Linear regression in R

Topic 4.2. Estimating the coefficients in R; Model analytics and diagnostics

Sonja Petrović
Created for ITMD/ITMS/STAT 514

Spring 2021.

Goals for this lecture

- Understand & interpret coefficient estimates in multiple and simple linear regression
- Fit a regression model in `R`
- Understand & interpret `R` output for linear models
- Model diagnostics & assessing model fit

Some important questions about linear regression model

1. Is at least one of the predictors X1,dots, Xp useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Section 1

Simple linear regression case

Is There a Relationship?

Question

Is there a relationship between the response Y and predictor X?

Recall from last lecture:

- check whether $\beta_1 = 0$
  - Hypothesis test: $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.
  - a t-statistic measures the number of standard deviations that β_1 is away from 0 (specifically, $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$)
  - p-value
    - this is defined - as usual! - the probability of seeing the data we saw, or more extreme, under the H_0.
    - **in practice, we just read off the t-test. *or* read off the output of linear models.**

Assessing model fit

Question

Suppose we have rejected the null hypothesis in favor of the alternative. Now what??

- Natural: quantify the extent to which the model fits the data.
- The quality of a linear regression fit is typically assessed using two related quantities:
  - the residual standard error (RSE) and
  - the R^2 statistic.

$\rightarrow$ advertising example - revisit the statistics output.

Assessing model fit

Question

Suppose we have rejected the null hypothesis in favor of the alternative. Now what??

- Natural: quantify the extent to which the model fits the data.
- The quality of a linear regression fit is typically assessed using two related quantities:
  - the residual standard error (RSE) and
  - the R^2 statistic.

→ advertising example - revisit the statistics output.

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

TABLE 3.1. *For the `Advertising` data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of $1,000 in the TV advertising budget is associated with an increase in sales by around 50 units (Recall that the `sales` variable is in thousands of units, and the `TV` variable is in thousands of dollars).*

Assessing model fit

Question

Suppose we have rejected the null hypothesis in favor of the alternative. Now what??

- Natural: quantify the extent to which the model fits the data.
- The quality of a linear regression fit is typically assessed using two related quantities:
  - the residual standard error (RSE) and
  - the R^2 statistic.

$\rightarrow$ advertising example - revisit the statistics output.

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

TABLE 3.1. *For the* Advertising *data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of $1,000 in the TV advertising budget is associated with an increase in sales by around 50 units (Recall that the* sales *variable is in thousands of units, and the* TV *variable is in thousands of dollars).*

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

RSE

A measure of the lack of fit of the model simple linear regression model to the data:

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RSE

A measure of the lack of fit of the model simple linear regression model to the data:

$$RSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

- If the predictions obtained using the model are very close to the true outcome values ($\hat{y}_i \approx y_i$ for i = 1, . . . , n), then RSE will be small
  - we can conclude that the model fits the data very well.
- If $\hat{y}_i$ is very far from y_i for one or more observations, then the RSE may be quite large
  - indicating that the model doesn't fit the data well.

RSE

A measure of the lack of fit of the model simple linear regression model to the data:

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- If the predictions obtained using the model are very close to the true outcome values ($\hat{y}_i \approx y_i$ for i = 1, . . . , n), then RSE will be small
  - we can conclude that the model fits the data very well.
- If $\hat{y}_i$ is very far from y_i for one or more observations, then the RSE may be quite large
  - indicating that the model doesn't fit the data well.

Interpretation

The RSE provides an absolute measure of lack of fit. But since it is measured in the units of Y , it is not always clear what constitutes a good RSE. . .

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

Figure 1: ISLR table 3.2. For the `Advertising` data, more information about the least squares model for the regression of number of units sold on TV advertising budget.

R^2

The R^2 statistic provides an alternative measure of fit (proportion):

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- TSS $=$ total sum of squares $\sum(y_i - \bar{y}_i)^2$
- RSS $=$ residual sum of squares $\sum(y_i - \hat{y}_i)^2$

Discuss: R^2 measures the proportion of variability in Y that can be explained using X

R^2

The R^2 statistic provides an alternative measure of fit (proportion):

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- TSS $=$ total sum of squares $\sum(y_i - \bar{y}_i)^2$
- RSS $=$ residual sum of squares $\sum(y_i - \hat{y}_i)^2$

Discuss: R^2 measures the proportion of variability in Y that can be explained using X

Interpretation

Proportion of variance explained.
Always between 0 and 1 (independent of scale of Y).

What's a good value?

Can be challenging to determine ... in general, depends on the application.

Example

Objective:

Use simple linear regression on the ‘Auto‘ data set.

- Use the `lm()` function to perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor.

```
require(ISLR)
```

```
Loading required package: ISLR
```

```
data(Auto)
fit.lm <- lm(mpg ~ horsepower, data=Auto)
```

$\rightarrow$ Where is the output??

- Let’s take a look at the `fit.lm` object.

Use the summary() function to print the results.

```
summary(fit.lm)
```

```
Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

- Is there a relationship between the predictor and the response?

```
Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

- Is there a relationship between the predictor and the response?
    - Yes

```
Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

- Is there a relationship between the predictor and the response?
  - Yes
- How strong is the relationship between the predictor and the response?

```
Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

- Is there a relationship between the predictor and the response?
  - Yes
- How strong is the relationship between the predictor and the response?
  - *p*-value is close to 0: relationship is strong

```
Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

- Is there a relationship between the predictor and the response?
  - Yes
- How strong is the relationship between the predictor and the response?
  - p-value is close to 0: relationship is strong
- Is the relationship between the predictor and the response positive or negative?

```
Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

- Is there a relationship between the predictor and the response?
  - Yes
- How strong is the relationship between the predictor and the response?
  - *p*-value is close to 0: relationship is strong
- Is the relationship between the predictor and the response positive or negative?
  - Coefficient is negative: relationship is negative

- What is the predicted `mpg` associated with a `horsepower` of 98? What are the associated 95% confidence and prediction intervals?

- What is the predicted `mpg` associated with a `horsepower` of 98? What are the associated 95% confidence and prediction intervals?

```
new <- data.frame(horsepower = 98)
predict(fit.lm, new) # predicted mpg
```

```
       1
24.46708
```

```
predict(fit.lm, new, interval="confidence") # conf interval
```

```
       fit      lwr      upr
1 24.46708 23.97308 24.96108
```

```
predict(fit.lm, new, interval="prediction") # pred interval
```

```
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```

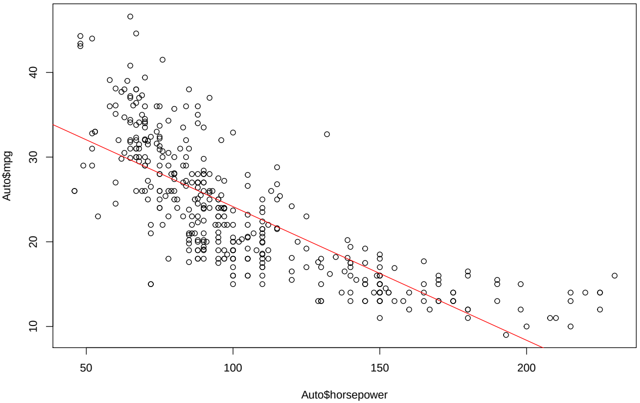
- $\rightarrow$ confidence interval vs. prediction interval $\leftarrow$

Confidence and prediction intervals!?

- next lecture!!

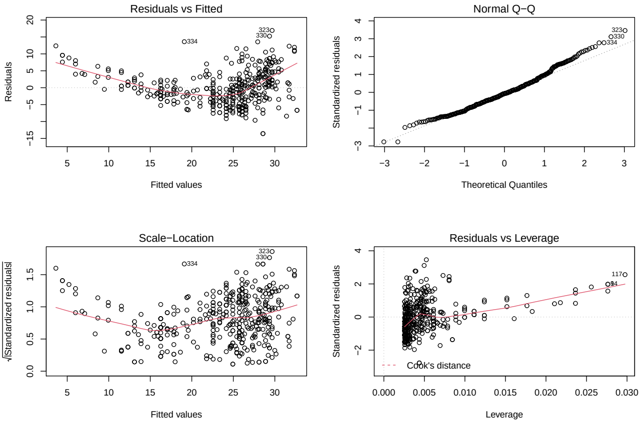
Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.

```
plot(Auto$horsepower, Auto$mpg)
abline(fit.lm, col="red")
```



Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

```r
par(mfrow=c(2,2))
plot(fit.lm)
```



- residuals vs fitted plot shows that the relationship is non-linear

Section 2

Multiple linear regression case

Is There a Relationship?

Q: is there a relationship between the Response and Predictor?

- Multiple case: p predictors; we need to ask whether all of the regression coefficients are zero: $\beta_1 = \cdots = \beta_p = 0$?

Is There a Relationship?

Q: is there a relationship between the Response and Predictor?

- Multiple case: p predictors; we need to ask whether all of the regression coefficients are zero: $\beta_1 = \cdots = \beta_p = 0$?
  - Hypothesis test: $H_0 : \beta_1 = \cdots = \beta_p = 0$ vs. H_1 : at least one $\beta_i \neq 0$.
  - Which statistic?

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}.$$

    - TSS and RSS defined as in simple case.
- when there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1. [this can be proved via expected values]
- else > 1.

$\rightarrow$ advertising example - revisit the statistics output.

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	-0.001	0.0059	-0.18	0.8599

TABLE 3.4. *For the* `Advertising` *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

	`TV`	`radio`	`newspaper`	`sales`
`TV`	1.0000	0.0548	0.0567	0.7822
`radio`		1.0000	0.3541	0.5762
`newspaper`			1.0000	0.2283
`sales`				1.0000

TABLE 3.5. *Correlation matrix for* `TV`, `radio`, `newspaper`, *and* `sales` *for the* `Advertising` *data.*

Warning

$\rightarrow$ in case of large p, may want to measure *partial effects*, and do some *variable selection* (out of scope Fall 2020).

Assessing model fit

Question

Suppose we have rejected the null hypothesis in favor of the alternative. Now what??

- Same story as for simple regression.
- Measuring the quality of a linear regression fit:
  - the residual standard error (RSE);
  - the R^2 statistic.

$\rightarrow$ advertising example - revisit the statistics output.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

TABLE 3.5. *Correlation matrix for* TV, radio, newspaper, *and* sales *for the* Advertising *data.*

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

Figure 2: ISLR Table 3.6: More information about the least squares model for the regression of number of units sold on TV, newspaper, and radio advertising budgets in the Advertising data. Other information about this model was displayed in Table 3.4.

In addition to looking at RSE and R^2 statistics, it can be useful to plot the data.

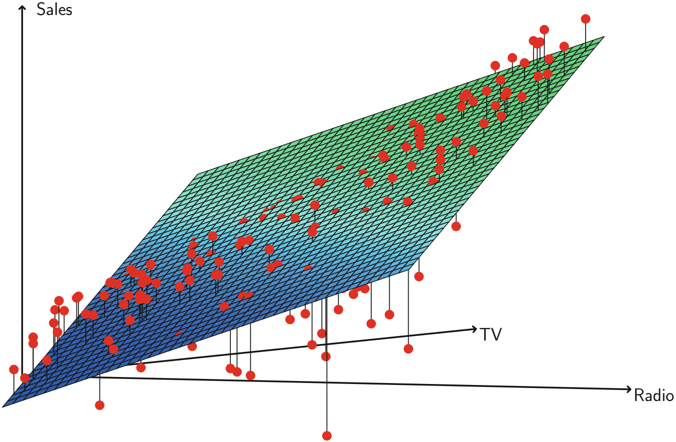


Figure 3: ISLR fig 3.5. For the Advertising data, a linear regression fit to sales using TV and radio as predictors. From the pattern of the residuals, we can see that there is a pronounced non-linear relationship in the data. The positive residuals (those visible above the surface), tend to lie along the 45-degree line

Example

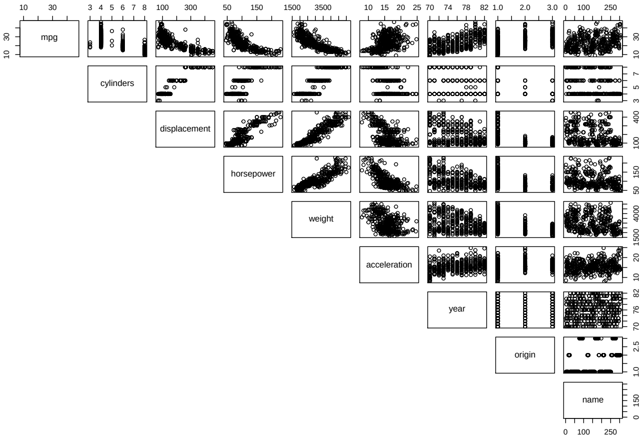
Objective

Use multiple linear regression on the ‘Auto‘ data set.

```r
require(ISLR)
data(Auto)
```

- Produce a scatterplot matrix which includes all of the variables in the data set.

```
pairs(Auto,lower.panel = NULL)
```



What are we looking at?!

- Compute the matrix of correlations between the variables using the function `cor()`. You will need to exclude the `name` variable, which is qualitative.

```r
cor(subset(Auto, select=-name))
```

```
                    mpg  cylinders displacement horsepower      weig
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.83224
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.89752
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.93299
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.86453
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.00000
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.41683
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.30911
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.58500
             acceleration       year     origin
mpg             0.4233285  0.5805410  0.5652088
cylinders      -0.5046834 -0.3456474 -0.5689316
displacement   -0.5438005 -0.3698552 -0.6145351
horsepower     -0.6891955 -0.4163615 -0.4551715
weight         -0.4168392 -0.3091199 -0.5850054
acceleration    1.0000000  0.2903161  0.2127458
```

- Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and *all* other variables except `name` as the predictors. Print results.

```
fit.lm <- lm(mpg~.-name, data=Auto)
summary(fit.lm)
```

```
Call:
lm(formula = mpg ~ . - name, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

Refer to the output on the previous slide:

- Is there a relationship between the predictors and the response?

Refer to the output on the previous slide:

- Is there a relationship between the predictors and the response?
    - There is a relationship between predictors and response.

Refer to the output on the previous slide:

- Is there a relationship between the predictors and the response?
    - There is a relationship between predictors and response.
- Which predictors appear to have a statistically significant relationship to the response?

Refer to the output on the previous slide:

- Is there a relationship between the predictors and the response?
  - There is a relationship between predictors and response.
- Which predictors appear to have a statistically significant relationship to the response?
  - `weight`, `year`, `origin` and `displacement` have statistically significant relationships

Refer to the output on the previous slide:

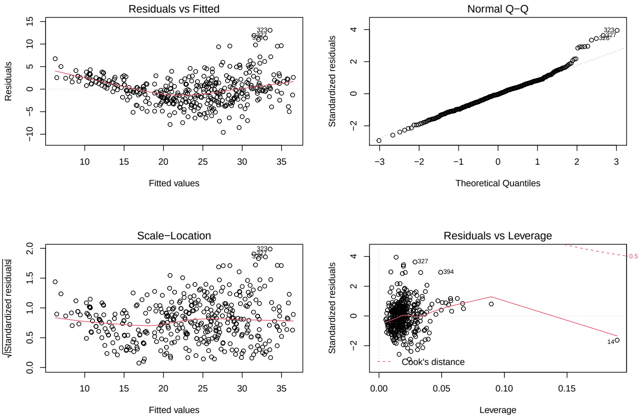
- Is there a relationship between the predictors and the response?
  - There is a relationship between predictors and response.
- Which predictors appear to have a statistically significant relationship to the response?
  - `weight`, `year`, `origin` and `displacement` have statistically significant relationships
- What does the coefficient for the `year` variable suggest?

Refer to the output on the previous slide:

- Is there a relationship between the predictors and the response?
  - There is a relationship between predictors and response.
- Which predictors appear to have a statistically significant relationship to the response?
  - `weight`, `year`, `origin` and `displacement` have statistically significant relationships
- What does the coefficient for the `year` variable suggest?
  - 0.75 coefficient for `year` suggests that later model year cars have better (higher) `mpg`.

- Use the ‘plot() function to produce diagnostic plots of the linear regression fit.

```
par(mfrow=c(2,2))
plot(fit.lm)
```



- Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers?
  - evidence of non-linearity
- Does the leverage plot identify any observations with unusually high leverage?
  - observation 14 has high leverage. . .

More info!

To view `lm {stats} R Documentation`, type:

```
help(lm)
```

and/or visit:

https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm

License

Contents of this lecture is based on the chapter 3 of the textbook Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *' An Introduction to Statistical Learning: with Applications in R'*.