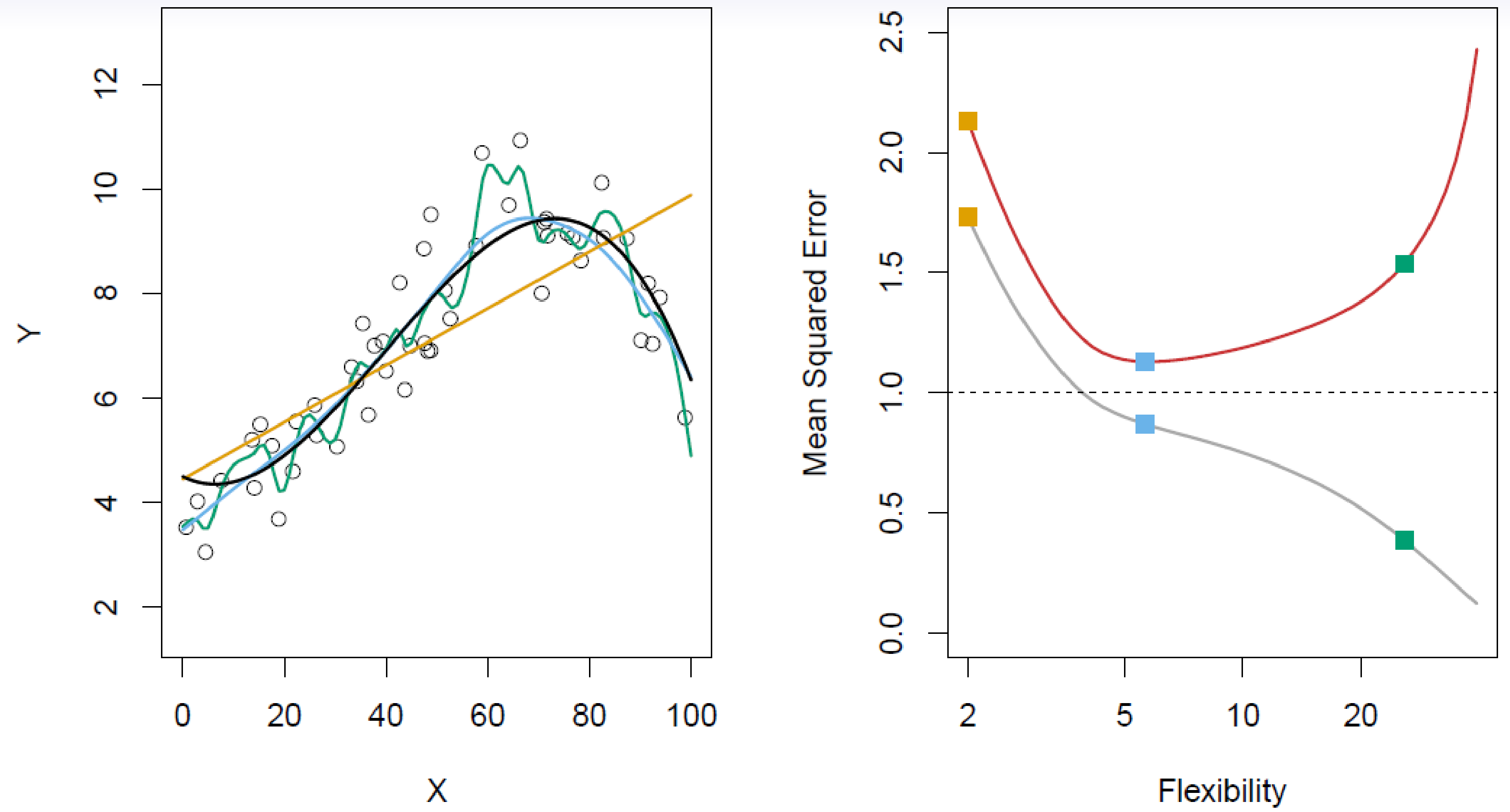
Class 3 – Regression

Pedram Jahangiry

Jon M. Huntsman School of Business
Utah State University

Training vs. Test MSE



Black curve is truth. Red curve on right is MSE_{Te}, grey curve is MSE_{Tr}. Orange, blue and green curves/squares correspond to fits of different flexibility.

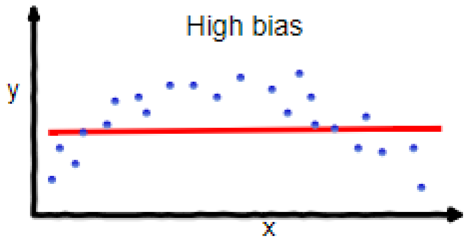
Bias-Variance Trade-off

Suppose we have fit a model $\hat{f}(x)$ to some training data Tr, and let (x_0, y_0) be a test observation drawn from the population. If the true model is $Y = f(X) + \epsilon$ (with $f(x) = E(Y|X = x)$), then

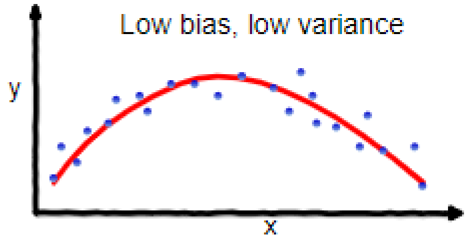
$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon).$$

The expectation averages over the variability of y_0 as well as the variability in Tr. Note that $\text{Bias}(\hat{f}(x_0))] = E[\hat{f}(x_0)] - f(x_0)$.

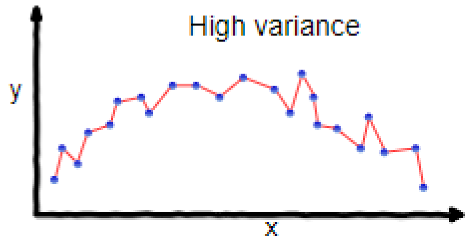
Typically as the *flexibility* of $\hat{f}$ increases, its variance increases, and its bias decreases. So choosing the flexibility based on average test error amounts to a *bias-variance trade-off*.



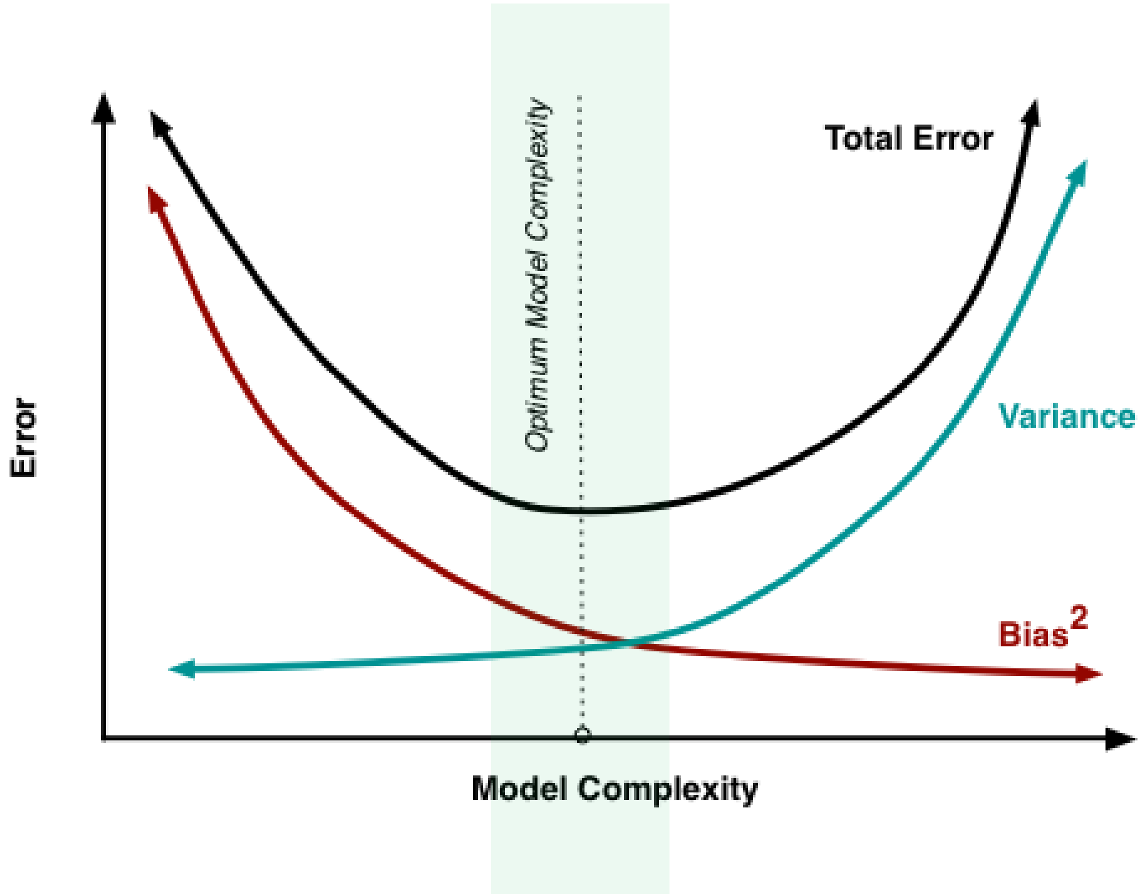
underfitting



Good balance

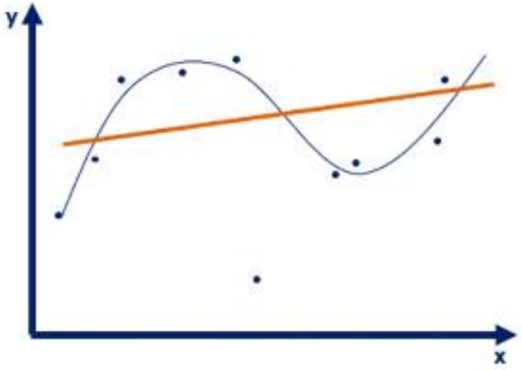


overfitting



Underfitting and overfitting

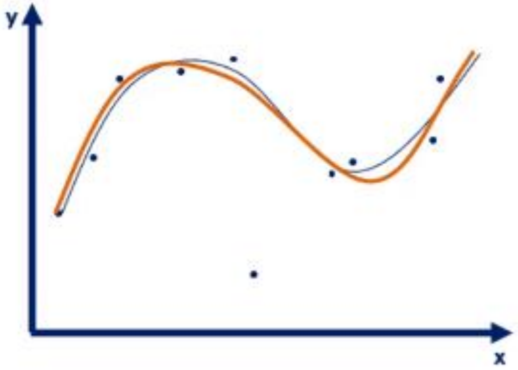
An **underfitted** model



Doesn't capture any logic

- Low train accuracy

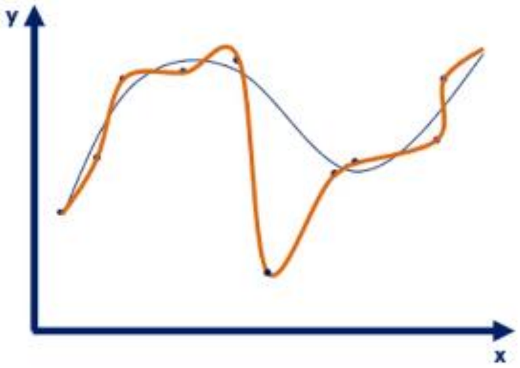
A **good** model



Captures the underlying logic of the dataset

- High train accuracy

An **overfitted** model



Captures all the noise, thus "missed the point"

- High train accuracy

Linear Models

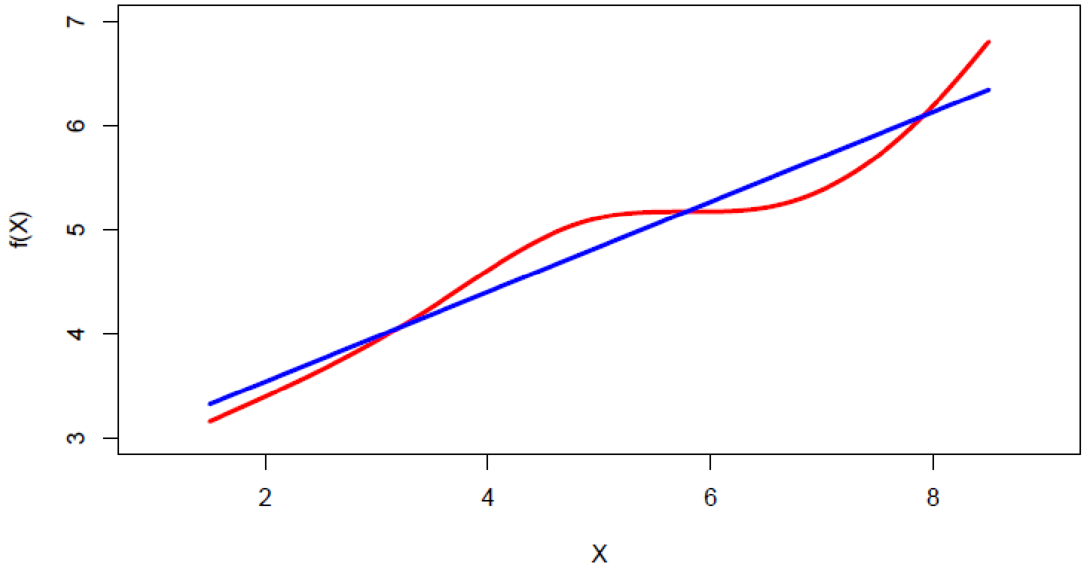
The *linear* model is an important example of a parametric model:

$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_p X_p.$$

- A linear model is specified in terms of $p + 1$ parameters $\beta_0, \beta_1, \dots, \beta_p$.
- We estimate the parameters by fitting the model to training data.
- Although it is *almost never correct*, a linear model often serves as a good and interpretable approximation to the unknown true function $f(X)$.

Linear Regression

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $X_1, X_2, \ldots X_p$ is linear.
- True regression functions are never linear!



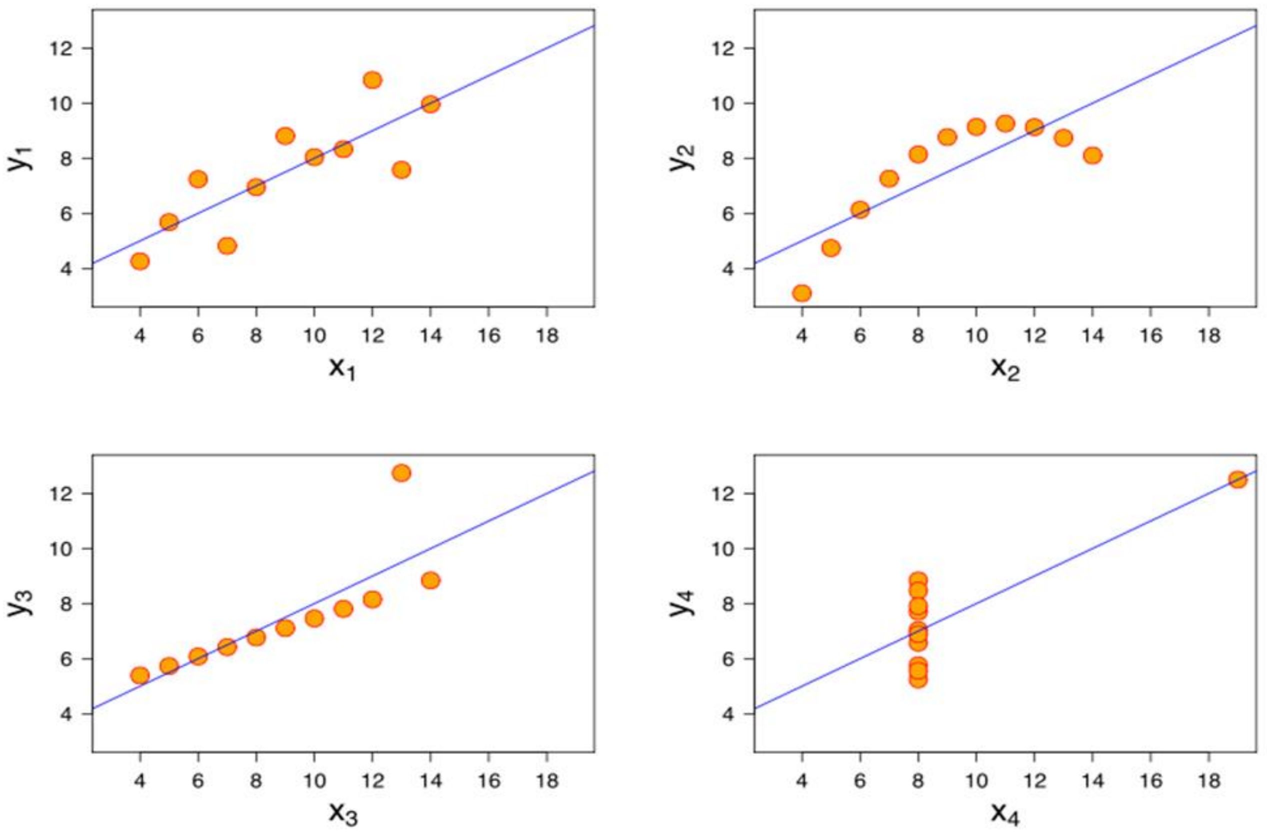
- although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

Assumptions of the OLS framework

1. Linearity $\gamma = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$
2. No endogeneity $\sigma_{X\varepsilon} = 0 : \forall\, x, \varepsilon$
3. Normality and homoscedasticity $\varepsilon \sim N(0, \sigma^2)$
4. No autocorrelation $\sigma_{\varepsilon_i \varepsilon_j} = 0 : \forall\, i \neq j$
5. No multicollinearity $\rho_{x_i x_j} \neq 1 : \forall\, i, j;\ i \neq j$

If any of these assumptions is violated, then you cannot use OLS

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	plus/minus 0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively



GAUSS MARKOV ASSUMPTIONS FOR REGRESSION

THE GAUSS-MARKOV ASSUMPTIONS

The following is a summary of the five Gauss-Markov assumptions that we used in this chapter. Remember, the first four were used to establish unbiasedness of OLS, whereas the fifth was added to derive the usual variance formulas and to conclude that OLS is best linear unbiased.

Assumption MLR.1 (Linear in Parameters)

The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u,$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters (constants) of interest and u is an unobserved random error or disturbance term.

Assumption MLR.2 (Random Sampling)

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i = 1, 2, \ldots, n\}$, following the population model in Assumption MLR.1.

Assumption MLR.3 (No Perfect Collinearity)

In the sample (and therefore in the population), none of the independent variables is constant, and there are no *exact linear* relationships among the independent variables.

Assumption MLR.4 (Zero Conditional Mean)

The error u has an expected value of zero given any values of the independent variables. In other words,

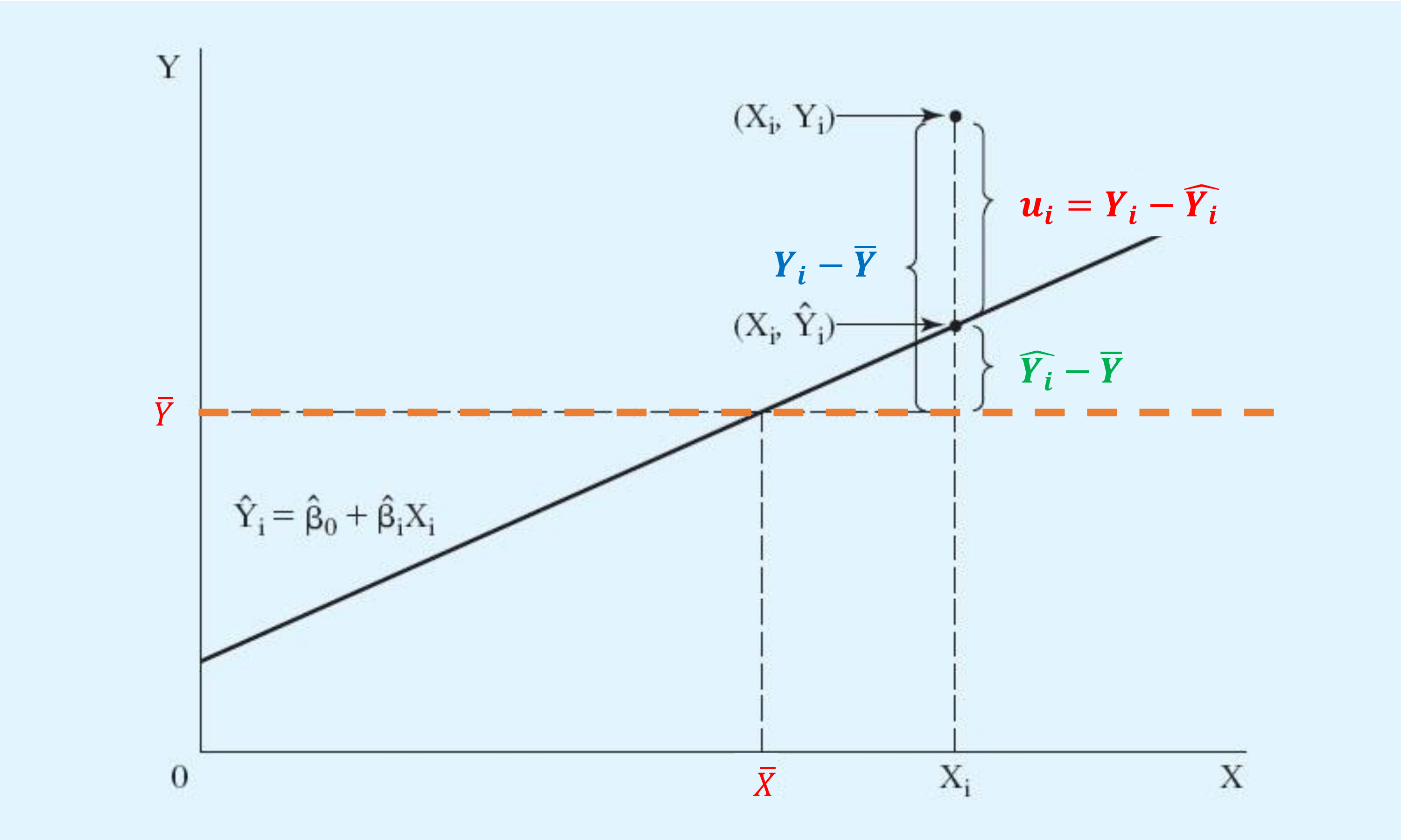
$$\mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0.$$

Assumption MLR.5 (Homoskedasticity)

The error u has the same variance given any value of the explanatory variables. In other words,

$$\mathrm{Var}(u|x_1, \ldots, x_k) = \sigma^2.$$

Decomposition of the variance in y



Measures of Variation

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

Total sum of squares,
Represents total sample variation in y

$$SSE \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Explained sum of squares,
Represents variation Explained by regression

$$SSR \equiv \sum_{i=1}^{n} \hat{u}_i{}^2$$

Residual sum of squares,
Represents variation **not** Explained by regression

Decomposition of Total Variation

$$SST = SSE + SSR$$

Total variation Explained part Unexplained part

Goodness-of-fit (R^2 or coefficient of determination)

How well does the explanatory variable explain the dependent variable?

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

R-squared measures the fraction of the total variation that is explained by the regression

Standard assumptions for the multiple regression model

Assumption MLR.1 **Linear in Parameters**

The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u, \quad [3.31]$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters (constants) of interest and u is an unobserved random error or disturbance term.

Assumption MLR.2 **Random Sampling**

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i = 1, 2, \ldots, n\}$, following the population model in Assumption MLR.1.

Standard assumptions for the multiple regression model

Assumption MLR.3 **No Perfect Collinearity**

In the sample (and therefore in the population), none of the independent variables is constant, and there are no *exact linear* relationships among the independent variables.

1. The assumption only rules out **perfect collinearity/correlation** between explanatory variables; **imperfect correlation is allowed**
2. If an explanatory variable is a perfect linear combination of other explanatory variables it is superfluous and may be eliminated
3. MLR.3 fails if $n < k + 1$. Intuitively, this makes sense: to estimate $k + 1$ parameters, we need at least $k + 1$ observations.

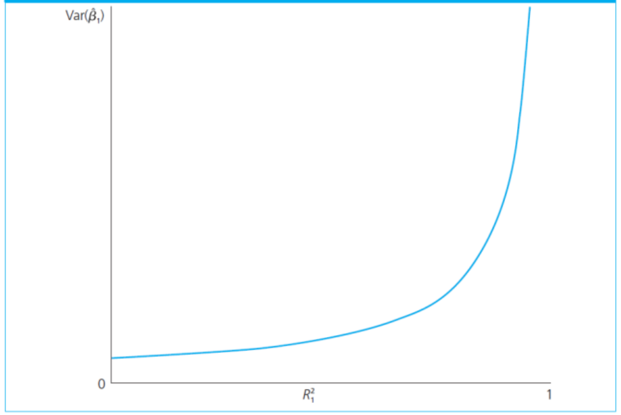
Detecting multicollinearity

Multicollinearity may be detected through Variance Inflation Factors:

$$VIF_j = 1/(1 - R_j^2)$$

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\mathrm{SST}_j(1 - R_j^2)} \longrightarrow \mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\mathrm{SST}_j} \cdot \mathrm{VIF}_j$$

FIGURE 3.1 $\mathrm{Var}(\hat{\beta}_1)$ as a function of R_1^2.



As an arbitrary rule of thumb, the variance inflation factor should not be larger than 10

Standard assumptions for the multiple regression model (cont.)

Assumption MLR.4 **Zero Conditional Mean**

The error u has an expected value of zero given any values of the independent variables. In other words,

$$E(u|x_1, x_2, \ldots, x_k) = 0. \quad [3.36]$$

- The value of the explanatory variables must contain no information about the mean of the unobserved factors
- In a multiple regression model, the **zero conditional mean assumption** is much **more likely to hold** because fewer things end up in the error.

Theorem 3.1 (Unbiasedness of OLS)

THEOREM 3.1

UNBIASEDNESS OF OLS

Under Assumptions MLR.1 through MLR.4,

$$\mathrm{E}(\hat{\beta}_j) = \beta_j,\ j = 0, 1, \ldots, k, \qquad [3.37]$$

for any values of the population parameter β_j. In other words, the OLS estimators are unbiased estimators of the population parameters.

Unbiasedness is an **average property in repeated samples**;
In a given sample, the estimates may still be far away from the true values!

Standard assumptions for the multiple regression model (cont.)

Assumption MLR.5	**Homoskedasticity**
	The error u has the same variance given any value of the explanatory variables. In other words, $\text{Var}(u \vert x_1, \ldots, x_k) = \sigma^2$.

- The value of the explanatory variables must contain no information about the **variance** of the unobserved factors

- Example: Wage equation

$$Var(u_i | educ_i, exper_i, tenure_i) = \sigma^2$$

This assumption may also be hard to justify in many cases

THEOREM 3.2

SAMPLING VARIANCES OF THE OLS SLOPE ESTIMATORS

Under Assumptions MLR.1 through MLR.5, conditional on the sample values of the independent variables,

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\mathrm{SST}_j(1 - R_j^2)}, \quad [3.51]$$

for $j = 1, 2, \ldots, k$, where $\mathrm{SST}_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the total sample variation in x_j, and R_j^2 is the R-squared from regressing x_j on all other independent variables (and including an intercept).

The sampling variability of the estimated regression coefficients depends on 4 things:

1. Variability of the unobserved factors (σ^2)
2. Variation in the explanatory variable $var(X_j)$ or SST_j
3. Number of observations n
4. Linear relationships among the independent variables (R^2)

Testing for Heteroskedasticity

There are many tests for heteroskedasticity; two popular:

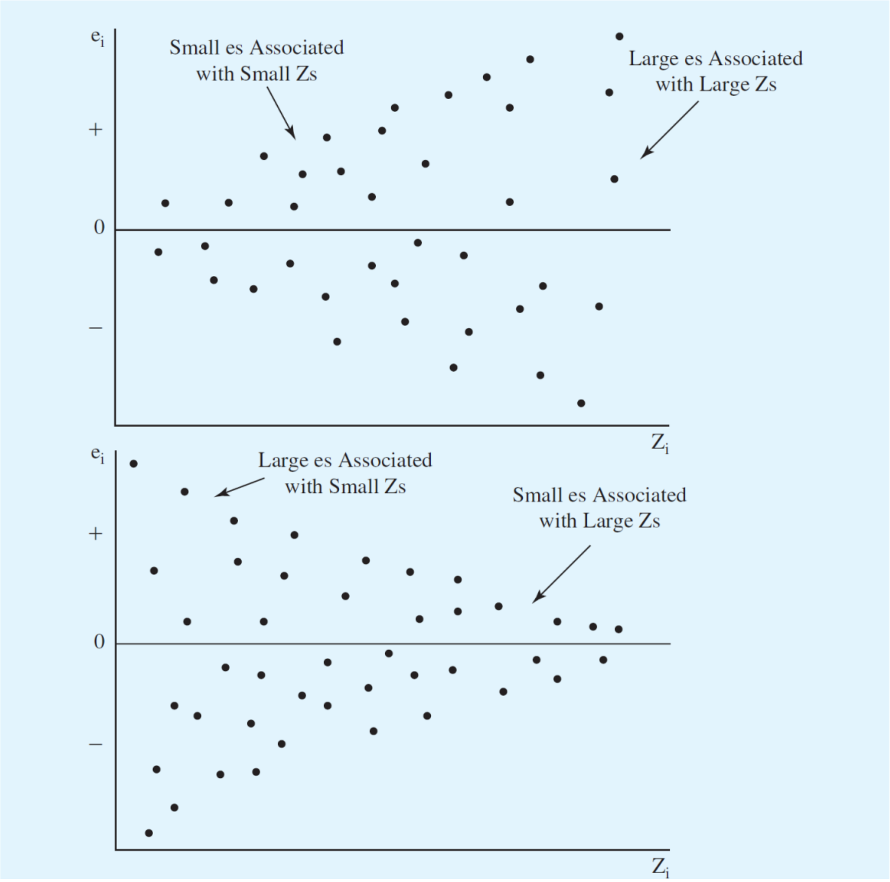
- Breusch-Pagan test
- White test

Before testing for heteroskedasticity, start with asking:

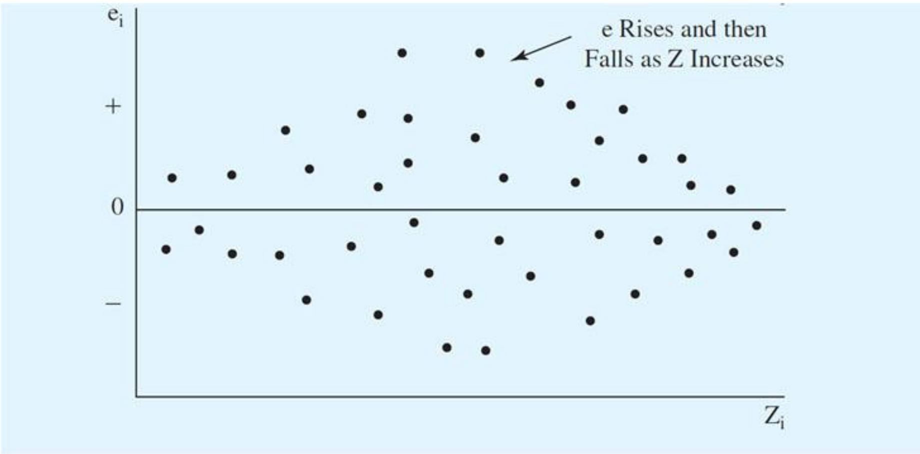
1. Are there any obvious specification errors?
2. Are there any early warning signs of heteroskedasticity?
3. Does a graph of the residuals show any evidence of heteroskedasticity?

Testing for Heteroskedasticity (cont’d)

Eyeballing Residuals for Possible Heteroskedasticity



If you plot the residuals of an equation with respect to a potential explanatory variable Z, a pattern in the residuals is an indication of possible heteroskedasticity.



The Breusch-Pagan Test for Heteroskedasticity:

Steps:

1. Estimate the model $\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u}$ by OLS, as usual. Obtain the squared OLS residuals $\hat{u}$
2. Run the regression in $\boxed{\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + error}$ Keep the R-squared from this regression $R^2_{\hat{u}^2}$
3. Form either the *F* statistic or the *LM* statistic and compute the *p*-value. If the *p*-value is sufficiently small, that is, below the chosen significance level, then we reject the **null hypothesis of homoskedasticity**.

$$\boxed{H_0 : Var(u|x_1, x_2, \ldots, x_k) = Var(u|\mathbf{x}) = \sigma^2} \longrightarrow H_0 : \delta_1 = \delta_2 = \cdots = \delta_k = 0$$

Regress squared residuals on all explanatory variables and test whether this regression has explanatory power.

$$F = \frac{R^2_{\hat{u}^2}/k}{1 - R^2_{\hat{u}^2}/(n-k-1)} \qquad LM = n \cdot R^2_{\hat{u}^2} \sim \chi^2_k$$

A large **F statistic** or a large **Lagrange multiplier** statistic, (LM) lead to rejection of the null hypothesis.

The White Test for Heteroskedasticity

Steps:

1. Estimate the model $\boxed{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u}$ by OLS, as usual. Obtain the squared OLS residuals $\hat{u}$
2. Run the regression in $\boxed{\begin{aligned}\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 \\ + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + error.\end{aligned}}$ Keep the R-squared from this regression $R^2_{\hat{u}^2}$
3. Form either the *F* statistic or the *LM* statistic and compute the *p*-value. If the *p*-value is sufficiently small, that is, below the chosen significance level, then we reject the **null hypothesis of homoskedasticity**.

$\boxed{H_0 : Var(u|x_1, x_2, \ldots, x_k) = Var(u|\mathbf{x}) = \sigma^2} \longrightarrow H_0 : \delta_1 = \delta_2 = \cdots = \delta_9 = 0$

Regress squared residuals on all explanatory variables, their squares, and interactions (here: example for k=3)

$$F = \frac{R^2_{\hat{u}^2}/k}{1 - R^2_{\hat{u}^2}/(n-k-1)} \qquad LM = n \cdot R^2_{\hat{u}^2} \sim \chi^2_k$$

A large **F statistic** or a large **Lagrange multiplier** statistic, (LM) lead to rejection of the null hypothesis.

Remedies for Heteroskedasticity

- If heteroskedasticity is found, the first thing to do is examine the equation carefully for specification errors.

- If there are no obvious specification errors, the heteroskedasticity is probably pure in nature and one of the following remedies should be considered.

1. Redefining the Variables
2. Heteroskedasticity-Corrected Standard Errors
3. Weighted Least Square Estimation!