Stat 140

Practice Midterm 1

Name: ____

Section: ____

You may use a calculator and one 8.5'‘by 11” sheet of notes (front and back), which you will turn in with your exam. I will not take the contents of your notes sheet into account when setting your grade; I collect them only because students occasionally make mistakes when putting together their notes sheets, and I want to fix those mistakes in case you use the notes sheet again on Exam 3.

Please show all your work, including all calculations, and explain your answers.

Whenever needed, please round numbers (including intermediate calculations) to the nearest 0.001.

Cell phones and any other electronic devices (aside from your calculator) are not permitted. No interaction of any sort is allowed with your classmates.

Conceptual Questions

Please answer the following in no more than 3-4 sentences each.

1. (3 points)

An article titled “Tough Times in Japan” stated the following:

> A recent survey by Japan’s Ministry of Health, Labor and Welfare indicates that as many as 60% of all households now have annual incomes below the national average of $52,339.

Explain how this can possibly make sense, even if a country was not having tough economic times.

2. (4 points)

In 1 or 2 sentences, give the definition and interpretation of interquartile range.

3. (4 points)

What are residual plots useful for?

4. (4 points)

Describe Simpson’s Paradox in 3 sentences or less.

Applied Problems

1. (10 points)

Students in a large statistics class each took one of 4 jigsaw puzzles to solve. The following data were recorded (showing the last 5 rows):

Student	Puzzle	Pieces	SolveTime
19AB	Plane	350	1.2
18CJ	Plane	350	2.1
21QX	SnowScene	1000	8.9
19JH	Lake	500	2.5
21AA	Kittens	300	1.9

Student is a student identifier, *Puzzle* is a label indicating which one of the 10 puzzles were solved, *Pieces* is the number of pieces in the puzzle, and *SolveTime* is the time taken to put together the puzzle (in hours).

(a) (2 points)

What are the quantitative variables?

(b) (2 points)

What are the categorical variables?

(c)

Pick the most appropriate plot for each task below. Circle the best choice.

i. (3 points) Examine the relationship between puzzle and solve time.

histogram scatterplot bargraph boxplot

ii. (3 points) Examine the relationship between the number of pieces and solve time.

histogram scatterplot bargraph boxplot

2. (12 points)

The following contingency table shows the breakdown of the adult survivors on the Titanic between `Class' and`Gender’.

Class/Gender	Male	Female
1st	57	140
2nd	14	80
3rd	75	76
Crew	192	20

(a) (4 points)

What is the marginal distribution of the class of adult survivors?

(b) (4 points)

What proportion of surviving adults were male crew members?

(c) (4 points)

What is the conditional distribution of a survivor’s gender, given that the survivor was in first class?

3. (12 points)

Suppose that in South Hadley, on winter days when it snows the total amount of snow fall follows a normal distribution with mean 6 inches and standard deviation 2 inches.

(a) (4 points)

On what proportion of snowy days does at least 9 inches of snow fall? You may use the following output from R in answering this question:

```
pnorm(q = 9, mean = 6, sd = 2)
```

```
## [1] 0.933
```

(b) (4 points)

What is the 84th percentile for the amount of snow that falls on snowy winter days? Use the 68-95-99.7 rule to find this.

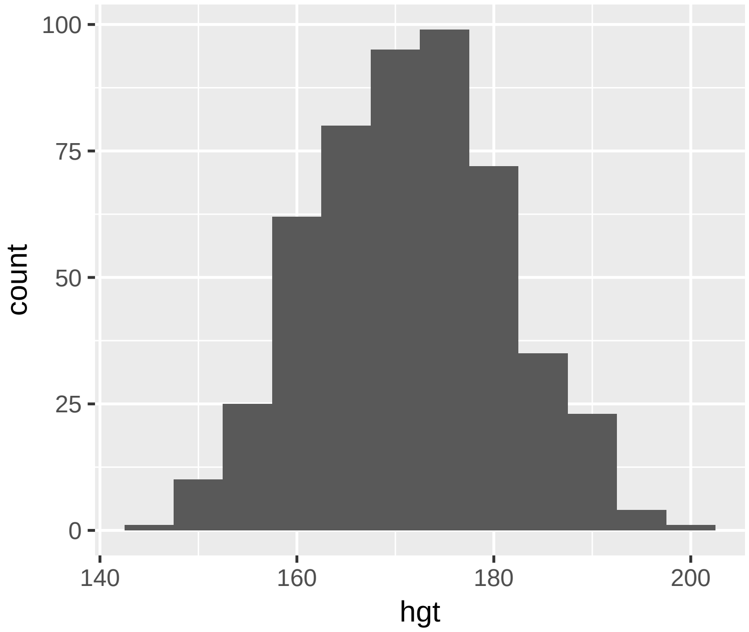
(c) (4 points)

On a particularly snowy day few years ago, 14 inches of snow fell. Find the z-score for this observation. What does the z-score measure?

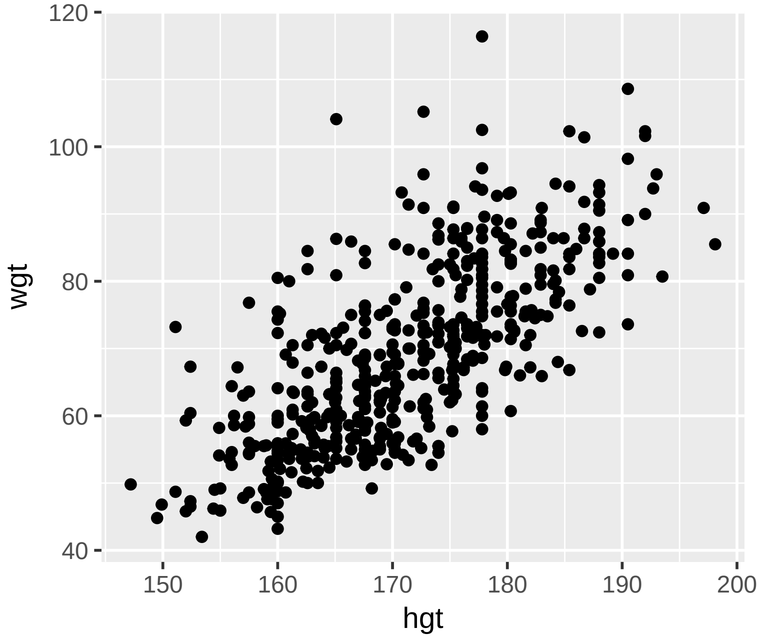
4. (32 points)

Let's look at some data showing the relationship between weight measured in kilograms (`wgt`) and height measured in centimeters (`hgt`) of 507 physically active individuals, stored in a data frame named `bdims`. Here are some plots of the data, summary statistics, and results from a linear model fit:

```r
ggplot() +
  geom_histogram(mapping = aes(x = hgt), binwidth = 5, data = bdims)
```



```r
ggplot() +
  geom_point(mapping = aes(x = hgt, y = wgt), data = bdims)
```



```
summarize(bdims,
  mean_hgt = mean(hgt),
  median_hgt = median(hgt),
  sd_hgt = sd(hgt),
  iqr_hgt = IQR(hgt)
)
```

```
##   mean_hgt median_hgt sd_hgt iqr_hgt
## 1      171        170   9.41      14
```

```
lm_weight_height <- lm(wgt ~ hgt, data = bdims)
coef(lm_weight_height)
```

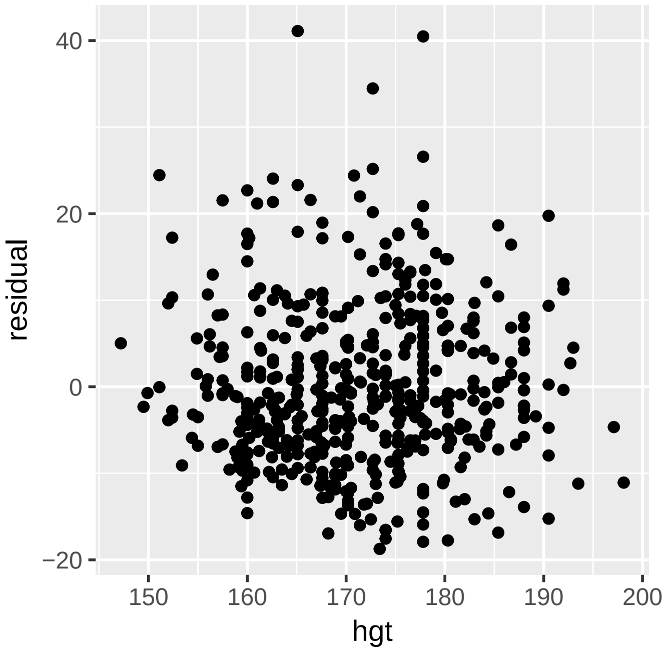
```
## (Intercept)         hgt
##     -105.01        1.02
```

```
summary(lm_weight_height)
```

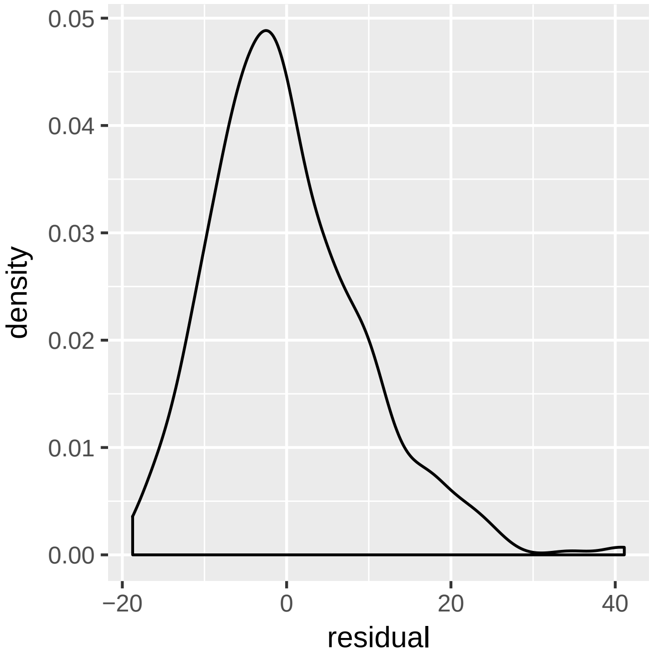
```
##
## Call:
## lm(formula = wgt ~ hgt, data = bdims)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18.74  -6.40  -1.23   5.06  41.10
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -105.011      7.539   -13.9   <2e-16 ***
## hgt            1.018      0.044    23.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.31 on 505 degrees of freedom
## Multiple R-squared:  0.515,  Adjusted R-squared:  0.514
## F-statistic:  535 on 1 and 505 DF,  p-value: <2e-16
```

```
bdims <- mutate(bdims,
  residual = residuals(lm_weight_height))

ggplot() +
  geom_point(mapping = aes(x = hgt, y = residual), data = bdims)
```



```
ggplot() +
  geom_density(mapping = aes(x = residual), data = bdims)
```



(a) (4 points)

Using the histogram of heights at the top of page 6, as well as the summary statistics calculated at the top of page 7, describe the **center**, **spread**, **shape**, and any **unusual features** of the distribution of heights in this data set. Be sure to include **units**. Justify your choice of statistics for summarizing the center and spread of the distribution.

(b) (7 points)

Using the scatter plot at the bottom of page 6, describe the relationship between height and weight. Address its **form**, **direction**, **strength**, and any **unusual features**. Based on just what you can see in that plot, would it be appropriate to use a linear model for this data set? Check all assumptions that you can check based on the scatter plot.

(c) (6 points)

Write the equation of the regression line. Interpret the slope and intercept in context.

(d) (3 points)

List all assumptions/conditions for the model that you could check using the residual plots on page 8. Based on the plots, do you see evidence of any potential problems with the linear model?

(e) (3 points)

Your friend has a height of 170 cm. What is the predicted value for her height based on the linear regression model? Show the setup/work in addition to your answer.

(f) (3 points)

Your friend's actual weight is 50 kg. What would the residual for this observation be? Show the setup/work in addition to your answer.

(g) (2 points)

Your 6 month old nephew has a height of 66 cm. Would it make sense to predict his weight using this linear model? Why or why not?

(h) (2 points)

What is the R^2 value for this linear model? Interpret it in context.

(i) (2 points)

What is the residual standard deviation for this linear model? Interpret it in context using the “95” part of the 68-95-99.7 rule.