2.3 Deviations from the sample average Sometimes it is convenient to write the simple linear regression model in a different form that is a little easier to manipulate. Taking equation (2.1), and adding $\beta_1\bar{x} - \beta_1\bar{x}$, which equals zero, to the right-hand side, and combining terms, we can write

$$\begin{aligned} y_i &= \beta_0 + \beta_1\bar{x} + \beta_1 x_i - \beta_1\bar{x} + e_i \\ &= (\beta_0 + \beta_1\bar{x}) + \beta_1(x_i - \bar{x}) + e_i \\ &= \alpha + \beta_1(x_i - \bar{x}) + e_i \end{aligned} \tag{2.29}$$

where we have defined $\alpha = \beta_0 + \beta_1\bar{x}$. This is called the *deviations from the sample average form for simple regression.*

2.3.1. What is the meaning of the parameter α?

Solution: α is the value of $\mathrm{E}(Y|X = \bar{x})$. ∎

2.3.2. Show that the least squares estimates are

$$\hat{\alpha} = \bar{y} \qquad \hat{\beta}_1 \text{ as given by (2.5)}$$

Solution: The residual sum of squares function can be written as

$$\begin{aligned} RSS(\alpha, \beta_1) &= \sum (y_i - \alpha - \beta_1(x_i - \bar{x}))^2 \\ &= \sum (y_i - \alpha)^2 - 2\beta_1 \sum (y_i - \alpha)(x_i - \bar{x}) + \beta_1^2 \sum (x_i - \bar{x})^2 \end{aligned}$$

We can write

$$\begin{aligned} \beta_1 \sum (y_i - \alpha)(x_i - \bar{x}) &= \sum y_i(x_i - \bar{x}) + \alpha \sum (x_i - \bar{x}) \\ &= SXY + 0 \\ &= SXY \end{aligned}$$

Substituting into the last equation,

$$RSS(\alpha, \beta_1) = \sum (y_i - \alpha)^2 - 2\beta_1 SXY + \beta_1^2 SXX$$

Differentiating with respect to α and β_1 immediately gives the desired result. ∎

2.3.3. Find expressions for the variances of the estimates and the covariance between them.

Solution:

$$\mathrm{Var}(\hat{\alpha}) = \frac{\sigma^2}{n}, \mathrm{Var}(\hat{\beta}_1) = \sigma^2/SXX$$

The estimates $\hat{\beta}_1$ and $\hat{\alpha}$ are uncorrelated. ∎

2.4 Heights of Mothers and Daughters

2.4.1. For the heights data in the file `heights.txt`, compute the regression of *Dheight* on *Mheight*, and report the estimates, their standard errors, the value of the coefficient of determination, and the estimate of variance. Give the analysis of variance table the tests the hypothesis that $\mathrm{E}(Dheight|Mheight) = \beta_0$ versus the alternative that $\mathrm{E}(Dheight|Mheight) = \beta_0 + \beta_1 Mheight$. Write a sentence or two that summarizes the results of these computations.

Solution:

```
> mean(heights)
Mheight Dheight
  62.45   63.75    Daughters are a little taller
> var(heights)
        Mheight Dheight
Mheight   5.547   3.005    Daughters are a little more variable
Dheight   3.005   6.760
```

```
> m1 <- lm(Dheight ~ Mheight, data=heights)
> summary(m1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   29.917      1.623    18.4   <2e-16 ***
Mheight        0.542      0.026    20.9   <2e-16 ***
---
Residual standard error: 2.27 on 1373 degrees of freedom
Multiple R-Squared: 0.241,      Adjusted R-squared: 0.24
F-statistic:  435 on 1 and 1373 DF,  p-value: <2e-16
```

The F-statistic has a p-value very close to zero, suggesting strongly that $\beta_1 \neq 0$. The value of $R^2 = 0.241$, so only about one-forth of the variability in daughter's height is explained by mother's height. ∎

2.4.2. Write the mean function in the deviations from the mean form as in Problem 2.3. For this particular problem, give an interpretation for the value of β_1. In particular, discuss the three cases of $\beta_1 = 1$, $\beta_1 < 1$ and $\beta_1 > 1$. Obtain a 99% confidence interval for β_1 from the data.

Solution: If $\beta_1 = 1$, then on average *Dheight* is the same as *Mheight*. If $\beta_1 < 1$, then, while tall mothers tend to have tall daughters, on average they are shorter than themselves; this is the idea behind the word *regression*, in which extreme values from one generation tend to produce values not so extreme in the next generation. $\beta_1 > 1$ would imply that daughters tend to be taller than their mothers, suggesting that, eventually, we will all be giants.

The base R function `vcov` returns the covariance matrix of the estimated coefficients from a fitted model, so the diagonal elements of this matrix gives the squares of the standard errors of the coefficient estimates. The `alr3` library adds this function for S-Plus as well. In addition, the function `confint` in the `alr3` package can be used to get the confidence intervals:

```
> confint(m1, level=0.99)
                 0.5 %     99.5 %
(Intercept) 25.7324151 34.1024585
Mheight      0.4747836  0.6087104
```

■

2.4.3. Obtain a prediction and 99% prediction interval for a daughter whose mother is 64 inches tall.

Solution: Using R,

```
> predict(m1,data.frame(Mheight=64),interval="prediction",level=.99)
       fit   lwr   upr
[1,] 64.59 58.74 70.44
```

2.8 Scale invariance

2.8.1. In the simple regression model (2.1), suppose the value of the predictor X is replaced by cX, where c is some non-zero constant. How are $\hat\beta_0$, $\hat\beta_1$, $\hat\sigma^2$, R^2, and the t-test of NH: $\beta_1 = 0$ affected by this change?

Solution: Write

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 X = \beta_0 + \frac{\beta_1}{c}(cX)$$

which suggests that the slope will change from β_1 to β_1/c, but no other summary statistics will change, and no tests will change. ■

2.8.2. Suppose each value of the response Y is replaced by dY, for some $d \neq 0$. Repeat 2.8.1.

Solution: Write

$$\begin{aligned}
\mathrm{E}(Y|X) &= \beta_0 + \beta_1 X \\
d\mathrm{E}(Y|X) &= d\beta_0 + d\beta_1 X \\
\mathrm{E}(dY|X) &= d\beta_0 + d\beta_1 X
\end{aligned}$$

and so the slope and intercept and their estimates are all multiplied by d. The variance is also multiplied by d. Scale-free quantities like R^2 and test statistics are unchanged. ■

note that the variance RSS/(n-2) is also multiplied by d^2.