Statistics 354 (Fall 2018)
Example: Temperature vs. Hardness

Consider the temperature (x) and hardness (y) data presented in Table 1.1 on page 5.

Run i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x_i	30	30	30	30	40	40	40	50	50	50	60	60	60	60
y_i	55.8	59.1	54.8	54.6	43.1	42.2	45.2	31.6	30.9	30.8	17.5	20.5	17.2	16.9

In order to fit a simple linear regression model to this data, we begin by calculating the following summary statistics.

- $n = 14$
- $\sum x_i = 630$
- $\sum x_i^2 = 30300$
- $s_{xx} = \sum (x_i - \overline{x})^2 = \left(\sum x_i^2\right) - \frac{1}{n}\left(\sum x_i\right)^2 = 30300 - \frac{630^2}{14} = 1950$
- $\sum y_i = 520.2$
- $\sum x_i y_i = 20940$
- $s_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right) = 20940 - \frac{630 \cdot 520.2}{14} = -2469$
- $\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}} = -\frac{2469}{1950} \doteq -1.266$
- $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = \frac{520.2}{14} - \frac{-2469}{1950} \cdot \frac{630}{14} \doteq 94.134$

  Note that the textbook answer of $\hat{\beta}_0 = 94.123$ is due to incorrect rounding.

Therefore, the equation of the simple linear regression line (also known as the line of best fit or the least squares line) is

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x = 94.134 - 1.266x.$$

The fitted values are $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and displayed in the following table.

Run i	1, 2, 3, 4	5, 6, 7	8, 9, 10	11, 12, 13, 14
x_i	30	40	50	60
$\hat{\mu}_i$	56.14945	43.48791	30.82637	18.16484

The residuals are $e_i = y_i - \hat{\mu}_i$ and the residual sum of squares is

$$\text{SSR} = S(\hat{\beta}_0, \hat{\beta}_1) = \sum (y_i - \hat{\mu}_i)^2 = \sum e_i^2 \doteq 26.820.$$

The mean square error is

$$\text{MSE} = s^2 = \frac{SSR}{n-2} = \frac{1}{n-2} \sum (y_i - \hat{\mu}_i)^2 = \frac{1}{n-2} \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \doteq 2.235.$$

In order to perform inferences about β_1 and β_0, we use appropriate t-tests. That is, since $n = 14$, we use t-tests with $df = n - 2 = 12$. For example, to compute 95% confidence intervals, we use the critical values corresponding to $\alpha = 0.05$, $df = 12$, namely $t(0.975; 12) = -t(0.025; 12) \doteq 2.18$.

This implies that a 95% confidence interval for β_1 is

$$\left[\hat{\beta}_1 - t(0.975; 12) \frac{s}{\sqrt{s_{xx}}},\ \hat{\beta}_1 + t(0.975; 12) \frac{s}{\sqrt{s_{xx}}} \right] \doteq [-1.33992,\ -1.19239]$$

and a 95% confidence interval for β_0 is

$$\left[\hat{\beta}_0 - t(0.975; 12) s \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{s_{xx}}},\ \hat{\beta}_0 + t(0.975; 12) s \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{s_{xx}}} \right] = [90.70242,\ 97.56571].$$

Note that the textbook answers are slightly different due to incorrect rounding.

To test the hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$, we compute the t-statistic

$$t_0 = \frac{\hat{\beta}_1}{s/\sqrt{s_{xx}}} \doteq -37.4.$$

This corresponds to a p-value of virtually 0. Therefore, there is overwhelming evidence to reject the null hypothesis and conclude that $\beta_1 \neq 0$. In other words, temperature has a major impact on hardness.

The R code that was used for the calculations is given below.

```
> x <- c(30, 30, 30, 30, 40, 40, 40, 50, 50, 50, 60, 60, 60, 60)
> y <- c(55.8, 59.1, 54.8, 54.6, 43.1, 42.2, 45.2, 31.6,
            30.9, 30.8, 17.5, 20.5, 17.2, 16.9)
> plot(x,y)
> n = 14
> sum(x)
[1] 630
> sum(x^2)
[1] 30300
> sxx = sum(x^2)-sum(x)^2/n
> sxx
[1] 1950
> sum(y)
[1] 520.2
> sum(x*y)
[1] 20940
> sxy = sum(x*y)-sum(x)*sum(y)/n
> sxy
[1] -2469
> hatbeta1 = sxy/sxx
> hatbeta1
[1] -1.266154
> hatbeta0 = sum(y)/n - hatbeta1*sum(x)/n
> hatbeta0
[1] 94.13407
> hatmu = hatbeta0 +hatbeta1*x
> hatmu
[1] 56.14945 56.14945 56.14945 56.14945 43.48791 43.48791 43.48791
           30.82637 30.82637 30.82637 18.16484 18.16484 18.16484 18.16484
> residuals = y - hatmu
> residuals
[1] -0.34945055  2.95054945 -1.34945055 -1.54945055 -0.38791209 -1.28791209
          1.71208791  0.77362637  0.07362637 -0.02637363 -0.66483516
               2.33516484 -0.96483516 -1.26483516
> SSR = sum(residuals^2)
> SSR
[1] 26.82044
> s2 = 1/(n-2)*sum((y-hatmu)^2)
> s2
[1] 2.235037
```

```
> qt(.975, df=12)
[1] 2.178813
> hatbeta1 - qt(.975, df=12)*sqrt(s2)/sqrt(sxx)
[1] -1.339918
> hatbeta1 + qt(.975, df=12)*sqrt(s2)/sqrt(sxx)
[1] -1.19239
> hatbeta0 - qt(.975, df=12)*sqrt(s2)*sqrt(1/n + mean(x)^2/sxx)
[1] 90.70242
> hatbeta0 + qt(.975, df=12)*sqrt(s2)*sqrt(1/n + mean(x)^2/sxx)
[1] 97.56571
> t0 = hatbeta1/(sqrt(s2)/sqrt(sxx))
> t0
[1] -37.39913
> pt(abs(t0), df=12)
[1] 1
> pt(t0, df=12)
[1] 4.289045e-14
```