Intro to Statistical Models: Linear Regression

Michael Malick

Updated: 15 May 2013

Linear Models

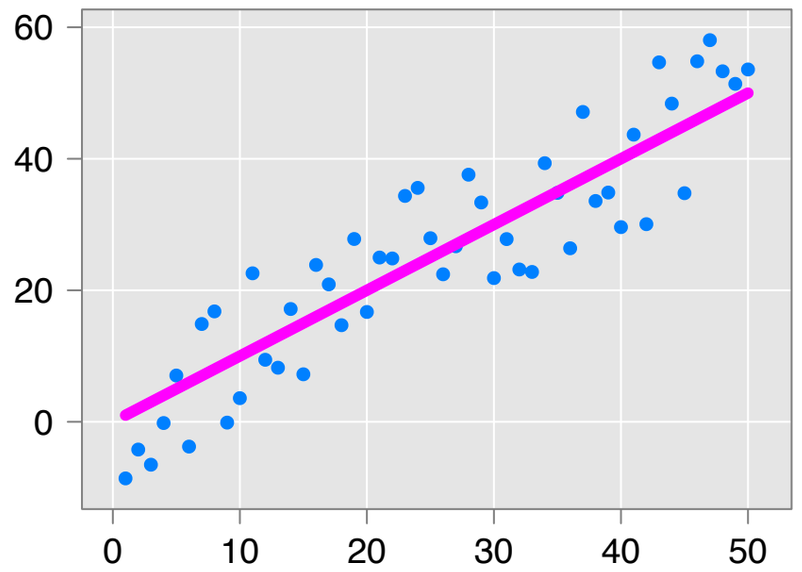
R has numerous built in functions to fit linear models

- Linear regression: `lm()`
- ANOVA: `aov()`
- Generalized linear model: `glm()`

Simple Linear Regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- y_i = response variable
- β_0 = intercept
- β_1 = slope
- x_i = predictor variable
- ϵ_i = error term



Linear Model Example

We will look at the relationship between miles per gallon and horsepower using data in the `mtcars` dataset

```
head(mtcars)
```

```
plot(mtcars$mpg ~ mtcars$hp)
```

```
cor(mtcars$mpg, mtcars$hp)
```

Linear Model Formula

The `lm()` function requires a formula to fit a model

```
y ~ x1 + x2 ...
```

- Response variable is on the left
- Predictor variables are on the right

Linear Model Example

We will fit a linear regression with **mpg** as the response variable and **horsepower** as the predictor variable

```
fit1 <- lm(mpg ~ hp, data = mtcars)

fit1
summary(fit1)

plot(mtcars$mpg ~ mtcars$hp)
abline(fit1, col = 2)
```

Extract Output

You can easily extract information from the fitted object

```
names(fit1)
fit1$residuals
fit1$coefficients

fit.sum <- summary(fit1)
names(fit.sum)
fit.sum$r.squared
```

Generic Modeling Functions

```
plot(fit1)

coef(fit1)

residuals(fit1)

fitted(fit1)

vcov(fit1)

confint(fit1)
```

Diagnostics: Histogram

```
hist(resid(fit1), col = 1, las = 1,
main = "Histogram of Residuals",
xlab = "Residuals")
```

Diagnostics: Residuals Plot

```
plot(resid(fit1) ~ fitted(fit1),
     col = 4, las = 1, pch = 19,
      main = "Residuals vs. Fitted Values",
      ylab = "Residuals",
      xlab = "Fitted Values")

abline(h = 0, lty = 2)
```

Diagnostics: QQ-Plot

```
qqnorm(resid(fit1), col = 4, pch = 19)
qqline(resid(fit1), col = 2)
```

Diagnostics: `lm.diag()`

- script_lm_diag.R

```
lm.diag(fit1)
```

Multiple Regression

The engine size or displacement might also be a good predictor of miles per gallon

```
plot(mtcars$mpg ~ mtcars$disp)
```

```
fit2  <- lm(mpg ~ disp, data = mtcars)
summary(fit2)
abline(fit2, col = 2)

fit3 <- lm(mpg ~ hp + disp, data = mtcars)
summary(fit3)
```

Model Comparisons: `AIC()`

```
fit1  <- lm(mpg ~ hp, data = mtcars)
summary(fit1)

fit2  <- lm(mpg ~ disp, data = mtcars)
summary(fit2)

fit3 <- lm(mpg ~ hp + disp, data = mtcars)
summary(fit3)

AIC(fit1, fit2, fit3)
```

You Try...

1. Using the `faithful` dataset, plot eruption length vs. waiting time
2. Fit a simple linear regression model with eruptions as the response variable (y) and waiting time as the predictor variable (x)
3. Add the fitted regression line to the plot
4. What is the R^2 value of the regression?