Stat 515: Introduction to Statistics

Chapter 11

Let’s Review Lines

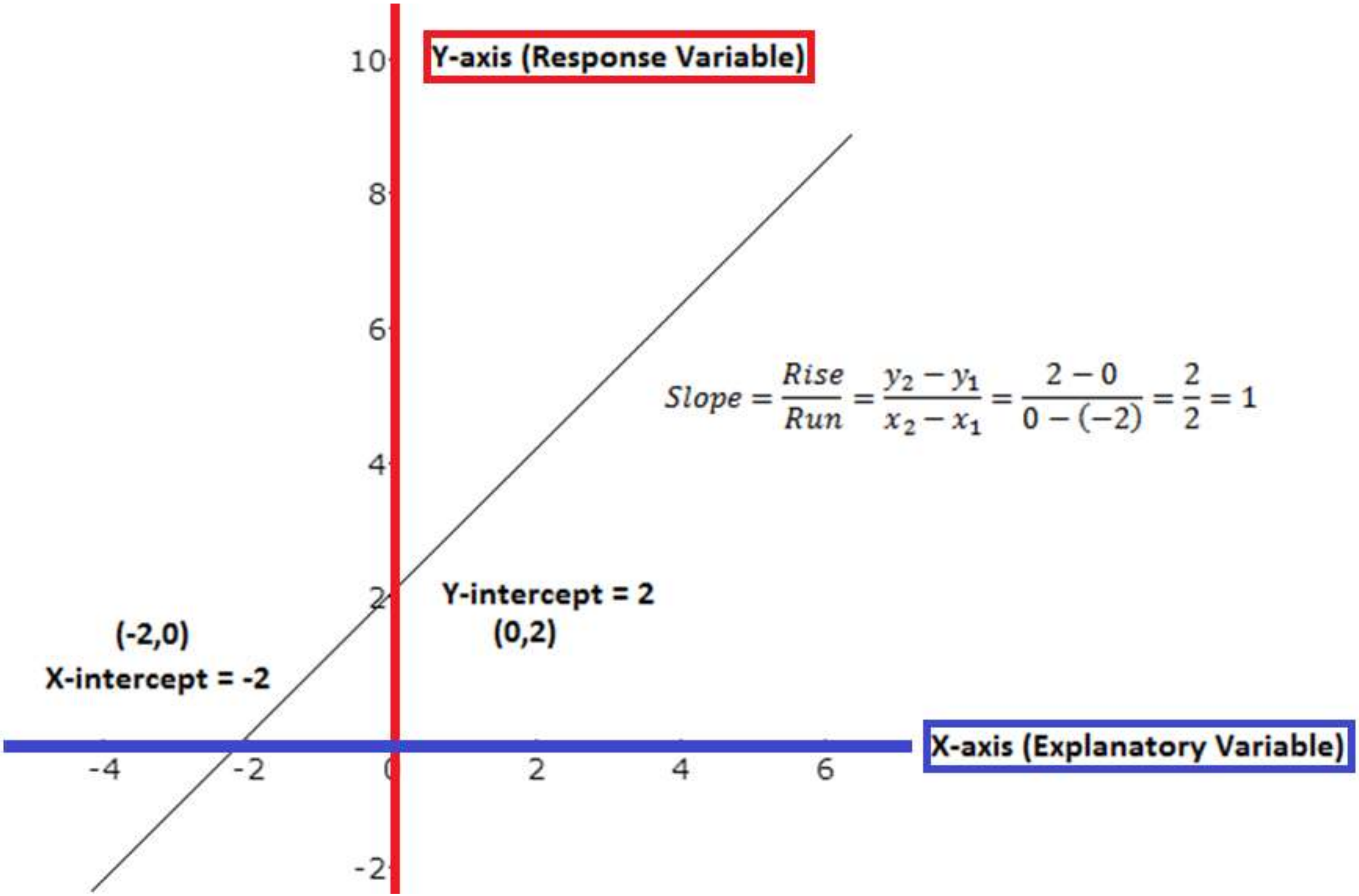
- A **line** is the shortest distance between two points. It has no curve, no thickness and it extends both ways indefinitely.

- The equation has the following forms
  - **Slope Intercept:** $y = m * x + b$
  - **Point-Slope:** $(y - y_1) = m * (x - x_1)$

Lines

- The **y-axis** runs vertically where x=0

- The **x-axis** runs horizontally where y=0.

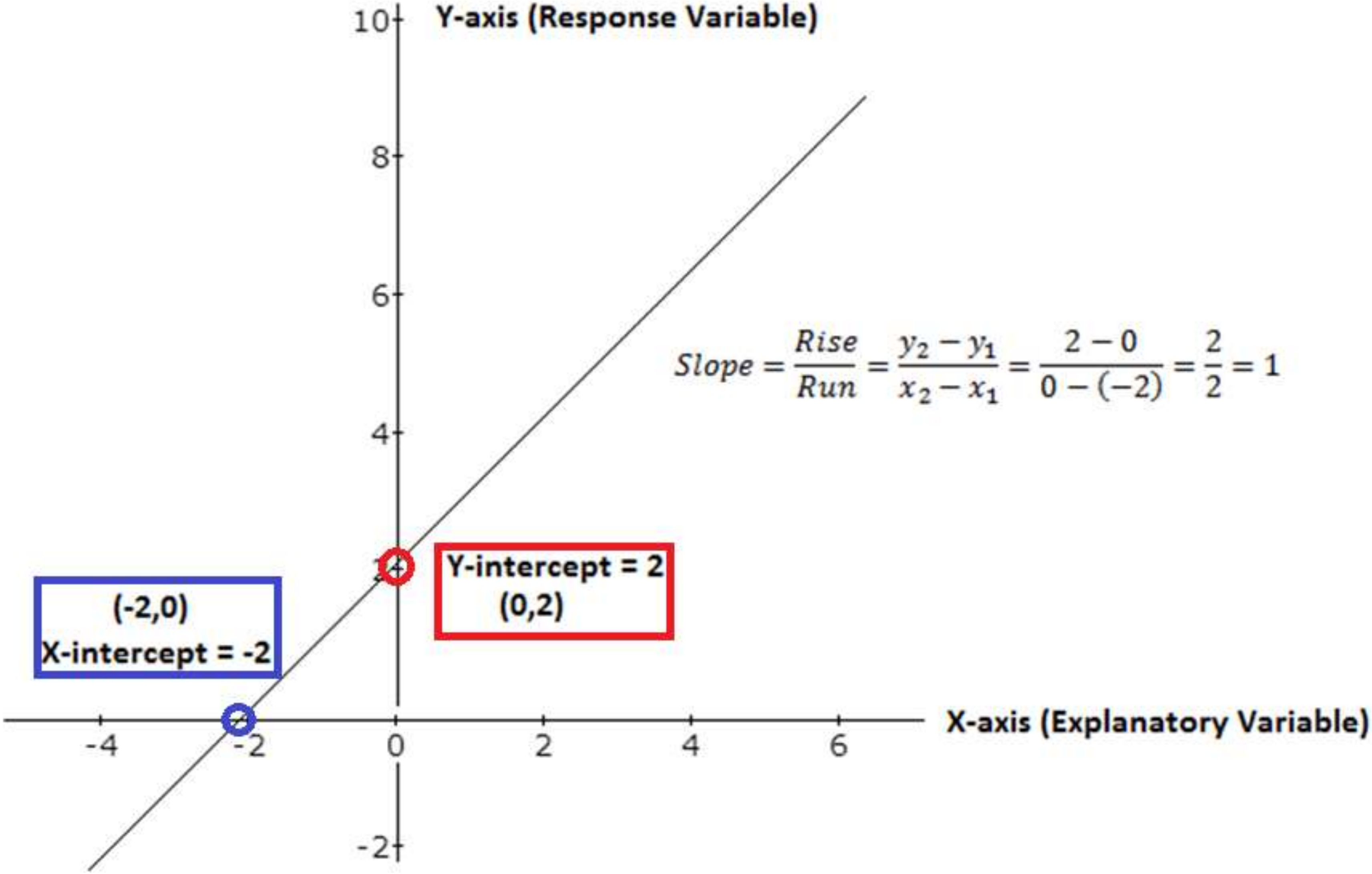


$$Slope = \frac{Rise}{Run} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{2-0}{0-(-2)} = \frac{2}{2} = 1$$

Lines

- The **Y-intercept** is where the line crosses the y-axis and can be found by plugging in $x=0$ → $y=m(0)+b=b$. So, b is the Y-intercept.
  - This is important because it is the value the dependent (response) variable takes when the independent (explanatory) variable is zero

- The **X-intercept** is where the line crosses the x-axis and can be found by plugging in $y=0$ → $0=mx+b$ →$mx=(-b)$→$x=(-b)/m$.

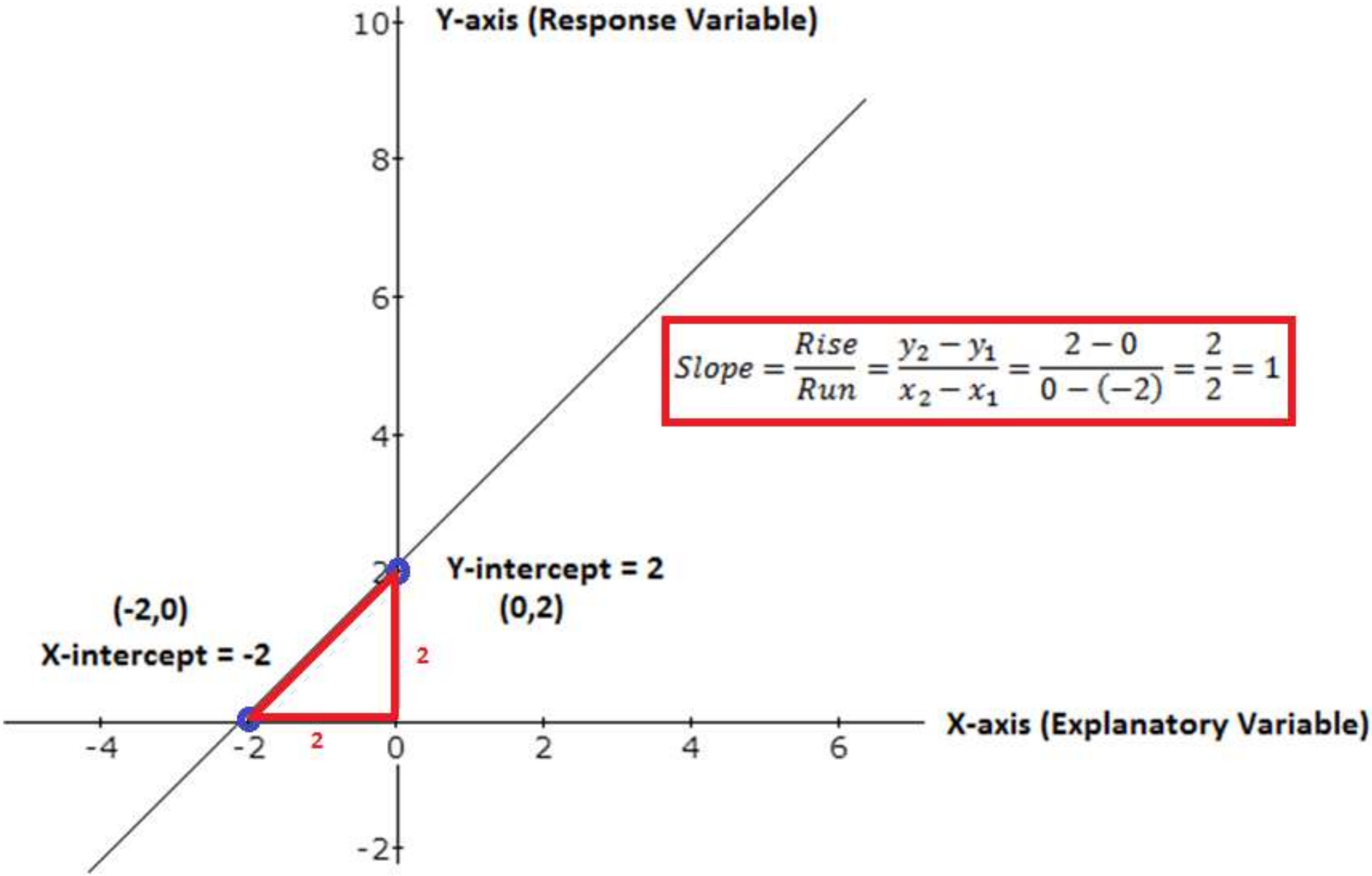


$$Slope = \frac{Rise}{Run} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{2-0}{0-(-2)} = \frac{2}{2} = 1$$

Lines

- The **slope (m)** is a measurement of how the line changes; it is the number that multiplies x. It can be thought of as the change in y for every unit change in x,
  - ie. the change in y for every increase of one in x.

- It can be calculated using any two points on the line (x_1, y_1) and (x_2, y_2) as below, but it is given by the m term in the equation for the line.
  - $Slope = m = \frac{Rise}{Run} = \frac{y_2 - y_1}{x_2 - x_1}$



$$Slope = \frac{Rise}{Run} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{2 - 0}{0 - (-2)} = \frac{2}{2} = 1$$

Regression – Making Lines Useful!

- Unlike the lines we learned in math, our data won't fit the line exactly
  - Math: deterministic model
  - Stats: probabilistic model

- **Regression Line –** predicts the value for the response variable y as a straight line function of the value of x, the explanatory variable, with some random error

Association of Variables – Two Categorical Variables

- **Response Variable –** this is our dependent variable, the outcome variable on which comparisons are made
- **Explanatory Variable –** this is our independent variable, the groups to be compared with respect to values on the response variable
- **Think "we use the explanatory variable to EXPLAIN what's going on with the response variable."**

Examples

- Example 1:
  - Response: Age of death (quantitative)
  - Explanatory: Cigarettes smoked per day (quantitative)

- The idea here is that an experimental unit's smoking status gives us some of the information about how long they will live
  - Actuaries do this sort of thing – evidence has shown that smoking decreases your life expectancy.

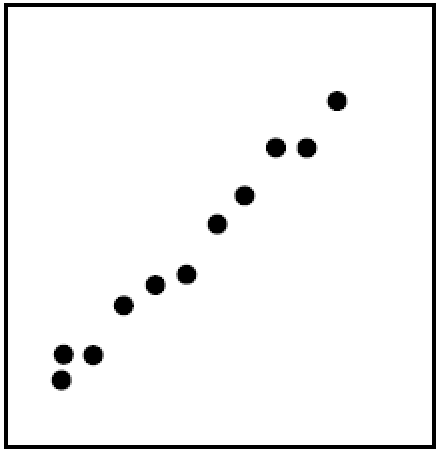
More Definitions

- An **association or correlation** exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable
- “Evidence has shown that smoking more decreases your life expectancy.”
  - Here we say that there is an **association** between smoking and life expectancy.

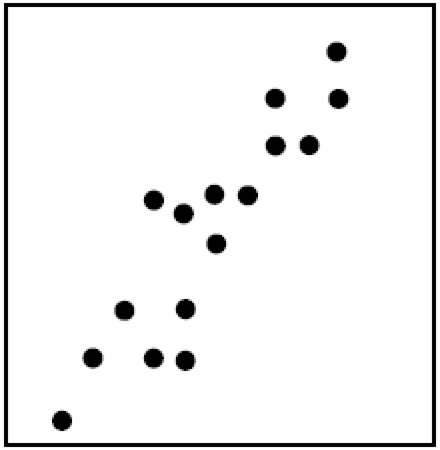
Scatterplots

- We can **compare** **two quantitative variables** and explore their association or correlation with a **scatterplot**

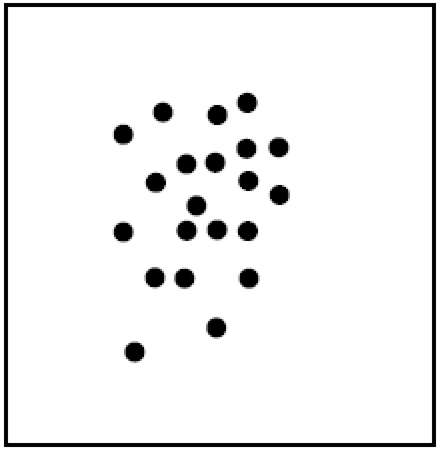
- To form a **scatterplot** we let the **response** variable be the y variable and the **explanatory** variable be the x variable and plot the points



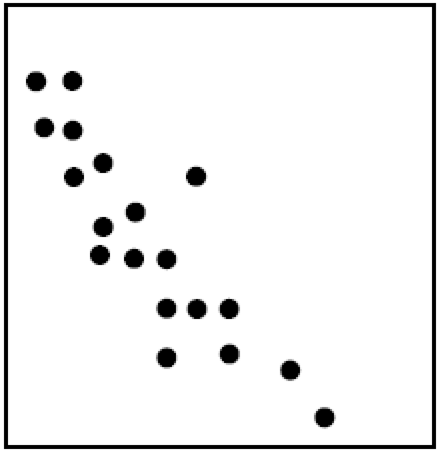
Strong positive correlation



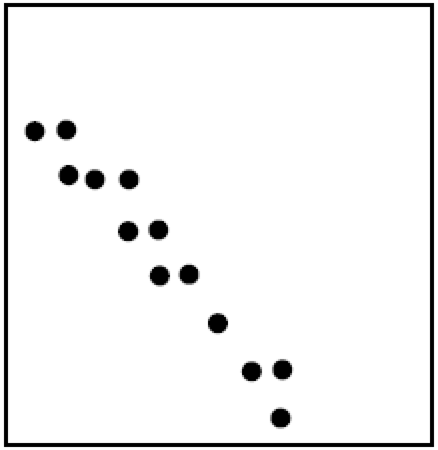
Moderate positive correlation



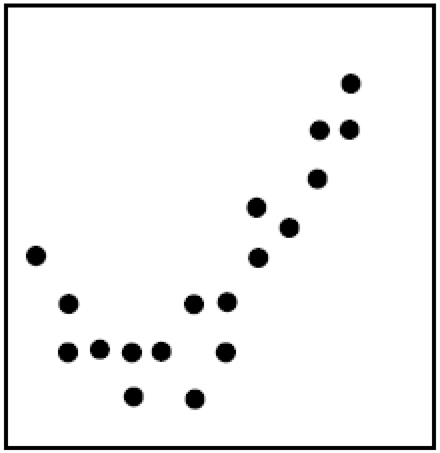
No correlation



Moderate negative correlation



Strong negative correlation



Curvilinear relationship

Coefficient of Correlation (r)

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

Where:

$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

$$SS_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$SS_{yy} = \sum (y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

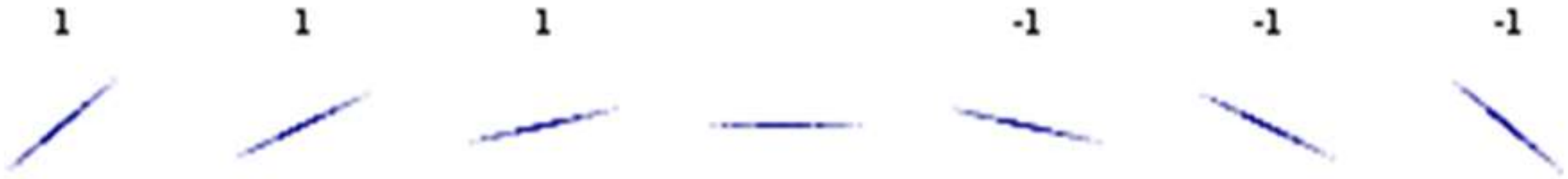
Coefficient of Correlation (r)

- **r** measures the **LINEAR** relationship between x and y [linear, linear, linear, linear!!!]
- **r > 0** → positive correlation or association
- **r < 0** → negative correlation or association
- **r=1** → perfect positive correlation or association
  - Here, all points would fit on a line
- **r=-1** → perfect negative correlation
  - Here, all points would fit on a line
- **r=0** → no correlation

Properties

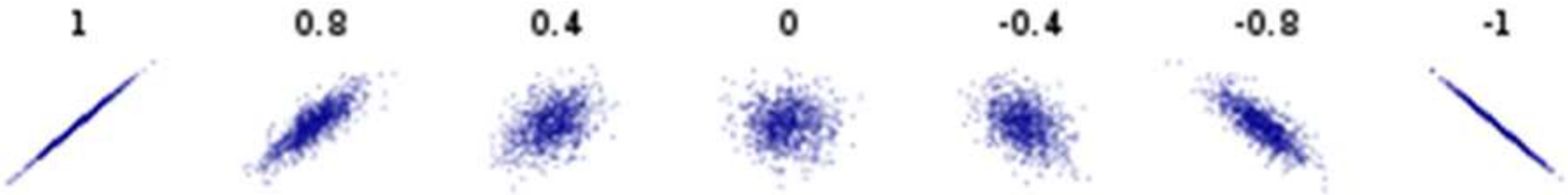
- $-1 \leq r \leq 1$
- The closer r is to 1 the stronger the evidence for positive association
- The closer r is to -1 the stronger the evidence for a negative association
- The closer r is to 0 the weaker the evidence for association
- **Affected by outliers so we have to be careful**

Coefficient of Correlation Examples



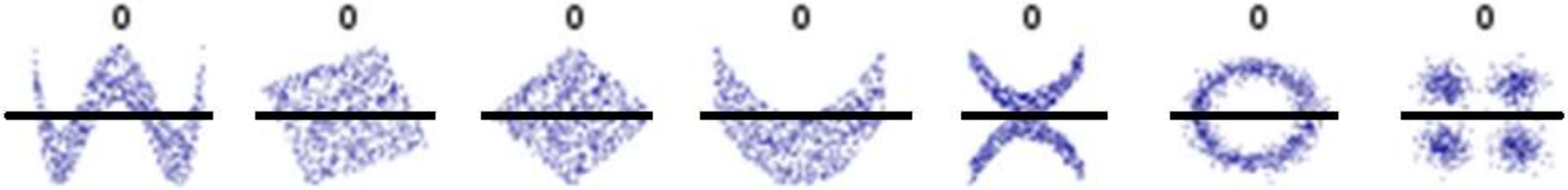
- The perfect lines have r=1 or r=-1 **depending on the sign of their slope <u>not</u> the magnitude of the slope**
- You might think that the plot in the middle has r=1 too because it too is fit perfectly by a line. **The catch** is that the line would be horizontal, thus having a slope value of zero (no sign.) These dots actually show that the explanatory variable provides no explanation for the response variable.

Coefficient of Correlation Examples



- Again, the perfect lines have r=1 or r=-1
- The points that don’t make perfect lines have decimal values depending on how close they are to a perfect line.
  - The closer r is to 1 the stronger the evidence for positive association
  - The closer r is to -1 the stronger the evidence for a negative association
  - The closer r is to 0 the weaker the evidence for association
- **Note: their value changes based on how close they are to forming a line <u>not</u> the magnitude of the slope**

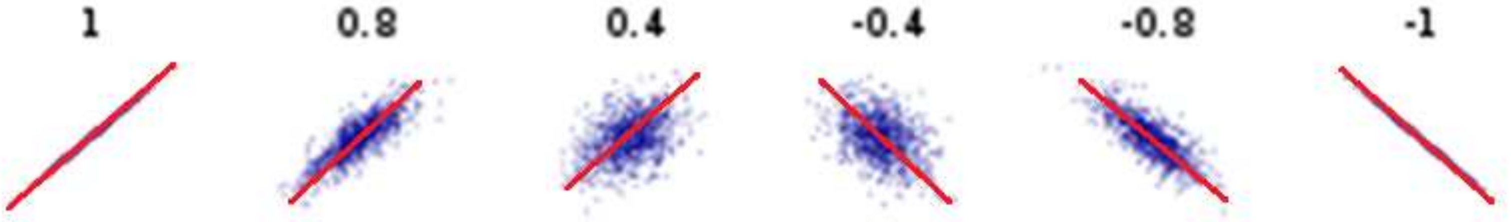
Coefficient of Correlation Examples



- Here there are obvious patterns here **but** they are not linear!
- Since, r, measures the **linear** relationship between two variables r=0 even though there are patterns!
  - In each case the best balanced line would be horizontal

Regression – Making Lines Useful!

- **Regression Line –** we make our regression line so that it best fits our data – unlike math it usually isn't a perfect



$$\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x} + \in$$

  - Note: we need to estimate these values with our data

Regression – Making Lines Useful!

- $\widehat{\boldsymbol{y}} = \boldsymbol{b_0} + \boldsymbol{b_1} * \boldsymbol{x} + \in$
  - $\boldsymbol{b_0}$ is the intercept – when x=0
    - This is important because it is the expected value of the response variable, y, when x=0
  - $\boldsymbol{b_1}$ is the slope of the line
    - This is important because it is the amount that $\hat{y}$ changes when x increases by one unit
  - $\widehat{\boldsymbol{y}}$ is the predicted value for some x
  - $\in$ **= the residual** = (the real y) – $\hat{y}$

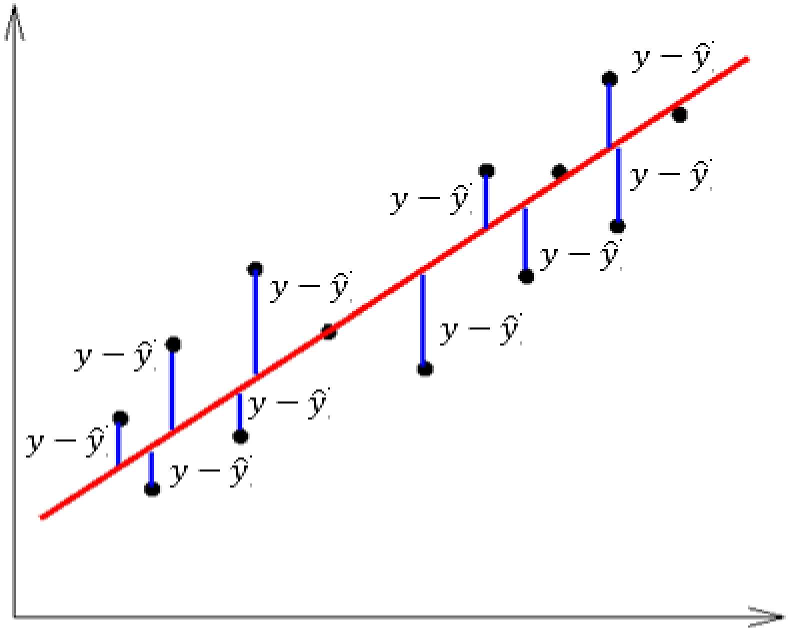
Regression – A Way to Find It

- **Least Squares** is the most popular method
  - It returns the line that has the smallest value for the residual sum of squares in using:

- $\boldsymbol{\widehat{y} = b_0 + b_1 * x} + \in$
- Residual Sum of Squares = $\sum(y - \hat{y})^2$

Regression – A Way to Find It

- We don’t just draw the ‘best-fit line’ like we might have before this class
- Least squares gives us the solution where the total length of blue lines the smallest



Regression – Least Squares

- We find b_0 and b_1 such that we minimize the sum of squared errors. These estimates are called the ordinary least squares estimators – we leave this up to software.

- In simple regression
  - $\boldsymbol{\hat{y} = b_0 + b_1 * x} + \in$
  - $b_0 = y\ intercept = \hat{y} - b_1 * \bar{x}$
  - $b_1 = slope = r\ * \left(\frac{s_y}{s_x}\right)$

Coefficient of Correlation (r)

$$r^2 = \frac{Explained\ sample\ variability}{total\ variability}$$

$$= \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

Where:

$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

$$SSE = \sum (y_i - \hat{y})^2$$

When Can We Use Regression?

- $\boldsymbol{R^2}$, given in the regression output, gives the percent of variation in the response variable explained by the explanatory variable

- **Note:** $R^2 = r^2$
- **Note:** $r = \sqrt{R^2}$

When Can We Use Regression?

- The **scatterplot** must show a fairly linear relationship
  - A rule of thumb is to look for a coefficient of correlation, **r > .7** or **r < -.7**
  - **Equivalently,** a rule of thumb is to look for a coefficient of correlation, R^2 **> .49**

When Can We Use Regression?

- Assumptions on ∈
  - ∈ is normally distributed
  - E(∈)=0
  - Var(∈)=$\sigma_{\in}^{2}$
  - Each error, ∈, is independent

- We makes these assumptions so that

$$E(y) = \beta_0 + \beta_1 x$$

Estimating σ for a Simple Linear Model

$$s_{\in}^{2} = \frac{SSE}{Degrees\ of\ Freedom\ for\ Error}$$

$$= \frac{\mathrm{SSE}}{\mathrm{n-2}}$$

Where: $SSE = \sum(y_i - \widehat{y_i})^2 = SS_{yy} - b_1 SS_{xy}$

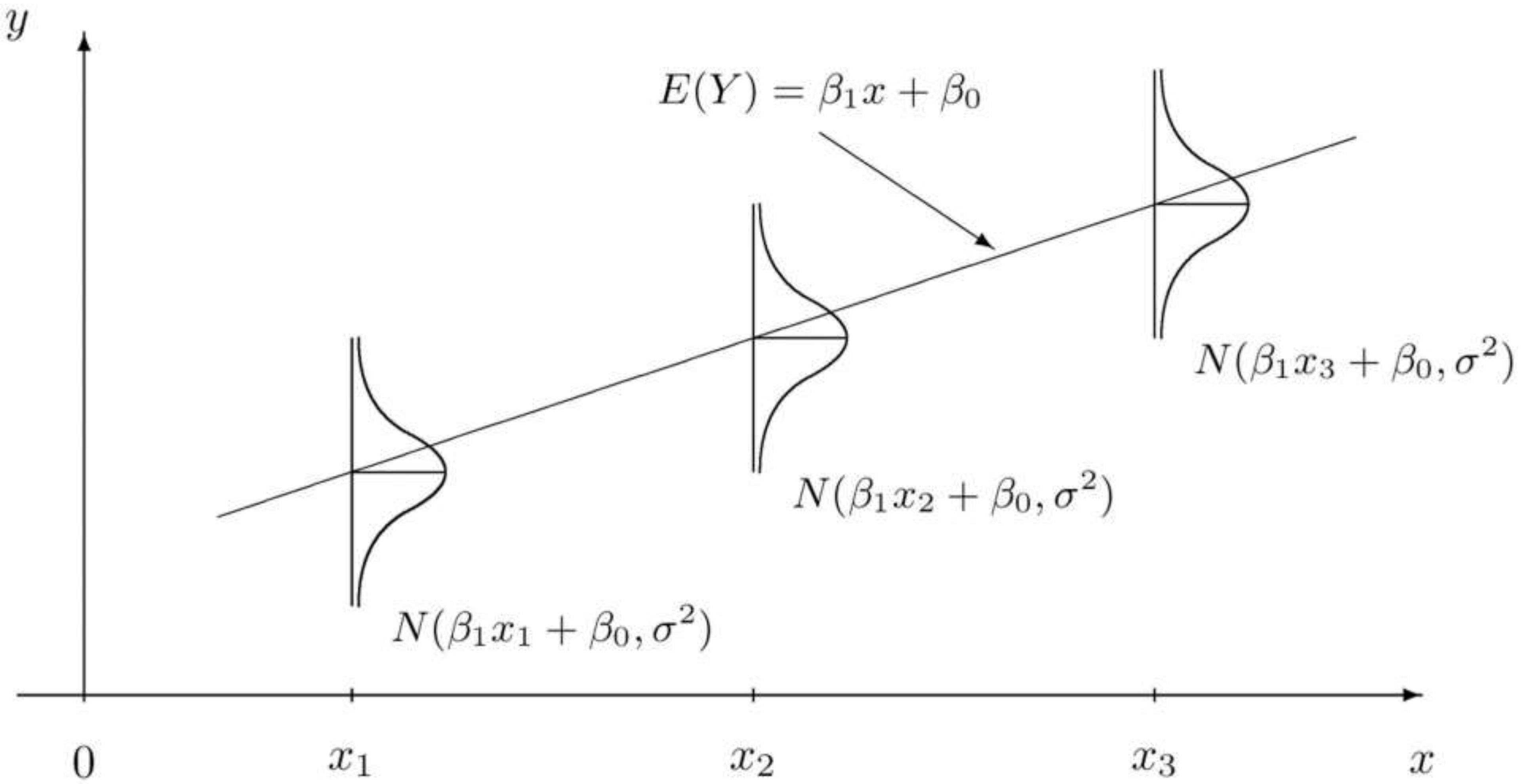
Then: $s_{\in} = \sqrt{s_{\in}^{2}} = \sqrt{\frac{\mathrm{SSE}}{\mathrm{n-2}}}$

Interpretation of our Estimate of σ

Applying the Empirical Rule:

- We expect approximately 90% of observed y values to lie within $1s_{\in}$ of the estimate $\hat{y}$
- We expect approximately 95% of observed y values to lie within $2s_{\in}$ of the estimate $\hat{y}$
- We expect approximately 99.7% of observed y values to lie within $3s_{\in}$ of the estimate $\hat{y}$

Interpretation of our Estimate of σ


y
$E(Y) = \beta_1 x + \beta_0$
$N(\beta_1 x_3 + \beta_0, \sigma^2)$
$N(\beta_1 x_2 + \beta_0, \sigma^2)$
$N(\beta_1 x_1 + \beta_0, \sigma^2)$
0
x_1
x_2
x_3
x

Hypothesis testing for r

Hypothesis Test for Correlation Coefficient: Step 1

- State Hypotheses to some value we're interested in, ρ_o- it's usually easier to start with H_a
  - **Null hypothesis**: we assume that the population correlation coefficient equals some ρ_0
    - $H_o: \rho \leq \rho_o = 0$ (one sided test)
    - $H_o: \rho \geq \rho_o = 0$(one sided test)
    - $H_o: \rho = \rho_o = 0$(two sided test)
  - **Alternative hypothesis:** What we're interested in
    - $H_a: \rho > \rho_o = 0$(one sided test)
    - $H_a: \rho < \rho_o = 0$(one sided test)
    - $Ha: \rho \neq \rho_o = 0$ (two sided test)

Hypothesis Test for Correlation Coefficient: Step 2

- **Check the assumptions:**
  - The variables must be normally distributed
  - There is a linear relationship between the two variables
  - Outliers are either kept to a minimum or removed from the data
  - Equal variance across random variables

Hypothesis Test for Coefficients: Step 3

- **Calculate Test Statistic, t***
  - The test statistic measures how different the sample correlation coefficient we have is from the null hypothesis
  - We calculate the t-statistic by assuming that ρ_o is the population coefficient (we use $\rho = \rho_o = 0$)
    - n-2 degrees of freedom

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{b_1}{\sqrt{\frac{\mathrm{SSE}}{\mathrm{n}-2}}}$$

Hypothesis Test for Coefficients: Step 4

- **Determine the P-value**
  - The P-value describes how unusual the sample data would be if we use $\rho = \rho_o$
  - t* is the test statistic from step 3 with n-2 dof

Alternative Hypothesis	Probability	Formula for the P-value
$H_a\!: \beta_1 > \beta_{1_o}$	Right tail	1-P(T<t*)
$H_a\!: \beta_1 < \beta_{1_o}$	Left tail	P(T<t*)
$H_a\!: \beta_1 \neq \beta_{1_o}$	Two-tail	2P(T<-|t*|)

Hypothesis Test for Coefficients: Step 5

- Summarize the test by reporting and interpreting the P-value
  - Smaller p-values give stronger evidence against H_o
- If p-value$\leq (1 - confidence) = \alpha$
  - Reject H_o, with a p-value = ____, we have sufficient evidence that the alternative hypothesis might be true
- If p-value$> (1 - confidence) = \alpha$
  - Fail to reject H_o, with a p-value = ____, we do not have sufficient evidence that the alternative hypothesis might be true

Using the Sampling Distribution of b_1

Confidence Interval for β_1

Sampling Distribution of b_1

- If the assumptions for $\in$ hold:

  $\mu_{b_1} = \beta_1$

  $\sigma_{b_1} = \frac{\sigma_\in}{\sqrt{SS_{xx}}}$ which we estimate with $\frac{s_\in}{\sqrt{SS_{xx}}}$

- We call σ_{b_1} the **estimated standard error of** b_1

Confidence Intervals for Coefficients

- **Check the assumptions:**
  - ∈ is normally distributed
  - E(∈)=0
  - var(∈)=σ^2
  - Each error, ∈, is independent

- **A 100*(1-α)% confidence interval for $\boldsymbol{\beta_1}$:**

$$b_1 \pm \left(t_{1-\frac{\alpha}{2}}\right)\widehat{s_{b_1}}$$

Using the Sampling Distribution of b_1

Hypothesis Test for β_1

Hypothesis Test for Coefficients: Step 1

- State Hypotheses to some value we're interested in, β_o- it's usually easier to start with H_a
  - **Null hypothesis**: we assume that the population coefficient equals someβ_{1_0}
    - H_o: $\beta_1 \leq \beta_{1_o}$ (one sided test)
    - H_o: $\beta_1 \geq \beta_{1_o}$ (one sided test)
    - H_o: $\beta_1 = \beta_{1_o}$ (two sided test)

  - **Alternative hypothesis:** What we're interested in
    - H_a: $\beta_1 > \beta_{1_o}$ (one sided test)
    - H_a: $\beta_1 < \beta_{1_o}$ (one sided test)
    - Ha: $\beta_1 \neq \beta_{1_o}$ (two sided test)

Hypothesis Test for Coefficients: Step 2

- **Check the assumptions:**
  - $\in$ is normally distributed
  - E($\in$)=0
  - Var($\in$)=$\sigma_{\in}^2$
  - Each error, $\in$, is independent

Hypothesis Test for Coefficients: Step 3

- **Calculate Test Statistic, t***
  - The test statistic measures how different the sample coefficient we have is from the null hypothesis
  - We calculate the t-statistic by assuming that ${\beta_1}_o$ is the population coefficient (we use $\beta_1 = {\beta_1}_o$)
    - n-2 degrees of freedom

$$t^* = \frac{(b_1 - \mu_{b_1})}{\sigma_{b_1}} = \frac{(b_1 - {\beta_1}_o)}{\frac{s_{b_1}}{\sqrt{SS_{xx}}}}$$

Hypothesis Test for Coefficients: Step 4

- **Determine the P-value**
  - The P-value describes how unusual the sample data would be if we use $\beta_1 = \beta_{1_0}$
  - t* is the test statistic from step 3 with n-2 dof

Alternative Hypothesis	Probability	Formula for the P-value
$H_a\colon \beta_1 > \beta_{1_o}$	Right tail	1-P(T<t*)
$H_a\colon \beta_1 < \beta_{1_o}$	Left tail	P(T<t*)
$H_a\colon \beta_1 \neq \beta_{1_o}$	Two-tail	2P(T<-|t*|)

Hypothesis Test for Coefficients: Step 5

- Summarize the test by reporting and interpreting the P-value
  - Smaller p-values give stronger evidence against H_o
- If p-value$\leq (1 - confidence) = \alpha$
  - Reject H_o, with a p-value = ____, we have sufficient evidence that the alternative hypothesis might be true
- If p-value$> (1 - confidence) = \alpha$
  - Fail to reject H_o, with a p-value = ____, we do not have sufficient evidence that the alternative hypothesis might be true

Hypothesis Test for Coefficients: Step 5

- We get the p-values from R for:
  - $H_o: \beta_1 = 0$
  - $H_a: \beta_1 \neq 0$

- This tests whether or not our explanatory variable is “statistically significant” in modeling the response variable

Using Regression Models for Prediction/Estimation Confidence Intervals

Prediction/Estimation: Sampling Errors

- $\sigma_y = \sigma_\in \sqrt{\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$

- $\sigma_{(y-\hat{y})} = \sigma_\in \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$

Prediction/Estimation: Confidence Interval

$$\hat{y} \pm t_{1-\frac{\alpha}{2},n-2} s_{\in} \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$$

- We call this a confidence interval for the observed y value of an observation with x=x_p
- This gives us a 100*(1- α) confidence interval for an observation with x=x_p.

Prediction/Estimation: Confidence Interval

$$\hat{y} \pm t_{1-\frac{\alpha}{2},n-2} s_{\in} \sqrt{\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$$

- We call this a confidence interval for the mean of observations with x=x_p **(Like sampling dist.)**
- This gives us a 100*(1- α) confidence interval for the mean of observations with x=x_p.

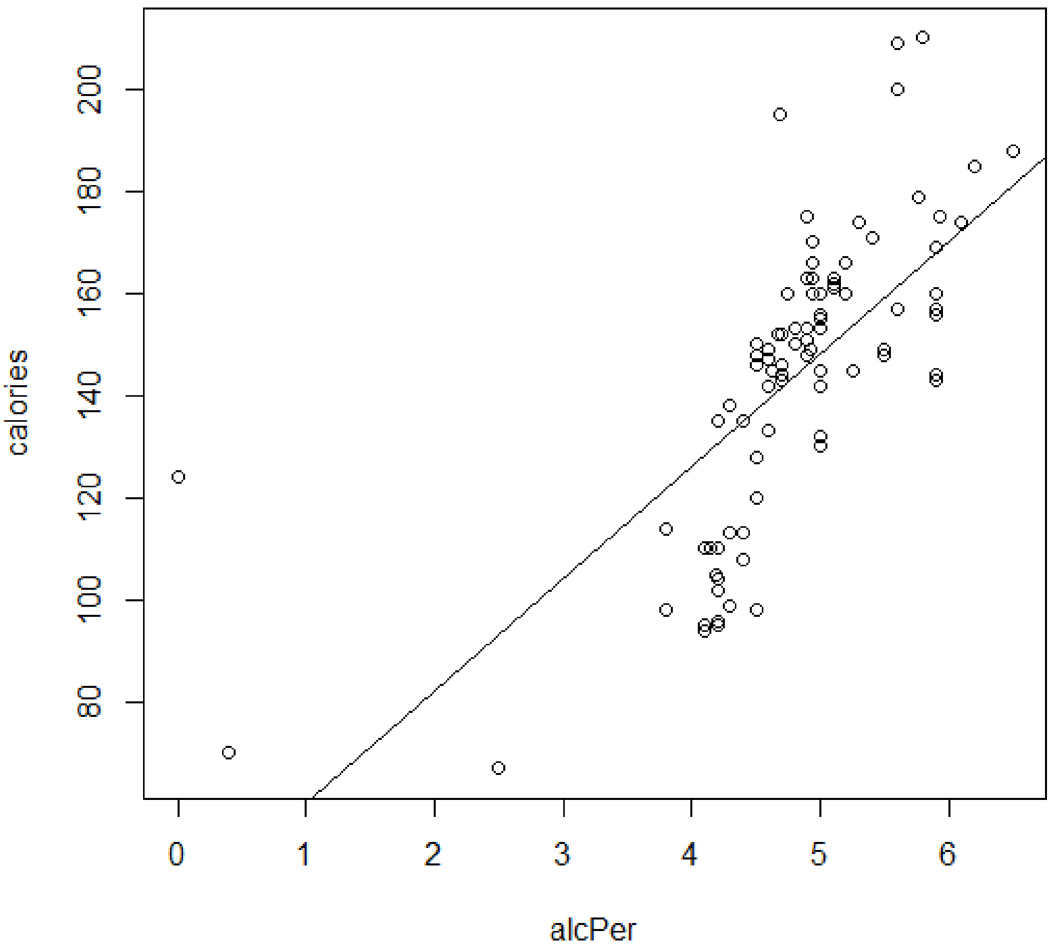
Regression Example

- In creating beer yeast and sugar react to create alcohol – the idea being, the more sugar and yeast you add the more alcohol the batch yields. It would then make sense that the more alcohol in the beer the more carbohydrates there are thus more calories – but who are we to make such assertions? Let us show it statistically.

Regression Example In R

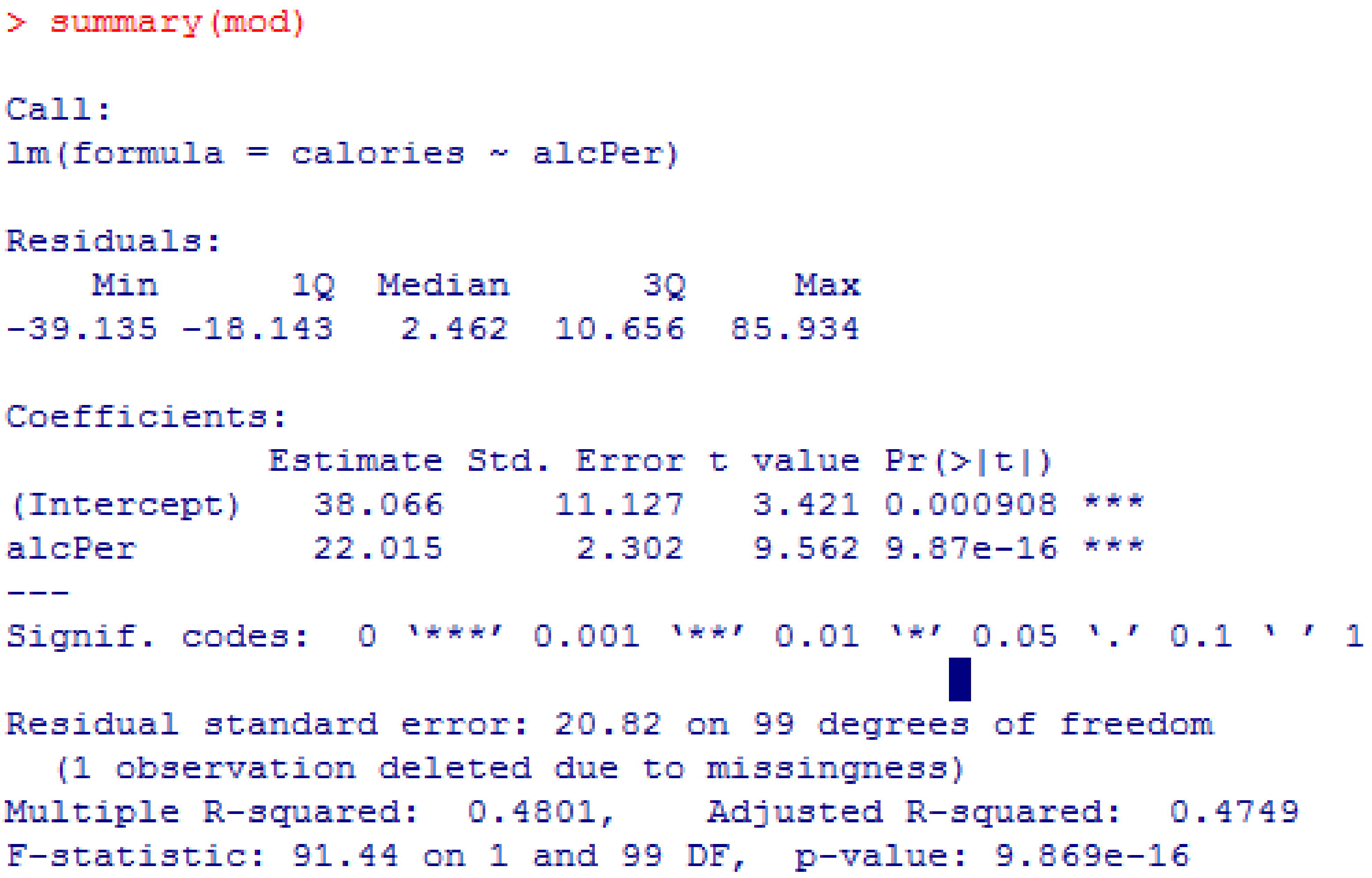
```
library('xlsx')
fileLocation="C:/Users/Will/Desktop/Beer.xlsx"
titles=TRUE
beerdata <- read.xlsx(fileLocation, 1,header=titles)

alcPer<-beerdata[,3]
calories<-beerdata[,4]

mod<-lm(calories~alcPer)
summary(mod)
anova(mod)

plot(alcPer, calories)
abline(mod)
```

Regression Example



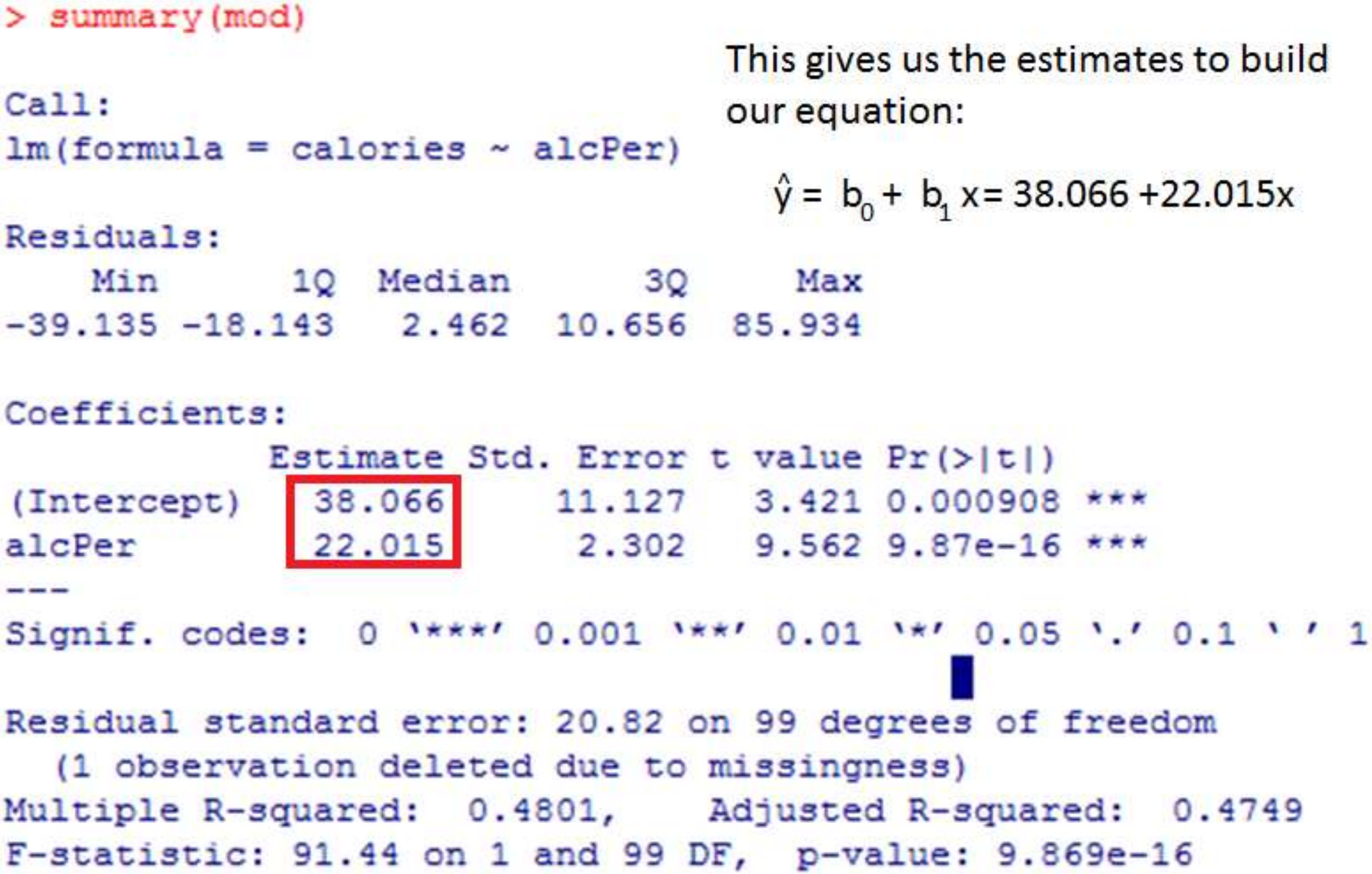
- The points are fit by the line pretty well, but we see that a couple points with alcPer below 1 are strange

Regression Example



```
> summary(mod)

Call:
lm(formula = calories ~ alcPer)

Residuals:
    Min       1Q   Median       3Q      Max 
-39.135 -18.143    2.462   10.656   85.934 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   38.066     11.127   3.421 0.000908 ***
alcPer        22.015      2.302   9.562 9.87e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 20.82 on 99 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4801,    Adjusted R-squared:  0.4749 
F-statistic: 91.44 on 1 and 99 DF,  p-value: 9.869e-16
```

Regression Example

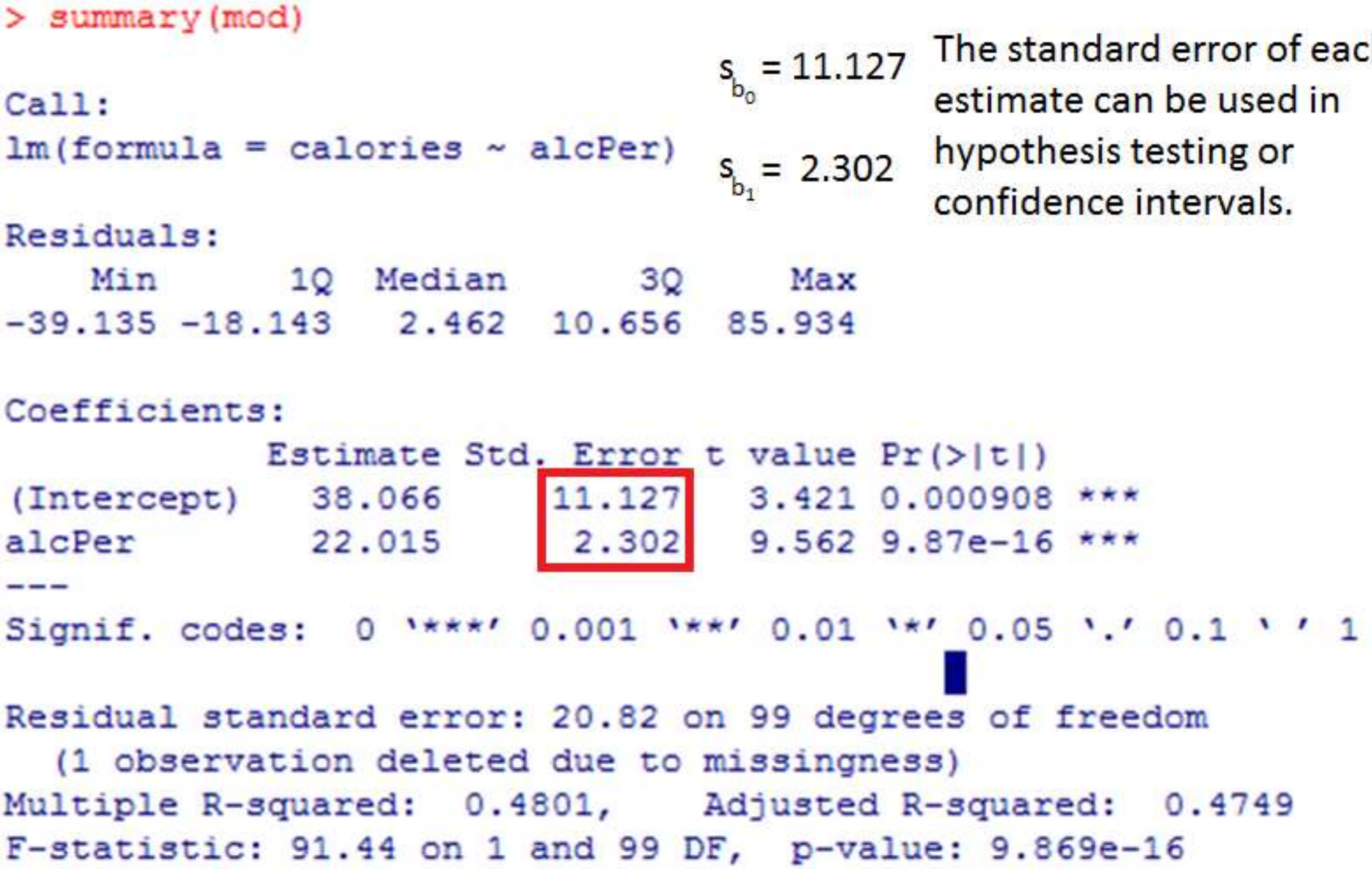


```
> summary(mod)

Call:
lm(formula = calories ~ alcPer)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.135 -18.143   2.462  10.656  85.934 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   38.066     11.127   3.421 0.000908 ***
alcPer        22.015      2.302   9.562 9.87e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.82 on 99 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4801,    Adjusted R-squared:  0.4749 
F-statistic: 91.44 on 1 and 99 DF,  p-value: 9.869e-16
```

This gives us the estimates to build our equation:

$$\hat{y} = b_0 + b_1 x = 38.066 + 22.015x$$

Regression Example



```
> summary(mod)

Call:
lm(formula = calories ~ alcPer)

Residuals:
    Min       1Q   Median       3Q      Max 
-39.135 -18.143    2.462   10.656   85.934 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   38.066     11.127   3.421 0.000908 ***
alcPer        22.015      2.302   9.562 9.87e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.82 on 99 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4801,    Adjusted R-squared:  0.4749 
F-statistic: 91.44 on 1 and 99 DF,  p-value: 9.869e-16
```

$s_{b_0} = 11.127$

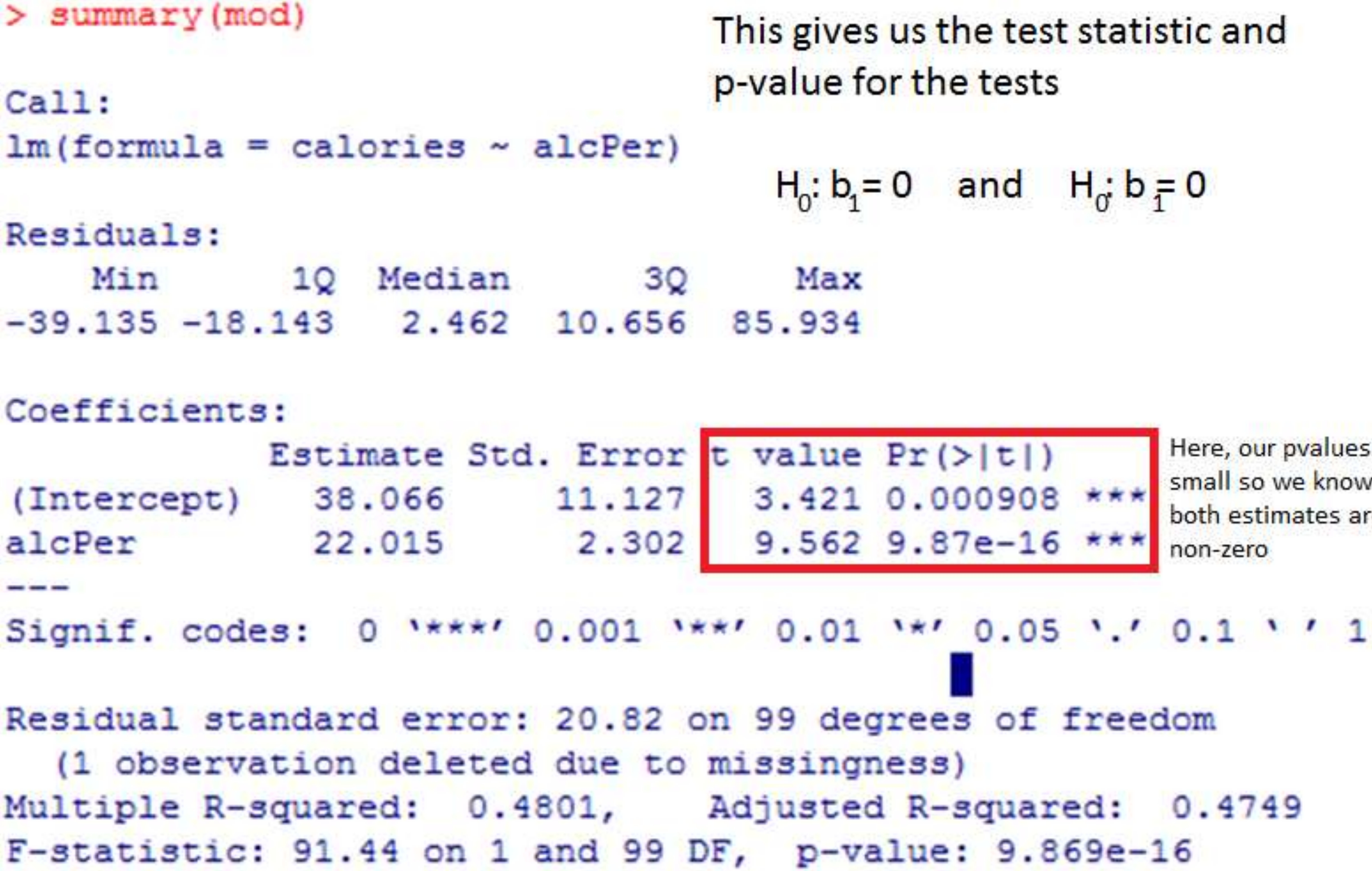
$s_{b_1} = 2.302$

The standard error of each estimate can be used in hypothesis testing or confidence intervals.

Regression Example

This gives us the test statistic and p-value for the tests

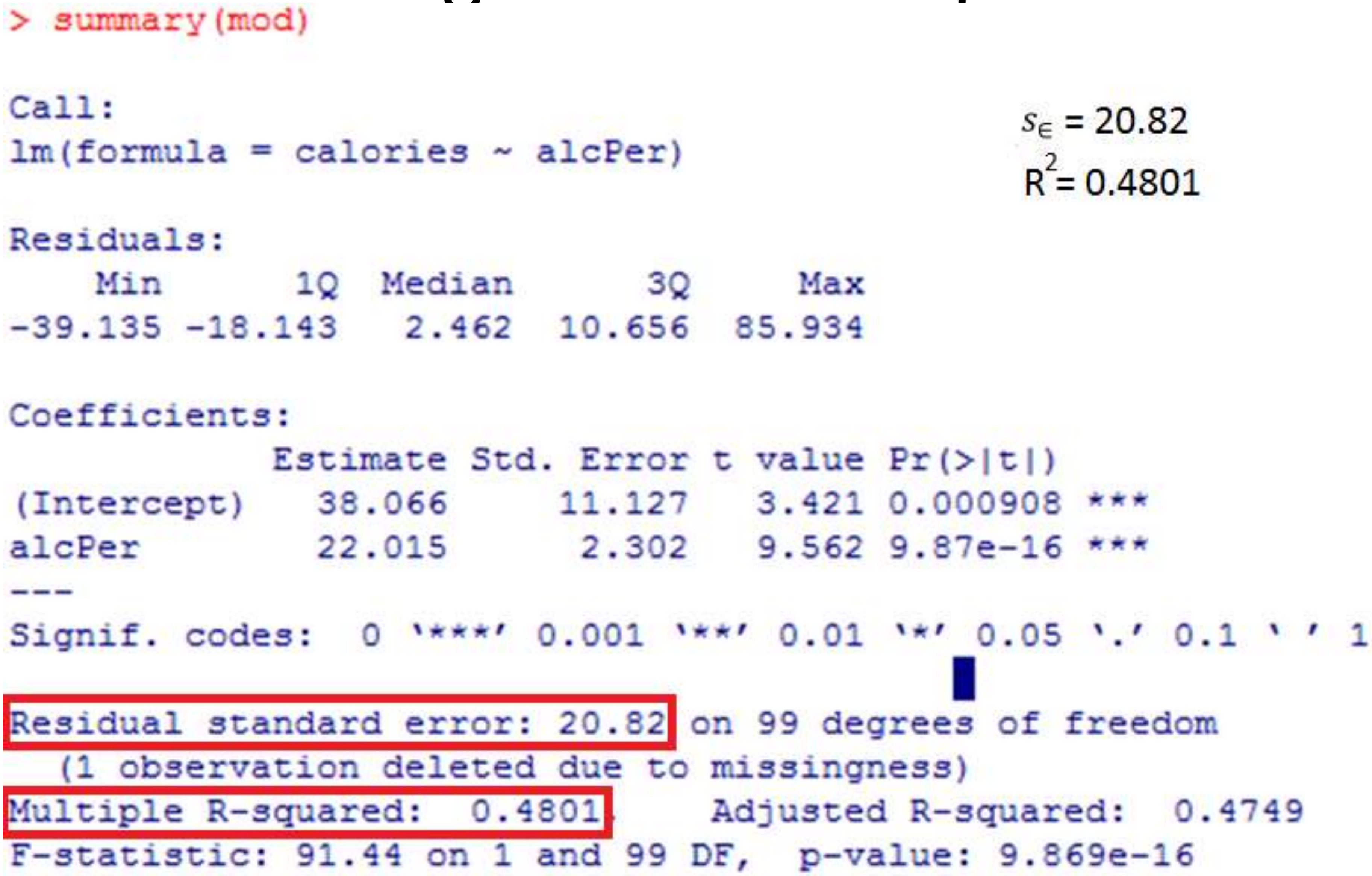
$H_0: b_1 = 0$ and $H_0: b_1 = 0$



```
> summary(mod)

Call:
lm(formula = calories ~ alcPer)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.135 -18.143   2.462  10.656  85.934 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   38.066     11.127   3.421 0.000908 ***
alcPer        22.015      2.302   9.562 9.87e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.82 on 99 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4801,    Adjusted R-squared:  0.4749 
F-statistic: 91.44 on 1 and 99 DF,  p-value: 9.869e-16
```

Here, our pvalues are small so we know both estimates are non-zero

Regression Example

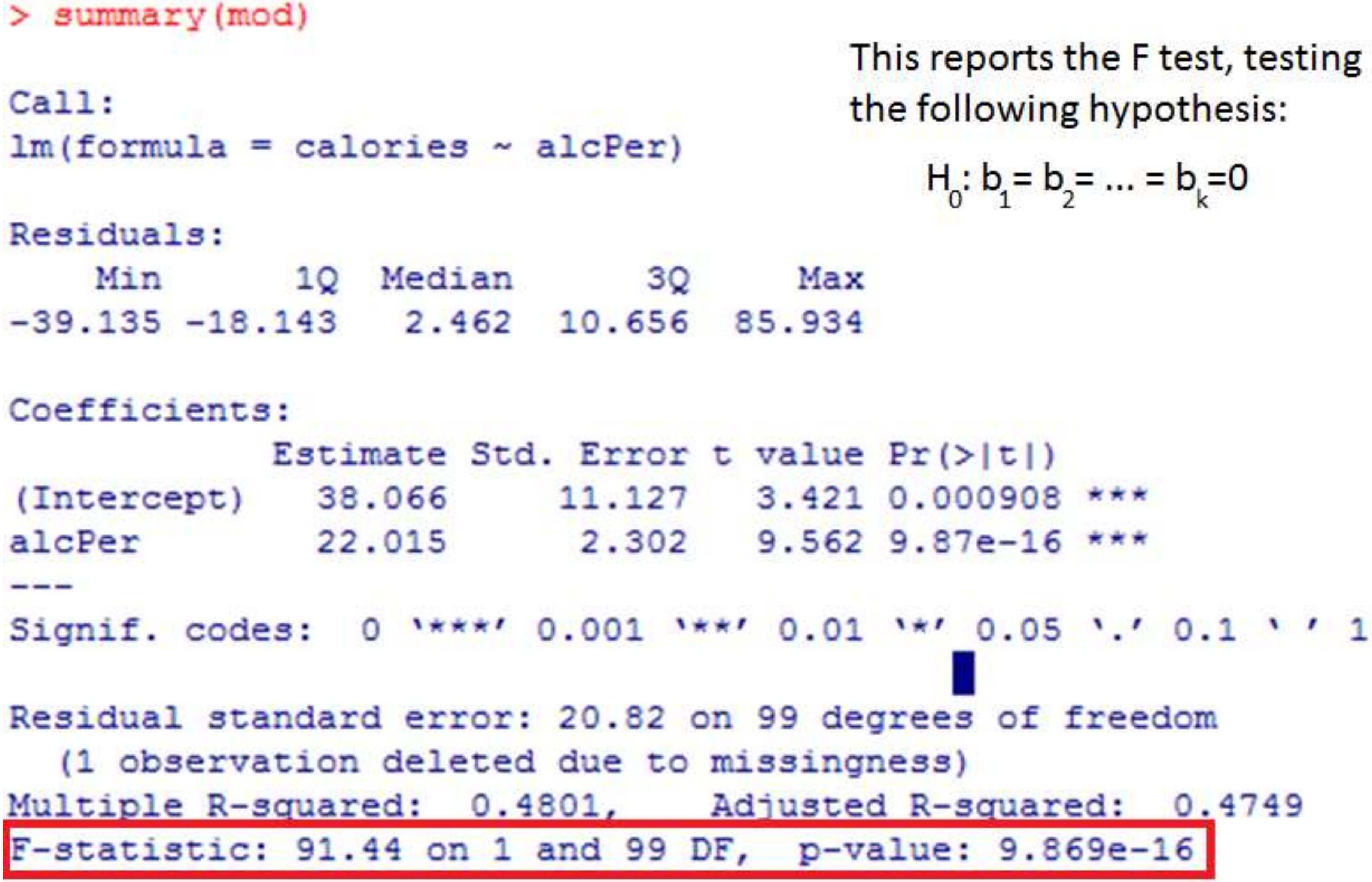


```
> summary(mod)

Call:
lm(formula = calories ~ alcPer)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.135 -18.143   2.462  10.656  85.934 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   38.066     11.127   3.421 0.000908 ***
alcPer        22.015      2.302   9.562 9.87e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.82 on 99 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4801,    Adjusted R-squared:  0.4749 
F-statistic: 91.44 on 1 and 99 DF,  p-value: 9.869e-16
```

$s_\epsilon = 20.82$

$R^2 = 0.4801$

Regression Example

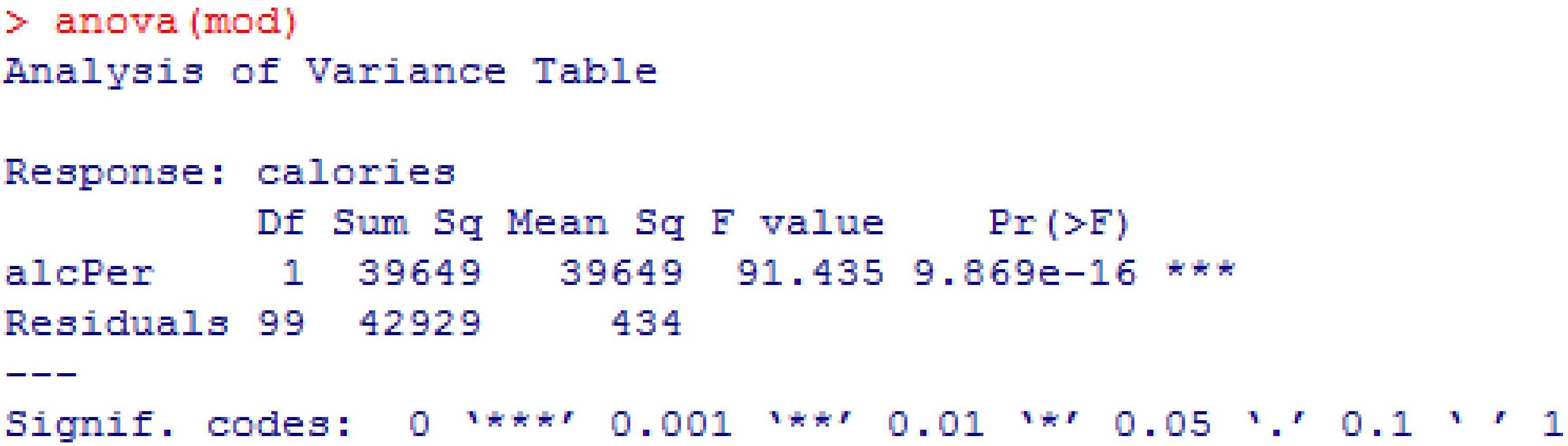


```
> summary(mod)

Call:
lm(formula = calories ~ alcPer)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.135 -18.143   2.462  10.656  85.934 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   38.066     11.127   3.421 0.000908 ***
alcPer        22.015      2.302   9.562 9.87e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.82 on 99 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4801,    Adjusted R-squared:  0.4749 
F-statistic: 91.44 on 1 and 99 DF,  p-value: 9.869e-16
```

This reports the F test, testing the following hypothesis:

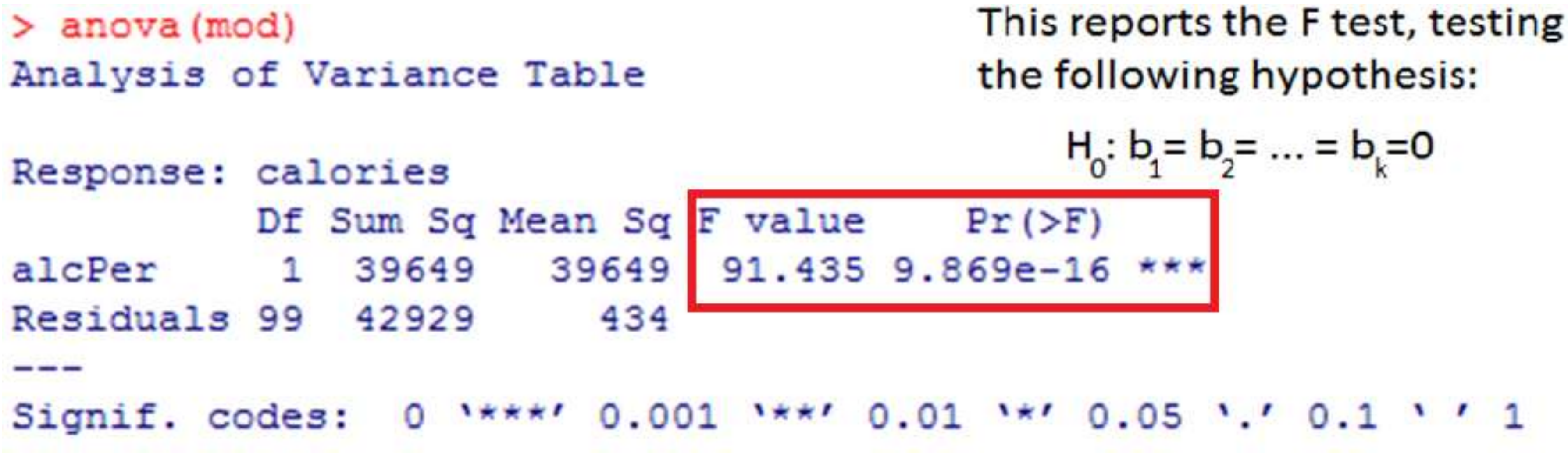
$$H_0: b_1 = b_2 = \ldots = b_k = 0$$

Regression Example



```
> anova(mod)
Analysis of Variance Table

Response: calories
          Df Sum Sq Mean Sq F value    Pr(>F)    
alcPer     1  39649   39649  91.435 9.869e-16 ***
Residuals 99  42929     434                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Regression Example

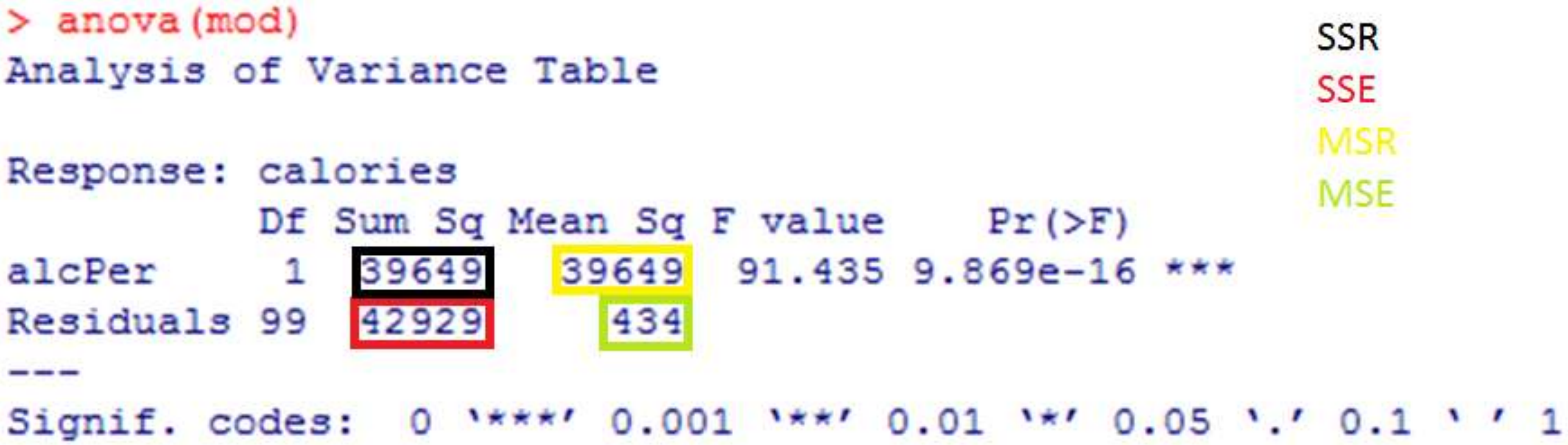


```
> anova(mod)
Analysis of Variance Table

Response: calories
          Df Sum Sq Mean Sq F value    Pr(>F)    
alcPer     1  39649   39649  91.435 9.869e-16 ***
Residuals 99  42929     434                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This reports the F test, testing the following hypothesis:

$$H_0: b_1 = b_2 = \dots = b_k = 0$$

Regression Example



```
> anova(mod)
Analysis of Variance Table

Response: calories
          Df Sum Sq Mean Sq F value    Pr(>F)
alcPer     1  39649   39649  91.435 9.869e-16 ***
Residuals 99  42929     434
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSR

SSE

MSR

MSE

Regression Example

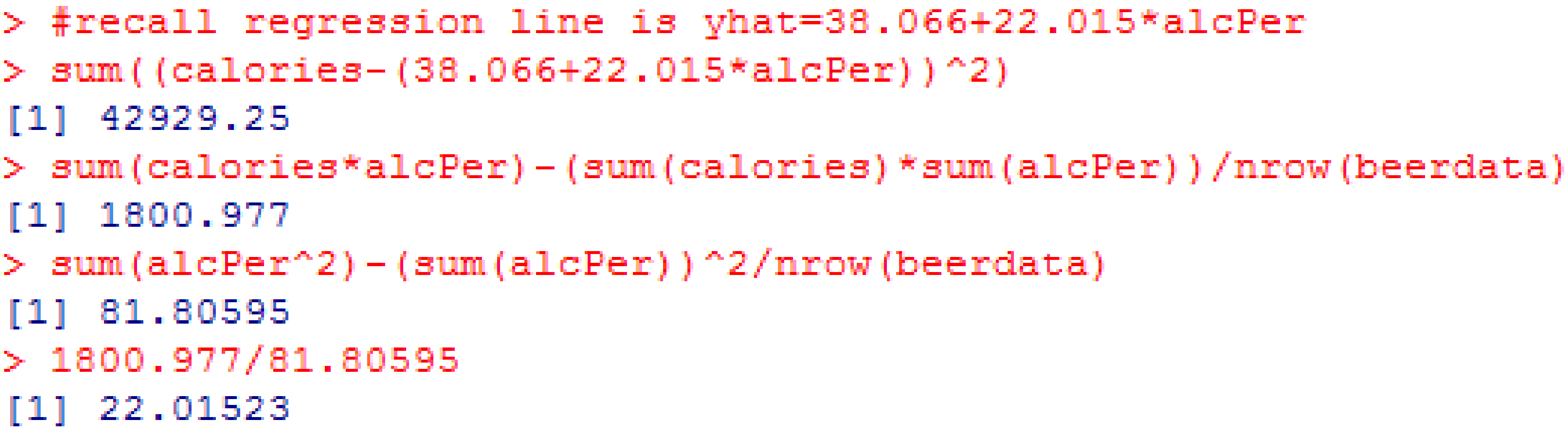
- $\boldsymbol{\widehat{y} = 38.066 + 22.015 * x} + \in$
    - Where $\boldsymbol{\widehat{y}}$ is the estimated calories and **x** is the alcPer
- The overall F test has a very low p-value so our model is capturing something
- $r\ = \sqrt{R^2} = \sqrt{.4801} = .6928925$ is close compared to our .7 benchmark so we have an “okay” model
- The P(>|t|) is small for alcPer so it is significant in terms of it’s ability to predict or model calories

Regression Example

- $SSE = \sum[y_i - (b_0 + b_1 x)]^2$
- $SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$
- $SS_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$

- Recall: $b_1 = slope = \frac{SS_{xy}}{SS_{xx}} = r \; * \left(\frac{s_y}{s_x}\right)$

Regression Example



```
> #recall regression line is yhat=38.066+22.015*alcPer
> sum((calories-(38.066+22.015*alcPer))^2)
[1] 42929.25
> sum(calories*alcPer)-(sum(calories)*sum(alcPer))/nrow(beerdata)
[1] 1800.977
> sum(alcPer^2)-(sum(alcPer))^2/nrow(beerdata)
[1] 81.80595
> 1800.977/81.80595
[1] 22.01523
```

```
sum((calories-(38.066+22.015*alcPer))^2)   #SSE
sum(calories*alcPer)-(sum(calories)*sum(alcPer))/nrow(beerdata) ##SSxy
sum(alcPer^2)-(sum(alcPer))^2/nrow(beerdata)  ##SSxx
1800.977/81.80595  ##B1
```

Regression Example

- (Est Calories) = 38.066+ 22.015 $*$ (Alcohol %)

  - $\boldsymbol{b_0}$**=** 38.066 is the intercept
    - The expected number of calories of a beer with 0% alcohol is 38.066.
  - $\boldsymbol{b_1}$=22.015 is the slope of the line
    - For every additional percent in alcohol percentage in beer we expect the number of calories to increase by 22.015 on average.

Confidence Intervals for Coefficients

- **A 100*(1-α)% confidence interval for $\boldsymbol{\beta_1}$:**

$$b_1 \pm \left(t_{1-\frac{.05}{2},101-2}\right)\widehat{s_{b_1}}$$

$$22.015 \pm (1.984217)(2.302)$$

$$(17.44733, 26.58267)$$

We are 95% confident that the true population coefficient, β_1, is between 17.45 and 26.58.

Hypothesis Testing for paired data μ unknown σ_d

Step One:	$H_0: \beta_1 = 23$ $H_a: \beta_1 \neq 23$
Step Two:	1. ∈ is normally distributed 2. E(∈)=0 3. Var(∈)=$\sigma_{\in}^2$ 4. Each error, ∈, is independent
Step Three:	$t^* = \frac{(22.015 - 23)}{2.302} = -.4278888$
Step Four:	$H_a: \beta_1$ ≠23→ p-value = 2*P(T<$-$|$-.4278888$|)=.669662
Step Five:	$.669662 > (1 - .95)$ =.05 → Fail to Reject H_0

Regression Example

- (Est Calories) = 38.066+ 22.015 $*$ (Alcohol %)

  - $\boldsymbol{R^2} = .4801$
    - 48.01% of the variation in calories in beer is explained by alcohol
  - $r\ = \sqrt{R^2} = \sqrt{.4801} = .6928925$
    - Since r is very close to one we have a **very strong** positive correlation

Regression Example

Step One:	$H_0: \rho \leq 0$ $H_a: \rho > 0$
Step Two:	1. The variables must be normally distributed 2. There is a linear relationship between the two variables 3. Outliers are either kept to a minimum or removed from the data 4. Equal variance across random variables
Step Three:	$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{(.6928925\sqrt{101-2})}{\sqrt{1-.4801}} = 9.561445$
Step Four:	$H_a: \rho > 0$ → p-value = P(T>9.561445) = 1-P(T<9.561445)$\approx$ 0
Step Five:	$0 \leq (1-.95) = .05$ → Reject H_0

Regression Example

- If we wanted to **estimate** the calories of my favorite beer, Rogue Dead Guy, we can plug in its alcohol percentage into the equation to find an estimate of the calories.
- If the Alcohol % = 6.5% for Rogue Dead Guy we can plug it in to find the estimated calories of Rogue Dead Guy.
- (Est Calories) = 38.066+ 22.015 $*$ (Alcohol %)

$$= 38.066 + 22.015 * (6.5)$$

$$= 181.1635$$

Regression Example

- So, the estimated amount of calories for Rogue Dead Guy is 181.163587.
- In fact, the actual amount of calories is 250 so our estimate isn’t very good, though it might be better than a random guess.
- **The residual** is the difference
    - true – estimate = 250-181.163587=68.83641.

Regression Example

$$\hat{y} \pm t_{1-\frac{.05}{2},101-2} s_{\in} \sqrt{\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$$

$$181.1635 \pm (1.984217)(20.82)\sqrt{\frac{1}{101} + \frac{(6.5 - 4.738416)^2}{81.80595}}$$

$$(172.1283, 190.1987)$$

We are 95% confident that the true mean, y, of all observations of alcPer 6.5 is between 172.1283 and 190.1987

Regression Example

$$\hat{y} \pm t_{1-\frac{.05}{2},101-2} s_{\in} \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$$

$$181.1635 \pm (1.984217)(20.82)\sqrt{1 + \frac{1}{101} + \frac{(6.5 - 4.738416)^2}{81.80595}}$$

$$(138.8756, 223.4514)$$

We are 95% confident that the true value, y, of an observation of alcPer 6.5 is between 138.8756 and 223.4514.

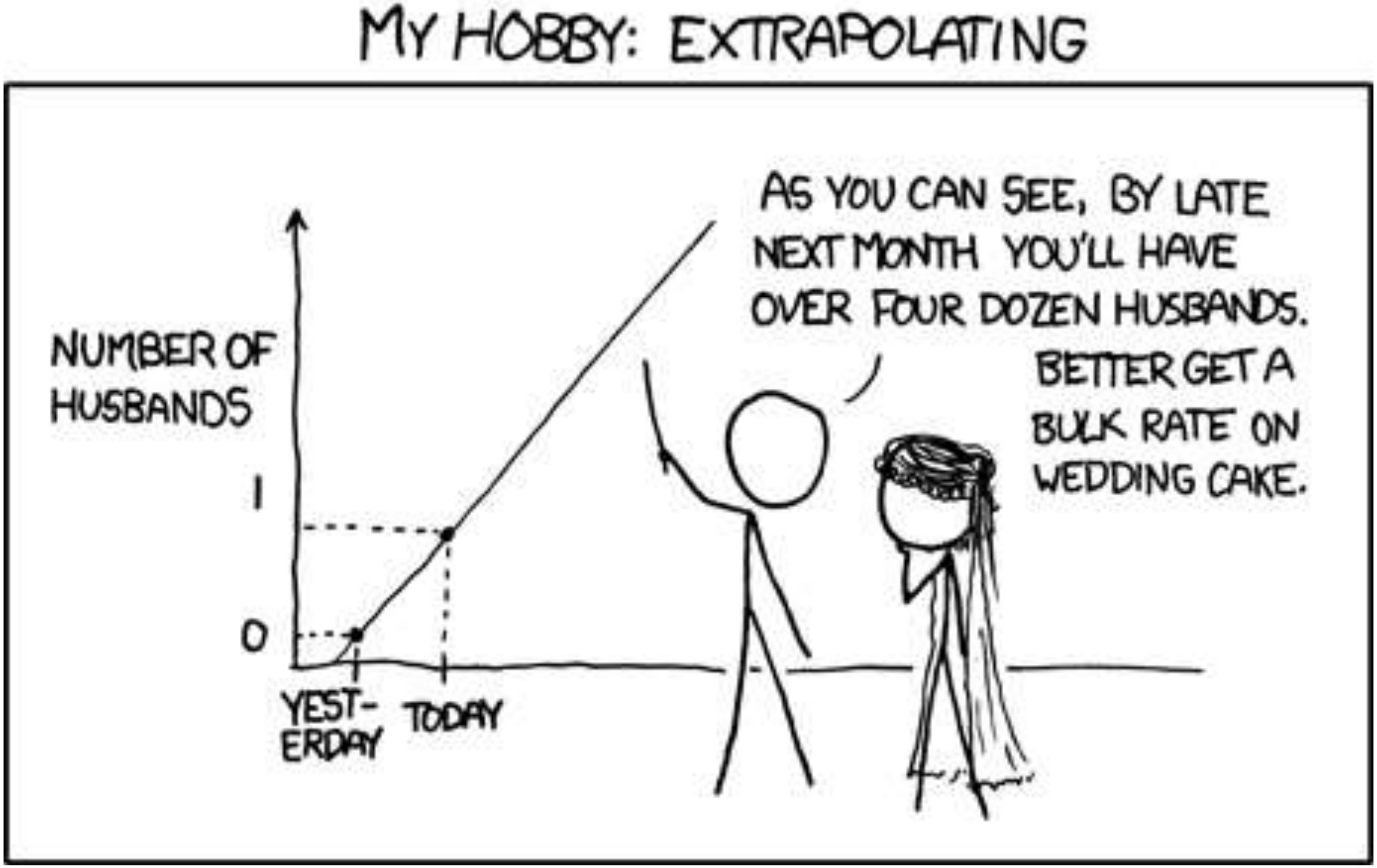
Regression Example

- Issues:
  - Look at the data: was that “weird point” an outlier? Is it representative?
  - What happens if we “fix” it or remove it from the data set?
  - Did we prove that alcPer causes calories in beer?
  - Is our model okay to use with a beer that has 6.5% alcohol percentage?

Regressions – Problems

- **Extrapolation** – We don’t want to predict using x values different than the known data
- **Influential Outliers** – a single point can really change the fit of the regression line – always check for stray points in the scatterplot
- **Correlation does not imply causation** – wait for it
- **Lurking Variables** – a variable that we don’t look at that causes the correlation (hot summers)

Regressions – Problems


MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS.
BETTER GET A BULK RATE ON WEDDING CAKE.

Credit: XKCD

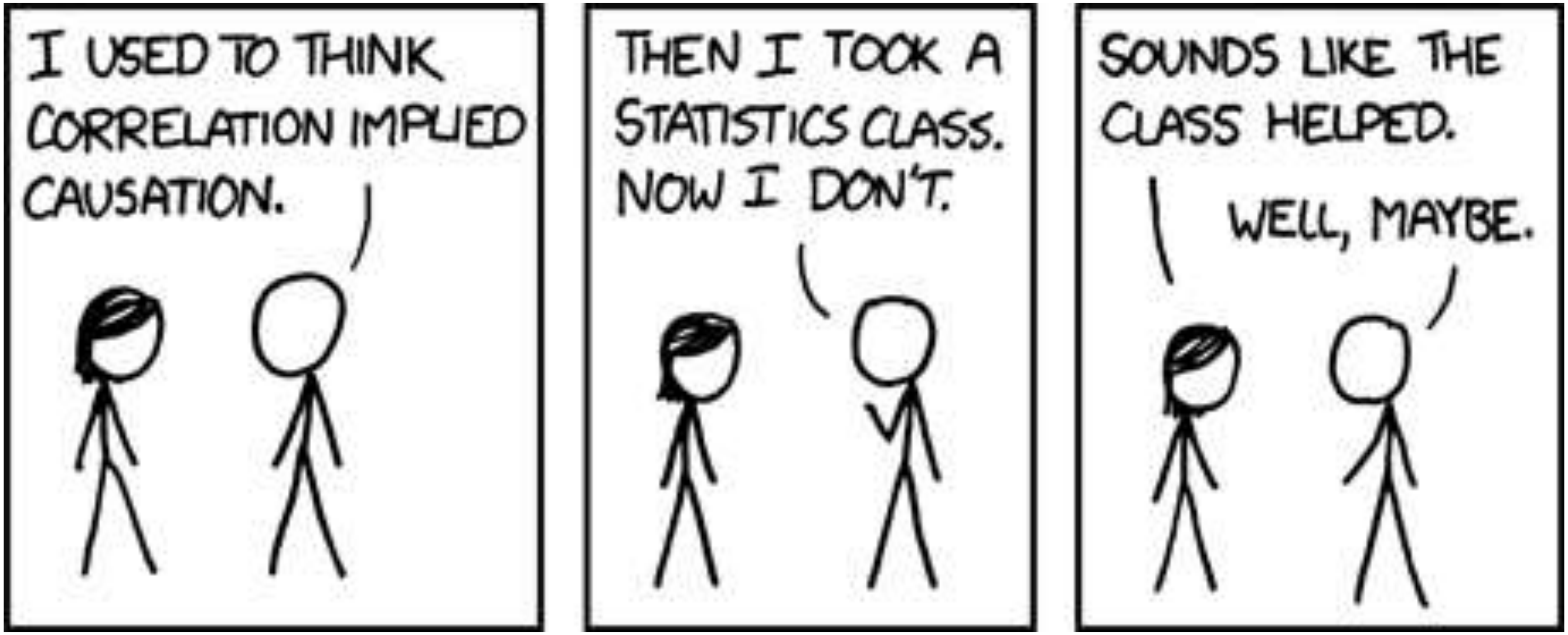
Correlation vs. Causation

- The idea here is that although some variables are correlated they one might not be the cause of the other.

- Let’s revisit http://www.tylervigen.com/?categoria=%22dinero%22

Regressions – Problems

- Correlation does not imply causation
  - We go from saying there exists a correlation to saying that one variable's change causes the other to change.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Credit: XKCD

Regression Introduction

- Set up and example using Facebook!*
  - https://www.youtube.com/watch?v=zPG4NjIkCjc

Summary!

Response Variable	this is our dependent variable, the outcome variable on which comparisons are made
Explanatory Variable	this is our independent variable, the groups to be compared with respect to values on the response variable
Association or Correlation	exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable

Comparing Two Quantitative Variables

Scatterplot	let the **response** variable be the y variable and the **explanatory** variable be the x variable and plot the points
Regression Line	predicts the value for the response variable y as a straight line function of the value of x, the explanatory variable $\boldsymbol{\hat{y} = b_1 * x + b_0}$
Intercept	$\boldsymbol{b_0}$ this is the expected value of y when x is zero
Slope	$\boldsymbol{b_1}$ this is the amount that $\boldsymbol{\hat{y}}$ changes by when x increases by one unit

Comparing Two Quantitative Variables

$\boldsymbol{R^2} = r^2$	gives the percent of variation in y explained by x
$\boldsymbol{r} = \sqrt{R^2}$	measures the **LINEAR** relationship between x and y
Estimate $\boldsymbol{\widehat{y}}$ for a given x	Plug x into the regression equation $\boldsymbol{\widehat{y} = b_1 * x + b_0}$
Residual	**Residual = (the real y) – $\boldsymbol{\widehat{y}}$**

Coefficient of Correlation (r)

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

Where:

$$SS_{xy} = \sum(x - \bar{x})(y - \bar{y})$$

$$SS_{xx} = \sum(x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$SS_{yy} = \sum(y - \bar{y})^2 = \sum y^2 - \frac{(\sum y)^2}{n}$$

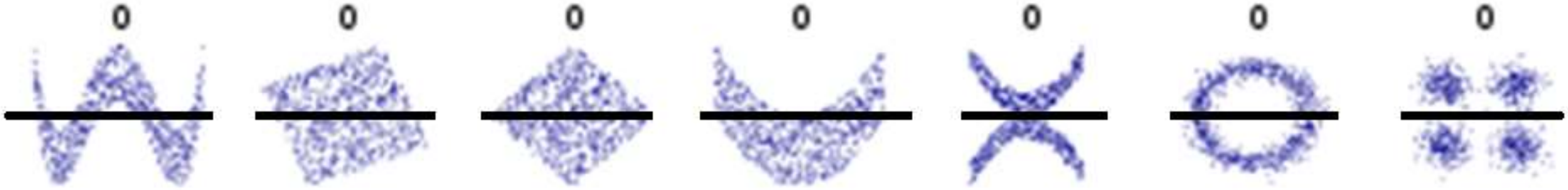
Coefficient of Correlation (r)

- **r** measures the **LINEAR** relationship between x and y [linear, linear, linear, linear!!!]
- **r > 0** → positive correlation or association
- **r < 0** → negative correlation or association
- **r=1** → perfect positive correlation or association
  - Here, all points would fit on a line
- **r=-1** → perfect negative correlation
  - Here, all points would fit on a line
- **r=0** → no correlation

Properties

- $-1 \leq r \leq 1$
- The closer r is to 1 the stronger the evidence for positive association
- The closer r is to -1 the stronger the evidence for a negative association
- The closer r is to 0 the weaker the evidence for association
- **Affected by outliers so we have to be careful**

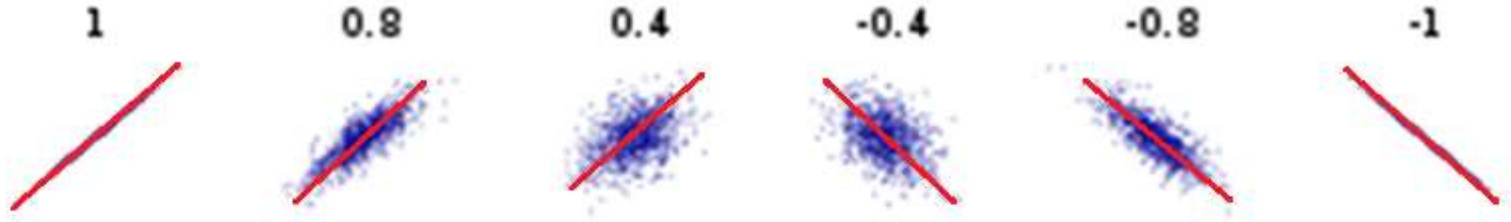
Coefficient of Correlation Examples



- Here there are obvious patterns here **but** they are not linear!
- Since, r, measures the **linear** relationship between two variables r=0 even though there are patterns!
  - In each case the best balanced line would be horizontal

Regression – Making Lines Useful!

- **Regression Line –** we make our regression line so that it best fits our data – unlike math it usually isn't a perfect



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

  - Note: we need to estimate these values with our data

Coefficient of Correlation (r)

$$r^2 = \frac{Explained\ sample\ variability}{total\ variability}$$

$$= \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

Where:

$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

$$SSE = \sum (y_i - \hat{y})^2$$

When Can We Use Regression?

- $\boldsymbol{R^2}$, given in the regression output, gives the percent of variation in the response variable explained by the explanatory variable

- **Note:** $R^2 = r^2$
- **Note:** $r = \sqrt{R^2}$

Estimating σ for a Simple Linear Model

$$s_{\in}^{2} = \frac{SSE}{Degrees\ of\ Freedom\ for\ Error}$$

$$= \frac{\mathrm{SSE}}{\mathrm{n{-}2}}$$

Where: $SSE = \sum(y_i - \widehat{y_i})^2 = SS_{yy} - b_1 SS_{xy}$

Then: $s_{\in} = \sqrt{s_{\in}^{2}} = \sqrt{\frac{\mathrm{SSE}}{\mathrm{n{-}2}}}$

Hypothesis Testing for paired data μ unknown σ_d

Step One:	(i) $\mathrm{H_0}: \rho \leq \rho_o$ & $\mathrm{H_a}: \rho > \rho_o$ (ii) $\mathrm{H_0}: \rho \geq \rho_o$ & $\mathrm{H_a}: \rho < \rho_o$ (iii) $\mathrm{H_0}: \rho = \rho_o$ & $\mathrm{H_a}: \rho \neq \rho_o$
Step Two:	1. The variables must be normally distributed 2. There is a linear relationship between the two variables 3. Outliers are either kept to a minimum or removed from the data 4. Equal variance across random variables
Step Three:	$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{b_1}{\sqrt{\frac{SSE}{n-2}}}$$
Step Four:	*(i)* $\mathrm{H_a}: \rho > \rho_o$→ p-value = P(T>t*) = 1-P(T<t*) (ii) $\mathrm{H_a}: \rho < \rho_o$→ p-value = P(T<t*) (iii) $\mathrm{H_a}: \rho \neq \rho_o$→p-value = 2*P(T<-|t*|)
Step Five:	If p-value$\leq (1 - \mathrm{confidene}) = \alpha$ → Reject $\mathrm{H_0}$ If p-value$> (1 - \mathrm{confidence}) = \alpha$ → Fail to Reject $\mathrm{H_0}$

Sampling Distribution of b_1

- If the assumptions for $\in$ hold:

  $\mu_{b_1} = \beta_1$

  $\sigma_{b_1} = \frac{\sigma_\in}{\sqrt{SS_{xx}}}$ which we estimate with $\frac{s_\in}{\sqrt{SS_{xx}}}$

- We call σ_{b_1} the **estimated standard error of** b_1

Confidence Intervals for Coefficients

- **Check the assumptions:**
  - ∈ is normally distributed
  - E(∈)=0
  - var(∈)=σ^2
  - Each error, ∈, is independent

- **A 100*(1-α)% confidence interval for $\boldsymbol{\beta_1}$:**

$$b_1 \pm \left(t_{1-\frac{\alpha}{2},n-2}\right)\widehat{s_{b_1}}$$

Hypothesis Testing for paired data
μ unknown σ_d

Step One:	(i) $H_0: \beta_1 \le \beta_{1_o}$ & $H_a: \beta_1 > \beta_{1_o}$ (ii) $H_0: \beta_1 \ge \beta_{1_o}$ & $H_a: \beta_1 < \beta_{1_o}$ (iii) $H_0: \beta_1 = \beta_{1_o}$ & $H_a: \beta_1 \ne \beta_{1_o}$
Step Two:	1. ∈ is normally distributed 2. E(∈)=0 3. Var(∈)=$\sigma_{\in}^2$ 4. Each error, ∈, is independent
Step Three:	$t^* = \frac{(b_1 - \mu_{b_1})}{\sigma_{b_1}} = \frac{(b_1 - \beta_{1_o})}{\frac{s_{b_1}}{\sqrt{SS_{xx}}}}$
Step Four:	(i) $H_a: \beta_1 > \beta_{1_o}$→ p-value = P(T>t*) = 1-P(T<t*) (ii) $H_a: \beta_1 < \beta_{1_o}$→ p-value = P(T<t*) (iii) $H_a: \beta_1 \ne \beta_{1_o}$→ p-value = 2*P(T<-|t*|)
Step Five:	If p-value$\le (1 - confidene) = \alpha$ → Reject H_0 If p-value$> (1 - confidence) = \alpha$ → Fail to Reject H_0

Prediction/Estimation: Sampling Errors

- $\sigma_y = \sigma_\in \sqrt{\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$

- $\sigma_{(y-\hat{y})} = \sigma_\in \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$

Prediction/Estimation: Confidence Interval

$$\hat{y} \pm t_{1-\frac{\alpha}{2},n-2} s_{\in} \sqrt{\frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$$

$$\hat{y} \pm t_{1-\frac{\alpha}{2},n-2} s_{\in} \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{SS_{xx}}}$$

Regression Example In R

```r
library('xlsx')
fileLocation="C:/Users/Will/Desktop/Beer.xlsx"
titles=TRUE
beerdata <- read.xlsx(fileLocation, 1,header=titles)

alcPer<-beerdata[,3]
calories<-beerdata[,4]

mod<-lm(calories~alcPer)
summary(mod)
anova(mod)

plot(alcPer, calories)
abline(mod)
```