Regression diagnostics & remedial techniques for violating basic assumptions
Geog 210B – Winter 2018

Notes from last time (matrix and deviations from the mean representation)

Properties of Linear Regression

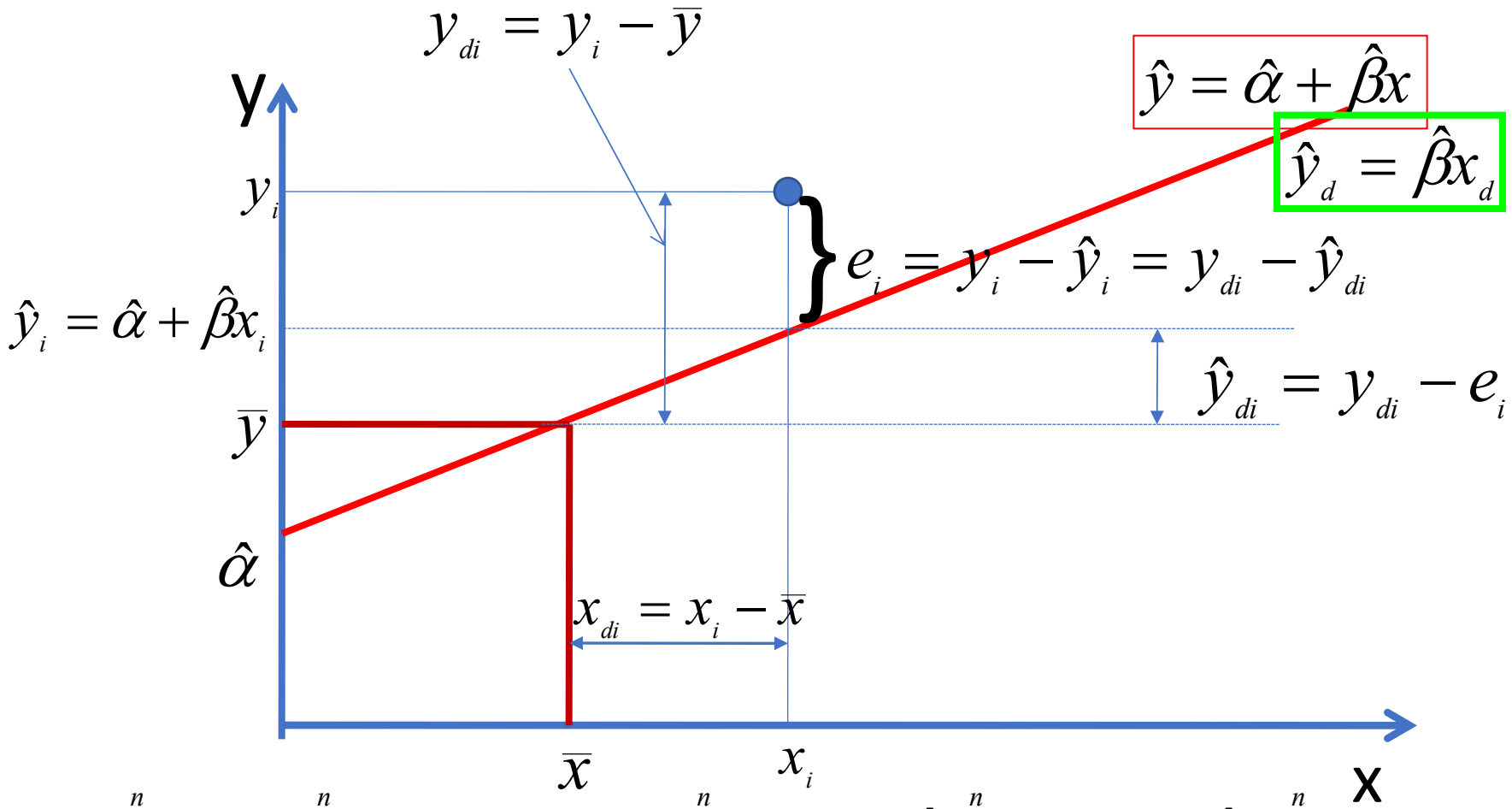
- The regression line passes through the point of x and y means
- The residuals have zero covariance with the sample x values

$$Cov(x\ e\ ) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(e_i - \bar{e}) = 0$$

- The residuals have zero covariance with the y hat values

$$Cov(\hat{y}\ e\ ) = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(e_i - \bar{e}) = 0$$

- The regression coefficients can be computed using deviations from the mean (next slide)
- The total variation of Y can be decomposed into explained and unexplained variation that can give us easy to understand “goodness of fit” statistics



$$\sum_{i=1}^{n} e^2 = \sum_{i=1}^{n} (y_{di} - \hat{y}_{di})^2 = \sum_{i=1}^{n} y_{di}^2 - 2\hat{\beta}\sum_{i=1}^{n} x_{di} y_{di} + \hat{\beta}^2 \sum_{i=1}^{n} x_{di}^2$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}} = 0 \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} x_{di} y_{di}}{\sum_{i=1}^{n} x_{di}^2}$$

Another way to think about the meaning/nature of a slope

Decomposition of sum of squares

$$\hat{y}_{di} = y_{di} - e_i \quad => \quad y_{di} = \hat{y}_{di} + e_i = \hat{\beta} x_{di} + e_i$$

Take the square and sum over all observations n

$$(y_{di})^2 = (\hat{y}_{di} + e_i)^2 = (\hat{\beta} x_{di} + e_i)^2$$

$$\sum_{i=1}^{n} (y_{di})^2 = \sum_{i=1}^{n} (\hat{y}_{di} + e_i)^2 = \sum_{i=1}^{n} (\hat{\beta} x_{di} + e_i)^2$$

$$\sum_{i=1}^{n} (y_{di})^2 = \sum_{i=1}^{n} (\hat{y}_{di})^2 + \sum_{i=1}^{n} (e_i)^2 + 2\sum_{i=1}^{n} (\hat{y}_{di} e_i)$$

$$\sum_{i=1}^{n} (\hat{y}_{di} e_i) = ?$$

This looks like a covariance

Decomposition of sum of squares

$$Cov(\hat{y}\ e\ ) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(e_i - \bar{e}) = 0$$

$$= y_{di} \qquad = e_i \qquad because\ \bar{e} = 0$$

$$Cov(\hat{y}\ e\ ) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_{di})(e_i) = 0$$

$$\sum_{i=1}^{n}(y_{di})^2 = \sum_{i=1}^{n}(\hat{y}_{di})^2 + \sum_{i=1}^{n}(e_i)^2 + 2\sum_{i=1}^{n}(\hat{y}_{di}e_i)$$

$$\sum_{i=1}^{n}(y_{di})^2 = \sum_{i=1}^{n}(\hat{y}_{di})^2 + \sum_{i=1}^{n}(e_i)^2$$

Decomposition of sum of squares

$$\sum_{i=1}^{n}(y_{di})^2 = \sum_{i=1}^{n}(\hat{y}_{di})^2 + \sum_{i=1}^{n}(e_i)^2$$

$$\sum_{i=1}^{n}(y_{di})^2 = TSS$$

Total sum of squares in the dependent variable measured about its mean

$$\sum_{i=1}^{n}(\hat{y}_{di})^2 = ESS$$

EXPLAINED sum of squares (also called Regression sum of squares

$$\sum_{i=1}^{n}(e_i)^2 = RSS$$

RESIDUAL sum of squares also called unexplained sum of squares

We like to work with proportions: ESS/TSS

$$\mathrm{var}(x) = s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x_{di})^2}{n}$$

Similar for y

The Pearson correlation coefficient:

$$r = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)}\sqrt{\mathrm{var}(y)}} = \frac{1/n\sum_{i=1}^{n} x_{di}y_{di}}{\sqrt{1/n\sum_{i=1}^{n} x_{di}^2}\sqrt{1/n\sum_{i=1}^{n} y_{di}^2}}$$

$$r = \frac{\sum_{i=1}^{n} x_{di}y_{di}}{n\sqrt{1/n\sum_{i=1}^{n} x_{di}^2}\sqrt{1/n\sum_{i=1}^{n} y_{di}^2}} = \left(\frac{\sum_{i=1}^{n} x_{di}y_{di}}{\sum_{i=1}^{n} x_{di}^2}\right)\frac{\sqrt{\sum_{i=1}^{n} x_{di}^2}}{\sqrt{\sum_{i=1}^{n} y_{di}^2}}$$

$$\mathrm{Re}member \quad that \quad \hat{\beta} = \frac{\sum_{i=1}^{n} x_{di}y_{di}}{\sum_{i=1}^{n} x_{di}^2}$$

Substitute for beta hat above

Correlation, β hat, and r square

$$r = \frac{\sum_{i=1}^{n} x_{di} y_{di}}{n\sqrt{1/n\sum_{i=1}^{n} x_{di}^2}\sqrt{1/n\sum_{i=1}^{n} y_{di}^2}} = \hat{\beta}\frac{\sqrt{\sum_{i=1}^{n} x_{di}^2}}{\sqrt{\sum_{i=1}^{n} y_{di}^2}} = \hat{\beta}\frac{s_x}{s_y}$$

$$r^2 = \frac{(\sum_{i=1}^{n} x_{di} y_{di})^2}{(\sqrt{\sum_{i=1}^{n} x_{di}^2}\sqrt{\sum_{i=1}^{n} y_{di}^2})^2} = \hat{\beta}\frac{\sum_{i=1}^{n} x_{di} y_{di}}{\sum_{i=1}^{n} y_{di}^2} = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

Note: $$ESS = \sum_{i=1}^{n} (\hat{y}_{di})^2 = \sum_{i=1}^{n} (\hat{\beta} x_{di})^2 = \hat{\beta}^2 \sum_{i=1}^{n} (x_{di})^2 = \hat{\beta}\frac{\sum_{i=1}^{n} x_{di} y_{di}}{\sum_{i=1}^{n} x_{di}^2}\sum_{i=1}^{n} x_{di}^2 = \hat{\beta}\sum_{i=1}^{n} x_{di} y_{di}$$

```
> summary(HHPMT.lm)

Call:
lm(formula = TotDist ~ HHSIZ, data = SmallHHfile)

Residuals:
   Min     1Q Median     3Q    Max
-186.1  -49.1  -26.5   11.4 5717.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.2118     1.1817   10.33   <2e-16 ***
HHSIZ        21.7305     0.4053   53.62   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 114.7 on 42429 degrees of freedom
Multiple R-squared:  0.06345,   Adjusted R-squared:  0.06343
F-statistic:  2875 on 1 and 42429 DF,  p-value: < 2.2e-16

> anova(HHPMT.lm) # anova table
Analysis of Variance Table

Response: TotDist
              Df    Sum Sq  Mean Sq F value    Pr(>F)
ESS HHSIZ      1  37815166 37815166  2874.6 < 2.2e-16 ***
RSS Residuals 42429 558154359    13155
TSS = ESS+RSS
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Some Notes on R output

- Mean Squares = the Sum of Squares divided by their respective Degrees of Freedom

- Adjusted R square. As Xs are added to the model, each X will explain some of the variance in the dependent variable Y. You can add Xs to the model and continue to improve the ability of the model to explain the dependent variable, but this increase in R-square would be simply due to chance variation (the coefficients will not be significantly different than zero but you get the illusion of better fit). The adjusted R-square attempts to yield a more honest value to estimate the R-squared for the population by penalizing models with many coefficients.

- Adjusted R-squared = 1 - ((1-R-sq)(N-1 / N - k - 1)). From this formula, you can see that when the number of observations is small and the number of Xs is large, there will be a much greater difference between R-square and adjusted R-square (because the ratio of (N-1 / N - k - 1) will be much less than 1.

- When the number of observations is very large compared to the number of Xs, the value of R-square and adjusted R-square will be much closer because the ratio of (N-1)/(N-k-1) is closer to 1.

Hypotheses Testing in CLR

Testing for significance of regression coefficients

$$\text{Remember that} \quad \hat{\beta} = \frac{\sum_{i=1}^{n} x_{di} y_{di}}{\sum_{i=1}^{n} x_{di}^2} \qquad \text{let's write } w_i = \frac{x_{di}}{\sum_{i=1}^{n} x_{di}^2}$$

It can be shown that:

$$\hat{\beta} = \sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n} w_i (\alpha + \beta x_i + \varepsilon_i)$$

Simplifying and taking the expectation:

$$E(\hat{\beta}) = \beta$$

Least squares = unbiased estimate of beta

$$\text{var}(\hat{\beta}) = E[(\hat{\beta} - \beta)^2] = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n} x_{di}^2}$$

Hypotheses Testing in CLR

Similar considerations for the intercept

$$E(\hat{\alpha}) = \alpha, \operatorname{var}(\hat{\alpha}) = \sigma_{\varepsilon}^{2}\left[\frac{1}{n} + \frac{\overline{x}^{2}}{\sum_{i=1}^{n} x_{di}}\right]$$

$$\operatorname{cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma_{\varepsilon}^{2}\overline{x}}{\sum_{i=1}^{n} x_{di}^{2}}$$

Think of circumstances when the covariance can be zero

Considering that ε is assumed normally distributed

$$\hat{\beta} \sim N(\beta, \sigma_{\varepsilon}^2 / \sum_{i=1}^{n} x_{di}^2)$$

We do not have the population value of sigma (σ) but we will use its estimate s

$$\frac{\hat{\beta} - \beta}{\frac{s}{\sqrt{\sum_{i=1}^{n} x_{di}^2}}} \sim t(n-2)$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{(n-2)}}$$

$$H_0 : \beta = \beta_0$$
$$H_a : \beta \neq \beta_0$$

Reject the null at the 95% confidence if

$$\left| \frac{\hat{\beta} - \beta_0}{s / \sqrt{\sum_{i=1}^{n} x_{di}^2}} \right| > t_{0.025}(n-2)$$

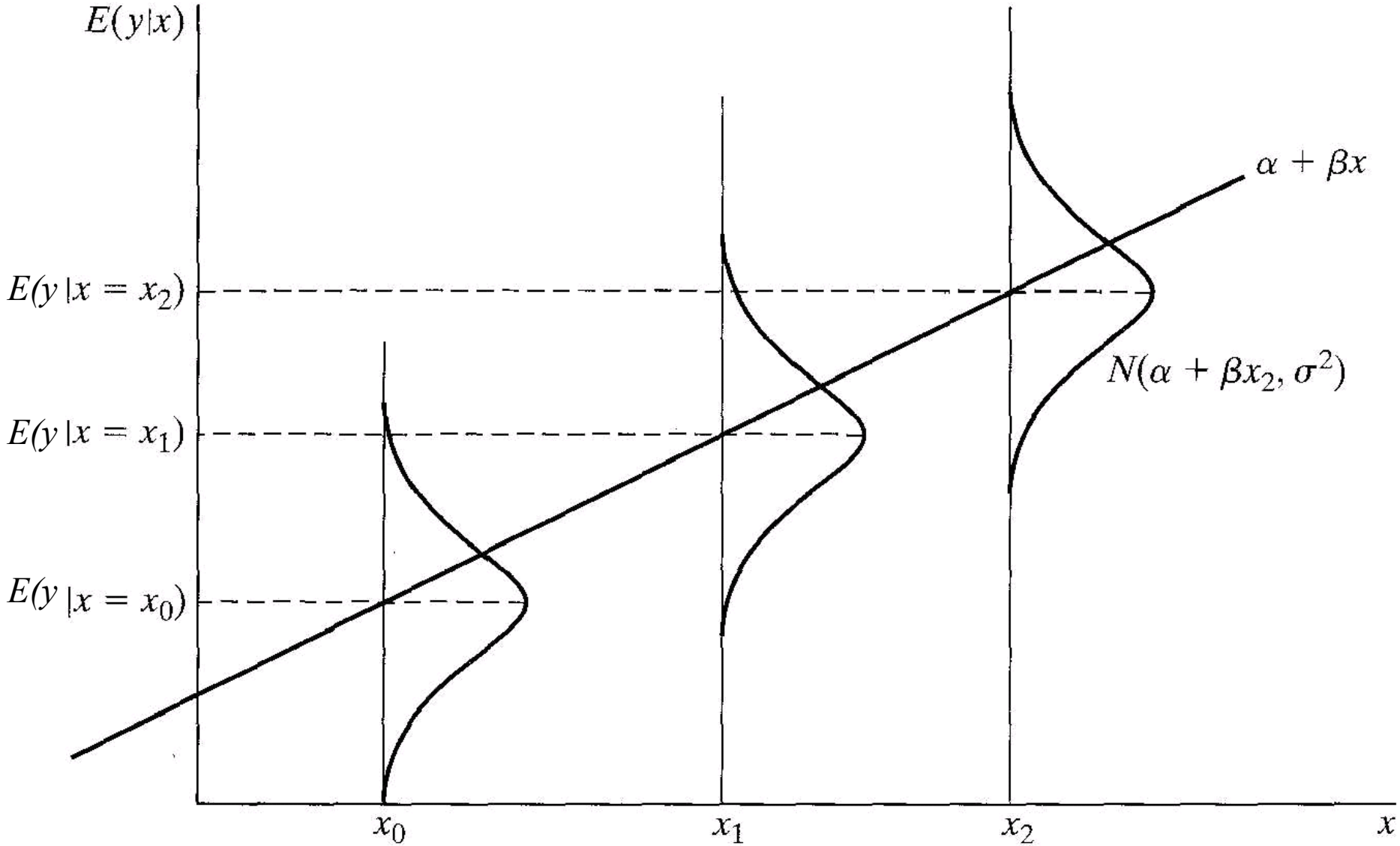
=0

We can also compute the following:

F statistic

$$F = \frac{ESS/1}{RSS/(n-2)} \sim F(1, n-2)$$

$$F = \frac{ESS/1}{RSS/(n-2)} > F_{0.95}(1, n-2)$$



Keep in mind always this representation of a linear model

Many x variables linear regression model

The dependent variable we try to explain

Random variable = error term

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

Intercept = constant to be estimated

Many "Slopes" = "sensitivity" of y to the values of each x_k

k = number of explanatory variables

Using an estimation method we try to find estimates of the parameter values for the intercept, slope, and variance of the error term using the sample data

$$\begin{aligned} E(y_i) &= E(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i) \\ &= E(\alpha) + E(\beta_1 x_{1i}) + E(\beta_2 x_{2i}) + \ldots + E(\beta_k x_{ki}) + E(\varepsilon_i) \\ &= \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} \end{aligned}$$

We estimate the parameters using the least (minimum) squares concept. The squared residuals:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_{2i} x_{2i} - \ldots - \hat{\beta}_k x_{ki})^2$$

Minimization is done by differentiation (Derivatives set to zero).

The solution in Matrix notation is:

$$\hat{\beta} = (X'X)^{-1} X'Y$$

What dimensions does this have?

Linear Model Assumptions

1. **The model is linear in the variables X and the disturbance ε**
2. **The random disturbance is centered at zero:**

$$E[\varepsilon] = \begin{bmatrix} E[\varepsilon_1] \\ E[\varepsilon_2] \\ \dots \\ E[\varepsilon_n] \end{bmatrix} = 0 \quad \text{implies} \quad E[Y] = X\beta$$

Linear Model Assumptions

3. Homoskedastic disturbances:

$$E[\varepsilon\varepsilon'] = \sigma^2 I$$

This is the variance covariance matrix of the disturbances. It is an n by n matrix with n the number of observations.

- **The disturbances are independently (covariance is zero) identically (variance is always the same) distributed – written as:**

$$E[\varepsilon\varepsilon'] = \begin{bmatrix} E[\varepsilon_1\varepsilon_1]\ E[\varepsilon_1\varepsilon_2] \dots E[\varepsilon_1\varepsilon_n] \\ E[\varepsilon_2\varepsilon_1]\ E[\varepsilon_2\varepsilon_2] \dots E[\varepsilon_2\varepsilon_n] \\ \dots \\ E[\varepsilon_n\varepsilon_1]\ E[\varepsilon_n\varepsilon_2] \dots E[\varepsilon_n\varepsilon_n] \end{bmatrix} = \begin{bmatrix} \sigma^2\ 0 \dots 0 \\ 0\ \sigma^2 \dots 0 \\ \dots \\ 0\ 0 \dots \sigma^2 \end{bmatrix}$$

This is called **homoskedasticity** of the error terms

Linear Model Assumptions

4. A reasonable assumption, but not necessary, is that the disturbances are also Normally distributed

$$\varepsilon \sim N[0, \sigma^2 I]$$

Linear Model Assumptions

5. There are no exact linear relationships among the variables

a) (full)Rank (X) = K = number of independent variables. That is the number of independent columns in X is K

b) Have at least K observations

In econometrics literature these five properties are the Gauss-Markov theorem assumptions

The β estimates using Least Squares are:

$$\hat{\beta} = (X'X)^{-1} X'Y$$

$$Y = X\beta + \varepsilon$$

Combine these equations to get:

$$\hat{\beta} = \beta + (X'X)^{-1} X'\varepsilon$$

When X is nonstochastic and E(X'ε) = 0 then the estimate of β is unbiased:

$$E(\hat{\beta}) = \beta$$

The variance of the parameter estimates is:

$$Var[\hat{\beta}] = \sigma^2 (X'X)^{-1}$$

The parameter estimates are a linear function of the disturbances ε and by virtue of assumption 4 we can say that:

$$\hat{\beta} \sim N[\beta, \sigma^2 (X'X)^{-1}]$$

multivariate normally distributed.

Note 1: Ideally we could use the standard normal distribution to test hypotheses regarding the estimates.

Note 2: Usually σ is not known and we need to estimate it.

An unbiased estimate of σ can be obtained from the residuals ei:

$$s^2 = \frac{e'e}{n - K}$$

Compare this with the 2 variable model

then the var-covariance matrix can be computed as (called the square of the standard error of the regression):

$$\hat{V}ar[\,\hat{\beta}\,] = s^2 (X'X)^{-1}$$

since we are using s instead of σ we cannot use the standard normal test and we use the t-test: T-statistic = estimate of β/(standard error of β).

Goodness of fit (how well our model replicates the data we use) is checked using indicators.

The most popular is called coefficient of determination or R-squared

Total Variation in the y variable is:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Total Sum of Squares (SST) = Regression Sum of Squares (SSR*) + Error Sum of Squares (SSE)

This is the same as:

Total Sum of Squares (TSS) = Explained Sum of Squares (ESS) + Residual Sum of Squares (RSS)

Regression Sum of Squares = variation we capture with xs and bs

*Note I am using different ss etc her because in literature you will find them both ways

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

Percent of variation we explain by exercising our model

Takes values between 0 and 1

Usually with large samples = lower values (harder to explain variation)

Small sample = higher values

Terminology

- Dependent variable
- Independent variables
- Explanatory variable
- Disturbance
- Random error term
- Coefficients
- Parameters to estimate
- Significance
- Goodness-of-fit measure

Some more terminology

- Autocorrelation
- Covariate (the x variables in regression)
- Deterministic relationship
- Disturbance (the error term epsilon)
- Exogeneity
- Explained variable
- Homoscedasticity
- Loglinear model
- Nonstochastic regressors
- Regressand (the dependent variable y)
- Regressor (the independent variables xs)
- Spherical disturbances (only same variance - no covariance)

Limitations due to assumptions

- **(1) Linearity in x and ε**
  - **Dummy variables and transformation of xs may not be enough**

- **(2) Do we know all relevant variables? (specification)**
  - **Do experiments with exclusion-inclusion of variables and use the theory**
- **(3) Prior notions regarding β**
  - **Incorporation of "prior" information can be done. Example: constrained least squares.**
- **(4) Observational errors**
  - **Use of a proxy for Y or X**
  - **Measurement errors in Y or X**
  - **The problem is more severe when the errors are for the Xs.**

Limitations due to assumptions

- **(5) Aggregation**
  - **Need to have information regarding aggregation used to obtain Xs. Spatial issues are particularly thorny.**
- **(6) X-non stochastic**
  - **Lagged dependent variable?**
  - **Errors in the Xs? Instrumental relationships?**
- **(7) Simultaneous equations**
  - **Interdependence among many ys**

- **(8) Heteroskedastic disturbances**
  - **Unequal variance**
- **(9) Correlated disturbances**
  - **Covariance different than zero**

Terminology: the multiple regression model that violates the classical assumptions of homoskedasticity and lack of correlation is called the generalized regression model

In general instead of an identify matrix for the variance covariance matrix of the disturbances we have Ω.

$$E[\varepsilon\varepsilon'] = \sigma^2\Omega$$

Heteroskedasticity: When the scale of the dependent variable and the model's explanatory power varies from observation to observation.

Example: High variability in trip making at high levels of income (stereotype: older people are richer but they make less trips and high income people tend to make more trips)

Heteroskedastic Disturbances

$$E[\varepsilon\varepsilon'] = \begin{bmatrix} E[\varepsilon_1\varepsilon_1] & E[\varepsilon_1\varepsilon_2] & \dots & E[\varepsilon_1\varepsilon_n] \\ E[\varepsilon_2\varepsilon_1] & E[\varepsilon_2\varepsilon_2] & \dots & E[\varepsilon_2\varepsilon_n] \\ & & \dots & \\ E[\varepsilon_n\varepsilon_1] & E[\varepsilon_n\varepsilon_2] & \dots & E[\varepsilon_n\varepsilon_n] \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ & & \dots & \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Autocorrelation: When the data contain "memory" effects.

Example: The y values of one row depend on the y values of another row (maybe they are households that live by or they belong to the same social network)

In spatial analysis we will see how we build neighborhoods

$$E[\varepsilon\varepsilon'] = \begin{bmatrix} E[\varepsilon_1\varepsilon_1]\ E[\varepsilon_1\varepsilon_2] \ldots E[\varepsilon_1\varepsilon_n] \\ E[\varepsilon_2\varepsilon_1]\ E[\varepsilon_2\varepsilon_2] \ldots E[\varepsilon_2\varepsilon_n] \\ \ldots \\ E[\varepsilon_n\varepsilon_1]\ E[\varepsilon_n\varepsilon_2] \ldots E[\varepsilon_n\varepsilon_n] \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 & 0 \ldots & \rho_{n-1} \\ \rho_1 & 1 & \ldots & \rho_{n-2} \\ & & \ldots & \\ \rho_{n-1} & \rho_{n-2} & \ldots & 1 \end{bmatrix}$$

The distribution of the estimator for beta is different now:

$$\hat{\beta} \sim N[\beta, \sigma^2 (X'X)^{-1} (X'\Omega X)(X'X)^{-1}]$$

If instead of the more complex equation for the variance covariance matrix we use the usual least squares estimate σ^2 (X'X)-1 we would get biased standard errors for the coefficient estimates and then the t-tests would be misleading.

Other problems may arise with the ordinary least squares depending on the particular form of Ω.

In general: Ordinary least squares, when $\Omega \neq I$, produces unbiased, consistent, and asymptotically normally distributed estimates.

The estimates are not efficient (i.e., minimum variance) and the usual inference procedures (t-tests etc) are not valid.

Usually we don't know the elements in Ω.

One way is to get an estimate of the Var-covariance matrix from Ordinary least squares and then produce weights for each observation.

Then, estimate a regression model on the weighted observations.

Testing the Overall Significance of the Multiple Regression Model

- Is using the regression equation to predict *y* better than using the mean of *y* ?

The Global *F*-Test

I. H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

The mean of *y* is doing as good a job at predicting the actual values of *y* as the regression equation.

H_1: At least one β_i does not equal 0.

The regression model is doing a better job of predicting actual values of *y* than using the mean of *y*.

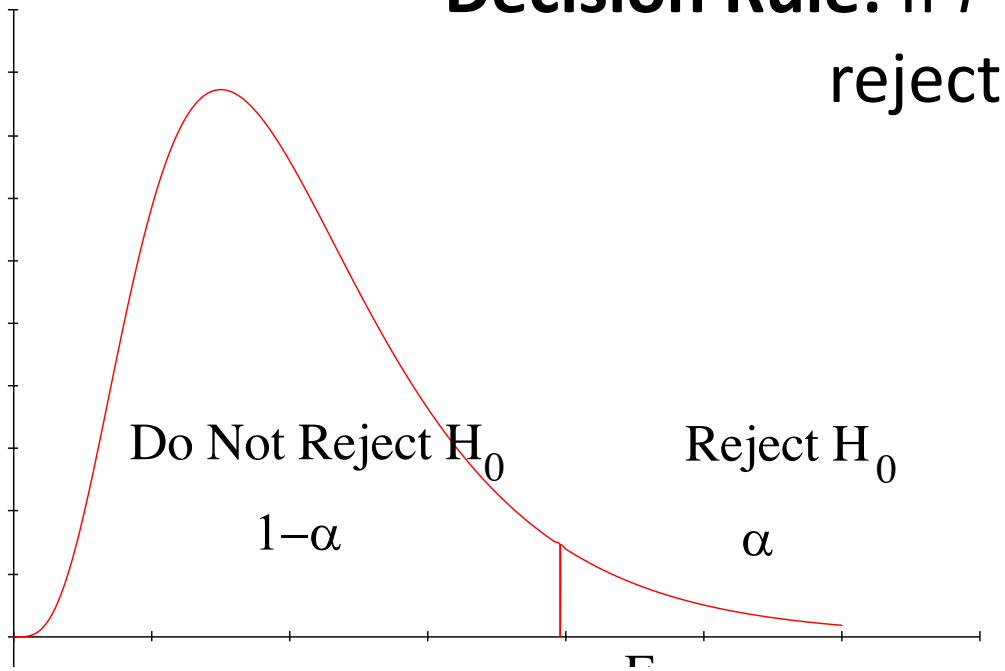
Testing Model Significance

II. **Rejection Region**

Given α and

numerator $df = k$, denominator $df = n - k - 1$

Decision Rule: If F > critical value, reject H_0.



Testing Model Significance

III. **Test Statistic**

$$F = \frac{SSR/k}{SSE/(n-k-1)}$$

where $SSR = SST - SSE$

$$SST = \Sigma(y_i - \bar{y})^2$$

$$SSE = \Sigma(y_i - \hat{y})^2$$

If H_0 is rejected:

- At least one β_i differs from zero.
- The regression equation does a better job of predicting the actual values of *y* than using the mean of *y*.

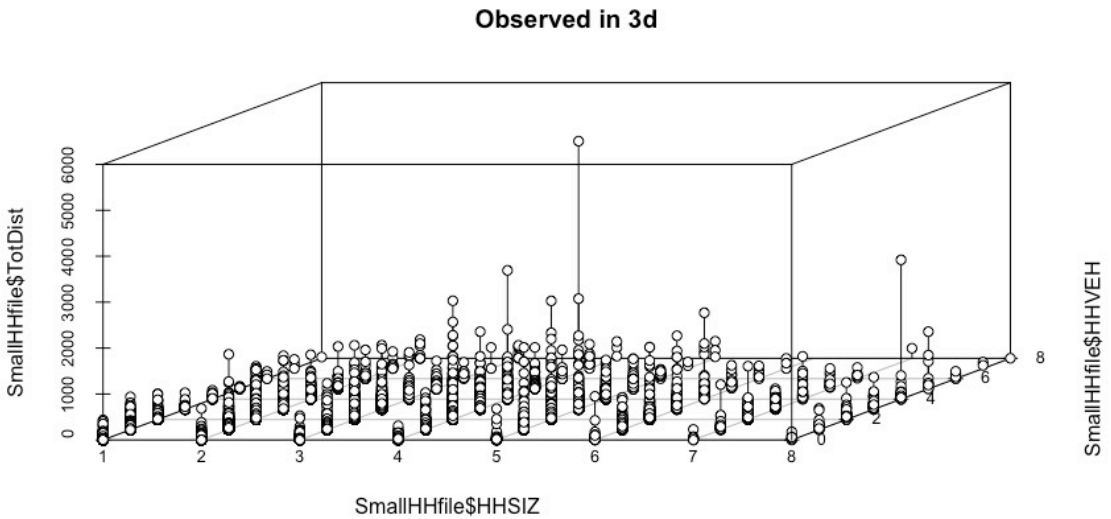
K is number of slopes

2 Models Comparison

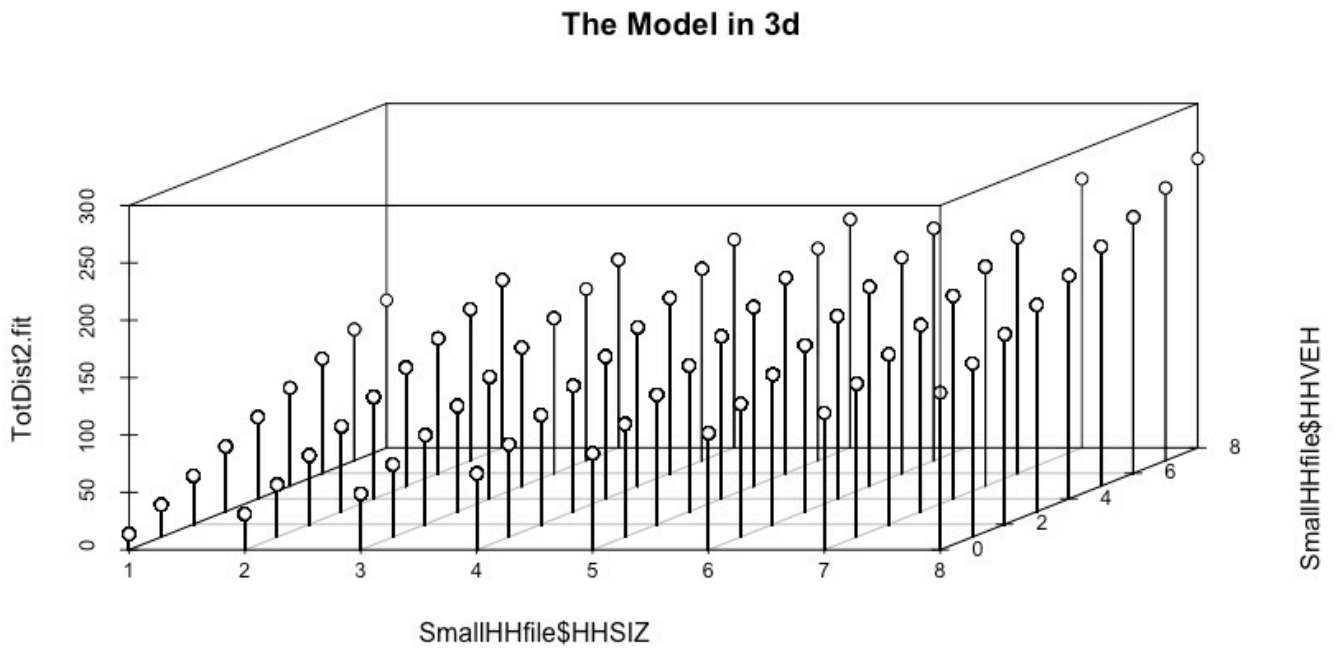
2Regression Results

	Dependent variable:	
	Number of Miles per Household	
	(1)	(2)
Household Size	21.730***	17.615***
	(0.405)	(0.438)
Household Cars		14.377***
		(0.604)
Constant	12.212***	-3.978***
	(1.182)	(1.357)
Observations	42,431	42,431
R^2	0.063	0.076
Adjusted R^2	0.063	0.076
Residual Std. Error	114.695 (df = 42429)	113.939 (df = 42428)
F Statistic	2,874.581*** (df = 1; 42429)	1,739.705*** (df = 2; 42428)
Note:		*p<0.1; **p<0.05; ***p<0.01

```
fit <- lm(TotDist~HHSIZ + HHVEH , data= SmallHHfile)

library(scatterplot3d)

s3d <- scatterplot3d(SmallHHfile$HHSIZ, SmallHHfile$HHVEH, SmallHHfile$TotDist, pch=16, highlight.3d = FALSE, type = "h", main = "3D Scatterplot")s3d$plane3d(fit)
```



2 Xs both continuous create a inclined “surface”



Model 2 is better:

1. R^2 and adjusted R^2 are higher

2. F-statistic is smaller

3. b of household cars significantly different than zero

2Regression Results

	Dependent variable:	
	Number of Miles per Household	
	(1)	(2)
Household Size	21.730***	17.615***
	(0.405)	(0.438)
Household Cars		14.377***
		(0.604)
Constant	12.212***	-3.978***
	(1.182)	(1.357)
Observations	42,431	42,431
R^2	0.063	0.076
Adjusted R^2	0.063	0.076
Residual Std. Error	114.695 (df = 42429)	113.939 (df = 42428)
F Statistic	2,874.581*** (df = 1 ; 42429)	1,739.705*** (df = 2; 42428)
Note:		*p<0.1; **p<0.05; ***p<0.01

The multiple regression model that considers the violation of the classical assumptions of homoskedasticity and lack of correlation is called the generalized regression model

In general instead of an identify matrix for the variance covariance matrix of the disturbances we have Ω.

$$E[\varepsilon\varepsilon'] = \sigma^2\Omega$$

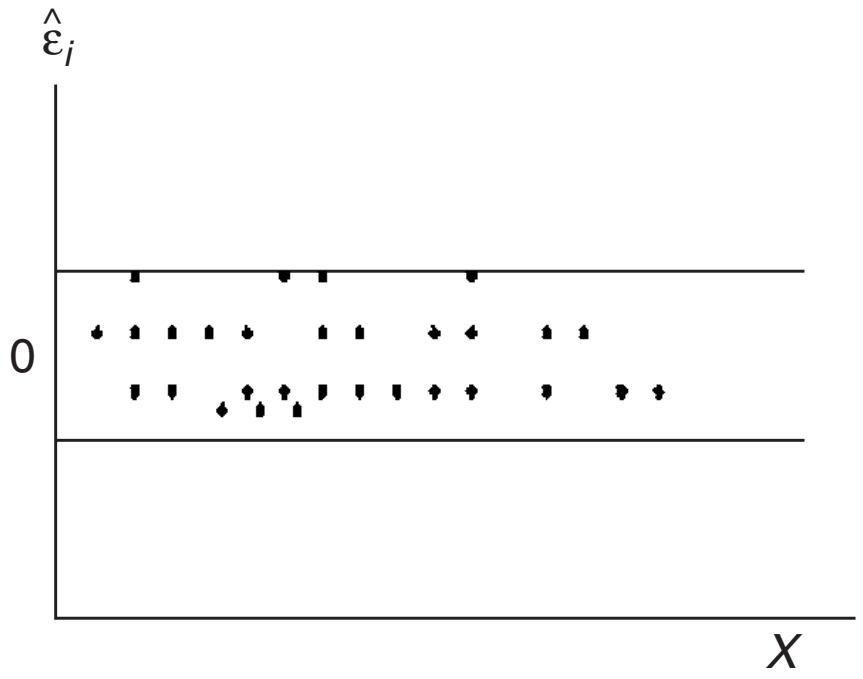
Heteroskedasticity: When the scale of the dependent variable and the model's explanatory power varies from observation to observation.

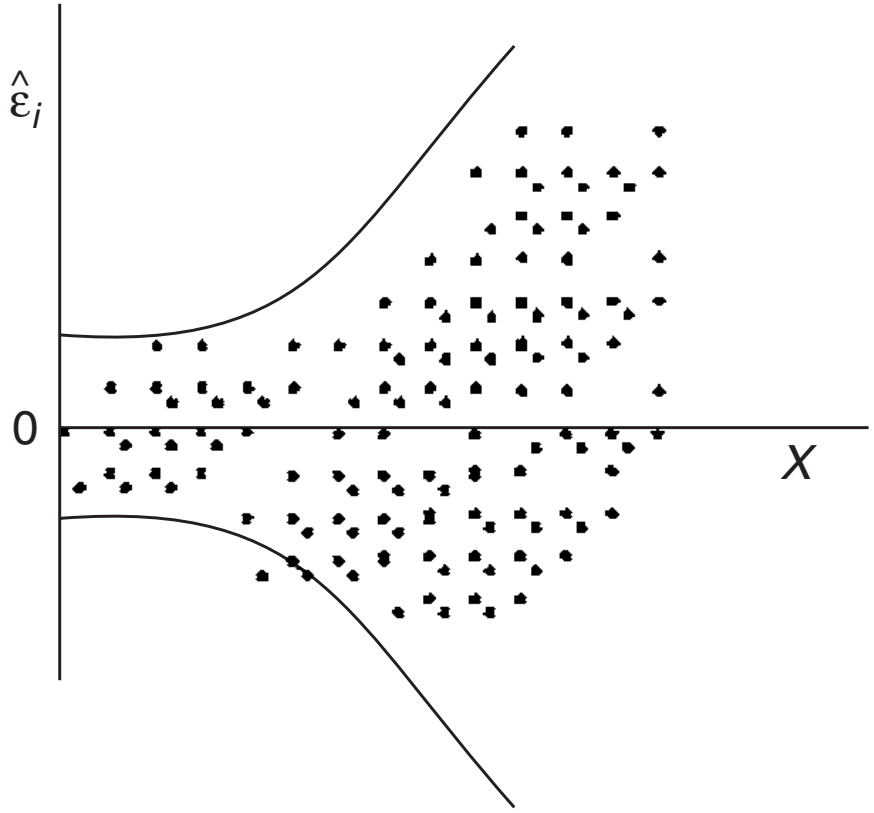
Example: High variability in trip making at high levels of income (stereotype: older people are richer but they make less trips and high income people tend to make more trips)

Heteroskedastic Disturbances

$$E[\varepsilon\varepsilon'] = \begin{bmatrix} E[\varepsilon_1\varepsilon_1] & E[\varepsilon_1\varepsilon_2] & \dots & E[\varepsilon_1\varepsilon_n] \\ E[\varepsilon_2\varepsilon_1] & E[\varepsilon_2\varepsilon_2] & \dots & E[\varepsilon_2\varepsilon_n] \\ & & \dots & \\ E[\varepsilon_n\varepsilon_1] & E[\varepsilon_n\varepsilon_2] & \dots & E[\varepsilon_n\varepsilon_n] \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ & & \dots & \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Homoskedastic vs Heteroskedastic residuals plots





Autocorrelation: When the data contain "memory" effects.

Example 1: Maybe we have observations that are also neighbors and share similar traits that we did not include in our xs

Example 2: Maybe we observe same people but at different days repeatedly

$$E[\varepsilon\varepsilon'] = \begin{bmatrix} E[\varepsilon_1\varepsilon_1] & E[\varepsilon_1\varepsilon_2] & \dots & E[\varepsilon_1\varepsilon_n] \\ E[\varepsilon_2\varepsilon_1] & E[\varepsilon_2\varepsilon_2] & \dots & E[\varepsilon_2\varepsilon_n] \\ & & \dots & \\ E[\varepsilon_n\varepsilon_1] & E[\varepsilon_n\varepsilon_2] & \dots & E[\varepsilon_n\varepsilon_n] \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 & 0 & \dots & \rho_{n-1} \\ \rho_1 & 1 & & \dots & \rho_{n-2} \\ & & \dots & & \\ \rho_{n-1} & \rho_{n-2} & & \dots & 1 \end{bmatrix}$$

The distribution of the estimator for beta is different now:

$$\hat{\beta} \sim N[\beta, \sigma^2 (X'X)^{-1} (X'\Omega X)(X'X)^{-1}]$$

INSTEAD OF:

$$\hat{\beta} \sim N[\beta, \sigma^2 (X'X)^{-1}]$$

If instead of the more complex equation for the variance covariance matrix we insist to use the usual least squares estimate $\sigma^2 (X'X)^{-1}$ we would get biased standard errors for the coefficient estimates and then the t-tests would be misleading.

Other problems may arise with the ordinary least squares depending on the particular form of Ω. We will see some of these problems in spatial statistics.

In general: Ordinary least squares, when $\Omega \neq I$, produces unbiased, consistent, and asymptotically normally distributed estimates.

The estimates are not efficient and the usual inference procedures (t-tests etc) are not valid.

Usually we don't know the elements in Ω.

One way is to get an estimate of the var-covariance matrix from Ordinary least squares and then produce weights for each observation.

Then, estimate a regression model on the weighted observations.

White's Estimator for the Variance of Least Squares

White's Estimator

White's estimator attempts to obtain a consistent estimator of the heteroskedastic variance covariance matrix.

$$(X'X)^{-1}[X'(\sigma^2\Omega)X](X'X)^{-1}$$

The true var-covariance matrix of the parameters estimates is:

To estimate this, White proposed using the ordinary least squares residuals e_i

$$Est.Var[\hat{\beta}] = n(X'X)^{-1}S_0(X'X)^{-1}$$

$$S_0 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}'_i$$

Page 163, Greene, Econometric Analysis, 2008

In essence we run our OLS. Then, get the residuals and compute a "better" standard error for the coefficient estimates to do the t-tests

In R using library sandwich:

```
library(sandwich) # I need this for the robust vcov matrix
vcov(HHPMT.lm)
```

```
##             (Intercept)      HHSIZ
## (Intercept)   1.3963356 -0.4224327
## HHSIZ        -0.4224327  0.1642723
```

```
vcovHC(HHPMT.lm)  # White's estimate of the variance -covariance matrix of the coefficient estimates
```

```
##             (Intercept)      HHSIZ
## (Intercept)   1.4699743 -0.6348233
## HHSIZ        -0.6348233  0.3183264
```

```
coef(HHPMT.lm)
```

```
## (Intercept)       HHSIZ
##    12.21181    21.73048
```

My note: explain the sqrt

```
sqrt(diag(vcovHC(HHPMT.lm, type = "const")))  # this is the same as in least squares with no White adjsustment
```

```
## (Intercept)       HHSIZ
##   1.1816664   0.4053052
```

```
sqrt(diag(vcovHC(HHPMT.lm, type = "HC0")))  # this is the traditional White's adjustment to the var-Cov of the co
efficient estimates
```

```
## (Intercept)       HHSIZ
##   1.2122251   0.5641077
```

In R another option is library lmtest

This has more variants of White's adjustments to account for heteroskedasticity

In lab we will look at these.

Bottom line=all adjustments lead to the same conclusion that

HHSIZ is a significant predictor of the dependent variable (y)

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 12.21181    1.21223  10.074 < 2.2e-16 ***
## HHSIZ       21.73048    0.56411  38.522 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# the following are other versions of computing the variance matrix of coefficient estimates
coeftest(HHPMT.lm, vcov = vcovHC(HHPMT.lm, type = "HC1")) # this is improved White's adjustment
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 12.21181    1.21225  10.074 < 2.2e-16 ***
## HHSIZ       21.73048    0.56412  38.521 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(HHPMT.lm, vcov = vcovHC(HHPMT.lm, type = "HC2")) # this is another improved White's adjustment
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 12.21181    1.21233  10.073 < 2.2e-16 ***
## HHSIZ       21.73048    0.56416  38.519 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(HHPMT.lm, vcov = vcovHC(HHPMT.lm, type = "HC3")) # this is the third improvement
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  12.2118     1.2124  10.072 < 2.2e-16 ***
## HHSIZ        21.7305     0.5642  38.515 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(HHPMT.lm, vcov = vcovHC(HHPMT.lm, type = "HC4")) # this is the fourth improvement
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 12.21181    1.21257  10.071 < 2.2e-16 ***
## HHSIZ       21.73048    0.56427  38.511 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hat values are products based on the explanatory variables that explore how influential an observation is to y hat

$$\begin{aligned}\hat{\mathbf{y}} &= \boldsymbol{X}\mathbf{b} \\ &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{y} \\ &= \mathbf{H}\mathbf{y}\end{aligned}$$

$$\underset{(n\times n)}{\mathbf{H}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

- If h_{ij} is large, the *i*th observation has a substantial impact on the *j*th fitted value
- Since $\mathbf{H}$ is symmetric and idempotent, the diagonal entries represent both the i_{th} row and the i_{th} column:

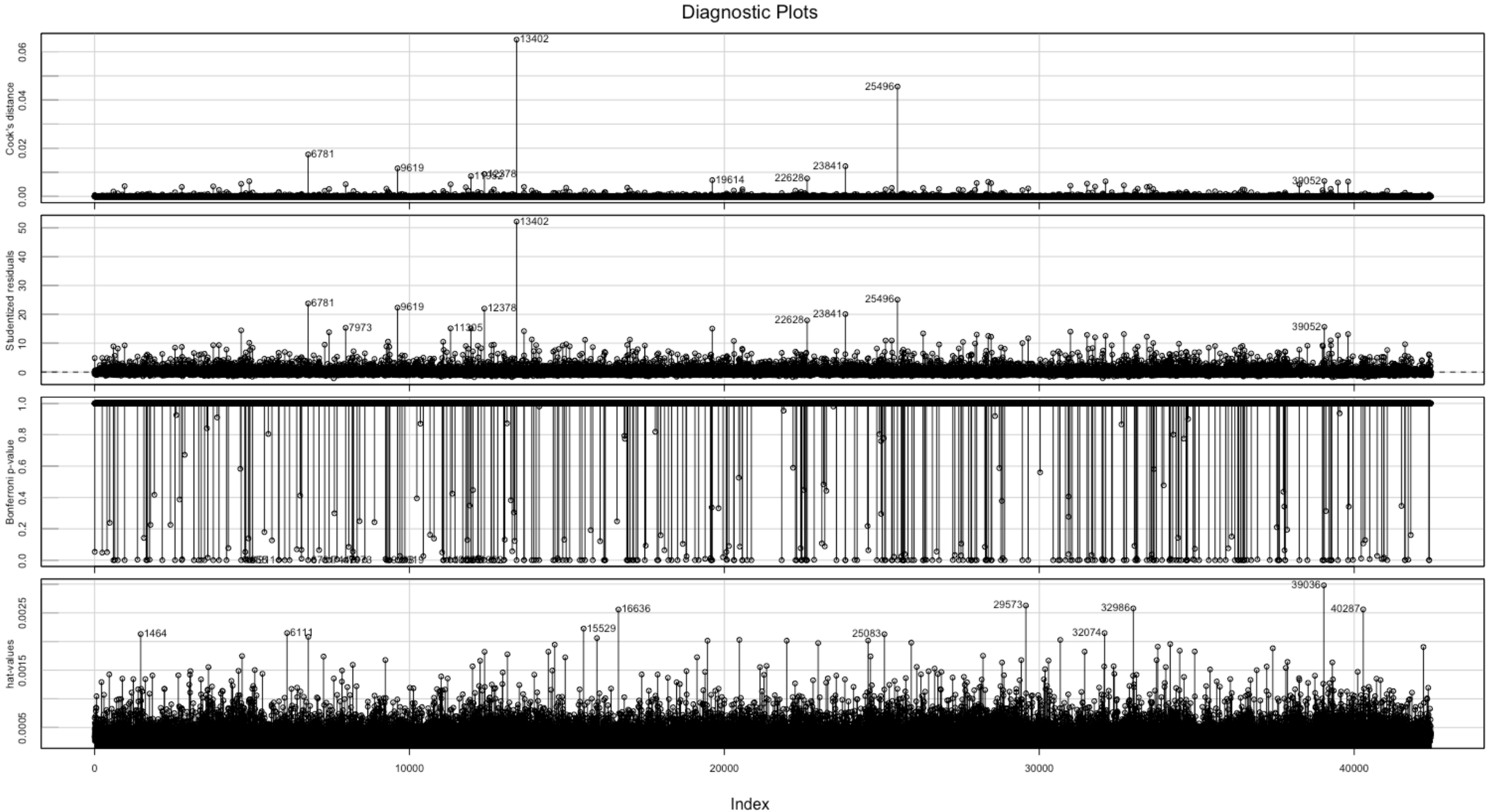
$$h_i = \boldsymbol{h}_i' \boldsymbol{h}_i = \sum_{j=1}^{n} h_{ij}^2$$

- This implies, then, that $h_i = h_{ii}$

- The average hat-value is: $\bar{h} = (k+1)/n$
- Hat values are bounded between $1/n$ and 1
- In simple regression hat values measure distance from the mean of X:

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

- In multiple regression, h_i measures the distance from the centroid point of X's (point of means)


Diagnostic Plots
Cook's distance
Studentized residuals
Bonferroni p-value
hat-values
Index
13402
25496
6781
9619
23841
19614
22628
39052
12378
7973
11305
1464
6111
15529
16636
25083
29573
32074
32986
39036
40287
0
10000
20000
30000
40000

In cross-section regressions: Assume that Ω is a diagonal matrix. A plug-in estimator for $\mathrm{Var}(\hat{\beta}|X)$ could use $\hat{\Omega} = \mathrm{diag}(\omega_1, \ldots, \omega_n)$ with:

From the AER package
Zeileis

$$\begin{aligned}
\text{const:} \quad \omega_i &= \hat{\sigma}^2 \\
\text{HC0:} \quad \omega_i &= \hat{\varepsilon}_i^2 \\
\text{HC1:} \quad \omega_i &= \frac{n}{n-k}\hat{\varepsilon}_i^2 \\
\text{HC2:} \quad \omega_i &= \frac{\hat{\varepsilon}_i^2}{1-h_{ii}} \\
\text{HC3:} \quad \omega_i &= \frac{\hat{\varepsilon}_i^2}{(1-h_{ii})^2} \\
\text{HC4:} \quad \omega_i &= \frac{\hat{\varepsilon}_i^2}{(1-h_{ii})^{\delta_i}}
\end{aligned}$$

Hat values identify influential observations that have a high impact on the predictor variables

We adjust slightly the ω_i when computing the standard error of the coefficient estimate

where h_{ii} are the hat values, $\bar{h}$ is their mean, and $\delta_i = \min\{4, h_{ii}/\bar{h}\}$.

Testing for heteroskedastic residuals

Breusch-Pagan Test

- This test first makes an assumption about the structure of the heteroscedasticity
- The assumption is that the error variance is a function of explanatory variables
- The Breusch-Pagan test assumes the error variance is a linear function of one or more variables.
- A related test called (the Harvey-Godfrey Test) assumes the error variance is an exponential function of one or more variables.
- The variables are usually assumed to be one or more of the explanatory variables in the regression equation

The Breusch-Pagan Test:

The null and the alternative hypotheses for the BP test are

$$H_0 : \sigma_i^2 = \sigma^2$$

$$\sigma_i^2 = F(\gamma + \delta Z)$$

F is a general function, Z are any explanatory variables.

If delta (δ) is equal to zero then we end up with a constant error term variance

Implementation of the hypothesis testing above is usually done by first getting residuals from the regression model as follows:

- Step 1: We regress Y_i against a constant and X_i using the OLS estimator.
- Step 2: We calculate the residuals from this regression, e_i.
- Step 3: Then we square these residuals, e_i^2 (for the Harvey-Godfrey Test, we take the logarithm of these squared residuals, $\ln(e_i^2)$)
- Step 4: For the Breusch-Pagan Test, we regress the squared residuals, e_i^2, on a constant and X_t,
- using OLS. For the Harvey-Godfrey Test, regress the logarithm of the squared residuals, $\ln(e_i^2)$,
- on a a constant and X_i, using OLS. This is called the **auxiliary** regression.

- Step 5: Find the unadjusted R^2 statistic and the number of observations, *n*, for the auxiliary regression.
- Step 6: Calculate the LM test statistic as follows (R in this case from the auxiliary regression): $LM = nR^2$.
- This LM statistic is chi-squared distributed with degrees of freedom the number of Xs
- Then, we compare the value of the test statistic to the critical value for some predetermined level of significance.
- If the calculated test statistic exceeds the critical value, then reject the null-hypothesis of constant error variance and conclude that **there is heteroscedasticity**.
- If not, do not reject the null-hypothesis and conclude that there is no evidence of heteroscedasticity.
- Bottom line: the larger this BP LM value is, the more likely to have heteroskedastic error terms in our model it is.
- Rule of thumb: Critical values of a chi-square distribution are usually approximately twice the degrees of freedom

- Note: in the literature this is also called Langrage Multiplier test (LM)

The null hypothesis is equivalent to δ=0. With the additional assumption that the e_i are normally distributed the following quantity is chi-square distributed with m (the number of variables in the function F).

$$LM = \frac{1}{2}(ExplainedSS)$$

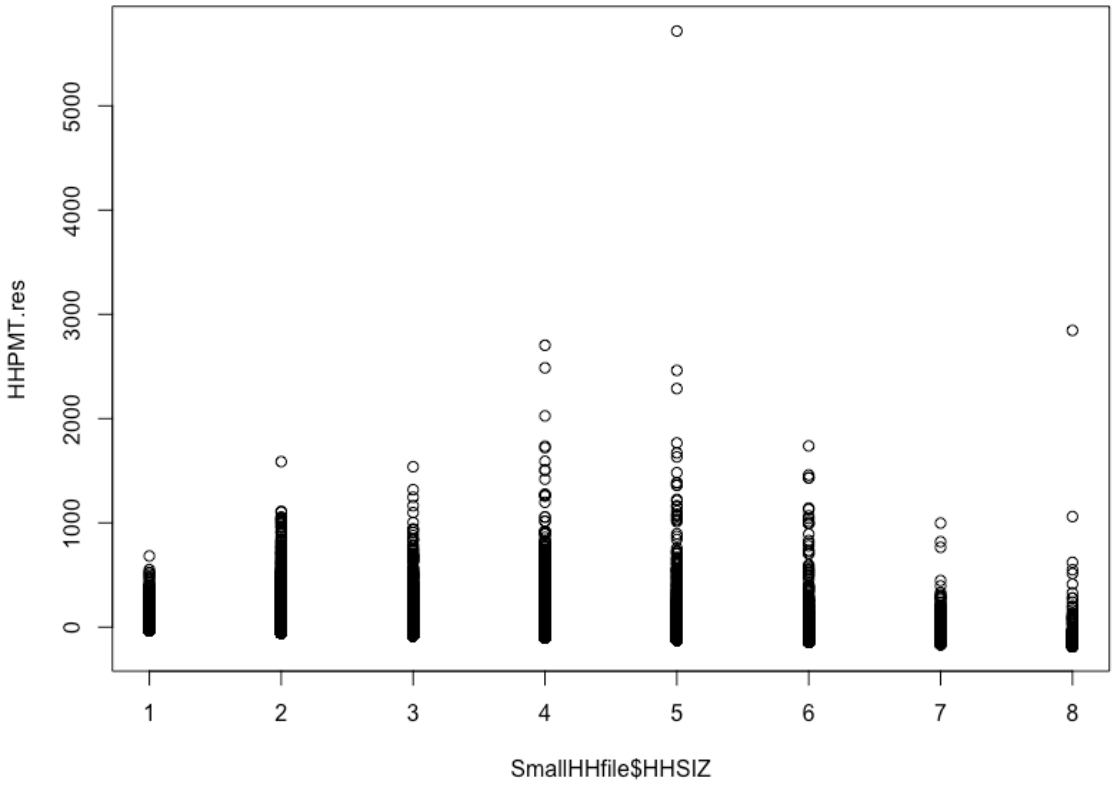
where the Explained SS comes from the regression of :

$$e_i^2 / (e'e/n)$$ **on z_i**

T.S. Breusch & A.R. Pagan (1979), A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica* **47**, 1287–1294

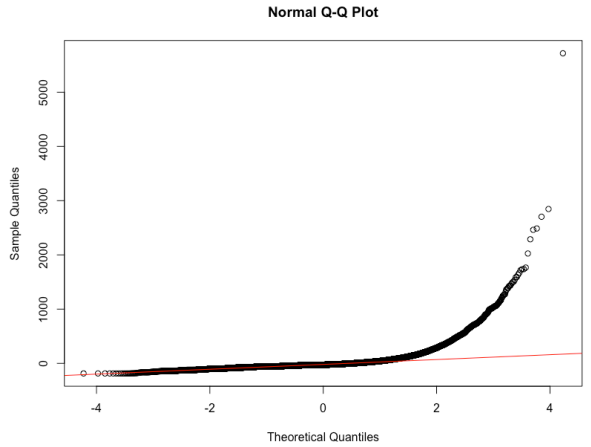
In R this is done by applying the test to a linear regression object (note the studentize = FALSE):
bptest(HHPMT.lm, studentize=FALSE)

```
> summary(HHPMT.lm)

Call:
lm(formula = TotDist ~ HHSIZ, data = SmallHHfile)

Residuals:
    Min      1Q  Median      3Q     Max
-186.1  -49.1  -26.5    11.4 5717.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.2118     1.1817   10.33   <2e-16 ***
HHSIZ        21.7305     0.4053   53.62   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 114.7 on 42429 degrees of freedom
Multiple R-squared:  0.06345,   Adjusted R-squared:  0.06343
F-statistic:  2875 on 1 and 42429 DF,  p-value: < 2.2e-16
```



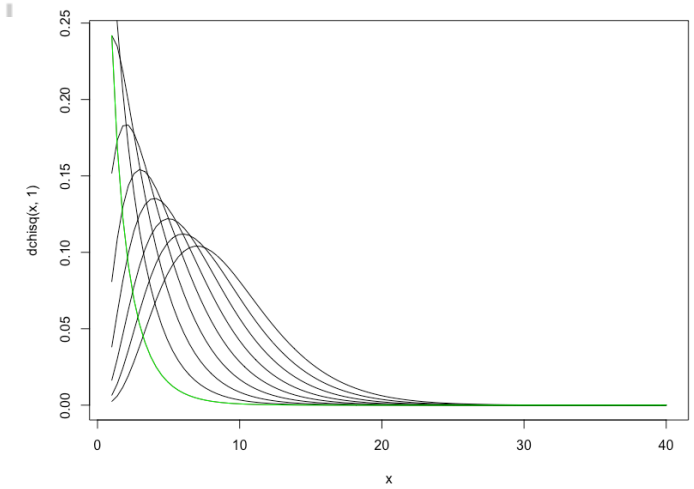
Most likely we have heteroskedastic error terms because the residuals seem to have different variance depending on the x variable (HHSIZ)

The qq plot to the right also shows non-normality



THE BREUSCH-PAGAN TEST FOR HETEROSKEDASTICITY

```
> bptest(HHPMT.lm, studentize=FALSE )

        Breusch-Pagan test

data:  HHPMT.lm
BP = 14060, df = 1, p-value < 2.2e-16

> bptest(HHPMT.lm, studentize=TRUE)

        studentized Breusch-Pagan test

data:  HHPMT.lm
BP = 131.67, df = 1, p-value < 2.2e-16
```



This test is a general test of heteroskedasticity
It is strongly based on the normality assumption regarding residuals!

But, we know they are not normally distributed so caution!

The studentized version is more robust to non normal residuals

The critical value of a chi-square distribution with 1 degree of freedom at the 95% confidence level is 3.841459

The BP numbers above are by far greater than 3.841459

WE HAVE HETEROSKEDASTIC ERRORS with this model

Maybe other variables are also influencing the residuals in this regression model

bptest(HHPMT.lm, ~HHVEH+HHSIZ+Fri+Sat+HHSTU, studentize=TRUE, data=SmallHHfile)

Note how we enter the other variables to check

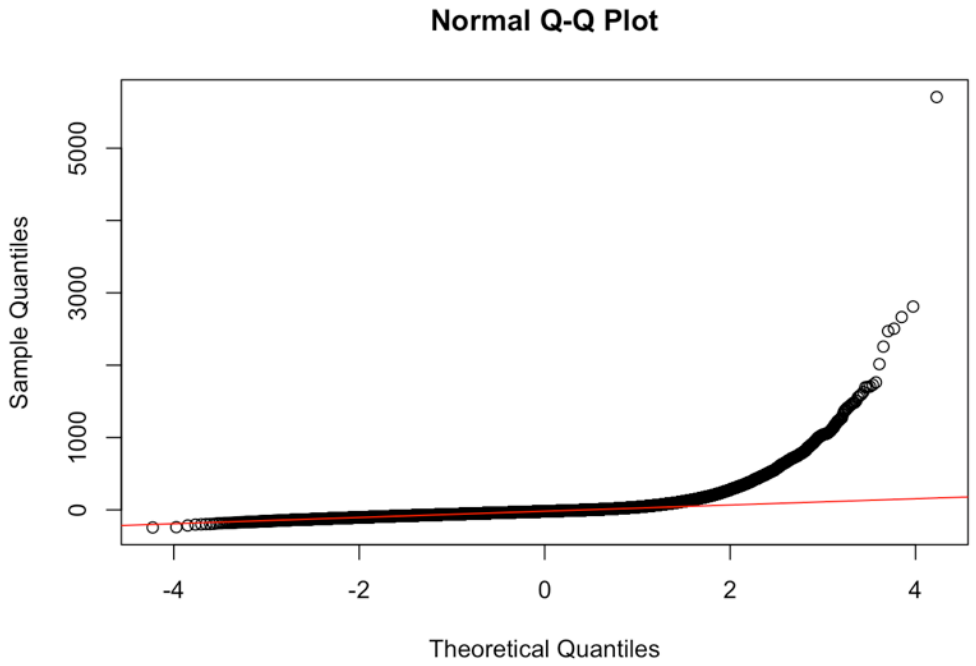
Breusch Pagan to test if other variables are also culprits in creating heteroskedasticity

```{r}
bptest(HHPMT.lm, ~HHVEH+HHSIZ+Fri+Sat+HHSTU, studentize=TRUE, data=SmallHHfile)     # the studentized version that is more robust to non-normal residuals
qchisq(.95, df=5)        # critical value at 95% confidence with 5 degrees of freedom (the five variables in auxiliary regression)
```

```
        studentized Breusch-Pagan test

data:  HHPMT.lm
BP = 150, df = 5, p-value < 2.2e-16

[1] 11.0705
```

In the lab we will also look at models that have many additional variables. One example:

```
##
## Call:
## lm(formula = TotDist ~ HHSIZ + HHVEH + highinc + Mon + Tue +
##     Wed + Thu + Fri + Sat + suburb + exurb + rural + HHEMP +
##     HHSTU + HHLIC, data = SmallHHfile)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -243.6  -47.5  -20.8   11.1 5707.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13.9504     2.1391  -6.522 7.03e-11 ***
## HHSIZ         9.2520     0.8256  11.206  < 2e-16 ***
## HHVEH         6.3211     0.7879   8.022 1.06e-15 ***
## highinc      17.4586     1.2102  14.426  < 2e-16 ***
## Mon          -3.9028     2.0618  -1.893 0.058374 .
## Tue          -1.5597     2.0339  -0.767 0.443180
## Wed           0.1132     2.0319   0.056 0.955560
## Thu          -2.2428     2.0229  -1.109 0.267575
## Fri           7.5513     2.0498   3.684 0.000230 ***
## Sat           7.3042     2.0448   3.572 0.000354 ***
## suburb        9.1248     1.4673   6.219 5.06e-10 ***
## exurb        18.0612     1.5643  11.546  < 2e-16 ***
## rural        22.8534     1.6348  13.980  < 2e-16 ***
## HHEMP         8.1317     0.7865  10.339  < 2e-16 ***
## HHSTU         8.6673     0.9328   9.292  < 2e-16 ***
## HHLIC         6.1081     1.0870   5.619 1.93e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 112.9 on 42415 degrees of freedom
## Multiple R-squared:  0.09218,    Adjusted R-squared:  0.09186
## F-statistic: 287.1 on 15 and 42415 DF,  p-value: < 2.2e-16
```




```
bptest(TotDist5.lm, studentize=TRUE)      # t
```

```
##
##  studentized Breusch-Pagan test
##
## data:  TotDist5.lm
## BP = 167.92, df = 15, p-value < 2.2e-16
```

```
qchisq(.95, df=15)        # critical value a
```

```
## [1] 24.99579
```

```
##
## Call:
## lm(formula = TotDist ~ HHSIZ + HHVEH + highinc + Mon + Tue +
##     Wed + Thu + Fri + Sat + suburb + exurb + rural + HHEMP +
##     HHSTU + HHLIC, data = SmallHHfile)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -243.6  -47.5  -20.8   11.1 5707.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13.9504     2.1391  -6.522 7.03e-11 ***
## HHSIZ         9.2520     0.8256  11.206  < 2e-16 ***
## HHVEH         6.3211     0.7879   8.022 1.06e-15 ***
## highinc      17.4586     1.2102  14.426  < 2e-16 ***
## Mon          -3.9028     2.0618  -1.893 0.058374 .
## Tue          -1.5597     2.0339  -0.767 0.443180
## Wed           0.1132     2.0319   0.056 0.955560
## Thu          -2.2428     2.0229  -1.109 0.267575
## Fri           7.5513     2.0498   3.684 0.000230 ***
## Sat           7.3042     2.0448   3.572 0.000354 ***
## suburb        9.1248     1.4673   6.219 5.06e-10 ***
## exurb        18.0612     1.5643  11.546  < 2e-16 ***
## rural        22.8534     1.6348  13.980  < 2e-16 ***
## HHEMP         8.1317     0.7865  10.339  < 2e-16 ***
## HHSTU         8.6673     0.9328   9.292  < 2e-16 ***
## HHLIC         6.1081     1.0870   5.619 1.93e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 112.9 on 42415 degrees of freedom
## Multiple R-squared:  0.09218,    Adjusted R-squared:  0.09186
## F-statistic: 287.1 on 15 and 42415 DF,  p-value: < 2.2e-16
```

Compare the same model with and without adjusting the standard errors of coefficient estimates using White's method

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -13.95036    2.21890 -6.2871 3.266e-10 ***
HHSIZ         9.25196    1.01837  9.0850 < 2.2e-16 ***
HHVEH         6.32110    0.80582  7.8443 4.454e-15 ***
highinc      17.45855    1.25879 13.8693 < 2.2e-16 ***
Mon          -3.90282    2.16357 -1.8039  0.071257 .
Tue          -1.55969    1.98252 -0.7867  0.431448
Wed           0.11323    2.06582  0.0548  0.956288
Thu          -2.24275    1.92388 -1.1657  0.243725
Fri           7.55128    2.38288  3.1690  0.001531 **
Sat           7.30420    2.34447  3.1155  0.001838 **
suburb        9.12479    1.36499  6.6849 2.340e-11 ***
exurb        18.06117    1.47030 12.2840 < 2.2e-16 ***
rural        22.85342    1.58534 14.4155 < 2.2e-16 ***
HHEMP         8.13166    0.81444  9.9843 < 2.2e-16 ***
HHSTU         8.66732    1.16038  7.4694 8.213e-14 ***
HHLIC         6.10812    1.40105  4.3597 1.306e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing for (auto)correlated residuals

Autocorrelated Disturbances

$$E[\varepsilon\varepsilon'] = \begin{bmatrix} E[\varepsilon_1\varepsilon_1]\ E[\varepsilon_1\varepsilon_2] \ldots E[\varepsilon_1\varepsilon_n] \\ E[\varepsilon_2\varepsilon_1]\ E[\varepsilon_2\varepsilon_2] \ldots E[\varepsilon_2\varepsilon_n] \\ \ldots \\ E[\varepsilon_n\varepsilon_1]\ E[\varepsilon_n\varepsilon_2] \ldots E[\varepsilon_n\varepsilon_n] \end{bmatrix} = \sigma^2 \begin{bmatrix} 1\ \rho_1\ \rho_2 \ldots \rho_{n-1} \\ \rho_1\ 1\ \ldots \rho_{n-2} \\ \ldots \\ \rho_{n-1}\ \rho_{n-2}\ \ldots 1 \end{bmatrix}$$

(When the data contain unobserved "memory" effects)

$$Cov[\varepsilon_t, \varepsilon_{t-s}] = Cov[\varepsilon_{t+s}, \varepsilon_t] = \gamma_s$$

If the autocovariance is:

Which means that Cov. is a function of the temporal "distance" s and not each t then the autoregressive process underlying the disturbances is called stationary.

The correlation between disturbances is:

$$Corr[\varepsilon_t, \varepsilon_{t-s}] = \frac{Cov[\varepsilon_{t+s}, \varepsilon_t]}{\sqrt{Var[\varepsilon_t] Var[\varepsilon_{t-s}]}} = \frac{\gamma_s}{\gamma_0} = \rho_s$$

where $\gamma_0 = \sigma^2$.

First Order Autoregressive Process

AR(1)

$$\varepsilon_t = \rho\, \varepsilon_{t-1} + u_t$$

It can be shown that for first order autoregressive processes:

$$\rho_s = \rho^s$$

Since the correlation is between 0 and 1 as s increases the correlation decreases really fast.

First Order Autoregressive Process

AR(1)

It can also be shown that the variance covariance matrix becomes:

$$E[\varepsilon\varepsilon'] = \sigma^2\Omega = \frac{\sigma_u^2}{1-\rho^2}\begin{bmatrix} 1 & \rho^1 & \rho^2 & \dots & \rho^{n-1} \\ \rho^1 & 1 & & \dots & \rho^{n-2} \\ & & \dots & & \\ \rho^{n-1} & \rho^{n-2} & & \dots & 1 \end{bmatrix}$$

Remember that:

$$\hat{\beta} \sim N[\beta, \sigma^2 (X'X)^{-1} (X'\Omega X)(X'X)^{-1}]$$

If instead of the more complex equation for the variance covariance matrix we use the usual least squares estimate, i.e., $\sigma^2 (X'X)^{-1}$, we would get **biased standard errors** for the coefficient estimates and then the **t-tests would be misleading**.

Usually we don't know the elements in the variance covariance matrix

But first we need a test for autocorrelation

Durbin-Watson Test for Autocorrelation

15.7.1. The Durbin–Watson Test

Most of the available tests for autocorrelation are based on the principle that if the true disturbances are autocorrelated, this will be revealed through the autocorrelations of the least squares residuals. By far the most widely used test is the Durbin–Watson test.[22] The test statistic is

$$d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}. \tag{15-37}$$

The statistic is closely related to the sample autocorrelation:

$$d = 2(1 - r) + \frac{e_1^2 + e_T^2}{\sum_{t=1}^{T} e_t^2}.$$

If the sample is reasonably large, the last term will be negligible, leaving

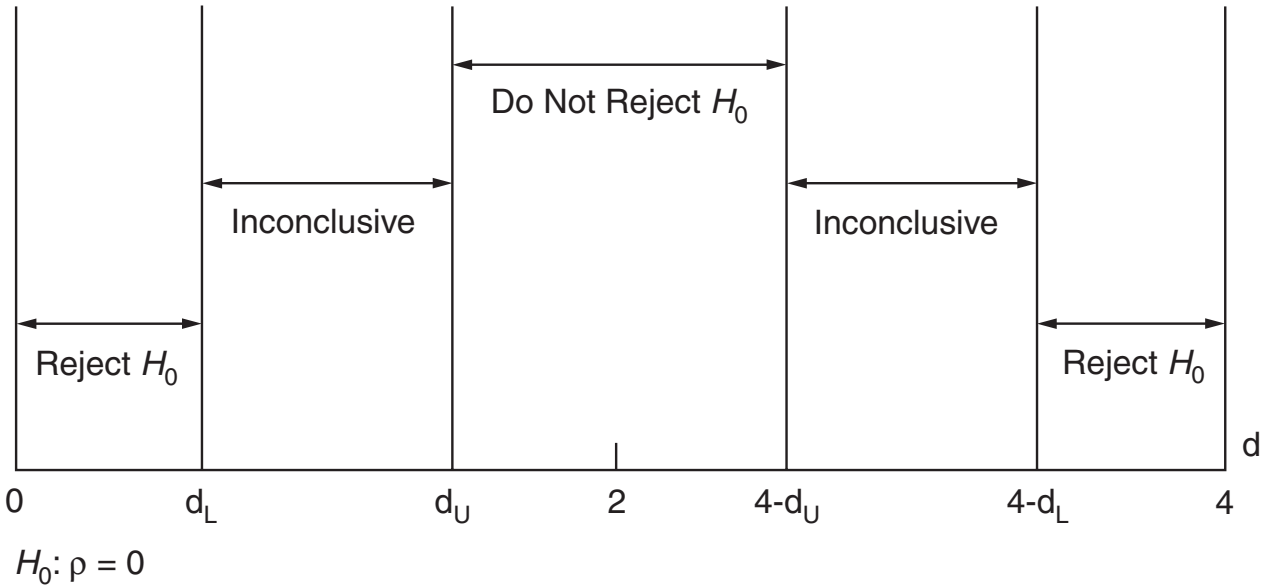
$$d \simeq 2(1 - r). \tag{15-38}$$

Values of d that differ significantly from two suggest autocorrelation of the disturbances.

From Greene's textbook -1990 – see also a WKM chapter on Gauchospace

How we use this test

- DW test statistics varies between 0 and 4. A value close to 2 indicates no autocorrelation. A value close to zero indicates positive correlation and a value close to 4 negative correlation.
- DW derived boundaries within which the DW statistic for a given model will need to be in order not to reject the null hypothesis that the residuals are not correlated ro maybe correlated and we need to do something about it.
- From the WKM book and published tables in https://www3.nd.edu/~wevans1/econ30331/Durbin_Watson_tables.pdf



- I get for one X (HHSIZ) and 5% confidence in correspondence of 100 observations (we have 42000+) so we are conservative using the table values of d_U = 1.69 and d_L=1.65

My DW statistic for the HHPMT.lm is 1.9724

This is between 1.69 and 4-1.69 = 2.31

I am very confident the residuals are not correlated because 1.9724 is within these boundaries

The default version of AR1 is as follows:
Compute an estimate of rho and then run a new regression by transforming y and x

$$Y^* = \begin{bmatrix} \sqrt{1-\hat{\rho}^2}\,y_1 \\ y_2 - \hat{\rho} y_1 \\ y_3 - \hat{\rho} y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n - \hat{\rho} y_n \end{bmatrix} \qquad X^* = \begin{bmatrix} \sqrt{1-\hat{\rho}^2} & \sqrt{1-\hat{\rho}^2}\,x_{11} & \sqrt{1-\hat{\rho}^2}\,x_{21} \\ 1-\hat{\rho} & x_{12} - \hat{\rho} x_{11} & x_{22} - \hat{\rho} x_{21} \\ 1-\hat{\rho} & x_{13} - \hat{\rho} x_{12} & x_{23} - \hat{\rho} x_{22} \\ \cdot & \cdot & \cdot \end{bmatrix}$$

Apply least squares to Y* and X* and compute a new set of residuals – compute a new rho hat and compare with the previously computed one. If the two rho hats are different continue – else stop

The DW statistic will tell you when to stop at DW=2

We will talk about this as spatial correlation later in 210B

Generalized Regression Weighted Least Squares

Estimation of a GLM with Heteroskedastic Error Term

$$E[\varepsilon\varepsilon'] = \sigma^2\Omega$$

Note: Instead of an identity matrix (I) for the variance covariance matrix of the disturbances we have Ω.

Heteroskedastic Disturbances

In matrix notation:

$$E[\varepsilon\varepsilon'] = \begin{bmatrix} E[\varepsilon_1\varepsilon_1] & E[\varepsilon_1\varepsilon_2] & \dots & E[\varepsilon_1\varepsilon_n] \\ E[\varepsilon_2\varepsilon_1] & E[\varepsilon_2\varepsilon_2] & \dots & E[\varepsilon_2\varepsilon_n] \\ & & \dots & \\ E[\varepsilon_n\varepsilon_1] & E[\varepsilon_n\varepsilon_2] & \dots & E[\varepsilon_n\varepsilon_n] \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ & & \dots & \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Note: We still assume no cross-error correlation!

If we neglect the presence of heteroskedasticity and use Ordinary least squares:

$$\hat{\beta} = (X'X)^{-1} X'Y$$

1. **The parameter estimates are still consistent**

$$V = s^2 (X'X)^{-1}$$

2. **The usual estimates of their variance-covariance matrix**

may not be consistent (i.e., if X and the elements of omega are correlated)

3. **The OLS standard errors of the coefficient estimates are BIASED - so what does this imply in practice? Can you trust the t-statistics?**

If we have heteroskedastic error terms, the estimator for regression coefficients is:

$$\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

This time the elements of Omega play a role in the determination of the values of the regression coefficients. The matrix Omega is:

$$\Omega = \begin{bmatrix} \omega_1 & 0 & \dots & 0 \\ 0 & \omega_2 & \dots & 0 \\ & & \dots & 0 \\ 0 & 0 & \dots & \omega_n \end{bmatrix}$$

The inverse of this is a diagonal matrix with elements $1/\omega_i$

If I multiply every x and every y by 1/sqrt(ω_i) and then regress the transformed (weighted) variables:

$$Py = \begin{bmatrix} y_1 / \sqrt{\omega_1} \\ y_2 / \sqrt{\omega_2} \\ \cdots \\ y_n / \sqrt{\omega_n} \end{bmatrix}$$

Regress this on PX using least squares:

$$PX = \begin{bmatrix} x_1 / \sqrt{\omega_1} \\ x_2 / \sqrt{\omega_2} \\ \cdots \\ x_n / \sqrt{\omega_n} \end{bmatrix}$$

This gives the Weighted Least Squares estimator (where $w_i = 1/\omega_i$):

$$\widehat{\beta} = \left[\sum_{i=1}^{n} w_i x_i x_i' \right]^{-1} \left[\sum_{i=1}^{n} w_i x_i y_i \right]$$

Observations with smaller variances receive a larger weight in computing the sums - so they influence more.

Getting estimates of the weights w_i

- They can be functions of other explanatory variables
- We can use some kind of iterative process in which we first compute the residuals from the unweighted regression and then build a regression model of the residuals squared as dependent variable and then create weights (see companion document)
- We can use external information about the origins of heteroskedasticity

- The nuclear option: think about the entire modeling process and if the linear model is appropriate from the background science viewpoint

Population equation: $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$

Sample equation = Estimate of the y variable using sample data: $\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_k x_{ki}$

Residual is the difference between observed dependent variable and estimated: $e_i = y_i - \hat{y}_i$

I prefer to run many regression models and look at them

- Check fit
- Check common sense of betas
- Compare with literature
- Tests – goodness of fit and significance

- What else?

For each of the following models we will write the equations

Regression Results

	Dependent variable:				
	Number of Miles per Household				
	(1)	(2)	(3)	(4)	(5)
Household Size	21.730***	17.615***	17.583***	17.767***	9.252***
	(0.405)	(0.438)	(0.437)	(0.436)	(0.826)
Household Cars		14.377***	11.376***	9.806***	6.321***
		(0.604)	(0.627)	(0.636)	(0.788)
High Income			19.744***	21.150***	17.459***
			(1.177)	(1.182)	(1.210)
Monday			-3.755*	-3.959*	-3.903*
			(2.072)	(2.067)	(2.062)
Tuesday			-2.079	-1.396	-1.560
			(2.043)	(2.039)	(2.034)
Wednesday			-0.345	0.310	0.113
			(2.041)	(2.037)	(2.032)
Thursday			-2.766	-2.042	-2.243
			(2.032)	(2.028)	(2.023)
Friday			7.448***	7.454***	7.551***
			(2.060)	(2.055)	(2.050)
Saturday			7.390***	7.323***	7.304***
			(2.054)	(2.050)	(2.045)
Residence in Suburb				8.264***	9.125***
				(1.468)	(1.467)
Residence in Exurb				16.848***	18.061***
				(1.564)	(1.564)
Residence in Rural Env				20.813***	22.853***
				(1.631)	(1.635)
Number of Employed					8.132***
					(0.787)
Number of Students					8.667***
					(0.933)
Number of Driver Licenses					6.108***
					(1.087)
Constant	12.212***	-3.978***	-7.276***	-16.105***	-13.950***
	(1.182)	(1.357)	(1.899)	(2.036)	(2.139)
Observations	42,431	42,431	42,431	42,431	42,431
R^2	0.063	0.076	0.083	0.087	0.092
Adjusted R^2	0.063	0.076	0.083	0.087	0.092
Residual Std. Error	114.695 (df = 42429)	113.939 (df = 42428)	113.500 (df = 42421)	113.235 (df = 42418)	112.941 (df = 42415)
F Statistic	2,874.581*** (df = 1; 42429)	1,739.705*** (df = 2; 42428)	426.900*** (df = 9; 42421)	338.475*** (df = 12; 42418)	287.110*** (df = 15; 42415)

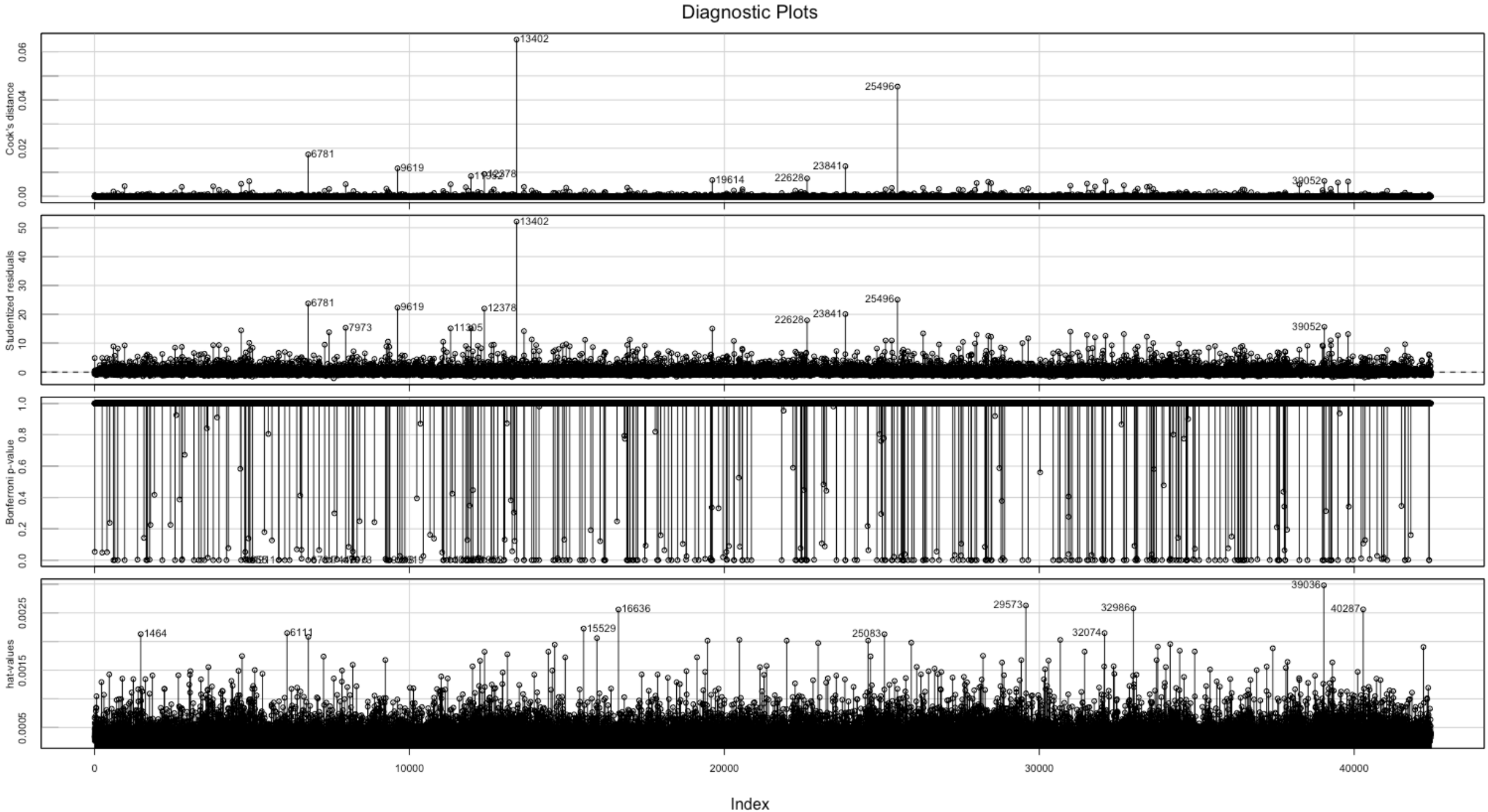
Note: *p<0.1; **p<0.05; ***p<0.01

Another cool package in R is car

- Cook's distance measures how much an observation influences the overall model or predicted values
- Studentized residuals are the residuals divided by their estimated standard deviation as a way to standardize them
- Bonferroni test to identify outliers
- Hat-points identify influential observations (have a high impact on the predictor variables)

- Next graphs show 10 possible “weird” households. The R code is:

```
install.packages("car")
library(car)
influenceIndexPlot(TotDist5.lm, id.n=10)
```


Diagnostic Plots
Cook's distance
Studentized residuals
Bonferroni p-value
hat-values
Index
13402
25496
6781
9619
23841
19614
22628
39052
12378
7973
11305
1464
6111
15529
16636
25083
29573
32074
32986
39036
40287
0
10000
20000
30000
40000