Chapter 2

Simple Linear Regression

2.1 Linear Correlation Coefficient

The *z-score* or *standardized value* of an observation x is the number of standard deviations it lies from the mean. So when standardizing with respect to a population,

$$z = \frac{x - \mu}{\sigma}$$

and when standardizing with respect to a sample,

$$z = \frac{x - \overline{x}}{s}$$

Facts about z-scores:

1. Their mean is always 0.

2. Their standard deviation is always 1.

3. If the data distribution is mound-shaped, the empirical rule (or the normal probability distribution) says about 68.27% fall between -1 and 1, 95.45% between -2 and 2, 99.73% between -3 and 3.

The *linear correlation coefficient* r measures the magnitude and direction of the linear association between two variables and is defined

$$\begin{aligned} r &:= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right) \\ &= \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i} \end{aligned}$$

Given a point pattern of data, r measures the average product of the z-scores. Data points above (below) average in the x-direction and above (below) average in the y-direction will have positive z-score products, whereas points above (below) average in the x-direction but below (above) average in the y-direction will have negative z-score products. The average of a bunch of mostly positive (negative) values will tend to be positive (negative). Of course it just takes one strange observation to screw this up.

2.2 Simple Linear Regression

Simple linear regression attempts to summarize the relationship between two quantitative continuous variables in a straight line.

Example 11 To estimate a fair used car price, you collect the following data:

Age (years)	3	6	8	4	2	11	4	7	7	9
Price ($100)	172	140	112	160	165	80	155	103	84	78

Here, *Age* is the *independent* or *explanatory* or *predictor* variable. *Price* is the *dependent* or *response* or *outcome variable*. Obviously there exist variables other than age that determine the price of a used car (e.g. mileage, make, model, etc.). For now we will focus on the case with one predictor variable.

Enter these data into a data frame in R:

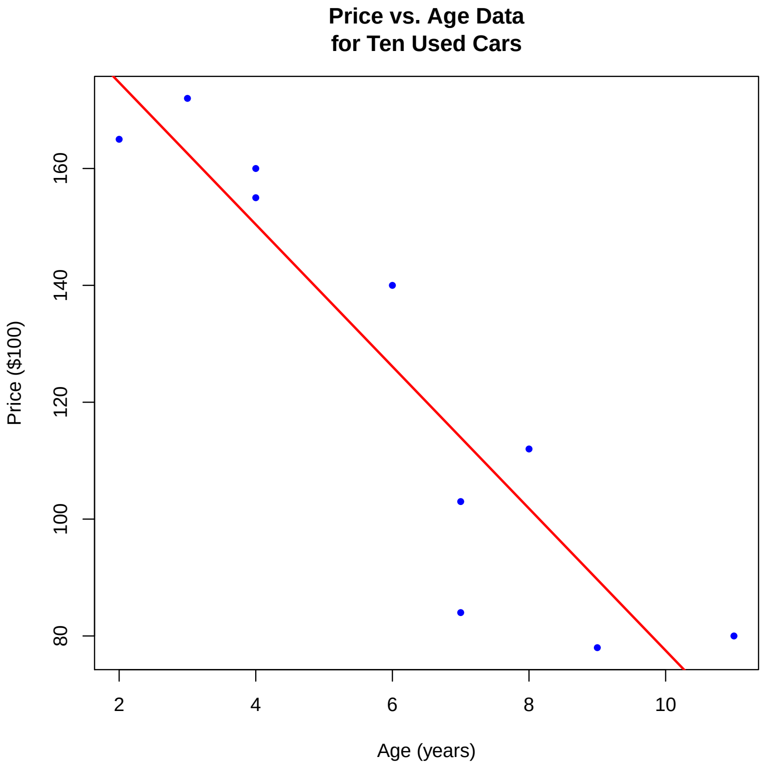
```
> age <- c(3, 6, 8, 4, 2, 11, 4, 7, 7, 9)
> price <- c(172, 140, 112, 160, 165, 80, 155, 103, 84, 78)
> used_cars <- data.frame(age, price)
> used_cars
      age   price
  1    3     172
  2    6     140
  3    8     112
  4    4     160
  5    2     165
  6   11      80
  7    4     155
  8    7     103
  9    7      84
 10    9      78
```

We suspect cars with higher age tend to be priced lower. We can "confirm" this by making a scatterplot of the data (see code below and Figure 2.1), along with a best-fit regression line (in red on the plot) of the form

$$\hat{y} = b_0 + b_1 x.$$

```
> regression <- lm(used_cars$price ~ used_cars$age)
> plot(used_cars$age, used_cars$price, main = "Price vs. Age Data
+ for Ten Used Cars", xlab = "Age (years)",
+ ylab = "Price ($100)", pch = 20, col="blue")
> regression


Call:
lm(formula = used_cars$price ~ used_cars$age)
Coefficients:
```

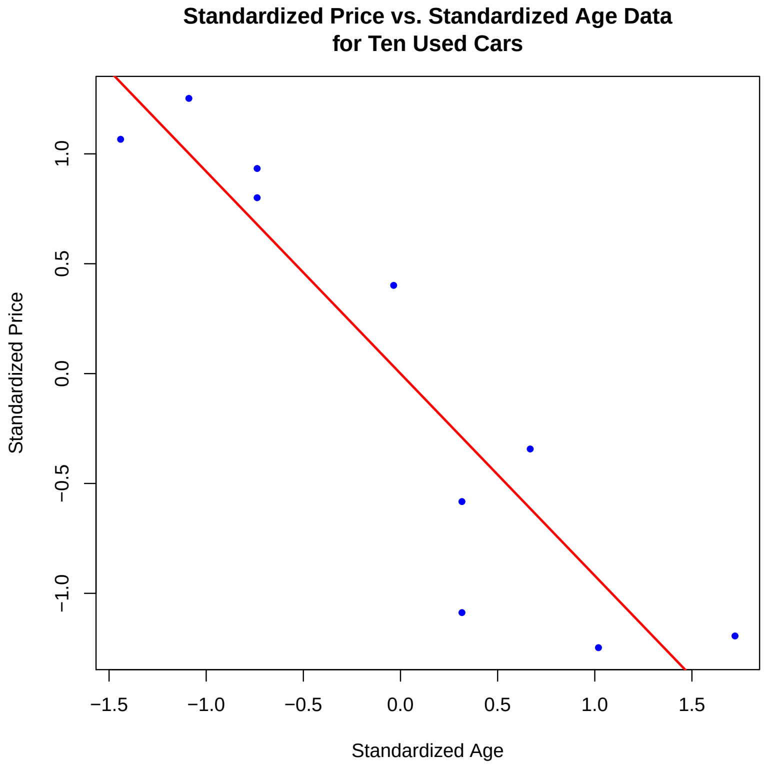


```
 (Intercept)  used_cars$age
      199.03         -12.15
```

```
> abline(regression, lwd=2, col="red")
```

What do you think would happen if we made a scatterplot and regression line for the standardized data?

```
> attach(used_cars)
> stdage <- (age - mean(age))/sd(age)
> stdprice <- (price - mean(price))/sd(price)
> stduc <- data.frame(stdage, stdprice)
> stduc
> windows()     # tells R to open multiple graph windows so it won't
> attach(stduc) # replace previous graphs with new ones
> regression_std <- lm(stdprice ~ stdage)
> plot(stdage, stdprice, main = "Standardized Price vs. Standardized Age Data
+ for Ten Used Cars", xlab = "Standardized Age",
```



```
+ ylab = "Standardized Price", pch = 20, col="blue")
> regression_std


Call:
lm(formula = stdprice ~ stdage)

Coefficients:
(Intercept)       stdage
 -1.517e-16   -9.197e-01


> abline(regression_std, lwd=2, col="red")
```

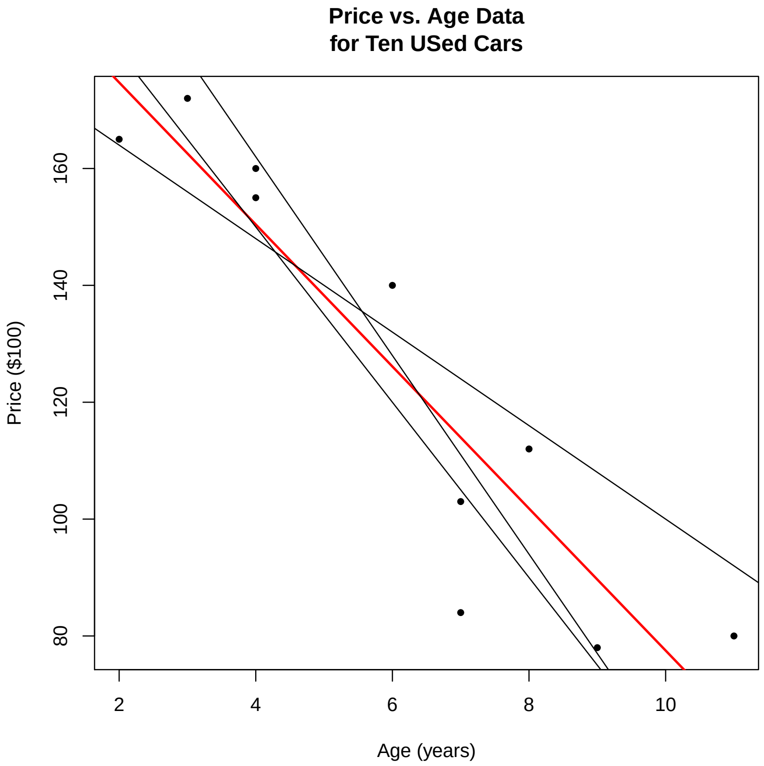


Figure 2.1: There are many different straight lines one could use to model the point pattern. The red line is best-fitting line, or least squares regression line, but how do we find it?

2.3 Least-Squares Regression Coefficients

There are many different straight lines one could use to model the point pattern. The red line is best-fitting line, or least squares regression line, but how do we find it? See Figure 2.1.

The coefficients in the least-squares regression take on values that minimize the cost function in (**??**), or equivalently, the sum of squared errors (SSE):

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_1 x_i - b_0)^2 \tag{2.1}$$

We want to minimize this quantity as a function of b_0 and b_1, so taking the first derivative with respect to b_0 and setting it equal to 0 gives

$$\frac{d}{db_0}\sum_{i=1}^{n} e_i^2 = -2\sum_{i=1}^{n}(y_i - b_1 x_i - b_0) = 0$$

$$\begin{aligned}
\Rightarrow \qquad & nb_0 + b_1\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0 \\
\Rightarrow \qquad & b_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - b_1\frac{1}{n}\sum_{i=1}^{n} x_i
\end{aligned}$$

or

$$\boxed{b_0 = \overline{y} - b_1\overline{x}}. \tag{2.2}$$

The first derivative of the sum of squared errors (2.1) with respect to b_1 is

$$\frac{d}{db_1}\sum_{i} e_i^2 = -2\sum_{i}(y_i - b_1 x_i - b_0)x_i.$$

Setting this derivative to 0 and subbing (2.2) for b_0 gives (you should verify this... just takes some algebra)

$$\boxed{b_1 = r\frac{s_y}{s_x}}. \tag{2.3}$$

The second derivatives of (2.1) with respect to b_0 and b_1 are indeed positive, so the squared error function is convex in b_0 and b_1 (you should check this).

Remark 12 Notice (2.2) nicely implies the regression line passes directly through the center of mass of the point pattern. Also (2.3) tells us the slope of the regression line equals the correlation coefficient, rescaled by the ratio of the standard deviations for the y- and x- values. In addition, r is the slope of the regression line through the standardized data points: the standardized data points will have the same point pattern as the original data, but centered at the origin with slope r.

Note also that one could multiply through (2.2) by n, and also multiply through (2.2) by x_i and sum over all i to obtain the so-called *normal equations*:

$$\begin{aligned} nb_0 + b_1 \sum_i x_i &= \sum_i y_i \\ b_0 \sum_i x_i + b_1 \sum_i x_i^2 &= \sum_i x_i y_i \end{aligned}$$

which are easily solved to give an alternative (but equivalent and useful) expression for the slope coefficient:

$$\boxed{b_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}}. \tag{2.4}$$

Example 13 Open the data set called CH01PR19 from ALSM. An admissions director wanted to know if an ACT score is a good predictor of first-semester college GPA.

```
> vars = c("GPA", "ACT")
> data <- read.table("CH01PR19.txt", header=FALSE, col.names = vars)
> head(data)
```

```
    GPA ACT
1 3.897  21
2 3.885  14
3 3.778  28
4 2.540  22
5 3.028  21
6 3.865  31
```

We can use (2.2), (2.3), and R to quickly obtain the regression coefficients:

```
> slope =  cor(data$GPA, data$ACT)*sd(data$GPA)/sd(data$ACT)
> intercept = mean(data$GPA) - slope*mean(data$ACT)
> slope
```

```
[1] 0.03882713
```

```
> intercept
```

```
[1] 2.114049
```

so the equation of the regression line is

$$\widehat{GPA} = 0.03882713 \times ACT + 2.114049.$$

That is, the data indicate a trend where GPA increases by about 0.03883 for every one point increase in ACT score. Also, the line predicts a GPA of about 2.114 for an ACT score of 0. Of course you can do this quicker:

```
> fit <- lm(data$GPA ~ data$ACT)
> summary(fit)


Call:
lm(formula = data$GPA ~ data$ACT)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.74004 -0.33827  0.04062  0.44064  1.22737 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
data$ACT     0.03883    0.01277   3.040  0.00292 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262,   Adjusted R-squared:  0.06476 
F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

R gives a lot of output here, all of which we will discuss. For now, note the estimates of the regression coefficients indeed match the ones we computed above.

2.4 The Regression Model

Think of the ith response value Y_i as a random variable that is a linear function of the ith predictor value X_i, plus some random error term ε_i, which can be positive or negative:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Here, the β_0 and β_1 are the *actual* intercept and slope parameters you would find if you could use all possible observations, and we estimate them with b_0 and b_1 in (2.2) and (2.3) or (2.4). We insist that

$$E[\varepsilon_i] = 0 \quad \text{and} \quad \mathrm{Var}(\varepsilon_i) = \sigma^2$$

for each i, and that the ε_i are independent. Later we may want to allow the explanatory variable to be a random variable, but for now and in much of what we do we want to think of it as a deterministic variable we can control. Definitely think of Y_i as a random variable whose distribution depends on the X_i variable.

Summary of facts about our model:

1. The errors ε_i are independent random variables with mean 0 and common variance σ^2. For now, and in much of our study of regression, we will assume they are normally distributed.

2. The mean of the response, $E[Y_i]$, at each value of the predictor, x_i, is a linear function of the x_i. In fact,

$$E[Y_i|X_i = x] = \tag{2.5}$$

3. The *Gauss-Markov theorem* says the estimators $b_0 = \hat{\beta}_0$ and $b_1 = \hat{\beta}_1$ given in (2.2) and (2.3) or (2.4) are *unbiased* and have the smallest variance among all unbiased linear estimators.

4. At some point we will relax the normality assumption on the ε_i and also allow the X_i to be random.

2.5 Estimating the Variance in the Error Terms

We want to be able to estimate the common variance σ^2 of the errors:

$$\sigma^2 = \mathrm{Var}(\varepsilon_i).$$

Small σ^2 is desirable, indicating the response can be predicted with some degree of certainty:

We cannot know σ^2, but we can estimate it with the *error mean square* or *mean squared error* MSE:

$$MSE = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{SSE}{n-2}.$$

More generally, a mean-square is a sum of squares divided by its degrees of freedom. We will encounter many different sums of squares and mean-squares in our study. The number of degrees of freedom is the number of data points minus the number of model parameters estimated (here we've used b_0 and b_1 to estimate β_0 and β_1, hence the '$n-2$').

It turns out that under our model assumptions, MSE is an unbiased estimator (see section 2.6) for the common variance σ^2, meaning $E[MSE] = \sigma^2$.

Example 14 Recall the regression output from the ACT vs. GPA example. There were 120 observations (nrow(data) gives the number of rows in the data frame), and so the degrees of freedom for the error sum of squares SSE is $n - 2 = 118$. The error mean square is also called the residual standard error (you might recall that standard error is the standard deviation of a mean), and was given in the output as 0.6231 of freedom. So our "best guess" for σ^2 is $s^2 = MSE = .6231^2 \approx .3883$, and $SSE = .6231^2 \times 118 \approx 45.81$.

2.6 Bias and Mean Squared Error

If we use the random quantity $\hat{\theta}$ as an estimator for a population parameter θ, the *bias* of $\hat{\theta}$ is

$$B(\hat{\theta}) := E[\hat{\theta}] - \theta.$$

An estimator is *unbiased* if its bias equals 0, or equivalently, if its expected value equals the parameter you're using it to estimate. For example, the sample mean is an unbiased estimator for the population mean because if you take many random samples from a population, and calculate the sample means for each one, and average them, the average tends toward the population average as the number of samples increases. Here is proof:

While unbiasedness is an attractive property for estimators to have, not all estimators are unbiased. The sample variance

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(Y_i - \overline{Y})^2$$

is an unbiased estimator for the population variance σ^2. The quantity

$$\frac{1}{n}\sum_{k=1}^{n}(Y_i - \overline{Y})^2$$

is a *biased* estimator for σ^2 having bias $-\sigma^2/n$ (it tends to fall a little short, on average). For another example, the sample maximum is a biased estimator for the population maximum. A little thought should reveal its bias is also negative.

Besides unbiasedness, there are other attractive qualities estimators can have (e.g., small variance, maximum likelihood, consistency, etc.), and sometimes you might wish to sacrifice one quality for another.

An estimated regression coefficient b_i is an unbiased estimate of the parameter β_i if the mean of all of the possible estimates b_i equals β_i. The predicted response $\hat{Y}_i$ is an unbiased estimator of μ_i if the mean of all of the possible predicted responses $\hat{Y}_i$ equals μ_i. Again, by the Gauss-Markov theorem, $b_0 = \hat{\beta}_0$ and $b_1 = \hat{\beta}_1$ given in (2.2) and (2.3) or (2.4) are attractive estimators for β_0 and β_1 because they are unbiased and have the smallest variance among all unbiased linear estimators.

The *total mean squared error* or simply *mean squared error* of an estimator is the expected squared difference between the estimator and the quantity being estimated:

$$MSE(\hat{\theta}) := E[(\hat{\theta} - \theta)^2].$$

It can be shown that

$$MSE(\hat{\theta}) = B(\hat{\theta})^2 + \mathrm{Var}(\hat{\theta}), \tag{2.6}$$

which is nice if you think in terms of accuracy and precision:

2.7 $SSTO$, SSR, SSE, and the Coefficient of Determination

The *total sum of squares* $SSTO$, *regression sum of squares* SSR, and *error sum of squares* SSE are

$$\begin{aligned} SSTO &:= \sum_{i=1}^{n}(y_i - \overline{y})^2, && (2.7)\\ SSR &:= \sum_{i=1}^{n}(\hat{y} - \overline{y})^2, && (2.8)\\ SSE &:= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. && (2.9) \end{aligned}$$

In the same spirit of the partitioning in (2.6), it turns out that

$$SSTO = SSR + SSE.$$

As you vary the explanatory variable, you'd like to know to what extent it seems to explain the variation in the response variable. One way to do this is with the *coefficient of determination* r^2:

$$r^2 := \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{2.10}$$

You should interpret r^2 as the proportion of variation in the data that is captured by the regression model:

It turns out the coefficient of determination equals the square of the linear correlation coefficient (which is why we call it "r^2").

CAUTION: the coefficient of determination r^2 is not necessarily a good way to evaluate a model!

Example 15 Pull up the data on Ten Used Cars.

```
> age <- c(3, 6, 8, 4, 2, 11, 4, 7, 7, 9)
> price <- c(172, 140, 112, 160, 165, 80, 155, 103, 84, 78)
> linreg <- lm(price ~ age)
> summary(linreg)


Call:
lm(formula = price ~ age)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.963 -10.653   7.004  10.037  14.646 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  199.029     12.233  16.270 2.05e-07 ***
age          -12.152      1.834  -6.627 0.000165 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.66 on 8 degrees of freedom
```

```
Multiple R-squared:  0.8459,    Adjusted R-squared:  0.8266
F-statistic: 43.91 on 1 and 8 DF,  p-value: 0.0001647
```

```
> summary(linreg)$r.squared
```

```
[1] 0.8459004
```

So r^2 is fairly high in this case, indicating a large proportion of variation in the data is captured by the regression line. However, suppose I found another data point... a car that is 55 years old selling for $99,000.00. Let's add this data point to our data set and reevaluate:

```
> uc <- data.frame(age, price)
> uc[11,1] <- 55
> uc[11,2] <- 990
> plot(uc[,2], uc[,1])
> out <- lm(uc[,2]~ uc[,1])
> summary(out)
```

```
Call:
lm(formula = uc[, 2] ~ uc[, 1])

Residuals:
    Min      1Q  Median      3Q     Max
-131.15  -54.62   12.47   63.41  104.36

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    27.20      31.79   0.855    0.414
uc[, 1]        16.72       1.79   9.343 6.28e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 84.84 on 9 degrees of freedom
Multiple R-squared:  0.9065,    Adjusted R-squared:  0.8961
F-statistic: 87.29 on 1 and 9 DF,  p-value: 6.282e-06
```

The coefficient of determination has increased to 90.65%. However, if you inspect the plot, you can see why this is, and how the new model is probably not so good. Note the regression line equation has changed substantially, as the sign of the slope has even changed. It also predicts a price of $2700.00 for a new car. Probably a curvilinear model (like a quadratic) would be a better model, which we will study in the general linear model.

2.8 Homework

1. Show that the standard deviation of z-scores equals 1.

2. Refer to the original *Used Cars* example with ten observations.

   (a) Use the regression line we obtained with R to predict the price of a 4.5 year-old car.

   (b) Use the regression line we obtained with R to predict the price of a 60 year-old car. Does this prediction make any sense? Explain. What went wrong?

   (c) Suppose your friend just sold a 70 year-old car for \$50,000, and so we add that to our data (70, 500). You can use the rbind() command.

      i. What is the new regression line equation?

      ii. What is the new r value?

      iii. Make a scatterplot of the new data, and include both regression lines (original and new one). Is the new regression line equation better or worse than the original one for predicting used car prices? Explain.

      iv. How is this new observation affecting our model?

      v. Remark on the fact that the new coefficient of determination is reasonably high, yet the new model is not very good.

3. Open the data set called *faithful* in R simply by typing *faithful* at the prompt. The first column contains eruption durations in minutes and the second contains the times between eruptions, also in minutes. The data are rounded quite a bit- a minute is a relatively long time. Use R to make a scatterplot of the duration vs. the times between. Also perform linear regression. Remark on the plot and the regression output, especially on r^2 and how it corresponds to the point pattern in the plot.

4. Show the slope coefficient is indeed $b_1 = r \cdot s_y/s_x$.

5. Explain how the ratio s_y/s_x affects the slope of the regression line through the original data.

6. Show that the regression line that runs through the standardized data values has the form
$$\hat{z}_y = r \cdot z_x.$$

7. Show that the coefficient of determination is indeed equal to the square of the linear correlation coefficient (hence the nickname "r^2").

8. Show that for any point estimator $\hat{\theta}$ that $MSE(\hat{\theta}) = B(\hat{\theta})^2 + \text{Var}(\hat{\theta})$.

9. Prove that the sample variance is an unbiased estimator for the population variance. That is, show that

$$E\left[\frac{1}{n-1}\sum_{k=1}^{n}(Y_i - \overline{Y})^2\right] = \sigma^2.$$

10. Verify the bias of

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

in estimating the population variance is $-\sigma^2/n$.