STAT 6315
Fall 2020
J. Reeves
Homework 7
Wes Bonelli
11/20/2020

1. Let

   X : water (inches)
   Y : flow rate

   and

   $$Y = -0.12 + 0.095X.$$

   In R:

   ```
   > y = function(x) { return(-0.12 + 0.095 * x) }
   ```

   Plugging into R:

   ```
   > y(10)
   [1] 0.83
   > y(15)
   [1] 1.305
   ```

a. The mean flow rates for a pressure drop of 10 and 15 inches are (respectively) 0.83 and 1.305.

b. The average change in flow rate associated with a 1 inch increase in pressure drop is the average of all consecutive differences in flow rate values predicted by the model. Since this model is linear, flow rate changes by the same amount for every equally-sized change in pressure drop: namely the model slope, 0.095.

   This can be verified in R:

   ```
   > ys = sapply(seq(5, 20), y)
   > mean(diff(ys))
   [1] 0.095
   ```

c. The mean difference between flow rate when pressure drop is 15 in. vs. 10 in. is given by

   $$1.305 - 0.83 = 0.475.$$

   Since this is a linear combination, we can estimate the standard error of the difference distribution as

   $$\sigma_D = \sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{0.2^2 + 0.2^2} = \sqrt{0.04 + 0.04} = \sqrt{0.08} \approx 0.2828$$

   and find the probability with the complement of the cumulative density function in R:

```
> 1 - pnorm(0, 0.475, 0.2828)
[1] 0.9534856
```

The probability that flow rate is greater when pressure drop is 15 in. than when it's 10 is approx. 95.3%.

2. Importing the data into R:

```
> year = c(1,2,3,4,5)
> hrt = c(46.30, 40.60, 39.50, 36.60, 30.00)
> bci = c(103.30, 105.00, 100.00, 93.80, 83.50)
> hrt_data = data.frame(Year = year, HRT = hrt, BCI = bci)
> hrt_model = lm(BCI ~ HRT, data = hrt_data)
```

a. In R:

```
> hrt_model

Call:
lm(formula = BCI ~ HRT, data = hrt_data)

Coefficients:
(Intercept)         HRT
     45.573       1.335
```

The equation of the estimated regression line is

$$Y = 45.573 + 1.335X.$$

b. The estimated average change in BCI associated with a 1 percentage point increase in HR use is 1.335.

c. Plugging into R:

```
> predict(hrt_model, data.frame(HRT = c(40)))[1]
       1
98.98959
```

When HRT use is 40%, we should predict BCI to be approx. 98.99.

d. The regression equation should not be used to make a prediction for HRT = 20% because it is an interpolation model. Extrapolation is not justified in this case; we can be relatively confident predicting values between (or very near) the minimum and maximum HRT values, but not beyond them.

e. In R:

```
> summary(hrt_model)

Call:
lm(formula = BCI ~ HRT, data = hrt_data)
```

```
Residuals:
      1       2       3       4       5
-4.1027  5.2092  1.6781 -0.6492 -2.1354

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.5727    13.5804   3.356   0.0439 *
HRT           1.3354     0.3485   3.832   0.0313 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.154 on 3 degrees of freedom
Multiple R-squared:  0.8303,       Adjusted R-squared:  0.7738
F-statistic: 14.68 on 1 and 3 DF,  p-value: 0.03132
```

This model's $R^2 = 0.8303$ and, since this is simple linear regression, $r = \sqrt{R^2} \approx 0.911$.

f. This model's $\hat{\sigma}_e = 4.154$.

3. Reading in the data:

```
> chi_data = data.frame(read_excel("CHI.xls")[,-4])
> chi_model = lm(CHI ~ Control, data = chi_data)
> chi_model
```

a. Fitting the linear regression:

```
Call:
lm(formula = CHI ~ Control, data = chi_data)

Coefficients:
(Intercept)      Control
    -96.671        1.595
```

b. In R:

```
> summary(chi_model)

Call:
lm(formula = CHI ~ Control, data = chi_data)

Residuals:
    Min      1Q  Median      3Q     Max
-57.520 -37.759  -1.848  10.450 114.379

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -96.6709    44.2955  -2.182   0.0606 .
Control       1.5946     0.0587  27.165 3.63e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53.67 on 8 degrees of freedom
```

```
Multiple R-squared:  0.9893,     Adjusted R-squared:  0.9879
F-statistic: 737.9 on 1 and 8 DF,  p-value: 3.633e-09
```

Our p-value (3.633e-09) is well below $\alpha = 0.05$. There is a very strong linear relationship between the mean response time for uninjured individuals and individuals with CHI.

c. In R:

```
> chi_model_zero = lm(CHI ~ Control + 0, data = chi_data)
> chi_model_zero

Call:
lm(formula = CHI ~ Control + 0, data = chi_data)

Coefficients:
Control
  1.476
```

The least-squares estimate $\hat{\beta} = 1.476$. This suggests that for every unit of time a non-injured individual requires to complete a task, a CHI individual requires about 1.5 times as long.

4. Read in the data:

```
> mba_data = data.frame(read_excel("MBA.xls"))
```

a. Plot the data:

```
> plot(mba_data$EXPER, mba_data$SALARY)
```

Yes, it looks like students with less experience also tend to have smaller salaries.

b. One student with 14 years of work experience has one of the lowest recorded salaries.

5.

a. In R:

```
> cor(mba_data$EXPER, mba_data$SALARY)
[1] 0.6946505
```

The correlation coefficient is approx. 0.695. Yes, both the sign (positive) and size seem to agree with the scatter-plot; there is some spread in the data but they show a clear linear trend.

b. In R:

```
> cor(mba_data$EXPER, mba_data$SALARY, method = "spearman")
[1] 0.7042325
```

The Spearman rank correlation coefficient is approx. 0.704.

c. The Spearman correlation measures fit to any monotonic function, while Pearson's correlation measures fit to a straight line, making Pearson's correlation less flexible and more sensitive to outliers.

6.

a. In R:

```
> mba_model = lm(SALARY ~ EXPER, data = mba_data)
> summary(mba_model)

Call:
lm(formula = SALARY ~ EXPER, data = mba_data)

Residuals:
    Min      1Q  Median      3Q     Max
-23.5844  -1.6891   0.2953   2.3335  10.7499

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 100.6157     1.2918  77.886  < 2e-16 ***
EXPER         1.4906     0.2183   6.828 1.11e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.556 on 50 degrees of freedom
Multiple R-squared:  0.4825,    Adjusted R-squared:  0.4722
F-statistic: 46.63 on 1 and 50 DF,  p-value: 1.114e-08
```

The estimate of the slope is 1.4906 and intercept 100.6157. The intercept can be interpreted as the expected starting salary for those with no previous work experience.

b. The residual standard deviation RSME $= 5.556$. The average amount by which salary predictions deviate from the underlying data is about $2.4 thousand.

c. Yes, with a p-value $= 1.114\text{e} - 8$ there seems to be a significant relationship between salary and experience.

d. About 48% of variability in salary is accounted for by years of experience.

7.

a. The data value associated with the student would be considered both high influence (because its omission would cause a relatively significant change in the model slope) and high leverage (because it is highly unusual compared to the rest of the dataset).

b. The slope of the model would increase.

c. The removal of this outlier would cause the residual standard deviation to decrease.

d. The removal of this outlier would cause the correlation to increase.

8.

a. Refitting the model:

```
> mba_data_trunc = mba_data[-c(11),]
> mba_model_trunc = lm(SALARY ~ EXPER, data = mba_data_trunc)
> summary(mba_model_trunc)

Call:
lm(formula = SALARY ~ EXPER, data = mba_data_trunc)

Residuals:
    Min      1Q  Median      3Q     Max
-10.7150  -2.2212  -0.0523   2.6225   9.5101

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  99.2651     1.0150    97.8  < 2e-16 ***
EXPER         1.8875     0.1798    10.5 3.93e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.257 on 49 degrees of freedom
Multiple R-squared:  0.6922,       Adjusted R-squared:  0.6859
F-statistic: 110.2 on 1 and 49 DF,  p-value: 3.929e-14
```

The change in slope is given by

$$1.8875 - 1.4906 = 0.3969.$$

The change in residual standard deviation is given by

$$4.257 - 5.556 = -1.299.$$

b. In R:

```
> cor(mba_data_trunc$EXPER, mba_data_trunc$SALARY)
[1] 0.8319708
```

The change in correlation coefficient is given by

$$0.8319708 - 0.6946505 \approx 0.137.$$

c. In R:

```
> cor(mba_data_trunc$EXPER, mba_data_trunc$SALARY, method = "spearman")
[1] 0.7993346
```

The change in Spearman rank correlation coefficient is given by

$$0.8319708 - 0.7042325 \approx 0.128.$$

d. The change in Spearman rank correlation was smaller than the change in standard correlation coefficient.

9. Reading the data in:

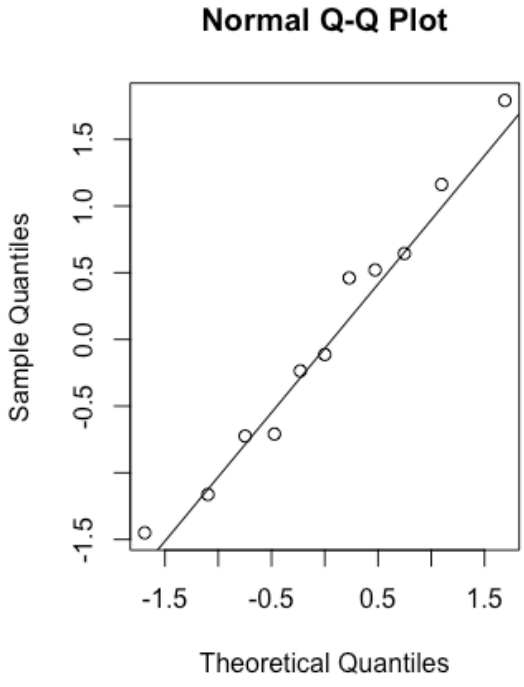
```
> bear_data = data.frame(read_excel('Bears.xls')[,-4])
```

a. Fitting the model:

```
> bear_model = lm(Range ~ Age + Weight, data = bear_data)
> summary(bear_model)

Call:
lm(formula = Range ~ Age + Weight, data = bear_data)

Residuals:
    Min      1Q  Median      3Q     Max
-20.820  -9.383  -1.491   7.310  25.183

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.4283    31.9113   0.828    0.432
Age           0.3499     0.8340   0.420    0.686
Weight        0.3033     0.6700   0.453    0.663

Residual standard error: 15.07 on 8 degrees of freedom
Multiple R-squared:  0.08618,    Adjusted R-squared:  -0.1423
F-statistic: 0.3772 on 2 and 8 DF,  p-value: 0.6973
```

b. Creating a residual plot:

```
> bear_resids = rstandard(bear_model)
> qqnorm(bear_resids)
> qqline(bear_resids)
```



Yes, the residuals can be fitted by an approximately straight line, so the random deviation distribution can be regarded as approximately normal.

c. Creating individual effects models and null model:

```
> bear_model_age = lm(Range ~ Age, data = bear_data)
> bear_model_weight = lm(Range ~ Weight, data = bear_data)
> bear_model_null = lm(Range ~ rep(mean(Range), times = length(Range)), data =
bear_data)
> summary(bear_model_age)

Call:
lm(formula = Range ~ Age, data = bear_data)

Residuals:
    Min      1Q  Median      3Q     Max
-21.078  -9.117  -2.916   6.690  25.370

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  40.3769     7.9267   5.094 0.000651 ***
Age           0.5370     0.6917   0.776 0.457433
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.39 on 9 degrees of freedom
Multiple R-squared:  0.06277,      Adjusted R-squared:  -0.04136
F-statistic: 0.6028 on 1 and 9 DF,  p-value: 0.4574

> summary(bear_model_weight)

Call:
lm(formula = Range ~ Weight, data = bear_data)

Residuals:
    Min      1Q  Median      3Q     Max
-20.725  -9.676  -1.542   7.971  24.018

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.5536    29.1135   0.775    0.458
Weight        0.4426     0.5546   0.798    0.445

Residual standard error: 14.37 on 9 degrees of freedom
Multiple R-squared:  0.06608,      Adjusted R-squared:  -0.03769
F-statistic: 0.6368 on 1 and 9 DF,  p-value: 0.4454

> summary(bear_model_null)

Call:
```

```
lm(formula = Range ~ rep(mean(Range), times = length(Range)),
    data = bear_data)

Residuals:
    Min      1Q  Median      3Q     Max
-21.127 -10.777  -2.427  11.773  23.173

Coefficients: (1 not defined because of singularities)
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                           45.527      4.252   10.71 8.47e-07 ***
rep(mean(Range), times = length(Range))       NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.1 on 10 degrees of freedom
```

At the $\alpha = 0.05$ level, none of our models can be judged useful to predict home-range size from weight or age (all have p-values much greater than 0.05).