Statistical modeling in logistics

MMS075

Lecture 2: Multiple linear regression

Department of Mechanics and Maritime Sciences
Division of Vehicle Safety

Acknowledgement: Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

András Bálint

Recommended resources

- Sections 3.2-3.4 in ISL and Sections 2.3 and 3.6 for R codes
- Sections 3.2-3.4 in the online course Statistical Learning
- Rougier, N.P., Droetboom, M., Bourne, P.E. (2014). Ten Simple Rules for Better Figures. PLoS Comput Biol. 10(9): e1003833, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4161295/
- Zook M, Barocas S, boyd d, Crawford K, Keller E, Gangadharan SP, et al. (2017) Ten simple rules for responsible big data research. PLoS Comput Biol 13(3): e1005399. https://doi.org/10.1371/journal.pcbi.1005399

Outline

- Simple linear regression continued
  - Revisiting Exercise 1
  - Testing relationship with response
  - Advertisement example & ethical aspects
- Multiple linear regression
  - Terminology
  - Relationship with response, variable significance
- Feedback

Simple linear regression continued

Revisiting Exercise 1 – reviewing relevant concepts in this context

Testing relationship with response – is there sufficient evidence of a relationship?

Advertisement example & ethical aspects – rules for good figures and ethical analysis

Recall exercise 1 background

A (hypothetical) very large company called Maintain-IT is responsible for a project task that needs to be repeated every year. They want to determine how the number of employees assigned to the project affects the completion time. An analyst at Maintain-IT decides to use simple linear regression to model this dependence, based on the observations shown in the table to the right.

Year	Employees on project	Completion time (days)
1	70	20
2	30	60
3	10	100
4	90	20

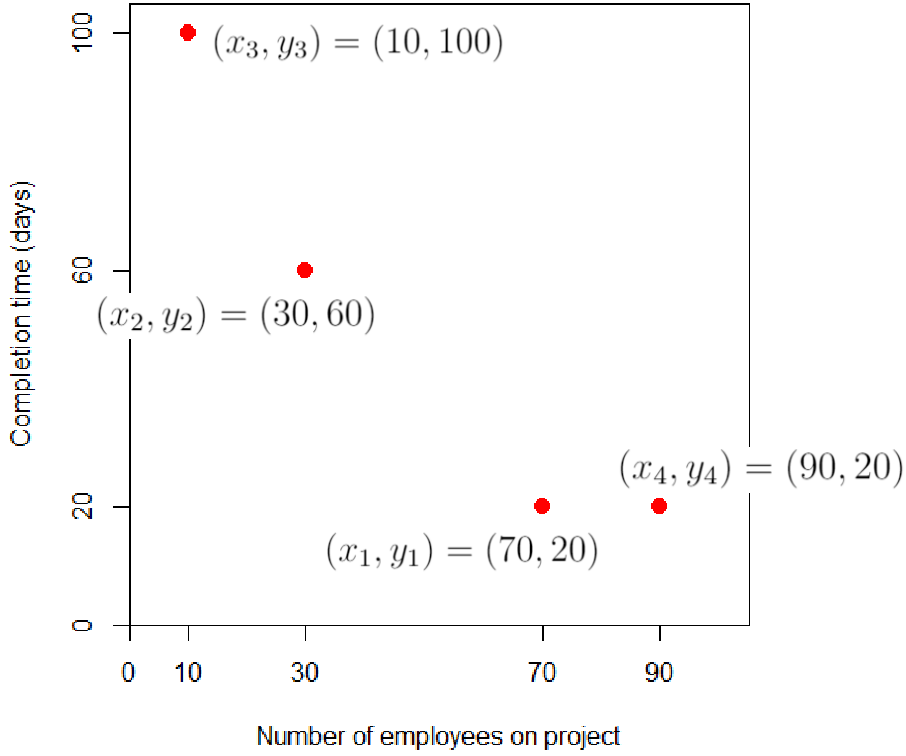
Step 1: specifying data points

Year	Employees on project	Completion time (days)
1	70	20
2	30	60
3	10	100
4	90	20

- The analyst wants to explain completion time by number of employees on project

→ The response, Y, is completion time, and the predictor, X, is the number of employees on project

- Data points in the sample are pairs of (number of employees, completion time) observed in years 1-4 (i.e., $n = 4$)

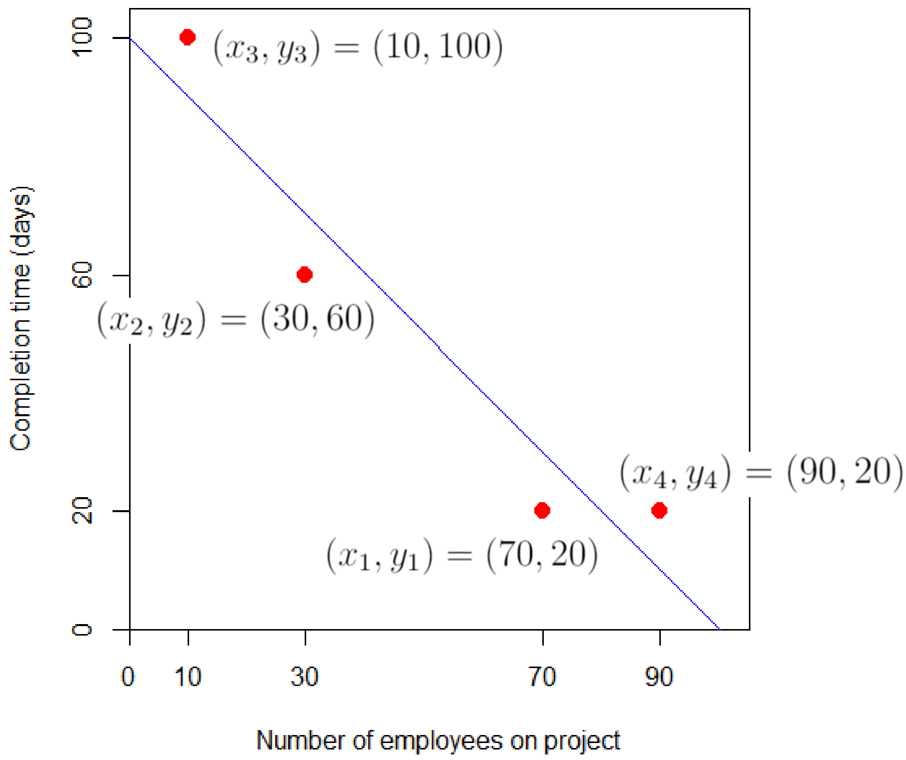


Step 2: fitting the least squares line

- Use formulas/software to determine the least squares coefficients:

$$\hat{\beta}_0 = 100$$
$$\hat{\beta}_1 = -1$$

- The least squares line is a line that crosses the y-axis at the value of $\hat{\beta}_0$, and has slope $\hat{\beta}_1$ indicating the change in the y-coordinate after a one-unit step to the right on the x-axis

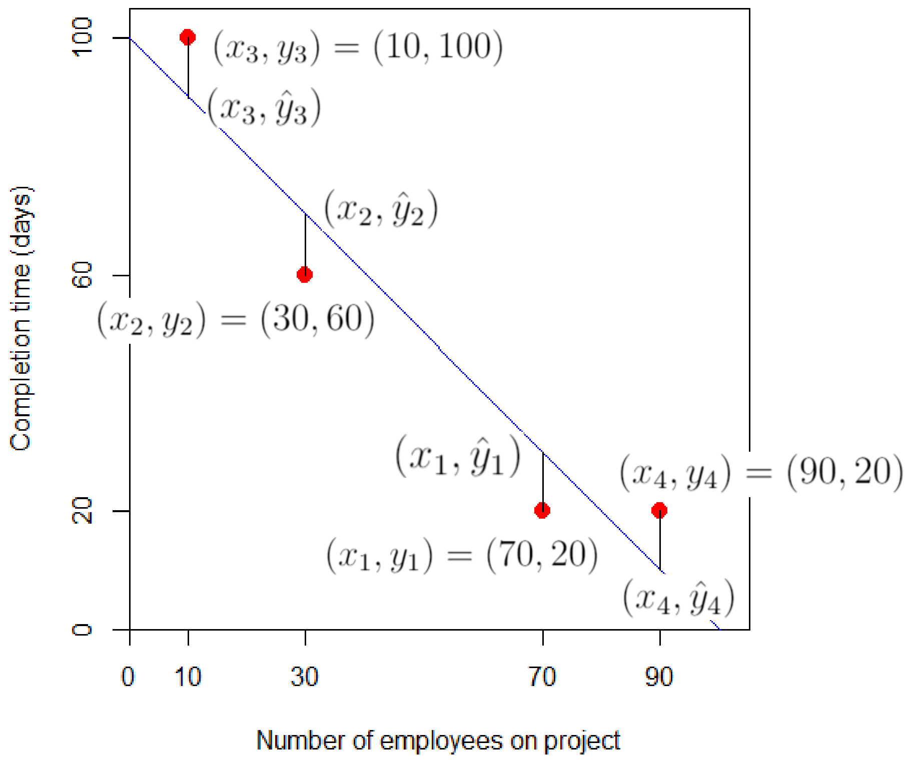


Step 3: predicted values for data points

- For each data value x_1, x_2, x_3, x_4 there is a prediction $\hat{y}_1, \hat{y}_2, \hat{y}_3, \hat{y}_4$ determined by the y-value of the point on the regression line at the given x-coordinate

- This is determined by substituting the x-values in the formula for the line:

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1 = 100 - 70 = 30$$
$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_2 = 100 - 30 = 70$$
$$\hat{y}_3 = \hat{\beta}_0 + \hat{\beta}_1 x_3 = 100 - 10 = 90$$
$$\hat{y}_4 = \hat{\beta}_0 + \hat{\beta}_1 x_4 = 100 - 90 = 10$$



Step 4: determining residuals and RSS

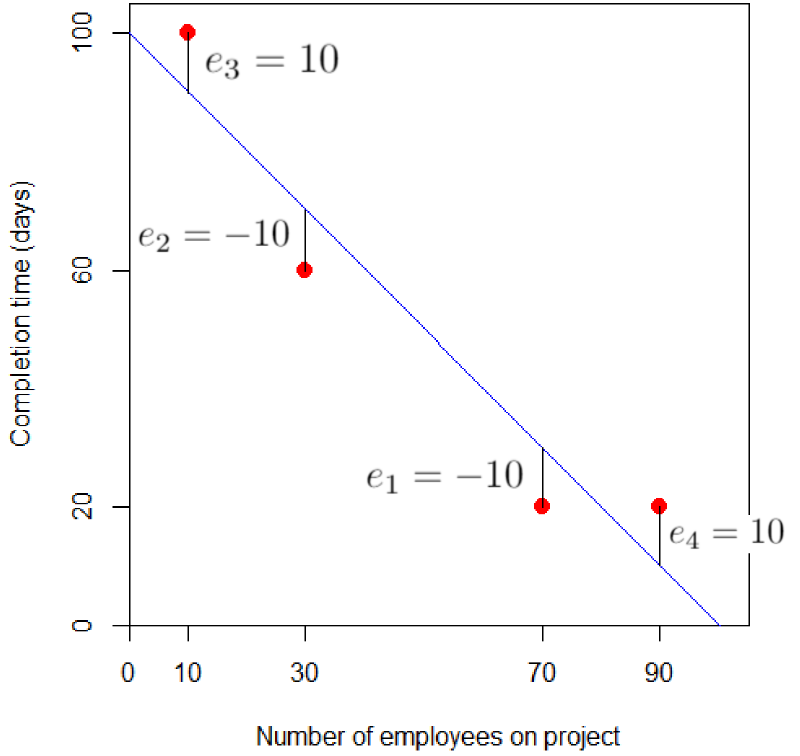
Differences between the observed and predicted responses are called residuals:

$$e_1 = y_1 - \hat{y}_1 = 20 - 30 = -10$$
$$e_2 = y_2 - \hat{y}_2 = 60 - 70 = -10$$
$$e_3 = y_3 - \hat{y}_3 = 100 - 90 = 10$$
$$e_4 = y_4 - \hat{y}_4 = 20 - 10 = 10$$

The sum of squared residuals is RSS:

$$\mathrm{RSS} = (-10)^2 + (-10)^2 + 10^2 + 10^2 = 400$$

(RSS is minimal for least squares line, i.e. sum of squared distances would be at least 400 for any other line)



Step 5: determining TSS & computing R^2

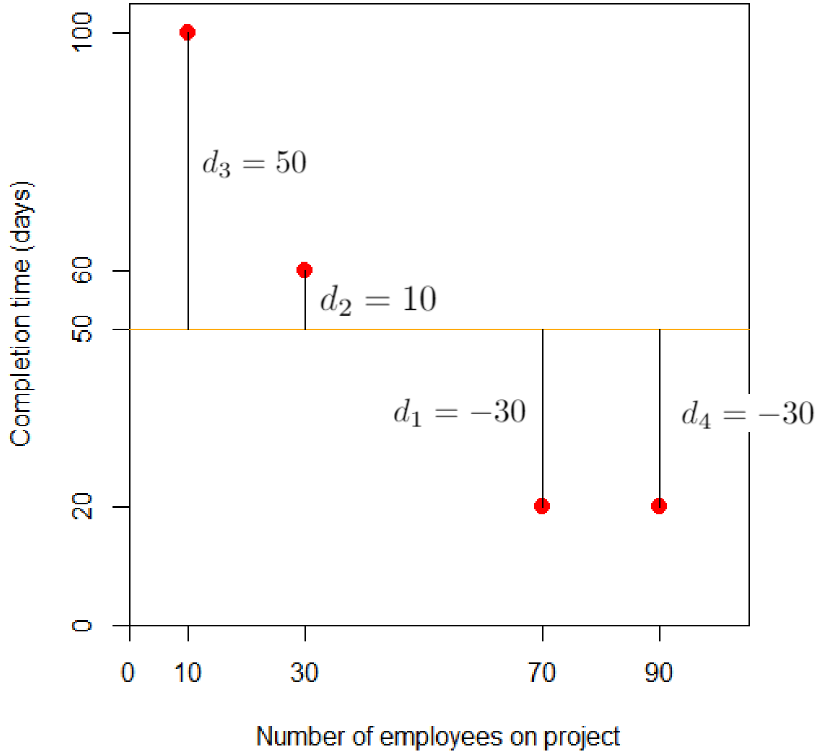
TSS measures the variation of responses compared to their average:

$$\begin{aligned} d_1 &= y_1 - \bar{y} = 20 - 50 = -30 \\ d_2 &= y_2 - \bar{y} = 60 - 50 = 10 \\ d_3 &= y_3 - \bar{y} = 100 - 50 = 50 \\ d_4 &= y_4 - \bar{y} = 20 - 50 = -30 \end{aligned}$$

The sum of squared differences is TSS:

$$\text{TSS} = (-30)^2 + (10)^2 + 50^2 + (-30)^2 = 4400$$

Have RSS & TSS → compute proportion of variation in completion time explained by the model: $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 0.9091$



Step 6: RSE & standard error of coefficients

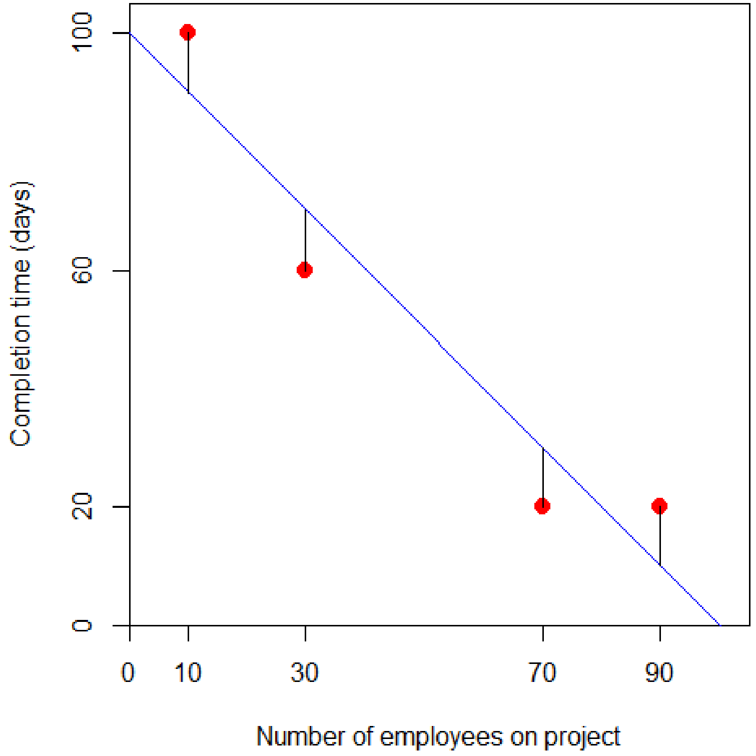
Residual standard error quantifies the lack of fit of the model:

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = 14.1421$$, i.e. 28% of $\bar{y}$

The standard errors of coefficients:

$$\mathrm{SE}(\hat{\beta}_0) = \mathrm{RSE} \times \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 13.2288$$

$$\mathrm{SE}(\hat{\beta}_1) = \mathrm{RSE} \times \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 0.2236$$

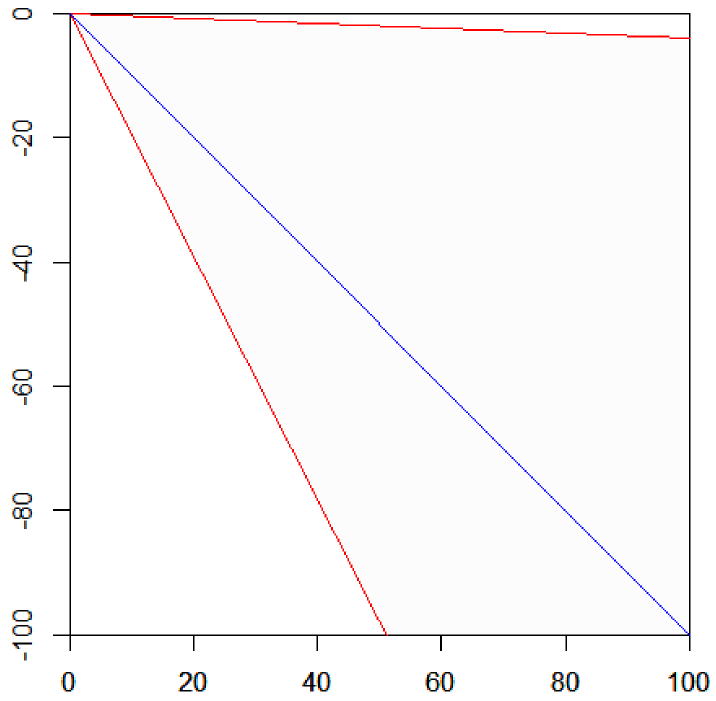


Step 7: Confidence intervals for coefficients

The 97.5% quantile of the t distribution with $n-2=2$ degrees of freedom is 4.3027, so the confidence intervals are:

$$\hat{\beta}_0 \pm 4.3027 \cdot \mathrm{SE}(\hat{\beta}_0) = 100 \pm 56.92$$
$$\hat{\beta}_1 \pm 4.3027 \cdot \mathrm{SE}(\hat{\beta}_1) = -1 \pm 0.96$$

Why wasn't the simple formula with 2 instead of 4.3027 good? Why are CIs so wide? Because of the small sample size ($n=4$), see next slides



The blue line has slope -1, like the least squares line, while the red lines have slopes corresponding to the endpoints of the confidence intervals for the slope, i.e. -1±0.96.

Effect of small sample size on CIs

- Recall from Lecture 1 that approximate confidence intervals are defined as follows:

$$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0)$$
$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

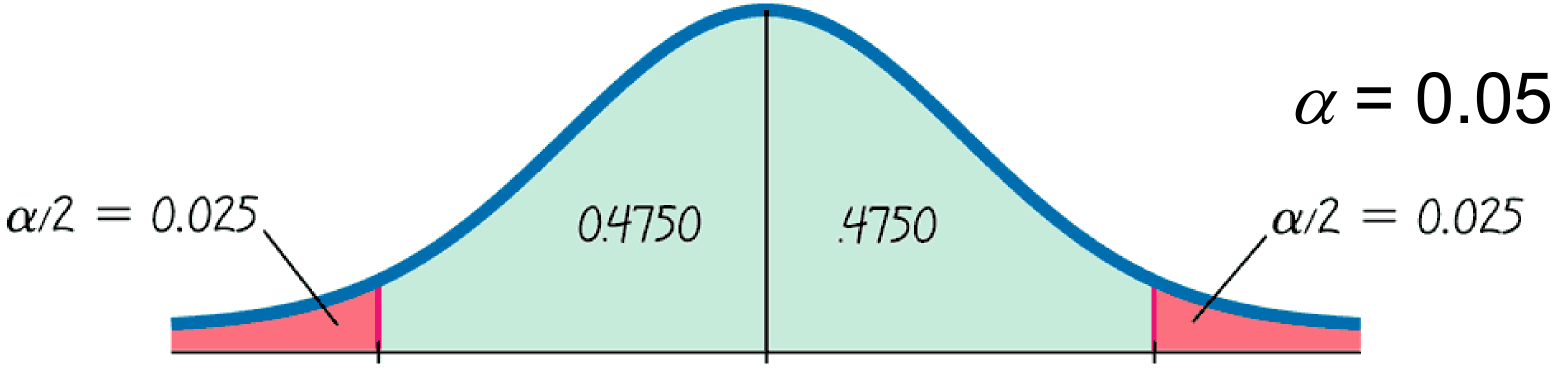
- This is almost like a 95% confidence interval for a standard normal variable – see next slide from SJO915

- We have a variable with Student's t distribution instead* with $n-2$ degrees of freedom – this is close to normal for large enough n

* Because we do not know the standard error of ε but rather had to estimate it by RSE

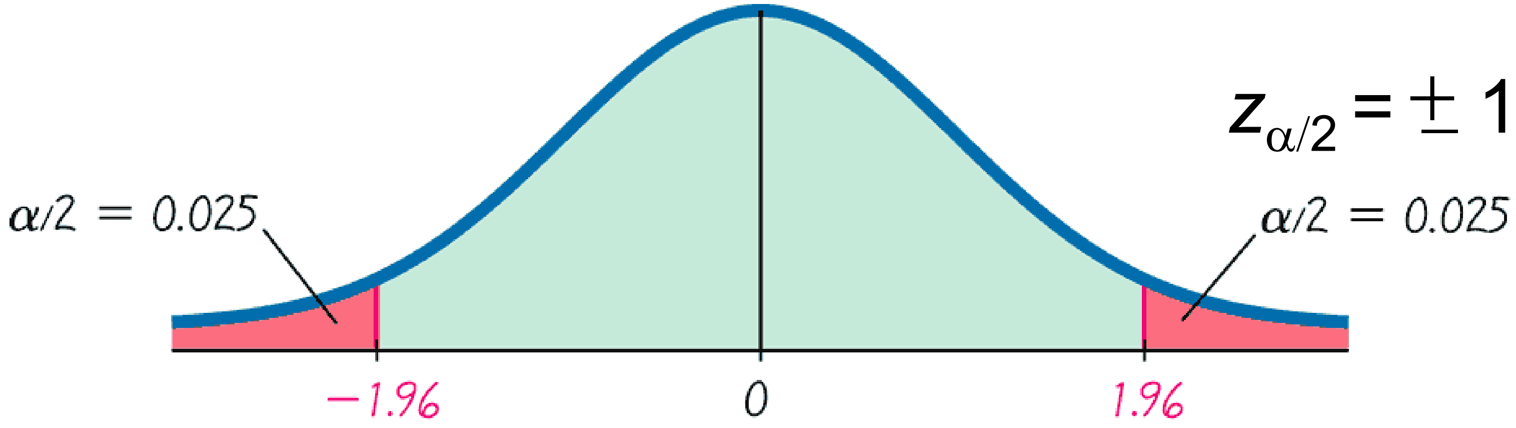
Finding $z_{\alpha/2}$ for a 95% Confidence Level

$\alpha = 0.05$

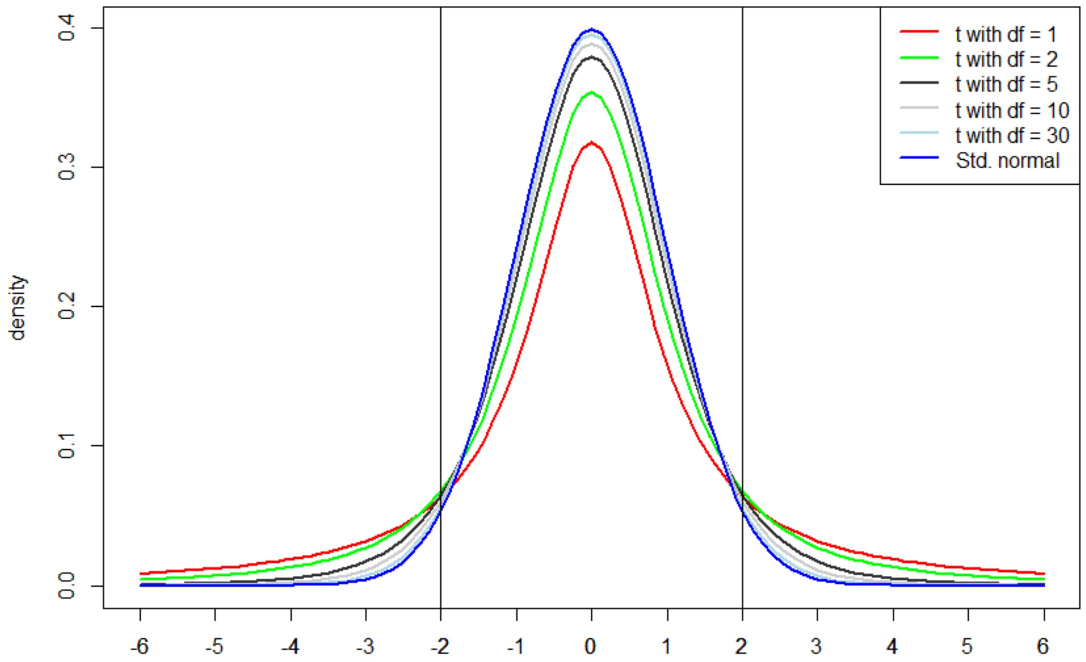


Use the standard table to find a *z*-score of 1.96

$z_{\alpha/2} = \pm 1.96$



t distribution is close to normal <u>for large df</u>



- 97.5% quantiles of the t distribution by degree of freedom:

df	1	2	3	4	5	10	20	30	100	1000	10000
97.5% quantile	12.71	4.30	3.18	2.78	2.57	2.23	2.09	2.04	1.98	1.96	1.96

Simple linear regression continued

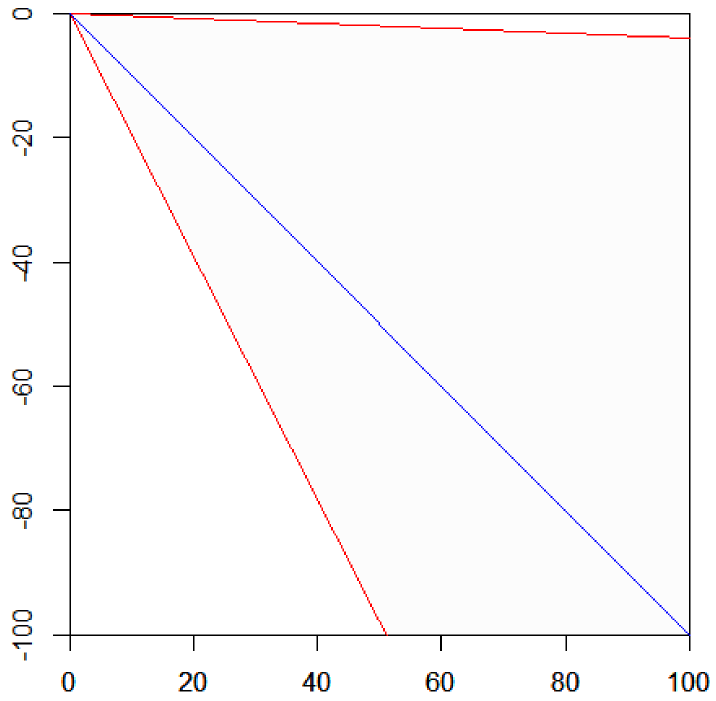
Revisiting Exercise 1 – reviewing relevant concepts in this context

Testing relationship with response – is there sufficient evidence of a relationship?

Advertisement example & ethical aspects – rules for good figures and ethical analysis

Slope CI as evidence of relationship

- Recall: in Excercise 1, the confidence interval for slope was -1±0.96 → we are quite confident that adding an extra person on the project reduces the completion time on average (by at least 0.04 days and at most 1.96 days)
- This is strong evidence of a relationship between the predictor and the response
- Can also do hypothesis testing – recall corresponding concept from SJO915:



The blue line has slope -1, like the least squares line, while the red lines have slopes corresponding to the endpoints of the confidence intervals for the slope, i.e. -1±0.96.

Key Concept

Individual components of a hypothesis testing:

- identify the **null hypothesis** and **alternative hypothesis** from a given claim, and how to express both in symbolic form
- identify the *Critical Value(s),* given a significance level
- calculate the value of *the test statistic*, given a claim and sample data
- identify the *P-value*, given a value of the test statistic
- state the conclusion about a claim in simple and nontechnical terms

p-value and corresponding decisions

- For a specific null hypothesis and test statistic, the corresponding p-value is the probability of having **at least as extreme values of the test statistic as the one observed**, **assuming that the null hypothesis is true**.

- If the p-value is very small*, that's strong evidence against the null hypothesis ➔ decision: **reject the null hypothesis**

- If the p-value is not very small, that's not enough evidence against the null hypothesis ➔ decision: **fail to reject the null hypothesis**

*Smaller than a significance level α which is typically chosen as 0.1, 0.05 (this is most common) or 0.01

Is there a relationship between X and Y?

- Recall the model equation: $Y = \beta_0 + \beta_1 X + \varepsilon$
- This describes a relationship indicating how Y changes as a result of changing X, **unless the coefficient β_1 is zero**
- If $\beta_1 = 0$, then the model equation reduces to $Y = \beta_0 + \varepsilon$, and there is indeed no relationship between X and Y
- How can we test whether this is the case?

Testing relevance of X in predicting Y

- The null and alternative hypotheses:

$$H_0 : \beta_1 = 0$$ ← No relationship is the null hypothesis – is there evidence against this?

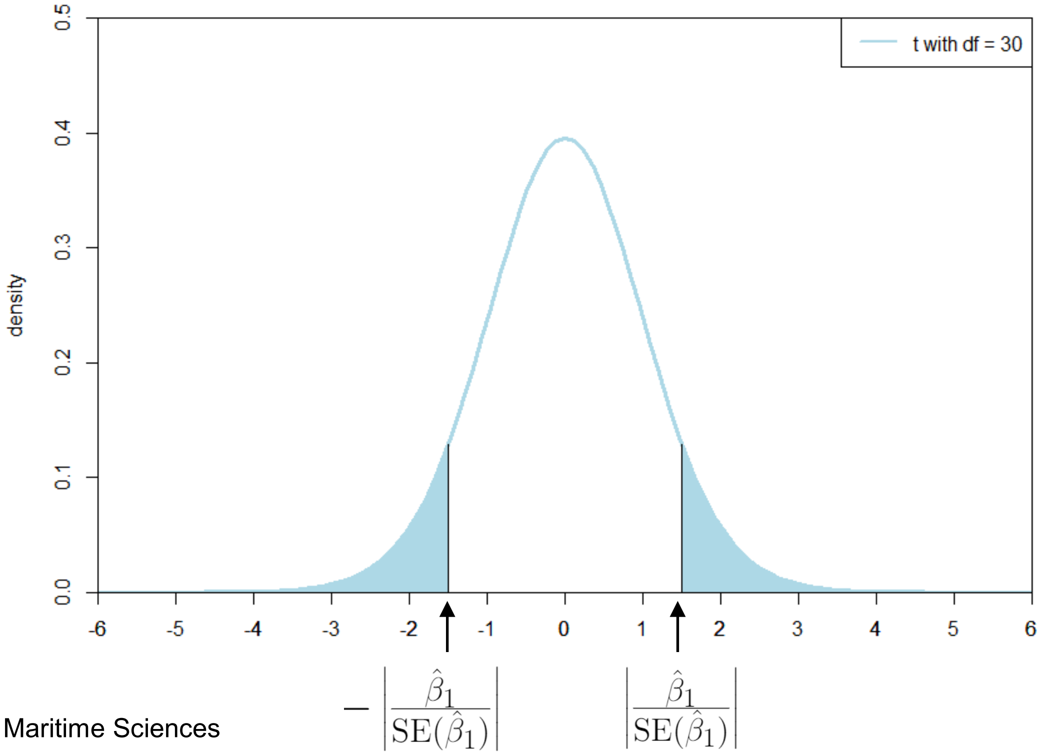
$$H_a : \beta_1 \neq 0$$

- Is the estimated slope large enough, compared to its standard error? If so, that would provide some evidence that the true slope parameter is different from 0

→ Test statistic: $t = \dfrac{\hat{\beta}_1}{\mathrm{SE}(\hat{\beta}_1)}$ ← If H_0 were true, then this statistic would have t distribution with $n-2$ degrees of freedom

Testing relevance of X in predicting Y (cont.)

- For p-value: which values of a t distribution with $n-2$ degrees of freedom are ”at least as extreme as the one observed”? It is those values that are at least as far from zero as the test statistic:



- Example with 30 degrees of freedom (corresponding to sample size 32) and the test statistic taking value -1.5
- Here the p-value is the sum of the shaded areas; here: 0.144
- This is not sufficiently small to reject the null hypothesis; for df = 30, test statistic values smaller than -2.04 or greater than 2.04 would be needed for rejection of H_0

Exercise 1 example

- Output from R: p-value of 0.0465 < 0.05 → decision: reject H_0, conclude: there is a relationship between number of employees on the project and completion time

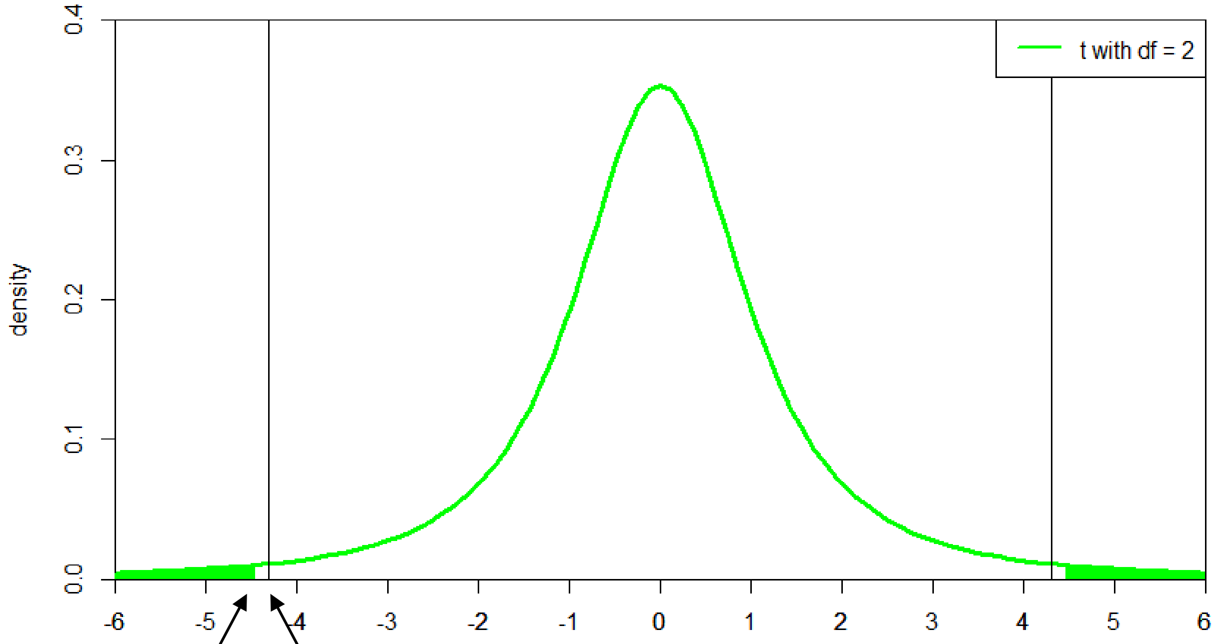
- **Important: p-value < 0.05 if and only if 95% confidence interval for the slope does not contain 0**

```
Call:
lm(formula = CompTime ~ Employees)

Residuals:
  1   2   3   4 
-10 -10  10  10 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 100.0000    13.2288   7.559   0.0171 *
Employees    -1.0000     0.2236  -4.472   0.0465 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.14 on 2 degrees of freedom
Multiple R-squared:  0.9091,	Adjusted R-squared:  0.8636 
F-statistic:    20 on 1 and 2 DF,  p-value: 0.04654
```



t-statistic: -4.472 < -4.3027, which is the 2.5% quantile of the t distribution with $n-2=2$ df

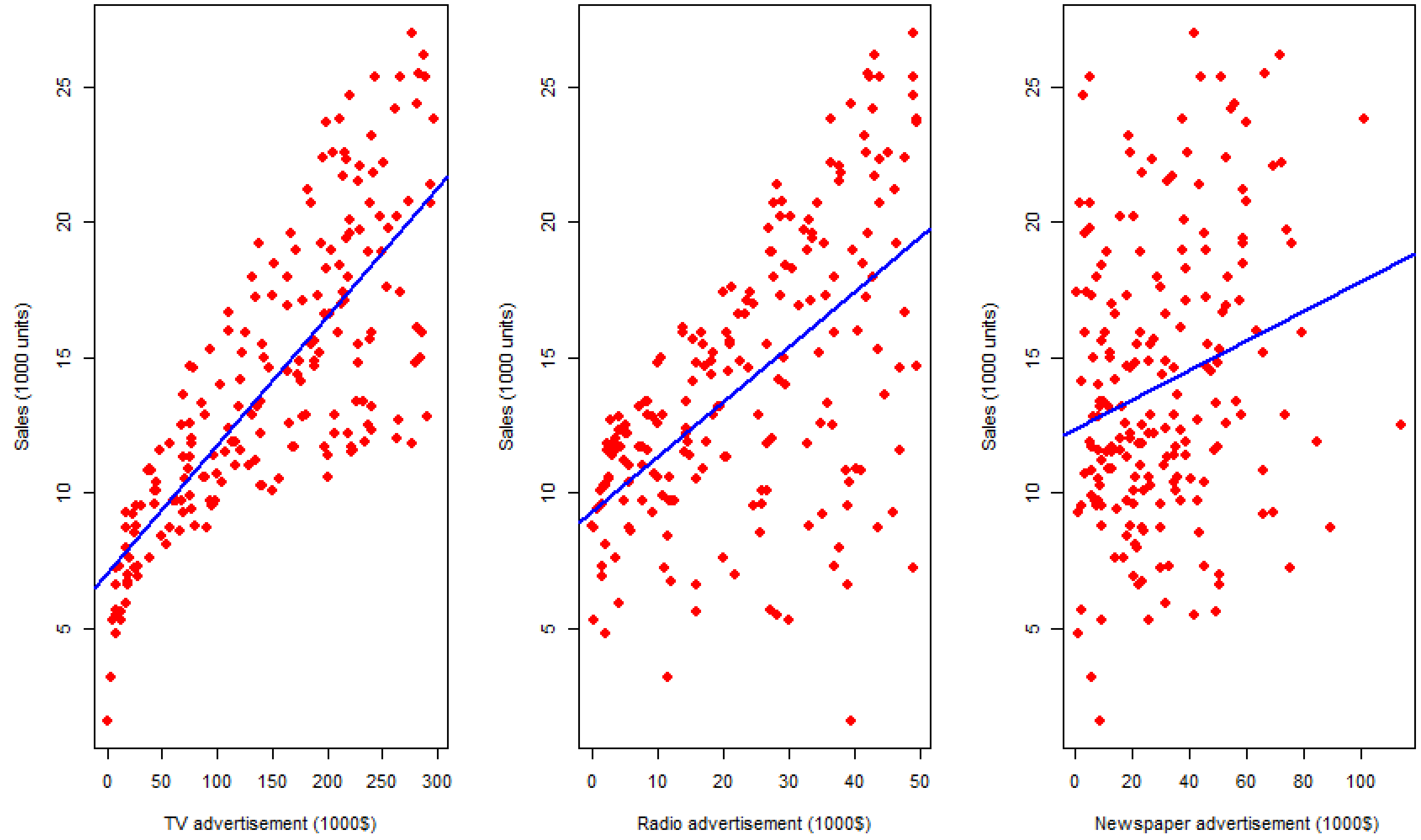
Simple linear regression continued

Revisiting Exercise 1 – reviewing relevant concepts in this context

Testing relationship with response – is there sufficient evidence of a relationship?

Advertisement example & ethical aspects – rules for good figures and ethical analysis

Advertising data, three separate LR models




```
Coefficients:
             Estimate
(Intercept) 7.032594
TV          0.047537

Coefficients:
             Estimate
(Intercept)  9.31164
radio        0.20250

Coefficients:
             Estimate
(Intercept) 12.35141
newspaper    0.05469
```

This is the largest slope by far – shouldn't regression line be steepest for radio??

Figure similar to: Figure 2.1 in ISL; figure created in R based on data downloaded from http://faculty.marshall.usc.edu/gareth-james/ISL/data.html

Same data, same x-axis range in all graphs

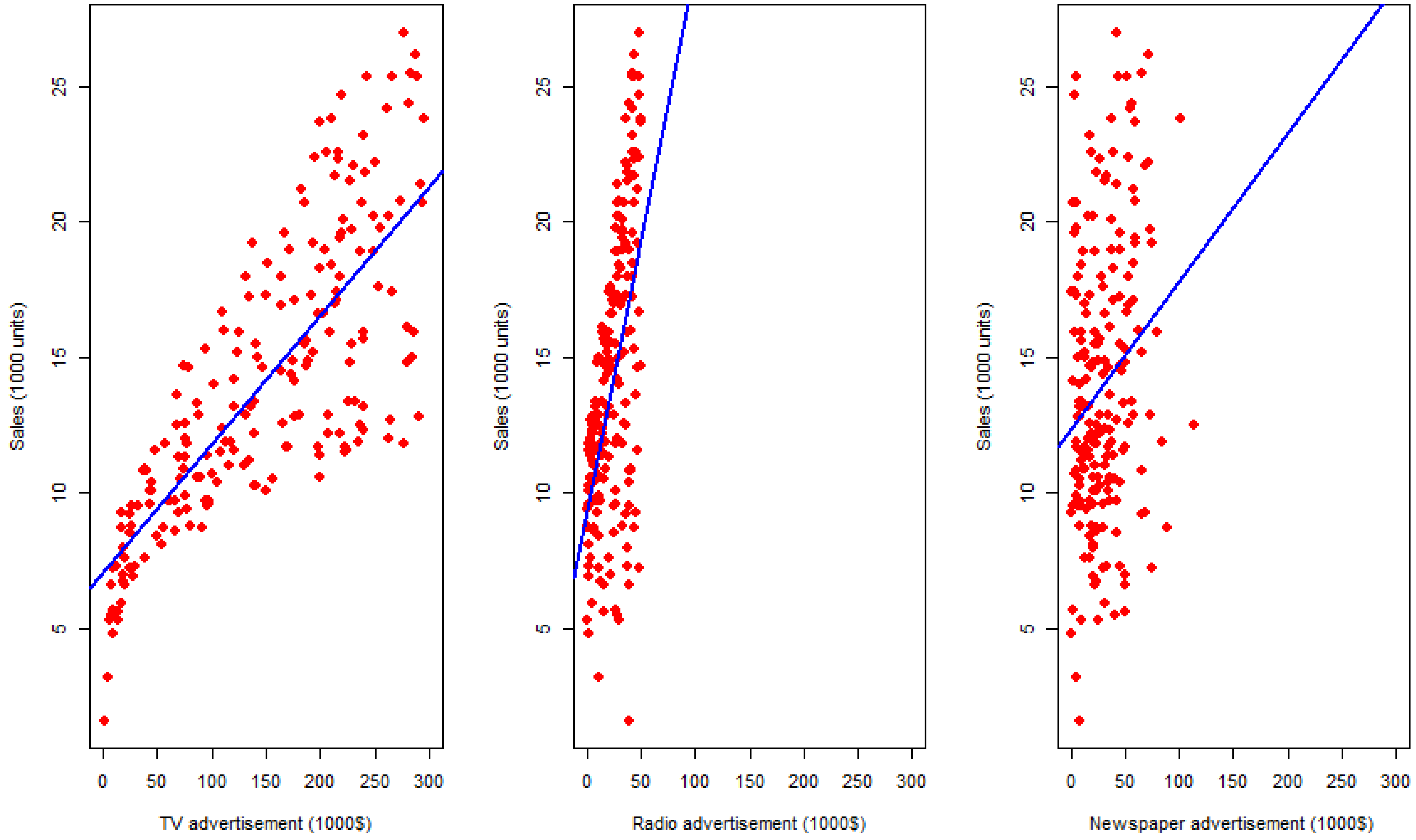


Figure created in R based on data downloaded from http://faculty.marshall.usc.edu/gareth-james/ISL/data.html

Is the first figure misleading?

Check the 10 rules described in Rougier et al. (2014)*. Are any of these recommendations violated? Consider individual figures and the three figures side-by-side.

When is it fine to use the first figure? Why? Vote at www.menti.com (code: 49 82 50) & discuss!

Check also "Ten simple rules for responsible big data research"** for a detailed discussion of ethical aspects.

*Rougier, N.P., Droetboom, M., Bourne, P.E. (2014). Ten Simple Rules for Better Figures. PLoS Comput Biol. 10(9): e1003833, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4161295/
** Zook M, Barocas S, boyd d, Crawford K, Keller E, Gangadharan SP, et al. (2017) Ten simple rules for responsible big data research. PLoS Comput Biol 13(3): e1005399. https://doi.org/10.1371/journal.pcbi.1005399

Multiple linear regression

Terminology (e.g. Slope, Intercept, RSS, TSS, RSE, R^2)

Relationship between predictors and response, variable significance

Advertising example: all media in same model

- We constructed three separate models to understand the effect of each advertisement form on sales

- A better approach: include all predictors in the same model!

- Give each predictor a separate slope, add a common intercept:

$$\text{sales} \approx \beta_0 + \beta_{TV} \times \text{TV} + \beta_R \times \text{radio} + \beta_N \times \text{newspaper}$$

Multiple linear regression: model definition

- Outcome linearly depends on $p \geq 2$ predictors:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Response
Output variable
Dependent variable

Coefficients
Parameters

Predictors
Features
Input variables
Independent variables

Error term

- All the above terms are used regularly in different contexts

Multiple linear regression: model definition

- Outcome linearly depends on $p \geq 2$ predictors:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

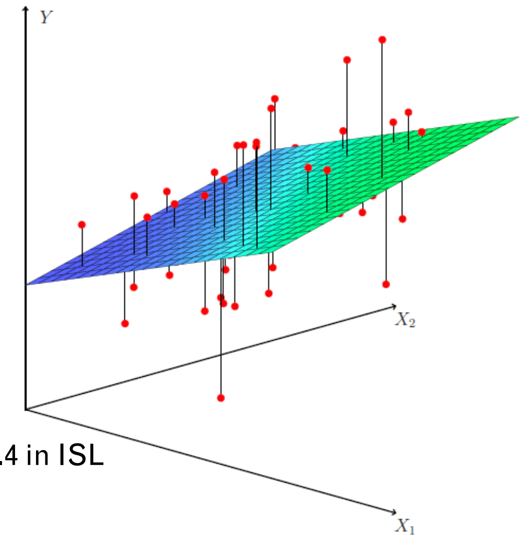
Intercept **Slope of X_1** **Slope of X_2** **Slope of X_p**

Coefficients
Parameters

Residuals & RSS

- Observed data: n observations containing values for each of the p predictors and the response:

$$(x_{1,1}, x_{1,2}, \ldots, x_{1,p}, y_1)$$
$$(x_{2,1}, x_{2,2}, \ldots, x_{2,p}, y_2)$$
$$\vdots$$
$$(x_{n,1}, x_{n,2}, \ldots, x_{n,p}, y_n)$$



Source:
Figure 3.4 in ISL

- For fixed coefficients, define residuals & residual sum of squares:

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \hat{\beta}_2 x_{i,2} - \ldots - \hat{\beta}_p x_{i,p}$$

$$\mathrm{RSS} = e_1^2 + e_2^2 + \ldots + e_n^2$$

Want: this is small

Coefficients from software output

- **Least squares coefficient estimates:** parameter values that minimize RSS; we get them from computer software

- Available in different software, R output is shown

```
Call:
lm(formula = AdData$sales ~ AdData$TV + AdData$radio + AdData$newspaper)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8277 -0.8908  0.2418  1.1893  2.8292 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       2.938889   0.311908   9.422   <2e-16 ***
AdData$TV         0.045765   0.001395  32.809   <2e-16 ***
AdData$radio      0.188530   0.008611  21.893   <2e-16 ***
AdData$newspaper -0.001037   0.005871  -0.177     0.86    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956 
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Interpretation for advertisement model

In the best linear model, $\beta_0 = 2.939, \beta_{TV} = 0.046, \beta_R = 0.189, \beta_N = -0.001$

→ $\text{sales} \approx 2.939 + 0.046 \times \text{TV} + 0.189 \times \text{radio} - 0.001 \times \text{newspaper}$

What do we learn from this?

1. Setting all predictors to 0 results in Sales=2.939 → without any advertisements, we would sell about 3000 units
2. If we increase ”TV” by 1 **while holding ”radio” and ”newspaper” constant**, then ”sales” increases by 0.046 → if we decided to invest 1000$ more in TV advertisements and the held radio and newspaper budgets fixed, we could expect to increase our sales by about 46 units
3. Having TV and radio in the model, newspaper has no added value on sales

Accuracy of the model – RSE

- Residual standard error, which is an estimate of the standard deviation of ε:

$$\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{RSS}}$$

- This is the average amount that the response deviates from the true regression line → measures the lack of fit of the model

- This is an absolute measure. For advertising example, RSE=1.686 → even if we knew the true parameter values, our prediction of sales may be off by 1686 units. Is this OK or too much?

Accuracy of the model – R^2

Measures the **proportion of variability in the response that is explained by the linear model including all predictors**:

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

where TSS is the **total sum of squares** and RSS is the residual sum of squares (as defined before):

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

$$\mathrm{RSS} = \Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2 = \Sigma_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \hat{\beta}_2 x_{i,2} - \ldots - \hat{\beta}_p x_{i,p})$$

Properties of R^2

- This is a proportion → its value is always between 0 and 1, larger R^2 indicates better fit of the model (but ”good enough” values depend on the application)

- Values close to 0 may indicate:
  - Linear model is wrong
  - Inherent error is high

  and/or

- Link between R^2 and correlation in the multiple variable case:
  $R^2 = \mathrm{Cov}(Y, \hat{Y})^2$

Multiple linear regression

Terminology (e.g. Slope, Intercept, RSS, TSS, RSE, R^2)

Relationship between predictors and response, variable significance

Testing relationsip between response and all predictors (model significance)

- Test null hypothesis that all coefficients are zero (and hence this collection of predictors has no relationship to response):

$$H_0 : \beta_1 = \beta_2 = \ldots \beta_p = 0$$

No relationship is the null hypothesis – is there evidence against this?

$$H_a : \text{at least one } \beta_j \text{ is not zero}$$

- Test statistic: $F = \frac{(\text{TSS}-\text{RSS})/p}{\text{RSS}/(n-p-1)}$

- Large values of F are evidence of a relationship

Testing relevance of X_j in predicting Y in the presence of all other predictors

- The null and alternative hypotheses:

$$H_0 : \beta_j = 0$$ ⟵ No relationship in the presence of all other predictors is the null hypothesis – is there evidence against this?

$$H_a : \beta_j \neq 0$$

- Is the estimated slope large enough, compared to its standard error? If so, that would provide some evidence that the true slope parameter is different from 0

→Test statistic: $t = \dfrac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)}$

Feedback

Feedback quiz

Feedback is essential to me so that I can improve the lectures during the course. Please take your chance to optimize your learning experience!

If you are willing to give feedback, please follow these steps:

1. Go to www.menti.com
2. Enter the code 48 61 22
3. Rate your experience today in slide 1
4. Wait until I change slide
5. Answer to the questions in slide 2