CHAPTER 27.3

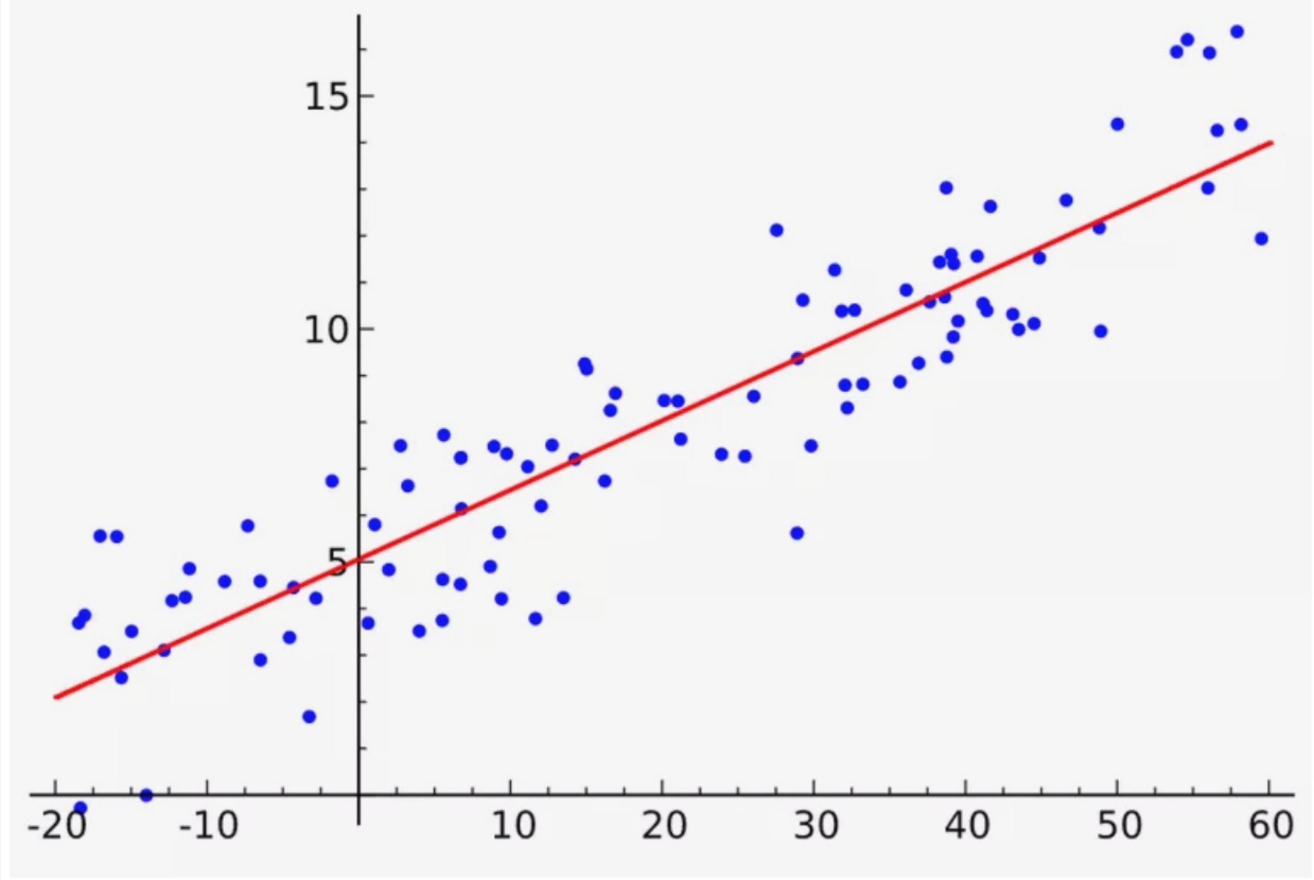
Inference in Simple Linear Regression

Agenda for the day:

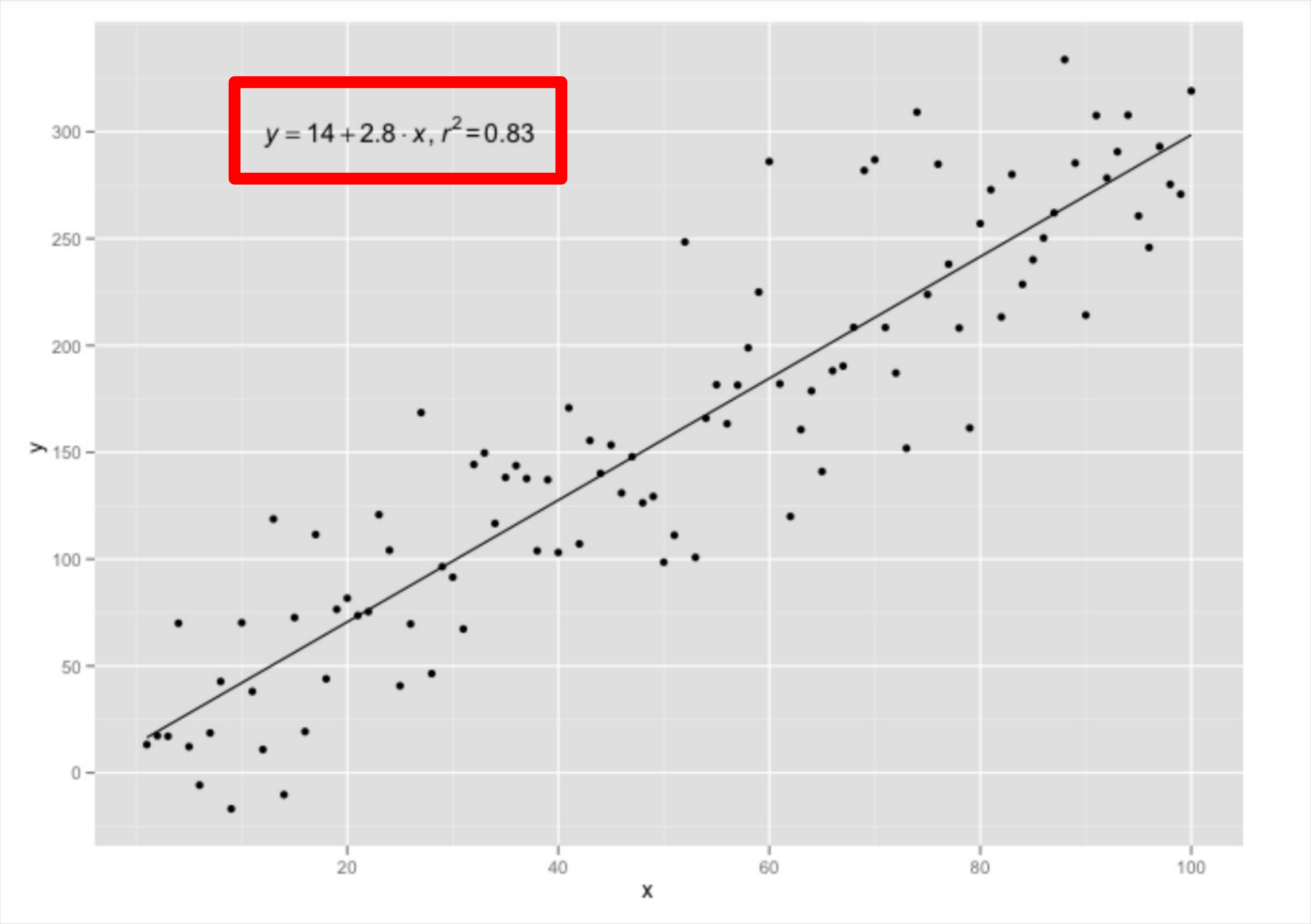
- Estimate the variance in data about the true regression line

- Quantify the goodness of fit of our SLR model.

- Perform inference on the regression parameter, β_1, slope.



So, now that we can create regression lines, we ask how good are they?



Four ideas to watch for throughout this slide deck:

1] r, the (Pearson's) correlation coefficient.

This quantifies the strength of the linear relationship (and direction) between two variables in a correlation analysis.

2] R^2, the coefficient of determination.

This is a goodness of fit measure. R^2 is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable in a regression model.

3] Residuals.

What are they? What is their distribution? $\epsilon \sim N(0, \sigma^2)$

4] Confidence intervals for the slope of a regression line.

The regression line is a statistic from a sample.

How well does it predict the parameter from the population? Ans: Build a CI.

We often hear/see headlines like:

> "Drinking a glass of red wine per day may decrease your chances of a heart attack."
>
> "Eating lots of fish may improve your brain health."
>
> "Driving slower reduces your chances of getting killed in an accident."
>
> "Smoking may lead to low birthweight babies."
>
> "The more you study, the higher your grades will likely be."

The correlation, denoted by r, measures the amount of linear association between two variables. It is always true that $-1 \leq r \leq 1$.

R^2 is the square of correlation.
It measures the proportion of variation in the dependent variable that can be attributed to the independent variable. It is always true that $0 \leq R^2 \leq 1$.

How high must a correlation be to be considered meaningful?

It depends on the area of study.

Roughly speaking:

Area of Study	r **is meaningful**	R^2 **is meaningful**
Physics	$r < -0.95$ or $0.95 < r$	$0.9 < R^2$
Chemistry	$r < -0.90$ or $0.90 < r$	$0.8 < R^2$
Biology	$r < -0.70$ or $0.70 < r$	$0.5 < R^2$
Social Sciences	$r < -0.60$ or $0.60 < r$	$0.35 < R^2$

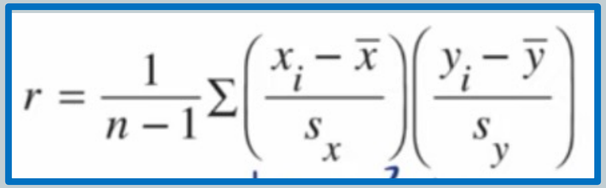
Physical processes are pretty consistent during experimentation.
Animals, people especially, are too flaky (variable).

We won't be calculating r or R^2 by hand; after all, we aren't cave people.
But it is wise to have some idea as to their origin/creation.

Suppose your data was: $\{(1,5),(3,9),(4,7),(5,1),(7,13)\}$

Correlation, r, is the mean of the z-scores of the x's and y's.

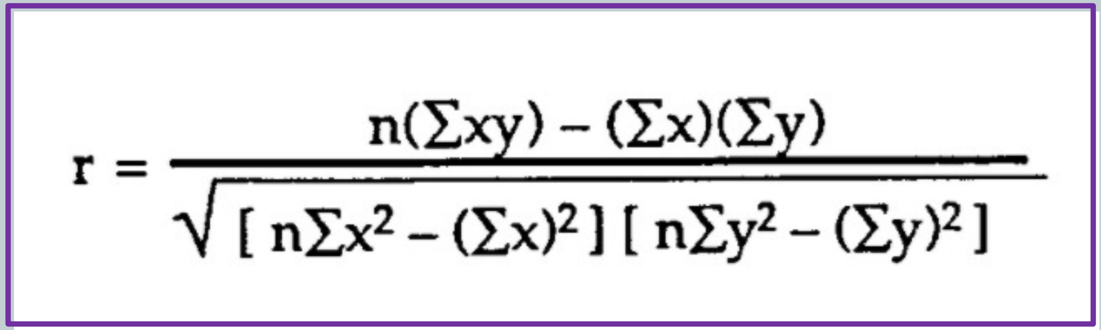
$$Z_{x_i} = \frac{(x_i - \bar{x})}{s_x} \quad \text{and} \quad Z_{y_i} = \frac{(y_i - \bar{y})}{s_y}$$



$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$r = \frac{\sum Z_{x_i} Z_{y_i}}{n-1} = 0.40$$

Alternatively



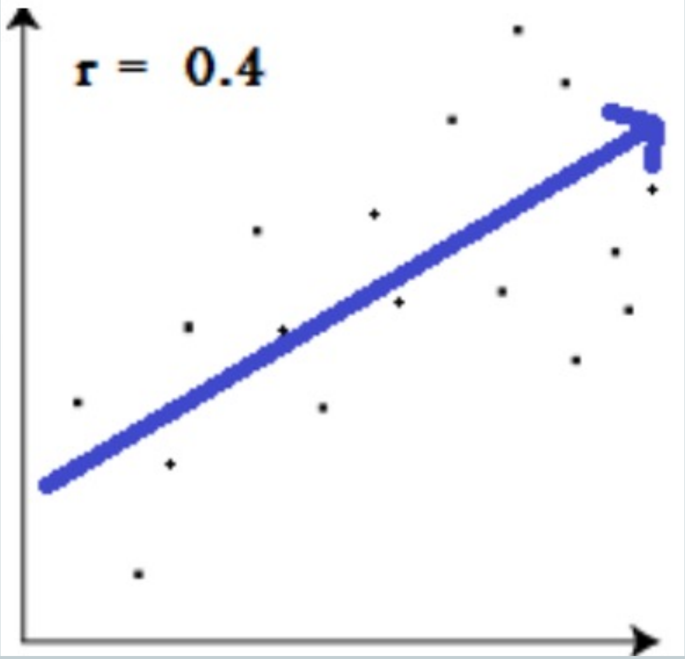
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

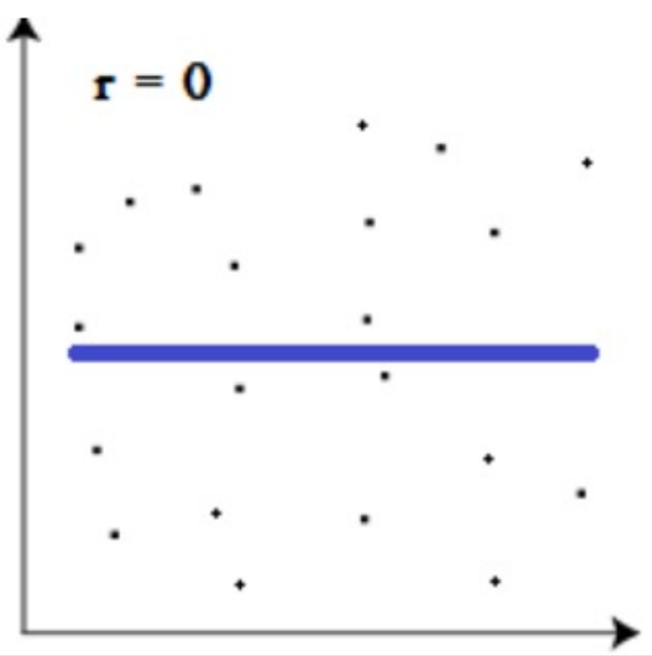
What does $r = 0.40$ mean?

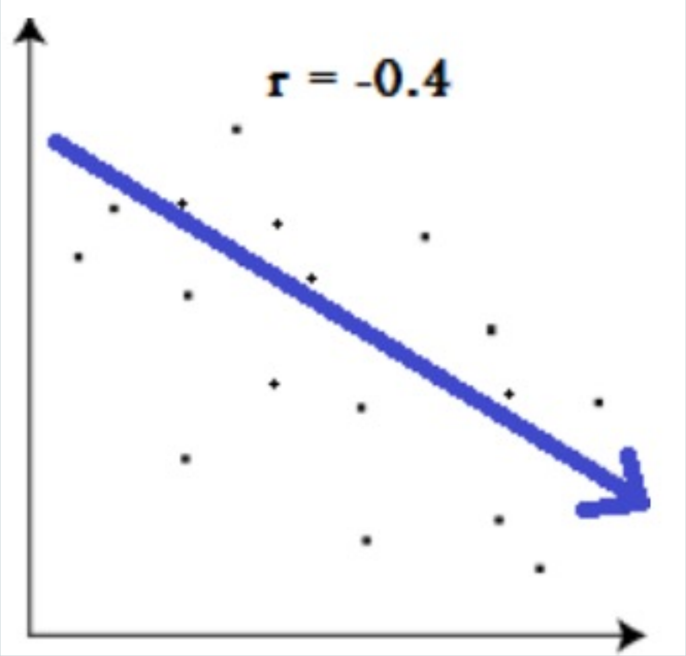
An $r = 1$ indicates a strong positive relationship

An $r = -1$ indicates a strong negative relationship

An $r = 0$ indicates no relationship at all







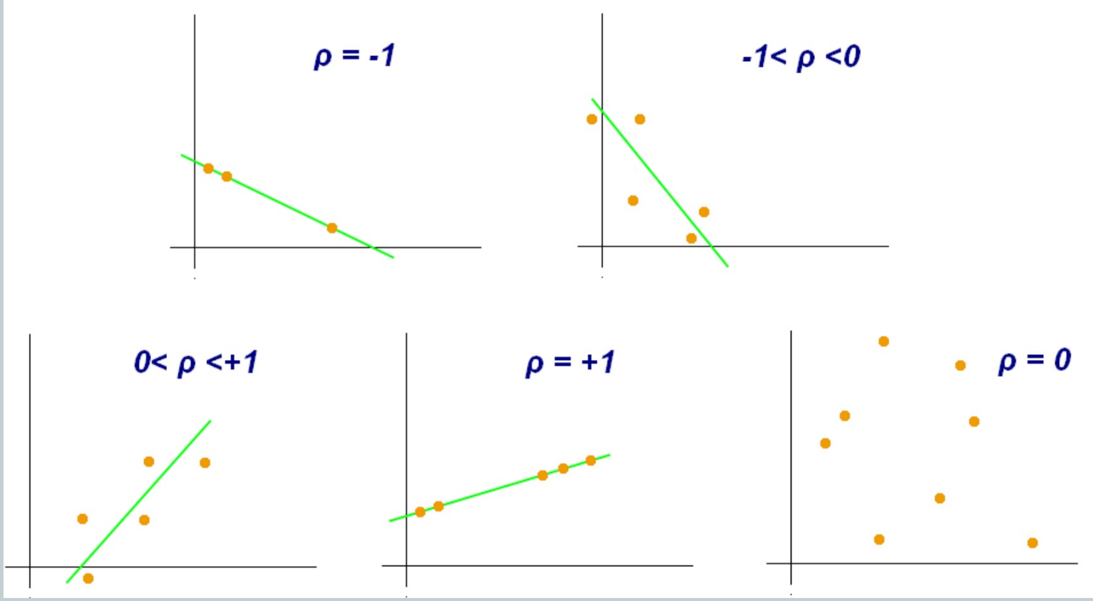
What would you expect r to be for correlating the price of gas and weekend traffic?

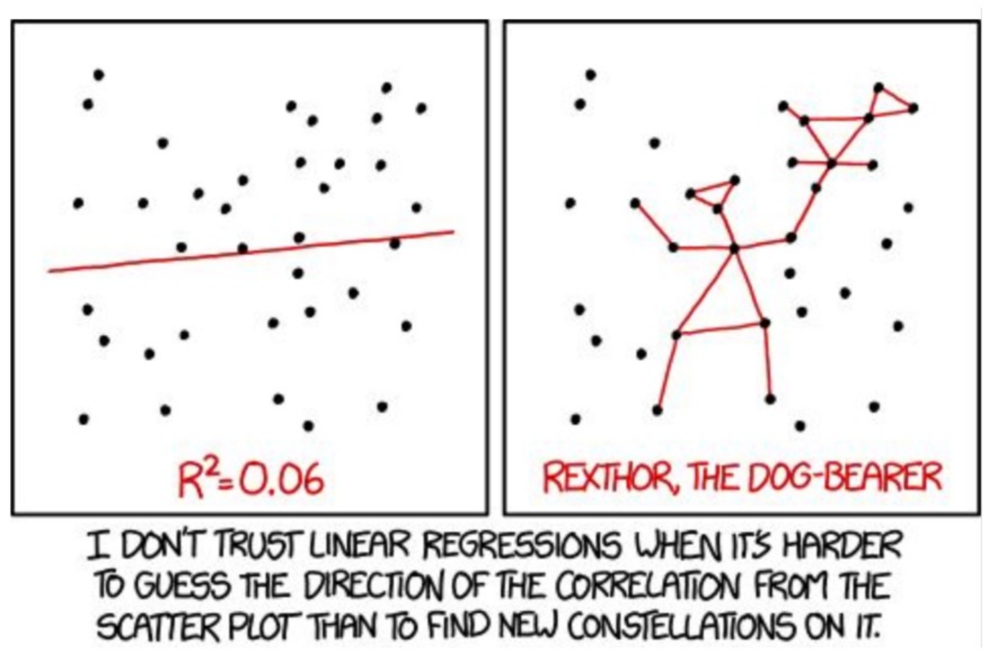
Positive, negative, or zero?

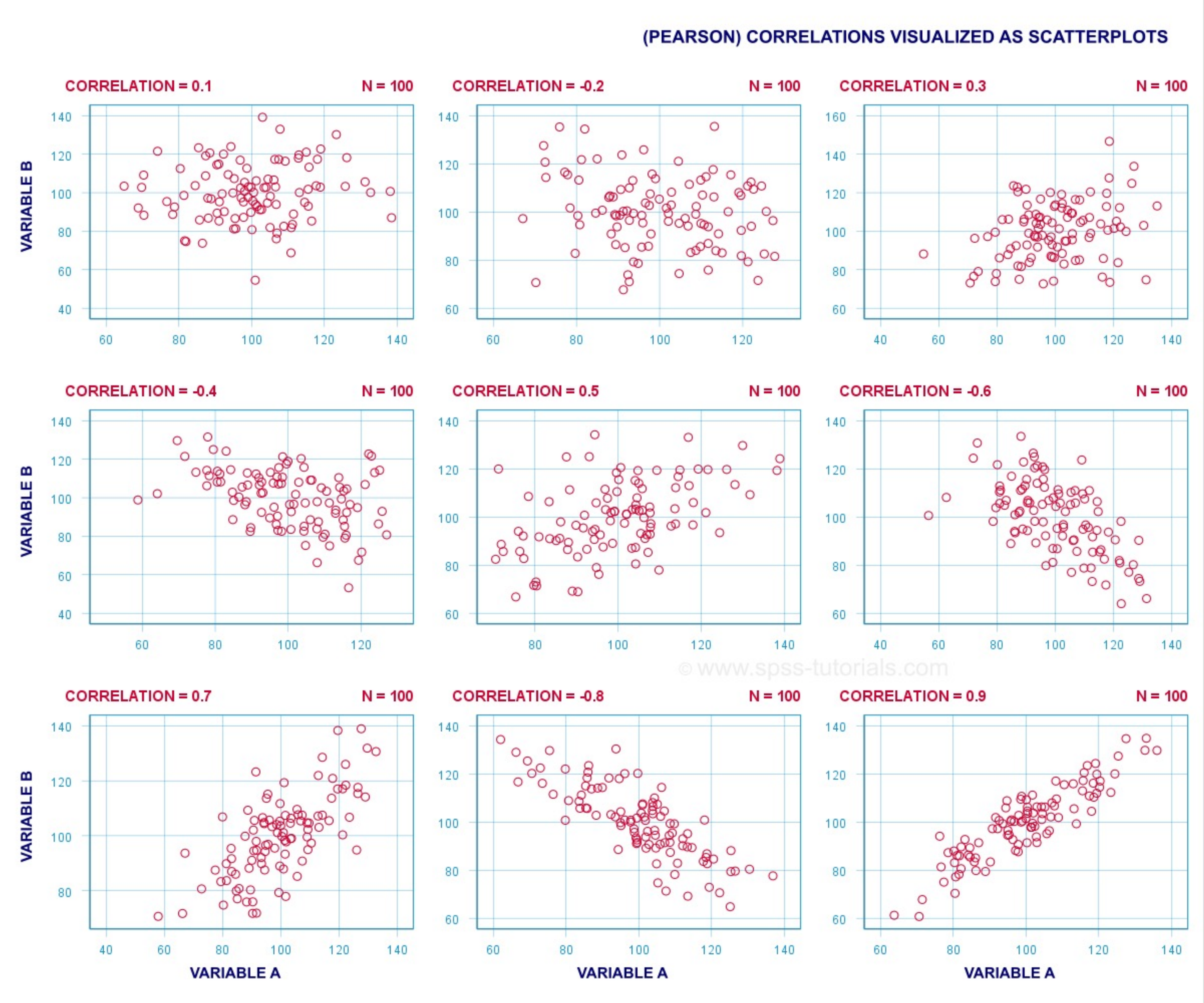
What would you expect r to be for correlating shoe size and intelligence?

Causation versus correlation.

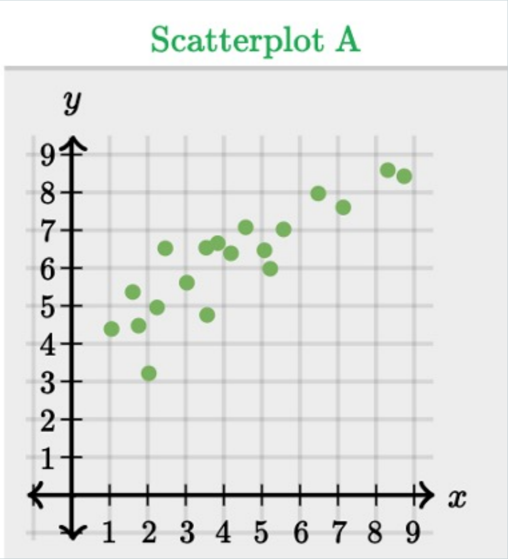
What does Pearson's r correlation look like?




R²=0.06
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.



Can you match the scatterplot to the correlation coefficient?

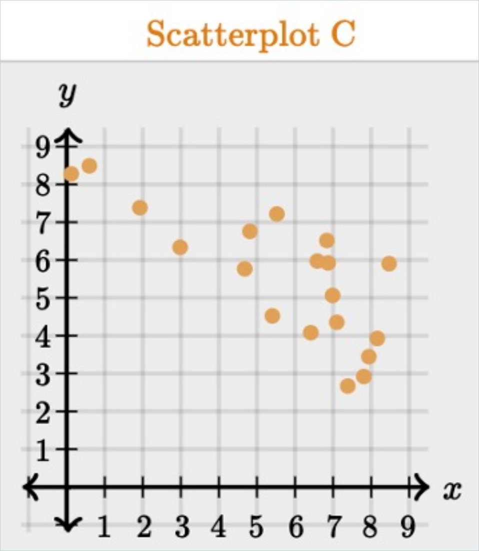


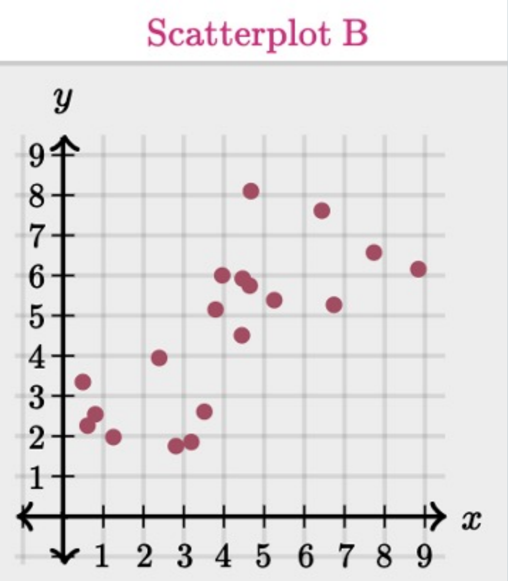
r= - 0.42

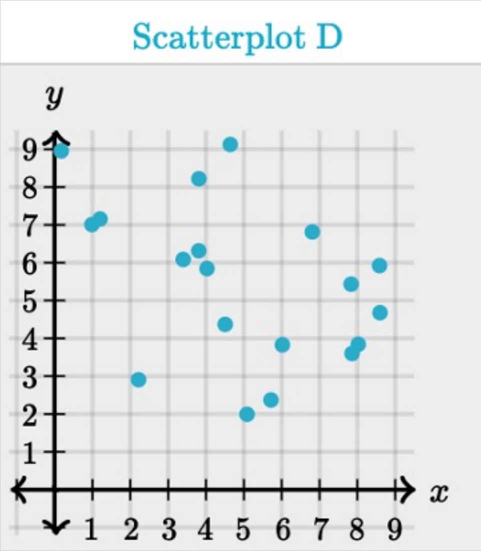
r= - 0.77

r=0.73

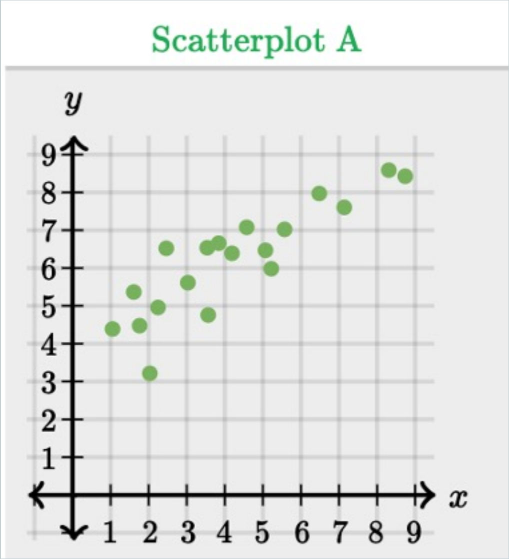
r=0.87

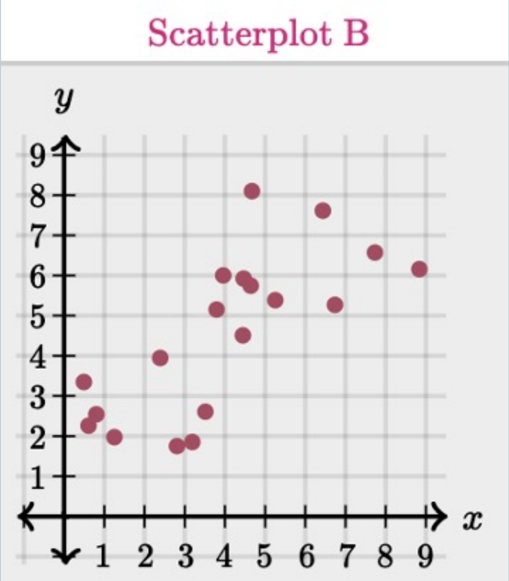






Can you match the scatterplot to the correlation coefficient?



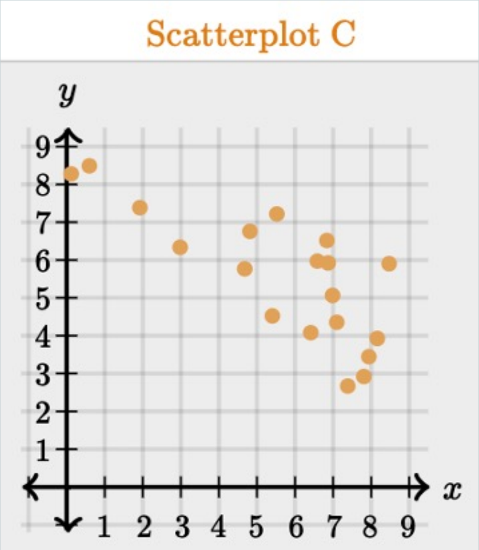


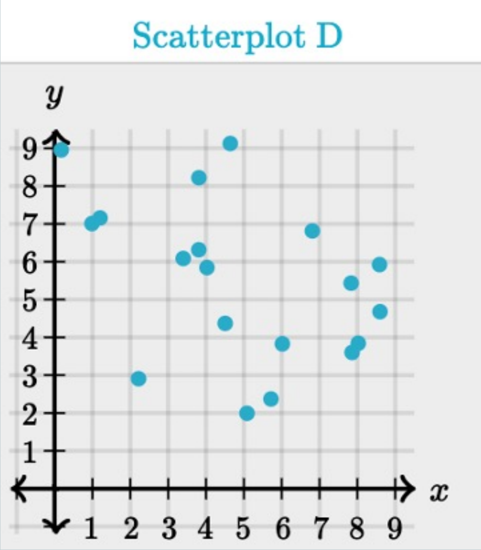
r= - 0.42

r= - 0.77

r=0.73

r=0.87





r is a correlation coefficient. What is R^2 ?

R^2 is the square of r

R^2 is a goodness-of-fit measure for our linear regression models.

R^2 indicates the percentage of the variance in the dependent variable that the independent variable(s) explain.

R^2 measures the strength of the relationship between the model and the dependent variable on a $0 – 100\%$ scale.

After fitting a linear regression model, you need to determine how well the model fits the data. How well does it explain changes in the dependent variable?

Note:

Small R^2 values are not always bad.

High R^2 values are not always good.

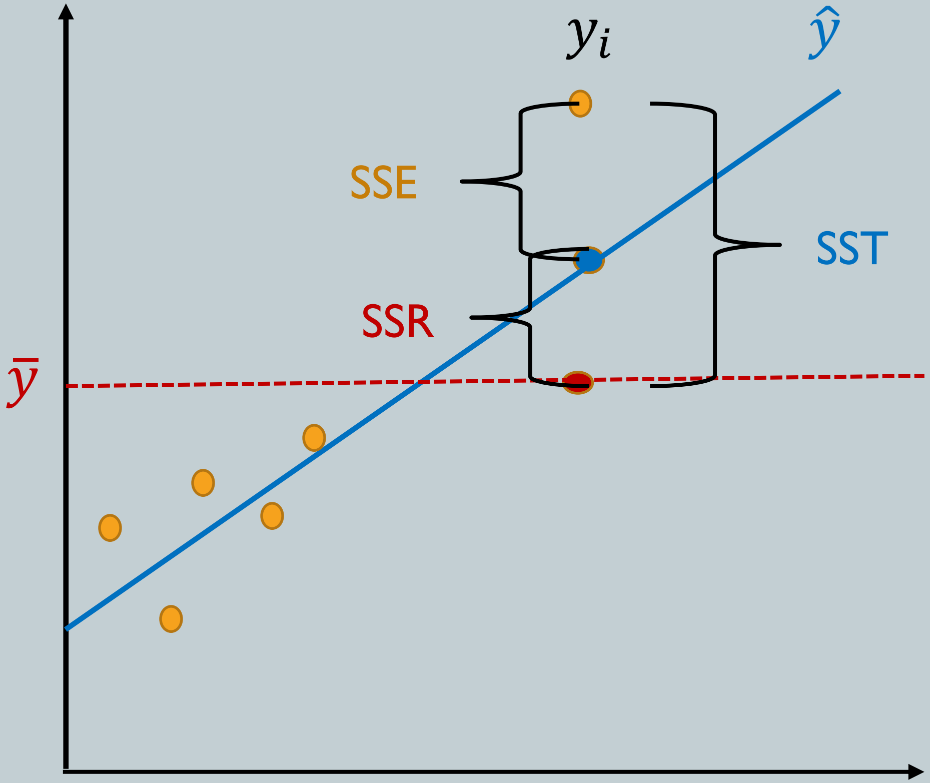
Once we have a regression line, it is important to quantify how well it fits the data.

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SST - SSE}{SST} = \frac{\text{Variance explained by model}}{\text{total variance}}$$

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

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

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



SSE is the sum of squares of errors, SSR is the sum of squares due to regression, SST is the sum of squares total

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

Gives a sense of how much variation in y is left unexplained by the model.

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

Gives a sense of how much variation in y is explained by our model.

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

Gives a sense of the total amount of variation observed in y values.

SST is what we would get for SSE if we used the mean of the data as our model.

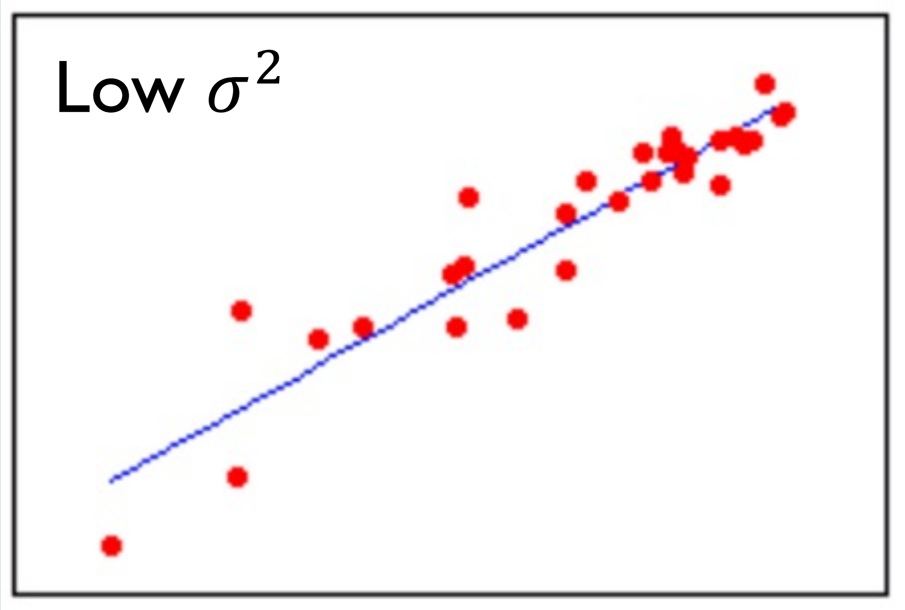
Estimating the variance

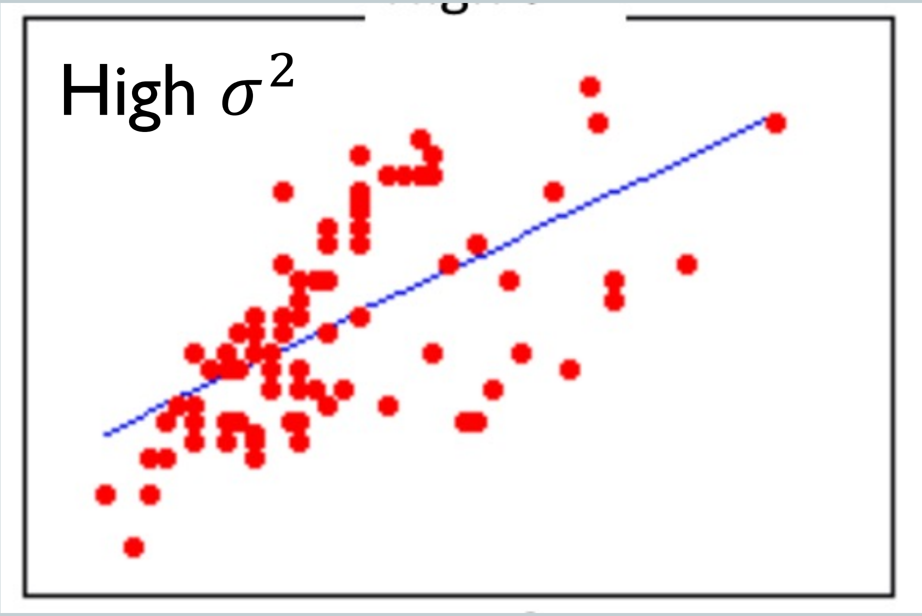
The parameters σ^2 determines the spread of the data about the true regression line.

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

An estimate of σ^2 will be used in computing confidence intervals and doing hypothesis testing on the estimated regression parameters.

Our estimate of the variance is $\boldsymbol{\hat{\sigma}^2 = \frac{SSE}{n-2}}$





R^2 is the coefficient of determination

The sum of squared deviations about the least-squares line is smaller than the sum of squared deviations about any other line: $SSE \leq SST$.

The ratio ${}^{SSE}/_{SST}$ is the proportion of total variation in the data (SST) that cannot be explained by the SLR model (SSE).

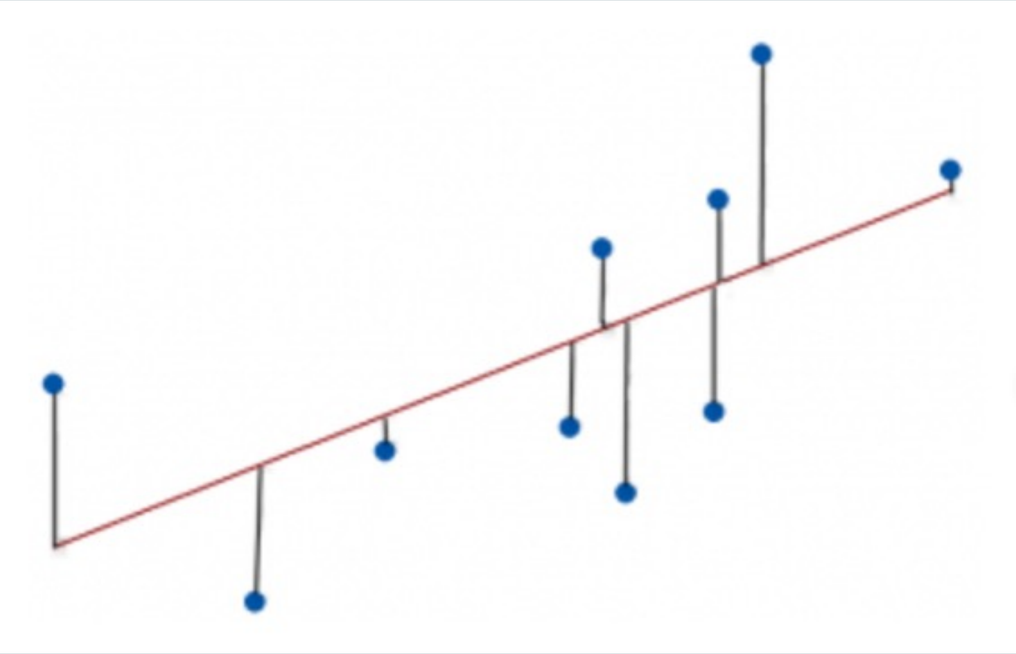
So, we define the coefficient of determination R^2 to be the proportion that can be explained by the model.

$$R^2 = 1 - \frac{SSE}{SST}$$

Remember, linear regression identifies the equation that produces the smallest difference between all the observed values and their fitted values.

Linear regression finds the smallest sum of squared residuals possible for the data.

Residuals are the distance between the observed value and the fitted value



The sum of the squares of the lengths of the vertical lines

The regression model fits the data well when the residuals are small and unbiased.

Small means short vertical lines.
unbiased means the fitted values are not systematically too high or too low anywhere in the observation space.

Note: Prior to assessing R^2, you should evaluate the residual plot.
Visually, these plots can expose biased models far more effectively.

If the model is biased, then do not trust the results; R^2.
If the residual plot looks good, then you can assess R^2.

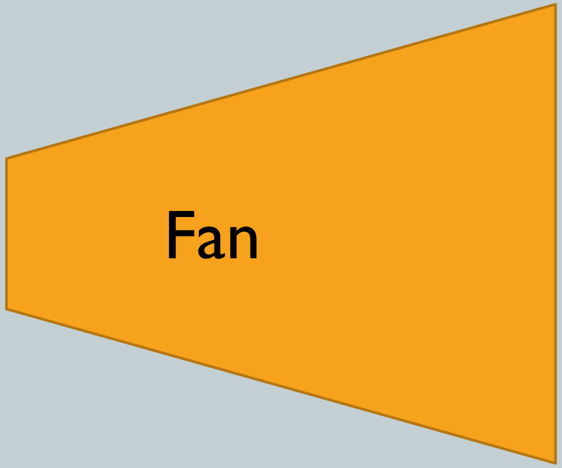
A residual plot puts the residuals on the y-axis and the predicted values of the dependent variable on the x-axis.

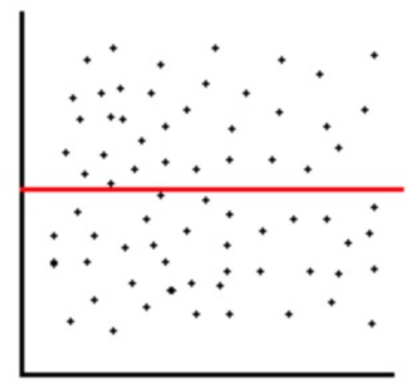
We want the residuals to be:

Unbiased – have an average value of 0 in any thin vertical strip

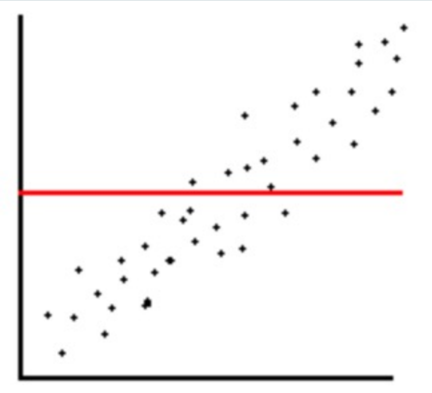
Homoscedastic – "same stretch" the spread of the residuals should be the same in any thin vertical strip.

The residuals are heteroscedastic if they are **not** homoscedastic.
Heteroscedastic graphs are recognizable due to their 'fan' shape.

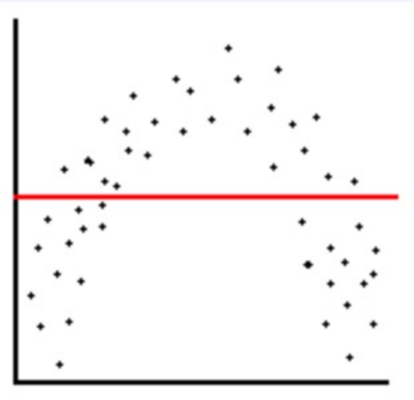




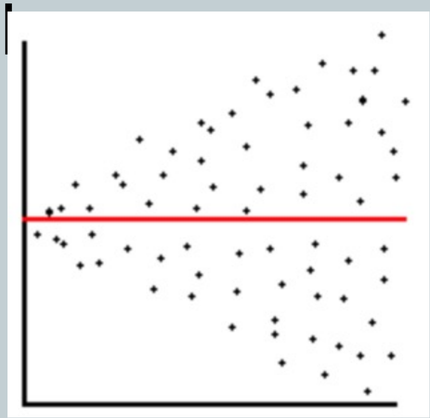
Unbiased and homoscedastic (desirable)
The residuals average to 0 in each thin vertical strip and the SD is the same all across the plot



Biased and homoscedastic
The residuals show a linear pattern, probably due to a lurking variable not included in the experiment.

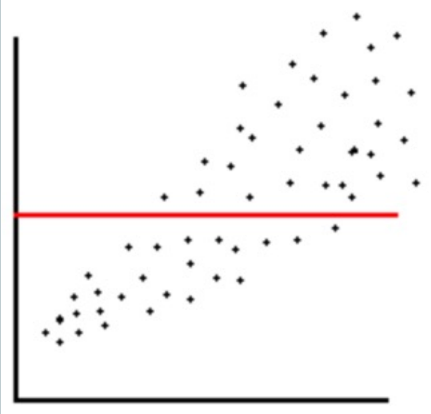


Also Biased and homoscedastic
The residuals show a quadratic pattern, possibly because of a nonlinear relationship. Sometimes a variable transform will eliminate the bias.



Unbiased but heteroscedastic

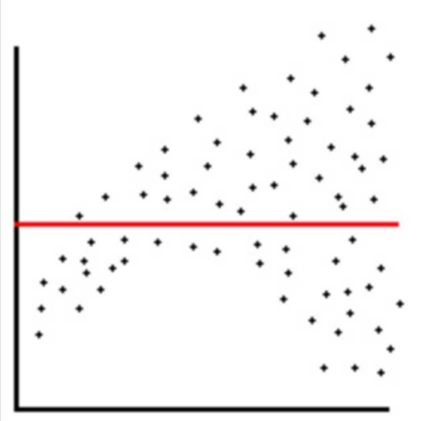
The SD is small to the left of the plot and large to the right: the residuals are heteroscadastic.



Biased and heteroscedastic.

The pattern is linear.

We would also like the residuals to be normally distributed.
We check this by looking at the <u>normal plot</u> of the residuals



Also Biased and heteroscedastic.

The pattern is quadratic

R^2 evaluates the scatter of the data points around the fitted regression line.

Higher R^2 values represent smaller differences in residuals.

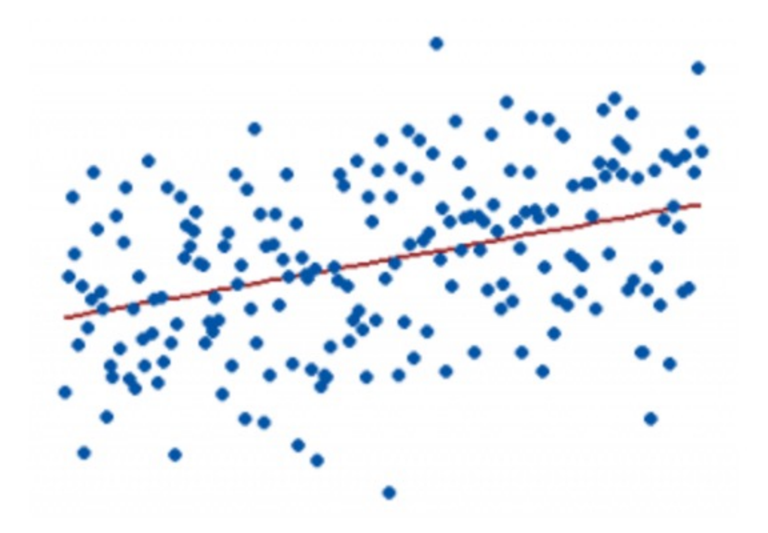
R^2 is the percentage of the dependent variables variation that a linear model explains.

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

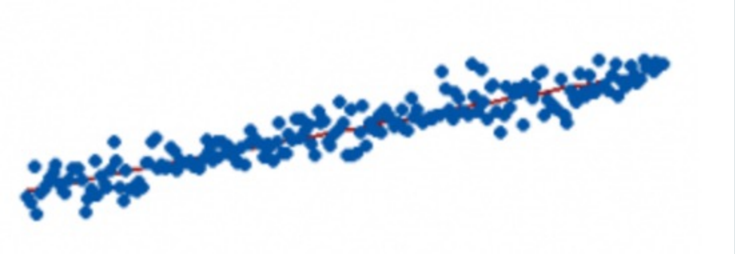
$R^2 = .3$ means that 30% of points fall on the regression line

R^2 graphically

R^2 for this regression model is15%



R^2 for this regression model is 85%



R^2 Caveats

Note that R^2 does not indicate if a regression model provides an adequate fit to your data.

A good model can have a low R^2 value, and a biased model can have a high R^2 value.

You cannot use R^2 to determine whether the coefficients estimates and predictions are biased, which is why you must first assess the residual plots.

If a person steps on a scale, and the scale reads 400 pounds, this does not necessarily mean they are large. First assess what they are wearing; backpack full of lead bricks.

If a professional boxer amasses a record of $75 – 0$, this does not necessarily mean they are a great boxer. First assess who their opponents were.
Great boxers have losses on their record, mediocre boxers can have spotless records.

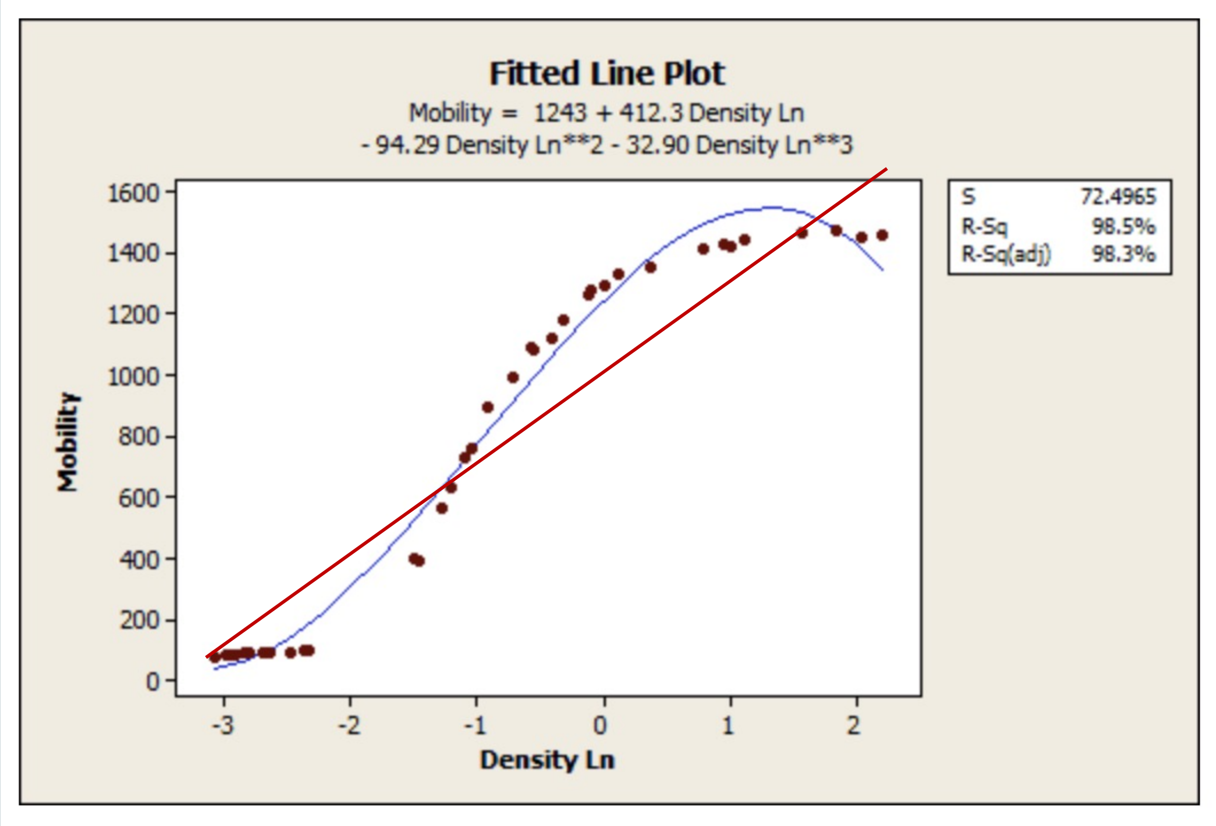
Regression models can be perfectly good models and have a low R^2.
Studies with humans have notoriously low R^2. Human behavior is hard to predict.

If you have a low R^2 but the independent variables are statistically significant, you can still draw important conclusions about the relationships between the variables.

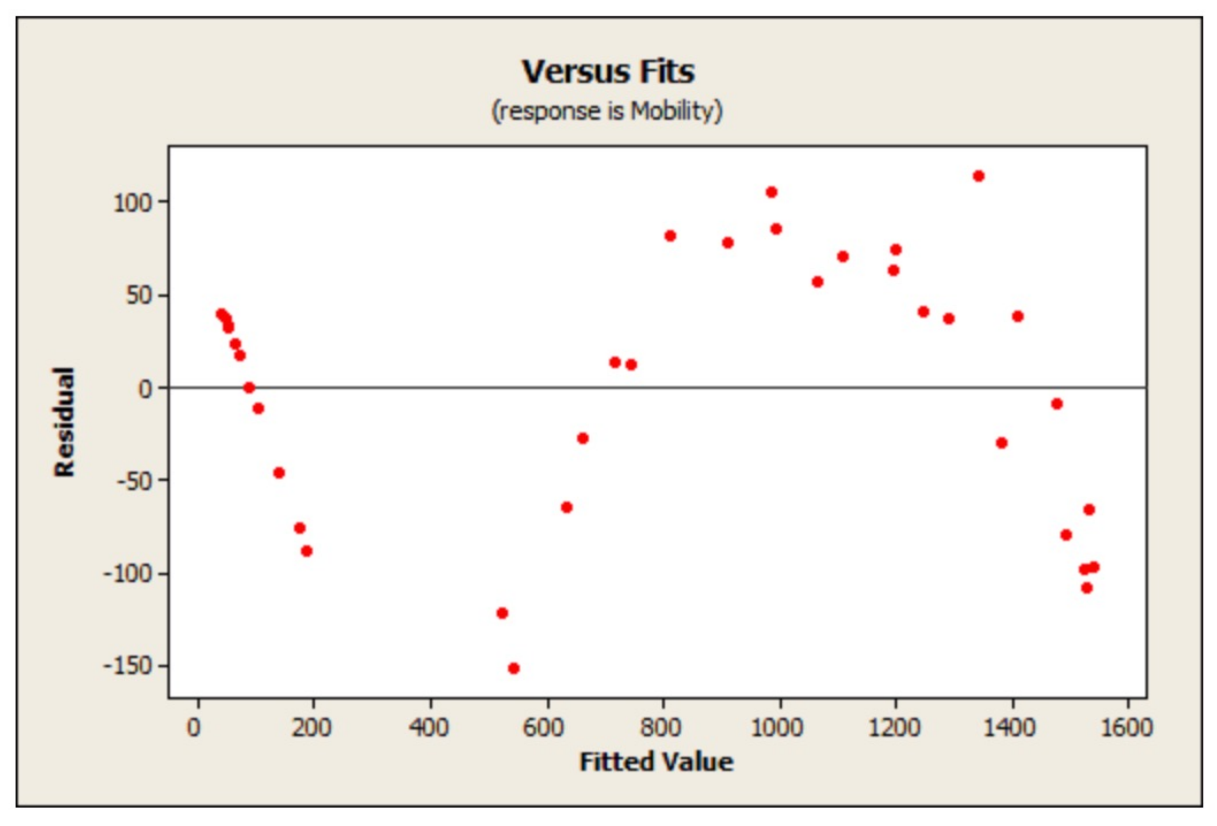
Alternatively, generating relatively precise predictions with low R^2 should not be attempted; it is problematic.

A high R^2 is necessary for precise predictions.

A high R^2 value does not mean there are no problems



$R^2 = 98.5\%$



Residual Plot

The data in the fitted line plot follow a very low noise relationship and $R^2 = 98.5\%$. This seems fantastic but the regression line consistently under and over-predicts the data along the curve, which is bias.

The residuals versus the fits plot emphasizes this unwanted pattern.
An unbiased model has residuals that are randomly scattered around zero.
Non-random residual patterns indicate a bad fit despite a high R^2.
Always check your residual plots!

This type of specification bias occurs when your linear model is underspecified. In other words, it is missing significant independent variables, polynomial terms, and interaction terms. (Next lecture: Multiple Linear Regression)

To produce random residuals, try adding terms to the model or fitting a **nonlinear** model.

A variety of other circumstances can artificially inflate R^2.
These reason include overfitting the model and data mining.

Either of these can produce a model that looks like it provides an excellent fit to the data but in reality the results can be entirely deceptive.

An overfit model is one where the model fits the random quirks of the sample. Data mining can take advantage of chance correlations.

In either case, you can obtain a model with a high R^2, **even for entirely random data.**

Data Mining: The process of digging through data to discover hidden connections and predict future trends. Sometimes referred to as "knowledge discovery in databases.”
Data mining is considered a mixture of three disciplines:
statistics, artificial intelligence, machine learning.

Now, let's make inferences from a regression line.

Why?

A regression line comes from a sample.

Recall that statistics come from a sample and they are used to estimate the parameters of a population.

$\hat{y}$ (the regression line) is a statistic.

It will be different for every sample set.

(Recall NB 20)

Recall that given data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ we fit a simple linear regression of the form $y_i = \alpha + \beta x_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$

Estimates of the intercept and slope parameters are given by minimizing:

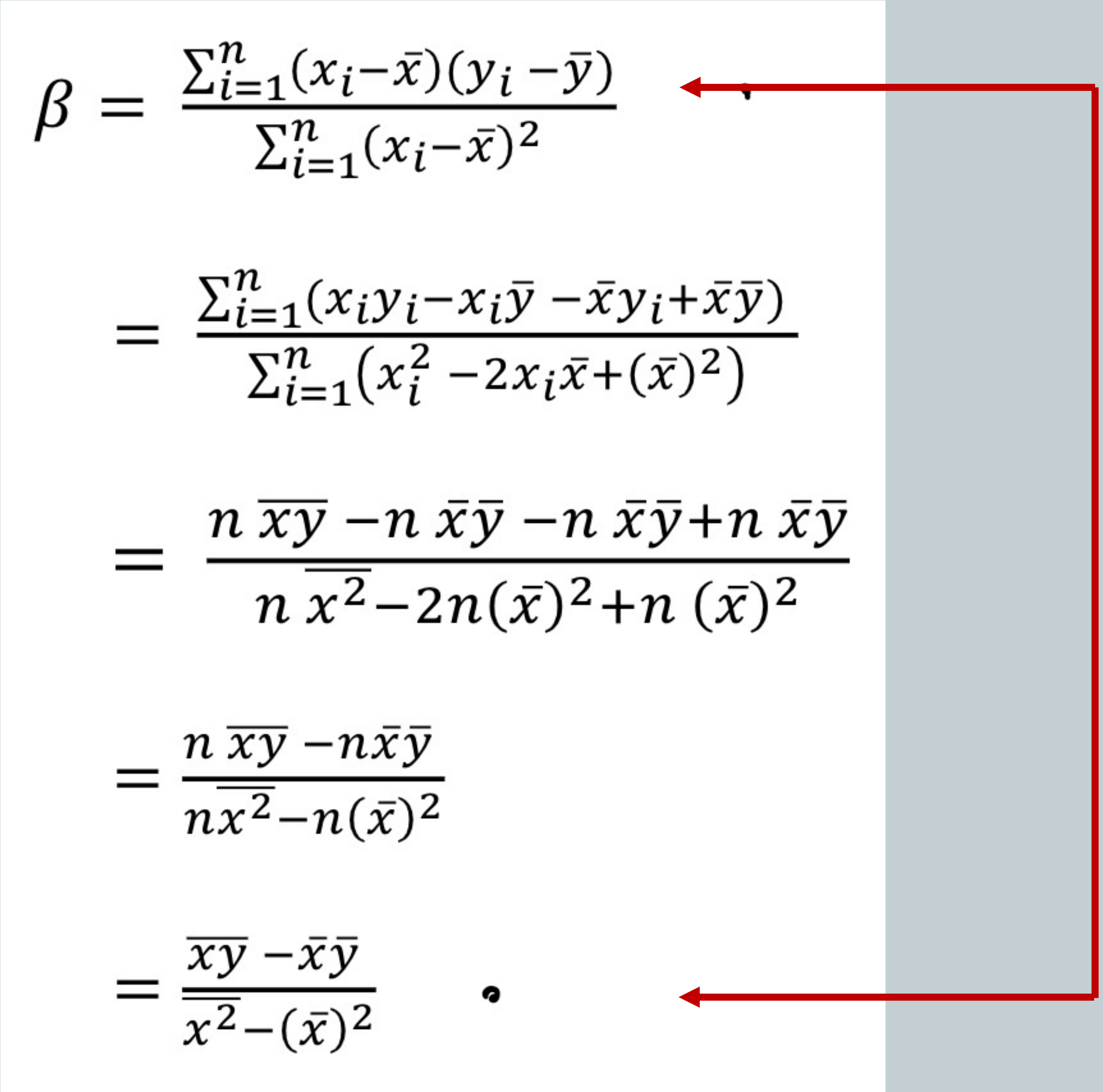
$$SSE = \sum_{i=1}^{n} \left(y_i - (\alpha + \beta x_i)\right)^2$$

The least-squares estimates of the parameters are:

$$\hat{\beta} = \frac{\bar{x}\cdot\bar{y} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

The regression line is: $\hat{y} = \hat{\alpha} + \hat{\beta} x$

Calculating β's different forms.



$$\beta = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y})}{\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + (\bar{x})^2\right)}$$

$$= \frac{n\,\overline{xy} - n\,\bar{x}\bar{y} - n\,\bar{x}\bar{y} + n\,\bar{x}\bar{y}}{n\,\overline{x^2} - 2n(\bar{x})^2 + n\,(\bar{x})^2}$$

$$= \frac{n\,\overline{xy} - n\bar{x}\bar{y}}{n\overline{x^2} - n(\bar{x})^2}$$

$$= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} \quad .$$

Our sample $\hat{y}$ is used to estimate the population line of best fit.

We optimize the line to fit the sample data,
but is this same line descriptive of the population?

To answer that, we create a confidence interval for the line's coefficients, α and β.

Inference about SLR parameters

Given data (x, y)

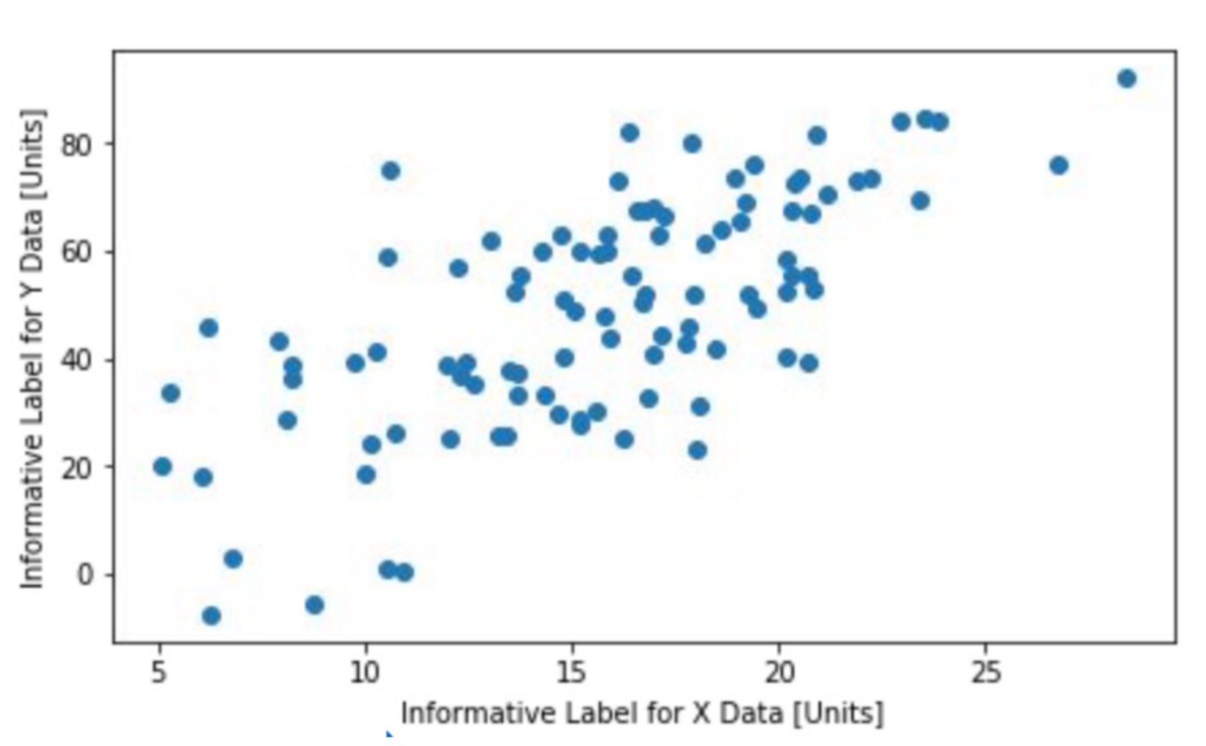
Explore/plot/histograms

Formulate null and alternative hypotheses

Fit a regression line

Compute CI or p-value/rejection region
and test statistic for testing your hypotheses.

Make conclusions



The parameters in the simple linear regression model have distributions.

From these distributions, we can construct CI's for the parameters, conduct hypothesis tests, etc.

We will focus mainly on the slope parameter β to determine if there really is a relationship between the feature (independent) and the response (dependent).

The distribution for the estimate of the slope is given by:

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \quad \longrightarrow \quad SE(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

The confidence interval for β is given by:

$$\hat{\beta} \pm t_{({}^{\alpha}/_{2}, df=n-2)} \cdot \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

And in the case of Hypothesis Testing:

$H_0\colon \beta = c$

$H_1\colon \beta \neq c$

Test statistic: $t = \frac{\widehat{\beta}-c}{SE(\widehat{\beta})}$ compare to $t_{{}^{\alpha}/_{2}, n-2}$ or compute p-value.

Every sample creates a different line with a slightly different intercept and slightly different slope.

These lines are supposed to represent the population.
The $\hat{y}$'s (equations) are just statistics; used to estimate the population parameters.

So, we can make inferences on our sample: $\hat{y} = b_0 + b_1 x$
about the population: $y = \alpha + \beta x.$

Q: How confident are we that b_1 is close to β?

A: Create a confidence interval for our b_1.

Conditions for inference on slope

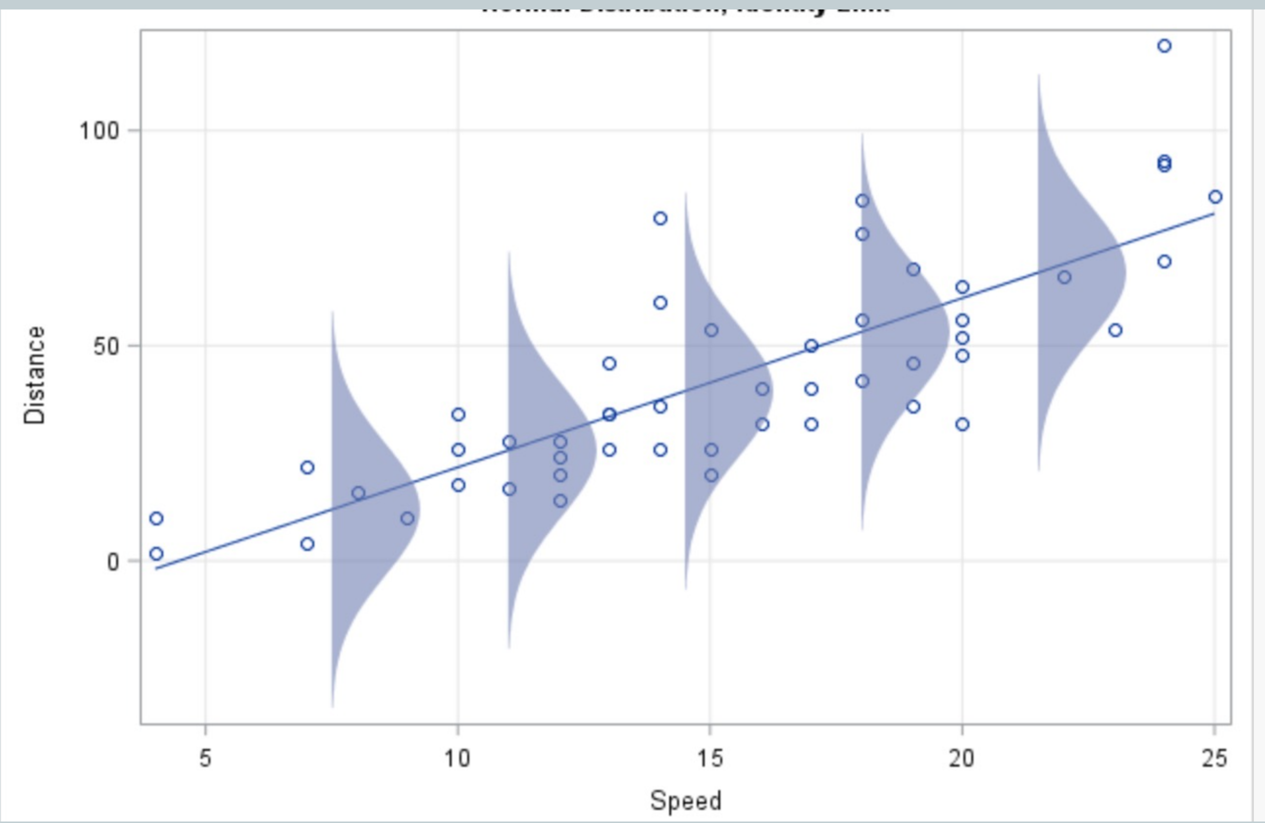
Linear : The relationship between the independent variable and the dependent variable is in fact linear.

Independence : The noise, errors, are independent.

Normal : The errors are normally distributed.

Equal Variance: Homoscedastic errors

Random: Random errors. unbiased.



A $(1-\alpha)100\%$ t-interval for slope parameter β (β is the population slope)

The formula for the confidence interval for β, in words is:

$$\text{Sample estimate } \pm (\text{t\_multiplier } \cdot \text{standard error})$$

In notation this is:

$$b_1 \pm t_{\left({}^{\alpha}/_{2}, n-2\right)} \cdot \left(\frac{\sqrt{MSE}}{\sqrt{\sum (x_i - \bar{x})^2}} \right)$$

The critical value will be a t-score, not a z-score.

$$b_1 \pm t_{\left(\alpha/_{2},n-2\right)} \cdot \left(\frac{\sqrt{MSE}}{\sqrt{\sum(x_i - \bar{x})^2}}\right)$$

Why a t-score?
Because the variance of the error must be estimated from the residuals.
This is then used to calculate the standard error of the regression coefficients and predictions. Where "standard error" is the standard deviation of the sampling distribution of the sample statistic. i.e., The sample statistic is the regression line slope.

But of course, we don't know the SD, so we estimate with standard error of the statistic.

Recall MSE $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$$b_1 \pm t_{(\alpha/_{2},n-2)} \cdot \left(\frac{\sqrt{MSE}}{\sqrt{\sum(x_i - \bar{x})^2}}\right)$$

This confidence interval not only gives us a range of values that is likely to contain the true unknown value β, it also allows us to answer the research question:

“Is the predictor x related to the response y?”

If the confidence interval for β contains 0, then we conclude that there is no evidence of a relationship between the predictor x and the response y in the population.

On the other hand, if the confidence interval for β does not contain 0, then we conclude that there is evidence of a relationship between the predictor x and the response y in the population.

How do we run a hypothesis test for the slope parameter β_1 ?

We follow standard hypothesis test procedures in conducting a hypothesis test for the slope β. First, we specify the null and alternative hypothesis:

$$H_0: \beta = \text{some number b}$$
$$H_1: \beta \neq \text{some number b}$$

The phrase "Some number b" means that you can test whether or not the population slope takes on any value.

Most often, however, we are interested in testing whether $\beta = 0$.

H_1 can also state that $\beta <$ some number b, or $\beta >$ some number b.

Second, we calculate the value of the test statistic using the formula:

$$t = \frac{b_1 - \beta}{\left(\frac{\sqrt{MSE}}{\sqrt{\sum(x_i - \bar{x})^2}}\right)} = \frac{b_1 - \beta}{SE(b_1)}$$

Third, use the resulting test statistic to calculate the p-value.
As always, the p-value is the answer to the question "How likely is it that we'd get a test statistic t as extreme as we did if the null hypothesis were true?"

The p-value is determined by referring to a t-distribution with $n - 2$ degrees of freedom.
This linear regression model has two degrees of freedom either because there are two parameters in the model that must be estimated from a training dataset, or because we subtract 1 degree of freedom due to regression from the total degrees of freedom; $n - 1$.

Finally, make a decision:

If the p-value is smaller than the significance level α, we reject the null hypothesis in favor of the alternative.

Conclude: "There is sufficient evidence at the α level to conclude that there is a relationship in the population between the predictor x and response y."

If the p-value is larger than the significance level alpha, we fail to reject the null hypothesis.

Conclude: "There is not enough evidence at the α level to conclude that there is a relationship in the population between the predictor x and response y."

Examples

Unless asked, most of your work will be ‘black box’. That is, Python will provide you with numbers that you will then need to interpret.

[1] James is interested in finding a relationship between diet and energy.
James collects data from 20 day-laborers.

Specifically, the data is hours spent working and calorie consumption.
After collecting data, James finds the line-of-best-fit.

The regression line, $\hat{y} = b_0 + b_1 x$ turns out to be $\hat{y} = 2.544 + 0.164x$.

Since this is a statistic from a sample that is used to estimate the population, James decides to create a confidence interval for the slope of the line.

As always, $CI = \text{statistic} \pm \text{score} \cdot SE$.

But what is the standard error for the slope of a regression line?

$$SE_{RegLine} = \frac{\sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}}{\sqrt{\sum(x_i - \bar{x})^2}}$$

$SE = 0.057$ (provided by stats package)

$R^2 = 60\%$

Assume that all conditions for inference have been met.

What is the 95% confidence interval for the slope of the least squares regression line? Recall $\hat{y} = 2.544 + 0.164x$

Answer:

$$0.164 \pm t \cdot 0.057$$

$$df = 20 - 2 = 18$$

$t = 2.101$ because `stats.t.ppf(.975,18)= 2.10092204024096`

$$0.164 \pm (2.101) \cdot 0.057 = [0.044, 0.284]$$

You want to know the best deal on soda.
So, you sample 24 different kinds/brands and rate them. You then look at the relationship between price and rating.

You generate the least squares regression line for the data:
$\hat{y} = -0.39 + 2.52x$ with $SE_{\beta} = 0.283$ (The return from stats package)

Assuming all conditions for inference have been met, what is a 90% CI for the slope of the least squares regression line?

Ans: ?

You want to know the best deal on soda.
So, you sample 24 different kinds/brands and rate them. You then look at the relationship between price and rating.

You generate the least squares regression line for the data:
$\hat{y} = -0.39 + 2.52x$ with $SE_\beta = 0.283$

Assuming all conditions for inference have been met, what is a 90% CI for the slope of the least squares regression line?

Ans: $t_{(.95,22)} = 1.717$

```
stats.t.ppf(.95,22) = 1.717144374380242
```

$$2.52 \pm 1.717(0.283) = [2.034, 3.006]$$

Barkley wants to study the relationship between age and probability of being in a bicycle accident.

He finds the number of accidents per 100 riders for ages 16 to 24 inclusive.

He finds the regression line to be: $\hat{y} = 39.11 - 1.133x$ with $SE_\beta = 0.27$.

If we assume that all conditions for inference have been met,
then what is a 95% CI for the slope of the least squares regression line?

Ans: ?

Barkley wants to study the relationship between age and probability of being in a bicycle accident.

He finds the number of accidents per 100 riders for ages 16 to 24 inclusive.

He finds the regression line to be: $\hat{y} = 39.11 - 1.133x$ with $SE_\beta = 0.27$.

If we assume that all conditions for inference have been met, then what is a 95% CI for the slope of the least squares regression line?

Ans: $t_{.975,7} = 2.365$

```
stats.t.ppf(.975,7) = 2.3646242510102993
```

$$-1.133 \pm 2.365(0.27) = [-1.772, -0.494]$$

Walt is looking for a used car, preferably a Pontiac Aztec.
He collected data on car mileage and value.

In a random sample of 11 Pontiac Aztecs he found the regression line to be:
$\hat{y} = 39.575 - 0.246x$ with $SE_{\beta} = 0.013.$

Assuming all conditions for inference have been met, then what is a 99% CI for the slope of the least squares regression line?

Ans: ?

Walt is looking for a used car, preferably a Pontiac Aztec.
He collected data on car mileage and value.

In a random sample of 11 Pontiac Aztecs he found the regression line to be:
$\hat{y} = 39.575 - 0.246x$ with $SE_\beta = 0.013$.

Assuming all conditions for inference have been met, then what is a 99% CI for the slope of the least squares regression line?

Ans: $t_{.005,9} = 3.25$

```
stats.t.ppf(.995,9) = 3.2498355440153697
```

$$-0.246 \pm 3.25(0.013) = [-0.288, -0.204]$$

Ernie is looking for a new roommate since he left S. street.
He collects data from college towns about population density and average rent for 1-bedroom apartments.

Ernie collects data from 25 different college towns in the US. Ernie discovers the regression line to be:

$$\hat{y} = 812 + 22.615 \text{ with } SE_{\beta} = 4.179$$

If we assume that all conditions for inference have been met,
then what is a 90% CI for the slope of the least squares regression line?

Ans: ?

Ernie is looking for a new roommate since he left S. street.
He collects data from college towns about population density and average rent for 1-bedroom apartments.

Ernie collects data from 25 different college towns in the US. Ernie discovers the regression line to be:

$$\hat{y} = 812 + 22.615 \quad \text{with} \quad SE_{\beta} = 4.179$$

If we assume that all conditions for inference have been met, then what is a 90% CI for the slope of the least squares regression line?

Ans: $t_{.95,23} = 1.714$

```
stats.t.ppf(.95,23) = 1.7138715277470473
```

$$22.615 \pm 1.714(4.179) = [15.452, 29.778]$$

Q: What is a residual?

A: Residual = actual – predicted

Residuals tell you how good your line is.

If your regression line is $\hat{y} = \frac{1}{3} + \frac{1}{3}x$ then what is the residual for an independent variable value of 155 and a dependent variable of 51?

Ans: ?

Q: What is a residual?

A: Residual = actual – predicted

Residuals tell you how good your line is.

If your regression line is $\hat{y} = \frac{1}{3} + \frac{1}{3}x$ then what is the residual for an independent variable value of 155 and a dependent variable of 51?

Ans: ?

residual = actual – predicted

$$51 - \left(\frac{1}{3} + \frac{1}{3} \cdot 155\right) = 51 \ - 52 = -1$$

As the CEO of BOA (Big Oil Airlines) you record the cost per barrel of crude oil and the cost per barrel of jet fuel each week.

After plotting the results, you decide a regression line is appropriate for predicting the cost per barrel of jet fuel from the cost per barrel of crude oil.

Your stats package provides you with $\hat{y} = 2 + \frac{7}{6}x$.

What is the residual of a week in which a barrel of oil costs $\$60$ and a barrel of jet fuel costs $\$78$?

Ans: ?

As the CEO of BOA (Big Oil Airlines) you record the cost per barrel of crude oil and the cost per barrel of jet fuel each week.

After plotting the results, you decide a regression line is appropriate for predicting the cost per barrel of jet fuel from the cost per barrel of crude oil.

Your stats package provides you with $\hat{y} = 2 + \frac{7}{6}x$.

What is the residual of a week in which a barrel of oil costs $\$60$ and a barrel of jet fuel costs $\$78$?

Ans: $\text{residual} = \text{actual} - \text{predicted}$

$$\text{residual} = 78 - 72 = 6$$

Next Time: Multiple Linear Regression