where

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

and

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

R 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 229.

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 R output is shown in **R 11.3** on page 209.

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

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

If you wish, you can use R to make the scatterplot with the command `plot(x,y)`. If you then use the command `abline(lm(y~x))`, R will plot the least squares regression line on your scatter plot. How cool is that!

(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 R,* $\widehat{se}(\hat{y}) = 0.448$*. Find* $\hat{y}$*,* $\widehat{se}(\hat{y})$*, and the CI on the R output.*

(i) Is there a significant linear relationship between years of experience and salary? *Hint: According to R,* $\widehat{se}(\hat{\beta}_1) = 0.0878$*. 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 R 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 R 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 R,* $\widehat{se}(\hat{\beta}_1) = 0.0878$*. Find* $\hat{\beta}_1$ *and* $\widehat{se}(\hat{\beta}_1)$ *on the R 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 R,* $\widehat{se}(\hat{\beta}_0) = 0.9208$*. Find* $\hat{\beta}_0$ *and* $\widehat{se}(\beta_0)$ *on the R 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 R 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 R analysis is shown in **R 11.4** on page 210.

(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:* $\widehat{se}(\hat{\beta}_1) = 0.04855$.

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 R output is shown in **R 11.5** on page 210.

(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 R 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* **R 11.6** *output on page 211.*

(c) Find $\text{Cov}(x_M, y_M)$, 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* **R 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 174). 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 R output

R 11.1 R output for the hours studied versus exam score example.

```
# Input the data.
> x <- c(10,14,2,10)
> y <- c(82,94,50,70)

# The "lm" command does simple linear regression. The "y~x"
# argument specifies the model "yhat = beta0hat + beta1hat x".
# We store the regression results in the object "study.results".
> study.results <- lm(y~x)

# Lets check out the results.
> summary(study.results)

Call:
lm(formula = y ~ x)

Residuals:
     1      2      3      4
 4.421  2.105  1.053 -7.579

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  41.7895     7.3684   5.671   0.0297
x             3.5789     0.7368   4.857   0.0399

Residual standard error: 6.424 on 2 degrees of freedom
Multiple R-squared:  0.9219,        Adjusted R-squared:  0.8828
F-statistic: 23.59 on 1 and 2 DF,  p-value: 0.03987

# Analysis of Variance (ANOVA) table summarizing regression.
> anova(study.results)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
x          1 973.47  973.47  23.592 0.03987
Residuals  2  82.53   41.26

# We want to get predictions and confidence intervals at
# x=4 and 11. The "fit" column specifies the yhat values,
# and the "lwr" and "upr" specify the CI.
> predict(study.results, list(x=c(4,11)), interval="confidence", se.fit=TRUE)
$fit
       fit      lwr      upr
1 56.10526 35.07537 77.13516
2 81.15789 65.95331 96.36248
$se.fit
[1] 4.89 3.53
```

R 11.2 R output for the salary/years of experience dataset.

```
Minority Employees
==================
> x <- c(3,3,5,6,8,8,10)
> y <- c(17,19,20,21,22,24,25)
> m.results <- lm(y˜x)
> summary(m.results)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.0000     0.9048  16.578 1.46e-05
x             1.0000     0.1366   7.319 0.000746

Residual standard error: 0.8944 on 5 degrees of freedom
Multiple R-squared:  0.9146,       Adjusted R-squared:  0.8976
F-statistic: 53.57 on 1 and 5 DF,  p-value: 0.0007461

> anova(m.results)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 42.857  42.857  53.571 0.0007461
Residuals  5  4.000   0.800

White Employees
===============
> x <- c(1,2,3,3,4,5)
> y <- c(18.0,19.3,21.6,19.6,21.9,23.2)
> w.results <- lm(y˜x)
> summary(w.results)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.7000     0.7303  22.867 2.17e-05
x             1.3000     0.2236   5.814  0.00436

Residual standard error: 0.7071 on 4 degrees of freedom
Multiple R-squared:  0.8942,       Adjusted R-squared:  0.8677
F-statistic:  33.8 on 1 and 4 DF,  p-value: 0.004357

> anova(w.results)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
x          1   16.9    16.9    33.8 0.004357
Residuals  4    2.0     0.5
```

R 11.3 R output for Exercise 11.1.

```
> x <- c(6,7,9,10,13,15)
> y <- c(40,41,43,45,46,49)

> mean(x)
[1] 10
> sd(x)
[1] 3.464102
> mean(y)
[1] 44
> sd(y)
[1] 3.34664
> cov(x,y)
[1] 11.4
> cor(x,y)
[1] 0.9833434

> salary.results <- lm(y~x)
> summary(salary.results)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.5000     0.9208   37.47 3.03e-06
x             0.9500     0.0878   10.82 0.000414

Residual standard error: 0.6801 on 4 degrees of freedom
Multiple R-squared:  0.967,Adjusted R-squared:  0.9587
F-statistic: 117.1 on 1 and 4 DF,  p-value: 0.0004139

> anova(salary.results)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1  54.15  54.150  117.08 0.0004139
Residuals  4   1.85   0.463

> predict(salary.results, list(x=c(6)), interval="confidence", se.fit=TRUE)
$fit
   fit      lwr      upr
1 40.2 38.95704 41.44296
$se.fit
[1] 0.448
```

R 11.4 R output for Exercise 11.2.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.9643     0.3043  108.33    0.000
x           -0.25000    0.04855   -5.15    0.004

Residual standard error: 0.485504 on 5 degrees of freedom
Multiple R-squared:  0.841,        Adjusted R-squared:  0.810
F-statistic: 26.52 on 1 and 5 DF,  p-value: 0.004

> anova(salary.results)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1  6.250   6.250   26.52     0.004
Residuals  5  1.179   0.236

> predict(salary.results, list(x=c(7)), interval="confidence", se.fit=TRUE)
$fit
     fit    lwr    upr
1 31.214 30.681 31.748
$se.fit
[1] 0.208
```

R 11.5 R output for Exercise 11.3.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.7032     0.7188   16.28    0.000
x           0.024842   0.004442    5.59    0.003

Residual standard error: 0.517484 on 5 degrees of freedom
Multiple R-squared:  0.862,        Adjusted R-squared:  0.835
F-statistic: 31.28 on 1 and 5 DF,  p-value: 0.003

> anova(salary.results)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
x          1 8.3753  8.3753   31.28   0.003
Residuals  5 1.3389  0.2678

> predict(salary.results, list(x=c(200)), interval="confidence", se.fit=TRUE)
$fit
   fit      lwr    upr
1 16.672 15.958 17.385
$se.fit
[1] 0.277
```

R 11.6 R output for Exercise 11.4.

```
R output for females:
=====================
> x.f <- c(6,7,9,10,13,15)
> y.f <- c(30,31,33,35,36,39)

> mean(x.f)
[1] 10
> sd(x.f)
[1] 3.464102
> mean(y.f)
[1] 34
> sd(y.f)
[1] 3.34664
> cov(x.f,y.f)
[1] 11.4
> cor(x.f,y.f)
[1] 0.9833434

> female.results <- lm(y.f~x.f)
> summary(female.results)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.5000     0.9208   26.61 1.19e-05
x.f           0.9500     0.0878   10.82 0.000414

Residual standard error: 0.6801 on 4 degrees of freedom
Multiple R-squared:  0.967,        Adjusted R-squared:  0.9587
F-statistic: 117.1 on 1 and 4 DF,  p-value: 0.0004139

> anova(female.results)

Analysis of Variance Table

Response: y.f
          Df Sum Sq Mean Sq F value    Pr(>F)
x.f        1  54.15  54.150  117.08 0.0004139
Residuals  4   1.85   0.462


R output for males:
===================
> x.m <- c(2,3,3,5,7)
> y.m <- c(28,29,30,32,34)

> mean(x.m)
[1] 4
> sd(x.m)
[1] 2
> mean(y.m)
[1] 30.6
> sd(y.m)
[1] 2.408319
> cov(x.m,y.m)
[1] 4.75
> cor(x.m,y.m)
[1] 0.9861651

> male.results <- lm(y.m~x.m)
> summary(male.results)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.8500     0.5050   51.19 1.64e-05
x.m           1.1875     0.1152   10.30  0.00195

Residual standard error: 0.461 on 3 degrees of freedom
Multiple R-squared:  0.9725,        Adjusted R-squared:  0.9634
F-statistic: 106.2 on 1 and 3 DF,  p-value: 0.001949

> anova(male.results)

Analysis of Variance Table

Response: y.m
          Df  Sum Sq Mean Sq F value   Pr(>F)
x.m        1 22.5625 22.5625  106.18 0.001949
Residuals  3  0.6375  0.2125

Can you figure out what these commands do? You will use
this result when finding the adjusted means.
> x.all <- c(x.f, x.m)
> mean(x.all)
[1] 7.272727

How about this? Note:
   "pch" stands for "plotting character"
   "xlim=c(0,15)" forces x-axis to be from 0 to 15
   "ylim=c(27,40)" forces y-axis to be from 27 to 40
> plot(x.f, y.f, col="red", pch=1, xlim=c(0,15), ylim=c(23,45), xlab="years" ylab="salary")
> abline(lm(y.f~x.f), col="red")
> points(x.m, y.m, col="blue", pch=2)
> abline(lm(y.m~x.m), col="blue")
```