STAT 215
Confidence and Prediction Intervals in Regression

Colin Reimer Dawson

Oberlin College

24 October 2016

Outline

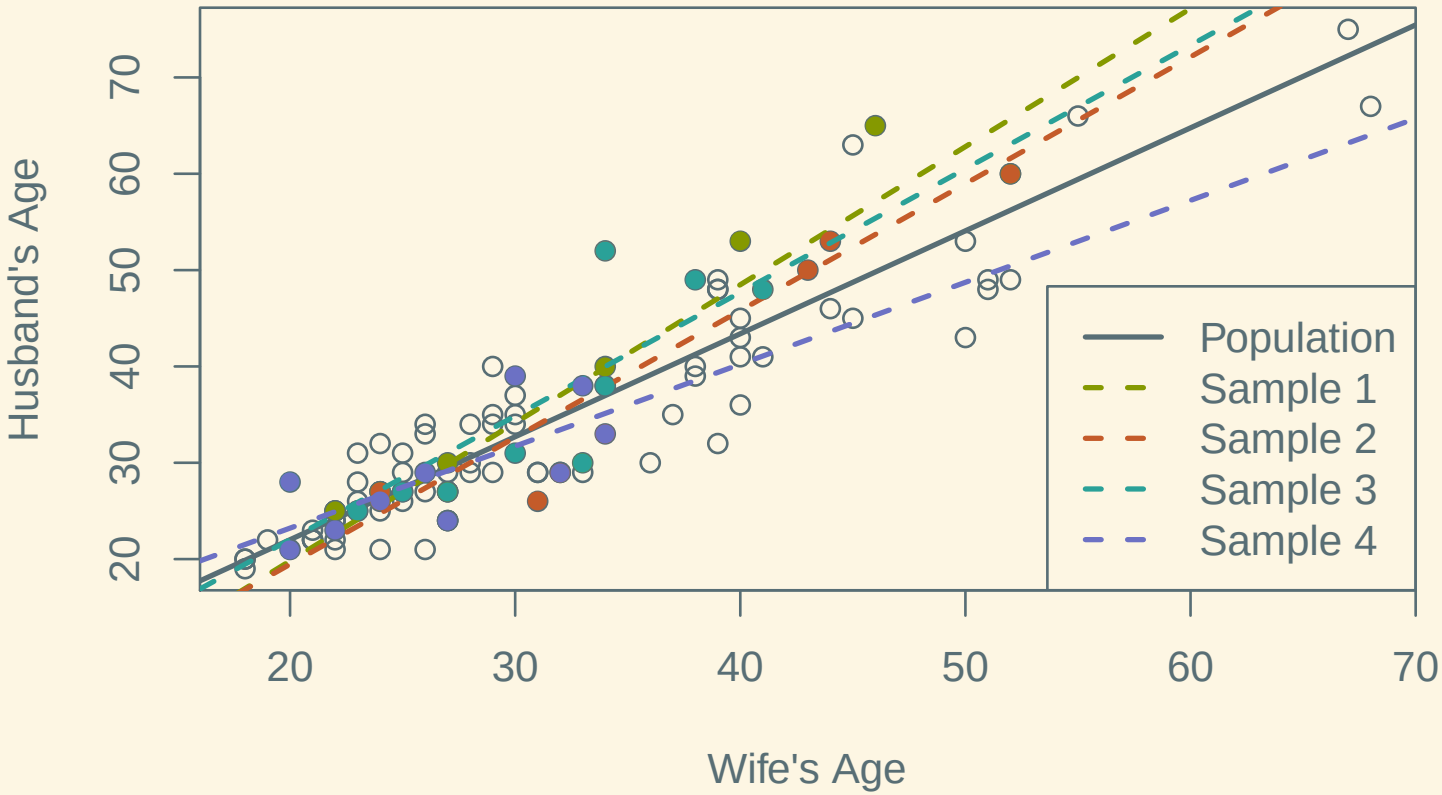
Reminder: Sampling Distributions

Sampling Distribution

The **sampling distribution** of a sample statistic (e.g., $\hat{\beta}_1$) for β_1, or $\bar{Y}$ for μ_Y) is the distribution that statistic has across all possible samples from the population.

Sample vs Population "Best-Fit" Line

- For a sample: choose intercept and slope to minimize sum of squared errors.
- But this does not yield the "correct" (or even "best") model for the population, due to sampling error.

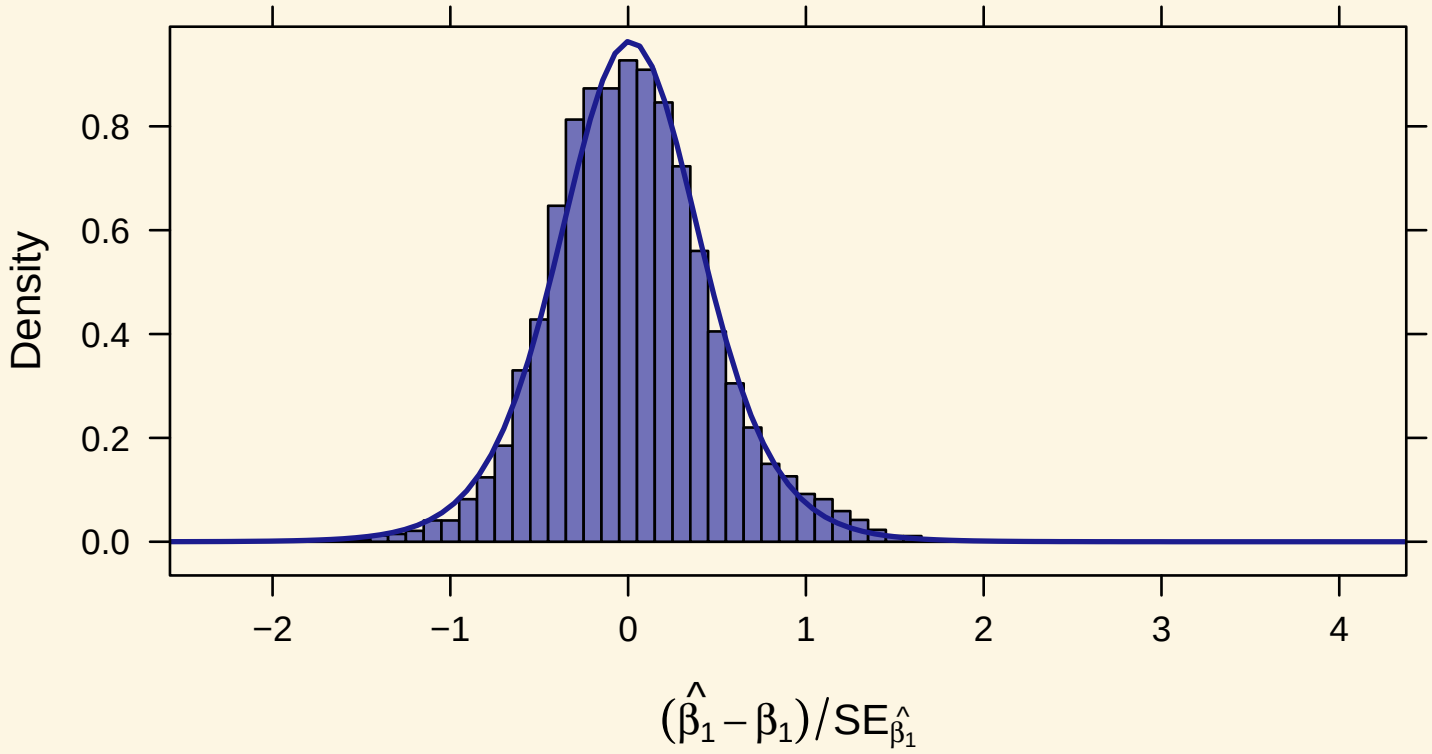


Two Methods for Estimating Sampling Distribution of $\hat{\beta}_1$

1. t-distribution: assumes Normal residuals (along with other regression conditions)
2. Bootstrap distribution: no Normal assumption needed

t-distribution model (assumes Normal residuals)

$$\text{If } \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon), \text{ then } \frac{\hat{\beta}_i - \beta_i}{SE_{\hat{\beta}_i}} \sim t_{n-2}$$



t-based Confidence Interval

$$CI_{1-\alpha} : \hat{\beta}_i \pm t_{n-2}^{*(1-\alpha/2)} \cdot SE_{\beta_i} \tag{1}$$

where $t_{n-2}^{*(1-\alpha/2)}$ represents the $1-\alpha/2$ quantile of the t_{n-2} distribution.

```
sample.model <- lm(Husband ~ Wife, data = sample1)
summary(sample.model)
```

```

    Call:
    lm(formula = Husband ~ Wife, data = sample1)

    Residuals:
        Min      1Q  Median      3Q     Max
    -7.3577 -4.9705  0.7395  3.3295  7.5107

    Coefficients:
                Estimate Std. Error t value Pr(>|t|)
    (Intercept) -13.7455     7.1773  -1.915   0.0918 .
    Wife          1.5486     0.1989   7.786  5.3e-05 ***
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Residual standard error: 5.774 on 8 degrees of freedom
    Multiple R-squared:  0.8834,Adjusted R-squared:  0.8689
    F-statistic: 60.63 on 1 and 8 DF,  p-value: 5.304e-05
```

```
MoE.95 <- qt(0.975, df = 8) * 0.1989
(CI.95 <- c(1.5486 - MoE.95, 1.5486 + MoE.95))
```

```

    [1] 1.089936 2.007264
```

```
confint(sample.model, level = 0.95)
```

```
                    2.5 %   97.5 %
(Intercept) -30.296476 2.805433
Wife          1.089947 2.007218
```

t-based Hypothesis Test

$$t_{obs} = \frac{\hat{\beta}_1 - 0}{SE_{\hat{\beta}_1}} \tag{2}$$

where $t_{n-2}^{*(1-\alpha/2)}$ represents the $1 - \alpha/2$ quantile of the t_{n-2} distribution.

Correlation Test and Interval

Can also estimate dist. for correlation r using t_{n-2}, where

$$SE_r = \sqrt{\frac{1-r^2}{n-2}} \tag{3}$$

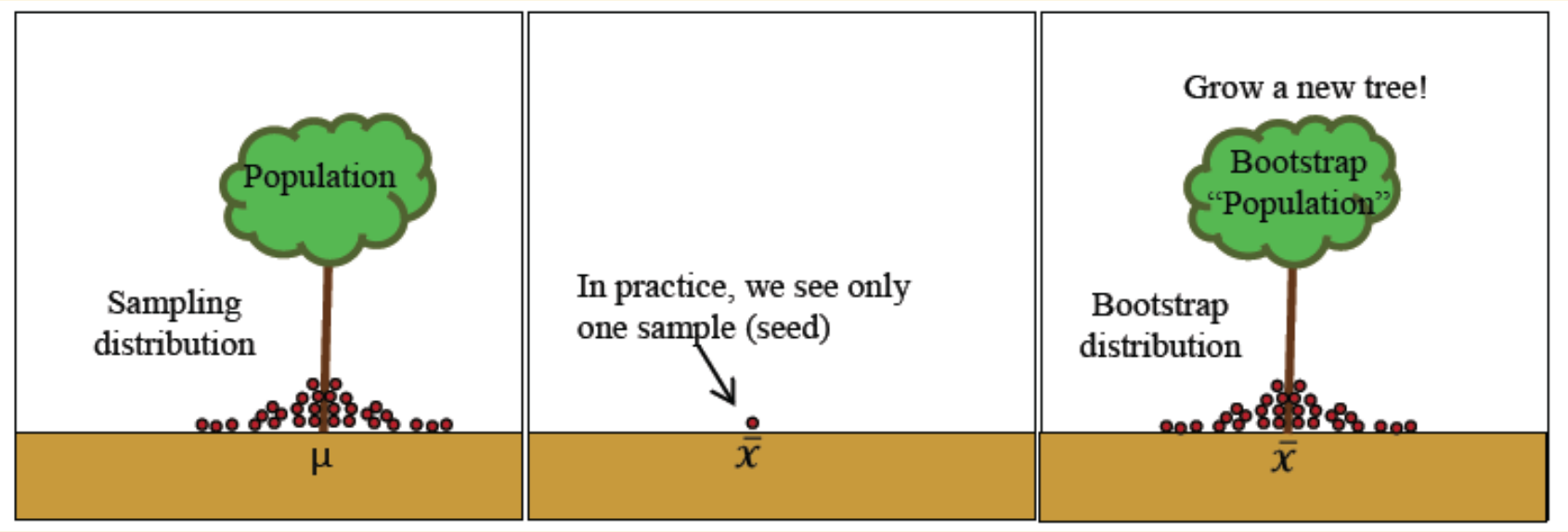
$$CI_{1-\alpha} : r \pm t_{n-2}^{*(1-\alpha/2)} \cdot SE_r \tag{4}$$

$$t_{obs} = \frac{r-0}{SE_r} \tag{5}$$

Interpretation of Tests and Intervals

- $100(1-\alpha)$% CI: We are $100(1-\alpha)$ confident that the population parameter (such as β_1 or ρ) lies in the interval. (Our interval will miss on 100α% of datasets)
- P-value: The likelihood of a sample (slope / correlation) with a magnitude this large if the population value is zero is P.
  - P small $\Rightarrow$ would be too surprising to get such a large value "by chance alone" $\Rightarrow$ Reject H_0

Bootstrap Distribution



Bootstrap Distribution

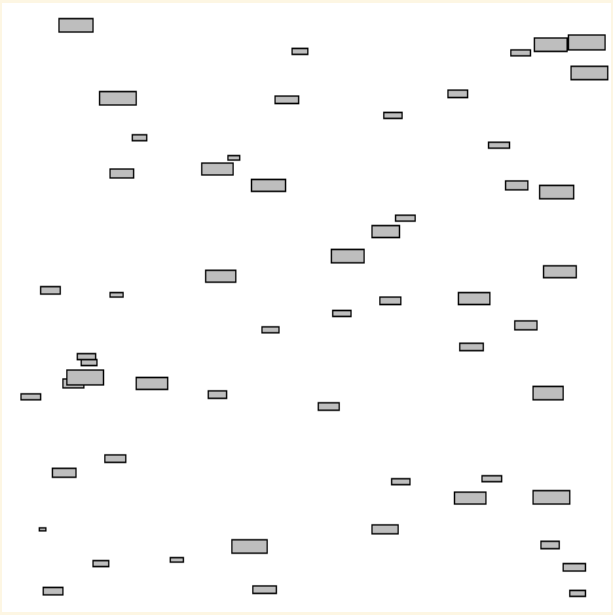


Figure: Our actual sample

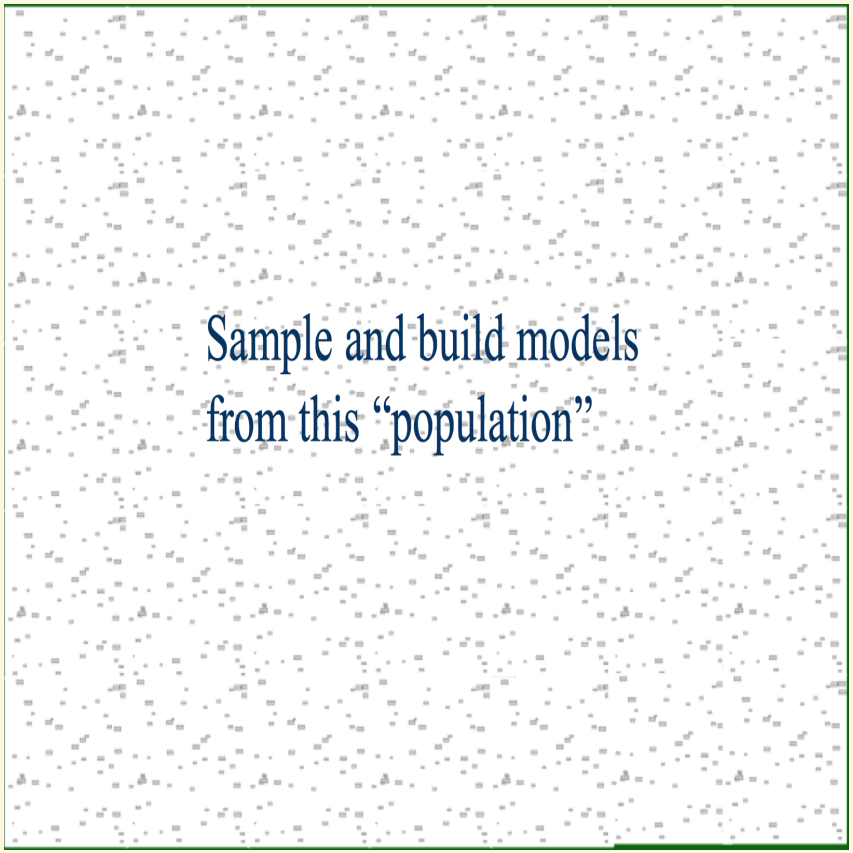


Figure: Our simulated population

Permutation Test: Slope (or correlation)

To test $H_0 : \beta_1 = 0$ (or $H_0 : \rho = 0$), we want probability that a random $\hat{\beta}_1$ (random r) is as large or larger than observed $\hat{\beta}_1$ (observed r), assuming H_0 true: $\beta_1 = 0$ ($\rho = 0$).

Permutation Test: Slope (Correlation)

1. Simulate H_0 by randomly pairing X and Y values, and computing $\hat{\beta}_1$ (or r) for each pseudodataset.
2. Repeat many times
3. Calculate proportion of random $\hat{\beta}_1$ (or r) that exceed actual $\hat{\beta}_1$ (or r). This is the P-value of the test.
4. If $P < \alpha$ for predetermined α, reject H_0.

Interpretation of Tests and Intervals

Exactly the same as before.

ANOVA for Regression

$$\begin{array}{ccccc} \text{DATA} & = & \text{PATTERN} & + & \text{IDIOSYNCRACIES} \\ Y & = & f(X) & + & \varepsilon \end{array}$$

$$\begin{array}{ccccc} \text{Total Variation} & = & \text{Explained} & + & \text{Unexplained} \\ Y - \bar{Y} & = & \hat{Y} - \bar{Y} & + & Y - \hat{Y} \\ \sum_i (Y_i - \bar{Y})^2 & = & \sum_i (\hat{Y}_i - \bar{Y})^2 & +0+ & \sum (Y_i - \hat{Y}_i)^2 \\ SSTotal & = & SSModel & + & SSError \end{array}$$

"Omnibus" F-test for a Regression Model

$$F = \frac{SSModel/df_{Model}}{SSError/df_{Error}} = \frac{MS_{Model}}{MS_{Error}}$$

This statistic has an F distribution with corresponding df if the null model is correct (i.e., $Y = \beta_0 + \varepsilon$)

```
brain.model <-
    lm(log(brain.weight.grams) ~ log(body.weight.kilograms),
       data = BrainBodyWeight)
anova(brain.model)


     Analysis of Variance Table

     Response: log(brain.weight.grams)
                               Df Sum Sq Mean Sq F value    Pr(>F)
     log(body.weight.kilograms)  1 336.19  336.19  697.42 < 2.2e-16 ***
     Residuals                  60  28.92    0.48
     ---
     Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Proportion of Variability Explained

The Coefficient of Determination (R^2)

The **coefficient of determination**, or R^2 value, associated with a linear model, is the percent reduction in prediction uncertainty achieved by the regression model compared to the null model

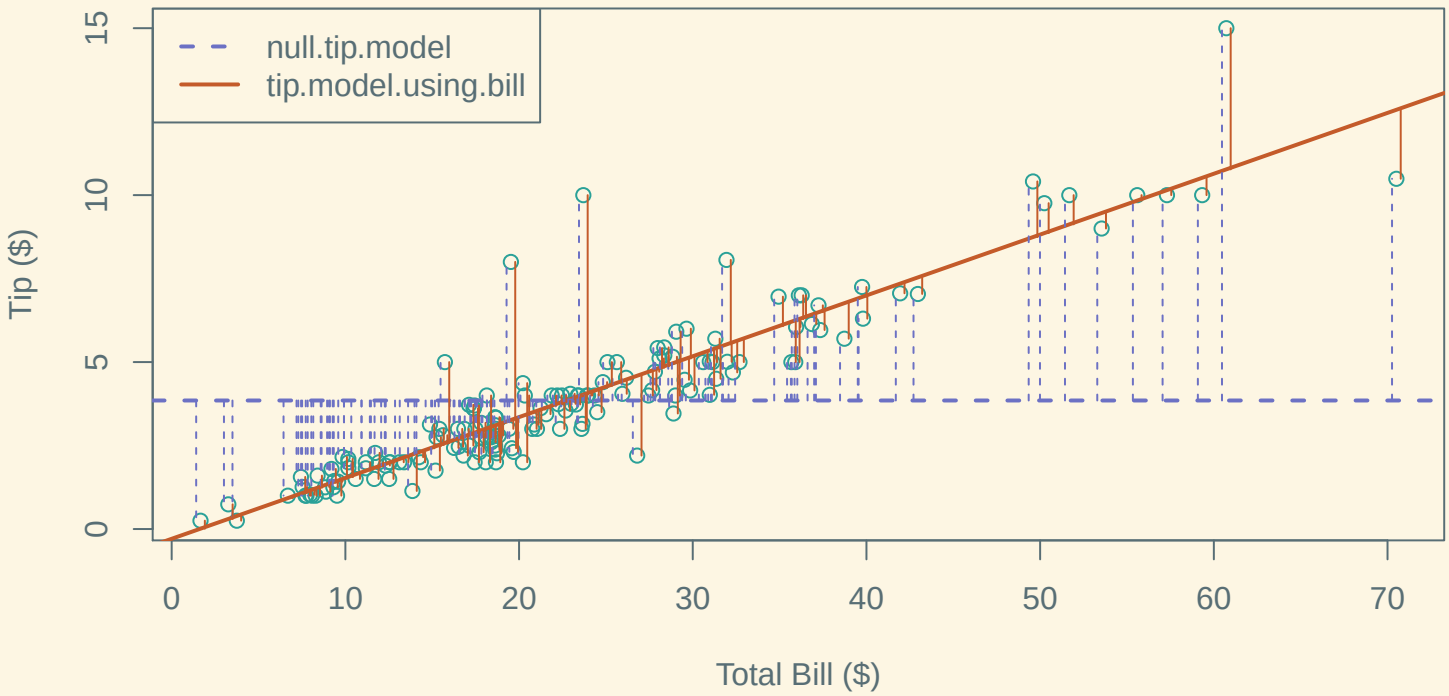
I.e., what proportion of the variation (variance) in y is "explained":

$$R^2 = \frac{SSModel}{SSTotal} \tag{6}$$

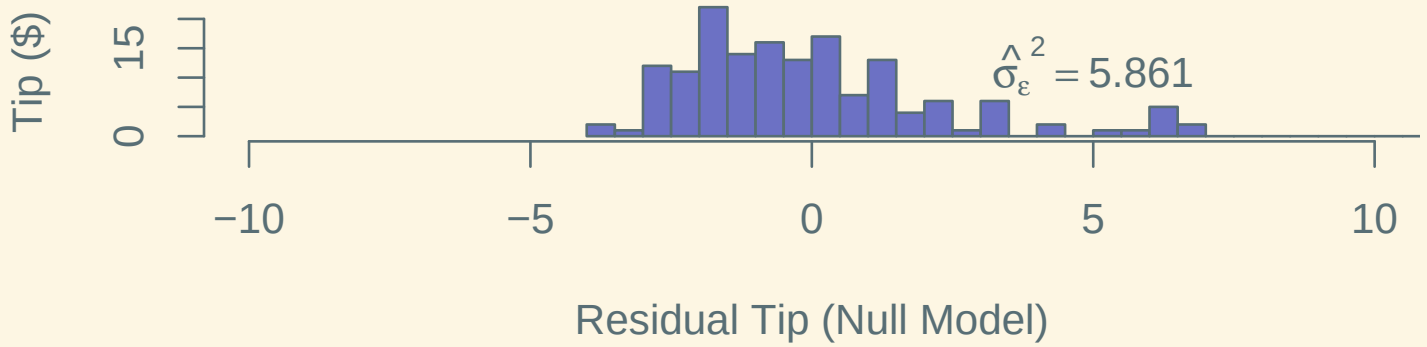
Turns out to just be the square of the correlation! (Show this algebraically)

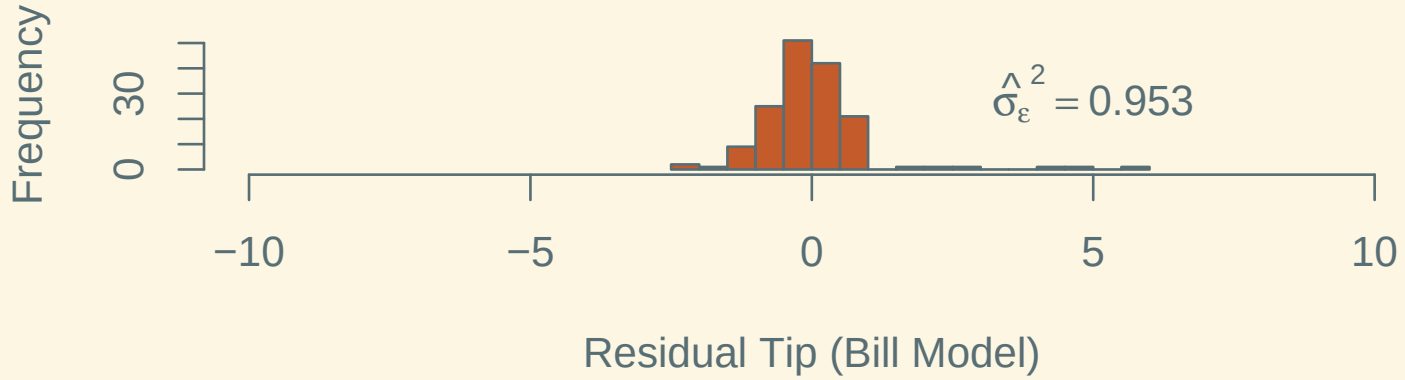
Example: Restaurant Tips

```
library("Lock5Data"); library("mosaic")
data("RestaurantTips")
null.tip.model <- lm(Tip ~ 1, data = RestaurantTips)
tip.model.using.bill <- lm(Tip ~ Bill, data = RestaurantTips)
```



Example: Restaurant Tips





Regression Summary

```
R.squared <- 1 - 0.953 / 5.861; R.squared
```

```
[1] 0.8373998
```

```
summary(tip.model.using.bill)
```

```
Call:
lm(formula = Tip ~ Bill, data = RestaurantTips)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3911 -0.4891 -0.1108  0.2839  5.9738

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.292267   0.166160  -1.759   0.0806 .
Bill         0.182215   0.006451  28.247   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9795 on 155 degrees of freedom
Multiple R-squared:  0.8373,Adjusted R-squared:  0.8363
F-statistic: 797.9 on 1 and 155 DF,  p-value: < 2.2e-16
```

Intervals at a particular X

- A confidence interval for the slope is useful, but if our goal is a predictive model, we want to be able to make statements about Y values at particular X values.
- I should be able to estimate
  1. What the mean Y value is at that X *in the population*
  2. Where the particular Y is likely to be for *this one new observation*
- Note: These are different things, in the same way that a 95% confidence interval does *not* tell us where 95% of the *individual cases* are.

Confidence and Prediction Intervals for a Linear Model

(Population) linear model:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \varepsilon \\ &= f(X) + \varepsilon \end{aligned}$$

1. A **confidence interval** (for a particular X) is an estimate (with a margin of error) of $f(X)$.
2. A **prediction interval** (for a particular X) is an estimate about Y

Confidence vs. Prediction Intervals

Which is wider? The prediction interval is wider, b/c it has uncertainty about ε *plus* the uncertainty about $f(X)$

A Subtlety Re: Prediction Intervals

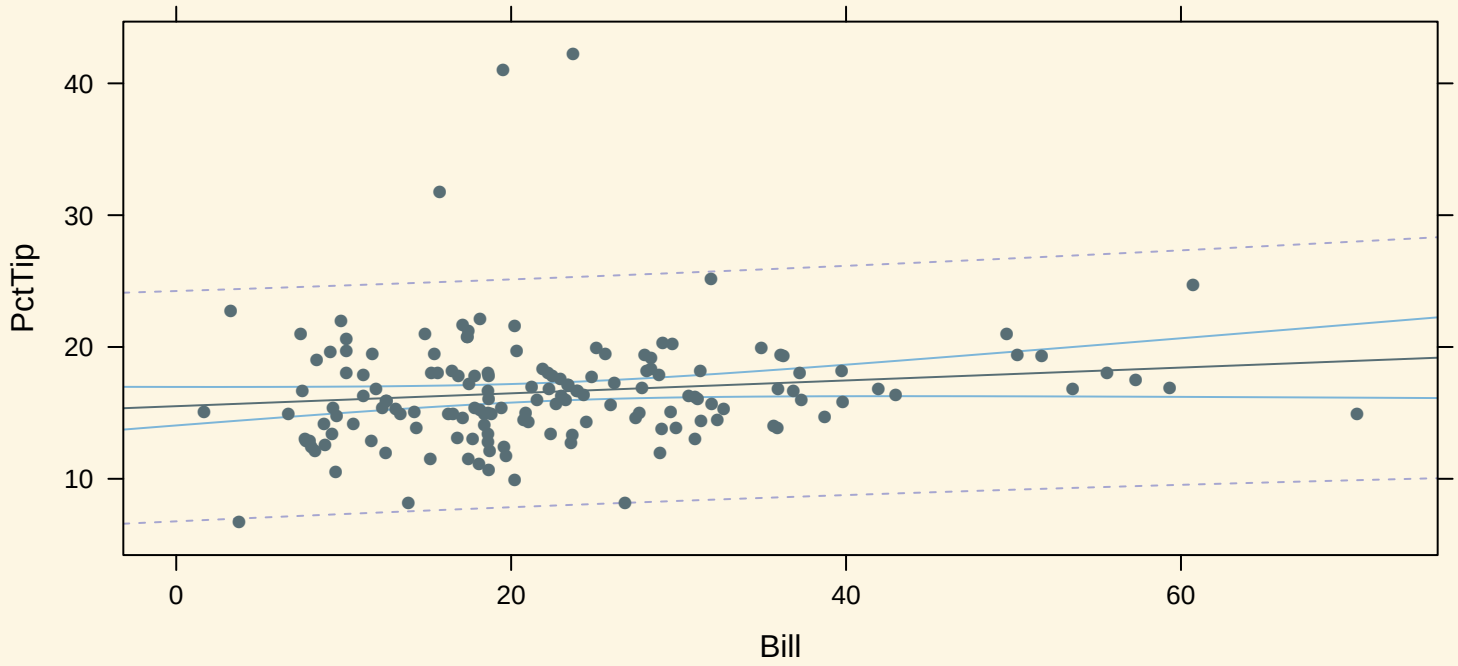
Interpreting Prediction Intervals

A coverage level of 95% for a prediction interval does *not* mean that, having fit a model from a *particular* sample, we will make successful predictions 95% of the time going forward. The worse our line, the lower the %.

What we can say is that the *average* success rate across *all possible samples* is 95%

Confidence and Prediction Bands

Intervals for all x in the range are called "confidence / prediction bands".



Why the hourglass shape? More leverage at extreme X^*: bigger change in line from one sample to the next

Calculating Confidence and Prediction Intervals

Both types of intervals are of the form

$$(1-\alpha) \text{ interval } = \text{Point Estimate} \pm t_{n-2}^{*(1-\alpha/2)} \cdot SE$$

Confidence Interval:

$$\hat{f}(X^*) \pm t_{n-2}^{*(1-\alpha/2)} \cdot \sqrt{\hat{\sigma}^2_{\hat{f}(X^*)}}$$

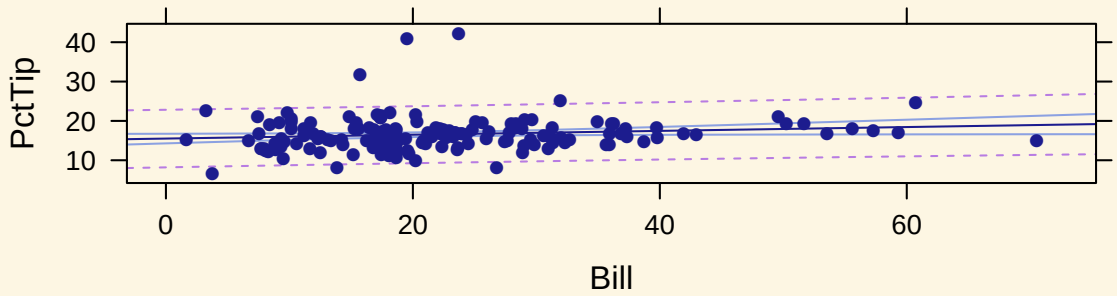
where $\hat{\sigma}^2_{\hat{f}(X^*)} = \hat{\sigma}^2_{\varepsilon} h(X^*)$ and $h(X^*) = \frac{1}{n} + \frac{(x*-\bar{x})^2}{\sum(x_i-\bar{x})^2}$ is the leverage at X^*.

Prediction Interval:

$$\hat{Y}* \pm t_{n-2}^{*(1-\alpha/2)} \cdot \sqrt{\hat{\sigma}^2_{\hat{f}(X^*)} + \hat{\sigma}^2_{\varepsilon}}$$

R code for a confidence/prediction bands plot:

```
library("mosaic"); library("Lock5Data")
data("RestaurantTips")
xyplot(PctTip ~ Bill, data = RestaurantTips,
       panel = panel.lmbands,  # Note, no quotes
       level = 0.90,           # The confidence level
       ## OPTIONAL: band.lty= what kind of lines to use
       ##    format: c(conf.linetype, pred.linetype), where
       ##    1 = solid, 2 = dashed, 3 = dotted
       band.lty = c(1,2),
       ## OPTIONAL: band.col: what color lines to use
       ##    format: c(conf.color, pred.color)
       band.col = c("royalblue", "blueviolet")
       )
```



We can get intervals for specific X values as follows:

```
tip.model.using.bill <- lm(PctTip ~ Bill, data = RestaurantTips)
## Creates a new function with the given name
f.hat <- makeFun(tip.model.using.bill)
## Use it like a regular function
##   First arg name: name of predictor variable
##      (= the desired x value to get the interval for)
##   interval="confidence" or interval="prediction"
##      controls which interval type to return
##      (or leave this out to just get the pt estimate)
##   level=confidence.level controls the confidence level
f.hat(Bill = 40, interval = "confidence", level = 0.90)

            fit      lwr      upr
    1 17.46215 16.45974 18.46455

f.hat(Bill = 40, interval = "prediction", level = 0.90)

            fit     lwr      upr
    1 17.46215 10.1786 24.74569
```