Week 3: Applied Review of Regression

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2020

Outline

- The linear/OLS model assumptions
- Inference: hypothesis tests and confidence intervals
- ANOVA is a just linear regression model
- t-tests and regression
- Linear probability model
- Nonparametric regression
- Heteroskedasticity

Linear/OLS model assumptions

- We saw informally that the linear/OLS model is a **conditional expectation function**, so we can interpreted our estimated linear model as

$$E[y_i|\boldsymbol{x}_i] = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_j x_{ji}$$

- This fact is not that intuitive if you learned to derive the linear model using ordinary least squares (OLS)
- In the class about maximum likelihood estimation, it will become crystal clear that we are modeling the mean of an outcome conditional on covariates assuming that the outcome distributes normal **conditional on covariates** (but the covariates are "fixed," although that can be misleading especially with observational data)
- The *conditional* on covariates part is key, that's why I told you that sometimes it's more helpful to say that the assumption $\epsilon_i \sim N(0, \sigma^2)$ is the same as $y_i \sim (\boldsymbol{X}_i'\boldsymbol{\beta}, \sigma^2)$

Data

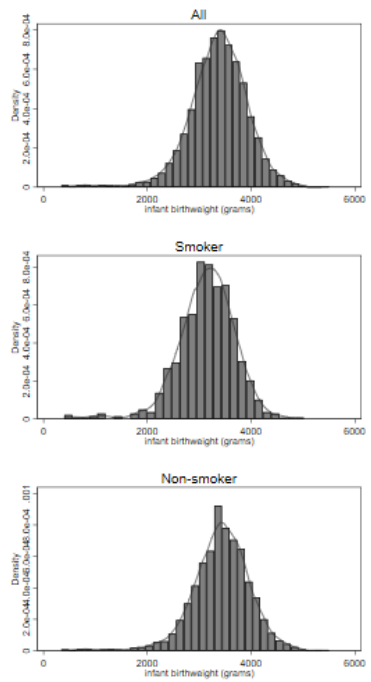
- We will use data from Cattaneo (2010) on birthweight as a function of mother and father covariates (see today's do file)

```
use https://www.stata-press.com/data/r16/cattaneo2, clear
(Excerpt from Cattaneo (2010) Journal of Econometrics 155: 138-154)
desc bweight lbweight mbsmoke msmoke mage medu mrace fbaby
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------------------------
bweight         int     %9.0g                 infant birthweight (grams)
lbweight        byte    %9.0g                 1 if low birthweight baby
mbsmoke         byte    %9.0g      mbsmoke    1 if mother smoked
msmoke          byte    %27.0g     smoke2     cigarettes smoked during pregnancy
mage            byte    %9.0g                 mother's age
medu            byte    %9.0g                 mother's education attainment
mrace           byte    %9.0g                 1 if mother is white
fbaby           byte    %9.0g      YesNo      1 if first baby

tabstat bweight, by(mbsmoke) stats(N mean sd min max)
Summary for variables: bweight
     by categories of: mbsmoke (1 if mother smoked)
  mbsmoke |         N      mean        sd       min       max
----------+--------------------------------------------------
nonsmoker |      3778  3412.912  570.6871       340      5500
   smoker |       864   3137.66  560.8931       397      5018
----------+--------------------------------------------------
    Total |      4642   3361.68  578.8196       340      5500
------------------------------------------------------------
```

Check distribution

- Birthweight distributes normal as many things in **nature** do. It's not random that statistics was developed around normality

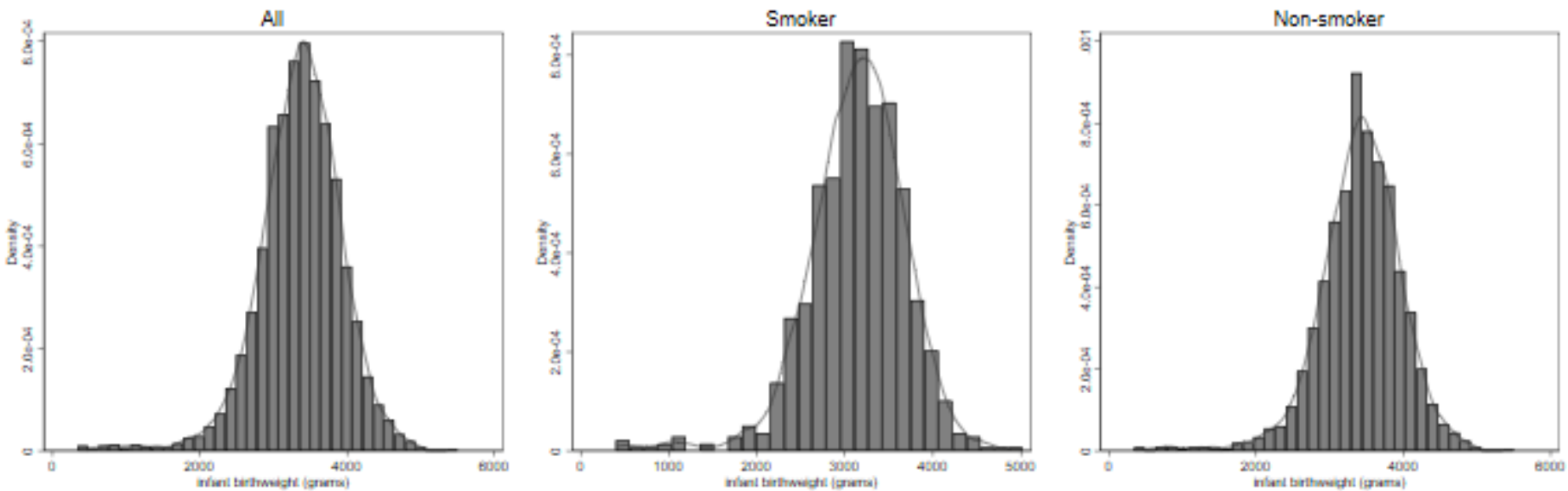
```
set scheme s1mono, permanently
hist bweight, kdensity saving(w_all.gph, replace) title("All")
hist bweight if mbsmoke == 1, kdensity saving(w_smok.gph, replace) title("Smoker")
hist bweight if mbsmoke == 0, kdensity saving(w_nonsmok.gph, replace) title("Non-smoker")
graph combine w_all.gph w_smok.gph w_nonsmok.gph, col(1) xcommon ysize(10)
graph export wg_graph.png, replace
```



Check distribution - better

- This way is better to see the data; note that I had to play with the ysize and xsize options
- Note too how helpful the smoother kdensity is to avoid the issue that the bin size affects the shape of the histogram

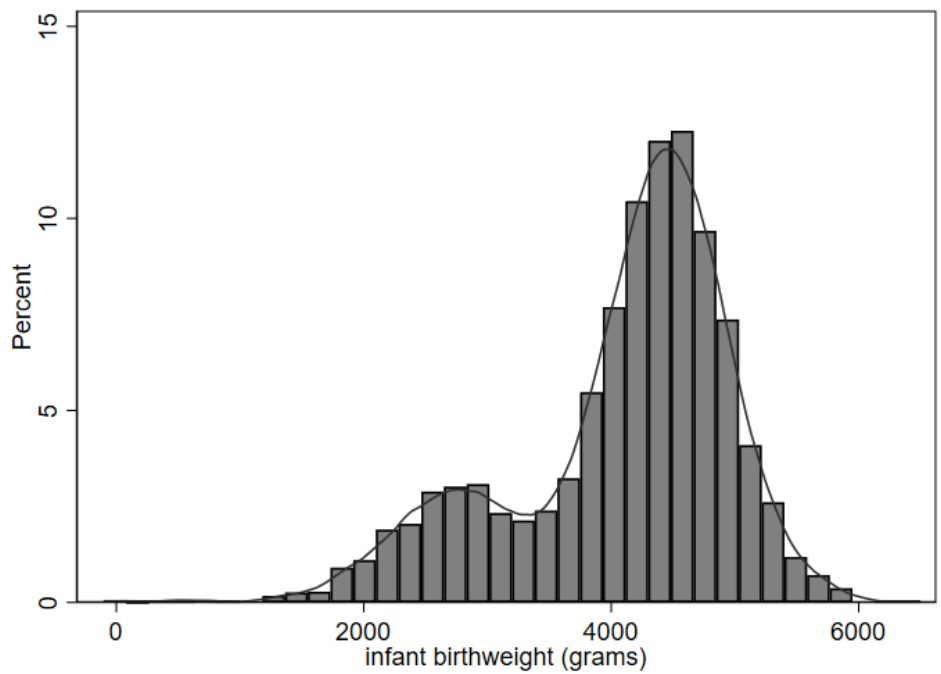
```
graph combine w_all.gph w_smok.gph w_nonsmok.gph, row(1) ysize(4) xsize(12)
graph export wg_graph1.png, replace
```



Let's fake some data

- I want to show you another way of understanding the assumption about the distribution of Y in linear/OLS models
- We will exaggerate (unrealistically) the difference in birthweight between mothers who smoke and mothers who do not

```
* Clonevar is handy
clonevar bweight_fake = bweight
replace bweight_fake = bweight + 1000 if mbsmoke == 0
replace bweight_fake = bweight - 500 if mbsmoke == 1
hist bweight_fake, kdensity percent saving(hist_bwfake, replace)
graph export bw_fake.png, replace
```



Can we use a linear/OLS model with a distribution like that?

- If you showed that distribution to a group of educated people, most likely the majority would agree that a linear/OLS model is not "appropriate" because clearly the outcome is not normally distributed
- But the assumption is not about the distribution of Y. The assumption is about the **conditional distribution** of Y: $Y \sim (\boldsymbol{\beta}'\boldsymbol{X}, \sigma^2)$
- In other words, if some of the covariates "explain" the non-normal distribution of the outcome, we are fine with our linear/OLS model
- That's why you were taught to **check the distribution of the residuals, not the the distribution of the outcome**
- (Digression: If you didn't have data on who smoked, you could estimate a finite mixture model. See here: `https://perraillon.com/zicen` Stata now incorporate Partha Deb's fmm command)

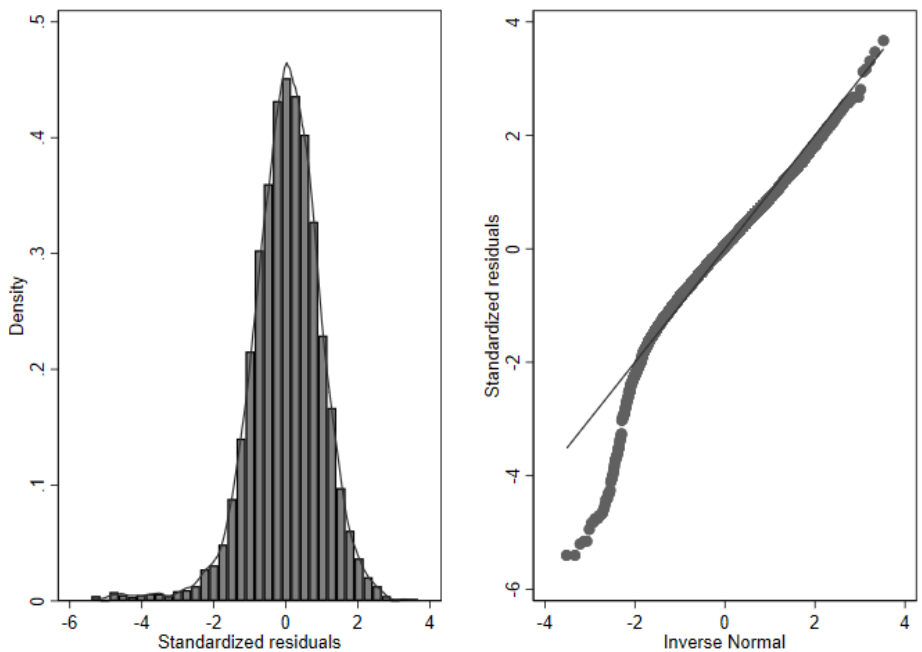
Checking model fit

- Remember from your intro classes that we want to standardize the residuals before checking the distribution (although, in practice, the non-standardized residuals are usually fine)
- qnorm compares the variable with a normal distribution (Q-Q plot)
- More precisely, qnorm compares *observed* quantiles of the variable to a *theoretical* normal with mean and standard deviation like those of the observed variable
- It's a graphical way of exploring deviations from a theoretical distribution

```
qui reg bweight_fake mbsmoke
predict res_std, rstandard

hist res_std, kdensity saving(rno.gph, replace)
qnorm res_std, saving(qno.gph, replace)
graph combine rno.gph qno.gph, row(1)
graph export nor.png, replace
```

Checking model fit

- The graph below shows a fairly decent fitting model, but not great
- R^2 is high: 0.5961. This of course because we faked the data so we made it so: smoking is a very strong predictor of birthweight
- Could we improve the model? Yes, there other variables that could explain the longer tail of the residuals. It could misspecification of functional form or other data problem



Digression: Meaning of R^2

- You learned that R^2 is the percent of the total variance explained by the model
- But it's also a measure of how well a model predicts the observed data: the R^2 is also the square of the correlation between observed Y and predicted $\hat{Y}$

```
reg bweight_fake mbsmoke
      Source |       SS           df       MS      Number of obs   =     4,642
-------------+----------------------------------   F(1, 4640)      =   6847.81
       Model |  2.2161e+09         1  2.2161e+09   Prob > F        =    0.0000
    Residual |  1.5016e+09     4,640  323622.478   R-squared       =    0.5961
-------------+----------------------------------   Adj R-squared   =    0.5960
       Total |  3.7177e+09     4,641  801058.875   Root MSE        =    568.88
------------------------------------------------------------------------------
bweight_fake |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mbsmoke |  -1775.252    21.4528   -82.75   0.000     -1817.31   -1733.194
       _cons |   4412.912   9.255254   476.80   0.000     4394.767    4431.056
------------------------------------------------------------------------------
capture drop yhat
predict yhat
corr yhat bweight_fake
(obs=4,642)
             |     yhat bweigh~e
-------------+------------------
        yhat |   1.0000
bweight_fake |   0.7721   1.0000
. di 0.7721 ^2
.59613841
```

Important: Unbiasness versus inference

- In the class about Stata and today we saw that the linear/OLS model is a conditional expectation function (CEF).
- Please go over Angrist and Pischke Chapter 3. In reading that chapter note that the justifications for CEF do not need the normality assumption
- We don't need the normality assumption to show that the linear/OLS model is a conditional expectation function and is the best one
- We DO NEED the normality of ϵ or the the conditional normality of Y for **inference**
- **In practical terms** this means that even if the normality assumption does not hold (as in the case of cost data), we can still get unbiased estimates the conditional expectation $E[Y_i|\boldsymbol{X}_i]$ using the linear model – of course we still make functional form assumptions and the overlap assumption
- But if we are going to use the model for inference –hypothesis tests, confidence intervals – **we do need to worry about the normality assumption**

Asymptotics

- Asymptotic theory (large sample theory) is a framework for evaluating the properties of estimators and statistical tests
- Usually we assume that sample size grows indefinitely because the population is infinitely large. Some properties of estimators are evaluated as the limit when $n \to \infty$. Properties that work asymptotically may not work in finite samples. We worry too about how fast estimators converge to their true values. Another important idea is that we want to know how the **estimator distributes** if we could repeat the experiment or data generating process many times
- This is where the math in statistics gets difficult. Angrist and Pischke Chapter 3 provide some intuition; Chapter 5 of Wooldridge has more details and needed assumptions
- The main result is this: the estimated coefficients $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_j)$ distribute **multivariate normal** with an estimated variance-covariance matrix. Each $\hat{\beta}_j$ has a marginal normal distribution
- Since theory gives us the distribution of $\hat{\boldsymbol{\beta}}$ and its standard deviation (the **standard error**), $SE(\hat{\beta}_j)$, we can can do hypothesis testing

Stata saves the variance-covariance matrix

- The square root of the variance-covariance matrix diagonal elements are the reported standard errors
- The off-diagonal of the variance-covariance has the covariance of the coefficients (**not the data**!)

```
qui reg bweight mbsmoke mage medu
* ereturn list

matrix list e(b)

e(b)[1,4]
      mbsmoke       mage       medu      _cons
y1  -249.5144  6.8227371  11.249044  3084.5422

matrix list e(V)

symmetric e(V)[4,4]
            mbsmoke        mage        medu       _cons
mbsmoke   475.56678
   mage   1.2833327   2.6301349
   medu   13.472218  -2.3549168   13.434309
  _cons  -293.48648  -40.066446  -110.56724   2588.7052

di sqrt(e(V)[2,2])
1.6217691
```

Big picture

- From now to eternity, think of the estimated coefficients $\hat{\beta}$ in any regression model as random variables that also have an estimated standard deviation (standard error)
- In applied work, we often use the asymptotic properties to understand their distributions
- Once we know the distribution, we can make probabilistic statements about their likely value in the **population**. That's **inference**
- Variance estimation (precision of estimates) is yet another important area of statistics. When theory doesn't help, we have the **bootstrap** (yet another invention that requires computing power) or the **delta method** (that requires normality and Taylor series)

Hypothesis testing reminder

- We start with a **null hypothesis** or a statement about the **population parameter** than you want to test for
- We **assume that the null hypothesis is true** until the data provides evidence that it is not; similar to the way the criminal justice system treats a defendant (innocent until proven guilty by evidence)
- We denote the null hypothesis as H_0. For example, we are interested in testing if the population parameter is equal to zero, so we write: $H_0 : \beta_1 = 0$ (note that we don't write $H_0 : \hat{\beta}_1 = 0$)
- We will **use** $\hat{\beta}_1$ and the *precision* of the estimate of $\hat{\beta}_1$, $SE(\hat{\beta}_1) = \sqrt{var(\hat{\beta}_1)}$, as **evidence**
- We also need an alternative hypothesis, which is usually $H_1 : \beta_1 \neq 0$ (two-tailed test) or $H_1 : \beta_1 > 0$ or $H_1 : \beta_1 < 0$ (one-tailed test)

Hypothesis testing

- We will either accept or reject the null hypothesis given the evidence, which doesn't necessarily make the alternative hypothesis true
- We can make **two mistakes**. We can reject the null when in fact it's true, which is analogous to the mistake of rejecting the innocence of a person on trial when in fact the person is innocent. We call this mistake **Type I error**
- The **significance level** of the test, α, is the **probability of Type I error**: $\alpha = P(rejectH_0|H_0true)$
- We want to **make this error as small as possible** and by convention it's often set at $\alpha = 5\%$, although $\alpha = 10\%$ is acceptable in some disciplines (like economics)

Hypothesis testing

- The other error is failing to reject the null when in fact is false; this is called the **Type II error**
- We also want to minimize this error
- As an alternative, we would like to maximize the **power of the test**, which is $1 - P(TypeII)$
- In other words, **we want to make sure that we have enough evidence to reject the null**
- Sample size plays a key role. The larger the sample size the more "power" we have, so we are confident that we will be able to reject the null
- The 5% has historical roots that go all the way back to Ronald Fischer. 5% is equivalent to 1 in 20

Hypothesis testing

- We also need to choose a **test statistic** and a **critical value**
- A **test statistic**, denoted by T, is a function of the random sample; a particular number for the test statistic is denoted by t
- A more intuitive explanation for the test statistic is that it's a **standardized value** calculated **from the sample data** that is used as **evidence** to accept or reject a null hypothesis
- We are essentially interested in figuring out a probability: What is the probability that we will find a test statistic t **given that the null hypothesis is true**?
- To answer this question given a level of α **we need to know the distribution of t under the null**
- Note the key ingredient of hypothesis testing: we need to understand the **probability distribution of the PARAMETER**, not the data, although that's needed too to understand how the parameter distributes. *Make sure you always think of parameters as random variables*

Hypothesis testing

- That's how you learned to do a test in stats 101: calculate a t statistic and compare it to a critical value from a table at the end of the book. If the t statistic falls outside a rejection region, you reject the null
- With the use of statistical software came the convenience of using **p-values** instead (and you don't need tables at the end of the book either)
- **P-value**: What is the largest significance level at which we could carry out the test and still fail to reject the null hypothesis?
- Now we just look at the p-value for a statistic: if p-value ≥ 0.05, then we do not reject the null at $\alpha = 0.05$
- **So, how do you choose a test statistics? Where do they come from?** From theory

Wald tests

- **Wald tests** come up in parametric models in which parameters are estimated to model a relationship
- Using the estimated parameters from the sample, we can perform statistical tests of the true population value of a parameter θ
- They have the form:

  $t = \frac{\hat{\theta} - \theta_0}{\sqrt{var(\hat{\theta})}}$,

  where θ_0 is the value of θ we want to test for
- Often, we want to test $\theta_0 = 0$ so the Wald test is

  $t = \frac{\hat{\theta}}{\sqrt{var(\hat{\theta})}} = \frac{\hat{\theta}}{se(\hat{\theta})}$

Intuition

- Look at the formula for the Wald test. When is the null more likely to be rejected? $t = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$
- The largest the t the more likely that the test is rejected given a significance level. So:
  1. If $\hat{\theta} - \theta_0$ is large, or alternatively, if the value of the null hypothesis is far from the estimated value from the sample
  2. If $se(\hat{\theta})$ is small; if $se(\hat{\theta})$ is large we will reject because we don't have enough solid evidence (the estimate of the true parameter is uncertain or imprecise)
- $se(\hat{\theta})$ is a **function of the sample size** n. With large datasets, $se(\hat{\theta})$ will be small so we will tend to reject the null even for small $\hat{\theta} - \theta_0$
- This is a current area of research (navel gazing?) in statistics and "big data." Should we use conventional statistical inference? *Most differences are statistically significant* when you have millions of observations

Wald tests

- Note that now we need to figure out how t distributes **under the null** hypothesis so we quantify the probability of observing the null hypothesis
- The test statistics distributes t-student in the linear/OLS model (in the logistic model it will distribute standard normal)
- How we know that? Theory. The numerator of the Wald test distributes normal, the denominator Chi-square. The ratio distributes t-student
- **Technical note**: The Wald test and CIs are based on the coefficients β_j distributing normal. Yet, this is an approximation that depends on sample size and other assumptions
- **Confusion alert**: "t" is for **test statistic** in this context, which in the linear/OLS model happens to distribute t-student

Digression: Basic results form mathematical statistics

- Other probability distributions will show up when we do hypothesistests: t-student, Chi-square, and the F-distribution
- The **Chi-square** (χ^2) is the square of a standard normal so only takes positive values, like a variance
- **F-distribution**: The F distribution originates from a ratio of two Chi-square distributions – like a ratio of variances. Hence it is sometimes "variance ratio distribution" or Fisher-Snedecor distribution (hence the F)
- **t-student**: The t-student looks like a normal distribution with "fatter" tails and originates from a ratio of a normal to a Chi-square distribution - like the Walt test in the linear/OLS model
- The square of the Student's t distribution with d degrees of freedom is equivalent to the F distribution with 1 numerator degree of freedom and d denominator degrees of freedom

Digression II

- The Wald test can also be written as $t = \frac{(\hat{\theta}-\theta_0)^2}{se(\hat{\theta})}$
- The intuition is the same. The difference is that the the test statistics distribution is different (F distribution). Without the square in the numerator, the test statistics distributes t-student
- Not the only type of hypothesis test. We will see another way of performing an asymptotically equivalent test using the **Likelihood Ratio Test** (LRT)

Examples

- Stata displays Wald tests (and CIs) for all estimation commands, but remember the null hypothesis of the reported Wald tests

```
------------------------------------------------------------------------------
     bweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mbsmoke |  -249.5144   21.80749   -11.44   0.000    -292.2675   -206.7613
        mage |   6.822737   1.621769     4.21   0.000     3.643298    10.00218
        medu |   11.24904   3.665284     3.07   0.002     4.063344    18.43474
       _cons |   3084.542   50.87932    60.62   0.000     2984.795     3184.29
------------------------------------------------------------------------------
```

- The $P > |t|$ column is a two-tailed Wald test for $\beta = 0$ item **Never forget that in that test the null is $H_0 : \beta = 0$**

t-student

- The t-student distribution depends on sample size (degrees of freedom), but with larger samples it **converges to a standard normal**
- If you recall stats 101, the two-tailed rejection at $\alpha = 0.05$ is 1.96 with a *standard normal*
- With the t-student and 120 df it is 1.98. With 10 it's 2.23
- Of course, we do not need tables now. We use p-values but we can also use Stata (remember the normal and the t-student are symmetric)

```
 di invttail(10,0.025)
2.2281389
di invttail(120,0.025)
1.9799304
di invttail(300,0.025)
1.967903
```

Tests in Stata

- Stata has a post-estimation command called "test" that is extremely flexible; you can do all sort of tests with the coefficients of a model (type "help test"). We will use the smoking categories in the dataset

```
. desc msmoke

              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------------------------
msmoke          byte    %27.0g     smoke2     cigarettes smoked during pregnancy

. label list smoke2
smoke2:
           0 0 daily
           1 1-5 daily
           2 6-10 daily
           3 11+ daily

. tab msmoke

 cigarettes |
     smoked |
     during |
  pregnancy |      Freq.     Percent        Cum.
------------+-----------------------------------
    0 daily |      3,778       81.39       81.39
  1-5 daily |        200        4.31       85.70
 6-10 daily |        337        7.26       92.96
  11+ daily |        327        7.04      100.00
------------+-----------------------------------
      Total |      4,642      100.00
```

Tests in Stata

```
reg bweight i.msmoke mage medu
<output omitted>
------------------------------------------------------------------------------
     bweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      msmoke |
  1-5 daily  |  -191.7307   41.28743    -4.64   0.000    -272.6737   -110.7877
 6-10 daily  |  -250.9248   32.50921    -7.72   0.000    -314.6583   -187.1913
  11+ daily  |  -283.7121   32.97667    -8.60   0.000    -348.3621   -219.0622
             |
        mage |   6.939716   1.622869     4.28   0.000      3.75812    10.12131
        medu |   11.00934   3.667179     3.00   0.003     3.819922    18.19875
       _cons |   3084.505   50.87333    60.63   0.000     2984.769    3184.241
------------------------------------------------------------------------------
test medu
 ( 1)  medu = 0
       F(  1,  4636) =    9.01
            Prob > F =    0.0027
*test medu =0
test medu = 1
 ( 1)  medu = 1
       F(  1,  4636) =    7.45
            Prob > F =    0.0064

* Joint: test if effect of 6-10 cigs is same as 11+ daily
test 2.msmoke= 3.msmoke
 ( 1)  2.msmoke - 3.msmoke = 0
       F(  1,  4636) =    0.56
            Prob > F =    0.4558
* Can't do: test medu > 12
```

Tests in Stata

- As you can tell, Stata is not showing a t-student tests but rather F tests. That's because Stata is performing a Wald test squaring the numerator
- The Wald test in outpout is an F test with 1 numerator degree of freedom and 4636 denominator degrees of freedom $(n - j)$, where j is the number of parameters
- The test command is more flexible. It can be used to test **linear combination** of parameters
- In the previous slide we used it to test if the effect of smoking 6-10 and 11+ is the same. Said another way, we are testing if expected value of birthweight is the same for mothers who smoke 6-10 and 11+
- What is the Wald test for the coefficient of 6-10 daily testing in the regression ouput?

Please review your stats notes

- Suppose we estimate the following model:

$$bweight = \beta_0 + \beta_1 smoked + \beta_3 mage + \beta_4 smoked * mage + \beta_5 medu + \epsilon$$

- How do you interpret β_0? β_1? β_4?
- How do you test that the affect of age on birthweight is significant? What is the null hypothesis?
- How do you test that the effect of smoking is significant? What is the null hypothesis?

Centering

- We will use centering with parametric regression discontinuity models, but it's very handy to use center with interactions when one or more variable is continuous
- Instead of using mother's age in the raw scale, we can center mother age at some value. For example, we can center mother's age at the mean, which in this dataset is 25.6
- But we may want to choose another number. Pregnancies for mothers age 35 or older are considered high risk, so let's use 35
- We define a new variable: $mage_c = mage - 35$. The model is now:

$$bweight = \gamma_0 + \gamma_1 smoked + \gamma_3 mage_c + \gamma_4 smoked * mage_c + \eta$$

Centering

- Same as

$$bweight = \gamma_0 + \gamma_1 smoked + \gamma_3(mage - 35) + \gamma_4 smoked * (mage - 35) + \eta$$

- Here is the best part. γ_1 is now the effect of smoking on birthweight for a 35 y/0 mother
- Notice the key part: we have the effect of smoking at a specific age, rather than at $mage = 0$, which is of no interest
- γ_3 is still the effect of mother's age on birthweight for a mother who did not smoke. Same interpretation centering or not (showing why is one of your homework questions)
- γ_0 is the average birthweigth for a 35 y/o mother who did not smoke

Check

- Note which coefficients change and which do not

```
gen mage_c = mage - 35

qui reg bweight i.msmoke##c.mage
est sto m1
qui reg bweight i.msmoke##c.mage_c
est sto m2
est table m1 m2

----------------------------------------
    Variable |     m1           m2
-------------+--------------------------
      msmoke |
  1-5 daily  |  195.35849   -357.50063
 6-10 daily  |  128.41648   -410.86158
  11+ daily  |  57.909727   -422.91716
             |
       mage  |  11.362577
             |
      msmoke#|
      c.mage |
  1-5 daily  | -15.795975
 6-10 daily  | -15.407944
  11+ daily  | -13.737911
             |
      mage_c |               11.362577
             |
      msmoke#|
    c.mage_c |
  1-5 daily  |              -15.795975
 6-10 daily  |              -15.407944
  11+ daily  |              -13.737911
             |
       _cons |  3108.2754    3505.9656
----------------------------------------
```

Tests: Linear combinations

```
qui reg bweight i.mbsmoke##c.mage medu
* to remember how to name the coefficient use the code below
matrix list e(b)
e(b)[1,7]
             0b.            1.                0b.mbsmoke#   1.mbsmoke#
        mbsmoke       mbsmoke          mage       co.mage       c.mage          medu
y1            0    119.49673    9.4224137             0   -14.540117    10.328387
          _cons
y1    3026.7476

test mage 1.mbsmoke#c.mage
 ( 1)  mage = 0
 ( 2)  1.mbsmoke#c.mage = 0

       F(  2,  4637) =   15.51
            Prob > F =    0.0000

* syntax can be confusing
 test mage = 1.mbsmoke#c.mage =0
 ( 1)  mage - 1.mbsmoke#c.mage = 0
 ( 2)  mage = 0

       F(  2,  4637) =   15.51
            Prob > F =    0.0000
```

Tests: Linear combinations - comparing two models

- The previous test is equivalent to comparing nested models using an F test. The intuition for comparing nested models is fairly simple: we will compare their SSEs. SSE is the sum of squares of the **residuals**, which gives a measure of the variance **not explained** by our model
- The full model (FM) has age interacted. The reduced model (RM) does not have age
- Define SSE(RM) as the sum of square of the residuals of the reduced model and SSE(FM) as the sum of square of the residuals of the full model
- We will use the ratio $F = \frac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}$
- The null is that that the reduced model is adequate. We reject the null if $F \geq F(p+1-k, n-p-1; \alpha)$
- $F(p+1-k, n-p-1; \alpha)$ is the critical value
- Note that p+1-k is just the number of additional parameters in the full model

F test for nested models

```
qui reg bweight i.mbsmoke##c.mage medu
scalar sse_f = e(rss)

qui reg bweight i.mbsmoke medu
scalar sse_r = e(rss)

di ((sse_r - sse_f)/2)/(sse_f/(4642-4-1))
15.512601
```

- More details on these lecture notes: `https://perraillon.com/s/week-5-MLR-II.pdf`
- Review Chapter 12 of Cameron and Trivedi

Likelihood ratio test (LRT)

- When we cover maximum likelihood estimation we will see yet another test to compare nested models: the likelihood ratio test
- There is a direct connection between SSE and the (log) likelihood function:

$$ll = -0.5N(ln(2\pi) + ln(\frac{SSE}{N}) + 1)$$

- The F test above and the LRT are **asymptotically equivalent**. But one drawback of the LRT for linear models is that it doesn't correct for heteroskedasticity. So if you must use robust SEs, then better to stick to F tests (test command). More on heteroskedasticity below

```
qui reg bweight i.mbsmoke##c.mage medu
est sto mf

qui reg bweight i.mbsmoke medu
est sto mr

lrtest mf mr

Likelihood-ratio test                                 LR chi2(2)  =     30.96
(Assumption: mr nested in mf)                         Prob > chi2 =    0.0000
```

Partitioning variance

- A key result of the linear/OLS model is that we can partition the variance
- The **observed outcome** Y has a variance. We often use the numerator of the variance, the total sum of squares: $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- The **total observed variance** can be divided into two components: the sum squared due to the **R**egression model (SSR) and the sum squared of the residual (**E**rror): $SST = SSR + SSE$
- In more intuitive words, the observed total variance is made of two parts. One part could be explained by a regression model (that is, by covariates) while the other part is unexplained by the model (the error)
- When we estimate a linear model we are trying to explain as much as possible the observed variability of Y. See details here: `https://perraillon.com/s/week-4-SLR-III.pdf`
- An example: traffic in Chicago

ANOVA

- You probably spent a lot of time in your stats classes doing ANOVA, ANCOVA tests
- If you come from an econ background, you may not even know what is ANOVA. ANOVA, Analysis of Variance, is used to compare **means** (yep, very intuitive)
- Let's say we estimate this model:
  $Y_i = \beta_0 + \beta_1 cigs1_i + \beta_2 cigs2_i + \beta_3 cigs3_i + \epsilon_i$, where cigs1 is the dummy variable for smoking 1-5 daily, cigs2 is 6-10, etc
- Remember, the linear/OLS model is a conditional expectation function, so $E[Y|cigs = 1] = \hat{\beta}_0 + \hat{\beta}_1$ is the mean birthweight if a mother smoked 1-5 cigs daily, while $\hat{\beta}_1$ is the **difference** in birthweight means between mothers who smoked 1-5 cigs and those who did not smoke ($\hat{\beta}_0$)
- If we test the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3$, we are essentially testing if the mean is the same for all smoking categories
- So **ANOVA is just a linear regression model** (see more nuance here: `https://perraillon.com/s/week-8-qualitative.pdf`)

ANOVA as a linear model

```
reg bweight i.msmoke
      Source |       SS           df       MS      Number of obs   =     4,642
-------------+----------------------------------   F(3, 4638)      =     55.90
       Model |  54262201.1         3  18087400.4   Prob > F        =    0.0000
    Residual |  1.5006e+09     4,638  323549.382   R-squared       =    0.0349
-------------+----------------------------------   Adj R-squared   =    0.0343
       Total |  1.5549e+09     4,641  335032.156   Root MSE        =    568.81

< output omitted >

. test (1.msmoke=0) (2.msmoke=0) (3.msmoke=0)

 ( 1)  1.msmoke = 0
 ( 2)  2.msmoke = 0
 ( 3)  3.msmoke = 0

       F(  3,  4638) =   55.90
            Prob > F =    0.0000

*anova bweight i.msmoke
*contrast a.msmoke
```

- Note the same test on top of the regression output. That's the "omnibus" test of significant
- Making connections: Analysis of Variance compare ratio of variances, so no surprise it's an F-test

ANOVA

- You probably spent a lot of time in your stats classes doing ANOVA, ANCOVA tests
- If you come from an econ background, you may not even know what is ANOVA. ANOVA, Analysis of Variance, is used to compare means
- Let's say we estimate this model: $Y_i = \beta_0 + \beta_1 cigs1_i + \beta_2 cigs2_i + \beta_3 cigs3_i + \epsilon_i$, where cigs1 is the dummy variable for smoking 1-5 daily, cigs2 is 6-10, etc
- Remember, the linear/OLS model is a conditional expectation function, so $E[Y|cigs = 1] = \hat{\beta}_0 + \hat{\beta}_1$ is the mean birthweight if a mother smoked 1-5 cigs daily, while $\hat{\beta}_1$ is the **difference** in birthweight means between mothers who smoked 1-5 cigs and those who did not smoke ($\hat{\beta}_0$)
- If we test the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3$, were are essentially testing that the mean birthweight is the same regardless of the category of smoking during pregnancy
- So **ANOVA is just a linear regression model** (see more nuance here: `https://perraillon.com/s/week-8-qualitative.pdf`)

t-tests

- In stats 101 you also learned to do t-test for independent samples to compare means. That, of course, has to be equivalent to a regression model

```
reg bweight i.mbsmoke
<... output omitted ... >
     smoker  |  -275.2519    21.4528   -12.83   0.000    -317.3096   -233.1942

ttest bweight, by(mbsmoke)
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
nonsmoke |   3,778    3412.912    9.284683    570.6871    3394.708    3431.115
  smoker |     864     3137.66    19.08197    560.8931    3100.207    3175.112
---------+--------------------------------------------------------------------
combined |   4,642     3361.68    8.495534    578.8196    3345.025    3378.335
---------+--------------------------------------------------------------------
    diff |            275.2519     21.4528                233.1942    317.3096
------------------------------------------------------------------------------
    diff = mean(nonsmoke) - mean(smoker)                          t =  12.8306
Ho: diff = 0                                     degrees of freedom =     4640

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Confidence Intervals

- The proper interpretation of a confidence interval in the frequentist approach is that if we repeated the experiment many times, about $x\%$ percent of the time the value of β would be within the confidence interval
- By convention, we build 95% confidence intervals, which implies $\alpha = 0.05$
- Intuitively, we need to know the distribution of $\hat{\beta}$ and its precision, the standard error, same as with hypothesis testing
- From theory, the confidence interval for $\hat{\beta}_j$ is: $\hat{\beta}_j \pm t_{(n-2,\alpha/2)} se(\hat{\beta}_j)$. $t_{(n-2,\alpha/2)}$ is about 1.98
- If the Wald test for $H_0 : \beta_j = 0$ is a probabilistic statement about how likely it is the the standardized beta is zero, then it's obvious that the confidence interval is in essence a range of values β_j could take in the population given the sample

Hypothesis test and confidence intervals

```
------------------------------------------------------------------------------
     bweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mbsmoke |
     smoker  |  -249.5144   21.80749   -11.44   0.000    -292.2675   -206.7613
        mage |   6.822737   1.621769     4.21   0.000     3.643298    10.00218
        medu |   11.24904   3.665284     3.07   0.002     4.063344    18.43474
       _cons |   3084.542   50.87932    60.62   0.000     2984.795     3184.29
------------------------------------------------------------------------------
test 1.mbsmoke = -292.2675
 ( 1)  1.mbsmoke = -292.2675
       F(  1,  4638) =    3.84
            Prob > F =    0.0500
test 1.mbsmoke = -206.7613
 ( 1)  1.mbsmoke = -206.7613
       F(  1,  4638) =    3.84
            Prob > F =    0.0500
test 1.mbsmoke = -250
 ( 1)  1.mbsmoke = -250
       F(  1,  4638) =    0.00
            Prob > F =    0.9822
test 1.mbsmoke = -300
 ( 1)  1.mbsmoke = -300
       F(  1,  4638) =    5.36
            Prob > F =    0.0207
```

- **Remember this**: If the number θ_0 in null $H_0 : \beta_j = \theta_0$ is within 95%CI, we won't reject a null for that value; if the number is outside CI, we will reject

Cool things you can do with simulations

- If we know from theory the distribution of the estimated coefficients and their standard errors, we could also use simulations to make probabilistic statements
- **Digression**: This approach could be viewed are reverse Bayesian statistics. In Bayesian analysis, estimation is done with simulations. You don't get just one $\hat{\beta}$, you get thousands. The mean (or mode) of the thousands betas is the point estimate. You use the thousands to do confidence intervals and other statistics juts by counting. If we assume uninformative ("flat" priors, we arrive to similar point estimates as OLS or MLE). See Bayes.do file if curious
- The code in the next slide uses this idea. We just simulate the possible values of $\hat{\beta}_{smoked}$ to recreate confidence intervals
- But we can do more. Now we could estimate the probability that, say, $\hat{\beta}_{smoked} \geq -300$

Simulation

```
preserve
   clear
   set seed 1234567066
   set obs 10000
   * The rt function generates t-student distribution with mean zero and sd 1
   gen zt = rt(4642-4)
   * Make it have a mean of bsmoke and sd of bsmoke_se
   gen beta_sm_sd = bsmoke_se*zt + bsmoke
   * Get the 2,5 and 97.5 percentile
   _pctile beta_sm_sd, p(2.5)
   local ci_lb=r(r1)
   _pctile beta_sm_sd, p(97.5)
   local ci_ub=r(r1)
   di "["  `ci_lb' ", " `ci_ub' "]"
   * Another way to get the percentile
   *centile beta_sm_sd, centile(2.5(5)97.5)
   * How many times are the simulated coefficients within the confidence interval?
   qui count if beta_sm_sd >= -292.2675 & beta_sm_sd <= -206.7613
   di r(N)/10000
   * More nifty. What is the probability that the difference in birthweight between
   * smokers and non-smokers is greater than -300 grams?
   qui count if beta_sm_sd >= -310
   di r(N)/10000
restore
```

- It may look complicated but the code is easier to follow. You need to remember the definition of centile

Linear Probability Models

- We know that the linear/OLS is an unbiased conditional expectation function regardless of the (conditional) distribution of the outcome
- So why not just use linear/OLS models for every type of outcome? Well, because we know that the distribution of ϵ or the (conditional) normality of Y are need for **inference**
- Nonetheless, using the linear/OLS model as a first pass is incredibly helpful regardless of how the outcome distributes
- In other words: the linear/OLS is the wrong but helpful model. You just need to be careful about the inference part (and other nuances)
- One main advantage of using the linaer/OLS model is that the **coefficient can be interpret in the scale of of interest**

Example: What is the probability of a low birthweight baby?

- The outcome is a dummy variable 1/0 indicating low birthweight. The only predictor for now will be mother's age. Remember that the **mean of a 1/0 variable is just the proportion of 1s**

```
sum lbweight mage
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    lbweight |      4,642    .0603188    .2381022          0          1
        mage |      4,642    26.50452    5.619026         13         45

reg lbweight mage, robust
...
------------------------------------------------------------------------------
             |               Robust
    lbweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mage |  -.0018922   .0006662    -2.84   0.005    -.0031982   -.0005862
       _cons |   .1104703   .0185064     5.97   0.000      .074189    .1467517
------------------------------------------------------------------------------
```

- The probability of low birthweight decreases as mothers get older: a 0.18 percent points decrease by each year or about 2 percent points in 10 years (note: this dataset is in part simulated; usually a U-shaped relationship)

Linear Probability Models

- Several considerations to keep in mind
  1. LPMs assume additive, linear effects (we could have added a quadratic term for mother's age)
  2. Predictions can be negative
  3. If the mean of the outcome Y is close to 0 or close to 1, the LPM could be misleading because in those areas effects are non-linear
  4. The standard errors are wrong – in my experience, though, they are often close enough
  5. If interactions between variables are considerably, effects could be different than when running the correct model: logit or probit
- Despite this drawbacks, LPMs are very handy. But eventually you need to estimate the correct model

The correct model

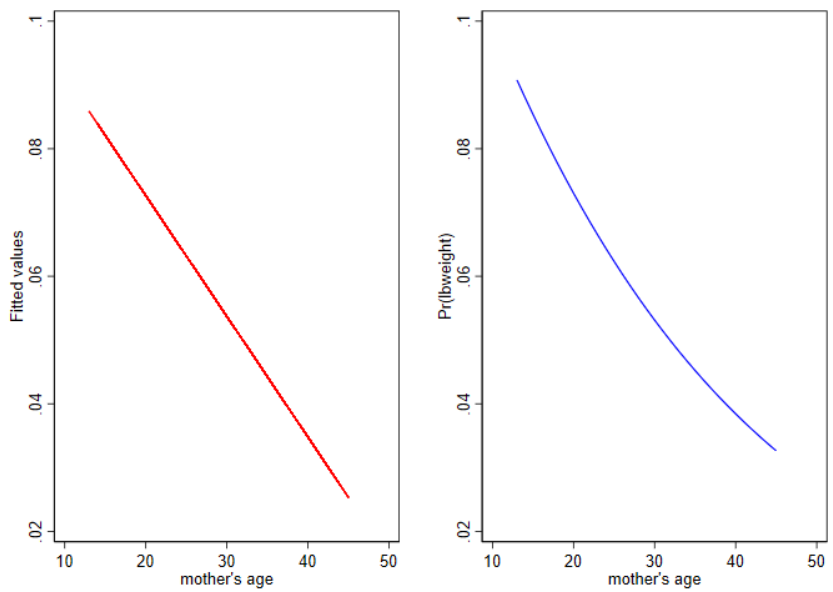
- Fairly similar results with logistic models

```
logit lbweight mage, nolog

Logistic regression                             Number of obs     =      4,642
                                                LR chi2(1)        =       9.35
                                                Prob > chi2       =     0.0022
Log likelihood = -1052.9756                     Pseudo R2         =     0.0044

------------------------------------------------------------------------------
    lbweight |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mage |   -.033934   .0111749    -3.04   0.002    -.0558363   -.0120317
       _cons |  -1.862322   .2924069    -6.37   0.000    -2.435429   -1.289214
------------------------------------------------------------------------------

margins, dydx(mage)
Average marginal effects                        Number of obs     =      4,642
Model VCE    : OIM
Expression   : Pr(lbweight), predict()
dy/dx w.r.t. : mage

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mage |  -.0019195   .0006381    -3.01   0.003    -.0031701   -.0006689
------------------------------------------------------------------------------
```

Intuition

- Compare predictions to understand why in this case effects are so similar. Note that we could have added a quadratic term for mage in the linear model to incorporate nonlinearity but fit could be worse, actually (check)

```
qui logit lbweight mage, nolog
predict yhat_ols
line yhat_ols mage, color(red) saving(pred_ols.gph, replace)
logit lbweight mage, nolog
predict yhat_logit
line yhat_logit mage, sort color(blue) saving(pred_logit.gph, replace)
graph combine pred_ols.gph pred_logit.gph, row(1)
```

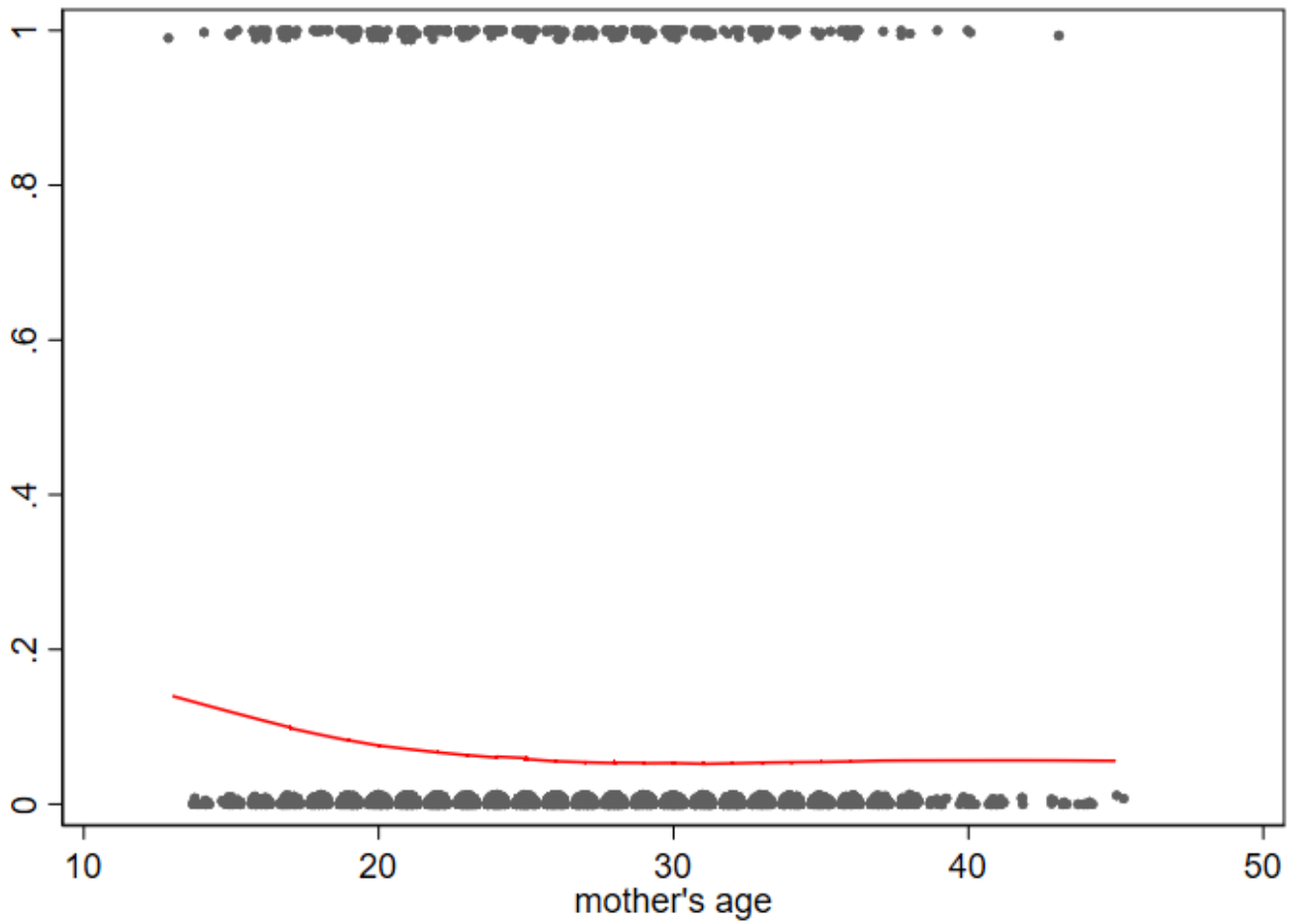


Cool way of graphing a dummy outcome

- We can use lowess or other nonparametric methods to see the shape of a 1/0 variable with respect to another variable

```
lowess lbweight mage, gen(y_smooth)

* make it pretty
scatter lbweight mage, jitter(2) msize(vsmall) legend(off) ///
        || line y_smooth mage, sort color(red)
graph export low.png, replace
```

Cool way of graphing a dummy outcome

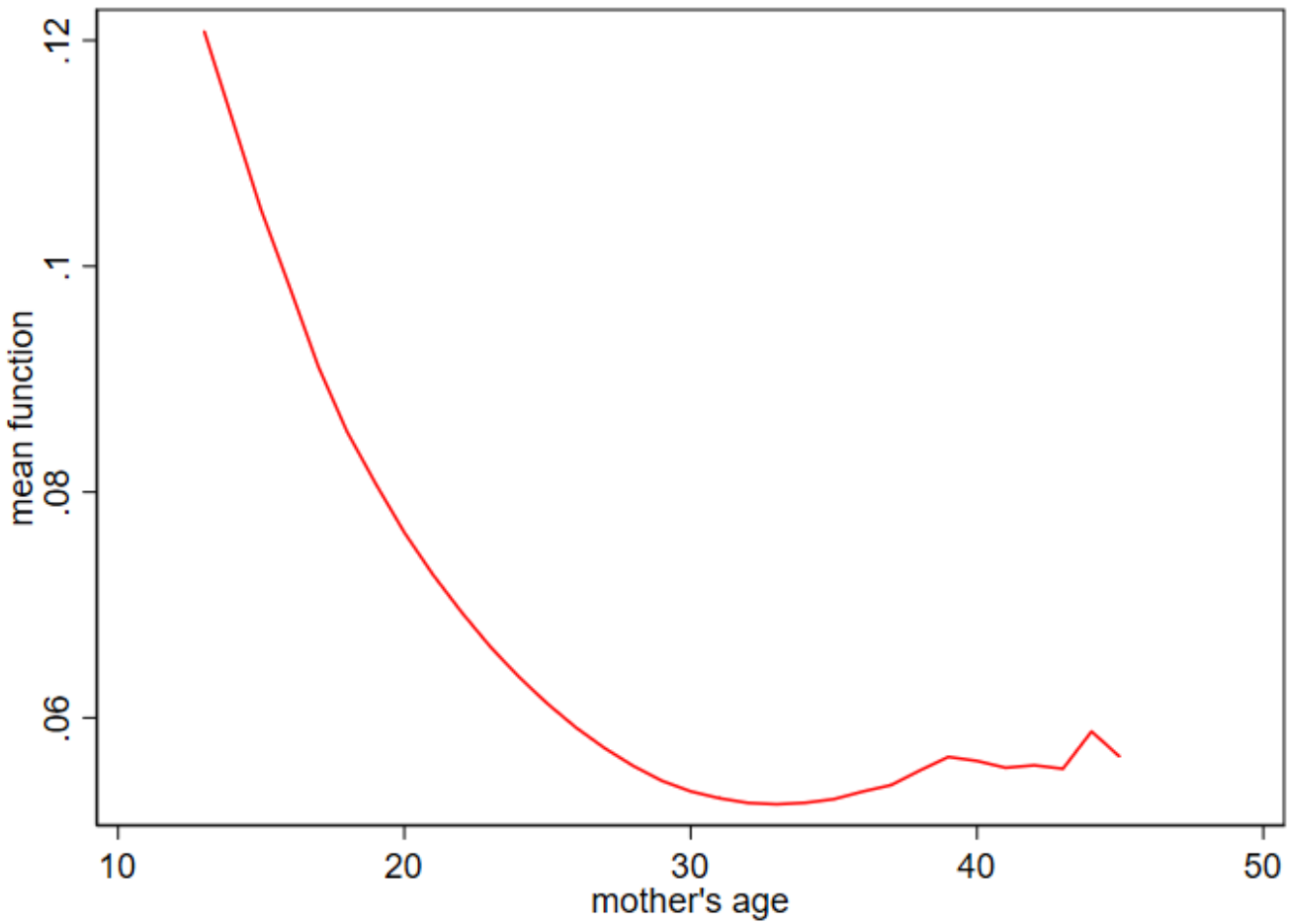


Another way: nonparametric regression

- Stata has new(ish) and more powerful command to perform nonparametric (kernel or series) regression: npregress. I'll focus on kernel regression
- Same idea as lowess, details are more complicated. You still need a bandwidth, although it's optimally determined, and a weighting formula that is called the kernel
- The npregress commands can have more than one covariate and you can do inference as well (SEs calculated using bootstrap). See do file

```
* One way with the npgrah option
npregress kernel lbweight mage
npgraph

* Or make your own graphs
npregress kernel lbweight mage, predict(y_kernel deriv)
sum y_kernel deriv
* note this matches the output
scatter lbweight mage, jitter(3) || line y_kernel mage, sort color(red)

* Or just by itself to see trend better
line y_kernel mage, sort color(red)
graph export ykernel.png, replace
```

Nonparametric regression graph



Hetoroskedasticity

- We wrote the assumptions as $\epsilon_i \sim N(0, \sigma^2)$ or $y_i \sim (\beta' \boldsymbol{X}_i, \sigma^2)$
- Note that we don't write σ_i. We assume that that all observations i have the same variance. This is a fairly unrealistic assumption. More likely that not, it doesn't hold in any dataset you will use
- For example, there is likely more variability on $E[birthweight|mage]$ for older mothers
- What is the problem of hetoroskedasticity? Nothing to do with the bias of the linear/OLS model. The problem is that the standard errors (and p-values, etc) are not correct (to high or too low)
- Thankfully, there is an easy solution: we have the sandwich estimator for the variance, which solves the problem (that's the **robust** option in regress)
- Remember that I told you to always ask yourself "robust to what?" when you the work robust: it's robust to hetoroskedasticity in this context

The good, great, fantastic, and (slightly) bad of the sandwich

- See this lecture for more: `https://perraillon.com/s/week-10-heteroskedasticity.pdf`
- **Good**: We do **not need to know the source** of unequal variance
- **Great**: The sandwhich estimator is **asymptotically unbiased**
- **Fantastic**: The sandwhich estimator is asymptotically unbiased **even in the presence of homoskedasticity**
- If we often suspect heteroskedasticity and the sandwich estimator is asymptotically valid *even in the presence of homoskedasticity*, **why not always use the robust SEs?**
- Well... many researchers add the option robust to every single model for "insurance"
- **The bad**: The only drawback is that if the homoskedasticity assumption is valid, in **smaller samples the robust SEs may be biased**. But we seldom work with "small" samples anymore

Interactions and stratification

- We are going to practice interpreting interactions because interactions can be confusing
- With interactions, we are essentially saying that the effect one variable on the outcome depends on the value of another variable. So **not a constant** effect or linear effect
- In the model $lbweight_i = \beta_0 + \beta_1 mage_i + \beta_2 smoked_i + \beta_3 mage * smoked_i$, we are assuming that the effect of mother's age on low birth weight depends on whether the mother smoked or not. But we can say it the other way around: the effect of smoking depends on mother's age
- Easier with derivatives: $\frac{\partial E[lbweight_i | mage_i, smoked_i]}{\partial mage_i} = \beta_1 + \beta_3 smoked_i$
- If a mother smoked, the effect if age is $\beta_1 + \beta_3$. If not, then it's just β_1
- We can't take derivatives of indicator variables because derivatives are limits and require small changes. But the formula works too in simple models:
- $\frac{\Delta E[lbweight_i | mage_i, smoked_i]}{\Delta smoked_i} = \beta_2 + \beta_3 mage_i$

The interaction term is a difference-in-difference

- Using the equations above, what is the effect of age for mothers who smoke? It's $\beta_1 + \beta_3$. So the difference in birthweight for a small change (or difference) in age is given by that expression
- What is the change or small difference in low birthweight for a change in age for those who do not smoke? It's just β_1
- What is the difference between those two differences? It's $\beta_1 + \beta_3 - \beta_1 = \beta_3$
- When we have interactions between two dummy variables, it's easier to see that the interaction is a difference-in-difference
- Obviously, there is a connection with difference-in-difference research designs. The effect of interest is an interaction or a difference-in-difference
- **Caution**: A model with an interaction is not a difference-in-difference research design

Not the only way to get there

- Another way. What is the predicted weight for a mother who smoked?:
  $E[lbweight_i|mage_i, smoked_i = 1] = \beta_0 + \beta_1 mage_i + \beta_2 + \beta_3 mage_i$
- What is the predicted value for a mother who did not smoke?
  $E[lbweight_i|mage_i, smoked_i = 0] = \beta_0 + \beta_1 mage$
- What is the difference in expected birth weight for those who smoked vs those who did not:
  $E[lbweight_i|mage_i, smoked_i = 1] - E[lbweight_i|mage_i, smoked_i = 0]$?:
  $\beta_0 + \beta_1 mage + \beta_3 mage_i + \beta_2 - (\beta_0 + \beta_1 mage) = \beta_2 + \beta_3 mage$
- One way to see the previous steps is that we held mage constant while changing smoking status

Stratification

- Note that with interactions we said that the effect of age is different depending on whether a mother smoked or not

  Model with interaction:
  $lbweight = \beta_0 + \beta_1 mage_i + \beta_2 smoked + \beta_3 mage * smoked$
- We could run **two models** instead:

  Model 1 for $smoked = 1$: $lbweight = \alpha_0 + \alpha_1 mage$

  Model 2 for $smoked = 0$: $lbweight = \gamma_0 + \gamma_1 mage$
- I changed the notation to emphasize that the coefficients in the three models are likely to be different
- As we may suspect, $\beta_3 = \alpha_1 - \gamma_1$
- With an interacted model, we get an statistical test with a null *equivalent* to $H_0 : \alpha_1 = \gamma_1$ or $H_0 : \alpha_1 - \gamma_1 = 0$
- **This is important**. We will practice many times

Verify

- Check that coefficients match

```
/// --- Interactions and stratification
* Run stratified models and save coefficients
qui reg lbweight mage if mbsmoke == 1, robust
local m1_alpha1 = _b[mage]
qui reg lbweight mage if mbsmoke == 0, robust
local m1_gamma1 = _b[mage]

* Difference
di `m1_alpha1' - `m1_gamma1'
.006432

* Should match interaction term
qui reg lbweight c.mage##i.mbsmoke, robust
di _b[1.mbsmoke#c.mage]
.006432
```