Violating Assumptions

Jerome Dumortier

26 February 2025

Required Packages

The following packages are needed for the material presented in the slides

- car
- lmtest
- MASS
- nlme
- orcutt
- prais
- sandwich

Assumptions

Key assumptions underlying the ordinary least square (OLS) model

1. **Linear in coefficients**: Linear relationship between y and $x_1, \dots, x_k$
2. **Zero mean of the error terms**: $E(\epsilon | x_1, \dots, x_k) = 0$ and normally distributed error terms
3. **Homoscedasticity**: $Var(\epsilon_i) = \sigma^2$
4. **No autocorrelation between error terms**: $Cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$
5. **Exogeneity of independent variables**: $E(\epsilon_i | x_1, \dots, x_k)$, i.e., independent variables contain no information to predict error terms
6. **Full rank** (linear independence of all columns) of X (matrix of independent variables): Perfect multicollinearity (i.e., one independent variable being perfectly predicted from a linear combination of one or more other independent variables) leads to a rank deficiency of X

Topics Covered

Non-constant error (ϵ_i) variance, i.e., heteroscedasticity

- Testing for heteroscedasticity using the Goldfeld-Quandt Test (1965) and the Breusch-Pagan-Godfrey Test (1979)
- Correcting for heteroscedasticity by using heteroskedasticity-consistent (robust) standard errors

Multicollinearity

- Detecting multicollinearity with Variance Inflation Factors (VIF)

Autocorrelation

Overview

Homoscedasticity

$$Var(\epsilon_i) = \sigma^2$$

Heteroscedasticity

$$Var(\epsilon_i) = \sigma_i^2$$

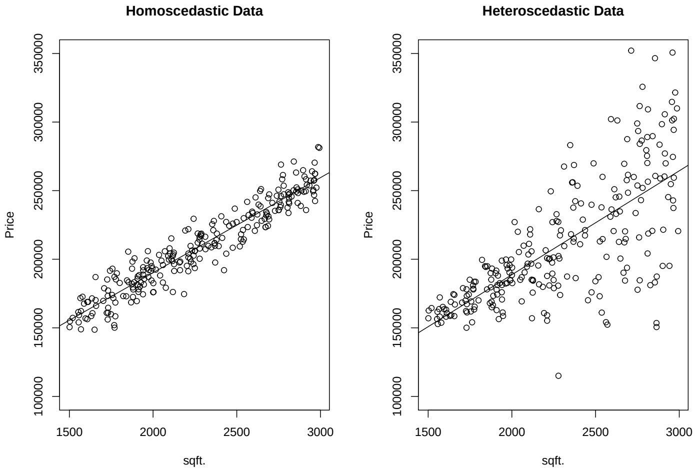
It can be shown that

$$Var(\hat{\beta}_1) = \underbrace{\frac{\sigma_i^2}{\sum x_i^2}}_{Hetero.} \neq \underbrace{\frac{\sigma^2}{\sum x_i^2}}_{Homo.}$$

Notes

- Coefficient estimates and R^2 are unaffected by heteroscedasticity
- Variance of β_1 is larger

Homoscedastic vs. Heteroscedastic Data



Examples and Effects of Heteroscedasticity I

Examples

- *Income, savings, and consumption*: People with higher incomes tend to have more variability in their savings and expenditures whereas low-income individuals spend close to their income
- *Firms and dividends*: Companies with larger profits show more variability in dividend payments
- *Education and income*: Wages may be more predictable for lower education levels while higher education degrees introduce greater variability due to differences in occupation, industry, and experience
- *House price and square footage*: Small price variations for smaller homes compared to larger homes

Examples and Effects of Heteroscedasticity II

Examples

- *Municipal budget variability and city size*: Larger cities experience greater fluctuations in budget expenditures due to the complexity and unpredictability of managing diverse public services
- *Public program effectiveness and demographics*: Policy interventions show more variable outcomes in diverse populations (e.g., in terms of socio-economics) compared to more homogeneous communities, leading to inconsistent program effectiveness

Effects of heteroscedasticity

- Requirement of homoscedasticity for t-test, F-test, and confidence intervals
- F-statistics no longer have the F-distribution
- Bottom line: Hypothesis tests on the β coefficients are no longer valid

Generalized Least Squares (GLS) I

If σ_i^2 was known:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

Dividing both sides by the known variance:

$$\frac{y_i}{\sigma_i} = \beta_0 \cdot \frac{1}{\sigma_i} + \beta_1 \cdot \frac{x_i}{\sigma_i} + \frac{\epsilon_i}{\sigma_i}$$

If $\epsilon_i^* = \epsilon_i / \sigma_i$, then it can be shown that $Var(\epsilon_i^*) = 1$, i.e., constant.

Generalized Least Squares (GLS) II

Regular OLS

$$\sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} (y_i - \beta_0 + \beta_1 \cdot x_i)^2$$

GLS with $w_i = 1/\sigma_i$

$$\sum_{i=1}^{N} w_i \cdot \epsilon_i^2 = \sum_{i=1}^{N} w_i \cdot (y_i - \beta_0 + \beta_1 \cdot x_i)^2$$

GLS: Minimization of the weighted sum of squared residuals

Generalized Least Squares (GLS) III

Implementation of GLS

1. Estimate the heteroscedasticity structure, e.g., using a White test or Breusch-Pagan test
2. Model the variance function σ_i^2, e.g., as a function of explanatory variables
3. Compute weights $w_i = 1/\hat{\sigma}_i$.
4. Transform the dependent and independent variables using those weights
5. Perform weighted least squares (WLS) regression on the transformed data

Goldfeld-Quandt Test: Steps

Steps for the Goldfeld-Quandt Test

1. Sorting observations in ascending order of an independent variable likely introducing heteroscedasticity
2. Choosing c as the number of central observations to drop resulting in sample sizes $n_1 = n_2 = (n-c)/2$
3. Running two separate regression equations
4. Compute λ with k as the number of coefficients to be estimated including the intercept
$$\lambda = \frac{RSS_2/(n_2-k)}{RSS_1/(n_1-k)}$$
5. λ follows F-distribution and a hypothesis test can be conducted

Goldfeld-Quandt Test: Manual Implementation

Setup

- Example using gqdata with *sqft* being sorted in ascending order,
- $C = 4$

```
gqdata1         = gqdata[1:20,]
gqdata2         = gqdata[31:50,]
bhat            = lm(price~sqft,data=gqdata)
bhat1           = lm(price~sqft,data=gqdata1)
bhat2           = lm(price~sqft,data=gqdata2)
sum(bhat2$residuals^2)/sum(bhat1$residuals^2)
```

```
## [1] 2.826607
```

Goldfeld-Quandt Test: R Function

```r
gqtest(bhat,fraction=10)
```

```
##
##  Goldfeld-Quandt test
##
## data:  bhat
## GQ = 2.8266, df1 = 18, df2 = 18, p-value = 0.01665
## alternative hypothesis: variance increases from segment 1 to 2
```

Breusch-Pagan-Godfrey Test: Steps

Steps for the Breusch-Pagan-Godfrey Test

1. Run a regular OLS model and obtain the residuals
2. Calculate
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \epsilon_i^2}{N}$$
3. Construct the variable $p_i = \epsilon_i^2 / \hat{\sigma}^2$
4. Run a regression as follows with x_i as the independent variables from the original regression
$$p = \alpha_0 + \alpha_1 \cdot x_1 + \alpha_2 \cdot x_2 + \ldots$$
5. Obtain the explained sum of squares (ESS) and define $\Theta = 0.5 \cdot ESS$. Then $\Theta \sim \chi^2_{m-1}$.

Or simply use `bptest(bhat)` in R

Breusch-Pagan-Godfrey Test: R Function

```r
bptest(bhat)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  bhat
## BP = 3.8751, df = 1, p-value = 0.04901
```

Robust Standard Errors: Steps

Robust standard (heteroscedasticity-consistent) errors

- Estimation of a covariance matrix (usually denoted Ω in books)

Steps in R

1. Estimation of a regular OLS model
2. Estimation of a covariance matrix using `vcovHC()` from the sandwich package
3. Applying the function `coeftest()` from the nlme package

Simultaneous execution of steps 2 and 3

Robust Standard Errors: Methods

HC0: Default heteroscedasticity-consistent (HC) standard error estimator

- Uses squared residuals without any adjustment
- Suitable for large samples

HC1: Adjusts HC0 for small sample bias by scaling the residuals

- Equivalent to HC0 multiplied by $n/(n-k)$ where k is the number of independent variables

HC2: Corrects for leverage effects in small samples

- Division of squared residuals by $1-h_i$ where $0 \leq h_i \leq 1$ is the leverage of observation i (i.e., influence of i on regression coefficients)

HC3: Additional adjustment compared to HC2 for small sample size

- Division of squared residuals by $(1-h_i)^2$

Robust Standard Errors: Implementation

```
bhat = lm(price~sqft,data=gqdata)
b1   = coeftest(bhat,vcov=vcovHC(bhat,type="HC0"))
b2   = coeftest(bhat,vcov=vcovHC(bhat,type="HC1"))
b3   = coeftest(bhat,vcov=vcovHC(bhat,type="HC2"))
b4   = coeftest(bhat,vcov=vcovHC(bhat,type="HC3"))
```

Robust Standard Errors: Implementation

```
##
## =================================================================================
##                                        Dependent variable:
##                     -------------------------------------------------------------
##                             price
##                              OLS                         coefficient
##                                                              test
##                              (1)              (2)       (3)       (4)       (5)
## ---------------------------------------------------------------------------------
## sqft                      73.911***        73.911*** 73.911*** 73.911*** 73.911***
##                            (8.062)          (6.894)   (7.036)   (7.088)   (7.289)
## ---------------------------------------------------------------------------------
## Observations                 50
## R2                          0.636
## Adjusted R2                 0.629
## Residual Std. Error  24,873.760 (df = 48)
## F Statistic         84.045*** (df = 1; 48)
## =================================================================================
## Note:                                              *p<0.1; **p<0.05; ***p<0.01
```

Problem with Multicollinearity

From the book Basic Econometrics by Gujarati:

"If multicollinearity is perfect [. . .], the regression coefficients of the X variables are indeterminate and their standard errors are infinite. If multicollinearity is less than perfect [. . .], the regression coefficients, although determinate, possess large standard errors (in relation to the coefficients themselves), which means the coefficients cannot be estimated with great precision or accuracy."

Overview

Perfect multicollinearity with λ_i representing constants that are not all zero simultaneously

$$\lambda_1 \cdot x_1 + \lambda_2 \cdot x_2 + \cdots + \lambda_k \cdot x_k = 0$$

Example

$$x_1 = \{8, 12, 15, 45\}$$

$$x_2 = \{24, 36, 15, 51\}$$

$\lambda_1 = 1$ and $\lambda_2 = -1/3$ are such that $x_1 - 1/3 \cdot x_2 = 0$. Multicollinearity refers to linear relationships and including a squared or cubed term does not represent multicollinearity

Examples

Estimation of energy consumption based on income and home size

- Likely high correlation between income and house size

Estimation of education quality (e.g., test scores, graduation rates) based on public spending (e.g., per-capita education budget, teacher salary, and number of schools)

- Correlation between education budget and teacher salary as well as education budget and number of schools

Estimation of crime based on crime prevention policies and public safety expenditures

- Likely correlation of public safety expenditures and the ability to fund crime prevention policies

Over-determined model

- Number of variables k larger than number of observations n

Indications of Multicollinearity

Signs of multicollinearity

- High R^2 but few significant variables
- Failure to reject H_0 (i.e.,$\beta_i = 0$) based on t-values but rejection of F-test (i.e., all slopes being simultaneously zero)
- High correlation among explanatory variables
- Variation of statistically significant variables between models that include different sets of independent variables

Consequences of multicollinearity

- Increase in variances of coefficients

VIF: Overview

Purpose

- Identification of possible correlation among multiple independent variables and not just two as in the case of a correlation coefficient
- Detect inflated variance based on multicollinearity

Theoretical aspects

- Existence of a VIF for each independent variable in the model

Regressing each independent variable on all other independent variables

VIF: Calculation and Interpretation

Calculation

- VIF for variable k

$$VIF_k = \frac{1}{1 - R_k^2}$$

Interpretation

- $VIF = 1$: No relationship between the variable x_k and the remaining independent variables
- $VIF > 1$: Some degree of multicollinearity
- $VIF > 4$: Warrants attention
- $VIF > 10$: Indication of serious problem

The latter two are rules of thumb

Setup

Data used: `bloodpressure`

- Patient ID (*pt*), blood pressure (*bp*), body surface area (*bsa*), and duration of hypertension (*dur*)

Correlation matrix

```
##          bp  age weight  bsa  dur pulse stress
## bp     1.00 0.66   0.95 0.87 0.29  0.72   0.16
## age    0.66 1.00   0.41 0.38 0.34  0.62   0.37
## weight 0.95 0.41   1.00 0.88 0.20  0.66   0.03
## bsa    0.87 0.38   0.88 1.00 0.13  0.46   0.02
## dur    0.29 0.34   0.20 0.13 1.00  0.40   0.31
## pulse  0.72 0.62   0.66 0.46 0.40  1.00   0.51
## stress 0.16 0.37   0.03 0.02 0.31  0.51   1.00
```

Regular OLS Regression Results

```
##
## ===========================================
##                    Dependent variable:
##                ----------------------------
##                            bp
## -------------------------------------------
## age                  0.703*** (0.050)
## weight               0.970*** (0.063)
## bsa                   3.776** (1.580)
## dur                     0.068 (0.048)
## pulse                  -0.084 (0.052)
## stress                  0.006 (0.003)
## -------------------------------------------
## Observations               20
## R2                        0.996
## F Statistic     560.641*** (df = 6; 13)
## ===========================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

VIF Calculation

Using the function `vif` from the package car:

```
vif(bhat1)
```

```
##       age   weight      bsa      dur    pulse   stress
## 1.762807 8.417035 5.328751 1.237309 4.413575 1.834845
```

VIF: Calculation of VIF for `weight`

```
##
## ========================================
##                   Dependent variable:
##              ---------------------------
##                         weight
## ----------------------------------------
## age                 -0.145 (0.206)
## bsa                21.422*** (3.465)
## dur                  0.009 (0.205)
## pulse              0.558*** (0.160)
## stress              -0.023 (0.013)
## ----------------------------------------
## Observations              20
## R2                      0.881
## F Statistic    20.768*** (df = 5; 14)
## ========================================
## Note:        *p<0.1; **p<0.05; ***p<0.01
```

VIF: Manual Calculation

The results indicate that $R^2 = 0.881$ then

$$VIF = \frac{1}{1 - 0.881} = 8.403361$$

Solution:

- Eliminate `BSA` because `weight` is easier to obtain.
- `Pulse` may be an issue as well.

VIF: Final Regression

```
##
## ============================================================
##                          Dependent variable:
##              -----------------------------------------------
##                                  bp
##                        (1)                     (2)
## ------------------------------------------------------------
## age              0.703*** (0.050)        0.732*** (0.056)
## weight           0.970*** (0.063)        1.099*** (0.038)
## bsa               3.776** (1.580)
## dur                0.068 (0.048)           0.064 (0.056)
## pulse             -0.084 (0.052)         -0.137** (0.054)
## stress             0.006 (0.003)          0.007* (0.004)
## ------------------------------------------------------------
## Observations             20                      20
## R2                      0.996                   0.994
## F Statistic  560.641*** (df = 6; 13) 502.503*** (df = 5; 14)
## ============================================================
## Note:                            *p<0.1; **p<0.05; ***p<0.01
```

Correlated Error Terms

Data available in research

- Cross-sectional: Multiple observations at same point in time
- Time series: One variable observed over time
- Pooled data: Multiple observations at different points in time (e.g., GSS)
- Panel data: Same observations at different points in time

Serial correlation versus autocorrelation

- Serial correlation: Correlation between two series
- Autocorrelation: Correlation with lagged variables

Consequences

- Unbiased OLS coefficients but no minimum variance since $E(\epsilon_i \epsilon_j) \neq 0$

Autocorrelation unlikely for cross-sectional data except for spatial auto-correlation

Causes of Autocorrelation

Multiple reasons for autocorrelation

1. Omitted variables
2. Incorrect function form

Lagged Dependent Variable as an Explanatory Variable When the dependent variable from previous periods is used as an explanatory variable, it can induce autocorrelation if there are other omitted dynamic effects.

Serial Correlation in Explanatory Variables If the independent variables themselves exhibit serial correlation, their effect can propagate into the residuals.

Measurement Errors Errors in measuring variables, particularly when they are persistent over time, can lead to autocorrelated errors.

Incorrect Specification of Dynamics In time series models, failing to account for dynamic relationships (e.g., failing to include lagged explanatory variables when necessary) can result in autocorrelation.

Omitted Variables

Correct equation

$$Q_{beef,t} = \beta_0 + \beta_1 \cdot P_{beef,t} + \beta_2 \cdot P_{income,t} + \beta_3 \cdot P_{pork,t} + \epsilon_t$$

Estimated equation

$$Q_{beef,t} = \beta_0 + \beta_1 \cdot P_{beef,t} + \beta_2 \cdot P_{income,t} + \upsilon_t$$

Systematic pattern in the error term υ_t

$$\upsilon_t = \beta_3 \cdot P_{pork,t} + \epsilon_t$$

Relevant variable(s) not included with persistent effects over time

Incorrect Functional Form: Setup

Correct equation

$$y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot x_i^2 + \epsilon_i$$

Estimated equation

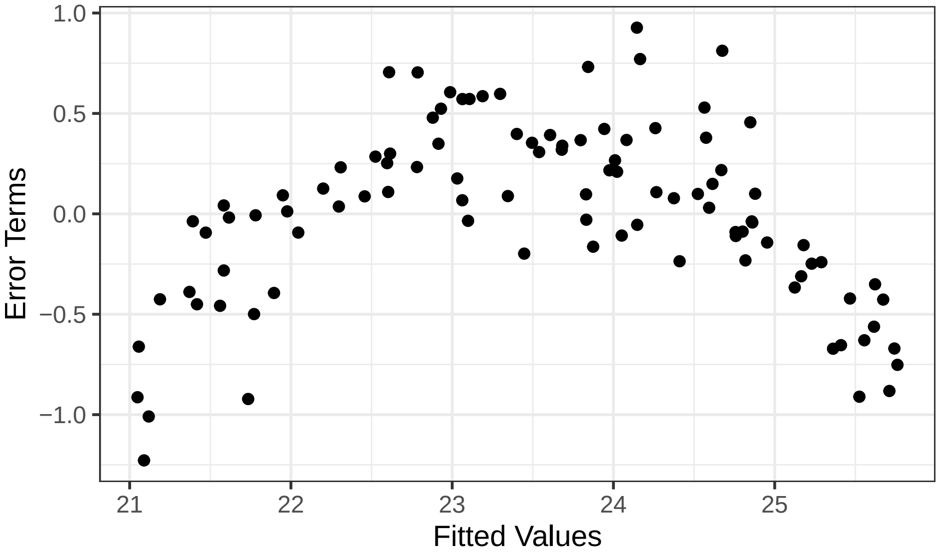
$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

Simulated data

- Coefficients: $\beta_0 = 5$, $\beta_1 = 0.4$ $\beta_2 = -0.002$
- Error term: $\epsilon \sim \mathcal{N}(0, 0.25^2)$

```
income     = runif(100,50,100)
error      = rnorm(100,0,0.25)
foodcons   = 5+0.4*income-0.002*income^2+error
bhat       = lm(foodcons~income)
```

Incorrect Functional Form: Plot



Cobbweb

Cobweb phenomenon (e.g., production decision before prices are observed such as in agriculture):

$$supply_t = \beta_0 + \beta_1 \cdot p_{t-1}$$

Lags:

bhat = lm(qpork~pbeef+pchicken+ppork+rdi,data=meatdemand) summary(bhat)

Cobweb Simulation

```r
set.seed(123)

# Simulation parameters
T <- 50        # Number of periods
alpha <- 0.8 # Price elasticity of supply (how much supply reacts to
beta <- -0.9 # Price elasticity of demand (negative: higher supply lo
p_eq <- 10    # Equilibrium price
q_eq <- 100   # Equilibrium quantity

# Initialize vectors
price <- numeric(T)
supply <- numeric(T)

# Initial values
price[1] <- 12  # Initial price
supply[1] <- q_eq + alpha * (price[1] - p_eq)  # Initial supply based
```

Autoregression

Lagged dependent variable as explanatory variable

$$consumption_t = \beta_0 + \beta_1 \cdot income_t + \beta_3 \cdot consumption_{t-1} + \epsilon_t$$

Notes:

First-Order Autoregressive Scheme

Consider the model:

$$y_t = \beta_0 + \beta_1 \cdot x_t + \upsilon_t$$

Assume the following form of υ:

$$\upsilon_t = \rho \cdot \upsilon_{t-1} + \epsilon_t$$

This last equation is called a first-order autoregressive AR(1) scheme. An AR(2) would be written as

$$\upsilon_t = \rho_1 \cdot \upsilon_{t-1} + \rho_2 \cdot \upsilon_{t-2} + \epsilon_t$$

AR(1): Numerical Example

Consider the model:

$$y_t = 10 + 2 \cdot x_t + \upsilon_t$$

Assume the following form of υ:

$$\upsilon_t = 0.75 \cdot \upsilon_{t-1} + \epsilon_t$$

Procedure

- Simulate the above model 100 times assuming $\epsilon \sim N(0,1)$
- Compare variance of coefficients under different two different methods: (1) OLS and (2) Cochrane-Orcutt

```
# https://onlinecourses.science.psu.edu/stat510/node/72
library(orcutt)
simulations          = 100
nobs                 = 50
beta0                = 10
beta1                = 2
```

Detecting Autocorrelation

Durbin-Watson d test

- Key assumption: First-order autoregressive error term, i.e., AR(1)

Breusch-Godfrey test

- Higher-order autoregressive error terms, e.g., AR(1), AR(2), AR(3)

Durbin Watson d Test

Test statistic:

$$d = \frac{\sum_{t=2}^{N}(e_t - e_{t-1})^2}{\sum_{t=1}^{N} e_t^2}$$

Assumptions

- AR(1) process, i.e., $\upsilon_t = \rho \cdot \upsilon_{t-1} + \epsilon_t$
- No lagged independent variables

Original papers derive lower (d_L) and upper (d_U) bounds, i.e., critical values, that depend on N and k only.

- $d \approx 2 \cdot (1 - \rho)$ and since $-1 \leq \rho \leq 1$, we have $0 \leq d \leq 4$.

Rule of thumb indicates that $d = 2$ signals no problems.

Breusch-Godfrey Test

Consider the following model $y_t = \beta_0 + \beta_1 x_t + \upsilon_t$ with the following error term structure:

$$\upsilon_t = \rho_1 \cdot \upsilon_{t-1} + \rho_2 \cdot \upsilon_{t-2} + \cdots + \rho_p \cdot \upsilon_{t-p} + \epsilon_t$$

The null hypothesis for the test is expressed as follows:

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_p = 0$$

When the following regression is executed:

$$\hat{\upsilon}_t = \alpha_0 + \alpha_1 x_t + \hat{\rho}_1 \hat{\upsilon}_{t-1} + \hat{\rho}_2 \hat{\upsilon}_{t-2} + \cdots + \hat{\rho}_p \hat{\upsilon}_{t-p} + \epsilon_t$$

Then

$$(n-p) \cdot R^2 \sim \chi^2_p$$

Pork Demand: Regular Estimation

```
##
## Call:
## lm(formula = qpork ~ pbeef + pchicken + ppork + rdi, data = meatdemand)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8025 -1.1379 -0.3808  1.1136  2.9714
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 67.2921311  9.7977901   6.868 6.58e-08 ***
## pbeef        0.0236448  0.0061255   3.860 0.000483 ***
## pchicken     0.0195407  0.0246975   0.791 0.434311
## ppork       -0.0652253  0.0144532  -4.513 7.29e-05 ***
## rdi         -0.0002451  0.0001122  -2.186 0.035817 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Pork Demand: Autocorrelation Tests

Durbin-Watson test

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.6128715     0.6606634       0
##  Alternative hypothesis: rho != 0
```

Breusch-Godfrey test

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  bhat
## LM test = 15.904, df = 1, p-value = 6.665e-05
```

Pork Demand: ρ-Estimation

```r
summary(lm(bhat$residuals~Hmisc::Lag(bhat$residuals)))
```

```
##
## Call:
## lm(formula = bhat$residuals ~ Hmisc::Lag(bhat$residuals))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0571 -1.0478 -0.0517  0.8261  3.2781
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -0.0265     0.1923  -0.138    0.891
## Hmisc::Lag(bhat$residuals)   0.6452     0.1260   5.122 1.04e-05 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Correction for Autocorrelation

Newey-West Standard Errors (HAC)

```
bhatHAC = coeftest(bhat,vcov=NeweyWest(bhat,lag=1))
```

Generalized Least Squares (GLS)

```
bhatGLS = gls(qpork~pbeef+pchicken+ppork+rdi,
              data=meatdemand,
              correlation=corAR1(form=~year))
```

Cochrane-Orcutt Estimation

```
bhatCO             = cochrane.orcutt(bhat)
bhatCO$rho
```

```
## [1] 0.6895679
```

Prais-Winsten Estimation

Results

```
stargazer(bhat,bhatHAC,bhatCO,bhatGLS,type="text",no.space=TRUE)
```

```
##
## ==========================================================
##                       Dependent variable:
##              ---------------------------------------------
##                     qpork                           qp
##                     OLS        coefficient    OLS
##                                   test                  l
##                     (1)           (2)         (3)
## ----------------------------------------------------------
## pbeef            0.024***       0.024**      0.017*
##                  (0.006)        (0.011)      (0.010)
## pchicken          0.020          0.020        0.010
##                  (0.025)        (0.013)      (0.020)
## ppork           -0.065***      -0.065***    -0.072***
```

Wages and Productivity in the United States 1959-1998

Consider the data in `business.csv` and do the following:

- Plot the data in a scatter plot
- Run the regression in level form as well as log format
- Plot the diagnostic plots.
- Run the Durbin-Watson test and the Breusch-Godfrey test. What do you conclude?
- Run the regression by (1) including a trend variable and (2) a squared term but no trend.

Other Issues and Problems with Data

More serious problems than heteroscedasticity:

- Functional form misspecification
- Measurement error
- Missing data, non-random samples, and outliers

Missing Data and Non-Random Samples

Consequences and remedies

- Standard regression model is not possible with missing values
- All statistical software packages ignore missing data

Missing data is a minor problem if it is due to random error. Missing data can be problematic if it is systematically missing

- Missing education data for people with lower education
- Missing IQ scores from people with higher IQ's

Examples of exogenous sample selection or sample selection based on the independent variable