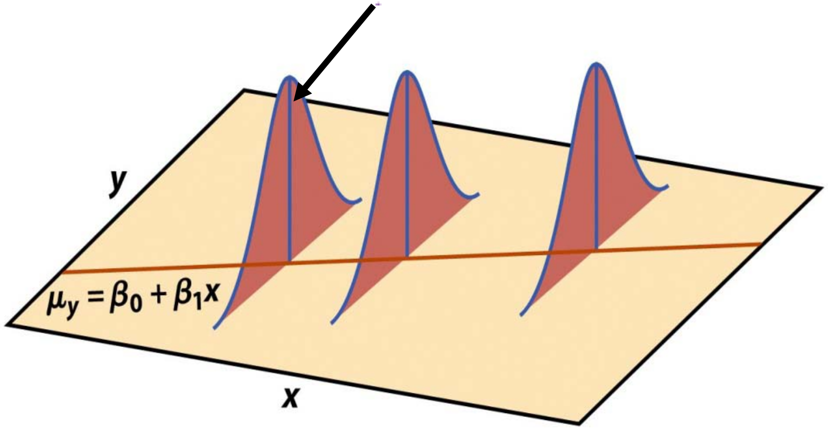
Lecture 9

Simple Linear Regression

ANOVA for regression (10.2)

Analysis of variance for regression

The regression model is:



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

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

Sums of squares measure the variation present in responses. It can be partitioned as:

SST = SS model + SS error

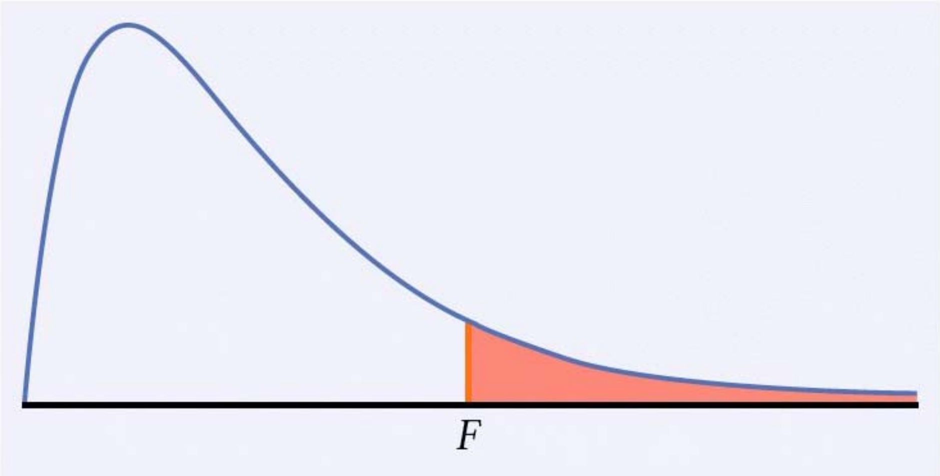
DFT = DF model + DF error

For a simple linear relationship, the ANOVA tests the hypotheses

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

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).



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	SSG/DFG	MSG/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 $e_i = y_i - \hat{y}_i$

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

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***.

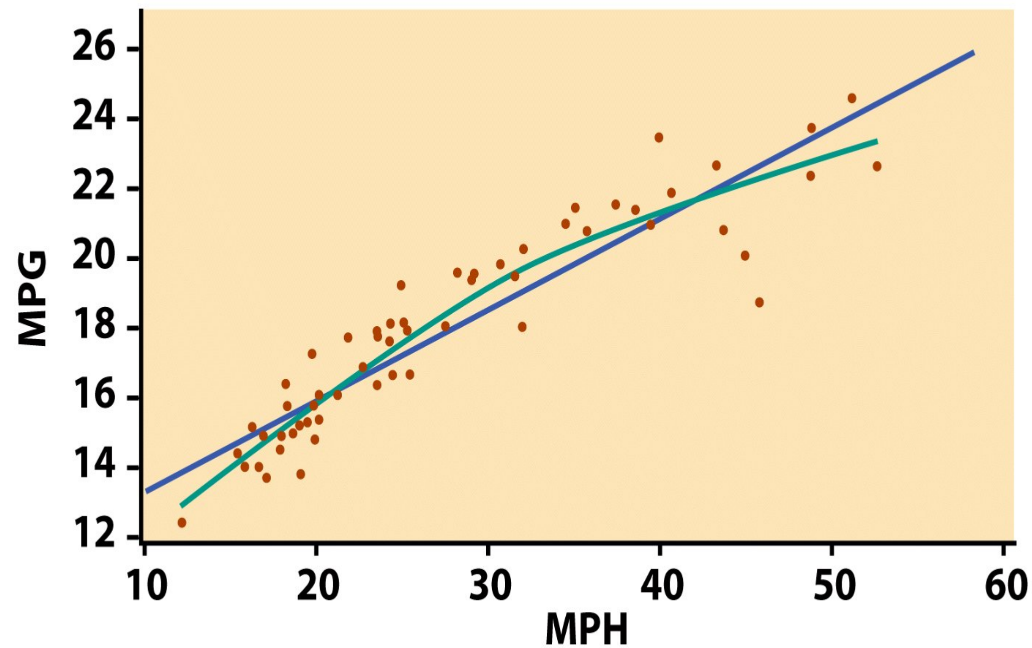
$$r^2 = \frac{\text{variation in } y \text{ caused by } x \quad \text{(i.e., the regression line)}}{\text{total variation in observed } y \text{ values around the mean}}$$

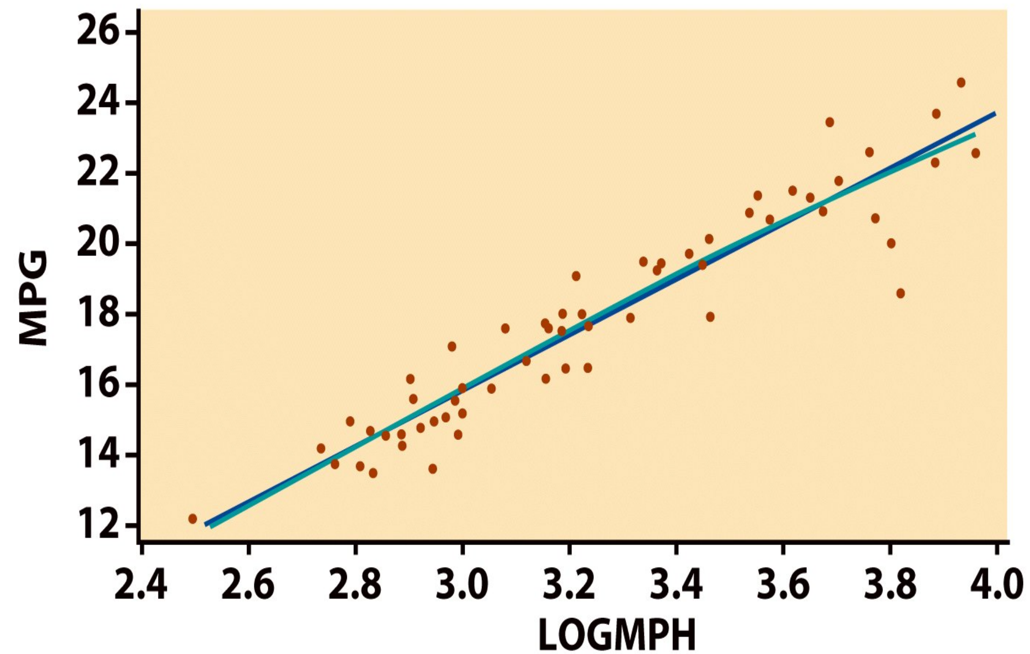
$$\boxed{r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\text{SSM}}{\text{SST}}}$$

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.





```
> anova(lm(MPG ~ LOGMPH, data=eg10.1))
Analysis of Variance Table

Response: MPG
          Df Sum Sq Mean Sq F value    Pr(>F)
LOGMPH     1 493.99  493.99   494.5 < 2.2e-16 ***
Residuals 58  57.94    1.00
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(lm(MPG ~ LOGMPH, data=eg10.1))

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
```

SST (sum of squares total) is the sum of the two

R-squared is the ratio: SSM/SST=494/552

In this case both tests check the same thing that is why the p-value is identical

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:

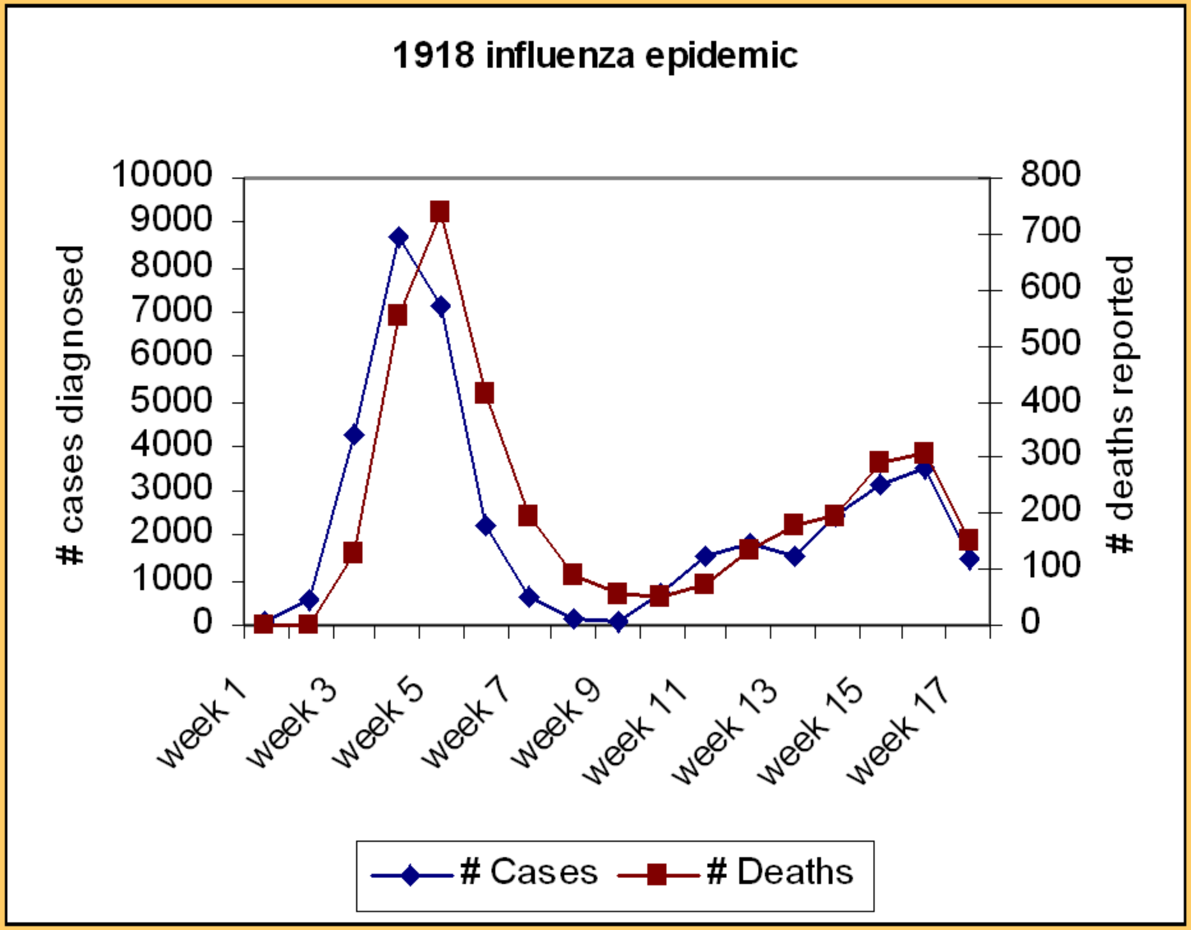
$$\mathrm{SE}_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x - \bar{x})^2}}$$

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

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

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.

```
> summary(lm(Deaths[-1]~Cases[-length(Cases)]))

Call:
lm(formula = Deaths[-1] ~ Cases[-length(Cases)])

Residuals:
     Min        1Q    Median        3Q       Max
-152.688   -23.998    -3.361    35.759   196.994

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           49.291806  29.845387   1.652    0.121
Cases[-length(Cases)]  0.072222   0.008741   8.263 9.38e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 85.07 on 14 degrees of freedom
Multiple R-Squared: 0.8298,     Adjusted R-squared: 0.8177
F-statistic: 68.27 on 1 and 14 DF,  p-value: 9.382e-07
```

SE_{b0}

SE_{b1}

$s = \sqrt{MSE}$

R^2 = SSM / SST

P-value for

H_0: $\beta = 0$; H_a: $\beta \neq 0$

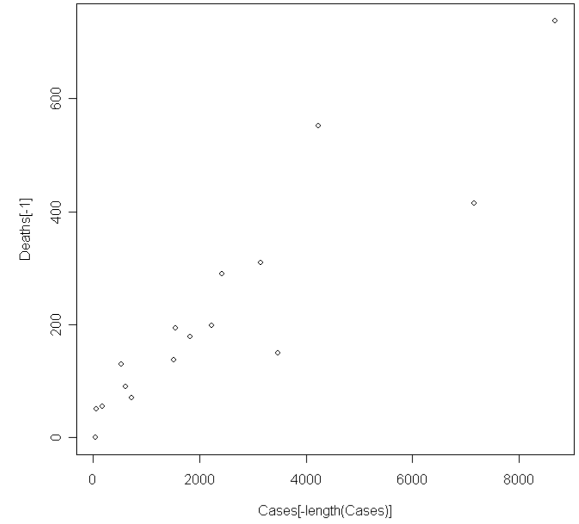
```
> anova(lm(Deaths[-1]~Cases[-length(Cases)]))
Analysis of Variance Table

Response: Deaths[-1]
                      Df Sum Sq Mean Sq F value    Pr(>F)
Cases[-length(Cases)]  1 494041  494041  68.273 9.382e-07 ***
Residuals             14 101308    7236
---
```

$MSE = s^2$

SSM

SST 595349



Inference for correlation

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

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

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

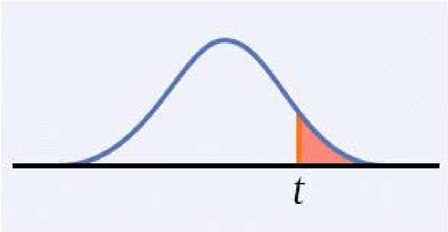
The test of significance for ρ uses the one-sample t-test for: H_0: $\rho = 0$.

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

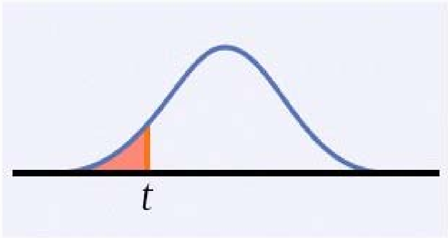
$$\boxed{t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}}$$

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

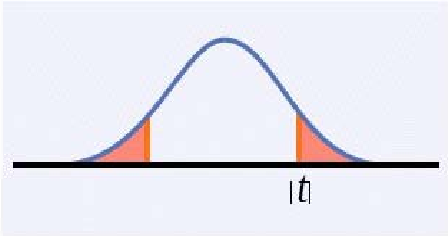
$H_a: \rho > 0$ is $P(T \geq t)$



$H_a: \rho < 0$ 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).