Solutions to Homework 11
Statistics 302 Professor Larget

Textbook Exercises

8.47 Body Mass Gain (Graded for Completeness) Computer output showing body mass gain (in grams) for the mice after four weeks in each of the three light conditions is shown in the book, along with relevant ANOVA output. Which light conditions give significantly different mean body mass gain?

Solution
We have three pairs to test. We first test $H_0 : \mu_{DM} = \mu_{LD}$ vs $H_a : \mu_{DM} \neq \mu_{LD}$. The test statistic is

$$t = \frac{\bar{x}_{DM} - \bar{x}_{LD}}{\sqrt{MSE\left(\frac{1}{n_{DM}} + \frac{1}{n_{LD}}\right)}} = \frac{7.859 - 5.987}{\sqrt{6.48\left(\frac{1}{10} + \frac{1}{9}\right)}} = 1.60.$$

This is a two-tail test so the p-value is twice the area above 1.60 in a t-distribution with $df = 25$. We see that the p-value is $2(0.061) = 0.122$. We don?t find convincing evidence for a difference in mean weight gain between the dim light condition and the light/dark condition.

We next test $H_0 : \mu_{DM} = \mu_{LL}$ vs $H_a : \mu_{DM} \neq \mu_{LL}$. The test statistic is

$$t = \frac{\bar{x}_{DM} - \bar{x}_{LL}}{\sqrt{MSD\left(\frac{1}{n_{DM}} + \frac{1}{n_{LL}}\right)}} = \frac{7.859 - 11.010}{\sqrt{6.48\left(\frac{1}{10} + \frac{1}{9}\right)}} = -2.69.$$

This is a two-tail test so the p-value is twice the area below -2.69 in a t-distribution with $df = 25$. We see that the p-value is $2(0.0063) = 0.0126$. At a 5% level, we do find a difference in mean weight gain between the dim light condition and the bright light condition, with higher mean weight gain in the bright light condition.

Finally, we test $H_0 : \mu_{LD} = \mu_{LL}$ vs $H_a : \mu_{LD} \neq \mu_{LL}$. The test statistic is

$$t = \frac{\bar{x}_{LD} - \bar{x}_{LL}}{\sqrt{MSE\left(\frac{1}{n_{LD}} + \frac{1}{n_{LL}}\right)}} = \frac{5.987 - 11.010}{\sqrt{6.48\left(\frac{1}{9} + \frac{1}{9}\right)}} = -4.19.$$

This is a two-tail test so the p-value is twice the area below -4.19 in a t-distribution with $df = 25$. We see that the p-value is $2(0.00015) = 0.0003$. There is strong evidence of a difference in mean weight gain between the light/dark condition and the bright light condition, with higher mean weight gain in the bright light condition.

8.48 When Calories Are Consumed (Graded for Accurateness) Researchers hypothesized that the increased weight gain seen in mice with light at night might be caused when the mice are eating. Computer output for the percentage of food consumed during the day (when mice would normally be sleeping) for each of the three light conditions is shown in the book, along with relevant ANOVA output. Which light conditions give significantly different mean percentage of calories consumed during the day?

Solution
We have three pairs to test. We first test $H_0 : \mu_{DM} = \mu_{LD}$ vs $H_a : \mu_{DM} \neq \mu_{LD}$. The test statistic is

$$t = \frac{\bar{x}_{DM} - \bar{x}_{LD}}{\sqrt{MSE\left(\frac{1}{n_{DM}} + \frac{1}{n_{LD}}\right)}} = \frac{55.516 - 36.485}{\sqrt{92.8\left(\frac{1}{10} + \frac{1}{9}\right)}} = 4.30.$$

This is a two-tail test so the p-value is twice the area above 4.30 in a t-distribution with $df = 25$. We see that the p-value is $2(0.0001) = 0.0002$. We find strong evidence of a difference in mean daytime consumption percent

between the dim light condition and the light/dark condition. A higher mean percentage of food is consumed during the day in the dim light condition.

We next test $H_0 : \mu_{DM} = \mu_{LL}$ vs $H_a : \mu_{DM} \neq \mu_{LL}$. The test statistic is

$$t = \frac{\bar{x}_{DM} - \bar{x}_{LL}}{\sqrt{MSE\left(\frac{1}{n_{DM}} + \frac{1}{n_{LL}}\right)}} = \frac{55.516 - 76.573}{\sqrt{92.8\left(\frac{1}{10} + \frac{1}{9}\right)}} = -4.76.$$

This is a two-tail test so the p-value is twice the area below -4.76 in a t-distribution with $df = 25$. We see that the p-value is $2(0.00003) = 0.00006$. We find strong evidence of a difference in mean daytime consumption percent between the dim light condition and the bright light condition. A higher mean percentage of food is consumed during the day in the bright light condition.

Finally, we test $H_0 : \mu_{LD} = \mu_{LL}$ vs $H_a : \mu_{LD} \neq \mu_{LL}$. The test statistic is

$$t = \frac{\bar{x}_{LD} - \bar{x}_{LL}}{\sqrt{MSE\left(\frac{1}{n_{LD}} + \frac{1}{n_{LL}}\right)}} = \frac{36.485 - 76.573}{\sqrt{92.8\left(\frac{1}{9} + \frac{1}{9}\right)}} = -8.83.$$

This is a two-tail test so the p-value is twice the area below -8.83 in a t-distribution with $df = 25$. We see that the p-value is essentially zero, so there is very strong evidence of a difference in mean daytime consumption percent between the light/dark condition and the bright light condition. A higher mean percentage of food is consumed during the day in the bright light condition.

9.18 Does When Food Is Eaten Affect Weight Gain? (Graded for Completeness) Data A.1 on page 136 introduces a study that examines the effect of light at night on weight gain in a sample of 27 mice observed over a four week period. The mice who had a light on at night gained significantly more weight than the mice with darkness at night, despite eating the same number of calories and exercising the same amount. Researchers noticed that the mice with light at night ate a greater percentage of their calories during the day (when mice are suppose to be sleeping). The computer output shown in the book allows us to examine the relationship between percent of calories eaten during the day, *DayPct*, and body mass gain in grams, *BMGain*. A scatterplot with regression line is shown in the book as well.

(a) Use the scatterplot to determine whether we should have any strong concerns about the conditions being met for using a linear model with these data.

(b) What is the correlation between these two variables? What is the p-value from a test of the correlation? What is the conclusion of the test, in context?

(c) What is the least squares line to predict body mass gain from percent daytime consumption? What gain is predicted for a mouse that eats 50% of its calories during the day (*DayPct*=50)?

(d) What is the estimated slope for this regression model? Interpret the slope in context.

(e) What is the p-value for a test of the slope? What is the conclusion of the test, in context?

(f) What is the relationship between the p-value of the correlation test and the p-value of the slope test?

(g) What is R^2 for this linear model? Interpret it in context.

(h) Verify that the correlation squared gives the coefficient of determination R^2.

Solution

(a) On the scatterplot, we have concerns if there is a curved pattern (there isn't) or variability from the line increasing or decreasing in a consistent way (it isn't) or extreme outliers (there aren't any). We do not have any serious concerns about using these data to fit a linear model.

(b) In the output the correlation is $r = 0.740$ and the p-value is 0.000. This small p-value gives strong evidence of a linear relationship between body mass gain and when food is eaten.

(c) From the computer output, the least squares line is $\hat{BMGain} = 1.11 + 0.127 \cdot DayPct$. For a mouse that eats 50% of calories during the day, we have

$$\hat{BMGain} = 1.11 + 0.127(50) = 7.46 \text{ grams}$$

A mouse that eats 50% of its calories during the day is predicted to gain 7.46 grams over a 4-week period.

(d) The estimated slope is $b_1 = 0.127$. For an additional 1% of calories eaten during the day, body mass gain is predicted to go up by 0.127 grams.

(e) For testing $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$ we see $t = 5.50$ and p-value ≈ 0. The percent of calories eaten during the day is an effective predictor of body mass gain.

(f) The p-values for testing the correlation and the slope for these two variables is the same: both are 0.000. In fact, if we calculate the t-statistic for testing the correlation using $r = 0.74$ and $n = 27$ we have

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.74\sqrt{27-2}}{1-(0.74)^2} = 5.50$$

which matches the t-statistic for the slope.

(g) We see that $R^2 = 54.7\%$. This tell us that 54.7% of the variability in body mass gain can be explained by the percent of calories eaten during the day. More than half of the variability in body mass gain can be explained simply by when the calories are eaten.

(h) Using $r = 0.740$, we find that $r^2 = (0.740)^2 = 0.5476$, matching $R^2 = 54.7\%$ up to round-off.

9.21 Using pH in Lakes as a Predictor of Mercury in Fish (Graded for Accurateness) The `FloridaLakes` dataset, introduced in Data 2.4, includes data on 53 lakes in Florida. Two of the variables are recorded are pH (acidity of the lake water) and $AvgMercury$ (average mercury level for a sample of fish from each lake). We wish to use the pH of the lake water (which is easy to measure) to predict average mercury levels in fish, which is harder to measure. A scatter lot of the data is shown in Figure 2.49(a) on page 106 and we see that the conditions for fitting a linear model are reasonably met. Computer output for the regression analysis is shown in the book.

(a) Use the fitted model to predict the average mercury level in fish for a lake with a pH of 6.0.

(b) What is the slop in the model? Interpret the slop in context.

(c) What is the test statistic for a test of the slope? What is the p-value? What is the conclusion of the test, in context?

(d) Compute and interpret a 95% confidence interval for the slope.

(e) What is R^2? Interpret it in context.

Solution

(a) For a pH reading of 6.0 we have

$$\hat{AvgMercury} = 1.53 - 0.152 \cdot pH = 1.53 - 0.152(6) = 0.618$$

The model predicts that fish in lakes with a pH of 6.0 will have an average mercury level of 0.618.

(b) The estimated slope is $b_1 = -0.152$. This means that as pH increases by one unit, predicted average mercury level in fish will go down by 0.152 units.

(c) The test statistic is $t = -5.02$, and the p-value is essentially zero. Since this is a very small p-value we have strong evidence that the pH of a lake is effective as a predictor of mercury levels in fish.

(d) The estimated slope is $b_1 = -0.152$ and the standard error is $SE = 0.03031$. For 95% confidence we use a t-distribution with $53 - 2 = 51$ degrees of freedom to find $t^* = 2.01$. The confidence interval for the slope is

$$\begin{array}{rcl} b_1 & \pm & t^* \cdot SE \\ -0.152 & \pm & 2.01(0.03031) \\ -0.152 & \pm & 0.0609 \\ -0.2129 & \text{to} & -0.0911 \end{array}$$

Based on these data we are 95% sure that the slope (increase in mercury for a one unit increase in pH) is somewhere between -0.213 and -0.091.

(e) We see that R^2 is 33.1%. This tells us that 33.1% of the variability in average mercury levels in fish can be explained by the pH of the lake water that the fish come from.

9.42 Predicting Prices of Printers (Graded for Accurateess) Data 9.1 on page 525 introduces the dataset **InkjetPrinters**, which includes information on all-in-one printers. Two of the variables are *Price* (the price of the printer in dollars) and *CostColor* (average cost per page in cents for printing in color). Computer output for predicting the prince from the cost of printing is shown in the book.

(a) What is the predicted price of a printer that costs 10 cents a page for color printing?

(b) According to the model, does it tend to cost more or less (per page) to do color printing on a cheaper printer?

(c) Use the information in the ANOVA table to determine the number of printers included in the dataset.

(d) Use the information in the ANOVA table to compute and interpret R^2.

(e) Is the linear model effective at predicting the price of a printer? Use information from the computer output and state the conclusion in context.

Solution

(a) We see that the predicted price when $CostColor = 10$ is given by

$$\hat{Price} = 378 - 18.6CostColor = 378 - 18.6(10) = 192$$

The predicted price for a printer where each color page costs about 10 cents to print is \$192.

(b) Since the slope ($b_1 = -18.6$) is negative, the price of a printer goes down as the cost of color printing increases. In other words, cheaper printers cost more to print in color.

(c) Since the total degrees of freedom is $n - 1 = 19$, the sample size is 20.

(d) To calculate R^2, we use

$$R^2 = \frac{SSModel}{SSTotal} = \frac{57604}{136237} = 0.423$$

We see that $R^2 = 42.3\%$, which tells us that 42.3% of the variability in prices of inkjet printers can be explained by the cost to print a page in color.
(e) The hypotheses are H_0 : The model is ineffective vs H_a: The model is effective. We see in the ANOVA table that the F-statistic is 13.19 and the p-value is 0.002. This p-value is quite small so we reject H_0. There is evidence that the linear model to predict price using the cost of color printing is effective.

When Calories Are Consumed and Weight Gain in Mice In Exercise 9.18 on page 535, we look at a model to predict weight gain (in grams) in mice based on the percent of calories the mice eat during the day (when mice should be sleeping instead of eating). In Exercises 9.59 and 9.60, we give computer output with two regression intervals and information about the percent of calories eaten during the day. Interpret each of the intervals in the context of this data situation.

(a) The 95% confidence interval for the mean response

(b) The 95% prediction interval for the mean response

9.59 (Graded for Accurateness) The intervals given in the book are for mice that eat 50% of their calories during the day.

Solution

(a) The 95% confidence interval for the mean response is 6.535 to 8.417. We are 95% confident that for mice that eat 50% of calories during the day, the average weight gain will be between 6.535 grams and 8.417 grams.

(b) The 95% prediction interval for the response is 2.786 to 12.166. We are 95% confident that a mouse that eats 50% of its calories during the day will gain between 2.786 grams and 12.166 grams.

9.60 (Graded for Completeness) The intervals given in the book are for mice that eat 10% of their calories during the day.

Solution

(a) The 95% confidence interval for the mean response is -0.013 to 4.783. We are 95% confident that for mice that eat 10% of calories during the day, the average weight change will be between losing 0.013 grams and gaining 4.783 grams.

(b) The 95% prediction interval for the response is -2.797 to 7.568. We are 95% confident that a mouse that eats 10% of its calories during the day will have a weight change between losing 2.797 grams and gaining 7.568 grams.

Computer Exercises

For each R problem, turn in answers to questions with the written portion of the homework. Send the R code for the problem to Katherine Goode. The answers to questions in the written part should be well written, clear, and organized. The R code should be commented and well formatted.

R problem 1 (Graded for Completeness) In this problem, you will do a simulation exercise to see the distribution of p-values for the t-test for the difference in population means when applied to the samples with the largest and smallest means for a setting where all sample sizes are 8, there are 7 populations (and so 21 possible pairwise comparisons), and the null hypothesis is true with all population means equal to 100 and all population standard deviations equal to 20.

The following function will do the following:

1. Generate random samples.
2. Use `lm()` to fit a linear model.
3. Determine which sample means are largest and smallest.
4. Find the two-sided p-value for the test of the difference of these two population means using

$$SE = \sqrt{MSE} \times \sqrt{\frac{1}{8} + \frac{1}{8}}$$

for the standard error and a t distribution with $n - k = 56 - 7 = 49$ degrees of freedom to find the p-value.

```
sim = function(npop = 7, ni = 8, mu = 100, sigma = 20) {
# Create data
group = factor(rep(LETTERS[1:npop],each=ni))
y = rnorm(n=npop*ni,mean=mu,sd=sigma)
# Find all sample means, and largest and smallest
all.means = as.vector( by(y,group,mean) )
max.mean = max(all.means)
min.mean = min(all.means)
# Fit the linear model (ANOVA)
fit = lm(y ~ group)
# Find the df and sqrt of MSE and the SE
mse = anova(fit)$Mean[2]
df = fit$df.residual
se = sqrt(mse) * sqrt(2/ni)
# Compute and return p-value
t.stat = (max.mean - min.mean) / se
```

```
p.value = 2*pt(-abs(t.stat),df)
return( p.value )
}
```

1. In R, create variables `npop`, `ni mu`, and `sigma` with the specified values. Then, enter each line of the function `sim()` into R separately and explain what it does. Display the result of each object created.

   Solution
   The first line of the code draws `ni` of each of the first `npop` letters. They are placed into a vector called group and told to be treated as factors. Thus, we will have `npop` number of populations with `ni` samples from each population. Thus, we will have a data set set with `npop*ni` total data points.

```
> group = factor(rep(LETTERS[1:npop],each=ni))
> group
[1] A A A A A A A A B B B B B B B B C C C C C C C C D D D D D D D D E E
E E E E E E F F F F F F F F G
[50] G G G G G G G
Levels: A B C D E F G
```

   We next obtain our data by drawing `npop*ni` values from a normal distribution with mean `mu` and standard deviation `sigma`.

```
> y = rnorm(n=npop*ni,mean=mu,sd=sigma)
> y
[1] 127.86120 113.00120  88.92601 107.93783 102.90153 107.48332  55.37059  83.94684  98.67751
[10]  81.51691 122.27146  68.76922  93.95310 109.20888  61.36001  40.55818  87.54020  76.44623
[19]  83.02394  74.35637  85.44577  95.31288  63.26493 101.92742 104.10770 106.43524 130.40960
[28] 113.35964 103.63373  67.86388  81.79665 109.40837  87.29306  92.88517 104.86553  95.12380
[37]  89.19441 119.45884 109.49923 133.07072 103.32258 110.63881  76.00023  92.16459 118.03652
[46]  99.47053 102.26723  31.19028  95.79515 111.08286 113.30926 112.37424 112.68995  89.76025
[55] 116.92439  95.93225
```

   We then take the data and sort them into the `npop` groups and find the means of each group. We do this using the `by` command in R. We tell R to take the values in `y`, divide them into groups based on the `group` vector we created, and then take the mean of each group. We place these values in a vector using the `as.vector` command.

```
> all.means = as.vector( by(y,group,mean) )
> all.means
[1]  98.42856  84.53941  83.41472 102.12685 103.92385  91.63634 105.98354
```

   We now find both the largest and smallest means and call them `max.mean` and `min.mean`.

```
> max.mean = max(all.means)
> max.mean
[1] 105.9835
> min.mean = min(all.means)
> min.mean
[1] 83.41472
```

Now we fit the linear model based on the data, so that we can eventually perform anova on the data.

```
> fit = lm(y ~ group)
> fit
Call:
lm(formula = y ~ group)
Coefficients:
(Intercept)       groupB       groupC       groupD       groupE       groupF       groupG
     98.429      -13.889      -15.014        3.698        5.495       -6.792        7.555
```

We now create an anova table using the `anova` command. However, we are only interested in the mean square error, so we use `$Mean[2]` to tell R to only output the mean square error from the table.

```
> mse = anova(fit)$Mean[2]
> mse
[1] 407.0427
```

Now we obtain the error degrees of freedom from the linear model that we fit.

```
> df = fit$df.residual
> df
[1] 49
```

With the values that we have obtained, we can calculate the standard error.

```
> se = sqrt(mse) * sqrt(2/ni)
> se
[1] 10.08765
```

Now we calculate the test statistic for our test for the difference in population means.

```
> t.stat = (max.mean - min.mean) / se
> t.stat
[1] 2.237273
```

Lastly, we obtain the p-value using the test statistic and the error degrees of freedom and tell R to return at the end of the function.

```
> p.value = 2*pt(-abs(t.stat),df)
> p.value
[1] 0.02984969
> return( p.value )
```

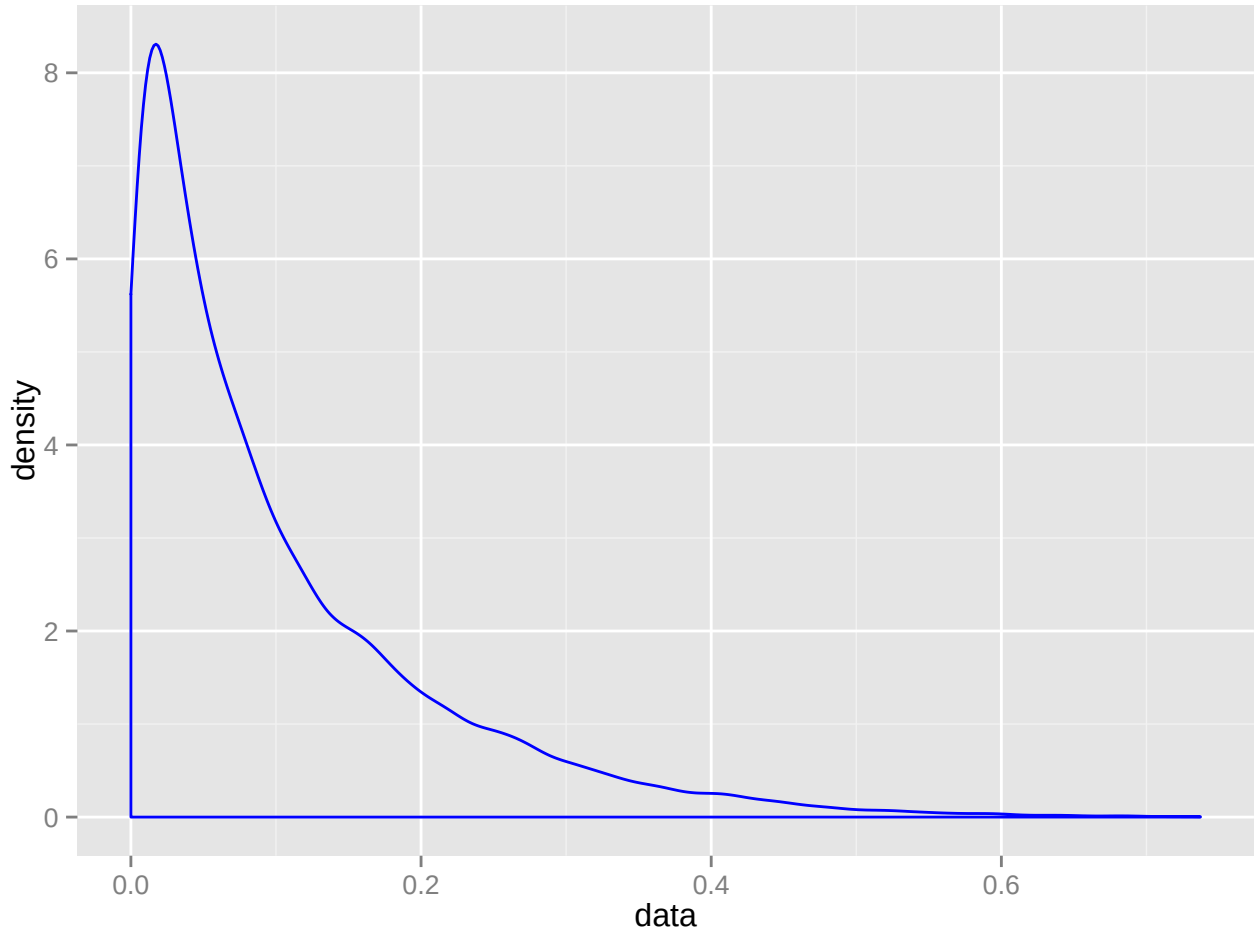
2. Write a `for()` loop to run `sim()` 10,000 times, saving the p-values into a vector. (This will take up to a minute to actually run.)

   Solution
   Below is the code that was used to run the loop to obtain the 10,000 p-values.

```
pvalues <- numeric(10000)
for(i in 1:10000)
{
  pvalues[i] <- sim()
}
```

3. Display the distribution of p-values with `ggplot2` and `geom_density()`. Describe the shape of p-values. Are they approximately uniform from 0 to 1, or is the center of the distribution shifted left or right?

   Solution
   Below is the density plot of the distribution of the p-values. We see that the distribution is skewed right, and it is not uniform from 0 to 1. The center is shifted to the left.



   The following code was used to obtain this plot.

```
library(ggplot2)
ggplot(data.frame(data=pvalues),aes(x=data)) +geom_density(color="blue")
```

4. What fraction of the p-values are less than 0.05?

Solution
We calculate that the faction of p-values that are less than 0.05 is 0.4281. We used the following code to obtain this result.

```
sum(pvalues<0.05)/10000
```

5. Explain what this simulation result means with respect to interpreting p-values from pairwise comparisons following an ANOVA analysis with regard to the issue of multiple testing.

   Solution
   We know that all of the data we used in this simulation came from the same population. Therefore, we would expect that when we tested to see if there was a difference in means, we would get that the result would not be significant. Nevertheless, when we did the simulation and tested this difference 10,000 times, we saw that 42.81% of the time, we would have rejected the null hypothesis that there was no difference between the groups. This is the problem with multiple testing. The more times the test is repeated, the higher the probability of making a type I error becomes. As we see in this simulation, a type one error was made many times.

R Problem 2 (Graded for Accurateness) This problem will teach you to do many steps in a simple linear regression analysis.

1. Load the data from **InkjetPrinters** into R (`library(Lock5Data)`; `data(InkjetPrinters)`.

   Solution
   I loaded the data in using the commands given above.

2. Plot a scatterplot of the data using *CostColor* as the explanatory variable and *Price* as the response variable. Use `ggplot()`.

   Solution

   The code used to obtain this graph is as follows.

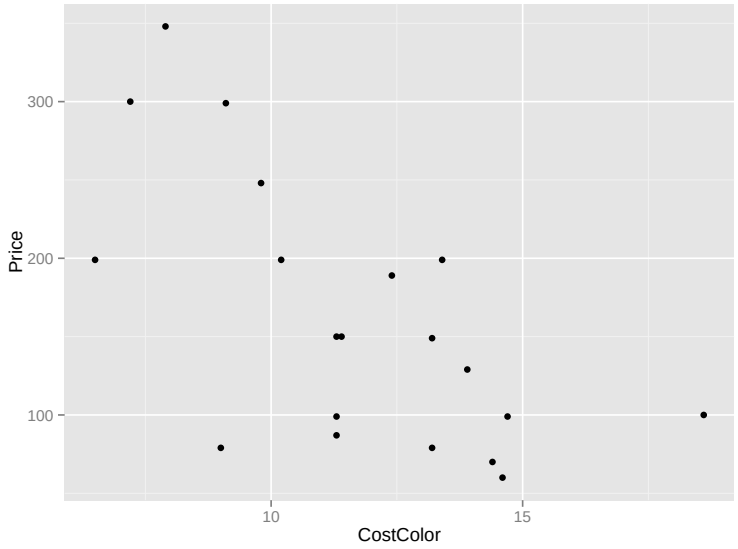
```
ggplot(InkjetPrinters, aes(x=CostColor,y=Price))+geom_point()
```

3. Fit the simple linear regression model. Print a simple summary. Pull data from the summary to write an expression for the regression line and an estimate of σ. Your expression should be like this, but with numbers instead of a and b.

$$(Price) = a + b(CostColor)$$

```
require(Lock5Data)
data(InkjetPrinters)
fit = lm(Price ~ CostColor, data = InkjetPrinters)
summary(fit)
```

   Solution
   Using the commands given above, we obtain the following output from R.




```
Call:
lm(formula = Price ~ CostColor, data = InkjetPrinters)

Residuals:
     Min       1Q   Median       3Q      Max
-132.155  -48.965    1.213   52.629  116.429

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  378.195     61.451   6.154 8.23e-06 ***
CostColor    -18.560      5.111  -3.631  0.00191 **
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 66.09 on 18 degrees of freedom
Multiple R-squared:  0.4228,Adjusted R-squared:  0.3908
F-statistic: 13.19 on 1 and 18 DF,  p-value: 0.00191
```

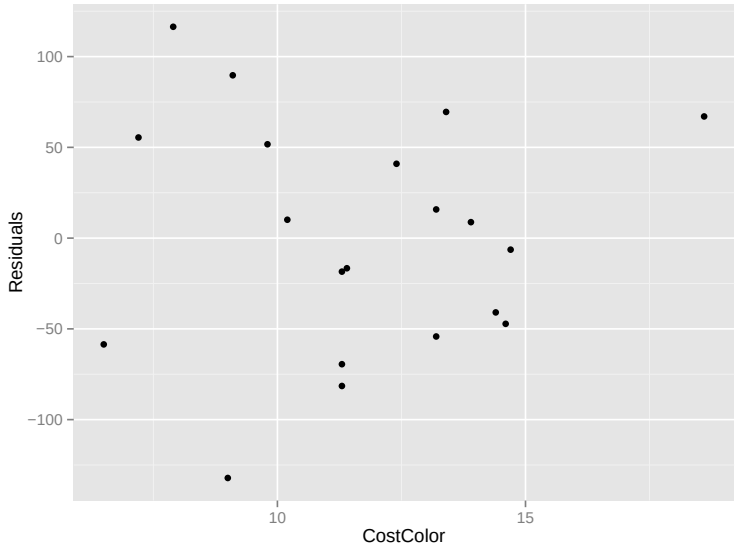
From this output, we find that the slop is -18.560, and y-intercept is 378.195. Thus, our regression line is:

$$(Price) = 378.195 - 18.560(CostColor)$$

4. Make a plot of residuals versus *CostColor*. Are there any patterns to suggest nonlinearity or nonconstant variance as x changes?

```
resid = residuals(fit)
d = data.frame(CostColor = InkjetPrinters$CostColor,Residuals=resid)
```

Solution
Below is the plot of the residuals versus *CostColor*. It appears that the data points become closer to zero as the value of *CostColor* increases. This indicates that the variance is not constant as is should be.



5. Use `predict()` to find a 95% confidence interval for the mean price of all inkjet printers where the cost per page of color printing is 10 cents. Verify that the numerical results match those from the equations on page 553.

```
predict(fit,data.frame(CostColor=10),interval="confidence")
```

Solution
Using the code provided above, we obtain the following output from R.

```
       fit      lwr      upr
1 192.5952 156.7387 228.4517
```

We are 95% confident that the mean price of all inkjet printers where the cost per page of color printing is 10 cents is between \$156.74 and \$228.45.

We now use the equation from the book.

$$\begin{aligned} \hat{y} &\pm t^* s_\epsilon \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}} \\ 192.5952 &\pm 2.100922 \cdot 66.09473 \sqrt{\frac{1}{20} + \frac{(10 - 11.67)^2}{(20-1) \cdot 8.801158}} \\ 156.7387 &\text{ to } 228.4517 \end{aligned}$$

We see that we get the same result when calculating it by hand. The R code that was used to obtain this answer is included below.

```
yhat <- predict(fit,data.frame(CostColor=10),interval="confidence")[1]
df <- 20-2
t <- qt(0.975,df)
se <- sqrt(anova(fit)$Sum[2]/(20-2))
mean <- mean(InkjetPrinters$CostColor)
var <- sd(InkjetPrinters$CostColor)^2
SE <- se*sqrt((1/20)+((10-mean)^2)/((20-1)*var))
yhat+c(-1,1)*t*SE
```

6. Use `predict()` to find a 95% prediction interval for the price of a single inkjet printer where the cost per page of color printing is 10 cents. Verify that the numerical results match those from the equations on page 553.

```
predict(fit,data.frame(CostColor=10),interval="prediction")
```

Solution
Using the code provided above, we obtain the following output from R.

```
       fit      lwr      upr
1 192.5952 49.18058 336.0098
```

We are 95% confident that the price of all inkjet printers where the cost per page of color printing is 10 cents is between \$156.74 and \$228.45.

Using the equation from the book, we get

$$\begin{aligned} \hat{y} &\pm t^* s_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}} \\ 192.5952 &\pm 2.100922 \cdot 66.09473 \sqrt{1 + \frac{1}{20} + \frac{(10 - 11.67)^2}{(20-1) \cdot 8.801158}} \\ 49.18058 &\text{ to } 336.00982 \end{aligned}$$

We see that we get the same result when calculating it by hand. The R code that was used to obtain this answer is included below.

```
SE.pi <- se*sqrt(1+(1/20)+((10-mean)^2)/((20-1)*var))
yhat+c(-1,1)*t*SE.pi
```

7. Briefly explain why the prediction interval is wider than the confidence interval.

Solution
We have that the prediction interval is trying to capture most of the response variables from a population for a particular value of the predictor variable instead of the possible values that a mean could take on. As a result, the prediction interval tends to be larger.