Section 2.2: Simple Linear Regression: Predictions and Inference

Jared S. Murray
The University of Texas at Austin
McCombs School of Business
Suggested reading: OpenIntro Statistics, Chapter 7.4

Simple Linear Regression: Predictions and Uncertainty

Two things that we might want to know:

- What value of Y can we *expect* for a given X?
- How sure are we about this prediction (or forecast)? That is, how different could Y be from what we expect?

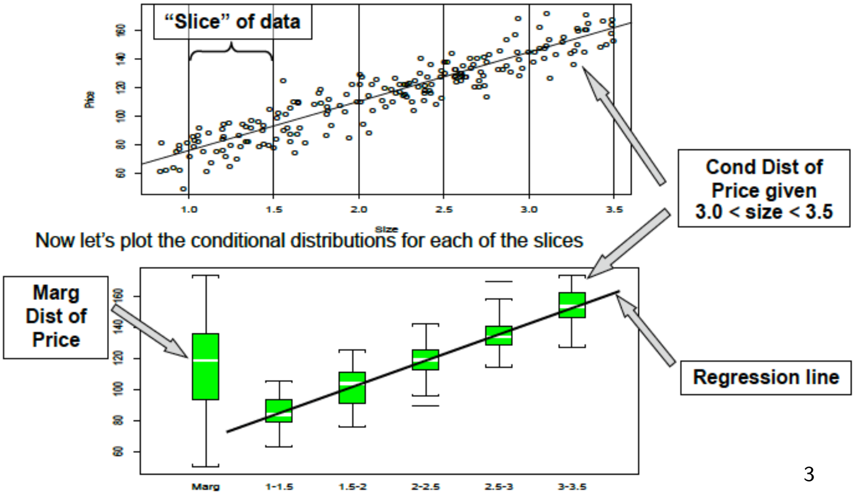
Our goal is to measure the accuracy of our forecasts or how much uncertainty there is in the forecast. One method is to specify a range of Y values that are likely, given an X value.

Prediction Interval: probable range of Y values for a given X

We need the conditional distribution of Y given X.

Conditional Distributions vs the Marginal Distribution

For example, consider our house price data. We can look at the distribution of house prices in "slices" determined by size ranges:



Conditional Distributions vs the Marginal Distribution

What do we see?

The conditional distributions are less variable (narrower boxplots) than the marginal distribution.

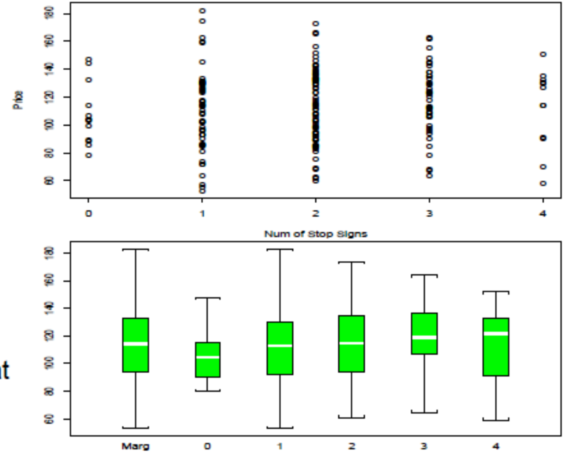
Variation in house sizes *expains* a lot of the original variation in price. What does this mean about SST, SSR, SSE, and R^2 from last time?

Conditional Distributions vs the Marginal Distribution

When X has no predictive power, the story is different:

House price (Y) vs. the number of stop signs within a two block radius of a house (X).

See that in this case, the marginal and the Conditionals are not that different!



Probability models for prediciton

"Slicing" our data is an awkward way to build a prediction and prediction interval (Why 500sqft slices and not 200 or 1000? What's the tradeoff between large and small slices?)

Instead we build a probability model (e.g., normal distribution).

Then we can say something like "with 95% probability the prediction error will be within $\pm\$28,000$".

We must also acknowledge that the "fitted" line may be fooled by particular realizations of the residuals (an unlucky sample)

The Simple Linear Regression Model

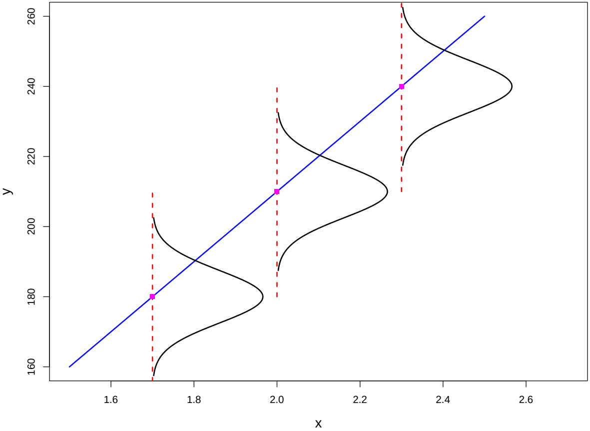
Simple Linear Regression Model: $Y = \beta_0 + \beta_1 X + \varepsilon$

$$\varepsilon \sim \mathrm{N}(0, \sigma^2)$$

- $\beta_0 + \beta_1 X$ represents the "true line"; The part of Y that depends on X.
- The error term ε is independent "idosyncratic noise"; The part of Y not associated with X.

The Simple Linear Regression Model

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



The conditional distribution for Y given X is Normal (why?):

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

The Simple Linear Regression Model – Example

You are told (without looking at the data) that

$$\beta_0 = 40;\ \beta_1 = 45;\ \sigma = 10$$

and you are asked to predict price of a 1500 square foot house.

What do you know about Y from the model?

$$\begin{aligned} Y &= 40 + 45(1.5) + \varepsilon \\ &= 107.5 + \varepsilon \end{aligned}$$

Thus our prediction for the price is $E(Y \mid X = 1.5) = 107.5$(the *conditional* expected value), and since $(Y \mid X = 1.5) \sim \mathrm{N}(107.5, 10^2)$ a 95% *Prediction Interval* for Y is $87.5 < Y < 127.5$

Summary of Simple Linear Regression

The model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$.

The SLR has 3 basic parameters:

- β_0, β_1 (linear pattern)
- σ (variation around the line).

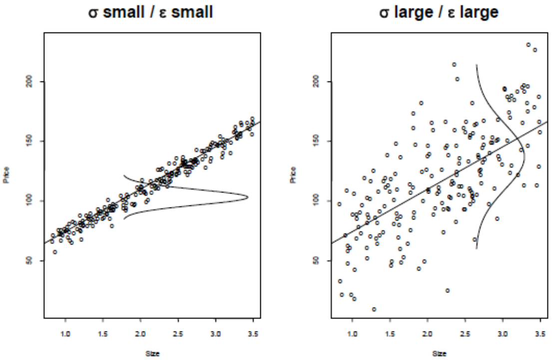
Assumptions:

- independence means that knowing ε_i doesn't affect your views about ε_j
- identically distributed means that we are using the same normal distribution for every ε_i

Conditional Distributions vs the Marginal Distribution

You know that β_0 and β_1 determine the linear relationship between X and the mean of Y given X.

σ determines the spread or variation of the realized values around the line (i.e., the *conditional* mean of Y)



Learning from data in the SLR Model

SLR assumes every observation in the dataset was generated by the model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

This is a model for the conditional distribution of Y given X.

We use Least Squares *to estimate* β_0 and β_1:

$$\hat{\beta}_1 = b_1 = r_{xy} \times \frac{s_y}{s_x}$$

$$\hat{\beta}_0 = b_0 = \bar{Y} - b_1 \bar{X}$$

Estimation of Error Variance

We estimate σ^2 with:

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

(2 is the number of regression coefficients; i.e. 2 for β_0 and β_1).

We have $n - 2$ degrees of freedom because 2 have been "used up" in the estimation of b_0 and b_1.

We usually use $s = \sqrt{SSE/(n-2)}$, in the same units as Y. It's also called the regression or residual standard error.

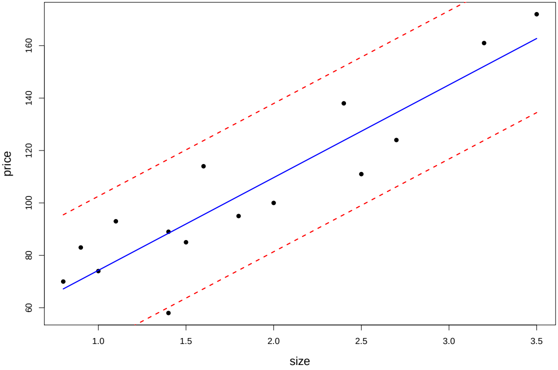
Finding s from R output

```
summary(fit)

##
## Call:
## lm(formula = Price ~ Size, data = housing)
##
## 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
```

One Picture Summary of SLR

- The plot below has the house data, the fitted regression line $(b_0 + b_1 X)$ and $\pm 2 * s$...
- From this picture, what can you tell me about b_0, b_1 and s^2?



How about β_0, β_1 and σ^2?

Sampling Distribution of Least Squares Estimates

How much do our estimates depend on the particular random sample that we happen to observe? Imagine:

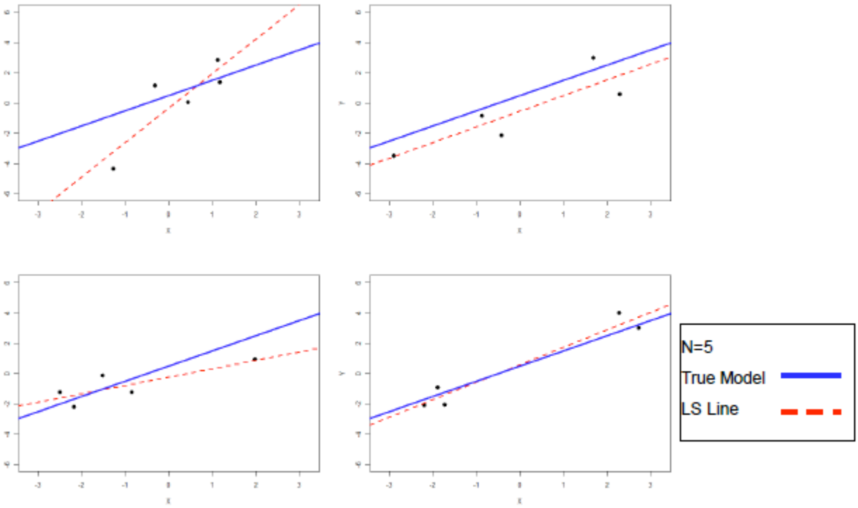
- Randomly draw different samples of the same size.
- For each sample, compute the estimates b_0, b_1, and s.

(just like we did for sample means in Section 1.4)

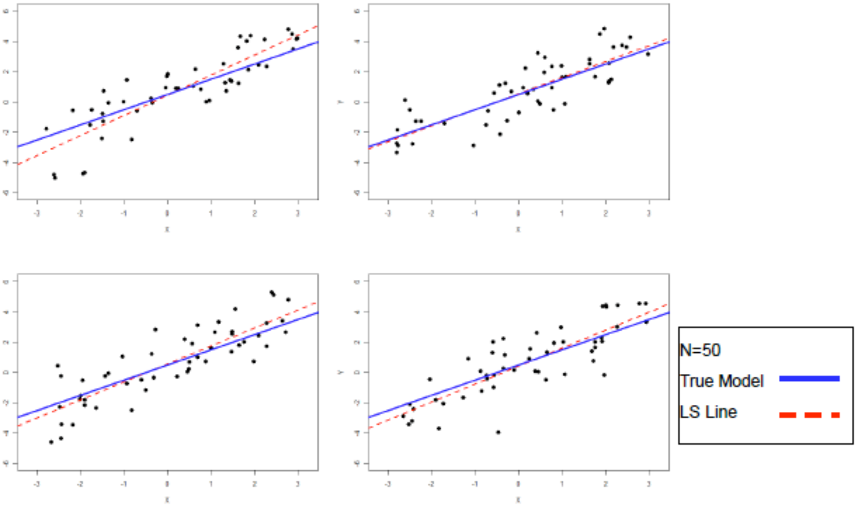
If the estimates don't vary much from sample to sample, then it doesn't matter which sample you happen to observe.

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

Sampling Distribution of Least Squares Estimates



Sampling Distribution of Least Squares Estimates



Sampling Distribution of b_1

The sampling distribution of b_1 describes how estimator $b_1 = \hat{\beta}_1$ varies over different samples with the X values fixed.

It turns out that b_1 is normally distributed (approximately): $b_1 \sim N(\beta_1, s_{b_1}^2)$.

- b_1 is unbiased: $E[b_1] = \beta_1$.
- s_{b_1} is the standard error of b_1. In general, the standard error of an estimate is its standard deviation over many randomly sampled datasets of size n. It determines how close b_1 is to β_1 on average.
- This is a number directly available from the regression output.

Sampling Distribution of b_1

Can we intuit what should be in the formula for s_{b_1}?

- How should s figure in the formula?
- What about n?
- Anything else?

$$s_{b_1}^2 = \frac{s^2}{\sum(X_i - \bar{X})^2} = \frac{s^2}{(n-1)s_x^2}$$

Three Factors:

sample size (n), error variance (s^2), and X-spread (s_x).

Sampling Distribution of b_0

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

$$s_{b_0}^2 = Var(b_0) = s^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2} \right)$$

What is the intuition here?

Confidence Intervals

Since $b_1 \sim N(\beta_1, s_{b_1}^2)$, Thus:

- 68% Confidence Interval: $b_1 \pm 1 \times s_{b_1}$
- 95% Confidence Interval: $b_1 \pm 2 \times s_{b_1}$
- 99% Confidence Interval: $b_1 \pm 3 \times s_{b_1}$

Same thing for b_0

- 95% Confidence Interval: $b_0 \pm 2 \times s_{b_0}$

The confidence interval provides you with a set of plausible values for the parameters

Finding standard errors from R output

```
summary(fit)

##
## Call:
## lm(formula = Price ~ Size, data = housing)
##
## 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
```

Confidence intervals in R

In R, you can extract confidence intervals easily:

```
confint(fit, level=0.95)

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

These are close to what we get by hand, but not exactly the same:

```
38.885 - 2*9.094; 38.885 + 2*9.094;

## [1] 20.697
## [1] 57.073

35.386 - 2*4.494; 35.386 + 2*4.494;
```

Confidence intervals in R

Why don't our answers agree?

R is using a slightly more accurate approximation to the sampling distribution of the coefficients, based on the t distribution.

The difference only matters in small samples, and if it changes your inferences or decisions then you probably need more data!

Testing

Suppose we want to assess whether or not β_1 equals a proposed value β_1^0. This is called hypothesis testing.

Formally we test the null hypothesis:

$H_0 : \beta_1 = \beta_1^0$

vs. the alternative

$H_1 : \beta_1 \neq \beta_1^0$

(For example, testing $\beta_1 = 0$ vs. $\beta_1 \neq 0$ is testing whether X is predictive of Y **under our SLR model assumptions.**)

Testing

That are 2 ways we can think about testing a regression coefficient:

1. Building a test statistic... the t-stat,

$$t = \frac{b_1 - \beta_1^0}{s_{b_1}}$$

   This quantity measures how many standard errors (SD of b_1) the estimate (b_1) is from the proposed value (β_1^0).

   If the absolute value of t is greater than 2, we need to worry (why?)... we reject the null hypothesis.

Testing

2. Looking at the confidence interval. If the hypothesized value is outside the confidence interval you reject the null hypothesis.

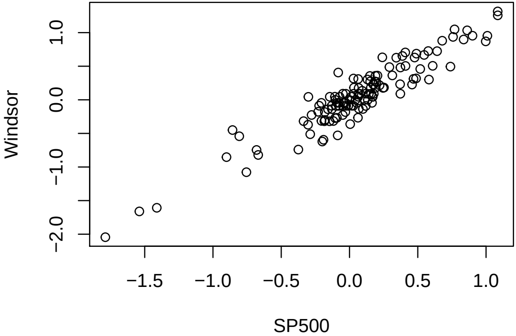
   Notice that this is equivalent to the t-stat. An absolute value for t greater than 2 implies that the proposed value is outside the confidence interval... therefore reject.

   In fact, a 95% confidence interval contains all the values for a parameter that are **not** rejected by hypothesis test with a false positive rate of 5%

   This is my preferred approach for the testing problem. You can't go wrong by using the confidence interval!

Example: Mutual Funds

Let's investigate the performance of the Windsor Fund, an aggressive large cap fund by Vanguard...



The plot shows 6mos of daily returns for Windsor vs. the S&P500

Example: Mutual Funds

Consider the following regression model for the Windsor mutual fund:

$$r_w = \beta_0 + \beta_1 r_{sp500} + \epsilon$$

Let's first test $\beta_1 = 0$

$H_0: \beta_1 = 0$. Is the Windsor fund related to the market?

$H_1: \beta_1 \neq 0$

Example: Mutual Funds

```
##
## Call:
## lm(formula = Windsor ~ SP500, data = windsor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42557 -0.11035  0.01057  0.11915  0.50539
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01027    0.01602  -0.641    0.523
## SP500        1.07875    0.03498  30.841   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1777 on 124 degrees of freedom
## Multiple R-squared:  0.8847,Adjusted R-squared:  0.8837
## F-statistic: 951.2 on 1 and 124 DF,  p-value: < 2.2e-16
```

Example: Mutual Funds

The approximate 95% confidence interval is
$1.079 \pm 2 \times 0.035 = (1.009.1.149)$, so we'd reject $H_0: \ \beta = 0$

```
confint(fit, level=0.95)

##                    2.5 %   97.5 %
## (Intercept) -0.04197045 0.021435
## SP500        1.00951622 1.147976
```

The $t-$ statistic is $(1.079 - 0)/0.035 = 30.8$ (see also the R output) - reject!

Example: Mutual Funds

Now let's test $\beta_1 = 1$. What does that mean?

$H_0: \ \beta_1 = 1$ Windsor is as risky as the market.

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

We are asking whether Windsor moves in a different way than the market (does it exhibit larger/smaller changes than the market, or about the same?).

Example: Mutual Funds

The approximate 95% confidence interval still $1.079 \pm 2 \times 0.035 = (1.009.1.149)$, so we'd reject $H_0: \ \beta = 1$ as well.

```
confint(fit, level=0.95)

##                   2.5 %   97.5 %
## (Intercept) -0.04197045 0.021435
## SP500        1.00951622 1.147976
```

The $t-$ statistic is $(1.079 - 1)/0.035 = 2.26$ - reject!

But...

Testing – Why I like giving an interval

- What if the Windsor beta estimate had been 1.07 with a CI of (0.99, 1.14)? Would our assessment of the fund's market risk really change?
- Now suppose in testing $H_0 : \beta_1 = 1$ you got a t-stat of 6 and the confidence interval was

  $$[1.00001, 1.00002]$$

  Do you reject $H_0 : \beta_1 = 1$ and conclude Windsor is riskier than the market? Could you justify that to your boss? Probably not! (why?)

Testing – Why I like giving an interval

- Now, suppose in testing $H_0 : \beta_1 = 1$ you got a t-stat of -0.02 and the confidence interval was

$$[-100, 100]$$

Do you "accept" $H_0 : \beta_1 = 1$? Could you justify that to you boss? Probably not! (why?)

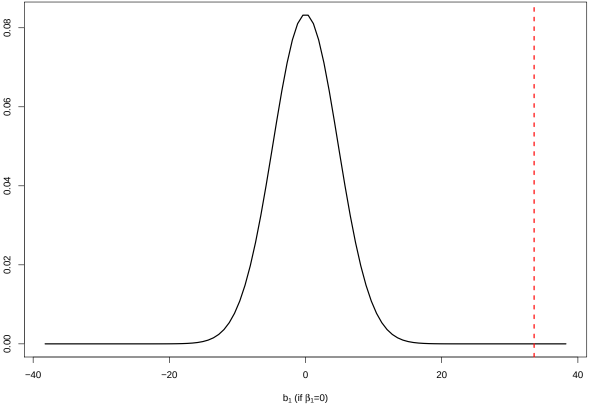
The confidence interval is your friend when it comes to testing regression coefficients

P-values

- The p-value provides a measure of how weird your estimate is **if** the null hypothesis is true
- Small p-values are evidence against the null hypothesis
- In the AVG vs. R/G example... $H_0 : \beta_1 = 0$. How weird is our estimate of $b_1 = 33.57$?
- Remember: $b_1 \sim N(\beta_1, s_{b_1}^2)$... If the null was true ($\beta_1 = 0$), $b_1 \sim N(0, s_{b_1}^2)$

P-values

- Where is 33.57 in the picture below?



The p-value is the probability of seeing b_1 equal or greater than 33.57 in absolute terms. Here, p-value$=0.000000124$!!

Small p-value = bad null

P-values - Windsor fund

R will report p-values for testing each coefficient at $\beta_j = 0$, in the column $\Pr(>|t|)$

```
##
## Call:
## lm(formula = Windsor ~ SP500, data = windsor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42557 -0.11035  0.01057  0.11915  0.50539
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01027    0.01602  -0.641    0.523
## SP500        1.07875    0.03498  30.841   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1777 on 124 degrees of freedom
## Multiple R-squared:  0.8847,Adjusted R-squared:  0.8837
## F-statistic: 951.2 on 1 and 124 DF,  p-value: < 2.2e-16
```

P-values for other null hypotheses

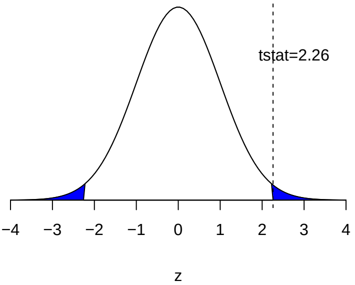
We have to do other tests ourselves: To get a p-value for $H_0 : \beta_1 = q$ versus $H_0 : \beta_1 \neq q$, note that $b_1 \sim N(q, se(b_1))$ (approximately) under the null, and

$$\frac{(b_1 - q)}{se(b_1)} \sim N(0, 1)$$

P-values for other null hypotheses

Under H_0, prob. of seeing a coefficient *at least* as extreme as b_1 is:

$$\Pr(|Z| > |t|), \quad t = (b_1 - q)/se(b_1)$$



The p-value for testing $H_0 : \beta = 1$ in the Windsor data is

```
2*pnorm(abs(1.079 - 1)/0.035, lower.tail=FALSE)

## [1] 0.02399915
```

Testing – Summary

- Large t or small p-value mean the same thing...
- p-value < 0.05 is equivalent to a t-stat > 2 in absolute value
- Small p-value means the data at hand are unlikely to be observed if the null hypothesis was true...
- Bottom line, small p-value $\rightarrow$ REJECT! Large t $\rightarrow$ REJECT!
- But remember, always look at the confidence interveal!

Prediction/Forecasting under Uncertainty

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: $\hat{Y}_f = b_0 + b_1 X_f$.

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

Forecasting: Plug-in Method

A common approach is to assume that $\beta_0 \approx b_0$, $\beta_1 \approx b_1$ and $\sigma \approx s$... in this case the 95% plug-in prediction interval is:

$$(b_0 + b_1 X_f) \pm 2 \times s$$

It's called "plug-in" because we just plug-in the estimates (b_0, b_1 and s) for the unknown parameters (β_0, β_1 and σ).

Forecasting: Better intervals in R

But remember that you are uncertain about b_0 and b_1! As a practical matter if the confidence intervals are big you should be careful! R will give you a larger (and correct) prediction interval.

A larger prediction 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 difference between X_f and $\bar{X}$.

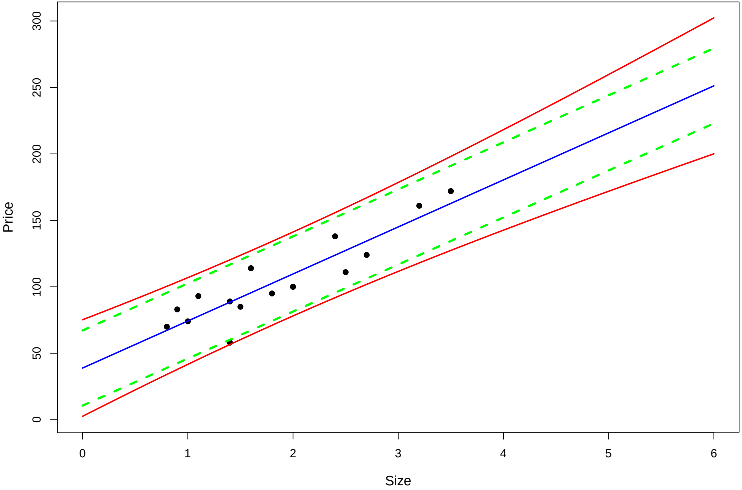
Forecasting: Better intervals in R

```
fit = lm(Price~Size, data=housing)

# Make a data.frame with some X_f values
# (X values where we want to forecast)
newdf = data.frame(Size=c(1, 1.85, 3.2, 4.1))
predict(fit, newdata = newdf,
        interval = "prediction", level = 0.95)

##         fit       lwr      upr
## 1  74.27065  41.65499 106.8863
## 2 104.34871  72.80283 135.8946
## 3 152.11976 117.97174 186.2678
## 4 183.96713 145.61441 222.3199
```

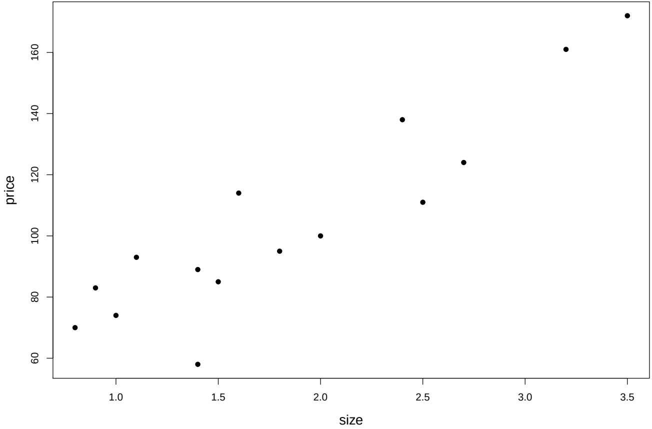
Forecasting: Better intervals in R



- Red lines: prediction intervals
- Green lines: "plug-in" prediction intervals

House Data – one more time!

- $R^2 = 82\%$
- Great R^2, we are happy using this model to predict house prices, right?



House Data – one more time!

- But, $s = 14$ leading to a predictive interval width of about US\$60,000!! How do you feel about the model now?
- As a practical matter, s is a much more relevant quantity than R^2. Once again, *intervals* are your friend!