Ordinary least squares regression II: The univariate affine regression.

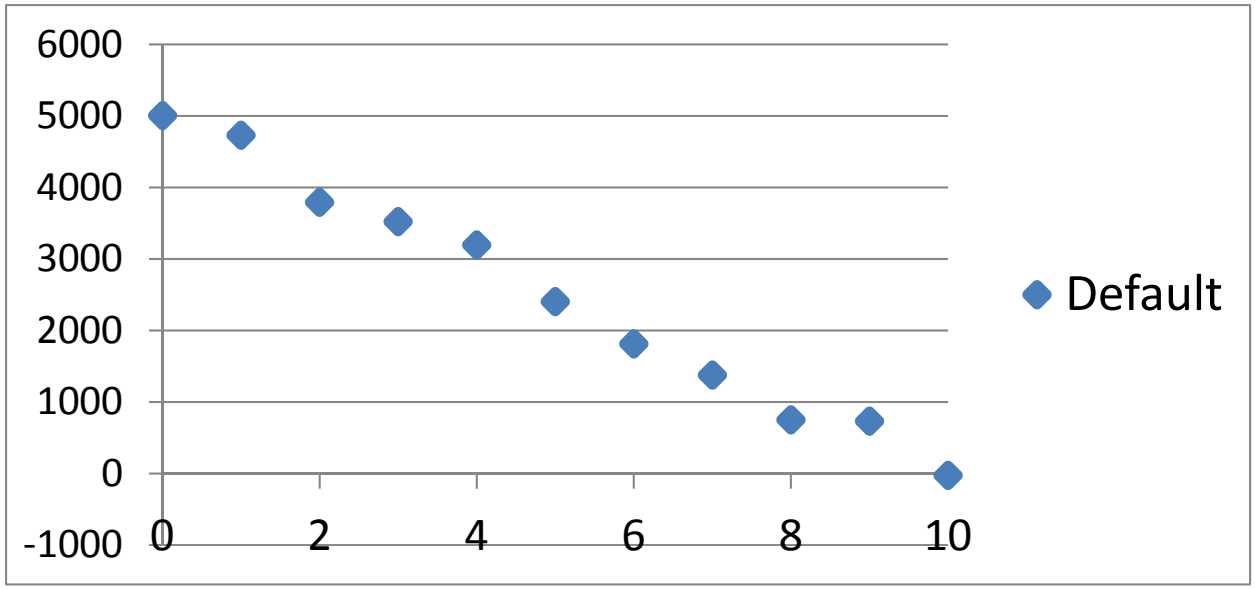
Clement de Chaisemartin, UCSB

Many people need to make predictions

- Traders: use today's GDP growth to predict tomorrow oil's price.
- Banks: use FICO score to predict the amount that a April 2018 applicant will fail to reimburse on her one-year loan in April 2018.
- Gmail: use whether incoming email has the word "free" in it to predict whether it's a spam.

The relationship between FICO score and default

- Assume that relationship between FICO score and amount people fail to repay looks like graph below: people with low FICO fail to repay more.
- If you use a univariate linear regression to predict the amount people fail to repay based on their FICO score, will you make good predictions? Discuss this question with your neighbor.



iClicker time

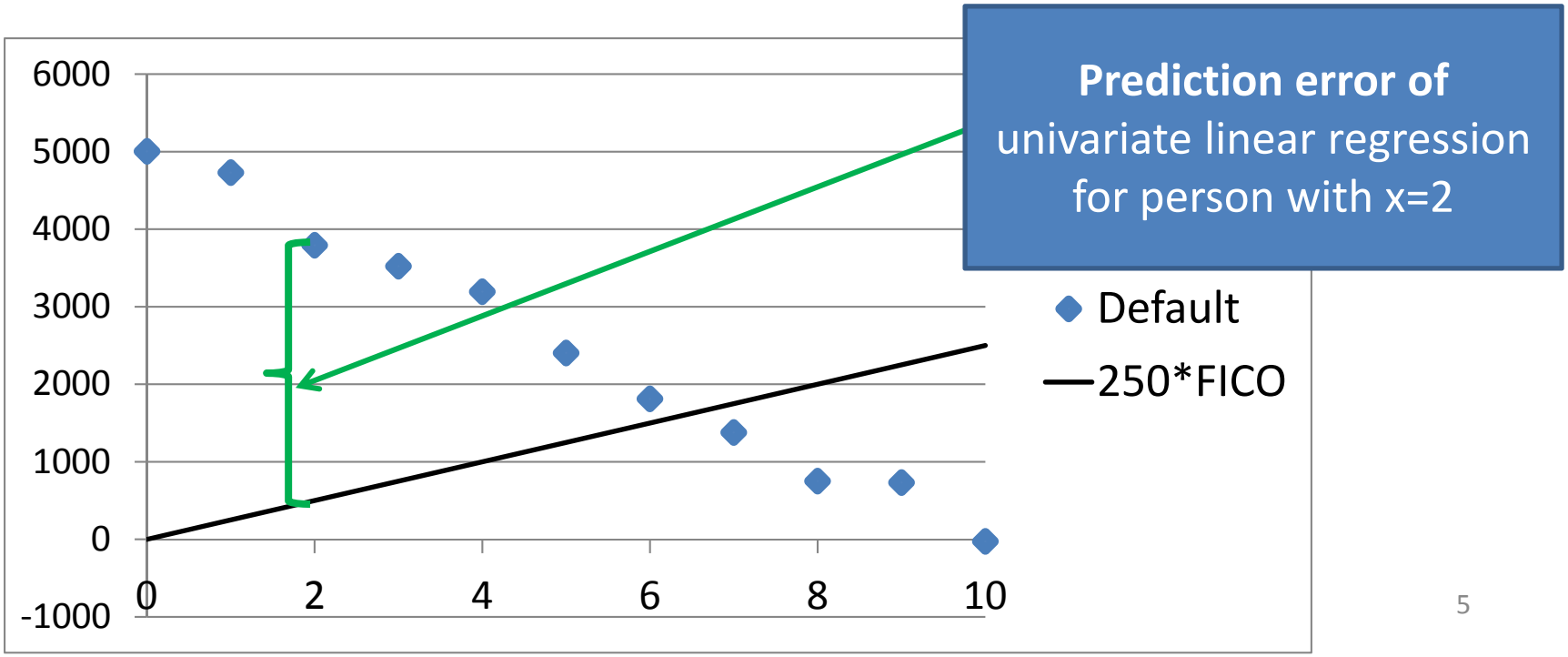
- If you use a univariate linear regression to predict the amount people fail to repay based on their FICO score, will you make good predictions?

a) Yes

b) No

No!

- In this example, OLS regression function is 250*FICO, increasing with FICO! We predict that people with better scores will fail to reimburse more.
- OLS regression makes large prediction errors.
- Why does regression make large prediction errors?



iClicker time

- Why does the univariate linear regression make large prediction errors?

a) Because the relationship between FICO and the amount people fail to repay is decreasing.

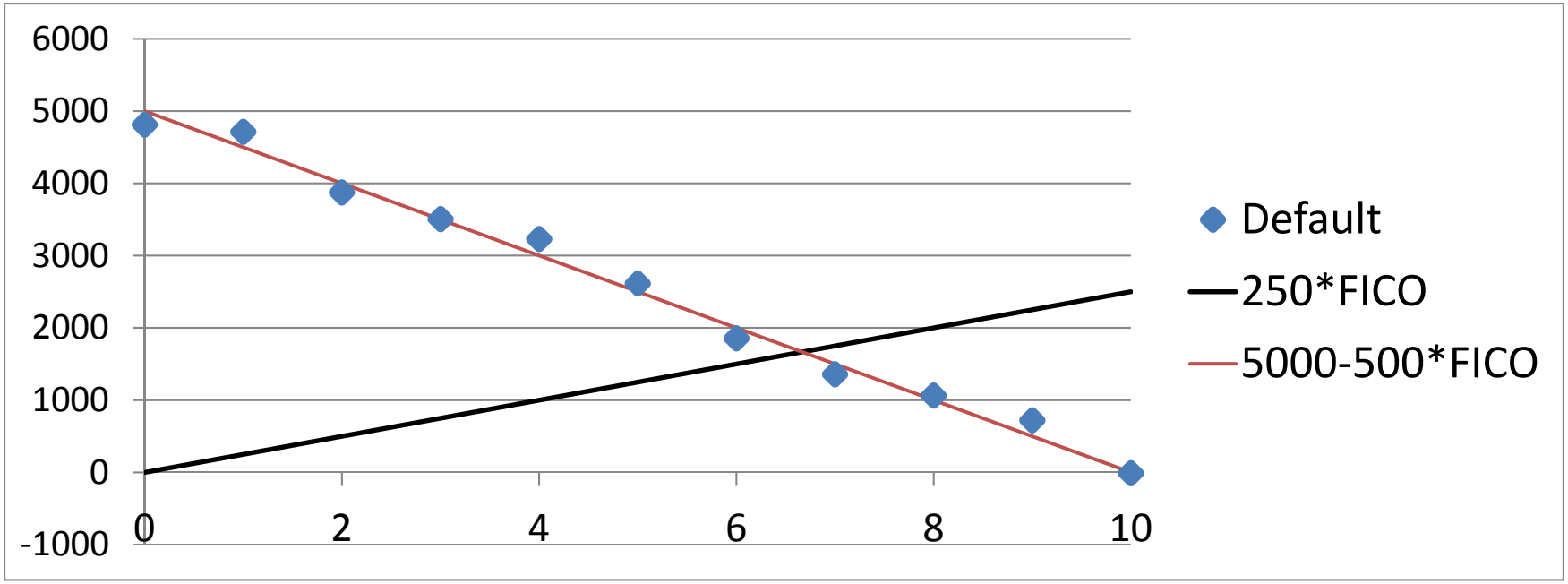
b) Because the amount that people with FICO score equal to 0 fail to repay is different from 0.

Because the amount that people with a FICO score equal to 0 fail to repay is different from 0.

- The univariate linear regression function is αx_k. Therefore, by construction, our prediction will be 0 for people with FICO score = 0.
- However, as you can see from the graph, people with a FICO score equal to 0 fail to reimburse a strictly positive amount on their loan, not a 0 amount.

You should use an affine prediction function.

- The graph below shows that the function 5000-500*FICO does a much better job at predicting the amount that people fail to repay than the univariate linear regression function 250*FICO
- 5000-500*FICO is a an affine function of FICO, with an intercept equal to 5000, and a slope equal to -500.
- In these lectures, we study OLS univariate affine regression.



Roadmap

1. The OLS univariate affine regression function.
2. Estimating the OLS univariate affine regression function.
3. Interpreting $\hat{\beta}_1$
4. OLS univariate affine regression in practice.

Set up and notation.

- We consider a population of N units.
  - N = number of people who apply for a one year-loan with bank A during April 2018.
  - N = number of emails reaching Gmail accounts in April 2018.
- Each unit k has a variable y_k attached to it that we do not observe. **We call this variable the dependent variable.**
  - In loan example, y_k is a variable equal to the amount of her loan applicant k will fail to reimburse when her loan expires in April 2018.
  - In email example, y_k = 1 if email k is a spam and 0 otherwise.
- Each unit k also has 1 variable x_k attached to it that we do observe. **We call this variable the independent variable.**
  - In loan example, x_k could be the FICO score of applicant k.
  - In email example, x_k =1 if the word "free" appears in the email.
- $\bar{y} = \frac{1}{N}\sum_{k=1}^{N} y_k$ and $\bar{x} = \frac{1}{N}\sum_{k=1}^{N} x_k$: average of y_ks and x_ks.

Your prediction should be a function of x_k

- Based on the value of x_k of each unit, we want to predict her y_k.
- E.g.: in the loan example, we want to predict the amount that unit k will fail to repay on her loan based on her FICO score.
- Assume that applicant 1 has a very high (good) credit score, while applicant 2 has a very low (bad) credit score.
- Should you predict the same value of y_k for applicants 1 and 2?
- No! Your prediction should a function of x_k, $f(x_k)$.
- **In these lectures, we focus on predictions which are a affine function of x_k: $f(x_k) = b_0 + b_1 x_k$, for two real numbers b_0 and b_1.**

Our prediction error is $y_k - (b_0 + b_1 x_k)$.

- Based on the value of x_k of each unit, we want to predict her y_k.
- Our prediction should a function of x_k, $f(x_k)$. We focus on predictions which are a affine function of x_k: $f(x_k) = b_0 + b_1 x_k$, for two real numbers b_0 and b_1.
- $y_k - (b_0 + b_1 x_k)$, the difference between our prediction and y_k, is our prediction error.
- In the loan example, if $y_k - (b_0 + b_1 x_k)$ is large and positive, our prediction is much below the amount applicant k will fail to reimburse.
- If $y_k - (b_0 + b_1 x_k)$ is large and negative, our prediction is much above the amount person k will fail to reimburse.
- Large positive or negative values of $y_k - (b_0 + b_1 x_k)$ mean bad prediction.
- $y_k - (b_0 + b_1 x_k)$ close to 0 means good prediction.

We want to find the value of (b_0, b_1) that minimizes $\sum_{k=1}^{N}\left(y_k - (b_0 + b_1 x_k)\right)^2$

- $\sum_{k=1}^{N}\left(y_k - (b_0 + b_1 x_k)\right)^2$ is positive. => minimizing it = same thing as making it as close to 0 as possible.
- If $\sum_{k=1}^{N}\left(y_k - (b_0 + b_1 x_k)\right)^2$ is as close to 0 as possible, means that the sum of the squared value of our prediction errors is as small as possible.
- => we make small errors. That's good, that's what we want!

The OLS univariate affine regression function in the population.

- Let

$$(\beta_0, \beta_1) = argmin_{(b_0,b_1)\in R^2} \sum_{k=1}^{N} \left(y_k - (b_0 + b_1 x_k)\right)^2$$

- We call $\beta_0 + \beta_1 x_k$ the **ordinary least squares (OLS) univariate OLS affine regression function of $\boldsymbol{y_k}$ on $\boldsymbol{x_k}$ in the population.**
- Affine: because the regression function is an affine function of x_k.
- Shortcut: **OLS regression of $\boldsymbol{y_k}$ on a constant and $\boldsymbol{x_k}$ in the population**.
- Constant: because there is the constant β_0 in our prediction function.

Decomposing y_k between predicted value and error.

- β_0 and β_1: coefficient of the constant and x_k in the OLS regression of y_k on a constant and x_k in the population.
- Let $\tilde{y}_k = \beta_0 + \beta_1 x_k$. $\tilde{y}_k$ is the predicted value for y_k according to the OLS regression of y_k on a constant and x_k in the population.
- Let $e_k = y_k - \tilde{y}_k$. e_k: error we make when we use OLS regression in the population to predict y_k.
- We have $y_k = \tilde{y}_k + e_k$.

y_k =predicted value + error.

$\beta_0 = \bar{y} - \beta_1 \bar{x}$…

- (β_0, β_1): (b_0, b_1) minimizing $\sum_{k=1}^{N} \left(y_k - (b_0 + b_1 x_k)\right)^2$.
- Derivative wrt to b_0 is: $\sum_{k=1}^{N} -2\left(y_k - (b_0 + b_1 x_k)\right)$. Why?
- Derivative wrt to b_1 is: $\sum_{k=1}^{N} -2x_k\left(y_k - (b_0 + b_1 x_k)\right)$. Why?
- (β_0, β_1): value of(b_0, b_1) for which 2 derivatives = 0.
- We use fact 1st derivative = 0 to write β_0 as function of β_1:

$$\sum_{k=1}^{N} -2\left(y_k - (\beta_0 + \beta_1 x_k)\right) = 0$$
$$\text{i}if\ -2\sum_{k=1}^{N}(y_k - \beta_0 - \beta_1 x_k) = 0$$
$$\text{i}if \sum_{k=1}^{N}(y_k - \beta_0 - \beta_1 x_k) = 0$$
$$\text{i}if \sum_{k=1}^{N} y_k - \sum_{k=1}^{N} \beta_0 - \sum_{k=1}^{N} \beta_1 x_k = 0$$
$$\text{i}if \sum_{k=1}^{N} y_k - \sum_{k=1}^{N} \beta_1 x_k = \sum_{k=1}^{N} \beta_0$$
$$\text{i}if \sum_{k=1}^{N} y_k - \beta_1 \sum_{k=1}^{N} x_k = N\beta_0$$
$$\text{i}if\ \frac{1}{N}\sum_{k=1}^{N} y_k - \beta_1 \frac{1}{N}\sum_{k=1}^{N} x_k = \beta_0$$
$$\text{i}if\, \beta_0 = \bar{y} - \beta_1 \bar{x}.$$

2 useful formulas for the next derivation.

- During the sessions, you have proven that

$\frac{1}{N}\sum_{k=1}^{N} {x_k}^2 - \bar{x}^2 = \frac{1}{N}\sum_{k=1}^{N}(x_k - \bar{x})^2$.

- Multiplying both sides by N, equivalent to saying that

$\sum_{k=1}^{N} {x_k}^2 - N\bar{x}^2 = \sum_{k=1}^{N}(x_k - \bar{x})^2$.

- Bear this 1st equality in mind, we use it in next derivation.
- Moreover,

$\sum_{k=1}^{N} \bar{x}(y_k - \bar{y}) = \bar{x}\sum_{k=1}^{N}(y_k - \bar{y}) = \bar{x}\sum_{k=1}^{N} y_k - \bar{x}\sum_{k=1}^{N} \bar{y}$
$= \bar{x}N\bar{y}$-$\bar{x}N\bar{y} = 0$.

- Therefore,

$$\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y}) = \sum_{k=1}^{N}[x_k(y_k - \bar{y}) - \bar{x}(y_k - \bar{y})]$$

$$= \sum_{k=1}^{N} x_k(y_k - \bar{y}) - \sum_{k=1}^{N} \bar{x}(y_k - \bar{y}) = \sum_{k=1}^{N} x_k(y_k - \bar{y})$$

- Bear this 2nd equality in mind, we use it in next derivation.

… and $\beta_1 = \frac{\sum_{k=1}^{N}(x_k-\bar{x})(y_k-\bar{y})}{\sum_{k=1}^{N}(x_k-\bar{x})^2}$

- Now, let's use fact 2nd derivative = 0 and formula for β_0 to find β_1.

$$\sum_{k=1}^{N} -2x_k\big(y_k-(\beta_0+\beta_1 x_k)\big)=0$$
$$\text{if } \sum_{k=1}^{N} x_k\big(y_k-(\beta_0+\beta_1 x_k)\big)=0$$
$$\text{if } \sum_{k=1}^{N}(x_k y_k-\beta_0 x_k-\beta_1 {x_k}^2)=0$$
$$\text{if } \sum_{k=1}^{N} x_k y_k-\sum_{k=1}^{N}\beta_0 x_k-\sum_{k=1}^{N}\beta_1 {x_k}^2=0$$
$$\text{if } \sum_{k=1}^{N} x_k y_k-\beta_0\sum_{k=1}^{N} x_k-\beta_1\sum_{k=1}^{N} {x_k}^2=0$$
$$\text{if } \sum_{k=1}^{N} x_k y_k=\beta_0\sum_{k=1}^{N} x_k+\beta_1\sum_{k=1}^{N} {x_k}^2$$
$$\text{if } \sum_{k=1}^{N} x_k y_k=(\bar{y}-\beta_1\bar{x})\sum_{k=1}^{N} x_k+\beta_1\sum_{k=1}^{N} {x_k}^2$$
$$\text{if } \sum_{k=1}^{N} x_k y_k=\bar{y}\sum_{k=1}^{N} x_k-\beta_1\bar{x}\sum_{k=1}^{N} x_k+\beta_1\sum_{k=1}^{N} {x_k}^2$$
$$\text{if } \sum_{k=1}^{N} x_k y_k-\sum_{k=1}^{N} x_k\bar{y}=\beta_1\Big(\sum_{k=1}^{N} {x_k}^2-\bar{x}\sum_{k=1}^{N} x_k\Big)$$
$$\text{if } \sum_{k=1}^{N}(x_k y_k-x_k\bar{y})=\beta_1\Big(\sum_{k=1}^{N} {x_k}^2-\bar{x}N\bar{x}\Big)$$
$$\text{if } \sum_{k=1}^{N} x_k(y_k-\bar{y})=\beta_1\Big(\sum_{k=1}^{N} {x_k}^2-N\bar{x}^2\Big)$$
$$\text{if } \sum_{k=1}^{N}(x_k-\bar{x})(y_k-\bar{y})=\beta_1\sum_{k=1}^{N}(x_k-\bar{x})^2$$
$$\text{if } \beta_1=\frac{\sum_{k=1}^{N}(x_k-\bar{x})(y_k-\bar{y})}{\sum_{k=1}^{N}(x_k-\bar{x})^2}.$$

Applying the formulas for β_0 and β_1 in an example.

- Assume for a minute that $N = 3$: there are only two units in the population.
- Assume that $y_1 = 2, x_1 = 0, y_2 = 3, x_2 = 1, y_3 = 7$, and $x_3 = 2$.
- Use the previous formulas to compute β_0 and β_1 in this example.

iClicker time

- If $N = 3$, $y_1 = 2$, $x_1 = 0$, $y_2 = 3$ and $x_2 = 1$, $y_3 = 7$ and $x_3 = 2$, then

a) $\beta_0 = \frac{3}{2}$ and $\beta_1 = \frac{7}{2}$

b) $\beta_0 = \frac{3}{2}$ and $\beta_1 = -\frac{7}{2}$

c) $\beta_0 = \frac{3}{2}$ and $\beta_1 = \frac{5}{2}$

$\beta_0 = \frac{3}{2}$ and $\beta_1 = \frac{5}{2}$!

- If $N = 3$, $y_1 = 2$, $x_1 = 0$, $y_2 = 3$, $x_2 = 1$, $y_3 = 7$, and $x_3 = 2$, then $\bar{y} = 4$ and $\bar{x} = 1$.
- Then,

$$\beta_1 = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + (x_3 - \bar{x})(y_3 - \bar{y})}{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2}$$

$$= \frac{(0-1)(2-4) + (1-1)(3-4) + (2-1)(7-4)}{(0-1)^2 + (1-1)^2 + (2-1)^2}$$

$$= \frac{5}{2}.$$

- And $\beta_0 = \bar{y} - \beta_1 \bar{x} = 4 - \frac{5}{2} = \frac{3}{2}$.

Two other useful formulas

- We let $e_k = y_k - (\beta_0 + \beta_1 x_k)$. e_k: error we make when we use a univariate affine regression to predict y_k.
- In the derivation of the formula of β_0, we have shown that $\sum_{k=1}^{N}(y_k - \beta_0 - \beta_1 x_k) = 0$
- This is equivalent to $\sum_{k=1}^{N} e_k = 0$, which is itself equivalent to $\frac{1}{N}\sum_{k=1}^{N} e_k = 0$: the average of our prediction errors is 0.
- In the derivation of the formula of β_0, we have also shown that $\sum_{k=1}^{N} x_k(y_k - \beta_0 - \beta_1 x_k) = 0$
- This is equivalent to $\sum_{k=1}^{N} x_k e_k = 0$, which is itself equivalent to saying $\frac{1}{N}\sum_{k=1}^{N} x_k e_k = 0$: the average of the product of our prediction errors and x_k is 0.

What you need to remember

- Population of N units. Each unit k has 2 variables attached to it: y_k is a variable we do not observe, x_k is a variable we observe.
- We want to predict the y_k of each unit based on her x_k.
- Our prediction should be function of x_k, $f(x_k)$.
- Focus on affine functions: $b_0 + b_1 x_k$, for 2 numbers b_0 and b_1.
- Best (b_0, b_1) is that minimizing $\sum_{k=1}^{N} \left(y_k - (b_0 + b_1 x_k)\right)^2$.
- We call that value (β_0, β_1), we call $\beta_0 + \beta_1 x_k$: OLS regression function of y_k on a constant and x_k, and we let $e_k = y_k - (\beta_0 + \beta_1 x_k)$.
- $\beta_0 = \bar{y} - \beta_1 \bar{x}$, and $\beta_1 = \frac{\sum_{k=1}^{N} (x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{N} (x_k - \bar{x})^2}$.
- We have $\frac{1}{N} \sum_{k=1}^{N} e_k = 0$: average prediction error is 0, and $\frac{1}{N} \sum_{k=1}^{N} x_k e_k = 0$.

Roadmap

1. The OLS univariate affine regression function.
2. Estimating the OLS univariate affine regression function.
3. Interpreting $\hat{\beta}_1$.
4. OLS univariate affine regression in practice.

Can we compute (β_0,β_1)?

- Our prediction for y_k based on a univariate linear regression is $\beta_0 + \beta_1 x_k$, the univariate linear regression function.
- => to be able to make a prediction for a unit's y_k based on her x_k, we need to know the value of (β_0, β_1).
- Under the assumptions we have made so far, can we compute(β_0, β_1)? Discuss this question with your neighbor during 1 minute.

iClicker time

- Can we compute (β_0, β_1)?

a) Yes

b) No

No!

- $\beta_0 = \bar{y} - \beta_1 \bar{x}$, and $\beta_1 = \frac{\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{N}(x_k - \bar{x})^2}$.
- Remember, we have assumed that we observe the x_ks of everybody in the population (e.g. applicants' FICO scores) but not the y_ks (e.g. the amount that a person applying for a one-year loan in April 2018 will fail to reimburse in April 2018 when that loan expires).
- => we cannot compute β_0, and β_1.

A method to estimate β_0 and β_1

- We draw n units from the population, and we measure the dependent and the independent variable of those units.
- For every i between 1 and n, Y_i and X_i = value of dependent and of independent variable of ith unit we randomly select.
- We want to use the Y_is and the X_is to estimate β_0 and β_1.
- (β_0, β_1), (b_0, b_1) minimizing $\sum_{k=1}^{N}\left(y_k - (b_0 + b_1 x_k)\right)^2$.
- => to estimate (β_0, β_1), we use (b_0, b_1) minimizing $\sum_{i=1}^{n}\left(Y_i - (b_0 + b_1 X_i)\right)^2$.
- Instead of finding (b_0, b_1) that minimizes sum of squared prediction errors in population, find (b_0, b_1) that minimizes sum of squared prediction errors in the sample.
- Intuition: if we find a method to predict well the dependent variable in the sample, method should work well in entire population, given that sample representative of population.

The OLS regression function in the sample.

- Let

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = argmin_{(b_0,b_1)\in R^2} \sum_{i=1}^{n} \left(Y_i - (b_0 + b_1 X_i)\right)^2$$

- We call $\hat{\beta}_0 + \hat{\beta}_1 X_i$ the OLS regression function of Y_i on a constant and X_i **in the sample.**
- In the sample: because we only use the Y_is and X_is of the n units in the sample we randomly draw from the population.
- $\left(\hat{\beta}_0, \hat{\beta}_1\right)$: coefficients of the constant and X_i in the OLS regression of Y_i on X_i **in the sample**.
- Let $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. $\hat{Y}_i$ is the predicted value for Y_i according to the OLS regression of Y_i on a constant and X_i **in the sample**.
- Let $\hat{e}_i = Y_i - \hat{Y}_i$. $\hat{e}_i$: error we make when we use OLS regression **in the sample** to predict Y_i.
- We have $Y_i = \hat{Y}_i + \hat{e}_i$.

Find the value of (b_0, b_1) that minimizes $\sum_{i=1}^{n}\left(Y_i - (b_0 + b_1 X_i)\right)^2$.

- Find a formula for the value of (b_0, b_1) that minimizes $\sum_{i=1}^{n}\left(Y_i - (b_0 + b_1 X_i)\right)^2$. Hint: the formula is "almost" the same as that for (β_0, β_1), except that you need to replace:
  - N, size of population by n, size of sample,
  - y_k, the dependent variable of unit k in the population, by Y_i, the dependent variable of unit i in the sample,
  - x_k, the independent variable of unit k in the population, by X_i, the independent variable of unit i in the sample.

iClicker time

- Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Let $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ denote the value of (b_0, b_1) that minimizes $\sum_{i=1}^{n}\left(Y_i - (b_0 + b_1 X_i)\right)^2$. We have:

a) $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, and $\hat{\beta}_1 = \frac{\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{N}(x_k - \bar{x})^2}$.

b) $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, and $\hat{\beta}_1 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{(X_i - \bar{X})^2}$.

c) $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, and $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, and $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$!

- Sketch of the proof.
- Differentiate $\sum_{i=1}^{n}\left(Y_i - (b_0 + b_1 X_i)\right)^2$ wrt to b_0 and b_1.
- $\hat{\beta}_0$ and $\hat{\beta}_1$: values of b_0 and b_1 that cancel these two derivatives. That gives us a system of 2 equations with 2 unknowns to solve.
- The steps to solve it are exactly the same as those we used to find $\beta_0 = \bar{y} - \beta_1 \bar{x}$, and $\beta_1 = \frac{\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{N}(x_k - \bar{x})^2}$, except that we replace:
  - N, the size of the population by n, the size of the sample,
  - y_k, the dependent variable of unit k in the population, by Y_i, the dependent variable of unit i in the sample,
  - x_k, the independent variable of unit k in the population, by X_i, the independent variable of unit i in the sample.

$\hat{\beta}_0$ converges towards β_0, and $\hat{\beta}_1$ converges towards β_1.

- Remember, when we studied the OLS regression of Y_i on X_i without a constant, we used the law of large numbers to prove that $\lim_{n \to +\infty} \hat{\alpha} = \alpha$.
- When the sample we randomly draw gets large, $\hat{\alpha}$, the sample coefficient of the regression, gets close to α, the population coefficient, so $\hat{\alpha}$ is a good proxy for α.
- Here, one can also use the law of large numbers to prove that $\lim_{n \to +\infty} \hat{\beta}_0 = \beta_0$ and $\lim_{n \to +\infty} \hat{\beta}_1 = \beta_1$.
- Take-away: when sample we randomly draw gets large, $\hat{\beta}_0$ and $\hat{\beta}_1$, sample coefficients of the regression of Y_i on a constant and X_i, get close to β_0 and β_1, the population coefficients.
- Therefore, $\hat{\beta}_0$ and $\hat{\beta}_1$ = good proxys of β_0 and β_1 when sample is large enough.

iClicker time

- We have shown that $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$
- Is $\hat{\beta}_1$ a real number, or is it a random variable? Discuss this question with your neighbour for 1mn, and then answer.

a) $\hat{\beta}_1$ a real number

b) $\hat{\beta}_1$ a random variable.

$\hat{\beta}_1$ is a random variable!

- We have shown that $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$
- X_is and Y_is are random variables: their value depends on which unit we randomly draw when we draw ith unit in sample.
- Therefore, $\hat{\beta}_1$ is a random variable, with a variance.
- Let $\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(e_k)^2$ denote the average of the squared of our prediction errors in the population.
- One can show that $V(\hat{\beta}_1) \approx \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$.
- $V(\hat{\beta}_1)$ small if average squared prediction error low, meaning that regression model makes small prediction errors in the population
- $V(\hat{\beta}_1)$ small if high variability of X_i.
- $V(\hat{\beta}_1)$ small if sample size is large.

Using central limit theorem for $\hat{\beta}_1$ to construct a test and a confidence interval.

- If $n \geq 100$, $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{V(\hat{\beta}_1)}}$ follows normal distribution with mean 0 and variance 1.
- We can use this to test null hypothesis on β_1.
- Often, we want to test $\beta_1 = 0$. If $\beta_1 = 0$, OLS regression function is $\beta_0 + 0 \times x_k = \beta_0$. Means that actually x_k is useless to predict y_k. E.g.: best prediction of amount people will fail to repay on their loan is actually not a function of their FICO score, it is just a constant.
- If we want to have 5% chances of wrongly rejecting $\beta_1 = 0$, test is:

Reject $\beta_1 = 0$ if $\frac{\hat{\beta}_1}{\sqrt{V(\hat{\beta}_1)}} > 1.96$ or $\frac{\hat{\beta}_1}{\sqrt{V(\hat{\beta}_1)}} < -1.96$.

Otherwise, do not reject $\beta_1 = 0$.

- We can also construct confidence interval for β_1:

$$\left[\hat{\beta}_1 - 1.96\sqrt{V(\hat{\beta}_1)}, \hat{\beta}_1 + 1.96\sqrt{V(\hat{\beta}_1)}\right].$$

For 95% of random samples we can draw, β_1 belongs to confidence interval.

Assessing quality of our predictions: the MSE

- For every individual in sample, $\hat{e}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)$: error we make when we use sample OLS regression to predict Y_i.
- We have $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{e}_i$.
- $e_k = y_k - (\beta_0 + \beta_1 x_k)$: population prediction errors. $\hat{e}_i$: sample prediction errors.
- Slide 25: we have shown that $\frac{1}{N}\sum_{k=1}^{N} e_k = 0$.
- Similarly, $\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i = 0$. Average sample prediction error=0.
- We cannot use $\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i$ to assess quality of our predictions. Even if our regression makes bad predictions, $\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i$ always equal to 0.
- Instead, we use $\frac{1}{n}\sum_{i=1}^{n} {\hat{e}_i}^2$: **mean-squared error** (MSE) of regression.
- Good to compare regressions: if regression A has a lower MSE than B, A better than B: makes smaller errors on average.
- However, $\frac{1}{n}\sum_{i=1}^{n} {\hat{e}_i}^2$ hard to interpret: if equal to 10, what does that mean? Does not have a natural scale to which we can compare it.

Assessing quality of our predictions: the R^2.

- Instead, we are going to use $R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i{}^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$
- 1 – MSE / sample variance of the Y_is

The R^2 has a natural scale (1/3)

- $\hat{e}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)$ = error we make when we use sample OLS regression to predict Y_i. We have $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{e}_i$.
- $e_k = y_k - (\beta_0 + \beta_1 x_k)$: population errors. $\hat{e}_i$: sample errors.
- Slide 25: we have shown that $\frac{1}{N}\sum_{k=1}^{N} e_k = 0$ and $\frac{1}{N}\sum_{k=1}^{N} x_k e_k = 0$. Average population error = 0, and average product of x_ks and e_ks = 0.
- Similarly, one can show that $\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i = 0$ and $\frac{1}{n}\sum_{i=1}^{n} X_i \hat{e}_i = 0$. Average sample error=0, and average product of the X_is and $\hat{e}_i$s = 0.
- Because of this, one can show that

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{e}_i - \left(\hat{\beta}_0 + \hat{\beta}_1 \bar{X} + \bar{\hat{e}}\right)\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 X_i - \left(\hat{\beta}_0 + \hat{\beta}_1 \bar{X}\right) + \hat{e}_i - \bar{\hat{e}}\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 X_i - \left(\hat{\beta}_0 + \hat{\beta}_1 \bar{X}\right)\right)^2 + \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i{}^2.$$

That's because $\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i = 0$ and $\frac{1}{n}\sum_{i=1}^{n} X_i \hat{e}_i = 0$ implies

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 X_i - \left(\hat{\beta}_0 + \hat{\beta}_1 \bar{X}\right)\right)\left(\hat{e}_i - \bar{\hat{e}}\right) = 0.$$

The R^2 has a natural scale (2/3)

- Let $\hat{e}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)$ denote the error we make when we use the sample OLS regression function to predict Y_i. We have $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{e}_i$.
- One can show that $\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i = 0$ and $\frac{1}{n}\sum_{i=1}^{n} X_i \hat{e}_i = 0$. The average sample prediction error is 0, and the average product of the X_is and $\hat{E}_i$s in the sample is 0.
- Because of this, one can show that

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 X_i - \left(\hat{\beta}_0 + \hat{\beta}_1 \bar{X}\right)\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\hat{e}_i - \bar{\hat{e}}\right)^2.$$

- $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is the sample variance of the Y_is,

$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 X_i - \left(\hat{\beta}_0 + \hat{\beta}_1 \bar{X}\right)\right)^2$ is sample variance of $\hat{\beta}_0 + \hat{\beta}_1 X_i$s, and $\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i{}^2$ is the MSE.

- **The sample variance of the Y_is is equal to the sample variance of $\hat{\beta}_0 + \hat{\beta}_1 X_i$, our predictions for Y_i, plus the MSE of the regression.**

iClicker time

- One can show that

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 X_i - \left(\hat{\beta}_0 + \hat{\beta}_1 \bar{X}\right)\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i{}^2.$$

- $R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i{}^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$.
- Based on the equality above, and based on its definition, which of the following properties should the number R^2 satisfy?

a) R^2 must be included between 0.5 and 1.

b) R^2 must be included between 0.5 and 1.5.

c) R^2 must be included between 0 and 1.

The R^2 has a natural scale: it must be included between 0 and 1 (3/3)

- One has:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0+\hat{\beta}_1X_i-\left(\hat{\beta}_0+\hat{\beta}_1\bar{X}\right)\right)^2+\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i{}^2.$$

- $R^2=1-\frac{\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i{}^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$, so $R^2\leq 1$
- Then, using the fact that

$$\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i{}^2=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2-\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0+\hat{\beta}_1X_i-\left(\hat{\beta}_0+\hat{\beta}_1\bar{X}\right)\right)^2,$$

one can show that

$$R^2=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_0+\hat{\beta}_1X_i-\left(\hat{\beta}_0+\hat{\beta}_1\bar{X}\right)\right)^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}\geq 0.$$

- => R^2= easily interpretable measure of quality of our predictions. If close to 1, our predictions make almost no error (MSE close to 0), so excellent prediction. If close to 0, poor prediction.

What you need to remember

- Prediction for y_k based on OLS regression of y_k on a constant and x_k in the population is $\beta_0 + \beta_1 x_k$, with $\beta_0 = \bar{y} - \beta_1 \bar{x}$, and $\beta_1 = \frac{\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{N}(x_k - \bar{x})^2}$.
- We can estimate (β_0, β_1) if we measure y_ks for random sample.
- For every i between 1 and n, Y_i and X_i = value of dependent and independent variables of ith unit we randomly select.
- (β_0, β_1) is (b_0, b_1) that minimizes $\sum_{k=1}^{N}\left(y_k - (b_0 + b_1 x_k)\right)^2$
- To estimate (β_0, β_1), find (b_0, b_1) minimizing $\sum_{i=1}^{n}\left(Y_i - (b_0 + b_1 X_i)\right)^2$.
- Yields $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, and $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$.
- $V(\hat{\beta}_1) \approx \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$, and if $n \geq 100$, $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{V(\hat{\beta}_1)}}$ follows N(0,1). We can use this to test $\beta_1 = 0$ and get 95% confidence interval for β_1.
- $R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^{\,2}}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$. Close to 1: good prediction. Close to 0: poor prediction.

Roadmap

1. The OLS univariate affine regression function.
2. Estimating the OLS univariate affine regression function.
3. Interpreting $\hat{\beta}_1$
4. OLS univariate affine regression in practice.

A useful reminder: the sample covariance

- Assume randomly draw a sample of and n units from a population, and for each unit observe variables X_i and Y_i.
- Sample covariance between X_i and Y_i is $\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})$.
- Example: $X_i =$ FICO score of ith person, Y_i: amount she defaults.
- If $X_i > \bar{X}$ (person i's FICO > average FICO in sample):
  - If $Y_i > \bar{Y}$ (amount i defaults > average default in sample) then $(X_i-\bar{X})(Y_i-\bar{Y}) > 0$,
  - If $Y_i < \bar{Y}$ then $(X_i-\bar{X})(Y_i-\bar{Y}) < 0$.
- If $X_i < \bar{X}$ (person i's FICO < average FICO in sample):
  - If $Y_i < \bar{Y}$ then $(X_i-\bar{X})(Y_i-\bar{Y}) > 0$.
  - If $Y_i > \bar{Y}$ then $(X_i-\bar{X})(Y_i-\bar{Y}) < 0$.
- When many people have $X_i > \bar{X}$ and $Y_i > \bar{Y}$, and many people have $X_i < \bar{X}$ and $Y_i < \bar{Y}$, then $\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) > 0$.

  X_i and Y_i move in the same direction.
- When many people have both $X_i > \bar{X}$ and $Y_i < \bar{Y}$, and many people have both $X_i < \bar{X}$ and $Y_i > \bar{Y}$, then $\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) < 0$.

  X_i and Y_i move in opposite directions.

iClicker time

- Let X_i = FICO score of ith person, Y_i: amount she defaults.
- Let $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ be their sample covariance.
- Which of the two statements sounds the most likely to you?

a) $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ is strictly positive

b) $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ is strictly negative

In example, likely $\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})<0$

- Typically, one would expect that people with a FICO score below average default more than the average on their loan.
- Similarly, one would expect that people with a FICO score above average default less than the average on their loan.
- Therefore, we expect that people with $X_i<\bar{X}$ also have $Y_i>\bar{Y}$, and people with $X_i>\bar{X}$ also have $Y_i<\bar{Y}$.
- Therefore, it is likely that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})<0$$

iClicker time

- Go back to the formula we derived for $\hat{\beta}_1$, the coefficient of X_i in the sample regression of Y_i on a constant and X_i.
- Which of the following statements is correct:

a) $\hat{\beta}_1$ is equal to the sample covariance between X_i and Y_i divided by the sample variance of X_i.

b) $\hat{\beta}_1$ is equal to the sample covariance between Y_i and Y_i divided by the sample variance of Y_i.

c) $\hat{\beta}_1$ is equal to the sample variance of X_i divided by the sample covariance between X_i and Y_i.

d) None of the three statements above is correct.

$\hat{\beta}_1 =$ sample covariance between X_i and Y_i divided by sample variance of X_i.

- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$
- Multiply numerator and denominator by $\frac{1}{n}$, yields:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- $\hat{\beta}_1 =$ sample covariance between X_i and Y_i divided by sample variance of X_i.
- Therefore, $\hat{\beta}_1 > 0$ if X_i and Y_i move in the same direction, $\hat{\beta}_1 < 0$ if move in opposite directions.
- In the regression of the amount defaulted on a constant and FICO, do you expect that $\hat{\beta}_1 > 0$ or $\hat{\beta}_1 < 0$?

For now, we can interpret the sign of $\hat{\beta}_1$, not its specific value.

- For now, we have seen that $\hat{\beta}_1 > 0$ means that X_i and Y_i move in the same direction, $\hat{\beta}_1 < 0$ means that move in opposite directions.
- Interesting, but does not tell us how we should interpret a specific value of $\hat{\beta}_1$.
- For instance, what does $\hat{\beta}_1 = 3$ mean?
- That's what we are going to see now.

Interpreting $\hat{\beta}_1$ when X_i is binary.

- Assume you run an OLS regression of Y_i on a constant and X_i, where X_i is a binary variable (variable either equal to 0 or to 1).
- Example: you regress Y_i, whether email i is a spam on a constant and X_i, a binary variable equal to 1 if the email has the word "free" in it, and to 0 if the email does not contain that word.
- Then, you have shown / will show during sessions that

$$\hat{\beta}_1 = \frac{1}{n_1}\sum_{i:X_i=1} Y_i - \frac{1}{n_0}\sum_{i:X_i=0} Y_i \,,$$

where n_1 is the number of units that have $X_i = 1$, n_0 is the number of units that have $X_i = 0$, $\sum_{i:X_i=1} Y_i$ is the sum of Y_i of all units with $X_i = 1$, and $\sum_{i:X_i=0} Y_i$ is the sum of Y_i of all units with $X_i = 0$.

- In the spam example, explain with words what $\frac{1}{n_1}\sum_{i:X_i=1} Y_i$, $\frac{1}{n_0}\sum_{i:X_i=0} Y_i$, and $\hat{\beta}_1$ respectively represent. Discuss this question with your neighbour for one minute.

iClicker time

Assume you regress Y_i, whether email i is a spam on a constant and X_i, a binary variable equal to 1 if the email has the word "free" in it, and to 0 if the email does not contain that word. You know that

$$\hat{\beta}_1 = \frac{1}{n_1}\sum_{i:X_i=1} Y_i - \frac{1}{n_0}\sum_{i:X_i=0} Y_i \,.$$

- Which of the following statements is correct?

a) $\frac{1}{n_1}\sum_{i:X_i=1} Y_i$ is the percentage of emails that have the word free among the emails that are spams, $\frac{1}{n_0}\sum_{i:X_i=0} Y_i$ is the percentage of emails that have the word free among the emails that are not spams, so $\hat{\beta}_1$ is the difference between the percentage of emails that have the word free across spams and non spams.

b) $\frac{1}{n_1}\sum_{i:X_i=1} Y_i$ is the percentage of emails that are spams among the emails that have the word free, $\frac{1}{n_0}\sum_{i:X_i=0} Y_i$ is the percentage of emails that are spams among the emails that do not have the word free, so $\hat{\beta}_1$ is the difference between the percentage of emails that are spams across emails that have and do not have the word free.

$\hat{\beta}_1 =$ difference between % of spams across emails with/without word free.

- $\sum_{i:X_i=1} Y_i$ counts the number of spams among emails that have the word free.
- n_1 is the number of emails that have the word free.
- Therefore, $\frac{1}{n_1}\sum_{i:X_i=1} Y_i$: percentage of spams among emails that have the word free.
- Similarly, $\frac{1}{n_0}\sum_{i:X_i=0} Y_i$: percentage of spams among emails that do not have the word free.
- $\hat{\beta}_1 =$ difference between % of spams across emails with/without word free.
- Outside of this example, we have following, very important result:

When you regress Y_i on a constant and X_i, where X_i is a binary variable, $\widehat{\beta}_1$ is the difference between the average value of Y_i among units with $X_i = 1$ and among units with $X_i = 0$.

Testing whether the average of a variable is significantly different between 2 groups.

- When you regress Y_i on a constant and X_i, where X_i is a binary variable, $\hat{\beta}_1$ is the difference between the average value of Y_i among units with $X_i = 1$ and among units with $X_i = 0$ in the sample.
- Similarly, β_1 is difference between the average Y_i among units with $X_i = 1$ and among units with $X_i = 0$ in the full population.
- Remember that if $\frac{\widehat{\beta}_1}{\sqrt{V(\widehat{\beta}_1)}} > 1.96$ or $\frac{\widehat{\beta}_1}{\sqrt{V(\widehat{\beta}_1)}} < -1.96$, we can reject at the 5% level the null hypothesis that $\beta_1 = 0$.
- When we reject $\beta_1 = 0$ in a regression of Y_i on a constant and X_i, where X_i is a binary variable, we reject the null hypothesis that the average of Y_i is the same among units with $X_i = 1$ and among units with $X_i = 0$ in the full population.
- The difference between the average of Y_i between the two groups in our sample is unlikely to be due to chance.
- **Groups have a significantly different average of Y_i at the 5% level.**

What about $\hat{\beta}_0$?

- Assume you run an OLS regression of Y_i on a constant and X_i, where X_i is a binary variable (variable either equal to 0 or to 1).
- Then, you have shown / will show during sessions that

$$\hat{\beta}_0 = \frac{1}{n_0} \sum_{i:X_i=0} Y_i.$$

- $\hat{\beta}_0$: average of Y_i among units with $X_i = 0$.
- $\hat{\beta}_1$ is the difference between the average value of Y_i among units with $X_i = 1$ and among units with $X_i = 0$.
- People sometimes call units with $X_i = 0$ the **reference category**, because $\hat{\beta}_1$ compares the average value of Y_i among units that do not belong to that reference category to units in that reference category.
- In the spam example, $\hat{\beta}_0$: percentage of spams among emails that do not have the word free in them, $\hat{\beta}_1$= difference between percentage of spams across emails that have the word free in them and emails that do not have that word.

To predict Y of a unit, OLS uses average Y_i among units with same X_i as that unit

- Now, let's consider some units j outside of our sample.
- We do not observe their Y_j but we observe their X_j.
- Predicted value of Y_j according to OLS regression: $\hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_j$.
- $\hat{\beta}_0 = \frac{1}{n_0}\sum_{i:X_i=0} Y_i$, and $\hat{\beta}_1 = \frac{1}{n_1}\sum_{i:X_i=1} Y_i - \frac{1}{n_0}\sum_{i:X_i=0} Y_i$.
- So $\hat{Y}_j = \frac{1}{n_0}\sum_{i:X_i=0} Y_i$ for units j such that $X_j = 0$.
- And $\hat{Y}_j = \frac{1}{n_0}\sum_{i:X_i=0} Y_i + \frac{1}{n_1}\sum_{i:X_i=1} Y_i - \frac{1}{n_0}\sum_{i:X_i=0} Y_i = \frac{1}{n_1}\sum_{i:X_i=1} Y_i$ for units j such that $X_j = 1$.
- To make prediction for unit with $X_j = 0$, we use average Y_i among units with $X_i = 0$ in sample.
- To make prediction for a unit with $X_j = 1$, we use average Y_i among units with $X_i = 1$ in sample.
- **Prediction = average Y_i among units with same X_i in sample.**
- In sessions: in regression of Y_i on a constant, OLS prediction = average Y_i among units in sample.

For now, we know how to interpret the value of $\hat{\beta}_1$, but only when X_i binary.

- When X_i binary,

$$\hat{\beta}_1 = \frac{1}{n_1} \sum_{i:X_i=1} Y_i - \frac{1}{n_0} \sum_{i:X_i=0} Y_i \,.$$

- In that special case, $\hat{\beta}_1$ has a very simple interpretation: difference between average Y_i among units with $X_i = 1$ and among units with $X_i = 0$.
- In other words, $\hat{\beta}_1$ measures by the difference between the average of Y_i across subgroups whose X_i differs by one (units with $X_i = 1$ versus units with $X_i = 0$).
- Does this result extend to the case where X_i not binary?

$\hat{\beta}_1$ measures difference between the average of Y_i across subgroups whose X_i differs by one

- When X_i binary, $\hat{\beta}_1$ measures diff. between average of Y_i across subgroups whose X_i differs by one (units with $X_i = 1$ versus $X_i = 0$).
- Now, assume that X_i can be equal to 0, 1, or 2.
- n_0: number of units with $X_i = 0$. n_1: number of units with $X_i = 1$. n_2: number of units with $X_i = 2$.

$$\hat{\beta}_1 = w\left(\frac{1}{n_1}\sum_{i:X_i=1} Y_i - \frac{1}{n_0}\sum_{i:X_i=0} Y_i\right) + (1-w)\left(\frac{1}{n_2}\sum_{i:X_i=2} Y_i - \frac{1}{n_1}\sum_{i:X_i=1} Y_i\right),$$

where w is number included between 0 & 1 that you don't need to know.

- $\hat{\beta}_1$: weighted average of diff. between average Y_i of units with $X_i = 1$ and $X_i = 0$, and of diff. between average Y_i of units with $X_i = 2$ and $X_i = 1$.
- Units with $X_i = 1$ and $X_i = 0$ have a value of X_i that differs by one.
- Units with $X_i = 2$ and $X_i = 1$ have a value of X_i that differs by one.
- => $\hat{\beta}_1$ measures the difference between the average of Y_i across subgroups whose X_i differs by one!

$\hat{\beta}_1$ measures difference between the average of Y_i across subgroups whose X_i differs by one

- When X_i binary, $\hat{\beta}_1$ measures diff. between average of Y_i across subgroups whose X_i differs by one (units with $X_i = 1$ versus $X_i = 0$).
- Now, assume that X_i can be equal to 0, 1, 2,…,K.
- n_0: number of units with $X_i = 0$, n_1: number of units with $X_i = 1$,…, n_K: number of units with $X_i = K$.

$$\hat{\beta}_1 = \sum_{k=1}^{K} w_k \left(\frac{1}{n_k} \sum_{i:X_i=k} Y_i - \frac{1}{n_{k-1}} \sum_{i:X_i=k-1} Y_i \right),$$

where w_k: positive weights summing to 1 that you do not need to know.

- $\hat{\beta}_1$: weighted average of diff. between average Y_i of units with $X_i = 1$ and $X_i = 0$, of diff. between average Y_i of units with $X_i = 2$ and $X_i = 1$,…, of diff. between average Y_i of units with $X_i = K$ and $X_i = K - 1$.
- Units with $X_i = 1$ and $X_i = 0$ have a value of X_i that differs by one.
- Units with $X_i = K$ and $X_i = K - 1$ have a value of X_i that differs by one.
- $\hat{\beta}_1 =$ diff. between average of Y_i across subgroups whose X_i differs by 1!

Logs versus levels

- Assume you regress Y_i on constant and X_i, $\hat{\beta}_1 = 0.5$: when you compare people whose X_i differs by 1, average Y_i 0.5 larger among people whose X_i is 1 unit larger.
- Assume you regress $\ln(Y_i)$ on constant and X_i, $\hat{\beta}_1 = 0.5$: when you compare people whose X_i differs by 1, average $\ln(Y_i)$ 0.5 larger among people whose X_i 1 unit larger.
- Due to properties $\ln$ function, if people whose X_i is 1 unit larger have an average $\ln(Y_i)$ 0.5 larger, average of Y_i 50% larger among those people.
- Assume you regress $\ln(Y_i)$ on constant and $\ln(X_i)$, $\hat{\beta}_1 = 0.5$: when you compare people whose X_i differs by 1%, average Y_i 0.5% larger among people whose X_i 1% larger.
- Regressing Y_i on constant and X_i is useful to study how the mean of Y_i differs **in levels** across units whose X_i differs by one.
- Regressing $\ln(Y_i)$ on constant and X_i is useful to study how the mean of Y_i differs **in relative terms** across units whose X_i differs by one.
- Regressing $\ln(Y_i)$ on constant and $\ln(X_i)$ is useful to study how the mean of Y_i differs **in relative terms** across units whose X_i differs by 1%.

iClicker time

- Assume you observe the wages of a sample of wage earners in the US. You regress Y_i, the monthly wage of person i, on a constant and X_i, a binary variable equal to 1 if i is a female and to 0 if i is a male. Assume that you find $\hat{\beta}_1 = -200$ and $\hat{\beta}_0 = 2000$
- Which of the following statements is correct?

a) In this sample, the average wage of females is 200 dollars higher than the average wage of males, and the average wage of females is 2000 dollars.

b) In this sample, the average wage of females is 200 dollars lower than the average wage of males, and the average wage of males is 2000 dollars.

Average wage of females is 200 dollars lower than average wage of males.

- X_i binary: $X_i = 0$ for males, $X_i = 1$ for females.
- $\hat{\beta}_1 = \frac{1}{n_1}\sum_{i:X_i=1} Y_i - \frac{1}{n_0}\sum_{i:X_i=0} Y_i$, Therefore, $\hat{\beta}_1 =$ difference between average wage of females and males.
- $\hat{\beta}_1 = -200$ means that females make 200 dollars less than males on average.
- $\hat{\beta}_0 = \frac{1}{n_0}\sum_{i:X_i=0} Y_i$, Therefore, $\hat{\beta}_0 =$ average wage of males.
- $\hat{\beta}_0 = 2000$ means that males make 2000 dollars on average.

iClicker time

- Assume you observe the wages of a sample of 5,000 wage earners in the US. You regress Y_i, the monthly wage of person i, on a constant and X_i, a binary variable equal to 1 if i is a female and to 0 if i is a male. Assume that Eviews or Stata tells you that $\hat{\beta}_1 = -200$ and $\sqrt{V(\hat{\beta}_1)} = 20$. Which of the following statements is correct?

a) In this sample, the average wage of females is 200 dollars lower than the average wage of males, and the difference between the average wage of the two groups is statistically significant at the 5% level.

b) In this sample, the average wage of females is 200 dollars lower than the average wage of males, and the difference between the average wage of the two groups is not statistically significant at the 5% level.

$$\frac{\widehat{\beta}_1}{\sqrt{V(\widehat{\beta}_1)}} = -10, \text{ so we reject } \beta_1 = 0 \text{ at 5\%}$$

- $\frac{\widehat{\beta}_1}{\sqrt{V(\widehat{\beta}_1)}} = -10$, so we reject $\beta_1 = 0$ at 5% level.
- The difference between the average wage of males and females is statistically significant at the 5% level.
- It is very unlikely (less than 5% chances) that in the US population males and females have the same average wages, but that we drew a random sample fairly different from the US population where males' average wage is 200 higher than that of female.
- Given that our random sample is quite large (5,000 people), the fact that in our sample the average wage of males is 200 dollars > than that of females indicates that in the US population, males also have a higher average wage than females.

iClicker time

- Assume you observe the wages of a sample of wage earners in the US. You regress Y_i, the monthly wage of person i, on a constant and X_i, a binary variable equal to 1 if i is a female and to 0 if i is a male. Assume that you find $\hat{\beta}_1 = -200$ and $\hat{\beta}_0 = 2000$
- Which of the following statements is correct?

a) To predict the wage of a female not in the sample, this regression model will use the average wage of females in the sample.

b) To predict the wage of a female not in the sample, this regression model will use the average wage of males and females in the sample.

To predict wage of a female not in sample, regression uses average wage of females in sample.

- Now, let's consider some units j outside of our sample => we do not observe their Y_j.
- Predicted value of Y_j according to OLS regression: $\hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_j$.
- Given that j female, $X_j = 1$, so predicted wage: $\hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1$.
- $\hat{\beta}_0 = \frac{1}{n_0}\sum_{i:X_i=0} Y_i$, and $\hat{\beta}_1 = \frac{1}{n_1}\sum_{i:X_i=1} Y_i - \frac{1}{n_0}\sum_{i:X_i=0} Y_i$, so

$$\hat{Y}_j = \frac{1}{n_0}\sum_{i:X_i=0} Y_i + \frac{1}{n_1}\sum_{i:X_i=1} Y_i - \frac{1}{n_0}\sum_{i:X_i=0} Y_i = \frac{1}{n_1}\sum_{i:X_i=1} Y_i$$

- Predicted wage: average wage of females in sample.

iClicker time

- Assume you observe the wages of a sample of wage earners in the US. You regress $\ln(Y_i)$, the monthly wage of person i, on a constant and X_i, a binary variable equal to 1 if i is a female and to 0 if i is a male. Assume that you find $\hat{\beta}_1 = -0.1$.
- Which of the following statements is correct?

a) In this sample, the average wage of females is 0.1 dollars lower than the average wage of males.

b) In this sample, the average wage of females is 10% lower than the average wage of males.

Average wage of females is 10% lower than average wage of males.

- X_i binary: $X_i = 0$ for males, $X_i = 1$ for females.
- $\hat{\beta}_1 = \frac{1}{n_1}\sum_{i:X_i=1} \ln(Y_i) - \frac{1}{n_0}\sum_{i:X_i=0} \ln(Y_i)$, Therefore, $\hat{\beta}_1 =$ difference between average ln(wage) of females and males.
- $\hat{\beta}_1 = -0.1$ means that the average ln(wage) of females 0.1 lower than the average ln(wage) of males.
- As we discussed a few slides ago, using some properties of the ln function, one can show that this implies that the average wage of females is 10% lower than the average wage of males.

iClicker time

- Assume you observe the wages of a sample of wage earners in the US. You regress Y_i, the monthly wage of person i, on a constant and X_i, their number of years of professional experience (from 0 for people who just started working to 50 for people who have worked for 50 years). Assume that you find $\hat{\beta}_1 = 100$.
- Which of the following statements is correct?

a) When we compare people whose years of experience differ by one, we find that on average, those who have one more year of experience earn 100 more dollars per month.

b) The covariance between years of experience and wage is equal to 100.

c) The covariance between years of experience and wage divided by the variance of years of experience is equal to 100.

Answers a) and c) both correct

- X_i can be equal to 0 (no experience), 1, 2,…50.
- Let n_0 be number of units with $X_i = 0$ (no experience),…, let n_{50} be number of units with $X_i = 50$ (50 years of experience).

$$\hat{\beta}_1 = \sum_{k=1}^{50} w_k \left(\frac{1}{n_k} \sum_{i:X_i=k} Y_i - \frac{1}{n_{k-1}} \sum_{i:X_i=k-1} Y_i \right),$$

where w_k are positive weights summing to 1 that you do not need to know.

- $\hat{\beta}_1$: weighted average of difference between average wage of people with 1 and 0 years of experience, of difference between average wage of people with 2 and 1 years of experience,…, of difference between average wage of units with 50 and 49 years of experience.
- $\hat{\beta}_1 = 100$ means that when we compare people whose years of experience differ by one, we find that on average, those who have one more year of experience earn 100 more dollars per month.
- Answer c) also correct. However, ratio of covariance and variance hard to interpret, while average difference of wages of people with one year of difference in their experience easy to interpret.

iClicker time

- Assume you observe the wages of a sample of wage earners in the US. You regress $\ln(Y_i)$, the ln(monthly wage) of person i, on a constant and $\ln(X_i)$, the ln(number of years of professional experience) of that person. Assume that you find $\hat{\beta}_1 = 0.5$.
- Which of the following statements is correct?

a) When we compare people whose years of experience differ by one, we find that on average, those who have one more year of experience earn 50% more.

b) When we compare people whose years of experience differ by 1%, we find that on average, those who have 1% more years of experience earn 0.5% more.

Answer b) correct

- We regress $\ln(Y_i)$, the ln(monthly wage) of person i, on a constant and $\ln(X_i)$, the ln(number of years of professional experience) of that person.
- Because $\ln(X_i)$ and not X_i in regression, $\hat{\beta}_1$ does not compare subgroups whose experience differ by 1 one year, but subgroups whose experience differ by 1%!
- In this sample, when we compare subgroups of people whose years of experience differ by 1%, we find that on average, those who have 1% more years of experience earn 0.5% more.

What you need to remember

- $\hat{\beta}_1 =$ sample covariance between X_i and Y_i / by sample variance of X_i.
- $\hat{\beta}_1 > 0$ (resp. $\hat{\beta}_1 < 0$): covariance between X_i and $Y_i > 0$ (resp. < 0): X_i and Y_i positively correlated, move in same (resp. opposite) direction.
- When X_i binary, $\hat{\beta}_1 = \frac{1}{n_1}\sum_{i:X_i=1} Y_i - \frac{1}{n_0}\sum_{i:X_i=0} Y_i$: difference between the average of Y_i among subgroups whose X_i differs by one (units with $X_i = 1$ versus units with $X_i = 0$).
- When X_i not binary, $\hat{\beta}_1$ still measures difference between average of Y_i among subgroups whose X_i differs by one.
- You need to know how to interpret $\hat{\beta}_1$ in a regression of Y_i on a constant and X_i, in a regression of $\ln(Y_i)$ on a constant and X_i, and in a regression of $\ln(Y_i)$ on a constant and $\ln(X_i)$.

Roadmap

1. The OLS univariate affine regression function.
2. Estimating the OLS univariate affine regression function.
3. Interpreting $\hat{\beta}_1$
4. OLS univariate affine regression in practice.

How Gmail uses OLS univariate affine regression

- Gmail wants to predict y_k: 1 if email k is spam, 0 otherwise.
- To do so, use x_k: 1 if "free" appears in email, 0 otherwise.
- x_k easy to measure (a computer can do it automatically, by searching for "free" in the email), but y_k is hard to measure: only a human can know whether an email is a spam or not. => cannot observe y_k for all emails.
- To make good predictions, would like to compute, (β_0, β_1), value of (b_0, b_1) minimizing $\sum_{k=1}^{N}\left(y_k - (b_0 + b_1 x_k)\right)^2$, and then use $\beta_0 + \beta_1 x_k$ to predict y_k. $\beta_0 + \beta_1 x_k$: affine function of x_k for which sum of squared prediction errors $\left(y_k - (b_0 + b_1 x_k)\right)^2$ minimized.
- Issue: $\beta_0 = \bar{y} - \beta_1 \bar{x}$, and $\beta_1 = \frac{\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{N}(x_k - \bar{x})^2}$ => they cannot compute these numbers because do not observe y_k.

How Gmail uses OLS univariate affine regression

- Instead Gmail draws random sample of, say, 5000 emails, ask humans to read them and determine whether spams or not.
- For i between 1 and 5000, Y_i: whether ith randomly drawn email is spam, X_i: whether ith randomly drawn email has free in it.
- (β_0, β_1) is value of (b_0, b_1) minimizing $\sum_{k=1}^{N}\left(y_k - (b_0 + b_1 x_k)\right)^2$
- Estimate (β_0, β_1): use (b_0, b_1) minimizing $\sum_{i=1}^{n}\left(Y_i - (b_0 + b_1 X_i)\right)^2$.
- Yields $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ and $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$.
- For emails not in sample, do not know if spam, but use $\hat{\beta}_0 + \hat{\beta}_1 x_k$ as their prediction of whether the email is a spam or not.
- Because their random sample of emails is large, $\hat{\beta}_0$ and $\hat{\beta}_1$ should be close to β_0 and β_1, and therefore $\hat{\beta}_0 + \hat{\beta}_1 x_k$ should be close to $\beta_0 + \beta_1 x_k$, the best univariate affine prediction of y_k given x_k.
- Use R^2 to assess whether regression makes good predictions.

Application to a data set of 4601 emails

- 4601 emails which have been read by humans. Variable spam = 1 if email = spam, 0 otherwise.
- We have another variable: number of times the word “free” appears in the email/number of words in the email *100. Ranges from 0 to 100: percentage points.
- We go to Eviews and write “ls spam c percent_word_free”.

Dependent Variable: SPAM
Method: Least Squares
Date: 04/26/17 Time: 15:56
Sample: 1 4601
Included observations: 4601

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PERCENT_WORD_FREE	0.201984	0.023411	8.627873	0.0000
C	0.372927	0.007555	49.35958	0.0000

R-squared	0.015928	Mean dependent var	0.394045
Adjusted R-squared	0.015714	S.D. dependent var	0.488698
S.E. of regression	0.484843	Akaike info criterion	1.390450
Sum squared resid	1081.098	Schwarz criterion	1.393247
Log likelihood	-3196.730	Hannan-Quinn criter.	1.391434
F-statistic	74.44020	Durbin-Watson stat	0.032029
Prob(F-statistic)	0.000000		

Interpretation of $\hat{\beta}_1$

- $\hat{\beta}_0 = 0.37$ and $\hat{\beta}_1 = 0.20$. Interpretation of $\hat{\beta}_1$:

When we compare emails whose percentage of words that are the word “free” differ by 1, percentage of spams is 20 points higher among emails whose percentage of the word free is 1 point higher.

- Emails where the word free appears more often are more likely to be spams!

Using $\hat{\beta}_0$ and $\hat{\beta}_1$ to make predictions

- $\hat{\beta}_0 = 0.37$ and $\hat{\beta}_1 = 0.20$. Assume you consider two emails outside of your sample, and therefore you do not know whether they are spams or not.
- In one email, the word “free” =0% of the words of the email, the other one where the word “free”=1% of the words of the email.
- According to the OLS affine regression function, what is your prediction for the first email being a spam? What is your prediction for the second email being a spam? Discuss this question with your neighbor for 2 minutes.

iClicker time

- $\hat{\beta}_0 = 0.37$ and $\hat{\beta}_1 = 0.20$. Assume you consider two emails, one where the word "free" =0% of the words of the email, the other one where the word "free"=1% of the words of the email.
- According to the OLS affine regression function, what is your prediction for the first email being a spam? What is your prediction for the second email being a spam?

a) The predicted value for the first email being a spam is 0.37, while the predicted value for the second email being a spam is 0.372.

b) The predicted value for the first email being a spam is 0.37, while the predicted value for the second email being a spam is 0.57.

Predicted value for 1st email being spam is 0.37, predicted value for 2nd email being spam is 0.57.

- $\hat{\beta}_0 = 0.37$ and $\hat{\beta}_1 = 0.20$. Assume you consider two emails, one where the word "free" =0% of the words of the email, the other one where the word "free"=1% of the words of the email.
- According to the OLS affine regression function, what is your prediction for the first email being a spam? What is your prediction for the second email being a spam?
- According to this regression, predicted value for whether email is a spam is $\hat{\beta}_0$+$\hat{\beta}_1 x$, where x is number of times "free" appears in the email/number of words in the email * 100.
- For first email $x = 0$=> predicted value = 0.37.
- For second email, $x = 1$ => predicted value = 0.57.

Testing $\beta_1 = 0$.

- $\hat{\beta}_1 = 0.20$, and $\sqrt{V(\hat{\beta}_1)} = 0.023$.
- Can we reject at the 5% level the null hypothesis that $\beta_1 = 0$? Discuss this question with your neighbor for 1 minute.

iClicker time

- $\hat{\beta}_1 = 0.20$, and $\sqrt{V(\hat{\beta}_1)} = 0.023$.
- Can we reject at the 5% level the null hypothesis that $\beta_1 = 0$?

a) Yes

b) No

Yes!

- If we want to have 5% chances of wrongly rejecting $\beta_1 = 0$, test is:

Reject $\beta_1 = 0$ if $\frac{\hat{\beta}_1}{\sqrt{V(\hat{\beta}_1)}} > 1.96$ or $\frac{\hat{\beta}_1}{\sqrt{V(\hat{\beta}_1)}} < -1.96$.

Otherwise, do not reject $\beta_1 = 0$.

- Here, $\frac{\hat{\beta}_1}{\sqrt{V(\hat{\beta}_1)}} = 8.63$ => we can reject $\beta_1 = 0$.
- The percentage of the words of the email that are the word "free" is a statistically significant predictor of whether the email is a spam or not!

- Find the 95% confidence interval for β_1. You have 2mns.

iClicker time

- $\hat{\beta}_1 = 0.20$, and $\sqrt{V(\hat{\beta}_1)} = 0.023$.
- The 95% confidence interval for β_1 is:

a) [0.155,0.245]

b) [0.143,0.228]

95% confidence interval for β_1 is [0.155,0.245]

- $\hat{\beta}_1 = 0.20$, and $\sqrt{V(\hat{\beta}_1)} = 0.023$.
- The 95% confidence interval for β_1 is $\left[\hat{\beta}_1 - 1.96\sqrt{V(\hat{\beta}_1)}, \hat{\beta}_1 + 1.96\sqrt{V(\hat{\beta}_1)}\right]$.
- Plugging in the values of $\hat{\beta}_1$ and $\sqrt{V(\hat{\beta}_1)}$ yields [0.155,0.245].

iClicker time

Dependent Variable: SPAM
Method: Least Squares
Date: 04/26/17 Time: 15:56
Sample: 1 4601
Included observations: 4601

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PERCENT_WORD_FREE	0.201984	0.023411	8.627873	0.0000
C	0.372927	0.007555	49.35958	0.0000

R-squared	0.015928	Mean dependent var	0.394045
Adjusted R-squared	0.015714	S.D. dependent var	0.488698
S.E. of regression	0.484843	Akaike info criterion	1.390450
Sum squared resid	1081.098	Schwarz criterion	1.393247
Log likelihood	-3196.730	Hannan-Quinn criter.	1.391434
F-statistic	74.44020	Durbin-Watson stat	0.032029
Prob(F-statistic)	0.000000		

- Does regression has a low or a high R-squared?

a) It has a low R-squared.

b) It has a high R-squared.

Our regression has a low R^2

- The R^2 of the regression is equal to 0.016.
- R^2 included between 0 and 1. Close to 0: bad prediction. Close to 1 good prediction.
- Here close to 0 => bad prediction.

If we use this regression to construct a spam filter, filter will be pretty bad.

- We can compute $\hat{\beta}_0+\hat{\beta}_1 x$ for each email in our sample.
- 39% of those 4601 emails are spams => we could say: we predict that the 39% of emails with highest value of $\hat{\beta}_0+\hat{\beta}_1 x$ are spams, while the other emails are not spams.
- We can look how this spam filter performs in our sample.
- Among the non-spams we correctly predict that 85% are not spams, but we wrongly predict that 15% are spams.
- Among the spams, we correctly predict that 35% are spams, but we wrongly predict that 65% are non-spams.
- => if Gmail used this spam filter, you would receive many spams, Gmail would send many true emails to your trash, and you would change your email account to Microsoft.
- In the homework, you will see how to construct a better spam filter.

What you need to remember, and what's next

- In practice, many instances where we can measure the $y_k s$, the variable we do not observe for everyone, for a sample of population.
- We can use that sample to compute $\hat{\beta}_0$ and $\hat{\beta}_1$, and then use $\hat{\beta}_0 + \hat{\beta}_1 x_k$ as our prediction of the $y_k s$ we do not observe.
- If that sample is a random sample from the population, $\hat{\beta}_0 + \hat{\beta}_1 x_k$ should be close to $\beta_0 + \beta_1 x_k$, the best affine prediction for y_k.
- But univariate affine regression might still not give great predictions: spam example.
- There are better prediction methods available. Next lectures: we see one of them.