Model Fit

Jacob M. Montgomery

Quantitative Political Methodology

Model Fit

Poster questions?

Road map

Where we have been:

- Single-variable regression
- Multivariate regression
- Regression and causal inference

Road map

Where we have been:

- Single-variable regression
- Multivariate regression
- Regression and causal inference

Today:

- Review of correlation (r)
- RMSE and Model fit (r^2)
- F-tests
- Multivariate model fit
- Time for posters

Review of Correlation

Pearson's r

Review of Correlation

Pearson's r (Standardized slope)

Review of Correlation

Pearson's r (Standardized slope)

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

Review of Correlation

Pearson's r (Standardized slope)

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

- $$S_X = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

Review of Correlation

Pearson's r (Standardized slope)

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

- $$S_X = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

- $$r = \left(\frac{S_x}{S_Y}\right)\hat{\beta}$$

Reminder: These are the main parameters

$$Y = \alpha + X\beta + \epsilon$$

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

Reminder: These are the main parameters

$$Y = \alpha + X\beta + \epsilon$$

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

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

Reminder: These are the main parameters

$$Y = \alpha + X\beta + \epsilon$$

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

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

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

How good is our model?: Thinking about variance

- Unconditional variance: Estimate of total variance in the population

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

How good is our model?: Thinking about variance

- Unconditional variance: Estimate of total variance in the population

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

- Sum of Squared Error: A measure of "spread" around the line

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

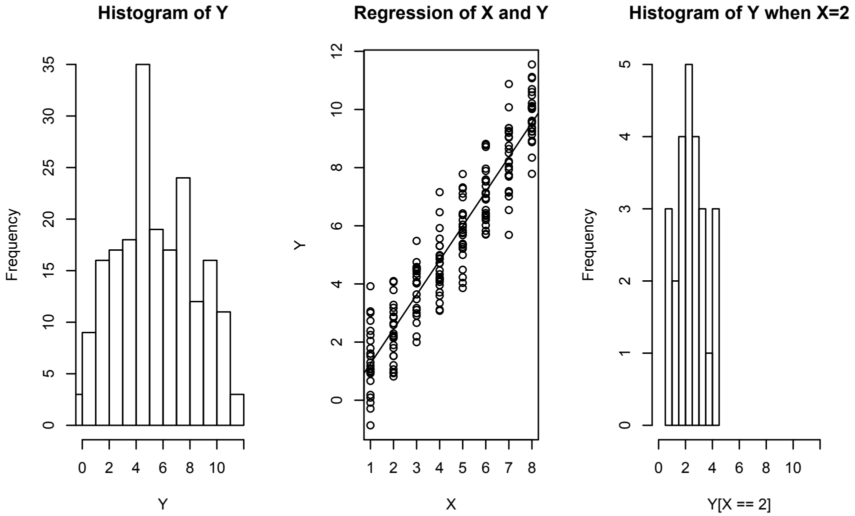
- Conditional Variance: Estimate of variance around line in population

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

- Conditional Variance: Estimate of variance around line in population

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

- $\hat{\sigma}^2$ is sometimes called "Mean squared error" (MSE) and $\hat{\sigma}$ is "Root mean squared error" (RMSE) or "Residual standard error" (in `R`) or "Standard error of the estimate" (in SPSS).



A really, really good line will have small conditional variance.

Evaluating model fit: Hold onto some basic ideas

- $\sum(Y_i - \bar{Y})^2$

Evaluating model fit: Hold onto some basic ideas

- $\sum(Y_i - \bar{Y})^2 =$ Total Sum of Squares
- Unconditional variance: $S^2 = \frac{\sum(Y_i - \bar{Y})^2}{n-1}$

Evaluating model fit: Hold onto some basic ideas

- $\sum(Y_i - \bar{Y})^2 =$ Total Sum of Squares
- Unconditional variance: $S^2 = \frac{\sum(Y_i - \bar{Y})^2}{n-1}$
- $\sum(Y_i - \hat{Y}_i)^2$

Evaluating model fit: Hold onto some basic ideas

- $\sum(Y_i - \bar{Y})^2 =$ Total Sum of Squares
- Unconditional variance: $S^2 = \frac{\sum(Y_i - \bar{Y})^2}{n-1}$
- $\sum(Y_i - \hat{Y}_i)^2 =$ Sum of Squared Error

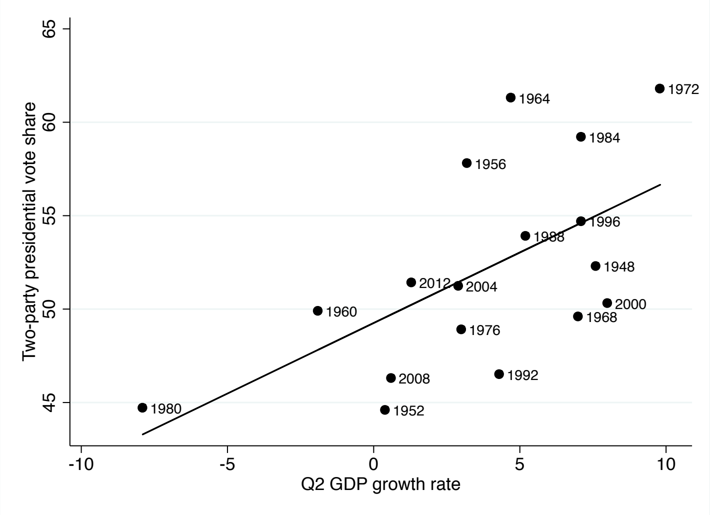
Evaluating model fit: Hold onto some basic ideas

- $\sum(Y_i - \bar{Y})^2 =$ Total Sum of Squares
- Unconditional variance: $S^2 = \frac{\sum(Y_i - \bar{Y})^2}{n-1}$
- $\sum(Y_i - \hat{Y}_i)^2 =$ Sum of Squared Error
- Conditional variance: $\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}$

Evaluating model fit: Hold onto some basic ideas

- $\sum (Y_i - \bar{Y})^2 =$ Total Sum of Squares
- Unconditional variance: $S^2 = \frac{\sum (Y_i - \bar{Y})^2}{n-1}$
- $\sum (Y_i - \hat{Y}_i)^2 =$ Sum of Squared Error
- Conditional variance: $\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}$
- We are going to say that IF we have a really good model, $\hat{\sigma}^2$ should be "a lot" smaller than S^2.

Let's go back: Regression between GDP growth and election outcomes



We draw the line that reduces SSE

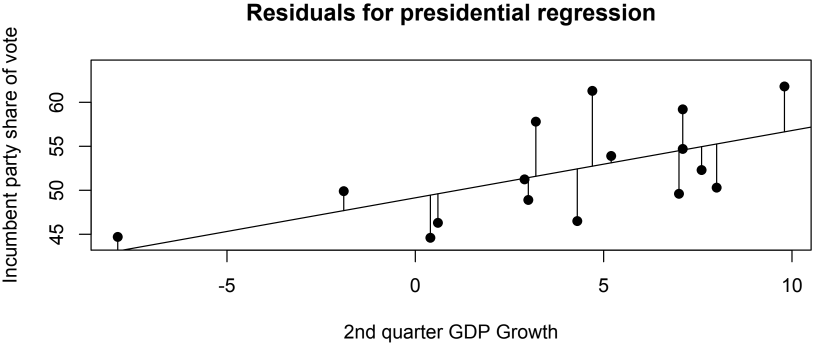
- Residuals:

$$e_i = (Y_i - \hat{Y}_i) = (y_i - \hat{\alpha} - \hat{\beta} X_i)$$

We draw the line that reduces SSE

- Residuals:

$$e_i = (Y_i - \hat{Y}_i) = (y_i - \hat{\alpha} - \hat{\beta} X_i)$$



So … how good is your model? r^2

Define some preliminary terms:

- $TSS = \sum(Y_i - \bar{Y})^2$
- $SSE = \sum(Y_i - \hat{Y}_i)^2$

So ... how good is your model? r^2

Define some preliminary terms:

- $TSS = \sum(Y_i - \bar{Y})^2$
- $SSE = \sum(Y_i - \hat{Y}_i)^2$

Let's define r^2

$$r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}} = \frac{\text{Total Variance - Unexplained Variance}}{\text{Total Variance}}$$

So . . . how good is your model? r^2

Define some preliminary terms:

- $TSS = \sum(Y_i - \bar{Y})^2$
- $SSE = \sum(Y_i - \hat{Y}_i)^2$

Let's define r^2

$$r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}} = \frac{\text{Total Variance - Unexplained Variance}}{\text{Total Variance}}$$

$$r^2 = \frac{TSS - SSE}{TSS}$$

Some notes on r^2

$$r^2 = \frac{TSS - SSE}{TSS}$$

- SSE=0 (Perfect fit) $r^2 = 1$

Some notes on r^2

$$r^2 = \frac{TSS - SSE}{TSS}$$

- SSE=0 (Perfect fit) $r^2 = 1$
- SSE = TSS (No fit) $r^2 = 0$

Some notes on r^2

$$r^2 = \frac{TSS - SSE}{TSS}$$

- SSE=0 (Perfect fit) $r^2 = 1$
- SSE = TSS (No fit) $r^2 = 0$
- Does not depend on units of measurement

Some notes on r^2

$$r^2 = \frac{TSS - SSE}{TSS}$$

- SSE=0 (Perfect fit) $r^2 = 1$
- SSE = TSS (No fit) $r^2 = 0$
- Does not depend on units of measurement
- Sometimes called the "coefficient of determination"

Some notes on r^2

$$r^2 = \frac{TSS - SSE}{TSS}$$

- SSE=0 (Perfect fit) $r^2 = 1$
- SSE = TSS (No fit) $r^2 = 0$
- Does not depend on units of measurement
- Sometimes called the "coefficient of determination"
- It does not penalize for "model complexity." Often "adjusted R-squared" is used.

Some notes on r^2

$$r^2 = \frac{TSS - SSE}{TSS}$$

- SSE=0 (Perfect fit) $r^2 = 1$
- SSE = TSS (No fit) $r^2 = 0$
- Does not depend on units of measurement
- Sometimes called the "coefficient of determination"
- It does not penalize for "model complexity." Often "adjusted R-squared" is used.
- In R-output this is labeled "Multiple R-squared"

Some notes on r^2

$$r^2 = \frac{TSS - SSE}{TSS}$$

- SSE=0 (Perfect fit) $r^2 = 1$
- SSE = TSS (No fit) $r^2 = 0$
- Does not depend on units of measurement
- Sometimes called the "coefficient of determination"
- It does not penalize for "model complexity." Often "adjusted R-squared" is used.
- In R-output this is labeled "Multiple R-squared"
- Why do we use it?
  - Gives us an overall impression for how well our model is doing.
  - We can *informally* compare models.

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   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
Multiple R-squared:  0.3637,    Adjusted R-squared:  0.3213
F-statistic: 8.573 on 1 and 15 DF,  p-value: 0.01039
```

Primer on the F-statistic for regression

Before we said that r^2 is intuitively $r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}$.

Primer on the F-statistic for regression

Before we said that r^2 is intuitively $r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}$. It makes sense then that $(1 - r^2)$ is the percent of variance we haven't explained.

Primer on the F-statistic for regression

Before we said that r^2 is intuitively $r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}$. It makes sense then that $(1 - r^2)$ is the percent of variance we haven't explained. It turns out that:

- F-statistic for regression

$$F = \frac{r^2/p}{(1 - r^2)/[n - (p + 1)]}$$

Primer on the F-statistic for regression

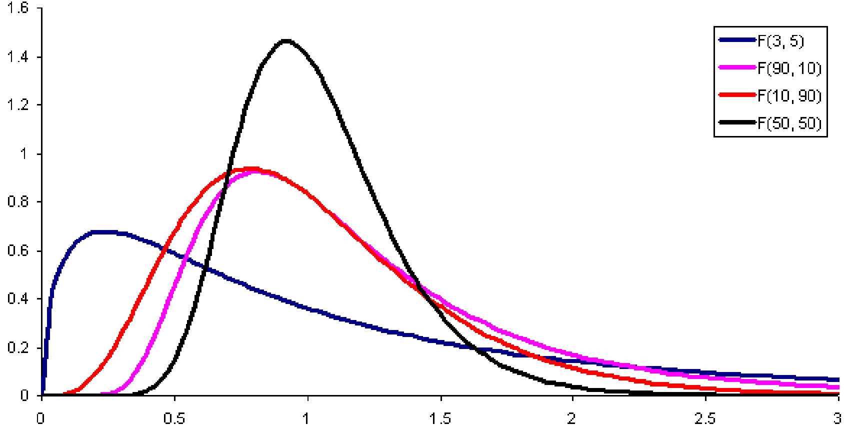
Before we said that r^2 is intuitively $r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}$. It makes sense then that $(1 - r^2)$ is the percent of variance we haven't explained. It turns out that:

- F-statistic for regression

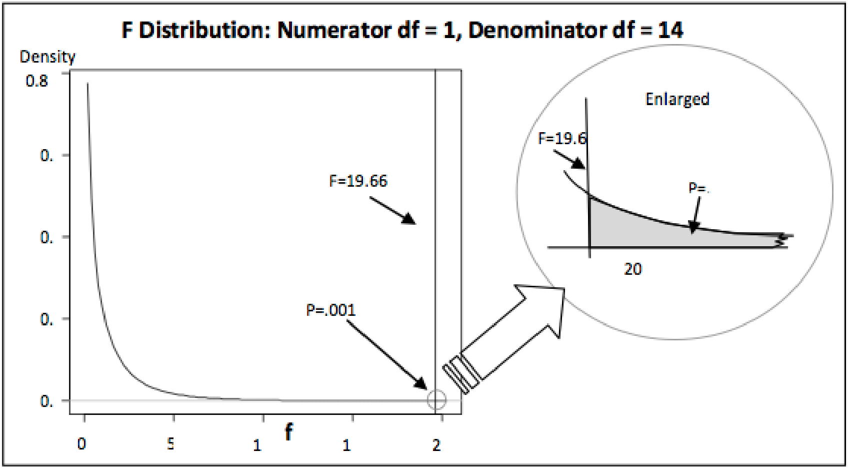
$$F = \frac{r^2/p}{(1 - r^2)/[n - (p + 1)]}$$

- Here p is the number of covariates (gdp, incumbent, etc.), and n is the number of observations. This will be distributed according to the F-distribution with $df_1 = p$, and $df_2 = n - (p + 1)$.

Example F-Distributions



And now you understand (almost) everything on a regression table.



Interpreting the F-test

- Is our model any good?

Interpreting the F-test

- Is our model any good?
- This is a *formalized* way of asking whether our model is any good.

Interpreting the F-test

- Is our model any good?
- This is a *formalized* way of asking whether our model is any good.
- We compare the amount of variance explained by the regression to the amount unexplained.

Interpreting the F-test

- Is our model any good?
- This is a *formalized* way of asking whether our model is any good.
- We compare the amount of variance explained by the regression to the amount unexplained.
- This is essentially a comparison of the following two models:
  - $H_0 : Y_i = \alpha + \epsilon_i$
  - $H_a : Y_i = \alpha + X_i\beta + \epsilon_i$

Interpreting the F-test

- Is our model any good?
- This is a *formalized* way of asking whether our model is any good.
- We compare the amount of variance explained by the regression to the amount unexplained.
- This is essentially a comparison of the following two models:
  - $H_0 : Y_i = \alpha + \epsilon_i$
  - $H_a : Y_i = \alpha + X_i\beta + \epsilon_i$
- This is more useful in multivariate regression:
  - $H_0 : Y_i = \alpha + \epsilon_i$
  - $H_a : Y_i = \alpha + X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + \ldots + \epsilon_i$

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   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
Multiple R-squared:  0.3637,    Adjusted R-squared:  0.3213
F-statistic: 8.573 on 1 and 15 DF,  p-value: 0.01039
```