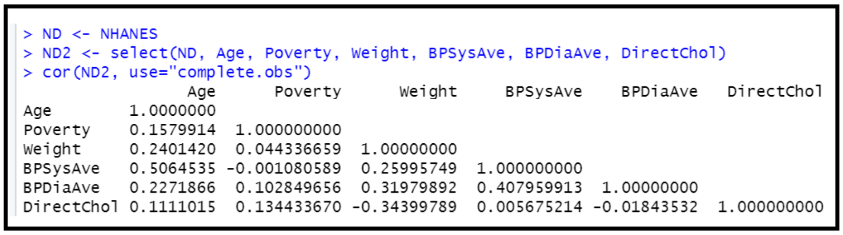
Intro to Stats, Day 12: More About Regression

Name: ____________________________________

Discussing and understanding Regression Towards the Mean

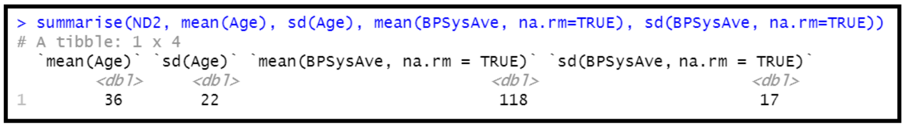
The theory of Regression Towards the Mean states that, if we want to use a x-variable to predict a y variable; and if an individual x-value is (s) standard deviations above the mean (where mean and SD are measured for the x-value); then we would predict the y-value is ($s \times r$) standard deviations above the mean (where mean and SD are measured for the y-value).



```
> ND <- NHANES
> ND2 <- select(ND, Age, Poverty, Weight, BPSysAve, BPDiaAve, DirectChol)
> cor(ND2, use="complete.obs")
                 Age      Poverty      Weight     BPSysAve    BPDiaAve   DirectChol
Age        1.0000000
Poverty    0.1579914  1.000000000
Weight     0.2401420  0.044336659  1.00000000
BPSysAve   0.5064535 -0.001080589  0.25995749  1.000000000
BPDiaAve   0.2271866  0.102849656  0.31979892  0.407959913  1.00000000
DirectChol 0.1111015  0.134433670 -0.34399789  0.005675214 -0.01843532  1.000000000
```

We can create a correlation table for several of our `NHANES` variables, as shown above.

1. Suppose we want to use Age to predict Average Systolic Blood Pressure. This suggests we need to know the correlation coefficient relating the two variables, and also the following information (recorded here):



```
> summarise(ND2, mean(Age), sd(Age), mean(BPSysAve, na.rm=TRUE), sd(BPSysAve, na.rm=TRUE))
# A tibble: 1 x 4
  `mean(Age)` `sd(Age)` `mean(BPSysAve, na.rm = TRUE)` `sd(BPSysAve, na.rm = TRUE)`
        <dbl>     <dbl>                          <dbl>                         <dbl>
1          36        22                            118                            17
```

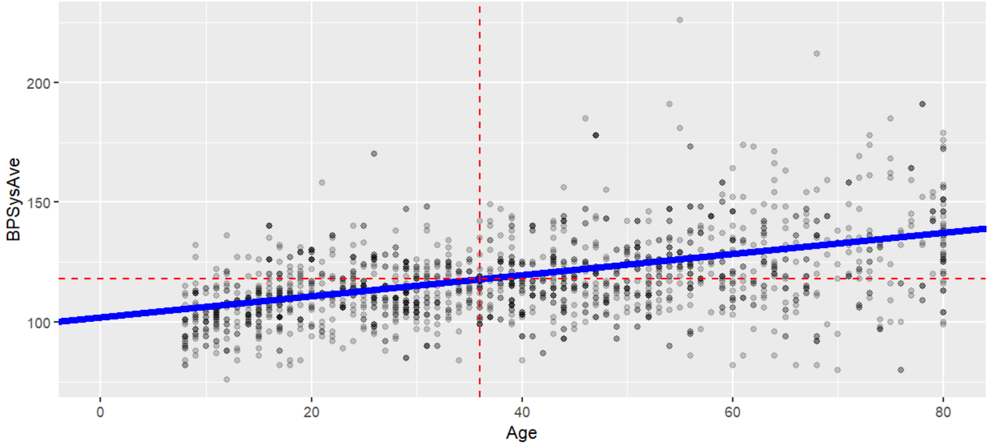
   (a) What is the correlation coefficient r between Age and Average Systolic Blood Pressure (i.e. `BPSysAve`)?

   (b) Suppose an individual is known to be 2 standard deviations above the average age. How many standard deviations above the average would we predict their blood pressure to be?

   (c) What would we predict the blood pressure of a 58-year-old person would be?

   (d) What would we predict the blood pressure of an 80-year-old person would be?

2. Age and BPSysAve are shown in the following scatteplot. Also pictured is the line of best fit (in solid blue), and also vertical and horizontal lines (red, dashed) that correspond to the mean for the x and y variables.



   (a) Try measuring out (on the x-axis) one and two standard deviations above the mean for age. Draw those on the image as vertical lines.

   (b) Similarly, measure up one standard deviation above the mean for BPSysAve. Draw a horizontal line at that height to indicate this.

   (c) Finally, verify visually the rule of regression towards the mean: that 1 and 2 standard deviations above the mean for Age correspond exactly to $1 \times r$ and $2 \times r$ standard deviations above the mean for BPSysAve.

Checking the appropriateness of a linear regression

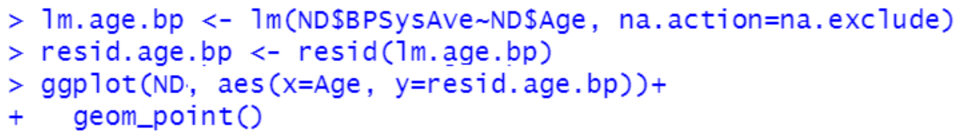
3. Recall that the *residuals* of a regression are defined as

$$\text{residual} = \text{observed} - \text{predicted}.$$

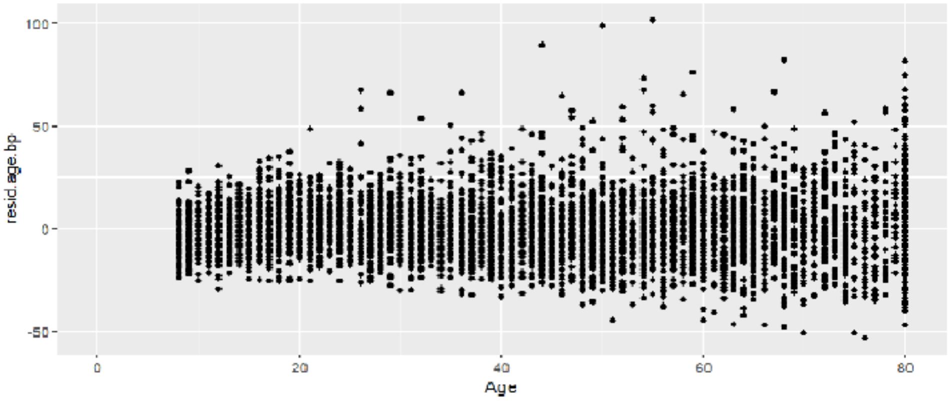
Writing this as a formula (where the letter e is used for the residual), we get

$$e = y - \hat{y}.$$

If our regression line is well-chosen and appropriate, we expect (or hope) that the residuals are relatively scattered, with no strong discernible patterns. For example, here is a "residual" plot for `Age` vs `BPSysAve`. We still plot age on the x-axis, but instead of plotting blood pressure on the y-axis, we plot the residuals on that axis (replacing each point's height with its *residual value*).



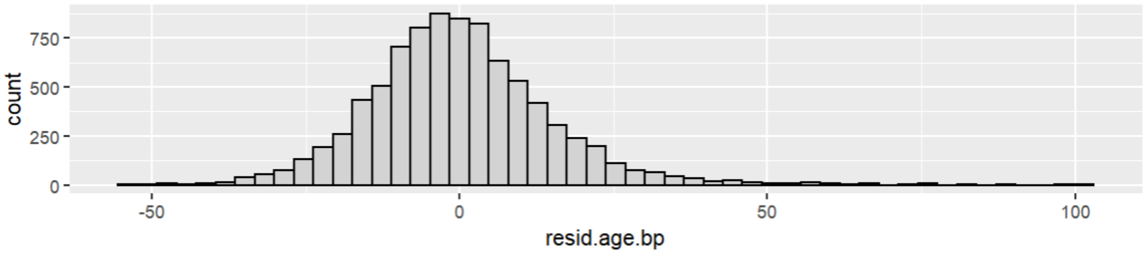
```
> lm.age.bp <- lm(ND$BPSysAve~ND$Age, na.action=na.exclude)
> resid.age.bp <- resid(lm.age.bp)
> ggplot(ND, aes(x=Age, y=resid.age.bp))+
+   geom_point()
```



What observations can we make about the residual plot? Is it boring? (hopefully!)

4. Below, we see the values of the residuals for each data point graphed on a histogram. Describe the shape and the mean of the histogram.

   Note: The standard deviation of the residuals is often called the *standard error*. It describes the amount that the actual data varies from the predicted model line. It's notation is often s_e. In this case, $s_e = 14.8$.

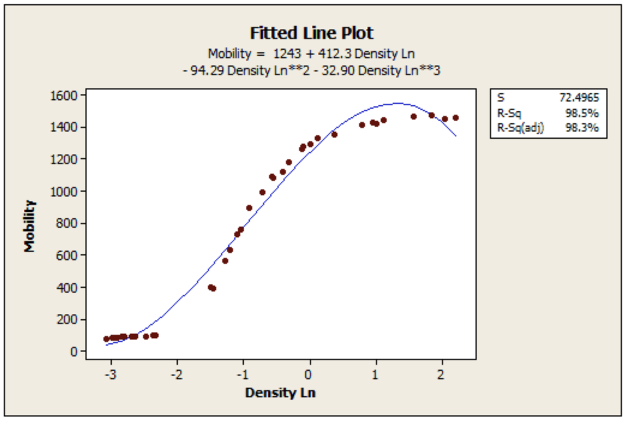


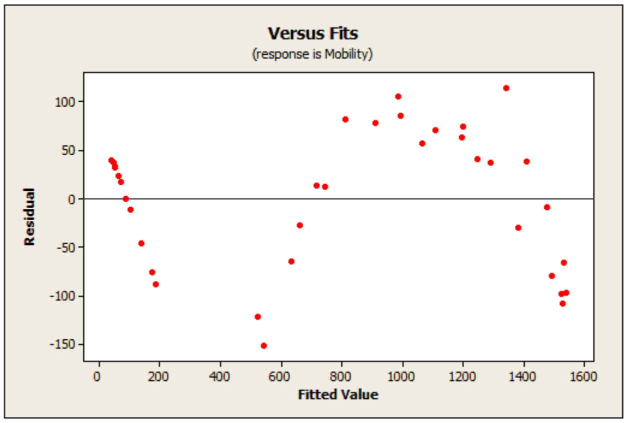
   If our model is a good fit for our data, our residual scatterplot should be boring, and our residual histogram should be unimodal and symmetric, with a mean at 0 (suggesting that there are as many positive residuals as negative residuals). Does that appear to be true in these pictures?

5. All regression models fall somewhere between the two extremes of zero correlation and perfect correlation. We'd often like to gauge how well our model fares. We do this by **taking the square of the correlation** r (which, for some reason, we denote as R^2 with a capital R). The quantity R^2 gives us the proportion (between 0 and 1) of the data's variation that is accounted for by the linear regression model. The quantity $1-R^2$ tells us what proportion of the data's variability is accounted for in the residuals (ie, what proportion fails to be accounted for by the model).

   (a) What proportion of the data's variation in our Blood Pressure example can be attributed to the model, which predicts blood pressure based on age? What proportion can be attributed to the residuals? As a reminder, the correlation coefficient of this data set was approximately 0.5).

6. The graph below compares two variables with a nonlinear relationship. The variables are **semiconductor electron mobility** and **the natural log of the density**.



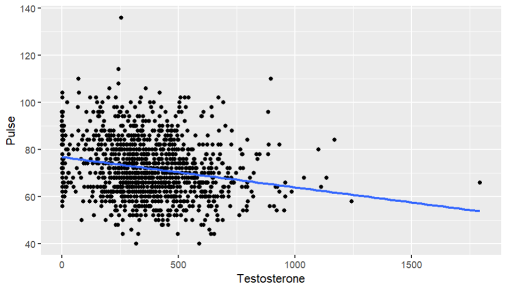


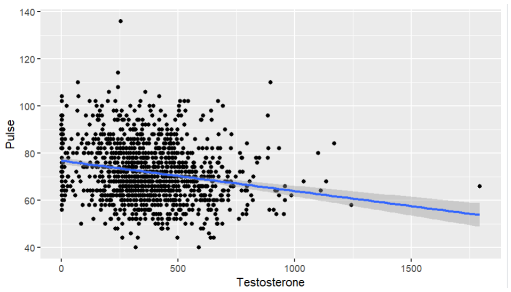
The graph uses a nonlinear regression model[1]. Compare the R^2 value and the residual plot. What can we conclude?

[1] which we won't explore in this class, but is useful for modeling the nonlinear relationships between variables

The se=TRUE option in the Linear Model

Below is a histogram comparing the Testosterone and Pulse for approximately 2000 individuals in the NHANES data set.[2] Two scatterplots and linear regression models are recorded here: one with `se = FALSE` and one with `se = TRUE`.





The second picture, you'll notice, has a shaded "zone" around the regression line. We won't discuss a super specific interpretation of this at the moment, but we will give a brief interpretation here. **We recognize that the individuals we are looking at here are from a sample of the total population, and the precise relationship we see in this sample might differ ever-so-slightly from the broader population. The solid blue line represents the line of best fit for this specific data set; the shaded area around the line represents where we are "reasonably confident" the true population's regression line will fall. It is calculated using the standard error and the residuals, which is why it is denoted as "`se =` " in the command.**[3]

7. In the picture above, we see evidence that we can be "reasonably confident" our linear relationship between testosterone and pulse is negative. How can we explain this?

8. Why do we think the grey zone tend to "flare out" to the far right of the picture?

[2] There are only 1978 men in the data set that had both their testosterone levels and their pulse recorded.

[3] In short: the grey zone represents a confidence interval. We will discuss confidence intervals in greater detail near the end of the semester.

Try your own!

9. Try finding your own interesting linear relationships between variables in either the `NHANES` or `longley` data set. In each case,

   - plot the two quantitative variables on a scatterplot.
   - calculate the linear regression coefficients (using the `lm()` command) and then graph the linear regression (using the `geom_smooth()` or the `geom_abline` commands)
   - Create two graphs of the residuals: the residual scatterplot, and also a histogram of the residual values. Verify whether the residuals are "appropriately boring."
   - Calculate the residual mean and standard deviation (i.e. the standard error).
   - Explain the relationship between the two variables you picked. What is the equation of the linear regression? How can you interpret the slope and intercept? Was the relationship suitably linear, and the residual plot suitably boring?

   Record your results here!