

## Day 2: Departures from OLS Assumptions and ANOVA

# Outline

- ▶ Recap
- ▶ Departures from Assumptions
- ▶ Regression Diagnostics
- ▶ One-Way ANOVA

## From Last Time...

Under certain conditions, we have the following results:

- ▶  $\hat{\beta} = (X^T X)^{-1} X^T Y$  is an unbiased estimator for regression coefficients  $\beta$  with variance  $\sigma^2 (X^T X)^{-1}$ .
- ▶  $S^2 = \frac{(Y - X\hat{\beta})^T (Y - X\hat{\beta})}{n-r}$  is an unbiased estimator for  $\sigma^2$ .
- ▶ If  $X$  has full rank, then  $\hat{\beta}$  is the BLUE.
- ▶ If  $\epsilon$  are normally distributed, then  $\hat{\beta} \sim \mathcal{N}((X^T X)^{-1} X^T Y, \sigma^2 (X^T X)^{-1})$ .

## From Last Time...

These results depend on four assumptions:

- ▶ Errors are unbiased:  $E[\epsilon] = 0$ .
- ▶ Errors have constant variance (homoscedasticity).
- ▶ Errors are uncorrelated.
- ▶ Errors are normally distributed (only necessary for distributional results).

## Departures from Assumptions

# What Happens When Assumptions Don't Hold?

There are a series of situations where the usual assumptions do not apply, and it is helpful to know how OLS estimates behave in those circumstances.

- ▶ Underfitting
- ▶ Overfitting
- ▶ Misspecified Covariance Matrix
- ▶ Non-normality

# Underfitting

Underfitting means that the model was fit with too few explanatory variables. This can be represented as fitting

$$E[Y] = X\beta$$

when the true model generating the data is

$$E[Y] = X\beta + Z\gamma.$$

# Underfitting

In this case,

$$\begin{aligned}E[\hat{\beta}] &= E[(X^T X)^{-1} X^T Y] \\&= (X^T X)^{-1} X^T (X\beta + Z\gamma) \\&= (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T Z\gamma \\&= \beta + (X^T X)^{-1} X^T Z\gamma,\end{aligned}$$

so  $\hat{\beta}$  is a biased estimator for  $\beta$ .



# Underfitting

The covariance matrix for  $\hat{\beta}$  is still the same:

$$\begin{aligned} V[\hat{\beta}] &= V[(X^T X)^{-1} X^T Y] \\ &= (X^T X)^{-1} X^T V[Y] X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

However, we find that underfitting causes  $S^2$  to be biased upwards.

# Overfitting

Overfitting involves including too many explanatory variables in the model, like fitting

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

when the true model is

$$Y = X_1\beta_1 + \epsilon$$

# Overfitting

We can show that  $\hat{\beta}$  is still unbiased:

$$\begin{aligned}E[\hat{\beta}] &= (X^T X)^{-1} X^T E[Y] \\&= (X^T X)^{-1} X^T X_1 \beta_1 \\&= (X^T X)^{-1} X^T \begin{pmatrix} X_1 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ 0 \end{pmatrix} \\&= (X^T X)^{-1} X^T X \begin{pmatrix} \beta_1 \\ 0 \end{pmatrix} \\&= \begin{pmatrix} \beta_1 \\ 0 \end{pmatrix}\end{aligned}$$

# Overfitting

We can also show that the error variance estimate  $S^2$  is unbiased, but the covariance matrix  $Var(\hat{\beta})$  is higher than the true variance.

## Misspecified Covariance

One way of misspecifying the covariance matrix is by assuming that  $Var(\epsilon) = \sigma^2 I$  when it is actually

$$Var(\epsilon) = \sigma^2 V.$$

In this case,

$$\begin{aligned} Var(\hat{\beta}) &= Var((X^T X)^{-1} X^T Y) \\ &= (X^T X)^{-1} X^T Var(Y) X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T (\sigma^2 V) X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} X^T V X (X^T X)^{-1} \\ &\neq \sigma^2 (X^T X)^{-1} \end{aligned}$$

In most cases,  $S^2$  is also a biased estimator for  $\sigma^2$ .

# Non-Normality

If we have correctly specified the model  $Y = X\beta + \epsilon$  with  $E[\epsilon] = 0$  and  $Var(\epsilon) = \sigma^2 I$ , but incorrectly assume that the errors are normally distributed:

- ▶  $\hat{\beta}$  is unbiased for  $\beta$
- ▶  $Var(\hat{\beta}) = \sigma^2(X^T X)^{-1}$
- ▶ If the sample size is large enough, then  $\hat{\beta} \approx \mathcal{N}(\beta, \sigma^2(X^T X)^{-1})$ .

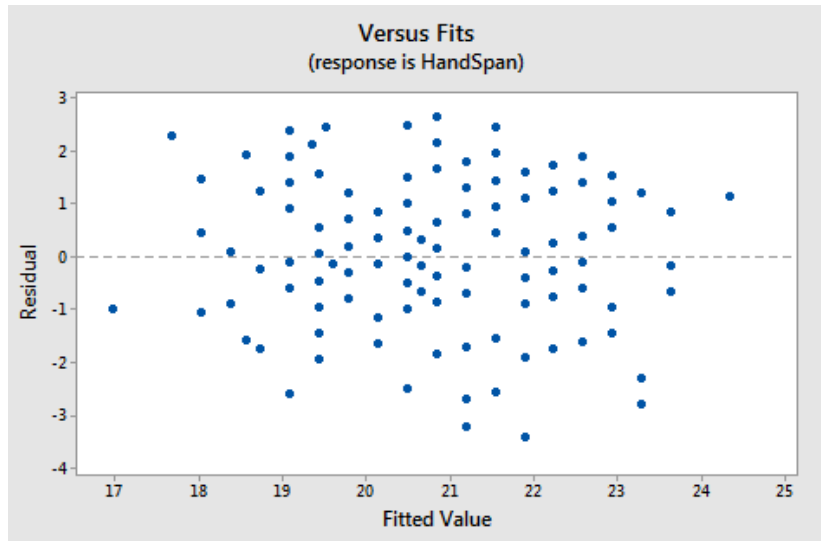
# How Can We Check Assumptions?

We are able to check the validity of some assumptions empirically.  
The main tools we use are:

- ▶ Residual Plots
- ▶ Normal Q-Q Plots

## Residual Plots

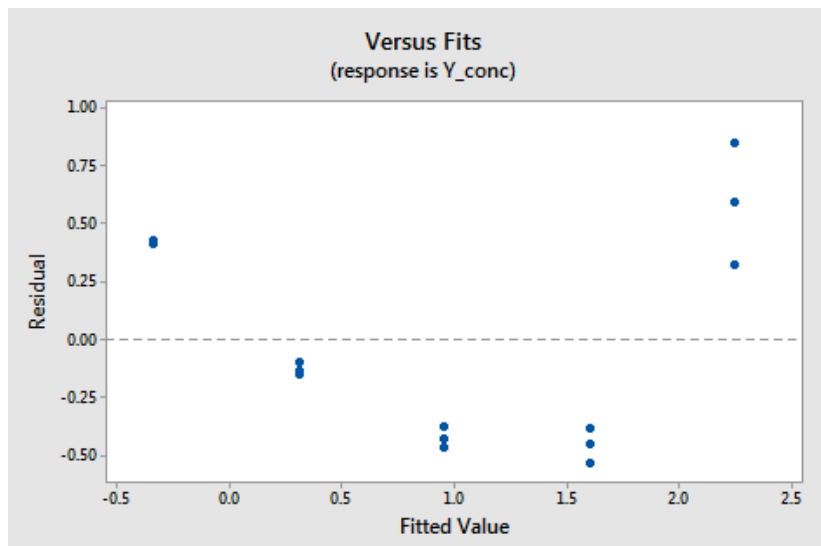
Comparing the model residuals to the fitted values allows us to check whether the linearity and homoscedasticity assumptions are correct.





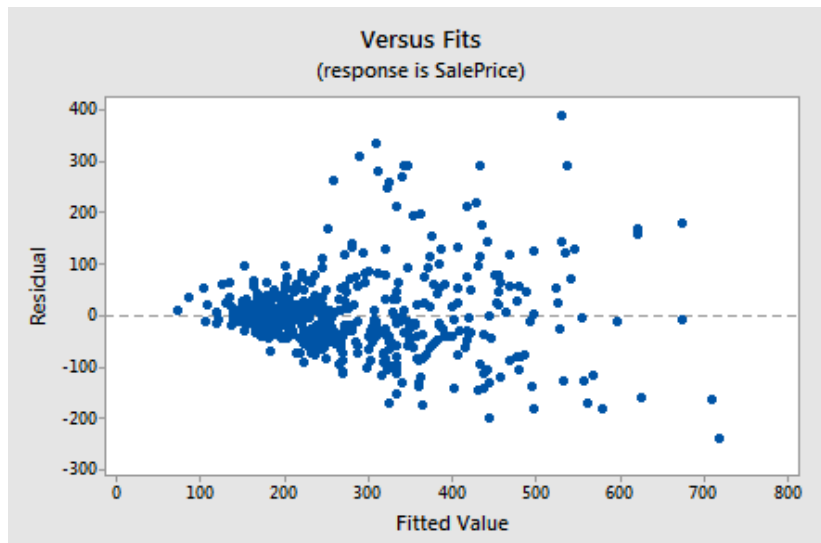
# Residual Plots

Here we see a case where linearity does not hold:



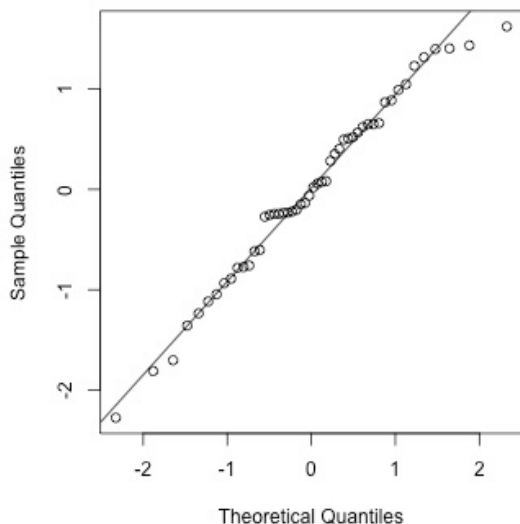
## Residual Plots

Here, homoscedasticity does not hold:



## Normal Q-Q Plots

Normal Q-Q plots allow us to verify the normality of the residuals:



# Domain Knowledge

Domain knowledge can be the most important way to verify the validity of the model

- ▶ Do the variables included make sense?
- ▶ Are there variables that should be included that are not?
- ▶ Is a linear model a reasonable assumption?

# ANOVA

# Regression with Categorical Predictors

We usually consider linear regression being used with continuous explanatory variables, but sometimes we have variables that indicate belonging to a certain group.

- ▶ Analysis of Variance (ANOVA) models are essentially linear regression models where the explanatory variables are indicator variables.
- ▶ Analysis of Covariance (ANCOVA) models are regression models using both continuous and indicator variables as explanatory variables.

# One-Way ANOVA

One-way ANOVA models are used in situations where we want to compare several independent samples. We have data where we assume:

- ▶ We have  $I$  independent samples
- ▶ Each sample is allowed to have a different number of observations, denoted  $J_i$

The simplest way to describe this data is with the model:

$$Y_{ij} = \theta_i + \epsilon_{ij}.$$

Here, we can assume that  $E[\epsilon_{ij}] = 0$ , so  $E[Y_{ij}] = \theta_i$ . This means that  $\theta_i$  represents the mean of the  $i$ th sample.

## Alternative Parameterization

Another common way to parameterize the one-way ANOVA model is:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}.$$



# One-Way ANOVA Assumptions

The validity of the ANOVA results depend on the following assumptions:

- ▶  $E[\epsilon_{ij}] = 0$ .
- ▶  $Var(\epsilon_{ij}) = \sigma_i^2 < \infty$  and  $Cov(\epsilon_{ij}, \epsilon_{i'j'}) = 0$  if  $i \neq i'$  or  $j \neq j'$ .
- ▶  $\sigma_i^2 = \sigma^2$  for all  $i$ : samples have constant variances (homoscedasticity).
- ▶  $\epsilon_{ij}$  are independent and normally distributed.

These assumptions are the same as the OLS assumptions.

# Inference Using ANOVA

One-way ANOVA models are traditionally used to test the hypotheses:

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_I$$

$$H_1 : \theta_i \neq \theta_j \text{ for some } i \neq j.$$

# One-Way ANOVA Table

Variation	DF	SSE	MSE	F Statistic
Between Groups	$I - 1$	$SSB = \sum_i J_i (\bar{y}_i - \bar{y})^2$	$MSB = \frac{SSB}{I-1}$	$F = \frac{MSB}{MSW}$
Within Groups	$N - I$	$SSW = \sum_j \sum_i (y_{ij} - \bar{y}_i)^2$	$MSW = \frac{SSW}{N-I}$	
Total	$N - 1$	$SST = \sum_j \sum_i (y_{ij} - \bar{y})^2$		