14.8 Appendix: R Output for the Samara Example
by EV Nordheim, MK Clayton & BS Yandell, December 11, 2003

In this appendix we will briefly illustrate some of the regression commands available in R by
using the samara data and the \texttt{lm} command. Note that \texttt{lm} allows for the possibility of having
several predictors. This is important in multiple regression, a topic we will not pursue in
this chapter.

We have entered the data with $x$ in column V1 and $y$ in column V2.

\begin{verbatim}
> samara = read.table("http://www.stat.wisc.edu/~st571-1/data/samara.dat")
> x = samara$V1
> y = samara$V2
\end{verbatim}

Alternatively, you can enter data as we have sometimes done:

\begin{verbatim}
> x = c(1.72, 1.72, 1.77, 1.78, 1.82, 1.85, 1.88, 1.93, 1.96, 1.96, 
+       2, 2, 2.03, 2.06)
> y = c(0.85, 0.86, 0.72, 0.79, 0.82, 0.8, 0.99, 0.94, 0.82, 0.89, 
+       0.95, 1, 0.98, 0.99)
\end{verbatim}

The R command \texttt{plot} produces a scatterplot of $y$ versus $x$. The line

\begin{verbatim}
> plot(x, y)
\end{verbatim}

To regress $y$ on $x$, we proceed as follows:
> samara.lm = lm(y ~ x)
> summary(samara.lm)

Call:
  lm(formula = y ~ x)

Residuals:
     Min      1Q  Median      3Q     Max -0.10345 -0.03416  0.00803  0.04917  0.11057

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1551  0.2992 -0.518 0.61355
          x     0.5503  0.1579  3.485 0.00451 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06605 on 12 degrees of freedom
Multiple R-Squared: 0.503,    Adjusted R-squared: 0.4616
F-statistic: 12.14 on 1 and 12 DF,  p-value: 0.004506

> anova(samara.lm)

Analysis of Variance Table

Response: y

          Df Sum Sq Mean Sq  F value Pr(>F)
          x     1 0.052985 0.052985 12.144  0.004506 **
Residuals 12 0.052357  0.004363
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The object samara.lm contains all the information about the regression fit. The summary command provides the form of the regression equation, the estimates of the intercept “(Intercept)” and slope (“x”). Also provided are the estimated standard deviations (“Std. Error”) of these quantities, T values corresponding to $H_0: b_0 = 0$ and $H_0: b_1 = 0$, and p-values for these tests. R then prints an estimate of $\sigma_e$ and $R^2$. The notation Adjusted R-squared refers to an adjusted version of $R^2$ important in multiple regression. Finally, the anova produces the ANOVA table for the regression, including the F value and p-value for $H_0: b_1 = 0$.

You can add the regression line to the data plot using the following command. However, we will not show the plot here.

> lines(x, predict(samara.lm))
R has a number of definitions of residual suitable for different purposes. To produce a residual plot as we have defined it, we use the `plot` command.

```r
> plot(samara.lm, which = 1)
```

If we want to see all the predicted values and residuals, we can use the commands `predict(samara.lm)` and `resid(samara.lm)`, respectively. Actually, there are four possible plots for `lm` objects. For instance, the second plot is the Q-Q plot:

```r
> plot(samara.lm, which = 2)
```

R can also be used to obtain $\hat{Y}_{est}$ and its estimated standard error and to obtain confidence intervals for $\hat{Y}_{est}$ and $\hat{Y}_{pred}$. We do this below for the value $x_\ast = 1.80$ by using the `predict` command. We use it twice to get confidence and prediction intervals.
> predict(samara.lm, data.frame(x = 1.8), se.fit = TRUE, interval = "confidence")

$fit
   fit   lwr   upr
[1,] 0.8354017 0.7857125 0.885091

$se.fit
[1] 0.02280564

$df
[1] 12

$residual.scale
[1] 0.0660539

> predict(samara.lm, data.frame(x = 1.8), se.fit = TRUE, interval = "prediction")

$fit
   fit   lwr   upr
[1,] 0.8354017 0.6831462 0.9876571

$se.fit
[1] 0.02280564

$df
[1] 12

$residual.scale
[1] 0.0660539