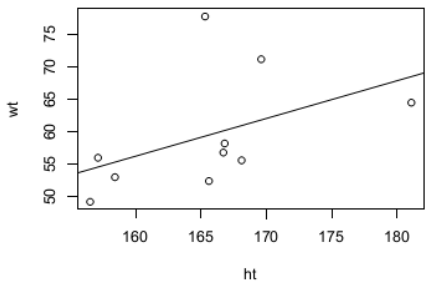
Sample solutions

Stat 8051 Homework 1

Problem 1: ALR Exercise 2.1

2.1.1 The scatterplot of Wt vs. Ht is as below:

```
> plot(Ht,Wt)
```



8 of the 10 data points are plotted more or less in a linear fashion, so linear regression can be a good idea. But the other two points are situated away from the others and they can possibly distort the linear fit that will be obtained.

2.1.2

```
> xbar = mean(Ht); ybar = mean(Wt)
> sxx = var(Ht)*(length(Ht)-1); > syy = var(Wt)*(length(Wt)-1)
> sxy = cov(Ht,Wt)*(length(Ht)-1))
```

and thus we have that $\bar{x} = 165.52, \bar{y} = 59.47, SXX = 472.076, SYY = 731.961, SXY = 274.786$. Hence the estimates for slope and intercept are:

$$\hat{\beta}_1 = \frac{SXY}{SXX} = 0.58, \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -36.88$$

After that we plot the regression line obtained in the above scatterplot:

```
> abline(beta1hat, beta0hat)
```

2.1.3 The estimate of σ^2 is $\hat{\sigma}^2 = \dfrac{RSS}{n-2} = \dfrac{1}{n-2}\left[SYY - \frac{(SXY)^2}{SXX}\right] = 71.502$, and the estimated standard errors of regression coefficients are:

$$S.E.(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SXX}} = 0.39, S.E.(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SXX}} = 64.49$$

The estimated covariance of regression coefficients

$$\hat{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\hat{\sigma}^2 \frac{\bar{x}}{SXX} = -25.07$$

The test statistics for t-tests for testing $\beta_0 = 0$ and $\beta_1 = 0$ are $T_0 = \hat{\beta}_0/SE(\hat{\beta}_0)$ and $T_1 = \hat{\beta}_1/SE(\hat{\beta}_1)$, respectively, both of which follow t-distributions with df $n-2$. For our given sample, these values come out to be

$$T_0 = \frac{-36.88}{64.49} = -0.57, T_1 = \frac{0.58}{0.39} = 1.49$$

thus, for a t_{n-2} distribution, the two-sided p-values come out to be $2(1 - F(|T_0|)) = 0.58$ and $2(1 - F(|T_1|)) = 0.17$, repectively, with $F(.)$ being the CDF for t_{n-2} distribution.

Problem 2: ALR Exercise 2.10

2.10.1

$$\bar{y} = \frac{1}{m_0 + m_1}\left[\sum_{i=1}^{m_0} y_i + \sum_{j=1}^{m_1} y_j\right] = \frac{m\bar{y}_0 + m\bar{y}_1}{2m} = \frac{\bar{y}_0 + \bar{y}_1}{2}$$

$$\bar{x} = \frac{1}{m_0 + m_1}\left[\sum_{i=1}^{m_0} x_i + \sum_{j=1}^{m_1} x_j\right] = \frac{0 + m}{2m} = \frac{1}{2}$$

$$SXX = \sum_{i=1}^{m_0}(x_i - \bar{x})^2 + \sum_{j=1}^{m_1}(x_j - \bar{x})^2 = m\left(0 - \frac{1}{2}\right)^2 + m\left(1 - \frac{1}{2}\right)^2 = \frac{2m}{4} = \frac{m}{2}$$

$$\begin{aligned} SXY &= \sum_{i=1}^{m_0}(x_i - \bar{x})(y_i - \bar{y}) + \sum_{j=1}^{m_1}(x_j - \bar{x})(y_j - \bar{y}) \\ &= -m\left(0 - \frac{1}{2}\right)\left[\bar{y}_0 - \frac{\bar{y}_0 + \bar{y}_1}{2}\right] + m\left(1 - \frac{1}{2}\right)\left[\bar{y}_1 - \frac{\bar{y}_0 + \bar{y}_1}{2}\right] \\ &= \frac{m(\bar{y}_1 - \bar{y}_0)}{2} \end{aligned}$$

2.10.2

$$\hat{\beta}_1 = \frac{SXY}{SXX} = \bar{y}_1 - \bar{y}_0, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \frac{\bar{y}_1 + \bar{y}_0}{2} - \frac{\bar{y}_1 - \bar{y}_0}{2} = \bar{y}_0$$

2.10.3 For the groups 0 and 1, fitted values are $\hat{\beta}_0 + 0.\hat{\beta}_1 = \bar{y}_0$ and $\hat{\beta}_0 + 1.\hat{\beta}_1 = \bar{y}_1$ respectively, which gives residuals $y_i - \bar{y}_0$ and $y_i - \bar{y}_1$, respectively. Hence we have

$$\hat{\sigma}^2 = \frac{\sum_{i:x_i=0}(y_i - \bar{y}_0)^2 + \sum_{i:x_i=1}(y_i - \bar{y}_1)^2}{2m - 1}$$

2.10.4 We have $Var(\hat{\beta}_1|X) = \hat{\sigma}^2/SXX = 2\hat{\sigma}^2/m$. Thus the t-statistic:

$$t = \frac{\hat{\beta}_1 - 0}{\sqrt{(Var(\hat{\beta}_1|X)}} = \frac{\bar{y}_1 - \bar{y}_0}{\hat{\sigma}\sqrt{\frac{2}{m}}}$$

which is same as the two-sample t-statistic with $m_0 = m_1 = m$.

2.10.5 For $x_i^* = ax_i + b$ with x_i being either 0 or 1, we have

$$E(Y|X^*) = \beta_0^* + \beta^* x^* = \beta_0^* + \beta^*(ax + b) = (\beta_0^* + \beta_1^* b) + a\beta_1^* x$$

Thus $\beta_0 = \beta_0^* + \beta_1^* b, \beta_1 = a\beta_1^*$, which gives

$$\beta_1^* = \beta_1/a, \quad \beta_0^* = \beta_0 - a\beta_1$$

Here $a = 2, b = -1$ (check). Thus the new coefficients are

$$\beta_1^* = \frac{\bar{y}_1 - \bar{y}_0}{2}, \quad \beta_0^* = \bar{y}_0 + \frac{\bar{y}_1 - \bar{y}_0}{2} = \frac{\bar{y}_1 + \bar{y}_0}{2}$$

Since the response has not changed, the estimate of σ^2 and the value of R^2 will be unchanged. Hence test of the slope equal to 0 will be unchanged.

2.10.6 Get the summary after applying `lm` and compare with coefficients calculated by hand.

Problem 3: ALR Exercise 2.13

2.13.1

```
> colMeans(Heights)
 mheight  dheight
62.45280 63.75105
> var(Heights)
         mheight  dheight
mheight 5.546511 3.004806
dheight 3.004806 6.760274
> m1 <- lm(dheight ~ mheight, data=Heights)
> summary(m1)
```

```
Call:
lm(formula = dheight ~ mheight, data = Heights)

Residuals:
   Min     1Q Median     3Q    Max
-7.397 -1.529  0.036  1.492  9.053

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.91744    1.62247   18.44   <2e-16 ***
mheight      0.54175    0.02596   20.87   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.266 on 1373 degrees of freedom
Multiple R-squared:  0.2408,Adjusted R-squared:  0.2402
F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
```

Residual variance $\hat{\sigma}^2 = 2.266^2 = 5.135$. The t-statistic for the slope has a p-value very close to 0, suggesting strongly that $\beta_1 = 0$. The value of $R^2 = 0.241$, so only about one-fourth of the variability in daughter's height is explained by mother's height.

2.13.2 Although the confidence intervals can be computed from the formulas in the text, most programs will produce them automatically. In R the function `confint` does this:

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

The second row gives 99% CI for $\hat{\beta}_1$.

2.13.3

```
> predict(m1, newdata=data.frame(mheight=64),
+         interval="prediction",
+         level=.99)
       fit      lwr      upr
1 64.58925 58.74045 70.43805
```

Problem 4: ALR Exercise 2.17

2.17.1 The least square estimator for β_1 is obtained by finding the value of $\hat{\beta}_1$ such that $RSS(\beta_1)$ is minimized. Taking the derivative of the given expression for $RSS(\hat{\beta}_1)$ with

respect to $\hat{\beta}_1$ and setting the resulting expression equal to zero we find

$$\frac{d}{d\hat{\beta}_1} RSS(\hat{\beta}_1) = 2\sum(y_i - \hat{\beta}_1 x_i)(-x_i) = 0 \quad \Rightarrow \quad -\sum y_i x_i + \hat{\beta}_1 \sum x_i^2 = 0$$

Solving this expression for $\hat{\beta}_1$ we find: $\hat{\beta}_1 = (\sum x_i y_i)/(\sum x_i^2)$.

To study the bias introduced by this estimator of β_1 we compute

$$E(\hat{\beta}_1) = \frac{\sum x_i E(y_i)}{\sum x_i^2} = \beta_1 \frac{\sum x_i^2}{\sum x_i^2} = \beta_1$$

showing that this estimator is unbiased. To study the variance of this estimator we compute

$$Var(\hat{\beta}_1) = \frac{\sum x_i^2}{(\sum x_i^2)^2} Var(y_i) = \frac{\sigma^2}{\sum x_i^2}$$

the requested expression. An estimate of $\hat{\sigma}$ is given by the usual $\hat{\sigma}^2 = RSS/(n-1)$ which has $n-1$ degrees of freedom.

2.17.2 Models are fit in R without the intercept by adding -1 to the formula.

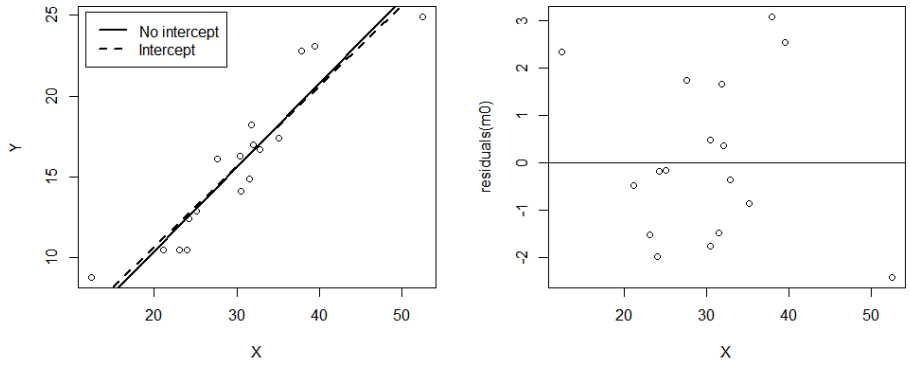
```
> m0 <- lm(Y ~ X - 1, data=snake)
> summary(m0)

Call:
lm(formula = Y ~ X - 1, data = snake)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4207 -1.4924 -0.1935  1.6515  3.0771

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
X  0.52039    0.01318   39.48   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.7 on 16 degrees of freedom
Multiple R-squared:  0.9898,Adjusted R-squared:  0.9892
F-statistic:  1559 on 1 and 16 DF,  p-value: < 2.2e-16

> confint(m0)
     2.5 %   97.5 %
X 0.492451 0.548337
> tval <- (coef(m0)[1] - 0.49)/ sqrt(vcov(m0)[1,1])
> df <- dim(snake)[1] - 1
```

```
> data.frame(tval = tval, df=df, pval = 1 - pt(abs(tval), df))
      tval df       pval
X 2.305853 16 0.01742104
```

Most programs won't automatically provide a test that the slope has any value other than 0, so we need to do the "hand" calculation. The `pt` function computes the area to the left of its argument, which would correspond to the lower tail. We subtract from 1 to get the upper tail.

2.17.3

```
> par(mfrow=c(1,2))
> plot(Y ~ X, snake)
> m1 <- lm(Y ~ X, snake)
> abline(m0, lwd=2)
> abline(m1, lty=2, lwd=2)
> legend("topleft", c("No intercept", "Intercept"),
+        lty=1:2, inset=0.02, lwd=2)
> plot(residuals(m0) ~ X, snake)
> abline(h=0)
> par(mfrow=c(1,1))
```



The plot at the left shows both the fit of the through-the-origin model (solid line) and the simple regression model (dashed line), suggesting little difference between them. The residual plot emphasizes the 2 points with the largest and smallest value of as somewhat separated from the other points, and fit somewhat less well. However, the through-the-origin model seems to be OK here.