# Inference for Regression

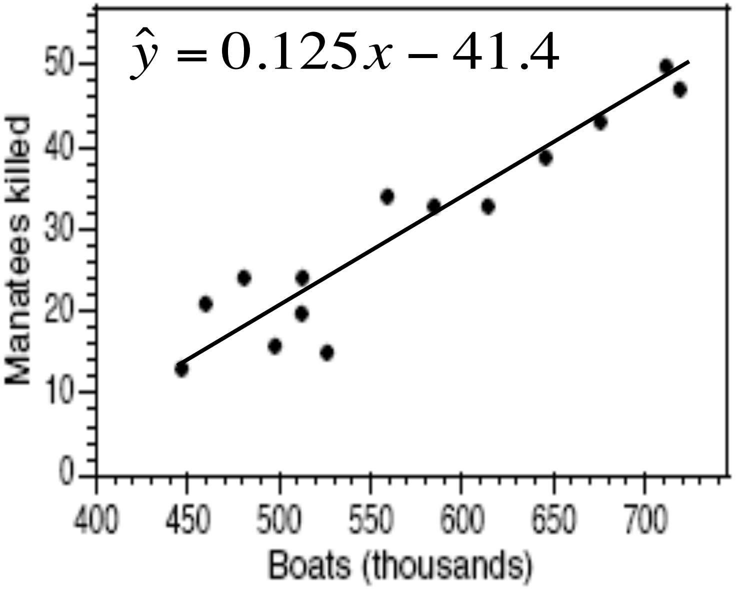
## Simple Linear Regression

IPS Chapter 10.1

# Objectives (IPS Chapter 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 $\mu_y$
- Prediction intervals



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

We want to describe the **population mean response $\mu_y$** 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 $\mu_y$ **= $\beta_0$ + $\beta_1 x$.**

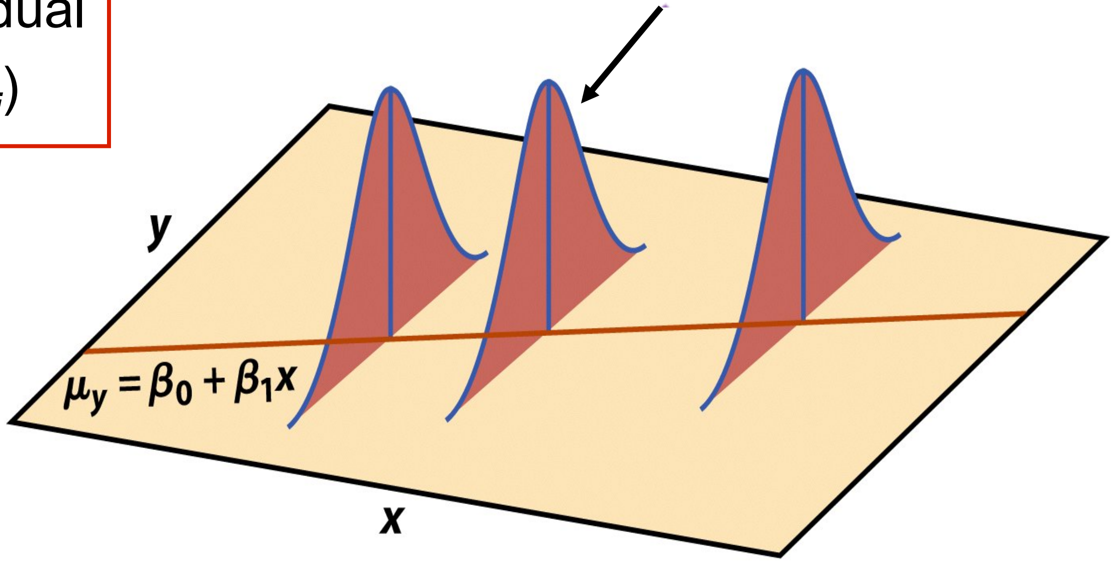
Sample data then fits the model:

$$\begin{array}{c}\text{Data} \;=\\ y_i \;=\end{array} \boxed{\begin{array}{c}\text{fit}\\ (\beta_0+\beta_1 x_i)\end{array}} \begin{array}{c}+\\ +\end{array} \boxed{\begin{array}{c}\text{residual}\\ (\varepsilon_i)\end{array}}$$

where the $\varepsilon_i$ are

**independent** and

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


For any fixed *x*, the responses *y* follow a Normal distribution with standard deviation σ.

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

(σ is the same for all values of *x*).

# Estimating the parameters

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

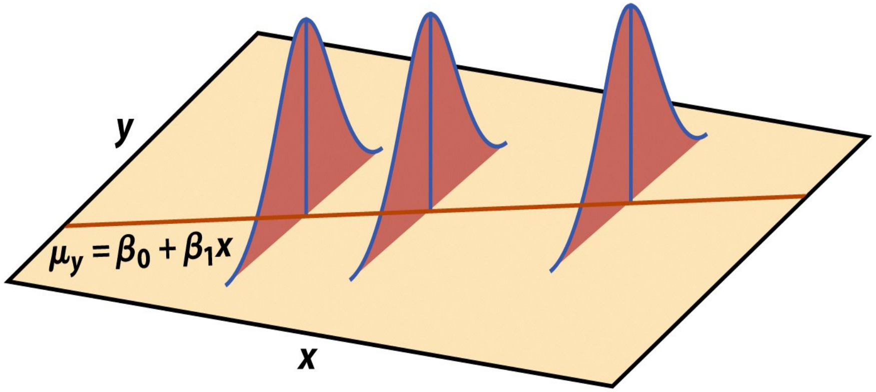
The intercept $\beta_0$, the slope $\beta_1$, and the standard deviation $\sigma$ of *y* are the unknown parameters of the regression model. We rely on the sample data to provide unbiased estimates of these parameters.

- The value of $\hat{y}$ from the least-squares regression line (remember Chapter 2?) is really a prediction of the mean value of *y* ($\mu_y$) 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 $\mu_y$

$b_0$: unbiased estimate for intercept $\beta_0$

$b_1$: unbiased estimate for slope $\beta_1$



The **population standard deviation σ** for *y* at any given value of *x* represents the spread of the normal distribution of the $\varepsilon_i$ around the mean $\mu_y$ .

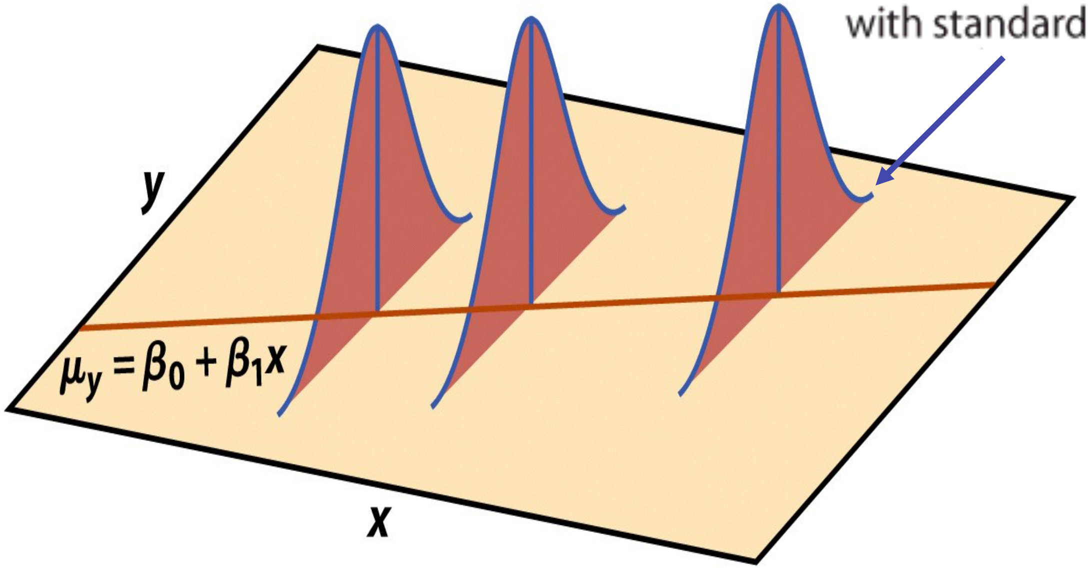
The estimate of σ is calculated from the **residuals,** $e_i = y_i - \hat{y}_i$ :

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

*s* is an estimate of the population standard deviation σ.

# Conditions for inference

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

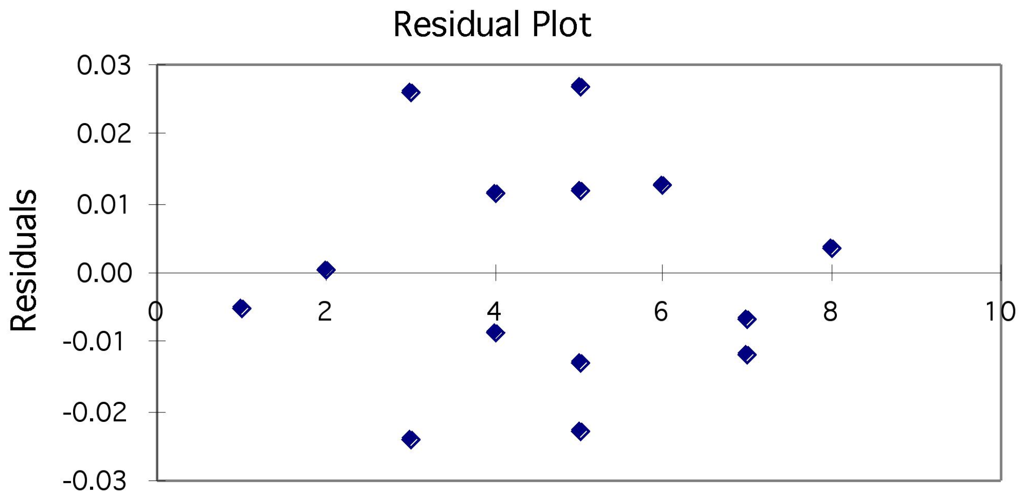


For any fixed *x*, the responses *y* follow a Normal distribution with standard deviation σ.

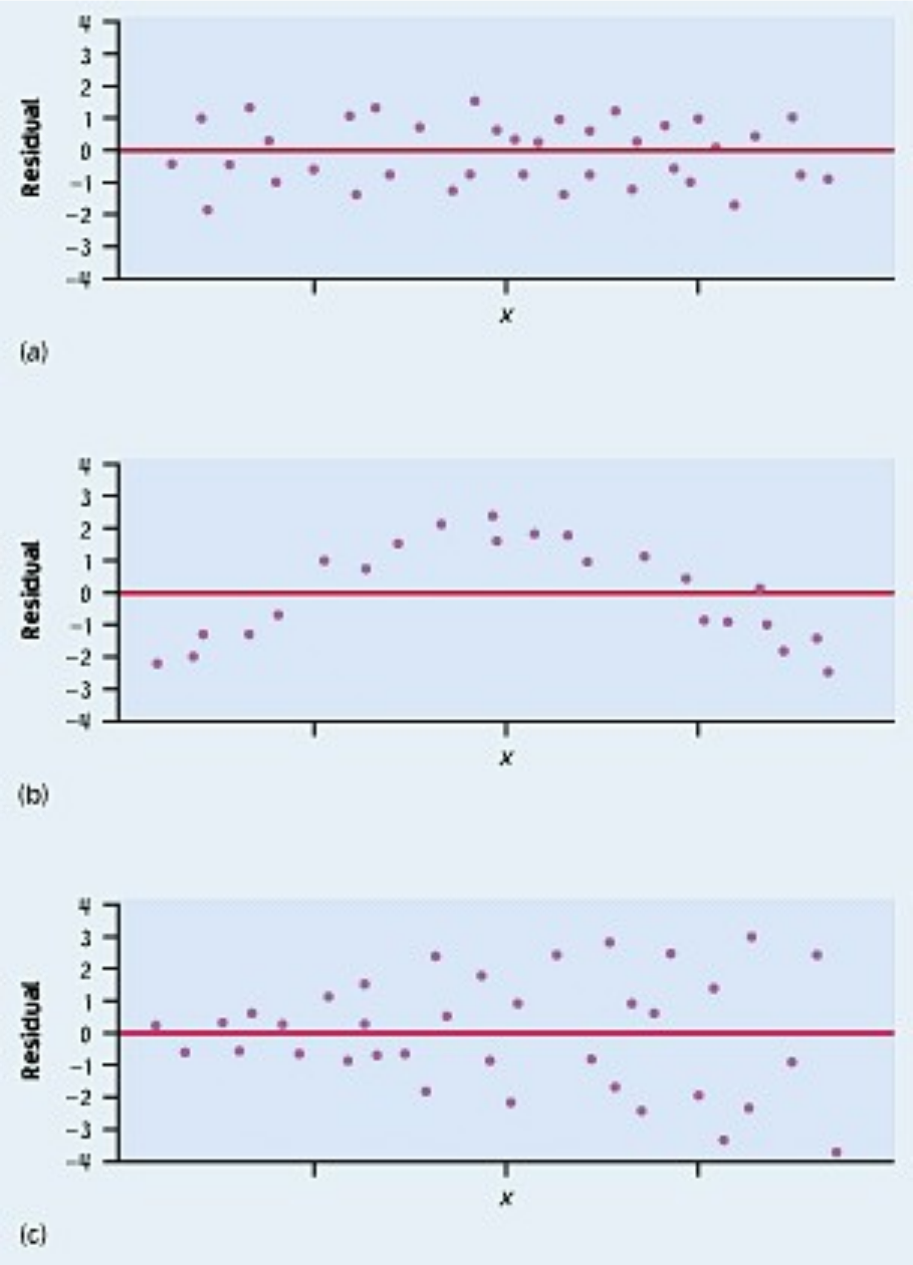
# Using residual plots to check for regression validity

The residuals give useful information about the contribution of individual data points to the overall pattern of scatter.

We view the residuals in a **residual plot** of $e_i$ *vs.* $x_i$



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*.

# Confidence intervals for regression parameters

Estimating the regression parameters $\beta_0$ and $\beta_1$ 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, $\beta_1$,** is based on an estimate of the standard deviation of the estimated slope, $b_1$:

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

A level C **confidence interval for the intercept, $\beta_0$,** is based on an estimate of the standard deviation of the estimated intercept, $b_0$:

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

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

We'll see formulas for the SE values in 10.2.

# Significance test for the slope

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

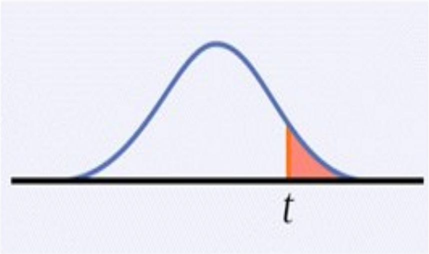
We calculate $\boldsymbol{t = b_1 / SE_{b1}}$,

which has a ***t* (*n* – 2)**

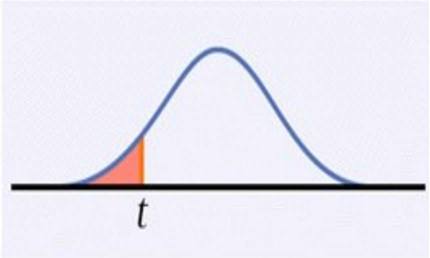
**distribution** to find the

p-value of the test.

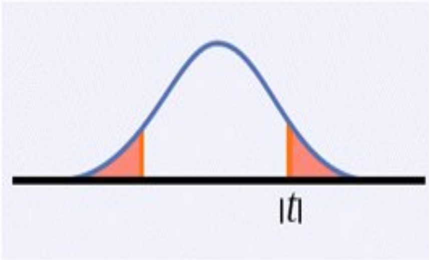
$H_a: \beta_1 > 0$ is $P(T \geq t)$



$H_a: \beta_1 < 0$ 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.*

# Confidence interval for $\mu_y$

We can calculate a **confidence interval for the population mean $\mu_y$** of all responses *y* when *x* takes the value *x**:

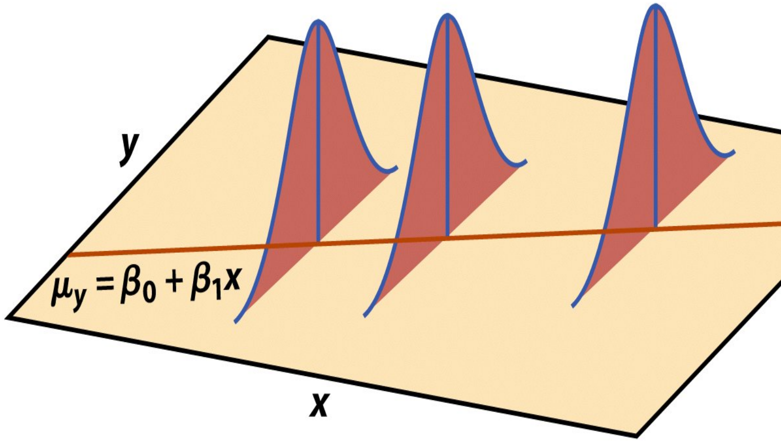
This interval is centered on $\hat{y}$, the unbiased estimate of $\mu_y$, and has

The typical form of a CI: estimate ± $t^*SE_{\text{estimate}}$.

The true value of the population mean $\mu_y$ when x = x*

will be within our confidence interval in

C% of samples taken from the population.

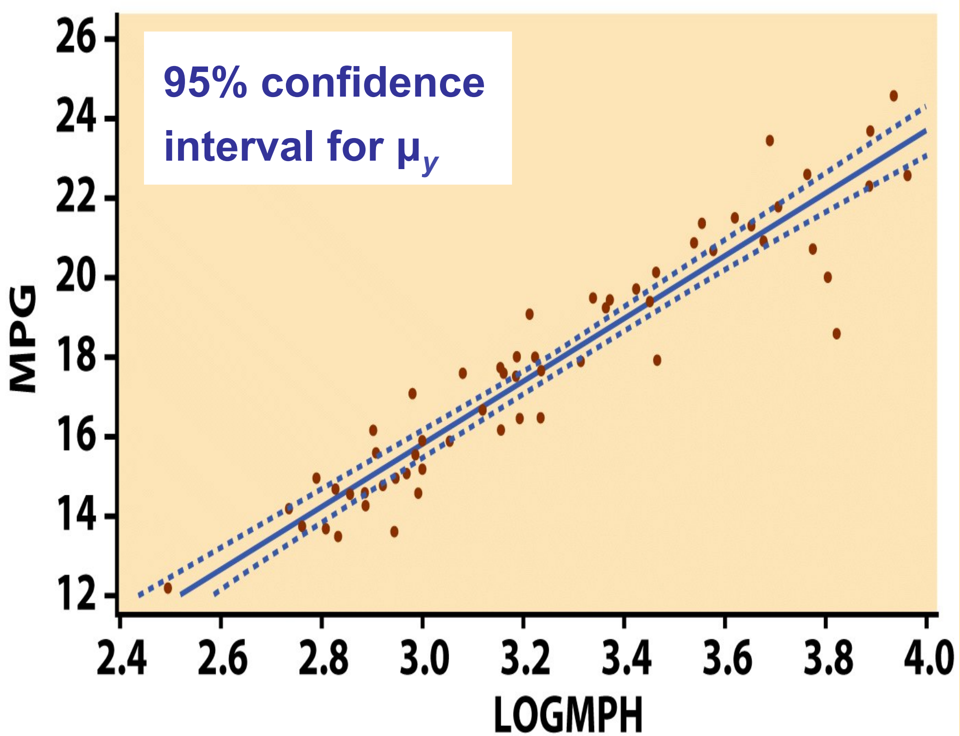


The **level *C* confidence interval for the mean response *$\mu_y$*** at a given value *x** of *x* is:

$$\hat{\mu}_y \pm t^* SE_{\hat{\mu}_y}$$

*t* is the value from the **t (n – 2)** distribution with area C between –t* and +t*. Again, we'll hold off on the SE term for now.*

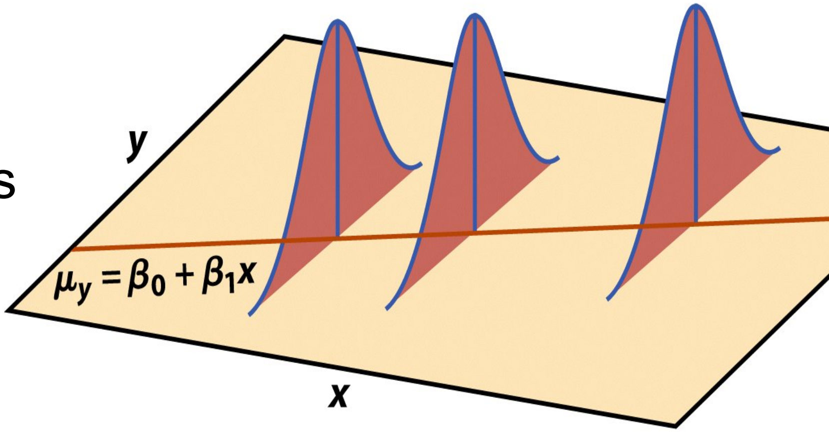
A confidence interval for $\mu_y$ can be found for any value, x*. Graphically, these confidence intervals are shown as a pair of smooth curves centered on the LS regression line. We often call these "confidence bands".



# Inference for prediction

We have seen that one use of regression is to **predict** the value of *y* for a particular value of x: $\hat{y} = b_0 + b_1x$. Now that we understand the ideas of statistical inference, we can employ them to help us understand the variability of that simple estimate.

To predict an individual response *y* for a given value of *x,* we use a **prediction interval.** The prediction interval is wider than the CI for mean response to reflect the fact that individual observations are more variable.

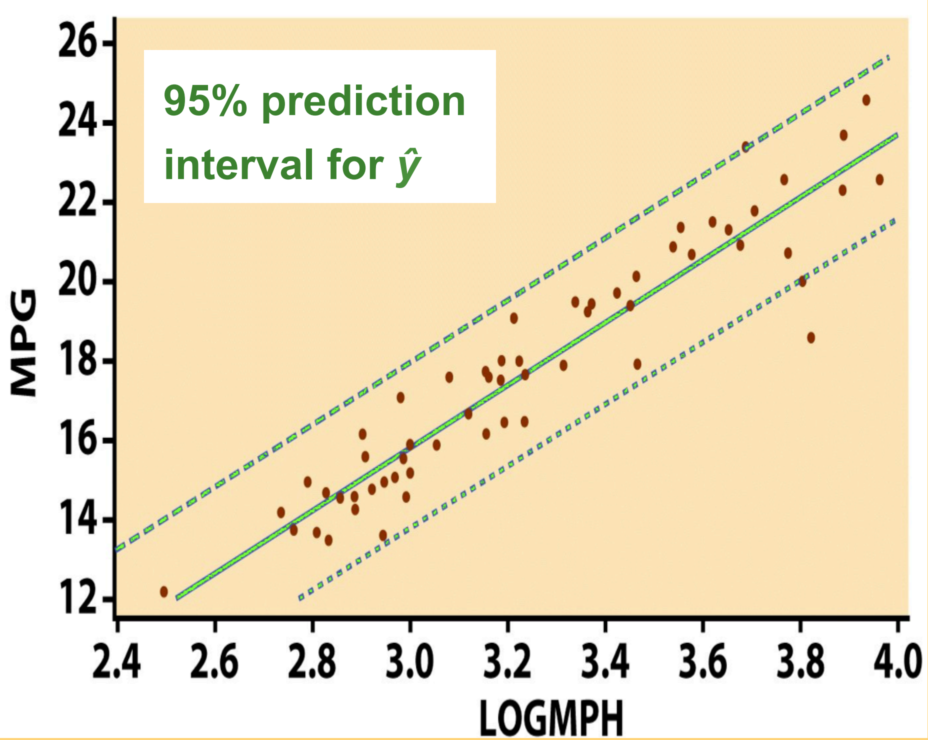


The **level *C* prediction interval for a single observation** of *y* when *x* takes the value *x** is:

$$\hat{y} \pm t^* SE_{\hat{y}}$$

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

We can display the prediction intervals in the same way as the CI for mean response.



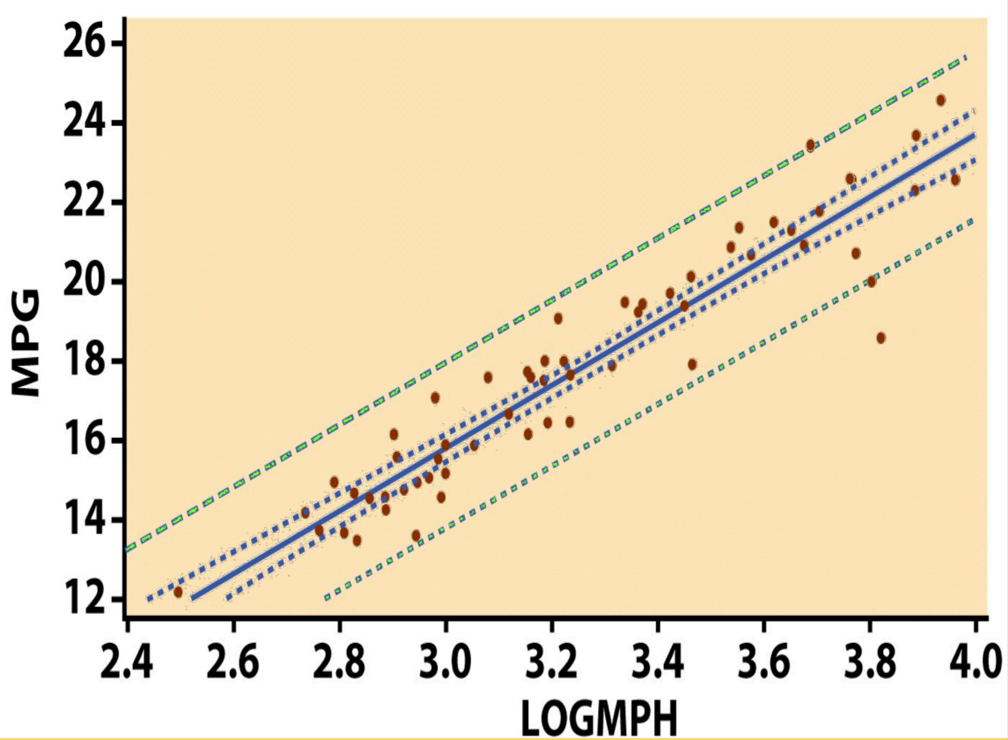
- The **confidence interval for $\mu_y$** estimates the mean value of y for all individuals in the population whose value of x is x*.

- The **prediction interval** predicts the value of y for one single individual whose value of x is x*.

95% prediction interval for *y*

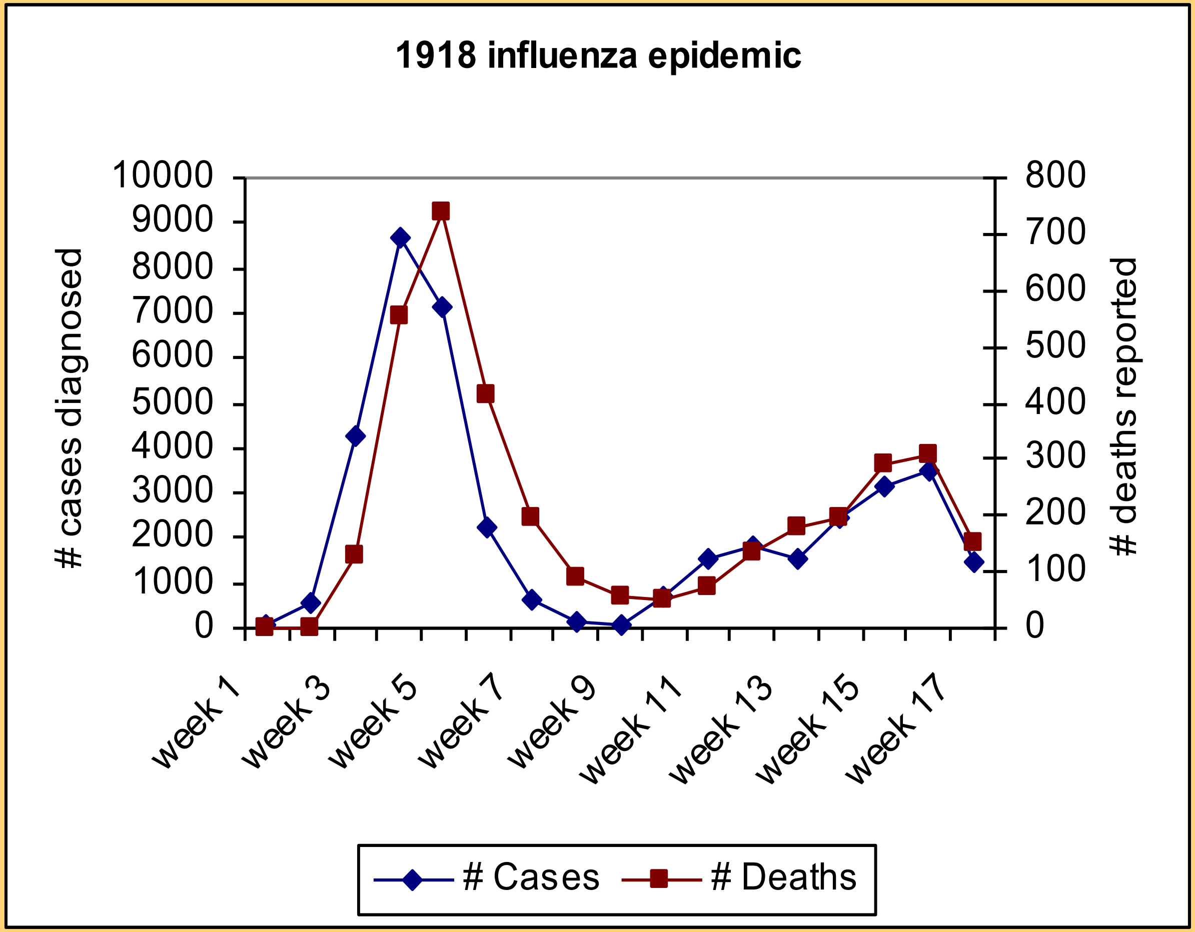
95% confidence interval for $\mu_y$

This figure illustrates the fact that the CI for mean response is narrower than the corresponding prediction interval. Notice that both intervals are most narrow for values near the center of the X distribution.



# 1918 flu epidemic

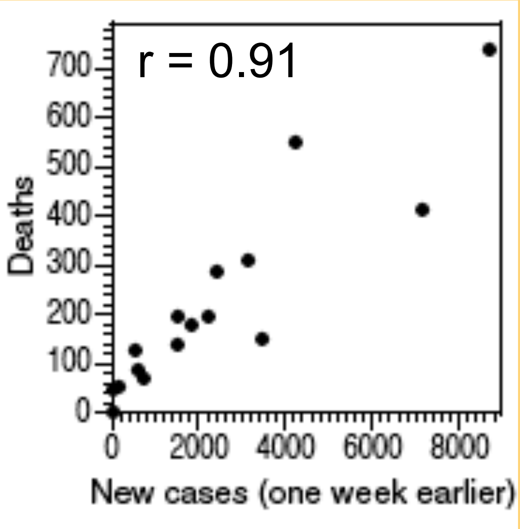
| 1918 influenza epidemic | | |
|---|---|---|
| **Date** | **# Cases** | **# Deaths** |
| week 1 | 36 | 0 |
| week 2 | 531 | 0 |
| week 3 | 4233 | 130 |
| week 4 | 8682 | 552 |
| week 5 | 7164 | 738 |
| week 6 | 2229 | 414 |
| week 7 | 600 | 198 |
| week 8 | 164 | 90 |
| week 9 | 57 | 56 |
| week 10 | 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 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**

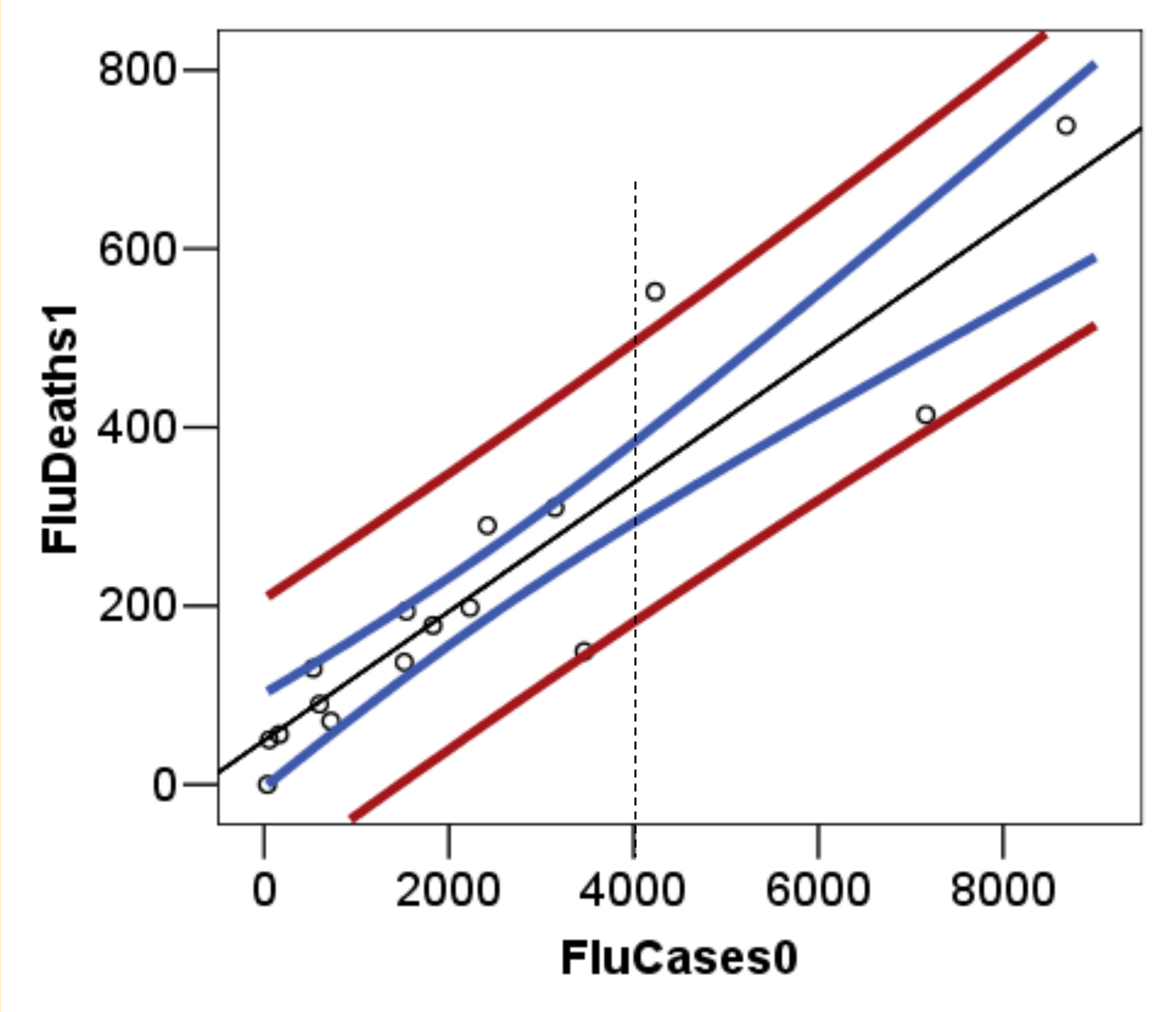
| **Regression Statistics** | |
|---|---|
| Multiple R | 0.911 |
| R Square | 0.830 |
| Adjusted R Square | 0.82 |
| Standard Error | 85.07 s |
| Observations | 16.00 |

| | *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 |
| | **$b_1$** | $SE_{b1}$ | | **P-value for $H_0$: $\beta_1 = 0$** | | |

P-value very small ➔ reject $H_0$ ➔ $\beta_1$ 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 $\mu_y$**

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

Prediction interval for a weekly death count one week after 4000 flu cases are diagnosed: *y* within about 180–500 deaths.

# Inference for Regression

## More Detail about Simple Linear Regression

IPS Chapter 10.2

# Objectives (IPS Chapter 10.2)

**Inference for regression—more details**

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

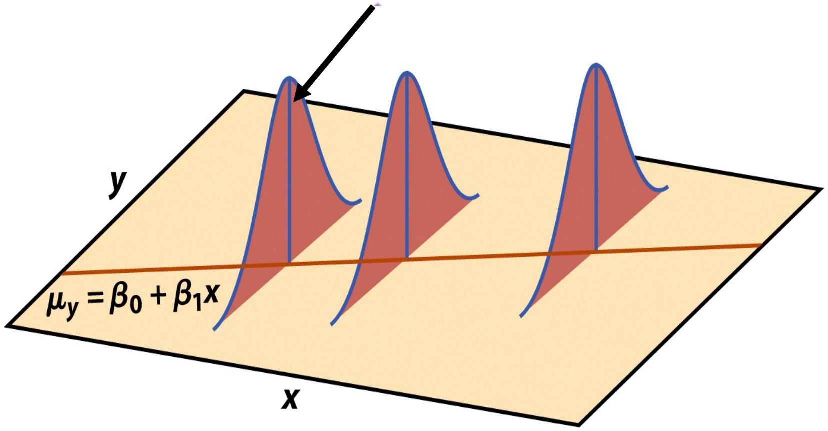
# Analysis of variance for regression

The regression model is:

Data = fit + residual

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

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


For any fixed *x*, the responses *y* follow a Normal distribution with standard deviation σ.

The calculations are based on a technique called **Analysis of Variance**, or **ANOVA**. ANOVA partitions the total amount of variation in a data set into two or more **components of variation**.

# Analysis of Variance (ANOVA)

In the linear regression setting, there are two reasons that values of y differ from one another:

- The individuals have different values of X
- Random variation among individuals with the same value of X

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

Total variation of DATA

Variation due to RESIDUALS

Variation due to model FIT

$$\text{SST} = \text{SSM} + \text{SSE}$$

# Sums of Squares, Degrees of Freedom, and Mean Squares

We have just seen that SSTotal = SSModel +SSError. Each sum of squares has an associated **degrees of freedom**, which also add,

$$\text{DFT} = \text{DFM} + \text{DFE}.$$

For SLR, DFT = n-1, DFM = 1, DFE = n-2

Each SS is also associated with **Mean Squares**, which do NOT add:

$$\text{MS} = \text{SS}/\text{DF}.$$

MST = SST/DFT, MSM = SSM/DFM, MSE = SSE/DFE

# ANOVA and correlation

Fact: the correlation between x and y is related to the percentage of variation in y explained by the FIT of the linear model:

$$r^2 = \frac{SSM}{SST}$$

If x and y are highly correlated, SST is dominated by the SSM term, with little contribution from the residuals (SSE); conversely, if x and y are weakly correlated, SST is dominated by SSE.

$r^2$ is often called the **coefficient of determination**.

# The ANOVA *F* test

The null hypotheses that *y* is not linearly related to *x*,

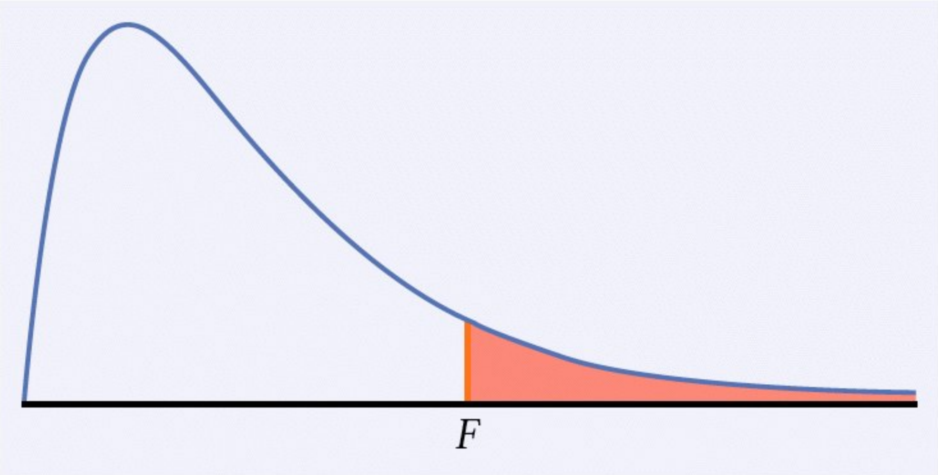
$$H_0\text{: } \beta_1 = 0 \text{ versus } H_a\text{: } \beta_1 \neq 0$$

can be tested by comparing MSM to MSE.

$$\mathbf{F = MSM / MSE}$$

When $H_0$ is true, *F* follows the *F*(1, *n* − 2) distribution.

The p-value is the probability of observing a value of *F* greater than the observed one.



# ANOVA table

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

The estimate of σ is calculated from the residuals $e_i = y_i – \hat{y}_i$:

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

# Calculations for regression inference

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

**The standard error of the least-squares slope $b_1$ is:**

$$SE_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x}_i)^2}}$$

**The standard error of the intercept $b_0$ is:**

$$SE_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x}_i)^2}}$$

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

**The standard error of the mean response $\mu_y$ 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 *y* is:**

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

# Inference for correlation

The sample correlation coefficient, r, can be used as an estimate of a population-level correlation, ρ. We can test $H_0$: ρ = 0 with a t statistic:

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

A P-value can be found using the t(n-2) distribution.

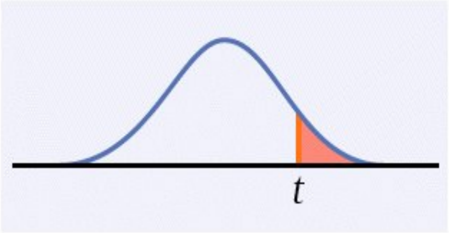
The test of significance for $\rho$ uses the one-sample *t*-test for: $H_0$: $\rho = 0$.

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

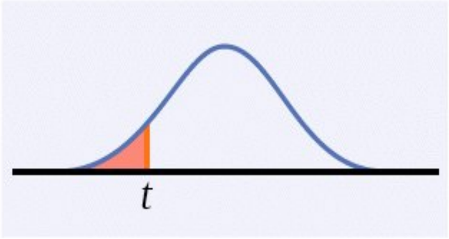
$$\boxed{t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}}$$

The p-value is the area under *t* (*n* – 2) for values of *T* as extreme as *t* 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 $2P(T \geq |t|)$