

Inference in linear regression: the 'popular' p-value

Week 2

What are the goals for today?

- We have learned how to get the regression equation, and verify assumptions of linear regression
 - How do we evaluate the effectiveness of the model?
 - How do we know that the relationship is significant?
 - How much of the variability in the response variable can be explained by its relationship to the predictor?
 - Make predictions about population based on our sample data – inference!

Going back to our first example

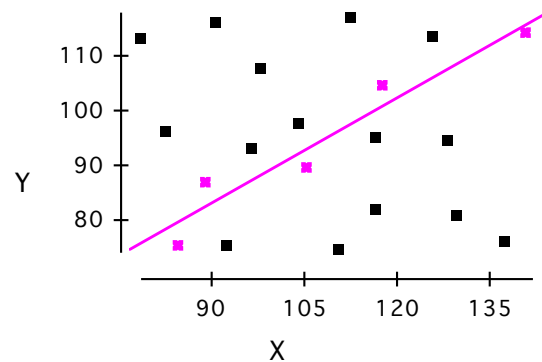
- Consider our earlier example: Medical researchers have noted that adolescent females are more likely to deliver low-birthweight babies than are 'adult' females. Because LBW babies tend to have higher mortality rates, studies have been conducted to examine the relationship between birthweight and the mother's age.

Observation	1	2	3	4	5	6	7	8	9	10
Maternal Age	15	17	18	15	16	19	17	16	18	19
Birthweight (g)	2289	3393	3271	2648	2897	3327	2970	2535	3138	3573

- We found the fitted regression equation:
 $\text{Weight}^{\wedge} = -1163.45 + 245.15\text{age}$
 - Weight should have a 'hat' \wedge because its fitted/predicted
 - Versus when talking about the population:
 - $Y = \beta_0 + \beta_1 x_1 + \varepsilon$

How can we assess the model?

- $\beta_1^{\wedge} = 245.15$, For each additional age of the mother, birthweight increases on average by 245.15 grams.
- How do we know if this is meaningful? Maybe just a fluke? Maybe not a large enough increase to be meaningful, useful or significant.



- Need a way to test this. We can do this by asking the question, is β_1 different from zero.
- To help answer this, we will need an estimate of the variability of the slope.

Testing the slope

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	-1163.450	783.138		-1.486	.176
age	245.150	45.908	.884	5.340	.001

a. Dependent Variable: birthweight

- Hypothesis test:

$H_0: \beta_1 = 0$

Null hypothesis, what is true if no relationship

$H_a: \beta_1 \neq 0$

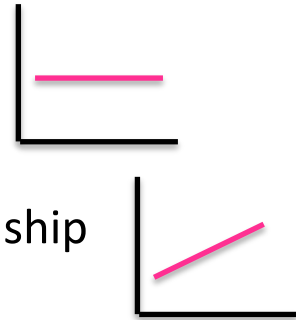
Alternative hypothesis, true if there is a relationship

- Test statistic:

With n-2 degrees of freedom

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

- If the conditions (assumptions) for linear regression are met, this test statistic follows a t-distribution, so we can put a pvalue to it



Testing the slope

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	-1163.450	783.138		-1.486	.176
age	245.150	45.908	.884	5.340	.001

a. Dependent Variable: birthweight

- Hypothesis test:

$H_0: \beta_1 = 0$ Null hypothesis, what is true if no relationship

$H_a: \beta_1 \neq 0$ Alternative hypothesis, true if there is a relationship

- Test statistic:

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

With n-2 degrees of freedom

t=

t= 245.15/45.908

t= 5.34

Can look this t value up, with n-2, or 8 degrees of freedom, to get a pvalue

What exactly is a pvalue?

- If we assume the null hypothesis is true, the p-value is the probability of getting a test statistic as large as you have in your sample.
 - So if we assume that there is no relationship between maternal age and birthweight, the probability of getting a slope of 245.15 is .001.
 - Not very likely!!!
 - Quiz question: The pvalue for the slope tells us how confident we are that the slope we found represents the true slope in the population of interest. 60% correct
- How small does it need to be?
 - Commonly people use .05 as a threshold, or α
- If it is small, we can REJECT the null hypothesis!
 - Never accept the alternative!!!
 - Never accept the null hypothesis either, we can only reject or fail to reject! (46% correct)

Its significant, so what?

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	-1163.450	783.138		-1.486	.176	-2969.369	642.469
Age	245.150	45.908	.884	5.340	.001	139.285	351.015

a. Dependent Variable: Birthweight

- If our goal is to be able to answers questions about the population
 - how close is our estimate to the 'real' slope value?
- What would we get if we took a different sample from our population?
- We can create a confidence interval to help answer that question

Its significant, so what?

Coefficients ^a							
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	-1163.450	783.138		-1.486	.176	-2969.369	642.469
Age	245.150	45.908	.884	5.340	.001	139.285	351.015

a. Dependent Variable: Birthweight

- $\text{Weight}^{\wedge} = -1163.45 + 245.15\text{age}$
- Confidence interval has the form: $\hat{\beta}_1 \pm t^* \cdot SE(\hat{\beta}_1)$

where t^* is the critical value for t needed to feel a certain level of confidence (95%...), with $n-2$ degrees of freedom. For $df=8$, $t^*=2.306$

$$= 245.15 \pm 2.306(45.908)$$

$$= 245.15 \pm 105.86 = (139.29, 351.01)$$

Its significant, so what?

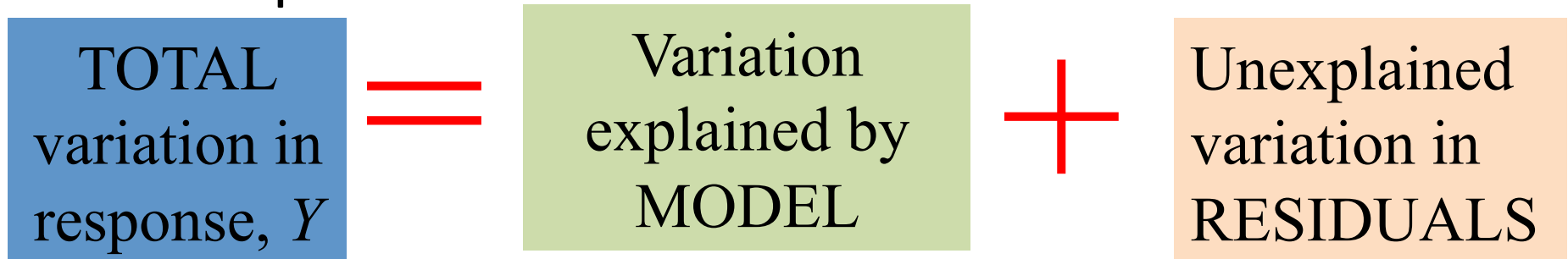
Coefficients ^a							
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	-1163.450	783.138		-1.486	.176	-2969.369	642.469
Age	245.150	45.908	.884	5.340	.001	139.285	351.015

a. Dependent Variable: Birthweight

- What does this interval tell us?
- Based on the sample data, we are 95% confident that the true average increase in the weight of a baby associated with a one-year increase in age of the mother is between 139.29 and 351.01 g.
- What does it mean to be 95% confident?
 - If we would collect 100 samples from the same population and calculated the CI, the true population parameter would be within the interval 95 times.
- Now we have assessed the significance of the slope, what about the overall model?

Assess the effectiveness of a model: ANOVA

- Measure how much of the variability in the response variable is explained by the predictions of the fitted model, and how much is left unexplained.

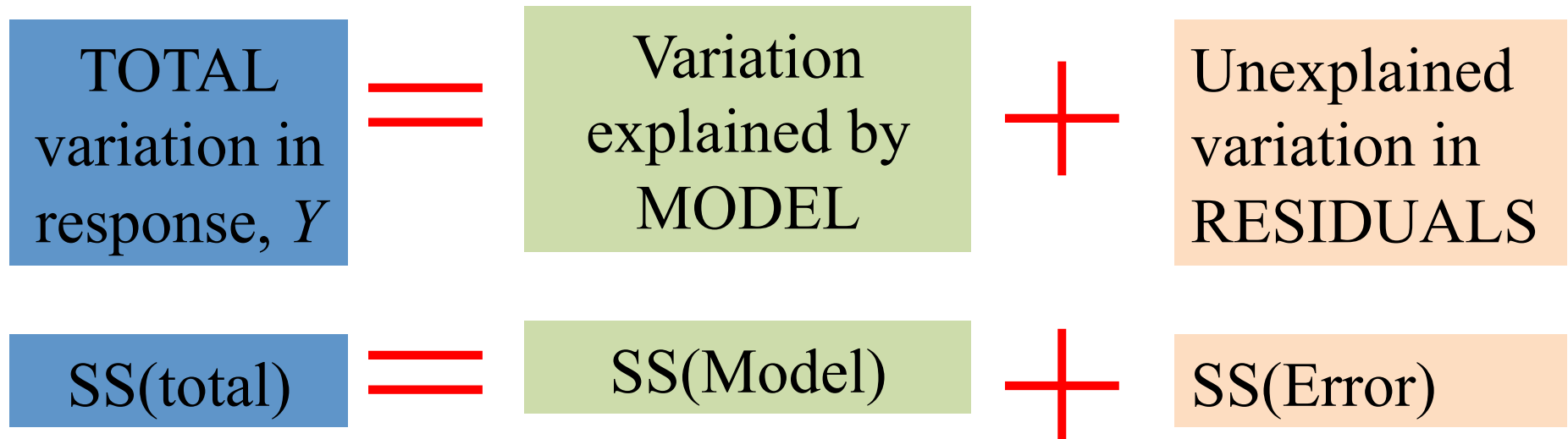


A diagram illustrating the ANOVA equation. It consists of three colored boxes connected by mathematical symbols. The first box is blue and contains the text 'TOTAL variation in response, Y'. To its right is a red equals sign. The second box is green and contains the text 'Variation explained by MODEL'. To its right is a red plus sign. The third box is orange and contains the text 'Unexplained variation in RESIDUALS'.

$$\text{TOTAL variation in response, } Y = \text{Variation explained by MODEL} + \text{Unexplained variation in RESIDUALS}$$

- Does the model explain a significant amount of the total variability?
- Quiz question: The variability due to error (SSE) is always smaller than the variation explained by the model (SSModel). 76% correct

Assess the effectiveness of a model: ANOVA



ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1201970.450	1	1201970.450	28.515	.001 ^b
	Residual	337212.450	8	42151.556		
	Total	1539182.900	9			

a. Dependent Variable: Birthweight

b. Predictors: (Constant), Age

Idea will become our test statistic

Test the effectiveness of the overall model

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

- In simple linear regression, this is essentially the same as testing the slope. It will make more sense to do in multiple linear regression
- The test statistic is essentially:
$$\frac{\text{The variability explained by the model}}{\text{The variability unexplained by the model}}$$
- Follows an F-distribution

Assess the overall model

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

- What is the pvalue?
 - .001
 - Decision:
 - Since p is $< .05$ and $F = 28.515$, reject the null hypothesis
 - Conclusion:
 - The data indicates that there is a linear relationship between the mother's age and the baby's birthweight.

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1201970.450	1	1201970.450	28.515	.001 ^b
	Residual	337212.450	8	42151.556		
	Total	1539182.900	9			

a. Dependent Variable: Birthweight

b. Predictors: (Constant), Age

Guideline for ‘assessing’ something

$$H_0: \beta_1 = 0$$

Hypothesis

$$H_a: \beta_1 \neq 0$$

– Decision:

Decision

- Since p is <.05 and F=28.515, reject the null hypothesis

– Conclusion:

Conclusion

- The data indicates that there is a linear relationship between the mother’s age and the baby’s birthweight.

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1201970.450	1	1201970.450	28.515	.001 ^b
	Residual	337212.450	8	42151.556		
	Total	1539182.900	9			

a. Dependent Variable: Birthweight

b. Predictors: (Constant), Age

Required SPSS output

One more way to assess the model!

- ANOVA:
$$\frac{\text{The variability explained by the model}}{\text{The variability unexplained by the model}}$$
- Coefficient of determination (r^2):
$$\frac{\text{The variability explained by the model}}{\text{The total variability in } y}$$
- Why called r^2 ?
 - For simple linear regression, it is the correlation coefficient squared.
- r^2 is the fraction of variability explained by the model, so its often converted to a percentage
- Quiz question: A regression equation was fit to a set of data, and the model was found to predict 60% of the variability in the response variable. That means that the correlation between the two variables is .6. 73% correct

Coefficient of determination

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.884 ^a	.781	.754	205.30844

a. Predictors: (Constant), age

b. Dependent Variable: birthweight

- $r^2 = .781$
- The data suggests that 78.1% of the variation in birthweights can be explained by the model, or the mother's age

Why 3 measures to assess the model?

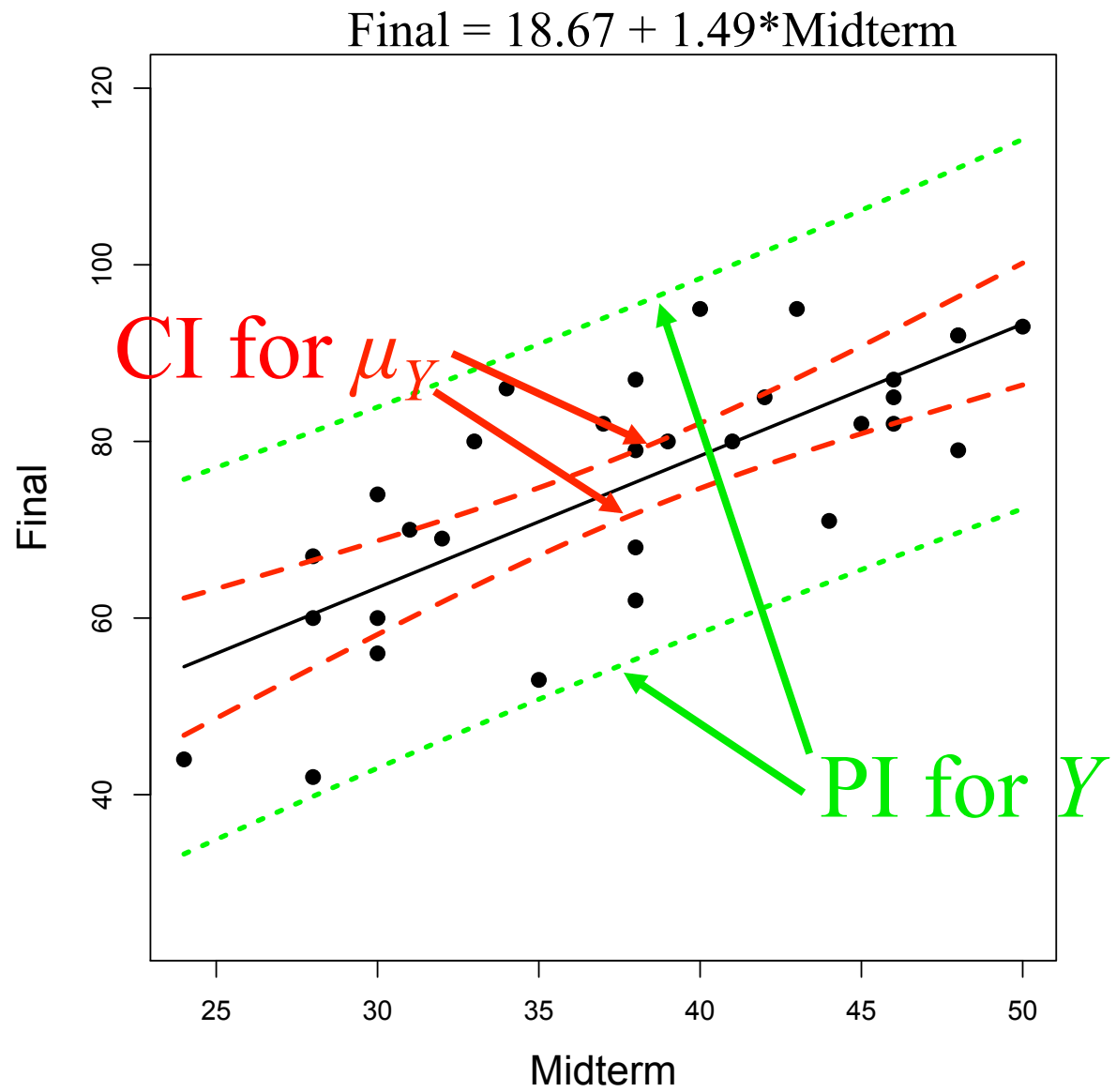
- What 3 do we have?
 - T-test of slope, ANOVA (F-test) of model, and coefficient of determination
- They are all the same in simple linear regression – always agree on effectiveness of model (75% correct)
- In multiple regression, they don't always have to agree – emphasize different aspects of the model's effectiveness
- Now that we have learned more ways to assess the model, let's learn some more about using the model!

Intervals for predictions

- One of the most common reasons to fit a model to data is to predict the response for a value of the explanatory variable.
 - How much is a baby expected to weigh if the mother is 13 years old?
 - We saw how to plug that into the prediction equation and get a predicted \hat{Y} .
- Don't want to know just a value, but also how confident or accurate our prediction is.

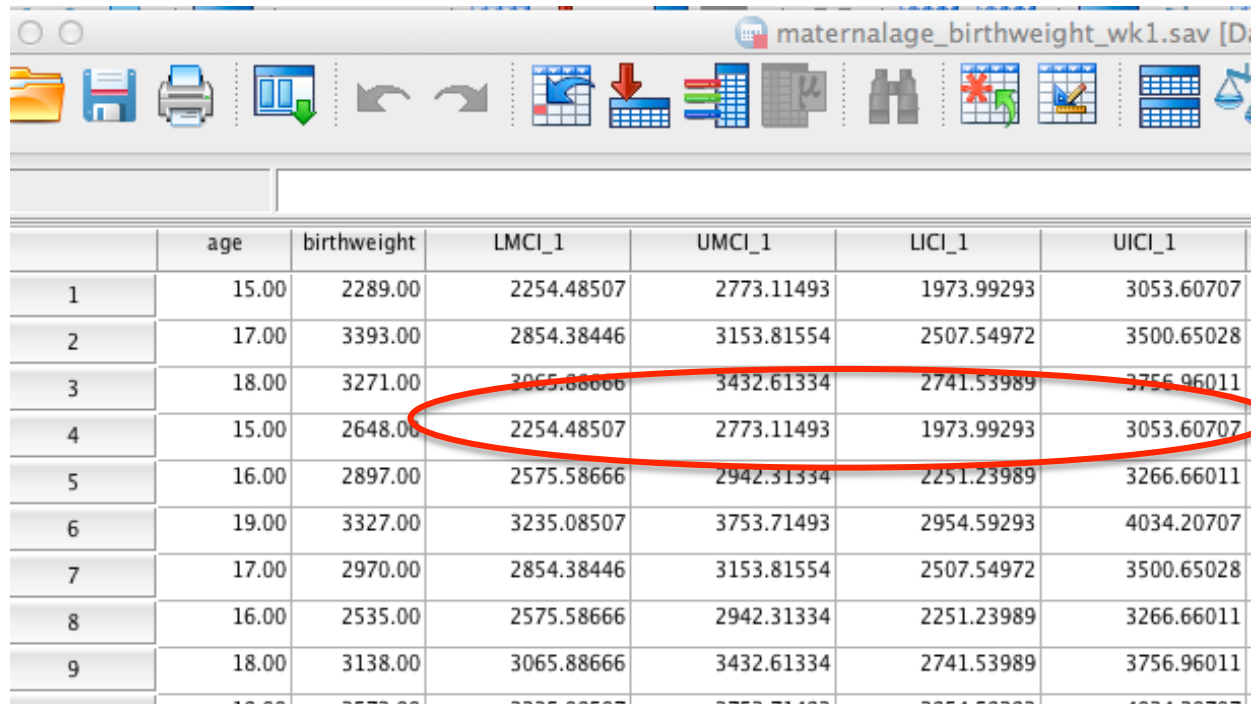
Intervals for predictions

- An interval can provide this measure of accuracy.
- Two kinds (65% correct): Do we want to make a prediction about the mean response for a particular value of x , or do we want to know the predicted Y for an individual case?
- Mean response: Confidence interval
 - What is the average predicted birthweight for mothers 15 years old
- Individual response: prediction interval
 - What is the predicted birthweight for a mother 15 years old
- We can be 95% confident for both
- Harder to predict value for an individual than to predict a mean response
 - So prediction interval larger!



Intervals for prediction in SPSS

- The predicted weight of a baby with a mother that is 15 years old is 2513.8, but how confident are we in that?
- Mean response: Confidence interval (LMCI, UMCI)
 - We can be 95% confident that the average birthweight for mothers 15 years old will be between 2254.5 and 2773.1g
- Individual response: prediction interval (LICI, UICI)
 - We can be 95% confident that the birthweight for a mother 15 years old will be between 1974.0 and 3053.6g.



maternalage_birthweight_wk1.sav [D:

	age	birthweight	LMCI_1	UMCI_1	LICI_1	UICI_1
1	15.00	2289.00	2254.48507	2773.11493	1973.99293	3053.60707
2	17.00	3393.00	2854.38446	3153.81554	2507.54972	3500.65028
3	18.00	3271.00	3065.88666	3432.61334	2741.53989	3756.96011
4	15.00	2648.00	2254.48507	2773.11493	1973.99293	3053.60707
5	16.00	2897.00	2575.58666	2942.31334	2251.23989	3266.66011
6	19.00	3327.00	3235.08507	3753.71493	2954.59293	4034.20707
7	17.00	2970.00	2854.38446	3153.81554	2507.54972	3500.65028
8	16.00	2535.00	2575.58666	2942.31334	2251.23989	3266.66011
9	18.00	3138.00	3065.88666	3432.61334	2741.53989	3756.96011

Now lets put it all together!

- A study at University College London (Science, 2004) investigated the relationship between brain activity and pain-related empathy in couples. They had the female partner watch while painful stimulation was applied to the finger of their partner. Two variables were measured. The pain-related brain activity, measured on a scale from -2 to 2, and a score on the Empathic Concern Scale that ranges from 0-25 points. The question of interest is whether people scoring higher in empathy show higher pain-related brain activity.

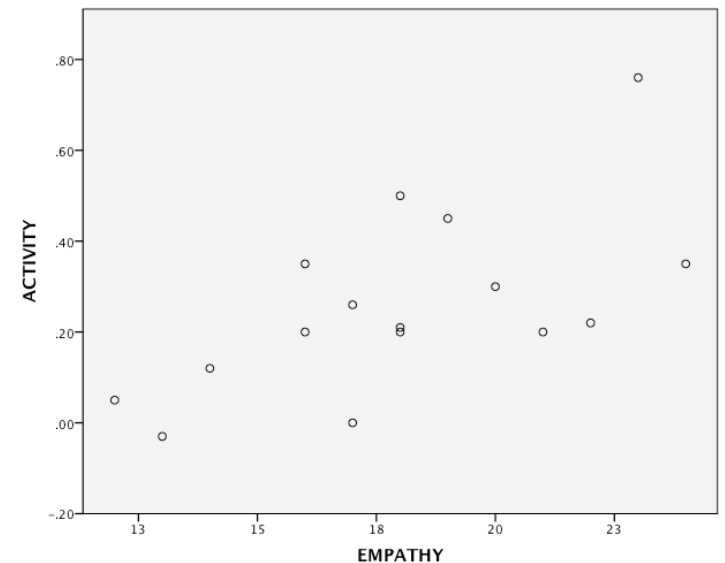
The Four-Step Process for statistical modeling:

1. **Choose** a form for the model
Identify the variables and their types:
Examine graphs to help identify the appropriate model
2. **Fit** the model to the data
Use the sample data to estimate the values of the model parameters
3. **Assess** how well the model fits the data
Verify assumptions
Examine the residuals
Investigate significance, refine model
4. **Use** the model to make predictions, explain relationships, assess differences

The appropriate model depends on the type of variables and the role each variable plays in the analysis.

Choose the form of the model

- Identify the variables and their types
 - What are the variables, and their types?
 - Brain activity, quantitative
 - Empathy rating, quantitative
- Is this consistent with regression?
 - Yes!
- Which is the response, and the predictor?
 - Predictor – empathy
 - response – brain-activity
- Look at the data!



The Four-Step Process for statistical modeling:

1. **Choose** a form for the model

Identify the variables and their types:

Examine graphs to help identify the appropriate model

2. **Fit** the model to the data

Use the sample data to estimate the values of the model parameters

3. **Assess** how well the model fits the data

Verify assumptions

Examine the residuals

Investigate significance, refine model

4. **Use** the model to make predictions, explain relationships, assess differences

The appropriate model depends on the type of variables and the role each variable plays in the analysis.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.627 ^a	.394	.350	.16008

a. Predictors: (Constant), EMPATHY

b. Dependent Variable: ACTIVITY

Fit!

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.233	1	.233	9.092	.009 ^b
	Residual	.359	14	.026		
	Total	.592	15			

a. Dependent Variable: ACTIVITY

b. Predictors: (Constant), EMPATHY

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-.392	.220		-1.787	.096	-.864	.079
	EMPATHY	.036	.012	.627	3.015	.009	.010	.062

a. Dependent Variable: ACTIVITY

What is the fitted regression equation?

$$\text{Activity}^{\wedge} = -.392 + .036\text{Empathy}$$

What does the value .036 tell you?

An increase of 1 point in empathy score is associated with an average increase in brain activity score of .036 points

The Four-Step Process for statistical modeling:

1. **Choose** a form for the model

Identify the variables and their types:

Examine graphs to help identify the appropriate model

2. **Fit** the model to the data

Use the sample data to estimate the values of the model parameters

3. **Assess** how well the model fits the data

Verify assumptions

Examine the residuals

Investigate significance, refine model

4. **Use** the model to make predictions, explain relationships, assess differences

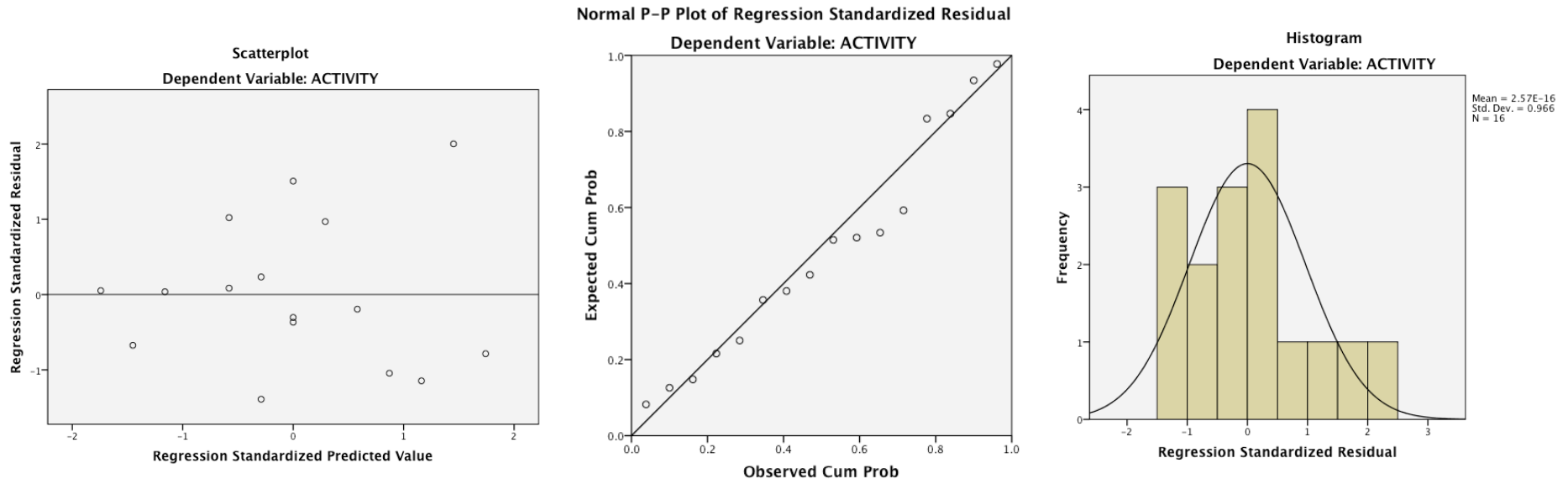
The appropriate model depends on the type of variables and the role each variable plays in the analysis.

Assess the model: verify assumptions

- **Linearity** - the scatterplot shows a general linear pattern
- Conditions that deal with the distribution of ERRORS
 - **Zero Mean** - the distribution of the errors is centered at zero – always true with least squares regression!
 - **Constant Variance** - the variability of the errors is the same for all values of the predictor variable
 - **Independence** - the errors are independent of each other
 - **Normality** – In order to use standard distributions for confidence intervals and hypothesis tests, we often need to assume the random errors follow a normal distribution.
- **Random** – the data are obtained using a random process, like random sampling from a population of interest.

Check residuals with Residual plot, NPP, and histogram

Assess: Look at the residuals



- Linearity?:
 - Residual plot: no clear shape in the residuals
- Constant variance:
 - Residual plot: The plot does not thicken, seems like similar spread for all values of x.
- Normality:
 - NPP falls mostly on the line, histogram is hard to tell....

Check the significance

- Check overall model: 3 parts: hypothesis, decision, conclusion

Hypothesis:

$$H_0: \beta_1 = 0 \quad H_a: \beta_1 \neq 0$$

- What is the pvalue?

.009

- Decision:

– Since p is <.05 and F=9.092, reject the null hypothesis

- Conclusion:

– The data indicates that there is a linear relationship between the empathy scores and brain activity.

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.233	1	.233	9.092	.009 ^b
	Residual	.359	14	.026		
	Total	.592	15			

a. Dependent Variable: ACTIVITY

b. Predictors: (Constant), EMPATHY

Put it in perspective

- What is the coefficient of determination?

$$R^2 = .394$$

- What does it mean?

This tells us that the model (amount of empathy) accounts for 39.4% of the variation in brain activity.

- What is the Standard error of the estimate and what does it represent?

.160, and it represents the typical error when using the model – used to make prediction/confidence intervals later on.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.627 ^a	.394	.350	.16008

a. Predictors: (Constant), EMPATHY

b. Dependent Variable: ACTIVITY

Investigate individual betas

- Check betas, just like model: 3 parts: hypothesis, decision, conclusion

Hypothesis:

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

- What is the pvalue?

.009

- Decision:

Since p is <.05 and t=3.015, reject the null hypothesis

- Conclusion:

The data indicates that there is a positive linear relationship between empathy scores and brain activity, such that every additional point in empathy will result in an increase of .036 points on average, on the brain activity scale.

- Put it in perspective:

We are 95% confident that a one unit change in empathy will result in an average increase in brain activity between .01 and .062 points.

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	-.392	.220		-1.787	.096	-.864	.079
1 EMPATHY	.036	.012	.627	3.015	.009	.010	.062

a. Dependent Variable: ACTIVITY

What about a 1-tailed test?

- Say we have a specific hypothesis about the relationship between empathy and brain activity, such that brain activity should increase as empathy increases?

Hypothesis: $H_0: \beta_1=0$ $H_a: \beta_1>0$

- Must divide pvalue in half to get 1-tailed pvalue.
- What is the pvalue?

$.009/2 = .0045$

- Decision: sign of the test statistic matters

Since p is $<.05$ and $t=3.015$, reject the null hypothesis

- Conclusion:

The data indicates that there is a *positive* linear relationship between the empathy scores and brain activity.

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	-.392	.220		-1.787	.096	-.864	.079
1 EMPATHY	.036	.012	.627	3.015	.009	.010	.062

a. Dependent Variable: ACTIVITY

The Four-Step Process for statistical modeling:

1. **Choose** a form for the model
Identify the variables and their types:
Examine graphs to help identify the appropriate model
2. **Fit** the model to the data
Use the sample data to estimate the values of the model parameters
3. **Assess** how well the model fits the data
Verify assumptions
Examine the residuals
Investigate significance, refine model
4. **Use** the model to make predictions, explain relationships, assess differences

The appropriate model depends on the type of variables and the role each variable plays in the analysis.

Predictions

- What would the average brain activity be for people with an empathy score of 17?
 - We are 95% confident that the average brain activity score for people with an empathy score of 17 will be between .132 and .312 points
- What would the predicted brain activity be for a person with an empathy score of 17?
 - We can be 95% confident that the brain activity score for someone with an empathy score of 17 will be between -.132 and .577

ACTIVITY	EMPATHY	LMCI_1	UMCI_1	LICI_1	UICI_1
.05	12	-.13499	.21834	-.34446	.42780
-.03	13	-.07683	.23253	-.29873	.45443
.12	14	-.02000	.24806	-.25455	.48261
.20	16	.08631	.28648	-.17125	.54403
.35	16	.08631	.28648	-.17125	.54403
.00	17	.13296	.31218	-.13228	.57742
.26	17	.13296	.31218	-.13228	.57742
.50	18	.17291	.34459	-.09516	.61266

Now for more variables!!

Multiple regression

- Essentially the same as simple linear regression, but with more than one predictor.
 - Can include categorical predictors
 - Can look at several different quantitative predictors
 - Can also have higher order terms, like square terms, interactions – anything that combines the original predictors

Form of the model

- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$
- Still uses the sum of squares procedure to minimize residuals – now just more Betas to predict

Predicting the amount (in³) of usable wood in cherry trees

- Might imagine this is related to the height (inches) of the tree.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.598 ^a	.358	.336	13.39698

a. Predictors: (Constant), height

b. Dependent Variable: vol

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2901.189	1	2901.189	16.164	.000 ^b
	Residual	5204.895	29	179.479		
	Total	8106.084	30			

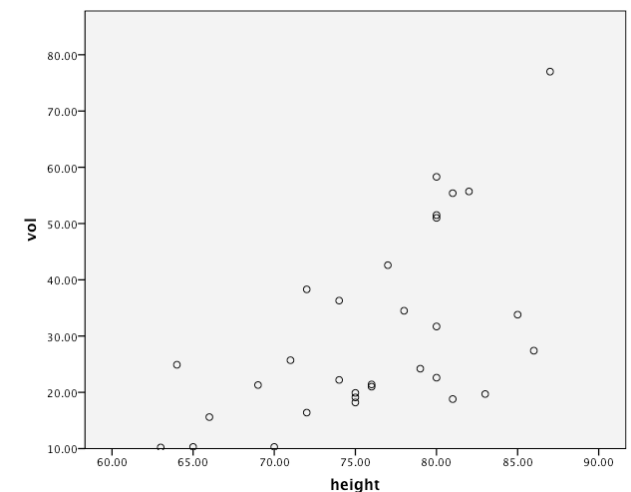
a. Dependent Variable: vol

b. Predictors: (Constant), height

Coefficients^a

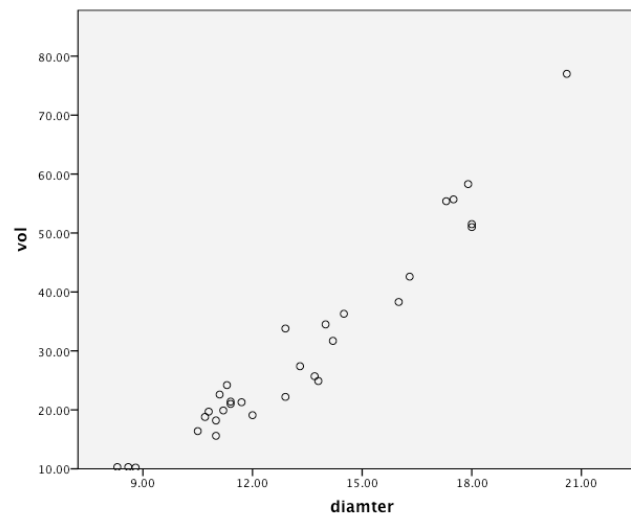
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-87.124	29.273		-2.976	.006
	height	1.543	.384	.598	4.021	.000

a. Dependent Variable: vol



But can we do better?

- Diameter of a tree might also be related to useable wood volume
- Will we do better at predicting wood volume if we use both variables? – Lets try!!

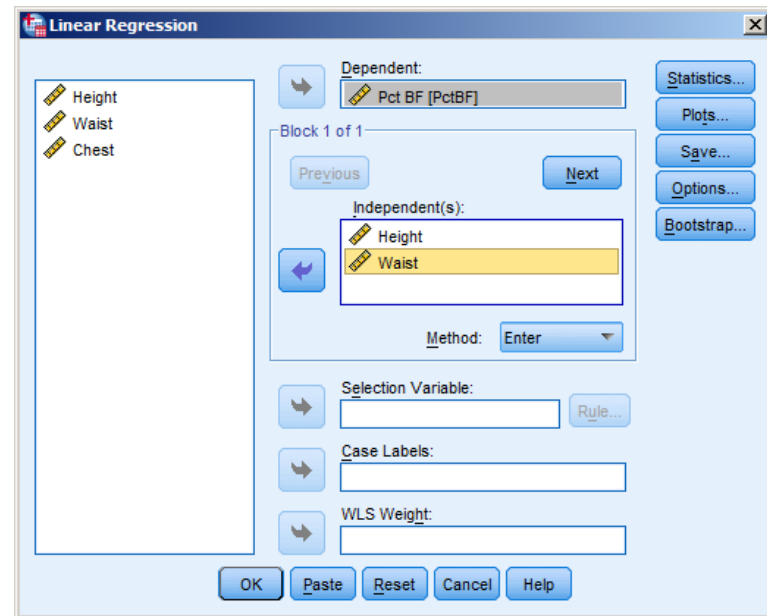
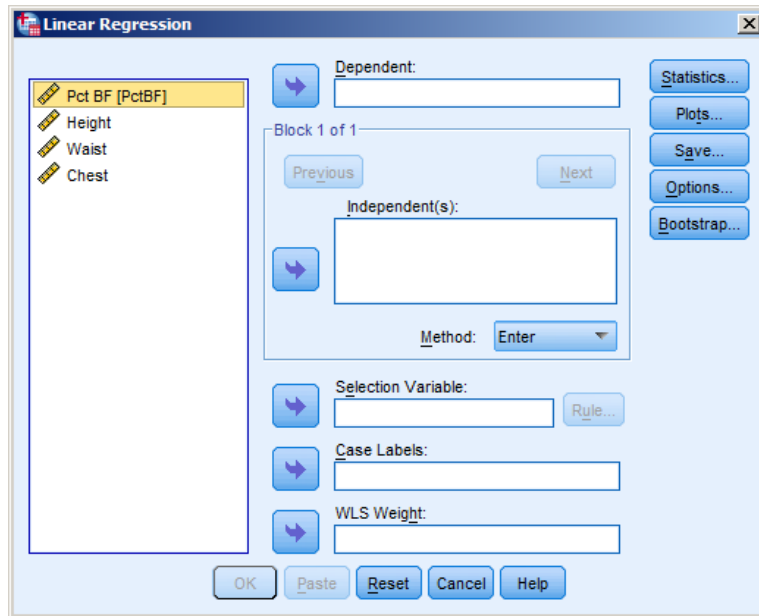


Estimating the Model

Click on **Analyze > Regression > Linear**

Drag the dependent variable and all independent variables to the appropriate locations.

Click on **OK**.



Regression output

- Looks mostly the same

Now each beta gets its own pvalue

And the ANOVA hypothesis doesn't have to match the individual beta tests

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.974 ^a	.948	.944	3.88183

a. Predictors: (Constant), height, diameter

b. Dependent Variable: vol

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7684.163	2	3842.081	254.972	.000 ^b
	Residual	421.921	28	15.069		
	Total	8106.084	30			

a. Dependent Variable: vol

b. Predictors: (Constant), height, diameter

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-57.988	8.638		-6.713	.000
	diameter	4.708	.264	.899	17.816	.000
	height	.339	.130	.132	2.607	.014

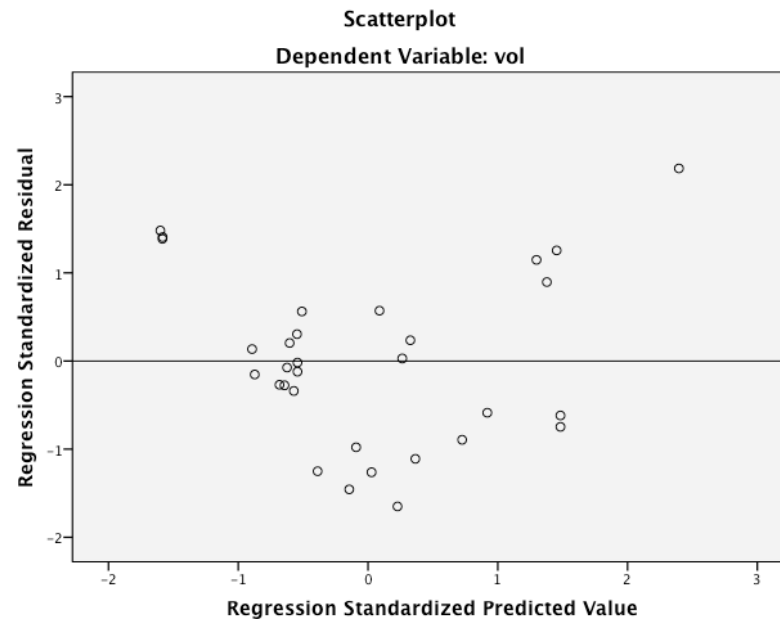
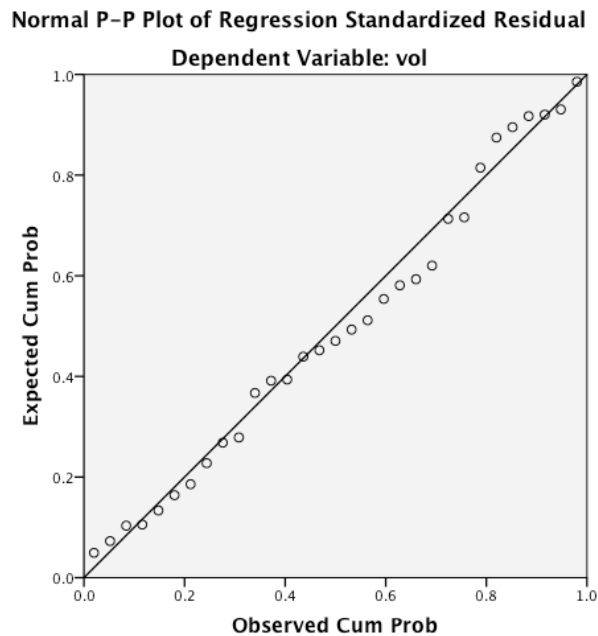
a. Dependent Variable: vol

Very similar assumptions

- But no linearity assumption – because we will see later that we can use higher order terms to represent non-linear relationships
- Conditions that deal with the distribution of ERRORS
 - **Zero Mean** - the distribution of the errors is centered at zero – always true with least squares regression!
 - **Constant Variance** - the variability of the errors is the same for all values of the predictor variable
 - **Independence** - the errors are independent of each other
 - **Normality** – In order to use standard distributions for confidence intervals and hypothesis tests, we often need to assume the random errors follow a normal distribution.
- **Random** – the data are obtained using a random process, like random sampling from a population of interest.

Check residuals

- Do we meet the assumptions?
 - Residuals look randomly distributed around 0, maybe a possible outlier, and the residuals look normal, as the NPP plot shows the data very close to the line



Assess the overall model

- Hypothesis test now differs from individual beta test

$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$$

H_a : The slopes are not all zero – or at least one slope is not zero

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7684.163	2	3842.081	254.972	.000 ^b
	Residual	421.921	28	15.069		
	Total	8106.084	30			

a. Dependent Variable: vol

b. Predictors: (Constant), height, diameter

Assess the overall model

- For our data, with 2 betas:

$$H_0: \beta_1 = \beta_2 = 0 \text{ or } H_0: \beta_{\text{diameter}} = \beta_{\text{height}} = 0$$

H_a : at least one slope is not zero (no easy way to write with symbols)

- Decision:

Pvalue is essentially zero, 0^+ , or $p < .001$ and $F = 254.972$, we can reject the null hypothesis

- Conclusion:

The data indicates that there is a relationship between volume of usable cherry wood and the predictor variables diameter and height. Together they account for a significant amount of the variability in cherry wood volume.

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7684.163	2	3842.081	254.972	.000 ^b
	Residual	421.921	28	15.069		
	Total	8106.084	30			

a. Dependent Variable: vol

b. Predictors: (Constant), height, diameter

Assess the overall model

- Coefficient of determination, R^2 – no longer directly related to r
- We could use, it means the same thing in multiple regression – but can be inflated when you have several predictors
- Instead, for multiple regression we use R^2_{adj}
 - Has a penalty for additional predictors, and considers sample size
- Can interpret the same way:

Height and diameter together account for 94.4% of the variability in volume of usable cherry wood, after correcting for the number of predictors and sample size.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.974 ^a	.948	.944	3.88183

a. Predictors: (Constant), height, diameter

b. Dependent Variable: vol

Are we doing 'better' by including two variables?

- One easy way to answer that is by looking at R^2_{adj} values
- What do you think?

— Yes

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.598 ^a	.358	.336	13.39698

a. Predictors: (Constant), height

b. Dependent Variable: vol

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.974 ^a	.948	.944	3.88183

a. Predictors: (Constant), height, diameter

b. Dependent Variable: vol

Are we doing 'better' by including two variables?

- Another way is the standard error of the estimate
- What do you think?

— Yes

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.598 ^a	.358	.336	13.39698

a. Predictors: (Constant), height

b. Dependent Variable: vol

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.974 ^a	.948	.944	3.88183

a. Predictors: (Constant), height, diameter

b. Dependent Variable: vol

Individual betas

- Now we know the model is good, but which betas are contributing to the 'goodness' of the model?
- Need to test each beta separately
- Generically:

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	-57.988	8.638	-6.713	.000
	diameter	4.708	.264	.899	.000
	height	.339	.130	.132	.014

a. Dependent Variable: vol

Individual betas

$$H_0: \beta_{\text{diameter}} = 0$$

$$H_a: \beta_{\text{diameter}} \neq 0$$

- Decision: Since p is less than .05 and $t=17.816$, we can reject the null hypothesis
- Conclusion:
 - The data suggests that there is a significant relationship between tree diameter and average volume of usable cherry wood, **after accounting for height**. (will interpret the slope value later)
- Some people say, while holding height constant, or for a given value of height.

Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	-57.988	8.638	-6.713	.000
	diameter	4.708	.264	.899	.000
	height	.339	.130	.132	.014

a. Dependent Variable: vol

Individual betas

- Important to note that the meaning of each coefficient depends on all of the predictors in the regression model
 - If we fail to reject the null hypothesis, it means that the corresponding predictor variable contributes nothing to the multiple regression model after allowing for all other predictors.

Coefficients ^a					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	-57.988	8.638	-6.713	.000
	diameter	4.708	.264	.899	.000
	height	.339	.130	.132	.014

a. Dependent Variable: vol

Individual betas

$$H_0: \beta_{\text{height}} = 0$$

$$H_a: \beta_{\text{height}} \neq 0$$

- Decision:

Since p is less than .05 and $t=2.607$, we can reject the null hypothesis

- Conclusion:

The data suggests that there is a significant relationship between tree height and volume of usable cherry wood, **after accounting for tree diameter.**

- These tests are independent.
- What is the fitted regression equation?

$$\text{Wood volume}^{\wedge} = -58.0 + 4.7\text{Diameter} + .34\text{Height}$$

- Interpret slope for height:

- For each additional inch in the height of a tree, the volume of wood will increase on average by .34 in³, for a given value of tree diameter. (while holding tree diameter constant)

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	-57.988	8.638		-6.713	.000
diameter	4.708	.264	.899	17.816	.000
height	.339	.130	.132	2.607	.014

a. Dependent Variable: vol

What if one wasn't significant?

- Then we can not reject the null hypothesis, and it suggests that there is no significant relationship between the predictor and the response variable, after accounting for all of the other predictors.
- This suggests we could take the variable out of the model, **and rerun it!!** – but depends on our question of interest.
 - Good to try. If everything still looks good, or better, then go with the model with less terms.

Making predictions

- Same as for simple linear regression, but need to have a value for each predictor in mind.
- What is the average predicted wood volume for trees with a height of 62 in and a diameter of 12 in?
 - We can be 95% confident that the average volume of usable wood will be between -3.5 and 20.6 for trees with a height of 62 in and a diameter of 12in
- What is the predicted wood volume for a tree with a height of 62 in and a diameter of 12 in?
 - We can be 95% confident that the volume of usable wood will be between -21.4 to 38.5 in³ for a tree with a height of 62 in and a diameter of 12in

diamter	height	vol	LMCI_1	UMCI_1	LICI_1	UICI_1
18.00	80.00	51.00	30.50656	42.18218	8.32946	64.35927
20.60	87.00	77.00	37.20798	57.08765	18.00069	76.29494
12.00	62.00	.	-3.47873	20.60687	-21.36558	38.49372