Session 1

Decision Sciences 6203

Statistics for Analytics I

Philip W. Wirtz
The George Washington University

Introduction to Blackboard Collaborate Ultra

Statistics for Analytics 1: Formalities and Expectations

- Class schedule: Wednesdays 4:30pm-7:00pm, 10/28->12/9 (NOTE: All times listed are Eastern US time)
  - No class 11/25 (Thanksgiving break); ***Makeup class Friday 11/20 (ADMINISTRATIVELY MANDATED!!!)***
  - Please download transparencies from Blackboard: have available during class sessions
    - Transparencies are released at midnight on the night prior to class
  - If we can afford the time, there will be a 5-minute break around 5:45pm
  - Sessions will be small part didactic, large part socratic. ***Highly*** interactive class. I will expect you to actively participate in class
- Virtual Office Hours: VOH 1 Fridays (8:30am), VOH 2 Mondays (8:30pm), VOH 3 Tuesdays (8:30am)
  - Attendance highly recommended. Quiz at beginning of Tuesday VOH session.
- All class sessions and VOH sessions are recorded and placed on Blackboard. All material presented/presented in VOH sessions is “fair game” for quizzes and examination
- 5 Weekly assignments, on Blackboard,
  - Assignments MUST be submitted by 4:25pm on the day of next class session
  - Focus is on R; occasionally, Excel
    - ***You must have a working knowledge of R prior to entering this course***
  - Submission protocol a little different from 6206: Word *templates*
  - Grading is numeric: 11=A, 10=A-, 9=B+, 8=B, 7=B-, 6=C+, 5=C, 4=C-, 0=F
  - PIN assigned, Master Keys uploaded
- Academic Integrity Policy
- Quizzes each week
  - Missed quiz policy: No makeups or rescheduling. If you are not present for a quiz, it increases your examination weight.
  - GWIDs − not names − on all quizzes
- If you have not yet completed the Student Information Record or uploaded your “headshot”, **now** is the time!
- Office Hours: By appointment (pww@gwu.edu)
- Final examination (Provisional): Wednesday 12/16 ***Regular Class Time***

A Preview of Today's Discussion

- **In Business Analytics, when we examine the relationship between variables, it is important to assess BOTH *statistical significance* and *relationship strength*.**

  - **Statistical tests that we will discuss (such as *F* and *t*, and the probability value (*p*-value)) all pertain to statistical significance;**
  - **We will also discuss today the Coefficient of Determination (also known as r^2 in a regression context), which is a measure of relationship strength.**

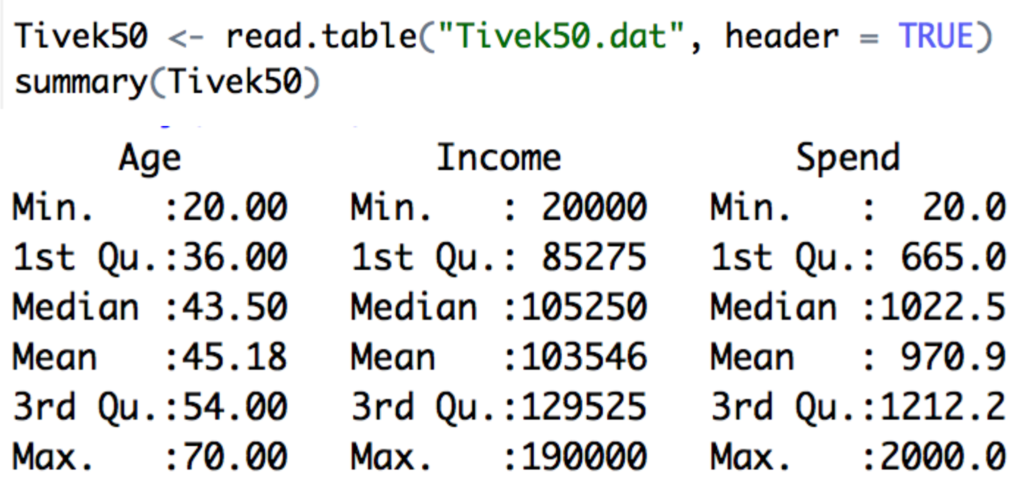
- **Measures of statistical significance address the question of whether ANY relationship exists between the variables; measures of relationship strength address the question of how strong the relationship is.**

- **The accuracy of both types of measures depend on sample size; even miniscule relationships can be "statistically significant" in the context of a large sample size, and even large Coefficients of Determination can be statistically non-significant in the context of a small sample size.**

- **Particularly when dealing with "big data", where nearly every relationship will be statistically significant, it is vitally important to assess the strength of the relationship as well as its statistical significance.**

Today's Case Study: Tivek, Inc.

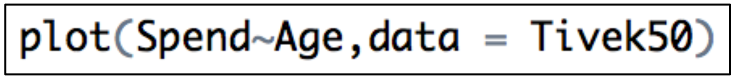
- Tivek, Inc.: retail cosmetics mail order firm
- Data collected from loyal (>1 year) customers
  - We will consider three samples: Tivek50.dat, Tivek200.dat, and Tivek2000.dat
  - Data are available on Blackboard (Outline/Session 1)
- Collect: age, income (…we will focus initially on age…), and amount spent by the customer on Tivek products in the past year
- Objective: Create linear regression model to predict annual "Spend" (i.e., amount spent by customers on Tivek products)
- Question: Why would the Tivek CEO be interested in this objective? What action would he/she take in response to the results?
- First foray: sample of 50 customers. Do the data "smell" right?



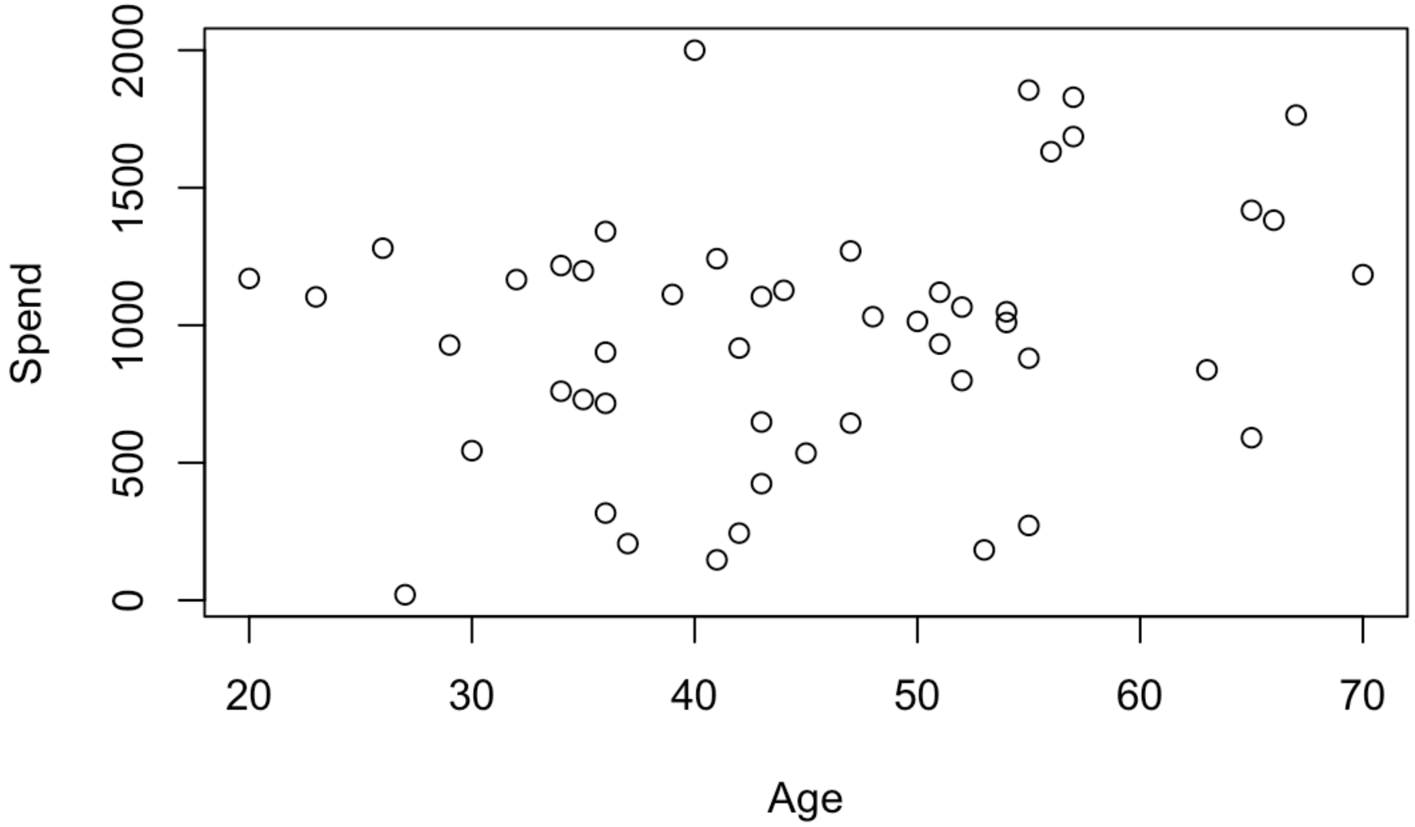
```r
Tivek50 <- read.table("Tivek50.dat", header = TRUE)
summary(Tivek50)
```

```
      Age             Income             Spend       
 Min.   :20.00   Min.   : 20000   Min.   :  20.0  
 1st Qu.:36.00   1st Qu.: 85275   1st Qu.: 665.0  
 Median :43.50   Median :105250   Median :1022.5  
 Mean   :45.18   Mean   :103546   Mean   : 970.9  
 3rd Qu.:54.00   3rd Qu.:129525   3rd Qu.:1212.2  
 Max.   :70.00   Max.   :190000   Max.   :2000.0  
```

Bivariate Relationship: Spend with Age



```
plot(Spend~Age,data = Tivek50)
```



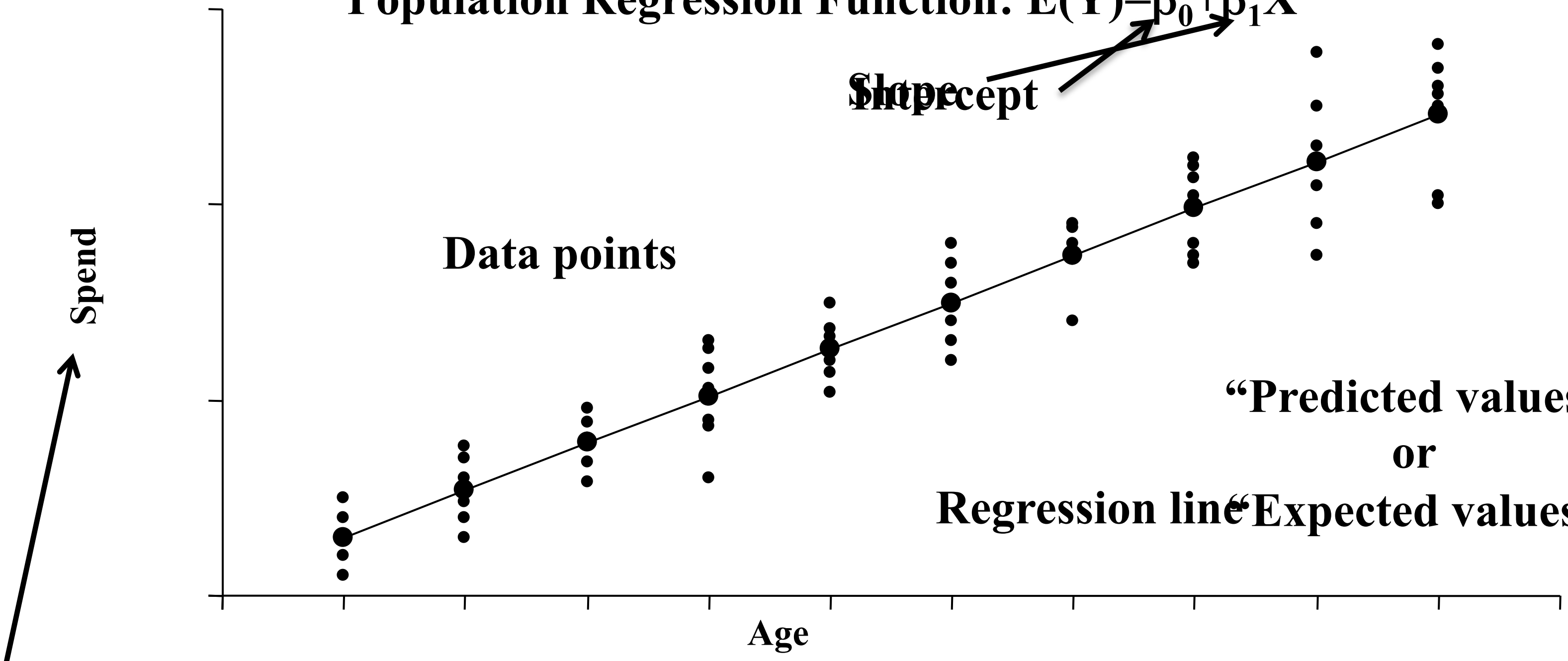
Any preliminary observations regarding the CEO’s objectives?

Simple Linear Regression

Population Regression Function: $E(Y)=\beta_0+\beta_1X$

Slope

Intercept



Data points

Regression line

“Predicted values”
or
“Expected values”

“Dependent variable” or “criterion” (Y)

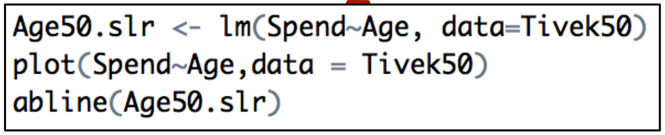
“Independent variable” or “predictor” (X)

Interpretation of the Intercept:
“Expected value of Y when X is exactly equal to zero”

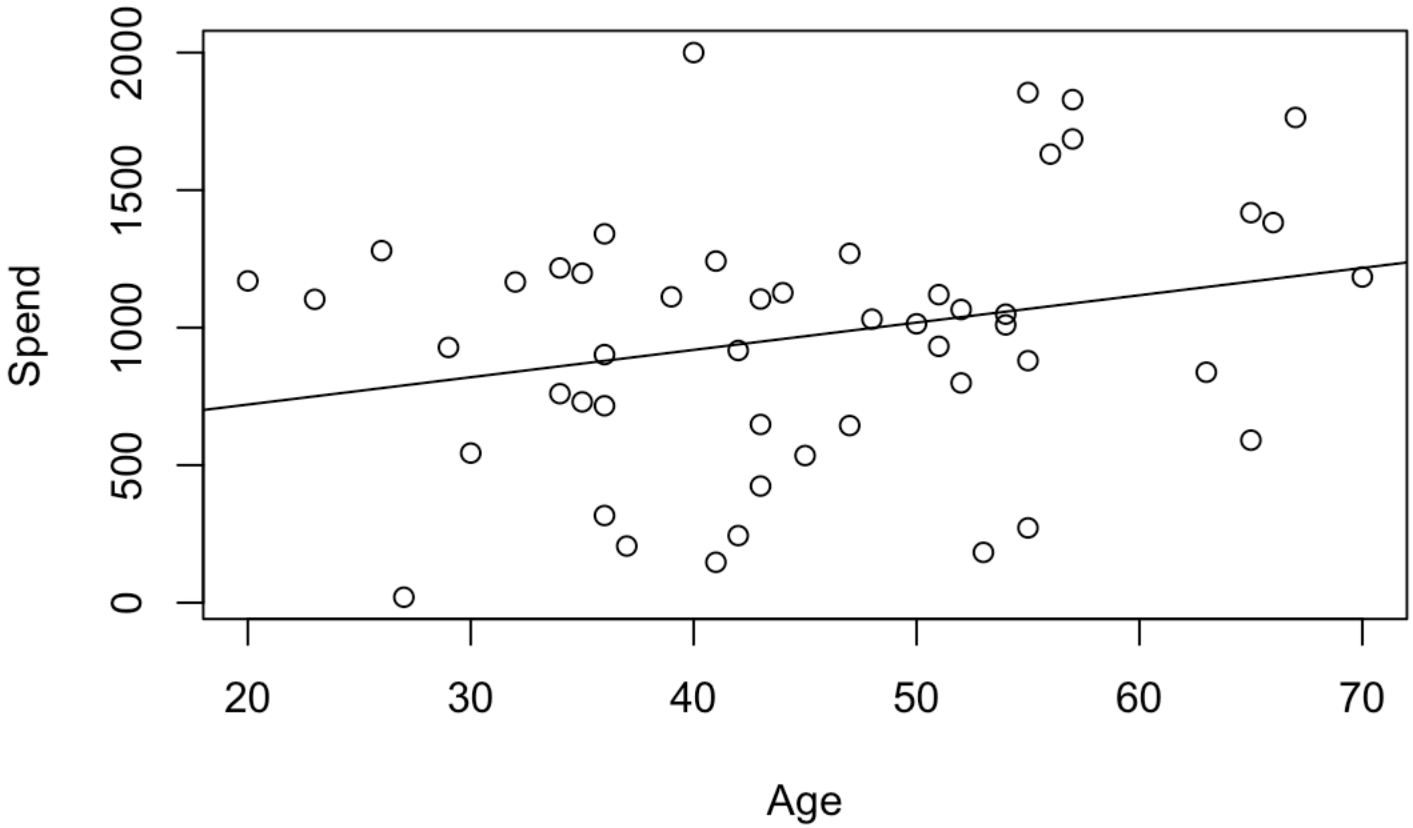
Interpretation of the Slope:
“Expected number of units of change in Y when X increases by one unit”

The Regression Line in Simple Linear Regression (Age)

Produces best estimates of intercept and slope

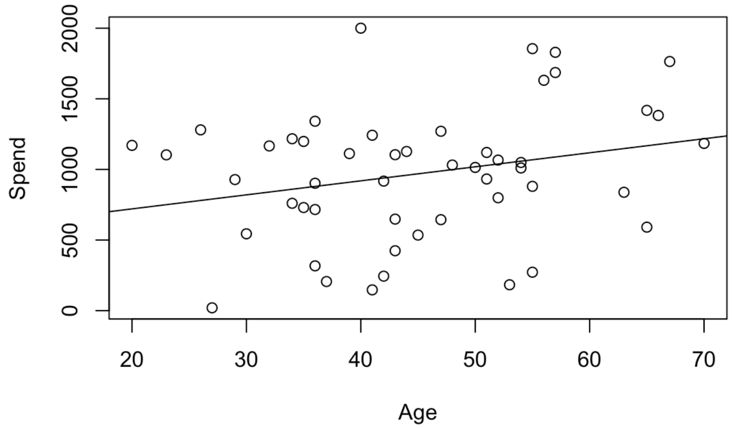


```
Age50.slr <- lm(Spend~Age, data=Tivek50)
plot(Spend~Age,data = Tivek50)
abline(Age50.slr)
```

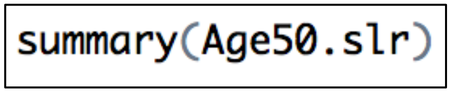


Are we more interested in the intercept, in the slope, or in both? Why?

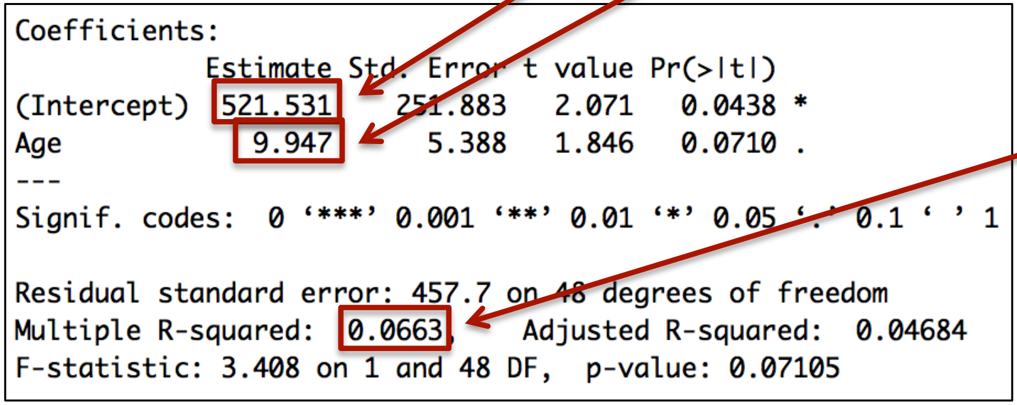
Predicting Spend From Age: The Regression Equation and an Assessment of the Strength of the Relationship



What is the “independent variable”?
What is the “dependent variable”?
The “Regression equation”:
E(Spend) = 521.531 + 9.947 * Age
What is the interpretation of the intercept?
What is the interpretation of the slope?
What is the meaning of “residual”?



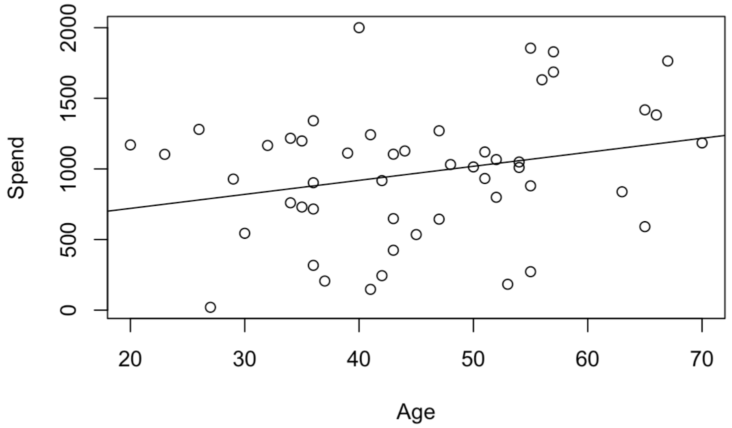
```
summary(Age50.slr)
```

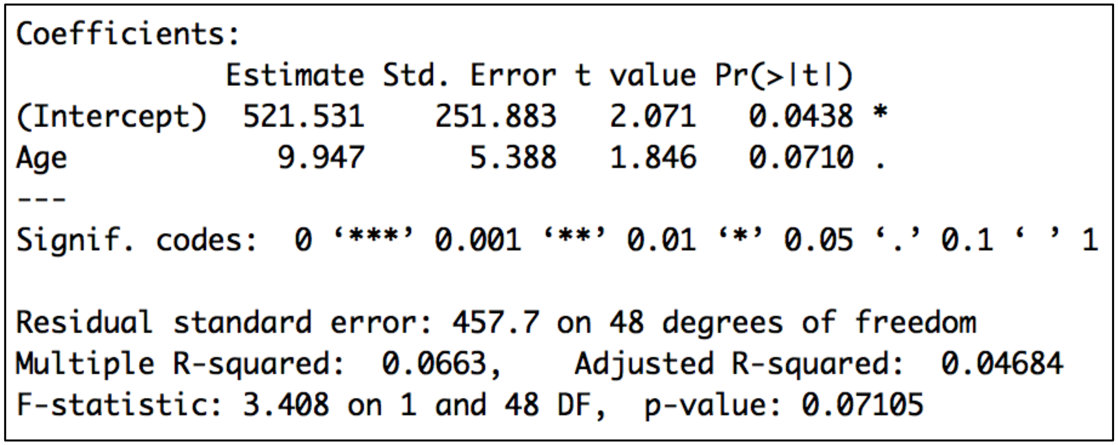


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  521.531    251.883   2.071   0.0438 *
Age            9.947      5.388   1.846   0.0710 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 457.7 on 48 degrees of freedom
Multiple R-squared: 0.0663,	Adjusted R-squared: 0.04684
F-statistic: 3.408 on 1 and 48 DF,  p-value: 0.07105
```

Strength of relationship:
Coefficient of Determination, a.k.a. (Multiple) R-squared

If the Data Comprise the Population …

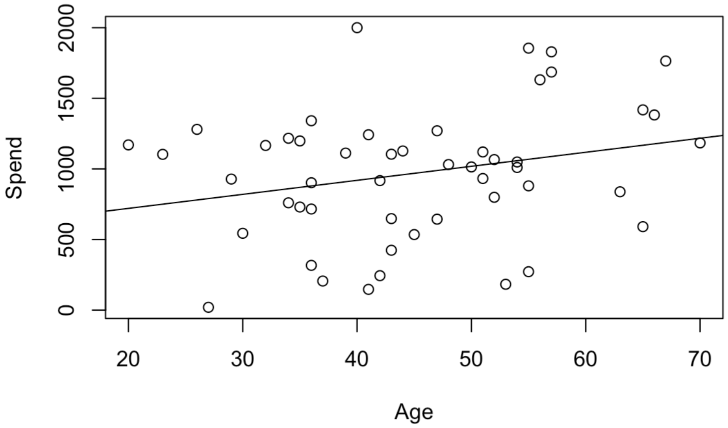


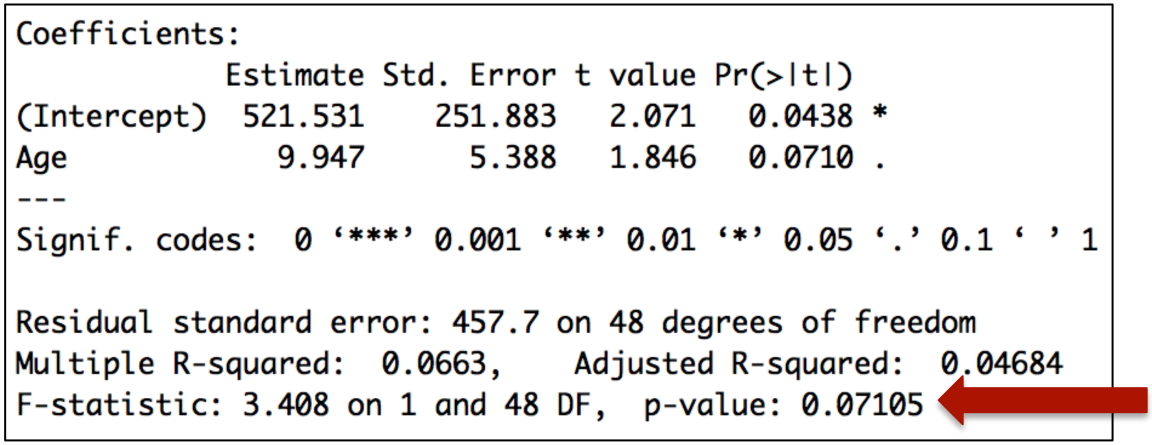


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  521.531    251.883   2.071   0.0438 *
Age            9.947      5.388   1.846   0.0710 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 457.7 on 48 degrees of freedom
Multiple R-squared:  0.0663,	Adjusted R-squared:  0.04684 
F-statistic: 3.408 on 1 and 48 DF,  p-value: 0.07105
```

- **Suppose these 50 individuals represented the entire population. What conclusions would you reach? What would you tell the CEO of Tivek?**
- **What implications do these results have that would be of interest to the CEO of Tivek?**
- **Would “statistical significance” be part of your report to the CEO?**

If the Data Comprise a Sample …





```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  521.531    251.883   2.071   0.0438 *
Age            9.947      5.388   1.846   0.0710 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 457.7 on 48 degrees of freedom
Multiple R-squared:  0.0663,    Adjusted R-squared:  0.04684
F-statistic: 3.408 on 1 and 48 DF,  p-value: 0.07105
```

Now, suppose these 50 individuals represented a random sample. Would your conclusions differ?

What implications do these results have that would be of interest to the CEO of Tivek? What would you tell him/her?

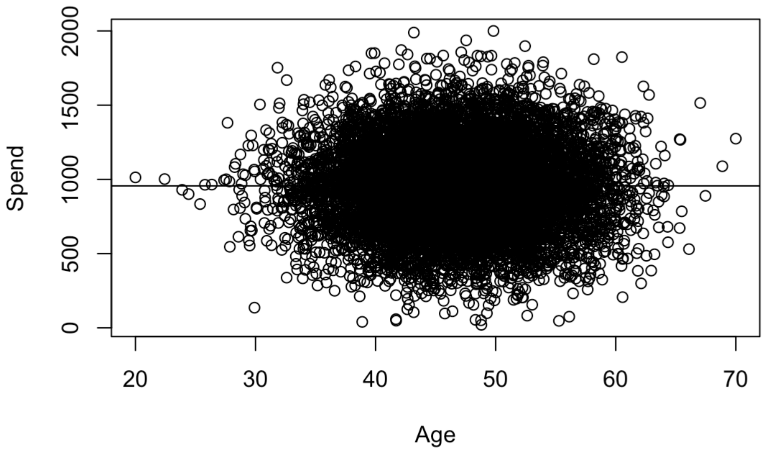
Would “statistical significance” play a role in your report?

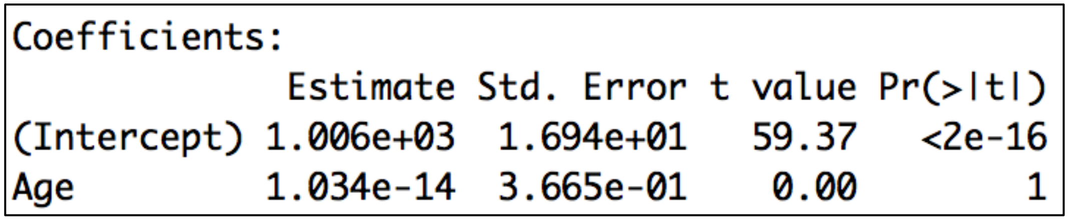
What is meant by the term “statistical significance”?

A Momentary Tangent: Random Sampling
Some Obvious (and Not So Obvious) Questions

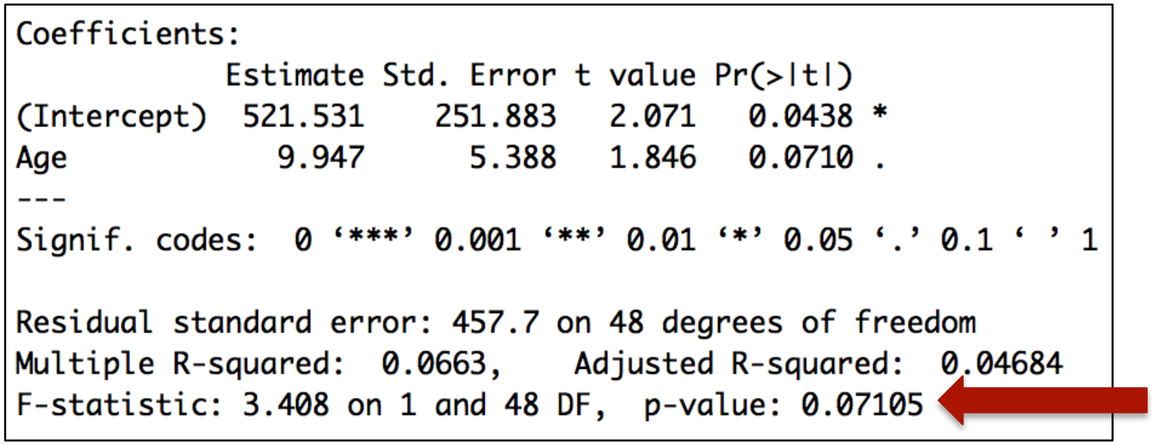
- **What is a “sample”?**
- **What makes a sample “random”?**
- **Why is it important for a sample to be “random”?**
- **Common forms of random sampling**
  - **Simple random sampling**
  - **Stratified random sampling**
  - **Clustered random sampling**
  - **Multi-stage (hybrid) random sampling**
- **What is the minimum number of units needed for a sample to be considered “random”?**
- **If you have a large database (i.e., “big data”), do you have to worry about the distinction between “samples” and “populations”?**
- **Is there a lower bound on the size of the sample needed in order for the sample to essentially replicate the characteristics of the population?**

The “Null” Case: Population Data





```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.006e+03  1.694e+01   59.37   <2e-16
Age         1.034e-14  3.665e-01    0.00        1
```



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  521.531    251.883   2.071   0.0438 *
Age            9.947      5.388   1.846   0.0710 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 457.7 on 48 degrees of freedom
Multiple R-squared:  0.0663,    Adjusted R-squared:  0.04684
F-statistic: 3.408 on 1 and 48 DF,  p-value: 0.07105
```

Suppose you had data from the entire population, and the results looked like this. What would you tell the CEO of Tivek? What implications would these results have that would be of interest to the CEO of Tivek?

Based on our random sample of 50, how sure are you that this is NOT the case for the population from which OUR sample was drawn??

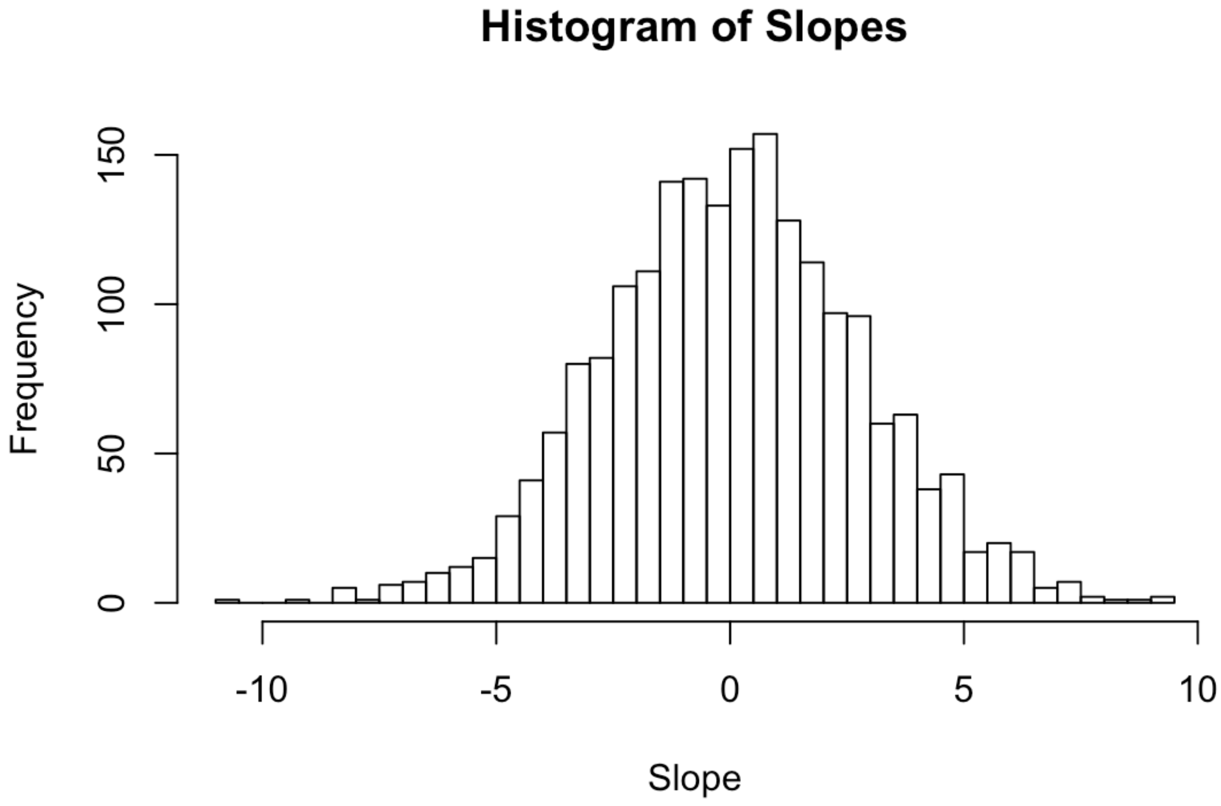
Formalizing the “Statistical Significance” Question: The Null Hypothesis and The Alternative Hypothesis

- Population Regression Equation: $Y = \beta_0 + \beta_1 * X + \varepsilon$
- Sample Regression Equation: $Y = b_0 + b_1 * X + e$
- Null Hypothesis (H_0): $\beta_1 = 0$
  The best-fitting regression line in the population has a slope of zero
- Alternative Hypothesis (H_A): $\beta_1 \neq 0$
  The best-fitting regression line in the population has a non-zero slope
- Relevant statistical theory: Statistical theory states that, given certain assumptions[1], when H_0 is true, the sampling distribution of the slope will be approximately normally distributed, with estimatable mean and standard error. This allows us to estimate seeing a slope as far from zero (or farther) as the slope from our sample when H_0 is true.
- Reject H0: If our sample slope is unlikely when H_0 is true, H_0 is likely to be false; we reject H_0
- Fail to reject H0: If our sample slope is NOT unlikely when H_0 is true, we FAIL TO reject H_0

[1] Assumptions are to be presented subsequently

The Sampling Distribution of the Slope (Empirically Derived)

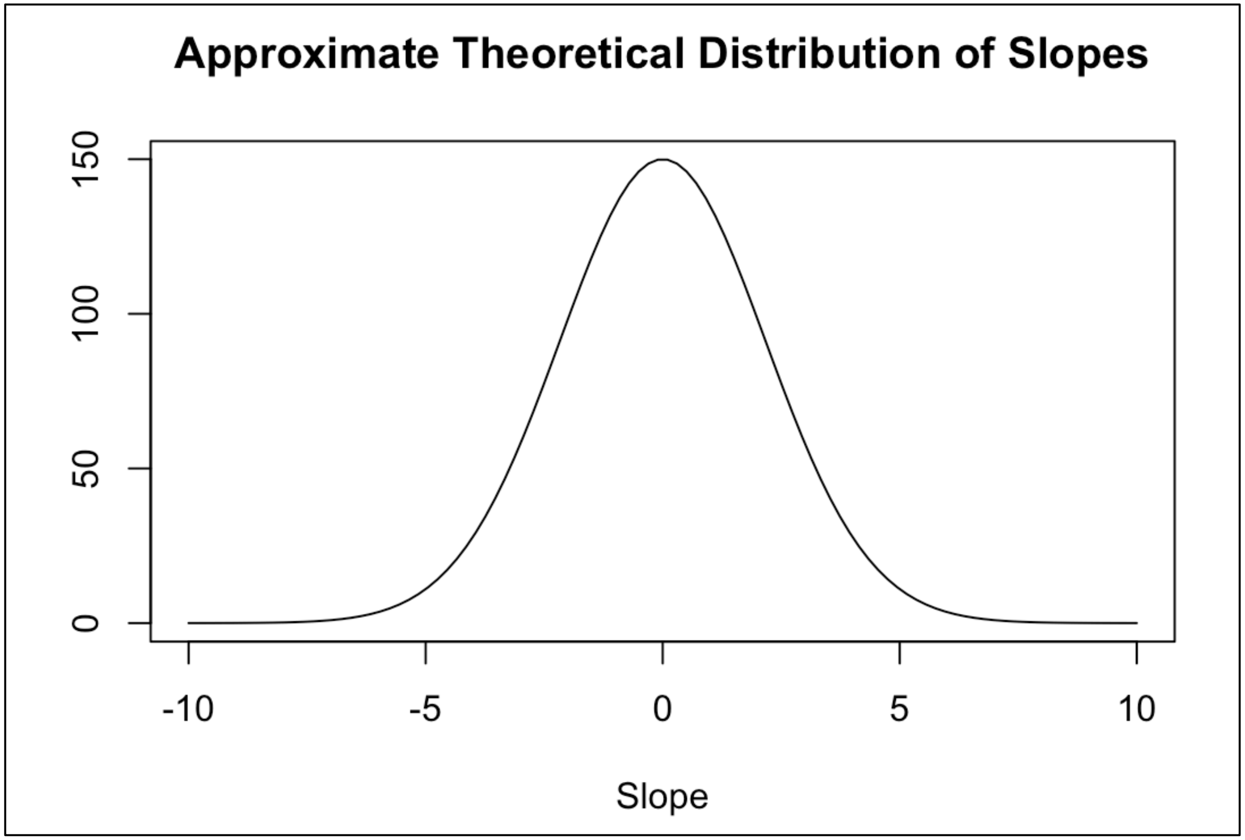
```
library(boot)
slope <- function(formula, data, indices) {
  d <- data[sample(1:10000,200,replace=TRUE),]
  fit <- lm(formula, data=d)
  return(summary(fit)$coefficients[2,1])}
results <- boot(data=df, statistic=slope,
               R=2000, formula=Spend~Age)

hist(results$t,breaks=50,main="Histogram of Slopes",xlab="Slope")
```



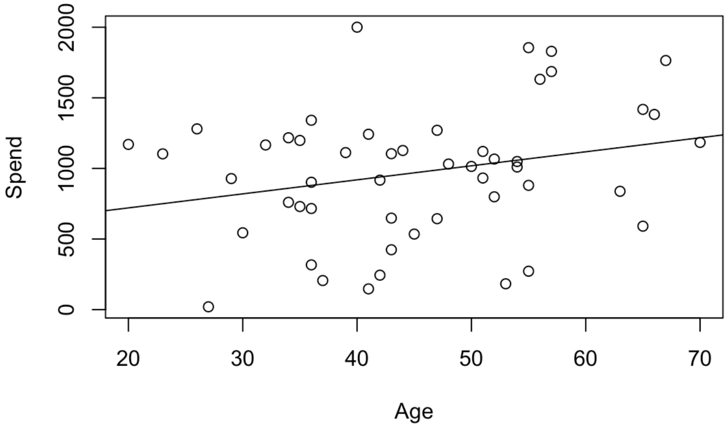
The Sampling Distribution of the Slope (Theoretically Derived)

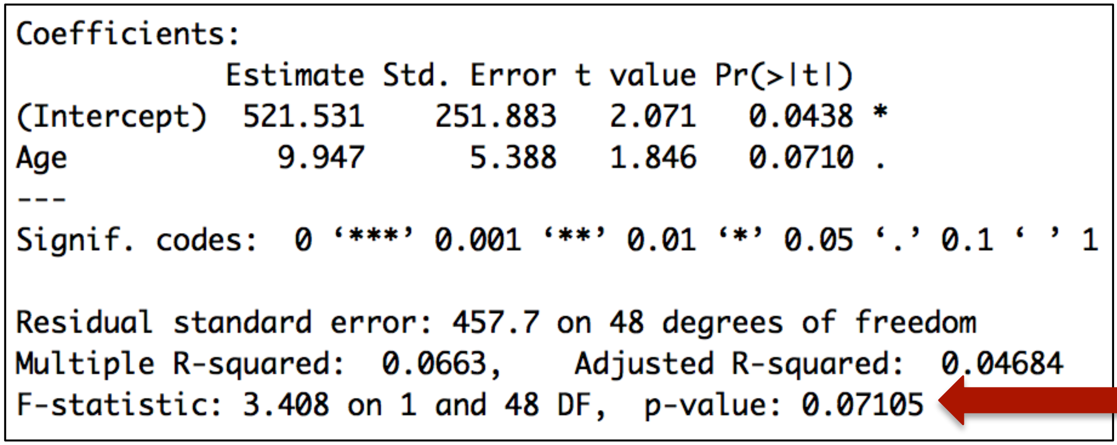
Question: If the population slope is zero, how likely is it that we would draw a sample with a slope as extreme as ours (i.e., more extreme than 1.9047) in either direction?

```
x <- seq(-10, 10, length=100)
plot(x, dt(x*.4592,198)*150/dt(0,198), type="l", lty=1, xlab="Slope", ylab="",
    main="Approximate Theoretical Distribution of Slopes")
```



Statistical Significance (revisited)





```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  521.531    251.883   2.071   0.0438 *
Age            9.947      5.388   1.846   0.0710 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 457.7 on 48 degrees of freedom
Multiple R-squared:  0.0663,    Adjusted R-squared:  0.04684
F-statistic: 3.408 on 1 and 48 DF,  p-value: 0.07105
```

Is this slope “statistically significant”?
Why/why not?

What does that term mean in the current context?

What is the null hypothesis?
What is the alternative hypothesis?

What would you tell the CEO of Tivek?

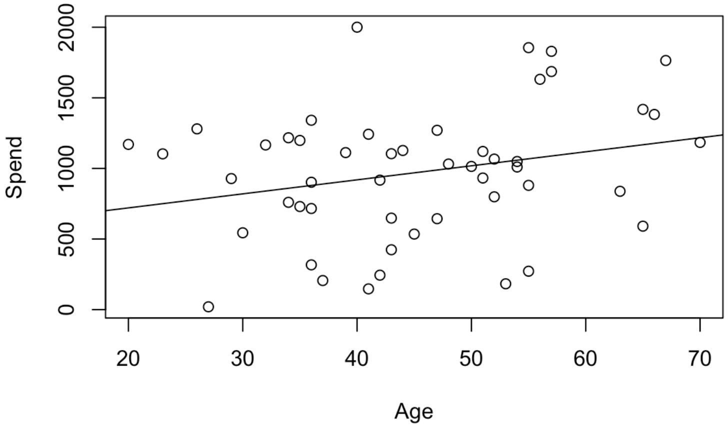
Type I and Type II Errors

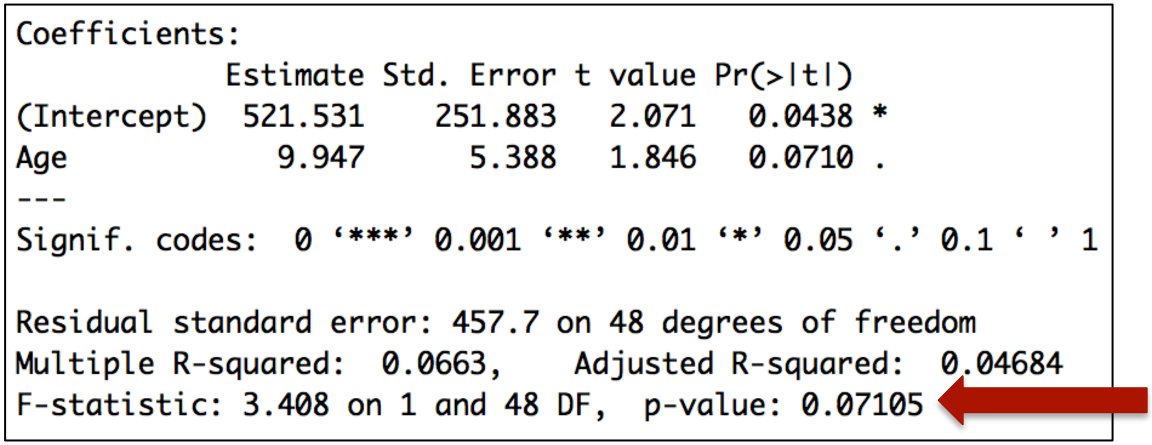
Type I error: Rejecting H_0 when H_0 is true

Type II errror: Failing to reject $H_0$when $H_0$is false

Convention: Reject $H_0$when the probability of making a Type I error (the *p* value) is less than .05 (the α level)

The Probability Value (“p-value”) Defined



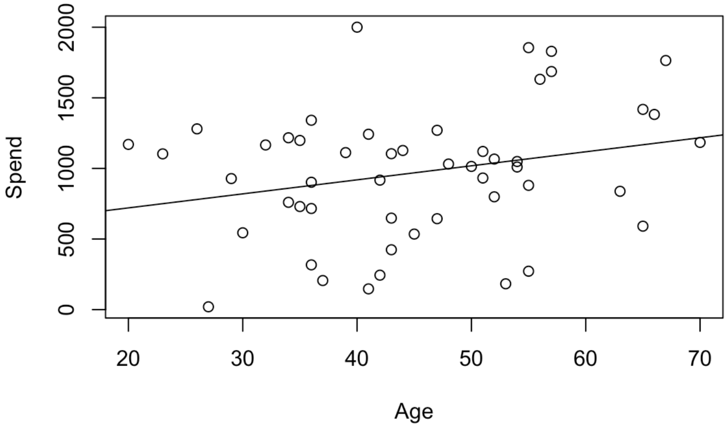


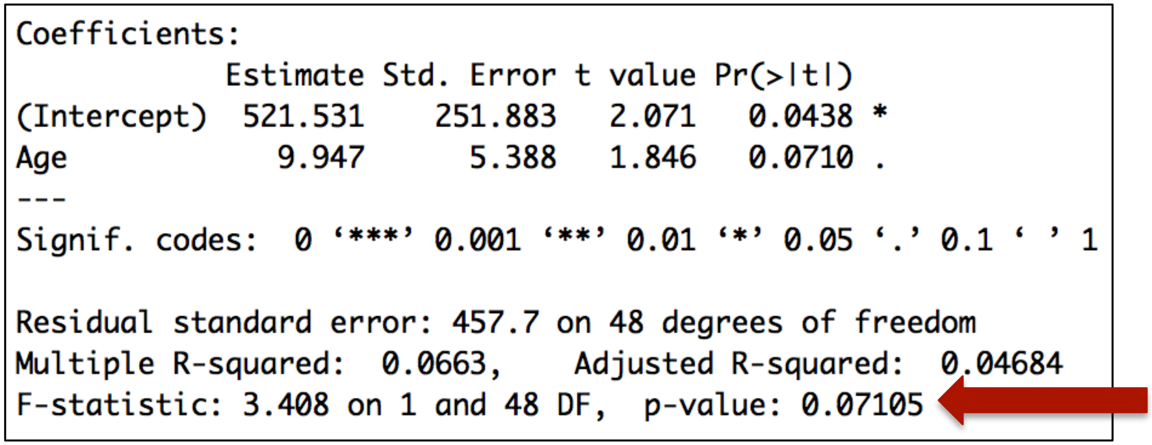
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  521.531    251.883   2.071   0.0438 *
Age            9.947      5.388   1.846   0.0710 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 457.7 on 48 degrees of freedom
Multiple R-squared:  0.0663,	Adjusted R-squared:  0.04684 
F-statistic: 3.408 on 1 and 48 DF,  p-value: 0.07105
```

p-value: Upon repeated sampling from a population where the population slope is exactly zero, we would see a slope as extreme as or more extreme than 9.947, in either direction, in approximately 7 samples out of every 100 we drew.

Question: Would you consider this a likely event or an unlikely event? At what threshold value would you consider this unlikely?

What Affects the p-value?



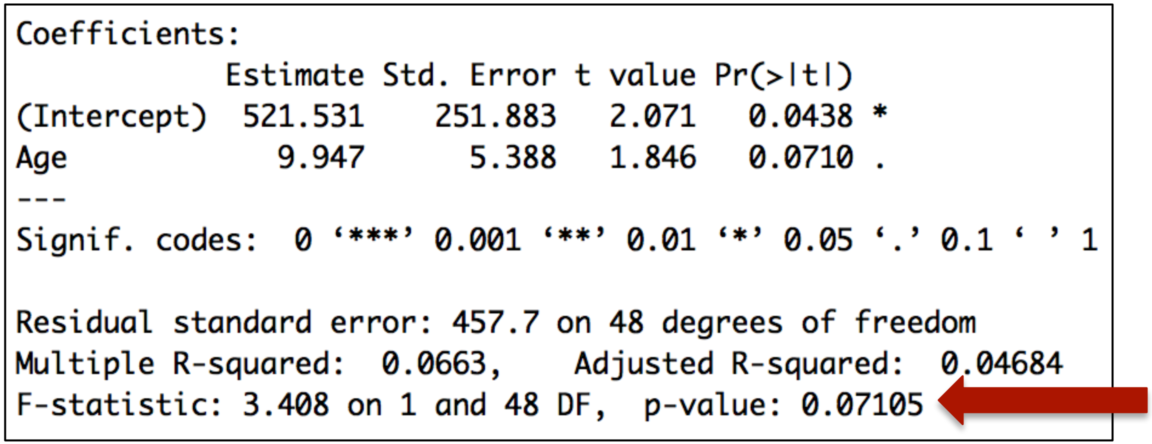


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  521.531    251.883   2.071   0.0438 *
Age            9.947      5.388   1.846   0.0710 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 457.7 on 48 degrees of freedom
Multiple R-squared:  0.0663,    Adjusted R-squared:  0.04684
F-statistic: 3.408 on 1 and 48 DF,  p-value: 0.07105
```

1. **The sample slope**
2. **The spread of the points around the regression line**
3. **The variance of the independent variable**
4. **The adequacy of the model (e.g., “linear”)**
5. **The sample size**

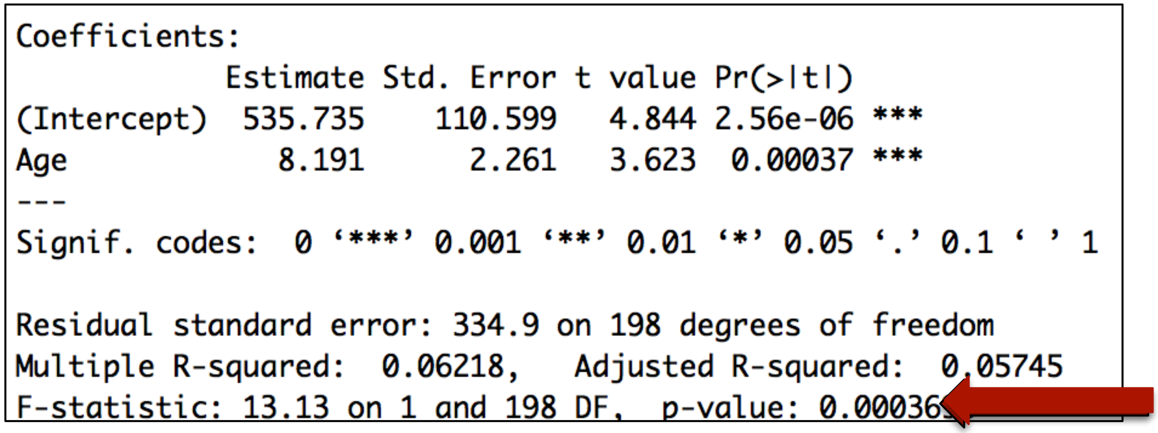
This fifth element (the sample size) can lead to an interesting paradox, which is largely responsible for the popularity of the Bayesian perspective on probability. Let’s consider an example. Please remember the slope (9.947) and your conclusion about the p-value.

The Paradox



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  521.531    251.883   2.071   0.0438 *
Age            9.947      5.388   1.846   0.0710 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 457.7 on 48 degrees of freedom
Multiple R-squared:  0.0663,    Adjusted R-squared:  0.04684
F-statistic: 3.408 on 1 and 48 DF,  p-value: 0.07105
```

n=50



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  535.735    110.599   4.844 2.56e-06 ***
Age            8.191      2.261   3.623  0.00037 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 334.9 on 198 degrees of freedom
Multiple R-squared:  0.06218,   Adjusted R-squared:  0.05745
F-statistic: 13.13 on 1 and 198 DF,  p-value: 0.00036
```

n=200

Suppose that drew a random sample of n=200 rather than n=50. The slope is likely to be very similar (as it is here).

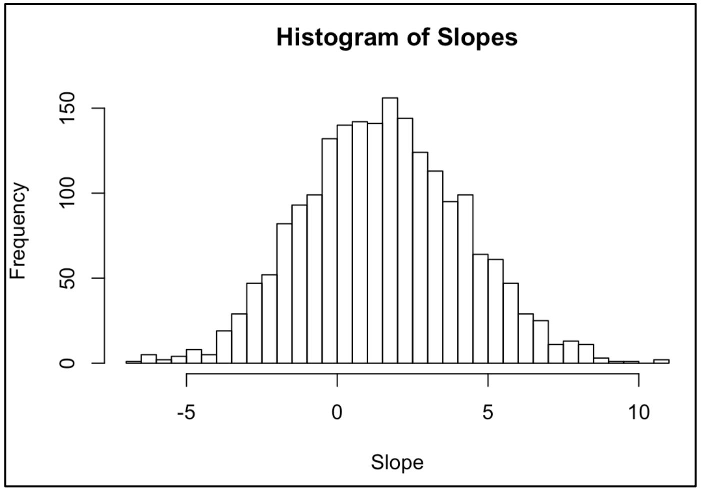
Statistically significant?

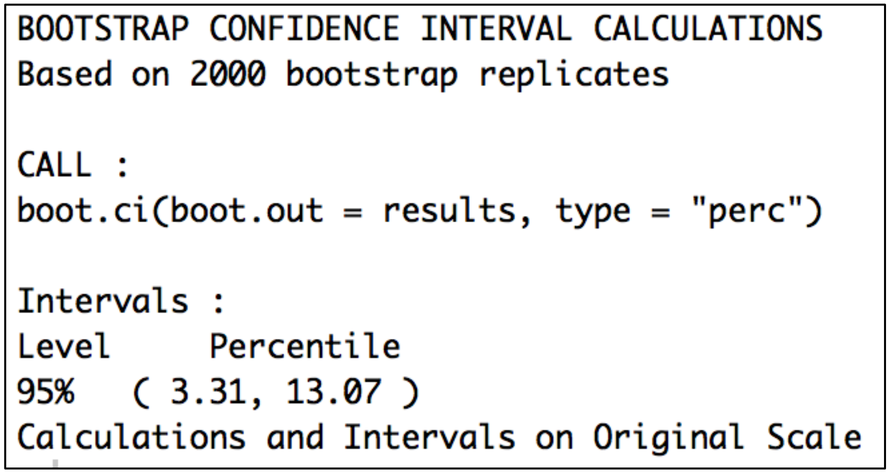
NOW what do you tell the CEO of Tivek?

Is this paradox likely to occur in an era of “big data”?

What is the way out of this paradox?

The True Value of the Population Slope: Bootstrapped Confidence Intervals

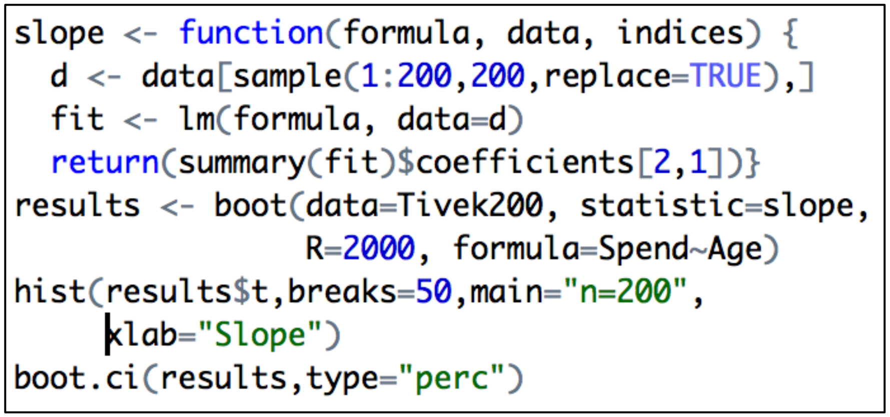




```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 2000 bootstrap replicates

CALL :
boot.ci(boot.out = results, type = "perc")

Intervals :
Level     Percentile
95%   ( 3.31, 13.07 )
Calculations and Intervals on Original Scale
```

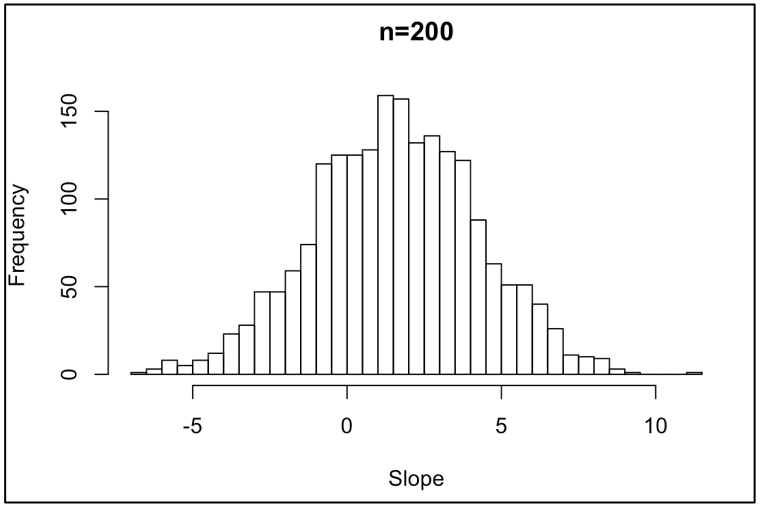
Questions:
Between what two approximate values can we be 95% confident that the population slope lies?

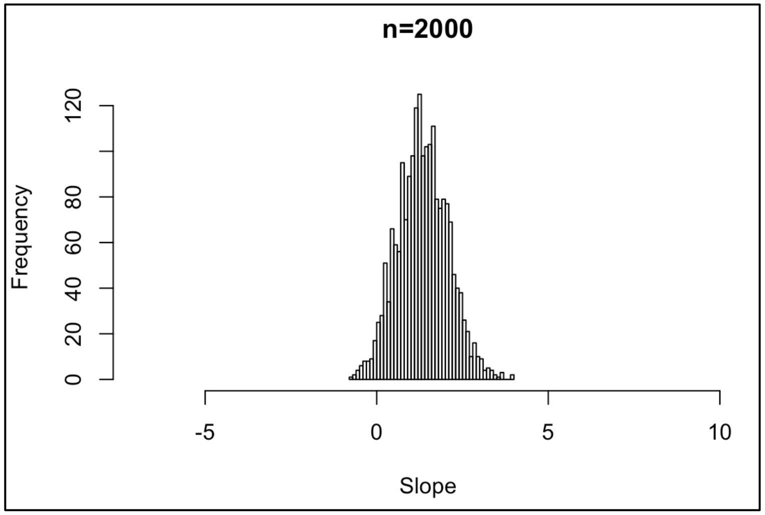
Would this information go into your report to the CEO of Tivek? Why/why not?

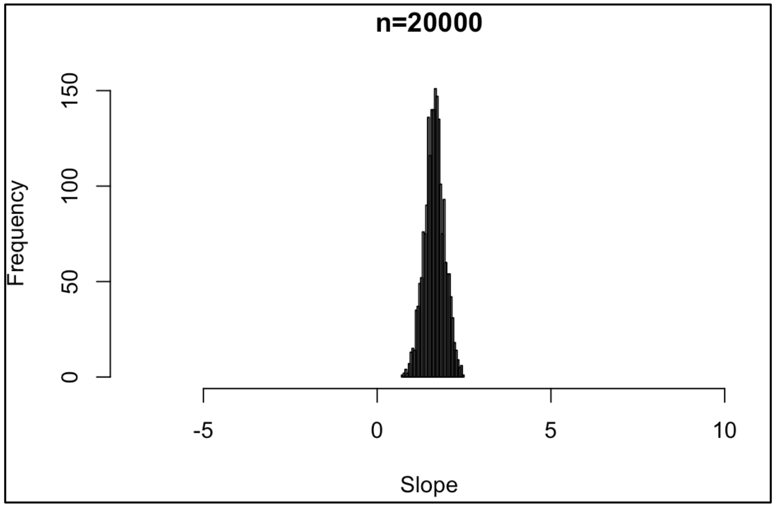


```
slope <- function(formula, data, indices) {
  d <- data[sample(1:200,200,replace=TRUE),]
  fit <- lm(formula, data=d)
  return(summary(fit)$coefficients[2,1])}
results <- boot(data=Tivek200, statistic=slope,
                R=2000, formula=Spend~Age)
hist(results$t,breaks=50,main="n=200",
     xlab="Slope")
boot.ci(results,type="perc")
```

How Does "Big Data" Affect All of This?





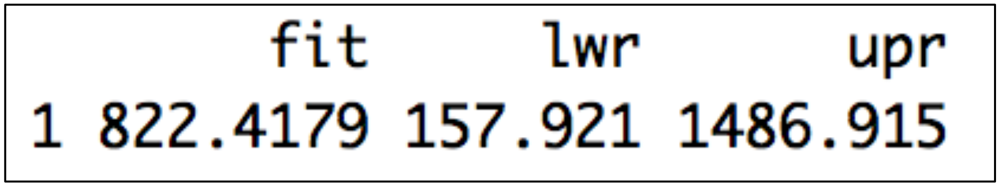


How does the context of "big data" affect "statistical significance"?

How large does the sample need to be in order for this effect to occur?

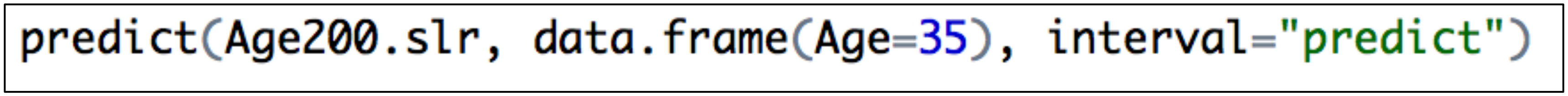
How does "big data" affect the nature of your report to the Tivek CEO?

Prediction and the “Prediction (Tolerance) Interval”



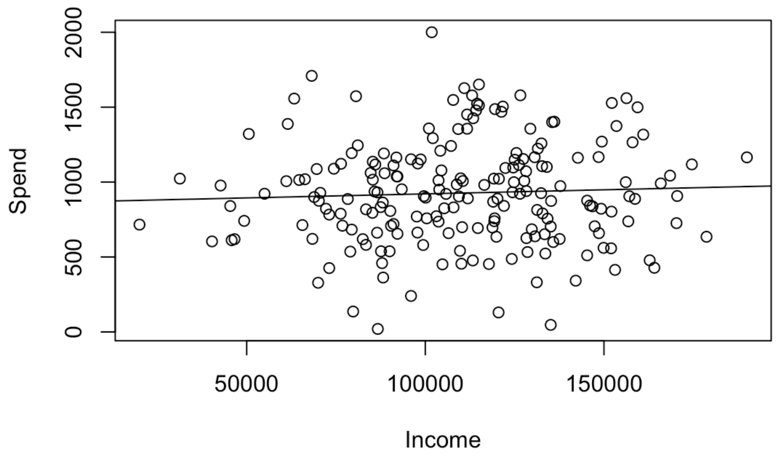
```
       fit     lwr      upr
1 822.4179 157.921 1486.915
```

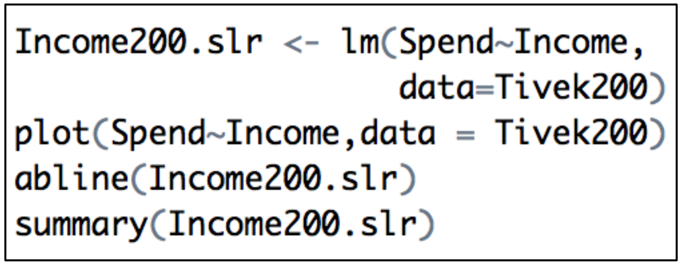
Assuming that the error term ϵ in the simple linear regression model is independent of x, and is normally distributed, with zero mean and constant variance, for a given value of x, what is the 95% interval estimate of the dependent variable?



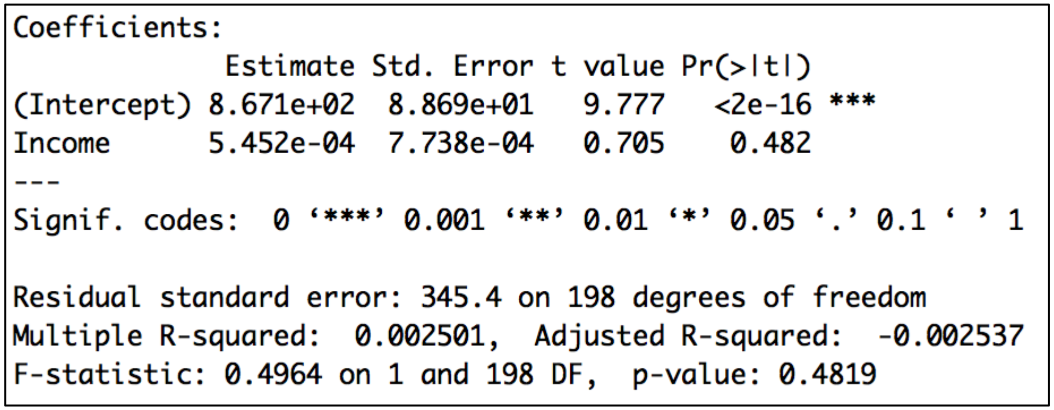
```
predict(Age200.slr, data.frame(Age=35), interval="predict")
```

Income: Tivek200 Sample





```
Income200.slr <- lm(Spend~Income,
                    data=Tivek200)
plot(Spend~Income,data = Tivek200)
abline(Income200.slr)
summary(Income200.slr)
```

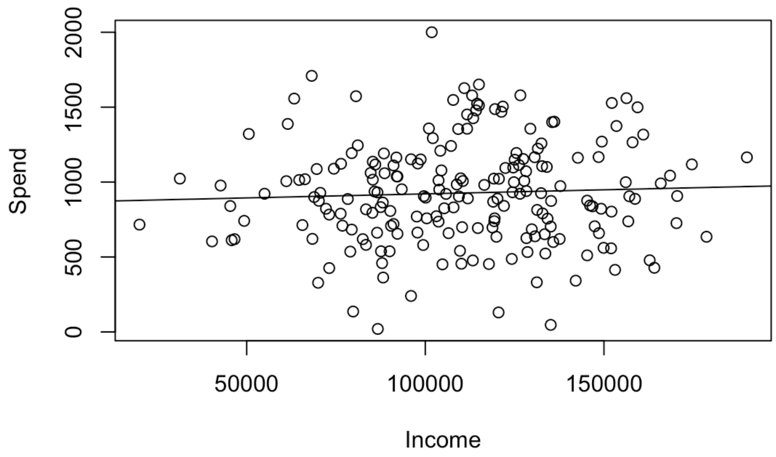


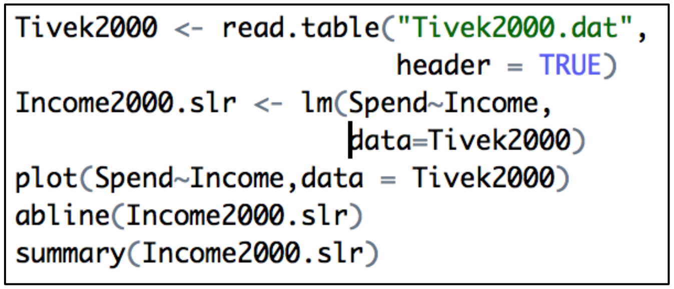
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 8.671e+02  8.869e+01   9.777   <2e-16 ***
Income      5.452e-04  7.738e-04   0.705    0.482    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 345.4 on 198 degrees of freedom
Multiple R-squared:  0.002501,	Adjusted R-squared:  -0.002537 
F-statistic: 0.4964 on 1 and 198 DF,  p-value: 0.4819
```

What implications do these results have that would be of interest to the CEO of Tivek?

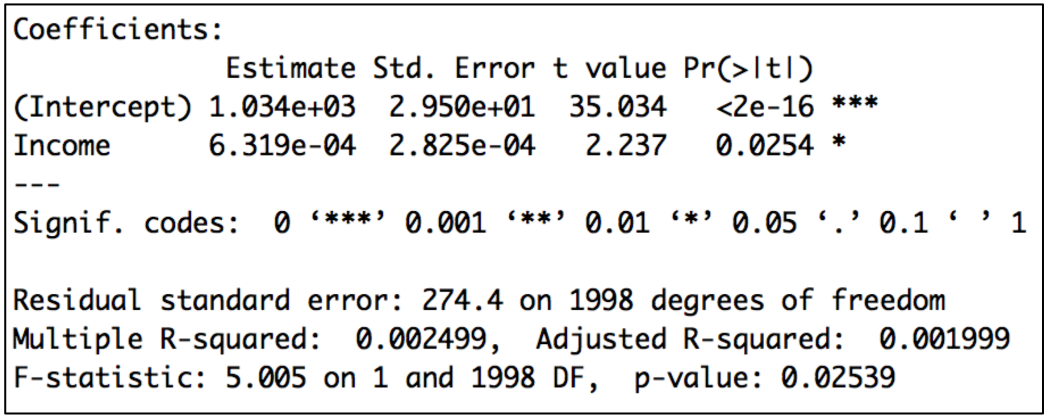
Would “statistical significance” be part of your report to the CEO?

Income: Tivek2000 Sample





```
Tivek2000 <- read.table("Tivek2000.dat",
                        header = TRUE)
Income2000.slr <- lm(Spend~Income,
                     data=Tivek2000)
plot(Spend~Income,data = Tivek2000)
abline(Income2000.slr)
summary(Income2000.slr)
```



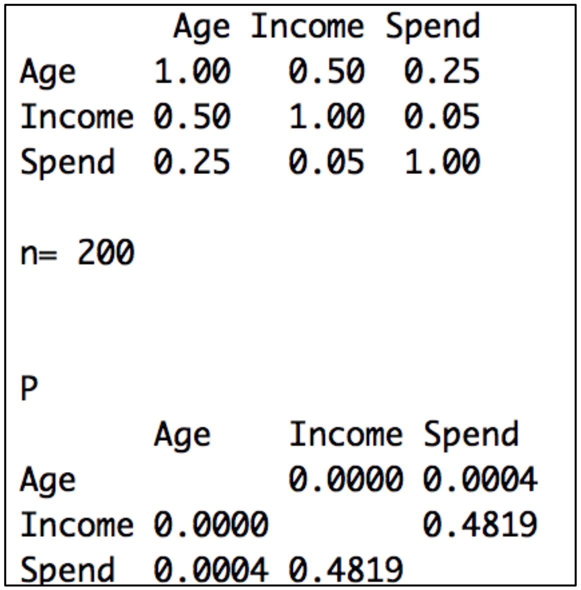
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.034e+03  2.950e+01  35.034   <2e-16 ***
Income      6.319e-04  2.825e-04   2.237   0.0254 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 274.4 on 1998 degrees of freedom
Multiple R-squared:  0.002499,  Adjusted R-squared:  0.001999
F-statistic: 5.005 on 1 and 1998 DF,  p-value: 0.02539
```

Note: Almost identical slope in Tivek200 sample.

What implications do these results have that would be of interest to the CEO of Tivek?

Would “statistical significance” be part of your report to the CEO?

A Preview of Next Week: Multicollinearity (Tivek200 Data)

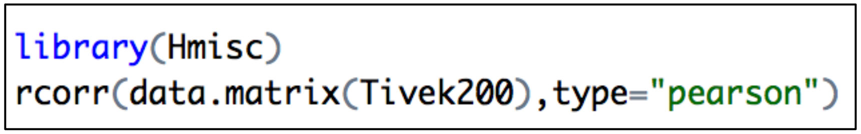


```
         Age  Income Spend
Age     1.00   0.50  0.25
Income  0.50   1.00  0.05
Spend   0.25   0.05  1.00

n= 200


P
        Age    Income Spend
Age            0.0000 0.0004
Income 0.0000         0.4819
Spend  0.0004 0.4819
```

Are Age and Income correlated? What does "being correlated" mean?

If so, is this correlation "significant"? If so, what does "significant" mean?

What implications do these results have that would be of interest to the CEO of Tivek?



```
library(Hmisc)
rcorr(data.matrix(Tivek200),type="pearson")
```

Key Points in Today's Discussion

- In Business Analytics, when we examine the relationship between variables, it is vitally important to assess BOTH *statistical significance* and *relationship strength*.
    - *F*, *t*, and the probability value (*p*-value) all pertain to statistical significance;
    - The Coefficient of Determination is a measure of relationship strength.
- Measures of statistical significance address the question of whether ANY relationship exists between the variables; measures of relationship strength address the question of how strong the relationship is.
- The accuracy of both types of measures depend on sample size; even miniscule relationships can be "statistically significant" in the context of a large sample size (e.g., "big data"), and even large Coefficients of Determination can be statistically non-significant in the context of a small sample size.
- Particularly when dealing with "big data", where nearly every relationship will be statistically significant, it is important to assess the strength of the relationship as well as its statistical significance.
- As we will see next week, when dealing with multiple independent variables, both the statistical significance and the comparative importance of each of the independent variables may be different at the multivariable level than what is observed at the bivariate level.

Administrivia

- **Reminder: Assignment 1 is due at 4:25pm next Wednesday (via Blackboard).** ***Blackboard will prohibit submitting assignments after 4:25pm*****. Assignments will NOT be accepted by email;**

- **If you would like to set up an appointment with me, please send email to pww@gwu.edu.**

VOH ON FRIDAY!