Multiple regression!

Start time stamper!
Eastern time zone

Week 3

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 he 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 (in^3) of usable wood in cherry trees

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

Model Summary[b]

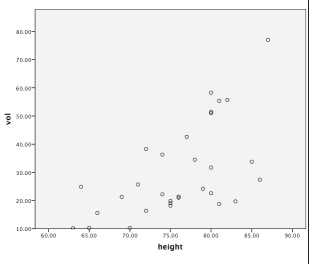
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.598[a]	.358	.336	13.39698

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



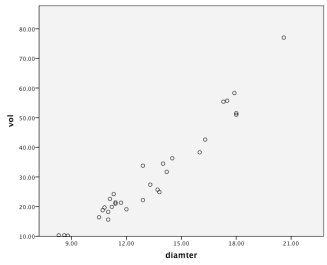
Coefficients[a]

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!!

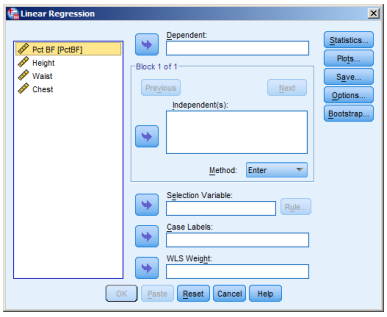


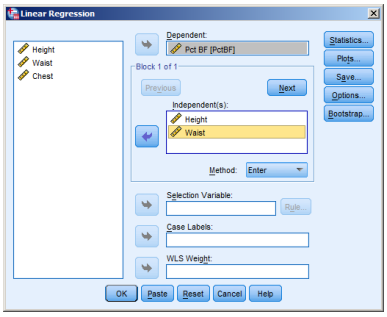
Estimating the Model

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, diamter
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, diamter

Coefficients[a]

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

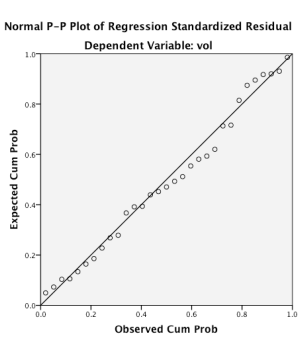
a. Dependent Variable: vol

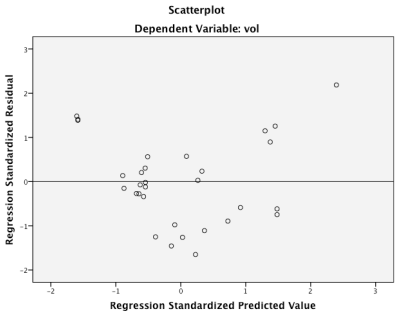
Very similar assumptions

- 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, diamter

Assess the overall model

- For our data, with 2 betas:

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

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

  Pvalue is essentially zero, 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 are useful in predicting the average volume of cherry wood.

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, diamter

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^2adj
  - Has a penalty for additional predictors, and considers sample size
  - Does R^2adj always increase when add more predictors? 75% corr
- 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, diamter

b. Dependent Variable: vol

Are we doing 'better' by including two variables?

- One easy way to answer that is by looking at R^2adj 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, diamter
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, diamter
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
	diamter	4.708	.264	.899	17.816	.000
	height	.339	.130	.132	2.607	.014

a. Dependent Variable: vol

Individual betas

$H_0: \beta_{diameter} = 0$

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

- 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
	diamter	4.708	.264	.899	17.816	.000
	height	.339	.130	.132	2.607	.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
	diamter	4.708	.264	.899	17.816	.000
	height	.339	.130	.132	2.607	.014

a. Dependent Variable: vol

Individual betas

$H_0: \beta_{height} = 0$

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

- 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 average volume of usable cherry wood, **after accounting for tree diameter.**

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

  Wood volume^= -58.0 +4.7Diameter +.34Height

- Interpret slope for height:
  - For each additional inch in the height of a tree, the volume of wood will increase on average by .34 in^3, for a given value of tree diameter. (while holding tree diameter constant, after accounting for tree diameter)

Coefficients[a]

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

a. Dependent Variable: vol

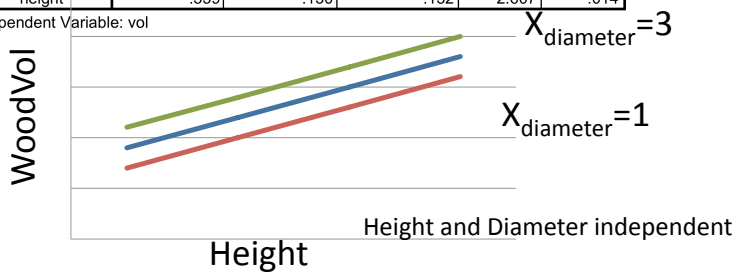
Individual betas

- Wood volume^= -58.0 +4.7Diameter +.34Height
  - For each additional inch in the height of a tree, the volume of wood will increase on average by .34 in^3, for a given value of tree diameter
- Wood volume^= -58.0 +4.7(1) +.34Height =-53.3 +.34Height
- Wood volume^= -58.0 +4.7(3) +.34Height =-43.9 +.34Height

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-57.988	8.638		-6.713	.000
	diamter	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 in 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 in^3 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

Categorical predictors

- Categorical (or qualitative) variables can also be included in multiple regression models. These variables are coded as numbers so that we can employ the methods we have discussed, but are not really the same as quantitative variables. These coded values are called *indicator variables* or *dummy variables*.
- They are coded using 0 and 1, where

  0 = absence or 0 = "no"

  1 = presence 1 = "yes"

Maybe where a tree comes from matters

- We could imagine having a categorical variable representing if a tree was from the east or west cost. We can call this variable FromWest, and it will be:

  FromWest = 0 if the tree is from the east

  1 if the tree is from the west

- You MUST define these indicator variables, as which one is coded as 0/1 makes a huge difference, and is often arbitrary.

Comparing two regression lines

- This is how your book describes what is being investigated when you include one of these binary categorical variables
- This is because of what it does to the regression equation

Tree vol^= $\beta_0 + \beta_1$Diameter $+\beta_2$Height $+\beta_3$ FromWest

When a tree is from the west:

Tree vol^= $\beta_0 + \beta_1$Diameter $+\beta_2$Height $+\beta_3$ (1)

Tree vol^= $\mathbf{\beta_0 + \beta_3} + \beta_1$Diameter $+\beta_2$Height

When a tree is from the east:

Tree vol^= $\beta_0 + \beta_1$Diameter $+\beta_2$Height $+\beta_3$ (0)

Tree vol^= $\beta_0 + \beta_1$Diameter $+\beta_2$Height

- So including this binary predictor essential results in testing the hypothesis of whether the intercept is the same for trees from the east and west coasts.

What this means in your graph

Tree vol^= $\beta_0 + \beta_1$Diameter +β_2Height +β_3 FromWest

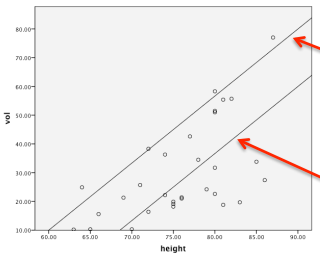
When a tree is from the west:

Tree vol^= $\mathbf{\beta_0 + \beta_3} + \beta_1$Diameter +$\beta_2$Height

When a tree is from the east:

Tree vol^= $\beta_0 + \beta_1$Diameter +β_2Height

It changes the INTERCEPT, not the slope -> so get parallel lines



So for a given value of diameter, the regression line for trees from the west might look like this

While for the same diameter, the line for trees from the east might look like this

Interpreting betas

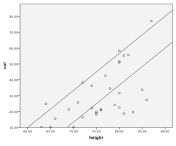
When a tree is from the west:

Tree vol^= $\mathbf{\beta_0 + \beta_3} + \beta_1$Diameter +$\beta_2$Height

When a tree is from the east:

Tree vol^= $\beta_0 + \beta_1$Diameter +β_2Height

- B_0= mean tree volume for trees from the east, after accounting for tree diameter and height – or can think of as the base or beginner tree vol (tree vol when diameter and height is zero)
- B_3= change in tree vol when go from a tree in the east to one in the west – for a given value of tree diameter and height (or after accounting for tree diameter and tree height)



Mac vs PC question, if the beta for th categorical variable is signficant, always a difference in price 78% corr

Interpreting betas from categorical variables

Tree vol^= 20 +15Height + 10Diameter +5FromWest

- What does 20 represent?

  the mean tree vol for a tree from the east with 0 height and 0 diameter, or the mean tree vol for trees from the east after accounting for tree height and diameter

- What does 5 represent?

  The average difference in tree vol when a tree is from the west compared to the east, after accounting for tree height and diameter

Can do this for more than binary categorical variables

- Lets investigate a cooling method for engines. The response variable is the heatrate of the engine (kilojoules per kilowatt per hour), which we will predict using the speed of the engine (revolutions per min) and the type of engine (Traditional, advanced, and aerodynamic).
- Engine type is categorical with 3 levels
- You would need to build n-1 indicator variables, where n=number of levels in your categorical variable
- One category isn't coded, this is the 'base' or reference category – its essentially the '0' category when you have a binary variable
  - IsAdvan= 1 when from advanced, 0 otherwise
  - IsAero = 1 when aerodynamic, 0 otherwise
- So this one variable of interest turns into 2 terms in the regression model

$Heatrate=\beta_0 + \beta_1 IsAdvan + \beta_2 IsAero + \beta_3 RPM + \varepsilon$

Model output

- Heatrate^ = 9,919 -901.04IsAdvan -692.17IsAero +.18RPM
- What does 9,919 represent?
  - The average heatrate for a traditional engine with 0RPM rate
- Interpret the beta for IsAero
  - On average, heatrate will decrease by 692.17 kj/kw/hr for aerodynamic engines compared to traditional engines for a given RPM
- What would the graph of this model look like?
  - Three parallel lines, one for each engine type.
- If ISAero had been nonsig, could we take it out of the model?
  - No! it is necessary to correctly code engine type

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9919.275	187.796		52.819	.000
	IsAdvanced	-901.041	220.607	-.264	-4.084	.000
	IsAero	-692.174	338.624	-.134	-2.044	.045
	RPM	.180	.016	.794	11.400	.000

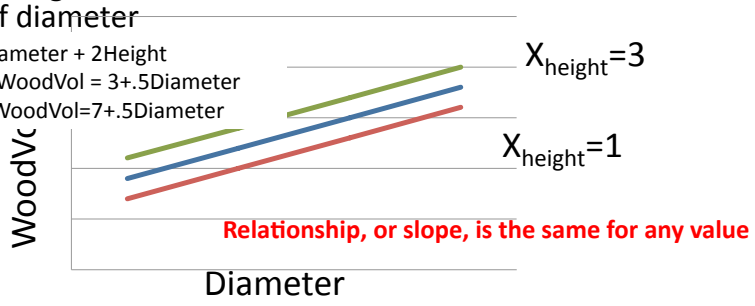
a. Dependent Variable: HEATRATE

Higher order terms: Interactions!

New predictors from old: Interactions

- Sometimes we may hypothesize that two predictors might interact with each other.
  - Short trees with a very large diameter might have a particularly high volume of usable wood, while a really tall tree with a small diameter might not have much usable wood.
- This suggests that there may be a different relationship between diameter and wood volume for different values of height
- This goes against what we concluded before about the slope of diameter

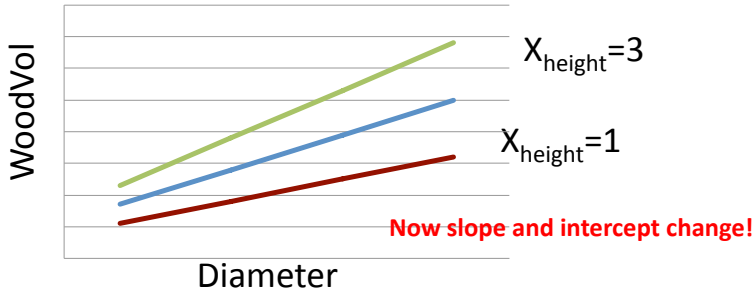
WoodVol^=1 +.5Diameter + 2Height
When height^ =1: WoodVol = 3+.5Diameter
When height^=3: WoodVol=7+.5Diameter



Let's add an interaction term! They way we model this is by adding an 'interaction' term, which is created by simply multiplying tree diameter by tree height.

How does this change the equation?

- WoodVol^=1 +.5Diameter + 2Height +3DiaHeight
- When height=1:
  WoodVo^l= 1+.5Diameter +2(1) +3Diameter(1)
  WoodVol^ = 3+.5Diameter +3Diameter
  WoodVol^ = 3 +3.5Diameter
- When Height = 2:
  WoodVol^= 1+.5Diameter +2(2) +3Diameter(2)
  WoodVol^ = 5+.5Diameter +6Diameter
  WoodVol^ = 5 +6.5Diameter



Interactions

- This means when you want to know the relationship between height and wood volume, you can't just look at β_2

$Y^ = \beta_0 + \beta_1 Diameter + \beta_2 Height + \mathbf{\beta_3 DiameterHeight}$

- Its really now for each unit change in height, wood volume is expected to change by:

$\beta_2 + \beta_3 Diameter$, after accounting for tree diameter

Means that the effect of one predictor on the response variable is different at different values of the other predictor.

How interaction terms change things

$\hat{Y} = \beta_0 + \beta_1 Diameter + \beta_2 Height + \mathbf{\beta_3 DiameterHeight}$

- 3 important consequences
  - You **can not infer the direction of the relationship from the sign of β_2.**
  - If β_2 is not significant, but the β_3 is, then you can't get rid of β_2
    - **Can't remove terms when they are used to create 'higher order' terms**
  - If higher order terms aren't significant – get rid of them!!
    - And lower order term?
      - Check significance without higher order term before removing.

Interaction data

- Output looks just like normal.
- What is the fitted regression equation?

WoodVol^=-15.2 +1.2Dia + .19Height + .03DiaHeight

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.997[a]	.993	.992	1.44169

a. Predictors: (Constant), diam_height, height, diamter
b. Dependent Variable: vol

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8049.966	3	2683.322	1291.017	.000[b]
	Residual	56.118	27	2.078		
	Total	8106.084	30			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-15.186	4.550		-3.338	.002
	diamter	1.165	.285	.222	4.095	.000
	height	.193	.050	.075	3.882	.001
	diam_height	.034	.003	.739	13.266	.000

a. Dependent Variable: vol

Interaction data

WoodVol^=-15.2 +1.2Dia + .19Height + .03DiaHeight

- Does the relationship between tree diameter and wood volume depend on height?

H_0: $\beta_{DiaHeight} = 0$

H_a: $\beta_{DiaHeight} \neq 0$

- Decision:
  - Since p <.05, reject null hypothesis
- Conclusion:
  - The relationship between tree diameter and average wood volume is different at different values of tree height.
- To find out how it is different, plug in a few values of height
- Height =0: 15.2 +1.2Dia, height =10: 15.2 +1.2Dia +1.9 +.3Dia, -> 17.1 +1.5Dia
  - The average increase in tree vol for a one unit increase in tree diameter will be larger for larger values of tree height

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-15.186	4.550		-3.338	.002
	diamter	1.165	.285	.222	4.095	.000
	height	.193	.050	.075	3.882	.001
	diam_height	.034	.003	.739	13.266	.000

a. Dependent Variable: vol

Interactions with categorical predictors

- Go back to our engine heatrate example, but lets hypothesize that the amount of heatrate varies not only by engine type, but for each engine type varies by RPMs.
- Old model:

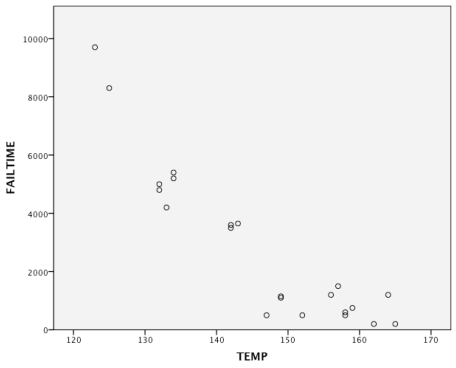
  Heatrate^=$\beta_0 + \beta_1$IsAdvan + β_2IsAero + β_3RPM
- New model?

  Heatrate^=$\beta_0 + \beta_1$IsAdvan + β_2IsAero + β_3RPM + β_4IsAdvanRPM + β_5IsAeroRPM
- How to interpret
  - Predicting Heatrate for Traditional:
    - β_0 = intercept, β_3 = slope
  - Predicting heatrate for Advanced:
    - β_1 = change in intercept, β_4 = change in slope from traditional
  - Predicting heatrate for Aerodynamic:
    - β_2 =change in intercept, β_5 = change in slope from traditional

Polynomial regression: quadratic terms

New predictors from old: Polynomial regression

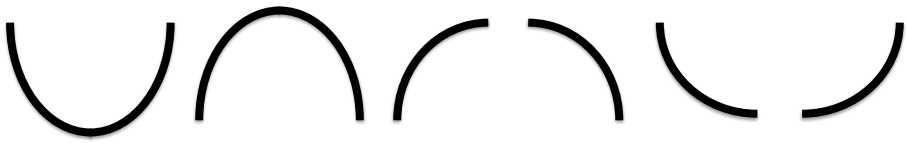
- Material scientists investigated how long it took microchips to fail (in hours), for different solder temperatures (degrees centigrade)
- Saw how we can transform data, but now we can actually model **curvilinea**r relationships



Quadratic relationships

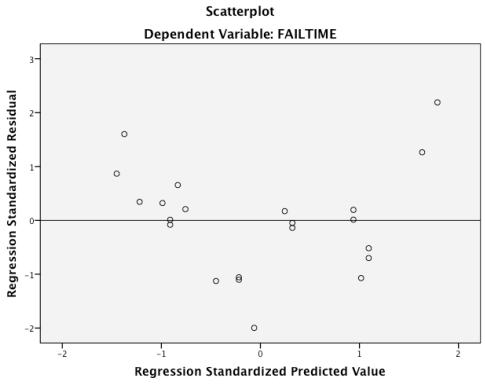
- Add 'powers' to the equation

  $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$



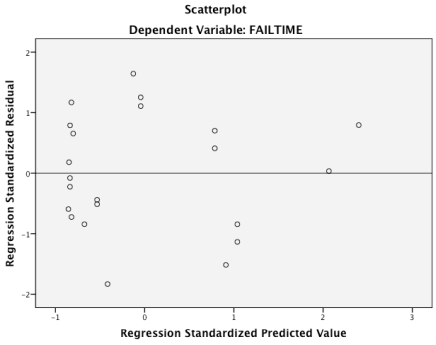
If we don't include the square term

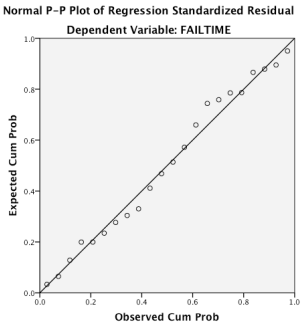
- $Y\hat{} = \beta_0 + \beta_1 X$
- We get a 'pattern' in the residuals



Check residuals

- $\hat{Y} = \beta_0 + \beta_1 temp + \beta_2 temp^2$





Output

- To test whether there is a quadratic relationship, need to test beta for higher order term

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.970[a]	.942	.935	688.137

a. Predictors: (Constant), tempsq, TEMP

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	144830279.619	2	72415139.810	152.926	.000[b]
	Residual	8997106.744	19	473531.934		
	Total	153827386.364	21			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	154242.914	21868.474		7.053	.000
	TEMP	-1908.850	303.664	-9.148	-6.286	.000
	tempsq	5.929	1.048	8.236	5.659	.000

a. Dependent Variable: FAILTIME

Is there a curvilinear relationship?

H_0: $\beta_{tempsq} = 0$

H_a: $\beta_{tempsq} \neq 0$

- Decision: Since p is less than .05, we can reject the null hypothesis
- Conclusion:

The data suggests that there is a significant **quadratic** relationship between solder temperature and the average time it takes for a microchip to fail.

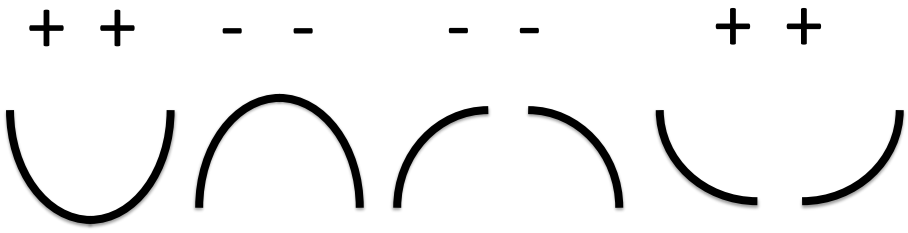
- Similar to when you add an interaction term, you can now no longer simply look at the sign of TEMP to determine the relationship.
- Higher order terms mess things up!!
- Also like in an interaction, you should leave lower order term in, even if not significant.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	154242.914	21868.474		7.053	.000
	TEMP	-1908.850	303.664	-9.148	-6.286	.000
	tempsq	5.929	1.048	8.236	5.659	.000

a. Dependent Variable: FAILTIME

Can get information about direction of relationship from the sign of quadratic beta



U-shaped

Inverted U-shaped

Increase at a decreasing rate

Decrease at an increasing rate

Decrease at a decreasing rate

Increase at an increasing rate

Putting relationship in words

- Look at your data, because you need to know more than just the sign to be accurate

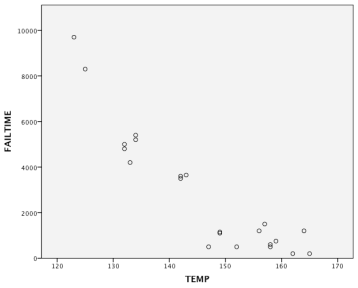
Failtime^=154,242.9 – 1,908.9temp + 5.9temp^2

The data suggests that there is a significant curvilinear relationship between solder temperature and fail time, such that as temperature increases, average fail time decreases at a decreasing rate.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	154242.914	21868.474		7.053	.000
	TEMP	-1908.850	303.664	-9.148	-6.286	.000
	tempsq	5.929	1.048	8.236	5.659	.000

a. Dependent Variable: FAILTIME



Summary

- Can have categorical variables via indicator variables
  - Essentially allow you to model multiple lines (with different intercept) for each level of categorical variable
- Can include higher order terms like interactions and quadratic terms
  - Complicate interpretation of lower-order terms
  - If higher order terms not sig, take them out
  - If higher order term is sig, must keep lower order terms in regardless

Multicollinarity

- When some of your predictor variables are correlated
- It's a problem because then they are 'fighting' over similar variability in response variable
  - Redundant
- Can make results weird…..
- Things to look for:
  - Predictors can be very significant alone, but not when combined
  - model can be significant while predictors are not
  - signs of the betas may not be what is predicted
- Can check for by looking at correlation between all your variables in the model

Dataset: cigarette content

- The Federal Trade Commission ranks cigarettes according to tar, nicotine, and carbon monoxide contents – three substances thought to be hazardous to health. Past studies have show that increases in tar and nicotine are accompanied by an increase in carbon monoxide. They also recorded weight. Want to model:

$$CarbonMonoxide=\beta_0 + \beta_1 X_{tar} + \beta_2 X_{nicotine} + \beta_3 X_{weight} + \varepsilon$$

- Would hypothesize that carbon monoxide should increase as tar, nicotine, and weight increases

Output

Does anything seem odd?

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.958[a]	.919	.907	1.4457

a. Predictors: (Constant), weight, tar, nicotine
b. Dependent Variable: co

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	495.258	3	165.086	78.984	.000[b]
	Residual	43.893	21	2.090		
	Total	539.150	24			

a. Dependent Variable: co
b. Predictors: (Constant), weight, tar, nicotine

Coefficients[a]

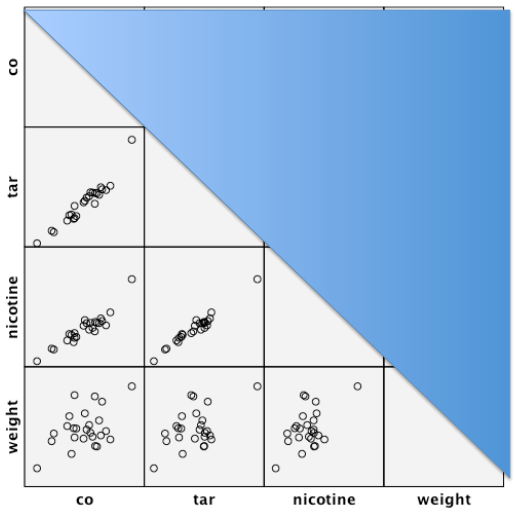
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.202	3.462		.925	.365
	tar	.963	.242	1.151	3.974	.001
	nicotine	-2.632	3.901	-.197	-.675	.507
	weight	-.130	3.885	-.002	-.034	.974

a. Dependent Variable: co

Model very significant, while only one Beta is, and sign of nicotine is opposite of what would predict

Investigate multicollinearity: Scatter plot matrix

- Only need to look at one half of the plot



If you are a numbers person: Correlation matrix

- Puts numbers to graphs, only shows predictors

Coefficient Correlations[a]

Model		weight	tar	nicotine
Correlations	weight	1.000		
	tar	-.012	1.000	
	nicotine	-.112	-.969	1.000

a. Dependent Variable: co

Variance inflation factor (VIF)

- A statistic to capture multicollinearity
- Helpful for more subtle dependence among predictors, such as when some combination of predictors taken together are strongly related to another predictor.
- VIF reflects the association between a predictor and ALL the other predictors
  - VIF > 5 can be an indicator of an issue, > 10, very concerning
- Tolerance is the inverse of VIF

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3.202	3.462		.925	.365		
	tar	.963	.242	1.151	3.974	.001	.046	21.631
	nicotine	-2.632	3.901	-.197	-.675	.507	.046	21.900
	weight	-.130	3.885	-.002	-.034	.974	.750	1.334

a. Dependent Variable: co

So what should you do?

- Depends on what you want to do with model
- If you just want to estimate/predict
  - you can leave all the predictors in, and as long as the population you will make predictions about has the same collinearity issues as your sample did.
  - Avoid making inferences about individual betas
- Generally, trying to answer a relationship question, and want to interpret individual betas
  - Drop predictors
    - If adjusted R^2 doesn't change much, you are probably good
  - Combine predictors
    - Can add predictors together (tar + nicotine)
    - Or sometimes a ratio (tar/nicotine)

Model building: how to choose predictors?

Model building: how to choose predictors

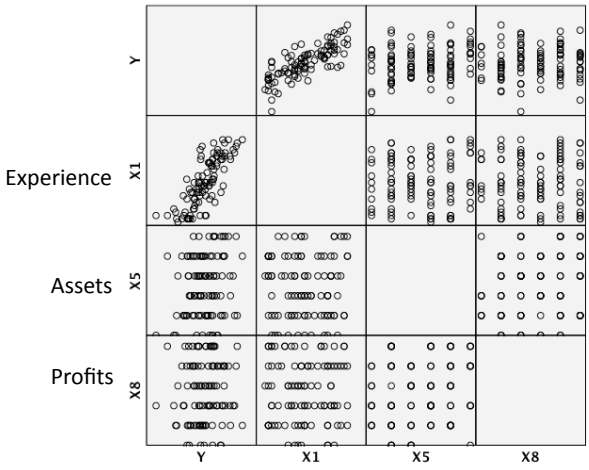
- Sometimes you might have many variables to choose from in building a model.
- Why not just include them all?
  - Usually models with less predictors and less higher order terms are easier to understand and use
  - So it's a balance between doing a good job at predicting and staying simple
- Best way to start -> look at correlations between variables and the response variable (and each other)

Predicting executive salaries

- Response variable: Salary, in $10,000
- Predictors
  - Years of experience
  - Corporate assets (in millions)
  - Company profits (past 12 months, in millions)

Look at your data!

- Might want to include sq of years of experience



Results

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.814[a]	.662	.648	.1541193

a. Predictors: (Constant), X1sq, X8, X5, X1
b. Dependent Variable: Y

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.426	4	1.107	46.586	.000[b]
	Residual	2.257	95	.024		
	Total	6.683	99			

a. Dependent Variable: Y
b. Predictors: (Constant), X1sq, X8, X5, X1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.421	.199		52.313	.000
	X1	.036	.009	1.011	4.092	.000
	X8	.002	.010	.014	.235	.814
	X5	.004	.001	.208	3.384	.001
	X1sq	.000	.000	-.232	-.938	.351

a. Dependent Variable: Y

We have a few nonsig factors, might want to eliminate – especially higher order terms

Results: removing nonsig terms

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.812[a]	.659	.652	.1532507

a. Predictors: (Constant), X5, X1
b. Dependent Variable: Y

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.405	2	2.202	93.770	.000[b]
	Residual	2.278	97	.023		
	Total	6.683	99			

a. Dependent Variable: Y
b. Predictors: (Constant), X5, X1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.500	.178		59.077	.000
	X1	.028	.002	.786	13.256	.000
	X5	.003	.001	.200	3.379	.001

a. Dependent Variable: Y

Now everything is significant, but are we still doing a good job?

Compare models

- With all terms

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.814[a]	.662	.648	.1541193

a. Predictors: (Constant), X1sq, X8, X5, X1
b. Dependent Variable: Y

- With only years of experience and corporate assets

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.812[a]	.659	.652	.1532507

a. Predictors: (Constant), X5, X1
b. Dependent Variable: Y

- What do you think?
  - Yes! Adjusted R^2 higher, stderr of the estimate smaller

But what if too many variables?

- Response variable: Salary, in $10,000
- Predictors
  - X_1 Years of experience
  - X_2 Education (years)
  - X_3 Gender (1 if male, 0 if female)
  - X_4 Number of employees supervised
  - X_5 Corporate assets (in millions)
  - X_6 Board member (1 if yes, 0 if no)
  - X_7 Age (years)
  - X_8 Company profits (past 12 months, in millions)
  - X_9 Has international responsibility (1 if yes, 0 if no)
  - X_{10} Company's total sales (past 12 months, in millions)

Stepwise regression

- Pretty much does the same process as we did by hand
- Finds the variable with the largest correlation (R^2)
  - Then tries each additional predictor, looking for the one that results in the largest increase in R^2.
  - If that variable isn't significant (α=.05), it stops
  - Otherwise keeps going
- One caveat, sometimes when additional predictors are added, ones that were significant in the beginning might become nonsig. (redundant), so keeps an eye out for this
  - Will delete any variables that become nonsig

Results

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.787[a]	.619	.615	.1611893
2	.866[b]	.749	.744	.1314457
3	.916[c]	.839	.834	.1058344
4	.953[d]	.907	.904	.0806758
5	.959[e]	.921	.916	.0751179

a. Predictors: (Constant), X1
b. Predictors: (Constant), X1, X3
c. Predictors: (Constant), X1, X3, X4
d. Predictors: (Constant), X1, X3, X4, X2

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.136	1	4.136	159.204	.000[b]
	Residual	2.546	98	.026		
	Total	6.683	99			
2	Regression	5.007	2	2.503	144.887	.000[c]
	Residual	1.676	97	.017		
	Total	6.683	99			
3	Regression	5.607	3	1.869	166.873	.000[d]
	Residual	1.075	96	.011		
	Total	6.683	99			
4	Regression	6.064	4	1.516	232.936	.000[e]
	Residual	.618	95	.007		
	Total	6.683	99			
5	Regression	6.152	5	1.230	218.061	.000[f]
	Residual	.530	94	.006		
	Total	6.683	99			

a. Dependent Variable: Y
b. Predictors: (Constant), X1
c. Predictors: (Constant), X1, X3
d. Predictors: (Constant), X1, X3, X4
e. Predictors: (Constant), X1, X3, X4, X2

Coefficients

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	11.091	.033		335.524	.000
	X1	.028	.002	.787	12.618	.000
2	(Constant)	10.968	.032		342.659	.000
	X1	.027	.002	.770	15.134	.000
	X3	.197	.028	.361	7.097	.000
3	(Constant)	10.783	.036		298.170	.000
	X1	.027	.001	.771	18.801	.000
	X3	.233	.023	.427	10.170	.000
	X4	.000	.000	.307	7.323	.000
4	(Constant)	10.278	.066		155.154	.000
	X1	.027	.001	.771	24.677	.000
	X3	.232	.017	.425	13.297	.000
	X4	.001	.000	.354	10.920	.000
	X2	.030	.004	.266	8.379	.000
5	(Constant)	9.962	.101		98.578	.000
	X1	.027	.001	.771	26.501	.000
	X3	.225	.016	.412	13.742	.000
	X4	.001	.000	.337	11.064	.000
	X2	.029	.003	.258	8.719	.000
	X5	.002	.000	.116	3.947	.000

a. Dependent Variable: Y

- All terms will be sig, as a result of the process
- What is the final model?

$Y=9.962 + .027X_1 + .225X_3 + .001X_4 + .029X_2 + .002X_5$

Final fitted equation

- $Y=9.962+ .027X_1 + .225X_3 +.001X_4 +.029X_2 + .002X_5$
- Predictors
  - **X_1 Years of experienc**e
  - **X_2 Education (years)**
  - **X_3 Gender (1 if male, 0 if female)**
  - **X_4 Number of employees supervised**
  - **X_5 Corporate assets (in millions)**
  - X_6 Board member (1 if yes, 0 if no)
  - X_7 Age (years)
  - X_8 Company profits (past 12 months, in millions)
  - X_9 Has international responsibility (1 if yes, 0 if no)
  - X_{10} Company's total sales (past 12 months, in millions)

Use model building with caution!

- Based on overall model goodness
- Doesn't know your question of interest
- Isn't checking residual plots and such
- SPSS can't tell what makes 'sense'
- Usually only done with first order terms – no interactions or quadratic terms
  - Need to think about whether should include some of these after you run stepwise
  - May want to test the complete second order model

Complete second order model

- $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$
- Predictors
  - **X_1 Years of experience**
  - **X_2 Education (years)**
  - **X_3 Gender (1 if male, 0 if female)**
  - **X_4 Number of employees supervised**
  - **X_5 Corporate assets (in millions)**
- All interactions
  - $\beta_6 X_1X_2 + \beta_7 X_1X_3 + \beta_8 X_1X_4 + \beta_9 X_1X_5 + \beta_{10} X_2X_3 + \beta_{11} X_2X_4 + \beta_{12} X_2X_5 + \beta_{13} X_3X_4 + \beta_{14} X_3X_5 + \beta_{15} X_4X_5$
- All square terms
  - $\beta_{16} X_1^2 + \beta_{17} X_2^2 + \beta_{18} X_4^2 + \beta_{19} X_5^2$
- $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_1X_2 + \beta_7 X_1X_3 + \beta_8 X_1X_4 + \beta_9 X_1X_5 + \beta_{10} X_2X_3 + \beta_{11} X_2X_4 + \beta_{12} X_2X_5 + \beta_{13} X_3X_4 + \beta_{14} X_3X_5 + \beta_{15} X_4X_5 + \beta_{16} X_1^2 + \beta_{17} X_2^2 + \beta_{18} X_4^2 + \beta_{19} X_5^2$