Regression Methods

ECON 5753 — University of Arkansas

Prof. Kyle Butts

March 2025

Theoretical Results from Last Time

1. Derive the OLS estimator $\hat{\beta}_{\text{OLS}} = (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'y$

2. Derive the sample distribution of $\hat{\beta}_{\text{OLS}}$

$$\hat{\beta}_{\text{OLS}} \sim \mathcal{N}(\beta_0,\ (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\Sigma\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1})$$

3. Forecasting $w'\hat{\beta}_{\text{OLS}}$ and $w'\hat{\beta}_{\text{OLS}} \sim \mathcal{N}(w'\beta_0,\ w'V_{\hat{\beta}_{\text{OLS}}}w)$

4. Marginal (predictive) effects, $\frac{\partial}{\partial x_\ell}\hat{f}(X) = \sum_{k=1}^{K}\frac{\partial}{\partial x_\ell}g_k(X)\hat{\beta}_{\text{OLS},k}$

This time

The rest of this topic will cover various practical uses of regression:

$\rightarrow$ Interpreting regression coefficients

$\rightarrow$ Different explanatory variables to include in W_i and how to interpret them

- Polynomials, Indicators, Discrete Variables, Bins, Splines

$\rightarrow$ Interactions

$\rightarrow$ Partially linear model

$\rightarrow$ $\log$ transformed variables

Basically, this set of slides is the *applied* half of cross-sectional regression.

Multivariate Regression – “All Else Equal”

Regressors

- Polynomials
- Indicators
- Multi-valued discrete variables
- Interactions
- Bins
- Splines

log transformations

Binary outcomes

Cluster-robust Standard Errors

Marginal Effects with Multiple Variables

Say we have two variables in our linear model $\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 + X_2$.

Our predictions for unit i and unit j differ by

$$\hat{y}_j - \hat{y}_i = \beta_1(X_{1,j} - X_{1,i}) + \beta_2(X_{2,j} - X_{2,i})$$

Marginal Effects with Multiple Variables

$$\hat{y}_j - \hat{y}_i = \beta_1(X_{1,j} - X_{1,i}) + \beta_2(X_{2,j} - X_{2,i})$$

Let's think about a simple version. Take two individuals with the same X_2, but X_1 that differs by 1 unit (say $X_{1,j} - X_{1,i} = 1$)

Marginal Effects with Multiple Variables

$$\hat{y}_j - \hat{y}_i = \beta_1(X_{1,j} - X_{1,i}) + \beta_2(X_{2,j} - X_{2,i})$$

Let's think about a simple version. Take two individuals with the same X_2, but X_1 that differs by 1 unit (say $X_{1,j} - X_{1,i} = 1$)

Then, our estimated change is

$$\Delta\hat{y} = \beta_1$$

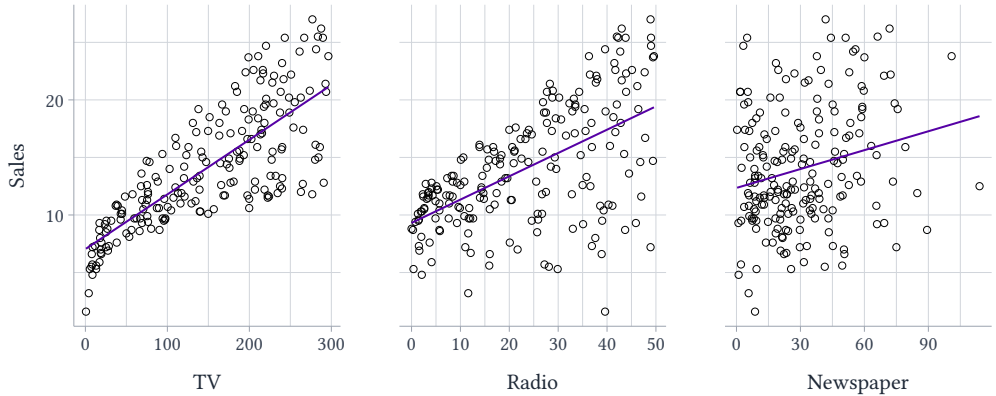
We refer to the β_k as "marginal effects", i.e. the predicted change in y from a 1 unit increase in X, holding *fixed* all the other variables

Advertising Example – Why Multiple Variable Regression?

Let's given an example. Say you're a business and you want to use advertising to boost sales. You have a bunch of different markets (e.g. cities) and you have data on how you've spent your advertising budget (TV, Radio, and Newspaper) and the sales in that market

Compare the results of regressing sales on all three variables to bivariate regressions

Advertising Example – Bivariate Regressions



Dependent Variable:	Sales			
Constant	7.033***	9.312***	12.35***	2.939***
	(0.4578)	(0.3882)	(0.6338)	(0.3365)
TV	0.0475***			0.0458***
	(0.0027)			(0.0019)
Radio		0.2025***		0.1885***
		(0.0217)		(0.0108)
Newspaper			0.0547***	-0.0010
			(0.0186)	(0.0064)
R^2	0.61188	0.33203	0.05212	0.89721

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Interpreting regression results

Notice on the last table, in the bivariate regression, higher newspaper ads spending is associated with higher sales

But, after controlling for TV and Radio ads, there is no relationship between Newspaper ads and sales

→ It seems like the bivariate relationship is being driven by newspaper ads being correlated with TV and Radio sales.

- Holding them fixed removed the relationship between Newspaper ads and sales

“All Else Equal”

The latin word you’ll sometimes see is *ceteris parabus* meaning “other things being equal”. This refers to variables *included* in your model

→ In this case, we are “holding fixed” TV and Radio advertising

→ Other things not included in the model still are moving as you compare areas with different Newspaper spending

For example, maybe cities with more newspaper spending have an older population than those with less

→ OLS is assigning ‘credit’ to Newspaper for the effect of different age distributions in a city on sales

Multivariate Regression – “All Else Equal”

Regressors

- Polynomials
- Indicators
- Multi-valued discrete variables
- Interactions
- Bins
- Splines

log transformations

Binary outcomes

Cluster-robust Standard Errors

Multivariate Regression – “All Else Equal”

Regressors

- Polynomials
- Indicators
- Multi-valued discrete variables
- Interactions
- Bins
- Splines

log **transformations**

Binary outcomes

Cluster-robust Standard Errors

Quadratic terms

Say you have the following model with wages as a quadratic function of age

$$w_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{age}_i^2 + u_i$$

Before we were discussing the idea of changing one variable, while holding the rest "equal"

$\rightarrow$ How does one change age without changing age^2?

This was why I focused on separating the variables you condition on X_i and the variables you include in your explanatory variables $W_i = (g_1(X_i), \ldots, g_K(X_i))'$

Marginal Effects

$$\frac{\partial}{\partial x_\ell}\hat{f}(X) = \sum_{k=1}^{K} \frac{\partial}{\partial x_\ell} g_k(X)\hat{\beta}_{\text{OLS},k}$$

This is holding fixed all the other variables at the original covariate values: $x_{1,i}, \ldots, x_{K,i}$ and only changing x_ℓ

$\rightarrow$ If a given g_k does not depend on x_ℓ, then $\frac{\partial}{\partial x_\ell} g_k(X) = 0$

$\rightarrow$ multiple W_k can change from changing a particular x_ℓ

Marginal effects with quadratic terms

$$\hat{w}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{age}_i + \hat{\beta}_2 \text{age}_i^2$$

From calculus, we know that the partial derivative of $\hat{w}_i$ with respect to age_i is given by

$$\frac{\partial}{\partial \text{age}_i} \hat{w}_i = \hat{\beta}_1 + 2\hat{\beta}_2 \text{age}_i$$

Marginal effects with quadratic terms

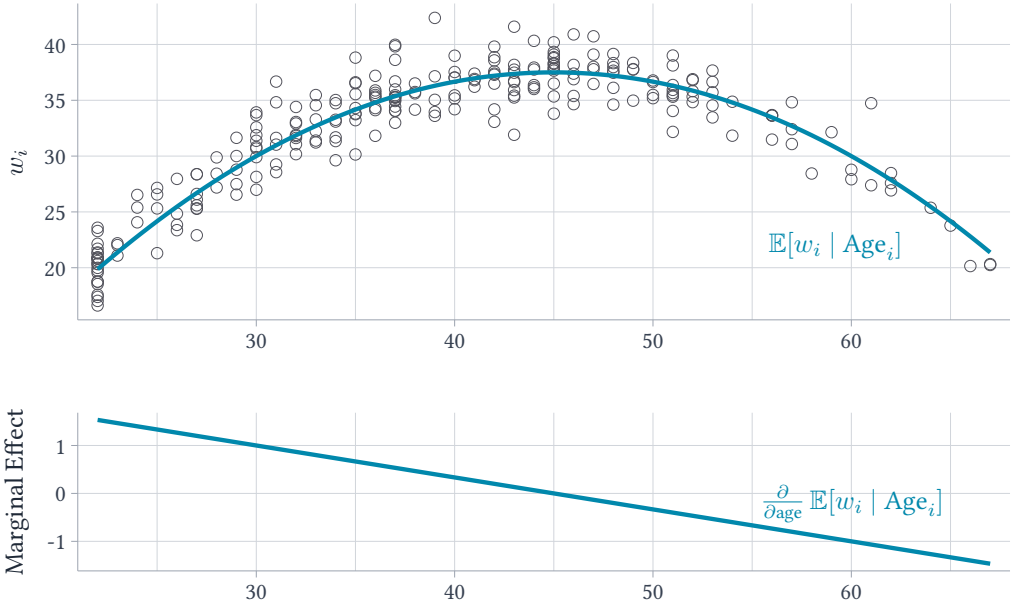
$$\hat{w}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{age}_i + \hat{\beta}_2 \text{age}_i^2$$

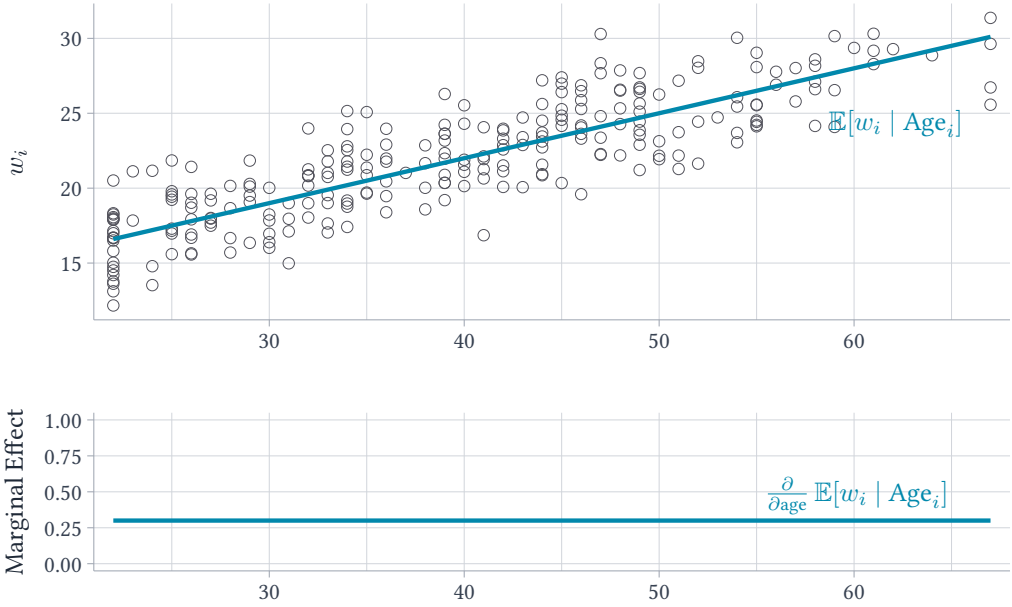
From calculus, we know that the partial derivative of $\hat{w}_i$ with respect to age_i is given by

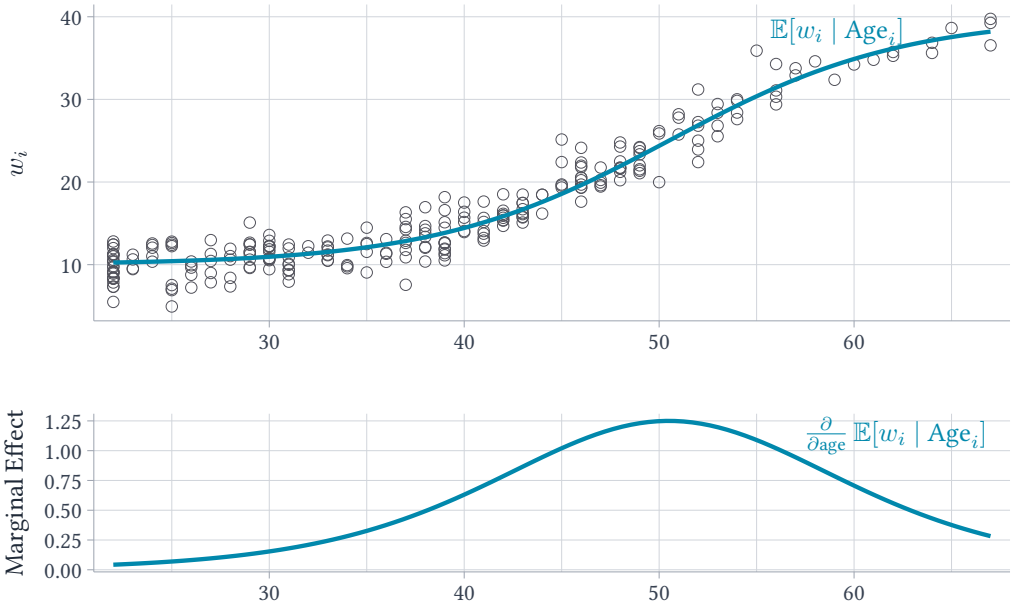
$$\frac{\partial}{\partial \text{age}_i} \hat{w}_i = \hat{\beta}_1 + 2\hat{\beta}_2 \text{age}_i$$

The change in predicted wage of a worker as they age is given by $\hat{\beta}_1 + 2\hat{\beta}_2 \text{age}_i$ which depends on their age

$\rightarrow$ In words, how much predicted wages change as a worker gets a year older changes over a worker's lifetime


w_i
40
35
30
25
20
30
40
50
60
$\mathbb{E}[w_i \mid \text{Age}_i]$
Marginal Effect
1
0
-1
30
40
50
60
$\frac{\partial}{\partial \text{age}} \mathbb{E}[w_i \mid \text{Age}_i]$


w_i
30
25
20
15
30
40
50
60
$\mathbb{E}[w_i \mid \text{Age}_i]$
Marginal Effect
1.00
0.75
0.50
0.25
0.00
30
40
50
60
$\frac{\partial}{\partial \text{age}} \mathbb{E}[w_i \mid \text{Age}_i]$


w_i
40
30
20
10
30
40
50
60
$\mathbb{E}[w_i \mid \text{Age}_i]$
Marginal Effect
1.25
1.00
0.75
0.50
0.25
0.00
30
40
50
60
$\frac{\partial}{\partial \text{age}} \mathbb{E}[w_i \mid \text{Age}_i]$

Testing quadratic term

$$\hat{w}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{age}_i + \hat{\beta}_2 \text{age}_i^2$$

In general, a linear term is much easier to interpret than a quadratic term: 'a one unit change is associated with a $\hat{\beta}_1$ unit change in w'

$\rightarrow$ We want to test whether the quadratic term is necessary or not

Testing quadratic term

$$\hat{w}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{age}_i + \hat{\beta}_2 \text{age}_i^2$$

You can test whether there is a significant quadratic relationship by testing the null that $H_0 : \beta_2 = 0$. Since we know $\hat{\beta}_2$ has a normal distribution, can use a t-test:

$$\hat{t} \equiv \frac{\hat{\beta}_2 - 0}{\text{SE}(\hat{\beta}_2)}$$

If the p-value associated with this is larger than the level of significance, then we can not reject the simple linear model.

Higher-order polynomials

This logic can be extended to higher-order polynomials to better estimate "wiggly" relationships between a given X and y

$\rightarrow$ A key property of polynomials is that they are smooth functions

$\rightarrow$ Any smooth CEF can be approximated well by a polynomial of high-enough order (Taylor expansion)

Higher-order polynomials

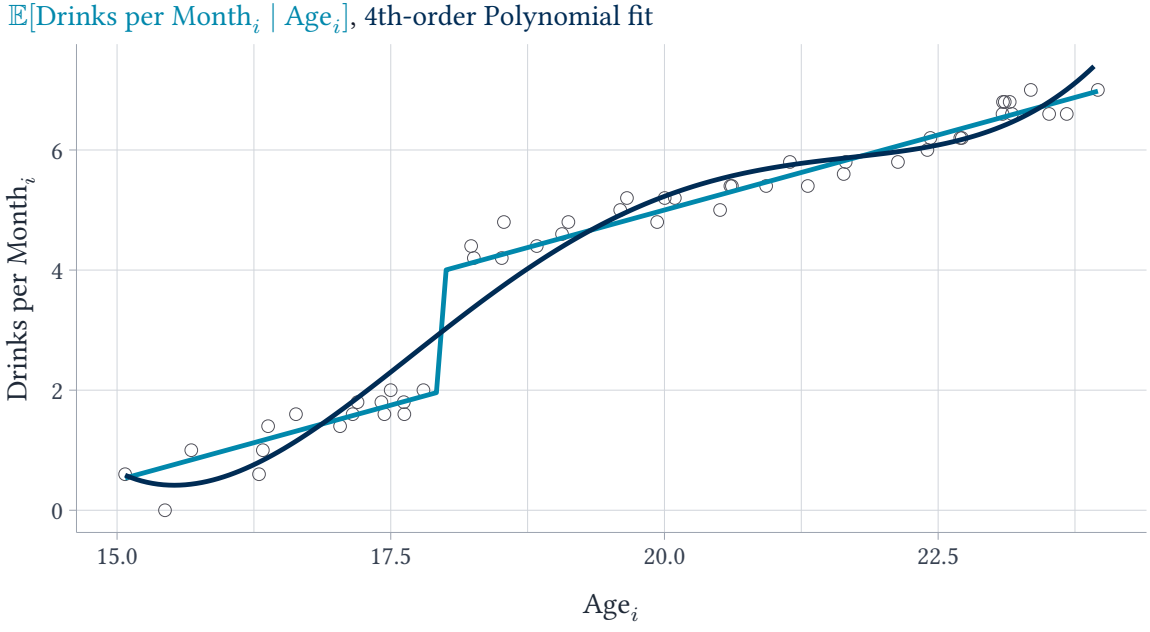
This logic can be extended to higher-order polynomials to better estimate "wiggly" relationships between a given X and y

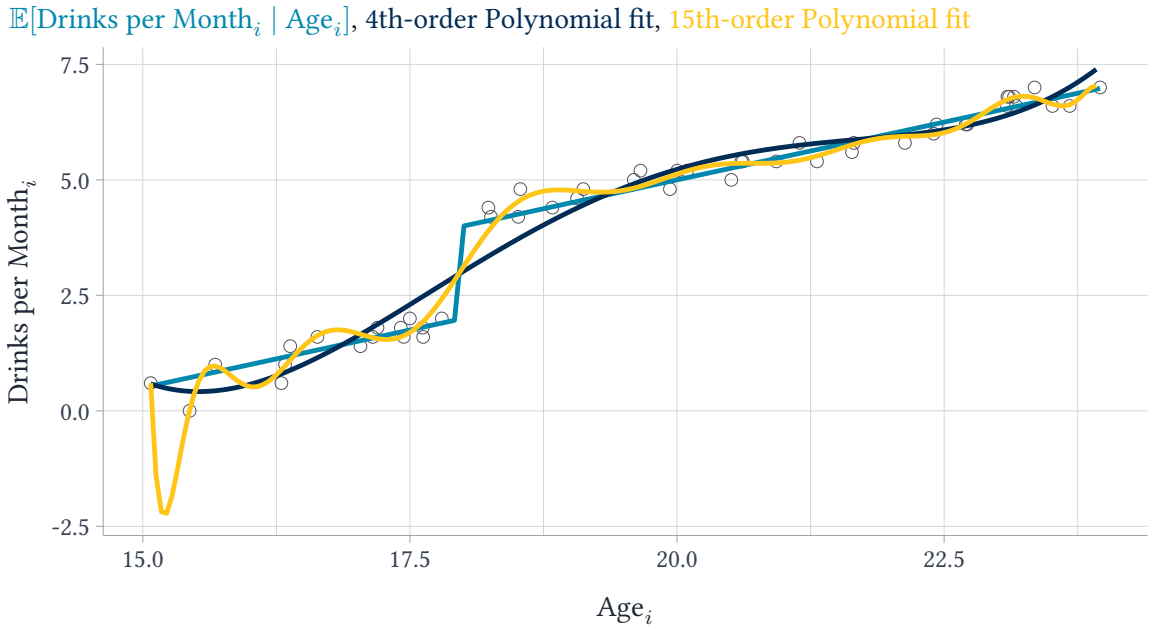
$\rightarrow$ A key property of polynomials is that they are smooth functions

$\rightarrow$ Any smooth CEF can be approximated well by a polynomial of high-enough order (Taylor expansion)

The smoothness can be a feature if you think outcomes evolve smoothly with a given X

$\rightarrow$ But, it can be a problem if you think there's a discontinuity present (e.g. turing 21 has a large jump in rates of DUI)


$\mathbb{E}[\text{Drinks per Month}_i \mid \text{Age}_i]$, 4th-order Polynomial fit
Drinks per Month$_i$
0
2
4
6
15.0
17.5
20.0
22.5
Age$_i$


$\mathbb{E}[\text{Drinks per Month}_i \mid \text{Age}_i]$, 4th-order Polynomial fit, 15th-order Polynomial fit
7.5
5.0
2.5
0.0
-2.5
Drinks per Month$_i$
15.0
17.5
20.0
22.5
Age$_i$

Polynomials and forecasting

One thing to note about polynomials is that they will always shoot off to $\pm\infty$ as you move $X \to \infty$ and $-\infty$

When you forecast using your model *outside* of the values of X you train on, this is called extrapolation

$\to$ Even for relatively small values outside the samples' domain of X can have very strange forecastes

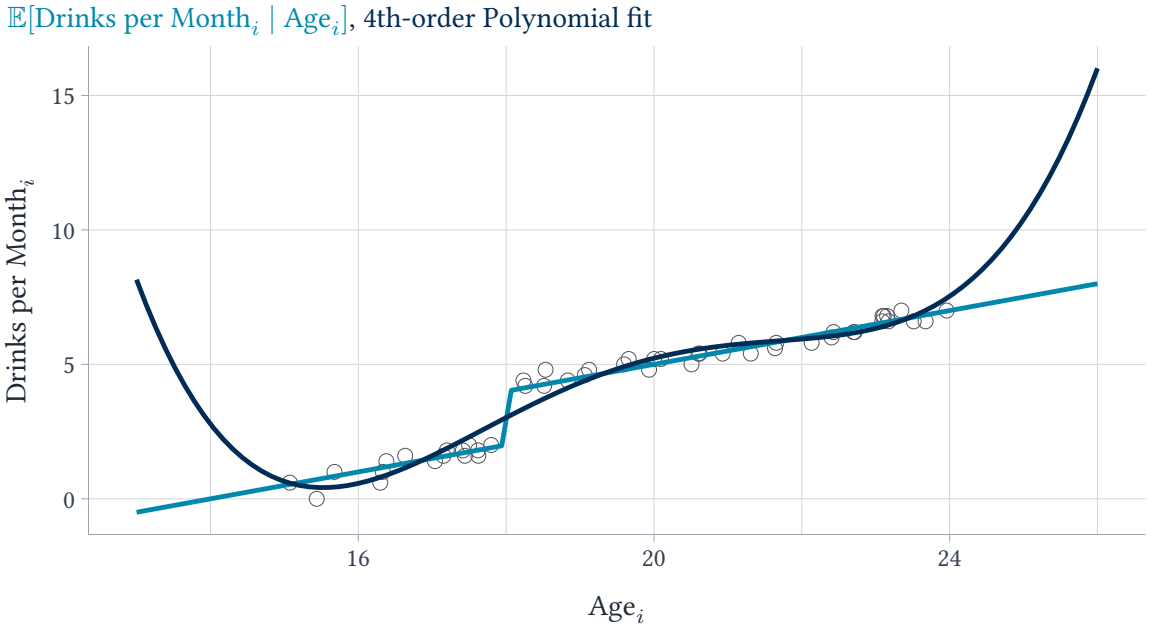
Polynomials and forecasting

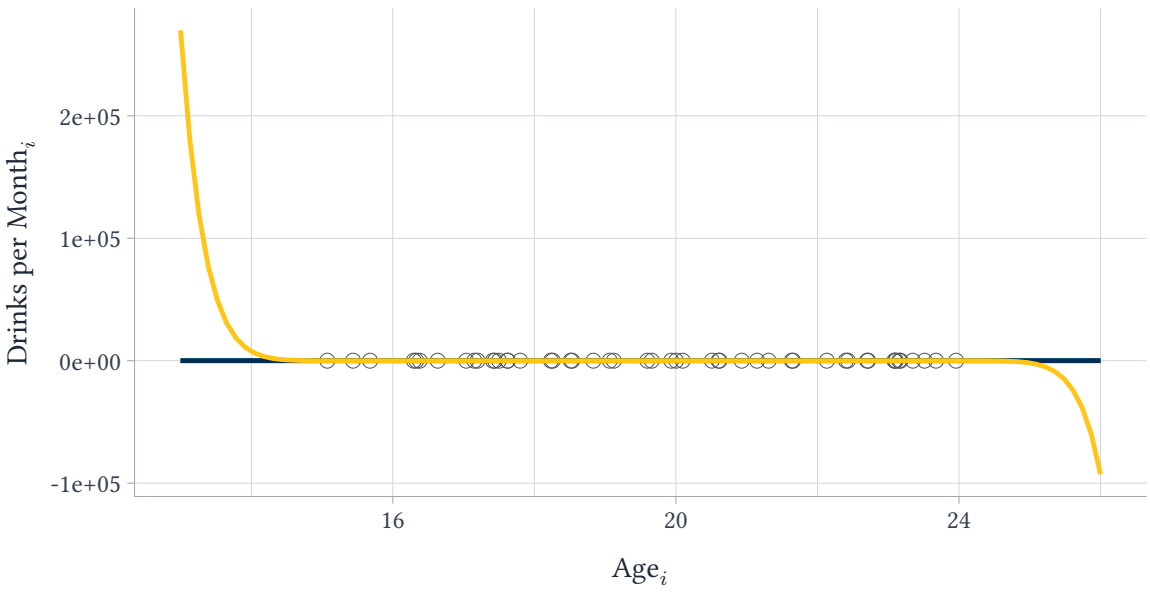
One thing to note about polynomials is that they will always shoot off to $\pm\infty$ as you move $X \to \infty$ and $-\infty$

When you forecast using your model *outside* of the values of X you train on, this is called extrapolation

$\to$ Even for relatively small values outside the samples' domain of X can have very strange forecastes

The higher the order of the polynomial, the worse this gets


$\mathbb{E}[\text{Drinks per Month}_i \mid \text{Age}_i]$, 4th-order Polynomial fit
Drinks per Month$_i$
15
10
5
0
16
20
24
Age$_i$


$\mathbb{E}[\text{Drinks per Month}_i \mid \text{Age}_i]$, 4th-order Polynomial fit, 15th-order Polynomial fit
Drinks per Month$_i$
2e+05
1e+05
0e+00
-1e+05
16
20
24
Age$_i$

Multivariate Regression – “All Else Equal”

Regressors

- Polynomials
- Indicators
- Multi-valued discrete variables
- Interactions
- Bins
- Splines

log **transformations**

Binary outcomes

Cluster-robust Standard Errors

Intercept-only regression estimates sample mean

```
feols(y ~ 1, data = df)
```

Running this regression is useful because it will estimate the mean of y and give us the standard error estimate $\frac{\hat{\sigma}}{\sqrt{n}}$

$\rightarrow$ This makes inference easier: hypothesis testing and confidence intervals

Comparing means using indicator variables

An *indicator variable* is a variable that can only equal 0 and 1, splitting units into groups

$\rightarrow$ X "indicates" when a unit is of type 0 or type 1

$\rightarrow$ Often, it is written as $\mathbb{1}[\cdot]$ where '$\cdot$' is a true/false condition

E.g. include

$\rightarrow$ being born male ($= 1$) or female ($= 0$)

$\rightarrow$ being White ($= 1$) or not ($= 0$)

$\rightarrow$ having a high-school degree ($= 1$) or not ($= 0$)

$\rightarrow$ $\mathbb{1}[\text{Height}_i \geq 6]$, being at least 6 foot tall

Indicator variable

Let's work through some properties of an indicator variable, D_i. First, the sample mean of an indicator variable is the proportion of units with a 1:

$$\frac{1}{n}\sum_{i=1}^{n} D_i = \frac{\#\text{ of } 1s}{n} = \%\text{ of sample with } D_i = 1$$

Indicator variable

Let's work through some properties of an indicator variable, D_i. First, the sample mean of an indicator variable is the proportion of units with a 1:

$$\frac{1}{n}\sum_{i=1}^{n} D_i = \frac{\#\text{ of } 1s}{n} = \%\text{ of sample with } D_i = 1$$

Define π as the (population) fraction of units with $D_i = 1$. Using the formula for variance, we can derive:

$$\text{Var}(D_i) = \pi(1-\pi)$$

Covariance with an indicator variable

What is Cov(D_i, Y_i)? Rembmer

$$\text{Cov}(D_i, Y_i) = \mathbb{E}[D_i Y_i] - \mathbb{E}[D_i]\mathbb{E}[Y_i]$$

Covariance with an indicator variable

What is $\text{Cov}(D_i, Y_i)$? Rembmer

$$\text{Cov}(D_i, Y_i) = \mathbb{E}[D_i Y_i] - \mathbb{E}[D_i]\mathbb{E}[Y_i]$$

Again, skipping the math:

$$\text{Cov}(D_i, Y_i) = \pi(1-\pi)\left(\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]\right)$$

Covariance with an indicator variable

Math details (if you're curious):

$$\begin{aligned}
\text{Cov}(D_i, Y_i) &= \mathbb{E}[D_i Y_i] - \mathbb{E}[D_i]\mathbb{E}[Y_i] \\
&= \mathbb{E}[D_i Y_i] - \pi \left(\pi \mathbb{E}[Y_i \mid D_i = 1] + (1-\pi)\mathbb{E}[Y_i \mid D_i = 0]\right) \\
&= \pi \mathbb{E}[Y_i \mid D_i = 1] - \pi\pi \mathbb{E}[Y_i \mid D_i = 1] - \pi(1-\pi)\mathbb{E}[Y_i \mid D_i = 0] \\
&= \pi(1-\pi)\left(\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]\right)
\end{aligned}$$

Regression with an indicator variable

Say you have a regression of $Y_i = \beta_0 + \beta_1 * D_i + u_i$. What does $\hat{\beta}_0$ and $\hat{\beta}_1$ equal?

$$\hat{\beta}_1 = \frac{\text{Cov}(D_i, Y_i)}{\text{Var}(D_i)}$$

Regression with an indicator variable

Say you have a regression of $Y_i = \beta_0 + \beta_1 * D_i + u_i$. What does $\hat{\beta}_0$ and $\hat{\beta}_1$ equal?

$$\hat{\beta}_1 = \frac{\text{Cov}(D_i, Y_i)}{\text{Var}(D_i)} = \frac{\pi(1-\pi)\left(\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]\right)}{\pi(1-\pi)}$$

$$= \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$$

The coefficient $\hat{\beta}_1$ tells me the difference in sample means between the group with $D_i = 1$ and the group with $D_i = 0$

Regression with an indicator variable

Say you have a regression of $Y_i = \beta_0 + \beta_1 * D_i + u_i$. From the last slide, we have:

$$\hat{\beta}_1 = \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$$

Solving our other first-order condition for $\hat{\beta}_0$, we have:

$$\begin{aligned}
\beta_0 &= \mathbb{E}[Y_i] - \hat{\beta}_1 \mathbb{E}[D_i] \\
&= \mathbb{E}[Y_i] - \hat{\beta}_1 \pi \\
&= \underbrace{\pi \mathbb{E}[Y_i \mid D_i = 1] + (1 - \pi)\mathbb{E}[Y_i \mid D_i = 0]} - \pi \mathbb{E}[Y_i \mid D_i = 1] - \pi \mathbb{E}[Y_i \mid D_i = 0] \\
&= \mathbb{E}[Y_i \mid D_i = 0]
\end{aligned}$$

Marginal effects

Our forecast is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i$. Since D_i can only be 0 or 1, our marginal effect just compares these values directly:

$$D_i = 0: \qquad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 * 0 = \hat{\beta}_0$$

$$D_i = 1: \qquad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 * 1 = \hat{\beta}_0 + \hat{\beta}_1$$

$\hat{\beta}_0$ is our predicted value for Y_i for the group with $D_i = 0$ and $\hat{\beta}_0 + \hat{\beta}_1$ is our predicted value for Y_i for the group with $D_i = 1$

$\rightarrow$ This makes $\hat{\beta}_1$ is the *difference in the means* between those with $D_i = 1$ compared to $D_i = 0$

Interpreting coefficient on an indicator

Our forecast is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i$. We can interpret $\hat{\beta}_1$ as follows:

> "On average, someone with $D_i = 1$ has a $\hat{\beta}_1$ larger/smaller value of $\hat{Y}$ compared to the units with $D_i = 0$ *holding all else equal*"

$\rightarrow$ Where you add the last part if you include additional covariates

Example

Let's revisit our example with the `mtcars` dataset. There is an indicator variable, `am` for being an automatic ($= 1$) or manual ($= 0$). Regress the miles per gallon a car gets, `mpg`, on `am`.

$\rightarrow$ In `fixest`, we can use `i(am)` to make it print out more nicely

```
with(mtcars, mean(mpg[am == 0]))
#> [1] 17.14737
with(mtcars, mean(mpg[am == 1]) - mean(mpg[am == 0]))
#> [1] 7.24494

feols(mpg ~ i(am), data = mtcars)
#> OLS estimation, Dep. Var.: mpg
#> Observations: 32
#> Standard-errors: IID
#>             Estimate Std. Error  t value   Pr(>|t|)
#> (Intercept) 17.14737    1.12460 15.24749 1.1340e-15 ***
#> am::1        7.24494    1.76442  4.10613 2.8502e-04 ***
```

Multivariate Regression – “All Else Equal”

Regressors

- Polynomials
- Indicators
- Multi-valued discrete variables
- Interactions
- Bins
- Splines

log **transformations**

Binary outcomes

Cluster-robust Standard Errors

Multi-valued discrete variables

This intuition will extend directly to settings where we have a discrete variable that obtains K distinct values:

$\rightarrow$ E.g. race/ethnicity, 10-year bins of age, number of cylinders in engine

Multi-valued discrete variables

This intuition will extend directly to settings where we have a discrete variable that obtains K distinct values:

$\rightarrow$ E.g. race/ethnicity, 10-year bins of age, number of cylinders in engine

We can construct a *set of* indicator variables for each value that X can obtain. For $k = 1, \ldots, K$

$$X_{ik} \equiv \mathbb{1}[X_i = x_k]$$

There are K such variables: $X_{i1}, \ldots, X_{iK}$

Multi-valued variable regression

$$y_i = \sum_{k=1}^{K} X_{ik}\beta_k + u_i$$

From the same intuition as before, we have $\hat{\beta}_k$ is the sample average of y_i for individuals with $X_i = x_k$

$\rightarrow$ We are in a very special case since these variables are mutually exclusive (only one of them is non-zero per unit), so this is easy to show with matrix algebra

Example

Let's revisit our example with the `mtcars` dataset. Let's see if `mpg` differs based on the number of cylinders a car has, `cyl`.

→ In `fixest`, we can use `i(cyl)` to make indicators for each value of a variable

→ Otherwise, we could for 4, 6, and 8 create the indicator variables with e.g.

```
mtcars$cyl4 = (mtcars$cyl == 4)
```

Interpret these coefficients:

```
feols(mpg ~ 0 + i(cyl), data = mtcars)
#> OLS estimation, Dep. Var.: mpg
#> Observations: 32
#> Standard-errors: IID
#>         Estimate Std. Error t value   Pr(>|t|)
#> cyl::4  26.6636   0.971801 27.4373  < 2.2e-16 ***
#> cyl::6  19.7429   1.218217 16.2064 4.4933e-16 ***
#> cyl::8  15.1000   0.861409 17.5294  < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> RMSE: 3.0683   Adj. R2: 0.704784
```

Intercept and Multicollinearity

In reality, we will want to include many predictor variables (beyond our single multi-valued discrete variable). In this case, we want to include an *intercept*

$$y_i = \beta_0 + \sum_{k=1}^{K} X_{ik}\beta_k + u_i$$

Multicollinearity

Our matrix $\boldsymbol{W}$ look like this. Note the 3 `cyl` indicator variables sum to the intercept

```
#>  (Intercept)   cyl::4   cyl::6   cyl::8
#>            1        0        1        0
#>            1        0        1        0
#>            1        1        0        0
#>            1        0        1        0
#>            1        0        0        1
#>            1        0        1        0
#>            1        0        0        1
#>            1        1        0        0
```

Multicollinearity

$$\hat{y}_i = \hat{\beta}_0 + \sum_{k=1}^{K} X_{ik}\hat{\beta}_k$$

It turns out that we face a non-uniqueness problem because of the multicollinearity we identified

$\rightarrow$ We can add 10 to $\hat{\beta}_0$ and subtract 10 from $\hat{\beta}_4$, $\hat{\beta}_6$, and $\hat{\beta}_8$ and get the exact same $\hat{y}$

Therefore, we will typically need to drop one of the X_{ik} variables (or R will do it for you)

```
feols(mpg ~ 1 + i(cyl), data = mtcars)
#> OLS estimation, Dep. Var.: mpg
#> Observations: 32
#> Standard-errors: IID
#>              Estimate Std. Error  t value   Pr(>|t|)
#> (Intercept)  26.66364   0.971801 27.43735  < 2.2e-16 ***
#> cyl::6       -6.92078   1.558348 -4.44110 1.1947e-04 ***
#> cyl::8      -11.56364   1.298623 -8.90453 8.5682e-10 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> RMSE: 3.0683   Adj. R2: 0.714009
```

Interpreting Multicollinearity

In the previous example, we dropped $\mathbb{1}[X_i = 4]$. This is the omitted group. What happened to our coefficient estimates?

Just like in the indicator variable case, $\hat{\beta}_0$ estimated the mean of `mpg` for cars with $X_i = 4$

Interpreting Multicollinearity

In the previous example, we dropped $\mathbb{1}[X_i = 4]$. This is the omitted group. What happened to our coefficient estimates?

Just like in the indicator variable case, $\hat{\beta}_0$ estimated the mean of `mpg` for cars with $X_i = 4$

The coefficients on the other $\hat{\beta}_k$ now represent the *difference* in means between the group for $X_i = 6$ and the 'omitted group' $X_i = 4$.

$\rightarrow$ The mean for $X_i = 6$ is $19.742 = 26.663 - 6.921$

Specifying `ref` option

```
feols(mpg ~ i(cyl, ref = 6), data = mtcars)
#> OLS estimation, Dep. Var.: mpg
#> Observations: 32
#> Standard-errors: IID
#>             Estimate Std. Error  t value   Pr(>|t|)
#> (Intercept) 19.74286    1.21822 16.20636 4.4933e-16 ***
#> cyl::4       6.92078    1.55835  4.44110 1.1947e-04 ***
#> cyl::8      -4.64286    1.49200 -3.11182 4.1522e-03 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#> RMSE: 3.0683   Adj. R2: 0.714009
```

Significance with indicator variables

```
#>              Estimate Std. Error  t value   Pr(>|t|)
#> (Intercept)  26.66364   0.971801 27.43735  < 2.2e-16 ***
#> cyl::6       -6.92078   1.558348 -4.44110 1.1947e-04 ***
#> cyl::8      -11.56364   1.298623 -8.90453 8.5682e-10 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Including an intercept also helps with certaint statistical inference. The estimates test that the average of y for the omitted group is *the same* for the other groups

$\rightarrow$ Rejecting this (p-value $< \alpha$) rejects the null that the two means are the same

Interpreting coefficient on indicators for a discrete variable

Our forecast is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_6 X_{i6} + \hat{\beta}_8 X_{i8}$. We can interpret $\hat{\beta}_6$ as follows:

> "On average, a car with 6 cylinders has a $\hat{\beta}_6$ larger/smaller value of mpg compared to cars with 4 cylinders *holding all else equal*"

$\rightarrow$ Where you add the last part if you include additional covariates

Multivariate Regression – “All Else Equal”

Regressors

- Polynomials
- Indicators
- Multi-valued discrete variables
- Interactions
- Bins
- Splines

log transformations

Binary outcomes

Cluster-robust Standard Errors

Why interactions

Wages Example

Consider a model where we want to understand how wages are influenced by both being a female and being a college-educated worker. We can write the model as:

$$w_i = \beta_0 + \beta_1 \text{female}_i + \beta_2 \text{college}_i + \beta_3 (\text{female}_i \times \text{college}_i) + u_i$$

Here, β_3 captures the interaction effect between female and college-education status on wages

$\rightarrow$ This means that the effect of females on wages may differ depending on whether the worker has a college-degree, and vice versa.

Interactions

Wages Example

Consider the difference in predicted wages for non-college educated male vs. non-college educated workers:

$$w_{i,NC,F} - w_{i,NC,M} = (\beta_0 + \beta_1) - \beta_0 = \beta_1$$

Compare this to the difference in predicted wages for college educated male vs. college educated workers:

$$w_{i,C,F} - w_{i,C,M} = (\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2) = \beta_1 + \beta_3$$

Interactions

Wages Example

Wage-gap for college educated workers is $\beta_1 + \beta_3$ and the wage-gap for non-college educated workers is β_1

$\rightarrow$ β_3 represents the difference in wage gaps of college-educated vs. non-college-educated workers.

Interactions

Wages Example

Wage-gap for college educated workers is $\beta_1 + \beta_3$ and the wage-gap for non-college educated workers is β_1

$\rightarrow$ β_3 represents the difference in wage gaps of college-educated vs. non-college-educated workers.

More generally, can interpret interactions how one variable changes the marginal effect of another variable

$\rightarrow$ This is similar to when you have a quadratic function of X, the marginal effect depends where you are along the distribution of X.

Interactions

Partial Derivative

$$\hat{w}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{female}_i + \hat{\beta}_2 \text{college}_i + \hat{\beta}_3 (\text{female}_i \times \text{college}_i)$$

We can derive this result using partial derivatives:

$$\frac{\partial \hat{w}_i}{\partial \text{female}_i} = \hat{\beta}_1 + \hat{\beta}_3 \text{college}_i$$

Interactions

Partial Derivative

$$\hat{w}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{female}_i + \hat{\beta}_2 \text{college}_i + \hat{\beta}_3 (\text{female}_i \times \text{college}_i)$$

We can derive this result using partial derivatives:

$$\frac{\partial \hat{w}_i}{\partial \text{female}_i} = \hat{\beta}_1 + \hat{\beta}_3 \text{college}_i$$

In this case, it's a little weird to think of "small" changes in the female variable. Instead, we will think of it as a 1 unit change (from 0 to 1)

Continuous interacted with a discrete variable

Let X_i be a continuous variable and D_i be a dummy variable and consider the regression

$$y_i = \beta_0 + D_i\beta_1 + X_i\beta_2 + X_iD_i\beta_3 + u_i$$

Continuous interacted with a discrete variable

Let X_i be a continuous variable and D_i be a dummy variable and consider the regression

$$y_i = \beta_0 + D_i\beta_1 + X_i\beta_2 + X_iD_i\beta_3 + u_i$$

In this case, the marginal effect of X_i is given by $\frac{\partial \hat{y}_i}{\partial X_i} = \hat{\beta}_2 + D_i\hat{\beta}_3$

$\rightarrow$ The marginal effect for group $D_i = 0$ is given by $\hat{\beta}_2$

$\rightarrow$ The marginal effect for group $D_i = 1$ is given by $\hat{\beta}_2 + \hat{\beta}_3$

Continuous interacted with a discrete variable

Let X_i be a continuous variable and D_i be a dummy variable and consider the regression

$$y_i = \beta_0 + D_i\beta_1 + X_i\beta_2 + X_iD_i\beta_3 + u_i$$

In this case, the marginal effect of X_i is given by $\frac{\partial \hat{y}_i}{\partial X_i} = \hat{\beta}_2 + D_i\hat{\beta}_3$

$\rightarrow$ The marginal effect for group $D_i = 0$ is given by $\hat{\beta}_2$

$\rightarrow$ The marginal effect for group $D_i = 1$ is given by $\hat{\beta}_2 + \hat{\beta}_3$

$\rightarrow$ Therefore, $\hat{\beta}_3$ is the difference in marginal effects between $D_i = 1$ relative to $D_i = 0$

Continuous interacted with a discrete variable

$$y_i = \beta_0 + D_i\beta_1 + X_i\beta_2 + X_iD_i\beta_3 + u_i$$

Exercise:

→ In words, how would you interpret a t-test for the null that $\hat{\beta}_2 = 0$?

→ In words, how would you interpret a t-test for the null that $\hat{\beta}_3 = 0$?

mtcars example

```
OLS estimation, Dep. Var.: mpg
              Estimate Std. Error   t value   Pr(>|t|)
(Intercept) 26.624848   1.346754 19.769644  < 2.2e-16 ***
am::1        5.217653   2.324898  2.244250 3.2904e-02 *
hp          -0.059137   0.008957 -6.602265 3.6781e-07 ***
am::1:hp     0.000403   0.013362  0.030152 9.7616e-01
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercise: Does the estimate relationship between being a car's horsepower and miles per gallon depend on whether it is an automatic? How do you know?

Interaction terms always should have ‘main effects’

When including an interaction term, it is important to (almost) always include the main effects

The main effects are the variables by themselves. E.g. if you interact gender with race, you want to include race and gender as separate terms as well

→ The main effects are what let us interpret the interaction term as the ‘difference’ in marginal effects

Continuous-Continuous interactions

Now consider two continuous variables being interacted:

$$y_i = \beta_0 + X_{1,i}\beta_1 + X_{2,i}\beta_2 + X_{1,i}X_{2,i}\beta_3 + u_i$$

This is common when you think there are complementarities between variables

→ E.g. y is crop-yield, X_1 is the amount of fertilizer applied, and X_2 is the amount of water given. Does it help to do more of both (complements)?

→ y is a measure of job performance, X_1 is a measure of on-the-job experience, and X_2 is a measure of training

Continuous-Continuous interactions

$$\frac{\partial \hat{y}_i}{\partial X_{1,i}} = \hat{\beta}_1 + X_{2,i}\hat{\beta}_3 \quad \text{and} \quad \frac{\partial \hat{y}_i}{\partial X_{2,i}} = \hat{\beta}_2 + X_{1,i}\hat{\beta}_3$$

Can interpret it in two ways:

1. The marginal effect of X_1 grows/shrinks with the value of X_2 (depending on the sign of $\hat{\beta}_3$)

Multivariate Regression – “All Else Equal”

Regressors

- Polynomials
- Indicators
- Multi-valued discrete variables
- Interactions
- Bins
- Splines

log transformations

Binary outcomes

Cluster-robust Standard Errors

Key explanatory variables

In a forecast model, we might have a bunch of different variables we want to use as predictors

Many times, we really want to see how the forecasted outcome varies with a *particular* variable of interest

$\rightarrow$ Want to "control" for a bunch of variables and then model the relationship between y and key variable, X_1, flexibly

Partially-linear model

The Partially linear model mixes high model flexibility in a key variable we care about and linear model for the rest of the covariates:

$$y_i = \mu(X_{1i}) + W_i'\beta + u_i$$

$\rightarrow$ $\mu(X_{1i})$ is a highly flexible function

$\rightarrow$ W_i is a set of *linear* control variables

This allows you to prevent the curse of dimensionality by linearly controlling for most of the variables. Allows a flexible model for the key variable of interest, X_i, that is good for graphing.

Chopping X_1 into bins

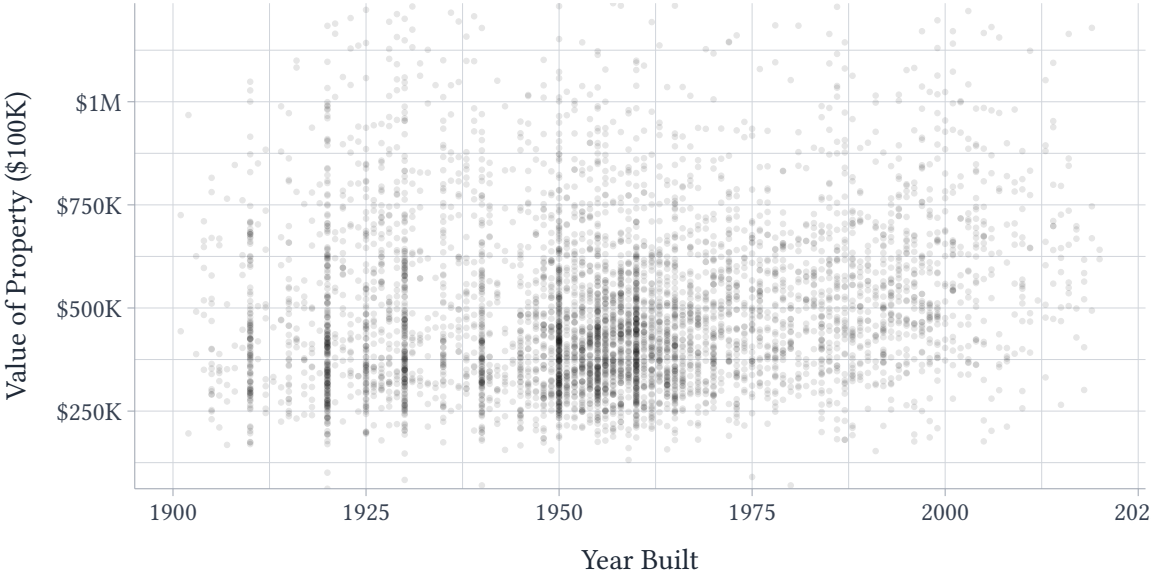
The first approach, is to take X_{1i} and make it into a discrete set of bins:

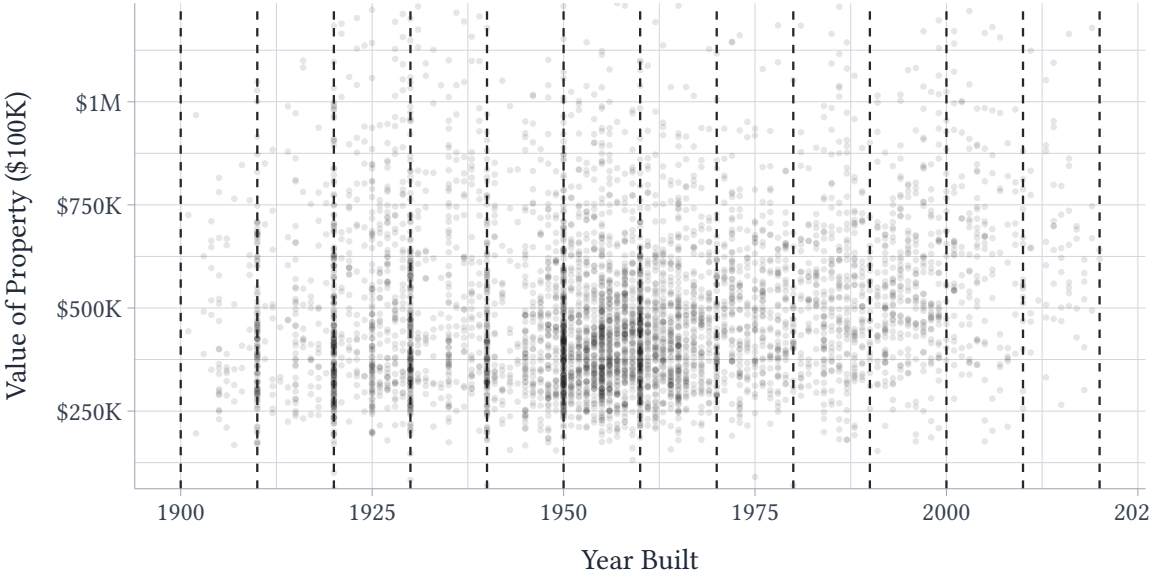
$\rightarrow$ E.g. Person's age into $20-24, 25-29, 30-34, \ldots$

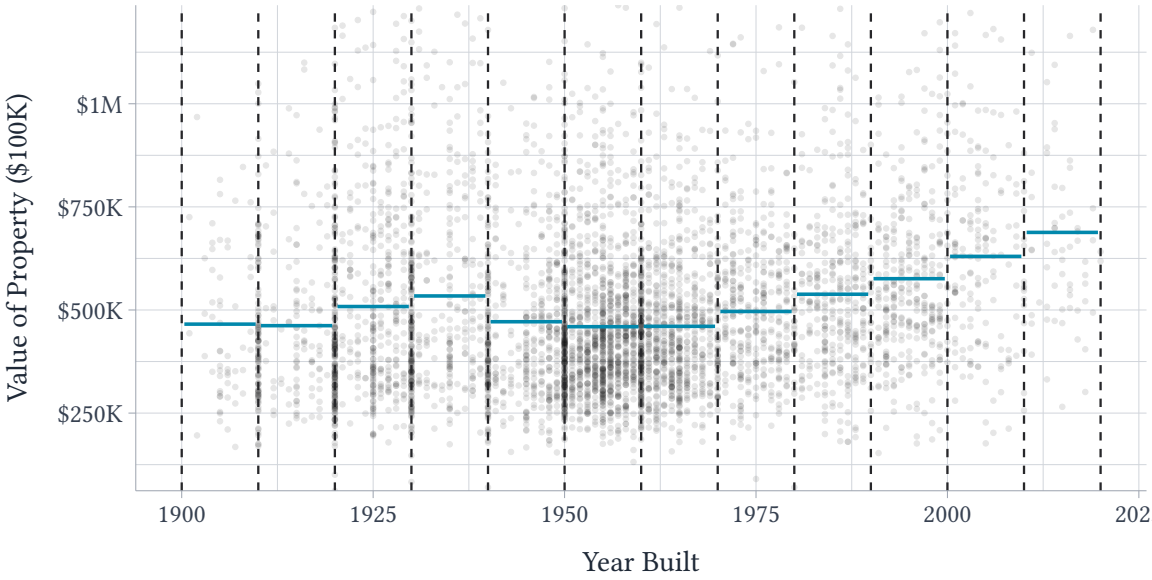
Then, we can treat this as a multi-valued discrete variable

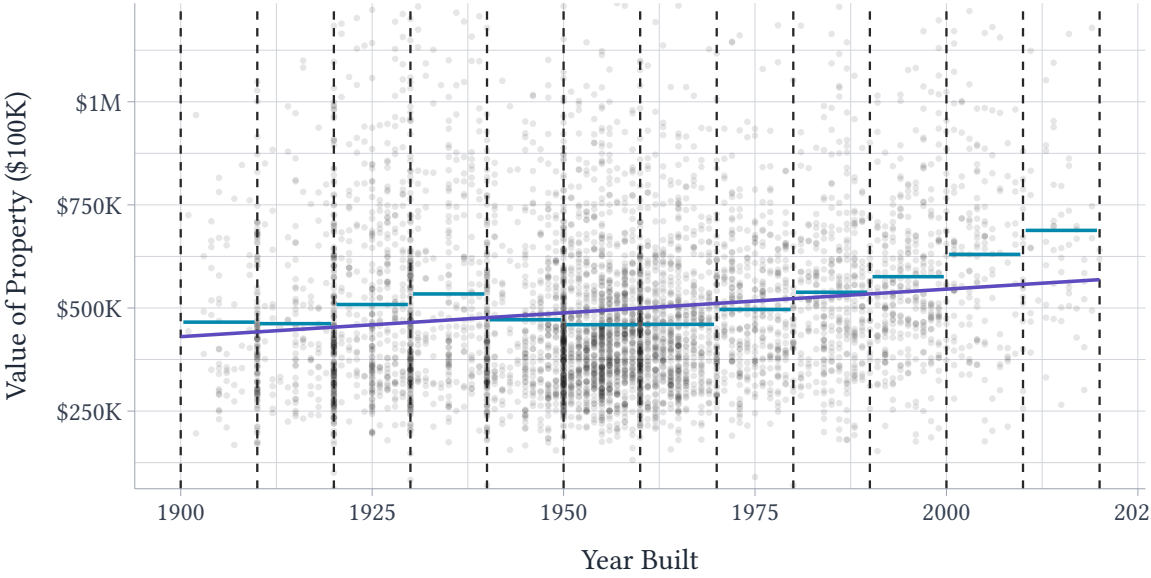
$\rightarrow$ Indicators for each bin and one omitted category

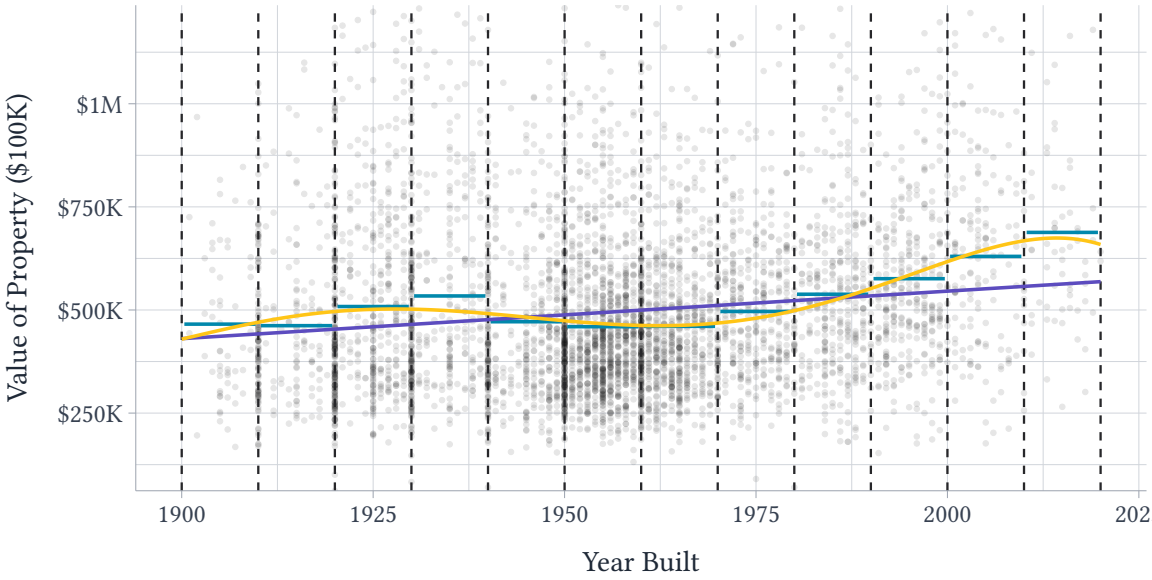
$\rightarrow$ We estimate sample means for each bin (relative to omitted group's mean)


Value of Property ($100K)
$1M
$750K
$500K
$250K
1900
1925
1950
1975
2000
2025
Year Built


Value of Property ($100K)
$1M
$750K
$500K
$250K
1900
1925
1950
1975
2000
2025
Year Built


Value of Property ($100K)
$1M
$750K
$500K
$250K
1900
1925
1950
1975
2000
2025
Year Built


Value of Property ($100K)
$1M
$750K
$500K
$250K
1900
1925
1950
1975
2000
2025
Year Built


Value of Property ($100K)
$1M
$750K
$500K
$250K
1900
1925
1950
1975
2000
2025
Year Built

Pros and Cons

One advantage is this is relatively easy to explain and quite flexible:

$\rightarrow$ "The average property value for homes built from 1950 – 1959 is $\$471,000$"

$\rightarrow$ If we want more flexibility, we can increase the number of bins (if we have enough data)

Pros and Cons

One advantage is this is relatively easy to explain and quite flexible:

→ "The average property value for homes built from 1950 – 1959 is $\$471,000$"

→ If we want more flexibility, we can increase the number of bins (if we have enough data)

One problem with this method is that the marginal effect estimates can be odd:

→ The estimated function is flat (0 marginal effect) except at the "knots" where there is instant jump

→ But, we are okay if we recognize this limitation

Multivariate Regression – “All Else Equal”

Regressors

- Polynomials
- Indicators
- Multi-valued discrete variables
- Interactions
- Bins
- Splines

log transformations

Binary outcomes

Cluster-robust Standard Errors

Previous approaches

$$y_i = \mu(X_{1i}) + W_i'\beta + u_i$$

We have discussed two ways to estimate $\mu(X_{1i})$ flexibly, both had pros and cons

$\rightarrow$ Polynomials allow for flexibility, but could become quite sensitive / noisy

$\rightarrow$ Bins are flexible and simple to explain, but created artificial discontinuities and non-smoothness in our estimate

Previous approaches

$$y_i = \mu(X_{1i}) + W_i'\beta + u_i$$

Splines are a way to try and blend the two approaches:

$\rightarrow$ Chop up the domain of X_1 into a set of bins

$\rightarrow$ Within each bin, estimate a polynomial of a given order (p)

$\rightarrow$ Possibly, you can require the end points to "connect" (s)

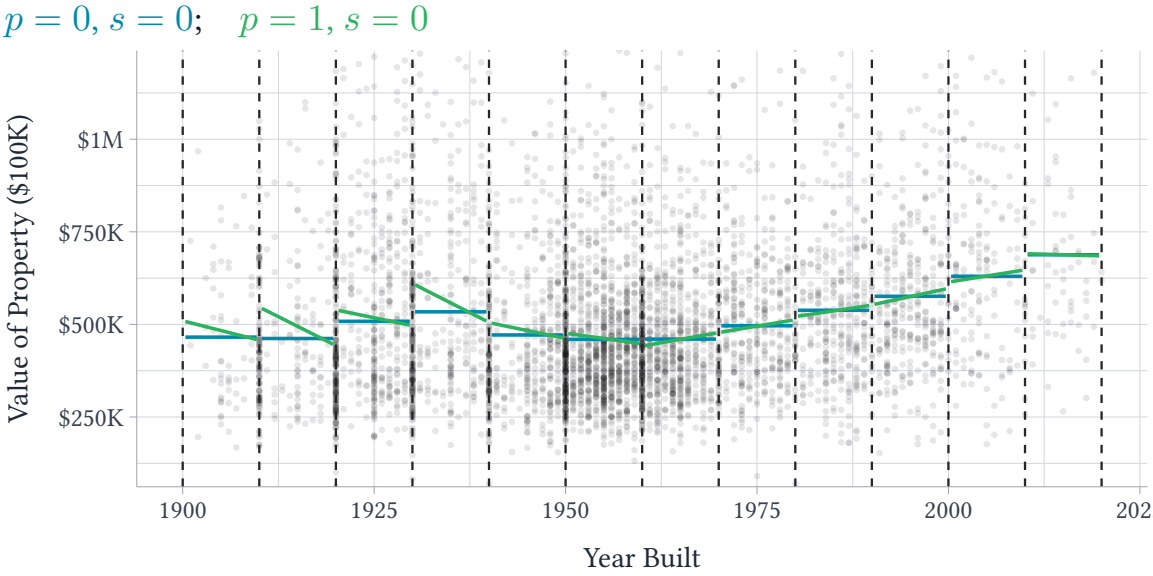
Polynomial order and smoothness

Our binned averages is a particular version of splines:

$\rightarrow$ Polynomial of order 0 ($p = 0$)

$\rightarrow$ Not requied to connect ($s = 0$)

We could increase the polynomial order to $p = 1$ to create linear functions within each bin


$p = 0, s = 0$;
$p = 1, s = 0$
Value of Property ($100K)
$1M
$750K
$500K
$250K
1900
1925
1950
1975
2000
2025
Year Built

Polynomial order and smoothness

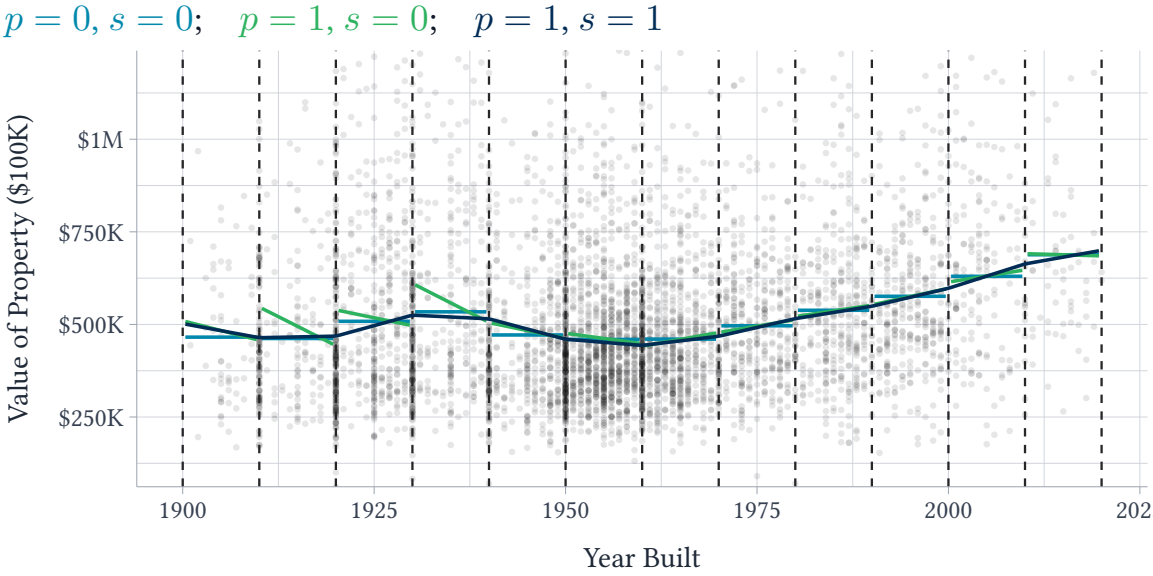
Our binned averages is a particular version of splines:

$\rightarrow$ Polynomial of order 0 ($p = 0$)

$\rightarrow$ Not requied to connect ($s = 0$)

We could increase the polynomial order to $p = 1$ to create linear functions within each bin

If we want the lines to connect, we can add a smoothness constraint $s = 1$


$p = 0, s = 0$;
$p = 1, s = 0$;
$p = 1, s = 1$
Value of Property ($100K)
$1M
$750K
$500K
$250K
1900
1925
1950
1975
2000
2025
Year Built

Polynomial order and smoothness

Our binned averages is a particular version of splines:

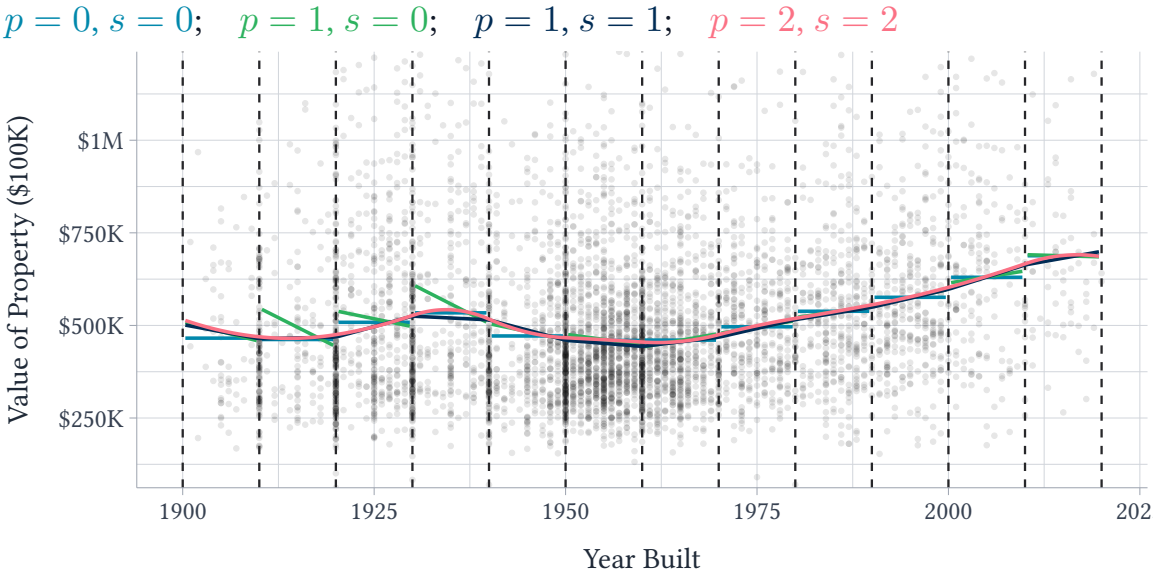
$\rightarrow$ Polynomial of order 0 ($p = 0$)

$\rightarrow$ Not requied to connect ($s = 0$)

We could increase the polynomial order to $p = 1$ to create linear functions within each bin

If we want the lines to connect, we can add a smoothness constraint $s = 1$

Or, we can set $p = 2$ and $s = 2$ to estimate quadratic functions and make them connect smoothly


$p = 0, s = 0$;
$p = 1, s = 0$;
$p = 1, s = 1$;
$p = 2, s = 2$
Value of Property ($100K)
$1M
$750K
$500K
$250K
1900
1925
1950
1975
2000
2025
Year Built

Spline advantages and costs

From the previous figure, it seems like we've hit a really nice *sweet spot*

→ Quite flexible model to estimate $\mu(X_{1i})$ and can easily add covariates W_i as controls

→ But, we can not really do this with *all* our covariates because splines add a lot of parameters (at least as many as bins you have)

Spline advantages and costs

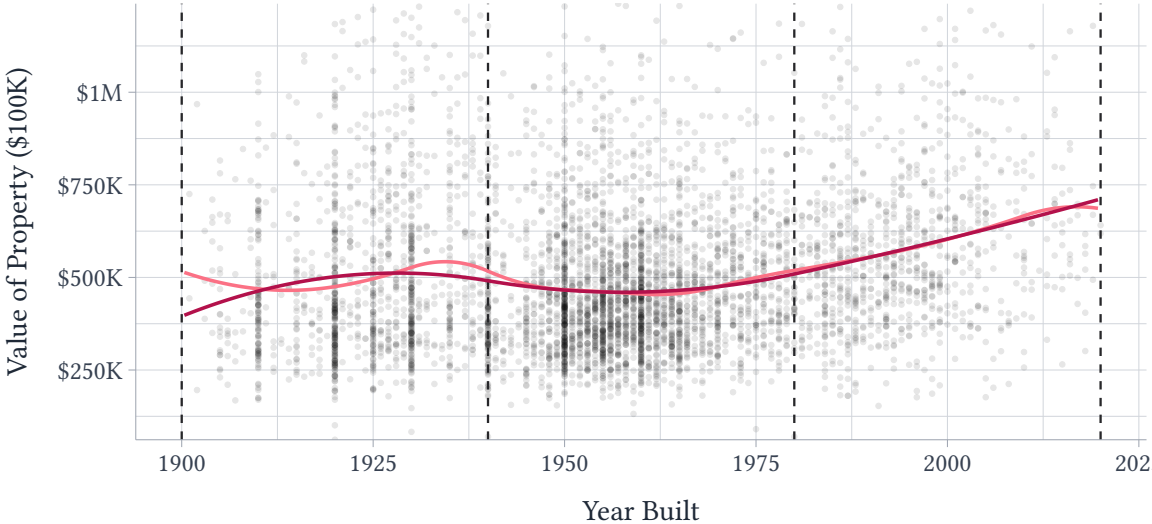
From the previous figure, it seems like we've hit a really nice *sweet spot*

→ Quite flexible model to estimate $\mu(X_{1i})$ and can easily add covariates W_i as controls

→ But, we can not really do this with *all* our covariates because splines add a lot of parameters (at least as many as bins you have)

It's not clear yet how we should chose the bins...

knots every 10 years; knots every 40 years



Choosing Bins

There is a *bias*/*variance* trade-off when selecting the number of bins

→ More bins means we can better approximate $\mu(X_{1i})$ (decreased bias)

→ But, with fewer observations per bin, our estimates will be more noisy (increased variance)

More, it is not clear why the bins should be evenly-space

→ It makes 'intuitive' sense, but maybe the bins should get smaller when the data is more 'wiggly' and larger where the data is less

Choosing Bins

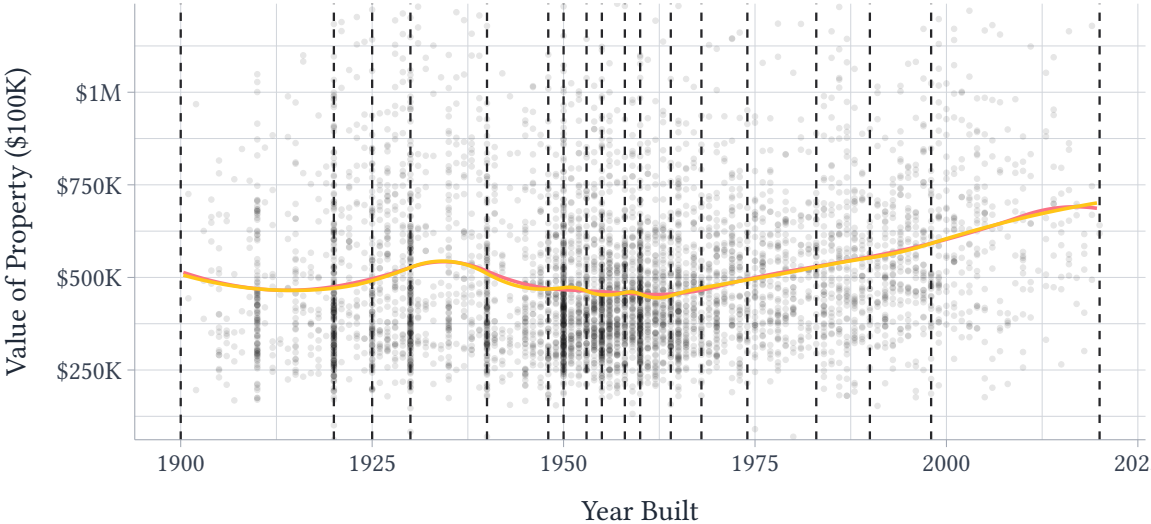
$$y_i = \mu(X_i) + W_i'\beta + u_i$$

One recent way of estimating this is using a 'binscatter' regression:

1. Chop X variable into J bins with an *equal number of observations into each bin*
2. Choose J *optimally* to minimize mean-squared prediction error (bias-variance trade-off)

Implemented in the `binsreg` package

knots every 10 years; binscatter selected



Multivariate Regression – “All Else Equal”

Regressors

- Polynomials
- Indicators
- Multi-valued discrete variables
- Interactions
- Bins
- Splines

log **transformations**

Binary outcomes

Cluster-robust Standard Errors

When to use $\log$ transformations

You should take the $\log$ of an outcome varible when you think a 1 unit change in X is related to a $\%$ change in Y.

When to use $\log$ transformations

You should take the $\log$ of an outcome varible when you think a 1 unit change in X is related to a $\%$ change in Y.

Settings with skewed distributions (e.g. home prices, GDP, population, income)

$\rightarrow$ Skewness makes a 'unit' change in X difficult to think about

When to use $\log$ transformations

You should take the $\log$ of an outcome varible when you think a 1 unit change in X is related to a $\%$ change in Y.

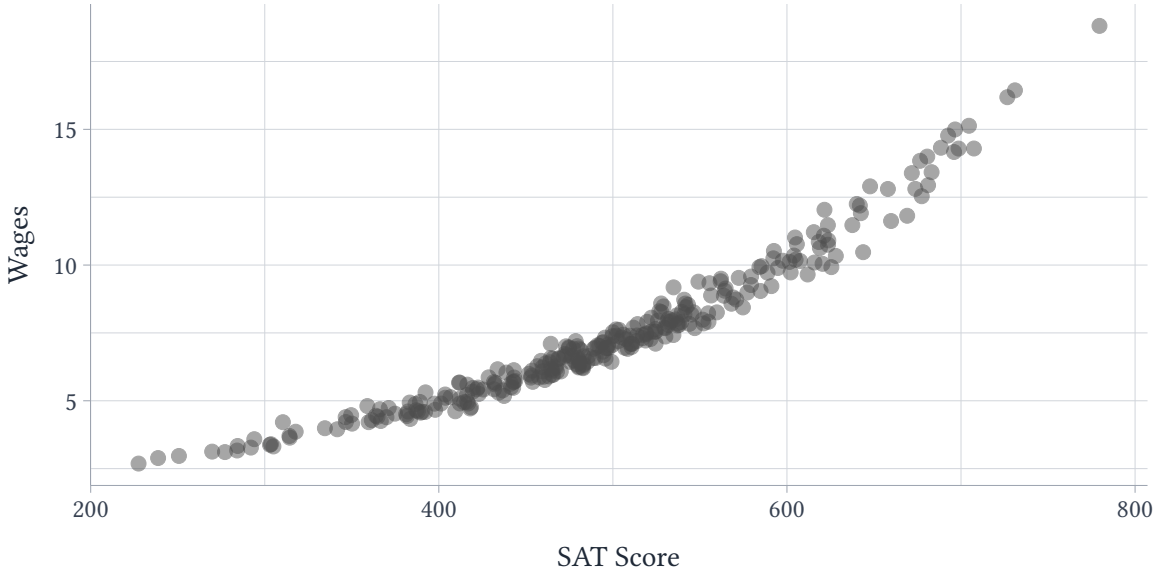
Settings with skewed distributions (e.g. home prices, GDP, population, income)

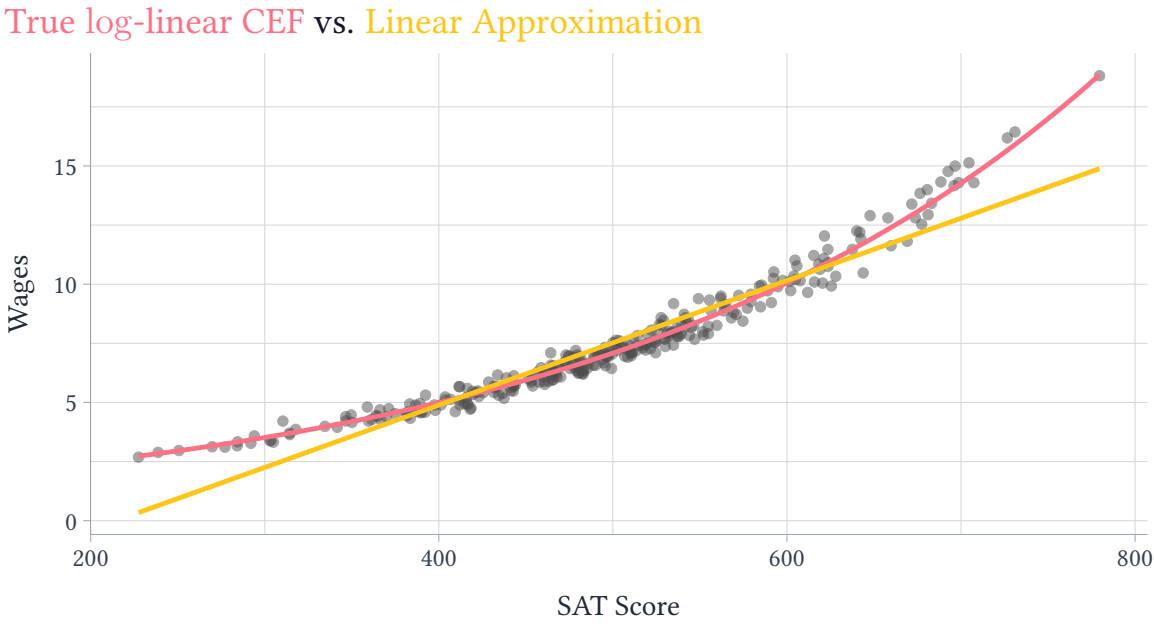
$\rightarrow$ Skewness makes a 'unit' change in X difficult to think about

$\log(Y) = X\beta \iff Y = e^{X\beta}$ is fitting an 'exponential' relationship. These are common in:

1. Financial markets where compounding returns imply $Y_t = Y_0 e^{rt}$
2. Epidemiology where disease growth rate is exponential (it is not actually, but early growth rate is approximately)

Data on SAT score and wages




True log-linear CEF vs. Linear Approximation
Wages
0
5
10
15
200
400
600
800
SAT Score

Extrapolating Exponential Relationship

If you thought extrapolating with polynomials with bad... Exponential is even more extreme

→ Things that have rapid growth often will tend to slow-down

→ Stonks don't go to the moon...

log-transformation

$$\log(\text{wages}_i) = \beta_0 + \beta_1 \text{College Degree}_i + u_i$$

This specification changes our interpretation of the slope coefficients:

> "Having a college degree is associated with an increase in wages of $\beta_1 * 100$ percent"

$\rightarrow$ E.g. if $\beta_1 = 0.02$, then a college degree is associated with a 2% increase in wages.

Derivation of $\log$-transformation interpretation

Compare two individuals: unit 1 with and unit 0 without a college degree. Then, we have

$$\log(\text{wages}_1) - \log(\text{wages}_0) = \beta_0 + \beta_1 - \beta_0$$

$$\implies \log(\text{wages}_1/\text{wages}_0) = \beta_1$$

$$\implies \log(1 + \frac{\text{wages}_1 - \text{wages}_0}{\text{wages}_0}) = \beta_1$$

Derivation of $\log$-transformation interpretation

Compare two individuals: unit 1 with and unit 0 without a college degree. Then, we have

$$\log(\text{wages}_1) - \log(\text{wages}_0) = \beta_0 + \beta_1 - \beta_0$$

$$\implies \log(\text{wages}_1/\text{wages}_0) = \beta_1$$

$$\implies \log(1 + \frac{\text{wages}_1 - \text{wages}_0}{\text{wages}_0}) = \beta_1$$

If you recall, exponentiating gets rid of the the $\log$

$$\frac{\text{wages}_1 - \text{wages}_0}{\text{wages}_0} = \exp(\beta_1) - 1$$

Derivation of $\log$-transformation interpretation

The left-hand side is our percent-change formula from high-school science class

$$\frac{\text{wages}_1 - \text{wages}_0}{\text{wages}_0} = \exp(\beta_1) - 1$$

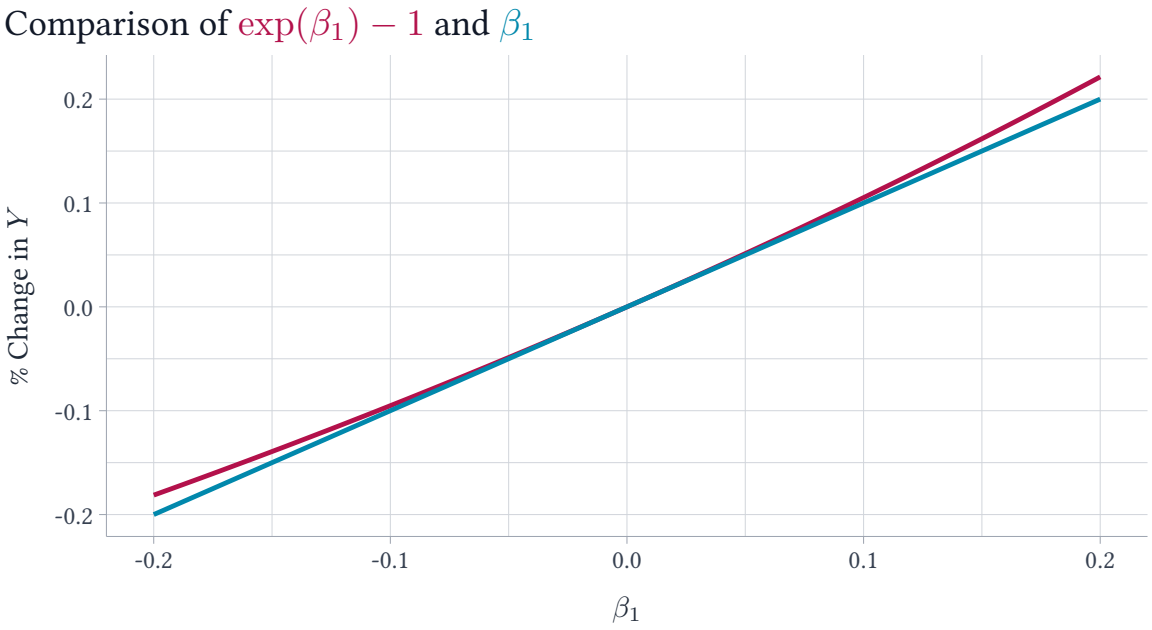
Derivation of $\log$-transformation interpretation

The left-hand side is our percent-change formula from high-school science class

$$\frac{\text{wages}_1 - \text{wages}_0}{\text{wages}_0} = \exp(\beta_1) - 1$$

In this case, the more *precise* answer is that having a college degree is associated with an $(\exp(\beta_1) - 1) * 100\%$ change in wages

$\rightarrow$ But for $-0.10 < \beta_1 < 0.10$, $\exp(\beta_1) - 1$ is approximately equal to β_1 so it's simpler to use the latter


Comparison of $\exp(\beta_1) - 1$ and β_1
0.2
0.1
0.0
-0.1
-0.2
% Change in Y
-0.2
-0.1
0.0
0.1
0.2
β_1

log-log transformations

Alternatively, you may see log transformations of both variables:

$$\log(Y_i) = \beta_0 + \beta_1 \log(X_i) + u_i$$

The interpretation is now simpler: a 1% change in X is associated with a β_1 % change in Y

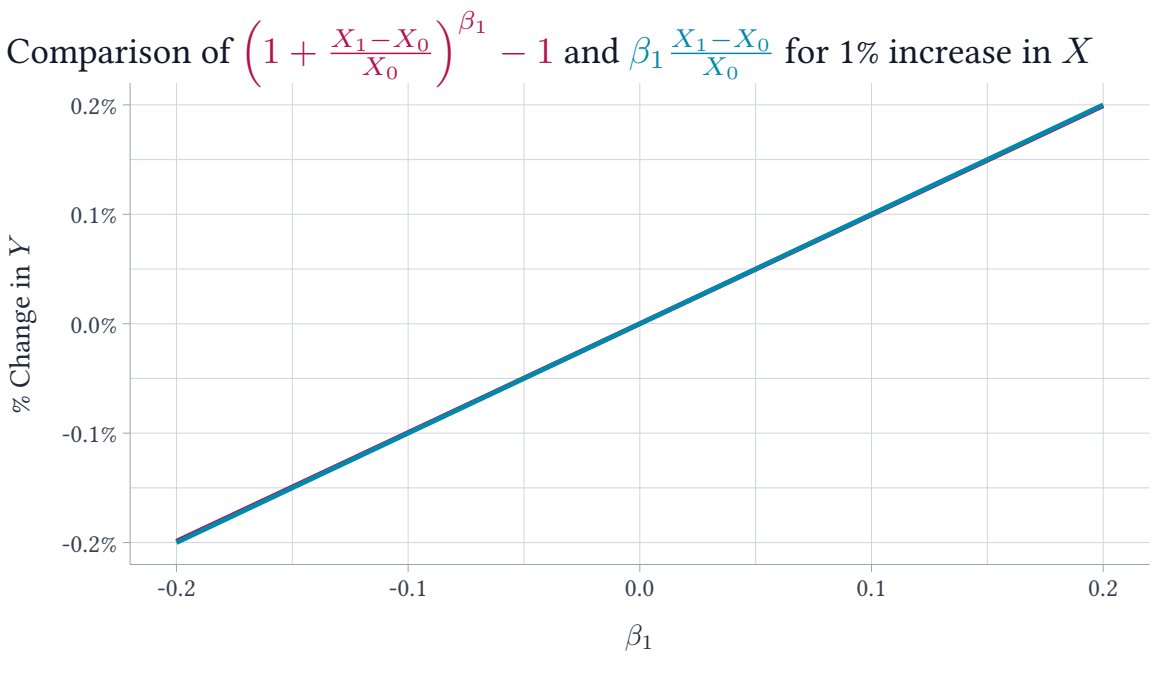
Derivation

$$\log(Y_1) - \log(Y_0) = (\beta_0 + \beta_1 \log(X_1)) - (\beta_0 + \beta_1 \log(X_0))$$

$$\implies \log(1 + \frac{Y_1 - Y_0}{Y_0}) = \beta_1 \log(1 + \frac{X_1 - X_0}{X_0})$$

$$\implies 1 + \frac{Y_1 - Y_0}{Y_0} = \left(1 + \frac{X_1 - X_0}{X_0}\right)^{\beta_1} \approx 1 + \beta_1 \frac{X_1 - X_0}{X_0}$$

$$\implies \frac{Y_1 - Y_0}{Y_0} = \beta_1 \frac{X_1 - X_0}{X_0}$$


Comparison of $\left(1 + \frac{X_1 - X_0}{X_0}\right)^{\beta_1} - 1$ and $\beta_1 \frac{X_1 - X_0}{X_0}$ for 1% increase in X
0.2%
0.1%
0.0%
-0.1%
-0.2%
% Change in Y
-0.2
-0.1
0.0
0.1
0.2
β_1

Near-perfect approximation

Our approximation is basically exact:

$$\frac{Y_1 - Y_0}{Y_0} = \beta_1 \frac{X_1 - X_0}{X_0}$$

So, we interpret a T percent increase in X as a $\beta_1 * T$ percent increase in Y

linear-log transformations

Last, we have the linear-log specification:

$$Y_i = \beta_0 + \beta_1 \log(X_i) + u_i$$

This has the interpretation a 1% change in X is associated with a β_1 unit change in Y

Multivariate Regression – “All Else Equal”

Regressors

- Polynomials
- Indicators
- Multi-valued discrete variables
- Interactions
- Bins
- Splines

log transformations

Binary outcomes

Cluster-robust Standard Errors

Binary Outcomes

In many instances, our outcome of interest is a binary outcome taking the value of 0 or 1

$\rightarrow$ An indicator variable, but people call it a binary variable when it's an outcome

E.g. Y_i denoting whether a customer purchased something from a store and X_i is demographic variables

$\rightarrow$ Want to predict the 'propensity'/'probability' that someone will buy a product

Expectation of a Binary Outcome

The expectation of a binary variable is just

$$\begin{aligned}\mathbb{E}[Y_i] &= 1 * \mathbb{P}(Y_i = 1) + 0 * \mathbb{P}(Y_i = 0) \\ &= \mathbb{P}(Y_i = 1)\end{aligned}$$

This can be estimated by the sample proportion of times $Y_i = 1$:

$$\hat{p} = \frac{1}{n} \sum_i Y_i$$

Conditional Expectation Function of a Binary Outcome

The conditional expectation of a binary outcome can be found similarly:

$$\begin{aligned}\mathbb{E}[Y_i \mid X_i = x] &= 1 * \mathbb{P}(Y_i = 1 \mid X_i = x) + 0 * \mathbb{P}(Y_i = 0 \mid X_i = x) \\ &= \mathbb{P}(Y_i = 1 \mid X_i = x)\end{aligned}$$

We can estimate the CEF of a binary outcome at a point x in the same way:

$\rightarrow$ Subset to people with $X_i = x$

$\rightarrow$ Take the sample proportion of $Y_i = 1$ *within that subsample*. Call this $\hat{f}(x)$

Conditional Expectation Function of a Binary Outcome

Note that this procedure faces the same benefits and costs as with a continuous outcome variable:

$\rightarrow$ The estimator is unbiased and consistent

$\rightarrow$ But the curse of dimensionality makes this estimator noisy when X_i has a large number of variables

Conditional Expectation Function of a Binary Outcome

Note that this procedure faces the same benefits and costs as with a continuous outcome variable:

→ The estimator is unbiased and consistent

→ But the curse of dimensionality makes this estimator noisy when X_i has a large number of variables

This motivates us to study a linear model for the conditional expectation function:

$$\mathbb{E}[Y_i \mid X_i = x] = \mathbb{P}(Y_i = 1 \mid X_i = x) = W_i'\beta$$

for some $W_i = g(X_i)$

Linear Probability Model

This motivates us to study a linear model for the conditional expectation function:

$$\mathbb{E}[Y_i \mid X_i = x] = \mathbb{P}(Y_i = 1 \mid X_i = x) = W_i'\beta$$

for some $W_i = g(X_i)$

This is called a linear probability model and predicts the proportion of 1s you would observe in the population given $X_i = x$

$\rightarrow$ Can be estimated in the exact same way $\hat{\beta}_{\text{OLS}} = (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{Y}$

Marginal Effect Interpretation

Our marginal predictive effects take the same form:

$$\frac{\partial}{\partial x_\ell} \hat{\mathbb{P}}[Y_i = 1 \mid X_i = x] = \sum_{k=1}^{K} \frac{\partial}{\partial x_\ell} g_k(X) \hat{\beta}_{\text{OLS},k}$$

You should reword the interpretation to reflect that you are estimating the conditional probability of $Y_i = 1$:

> "On average, a one unit increase in x_ℓ is associated with a $\hat{\beta}_1$ larger/smaller probability of $Y_i = 1$ *holding all else equal*"

Problems with Linear Probability Model

$$\mathbb{E}[Y_i \mid X_i = x] = \mathbb{P}(Y_i = 1 \mid X_i = x) = W_i'\beta$$

Probabilities must fall between 0 and 1. When we estimate our model, we do not impose that *restriction*

$\rightarrow$ For given values of x, $w'\hat{\beta}_{\text{OLS}}$ can be less than 0 or greater than 1

Though, it is worth saying that this is *not* true when you *only* use indicator variables in W_i

Logistic Regression

Instead, we will use a logistic transformation to model the conditional probability:

$$\mathbb{E}[Y_i \mid X_i = x] = \mathbb{P}(Y_i = 1 \mid X_i = x) = \text{logistic}\left(W_i'\beta\right)$$

"Linear Index"

"transformation"

where $\text{logistic}(W_i'\beta) = \frac{e^{W_i'\beta}}{e^{W_i'\beta}+1}$ and $W_i'\beta$ is called the (linear) 'index function'

Logistic Regression

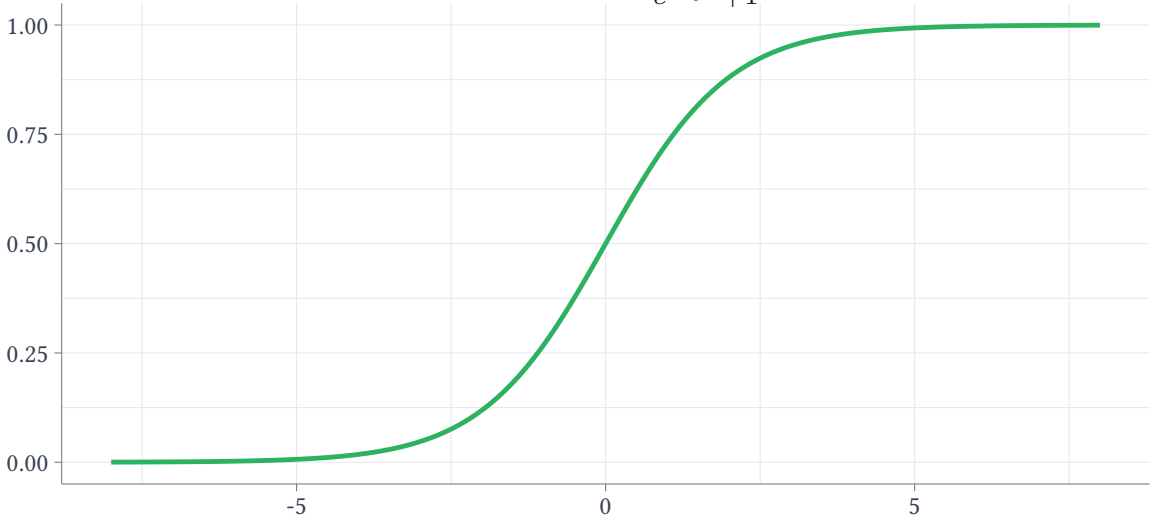
Instead, we will use a logistic transformation to model the conditional probability:

$$\mathbb{E}[Y_i \mid X_i = x] = \mathbb{P}(Y_i = 1 \mid X_i = x) = \underbrace{\text{logistic}}_{\text{“transformation”}} \left(\overbrace{W_i'\beta}^{\text{“Linear Index”}} \right)$$

where $\text{logistic}(W_i'\beta) = \frac{e^{W_i'\beta}}{e^{W_i'\beta}+1}$ and $W_i'\beta$ is called the (linear) ‘index function’

$\rightarrow$ The logistic function ensures that the forecasted probabilities fall within 0 and 1

Logistic transformation: logistic $(W_i'\beta) = \frac{e^{W_i'\beta}}{e^{W_i'\beta}+1}$



Binary Choice Model

One way to arrive at the logistic model is to assume a latent index model where:

$$Y_i^* = W_i'\beta + \varepsilon_i$$

When the 'utility' Y_i^* is high-enough, we switch from 0 to 1. We observe $Y_i \equiv \mathbb{1}[Y_i^* \geq 0]$.

Binary Choice Model $\Longrightarrow$ Logistic CEF

$$Y_i^* = W_i'\beta + \varepsilon_i \quad \text{and} \quad Y_i \equiv \mathbb{1}[Y_i^* \geq 0]$$

When ε_i follows the logistic distribution, this implies

$$\mathbb{P}(Y_i = 1 \mid X_i) = \mathbb{P}\Big(\varepsilon_i \leq -W_i'\beta\Big) = \frac{e^{W_i'\beta}}{e^{W_i'\beta} + 1}$$

Binary Choice Model $\Longrightarrow$ Logistic CEF

$$Y_i^* = W_i'\beta + \varepsilon_i \quad \text{and} \quad Y_i \equiv \mathbb{1}[Y_i^* \geq 0]$$

When ε_i follows the logistic distribution, this implies

$$\mathbb{P}(Y_i = 1 \mid X_i) = \mathbb{P}\Big(\varepsilon_i \leq -W_i'\beta\Big) = \frac{e^{W_i'\beta}}{e^{W_i'\beta} + 1}$$

That is, our latent-index model and a logistic error term implies our logistic transformation

$\rightarrow$ If we assume ε_i is normally distributed, then we end up with a "probit"

“Likelihood”

$$\mathbb{P}(Y_i = 1 \mid X_i) = \mathbb{P}\Big(\varepsilon_i \leq -W_i'\beta\Big) = \frac{e^{W_i'\beta}}{e^{W_i'\beta} + 1}$$

With this, we can think of each Y_i as a Bernoulli random variable with probability of success of $\pi_i \equiv \frac{e^{W_i'\beta}}{e^{W_i'\beta}+1}$

“Likelihood”

$$\mathbb{P}(Y_i = 1 \mid X_i) = \mathbb{P}\Big(\varepsilon_i \leq -W_i'\beta\Big) = \frac{e^{W_i'\beta}}{e^{W_i'\beta} + 1}$$

With this, we can think of each Y_i as a Bernoulli random variable with probability of success of $\pi_i \equiv \frac{e^{W_i'\beta}}{e^{W_i'\beta}+1}$

The likelihood of observing $Y_i = 1$ is π_i and the likelihood of observing $Y_i = 0$ is $1 - \pi_i$

$\rightarrow$ The π_i depends on our choice of β

Maximum-Likelihood Estimation

If our outcomes, Y_i, are independent, then our likelihood of observing the full vector Y given β is

$$\mathcal{L}(\beta \mid Y, \boldsymbol{X}) = \prod_{i=1}^{n} \underbrace{\pi_i^{Y_i}}_{=1 \text{ when } Y_i = 0} \underbrace{(1-\pi_i)^{1-Y_i}}_{=1 \text{ when } Y_i = 1}$$

Recall $\pi_i \equiv e^{W_i'\beta}/(e^{W_i'\beta}+1)$

Maximum-Likelihood Estimation

If our outcomes, Y_i, are independent, then our likelihood of observing the full vector Y given β is

$$\mathcal{L}(\beta \mid Y, \boldsymbol{X}) = \prod_{i=1}^{n} \underbrace{\pi_i^{Y_i}}_{=1 \text{ when } Y_i = 1} \underbrace{(1-\pi_i)^{1-Y_i}}_{=1 \text{ when } Y_i = 0}$$

Recall $\pi_i \equiv e^{W_i'\beta}/(e^{W_i'\beta}+1)$

The goal then is to find the best β to *maximize* the *likelihood* of observing the data Y we did observe

Maximum-Likelihood Estimation In R

Luckily, our `fixest` package also makes performing logistic regression simple:

```
feglm(fmla, data, family = "logit")
```

This performs maximum-likelihood estimation of $\hat{\beta}$

Forecasting probabilities

With our estimated $\hat{\beta}$, we can form forecasts of

$$\hat{\pi}_i \equiv e^{W_i'\hat{\beta}}/(e^{W_i'\hat{\beta}}+1)$$

Note here, we are predicting *the probability* that $Y_i = 1$ given that $X_i = x$.

$\rightarrow$ Often, this is interesting (60% of people buy this product when $X_i = x$)

$\rightarrow$ If we were forced, could chose a rule like $\hat{Y}_i = 1$ if $\hat{\pi}_i \geq 0.5$

Forecasting probabilities

With our estimated $\hat{\beta}$, we can form forecasts of

$$\hat{\pi}_i \equiv e^{W_i'\hat{\beta}}/(e^{W_i'\hat{\beta}}+1)$$

Note here, we are predicting *the probability* that $Y_i = 1$ given that $X_i = x$.

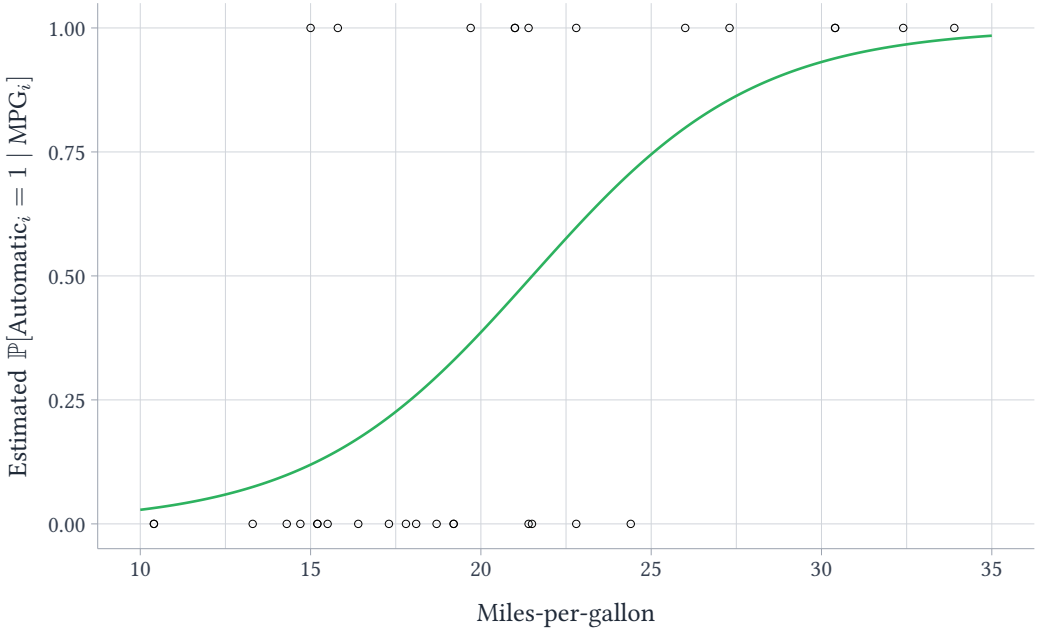
$\rightarrow$ Often, this is interesting (60% of people buy this product when $X_i = x$)

$\rightarrow$ If we were forced, could chose a rule like $\hat{Y}_i = 1$ if $\hat{\pi}_i \geq 0.5$

Forecasting can be done easily as well with our `predict` function and possibly with the `newdata` argument

R Example – Estimation

```
library(fixest)
est <- feglm(am ~ mpg, data = mtcars, family = "logit")
print(est)
#> GLM estimation, family = binomial(link = "logit"), Dep. Var.: am
#> Observations: 32
#> Standard-errors: IID
#>              Estimate Std. Error  z value  Pr(>|z|)
#> (Intercept) -6.603527   2.390311 -2.76262 0.0057339 **
#> mpg          0.307028   0.116740  2.63002 0.0085380 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


1.00
0.75
0.50
0.25
0.00
Estimated $\mathbb{P}[\text{Automatic}_i = 1 \mid \text{MPG}_i]$
10
15
20
25
30
35
Miles-per-gallon

R Example – Forecasting

```
head(predict(est))
#> [1] 0.46109512 0.46109512 0.59789839 0.49171990 0.29690087

predict(est, newdata = data.frame(mpg = c(20, 25, 30)))
#> [1] 0.3862832 0.7450109 0.9313311
```

Classification

$$\hat{\pi}_i \equiv \hat{\mathbb{P}}[Y_i = 1 \mid X_i = x]$$

With our predicted probabilities, we can perform classifcation using some *rule*

→ The obvious rule is forecast $\hat{Y}_i = \mathbb{1}[\hat{\pi}_i \geq 0.5]$

→ Or, could use some other cutoff $\hat{Y}_i = \mathbb{1}[\hat{\pi}_i \geq 0.25]$

Evalutating classification

Table: Confusion Matrix

	Y_i	
$\hat{Y}_i$	0	1
0	135	29
1	31	108

Four possibilities:

1. $Y_i = 0, \hat{Y}_i = 0$: true negative
2. $Y_i = 1, \hat{Y}_i = 0$: false negative
3. $Y_i = 0, \hat{Y}_i = 1$: false positive
4. $Y_i = 1, \hat{Y}_i = 1$: true positive

Usually *rates* are presented:

$\rightarrow$ E.g. false-positive rate is $\frac{\text{\# false positives}}{n}$

Deciding on Classification Rules

$$\hat{Y}_i = \mathbb{1}[\hat{\pi}_i \geq \text{cutoff}]$$

How do you choose a cutoff rule?

$\rightarrow$ Given an unbiased logistic model, a cutoff of 0.5 would maximize the % of correct classifications (in expectation)

Deciding on Classification Rules

$$\hat{Y}_i = \mathbb{1}[\hat{\pi}_i \geq \text{cutoff}]$$

If you are a medical doctor, you might want to have few false-negatives

$\rightarrow$ Would you want a smaller or larger cutoff?

Deciding on Classification Rules

$$\hat{Y}_i = \mathbb{1}[\hat{\pi}_i \geq \text{cutoff}]$$

If you are a medical doctor, you might want to have few false-negatives

$\rightarrow$ Would you want a smaller or larger cutoff?

$\rightarrow$ A smaller cut-off would yield fewer false negatives, at the cost of more false positives

Deciding on Classification Rules

$$\hat{Y}_i = \mathbb{1}[\hat{\pi}_i \geq \text{cutoff}]$$

If you are targeting ads at consumers, it is costly to buy ads, so you want to have few false positives

$\rightarrow$ Would you want a smaller or larger cutoff?

Deciding on Classification Rules

$$\hat{Y}_i = \mathbb{1}[\hat{\pi}_i \geq \text{cutoff}]$$

If you are targeting ads at consumers, it is costly to buy ads, so you want to have few false positives

$\rightarrow$ Would you want a smaller or larger cutoff?

$\rightarrow$ A larger cut-off would yield fewer false positives, at the cost of more false negatives

Marginal Effects Of Logistic Transformation

Remember that our "index" is given by our linear function $U_i = W_i'\beta$.

We know how to calculate marginal effects of U_i with respect to some $x_{i,\ell}$:

$$\frac{\partial}{\partial x_\ell}\hat{U}_i = \sum_k \frac{\partial}{\partial x_\ell} g_k(\ell)\hat{\beta}_\ell$$

Marginal Effects Of Logistic Transformation

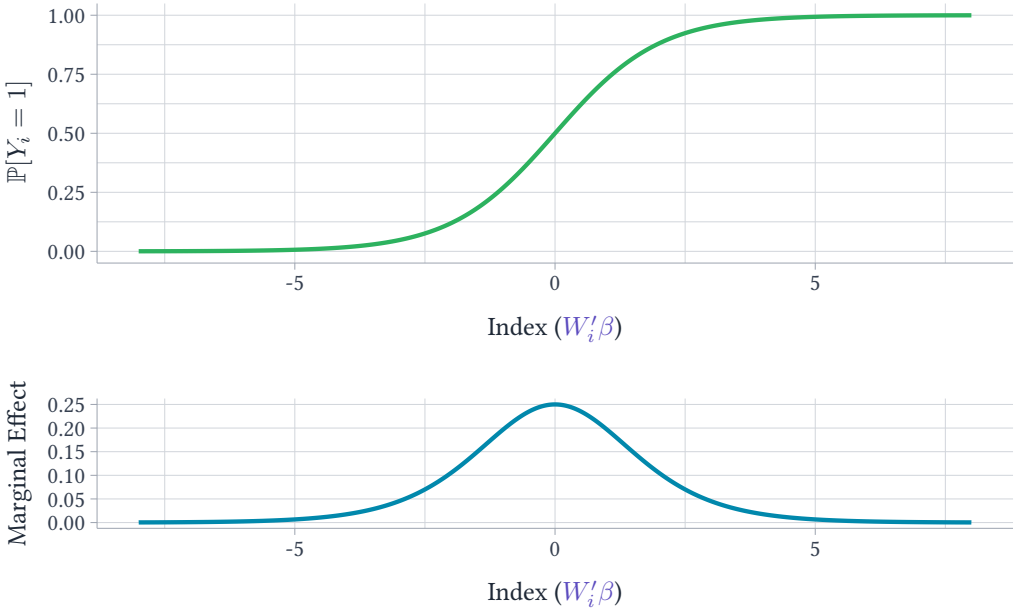
Remember that our "index" is given by our linear function $U_i = W_i'\beta$.

We know how to calculate marginal effects of U_i with respect to some $x_{i,\ell}$:

$$\frac{\partial}{\partial x_\ell}\hat{U}_i = \sum_k \frac{\partial}{\partial x_\ell} g_k(\ell)\hat{\beta}_\ell$$

But, $\hat{\pi}_i = e^{\hat{U}_i}/(e^{\hat{U}_i}+1)$ is a transformation of $\hat{U}_i$

$\rightarrow$ How does $\hat{\pi}_i$ change with $\hat{U}_i$?


1.00
0.75
0.50
0.25
0.00
$\mathbb{P}[Y_i = 1]$
-5
0
5
Index ($W_i'\beta$)
0.25
0.20
0.15
0.10
0.05
0.00
Marginal Effect
-5
0
5
Index ($W_i'\beta$)

Marginal Effects Of Logistic Transformation

$$\hat{\pi}_i = e^{\hat{U}_i}/(e^{\hat{U}_i} + 1)$$

The partial derivative can be estimated via the chain-rule:

$$\frac{\partial}{\partial x_\ell}\hat{\pi}_i = \frac{e^{\hat{U}_i}}{(e^{\hat{U}_i} + 1)^2} * \frac{\partial}{\partial x_\ell}\hat{U}_i$$

$\rightarrow$ The second term is the marginal effect on the *index*, U_i

Marginal Effects

Once we calculate the marginal effect, $\frac{\partial}{\partial x_\ell}\hat{\pi}_i$, we would interpret it as

> "On average, for a unit with $X_i = x$, a one unit larger $x_{i,\ell}$ is associated with an increased probability of $Y_i = 1$ of # *holding all else equal*"

Marginal Effects

Once we calculate the marginal effect, $\frac{\partial}{\partial x_\ell}\hat{\pi}_i$, we would interpret it as

> "On average, for a unit with $X_i = x$, a one unit larger $x_{i,\ell}$ is associated with an increased probability of $Y_i = 1$ of # *holding all else equal*"

Notice now the marginal effect depends on the full set of covariates $X_i = x$ (because they all impact $W_i'\beta$), we need to specify all of this when we talk about marginal effects

Interpreting Marginal Effects

Notice now the marginal effect depends on the full set of covariates $X_i = x$ (because they all impact $W_i'\beta$), we need to specify all of this when we talk about marginal effects

Two options people use:

→ Specify x to be the sample mean of each X_i;

- "At the mean of X, the marginal effect is . . . "

→ Average the marginal effect across the sample or a particular sub-sample

- "On average for men, the marginal effect of increasing $x_{i,\ell}$ is . . . "

marginaleffects Package In R

This would be a bit of a pain to estimate by hand

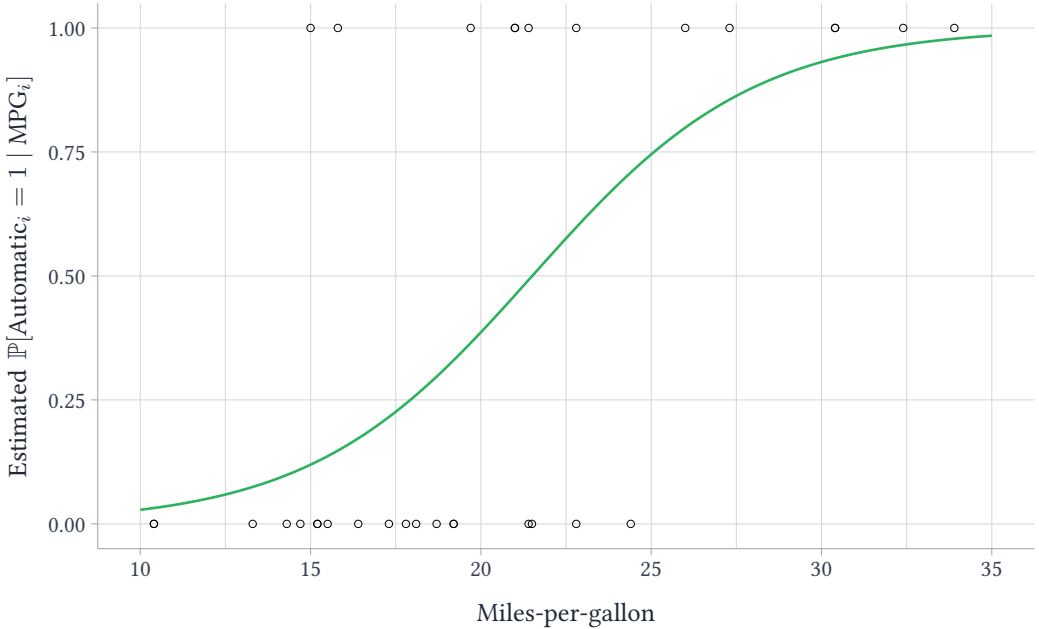
→ And even more annoying would be getting standard errors!

But, the marginaleffects package makes this really easy (and works with fixest!)

→ See https://marginaleffects.com/chapters/slopes.html for details

R Example – Marginal Effects

```
library(marginaleffects)
slopes(
  est, variables = "mpg",
  newdata = data.frame(mpg = c(20, 25, 30))
)
#> Estimate  Std. Error     z  Pr(>|z|)     S     2.5 %  97.5 %
#>   0.0728      0.0285  2.55    0.0108   6.5   0.01684  0.1287
#>   0.0583      0.0144  4.05    <0.001  14.2   0.03007  0.0866
#>   0.0196      0.0144  1.36    0.1738   2.5  -0.00866  0.0479
```


1.00
0.75
0.50
0.25
0.00
Estimated $\mathbb{P}[\text{Automatic}_i = 1 \mid \text{MPG}_i]$
10
15
20
25
30
35
Miles-per-gallon

Multivariate Regression – “All Else Equal”

Regressors

- Polynomials
- Indicators
- Multi-valued discrete variables
- Interactions
- Bins
- Splines

log transformations

Binary outcomes

Cluster-robust Standard Errors

‘Robust’ Inference

In topic 3, we discussed ‘robust’ or ‘HC1’ standard errors for our OLS coefficients.

It depended on independent errors in the regression model: $y_i = W_i'\beta + u_i$

$\rightarrow$ A shock to one unit is unrelated to *all* other units

Independent Errors

For independent errors, we have:

$$\boldsymbol{\Sigma} = \mathbb{E}\left[uu'\right] = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

$\rightarrow$ σ_i^2 can vary across units (heteroskedasticity)

$\rightarrow$ $\sigma_{i,j} = 0$ whenever $i \neq j$ (independence)

Is independence plausible?

In many settings, we do not believe in independent errors because unit-level shocks are correlated, e.g.

→ students in same school,

→ workers in the same labor market,

→ and patients at the same hospital

Is independence plausible?

In many settings, we do not believe in independent errors because unit-level shocks are correlated, e.g.

→ students in same school,

→ workers in the same labor market,

→ and patients at the same hospital

When errors are correlated across units, our 'robust' standard errors are going to typically be *too small*

→ Intuition: correlated errors imply we have fewer 'observations' than n

Within-cluster correlation

The most common way of addressing this is to model the errors as being clustered

→ Units within the same ‘cluster’ are allowed to have arbitrarily correlated errors

→ But, units in different cluster have 0 correlation

Within-cluster correlation

The most common way of addressing this is to model the errors as being clustered

→ Units within the same 'cluster' are allowed to have arbitrarily correlated errors

→ But, units in different cluster have 0 correlation

Let g_i denote the cluster / group that unit i belongs to.

→ $\sigma_{i,j} = 0$ whenever $g_i \neq g_j$ (cluster independence)

→ $\sigma_{i,j}$ is unrestricted when $g_i = g_j$

Clustered Σ

Rearranging rows of our data so that units in the same group are blocked together, we can write out error-term variance covariance matrix as

$$\mathbf{\Sigma} = \mathbb{E}\left[uu'\right] = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{\Sigma}_G \end{bmatrix},$$

where $\mathbf{\Sigma}_g$ is the $n_g \times n_g$ matrix of $\sigma_{i,j}$

Cluster Robust Standard Errors

$\mathbf{\Sigma}_g$ is the $n_g \times n_g$ matrix of $\sigma_{i,j}$

$\rightarrow$ We can estimate each $\sigma_{i,j}$ as the product of regression residuals $\hat{u}_i \hat{u}_j$

For example,

$$\hat{\mathbf{\Sigma}}_1 = \begin{bmatrix} \hat{u}_1^2 & \hat{u}_1\hat{u}_2 & \dots & \hat{u}_1\hat{u}_{n_1} \\ \hat{u}_2\hat{u}_1 & \hat{u}_2^2 & \dots & \hat{u}_2\hat{u}_{n_1} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{u}_{n_1}\hat{u}_1 & \hat{u}_{n_1}\hat{u}_2 & \dots & \hat{u}_{n_1}^2 \end{bmatrix},$$

Cluster Robust Standard Errors

$$\hat{\beta}_{\text{OLS}} \sim \mathcal{N}\left(\beta_0, \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \boldsymbol{W}'\hat{\boldsymbol{\Sigma}}\boldsymbol{W} \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\right)$$

The cluster robust standard errors are formed using the block-diagonal matrix

$$\begin{bmatrix} \hat{\boldsymbol{\Sigma}}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Sigma}}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \hat{\boldsymbol{\Sigma}}_G \end{bmatrix}$$

Fixed effects vs. clustering

The motivation for using clustered standard errors is to allow for common shocks to the units within a cluster (class, labor market, etc.)

We could include a set of indicator variables for cluster membership

$$y_i = \alpha_{g_i} + \beta x_i + \epsilon_i$$

$\rightarrow$ α_g is the common effect impacting all units in the same group ('fixed effect')

So why would we cluster our standard errors?

Fixed effects vs. clustering

Fixed effects are a very specific type of effect

→ *All units* in the group are effected and *by the same amount*

Fixed effects vs. clustering

Fixed effects are a very specific type of effect

→ *All units* in the group are effected and *by the same amount*

But, there can be more general forms of correlated errors in a group

→ Might have within-cluster shocks (e.g. clustered at school level, common shocks by classroom)

→ Some people are more impacted by shocks than other

Clustered standard errors can still be necessary *even if* using cluster fixed-effects ‼

IID Standard Errors

```
feols(mpg ~ i(am), data = mtcars)
#> Observations: 32
#> Standard-errors: IID
#>             Estimate Std. Error  t value   Pr(>|t|)
#> (Intercept) 17.14737    1.12460 15.24749 1.1340e-15 ***
#> am::1        7.24494    1.76442  4.10613 2.8502e-04 ***
```

HC1 Standard Errors

```
feols(mpg ~ i(am), data = mtcars, vcov = "HC1")
#> Observations: 32
#> Standard-errors: Heteroskedasticity-robust
#>             Estimate Std. Error  t value   Pr(>|t|)
#> (Intercept) 17.14737   0.884189 19.39332  < 2.2e-16 ***
#> am::1        7.24494   1.913596  3.78603 0.00068505 ***
```

Clustered Standard Errors

```
feols(mpg ~ i(am), data = mtcars, cluster = ~cyl)
#> Observations: 32
#> Standard-errors: Clustered (cyl)
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 17.14737    2.06574 8.30085 0.014204 *
#> am::1        7.24494    2.34091 3.09492 0.090458 .
```

Correct "n"

The intuition for why standard errors are typically larger for cluster-robust standard errors is that we have fewer "distinct" observations when error terms can be correlated within group

That raises the question, how many 'n' do we have?

$\rightarrow$ In the worst-case scenario (extreme correlation), you only have G observations (one for each group)

$\rightarrow$ In the best-case scenario (independence within cluster), you have n observations

Correct "n"

The rule of thumb is you want at least 20-30 groups, each with multiple observations

$\rightarrow$ If you have a smaller number of groups, then you want to use 'wild cluster bootstrap'

See `fwildclusterboot` package when you have a few groups