MATH 1051H S61: Lecture 07

Statistics

Transition

In the first few lectures, we discussed some aspects of statistics (data, variable types, sampling concerns, and experimental design). We also touched on some aspects of plotting, which we then used in workshops.

Then we went off on a seeming tangent, and talked about probability for a while!

Now, we're back to statistics, and will stay with it for the rest of the semester.

Data and Modeling

In science, we often obtain data from experiments or observational studies, and then are interested in **modeling** it. How we model it varies from field to field, but underlying many of the models used in science are **statistical models**. In this lecture, we will be examining one of the foundational models used all through every scientific field.

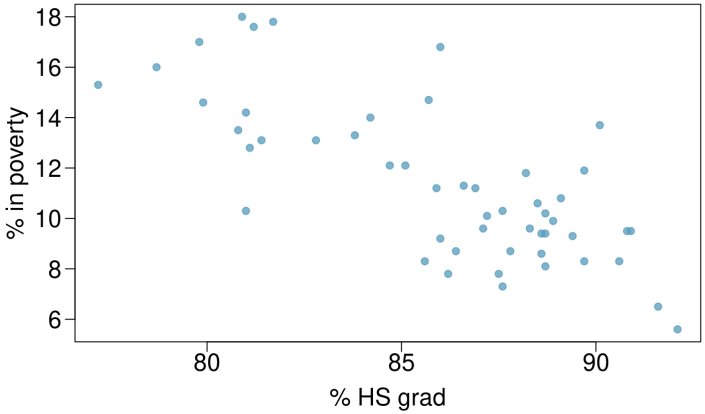
The material for this lecture comes from Chapters 2.1.1 and 8.

Line fitting, residuals, and correlation

Modeling numerical variables

We will begin by quantifying the relationship between two numerical variables, as well as modeling numerical response variables using a numerical or categorical explanatory variable. The model is therefore “find the relationship between two variables”.

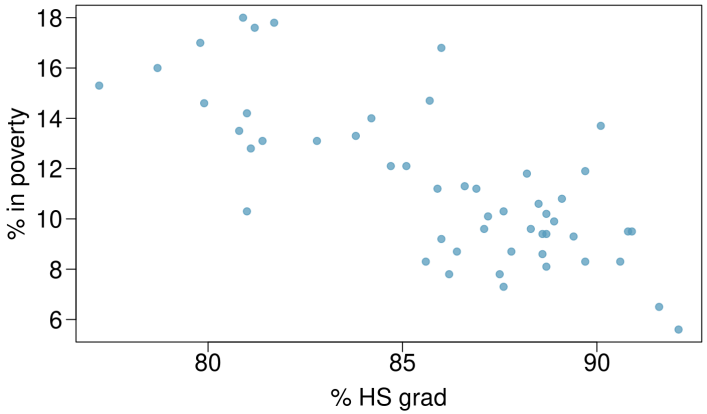
Poverty vs. HS graduate rate



This **scatterplot** shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).

Poverty vs. HS graduate rate (code)

```
par(mar=c(4,4,1,1), las=1, mgp=c(2.5,0.7,0), cex.lab = 1.5, cex.axis = 1.5)
plot(poverty$Poverty ~ poverty$Graduates, ylab = "% in poverty",
     xlab = "% HS grad", pch = 19, col = COL[1,2])
```



Poverty vs. HS graduate rate

- **Response variable?**
- **Explanatory variable?**
- **Relationship?**

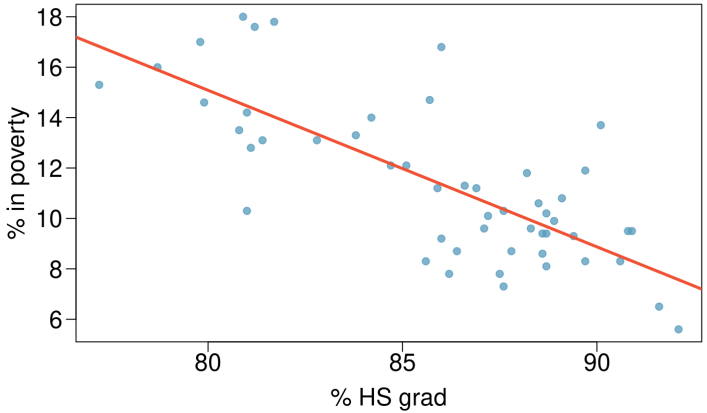
Poverty vs. HS graduate rate

- **Response variable?** % in poverty
- **Explanatory variable?** % HS grad
- **Relationship?** linear, negative, moderately strong

Quantifying the relationship

- **Correlation** describes the strength of the *linear* association between two variables.
- It takes values between -1 (perfect negative) and +1 (perfect positive).
- A value of 0 indicates no linear association.

Guessing the correlation

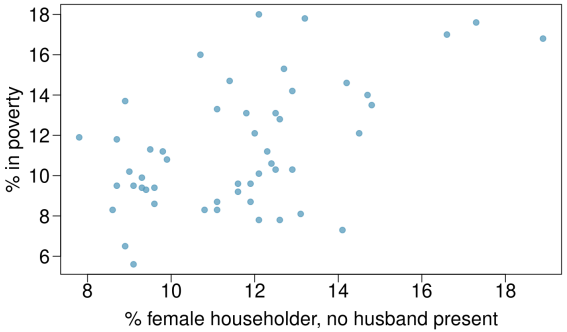


Which of the following is the best guess for the correlation between % in poverty and % HS grad?

0.6 -0.75 -0.1 0.02 -1.5

Guessing the correlation

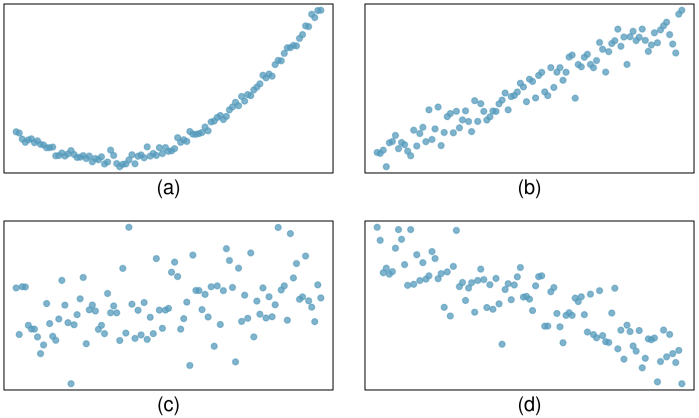
```
# scatterplot, %poverty vs.%no husband
par(mar=c(4,4,1,1), las=1, mgp=c(2.5,0.7,0), cex.lab = 1.5, cex.axis = 1.5)
plot(poverty$Poverty ~ poverty$PercentFemaleHouseholderNoHusbandPresent, ylab = "% in poverty"
```



Which of the following is the best guess for the correlation between % in poverty and % female householder (no husband present)?

0.1 -0.6 -0.4 0.9 0.5

Assessing the correlation



Which of the given plots has the strongest correlation, i.e., correlation coefficient closest to +1 or -1?

Answer

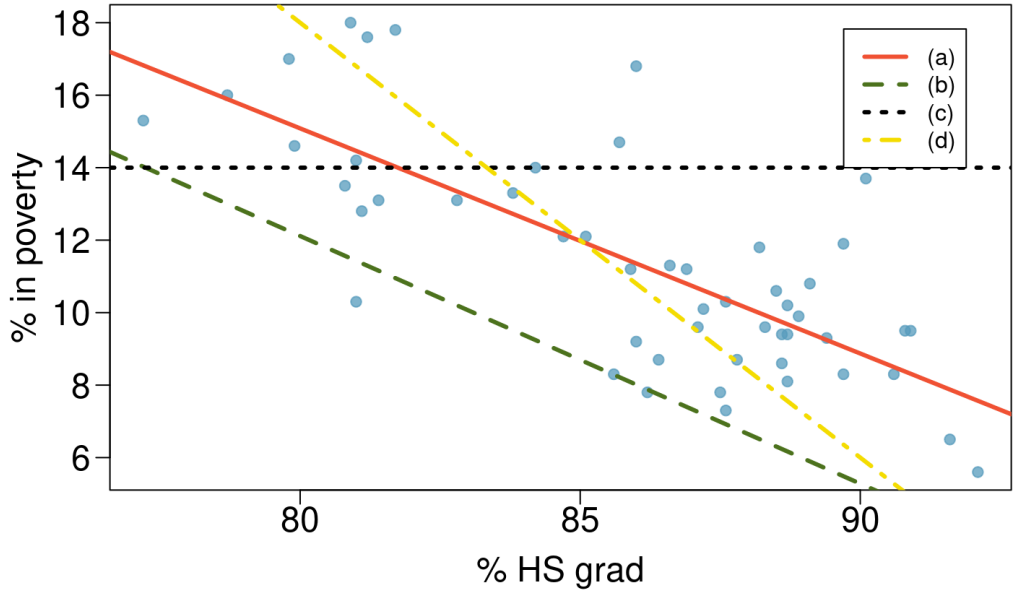
Correlation means **linear** association, and the strongest linear association is Figure (b). We can actually **compute** the correlations, and they are:

- 0.834 (a)
- 0.966 (b)
- 0.357 (c)
- -0.852 (d)

We'll show you how to do this in workshop!

Fitting a line by least squares regression

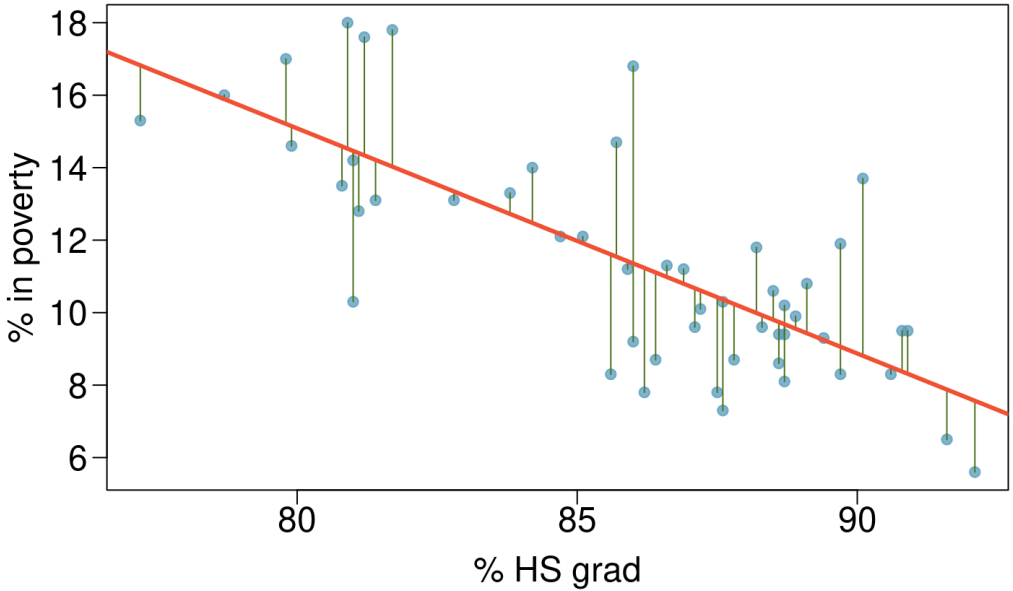
Eyeballing the line



Which of the lines on the figure appears to be the line that best fits the linear relationship between % in poverty and % HS grad? Choose one.

Residuals

Residuals are the leftovers from the model fit: Data = Fit + Residual

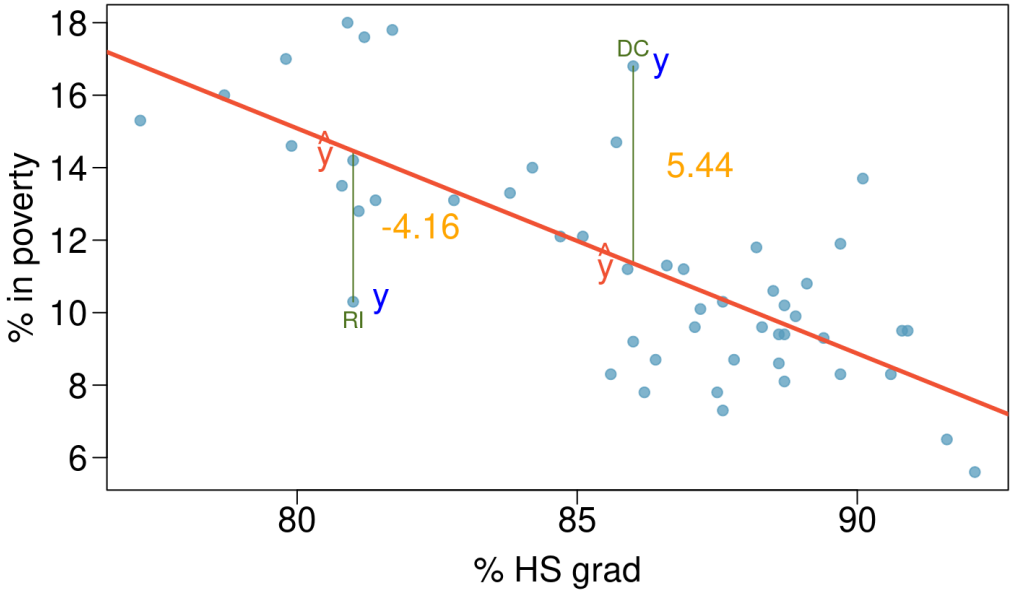


Residuals (cont.)

Formally, residuals are the difference between the observed (y_i) and predicted $\hat{y}_i$.

$$e_i = y_i - \hat{y}_i$$

Specific Residuals



- % living in poverty in DC is 5.44% more than predicted.
- % living in poverty in RI is 4.16% less than predicted.

A measure for the best line

- We want a line that has small residuals:
  - Option 1: Minimize the sum of magnitudes (absolute values) of residuals

$$|e_1| + |e_2| + \cdots + |e_n|$$

  - Option 2: Minimize the sum of squared residuals – **least squares**

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

  - Option 3 … 100: actual topics of research!
- Why least squares?
  - Most commonly used
  - Easier to compute by hand and using software
  - In many applications, a residual twice as large as another is usually more than twice as bad

The least squares line

$$\hat{y} = \beta_0 + \beta_1 x$$

Notation:

- Intercept:
  - Parameter: β_0
  - Point estimate: b_0
- Slope:
  - Parameter: β_1
  - Point estimate: b_1

The least squares line

```
mod <- lm(Poverty ~ Graduates, data = poverty)
summary(mod)
```

```
##
## Call:
## lm(formula = Poverty ~ Graduates, data = poverty)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1624 -1.2593 -0.2184  0.9611  5.4437
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 64.78097    6.80260   9.523 9.94e-13 ***
## Graduates   -0.62122    0.07902  -7.862 3.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.082 on 49 degrees of freedom
## Multiple R-squared:  0.5578, Adjusted R-squared:  0.5488
## F-statistic: 61.81 on 1 and 49 DF,  p-value: 3.109e-10
```

Slope

The **slope** of the regression can be calculated as

$$b_1 = \frac{s_y}{s_x} R$$

but we will just use R, and identify it from the summary table:

```
mod$coefficients
```

```
## (Intercept)   Graduates
##  64.7809658  -0.6212167
```

(so $b_1 = -0.62$)

Interpretation For each additional % point in HS graduate rate, we would expect the % living in poverty to be lower on average by 0.62% points.

Intercept

The **intercept** is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line always passes through $(\bar{x}, \bar{y})$:

$$b_0 = \bar{y} - b_1 \bar{x},$$

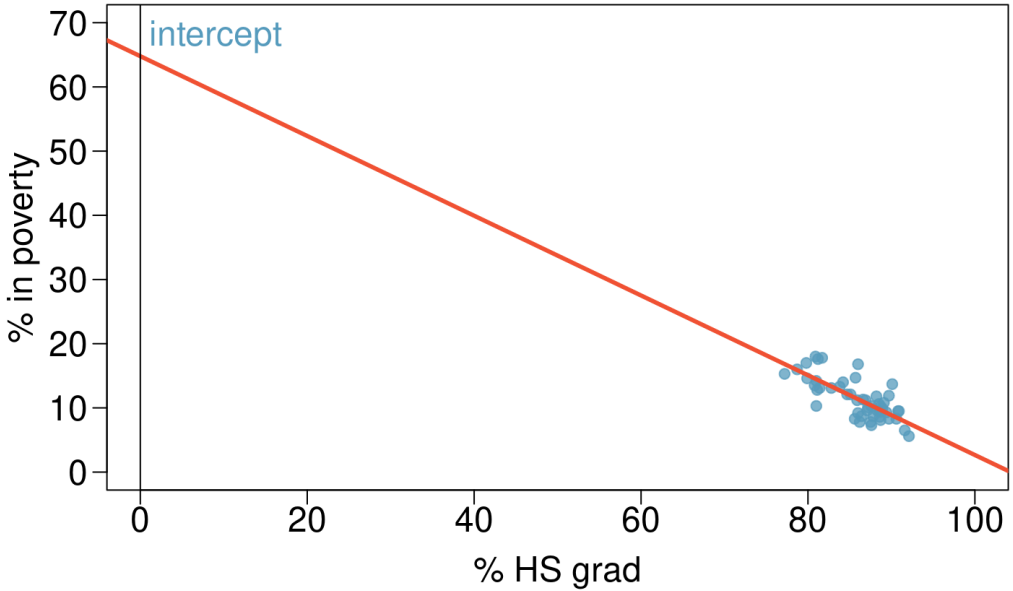
but again, we will just use R and identify it:

```
mod$coefficients
```

```
## (Intercept)   Graduates
##  64.7809658  -0.6212167
```

(so $b_0 = 64.78$).

Intercept


intercept
% in poverty
% HS grad
70
60
50
40
30
20
10
0
0
20
40
60
80
100

Which of the following is the correct interpretation of the intercept?

- For each % point increase in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.
- For each % point decrease in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.
- Having no HS graduates leads to 64.68% of residents living below the poverty line.
- States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line.
- In states with no HS graduates % living in poverty is expected to increase on average by 64.68%.

More on the intercept

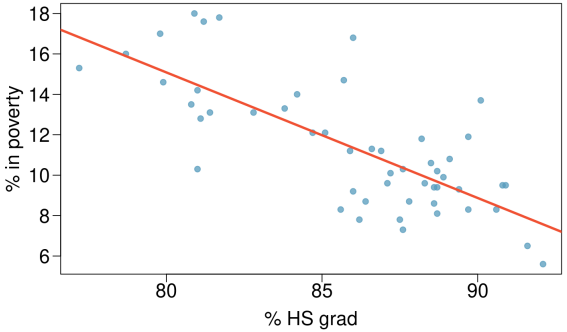
Since there are no states in the dataset with no HS graduates, the intercept is of no interest, not very useful, and also not reliable since the predicted value of the intercept is so far from the bulk of the data.

(see previous figure)

Regression line

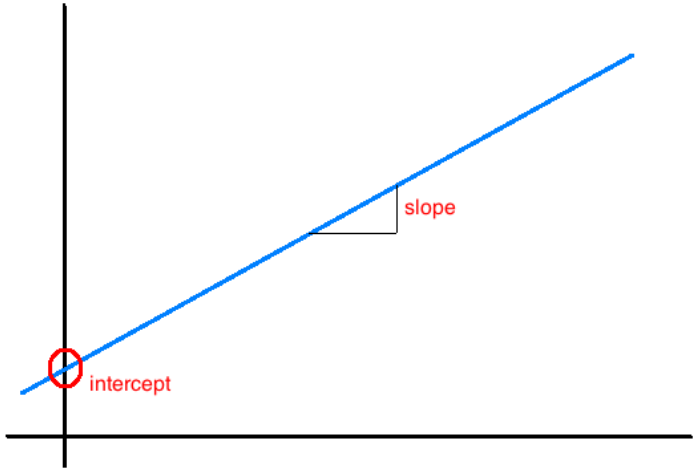
$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ \%\ HS\ grad$$

```
par(mar=c(4,4,1,1), las=1, mgp=c(2.5,0.7,0), cex.lab = 1.5, cex.axis = 1.5)
plot(poverty$Poverty ~ poverty$Graduates, ylab = "% in poverty",
     xlab = "% HS grad", pch=19, col=COL[1,2])
lm_pov_grad = lm(Poverty ~ Graduates, data = poverty)
abline(lm_pov_grad, col = COL[4], lwd = 3)
```



Interpretation of slope and intercept

- **Intercept:** When $x = 0$, y is expected to equal the intercept.
- **Slope:** For each unit in x, y is expected to increase / decrease on average by the slope.

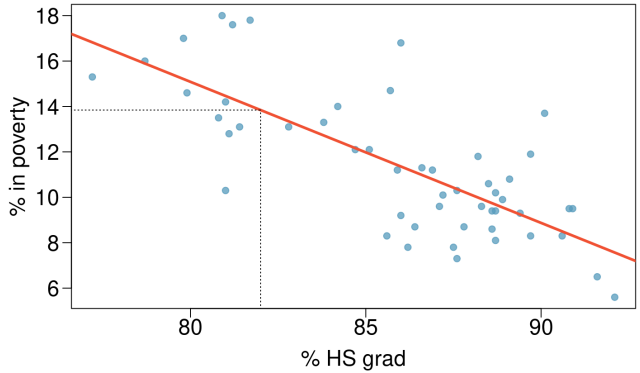


These statements are not causal, unless the study is a randomized controlled experiment.

Prediction & extrapolation

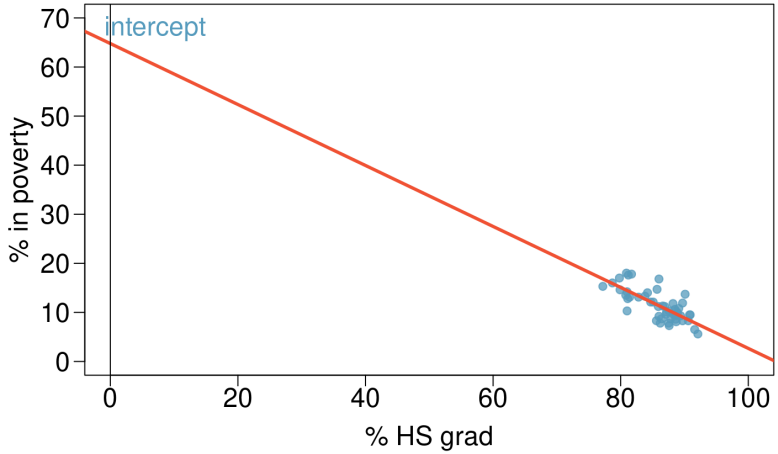
Prediction

- Using the linear model to predict the value of the response variable for a given value of the explanatory variable is called **prediction**, simply by plugging in the value of x in the linear model equation.
- There will be some uncertainty associated with the predicted value.

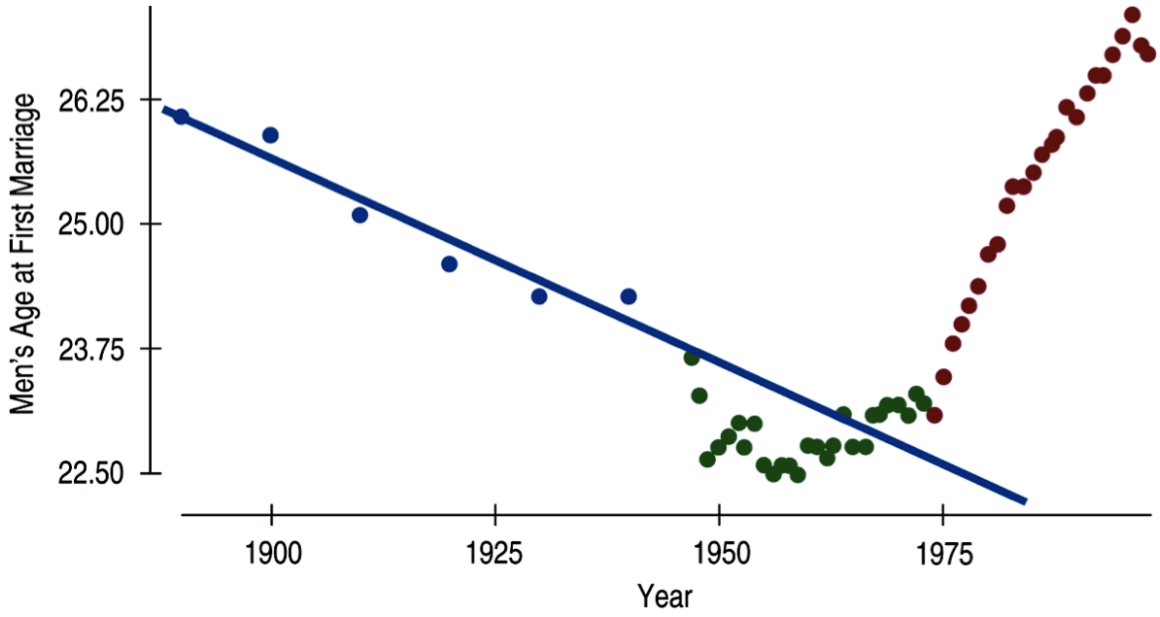


Extrapolation

- Applying a model estimate to values outside of the realm of the original data is called **extrapolation**.
- Sometimes the intercept might be an extrapolation.



Examples of extrapolation



Examples of extrapolation

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

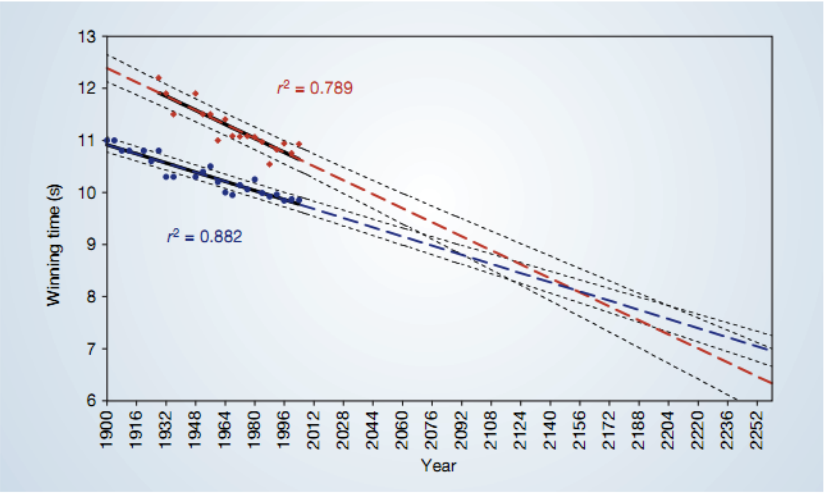


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

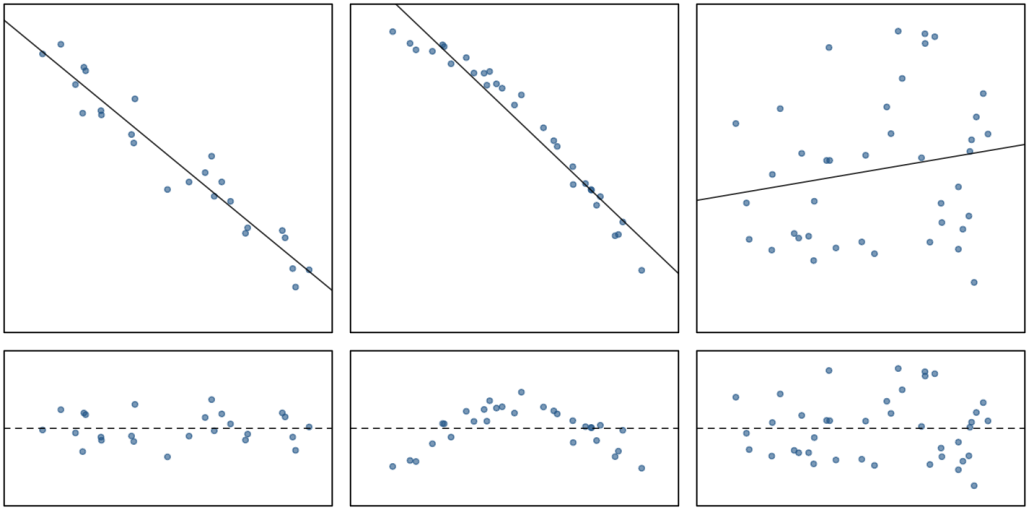
Conditions for the least squares line

- Linearity
- Nearly normal residuals
- Constant variability

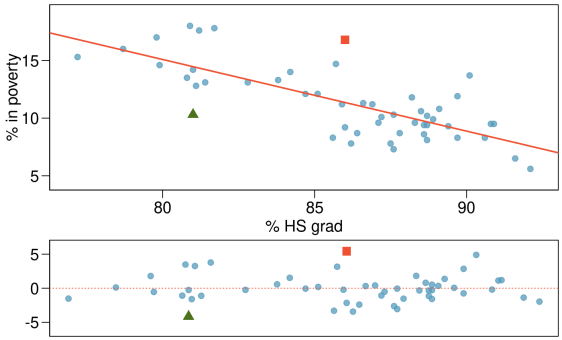
Conditions: (1) Linearity

- The relationship between the explanatory and the response variable should be linear.
- Methods for fitting a model to non-linear relationships exist, but are beyond the scope of this class. One such method will be discussed in MATH 1052H.
- Check using a scatterplot of the data, or a **residuals plot**.

Conditions: (1) Linearity (Examples)



Anatomy of a residuals plot



RI: % HS grade (81), % poverty (10.3)

$$\widehat{\%\ in\ poverty} = 64.68 - 0.62 * 81 = 14.46$$

$$e = \%\ in\ poverty - \widehat{\%\ in\ poverty}$$

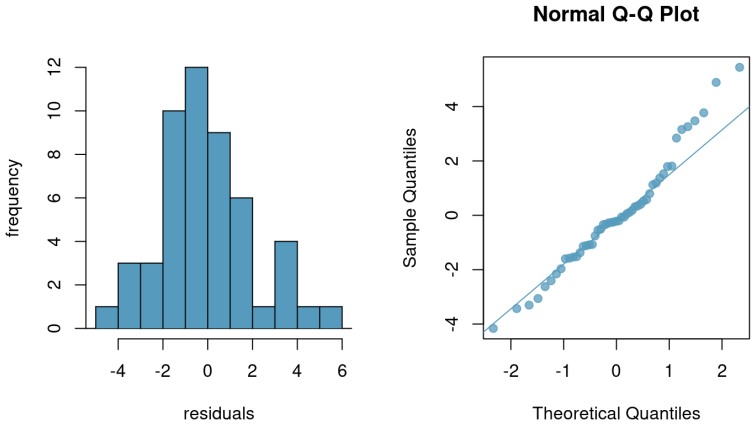
$$e = 10.3 - 14.46 = -4.16$$

Conditions: (2) Nearly normal residuals

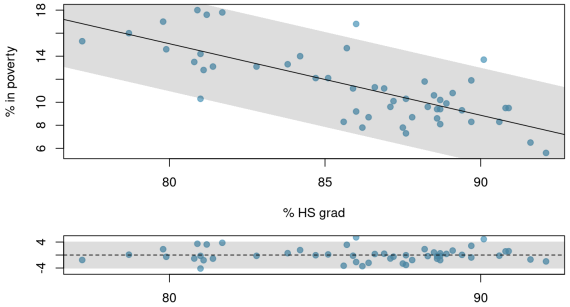
- The residuals should be nearly normal.
- This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.
- Check using a histogram or normal probability plot of residuals.

Conditions: (2) Nearly normal residuals

```
par(mfrow=c(1,2))
histPlot(lm_pov_grad$residuals, col = COL[1], xlab = "residuals")
qqnorm(lm_pov_grad$residuals, pch = 19, col = COL[1,2])
qqline(lm_pov_grad$residuals, pch = 19, col = COL[1])
```

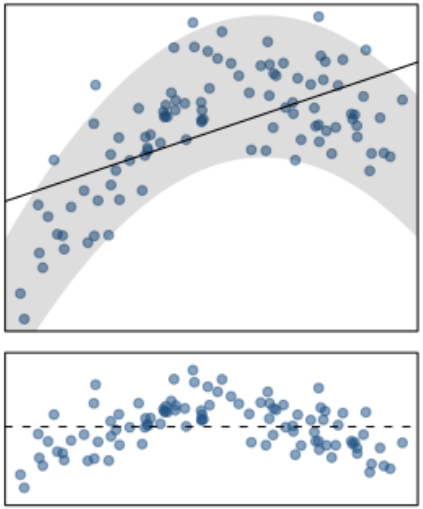


Conditions: (3) Constant variability



- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the 0 line should be roughly constant as well.
- Also called **homoscedasticity**.
- Check using a histogram or normal probability plot of residuals.

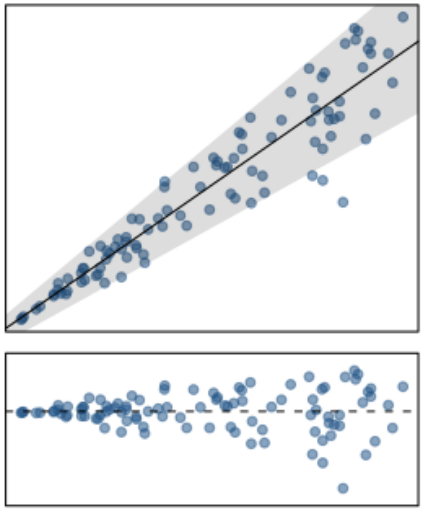
Checking conditions



What condition is this linear model obviously violating?

- Constant variability
- Linear relationship
- Normal residuals
- No extreme outliers

Checking conditions



What condition is this linear model obviously violating?

- Constant variability
- Linear relationship
- Normal residuals
- No extreme outliers

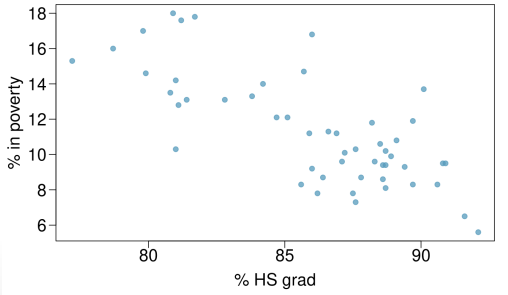
R^2

- The strength of the fit of a linear model is most commonly evaluated using R^2
- R^2 is calculated as the square of the correlation coefficient.
- It tells us what percent of variability in the response variable is explained by the model.
- The remainder of the variability is explained by variables not included in the model or by inherent randomness in the data.
- For the model we've been working with, $R^2 = -0.62^2 = 0.38$.

Interpretation of R^2

Which of the below is the correct interpretation of $R = -0.62$, $R^2 = 0.38$?

- 38% of the variability in the % of HG graduates among the 51 states is explained by the model.
- 38% of the variability in the % of residents living in poverty among the 51 states is explained by the model.
- 38% of the time % HS graduates predict % living in poverty correctly.
- 62% of the variability in the % of residents living in poverty among the 51 states is explained by the model.



To Come

We now have our first statistical model. In the coming lectures, we will talk about what this model means, how we use it, and how we can determine uncertainty in the model (since it is statistical, after all!).