Class 3: Error Variance & OLS Variance

MFE 402

Dan Yavorsky

Last Class

- Discussed the CEF $m(X) = \mathbb{E}[Y|X]$ as a property of a joint distribution
  - It is our quantity of interest because it is the best (min MSE) regression function for Y
- Introduced the Linear CEF Model: $Y = X'\beta + e$ with $\mathbb{E}[e|X] = 0$
- Provided three reasons why the Linear CEF Model may be a good model
  - The linear regression function $(X'\beta)$ *is* the CEF in discrete or MVN cases
  - The linear regression function $(X'\beta)$ is the best *linear* predictor of Y given X
  - The linear regression function $(X'\beta)$ is the best *linear* approximation to the CEF
- Derived β and provided two approaches to find (the same) estimator $\hat{\beta}$ for β
  - As a Method of Moments estimator for $\mathbf{Q}_{XX}^{-1}\mathbf{Q}_{XY}$ or $S(\beta)$
  - As the minimizer of the sum of squared errors (where the OLS estimator $\hat{\beta}$ gets its name)

Topics for Today

1. **OLS Estimator Mean**
2. CEF Error Variance
3. **OLS Estimator Variance**
4. Residuals
5. Projections
6. Estimators of CEF Error Variance
7. **Estimators of OLS Estimator Variance**
8. Coefficient of Determination (**R-Squared**)
9. Computation in `R`

Mean of $\hat{\beta}$

Unbiasedness of the OLS Estimator

$\hat{\beta}$ is a function of random variables X and Y and so it is a random variable.

This means that it has a distribution, which we call the **sampling distribution** of $\hat{\beta}$.

If the mean of the sampling distribution is centered over the value we seek to estimate, then the estimator is said to be **unbiased**.

$$\begin{aligned}
\mathbb{E}[\hat{\beta}|\mathbf{X}] &= \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X}] \\
&= \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{e})|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{e}|\mathbf{X}] \\
&= \beta + \mathbf{0}
\end{aligned}$$

Notice this requires our assumption about the error term: $\mathbb{E}[e|X] = 0$

Use LIE to find that $\mathbb{E}[\hat{\beta}] = \mathbb{E}[\mathbb{E}[\hat{\beta}|\mathbf{X}]] = \mathbb{E}[\beta] = \beta$

$\Rightarrow$ $\hat{\beta}$ is an **unbiased** estimator for β.

Error Variance

Unconditional Error Variance

An important measure of the dispersion about the CEF function is the unconditional (on X) variance of the CEF error e:

$$\sigma^2 = \text{var}(e) = \mathbb{E}\left[(e - \mathbb{E}[e])^2\right] = \mathbb{E}[e^2]$$

Econometricians have several names for this:

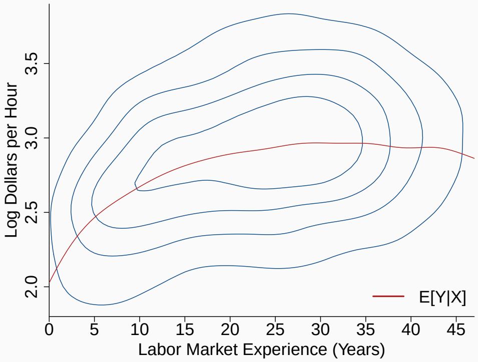
- Error variance
- Variance of the regression error
- Regression variance

σ^2 measures the amount of variation in Y which is not "explained" by the CEF

Note that $\sigma_Y^2 = \text{Var}(m(X)) + \text{Var}(e) \geq \sigma^2$, with equality only when $\text{Corr}(X, Y) = 0$ or equivalently when $m(X)$ is a constant.

Recall these two Joint Distributions

$F_{X,Y}(X, Y)$



$F_{X,e}(X, e)$

Adding Regressors Changes the Regression Variance

Think of Y as the combination of an "explained" (by X) portion and an unexplained (by X) portion:

$$Y = \underbrace{m(X)}_{\text{explained}} + \underbrace{e}_{\text{unexplained}}$$

Changing the conditioning information (the X's in X)

- changes the CEF $m(X)$
- and thus changes the error e
- and thus changes the variance of the error σ^2

The relationship is monotonic: more info $\Rightarrow$ smaller σ^2

Conditional Error Variance

Consider the conditional variance of Y given $X = x$:

$$\sigma_Y^2(x) = \text{Var}(Y|X = x) = \mathbb{E}\left[(Y - \mathbb{E}[Y|X = x])^2 \,|X = x\right]$$

The conditional variance $\sigma_Y^2(x)$ is a function of the conditioning variables (the X's), much like how CEF $m(x)$ is a function of the X vector.

Now, consider the conditional variance of the CEF error e given $X = x$:

$$\sigma_e^2(x) = \text{Var}(e|X = x) = \mathbb{E}[e^2|X] = \mathbb{E}\left[(Y - \mathbb{E}[Y|X = x])^2 \,|X = x\right]$$

They're equal! $\sigma_e^2(x) = \sigma_Y^2(x) = \sigma^2(x)$

Mean-Variance Representation of the CEF

$\sigma^2(x)$ is in a different unit of measurement than Y. To convert it to the same unit of measure, define the conditional standard deviation: $\sigma(x) = \sqrt{\sigma^2(x)}$.

Consider the re-scaled error $u = e/\sigma(x)$. Notice:

$$\begin{aligned}
\mathbb{E}[u|X] &= \mathbb{E}[e/\sigma(x)|X] = (1/\sigma(x))\mathbb{E}[e|X] = 0 \\
\mathrm{Var}(u|X) &= \mathbb{E}[u^2|X] = \mathbb{E}[e^2/\sigma^2(x)|X] = (1/\sigma^2(x))\mathbb{E}[e^2|X] = 1
\end{aligned}$$

So we can write the CEF Model in a mean-variance representation:

$$Y = m(X) + \sigma(X)u$$

Most econometric studies focus on $m(x)$ and either treat $\sigma(x)$ as a constant ($\sigma(x) = \sigma$) or treat it as a nuisance parameter by ignoring it.

Homoskedasticity & Heterskedasticity

Two terms are used to summarize assumptions about the conditional variance:

- The error is **homoskedastic** if the conditional variance does not depend on X: $\sigma^2(x) = \sigma^2$
- The error is **heteroskedastic** if the conditional variance depends on X: $\sigma^2(x)$
  - It is not entirely correct to think of heteroskedasticity as "varying by observation" because the conditional variance is a function of X, not i.

Heteroskedasticity is typically a *more correct* model specification!

Homoskedasticity is useful for:

- Simplifying calculations
- Teaching and learning
- Understanding a specific, unusual, and exceptional special case
- Understanding the default output of most statistical software packages

Variance of OLS Estimator

Variance of a Random Vector

Let $Z = [Z_1, Z_2, Z_3]'$ be a random vector. Then the variance of Z is defined as the (variance-) covariance matrix:

$$
\begin{aligned}
\text{Var}(Z) &= \mathbb{E}[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])'] \\
&= \mathbb{E}\begin{bmatrix} (Z_1 - \mathbb{E}[Z_1])(Z_1 - \mathbb{E}[Z_1]) & (Z_1 - \mathbb{E}[Z_1])(Z_2 - \mathbb{E}[Z_2]) & (Z_1 - \mathbb{E}[Z_1])(Z_3 - \mathbb{E}[Z_3]) \\ (Z_2 - \mathbb{E}[Z_2])(Z_1 - \mathbb{E}[Z_1]) & (Z_2 - \mathbb{E}[Z_2])(Z_2 - \mathbb{E}[Z_2]) & (Z_2 - \mathbb{E}[Z_2])(Z_3 - \mathbb{E}[Z_3]) \\ (Z_3 - \mathbb{E}[Z_3])(Z_1 - \mathbb{E}[Z_1]) & (Z_3 - \mathbb{E}[Z_3])(Z_2 - \mathbb{E}[Z_2]) & (Z_3 - \mathbb{E}[Z_3])(Z_3 - \mathbb{E}[Z_3]) \end{bmatrix} \\
&= \begin{bmatrix} \text{Var}(Z_1) & \text{Cov}(Z_1, Z_2) & \text{Cov}(Z_1, Z_3) \\ \text{Cov}(Z_2, Z_1) & \text{Var}(Z_2) & \text{Cov}(Z_2, Z_3) \\ \text{Cov}(Z_3, Z_1) & \text{Cov}(Z_3, Z_2) & \text{Var}(Z_3) \end{bmatrix}
\end{aligned}
$$

Additionally,

- $\text{Var}(Z) = \mathbb{E}[ZZ'] - \mathbb{E}[Z]\mathbb{E}[Z]'$
- $\text{Var}(a + bZ) = b\text{Var}(Z)b'$ for any scalars or vectors a and b

Variance of the OLS Estimator

$$\text{Recall:} \quad \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Define the $k \times k$ conditional variance-covariance matrix of the OLS estimator to be:

$$\begin{aligned}
\mathbf{V}_{\hat{\beta}} = \text{Var}(\hat{\beta}|\mathbf{X}) &= \text{Var}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X}\right) \\
&= \text{Var}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{e})|\mathbf{X}\right) \\
&= \text{Var}\left(\beta|\mathbf{X}\right) + \text{Var}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}|X\right) \\
&= 0 + \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\text{Var}\left(\mathbf{e}|\mathbf{X}\right)\ \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbb{E}[\mathbf{e}\mathbf{e}'|\mathbf{X}]\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

Variance of the OLS Estimator (cont.)

Let's explore the "meat" of the sandwich. Define the $n \times n$ matrix $\mathbf{D}$:

$$\mathbf{D} = \text{Var}(\mathbf{e}|X) = \mathbb{E}[\mathbf{e}\mathbf{e}'|X] = \begin{bmatrix} \sigma_1^2(x) & 0 & \cdots & 0 \\ 0 & \sigma_2^2(x) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2(x) \end{bmatrix}$$

Because

- the i^{th} diagonal element of $\mathbf{D}$ is $\mathbb{E}[e_i^2|\mathbf{X}] = \mathbb{E}[e_i^2|X_i] = \sigma_i^2(x)$
- the ij^{th} off-diagonal element of $\mathbf{D}$ is $\mathbb{E}[e_i e_j|\mathbf{X}] = \mathbb{E}[e_i|X_i]\mathbb{E}[e_j|X_j] = 0$ by independence

Variance of the Estimator Under Homoskedasticity

Under an assumption of homoskedasticity, we have $\sigma^2(x) = \mathbb{E}[e_i^2|\mathbf{X}] = \sigma^2$ for $i = 1, \ldots, n$

Then D simplifies to:

$$\mathbf{D} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \sigma^2 \mathbf{I}_n$$

And the variance-covariance matrix of the OLS estimator simplifies to

$$\mathbf{V}_{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathbb{E}[\mathbf{e}\mathbf{e}'|\mathbf{X}]\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\sigma^2\mathbf{I}_n\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Gauss-Markov Theorem

For the homoskedastic Linear Regression Model

$$Y = X'\beta + e \quad \text{with} \quad \mathbb{E}[e|X] = 0 \quad \text{and} \quad \mathbb{E}[\mathbf{ee}'|\mathbf{X}] = \sigma^2 \mathbf{I}_n$$

the OLS estimator $\hat{\beta}$ is the Best (lowest variance) Linear Unbiased Estimator (BLUE).

In other words, suppose $\tilde{\beta} = \mathbf{A}'\mathbf{y}$ is unbiased, then $\text{Var}(\tilde{\beta}|\mathbf{X}) \geq \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

A new paper by Hansen (2022) in *Econometrica* shows $\hat{\beta}$ is BUE – future textbooks might call it the Gauss-Markov-Hansen Theorem!

Gauss-Markov Theorem Proof

$$\mathbb{E}[\tilde{\beta}|\mathbf{X}] = \mathbf{A}'\mathbb{E}[\mathbf{y}|\mathbf{X}] = \mathbf{A}'\mathbf{X}\beta \quad \Rightarrow \quad \mathbf{A}'\mathbf{X} = \mathbf{I}_n$$
$$\text{Var}(\tilde{\beta}|\mathbf{X}) = \text{Var}(\mathbf{A}'\mathbf{y}|\mathbf{X}) = \mathbf{A}'\mathbf{D}\mathbf{A} = \sigma^2\mathbf{A}'\mathbf{A}$$

What's left to show is that $\mathbf{A}'\mathbf{A} \geq (\mathbf{X}'\mathbf{X})^{-1}$

Define $\mathbf{C} = \mathbf{A} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ such that $\mathbf{A} = \mathbf{C} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ and notice that $\mathbf{X}'\mathbf{C} = \mathbf{0}$. Then:

$$\begin{aligned}
\mathbf{A}'\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1} &= \left(\mathbf{C} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right)' \left(\mathbf{C} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right) - (\mathbf{X}'\mathbf{X})^{-1} \\
&= \mathbf{C}'\mathbf{C} + \mathbf{C}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1} \\
&= \mathbf{C}'\mathbf{C} \\
&\geq \mathbf{0} \qquad \text{meaning positive semi-definite}
\end{aligned}$$

Residuals

OLS Fitted Values and Residuals

As a by-product of estimation, we obtain two useful quantities for each observation i:

- $\hat{Y}_i = X_i'\hat{\beta}$ are fitted value (not predicted values)
- $\hat{e}_i = Y_i$ - $\hat{Y}_i$ are residuals (not errors)

Thus, we have:

$$Y_i = X_i'\hat{\beta} + \hat{e}_i \quad \text{or equivalently} \quad \mathbf{y} \text{ - } \mathbf{X}\hat{\beta} + \hat{\mathbf{e}}$$

which, to be clear, is different from

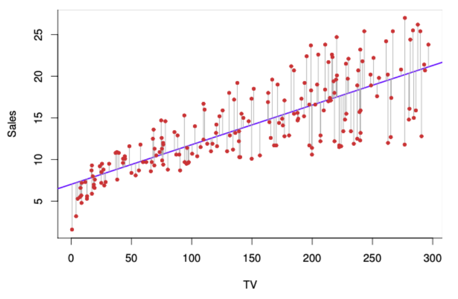
$$Y_i = X_i'\beta + e_i \quad \text{or equivalently} \quad \mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

Note that:

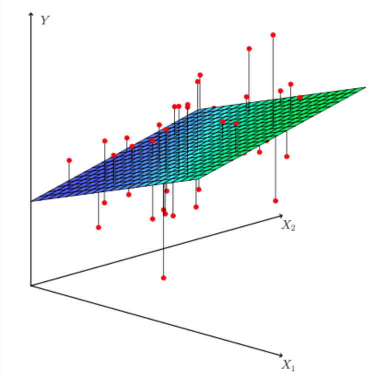
- The error e_i is unobservable
- The residual $\hat{e}_i$ is a statistic (a function of the data) and thus observable
- We will use $\hat{e}_i$ as an estimator of e_i, hence the hat notation

Visualizing Residuals

When $X \in \mathbb{R}$



When $X \in \mathbb{R}^2$



Two Algebraic Properties of Residuals

The sample correlation between the regressors and the residuals is the zero vector:

$$\sum_{i=1}^{n} X_i \hat{e}_i = \mathbf{X}'\hat{\mathbf{e}} = \mathbf{0}$$

When X_i contains a constant for the intercept, then

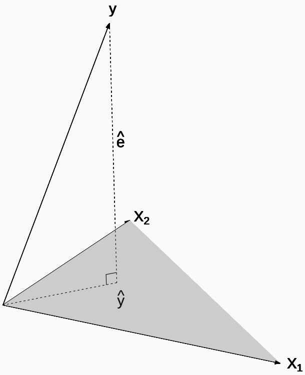
$$\sum_{i=1}^{n} \hat{e}_i = 0$$

Notice:

- these offer a nice parallel to the moment conditions $\mathbb{E}[Xe] = 0$ and $\mathbb{E}[e] = 0$
- in fact, they are the first-order conditions when solving for the OLS estimator
- so, you could derive $\hat{\beta}$ by using method of moments with these moment conditions

Projection Matrices

Visualizing Least Squares as Projection



Column vectors:

- $\mathbf{y}$ is the length-n vector in $\mathbb{R}^n$
- The k regressors (X_j for $j = 1, \ldots, k$) are also length-n vectors in $\mathbb{R}^n$
- When rank($\mathbf{X}$) $= k$, the k regressors are linearly independent and span the subspace $\mathbb{R}^k$
- $\hat{\mathbf{y}}$ is the projection of $\mathbf{y}$ onto the subspace spanned by the regressors
- $\hat{\mathbf{e}}$ is the residual vector, a project of $\mathbf{y}$ onto the n - k subspace orthogonal to the subspace spanned by the regressors

Projection Matrix

Define the $n \times n$ projection matrix $\mathbf{P}$:

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

This is sometimes called the “hat” matrix because

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

Some important properties:

- $\mathbf{P}$ is symmetric ($\mathbf{P}' = \mathbf{P}$)
- $\mathbf{P}$ is idempotent ($\mathbf{P}\mathbf{P} = \mathbf{P}$)
- $\mathbf{P}$ has k eigenvalues equaling 1 and $n - k$ equaling 0
- $\text{trace}(\mathbf{P}) = k$

Annihilator Matrix

Define the $n \times n$ annihilator matrix $\mathbf{M}$:

$$\mathbf{M} = \mathbf{I}_n - \mathbf{P} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

It gets its name from the calculation of $\mathbf{MX}$:

$$\mathbf{MX} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X} = \mathbf{X} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{X} - \mathbf{X}\mathbf{I}_n = \mathbf{0}$$

A useful relationship with $\mathbf{M}$ is:

$$\mathbf{My} = \mathbf{y} - \mathbf{Py} = \mathbf{y} - \hat{\mathbf{y}} = \hat{\mathbf{e}}$$
$$\mathbf{My} = M(X\beta + \mathbf{e}) = \mathbf{MX}\beta + \mathbf{Me} = \mathbf{Me}$$

$\mathbf{M}$ is symmetric, idempotent, and has trace$(\mathbf{M}) = n - k$

Estimate Error Variance

Estimate the Error Variance

The unconditional error variance is a moment:

$$\sigma^2 = \mathbb{E}[e^2]$$

So a natural (analog, plug-in, or method of moments) estimator would be:

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

But the errors e_i are not observed, so we first estimate them with the residuals $\hat{e}_i$:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2$$

$\hat{\sigma}^2 \leq \tilde{\sigma}^2$

The feasible estimator ($\hat{\sigma}^2$) is smaller than the idealized estimator ($\tilde{\sigma}^2$):

Rewrite the feasible estimator as:

$$\begin{aligned}\hat{\sigma}^2 &= n^{-1}\hat{\mathbf{e}}'\hat{\mathbf{e}} \\ &= n^{-1}(\mathbf{Me})'\mathbf{Me} \\ &= n^{-1}\mathbf{e}'\mathbf{Me}\end{aligned}$$

Then take the difference:

$$\begin{aligned}\tilde{\sigma}^2 - \hat{\sigma}^2 &= n^{-1}\mathbf{e}'\mathbf{e} - n^{-1}\mathbf{e}'\mathbf{Me} \\ &= n^{-1}\mathbf{e}'(\mathbf{I} - \mathbf{M})\mathbf{e} \\ &= n^{-1}\mathbf{e}'\mathbf{Pe}\end{aligned}$$

Since $\mathbf{e}'\mathbf{Pe}$ is quadratic form, $\mathbf{e}'\mathbf{Pe} \geq 0$ which implies $\hat{\sigma}^2 \leq \tilde{\sigma}^2$

$\hat{\sigma}^2$ is biased

Recall two special properties of the trace operator:

- $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$ when $\dim(\mathbf{A}) = \dim(\mathbf{B}')$
- $\text{tr}(\mathbf{A}) = \sum_{i=1}^{k} \lambda_k$ for square $k \times k$ matrix $\mathbf{A}$ and eigenvalues λ_i for $i = 1, \ldots, k$.

Then we can show:

$$\hat{\sigma}^2 = \frac{1}{n}\mathbf{e}'\mathbf{M}\mathbf{e} = \frac{1}{n}\text{tr}(\mathbf{e}'\mathbf{M}\mathbf{e}) = \frac{1}{n}\text{tr}(\mathbf{M}\mathbf{e}\mathbf{e}')$$

Taking the conditional expected value:

$$\mathbb{E}[\hat{\sigma}^2|\mathbf{X}] = \frac{1}{n}\text{tr}\left(\mathbb{E}[\mathbf{M}\mathbf{e}\mathbf{e}'|\mathbf{X}]\right) = \frac{1}{n}\text{tr}\left(\mathbf{M}\mathbb{E}[\mathbf{e}\mathbf{e}'|\mathbf{X}]\right)$$

$\hat{\sigma}^2$ is biased (Cont.)

Under an assumption of homoskedasticity, $\mathbb{E}[\mathbf{ee}'|X] = \sigma^2\mathbf{I}_n$ so that

$$\mathbb{E}[\hat{\sigma}^2|\mathbf{X}] = \frac{1}{n}\text{tr}\left(\mathbf{M}\mathbb{E}[\mathbf{ee}'|\mathbf{X}]\right) = \frac{1}{n}\text{tr}(\mathbf{M})\sigma^2 = \frac{n-k}{n}\sigma^2$$

The "fix" is to propose an unbiased estimator s^2

$$s^2 = \frac{n}{n-k}\hat{\sigma}^2 = \frac{n}{n-k}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2\right) = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n-k}$$

Terminology:

- The `summary()` command in R calls $\sqrt{s^2}$ the **Residual Standard Error**
- Some textbooks call $\sqrt{s^2}$ the **Standard Error of the Regression**

Estimate Error Variance

OLS Covariance Matrix Estimation Under Homoskedasticity

Under the assumption of homoskedasticity, the var-cov matrix of the OLS estimator is:

$$\mathbf{V}^0_{\hat{\beta}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

The most common estimator for $\mathbf{V}^0_{\hat{\beta}}$ replaces σ_2 with its unbiased estimator s^2:

$$\hat{\mathbf{V}}^0_{\hat{\beta}} = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

$\hat{\mathbf{V}}^0_{\hat{\beta}}$ is conditionally unbiased for $\mathbf{V}^0_{\hat{\beta}}$ under homoskedasticity:

$$\mathbb{E}\left[\hat{\mathbf{V}}^0_{\hat{\beta}}|\mathbf{X}\right] = \mathbb{E}[s^2|\mathbf{X}](\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{V}^0_{\hat{\beta}}$$

OLS Covariance Matrix Estimation Under Heteroskedasticity

Without the assumption of homoskedasticity, the var-cov matrix of $\hat{\beta}$ is

$$\mathbf{V}_{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{D}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n} X_i X_i' \mathbb{E}[e_i^2|X]\right)(\mathbf{X}'\mathbf{X})^{-1}$$

An idealized estimator would be:

$$\hat{\mathbf{V}}_{\hat{\beta}}^{\text{ideal}} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n} X_i X_i' e_i^2\right)(\mathbf{X}'\mathbf{X})^{-1}$$

Two feasible estimators (called White, Eicker-White, robust, heteroskedasticity-consistent) are:

$$\hat{\mathbf{V}}_{\hat{\beta}}^{\text{HC0}} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n} X_i X_i' \hat{e}_i^2\right)(\mathbf{X}'\mathbf{X})^{-1}$$

$$\hat{\mathbf{V}}_{\hat{\beta}}^{\text{HC1}} = \frac{n}{n-k}(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n} X_i X_i' \hat{e}_i^2\right)(\mathbf{X}'\mathbf{X})^{-1}$$

Standard Errors of the OLS Estimator

A **standard error** $s(\hat{\beta})$ for an estimator $\hat{\beta}$ is an estimator of the standard deviation of the sampling distribution of $\hat{\beta}$.

When β is a vector with estimator $\hat{\beta}$ and variance-covariance matrix estimator $\hat{\mathbf{V}}_{\hat{\beta}}$, the standard errors are the square roots of the diagonal elements of $\hat{\mathbf{V}}_{\hat{\beta}}$:

$$s(\hat{\beta}_j) = \sqrt{\left[\hat{\mathbf{V}}_{\hat{\beta}}\right]_{jj}}$$

Variance of the Estimator Under Homoskedasticity (one X)

Suppose X is univariate. Define the sample variance of X as $s_X^2 = (n - 1)^{-1} \sum_{i=1}^{n}(X_i - \bar{X})^2$.

Then for a simple (X is a scalar random variable) linear regression model, the standard deviation of the slope coefficient can be written as:

$$\left(s(\hat{\beta}_1)\right)^2 = \mathrm{Var}(\hat{\beta}_1|X) = \frac{\sigma^2}{(n-1)s_X^2}$$

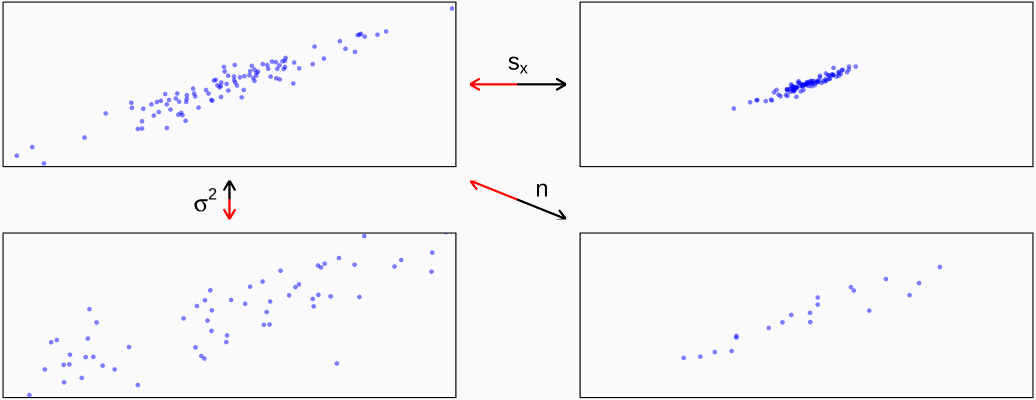
This equation makes it clear that the stand deviation of the slope:

- increases when the error variance σ^2 increases
- decreases when the sample size n increases
- decreases when the spread of the X values s_X^2 increases

Variance of the Estimator Under Homoskedasticity (one X): Graphically

The red side of arrows indicates an increase in the parameter (ie, either σ^2, n, or s_X^2).

Relative to the top-left plot, each plot has an increase in $\text{Var}(\hat{\beta}_1)$



R-Squared

Analysis of Variance

The matrices $\mathbf{P}$ and $\mathbf{M}$ make it easy to show that the decomposition of $\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}}$ into fitted values $\hat{\mathbf{y}}$ and residuals $\hat{\mathbf{e}}$ is orthogonal:

$$\hat{\mathbf{y}}'\hat{\mathbf{e}} = (\mathbf{P}\mathbf{y})'(\mathbf{M}\mathbf{y}) = \mathbf{y}'\mathbf{P}\mathbf{M}\mathbf{y} = \mathbf{y}'(\mathbf{P} - \mathbf{P})\mathbf{y} = 0$$

It follows that

$$\mathbf{y}'\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + 2\hat{\mathbf{y}}'\hat{\mathbf{e}} + \hat{\mathbf{e}}'\hat{\mathbf{e}} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\mathbf{e}}'\hat{\mathbf{e}} \quad \text{or that} \quad \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} \hat{e}_i^2$$

Replace $\mathbf{y}$ with $(\mathbf{y} - \mathbf{i}_n\bar{\mathbf{y}})$ and do some algebra to show

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n} \hat{e}_i^2}_{\text{SSE}}$$

Coefficient of Determination (R^2)

A commonly reported statistic is the Coefficient of Determination (or R^2):

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

Interpretation: The fraction of the sample variance of Y explained by the least squares fit

Notice:

- Minimizing SSE is the same as maximizing R^2
- R^2 (weakly) increases as more regressors are included in a regression model
- The notation comes from the fact that R^2 is the square of the sample correlation between $\mathbf{y}$ and $\hat{\mathbf{y}}$, and also the square of the sample correlation between X and Y when X is univariate

A "high" value of R^2 is sometimes used to claim that a regression model is "valid" or correctly specified or highly accurate for prediction – none of these are necessarily true.

Adjusted R^2

Define $\hat{\sigma}_Y^2 = (1/n)\sum_{i=1}^n (Y_i - \bar{Y})^2$. Then

$$R^2 = 1 - \frac{\text{SSE}}{\text{TSS}} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_Y^2}$$

Since $\hat{\sigma}_Y^2$ is biased for the variance of Y and $\hat{\sigma}^2$ is biased for the error variance, an "adjusted" R^2 measure was proposed using unbiased estimators, often denoted $\bar{R}^2$:

$$\bar{R}^2 = 1 - \frac{s^2}{s_Y^2} = 1 - \left(\frac{n - 1}{n - k}\right)\frac{\sum_{i=1}^n \hat{e}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

where s_Y^2 is the sample variance of Y: $s_Y^2 = (n - 1)^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$

Computation

Computation in R: `lm()`

```r
dat <- read.table("support/cps09mar.txt")
exper <- dat[,1] - dat[,4] - 6
lwage <- log( dat[,5]/(dat[,6]*dat[,7]) )
sam <- dat[,11]==4 & dat[,12]==7 & dat[,2]==0
```

```r
out <- lm(lwage[sam] ~ exper[sam])
summary(out)
```

```
Call:
lm(formula = lwage[sam] ~ exper[sam])

Residuals:
    Min      1Q  Median      3Q     Max
-2.3583 -0.4215  0.0042  0.4718  2.3569

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.876515   0.067631  42.532   <2e-16 ***
exper[sam]  0.004776   0.004335   1.102    0.272
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7122 on 266 degrees of freedom
Multiple R-squared:  0.004542,  Adjusted R-squared:  0.0007998
F-statistic: 1.214 on 1 and 266 DF,  p-value: 0.2716
```

Computation in R: $\hat{y}$ and $\hat{e}$

```r
y <- matrix(lwage[sam], ncol=1)
x <- cbind(1, exper[sam])

xxi <- solve(crossprod(x))
xy <- crossprod(x,y)
betahat <- xxi %*% xy

yhat <- x %*% betahat # fitted values
ehat <- y - yhat # residuals
```

```r
# check y = yhat + resids
sum(y - (yhat + ehat))
```

```
[1] 1.110223e-16
```

```r
# check sum(resids)=0
sum(ehat)
```

```
[1] -3.819167e-13
```

```r
# check sum(x_ie_i)=0
crossprod(x,ehat)
```

```
              [,1]
[1,] -3.819167e-13
[2,] -4.680700e-13
```

Computation in R: $s(\hat{\beta})$

```r
n <- nrow(y)
k <- ncol(x)

# residual standard error
s2 <- (1/(n-k)) * t(ehat) %*% ehat
s2 <- as.vector(s2)
sqrt(s2)
```

```
[1] 0.712242
```

```r
# std err (homosk)
V0 <- s2*xxi
sqrt(diag(V0))
```

```
[1] 0.067631401 0.004335196
```

```r
# std err (heterosk)
u <- x*(ehat %*% matrix(1, ncol=k))
VHC0 <- xxi %*% (t(u) %*% u) %*% xxi
VHC1 <- (n / (n-k)) * VHC0

sqrt(diag(VHC0))
```

```
[1] 0.071346291 0.004295331
```

```r
sqrt(diag(VHC1))
```

```
[1] 0.071614008 0.004311449
```

Computation in R: R^2 and $\bar{R}^2$

```
# R-squared
ybar <- mean(y)
TSS <- sum((y - ybar)^2)
SSE <- t(ehat) %*% ehat

1 - SSE/TSS
```

```
            [,1]
[1,] 0.004542129
```

```
sig2hat <- t(ehat) %*% ehat / n
sigYtilde <- sum((y - ybar)^2) / n

1 - sig2hat/sigYtilde
```

```
            [,1]
[1,] 0.004542129
```

```
#adjusted R-squared
1 - s2/var(y)
```

```
             [,1]
[1,] 0.0007998062
```

Next Time

Asymptotic Distribution of $\hat{\beta}$

- Tools for Asymptotics
- Consistency of $\hat{\beta}$

Inference, once we have the asymptotic distribution

- Hypothesis Tests
- Confidence Intervals

Revisit everything Assuming Errors are iid Normal