QUANTITATIVE RESEARCH METHODS

DR. MEIKE MORREN

Lecture 4

contents

- Linear regression
  - Deterministic vs Probabilistic
  - Simple regression with nominal, ordinal and interval variables
  - T-test

- Estimating the coefficients
- 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:

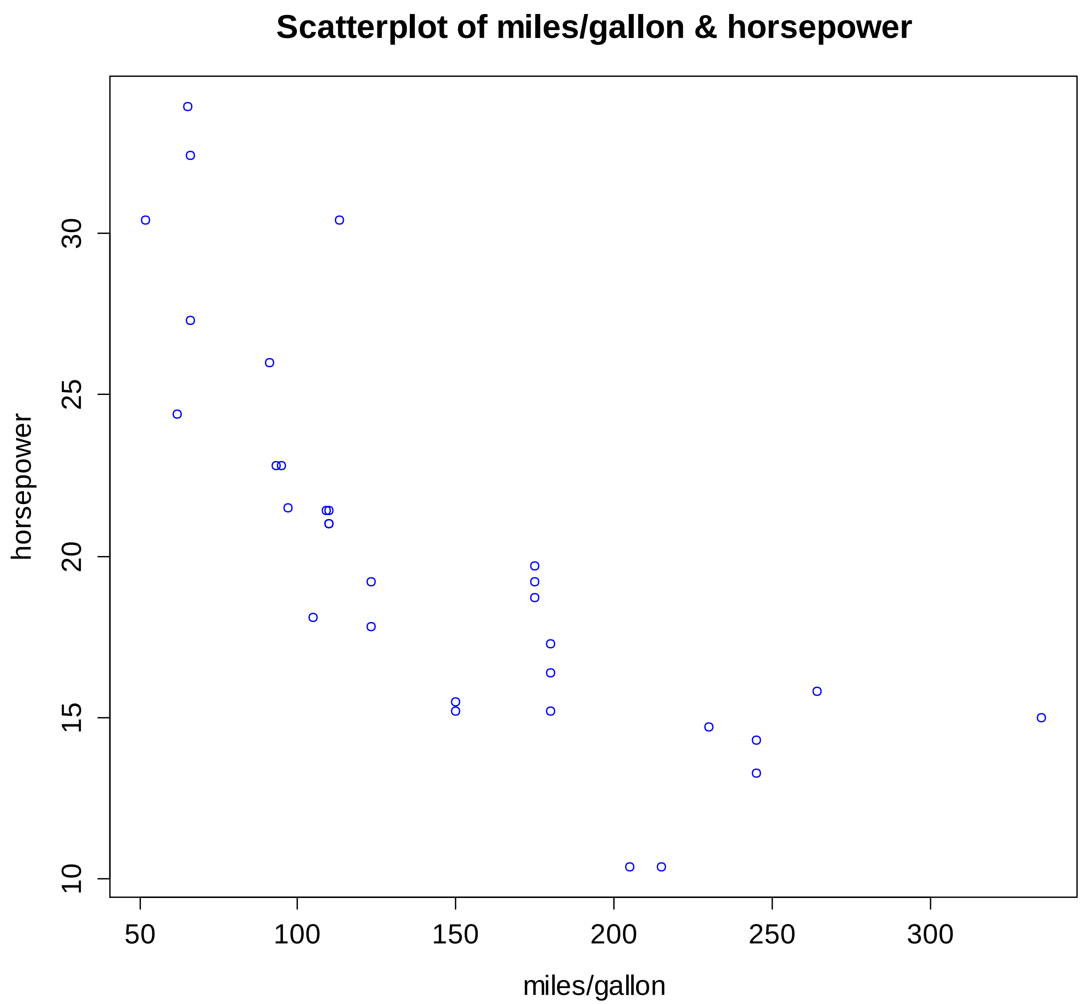
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

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

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

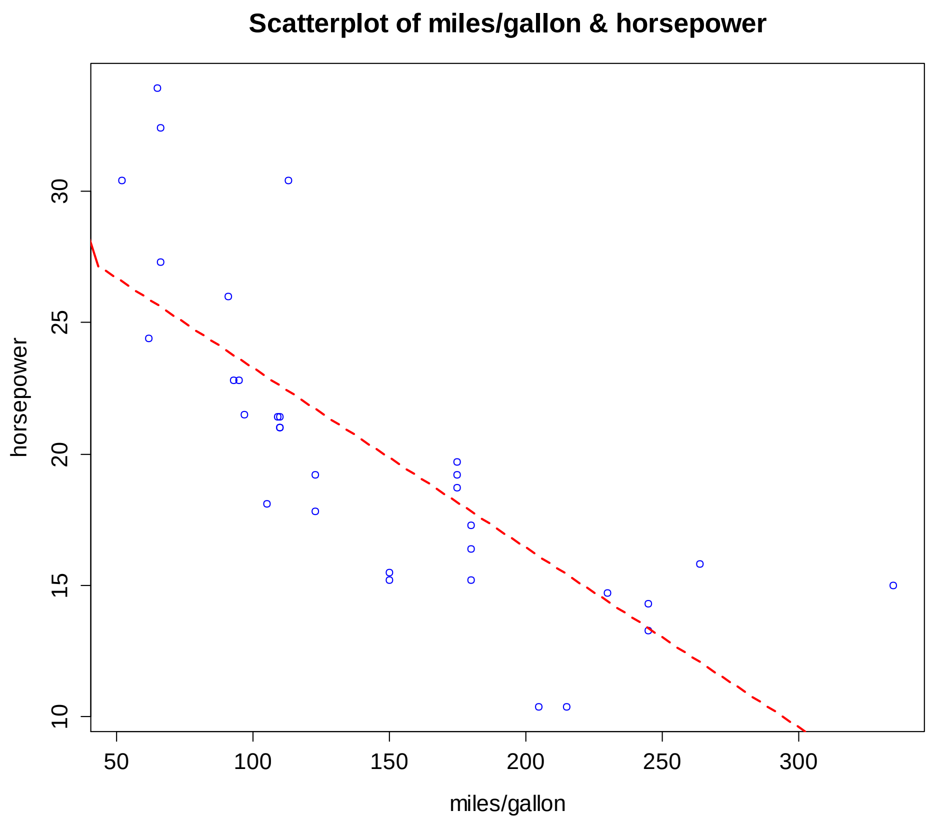
$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- 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)

```
z <- lm(mpg ~ hp, data = mtcars)
plot(mtcars$hp,mtcars$mpg, col="blue")
abline(z,lty="dashed", col="red")
```



Equation (1/3)

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

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

Equation (2/3)

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

The mean level of Y where X1 is zero

Equation (3/3)

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

The predicted level of Y (=mean of 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

R^2 = 1- SSE/SST

R^2 adjusted =

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

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

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

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

NOMINAL INDEPENDENT VARIABLES

Nominal variables

Is equal to a dummy variable:

Ex. Female (1) and male (0)

Coefficient interpretation for dummy variables

The coefficient still represents a one-point increase, but now this means the effect of being male on the dependent variable

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

The effect of being male (i.e. X1=1) on Y

Intercept interpretation for dummy variables

The constant still represents the mean level of the dependent variable where the independent variables are zero, now this is being female

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

The mean level of Y for females (i.e. X1=0)

Example mtcars

```
F-statistic: 23.66 on 1 and 30 DF,  p-value: 3.416e-05

> summary(lm(mpg ~ vs, data = mtcars))

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

Residuals:
   Min     1Q Median     3Q    Max 
-6.757 -3.082 -1.267  2.828  9.383 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   16.617      1.080  15.390 8.85e-16 ***
vs             7.940      1.632   4.864 3.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.581 on 30 degrees of freedom
Multiple R-squared:  0.4409,	Adjusted R-squared:  0.4223 
F-statistic: 23.66 on 1 and 30 DF,  p-value: 3.416e-05

> 
```

Example mtcars

```
F-statistic: 23.66 on 1 and 30 DF,  p-value: 3.416e-05

> summary(lm(mpg ~ vs, data = mtcars))

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

Residuals:
   Min     1Q Median     3Q    Max 
-6.757 -3.082 -1.267  2.828  9.383 

Coefficients:
            Estimate Std. Er
(Intercept)   16.617       
vs             7.940      1.
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.581 on 30 degrees of freedom
Multiple R-squared:  0.4409,    Adjusted R-squared:  0.4223 
F-statistic: 23.66 on 1 and 30 DF,  p-value: 3.416e-05

> 
```

The mean level of Y where X is 0

Example mtcars

```
F-statistic: 23.66 on 1 and 30 DF,  p-value: 3.416e-05

> summary(lm(mpg ~ vs, data = mtcars))

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

Residuals:
   Min     1Q Median     3Q    Max 
-6.757 -3.082 -1.267  2.828  9.383 

Coefficients:
            Estimate Std. Erro  t value Pr(>|t|)
(Intercept)   16.617       1
vs             7.940
---
Signif. codes:  0 '***' 0.001      *' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.581 on 30 degrees of freedom
Multiple R-squared:  0.4409,    Adjusted R-squared:  0.4223 
F-statistic: 23.66 on 1 and 30 DF,  p-value: 3.416e-05

> 
```

The increase in Y where X is 1 (=USA)

t.test()

- When you ad a dummy variable to the model, you compare two means

- The mean of Y when X is zero
- The mean of Y when X is one

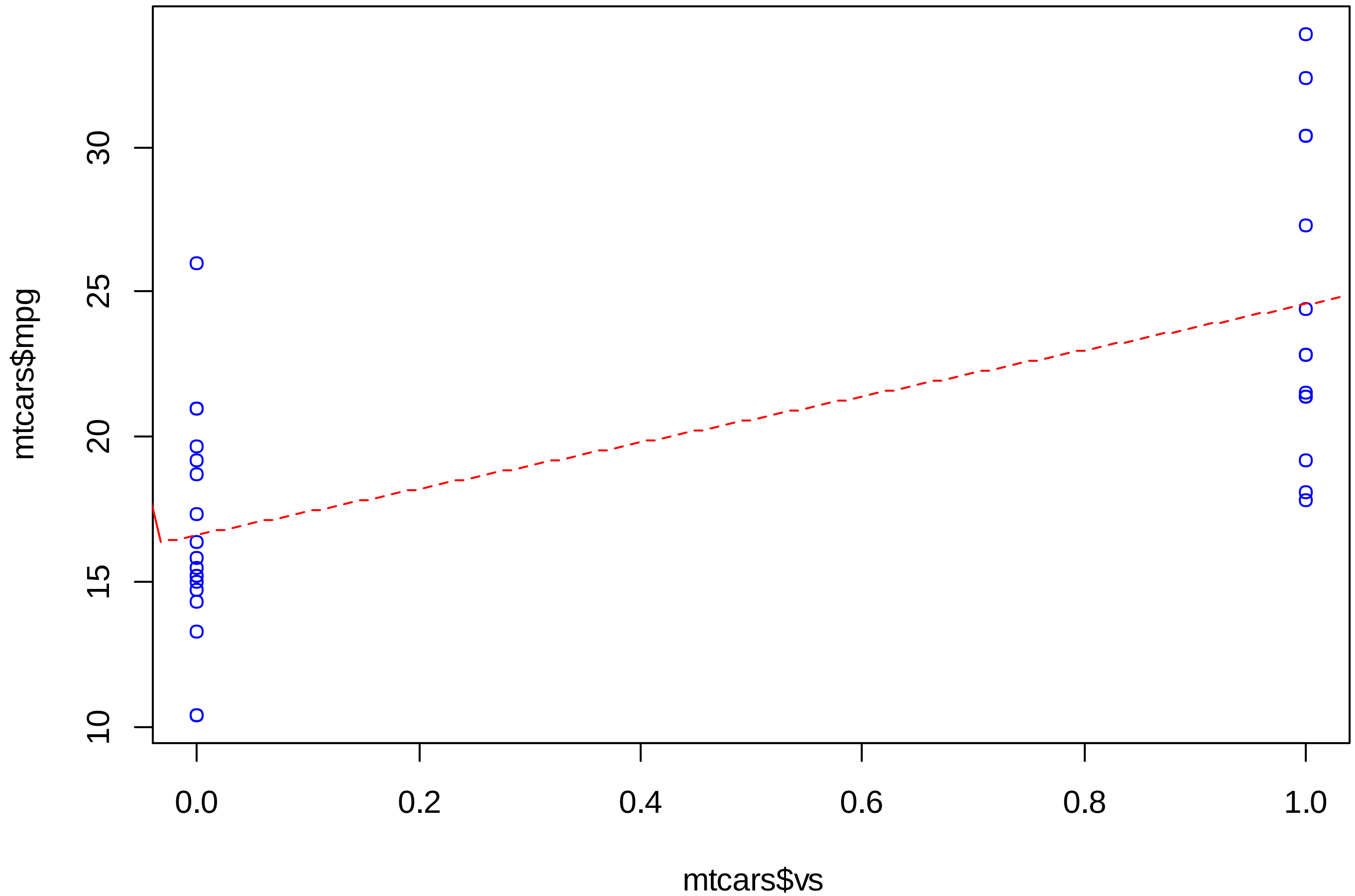
- This is exactly the same as a t-test!

Example mtcars

```
summary(lm(mpg ~ vs, data = mtcars))
plot(mtcars$vs,mtcars$mpg, col="blue")
abline(z,lty="dashed", col="red")
```

```
F-statistic: 23.66 on 1 and 30 DF,  p-value: 3.416e-05

> summary(lm(mpg ~ vs, data = mtcars))

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

Residuals:
   Min     1Q Median     3Q    Max 
-6.757 -3.082 -1.267  2.828  9.383 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   16.617      1.080  15.390 8.85e-16 ***
vs             7.940      1.632   4.864 3.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example mtcars



Exercise 4_1.r

Use the WVS dataset

- *Relate happiness to country*

a) Estimate a model with dummy variable (country)
b) Change the dummy variable into a factor
c) Change reference group
d) Check with t.test function

ORDINAL INDEPENDENT VARIABLES

Ordinal variables

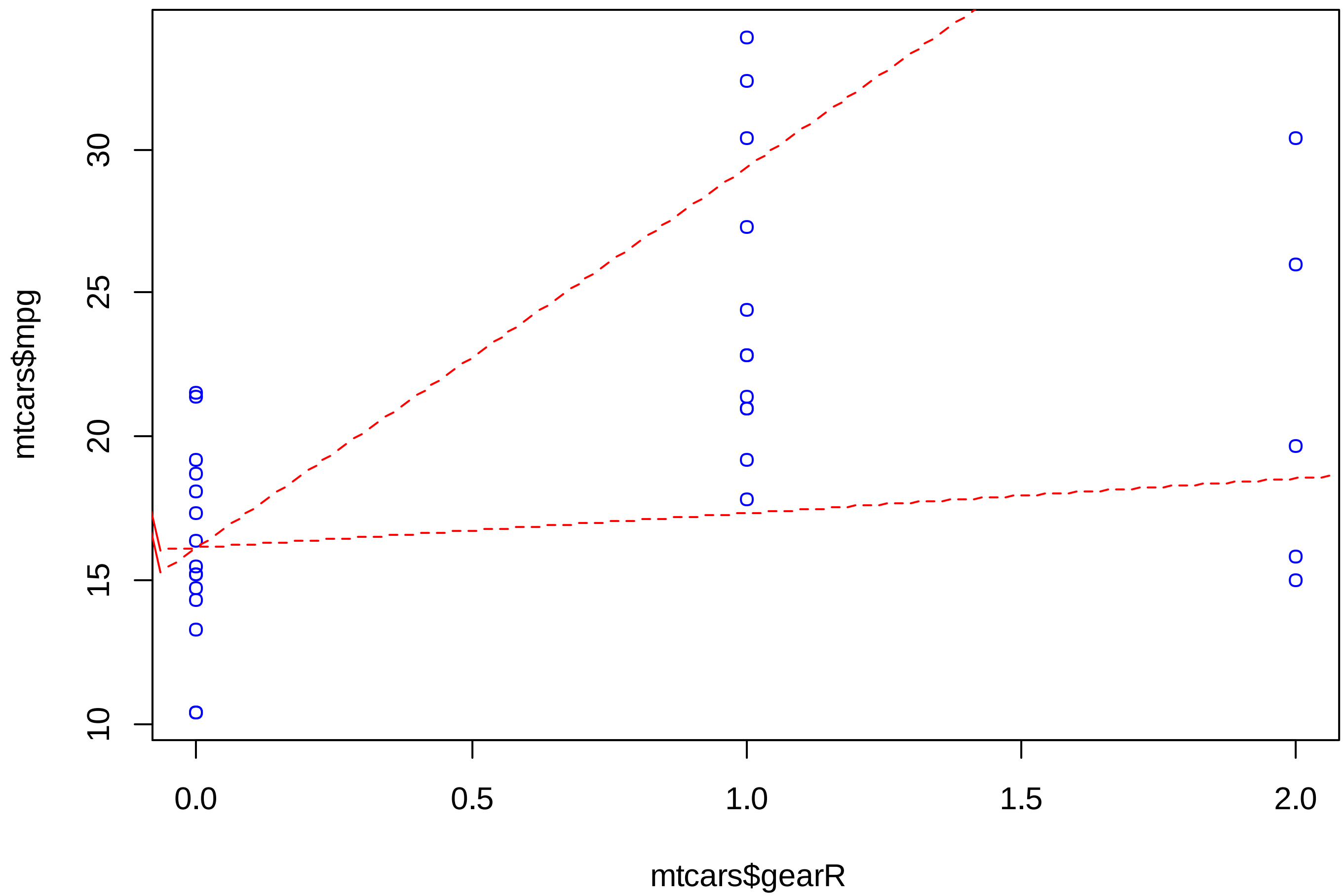
- You should regard a variable ordinal when you can assume order, but you are not sure about the equal distances

- Compare two models in which
  - (1) you include this variable as a multinomial variable (and explore each category separately)
  - (2) you include the variable as interval variable

Ordinal variables (factor())

```
> summary(lm(mpg ~ gear, data = mtcars))

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

Residuals:
    Min      1Q  Median      3Q     Max 
-10.240  -2.793  -0.205   2.126  12.583 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)    5.623      4.916   1.144   0.2618   
gear           3.923      1.308   2.999   0.0054 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.374 on 30 degrees of freedom
Multiple R-squared:  0.2307,	Adjusted R-squared:  0.205 
F-statistic: 8.995 on 1 and 30 DF,  p-value: 0.005401

> summary(lm(mpg ~ factor(gearR), data = mtcars))

Call:
lm(formula = mpg ~ factor(gearR), data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.7333 -3.2333 -0.9067  2.8483  9.3667 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      16.107      1.216  13.250 7.87e-14 ***
factor(gearR)1    8.427      1.823   4.621 7.26e-05 ***
factor(gearR)2    5.273      2.431   2.169   0.0384 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Plot two lines



Plot two lines

```
z<-summary(lm(mpg ~ factor(gearR), data = mtcars))

plot(mtcars$gearR,mtcars$mpg, col="blue")

abline(a=z$coef[1,1],b=z$coef[1,2],lty="dashed",
col="red")

abline(a=z$coef[1,1],b=z$coef[1,3],lty="dashed",
col="red")
```

Exercise 4_2.r

Use the WVS dataset

- *Relate happiness to education level or age*

a) Estimate a model with an ordinal variable (eduR or ageR)
b) Recode the ordinal variable so that the lowest level is zero
c) Compare with factor variable
d) Plot the lines (optional)

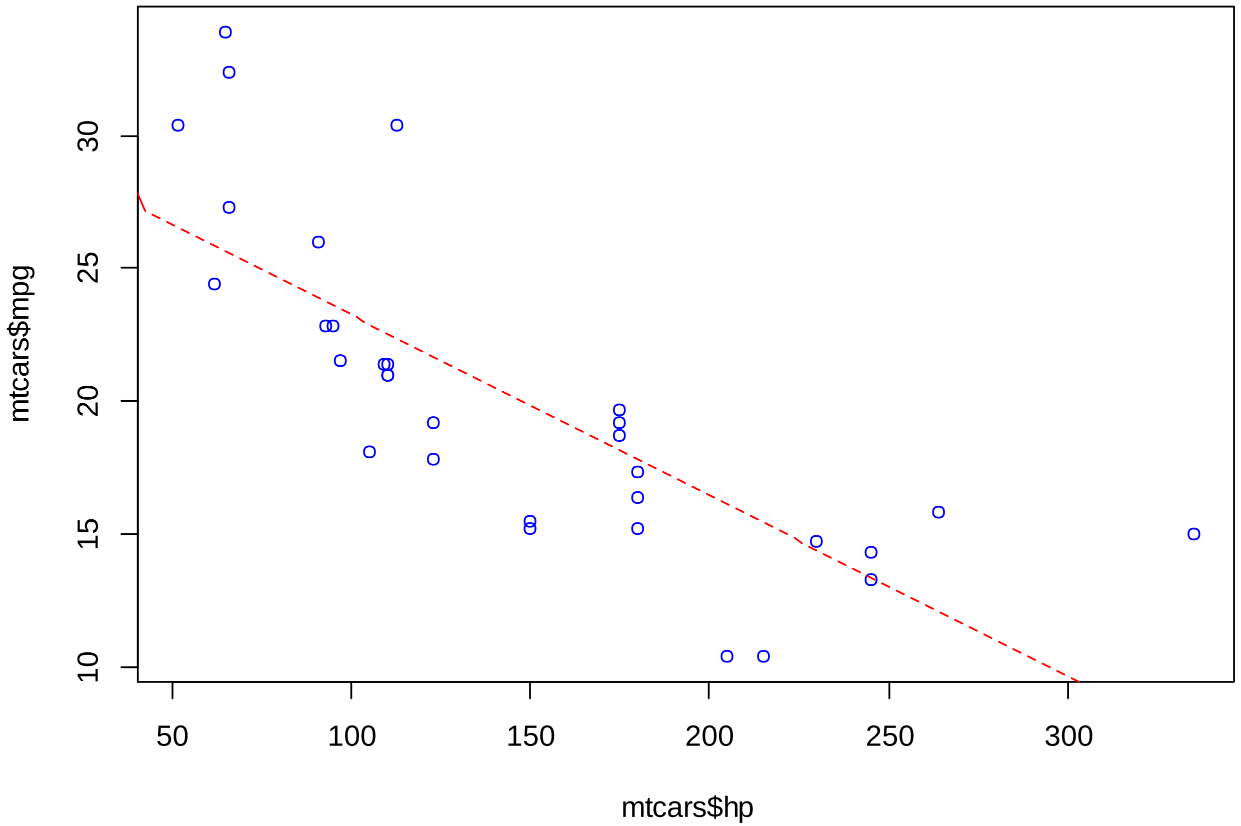
INTERVAL INDEPENDENT VARIABLES

Interval variables

- This is the level usually assumed
- A one-point increase is the same across all levels of the variable
- One straight line is estimated

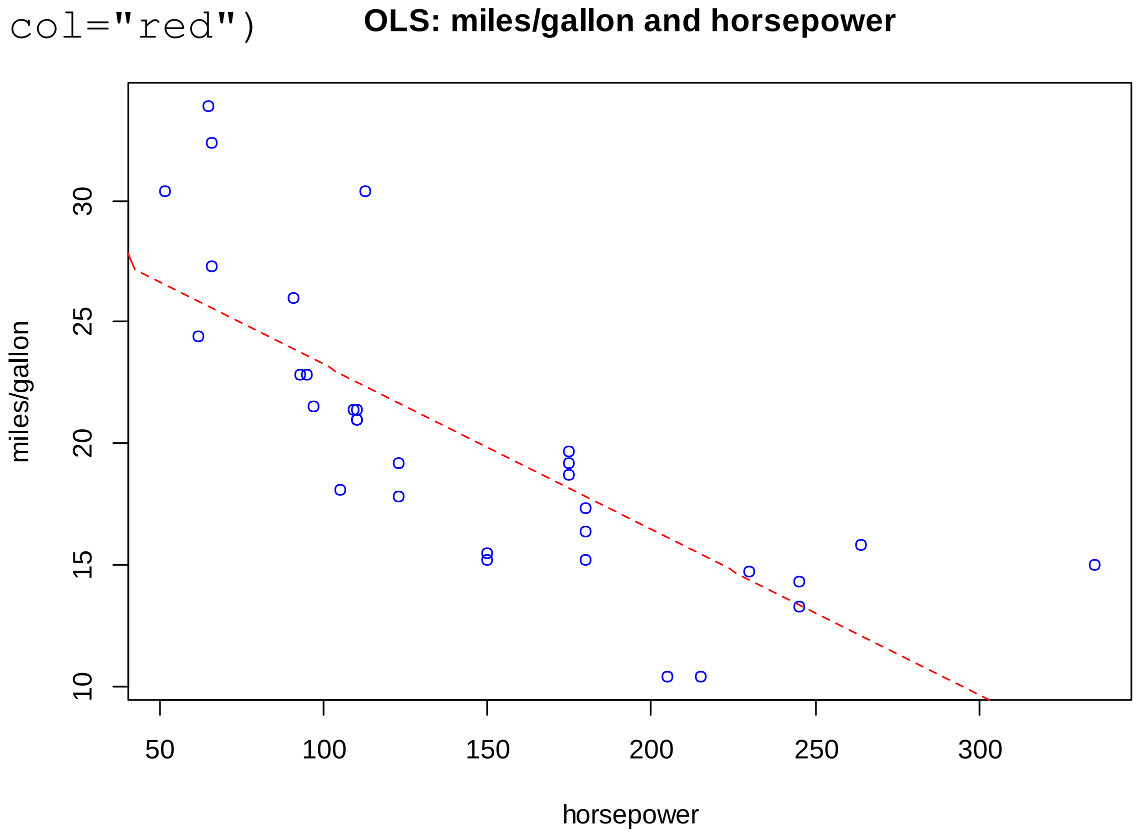
Example mtcars

```
z <- lm(mpg ~ hp, data = mtcars)
plot(mtcars$hp,mtcars$mpg, col="blue")
abline(z,lty="dashed", col="red")
```



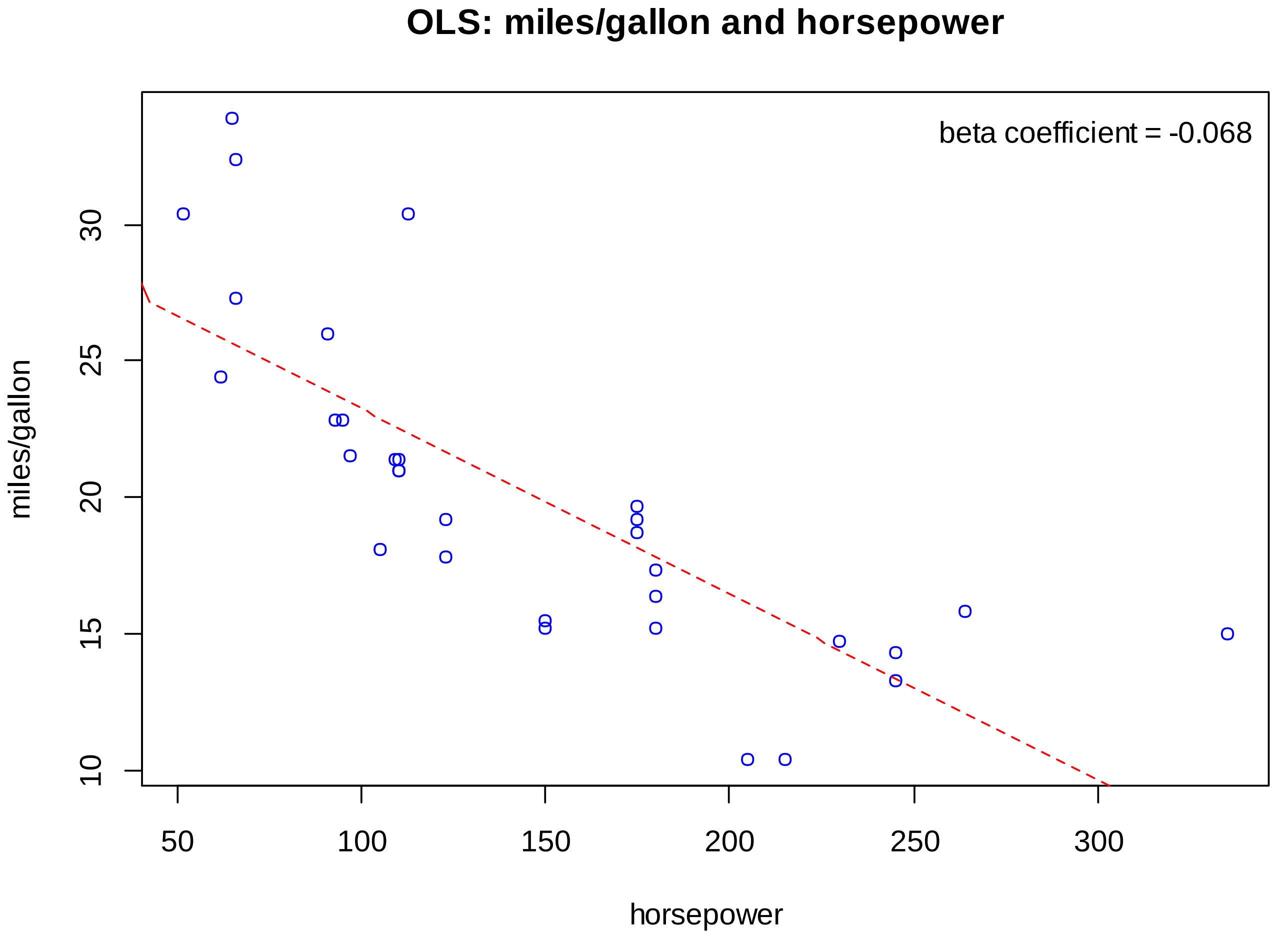
Ad axis labels

```
z <- lm(mpg ~ hp, data = mtcars)
plot(mtcars$hp,mtcars$mpg,col="blue",
      main = "OLS: miles/gallon and horsepower",
      xlab="horsepower",ylab="miles/gallon")
abline(z,lty="dashed", col="red")
```



Ad legend

```
legend("topright", bty="n", legend=paste("beta coefficient
=", round(z$coef[2], digits=3)))
```



Exercise 4_3.r

Use the WVS dataset

- *Relate happiness to income*

a) Estimate a model with income
b) Plot the line

Next lecture

- Multiple regression