# Lecture 23: Residual Analysis and Type of Outliers

- Conditions for the least squares line
- Categorical explanatory variables
- Types of outliers in linear regression
- Log transformation

# Introduction

In the last lecture, we talked about **least square methods** and **simple linear regression**,

- The **response** variable is labeled $y$, the variable we are often interested in explaining or predicting. This is sometimes called the **dependent** variable (or **outcome**).
- The **explanatory** variable is labeled $x$, the variable we think/hope explains $y$. This is sometimes called the **independent** variable, or the **covariate**, or **predictor**.

The **regression model** assumes that the observed response is :

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

where the $\epsilon_i$'s are random error (due to genetics, measurement error, etc.). We assume that these errors are independent with mean 0 and unknown sd $\sigma$. We assume the $x_i$'s are measured without error.

– **Sample:** $(x_1, Y_1 = y_1), (x_2, Y_2 = y_2), ..., (x_n, Y_n = y_n)$

– **Least square estimators:**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

– Assumptions on the random errors $\epsilon_i$:

$$E[\epsilon_i] = 0 \quad V[\epsilon_i] = \sigma^2$$

– Implications for the response variable $Y_i$:

$$E[Y_i] = \beta_0 + \beta_1 x_i \quad V[Y_i] = \sigma^2$$

– The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are **unbiased**, that is

$$E[\hat{\beta}_0] = \beta_0, \quad E[\hat{\beta}_1] = \beta_1$$

– **Variances** of estimators

$$V[\hat{\beta}_0] = \sigma^2 \frac{\sum_{i=1}^n x_i^2}{nSS_{xx}}, \quad V[\hat{\beta}_1] = \frac{1}{SS_{xx}}\sigma^2$$

– In practice, the variance $\sigma^2$ of the random error is usually unknown. So it is necessary to estimate it. The unbiased estimator of $\sigma^2$ is

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\text{SSE}$$

# Normal Distributed Random Errors

In regression, we assume that the $y$ values are observed values of a collection of random variables. In this case, it turns out that assuming a normal distribution on $\epsilon$ makes the maximum likelihood estimates of $\beta_0$ and $\beta_1$ the same as what we had for the method of least squares.

In fact, we have that $y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma_\epsilon^2\right)$. By MLE, it turns out that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}; \qquad \hat{\beta}_1 = SS_{xy}/SS_{xx}$$

just like we had before using the method of least squares. By the way, how should we interpret $\beta_0$ and $\beta_1$? $\beta_0$ is usually the average of $y$ when $x$ is zero, thus it is only meaningful when $x$ can be zero (think about when $x$ represents the height of a person). For $\beta_1$, every unit increase in $x$ corresponds to an increase in $y$ by $\beta_1$.

The mathematical model for regression assumes that:

1. Each point $(x_i, y_i)$ in the scatterplot satisfies:

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

   where the $\epsilon_i$'s have a normal distribution with mean zero and (usually) unknown standard deviation.

2. The errors $\epsilon_i$ have nothing to do with one another; they are independent.
3. The $X_i$ values are measured without error. Thus all the error occurs in the $y$, and we do not need to minimize perpendicular distance to the line.

**Implications:** $Y_i, \hat{\beta}_0, \hat{\beta}_1$ are all normally distributed.

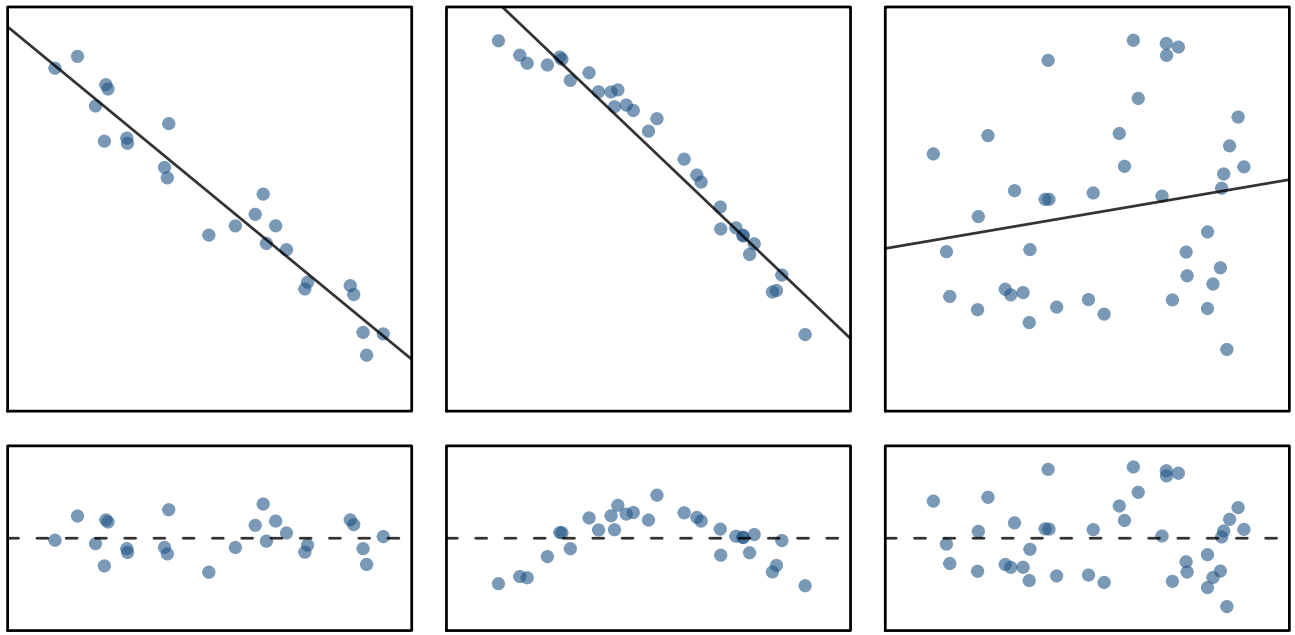
# Conditions for the least squares line

When fitting a least square line, we generally require

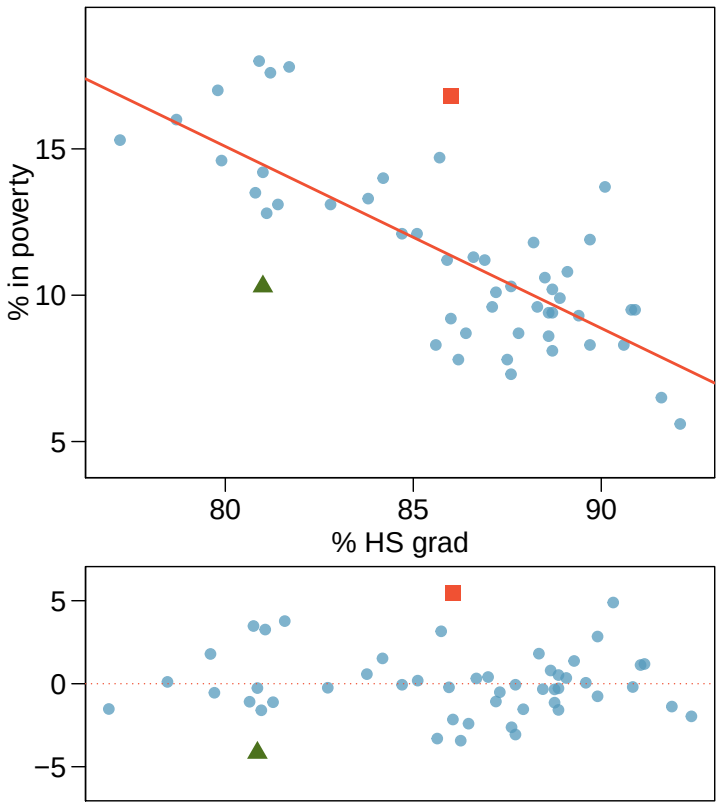
1. **Linearity:** The data should show a linear trend
2. **Nearly normal residuals:** the residuals must be nearly normally distributed
3. **Constant variability:** the variability of points around the least squares line remains roughly constant
4. **Independence observations:** depends on data collection method, often violated for time-series data. (We suspect that order of data collection may influence the outcome.)

# 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. If this topic is of interest, an Online Extra is available on openintro.org covering new techniques.
- Check using a scatterplot of the data, or a *residuals plot* where the residuals should be scattered around 0



# Anatomy of a residuals plot




▲ *RI:*

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

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

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

$$= 10.3 - 14.46 = -4.16$$

■ *DC:*

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

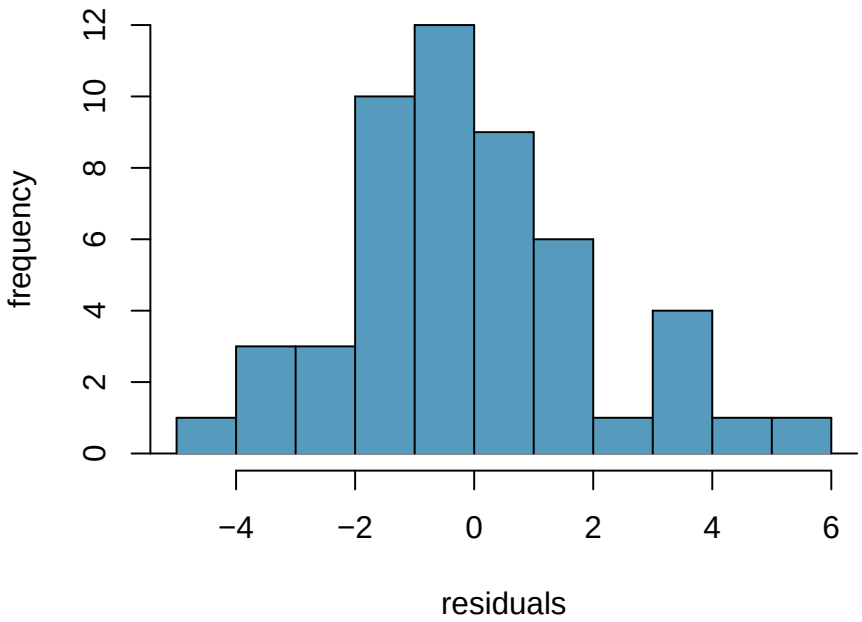
$$\widehat{\%\ in\ poverty} = 64.68 - 0.62 * 86 = 11.36$$

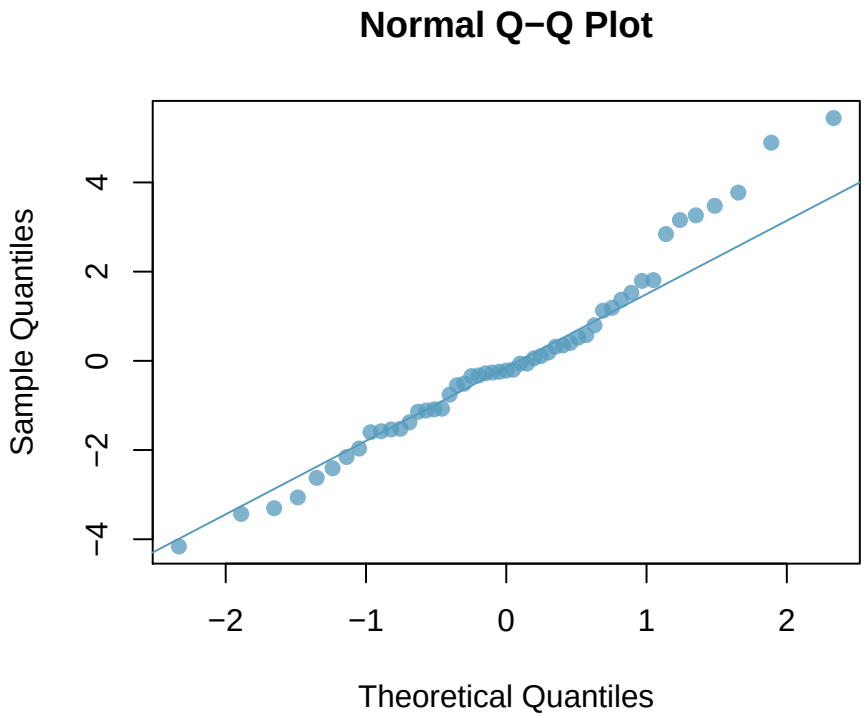
$$e = \%\ in\ poverty - \widehat{\%\ in\ poverty}$$

$$= 16.8 - 11.36 = 5.44$$

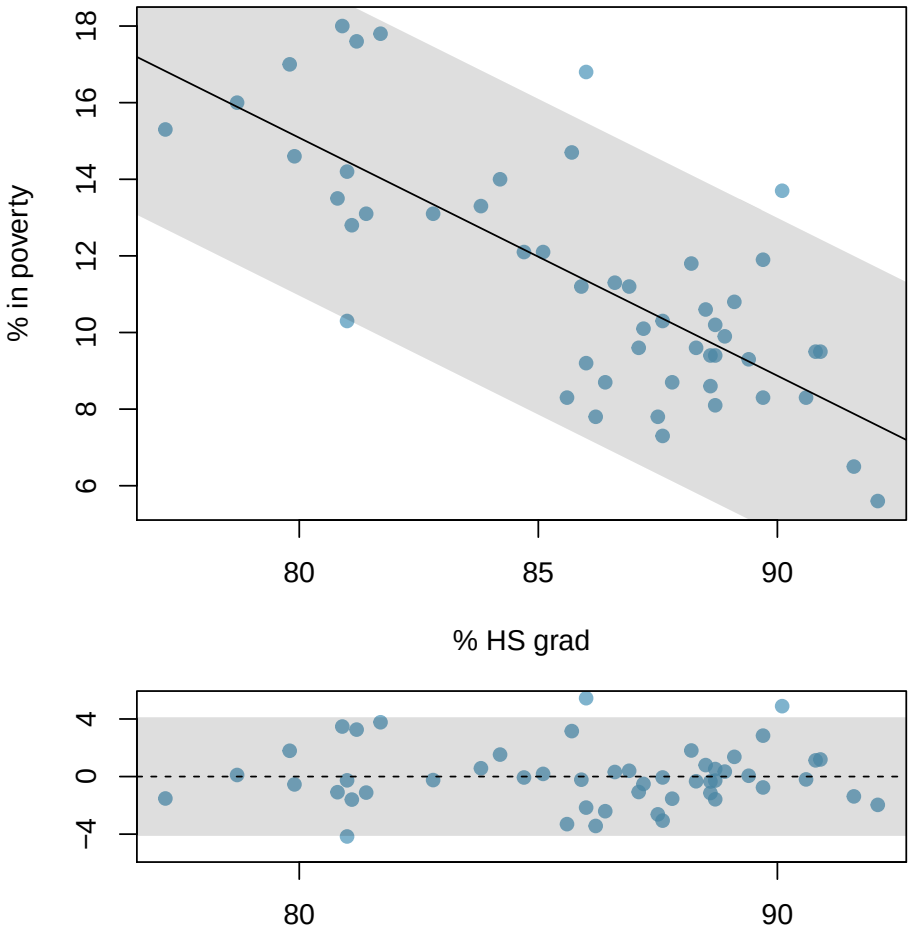
# Conditions: (2) Nearly normal residuals

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



# Conditions: (3) Constant variability

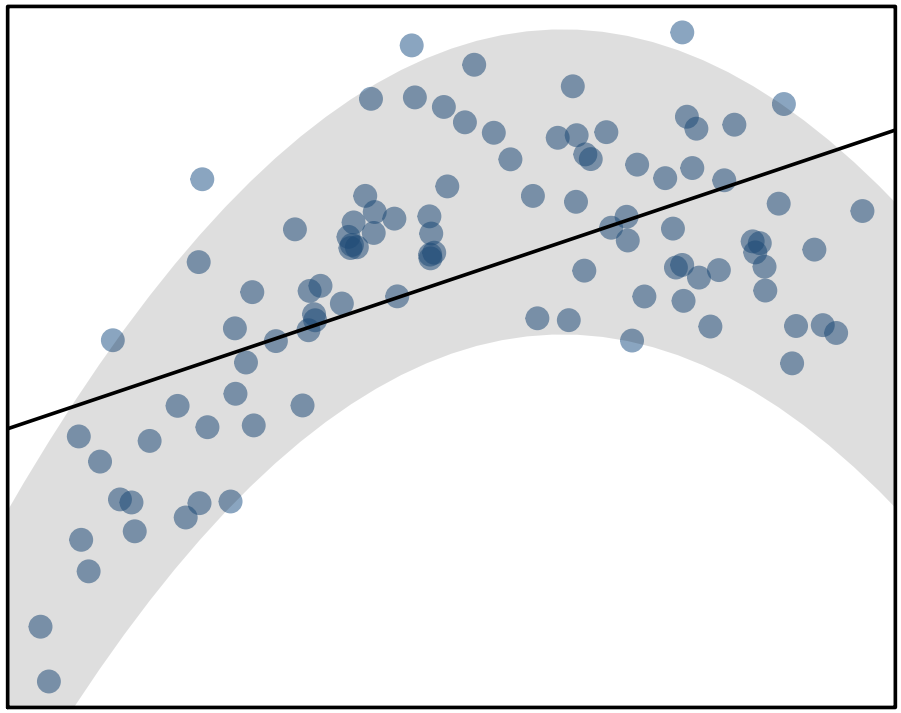


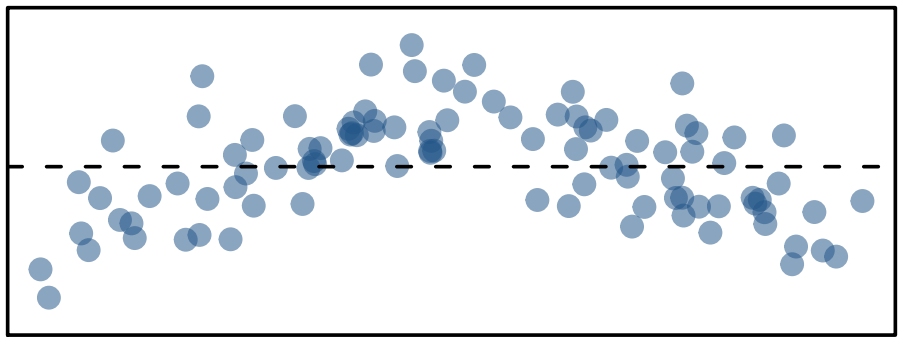

- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the 0 line should be roughly constant as well.
- Also called *homoscedasticity*.

# Checking conditions

What condition is this linear model obviously violating?

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





# Checking conditions

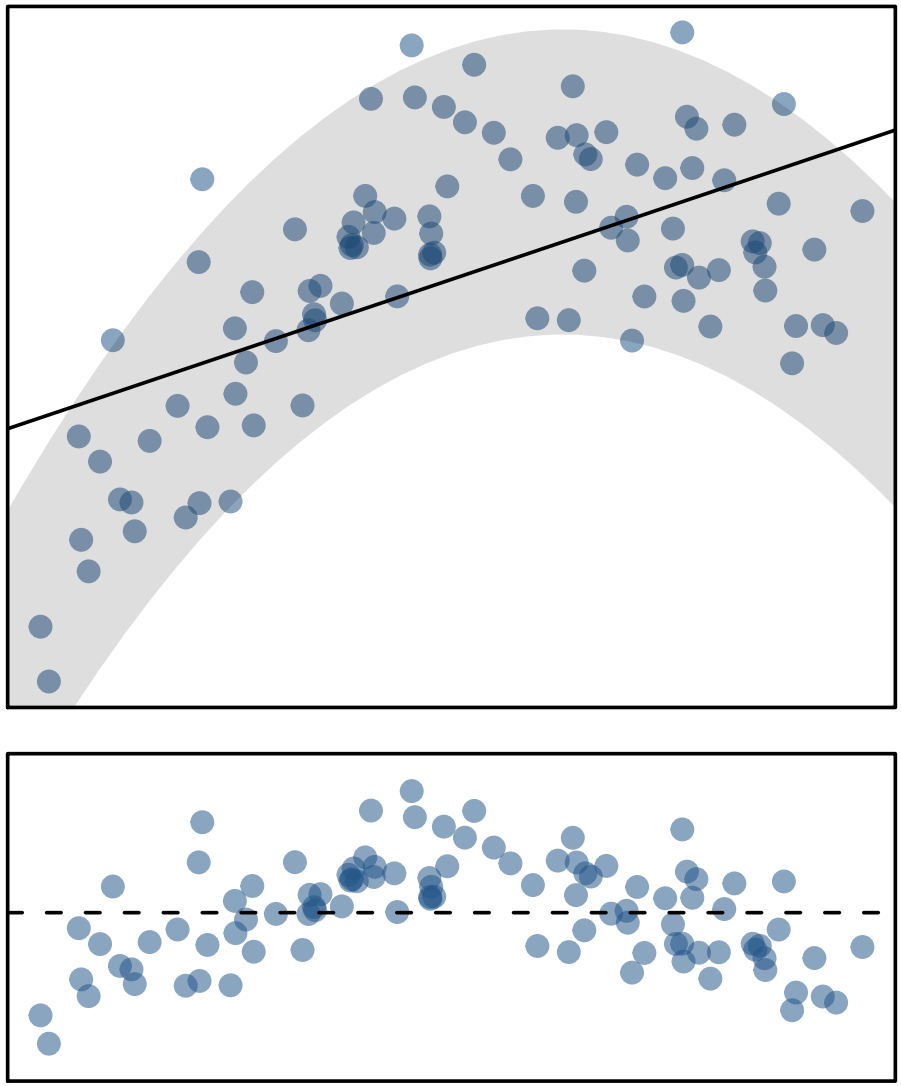
What condition is this linear model obviously violating?

(a) Constant variability

(b) *Linear relationship*

(c) Normal residuals

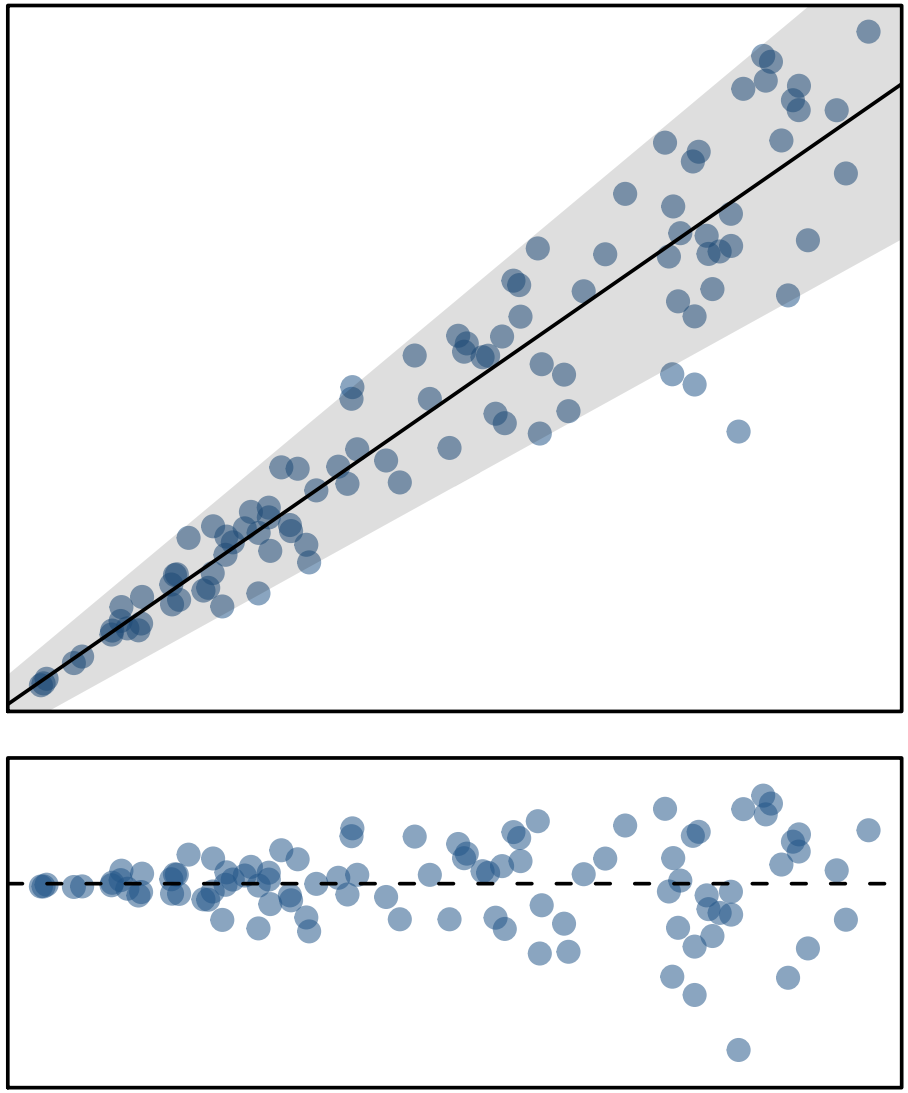
(d) No extreme outliers



# Checking conditions

What condition is this linear model obviously violating?

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



# Checking conditions

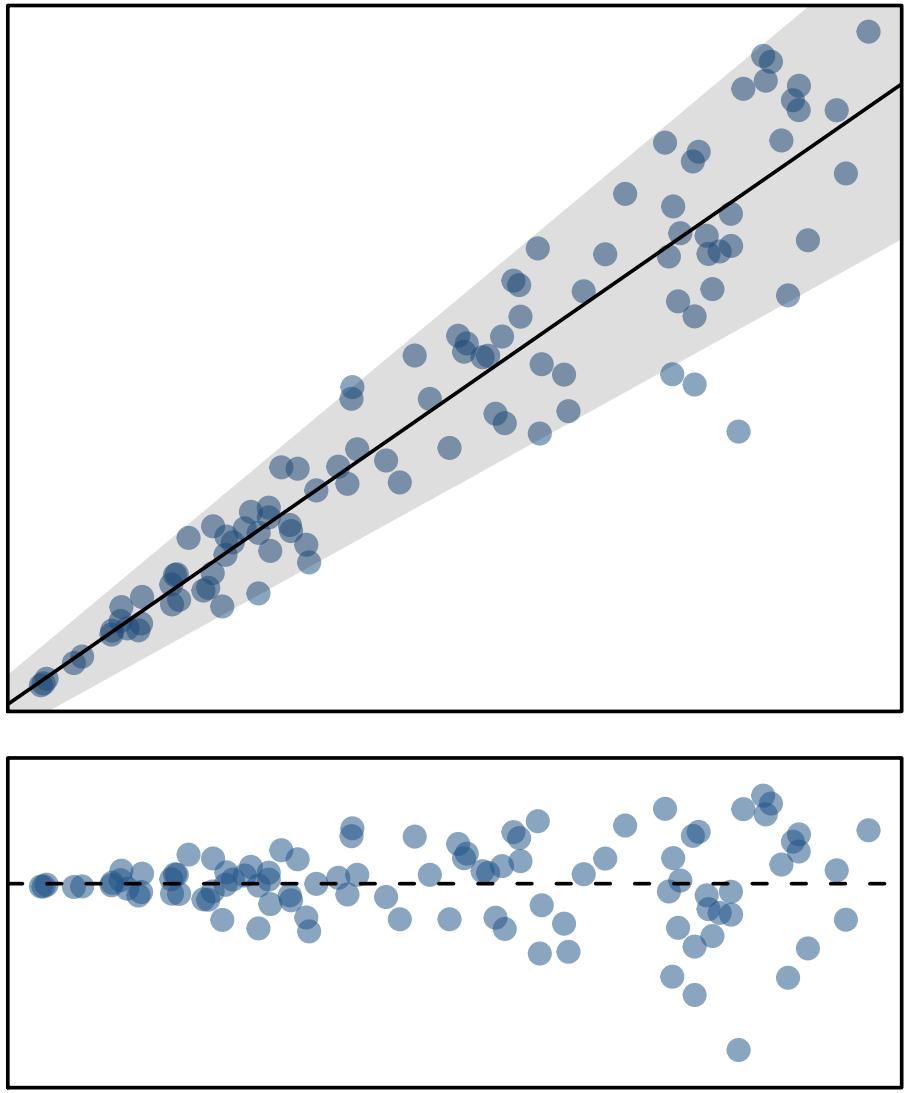
What condition is this linear model obviously violating?

(a) *Constant variability*

(b) Linear relationship

(c) Normal residuals

(d) No extreme outliers



# The variability in response

- Total sum of squares

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Regression sum of squares

$$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Residual sum of squares

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Coefficient of determination

$$r^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$$

# Relationship with correlation coefficient

- Simple linear regression

$$r^2 = \frac{\text{SS}_{\text{reg}}}{\text{SS}_{\text{tot}}} = \widehat{\text{corr}}(y, x)^2 = \frac{\text{SS}_{xy}^2}{\text{SS}_{xx}\text{SS}_{yy}}$$

- General linear regression

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in} y} = \frac{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}{\sum_{i=1}^{n}(y - \bar{y})^2} = \widehat{\text{corr}}(y, \hat{y})^2$$

# $R^2$

- 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.

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in} y} = \frac{\sum_{i=1}^{n}(\hat{y} - \bar{y})^2}{\sum_{i=1}^{n}(y - \bar{y})^2}$$

- 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 $R^2$

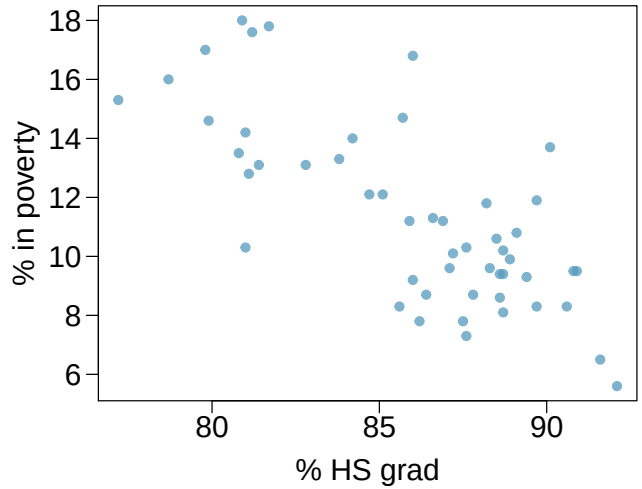
Which of the below is the correct interpretation of $R = -0.62$, $R^2 = 0.38$?

(a) 38% of the variability in the % of HG graduates among the 51 states is explained by the model.

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

(c) 38% of the time % HS graduates predict % living in poverty correctly.

(d) 62% of the variability in the % of residents living in poverty among the 51 states is explained by the model.



# Interpretation of $R^2$

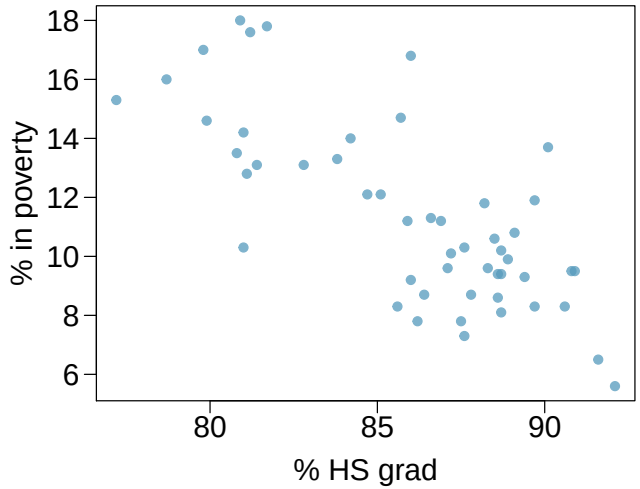
Which of the below is the correct interpretation of $R = -0.62$, $R^2 = 0.38$?

(a) 38% of the variability in the % of HG graduates among the 51 states is explained by the model.

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

(c) 38% of the time % HS graduates predict % living in poverty correctly.

(d) 62% of the variability in the % of residents living in poverty among the 51 states is explained by the model.



# Poverty vs. region (east, west)

$$\widehat{poverty} = 11.17 + 0.38 \times west$$

- Explanatory variable: region, *reference level:* east
- *Intercept:* The estimated average poverty percentage in eastern states is 11.17%
  - This is the value we get if we plug in *0* for the explanatory variable
- *Slope:* The estimated average poverty percentage in western states is 0.38% higher than eastern states.
  - Then, the estimated average poverty percentage in western states is 11.17 + 0.38 = 11.55%.
  - This is the value we get if we plug in *1* for the explanatory variable

# Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) is the reference level?

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 9.50 | 0.87 | 10.94 | 0.00 |
| region4midwest | 0.03 | 1.15 | 0.02 | 0.98 |
| region4west | 1.79 | 1.13 | 1.59 | 0.12 |
| region4south | 4.16 | 1.07 | 3.87 | 0.00 |

(a) northeast

(b) midwest

(c) west

(d) south

(e) cannot tell

# Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) is the reference level?

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 9.50 | 0.87 | 10.94 | 0.00 |
| region4midwest | 0.03 | 1.15 | 0.02 | 0.98 |
| region4west | 1.79 | 1.13 | 1.59 | 0.12 |
| region4south | 4.16 | 1.07 | 3.87 | 0.00 |

(a) *northeast*

(b) midwest

(c) west

(d) south

(e) cannot tell

# Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) has the lowest poverty percentage?

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 9.50 | 0.87 | 10.94 | 0.00 |
| region4midwest | 0.03 | 1.15 | 0.02 | 0.98 |
| region4west | 1.79 | 1.13 | 1.59 | 0.12 |
| region4south | 4.16 | 1.07 | 3.87 | 0.00 |

(a) northeast

(b) midwest

(c) west

(d) south

(e) cannot tell

# Poverty vs. region (northeast, midwest, west, south)

Which region (northeast, midwest, west, or south) has the lowest poverty percentage?

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 9.50 | 0.87 | 10.94 | 0.00 |
| region4midwest | 0.03 | 1.15 | 0.02 | 0.98 |
| region4west | 1.79 | 1.13 | 1.59 | 0.12 |
| region4south | 4.16 | 1.07 | 3.87 | 0.00 |

(a) *northeast*

(b) midwest

(c) west

(d) south

(e) cannot tell

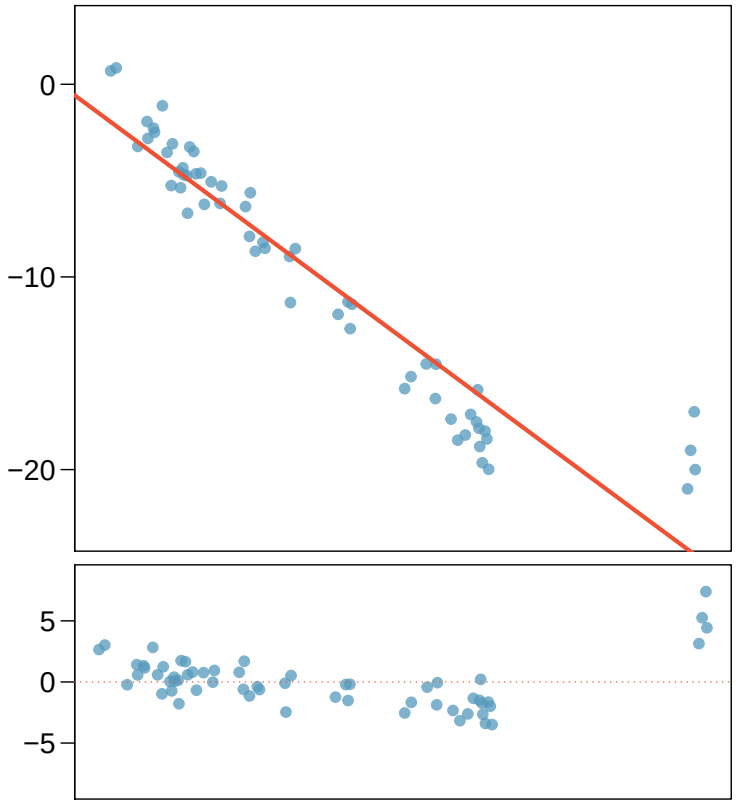
# Types of outliers

1. **Outliers** in regression are observations that fall far from the "cloud" of points
2. These points are especially important because they can have a strong influence on the least square line
3. We identify criteria for determining which outliers are important and influential
4. Type of outlier determines how it should be handled

# Types of outliers

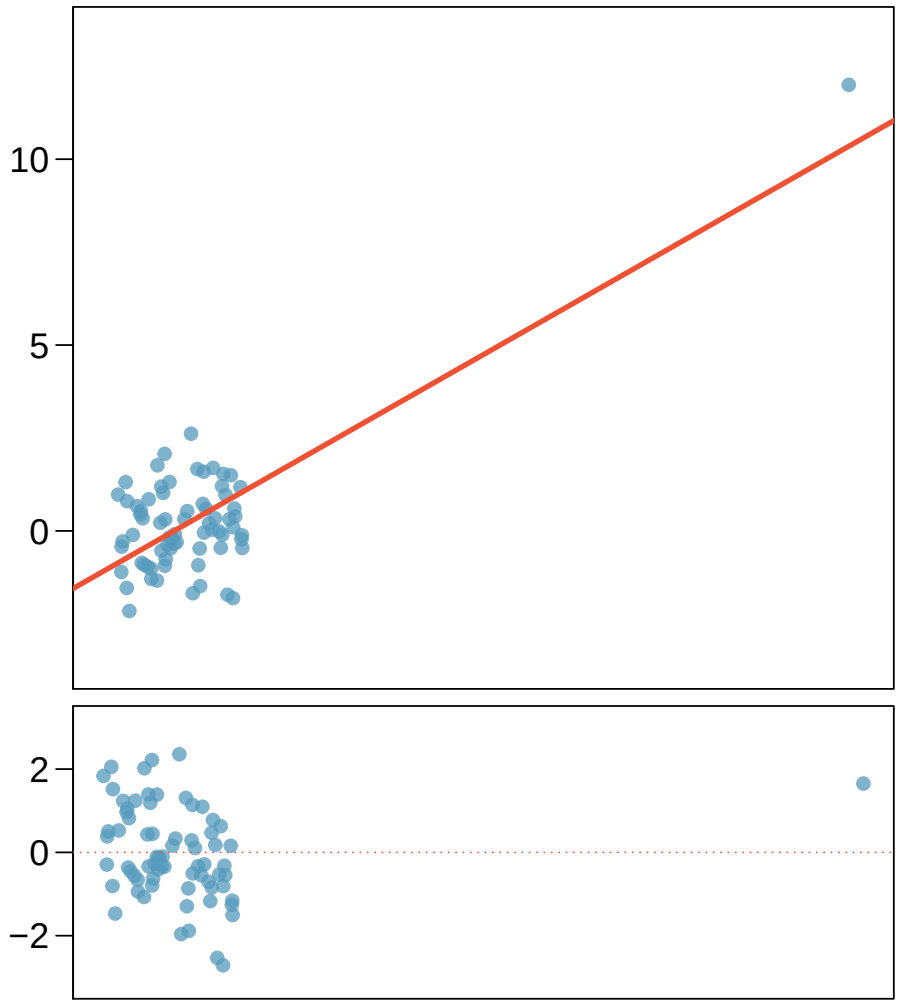
How do outliers influence the least squares line in this plot?

To answer this question think of where the regression line would be with and without the outlier(s). Without the outliers the regression line would be steeper, and lie closer to the larger group of observations. With the outliers the line is pulled up and away from some of the observations in the larger group.



# Types of outliers

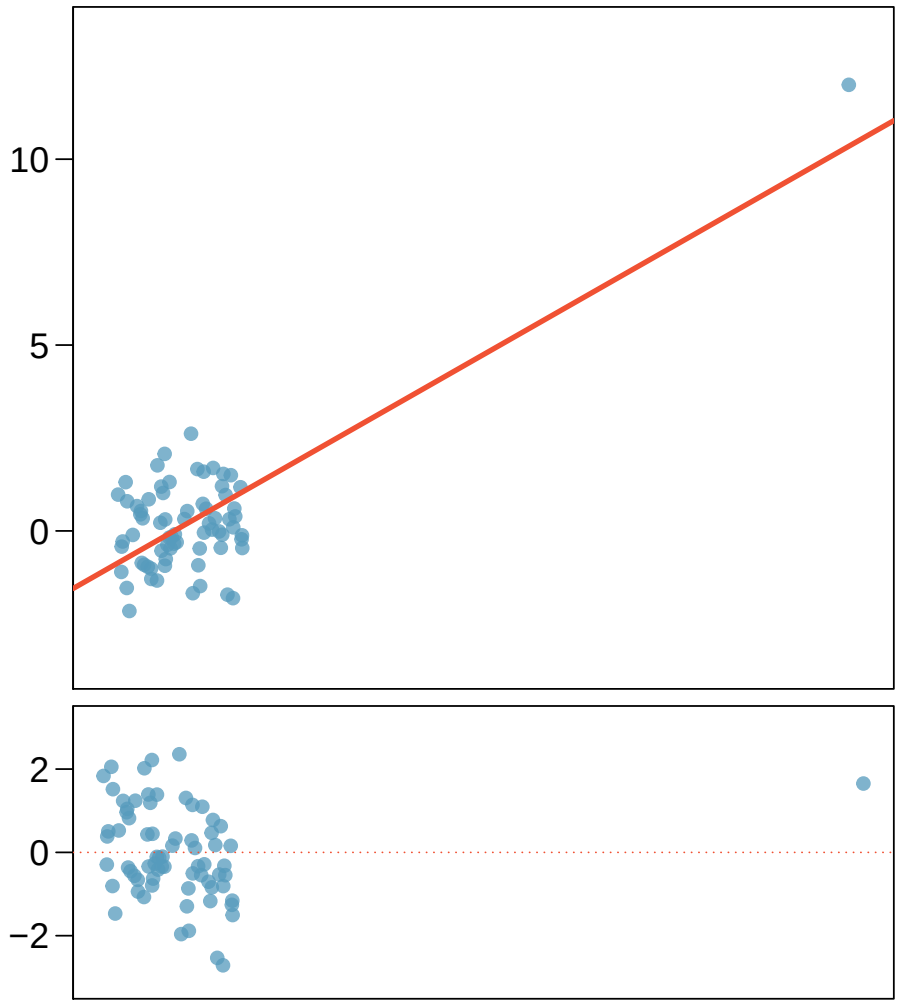
How do outliers influence the least squares line in this plot?



# Types of outliers

How do outliers influence the least squares line in this plot?

*Without the outlier there is no evident relationship between x and y.*

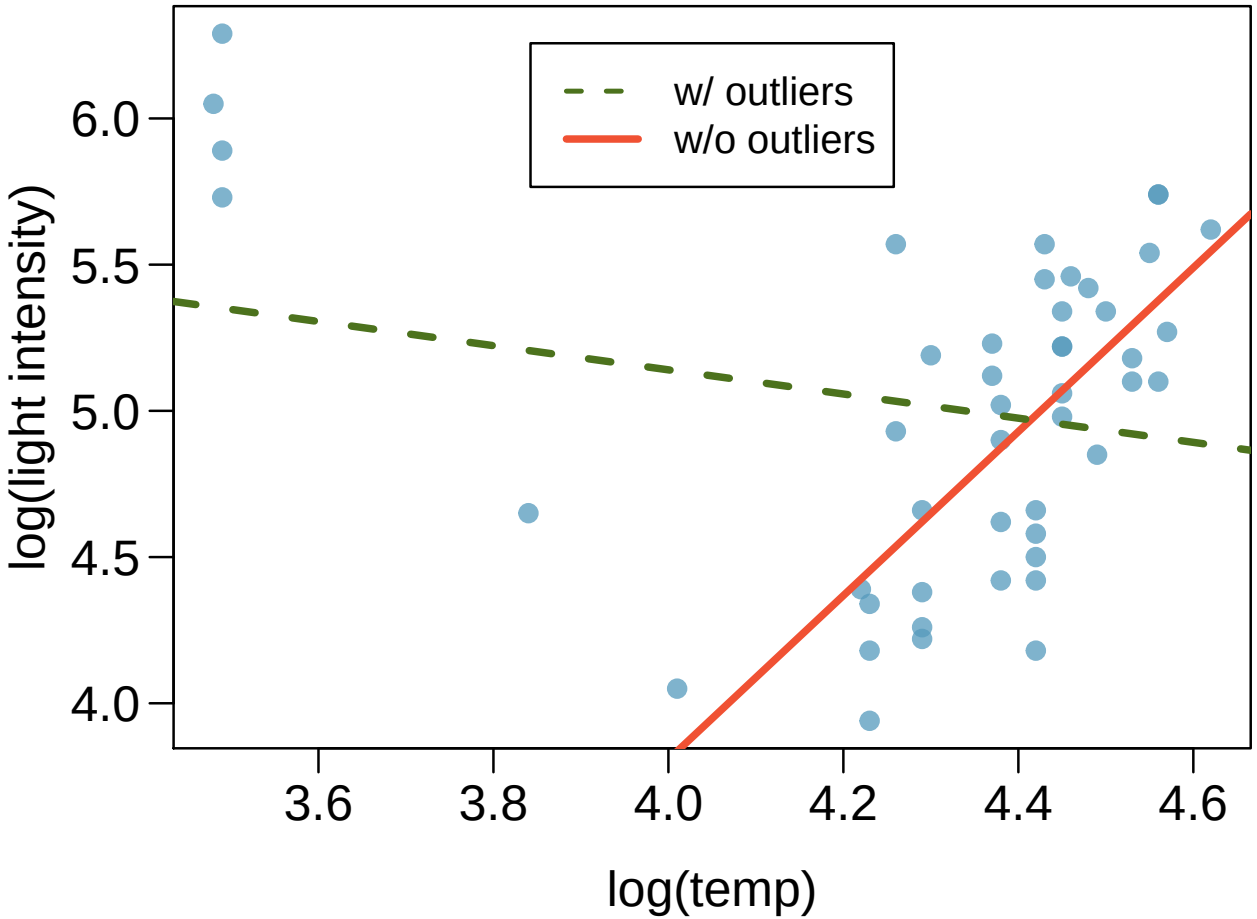


# Some terminology

- *Outliers* are points that fall away from the cloud of points.
- Outliers that lie horizontally away from the center of the cloud are called *high leverage* points (i.e. it has extreme predictor $x$ values).
- High leverage points that actually influence the slope of the regression line are called *influential* points.
- In order to determine if a point is influential, visualize the regression line with and without the point. Does the slope of the line change considerably? If so, then the point is influential. If not, then it's not an influential point.

# Influential points

Data are available on the log of the surface temperature and the log of the light intensity of 47 stars in the star cluster CYG OB1.




# Types of outliers

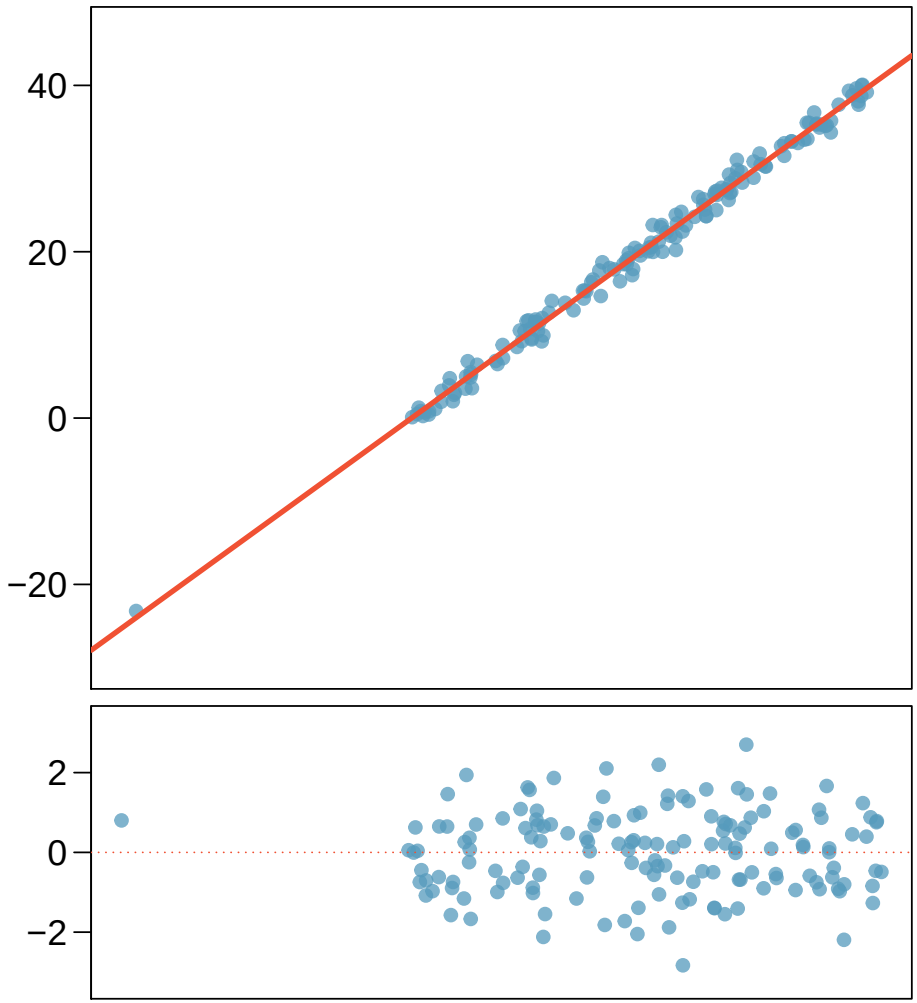
Which of the below best describes the outlier?

(a) influential

(b) high leverage

(c) none of the above

(d) there are no outliers



# Types of outliers

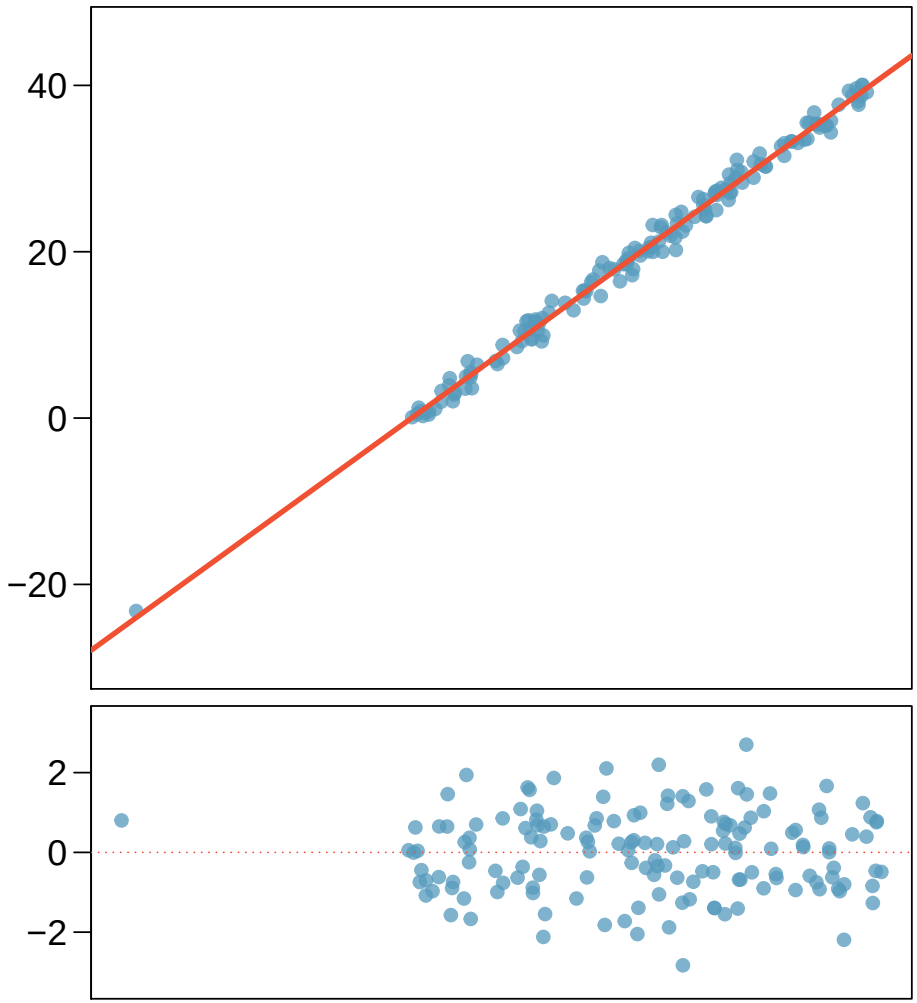
Which of the below best describes the outlier?

(a) influential

(b) *high leverage*

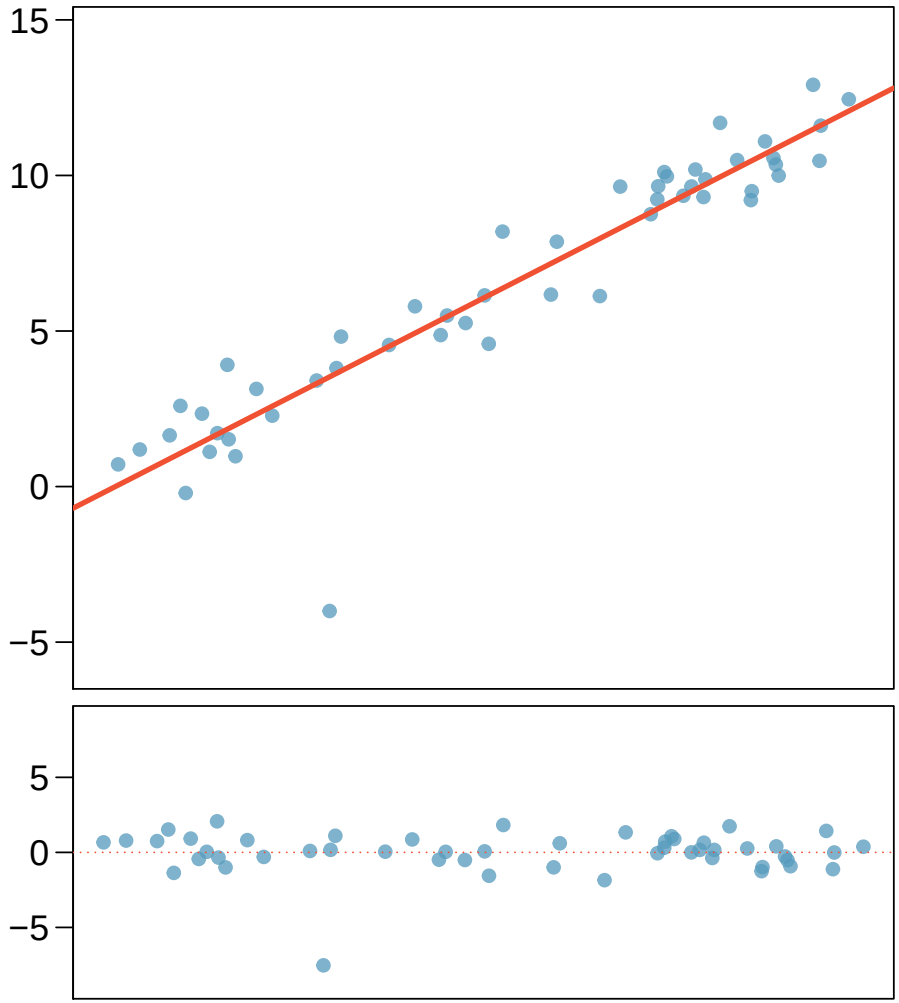
(c) none of the above

(d) there are no outliers



# Types of outliers

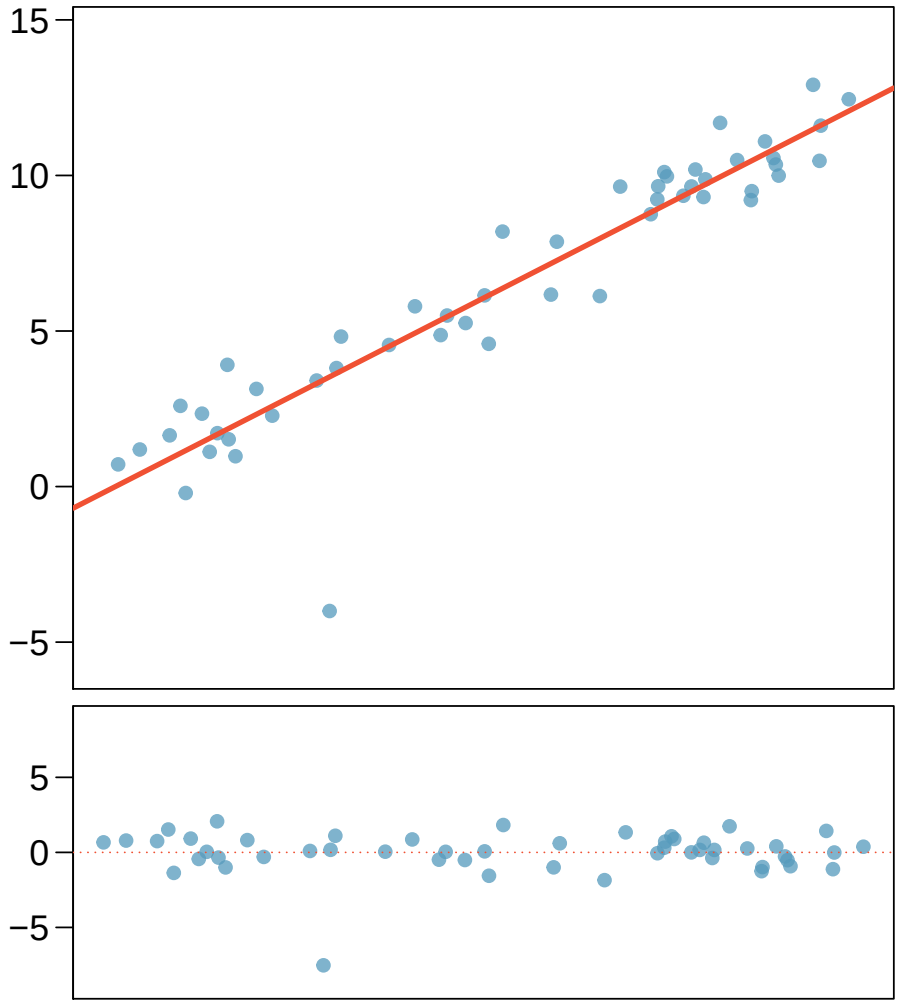
Does this outlier influence the slope of the regression line?



# Types of outliers

Does this outlier influence the slope of the regression line?

*Not much...*



# Recap

Which of following is true?

(a) Influential points always change the intercept of the regression line.

(b) Influential points always reduce $R^2$.

(c) It is much more likely for a low leverage point to be influential, than a high leverage point.

(d) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.

(e) None of the above.

# Recap

Which of following is true?

(a) Influential points always change the intercept of the regression line.

(b) Influential points always reduce $R^2$.

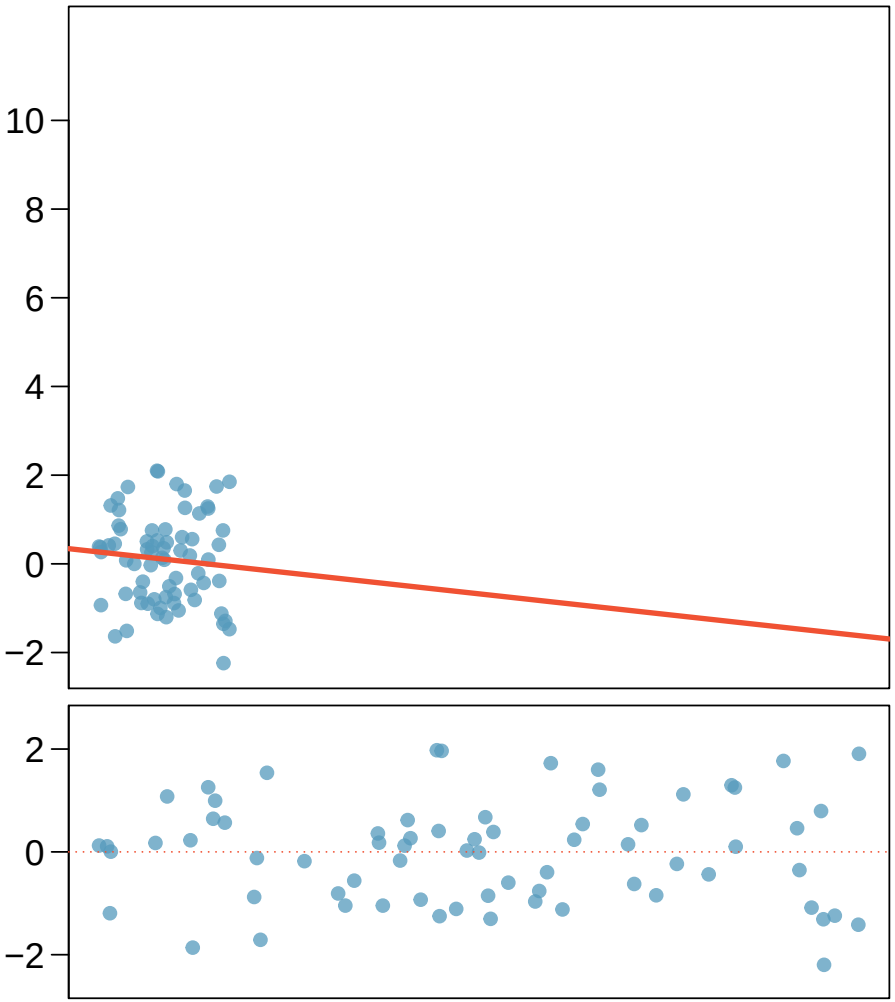
(c) It is much more likely for a low leverage point to be influential, than a high leverage point.

(d) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.

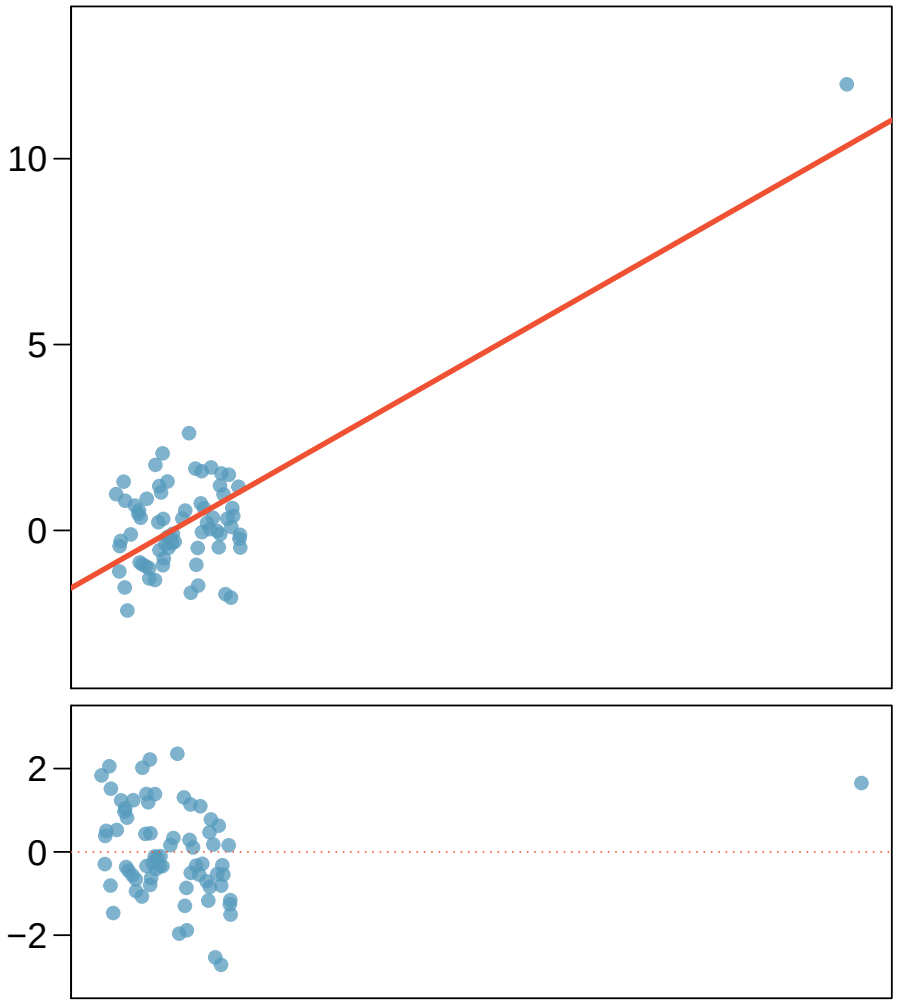
(e) *None of the above.*

# Recap (cont.)

$R = -0.08, R^2 = 0.0064$




$R = 0.79, R^2 = 0.6241$

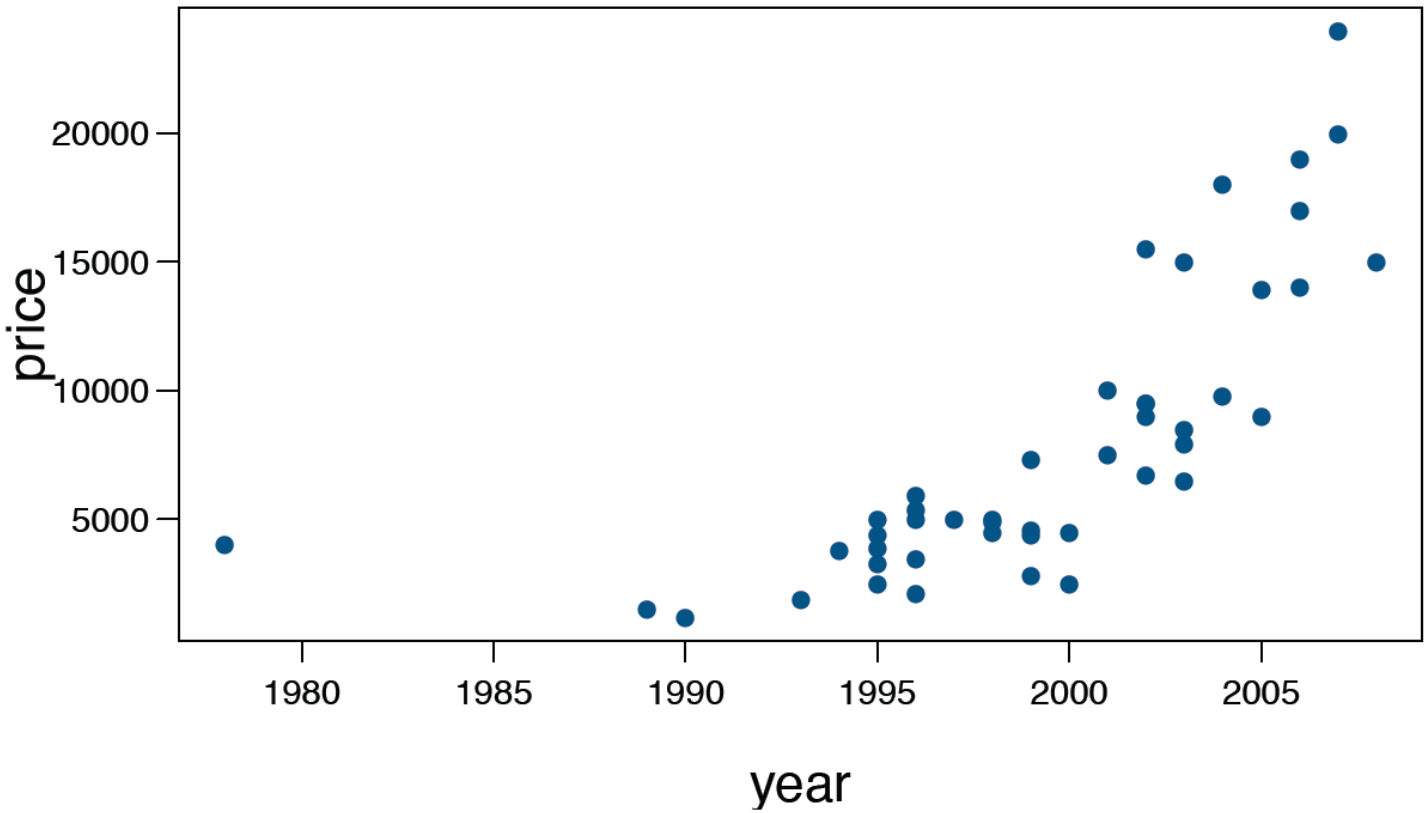


# Summary

1. **Leverage** point is away from the cloud of points horizontally, does not necessarily change the slope
2. **Influential** point changes the slope (most likely also has high leverage) - run the regression with and without that point to determine
3. Don't remove **outliers** without a good reason. Models that ignore exceptional cases often perform poorly
4. If clusters (group of points) are apparent in the data, it might be worthwhile to model the group separately

# Truck prices

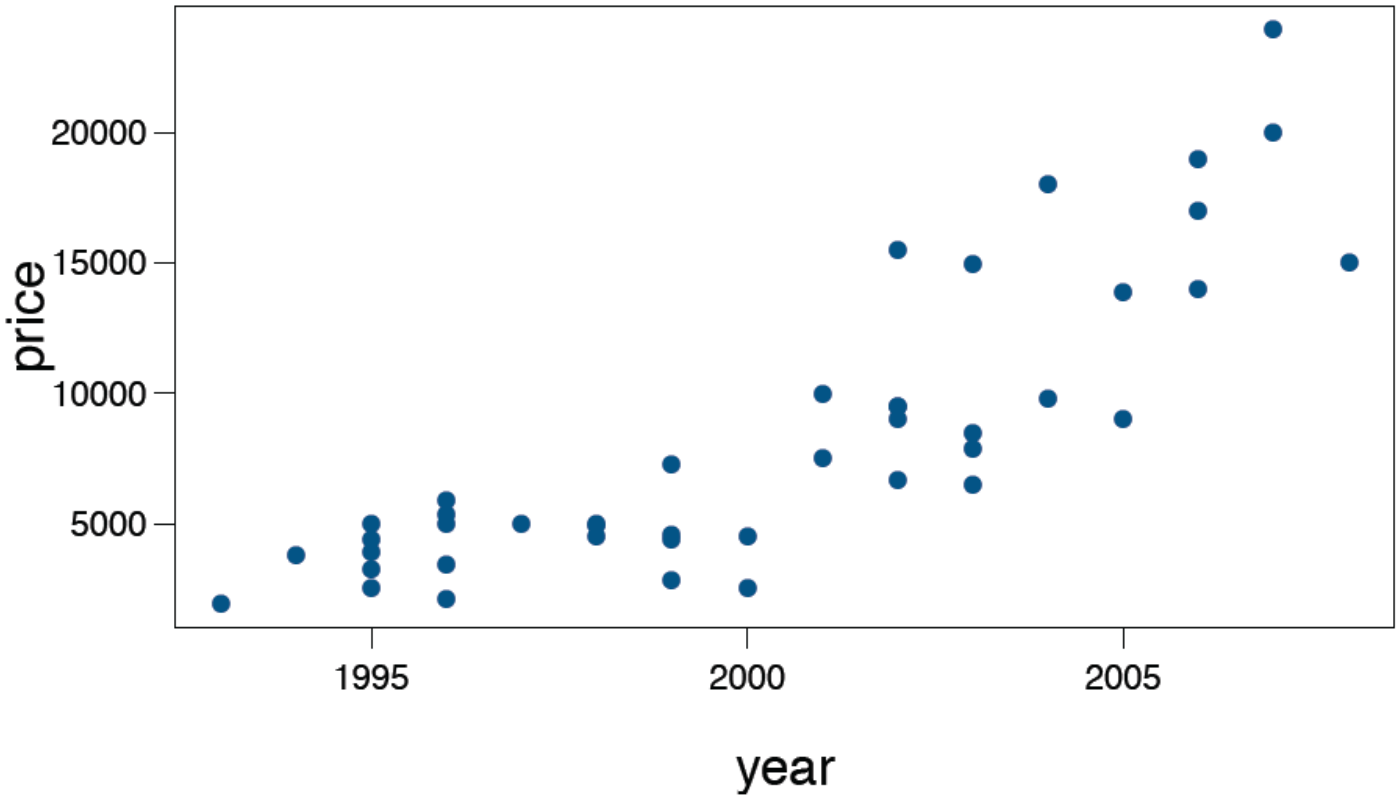
The scatterplot below shows the relationship between year and price of a random sample of 43 pickup trucks. Describe the relationship between these two variables.



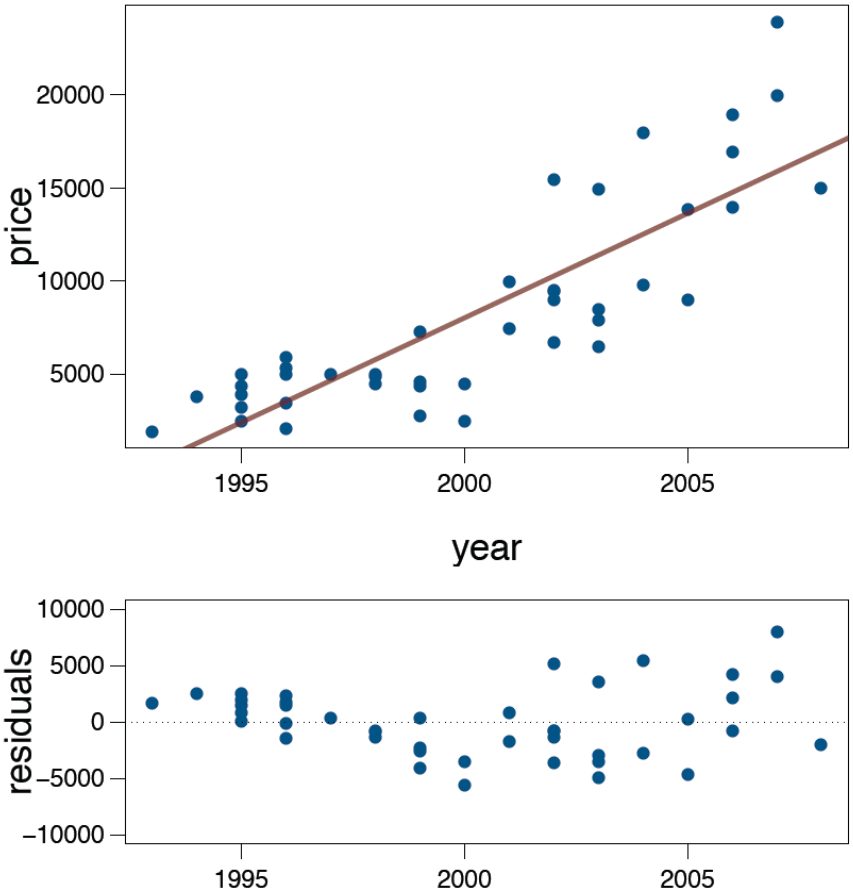

From: `http://faculty.chicagobooth.edu/robert.gramacy/teaching.html`

# Remove unusual obserations

Let's remove trucks older than 20 years, and only focus on trucks made in 1992 or later.
Now what can you say about the relationship?



# Truck prices - linear model?

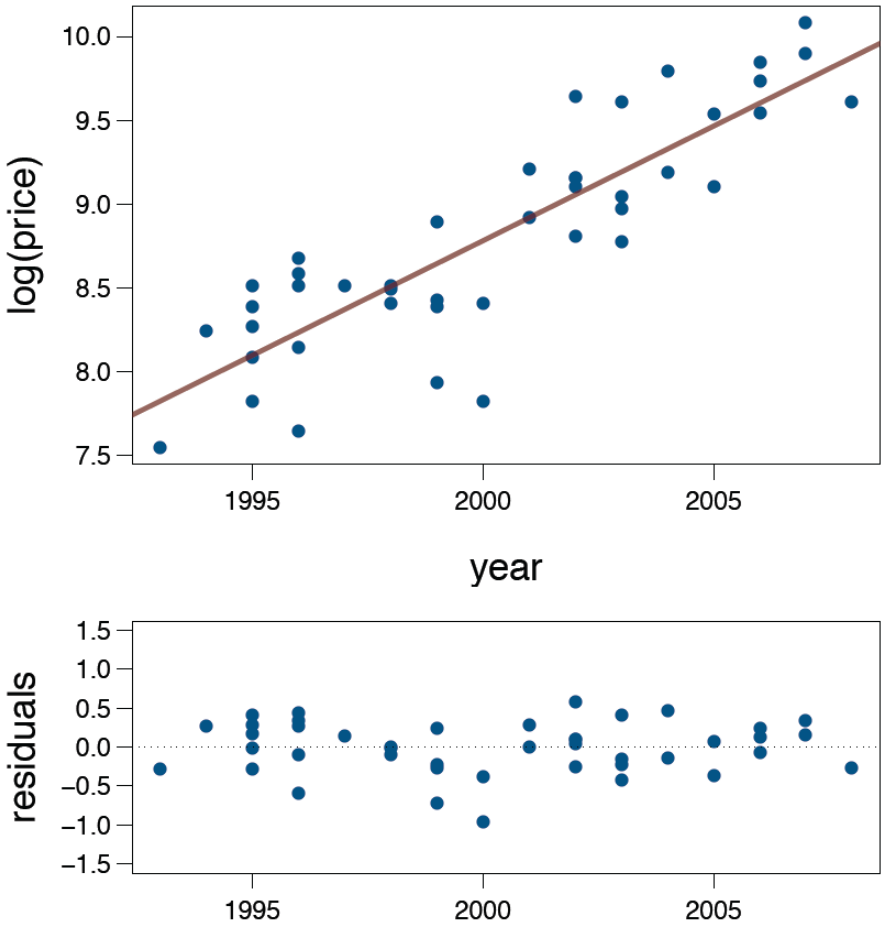



Model:

$$\widehat{\text{price}} = b_0 + b_1 \text{year}$$

The linear model doesn't appear to be a good fit since the residuals have non-constant variance.

# Truck prices - log transform of the response variable




Model:

$$\widehat{\log(\text{price})} = b_0 + b_1 \text{year}$$

We applied a log transformation to the response variable. The relationship now seems linear, and the residuals no longer have non-constant variance.

# Interpreting models with log transformation

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | -265.07 | 25.04 | -10.59 | 0.00 |
| pu$year | 0.14 | 0.01 | 10.94 | 0.00 |

$$\text{Model:} \quad \widehat{\log(\text{price})} = -265.07 + 0.14\text{year}$$

- For each additional year the car is newer (for each year decrease in car's age) we would expect the log price of the car to increase on average by 0.14 log dollars
- which is not very useful...

# Working with logs

- Subtraction and logs: $\log(a) - \log(b) = \log(a/b)$
- Natural logarithm: $\exp(\log(x)) = x$
- We can use these identities to "undo" the log transformation

# Interpreting models with log transformation

The slope coefficient for the log transformed model is 0.14, meaning the log price difference between cars are one year apart is predicted to be 0.14 log dollars.

$$\log\left(\text{price at year } x+1\right) - \log\left(\text{price at year } x\right) = 0.14$$

$$\log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right) = 0.14$$

$$\exp\left[\log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right)\right] = \exp\left[0.14\right]$$

$$\frac{\text{price at year } x+1}{\text{price at year } x} = 1.15$$

For each additional year the car is newer (for each year decrease in car's age) we would expect the price of the car to increase on average by a factor of 1.15.

# Dealing with non-constant variance

- Non-constant variance is one of the most common model violations, however it is usually fixable by transforming the response $y$ variable
- The most common variance stabilizing transform is the log transformation: $\log(y)$, especially useful when the response variable is (extremely) right skewed
- When using a log transformation on the response variable the interpretation of the slope changes: For each unit increase in $x$, $y$ is expected on average to decrease/increase by a factor of $e^{b_1}$
- Another useful transformation is the square root: $\sqrt{y}$, especially useful when the response variable is counts.
- These transformation may also be useful when the relationship is non-linear, but in those cases a polynomial regression may also be needed.

# Recap

Today we talked about,

- Conditions for the least squares line
- Categorical explanatory variables
- $R^2$
- Types of outliers in linear regression
- Log transformation

Suggested reading:

- D.S. Sec. 11.1, 11.2
- OpenIntro3: 7.3