Simonson_HW8

Martin A. Simonson

March 10, 2019

1.) Rowan Adaptation

The rowan (*Sorbus aucuparia*) is a tree that grows in a wide range of altitudes. To study how the tree adapts toits varying habitats, researchers collected twigs with attached buds from 12 trees growing at various altitudes in North Angus, Scotland. The buds were brought back to the laboratory and measurements were made on the dark respiration rate. The average altitude from the origin for the 12 trees is 433.33 meters. The average respiration rate computed for the 12 trees is 0.21 μl oxygen per hour per mg dry weight of tissue. The standard deviation of the altitude for the 12 trees is 214.62 meters and for respiration rate is 0.077 $\mu l/(hour * mg^{-1})$ mg.The correlation between altitude from the origin and dark respiration rate(r_{xy}) was 0.887.

(a)

We will use linear regression to describe the relationship between altitude and dark respiration. Which of the two variables would most naturally be considered the explanatory variable and which would be the response variable?

- **Answer:** The explanatory variable is the elevation from which the sample buds were taken, in meters. The response variable is the dark respiration rate in micrograms oxygen per hour per mg of dry bud tissue [$\mu l/(hour * mg^{-1})$].

(b)

Compute the equation of the least squares regression line.

- **Answer:** We want to fit the model $y = \hat{\beta}_0 + \hat{\beta}_1 * x$ where $\hat{\beta}_0$ is constant and represents the least squares regression incercept, and $\hat{\beta}_1$ is constant and represents the least squares regression slope. To calculate this from available data, we first estimate $\hat{\beta}_1$ by:

$$\hat{\beta}_1 = r_{xy} * \frac{S_y}{S_x}$$

where r_{xy} is the correlation coefficient of the two variables; the sample standard deviations for y (dark respiration) and x (elevation) are represented by S_Y and S_x, respectively.

```
r.xy<-0.887
S.y<-0.077
S.x<-214.62

beta.1<-r.xy*(S.y/S.x)
beta.1
```

```
## [1] 0.0003182322
```

The formula for β_0 is given by:

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1} * \bar{x}$$

```
y.bar<-0.21
x.bar<-433.33

beta.0<-y.bar-(beta.1*x.bar)
beta.0
```

```
## [1] 0.07210043
```

Therefore, the equation of the least squares regression line for the respiration rate in response to tree altitude is:

$$y = 0.07210043 + 0.0003182322 * x$$

(c)

Compute a 95% confidence interval for the slope of the regression line.

- **Answer:** We want to test $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$. First we need to compute the root mean squared error (RMSE, $\hat{\sigma}$):

$$RMSE = S_y * \sqrt{1 - r_{xy}^2}$$

```
rmse<-S.y * sqrt(1-(r.xy^2))
rmse
```

```
## [1] 0.03555625
```

Then, we compute the SE of β_1 with the following formula:

$$SE(\hat{\beta_1}) = \frac{\hat{\sigma}}{\sqrt{(n-1) * S_x^2}}$$

```
se.beta.1<-rmse/sqrt((12-1)*(S.x^2))
se.beta.1
```

```
## [1] 4.99516e-05
```

And compute the T-ratio

$$T = \frac{\hat{\beta_1}}{SE(\hat{\beta_1})}$$

```
t.statistic<-beta.1/se.beta.1
t.statistic
```

```
## [1] 6.370812
```

We observe a t-statistic of 6.370812 with 10 degrees of freedom. From the t-table we find a twp-tailed p-value of less than 0.001, providing strong evidence to reject the null hypothesis. To complete the 95% confidnce interval, we use the formula:

$$95\%CI = \hat{\beta_1} \pm t^* * SE(\hat{\beta_1})$$

From the t-table, using 10 degrees of freedom, we find the t^* from the t-table is 2.228.

```
lower<-beta.1-(2.228*se.beta.1)
upper<-beta.1+(2.228*se.beta.1)
CI<-data.frame(lower,upper)
CI
```

```
##          lower        upper
## 1 0.0002069401 0.0004295244
```

We are 95% confident that a 1 meter rise in altitude is accompanied by between 0.0002069401 and 0.0004295244 increase in respiration, measured as $\mu l/(hour * mg^{-1})$. The fact that the confdiecnet inverval for $\hat{\beta}_1$ does not include zero, along with the p-value above, provide very strong evidence that the slope of the the least squares regression line does not equal 0.

(d)

Estimate the mean respiration rate for a tree growing at 300 meters above origin.

```
x<- 300
mu.300<- 0.07210043 + 0.0003182322*x
mu.300
```

```
## [1] 0.1675701
```

- **Answer:** We would expect the respiration rate for a tree growing at an elevation of 300 meters to be 0.1675701 $\mu l/(hour * mg^{-1})$.

(e)

If the value of the new explanatory variable lies within one s.d. of the mean of the values of explanatory variables, the prediction is called interpolation, otherwise it is called extrapolation. Do you think your estimate in part (d) involved interpolation or extrapolation? Explain.

```
x.range<-c((433.33 - 214.62),(433.33 + 214.62))
x.range
```

```
## [1] 218.71 647.95
```

- **Answer:** The value provided in part (d), 300 meters, is within one standard deviation of the mean elevation (218.71 meters to 647.95 meters), therefore, this estimate involved interpolation.

(f)

Compute the s.e. of the mean respiration rate at 300 meters elevation.

- **Answer:** Let $\hat{\mu}_{300}$ represent the mean predicted respiration rate at 300 meters elevation, as calculated by $\hat{\mu}_{300} = \hat{\beta}_0 + \hat{\beta}_1 * 300$. We want to compute the standard error for $\hat{\mu}_{300}$:

$$SE(\hat{\mu}_{300}) = \hat{\sigma} * \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1) * S_x^2}}$$

Where x is equal to 300, $\bar{x}$ is equal to 433.33, n is equal to 12, and S_x is equal to 214.62.

```
se.mu.300<-rmse*sqrt((1/12)+(((300-433.33)^2) / ((12-1)*(S.x^2))))
se.mu.300
```

```
## [1] 0.01223561
```

The standard error of the mean respiration rate at 300 meters elevation is 0.01223561 $\mu l/(hour * mg^{-1})$.

2.) Old Faithful

The data in *eruption.csv* are the interval waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming. Ignore the date variable and consider all observations as one sample of X and Y values. Answer the following questions:

(a)

Fit the regression model that predicts interval from the duration of the last eruption and report the estimated intercept and slope.

```
df<-read.csv("Data/eruption.csv",header=T)
summary(df)
```

```
##       date          interval       duration
##  Min.   : 1.00   Min.   :42.0   Min.   :1.700
##  1st Qu.:15.50   1st Qu.:59.0   1st Qu.:2.300
##  Median :42.00   Median :75.0   Median :3.800
##  Mean   :42.73   Mean   :71.0   Mean   :3.461
##  3rd Qu.:68.50   3rd Qu.:80.5   3rd Qu.:4.300
##  Max.   :95.00   Max.   :95.0   Max.   :4.900
```

```
df$interval<-as.numeric(as.character(df$interval))
model<-lm(interval~duration, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = interval ~ duration, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.644  -4.440  -1.088   4.467  15.652
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.8282     2.2618   14.96   <2e-16 ***
## duration     10.7410     0.6263   17.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.683 on 105 degrees of freedom
## Multiple R-squared:  0.7369, Adjusted R-squared:  0.7344
## F-statistic: 294.1 on 1 and 105 DF,  p-value: < 2.2e-16
```

- **Answer:** The estimated interval length intercept is 33.8282 minutes with a slope of 10.7410 minutes.

(b)

Calculate and report a 95% confidence interval for the slope

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

```
##                 2.5 %   97.5 %
## (Intercept) 29.343441 38.31297
## duration     9.499061 11.98288
```

- **Answer:** We are 95% confident that the mean increase in chirping rate lies between 9.499 and 11.983 minutes per minute of eruption duration.

(c)

Test whether the slope equals 0. Report your test statistic, p-value and a one-sentence conclusion.

```
summary(model)
```

```
##
## Call:
## lm(formula = interval ~ duration, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.644  -4.440  -1.088   4.467  15.652
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.8282     2.2618   14.96   <2e-16 ***
## duration     10.7410     0.6263   17.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.683 on 105 degrees of freedom
## Multiple R-squared:  0.7369, Adjusted R-squared:  0.7344
## F-statistic: 294.1 on 1 and 105 DF,  p-value: < 2.2e-16
```

- **Answer:** Let β represent the slope of the relationship between interval time and eruption time (both in minutes). We want to test: $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$. The t-statistic for the effect of duration from our model summary is 17.15 on 105 degrees of freedom. The associated p-value less than 0.0001 provides very strong evidence to reject the null hypothesis and conclude the true slope of the relationship betweene eruption duration and interval is not equal to 0.

(d)

Rangers at Yellowstone report a interval describing when they expect the next eruption to occur. Is it more appropriate to provie a confidence interval or a prediction interval? Briefly explain why.

- **Answer:** The rangers should use a prediction interval because it will account for the scatter of data, and estimate variance around a predicted value.

3.) Wormyfruit

It is generally thought that the percentage of fruit attacked by codling moth larvae is greater on apple trees bearing a small crop. Apparently the density of the flying moths is unrelated to the size of the crop on a tree, so the chance of attack for any particular fruit is increased if few fruits are on the tree. Data collected for a random sample of 10 trees gives a sample linear correlation coefficient of -0.77. Other summary statistics obtained from the sample are provided below. The data are available in *wormyfruit.csv*.

Variable	Sample Mean	Sample Standard Deviation
Crop Size (#fruit)	126.702	59.924
% wormy fruits	41.539	12.070

(a)

Estimate the least-squares regression line for predicting percentage of wormy fruits from crop size. Show all your work in addition to the equation of the estimated model.

```
df<-read.csv("Data/wormyfruit.csv",header=T)
str(df) # size of crop x; percent of fruit attacked is y
```

```
## 'data.frame':    10 obs. of  2 variables:
##  $ Perc: num  50.8 66.9 22.5 43.3 45.1 ...
##  $ Size: num  66.7 63.5 189.9 171 51.7 ...
```

```
model<-lm(Perc~Size,data=df)
summary(model)
```

```
##
## Call:
## lm(formula = Perc ~ Size, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.125 -6.058 -1.196  3.628 14.619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 62.96802    6.79719   9.264  1.5e-05 ***
## Size        -0.16913    0.04962  -3.409  0.00924 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.175 on 8 degrees of freedom
## Multiple R-squared:  0.5922, Adjusted R-squared:  0.5413
## F-statistic: 11.62 on 1 and 8 DF,  p-value: 0.00924
```

- **Answer:** Let Y be the percent of fruit attacked and let X be the size of a crop. The equation for the estimated linear model of Y in response to X is:

$$y = 62.96802 + (-0.16913 * x)$$

(b)

Write down the ANOVA table for the simple linear regression of percentage of wormy fruit on crop size. Compute the F statistic and find a p-value.

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Perc
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Size       1 776.60  776.60   11.62 0.00924 **
## Residuals  8 534.67   66.83
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Answer:** The F-statistic is 11.62 with 8 and 1 degrees of freedom, and an observed p-value 0.00924. Therefore, we have strong evidence against the null hypothesis that the true effect of crop size is not equal to zero.

(c)

Based on the F statistic and p-value computed in part (b), is there statistically significant evidence that the slope of the regression line is significantly different from 0?

- **Answer:** There is strong evidence to reject the null hypothesis that the slope of the regression is equal to 0, and therefore conclude that the mean percentage of wormy fruit decreases with crop size.

(d)

What proportion of the variability in the percentage of wormy fruit is explained by the regression of percentage of wormy fruit on crop size?

- **Answer:** The model summary from (a) returns an adjusted R^2 of 0.5413, meaning about 54.13% of the variability in the percentage of wormy fruit is explained by the regression of percentage of wormy fruit on crop size.

(e)

Estimate the mean percentage of wormy fruits for trees with a crop size of 150 fruit.

```
new.data = data.frame(Size=c(150))

perc<-predict(model, newdata = new.data,se.fit = T,interval = "conf",level=0.95)
perc$fit # estimated percentage infestation at 150 apples
```

```
##        fit      lwr      upr
## 1 37.59886 31.06851 44.12922
```

- **Answer:** We expect 37.59 percent of apples would be infested with worms would if the crop size were 150 apples.

(f)

Provide a 95% confidence interval for the mean percentage of wormy fruits for trees witha crop size of 150 fruit.

```
perc$fit
```

```
##        fit      lwr      upr
## 1 37.59886 31.06851 44.12922
```

- **Answer:** We are 95% confident that the interval for the mean percentage of wormy fruits for trees with a crop size of 150 fruit is between 31.07% and 44.13%.

(g)

Is there a statistically significant evidence from these data that trees with a crop size of 150 have, on average, 50% wormy fruit?

- **Answer:** No, we would not expect that another sample with a crop size of 150 apples would have 50% wormy fruit.

(f)

Provide a 90% prediction interval for the average percentage of wormy fruit for the first tree with 150 fruit that was selected to be part of the study.

```
perc2<-predict(model, newdata = new.data, se.fit = T, interval = "predict",level=0.90)
perc2$fit
```

```
##        fit      lwr      upr
## 1 37.59886 21.51042 53.68731
```

- **Answer:** The 90% prediction interval for the mean percentage of wormy fruit in a tree with 150 fruit is between 21.51% and 53.69%.

4.) Butterfly Diversity

The data in *diversity.csv* are decribed in Chapter 8, problem 22 (both editions). Ecological theory suggests that the appropriate regression model is diversity $= \hat{\beta}_0 + \hat{\beta}_1 * log(area)$. Use this model for all questions below. The data come from an experimental study where the area of the patch was randomly assigned to plots of land.

(a)

Estimate the regression coefficients of the linear regression model suggested by ecological theory, then test whether the data provide evidence of a positive linear relationship between diversity and log(area). Report the estimates, the test statistic, and p-value for the test.

```
df<-read.csv("Data/diversity.csv",header=T)
str(df) # area x; number of species is y
```

```
## 'data.frame':    16 obs. of  2 variables:
##  $ area   : int  1 1 1 1 1 1 10 10 10 10 ...
##  $ species: int  14 50 55 34 40 57 43 103 33 53 ...
```

```
model<-lm(species~area,data=df)
summary(model)
```

```
##
## Call:
## lm(formula = species ~ area, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.860 -14.276  -5.749   1.945  53.361
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 55.76094    7.30351   7.635 2.35e-06 ***
## area         0.09878    0.02864   3.449  0.00391 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.29 on 14 degrees of freedom
## Multiple R-squared:  0.4594, Adjusted R-squared:  0.4208
## F-statistic:  11.9 on 1 and 14 DF,  p-value: 0.003912
```

- **Answer:** The estimate for $\hat{\beta}_0$ is 55.76094; the estimate for $\hat{\beta}_1$ is 0.09878. To test whether or not the data provide evidence of a positive linear relationship we observe a t-statistic for $\hat{\beta}_1$ of 3.449 on 14 degrees of freedom, with a p-value of 0.003912.

(b)

Does the mean diversity vary linearly with log(area), or does a model that allows the mean diversity to follow some other pattern appear to fit the data better? Provide the hypotheses, test statistic and a p-value in addition to your conclusion.

- **Answer:**

(c)

The investigators are interested in using this model to predict the number of species on new patches of forest. What is the area of a patch that has the smallest standard error for the predictions of the mean number of species? What is the area of a patch that has the smallest standard error for predictions of number of species in an individual patch?

- **Answer:**

(d)

The investigators would really like to make predictions of diversity in an individual patch that has standard errors less than 20 species. Is that possible with this model and data? Explain why or why not.

- **Answer:**

(e)

The study that produced these data has attracted a lot of international attention. Imagine that this study can be repeated with 10 times as many plots (160 instead of the 16 here). All else about the system remains the same: the regression intercept, slope, and error variance (i.e. $\hat{\sigma}$) are the same as here. Will this study be able to predict diversity in an individual patch with a standard error less than 20 species? Briefly explain why or why not.

- **Answer:**

5.) Manatees

Manatees (a.k.a. sea cows) live off the coast of Florida and unfortunately, many manatees are killed or injured by powerboats every year. The file *manatee.csv* contains data on X = the number of Florida powerboat registrations (in 1000s) and Y = number of manatees killed near Florida. There is one point for each year from 1977 to 1990.

(a)

Use software to construct a scatterplot between the number of manates killed and number of Florida power boat registrations. Is the relationship roughly linear?

- **Answer:**

(b)

Compute the correlation coefficient between two variables. Describe what this value says about the relationship between X and Y, in the context of the study.

- **Answer:**

(c)

Give the equation of the least-squares regression line for predicting the number of manatee deaths as a function of powerboat registrations.

- **Answer:**

(d)

Predict the number of manatee deaths for a year in which 600,000 powerboats are registered in Florida (X=600).

- **Answer:**

(e)

Provide a 95% prediction interval for the number of manatee deaths in a year in which 600,000 powerboats are registered in Florida.

- **Answer:**

(f)

Examine the residual plot and the normal probability plot of the residuals from the simple linear regression model. is there an indication that the assumptions of the model are violated? Explain why or why not.

- **Answer:**