Lesson 2

Nicholas Clark

In the United States, the 1963 Equal Pay Act requires that men and women be given equal pay for equal work and Title VII of the Civil Rights Act of 1964 prohibits discrimination on the basis of race, color, relgion, sex, and national origin. How successful have these acts been?

WageRace contains observations from 1987 for a sample of 25,632 males between the age of 18 and 70 who worked full-time along with their years of educaiton, years of experience, race, whether they worked in a standard metropolitan area, and the region of US where they worked.

Primary research question is whether wages for blacks differ significantly from wages for non-blacks?

```
library(tidyverse)
wage.dat<-read.table("http://www.isi-stats.com/isi2/data/Wages.txt",header=T)
```

Identify the observational units in the study. How many are there?

```
nrow(wage.dat)
```

```
## [1] 25631
```

```
#head(wage.dat) This gives the first couple of entries
```

Is the wages variable a quantitative or categorical variable?

```
ggplot(wage.dat,aes(x=wage))+geom_histogram(bins=100)
```

```
ggplot(wage.dat,aes(y=wage))+geom_boxplot()+coord_flip()
```

Why are we looking at histograms and boxplots rater than a bar graph?

Does anything stand out to you about the boxplot that is less obvious in the histogram?

Which visual, the histogram or boxplot, do you like better? Why?

Which is larger, the mean or the median? How do you know?

Do the wages appear to follow a normal distribution? How do you know?

In this study, the researchers were most interested in whether race explained differences in wages.

Which variable is the explanatory variable? Which is the response variable?

Do you think the explanatory variable explains some variation in the response variable? Do you think it explains all of the variation in the response variable? Why or why not?

```
ggplot(wage.dat,aes(y=wage,x=race))+geom_boxplot()+coord_flip()
```

```
wage.dat%>%group_by(race)%>%
  summarise(n=n(),mean=mean(wage),StDev=sd(wage),Minimum=min(wage),Median=median(wage),Maximum=max(wage
```

```
## # A tibble: 2 x 7
##   race         n  mean StDev Minimum Median Maximum
##   <fct>    <int> <dbl> <dbl>   <dbl>  <dbl>   <dbl>
## 1 black     1988  479.  308.    53.8   412.   3527.
## 2 nonblack 23643  654.  451.    50.4   570.  18777.
```

Consider whether there appears to be an association between wage and race: Does the wage distribution differ substantially between blacks and non-blacks? What is the difference in the mean weekly wages? Can we conclude wage discrimination?

```
ggplot(wage.dat,aes(y=wage,x=educ))+geom_boxplot()+coord_flip()
```

```
wage.dat%>%group_by(educ)%>%
  summarise(n=n(),mean=mean(wage),StDev=sd(wage),Minimum=min(wage),Median=median(wage),Maximum=max(wage
```

```
## # A tibble: 4 x 7
##   educ              n  mean StDev Minimum Median Maximum
##   <fct>         <int> <dbl> <dbl>   <dbl>  <dbl>   <dbl>
## 1 belowHS        1509  414.  282.    61.7   356.   5144.
## 2 beyondCollege  2925  966.  499.    59.3   878.   3601.
## 3 college        8977  703.  476.    54.2   636.  18777.
## 4 HS            12220  544.  367.    50.4   480.  15124.
```

Suggest an easy way to improve this graphical display to better focus on a trend of increasing salaries with increasing education.

Describe the association between education and wage. Is it as you would have predicted? Explain.

What would need to be true for education level to provide an alternative explanation for why non-blacks in this sample tended to earn more than blacks?

```
ggplot(wage.dat,aes(y=wage,x=educ,fill=race))+geom_boxplot()+coord_flip()
```

```
wage.dat%>%group_by(educ,race)%>%
  summarise(mean=mean(wage),StDev=sd(wage))
```

```
## # A tibble: 8 x 4
## # Groups:   educ [4]
##   educ          race      mean StDev
##   <fct>         <fct>    <dbl> <dbl>
## 1 belowHS       black     368.  208.
## 2 belowHS       nonblack  419.  289.
## 3 beyondCollege black     847.  487.
## 4 beyondCollege nonblack  971.  499.
## 5 college       black     544.  315.
## 6 college       nonblack  714.  483.
## 7 HS            black     425.  257.
## 8 HS            nonblack  556.  374.
```

Is there a difference in the average wage between blacks and non-blacks in the "beyond college" group? Is this difference larger or smaller than when we did not take the education level into account?

Do the lower average wages for blacks compared to non-blacks appear to be consistent across each of the education levels?

If you were to compare the average weekly wage for blacks to the average weekly wage for non-blacks in the same education group, roughly how large would you say that difference is?

How do you respond to the argument that the wage disparity between blacks and non-blacks is really an issue of education level?

Sources of variation diagram:

```
birthwt.dat<-read.csv("births.csv")
```

Explore:

```
ggplot(aes(x=weight),data=birthwt.dat)+geom_histogram(bins=100)
```

Filter out the unknowns

```
birthwt.clean<-birthwt.dat %>% filter(weight < 8166)
ggplot(aes(x=weight),data=birthwt.clean)+geom_histogram(bins=100)
```

summary statistics

```
birthwt.clean%>%summarise(N=n(),Mean=mean(weight),StDev=sd(weight),Min=min(weight),Max=max(weight))
```

```
##        N     Mean   StDev Min  Max
## 1 317038 3259.127 592.212 227 8165
```

If we used the mean to predict future newborn weight how well would we do?

The statistical model would be:

A residual is the value $y_i - \hat{y}_i$ for $i = 1, \cdots, n$. We can find the residuals two different ways:

```
birthwt.resid<-birthwt.clean%>% mutate(resid=weight-mean(weight))%>%select(resid)
ggplot(aes(x=resid),data=birthwt.resid)+geom_histogram(bins=100)
```

```
birthwt.resid%>%summarise(Mean=mean(resid),StdDev=sd(resid))
```

```
##          Mean  StdDev
## 1 1.16267e-14 592.212
```

What is going on? Have we explained any variation?

```
ggplot(aes(x=weight,color=full.term),data=birthwt.clean)+geom_histogram(fill="white", alpha=0.5, positi
```

```
birthwt.clean%>%group_by(full.term)%>%
  summarise(N=n(),Mean=mean(weight),StDev=sd(weight),Min=min(weight),Max=max(weight))
```

```
## # A tibble: 2 x 6
##   full.term      N  Mean StDev   Min   Max
##   <lgl>      <int> <dbl> <dbl> <int> <int>
## 1 FALSE      36963 2494.  796.   227  5670
## 2 TRUE      280075 3360.  475.   320  8165
```

Our predicted model becomes:

Standard error:

```
model<-lm(weight~0+full.term,data=birthwt.clean)
summary(model)
```

```
##
## Call:
## lm(formula = weight ~ 0 + full.term, data = birthwt.clean)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3040.1  -320.1    -0.1   324.9  4804.9
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## full.termFALSE 2493.793      2.720   916.9   <2e-16 ***
## full.termTRUE  3360.132      0.988  3400.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 522.9 on 317036 degrees of freedom
## Multiple R-squared:  0.9751, Adjusted R-squared:  0.9751
## F-statistic: 6.203e+06 on 2 and 317036 DF,  p-value: < 2.2e-16
```

Does Mom's BMI impact weight?

```
birthwt.bmi<-birthwt.clean%>%filter(mom.BMI < 90)
model<-lm(weight~0+full.term*mom.BMI,data=birthwt.bmi)
summary(model) #I think there's an error in book
```

```
##
## Call:
## lm(formula = weight ~ 0 + full.term * mom.BMI, data = birthwt.bmi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2998.8  -313.1     0.2   321.5  3558.3
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## full.termFALSE        2409.1457    11.0257 218.502   <2e-16 ***
## full.termTRUE         3130.0804     4.1467 754.838   <2e-16 ***
## mom.BMI                  3.4196     0.3922   8.719   <2e-16 ***
## full.termTRUE:mom.BMI    5.2366     0.4202  12.463   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 518.1 on 308021 degrees of freedom
## Multiple R-squared:  0.9756, Adjusted R-squared:  0.9756
## F-statistic: 3.075e+06 on 4 and 308021 DF,  p-value: < 2.2e-16
```

Sources of variation diagram:

Memorizing Levels

1-20

11 - Consider histogram, mean, standard deviation, median

13 - Consider side by side boxplots