Simple Linear Regression

Nathaniel E. Helwig

Assistant Professor of Psychology and Statistics
University of Minnesota (Twin Cities)

Updated 04-Jan-2017

Copyright

Outline of Notes

1) Overview of SLR Model:

- Model form (scalar)
- SLR assumptions
- Model form (matrix)

2) Estimation of SLR Model:

- Ordinary least squares
- Maximum likelihood
- Estimating error variance

3) Inferences in SLR:

- Distribution of estimator
- ANOVA table and *F* test
- CIs and prediction

4) SLR in R:

- The `lm` Function
- Example A: Alcohol
- Example B: GPA

Overview of SLR Model

SLR Model: Form

The simple linear regression model has the form

$$y_i = b_0 + b_1 x_i + e_i$$

for $i \in \{1, \ldots, n\}$ where

- $y_i \in \mathbb{R}$ is the real-valued response for the i-th observation
- $b_0 \in \mathbb{R}$ is the regression intercept
- $b_1 \in \mathbb{R}$ is the regression slope
- $x_i \in \mathbb{R}$ is the predictor for the i-th observation
- $e_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is a Gaussian error term

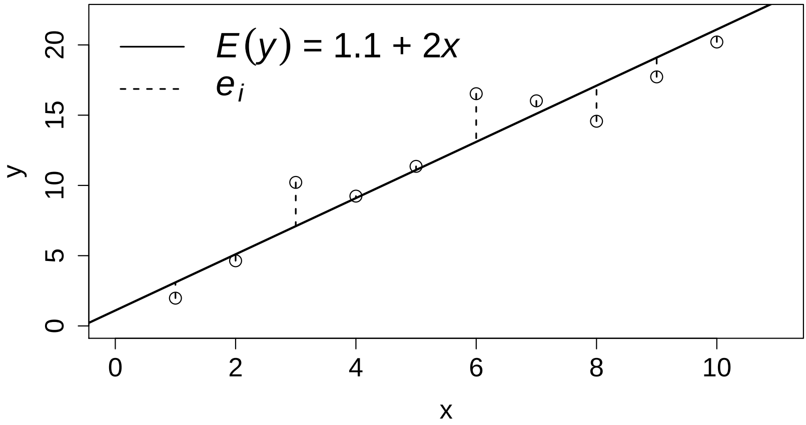
SLR Model: Name

The model is simple because we have only one predictor.

The model is linear because y_i is a linear function of the parameters (b_0 and b_1 are the parameters).

The model is a regression model because we are modeling a response variable (Y) as a function of a predictor variable (X).

SLR Model: Visualization

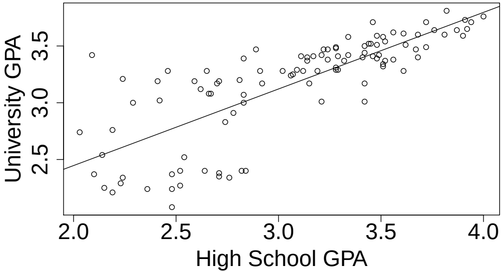


SLR Model: Example

Have GPA from high school and university for $n = 105$ students.

Simple linear regression equation for modeling university GPA:

$$(\mathrm{U_{gpa}})_i = 1.0968 + 0.6748(\mathrm{H_{gpa}})_i + (\mathrm{error})_i$$



Data from http://onlinestatbook.com/2/regression/intro.html

SLR Assumptions: Overview

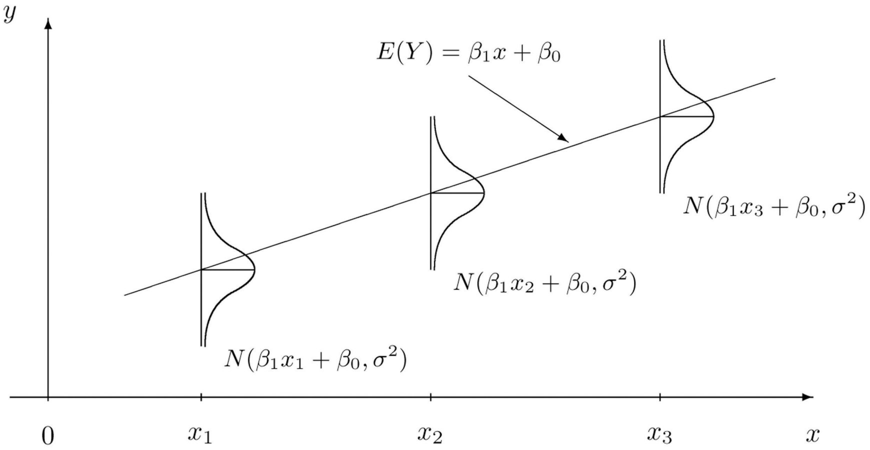
The fundamental assumptions of the SLR model are:

1. Relationship between X and Y is linear
2. x_i and y_i are observed random variables (known constants)
3. $e_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is an unobserved random variable
4. b_0 and b_1 are unknown constants
5. $(y_i | x_i) \overset{\text{ind}}{\sim} \mathrm{N}(b_0 + b_1 x_i, \sigma^2)$; note: homogeneity of variance

Note: b_1 is expected increase in Y for 1-unit increase in X

Which assumption may be violated in the GPA example??

SLR Assumptions: Visualization



http://2012books.lardbucket.org/books/beginning-statistics/s14-03-modelling-linear-relationships.html

SLR Model: Form (revisited)

The simple linear regression model has the form

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

where

- $\mathbf{y} = (y_1, \ldots, y_n)' \in \mathbb{R}^n$ is the $n \times 1$ response vector
- $\mathbf{X} = [\mathbf{1}_n, \mathbf{x}] \in \mathbb{R}^{n \times 2}$ is the $n \times 2$ design matrix
  - $\mathbf{1}_n$ is an $n \times 1$ vector of ones
  - $\mathbf{x} = (x_1, \ldots, x_n)' \in \mathbb{R}^n$ is the $n \times 1$ predictor vector
- $\mathbf{b} = (b_0, b_1)' \in \mathbb{R}^2$ is the 2×1 regression coefficient vector
- $\mathbf{e} = (e_1, \ldots, e_n)' \in \mathbb{R}^n$ is the $n \times 1$ error vector

SLR Model: Form (another look)

Matrix form writes SLR model for all *n* points simultaneously

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{pmatrix}$$

SLR Model: Assumptions (revisited)

In matrix terms, the error vector is multivariate normal:

$$\mathbf{e} \sim \mathrm{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$$

In matrix terms, the response vector is multivariate normal given $\mathbf{x}$:

$$(\mathbf{y}|\mathbf{x}) \sim \mathrm{N}(\mathbf{X}\mathbf{b}, \sigma^2 \mathbf{I}_n)$$

Estimation of SLR Model

Ordinary Least Squares: Scalar Form

The ordinary least squares (OLS) problem is

$$\min_{b_0, b_1 \in \mathbb{R}} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

and the OLS solution has the form

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$
$$\hat{b}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where $\bar{x} = (1/n) \sum_{i=1}^{n} x_i$ and $\bar{y} = (1/n) \sum_{i=1}^{n} y_i$

▸ Calculus derivation

Ordinary Least Squares: Matrix Form

The ordinary least squares (OLS) problem is

$$\min_{\mathbf{b}\in\mathbb{R}^2} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$$

where $\|\cdot\|$ denotes the Frobenius norm; the OLS solution has the form

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

where

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n\sum_{i=1}^n (x_i - \bar{x})^2}\begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix}$$

Calculus derivation

Fitted Values and Residuals

SCALAR FORM:

Fitted values are given by

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$$

and residuals are given by

$$\hat{e}_i = y_i - \hat{y}_i$$

MATRIX FORM:

Fitted values are given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$$

and residuals are given by

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$$

Hat Matrix

Note that we can write the fitted values as

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

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat matrix.

$\mathbf{H}$ is a symmetric and idempotent matrix: $\mathbf{H}\mathbf{H} = \mathbf{H}$

$\mathbf{H}$ projects $\mathbf{y}$ onto the column space of $\mathbf{X}$.

Properties of OLS Estimators

Some useful properties of OLS estimators include:

1. $\sum_{i=1}^{n} \hat{e}_i = 0$
2. $\sum_{i=1}^{n} \hat{e}_i^2$ is minimized with $\mathbf{b} = \hat{\mathbf{b}}$
3. $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(\hat{y}_i + \hat{e}_i) = \sum_{i=1}^{n} \hat{y}_i$
4. $\sum_{i=1}^{n} x_i \hat{e}_i = \sum_{i=1}^{n} x_i(y_i - \hat{b}_0 - \hat{b}_1 x_i) = 0$
5. Regression line passes through center of mass: $(\bar{x}, \bar{y})$

Example #1: Pizza Data

The owner of Momma Leona's Pizza restaurant chain believes that if a restaurant is located near a college campus, then there is a linear relationship between sales and the size of the student population. Suppose data were collected from a sample of 10 Momma Leona's Pizza restaurants located near college campuses.

Population (1000s): x	2	6	8	8	12	16	20	20	22	26
Sales ($1000s): y	58	105	88	118	117	137	157	169	149	202

We want to find the equation of the least-squares regression line predicting quarterly pizza sales (y) from student population (x).

Example #1: OLS Estimation

First note that $\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x}$ and $\hat{b}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Next note that…

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

We only need to find means, sums-of-squares, and cross-products.

Example #1: OLS Estimation (continued)

	x	y	x^2	y^2	xy
	2	58	4	3364	116
	6	105	36	11025	630
	8	88	64	7744	704
	8	118	64	13924	944
	12	117	144	13689	1404
	16	137	256	18769	2192
	20	157	400	24649	3140
	20	169	400	28561	3380
	22	149	484	22201	3278
	26	202	676	40804	5252
$\sum$	140	1300	2528	184730	21040

Example #1: OLS Estimation (continued)

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 2528 - 10(14^2) = 568$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 21040 - 10(14)(130) = 2840$$

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 2840/568 = 5$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x} = 130 - 5(14) = 60$$

$$\hat{y} = 60 + 5x$$

Regression Sums-of-Squares

In SLR models, the relevant sums-of-squares (SS) are

- Sum-of-Squares Total: $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- Sum-of-Squares Regression: $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- Sum-of-Squares Error: $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

The corresponding degrees of freedom (df) are

- SST: $df_T = n - 1$
- SSR: $df_R = 1$
- SSE: $df_E = n - 2$

Partitioning the Variance

We can partition the total variation in y_i as

$$
\begin{aligned}
SST &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n} (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) \\
&= SSR + SSE + 2\sum_{i=1}^{n} (\hat{y}_i - \bar{y})\hat{e}_i \\
&= SSR + SSE
\end{aligned}
$$

▸ Partition proof

Coefficient of Determination

The coefficient of determination is defined as

$$\begin{aligned} R^2 &= \frac{SSR}{SST} \\ &= 1 - \frac{SSE}{SST} \end{aligned}$$

and gives the amount of variation in y_i that is explained by the linear relationship with x_i.

When interpreting R^2 values, note that...

- $0 \leq R^2 \leq 1$
- Large R^2 values do not necessarily imply a good model

Example #1: Fitted Values and Residuals

Returning to the Momma Leona's Pizza example: $\hat{y} = 60 + 5x$

	x	y	$\hat{y}$	$\hat{e}$	$\hat{e}^2$	y^2
	2	58	70	-12	144	3364
	6	105	90	15	225	11025
	8	88	100	-12	144	7744
	8	118	100	18	324	13924
	12	117	120	-3	9	13689
	16	137	140	-3	9	18769
	20	157	160	-3	9	24649
	20	169	160	9	81	28561
	22	149	170	-21	441	22201
	26	202	190	12	144	40804
$\sum$	140	1300	1300	0	1530	184730

Example #1: Sums-of-Squares and R^2

Using the results from the previous table, note that

$$SST = \sum_{i=1}^{10}(y_i - \bar{y})^2 = \sum_{i=1}^{10} y_i^2 - 10\bar{y}^2 = 184730 - 10(130^2) = 15730$$
$$SSE = \sum_{i=1}^{10}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{10} \hat{e}_i^2 = 1530$$
$$SSR = SST - SSE = 15730 - 1530 = 14200$$

which implies that

$$R^2 = SSR/SST = 14200/15730 = 0.9027336$$

so the student population can explain about 90% of the variation in Momma Leona's pizza sales.

Example #1: SS Partition Trick

Note that $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i = \bar{y} + \hat{b}_1(x_i - \bar{x})$ because $\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x}$

Plugging $\hat{y}_i = \bar{y} + \hat{b}_1(x_i - \bar{x})$ into the definition of SSR produces

$$\begin{aligned} SSR &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} \hat{b}_1^2 (x_i - \bar{x})^2 \\ &= 5^2(568) \\ &= 14200 \end{aligned}$$

so do not need the sum-of-squares for y_i

Relation to ML Solution

Remember that $(\mathbf{y}|\mathbf{x}) \sim \mathrm{N}(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I}_n)$, which implies that $\mathbf{y}$ has pdf

$$f(\mathbf{y}|\mathbf{x}, \mathbf{b}, \sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})}$$

As a result, the log-likelihood of $\mathbf{b}$ given $(\mathbf{y}, \mathbf{x}, \sigma^2)$ is

$$\ln\{L(\mathbf{b}|\mathbf{y}, \mathbf{x}, \sigma^2)\} = -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) + c$$

where c is a constant that does not depend on $\mathbf{b}$.

Relation to ML Solution (continued)

The maximum likelihood estimate (MLE) of **b** is the estimate satisfying

$$\max_{\mathbf{b}\in\mathbb{R}^2} -\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})$$

Now, note that...

- $\max_{\mathbf{b}\in\mathbb{R}^2} -\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b}) = \max_{\mathbf{b}\in\mathbb{R}^2} -(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})$
- $\max_{\mathbf{b}\in\mathbb{R}^2} -(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b}) = \min_{\mathbf{b}\in\mathbb{R}^2}(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})$

Thus, the OLS and ML estimate of **b** is the same: $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

Estimated Error Variance (Mean Squared Error)

The estimated error variance is

$$\begin{aligned}\hat{\sigma}^2 &= SSE/(n-2)\\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)\\ &= \|(\mathbf{I}_n - \mathbf{H})\mathbf{y}\|^2/(n-2)\end{aligned}$$

which is an unbiased estimate of error variance σ^2. Unbiased proof

The estimate $\hat{\sigma}^2$ is the mean squared error (MSE) of the model.

Maximum Likelihood Estimate of Error Variance

$\tilde{\sigma}^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n$ is the MLE of σ^2. Calculus derivation

From our previous results using $\hat{\sigma}^2$, we have that

$$\mathrm{E}(\tilde{\sigma}^2) = \frac{n-2}{n}\sigma^2$$

Consequently, the bias of the estimator $\tilde{\sigma}^2$ is given by

$$\frac{n-2}{n}\sigma^2 - \sigma^2 = -\frac{2}{n}\sigma^2$$

and note that $-\frac{2}{n}\sigma^2 \to 0$ as $n \to \infty$.

Comparing $\hat{\sigma}^2$ and $\tilde{\sigma}^2$

Reminder: the MSE and MLE of σ^2 are given by

$$\hat{\sigma}^2 = \|(\mathbf{I}_n - \mathbf{H})\mathbf{y}\|^2/(n-2)$$
$$\tilde{\sigma}^2 = \|(\mathbf{I}_n - \mathbf{H})\mathbf{y}\|^2/n$$

From the definitions of $\hat{\sigma}^2$ and $\tilde{\sigma}^2$ we have that

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

so the MLE produces a smaller estimate of the error variance.

Example #1: Calculating $\hat{\sigma}^2$ and $\tilde{\sigma}^2$

Returning to Momma Leona's Pizza example:

$$SSE = \sum_{i=1}^{10}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{10} \hat{e}_i^2 = 1530$$
$$df_E = 10 - 2 = 8$$

So the estimates of the error variance are given by

$$\hat{\sigma}^2 = MSE = 1530/8 = 191.25$$
$$\tilde{\sigma}^2 = (8/10)MSE = 153$$

Inferences in SLR

OLS Coefficients are Random Variables

Note that $\hat{\mathbf{b}}$ is a linear function of $\mathbf{y}$, so $\hat{\mathbf{b}}$ is multivariate normal.

The expectation of $\hat{\mathbf{b}}$ is given by

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

and the covariance matrix is given by

$$\begin{aligned} \mathrm{V}(\hat{\mathbf{b}}) &= \mathrm{V}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{V}[\mathbf{y}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I}_n)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

OLS Coefficients are Random Variables (continued)

Given the results on the previous slide, have that $\hat{\mathbf{b}} \sim \mathrm{N}(\mathbf{b}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.

Remembering the form of $(\mathbf{X}'\mathbf{X})^{-1}$, we have that

$$\mathrm{V}(\hat{b}_0) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\mathrm{V}(\hat{b}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Fitted Values are Random Variables

Similarly $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$ is a linear function of $\mathbf{y}$, so $\hat{\mathbf{y}}$ is multivariate normal.

The expectation of $\hat{\mathbf{y}}$ is given by

$$\begin{aligned}
\mathrm{E}(\hat{\mathbf{y}}) &= \mathrm{E}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\
&= \mathrm{E}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\mathbf{b} + \mathbf{e})] \\
&= \mathrm{E}[\mathbf{X}\mathbf{b}] + \mathbf{H}\mathrm{E}[\mathbf{e}] \\
&= \mathbf{X}\mathbf{b}
\end{aligned}$$

and the covariance matrix is given by

$$\begin{aligned}
\mathrm{V}(\hat{\mathbf{y}}) &= \mathrm{V}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\
&= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{V}[\mathbf{y}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
&= \mathbf{H}(\sigma^2\mathbf{I}_n)\mathbf{H} \\
&= \sigma^2\mathbf{H}
\end{aligned}$$

Residuals are Random Variables

Also $\hat{\mathbf{e}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$ is a linear function of $\mathbf{y}$, so $\hat{\mathbf{e}}$ is multivariate normal.

The expectation of $\hat{\mathbf{e}}$ is given by

$$\begin{aligned} \mathrm{E}(\hat{\mathbf{e}}) &= \mathrm{E}[(\mathbf{I}_n - \mathbf{H})\mathbf{y}] \\ &= (\mathbf{I}_n - \mathbf{H})\mathrm{E}[\mathbf{y}] \\ &= (\mathbf{I}_n - \mathbf{H})\mathbf{X}\mathbf{b} \\ &= \mathbf{0} \end{aligned}$$

and the covariance matrix is given by

$$\begin{aligned} \mathrm{V}(\hat{\mathbf{e}}) &= \mathrm{V}[(\mathbf{I}_n - \mathbf{H})\mathbf{y}] \\ &= (\mathbf{I}_n - \mathbf{H})\mathrm{V}[\mathbf{y}](\mathbf{I}_n - \mathbf{H}) \\ &= (\mathbf{I}_n - \mathbf{H})(\sigma^2\mathbf{I}_n)(\mathbf{I}_n - \mathbf{H}) \\ &= \sigma^2(\mathbf{I}_n - \mathbf{H}) \end{aligned}$$

Summary of Results

Summarizing the results on the previous slides, we have

$$\hat{\mathbf{b}} \sim \mathrm{N}(\mathbf{b}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

$$\hat{\mathbf{y}} \sim \mathrm{N}(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{H})$$

$$\hat{\mathbf{e}} \sim \mathrm{N}(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H}))$$

Typically σ^2 is unknown, so we use the MSE $\hat{\sigma}^2$ in practice.

ANOVA Table and Regression F Test

We typically organize the SS information into an ANOVA table:

Source	SS	df	MS	F	p-value
SSR	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	1	MSR	F^*	p^*
SSE	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-2$	MSE		
SST	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$			

$MSR = \frac{SSR}{1}$, $MSE = \frac{SSE}{n-2}$, $F^* = \frac{MSR}{MSE} \sim F_{1,n-2}$, $p^* = P(F_{1,n-2} > F^*)$

F^*-statistic and p^*-value are testing $H_0 : b_1 = 0$ versus $H_1 : b_1 \neq 0$

Example #1: ANOVA Table and R^2

Using the results from the previous table, note that

$$SST = \sum_{i=1}^{10}(y_i - \bar{y})^2 = \sum_{i=1}^{10} y_i^2 - 10\bar{y}^2 = 184730 - 10(130^2) = 15730$$
$$SSE = \sum_{i=1}^{10}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{10} \hat{e}_i^2 = 1530$$
$$SSR = SST - SSE = 15730 - 1530 = 14200$$

Source	SS	df	MS	F	p-value
SSR	14200	1	14200.00	74.24837	$< .0001$
SSE	1530	8	191.25		
SST	15730	9			

Reject $H_0 : b_1 = 0$ at any typical α level.

Inferences about $\hat{\mathbf{b}}$ with σ^2 Known

If σ^2 is known, form $100(1-\alpha)\%$ CIs using

$$\hat{b}_0 \pm Z_{\alpha/2}\sigma_{b_0} \qquad \hat{b}_1 \pm Z_{\alpha/2}\sigma_{b_1}$$

where

- $Z_{\alpha/2}$ is normal quantile such that $P(X > Z_{\alpha/2}) = \alpha/2$
- σ_{b_0} and σ_{b_1} are square-roots of diagonals of $\mathrm{V}(\hat{\mathbf{b}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

To test $H_0 : b_j = b_j^*$ vs. $H_1 : b_j \neq b_j^*$ (for $j \in \{0,1\}$) use test statistic

$$Z = (\hat{b}_j - b_j^*)/\sigma_{b_j}$$

which follows a standard normal distribution under H_0.

Inferences about $\hat{\mathbf{b}}$ with σ^2 Unknown

If σ^2 is unknown, form $100(1-\alpha)\%$ CIs using

$$\hat{b}_0 \pm t_{n-2}^{(\alpha/2)} \hat{\sigma}_{b_0} \qquad \hat{b}_1 \pm t_{n-2}^{(\alpha/2)} \hat{\sigma}_{b_1}$$

where

- $t_{n-2}^{(\alpha/2)}$ is t_{n-2} quantile such that $P\left(T > t_{n-2}^{(\alpha/2)}\right) = \alpha/2$
- $\hat{\sigma}_{b_0}$ and $\hat{\sigma}_{b_1}$ are square-roots of diagonals of $\hat{\mathrm{V}}(\hat{\mathbf{b}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$

To test $H_0 : b_j = b_j^*$ vs. $H_1 : b_j \neq b_j^*$ (for $j \in \{0, 1\}$) use test statistic

$$T = (\hat{b}_j - b_j^*)/\hat{\sigma}_{b_j}$$

which follows a t_{n-2} distribution under H_0.

Confidence Interval for σ^2

Note that $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{SSE}{\sigma^2} = \frac{\sum_{i=1}^n \hat{e}_i^2}{\sigma^2} \sim \chi^2_{n-2}$

This implies that

$$\chi^2_{(n-2;1-\alpha/2)} < \frac{(n-2)\hat{\sigma}^2}{\sigma^2} < \chi^2_{(n-2;\alpha/2)}$$

where $P(Q > \chi^2_{(n-2;\alpha/2)}) = \alpha/2$, so a $100(1-\alpha)\%$ CI is given by

$$\frac{(n-2)\hat{\sigma}^2}{\chi^2_{(n-2;\alpha/2)}} < \sigma^2 < \frac{(n-2)\hat{\sigma}^2}{\chi^2_{(n-2;1-\alpha/2)}}$$

Example #1: Inference Questions

Returning to Momma Leona's Pizza example, suppose we want to...

(a) Construct a 90% CI for b_1

(b) Test $H_0 : b_0 = 0$ vs. $H_1 : b_0 \neq 0$. Use $\alpha = 0.01$ for the test.

(c) Test $H_0 : b_0 = 75$ vs. $H_1 : b_0 < 75$. Use a 5% level of significance.

(d) Construct a 95% confidence interval for σ^2.

Example #1: Answer 1a

Question: Construct a 90% CI for b_1.

The variance of $\hat{b}_1$ is given by

$$\begin{aligned}\hat{\mathrm{V}}(\hat{b}_1) &= \hat{\sigma}^2 / \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= 191.25/568 \\ &= 0.3367077\end{aligned}$$

and the critical t_8 values are $t_{(8;.95)} = -1.85955$ and $t_{(8;.05)} = 1.85955$

So the 90% CI for b_1 is given by

$$\begin{aligned}\hat{b}_1 \pm t_{(8;.05)} \sqrt{\hat{\mathrm{V}}(\hat{b}_1)} &= 5 \pm 1.85955 \sqrt{191.25/568} \\ &= [3.920969;\ 6.079031]\end{aligned}$$

Example #1: Answer 1b

Question: Test $H_0 : b_0 = 0$ vs. $H_1 : b_0 \neq 0$. Use $\alpha = 0.01$ for the test.

The variance of $\hat{b}_0$ is given by

$$\begin{aligned}\hat{\mathrm{V}}(\hat{b}_0) &= \frac{\hat{\sigma}^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{191.25(2528)}{10(568)} \\ &= 85.11972\end{aligned}$$

and the critical t_8 values are $t_{(8;.995)} = -3.3554$ and $t_{(8;.005)} = 3.3554$

Observed t test statistic is $T = \frac{60-0}{\sqrt{85.11972}} = 6.503336$, so decision is

$$t_{(8;.005)} = 3.3554 < 6.503336 = T \implies \text{Reject } H_0$$

Example #1: Answer 1c

Question: Test $H_0 : b_0 = 75$ vs. $H_1 : b_0 < 75$. Use $\alpha = 0.05$ for the test.

The variance of $\hat{b}_0$ is given by

$$\hat{\mathrm{V}}(\hat{b}_0) = \frac{\hat{\sigma}^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}$$
$$= 85.11972$$

and the critical t_8 value is $t_{(8;.95)} = 1.859548$

Observed t test statistic is $T = \frac{60-75}{\sqrt{85.11972}} = -1.625834$, so decision is

$$t_{(8;.95)} = -1.859548 < -1.625834 = T \Longrightarrow \text{ Retain } H_0$$

Example #1: Answer 1d

Question: Construct a 95% confidence interval for σ^2.

Using $\alpha = .05$, the critical χ^2_8 values are

$$\chi^2_{(8;.975)} = 2.179731 \qquad \text{and} \qquad \chi^2_{(8;.025)} = 17.53455$$

So the 95% confidence interval for σ^2 is given by

$$\left[\frac{8\hat{\sigma}^2}{\chi^2_{(8;.025)}};\ \frac{8\hat{\sigma}^2}{\chi^2_{(8;.975)}}\right] = \left[\frac{1530}{17.53455};\ \frac{1530}{2.179731}\right]$$
$$= [87.2563;\ 701.9215]$$

Interval Estimation

Idea: estimate expected value of response for a given predictor score.

Given x_h, the fitted value is $\hat{y}_h = \mathbf{x}_h \hat{\mathbf{b}}$ where $\mathbf{x}_h = (1 \ x_h)$.

Variance of $\hat{y}_h$ is given by $\sigma^2_{\bar{y}_h} = \mathrm{V}(\mathbf{x}_h \hat{\mathbf{b}}) = \mathbf{x}_h \mathrm{V}(\hat{\mathbf{b}}) \mathbf{x}'_h = \sigma^2 \mathbf{x}_h (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}'_h$

- Use $\hat{\sigma}^2_{\bar{y}_h} = \hat{\sigma}^2 \mathbf{x}_h (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}'_h$ if σ^2 is unknown

We can test $H_0 : \mathrm{E}(y_h) = y^*_h$ vs. $H_1 : \mathrm{E}(y_h) \neq y^*_h$

- Test statistic: $T = (\hat{y}_h - y^*_h)/\hat{\sigma}_{\bar{y}_h}$, which follows $t_{(n-2)}$ distribution
- $100(1-\alpha)$% CI for $E(y_h)$: $\hat{y}_h \pm t^{(\alpha/2)}_{n-2} \hat{\sigma}_{\bar{y}_h}$

Predicting New Observations

Idea: estimate observed value of response for a given predictor score.

- Note: interested in actual $\hat{y}_h$ value instead of $\mathrm{E}(\hat{y}_h)$

Given x_h, the fitted value is $\hat{y}_h = \mathbf{x}_h \hat{\mathbf{b}}$ where $\mathbf{x}_h = (1 \ x_h)$.

- Note: same as interval estimation

When predicting a new observation, there are two uncertainties:

- location of the distribution of Y for X_h (captured by $\sigma^2_{\hat{y}_h}$)
- variability within the distribution of Y (captured by σ^2)

Predicting New Observations (continued)

Two sources of variance are independent so $\sigma^2_{y_h} = \sigma^2_{\bar{y}_h} + \sigma^2$

- Use $\hat{\sigma}^2_{y_h} = \hat{\sigma}^2_{\bar{y}_h} + \hat{\sigma}^2$ if σ^2 is unknown

We can test $H_0 : y_h = y_h^*$ vs. $H_1 : y_h \neq y_h^*$

- Test statistic: $T = (\hat{y}_h - y_h^*)/\hat{\sigma}_{y_h}$, which follows $t_{(n-2)}$ distribution
- $100(1-\alpha)\%$ Prediction Interval (PI) for y_h: $\hat{y}_h \pm t_{n-2}^{(\alpha/2)} \hat{\sigma}_{y_h}$

Familywise Confidence Intervals

Returning to the idea of interval estimation, we could construct a $100(1-\alpha)\%$ CI around $E(y_h)$ for $g > 1$ different x_h values.

- Note: we have an error rate of α for each individual CI

The familywise error rate is the probability that we make one (or more) errors among all g predictions simultaneously.

If predictions are independent, we have that $FWER = 1 - (1-\alpha)^g$.

- Note: familywise error rate increases as g increases
- With $g = 1$ and $\alpha = .05$, $FWER = 1 - (1 - .05) = .05$
- With $g = 2$ and $\alpha = .05$, $FWER = 1 - (1 - .05)^2 = 0.0975$

Familywise Confidence Intervals (continued)

There are many options (corrections or adjustments) we can use.

Bonferroni adjustment controls FWER at α by using $\alpha^* = \alpha/g$ as significance level for each of the g CIs.

Bonferroni's adjustment is very simple, but is conservative

- Does not assume independence between g predictions
- Will be overly conservative if predictions are independent

Simultaneous Confidence Bands

In SLR we typically want a confidence band, which is similar to a CI but holds for multiple values of x.

Given the distribution of $\hat{\mathbf{b}}$ (and some probability theory), we have that

$$\frac{(\hat{\mathbf{b}} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b})}{\sigma^2} \sim \chi_2^2$$
$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-2}^2$$

which implies that

$$\frac{(\hat{\mathbf{b}} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b})}{2\hat{\sigma}^2} \sim \frac{\chi_2^2/2}{\chi_{n-2}^2/(n-2)} \equiv F_{2,n-2}$$

Simultaneous Confidence Bands (continued)

To form a $100(1-\alpha)\%$ confidence band (CB) use limits such that

$$(\hat{\mathbf{b}} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\hat{\mathbf{b}} - \mathbf{b}) \leq 2\hat{\sigma}^2 F_{2,n-2}^{(\alpha)}$$

where $F_{2,n-2}^{(\alpha)}$ is the critical value corresponding to significance level α.

For the SLR model we can form a $100(1-\alpha)\%$ CB using

$$\hat{b}_0 + \hat{b}_1 x \pm \sqrt{2F_{2,n-2}^{(\alpha)}\hat{\sigma}^2 \begin{pmatrix} 1 & x \end{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} \begin{pmatrix} 1 \\ x \end{pmatrix}}$$

Example #1: Prediction Questions

Returning to Momma Leona's Pizza example, suppose we want to...

(e) Construct a 95% confidence interval for $E(Y|X = 48)$.

(f) Construct a 95% prediction interval for a future value of Y corresponding to $X = 48$.

(g) University of Minnesota has 48 thousand students. Momma Leona would agree to open a restaurant near the UMN campus, but only if there is enough evidence that the average quarterly sales would be over \$250,000. Using $\alpha = 0.05$, test $H_0 : E(Y|X = 48) = 250$ vs. $H_1 : E(Y|X = 48) > 250$.

Example #1: Answer 1e

Question: Construct a 95% confidence interval for $E(Y|X = 48)$.

The fitted value is $\hat{y} = 60 + 5(48) = 300$ and the variance of $E(Y|X = 48)$ is given by

$$
\begin{aligned}
\sigma^2_{\bar{y}} &= \hat{\sigma}^2 \begin{pmatrix} 1 & 48 \end{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} \begin{pmatrix} 1 \\ 48 \end{pmatrix} \\
&= \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(48 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= 191.25 \left(\frac{1}{10} + \frac{(48 - 14)^2}{568} \right) \\
&= 191.25(2.135211) \\
&= 408.3592
\end{aligned}
$$

and the critical t_8 values are $t_{(8;.975)} = -2.306$ and $t_{(8;.025)} = 2.306$

Example #1: Answer 1e (continued)

Question: Construct a 95% confidence interval for $E(Y|X = 48)$.

Note that $\hat{y} = 60 + 5(48) = 300$, $\sigma^2_{\bar{y}} = 408.3592$, and $t_{(8;.025)} = 2.306$

So the 95% CI for $E(Y|X = 48)$ is given by

$$\begin{aligned} \hat{y} \pm t_{(8;.025)}\sigma_{\bar{y}} &= 300 \pm 2.306\sqrt{408.3592} \\ &= [253.4005;\ 346.5995] \end{aligned}$$

Example #1: Answer 1f

Question: Construct a 95% prediction interval for a future value of Y corresponding to $X = 48$.

The fitted value is $\hat{y} = 60 + 5(48) = 300$ and the variance of a predicted value corresponding to $X = 48$ is given by

$$\begin{aligned} \sigma^2_{\hat{y}} &= \hat{\sigma}^2 \left[1 + \begin{pmatrix} 1 & 48 \end{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} \begin{pmatrix} 1 \\ 48 \end{pmatrix} \right] \\ &= 191.25\,[1 + 2.135211] \\ &= 599.6092 \end{aligned}$$

So, given $X = 48$, the 95% PI for Y would be

$$\begin{aligned} \hat{y} \pm t_{(8;.025)}\sigma_{\hat{y}} &= 300 \pm 2.306\sqrt{599.6092} \\ &= [243.5331;\ 356.4669] \end{aligned}$$

Example #1: Answer 1g

Question: Test $H_0 : E(Y|X = 48) = 250$ vs. $H_1 : E(Y|X = 48) > 250$ using significance level of $\alpha = 0.05$.

The fitted value is $\hat{y} = 60 + 5(48) = 300$ and the variance of $E(Y|X = 48)$ is given by

$$\begin{aligned} \sigma^2_{\bar{y}} &= \hat{\sigma}^2 \begin{pmatrix} 1 & 48 \end{pmatrix} (\mathbf{X}'\mathbf{X})^{-1} \begin{pmatrix} 1 \\ 48 \end{pmatrix} \\ &= 408.3592 \end{aligned}$$

and the critical t_8 value is $t_{(8;.05)} = 1.859548$

Observed t test statistic is $T = \frac{300-250}{\sqrt{408.3592}} = 2.47428$, so decision is

$$t_{(8;.95)} = 1.859548 < 2.47428 = T \implies \text{Reject } H_0$$

SLR in R

Linear Models in R using `lm` Function

In R linear models are fit using the `lm` function.

For SLR the basic syntax of the `lm` function is

```
lm(y ~ x, data=mydata)
```

where

- `y` is the response variable
- `x` is the predictor variable
- `~` separates response and predictors
- `mydata` is the data frame containing `y` and `x`

Note: if `y` and `x` are defined in workspace, you can ignore `data` input.

Output from `lm` Function

We fit and save a linear model using the code

```
mymod = lm(y ~ x, data=mydata)
```

where `mymod` is the object produced by the `lm` function.

Note that `mymod` is an object of class `lm`, which is a list containing many pieces of information about the fit model:

- `coefficients`: $\hat{b}_0$ and $\hat{b}_1$ estimates
- `residuals`: $\hat{e}_i = y_i - \hat{y}_i$ estimates
- `fitted.values`: $\hat{y}_i$ estimates
- And more…

`print` and `summary` of `lm` Output

We can input an object output from the `lm` function into...

- `print` function to see formula and coefficients
- `summary` function to see formula, coefficients, and some basic inference information (R^2, $\hat{\sigma}$, $\hat{\sigma}_{b_0}$, $\hat{\sigma}_{b_1}$, etc.)

Note 1: `print(mymod)` produces same result as typing `mymod`

Note 2: `summary` is typically more useful than `print`

Example A: Drinking Data

This example uses the drinking data set from A Handbook of Statistical Analyses using SAS, 3rd Edition (Der & Everitt, 2008).

Y: number of cirrhosis deaths per 100,000 people (`cirrhosis`).

X: average yearly alcohol consumption in liters/person (`alcohol`).

Have data from $n = 15$ different countries (note: these data are old).

Example A: Drinking Data (continued)

```
> drinking
       country alcohol cirrhosis
1       France    24.7      46.1
2        Italy    15.2      23.6
3    W.Germany    12.3      23.7
4      Austria    10.9       7.0
5      Belgium    10.8      12.3
6          USA     9.9      14.2
7       Canada     8.3       7.4
8          E&W     7.2       3.0
9       Sweden     6.6       7.2
10       Japan     5.8      10.6
11 Netherlands     5.7       3.7
12     Ireland     5.6       3.4
13      Norway     4.2       4.3
14     Finland     3.9       3.6
15      Israel     3.1       5.4
```

Example A: Analyses and Results

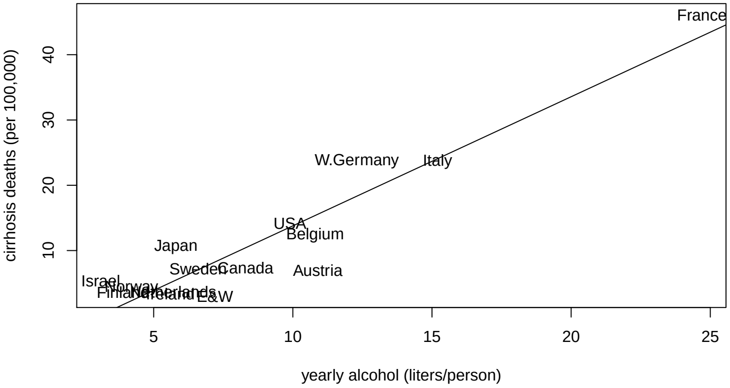
```
> drinkmod = lm(cirrhosis ~ alcohol, data=drinking)
> summary(drinkmod)

Call:
lm(formula = cirrhosis ~ alcohol, data = drinking)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5635 -2.3508  0.1415  2.6149  5.3674

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.9958     2.0977  -2.858   0.0134 *
alcohol       1.9779     0.2012   9.829  2.2e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.17 on 13 degrees of freedom
Multiple R-squared:  0.8814, Adjusted R-squared:  0.8723
F-statistic: 96.61 on 1 and 13 DF,  p-value: 2.197e-07
```

Example A: Manual Calculations

```
> X = cbind(1, drinking$alcohol)
> y = drinking$cirrhosis
> XtX = crossprod(X)
> Xty = crossprod(X, y)
> XtXi = solve(XtX)
> bhat = XtXi %*% Xty
> yhat = X %*% bhat
> ehat = y - yhat
> sigsq = sum(ehat^2) / (nrow(X)-2)
> bhatse = sqrt(sigsq*diag(XtXi))
> tval = bhat / bhatse
> pval = 2*(1-pt(abs(tval),nrow(X)-2))
> data.frame(bhat=bhat, se=bhatse, t=tval, p=pval)
       bhat        se         t           p
1 -5.995753 2.0977480 -2.858186 1.34443e-02
2  1.977916 0.2012283  9.829211 2.19651e-07
```

Example A: Visualization

```
plot(drinking$alcohol, drinking$cirrhosis, type="n",
     xlab="yearly alcohol (liters/person)", ylab="cirrhosis deaths (per 100,000)")
text(drinking$alcohol, drinking$cirrhosis, drinking$country)
abline(drinkmod$coef[1], drinkmod$coef[2])
```



Example A: Prediction

Suppose we have the following data from four countries

```
> drinknew
     country alcohol cirrhosis
1  Lithuania    12.6        NA
2    Romania    12.7        NA
3     Latvia    13.2        NA
4 Luxembourg    15.3        NA
```

To get the associated $\hat{y}_h$ values use the `predict` function:

```
> predict(drinkmod, newdata=drinknew)
       1        2        3        4
18.92599 19.12378 20.11274 24.26636
```

Example A: Prediction (continued)

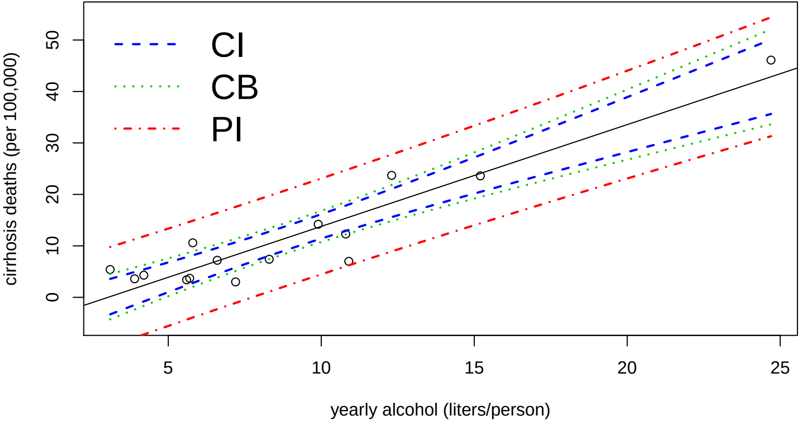
You can use the `predict` function to make CIs around $\mathrm{E}(\hat{y}_h)$:

```
> predict(drinkmod, newdata=drinknew, interval="confidence", level=0.9)
       fit      lwr      upr
1 18.92599 16.61708 21.23489
2 19.12378 16.79459 21.45296
3 20.11274 17.67686 22.54861
4 24.26636 21.30626 27.22645
```

Or you can use the `predict` function to make PIs around $\hat{y}_h$:

```
> predict(drinkmod, newdata=drinknew, interval="prediction", level=0.9)
       fit      lwr      upr
1 18.92599 11.18828 26.66369
2 19.12378 11.38000 26.86755
3 20.11274 12.33620 27.88927
4 24.26636 16.31003 32.22269
```

Example A: Visualization (revisited)



Example A: Visualization (R code)

```
drng = range(drinking$alcohol)
drinkseq = data.frame(alcohol=seq(drng[1],drng[2],length.out=100))
civals = predict(drinkmod,newdata=drinkseq,interval="confidence")
pivals = predict(drinkmod,newdata=drinkseq,interval="prediction")
sevals = predict(drinkmod,newdata=drinkseq,se.fit=T)
plot(drinking$alcohol,drinking$cirrhosis,ylim=c(-5,55),
     xlab="yearly alcohol (liters/person)",
     ylab="cirrhosis deaths (per 100,000)")
abline(drinkmod$coef[1],drinkmod$coef[2])
W = sqrt(2*qf(.95,2,13))
lines(drinkseq$alcohol,civals[,2],lty=2,col="blue",lwd=2)
lines(drinkseq$alcohol,civals[,3],lty=2,col="blue",lwd=2)
lines(drinkseq$alcohol,sevals$fit+W*sevals$se.fit,
      lty=3,col="green3",lwd=2)
lines(drinkseq$alcohol,sevals$fit-W*sevals$se.fit,
      lty=3,col="green3",lwd=2)
lines(drinkseq$alcohol,pivals[,2],lty=4,col="red",lwd=2)
lines(drinkseq$alcohol,pivals[,3],lty=4,col="red",lwd=2)
legend("topleft",c("CI","CB","PI"),lty=2:4,cex=2,
       lwd=rep(2,3),col=c("blue","green3","red"),bty="n")
```

Example B: GPA Data

This example uses the GPA data set that we examined before.

- From http://onlinestatbook.com/2/regression/intro.html

Y: student's university grade point average.

X: student's high school grade point average.

Have data from $n = 105$ different students.

Example B: GPA Data (continued)

GPAs for the first 10 students in data set:

```
> gpa[1:10,]
   high_GPA math_SAT verb_SAT comp_GPA univ_GPA
1      3.45      643      589     3.76     3.52
2      2.78      558      512     2.87     2.91
3      2.52      583      503     2.54     2.40
4      3.67      685      602     3.83     3.47
5      3.24      592      538     3.29     3.47
6      2.10      562      486     2.64     2.37
7      2.82      573      548     2.86     2.40
8      2.36      559      536     2.03     2.24
9      2.42      552      583     2.81     3.02
10     3.51      617      591     3.41     3.32
```

Example B: Analyses and Results

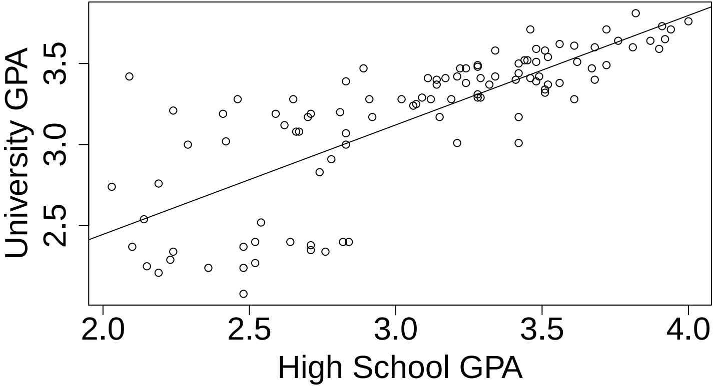
```
> gpamod = lm(univ_GPA ~ high_GPA, data=gpa)
> summary(gpamod)

Call:
lm(formula = univ_GPA ~ high_GPA, data = gpa)

Residuals:
     Min       1Q   Median       3Q      Max
-0.69040 -0.11922  0.03274  0.17397  0.91278

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.09682    0.16663   6.583 1.98e-09 ***
high_GPA     0.67483    0.05342  12.632  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2814 on 103 degrees of freedom
Multiple R-squared:  0.6077, Adjusted R-squared:  0.6039
F-statistic: 159.6 on 1 and 103 DF,  p-value: < 2.2e-16
```

Example B: Manual Calculations

```
> X = cbind(1, gpa$high_GPA)
> y = gpa$univ_GPA
> XtX = crossprod(X)
> Xty = crossprod(X,y)
> XtXi = solve(XtX)
> bhat = XtXi %*% Xty
> yhat = X %*% bhat
> ehat = y - yhat
> sigsq = sum(ehat^2) / (nrow(X)-2)
> bhatse = sqrt(sigsq*diag(XtXi))
> tval = bhat / bhatse
> pval = 2*(1-pt(abs(tval),nrow(X)-2))
> data.frame(bhat=bhat, se=bhatse, t=tval, p=pval)
       bhat         se        t            p
1 1.0968233 0.16662690  6.58251 1.976679e-09
2 0.6748299 0.05342238 12.63197 0.000000e+00
```

Example B: Visualization

```
par(mar=c(5,5.4,4,2)+0.1)
plot(gpa$high_GPA, gpa$univ_GPA, xlab="High School GPA",
     ylab="University GPA", cex.lab=2, cex.axis=2)
abline(a=gpamod$coef[1],gpamod$coef[2])
```

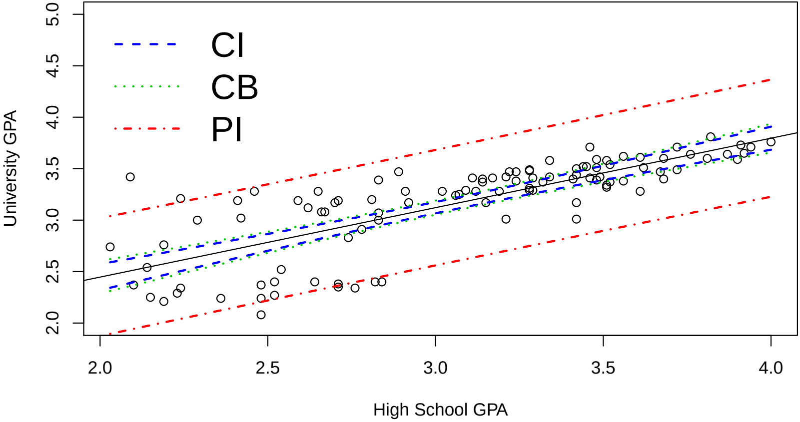


Example B: Prediction

Predicted university GPA for data from five new students

```
> gpanew = data.frame(high_GPA=c(2.4,3,3.1,3.3,3.9),
+                     univ_GPA=rep(NA,5))
> gpanew
  high_GPA univ_GPA
1      2.4       NA
2      3.0       NA
3      3.1       NA
4      3.3       NA
5      3.9       NA
> predict(gpamod, newdata=gpanew)
       1        2        3        4        5
2.716415 3.121313 3.188796 3.323762 3.728660
```

Example B: Visualization (revisited)



Example B: Visualization (R code)

```
drng = range(gpa$high_GPA)
gpaseq = data.frame(high_GPA=seq(drng[1],drng[2],length.out=100))
civals = predict(gpamod,newdata=gpaseq,interval="confidence")
pivals = predict(gpamod,newdata=gpaseq,interval="prediction")
sevals = predict(gpamod,newdata=gpaseq,se.fit=T)
plot(gpa$high_GPA, gpa$univ_GPA, ylim=c(2,5),
     xlab="High School GPA", ylab="University GPA")
abline(gpamod$coef[1],gpamod$coef[2])
W = sqrt(2*qf(.95,2,103))
lines(gpaseq$high_GPA,civals[,2],lty=2,col="blue",lwd=2)
lines(gpaseq$high_GPA,civals[,3],lty=2,col="blue",lwd=2)
lines(gpaseq$high_GPA,sevals$fit+W*sevals$se.fit,
      lty=3,col="green3",lwd=2)
lines(gpaseq$high_GPA,sevals$fit-W*sevals$se.fit,
      lty=3,col="green3",lwd=2)
lines(gpaseq$high_GPA,pivals[,2],lty=4,col="red",lwd=2)
lines(gpaseq$high_GPA,pivals[,3],lty=4,col="red",lwd=2)
legend("topleft",c("CI","CB","PI"),lty=2:4,cex=2,
       lwd=rep(2,3),col=c("blue","green3","red"),bty="n")
```

Appendix

OLS Problem (revisited)

The OLS problem is to find the $b_0, b_1 \in \mathbb{R}$ that minimize

$$
\begin{aligned}
SSE &= \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 \\
&= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \\
&= \sum_{i=1}^{n} \left\{ y_i^2 - 2y_i(b_0 + b_1 x_i) + (b_0 + b_1 x_i)^2 \right\} \\
&= \sum_{i=1}^{n} \left\{ y_i^2 - 2y_i(b_0 + b_1 x_i) + b_0^2 + 2b_0 b_1 x_i + b_1^2 x_i^2 \right\}
\end{aligned}
$$

Solving for Intercept

Taking the derivative of the SSE with respect to b_0 gives

$$\begin{aligned}\frac{\partial SSE}{\partial b_0} &= \sum_{i=1}^{n}\{-2y_i + 2b_0 + 2b_1 x_i\} \\ &= -2n\bar{y} + 2nb_0 + 2nb_1\bar{x}\end{aligned}$$

and setting to zero and solving for b_0 gives

$$\hat{b}_0 = \bar{y} - b_1\bar{x}$$

Solving for Slope

Taking the derivative of the SSE with respect to b_1 gives

$$\begin{aligned}\frac{\partial SSE}{\partial b_1} &= \textstyle\sum_{i=1}^n \left\{-2y_i x_i + 2b_0 x_i + 2b_1 x_i^2\right\} \\ &= \textstyle\sum_{i=1}^n \left\{-2y_i x_i + 2(\bar{y} - b_1\bar{x})x_i + 2b_1 x_i^2\right\} \\ &= -2\textstyle\sum_{i=1}^n y_i x_i + 2n\bar{x}\bar{y} - 2nb_1\bar{x}^2 + 2b_1\sum_{i=1}^n x_i^2 \\ &= -2\textstyle\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + 2b_1\sum_{i=1}^n (x_i - \bar{x})^2\end{aligned}$$

and setting to zero and solving for b_1 gives

$$\hat{b}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Return

Vector Calculus: Derivative of Matrix-Vector Product

Given $\mathbf{A} = \{a_{ij}\}_{n \times p}$ and $\mathbf{b} = \{b_j\}_{p \times 1}$, we have that

$$
\begin{aligned}
\frac{\partial \mathbf{A}\mathbf{b}}{\partial \mathbf{b}'} &= \begin{pmatrix}
\frac{\partial \sum_{j=1}^{p} a_{1j}b_j}{\partial b_1} & \frac{\partial \sum_{j=1}^{p} a_{1j}b_j}{\partial b_2} & \cdots & \frac{\partial \sum_{j=1}^{p} a_{1j}b_j}{\partial b_p} \\
\frac{\partial \sum_{j=1}^{p} a_{2j}b_j}{\partial b_1} & \frac{\partial \sum_{j=1}^{p} a_{2j}b_j}{\partial b_2} & \cdots & \frac{\partial \sum_{j=1}^{p} a_{2j}b_j}{\partial b_p} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial \sum_{j=1}^{p} a_{nj}b_j}{\partial b_1} & \frac{\partial \sum_{j=1}^{p} a_{nj}b_j}{\partial b_2} & \cdots & \frac{\partial \sum_{j=1}^{p} a_{nj}b_j}{\partial b_p}
\end{pmatrix}_{n \times p} \\
&= \mathbf{A}
\end{aligned}
$$

Vector Calculus: Derivative of Quadratic Form

Given $\mathbf{A} = \{a_{ij}\}_{p \times p}$ and $\mathbf{b} = \{b_i\}_{p \times 1}$, we have that

$$\begin{aligned}
\frac{\partial \mathbf{b}'\mathbf{A}\mathbf{b}}{\partial \mathbf{b}'} &= \begin{pmatrix} \frac{\partial \sum_{i=1}^{p} \sum_{j=1}^{p} b_i b_j a_{ij}}{\partial b_1} & \frac{\partial \sum_{i=1}^{p} \sum_{j=1}^{p} b_i b_j a_{ij}}{\partial b_2} & \cdots & \frac{\partial \sum_{i=1}^{p} \sum_{j=1}^{p} b_i b_j a_{ij}}{\partial b_p} \end{pmatrix}_{1 \times p} \\
&= \begin{pmatrix} 2\sum_{i=1}^{p} b_i a_{i1} & 2\sum_{i=1}^{p} b_i a_{i2} & \cdots & 2\sum_{i=1}^{p} b_i a_{ip} \end{pmatrix}_{1 \times p} \\
&= 2\mathbf{b}'\mathbf{A}
\end{aligned}$$

Solving for Intercept and Slope Simultaneously

Note that we can write the OLS problem as

$$\begin{aligned} SSE &= \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 \\ &= (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \end{aligned}$$

Taking the first derivative of *SSE* with respect to **b** produces

$$\frac{\partial SSE}{\partial \mathbf{b}'} = -2\mathbf{y}'\mathbf{X} + 2\mathbf{b}'\mathbf{X}'\mathbf{X}$$

Setting to zero and solving for **b** gives

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

Return

Partitioning the Variance: Proof

To show that $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})\hat{e}_i = 0$, note that

$$\begin{aligned}
\sum_{i=1}^{n}(\hat{y}_i - \bar{y})\hat{e}_i &= (\mathbf{Hy} - n^{-1}\mathbf{1}_n\mathbf{1}_n'\mathbf{y})'(\mathbf{y} - \mathbf{Hy}) \\
&= \mathbf{y}'\mathbf{Hy} - \mathbf{y}'\mathbf{H}^2\mathbf{y} - n^{-1}\mathbf{y}'\mathbf{1}_n\mathbf{1}_n'\mathbf{y} + n^{-1}\mathbf{y}'\mathbf{1}_n\mathbf{1}_n'\mathbf{Hy} \\
&= \mathbf{y}'\mathbf{Hy} - \mathbf{y}'\mathbf{H}^2\mathbf{y} - n^{-1}\mathbf{y}'\mathbf{1}_n\mathbf{1}_n'\mathbf{y} + n^{-1}\mathbf{y}'\mathbf{H}\mathbf{1}_n\mathbf{1}_n'\mathbf{y} \\
&= 0
\end{aligned}$$

given that $\mathbf{H}^2 = \mathbf{H}$ (because $\mathbf{H}$ is idempotent) and $\mathbf{H}\mathbf{1}_n\mathbf{1}_n' = \mathbf{1}_n\mathbf{1}_n'$ (because $\mathbf{1}_n\mathbf{1}_n'$ is within the column space of $\mathbf{X}$ and $\mathbf{H}$ is the projection matrix for the column space of $\mathbf{X}$).

Return

Proof MSE is Unbiased

First note that we can write *SSE* as

$$\begin{aligned}\|(\mathbf{I}_n - \mathbf{H})\mathbf{y}\|^2 &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{H}\mathbf{y} + \mathbf{y}'\mathbf{H}^2\mathbf{y} \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{H}\mathbf{y}\end{aligned}$$

Now define $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b}$ and note that

$$\begin{aligned}\tilde{\mathbf{y}}'\tilde{\mathbf{y}} - \tilde{\mathbf{y}}'\mathbf{H}\tilde{\mathbf{y}} &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} - \mathbf{y}'\mathbf{H}\mathbf{y} + 2\mathbf{y}'\mathbf{H}\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{H}\mathbf{X}\mathbf{b} \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{H}\mathbf{y} \\ &= SSE\end{aligned}$$

given that $\mathbf{H}\mathbf{X} = \mathbf{X}$ (note $\mathbf{H}$ is projection matrix for column space of $\mathbf{X}$).

Now use the trace trick

$$\begin{aligned}\tilde{\mathbf{y}}'\tilde{\mathbf{y}} - \tilde{\mathbf{y}}'\mathbf{H}\tilde{\mathbf{y}} &= \mathrm{tr}(\tilde{\mathbf{y}}'\tilde{\mathbf{y}}) - \mathrm{tr}(\tilde{\mathbf{y}}'\mathbf{H}\tilde{\mathbf{y}}) \\ &= \mathrm{tr}(\tilde{\mathbf{y}}\tilde{\mathbf{y}}') - \mathrm{tr}(\mathbf{H}\tilde{\mathbf{y}}\tilde{\mathbf{y}}')\end{aligned}$$

Proof MSE is Unbiased (continued)

Plugging in the previous results and taking the expectation gives

$$\begin{aligned}
\mathrm{E}(\hat{\sigma}^2) &= \frac{\mathrm{E}\left[\mathrm{tr}(\tilde{\mathbf{y}}\tilde{\mathbf{y}}')\right]}{n-2} - \frac{\mathrm{E}\left[\mathrm{tr}(\mathbf{H}\tilde{\mathbf{y}}\tilde{\mathbf{y}}')\right]}{n-2} \\
&= \frac{\mathrm{tr}(\mathrm{E}\left[\tilde{\mathbf{y}}\tilde{\mathbf{y}}'\right])}{n-2} - \frac{\mathrm{tr}(\mathbf{H}\mathrm{E}\left[\tilde{\mathbf{y}}\tilde{\mathbf{y}}'\right])}{n-2} \\
&= \frac{\mathrm{tr}(\sigma^2\mathbf{I}_n)}{n-2} - \frac{\mathrm{tr}(\mathbf{H}\sigma^2\mathbf{I}_n)}{n-2} \\
&= \frac{n\sigma^2}{n-2} - \frac{2\sigma^2}{n-2} \\
&= \sigma^2
\end{aligned}$$

which completes the proof; to prove that $\mathrm{tr}(\mathbf{H}) = 2$, note that

$$\mathrm{tr}(\mathbf{H}) = \mathrm{tr}\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) = \mathrm{tr}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\right) = \mathrm{tr}(\mathbf{I}_2) = 2$$

▸ Return

ML Estimate of σ^2: Overview

Remember that the pdf of **y** has the form

$$f(\mathbf{y}|\mathbf{x}, \mathbf{b}, \sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})}$$

As a result, the log-likelihood of σ^2 given $(\mathbf{y}, \mathbf{x}, \hat{\mathbf{b}})$ is

$$\ln\{L(\sigma^2|\mathbf{y}, \mathbf{x}, \hat{\mathbf{b}})\} = -\frac{n\ln(\sigma^2)}{2} - \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{2\sigma^2} + d$$

where d is a constant that does not depend on σ^2.

Solving for Error Variance

The MLE of σ^2 is the estimate satisfying

$$\max_{\sigma^2 \in \mathbb{R}^+} -\frac{n \ln(\sigma^2)}{2} - \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{2\sigma^2}$$

Taking the first derivative with respect to σ^2 gives

$$\frac{\partial \left\{ -\frac{n \ln(\sigma^2)}{2} - \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{2\sigma^2} \right\}}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{2\sigma^4}$$

Setting to zero and solving for σ^2 gives

$$\tilde{\sigma}^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}}/n$$

Return