GLIB DOLOTOV (U1473437 1) MA214 SPRG. 2019

Problem Set 7 Due Monday 3/18

THESE TWO PROBLEMS WILL BE GRADED FOR ACCURACY LIKE A TYPICAL HOMEWORK ASSIGNMENT

BOILERS 12.40 **Boiler drum production.** In a production facility, an accurate estimate of hours needed to complete a task is crucial to management in making such decisions as hiring the proper number of workers, quoting an accurate deadline for a client, or performing cost analyses regarding budgets. A manufacturer of boiler drums wants to use regression to predict the number of hours needed to erect the drums in future projects. To accomplish this task, data on 36 boilers were collected. In addition to hours (y), the variables measured were boiler capacity (x_1=lb/hr), boiler design pressure (x_2=poundsper square inch, or psi), boiler type (x_3=1 if industry field erected, 0 if utility field erected), and drum type (x_4=1 if steam, 0 if mud). The data are saved in the **BOILERS** file.

$x_3 = 1$, $x_4 = 0$

a. Fit the model $E(y)=\beta_0+\beta_1x1+\beta_2x_2+\beta_3x_3+\beta_4x_4$ to the data and give the prediction equation.
b. Conduct a test for the global utility of the model. Use α=.01.
c. Find a 95% confidence interval for $E(y)$ when $x1$=150,000,$x2$=500,$x3$=1,and $x4$=0. Interpret the result.
d. What type of interval would you use if you want to estimate the average number of hours required to erect all industrial mud boilers with a capacity of 150,000 lb/hr and a design pressure of 500 psi?

```
> data <- read.csv("BOILERStxt.csv")
> attach(data)
The following objects are masked from data (pos = 3):

    Boiler, Capacity, Drum, ManHours, Pressure

> model <- lm(ManHours ~ Capacity + Pressure + Boiler + Drum)
> summary(model)

Call:
lm(formula = ManHours ~ Capacity + Pressure + Boiler + Drum)

Residuals:
     Min       1Q   Median       3Q      Max
-1612.66  -549.18   -12.38   406.97  2768.66

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.783e+03  1.205e+03  -3.139 0.003711 **
Capacity     8.749e-03  9.035e-04   9.684 6.86e-11 ***
Pressure     1.926e+00  6.489e-01   2.969 0.005723 **
Boiler       3.444e+03  9.117e+02   3.778 0.000675 ***
Drum         2.093e+03  3.056e+02   6.849 1.12e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 894.6 on 31 degrees of freedom
Multiple R-squared:  0.903,     Adjusted R-squared:  0.8904
F-statistic: 72.11 on 4 and 31 DF,  p-value: 2.977e-15
```

(a)

$$E(y) = -3.783\times10^{3} + 8.749\times10^{-3}\cdot x_1 + 1.926\cdot x_2 + 3.444\times10^{3}\cdot x_3 + 2.093\times10^{3}\cdot x_4$$

(c) 95% CI

```
> predict(model,
+         newdata = data.frame(
+           Capacity = 150000,
+           Pressure = 500,
+           Boiler = 1,
+           Drum = 0),
+         interval = "confidence",
+         level = 0.95)
       fit     lwr      upr
1 1936.412 1448.65 2424.174
```

(1448.65, 2424.174)

(b)

$$\text{Test statistic: } F_c = \frac{(SS_{yy} - SSE)/k}{SSE/[n-(k+1)]} = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} = \frac{\text{Mean square (Model)}}{\text{Mean square (Error)}}$$

where n is the sample size and k is the number of terms in the model.
Rejection region: $F_c > F_\alpha$
p-value: $P(F > F_c)$
where the F-distribution has k numerator degrees of freedom and $[n-(k+1)]$ denominator degrees of freedom.

$F_\alpha = F(p=0.01,\ df1 = k = 4,\ df2 = n-(k+1) = 36-5 = 31)$

$= 3.993 < F_c = 72.11$.

$\Rightarrow$ reject null ($\beta_1=\beta_2=\beta_3=\beta_4=0$) accept Ha, at least one $\beta_i \neq 0$

```
> anova(model)
Analysis of Variance Table

Response: ManHours
          Df    Sum Sq   Mean Sq  F value    Pr(>F)
Capacity   1 175007141 175007141 218.6729 1.369e-15 ***
Pressure   1    490357    490357   0.6127    0.4397
Boiler     1  17813091  17813091  22.2576 4.815e-05 ***
Drum       1  37544266  37544266  46.9119 1.124e-07 ***
Residuals 31  24809761    800315
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$SS_{yy} - SSE$
$= 175007141 + 490357 + 17813091 + 37544266$
$SSE/[n-(k+1)] = 800315$
$\frac{(SS_{yy}-SSE)/k}{SSE/[n-(k+1)]} = 72.11 = F_c$

(d) We use a Confidence interval for estimates of the mean.

12.50 **Role of retailer interest in shopping behavior.** Retail interest is defined by marketers as the level of interest a consumer has in a given retail store. Marketing professors at the University of Tennessee at Chattanooga and the University of Alabama investigated the role of retailer interest in consumers' shopping behavior (*Journal of Retailing*, Summer 2006). Using survey data collected on n=375 consumers, the professors developed an interaction model for y=willingness of the consumer to shop at a retailer's store in the future (called "repatronage intentions") as a function of $x1$=consumer satisfaction and $x2$=retailerinterest. The regression results are shown below.

Variable	$\hat{\beta}$	t-Value	p-Value
Satisfaction (x_1)	.426	7.33	$<.01$
Interest (x_2)	.044	0.85	$>.10$
Satisfaction $\times$ Interest (x_1x_2)	$-.157$	-3.09	$<.01$
$R^2 = .65, F = 226.35, p\text{-value} < .001$			

$n = 375$

1. Is the overall model statistically useful in predicting y? Test, using α=.05.
2. Conduct a test for interaction at α=.05.
3. Use the β-estimates to sketch the estimated relationship between repatronage intentions (y) and satisfaction ($x1$)when retailer interest is $x2$=1(a low value).
4. Repeat part **c** for the case when retailer interest is $x2$=7 (a high value).
5. Put the two lines you sketched in parts **c** and **d** on the same graph to illustrate the nature of the interaction.

1. $F_\alpha = F(p = 0.05, df1 = 3, df2 = 371) = 2.629$

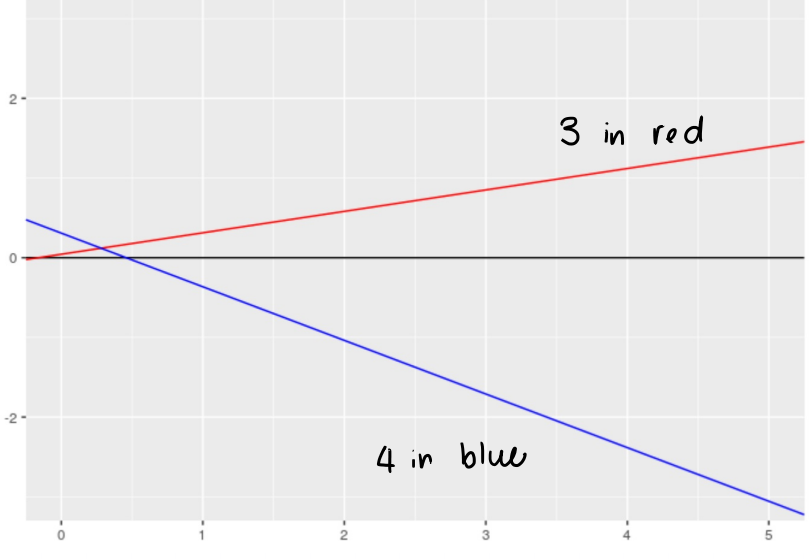
$F_c = 226.35 > F_\alpha = 2.629 \Rightarrow$ model is statistically useful.

2. The p-value on (x_1x_2) is $< 0.01 \Rightarrow$ we will reject the null hypothesis $\beta_3 = 0$ at any confidence level over 0.01. $\alpha = 0.05$
$\Rightarrow$ for such a test we reject the null and accept $\beta_3 \neq 0$.

3. $y = 0.426 \cdot x_1 + 0.044 \cdot x_2 - 0.157 x_1 \cdot x_2$; $x_2 = 1 \Rightarrow$
$y = 0.426 x_1 + 0.044(1) - 0.157 x_1 \cdot (1)$
$= 0.269 x_1 + 0.044$

4. $y = 0.426 \cdot x_1 + 0.044 \cdot x_2 - 0.157 x_1 \cdot x_2$;
$= 0.426 x_1 + 0.044(7) - 0.157 x_1 \cdot (7)$
$= -0.673 \cdot x_1 + 0.308$

```
> ggplot() +
+   geom_abline(slope = 0) +
+   scale_y_continuous(limits = c(-3,3)) +
+   scale_x_continuous(limits = c(0,5)) +
+   geom_abline(slope = 0.269, intercept = 0.044, color = "red") +
+   geom_abline(slope = -0.673, intercept = 0.308, color = "blue")
```



← (5)

DO AT LEAST TWO PROBLEMS FROM EACH CHAPTER. THESE PROBLEMS WILL BE GRADED ON COMPLETION FOR 5 POINTS EACH. YOU MAY DO MORE PROLEMS FOR MORE POINTS. THE ENTIRE ASSIGNMENT (INCLUDING THE TWO ABOVE) IS GRADED OUT OF 60 POINTS SO IT'S POSSIBLE TO GET MORE THAN 100%

9.126 CRASH: NHTSA new car crash test

9.127 EVOS: Oil spill impact on seabirds

9.140 TWINS: Identical twins reared apart

9.150 MS: MS and exercise

9.151 SMWT: Self-managed work team and family life

11.23 MOON: Measuring the moon's orbit

11.26 ANTS: Mongolian desert ants

11.122 PONG: Impact of dropping ping-pong balls

11.130

11.142 RAIN: New method of estimating rainfall

11.151 FLOUR: Regression through the origin

9.126 NHTSA new car crash tests. Refer to the National Highway Traffic Safety Administration (NHTSA) crash test data on new cars, saved in the **CRASH** file. Crash test dummies were placed in the driver's seat and front passenger's seat of a new car model, and the car was steered by remote control into a head-on collision with a fixed barrier while traveling at 35 miles per hour. Two of the variables measured for each of the 98 new cars in the data set are (1) the severity of the driver's chest injury and (2) the severity of the passenger's chest injury. (The more points assigned to the chest injury rating, the more severe the injury is.) Suppose the NHTSA wants to determine whether the true mean driver chest injury rating exceeds the true mean passenger chest injury rating and, if so, by how much.

a. State the parameter of interest to the NHTSA.
b. Explain why the data should be analyzed as matched pairs.
c. Find a 99% confidence interval for the true difference between the mean chest injury ratings of drivers and front-seat passengers.
d. Interpret the interval you found in part **c.** Does the true mean driver chest injury rating exceed the true mean passenger chest injury rating? If so, by how much?
e. What conditions are required for the analysis to be valid? Do these conditions hold for these data?

```
> #3) 9.126
> data <- read.csv("CRASHtxt.csv")
> attach(data)
> DIF <- DRIVCHST - PASSCHST
> setNames(c(length(DIF),mean(DIF),sd(DIF)),
+          c("n","xbar.diff","sd.diff"))
         n  xbar.diff    sd.diff
98.0000000 -0.5612245  5.5167343
> qnorm(0.995)
[1] 2.575829
```

a. parameter of interest: mean of population differences

$\mu_d = \mu_1 - \mu_2$

b. Data should be analyzed in matched pairs because the samples for each parameter WERE NOT collected independently.

c. **Confidence Interval**

$\bar{x}_d \pm z_{\alpha/2}\left(\frac{s_d}{\sqrt{n_d}}\right)$

$= -0.561 \pm (2.576) \cdot \frac{5.517}{\sqrt{98}}$

$= -0.561 \pm 1.436 = (-1.997, 0.875)$

d. We do not have enough evidence to conclude that the true mean driver chest injury rating exceeds the true mean passenger chest injury rating.

e. RANDOM SAMPLE; $n_d \geq 30$.

Conditions Required for Valid Large-Sample Inferences about μ_d

1. A random sample of differences is selected from the target population of differences.
2. The sample size n_d is large (i.e., $n_d \geq 30$). (By the Central Limit Theorem, this condition guarantees that the test statistic will be approximately normal, regardless of the shape of the underlying probability distribution of the population.)

9.127 Oil spill impact on seabirds. Refer to the *Journal of Agricultural, Biological, and Environmental Statistics* (Sept. 2000) study of the impact of a tanker oil spill on the seabird population in Alaska, presented in Exercise 2.205 (p. 111). Recall that for each of 96 shoreline locations (called transects), the number of seabirds found, the length (in kilometers) of the transect, and whether the transect was in an oiled area were recorded. (The data are saved in the **EVOS** file.) *Observed seabird density* is defined as the observed count divided by the length of the transect. A comparison of the mean densities of oiled and unoiled transects is displayed in the MINITAB printout on the next page. Use this information to make an inference about the difference in the population mean seabird densities of oiled and unoiled transects.

We do not have enough evidence to suggest a relationship between the population mean seabird densities of transects and the oiled/unoiled status of the transect.

MINITAB output for Exercise 9.127

Two-Sample T-Test and CI: Density, Oil

```
Two-sample T for Density

Oil   N  Mean  StDev  SE Mean
no   36  3.27   6.70      1.1
yes  60  3.50   5.97     0.77


Difference = mu (no ) - mu (yes)
Estimate for difference:  -0.221165
95% CI for difference:  (-2.927767, 2.485436)
T-Test of difference = 0 (vs not =): T-Value = -0.16  P-Value = 0.871  DF = 67
```

87.1% probability that NO relationship exists between the oiled/unoiled status and population mean seabird densities at transects.

9.150 MS and exercise study. A study published in *Clinical Kinesiology* (Spring 1995) was designed to examine the metabolic and cardiopulmonary responses during exercise of persons diagnosed with multiple sclerosis (MS). Leg-cycling and arm-cranking exercises were performed by 10 MS patients and 10 healthy (non-MS) control subjects. Each member of the control group was selected on the basis of gender, age, height, and weight to match (as closely as possible) with one member of the MS group. Consequently, the researchers compared the MS and non-MS groups by matched-pairs *t*-tests on such outcome variables as oxygen uptake, carbon dioxide output, and peak aerobic power. The data on the matching variables used in the experiment are shown in the table on the next page. Have the researchers successfully matched the MS and non-MS subjects?

D
MS

Data for Exercise 9.150

	MS Subjects				Non-MS Subjects			
Matched Pair	Gender	Age (years)	Height (cm)	Weight (kg)	Gender	Age (years)	Height (cm)	Weight (kg)
1	M	48	171.0	80.8	M	45	173.0	76.3
2	F	34	158.5	75.0	F	34	158.0	75.6
3	F	34	167.6	55.5	F	34	164.5	57.7
4	M	38	167.0	71.3	M	34	161.3	70.0
5	M	45	182.5	90.9	M	39	179.0	96.0
6	F	42	166.0	72.4	F	42	167.0	77.8
7	M	32	172.0	70.5	M	34	165.8	74.7
8	F	35	166.5	55.3	F	43	165.1	71.4
9	F	33	166.5	57.9	F	31	170.1	60.4
10	F	46	175.0	79.9	F	43	175.0	77.9

From "Maximal aerobic exercise of individuals with multiple sclerosis using three modes of ergometry." *Clinical Kinesiology,* Vol. 49, No. 1, Spring 1995, p. 7. Reprinted with permission from W. Jeffrey Armstrong.

9.151 Self-managed work teams and family life. To improve quality, productivity, and timeliness, more and more American industries are utilizing self-managed work teams (SMWTs). A team typically consists of 5 to 15 workers who are collectively responsible for making decisions and performing all tasks related to a particular project. Researchers L. Stanley-Stevens (Tarleton State University), D. E. Yeatts, and R. R. Seward (both from the University of North Texas) investigated the connection between SMWTs, work characteristics, and workers' perceptions of positive spillover into family life (*Quality Management Journal*, Summer 1995). Survey data were collected from 114 AT&T employees who worked on 1 of 15 SMWTs at an AT&T technical division. The workers were divided into two groups: (1) those who reported a positive spillover of work skills to family life and (2) those who did not report any such positive work spillover. The two groups were compared on a variety of job and demographic characteristics, several of which are shown in the table (next column). All but the demographic characteristics were measured on a seven-point scale, ranging from 1 = "strongly disagree" to 7 = "strongly agree"; thus, the larger the number, the more the characteristic was indicated. The file named **SMWT** includes the values of the variables listed in the table for each of the 114 survey participants. The researchers' objectives were to compare the two groups of workers on each characteristic. In particular, they wanted to know which job-related characteristics are most highly associated with positive work spillover. Conduct a complete analysis of the data for the researchers.

Characteristic	Variable
Information Flow	Use of creative ideas (seven-point scale)
Information Flow	Utilization of information (seven-point scale)
Decision Making	Participation in decisions regarding personnel matters (seven-point scale)
Job	Good use of skills (seven-point scale)
Job	Task identity (seven-point scale)
Demographic	Age (years)
Demographic	Education (years)
Demographic	Gender (male or female)
Comparison	Group (positive spillover or no spillover)

11.23 Measuring the moon's orbit. A handheld digital camera was used to photograph the moon's orbit and the results summarized in the *American Journal of Physics* (Apr. 2014). The pictures were used to measure the angular size (in pixels) of the moon at various distances (heights) above the horizon (measured in degrees). The data for 13 different heights are illustrated in the graph below and saved in the **MOON** file.

MOON

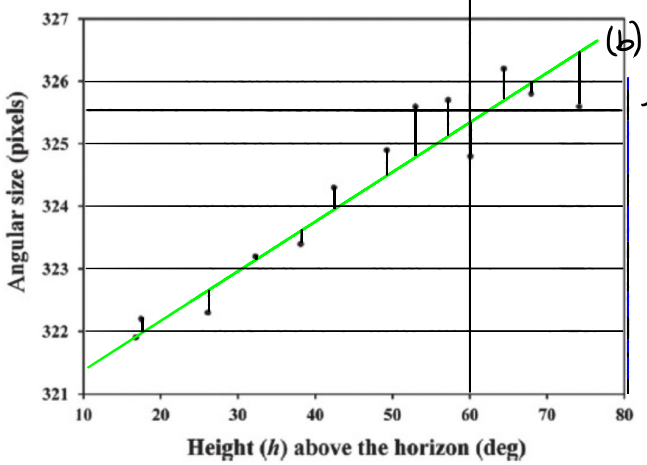
a. Is there visual evidence of a linear trend between angular size (y) and height above horizon (x)? If so, is the trend positive or negative? Explain.

b. Draw what you believe is the best-fitting line through the data.

c. Draw vertical lines from the actual data points to the line, part **b**. Measure these deviations and then compute the sum of squared deviations for the visually fitted line.

d. An SAS simple linear regression printout for the data is shown in the next column. Compare the y-intercept and slope of the regression line to the visually fitted line, part **b**.

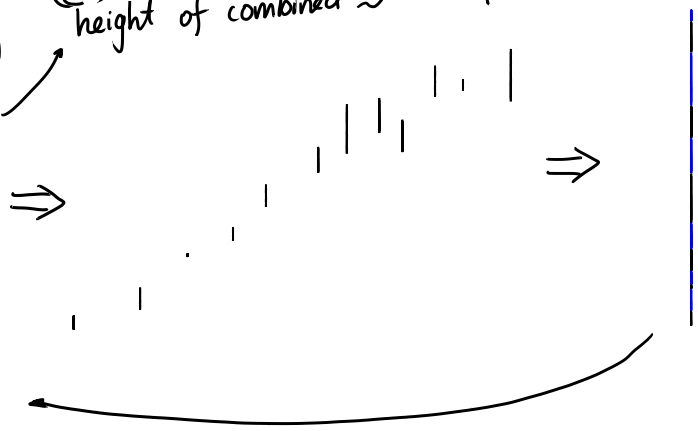
e. Locate SSE on the printout. Compare this value to the result in part **c**. Which value is smaller?



(a) There is visual evidence of a positive-trend (direct) relationship between the moons height above the horizon and the angular size. As seen in the scatterplot, as we move along the x-axis, the points tend to rise on the y-axis.

(b)

(c) height of combined ≈ 5.1 pixels



SAS Output for Exercise 11.23

The REG Procedure
Model: MODEL1
Dependent Variable: ANGLE

Number of Observations Read	13
Number of Observations Used	13

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	28.64766	28.64766	88.40	<.0001
Error	11	3.56465	0.32406		
Corrected Total	12	32.21231			

Root MSE	0.56926	R-Square	0.8893
Dependent Mean	324.44615	Adj R-Sq	0.8793
Coeff Var	0.17546		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	320.63639	0.43487	737.32	<.0001
HEIGHT	1	0.08338	0.00887	9.40	<.0001

(d) $y = 0.08338x + 320.63$

$\Rightarrow y(x=10) = 0.8338 + 320.63$

≈ 321.46

↳ my intercept @ $x=10$

≈ 321.5.

very close!

my slope: $\frac{321.5 - 325.3}{10 - 60} = 0.076$

(e)

	Slope	Intercept	SSE
visual	0.076	321.5	5.1
fitted	0.083	321.5	3.6

↳ SSE of the SAS-fitted model is smaller!

11.26 Mongolian desert ants. Refer to the *Journal of Biogeography* (Dec. 2003) study of ants in Mongolia, presented in Exercise 2.167 (p. 97). Data on annual rainfall, maximum daily temperature, and number of ant species recorded at each of 11 study sites are listed in the table.

D ANTS

Site	Region	Annual Rainfall (mm)	Max. Daily Temp. (°C)	Number of Ant Species
1	Dry Steppe	196	5.7	3
2	Dry Steppe	196	5.7	3
3	Dry Steppe	179	7.0	52
4	Dry Steppe	197	8.0	7
5	Dry Steppe	149	8.5	5
6	Gobi Desert	112	10.7	49
7	Gobi Desert	125	11.4	5
8	Gobi Desert	99	10.9	4
9	Gobi Desert	125	11.4	4
10	Gobi Desert	84	11.4	5
11	Gobi Desert	115	11.4	4

a. Consider a straight-line model relating annual rainfall (y) and maximum daily temperature (x). A MINITAB printout of the simple linear regression is shown below. Give the least squares prediction equation.

Regression Analysis: Rain versus Temp

```
The regression equation is
Rain = 295 - 16.4 Temp

Predictor      Coef  SE Coef       T      P
Constant     295.25    22.41   13.18  0.000
Temp        -16.364    2.346   -6.97  0.000

S = 17.5111   R-Sq = 84.4%   R-Sq(adj) = 82.7%

Analysis of Variance
Source           DF     SS     MS      F      P
Regression        1  14915  14915  48.64  0.000
Residual Error    9   2760    307
Total            10  17675
```

b. Construct a scatterplot for the analysis you performed in part **a.** Include the least squares line on the plot. Does the line appear to be a good predictor of annual rainfall?

c. Now consider a straight-line model relating number of ant species (y) to annual rainfall (x). On the basis of the MINITAB printout below. repeat parts **a** and **b.**

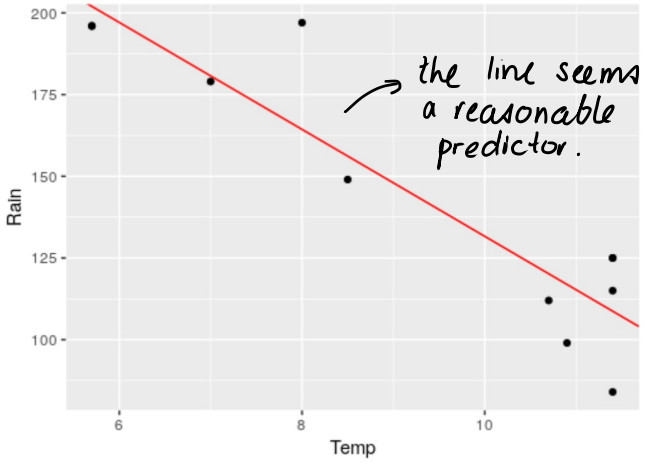
Regression Analysis: AntSpecies versus Rain

```
The regression equation is
AntSpecies = 10.5 + 0.016 Rain

Predictor     Coef  SE Coef      T      P
Constant     10.52    22.03   0.48  0.644
Rain        0.0160   0.1480   0.11  0.916

S = 19.6726   R-Sq = 0.1%   R-Sq(adj) = 0.0%

Analysis of Variance
Source           DF     SS     MS     F      P
Regression        1    4.5    4.5  0.01  0.916
Residual Error    9 3483.1  387.0
Total            10 3487.6
```

(a) $y = 295.25 - 16.364x$

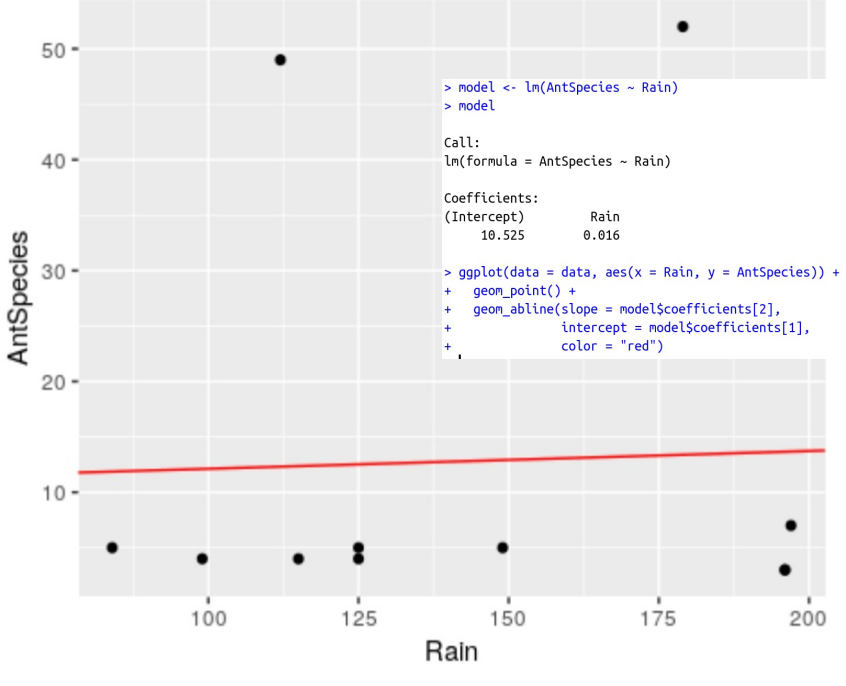
(b)

```
> # 6) 11.26
> data <- read.csv("ANTStxt(1).csv")
> attach(data)
The following objects are masked from data (pos = 3):

    AntSpecies, Diversity, PlantCov, Rain, Region, Site, Temp

> model <- lm(Rain ~ Temp)
> model

Call:
lm(formula = Rain ~ Temp)

Coefficients:
(Intercept)         Temp
     295.25       -16.36

> library(ggplot2)
> ggplot(data = data, aes(x = Temp, y = Rain)) +
+   geom_point() +
+   geom_abline(slope = model$coefficients[2],
+               intercept = model$coefficients[1],
+               color = "red")
```



(c) $y = 10.52 + 0.016x$

```
> model <- lm(AntSpecies ~ Rain)
> model

Call:
lm(formula = AntSpecies ~ Rain)

Coefficients:
(Intercept)         Rain
     10.525        0.016

> ggplot(data = data, aes(x = Rain, y = AntSpecies)) +
+   geom_point() +
+   geom_abline(slope = model$coefficients[2],
+               intercept = model$coefficients[1],
+               color = "red")
```



11.122

Applying the Concepts—Intermediate

11.122 Impact of dropping ping-pong balls. The impact of dropping hollow balls was investigated in the *American Journal of Physics* (Mar. 2014). Standard ping-pong balls were dropped vertically onto a force plate. Upon impact, two variables were measured: y = coefficient of restitution, COR (measured as a ratio of the speed at impact and rebound speed) and x = speed at impact (meters/second). Of the 19 balls dropped, 10 buckled at impact. The data (simulated from information provided in the article) are listed in the table.

D PONG

a. Conduct a complete simple linear regression analysis of the relationship between coefficient of restitution (y) and impact speed (x). Write all your conclusions in the words of the problem.

b. The researcher believes that the rate of increase in the coefficient of restitution with impact speed differs depending on whether the ping-pong ball buckles. Do the data support this hypothesis? Explain.

Ball	COR y	Speed x	Buckle
1	.945	0.8	No
2	.950	1.0	No
3	.930	1.5	No
4	.920	1.8	No
5	.920	3.0	No
6	.930	3.4	No
7	.905	4.4	No
8	.915	5.0	No
9	.910	6.4	No
10	.900	4.4	Yes
11	.885	5.3	Yes
12	.870	5.4	Yes
13	.850	7.4	Yes
14	.795	7.2	Yes
15	.790	7.2	Yes
16	.800	8.0	Yes
17	.820	8.5	Yes
18	.810	9.4	Yes
19	.780	9.0	Yes

11. 130

11.130 In fitting a least squares line to $n = 15$ data points, the following quantities were computed: $SS_{xx} = 55$, $SS_{yy} = 198$, $SS_{xy} = -88$, $\bar{x} = 1.3$, and $\bar{y} = 35$.

a. Find the least squares line.

b. Graph the least squares line.

c. Calculate SSE.

d. Calculate s^2.

e. Find a 90% confidence interval for β_1. Interpret this estimate.

f. Find a 90% confidence interval for the mean value of y when $x = 15$.

g. Find a 90% prediction interval for y when $x = 15$.

11.142

Applying the Concepts—Intermediate

11.142 New method of estimating rainfall. Accurate measurements of rainfall are critical for many hydrological and meteorological projects. Two standard methods of monitoring rainfall use rain gauges and weather radar. Both, however, can be contaminated by human and environmental interference. In the *Journal of Data Science* (Apr. 2004), researchers employed artificial neural networks (i.e., computer-based mathematical models) to estimate rainfall at a meteorological station in Montreal. Rainfall estimates were made every 5 minutes over a 70-minute period by each of the three methods. The data (in millimeters) are listed in the table.

D
RAIN

Time	Radar	Rain Gauge	Neural Network
8:00 a.m.	3.6	0	1.8
8:05	2.0	1.2	1.8
8:10	1.1	1.2	1.4
8:15	1.3	1.3	1.9
8:20	1.8	1.4	1.7
8:25	2.1	1.4	1.5
8:30	3.2	2.0	2.1
8:35	2.7	2.1	1.0
8:40	2.5	2.5	2.6
8:45	3.5	2.9	2.6
8:50	3.9	4.0	4.0
8:55	3.5	4.9	3.4
9:00 a.m.	6.5	6.2	6.2
9:05	7.3	6.6	7.5
9:10	6.4	7.8	7.2

Based on Hessami, M., et al. "Selection of an artificial neural network model for the post-calibration of weather radar rainfall estimation." *Journal of Data Science*, Vol. 2, No. 2, Apr. 2004. (Adapted from Figures 2 and 4.)

a. Propose a straight-line model relating rain gauge amount (y) to weather radar rain estimate (x).
b. Use the method of least squares to fit the model.
c. Graph the least squares line on a scatterplot of the data. Is there visual evidence of a relationship between the two variables? Is the relationship positive or negative?
d. Interpret the estimates of the y-intercept and slope in the words of the problem.
e. Find and interpret the value of s for this regression.
f. Test whether y is linearly related to x. Use $\alpha = .01$.
g. Construct a 99% confidence interval for β_1. Interpret the result practically.
h. Now consider a model relating rain gauge amount (y) to the artificial neural network rain estimate (x). Repeat parts **a–g** for this model.

11.151

Applying the Concepts—Advanced

11.151 Regression through the origin. Sometimes it is known from theoretical considerations that the straight [illegible] between two variables x and y passes thro[illegible] the xy-plane. Consider the relationship b[illegible] weight y of a shipment of 50-pound bags [illegible] number x of bags in the shipment. Since a s[illegible] ing $x = 0$ bags (i.e., no shipment at all) has $y = 0$, a straight-line model of the relatic[illegible] and y should pass through the point $x =$ [illegible] a case, you could assume that $\beta_0 = 0$ and relationship between x and y with the follc[illegible]

$$y = \beta_1 x + \varepsilon$$

The least squares estimate of β_1 for this r[illegible]

$$\hat{\beta}_1 = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

From the records of past flour shipmen[illegible] were randomly chosen and the data shov[illegible] ing table were recorded.

Weight of Shipment	Number of 50-Pound Bags in Shipment
5,050	100
10,249	205
20,000	450
7,420	150
24,685	500
10,206	200
7,325	150
4,958	100
7,162	150
24,000	500
4,900	100
14,501	300
28,000	600
17,002	400
16,100	400

a. Find the least squares line for the given data under the assumption that $\beta_0 = 0$. Plot the least squares line on a scatterplot of the data.

b. Find the least squares line for the given data, using the model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

(i.e., do not restrict β_0 to equal 0). Plot this line on the same scatterplot you constructed in part **a**.

c. Refer to part **b**. Why might $\hat{\beta}_0$ be different from 0 even though the true value of β_0 is known to be 0?

d. The estimated standard error of $\hat{\beta}_0$ is equal to

$$s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}}$$

Use the t-statistic

$$t = \frac{\hat{\beta}_0 - 0}{s\sqrt{(1/n) + (\bar{x}^2/SS_{xx})}}$$

to test the null hypothesis $H_0: \beta_0 = 0$ against the alternative $H_a: \beta_0 \neq 0$. Take $\alpha = .10$. Should you include β_0 in your model?