Business Statistics 41000

Simple Linear Regression

Mladen Kolar

CAPM Example

Another example of conditional distributions:

Individual returns given market return.

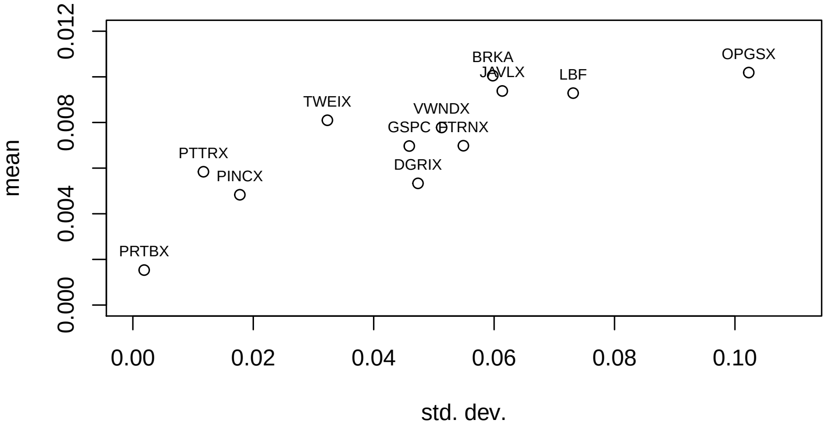
The Capital Asset Pricing Model (CAPM) for asset A relates

return $R_{At} = \dfrac{V_{At} - V_{At-1}}{V_{At-1}}$ to the "market" return, R_{Mt}.

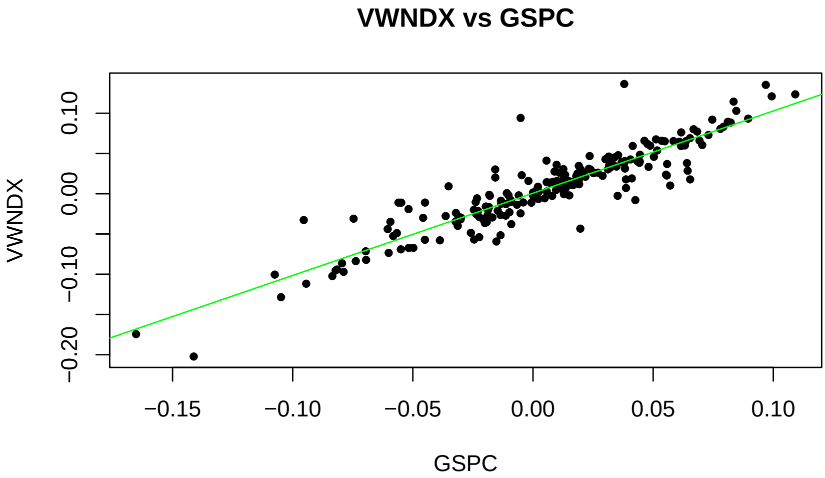
In particular, the relationship is given by the regression model $R_{At} = \alpha + \beta R_{Mt} + \varepsilon$ with observations at times $t = 1 \ldots T$ (and where $[\alpha, \beta] \equiv [\beta_0, \beta_1]$).

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

```r
mfund = read.csv("mutualFundReturn.csv", row.names=1)
mean.mfr = apply(mfund, 2, mean)
sd.mfr = apply(mfund, 2, sd)
plot(sd.mfr, mean.mfr, type="p", xlim=c(0, 0.11), ylim=c(0, 0.012),
     main="", xlab = "std. dev.", ylab = "mean")
text(sd.mfr, mean.mfr, labels=names(mfund), cex= 0.7, pos=3)
```



```r
plot(mfund$GSPC, mfund$VWNDX, pch=20, xlab="GSPC", ylab="VWNDX", main="VWNDX vs GSPC")
VWNDX.reg = lm(mfund$VWNDX ~ mfund$GSPC)
abline(VWNDX.reg, col="green")
```




```
## [1] "b_0 = 7e-04"

## [1] "b_1 = 1.0218"
```

Modeling goals

Prediction

$$\hat{Y} = b_0 + b_1 X$$
$$Y = b_0 + b_1 X + e$$

Model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

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

Here it is (again!):

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

What's important?

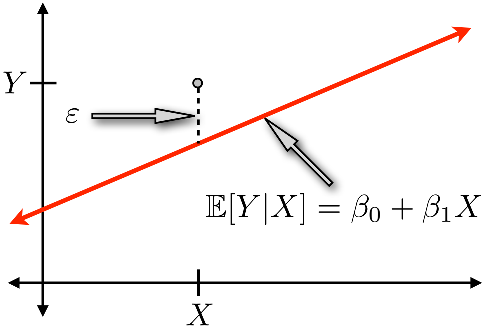
- It is a model, so we are *assuming* this relationship holds for some fixed but unknown values of β_0, β_1.
- It is linear.
- The error ε is independent, additively separable, idiosyncratic noise.
  1. $E[\varepsilon] = 0 \Leftrightarrow E[Y \mid X] = \beta_0 + \beta_1 X$
  2. Fixed but unknown variance σ^2; constant over X
  3. Most things are approx. Normal (Central Limit Theorem)
- It just works! *This is a very robust model for the world.*

Before looking at any data, the model specifies

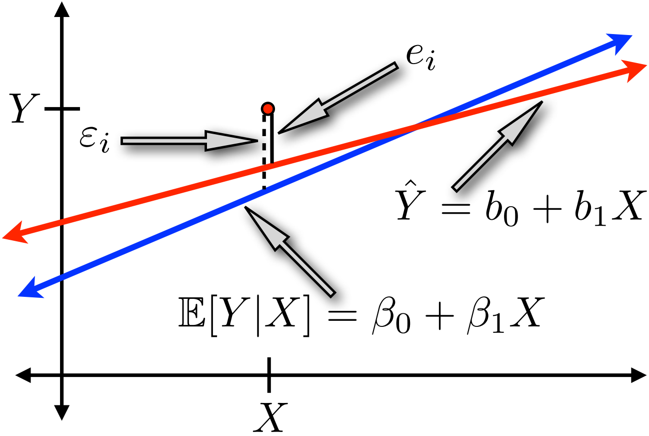
- how Y varies with X on average: $E[Y|X] = \beta_0 + \beta_1 X$;

  i.e. what's the trend?

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



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

$E[Y \mid X]$ 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 but
  - is independent of the `size`,
  - and is Normal.

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

Context from the CAPM example

$E[Y|X]$ 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 but
  - is independent of the `market return`,
  - and is Normal.

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

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

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

- 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?

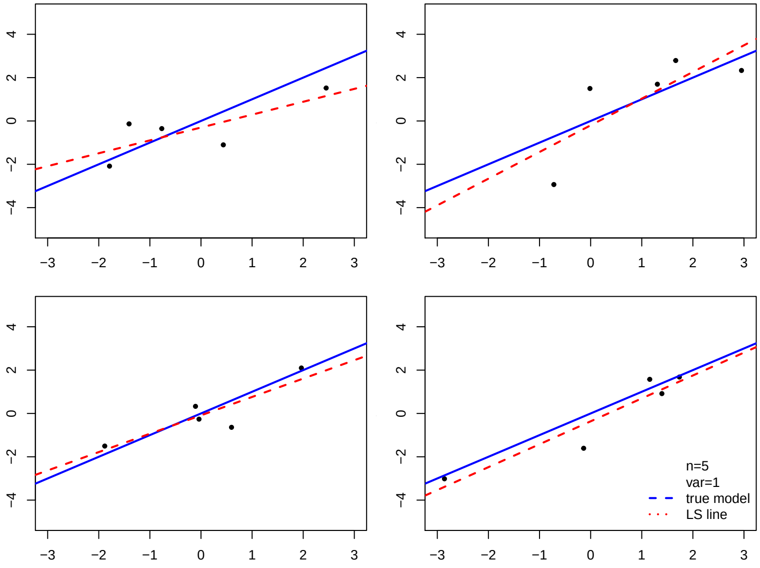
We pretend!

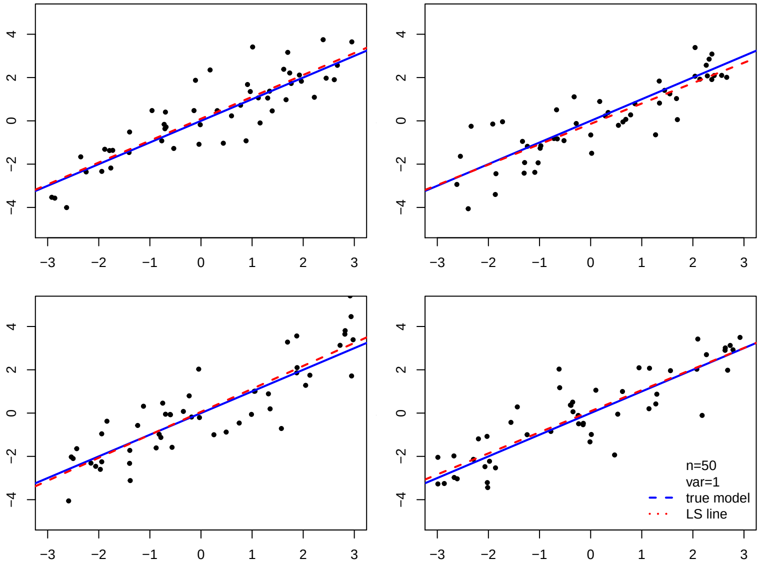
1. Randomly draw **new** data
2. Compute the **estimates** b_0 and b_1
3. Repeat

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=1
true model
LS line


n=50
var=1
true model
LS line

Sampling distribution of LS estimates

What did we just do?

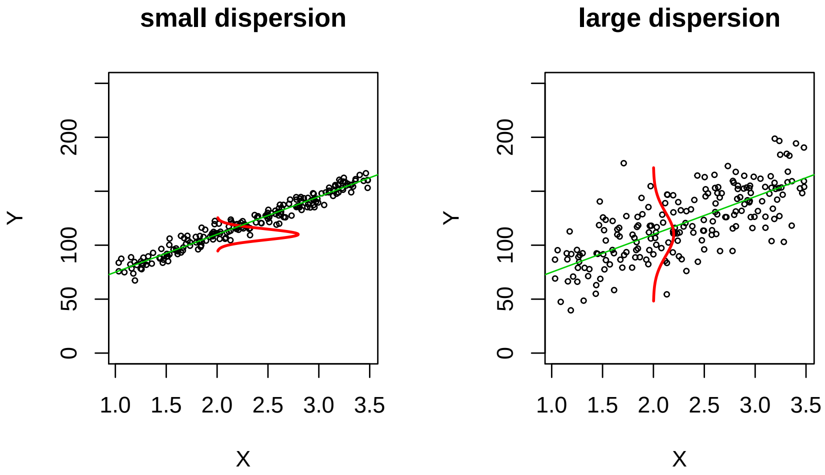
- We "imagined" through simulation the sampling distribution of a LS line.

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!

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

- think signal-to-noise



Sampling distribution of LS estimates

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.

Sampling distribution of b_1 and b_0

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

- b_1 is unbiased: $E[b_1] = \beta_1$.
- The sampling sd σ_{b_1} determines precision of b_1.
  It depends on three factors:
  1. sample size (n)
  2. error variance ($\sigma^2 = \sigma^2_\varepsilon$), and
  3. X-spread (s_x).

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

Confidence intervals

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

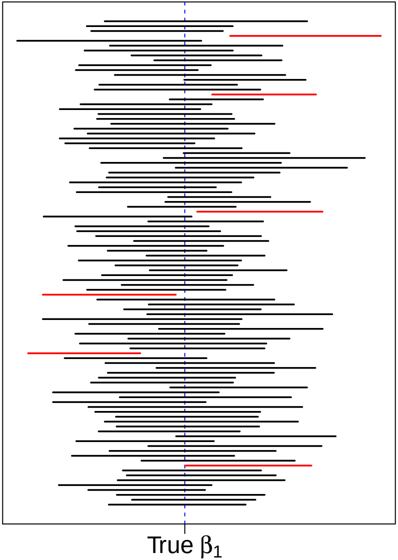
$$
\begin{aligned}
1 - \alpha &= \mathbb{P}\left[z_{\alpha/2} < \frac{b_j - \beta_j}{\sigma_{b_j}} < z_{1-\alpha/2} \right] \\
&= \mathbb{P}\Big[\beta_j \in (b_j \pm z_{\alpha/2}\sigma_{b_j}) \Big]
\end{aligned}
$$

(just replace $j = 0$ or $j = 1$ above for β_0 and β_1)

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

Confidence intervals: $\mathbb{P}\left[\beta_1 \in (b_1 \pm 2\sigma_{b_1})\right] = 95\%$



Estimation of error variance

The last parameter that we have not talked about in the model

$$Y \mid X \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

is the error variance $\sigma^2 = \sigma_\varepsilon^2$.

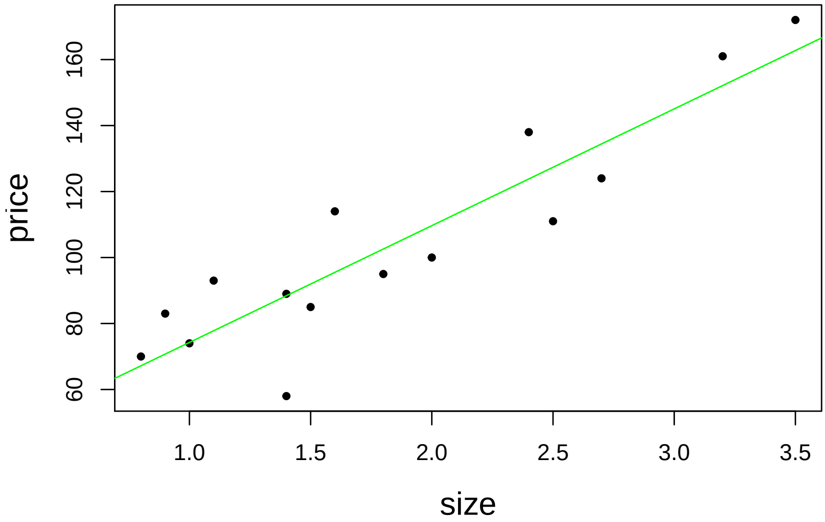
We estimate σ^2 with sample variance of the residuals

$$s^2 = \frac{1}{n-p} \sum_{i=1}^{n} e_i^2 = \frac{SSE}{n-p}$$

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

It is often convenient to report s, which are in the same units as Y.

Example: revisit the house price/size data



Example: revisit the house price/size data

```r
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
```

```r
confint(house.reg,level=0.95)
```

```
##                2.5 %   97.5 %
## (Intercept) 19.23850 58.53087
## size        25.67709 45.09484
```

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 alternative claim
- Gives only a "yes" or "no" answer!

For example, is there any evidence in the data to support the existence of a relationship between X and Y? Then $\beta_1 = 0$ is the null.

We use b_j for our test about β_j.

- Reject H_0 when 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 test statistic (z-score) is going to tell us that:

$$z_{\beta_j} = \frac{b_j - \beta_j^0}{s_{b_j}}.$$

- If H_0 is true, then $z_{\beta_j} \sim N(0,1)$. $\left(\mathbb{P}[|z_{b_j}|>2] \ < \ 0.05 = \alpha\right)$
- So "large" $|z_{\beta_j}|$ makes our guess β_j^0 look silly $\Rightarrow$ reject

But: $\qquad |z_{\beta_j}| > 2 \qquad \Leftrightarrow \qquad \beta_j^0 \notin (b_j \pm 2s_{b_j})$

$\Rightarrow$ Reject at the α level any β_j^0 outside the $1-\alpha$ CI!

Example: revisit the CAPM regression for the Windsor fund.

Does Windsor have a non-zero intercept?
(that is, 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"

```r
summary(VWNDX.reg)
```

```
##
## Call:
## lm(formula = mfund$VWNDX ~ mfund$GSPC)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.064153 -0.008549 -0.000471  0.008385  0.098748
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0006605  0.0014610   0.452    0.652
## mfund$GSPC  1.0218342  0.0315369  32.401   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02078 on 205 degrees of freedom
## Multiple R-squared:  0.8366, Adjusted R-squared:  0.8358
## F-statistic:  1050 on 1 and 205 DF,  p-value: < 2.2e-16
```

It turns out that we fail reject the null at $\alpha = .05$

- Thus we do not have evidence that Windsor does have an "alpha" over the market.

Looking now at the slope, 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.

We are asking 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 time, R's output (z-score and p values) are not what we want (why?).

```
summary(VWNDX.reg)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.00066    0.00146    0.45     0.65
## mfund$GSPC   1.02183    0.03154   32.40   <2e-16 ***
```

But we can get the appropriate values easily:

```
zb1 = (1.02183 - 1) / 0.03154
2*pnorm(-abs(zb1))
```

```
## [1] 0.4888513
```

```
confint(VWNDX.reg, level=0.95)
```

```
##                    2.5 %      97.5 %
## (Intercept) -0.002219965 0.003540918
## mfund$GSPC   0.959655856 1.084012450
```

We thus fail to reject the null at $\alpha = .05$.

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.

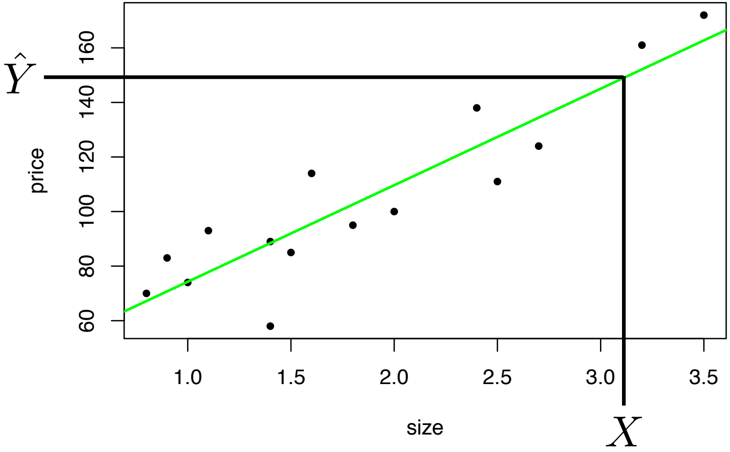
One method is to specify a prediction interval

- 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.



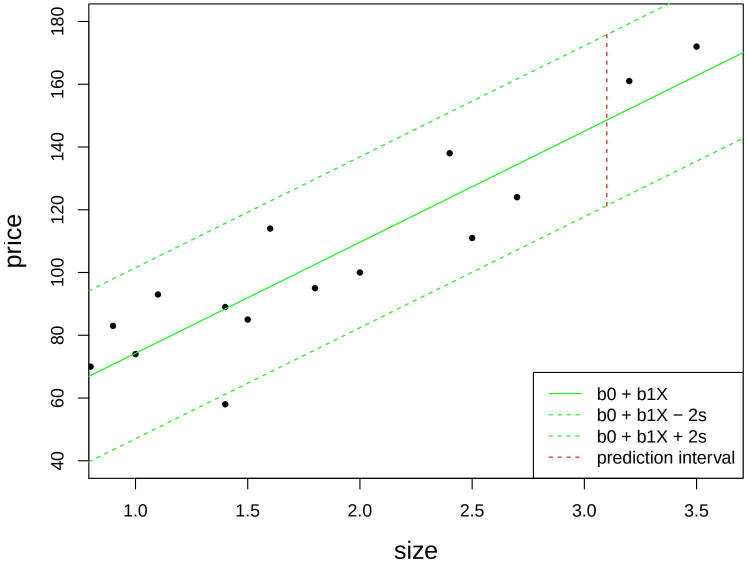
We will use our model to create a 95% prediction interval:

$$Y \mid X \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

Using the model, we form a 95% prediction interval as $\beta_0 + \beta_1 X \pm 1.96\sigma$.

Since we do not know the true parameters, we plug-in the estimated values

$$(b_0 + b_1 X - 1.96s, b_0 + b_1 X + 1.96s)$$


price
size
40
60
80
100
120
140
160
180
1.0
1.5
2.0
2.5
3.0
3.5
b0 + b1X
b0 + b1X − 2s
b0 + b1X + 2s
prediction interval

```
Xf <- data.frame(size=c(1, 2.5, 3))
cbind(Xf,predict(house.reg, newdata=Xf, interval="prediction"))
```

```
##   size       fit       lwr      upr
## 1  1.0  74.27065  41.65499 106.8863
## 2  2.5 127.34959  95.18501 159.5142
## 3  3.0 145.04257 111.58989 178.4952
```

- `interval="prediction"` gives `lwr` and `upr`, otherwise we just get `fit`

A (bad) goodness of fit measure: R^2

How well does the least squares fit explain variation in Y?

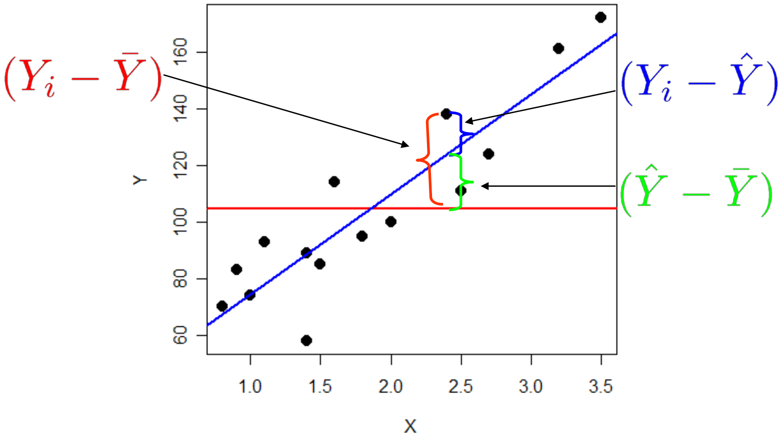
$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\substack{\text{Total} \\ \text{sum of squares} \\ \text{(SST)}}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\substack{\text{Regression} \\ \text{sum of squares} \\ \text{(SSR)}}} + \underbrace{\sum_{i=1}^{n} e_i^2}_{\substack{\text{Error} \\ \text{sum of squares} \\ \text{(SSE)}}}$$

SSR: Variation in Y explained by the regression.
SSE: Variation in Y that is left unexplained.

$$\text{SSR} = \text{SST} \Rightarrow \text{perfect fit.}$$

Be careful of similar acronyms; for example, SSR for "residual" SS.

How does that breakdown look on a scatterplot?



A (bad) goodness of fit measure: R^2

The coefficient of determination, denoted by R^2, measures goodness-of-fit:

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}}$$

- SLR or MLR: same formula.
- $R^2 = \mathrm{corr}^2(\hat{Y}, Y) = r^2_{\hat{y}y}$ $(= r^2_{xy}$ in SLR)
- $0 < R^2 < 1$.
- The closer R^2 is to 1, the better the fit.
  - No surprise: the higher the sample correlation between X and Y, the better you are doing in your regression.
  - So what? What's a "good" R^2?

```r
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
```

```r
summary(VWNDX.reg)
```

```
##
## Call:
## lm(formula = mfund$VWNDX ~ mfund$GSPC)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.064153 -0.008549 -0.000471  0.008385  0.098748
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0006605  0.0014610   0.452    0.652
## mfund$GSPC  1.0218342  0.0315369  32.401   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02078 on 205 degrees of freedom
## Multiple R-squared:  0.8366, Adjusted R-squared:  0.8358
## F-statistic:  1050 on 1 and 205 DF,  p-value: < 2.2e-16
```