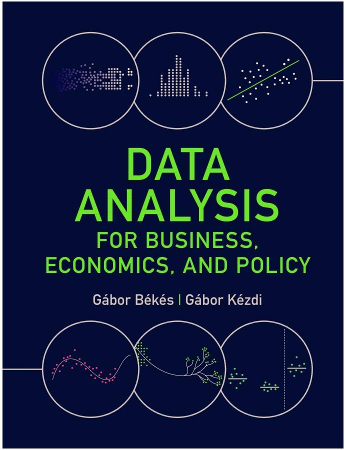
# 10. Multiple regression

**Gábor Békés**

Data Analysis 2: Regression analysis

2020

# Slideshow for the Békés-Kézdi Data Analysis textbook




- Cambridge University Press, 2021
- **gabors-data-analysis.com**
  - Download all data and code `gabors-data-analysis.com/data-and-code/`
- This slideshow is for **Chapter 10**

# Motivation

- You want to find out how running time, distance and altitude are associated with each other to evaluate your local running time.
- Interested in finding evidence for or against labor market discrimination of women. Compare wages for men and women who share similarities in wage relevant factors such as experience.

# Topics for today:

## Topics for today

# Multiple regression analysis

- Multiple regression analysis uncovers average $y$ as a function of more than one $x$ variable: $y^E = f(x_1, x_2, ...)$.
- It can lead to better predictions $\hat{y}$ by considering more explanatory variables.
- It may improve the interpretation of slope coefficients by comparing observations that are different in terms of one of the $x_i$ variable but similar in terms of other $x_{-i}$ variables ($-i$ means all other variable except $i$).
- Multiple linear regression specifies a linear function of the explanatory variables for the average $y$.

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_k x_k$$

# Multiple regression - case of two regressors

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- $\beta_1$: the slope coefficient on $x_1$ shows difference in average $y$ across observations with unit difference in $x_1$, *but the same value of* $x_2$.
  - $\beta_2$ shows difference in average $y$ across observations with with unit difference in $x_2$, *but the same value of* $x_1$.
- Can compare observations that are similar in one explanatory variable to see the differences related to the other explanatory variable.

# Multiple regression - visual representation

With two explanatory variables visually it means to fit linear plane:

- We are still minimizing the sum of squared errors:

$$\arg \min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{N} (y - \beta_0 - \beta_1 x_1 - \beta_2 x_2)^2$$

- For $K$ variables you fit a $K$ dimensional linear plane!
- It is tricky how to visualize multiple regression...
- We cover some of those possibilities.

# Multiple regression vs single regression

Compare slope coefficient in simple ($\beta$) and in multiple ($\beta_1$) linear regression:

$$\text{Simple: } y^E = \alpha + \beta x_1$$

$$\text{Multiple: } y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

To connect $\beta$ and $\beta_1$ you need to regress $x_2$ on $x_1$ (called: "$x - x$ regression"):

$$x_2^E = \gamma + \delta x_1$$

# Multiple regression vs single regression

Compare slope coefficient in simple ($\beta$) and in multiple ($\beta_1$) linear regression:

$$\text{Simple: } y^E = \alpha + \beta x_1$$

$$\text{Multiple: } y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

To connect $\beta$ and $\beta_1$ you need to regress $x_2$ on $x_1$ (called: "$x - x$ regression"):

$$x_2^E = \gamma + \delta x_1$$

Plug this into the multiple regression:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2(\gamma + \delta x_1) = \beta_0 + \beta_2\gamma + (\beta_1 + \beta_2\delta) x_1 \,.$$

It turns out:

$$\beta - \beta_1 = \delta\beta_2$$

# Difference in slopes - in words...

- The slope of $x_1$ in a simple regression is different from its slope in the multiple regression, the difference being the product of its slope in the regression of $x_2$ on $x_1$ and the slope of $x_2$ in the multiple regression.
- The slope coefficient on $x_1$ in the two regressions is different
  - unless $x_1$ and $x_2$ are uncorrelated ($\delta = 0$) OR
  - the coefficient on $x_2$ is zero in the multiple regression ($\beta_2 = 0$).
- The slope in the simple regression is larger if $x_2$ and $x_1$ are positively correlated and $\beta_2$ is positive
  - or $x_2$ and $x_1$ are negatively correlated and $\beta_2$ is negative

# Multiple regression - why different?

- If $x_1$ and $x_2$ are correlated, comparing observations with or without the same $x_2$ value makes a difference.
- If they are positively correlated, observations with higher $x_2$ tend to have higher $x_1$.
- In the simple regression we ignore differences in $x_2$ and compare observations with different values of $x_1$.
- But higher $x_1$ values mean higher $x_2$ values, too.
- Corresponding differences in $y$ may be due to differences in $x_1$ but also differences in $x_2$.
  - Neglecting $x_2$, when it is important leads to 'omitted variable bias'.

# Multiple regression - omitted variable

- Omitted variables are important, if you are interested in a coefficient value:
  - If you have a measure/variable on $x_2$ use it and you are done.
  - If you do not have a measure/variable on $x_2$:
    - similar to measurement errors: think and argue!
    - Is your 'true' parameter smaller or larger than what you estimated?
- Language: The slope on $x_1$ in the sample is confounded by omitting the $x_2$ variable, and thus $x_2$ is a **confounder**.
  - When you see/report coefficient values with adding more and more other variables to the model:
    - Want to show parameter stability - there is no other important confounder.
    - If your coefficient value changes by adding other variable(s), then you most likely have omitted variable bias problem.

# Multiple regression - some language

- Multiple regression with two explanatory variables ($x_1$ and $x_2$),
- We measure differences in expected $y$ across observations that differ in $x_1$ but are similar in terms of $x_2$.
- Difference in $y$ by $x_1$, *conditional on* $x_2$. OR *controlling for* $x_2$.
- We condition on $x_2$, or control for $x_2$, when we include it in a multiple regression that focuses on average differences in $y$ by $x_1$.

# OLS estimator - to see such formulation

For multiple regression usually we use matrix notation:

$$y = \boldsymbol{x}'\boldsymbol{\beta}$$

where, $\boldsymbol{x} = [1, x_1, x_2, \ldots, x_k]$ and $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \ldots, \beta_k]'$.
OLS has a closed form solution in matrix form:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{x}'\boldsymbol{x}\right)^{-1} \boldsymbol{x}'y$$

# Standard Error of Beta

- Inference, confidence intervals in multiple regressions is analogous to those in simple regressions.

$$SE(\hat{\beta}_1) = \frac{Std[e]}{\sqrt{n}Std(x_1)\sqrt{1 - R_1^2}}$$

- Behaviour is the same, the SE is small IF: small Std of the residuals (the better the fit of the regression); large sample, large the Std of $x_1$.
- New element: $\sqrt{1 - R_1^2}$ term in the denominator - the R-squared of the regression of $x_1$ on $x_2$ - refers to the correlation between $x_1$ and $x_2$.
- The stronger the correlation between $x_1$ and $x_2$ the larger the SE of $\hat{\beta}_1$.
- Note the symmetry: the same applies to the SE of $\hat{\beta}_2$.
- As usual, in practice, use robust SE.

# Collinearity of explanatory variables

- **Perfectly collinearity** is when $x_1$ is a linear function of $x_2$.
- Consequence: cannot calculate coefficients (reason: linearly dependent matrix: inverse does not exists...)
  - One will be dropped by software
- Strong but imperfect correlation between explanatory is called *multicollinearity*.
  - Consequence: We can still get the slope coefficients and their standard errors, but:
    - Standard errors may be large.
    - Does not affect the value of $\beta$

# Multicollinearity and SE of beta

- As a consequence of multicollinearity the standard errors may be large.
  - Concept: Few variables that are different in $x_1$ but not in $x_2$. Not enough observations for comparing average $y$ when $x_1$ is different but $x_2$ remains the same.
  - Math: $R_1^2$ is high ($x_2$ is a good predictor of $x_1$), thus $\sqrt{1 - R_1^2}$ is (really) small, which makes $SE(\beta_1)$ (very) large.
- This is a small sample problem.
  - May look at pair-wise correlations when start working with data
  - Drop one or the other, or combine them (use z-score/average/PCA).

# F-test: joint significance

- *Testing joint hypotheses*: null hypotheses that contain statements about more than one regression coefficient.
- We aim at testing whether a subset of the coefficients (such as all geographical variables) are all zero.
- F-test answers this.
  - Individually they are not all statistically different from zero, but together they may be.
  - Everything is similar to t-tests, but the sampling distribution here is a 'F-distribution'

- We may ask if *all slope coefficients are zero* in the regression.
  - "Global F-test", and its results are often shown by statistical software by default.

# Many explanatory variables

- Having more explanatory variables is straightforward extension:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots$$

- Interpreting the slope of $x_1$: on average, $y$ is $\beta_1$ units larger in the data for observations with one unit larger $x_1$ but the same value for all other $x$ variables.
- SE formula - small when $R_k^2$ is small - $R^2$ of regression of $x_k$ on all *other* $x$ variables.

$$SE(\hat{\beta}_k) = \frac{Std[e]}{\sqrt{n}Std[x_k]\sqrt{1 - R_k^2}}$$

# Non-linear patterns with multiple regression

- Uses splines, polynomials - actually like multiple regression - we have multiple coefficient estimates.
- Multicollinearity - not (perfect) *linear* combinations, but keep in mind...
  - Remember the 'poly()' function? $\rightarrow$ it handles this issue!
- Non-linear function of various $x_i$ variables may be combined.

# Understanding the gender difference in earnings

- In the USA (2014), women tend to earn about 20% less than men
- Aim 1: Find patterns to better understand the gender gap.
  - Our focus is the interaction with age.
- Later - Aim 2: Think about if there is a causal link from being female to getting paid less.

# Gender gap in earnings - data

- 2014 census data
  - Age between 15 to 65
  - Exclude self-employed (earnings is difficult to measure)
  - Include those who reported 20 hours more as their usual weekly time worked
- Employees with a graduate degree (higher than 4-year college)
- Use log hourly earnings ($\ln(w)$) as dependent variable
- Use gender and add age as explanatory variables

# Basic models for gender gap

We are quite familiar with the relation between earnings and gender:

$$\ln w^E = \alpha + \beta \mathit{female}, \qquad \beta < 0$$

Let's extend the model with age:

$$\ln w^E = \beta_0 + \beta_1 \mathit{female} + \beta_2 \mathit{age}$$

We can calculate the correlation between female and age, which is in fact negative.

What do you expect about $\beta, \beta_1, \delta$?
Reminder:

$$\mathit{age}^E = \gamma + \delta \mathit{female}$$

# Gender gap regression - baseline

| VARIABLES | (1) lnw | (2) lnw | (3) age |
|---|---|---|---|
| female | -0.195** | -0.185** | -1.484** |
| | (0.008) | (0.008) | (0.159) |
| age | | 0.007** | |
| | | (0.000) | |
| Constant | 3.514** | 3.198** | 44.630** |
| | (0.006) | (0.018) | (0.116) |
| Observations | 18,241 | 18,241 | 18,241 |
| R-squared | 0.028 | 0.046 | 0.005 |

Note: *All employees with a graduate degree. Robust standard errors in parentheses*
Source: `cps-earnings` dataset. 2014 CPS Morg.

# Age is a confounder variable

Remember: the omitted variable bias is given by:

$$\beta - \beta_1 = \delta\beta_2$$

which can be calculated easily:

- $\beta - \beta_1 = -0.195 - (-0.185) = -0.01$
- $\delta\beta_2 = -1.48 \times 0.007 \approx -0.01$

Interpretation:

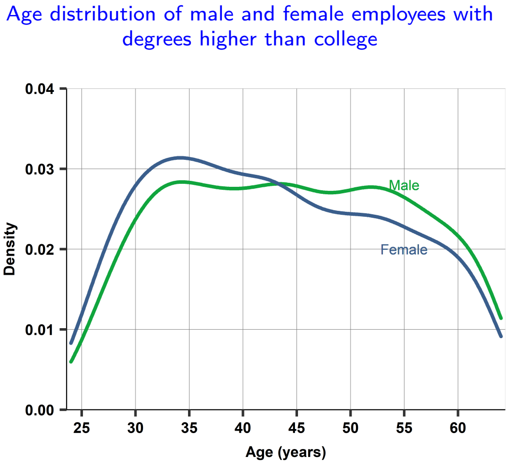
- Age is a confounder, it is different from zero and the value of beta coefficient changes.
- But a weak one: the magnitude of the change is not really large.

# Interpretations and connections of the basic model

Interpretation of model coefficients:

- Women of the same age have a slightly smaller earnings disadvantage in this data because they are somewhat younger, on average
- employees that are younger tend to earn less
- part of the earnings disadvantage of women is thus due to the fact that they are younger.
  - This is a small part: around 1 percentage points of the 20% difference,
  - Overall this is only a 5% share of the entire difference.
    - This is the difference if we control for age or not.
- A single linear variable for age may not be enough.
  - Investigate the impact of age.

# Conditional distribution of age based on gender




- Relatively few below age 30
- Above 30
  - close to uniform for men
  - for women, the proportion of female employees with graduate degrees drops above age 45, and again, above age 55
- Two possible things
  - fewer women with graduate degrees among the 45+ old than among the younger ones
  - fewer of them are employed

# Non-linearity in age, but same effect on gender

| VARIABLES | (1) lnw | (2) lnw | (3) lnw | (4) lnw |
|---|---|---|---|---|
| female | -0.195** | -0.185** | -0.183** | -0.183** |
| | (0.008) | (0.008) | (0.008) | (0.008) |
| age | | 0.007** | 0.063** | 0.572** |
| | | (0.000) | (0.003) | (0.116) |
| agesq | | | -0.001** | -0.017** |
| | | | (0.000) | (0.004) |
| agecu | | | | 0.000** |
| | | | | (0.000) |
| agequ | | | | -0.000** |
| | | | | (0.000) |
| Constant | 3.514** | 3.198** | 2.027** | -3.606** |
| | (0.006) | (0.018) | (0.073) | (1.178) |

# Using qualitative variables

- Can have binary variables as well as other qualitative variables (factors) .
- Consider a qualitative variable like income categories or continents. How to add it to the regression model?
  - Create binary variables (dummy variables) for all options. Add them - all but one. (Why? $\rightarrow$ linear dependence with the intercept!)
  - Left out one will be the base/reference!

# Qualitative variables - example I.

- $x$ is a categorical variable with three values *low*, *medium* and *high*
- binary variable $x_m$ denote if $x = medium$, $x_h$ variable denote if $x = high$.
- for $x = low$ is not included. It is called the *reference category* or left-out category.

$$y^E = \beta_0 + \beta_1 x_m + \beta_2 x_h$$

# Qualitative variables - example II.

$$y^E = \beta_0 + \beta_1 x_m + \beta_2 x_h$$

- Pick $x = low$ as the reference category. Other values compared to this.
  - This is the left out variable
- $\beta_0$ shows average $y$ in the reference category. Here, $\beta_0$ is average $y$ when both $x_m = 0$ and $x_h = 0$: this is the case of $x = low$.
- $\beta_1$ shows the difference of average $y$ between observations with $x = medium$ and $x = low$
- $\beta_2$ shows the difference of average $y$ between observations with $x = high$ and $x = low$.

# Interactions

- Many cases, data is made up of important groups: male and female workers or countries in different continents.
- Some of the patterns we are after may vary across these groups.
- The strength of a relation may also be altered by a special variable.
- In medicine, a *moderator variable* can reduce / amplify the effect of a drug on people.
- In business, financial strength can affect how firms/countries may weather a recession.
- All of these mean different patterns for subsets of observations.

# Interactions - when to use?

- Regression with two explanatory variables: $x_1$ is continuous, $D$ is binary denoting two groups in the data (e.g., male or female employees).
- We wonder if the relationship between average $y$ and $x_1$ is different for observations with $D = 1$ than for $D = 0$. How to test?

# Interaction - parallel lines

- Option 1: Two *parallel lines* for the $y$ - $x_1$ pattern: one for those with $D = 0$ and one for those with $D = 1$.
- Similar to qualitative variables plus a continuous variable $x_1$

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 D$$

- The predicted/expected values for the two groups ($y_0^E = E[y^E|D = 0], y_1^E = E[y^E|D = 1]$) can be written as,

$$y_0^E = \beta_0 + \beta_2 \times 0 + \beta_1 x_1$$

$$y_1^E = \beta_0 + \beta_2 \times 1 + \beta_1 x_1$$

# Interaction - different slopes

- Option 2: *Allow for different slopes* in the two $D$ groups we have to add an interaction term directly to $x_1$ as well:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 D + \beta_3 (x_1 \times D)$$

- Intercepts are kept different by $\beta_2$ AND slopes different by $\beta_3$. The two slopes are given by,

$$y_0^E = \beta_0 + \beta_1 x_1$$

$$y_1^E = \beta_0 + \beta_2 + (\beta_1 + \beta_3) x_1$$

# Interactions vs separate regressions

- Separate regressions in the two groups and the regression that pools observations but includes an interaction term, yield *exactly the same* coefficient estimates.
  - The coefficients of the separate regressions are easier to interpret.
  - The pooled regression with interaction allows for a direct test of whether the slopes are the same.

# Interaction with many groups

- You can generalize to three groups
  - Let: $D_1$, $D_2$ are binaries and $x$ is continuous:

$$y^E = \beta_0 + \beta_1 x + \beta_2 D_1 + \beta_3 D_2 + \beta_4 (D_1 \times x) + \beta_5 (D_2 \times x)$$

- In general, if you have $K$ groups

$$y^E = \beta_0 + \beta_1 x + \sum_{k=2}^{K} \beta_k D_{k-1} + \beta_{K+k} (D_{k-1} \times x)$$

# Interaction with two continuous variable

- Same model used for two continuous variables, $x_1$ and $x_2$:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- Example: Firm level data, 100 industries.
  - $y$ is change in revenue, $x_1$ is change in global demand, $x_2$ is firm's financial health
  - The interaction can capture that drop in demand can cause financial problems in firms, but less so for firms with better balance sheet.
- Note: interpretation is tricky! Use the derivative to see why!

# Interaction between gender and age

- Why we assume that age has the same slope regardless of gender? We might want to check, whether they are different!
- Are the slopes significantly different?
- Can one get the slope for age for female only from the regression with the interaction?
- How the gender dummy's coefficient changed?

# Interaction between gender and age

- Earning for men rises faster with age.
- Pooled EQ with interaction: interaction + age coefficient is the SAME as women's age coefficient.
- $\beta_3$ is significant: earning growth by age is different for male and female.
- Constant dummy is close to zero and seems insignificant
  - at birth there would be no difference,
  - but at 25, there is already a significant difference → interaction term

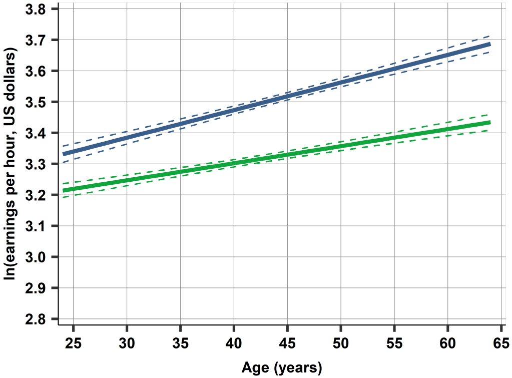
| VARIABLES | (1) WOMEN lnw | (2) MEN lnw | (3) ALL lnw |
|---|---|---|---|
| female | | | -0.036 |
| | | | (0.035) |
| age | 0.006** | 0.009** | 0.009** |
| | (0.001) | (0.001) | (0.001) |
| female X age | | | -0.003** |
| | | | (0.001) |
| Constant | 3.081** | 3.117** | 3.117** |
| | (0.023) | (0.026) | (0.026) |
| Observations | 9,685 | 8,556 | 18,241 |
| R-squared | 0.011 | 0.028 | 0.047 |

# Nonlinearities and interactions

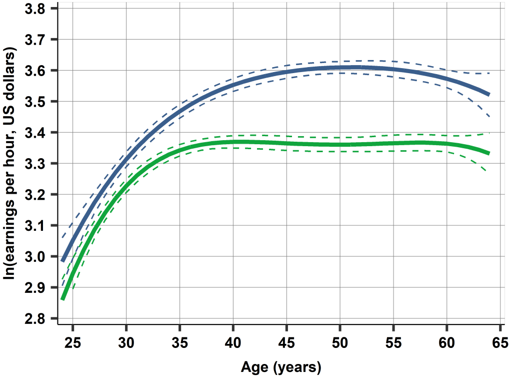
We can estimate interactions with non-linear terms as well:

$$
\begin{aligned}
lnw^E = \beta_0 &+ \beta_1 age + \beta_2 age^2 + \beta_3 age^3 + \beta_4 age^4 \\
&+ \beta_5 female + \beta_6 female \times age + \beta_7 female \times age^2 \\
&+ \beta_8 female \times age^3 + \beta_9 female \times age^4
\end{aligned}
$$

# Nonlinearities and interactions




*Log earnings per hour and age by gender: predicted values and confidence intervals from a linear regression interacted with gender.*




*Log earnings per hour and age by gender: predicted values and confidence intervals from a regression with 4th-order polynomial interacted with gender.*

# Visual inspection in the regression lines

- The average earnings difference is around 10% between ages 25 and 30
- increases to around 15% by age 40, and reaches 22% by age 50,
- from where it decreases slightly to age 60 and more by age 65.
- confidence intervals around the regression curves are rather narrow, except at the two ends.
- Conclusion?

# Causal analysis with multiple regression

- One main reason to estimate multiple regressions is to get closer to a causal interpretation.
- Called: Causal analysis or causal inference
- By conditioning on other observable variables, we can get closer to comparing similar objects – "apples to apples" – even in observational data.
- But getting closer is not the same as getting there.
- In principle, one may help that by conditioning on *every* potential confounder: variables that would affect $y$ and the causal variable $x_1$ at the same time.
  - Ceteris paribus = conditioning on *every* such relevant variable.

# Causal analysis - ceteris paribus

- Ceteris paribus = conditioning on **every** such relevant variable.
- *Ceteris paribus* prescribes what we want to condition on.
  - A multiple regression can condition on **what's in the data** the way it is measured.
- Importantly, conditioning on everything is impossible in general.
- Multiple regression is never (hardly ever) ceteris paribus.

# Causal analysis

- A multiple regression on observational data is rarely capable of uncovering a causal relationship.
  - Cannot capture all potential confounder. (No ceteris paribus comparison)
  - We can never really know. BUT
- multiple regression can get us **closer** to uncovering a causal relationship
  - Compare units that are the same in many respects - controls
- More on causal inference in Chapters 19-24

# Gender difference in earnings - causality?

What may cause the difference in wages?

- Labor discrimination - one group earns less even if they have the same *marginal product*
- Try control for marginal product (or for variables which matters to marginal product)
  - Eg.: occupation (as an indicator for inequality in gender roles), or industry, union status, hours worked and other socio-economic characteristics
- Use variables as controls - does comparing apples to apple change coefficient of female variable?
  - Practice: add more variables if coefficient is the same you are good. Otherwise need to think about OVB...

# Causal analysis - results

- More and more confounders added
- Female coefficient reduced from 22% to 14%
- Compare two people, with same age, hours, industry, occupation, geography, background (=confounders) - women earn 14% less, on average.

| VARIABLES | (1) ln wage | (2) ln wage | (3) ln wage | (4) ln wage |
|---|---|---|---|---|
| Female | -0.224** | -0.212** | -0.151** | -0.141** |
| | (0.012) | (0.012) | (0.012) | (0.012) |
| Age and education | | YES | YES | YES |
| Family background | | | YES | YES |
| Hours worked | | | YES | YES |
| Government or private | | | YES | YES |
| Union member | | | YES | YES |
| Not born in USA | | | | YES |
| Age in polynomial | | | | YES |
| Hours in polynomial | | | | YES |
| Observations | 9,816 | 9,816 | 9,816 | 9,816 |
| R-squared | 0.036 | 0.043 | 0.182 | 0.195 |

# Discussion

- Could not safely pin down the role of labor market discrimination and broader gender inequality

# Prediction with multiple regression

- Reason to estimate a multiple regression is to make a *prediction*.
  - find the best guess for the dependent variable $y_j$ for a particular *target observation* $j$

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} + \ldots$$

- When the goal is prediction we want the regression to produce as good a fit as possible.
  - 'good fit' in the general pattern that is representative of the target observation $j$.
- A common danger is *overfitting* the data: finding patterns in the data that are not true in the general pattern, only for your sample.
- More on prediction in Chapters 13-18

# Visualization of fit for multiple regression

- The $\hat{y} - y$ *plot* has $\hat{y}$ on the horizontal axis and $y$ on the vertical axis.
  - The plot features the 45 degree line and the scatterplot around it = the regression line of $y$ regressed on $\hat{y}$.
- The scatterplot around this line shows how actual values of $y$ differ from their predicted value $\hat{y}$.
- Review case study in Chapter 10

# Summary take-away

- Multiple regression are linear models with several x variables.
- May include binary variables and interactions
- Multiple regression can take us closer to a causal interpretation and help make better predictions.