BUS41100 Applied Regression Analysis

Week 2: Inference for SLR

Inference: sampling distributions, testing confidence intervals, and prediction intervals

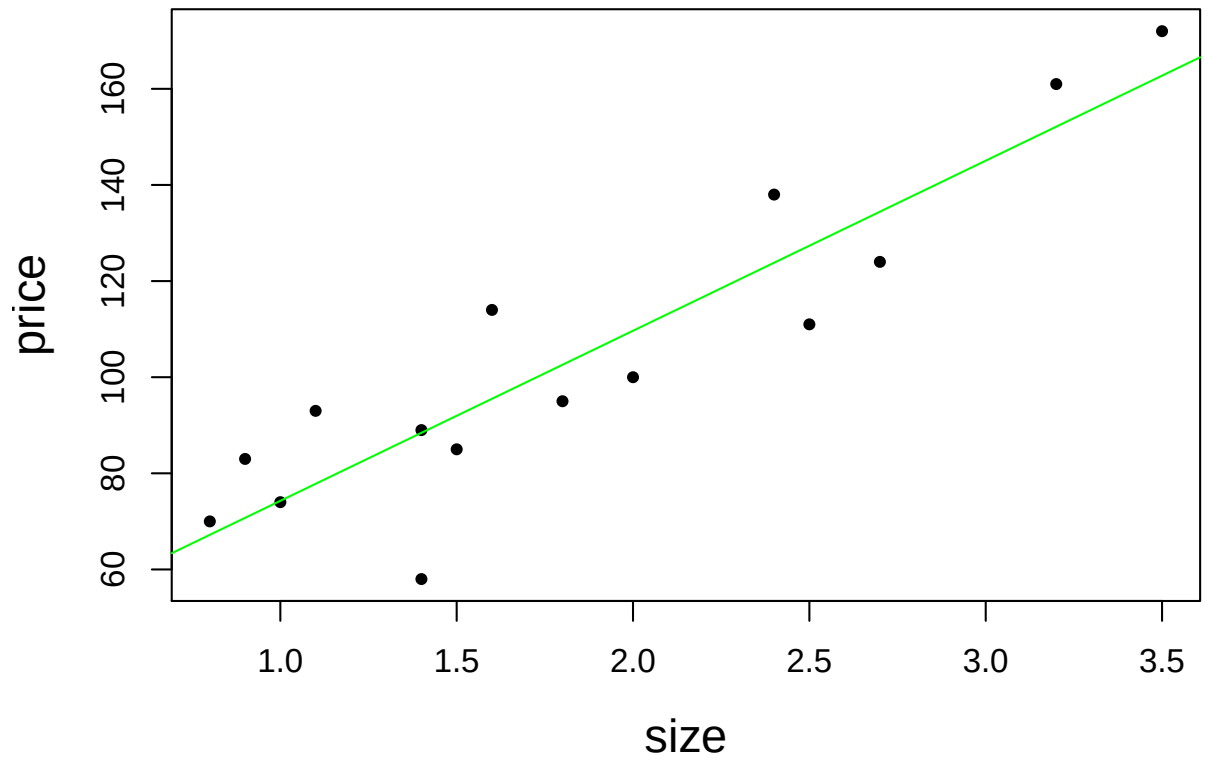
Max H. Farrell
The University of Chicago Booth School of Business

Back to House Prices

Understand the relationship between `price` and `size`. How?

Last week we fit a line through a bunch of points:

$$\texttt{price} = 39 + 35 \times \texttt{size}.$$



CAPM

Another example of conditional distributions:

Individual returns given market return.

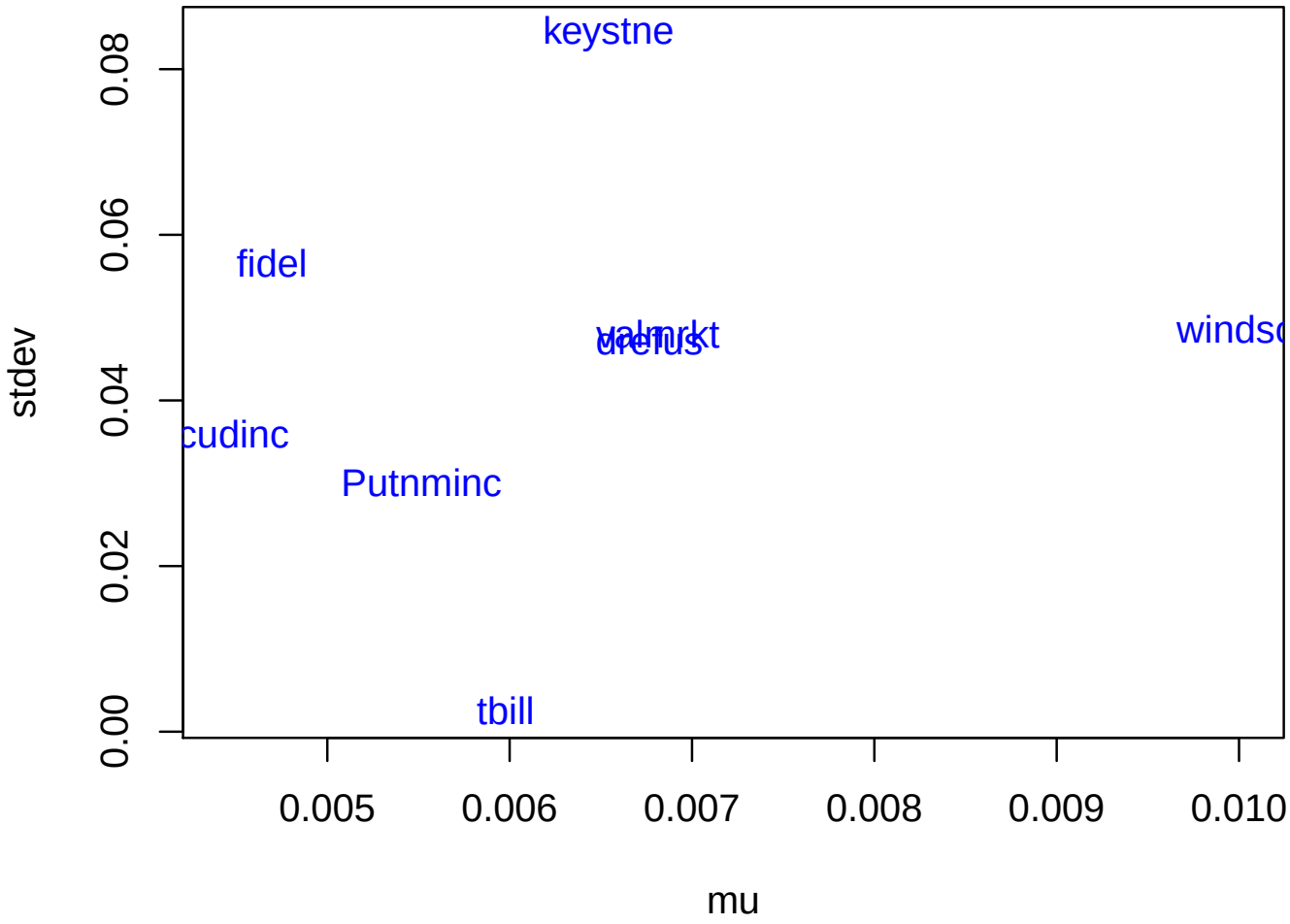
The Capital Asset Pricing Model (CAPM) for asset A relates return $R_{At} = \dfrac{V_{At} - V_{At-1}}{V_{At-1}}$ to the "market" return, R_{Mt}.

In particular, the relationship is given by the regression model $R_{At} = \alpha + \beta R_{Mt} + \varepsilon$ with observations at times $t = 1 \ldots T$ (and where $[\alpha, \beta] \equiv [\beta_0, \beta_1]$).

When asset A is a mutual fund, this CAPM regression can be used as a performance benchmark for fund managers.

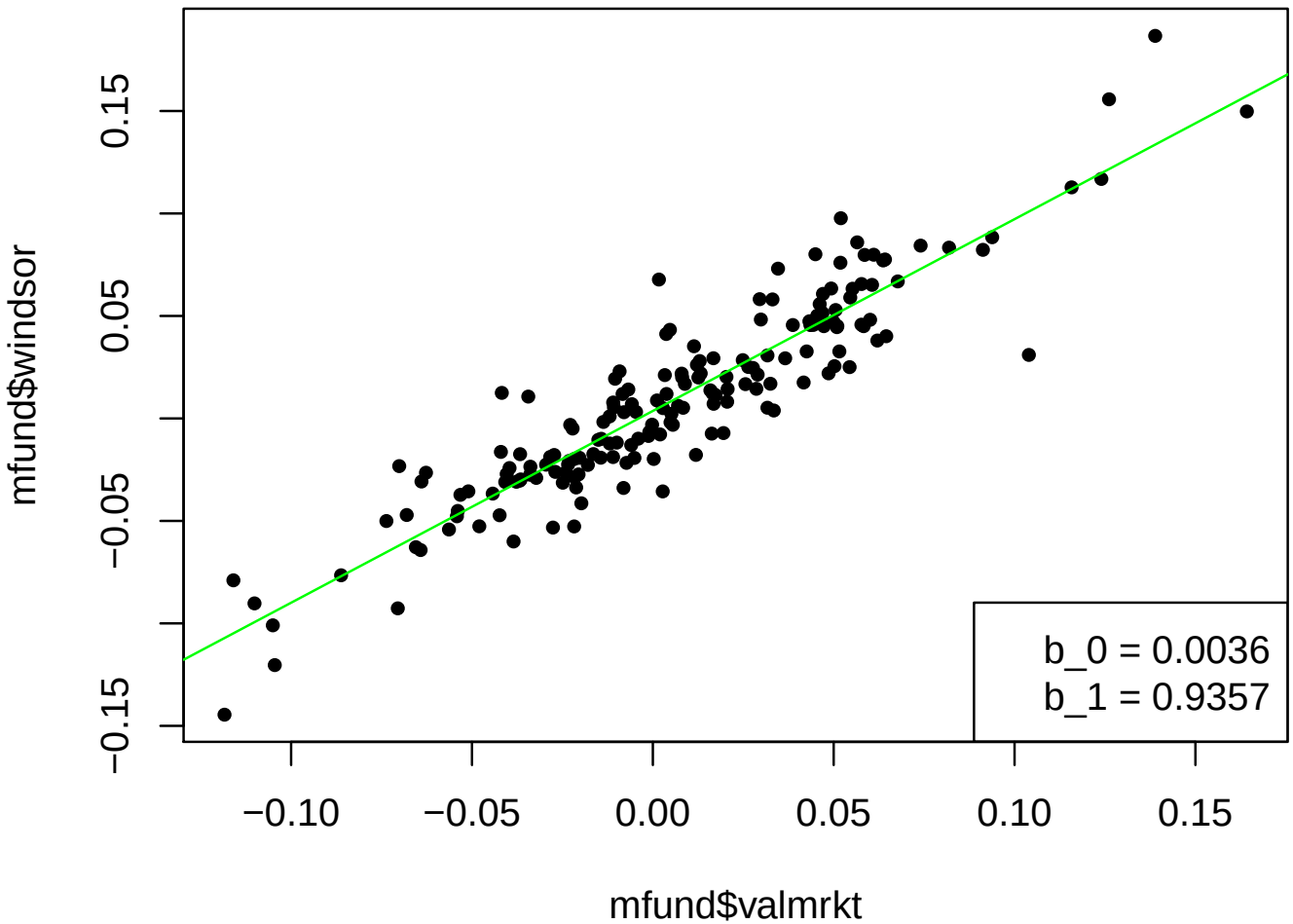
```
> mfund <- read.csv("mfunds.csv")
> mu <- apply(mfund, 2, mean)
> mu
     drefus       fidel     keystne    Putnminc     scudinc
0.006767000 0.004696739 0.006542550 0.005517072 0.004432333
    windsor     valmrkt       tbill
0.010021906 0.006812983 0.005978333

> stdev <- apply(mfund, 2, sd)
> stdev
     drefus       fidel     keystne    Putnminc     scudinc
0.047237111 0.056587091 0.084236450 0.030079074 0.035969261
    windsor     valmrkt       tbill
0.048639473 0.048000146 0.002522863
```

```
> plot(mu, stdev, col=0)
> text(x=mu, y=stdev, labels=names(mfund), col=4)
```



Lets look at just windsor (which dominates the market).

```
> windsor.reg <- lm(mfund$windsor ~ mfund$valmrkt)
> plot(mfund$valmrkt, mfund$windsor, pch=20)
> abline(windsor.reg, col="green")
```



Modeling goals

Prediction	Model
$\hat{Y} = b_0 + b_1 X$	$Y = \beta_0 + \beta_1 X + \varepsilon$
$Y = b_0 + b_1 X + e$	

Why are we running regressions anyway?

1. Properties of β_k
   - Sign: Does Y go up when X goes up?
   - Magnitude: By how much?
2. Predicting Y
   - Best guess for Y given X.

Key question today: how uncertain are our answers?

- First we must formalize our model.

Simple linear regression (SLR) model

$$Y = \beta_0 + \beta_1 X + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

What's important?

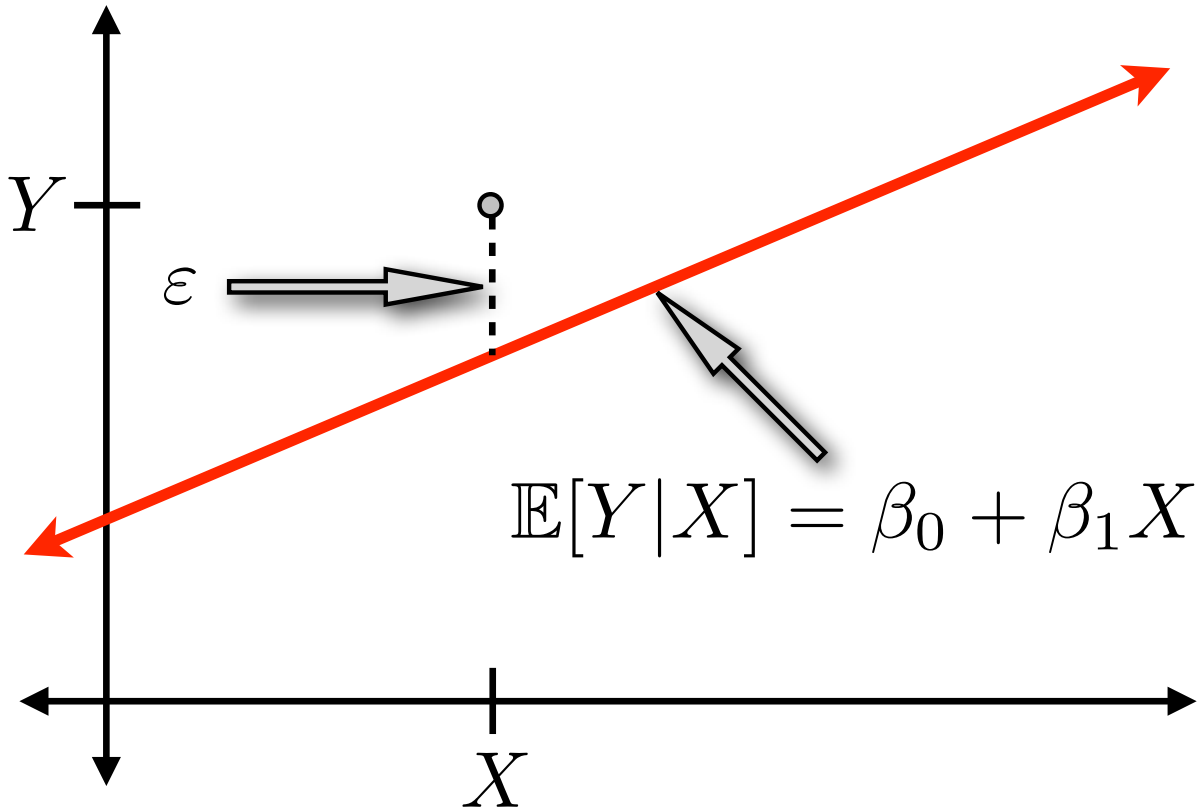
- It is a model, so we are *assuming* this relationship holds for some fixed but unknown values of β_0, β_1.
- It is linear.
- The error ε is independent & mean zero
  1. $\mathbb{E}[\varepsilon] = 0 \Leftrightarrow \mathbb{E}[Y|X] = \beta_0 + \beta_1 X$
  2. Fixed but unknown variance σ^2; constant over X
  3. Most things are approx. Normal (Central Limit Theorem)
  4. ε represents anything left, not captured in linear fcn of X
- It just works! *This is a very robust model for the world.*

Before looking at any data, the model specifies

- how Y varies with X on average: $\mathbb{E}[Y|X] = \beta_0 + \beta_1 X$;

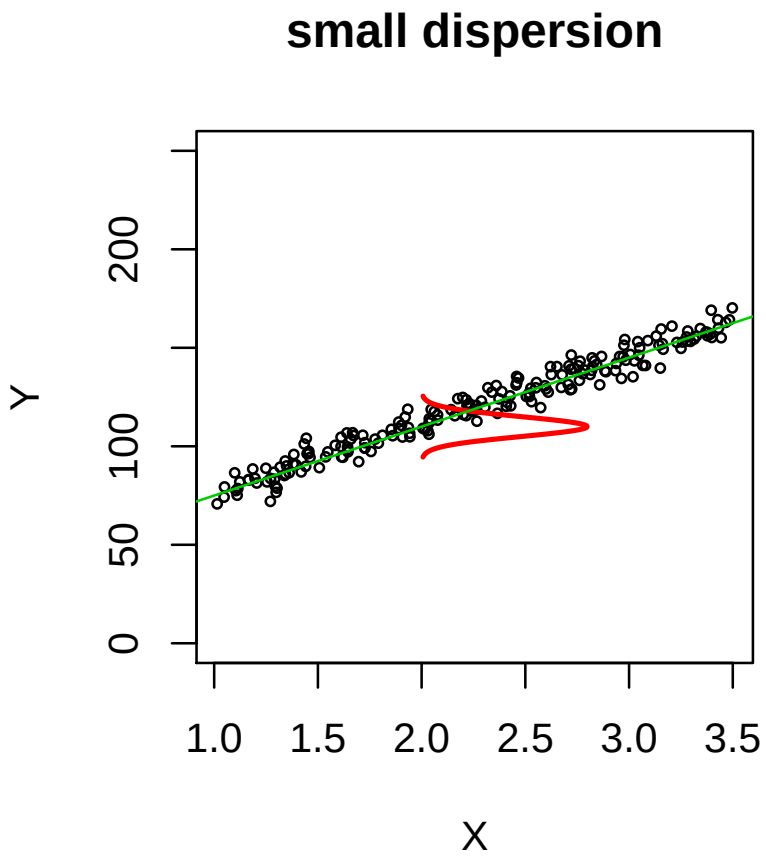
  i.e. what's the trend?

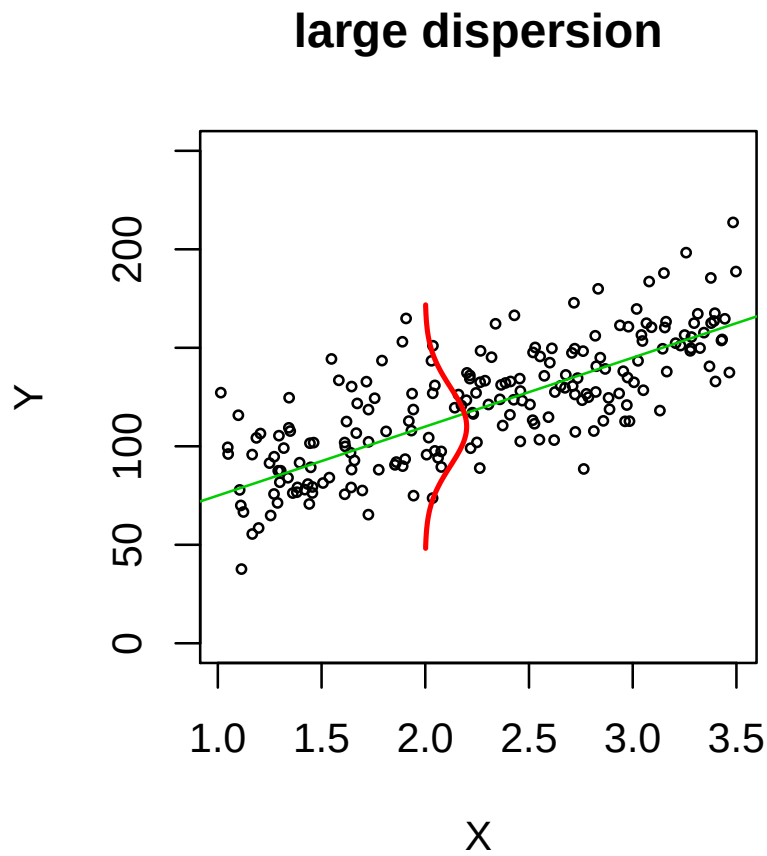
- and the influence of factors other than X, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ independently of X.



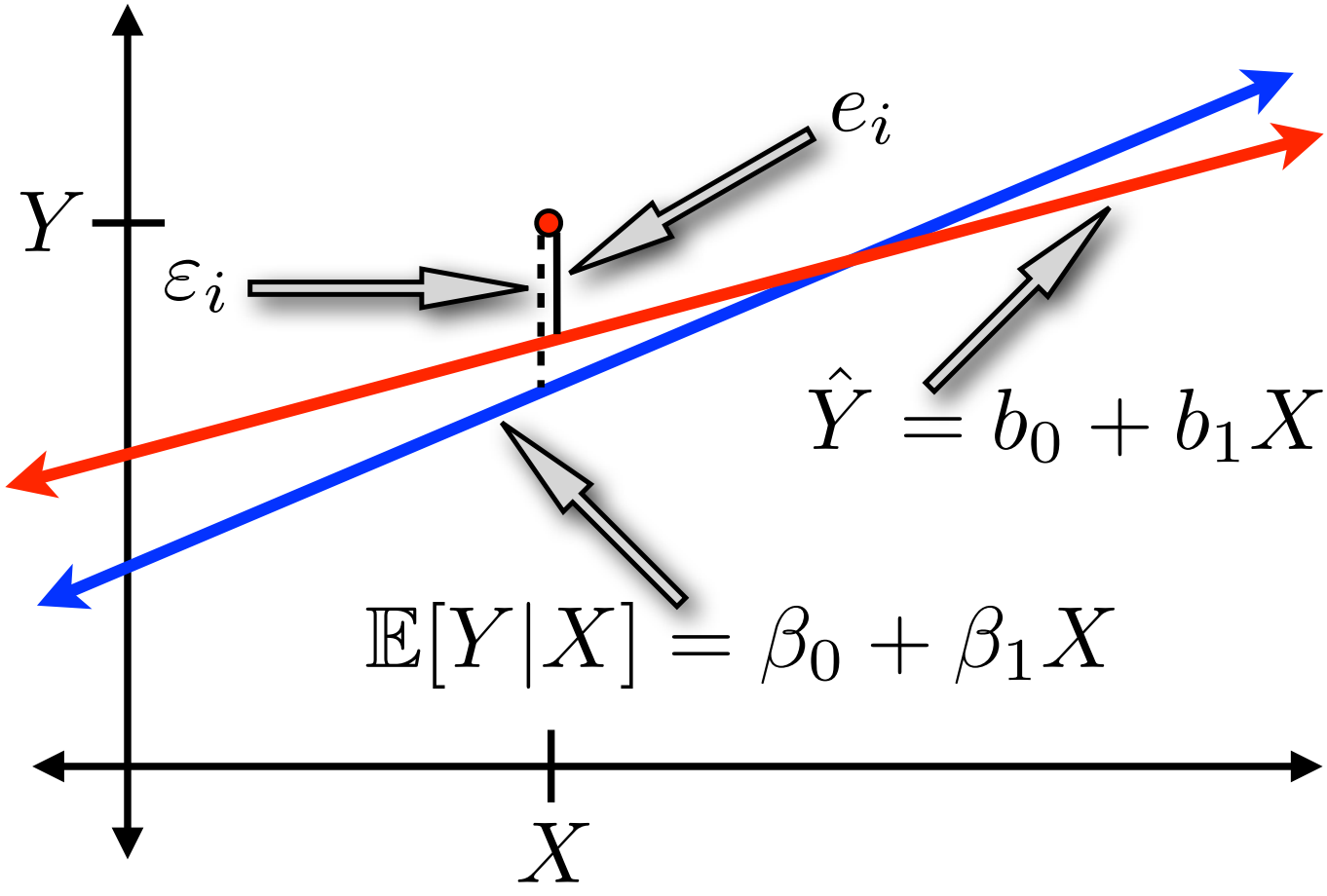
The variance σ^2 controls the dispersion of Y around $\beta_0 + \beta_1 X$

- think signal-to-noise





IMPORTANT! β_0 is not b_0, β_1 is not b_1, and ε_i is not e_i



(We use Greek letters remind to us.)

Context from the house data example

$\mathbb{E}[Y|X]$ is the average `price` of houses with `size` X, and σ^2 is the spread around that average.

When we specify the SLR model we say that

- the average house `price` is linear in its `size`, but we don't know the coefficients.
- Some houses could have a higher than expected value, some lower, but the amount by which they differ from average is unknown and
  - is independent of the `size`,
  - and is Normal.

Question: At an open house: is this house priced fairly?

Context from the CAPM example

$\mathbb{E}[Y|X]$ is the average `return` of the asset when the `market return` is X, and σ^2 is the spread around that average.

When we specify the SLR model we say that

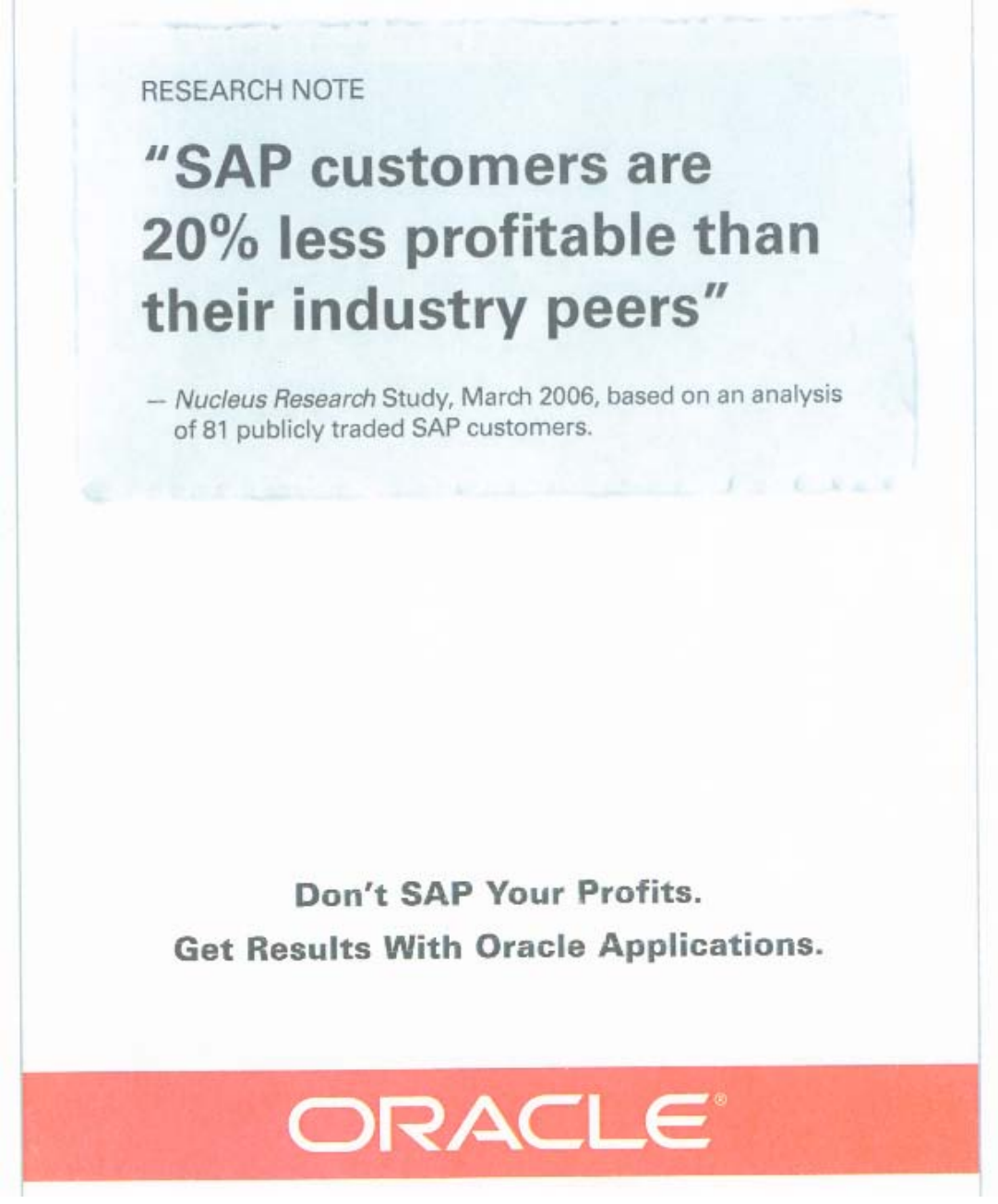
- the average asset `return` is linear in the market `return`, but we don't know the coefficients.
- Some days could have a higher than expected value, some lower, but the amount by which they differ from average is unknown and
  - is independent of the `market return`,
  - and is Normal.

Question: Does this asset follow the market? (Is $\beta = 1$?)

Detour / example:

Oracle v. SAP

Uncertainty Matters!



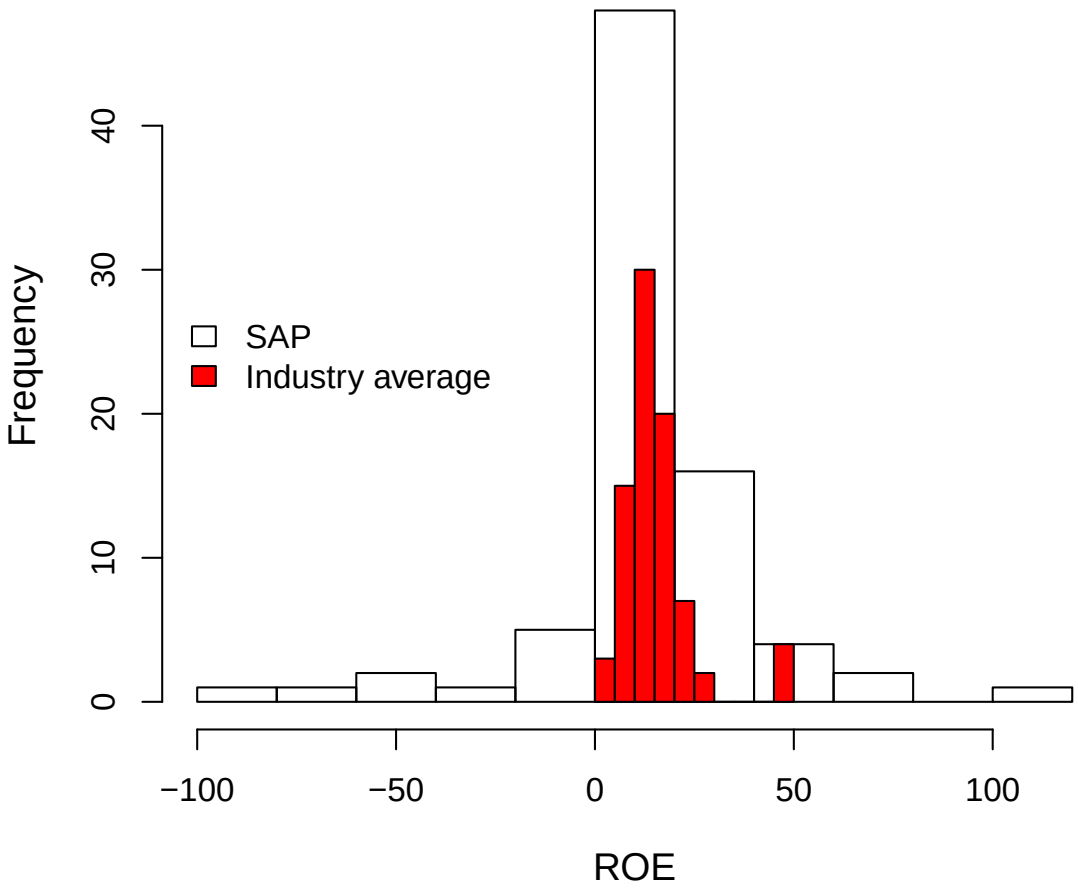
```
> sap <- read.csv("sap.csv")
> m.sap <- mean(sap$ROE)
> m.I <- mean(sap$IndustryROE)
> m.sap / m.I
[1] 0.8049701
```

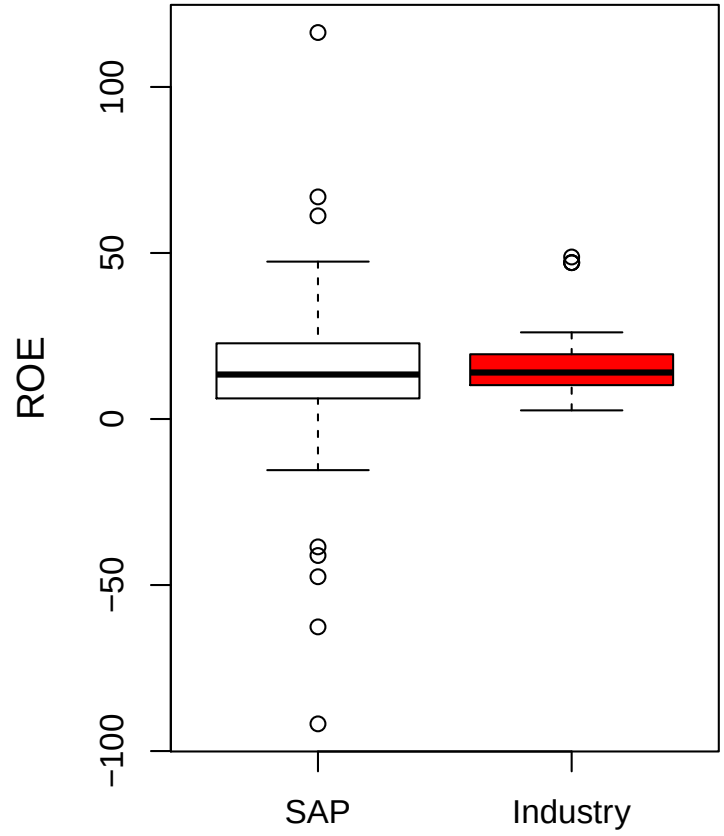
That's the mean, what about the spread?

```
> summary(sap[,4:5])
      ROE             IndustryROE
 Min.   :-91.80     Min.   : 2.6
 1st Qu.:  6.20     1st Qu.:10.2
 Median : 13.40     Median :14.0
 Mean   : 12.64     Mean   :15.7
 3rd Qu.: 22.80     3rd Qu.:19.5
 Max.   :116.40     Max.   :48.8
```

What's going on here?

- SAP ROE is more variable than average Industry ROE.
  $\hookrightarrow$ Makes sense, averages are less variable than atoms
- What about large values (positive and negative)?





Uncertainty matters!

Do we even think that SAP use *is correlated with* lower ROE?

- Probably not, given the above results

But even beyond statistical uncertainty:

- Does SAP use *cause* ROE to fall?
- Were the SAP ROEs selected at random in the industry?

Statistical uncertainty is the only kind we can quantify. In any analysis there is a lot we aren't sure about:

- Do we have the right data?
- Do we have the "right" (useful?) model?
- What assumptions are we making?

Sampling distribution of LS estimates

We think of the data as being one possible realization of data that *could* have been generated from the model

$$Y|X \sim \mathcal{N}(\beta_0 + \beta_1 X, \sigma^2).$$

- How much do our estimates depend on the particular random sample that we happen to observe?
  - Different data $\Rightarrow$ different b_0 and b_1
  - Always the same β_0 and β_1.

If the estimates don't vary much from sample to sample, then it doesn't matter which sample you happen to observe.

If the estimates do vary a lot, then it matters which sample you happen to observe.

How do we know what would happen with other realizations?

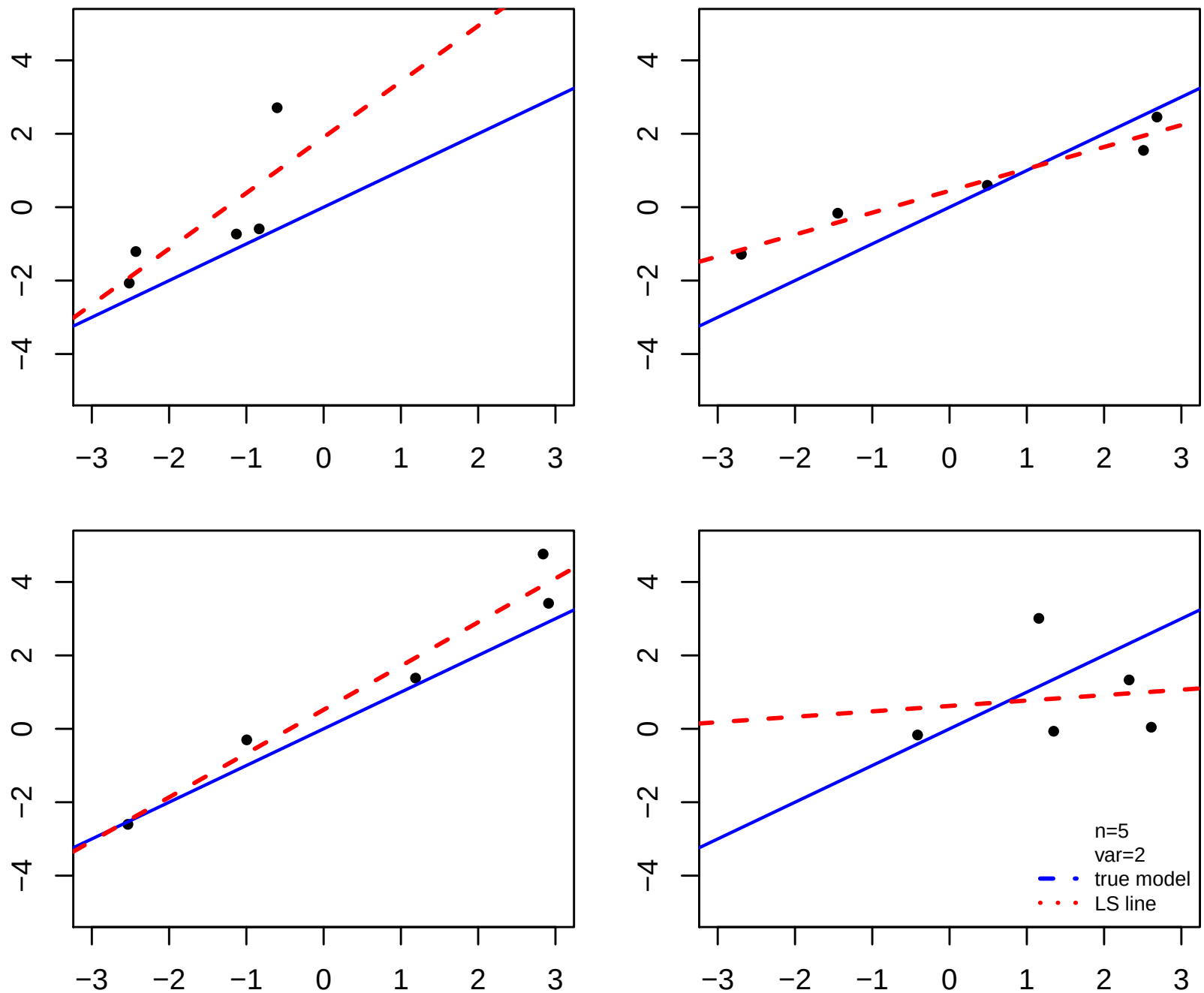
We pretend!

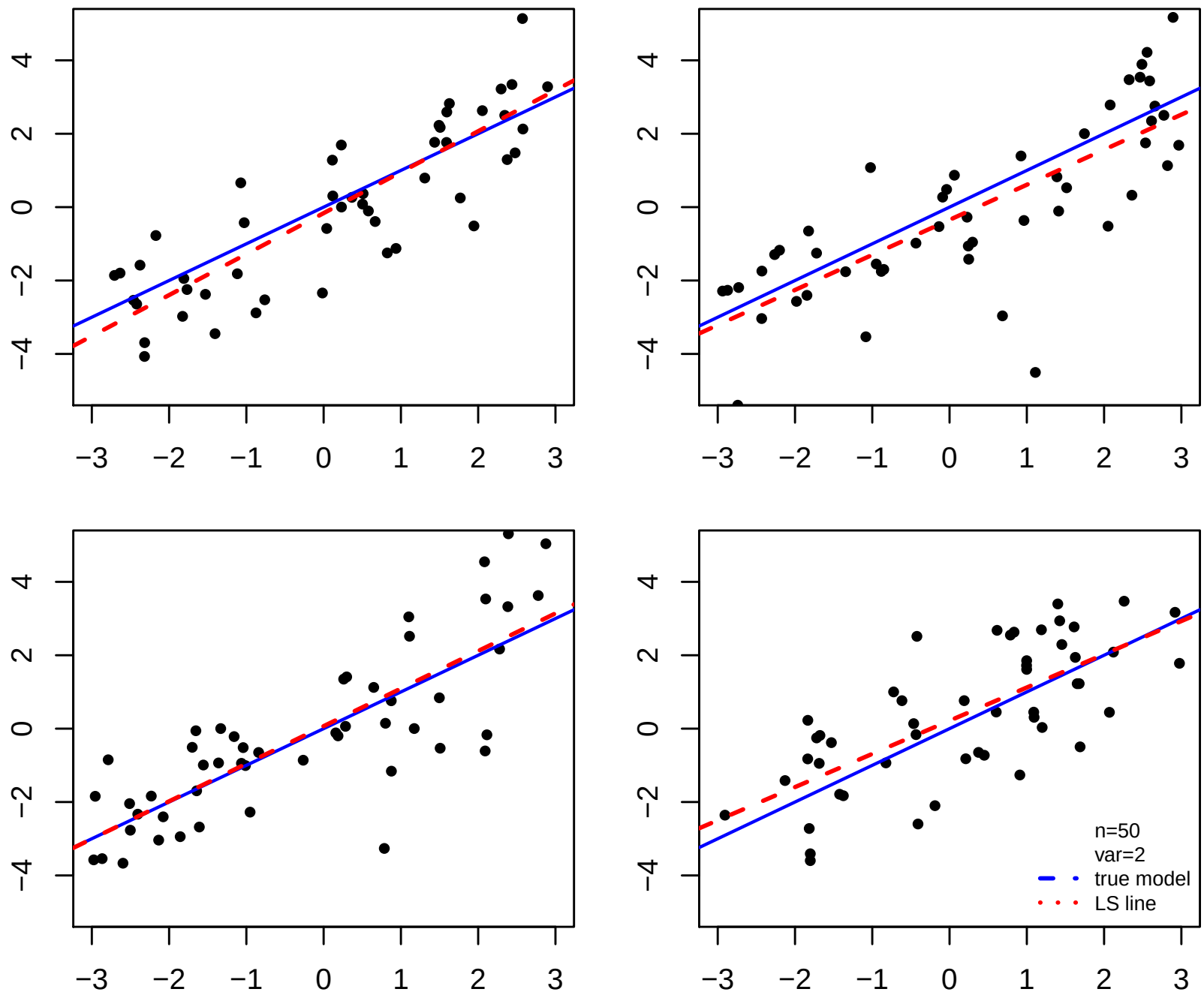
1. Randomly draw **new** data
2. Compute the **estimates** b_0 and b_1
3. Repeat

Or we use statistics to tell us:

- What the sampling distribution is . . .
- . . . and how to use it to measure uncertainty.
  - Testing, confidence intervals, etc.

But first let's see it!


n=5
var=2
true model
LS line


n=50
var=2
true model
LS line

Sampling distribution of LS estimates

What did we just do?

- We "imagined" through simulation the sampling distribution of a LS line.

What did we learn?

- Looked pretty Normal!
- When $n = 5$, some lines are close, others aren't:

  we need to get lucky.
- The lines are much closer to the truth when $n = 50$.
- The variance σ^2 matters a lot!

What happens in real life?

- We get just one data set, and we don't know the true generating model.
- But we can still imagine ...

...and use statistics!

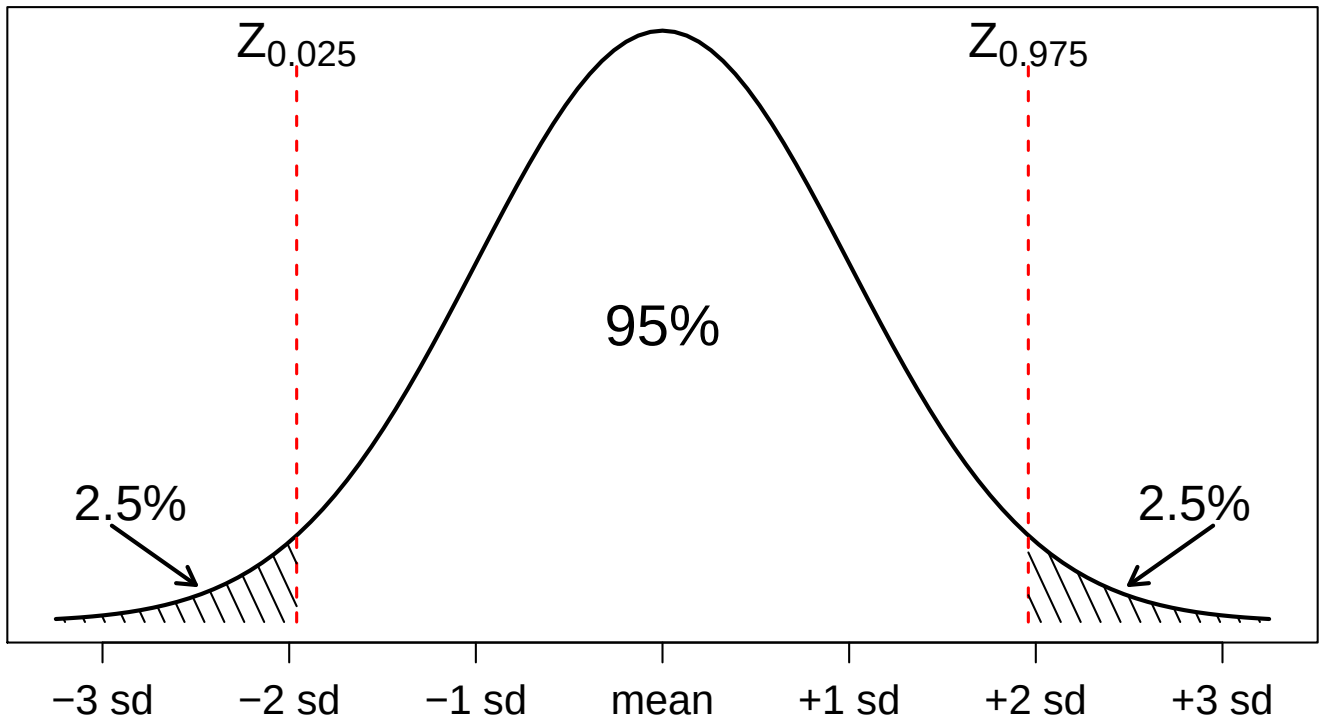
- Quantify how n and σ^2 matter
- Quantify uncertainty

only within our model.

Normal Distribution – Quick Review

Why do we like the Normal distribution?

- Symmetric
- Concentration around the mean!
  - ↪ 95% of the data within 2 s.d.



Sampling distribution of b_1

It turns out that b_1 is Normally distributed: $b_1 \sim \mathcal{N}(\beta_1, \sigma^2_{b_1})$.

- b_1 is unbiased: $\mathbb{E}[b_1] = \beta_1$.
- The sampling sd σ_{b_1} determines precision of b_1:

$$\sigma^2_{b_1} = \text{var}(b_1) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2} = \frac{\sigma^2}{(n-1)s^2_x}.$$

It depends on three factors:

1. sample size (n)
2. error variance ($\sigma^2 = \sigma^2_\varepsilon$), and
3. X-spread (s^2_x).

(*We don't have time to do detailed proofs, but there is an extensive handout on my website; see also the Sheather book.*)

Sampling distribution of b_0

The intercept is also normal and unbiased: $b_0 \sim \mathcal{N}(\beta_0, \sigma^2_{b_0})$, where

$$\sigma^2_{b_0} = \text{var}(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s^2_x} \right).$$

What is the intuition here?

$$\text{var}(\bar{Y} - \bar{X}b_1) = \text{var}(\bar{Y}) + \bar{X}^2\text{var}(b_1) \; - \; 2\bar{X}\text{cov}(\bar{Y}, b_1)$$

- $\bar{Y}$ and b_1 are uncorrelated because the slope (b_1) is invariant if you shift the data up or down $(\bar{Y})$.

Joint distribution of b_0 and b_1

We know that b_0 and b_1 *can be* dependent, i.e.,

$$\mathbb{E}[(b_0 - \beta_0)(b_1 - \beta_1)] \neq 0.$$

This means that estimation error in the slope is correlated with the estimation error in the intercept.

$$\operatorname{cov}(b_0, b_1) = -\sigma^2 \left(\frac{\bar{X}}{(n-1)s_x^2} \right)$$

- Usually, if the slope estimate is too high, the intercept estimate is too low (negative correlation).
- The correlation decreases with more X spread (s_x^2).

Estimation of error variance

The formulas aren't practicable since they involve an unknown quantity: $\sigma = \sigma_\varepsilon$. Replace with:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \qquad \text{or} \qquad s^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2 = \frac{SSE}{n-p}$$

(p is the number of regression coefficients; i.e. 2 for $\beta_0 + \beta_1$).

It is often convenient to report $\hat{\sigma}$ or s, which are in the same units as Y.

Plug in for σ in any formula, e.g.

$$\sigma^2_{b_1} = \frac{\sigma^2}{(n-1)s_x^2} \qquad \Rightarrow \qquad s^2_{b_1} = \frac{s^2}{(n-1)s_x^2}$$

- Small $s^2_{b_j}$ values mean high info/precision/accuracy.

Example: revisit the house price/size data

```
> summary(house.reg)

Call:
lm(formula = price ~ size)

Residuals:
    Min      1Q  Median      3Q     Max
-30.425  -8.618   0.575  10.766  18.498

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.885      9.094   4.276 0.000903 ***
size          35.386      4.494   7.874 2.66e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.14 on 13 degrees of freedom
Multiple R-squared:  0.8267,    Adjusted R-squared:  0.8133
F-statistic:    62 on 1 and 13 DF,  p-value: 2.66e-06
```

Testing

Suppose we think that the true β_j is equal to some value β_j^0 (often 0). Does the data support that guess?

We can rephrase this in terms of competing hypotheses.

$$\text{(Null) } H_0 : \beta_j = \beta_j^0$$
$$\text{(Alternative) } H_1 : \beta_j \neq \beta_j^0$$

Our hypothesis test will either reject or fail to reject the null hypothesis

- If the hypothesis test rejects the null hypothesis, we have statistical support for our claim
- Gives only a "yes" or "no" answer!
- You choose the "probability" of false rejection: α

We use b_j for our test about β_j.

- Reject H_0 if b_j is "far" from β_j^0; assume H_0 when close
- What we really care about is:

how many standard errors b_j is away from β_j^0

The t-statistic for this test is

$$z_{b_j} = \frac{b_j - \beta_j^0}{s_{b_j}} \overset{H_0}{\sim} \mathcal{N}(0,1).$$

"Big" $|z_{\beta_j}|$ makes our guess β_j^0 look silly $\Rightarrow$ reject

- If H_0 is true, then $\mathbb{P}[|z_{b_j}| > 2] < 0.05 = \alpha$

But:

$$|z_{\beta_j}| = \left|\frac{b_j - \beta_j^0}{s_{b_j}}\right| > 2 \quad \Leftrightarrow \quad \beta_j^0 \notin (b_j \pm 2s_{b_j})$$

Confidence intervals

Since $b_j \sim \mathcal{N}(\beta_j, \sigma^2_{b_j})$,

$$1 - \alpha = \mathbb{P}\left[z_{\alpha/2} < \frac{b_j - \beta_j}{s_{b_j}} < z_{1-\alpha/2}\right]$$
$$= \mathbb{P}\Big[\beta_j \in (b_j \pm z_{\alpha/2} s_{b_j})\Big]$$

Why should we care about confidence intervals?

- The confidence interval completely captures the information in the data about the parameter.
  - Center is your estimate
  - Length is how sure you are about your estimate
  - *Any value outside would be rejected by a test!*

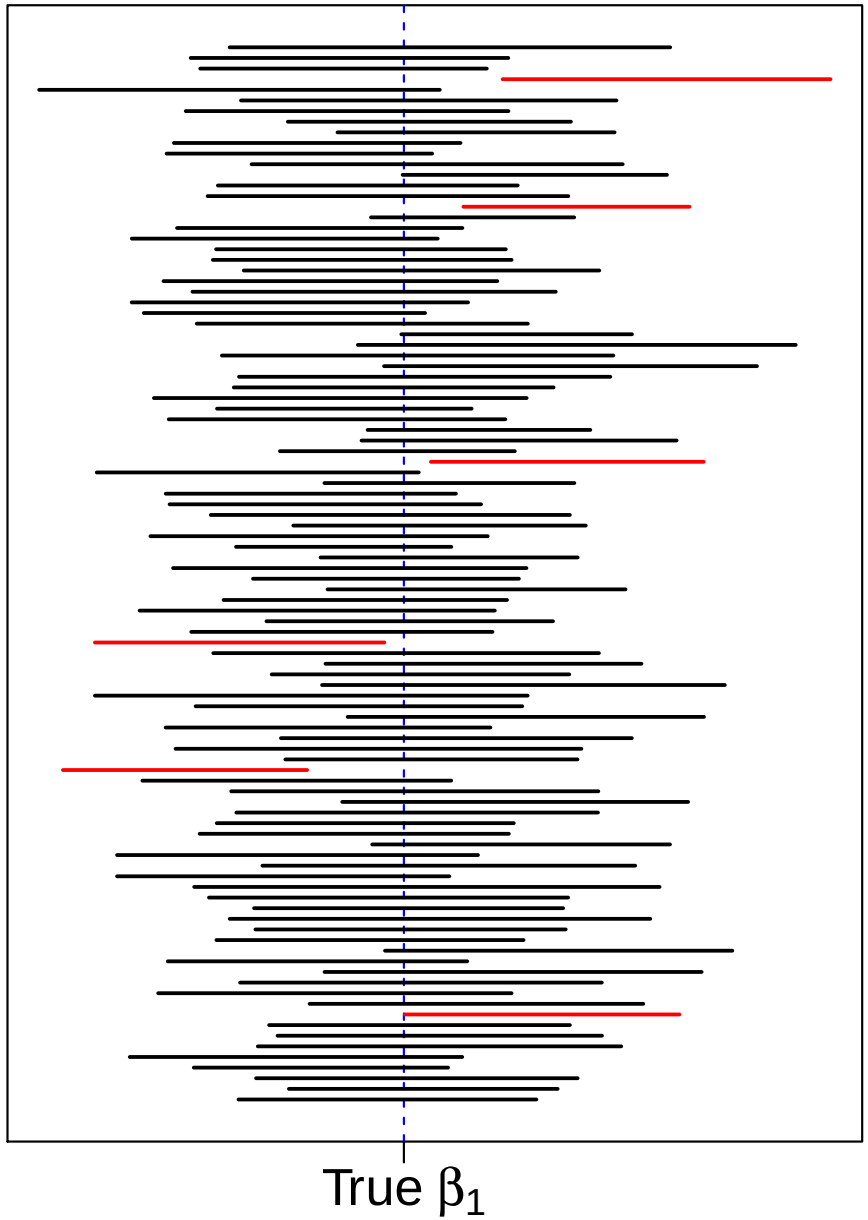
Real life or pretend?

$$\mathbb{P}\Big[\beta_1 \in (b_1 \pm 2\sigma_{b_1})\Big] = 95\%$$

or

$$\mathbb{P}\Big[\beta_1 \in (b_1 \pm 2\sigma_{b_1})\Big] = 0 \text{ or } 1$$

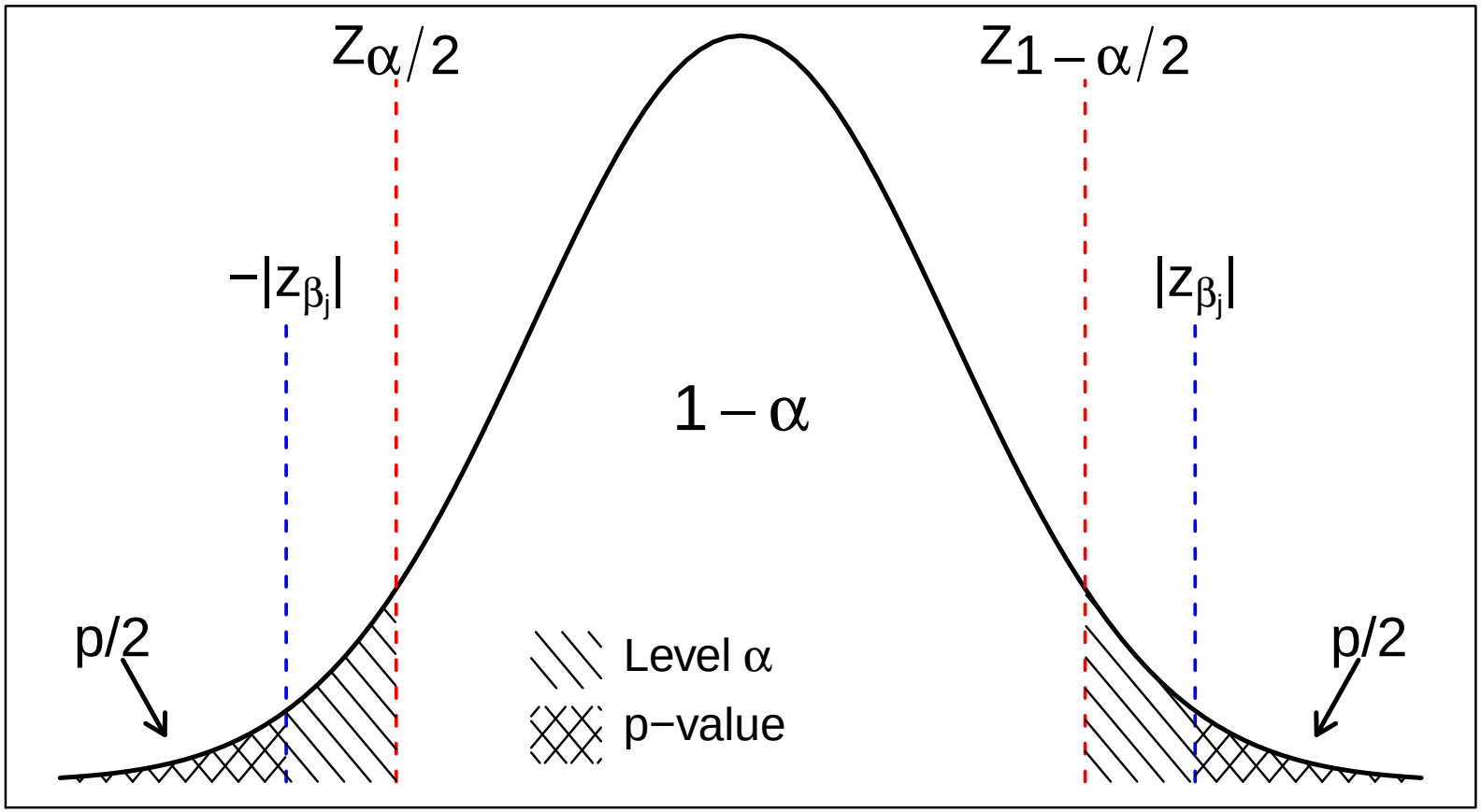
?



Level, Size, and p-values

The p-value is $\mathbb{P}[|Z| > |z_{\beta_j}|]$.

- Test with size/level $=$ p-value *almost* rejects
- CI of level $1 - (p\text{-value})$ *just* excludes $|z_{\beta_j}|$



Example: revisit the CAPM regression for the Windsor fund.

Does Windsor have a non-zero intercept?
(i.e., does it make/lose money independent of the market?).

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

- Recall: the intercept estimate b_0 is the stock's "alpha"

```
> summary(windsor.reg)  ## output abbreviated
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.003647   0.001409   2.588   0.0105 *
mfund$valmrkt 0.935717   0.029150  32.100   <2e-16 ***
> 2*pnorm(-abs(0.003647/.001409))
[1] 0.009643399
```

We reject the null at $\alpha = .05$, Windsor does have an "alpha" over the market.

- Why set $\alpha = .05$? What about at $\alpha = 0.01$?

Now let's ask whether or not Windsor moves in a different way than the market (e.g., is it more conservative?).

- Recall that the estimate of the slope b_1 is the "beta" of the stock.

This is a rare case where the null hypothesis is not zero:

$H_0 : \beta_1 = 1$, Windsor is just the market ($+$ alpha).

$H_1 : \beta_1 \neq 1$, Windsor softens or exaggerates market moves.

This time, R's output t/p values are not what we want (why?).

```
> summary(windsor.reg)  ## output abbreviated
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.003647   0.001409   2.588   0.0105 *
mfund$valmrkt 0.935717   0.029150  32.100   <2e-16 ***
```

But we can get the appropriate values easily:

- Test and p-value:

```
> b1 <- 0.935717; sb1 <- 0.029150
> zb1 <- (b1 - 1)/sb1
[1] -2.205249
> 2*pnorm(-abs(zb1))
[1] 0.02743665
```

- Confidence Interval

```
> confint(windsor.reg, level=0.95)
                     2.5 %      97.5 %
(Intercept)   0.000865657 0.006428105
mfund$valmrkt 0.878193149 0.993240873
```

Reject at $\alpha = .05$, so Windsor softens than the market.

- What about other values of α?

```
confint(windsor.reg, level=0.99)
confint(windsor.reg, level=(1-2*pt(-abs(zb1), df=178)))
```

Forecasting & Prediction Intervals

The conditional forecasting problem:

- Given covariate X_f and sample data $\{X_i, Y_i\}_{i=1}^n$, predict the "future" observation Y_f.

The solution is to use our LS fitted value: $\hat{Y}_f = b_0 + b_1 X_f$.

- That's the easy bit.

The hard (and very important!) part of forecasting is assessing uncertainty about our predictions.

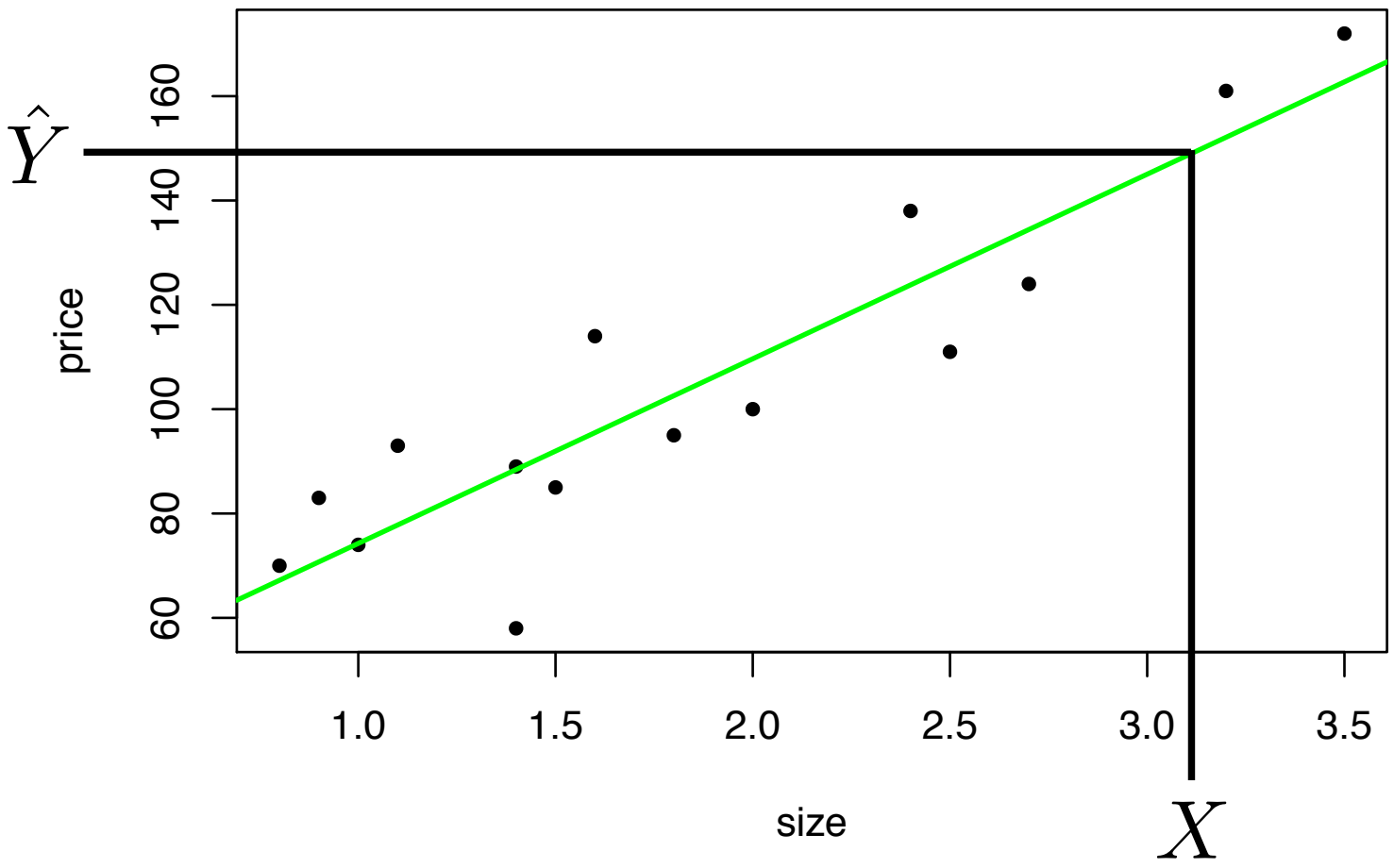
One method is to specify a prediction interval

- a range of Y values that are likely, given an X value.

The least squares line is a prediction rule:

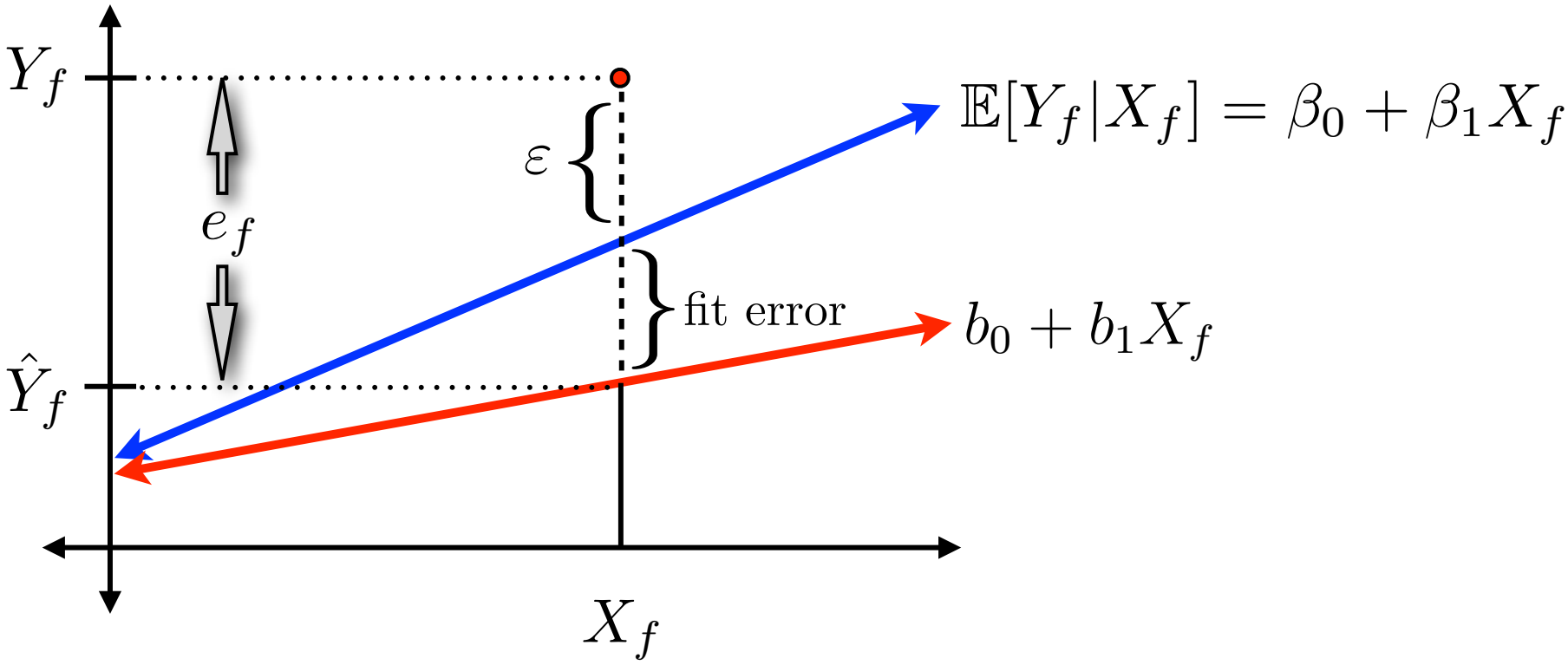
Read $\hat{Y}$ off the line for a new X.

- It's not a perfect prediction: $\hat{Y}$ is what we expect.



If we use $\hat{Y}_f$, our prediction error has two pieces

$$e_f = Y_f - \hat{Y}_f = Y_f - b_0 - b_1 X_f$$



We can decompose e_f into two sources of error:

- Inherent idiosyncratic randomness (due to ε).
- Estimation error in the intercept and slope (i.e., discrepancy between our line and "the truth").

$$\begin{aligned} e_f &= Y_f - \hat{Y}_f = (Y_f - \mathbb{E}[Y_f|X_f]) + \mathbb{E}[Y_f|X_f] - \hat{Y}_f \\ &= \varepsilon_f + (\mathbb{E}[Y_f|X_f] - \hat{Y}_f) \\ &= \varepsilon_f + (\beta_0 - b_0) + (\beta_1 - b_1)X_f. \end{aligned}$$

The variance of our prediction error is thus

$$\mathsf{var}(e_f) = \mathsf{var}(\varepsilon_f) + \mathsf{var}(\mathbb{E}[Y_f|X_f] - \hat{Y}_f) = \sigma^2 + \mathsf{var}(\hat{Y}_f)$$

From the sampling distributions derived earlier, $\text{var}(\hat{Y}_f)$ is

$$\begin{aligned}\text{var}(b_0 + b_1 X_f) &= \text{var}(b_0) + X_f^2 \text{var}(b_1) + 2X_f \text{cov}(b_0, b_1) \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2} \right].\end{aligned}$$

Replacing σ^2 with s^2 gives the standard error for $\hat{Y}_f$.

And hence the variance of our predictive error is

$$\text{var}(e_f) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2} \right].$$

Putting it all together, we have that

$$\hat{Y}_f \sim \mathcal{N}\left(Y_f, \sigma^2\left[1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2}\right]\right)$$

A $(1-\alpha)100\%$ confidence/prediction interval for Y_f is thus

$$b_0 + b_1 X_f \quad \pm \quad z_{\alpha/2} \times \left(s\sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2}}\right).$$

Looking closer at what we'll call

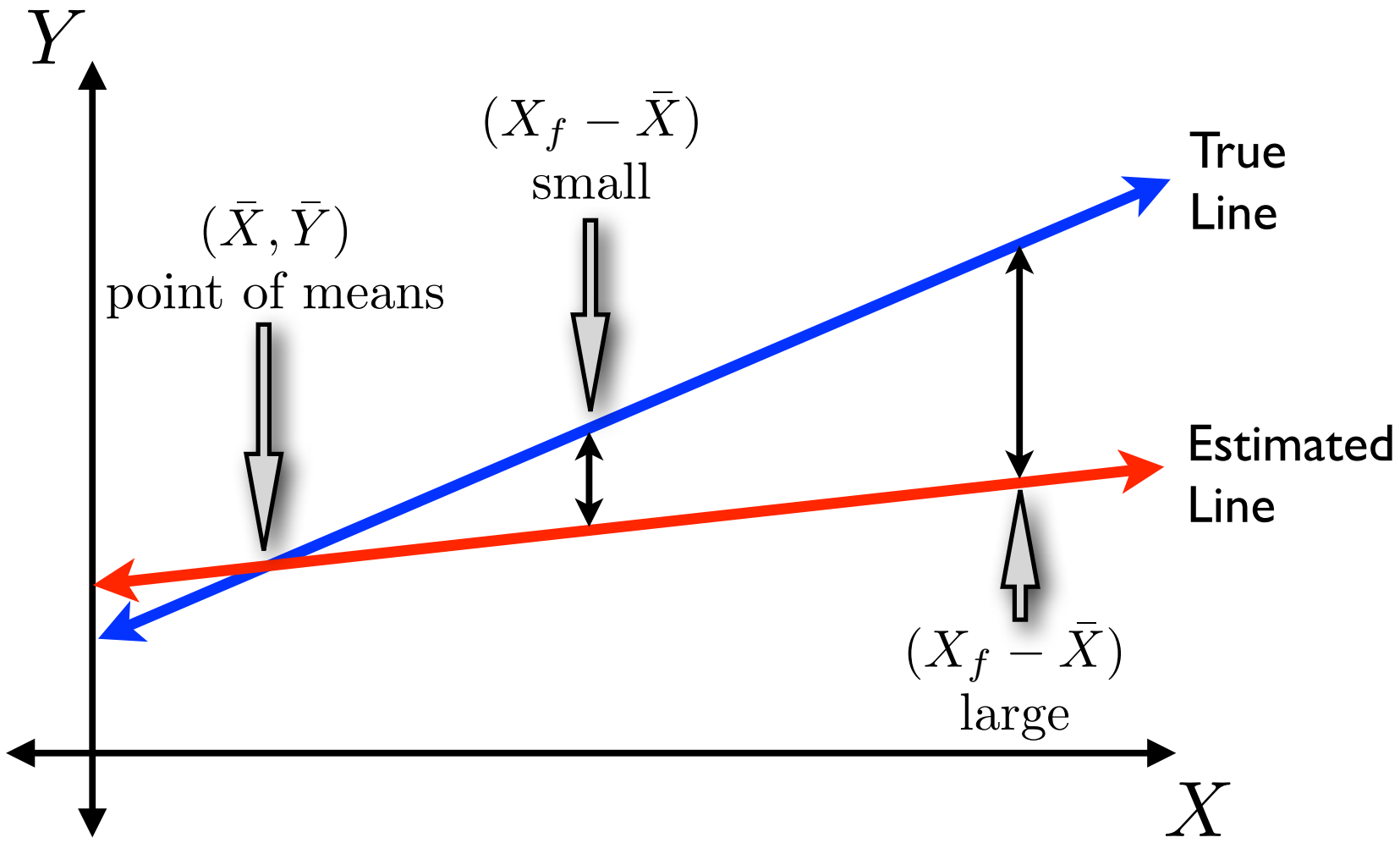
$$s_{\text{pred}} = s\sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2}} = \sqrt{s^2 + s_{\text{fit}}^2}.$$

A large predictive error variance (high uncertainty) comes from

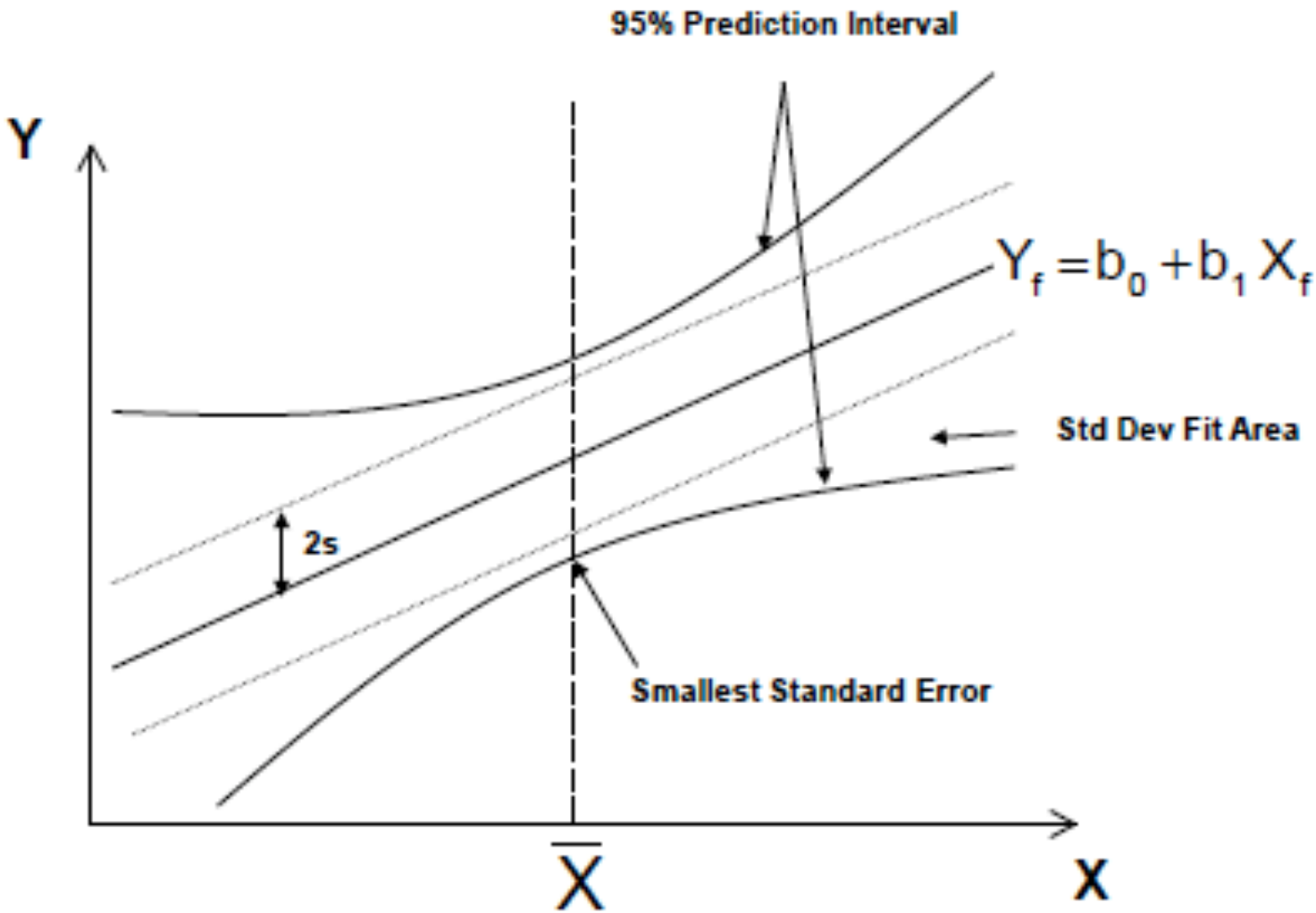
- Large s (i.e., large ε's).
- Small n (not enough data).
- Small s_x (not enough observed spread in covariates).
- Large $(X_f - \bar{X})$.

The first three are familiar... what about the last one?

For X_f far from our $\bar{X}$, the space between lines is magnified ...



$\Rightarrow$ The prediction (conf.) interval needs to widen away from $\bar{X}$



Returning to our housing data for an example ...

```
> Xf <- data.frame(size=c(mean(size), 2.5, max(size)))
> cbind(Xf,predict(reg, newdata=Xf, interval="prediction"))
      size      fit       lwr      upr
1 1.853333 104.4667  72.92080 136.0125
2 2.500000 127.3496  95.18501 159.5142
3 3.500000 162.7356 127.36982 198.1013
```

- `interval="prediction"` gives `lwr` and `upr`, otherwise we just get `fit`
- s_{pred} is not shown in this output

We can get s_{pred} from the `predict` output.

```
> p <- predict(reg, newdata=Xf, se.fit=TRUE)
> s <- p$residual.scale
> sfit <- p$se.fit
> spred <- sqrt(s^2+sfit^2)
> b <- reg$coef
> b[1] + b[2]*Xf[1,]+ c(0,-1, 1)*qnorm(.975)*spred[1]
         [,1]      [,2]     [,3]
[1,] 104.4667  75.84713 133.0862
> b[1] + b[2]*Xf[1,]+ c(0,-1, 1)*qt(.975, df=n-2)*spred[1]
[1,] 104.4667  72.92080 136.0125
```

- Or, we can calculate it by hand [see R code].

Notice that $s_{\text{pred}} = \sqrt{s^2 + s^2_{\text{fit}}}$; you need to square before summing.

Summary

Uncertainty matters!

Captured by the Sampling Distribution.

- Quantifies uncertainty from the data
- . . . only within the model, assumed **before** we see data.
- Which factors matter for signal-to-noise?

Reporting

- Confidence Interval: **completely** captures the information in the data about the parameter.
- Testing/p-value: only a yes/no answer.

(*Don't abuse p-values*)

Glossary and Equations

- LS Estimators: $b_1 = r_{xy}\frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2}$ and $b_0 = \bar{Y} - b_1\bar{X}$.
- $\hat{Y}_i = b_0 + b_1 X_i$ is the ith fitted value.
- $e_i = Y_i - \hat{Y}_i$ is the ith residual.
- $\hat{\sigma}$, s: standard error of regression residuals ($\approx \sigma = \sigma_\varepsilon$).
- s_{b_j}: standard error of regression coefficients.

$$s_{b_1} = \sqrt{\frac{s^2}{(n-1)s_x^2}} \qquad s_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2}}$$

- α is the significance level (prob of type 1 error).
- $z_{\alpha/2}$ is the value such that for $Z \sim \mathcal{N}(0,1)$,

$$\mathbb{P}[Z > -z_{\alpha/2}] = \mathbb{P}[Z < z_{\alpha/2}] = \alpha/2.$$

- z_{b_j} is the standardized coefficient:

$$z_{b_j} = \frac{b_j - \beta_j^0}{s_{b_j}} \overset{H_0}{\sim} \mathcal{N}(0,1).$$

- The $(1-\alpha) * 100\%$ confidence interval for β_j is $b_j \pm z_{\alpha/2} s_{b_j}$

- $\hat{Y}_f = b_0 + X_f b_1$ is a forecast prediction.

$$\text{se}(\hat{Y}_f) = s_{\text{fit}} = s\sqrt{\frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2}}$$

- Forecast residual is $e_f = Y_f - \hat{Y}_f$ and $\text{var}(e_f) = s^2 + s_{\text{fit}}^2$. That is, the predictive standard error is

$$s_{\text{pred}} = s\sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2}}.$$

and $\hat{Y}_f \pm z_{\alpha/2} s_{\text{pred}}$ is the $(1-\alpha)100\%$ prediction interval at X_f.