Testing the Usefulness of the Model

For the SLR model, $Y = \beta_0 + \beta_1 X + \varepsilon$.

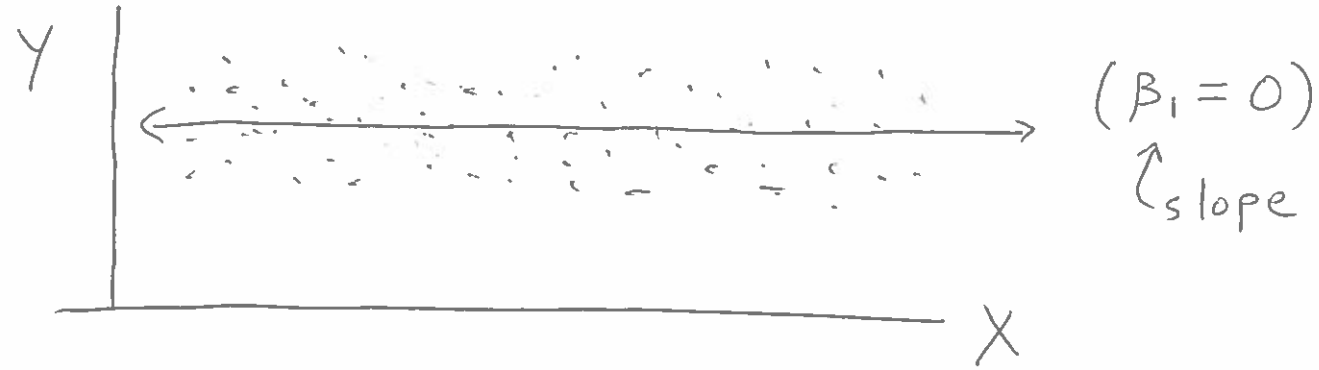
Note: *X* is completely useless in helping to predict *Y* if and only if $\beta_1 = 0$.

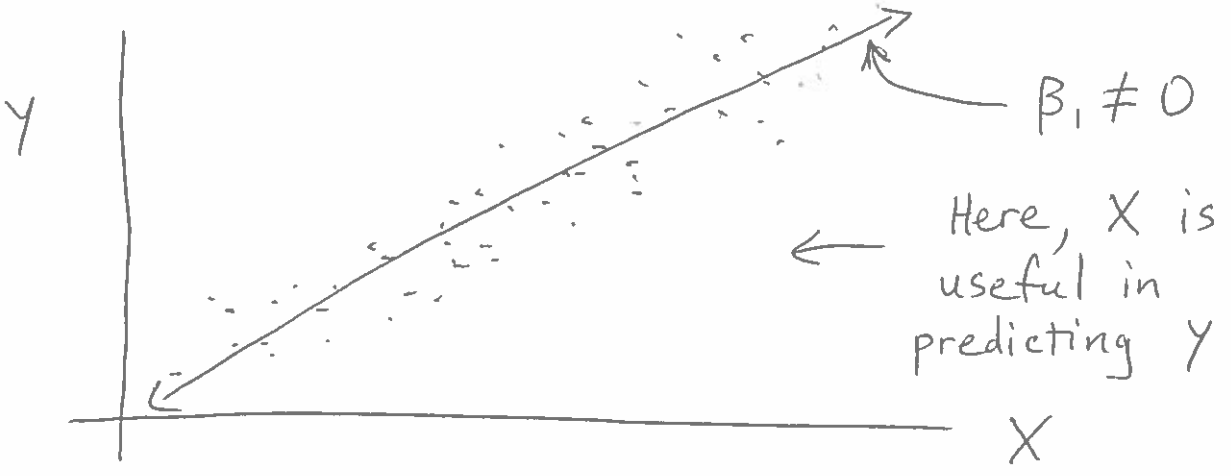
So to test the usefulness of the model for predicting *Y*, we test: $H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

If we reject H_0 and conclude H_a is true, then we conclude that *X* does provide information for the prediction of *Y*. (X is linearly related to Y)

Picture:





Recall that the estimate $\hat{\beta}_1$ is a statistic that depends on the sample data.

This $\hat{\beta}_1$ has a sampling distribution.

If our four SLR assumptions hold, the sampling distribution of $\hat{\beta}_1$ is normal with mean β_1 and standard deviation $\frac{\sigma}{\sqrt{SS_{xx}}}$ which we estimate by $\frac{s}{\sqrt{SS_{xx}}}$

Under H_0: $\beta_1 = 0$, the statistic $\frac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}}$

has a t-distribution with $n - 2$ d.f.

Test for Model Usefulness

One-Tailed Tests		Two-Tailed Test
H_0: $\beta_1 = 0$	H_0: $\beta_1 = 0$	H_0: $\beta_1 = 0$
H_a: $\beta_1 < 0$	H_a: $\beta_1 > 0$	H_a: $\beta_1 \neq 0$

Test statistic: $t = \frac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}}$

Rejection region:

$t < -t_\alpha$	$t > t_\alpha$	$t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$

$df = n - 2$

P-value:

left tail area outside t	right tail area outside t	2*(tail area outside t)

Is reaction time truly linearly related to drug amount?

Example: In the drug reaction example, recall $\hat{\beta}_1 = 0.7$. Is the real β_1 significantly different from 0? (Use $\alpha = .05$.)

$H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}}$$

To calculate $s = \sqrt{MSE}$, note $\sum Y_i^2 = 1^2 + 1^2 + 2^2 + 2^2 + 4^2 = 26$

$$SS_{YY} = 26 - \frac{(10)^2}{5} = 26 - 20 = 6$$

$$SSE = SS_{YY} - \hat{\beta}_1 SS_{XY} = 6 - (0.7)(7) = 1.1$$

(0.7 and 7 are from previous calculations)

$$MSE = \frac{SSE}{n-2} = \frac{1.1}{3} = 0.3667$$

$$s = \sqrt{MSE} = \sqrt{.3667} = .606$$

Recall $SS_{xx} = 10$, so $t = \frac{0.7}{.606/\sqrt{10}} = \frac{0.7}{.19} = \underline{\underline{3.68}}$

We reject H_0 if $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$

$t_{\alpha/2} = t_{.025}$ (3 d.f.) $= 3.182$ (t-table)

$|t| = 3.68 > 3.182$, so we reject H_0. We conclude the true slope is not zero and that drug amount <u>is</u> a useful predictor of reaction time. (they <u>are</u> linearly related). P-value $\approx 2(.02) = .04$

A 100(1 – α)% Confidence Interval for the true slope β_1 is given by:

$$\hat{\beta}_1 \pm t_{\alpha/2}\left(\frac{s}{\sqrt{SS_{xx}}}\right)$$

where $t_{\alpha/2}$ is based on $n - 2$ d.f.

In our example, a 95% CI for β_1 is: $1-\alpha = .95 \Rightarrow \alpha = .05$

$\alpha/2 = .025$

$t_{.025} = 3.182$, for 3 d.f.

$$0.7 \pm (3.182)\left(\frac{.606}{\sqrt{10}}\right)$$

$$(0.09, 1.31)$$

With 95% confidence, for each one-percent increase in drug amount, the expected reaction time increases by between 0.09 seconds and 1.31 seconds.

Correlation

The scatterplot gives us a general idea about whether there is a linear relationship between two variables.

More precise: The coefficient of correlation (denoted *r*) is a numerical measure of the strength and direction of the linear relationship between two variables.

Formula for *r* (the correlation coefficient between two variables *X* and *Y*):

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

Most computer packages will also calculate the correlation coefficient.

Interpreting the correlation coefficient:

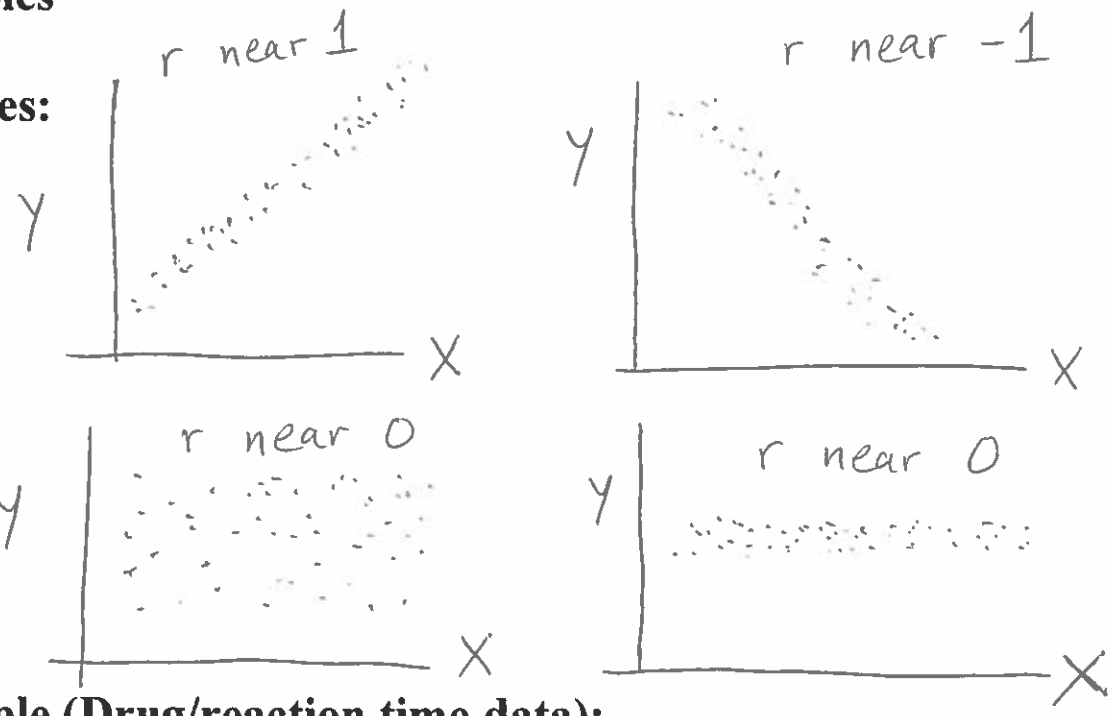
- **Positive *r* => The two variables are positively associated (large values of one variable correspond to large values of the other variable)**
- **Negative *r* => The two variables are negatively associated (large values of one variable correspond to small values of the other variable)**
- **$r = 0$ => No linear association between the two variables.**

Note: $-1 \leq r \leq 1$ always.

How far *r* is from 0 measures the *strength* of the linear relationship:

- ***r*** **nearly 1 => Strong positive relationship between the two variables**
- ***r*** **nearly -1 => Strong negative relationship between the two variables**
- ***r*** **near 0 => Weak linear relationship between the two variables**

Pictures:



Example (Drug/reaction time data):

Recall $SS_{xx} = 10$, $SS_{xy} = 7$, $SS_{yy} = 6$

$$r = \frac{7}{\sqrt{(10)(6)}} = \frac{7}{7.746} = 0.9037$$

Interpretation? For these five subjects, there is a strong, positive linear relationship between drug amount and reaction time.

Notes: (1) Correlation makes no distinction between predictor and response variables.
(2) Variables must be numerical to calculate *r*.

Examples: What would we expect the correlation to be if our two variables were:
(1) Work Experience & Salary? Positive

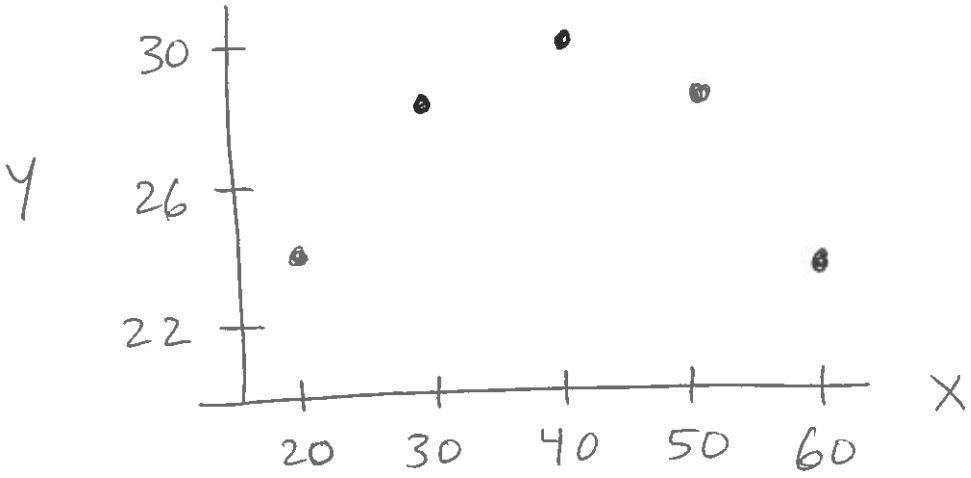
(2) Weight of a Car & Gas Mileage? Negative

Some Cautions

Example:

Speed of a car (*X*)	20	30	40	50	60
Mileage in mpg (Y)	24	28	30	28	24

Scatterplot of these data:



Calculation will show that *r* = 0 for these data.

Are the two variables related? Yes, but it is not a linear association. r does not measure curvilinear association between 2 variables.

Another caution: Correlation between two variables does not automatically imply that there is a cause-effect relationship between them.

Note: The population correlation coefficient between two variables is denoted ρ. To test H_0: $\rho = 0$, we simply use the equivalent test of H_0: $\beta_1 = 0$ in the SLR model. If this null hypothesis is rejected, we conclude there is a significant correlation between the two variables.

The square of the correlation coefficient is called the coefficient of determination, r^2.

Interpretation: r^2 represents the proportion of sample variability in Y that is explained by its linear relationship with X.

$$r^2 = 1 - \frac{SSE}{SS_{yy}}$$ **(r^2 always between 0 and 1)**

For the drug/reaction time example, $r^2 =$ $(.9037)^2 = .8167$

Interpretation: About 82% of the sample variation in reaction time can be explained by its linear relationship with drug amount.