Calculus: Homework #2

Due on February 12, 2014 at 3:10pm

Professor Isaac Newton Section A

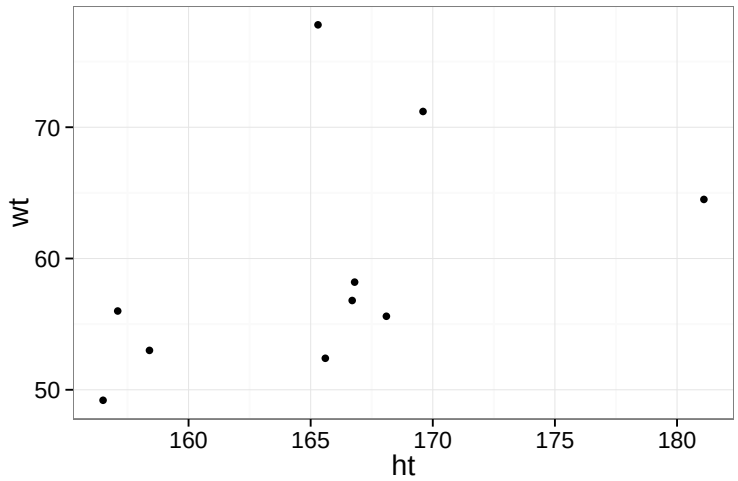
Josh Davis

Problem 2.1. Height and weight data (Data file: Htwt) The table below and the data file give ht = height in centimeters and wt = weight in kilograms for a sample of n = 10 18-year-old girls. The data are taken from a larger study described in Problem 3.3. Interest is in predicting weight from height.

Table 1: Problem 1

ht	wt
169.6	71.2
166.8	58.2
157.1	56
181.1	64.5
158.4	53
165.6	52.4
166.7	56.8
156.5	49.2
168.1	55.6
165.3	77.8

Problem 2.1.1. Draw a scatterplot of wt on the vertical axis versus ht on the horizontal axis. On the basis of this plot, does a simple linear regression model make sense for these data? Why or why not?



Yes, except for some extreme outliers, the linear ascending tendency of wt over ht is obvious.

Problem 2.1.2. Show that $\bar{x} = 165.52$, $\bar{y} = 59.47$, $S_{XX} = 472.08$, $S_{YY} = 731.96$, and $S_{XY} = 274.79$. Compute estimates of the slope and the intercept for the regression of Y on X. Draw the fitted line on your scatterplot.

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{n} = 165.52$$

$$\bar{y} = \frac{\sum_{i=1}^{10} y_i}{n} = 59.47$$

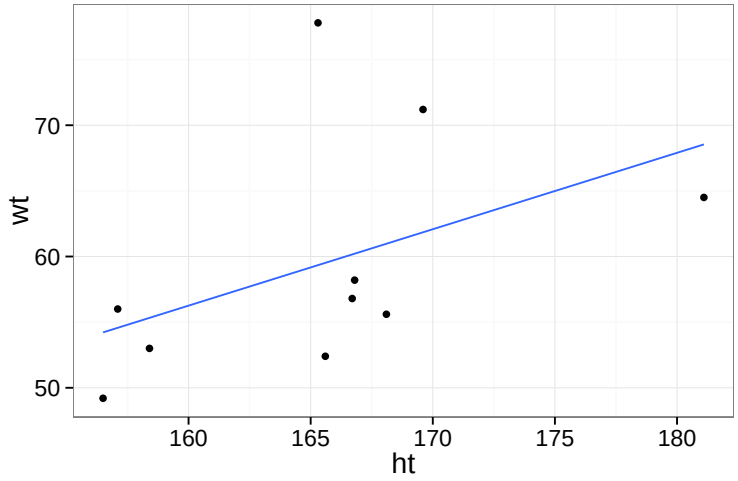
$$S_{XX} = \sum_{i=1}^{10} (x_i - \bar{x})^2 = 472.08$$

$$S_{YY} = \sum_{i=1}^{10} (y_i - \bar{y})^2 = 731.96$$

$$S_{XY} = \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = 274.786 \simeq 274.79$$

$$\hat{\beta_1} = \frac{S_{XY}}{S_{XX}} = \frac{274.79}{472.08} = 0.5821$$

$$\hat{\beta_0} = \bar{y} - \bar{x} \cdot \hat{\beta_1} = 59.47 - 165.52 \cdot 0.5821 = -36.88$$



Problem 2.1.3. Obtain the estimate of σ^2 and find the estimated standard errors of β_0 and β_1. Also find the estimated covariance between β_0 and β_1. Compute the t-tests for the hypotheses that $\beta_0 = 0$ and that $\beta_1 = 0$ and find the appropriate p-values using two-sided tests.

$$\hat{\sigma^2} = \frac{\sum_i e_i^2}{n-2} = 71.5017$$

$$\hat{Se}(\hat{\beta_1}) = \sqrt{\hat{\sigma^2} \cdot \frac{1}{S_{XX}}} = 0.3892$$

$$\hat{Se}(\hat{\beta_0}) = \sqrt{\hat{\sigma^2} \cdot (\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}})} = 64.4728$$

$$Cov(\hat{\beta_1}, \hat{\beta_0}) = -\hat{\sigma^2} \frac{\bar{x}}{S_{XX}} = -25.07$$

$$t_{\beta_0} = \frac{\hat{\beta_0} - 0}{\hat{Se}(\hat{\beta_0})} = \frac{-36.8759}{64.4728} = -0.572$$

$$t_{\beta_1} = \frac{\hat{\beta_1} - 0}{\hat{Se}(\hat{\beta_1})} = \frac{0.5821}{0.3892} = 1.496$$

$$p_{\beta_0} = P(x > |t_{\beta_0}|; x \sim t(8)) = 0.583$$
$$p_{\beta_1} = P(x > |t_{\beta_1}|; x \sim t(8)) = 0.173$$

Problem 2.10. Two-sample tests One of the basic problems in elementary statistics is testing for equality of two means. If $\bar{y}_j, j = 0, 1$, are the sample means, the sample sizes are $m_j, j = 0, 1$, and the sample standard deviations are $SD_j, j = 0, 1$, then under the assumption that sample j is $NID(\mu_j, \sigma^2)$, the statistic

$$t = \frac{\bar{y}_1 - \bar{y}_0}{\hat{\sigma}\sqrt{1/m_0 + 1/m_1}}$$

with $\hat{\sigma}^2 = [(m_0 - 1)SD_0^2 + (m_1 - 1)SD_1^2]/[m_0 + m_1 - 2]$ is used to test $\mu_0 = \mu_1$ against a general alternative. Under normality and the assumptions of equal variance in each population, the null distribution is $t \sim t(m_0 + m_1 - 2)$.

For simplicity assume $m_0 = m_1 = m$, although the results do not depend on the equal sample sizes. Define a predictor X with values $x_i = 0$ for $i = 1, \ldots, m$ and $x_i = 1$ for $i = m + 1, \ldots, 2m$. Combine the response y_i into a vector of length $2m$, the first m observations corresponding to population 0 and the remaining to population 1. In this problem we will fit the simple linear regression model for this X and Y, and show that it is equivalent to the two-sample problem.

Problem 2.10.1. Show that $\bar{y} = (\bar{y}_0 + \bar{y}_1)/2$, $\bar{x} = 1/2$, $S_{XX} = m/2$, and $S_{XY} = m(\bar{y}_1 - \bar{y}_0)/2$.

$$\bar{y} = \frac{\sum_{i=1}^{m} y_i + \sum_{i=m+1}^{2m} y_i}{2m} = \frac{\bar{y}_0 m + \bar{y}_1 m}{2m} = \frac{\bar{y}_0 + \bar{y}_1}{2}$$
$$\bar{x} = \frac{\sum_{i=1}^{m} x_i + \sum_{i=m+1}^{2m} x_i}{2m} = \frac{m}{2m} = \frac{1}{2}$$
$$S_{XX} = \sum_{i=1}^{m} (x_i - \frac{1}{2})^2 + \sum_{i=m+1}^{2m} (x_i - \frac{1}{2})^2 = \frac{1}{4} \cdot 2m = \frac{m}{2}$$
$$S_{XY} = \sum_{i=1}^{m} (0 - \frac{1}{2})(y_i - \frac{\bar{y}_0 + \bar{y}_1}{2}) + \sum_{i=m+1}^{2m} (1 - \frac{1}{2})(y_i - \frac{\bar{y}_0 + \bar{y}_1}{2}) = \frac{m}{2}(\bar{y}_1 - \bar{y}_0)$$

Problem 2.10.2. Give the formulas for the OLS estimates of β_0 and β_1 in the simple linear regression model with the Y and X as specified in this problem. Interpret the estimates.

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\frac{m}{2}(\bar{y}_1 - \bar{y}_0)}{\frac{m}{2}} = \bar{y}_1 - \bar{y}_0$$
$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 = \frac{\bar{y}_0 + \bar{y}_1}{2} - \frac{1}{2}(\bar{y}_1 - \bar{y}_0) = \bar{y}_0$$

$E(Y_i|X_i = 1) = \bar{y}_1$, $E(Y_i|X_i = 0) = \bar{y}_0$. Given $x = 1$, the estimated value of y is equal to the mean value of population 1; given $x = 0$, the estimated value of y is equal to the mean value of population 0;

Problem 2.10.3. Find the fitted values and the residuals. Give an expression for RSS obtained by squaring and adding up the residuals and then dividing by the df.

$$\hat{y}_i = \bar{y}_0 + (\bar{y}_1 - \bar{y}_0)x_i = \begin{cases} \bar{y}_0 & x_i = 0 \\ \bar{y}_1 & x_i = 1 \end{cases}$$

$$RSS = \sum_{i=1}^{2m} (y_i - \hat{y}_i) = \sum_{i=1}^{m} (y_i - \bar{y}_0)^2 + \sum_{i=m+1}^{2m} (y_i - \bar{y}_1)^2 = (m-1)SD_0^2 + (m-1)SD_1^2$$

$$\hat{\sigma^2} = \frac{RSS}{df} = \frac{(m-1)SD_0^2 + (m-1)SD_1^2}{2m-2} = \frac{SD_0^2 + SD_1^2}{2}$$

Problem 2.10.4. Show that the t-statistic for testing $\beta_1 = 0$ is exactly the same as the usual two-sample t-test for comparing two groups with an assumption of equal within-group variance.

$$\hat{Se}(\hat{\beta}_1) = \sqrt{\frac{1}{S_{XX}}\hat{\sigma^2}} = \sqrt{\frac{1}{m}(SD_0^2 + SD_1^2)}$$

$$t_{\beta_1} = \frac{\hat{\beta}_1}{\hat{Se}(\hat{\beta}_1)} = \frac{\bar{y}_1 - \bar{y}_0}{\sqrt{\frac{1}{m}(SD_0^2 + SD_1^2)}} \sim t(2m-2) = \frac{\bar{y}_1 - \bar{y}_0}{\hat{\sigma}\sqrt{\frac{2}{m}}}$$

which is the same as the usual two-sample t-test for comparing two groups with an assumption of equal within-group variance.

Problem 2.10.5. The group indicator is set to $x_i = 0$ for one group and $x_i = 1$ for the other group. Suppose we used as the group indicator $x_i^* = -1$ for the first group and $x_i^* = +1$ for the second group. How will this change the estimates of β_0 and β_1 and the meaning of the test that $\beta_1 = 0$? (Hint: Find values a and b such that $x_i^* = a(x_i + b)$, and then apply Problem 2.9.1.)

$$x_i^* = 2(x_i - 0.5) \rightarrow x_i = 0.5 \cdot x_i^* + 0.5$$

$$\beta_1^* = \beta_1 \cdot 0.5$$

$$\beta_0^* = \beta_0 + 0.5 \cdot \beta_1$$

$$\beta_1^* = 0 \iff \beta_1 = 0$$

Since the response Y has not changed, the estimate of σ^2 will be unchanged. The test of the β_1^* equal to 0 will be the same.

Problem 2.10.6. (Data file: cathedral) The datafile contains the Height and Length in feet of 25 cathedrals, nine in the Romanesque style, and 16 in the later Gothic style. Consider only the first 18 rows of this data file, which contain all the Romanesque cathedrals and nine of the Gothic cathedrals, and consider testing the hypothesis that the mean length is the same for Romanesque and Gothic cathedrals against the alternative that they are different. Use these data to verify all the results of the preceding sections of this problem. (Hint: In the

data file the group indicator Type is a text variable with values Romanesque and Gothic that you may need to convert to zeros and ones. In R, for example, the statement

> cathedral$group <- ifelse(cathedral$Type==Romanesque, 0, 1)

will do it. Dont forget to remove the last seven rows of the file, although if you do forget the test computed will still be a t-test of the hypothesis that the two types of cathedrals have the same mean height, but based on different data.)

Comparison between two-sample t-test (model_1) and linear regression(model_2)

```
data <- cathedral
data$group <- ifelse (data$Type == "Romanesque", 0, 1)
data <- data[1:(nrow(data)-7), ]
model_1 <- t.test (Length ~ Type, data = data, var.equal = TRUE)
model_2 <- lm (Length ~ group, data = data)
model_1
```

```
##
##  Two Sample t-test
##
## data:  Length by Type
## t = -1.634, df = 16, p-value = 0.1218
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -153.43   19.87
## sample estimates:
##     mean in group Gothic mean in group Romanesque
##                    408.7                    475.4
```

```
summary (model_2)
```

```
##
## Call:
## lm(formula = Length ~ group, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -183.7  -47.5   12.8   67.3  110.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    475.4       28.9   16.45  1.9e-11 ***
## group          -66.8       40.9   -1.63     0.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 86.7 on 16 degrees of freedom
## Multiple R-squared:  0.143,Adjusted R-squared:  0.0894
## F-statistic: 2.67 on 1 and 16 DF,  p-value: 0.122
```

Note that the t statistic and the P-value of this two models are the same.

Comparison between linear regression (model_2) and regression with scaled X (model_3)

```
data <- cathedral
data$group <- ifelse (data$Type == "Romanesque", 0, 1)
data <- data[1:(nrow(data)-7), ]
data$group_1 <- (data$group - 0.5) * 2
model_3 <- lm (Length ~ group_1, data = data)
summary(model_3)
```

```
##
## Call:
## lm(formula = Length ~ group_1, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -183.7  -47.5   12.8   67.3  110.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    442.1       20.4   21.63  2.9e-13 ***
## group_1        -33.4       20.4   -1.63     0.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 86.7 on 16 degrees of freedom
## Multiple R-squared:  0.143,Adjusted R-squared:  0.0894
## F-statistic: 2.67 on 1 and 16 DF,  p-value: 0.122
```

Note that the relationship between two β_1, β_0 is the same as what we prove in 2.10.5, and the t statistic and the t-test remain the same.

Problem 2.13. Heights of mothers and daughters (Data file: Heights)

Problem 2.13.1. 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. Write a sentence or two that summarizes the results of these computations.

```
data <- Heights
model <- lm(dheight ~ mheight, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = dheight ~ mheight, data = data)
##
## 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)   29.917      1.623    18.4   <2e-16 ***
## mheight        0.542      0.026    20.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 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 OLS estimate of β_1 is 0.54175, the OLS estimate of β_0 is 29.91744, the standard error of the estimate of β_1 is 0.02596, the standard error of the estimate of β_0 is 1.62247. The coefficient of determination of this OLS model is $R^2 = 0.2408$, the estimate of variance of the OLS model is $\hat{\sigma}^2 = 2.266^2 = 5.136$

The p-value of t-test is almost 0, suggesting that $\beta_1 \neq 0$. $R^2 = 0.241$, so only about one-fourth of the variability in daughters height is explained by mothers height.

A 1 inch increase in mother's height will result in a 0.54 increase in daughter's height on average. (given all other predictors have the same value)

Problem 2.13.2. Obtain a 99% confidence interval for β_1 from the data.

```
confint(model, level=0.99)
```

```
##                0.5 %   99.5 %
## (Intercept) 25.7324 34.1025
## mheight      0.4748  0.6087
```

$\beta_1 \in [0.475,\ 0.609]$

Problem 2.13.3. Obtain a prediction and 99% prediction interval for a daughter whose mother is 64 inches tall.

```
predict(model, data.frame(mheight=64), interval="prediction", level=.99)
```

```
##     fit   lwr   upr
## 1 64.59 58.74 70.44
```

$y \in [58.740,\ 70.438]$

Problem 2.17. Regression through the origin Occasionally, a mean function in which the intercept is known a priori to be 0 may be fit. This mean function is given by

$$E(y|x) = \beta_1 x$$

The residual sum of squares for this model, assuming the errors are independent with common variance σ^2 , is $RSS = \sum (y_i - \beta x_i)^2$

Problem 2.17.1. Show that the least squares estimate of β_1 is $\hat{\beta}_1 = \sum x_i y_i / \sum x_i^2$. Show that $\hat{\beta}_1$ is unbiased and that $Var(\hat{\beta}_1|X) = \sigma^2 / \sum x_i^2$. Find an expression for σ^2. How many df does it have?

$$\hat{\beta}_1 = \min_{\beta} \sum (y_i - x_i\beta)^2 = \min_{\beta} f(\beta)$$

$$\frac{\partial f}{\partial \beta} = 0 \Rightarrow \sum (y_i - x_i\beta)x_i \Rightarrow \hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

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

$$Var(\hat{\beta}_1) = Var(\sum \frac{x_i}{\sum x_i^2} \cdot y_i) = \sum \frac{x_i^2}{(\sum x_i^2)^2} Var(y_i) = \frac{1}{\sum x_i^2}\sigma^2$$

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

$$df = n - 1$$

Problem 2.17.2. (Data file: snake) The data file gives X = water content of snow on April 1 and Y = water yield from April to July in inches in the Snake River watershed in Wyoming for $n = 17$ years from 1919 to 1935 (Wilm, 1950). Fit a regression through the origin and find $\hat{\beta}_1$ and $\hat{\sigma^2}$. Obtain a 95% confidence interval for β_1 . Test the hypothesis that the slope $\beta_1 = 0.49$, against the alternative that $\beta_1 > 0.49$.

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

```
##
## Call:
## lm(formula = Y ~ X - 1, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.421 -1.492 -0.194  1.651  3.077
```

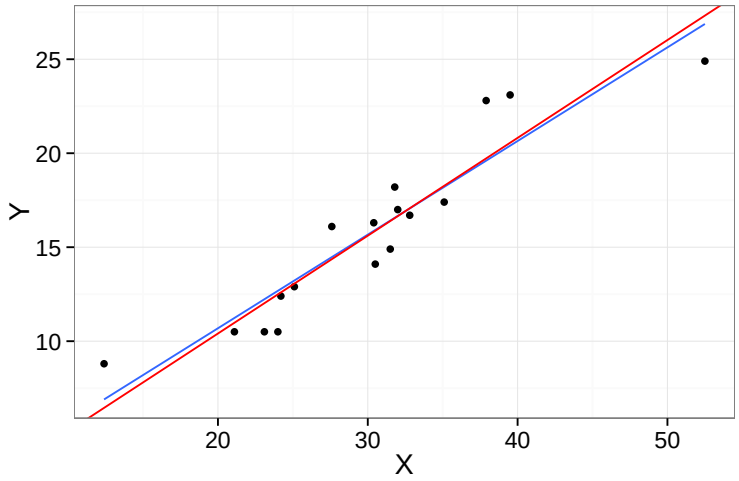
```
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## X   0.5204     0.0132    39.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.7 on 16 degrees of freedom
## Multiple R-squared:  0.99,Adjusted R-squared:  0.989
## F-statistic: 1.56e+03 on 1 and 16 DF,  p-value: <2e-16
```

```
confint(model, level = 0.95)
```

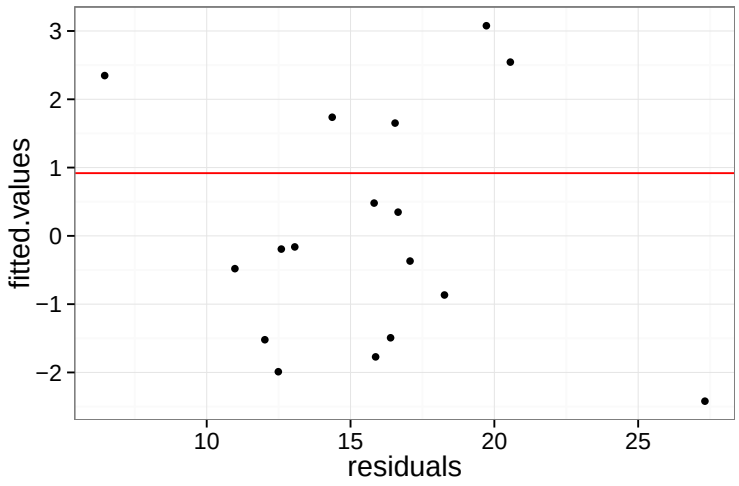
```
##    2.5 % 97.5 %
## X 0.4925 0.5483
```

$\hat{\beta}_1 = 0.52039$, $\hat{\sigma^2} = 1.7^2 = 2.89$, $\beta_1 \in [0.4925,\ 0.5483]$
Let $Z \sim t(16)$, then $P(Z > t) = P(Z > \frac{\hat{\beta}_1 - 0.49}{\hat{Se}(\hat{\beta}_1)}) = 0.017$

Problem 2.17.3. Plot the residuals versus the fitted values, and comment on the adequacy of the mean function with 0 intercept. In regression through the origin, $\sum \hat{e}_i \neq 0$.



the Red line in this graph is the regression model fitted where intercept is known a priori to be 0, Blue line is the model fitted via standard OLS, Note that there exists little difference between this two line, which indicates that regression through origin point is OK in this problem.



Red line in this graph illustrates the level of sum value of all residuals. Note that the Red line is far from $y = 0$, suggesting that $\sum \hat{e}_i \neq 0$.