STAT 135 – Problem Set 10

Quincy Hughes

due by Tuesday, April 23 at 10 AM

I *strongly* encourage that you read the questions as soon as you get the assignment. This will help you to start thinking how to solve them! In case of questions, or if you get stuck, please don't hesitate to email me (though I'm considerably less sympathetic to questions when I receive emails within 24 hours of the due date for the assignment). Better yet, come visit me in my office hours:

Mondays 1:00-2:30 PM
Thursdays 1:00-2:30 PM

Remember that the Statistics & Data Science Fellows have drop-in hours Sunday-Thursday evenings from 7:00-9:00 PM in the new science center (room E208). I encourage you to use this resource; the fellows are able to help with questions regarding conceptual understanding of the course material, R and RMarkdown. Datasets for most of the tables, examples and exercises in IS5 are available online (https://nhorton.people.amherst.edu/is5).

Steps to proceed:

1. Download the file PS10.Rmd from Moodle
2. Upload the file to the RStudio server (r.amherst.edu)
3. Replace "YOUR NAME HERE" with your name
4. Add in your responses where it is marked SOLUTION:
5. Run "Knit PDF"
6. Upload the pdf to Gradescope

PRACTICE PROBLEMS (not to be submitted)

IS5 20.39, 20.49, 20.55, 20.57

PROBLEMS TO TURN IN (use PS10.Rmd template)

IS5 20.44, 20.52, 20.54 + 2 non-textbook problems

If you discussed this assignment with any of your peers, please list who here:

SOLUTION:

IS5 20.44 Fuel economy, part III

Consider the data in Exercise 40 about the gas mileage and weight of cars.

a) Create a 95% confidence interval for the average fuel efficiency among cars weighing 2500 pounds, and explain what your interval means.] > SOLUTION: We are 95% confident that the mean fuel efficiency for cars weighing 2500 pounds is between 27.34 and 29.07 mpg, according to our model.

```
Fuel <- read.csv("https://nhorton.people.amherst.edu/is5/data/Fuel_econom
  rename(Weight = wt.1000s.)

# fit the linear model (same output as displayed in Exercise 40)
model1 <- lm(MPG ~ Weight, data=Fuel)
msummary(model1)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  48.7393     1.9756   24.67  < 2e-16 ***
## Weight       -8.2136     0.6738  -12.19 2.64e-16 ***
##
## Residual standard error: 2.413 on 48 degrees of freedom
## Multiple R-squared:  0.7558, Adjusted R-squared:  0.7507
## F-statistic: 148.6 on 1 and 48 DF,  p-value: 2.641e-16
```

```
predict(model1,newdata=data.frame(Weight=2.5), int="confidence", level=0.
```

```
##        fit      lwr      upr
## 1 28.20524 27.34097 29.06951
```

```
# use the predict() function to generate confidence intervals for means

predict(model1,newdata=data.frame(Weight=3.45), int="prediction", level=0
```

```
##       fit      lwr      upr
## 1 20.4023 15.44277 25.36183
```

b) Create a 95% prediction interval for the gas mileage you might get driving your new 3450-pound SUV, and explain what that interval means in context.

SOLUTION: According to our model, we are 95% confident the 3450 pound SUV will have have a fuel efficiency between 15.44 mpg and 25.36 mpg.

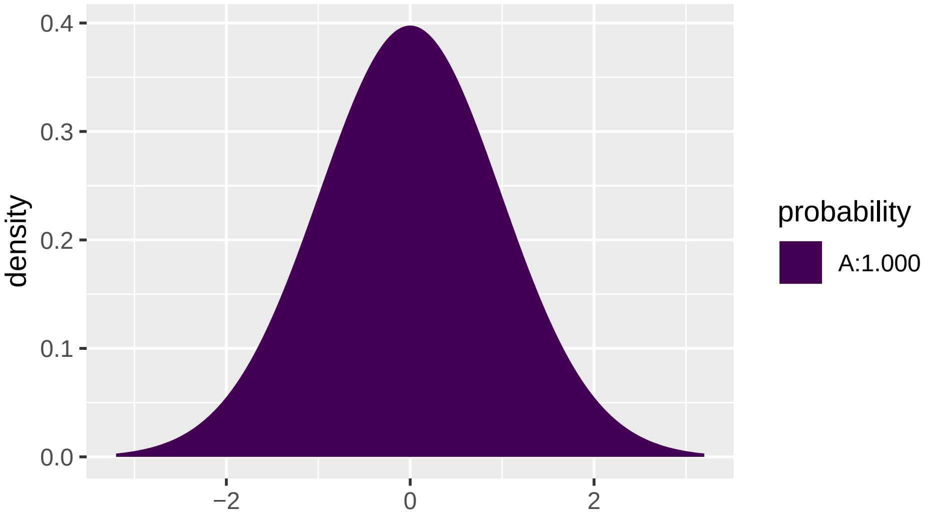
IS5 20.52 Sales and profits

A business analyst was interested in the relationship between a company's sales and its profits. She collected data (in milions of dollars) from a random sample of Fortune 500 companies and created the regression analsis and summary statistics shown (see page 680 for the chart). The assumptions for regression inference appeared to be satisfied.

a) Is there a statistically significant association between sales and profits? Test an appropriate hypothesis and state your conclusion in context.

SOLUTION: H0: b1=0. HA: b1=/=0. The test statistic for the observed result is a t value of 12.26. There is a 4e-20 chance of observing these results if the null hypothesis is true, so we reject the null hypothesis and say there is an association.

```
t<-0.092/0.0075
xpt(t,df=77,ncp=0)
```



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

```
t
```

```
## [1] 12.26667
```

b) Do you think the company's sales serve as a useful predictor of its profits? Use the values of both R^2 and s in your explanation.

SOLUTION: 66.2% of the variability in profits can be attributed to variability in sales. That is pretty high all things considered, so suggests that sales is useful as a predictor. The standard deviation of the residuals is 466.2, which is rather high (the standard deviation of the profits is 796.98) so it's not the best predictor. A prediction interval would have a very wide spread.

IS5 20.54 More sales and profits

Consider again the relationship between the sales and profits of Fortune 500 companies that you analyzed in Exercise 52.

a) Find a 95% confidence interval for the slope of the regression line. Interpret your interval in context.

   SOLUTION: 0.077-0.107. We are 95% confident that the true slope of the regression line of the relationship between sales and profits is between 0.077 and 0.107.

b) Last year, the drug manufacturer Eli Lilly, Inc., reported gross sales of $23 billion (that's $23,000 million). Create a 95% prediction interval for the company's profits, and interpret your interval in context.

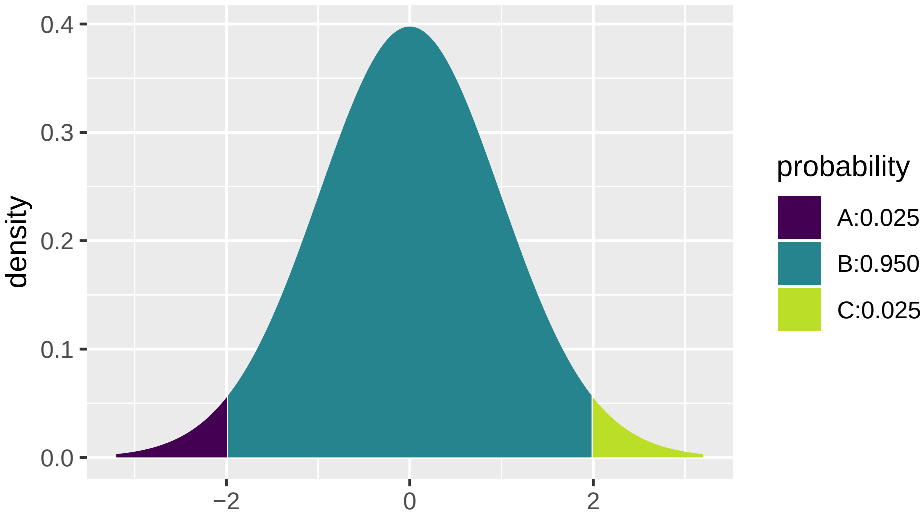
   SOLUTION: 975.0-2926.6. We are 95% confident that Eli Lilly Inc makes between 975.0 million and 2926 million in profit according to our model.

```
Fortune500 <- read.csv("https://nhorton.people.amherst.edu/is5/data/Sales

# fit the linear model (same output as displayed in Exercise 40)
model2 <- lm(Profits ~ Sales, data=Fortune500)
msummary(model2)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.766e+02  6.116e+01  -2.888  0.00503 **
## Sales        9.250e-02  7.528e-03  12.287  < 2e-16 ***
##
## Residual standard error: 466.2 on 77 degrees of freedom
## Multiple R-squared:  0.6622, Adjusted R-squared:  0.6578
## F-statistic:   151 on 1 and 77 DF,  p-value: < 2.2e-16
```

```
xqt(c(0.025,0.975),df=77)
```



```
## [1] -1.991254  1.991254
```

```
9.250e-02+1.99*7.528e-03
```

```
## [1] 0.1074807
```

```
9.250e-02-1.99*7.528e-03
```

```
## [1] 0.07751928
```

```
predict(model2,newdata=data.frame(Sales=23000), int="prediction", level=0
```

```
##        fit     lwr      upr
## 1 1950.806 974.978 2926.634
```

Non-textbook problem 1: Interaction

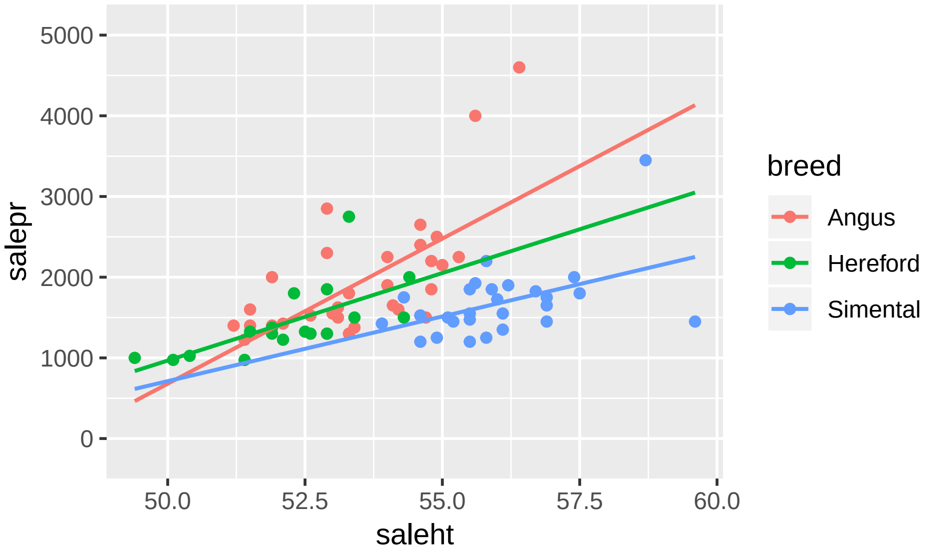
Does the association between the sale height of bulls and the sale price of bulls at auction depend on breed of th bull?

(a) Use the **Bulls** dataset to visualize the relationship of sale height and price by breed. What do you notice?

SOLUTION: The slopes differ based on the type of bull. Height has a stronger association with price (stronger meaning the effect is greater) for Angus bulls than the others, and a slighter stronger association for Hereford than for Simental.

```
Bulls <- read.table("https://awagaman.people.amherst.edu/stat230/bulls.tx
  mutate(breed = cut(breed, breaks = c(0, 4, 6, 10)
                                   , labels = c("Angus", "Hereford", "Sim
                                   , include.lowest = TRUE))

gf_point(salepr ~ saleht, color=~breed, data=Bulls) %>%
  gf_lm()
```



b) Fit an interaction model that allows the slope of saleht to vary by breed. What is the equation for predicting price from sale height among the Simental breed? What is the equation for predicting price from sale height among the Angus breed?

SOLUTION: Simental: sale price = -17302.88 + 160.35(sale height)

Angus: sale price = -17302.88 + 359.65(sale height)

```
model3 <- lm(salepr ~ saleht + breed + saleht*breed, data=Bulls)
msummary(model3)
```

```
##                        Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           -17302.88    3392.00  -5.101 2.77e-06 ***
## saleht                   359.65      63.32   5.680 2.85e-07 ***
## breedHereford           7423.34    5694.36   1.304   0.1966    
## breedSimental           9997.08    5335.61   1.874   0.0652 .  
## saleht:breedHereford    -142.72     108.09  -1.320   0.1910    
## saleht:breedSimental    -199.30      97.01  -2.054   0.0437 *  
## 
## Residual standard error: 479.4 on 70 degrees of freedom
## Multiple R-squared:  0.4473, Adjusted R-squared:  0.4078 
## F-statistic: 11.33 on 5 and 70 DF,  p-value: 5.015e-08
```

```
359.65-199.30
```

```
## [1] 160.35
```

c) Is there significant interaction between sale height and breed by price? Report the relevant test statistic(s) and p-value(s), and state your conclusion in context.

SOLUTION: There is interaction. The interaction term for Simental bulls has a t value of -2.054 and a p value of 0.04. This is less than 0.05, so I consider it statistically significant. the interaction term for Hereford bulls has a t value of -1.32 and a p value of 0.19. The p value is not high enough to be significant, as there is a 0.19 chance of getting these results if the actual interaction term has a value of 0. The effect of the height on the sale price of Simental bulls is significantly different from the effect of height on the sale price of Angus bulls.

Non-textbook problem: Linear regression in use

Find a peer-reviewed journal article that uses multiple linear regression. On the Moodle course site, there is a forum “Linear regression in use”. Click “Add a new discussion topic” and write a post that briefly summarizes what the study question is, why multiple linear regression was used, and whether you think the article correctly interprets the linear regression model. Attach a PDF of the article to the post.

> SOLUTION: DO NOT WRITE ANYTHING HERE. INSTEAD, CREATE A POST IN THE “LINEAR REGRESSION IN USE” DISCUSSION FORUM ON THE MOODLE COURSE WEBSITE.