lab 4

Zeren Li

9/18/2019

Roadmap

- Compute se, t-stat, and confidence interval
- Heteroskedasticity

a. RVF plot: residual vs fitted
b. Lvr2 plot: residuals to detect outliers (leverage vs magnitude)

- Non-linearity
  - Log regression
  - Bivariate quadratic equation

Standard Error of Regression

(SER) is an estimator of the standard deviation of the residuals $\hat{u}_i$. As such it measures the magnitude of a typical deviation from the regression line, i.e., the magnitude of a typical residual.

$$SER = s_{\hat{u}} = \sqrt{s^2_{\hat{u}}} \quad \text{where} \quad s^2_{\hat{u}} = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-2}$$

As $TSS = ESS + SSR$

$$R^2 = 1 - \frac{SSR}{TSS}$$

Assumptions of OLS

Assumption 1: The Error Term has Conditional Mean of Zero

Assumption 2: Independently and Identically Distributed Data

Assumption 3: Large Outliers are Unlikely

The Sampling Distribution of the OLS Estimator

- As $\hat{\beta}_0$ and $\hat{\beta}_1$ are computed from a sample, estimators themselves are random variables with a probability distribution — the so-called sampling distribution of the estimators — which describes the values they could take on over different samples.
- Sampling distribution can be complicated when the sample size is small and generally changes with the number of observations, n
- When Sample is sufficiently large, by the central limit theorem the *joint* sampling distribution of the estimators is well approximated by the bivariate normal distribution

$$E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad E(\hat{\beta}_1) = \beta_1,$$

Hypothesis Tests

- Testing Hypotheses regarding regression coefficients.
- Confidence intervals for regression coefficients.
- Heteroskedasticity and Homoskedasticity.

Testing Two-Sided Hypotheses Concerning the Slope Coefficient

A general t-statistic has the form

$$t = \frac{\text{estimated value} - \text{hypothesized value}}{\text{standard error of the estimator}}.$$

Testing Two-Sided Hypotheses Concerning the Slope Coefficient

For testing the hypothesis $H_0 : \beta_1 = \beta_{1,0}$, we need to perform the following steps:

1. Compute the standard error of $\hat{\beta}_1$, $SE(\hat{\beta}_1)$

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}, \hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \hat{u_i}^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]^2}.$$

2. Compute the t-statistic

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}.$$

3. Given a two sided alternative ($H_1 : \beta_1 \neq \beta_{1,0}$) we reject at the 5% level if $|t^{act}| > 1.96$ or, equivalently, if the p-value is less than 0.05.
   Recall the definition of the p-value:

$$\begin{aligned} p\text{-value} &= \Pr_{H_0}\left[\left|\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right| > \left|\frac{\hat{\beta}_1^{act} - \beta_{1,0}}{SE(\hat{\beta}_1)}\right|\right] \\ &= \Pr_{H_0}(|t| > |t^{act}|) \\ &\approx 2 \cdot \Phi(-|t^{act}|) \end{aligned}$$

The last transformation is due to the normal approximation for large samples.

Example: Returns to Performance

```
# estimate the model
m1 <- lm(salary ~ sales, data = ceo)

summary(m1)
```

```
##
## Call:
## lm(formula = salary ~ sales, data = ceo)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -735.4 -340.2 -125.7  236.5 4474.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.364e+02  4.738e+01  15.540  < 2e-16 ***
## sales       3.669e-02  6.747e-03   5.438 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 545 on 175 degrees of freedom
## Multiple R-squared:  0.1446, Adjusted R-squared:  0.1397
## F-statistic: 29.58 on 1 and 175 DF,  p-value: 1.788e-07
```

Confidence Intervals

A 95% confidence interval for β_i has two equivalent definitions:

- The interval has a probability of 95% to contain the true value of β_i. So in 95% of all samples that could be drawn, the confidence interval will cover the true value of β_i.
- The interval is the set of values for which a hypothesis test to the level of 5% cannot be rejected.

Heteroskedasticity and Homoskedasticity

All inference made in the previous discussion relies on the assumption that the error variance does not vary as regressor values change. But this will often not be the case in empirical applications

- The error term of our regression model is homoskedastic if the variance of the conditional distribution of u_i given X_i, $Var(u_i|X_i = x)$, is constant *for all* observations in our sample:

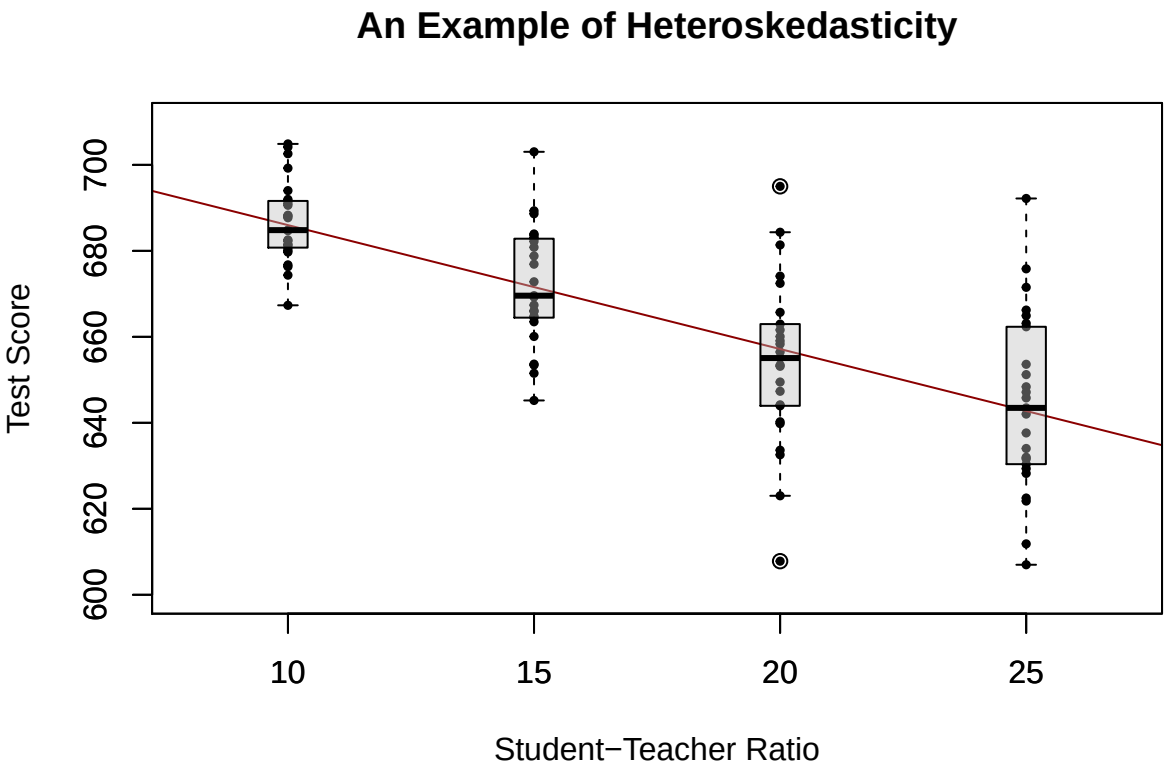
$$\text{Var}(u_i|X_i = x) = \sigma^2 \ \forall \ i = 1, \ldots, n.$$

- If instead there is dependence of the conditional variance of u_i on X_i, the error term is said to be heteroskedastic. We then write

$$\text{Var}(u_i|X_i = x) = \sigma_i^2 \ \forall \ i = 1, \ldots, n.$$

- Homoskedasticity is a *special case* of heteroskedasticity.

A better understanding of heteroskedasticity



A real-world example

Think about the economic value of education:

$$wage_i = \beta_0 + \beta_1 \cdot education_i + u_i.$$

- On average, higher educated workers earn more than workers with less education -> an upward sloping regression line.
- Also, it seems plausible that earnings of better educated workers have a higher dispersion than those of low-skilled workers
- Solid education is not a guarantee for a high salary so even highly qualified workers take on low-income jobs

A real-world example

```
# load package and attach data
library(AER)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

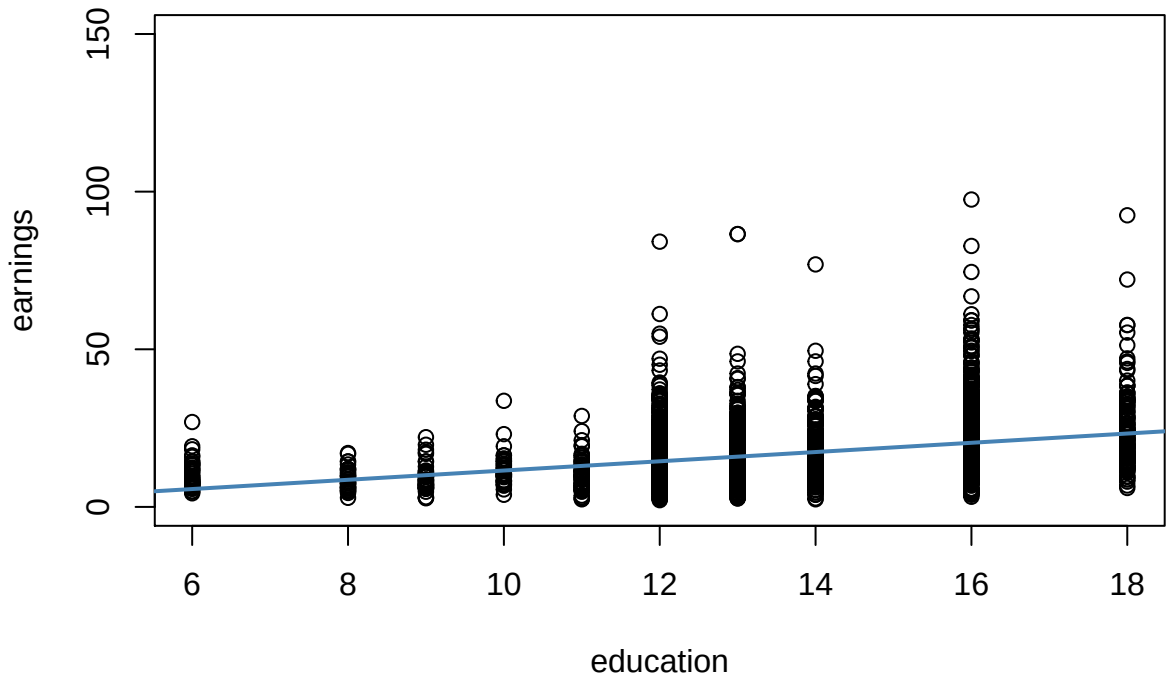
```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

## The following object is masked from 'package:openintro':
##
##     densityPlot

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival

##
## Attaching package: 'survival'

## The following object is masked from 'package:openintro':
##
##     transplant
```

```
data("CPSSWEducation")
attach(CPSSWEducation)

# get an overview
summary(CPSSWEducation)
```

```
##       age          gender        earnings        education
##  Min.   :29.0   female:1202   Min.   : 2.137   Min.   : 6.00
##  1st Qu.:29.0   male  :1748   1st Qu.:10.577   1st Qu.:12.00
##  Median :29.0                 Median :14.615   Median :13.00
##  Mean   :29.5                 Mean   :16.743   Mean   :13.55
##  3rd Qu.:30.0                 3rd Qu.:20.192   3rd Qu.:16.00
##  Max.   :30.0                 Max.   :97.500   Max.   :18.00
```

```
# estimate a simple regression model
labor_model <- lm(earnings ~ education)
```



Gauss-Markov theorem

The best (in the sense of smallest variance) linear conditionally unbiased estimator (BLUE) in this setting.

- Estimators of β_1 that are linear functions of the $Y_1, \ldots, Y_n$ and that are unbiased conditionally on the regressor $X_1, \ldots, X_n$ can be written as

$$\widetilde{\beta}_1 = \sum_{i=1}^{n} a_i Y_i$$

  where the a_i are weights that are allowed to depend on the X_i but *not* on the Y_i.

- We already know that $\widetilde{\beta}_1$ has a sampling distribution: $\widetilde{\beta}_1$ is a linear function of the Y_i which are random variables. If now

$$E(\widetilde{\beta}_1 | X_1, \ldots, X_n) = \beta_1,$$

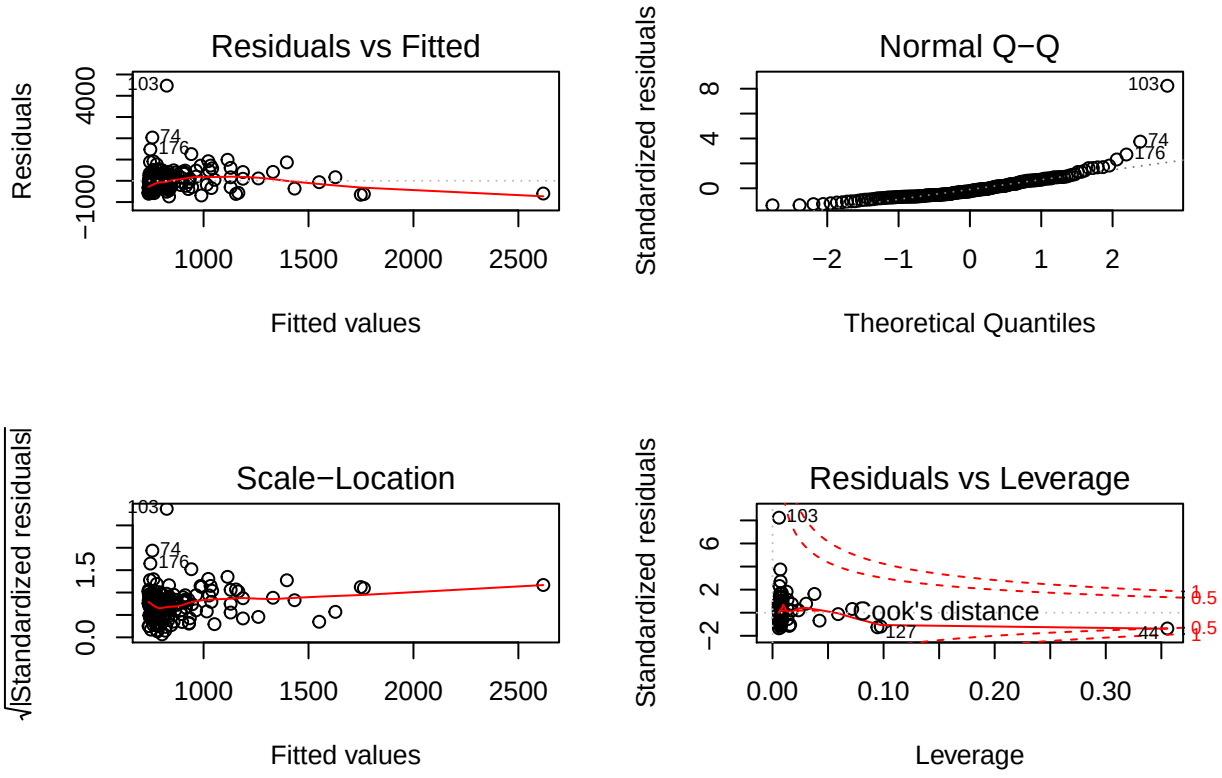
  $\widetilde{\beta}_1$ is a linear unbiased estimator of β_1, conditionally on the $X_1, \ldots, X_n$.

- We may ask if $\widetilde{\beta}_1$ is also the *best* estimator in this class, i.e., the most efficient one of all linear conditionally unbiased estimators where "most efficient" means smallest variance. The weights a_i play an important role here and it turns out that OLS uses just the right weights to have the BLUE property.

Residual Plots

1. Residuals vs Fitted: shows if residuals have non-linear patterns
2. Normal Q-Q: shows if residuals are normally distributed 3.Scale-Location: shows if residuals are spread equally along with the ranges of predictors
3. Residuals vs Leverage: helps us to find influential cases

```
par(mfrow=c(2,2))
plot(m1, ask=F)
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
103
74
176
127
44