# Single Linear Regression

Dr. Patrick Toche

Textbook:

**James H. Stock and Mark W. Watson**, *Introduction to Econometrics*, 4th Edition, Pearson.

Other references:

**Joshua D. Angrist and Jörn-Steffen Pischke**, *Mostly Harmless Econometrics: An Empiricist's Companion*, 1st Edition, Princeton University Press.
**Jeffrey M. Wooldridge**, *Introductory Econometrics: A Modern Approach*, 7th Edition, Cengage Learning.

The textbook comes with online resources and study guides. Other references will be given from time to time.

## In this lesson you will learn ...

- about linear regression models
- the ordinary least squares estimator
- measures of fit, R-squared, standard error of the regression
- BLUE and the assumptions of the Gauss-Markov theorem
- the sampling distribution of the OLS estimators

## Linear Regression with One Regressor

- **Problem of Regression:** A linear regression model relates one variable, $X$, to another variable, $Y$. The intercept and slope of the line relating $X$ and $Y$ are unknown characteristics of the population joint distribution of $X$ and $Y$. The econometric problem is to estimate the intercept and slope using a sample of data on these two variables.
- Linear regression is a statistical procedure that can be used for causal inference and for prediction.
  - **Causal inference:** using data to estimate the effect on an outcome of interest of an intervention that changes the value of another variable.
  - **Prediction:** using the observed value of some variable to predict the value of another variable.

## Population Regression Function

- **Population Regression Function:** The linear population regression function assumes that the expectation of $Y_i$ conditional on $X_i$ is linear in $X_i$:

$$\mathsf{E}[Y_i|X_i] = \beta_0 + \beta_1 X_i$$

- Let $u_i$ denote the error made by predicting $Y_i$ using its conditional mean $\mathsf{E}[Y_i|X_i]$.
- The actual regression relation can be written:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- $Y$ is the dependent variable, with realizations $Y_1, \ldots, Y_n$.
- $X$ is the independent variable, with realizations $X_1, \ldots, X_n$.
- $u$ is the error term, assumed random. The realized errors are the difference between the actual values of the dependent variable and their predicted values:

$$u_i = Y_i - \mathsf{E}[Y_i|X_i]$$

- $\beta_0$ and $\beta_1$ are the **coefficients** of the population regression: $\beta_0$ the intercept, $\beta_1$ the slope.
- The slope $\beta_1$ is the difference in $Y$ associated with a unit difference in $X$. The intercept is the value of the population regression line when $X = 0$: it is the point at which the population regression line intersects the $Y$ axis.

## Sample Regression Function

- **Sample Regression Function:** (*aka* sample regression line) The straight line constructed using the OLS estimators $\hat{\beta}_0 + \hat{\beta}_1 X$.
- **Predicted values:** The predicted value of $Y_i$ given $X_i$ is:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

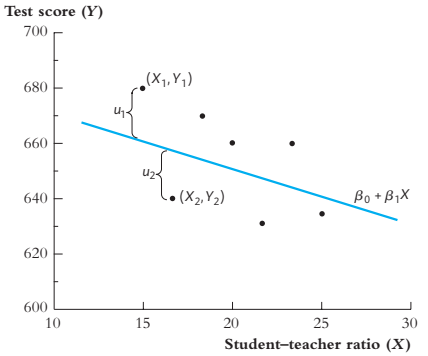
- **Residuals:** The residuals $\hat{u}_i$ are sample estimates of the population errors $u_i$, that is the difference between $Y_i$ and its predicted value $\hat{Y}_i$:

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

- **Estimators:** $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimators of the unknown coefficients of the population regression, $\beta_0, \beta_1$. Each estimate depends on the particular sample drawn.

## Scatter Plot with Regression Line

1. The population regression line has a negative slope, which means that districts with lower student-teacher ratios tend to have higher test scores.
2. The observations do not fall exactly on the population regression line. For example, the value of $Y$ for district $1$, $Y_1$, is greater than the population regression line: Test scores in district $1$ were better than predicted by the population regression line.
3. The error term for district $1$, $u_1$, is positive.
4. The value for district $2$, $Y_2$, is smaller than the population regression line. The error term $u_2$ is negative.




**Scatterplot of Student-Teacher Ratio and Test Scores.**

## Least Squares Estimators

- **OLS estimators of $\beta_0$ and $\beta_1$:**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2}$$
$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

- **OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$:**

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \ldots, n$$
$$\hat{u}_i = Y_i - \hat{Y}_i, \quad i = 1, \ldots, n$$

- The estimated intercept $\hat{\beta}_0$, slope $\hat{\beta}_1$, and residual $\hat{u}_i$ are computed from a sample of $n$ observations of $X_i$ and $Y_i, i = 1, \ldots, n$. These are estimates of the unknown true population intercept $\beta_0$, slope $\beta_1$, and error term $u_i$.

## Ordinary Least Squares Estimator

- The sample average, $\overline{Y}$, is the least squares estimator of the population mean, E$(Y)$. That is, $\overline{Y}$ minimizes the total squared estimation errors $\sum_{i=1}^n (Y_i - m)^2$ among all possible estimators $m$. The OLS estimator extends this idea to the linear regression model.
- **Ordinary least squares estimators of $\beta_0$ and $\beta_1$:** Estimators of the intercept and slope that minimize the sum of squared errors:

$$\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2 \rightarrow \min_{\beta_0, \beta_1}$$

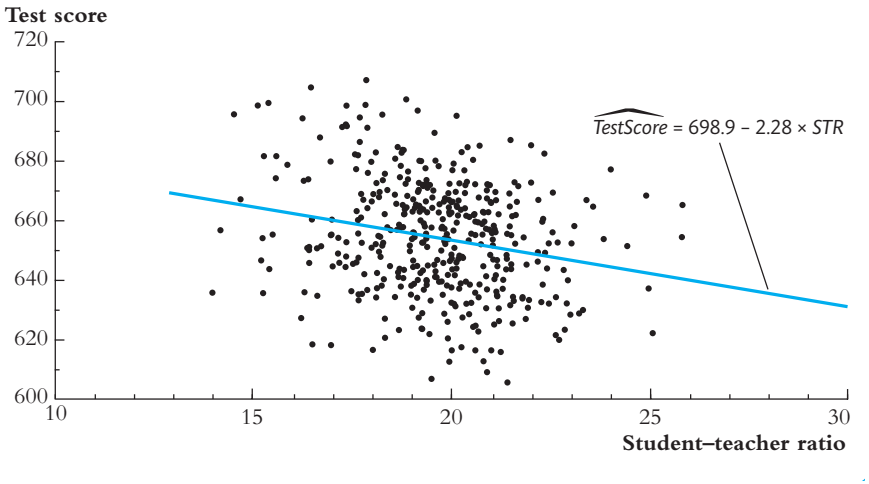
- **Estimation:** On the California School District Data, where $TestScore$ is the average test score in the $420$ districts and $STR$ is the student-teacher ratio, the estimated slope is $-2.28$ and intercept $698.9$:

$$\widehat{TestScore} = 698.9 - 2.28 \times STR$$

- **Prediction:** For a district with $20$ students per teacher,

$$\text{Predict}(TestScore | STR = 20) = 698.9 - 2.28 \times 20 = 653.3.$$

## Estimated Regression Line for the California School Districts Dataset




x

## Measures of Fit

- **Explained Sum of Squares (ESS):**
  The sum of squared deviations of the predicted value, $\hat{Y}_i$, from its average:

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

- **Total Sum of Squares (TSS):**
  The sum of squared deviations of $Y_i$ from its average:

$$TSS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

- **Regression $R^2$:**
  The fraction of the sample variance of $Y$ explained by $X$.

$$R^2 = \frac{ESS}{TSS}$$

- The $R^2$ ranges between $0$ and $1$.

## Measures of Fit

- **Sum of Squared Residuals (SSR):**
  The sum of squared residuals from a least squares regression:

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

- **Equivalent Definition of $R^2$:**

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

- Follows from the identity:

$$TSS = ESS + SSR$$

- If $X_i$ explains none of the variation of $Y_i$, then
  $\hat{\beta}_1 = 0$ and $\hat{Y}_i = \hat{\beta}_0 = \overline{Y}$, and $ESS = 0, SSR = TSS$, and $R^2 = 0$.
- If $X_i$ explains all of the variation of $Y_i$, then
  $Y_i = \hat{Y}_i$ for all $i$, and $\hat{u}_i = 0$, and $ESS = TSS$ and $R^2 = 1$.
- The $SSR$ is sometimes denoted $RSS$.

## Standard Error of the Regression

- **Standard Error of the Regression (SER):**
  An estimator of the standard deviation of the regression error measured in the units of the dependent variable.
- The SER is computed with the sample counterparts:

$$SER = s_{\hat{u}}$$

$$s_{\hat{u}}^2 = \frac{SSR}{n-2} = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2$$

- The adjusted divisor $n-2$ is an adjustment for the lost "degrees of freedom" when estimating the two coefficients $\beta_0$ and $\beta_1$.

## OLS Prediction

- **In-Sample Prediction:**
  The predicted value $\hat{Y}_i$ for the $i$th observation is the value of $Y$ predicted by the OLS regression line when $X$ takes on its value $X_i$ for that observation.
- **Out-of-Sample Prediction:**
  The predicted value $\hat{Y}$ for some value $X$ not in the estimation sample.
- Any prediction should be accompanied by an estimate of its accuracy, for instance $\hat{Y} \pm SER$.
- The California test score data regression reports $R^2 = 0.051$ and $SER = 18.6$. Thus, the regressor $STR$ explains $5.1\%$ of the variance of the dependent variable $TestScore$. The standard deviation of the regression residuals is large.

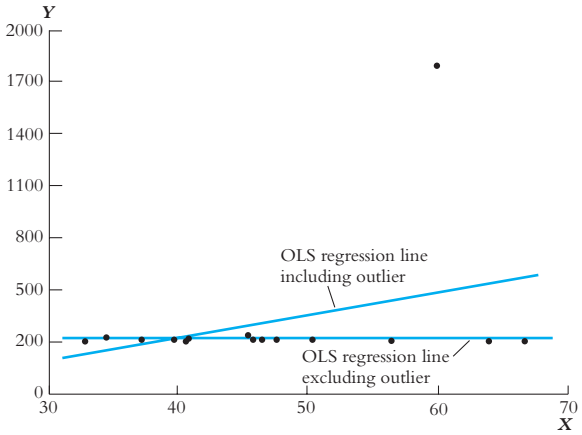
## Least Squares Assumptions for Causal Inference

- **Assumption 1: The conditional distribution of $u_i|X_i$ has zero mean:** $X$ must be randomly assigned or as-if randomly assigned.
- In a randomized controlled experiment with binary treatment, subjects are randomly assigned to the treatment group $X = 1$ or to the control group $X = 0$. It means that other factors contained in $u_i$ are unrelated to $X_i$.
- If $\mathsf{E}[u_i|X_i] = 0$, then $X_i$ and $u_i$ are uncorrelated.
- **Assumption 2: $X$ and $Y$ are independently and identically distributed across observations.**
- This assumption is a statement about how the sample is drawn. Not all samples are randomly selected. Biases can be introduced inadvertently. See the "hot hand fallacy" fallacy.
- **Assumption 3: $X$ and $Y$ have non-zero finite fourth moments.**

$$0 < \mathsf{E}[X_i^4] < \infty, \quad 0 < \mathsf{E}[Y_i^4] < \infty$$

- Finite kurtosis implies that outliers are very unlikely. OLS estimates are very sensitive to outliers. Commonly used distributions such as the normal distribution have four moments.

## Sensitivity of OLS to Large Outliers




The OLS regression line estimated with the outlier shows a strong positive relationship between X and Y, but the OLS regression line estimated without the outlier shows no relationship.

## Properties of Residuals

- **The sum of squared residuals is minimized:**

$$\hat{u}_i = Y_i - \hat{Y}_i, \qquad \sum_i \hat{u}_i^2 \to \min$$

  The residuals are the prediction errors of the estimates.
- **The mean residuals are zero:**

$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i = 0$$

- **The residuals are uncorrelated with the predictor:**

$$\mathsf{cov}(X_i, \hat{u}_i) = 0$$

- **The residuals are uncorrelated with the fitted values:**

$$\mathsf{cov}(\hat{Y}_i, \hat{u}_i) = 0$$

## Sampling Distribution

- **Sampling Distribution of $\overline{Y}$:**
  For large $n$, the central limit theorem states that this distribution is approximately normal, with $\overline{Y} \sim \mathcal{N}(\mu_Y, \sigma_Y^2/n)$. That is, $\mathsf{E}[\overline{Y}] = \mu_Y$
- **Sampling Distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$:**

$$\mathsf{E}[\hat{\beta}_0] = \beta_0, \qquad \mathsf{E}[\hat{\beta}_1] = \beta_1$$

For large $n$, the bivariate distribution of $(\hat{\beta}_0, \hat{\beta}_1)$ is approximately normal,

$$\hat{\beta}_0 \sim \mathcal{N}(\beta_0, \sigma^2_{\hat{\beta}_0}), \qquad \hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma^2_{\hat{\beta}_1})$$

where (formula valid for both homoskedastic and heteroskedastic errors):

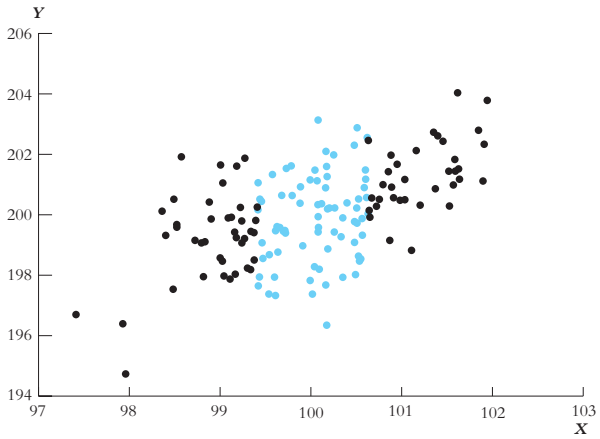
$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n} \frac{\mathsf{var}[(X_i - \mu_X)u_i]}{(\mathsf{var}[X_i])^2}$$

$$\sigma^2_{\hat{\beta}_0} = \frac{1}{n} \frac{\mathsf{var}[H_i u_i]}{(\mathsf{E}[H_i^2])^2},$$

$$H_i = 1 - \frac{\mu_X}{\mathsf{E}[X_i^2]} X_i$$

## Variance of $\hat{\beta}_1$ and Variance of $X$

The colored dots represent a set of $X_i$'s with a small variance. The black dots represent a set of $X_i$'s with a large variance. The regression line can be estimated more accurately with the black dots than with the colored dots.




The colored dots represent a set of $X_i$s with a small variance. The black dots represent a set of $X_i$s with a large variance. The regression line can be estimated more accurately with the black dots than with the colored dots.

## Summary

- The population regression line, $\beta_0 + \beta_1 X$, is the mean of $Y$ as a function of the value of $X$.
- The slope, $\beta_1$, is the expected difference in $Y$ values between two observations with $X$ values that differ by one unit.
- The intercept, $\beta_0$, determines the level of the regression line.
- The population regression line can be estimated using sample observations $(Y_i, X_i)$ by ordinary least squares (OLS) or another suitable method.
- The OLS estimators of the regression intercept and slope are denoted $\hat{\beta}_0$ and $\hat{\beta}_1$.
- The predicted value of $Y$ given $X$ is $\hat{\beta}_0 + \hat{\beta}_1 X$.
- The $R^2$ and standard error of the regression (SER) are measures of how close the values of $Y_i$ are to the estimated regression line. The $R^2$ is between $0$ and $1$, with a larger value indicating that the $Y_i$s are closer to the line.

## Summary

- A regression error is the deviation of the observed value from the true value. A regression residual is the difference between the observed value and the estimated value. The standard error of the regression estimates the standard deviation of the regression errors.
- There are three key assumptions for estimating causal effects using the linear regression model: (1) The regression errors, $u_i$, have a mean of $0$, conditional on the regressors $X_i$: $\mathsf{E}[u_i|X_i] = 0$. (2) The sample observations are i.i.d. random draws from the population; (3) Large outliers are unlikely: $X_i$ and $Y_i$ have nonzero finite fourth moments. If these assumptions hold, the OLS estimator $\hat{\beta}_1$ is (1) an unbiased estimator of the causal effect $\beta_1$, (2) consistent, and (3) normally distributed when the sample is large.

## Problems and Applications

Stock & Watson, Introduction (4th), Chapter 4, Review 3.

$SER$ and $R^2$ are "measures of fit" for a regression. Explain how $SER$ measures the fit of a regression. What are the units of $SER$? Explain how $R^2$ measures the fit of a regression. What are the units of $R^2$?

Stock & Watson, Introduction (4th), Chapter 4, Exercise 6.

Show that the first least squares assumption, $\mathrm{E}(u_i|X_i) = 0$, implies that $\mathrm{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i$.

Stock & Watson, Introduction (4th), Chapter 4, Exercise 14.

Show that the sample regression line passes through the point $(\overline{X}, \overline{Y})$.

## Problems and Applications

Stock & Watson, Introduction (4th), Chapter 4, Exercise 1.

A researcher, using data on class size (CS) and average test scores from $100$ third-grade classes, estimates the OLS regression:

$$\widehat{TestScore} = 520.4 - 5.82 \times CS, \quad R^2 = 0.08, \quad SER = 11.5$$

1. A classroom has $22$ students. What is the regression's prediction for that classroom's average test score?
2. Last year a classroom had $19$ students, and this year it has $23$ students. What is the regression's prediction for the change in the classroom average test score?
3. The sample average class size across the $100$ classrooms is $21.4$. What is the sample average of the test scores across the $100$ classrooms?
4. What is the sample standard deviation of test scores across the $100$ classrooms?

## Problems and Applications

Stock & Watson, Introduction (4th), Chapter 4, Exercise 2.

Suppose a random sample of $200$ 20-year-old men is selected from a population and their heights and weights are recorded. A regression of weight on height yields:

$$\widehat{Weight} = -99.41 + 3.94 \times Height, \quad R^2 = 0.81, \quad SER = 10.2$$

where $Weight$ is measured in pounds and $Height$ is measured in inches.

1. What is the regression's weight prediction for someone who is $70$ in. tall? $65$ in. tall? $74$ in. tall?
2. A man has a late growth spurt and grows $1.5$ in. over the course of a year. What is the regression's prediction for the increase in this man's weight?
3. Suppose that instead of measuring weight and height in pounds and inches, these variables are measured in centimeters and kilograms. What are the regression estimates from this new kilogram–centimeter regression? (Give all results, estimated coefficients, $R^2$, and $SER$.)

## Problems and Applications

Stock & Watson, Introduction (4th), Chapter 4, Exercise 12.

1. Show that the regression $R^2$ in the regression of $Y$ on $X$ is the squared value of the sample correlation between $X$ and $Y$. That is, show that $R^2 = r_{XY}^2$.
2. Show that the $R^2$ from the regression of $Y$ on $X$ is the same as the $R^2$ from the regression of $X$ on $Y$.
3. Show that $\hat{\beta}_1 = r_{XY}(s_Y/s_X)$, where $r_{XY}$ is the sample correlation between $X$ and $Y$, and $s_X$ and $s_Y$ are the sample standard deviations of $X$ and $Y$.

## Deriving the OLS Estimators

- To illustrate the principle, ignore $\beta_0$. Minimize the sum of squared residuals:

$$\min_{\beta_1} \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2$$

- Calculate the derivative with respect to the parameter $\beta_1$ and set it equal to zero:

$$\frac{d}{d\hat{\beta}_1} \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 X_i)^2 = 0$$

$$-2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 X_i) X_i = 0$$

$$\frac{1}{n} \sum_{i=1}^{n} X_i Y_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} X_i^2 = 0$$

$$\implies \hat{\beta}_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} X_i Y_i}{\frac{1}{n} \sum_{i=1}^{n} X_i^2}$$

## Deriving the OLS Estimators

- Minimize the sum of squared residuals:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- Calculate the derivative with respect to the parameter $\beta_1$ and set it equal to zero:

$$\frac{\partial}{\partial\hat{\beta}_1} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = 0$$

$$-2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i = 0$$

$$\frac{1}{n} \sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \frac{1}{n} \sum_{i=1}^{n} X_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} X_i^2 = 0$$

$$\implies \hat{\beta}_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \frac{1}{n} \sum_{i=1}^{n} X_i}{\frac{1}{n} \sum_{i=1}^{n} X_i^2}$$

- Thus, the partial derivative with respect to $\hat{\beta}_1$ yields a linear relation between $\hat{\beta}_0$ and $\hat{\beta}_1$.

## Deriving the OLS Estimators

- Calculate the derivative with respect to the parameter $\beta_0$ and set it equal to zero:

$$\frac{\partial}{\partial\hat{\beta}_0} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = 0$$

$$-2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\frac{1}{n} \sum_{i=1}^{n} Y_i - \hat{\beta}_0 \frac{1}{n} \sum_{i=1}^{n} 1 - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^{n} X_i = 0$$

$$\implies \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

- The partial derivative with respect to $\hat{\beta}_1$ has an $X_i$ term which here is simply $1$.
- Putting it together:

$$\hat{\beta}_1 = \frac{\frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

## Keywords

causal inference prediction regressor dependent variable independent variable population regression function intercept slope coefficients parameters error term ordinary least squares (OLS) estimators sample regression function predicted value residual regression R-squared explained sum of squares (ESS) total sum of squares (TSS) sum of squared residuals (SSR) standard error of the regression (SER) in-sample prediction out-of-sample prediction Gauss-Markov theorem