Simple Linear Regression

Rob McCulloch

1. The Simple Linear Regression Model
2. Estimates and Plug-in Prediction
3. Confidence Intervals, Prediction, and Hypothesis Tests

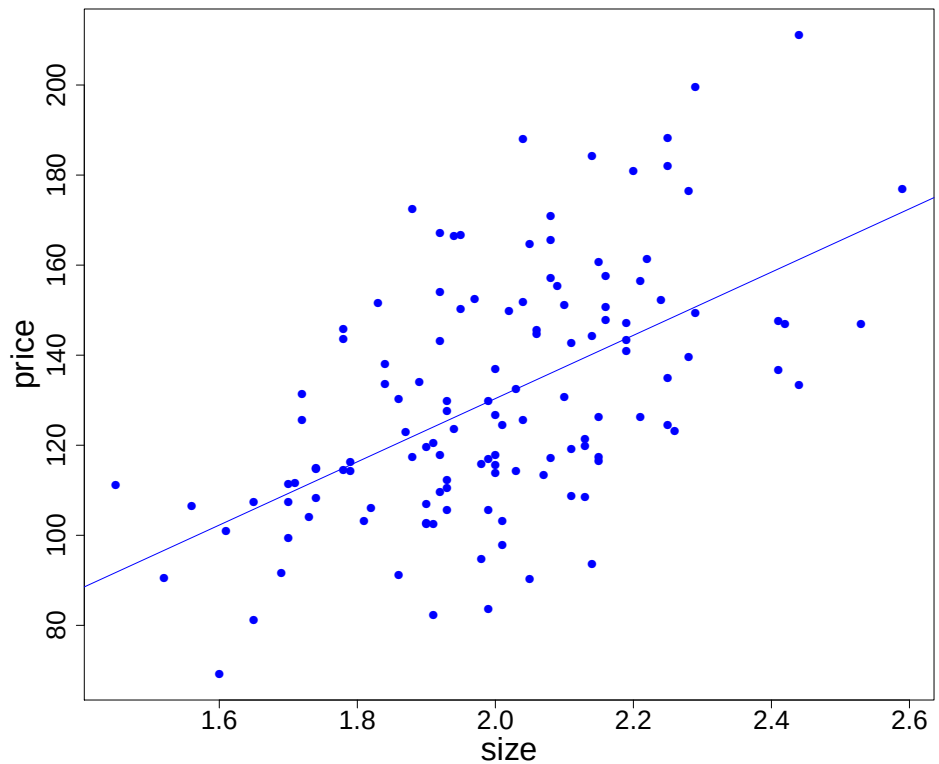
1. The Simple Linear Regression Model

Recall our house price data:
Each observation corresponds to a house.

x = size
(thousands of square feet)

versus

y = price
(thousands of dollars)

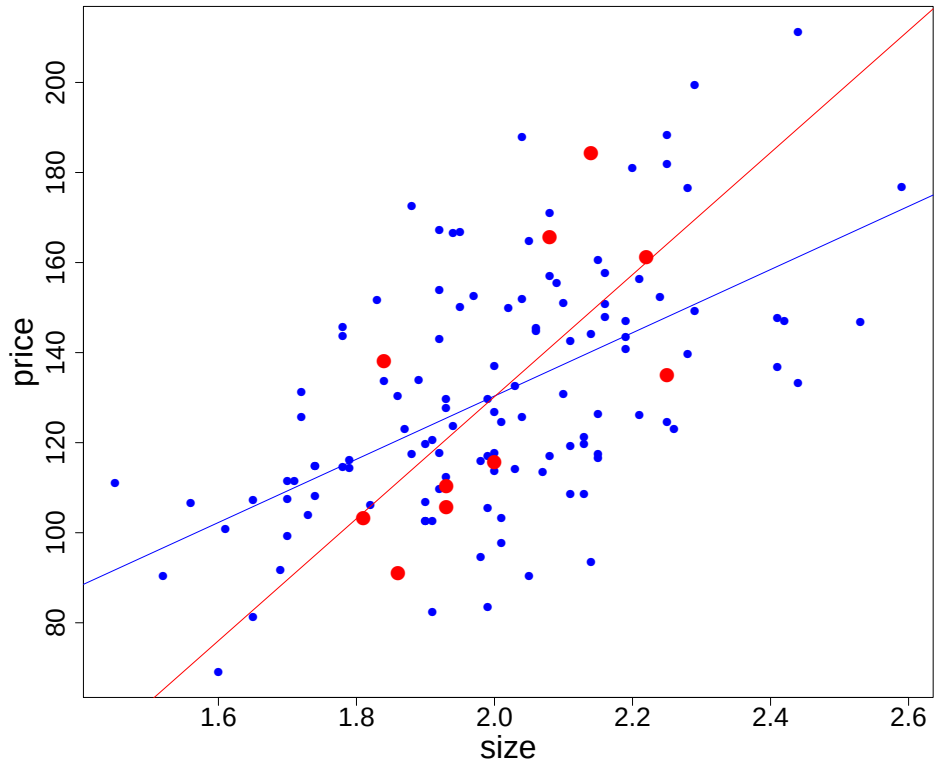


Previously we used the least-squares regression line to summarize the relationship between size and price and predict price given size.

Now we would like to think more deeply about about the relationship.

There are some basic issues that need to be addressed, mostly having to do with assessing the uncertainty in our fit of the line and the corresponding prediction.

As an example, I randomly sampled 10 of the houses and fit a regression line using just those 10 houses.



Which line is better for prediction, the line based on all the houses or the line using just the sample of 10 houses?

We would rather use the line chosen with all the data because it is probably closer to the the "true line" that works for all houses!! All the houses we have seen in the past, and will see in the future.

We feel like we *know more* with $n = 128$ houses than just 10 houses in our sample: *how do we quantify our uncertainy*.

We need a probability model to describe the true relationship between Y and x.

The probability model has to capture the amount of information x tells us about Y.

What kind of model should we use?

In the housing data, the "overall linear relationship" is striking.

Given x, *y is approximately a linear function of x*:

$$Y = \text{ linear function of x} + \text{error}$$

The Simple Linear Regression Model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2), IID.$$

If you knew the values of the parameters $(\beta_0, \beta_1, \sigma)$:

$\beta_0 + \beta_1 x$: the part of Y you learn from x, the "signal" from x.

ϵ: the part of Y you can't tell from x, the "noise".

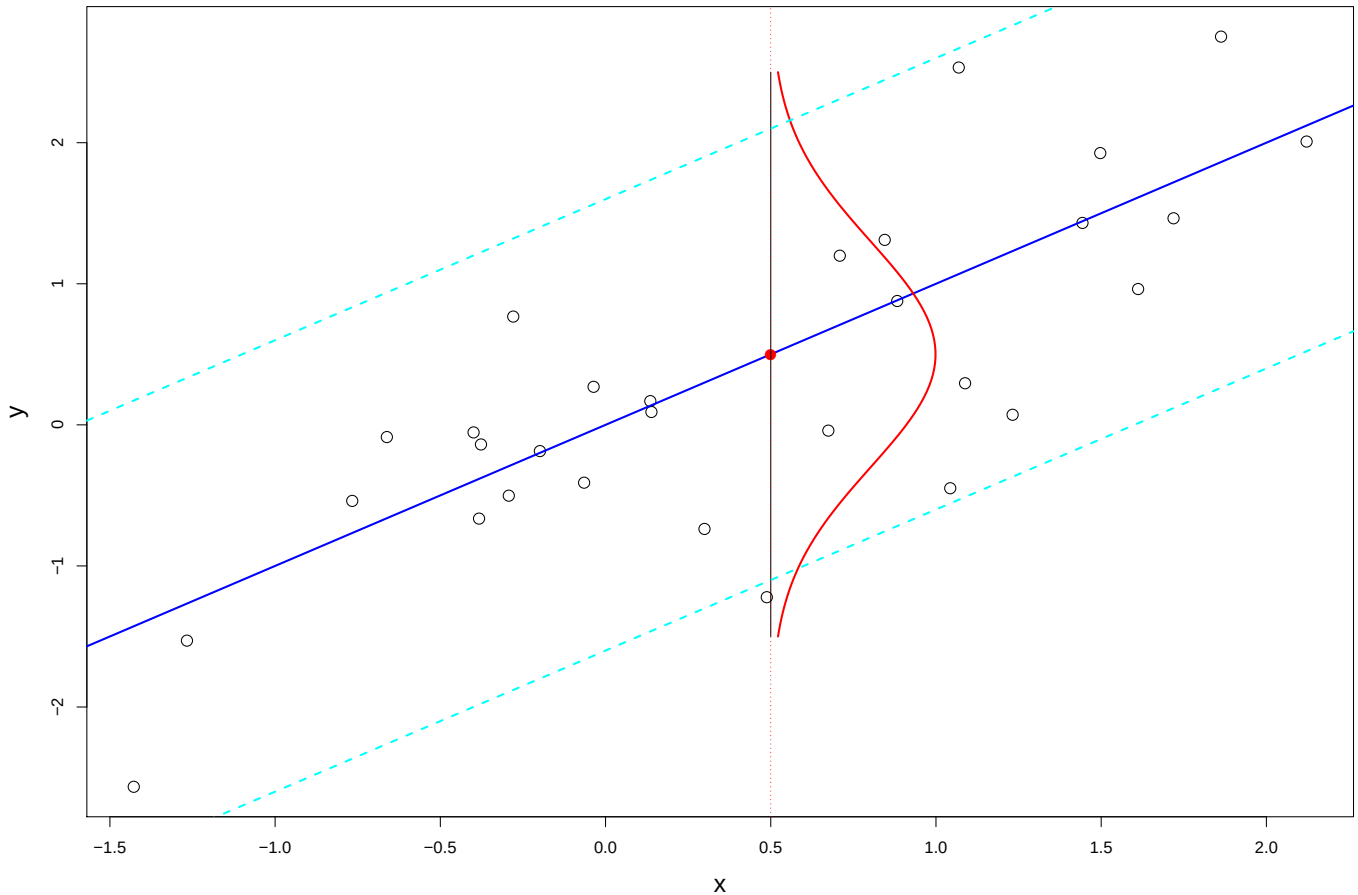
What is the conditional distribution of Y given x ???

$$Y =$$

$$\beta_0 + \beta_1 x$$

$$+$$

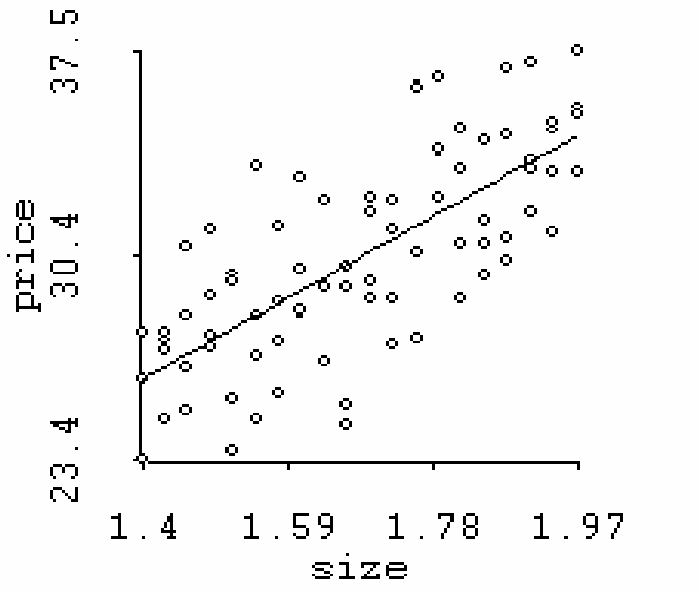
$$\epsilon \sim N(0, \sigma^2).$$



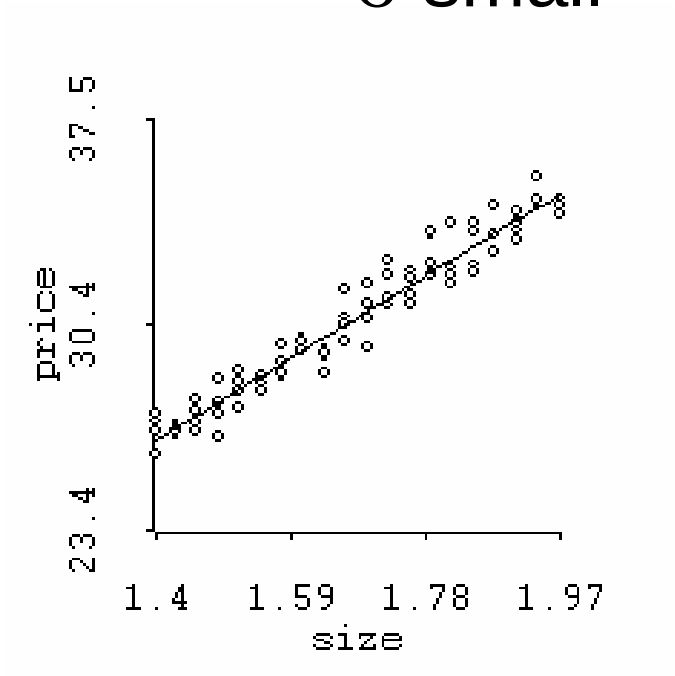
$$Y \mid x \sim N(\beta_0 + \beta_1 x, \sigma^2) \approx \beta_0 + \beta_1 x \pm 2 * \sigma$$

The role of σ

σ large



σ small



We need σ in the model to describe how close the relationship is to linear, how big the errors are.

2. Estimates and Plug-in Prediction

Our simple linear regression model has three parameters: $(\beta_0, \beta_1, \sigma)$.

When we "run" a regression using software we get estimates using the information in the data.

Previously we had:

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \quad a = \bar{y} - b\,\bar{x}.$$

for the least-squares slope and intercept.

Now we want to think b as an *estimate* of β_1 !!!!

Now we want to think a as an *estimate* of β_0 !!!!

To emphasize this we rename a and b:

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

$$\hat{\beta}_0 = \bar{y} - b\,\bar{x}.$$

How do we estimate σ^2?

Since σ^2 is the variance of the of the errors we might think about using the sample variance of the errors:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum(\epsilon_i - \bar{\epsilon})^2$$

However, we don't directly observe the ϵ_i.
But we can estimate each error with:

$$\begin{aligned}\epsilon_i &= y_i - \beta_0 - \beta_1 x_i \\ &\approx y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ &= e_i\end{aligned}$$

Recall that e_i is the *residual*.

So,

$$\begin{aligned}\hat{\sigma}^2 &\approx \frac{1}{n-1}\sum(\epsilon_i - \bar{\epsilon})^2 \\ &\approx \frac{1}{n-1}\sum(e_i - \bar{e})^2 \\ &\approx \frac{1}{n-2}\sum e_i^2\end{aligned}$$

Where we have used $\bar{e} = 0$ (make sense??) and $n-2$ instead of $n-1$ to adjust for the estimation of both β_0 and β_1.

The estimator for σ^2 is:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum e_i^2$$

R Regression output for the housing data:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -10.091     18.966  -0.532    0.596
sizethou      70.226      9.426   7.450  1.3e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 22.48 on 126 degrees of freedom
Multiple R-squared:  0.3058,    Adjusted R-squared:  0.3003
F-statistic:  55.5 on 1 and 126 DF,  p-value: 1.302e-11
```

Our estimate of β_0 is $\hat{\beta}_0 =$ -10.091.
Our estimate of β_1 is $\hat{\beta}_1 = 70.226$.
Our estimate of σ is $\hat{\sigma} = 22.48$.

The estimated model is:

$$\textit{price} = -10.1 + 70.2\,\textit{size} + \epsilon, \quad \epsilon \sim N(0, 22.48^2).$$

Interpret:

$\hat{\beta}_1 = 70.226$.

If one house is $\Delta x = .5$ thousand square feet bigger than another, we expect the price to be bigger by $\Delta y = 70.226 * .5 = 35.113$.

In general, your interpretation of the intercept is the Y you expect when $x = 0$.

In this application, we do not want to consider a house of size 0, so it does not make a lot of sense on its own.

Predict:

Our estimated model is:

$$price = -10.1 + 70.2\, size + \epsilon, \quad \epsilon \sim N(0, 22.48^2).$$

Suppose $x = size = 2.2$.
What is our prediction for $y = price$?

$$\begin{aligned} price &= -10.1 + 70.2\,(2.2) + \epsilon \\ &= 144.34 + \epsilon \\ &\sim N(144.34, 22.48^2) \\ &\approx 144.34 \pm 2(22.48) \\ &\approx 144.34 \pm 45 \end{aligned}$$

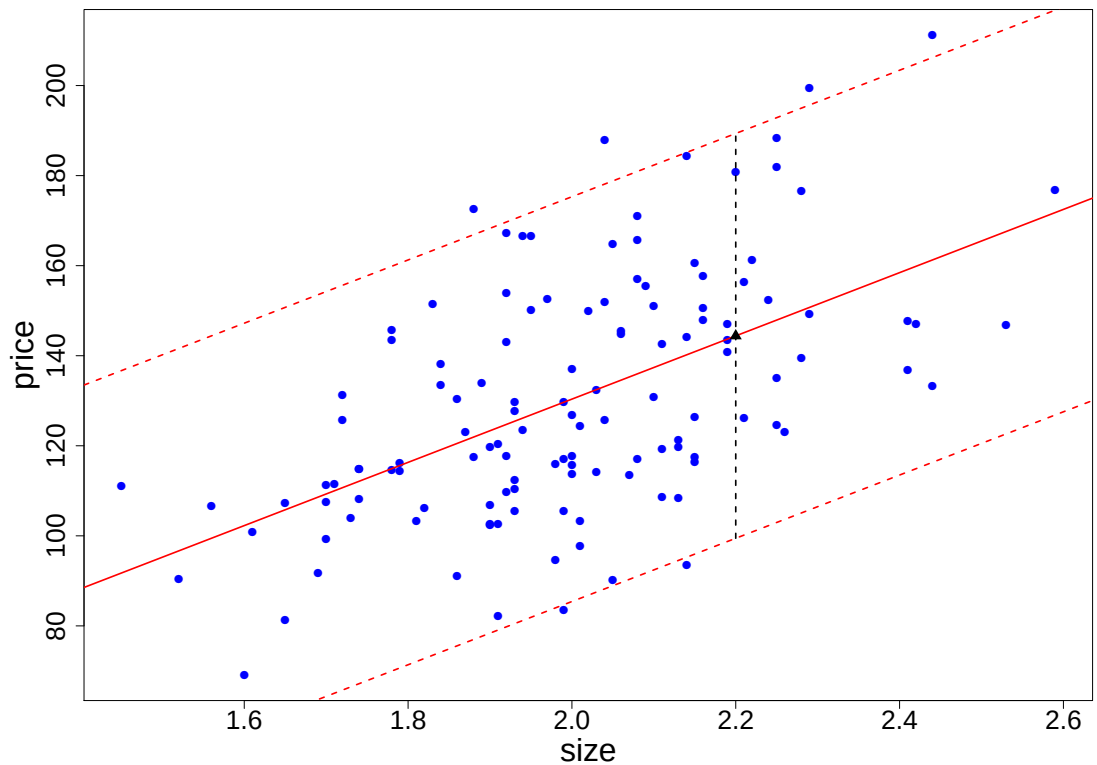
We call this it plug-in prediction since we just "plug in" our parameters estimates without worrying about possible estimation error.

The estimated plug-in conditional distribution is

$$Y \mid x \sim N(\hat{\beta}_0 + \hat{\beta}_1 x, \hat{\sigma}^2)$$

with plug-in 95% prediction interval

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm 2 * \hat{\sigma}$$



Excel regression output for the housing data.

StErr of Est is $\hat{\sigma}$ is 22.48.

Coefficient estimates are in the usual place.

Intercept estimate is -10.09 and slope estimate is 70.2263.

Results of multiple regression for pricethou

Summary measures

Multiple R	0.5530
R-Square	0.3058
Adj R-Square	0.3003
StErr of Est	22.4755

ANOVA Table

Source	df	SS	MS	F	p-value
Explained	1	28036.3627	28036.3627	55.5011	0.0000
Unexplained	126	63648.8516	505.1496		

Regression coefficients

	Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
Constant	-10.0911	18.9661	-0.5321	0.5956	-47.6245	27.4422
sizethou	70.2263	9.4265	7.4499	0.0000	51.5716	88.8810

3. Confidence Intervals, Prediction, and Hypothesis Tests

With more data we expect we have a better chance that our estimates will be close to the true (or "population") values.

The "true line" is the one that "generalizes" to the size and price of future houses, not just the ones in our current data.

How big is our error in estimating β_0 and β_1?

We have standard errors and confidence intervals for our estimates of the true slope and intercept.

95% Confidence interval for β_0:

$$\hat{\beta}_0 \pm 2\, se(\hat{\beta}_0)$$

95% Confidence interval for β_1:

$$\hat{\beta}_1 \pm 2\, se(\hat{\beta}_1)$$

Let's skip the formulas for the standard errors.

```
Call:
lm(formula = price ~ size, data = ddf)

Residuals:
   Min     1Q Median     3Q    Max
-46.59 -16.64  -1.61  15.12  54.83

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -10.091     18.966  -0.532    0.596
size          70.226      9.426   7.450  1.3e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 22.48 on 126 degrees of freedom
Multiple R-squared:  0.3058,Adjusted R-squared:  0.3003
F-statistic:  55.5 on 1 and 126 DF,  p-value: 1.302e-11
```

$se(\hat{\beta}_0) = 18.966.$
$se(\hat{\beta}_1) = 9.426.$

Confidence interval for the slope (β_1): $70.226 \pm 2(9.426) = 70.226 \pm 18.9$

Here is the regression output using just the sample of 10 houses:

```
Call:
lm(formula = price ~ size, data = ddfs)

Residuals:
    Min      1Q  Median      3Q     Max
-29.093 -14.978  -5.801  18.725  35.112

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -140.79     100.12  -1.406   0.1973
size          135.50      49.77   2.723   0.0261 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 23.75 on 8 degrees of freedom
Multiple R-squared:  0.4809,Adjusted R-squared:  0.4161
F-statistic: 7.412 on 1 and 8 DF,  p-value: 0.02615
```

What happens to the standard errors for the coefficient estimates when we go from $n = 128$ to $n = 10$?

The Predictive Interval:

If the confidence intervals for the slope and intercept are big the plug-in predictive interval can be misleading!!

The *predictive interval* accounts for both our uncertainty about the parameters (β_0,β_1,σ) and the part of price
not explained by size (ϵ).

The predictive interval is bigger than the plug-in predictive interval!

Some software (e.g. `R`) will give you the predictive interval.

Suppose we are predicting Y given x.

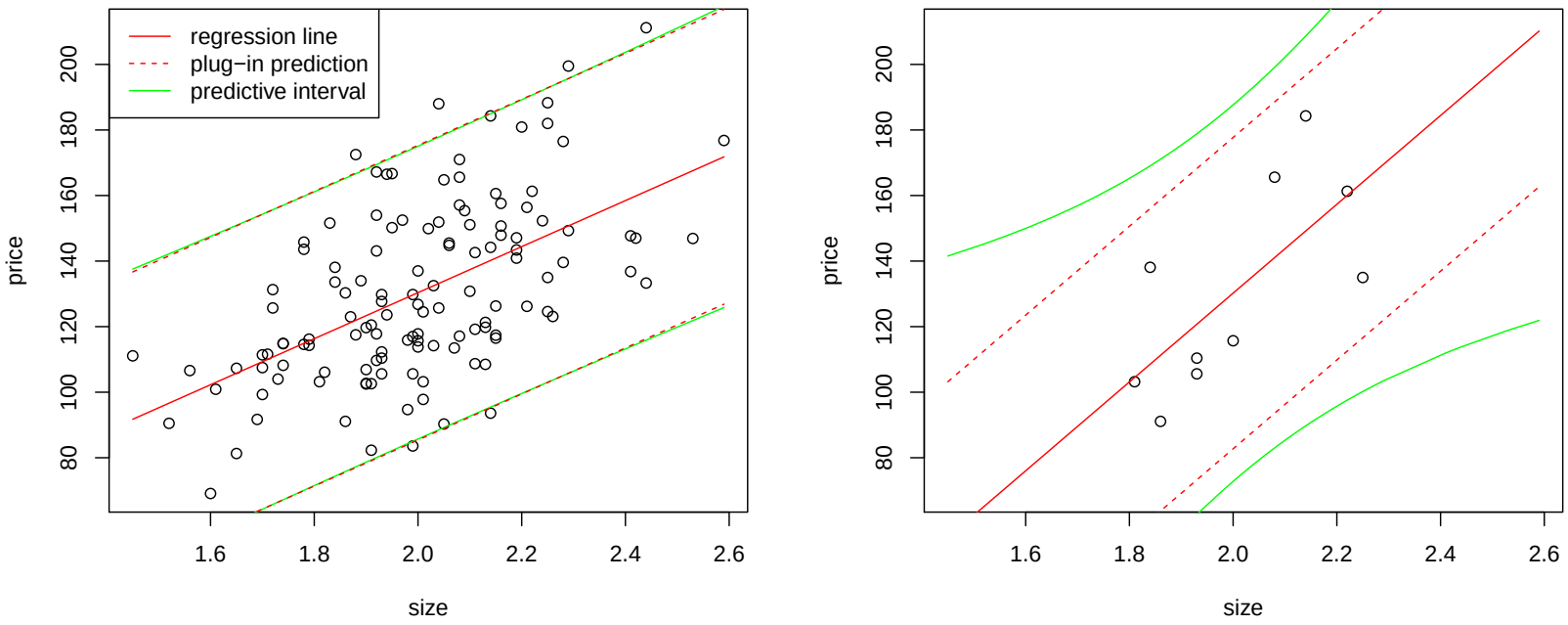
Our error in prediction is

$$\begin{aligned} E &= Y - \hat{Y} \\ &= (\beta_0 + \beta_1 x + \epsilon) - (\hat{\beta}_0 + \hat{\beta}_1 x) \\ &= (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) x + \epsilon \end{aligned}$$

The plug-in interval ignores the error due to estimation of the coefficients and just says $\epsilon \approx \pm 2\hat{\sigma}$.

The predictive interval accounts for all sources of uncertainty (assuming the model is correct).

Predictive and plug-in predictive intervals for the full data set (left) and the subset of size 10 (right).



When there is a lot uncertainty about the coefficents, the predictive interval can be much bigger than the plug-in interval.

In practice, I use the plug-in interval *a lot*.

It gives me a simple quick way to see how well my regression is working (bearing in mind that the full predictive interval may be bigger).

Hypothesis Tests in Simple Linear Regression

For i equal 0 or 1, to test the null hypothesis:

$$\text{Ho: } \beta_i = \beta_i^o \quad \textit{vs.} \quad \text{Ha: } \beta_i \neq \beta_i^o$$

We reject at level .05 if

$$|t| > 2, \quad \text{where} \quad t = \frac{\hat{\beta}_i - \beta_i^o}{se(\hat{\beta}_i)}.$$

Otherwise, we fail to reject.

Note:

(1)

The t thing is called the *t statistic*, it is our test statistic.

(2)

If the null hypothesis is true, the t should look like a draw from the standard normal (the t should look like a z).

(3)

(2) is actually an approximation that works for larger n (e.g. > 20).

For smaller n, the t actually is a draw from a *t distribution* but we are skipping this.

Note:

It is very common to let $\beta_1^o = 0$.

We test,

$$Ho : \beta_1 = 0.$$

Why is this important?

$$Y|X = x \sim N(\beta_0 + \beta_1 x, \sigma^2).$$

If $\beta_1 = 0$ then the conditional of Y does not depend on x
so they are independent!!

Housing regression in R:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -10.091     18.966  -0.532    0.596
sizethou      70.226      9.426   7.450  1.3e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 22.48 on 126 degrees of freedom
Multiple R-squared:  0.3058,Adjusted R-squared:  0.3003
F-statistic:  55.5 on 1 and 126 DF,  p-value: 1.302e-11
```

To test $\beta_1 = 0$ we have: $t = \frac{70.226-0}{9.426} = 7.45$.

We reject the null hypothesis $\beta_1 = 0$.

Information related to testing $\beta_1 = 0$ and $\beta_0 = 0$ are commonly included in regression output.

The t-value for testing $\beta_0 = 0$ is -.532. Fail to reject.

p-values:

Most regression packages automatically print out the p-values for the hypotheses that $\beta_0 = 0$ or that $\beta_1 = 0$.

In the R and excel output we have:

Is the intercept 0?: p-value = .59, fail to reject.
Is the slope 0?: p-value = .0000, reject.

Note: 2* (standard normal cdf at -.532) = .594726.

Example (the market model)

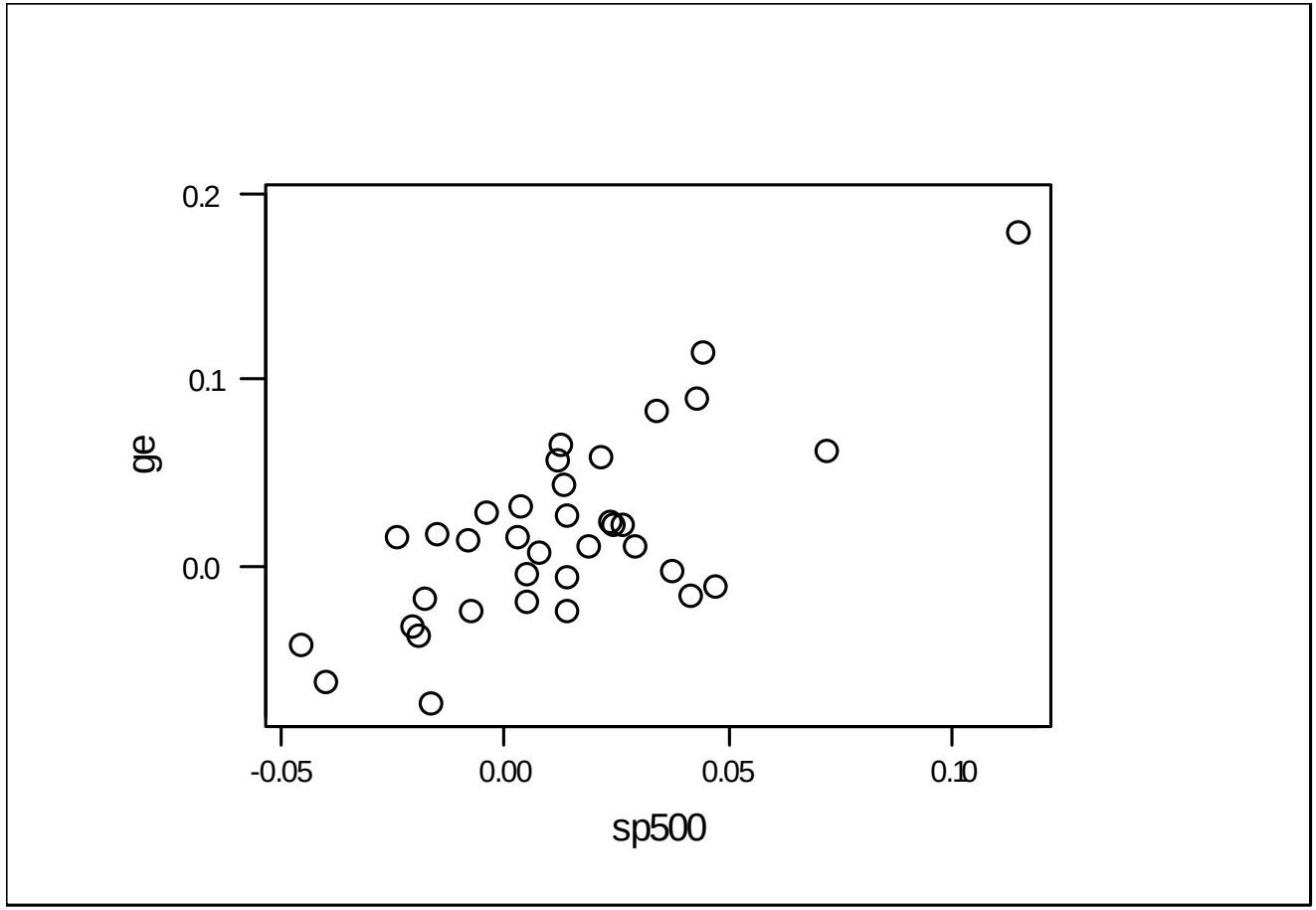
In finance, a popular model is to regress stock returns against returns on some market index, such as the S&P 500.

The slope of the regression line, referred to as “beta”, is a measure of how sensitive a stock is to movements in the market.

Usually, a beta less than 1 means the stock is less risky than the market, equal to 1 same risk as the market and greater than 1, riskier than the market.

We will examine the market model for the stock General Electric, using the S&P 500 as a proxy for the market.

Three years of monthly data give 36 observations.



minitab ouput:

```
The regression equation is
ge = 0.00301 + 1.20 sp500

Predictor       Coef      Stdev    t-ratio       p
Constant    0.003013   0.006229       0.48   0.632
sp500         1.1995     0.1895       6.33   0.000

s = 0.03454     R-sq = 54.1%     R-sq(adj) = 52.7%
```

We can test the hypothesis that the slope is zero:
that is, **are GE returns related to the market?**

The test statistic is

$$t = \frac{1.2 - 0}{.19} = 6.3$$

Clear reject.

This is the same value as in the regression output and the associated p-value is basically 0.

Now let's test the hypothesis that GE has the "same risk" as the market, that is, that the slope $= 1$.

We test: $H_0 : \beta_1 = 1$

$$t = \frac{1.2 - 1}{.19} = 1.05.$$

So, we fail to reject.

What would the p-value be?

Would we want to "accept" the hypothesis that $\beta_1 = 1$?

The confidence interval is

$$1.2 \pm 2(.2) = (.8, 1.6)$$

Is this a big interval?

Why do we "fail to reject"?