Chapter 4: Correlation and regression

4.1 a) Two variables can either be positively related, not related, or negatively related. The most logical relationship is that the distance a customer lives from the store is negatively related to how many times they visit the store.

4.1 b)

i	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	4.87	2.90	1.868	-1.09	-2.053
2	3.04	4.50	0.038	0.501	0.0194
3	1.65	4.94	-1.351	0.941	-1.272
4	2.88	3.28	-0.121	-0.718	0.087
5	2.31	4.73	0.691	0.731	-0.505
6	3.96	2.64	0.958	-1.358	-1.303
7	2.70	3.70	-0.301	-0.298	0.089
8	2.60	5.30	-0.401	1.301	-0.522

$\bar{x} = 3.001$ $\sum(x_i - \bar{x})(y_i - \bar{y}) = -5.458$

$\bar{y} = 3.998$ $n - 1 = 7$

$s_{xy} = -0.779$

4.1 c) The covariance is negative. A negative covariance indicates that that as one variable deviates from the mean, the other variable deviates in the other direction. This means that, when a customer's distance from the store in kilometers is higher than the mean, their average visits per week will likely be lower than the mean.

4.1 d) The disadvantage of using the covariance as a measure for the strength of this relationship is that it depends on the measurement unit (kilometers vs. meters) that the co-worker asks the questions in. If the co-worker would have asked the question in meters the covariance would have increased by a 1000 times, namely -779.84.

4.1 e) $s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} = \sqrt{\frac{6.977}{7}} = 0.998$ $s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}} = \sqrt{\frac{7}{7}} = 1$

4.1 f) $r_{xy} = \frac{s_{xy}}{s_x \times s_y} = \frac{-0.779}{0.998 \times 1} = -0.779$

4.1 g) The coefficient is -0.779, which represents is a relatively strong negative relationship.

4.2 a) $H_0: \quad \rho_{xy} \geq 0$ $\qquad H_1: \quad \rho_{xy} < 0$

4.2 b) $N =$ unknown $\qquad r_{xy} = -0.779$
$n = 8$ $\qquad \rho_{xy} =$ unknown

4.2 c) $z_{0.05} = -1.645$

4.2 d) $z_r = \frac{1}{2} \times log_e(\frac{1+r}{1-r}) = \frac{1}{2} \times log_e(\frac{1-0.779}{1+0.779}) = -1.043$

$SE_r = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{8-3}} = 0.447$

$z_{xy} = \frac{z_r}{SE_r} = \frac{-1.043}{0.447} = -2.33$

4.2 e) The observed z-value is more extreme (lower) than the critical z-value. H_0 is rejected with 95% confidence. You can be 95% confident that ρ_{xy} is negative in the population. There is a 5% change of a type-I error.

4.3 a)

```
# Be sure to set your working directory when providing a relative path
dataset6 <- read.csv('localSupermarket.csv')
```

4.3 b) Covariance: -0.30
Correlation: -0.30

```
cov(dataset6$Distance, dataset6$AvgVisits) # Covariance -0.3000603
cor(dataset6$Distance, dataset6$AvgVisits) # Correlation: -0.3000382
```

4.3 c)

```
cor.test(dataset6$Distance, dataset6$AvgVisits, alternative = 'less')
# Correlation: -0.30
# t-value: -9.936
# p-value: < 2.2e-16
```

4.3 d) The black line represents the normal distribution. The red line represents the t-distribution. The difference between the two distributions, in terms of their shape, is that the t-distribution has slightly thicker tails. When you increase the degrees of freedom of the t-distribution, it will start to look more like the normal distribution.

```
curve(dnorm(x, mean = 0, sd = 1), from = -3, to = 3, ylab = 'Density')
curve(dt(x, df = 3), from = -3, to = 3, add = TRUE, col = 'red')
```

4.3 e) $df = 998$ $t_{xy} = -9.936$

```
n <- nrow(dataset6)
dft <- n - 2 # 998

r <- cor(dataset6$Distance, dataset6$AvgVisits)
tscore <- r * sqrt(n - 2) / sqrt(1 - r^2)
# t-score: -9.936 so you can confirm the value in 4.3c
```

4.3 f) You can find the t-value in the bottom line of the output in the console.

4.3 g) The p-value is < 2.2e-16, which is lower than the significance level of 5%. This means that H_0 can be rejected with 95% confidence.

4.4 a)

```
# Be sure to set your working directory when providing a relative path
dataset7 <- read.csv('nationalSupermarket.csv')
```

4.4 b)

```
plot(x = dataset7$Price,
     y = dataset7$AvgWasted,
     main = 'Scatter plot of Price vs. AvgWasted',
     ylab = 'Average number of cartons wasted',
     xlab = 'Price of a carton of milk',
     las = 1,
     col = 'orange',
     pch = 19,
     bty = 'n')
```

4.4 c) AvgWasted $= \beta_0 + \beta_1 \times$ Price

4.4 d)

```
lmfit <- lm(formula = AvgWasted ~ Price, data = dataset7)
```

4.4 e) AvgWasted $= 0.236 + 2.995 \times$ Price

```
summary(lmfit)
# b0 = 0.236
# b1 = 2.995
# R-squared: 0.64
```

4.4 f)

```r
abline(lmfit)
```

4.4 g) $R^2 = 0.64$

Interpretation: The multiple R^2 is 0.64, meaning that 64% of the variation in the number of milk cartons that are thrown away each day can be explained by the price of the milk cartons.

4.4 h) H_0: $\beta_1 \leq 0$ H_1: $\beta_1 > 0$

4.4 i) The p-value for the regression coefficient is < 2e-16, which is lower than the significance level of 5%. H_0 can be rejected with 95% confidence. You can be 95% sure that β_1 is positive in the population. The price contributes significantly to the average number of milk cartons thrown away. There is a 5% risk of a type-I error.

4.5 a)

```r
newdata <- data.frame(Price = 0.70)
```

4.5 b) Predicted value: 2.33

```r
predict(object = lmfit,
        newdata = newdata) # Prediction: 2.33
```

4.5 c) Predicted value: $0.236 + 2.995 \times 0.70 = 2.33$

4.5 d)

```r
predict(object = lmfit, newdata = newdata,
        interval = 'prediction', level = 0.90)
# Lower bound: 0.734
# Upper bound: 3.931
```

4.5 e) The supermarket will throw away fewer cartons of milk.

Explanation: The current number of milk cartons thrown away (4) lies outside the bounds of the 90% confidence interval for the prediction.