

REGRESSION INTRODUCTION

Geog 210B Analytical Methods II
Winter 2018

+ January 15 - January 21

[Edit](#)

+  LECTURE 1 

[Edit](#)

+  Data for Lab SmallHHfile 

[Edit](#)

+  Intro210B.R 

[Edit](#)

+  Codebook for SmallHHfile.csv 

[Edit](#)

+  LECTURE 2 

[Edit](#)[+ Add an activity or resource](#)

+ January 22 - January 28

[Edit](#)

+  LECTURE 3 Intro to Regression January 23 

[Edit](#)

+  LECTURE 4 Matrix basics and regression 

[Edit](#)

+  R Script for January 23 

[Edit](#)

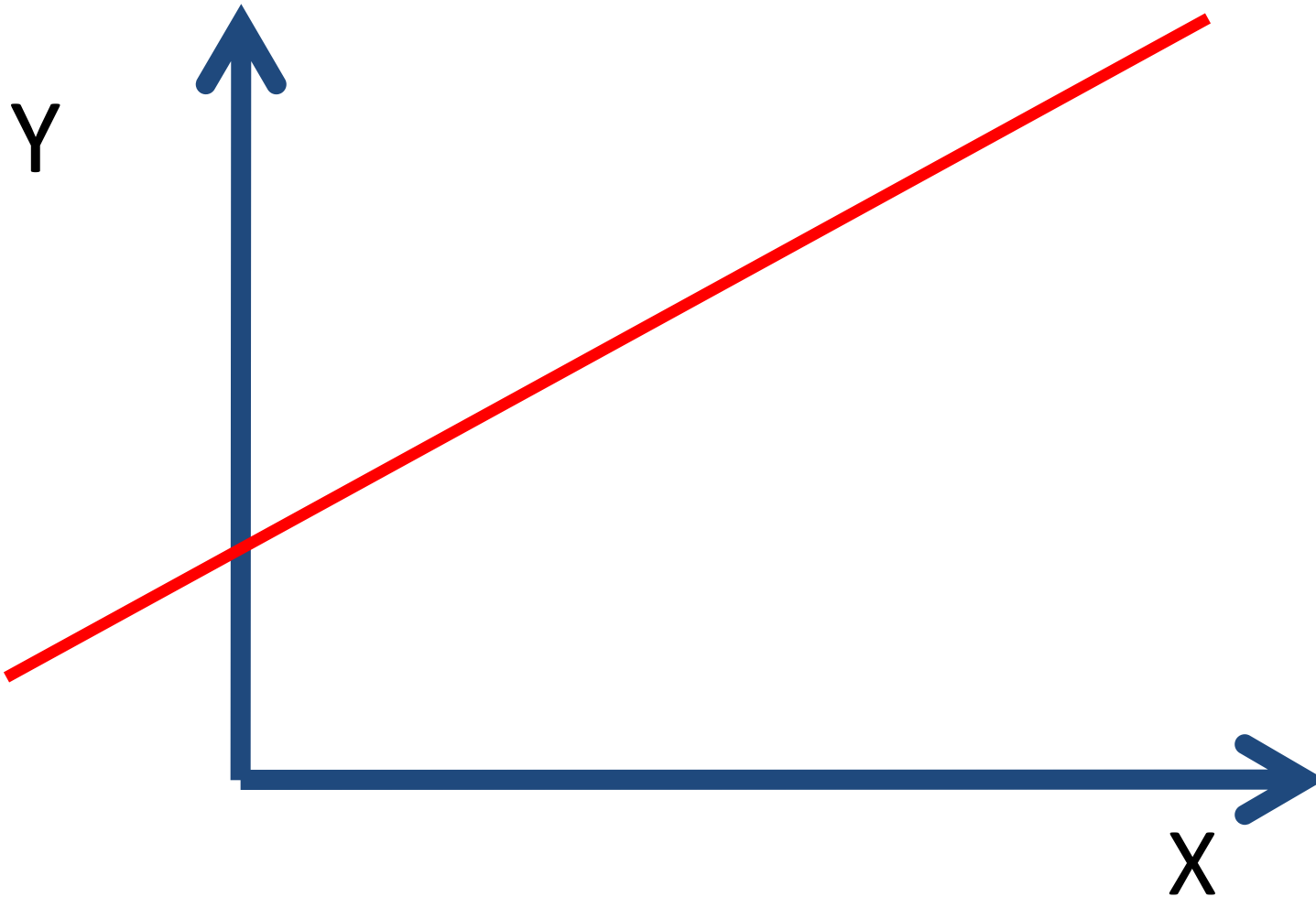
+  DISCRETIONARY READING on Linear Regression 

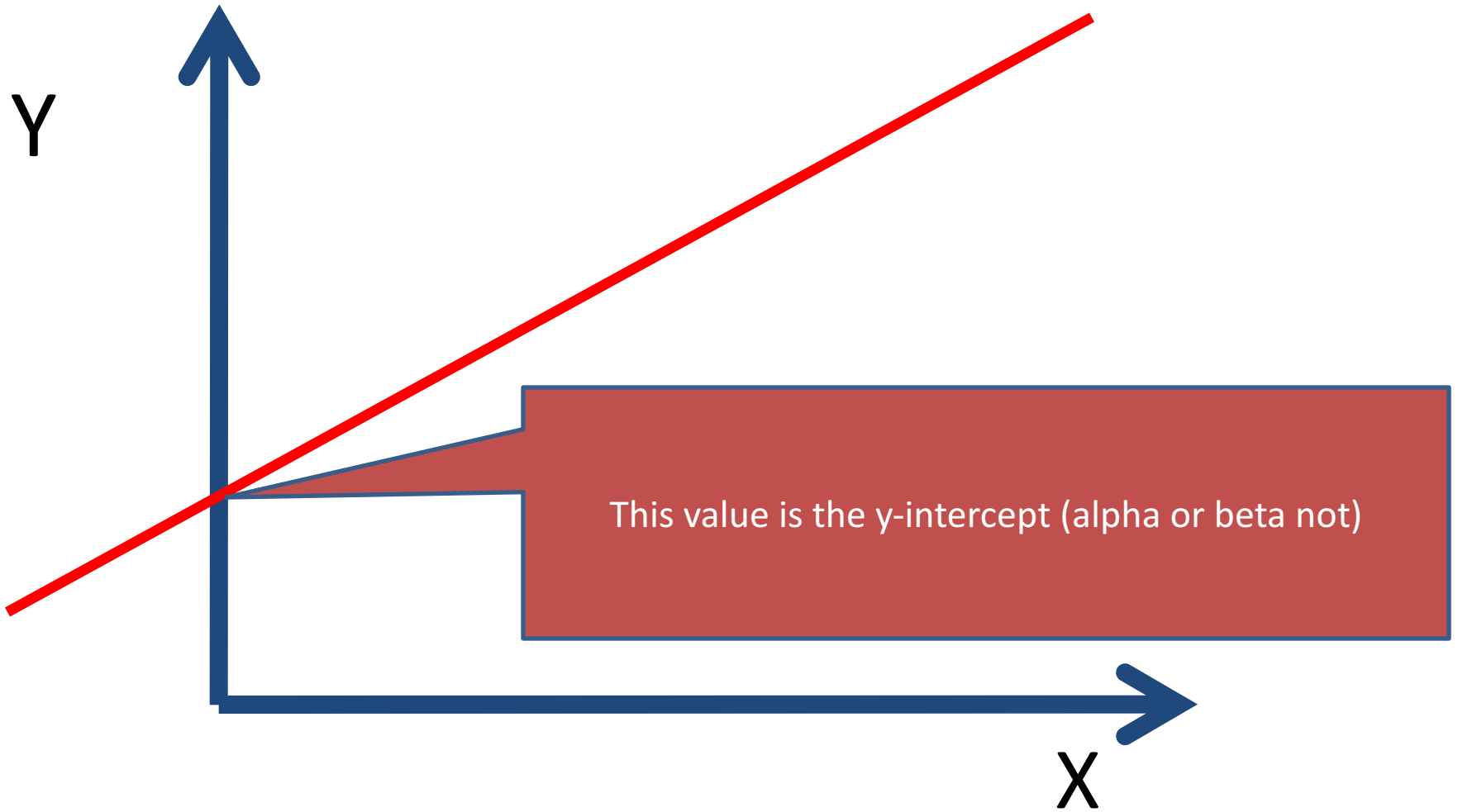
[Edit](#)[+ Add an activity or resource](#)

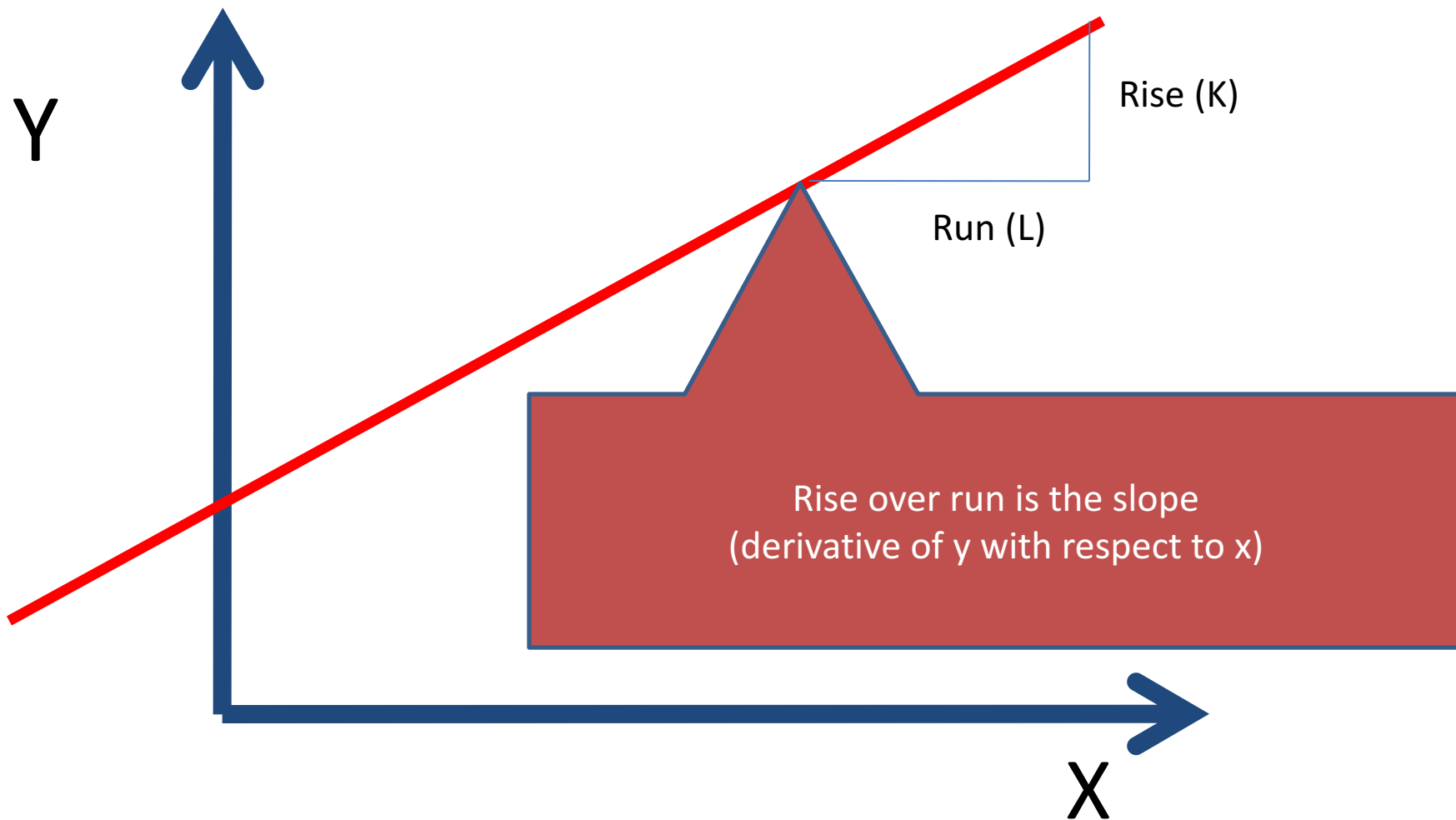
+ January 29 - February 4

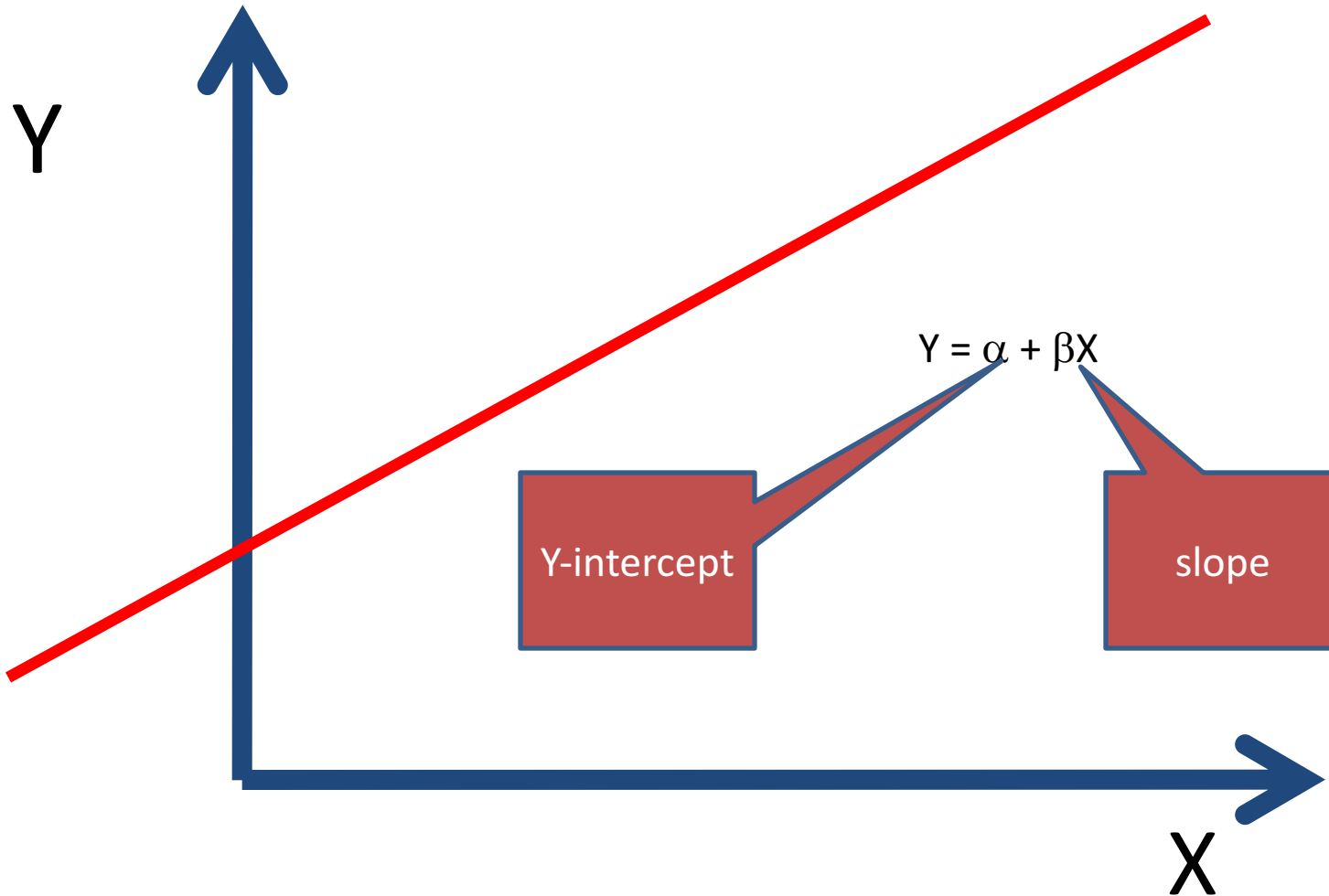
[Edit](#)

The simple model



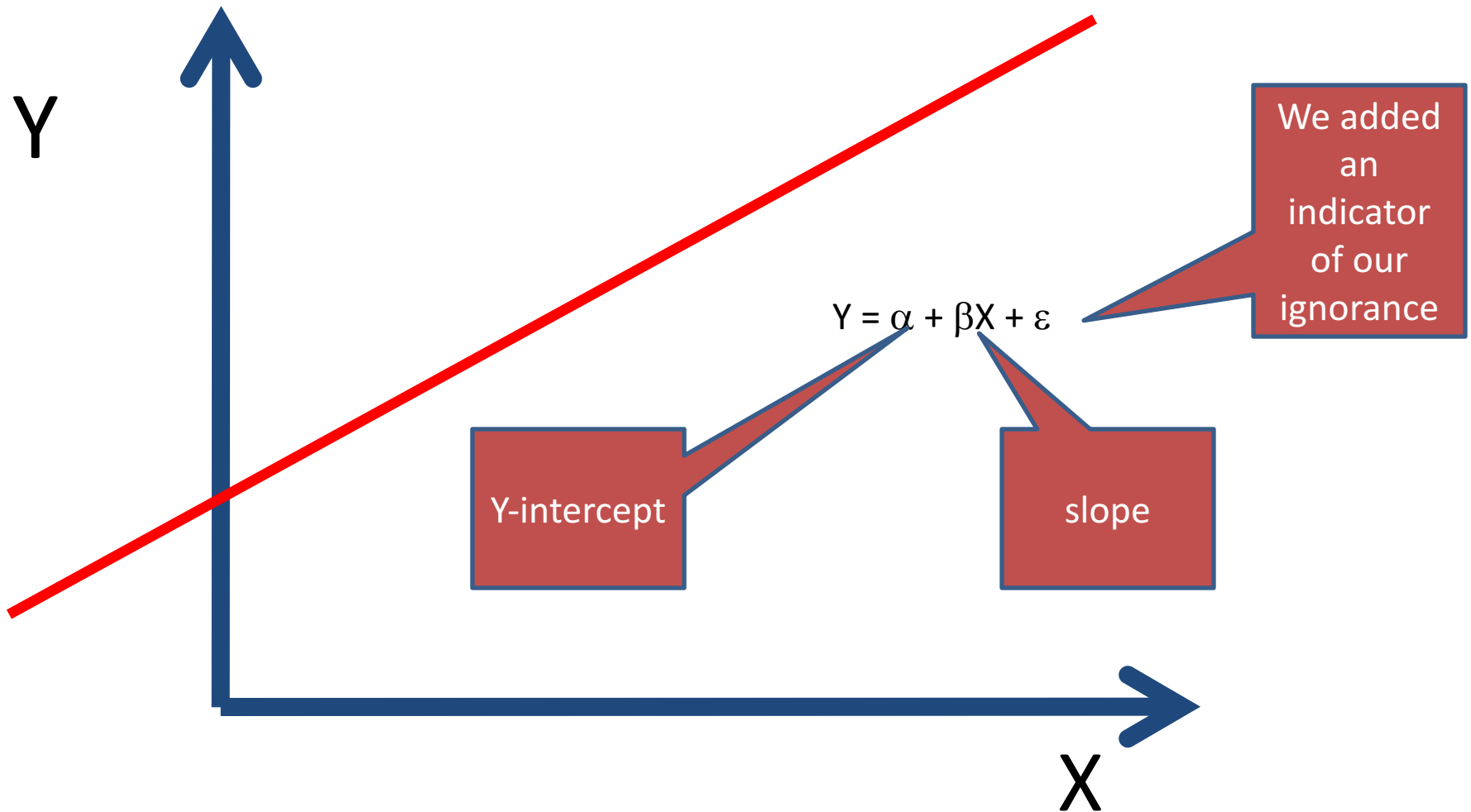




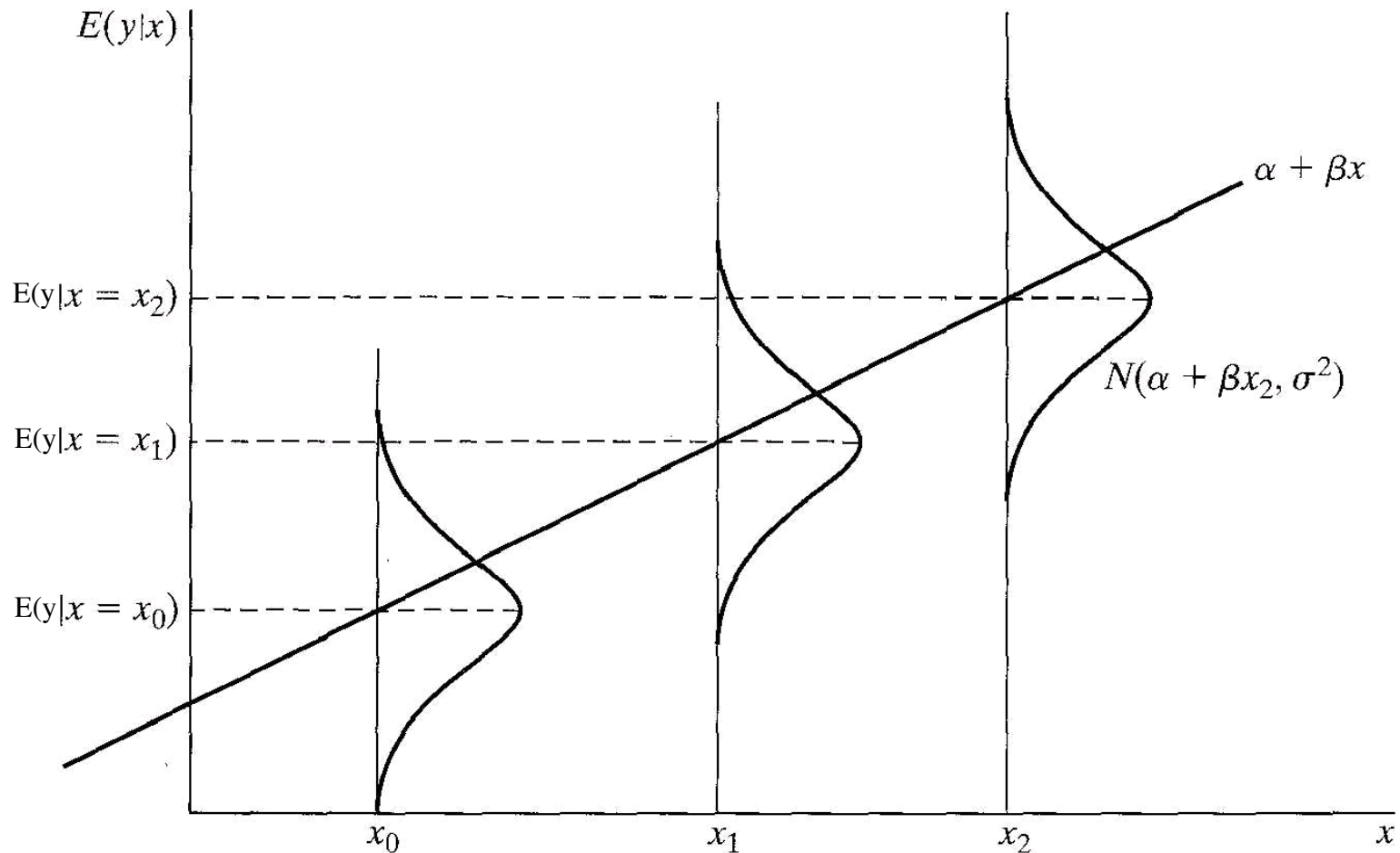


In this example this relationship is deterministic (no allowance for error) = We do not know this precisely and we made mistakes measuring Y and we cannot include all the Xs

From a 2-variable regression



This implies that for each x-value we determine a y-value that is variable
=> there are many observations with the same x but different y values



DIGRESSION ON GREEK

A α	B β	Γ γ	Δ δ	E ε	Z ζ	H η	Θ θ
-----	-----	-----	-----	-----	-----	-----	-----

ἄλφα	βῆτα	γάμμα	δέλτα	ἕψιλόν	ζῆτα	ἦτα	θῆτα
------	------	-------	-------	--------	------	-----	------

alpha	beta	gamma	delta	epsilon	zeta	eta	theta
-------	------	-------	-------	---------	------	-----	-------

a	b	g	d	e	z	ē	th
---	---	---	---	---	---	---	----

[a/a:]	[b]	[g]	[d]	[e]	[zd/dz]	[ε:]	[tʰ]
--------	-----	-----	-----	-----	---------	------	------

I ι	K κ	Λ λ	M μ	N ν	Ξ ξ	O ο	Π π
-----	-----	-----	-----	-----	-----	-----	-----

ἰῶτα	κάππα	λάμβδα	μῦ	νῦ	ξεῖ	ὀμικρόν	πεῖ
------	-------	--------	----	----	-----	---------	-----

iota	kappa	lambda	mu	nu	xi	omikron	pi
------	-------	--------	----	----	----	---------	----

i	k	l	m	n	ks/x	o	p
---	---	---	---	---	------	---	---

[i/i:]	[k]	[l]	[m]	[n]	[ks]	[o]	[p]
--------	-----	-----	-----	-----	------	-----	-----

P ρ	Σ σ/ς	T τ	Υ υ	Φ φ	X χ	Ψ ψ	Ω ω
-----	-------	-----	-----	-----	-----	-----	-----

ῥῶ	σῖγμα	ταῦ	ὔψιλόν	φεῖ	χεῖ	ψεῖ	ὦμέγα
----	-------	-----	--------	-----	-----	-----	-------

rho	sigma	tau	upsilon	phi	chi	psi	omega
-----	-------	-----	---------	-----	-----	-----	-------

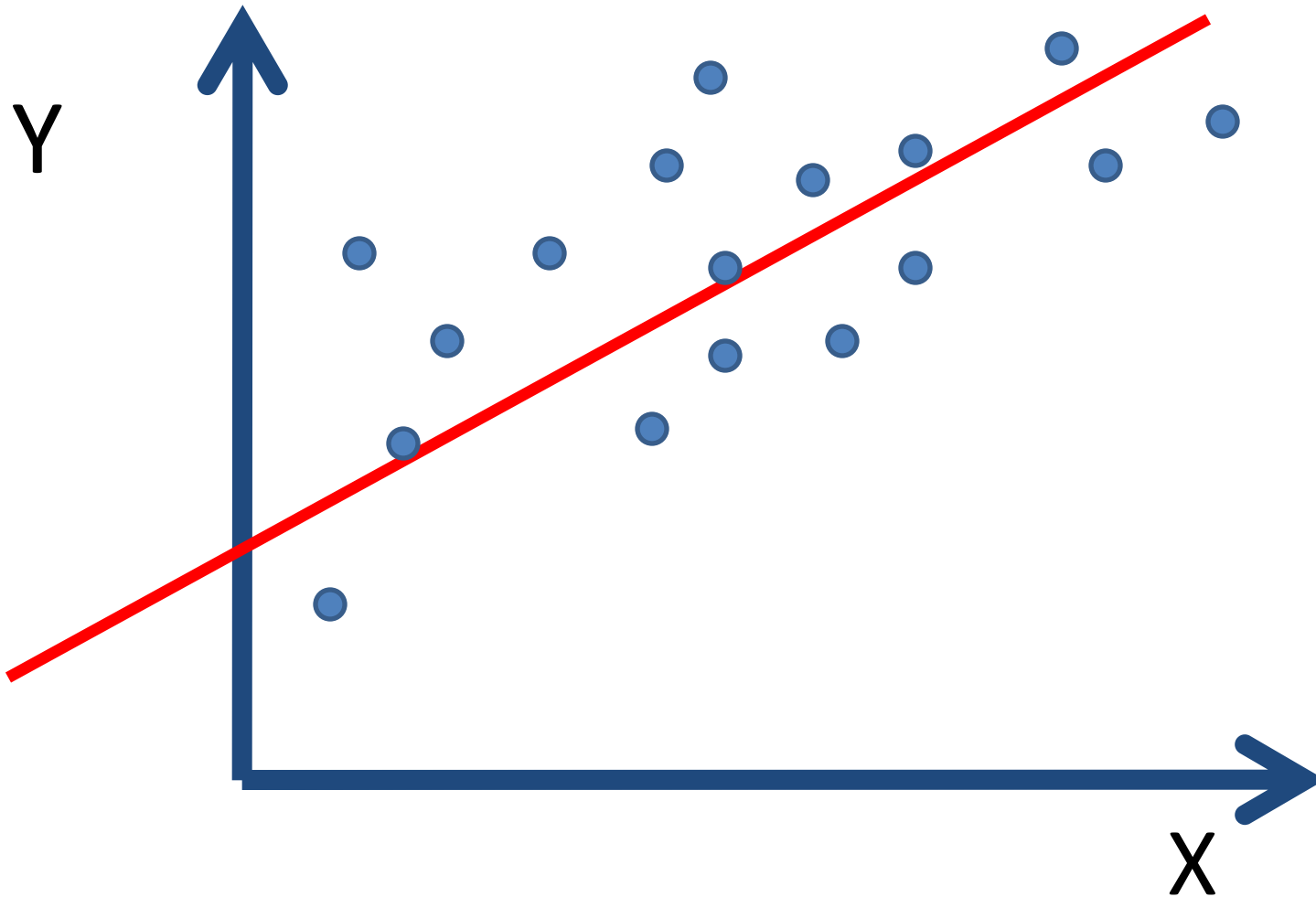
r/rh	s	t	u/y	ph	kh/ch	ps	ō
------	---	---	-----	----	-------	----	---

[r]	[s/z]	[t]	[y/y:]	[pʰ]	[kʰ]	[ps]	[ɔ:]
-----	-------	-----	--------	------	------	------	------

For pronunciation see:
<http://www.historyforkids.net/greek-alphabets.html>

Α α	Β β	Γ γ	Δ δ
Ε ε	Ζ ζ	Η η	Θ θ
Ι ι	Κ κ	Λ λ	Μ μ
Ν ν	Ξ ξ	Ο ο	Π π
Ρ ρ	Σ σς	Τ τ	Υ υ
Φ φ	Χ χ	Ψ ψ	Ω ω

Remember our task in stats is:



Given the data points -> find some line that provides efficient summaries of y-x relationship and then use it to describe behaviors and design things

We can have many xs that explain the variable y

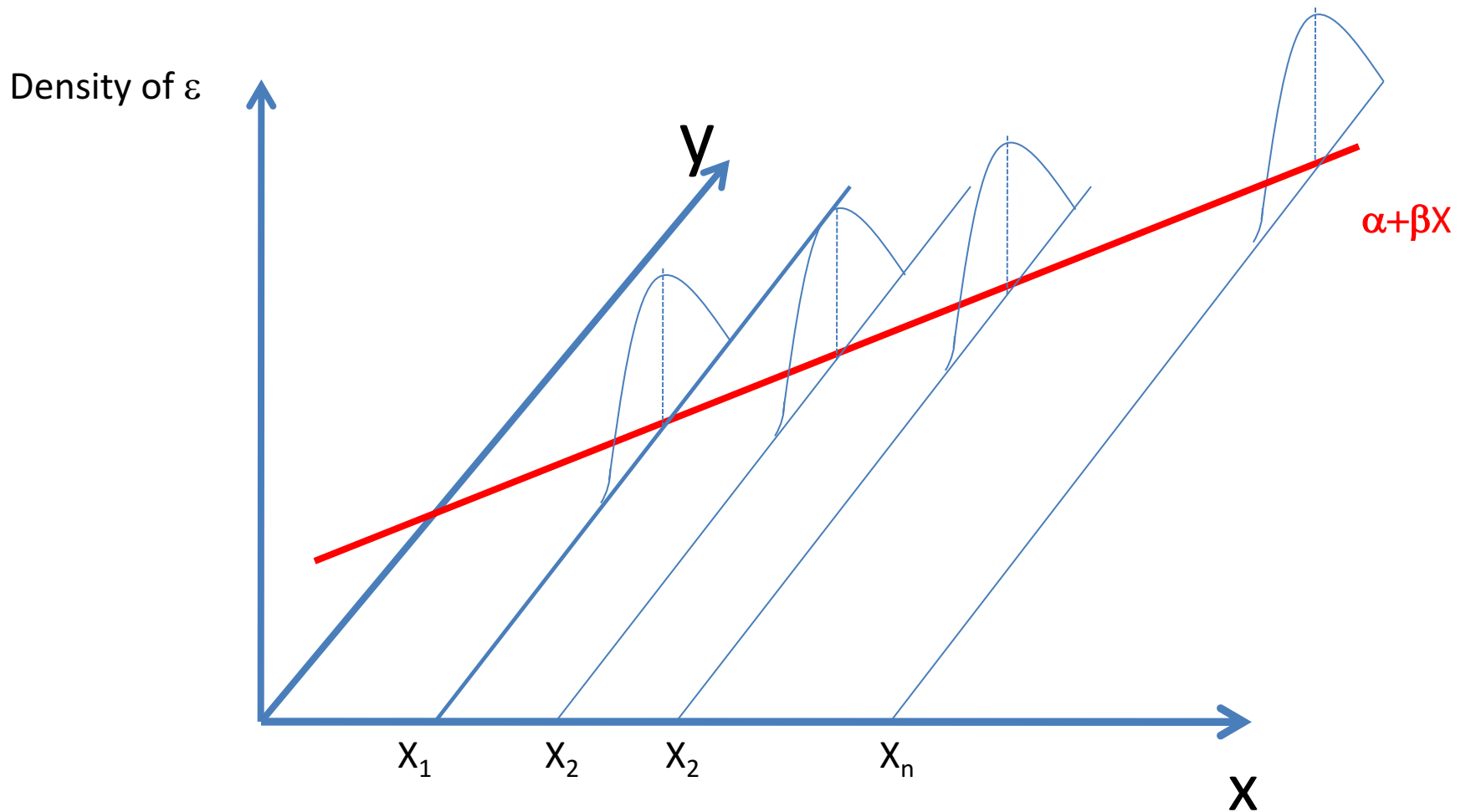
The multiple linear cross-sectional regression model with k independent variables:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

- Notation:**
- 1. Lower case letters for elements and upper case letters for vectors**
 - 2. Greek letters for population values and latin letters or greek with hat for estimates**

Basic Model

- One x variable and one y variable
- We need to make some assumptions for practical reasons
- Usually “assumptions” means we restrict the applicability of the model
- We will relax some of them later and derive different models
- Using a single x variable helps us visualize the model without loss of generality



$$y_i = \alpha + \beta x_i + \varepsilon_i \quad i=1, \dots, n$$

$$E(\varepsilon) = 0$$

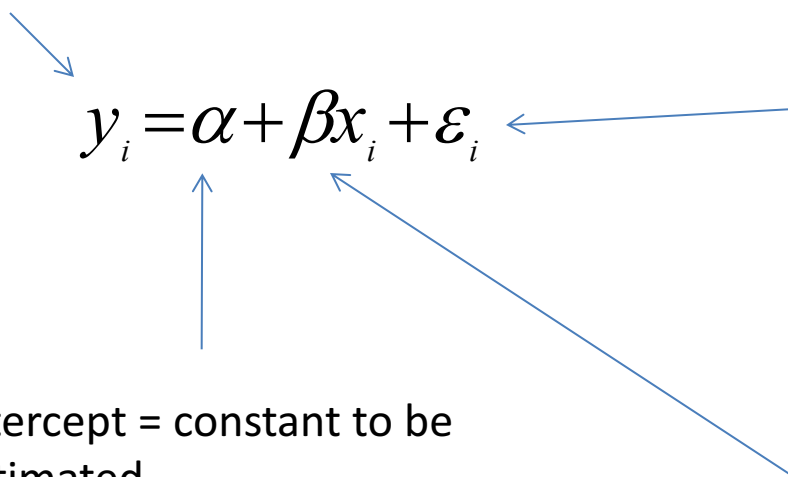
$$Var(\varepsilon) = \sigma_\varepsilon^2$$

The error terms
are independent
& identically
distributed = iid

©Konstadinos Goulas
WE HAVE 3 UNKNOWN PARAMETERS IN THIS MODEL:

$$\alpha, \beta, \sigma_\varepsilon$$

The dependent variable we try to explain


$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Random variable = error term

Intercept = constant to be estimated

Slope = “sensitivity” of y to the values of x

Using an estimation method we try to find estimates of the parameter values for the intercept, slope, and variance of the error term using the sample data

We also can test hypotheses about some of the model assumptions

$$E(y_i) = E(\alpha + \beta x_i + \varepsilon_i) = E(\alpha) + E(\beta x_i) + E(\varepsilon_i) = \alpha + \beta x_i$$

Terminology

- Dependent variable (Y)
- Independent variables (X)
- Explanatory variable (X)
- Disturbance (ε)
- Random error term (ε)
- Coefficients (β, α)
- Parameters to estimate (β, α)
- Significance (alpha, 1- alpha)
- Goodness-of-fit measure (R-square)

In Scalar Notation

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

$$y_1 = \alpha + \beta x_1 + \varepsilon_1$$

$$y_2 = \alpha + \beta x_2 + \varepsilon_2$$

$$y_3 = \alpha + \beta x_3 + \varepsilon_3$$

.

.

.

$$y_n = \alpha + \beta x_n + \varepsilon_n$$

©Konstadinos Goulias

In Matrix Notation

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} \alpha + \beta x_1 \\ \alpha + \beta x_2 \\ \cdot \\ \cdot \\ \cdot \\ \alpha + \beta x_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

n by 1 n by 2 2 by 1 n by 1

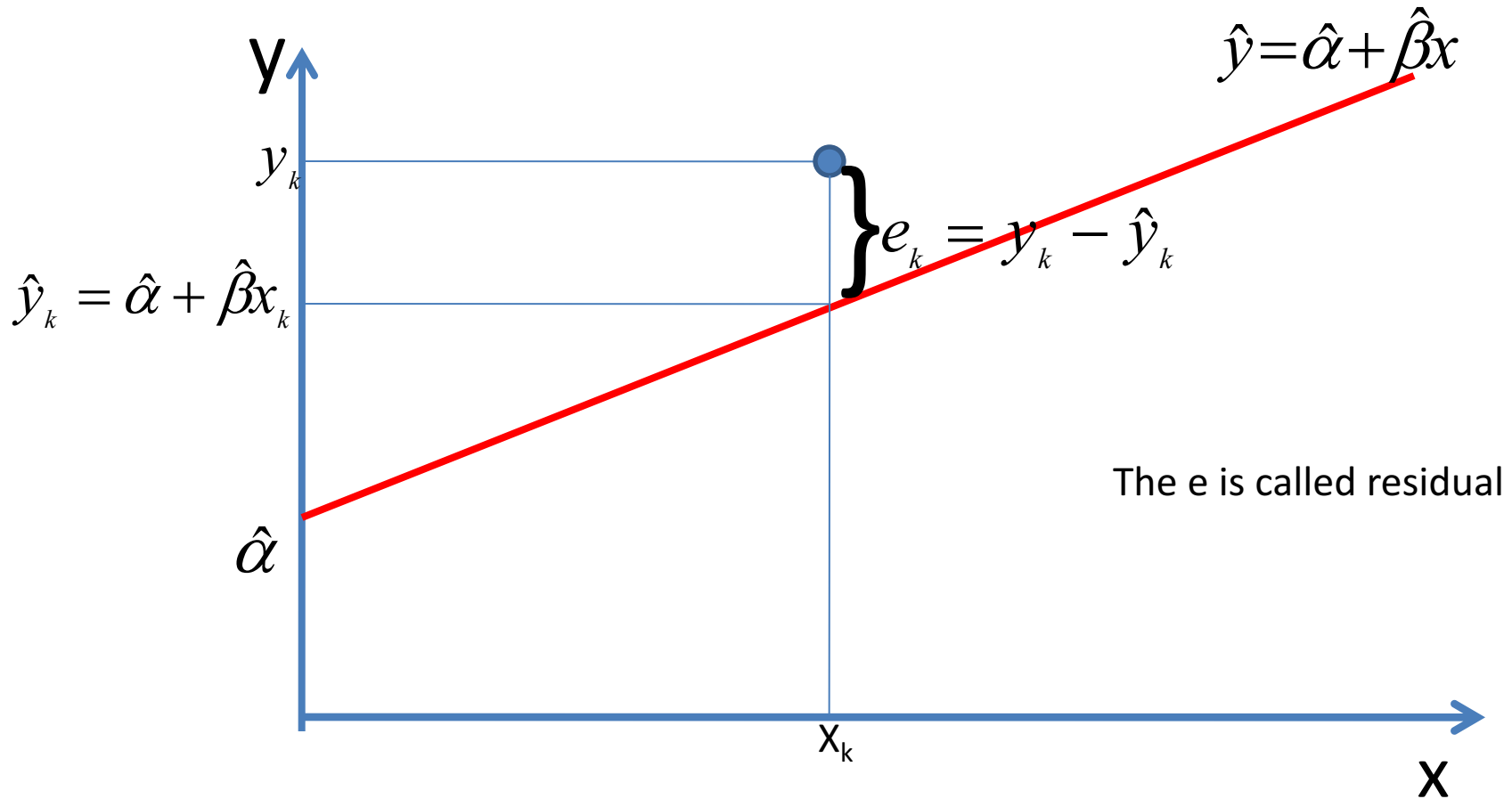
Simple Example 1

Observed data

x_i	2	3	1	5	9
y_i	4	7	3	9	17

Linear regression model

$$\text{Mean response} = E(y) = \alpha + \beta x$$



For any k in the sample

We sum all those squared distances e and minimize their sum over the sample to find the regression coefficients

Least Squares Estimation

$$e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}x_i$$

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

$$\sum_{i=1}^n e_i^2 = f(\hat{\alpha}, \hat{\beta})$$

Select $\hat{\alpha}$ and $\hat{\beta}$ to minimize $\sum_{i=1}^n e_i^2$

This estimator has many convenient and desirable properties

How to Solve

Select $\hat{\alpha}$ and $\hat{\beta}$ to minimize $\sum_{i=1}^n e_i^2$

$$\frac{\partial \sum_{i=1}^n e_i^2}{\partial \hat{\alpha}} = 0$$

2 equations and 2
unknowns

$$\frac{\partial \sum_{i=1}^n e_i^2}{\partial \hat{\beta}} = 0$$

Solution

$$\frac{\partial \sum_{i=1}^n e_i^2}{\partial \hat{\alpha}} = 0$$

$$\sum_{i=1}^n y_i = n\hat{\alpha} + \hat{\beta} \sum_{i=1}^n x_i$$

Interesting fact: if you divide this by n what do you get?

$$\frac{\partial \sum_{i=1}^n e_i^2}{\partial \hat{\beta}} = 0$$

$$\sum_{i=1}^n x_i y_i = \hat{\alpha} \sum_{i=1}^n x_i + \hat{\beta} \sum_{i=1}^n x_i^2$$

These are called the **Normal Equations** for the linear regression line

Numerical example

ID	X	Y	XY	X	Y hat	e=Y-Yhat
1	2	4	8	4	4.5	-0.5
2	3	7	21	9	6.25	0.75
3	1	3	3	1	2.75	0.25
4	5	9	45	25	9.75	-0.75
5	9	17	153	81	16.75	0.25
Sums	20	40	230	120	40	0
Average	4	8			8	0

Note to me:
Explain all items!

$$\sum_{i=1}^n y_i = n\hat{\alpha} + \hat{\beta} \sum_{i=1}^n x_i \Rightarrow 40 = 5\hat{\alpha} + \hat{\beta}20$$

$$\sum_{i=1}^n x_i y_i = \hat{\alpha} \sum_{i=1}^n x_i + \hat{\beta} \sum_{i=1}^n x_i^2 \Rightarrow 230 = 20\hat{\alpha} + 120\hat{\beta}$$

$$\text{Solution of this} \Rightarrow \hat{\alpha} = 1, \quad \hat{\beta} = 1.75$$

$$\hat{y} = 1 + 1.75x$$

- This was the Ordinary Least Squares (OLS) estimation
- There are other estimation methods to find the regression coefficients!
- Remember this is called “least squares” because we minimize the sum of squared residuals to find the values of the coefficients

EXAMPLE FROM THE HOUSEHOLD FILE SMALLHHFILE

Descriptive statistics

Statistic	N	Mean	St. Dev.	Min	Median	Max
SAMPN	42,431	2,588,379.00	1,641,345.00	1,031,985	1,971,814	7,212,388
INCOM	42,431	13.18	26.29	1	5	99
HHSIZ	42,431	2.57	1.37	1	2	8
HHEMP	42,431	1.22	0.88	0	1	6
HHSTU	42,431	0.64	1.02	0	0	8
HHLIC	42,431	1.86	0.85	0	2	8
DOW	42,431	4.02	1.99	1	4	7
HTRIPS	42,431	8.29	7.78	0	6	99
Mon	42,431	0.14	0.34	0	0	1
Tue	42,431	0.14	0.35	0	0	1
Wed	42,431	0.14	0.35	0	0	1
Thu	42,431	0.15	0.35	0	0	1
Fri	42,431	0.14	0.35	0	0	1
Sat	42,431	0.14	0.35	0	0	1
Sun	42,431	0.15	0.35	0	0	1
TotDist	42,431	68.09	118.52	0.00	33.89	5,838.26
center	42,431	0.28	0.45	0	0	1
suburb	42,431	0.29	0.45	0	0	1
exurb	42,431	0.23	0.42	0	0	1
rural	42,431	0.20	0.40	0	0	1
other	42,431	0.00	0.00	0	0	0
highinc	42,431	0.41	0.49	0	0	1
HHVEH	42,431	1.86	1.00	0	2	8
HHBIC	42,431	1.58	3.79	0	1	99
VEHNEW	42,431	2.15	2.02	1	2	9
OWN	42,431	1.24	0.56	1	1	9
CarBuy	42,431	0.45	0.50	0	0	1
snglhm	42,431	0.82	0.39	0	1	1
ownhm	42,431	0.77	0.42	0	1	1
MilesPr	42,431	27.12	43.46	0.00	14.50	1,167.65
TrpPrs	42,431	3.28	2.58	0.00	3.00	32.00

Discussion:

In Linear Regression

- a) Parameters to estimate ?
- b) What are the data given to us?
- c) What is considered random?

In R

- `lm` is used to fit linear models.
- `lm(formula, data, subset, weights, na.action, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, ...)`
- `HHPMT.lm = lm(TotDist ~ HHSIZ , data=SmallHHfile)`
 - This asks to run a regression model and create the object `HHPMT.lm`
 - (I can use any name I want)

output

```
> HHPMT.lm = lm(TotDist ~ HHSIZ , data=SmallHHfile)
> summary(HHPMT.lm)
```

Call:

```
lm(formula = TotDist ~ HHSIZ, data = SmallHHfile)
```

Residuals:

Min	1Q	Median	3Q	Max
-186.1	-49.1	-26.5	11.4	5717.4

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	12.2118	1.1817	10.33	<2e-16 ***
HHSIZ	21.7305	0.4053	53.62	<2e-16 ***

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

Residual standard error: 114.7 on 42429 degrees of freedom

Multiple R-squared: 0.06345, Adjusted R-squared: 0.06343

F-statistic: 2875 on 1 and 42429 DF, p-value: < 2.2e-16



```

> HHPMT.lm = lm(TotDist ~ HHSIZ , data=SmallHHfile)
> summary(HHPMT.lm)

Call:
lm(formula = TotDist ~ HHSIZ, data = SmallHHfile)

Residuals:
    Min       1Q   Median       3Q      Max
-186.1   -49.1   -26.5    11.4   5717.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.2118     1.1817   10.33  <2e-16 ***
HHSIZ        21.7305     0.4053   53.62  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 114.7 on 42429 degrees of freedom
Multiple R-squared:  0.06345,    Adjusted R-squared:  0.06343
F-statistic: 2875 on 1 and 42429 DF,  p-value: < 2.2e-16

```

Daily miles traveled by a household = $12.2118 + 21.7305 \text{ HHSIZ}$

What happens when the household is a single person?

What happens when the household is a couple?

Derivative of the daily miles traveled by a household with respect to household size is the coefficient $\beta_{\text{hat}} = 21.7305$

Goodness of fit (how well our model replicates the data we use) is checked using indicators.

The most popular is called coefficient of determination or R-squared

Total Variation in the y variable is:

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

**Total Sum of Squares (SST) = Regression Sum of Squares (SSR*)
+ Error Sum of Squares (SSE)**

Regression Sum of Squares = variation we capture with xs and bs

*Note I am using different ss etc here because in literature you will find them both ways
©Konstadinos Goulas

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Percent of variation we explain by exercising our model

R^2 Takes values between 0 and 1

Usually with large samples = lower values (harder to explain variation)

Small sample = higher values

Goodness of fit and sum of squares

```
> anova(HHPMT.lm) # anova table
```

```
Analysis of Variance Table
```

```
Response: TotDist
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
HHSIZ	1	37815166	37815166	2874.6	< 2.2e-16 ***
Residuals	42429	558154359	13155		

```
---
```

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

```
> |
```

R-square = $37815166 / (37815166 + 558154359) = 0.063$ (I explain only 6.3% of total variation!)

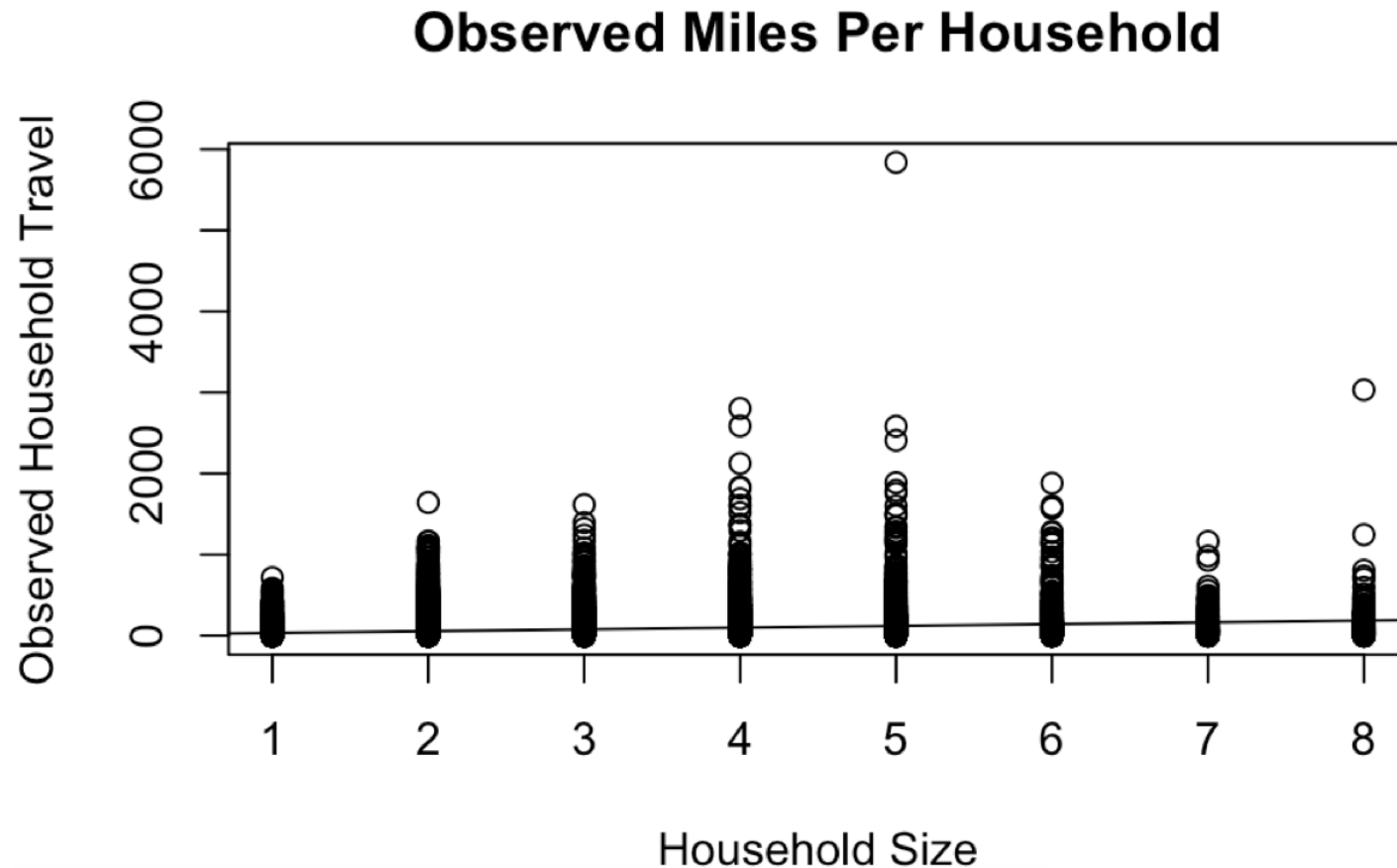
The more Sum of Squares our explanatory variables capture -> the higher R-square becomes

Used stargazer package to create a summary table

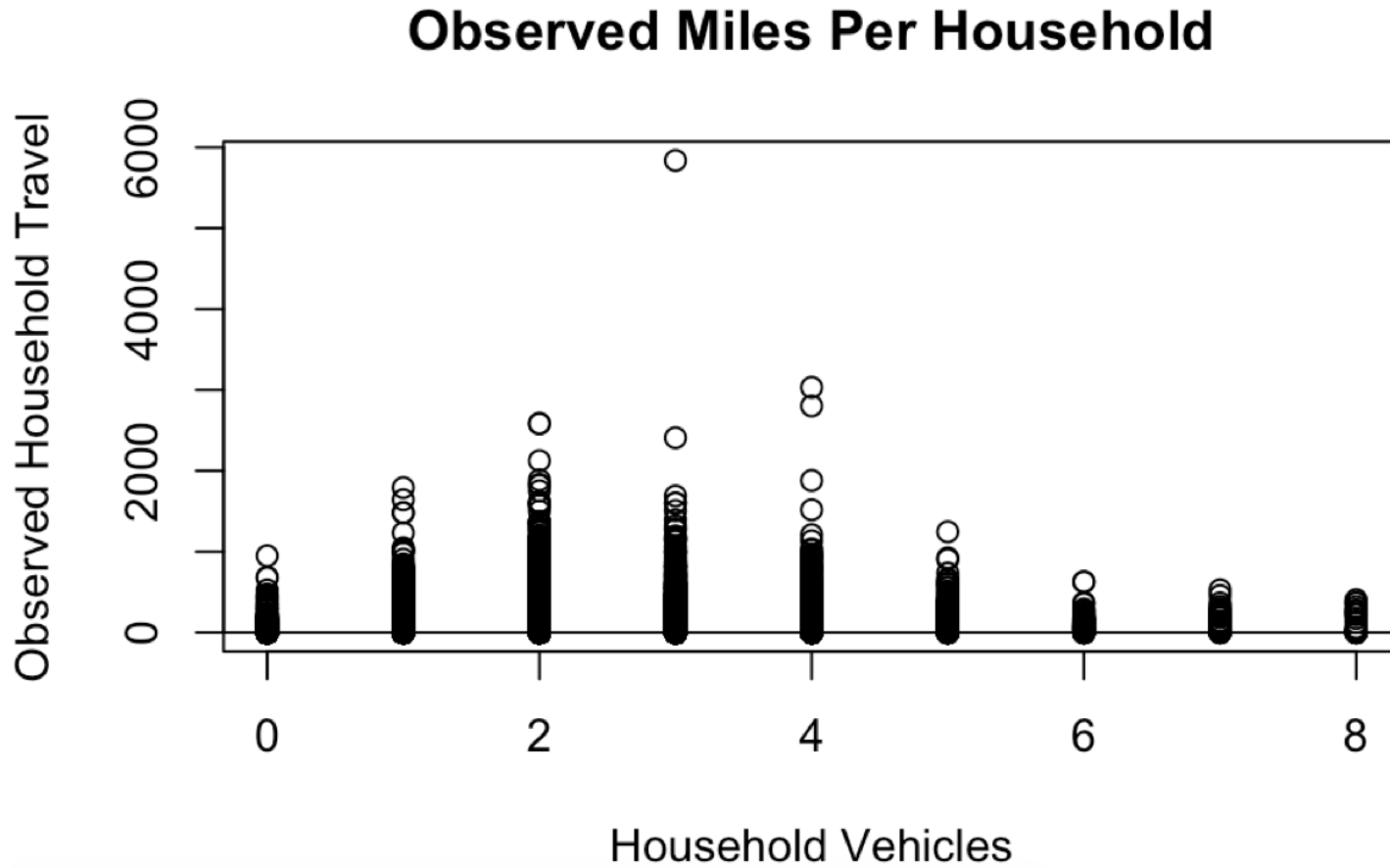
Regression Results

Dependent variable:	
Total Distance per Household	
Household Size	21.730*** (0.405)
Constant	12.212*** (1.182)
Observations	42,431
R2	0.063
Adjusted R2	0.063
Residual Std. Error	114.695 (df = 42429)
F Statistic	2,874.581*** (df = 1; 42429)
Note:	*p<0.1; **p<0.05; ***p<0.01

Observed vs Model & X variable

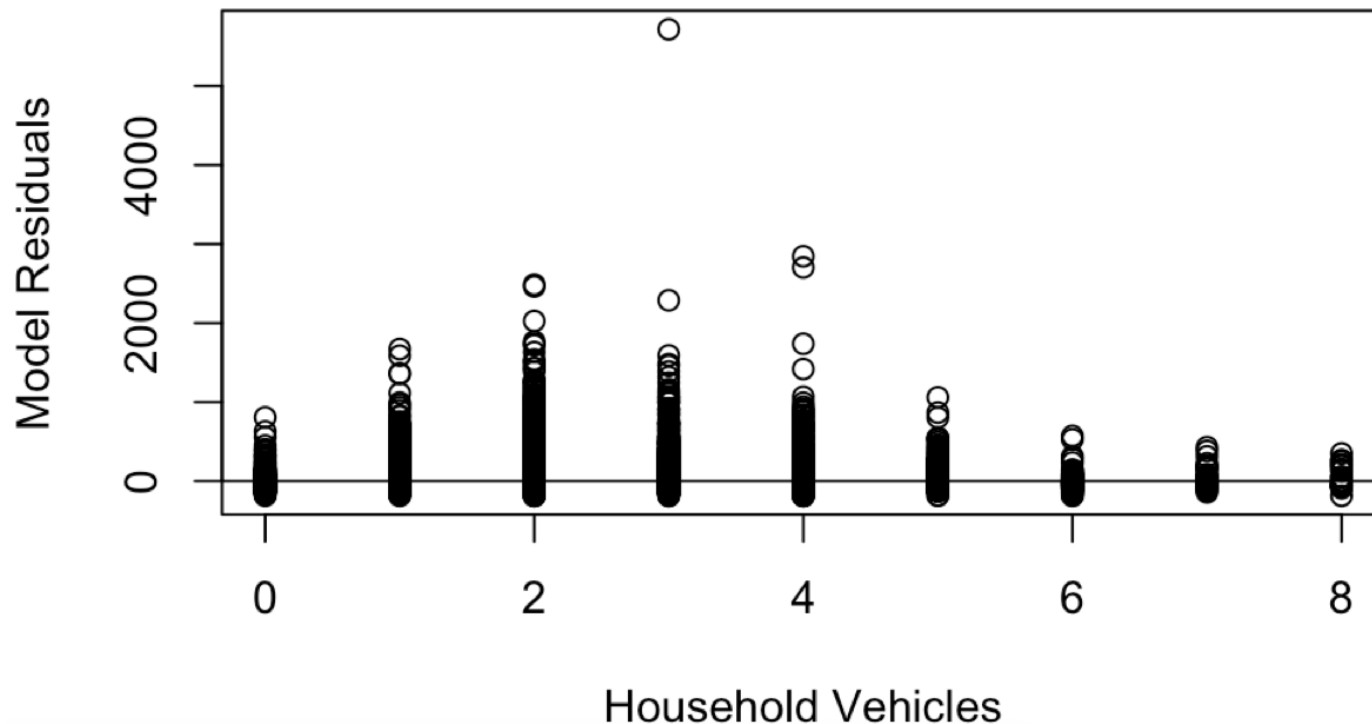


The effect of the excluded variable household vehicles?



Whatever I exclude and is important ends up
“inflating” residuals = misspecification

Is there a relationship?



Basic Concepts

DIGRESSION ON STATISTICAL INFERENCE

Statistical Inference & Estimation

- Deals with methods for making statements about a population based on a sample drawn from the population
 - **Point Estimation:** Find an unknown population parameter using a single statistic calculated from the sample data.
 - **Confidence Interval Estimation:** Find an interval from sample data that includes the unknown population parameter with a *pre-assigned* probability
 - **Hypothesis testing:** Test the potential of a value assigned to a statistic based on some risk acceptance

Textbook Definition

- Statistical estimator is a FUNCTION of the n observed values, x_1, \dots, x_n , in the sample from a random variable x .
- The estimator is also a random variable!
- The value of the estimator is theta hat $\hat{\theta}$
- It has its own distribution and it can be calculated
- The population parameter corresponding to this estimator is theta

Difference among the three

- Point Estimation: estimate the mean income of the population in Santa Barbara
- Confidence Interval Estimation: Find an interval [Lower, Upper] based on the data that includes the mean income of SB with a specified probability
- Hypothesis testing: Is the mean income in SB higher than \$67,000? Is the mean income equal to something?

Example & Definitions

- Estimator = the random variable $\hat{\theta}$ which is a function of the data we collect = the formula and/or the rule to be computed from the data
- Estimate = the specific numerical value of $\hat{\theta}$ calculated from the observed sample data
- Example: $X_i \sim N(\mu, \sigma^2)$
- Estimator = $\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$ is an **estimator** of μ
- Estimate = value = 107.3 is an **estimate** of μ

Methods of Evaluating Estimators

Bias and Variance

$$\text{Bias}(\hat{\theta}) = \text{Mean error} = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$$

- This bias measures the **accuracy** of an estimator.
- An estimator whose bias is zero is called **unbiased**.
- An unbiased estimator may, nevertheless, fluctuate greatly from sample to sample.

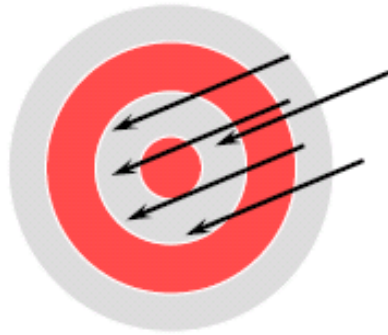
$$\text{Var}(\hat{\theta}) = E\{[\hat{\theta} - E(\hat{\theta})]^2\}$$

- Lower variance = more **precise** the estimator
- Some low-variance estimators are biased (precise but inaccurate)
- Among unbiased estimators, the one with the lowest variance should be chosen
- Efficient means low variance
- “Best” = minimum variance.

Accuracy and Precision



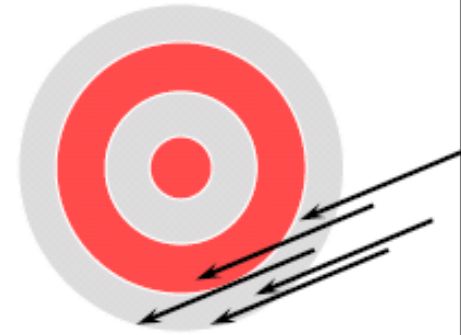
accurate and precise



accurate,
not precise



precise,
not accurate



not accurate,
not precise

Zero bias and low
variance

Zero bias and high
variance

Non zero bias and
low variance

Non zero bias and
high variance

Best estimator

Worst estimator

Mean Squared Error (MSE)

- To select among all estimators (biased and unbiased), minimize a measure that combines both **bias** and **variance**.
- A “good” estimator should have low bias (accurate) AND low variance (precise).

$$MSE(\hat{\beta}) = E(\hat{\beta} - \beta)^2 = Bias(\hat{\beta})^2 + Var(\hat{\beta})$$

- Minimizing this attempts to minimize bias and variance

Root MSE (RMSE) is:

$$\sqrt{MSE(\hat{\beta})}$$

Consistency

- Definition: probability limit of estimated β is equal to true β as n approaches infinity. The probability of the absolute value of the difference between the estimate and true value will be less than an arbitrarily selected small positive number will approach 1.

$$\lim_{n \rightarrow \infty} \text{Prob}(|\beta - \hat{\beta}| < \delta) = 1$$

Consistency Criterion

- An estimate is a consistent estimator of a true value if the:
-

$$P \lim \hat{\beta} = \beta$$

An estimator with a MSE that approaches zero as the sample size increases is also a consistent estimator – the reverse may not be true

Degrees of Freedom

- Estimates of statistical parameters are based on different amounts of information (the data)
- The number of independent pieces of information that go into the estimate of a parameter is called the degrees of freedom (df).
- $\text{df of an estimate} = (\text{number of independent scores that go into the estimate}) - (\text{the number of parameters estimated})$

Example: estimators of variance

Two estimators of variance:

$$S_1^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{(n-1)}$$

is unbiased

We used one degree
of freedom
calculating \bar{x}

$$S_2^2 = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n}$$

is biased but has smaller MSE

In spite of larger MSE, we almost always use the one with $n-1$

As $n \rightarrow \infty$ the two coincide in value

Standard Error (SE)

- The standard deviation of an estimator is called the standard error of the estimator (SE).
- The estimated standard error is also called standard error (Confusing!).
- The precision of an estimator is measured by the SE.

1. \bar{X} is an unbiased estimator of μ

$$\left. \begin{aligned} SE(\bar{X}) &= \sigma / \sqrt{n} \\ se(\bar{X}) &= s / \sqrt{n} \end{aligned} \right\}$$

are called the standard error of the mean

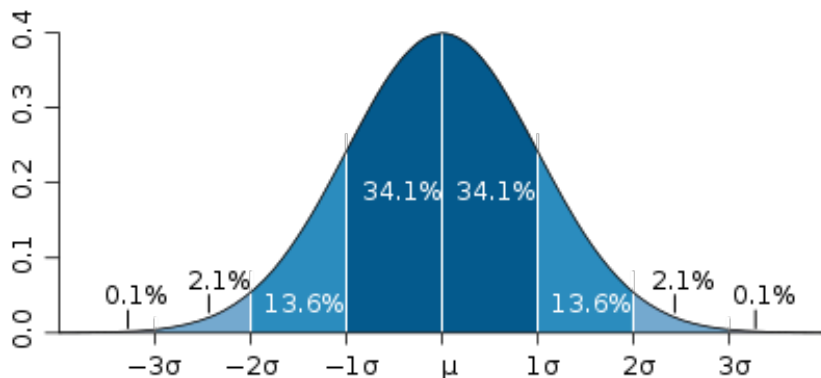
Precision and Standard Error

- A precise estimate has a small standard error, and the **precision** and **standard error** are related.
- If the sampling distribution of an estimator is normal with mean equal to the true parameter value (i.e., unbiased), about 95% of the time the estimator will be within two SE's from the true parameter value.

CONFIDENCE INTERVALS

An example

- Suppose we are interested in the population mean μ
- The population distribution is normal
- The value of the population standard deviation is known and σ
- The sample observations x_1, \dots, x_n are from this normal distribution
- The sample mean is $N(\mu, \sigma/\sqrt{n})$
- Let's work with the standard normal:



$$P(-1.96 < \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < 1.96) = 0.95$$

Let's manipulate this equation

$$P(-1.96 < \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} < 1.96) = 0.95$$

$$P(-1.96 \sigma / \sqrt{n} < \bar{X} - \mu < 1.96 \sigma / \sqrt{n}) = 0.95$$

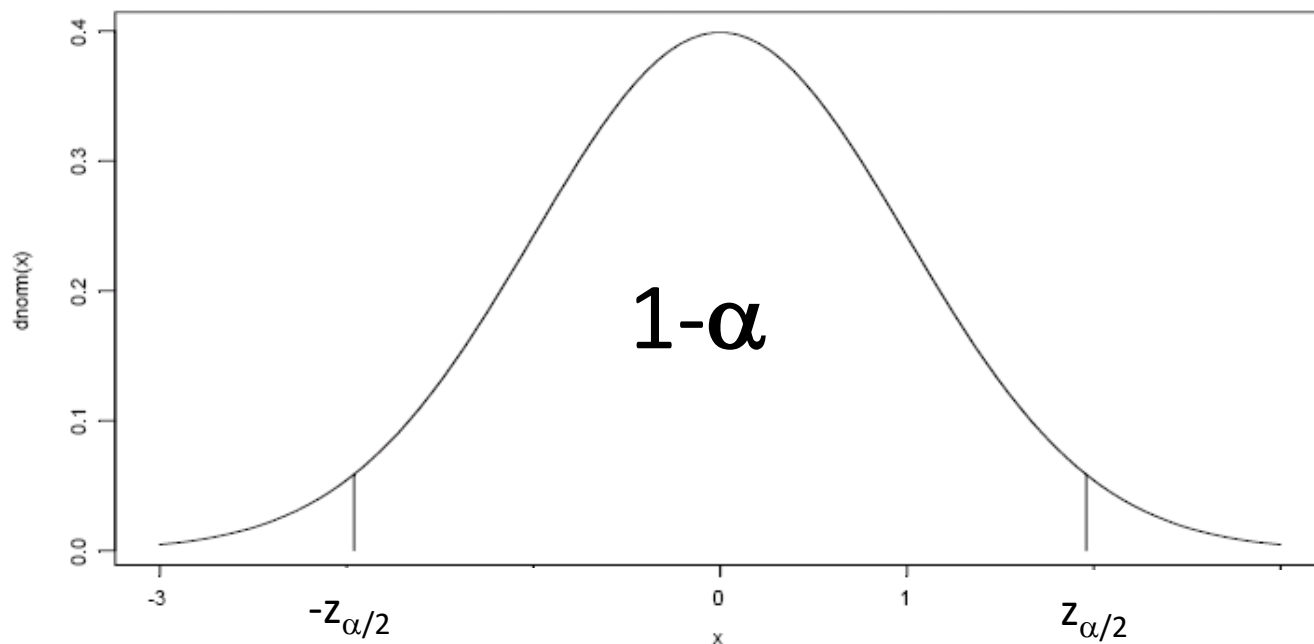
$$P(\bar{X} - 1.96 \sigma / \sqrt{n} < \mu < \bar{X} + 1.96 \sigma / \sqrt{n}) = 0.95$$

Given the observations and the population σ is given to us,

We can say with 95% probability the true population mean μ is between the lower and the upper values







Confidence Interval

$$(\bar{X} - 1.96 \sigma / \sqrt{n}, \bar{X} + 1.96 \sigma / \sqrt{n})$$



$$(\bar{X} - z_{\alpha/2} \sigma / \sqrt{n}, \bar{X} + z_{\alpha/2} \sigma / \sqrt{n})$$

TABLE A-3
Standard Normal Probabilities

z	 $P(-z < Z < z)$	 $P(Z > z)$	 $P(Z > z)$	 $P(Z < -z)$	 $P(Z < z)$	 $P(Z > -z)$
0.50	0.383	0.617	0.309	0.309	0.691	0.691
0.60	0.451	0.549	0.274	0.274	0.726	0.726
0.70	0.516	0.484	0.242	0.242	0.760	0.758
0.80	0.576	0.424	0.212	0.212	0.788	0.788
0.90	0.632	0.368	0.184	0.184	0.816	0.816
1.00	0.683	0.317	0.159	0.159	0.841	0.841
1.28	0.800	0.200	0.100	0.100	0.900	0.900
1.50	0.866	0.134	0.067	0.067	0.933	0.933
1.60	0.890	0.110	0.055	0.055	0.945	0.945
1.65	0.900	0.100	0.050	0.050	0.950	0.950
1.70	0.911	0.089	0.045	0.045	0.955	0.955
1.80	0.928	0.072	0.036	0.036	0.964	0.964
1.90	0.943	0.057	0.029	0.029	0.971	0.971
1.96	0.950	0.050	0.025	0.025	0.975	0.975
2.00	0.954	0.046	0.023	0.023	0.977	0.977
2.10	0.964	0.036	0.018	0.018	0.982	0.982
2.20	0.972	0.028	0.014	0.014	0.986	0.986
2.30	0.979	0.021	0.011	0.011	0.989	0.989
2.40	0.984	0.016	0.008	0.008	0.992	0.992
2.50	0.988	0.012	0.006	0.006	0.994	0.994
2.58	0.990	0.010	0.005	0.005	0.995	0.995
2.60	0.991	0.009	0.005	0.005	0.995	0.995
2.70	0.993	0.007	0.003	0.003	0.997	0.997
2.80	0.995	0.005	0.003	0.003	0.997	0.997
2.90	0.996	0.004	0.002	0.002	0.998	0.998
3.00	0.997	0.003	0.001	0.001	0.999	0.999
3.10	0.998	0.002	0.001	0.001	0.999	0.999
3.20	0.999	0.001	0.001	0.001	0.999	0.999
3.30	0.999	0.001	0.000	0.000	1.000	1.000
3.40	0.999	0.001	0.000	0.000	1.000	1.000
3.50	1.000	0.000	0.000	0.000	1.000	1.000

What is the value of $1-\alpha$ when $z = 1.96$?

What is the value of $1-\alpha$ when $z = 3.5$?

Frequentist Interpretation of Confidence Intervals

In an infinitely long series of trials in which repeated samples of size n are drawn from the same population and 95% CI's for mean are calculated using the same method, the proportion of intervals that actually include μ will be 95% (coverage probability). **BELOW are 50 trials: find out in how many samples we missed including 0? I would expect in 100 to miss how many?**

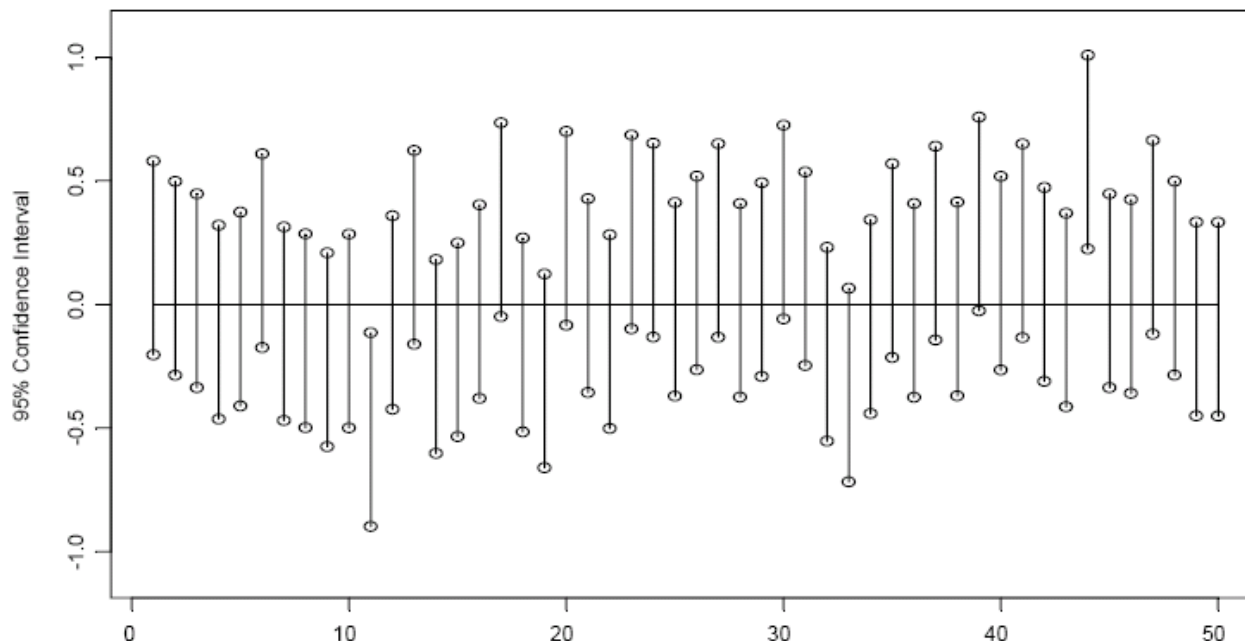








TABLE A-3
Standard Normal Probabilities

z	 $P(-z < Z < z)$	 $P(Z > z)$	 $P(Z > z)$	 $P(Z < -z)$	 $P(Z < z)$	 $P(Z > -z)$
0.50	0.383	0.617	0.309	0.309	0.691	0.691
0.60	0.451	0.549	0.274	0.274	0.726	0.726
0.70	0.516	0.484	0.242	0.242	0.760	0.758
0.80	0.576	0.424	0.212	0.212	0.788	0.788
0.90	0.632	0.368	0.184	0.184	0.816	0.816
1.00	0.683	0.317	0.159	0.159	0.841	0.841
1.28	0.800	0.200	0.100	0.100	0.900	0.900
1.50	0.866	0.134	0.067	0.067	0.933	0.933
1.60	0.890	0.110	0.055	0.055	0.945	0.945
1.65	0.900	0.100	0.050	0.050	0.950	0.950
1.70	0.911	0.089	0.045	0.045	0.955	0.955
1.80	0.928	0.072	0.036	0.036	0.964	0.964
1.90	0.943	0.057	0.029	0.029	0.971	0.971
1.96	0.980	0.020	0.010	0.010	0.975	0.975
2.00	0.954	0.046	0.023	0.023	0.977	0.977
2.10	0.964	0.036	0.018	0.018	0.982	0.982
2.20	0.972	0.028	0.014	0.014	0.986	0.986
2.30	0.979	0.021	0.011	0.011	0.989	0.989
2.40	0.984	0.016	0.008	0.008	0.992	0.992
2.50	0.988	0.012	0.006	0.006	0.994	0.994
2.58	0.990	0.010	0.005	0.005	0.995	0.995
2.60	0.991	0.009	0.005	0.005	0.995	0.995
2.70	0.993	0.007	0.003	0.003	0.997	0.997
2.80	0.995	0.005	0.003	0.003	0.997	0.997
2.90	0.996	0.004	0.002	0.002	0.998	0.998
3.00	0.997	0.003	0.001	0.001	0.999	0.999
3.10	0.998	0.002	0.001	0.001	0.999	0.999
3.20	0.999	0.001	0.001	0.001	0.999	0.999
3.30	0.999	0.001	0.000	0.000	1.000	1.000
3.40	0.999	0.001	0.000	0.000	1.000	1.000
3.50	1.000	0.000	0.000	0.000	1.000	1.000

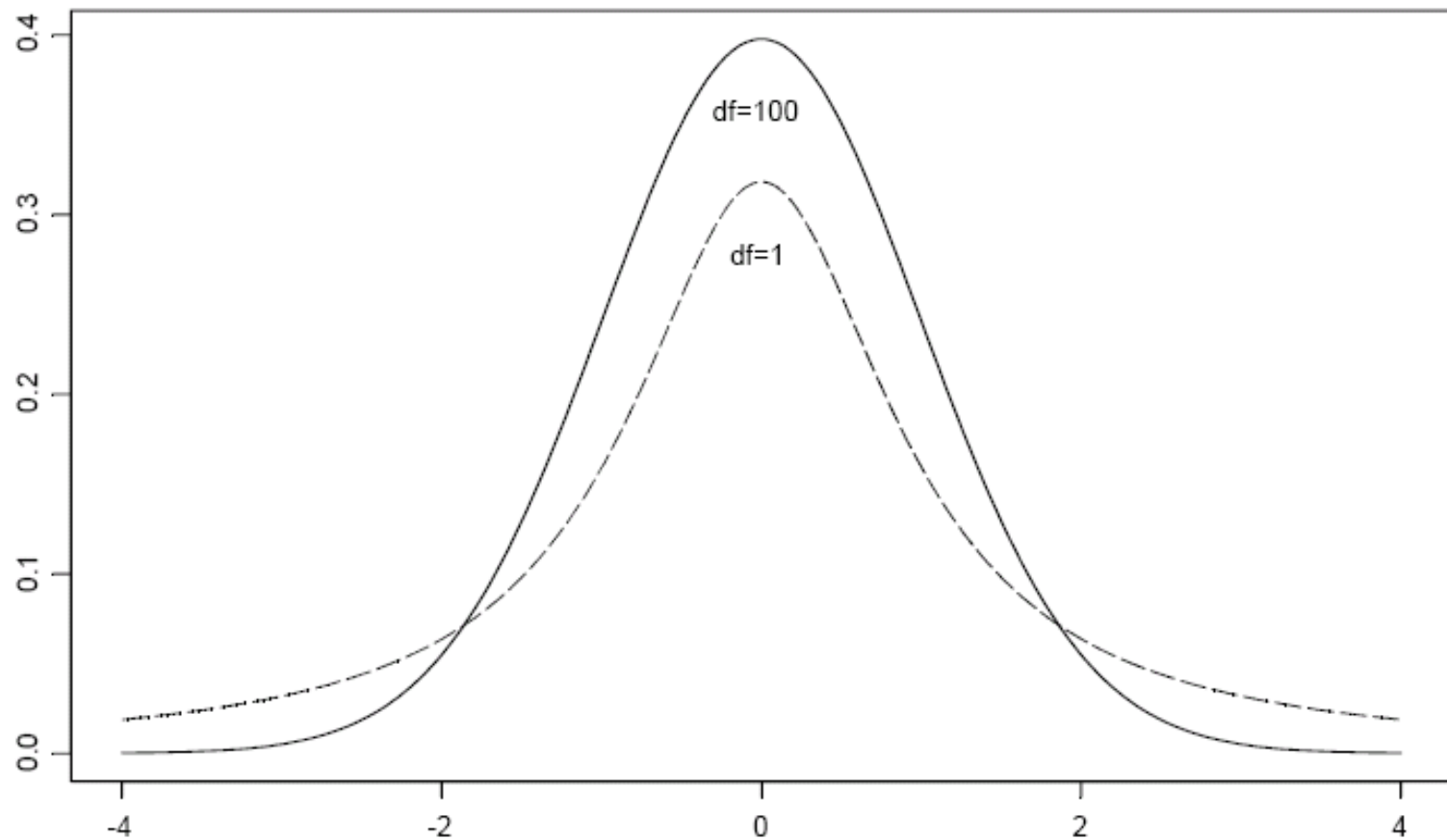
Change
confidence
interval find
the critical
value

Hint: 2.576 -> 99%

t-distribution

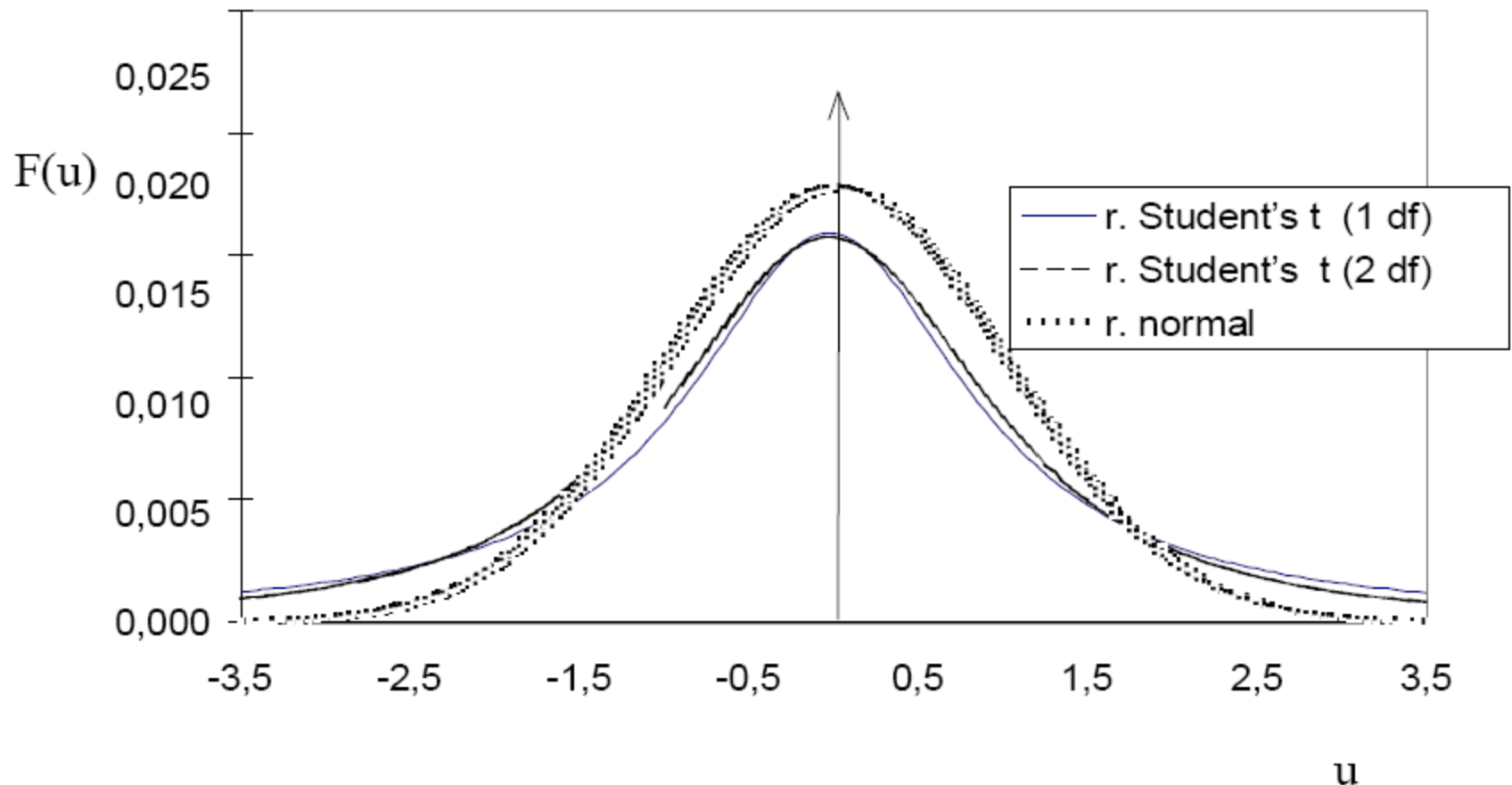
- It is unlikely that somebody will give you σ
- Most likely you will use the sample to compute s instead of σ
- We know something about the following:

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t(n - 1)$$



One parameter determines shape (called degrees of freedom)

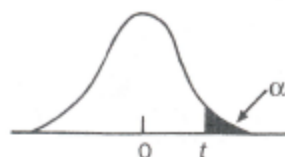
See also: http://en.wikipedia.org/wiki/Student's_t-distribution



For sample size > 100 s the t-student is almost exactly the same as the $N(0,1)$

Critical values for 5%, 1% are extremely close and for this reason we use for
 5% about 2 for critical value
 ©Konstadinos Goulias

TABLE A-4
t Distribution



df \ α	.10	.05	.025	.01	.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

HYPOTHESIS TESTING

Basics 1

- Make a proposition (claim) = null hypothesis H_0
 - Make a counter proposition (counter claim) = alternative hypothesis H_A
 - Assess the validity of a proposition against the counter proposition using sample data
1. We begin by assuming that H_0 is true.
 2. We also define an amount of acceptable risk to be wrong.
 3. If the data fails to contradict H_0 beyond a reasonable doubt, then H_0 is not rejected.
 4. Failing to reject H_0 does not mean that we accept it as true.
 5. It simply means that H_0 cannot be ruled out as a possible explanation for the observed data.

Basics 2

- A **hypothesis test** is a data-based rule to decide between H_0 and H_A .
- A **test statistic** calculated from the data is used to make this decision.
- The values of the test statistics for which the test rejects H_0 comprise the **rejection region** of the test.
- The complement of the rejection region is called the **non rejection region** (some people call it acceptance).
- The boundaries of the rejection region are defined by one or more **critical constants** also known as **critical values**.

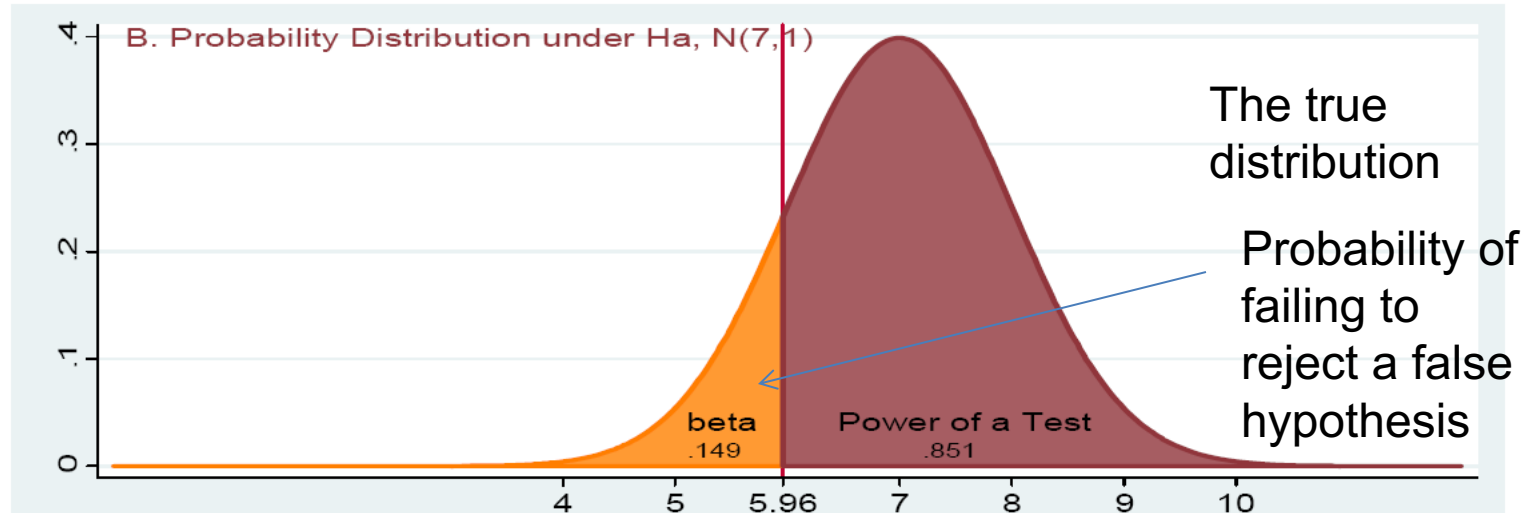
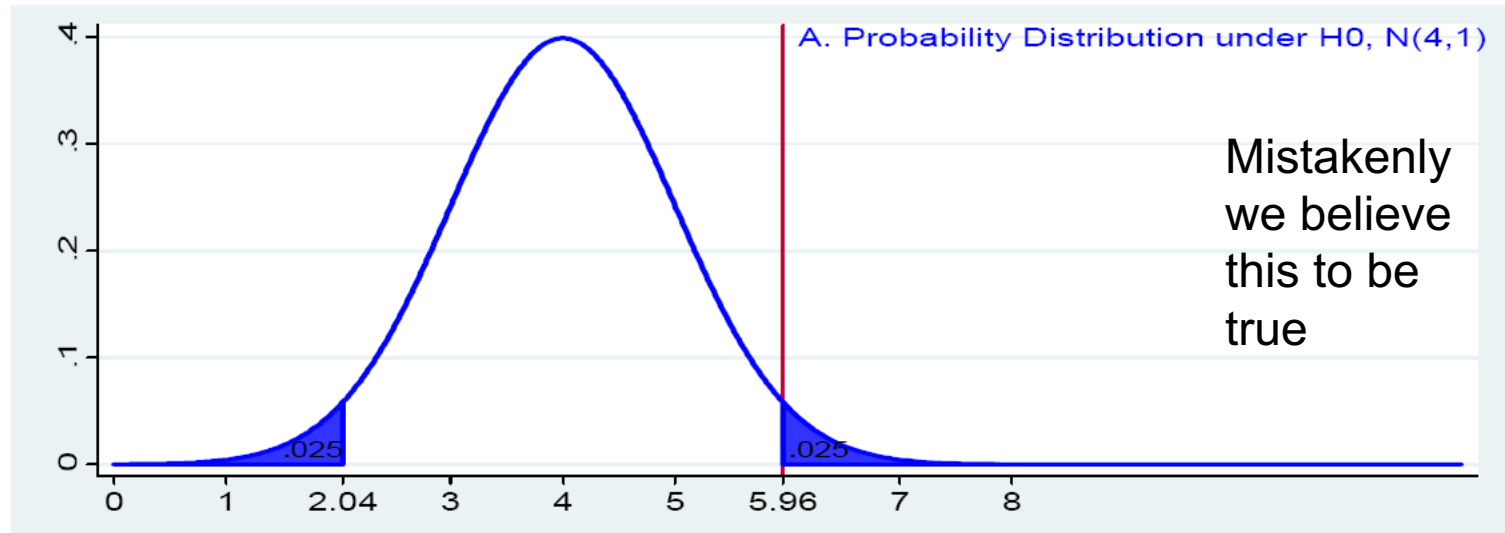
Types of Errors in Hypothesis Testing

- I. Suppose we test that θ is zero ($H_0 : \theta = 0$)
 - At the 5% level of significance we reject the H_0
 - It is possible that we incorrectly rejected the H_0
 - This is the Type I error
 - The probability of its occurrence is 5%

- II. Suppose we collect another sample and the lower and upper are -0.01 and 0.2
 - We cannot reject H_0
 - But the true value of θ is 0.145, which is not 0!
 - This mistake is a Type II error

	Do not Reject H_0	Reject H_0	Sum of probs
H_0 True	Correct Decision “confidence” $1-\alpha$	Type I Error α	1
H_0 False	Type II Error “Failure to detect” β	Correct decision “Prob of detection” $1-\beta$	1
Sum of probs	$1-\alpha+\beta$ not equal 1	$\alpha+1-\beta$ not equal 1	

$\alpha = P\{\text{Type I error}\} = P\{\text{Reject } H_0 \text{ when } H_0 \text{ is true}\} = P\{\text{Reject } H_0 | H_0\}$ also called α -risk or producer's risk or false alarm rate



$\beta = P\{\text{Type II error}\} = P\{\text{Fail to reject } H_0 \text{ when } H_1 \text{ is true}\} = P\{\text{Fail to reject } H_0 | H_1\}$ also called β -risk or consumer's risk or prob. of not detecting $\pi = 1 - \beta = P\{\text{Reject } H_0 | H_1\}$ is prob. of detection or power of the test

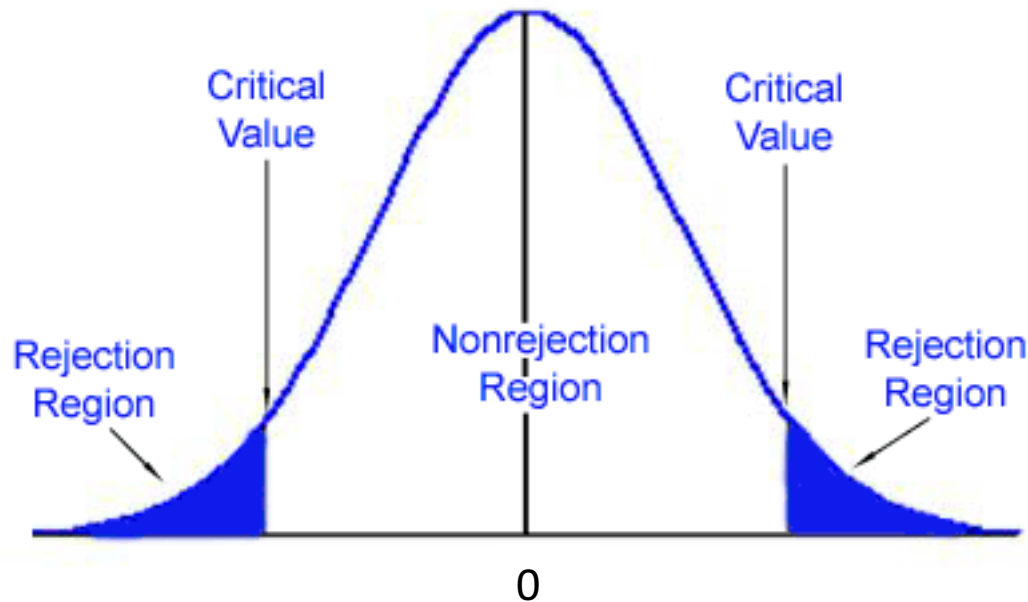
Level of Significance

- The practice of test of hypothesis is to put an upper bound on the P(Type I error) and, subject to that constraint, find a test with the lowest possible P(Type II error).
- The upper bound on P(Type I error) is called the level of significance of the test and is denoted by α (usually some small number such as 0.01, 0.05, or 0.10).
- The test is required to satisfy:
- $P\{\text{Type I error}\} = P\{\text{Test Rejects } H_0 \mid H_0\} \leq \alpha$
- Note that α is now used to denote an upper bound on P(Type I error).
- Motivated by the fact that the Type I error is usually the more serious.
- A hypothesis test with a significance level α is called an α -level test (5% or 1%).

Choice of Significance Level

- What α level should one use?
- Recall that as $P(\text{Type I error})$ decreases $P(\text{Type II error})$ increases.
- A proper choice of α should take into account the relative costs
 - of Type I and Type II errors. (These costs may be difficult to determine in practice, but must be considered!)
- Fisher said: $\alpha = 0.05$
- Today $\alpha = 0.10, \mathbf{0.05}, 0.01$ depending on how much proof
 - against the null hypothesis we want to have before rejecting it.
- P-values have become popular with the advent of computer programs.

The Same Concept



$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1) \quad T = \frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t(n-1)$$

Example with three databases of different sample sizes

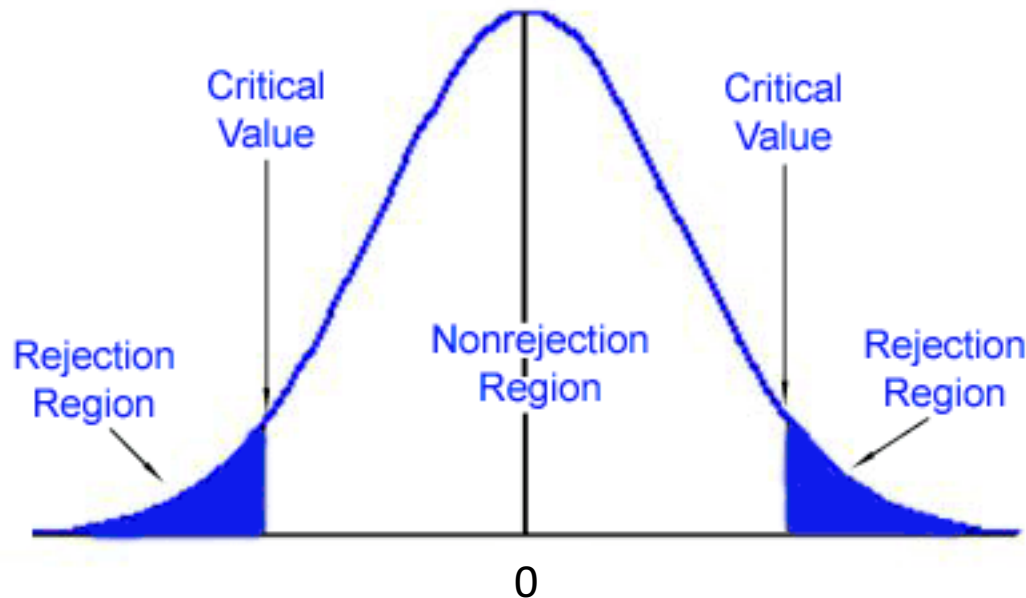
We treat our rows of 42431 observations as the statistical population

```
> describe(SmallHHfile$MilesPr) # descriptive stats for variable MilesPr
  vars      n mean      sd median trimmed  mad min      max range skew kurtosis  se
X1     1 42431 27.12 43.46   14.5    18.4 18.19   0 1167.65 1167.65 5.15    47.24 0.21
> TEN <- SmallHHfile[sample(nrow(SmallHHfile), 10), ] # randomly draw a sample of ten rows from SmallHHfile
> describe(TEN$MilesPr)
  vars  n mean      sd median trimmed  mad min      max range skew kurtosis  se
X1     1  10 9.53 10.44   4.45    8.32 6.56   0 28.73 28.73 0.57    -1.39 3.3
> HUNDRED <- SmallHHfile[sample(nrow(SmallHHfile), 100), ] # randomly draw a sample of 100 rows from SmallHHfile
> describe(HUNDRED$MilesPr)
  vars  n mean      sd median trimmed  mad min      max range skew kurtosis  se
X1     1 100 31.88 68.16  12.18    17.7 14.51   0 559.84 559.84 5.37    35.19 6.82
> THOUSAND <- SmallHHfile[sample(nrow(SmallHHfile), 1000), ] # randomly draw a sample of 1000 rows from SmallHHfile
> describe(THOUSAND$MilesPr)
  vars  n mean      sd median trimmed  mad min      max range skew kurtosis  se
X1     1 1000 25.34 36.09  14.27    17.91 17.84   0 312.29 312.29 3.55    17.34 1.14
> |
>
```

Let's compute confidence intervals and test simple hypotheses

©Konstadinos Goulias

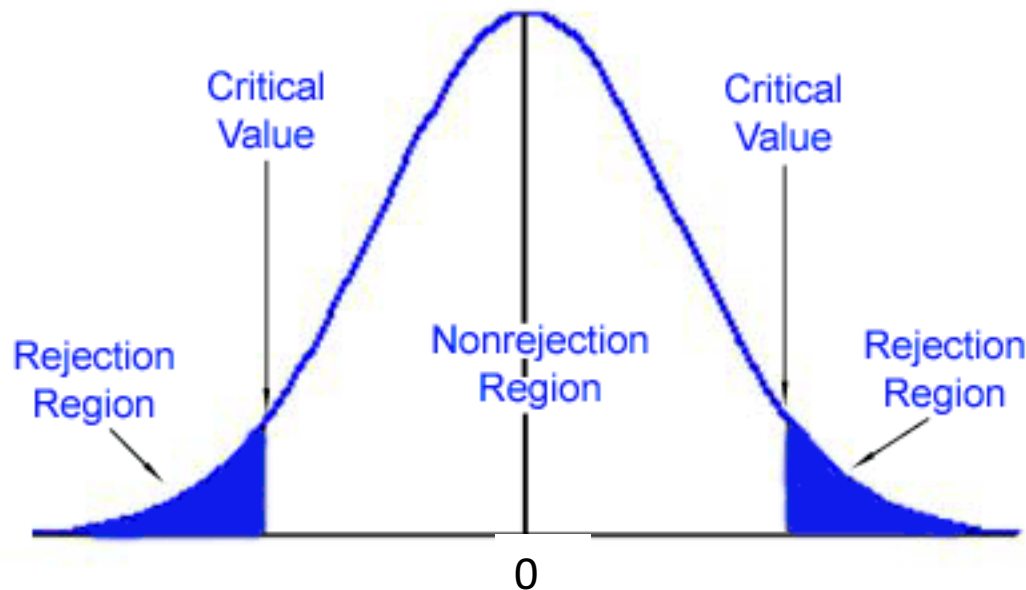
Is the mean of the Ten Row Sample equal to Population mean?
(H0: sample mean = population mean)?



At 5%
confidence I
conclude one
thing and at
10%
something
different

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{18.05 - 27.12}{43.46 / \sqrt{10}} = -0.6599599$$

Is the mean of the Thousand Rows Sample equal to
Population mean
(H0: sample mean = population mean)?



$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{24.76 - 27.12}{43.46 / \sqrt{1000}} = -1.717206$$

Under which circumstances would end up in rejection region? (think of combinations of sample size and differences between \bar{x} and μ)

In the R code (intro to Linear Regression) lines 45 to 180

- What would you do if we do not know the variance in population?
- Use an estimated sigma (s)
- Instead of Normal distribution use T-student

$$T = \frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t(n - 1)$$

- My note: go to R-studio and do hypothesis testing and first linear model

HYPOTHESIS TESTING IN REGRESSION MODELS

$$y = \alpha + \beta x + \varepsilon$$

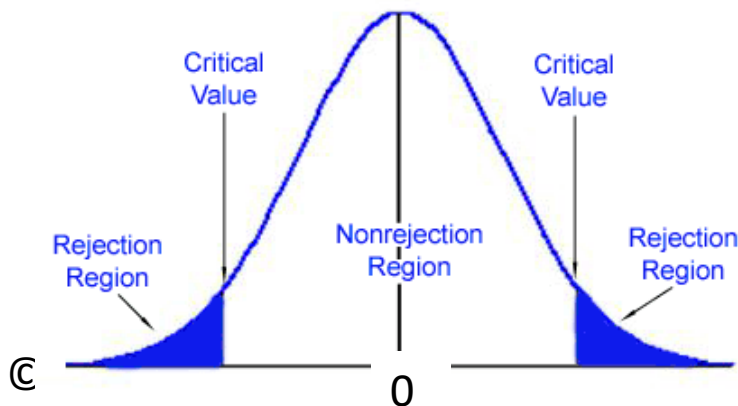
Is x influencing y?

If the coefficient β is zero, x does not influence y

We do not observe β but we have a sample from which we can obtain an estimate of β (with the method of minimizing the squared residuals we examined before)

We can also compute its standard deviation and build a statistic that will tell us if based on the data we have we can conclude that the population coefficient β is significantly different than zero and therefore x influences y

The statistic is the critical value of the T-student distribution



In regression

- Estimate of slope (b)
- Estimate of standard error of coefficient estimate
- Compute degrees of freedom
- Take the ratio between the coefficient estimate and its standard error and compare it to a T-student critical value
- Draw conclusions about the x and its relationship with y


```
> HHPMT.lm = lm(TotDist ~ HHSIZ , data=SmallHHfile)
> summary(HHPMT.lm)
```

```
Call:
lm(formula = TotDist ~ HHSIZ, data = SmallHHfile)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-186.1  -49.1  -26.5   11.4  5717.4
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.2118     1.1817   10.33  <2e-16 ***
HHSIZ        21.7305     0.4053   53.62  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

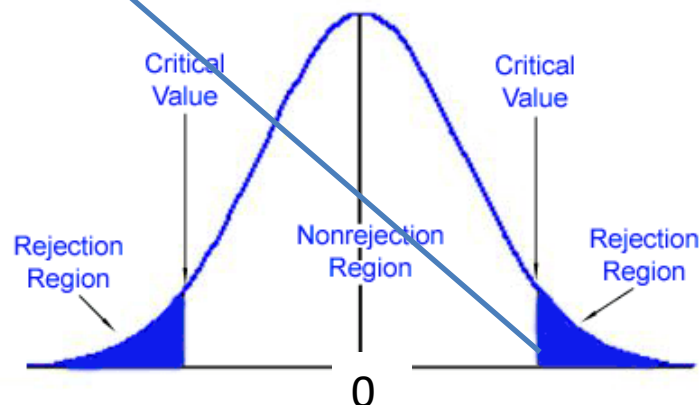
```
Residual standard error: 114.7 on 42429 degrees of freedom
Multiple R-squared:  0.06345,    Adjusted R-squared:  0.06343
F-statistic: 2875 on 1 and 42429 DF,  p-value: < 2.2e-16
```

The model is:

Number of miles per household (y) =
12.2118 + 21.7305* Household Size

Household Size is significantly influencing the number of trips a person makes.

Every additional person in the household increases the number of miles by 21.7305 units. This is BIG positive contribution to the total number of miles a household makes in a day



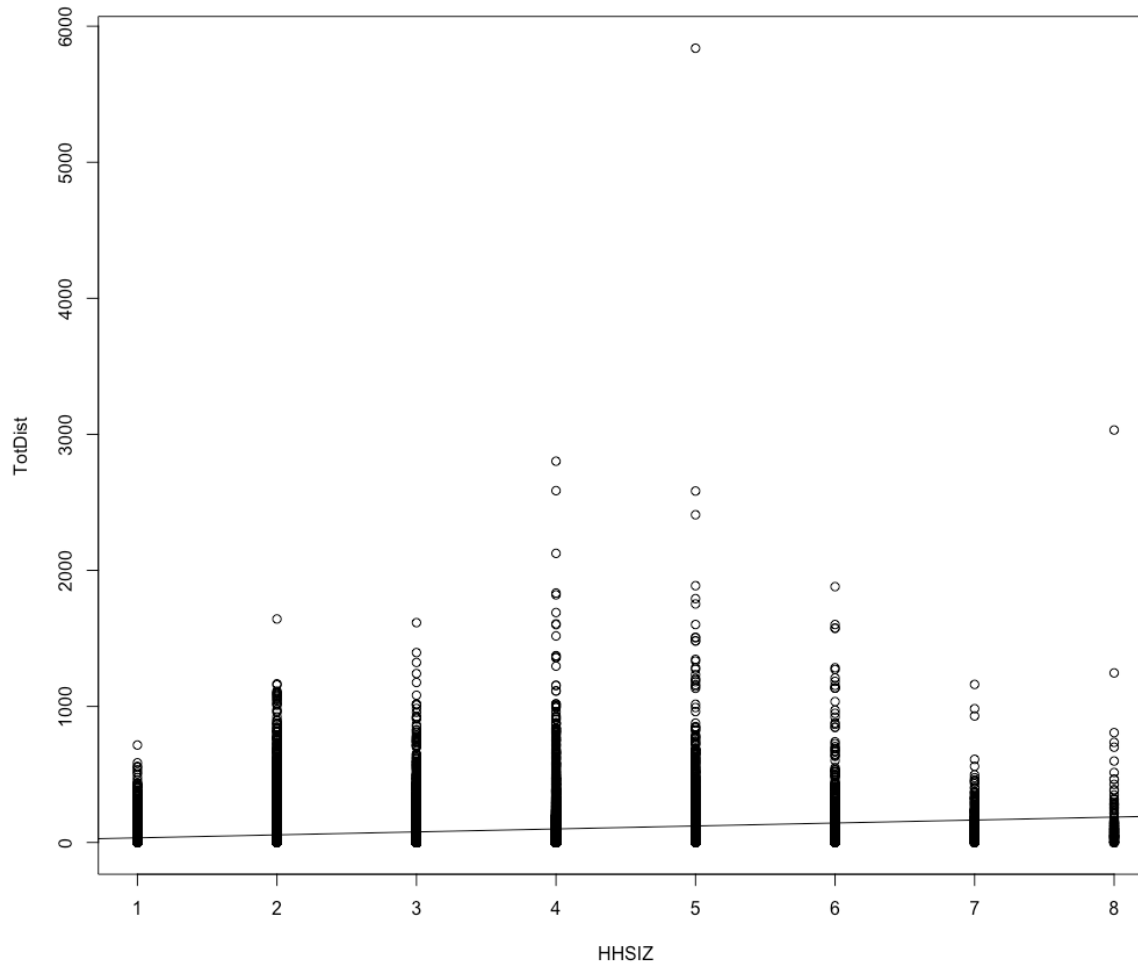
The software reports the exact alpha corresponding to a critical value = 53.62

Contents of lm object

- `output <- summary(result)` most important stats
- `SSR <- deviance(result)` sum of squared residuals
- `LL <- logLik(result)` log likelihood(later)
- `DegreesOfFreedom <- result$df` degrees of freedom
- `Yhat <- result$fitted.values` the dep. var. predictions
- `Coef <- result$coefficients` coefficient estimates
- `Resid <- result$residuals` residuals
- `s <- output$sigma` estimate of stand.Err of resid
- `RSquared <- output$r.squared` Goodness of Fit
- `CovMatrix <- s^2*output$cov` covariance of coef.est.
- `aic <- AIC(result)` $AIC = -2\log L(p) + 2p$
- `sbc <- AIC(result, k=log(NROW(smokerdata)))` $SBC = -2\log L(p) + p\log(N)$

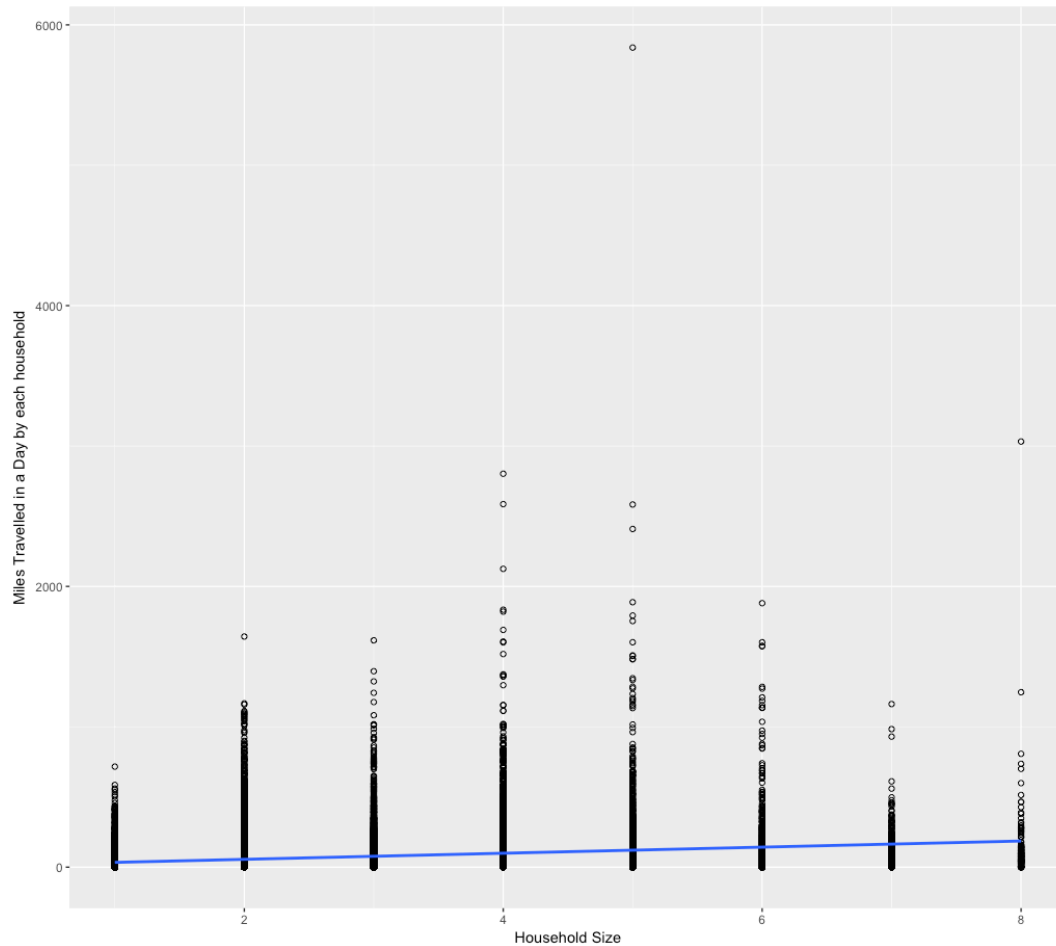
We will define and use some of these later

```
plot(TotDist ~ HHSIZ, data = SmallHHfile)  
abline(HHPMT.lm)
```

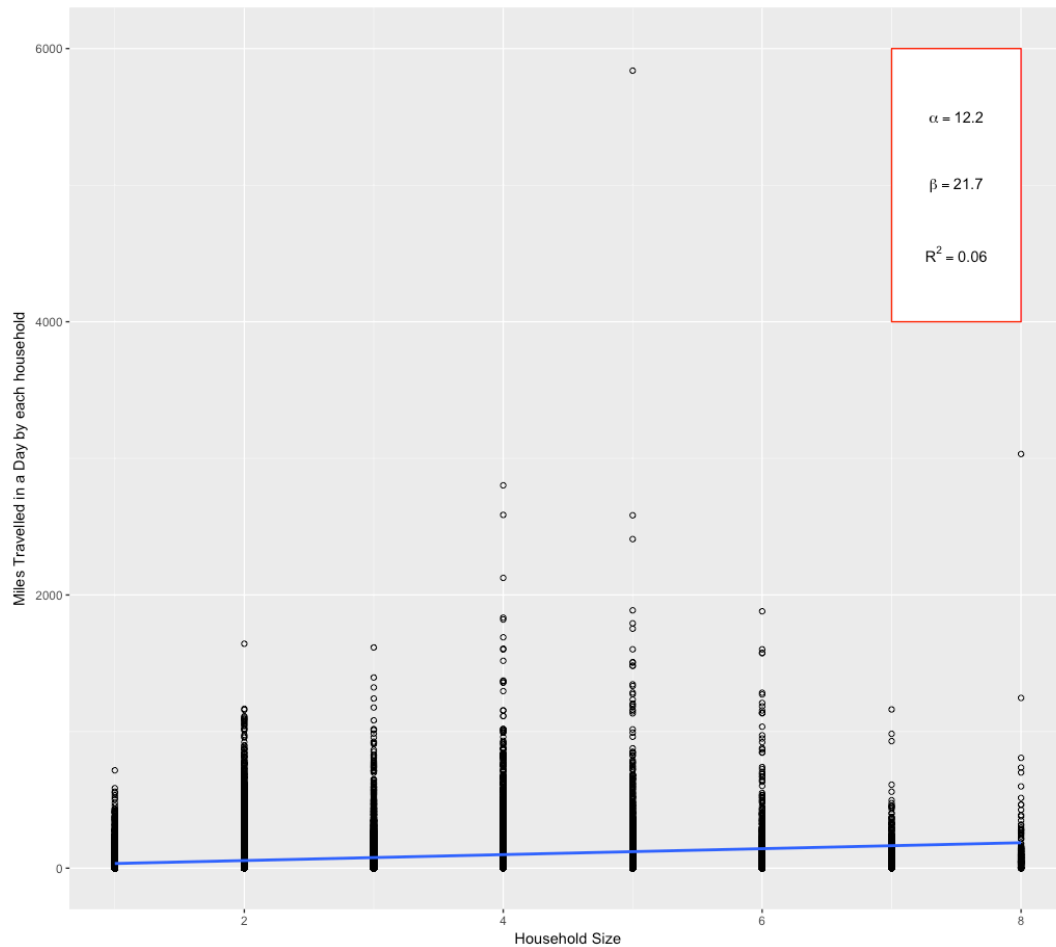


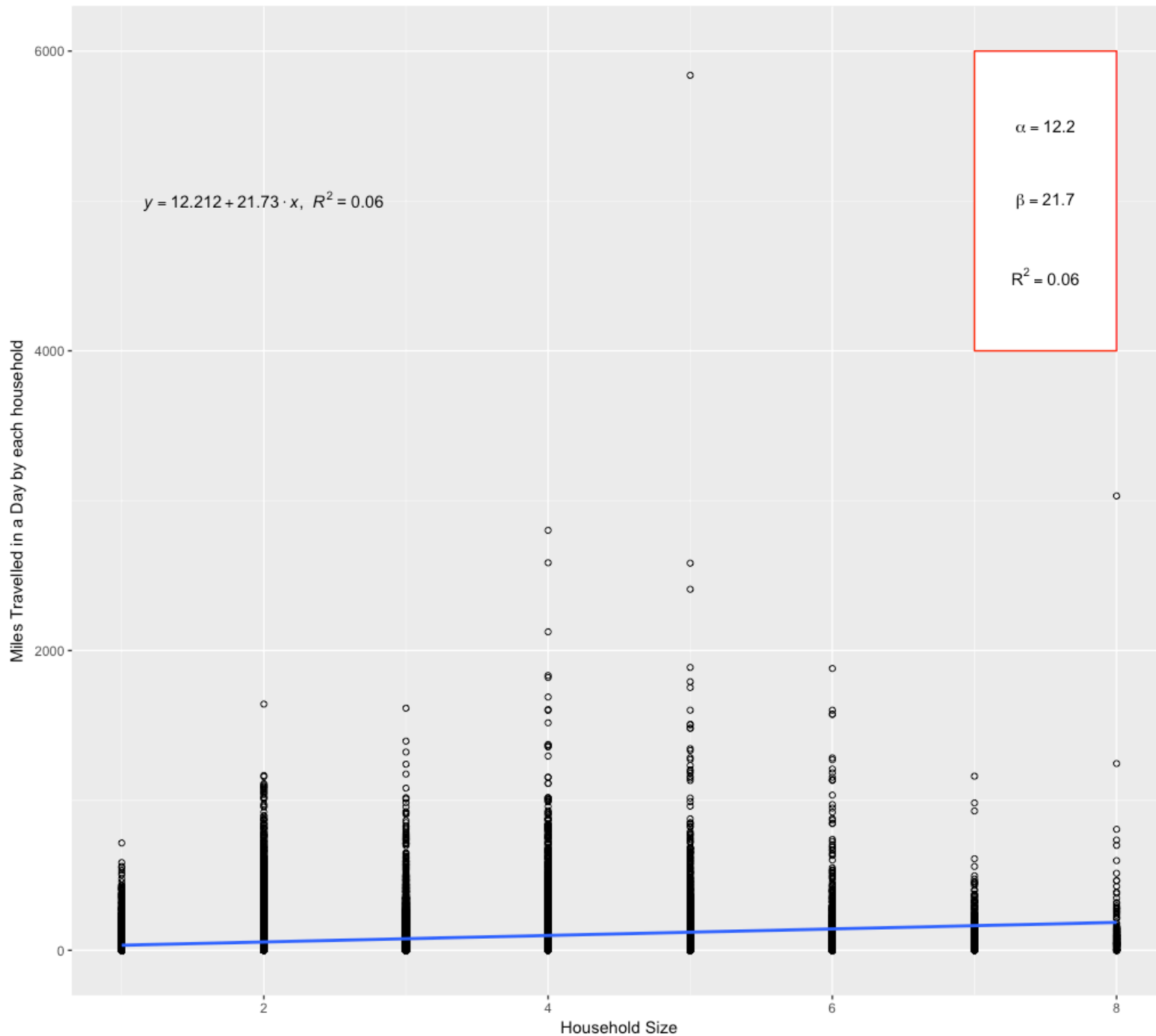
I used copy to
clipboard in r-
studio

Using ggplot (line 230 in R script)



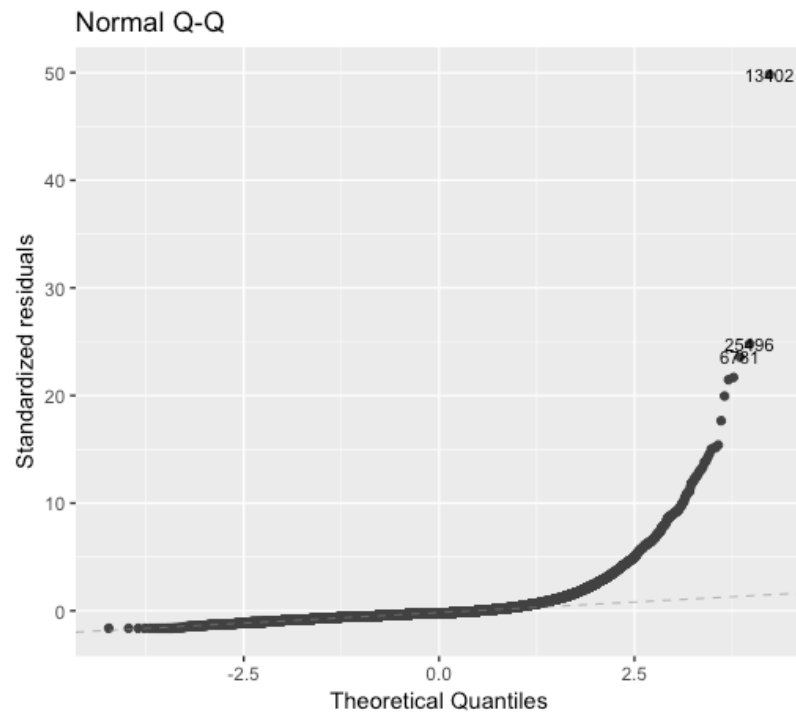
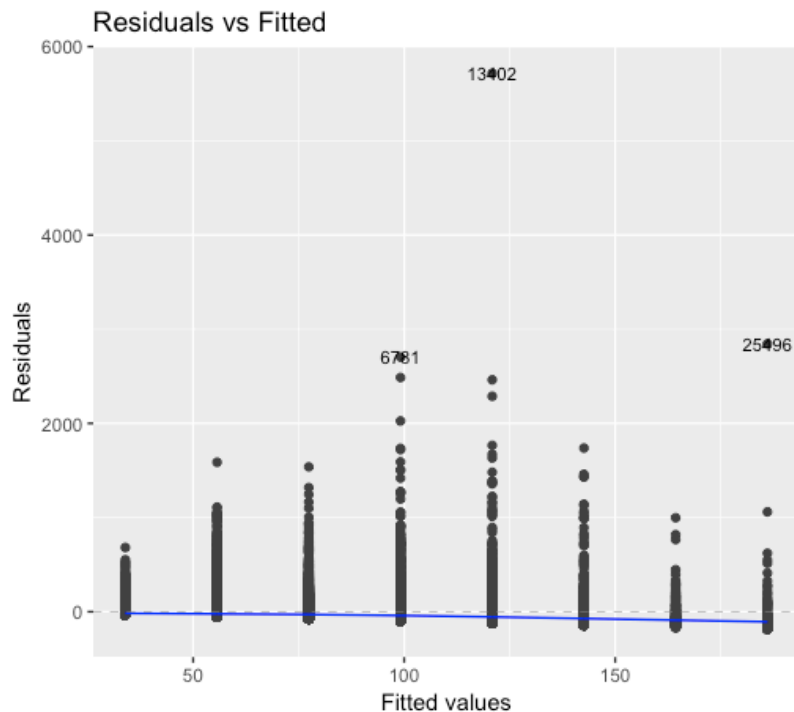
ggplot with annotate (line 239 R script)





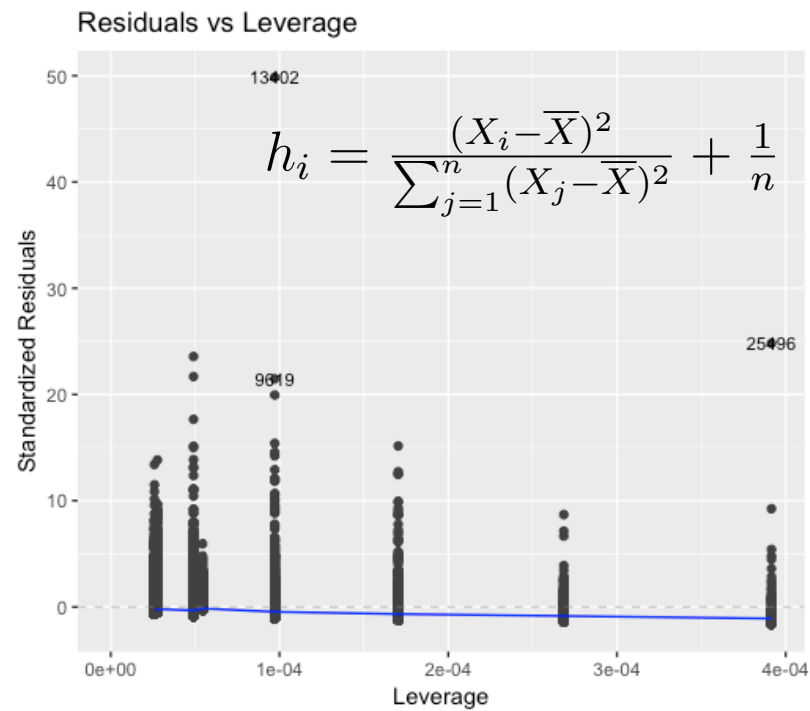
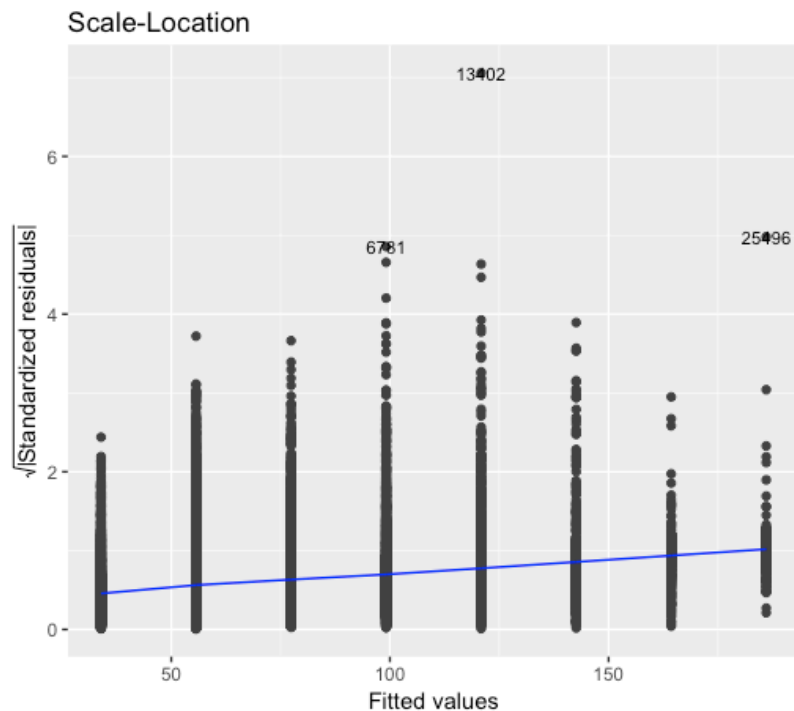
Ggplot with
annotate
and define
an object
that
extracts
values from
an lm
object and
insert it
into a
graph (line
248 to 259
r script)

DIAGNOSTICS WITH GRAPHICS (FOLLOWING GRAPHS FROM LINE 264- 266 R SCRIPT)



- Residuals vs Fitted – are residuals related to the fitted values in a linear way? (remember the fitted values are a linear function of x) - If they are equally spread residuals around a horizontal line without distinct patterns we don't have non-linear relationships.
- Residuals vs Normal Q-Q - check if residual resemble the normal distribution - use standardized = $\text{resid}_i / \text{standard deviation of resid}_i$
- Good if residuals are lined well on the straight dashed line.

- Sqrt of standardized residual vs Fitted – are residuals related to the fitted values (remember the fitted values are a function of x) -> maybe the variance of the residuals is not constant?
- Good if I see equally (randomly) spread points around the horizontal line (we will use a formal statistical test later).
- Standardized Residuals vs Leverage (h_i)- check if some observations with very high x influence our residuals (rule of thumb $h > 0.5$)



Findings

- Residuals change with fitted but seem ok
- Residual variance changes with fitted maybe heteroskedastic?
- Not normally distributed (see extremes)
- Not clear indications about influential observations with high or low x values

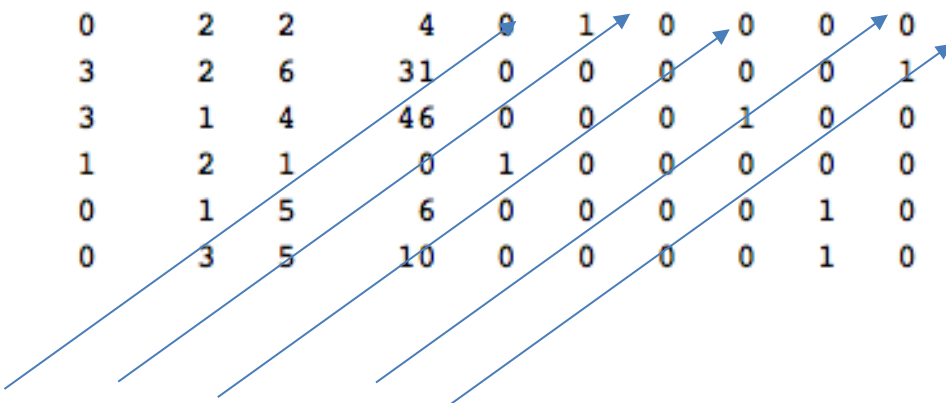
DUMMY VARIABLES (ALSO CALLED INDICATOR VARIABLES)

Dummy Variable

- Takes the value 1 if a condition is satisfied; 0 otherwise.

> SmallHHfile

	SAMPN	INCOM	HHSIZ	HHEMP	HHSTU	HHLIC	DOW	HTRIPS	Mon	Tue	Wed	Thu	Fri	Sat	Sun	TotDist
1	1031985	3	2	0	0	2	2	4	0	1	0	0	0	0	0	3.627588e+01
2	1032036	7	5	1	3	2	6	31	0	0	0	0	0	1	0	1.648952e+02
3	1032053	2	6	1	3	1	4	46	0	0	0	1	0	0	0	4.244294e+01
4	1032425	7	2	2	1	2	1	0	1	0	0	0	0	0	0	0.000000e+00
5	1032558	1	1	0	0	1	5	6	0	0	0	0	1	0	0	2.980830e+00
6	1033586	3	3	1	0	3	5	10	0	0	0	0	1	0	0	6.257646e+02

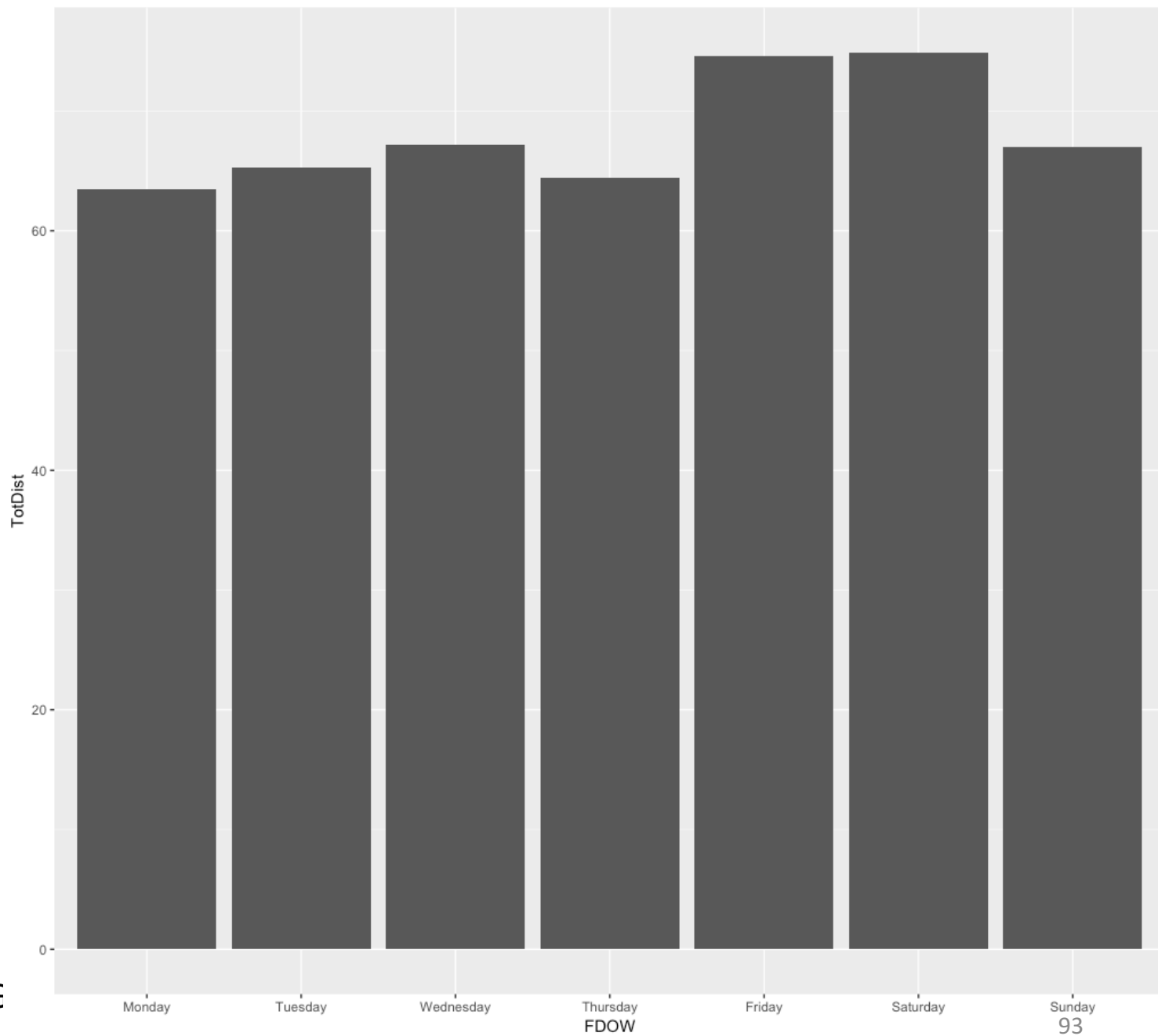


All these are dummies and were defined using the variable DOW

They are mutually exclusive (you cannot be a respondent of interview on a Monday and a Tuesday by the survey design)

Average TotDist
for each day of the
week

Households
interviewed on Fridays
or Saturdays travel
more



```
HHPMT2.lm = lm(TotDist ~ Mon + Tue + Wed + Thu + Fri+ Sat , data=SmallHHfile)
summary(HHPMT2.lm)
```

```
> summary(HHPMT2.lm)
```

Call:

```
lm(formula = TotDist ~ Mon + Tue + Wed + Thu + Fri + Sat, data = SmallHHfile)
```

Residuals:

Min	1Q	Median	3Q	Max
-74.9	-59.2	-33.9	14.9	5763.7

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	66.9746	1.4999	44.653	< 2e-16 ***
Mon	-3.4986	2.1618	-1.618	0.105601
Tue	-1.6596	2.1313	-0.779	0.436166
Wed	0.2083	2.1296	0.098	0.922087
Thu	-2.5304	2.1198	-1.194	0.232617
Fri	7.6030	2.1495	3.537	0.000405 ***
Sat	7.9258	2.1439	3.697	0.000218 ***

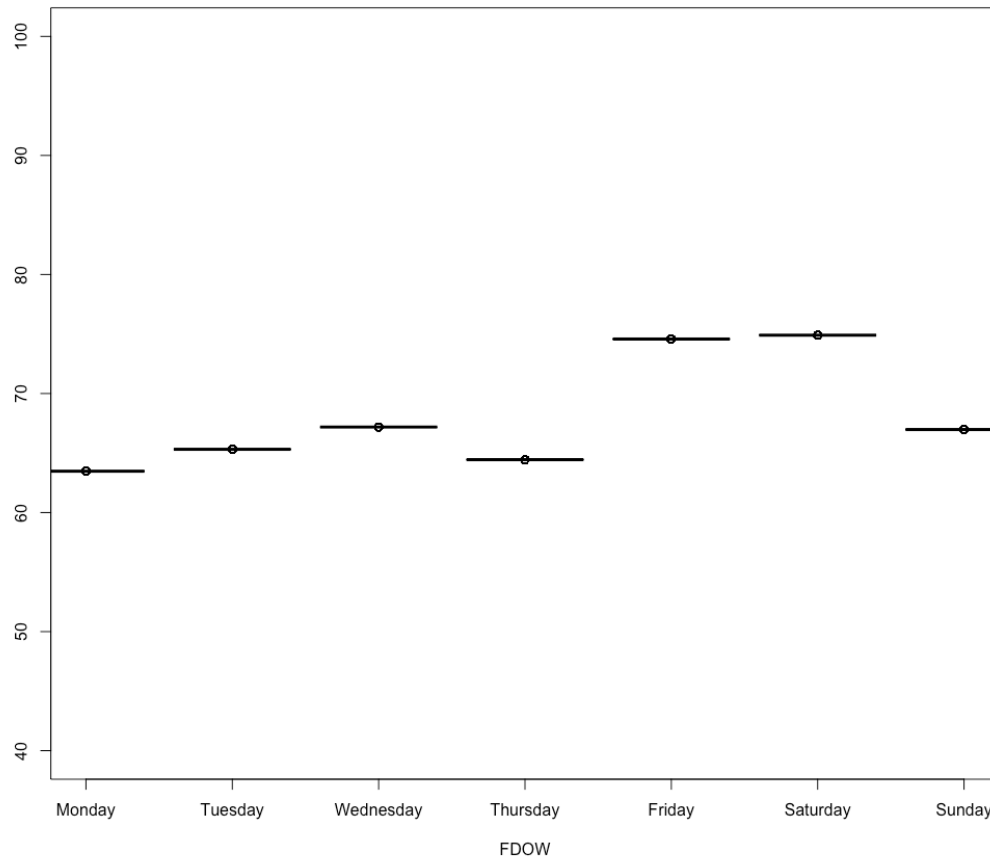
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 118.4 on 42424 degrees of freedom

Multiple R-squared: 0.00133, Adjusted R-squared: 0.001189

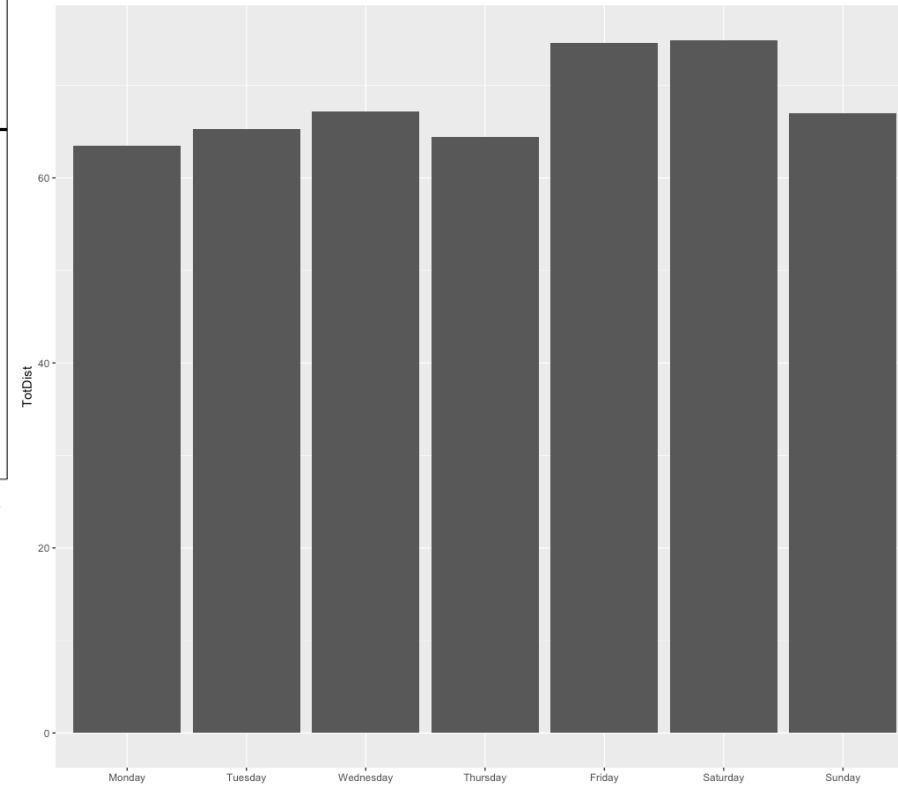
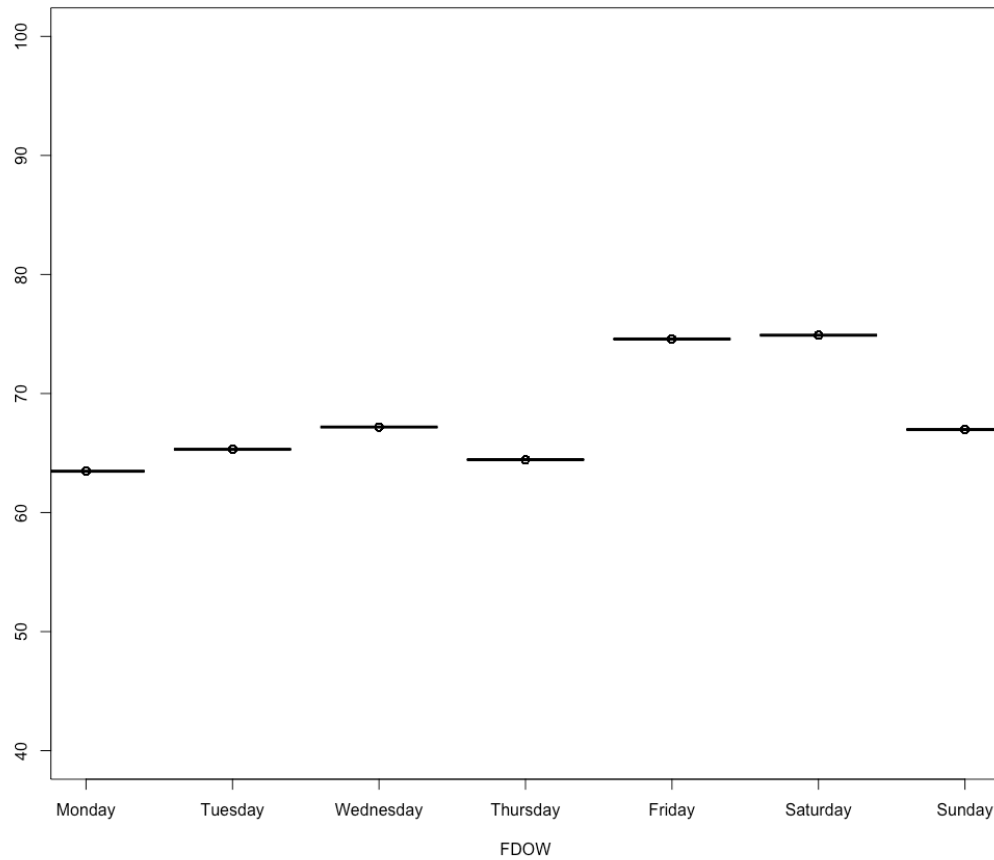
F-statistic: 9.417 on 6 and 42424 DF, p-value: 2.345e-10

Plot of fitted model vs day of the week

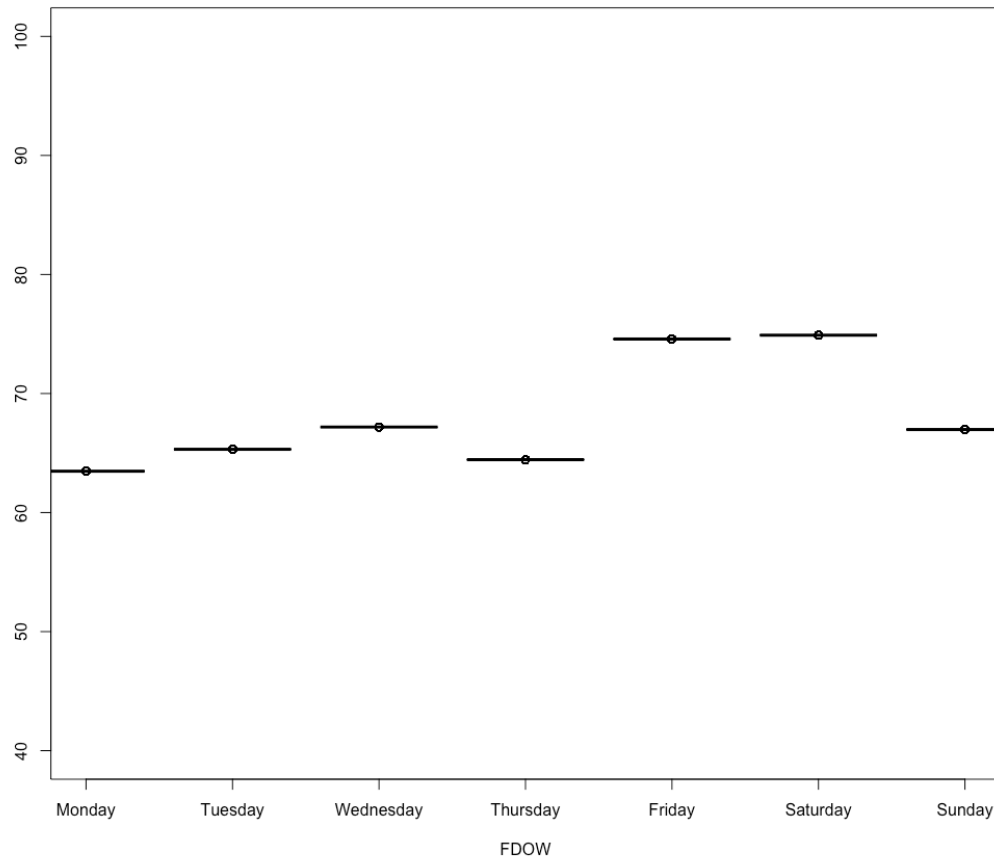


Intercept is
about 67

What do you observe?



What do you observe?



```
> summary(HHPMT2.lm)
```

Call:

```
lm(formula = TotDist ~ Mon + Tue + Wed + Thu + Fri + Sat, data = SmallHHf)
```

Residuals:

Min	1Q	Median	3Q	Max
-74.9	-59.2	-33.9	14.9	5763.7

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	66.9746	1.4999	44.653	< 2e-16 ***
Mon	-3.4986	2.1618	-1.618	0.105601
Tue	-1.6596	2.1313	-0.779	0.436166
Wed	0.2083	2.1296	0.098	0.922087
Thu	-2.5304	2.1198	-1.194	0.232617
Fri	7.6030	2.1495	3.537	0.000405 ***
Sat	7.9258	2.1439	3.697	0.000218 ***

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

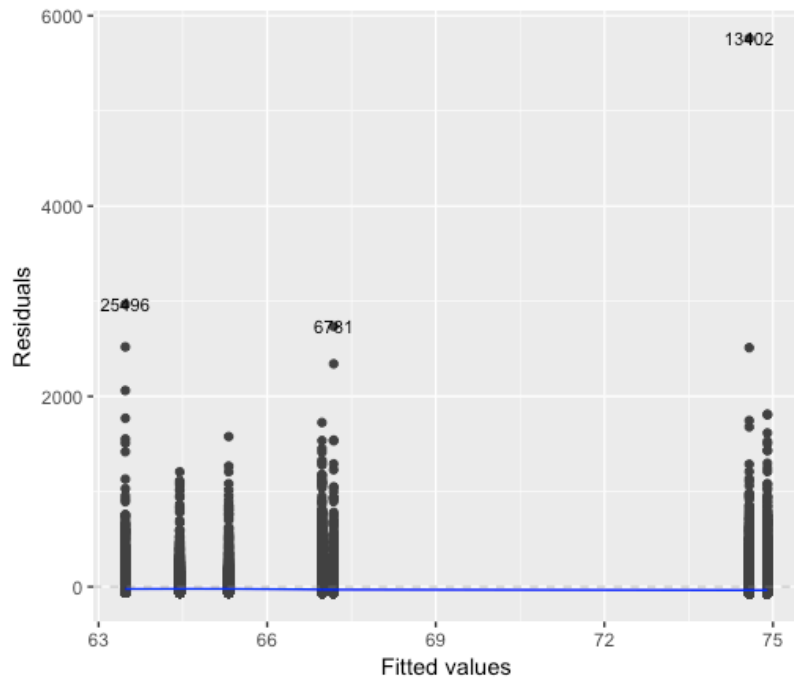
Residual standard error: 118.4 on 42424 degrees of freedom

Multiple R-squared: 0.00133, Adjusted R-squared: 0.001189

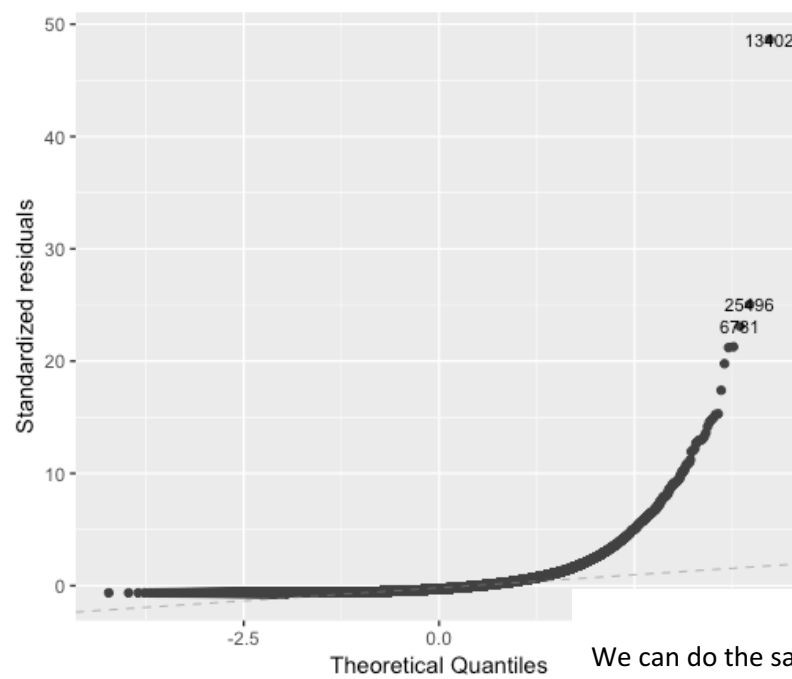
F-statistic: 9.417 on 6 and 42424 DF, p-value: 2.345e-10

The coefficient associated with each day of the week dummy adds or subtracts from the intercept to reproduce the overall mean

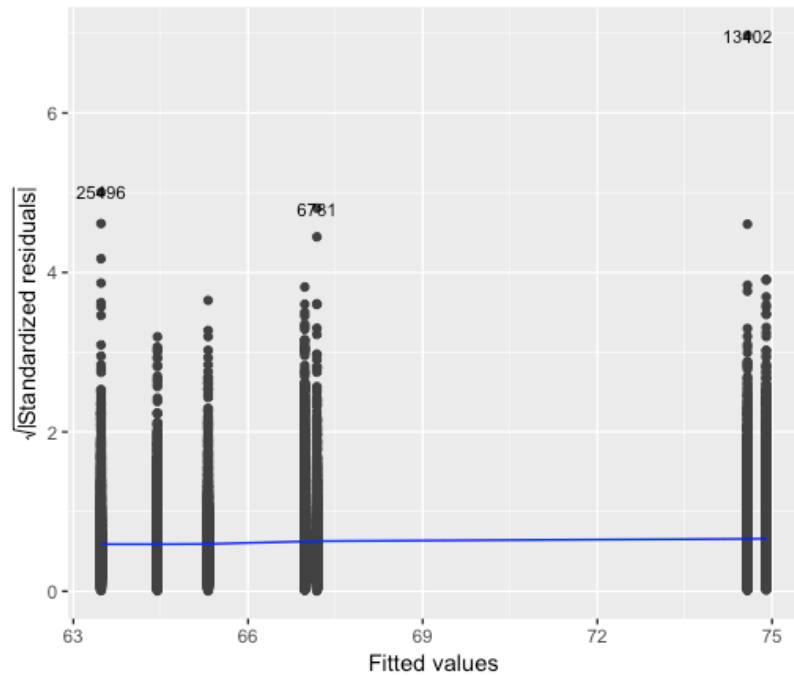
Residuals vs Fitted



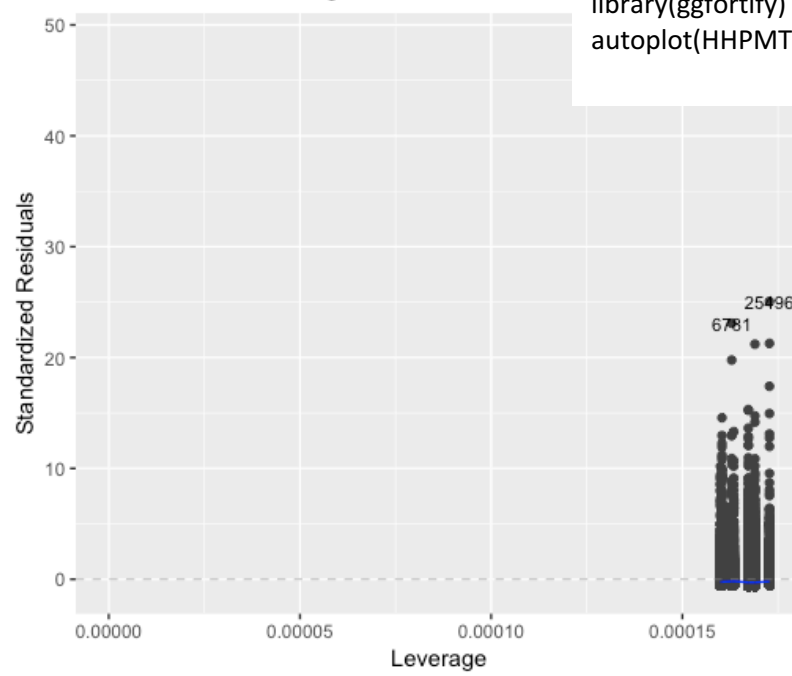
Normal Q-Q



Scale-Location



Residuals vs Leverage



We can do the same plotting of residuals as with the HHSIZ variable
`library(ggfortify)`
`autoplot(HHPMT2.lm, label.size = 3)`

My note: go to R-studio and do the rest