Homework 1 - Key

Math 530/630

Setup

Install and load the `MASS` package in R so that you can use the `cats` dataset.Using the `cats` dataset, we'll explore the difference between *covariance* and *correlation*.

```
#install.packages("MASS")
library(MASS)
library(dplyr)
library(ggplot2)
head(cats)
```

```
##   Sex Bwt Hwt
## 1   F 2.0 7.0
## 2   F 2.0 7.4
## 3   F 2.0 9.5
## 4   F 2.1 7.2
## 5   F 2.1 7.3
## 6   F 2.1 7.6
```

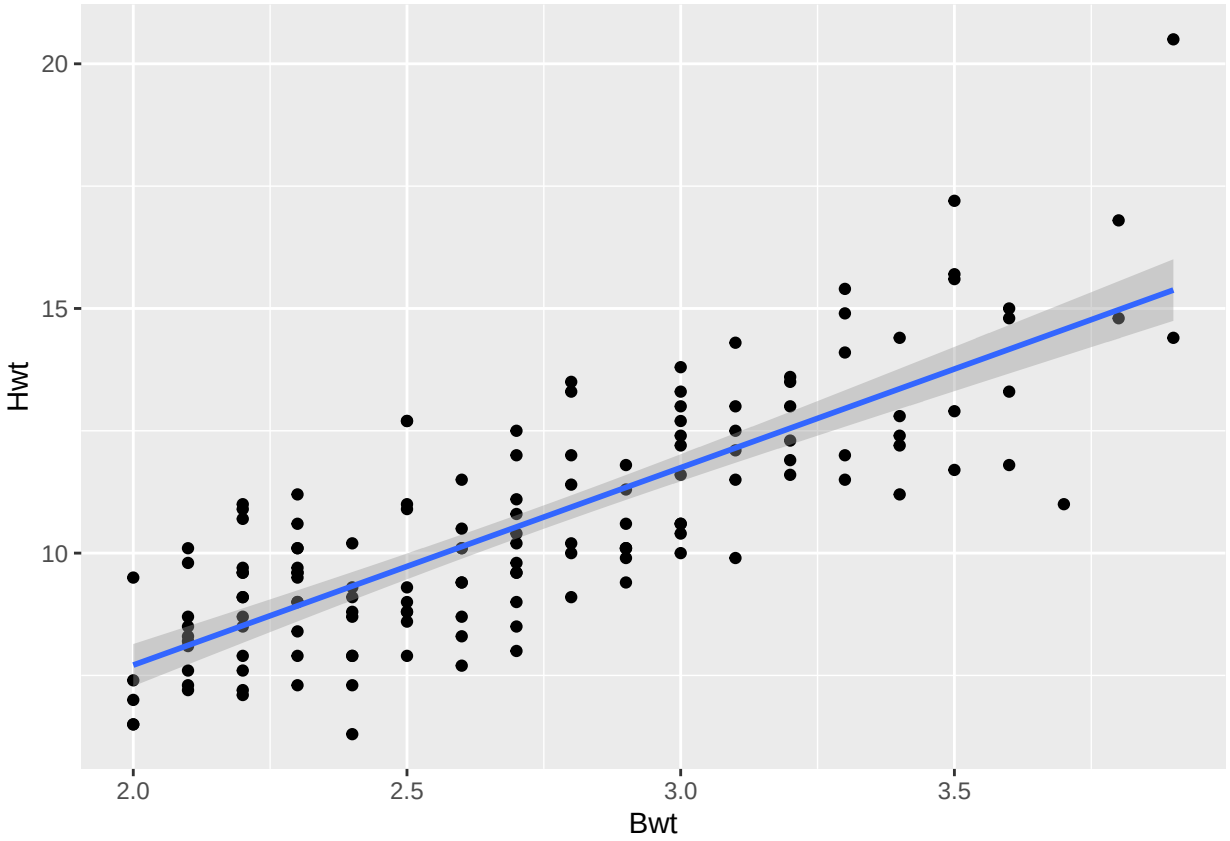
```
glimpse(cats)
```

```
## Observations: 144
## Variables: 3
## $ Sex <fct> F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F...
## $ Bwt <dbl> 2.0, 2.0, 2.0, 2.1, 2.1, 2.1, 2.1, 2.1, 2.1, 2.1, 2.1, 2.1...
## $ Hwt <dbl> 7.0, 7.4, 9.5, 7.2, 7.3, 7.6, 8.1, 8.2, 8.3, 8.5, 8.7, 9.8...
```

Problems

1. Make a scatterplot of body weight versus heart weight in this sample of cats, and add the linear regression line. What do you see? (hint: comment on linearity of association and strength of that association) (See also: http://guessthecorrelation.com). Using the built-in R function `cov`, calculate the covariance of body weight and heart weight in this sample of cats.

```
ggplot(cats, aes(x = Bwt, y = Hwt)) +
  geom_point() +
  geom_smooth(method = "lm")
```




```
# Body and heart weight appear to have a strong linear association,
# with heart weight increasing consistently as body weight increases.

cats %>%
  summarize(cov(Hwt,Bwt))
```

```
##   cov(Hwt, Bwt)
## 1     0.9501127
```

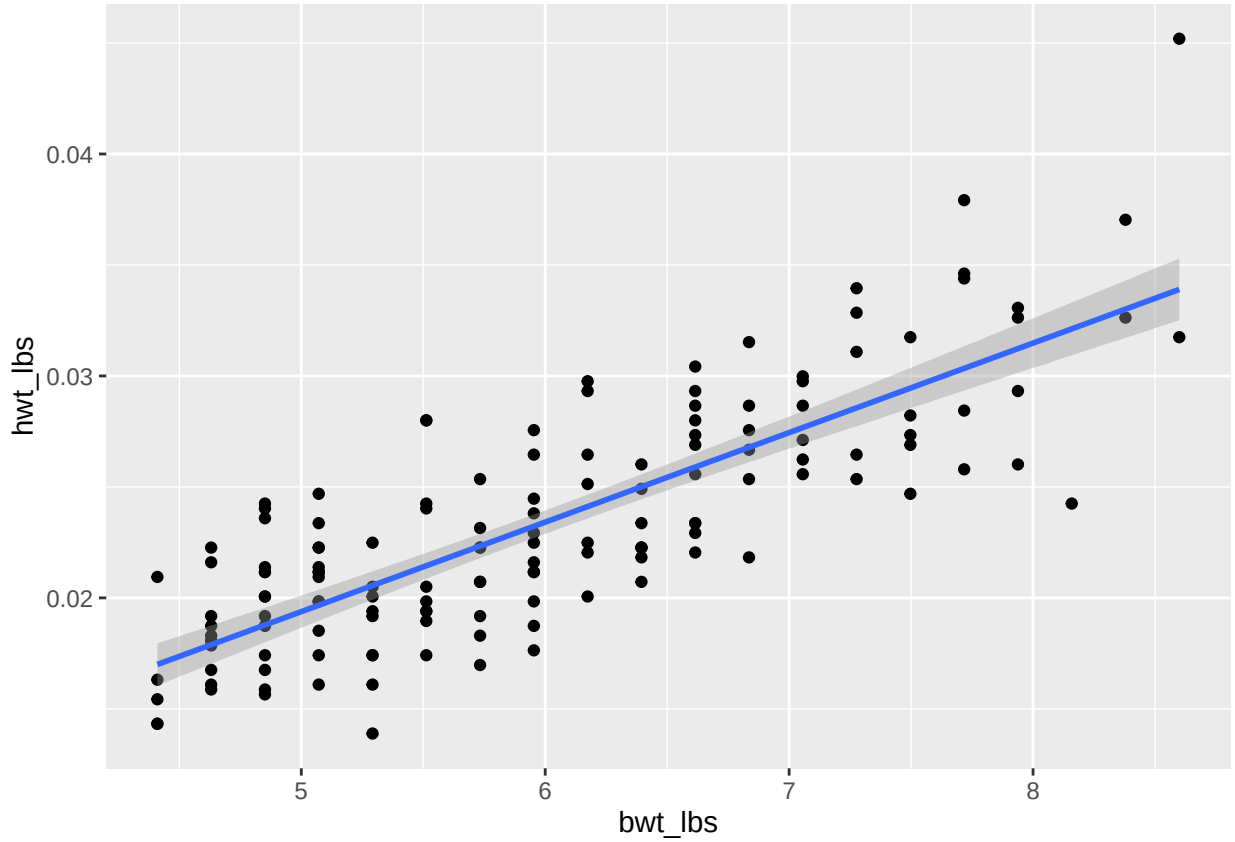
2. Convert the units of both body and heart weight from the metric system to the imperial system, using the code provided below, and re-make your scatterplot. What changed? What did not change? Also do the following:

    - Re-calculate the covariance between body and heart weight in pounds. What do you notice? Does it make sense?
    - Using the built-in R function `cor`, calculate two correlations between body and heart weight- one on the raw values in kg and g, respectively, and one using the same measurements after converting to pounds.
    - Compare and contrast the two correlations you calculated to the two covariances you have calculated.

```
# convert units for part c
cats_imperial <- cats %>%
  mutate(bwt_lbs = Bwt * 2.205,
         hwt_lbs = Hwt * 0.0022046)
```

```
ggplot(cats_imperial, aes(x = bwt_lbs, y = hwt_lbs)) +
  geom_point() +
```

```
  geom_smooth(method = "lm")
```




```
cats_imperial %>%
  summarise(cov_metric = cov(Bwt, Hwt),
            cov_imperial = cov(bwt_lbs, hwt_lbs))
```

```
##   cov_metric cov_imperial
## 1  0.9501127  0.004618634
```

```
# The covariance plummeted, but the linear relationship looks the same
# Hey! The new covariance is exactly... 2.205 * 0.0022046 * .95


cats_imperial %>%
  summarise(cov_metric = cor(Bwt, Hwt),
            cov_imperial = cor(bwt_lbs, hwt_lbs))
```

```
##   cov_metric cov_imperial
## 1  0.8041274    0.8041274
```

```
# So, covariance changes when converting the units, but correlation does not.
```

3. Finally, go back to the original raw measures of body and heart weight (in kg and g, respectively).
    - Transform each variable into z-score form (you may wish to confirm for yourself that the mean = 0 and sd = 1 for each).
    - Calculate the covariance and the correlation between the two variables in z-score form.
    - Use this example to explain in words how a correlation is different from covariance (hint: do

you think you want a measure of association between two variables that is sensitive to linear transformations? Why or why not? Do either of these two statistics appear to be insensitive to linear transformation? If so, look carefully at the formulas and try to explain in words why.)

```
# convert to z-scores
zcats <- cats_imperial %>%
  mutate(bwt_z = (Bwt - mean(Bwt))/sd(Bwt),
         hwt_z = (Hwt - mean(Hwt))/sd(Hwt))

zcats %>%
  summarise(cov_metric = cov(Bwt, Hwt),
            z_cov_metric = cov(bwt_z, hwt_z),
            z_cor_metric = cor(bwt_z, hwt_z))
```

```
##   cov_metric z_cov_metric z_cor_metric
## 1  0.9501127    0.8041274    0.8041274
```

```
# note covariance of z-scores and correlation are the same
```

Key takeaway:

Pearson's product moment correlation coefficient is invariant under linear transformation of x, y, or both. This is because correlation is the covariance of the z-scores. A measure of association between two variables is not very useful if linear transformation of either variable changes the statistic.