Practice problems for introduction to R with partial answers

1. The purpose of this problem is for you to gain experience with finding quantiles and probabilities for particular named distributions. Most of this problem is taken from a STAT 801 exam that I gave in the past.
	1. Suppose x has a normal distribution with a mean of 6 and a variance of 10. Find the 0.9 quantile from the probability distribution.

P(x < 10.05) = 0.9

> qnorm(p = 0.9, mean = 6, sd = sqrt(10))

[1] 10.05262

Here’s a plot to help demonstrate it:

> dnorm(x = 10.05, mean = 6, sd = sqrt(10)) #f(x) where x = 10.05

[1] 0.05555643

> curve(expr = dnorm(x = x, mean = 6, sd = sqrt(10)), xlim = c(-3, 15), ylab =

 "f(x)", xlab = "x")

> segments(x0 = qnorm(p = 0.9, mean = 6, sd = sqrt(10)), x1 = qnorm(p = 0.9, mean =

 6, sd = sqrt(10)), y0 = 0, y1 = dnorm(x = qnorm(p = 0.9, mean = 6, sd =

 sqrt(10)), mean = 6, sd = sqrt(10)), col = "red") #Line from (10.05, 0) to

 (10.05, 0.0556)

> abline(h = 0) #Horizontal line at 0

> text(x = 5, y = 0.02, labels = "Area \n is 0.9")

> text(x = 11, y = 0.02, labels = "Area \n is 0.1")



* 1. Suppose x has a normal distribution with a mean of 6 and a variance of 10. Find the probability that x is less than 10.05.

> pnorm(q = 10.05, mean = 6, sd = sqrt(10))

[1] 0.8998544

* 1. Suppose x has a chi-square distribution with a degrees of freedom of 10. Find the 0.1 quantile from the probability distribution.

P(x < 4.87) = 0.1

> qchisq(p = 0.1, df = 10)

[1] 4.865182

* 1. Suppose x has a t-distribution with a degrees of freedom of 1000. Find the 0.9 quantile from the probability distribution.

P(x < 1.28) = 0.9

> qt(p = 0.9, df = 1000)

[1] 1.282399

* 1. Explore the named distributions available to you in the stats package. One way to see a list of them is by typing help(Distributions) at an R Console prompt. Try finding quantiles and probabilities for your “favorite” distribution given in the list.
1. Describe what the head() and tail() functions do (examine the help for these functions, if needed). Apply them to a data set (see the upcoming diamond price problem for an example).
2. Install the rgl and mvtnorm packages. What is the purpose of these packages (look at the help documentation to make this determination)? We will use these packages later in the course.
3. Below is part of a project that I gave my STAT 870 (regression analysis) students a few years ago. This project gave a review of methods from STAT 801. I have included additional notes in the text for STAT 873 students as tracked changes.

What factors help determine a diamond’s price? The purpose of this project is to determine which factors and to predict price. Below is a partial listing of a data set that contains the price, carats, color, clarity, and certification body of 308 round cut diamonds sampled from a publication in the year 2000.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Carat** | **Color** | **Clarity** | **Certify** | **Price** |
| 0.3 | D | VS2 | GIA |  $745.92  |
| 0.3 | E | VS1 | GIA |  $865.08  |
| 0.3 | G | VVS1 | GIA |  $865.08  |
| 0.3 | G | VS1 | GIA |  $721.86  |
| 0.31 | D | VS1 | GIA |  $940.13  |
|  |  |  |  |  |
| 1.09 | I | VVS2 | HRD |  $5,217.42  |

For a description of carats, color, and clarity, see the diamond guide included at the end of this project. Certify denotes the certification body which determines the carat, clarity, and color. In this data set, color, clarity, and certification body takes on the following values:

|  |  |
| --- | --- |
| **Variable** | Values |
| Color | D, E, F, G, H, or I |
| Clarity | IF, VVS1, VVS2, VS1, or VS2 |
| Certify | Gemmological Institute of America (GIA), International Gemmological Institute (IGI), or Hoge Raad Voor Diamant (HRD) |

The data is stored in the file DIamondPrices.csv which can be downloaded from the Projects web page of the STAT 870 website (This is now available from the STAT 873 website).

Complete the following problems below. Within each part, include your R program output with code inside of it and any additional information needed to explain your answer. You may need to edit your output and code in order to make it look nice after you copy and paste it into your Word document. Use α = 0.05 for all hypothesis tests, confidence intervals, and prediction intervals.

1. What is the population from which the sample is taken? Describe possible implications of this population with regards to making inferences to round cut diamonds sold in Lincoln jewelry stores currently.

This is a difficult question given the lack of information provided for the project. What I was looking for was some general discussion of what the population may be and how this will affect whether or not inferences can be made regarding diamonds in Lincoln jewelry stores. Most likely, the year the data was collected would cause problems with using the regression model for Lincoln jewelry stores now. Also, one would need to know if the prices in the publication were only for a particular location or if they would be representative of the entire world’s diamond market.

1. Construct a scatter plot for price (y-axis) vs. carat (x-axis). Does there appear to be a relationship between price and carat? Explain.

The data file is located in the data folder of the c drive of my computer. You will need to change this location corresponding to where you saved the file!

> diamonds <- read.csv(file = "c:\\data\\DiamondPrices.csv")

> head(diamonds)

 carat color clarity certify price

1 0.30 D VS2 GIA 745.9184

2 0.30 E VS1 GIA 865.0820

3 0.30 G VVS1 GIA 865.0820

4 0.30 G VS1 GIA 721.8565

5 0.31 D VS1 GIA 940.1322

6 0.31 E VS1 GIA 890.8626

> #Scatter plot

> plot(x = diamonds$carat, y = diamonds$price, xlab = "Carat", ylab = "Price", main

 = "Price vs. Carat", col = "black", pch = 1, lwd = 1, panel.first = grid(col

 = "gray", lty = "dotted"))



Yes, there appears to be a positive relationship because as carat increases the price tends to increase as well. There may be a quadratic relationship too.

1. Find the sample regression models individually for carat, color, clarity, and certify (explanatory variables) with the price (response variable). State the four models.

 = -1316.73 + 6645.02Carat

 = 4067.5 – 912.0E – 1325.4F – 1531.8G – 1223.1H – 1102.8\*I

 = 1543.8 + 1353.3VS1 + 1812.3VS2 + 1645.9VVS1 + 1524.9VVS2

 = 3042.3 + 1071.6HRD – 1743.5IGI

If you are uncomfortable with how to use qualitative variables (like color) in a regression model, please see Chapter 8 (Section 8.3) of my STAT 870 lecture notes (http://www.chrisbilder.com/regression).

> #Find models for each explanatory variable

> mod.fit.carat <- lm(formula = price ~ carat, data = diamonds)

> summary(mod.fit.carat)

Call:

lm(formula = price ~ carat, data = diamonds)

Residuals:

 Min 1Q Median 3Q Max

-1297.47 -346.19 -66.53 249.30 3776.27

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) -1316.73 90.82 -14.50 <2e-16 \*\*\*

carat 6645.02 131.83 50.41 <2e-16 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 640.3 on 306 degrees of freedom

Multiple R-Squared: 0.8925, Adjusted R-squared: 0.8922

F-statistic: 2541 on 1 and 306 DF, p-value: < 2.2e-16

> mod.fit.color <- lm(formula = price ~ color, data = diamonds)

> summary(mod.fit.color)

Call:

lm(formula = price ~ color, data = diamonds)

Residuals:

 Min 1Q Median 3Q Max

-3563.4 -1806.3 -400.3 1514.2 5228.6

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 4067.5 483.9 8.406 1.70e-15 \*\*\*

colorE -912.0 565.1 -1.614 0.10757

colorF -1325.4 529.0 -2.505 0.01276 \*

colorG -1531.8 540.2 -2.836 0.00488 \*\*

colorH -1223.1 543.7 -2.250 0.02519 \*

colorI -1102.8 572.6 -1.926 0.05504 .

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1936 on 302 degrees of freedom

Multiple R-Squared: 0.03044, Adjusted R-squared: 0.01439

F-statistic: 1.896 on 5 and 302 DF, p-value: 0.09476

> mod.fit.clarity <- lm(formula = price ~ clarity, data = diamonds)

> summary(mod.fit.clarity)

Call:

lm(formula = price ~ clarity, data = diamonds)

Residuals:

 Min 1Q Median 3Q Max

-2990.6 -1111.4 -567.7 1182.2 6426.9

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 1543.8 283.1 5.453 1.03e-07 \*\*\*

clarityVS1 1353.3 351.7 3.848 0.000145 \*\*\*

clarityVS2 1812.3 383.0 4.732 3.42e-06 \*\*\*

clarityVVS1 1645.9 384.7 4.279 2.52e-05 \*\*\*

clarityVVS2 1524.9 354.1 4.307 2.24e-05 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1878 on 303 degrees of freedom

Multiple R-Squared: 0.08428, Adjusted R-squared: 0.0722

F-statistic: 6.972 on 4 and 303 DF, p-value: 2.216e-05

> mod.fit.certify<-lm(formula = price ~ certify, data = diamonds)

> summary(mod.fit.certify)

Call:

lm(formula = price ~ certify, data = diamonds)

Residuals:

 Min 1Q Median 3Q Max

-2413.3 -964.1 -416.1 933.5 6128.7

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 3042.3 135.7 22.425 < 2e-16 \*\*\*

certifyHRD 1071.6 231.5 4.629 5.44e-06 \*\*\*

certifyIGI -1743.5 232.5 -7.500 7.00e-13 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1667 on 305 degrees of freedom

Multiple R-Squared: 0.2736, Adjusted R-squared: 0.2688

F-statistic: 57.44 on 2 and 305 DF, p-value: < 2.2e-16

1. (3 points) Is there a linear relationship between carat and price? Perform both a t-test and a F-test to answer this question. Make sure to include all 5 steps for each test. I recommend using the p-value method to make the test easier.

The answer below is the same for both tests.

* + 1. Ho:βCarat=0 vs. Ha:βCarat≠0
		2. p-value < 2\*10-16
		3. =0.05
		4. Since p-value < 0.05, reject Ho.
		5. There is a linear relationship between carat and price.
1. (3 points) Is there a linear relationship between color and price, clarity and price, and certify and price? Explain your answers using the appropriate hypothesis tests. Note that you do not need to repeat all 5 steps for a hypothesis test here. I am only interested in determining if you can correctly interpret the hypothesis test results.

|  |  |
| --- | --- |
| Variable | **F-test p-value**  |
| Color | 0.09476 |
| Clarity | 2.216\*10-5 |
| Certify | 2.2\*10-16 |

There is a linear relationship for clarity and certify since the p-values are small. The p-value for color is only a little above the chosen α = 0.05 for this project. Thus, it is difficult to conclude if there is or is not a relationship.

1. (3 points) Which explanatory variable (carat, color, clarity, or certify) is doing the best job of estimating the price? Explain your answer using R2.

|  |  |
| --- | --- |
| Variable | **R2** |
| Carat | 89.25% |
| Color | 3.04% |
| Clarity | 8.42% |
| Certify | 27.36% |

Carat is the best since it has the largest R2.

1. (3 points) What characteristics of a diamond tend to cause it to be highly priced? Specifically state these characteristics and justify them with the use of your sample models.

|  |  |
| --- | --- |
| Variable | **Characteristic** |
| Carat | Large value since  >0 |
| Color | D produces the largest price (note that all of the other  are < 0). Care should be taken with interpreting this result since there was not sufficient evidence to conclude a linear relationship between color and price. |
| Clarity | VS2 produces the largest price. This is counterintuitive because one would expect the better the clarity, the higher price. There could be other characteristics of the data that we have not considered which could be the cause (perhaps the VS2 diamonds also tend to be the largest in carat size). This will be discussed more in future projects.  |
| Certify | HRD produces the largest price. This does not necessarily mean that HRD prices the same exact type of diamond higher than the other 2! We are not adjusting for other factors that affect price here – like carat. Maybe HRD happened to price most of the larger carat diamonds??? This will be discussed more in future projects. |

1. (3 points) Suppose two people are about to get engaged and they plan to purchase a diamond engagement ring. You offer your statistical expertise to aid them in pricing diamonds. Suppose the couple is interested in a 0.5 carat diamond. Find the estimated price of a 0.5 carat diamond using the appropriate sample regression model. Show how you can get the answer using by-hand calculations and through the use of R.

The predict() function in R is commonly used to obtain estimates of a response from a model:

> predict(object = mod.fit.carat, newdata = data.frame(carat = 0.5), interval =

 "prediction", level = 0.95)

 fit lwr upr

[1,] 2005.778 743.4186 3268.138

 = -1316.73 + 6645.02Carat = -1316.73 + 6645.02\*0.5 = $2,005.78

1. (3 points) Continuing 8), which is the more appropriate interval to use with estimating the price: confidence or prediction? Explain your answer and find the interval.

Prediction interval since we want to price the diamond for one couple’s engagement ring. The interval from the output is $743.42 < Price < $3,268.14.

1. (3 points) When shopping for a diamond engagement ring, a local jewelry store quotes the couple a price of $2,500 for a 0.5 carat, round cut diamond that can be placed in a ring. Is this price fair? Explain your answer using the correct information found in this project. Assume your regression model can be used to estimate prices of diamonds in Lincoln.

Yes, since the $2,500 price is inside the prediction interval.

1. (4 points) Using carat as the explanatory variable, construct a scatter plot with the sample model, 95% confidence interval bands, and 95% prediction interval bands plotted upon it. Show your interval found in 9) on the plot at carat = 0.5.

Because this is a more complicated plot, you are not responsible for it until after the graphics part of the course.

plot(x = diamonds$carat, y = diamonds$price, xlab = "Carat", ylab = "Price", main =

 "Price vs. Carat", col = "black", pch = 1, lwd = 1, panel.first = grid(col =

 "gray", lty = "dotted"), xlim = c(0,1.25), ylim = c(-1000, 10000))

curve(expr = predict(object = mod.fit.carat, newdata = data.frame(carat = x)), col

 = "red", lty = "solid", lwd = 2, add = TRUE, from = min(diamonds$carat), to =

 max(diamonds$carat))

curve(expr = predict(object = mod.fit.carat, newdata = data.frame(carat = x),

 interval = "confidence", level = 0.95)[,2], col = "darkgreen", lty =

 "dashed", lwd = 1, add = TRUE, from = min(diamonds$carat), to =

 max(diamonds$carat))

curve(expr = predict(object = mod.fit.carat, newdata = data.frame(carat = x),

 interval = "confidence", level = 0.95)[,3], col = "darkgreen", lty =

 "dashed", lwd = 1, add = TRUE, from = min(diamonds$carat), to =

 max(diamonds$carat))

curve(expr = predict(object = mod.fit.carat, newdata = data.frame(carat = x),

 interval = "prediction", level = 0.95)[,2], col = "blue", lty = "dashed", lwd

 = 1, add = TRUE, from = min(diamonds$carat), to = max(diamonds$carat))

curve(expr = predict(object = mod.fit.carat, newdata = data.frame(carat = x),

 interval = "prediction", level = 0.95)[,3], col = "blue", lty = "dashed", lwd

 = 1, add = TRUE, from = min(diamonds$carat), to = max(diamonds$carat))

legend(locator(1), legend = c("Sample model", "95% C.I.", "95% P.I."), col =

 c("red", "darkgreen", "blue"), lty = c("solid", "dashed", "dashed"), bty =

 "n", cex = 0.75)

save.pred<-predict(object = mod.fit.carat, newdata = data.frame(carat = 0.5),

 interval = "prediction", level = 0.95)

# Lower

segments(x0 = 0.5, y0 = -2000, x1 = 0.5, y1 = save.pred[1,2], lty = "dotted", col =

 "black", lwd = 2)

segments(x0 = 0.5, y0 = save.pred[1,2], x1 = 0, y1 =save.pred[1,2] , lty =

 "dotted", col = "black", lwd = 2)

# Upper

segments(x0 = 0.5, y0 = -2000, x1 = 0.5, y1 = save.pred[1,3], lty = "dotted", col =

 "black", lwd = 2)

segments(x0 = 0.5, y0 = save.pred[1,3], x1 = 0, y1 =save.pred[1,3] , lty =

 "dotted", col = "black", lwd = 2)

mtext(text = round(save.pred[2],2), side=2, cex = 0.75, at = save.pred[2], las = 2)

 #las makes perpendicular to axis

mtext(text = round(save.pred[3],2), side=2, cex = 0.75, at = save.pred[3], las = 2)



