Lecture 21 - Simple Linear Regression

Sta 111

Colin Rundel

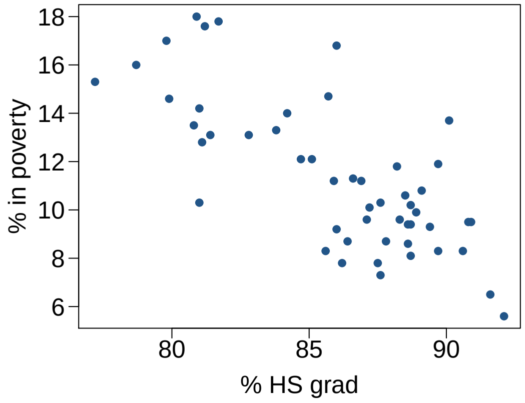
June 13, 2014

Modeling numerical variables

- So far we have worked with single numerical and categorical variables, and explored relationships between numerical and categorical, and two categorical variables.
- Today we will discuss how to quantify the relationship between two numerical variables, as well as modeling numerical response variables using a numerical or categorical explanatory variable.
- Next week we will learn to model numerical variables using many explanatory variables at once.

Poverty vs. HS graduate rate

The *scatterplot* below 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).



Response?

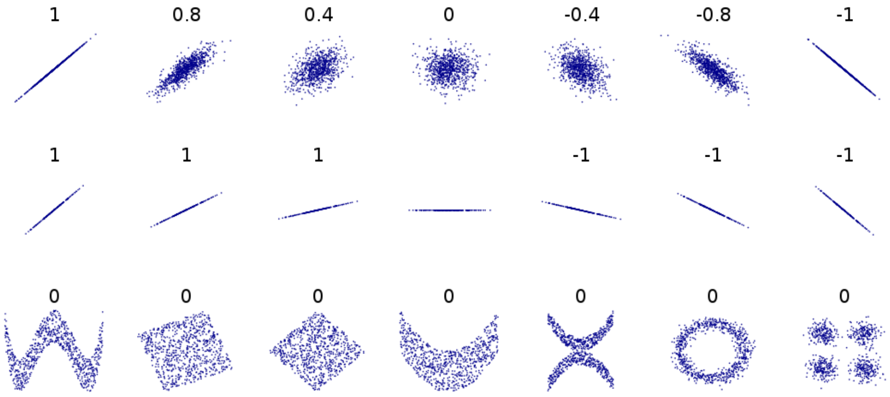
Explanatory?

Relationship?

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.
- We use ρ to indicate the population correlation coefficient, and R or r to indicate the sample correlation coefficient.

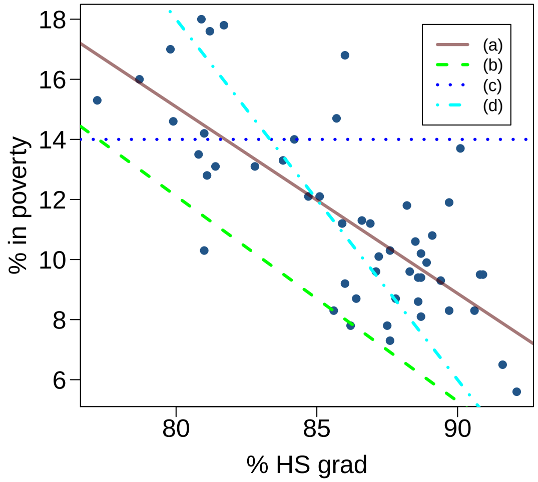
Correlation Examples



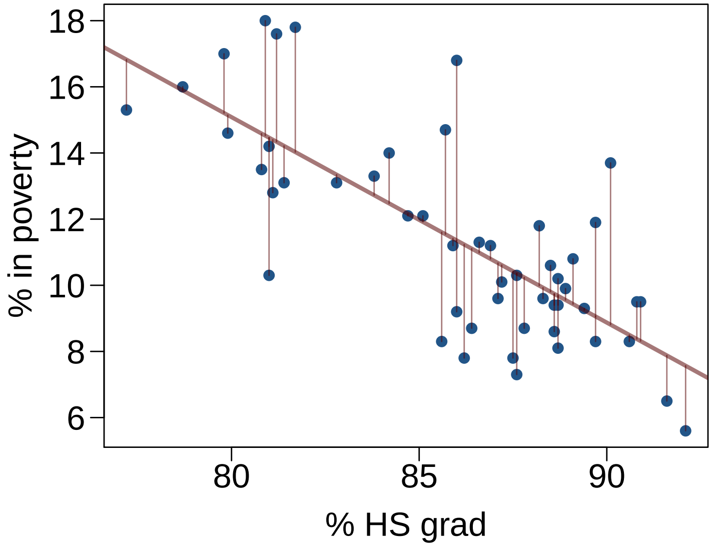
From http://en.wikipedia.org/wiki/Correlation

Eyeballing the line

Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad?



Quantifying best fit

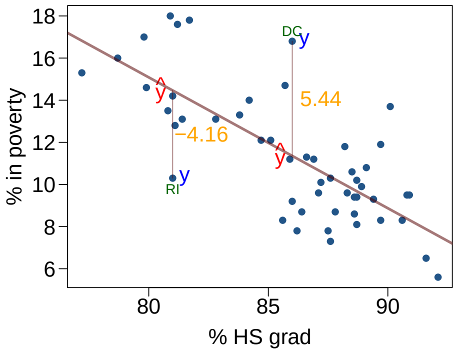


Residuals

Residual

Residual is the difference between the observed and predicted y.

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



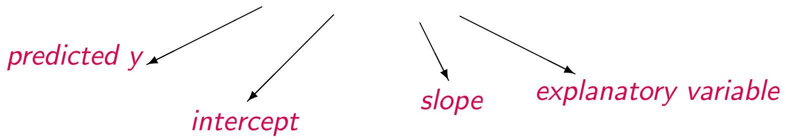
- % living in poverty in DC is 5.44% more than predicted.
- % living in poverty in RI is 4.16% less than predicted.

A measure for the best line

- We want a line that has small residuals:
  1. Option 1: Minimize the sum of magnitudes (absolute values) of residuals
$$|e_1| + |e_2| + \cdots + |e_n|$$
  2. Option 2: Minimize the sum of squared residuals – *least squares*
$$e_1^2 + e_2^2 + \cdots + e_n^2$$
- Why least squares?
  1. Most commonly used
  2. Easier to compute by hand and using software
  3. In many applications, a residual twice as large as another is more than twice as bad

The least squares line

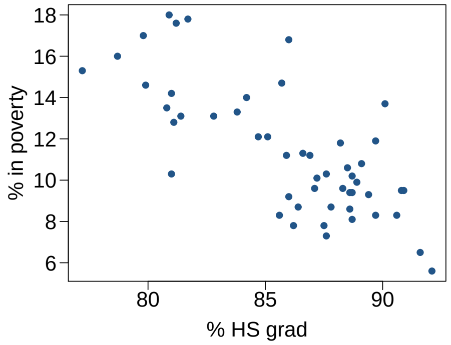
$$\hat{y} = \beta_0 + \beta_1 x$$



Notation:

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

Given...



	% HS grad (x)	% in poverty (y)
mean	$\bar{x} = 86.01$	$\bar{y} = 11.35$
sd	$s_x = 3.73$	$s_y = 3.1$
	correlation	$R = -0.75$

Find b_0 and b_1 that minimize,

$$\sum_i (\hat{y}_i - y_i)^2 = \sum_i (b_0 + b_1\, x_i - y_i)^2$$

Slope

The slope value that minimizes the sum of square residuals is

$$b_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{Cov(X,Y)}{Var(X)} = \frac{s_y}{s_x} R$$

In context...

$$b_1 = \frac{3.1}{3.73} \times -0.75 = -0.62$$

Interpretation

For each % point increase in HS graduate rate, we would *expect* the % living in poverty to decrease *on average* by 0.62% points.

Intercept

The intercept value that minimizes the sum of square residuals is

$$b_0 = \frac{\sum_i y_i - b_1 \sum_i x_i}{n} = \bar{y} - b_1 \bar{x}$$

In context...

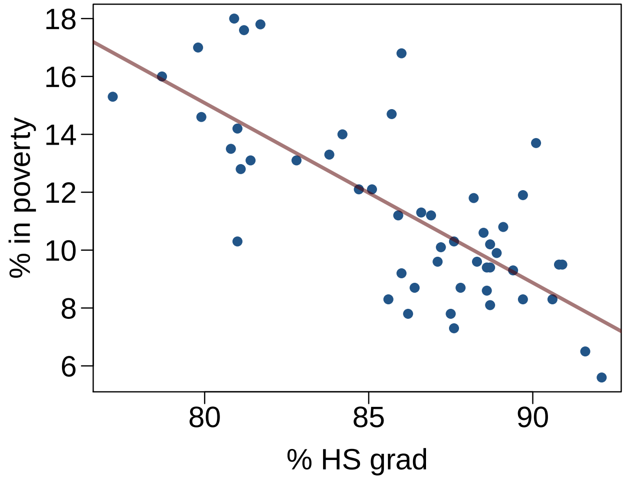
$$b_0 = 11.35 - (-0.62) \times 86.01 = 64.68$$

Interpretation

States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line.

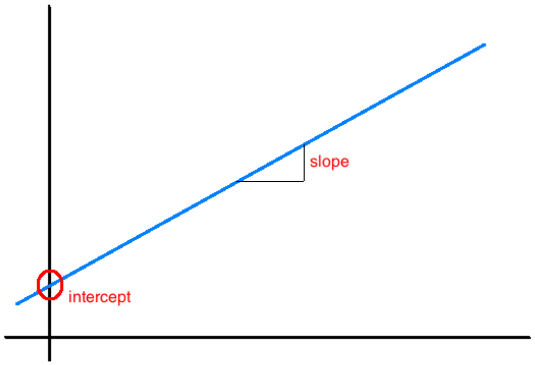
Regression line

$$\hat{y} = 64.68 - 0.62\ x$$



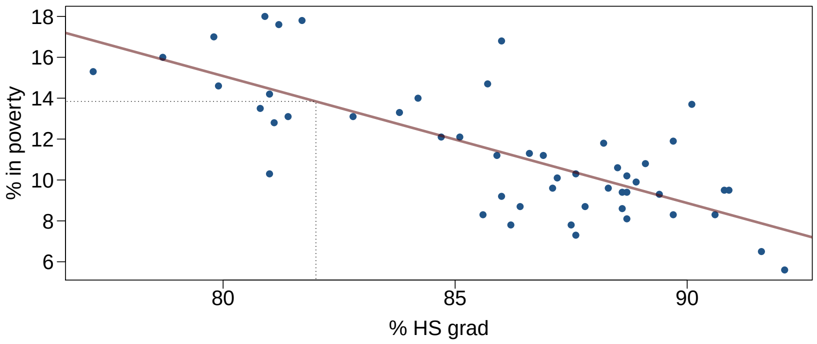
Interpretation of slope and intercept

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



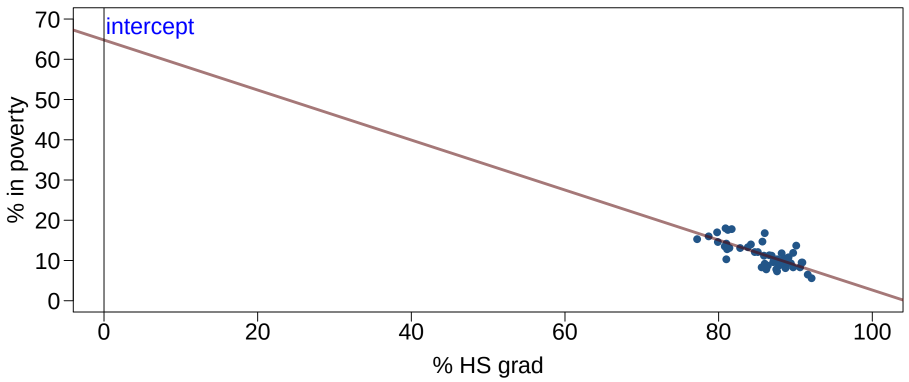
Prediction

- 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 - we'll talk about this in a little bit.

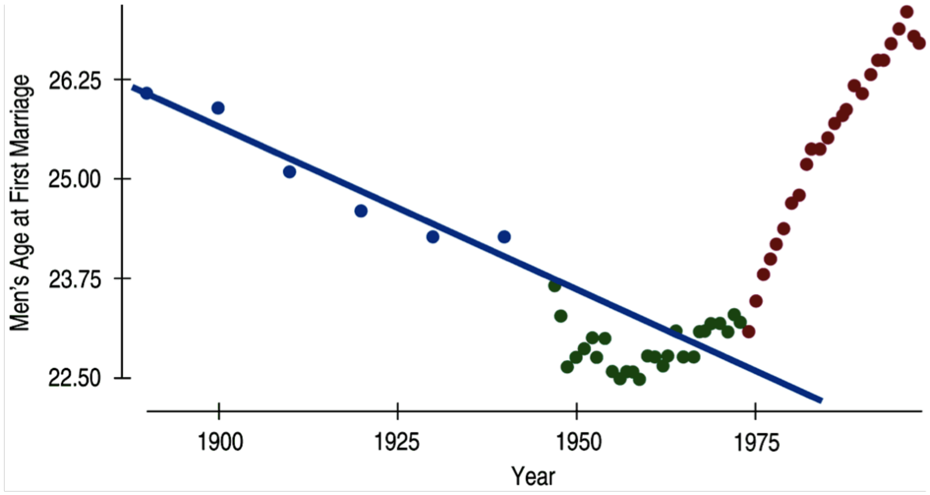


Extrapolation

- 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.

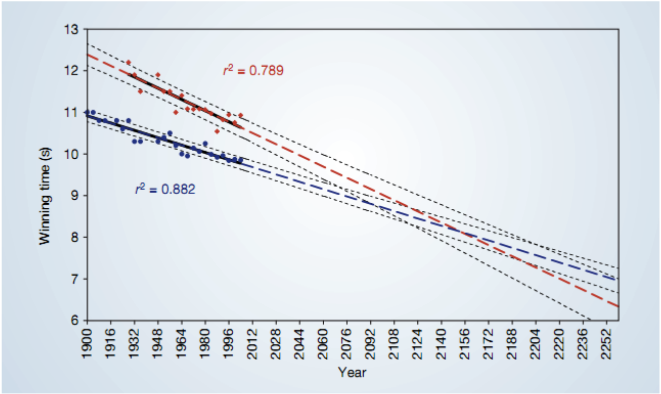


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and

sect 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.

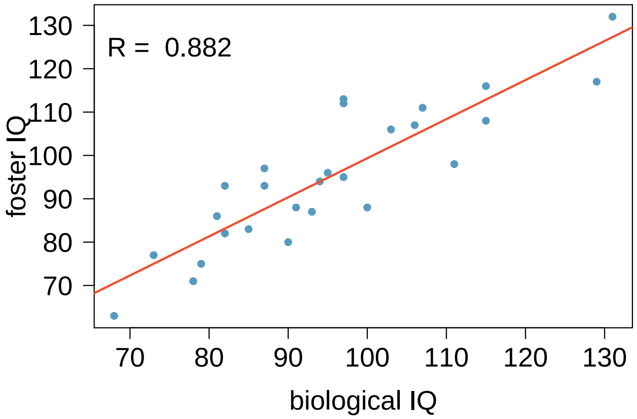
R^2

- 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.
- Sometimes called the coefficient of determination.
- For the model we've been working with, $R^2 = -0.62^2 = 0.38$.

38% of the variability in the % of residents living in poverty among the 51 states is explained by the model.

Nature vs. nurture?

In 1966 Cyril Burt published a paper called “The genetic determination of differences in intelligence: A study of monozygotic twins reared apart” The data consist of IQ scores for [an assumed random sample of] 27 identical twins, one raised by foster parents, the other by the biological parents.



Regression Output

```
summary(lm(twins$Foster ~ twins$Biological))

## Call:
## lm(formula = twins$Foster ~ twins$Biological)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3512  -5.7311   0.0574   4.3244  16.3531
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.20760    9.29990   0.990    0.332
## twins$Biological  0.90144    0.09633   9.358  1.2e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.729 on 25 degrees of freedom
## Multiple R-squared: 0.7779, Adjusted R-squared: 0.769
## F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

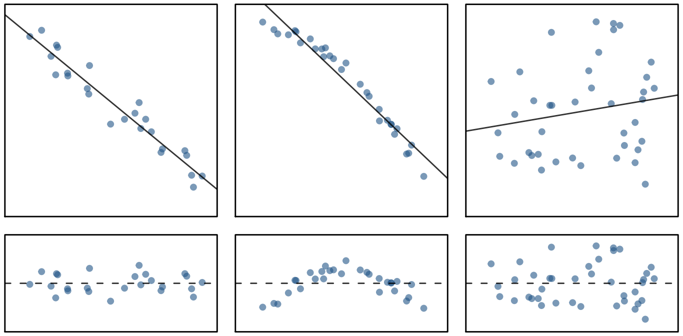
Conditions for inference

In order to perform inference, the following conditions must be met:

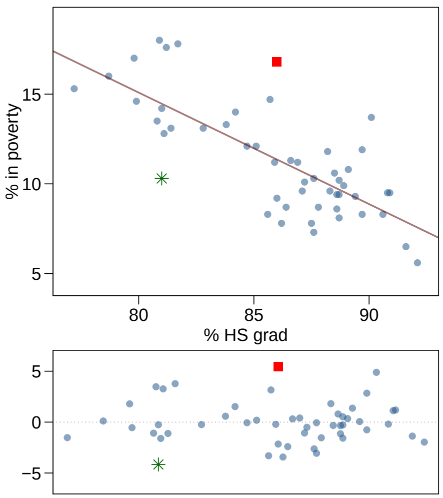
1. Linearity
2. Nearly normal residuals
3. Constant variability

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.
- Check using a *scatterplot* or a *residual plot*.



Anatomy of a residuals plot



✳ *Rhode Island:*

$$\%\ HS\ grad = 81 \qquad \%\ in\ poverty = 10.3$$

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

$$e_{RI} = \%\ in\ poverty - \widehat{\%\ in\ poverty} = 10.3 - 14.46 = -4.16$$

■ *Washington, DC:*

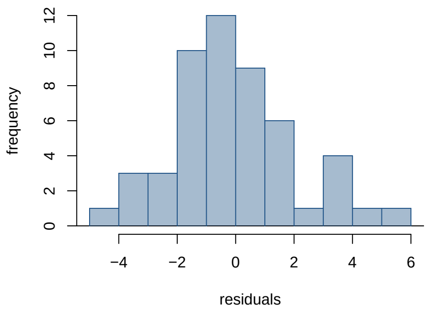
$$\%\ HS\ grad = 86 \qquad \%\ in\ poverty = 16.8$$

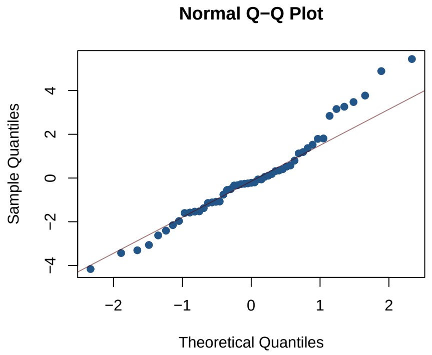
$$\widehat{\%\ in\ poverty} = 64.68 - 0.62 * 86 = 11.36$$

$$e_{DC} = \%\ in\ poverty - \widehat{\%\ in\ poverty} = 16.8 - 11.36 = 5.44$$

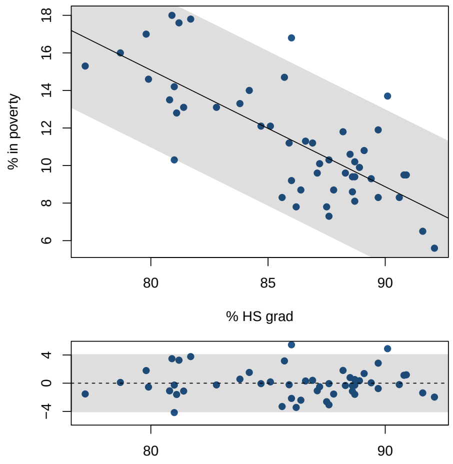
Conditions: (2) Nearly normal residuals

- 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.
- Checked using a histogram or normal probability plot of residuals.





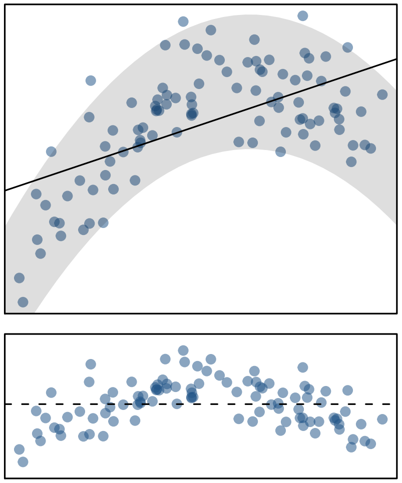
Conditions: (3) Constant variability



- 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 residuals plot.

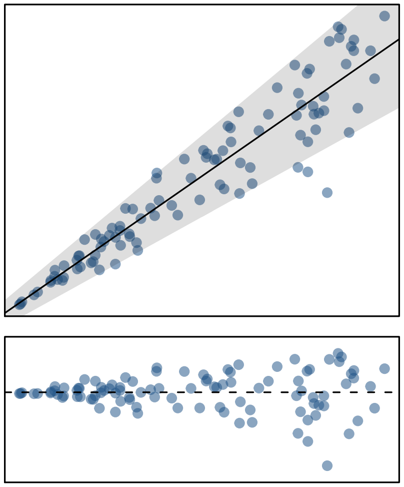
Checking conditions

What condition is this linear model violating?

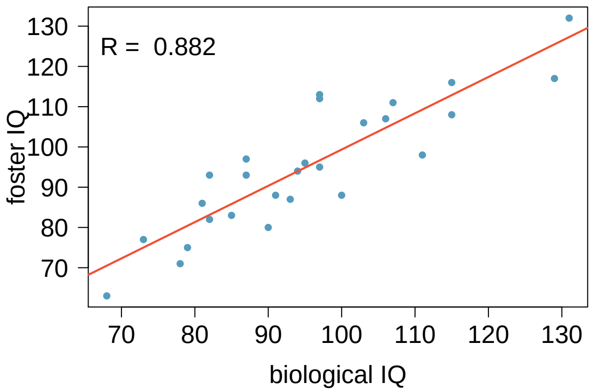


Checking conditions (II)

What condition is this linear model obviously violating?



Back to Nature vs nurture



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

Testing for the slope

Assuming that these 27 twins comprise a representative sample of all twins separated at birth, we would like to test if these data provide convincing evidence that the IQ of the biological twin is a significant predictor of IQ of the foster twin.

What are the appropriate hypotheses?

First consider what the null hypothesis should be, if there is no relationship between the two variables what value of the slope would we expect to see?

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

- We always use a *t*-test in inference for regression parameters.
  Remember: Test statistic, $T = \frac{point\ estimate - null\ value}{SE}$
- Point estimate $= b_1$ is the observed slope.
- SE_{b_1} is the standard error associated with the slope.
- Degrees of freedom associated with the slope is $df = n - k$, where n is the sample size and k is the number of parameters being estimated (k=2 here).
  Remember: We lose 1 degree of freedom for each parameter we estimate, and in simple linear regression we estimate 2 parameters, β_0 *and* β_1*.*

Testing for the slope (cont.)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	*0.9014*	*0.0963*	*9.36*	0.0000

$$s_Y = 16.082 \quad s_X = 15.735 \quad R = 0.882$$

$$\begin{aligned}
b_1 &= \frac{s_Y}{s_X} R = \frac{16.082}{15.735}\, 0.882 = 0.9014 \\
SE_{b_1} &= \sqrt{\frac{\frac{1}{n-2}\sum(\hat{y}_i - y_i)^2}{\sum_i (x_i - \bar{x})^2}} = 0.0963 \\
T &= \frac{0.9014 - 0}{0.0963} = 9.36 \\
df &= 27 - 2 = 25 \\
p - value &= P(|T| > 9.36) < 0.01
\end{aligned}$$

Confidence interval for the slope

Remember that a confidence interval is calculated as *point estimate* ± *ME* and the degrees of freedom associated with the slope in a simple linear regression is $n - 2$. What is the correct 95% confidence interval for the slope parameter? Note that the model is based on observations from 27 twins.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2076	9.2999	0.99	0.3316
bioIQ	0.9014	0.0963	9.36	0.0000

Recap

- Inference for the slope for a SLR model (only one explanatory variable):
  - Hypothesis test:

$$T = \frac{b_1 - null\ value}{SE_{b_1}} \qquad df = n - 2$$

  - Confidence interval:

$$b_1 \pm t^{\star}_{df=n-2} \times SE_{b_1}$$

- The null value is often 0 since we are usually checking for *any* relationship between the explanatory and the response variable.
- The regression output gives b_1, SE_{b_1}, and *two-tailed* p-value for the t-test for the slope where the null value is 0.
- We rarely do inference on the intercept, so we'll be focusing on the estimates and inference for the slope.

Caution

- Always be aware of the type of data you're working with: random sample, non-random sample, or population.
- Statistical inference, and the resulting p-values, are meaningless when you have population data.
- If you have a sample that is non-random (biased), the results will be unreliable.
- The ultimate goal is to have independent observations – and you know how to check for those by now.

Confidence intervals for average values

A confidence interval for the average (expected) value of y for a given $x^\star$, is given by

$$\hat{y} \pm t^\star_{n-2} s_e \sqrt{\frac{1}{n} + \frac{1}{n-1}\frac{(x^\star - \bar{x})^2}{s_x^2}}$$

where s_e is the standard deviation of the residuals

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-1}}$$

Note that when $x^\star = \bar{x}$ this reduces to

$$\hat{y} \pm t^\star_{n-2} \frac{s_e}{\sqrt{n}}$$

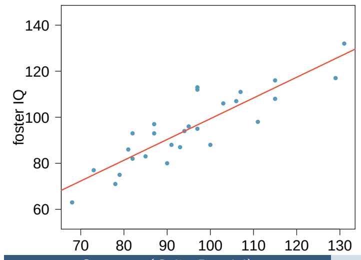
Example Calculation

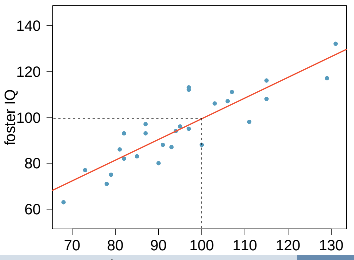
Calculate a 95% confidence interval for the average IQ score of foster twins whose biological twins have IQ scores of 100 points. Note that the average IQ score of 27 biological twins in the sample is 95.3 points, with a standard deviation is 15.74 points.

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.20760    9.29990   0.990    0.332
bioIQ            0.90144    0.09633   9.358 3.22e-09

Residual standard error: 7.729 on 25 degrees of freedom
```

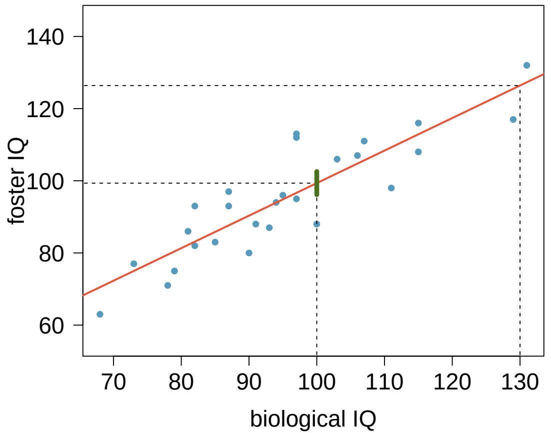
$$\begin{aligned} \hat{y} &= 9.2076 + 0.90144 \times 100 \approx 99.35 \\ df &= n-2 \qquad t^\star = 2.06 \\ ME &= 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(100-95.3)^2}{26 \times 15.74^2}} \\ &\approx 3.2 \\ CI &= 99.35 \pm 3.2 \\ &= (96.15, 102.55) \end{aligned}$$





Distance from the mean

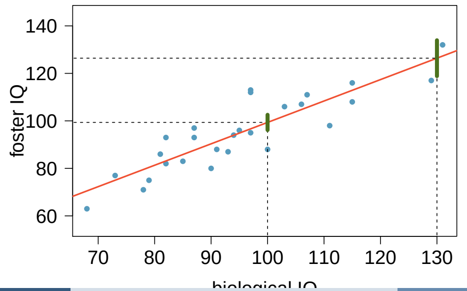
How would you expect the width of the 95% confidence interval for the average IQ score of foster twins whose biological twins have IQ scores of 130 points ($x^\star = 130$) to compare to the previous confidence interval (where $x^\star = 100$)?



How do the confidence intervals where $x^\star = 100$ and $x^\star = 130$ compare in terms of their widths?

$$x^\star = 100 \qquad ME_{100} = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(100 - 95.3)^2}{26 \times 15.74^2}} = 3.2$$

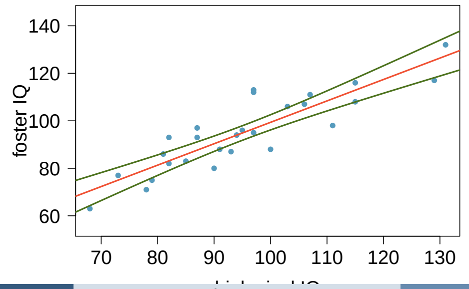
$$x^\star = 130 \qquad ME_{130} = 2.06 \times 7.729 \times \sqrt{\frac{1}{27} + \frac{(130 - 95.3)^2}{26 \times 15.74^2}} = 7.53$$



Recap

The width of the confidence interval for $E(y)$ increases as $x^\star$ moves away from the center.

- Conceptually: We are much more certain of our predictions at the center of the data than at the edges (and our level of certainty decreases even further when predicting outside the range of the data – extrapolation).
- Mathematically: As $(x^\star - \bar{x})^2$ term increases, the margin of error of the confidence interval increases as well.



Predicting a value, not an average

Earlier we learned how to calculate a confidence interval for average y, $E(y)$, for a given $x^\star$.

Suppose that we are not interested in the average, but instead we want to predict a future value of y for a given $x^\star$.

Would you expect there to be more uncertainty around an average or a specific predicted value?

Prediction intervals

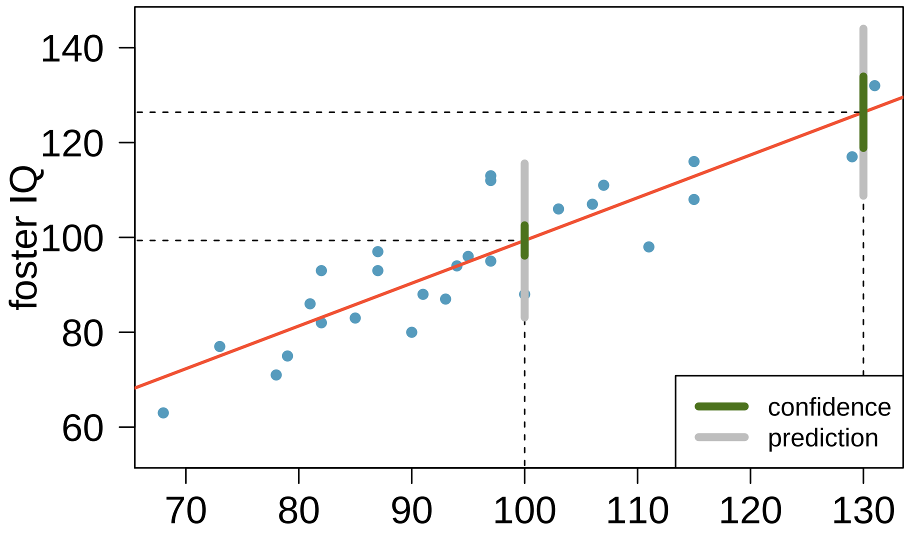
A *prediction interval* for y for a given $x^\star$ is

$$\hat{y} \pm t^\star_{n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{(n-1)s_x^2}}$$

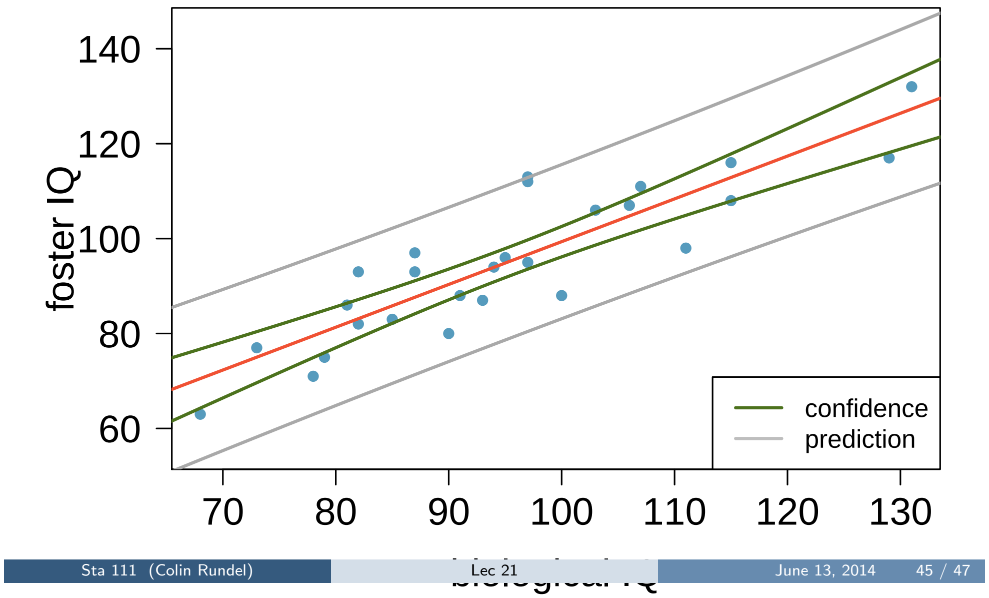
where s is the standard deviation of the residuals.

- The formula is very similar, except the variability is higher since there is a 1 added in the formula.
- Prediction level: If we repeat the study of obtaining a regression data set many times, each time forming a XX% prediction interval at $x^\star$, and wait to see what the future value of y is at $x^\star$, then roughly XX% of the prediction intervals will contain the corresponding actual value of y.

Confidence interval for $E(y)$ vs. prediction interval for y



CI for $E(y)$ vs. PI for y



CI for $E(y)$ vs. PI for y - differences

- A prediction interval is similar in spirit to a confidence interval, except that
  - the prediction interval is designed to cover a “moving target”, the random future value of y
  - the confidence interval is designed to cover the “fixed target”, the average (expected) value of y, $E(y)$,
- Although both are centered at $\hat{y}$, the prediction interval is wider than the confidence interval, for a given $x^\star$ and confidence level. This makes sense, since
  - the prediction interval must take account of the tendency of y to fluctuate from its mean value
  - the confidence interval simply needs to account for the uncertainty in estimating the mean value.

CI for $E(y)$ vs. PI for y - similarities

- For a given data set, the error in estimating $E(y)$ and $\hat{y}$ grows as $x^\star$ moves away from $\bar{x}$. Thus, the further $x^\star$ is from $\bar{x}$, the wider the confidence and prediction intervals will be.
- If any of the conditions underlying the model are violated, then the confidence intervals and prediction intervals may be invalid as well. This is why it’s so important to check the conditions by examining the residuals, etc.