Chapter 6: Diagnostics

Stats 500, Fall 2017
Brian Thelen, University of Michigan
443 West Hall, bjthelen@umich.edu

Diagnostics

- Checking error assumptions
- Finding unusual points
- Checking the structure of the model

Checking Error Assumptions

Assumption made so far: $\epsilon \sim N(0, \sigma^2 I)$
This includes

- $E(\epsilon) = 0$
- $var(\epsilon) = \sigma^2 I$
- ϵ's are independent, identically distributed, normal

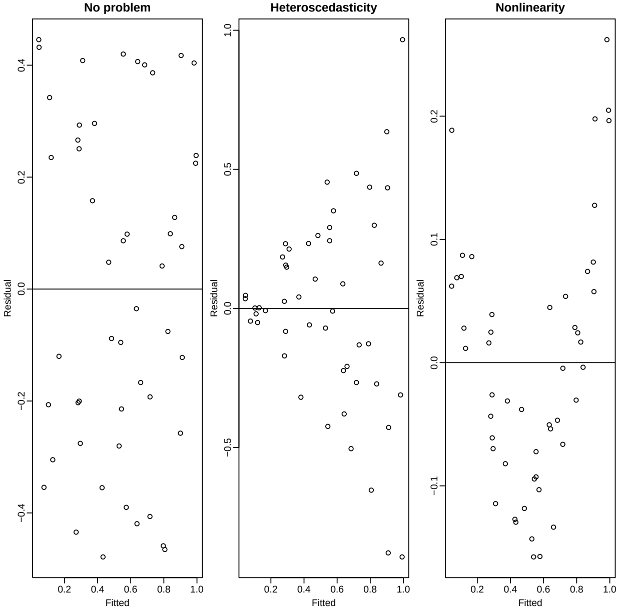
Graphical and numerical diagnostic methods

Constant Variance

Plot $\hat{\epsilon}$ against $\hat{y}$. Can show

- Homoscedasticity (constant variance)
- Heteroscedasticity (non-constant variance)
- Non-linearity

Constant Variance Illustration



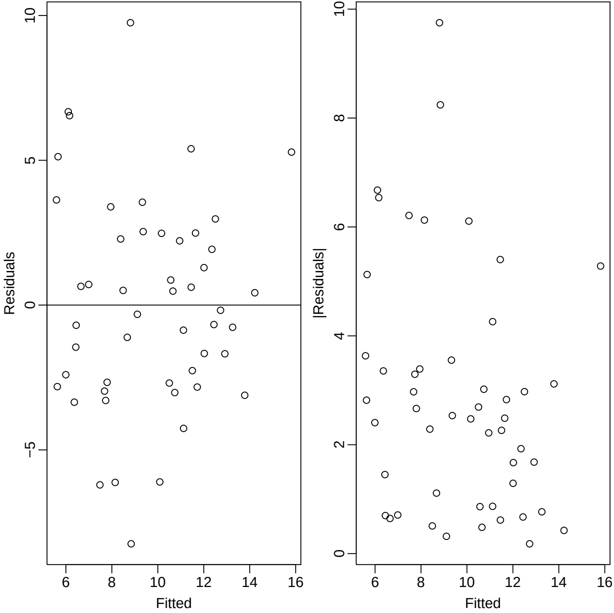
Checking Constant Variance: Example

- 50 different countries, $1960 - 1970$
- Response: aggregate personal saving divided by disposable income (*sr*)
- Predictors: per capital disposable income (*dpi*), percentage rate of change in per capita disposable income (*ddpi*), percentage of population under 15 (*pop15*), percentage of population over 75 (*pop75*)

```
> data(savings)
> result <- lm(sr ~ pop15 + pop75 + dpi + ddpi,
    savings)

## Plot residuals vs fitted values
> plot(result$fitted, result$residual,
       xlab="Fitted", ylab="Residuals")
> abline(h=0)
## Plot absolute values of residuals vs
## fitted values
> plot(result$fitted, abs(result$residual),
       xlab="Fitted", ylab="|Residuals|")
> summary(lm(abs(result$residual) ~
    result$fitted))
Coefficients:
            Estimate Std.Error t value Pr(>|t|)
(Intercept)   4.8397    1.1865   4.079 0.000170
result$fitted-0.2035    0.1185  -1.717 0.092506
```

Savings Example Ctd



What to Do

- Heteroscedasticity
  - Weighted least squares (Ch 8)
  - Transformation of the response (Ch 9)
- Nonlinearity: change the model (Ch 9)

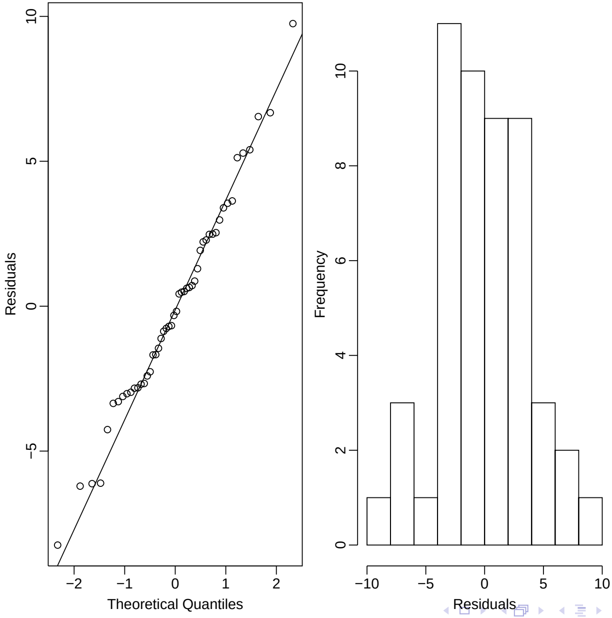
Checking Normality

QQ-plot

1. Sort the residuals $\hat{\epsilon}_{[1]} \leq \hat{\epsilon}_{[2]} \cdots \leq \hat{\epsilon}_{[n]}$
2. Compute $u_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$
3. Plot $\hat{\epsilon}_{[i]}$ against u_i.

```
## QQ-plot
> qqnorm(result$residual, ylab="Residuals")
> qqline(result$residual)
## Histogram
> hist(result$residual, xlab="Residuals")
```

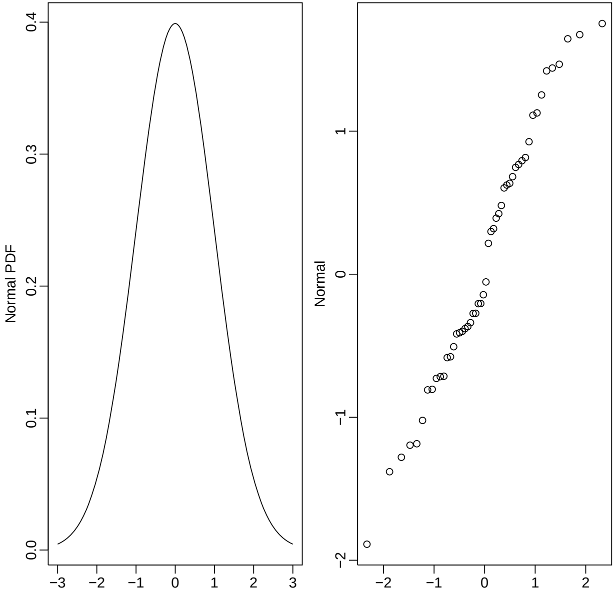
QQ-plot Example



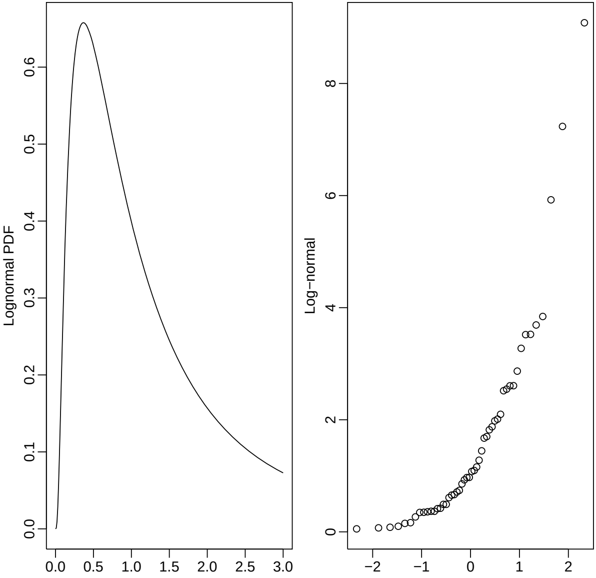
Non-Normality

- Skewed distribution (e.g., log-normal)
- Long-tailed distribution (e.g., Cauchy)
- Short-tailed distribution (e.g., uniform)

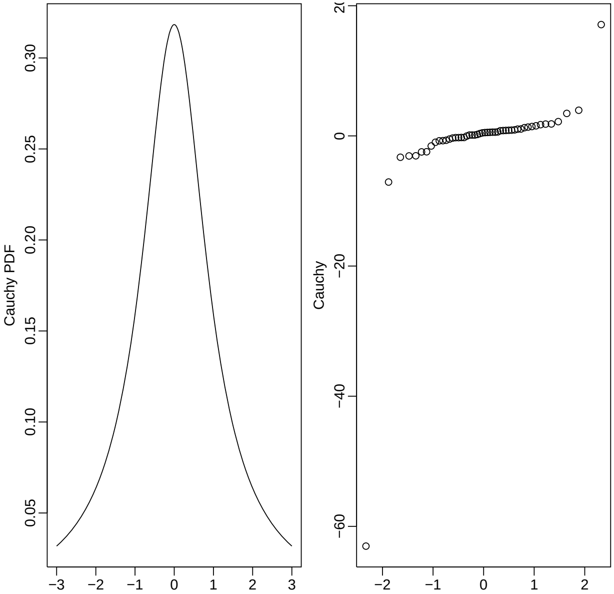
QQ-plot of Normal



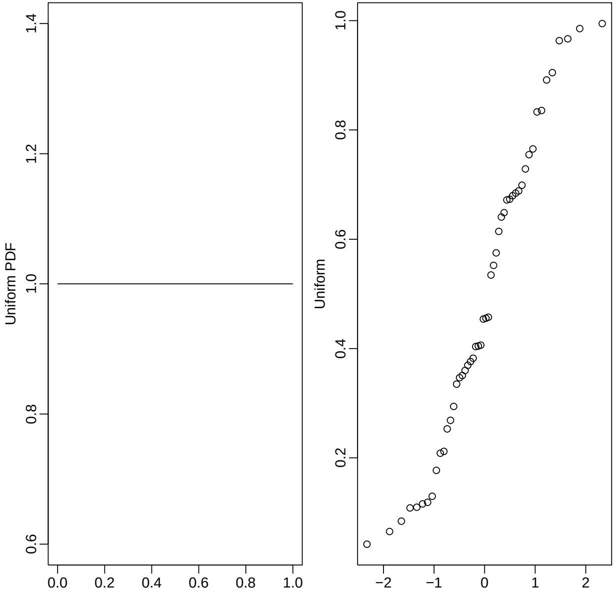
QQ-plot of Log-normal



QQ-plot of Cauchy



QQ-plot of Uniform



Shapiro-Wilk test for normality

```
> shapiro.test(result$residual)
        Shapiro-Wilk normality test
data:  result$residual
W = 0.987, p-value = 0.8524
```

Not very helpful (QQ plots are better).

- Small n – little power
- Large n – non-normality is less important

What to do about non-normal errors

- Transformation of the response (skewed errors)
- Robust methods (long-tailed distribution)
- Inference based on other distributions

Ch 8 & 9

Correlated Errors

Temporally related data

- Plot $\hat{\epsilon}$ against time
- Plot $\hat{\epsilon}_i$ against $\hat{\epsilon}_{i-1}$
- Time series analysis is probably more appropriate than regression

No temporal relationship or other ordering in the variables $\Rightarrow$ checking independence is very hard.

Finding Unusual Points

1. Outliers – do not fit the model well
2. Influential points – affect the fit of the model substantially

A point can be none, one, or both of these.

Leverage

Recall the hat matrix $H = X(X^T X)^{-1} X^T$.
Leverage of point i: $h_i = H_{ii}$.

- h_i depends only on X
- $var(\hat{\epsilon}_i) = \sigma^2(1 - h_i)$
- $\sum_i h_i = p + 1$
- $h_i \geq 1/n$

Rule of thumb: Leverages greater than $2(p+1)/n$ are considered high.

Half-normal Plot

Remark. Half-normal plots can be used to assess outliers, relative to the pattern.

Steps for assessing "outlier" leverages

- Sort $h_{[1]} \leq h_{[2]} \leq \cdots \leq h_{[n]}$
- Compute $u_i = \Phi^{-1}\left(\frac{n+i}{2n+1}\right)$
- Plot $h_{[i]}$ against u_i

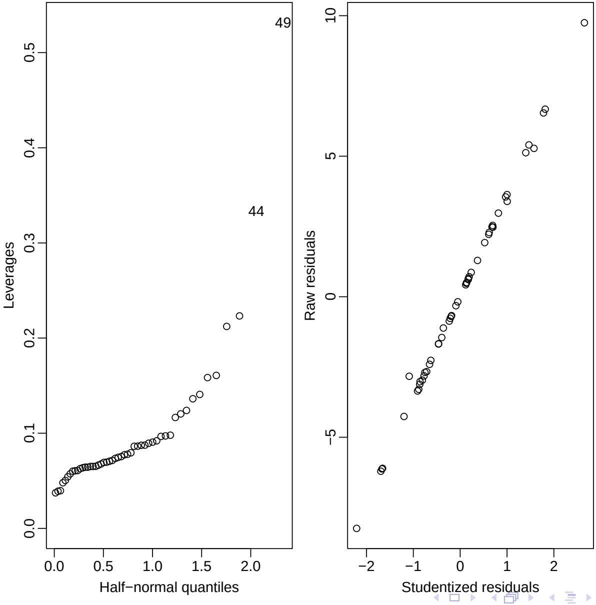
Unlike QQ-plot, not looking for a straight line, looking for points that diverge from the rest of the points.

Need to use `library(faraway)` in R.

Savings Example

```
> data(savings)
> result <- lm(sr ~ ., data=savings)
##  Half-normal plot for leverages
> halfnorm(lm.influence(result)$hat, nlab = 2,
           ylab="Leverages")
> savings[c(44,49),]
                sr pop15 pop75     dpi  ddpi
United States 7.56 29.81  3.43 4001.89  2.45
Libya         8.89 43.69  2.07  123.58 16.71
```

Savings Example Continued



Studentized Residuals

Since $var(\hat{\epsilon}_i) = \sigma^2(1 - h_i)$, let

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

These are called (internally) studentized residuals

- It is better to use studentized residuals for diagnostic plots (QQ-plot and testing constant variance)
- In practice, usually little difference (see plot on previous page)

Savings Example

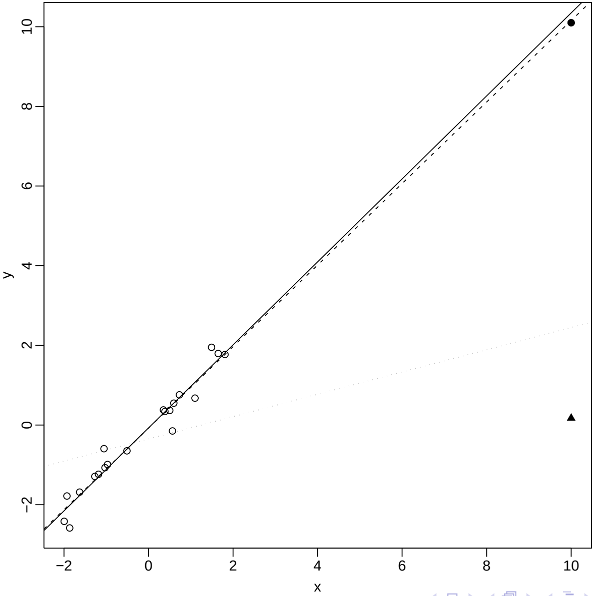
```
## Compute studentized residuals
> result.s <- summary(result)
> sigma.s <- result.s$sig
> hat.s <- lm.influence(result)$hat
> stud.res <- result$residuals/(sigma.s * sqrt(1-hat.s))
> plot(stud.res, result$residuals,
    xlab="Studentized residuals",
    ylab="Raw residuals")
```

Outliers

How do we distinguish between truly unusual points and large residuals?

- Exclude point i, recompute $\hat{\beta}_{(i)}$ and $\hat{y}_{(i)} = x_i^T \hat{\beta}_{(i)}$.
- If $|y_i - \hat{y}_{(i)}|$ is large, then observation i is an outlier; but how large is large?

Which Point is an Outlier?



Externally Studentized Residuals

It turns out

$$
\begin{aligned}
t_i &= \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)} \sqrt{1 + x_i^T \left(X_{(i)}^T X_{(i)} \right)^{-1} x_i}} \\
&= r_i \left(\frac{n - (p+1) - 1}{n - (p+1) - r_i^2} \right)^{1/2} \\
&\sim t_{n-(p+1)-1}
\end{aligned}
$$

The book also calls these jackknife residuals.

Multiple Hypothesis Tests

- If $|t_i|$ is too large, reject and conclude observation i is an outlier.
- For each observation i, compare $|t_i|$ with $t_{n-(p+1)-1}^{\alpha/2}$.
- Will reject too many points. Why?

Bonferroni Correction

$$\begin{aligned} \text{Type I Error} &= Pr_{H_0}(\text{reject at least one test}) \\ &\leq \sum_i Pr_{H_0}(\text{reject test } i) \\ &= n\alpha \end{aligned}$$

Bonferroni correction: test each hypothesis at level α/n

Savings Example

```
## Compute (externally) studentized residuals
> ti <- rstudent(result)
> max(abs(ti))
[1] 2.853558
> which(ti == max(abs(ti)))
Zambia
    46
## Compute p-value
> 2*(1-pt(max(abs(ti)), df=50-5-1))
[1] 0.006566663
## compare to alpha/n
> 0.05/50
[1] 0.001
```

Remarks on Outliers

- Two or more outliers can hide each other.
- Cluster of outliers: consider using robust methods.
- Examine the context – what could it mean?
  - Occasionally data entry errors occur
  - Lurking variables may be part of the explanation
  - Something going wrong: e.g., fraudulent use of credit cards
  - A new unknown effect (you may get a Nobel prize if you can explain it!)
  - Some patterns just have exceptions...

Influential Points

An influential point is one whose removal from the dataset would cause a large change in the fit. At least one of the following:

- Outlier
- Large leverage

Measure the influence:

- Change in the coefficients $\hat{\beta} - \hat{\beta}_{(i)}$
- Change in the fit $X^T(\hat{\beta} - \hat{\beta}_{(i)}) = \hat{y} - \hat{y}_{(i)}$

Cook's Distance

Cook statistic:

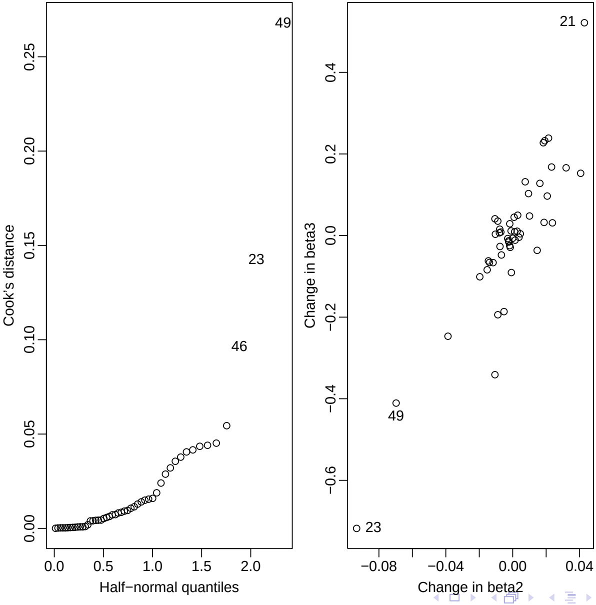
$$
\begin{aligned}
D_i &= \frac{(\hat{y} - \hat{y}_{(i)})^T(\hat{y} - \hat{y}_{(i)})}{(p+1)\hat{\sigma^2}} \\
&= \frac{1}{p+1} r_i^2 \frac{h_i}{1-h_i}
\end{aligned}
$$

Combination of residual effect and leverage effect

Savings Example

```
## Compute Cook's distance
> cook <- cooks.distance(result)
> halfnorm(cook, nlab = 3, ylab="Cook's distance")
> country[c(46, 23, 49)]
[1] Zambia Japan  Libya
## Fit the model w/o Libya
> result.libya <- lm(sr ~ pop15 + pop75
    + dpi + ddpi, data=savings,
    subset=(cook < max(cook)))
```

Savings Example Continued



```
> summary(result.libya)
Coefficients:
            Estimate Std.Error t value Pr(>|t|)
Intercept 24.5247126 8.2239839   2.982  0.00465
pop15     -0.3914544 0.1579094  -2.479  0.01708
pop75     -1.2809610 1.1451679  -1.119  0.26939
dpi       -0.0003188 0.0009294  -0.343  0.73323
ddpi       0.6102784 0.2687765   2.271  0.02812
---
Residual standard error:  3.795 on 44 degrees of freedom
Multiple R-Squared: 0.3554    Adjusted R-squared: 0.2968
F-statistic: 6.066 on 4 and 44 DF     p-value: 0.0005616
```

```
> summary(result)
Coefficients:
            Estimate Std.Error t value Pr(>|t|)
Intercept 28.5666100 7.3544986   3.884 0.000334
pop15     -0.4612050 0.1446425  -3.189 0.002602
pop75     -1.6915757 1.0835862  -1.561 0.125508
dpi       -0.0003368 0.0009311  -0.362 0.719296
ddpi       0.4096998 0.1961961   2.088 0.042468
---
Residual standard error:  3.803 on 45 degrees of freedom
Multiple R-Squared: 0.3385    Adjusted R-squared: 0.2797
F-statistic: 5.756 on 4 and 45 DF     p-value: 0.0007902

## Compute changes in coefficients
> result.inf <- lm.influence(result)
> plot(result.inf$coef[,2], result.inf$coef[,3],
    xlab="Change in beta2",
    ylab="Change in beta3")
```

```
## interactive tool to identify points by clicking
> identify(result.inf$coef[, 2], result.inf$coef[, 3])
> country[c(21, 23, 49)]
[1] Ireland Japan   Libya
## Fit the model w/o Japan
> result.japan <- lm(sr ~ pop15 + pop75
    + dpi + ddpi, data=savings,
    subset=(country!="Japan"))
> summary(result.japan)
Coefficients:
            Estimate Std.Error t value Pr(>|t|)
Intercept 23.9408334 7.7840159   3.076  0.00360
pop15     -0.3679159 0.1536306  -2.395  0.02095
pop75     -0.9737939 1.1554392  -0.843  0.40390
dpi       -0.0004705 0.0009191  -0.512  0.61130
ddpi       0.3347586 0.1984449   1.687  0.09870
---
Residual standard error:  3.738 on 44 degrees of freedom
Multiple R-Squared: 0.277     Adjusted R-squared: 0.2113
F-statistic: 4.214 on 4 and 44 DF      p-value: 0.005648
```

Checking the Structure of the Model

Plot $\hat{\epsilon}$ against $\hat{y}$ and x_j, but other predictors impact the relationship. Consider

- Partial regression plots
- Partial residual plots

Isolate the effect of x_j on y

Partial Regression Plots

1. Regress y on all x except x_j, get residuals $\hat{\delta}$
2. Regress x_j on all x except x_j, get residuals $\hat{\gamma}$
3. Plot $\hat{\delta}$ against $\hat{\gamma}$

The slope is $\hat{\beta}_j$. Look for non-linearity and outliers and influential points.

Partial Residual Plots

- Plot $\hat{\epsilon} + \hat{\beta}_j x_j$ against x_j

Where does this come from?

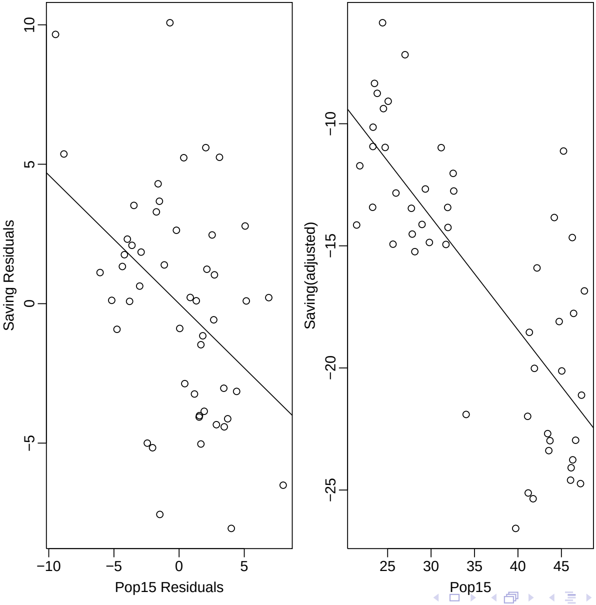
$$\begin{aligned} y - \sum_{j' \neq j} x_{j'} \hat{\beta}_{j'} &= \ldots \\ &= x_j \hat{\beta}_j + \hat{\epsilon} \end{aligned}$$

Savings Example

```
> result <- lm(sr ~ pop15 + pop75 + dpi
    + ddpi, data=savings)
## Partial regression plot
> delta <- residuals(lm(sr ~ pop75 + dpi
    + ddpi, data=savings))
> gamma <- residuals(lm(pop15 ~ pop75 + dpi
    + ddpi, data=savings))
> plot(gamma,delta, xlab="Pop15 Residuals",
    ylab="Saving Residuals")
> temp <- lm(delta ~ gamma)
> abline(reg=temp)
```

```
> coef(temp)
  (Intercept)         gamma
-7.049015e-17 -4.612050e-01
> coef(result)
  (Intercept)         pop15         pop75
28.5666100496 -0.4612050200 -1.6915756936
          dpi          ddpi
-0.0003367615  0.4096997730
## Partial residual plot
> plot(pop15, result$residuals +
    coef(result)['pop15']*pop15, xlab="Pop15",
    ylab="Savings (adjusted for pop15)")
> abline(a=0, b=coef(result)['pop15'])
```

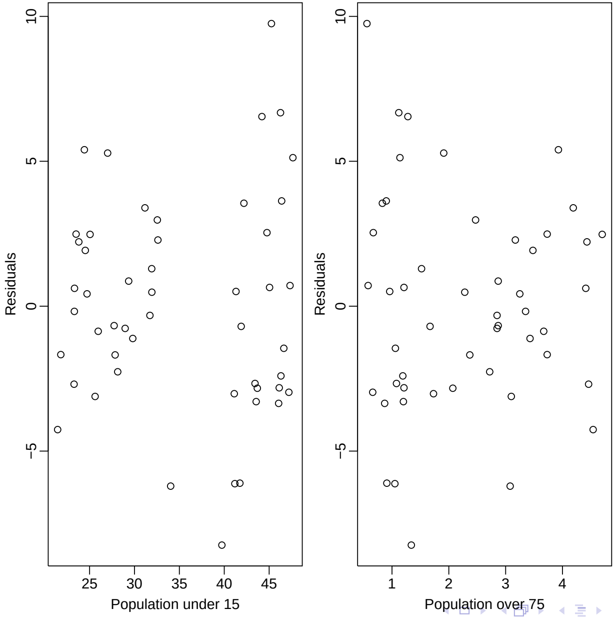
Savings Example Continued



Savings Example Continued

```
## Plot residuals vs predictors
> plot(pop15, result$residual,
    xlab="Population under 15",
    ylab="Residuals")
> plot(pop75, result$residual,
    xlab="Population over 75",
    ylab="Residuals")
```

Savings Example Continued



Savings Example Continued

```
## Two separate regressions on two groups
> temp1 <- lm(sr ~ pop15 + pop75 + dpi
    + ddpi, data=savings, subset=(pop15 > 35))
> temp2 <- lm(sr ~ pop15 + pop75 + dpi
    + ddpi, data=savings, subset=(pop15 < 35))
```

```
> summary(temp1)
Coefficients:
            Estimate Std.Error t value Pr(>|t|)
Intercept -2.4339689 21.155028  -0.115    0.910
pop15      0.2738537 0.4391910   0.624    0.541
pop75     -3.5484769 3.0332806  -1.170    0.257
dpi        0.0004208 0.0050001   0.084    0.934
ddpi       0.3954742 0.2901012   1.363    0.190
---
Residual standard error:  4.454 on 18 degrees of freedom
Multiple R-Squared: 0.156   Adjusted R-squared: -0.03185
F-statistic: 0.8302 on 4 and 18 DF       p-value: 0.5233
```

```
> summary(temp2)
Coefficients:
             Estimate Std.Error t value Pr(>|t|)
Intercept 23.9637508 8.0836079   2.964  0.00716
pop15     -0.3859519 0.1953668  -1.976  0.06089
pop75     -1.3278580 0.9260337  -1.434  0.16566
dpi       -0.0004587 0.0007237  -0.634  0.53271
ddpi       0.8843841 0.2953329   2.995  0.00668
---
Residual standard error:  2.772 on 22 degrees of freedom
Multiple R-Squared: 0.5073    Adjusted R-squared: 0.4177
F-statistic: 5.663 on 4 and 22 DF      p-value: 0.002733
```

Summary of Diagnostics

- Just fitting a model is not enough
- Graphical diagnostics are more informative but also more subjective
- Diagnostics often suggest a change in the model and then the whole process is repeated
- Time-consuming... but worth it