9) Simple Linear Regression

Vitor Kamada

January 2018

Reference

Tables, Graphics, and Figures from

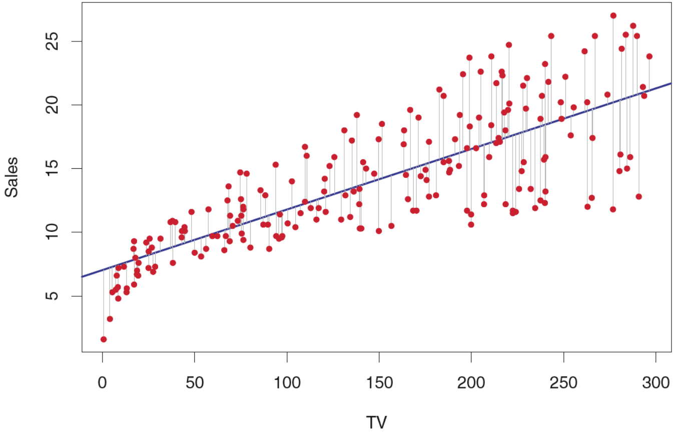
An Introduction to Statistical Learning

James et al. (2017): Chapters: 3.1, 3.6.1, 3.6.2

Simple Linear Regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 7.03 + 0.0475x$$



Residual Sum of Squares (RSS) $= e_1^2 + e_2^2 + ... + e_n^2$

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 = \sum_{i=1}^{n} e_i$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} [x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)] = 0$$

$$\sum_{i=1}^{n} x_i e_i = 0$$

Estimating the Coefficients

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

95% Confidence Interval

$$[\hat{\beta}_1 - 2SE(\hat{\beta}_1), \hat{\beta}_1 + 2SE(\hat{\beta}_1)]$$

CI for β_1 is [0.042, 0.053]

$$\hat{\beta}_0 \pm 2SE(\hat{\beta}_0)$$

CI for β_0 is [6.130, 7.935]

Standard Errors

$$SE(\hat{\beta}_0) = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}]$$

$$SE(\hat{\beta}_1) = \frac{\sigma^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\sigma^2 = Var(\epsilon)$$

$$\sigma = RSE = \sqrt{\frac{RSS}{n-2}}$$

Hypothesis Tests

$$H_0 : \beta_1 = 0 \text{ vs } H_a : \beta_1 \neq 0$$

$$t_{n-2} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

	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

Residual Standard Error

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

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

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

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

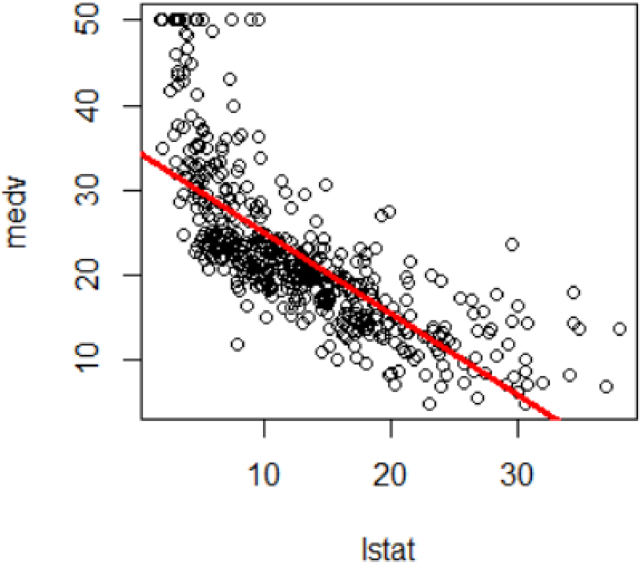
Boston Data Set

medv (median house value)

lstat (percent of households with low socioeconomic status)

Statistic	N	Mean	St. Dev.	Min	Max
crim	506	3.614	8.602	0.006	88.976
zn	506	11.364	23.322	0.000	100.000
indus	506	11.137	6.860	0.460	27.740
chas	506	0.069	0.254	0	1
nox	506	0.555	0.116	0.385	0.871
rm	506	6.285	0.703	3.561	8.780
age	506	68.575	28.149	2.900	100.000
dis	506	3.795	2.106	1.130	12.127
rad	506	9.549	8.707	1	24
tax	506	408.237	168.537	187	711
ptratio	506	18.456	2.165	12.600	22.000
black	506	356.674	91.295	0.320	396.900
lstat	506	12.653	7.141	1.730	37.970
medv	506	22.533	9.197	5.000	50.000

plot(lstat,medv); abline(lm.fit,lwd=3,col="red")



summary(lm.fit)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,	Adjusted R-squared:  0.5
432
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

confint(lm.fit)

```
                2.5 %     97.5 %
(Intercept) 33.448457 35.6592247
lstat       -1.026148 -0.8739505
```

95% Confidence and Prediction Interval

predict(lm.fit,data.frame(lstat=(c(5,10,15))),
interval="confidence")

```
       fit      lwr      upr
1 29.80359 29.00741 30.59978
2 25.05335 24.47413 25.63256
3 20.30310 19.73159 20.87461
```

predict(lm.fit,data.frame(lstat=(c(5,10,15))),
interval="prediction")

```
       fit       lwr      upr
1 29.80359 17.565675 42.04151
2 25.05335 12.827626 37.27907
3 20.30310  8.077742 32.52846
```

plot(predict(lm.fit), residuals(lm.fit))