Advanced Data Analysis I
Simple Regression

PA 541 Week 3

Michael D. Siciliano
Department of Public Administration
College of Urban Planning and Public Affairs

Overview

- Today's class will provide:
  - a look at different types of data
  - brief coverage of basic data handling techniques in R
  - a review of simple regression, and
  - a look at simple regression assumptions.

- Note, some of the items discussed by Wooldridge in ch. 2 will not be covered this week. Rather, these items will be covered in more detail later in the semester. This includes the topics on unit of measurement and functional form that were presented in section 2.4.

- At the end of class, we will spend some time working with data and running simple regressions. We will assign you to breakout rooms based on your final project groups.

Admin Stuff

- I will post first **homework** this week. It will be due on February 8th.

- Homework will cover data wrangling and basic regression models and interpretation.

Starter Question

- Assume I have some outcome variable Y, and I generate 100 independent variables at random. **How many of those 100 independent variables will have a significant correlation ($p < .05$) with Y?**

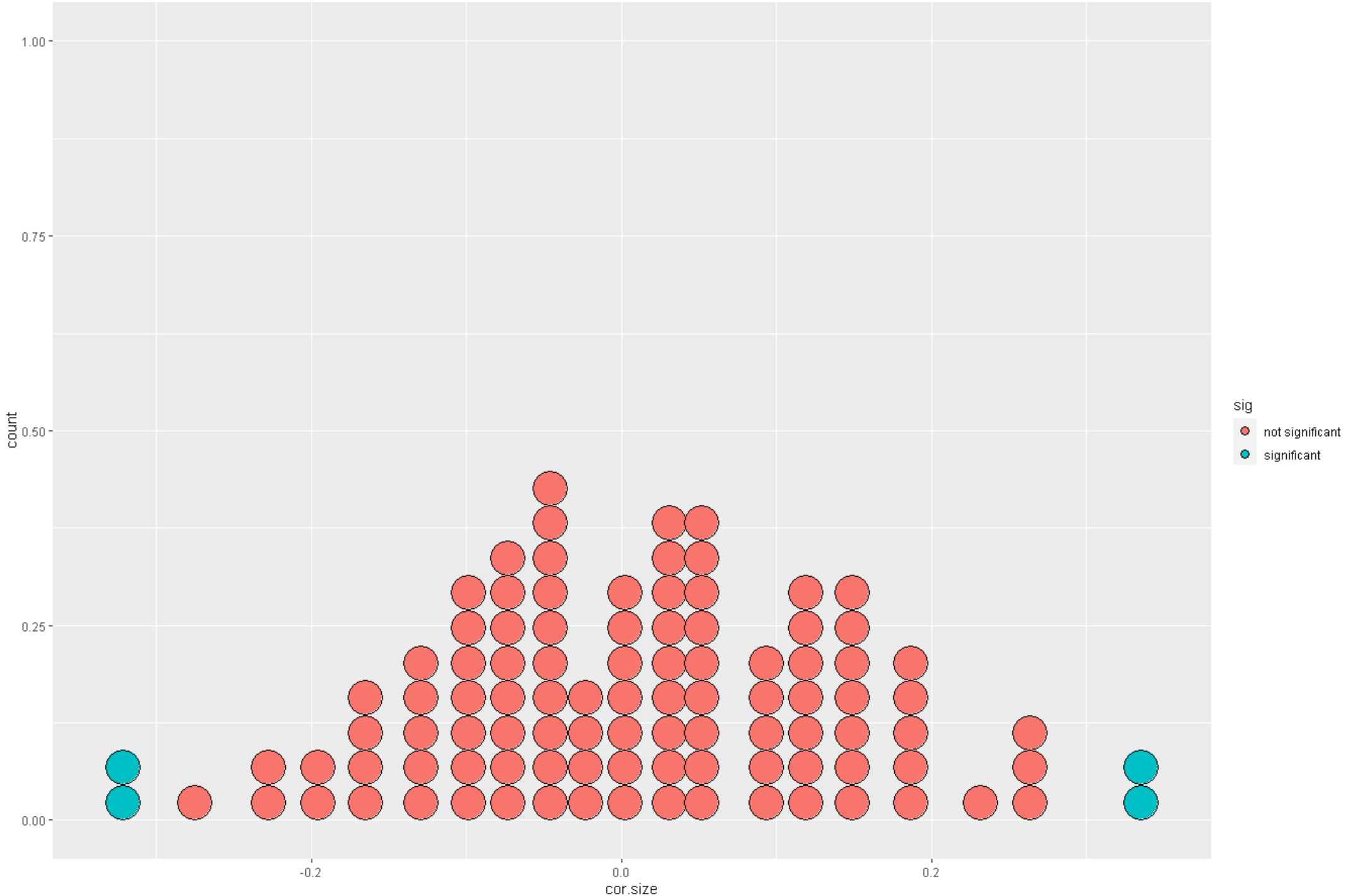
- About "snooping" and spurious findings.
  - This also highlights potential problems with "big data" and data mining.
  - And connects to how we understand a p-value.

```r
set.seed(1827)
a.random = rnorm(50)

b.data = data.frame(replicate(100, rnorm(50)))

cor.size = vector()
for (i in 1: 100)
  cor.size[i] = cor.test(a.random, b.data[,i])$estimate

cor.results = vector()
for (i in 1: 100)
   cor.results[i] = cor.test(a.random, b.data[,i])$p.value
```


1.00
0.75
0.50
0.25
0.00
count
-0.2
0.0
0.2
cor.size
sig
not significant
significant

Types of datasets

- **Econometric/Statistical analysis requires data!**
- **Different kinds of datasets exist and include:**
  - Cross-sectional data
  - Time series data
  - Pooled cross-sectional data
  - Panel/Longitudinal data
- **Econometric methods depend on the nature of the data used.**
  - Use of inappropriate methods may lead to misleading results.

Cross-sectional data sets

- Sample of individuals, households, firms, cities, states, countries, or other units of interest at a given point of time/in a given period.

- Cross-sectional observations are more or less **independent.**
  - For example, **pure random sampling** from a population.

- Sometimes pure random sampling is violated, e.g. units refuse to respond in surveys, or if sampling is characterized by clustering.
  - For example, we sample students in 10 different classrooms or 10 different neighborhoods.

- Cross-sectional data are typically the most commonly found data in various fields of research.
  - We will work with cross-sectional data for both continuos and discrete outcomes.

- **Cross-sectional data set on wages and other characteristics**

TABLE 1.1 A Cross-Sectional Data Set on Wages and Other Individual Characteristics

obsno	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

Indicator variables (1=yes, 0=no)

Observation number

Hourly wage

Time series data

- Time Series data consists of observations on a variable or several variables over time. Examples of time series data include stock prices, money supply, consumer price index, gross domestic product, annual homicide rates, and automobile sales.

- Because past events can influence future events and lags in behavior are prevalent in the social sciences, time is an important dimension in a time series data set. Unlike the arrangement of cross-sectional data, the chronological ordering of observations conveys important information.

- The key feature that makes time series data more difficult to analyze than cross-sectional data is the fact that **observations of social phenomena are rarely, if ever, assumed to be independent across time**.

- **Time series data on minimum wages and related variables**

TABLE 1.3 Minimum Wage, Unemployment, and Related Data for Puerto Rico

obsno	year	avgmin	avgcov	prunemp	prgnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
.	.	.	.	.	.
.	.	.	.	.	.
.		.	.	.	.
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

Average minimum wage for given year

Average coverage rate

Unemployment rate

Gross national product

Pooled cross-sectional data

- Pooled Cross Section have both cross-sectional and time series features. For example, suppose that two cross-sectional surveys of nonprofits are taken in the US, one in 2008 and one in 2018.

- Pooling cross sections from different years is often an effective way of analyzing the effects of a new government policy. The idea is to collect data from the years before and after a key policy change.
  - We will look at examples of after the midterm.

- A Pooled cross-section is analyzed much like a standard cross-section, except that we often need to account for differences in the variables across time.

Pooled cross sections on housing prices

TABLE 1.4 Pooled Cross Sections: Two Years of Housing Prices

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85500	42	1600	3	2.0
2	1993	67300	36	1440	3	2.5
3	1993	134000	38	2000	4	2.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243600	41	2600	4	3.0
251	1995	65000	16	1250	2	1.0
252	1995	182400	20	2200	4	2.0
253	1995	97500	15	1540	3	2.0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57200	16	1100	2	1.5

Property tax

Size of house in square feet

Number of bathrooms

Before reform

After reform

Panel or longitudinal data

- Panel data consists of a time series for each cross-sectional member in the data set.

- The key feature of panel data that distinguishes it from a pooled cross section is the fact that the same cross-sectional units (individuals, firms, counties, etc…) are followed over a given time period.
  - Example:
    - City crime statistics; each city is observed over several years
    - Time-invariant unobserved city characteristics may be modeled
    - Effect of police on crime rates may exhibit time lag
  - We will work with panel data and panel data models in week 14.

Two-year panel data on city crime statistics

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

Each city has two time series observations

Number of police in 1986

Number of police in 1990

Data Management

Topics

- Let's make sure you can do the following steps in R:
  - Set working directory
  - Load packages you will need
  - Read in the data

- Then we will go over some critical data management techniques in R such as:
  - Merging two dataframes
  - Creating new variables (already discussed with mutate)
  - Recoding variables
  - Creating a factor variable
  - Missing values and recoding missing values

Dataset we will work with: ‘pima’

Description

- The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix.

Format

The dataset contains the following variables:

- ***pregnant*** Number of times pregnant
- ***glucose*** Plasma glucose concentration at 2 hours in an oral glucose tolerance test
- ***diastolic*** Diastolic blood pressure (mm Hg)
- **triceps** Triceps skin fold thickness *(mm)*
- ***insulin*** 2-Hour serum insulin (mu U/ml)
- ***bmi*** Body mass index (weight in kg/(height in metres squared))
- ***diabetes*** Diabetes pedigree function (based on relatives or siblings having diabetes)
- ***age*** Age (years)
- ***test*** test whether the patient shows signs of diabetes (coded 0 if negative, 1 if positive)

- Let’s switch over to R

Bivariate Regression and Interpretation

Simple Regression

- "What is the relationship between variable X and variable Y?"
  - We can use regression to answer this question.

- Two variables X and Y may be related to each other exactly (as often in the physical sciences) or inexactly (as so often in the social sciences).

- $Y = a + bX$ where the values of the coefficients, a and b, determine the intercept and slope of the line relating X to Y.

Simple Regression

- Y = a + bX
- ‘a’ is referred to as the constant or intercept term and ‘b’ is referred to as the slope. It is just the formula for a line.
- Because relationships in the social science are inexact the equation is more realistically written as Y = a +bX + e where ‘e’ simply represents the presence of error.
- In this class, we see it written as:

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

Bivariate Regression Model

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

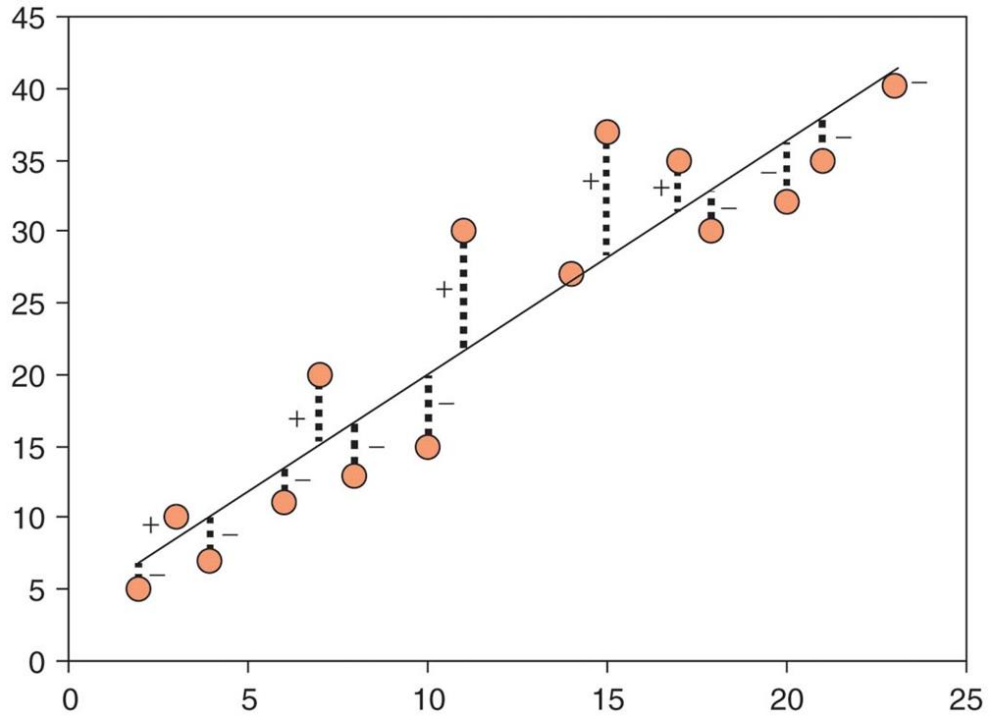
- The slope parameter, β_1, indicates the change in *Y* associated with a one unit increase in *X*.
- The intercept parameter, β_0, indicates the expected value of *Y* when *X* is zero.
- The estimates of these parameters are determined by minimizing the sum of squared residuals.

- Finding the optimal values for our coefficients requires us to minimize the sum of squared residuals (this is the ordinary least squares estimate).

$$SSR = \sum_{i=1}^{n} e_i^2$$

$$SSR = \sum_{i=1}^{n} (Yi - \widehat{Yi})^2$$

$$SSR = \sum_{i=1}^{n} (Yi - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$



Calculating the coefficients

- The values of β_0 and β_1 below are our least squares estimates for bivariate regressions:

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Two things to remember about the beta-hats: they are **random** variables and are **normally distributed** (due to the CLT).

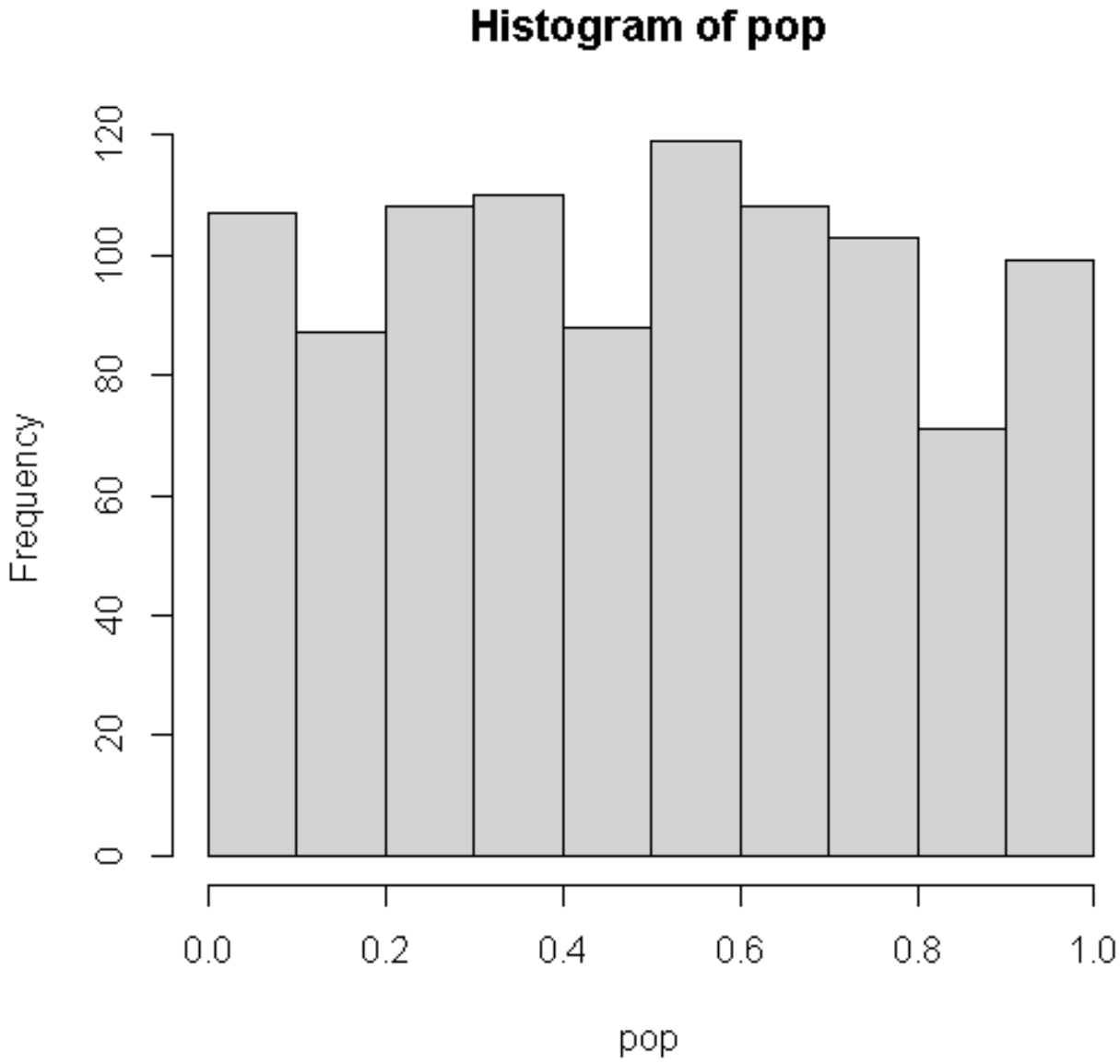
Central Limit Theorem

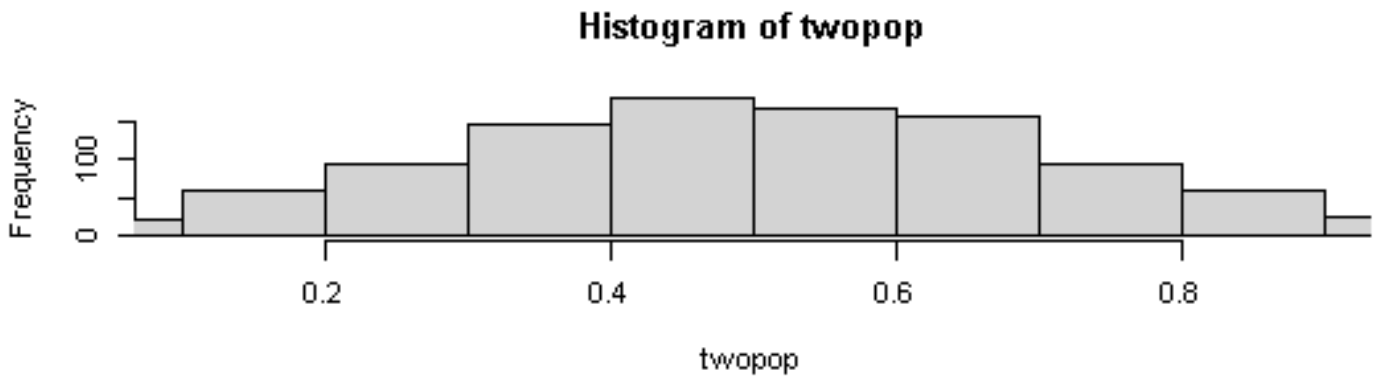
- Regardless of the distribution of the original variable X from which the statistic $\bar{X}$ was generated, the asymptotic sampling distribution of the statistic will be normal.
- Again:
  - Consider that when we take a random sample of size *n*, this sample is only one of many possible samples of size n that could be drawn from the population.
  - Obviously, if we took a second sample of size *n*, we would not end up with the same respondents and so the mean of the new sample would be different from the first sample.
  - The CLT, however, says that if we were to take all possible random samples of a given size from the population and compute the mean for each one, the distribution of the calculated means – the sampling distribution – would be normal (assuming the sample size is large enough).
    - The standard deviation of this sampling distribution is our standard error.

Demonstrating the CLT

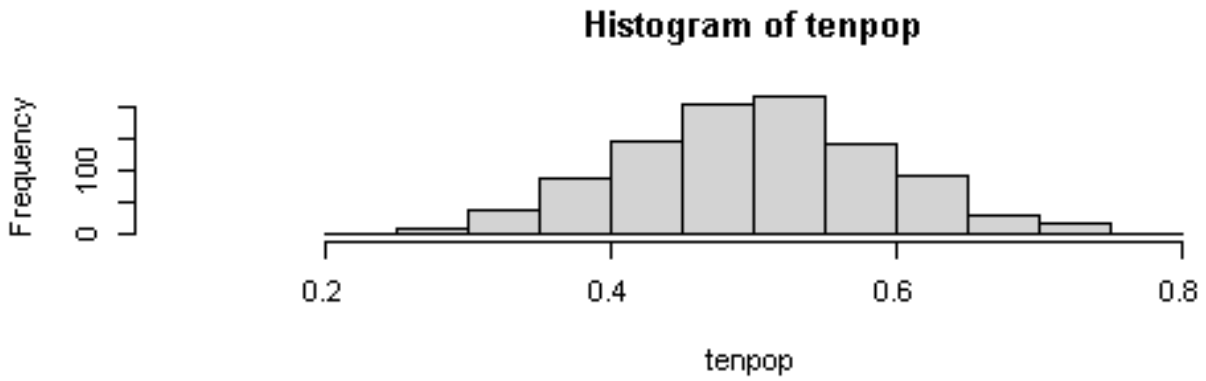
pop=runif(1000, 0,1)

hist(pop)

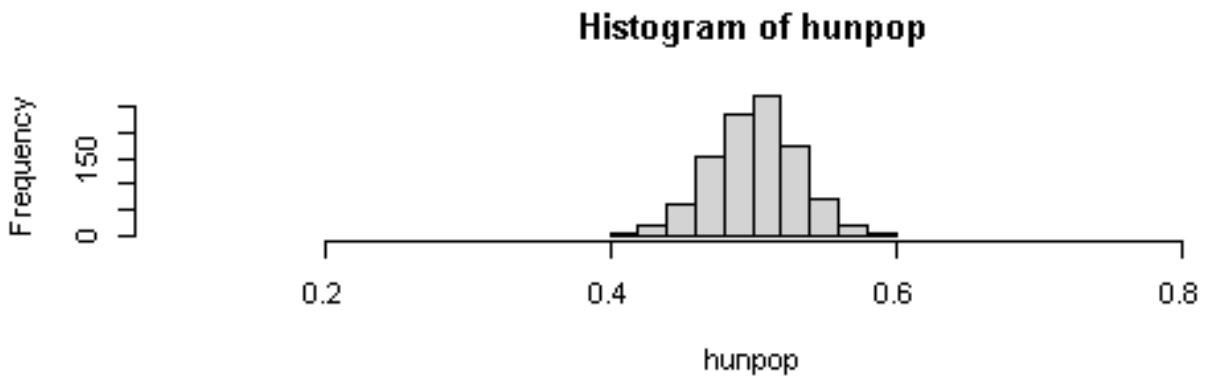




twopop=vector()
for (i in 1:1000) twopop[i]= mean(runif(**2**, 0,1))



tenpop=vector()
for (i in 1:1000) tenpop[i]= mean(runif(**10**, 0,1))



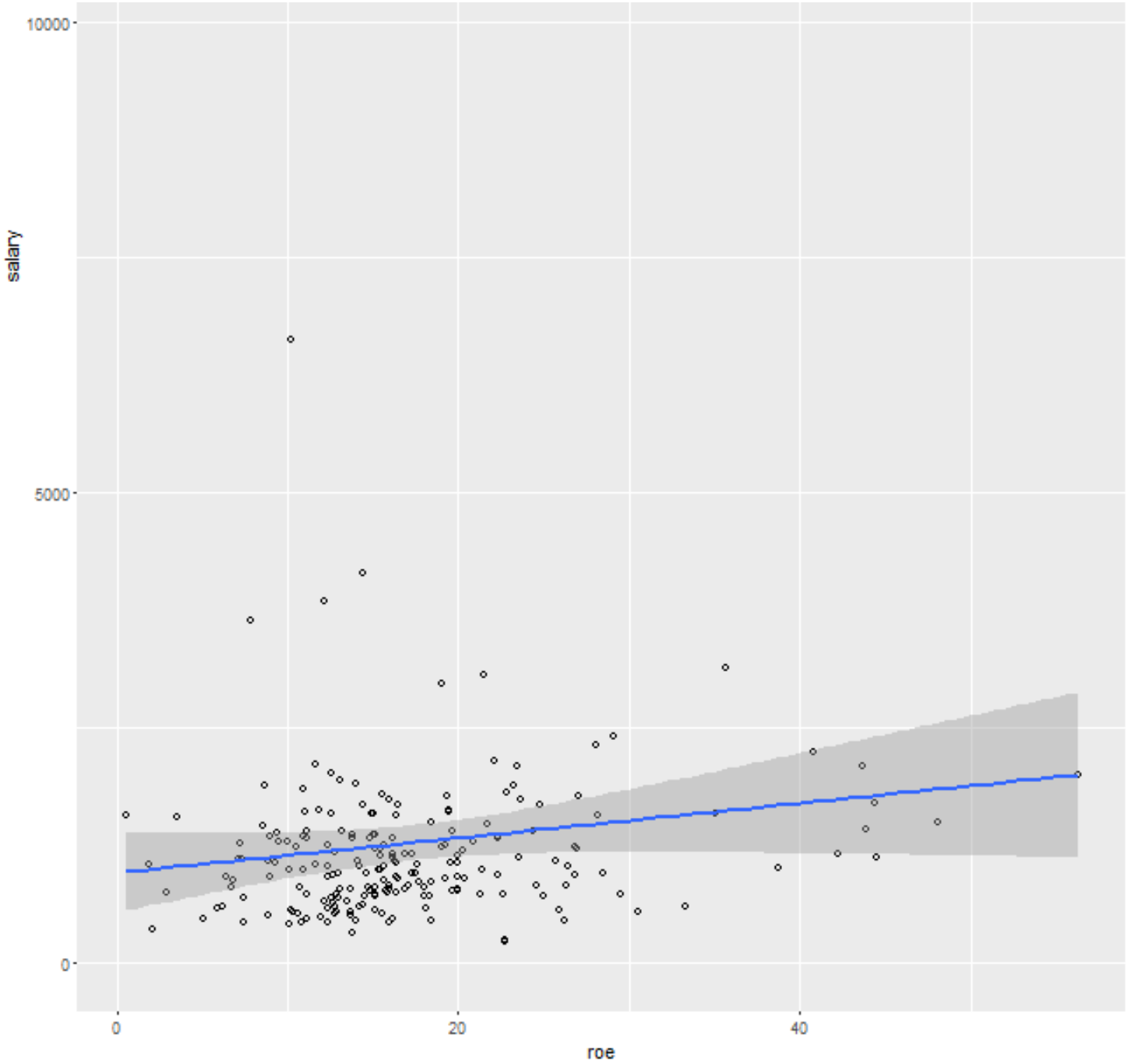
hunpop=vector()
for (i in 1:1000) hunpop[i]= mean(runif(**100**, 0,1))

Simple Regression Example

- What is the relationship between CEO salary and the company's average return on equity. Data consists of 209 observations on companies taken from Business Week in 1990.
  - Salary is measured in $1,000s.
  - ROE is defined in terms of net income as a percentage of common (shareholder) equity. Measured as a percent, not a decimal.

$$\widehat{salary} = \beta_0 + \beta_1 roe + \varepsilon$$

Dataset and discussion of data adapted from Wooldridge "Introductory Econometrics".



Looking at the Data Visually

Run a simple regression

```
> lm1=lm(salary ~ roe, data=ceo)
> summary(lm1)

Call:
lm(formula = salary ~ roe, data = ceo)

Residuals:
    Min      1Q  Median      3Q     Max
-1160.2  -526.0  -254.0   138.8 13499.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   963.19     213.24   4.517 1.05e-05 ***
roe            18.50      11.12   1.663   0.0978 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1367 on 207 degrees of freedom
Multiple R-squared: 0.01319,   Adjusted R-squared: 0.008421
F-statistic: 2.767 on 1 and 207 DF,  p-value: 0.09777
```

$$\hat{salary} = \beta_0 + \beta_1 roe + \varepsilon$$

$$\hat{salary} = 964.268 + 18.465 roe + \varepsilon$$

- What is the predicted salary for a CEO in a company with an ROE of 30%. 963.18 + 18.50(30) = 1,518.18 = $1,518,180
- Clearly, this does not mean that any CEO in a company with an ROE of 30 will earn that salary. This is simply our prediction given our model.
  - There may be many other things that can impact salary.

Model evaluation (Pardoe 2012)

- There are three standard and interrelated ways for evaluating numerically how well a simple linear regression fits our sample data.
- The methods can be categorized by the type of question they were designed to answer:
  1. Coefficient of determination (R^2): How much of the variability in Y have we been able to explain with our model?
  2. Residual or regression standard error: How close are the actual observed Y-values to the model-based fitted values?
  3. Slope parameter (and p-value): How strong is the evidence of a linear association between Y and X?

1. Coefficient of determination (R^2)

- If we have no predictor variables, then our best guess for predicting an individual Y-value is simply the mean.
- When this is our model, we can calculate the total amount of error in our model – one approach for doing this is the sum of squares total (SST).
- SST is simply a measure of the amount of variability in our dependent variable. It is this variability that we hope to explain through an improved model (i.e., one that just doesn't guess the mean).

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

Predicted Values and Residuals

- When we add predictor variables to our model, we are provided with a more informed prediction of a given Y-value.
- When we run a regression, we obtain a predicted or fitted value for each of our observations based on the sample of data we have collected.
  - By definition, each of our fitted values falls on our regression line – remember we are establishing a linear relationship between two variables.
- If the *residual* is negative, then the *observed* value falls below the regression line; if it is positive, the observed value falls above the regression line.

Predicted Values and Residuals cont…

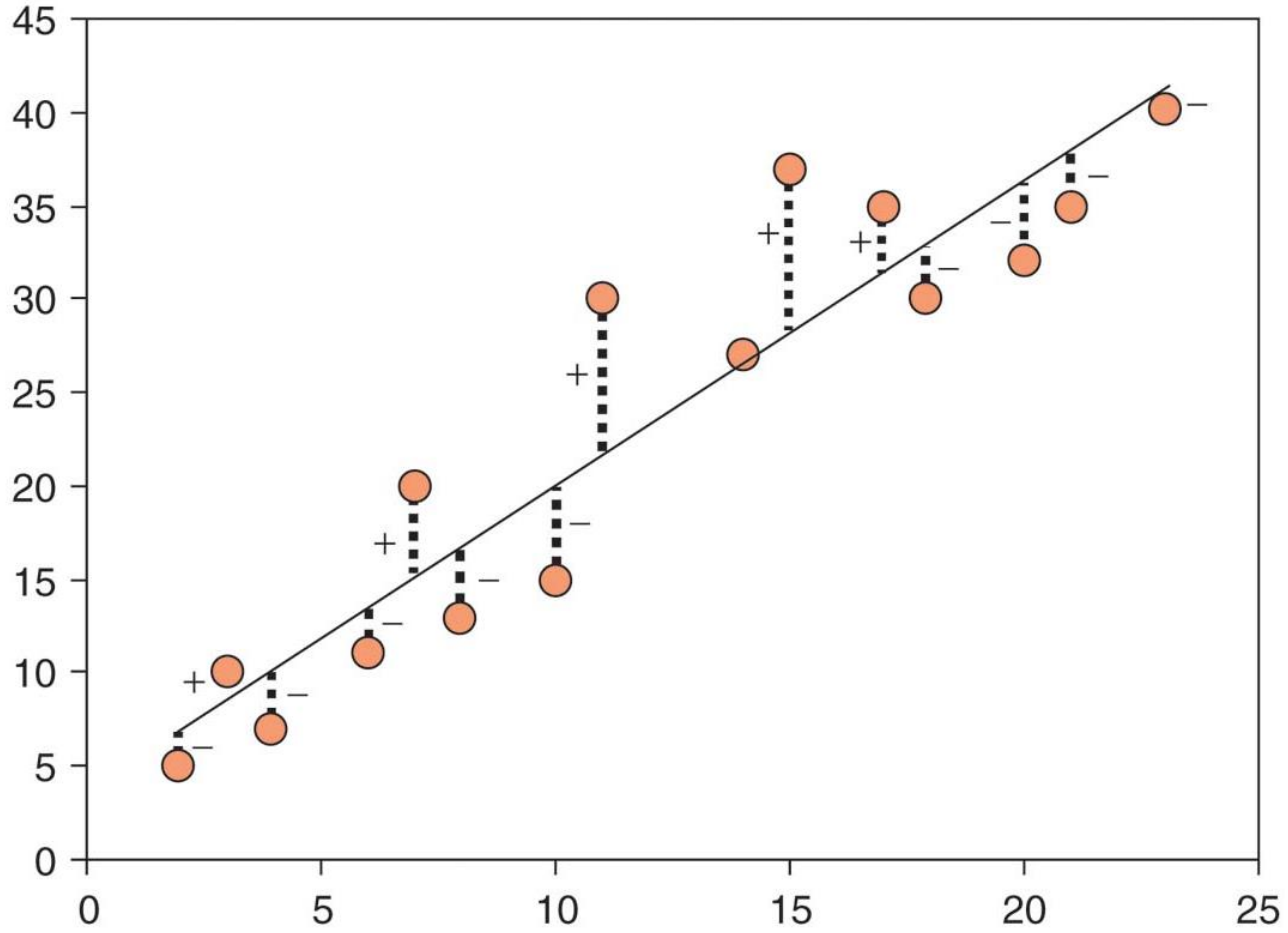


Diagram taken from Andy Field's "Discovering Statistics Using SPSS".

Breaking Down the Variation in The Dependent Variable

- Three measures of variation in our simple regression model:
  - **SST**: Total sum of squares
  - **SSM**: Model Sum of Squares (Explained Sum of Square)
  - **SSR**: Residual Sum of Squares (Error Sum of Squares)
  - SST = SSM + SSR

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

Total sum of squares, represents total variation in dependent variable

$$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Model/Explained sum of squares, represents variation explained by regression

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2$$

Residual sum of squares, represents variation not explained by regression

Goodness of Fit

- R-Squared of the regression is often referred to as the **coefficient of determination**.

- R^2 = SSM/SST or equivalently SST-SSR / SST

- R-squared is the ratio of explained variation to the total variation. In other words, it is the fraction of the sample variation in Y (salary) that can be explained by X (roe).

Looking at these values using the CEO data

```
> #SST or the total variation in salary is:
>
> SST = sum ( (ceo$salary - mean(ceo$salary))^2 )
> SST
[1] 391732982
>
> #SSR or the residual sum of squares, concerns the difference between
> #the predicted values from the model, and the observed values
>
> pred.y=predict(lm1)#predicted values from the model
> SSR=sum((ceo$salary - pred.y)^2)
> SSR
[1] 386566563
>
> #SSM - is what is explained by the model
> SSM = sum ( (pred.y - mean(ceo$salary))^2 )
> SSM
[1] 5166419
```

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

Total sum of squares, represents total variation in dependent variable

$$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Explained sum of squares, represents variation explained by regression

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2$$

Residual sum of squares, represents variation not explained by regression

R-Squared in our data set

```
> lm1=lm(salary ~ roe, data=ceo)
> summary(lm1)

Call:
lm(formula = salary ~ roe, data = ceo)

Residuals:
    Min      1Q  Median      3Q     Max
-1160.2  -526.0  -254.0   138.8 13499.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   963.19     213.24   4.517 1.05e-05 ***
roe            18.50      11.12   1.663   0.0978 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1367 on 207 degrees of freedom
Multiple R-squared: 0.01319,   Adjusted R-squared: 0.008421
F-statistic: 2.767 on 1 and 207 DF,  p-value: 0.09777
```

The R-square here is simply:

SSM/SST = 5166419/391732982

- Firms return on equity is only explaining 1.3% of the variation in salaries among the 209 CEOs. Thus, 98.7% of the variation remains unexplained.
- This lack of explanatory power is not surprising – many other characteristics of the firm and the CEO affect salary.
- Note: In the social sciences low R-squares are not uncommon – especially in cross sectional data.

- Another way to calculate the R^2, and where it derives its name, is the value is the square of the correlation between the fitted values and the actual values.
- Correlation is often indicated with an "*r*", so R^2 is simply the square of this value.

2. Residual standard error

```
> lm1=lm(salary ~ roe, data=ceo)
> summary(lm1)

Call:
lm(formula = salary ~ roe, data = ceo)

Residuals:
    Min      1Q  Median      3Q     Max
-1160.2  -526.0  -254.0   138.8 13499.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   963.19     213.24   4.517 1.05e-05 ***
roe            18.50      11.12   1.663   0.0978 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1367 on 207 degrees of freedom
Multiple R-squared: 0.01319,   Adjusted R-squared: 0.008421
F-statistic: 2.767 on 1 and 207 DF,  p-value: 0.09777
```

$$s = \sqrt{\frac{SSR}{n-p}}$$

- The residual standard error (or regression standard error or root mean squared error) is the standard deviation of the observed Y values from their fitted values.
- The value of *s*, the residual standard error, is in the same units of measurement as our dependent variable.
- We can calculate a rough prediction uncertainty - approximately 95% of the observed Y-values lie within plus or minus 2s of their fitted values.

3. Significance of slope parameters

```
> lm1=lm(salary ~ roe, data=ceo)
> summary(lm1)

Call:
lm(formula = salary ~ roe, data = ceo)

Residuals:
    Min      1Q  Median      3Q     Max
-1160.2  -526.0  -254.0   138.8 13499.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   963.19     213.24   4.517 1.05e-05 ***
roe            18.50      11.12   1.663   0.0978 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1367 on 207 degrees of freedom
Multiple R-squared: 0.01319,   Adjusted R-squared: 0.008421
F-statistic: 2.767 on 1 and 207 DF,  p-value: 0.09777
```

$$t = \frac{\beta_{observed} - \beta_{expected}}{SE_{\beta}}$$

- If a predictor variable significantly predicts the dependent variable than it should have a coefficient that is significantly different from zero. This hypothesis is tested using a t-test as noted by the output.
- The t-statistic tests the null hypothesis that the value of the coefficient is zero; therefore, if it is significant, we conclude that it contributes to our ability to estimate the outcome.
- How do we interpret the intercept and slope here?

Precision of the slope estimates

- The variance of the slope estimate, $\widehat{\beta}_1$, is the width of the $\hat{\beta}_1$ distribution. When certain regression assumptions are met the variance is calculated as:

$$var(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{N\ X\ var(X)}$$

- Where:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N-k}$$

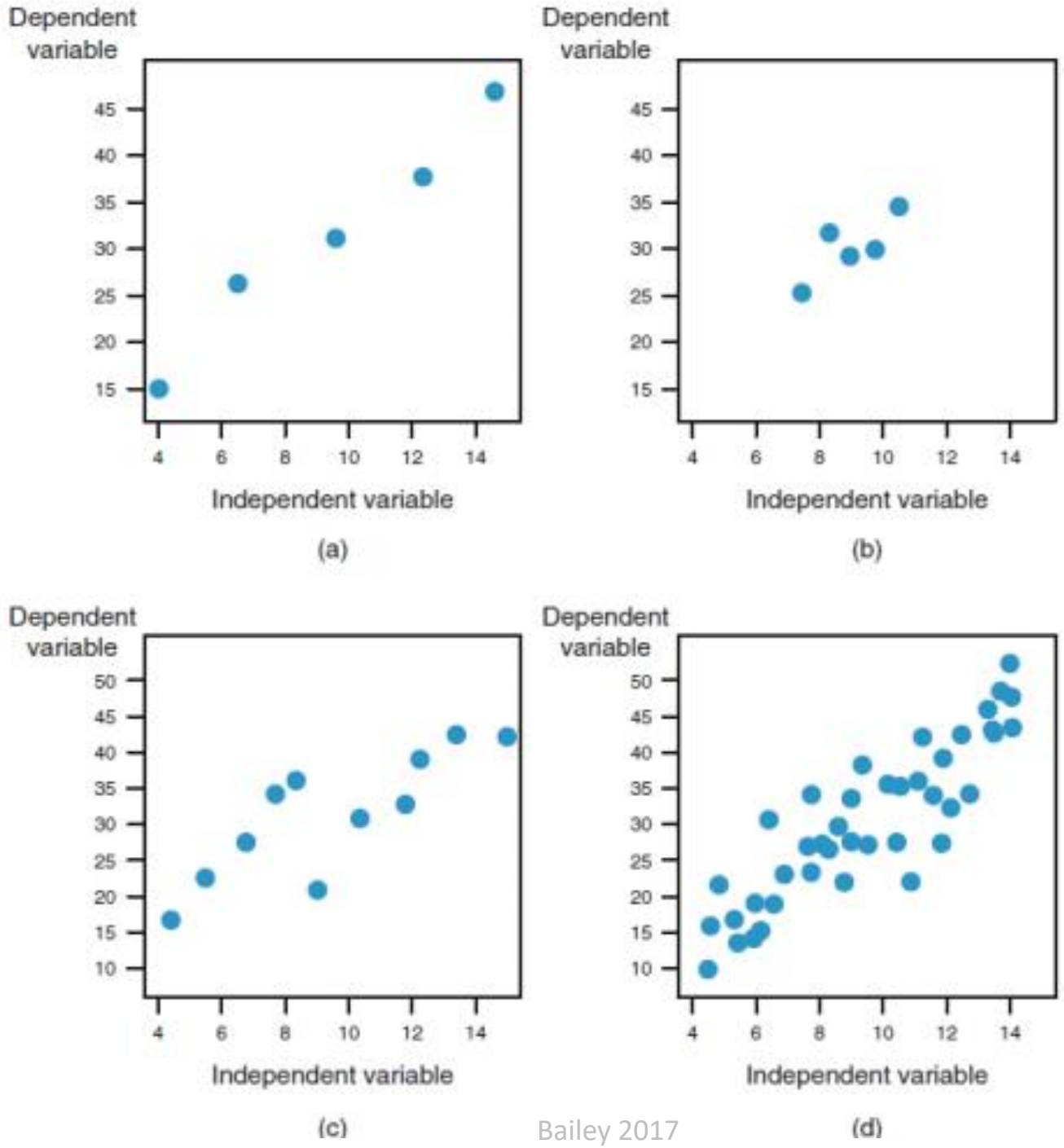
Precision of the slope estimates

- Alternative formula:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x}$$

$$var(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{N\ X\ var(X)}$$

- Will the variance of $\hat{\beta}_1$ be smaller in panel (a) or panel (b)? Why?
- Will the variance of $\hat{\beta}_1$ be smaller in panel (c) or panel (d)? Why?



Bailey 2017

Distribution of coefficient estimates

- The distribution of $\widehat{\beta_1}$ will be normally distributed if either

a) Sample size is large
   - Central Limit Theorem: the mean of a sufficient number of independent draws from any distribution will be normally distributed.
   - OLS estimates are weighted averages of Y, which implies $\widehat{\beta_1}$ will be normally distributed

b) Errors are normally distributed

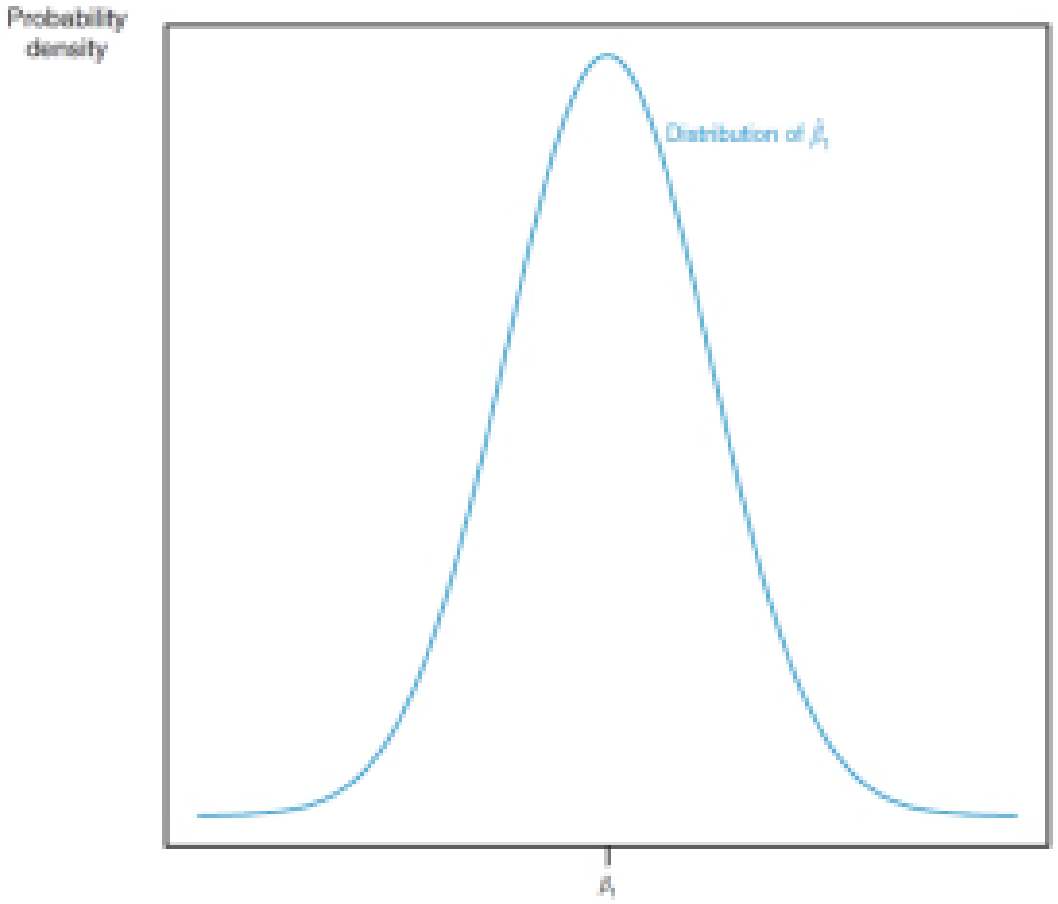


FIGURE 3.5: Distribution of $\hat{\beta}_1$ Bailey 2017

A quick note on linear regression

- In a linear model the parameters enter linearly – but the predictors themselves do not have to have a linear relationship. For example, the following is a linear model:

- $Y = \beta_0 + \beta_1 X_1^2 + \beta_2 \log X_2 + \beta_3 X_1 X_2 + \varepsilon$

- Some relationships can be transformed to linearity.
- While linear models may seem restrictive, they are quite flexible.
  - The predictors themselves can be transformed and combined in any way.
  - Linear models can handle complex datasets.
  - Linear models can also be curved (the relationship between X and Y does not need to be a straight line).

- We will talk more about transformations and their interpretations later this semester.

Simple Regression Assumptions

- Assumptions:
  - SLR.1 – Linear in parameters
  - SLR.2 – Data drawn from a random sample (i.e., the errors are independent - no autocorrelation in the data)
  - SLR.3 – Sample variation in the explanatory variable
  - SLR.4 – Zero conditional mean for the error term (most critical assumption of all and cannot be tested)
  - SLR.5 – Homoskedasticity (i.e., the errors have equal variance)

Standard assumptions for the linear regression model

- **Assumption SLR.1 (Linear in parameters)**

$$y = \beta_0 + \beta_1 x + u$$

In the population, the relationship between y and x is linear

- **Assumption SLR.2 (Random sampling)**

$$\{(x_i, y_i) : \ i = 1, \ldots n\}$$

The data is a random sample drawn from the population

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Each data point therefore follows the population equation

Assumptions for the linear regression model (cont.)

- **Assumption SLR.3 (Sample variation in explanatory variable)**

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$$

The values of the explanatory variables are not all the same (otherwise it would be impossible to study how different values of the explanatory variable lead to different values of the dependent variable)

- **Assumption SLR.4 (Zero conditional mean)**

$$E(u_i|x_i) = 0$$

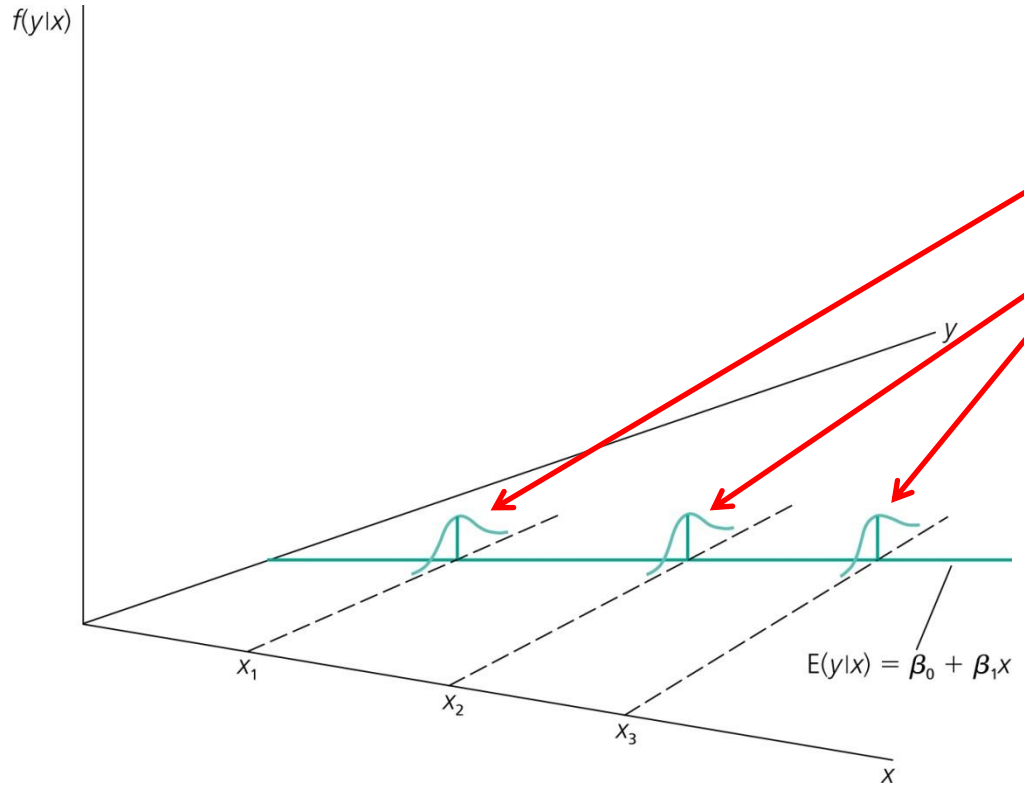
The value of the explanatory variable must contain no information about the mean of the unobserved factors

Assumption SLR.5 (Homoskedasticity)

$$Var(u_i|x_i) = \sigma^2$$

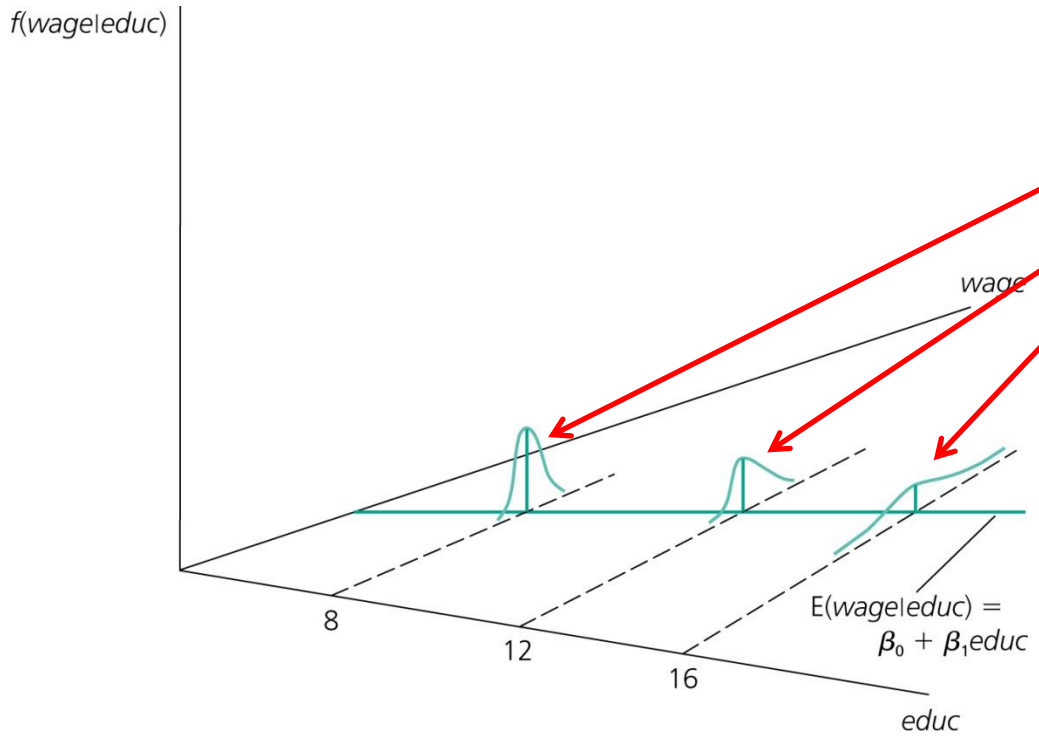
The value of the explanatory variable must contain no information about the variability of the unobserved factors

Graphical illustration of homoskedasticity



The variability of the unobserved influences does not depend on the value of the explanatory variable

An example for heteroskedasticity: Wage and education



The variance of the unobserved determinants of wages increases with the level of education

Two additional assumptions – validity and normality

- According to Gelman and Hill (2007) the most important assumption of regression modeling is **validity.**
- This means that the data you are analyzing should map to the research question you are trying to answer.
  - This sounds obvious but is often overlooked or ignored because it can be inconvenient. . . .

Normality

- It is often assumed that our *dependent variable* needs to be normally distributed. This is not true. We want our *errors* to be normally distributed.
  - Note, we never actually see these errors as will be discussed below.
- We can check and see if our *residuals* around the regression line are normally distributed (Cohen et al. 2003, p. 120).
  - Violations of the normality assumption do not lead to bias estimates of the regression coefficients.
  - The effect of violation of normality on significance tests and confidence intervals depends on the sample size, with problems occurring in small samples.
  - In large samples, nonnormality of the residuals does not lead to serious problems with the interpretation of either significance tests or confidence intervals.
    - However, nonnormal residuals are often an important sign of other problems in the

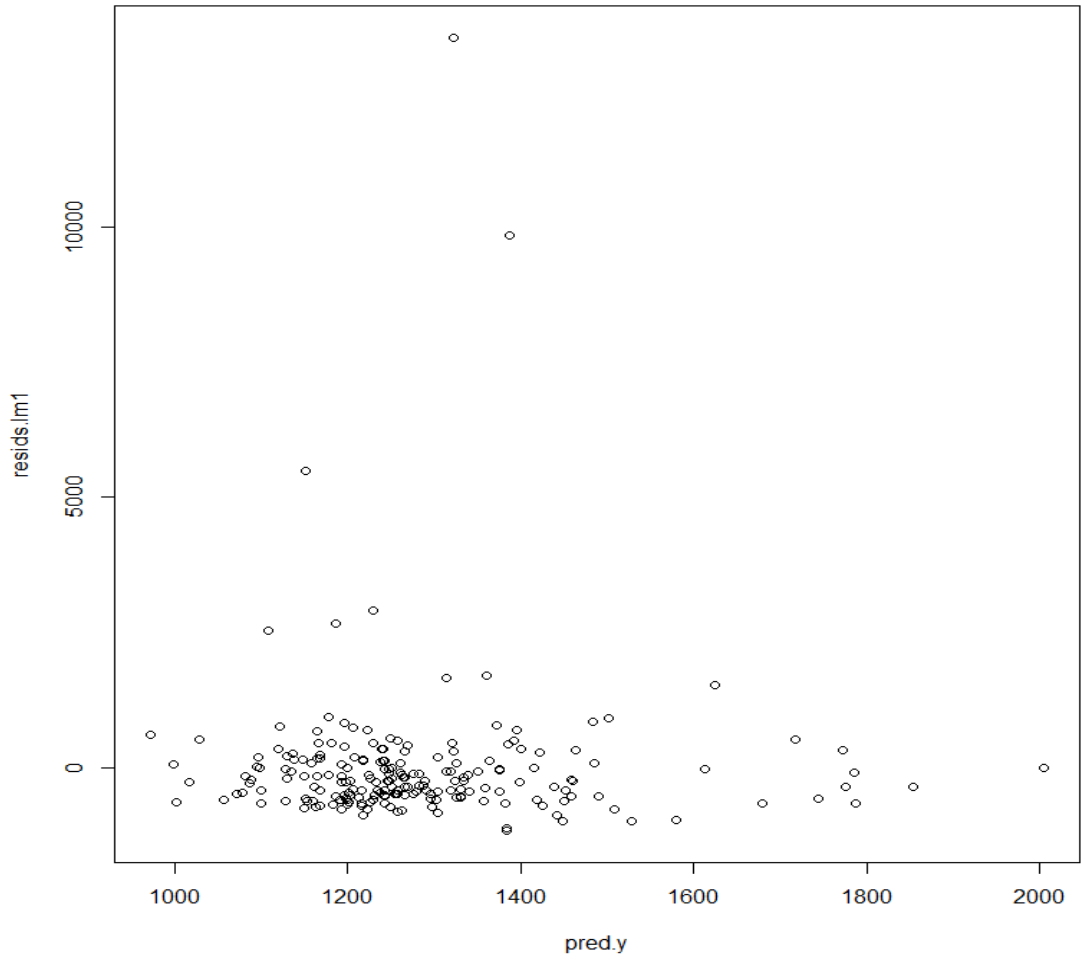
Visually inspecting regression assumptions

Visually inspecting regression assumptions

- Through visualization we can inspect 4 key assumptions about the random errors in our model:
  - SLR.1 - linearity
  - SLR.2 – Data drawn from a random sample (i.e., the errors are independent - no serial correlation in the data)
  - SLR.5 – Homoskedasticity (i.e., the errors have equal variance)
  - Normality
- SLR.4 is about variables in the error term that we do not observe and SLR.3 concerns that there is actual variance in our predictor variable.

```
lm1=lm(salary ~ roe, data=ceo)
> pred.y=predict(lm1) #obtain the predicted values or fitted values
> resids.lm1 = resid(lm1) #obtain the residuals
> plot(resids.lm1 ~ pred.y) #plot resids versus predicted values
```

- We can test the first three via a residuals versus fitted values plot.



Assumption Testing – What we are looking for from residual plot

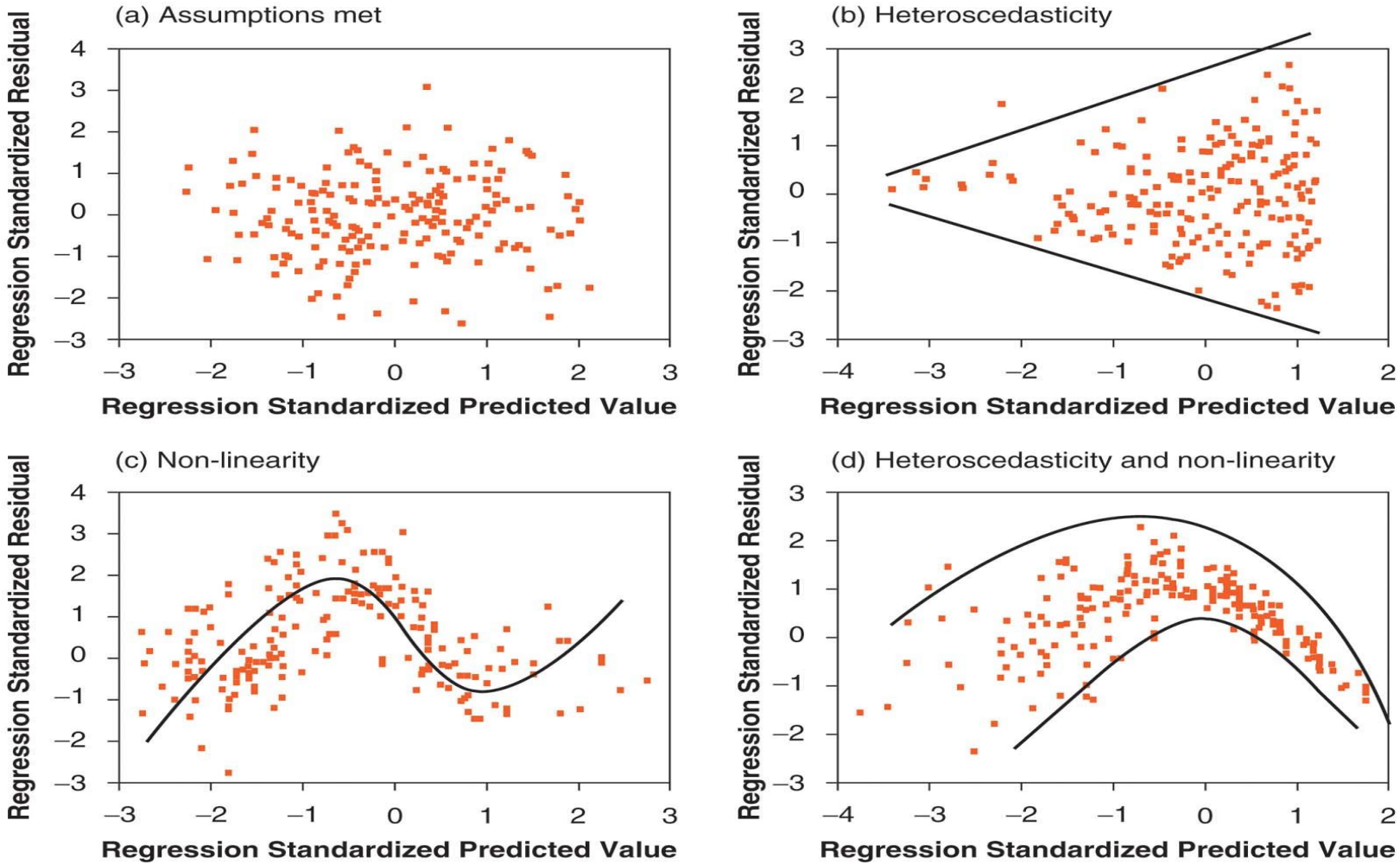
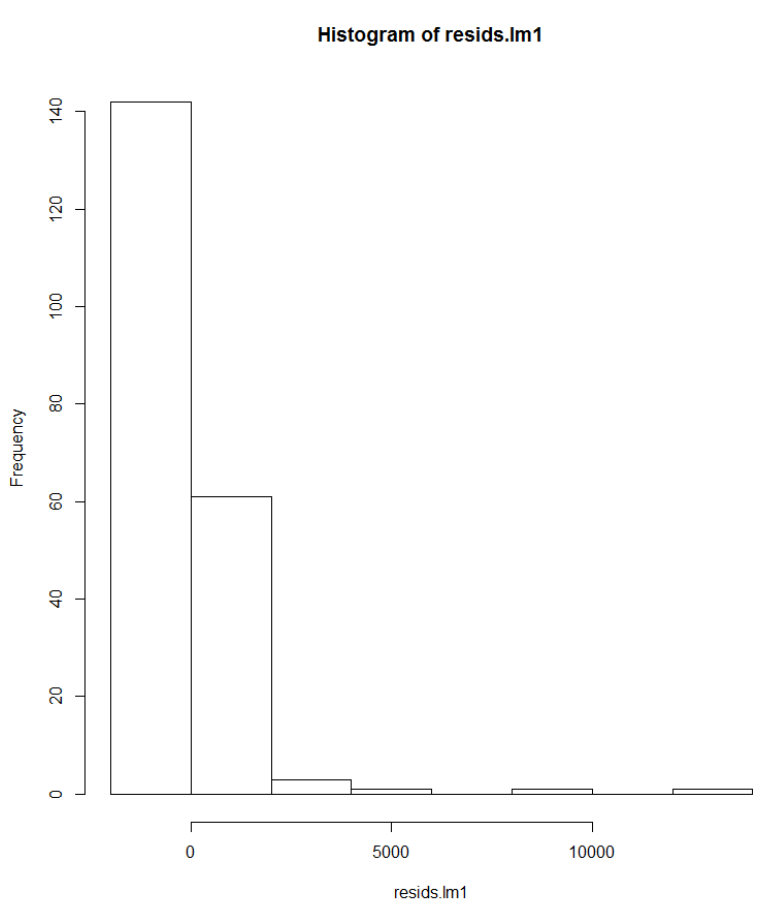
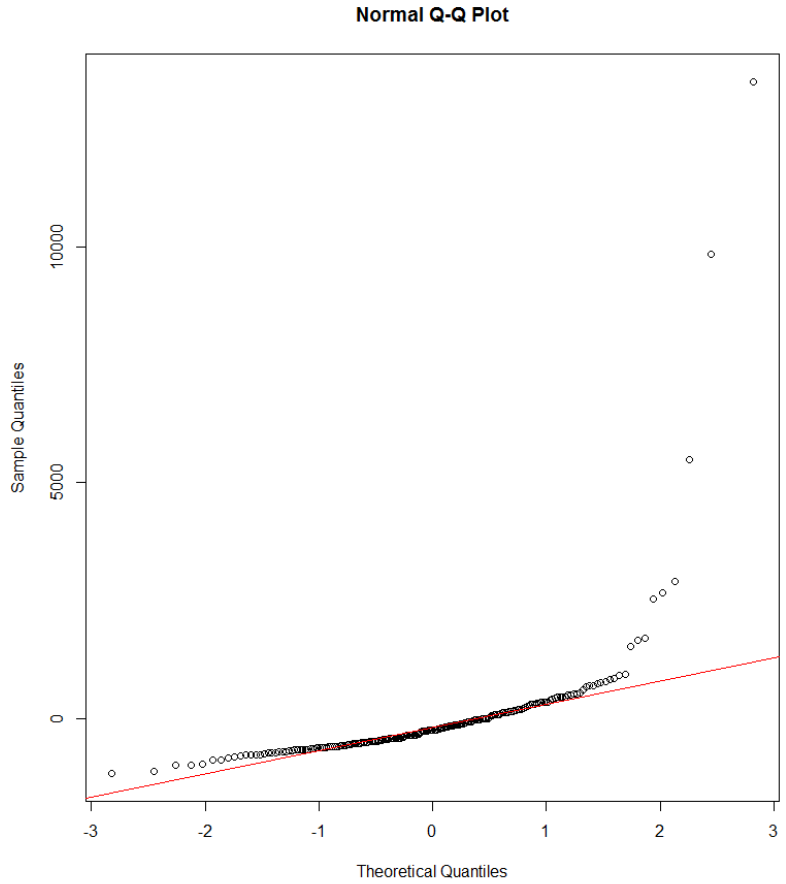


Diagram taken from Andy Field's "Discovering Statistics Using SPSS".

```
qqnorm(resids.lm1)
> qqline(resids.lm1, col=2)#col = 2 just makes the line r
```

- Normality is best assessed via a histogram of the residual and through a QQ-plot.





In Class Exercise

- Open the data set ceosal2.RData. The variable salary is the annual compensation of CEOs in thousands of dollars and ceoten is the prior number of years a CEO has been with the firm.
  - Find the average salary and the average tenure (years with firm) in the sample.
  - How many CEOs are in their first year as CEO (that is ceoten = 0)? What is the longest tenure?
  - Estimate the simple regression model:
    $Salary = B_0 + B_1(ceoten_i) + e_i$
    What is the approximate increase in salary given one more year as a CEO.
  - What is the r-squared for this model – how do you interpret it?
  - What is the predicted salary of an individual who has 0 years with a firm; what is the predicted salary for an individual who has been with a firm for 10 years?
  - Create a residual versus predicted values plot – do you have any concerns regarding the regression assumptions?