Linear Regression

Fitting a line

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

predicted y

intercept

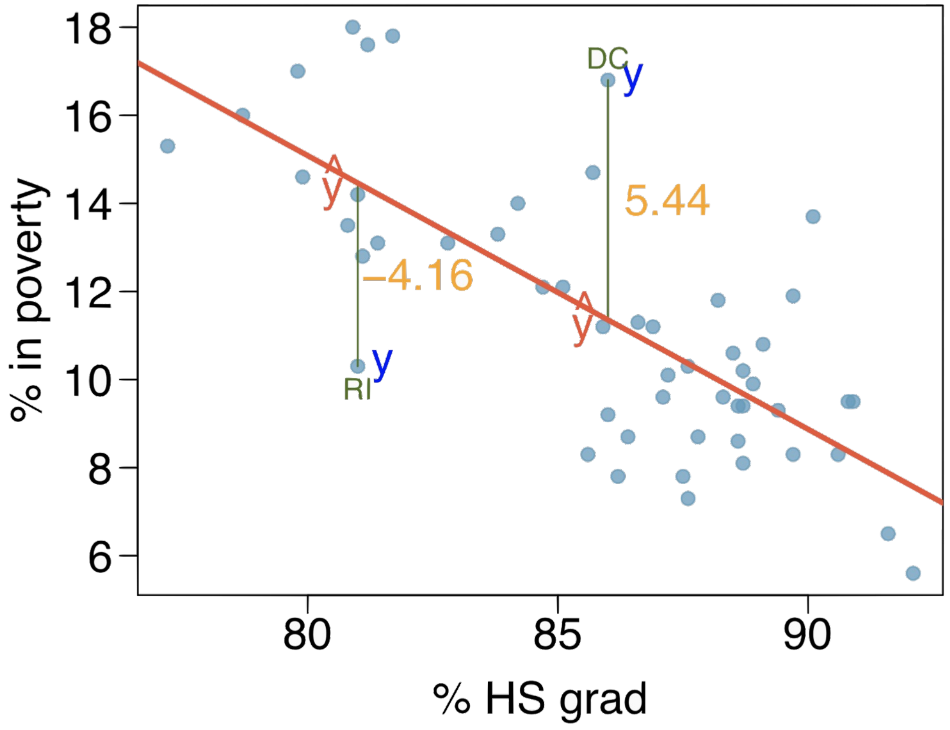
slope

explanatory variable

Choosing a line

Residual is the difference between the observed (y_i) and predicted $\hat{y}_i$.

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



% 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
     $$|e_1| + |e_2| + \ldots + |e_n|$$
  2. Option 2: Minimize the sum of squared residuals -- *least squares*
     $$e_1^2 + e_2^2 + \ldots + e_n^2$$

- 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 usually more than twice as bad

Finding the least squares line

Find b_0, b_1 that minimize the sum of squared residuals

$$RSS = \sum_i (\hat{y}_i - y_i)^2$$

To compute the distribution of the estimators $b_0 = \widehat{\beta_0}, b_1 = \widehat{\beta_1}$ we need to make some assumptions.

Slope

The slope of the regression can be calculated as

$$\widehat{\beta_1} = b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Slope

The slope of the regression can be calculated as

$$\widehat{\beta_1} = b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} =$$

$$= \frac{\sqrt{\sum(y_i - \bar{y})^2/n}}{\sqrt{\sum(x_i - \bar{x})^2/n}} \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \frac{s_y}{s_x} R$$

Interpretation: How many standard deviations do you expect y to change, if you increase x by one standard deviation.

Intercept

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

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

Interpretation: Average y for $x = 0$.

Linear Regression as conditional distribution of Y|X=x

If the values y_i correspond to a random variable Y, then we are interested in describing the conditional distribution of Y given the values of the predictors x.

Then $y_i = \ \beta_0 + \beta_1 x_i + \epsilon_i$

$$f(Y|X = x) \sim ?$$

Conditions for the least squares line

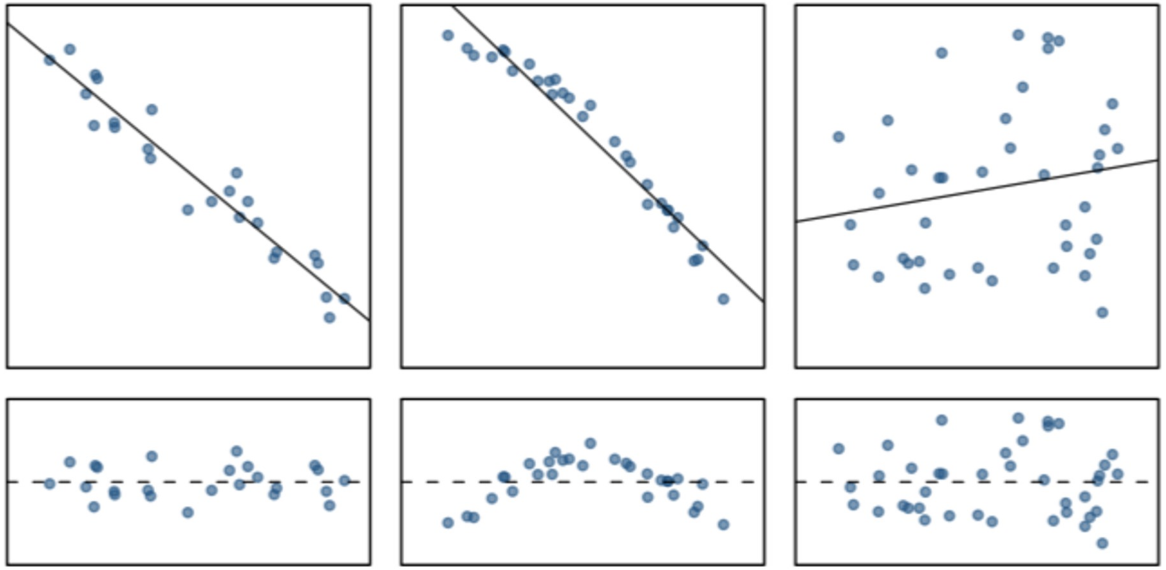
1. Linearity
2. Normality
3. Constant variance (homoskedasticity)
4. Independence

Conditions: (1) Linearity

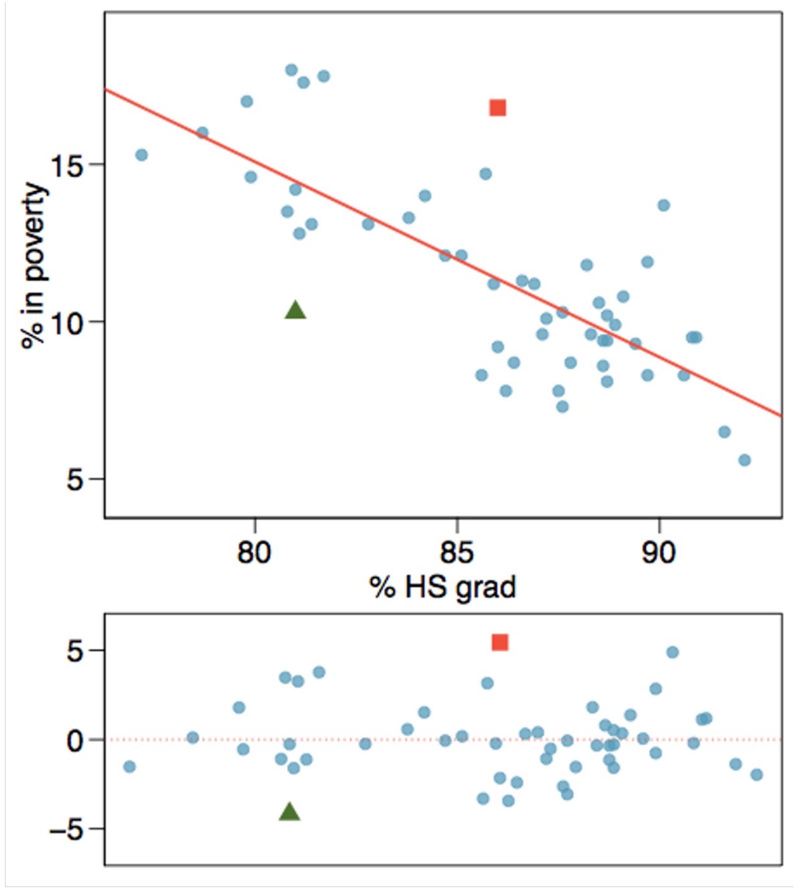
- The relationship between the explanatory and the mean of the response variable is linear
- There exist parameters β_0, β_1 such that $E(Y_i|x_i) = \beta_0 + \beta_1 x_i$

Conditions: (1) Linearity

- The relationship between the explanatory and the mean of the response variable is linear
- There exist parameters β_0, β_1 such that $E(Y_i|x_i) = \beta_0, +\beta_1 x_i$
- Check using a scatterplot of the data, or a *residuals plot*.



Anatomy of a residuals plot



▲ *RI:*

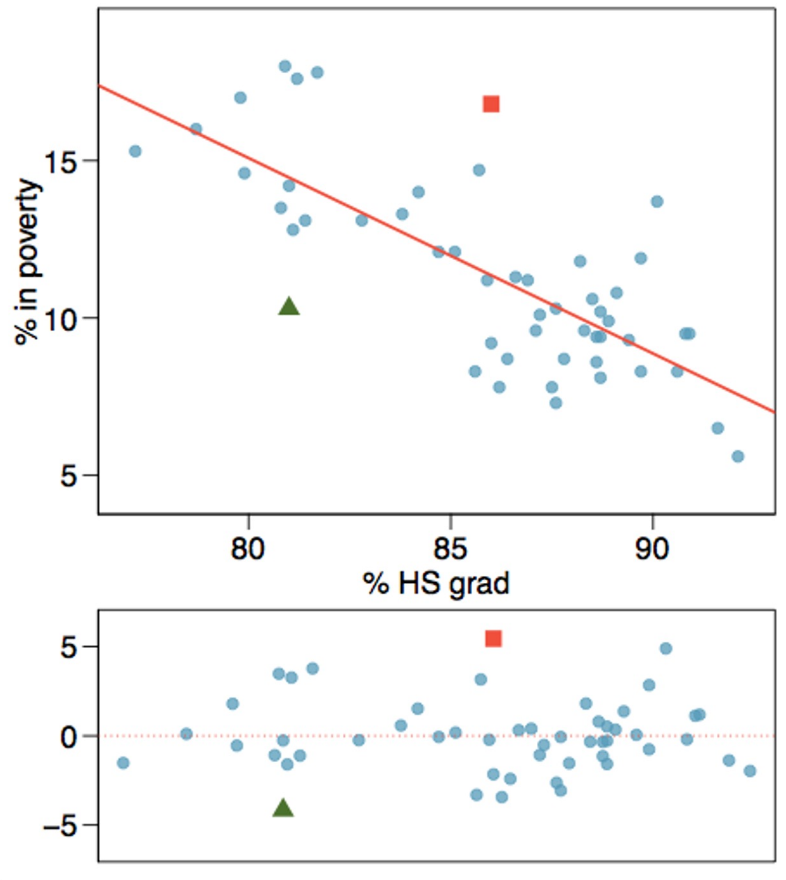
$$\% \; HS \; grad = 81 \qquad \% \; in \; poverty = 10.3$$

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

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

$$= 10.3 - 14.46 = -4.16$$

Anatomy of a residuals plot



▲ *RI:*

$$\% \text{ HS grad} = 81 \qquad \% \text{ in poverty} = 10.3$$

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

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

$$= 10.3 - 14.46 = -4.16$$

■ *DC:*

$$\% \text{ HS grad} = 86 \qquad \% \text{ in poverty} = 16.8$$

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

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

$$= 16.8 - 11.36 = 5.44$$

Conditions: (2) Normality

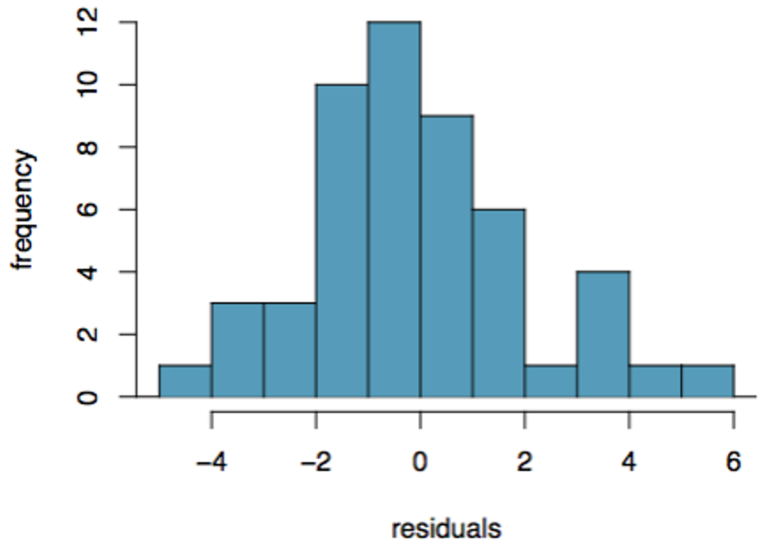
- For any fixed value of X, Y is normally distributed.

Conditions: (2) Normality

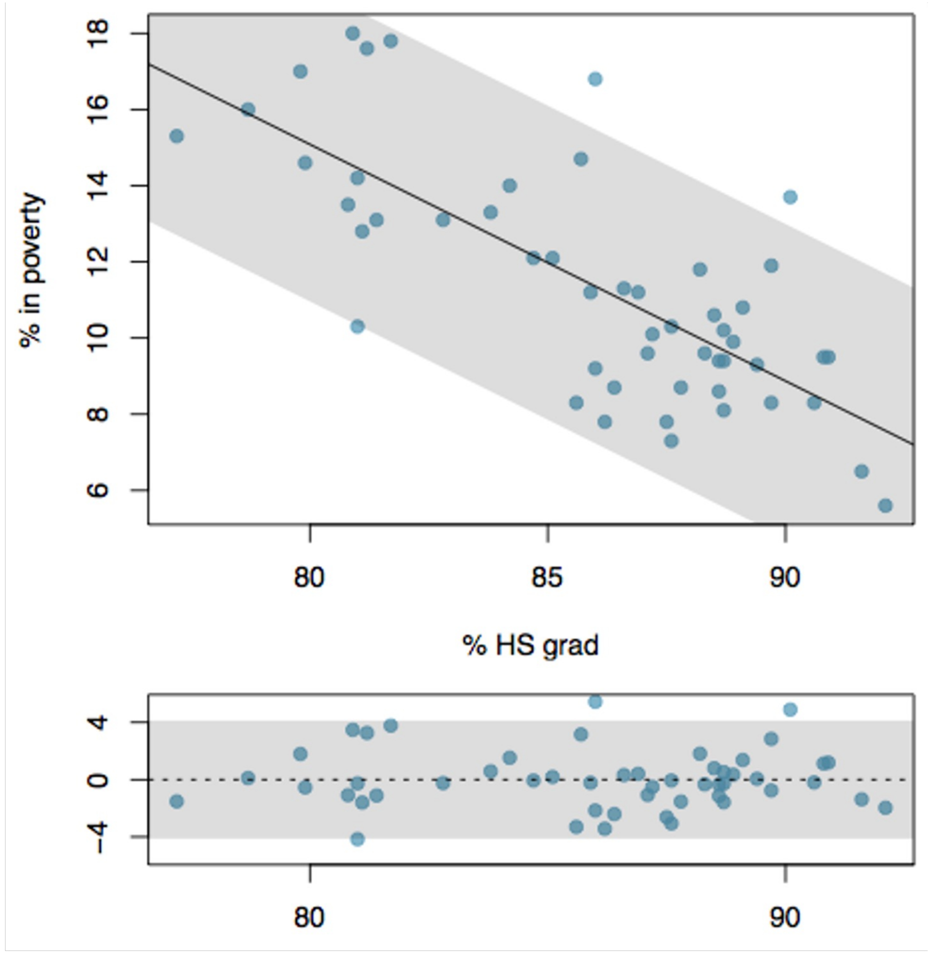
- For any fixed value of X, Y is normally distributed.
- The residuals should be nearly normal.
- This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.

Conditions: (2) Normality

- For any fixed value of X, Y is normally distributed.
- 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.

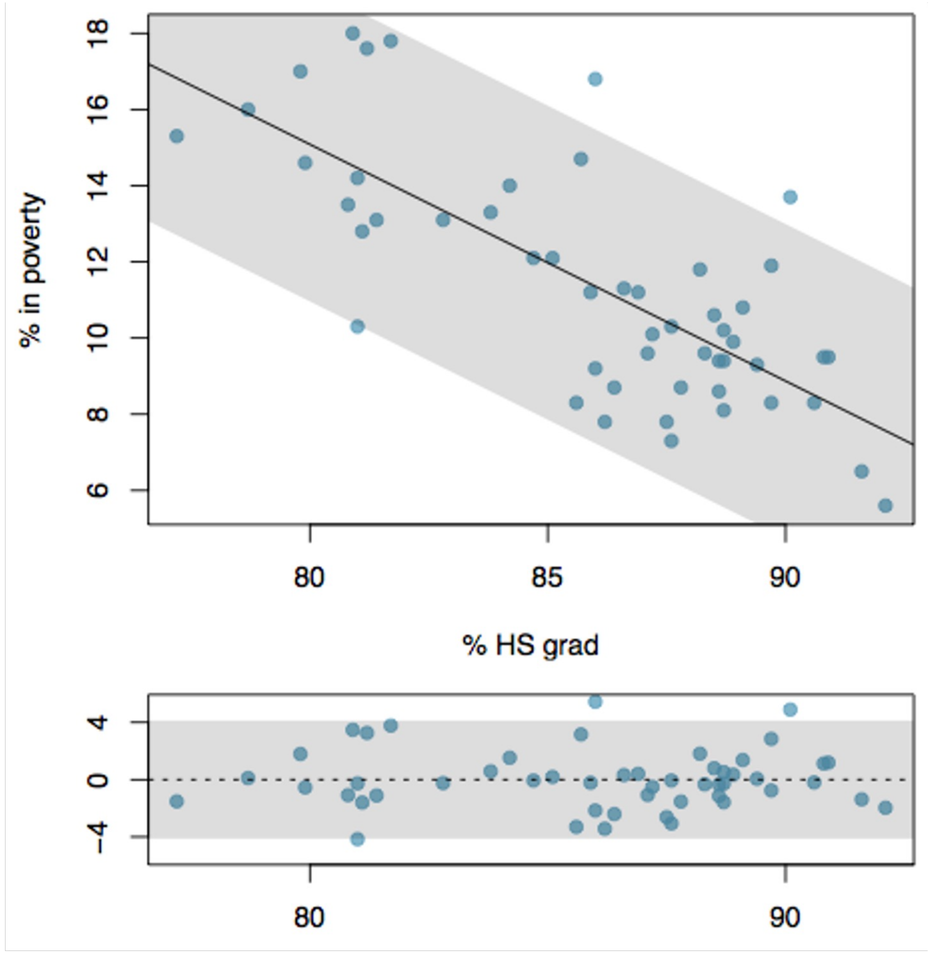


Conditions: (3) Constant variance



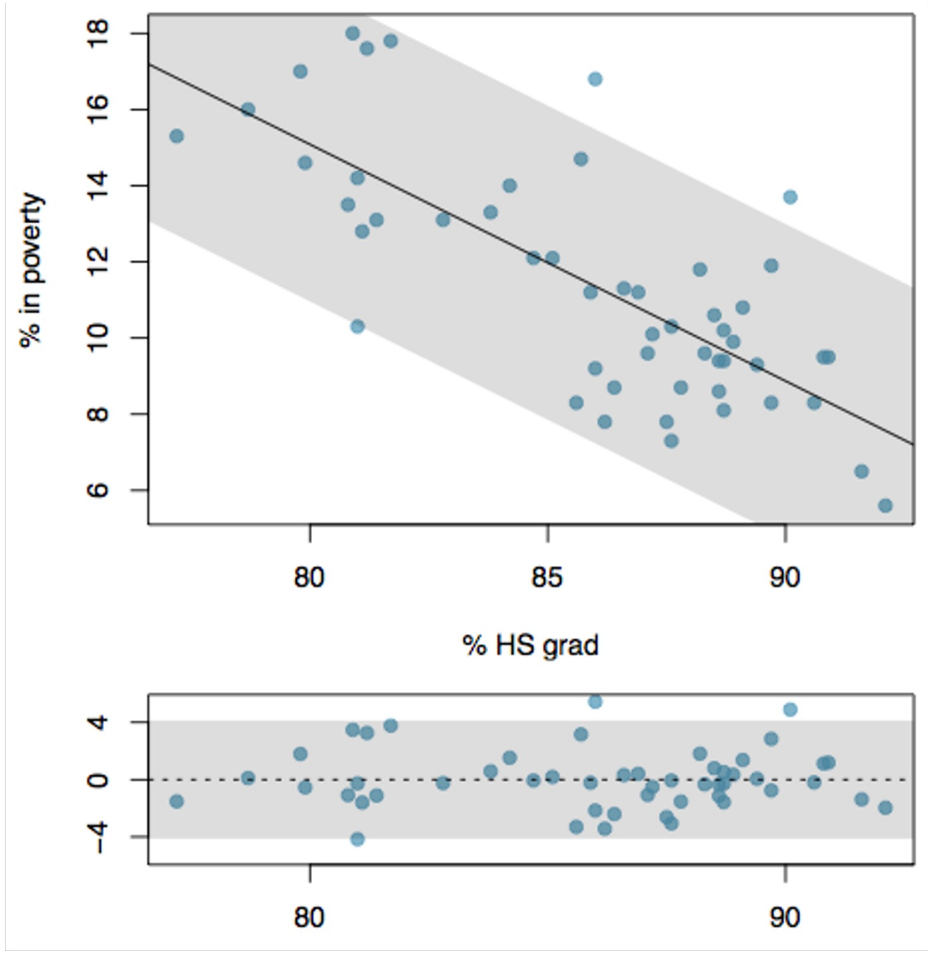
- The variance of points around the least squares line should be roughly constant.

Conditions: (3) Constant variance



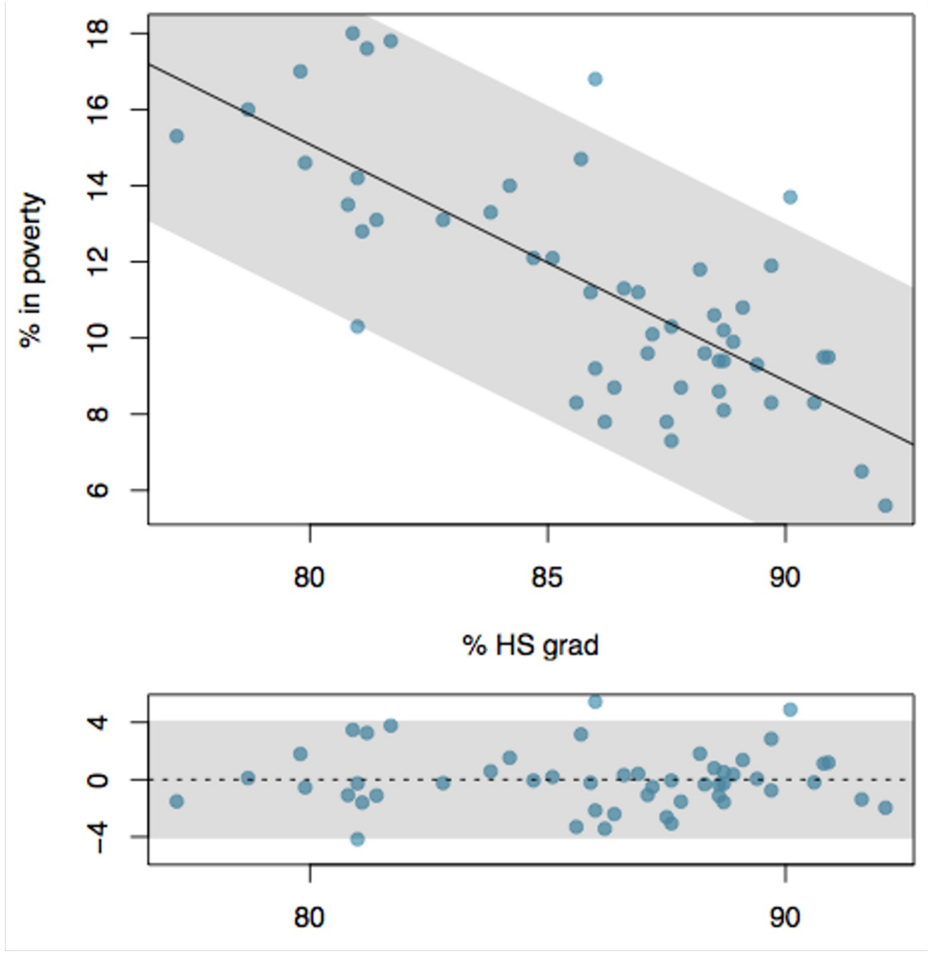
- The variance of points around the least squares line should be roughly constant.
- This implies that the variance of residuals around the 0 line should be roughly constant as well.

Conditions: (3) Constant variance



- The variance of points around the least squares line should be roughly constant.
- This implies that the variance of residuals around the 0 line should be roughly constant as well.
- Also called *homoscedasticity*.

Conditions: (3) Constant variance

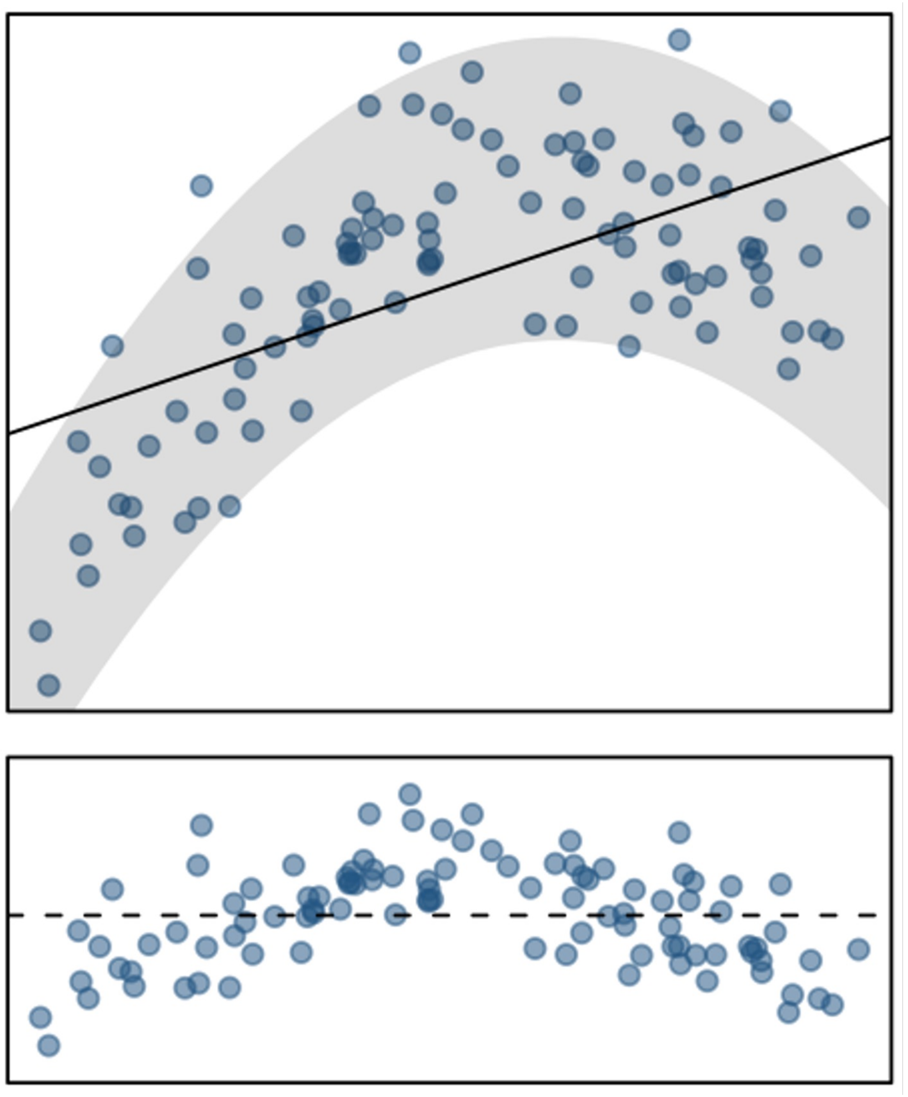


- The variance of points around the least squares line should be roughly constant.
- This implies that the variance of residuals around the 0 line should be roughly constant as well.
- Also called *homoscedasticity*.
- Check using a histogram or normal probability plot of residuals.

Checking conditions

What condition is this linear model obviously violating?

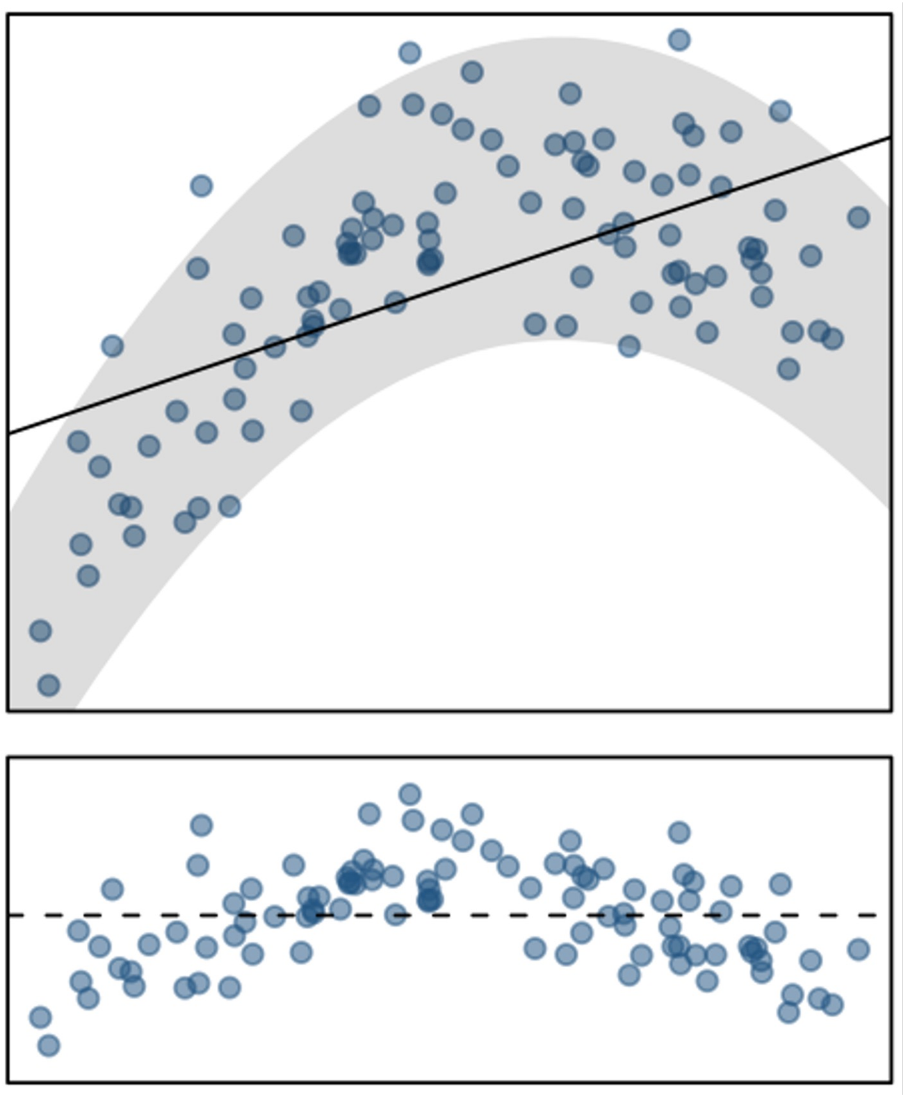
(a) Constant variance
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers



Checking conditions

What condition is this linear model obviously violating?

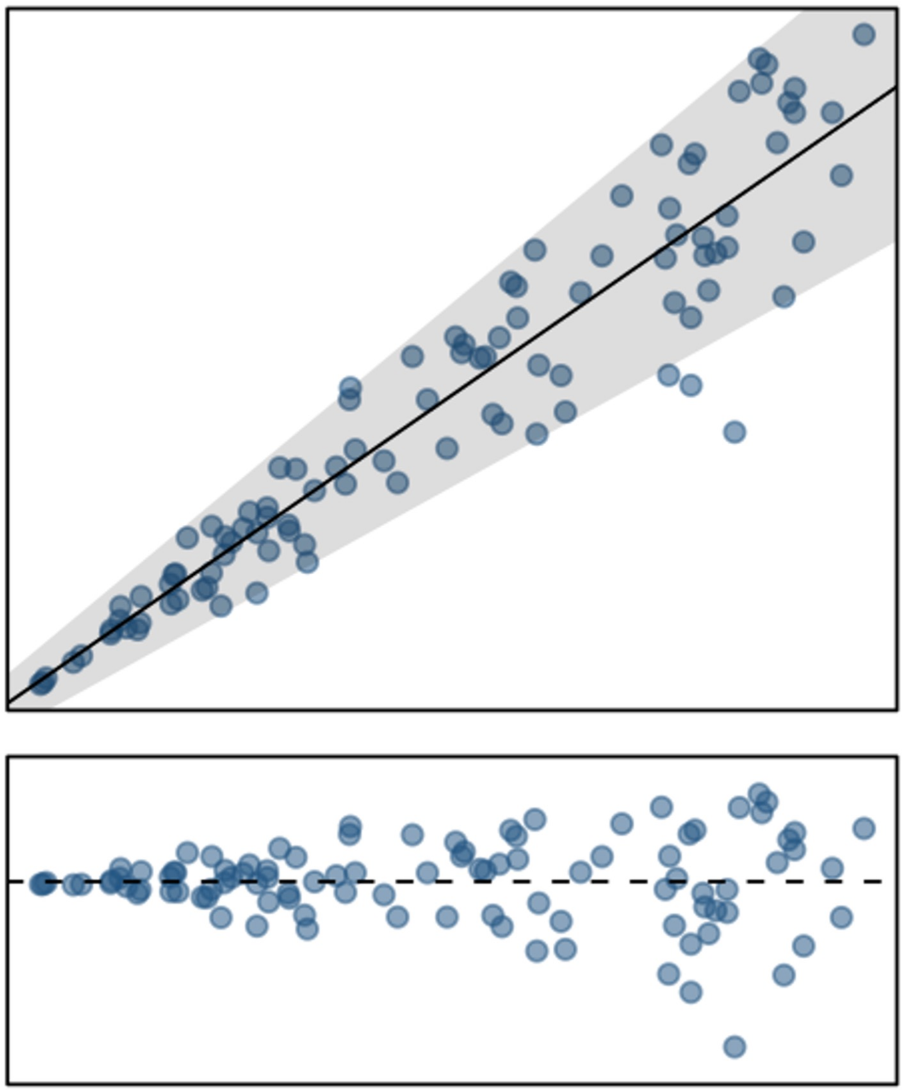
(a) Constant variance
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers



Checking conditions

What condition is this linear model obviously violating?

(a) Constant variance
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers



Checking conditions

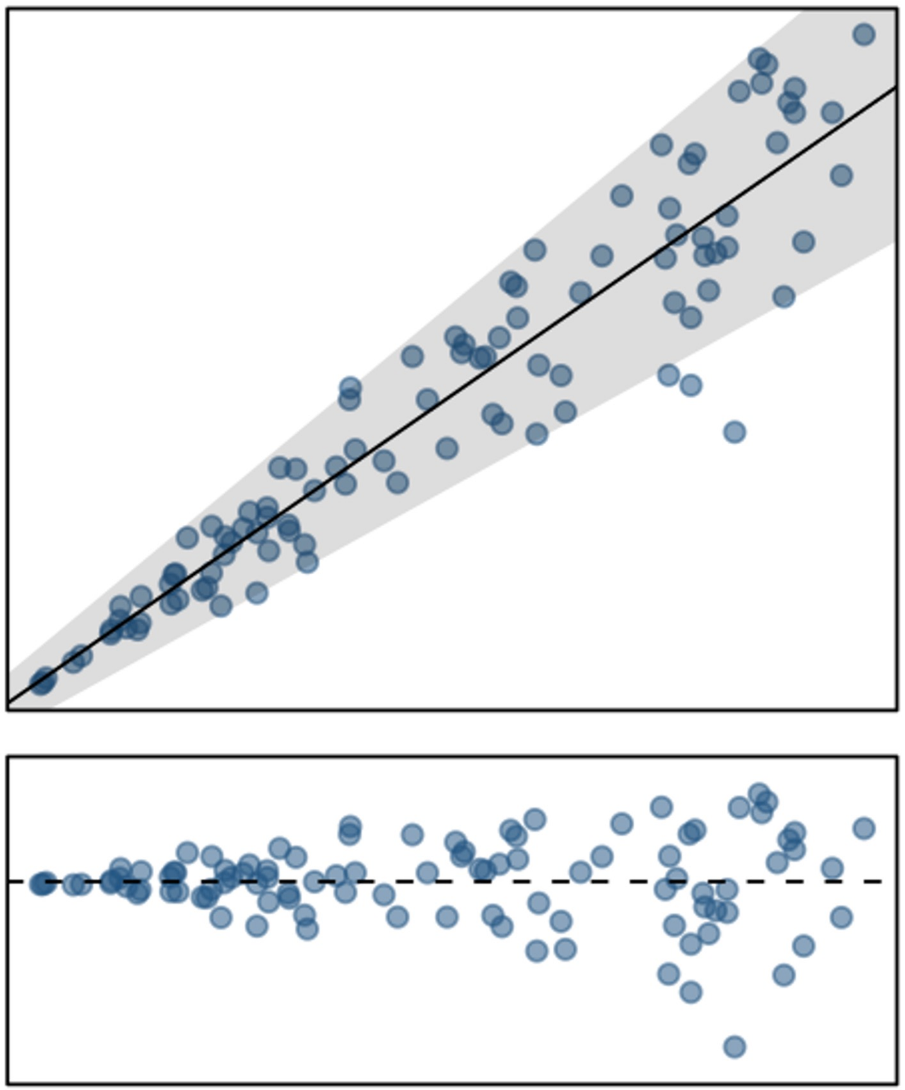
What condition is this linear model obviously violating?

(a) Constant variance

(b) Linear relationship

(c) Normal residuals

(d) No extreme outliers



Why do we need these assumptions

- These assumptions allow us to specify the conditional joint distribution of Y given the values x_i of X and the parameters $\beta_0, \beta_1, \sigma^2$

$$f(y|x, \beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp[-\frac{1}{2\sigma^2} \sum_i (y_i - \beta_0 - \beta_1 x_i)^2]$$

MLE estimates for $\beta_0, \beta_1, \sigma^2$:

$$b_o, b_1, \frac{1}{n} \sum_i (y_i - b_0 - b_1 x_i)^2$$

Least squares estimates for $\boldsymbol{\beta_0, \beta_1}$

- $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$

Sample standard deviation of y

- $\widehat{\beta_1} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{S_y}{S_x} R$

Sample standard deviation of x

Correlation coefficient

Fitted linear model:

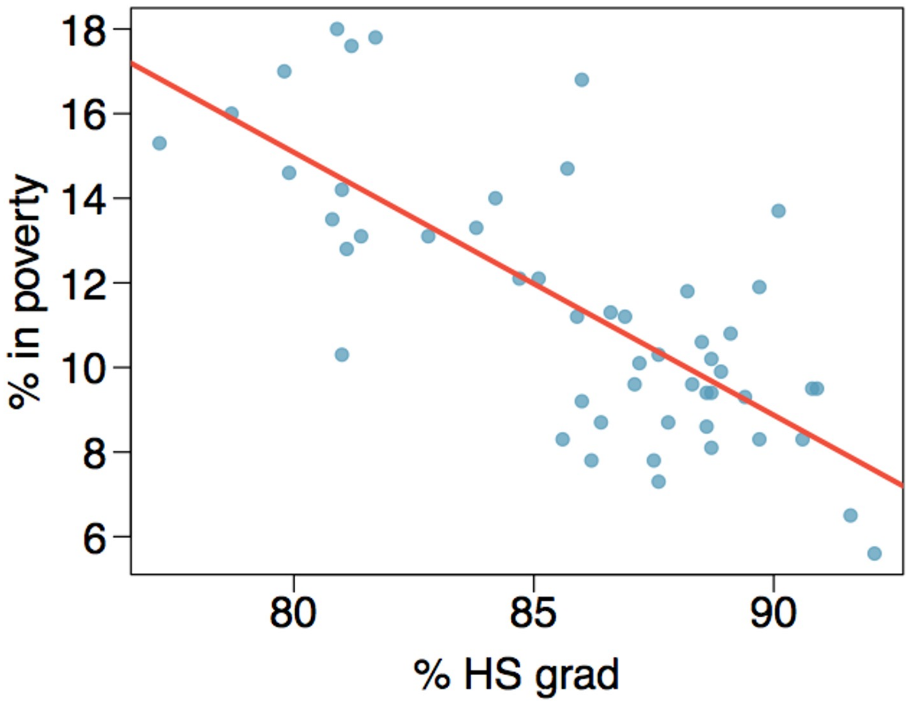
$$\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

Residuals:

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

Regression line

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



Coefficient of determination

R^2

- The strength of the fit of a linear model is commonly evaluated using R^2.

$$SS_{RES} = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$$

$$SS_{TOT} = \sum_i (y_i - \bar{y})^2$$

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}}$$

R^2

- The strength of the fit of a linear model is commonly evaluated using R^2.
- R^2 is calculated as the square of the correlation coefficient.
- It tells us what percent of variance in the response variable is explained by the model.
- The remainder of the variance is explained by variables not included in the model or by inherent randomness in the data.

Interpretation of R^2

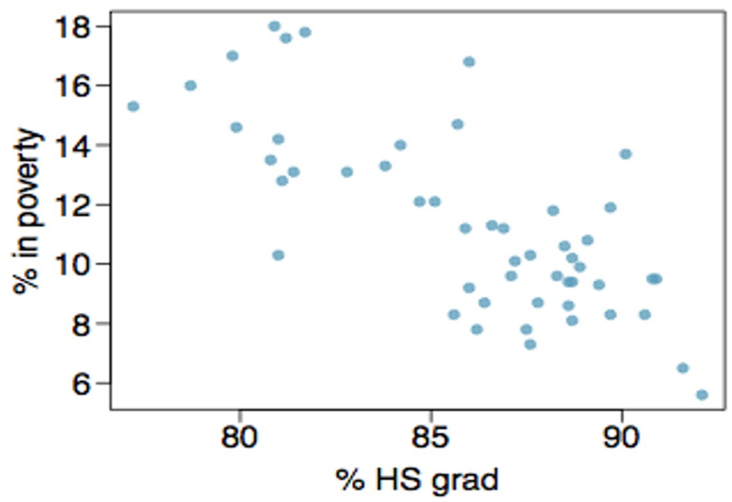
Which of the below is the correct interpretation of $R = -0.75$, $R^2 = 0.56$?

(a) 56% of the variance in the % of HG graduates among the 51 states is explained by the model.

(b) 56% of the variance in the % of residents living in poverty among the 51 states is explained by the model.

(c) 56% of the time % HS graduates predict % living in poverty correctly.

(d) 75% of the variance in the % of residents living in poverty among the 51 states is explained by the model.



Interpretation of R^2

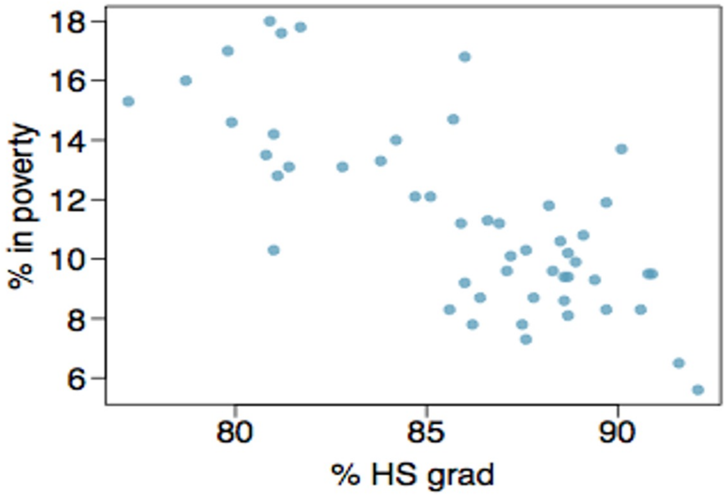
Which of the below is the correct interpretation of $R = -0.75$, $R^2 = 0.56$?

(a) 56% of the variance in the % of HG graduates among the 51 states is explained by the model.

(b) 56% of the variance in the % of residents living in poverty among the 51 states is explained by the model.

(c) 56% of the time % HS graduates predict % living in poverty correctly.

(d) 75% of the variance in the % of residents living in poverty among the 51 states is explained by the model.



Distribution of Estimators

Least squares/MLE estimates for $\boldsymbol{\beta_0, \beta_1}$

- $b_o = \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$

Sample standard deviation of y

- $b_1 = \widehat{\beta_1} = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x}) x_i} = \frac{S_y}{S_x} R$

Correlation coefficient

Sample standard deviation of x

Under the assumptions of linear regression, these are also the MLE estimates for β_0, β_1

MLE estimate for σ^2

- $\hat{\sigma}^2 = \frac{1}{n}\sum_i \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)^2$

Bias?

- $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$
- $\widehat{\beta_1} = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x}) x_i}$
- $\hat{\sigma}^2 = \frac{1}{n} \sum_i \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i \right)^2$

Bias?

- $\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$ No
- $\widehat{\beta_1} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$ No
- $\hat{\sigma}^2 = \frac{1}{n}\sum_i \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)^2$ Yes –divide by n-2 for unbiased estimator.

Distribution of the least squares estimators.

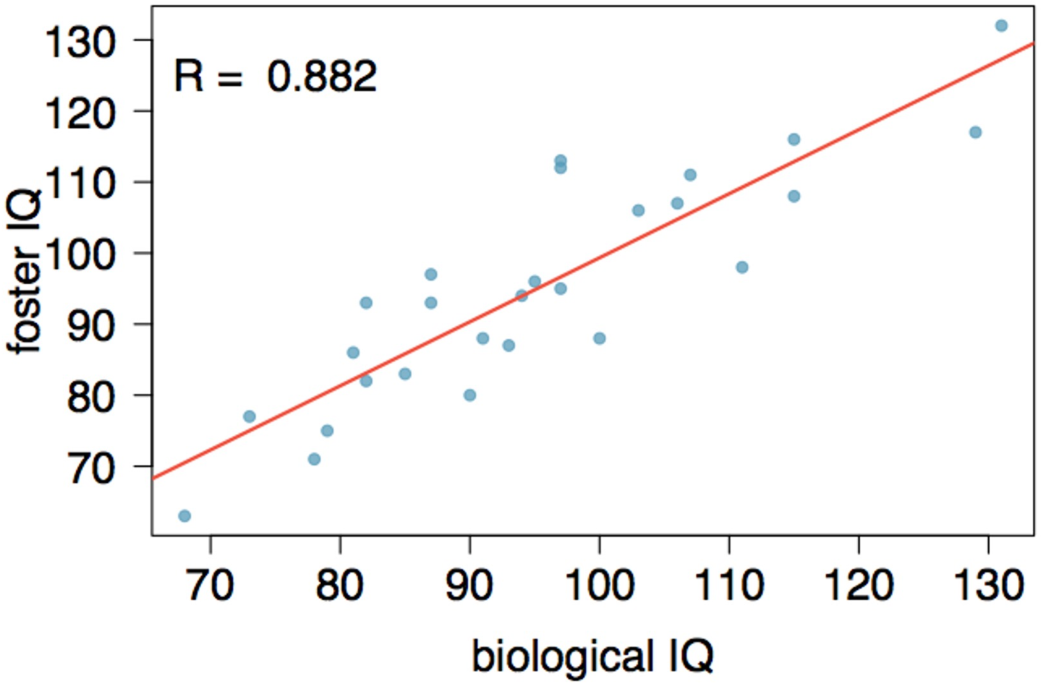
- $\widehat{\beta_0} \sim N\left(\beta_0, \sigma^2 \frac{\sum_i x_i^2}{n \sum_i (x_i - \bar{x})^2}\right)$
- $\widehat{\beta_1} \sim N(\beta_1, \frac{\sigma^2}{n \sum_i (x_i - \bar{x})^2})$
- $Cov(\widehat{\beta_0}, \widehat{\beta_1}) = \frac{\bar{x}\sigma^2}{\sum_i (x_i - \bar{x})^2}$

From CLT! Requires large samples

Inference for Linear Regression

Nature or 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 a random sample of 27 identical twins, one raised by foster parents, the other by the biological parents.



Nature or nurture?

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        9.20760    9.29990   0.990    0.332
bioIQ              0.90144    0.09633   9.358  1.2e-09

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
```

Practice

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       9.20760    9.29990   0.990    0.332
bioIQ             0.90144    0.09633   9.358  1.2e-09

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
```

(a) Additional 10 points in the biological twin's IQ is associated with additional 9 points in the foster twin's IQ, on average.

(b) The linear model is $\widehat{fosterIQ} = 9.2 + 0.9 \times bioIQ$

(c) Foster twins with IQs higher than average IQs tend to have biological twins with higher than average IQs as well.

Nature or nurture?

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.20760    9.29990   0.990    0.332
bioIQ         0.90144    0.09633   9.358  1.2e-09

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
```

$\widehat{\beta_0} = 9.21$

$\widehat{\beta_1} = 0.90$

$$\widehat{SE}\left(\widehat{\beta_0}\right) = \hat{\sigma}\sqrt{\frac{\sum_i x_i^2}{\sum_i (x_i - \bar{x})^2}}$$

$$\widehat{SE}\left(\widehat{\beta_1}\right) = \frac{\hat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}}$$

$$Var\left(\widehat{\beta_0}\right) = \sigma^2 \frac{\sum_i x_i^2}{n\sum_i (x_i - \bar{x})^2}$$

$$Var\left(\widehat{\beta_1}\right) = \frac{\sigma^2}{n\sum_i (x_i - \bar{x})^2}$$

$$SE(X) = \sigma_x/\sqrt{n}$$

Testing

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?

Testing

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?

(a) H_0: $b_0 = 0$; H_A: $b_0 \neq 0$
(b) H_0: $\beta_0 = 0$; H_A: $\beta_0 \neq 0$
(c) H_0: $b_1 = 0$; H_A: $b_1 \neq 0$
(d) H_0: $\beta_1 = 0$; H_A: $\beta_1 \neq 0$

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?

(a) H_0: $b_0 = 0$; H_A: $b_0 \neq 0$

(b) H_0: $\beta_0 = 0$; H_A: $\beta_0 \neq 0$

(c) H_0: $b_1 = 0$; H_A: $b_1 \neq 0$

(d) 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

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.

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.

Remember: test statistic T = (point estimate - null value) / SE

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.

  Remember: test statistic T = (point estimate - null value) / SE

- Point estimate = b_1 is the observed slope.

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.

  Remember: test statistic T = (point estimate - null value) / SE

- Point estimate = b_1 is the observed slope.
- SE_{b1} is the standard error associated with the slope.

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.

  Remember: test statistic T = (point estimate - null value) / SE

- Point estimate = b_1 is the observed slope.
- SE_{b1} is the standard error associated with the slope.
- Degrees of freedom associated with the slope is $df = n - 2$, where n is the sample size.

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.

  Remember: test statistic T = (point estimate - null value) / SE

- Point estimate = b_1 is the observed slope.
- SE_{b1} is the standard error associated with the slope.
- Degrees of freedom associated with the slope is $df = n - 2$, where n is the sample size.

  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

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

$$T = \frac{0.9014 - 0}{0.0963} = 9.36$$

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

$$T = \frac{0.9014 - 0}{0.0963} = 9.36$$

$$df = 27 - 2 = 25$$

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

$$T = \frac{0.9014 - 0}{0.0963} = 9.36$$

$$df = 27 - 2 = 25$$

$$p - value = P(|T| > 9.36) < 0.01$$

Confidence interval for the slope

Remember that a confidence interval is calculated as *point estimate* ± $t_{df} * SE$ and the degrees of freedom associated with the slope in a simple linear regression is *n* - 2. Which of the below 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

(a) 9.2076 ± 1.65 x 9.2999
(b) 0.9014 ± 2.06 x 0.0963
(c) 0.9014 ± 1.96 x 0.0963
(d) 9.2076 ± 1.96 x 0.0963

m	$p = .55$	.60	.65	.70	.75	.80	.85	.90	.95	.975	.99	.995
21	.127	.257	.391	.532	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831
22	.127	.256	.390	.532	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819
23	.127	.256	.390	.532	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807
24	.127	.256	.390	.531	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797
25	.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787
26	.127	.256	.390	.531	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779
27	.127	.256	.389	.531	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771
28	.127	.256	.389	.530	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763
29	.127	.256	.389	.530	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756
30	.127	.256	.389	.530	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750

Confidence interval for the slope

Remember that a confidence interval is calculated as *point estimate* $\pm\, t_{df} * SE$ and the degrees of freedom associated with the slope in a simple linear regression is *n* - 2. Which of the below 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

(a) 9.2076 ± 1.65 x 9.2999
(b) 0.9014 ± 2.06 x 0.0963
(c) 0.9014 ± 1.96 x 0.0963
(d) 9.2076 ± 1.96 x 0.0963

m	$p = .55$	.60	.65	.70	.75	.80	.85	.90	.95	.975	.99	.995
21	.127	.257	.391	.532	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831
22	.127	.256	.390	.532	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819
23	.127	.256	.390	.532	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807
24	.127	.256	.390	.531	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797
25	.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787
26	.127	.256	.390	.531	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779
27	.127	.256	.389	.531	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771
28	.127	.256	.389	.530	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763
29	.127	.256	.389	.530	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756
30	.127	.256	.389	.530	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750

n = 27 *df = 27 - 2 = 25*

For a γ – confidence interval we look at $(1+\gamma)/2$ T-quantile

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. Which of the below 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

(a) 9.2076 ± 1.65 x 9.2999

(b) 0.9014 ± 2.06 x 0.0963

(c) 0.9014 ± 1.96 x 0.0963

(d) 9.2076 ± 1.96 x 0.0963

$n = 27$ $df = 27 - 2 = 25$

95%: $t_{25}^* = 2.06$

0.9014 ± 2.06 x 0.0963

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. Which of the below 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

(a) 9.2076 ± 1.65 x 9.2999

(b) 0.9014 ± 2.06 x 0.0963

(c) 0.9014 ± 1.96 x 0.0963

(d) 9.2076 ± 1.96 x 0.0963

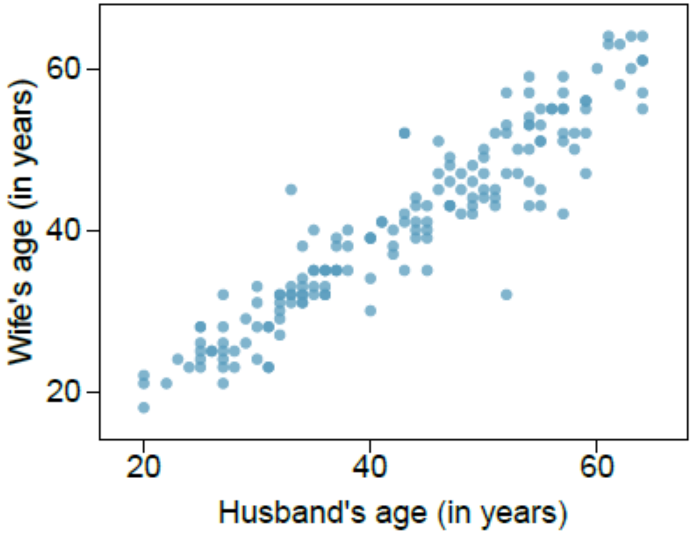
$n = 27$ $df = 27 - 2 = 25$

95%: $t_{25}^* = 2.06$

0.9014 ± 2.06 x 0.0963

(0.7, 1.1)

8.39 Husbands and wives, Part III. Exercise 8.33 presents a scatterplot displaying the relationship between husbands' and wives' ages in a random sample of 170 married couples in Britain, where both partners' ages are below 65 years. Given below is summary output of the least squares fit for predicting wife's age from husband's age.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.5740	1.1501	1.37	0.1730
age_husband	0.9112	0.0259	35.25	0.0000
				$df = 168$

(a) We might wonder, is the age difference between husbands and wives consistent across ages? If this were the case, then the slope parameter would be $\beta_1 = 1$. Use the information above to evaluate if there is strong evidence that the difference in husband and wife ages differs for different ages.

(b) Write the equation of the regression line for predicting wife's age from husband's age.

(c) Interpret the slope and intercept in context.

(d) Given that $R^2 = 0.88$, what is the correlation of ages in this data set?

(e) You meet a married man from Britain who is 55 years old. What would you predict his wife's age to be? How reliable is this prediction?

(f) You meet another married man from Britain who is 85 years old. Would it be wise to use the same linear model to predict his wife's age? Explain.

[20]Source: R Dataset, stat.ethz.ch/R-manual/R-patched/library/datasets/html/trees.html.