

STAT 213

Regression Inference

Colin Reimer Dawson

Oberlin College

February 14, 2018

Outline

Regression Inference

Partitioning Variability

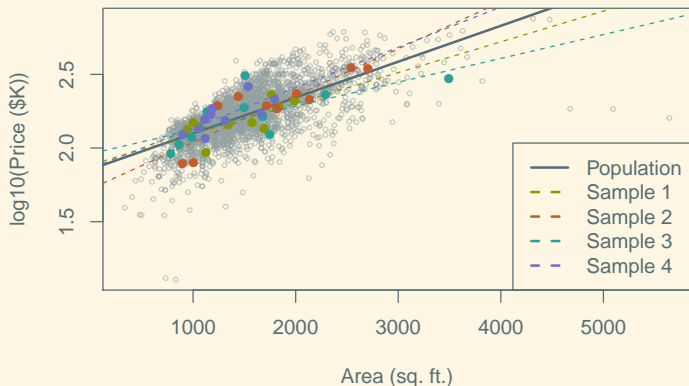
Outline

Regression Inference

Partitioning Variability

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.

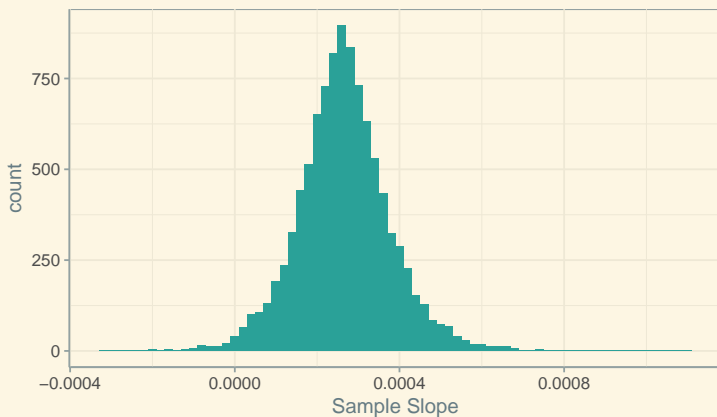


Reminder: Sampling Distributions

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.

Predicting Home Prices in Ames, Iowa

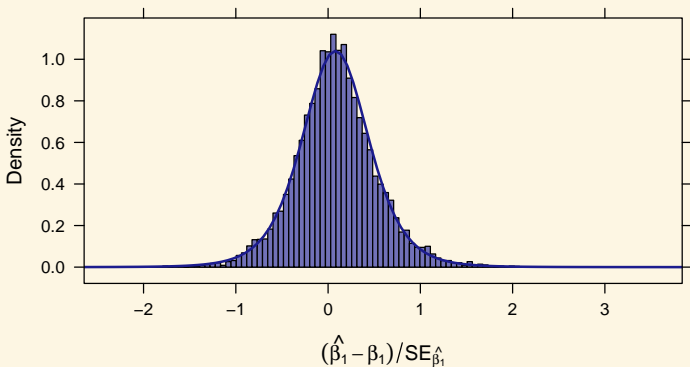


Two Methods for Estimating Sampling Distribution

1. t -distribution: assumes Normal residuals (along with other regression conditions)
2. Bootstrap distribution: no Normal assumption needed (won't use in this class)

Normal Residuals

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



t -based Confidence Interval

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

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

```
sample10 <- sample(Ames, 10) ## This would just be our dataset
sample.model <- lm(Price ~ Area, data = sample10)
summary(sample.model)$coefficients %>% round(digits = 3)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	154146.420	34633.509	4.451	0.002
Area	16.231	15.503	1.047	0.326

```
MoE.95 <- qt(0.975, df = 10 - 2) * 15.503
CI.95 <- c(16.231 - MoE.95, 16.231 + MoE.95)
CI.95
```

```
[1] -19.51898  51.98098
```

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

	2.5 %	97.5 %
(Intercept)	74281.40618	234011.43423
Area	-19.51942	51.98123

Outline

Regression Inference

Partitioning Variability

ANOVA for Regression

$$Y = f(X) + \varepsilon$$

DATA = PATTERN + IDIOSYNCRACIES

Total Variation = Explained Variation + Unexplained Variation

$$Y - \bar{Y} = \hat{Y} - \bar{Y} + Y - \hat{Y}$$

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + 0 + \sum_i (Y_i - \hat{Y}_i)^2$$

$$SSTotal = SSModel + SSEerror$$

“Omnibus” F -test for a Regression Model

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

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

```
BrainBodyWeight <- read.file("http://colindawson.net/data/BrainBodyWeight.csv")
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

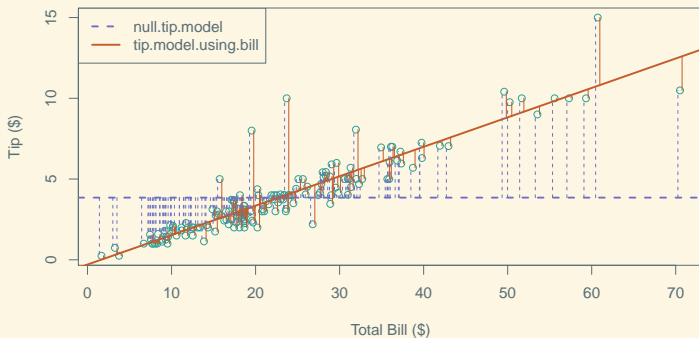
I.e., what proportion of the variation (variance) in y is “explained”:

$$R^2 = \frac{SS_{Model}}{SS_{Total}} \quad (1)$$

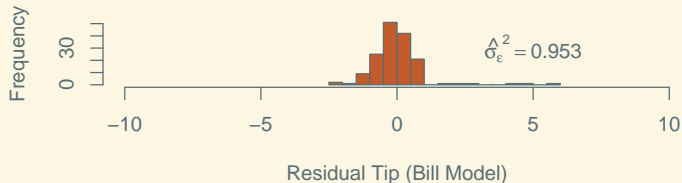
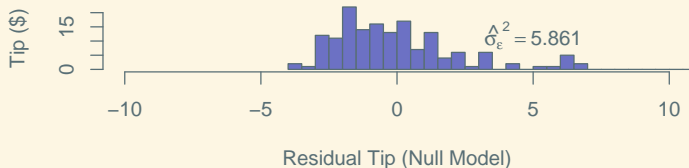
Turns out to just be the square of the correlation! (Show this algebraically)

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

```
summary(brain.model)
```

Call:

```
lm(formula = log(brain.weight.grams) ~ log(body.weight.kilograms),
    data = BrainBodyWeight)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.71550	-0.49228	-0.06162	0.43597	1.94829

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.13479	0.09604	22.23	<2e-16 ***
log(body.weight.kilograms)	0.75169	0.02846	26.41	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6943 on 60 degrees of freedom

Multiple R-squared: 0.9208, Adjusted R-squared: 0.9195

F-statistic: 697.4 on 1 and 60 DF, p-value: < 2.2e-16