

Unit 5: Linear Model Specification

Chapter 6-7

Introducing Specification

Introducing Specification

- What is Specification?
 - The process of selecting a population model
 - Remember that we begin by assuming our population model is true, then estimating its parameters.
 - You can't measure anything about a population without making at least some assumptions
 - Choosing which variables to include, and which to omit
 - Choosing what transformation to apply to each variable, or what form they should appear in.
- What are our goals when specifying a model?
 - We want our model to be realistic
 - We've seen that we need to make certain assumptions to perform inference.
 - We must test these assumptions, or justify them where appropriate
 - We may want our model to be understandable
 - We'd like to have an intuitive understanding of what our parameters mean
 - We may want our model to make accurate predictions
 - This seems obvious, but there's usually a tradeoff between prediction and understandability.
 - We may also have an interest in highlighting a specific relationship, testing a specific hypothesis, or measuring a specific effect and these goals must guide our specification.

Specifying a Linear Model

- Recall that our linear population model takes the form,
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + u$$
- At first glance, it may seem like we have no choices left to make.
 - Once we decide to use this linear structure, haven't we finished selecting our model?
- But in fact, there is a great deal of flexibility hidden within the linearity assumption.
 - Remember that we have complete flexibility to design our x 's and our y .
 - You can transform a variable, or create entirely new ones from existing variables.
 - You can include the same variable in multiple forms, modeling data with parabolas or other curves.
 - You can allow the value of one variable to alter the effect of another variable.
- In fact, the assumption of linearity is only about how the terms in our regression combine.
 - We're not allowed to do things like raise x_j to the power of β_j . $y = \beta_0 + x_1^{\beta_1}$
 - The β 's can only scale terms we're adding together.
- But this usually isn't much of a limitation.
- Today's lecture is about all the different specification strategies that we can use within the linear model.

More Hypothesis Testing

Hypothesis Testing

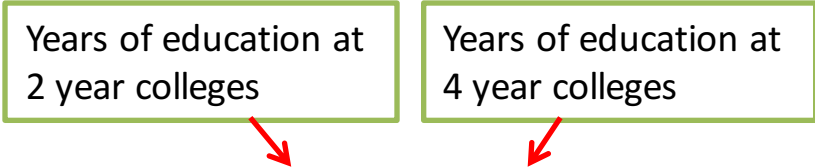
- One of the first factors to consider when choosing a specification is the hypotheses that we want to test.
- Last week, we saw how to test whether a coefficient in our population model was equal to zero.
- We gave conditions under which the coefficients had a normal sampling distribution.
- Since we have to estimate the standard deviation of the coefficient, we use a t-distribution.
- We saw that we can use a simple t-test to see if each coefficient is different than zero.
 - R provides this automatically in regression output.
- In fact, we can use specification to test a much wider variety of hypotheses.

Testing if Parameters are Different

- In this example, we're modeling wage as a function of years spent at a 2-year college (junior college) and years at a 4-year university.
- An interesting question is whether the two coefficients are different
 - Is a year at a 4 year university associated with more of a wage increase than a year at a 2-year college?

Years of education at
2 year colleges

Years of education at
4 year colleges


$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

- We want to test $H_0: \beta_1 - \beta_2 = 0$ against $H_a: \beta_1 - \beta_2 \neq 0$
 - You might be able to argue that a one-sided test is appropriate here, but we'll be conservative and use a two-sided one.
- Here's how we'd like to set up our test statistic
 - The difference in the coefficients divided by the standard error of their difference.
 - We're estimating the standard error, so this statistic has a t-distribution
 - But R won't give us this statistic if we run the linear regression.

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)}$$

Testing if Parameters are Different

- We also can't compute the t-statistic we want from the output in R.
 - We know the numerator, but what's the standard error of the difference in the denominator?
 - Let's use the basic properties of variance to write this out:

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{\widehat{Var}(\hat{\beta}_1 - \hat{\beta}_2)} = \sqrt{\widehat{Var}(\hat{\beta}_1) + \widehat{Var}(\hat{\beta}_2) - 2\widehat{Cov}(\hat{\beta}_1, \hat{\beta}_2)}$$

- R tells us the variance of $\hat{\beta}_1$ and $\hat{\beta}_2$, but not the covariance.
- Luckily, there's another method we can use to test this hypothesis: let's change the variables.
 - Define $\theta_1 = \beta_1 - \beta_2$. Then our null hypothesis, $H_0: \beta_1 - \beta_2 = 0$ is equivalent to $H_0: \theta_1 = 0$
- Now we can write our population model as:

$$\begin{aligned}\log(wage) &= \beta_0 + (\theta_1 + \beta_2)jc + \beta_2univ + \beta_3exper + u \\ &= \beta_0 + \theta_1jc + \beta_2(jc + univ) + \beta_3exper + u\end{aligned}$$

- Now we have a new variable, $jc + univ$, which represents total years in any college.
 - We create this variable in our data table, and then estimate the regression.
 - R then estimates θ_1 for us and also tests whether it's statistically significant. If θ_1 is statistically significant, we can reject our original null hypothesis.

Example

- Here are our results from an OLS regression.
- You can eyeball the coefficient for *jc* – notice that it's definitely less than twice its standard error, so we don't think it'll be significant.

$$\widehat{\log(wage)} = 1.472 - .0102 \textit{jc} + .0769 \textit{totcoll} + .0049 \textit{exper}$$
$$(.021) \quad (.0069) \quad (.0023) \quad (.0002)$$

$$n = 6,763, \quad R^2 = .222$$

$$t = -.0102/.0069 = -1.48$$

$$p = \Pr(|t| > 1.48) = 0.139$$

- Indeed, $p = 0.139$, so we know we can't reject the hypothesis that the effects of either type of college are equal.
- This method can be generalized to test any linear combination of parameters, not just one minus the other.

Joint Significance

Joint Significance


- We've seen how to test if a single coefficient in our OLS regression is different than zero.
- There are times when we want to test several coefficients at the same time.
- In this example from Wooldridge, we're modeling the salary of major league baseball players.
- On the right, we have years in the league, average games per year, and then a number of variables that measure performance: batting average, home runs per year, and runs batted in a year.

$$\log(\text{salary}) = \beta_0 + \beta_1 \text{years} + \beta_2 \text{gamesyr} \\ + \beta_3 \text{bavg} + \beta_4 \text{hrunsyr} + \beta_5 \text{rbisyr} + u$$

- An interesting question is whether performance has an effect on salary.
- We could formulate this as a joint null hypothesis: $H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$
- Economists like Wooldridge might call this an exclusion restriction – we're testing whether the variables could be excluded from the model.
- This means that we're interested in whether, collectively, the three performance indicators are associated with a change in salary.

Joint Significance

- Before we test our joint hypothesis, let's look at the usual regression output.

$$\widehat{\log}(\text{salary}) = 11.19 + .0689 \text{ years} + .0126 \text{ gamesyr} \\ (0.29) \quad (.0121) \quad \quad (.0026) \\ + .00098 \text{ bavg} + .0144 \text{ hrunsyr} + .0108 \text{ rbisyr} \\ (.00110) \quad \quad (.0161) \quad \quad (.0072)$$


Just by eyeballing the standard errors, you can see that none of the 3 performance variables is significant.

$$n = 353, SSR = 183.186, R^2 = .6278$$

- How can we test these coefficients jointly?
- One strategy is to remove them from the regression, and see how much worse the model fit is.
- As a measure of model fit, we can use the sum of squared residuals (SSR)
 - Here, SSR = 183

Joint Significance

- Here's the model with the performance variables taken out – we call this the restricted model

$$\widehat{\log}(\text{salary}) = 11.22 + .0713 \text{ years} + .0202 \text{ gamesyr}$$

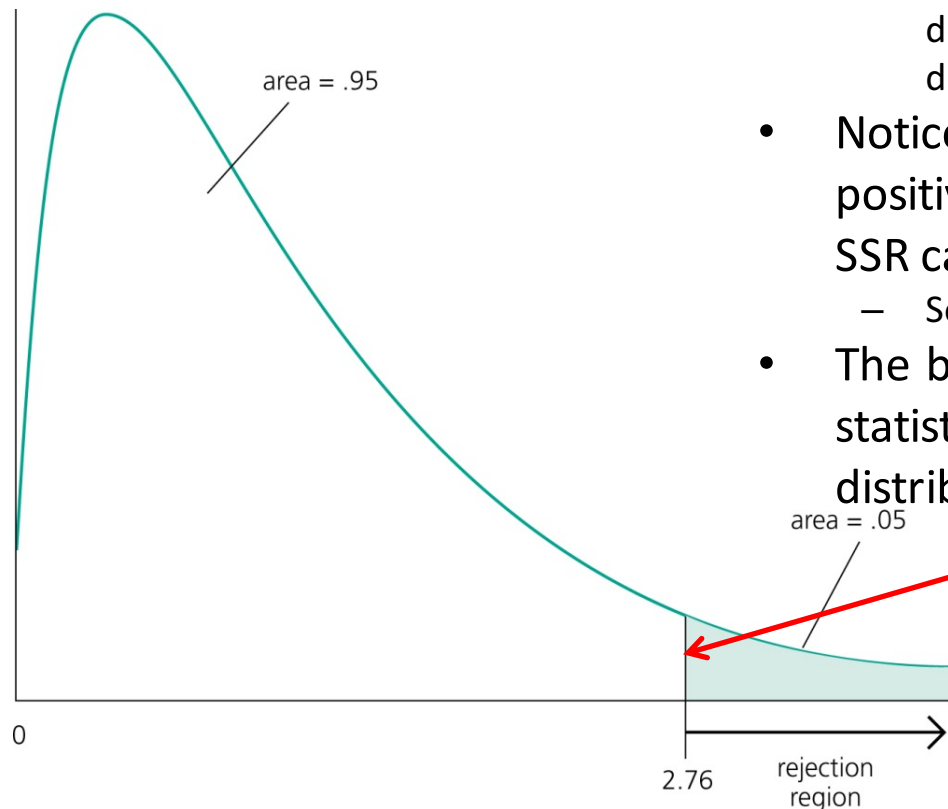
$$(0.11) \quad (.0125) \quad (.0013)$$

$$n = 353, SSR = 198.311, R^2 = .5971$$

- The SSR went up, it's now about 198
 - This is no surprise, taking out variables can only make the fit worse
 - The question is whether the increase is statistically significant.
- We can form a test statistic using the following formula.
- In the denominator, we have the unrestricted SSR, over degrees of freedom in the unrestricted model.
- In the numerator, we have the change in SSR, divided by the number of variables we're testing – you can think of this as the extra degrees of freedom.
- So we're measuring the relative change in SSR, with some constant scaling factors.
- Under the null hypothesis, and assuming the CLM assumptions (MLR.1-6) this test statistic follows a distribution known as the F distribution.

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n - k - 1)} \sim F_{q, n-k-1}$$

The F-Distribution



- There is actually an entire family of F-distributions.
 - To identify a single distribution, we have to specify both degrees of freedom for the numerator and for the denominator.
- Notice that the F-distribution only takes on positive values. This corresponds to the fact that SSR can only increase if we remove variables.
 - So the numerator of our F-statistic will be positive.
- The bigger the increase in SSR, the bigger our F-statistic will be, and the further to the right of the distribution.

As always, we choose the critical value so that the null hypothesis is rejected in 5% of the cases, assuming it is true.
In this case, that critical value is 2.76.

Joint Significance

- Here's our F-statistic for the baseball example.

$$F = \frac{(198.311 - 183.186)/3}{183.186/(353 - 5 - 1)} \approx 9.55$$

- Under an F-distribution with 3 and 347 degrees of freedom, we get a p-value of 4.48×10^{-6} .

$$F \sim F_{3,347} \Rightarrow c_{0.01} = 3.78$$

$$\Pr(F > 9.55) = 4.48 \cdot 10^{-6}$$

- The Null hypothesis can be rejected, even at the 0.001 level.
 - We say that the three variables are **jointly significant**
- Remember that they were not significant when tested individually.
 - The likely reason is multicollinearity between them. Performance metrics tend to move up and down together, so there is no much unique variation for OLS to work with.

Model Significance

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + u$$

- One application of joint significance is the overall test of the regression model
 - Here, we test to see if we can exclude every x variable at the same time.
 - We call this an omnibus test.
 - The null hypothesis is $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$
 - That means that the restricted model is just the mean: $y = \beta_0 + u$.
 - And SSR in the restricted model is just the total sum of squares.
 - We're testing if the model has any predictive power on the whole, or if it could all just be noise.

- We can then rewrite the F-statistic in the more compact form,

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n - k - 1)} = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} \sim F_{k, n-k-1}$$

- This test of overall significance is reported automatically in R
- Most of the time, the null hypothesis is overwhelmingly rejected.
 - If the null can't be rejected, we may have very little data, or we've really chosen variables that don't have predictive power
 - It's also possible for the model to be non significant, but one coefficient to have a significant t-statistic. The conclusion isn't clear here, but the evidence that the coefficient is not zero is marginal at best here.

Variable Transformations

Variable Transformations

- Let's begin with some common techniques for transforming variables.
 - That means that we're replacing some x_j with a function $f(x_j)$.
 - Or replacing y with a function $f(y)$
- By far the most common transformation is the logarithm.
 - It's simple
 - It makes results easy to interpret
 - It can occasionally correct problems with our OLS assumptions

Semi-Logarithmic Form

- If we take the log of our outcome variable, we have what's called semi-logarithmic form.
- This is very common for monetary measures like income and GDP.
- Here's the famous wage equation from labor economics, but we're modeling the log of wage instead of nominal wage.

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

There are two common choices for the base of the logarithm.

- Many researchers use a base 10 log
 - $\log_{10}(y)$ in R.
- This is helpful when we want to think about y in terms of powers of 10.
 - If the right side of the equation is close to, say, 3, the y variable will be about 1,000
 - If a slope coefficient is close to 1, every unit increase in x will multiply wage y by about 10.

The Natural Log

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

- Most people like to think in base 10, but in many circumstances, a better choice is to use the natural logarithm – the log base e.
 - $\log(y)$ in R
- When using a natural log, there's an elegant interpretation for our slope coefficients.

- Take the partial derivative of the population model to get

$$\beta_1 = \frac{\partial \log(wage)}{\partial educ} = \frac{1}{wage} \cdot \frac{\partial wage}{\partial educ} = \frac{\frac{\partial wage}{wage}}{\partial educ}$$

- In the denominator, we have a change in education.
 - In the numerator, we have a change in wage, but divided by the wage level – this is a proportional change in wage.
- So we can think about β_1 as the proportional increase in wage, as a result of an extra unit of education.
 - One nice thing is that if we multiply Y by a constant, say to change units, β_1 doesn't change.
- Let's say $\beta_1 = 0.15$. Then we expect an extra year of education to result in a 15% higher wage.
 - Technically, the changes in our equation are differential changes. Our interpretation is only exact in the limit as our changes become small.
 - But for reasonably small percentage changes (10% or even 20%) the increase in the log is pretty close to the proportional increase in the variable.

Graphing Semi-Logarithmic Form

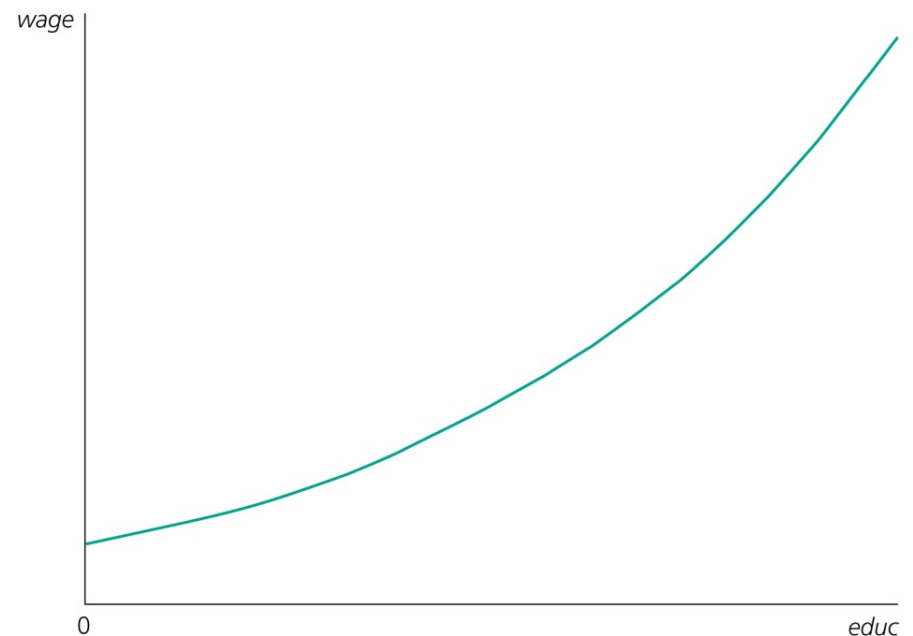
What does semi-log form look like when we graph it?

Take the exponent of both sides of the population model to get

$$wage = e^{\beta_0 + \beta_1 educ} = e^{\beta_0} e^{\beta_1 educ}$$

This tells us that wage is an exponential function of educ.

If $\beta_1 > 0$, wage increases to the right, otherwise it decreases to the right.



Log-Log Form

- There are times when we'll take the log of both our y and x variable. This is what we call log-log form.
- In this population model, the log of a CEO's salary is a linear function of the log of the firm's sales.

$$\log(\textit{salary}) = \beta_0 + \beta_1 \log(\textit{sales}) + u$$

- This changes our interpretation of the regression coefficient
- As we did before, take the partial of the population model to get this equation.
 - Each log results in one over a variable.
 - In the numerator, we have change in salary, as a fraction of the salary level.
 - In the denominator, change in sales, as a fraction of the sales level.
- So we're measuring the percentage increase in salary, per 1% increase in sales
 - This is only strictly true in the limit for small changes, but reasonably close for 10% or 20% changes.

$$\beta_1 = \frac{\partial \log(\textit{salary})}{\partial \log(\textit{sales})} = \frac{\frac{\partial \textit{salary}}{\textit{salary}}}{\frac{\partial \textit{sales}}{\textit{sales}}}$$

Log-Log Form

- Here's a fitted regression for our salary equation.
 - Each percentage increase sales is associated with a .257% increase in salary.

$$\widehat{\log}(\text{salary}) = 4.822 + 0.257 \log(\text{sales})$$

- **Aside:** In economics, the coefficient in a log-log model has a special name: the elasticity
 - Elasticity is the slope of a log-log plot of y against x.
 - By choosing the log-log form, we're assuming a constant-elasticity relationship, instead of a constant-slope relationship

Logarithm Rules of Thumb

- When should you take the log of a variable? Here are some rules of thumb:
- Look for variables that are naturally always positive.
 - The log is not defined for negative numbers.
 - Do **NOT** add a constant to make your variable positive, then take the log.
 - This breaks the intuitive properties of the log and your coefficients become very difficult to interpret.
- Look for variables that have a meaningful zero-point but no obvious maximum
 - Wage and Income do.
 - For year and temperature, we might disagree about the best zero point, so you probably shouldn't take logs of them.
- Look for variables where a percent change is meaningful.
 - We have a good sense of what a 20% increase in wage means.
 - A \$1 increase in wage is very different if you live in the US versus in Mali.
- Taking logs mitigates the influence of outliers in the positive direction.
 - These points are “squashed” downwards.
 - This is useful, again, for variables like income with some large outliers
- Taking logs can help to secure normality and homoscedasticity for OLS
 - However, this should only be a concern for small samples when you can't rely on asymptotics.
 - If you have a large sample, the decision of whether to take a log should be guided by what's more intuitive, or what gives a better model fit.

Other Transformations

- There are a few other transformations that are worth mentioning.
 - Quadratic / higher order polynomials
 - i.e., regress Y on X^2
 - Occasionally, you may see powers less than 1.
 - This corrects negative skew if you need a normal variable distribution
 - Indicator functions convert metric variables to binary ones.
 - We may assign a value of 1 whenever a variable is greater than its mean value.
 - This sometimes helps us deal with nonlinearities and may result in a closer model fit.
 - The logit function takes variables that are bounded by $[0,1]$ and maps them to the entire real line.
 - A common application is in modeling probabilities as an outcome variable.
 - This is the idea behind logistic regression.
- We'll look at some of these next.

Quadratic and Polynomial Specifications

Quadratics and higher-order polynomials

- Sometimes we want a variable to be entered into our regression as a polynomial
 - i.e., not just as x , but as x^2 or x^3 and so on
- This may be because
 - Our guiding theory requires it
 - Previous data suggests that the relationship is nonlinear
 - We want to allow the effect of x on y to change with the value of x
 - We're looking to reduce error in our prediction by allowing functional form to be flexible
 - Especially in control variables, where we care more about removing the effect than in understanding it on an intuitive level.
- A few guidelines:
 - When we have a quadratic form, OLS will find the parabola with the least squared error. For a cubic form, OLS will find the best cubic function, and so on.
 - Fit always improves when you add a higher-order term.
 - Make sure you include all lower-order terms
 - i.e., if you're running x^2 , make sure x is in there too
 - We'll need to interpret all coefficients simultaneously to understand the effect of the variable.

Quadratic Form Example

- Here's an example, a fitted wage equation, using both experience and experience squared as predictors.
- Notice that the coefficient for $exper^2$ is negative. This means that the function is **concave**
- Every extra year of experience adds less to the wage than the previous year.

$$\widehat{wage} = \underset{(.35)}{3.73} + \underset{(.041)}{.298} \text{ exper} - \underset{(.0009)}{.0061} \text{ exper}^2$$

$$n = 526, R^2 = .093$$

- We can see this by taking the derivative of both sides:

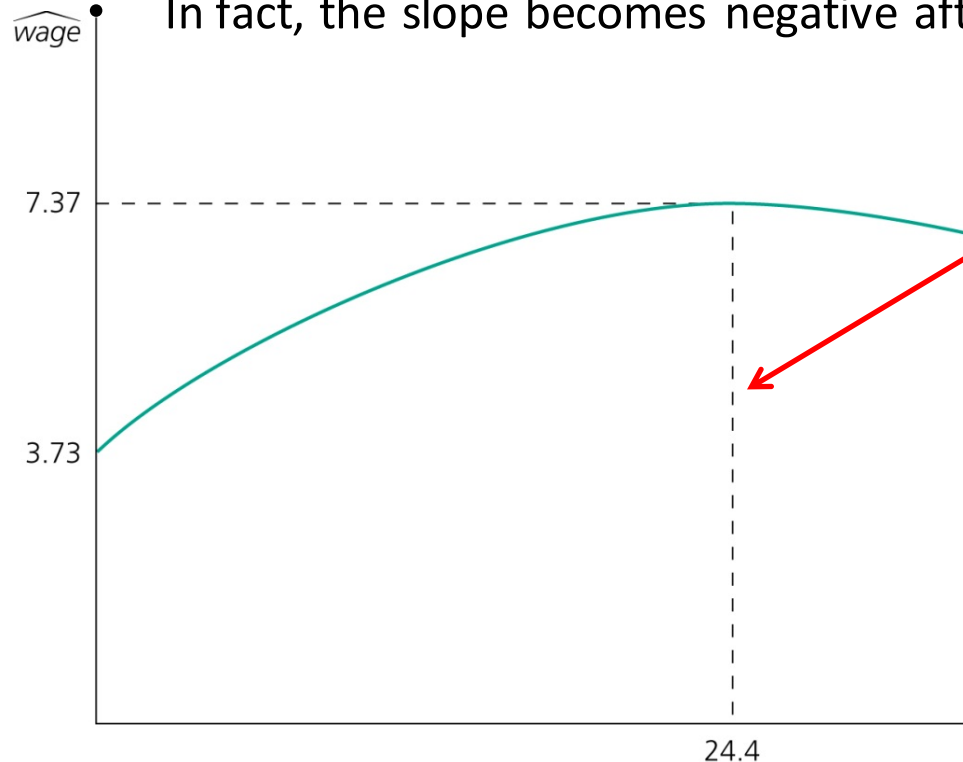
$$\frac{\partial wage}{\partial exper} = .298 - 2(.0061)exper$$

- Exper appears in the equation for the slope – and the slope is decreasing.
- The first year of experience increases wage by about .30\$, the second year by $.298 - 2(.0061)(1) = .29\$$ etc.
 - We say that experience has decreasing marginal benefit.

Extrapolating to Extreme Values

- Here's a graph of the fitted quadratic wage model. You can see the concavity.

- In fact, the slope becomes negative after 24.4 years.



$$x^* = \left| \frac{\hat{\beta}_1}{2\hat{\beta}_2} \right| = \left| \frac{.298}{2(.0061)} \right| \approx 24.4$$

Does this mean that your wage will start falling after you work for 24.4 years?

Not necessarily. For one thing, we're fitting a parabola to real data, and there may not be many data points above 24.4 years.

We also don't know if our model specification is correct. **Any** concave parabola will have negative slope after some point. So assuming a parabolic population model pretty much guarantees this will happen.

This example shows that we need to be careful when extrapolating to extreme values.

Indicator Variables

Qualitative data

- Many variables we want to study are categorical in nature
 - Examples: gender, race, industry, region, letter grade...
 - These take on a limited number of values, assigning each observation to a “category”.
 - In experimental traditions and in R, these variables are called factors.
- Categorical variables may be nominal or ordinal
- A nominal variable is a categorical variable with no sense of ordering between the categories
 - Hair color – there’s no clear way to order red, blonde, black.
 - Race, industry, etc.
- An ordinal variable is a categorical variable in which the categories have an ordering. These may also be called rank variables, since there’s information in the ranking
 - We might record education with the levels, “high school graduate”, “some college”, “college graduate” We know that “college graduate” represents more education than “some college.”
 - But notice that the intervals between categories aren’t equivalent. The jump from high school graduate to some college may not be the same as the jump from some college to college graduate.
 - We say that this variable doesn’t have an interval structure
 - If the intervals were the same, we’d really have a quantitative variable - we could assign integers to each category since adding 1 would always mean the same thing.
 - Since we don’t have an interval structure, we shouldn’t represent this variable as numerical in our population model
 - Our linear model would force the effect of moving from high school to some college to equal the effect of moving from some college to college graduate.

Incorporating Qualitative Data

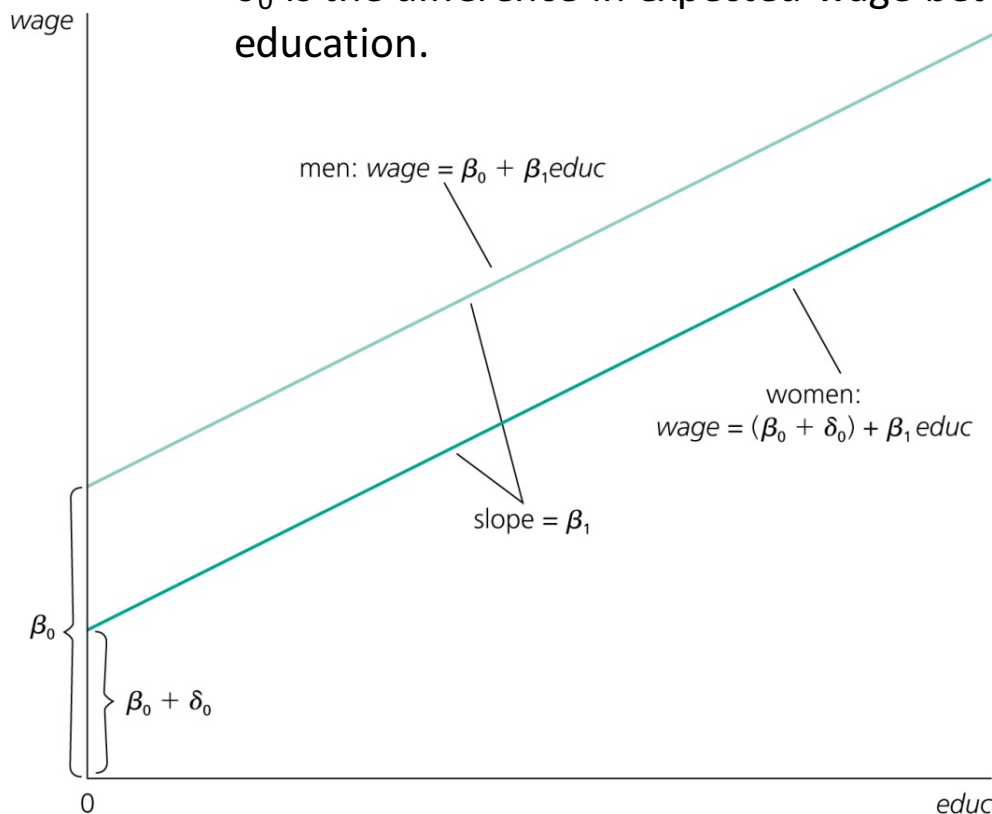
- To put categorical variables into our regression model, we typically use indicator variables (AKA dummy variables).
 - An indicator variable takes on the value 1 for certain states (e.g., winning an election, clicking on an advertisement, being over 6 feet tall) and 0 otherwise.
- Here's a population model that predicts wage as a function of education again, but we've added an indicator variable for female.

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u$$

- This means that for a male subject with a given value of *educ*, the expected wage will be $\beta_0 + \beta_1 educ$.
- For a female subject, the expected wage will be $\beta_0 + \delta_0 + \beta_1 educ$.

Qualitative data

- If we graph the expected wage for men and for women, we get two lines.
 - Notice that both lines have the same slope, β_1 .
 - If we assume this population model and then fit it to data, we should check to see how realistic this assumption is.
 - The intercepts are different. The intercept for men is β_0 , and for women it's $\beta_0 + \delta_0$.
 - In fact the two lines are always δ_0 apart.
 - δ_0 is the difference in expected wage between men and women of the same level of education.



$$\delta_0 = E(wage|female = 1, educ) - E(wage|female = 0, educ)$$

Omitting the Base Category

- Notice that we didn't include indicator variables for both male and female.

$$wage = \beta_0 + \gamma_0 \text{male} + \delta_0 \text{female} + \beta_1 educ + u$$

- If we did this, we wouldn't be able to estimate the model because they would be perfectly collinear.

- So one category always has to be omitted, we call this the base category. We could have chosen either male or female, and the model would be equivalent..

$$wage = \beta_0 + \delta_0 \text{female} + \beta_1 educ + u$$

$$wage = \beta_0 + \gamma_0 \text{male} + \beta_1 educ + u$$

- We could also leave both categories by omit the intercept
 - But if you do this, it's harder to test if the categories are different.
 - Also, the usual formula for R-squared is no longer valid.

$$wage = \gamma_0 \text{male} + \delta_0 \text{female} + \beta_1 educ + u$$

Interpreting coefficients

- Here's a fitted wage equation including the female indicator variable.

$$\widehat{wage} = -1.57_{(.72)} - 1.81_{(.26)} female + .572_{(.049)} educ \\ + .025_{(.012)} exper + .141_{(.021)} tenure$$

$$n = 526, R^2 = .364$$

Holding education, experience, and tenure fixed, women earn 1.81\$ less per hour than men

Comparing Group Means

- As a special case, we sometimes want to compare the mean of a variable for two different groups.
 - In that case, we can put an indicator variable for one of the categories in our population model by itself.

$$\widehat{wage} = \begin{matrix} 7.10 & - & 2.51 & female \\ (.21) & & (.26) \end{matrix}$$

$$n = 526, R^2 = .116$$

Not holding other factors constant, women earn 2.51\$ per hour less than men, i.e. the difference between the mean wage of men and that of women is 2.51\$.

- The t-statistic in this case is a test of whether two group means are equal
 - The t-test to compare group means is often presented in intro statistics classes. The test here is exactly the same, only we've placed it in a regression framework.

Treatment as a Dummy Variable

- In an experiment, we would randomly assign subjects to a control group and one or more treatment groups.
- In this case, it would be natural to treat the control group as a base category, and create dummies for each treatment.
- Suppose we run a clinical trial, in which subjects are randomly assigned to take a new blood pressure medication or a placebo. Our model might be
$$\text{Blood pressure} = \beta_0 + \beta_1 \text{medication} + u.$$
- β_1 would represent the difference in blood pressure between treatment and control, and our usual t-test would test the hypothesis that treatment has no effect.

Ordinal variables

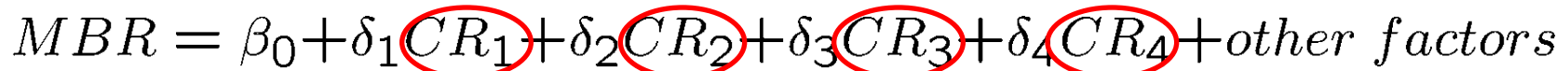
- How do we put ordinal variables into a regression model?
 - Remember that ordinal variables are categorical variables with a natural ordering. They may even be encoded in your dataset as 1,2,3,4...
- Generally speaking, it would be wrong to place an ordinal variable that's coded this way directly into your population model.
 - This would impose / assume a linear structure on the variable
 - i.e., the change in y from going from one category to the next would always be the same.
- So for the most part, we will use indicator variables for each category, allowing the effect of each one to vary independently.

Municipal bond rate

Credit rating from 0-4 (0=worst, 4=best)


$$MBR = \beta_0 + \beta_1 CR + other\ factors$$

This specification would probably not be appropriate as the credit rating only contains ordinal information. A better way to incorporate this information is to define dummies:


$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + other\ factors$$

Dummies indicating whether the particular rating applies, e.g. $CR_1=1$ if $CR=1$ and $CR_1=0$ otherwise. All effects are measured in comparison to the worst rating (= base category).

Interaction Terms for Indicator Variables

The middle 4 slides of this oyster should be replaced by the lightboard we already filmed on the topic.

Interaction Terms

- Consider the partial effect of a predictor x_j on y : $\partial y / \partial x_j$
- If we put x_j into our linear population model directly, the marginal effect of x on y is just a constant.
- But we've been exploring ways to transform x_j , and we've seen that the marginal effect of x_j can depend on the value of x_j .
- We might want even more flexibility than this: we might want the effect of x to depend on the values of other variables.
- Here are some motivating examples, some scenarios that we'd like to model with OLS:
 - The health risks of a genetic marker may differ by gender
 - The effect of a teacher training program may differ by grade level
 - The effectiveness of a promotion on customer retention may differ by contract length.
- These are all scenarios in which one variable changes the effect of another variable.
- The main way that we account for effects like these is with interaction terms.
- An interaction term is a term in a regression where two independent variables are multiplied together
- It's a term like $\beta_k x_i x_j$ where $i \neq j$.

$\partial y / \partial x_j = \beta_k x_i$

 - Notice that the effect of x_j on y is now linear in x_i .
 - So this is a pretty simple way to make one variable's effects depend on another variable, but it often helps to make population models more realistic.
 - As we'll see, we'll have to be careful in interpreting our coefficients. Our understanding will depend on whether we have continuous or indicator variables.

Interacting Dummy Variables

- The simplest interaction term to understand is one that has two indicator variables.
- Here's a model in which blood pressure depends on treatment by a new drug, and on a dummy for male.
- $Bp = \beta_0 + \beta_1 \text{drug} + \beta_2 \text{male} + u$.
- Notice that we did not include an interaction term. Here's a table that shows the four groups of people and the expected blood pressure for each one:
- The effects of each dummy are independent, so notice that if we move, say, from female to male, we add β_2 to our expected output, whether we're in the control or the drug group.
- Put another way, if we know what the expected values are for three of the groups, we can use arithmetic to compute the expected value for the last group.
 - This are “linked” through our additive structure.
 - This also means that if we fit this model to real data, our estimated means will not equal the actual means for each group, because the means in our sample are unlikely to follow this additive structure.

	Female	Male
Control	β_0	$\beta_0 + \beta_2$
Drug	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2$

Interacting Dummy Variables

- Now let's add an interaction term:
- $B_p = \beta_0 + \beta_1 \text{drug} + \beta_2 \text{male} + \beta_3 \text{drug} \times \text{male} + u.$
- Our table of expected values looks the same, except we've added a β_3 in the lower right corner.
- This new parameter represents an extra degree of freedom – that means that we can model any set of group means with an appropriate choice of β 's.
- This is what we call a saturated model – a model with all possible interaction terms included so that all groups means can vary independently
- Let's see this with an example...

	Female	Male
Control	β_0	$\beta_0 + \beta_2$
Drug	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

Interacting Dummy Variables

- Suppose these are our observed means from our study.
- We know from the upper right that $\beta_0 = 100$.
- To get β_2 , we subtract the Female-Control average from the Male-Control average to get $90 - 100 = -10$.
 - Intuitively, β_2 represents the effect of Male when the other variable is in its base category.
- To get β_1 , we subtract Female-Control from Female-Drug to get $110 - 100 = 10$.
 - Again, β_1 represents the effect of Drug when the other variable is in its base category.

	Female	Male
Control	$\beta_0 = 100$	$\beta_0 + \beta_2 = 90$
Drug	$\beta_0 + \beta_1 = 110$	$\beta_0 + \beta_1 + \beta_2 + \beta_3 = 120$

Interacting Dummy Variables

- Now to get β_3 , we could do two things.
- Let's first look at the marginal effect of being Male.
 - In the control group, this is $\beta_0 + \beta_2 - \beta_0 = \beta_2 = 90 - 100 = -10$
 - In the drug group, this is $\beta_0 + \beta_1 + \beta_2 + \beta_3 - \beta_0 + \beta_1 = \beta_2 + \beta_3 = 120 - 110 = 10$
- Now we take the difference of those marginal effects: $\beta_2 + \beta_3 - \beta_2 = \beta_3 = 10 + (-10) = 20$.
- We could have done this the other way around.
- Look at the marginal effect of the drug.
 - In the Female group, this is $\beta_0 + \beta_1 - \beta_0 = \beta_1 = 110 - 100 = 10$.
 - In the Male group, this is $\beta_0 + \beta_1 + \beta_2 + \beta_3 - \beta_0 + \beta_2 = \beta_1 + \beta_3 = 120 - 90 = 30$.
- Again, take the difference of the marginal effects to get,
 - $\beta_1 + \beta_3 - \beta_1 = 30 - 10 = 20$.
- We get the same number. So the interaction effect is both the change in the effect of drug from being Male, as well as the change in the effect of being male from taking the drug.

	Female	Male
Control	$\beta_0 = 100$	$\beta_0 + \beta_2 = 90$
Drug	$\beta_0 + \beta_1 = 110$	$\beta_0 + \beta_1 + \beta_2 + \beta_3 = 120$

Three-Way Interactions

- Interaction terms don't just have to include two variables. There are times when we want to interact 3 variables together
 - We could even do more, but it's hard enough to understand the meaning of the 3-way coefficient, so we usually stop there.
- In the drug example, say we also record whether a participant exercises in a dummy variable, *exer*
- The 3-way interaction, *drug*×*male*×*exer*, is a variable that only equals 1 if all three indicator variables are 1.
- Recall that when we include a power of one variable, we should also include all lower-order powers.
- Similarly, when we include an interaction term, we must include all interactions that include subsets of those variables.
- To include our 3-way interaction, our model would look like:
$$\begin{aligned} B_p = & \beta_0 + \beta_1 \text{drug} + \beta_2 \text{male} + \beta_3 \text{exer} \\ & + \beta_4 \text{drug} \times \text{male} + \beta_5 \text{drug} \times \text{exer} + \beta_6 \text{male} \times \text{exer} \\ & + \beta_7 \text{drug} \times \text{male} \times \text{exer} + u. \end{aligned}$$
- In this case, this is the saturated model, since we've included all variables and their interactions
 - There are now 8 groups of people described by our dummy variables and each one can have any expected blood pressure without any restriction.
- The coefficient β_7 must be interpreted cautiously. If it's non-zero, there may be a unique combination of the dummies – say women who exercise but don't take the drug – that stands out for having an unusually high or low mean value. It could be any of the 8 categories here.
 - Or there may be more than one combination that has a high or low value. In general, this just means that there are combinations that can't be predicted based on two-way interactions and something more complex is going on.

Interaction Terms with an Indicator and a Metric Variable

Interaction Terms with an Indicator and a Continuous Variable

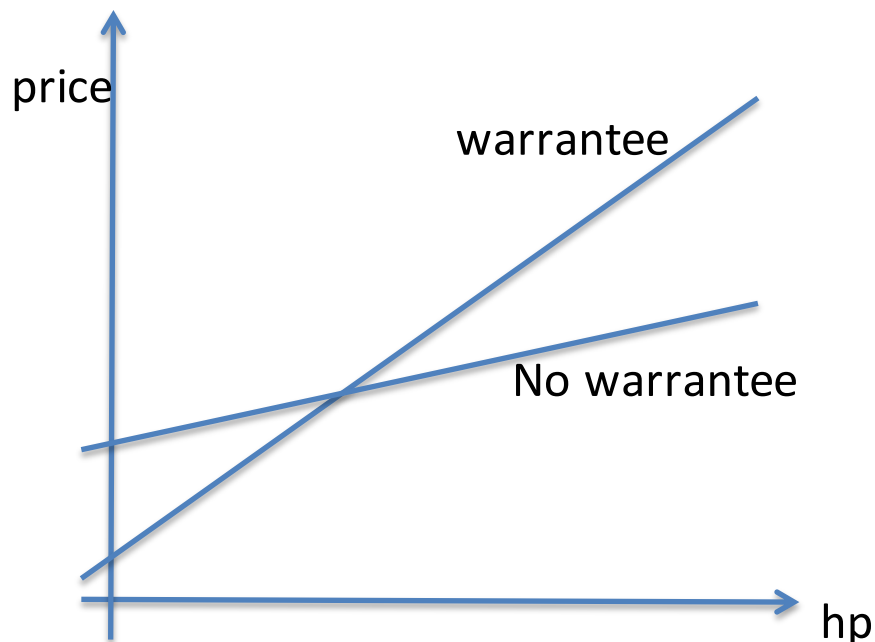
- Let's see what happens when we interact a continuous variable with an indicator.
- Our example will be this model of car price as a function of horsepower and a dummy for warrantee.
- $\text{Price} = \beta_0 + \beta_1 \text{hp} + \beta_2 \text{warrantee} + \beta_3 \text{hp} \times \text{warrantee} + u$.
- As we saw before, if we didn't put the interaction term in, β_2 would act as an intercept shift
 - We'd get two lines with the same slope, but with two different intercepts to fit each group as closely as possible.
- Now look at the slope of price with respect to hp:

$$\partial \text{price} / \partial \text{hp} = \beta_1 + \beta_3 \text{warrantee}$$

- Notice that it takes on two values: β_1 for cars with no warrantee, and $\beta_1 + \beta_3$ for cars with a warrantee.

Different Slopes

- Here's a graphical illustration of our model. We have one line for the cars with warranty and a separate line for those without.
- The interaction term allows both the intercepts and slopes to vary independently for the two groups – not just the intercept as we had previously.
- The interaction coefficient is the difference between these two slopes.
- When we look at real data, OLS will fit the best line to the warranty group, and the best line to the no warranty group.



- An interesting hypothesis we could test is whether the two groups have the same slope.
- This would mean that our interaction coefficient is zero, $\beta_3 = 0$.
- This is simply the standard t-test for that coefficient that's included with our R regression output.

Testing for Equal Lines

- We might also want to test if our two groups have the exact same lines – that is, the same slopes and the same intercepts.
- Our hypothesis is $\beta_2 = \beta_3 = 0$.
- To test this, we need to run an F-test between the full model, and one without the terms including warrantee
- Full model (contains full set of interactions)
$$\text{Price} = \beta_0 + \beta_1 \text{hp} + \beta_2 \text{warrantee} + \beta_3 \text{hp} \times \text{warrantee} + u.$$
- Restricted model (same regression for both groups)
$$\text{Price} = \beta_0 + \beta_1 \text{hp} + u.$$
- We compute an F-statistic to see if the saturated model yields a statistically significant improvement in model fit.

Interaction Terms between Metric Variables

(maybe for Jeffrey)

Interaction Terms between Metric Variables

- We've talked about interaction terms with two indicator variables, or one metric variable and one indicator variable. We can also create an interaction between two metric variables.
- This case is a bit less intuitive, but it's an important one to understand.
- Here's an example from Wooldridge:

$$\log(\text{price}) = \beta_0 + \beta_1 \text{sqrft} + \beta_2 \text{bdrms} \\ + \beta_3 \text{sqrft} \cdot \text{bdrms} + \beta_4 \text{bthrms} + u$$

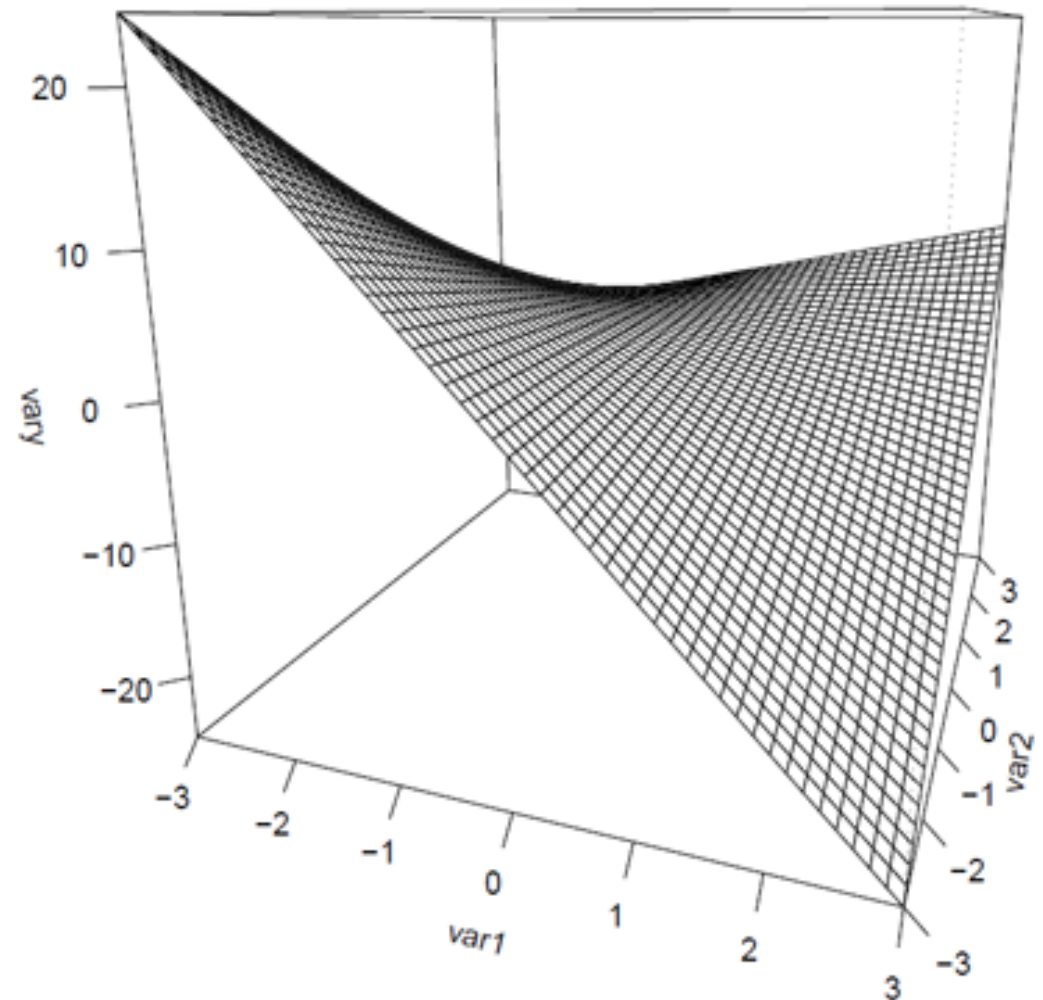
- We're modeling the log of the price of a home as a function of number of square feet, bedrooms, and bathrooms, and we've added an interaction term between square feet and bedrooms.
- Consider the effect of the number of bedrooms:

$$\Rightarrow \frac{\partial \log(\text{price})}{\partial \text{bdrms}} = \beta_2 + \beta_3 \text{sqrft}$$

- This slope varies continuously with the number of square feet.
- Now β_2 represents the effect of number of bedrooms, but only for a square footage of zero

Interaction Terms between Metric Variables

- Here's a graph that shows what a model with an interaction term could look like.
- For any one value of var1, you can look at the cross section and see that y is linear in var 2.
 - But the slope of y with respect to var2 changes for different values of var 1.
- The same is true in reverse if we fix different values of var 2.
- Notice that the resulting plot has curvature, but only if we travel in a direction that's not along an axis. This is because the interaction term is parabolic if we increase both var 1 and var 2.
- Graphic from http://www3.i-med.ac.at/genepi/images/stories/projects/interaction_figure1a.png



Reparametrization of interaction effects

- In the home price example, we found that our coefficient was the effect of number of bedrooms for a house with zero-square feet.
- That's not the most intuitive statistic, so we sometimes normalize our variables before fitting a regression.
- Suppose our original model has two variables and an interaction term

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$

- We subtract the mean of each variable in the interaction term to get a model like.

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + u$$

- These models are equivalent for an appropriate choice of coefficients.
- But now δ_1 and δ_2 represent the effects of x_1 and x_2 at their mean values.
 - Another advantage is we get standard errors and significance levels for partial effects at the mean values.
 - If desired, interaction may be centered at other interesting values

Guidelines for Polynomial and Interaction Terms

When to Include Interactions and Higher Powers

- When should we include interaction terms and higher powers of our variables?
- This is a question of research context. First, are we primarily interested in inference or prediction?
- If we want to improve our understanding of the system we're studying, we have to look to our guiding theory.
 - Do we have a theoretical reason to expect an interaction effect?
 - Are we interested in measuring the interaction effect, or are we most interested in measuring the average effect of one variable?
 - What hypotheses would be most useful for us to test?
 - Can our audience understand the interaction coefficients we estimate? (especially for 3-way interactions)
- If prediction is at least somewhat important to us, we may look to the data to see if adding terms improves our models.
 - And there are statistical tests that can help us with this.

Specification tests

- First, If we have specific interaction or higher power terms in mind, we can directly test if they improve the fit of the model (more than we'd expect by chance)
 - We would run an f-test between the model including our terms and the restricted model without them.
- There is also a general way to see if our model might be missing interaction and higher-order terms: the regression specification error test (RESET)
 - The idea of RESET is that instead of testing potential terms directly, we use the residuals in our regression to stand in for unexplained factors.
 - So we run a regression with our restricted model, and compute residuals, \hat{y} . We then put the square of the residuals and the cube into our model as new variables, and test to see if model fit is improved with anova.
 - If our F-test is significant, we have evidence that we've omitted higher order terms or interaction terms.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_3 \hat{y}^3 + error$$

- These are statistical tests, so you have to evaluate the results in context. They say nothing about the practical significance of your interaction effect.
 - If you have a large data set of, say, a few thousand observations, the RESET test will almost always be significant, but that doesn't mean that you should necessarily add more terms to the model.
 - Consider your sample size when you look at tests like these.
- A further concern arises with wide datasets having more than a dozen or so variables.
 - Some interaction terms are likely to appear significant due to chance, even if none are significant, so there is the potential for a fishing expedition. This is a central concern of machine learning, and you'll find many techniques for handling it in that class.