QUANTITATIVE RESEARCH METHODS

DR. MEIKE MORREN

Lecture 5

contents

- Multiple regression
- Moderation
- Mean centering
- Standardization

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

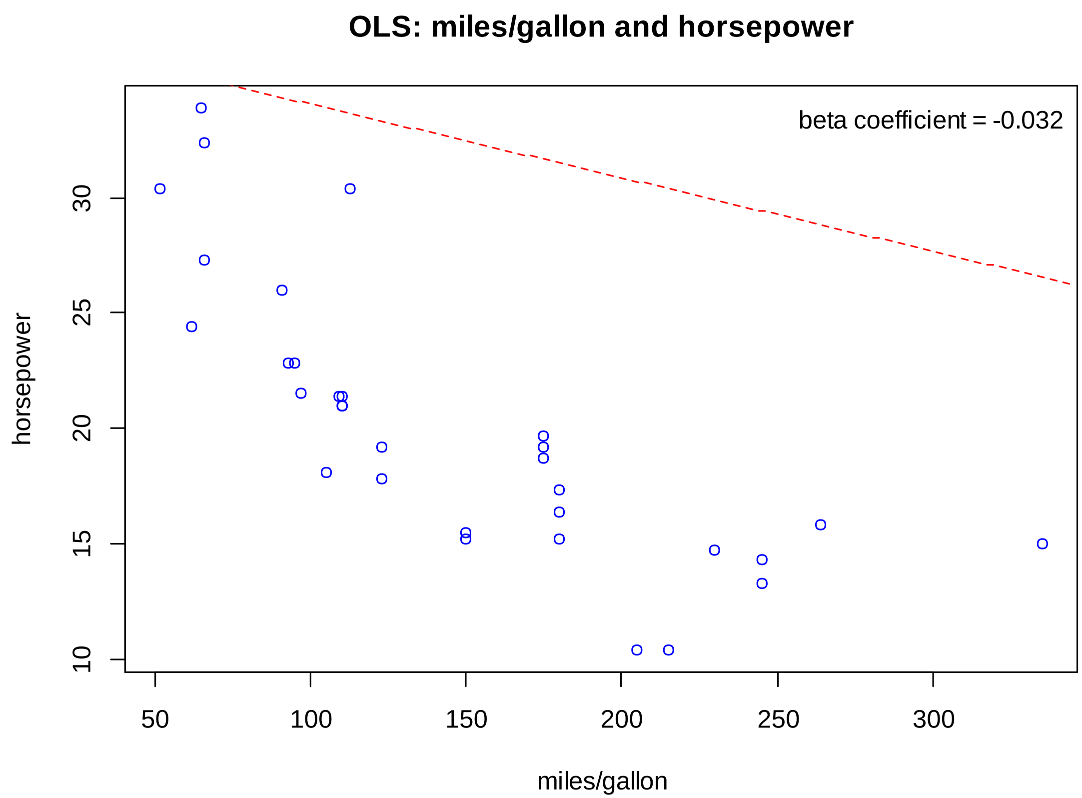
Y and X1 and X2 (hp and wt)

```
> 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
>
```

$$\widehat{Y_i} = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon_i$$

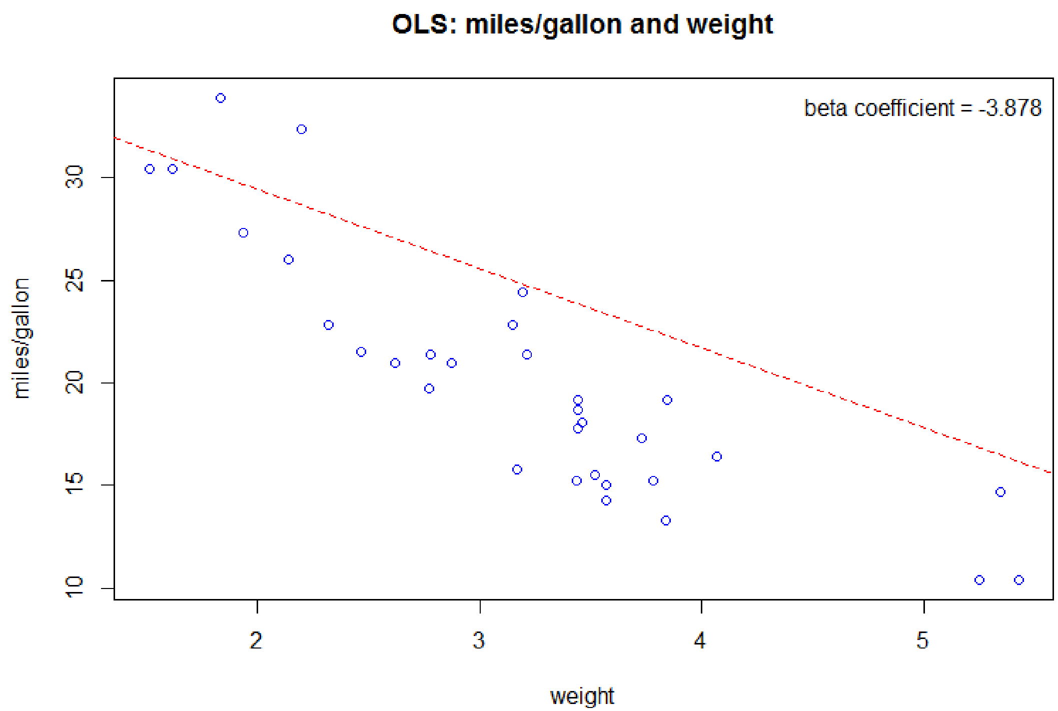
Scatterplot of y and X1

```
> 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 14.3 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 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 14.3 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
>
```



Equation (1)

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

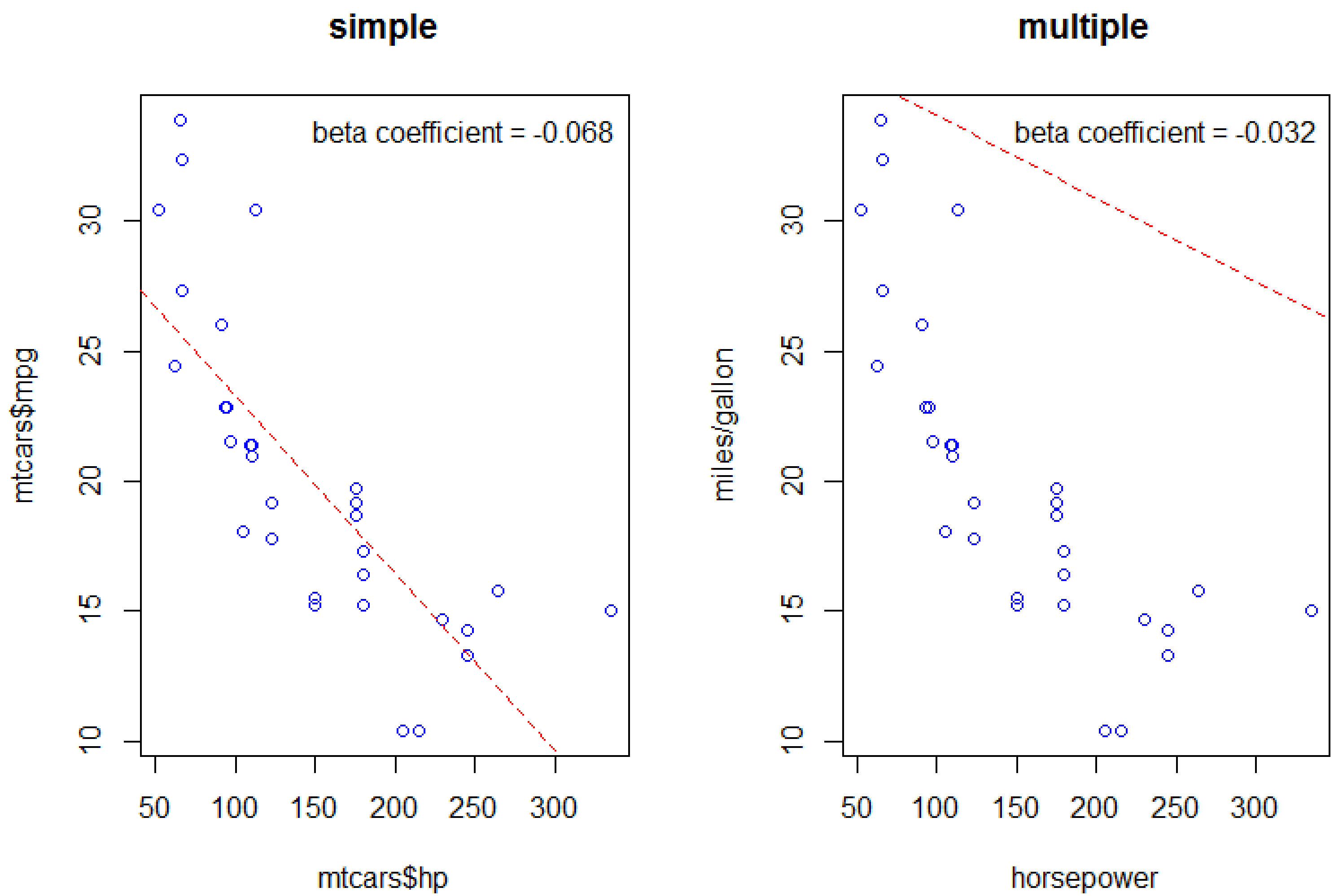
An increase of 1pt in X1 leads to increase of b1 in Y

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

Effect of controlling



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

Exe 5_1.r

- Estimate a multiple regression using two interval predictors (or ordinal treated as interval)
- Compare your results with the model where you do not control for the second variable
- Interpret the intercept and coefficients

MULTIPLE REGRESSION

Mean centering

Mean centering

- Usually a zero has no meaning for the independent variables and does not even occur in your data
- This makes the interpretation of the intercept difficult

Mean centering

- Mean centering allows you to interpret the intercept
- By subtracting the mean from all values on the independent variable, you make the mean zero
- The coefficients remain the same: a one point increase is still a one point increase

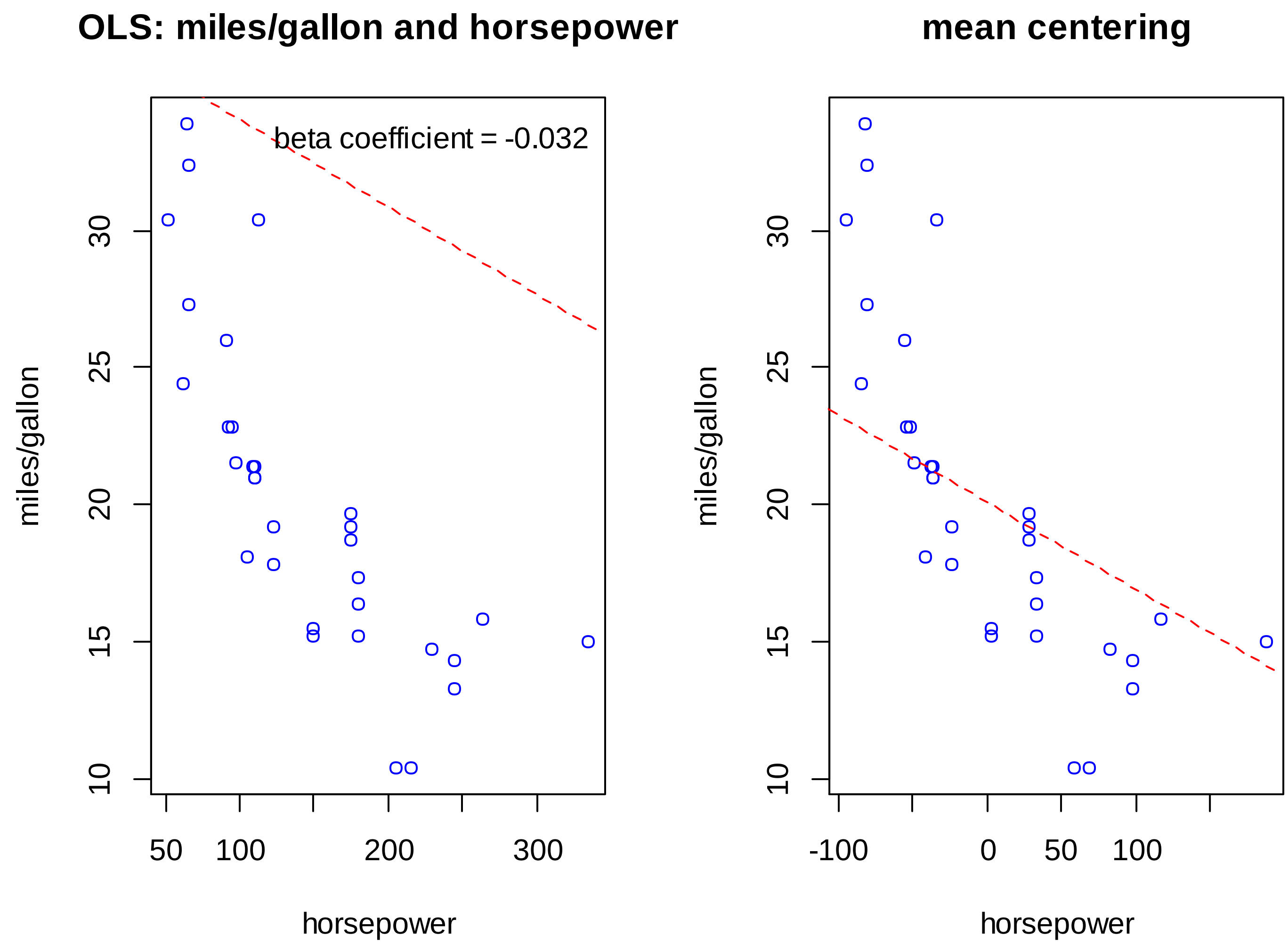
Mean centering

X	mean		X mean centered
110	110	110 – 110	0
110	110	110 – 110	0
931	110	931 – 110	821
110	110	110 – 110	0
175	110	175 – 110	65
105	110	105 – 110	-5
245	110	245 – 110	135
62	110	62 – 110	-48
95	110	95 – 110	-15

Example mtcars

```
> summary(lm(mpg ~ mhp + mwt, data = mtcars))

Call:
lm(formula = mpg ~ mhp + mwt, 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) 20.09062    0.45846  43.822  < 2e-16 ***
mhp         -0.03177    0.00903  -3.519  0.00145 ** 
mwt         -3.87783    0.63273  -6.129 1.12e-06 ***
---
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

> summary(lm(mpg ~ hp + wt, data = mtcars))

Call:
lm(formula = mpg ~ hp + wt, 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 ***
hp          -0.03177    0.00903  -3.519  0.00145 ** 
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example mtcars horsepower



Exe 5_2.r

- Center all your independent variables

- Estimate the model again, and compare with your result obtained in exe 5_1.r

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

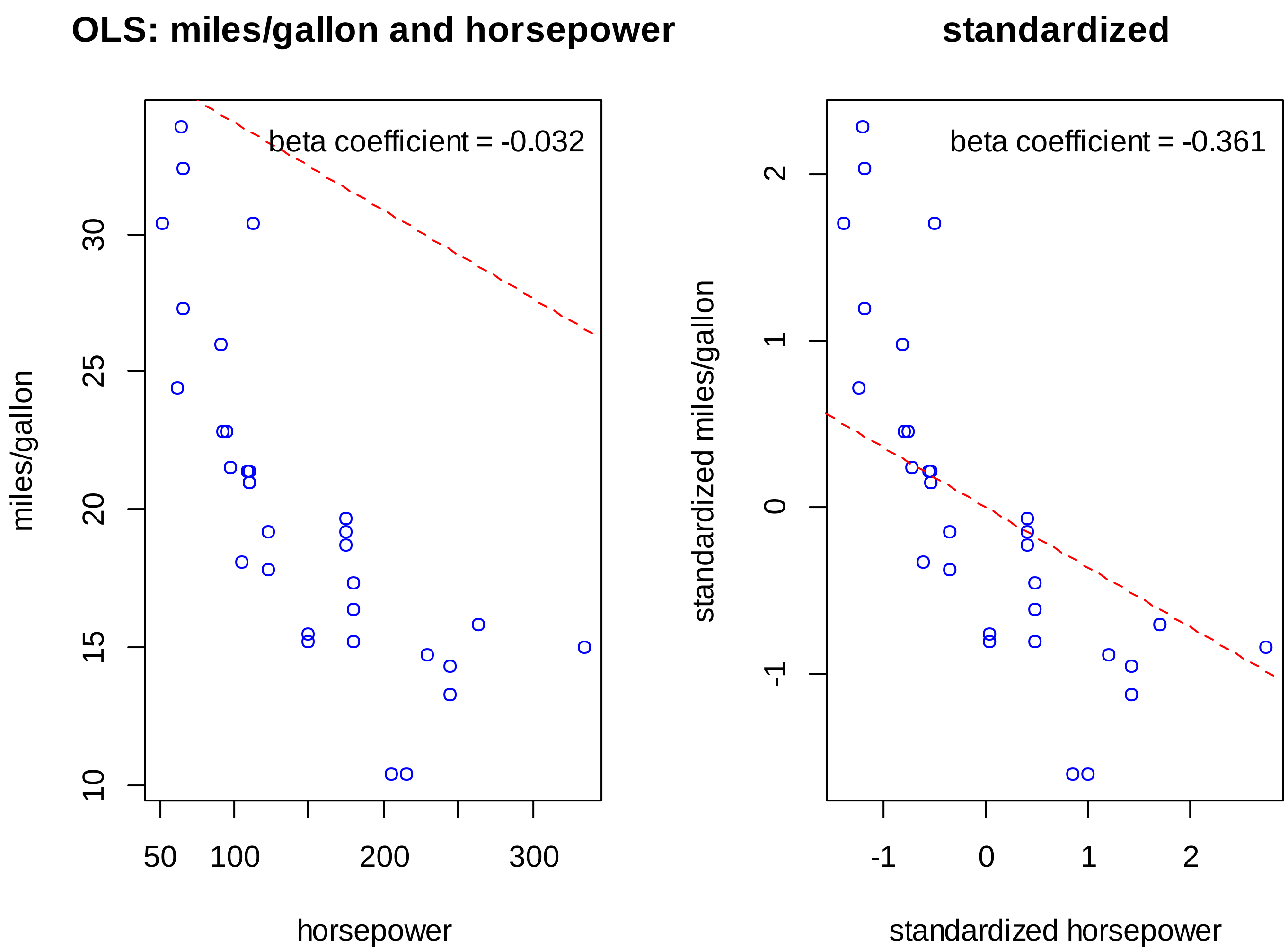
Z-scores have a

mean of zero and standard deviation of 1

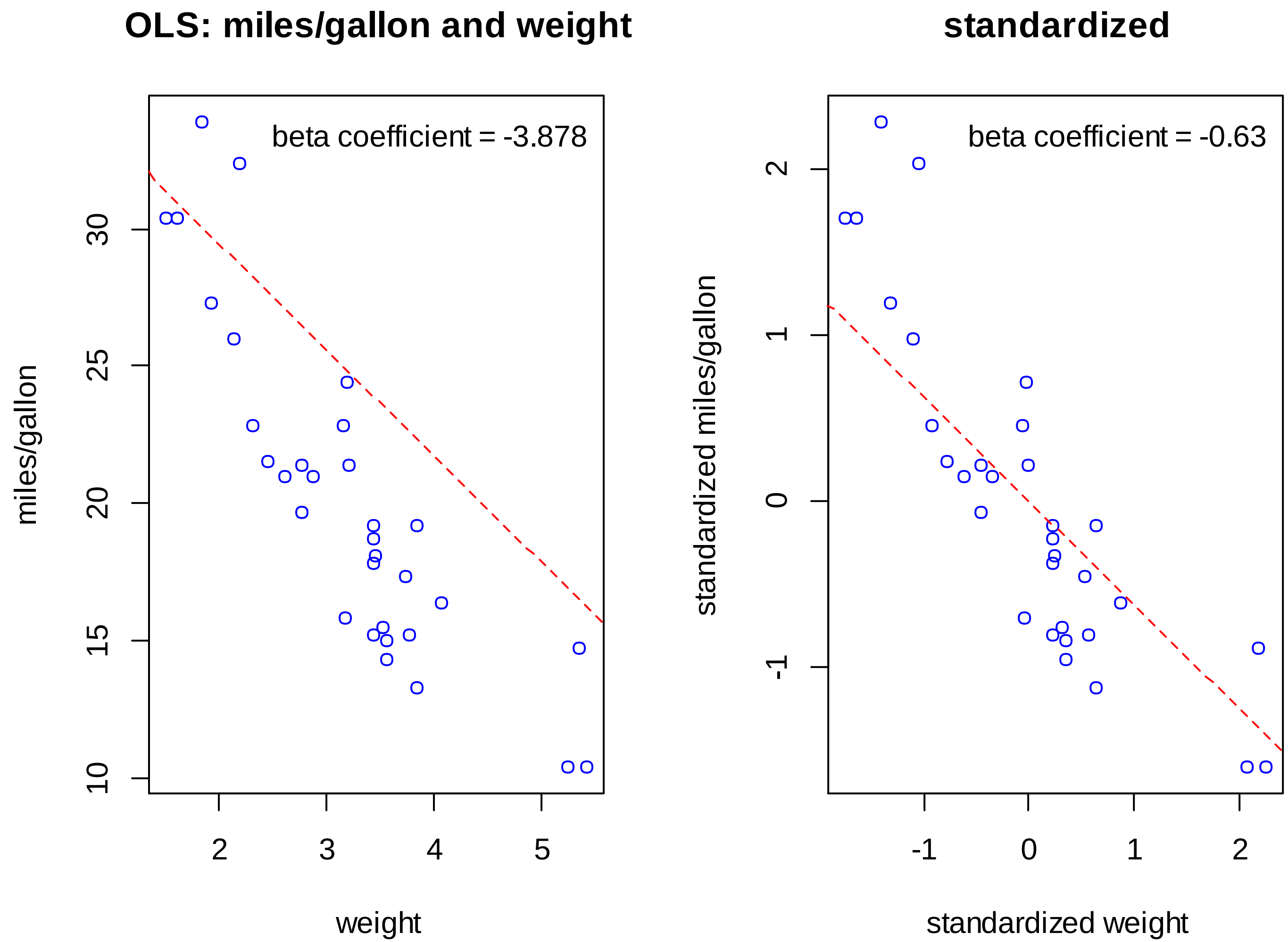
Example mtcars

- First calculate z-scores
- Create vectors of standardized scores
- Solve equation with standardized values
- Plot the line
- Compare with previous results

Example mtcars : horsepower



Example mtcars : weight



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

Exe 5_3.r

- Standardize all your variables (independent and dependent!)

- Estimate the model again, and compare with your result obtained in exe 5_1.r and exe 5_2.r

Next lecture

- moderation

- We will use the other dataset on tablets!