1. Scaling and Standardizing Variables

A. Scaling

i. Scaling the Dependent Variable

Let's look at sleep data, where we have the following population regression equation in mind:

$$sleep = \beta_0 + \beta_1 educ + \beta_2 age + u$$

Here are the results from our estimation :

$$\widehat{sleep} = 3315.574 - 12.189 educ + 2.7454 age$$

where *sleep* is sleep measured in minutes per week, *educ* is years of education, and *age* is age. What's the one-sentence size interpretation of the coefficient on education? **One more year of education is estimated to decrease predicted sleep by 12.189 minutes per week, holding age constant.**

Minutes per week may not be the most interesting measure of sleep, and you might wonder what the regression results would look like if we had changed our dependent variable to hours a week instead. We can rewrite that one-sentence interpretation in terms of *hours* of sleep instead of minutes: **One more year of education is estimated to decrease predicted sleep by $\frac{12.189}{60} = 0.2$ hours per week, holding age constant.**

So in essence, we could rewrite all of our interpretations in terms of hours instead of minutes by dividing the old coefficients by 60.

$$\frac{\hat{\beta}_0}{60}, \frac{\hat{\beta}_1}{60}, \frac{\hat{\beta}_2}{60}$$

Alternatively, it's often easier to just run the regression with our *sleep* variable in terms of hours i.e dividing the dependent by 60 and re-running the regression. We get the following results

$$\widehat{sleep} = 55.260 - .2032 educ + .0458 age$$

In general, if we rescale the dependent variable, y, by a factor α, then the equation we estimate becomes:

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + ... + \tilde{\beta}_k x_k + u$$
$$\alpha y = \alpha\beta_0 + \alpha\beta_1 x_1 + ... + \alpha\beta_k x_k + u$$

In the above example, $\alpha = \frac{1}{60}$, so the new $\hat{\beta}$s will be divided by 60 too. Note that nothing else about the regression changes (R^2, t-stats, p-values, etc.).

ii. Scaling the Independent Variable

We do this slightly differently if we scale one of the x variables instead of y. Suppose we'd rather think about education in units of 6 months (for some odd reason). Our initial estimates indicate that if education increases by a whole year (still holding age fixed), $\widehat{sleep}$ decreases by 12.189 minutes per week. This is clearly equivalent to saying that if education increases by 6 months, $\widehat{sleep}$ decreases by 6.095 minutes per week. I.e we can just divide our $\hat{\beta}$ estimate by 2:

$$\frac{\hat{\beta}_1}{2}$$

We can also re-scale the education variable in R to be in units of half-years (or 6 months) and get:

$$\widehat{sleep} = 3315.574 - 6.095 educ + 2.7454 age$$

where *educ* is now in terms of six months. Thus, the intercept and slope coefficeint on age are unchanged, but the coefficient on education is half of that on education.

Generally, if we scale x by α, the equation becomes:

$$\begin{aligned} y &= \beta_0 + \tilde{\beta}_1 \tilde{x}_1 + ... + \beta_k x_k + u \\ &= \beta_0 + \frac{\beta_1}{\alpha}(\alpha x_1) + ... + \beta_k x_k + u \end{aligned}$$

In the above example, we had $\alpha = 2$, which meant we had to scale our estimate of $\hat{\beta}_{educ}$ by $\frac{1}{2}$.

$$\widehat{sleep} = \hat{\beta}_0 + \left(\frac{1}{2}\hat{\beta}_1\right) 2educ$$

B. Standardizing

Standardizing variables eliminates the units in order to be able to compare the magnitude of estimates across independent variables. For example, if I wanted to compare $\hat{\beta}_{educ}$ to $\hat{\beta}_{age}$, I would be comparing a number that is in (minutes per week)/(years of education) units to a number that is in (minutes per week)/(years of age) units. We can solve this issue by standardizing the variables.

Suppose we have a regression with two variables, x_1 and x_2:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{u}$$

We know that our regression must go through the point of averages; or think if we average the previous equation, and use the fact that the $\hat{u}_i$'s have a zero sample average; or if we plugged in $\bar{x}_1$ and $\bar{x}_2$, we would predict $\bar{y}$ [1] :

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2$$

We can subtract the second equation from the first to get:

$$\begin{aligned} y - \bar{y} &= (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{u}) - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2) \\ &= \hat{\beta}_1 (x_1 - \bar{x}_1) + \hat{\beta}_2 (x_2 - \bar{x}_2) + \hat{u} \end{aligned}$$

Now we can throw some algebra at this to get it into standard units:[2]

$$\left(\frac{y - \bar{y}}{\hat{\sigma}_y}\right) = \frac{\hat{\sigma}_{x_1}}{\hat{\sigma}_y}\hat{\beta}_1 \left(\frac{x_1 - \bar{x}_1}{\hat{\sigma}_{x_1}}\right) + \frac{\hat{\sigma}_{x_2}}{\sigma_y}\hat{\beta}_2 \left(\frac{x_2 - \bar{x}_2}{\hat{\sigma}_{x_2}}\right) + \frac{\hat{u}}{\hat{\sigma}_y}$$

Now we can say that controlling for x_2, a one standard deviation increase in x_1 leads to a $\frac{\sigma_{x_1}}{\sigma_y}\hat{\beta}_1$ standard deviation increase in predicted y. We call this new term the **standardized coefficient** or

[1] This follows from the OLS first order conditions when we remember that the residuals are defined by $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. In other words $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to make the residuals add up to zero. From this $\sum y_i = \sum \hat{y}_i$ and dividing both sides by $1/n$ you get that $\bar{\hat{y}} = \bar{y}$

[2] What we're really doing: dividing both sides of this equation by the standard deviation of y, σ_y, and multiplying each independent variable x_k by $1 = \frac{\sigma_x}{\sigma_x}$.

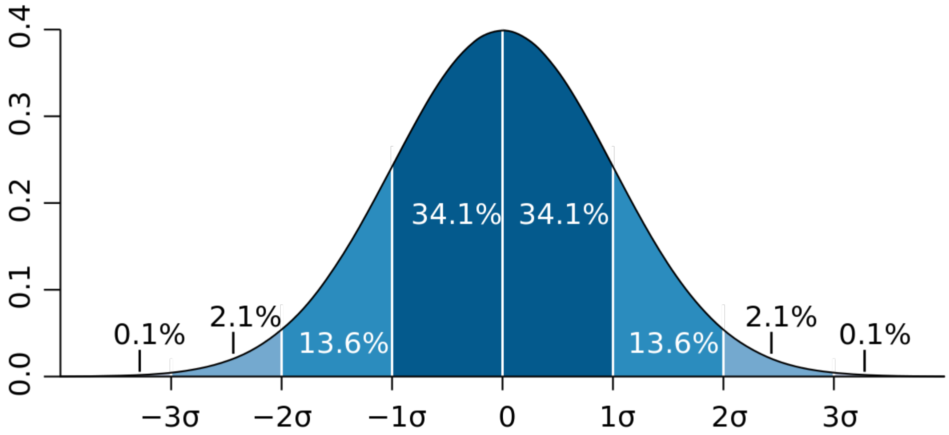
beta coefficient. In R, we can get these coefficients by performing the “scale()” function on all of our variables when we enter them in our “lm()” equation. The scale function standardizes each variable, so that you do not have to do it manually. For example, if we wanted standardized coefficients for our regression of sleep on age and education:

```
Call:
lm(formula = scale(sleep) ~ scale(educ) + scale(age), data = sleep75)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5746 -0.5290  0.0047  0.5869  3.1118

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.909e-16  3.743e-02   0.000   1.0000
scale(educ) -7.638e-02  3.886e-02  -1.966   0.0497 *
scale(age)   7.007e-02  3.886e-02   1.803   0.0718 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9946 on 703 degrees of freedom
Multiple R-squared:  0.01359,   Adjusted R-squared:  0.01078
F-statistic: 4.842 on 2 and 703 DF,  p-value: 0.008155
```

With these standardized results, we see that a one standard deviation increase in either years of education or years in age (holding the other fixed) will decrease/increase predicted minutes of sleep per week by about $-0.08/0.07$ standard deviations, respectively.

Note that when you standardize the variables, this mechanically makes it such that the intercept is zero. We can see this is the case in our regression above.

What do standard deviations mean? They are a measure of the variance/spread of your variable.



- 68.2% of data are within 1 SD
- 95.4% are within 2 SD
- 99.7% within 3 SD

2. Functional Form and Adjusted R^2

A. Choosing between functional forms/non-nested models

We've gone over how to do the following:

1. **Deciding if one of your x variables significantly affects the y variable in your estimation—for this we used a t-test.**
2. **Deciding if multiple variables *together* significantly affect the y variable in your estimation—for this we used an F-test.**

But these tests compare *nested* models (nested models refer to the case where one equation is just a special case of the other). What about *non-nested* models? This could two population regression equations, that differ by one variable - one has mother's education while the other has father's education - or that differ by the functional form for a variable - one has log(age) and one has age plus age squared.

Choosing between functional forms can involve a few considerations:

- Do we have any theoretical reason to believe a particular functional form is correct? For example, is there some biological or mechanical reasons that some independent variable x has a particular functional relationship to a dependent variable y?
- Which form fits the data best? If we aren't sure about what functional form is correct, the $Adj - R^2$ can help us understand which form does the best job of representing the data.
- What does the density of your X variable look like?
  - For example, logs place higher weight on lower values, so they are useful for X variables with long right tails. They are undefined for X$\leq$0.
- There are lots of functional forms you can consider, but it's rare to go beyond linear, logs, or quadratic (sometimes higher-order polynomial), or including interaction terms in your model.
- Note that you make decisions about each of your independent variables. They do not all need to have the same relationship with your dependent variable.

B. Adjusted-R^2

Let's focus on making comparison based on **how well the model fits the data**. You might look to R^2 to compare *non-nested* models. Indeed, R^2 is a measure of the "goodness of fit", or how well our regression line fits the data (it is the proportion of variation in our dependent variable, y, that is explained by our model, $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$) . This is a good instinct, but there's one problem with doing this: The usual R-squared doesn't penalize more complicated models. Since the SSR never goes up and usually falls as more independent variables are added, R-squared never falls when a new variable is added to the regression. So you could just keep adding variables to the model and artificially inflate your R^2.

What's called the *adjusted* R^2 (denoted $\bar{R}^2$) accounts for this problem, and we use this value to compare non-nested models instead of the R^2, t-tests, or F-tests. Now let's look at the formula for $\bar{R}^2$ to see how it makes up for this less useful quality of R^2. Recall

$$\begin{aligned} SST &= \frac{1}{n}\sum_i (y_i - \bar{y})^2 = Var(y) \\ SSR &= \frac{1}{n}\sum_i \hat{u}_i^2 = \frac{1}{n}\sum_i (y_i - \hat{y})^2 = Var(\hat{u}) \end{aligned}$$

This way of writing SST and SSR helps with the intuition, but isn't completely accurate: we don't usually define SST and SSR with the $\frac{1}{n}$ term in there. But when we look at our definition of R^2:

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\frac{1}{n}\sum_i (y_i - \hat{y})^2}{\frac{1}{n}\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2}$$

The $\frac{1}{n}$ cancels out for R^2 anyway, so we essentially ignored this term in the SST and SSR. Taking these terms, it is possible to show that:

1. $\frac{1}{n}\sum_i (y_i - \bar{y})^2$ is a biased estimator of the population variance of y (call it σ_y^2)
2. $\frac{1}{n}\sum_i (y_i - \hat{y})^2$ is a biased estimator of the variance of the true u_i (call it σ_u^2)

Recall that if an estimator is **biased**, that means that it does not equal the true population parameter in expectation.

So what would R^2 look like if we used **unbiased estimators of these variances**?

$$\begin{aligned} \bar{R}^2 &= 1 - \frac{\frac{1}{n-k-1}\sum_i \hat{u}_i^2}{\frac{1}{n-1}\sum_i (y_i - \bar{y})^2} \\ &= 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} \\ &= 1 - \frac{SSR}{SST}\left(\frac{n-1}{n-k-1}\right) \end{aligned}$$

Looking at this formula, you should notice three things:

1. **When we add variables to a regression, SSR cannot increase so $\frac{SSR}{SST}$ will (weakly) decrease.**

2. **However, adding variables to a regression will make k bigger (since k is the number of variables), so $\frac{n-1}{n-k-1}$ will increase.**

3. **As the sample size n increases, $\frac{n-1}{n-k-1}$ gets closer to 1.**

Bottom line: the adjusted R^2, which is sometimes denoted $\bar{R}^2$, includes a "**penalty**" to including variables, so we don't *always* conclude that adding variables improves the fit. However, the R^2 and the adjusted R^2 will be very similar for very large samples.

C. Example

When does adjusted R^2 come in handy? When we want to compare non-nested models. Suppose your friend thinks that older people sleep more at night, but the increase in sleep over time is diminishing, i.e. thinks the relationship between sleep and age is logarithmic and they show you the results from their estimation:

```
Call:
lm(formula = sleep ~ log(age), data = sleep75)

Residuals:
     Min       1Q   Median       3Q      Max
-2452.18  -258.46    11.21   269.80  1387.55

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2821.78     209.42   13.47   <2e-16 ***
log(age)      122.92      57.72    2.13   0.0335 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 443.3 on 704 degrees of freedom
Multiple R-squared:  0.006401,Adjusted R-squared:  0.00499
F-statistic: 4.535 on 1 and 704 DF,  p-value: 0.03355
```

However, you have particularly strong feelings about these functional form assumptions. Having gotten very little sleep these past weeks studying for several midterms, you think that kids sleep a lot, young adults probably sleep less (graduate students sleep even less), and old people sleep a lot. You think the relationship between sleep and age is quadratic, and you show your friend your results:

```
Call:
lm(formula = sleep ~ age + agesq, data = sleep75)

Residuals:
    Min      1Q  Median      3Q     Max
-2518.1  -250.4     2.6   276.8  1390.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3608.0297   230.6457  15.643   <2e-16 ***
age         -21.4904    11.7367  -1.831   0.0675 .
agesq          0.3012     0.1401   2.150   0.0319 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 441.8 on 703 degrees of freedom
```

```
Multiple R-squared:  0.01464,Adjusted R-squared:  0.01184
F-statistic: 5.224 on 2 and 703 DF,  p-value: 0.005598
```

These models are "non-nested" because one cannot be written as a special case of the other. Since there are more variables in your specification, we'd expect R^2 to mechanically increase, so it's not the best way to settle this dispute between you and your friend. However, the adjusted R^2, which will take the different number of variables into account, is still on your side. Note that the minimum of the quadratic function is at 35.6 years (turning point is at $x = -\frac{\beta_1}{2\beta_2}$). I guess we're all going to have to wait a long time for more sleep!

Note on interpreting quadratic variables:

- A one unit increase in x increases y by $\beta_1 + 2\beta_2 x$ units.
- $\beta_1 < 0$ and $\beta_2 > 0$: y decreases with x until $x = -\frac{\beta_1}{2\beta_2}$ and increases thereafter
- $\beta_1 > 0$ and $\beta_2 < 0$: y increases with x until $x = -\frac{\beta_1}{2\beta_2}$ and decreases thereafter

3. Dummy variables

In econometrics, binary variables (or zero-one variables) are referred to as dummy variables. We must decide which event is associated to the value of 1 and which event is associated the value of 0. Ex: we often include a "female" dummy in our regression which takes a value of 1 if the person is a female and 0 if the person is a male. Consider the following model (p.226 Woolridge):

$$wage = \beta_0 + \beta_1 female + \beta_2 educ + u$$

How do we interpret the β_1) parameter? Well, because $female = 1$ corresponds to females and $female = 0$ corresponds to males:

$$\begin{aligned}
E[wage|female = 0, educ] &= \beta_0 + \beta_2 educ \\
E[wage|female = 1, educ] &= \beta_0 + \beta_1 + \beta_2 educ \\
E[wage|female = 1, educ] - E[wage|female = 0, educ] &= \beta_1
\end{aligned}$$

For a given level of education, the difference between females and males wages is captured by β_1.

Recall, why don't we include both female and male in the regression?

- This would be redundant: when we write the regression as we do above, the intercept for males is β_0 and the intercept for female is $\beta_0 + \beta_1$. Because there are only two groups, we only need to have two intercepts.
- Using two dummies introduces perfect collinearity because $female + male = 1$: female is a perfect linear function of female.

We can think of this dummy as introducing an intercept shift between males and females.

4. Models with interactions

A. Interacting two continuous variables

Let's start with the following model:

$$wage = \beta_0 + \beta_1 age + \beta_2 educ + \beta_3 age \times educ + u$$

i. What is the marginal effect of age?

Step 1:

Intuitively, think about regrouping all the terms that have age in them

$$E[wage|educ, age] = \beta_0 + \underbrace{(\beta_1 + \beta_3 educ)}_{regrouped} age + \beta_2 educ$$

So the "age effect" i.e the marginal effect of age is $\beta_1 + \beta_3 educ$.

Step 2:

Rigorously, we can express the marginal effect of age is:

$$\frac{\partial E[wage]}{\partial age} = \beta_1 + \beta_3 educ$$

Then you might be asked to substitute a particular value of educ (usually we select the median). Substituting $educ = 10$ for example gives:

$$\frac{\partial E[wage]}{\partial age} = \beta_1 + \beta_3 * 10$$

Therefore we say that the marginal effect of age on expected wage for people with 10 years of education is $\beta_1 + \beta_3 * 10$.

ii. What is the marginal effect of education?

Step 1:

Intuitively, think about regrouping all the terms that have educ in them

$$E[wage|educ, age] = \beta_0 + \underbrace{(\beta_2 + \beta_3 age)}_{regrouped} educ + \beta_1 age$$

So the "education effect" i.e the marginal effect of education is $\beta_2 + \beta_3 age$.

Step 2:

Rigorously, we can express the marginal effect of education is:

$$\frac{\partial E[wage]}{\partial educ} = \beta_2 + \beta_3 age$$

Then you might be asked to substitute a particular value of educ (usually we select the median). Substituting $age = 20$ for example gives:

$$\frac{\partial E[wage]}{\partial educ} = \beta_2 + \beta_3 * 20$$

Therefore we say that the marginal effect of education on expected wage for people with 20 years of age is $\beta_2 + \beta_3 * 20$.

B. Interacting a continuous variable and a dummy

$$wage = \beta_0 + \beta_1 female + \beta_2 educ + \beta_3 female \times educ + u$$

i. What is the marginal effect of education?

Step 1:

Intuitively, think about regrouping all the terms that have education in them

$$E[wage|educ, female] = \beta_0 + \beta_1 female + \underbrace{(\beta_2 + \beta_3 female)}_{regrouped} educ$$

So the "education effect" i.e the marginal effect of education is $\beta_2 + \beta_3 female$.

Step 2:

Rigorously, we can express the marginal effect of educ is:

$$\frac{\partial E[wage]}{\partial educ} = \beta_2 + \beta_3 female$$

Now just like before, we might ask you to substitute a value for female. This is easy, there are only two possible values for female: 0 , 1.

Substituting $female = 0$ (i.e., you are a male) gives:

$$\frac{\partial E[wage]}{\partial educ} = \beta_2$$

Therefore we say that the marginal effect of education on expected wage for **males** is β_2

Substituting $female = 1$ (i.e., you are a female) gives:

$$\frac{\partial E[wage]}{\partial educ} = \beta_2 + \beta_3$$

Therefore we say that the marginal effect of education on expected wage for **females** is $\beta_2 + \beta_3$

ii. What is the effect of female?

Step 1:

Intuitively, think about regrouping all the terms that have female in them

$$E[wage|educ, age] = \beta_0 + \underbrace{(\beta_1 + \beta_3 educ)}_{regrouped} female + \beta_2 educ$$

So the "female effect" is $\beta_1 + \beta_3 educ$.

Step 2: Semi-Rigorously we can express the additional effect of being female

$$\frac{\Delta E[wage]}{\Delta fem} = \beta_2 + \beta_3 educ$$

where Δfem represents moving from 0 to 1 (i.e moving from male to female i.e. being female instead of male).

Then you might be asked to substitute a particular value of education (usually we select the median). Substituting $educ = 10$ for example gives:

$$\frac{\Delta E[wage]}{\Delta fem} = \beta_1 + \beta_3 * 10$$

Therefore we say that the effect of being female on expected wage for people with 10 years of education is $\beta_1 + \beta_3 * 10$.

iii. Interpreting each coefficient

$$wage = \beta_0 + \beta_1 female + \beta_2 educ + \beta_3 female \times educ + u$$

Interpreting β_0: This parameter reflects the intercept (value of wages) for males with no education.
Interpreting β_1: This parameter reflects the difference in the intercepts between women and men with no education.
Interpreting β_2: This parameter reflects the effect of an additional year of education for males.
Interpreting β_3: This parameter reflects the difference in the returns to education (differential effect of education on wage) for males and females.

1. $\beta_1 < 0, \beta_3 < 0$: the intercept for women is below that of men, and the slope on education is larger for men: women earn less than men at all levels of education, and the gap increases as education gets larger
2. $\beta_1 < 0, \beta_3 > 0$: the intercept for women is below that for men, but the slope on education is larger for women: women earn less than men at low levels of education, but the gap narrows as education increases.
3. $\beta_1 > 0, \beta_3 < 0$: the intercept for women is above that for men, and the slope on education is larger for men: women earn more than men at low levels of education, but the gap narrows as education increases
4. $\beta_1 > 0, \beta_3 > 0$ the intercept for women is above that for men, and the slope on education is larger for women: women earn more than men at all levels of education, and the gap increases as education gets larger

For those who want a bit more math about how we derive the interpretation above more formally. Consider the model:

$$wage = \beta_0 + \beta_1 female + \beta_2 educ + \beta_3 female \times educ + u \tag{1}$$

Then

$$\begin{aligned} E[wage|female = 0, educ] &= \beta_0 + \beta_2 educ \\ E[wage|female = 1, educ] &= \beta_0 + \beta_1 + (\beta_2 + \beta_3) educ \end{aligned}$$

If we let $female = 0$, then the intercept for males is β_0 and the slope on education for males is β_2.

If we let $female = 1$, then the intercept for females is $\beta_0 + \beta_1$ and the slope on education for females is $\beta_2 + \beta_3$.

Taken together :

$$\begin{aligned} E[wage|female = 1, educ] - E[wage|female = 0, educ] &= (\beta_0 + \delta_0 + (\beta_1 + \delta_1)educ) - (\beta_0 + \beta_1 educ) \\ &= \delta_0 + \delta_1 educ \end{aligned}$$

Therefore δ_0 measures the difference in intercepts between women and men, and δ_1 measures the difference in the return to education between men and women.

iv. Some hypotheses we may want to test

1. You hypothesize that the return to education is the same for women and men:

$$H_0 : \beta_3 = 0 \quad vs. \quad H_1 : \beta_3 \neq 0$$

which says that the slope of *wage* with respect to *educ* is the same for men and women. NB: this hypothesis doesn't say anything about the difference in intercepts β_1. So a wage differential between men and women is possible under this null, but the differential must be the same at all levels of education.

2. You hypothesize that average wages are identical for men and women who have the same levels of education

$$H_0 : \beta_1 = 0 \;\;\&\;\; \beta_3 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0 \;\;\&/or\;\; \beta_3 \neq 0$$

And you will use the standard F-test

C. Interacting two dummy variables

$$wage = \beta_0 + \beta_1 married + \beta_2 female + \beta_3 female \times married + u$$

i. What is the effect of married ?

Step 1:

Intuitively, think about regrouping all the terms that have married in them

$$E[wage|educ, age] = \beta_0 + \underbrace{(\beta_1 + \beta_3 female)}_{regrouped} married + \beta_2 female$$

So the "married effect" is $\beta_1 + \beta_3 female$.

Step 2:

Semi-rigorously, we can express the additional effect of being married as:

$$\frac{\Delta E[wage]}{\Delta mar} = \beta_1 + \beta_3 female$$

where Δmar represents moving from 0 to 1 (i.e moving from single to married i.e. being married instead of single).

Now just like before, we might ask you to substitute a value for female. This is easy, there are only two possible values for female: 0 , 1.

Substituting $female = 0$ (i.e you are a male) gives:

$$\frac{\Delta E[wage]}{\Delta mar} = \beta_1$$

Therefore we say that the effect of being married on expected wage for **males** is β_1

Substituting $female = 1$ (i.e you are a female) gives:

$$\frac{\Delta E[wage]}{\Delta mar} = \beta_1 + \beta_3$$

Therefore we say that the effect of being married on expected wage for **females** is $\beta_1 + \beta_3$

ii. What is the effect of female?

Step 1:

Intuitively, think about regrouping all the terms that have female in them

$$E[wage|educ, age] = \beta_0 + \underbrace{(\beta_2 + \beta_3 mar)}_{regrouped} female + \beta_1 mar$$

So the "female effect" is $\beta_2 + \beta_3 age$.

Step 2: Semi-rigorously we can express the additional effect of being female as:

$$\frac{\Delta E[wage]}{\Delta fem} = \beta_2 + \beta_3 mar$$

where Δfem represents moving from 0 to 1 (i.e moving from male to female i.e. being female instead of male).

Now just like before, we might ask you to substitute a value for married. This is easy, there are only two possible values for married: 0 , 1.

Substituting $married = 0$ (i.e you are a male) gives:

$$\frac{\Delta E[wage]}{\Delta fem} = \beta_2$$

Therefore we say that the effect of being female on expected wage for **single** is β_2

Substituting $married = 1$ (i.e you are a female) gives:

$$\frac{\Delta E[wage]}{\Delta fem} = \beta_2 + \beta_3$$

Therefore we say that the effect of being female on expected wage for **married** is $\beta_2 + \beta_3$

iii. Interpreting each coefficient

$$wage = \beta_0 + \beta_1 married + \beta_2 female + \beta_3 female \times married + u$$

Interpreting β_0: This reflects the average wage for single males.

Interpreting β_1: This reflects the average effect of being married for a male.

Interpreting β_2: This reflects the average effect of being female for a single individual.

Interpreting β_3: This reflects the differential effect of being married for a woman relative to what it is for a man.

For those who want a bit more math about how we derive the interpretation above more formally:

$$\begin{aligned}
E[wage|female = 0, married = 0] &= \beta_0 \\
E[wage|female = 0, married = 1] &= \beta_0 + \beta_1 \\
E[wage|female = 1, married = 0] &= \beta_0 + \beta_2 \\
E[wage|female = 1, married = 1] &= \beta_0 + \beta_1 + \beta_2 + \beta_3
\end{aligned}$$

Now let's think about what each coefficient in the regression:

$$\beta_0 : \underbrace{E[wage|female = 0, married = 0]}_{\text{average wage for single males.}} = \beta_0$$

$$\beta_1 : \underbrace{E[wage|female = 0, married = 1] - E[wage|female = 0, married = 0]}_{\text{effect of being married for a male}} = \beta_0 + \beta_1 - \beta_0 = \beta_1$$

$$\beta_2 : \underbrace{E[wage|female = 1, married = 0] - E[wage|female = 0, married = 0]}_{\text{effect of being female for single individuals}} = \beta_0 + \beta_2 - \beta_0 = \beta_2$$

$$\begin{aligned}
\beta_3 : E[wage|female = 1, married = 1] - &\underbrace{E[wage|female = 1, married = 0]}_{\text{subtracting the incremental effect of being married}} \\
\underbrace{- \underbrace{E[wage|female = 0, married = 1]}_{\text{subtracting the incremental effect of being female}} + \underbrace{E[wage|female = 0, married = 0]}_{\text{adding back in the base wage}}}_{\text{differential effect of being married for a woman relative to what it is for a man}} \\
= (\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2) - (\beta_0 + \beta_1) + \beta_0 = \beta_3
\end{aligned}$$

D. Examples

i. Example 1

We have the following model in mind

$$colGPA = \beta_0 + \beta_1 female + \beta_2 athlete + \beta_3 female \times athlete$$

R outputs is as follows

```
Call:
lm(formula = colgpa ˜ female + athlete + female * athlete, data = gpa2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.60856 -0.43341 -0.00341  0.44659  1.69302

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.60856    0.01415 184.380  < 2e-16 ***
female          0.12485    0.02085   5.987 2.31e-09 ***
athlete        -0.30158    0.05531  -5.453 5.25e-08 ***
female:athlete  0.19639    0.11295   1.739   0.0822 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6526 on 4133 degrees of freedom
Multiple R-squared:  0.01884,Adjusted R-squared:  0.01813
F-statistic: 26.46 on 3 and 4133 DF,  p-value: < 2.2e-16
```

- What's the effect of `female` on predicted GPA?

  $0.12485 + 0.19639 athlete$

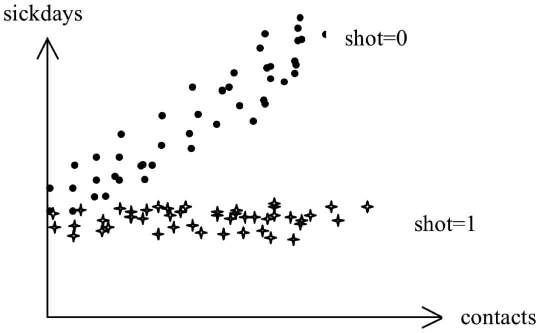
- What's the effect of `athlete` on predicted GPA?

  $-0.30158 + 0.19639 female$

- Interpret each coefficient in your model :
  - β_0: Average GPA for male non-athletes
  - β_1: Average effect of being a female for non-athletes
  - β_2: Average effect of being an athlete for males
  - β_3: Differential effect of being an athlete for a woman relative to what it is for a man

ii. Example 2 (Final 2007)

Suppose you have data for a sample of employees of a firm on the number of sick leave days taken in one year (sickdays), a rough estimate of how many people the employee comes into contact with during his/her workday (contacts), and whether or not the they got a flu shot (shot=1 for those who got a flu shot and 0 for those who didn't). The data give you the following plot:



1. Reading from the graph, explain in words the effect of flu shot on sick leave days.

   The effect of the flu shot reduces the number of sickdays regardless of the number of people the employee comes into contact during his/her workday
2. You estimate the following model:

$$\widehat{sickdays} = \hat{\beta}_0 + \hat{\beta}_1 contacts + \hat{\beta}_2 shot + \hat{\beta}_3(contacts * shot)$$

   What do you expect to find for the sign and significance of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ given the graph of your data above?

To start:

$$E[sickdays|shot = 0, contacts] = \beta_0 + \beta_1 contact$$
$$E[sickdays|shot = 1, contacts] = \beta_0 + \beta_2 + (\beta_1 + \beta_3)contact$$

So if we let $shot = 0$, then the intercept for those without shots is β_0 and the slope on contact for those without shots is β_1. Now if we let $shot = 1$, then the intercept for those with shots is $\beta_0 + \beta_2$ and the slope on contacts for those with shots is $\beta_1 + \beta_3$. Moreover

$$E[sickdays|shot = 1, contacts] - E[sickdays|shot = 0, contacts] = \beta_2 + \beta_3 contact$$

Therefore β_2 measures the difference in intercepts between those with and without flu shots, and β_3 measures the difference in the effect of contacts on sickdays between those with and without shots. Then we might expect:

- $\hat{\beta}_0 > 0$: measures the number of sickdays for an employee with 0 contact and no shots. Might expect those without shots to get sick, even though they don't have any contact with other individuals, which is why we might expect a positive effect.
- $\hat{\beta}_1 > 0$: measures the effect of number of contacts on sickdays for those who don't take shots.
- $\hat{\beta}_2 < 0$: measures the difference in intercepts between those who take a flu shot and those who dont. Would expect those who take shots to have a lower number of sickdays.
- $\hat{\beta}_3 < 0$: measures the difference in the effect of contacts on sickdays between those with and without shots. Would expect the effect of contacts on sickdays to be less for those with a shot.

iii. Example 3

Suppose we have data on the number of extramarital affairs a person has had, in addition to gender, age, the number of years he/she has been married, an indicator for a "happy" marriage, number of children, etc. We suspect that the number of years a person has been married will affect the number of affairs that person has *differently* for very happy marriages compared to very unhappy marriages. Which model below tests this hypothesis?

1. $affairs = \beta_0 + \beta_1 myears + \beta_2 unhappy$
2. $affairs = \beta_0 + \beta_1 myears + \beta_2 unhappy + \beta_3 (myears * unhappy)$

Remember that a dummy variable will identify different subpopulations, or groups, in our data (e.g. women, rural areas, poor households, treatment group, etc.), and we can interpret β_1 above as the slope of affairs with respect to the number of years married. When we interact a continuous variable with a dummy variable, we define different slopes (with respect to the continuous variable) for each group identified by the dummy variable. In other words, we say that the slope with respect to years married is different for unhappily married people and happily married people.

	...in Model (1)	...in Model(2)
What's the effect of an additional year of marriage?	β_1	$\beta_1 + \beta_3 * unhappy$
What's the effect of an unhappy marriage?	β_2	$\beta_3 * myears$

Next we're going to translate our intuition about the relationships between these variables into hypotheses about their signs:

Hypothesis	...expressed in words	...expressed in parameters
1	unhappy *newlyweds* have **more** affairs than happy ones	
2	people who've been happily married for a long time have **fewer** affairs	
3	people who've been *unhappily* married for a long time have **more** affairs	

Hypothesis 1:	unhappy newlyweds have more affairs	$\Rightarrow \beta_2 > 0$
Hypothesis 2:	people who've been happily married for a long time have fewer affairs	$\Rightarrow \beta_1 < 0$
Hypothesis 3:	people who've been unhappily married for a long time have more affairs	$\Rightarrow \beta_3 > 0$