dummy variable
e) Check with t.test function

MULTIPLE REGRESSION

Multiple vs simple regression

- Estimation becomes more complicated when multiple explanatory variables are included

- A general method would be the least squares (ordinary least squares – OLS) where one obtains the regression coefficients by minimizing the errors

- In order to compute the coefficients. we need to use derivations

OLS (1)

- Minimizing the sum of the squared deviations of the Y_i's
- This minimized solution provides reliable and stable estimates of β_n
- The estimated regression function is written

$$\widehat{Y}_i = b_0 + b_1 x_1 + \cdots + b_n x_n + \varepsilon_i$$

- Another way to model this the relationship is

$$f_\theta(x) = \theta_1 x_1 + \cdots + \theta_n x_n$$

OLS(2)

- We want to minimize the least-squares cost function

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(f_\theta\,(x^{(i)}) - y^{(i)})^2$$

Where $x^{(i)}$ is the *i*th observation and

$y^{(i)}$ is the *i*th expected result

OLS (3)

- We can rewrite this loss function J as

$$f_\theta(x) = \theta^T x$$

- With this we can rewrite the least-squares cost function using matrix multiplication

$$J(\theta_{0..n}) = \frac{1}{2m}(X\theta - y)^T(X\theta - y)$$

Derivatives

$$\partial J / \partial \theta = 2X^T X\theta - 2X^T y = 0$$

$$X^T X\theta = X^T y$$

If the matrix $X^T X$ is invertible. we can multiply both sides by $(X^T X)^{-1}$ and get

$$\theta = (X^T X)^{-1} X^T y$$

OLS estimation: step by step

- This formula can be used to estimate multiple regression coefficients

1. Combine all *k* independent variables in columns in a matrix (n x k)
2. Add a vector of 1s to estimate the intercept
3. Make a vector of the dependent variable
4. Solve the formula

Example mtcars

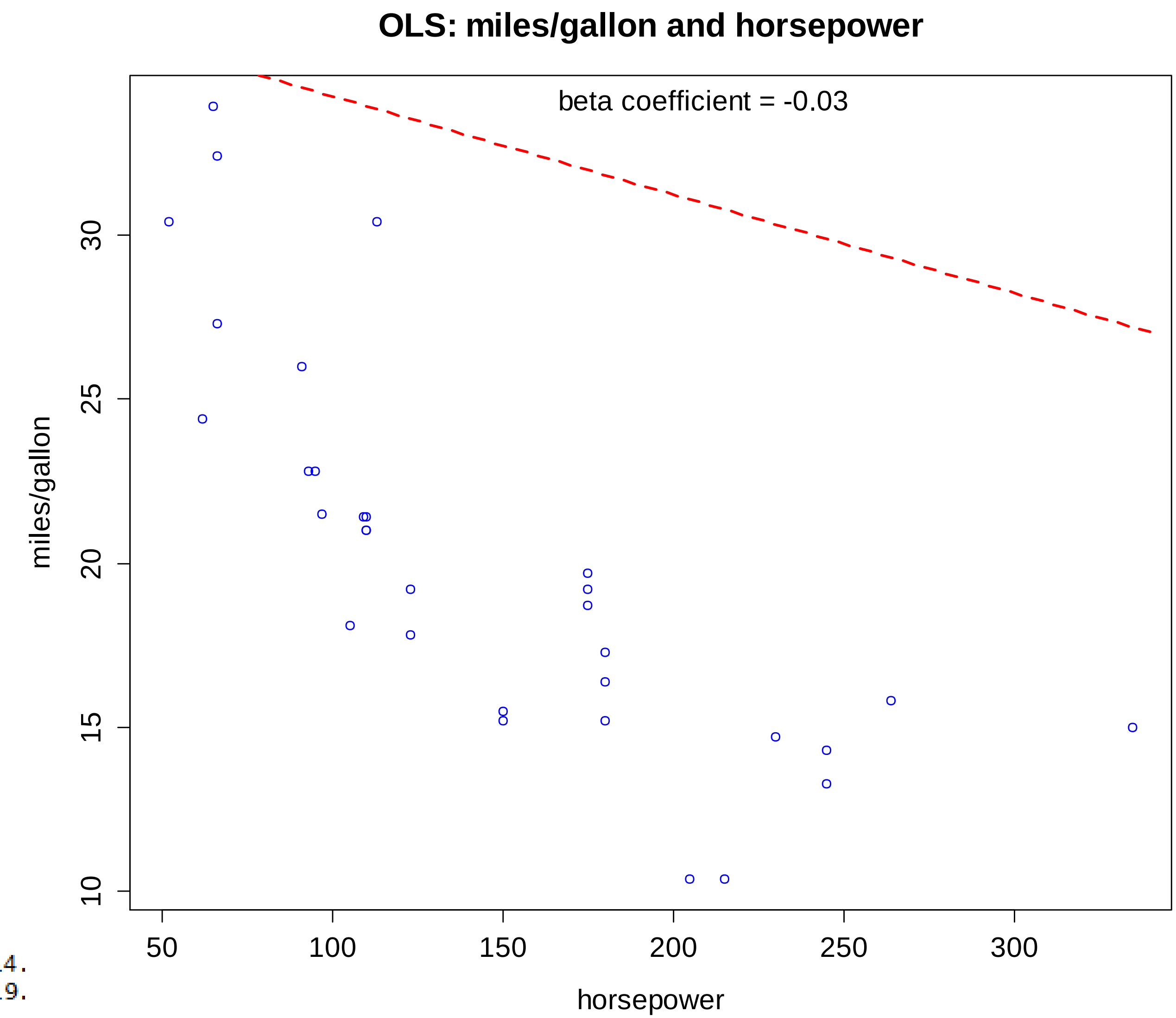
- Include both horsepower and weight
- Estimate linear regression using the formula
- Plot two variables
- Add regression line using the coefficients that you found

Y and X

```
> x
      [,1] [,2]  [,3]
 [1,]    1  110 2.620
 [2,]    1  110 2.875
 [3,]    1   93 2.320
 [4,]    1  110 3.215
 [5,]    1  175 3.440
 [6,]    1  105 3.460
 [7,]    1  245 3.570
 [8,]    1   62 3.190
 [9,]    1   95 3.150
[10,]    1  123 3.440
[11,]    1  123 3.440
[12,]    1  180 4.070
[13,]    1  180 3.730
[14,]    1  180 3.780
[15,]    1  205 5.250
[16,]    1  215 5.424
[17,]    1  230 5.345
[18,]    1   66 2.200
[19,]    1   52 1.615
[20,]    1   65 1.835
[21,]    1   97 2.465
[22,]    1  150 3.520
[23,]    1  150 3.435
[24,]    1  245 3.840
[25,]    1  175 3.845
[26,]    1   66 1.935
[27,]    1   91 2.140
[28,]    1  113 1.513
[29,]    1  264 3.170
[30,]    1  175 2.770
[31,]    1  335 3.570
[32,]    1  109 2.780
> y
 [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.
[24] 13.3 19.2 27.3 26.0 30.4 15.8 19.7 15.0 21.4
>
```

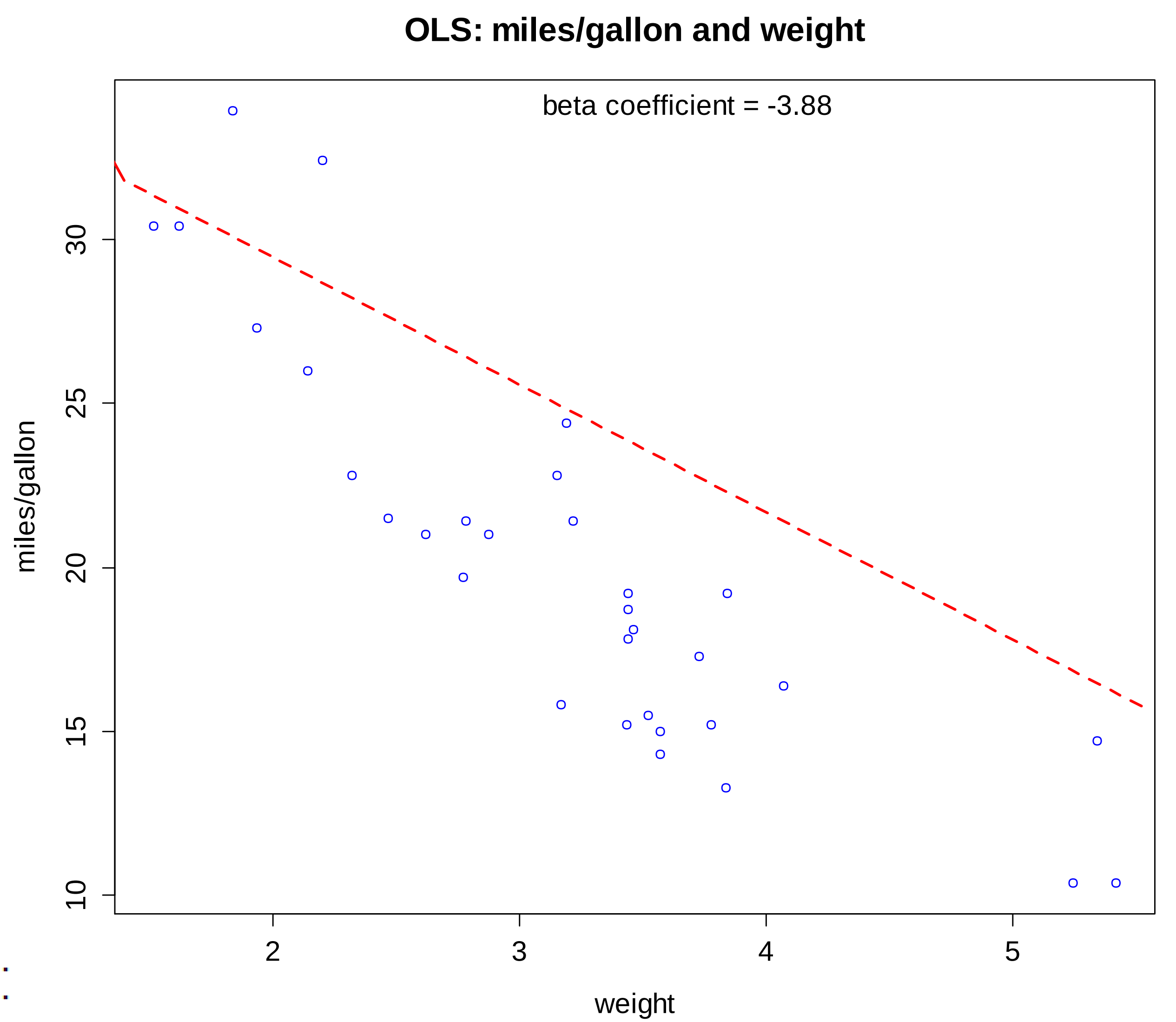
Scatterplot of y and X1

```
> x
      [,1] [,2]  [,3]
 [1,]    1  110 2.620
 [2,]    1  110 2.875
 [3,]    1   93 2.320
 [4,]    1  110 3.215
 [5,]    1  175 3.440
 [6,]    1  105 3.460
 [7,]    1  245 3.570
 [8,]    1   62 3.190
 [9,]    1   95 3.150
[10,]    1  123 3.440
[11,]    1  123 3.440
[12,]    1  180 4.070
[13,]    1  180 3.730
[14,]    1  180 3.780
[15,]    1  205 5.250
[16,]    1  215 5.424
[17,]    1  230 5.345
[18,]    1   66 2.200
[19,]    1   52 1.615
[20,]    1   65 1.835
[21,]    1   97 2.465
[22,]    1  150 3.520
[23,]    1  150 3.435
[24,]    1  245 3.840
[25,]    1  175 3.845
[26,]    1   66 1.935
[27,]    1   91 2.140
[28,]    1  113 1.513
[29,]    1  264 3.170
[30,]    1  175 2.770
[31,]    1  335 3.570
[32,]    1  109 2.780
> y
 [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.
[24] 13.3 19.2 27.3 26.0 30.4 15.8 19.
>
```



Scatterplot of y and X2

```
> X
      [,1] [,2]  [,3]
 [1,]    1  110 2.620
 [2,]    1  110 2.875
 [3,]    1   93 2.320
 [4,]    1  110 3.215
 [5,]    1  175 3.440
 [6,]    1  105 3.460
 [7,]    1  245 3.570
 [8,]    1   62 3.190
 [9,]    1   95 3.150
[10,]    1  123 3.440
[11,]    1  123 3.440
[12,]    1  180 4.070
[13,]    1  180 3.730
[14,]    1  180 3.780
[15,]    1  205 5.250
[16,]    1  215 5.424
[17,]    1  230 5.345
[18,]    1   66 2.200
[19,]    1   52 1.615
[20,]    1   65 1.835
[21,]    1   97 2.465
[22,]    1  150 3.520
[23,]    1  150 3.435
[24,]    1  245 3.840
[25,]    1  175 3.845
[26,]    1   66 1.935
[27,]    1   91 2.140
[28,]    1  113 1.513
[29,]    1  264 3.170
[30,]    1  175 2.770
[31,]    1  335 3.570
[32,]    1  109 2.780
> y
 [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.
[24] 13.3 19.2 27.3 26.0 30.4 15.8 19.
>
```



Equation (1)

$$Y_i = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon_i$$

An increase of 1pt in X1 leads to increase of b1 in Y

Interpretation beta coefficients:

- A 1-point increase in weight (measured in 1000lbs) leads to a 3.88 decrease in miles per gallon
- Thus, the heavier the car, the fewer miles you can drive with a gallon of gasoline

- Controlling for horsepower:
- This effect holds for all values of horsepower. So irrespective of how fast the car can drive, an 1pt increase in weight will always lead to a 3.88 decrease in miles per gallon

Equation (2)

$$Y_i = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon_i$$

= Constant

= Intercept

Interpretation constant :

The mean level of the dependent variable where ALL the independent variables are 0

Thus…

The mean level of miles per gallon for

0 weight and 0 horsepower

MULTIPLE REGRESSION

Standardized coefficients

Standardized regression

- If you have multiple variables that have a different range of values, the unstandardized coefficients are hard to compare in terms of strength
- A 1-point increase in one variable means something else than a 1-point increase in another variable

- Therefore, ALL variables are standardized
- The 1-point increase becomes a 1-standard deviation increase

Z-scores (centered and standardized)

- Calculate z-scores:

$$z_i = \frac{x_i - \bar{X}}{\sigma_x}$$

1. Calculate mean
2. Calculate standard deviation (sd)
3. Calculate z scores

Z-scores have a

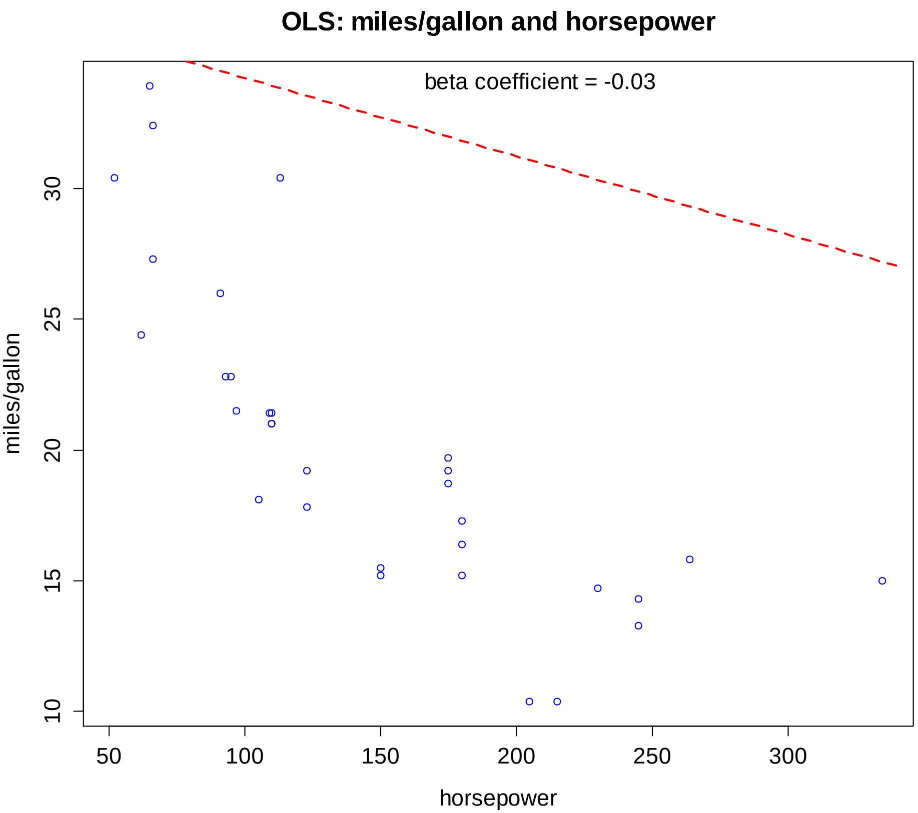
mean of zero and standard deviation of 1

Example mtcars

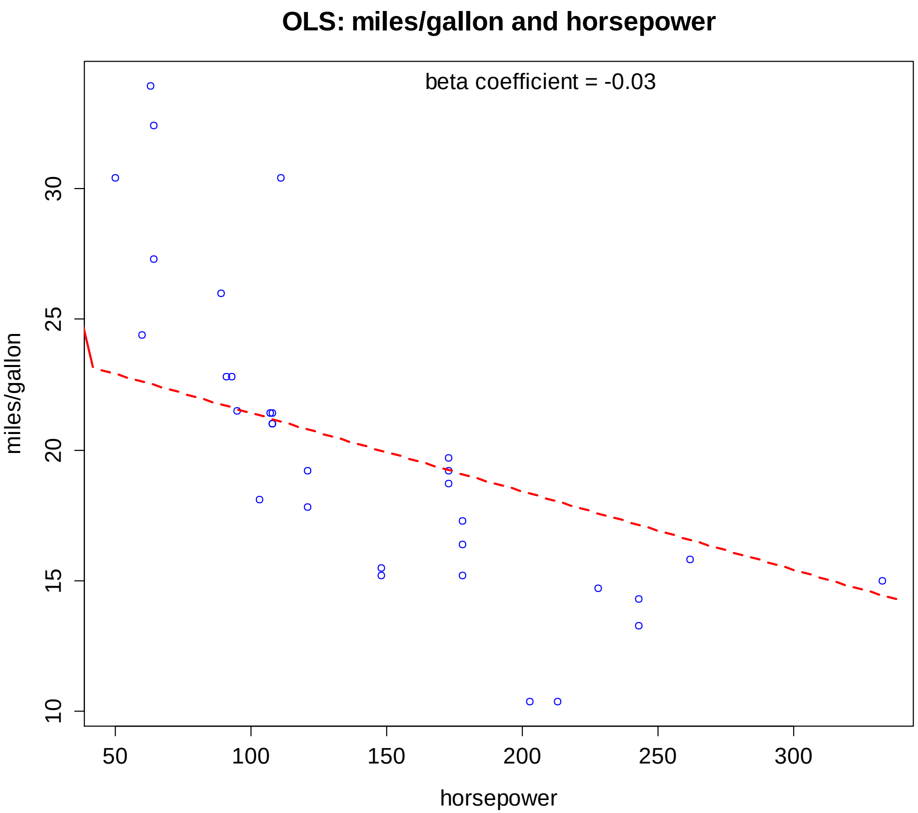
- First calculate z-scores (for loop)
- Create X vectors of standardized scores
- Solve equation with standardized values
- Plot the line
- Compare with previous results

Example mtcars: X1

Unstandardized

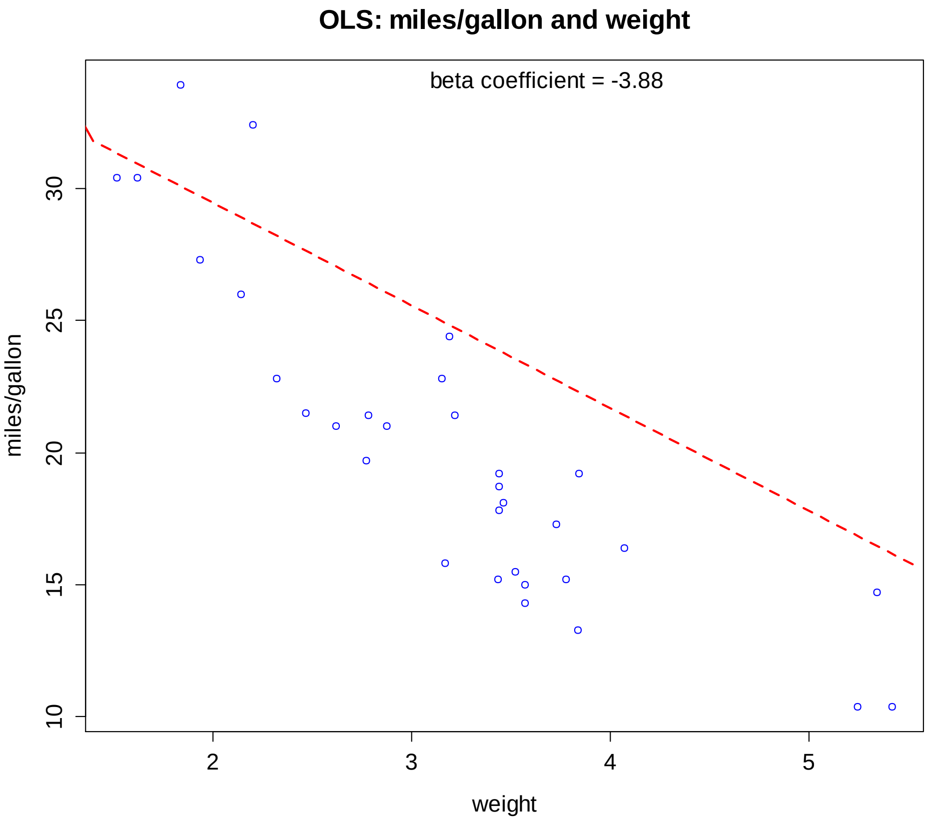


Standardized

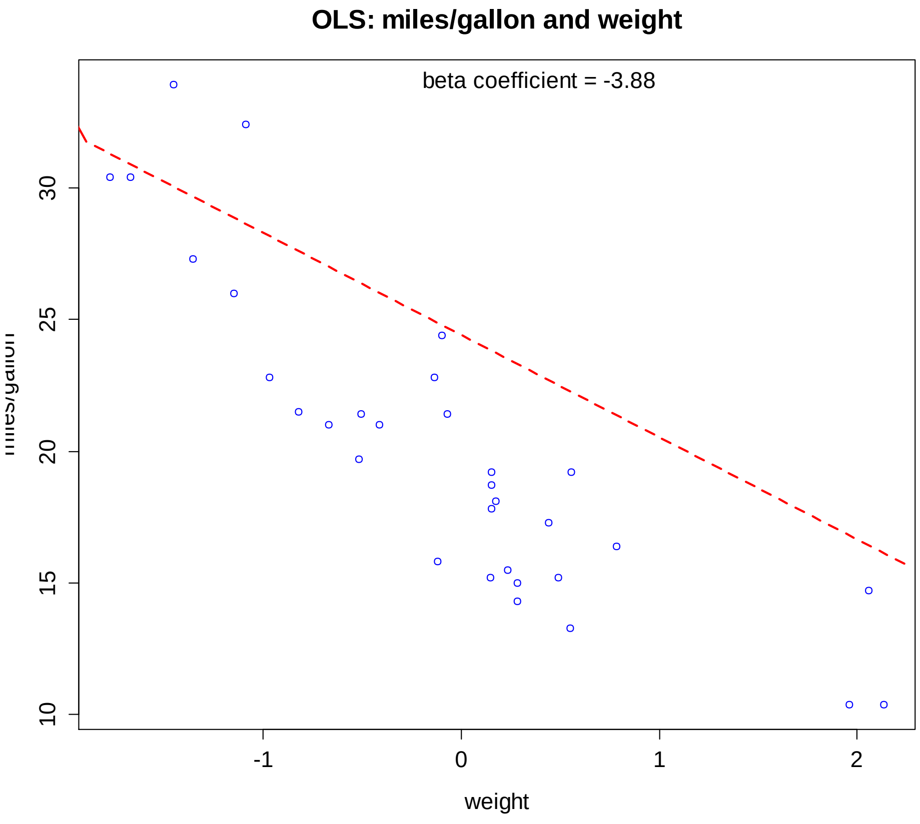


Example mtcars: X2

Unstandardized



Standardized



Interpretation effect weight:

- A 1 standard deviation increase in weight (measured in 1000lbs) leads to a 3.88 standard deviation decrease in miles per gallon
- Thus, the heavier the car, the fewer miles you can drive with a gallon of gasoline

- Controlling for horsepower:
- This effect holds for all values of horsepower. So irrespective of how fast the car can drive, an increase in weight will always lead to a decrease in miles per gallon

Exercise 5_2.r

- Include country in your model using a dummy
- Estimate manually the regression coefficients of a multiple regression equation
- Use the function `solve` for to solve the derivations
- Calculate standardized values (**not** dummy vars!)
- Check results with lm
  - Unstandardized regression vs standardized
  - Your own code vs lm

Next lecture

- moderation
- (if we have the time) mediation