Stat 608 Chapter 2 (and 5)

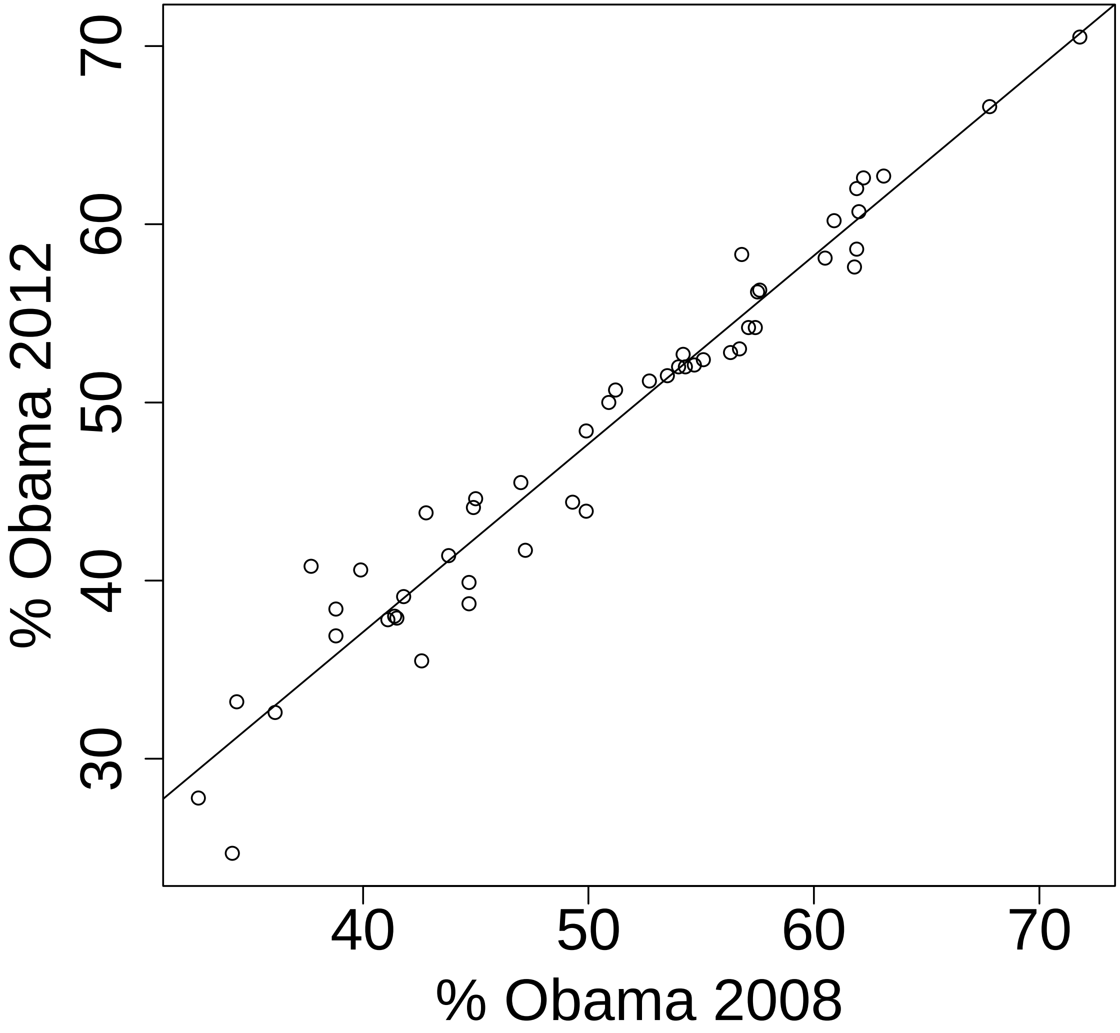
Simple Linear regression

Simple Linear Regression (SLR) Models

We always draw a scatterplot first to obtain an idea of the strength (measured by correlation), form (e.g., linear, quadratic, exponential), and direction (positive or negative, in the linear case) of any relationship that exists between two variables.

On the next slide we see a strong, positive linear relationship between the percentage of voters of each state who voted for President Obama in 2008 and in 2012.

Simple Linear Regression (SLR) Models



Simple Linear Regression (SLR) Models

When data are collected in pairs the standard notation used to designate this is:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

where x_1 denotes the first value of the X-variable and y_1 denotes the first value of the Y-variable. The X-variable is called the **explanatory variable,** while the Y-variable is called the **response variable.**

Simple Linear Regression (SLR) Models

- For example, the first few entries in the election data set above were

  (37.7, 40.8), (38.8, 38.4), (38.8, 36.9)

  meaning 37.7% of Alaska voters voted for President Obama in 2008, while 40.8% did so in 2012, and so on.

Simple Linear Regression (SLR) Models

- We could use a matched pairs t-test to test whether the same percentage of voters voted for President Obama in 2008 and 2012 in the 50 states, on average. If the percentages were indeed equal, what would that mean for our regression model?

+

S.L.R. Models

The regression of *Y* on *X* is linear if

$$E[Y|X = x] = \beta_0 + \beta_1 x$$

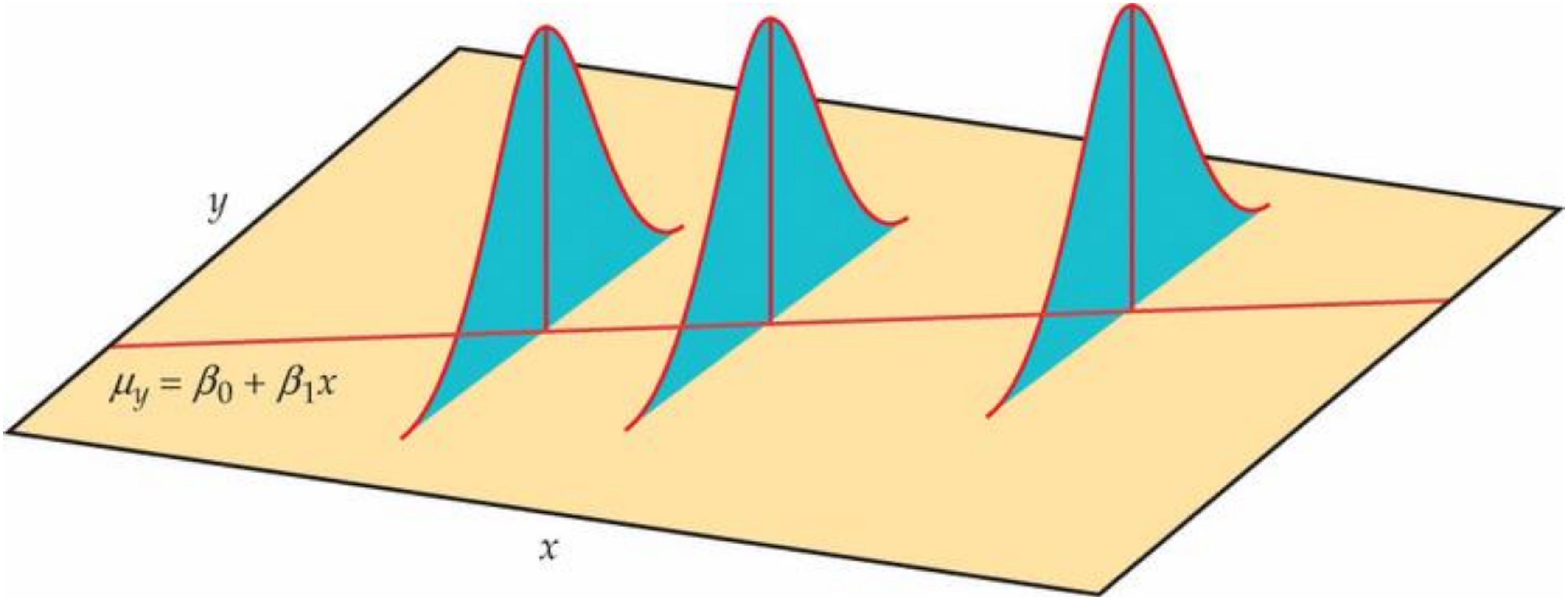
where the unknown parameters β_0 and β_1 determine the intercept and the slope, respectively, of a specific straight line.

Note: the parabolic model

$$E[Y|X = x] = \beta_0 + \beta_1 x + \beta_2 x^2$$

is also *l* nean of Y given X is related to the parameters via a linear function.

S.L.R. Models



S.L.R. Models

There will almost certainly be some variation in Y due strictly to random phenomenon that cannot be measured, predicted, or explained. This variation is called **random error**, and denoted e_i.

In other words, the difference between what actually occurs and our model is the error. Our model doesn't explain everything; the amount of variability in our response variable that remains unexplained is measured by the error.

S.L.R. Models

Suppose that Y_1, Y_2 , …, Y_n are independent realizations of the random variable *Y* that are observed at the values x_1 , x_2 , …, x_n *of a* random variable *X* . If the regression of *Y* on *X* is linear, then for *i = 1, 2, …, n;*

$$Y_i = \beta_0 + \beta_1 x_i + e_i = E[Y_i|X_i = x_i] + e_i$$

where e_i is the random error in Y_i and is such that $E(e_i \mid X) = 0$.

Continued

Notice that the random error term does not depend on *x* , nor does it contain any information about Y (otherwise it would be a systematic error).

For now, we assume the errors have constant variance:

$$Var(e_i) = \sigma^2$$

+

Introduction to Least Squares

- Don’t get these ideas confused:

$$y = \beta_0 + \beta_1 x$$

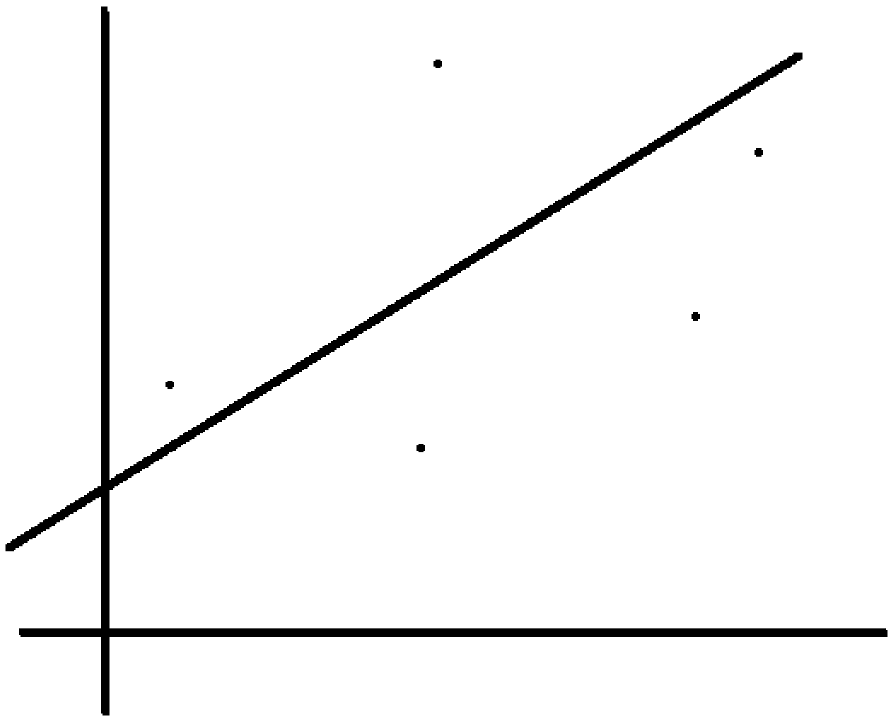
$$y_i = \beta_0 + \beta_1 x_i + e_i$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

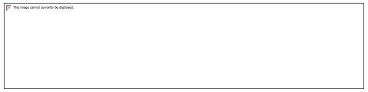
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{e}_i$$

- Applet: http://serc.carleton.edu/sp/cause/interactive/examples/17126.html

Introduction to Least Squares



Least Squares



Least Squares: Derive parameter estimates

$$R(\beta_0, \beta_1) = \sum (y_i - [\beta_0 + \beta_1 x_i])^2$$

Goal: Minimize with respect to β_0 and β_1.

$$\frac{\partial R(\beta_0, \beta_1)}{\partial \beta_0} = -\sum 2(y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial R(\beta_0, \beta_1)}{\partial \beta_1} = -\sum 2(y_i - \beta_0 - \beta_1 x_i) x_i$$

We have two equations with two unknowns. Set both equal to 0 and solve.

Least Squares: Derive parameter estimates

$$-\sum 2(y_i - \beta_0 - \beta_1 x_i) := 0$$

$$\sum y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum x_i = 0$$

$$n\hat{\beta}_0 = \sum y_i - \hat{\beta}_1 \sum x_i$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

First: derivative with respect to the intercept. Solve for the y-intercept.

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

The least squares regression line goes through the point $(\bar{x}, \bar{y})$

Least Squares: Derive parameter estimates

$$-\sum 2(y_i - \beta_0 - \beta_1 x_i)x_i := 0$$

$$\sum y_i x_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2 = 0$$

$$\sum y_i x_i - [\bar{y} - \hat{\beta}_1 \bar{x}] \sum x_i - \hat{\beta}_1 \sum x_i^2 = 0$$

$$\sum y_i x_i - \frac{\sum y_i \sum x_i}{n} + \hat{\beta}_1 \frac{(\sum x_i)^2}{n} - \hat{\beta}_1 \sum x_i^2 = 0$$

$$\hat{\beta}_1 \left(\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right) = \sum y_i x_i - \frac{\sum y_i \sum x_i}{n}$$

$$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{SXY}{SXX}$$

Second: Derivative with respect to the slope. Solve for the slope.

Applet: http://www.stat.tamu.edu/~west/ph/coreye.html

Example: R Code

```
election<-read.csv("/Users/Elizabeth/…/stat608/sp13/pct_obama.csv”)
my.lm<-lm(election$Obama_12 ~ election$Obama_08)
my.lm
```

For more information about this linear model, type

```
summary(my.lm)
anova(my.lm)
```

To plot your x and y variables on a scatterplot, use:

```
plot(x,y)
```

For more information, syntax, and sometimes examples on a function (e.g. the lm function), type:

```
?lm
```

Example: SAS Code

```
data election;
 infile ‘/Users/Elizabeth/…/stat608/sp13/pct_obama.csv’;
 input state $ Pollclose $ Electoral Obama12 Romney12 Pred $ Obama08
Mccain08;
run;

proc reg data = election;
 model Obama12 = Obama08;
run;

proc gplot data = election;
 plot y*x;
run;
```

The least squares line of best fit for the election data

```
Coefficients:
    (Intercept)  election$Obama_08
       -4.461            1.042
```

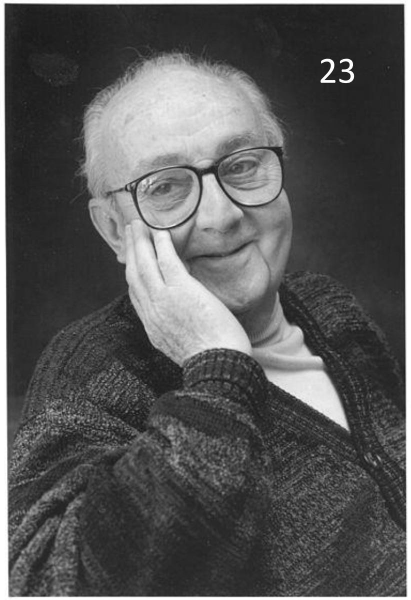
$$\hat{y} = -4.461 + 1.042x$$

Interpret the slope and intercept in context.

Least Squares: Matrix Setup

Least Squares: Derive parameter estimates in matrix notation

+ George Box



"All models are wrong, but some are useful."

+

Objectives: Slope & Intercept

Develop hypothesis tests and confidence intervals for the slope and intercept of a least squares regression model:

1. Assumptions
2. Bias of Estimates
3. Variability of Estimates

In English: Is the percent of the states that voted for Obama in 2008 linearly related to the percentage in 2012?

+

Expected values of matrices

$$\operatorname{E}[X] = \operatorname{E}\left[\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix}\right] = \begin{pmatrix} \operatorname{E}[x_{1,1}] & \operatorname{E}[x_{1,2}] & \cdots & \operatorname{E}[x_{1,n}] \\ \operatorname{E}[x_{2,1}] & \operatorname{E}[x_{2,2}] & \cdots & \operatorname{E}[x_{2,n}] \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{E}[x_{m,1}] & \operatorname{E}[x_{m,2}] & \cdots & \operatorname{E}[x_{m,n}] \end{pmatrix}.$$

Source: http://en.wikipedia.org/wiki/Expected_value#Expectation_of_matrices

Covariance Matrix

$$\Sigma_{ij} = \mathrm{cov}(X_i, X_j) = \mathrm{E}\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\ \mathrm{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \mathrm{E}[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \mathrm{E}[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}.$$

Source: http://en.wikipedia.org/wiki/Covariance_matrix

Usual Assumptions

Four assumptions:

L: Y and x are **Linearly** related. (The model must be valid! If we should be fitting a parabola, all bets are off.)

I: The errors are **Independent** of each other (e.g. random samples or randomized experiments)

N: The errors are **Normally** distributed with mean 0.

E: **Equal** variance of the errors, σ^2.

Usual Assumptions

In matrix notation:

Inferences About the Slope and Intercept

- Unbiased Estimates

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

+ Inferences About the Slope and Intercept

- Variance of the estimates

$$
\begin{aligned}
Var(\hat{\boldsymbol{\beta}}|X) &= Var((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\,Var(\mathbf{Y})\,\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\,Var(\mathbf{X}\boldsymbol{\beta}+\mathbf{e})\,\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\,\Sigma\,\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\,\sigma^2\mathbf{I}\,\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}
$$

+ Inferences About the Slope

- Distribution of slope:

$$E[\hat{\beta}_1|X] = \beta_1 \qquad Var(\hat{\beta}_1|X) = \frac{\sigma^2}{SXX}$$

Since the errors are normally distributed, $y_i = \beta_0 + \beta_1 x_i + e_i$ is also normally distributed, **given x**. Since the sample slope is a linear combination of the y_i's, it is also normally distributed.

$$z = \frac{R.V. - mean}{stdev.} = \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{SXX}}$$

+ Inferences About the Slope

Of course, σ is unknown, so we estimate it using the sample standard deviation. Then our test statistic has the t-distribution (with n – k degrees of freedom in general, k being the number of columns of X).

$$t_{n-2} = \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{SXX}}$$

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \frac{s}{\sqrt{SXX}}$$

+ Inferences About the Slope

- For our election example, R output gives:

```
Coefficients:         Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)          -5.13608   1.72705     -2.974   0.00459 **
states50$Obama_08  1.05610   0.03363     31.401   < 2e-16 ***
```

- Confidence Interval for the slope:

Inferences About the Slope

- Hypothesis Test for whether the slope = 0:

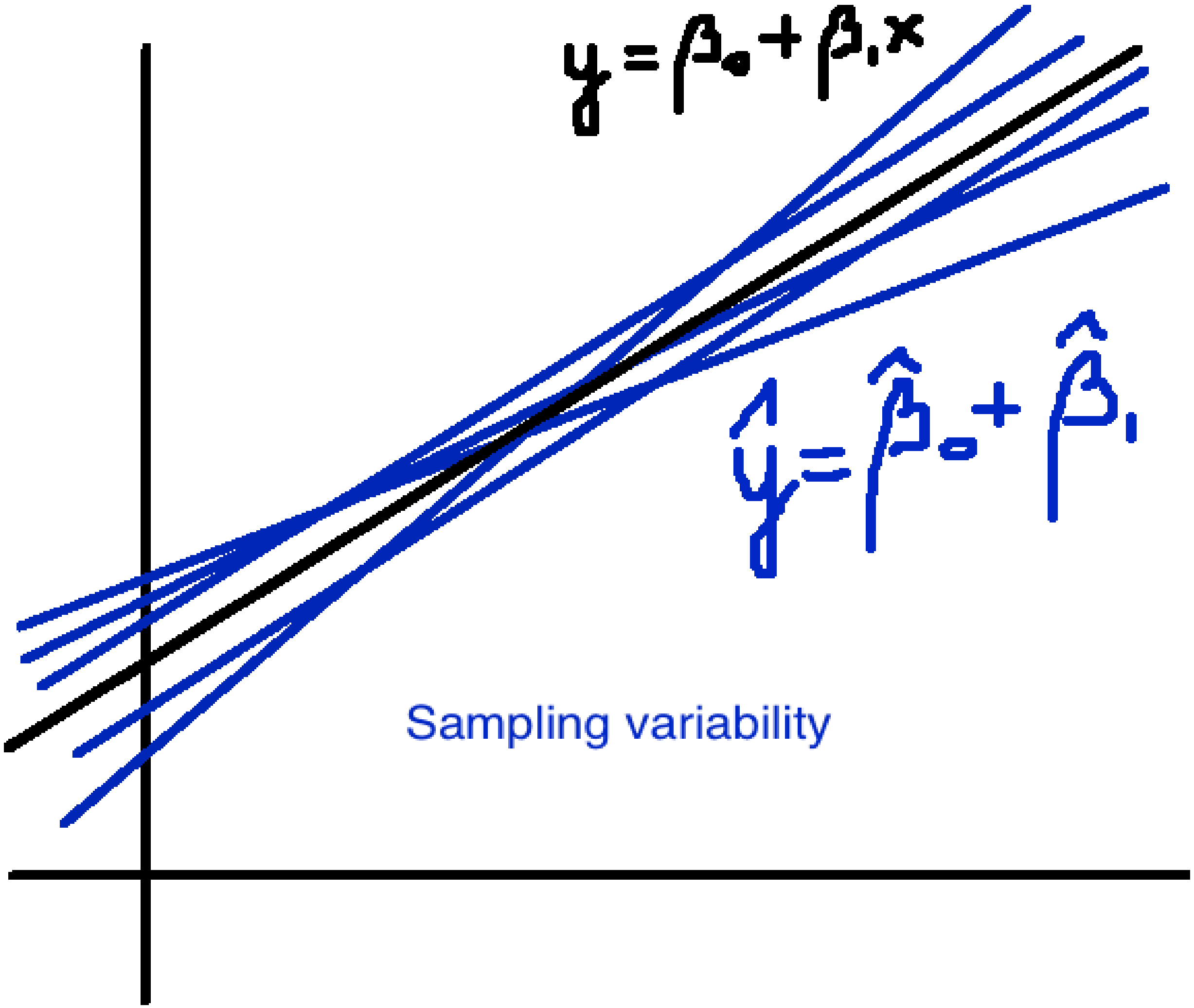
Objectives: Regression Line

Develop hypothesis tests and confidence intervals for the regression line:

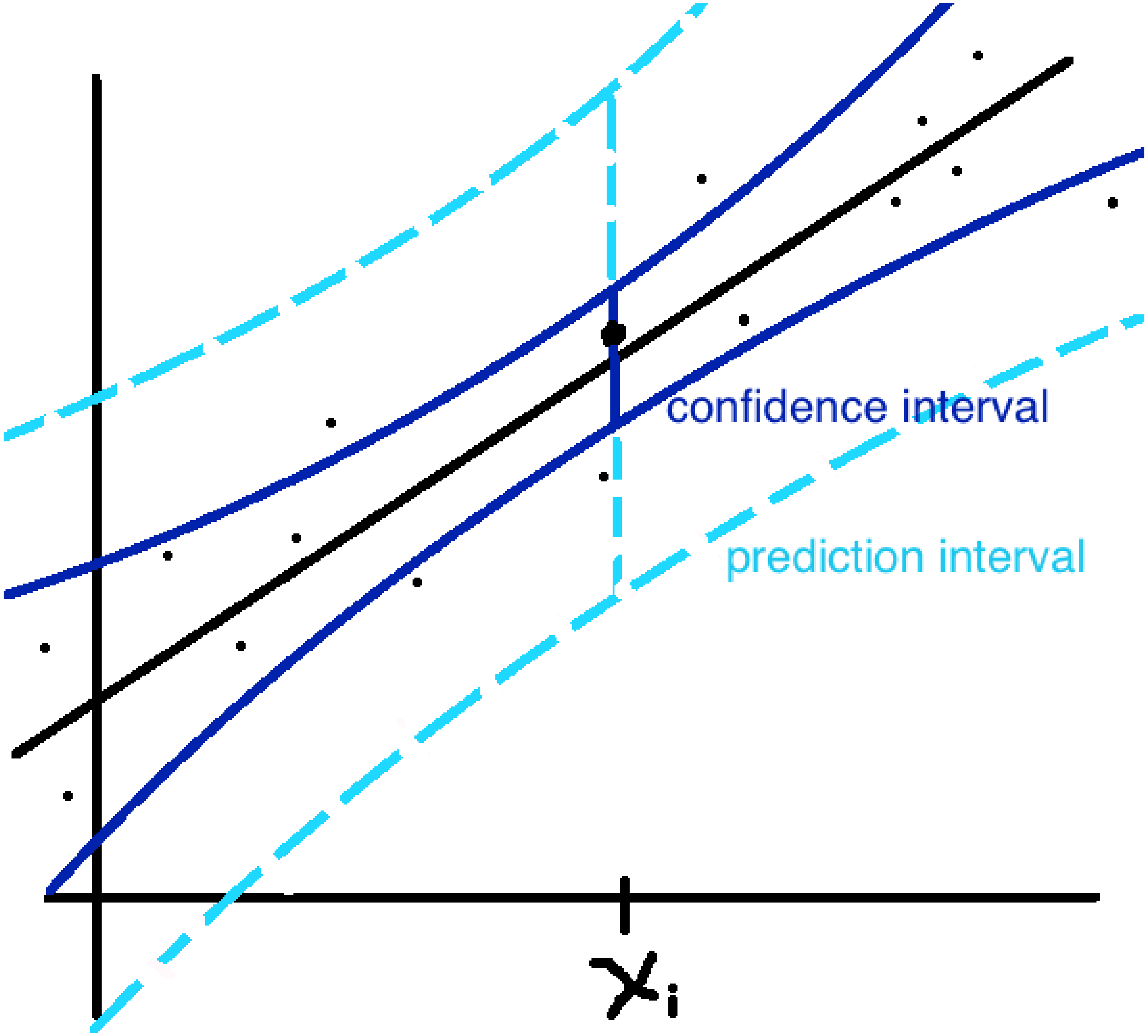
1. Confidence intervals for the mean
2. Prediction intervals for individuals

In English: If 48% of a state voted for Obama in 2008, what percentage voted for him in 2012?

Slope vs. regression line



+ Slope vs. **regression line**



+

Slope vs. regression line

- **Confidence interval for slope:**
  - If x increases by 1 unit, by how much does y increase or decrease?
  - Deals with sampling distribution of sample slope from one sample to another.
  - Can use CLT to say the sample slope is approximately normal for large n.
- **Confidence interval for the regression line (mean):**
  - For a given value of x, what is the mean value of y?
  - Deals with sampling distribution of sample mean from one sample to another.
  - Can use CLT to say the sample mean is approximately normal for large n.
- **Prediction interval for the regression line:**
  - For a given value of x, what are the individual values of y?
  - Deals with distribution of individuals about the regression line.
  - The CLT is NOT useful in this case. We need to know the distribution of the errors.

Confidence Intervals for the Population Regression Line

We consider the problem of finding a confidence interval for the unknown population mean at a given value of X, which we shall denote by x*.

First, recall that the population mean at *X* = *x** is given by

$$E[Y|X = x^*] = \beta_0 + \beta_1 x^*$$

$$E[\mathbf{Y}|\mathbf{X} = \mathbf{X}^*] = \mathbf{X}^*\boldsymbol{\beta}$$

Confidence Intervals for the Population Regression Line

- An estimator of this unknown quantity is the value of the estimated regression equation at *X* = *x**, namely*:*

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

$$E[\hat{y}^*] = \beta_0 + \beta_1 x^*$$

How many Populations are Sampled?

- One for each unique value of X = x^*
- How many populations were sampled in the election data?

How Many Means and Variances do we Have?

- ___ means
- ___ variances
- However, we assume that

$$\mu_{x^*} = \beta_0 + \beta_1 x^*$$

$$\sigma^2_{x^*} = \sigma^2$$

+ Variance of mean

$$\hat{\mathbf{y}^*} = \mathbf{X}^*\hat{\boldsymbol{\beta}}$$

$$\hat{y^*} = \begin{bmatrix} 1 & x^* \end{bmatrix} \hat{\beta}$$

$$\begin{aligned} V(\hat{\mathbf{y}^*}) &= \mathbf{X}^* V(\hat{\boldsymbol{\beta}})\mathbf{X}'^* \\ &= \sigma^2 \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'^* \end{aligned}$$

$$\begin{aligned} V(\hat{y^*}) &= \begin{bmatrix} 1 & x^* \end{bmatrix} V(\hat{\beta}) \begin{bmatrix} 1 \\ x^* \end{bmatrix} \\ &= \sigma^2 \begin{bmatrix} 1 & x^* \end{bmatrix} (\mathbf{X}'\mathbf{X})^{-1} \begin{bmatrix} 1 \\ x^* \end{bmatrix} \end{aligned}$$

+ Estimated Variance of mean

$$\hat{\mathbf{y}^*} = \mathbf{X}^*\hat{\boldsymbol{\beta}}$$

$$\hat{y^*} = \begin{bmatrix} 1 & x^* \end{bmatrix} \hat{\beta}$$

$$\begin{aligned} \hat{V}(\hat{\mathbf{y}^*}) &= \mathbf{X}^*\hat{V}(\hat{\boldsymbol{\beta}})\mathbf{X}'^* \\ &= s^2\mathbf{X}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'^* \end{aligned}$$

$$\begin{aligned} \hat{V}(\hat{y^*}) &= \begin{bmatrix} 1 & x^* \end{bmatrix} \hat{V}(\hat{\beta}) \begin{bmatrix} 1 \\ x^* \end{bmatrix} \\ &= s^2 \begin{bmatrix} 1 & x^* \end{bmatrix} (\mathbf{X}'\mathbf{X})^{-1} \begin{bmatrix} 1 \\ x^* \end{bmatrix} \end{aligned}$$

Confidence Intervals for the Population Regression Line

A 100(1 – *α)%* **confidence interval** for

$$E[Y|X = x^*] = \beta_0 + \beta_1 x^*$$

the population mean of Y at X = x*, is given by

$$\hat{y}^* \pm t_{\alpha/2,n-2} s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX}}$$

where $t_{\alpha/2,\, n-2}$ is the 100(1 – α/2)th quantile of the t-distribution with n - 2 degrees of freedom.

Confidence Intervals for the Population Regression Line

Does the confidence interval for the mean y-value get narrower as n gets bigger?

What happens to the width of the confidence interval as your desired x-value is farther from the mean for x?

$$\hat{y}^* \pm t_{\alpha/2,n-2} s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX}}$$

Prediction Intervals for the Actual Value of *Y*

$$\hat{\mathbf{e}}_i = \mathbf{y}_i^* - \hat{\mathbf{y}}_i$$

$$\mathbf{y}_i^* = \hat{\mathbf{e}}_i + \hat{\mathbf{y}}_i$$

$$\begin{aligned} Var(\hat{\mathbf{e}}_i + \hat{\mathbf{y}}_i) &= Var(\hat{\mathbf{e}}_i) + Var(\hat{\mathbf{y}}_i) \\ &= \sigma^2\mathbf{I} + \sigma^2\mathbf{X}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'^* \\ &= \sigma^2(\mathbf{I} + \mathbf{X}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'^*) \end{aligned}$$

+

Prediction Intervals for the actual value of Y

A 100(1–α)% **prediction interval** for Y^*, the value of Y at $X = x^*$, is given by

$$\hat{y}^* \pm t(\alpha/2, n-2) S \sqrt{(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX})}$$

$$= \hat{\beta}_0 + \hat{\beta}_1 \, x^* \pm t(\alpha/2, n-2) S \sqrt{(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX})}$$

$$= \hat{\beta}_0 + \hat{\beta}_1 \, x^* \pm t(\alpha/2, n-2) \sqrt{S^2(1 + [1 \;\; x^*](X^t X)^{-1}[1 \;\; x^*]^t)}$$

+

Dummy Variable Regression

Experiment: Does increasing calcium intake decrease blood pressure?

$$y_i = \alpha_0 + \alpha_1 x_i + e_i$$

$y_i =$ change in blood pressure

$$x_i = \begin{cases} 0 & \text{if placebo} \\ 1 & \text{if calcium} \end{cases}$$

For placebo group: $$y_i = \alpha_0 + e_i$$

For calcium group: $$y_i = (\alpha_0 + \alpha_1) + e_i$$

Dummy Variable Regression

For placebo group: $y_i = \alpha_0 + e_i$

For calcium group: $y_i = (\alpha_0 + \alpha_1) + e_i$

$$\mu_{placebo} = \alpha_0$$

$$\mu_{calcium} = \alpha_0 + \alpha_1$$

$$\mu_{calcium} = \mu_{placebo} + \alpha_1$$

$$\alpha_1 = \mu_{calcium} - \mu_{placebo}$$

Dummy Variable Regression

$$y_i = \alpha_0 + \alpha_1 x_i + e_i$$

$$x_i = \begin{cases} 0 & \text{if placebo} \\ 1 & \text{if calcium} \end{cases}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \qquad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix}$$

21 total men; 10 took the calcium supplement, and 11 took the placebo.

Dummy Variable Regression

- What if we had the intercept plus two variables, one an indicator of taking the placebo and one an indicator of taking the calcium supplement?

- The columns of **X** must be linearly independent; that is, the matrix **X** must be of full rank.

Dummy Variable Regression

$$\hat{\boldsymbol{\alpha}} =$$

Dummy Variable Regression

Dummy Variable Regression

```
my.lm<-lm(y~x)
summary(my.lm)
```

- Output:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2727     2.2266  -0.122    0.904
x             5.2727     3.2267   1.634    0.119
```

- Calcium group mean = 5, Placebo group mean = -0.27. Difference between means = 5.27, as expected.

Geometric Interpretation of ŷ

- Let V be the vector space spanned by the columns of our design matrix X.
- Then ŷ is the projection of the vector y down into the vector space V.
- Is **a** in V?

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad \mathbf{a} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

Geometric Interpretation of ŷ

- Projection (Hat) Matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

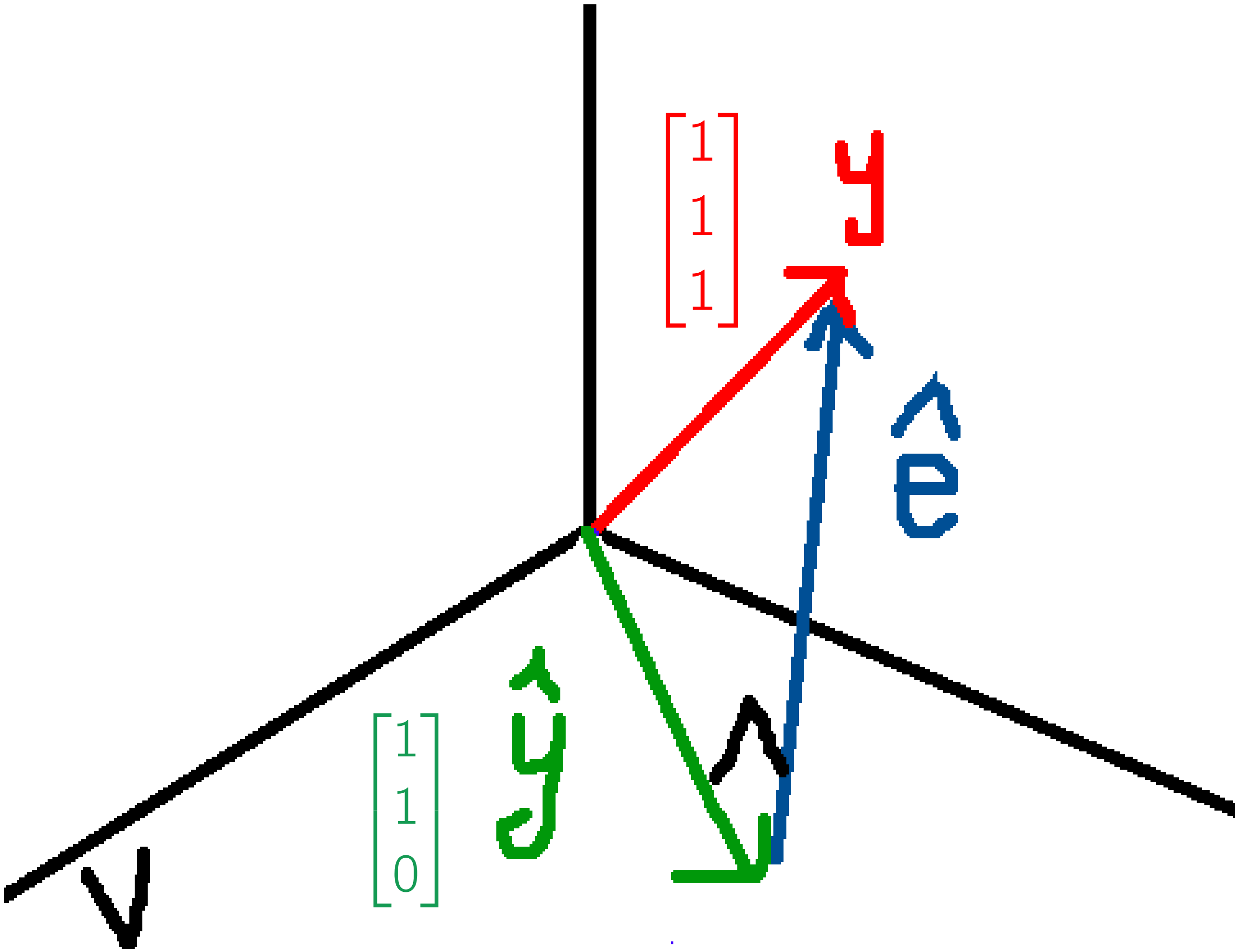
$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{H}\mathbf{y} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{X}\hat{\beta} \end{aligned}$$

- Projection is like a shadow of the real thing

Geometric Interpretation of ŷ

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

Geometric Interpretation of ŷ


$\begin{bmatrix}1\\1\\1\end{bmatrix}$
y
ê
$\begin{bmatrix}1\\1\\0\end{bmatrix}$
ŷ
V

Geometric Interpretation of ŷ

- Any vector in V multiplied by the error vector is 0:

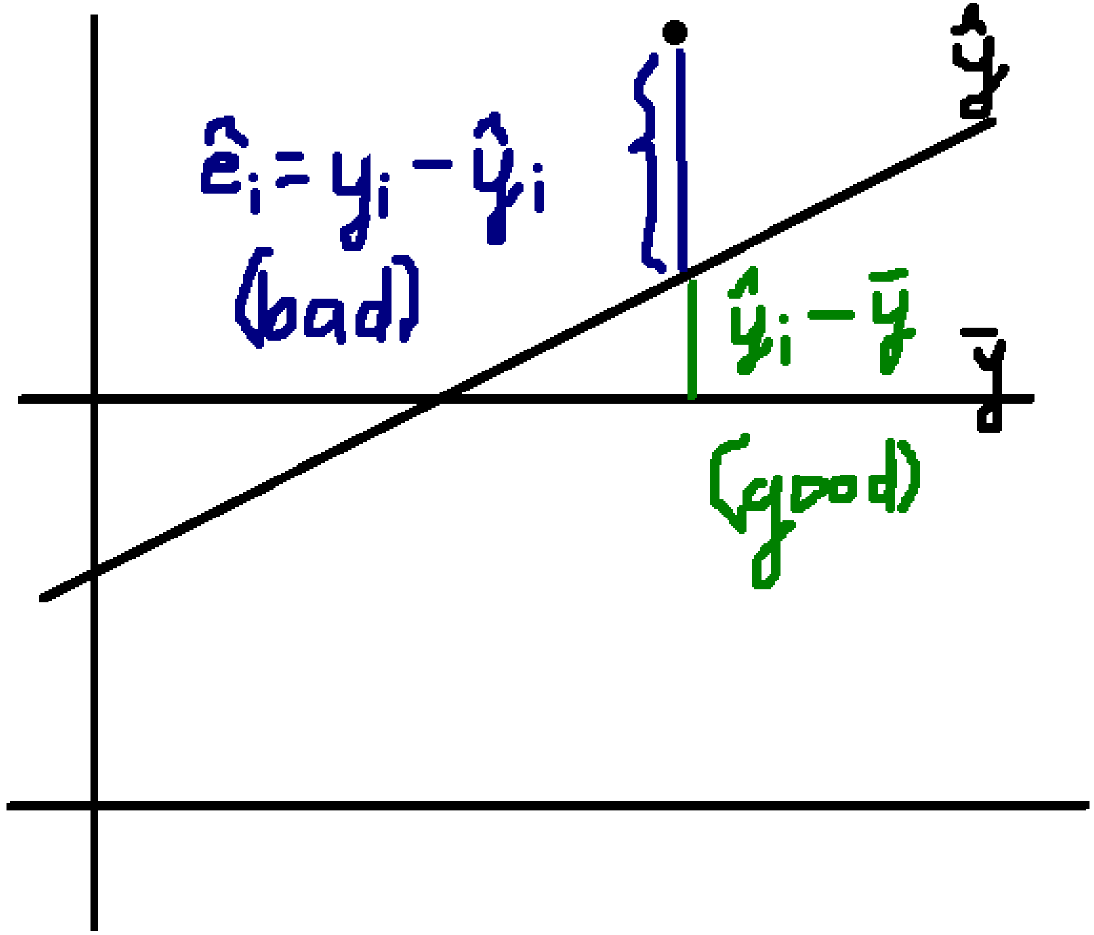
Any $\mathbf{v} \in V$ can be written as $\mathbf{Xa}$.

$$\begin{aligned}
\hat{\mathbf{e}}'\mathbf{v} &= (\mathbf{y} - \hat{\mathbf{y}})'\mathbf{Xa} \\
&= \mathbf{y}'\mathbf{Xa} - \left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right]'\mathbf{Xa} \\
&= \mathbf{y}'\mathbf{Xa} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Xa} \\
&= \mathbf{y}'\mathbf{Xa} - \mathbf{y}'\mathbf{Xa} \\
&= \mathbf{0}
\end{aligned}$$

Geometric Interpretation of ŷ

- Is it true that $\sum \hat{e}_i x_i = 0?$
- What is $\sum \hat{e}_i?$

ANalysis Of VAriance (ANOVA) Table



ANOVA Table

We’re considering two sources of variation using this regression model:

- Variation that can be explained by the model
- Variation due to randomness or other variables not considered.

Data = **Fit** **+** **Residual**

$$y_i = (\beta_0 + \beta_1 x_i) + (\varepsilon_i)$$

The regression model, whether using dummy variables or not, is not really about variances; the point in a regression model is to study how y changes with x.

ANOVA Table

- SST = RSS + SSReg

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$