Correlation strength table

0.91 – 1.0	→	very strong.
0.71 – 0.9	→	strong.
0.51 – 0.7	→	Medium.
0.31 – 0.50	→	Low
0.01 – 0.30	→	very Low.

CHAPTER 8

Least Square Regression

Introduction

When a scatterplot shows a linear relationship between a quantitative explanatory variable *x* and a quantitative response variable *y*, we can use the least-squares line fitted to the data to predict *y* for a given value of *x*. If the data are a random sample from a larger population, we need statistical inference to answer questions like these:

- ✓ Is there really a linear relationship between *x* and *y* in the population, or could the pattern we see in the scatterplot plausibly happen just by chance?
- ✓ What is the slope (rate of change) that relates *y* to *x* in the population, including a margin of error for our estimate of the slope?
- ✓ If we use the least-squares regression line to predict *y* for a given value of *x*, how accurate is our prediction (again, with a margin of error)?

The LSRL was defined as:

- The slope and intercept of the least-squares line are *statistics* and are calculated from sample data.
- These statistics would take somewhat different values if we repeated the data production process.

Now we are going to think about the LSRL computed from a sample as an estimate of a true regression line for the population.

- Population line: $\beta_0 + \beta_1 x$.
- To do inference, think of b_0 and b_1 as estimates of unknown parameters β_0 and β_1 that describe the population of interest.

Conditions for Regression inference

Conditions for Regression Inference

To use the least-squares line as a basis for inference about a population, each of the following conditions should be approximately met:

- The sample is SRS from the population.
- There is a linear relationship between x and y.
- The standard deviation of the responses y about the population regression line is the same for all x.
- The model deviations are Normally distributed.

Simple Linear Regression Model

Given n observations of the explanatory variable x and the response variable y. The **statistical model for simple linear regression** states that the observed response y_i when the explanatory variables takes the value x_i is :

$$DATA = FIT + RESIDUAL$$

$$y_i = (\beta_0 + \beta_1 x_i) + \varepsilon_i$$

Here, $\beta_0 + \beta_1 x_i$ is the mean response when $x = x_i$. The deviation ϵ_i are assumed to be independent and normally distributed with mean 0 and standard deviation σ.

Estimate the Regression Parameters

The intercept β_0, the slope β_1, and the standard deviation σ of y are the unknown parameters of the population regression line. We can use random sample data to provide unbiased estimates of these parameters.

- The least-squares regression line $\hat{y} = b_0 + b_1 x$ obtained from sample data is the best estimate of the true population regression line $\mu_y = \beta_0 + \beta_1 x$.
- The value of $\hat{y}$ from the least-squares regression line is really a prediction of the mean value of y (μ_y) for a given value of x.

Estimating Model Standard Deviation

From the LSRL the predicted values are denoted as $\hat{y}_i$ and the actual values are y_i, then the residuals are defined as:

$$e_i = y_i - \widehat{y_i} = y_i - b_0 - b_1 x_i$$

The estimate of the model standard deviation (σ) is given by the **regression standard error, *(s)*:**

$$s = \sqrt{\frac{\sum e_i^2}{n-2}} = \sqrt{\frac{\sum (y_i - \widehat{y_i})^2}{n-2}}$$

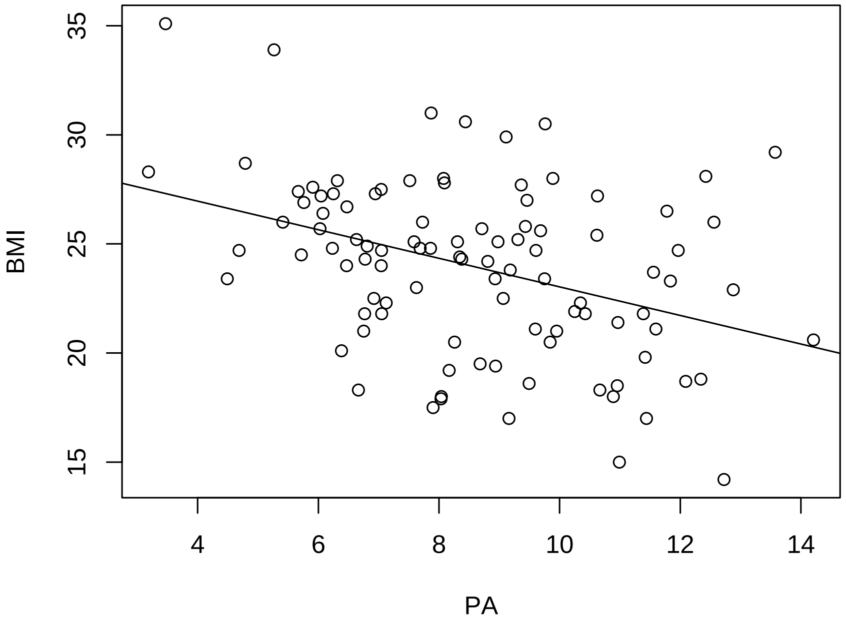
Example

Relationship between Body mass index(BMI) and Physical Activity.

Response variable: Body mass index (BMI)

Explanatory variable: Physical activity (PA) – measured with a pedometer

Consider a SRS of 100 female undergraduates



Example Cont.

```
➢ model <- lm(BMI~PA, data = dat)
➢ summary(model)

Call:
lm(formula = BMI ~ PA, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.3819 -2.5636  0.2062  1.9820  8.5078 

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)    
(Intercept)  29.5782     1.4120  20.948  < 2e-16 ***
PA           -0.6547     0.1583  -4.135  7.5e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 3.655 on 98 degrees of freedom
Multiple R-squared:  0.1485, Adjusted R-squared:  0.1399
F-statistic:  17.1 on 1 and 98 DF,  p-value: 7.503e-05
```

1) Write the equation of the least- square regression line.

2) What is the predicted BMI for a female college student who averages 8000 steps per day?

3) If her actual BMI is 25.655 what would the residual be?

Confident Intervals for Regression Slope

Confidence Interval for Regression Slope

A level C **confidence interval for the slope β_1** of the population regression line is:

$$\boldsymbol{b_1 \pm t^* \mathrm{SE}_{b1}}$$

Here t^* is the critical value for the t distribution with df = $n - 2$ having area C between $-t^*$ and t^*.

Example Cont.

Compute the 95% confidence interval for β_1 for BMI and PA.

```
Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  29.5782     1.4120  20.948  < 2e-16 ***
PA           -0.6547     0.1583  -4.135  7.5e-05 ***
---
```

Significance Test for Regression Slope

We may look for evidence of a **significant relationship** between variables x and y in the population from which our data were drawn.

For that, we can test the hypothesis that the regression slope parameter β is equal to zero.

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

Testing H_0: $\beta_1 = 0$ is equivalent to testing the **hypothesis of no correlation** between x and y in the population.

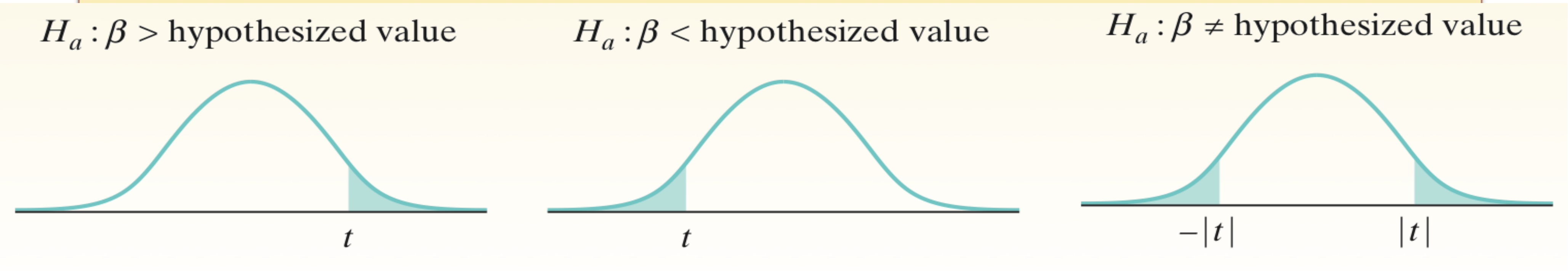
Significance Test for Regression Slope

Significance Test for Regression Slope

To test the hypothesis H_0: β_1 = hypothesized value, compute the test statistic:

$$t = \frac{b_1 - \text{hypothesized value}}{SE_{b_1}}$$

Find the P-value by calculating the probability of getting a t statistic this large or larger in the direction specified by the alternative hypothesis H_a. Use the t distribution with df = n – 2.



Example Cont.

Use significance test to check if there is a linear relationships between PA and BMI.

```
➤ model <- lm(BMI~PA, data = dat)
➤ summary(model)

Call:
lm(formula = BMI ~ PA, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.3819 -2.5636  0.2062  1.9820  8.5078 

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)    
(Intercept)  29.5782     1.4120  20.948  < 2e-16 ***
PA           -0.6547     0.1583  -4.135  7.5e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 3.655 on 98 degrees of freedom
Multiple R-squared:  0.1485, Adjusted R-squared:  0.1399 
F-statistic:  17.1 on 1 and 98 DF,  p-value: 7.503e-05
```

Analysis of Variance for Regression

The regression model is:

Data = Fit + Error

$$y_i = (\beta_0 + \beta_1 X_i) + (\varepsilon_i)$$

It resembles an ANOVA, which also assumes equal variance, where

SST = SSM + SSE

DFT = DFM + DFE

The ANOVA F Test

1) For a simple linear relationship, the ANOVA tests the hypotheses

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

2) Test statistic; F = MSM/MSE

3) When H_0 is true, F follows the $F(1, n - 2)$ distribution. The P-value is $P(F \geq f)$.

Note: The ANOVA test and the two-sided t-*test for* H_0: $\beta_1 = 0$ *yield the same P-value*

4) conclusion

The ANOVA Table

Source	Sum of squares SS	DF	Mean square MS	*F*	*P*-value
Model	$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	1	*MSM=SSM/DFM*	*MSM/MSE*	Tail area above F
Error	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	*n* − 2	*MSE=SSE/DFE*		
Total	$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$	*n* − 1			

SST = SSM + SSE

DFT = DFM + DFE

F=MSM/MSE

Example Cont.

Use significance test to check if there is a linear relationships between PA and BMI.

> anova(model)

Analysis of Variance Table

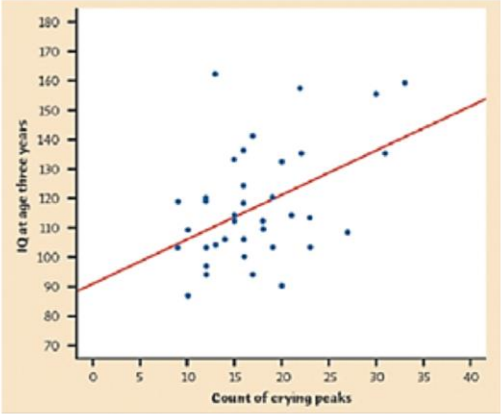
Response: BMI

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
PA	1	228.38	228.377	17.096	7.503e-05
Residuals	98	1309.10	13.358		

Practice Problem 1

Infants who cry easily may be more easily simulated than others. This may be a sign of higher IQ. Child development researchers explored the relationship between the crying of infants 4 to 10 days old and their later IQ test scores. A scatterplot and Minitab output for the data from a random sample of 38 infants is below.

1) write the equation for the LSRL.



Regression Analysis: IQ versus Crycount

Predictor	Coef	SE Coef	T	P
Constant	91.268	8.934	10.22	0.000
Crycount	1.4929	0.4870	3.07	0.004

S = 17.50 R-Sq = 20.7% R-Sq(adj) = 18.5%

Practice Problem Cont.

Regression Analysis: IQ versus Crycount

Predictor	Coef	SE Coef	T	P
Constant	91.268	8.934	10.22	0.000
Crycount	1.4929	0.4870	3.07	0.004

S = 17.50 R-Sq = 20.7% R-Sq(adj) = 18.5%

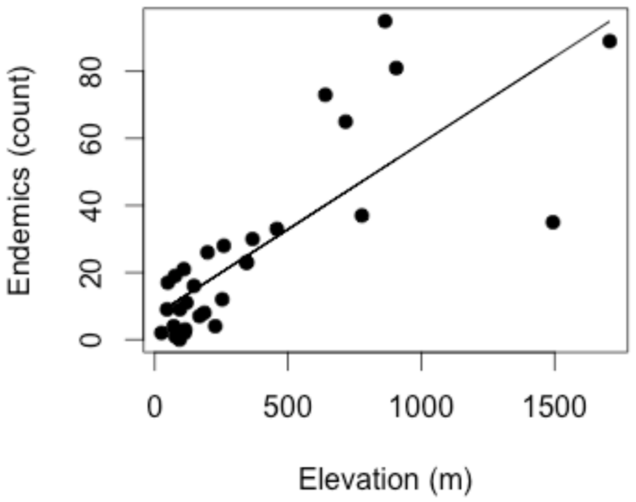
3) Calculate the 95% confidence interval for the slope ($t^* = 2.028$)

3) Perform a hypothesis test to determine if cry count is significant.

Practice Problem 2

Consider the following data set labeled Gala, which describe the number of species of turtles on the various Galapagos Islands. There are 30 cases and 7 variables in the dataset. In the following analysis, we consider the linear relationship between Elevation and Endemics.

1) What is the explanatory and response variable.



Practice Problem 2 Cont.

2) Use the RStudio below to perform a hypothesis test for the slope parameter.

```
> turtle.reg = lm(gala$Endemics ~ gala$Elevation)
> summary(turtle.reg)

Call:
lm(formula = gala$Endemics ~ gala$Elevation)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.976  -8.799  -2.133   7.453  43.407 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    7.182682   4.138088   1.736   0.0936 .  
gala$Elevation 0.051401   0.007465   6.886 1.75e-07 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.95 on 28 degrees of freedom
Multiple R-squared:  0.6287,	Adjusted R-squared:  0.6154 
F-statistic: 47.41 on 1 and 28 DF,  p-value: 1.751e-07
```

Practice Problem 2 Cont.

3) State and interpret the meaning of the coefficient of determinate .

4) Provide the 95% CI for the slope.

5) Write the equation for the LSRL and predict the Endemics of 500 meters.

Practice Problem 2 Cont.

6) Use the RStudio below to perform a Significance F test.

```
> anova(turtle.reg)
Analysis of Variance Table

Response: gala$Endemics
               Df  Sum Sq Mean Sq F value    Pr(>F)    
gala$Elevation  1 13619.3 13619.3   47.41 1.751e-07 ***
Residuals      28  8043.4   287.3                      
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```