

STATS 210P

Lecture 3

Sevan Koko Gulesserian

University of California, Irvine

Simple Linear Regression: Specifications

The simple linear regression model for the population is:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

where the ε 's are independent and ε follows a Normal distribution with mean/expectation 0 and variance σ_ε^2 (can also denote this as $\varepsilon \sim \text{Normal}(0, \sigma_\varepsilon^2)$).

σ_ε = standard deviation of the errors = standard deviation of the Y values at each X value.

- The above is meant to imply that $Y|X \sim N(\beta_0 + \beta_1 X, \sigma_\varepsilon^2)$.
- **Remember the $Y|X$ notation is meant to signify "Y given or conditional on X".**
- This is to say Y (given X) follows a normal distribution with expectation $\beta_0 + \beta_1 X$ (which we can think of as μ_Y) and variance σ_ε^2 .
- At the unit level this population model is: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where the ε_i 's are independent and $\varepsilon_i \sim \text{Normal}(0, \sigma_\varepsilon^2)$.

Simple Linear Regression: Assumptions

The simple linear regression model for the population is:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Note that this is just a general form of the population model so far (and we want to estimate it).
- The model holds for each individual observations, i .
 - $i = 1, 2, 3, \dots, n$. That is to say we have n many units.
- Therefore $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
 - $\varepsilon_i \stackrel{iid}{\sim} \text{Normal}(0, \sigma_\varepsilon^2)$
 - iid stands for identical independently distributed.
 - And the symbol \sim means "follows" ($z \sim \text{Normal}$ is to say z follows a Normal distribution).
- And so each i -th unit can have its own predicted value for Y as: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ (this is the estimated version of the population model).

Simple Linear Regression: Specifications

The simple linear regression model for the population is:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

where the ε 's are independent and $\varepsilon \sim \text{Normal}(0, \sigma_\varepsilon^2)$.

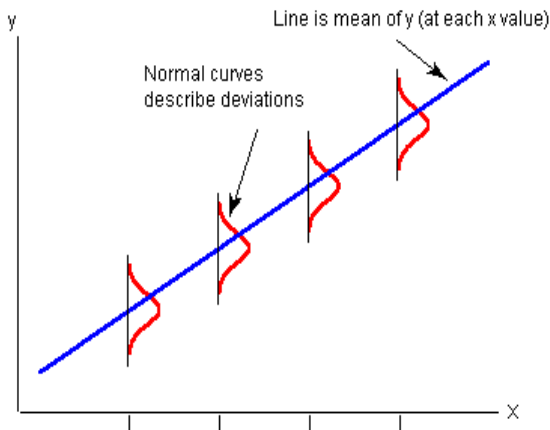
At the unit level this population model is: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$,
where the ε_i 's are independent and $\varepsilon_i \sim \text{Normal}(0, \sigma_\varepsilon^2)$.

- Note that the expectation of Y is: $E(Y) = \beta_0 + \beta_1 X$ (or $E(Y_i) = \beta_0 + \beta_1 X_i$).
- So when we have our estimated regression equation, what we are doing is estimating the expectation (or mean or average) of the Y at a certain value of X :

$$\widehat{E(Y)} = \hat{\beta}_0 + \hat{\beta}_1 X \text{ (can think of this as } \hat{\mu}_Y \text{)}.$$

Simple Linear Regression: Assumptions

A visualization of this model:



The blue line is the linear regression line (the estimated expected value of Y at that value of X). And the red curve is the normal distribution curve (with variance σ_{ϵ}^2).

Simple Linear Regression: Fitting the Model

Estimating σ_ϵ

- We use the residuals to estimate σ_ϵ .
- Call the estimate of σ_ϵ the *regression standard error* (or *residual standard error*).

$$\begin{aligned}s &= \hat{\sigma}_\epsilon = \sqrt{\frac{\text{Sum of Squared Residuals}}{n-2}} \\&= \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\text{MSE}}\end{aligned}$$

where MSE=Mean Squared Error.

- Note that n is always given, it is the sample size (number of observations in our dataset).
- Can compute SSE when given s , and vice-versa.
- **This formula is only for simple linear regression (a single X covariate, thus only β_0 and β_1 in the equation).**

Simple Linear Regression: Fitting the Model

The take away is that $\hat{\sigma}_\varepsilon$ (our estimate of σ_ε) is called the "residual standard error" given in the R output.

With the preceding, can approximate the distribution of Y for a given value of X as follows:

$$Y|X \sim N(\hat{\beta}_0 + \hat{\beta}_1 X, \hat{\sigma}_\varepsilon^2)$$

* The notation \sim means to approximately follow.

Simple Linear Regression: Fitting the Model

Example: In the highway sign distance example.

Let us fit a simple linear regression model where the response variable is how far away someone can read the sign (in feet) and the explanatory variable is age in years.

Say you import the dataset and call it "HighwaySign" (note our dataset does not have names for the variables).

To add names and then run the model, the following lines need to be run.

```
names(HighwaySign) = c("Age", "Distance")  
model = lm(Distance~Age, data=HighwaySign)  
summary(model)
```


Simple Linear Regression: Fitting the Model

Output from R's summary(model) call:

Residuals:

Min	1Q	Median	3Q	Max
-78.231	-41.710	7.646	33.552	108.831

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	576.6819	23.4709	24.570	< 2e-16 ***
Age	-3.0068	0.4243	-7.086	1.04e-07 ***

Residual standard error: 49.76 on 28 degrees of freedom

Multiple R-squared: 0.642, Adjusted R-squared: 0.6292

F-statistic: 50.21 on 1 and 28 DF, p-value: 1.041e-07

Simple Linear Regression: Fitting the Model

- The fitted regression equation is $\hat{Y} = 577 - 3 * X$ (rounding to the nearest whole number).
- Note that the residual standard error, s , is equal to 49.76.
- This is to say that for someone who is 30 years old, the mean of the approximate distribution of Y is $577 - 3 * 30 = 487$.
- The variance of the approximate distribution of Y is 49.76^2 (since the standard deviation is 49.76).
- Therefore, for all those who are 30 years old, the distance they can read a highway sign is approximately distributed as $\text{Normal}(487, 50^2)$.

Simple Linear Regression: Assumptions

The following assumptions are needed to confirm the validity of using the simple linear regression to obtain inference of the association between two quantitative variables.

- 1. Linearity: There is a linear relationship between the X and Y variables. Fitting a straight line through the data points is appropriate.
- 2. The variance (or standard deviation) of the Y values is constance for all values of X in the range of the data.
- 3. Independence: The errors are independent of each other. That is to say knowing the value of one does not help with knowing the value of the others.
- 4. The errors are normally distributed.
- 5. If we want to extend the results to the population, need the sample to be randomly selected from the population.

Simple Linear Regression

We will come to checking the assumptions soon.

For now, let us see what other quantities are estimated in the simple linear regression model besides β_0 and β_1

- Already noted that σ_ε is estimated.
 - This is known as the regression standard error or residual standard error (**In the R output, this is called the "residual standard error", and the degrees of freedom is equal to $n-2$**).
 - It can be used to estimate how much the response variable Y varies at a given value of X .

Simple Linear Regression: Fitting the Model

Example of fitting a model to $Y = \text{MCAT scores}$ and $X = \text{GPA}$ from the Medical school admission data.

```
lm(formula = MCAT ~ GPA, data = mcat)
```

Residuals:

Min	1Q	Median	3Q	Max
-11.4148	-2.5168	-0.1519	2.6653	8.6616

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.923	6.922	0.567	0.573
GPA	9.104	1.942	4.688	1.97e-05 ***

Residual standard error: 4.088 on 53 degrees of freedom
Multiple R-squared: 0.2931, Adjusted R-squared: 0.2798
F-statistic: 21.98 on 1 and 53 DF, p-value: 1.969e-05

Simple Linear Regression

- Note the residual standard error is 4.088
 - Degrees of freedom is meant to quantify the $n - 2$ in the formula of MSE shown in a previous slide.
 - 53 degrees of freedom means $n = 53 + 2 = 55$ observations are in the dataset.
 - At any given value of $X = \text{GPA}$, we say the estimated variance of the normal distribution of Y is 4.088^2 .

Simple Linear Regression

Coefficient of determination.

- Now, another quantity of interest that is estimated in a simple linear regression is presented.
- The *coefficient of determination* is denoted as R^2 (R-squared).
- It estimates how much of the variation in Y is explained or predicted by X .
- It is a number between 0 and 1 that is commonly converted to a percentage (0 to 100%).
- For now in the R output, we will look at the multiple R-squared value for R^2 and will cover adjusted R^2 when we begin the multiple linear regression material.

Simple Linear Regression

Before we show R^2 , need to define a few useful quantities.

- SSR is defined to be the sum of square regression.
 - $SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$
 - \bar{Y} is the sample mean of all the Y 's in the data (add up the n many Y_i 's and divide that sum by n).
- SSE is defined to be the sum of square errors.
 - $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$
- SSTO is defined to be the sum of square total.
 - $SSTO = \sum_{i=1}^n (Y_i - \bar{Y})^2$
- A result we will use (but not show) is that $SSTO = SSE + SSR$.

Simple Linear Regression

- $R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$.
- This can be noted since $SSTO = SSE + SSR$.
- R^2 can be viewed as comparing the variation of the predictions (\hat{Y}) about the mean of Y (\bar{Y}) to the variation of the true values (Y) about the mean of Y .
- Can be interpreted as the variation of Y explained or predicted by X (or the variate of Y explained by our model).
- Will be a number between 0 and 1.
- Example: An R^2 of 0.50 means that 50% of the variation of Y is explained by X .

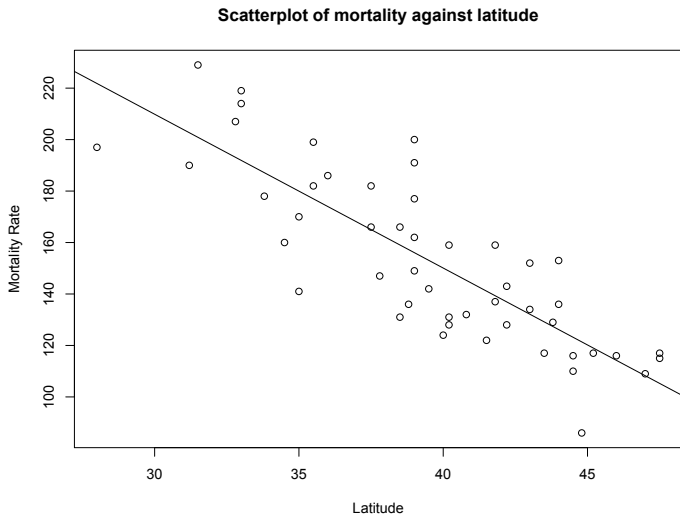
Simple Linear Regression

Example: the skincancer.txt dataset contains the mortality rates due to skin cancer from 49 states. The goal is to assess if the latitude of the state predicts (or explains) the mortality rate of skin cancer.

The closer the latitude is to 0, the closer you are to the equator. This implies the lower the latitude, to higher sun exposure you experience.

- Y = mortality rate (per 1,000,000 per year).
- X = latitude.

Simple Linear Regression



Simple Linear Regression: Fitting the Model

Fitting the model using the data, the output is:

```
lm(formula = Mort ~ Lat, data = skincancer)
```

Residuals:

Min	1Q	Median	3Q	Max
-38.972	-13.185	0.972	12.006	43.938

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	389.1894	23.8123	16.34	< 2e-16 ***
Lat	-5.9776	0.5984	-9.99	3.31e-13 ***

Residual standard error: 19.12 on 47 degrees of freedom

Multiple R-squared: 0.6798, Adjusted R-squared: 0.673

F-statistic: 99.8 on 1 and 47 DF, p-value: 3.309e-13

Simple Linear Regression

Example: the skincancer.txt dataset contains the mortality rates due to skin cancer from all 49 states. The goal is to assess if the latitude of the state predicts (or explains) the mortality rate of skin cancer.

- The regression equation is $\hat{Y}_i = 389.18 - 5.97X_i$.
- The residual standard error, $\hat{\sigma}_\varepsilon$ is 19.12.
- Degrees of freedom, $n - 2$, is equal to 47. Therefore $n = 47 + 2 = 49$ observations.

For now in the R output, we will look at the multiple R-squared value for R^2 and will cover adjusted R^2 when we begin the multiple linear regression material.

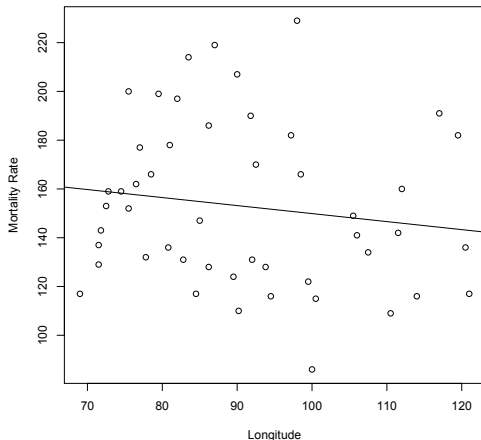
- $R^2 = 0.6798$.
- This to say that 68% of the variation in Y is explained by X .
 - 68% of the variation in mortality is explained by the latitude.

Simple Linear Regression

Using the skin cancer data again, let's look at the case where $X =$ Longitude (instead of latitude).

(Longitude is a measure of the East-West location of a place relative to London, England.)

Scatterplot of mortality against latitude



Simple Linear Regression: Fitting the Model

Fitting the model using the data, the output is:

```
lm(formula = Mort ~ Long, data = skincancer)
```

Residuals:

Min	1Q	Median	3Q	Max
-63.898	-25.995	-5.952	21.856	78.444

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	182.7696	29.8893	6.115	1.8e-07 ***
Long	-0.3287	0.3245	-1.013	0.316

Residual standard error: 33.42 on 47 degrees of freedom

Multiple R-squared: 0.02137, Adjusted R-squared: 0.0005491

F-statistic: 1.026 on 1 and 47 DF, p-value: 0.3162

Simple Linear Regression

- Can see that longitude is not nearly as good a predictor as latitude.
- $R^2 = 0.02$.
- This to say that 2% of the variation in Y is explained by X .
 - 2% of the variation in mortality is explained by the longitude.

Quick review using the skin cancer dataset.

- Hospital records were used to record the mortality rate for each state.
- This is an observational study, since subjects were not randomized to live in a state.
- Can we say that latitude causes mortality rates to increase?
- Any possible confounders? (Think lifestyle differences)

We have defined R^2 , the coefficient of determination, and what it is useful for.

We now look at what is called the *correlation coefficient*, ρ (which is estimated using R).

- R is an estimate of the linear correlation between Y and X (the true value being ρ).
- As R^2 was between 0 and 1, R is between -1 and 1.
- ρ is the true value of the linear correlation coefficient between X and Y , while R is the estimate of ρ .

Simple Linear Regression

- If R is negative, then Y and X are negatively related, and similarly for positive values of R (X and Y are positively related).
- The closer R is to -1 or 1 , the stronger the relationship is between Y and X .
- Given R^2 , can calculate $|R| = \sqrt{R^2}$.
 - But need to determine if it is $\sqrt{R^2}$ or $-\sqrt{R^2}$.
 - If association is negative (i.e. β_1 is negative, or looking at scatterplot and seeing negative trend), then $R = -\sqrt{R^2}$.
 - If association is positive (i.e. β_1 is positive, or looking at scatterplot and seeing positive trend), then $R = \sqrt{R^2}$.

Simple Linear Regression

- ρ was defined to be $\rho = \frac{\text{cov}(X,Y)}{\sigma_x \sigma_y}$ (this is the population formula).
- $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$ was the covariance between X and Y .
- σ_x and σ_y were the standard deviation of X and Y respectively.
- The sample based estimate is (for a sample size n):

$$R = \frac{\sum_i X_i Y_i - (\sum_i X_i)(\sum_i Y_i)/n}{\sqrt{(\sum_i X_i^2 - \frac{(\sum_i X_i)^2}{n})(\sum_i Y_i^2 - \frac{(\sum_i Y_i)^2}{n})}}.$$

- It can be show that this R when squared (R^2) is equal to $\frac{SSR}{SSTO}$ from slide 19.

Simple Linear Regression

Example: Medical school data. $Y = \text{MCAT}$ and $X = \text{GPA}$

- $R^2 = 0.29$. Thus 29% of the variation of MCAT score is explained by GPA.
- Note that slope of GPA (9.10) is positive.
- Therefore $R = \sqrt{0.29} = 0.53$.

Example: Skin cancer data. $Y = \text{Mortality}$ and $X = \text{Latitude}$.

- $R^2 = 0.67$. 67% of the variation of mortality is explained by latitude.
- Slope on latitude is negative, -5.97.
- $R = -\sqrt{0.67} = -0.81$.

Simple Linear Regression

- What is useful about R is that it is unit less (it does not matter the units used to measure X or Y).
- Example: The correlation between (X) your weight in pounds and (Y) your height in inches is the same as the correlation between (X) your weight in kilograms and (Y) your height in inches.
- The idea is that if you linearly transform one of the variables, or both, then the correlation will still capture the same relationship between the variables.
- To linearly transform means to multiply by a number and/or add a number to the variable (so new Y is $Y = a + b\tilde{Y}$, where \tilde{Y} is the old Y , and a and b are constants).
- Knowing the value of R will quantify the strength of the linear association between Y and X , and also quantify the direction of the association, negative or positive.
- Note that in the simple linear regression (with X and Y being quantitative) case that the correlation (R or ρ) between Y and X is the same as the correlation between X and Y .
 - That is $\text{corr}(Y, X) = \text{corr}(X, Y)$.

Simple Linear Regression: Review of Hypothesis Test

Overview of hypothesis testing.

- In a hypothesis test we have the null (H_0) and the alternative (H_a) hypothesis.
- We try to either reject or fail to reject the null hypothesis.
- What researchers want inference on is H_A , the alternative.
- The alternative can be of the form $\mu \neq \mu_0$, $\mu < \mu_0$, or $\mu > \mu_0$ (for some value of μ_0 that you specify).
- The null for a parameter μ will be the complement of the alternative (so $\mu = \mu_0$, $\mu \geq \mu_0$ or $\mu \leq \mu_0$ respectively for the alternatives in the previous bullet point).

Simple Linear Regression: Review of Hypothesis Test

In summary, the null and alternative hypothesis scenario will be one of the following forms:

1. $H_0: \mu = \mu_0$ versus $H_a: \mu \neq \mu_0$ (two-sided)
2. $H_0: \mu \geq \mu_0$ versus $H_a: \mu < \mu_0$ (one-sided)
3. $H_0: \mu \leq \mu_0$ versus $H_a: \mu > \mu_0$ (one-sided)

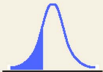
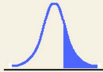
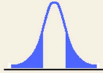
Simple Linear Regression: Review of Hypothesis Test

Once the null and alternative hypothesis are set and the test statistic t is calculated, obtain the p -value as follows:

For H_a ***less than***, the p -value is the area below t , even if t is positive.

For H_a ***greater than***, the p -value is the area above t , even if t is negative.

For H_a ***two-sided***, p -value is $2 \times$ area above $|t|$.

Statement of H_a	p -Value Area	t -Curve Region
$\mu < \mu_0$ (less than)	Area to the left of t (even if $t > 0$)	
$\mu > \mu_0$ (greater than)	Area to the right of t (even if $t < 0$)	
$\mu \neq \mu_0$ (not equal)	$2 \times$ area to the right of $ t $	

Simple Linear Regression: Review of Hypothesis Test

- Given a specific null hypothesis, we can construct our test statistic.
 - Example is a standardized estimate, such as a the t statistic for μ test, $t = \frac{\bar{x} - \mu_0}{s_x}$.
 - \bar{x} is the sample mean and s_x is sample mean standard deviation (this was $\frac{s}{\sqrt{n}}$ where s is the sample standard deviation).
- Given a test statistic, such as t , compute the p-value based on the alternative hypothesis.
 - P-value was the probability of observing a test statistic as extreme as the one you calculated, under the scenario of the null hypothesis.
 - Can also think of p-value as the probability of observing data (sample) as you did given the scenario of the null hypothesis.

Simple Linear Regression: Review of Hypothesis Test

- With a specific null hypothesis and alternative hypothesis, we can make a conclusion based on the p-value.
- Given the significance level of the test, α which is between 0 and 1 (this is given to you), compare p-value to α .
 - If p-value is less than α , reject the null and conclude evidence for the alternative. We have statistically significant results.
 - If p-value is greater than α , fail to reject the null and don't conclude evidence for the alternative.

Simple Linear Regression: Review of Hypothesis Test

Example: Review of one population mean test (one sample t-test).

A study where we want to test if the mean internal temperature for an adult is 98.6 fahrenheit. Researchers believe the average internal temperature is lower than 98.6 f. Say a 100 adults are sampled and have their internal temperature recorded.

Let μ denote the average internal temperature for an adult.

- The null is $H_0 : \mu \geq 98.6$ and alternative is $H_a : \mu < 98.6$
- The average of the 100 peoples temperature is 98.2 f.
- The sample mean standard deviation is 0.12.
- $t^* = \frac{\bar{x} - \mu_0}{s_x} = \frac{98.2 - 98.6}{0.12} = -3.33$
- P-value is area below -3.33 using a t distribution with $n-1=99$ degrees of freedom.
- We can approximate this p-value by using the standard normal distribution (instead of a t distribution). In R this will be `pnorm(-3.33)=0.0004`.
- Say $\alpha = 0.05$. Since p-value is about 0, we reject the null hypothesis and conclude evidence for the alternative.

Can summarize a hypothesis test into steps.

- 1. Determine the null and alternative hypothesis.
- 2. Verify if conditions of the test are satisfied.
- 3. Compute the test statistic.
- 4. Find the p-value.
- 5. See if results are statistically significant based on p-value, and make a conclusion in context of the problem.

Simple Linear Regression: Testing ρ

- In certain cases of simple linear regression models, it is not necessary that a certain variable is the response and the other is the explanatory.
- That is to say the relationship under study can go either way, from X to Y or Y to X (since $\text{corr}(X,Y)=\text{corr}(Y,X)$).
- In the example of the relationship between latitude and mortality, clearly the relationship that we are interested in is where X is latitude and Y is mortality.
- But for example wanting to study the relationship between the ages of married couples.
 - Want to study how the age of a husband is associated with the age of a wife.
 - Similarly, can study how the age of the wife is associated with the age of a husband.

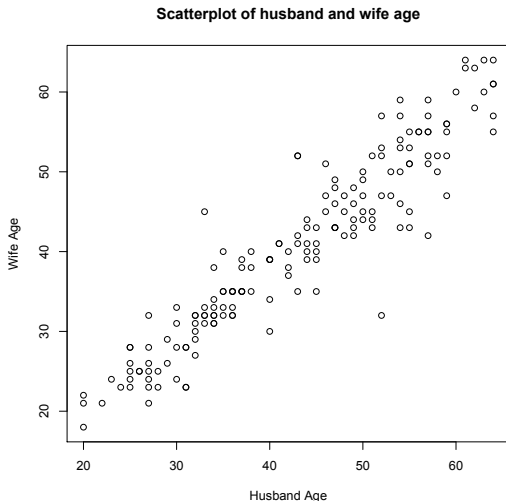
Simple Linear Regression: Testing ρ

Testing the correlation coefficient.

- Can do a formal test about ρ compared to a value you specify, ρ_0 (usually 0, thus we have $H_0 : \rho = 0$ vs $H_a : \rho \neq 0$).
- The estimate that will be used is R , and the population parameter is ρ .
- The test statistic is created as follows: $t^* = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$.
- This test statistic, t^* , follows a t distribution with d.f. = $n - 2$ (degrees of freedom).

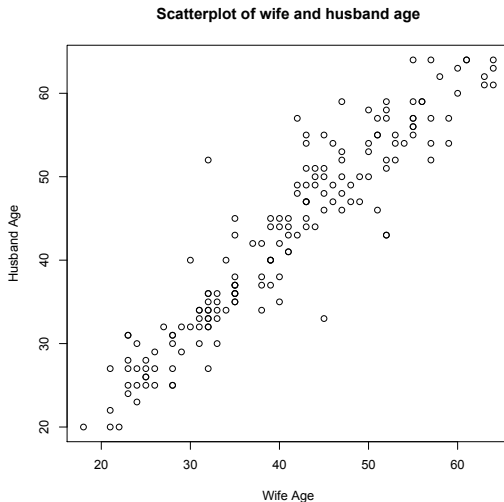
Simple Linear Regression: Testing ρ

In this example of couple age, it is not necessary that Y is a wife's age and X is the husband's age.



Simple Linear Regression: Testing ρ

Now if Y is husband's age and X is wife's age.



Simple Linear Regression: Fitting the Model

Fitting the model using the couples data, the output is:

```
lm(formula = HAge ~ WAge, data = couple)
```

Residuals:

Min	1Q	Median	3Q	Max
-14.0984	-2.2298	-0.5253	2.0986	17.4747

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.5760	1.1639	3.072	0.00248	**
WAge	0.9672	0.0276	35.044	< 2e-16	***

Residual standard error: 4.081 on 167 degrees of freedom

Multiple R-squared: 0.8803, Adjusted R-squared: 0.8796

F-statistic: 1228 on 1 and 167 DF, p-value: < 2.2e-16

Simple Linear Regression: Testing ρ

- $R^2 = 0.8803$ and $\hat{\beta}_1 = 0.96$.
- Therefore $R = \sqrt{0.8803} = 0.9382$.
- Can now test things of the form $\rho = 0$. Say $H_0 : \rho = 0$ and $H_a : \rho \neq 0$.
- Create a test statistic as follows: $t^* = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$.
- Here $t^* = 35.05$.
- The alternative is two sided. Two times the area above $|35.05| = 35.05$ using a t-distribution with 167 degrees of freedom (or approximating it using a standard normal distribution) is essentially 0 (we can get this using R or by using a table).
- Assuming a 5% significance level, we reject the null, and conclude evidence that $\rho \neq 0$ (the two variables are correlated).

Simple Linear Regression: Obtaining p-values

Getting p-value based on a test statistic t^* with $n-k$ degrees of freedom (our examples had $k=2$ and $n=169$).

- To get the area below the test statistic using a t-distribution, we would run the following in R: `pt(t^* , $n-k$)`.
- Thus to get the area above you would do: `1-pt(t^* , $n-k$)`.
- And so the two sided p-value would be (where $|t^*|$ is the absolute value of the test statistic): `2(1-pt($|t^*|$, $n-k$))`.
- To approximate the previous p-values using the standard normal distribution, we would do the following to get the area below: `pnorm(t^*)`.
- The idea is based on the result where as the degrees of freedom gets larger and larger (that is to say n gets larger) the t-distribution will converge to the standard normal distribution.

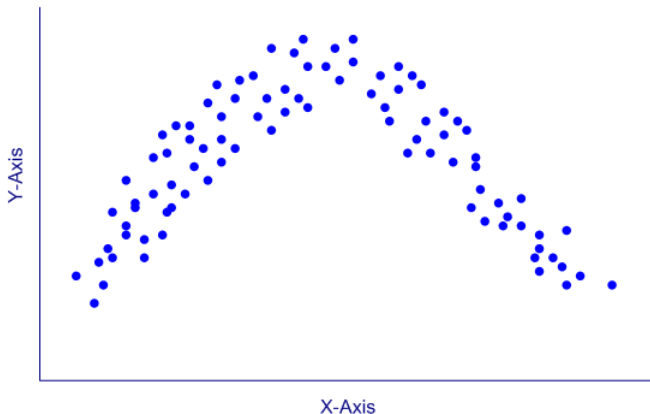
Simple Linear Regression: Correlation Coefficient

Note that the correlation coefficient mentioned earlier is the *linear* correlation coefficient.

Therefore it is only suitable for when there is a linear relationship between the two variables.

Simple Linear Regression: Correlation Coefficient

Example of a non-linear relationship which would lead to a correlation coefficient that is close to 0.



Simple Linear Regression: Correlation Coefficient

Example of a non-linear relationship which would lead to a correlation coefficient that is lower than what would expect based on the association between X and Y .

