Inference for Regression

IPS Chapter 10

- 10.1: Simple Linear Regression
- 10.2: More Detail about Simple Linear Regression

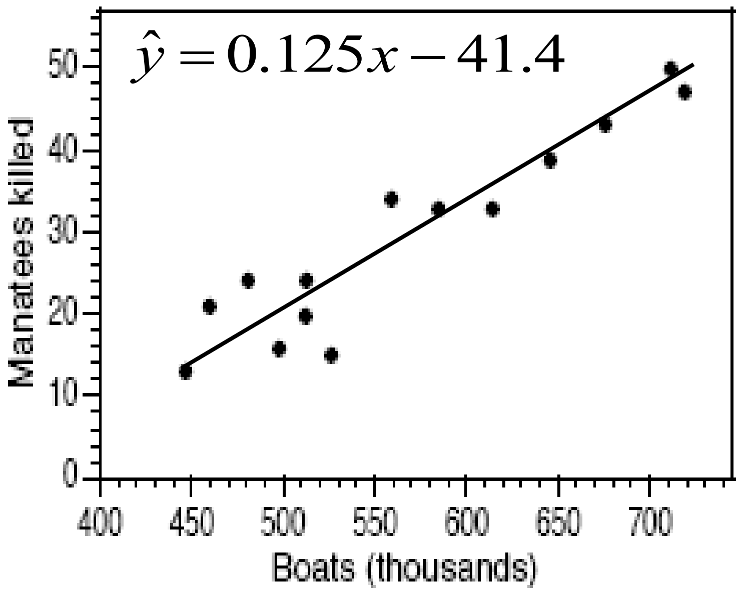
Inference for Regression

10.1 Simple Linear Regression

Objectives

10.1 Simple linear regression

- Statistical model for linear regression
- Estimating the regression parameters
- Confidence interval for regression parameters
- Significance test for the slope
- Confidence interval for μ_y
- Prediction intervals



The data in a scatterplot are a random **sample** from a population that may exhibit a linear relationship between *x* and *y*. Different sample ➔ different plot.

Now we want to describe the **population mean response** μ_y as a function of the explanatory variable *x*: $\mu_y = \beta_0 + \beta_1 x$.

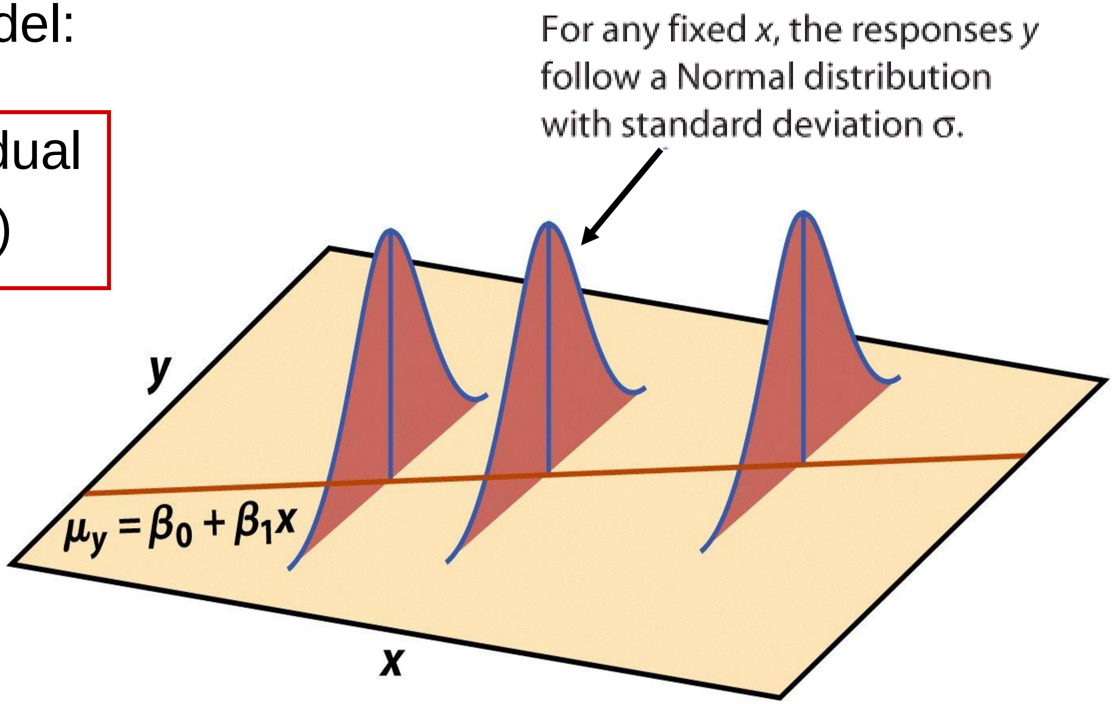
And to assess whether the observed **relationship** is **statistically significant** (not entirely explained by chance events due to random sampling).

Statistical model for linear regression

In the population, the linear regression equation is $\mu_y = \beta_0 + \beta_1 x$.

Sample data then fits the model:

Data =	fit	+	residual	
y_i =	$(\beta_0 + \beta_1 x_i)$	+	(ε_i)	



where the ε_i are

independent and

Normally distributed $N(0,\sigma)$.

Linear regression assumes **equal variance of *y***

(σ is the same for all values of x).

Estimating the parameters

$$\mu_y = \beta_0 + \beta_1 x$$

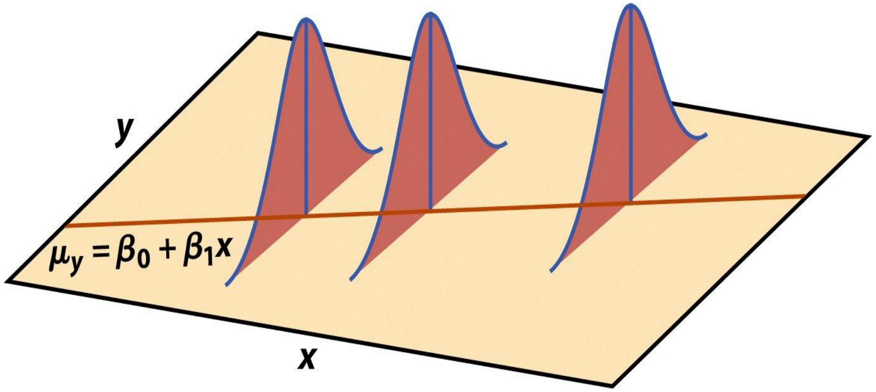
The intercept β_0**,** the slope β_1, and the standard deviation σ of *y* are the unknown parameters of the regression model. We rely on the random sample data to provide unbiased estimates of these parameters.

- The value of $\hat{y}$ from the least-squares regression line is really a prediction of the mean value of *y* (μ_y) for a given value of *x*.
- The least-squares regression line ($\hat{y} = b_0 + b_1 x$) obtained from sample data is the best estimate of the true population regression line ($\mu_y = \beta_0 + \beta_1 x$).

$\hat{y}$ unbiased estimate for mean response μ_y

b_0 unbiased estimate for intercept β_0

b_1 unbiased estimate for slope β_1



The **population standard deviation** σ for *y* at any given value of *x* represents the spread of the normal distribution of the ε_i around the mean μ_y .

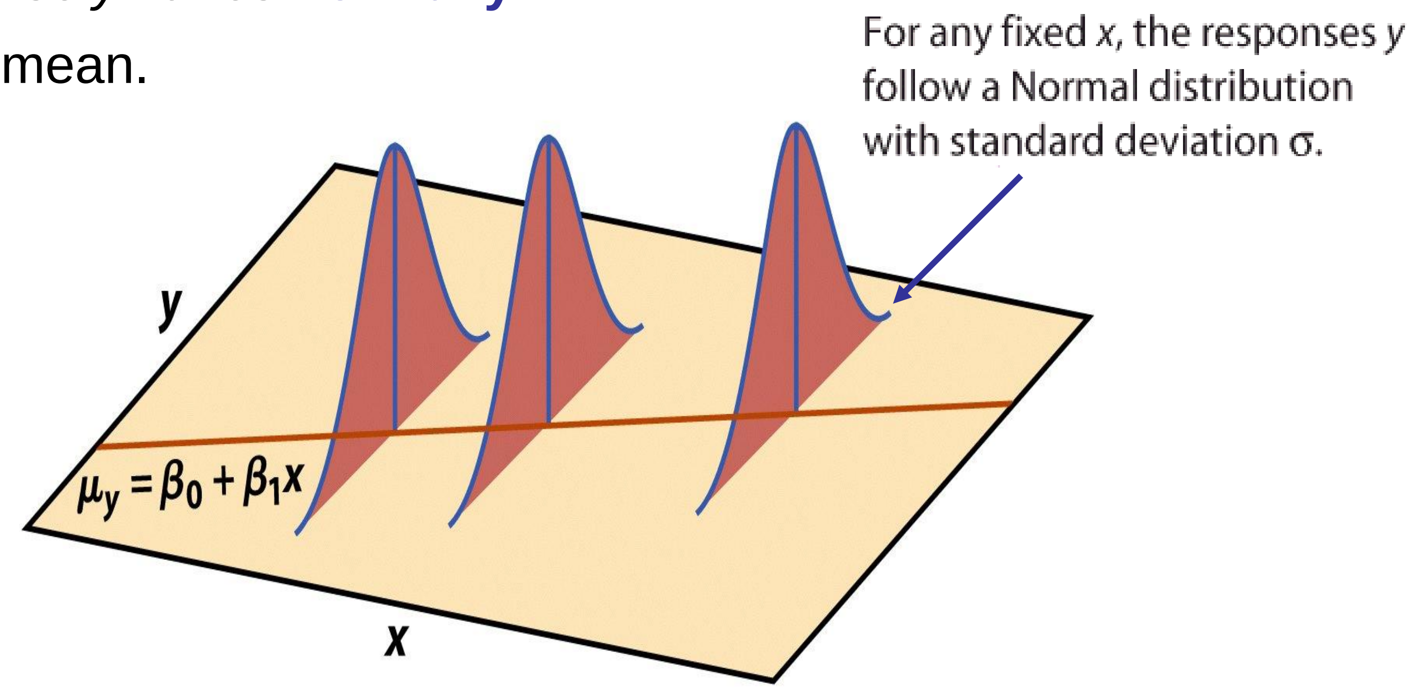
The **regression standard error, *s*,** for *n* sample data points is calculated from the residuals ($y_i - \hat{y}_i$):

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

s is an unbiased estimate of the regression standard deviation σ.

Conditions for inference

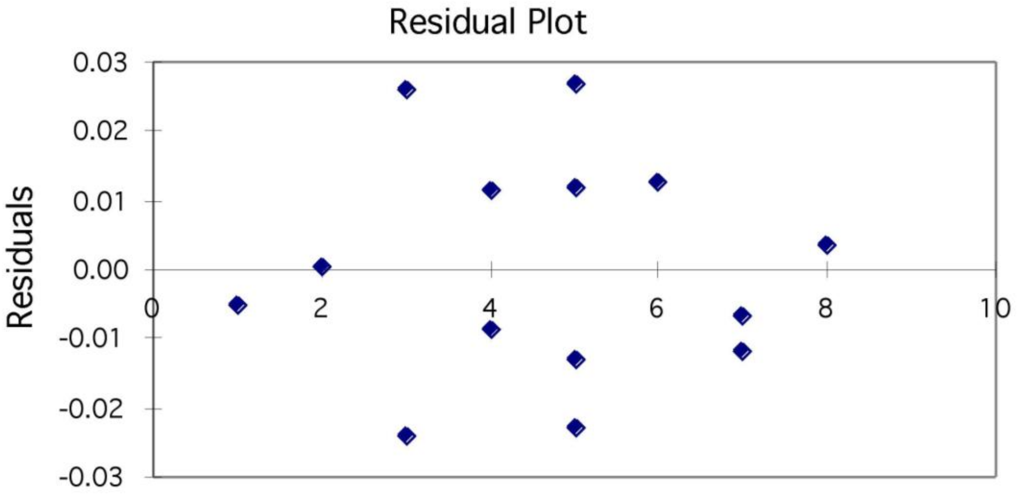
- The observations are **independent.**
- The relationship is indeed **linear.**
- The standard deviation of *y*, ***σ***, is the same for all values of *x*.
- The response *y* varies **normally** around its mean.



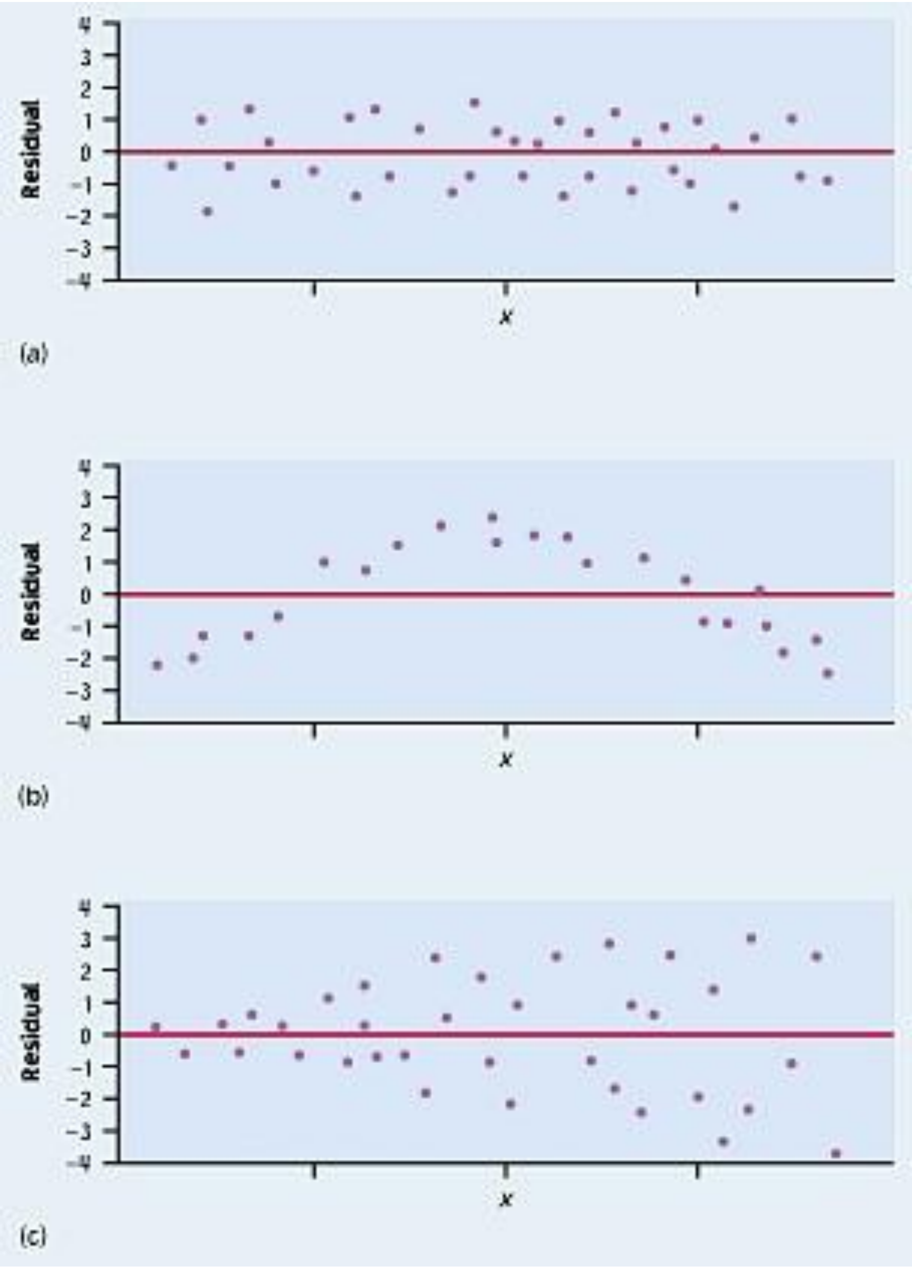
Using residual plots to check for regression validity

The residuals ($y - \hat{y}$) give useful information about the contribution of individual data points to the overall pattern of scatter.

We view the residuals in a **residual plot:**



If residuals are scattered randomly around 0 with uniform variation, it indicates that the data fit a linear model, have normally distributed residuals for each value of *x*, and constant standard deviation σ.



Residuals are randomly scattered

→ **good!**

Curved pattern

→ the relationship is **not linear.**

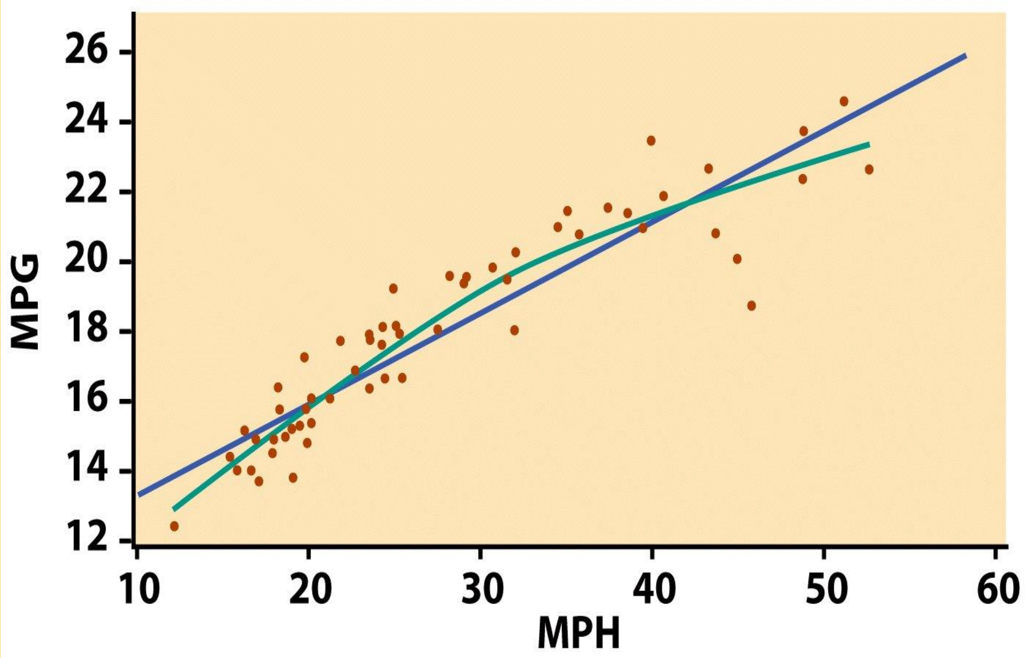
Change in variability across plot

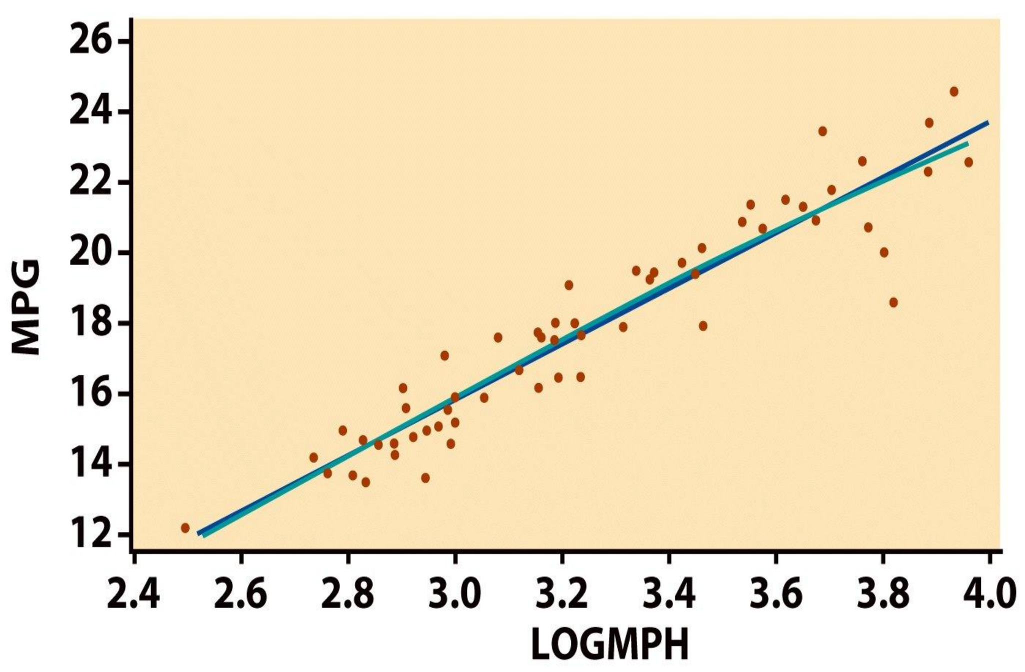
→***σ* not equal** for all values of *x*.

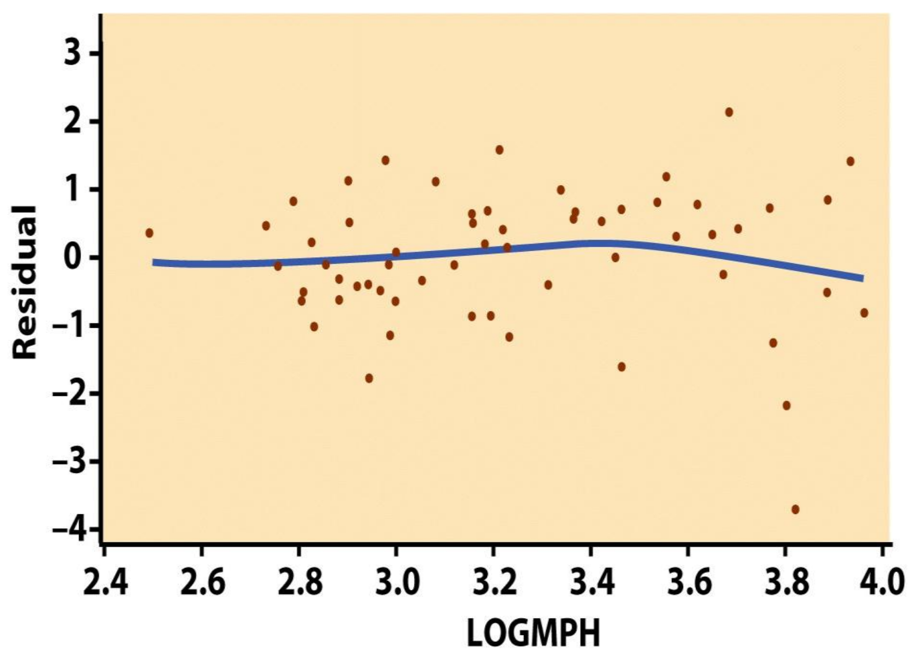
What is the relationship between the average speed a car is driven and its fuel efficiency?

We plot fuel efficiency (in miles per gallon, MPG) against average speed (in miles per hour, MPH) for a random sample of 60 cars. The relationship is curved.

When speed is log transformed (log of miles per hour, LOGMPH) the new scatterplot shows a positive, **linear** relationship.







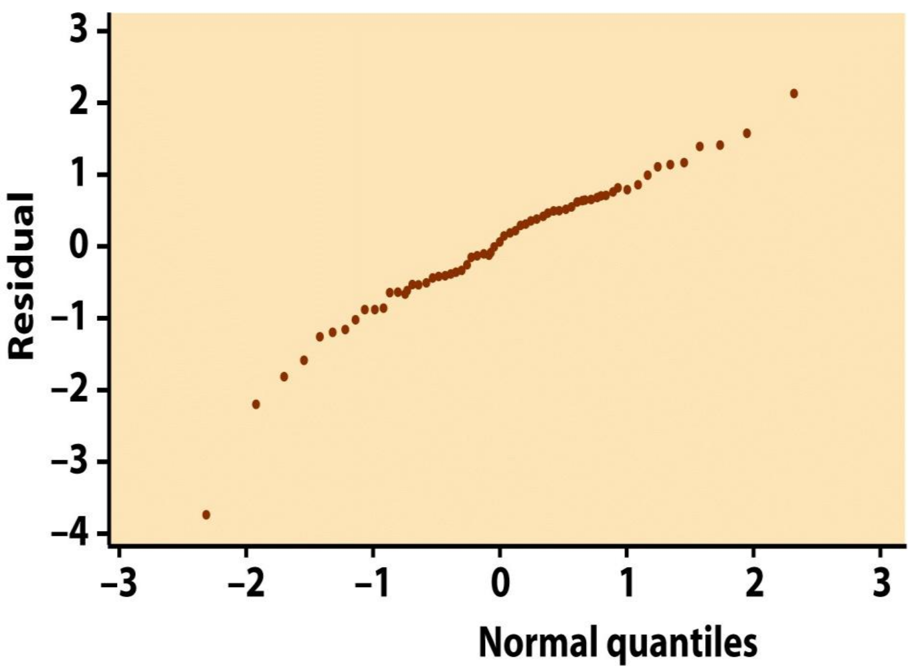
Residual plot:

The spread of the residuals is reasonably random—no clear pattern. The relationship is <u>indeed linear</u>.

But we see one low residual (3.8, −4) and one potentially influential point (2.5, 0.5).

Normal quantile plot for residuals:

The plot is fairly straight, supporting the assumption of normally distributed residuals.



➔ Data okay for inference.

Confidence interval for regression parameters

Estimating the regression parameters β_0, β_1 is a case of one-sample inference with unknown population variance.

➔ We rely on the *t* distribution, with ***n* – 2 degrees of freedom**.

A level *C* **confidence interval for the slope, β_1,** is proportional to the standarderror of the least-squares slope:

$$b_1 \pm t^* \, SE_{b1}$$

A level C **confidence interval for the intercept, β_0 ,** is proportional to the standard error of the least-squares intercept:

$$b_0 \pm t^* \, SE_{b0}$$

t is the t critical value for the t (n – 2) distribution with area C between –t* and +t*.*

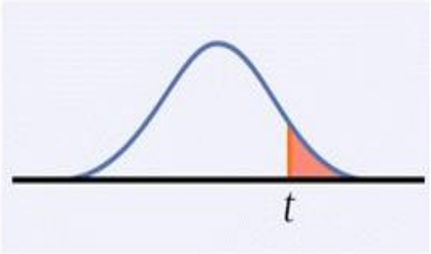
Significance test for the slope

We can test the hypothesis H_0: $\beta_1 = 0$ versus a 1 or 2 sided alternative.

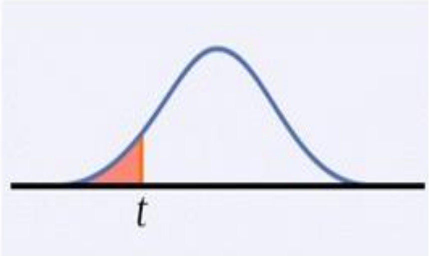
We calculate $t = b_1 / SE_{b1}$

which has the ***t* (*n* – 2) distribution** to find the p-value of the test.

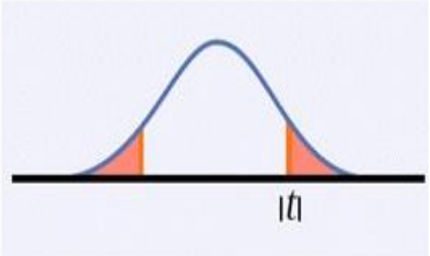
H_a: $\beta_1 > 0$ is $P(T \geq t)$



H_a: $\beta_1 < 0$ is $P(T \leq t)$



H_a: $\beta_1 \neq 0$ is $2P(T \geq |t|)$



<u>Note</u>: Software typically provides two-sided p-values.

Testing the hypothesis of no relationship

We may look for evidence of a **significant relationship** between variables *x* and *y* in the population from which our data were drawn.

For that, we can test the hypothesis that the regression slope parameter β is equal to zero.

$$H_0: \beta_1 = 0 \text{ vs. } H_0: \beta_1 \neq 0$$

$$\text{slope } b_1 = r \frac{s_y}{s_x}$$

Testing H_0: $\beta_1 = 0$ also allows to test the **hypothesis of no correlation** between *x* and *y* in the population.

Note: A test of hypothesis for β_0 is irrelevant (β_0 is often not even achievable).

Using technology

Computer software runs all the computations for regression analysis.

Here is some software output for the car speed/gas efficiency example.

SPSS

Model Summary

Model	R	R Square	Std. Error of the Estimate
1	.946	.895	.9995

a Predictors: (Constant), LOGMPH

		Coefficients		t	Sig.	95% Confidence Interval for B	
Model		B	Std. Error			Lower Bound	Upper Bound
1	(Constant)	−7.796	1.155	−6.750	.000	−10.108	−5.484
	LOGMPH	7.874	.354	22.237	.000	7.165	8.583

a Dependent Variable: MPG

Slope

Intercept

p-value for tests of significance

Confidence intervals

The *t*-test for regression slope is highly significant ($p < 0.001$). There is a significant relationship between average car speed and gas efficiency.

Excel

Regression Statistics						
Multiple R	0.946053015					
R Square	0.895016308					
Adjusted R Square	0.893206244					
Standard Error	0.999516364					
Observations	60					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	493.9885883	493.9886	494.4668	4.50949E-30	
Residual	58	57.94391174	0.999033			
Total	59	551.9325				
	Coefficients	Standard Error	tStet	P-value	Lower 95%	Upper 95%
intercept	−7.796250129	1.154944262	−6.75033	7.69E-09	−10.10812052	−5.48437974
logmph	7.874219013	0.354110611	22.23661	4.51E-30	7.165390143	8.583047883

"intercept": intercept
"logmph": slope

P-value for tests of significance

confidence intervals

SAS

Root MSE	0.99952	R-Square	0.8950
Dependent Mean	17.72500	Adj R-Sq	0.8932
Coeff Var	5.63902		

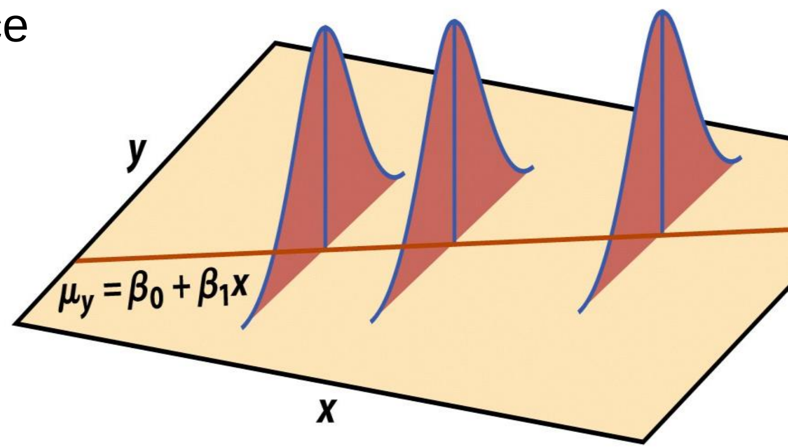
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	−7.79625	1.15494	−6.75	<.0001	−10.10812	−5.48438
logmph	1	7.87422	0.35411	22.24	<.0001	7.16539	8.58305

Confidence interval for μ_y

Using inference, we can also calculate a **confidence interval for the population mean μ_y** of all responses *y* when *x* takes the value *x** (within the range of data tested):

This interval is centered on $\hat{y}$, the unbiased estimate of μ_y.

The true value of the population mean μ_y at a given value of *x*, will indeed be within our confidence interval in *C*% of all intervals calculated from many different random samples.



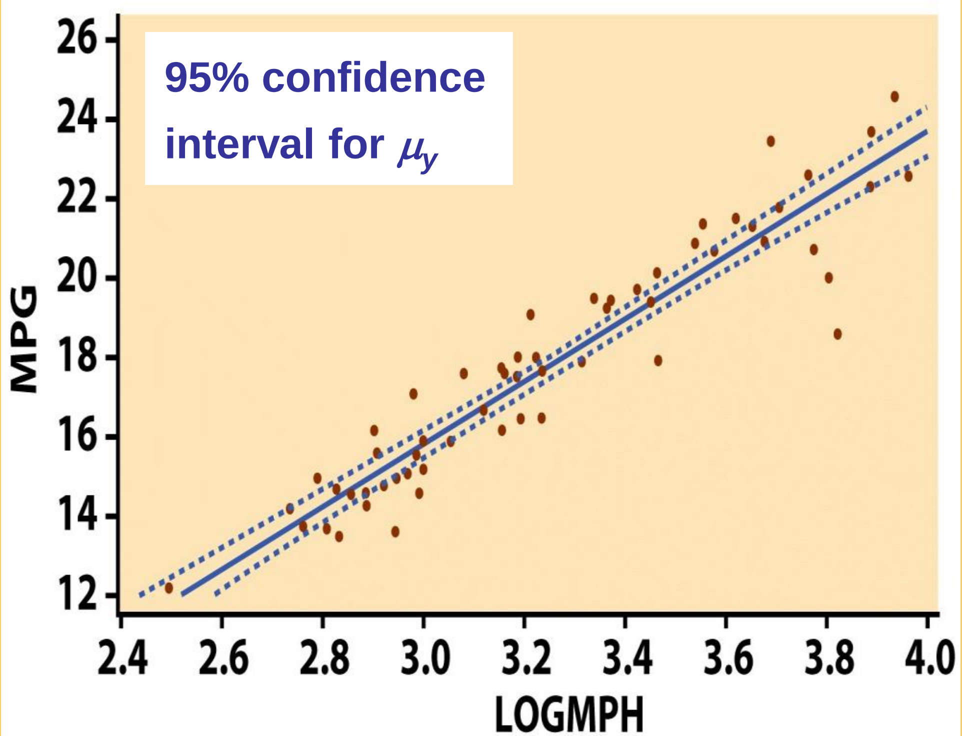
The **level *C* confidence interval for the mean response μ_y** at a given value x^* of x is:

$$\hat{\mu}_y \pm t_{n-2}^* \, SE_{\hat{\mu}}$$

*t^* is the t critical value for the **t (n – 2)** distribution with area C between –t^* and +t^*.*

A separate confidence interval is calculated for μ_y along all the values that x takes.

Graphically, the series of confidence intervals is shown as a continuous interval on either side of $\hat{y}$.



Inference for prediction

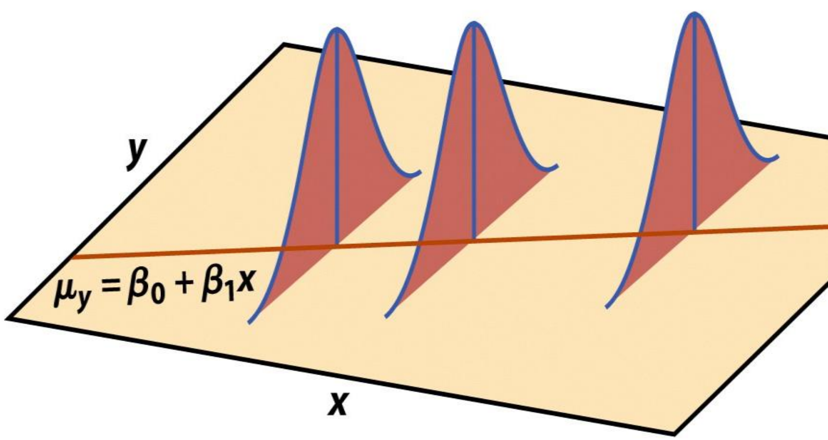
One use of regression is for **predicting** the value of y, $\hat{y}$, for any value of x within the range of data tested: $\hat{y} = b_0 + b_1 x$.

But the regression equation depends on the particular sample drawn.

More reliable predictions require statistical inference:

To estimate an *individual* response y for a given value of x, we use a **prediction interval.**

If we randomly sampled many times, there would be many different values of y obtained for a particular x following $N(0, \sigma)$ around the mean response μ_y.



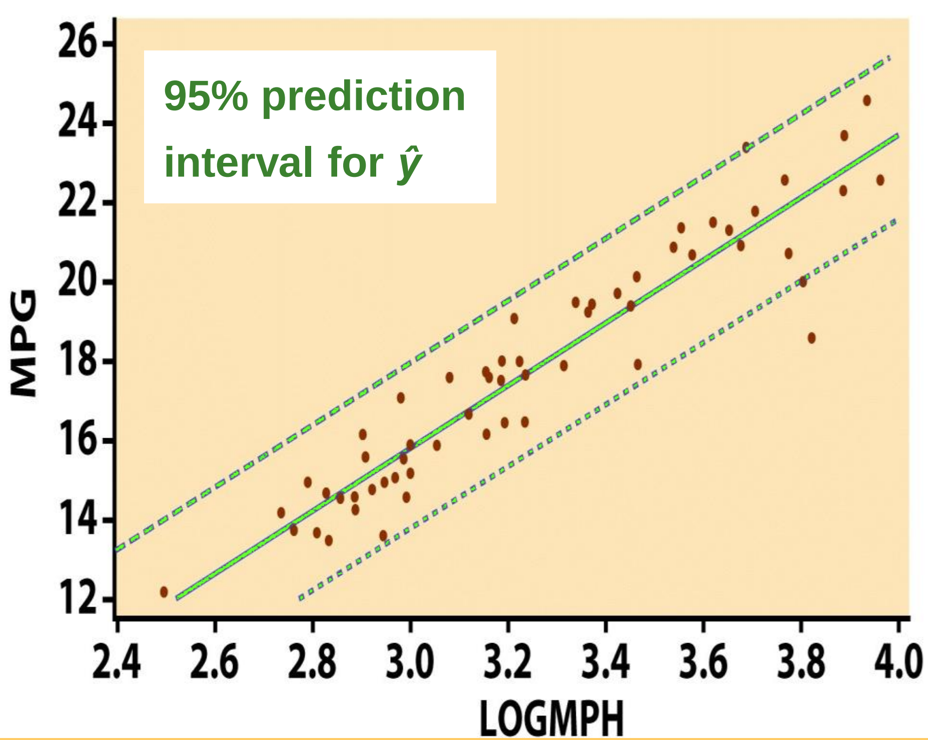
The **level C prediction interval for a single observation** on *y* when *x* takes the value *x** is:

$$\hat{y} \pm t^*_{n-2} SE_{\hat{y}}$$

*t^* is the t critical value for the **t (n – 2)** distribution with area C between $-t^*$ and $+t^*$.*

The prediction interval represents mainly the error from the normal distribution of the residuals ε_i.

Graphically, the series confidence intervals are shown as a continuous interval on either side of $\hat{y}$.



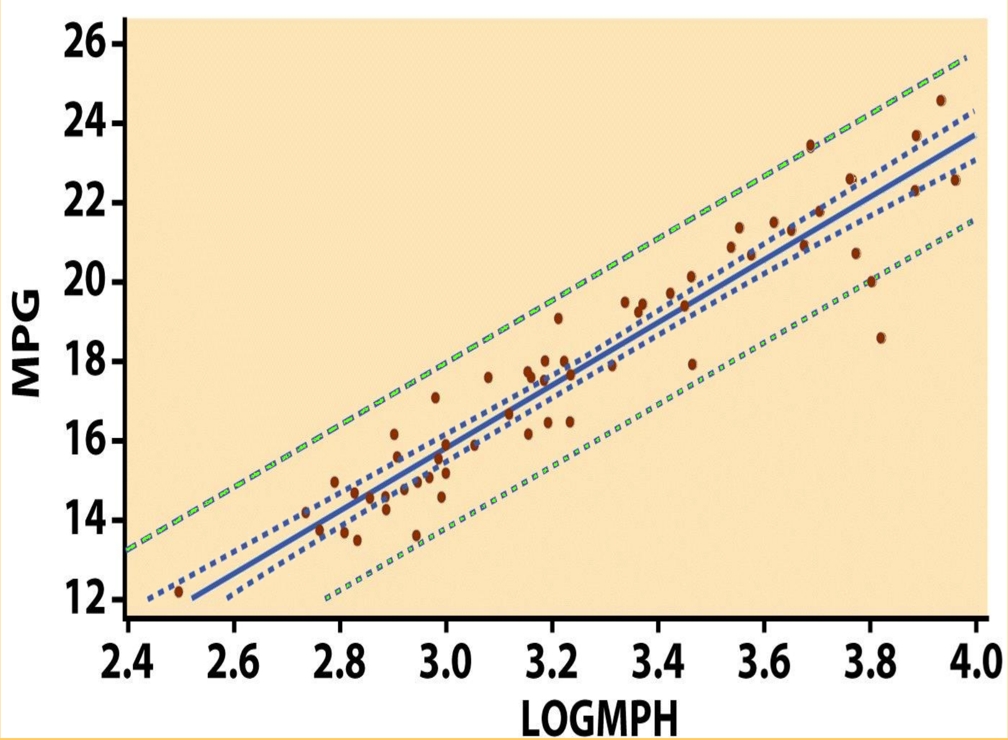
- The **confidence interval for μ_y** contains with *C*% confidence the population mean μ_y of all responses at a particular value of *x.*

- The **prediction interval** contains *C*% of all the individual values taken by *y* at a particular value of *x*.

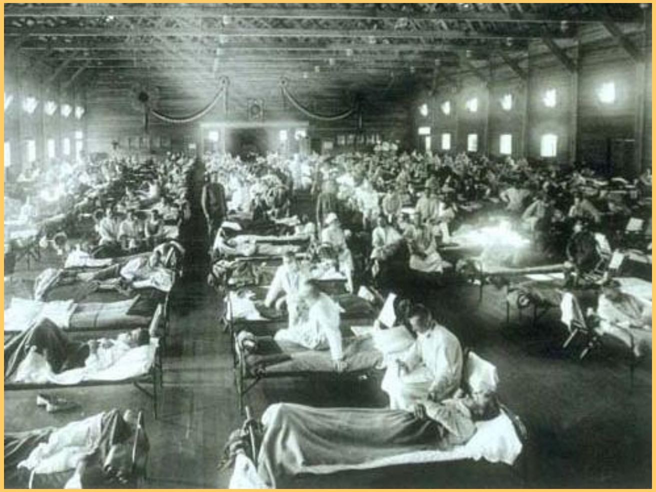
95% prediction interval for $\hat{y}$

95% confidence interval for μ_y

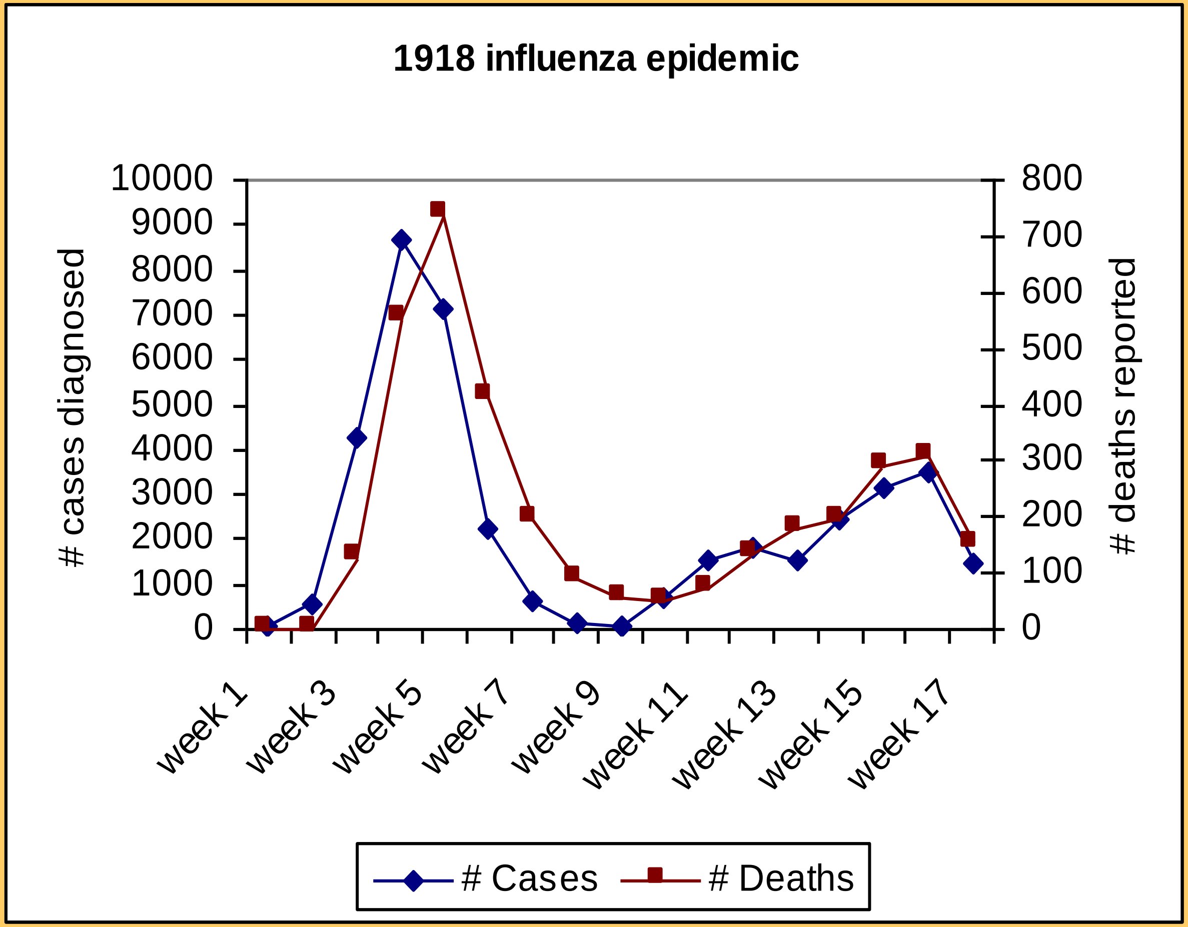
Estimating μ_y uses a smaller confidence interval than estimating an individual in the population (sampling distribution narrower than population distribution).



1918 flu epidemics



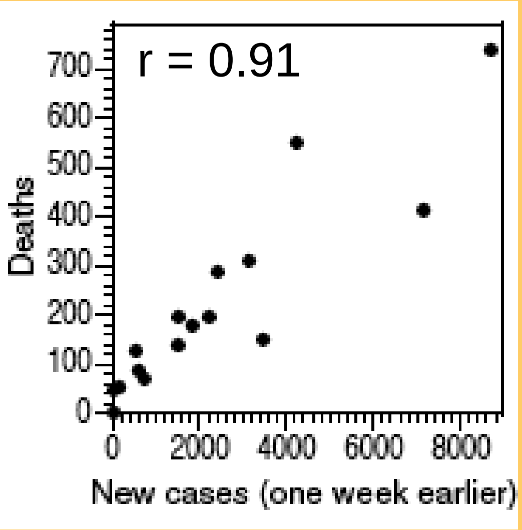
1918 influenza epidemic		
Date	**# Cases**	**# Deaths**
week 1	36	0
week 2	531	0
week 3	4233	130
week 4	8682	552
week 5	7164	738
week 6	2229	414
week 7	600	198
week 8	164	90
week 9	57	56
week 10	722	50
week 11	1517	71
week 12	1828	137
week 13	1539	178
week 14	2416	194
week 15	3148	290
week 16	3465	310
week 17	1440	149



The line graph suggests that 7% to 9% of those diagnosed with the flu died within about a week of diagnosis.

We look at the relationship between the number of deaths in a given week and the number of new diagnosed cases one week earlier.

1918 flu epidemic: Relationship between the number of deaths in a given week and the number of new diagnosed cases one week earlier.



EXCEL

Regression Statistics

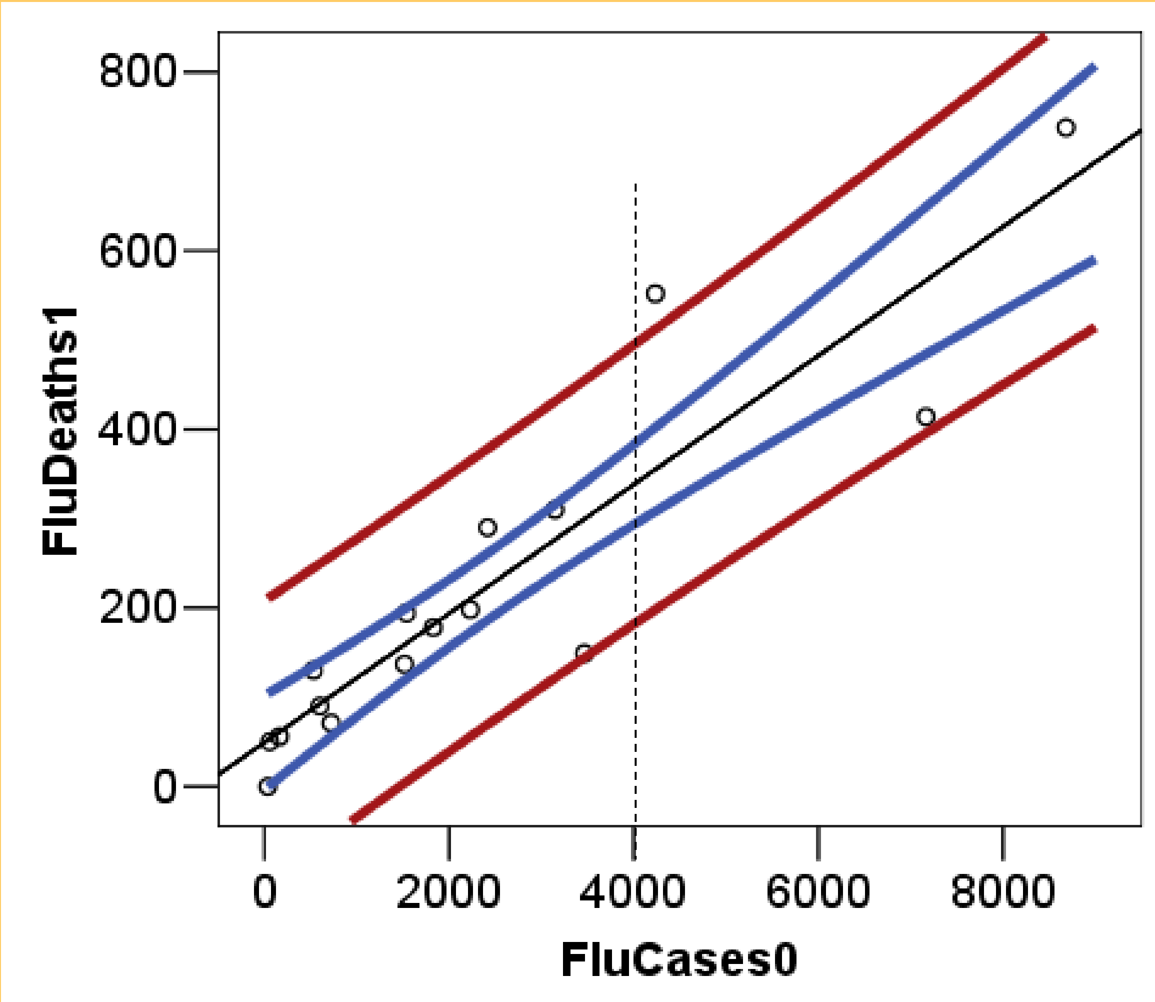
Multiple R	0.911
R Square	0.830
Adjusted R Square	0.82
Standard Error	85.07 s
Observations	16.00

	Coefficients	*St. Error*	*t Stat*	***P-value***	*Lower 95%*	*Upper 95%*
Intercept	49.292	29.845	1.652	0.1209	(14.720)	113.304
FluCases0	0.072	0.009	8.263	0.0000	0.053	0.091
	$\mathbf{b_1}$	SE_{b1}		**P-value for H_0: $\beta_1 = 0$**		

P-value very small ➔ reject H_0 ➔ β_1 significantly different from 0

There is a **significant relationship** between the number of flu cases and the number of deaths from flu a week later.

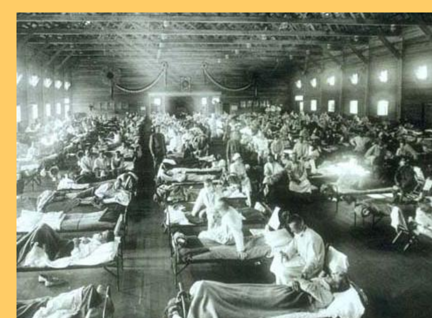
SPSS



Least squares regression line
95% prediction interval for *y*
95% confidence interval for μ_y

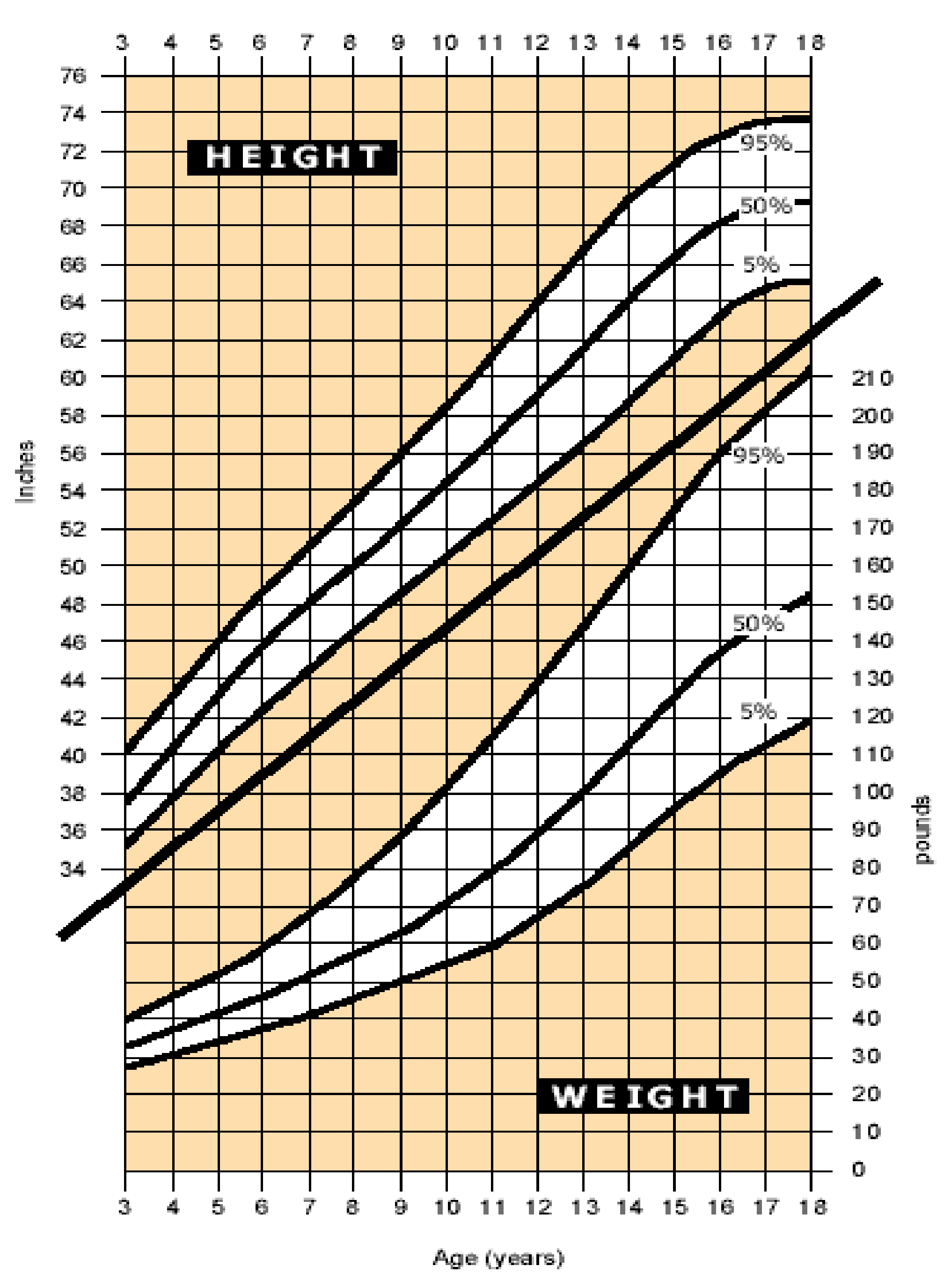
CI for mean weekly death count one week after 4,000 flu cases are diagnosed: μ_y within about 300–380.

Prediction interval for a weekly death count one week after 4,000 flu cases are diagnosed: $\hat{y}$ within about 180–500 deaths.



What is this?

A 90% prediction interval for the height (above) and a 90% prediction interval for the weight (below) of male children, ages 3 to 18.



Inference for Regression

10.2 More Detail about Simple Linear Regression

Objectives

10.2 More detail about simple linear regression

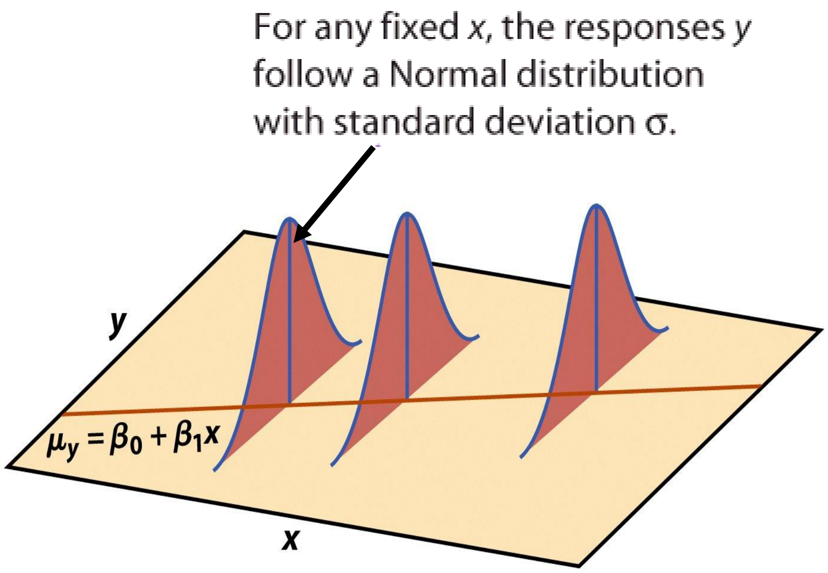
- Analysis of variance for regression
- The ANOVA *F* test
- Calculations for regression inference
- Inference for correlation

Analysis of variance for regression

The regression model is:

Data = fit + residual

$y_i = (\beta_0 + \beta_1 x_i) + (\varepsilon_i)$



where the ε_i are **independent** and **normally** distributed $N(0,\sigma)$, and σ is the same for all values of x.

It resembles an ANOVA, which also assumes equal variance, where

SST = SS model + SS error and

DFT = DF model + DF error

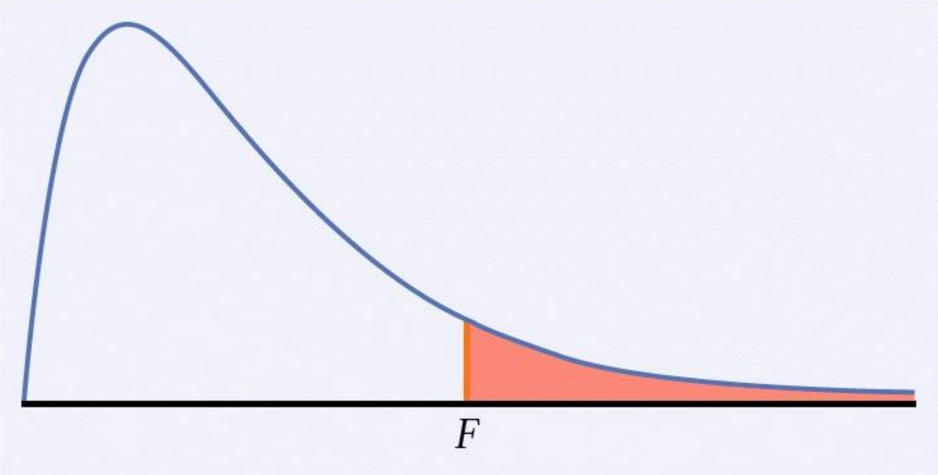
The ANOVA F test

For a simple linear relationship, the ANOVA tests the hypotheses

$$H_0\text{: } \beta_1 = 0 \text{ versus } H_a\text{: } \beta_1 \neq 0$$

by comparing MSM (model) to MSE (error): F = MSM/MSE

When H_0 is true, F follows the $F(1, n - 2)$ distribution.
The p-value is $P(F \geq f)$.



The ANOVA test and the two-sided t-test for H_0: $\beta_1 = 0$ yield the same p-value.

Software output for regression may provide t, F, or both, along with the p-value.

ANOVA table

Source	Sum of squares SS	DF	Mean square MS	F	P-value
Model	$\sum(\hat{y}_i - \bar{y})^2$	1	SSM/DFM	MSM/MSE	Tail area above F
Error	$\sum(y_i - \hat{y}_i)^2$	$n - 2$	SSE/DFE		
Total	$\sum(y_i - \bar{y})^2$	$n - 1$			

SST = SSM + SSE

DFT = DFM + DFE

The **standard deviation of the sampling distribution, *s*,** for *n* sample data points is calculated from the residuals $e_i = y_i - \hat{y}_i$

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{DFE} = MSE$$

s is an unbiased estimate of the regression standard deviation ***σ*****.**

Coefficient of determination, r^2

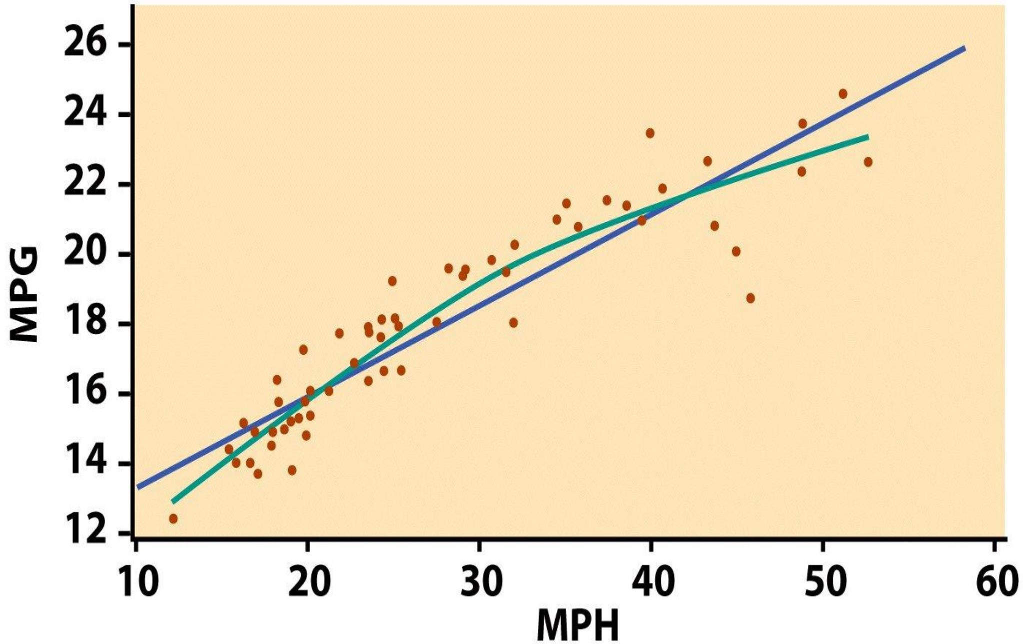
The coefficient of determination, *r*², square of the correlation coefficient, **is the percentage of the variance in *y*** (vertical scatter from the regression line) **that can be explained by changes in *x*.**

$$r^2 = \frac{\text{variation in } y \text{ caused by } x \quad \text{(i.e., the regression line)}}{\text{total variation in observed } y \text{ values around the mean}}$$

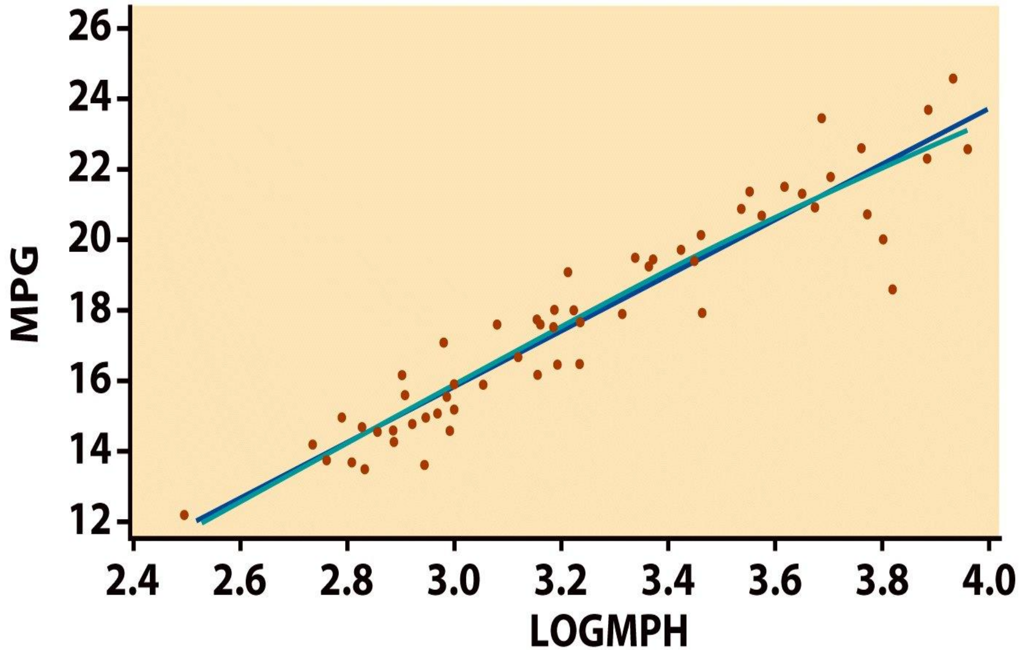
$$r^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\text{SSM}}{\text{SST}}$$

What is the relationship between the average speed a car is driven and its fuel efficiency?

We plot fuel efficiency (in miles per gallon, MPG) against average speed (in miles per hour, MPH) for a random sample of 60 cars. The relationship is curved.



When speed is log transformed (log of miles per hour, LOGMPH) the new scatterplot shows a positive, **linear** relationship.



Using software: SPSS

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	493.989	1	493.989	494.467	.000[a]
	Residual	57.944	58	.999		
	Total	551.932	59			

a. Predictors: (Constant), LOGMPH

b. Dependent Variable: MPG

r^2 =SSM/SST = 494/552

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.946[a]	.895	.893	.9995

a. Predictors: (Constant), LOGMPH

ANOVA and *t*-test give same p-value.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
	(Constant)	-7.796	1.155		-6.750	.000
	LOGMPH	7.874	.354	.946	22.237	.000

Dependent Variable: MPG

Calculations for regression inference

To estimate the parameters of the regression, we calculate the standard errors for the estimated regression coefficients.

The standard error of the least-squares slope β_1 is:

$$SE_{b1} = \frac{s}{\sqrt{\sum (x_i - \bar{x}_i)^2}}$$

The standard error of the intercept β_0 is:

$$SE_{b0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x}_i)^2}}$$

To estimate or predict future responses, we calculate the following standard errors

The standard error of the mean response μ_y is:

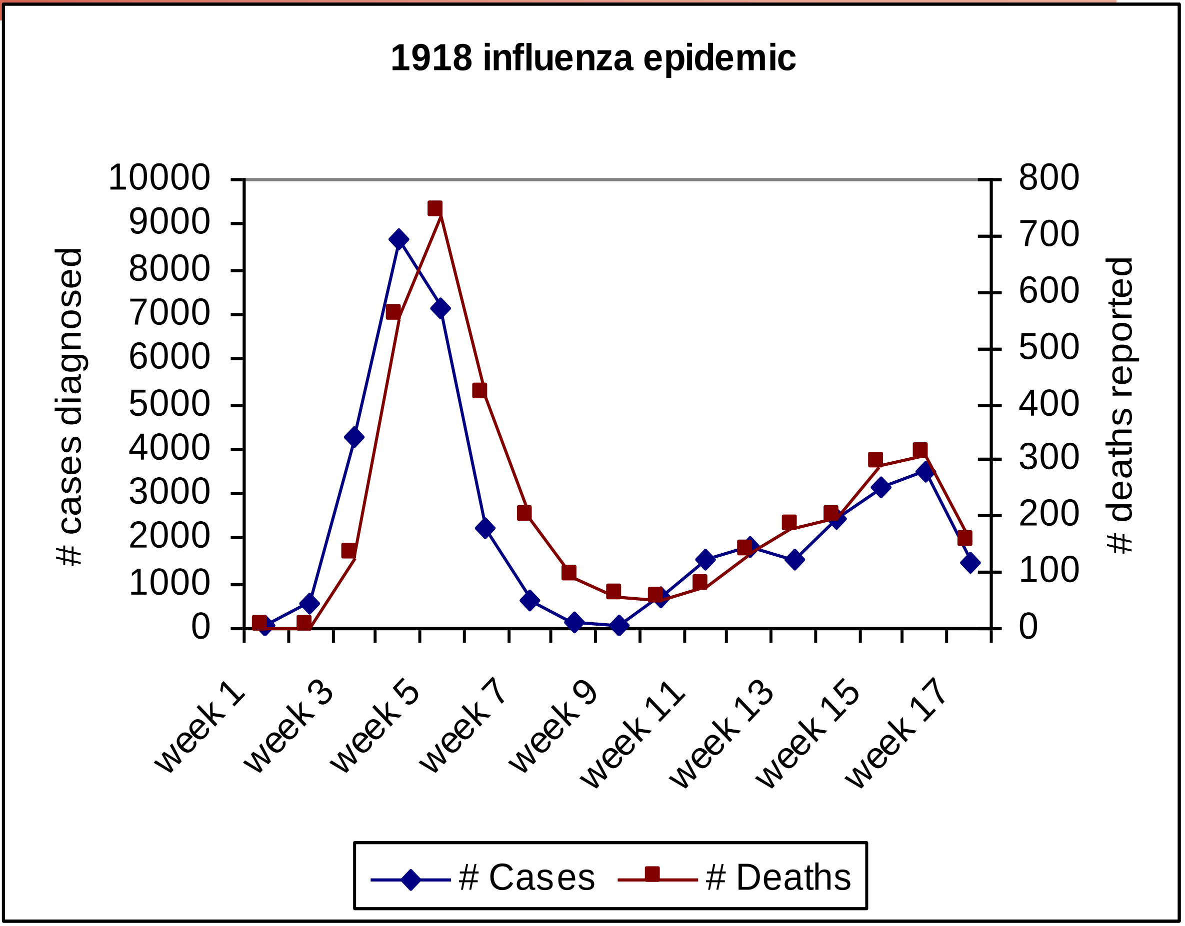
$$\mathrm{SE}_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum(x - \overline{x})^2}}$$

The standard error for predicting an individual response $\hat{y}$ is:

$$\mathrm{SE}_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum(x - \overline{x})^2}}$$

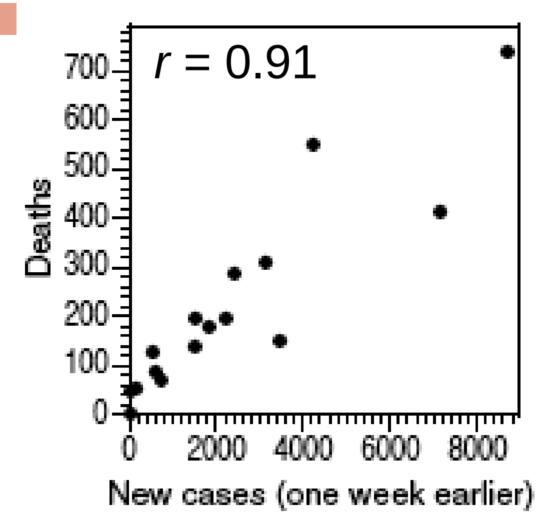
1918 flu epidemics

1918 influenza epidemic		
Date	**# Cases**	**# Deaths**
week 1	36	0
week 2	531	0
week 3	4233	130
week 4	8682	552
week 5	7164	738
week 6	2229	414
week 7	600	198
week 8	164	90
week 9	57	56
week 10	722	50
week 11	1517	71
week 12	1828	137
week 13	1539	178
week 14	2416	194
week 15	3148	290
week 16	3465	310
week 17	1440	149



The line graph suggests that about 7% to 8% of those diagnosed with the flu died within about a week of diagnosis. We look at the relationship between the number of deaths in a given week and the number of new diagnosed cases one week earlier.

1918 flu epidemic: Relationship between the number of deaths in a given week and the number of new diagnosed cases one week earlier.



MINITAB - Regression Analysis:

FluDeaths1 versus FluCases0

The regression equation is

FluDeaths1 = 49.3 + 0.0722 FluCases0

Predictor	Coef	SE Coef	T	P
Constant	49.29	29.85	1.65	0.121
FluCases	0.072222	0.008741	8.26	0.000

S = 85.07 $s = \sqrt{MSE}$

R-Sq = 83.0% r^2 = SSM / SST

R-Sq(adj) = 81.8%

P-value for H_0: $\beta_1 = 0$; H_a: $\beta_1 \neq 0$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	494041 SSM	494041	68.27	0.000
Residual Error	14	101308	7236 $MSE = s^2$		
Total	15	595349 SST			

Inference for correlation

To test for the null hypothesis of no linear association, we have the choice of also using the **correlation parameter *ρ***.

- When x is clearly the explanatory variable, this test is equivalent to testing the hypothesis H_0: $\beta = 0$.

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

- When there is no clear explanatory variable (e.g., arm length vs. leg length), a regression of x on y is not any more legitimate than one of y on x. In that case, the correlation test of significance should be used.

- When both x and y are normally distributed H_0: $\rho = 0$ tests for no association of any kind between x and y—not just linear associations.

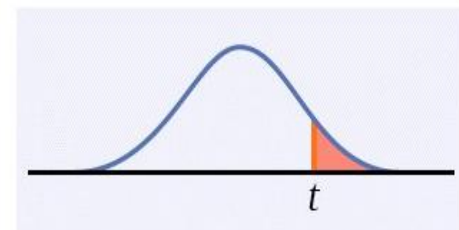
The test of significance for ρ uses the one-sample t-test for: H_0: $\rho = 0$.

We compute the t statistics for sample size n and correlation coefficient r.

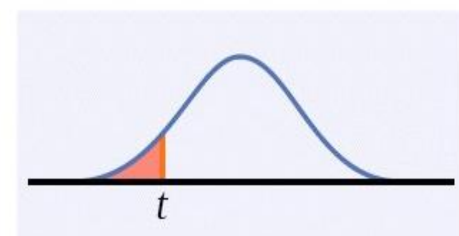
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

The p-value is the area under t (n – 2) for values of T as extreme as t or more in the direction of H_a:

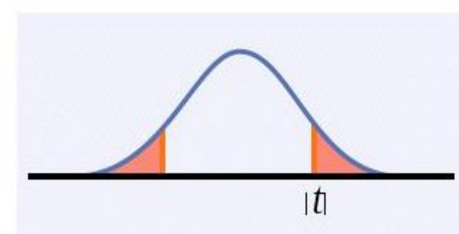
H_a: $\rho > 0$ is $P(T \geq t)$



H_a: $\rho < 0$ is $P(T \leq t)$



H_a: $\rho \neq 0$ is $2P(T \geq |t|)$



Relationship between average car speed and fuel efficiency

Correlations

		LOGMPH	MPG	
LOGMPH	Pearson Correlation	1	.946**	*r*
	Sig. (2-tailed)	.	.000	p-value
	N	60	60	*n*
MPG	Pearson Correlation	.946**	1	
	Sig. (2-tailed)	.000	.	
	N	60	60	

**. Correlation is significant at the 0.01 level (2-tailed).

There is a significant correlation (*r* is not 0) between fuel efficiency (MPG) and the logarithm of average speed (LOGMPH).

Cautions for Regression Inference

1. The observations must be independent.

   Repeated observations on the same cases or individuals.

2. The true relationship must be linear.

   Always plot your data. Look at the scatterplot to check that the overall pattern is roughly linear and that there are no outliers or influential points.

3. The standard deviation of the response about the true line is the same everywhere.

   Look at the scatterplot again. The scatter of the points about the line should be roughly the same over the entire range of the data. This is easier to check on a residual plot.

4. The response varies Normally about the true regression line.

   Make a histogram or stemplot of the residuals and check for skewness or other major departures from Normality.

Alternate Slides

The following slides offer alternate software output data and examples for this presentation.

Using technology

Computer software runs all the computations for regression analysis.

Here is some software output for the car speed/gas efficiency example.

Linear Fit

JMP

MPG = −7.79632 + 7.8742447 LOGMPH

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	−7.79632	1.154912	−6.75	<.0001*
LOGMPH	7.8742447	0.354101	22.24	<.0001*

Slope

Intercept

Standard error

p-value for tests of significance

The *t*-test for regression slope is highly significant ($p < 0.0001$). There is a significant relationship between average car speed and gas efficiency.

$MPG = -7.79632 + 7.8742447\ LOGMPH$

JMP

Summary of Fit

RSquare	0.895022
RSquare Adj	0.893212
Root Mean Square Error	0.999489
Mean of Response	17.725
Observations (or Sum Wgts)	60

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	493.99177	493.992	494.4971
Error	58	57.94073	0.999	**Prob > F**
C. Total	59	551.93250		<.0001*

Parameter Estimates

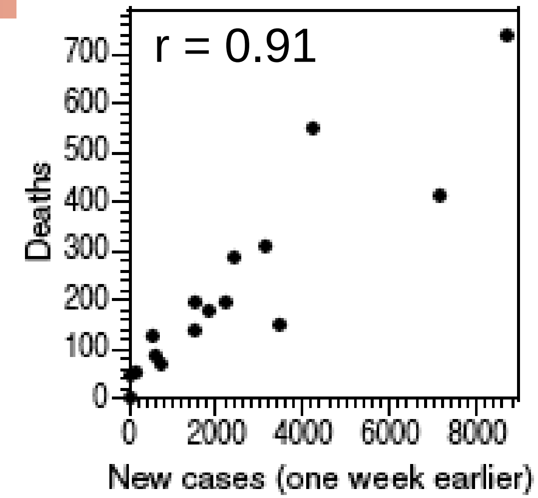
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	−7.79632	1.154912	−6.75	<.0001*
LOGMPH	7.8742447	0.354101	22.24	<.0001*

"intercept": intercept

"logmph": slope

P-value for tests of significance

1918 flu epidemic: Relationship between the number of deaths in a given week and the number of new diagnosed cases one week earlier. ***JMP***



Summary of Fit

RSquare	0.830	
RSquare Adj	0.818	
Root Mean Square Error	85.066	s_e
Mean of Response	222.313	
Observations (or Sum Wgts)	16.000	

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	49.292	29.8454	1.652	0.1209
New cases (-1)	0.072	0.0087	8.263	<.0001*

b_1

P-value for H_0: $\beta_1 = 0$

P-value very small ➔ reject H_0 ➔ β_1 significantly different from 0

There is a **significant relationship** between the number of flu cases and the number of deaths from flu a week later.

Using software: *JMP 6 SE*

Summary of Fit

RSquare	0.895
RSquare Adj	0.893
Root Mean Square Error	0.999
Mean of Response	17.725
Observations (or Sum Wgts)	60.000

r^2 =SSM/SST
= 494/552

Analysis of Variance

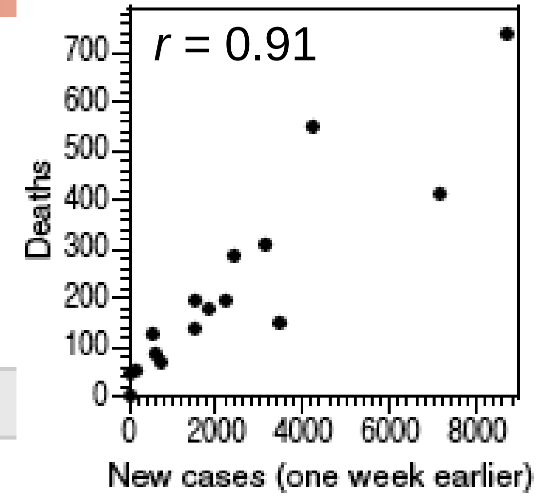
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	493.992	493.992	494.4971
Error	58	57.941	0.999	**Prob > F**
C. Total	59	551.932		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	–7.796	1.155	–6.751	<.0001*
LOGMPH	7.874	0.354	22.237	<.0001*

ANOVA and *t*-test give same p-value.

1918 flu epidemic: Relationship between the number of deaths in a given week and the number of new diagnosed cases one week earlier.



JMP – **Regression Analysis**

Linear Fit

Deaths = 49.29 + 0.0722 New cases (–1)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	SSM 494041	494041	68.27
Error	14	101308	7236	**Prob > F**
C. Total	15	SST 595349		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	49.2918	29.8454	1.65	0.1209
New cases (–1)	0.0722	0.0087	8.26	<.0001*

P-value for H_0: $\beta_1 = 0$; H_a: $\beta_1 \neq 0$

r^2 = SSM / SST = 0.8298

Relationship between average car speed and fuel efficiency

H_o: ρ = 0
H_a: ρ ≠ 0

We had n = 60 and r = 0.946

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$ = 22.225, df = 60-2 = 58.

p-value < 0.0001

There is a significant correlation (*r* is not 0) between fuel efficiency (MPG) and the logarithm of average speed (LOGMPH).