Inference for regression

Prof. Jacob M. Montgomery

Quantitative Political Methodology (L32 363)

November 1, 2017

Overview

Last time:

- How to think about regression

Overview

Last time:

- How to think about regression
- "Best" parameters for drawing a line through data

Overview

Last time:

- How to think about regression
- "Best" parameters for drawing a line through data
- Estimating regression model parameters

Overview

Last time:

- How to think about regression
- "Best" parameters for drawing a line through data
- Estimating regression model parameters

Overview

Last time:

- How to think about regression
- "Best" parameters for drawing a line through data
- Estimating regression model parameters

This time:

- Conditional variance

Overview

Last time:

- How to think about regression
- "Best" parameters for drawing a line through data
- Estimating regression model parameters

This time:

- Conditional variance
- Confidence intervals and hypothesis testing with regression

Overview

Last time:

- How to think about regression
- "Best" parameters for drawing a line through data
- Estimating regression model parameters

This time:

- Conditional variance
- Confidence intervals and hypothesis testing with regression
- Reading regression tables

Where did that statistic come from again?

We have a statistical model:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

Where did that statistic come from again?

We have a statistical model:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

We have "derived" the best statistic for estimating these parameters.

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

Where did that statistic come from again?

We have a statistical model:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

We have "derived" the best statistic for estimating these parameters.

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

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

Just one more statistic

Conditional standard deviation

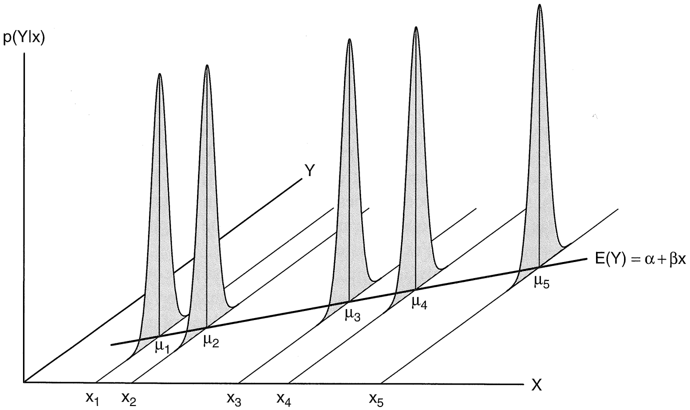
$$\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}}$$

Unconditional standard deviation

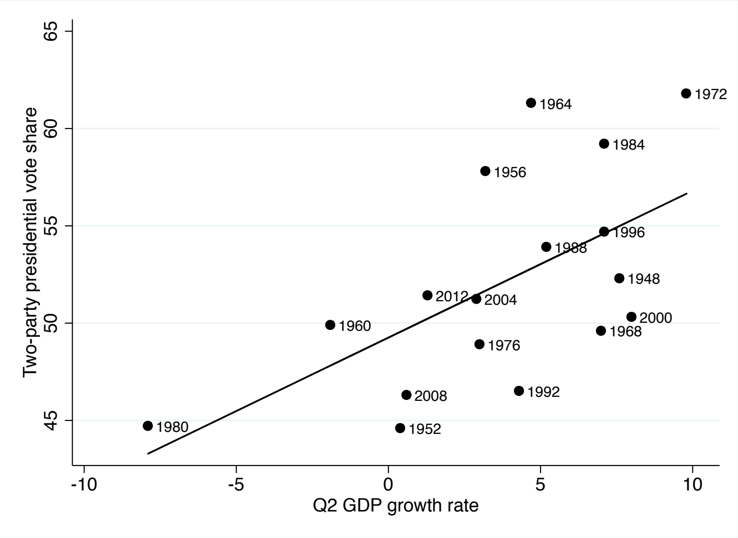
$$\hat{\sigma}_Y = \sqrt{\frac{\sum(Y_i - \bar{Y})^2}{n-1}}$$

Conditional variance will be smaller. Why is there a difference?

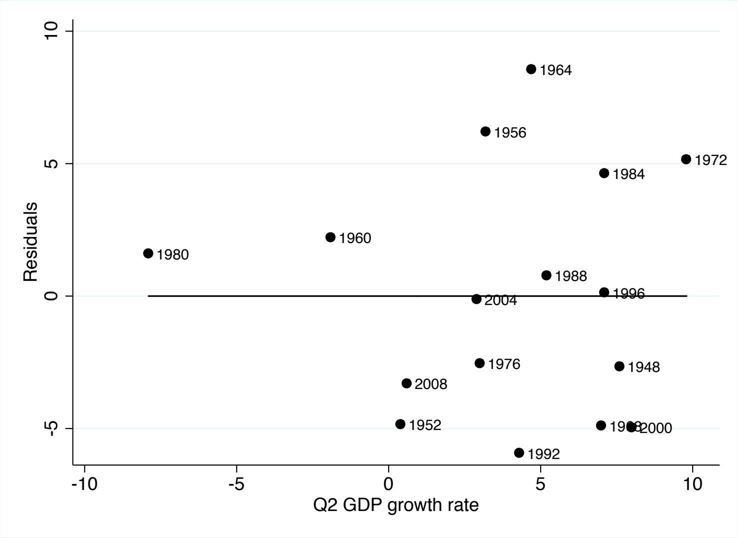
Conditional variance gives a measure of spread of the residuals.



Conditional variance gives a measure of spread of the residuals.



Conditional variance gives a measure of spread of the residuals.



You have a statistic. So what?

We used the *sample* data to make estimates about *population parameters*.

You have a statistic. So what?

We used the *sample* data to make estimates about *population parameters*. How can we make an inference about the population?

You have a statistic. So what?

We used the *sample* data to make estimates about *population parameters*. How can we make an inference about the population?

Hypothesis testing:

1. Stating assumptions
2. Specifying hypotheses
3. Calculating a test statistic
4. Calculating a p-value
5. Drawing conclusions

Assumptions

- Random sample from population

Assumptions

- Random sample from population
- Linear relationship

Assumptions

- Random sample from population
- Linear relationship

Assumptions

- Random sample from population
- Linear relationship
  (NOTE: this does **not** mean we assume they are all on the line)
- Errors around the line are normally distributed

Assumptions

- Random sample from population
- Linear relationship
  (NOTE: this does **not** mean we assume they are all on the line)
- Errors around the line are normally distributed
- Variance is constant

Assumptions

- Random sample from population
- Linear relationship
  (NOTE: this does **not** mean we assume they are all on the line)
- Errors around the line are normally distributed
- Variance is constant

Assumptions

- Random sample from population
- Linear relationship
  (NOTE: this does **not** mean we assume they are all on the line)
- Errors around the line are normally distributed
- Variance is constant (and some other stuff ... next week)
- X "fixed" and exogenous

There are thousands (millions?) of pages on what to do when these assumptions don't hold.

Assumptions

- Random sample from population
- Linear relationship
  (NOTE: this does **not** mean we assume they are all on the line)
- Errors around the line are normally distributed
- Variance is constant (and some other stuff ... next week)
- X "fixed" and exogenous

There are thousands (millions?) of pages on what to do when these assumptions don't hold.

Assumptions

- Random sample from population
- Linear relationship
  (NOTE: this does **not** mean we assume they are all on the line)
- Errors around the line are normally distributed
- Variance is constant (and some other stuff ... next week)
- X "fixed" and exogenous

There are thousands (millions?) of pages on what to do when these assumptions don't hold. You don't have to worry about that (for now).

State hypotheses

We have parameter estimates $(\hat{\beta}, \hat{\alpha})$. So we can make some hypotheses.

$$H_0 : \beta = 0$$
$$H_a : \beta \neq 0$$

Or, for the intercept:

$$H_0 : \alpha = 0$$
$$H_a : \alpha \neq 0$$

State hypotheses

We have parameter estimates $(\hat{\beta}, \hat{\alpha})$. So we can make some hypotheses.

$$H_0 : \beta = 0$$
$$H_a : \beta \neq 0$$

Or, for the intercept:

$$H_0 : \alpha = 0$$
$$H_a : \alpha \neq 0$$

What do all of these mean?

Calculate test statistic

By now, you should be able to guess generally.

Calculate test statistic

By now, you should be able to guess generally. It will be the point estimate divided by the standard error.

For β:

$$t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$$

Calculate test statistic

By now, you should be able to guess generally. It will be the point estimate divided by the standard error.

For β:

$$t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$$

$$\hat{\sigma}_{\hat{\beta}} = \frac{\hat{\sigma}}{\sqrt{\sum (X_i - \bar{X})^2}}$$

Calculate test statistic

By now, you should be able to guess generally. It will be the point estimate divided by the standard error.

For β:

$$t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$$

$$\hat{\sigma}_{\hat{\beta}} = \frac{\hat{\sigma}}{\sqrt{\sum(X_i - \bar{X})^2}}$$

$$d.f. = n - 2$$

This will always, always be the t-distribution.

Calculate test statistic

By now, you should be able to guess generally. It will be the point estimate divided by the standard error.

For β:

$$t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$$

$$\hat{\sigma}_{\hat{\beta}} = \frac{\hat{\sigma}}{\sqrt{\sum(X_i - \bar{X})^2}}$$

$$d.f. = n - 2$$

This will always, always be the t-distribution.

$\Rightarrow CI = \hat{\beta} \pm t \times \hat{\sigma}_{\hat{\beta}}$
Where t has $n-2$ degrees of freedom, and we look at $t_{\frac{\alpha}{2}}$

Calculate test statistic

By now, you should be able to guess generally. It will be the point estimate divided by the standard error.

For β:

$$t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$$

$$\hat{\sigma}_{\hat{\beta}} = \frac{\hat{\sigma}}{\sqrt{\sum(X_i - \bar{X})^2}}$$

$$d.f. = n - 2$$

This will always, always be the t-distribution.

$\Rightarrow CI = \hat{\beta} \pm t \times \hat{\sigma}_{\hat{\beta}}$
Where t has $n-2$ degrees of freedom, and we look at $t_{\frac{\alpha}{2}}$
P-values and decision-making are the same.

Calculate test statistic

For α:

$$t = \frac{\hat{\alpha}}{\hat{\sigma}_{\hat{\alpha}}}$$

Calculate test statistic

For α:

$$t = \frac{\hat{\alpha}}{\hat{\sigma}_{\hat{\alpha}}}$$

$$\hat{\sigma}_{\hat{\alpha}} = \frac{\hat{\sigma}}{Something}$$

Calculate test statistic

For α:

$$t = \frac{\hat{\alpha}}{\hat{\sigma}_{\hat{\alpha}}}$$

$$\hat{\sigma}_{\hat{\alpha}} = \frac{\hat{\sigma}}{Something}$$

$$d.f. = n - 2$$

This will always, always be the t-distribution.

Calculate test statistic

For α:

$$t = \frac{\hat{\alpha}}{\hat{\sigma}_{\hat{\alpha}}}$$

$$\hat{\sigma}_{\hat{\alpha}} = \frac{\hat{\sigma}}{Something}$$

$$d.f. = n - 2$$

This will always, always be the t-distribution.

$\Rightarrow CI = \hat{\alpha} \pm t \times \hat{\sigma}_{\hat{\alpha}}$
Where t has $n - 2$ degrees of freedom, and we look at $t_{\frac{\alpha}{2}}$

Calculate test statistic

For α:

$$t = \frac{\hat{\alpha}}{\hat{\sigma}_{\hat{\alpha}}}$$

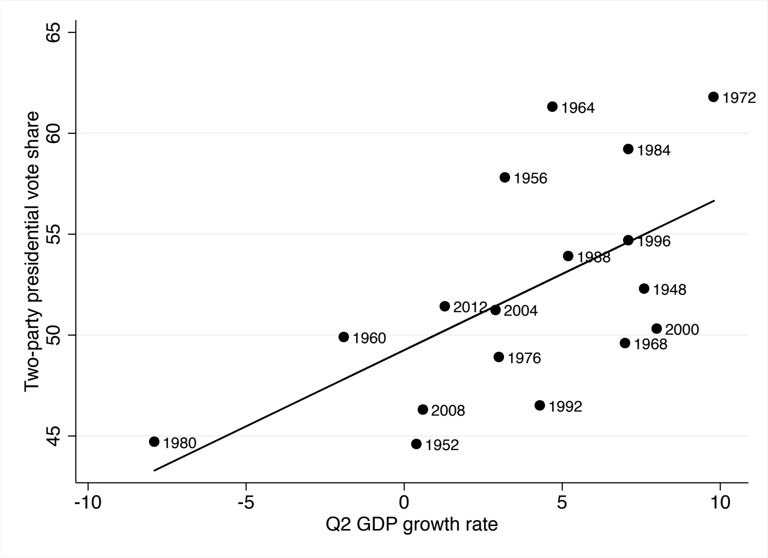
$$\hat{\sigma}_{\hat{\alpha}} = \frac{\hat{\sigma}}{Something}$$

$$d.f. = n - 2$$

This will always, always be the t-distribution.

$\Rightarrow CI = \hat{\alpha} \pm t \times \hat{\sigma}_{\hat{\alpha}}$
Where t has $n-2$ degrees of freedom, and we look at $t_{\frac{\alpha}{2}}$
P-values and decision-making are the same.

GDP Growth and Presidential Elections



R output

```
Call:
lm(formula = vote ~ q2gdp, data = Abram)

Residuals:
   Min     1Q Median     3Q    Max
-6.002 -3.409  0.084  2.078  8.496

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.2560     1.4411  34.179 1.21e-15 ***
q2gdp         0.7549     0.2578   ?  ?
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.481 on 15 degrees of freedom
Multiple R-squared:  0.3637,Adjusted R-squared:  0.3213
F-statistic: 8.573 on 1 and 15 DF,  p-value: 0.01039
```

R output

```
Call:
lm(formula = vote ~ q2gdp, data = Abram)

Residuals:
   Min     1Q Median     3Q    Max
-6.002 -3.409  0.084  2.078  8.496

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.2560     1.4411  34.179 1.21e-15 ***
q2gdp         0.7549     0.2578    ?  ?
---
```

With your team, calculate the p-value for q2gdp.

R output

```
summary(lm(formula = vote ~ q2gdp, data = Abram))

Call:
lm(formula = vote ~ q2gdp, data = Abram)

Residuals:
   Min     1Q Median     3Q    Max
-6.002 -3.409  0.084  2.078  8.496

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.2560     1.4411  34.179 1.21e-15 ***
q2gdp         0.7549     0.2578   2.928   0.0104 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.481 on 15 degrees of freedom
```

Misc.

- If there are lots of variables, hypothesis testing for the β's are about the same.

Misc.

- If there are lots of variables, hypothesis testing for the β's are about the same.

Misc.

- If there are lots of variables, hypothesis testing for the β's are about the same. Although the interpretation is different.
- If your t-statistic is near 2, you have a pretty strong relationship.

Misc.

- If there are lots of variables, hypothesis testing for the β's are about the same. Although the interpretation is different.
- If your t-statistic is near 2, you have a pretty strong relationship.
- If your t-statistic is near 2, you have a pretty strong relationship.

Misc.

- If there are lots of variables, hypothesis testing for the β's are about the same. Although the interpretation is different.
- If your t-statistic is near 2, you have a pretty strong relationship.
- If your t-statistic is near 2, you have a pretty strong relationship.
- If your t-statistic is near 2, you have a pretty strong relationship

Misc.

- If there are lots of variables, hypothesis testing for the β's are about the same. Although the interpretation is different.
- If your t-statistic is near 2, you have a pretty strong relationship.
- If your t-statistic is near 2, you have a pretty strong relationship.
- If your t-statistic is near 2, you have a pretty strong relationship
- Not all tables will include t-statistics