Chapter 11

Correlation & Simple Regression

The previous chapter dealt with inference for two categorical variables. In this chapter, we would like to examine the relationship between two *quantitative* variables. A common summary statistic describing the linear association between two quantitative variables is *Pearson's sample correlation coefficient.* More detailed inferences between two quantitative random variables is provided by a framework called *simple regression.*

11.1 Pearson's sample correlation coefficient

Definition 11.1: Pearson's Sample Correlation Coefficient
Pearson's sample correlation coefficient is

$$r = \frac{Cov(x,y)}{s_x s_y}$$

where the *sample covariance* between x and y is

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Equivalently, the correlation is sometimes computed using

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right).$$

Note:

1. r measures the strength of the *linear* association between two variables, say x and y.
2. $r > 0 \implies$ as x increases, y tends to increase.
3. $r < 0 \implies$ as x increases, y tends to decrease.
4. $-1 \leq r \leq 1$
5. r is affected by outliers.
6. $Cov(x,y)$ describes how x and y vary together (i.e. how they "co-vary").
7. $-\infty < Cov(x,y) < \infty$

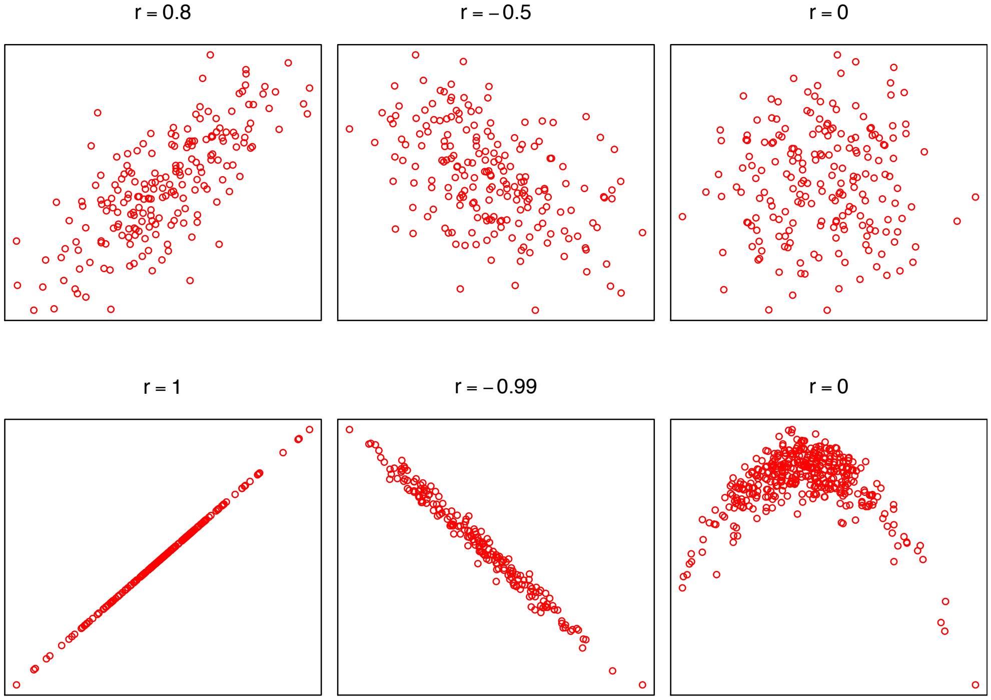

Figure 11.1: Scatterplots with various correlation coefficients.

8. $Cov(x, y)$ indicates a positive or negative association, not the strength of the association (i.e. a larger covariance doesn't necessarily indicate a stronger association/correlation).

Example: *Pearson's Sample Correlation Coefficient*

(a) Weight of car vs. mileage $\Longrightarrow r < 0$

(b) Weight of car vs. cost $\Longrightarrow r > 0$

(c) Natural gas usage vs. outside temperature $\Longrightarrow r < 0$

(d) Hours studied vs. exam score $\Longrightarrow r > 0$

Example: Scatterplots with $r = 0.8, -0.5, 0, 1, -0.99, 0$ are depicted in Figure 11.1 on page 184. The bottom right figure plots rainfall on the horizontal axis and crop yield on the vertical axis; because the correlation coefficient only detects *linear* associations, the correlation coefficient is 0 (there is a strong *quadratic* relationship, however).

Example: *Correlation*
We have data on the study habits and exam score of 4 students.

x = hours studied:	10	14	2	10
y = exam score:	82	94	50	70

A scatter plot of the data is shown in Figure 11.2.

(a) Compute r.

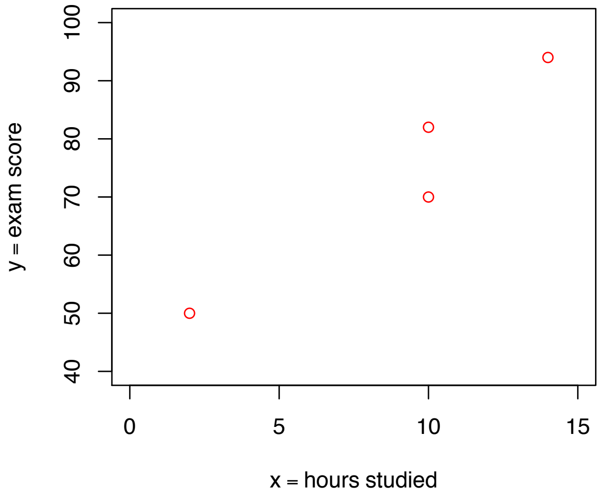

Figure 11.2: Scatterplot of x = hours studied versus y = exam score.

We have $n = 4$, $\bar{x} = 9$, $\bar{y} = 74$,

$$\begin{aligned} s_x &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ &= \sqrt{\frac{(10-9)^2+(14-9)^2+(2-9)^2+(10-9)^2}{4-1}} \\ &= 5.033 \end{aligned}$$

and

$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2} = 18.762$$

The covariance is

$$\begin{aligned} Cov(x,y) &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\ &= \frac{1}{4-1}[(10-9)(82-74)+(14-9)(94-74) \\ &\qquad +(2-9)(50-74)+(10-9)(70-74)] \\ &= \frac{272}{3} = 90.667 \end{aligned}$$

Therefore, Pearson's sample correlation coefficient is

$$r = \frac{Cov(x,y)}{s_x s_y} = \frac{90.667}{5.033 \cdot 18.762} = 0.960$$

Note: If two variables are correlated, one does not necessarily cause the other (i.e. correlation does not imply causation).

- Ice cream sales vs. number of drownings
- Amount of hair vs. running speed

11.2 Simple regression

Definition 11.2: Response and explanatory variables, regression line

- Response variable – measures the outcome of an individual. The response variable is denoted by y.
- Explanatory variable – explains (or influences) changes in the response variable. The explanatory variable is denoted by x. It is possible to have more than 1 explanatory variable; this is called *multiple regression*.
- A *regression line* describes how the mean of the response variable y changes as the explanatory variable x changes.

Theorem 11.1. *Least squares regression line*
The least squares regression line is the line that minimizes the sum of the squared vertical distances from the data points to the line (we use calculus to find this minimum). «show graph» The least squares regression line is

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

where (after some calculus)

$$\hat{\beta}_1 = r\frac{s_y}{s_x}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

The slope of the line is $\hat{\beta}_1$ *and the intercept is* $\hat{\beta}_0$*.*

Definition 11.3: Population regression line
The population regression line can be thought of as the "true" underlying regression line that we are trying to infer about. The population regression line is denoted as

$$\mu_{y|x} = \beta_0 + \beta_1 x$$

where $\mu_{y|x}$ is the population mean of y when the explanatory variable is equal to x. In theory, we could determine the population regression line if we collected data on all individuals in the population and proceeded to find the corresponding regression line. In reality, however, we can not collect data on the entire population; we only have a *sample* from the population. The least squares regression line is determined from this sample data. We believe that the least squares regression line

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

is reasonably "close" to the population regression line; i.e. $\hat{\beta}_0$ is close to β_0, $\hat{\beta}_1$ is close to β_1, and, therefore, $\hat{y}$ is close to $\mu_{y|x}$. As such, we use the data in the sample, and the resultant least squares regression line, to infer about the underlying (unknown) population regression line.

Note: Simple regression assumptions

1. The responses $y_1, \ldots, y_n$ are independent.
2. The relationship between x and y is linear. In other words, the population regression equation is a *line* (i.e. $\mu_{y|x} = \beta_0 + \beta_1 x$).

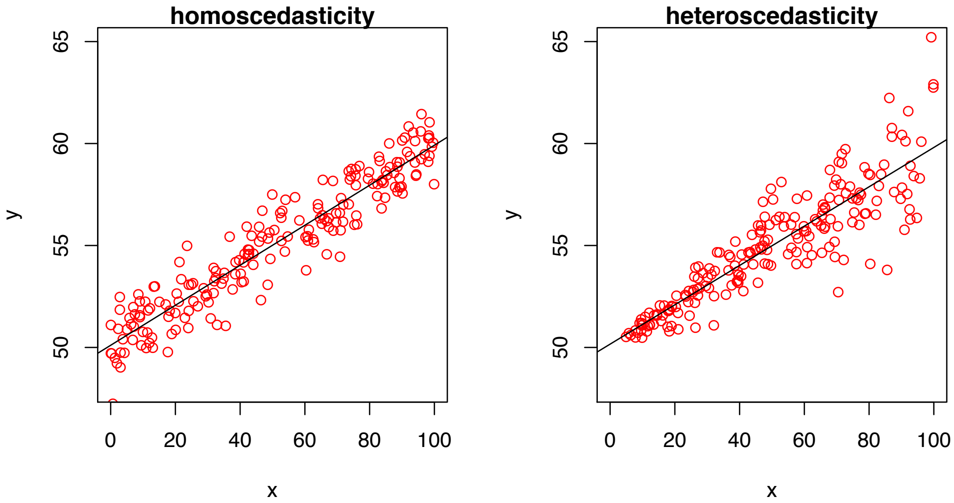

Figure 11.3: Left graph: $\sigma_{y|x}$ is the same for all x. Right graph: $\sigma_{y|x}$ increases in x (i.e. $\sigma_{y|x}$ is large when x is large); this is a violation of the simple regression assumptions.

3. For a given value of x, the distribution of Y is $N(\mu_{y|x}, \sigma_{y|x})$. Note that $\sigma_{y|x}$ describes how much variability (in the y direction) the data has around the regression line for a given value of x. If $\sigma_{y|x}$ is small, then the points will tightly cluster around the regression line; when it is large, the points will be widely spread around the regression line.

4. The standard deviation of Y given x, $\sigma_{y|x}$, must be the same for all x. This is called *homoscedasticity*. If $\sigma_{y|x}$ is not the same for all x, this is called *heteroscedasticity* and is a violation of the required assumptions. See Figure 11.3.

Example (continued): Recall that x = hours studied, y = exam score, $\bar{x} = 9$, $\bar{y} = 74$, $s_x = 5.033$, $s_y = 18.762$, and $r = 0.960$.

(b) Determine the least squares regression line.

The regression coefficients are

$$\hat{\beta}_1 = r\frac{s_y}{s_x} = 0.960\frac{18.762}{5.033} = 3.58$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 74 - 3.58(9) = 41.78$$

therefore the least squares regression line is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 41.78 + 3.58x$$

Be sure you are able find the least squares regression line in the **MTB 11.1** *output on page 200.*

(c) Plot the least squares regression line.

To graph a line, we only need to determine two points on the line and then "connect the dots". For example, when $x = 0$, the height of the regression line is

$$\hat{y} = 41.78 + 3.58(0) = 41.78$$

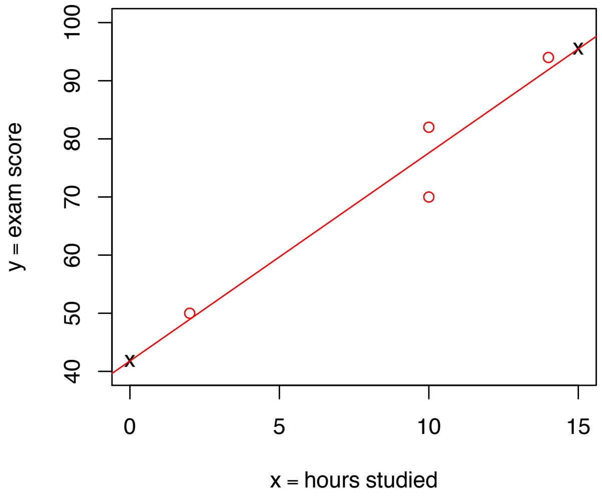

Figure 11.4: Scatterplot of x = hours studied versus y = exam score with the least squares regression line $\hat{y} = 41.78 + 3.58x$.

which, of course, is simply the intercept. When $x = 15$*,*

$$\hat{y} = 41.78 + 3.58(15) = 95.48.$$

*These two points are plotted as "*x*" in Figure 11.4.*

(d) Approximate the population mean exam score for students that studied 5 hours; i.e. approximate $\mu_{y|x=5}$.

Because the population regression line is unknown, we will estimate this unknown population mean using the least squares regression line:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 41.78 + 3.58(5) = 59.68.$$

Note that 59.68 is the height of the regression line at $x = 5$*.*

(e) Approximate the population mean exam score for students that studied 80 hours; i.e. approximate $\mu_{y|x=80}$.

$$\hat{y} = 41.78 + 3.58(80) = 328.18$$

This predicted value $\hat{y}$ *makes no sense since the highest exam score is 100! We tried to make a prediction far outside the range of our original* x *values (which ranged from 2 to 14). Making such predictions is called* extrapolation*; these predictions typically are* extremely *unreliable and should not be trusted.*

(f) Approximate the mean exam score for students that studied 0 hours; i.e. approximate $\mu_{y|x=0} = \beta_0$.

$$\hat{y} = 41.78 + 3.58(0) = 41.78 \quad (= \hat{\beta}_0)$$

(g) Approximate the population mean increase in exam score for each extra hour studied; i.e. approximate β_1.

$$\hat{\beta}_1 = 3.58$$

In other words, each extra hour studied yields an increase of 3.58 in the exam score, on average.

Definition 11.4: Estimated standard error of the regression coefficients
If we collected data on another 4 students, $\hat{\beta}_0$ and $\hat{\beta}_1$ would change. The estimated standard error of $\hat{\beta}_0$, $\widehat{se}(\hat{\beta}_0)$, and the estimated standard error of $\hat{\beta}_1$, $\widehat{se}(\hat{\beta}_1)$, describe how much the intercept $\hat{\beta}_0$ and the slope $\hat{\beta}_1$ change from sample to sample, respectively.

Example (continued):

(h) Is there a significant linear relationship between hours studied x and exam score y? In other words, is there evidence that $\beta_1 \neq 0$ in the population regression equation $\mu_{y|x} = \beta_0 + \beta_1 x$?

To answer this question, we need to test

$$\begin{aligned} H_0 &: \beta_1 = 0 \implies \text{not a significant linear relationship between } x \text{ and } y \\ H_a &: \beta_1 \neq 0 \implies \text{significant linear relationship between } x \text{ and } y \end{aligned}$$

at, say, the $\alpha = 0.05$ significance level. From Minitab we have $\widehat{se}(\hat{\beta}_1) = 0.7368$ (You will always be given the estimated standard errors; we won't learn how compute these. See Section 11.5 if you want the details).

(1) Test Statistic:

$$t^* = \frac{\hat{\beta}_1 - \beta_1}{\widehat{se}(\hat{\beta}_1)} = \frac{3.58 - 0}{0.7368} = 4.86$$

(2) Critical Value: We let p denote the number of parameters we have in our regression model. There are 2 parameters in our model, β_0 and β_1. Thus $p = 2$ and the critical value is

$$t_{\alpha/2,n-p} = t_{0.05/2,4-2} = t_{0.025,2} = 4.303$$

(3) Decision: See Figure 11.5. Reject H_0. Evidence that $\beta_1 \neq 0$. Hence, there is a significant linear relationship between hours studied x and exam score y.

(i) Find the p–value for the test in (i).

$$p-\text{value} = 2P(t_{(n-p)} > |t^*|) = 2P(t_{(2)} > 4.86) \in (0.02, 0.04)$$

Using the applet at

`http://www.stat.uiowa.edu/~mbognar/applets/t.html`

the actual p–value for this two-sided test is 0.0398.

(j) Find a 95% confidence interval for β_1.

$$\hat{\beta}_1 \pm t_{\alpha/2,n-p}\widehat{se}(\hat{\beta}_1) = 3.58 \pm 4.303(0.7368) = 3.58 \pm 3.17 = (0.41, 6.75)$$

Since the CI excludes 0, then there is a significant linear relationship between hours studied and exam score.

Be sure you are able find $\hat{\beta}_1$, $\widehat{se}(\hat{\beta}_1)$, the test statistic t^*, and the p–value in the **MTB 11.1** output on page 200.

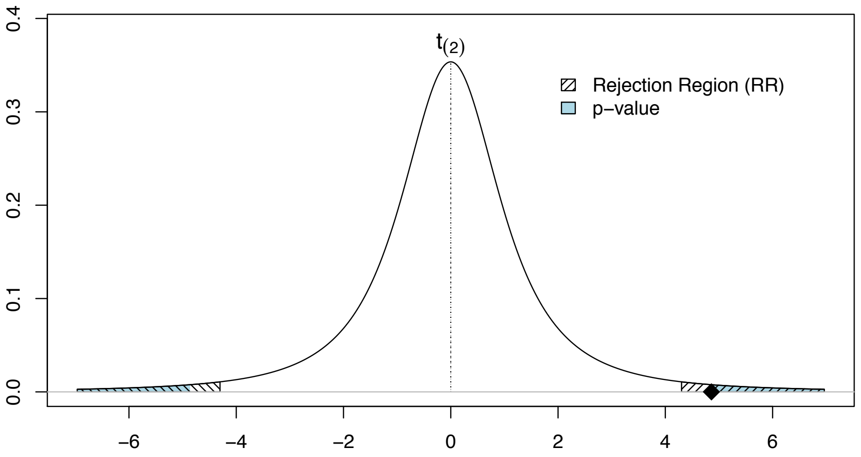

Figure 11.5: Exam Scores: Test statistic t^* is denoted by ◆, the rejection region and p−value are also shown. Note that the total area in the rejection region is equal to $\alpha = 0.05$.

Definition 11.5: Estimated standard error of $\hat{y}$
Recall that $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is an estimate of $\mu_{y|x} = \beta_0 + \beta_1 x$. Because $\hat{\beta}_0$ and $\hat{\beta}_1$ change from sample to sample, $\hat{y}$ will change from sample to sample (for a given x). The estimated standard error of $\hat{y}$, $\widehat{se}(\hat{y})$, describes how much $\hat{y}$ changes (for a given value of x) from sample to sample.

Example (continued):

(k) Find a 95% confidence interval for $\mu_{y|x=4}$, the population mean exam score for students that studied 4 hours.

When $x = 4$, Minitab indicates that $\widehat{se}(\hat{y}) = 4.89$. Now,

$$\hat{y} = 41.78 + 3.58(4) = 56.10$$

and therefore a 95% confidence interval for $\mu_{y|x=4}$ is

$$\hat{y} \pm t_{\alpha/2,n-p}\widehat{se}(\hat{y}) = 56.10 \pm 4.303(4.89) = 56.10 \pm 21.03 = (35.07, 77.13)$$

We are 95% confident that the population mean exam score for students that studied 4 hours is between 35.07 and 77.13.

- Are we at least 95% confident that $\mu_{y|x=4}$ significantly differs from 80? *Yes, since the CI excludes 80.*
- Are we at least 95% confident that $\mu_{y|x=4}$ significantly differs from 70? *No, since the CI includes 70.*

Be sure you are able find $\hat{y}$, $\widehat{se}(\hat{y})$, and the 95% confidence interval for $\mu_{y|x=4}$ in the **MTB 11.1** output on page 200.

Note: The estimated standard error of $\hat{y}$ depends upon x. In fact, $\widehat{se}(\hat{y})$ is *smallest* when $x = \bar{x}$, and becomes larger as x moves away from the mean. « show graph »

Definition 11.6: Coefficient of Determination R^2
The coefficient of determination R^2 describes the proportion of the variability in y that can be explained by the linear relationship with x (it indicates how well the regression line fits the data). For simple regression, the coefficient of determination is simply the square of the correlation:

$$R^2 = r^2$$

Example (continued):

(l) Find R^2 and interpret.

$$R^2 = r^2 = 0.960^2 = 0.922 = 92.2\%$$

Thus, 92.2% of the variability in exam scores (y) is explained via the linear relationship with hours studied (x). Be sure you are able find R^2 in the **MTB 11.1** *output on page 200.*

11.3 Assessing significance via analysis of variance (ANOVA) (optional)

The ANalysis Of VAriance (ANOVA) approach can also be used to assess significance. The total variability in the response variable y can be written via the following identity

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

which is commonly written as

$$SS_T = SS_R + SS_E.$$

The *total sum of squares*, SS_T, measures the total variability in the response variable y. SS_T can be broken down into two parts. The *regression sum of squares*, SS_R, is the amount of variability accounted by the least squares regression line, while the *error sum of squares*, SS_E, is the amount of variability not explained by the regression line. The sum of squares are typically written in an ANOVA table:

Source of Variation	**Degrees of Freedom**	**Sum of Squares**	**Mean Square**	**F**
Regression	$df_R = p - 1$	$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$MS_R = SS_R/df_R$	MS_R/MS_E
Error	$df_E = n - p$	$SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$MS_E = SS_E/df_E$	
Total	$df_T = n - 1$	$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$		

The variance of the points around the regression line, σ^2, is estimated by $s^2 = MS_E$. If s^2 is small, then the points tightly cluster around the least squares regression line; if s^2 is large, then the points loosely cluster around the line. Note that Minitab reports s instead of s^2.

Example (continued):
The Minitab ANOVA table for the exam score example is below.

```
S = 6.42364   R-Sq = 92.2%   R-Sq(adj) = 88.3%

Analysis of Variance
Source          DF       SS      MS      F      P
Regression       1   973.47  973.47  23.59  0.040
Residual Error   2    82.53   41.26
Total            3  1056.00
```

The total variation in exam scores is described by the total sum of squares, SS_T. The regression sum of squares, SS_R, is quite large (as a proportion of SS_T) indicating that a large proportion of the variability in exam scores is explained by least squares regression line. In fact, the coefficient of determination, R^2, describes the proportion of the variability in the response variable y explained by the least squares regression line. Specifically,

$$R^2 = \frac{SS_R}{SS_T} = \frac{973.47}{1056.00} = 0.922 = 92.2\%$$

which, when doing simple regression, is the same as the correlation squared (i.e. $R^2 = r^2$; we verified this earlier).

For simple regression, if we wish to test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ at the $\alpha = 0.05$ significance level, we could either do a $t-$test (as demonstrated in part (i) on page 189) or an $F-$test. The test statistic for the $F-$test is simply

$$F^* = \frac{MS_R}{MS_E} = \frac{973.47}{41.26} = 23.59$$

which is listed in the "F" column of the Minitab output. An $F-$distribution exists on the positive real line and has two parameters: the *numerator degrees of freedom* and the *denominator degrees of freedom*. Recall that p equals the number of parameters in our regression model. For simple regression, our model contains 2 parameters (β_0 and β_1), thus $p = 2$. The numerator degrees of freedom is $df_R = p - 1 = 2 - 1 = 1$ and the denominator degrees of freedom is $df_E = n - p = 4 - 2 = 2$. We will reject H_0 when MS_R is large relative to MS_E; in other words, we only reject in the right tail. Hence, the $p-$value for the $F-$test is

$$P(F_{(df_R,df_E)} > F^*) = P(F_{(1,2)} > 23.59) = 0.040$$

See Figure 11.6. Minitab automatically computes the $p-$value; it is shown in the "P" column. Because the $p-$value is less than our significance level α, then we reject H_0 and conclude that we have evidence that $\beta_1 \neq 0$. There is a significant linear relationship between hours studied and exam score.

The $p-$value can also be found using the applet at

`http://www.stat.uiowa.edu/~mbognar/applets/f.html`

Enter 1 in the df_1 box, enter 2 in the df_2 box, and enter 23.59 in the x box. The probability $P(F_{(1,2)} > 23.59)$ is computed and displayed in the pink box. Note that the $p-$value for this ANOVA test is *identical* to the $p-$value from the $t-$test in part (i) on page 189 (both tests will always match when doing simple regression). Interesting fact: notice that squaring the test statistic from part (i) yields the test statistic from the $F-$test, i.e. $23.59 = F^* = (t^*)^2 = 4.86^2$ (take more classes to get the details).

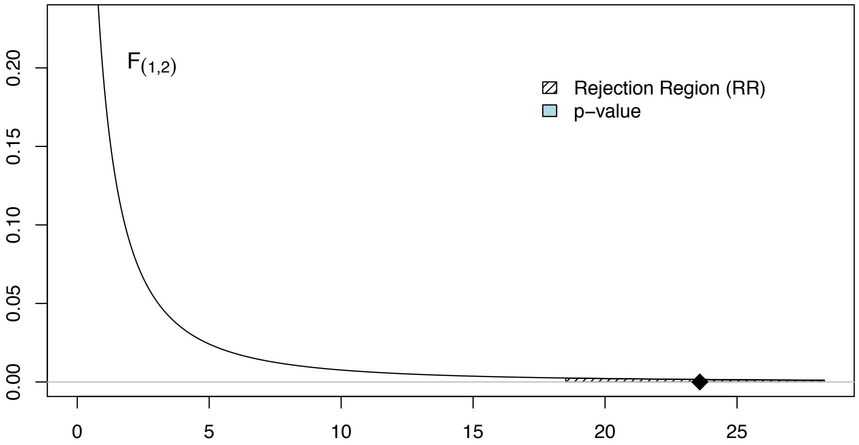

Figure 11.6: Exam Scores: Test statistic F^* is denoted by ◆, the rejection region and p−value are also shown (the critical value is $F_{\alpha;df_R,df_E} = F_{0.05;1,2} = 18.513$). Note that the total area in the rejection region is equal to $\alpha = 0.05$.

11.4 Statistical adjustment

Example: *Statistical Adjustment*
We have salary data on white and minority employees at a large company. The years of experience x and salary y (in thousands) of 7 minority employees and 6 white employees is

Minority	x = years:	3	3	5	6	8	8	10
	y = salary:	17	19	20	21	22	24	25

White	x = years:	1	2	3	3	4	5
	y = salary:	18.0	19.3	21.6	19.6	21.9	23.2

The summary statistics are

Minority	$\bar{x} = 6.143$	$s_x = 2.672$	$r = 0.956$
	$\bar{y} = 21.143$	$s_y = 2.795$	

White	$\bar{x} = 3.0$	$s_x = 1.414$	$r = 0.946$
	$\bar{y} = 20.6$	$s_y = 1.944$	

Based upon the mean salaries (i.e. the $\bar{y}$'s), minority employees make more than white employees. However, minority employees have more years of experience! This makes a comparison of these "unadjusted" mean salaries unfair. We want to adjust/account for years of experience before making salary comparisons.

(a) Determine the least squares regression line for each group.

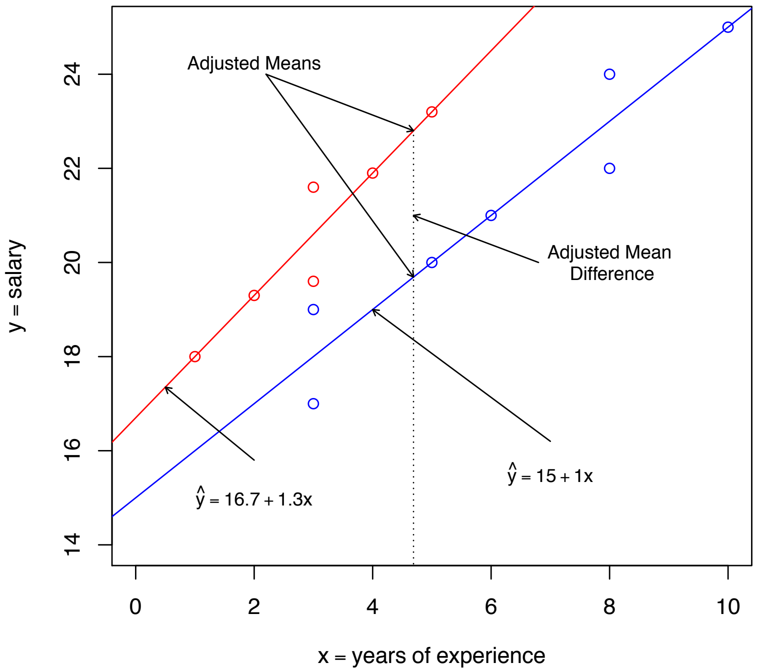

Figure 11.7: Nursing Salaries: Scatterplot and least squares regression line for the Minority nurses (in blue) and White nurses (in red).

$$\begin{aligned} \textit{Minority}: \quad \hat{\beta}_1 &= r\frac{s_y}{s_x} = 0.956\frac{2.795}{2.672} = 1.0 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} = 21.143 - 1.0(6.143) = 15.0 \\ \hat{y} &= 15.0 + 1.0x \end{aligned}$$

$$\begin{aligned} \textit{White}: \quad \hat{\beta}_1 &= r\frac{s_y}{s_x} = 0.946\frac{1.944}{1.414} = 1.3 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} = 20.6 - 1.3(3.0) = 16.7 \\ \hat{y} &= 16.7 + 1.3x \end{aligned}$$

The Minitab output for the two regression analyses is called **MTB 11.2** *on page 201. Scatterplots and the least squares regression lines for each group is shown in Figure 11.7.*

(b) Do white or minority employees have a higher mean (average) starting (i.e. no years of experience) salary?

$$\begin{aligned} \textit{Minority} &\rightarrow \hat{y} = 15.0 + 1.0(0) = 15.0 \\ \textit{White} &\rightarrow \hat{y} = 16.7 + 1.3(0) = 16.7 \end{aligned}$$

The mean starting pay of white nurses is approximately $1700 more than minority nurses.

(c) Do white or minority employees get pay increases at a faster rate?

$$\begin{aligned} \textit{Minority} &\rightarrow \hat{\beta}_1 = 1.0 \\ \textit{White} &\rightarrow \hat{\beta}_1 = 1.3 \end{aligned}$$

White nurses get pay raises at a faster rate. On average, minority nurses get approximately $1000 more for every extra year worked, while white nurses get $1300.

(d) After 5 years, do white or minority employees have a higher mean salary?

$$\begin{aligned} \textit{Minority} &\rightarrow \hat{y} = 15.0 + 1.0(5) = 20.0 \\ \textit{White} &\rightarrow \hat{y} = 16.7 + 1.3(5) = 23.2 \end{aligned}$$

After 5 years white nurses are making approximately $3200 more than minority nurses, on average.

(e) Determine the *adjusted mean salaries.*

Overall, the average amount of experience is

$$\frac{(3+3+5+6+8+8+10)+(1+2+3+3+4+5)}{13} = 4.69$$

Therefore, the adjusted mean salaries are

$$\begin{aligned} \textit{Minority} &\rightarrow \hat{y} = 15.0 + 1.0(4.69) = 19.69 \\ \textit{White} &\rightarrow \hat{y} = 16.7 + 1.3(4.69) = 22.80 \end{aligned}$$

See Figure 11.7.

(f) What is the *adjusted mean difference*?

The adjusted mean difference is the difference between the adjusted means: $22.80 - 19.69 = 3.11$. *Hence, the adjusted mean difference is $3110. See Figure 11.7.*

(g) In summary, after adjusting for years of experience, does there appear to be salary discrimination?

Yes. After adjusting for years of experience, white nurses are making approximately $3110 more than minority nurses, on average.

To learn how to assess statistical significance in this example, take another class.

11.5 Computational formulas (optional)

The estimated standard error of $\hat{\beta}_1$ is

$$\widehat{se}(\hat{\beta}_1) = \sqrt{\frac{s^2}{s_{xx}}}$$

where

$$s^2 = \frac{SSE}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

and

$$s_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Minitab computes and reports s instead of s^2. The estimated standard error of $\hat{\beta}_0$ is

$$\widehat{se}(\hat{\beta}_0) = \sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}.$$

The estimated standard error of $\hat{y}$ at $x = x_0$ is

$$\widehat{se}(\hat{y}) = \sqrt{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)}.$$

Note that $\widehat{se}(\hat{y})$ is smallest when $x = \bar{x}$. A $(1-\alpha)100\%$ prediction interval on a new observation at x_0 is

$$\hat{y}_0 \pm t_{\alpha/2,n-p}\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)}$$

where $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

11.6 Exercises

♥ = answers are provided beginning on page 213.

11.1 At a large hospital, the salaries (y, in thousands of dollars) and years of experience (x) of six randomly chosen female nurses are

x = experience:	6	7	9	10	13	15
y = salary:	40	41	43	45	46	49

The Minitab output is shown in **MTB 11.3** on page 202.

(a) By hand, compute Pearsons sample correlation coefficient r. *Be sure you can find r on the Minitab output.*

(b) By hand, determine least squares regression line. *Find the least squares regression line on the Minitab output.*

(c) Carefully make a scatter-plot of the dataset and draw the regression line (place the explanatory variable x on the horizontal axis, and the response variable y on the vertical axis). You can get graphpaper at

`http://www.stat.uiowa.edu/~mbognar/applets/graphpaper.pdf`

(d) On average, each extra year of experience yields how much extra pay?

(e) What is the approximate average starting pay?

(f) Approximate the mean salary for female nurses with 12 years of experience, i.e. approximate $\mu_{y|x=12}$.

(g) By hand, approximate the mean salary for female nurses with 6 years of experience, i.e. approximate $\mu_{y|x=6}$.

(h) By hand, find a 95% confidence interval for the population mean salary of female nurses with 6 years of experience, i.e. find a 95% CI for $\mu_{y|x=6}$. Interpret the CI. *Hint: According to Minitab,* $\widehat{se}(\hat{y}) = 0.448$. *Find* $\hat{y}$, $\widehat{se}(\hat{y})$, *and the CI on the Minitab output.*

(i) Is there a significant linear relationship between years of experience and salary? *Hint: According to Minitab,* $\widehat{se}(\hat{\beta}_1) = 0.0878$. *You must state* H_0 *and* H_a *(use* $\alpha = 0.05$*), find the test statistic and critical value, plot the rejection region, and state your decision and final conclusion. Find* $\hat{\beta}_1$, $\widehat{se}(\hat{\beta}_1)$, *and the test statistic* t^* *on the Minitab output.*

(j) Approximate the $p-$value for the test in 11.1(i) using the $t-$table. Based upon your $p-$value, is there a significant linear relationship between years of experience and salary? Why? *Find* $p-$*value on the Minitab output.*

(k) Use the $t-$Probability Applet at

`http://www.stat.uiowa.edu/~mbognar/applets/t.html`

to precisely determine the $p-$value for the test in 11.1(i).

(l) Find a 95% confidence interval for β_1. Based upon your CI, is there a significant linear relationship between years of experience and salary? Why? *Hint: According to Minitab,* $\widehat{se}(\hat{\beta}_1) = 0.0878$. *Find* $\hat{\beta}_1$ *and* $\widehat{se}(\hat{\beta}_1)$ *on the Minitab output.*

(m) Find a 95% confidence interval for the (population) mean starting salary, i.e. find a 95% CI for $\beta_0 = \mu_{y|x=0}$. *Hint: According to Minitab,* $\widehat{se}(\hat{\beta}_0) = 0.9208$. *Find* $\hat{\beta}_0$ *and* $\widehat{se}(\beta_0)$ *on the Minitab output.*

(n) In reference to question 11.1(m), is the population mean starting salary significantly different than 40 (i.e. $40,000)? Why?

(o) By hand, find the coefficient of determination, R^2. Interpret. *Find* R^2 *on the Minitab output.*

11.2 ♥ Because ethanol contains less energy than gasoline, a researcher wants to determine if the mileage of a car (y) is affected by the percent ethanol in the gasoline (x). The true population regression line relating the mean mileage y for a given ethanol content x (in percent) is $\mu_{y|x} = \beta_0 + \beta_1 x$. In a controlled environment, the mileage of a car is recorded when refueled 7 times using between 0% and 10% ethanol. The results from the Minitab analysis is shown in **MTB 11.4** on page 202.

(a) Interpret the slope $\hat{\beta}_1$.

(b) Interpret the intercept $\hat{\beta}_0$.

(c) We would like to test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 < 0$ at the $\alpha = 0.05$ significance level (i.e. we want to test if the addition of ethanol significantly *decreases* gas mileage). What is the $p-$value for this test? What is your conclusion?

(d) Approximate the mean mileage when 7% ethanol is used.

(e) Find a 95% confidence interval for $\mu_{y|x=7}$. Interpret. *Hint:* $\widehat{se}(\hat{y}) = 0.208$.

(f) Find a 95% confidence interval for β_1. Interpret. *Hint:* $\widehat{se}(\hat{\beta}_1) = 0.0485$.

11.3 ♥ A student seeks to determine if weight (x) affects the time it takes adults to run 100 yards (y). The true population regression line relating the mean time y for a given weight x (in pounds) is $\mu_{y|x} = \beta_0 + \beta_1 x$. A random sample of 7 adults were weighed and timed. The Minitab output is shown in **MTB 11.5** on page 203.

(a) Suppose we wish to determine if there is a significant linear relationship between weight and time. To answer this question, what test do we have to perform? Assume $\alpha = 0.05$. What is the p–value for this test?

(b) Find a 95% confidence interval for $\mu_{y|x=200}$. *Hint:* $\widehat{se}(\hat{y}) = 0.277$. *Find* $\widehat{se}(\hat{y})$ *on the Minitab output.*

(c) Is the population mean time for adults weighing 200 pounds significantly less than 20 seconds? Why?

11.4 At a large hospital, the salaries (y, in thousands of dollars) and years of experience (x) of six randomly chosen female nurses were

x_F = experience:	6	7	9	10	13	15
y_F = salary:	30	31	33	35	36	39

while five randomly chosen male nurses yielded

x_M = experience:	2	3	3	5	7
y_M = salary:	28	29	30	32	34

(a) Find the un-adjusted mean salaries for each gender (i.e. find $\bar{y}_F$ and $\bar{y}_M$). Based upon the un-adjusted means, is it fair to claim gender discrimination against males? Why?

(b) By hand, find r_F, the correlation coefficient for the female nurses. *Find the correlation coefficient for the female nurses on the* **MTB 11.6** *output on page 204.*

(c) Find Cov(x_M, y_M), the covariance between experience and salary for the male nurses. Hint: $r_M = 0.986165$, $s_{x_M} = 2$, and $s_{y_M} = 2.408319$.

(d) By hand, determine least squares regression line for each gender. *Compare your regression lines to the* **MTB 11.6** *output.*

(e) Carefully and accurately make a scatter-plot using different plotting symbols for each gender (place the explanatory variable x on the horizontal axis, and the response variable y on the vertical axis). Plot both regression lines. You can get graphpaper at

`http://www.stat.uiowa.edu/~mbognar/applets/graphpaper.pdf`

(f) Which gender has the higher average starting pay (i.e. when years of experience is 0)? How much difference exists in average starting pay? Mark this difference in your scatterplot.

(g) Which gender accumulates yearly pay increases at a faster rate? In a detailed fashion, describe the difference in yearly pay increases.

(h) For nurses with 6 years of experience, what is the difference in average pay between the genders? Mark this difference in your graph.

(i) Find the adjusted means. Mark the adjusted means on your graph.

(j) Find the adjusted mean difference. After adjusting for experience, which gender has the higher salary? Mark the adjusted mean difference on your graph.

(k) In summary, do the un-adjusted mean salaries in 11.4(a) provide a clear/fair picture of salary structure? Explain.

This exercise highlighted the fact that ignoring important factors in any *study can yield misleading results (this was also demonstrated in the* Simpson's Paradox *example on page 170). Don't blindly accept the results from studies at face value; you can very easily come to the wrong conclusion. If you ask the proper probing questions (such as determining if all important factors were accounted for), then you can make a better assessment of the quality of the study (and its conclusions).*

11.7 Minitab output

MTB 11.1 Minitab output for the hours studied versus exam score example.

```
The regression equation is
y (score) = 41.78 + 3.58 x (hours)

Predictor    Coef  SE Coef     T      P
Constant    41.78    7.368  5.67  0.030
x (hours)    3.58   0.7368  4.86  0.040

S = 6.42364   R-Sq = 92.2%   R-Sq(adj) = 88.3%

Analysis of Variance
Source          DF       SS      MS      F      P
Regression       1   973.47  973.47  23.59  0.040
Residual Error   2    82.53   41.26
Total            3  1056.00

Predicted Values for New Observations
New
Obs    Fit  SE Fit      95% CI           95% PI
  1  56.11    4.89  (35.08, 77.14)  (21.38,  90.83)
  2  81.16    3.53  (65.95, 96.36)  (49.61, 112.70)

Values of Predictors for New Observations
New
Obs  x (hours)
  1        4.0
  2       11.0
```

MTB 11.2 Minitab output for the salary/years of experience dataset.

```
Minority Employees
==================
The regression equation is
y (salary) = 15.0 + 1.00 x (years)

Predictor     Coef  SE Coef      T      P
Constant   15.0000   0.9048  16.58  0.000
x (years)   1.0000   0.1366   7.32  0.001

S = 0.894427   R-Sq = 91.5%   R-Sq(adj) = 89.8%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       1  42.857  42.857  53.57  0.001
Residual Error   5   4.000   0.800
Total            6  46.857


White Employees
===============
The regression equation is
y (salary) = 16.7 + 1.30 x (years)

Predictor     Coef  SE Coef      T      P
Constant   16.7000   0.7303  22.87  0.000
x (years)   1.3000   0.2236   5.81  0.004

S = 0.707107   R-Sq = 89.4%   R-Sq(adj) = 86.8%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       1  16.900  16.900  33.80  0.004
Residual Error   4   2.000   0.500
Total            5  18.900
```

MTB 11.3 Minitab output for Exercise 11.1.

```
Pearson correlation of x and y = 0.983
P-Value = 0.000

The regression equation is
y = 34.5 + 0.950 x

Predictor     Coef  SE Coef      T      P
Constant   34.5000   0.9208  37.47  0.000
x          0.95000  0.08780  10.82  0.000

S = 0.680074   R-Sq = 96.7%   R-Sq(adj) = 95.9%

Analysis of Variance
Source          DF      SS      MS       F      P
Regression       1  54.150  54.150  117.08  0.000
Residual Error   4   1.850   0.463
Total            5  56.000

Predicted Values for New Observations
New Obs     Fit  SE Fit       95% CI            95% PI
      1  40.200   0.448  (38.957, 41.443)  (37.939, 42.461)

Values of Predictors for New Observations
New Obs     x
      1  6.00
```

MTB 11.4 Minitab output for Exercise 11.2.

```
The regression equation is
y = 33.0 - 0.250 x

Predictor      Coef  SE Coef       T      P
Constant    32.9643   0.3043  108.33  0.000
x          -0.25000  0.04855   -5.15  0.004

S = 0.485504   R-Sq = 84.1%   R-Sq(adj) = 81.0%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       1  6.2500  6.2500  26.52  0.004
Residual Error   5  1.1786  0.2357
Total            6  7.4286

Predicted Values for New Observations
New Obs     Fit  SE Fit       95% CI            95% PI
      1  31.214   0.208  (30.681, 31.748)  (29.857, 32.572)

Values of Predictors for New Observations
New Obs     x
      1  7.00
```

MTB 11.5 Minitab output for Exercise 11.3.

```
The regression equation is
y = 11.7 + 0.0248 x

Predictor      Coef   SE Coef      T      P
Constant    11.7032    0.7188  16.28  0.000
x          0.024842  0.004442   5.59  0.003

S = 0.517484   R-Sq = 86.2%   R-Sq(adj) = 83.5%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       1  8.3753  8.3753  31.28  0.003
Residual Error   5  1.3389  0.2678
Total            6  9.7143

Predicted Values for New Observations
New Obs     Fit  SE Fit       95% CI            95% PI
      1  16.672   0.277  (15.958, 17.385)  (15.162, 18.181)

Values of Predictors for New Observations
New Obs    x
      1  200
```

MTB 11.6 Minitab output for Exercise 11.4.

```
Minitab output for females:
===========================
Correlations: xf, yf
Pearson correlation of xf and yf = 0.983

The regression equation is
yf = 24.5 + 0.950 xf

Predictor     Coef  SE Coef      T      P
Constant   24.5000   0.9208  26.61  0.000
xf         0.95000  0.08780  10.82  0.000

S = 0.680074   R-Sq = 96.7%   R-Sq(adj) = 95.9%

Analysis of Variance
Source          DF      SS      MS       F      P
Regression       1  54.150  54.150  117.08  0.000
Residual Error   4   1.850   0.463
Total            5  56.000


Minitab output for males:
=========================
Correlations: xm, ym
Pearson correlation of xm and ym = 0.986

The regression equation is
ym = 25.8 + 1.19 xm

Predictor     Coef  SE Coef      T      P
Constant   25.8500   0.5050  51.19  0.000
xm          1.1875   0.1152  10.30  0.002

S = 0.460977   R-Sq = 97.3%   R-Sq(adj) = 96.3%

Analysis of Variance
Source          DF      SS      MS       F      P
Regression       1  22.563  22.563  106.18  0.002
Residual Error   3   0.638   0.213
Total            4  23.200
```