Lecture 8

Simple Linear Regression (cont.)

Section 10.1. Objectives:

- ✓Statistical model for linear regression
- ✓Data for simple linear regression
- ✓Estimation of the parameters
- ✓Confidence intervals and significance tests
- ✓Confidence intervals for mean response

vs.

- ✓Prediction intervals (for future observation)

Settings of Simple Linear Regression

- Now we will think of the least squares regression line computed from the sample as an estimate of the true regression line for the population.
- Different Notations than Ch. 2.Think b_0=a, b_1=b.

Type of line	**Least Squares Regression equation of line**	**slope**	**y-intercept**
Ch. 2 General	$\hat{y} = a + bx$	b	a
Ch. 10 Sample	$\mu_{\hat{y}} = b_0 + b_1 x$	b_1	b_0
Ch. 10 Population	$\mu_y = \beta_0 + \beta_1 x$	β_1	β_0

The statistical model for simple linear regression:

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

- Data: n observations in the form (x_1, y_1), (x_2, y_2), … (x_n, y_n).
- The **deviations** ε_i are assumed to be independent and normally distributed with mean 0 and constant standard deviation σ.
- The parameters of the model are: β_0, β_1, and σ.

ANOVA: groups with same SD and different means:

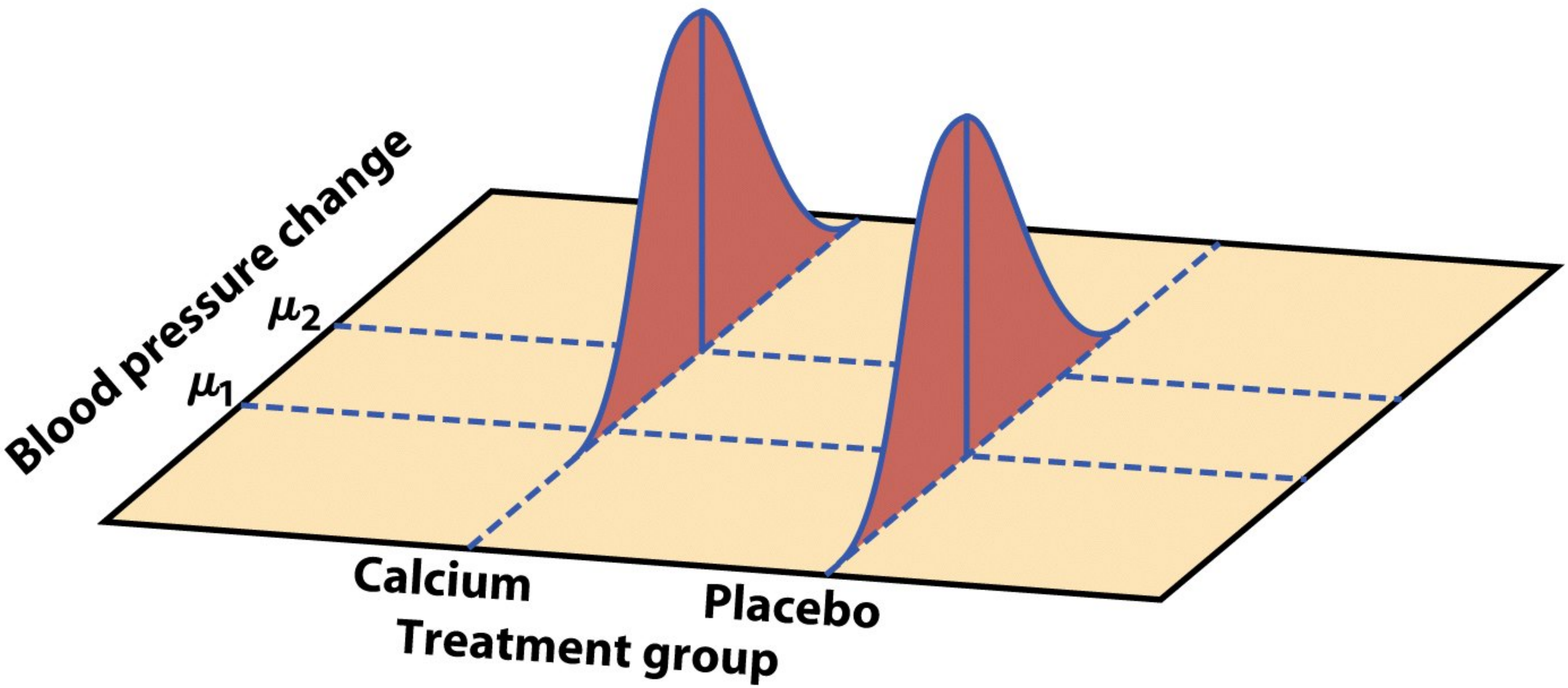


Figure 10-1
Introduction to the Practice of Statistics, Fifth Edition

Linear regression: many groups with means depending linearly on quantitative x

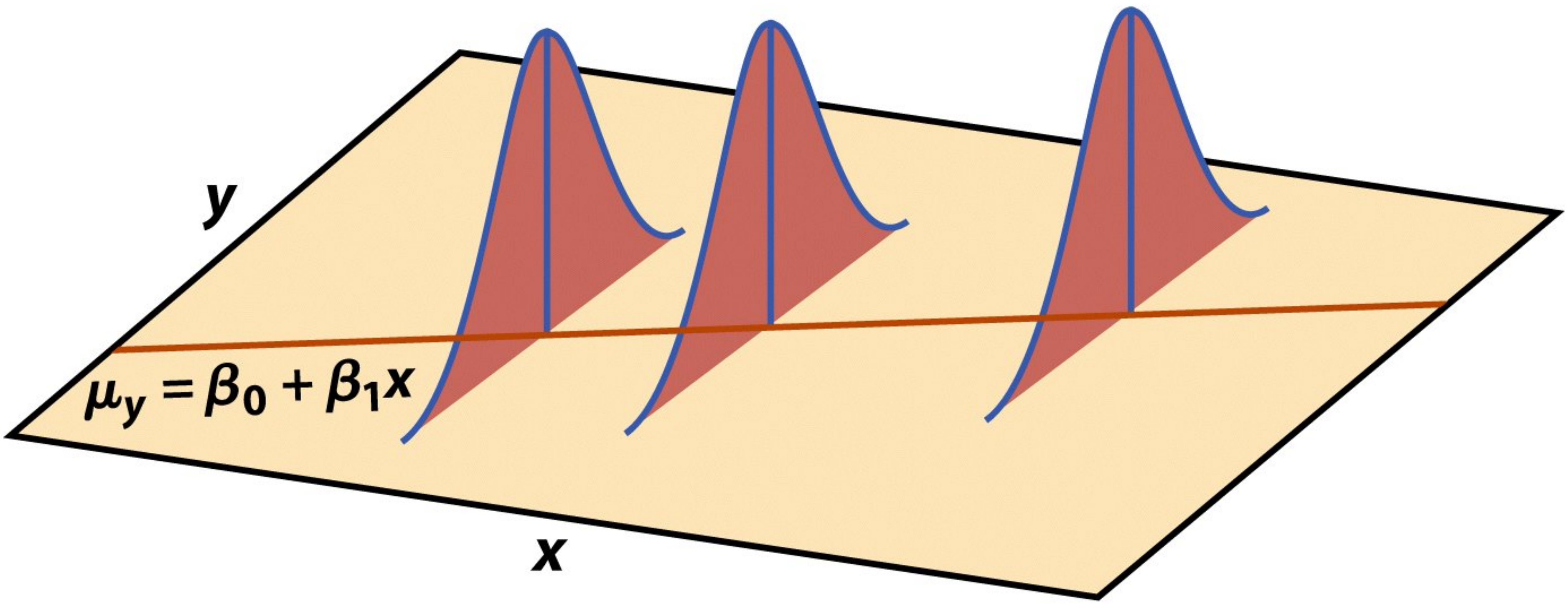


Figure 10-2
Introduction to the Practice of Statistics, Fifth Edition

Example: 10.1 page 636

- See R code.

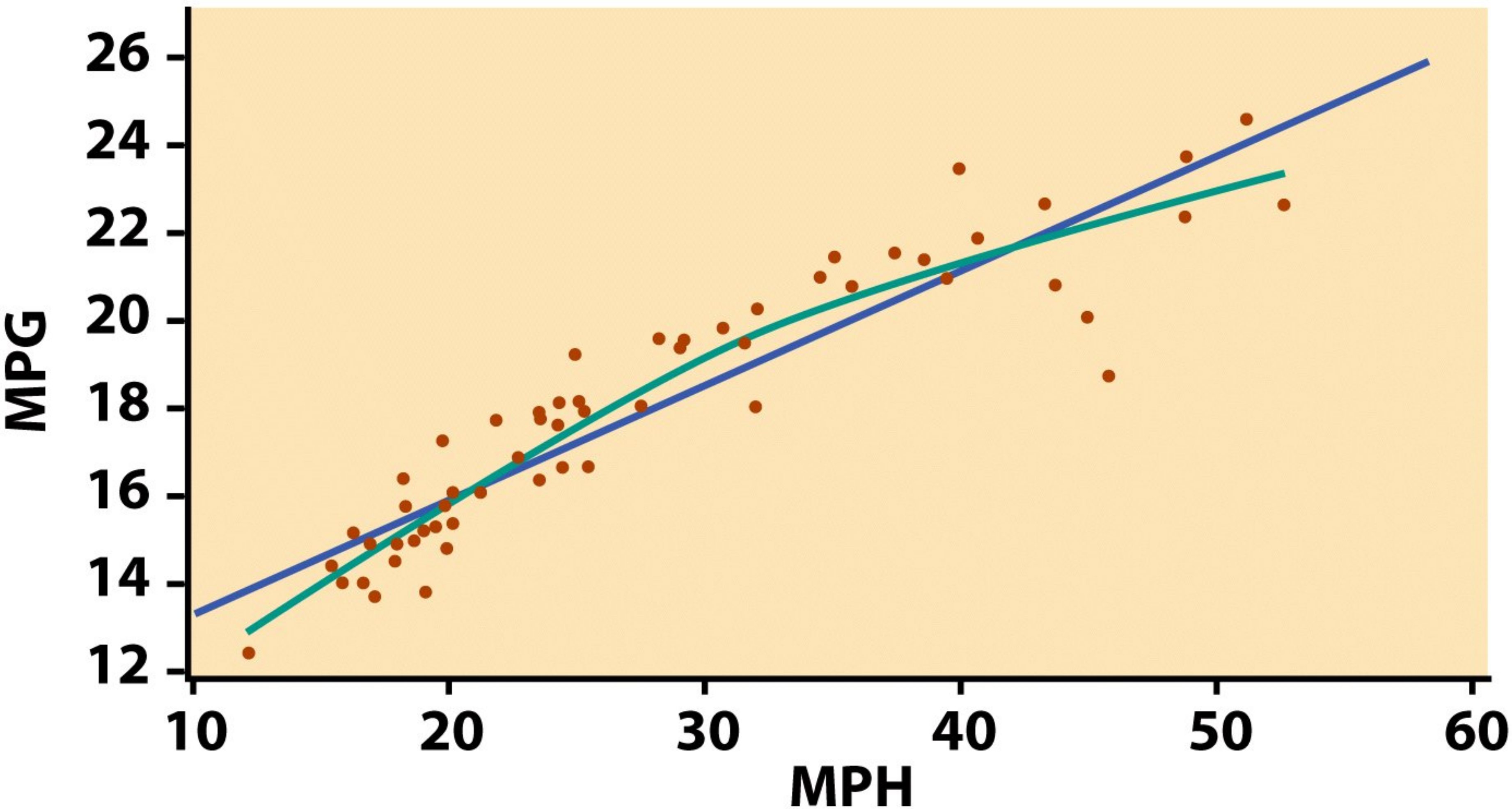


Figure 10-3
Introduction to the Practice of Statistics, Fifth Edition

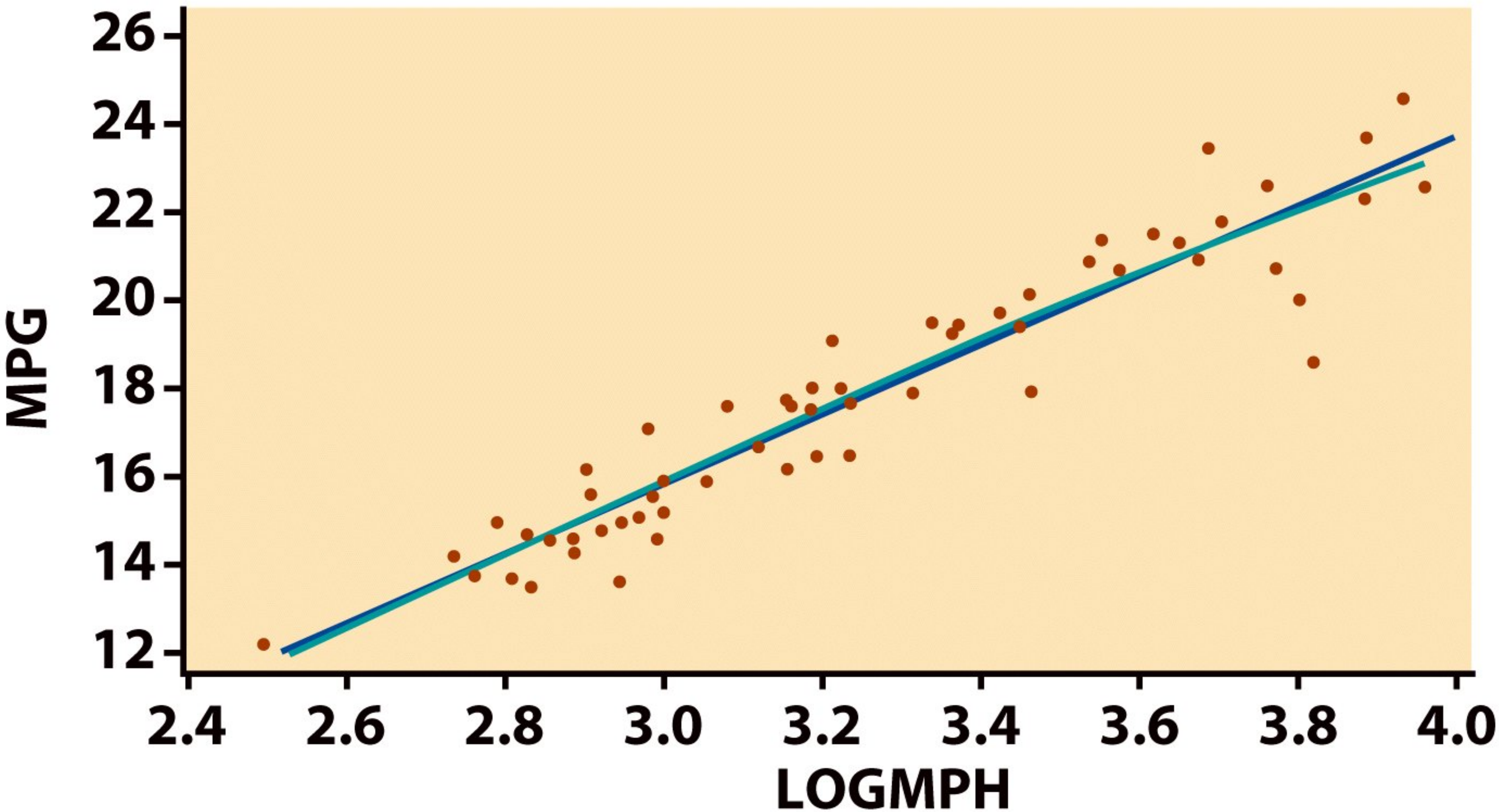


Figure 10-4
Introduction to the Practice of Statistics, Fifth Edition

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

Figure 10-5a
Introduction to the Practice of Statistics, Fifth Edition

```
The regression equation is
MPG = - 7.80 + 7.87 logmph

Predictor          Coef        StDev          T        P
Constant         -7.796        1.155      -6.75    0.000
logmph           7.8742       0.3541      22.24    0.000

S = 0.9995      R-Sq = 89.5%     R-Sq(adj) = 89.3%
```

Figure 10-5b
Introduction to the Practice of Statistics, Fifth Edition

Simple linear regression results:

Dependent Variable: MPG

Independent Variable: logmph

MPG = 7.7962503 + 7.874219 logmph

Sample size: 60

R (correlation coefficient) = 0.9461

Rñsq = 0.8950163

Estimate of error standard deviation: 0.99951637

Parameter estimates:

Parameter	Estimate	Std. Err.	DF	T -Stat	P-Value
Intercept	-7.7962503	1.1549443	58	-6.7503257	<0.0001
Slope	7.874219	0.3541106	58	22.236609	<0.0001

Figure 10-5c
Introduction to the Practice of Statistics, Fifth Edition

mpgmph60.xls

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.946053015					
5	R Square	0.895016308					
6	Adjusted R Square	0.893206244					
7	Standard Error	0.999516364					
8	Observations	60					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	493.9885883	493.9886	494.4668	4.50949E-30	
13	Residual	58	57.94391174	0.999033			
14	Total	59	551.9325				
15							
16		Coefficients	Standard Error	tStet	P-value	Lower 95%	Upper 95%
17	intercept	–7.796250129	1.154944262	–6.75033	7.69E-09	–10.10812052	–5.48437974
18	logmph	7.874219013	0.354110611	22.23661	4.51E-30	7.165390143	8583047883

Sheet1 Sheet2 Sheet3

Figure 10-5d

Introduction to the Practice of Statistics, Fifth Edition

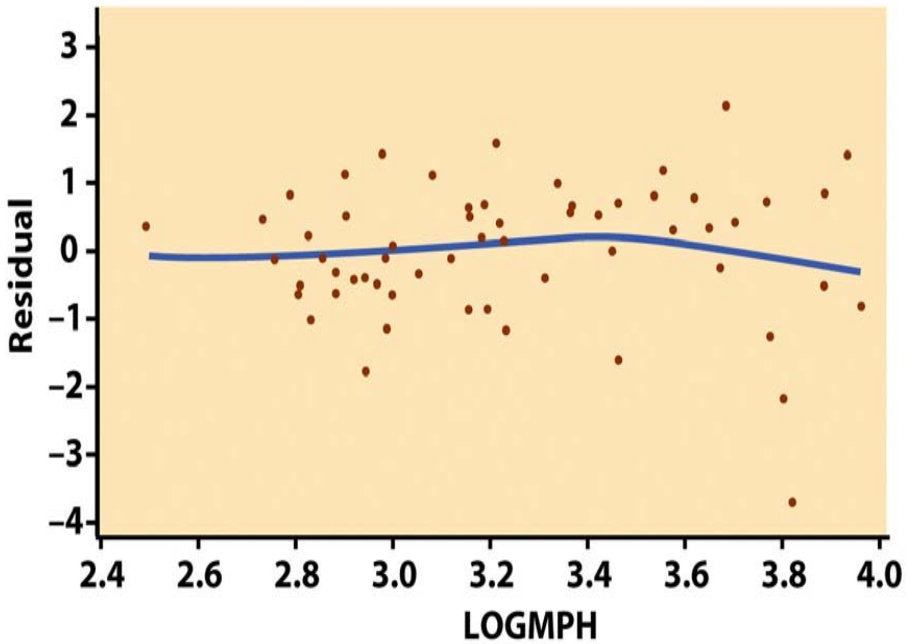
Root MSE	0.99952	R-Square	0.8950
Dependent Mean	17.72500	Adj R-Sq	0.8932
Coeff Var	5.63902		

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

Figure 10-5e
Introduction to the Practice of Statistics, Fifth Edition

Verifying the Conditions for inference:

- Look to the errors. They are supposed to be: -independent, normal and have the same variance.
- The errors are estimated using residuals: $(y - \hat{y})$



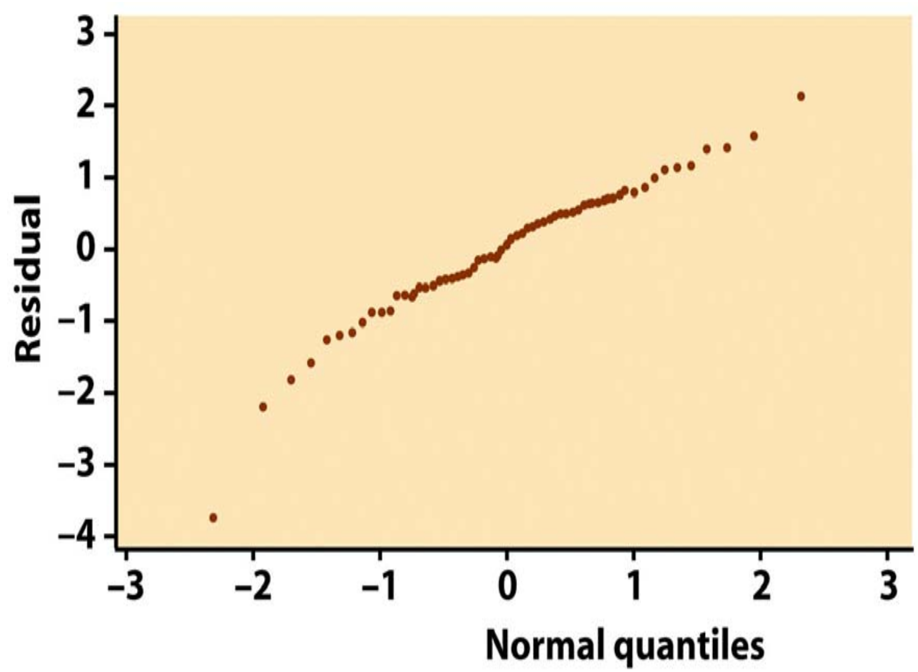
Residual plot:

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

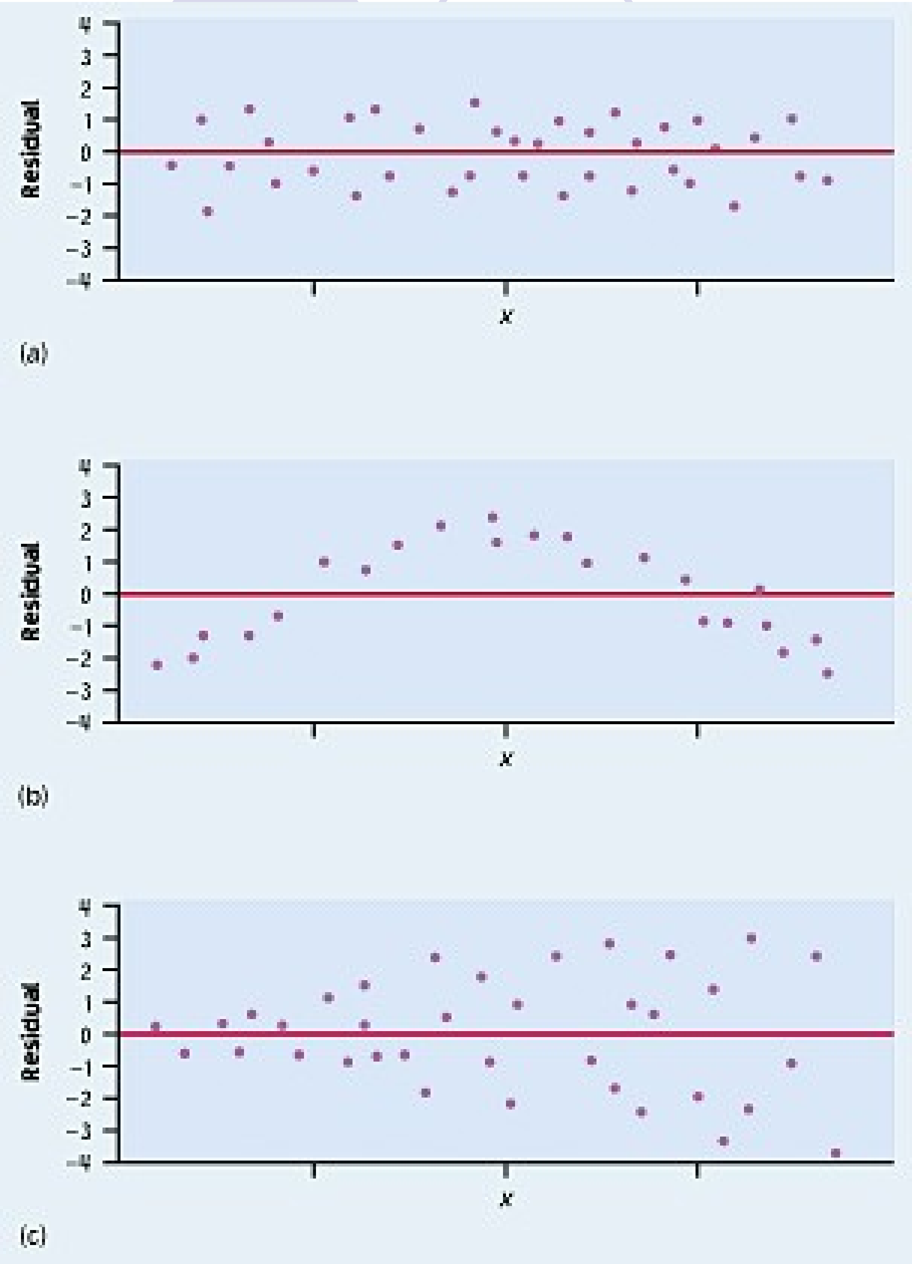
But we see one low residual (3.8, −4) and one potentially influential point (2.5, 0.5).

Normal quantile plot for residuals:

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



➔ Data okay for inference.



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*.

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 standard error of the least-squares slope:

$$b_1 \pm t^* \, \mathrm{SE}_{b1}$$

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

$$b_0 \pm t^* \, \mathrm{SE}_{b0}$$

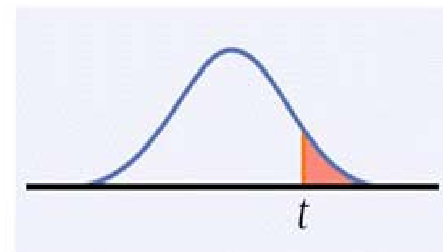
t^* is the critical value for the $t_{(n-2)}$ distribution with area C between $-t^*$ and $+t^*$.

Significance test for the slope

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

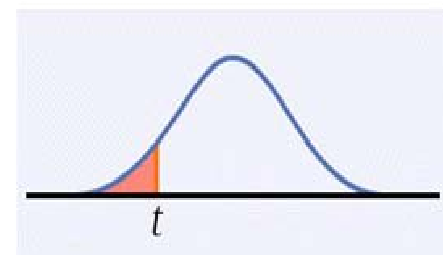
We calculate $\boxed{t = b_1 / SE_{b1}}$

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



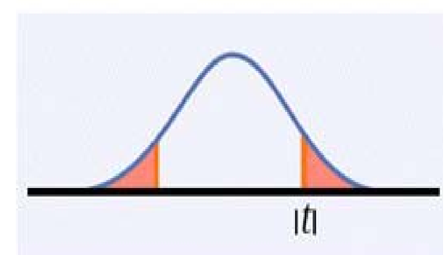
which has the ***t*** **(*n* – 2) distribution** to find the p-value of the test.

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



Note: Software typically provides two-sided p-values.

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



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.

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

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

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

<u>Note</u>: 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 software output for the car speed/gas efficiency example.

R Console

File Edit Misc Packages Help

```
> summary(model.2_logmodel)

Call:
lm(formula = MPG ~ LOGMPH, data = eg10.1)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7172 -0.5187  0.1121  0.6593  2.1490

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.7963     1.1549  -6.751 7.68e-09 ***
LOGMPH        7.8742     0.3541  22.237  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9995 on 58 degrees of freedom
Multiple R-Squared: 0.895,     Adjusted R-squared: 0.8932
F-statistic: 494.5 on 1 and 58 DF,  p-value: < 2.2e-16
```

Slope

Intercept

p-values for tests of significance

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

To obtain confidence intervals use the function confint()

Exercise: Calculate (manually) confidence intervals for the mean increase in gas consumption with every unit of (logmph) increase. Compare with software.

- confint(model.2_logmodel)

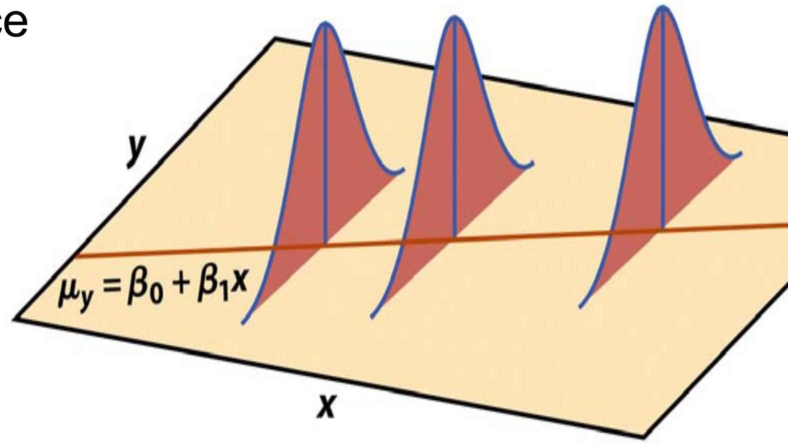
	2.5 %	97.5 %
LOGMPH	7.165435	8.583055

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.



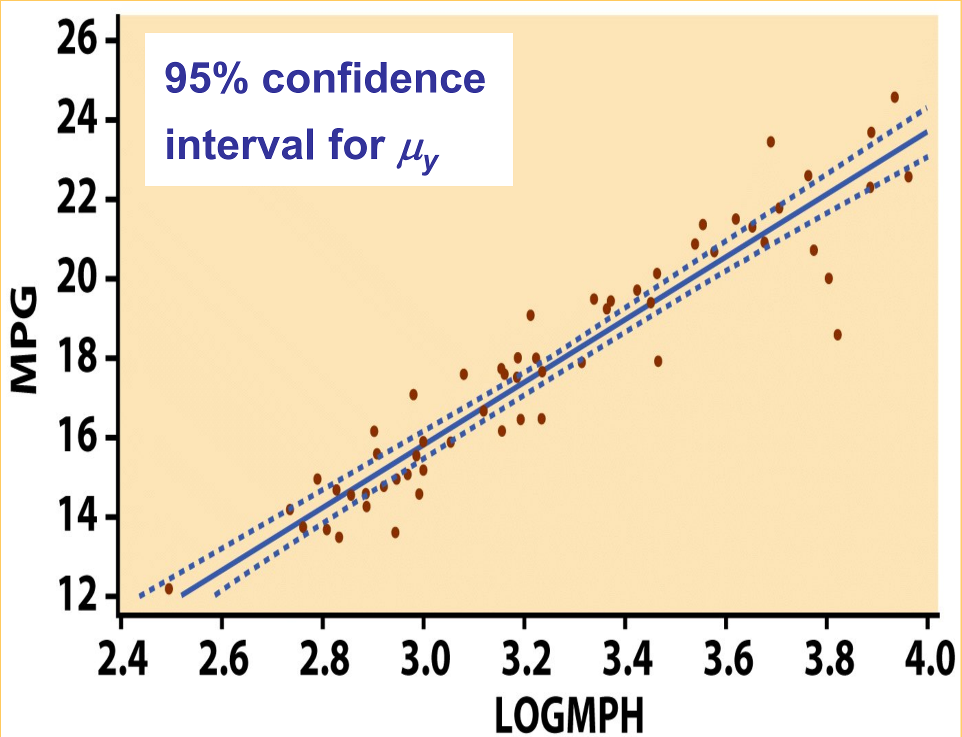
The **level C confidence interval for the mean response μ_y** at a given value x^* of x is centered on $\hat{y}$ (unbiased estimate of μ_y):

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

*t^* is the t critical 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

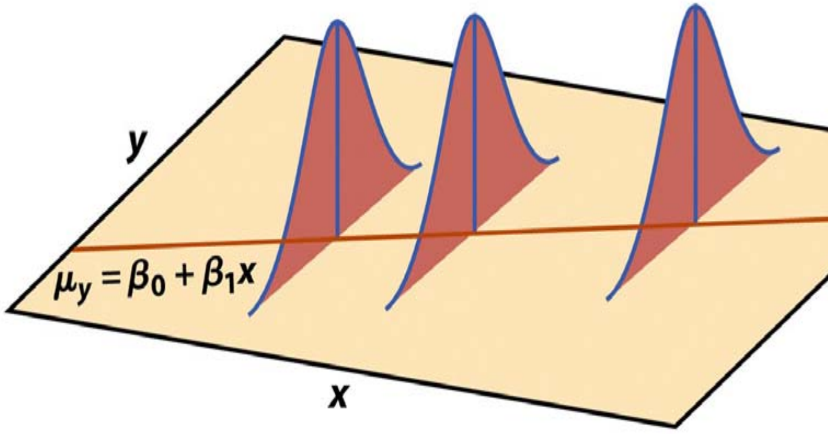
One use of regression is for **predicting** the value of y, $\hat{y}$, for any value of x within the range of data tested: $\hat{y} = b_0 + b_1x$.

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.



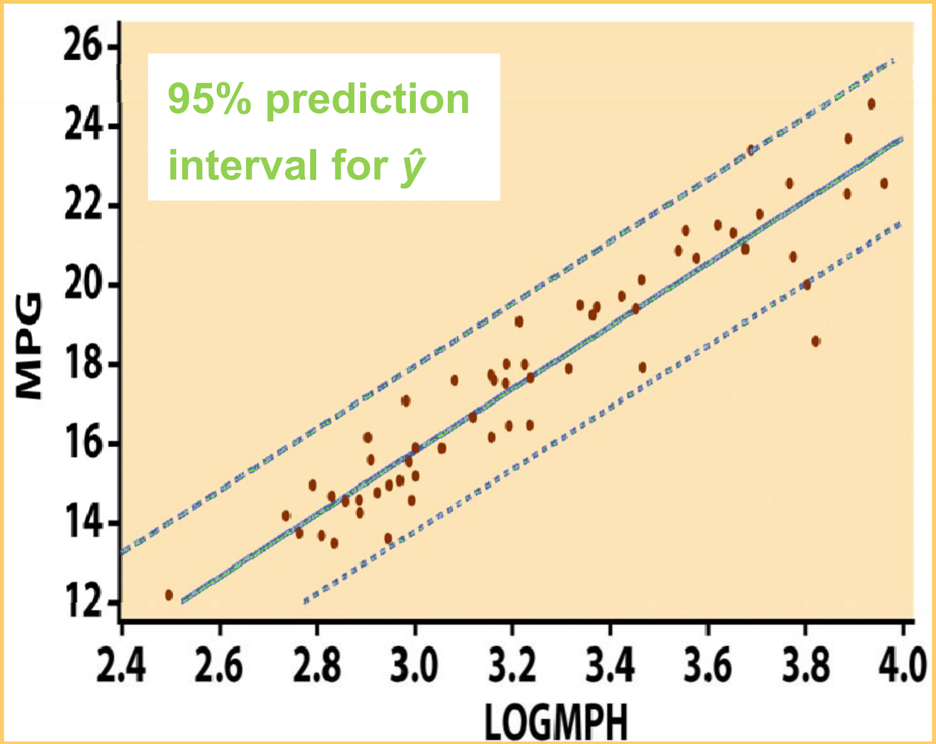
The **level C prediction interval for a single observation** on *y* when *x* takes the value *x** is:

$$\boxed{C \pm t^*_{n-2}\,\mathrm{SE}_{\hat{y}}}$$

t is the t critical 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 is shown as a continuous interval on either side of $\hat{y}$.



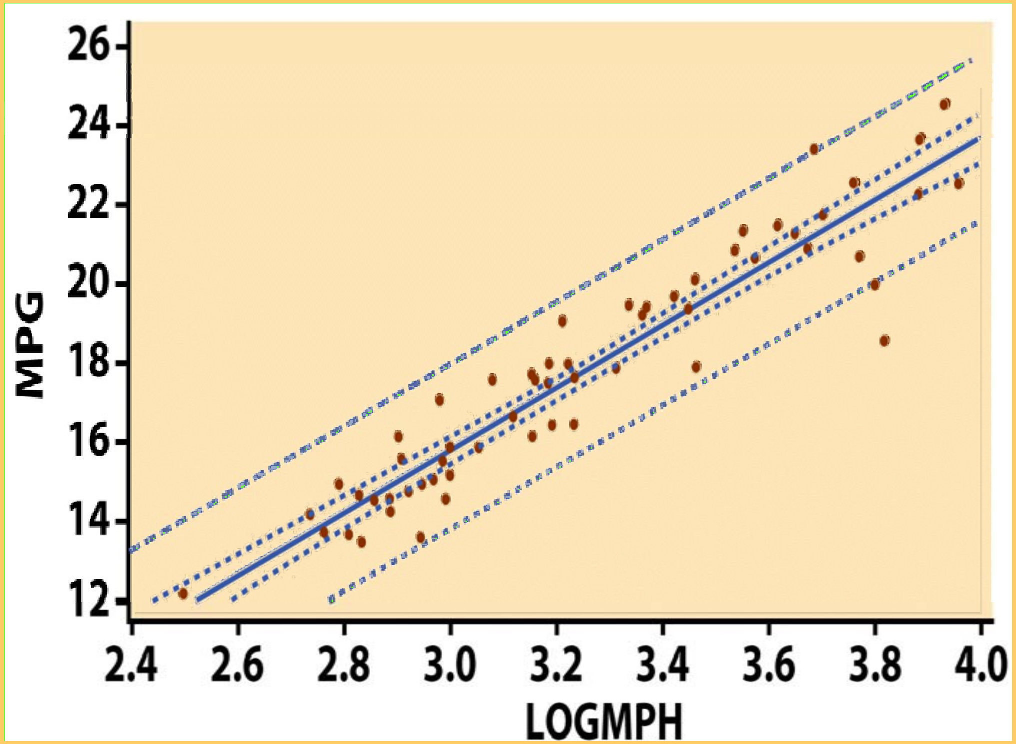
- The **confidence interval for μ_y** contains with *C*% confidence the population mean μ_y of all responses at a particular value of *x*.

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

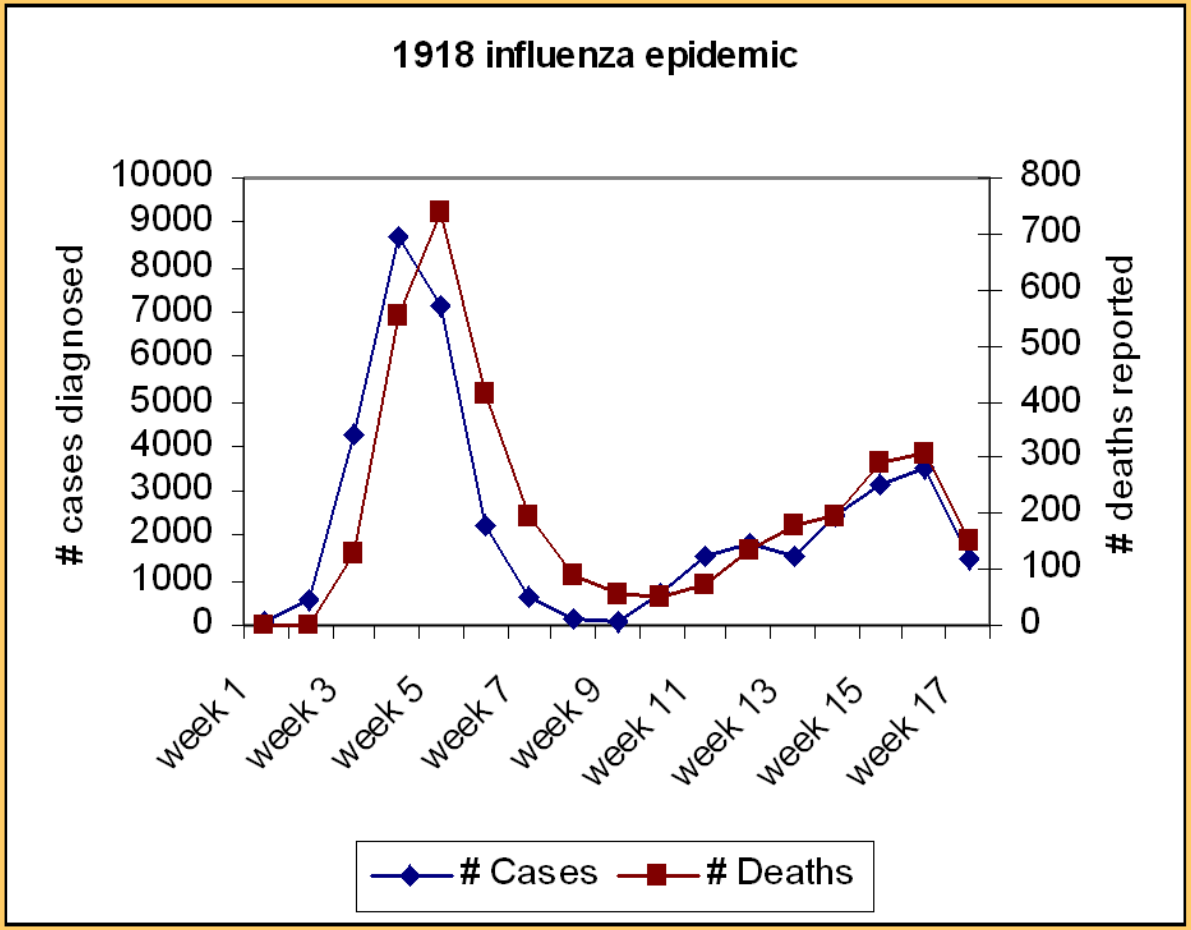
95% prediction interval for $\hat{y}$

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

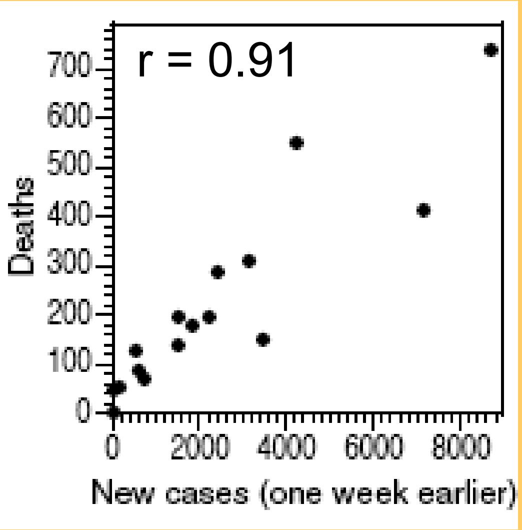


1918 influenza epidemic		
Date	# Cases	# Deaths
week 1	36	0
week 2	531	0
week 3	4233	130
week 4	8682	552
week 5	7164	738
week 6	2229	414
week 7	600	198
week 8	164	90
week 9	57	56
week 10	722	50
week 11	1517	71
week 12	1828	137
week 13	1539	178
week 14	2416	194
week 15	3148	290
week 16	3465	310
week 17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

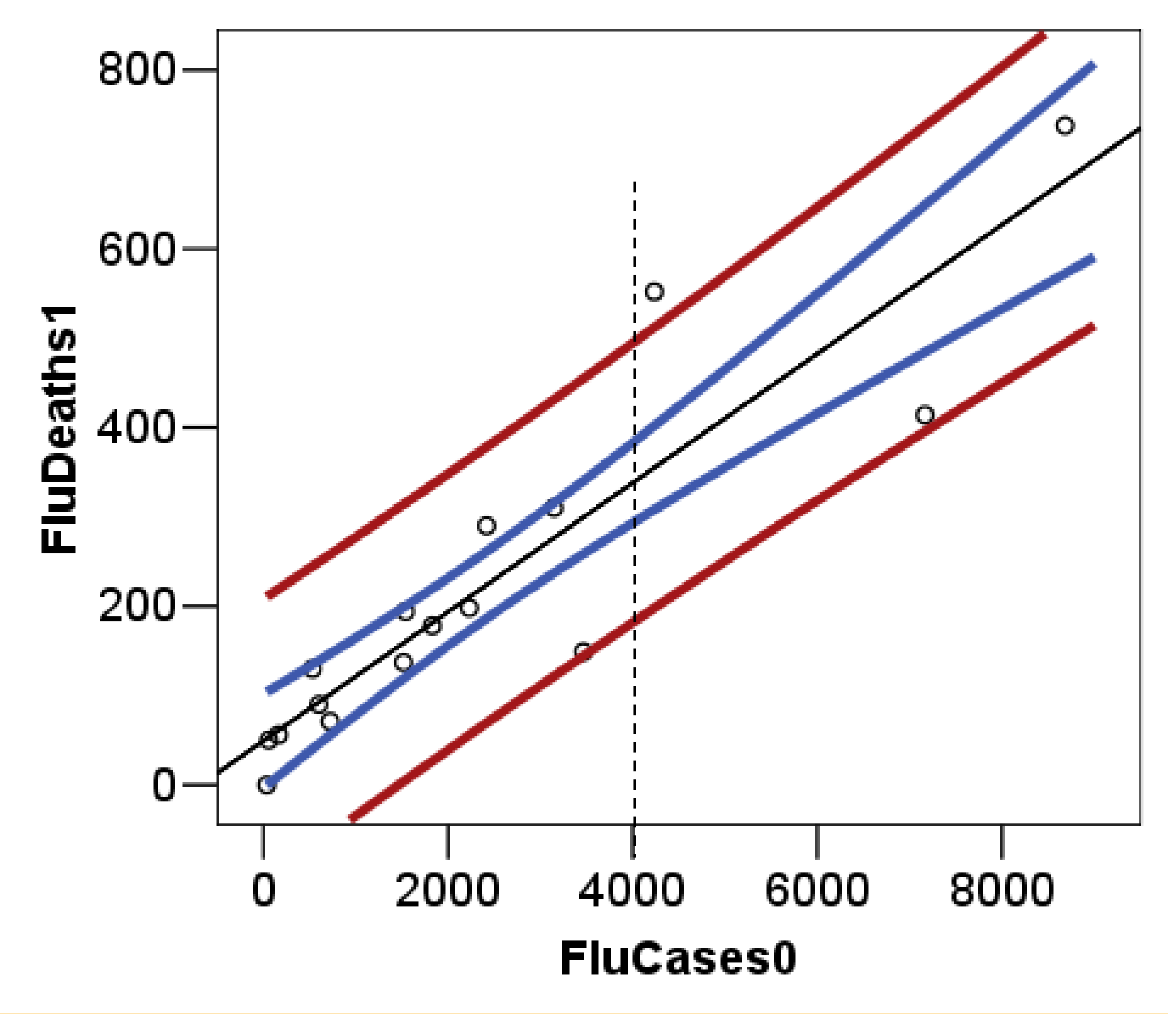
Regression Statistics

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.



Least squares regression line

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

95% confidence interval for μ_y

CI for mean weekly death count one week after 4000 flu cases are diagnosed: μ_y within about 300–380.

Prediction interval for a weekly death count one week after 4000 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.