Contents

Chapter 3

Simple Linear Regression

In this chapter, the mean of a population is not a constant and depends on the value of another variable. We use the sample data to estimate the straight-line relationship between the mean value of one quantitative variable, y, as it relates to a second quantitative variable, x.

Regression Analysis

Regression analysis is the process of finding a mathematical model or equation that best fits the sample data.

Dependent or Predicted or Response Variable, y: is the variable that we would like to predict.

Independent or Predictor or Explanatory Variable, x: is the variable that we use to predict y.

General Form of Probabilistic Model in Regression

$$y = E(y) + \epsilon$$

where:

$y =$ **Depndent Variable**

$E(y) =$ Mean (or expected) value of y

$\epsilon =$ Unexplainable, or random error which represents all unexplained variations in the depndent variable, y that caused by important, but unincluded, variables or by random phenomena.

3.1 Introduction

We start with the simplest of probabilistic models—the **straight-line model** using method of **least squares**.

This chapter is about **a simple linear regression analysis** which we can apply to:

- Determine the existence of a relationship between y and x.
- Estimate $E(y)$ using the model.
- Predict a future value of y for a given value of x.

3.2 The Straight-Line Probabilistic Model

A First-Order (Straight-Line) Probabilistic Model

$$y = \beta_0 + \beta_1 x + \epsilon$$

where

$y =$ Dependent or response variable (quantitative variable to be modeled or predicted)

$x =$ Independent or predictor variable (quantitative variable used as a predictor of y)

$\beta_0 + \beta_1 x = E(y) =$ Deterministic component

ϵ(epsilon)= Random error component

β_0= y-intercept of the line—that is, the point at which the line intersects, or cuts through, the y-axis

β_1= Slope of the line—that is, the change (amount of increase or decrease) in the deterministic component of y for every one-unit increase in x.

Note:

A positive slope implies that $E(y)$ increases by the amount β_1. A negative slope implies that $E(y)$ decreases by the amount β_1.

Standard Assumption: The mean value of the random error equals 0. $E(\epsilon) = 0$

See the following figure:

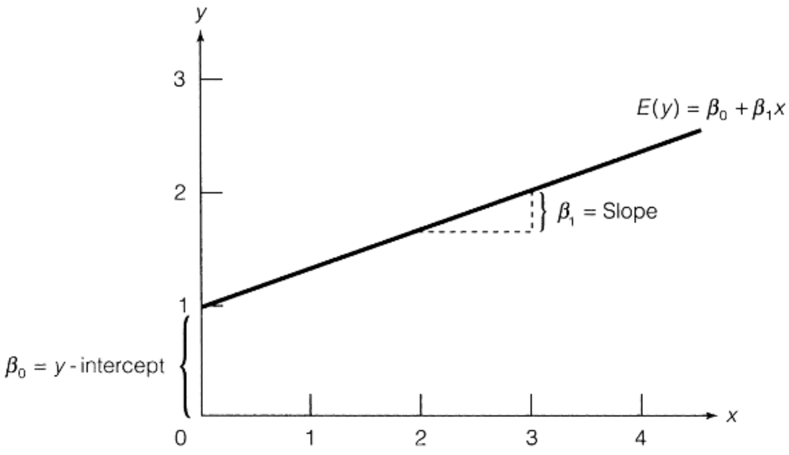


Figure 3.1: The straight-line model

Steps in Regression Analysis

1. Hypothesize the form of the model for $E(y)$ (using a scatterplot).

2. Collect the sample data.

3. Estimate unknown parameters in the model using the sample data.

4. Check the validity assumptions.

5. Statistically evaluate the usefulness of the model.

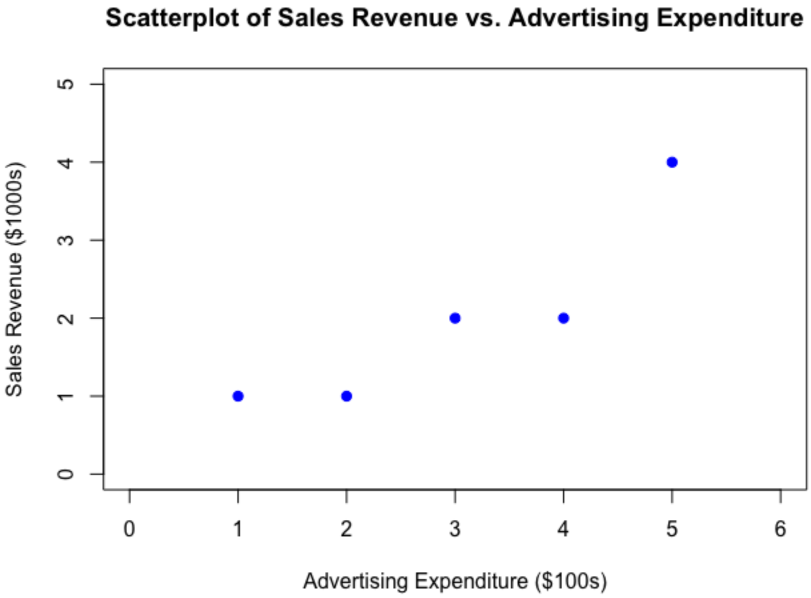
6. Apply the model for prediction and estimation.

3.3 Fitting the Model: The Method of Least Squares

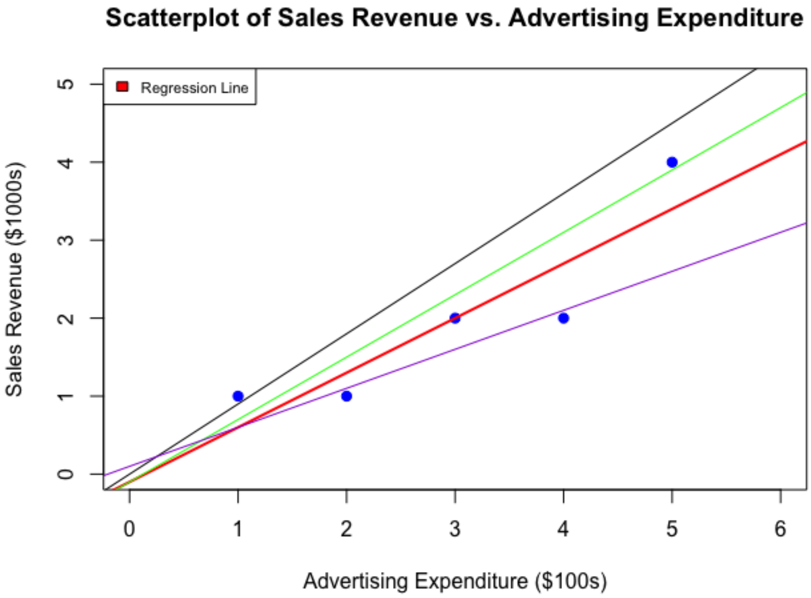
Example 3.3.1 *Suppose an appliance store conducts a 5-month experiment to determine the effect of advertising on sales revenue. (Data set: ADSALES)*

Hypothesized straight-line model: Sales Revenue $= \beta_0 + \beta_1$Advertising Expenditure $+ \epsilon$

To increase our information about the sample data, we look at the **scatterplot**.

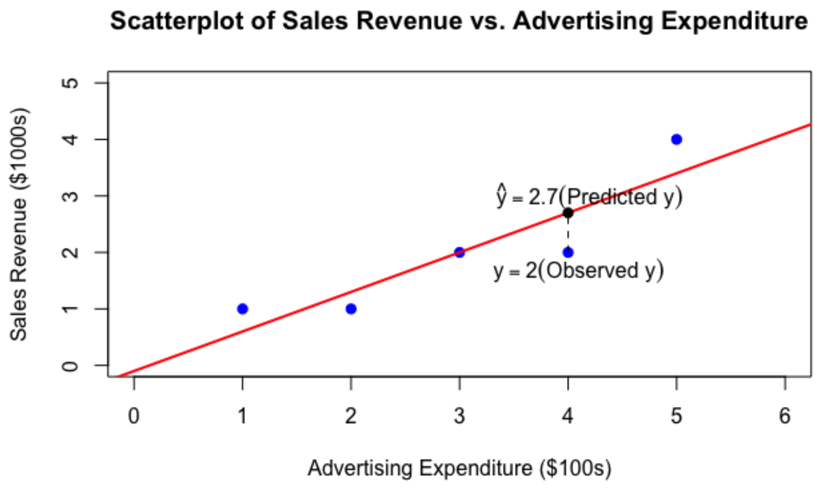


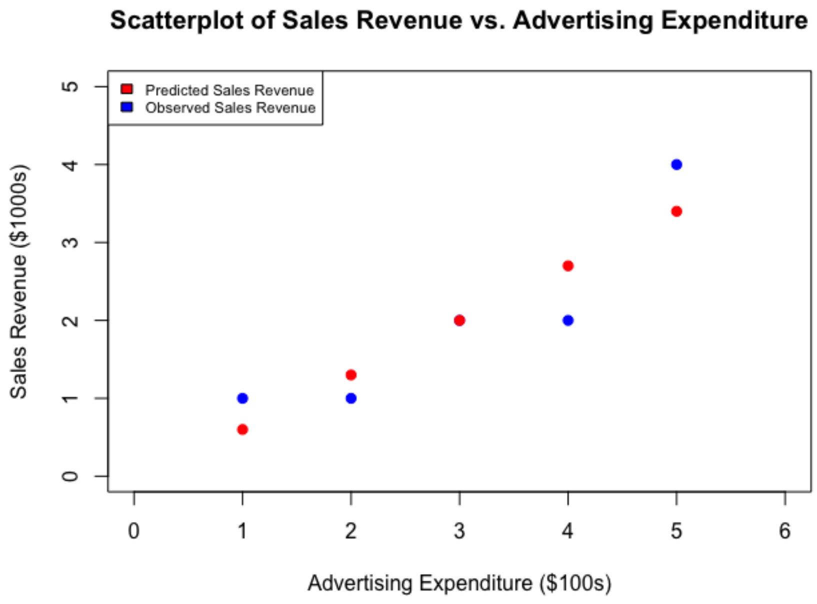
What does the scatterplot suggest? Can we plot a line that passes exactly all the points in the scatterplot? How many lines can we plot?



We calculate the magnitude of the deviations which are the differences between the observed and the predicted values of y suggested by the lines.

These deviations or **errors of prediction**, are the vertical distances between observed and predicted values of y.





Note: Sum of the Errors, $SE = 0$. It can be shown that there is one (and only one) line for which the SSE is a **minimum**. This line is called the **least squares line**, regression line, or least squares prediction equation.

Residual or Error of Prediction :

The deviation of the ith value of y from its predicted value, called the ith residual, is:

$$e_i = (y_i - \hat{y}_i) = [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]$$

Hence, the least squares line has the following two properties:

- $SE = \sum (y_i - \hat{y}_i) = 0$

- Sum of Squares of Residuals $SSE = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$ is smaller than for any other straight-line model with $SE = 0$.

Formulas for the Least Squares Estimates

The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the SSE are called the least squares estimates of the population parameters β_0 and β_1, and the prediction equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ called the least squares line.

Slope: $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$

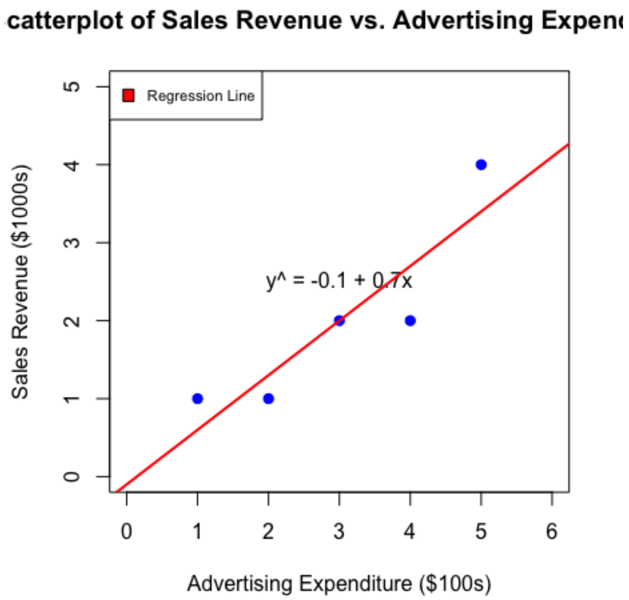
y-intercept: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

where

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

n = sample size

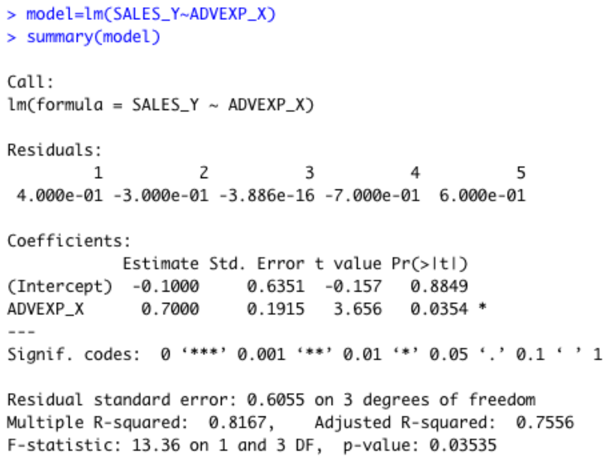


Question: Provide an interpretation of the slope estimate and intercept of the Least Square Regression Line in the context of the question.

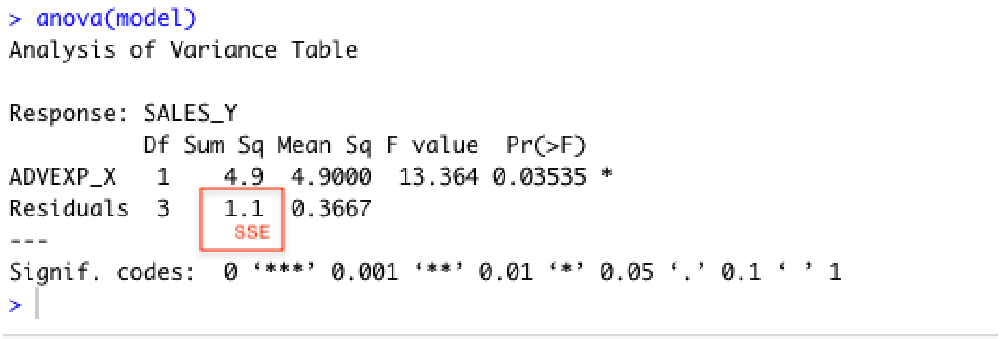
a) The mean monthly sales revenue increases $700 for every $100 increase in monthly advertising expenditure.

b) It is not appropriate. (Why?)

Extrapolation: Predicting outside the range of the sample data is called extrapolation.



```
> model=lm(SALES_Y~ADVEXP_X)
> summary(model)

Call:
lm(formula = SALES_Y ~ ADVEXP_X)

Residuals:
         1          2          3          4          5 
 4.000e-01 -3.000e-01 -3.886e-16 -7.000e-01  6.000e-01 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  -0.1000     0.6351  -0.157   0.8849  
ADVEXP_X      0.7000     0.1915   3.656   0.0354 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6055 on 3 degrees of freedom
Multiple R-squared:  0.8167,	Adjusted R-squared:  0.7556 
F-statistic: 13.36 on 1 and 3 DF,  p-value: 0.03535
```



```
> anova(model)
Analysis of Variance Table

Response: SALES_Y
          Df Sum Sq Mean Sq F value  Pr(>F)  
ADVEXP_X   1    4.9  4.9000  13.364 0.03535 *
Residuals  3    1.1  0.3667                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

SSE

3.4 Model Assumptions

Recall the probabilistic model $y = \beta_0 + \beta_1 x + \epsilon$ and the least squares estimate of the deterministic component of the model, $\beta_0 + \beta_1 x$, is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

We will use a probability distribution to characterize the behavior of ϵ. We will see how the probability distribution of ϵ determines how well the model describes the relationship between the dependent variable y and the independent variable x.

Basic Assumptions about the Probability Distribution of ϵ:

Assumption 1:

The mean of the probability distribution of ϵ is 0. $E(\epsilon) = 0 \longrightarrow E(y) = \beta_0 + \beta_1 x$

Assumption 2:

The variance of the probability distribution of ϵ is constant for all settings of the independent variable x.

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

Assumption 3:

The probability distribution of ϵ is normal.

$$\epsilon \sim N(0, \sigma)$$

Assumption 4:

The values of ϵ associated with any two observed values of y are independent. That is, the value of ϵ associated with one value of y has no effect on any of the values of ϵ associated with any other y values.

The implications of the first three assumptions can be seen in Figure 3.2, which shows distributions of errors for three values of x, namely, x_1, $x2$, and x_3.

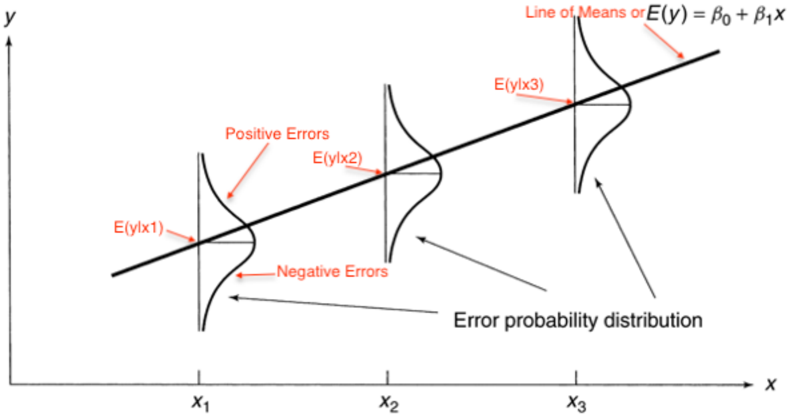

y
Line of Means or $E(y) = \beta_0 + \beta_1 x$
E(y|x3)
Positive Errors
E(y|x2)
E(y|x1)
Negative Errors
Error probability distribution
x_1
x_2
x_3
x

Figure 3.2: The probability distribution of ϵ

3.5 An Estimator of σ^2

Since the greater the variability of the random error ϵ (which is measured by its variance σ^2), the greater will be the errors in the estimation of the model parameters β_0 and β_1, and in the error of prediction when $\hat{y}$ is used to predict y for some value of x, we estimate σ^2 by s^2 and use it in all infrences.

Estimation of σ^2 for a (First-Order) Straight-Line Model

$$s^2 = \frac{SSE}{\text{Degrees of freedom for error}} = \frac{SSE}{n-2} = \text{Mean Square Error}, MSE$$

where

$SSE = \sum (y_i - \hat{y}_i)^2 = SS_{yy} - \hat{\beta}_1 SS_{xy}$

in which

$$SS_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

To estimate the standard deviation σ of ϵ, we calculate

$s = \sqrt{s^2} = \sqrt{\frac{SSE}{n-2}}$

We will refer to s as the estimated standard error of the regression model.

Interpretation of s, the Estimated Standard Deviation of ϵ

We expect most ($\approx 95\%$) of the observed y values to lie within $2s$ of their respective least squares predicted values, $\hat{y}$.

Example 3.5.1 *Refer to Example 3.3.1:*

a. Compute an estimate of σ.

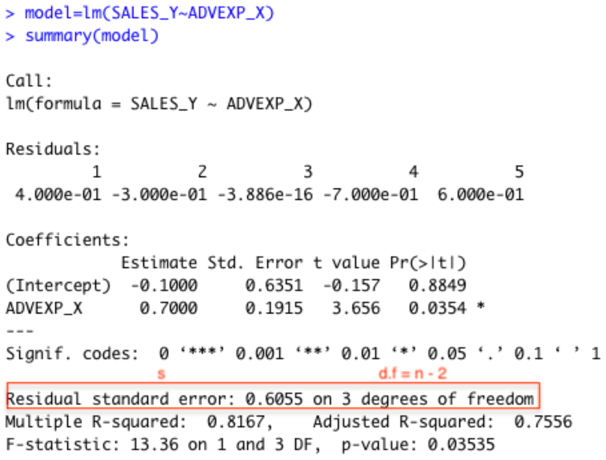
b. *Give a practical interpretation of the estimate.*

Solution

a. $s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1.1}{3}} = .61$ and $s^2 = 0.367$.

b. Since s measures the spread of the distribution of y values about the least squares line and these errors of prediction are assumed to be normally distributed, we should not be surprised to find that most (about 95%) of the observations lie within $2s$, or $2(.61) = 1.22$, of the least squares line. For this simple example (only five data points), all five data points fall within $2s$ of the least squares line. In this example; most of the monthly sales revenue values fall within \$1,220 of their respective predicted values using the least squares line.

That means in this example, an error of prediction of \$1,220 is probably acceptable if monthly sales revenue values are relatively large (e.g., \$100,000). On the other hand, an error of \$1,220 is undesirable if monthly sales revenues are small (e.g., \$1,000–5,000).

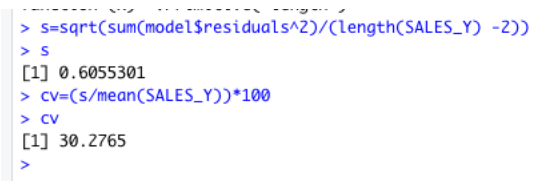


```
> model=lm(SALES_Y~ADVEXP_X)
> summary(model)

Call:
lm(formula = SALES_Y ~ ADVEXP_X)

Residuals:
         1          2          3          4          5 
 4.000e-01 -3.000e-01 -3.886e-16 -7.000e-01  6.000e-01 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  -0.1000     0.6351  -0.157   0.8849  
ADVEXP_X      0.7000     0.1915   3.656   0.0354 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6055 on 3 degrees of freedom
Multiple R-squared:  0.8167,	Adjusted R-squared:  0.7556 
F-statistic: 13.36 on 1 and 3 DF,  p-value: 0.03535
```

Coefficient of Variation, CV

The ratio of the estimated standard deviation, s to the sample mean of the dependent variable, $\bar{y}$, in a percentage:

$$CV = \frac{s}{\bar{y}} \times 100$$



```
> s=sqrt(sum(model$residuals^2)/(length(SALES_Y) -2))
> s
[1] 0.6055301
> cv=(s/mean(SALES_Y))*100
> cv
[1] 30.2765
> 
```

The value of s for the least squares line is 30% of the value of the sample mean sales revenue, $\bar{y}$.

Note:

As a rule of thumb, Regression models with CV values of **10% or smaller** usually lead to more precise predictions.

3.6 Assessing the Utility of the Model: Making Inferences About the Slope β_1

If x contributes no information for the prediction of y, $E(y)$ ($E(y) = \beta_0 + \beta_1 x$) does not change as x changes. Hence the predicted value of y is the same regardless of the value of x or in the straight-line model, $\beta_1 = 0$.

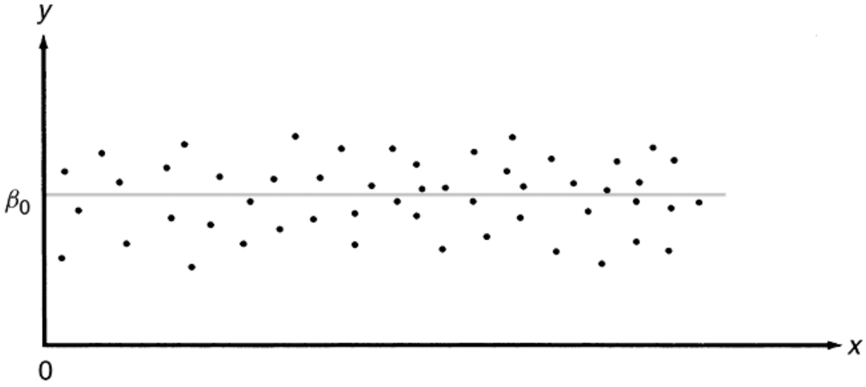


Figure 3.3: Graphing the model with $\beta_1 = 0 : y = \beta_0 + \epsilon$

The corresponding test is:

Test of the Usefulness of the Hypothesized Model

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

In order to find the test statistic, we consider the sampling distribution of the slope:

Sampling Distribution of $\hat{\beta}_1$

If we make the four assumptions about ϵ, then the sampling distribution of $\hat{\beta}_1$, the least squares estimator of the slope, will be a normal distribution with mean β_1 (the true slope) and standard deviation:

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SS_{xx}}}$$

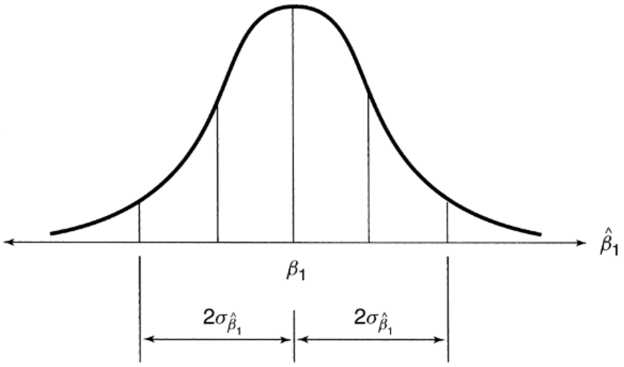


Figure 3.4: Sampling Distribution of $\hat{\beta}_1$

We estimate $\sigma_{\hat{\beta}_1}$ by $s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$ and refer to $s_{\hat{\beta}_1}$ as the estimated standard error of the least suares slope $\hat{\beta}_1$.

A Test of Model Utility: Simple Linear Regression

Test statistic:

$$t_c = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\frac{s}{\sqrt{SS_{xx}}}}$$

Lower-Tailed Tests

$H_0 : \beta_1 = 0$

$H_a : \beta_1 < 0$

Rejection region: $t < -t_\alpha$

p-value: $P(t < t_c)$

Upper-Tailed Tests

$H_0 : \beta_1 = 0$

$H_a : \beta_1 > 0$

Rejection region: $t > t_\alpha$

p-value: $P(t > t_c)$

Two-Tailed Test

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

Rejection region: $|t| > t_{\frac{\alpha}{2}}$

p-value: $2P(t > t_c)$ if t_c is positive

p-value: $2P(t < t_c)$ if t_c is negative

Decision

Reject H_0 if p-value $< \alpha$ or if test statistic (t_c) falls in rejection region.

Degrees of Freedom

t is based on $(n-2)$ degrees of freedom.

Assumptions: The four assumptions about the probability distribution ϵ.

A 100$(1-\alpha)$% Confidence Interval for the Simple Linear Regression Slope β_1

$$\hat{\beta}_1 \pm (t_{\frac{\alpha}{2}}) s_{\hat{\beta}_1}$$

where the estimated standard error of $\hat{\beta}_1$ is calculated by

$s_{\hat{\beta}_1} = s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$

and $t_{\frac{\alpha}{2}}$ is based on (n - 2) degrees of freedom.

Example 3.6.1 *Refer to the simple linear regression analysis for the advertising – sales example. Conduct a test (at $\alpha=$ .05) to determine whether the sales revenue (y) is linearly related to the advertising expenditure (x). Find a 95% confidence interval for the slope β_1.*

Solution:

Refer to the RSudio output. The t-value (test statistic) and p-value are 3.656 and 0.0354 respectively.

1. p-value Approach:

   Since the p-value $=$ .0354 is smaller than $\alpha =$ .05, we will reject H_0 and conclude that the slope β_1 is not 0.

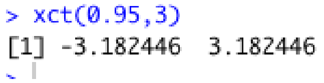
   Conclusion:

   The sample evidence indicates that the advertising expenditure, x contributes information for the prediction of the sales revenue, y when a linear model is used.

2. Rejection Region Approach

   Since $n = 5 \longrightarrow d.f = 3$ and the rejection region t (at $\alpha = .05$) will be

   $|t| > t_{\frac{0.05}{2}} = 3.182$



```
> xct(0.95,3)
[1] -3.182446  3.182446
>
```

Since this calculated t-value falls into the upper-tail rejection region, we reject the null hypothesis and conclude that the slope β_1 is not 0.

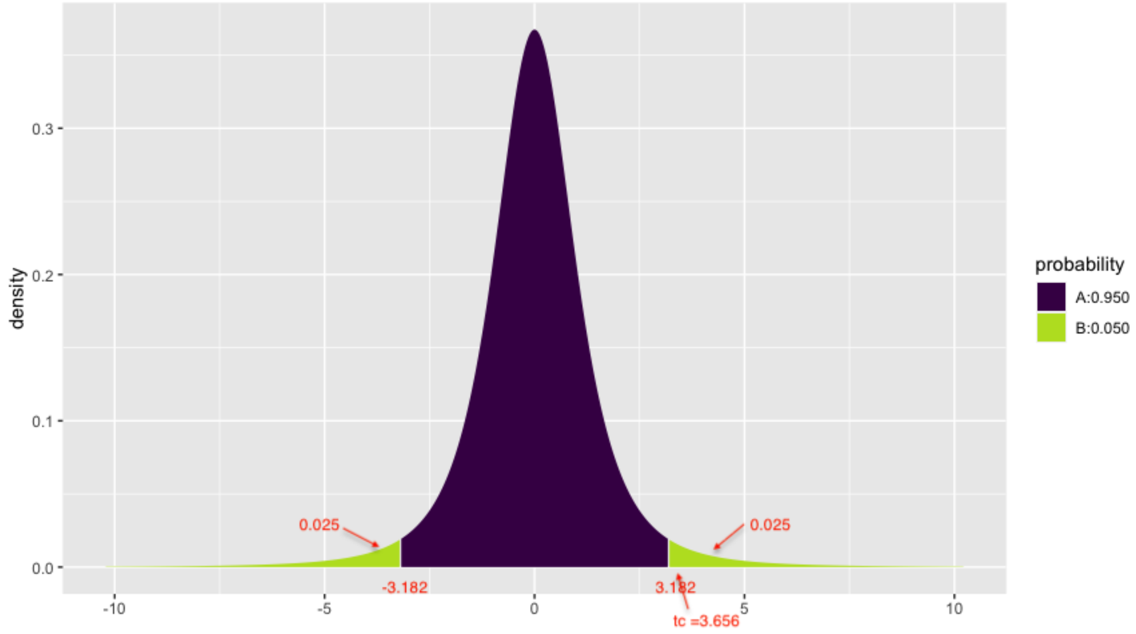


Figure 3.5: Rejection region and calculated t-value for testing whether the slope $\beta_1 = 0$

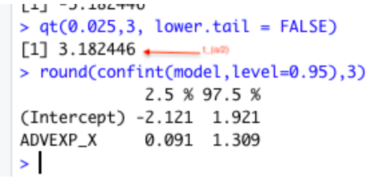
Confidence interval for the slope β_1:

$$\hat{\beta}_1 \pm (t_{\frac{\alpha}{2}}) s_{\hat{\beta}_1}$$

where from the RSudio output:

$\hat{\beta}_1 = 0.7$, $s_{\hat{\beta}_1} = 0.1915$ and $t_{\frac{0.05}{2}} = 3.182$

This interval can also be obtained with RStudio:



```
[1] -3.182446
> qt(0.025,3, lower.tail = FALSE)
[1] 3.182446
> round(confint(model,level=0.95),3)
             2.5 % 97.5 %
(Intercept) -2.121  1.921
ADVEXP_X     0.091  1.309
> 
```

Figure 3.6: RStudio output with 95% confidence intervals for the regression betas

We are 95% confident that the mean monthly sales revenue will increase between \$91 and \$1,309 for every \$100 increase in monthly advertising expenditure.

Question: How can we describe the relationship between $E(y)$ and x from the calculated confidence interval above?

3.7 The Coefficient of Correlation

A **bivariate relationship** describes an "association" or a "relationship" or a "correlation" between two variables x and y. Scatterplots are used to describe a bivariate relationship graphically. In this section, we will discuss the concept of correlation and how it can be used to measure the linear relationship between two variables x and y.

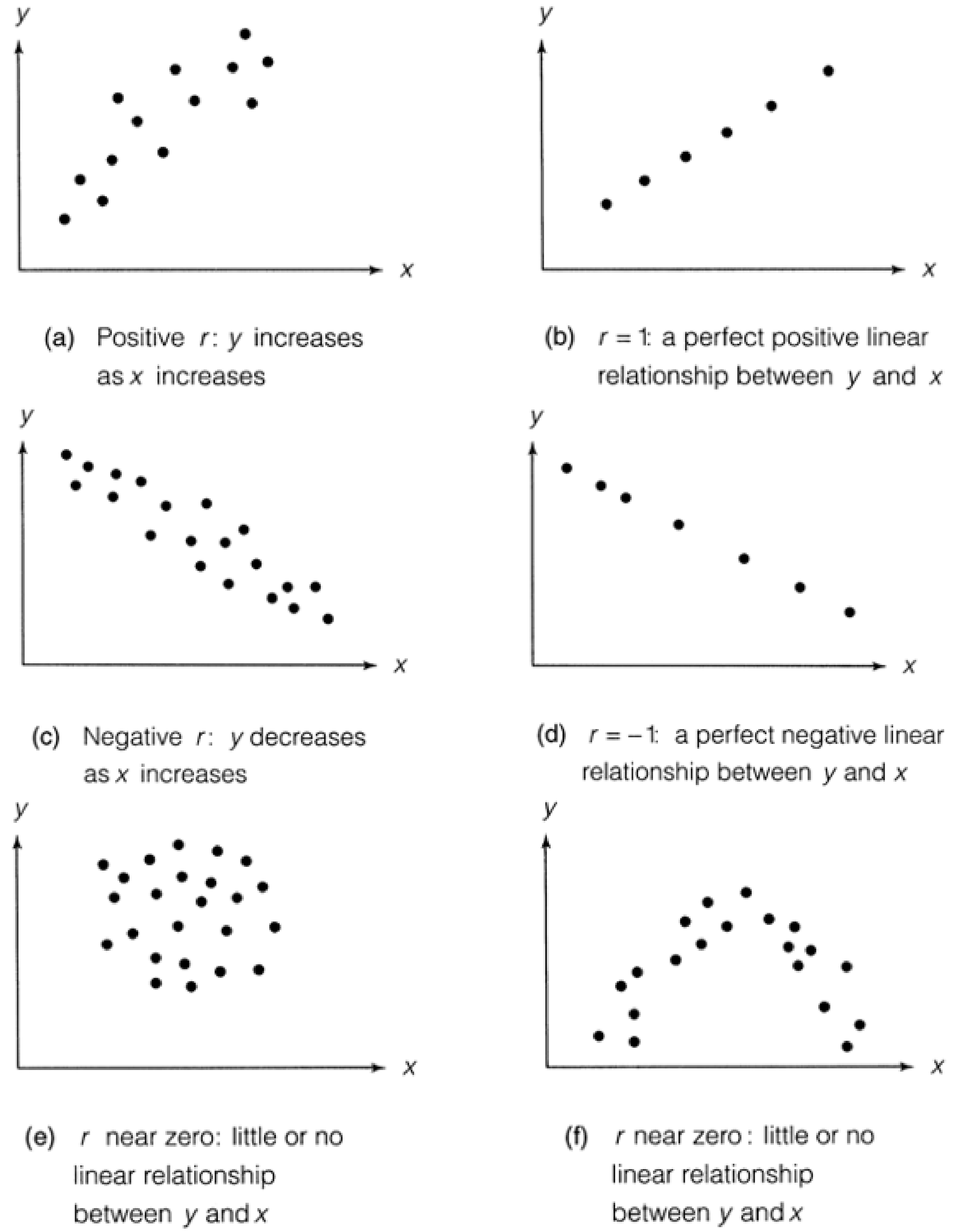
A numerical or quantitative descriptive measure of the strength of the **linear** relationship between x and y is provided by the coefficient of correlation, r or **Pearson product moment correlation coefficient**.

It is computed (for a sample of n measurements on x and y) as follows:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

Since the numerators of the expressions for $\hat{\beta}_1$ and r are identical, it is clear that $r = 0$ when $\hat{\beta}_1 = 0$ (the case where x contributes no information for the prediction of y) and that r is positive when the slope is positive and negative when the slope is negative. Unlike $\hat{\beta}_1$, the correlation coefficient r is scaleless and assumes a value between -1 and +1, regardless of the units of x and y.

A value of r near or equal to 0 implies little or no linear relationship between y and x. In contrast, the closer r comes to 1 or -1, the stronger is the linear relationship between y and x.And if $r = 1$ or $r = -1$, all the sample points fall exactly on the



(a) Positive r: y increases as x increases

(b) $r = 1$: a perfect positive linear relationship between y and x

(c) Negative r: y decreases as x increases

(d) $r = -1$: a perfect negative linear relationship between y and x

(e) r near zero: little or no linear relationship between y and x

(f) r near zero : little or no linear relationship between y and x

Figure 3.7: Types of Relationships Depicted by Scatterplots with different values of r

least squares line. Positive values of r imply a positive linear relationship between y and x; that is, y increases as x increases. Negative values of r imply a negative linear relationship between y and x; that is, y decreases as x increases.

Note: High absolute value of correlation r does not imply **causality**. The only valid conclusion is that a **linear** trend may exist between x and y.

Since the correlation coefficient r measures the correlation between x-values and y-values in the sample, we conduct the following test to see whether a similar linear coefficient of correlation exists for the population from which the data points were selected. We denote the population correlation coefficient by ρ.

A Test for Linear Correlation

Test statistic: $t_c = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$

Lower-Tailed Tests

$H_0 : \rho = 0$

$H_a : \rho < 0$

Rejection region: $t < -t_\alpha$

p-value: $P(t < t_c)$

Upper-Tailed Tests

$H_0 : \rho = 0$

$H_a : \rho > 0$

Rejection region: $t > t_\alpha$

p-value: $P(t > t_c)$

Two-Tailed Test

$H_0 : \rho = 0$

$H_a : \rho \neq 0$

Rejection region: $|t| > t_{\frac{\alpha}{2}}$

p-value: $2P(t > t_c)$ if t_c is positive

p-value: $2P(t < t_c)$ if t_c is negative

Decision

Reject H_0 if p-value $< \alpha$ or if test statistic (t_c) falls in rejection region.

Degrees of Freedom

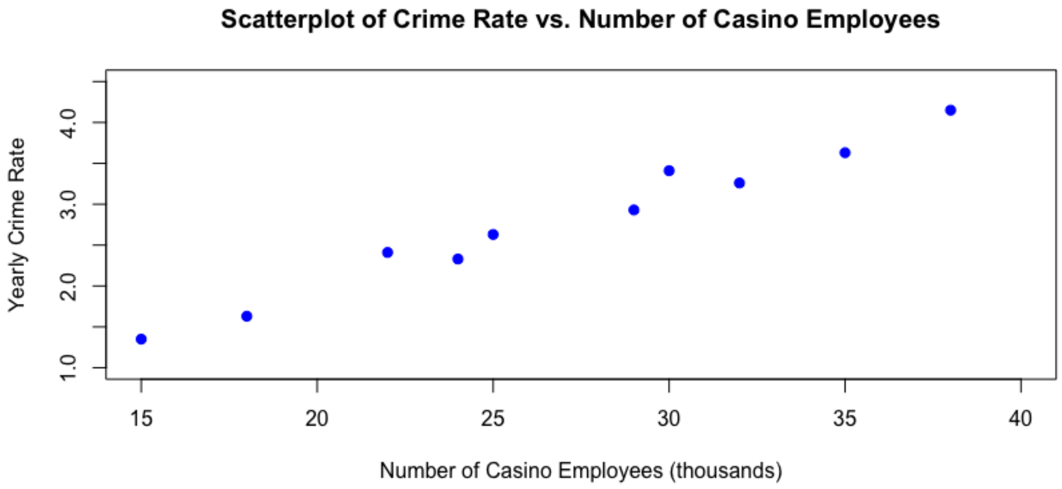
t is based on $(n-2)$ degrees of freedom.

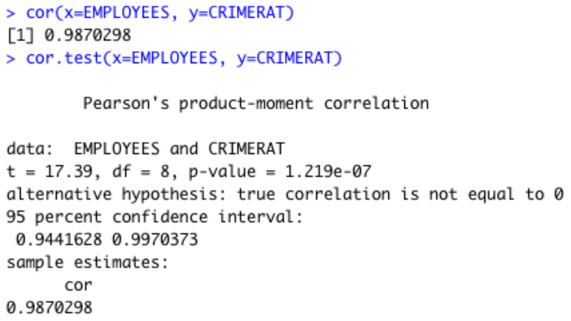
Assumptions:

The sample of (x, y) values is randomly selected from a **normal** population.

Example 3.7.1 *Refer to Example 3.1 from the textbook. (Data set: CASINO)*

Describe the following outputs.





```
> cor(x=EMPLOYEES, y=CRIMERAT)
[1] 0.9870298
> cor.test(x=EMPLOYEES, y=CRIMERAT)

	Pearson's product-moment correlation

data:  EMPLOYEES and CRIMERAT
t = 17.39, df = 8, p-value = 1.219e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9441628 0.9970373
sample estimates:
      cor
0.9870298
```

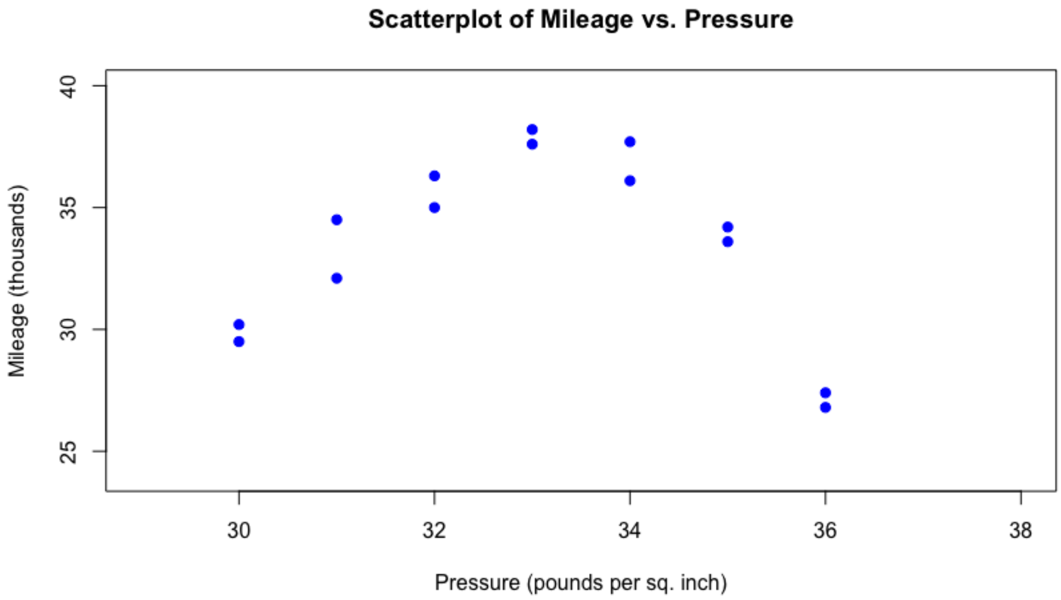
a) The scatterplot shows....

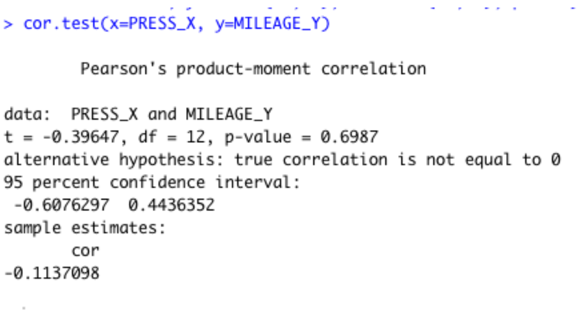
b) The value of $r = 0.9870298$ indicates...

c) The $p-$value $= 1.219\text{e-}07$ means ...

Example 3.7.2 *Refer to Example 3.2 from the textbook. (Data set: TIRES)*

Describe the following outputs.





```
> cor.test(x=PRESS_X, y=MILEAGE_Y)

	Pearson's product-moment correlation

data:  PRESS_X and MILEAGE_Y
t = -0.39647, df = 12, p-value = 0.6987
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6076297  0.4436352
sample estimates:
       cor
-0.1137098
```

a) The scatterplot shows....

b) The value of $r = -0.1137098$ indicates...

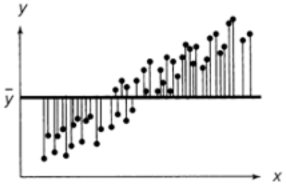
c) The $p-$value $= 0.6987$ means ...

3.8 The Coefficient of Determination

We also quantify the contribution of x in predicting y to measure the utility of the regression model. Hence, we compute how much the errors of prediction of y were reduced by using the information provided by x.

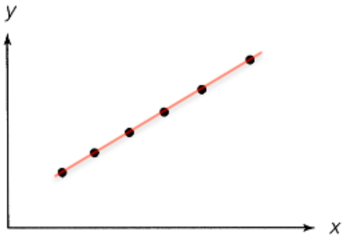
Consider the following cases:

1. **Question:** What would be the best prediction for a value of y and the sum of squares of deviations, If we assume x contributes no information for the prediction of y.



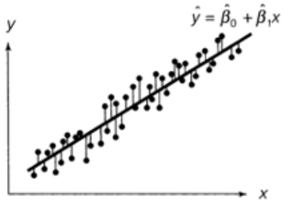
(b) Assumption: x contributes no information for predicting y; $\hat{y} = \bar{y}$

2. **Question:** What would be the best prediction for a value of y and the sum of squares of deviations, If we assume x contributes **all** information for the prediction of y.



(b) $r = 1$: a perfect positive linear relationship between y and x

3. **Question:** What would be the best prediction for a value of y and the sum of squares of deviations, If we assume x contributes **some** information for the prediction of y.



(c) Assumption: x contributes information for predicting y; $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

The **coefficient of determination** is

$$r^2 = \frac{(SS_{yy} - SSE)}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}} = \frac{SSR}{SS_{yy}}$$

and represents the proportion of the total sample variability around y that is explained by the linear relationship between $\bar{y}$ and x. (In simple linear regression, it may also be computed as the square of the coefficient of correlation, r.)

Where:

Sum of Squares of Total (Total Sample Variability): $SST = SS_{yy} = \sum (y_i - \bar{y})^2$

Sum of Squares of Residuals (Unexplained Sample Variability: $SSE = \sum (y_i - \hat{y}_i)^2$

Sum of Squares Regression (Explained Sample Variability) $SSR = \sum (\hat{y}_i - \bar{y})^2$

$$SST = SSE + SSR$$

Practical Interpretation of the Coefficient of Determination, r^2

About $100(r^2)\%$ of the sample variation in y (measured by the total sum of the squares of the deviations of the sample y values about their mean $\bar{y}$) can be explained by (or attributed to) using x to predict y in the straight-line model.

Example 3.8.1 *Find the coefficient of determination for the advertising–sales example from the RSudio output and interprete the value.*

3.9 Using the Model for Estimation and Prediction

Useful models can be applied to estimate or to predict the dependent variable, y for a given value of the independent variable, x.

Recall we use the model mainly for:

- estimating the mean value of y, $E(y)$, for a specific value of x. (When we have a very large number of experiments at the given x-value.)
- predicting a particular y value for a given x. (When we have a single experiment at the given x-value.)

In both cases we use the least squares model, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ to estimate the mean value of y and to predict a particular value of y for a given value of x.

$\widehat{E(y)} = 2.7$ or \$2700 means the estimated mean value of sales revenue for all months during which \$400 ($x = 4$) is spent on advertising.

$\hat{y} = 2.7$ or \$2700 means the predicted value of sales revenue if we spend \$400 on on advertising.

The difference is in the relative accuracy of the estimate and the prediction.

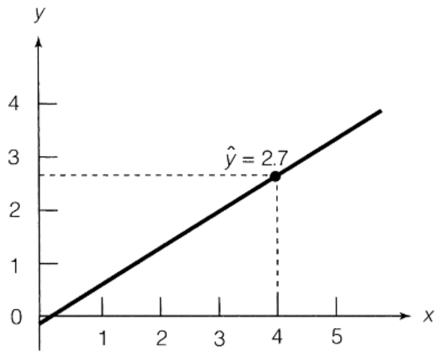


Figure 3.8: Estimated mean value and predicted individual value of sales revenue y for $x = 4$

Sampling Errors for the Estimator of the Mean of y and the Predictor of an Individual y for $x = x_p$

1. The satndard deviation of the sampling distribution of the estimator $\hat{y}$ of the mean value of y at a specific value of x, say x_p, is

$$\sigma_{\hat{y}} = \sigma\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

   where σ is the standard deviation of the random error ϵ. We refer to $\sigma_{\hat{y}}$ as the standard error of $\hat{y}$.

2. The satndard deviation of the prediction error for the predictor $\hat{y}$ of an individual new y value at a specific value of x is

$$\sigma_{(y-\hat{y})} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

   where σ is the standard deviation of the random error ϵ. We refer to $\sigma_{(y-\hat{y})}$ as standard error of prediction.

The true value of σ is rarely known, so we estimate σ by s and calculate the estimation and prediction intervals as shown in the next two boxes:

A 100(1 - α)% Confidence Interval for the Mean Value of y at $x = x_p$

$$\hat{y} \pm t_{\frac{\alpha}{2}} s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

where $t_{\frac{\alpha}{2}}$ is based on $(n-2)$ degrees of freedom.

A 100(1 - α)% Prediction Interval for an Individual y at $x = x_p$

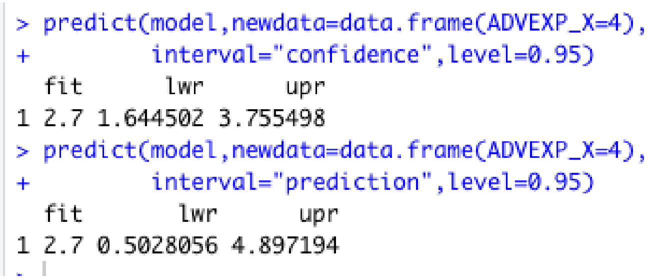
$$\hat{y} \pm t_{\frac{\alpha}{2}} s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

where $t_{\frac{\alpha}{2}}$ is based on $(n-2)$ degrees of freedom.

Example 3.9.1 *a) Find a 95% confidence interval for mean monthly sales when the appliance store spends \$400 on advertising.*

b) Predict the monthly sales for next month if a \$400 expenditure is to be made on advertising. Use a 95% prediction interval.

c) Interpret the calculated intervals.



```
> predict(model,newdata=data.frame(ADVEXP_X=4),
+         interval="confidence",level=0.95)
  fit      lwr      upr
1 2.7 1.644502 3.755498
> predict(model,newdata=data.frame(ADVEXP_X=4),
+         interval="prediction",level=0.95)
  fit       lwr      upr
1 2.7 0.5028056 4.897194
```

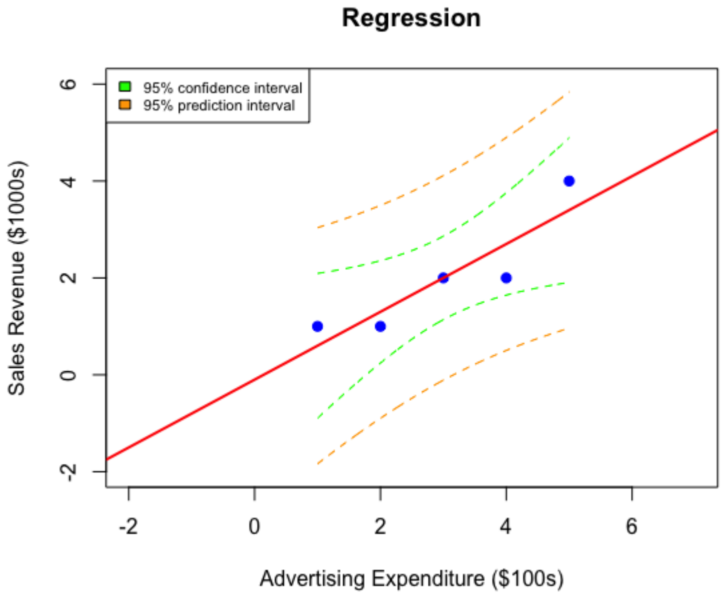
For CI: We are 95% confident that the mean value of sales revenue for all months when the store spends \$400 on advertising falls between \$1,645 to \$3,755.

For PI: With 95% confidence, we predict that the sales for the month we spend \$400 in advertising falls between \$503 and \$48972.

We can see that the confidence interval is **narrower** than the prediction interval.

For both uses:

- the error is minimum if $x_p = \bar{x}$
- the error is larger as x_p away from $\bar{x}$



Question: Why the confidence interval is always **narrower** than the prediction interval?

$\hat{y} - E(y)$: The error in estimating the mean value of y, $E(y)$, for a given value of x, say, x_p, is the distance between the least squares line and the true line of means, $E(y) = \beta_0 + \beta_1 x$.

$y_p - \hat{y}$: The error in predicting some future value of y is the sum of two errors—the error of estimating the mean of y, $E(y)$ plus the random error that is a component of the value of y to be predicted.

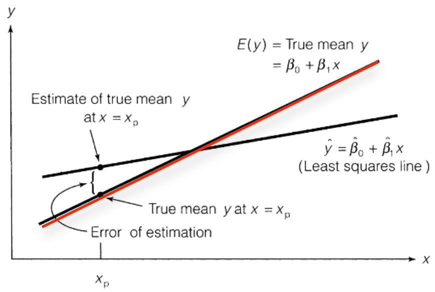


Figure 3.9: Error of estimating the mean value of y for a given value of x

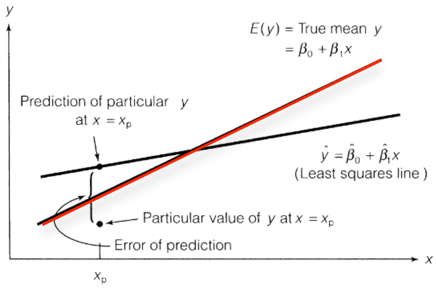


Figure 3.10: Error of predicting a future value of y for a given value of x

Caution Making any inferences about $E(y)$ or y (estimation or predection) when x is far away from $\bar{x}$ so that it falls outside the range of the sample data is dangerous and called **extrapolation**.

Note:

- The width of the confidence interval decreases as the sample size, n increases.
- To obtain more accurate predictions for new values of y, the standard deviation of the regression model, σ should be reduced and can be accomplished only by improving the model, either by using a curvilinear (rather than linear) relationship with x or by adding new independent variables to the model, or both.

Acknowledgement

The core content of the slides are from the textbook of this course;

A Second Course in Statistics: Regression Analysis (7th Edition)

by

Mendenhall, William and Sincich, Terry; Pearson Education.