Author: Brenda Gunderson, Ph.D., 2015

Stat 250 Gunderson Lecture Notes
11: Regression Analysis

> The invalid assumption that correlation implies cause is probably among the two or three most serious and common errors of human reasoning.
>
> *--Stephen Jay Gould, The Mismeasure of Man* © FAIR USE

Describing and assessing the significance of **relationships between variables** is very important in research. We will first learn how to do this in the case when the two variables are quantitative. Quantitative variables have numerical values that can be ordered according to those values.

Main idea

We wish to study the relationship between two quantitative variables.

Generally one variable is the _________________________ ***variable***, denoted by *y*.
This variable measures the outcome of the study
and is also called the _________________________ variable.

The other variable is the _________________________ ***variable***, denoted by *x*.
It is the variable that is thought to explain the changes we see in the response variable.

The explanatory variable is also called the _________________ variable.

The first step in examining the relationship is to use a graph - a ***scatterplot*** - to display the relationship. We will look for an overall pattern and see if there are any departures from this overall pattern.

If a ***linear*** relationship appears to be reasonable from the scatterplot, we will take the next step of finding a model (an equation of a line) to summarize the relationship. The resulting equation may be used for predicting the response for various values of the explanatory variable. If certain assumptions hold, we can assess the significance of the linear relationship and make some confidence intervals for our estimations and predictions.

Let's begin with an example that we will carry throughout our discussions.

Graphing the Relationship: Restaurant Bill vs Tip

How well does the size of a restaurant bill predict the tip the server receives? Below are the bills and tips from six different restaurant visits in dollars.

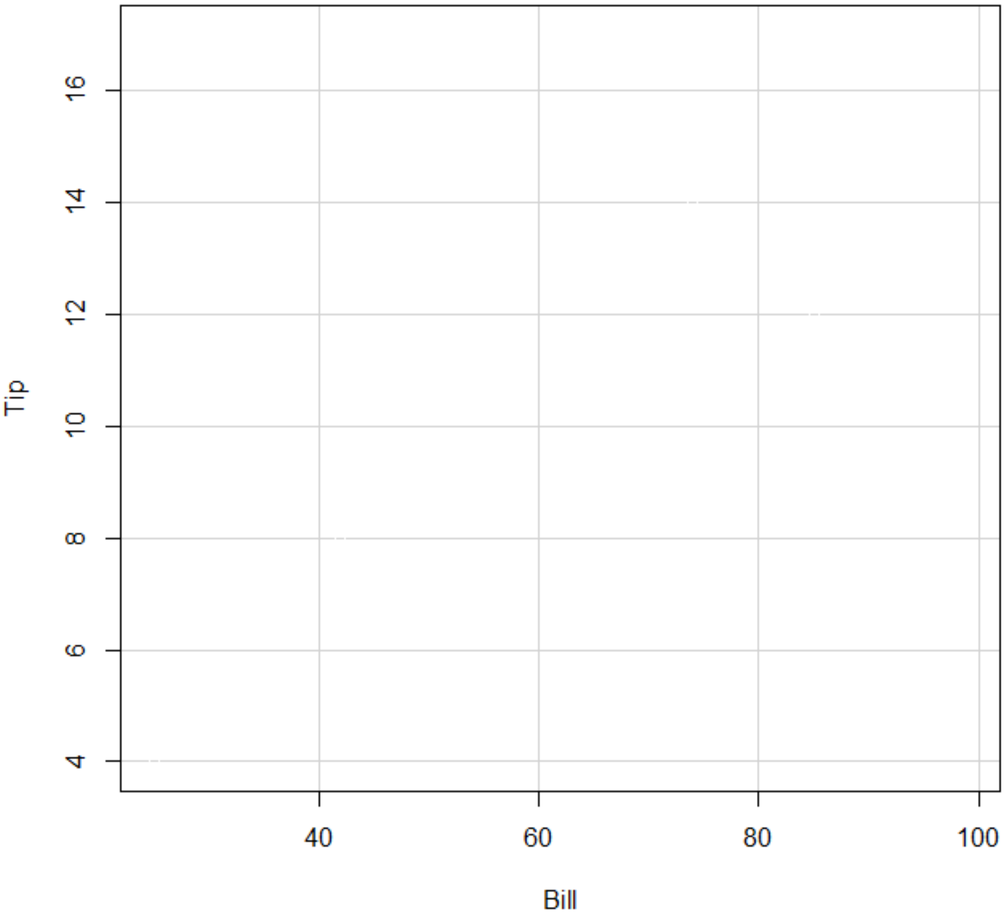
Bill	41	98	25	85	50	73
Tip	8	17	4	12	5	14

Response (dependent) variable *y* = ______________________________ .

Explanatory (independent) variable *x* = ______________________________ .

Step 1: Examine the data graphically with a scatterplot.

Add the points to the scatterplot below:



Interpret the scatterplot in terms of ...

- **overall form** (is the average pattern look like a straight line or is it curved?)
- **direction** of association (positive or negative)
- **strength** of association (how much do the points vary around the average pattern?)
- any **deviations** from the overall form?

Describing a Linear Relationship with a Regression Line

Regression analysis is the area of statistics used to examine the relationship between a quantitative response variable and one or more explanatory variables. A key element is the **estimation of an equation** that describes how, on average, the response variable is related to the explanatory variables. A regression equation can also be used to make predictions.

The simplest kind of relationship between two variables is a straight line, the analysis in this case is called **linear regression.**

Regression Line for Bill vs. Tip

Remember the equation of a line?
In statistics we denote the **regression line for a sample** as:
where:

$\hat{y}$

b_0

b_1

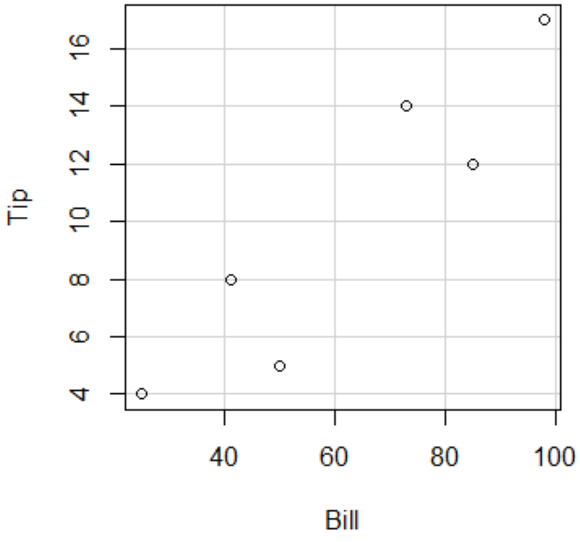
Goal:
To find a line that is "close" to the data points - find the "best fitting" line.

How?
What do we mean by best?
One measure of how good a line fits is to look at the "observed errors" in prediction.

Observed errors = ______________________

are called _____________________

So we want to choose the line for which the sum of squares of the observed errors (the sum of squared residuals) is the ***least***.



The line that does this is called:____

The equations for the estimated slope and intercept are given by:

b_1 =

b_0 =

The least squares regression line (estimated regression function) is: $\hat{y} = \hat{\mu}_y(x) = b_0 + b_1 x$

To find this estimated regression line for our exam data by hand, it is easier if we set up a calculation table. By filling in this table and computing the column totals, we will have all of the main summaries needed to perform a complete linear regression analysis. Note that here we have *n* = 6 observations. The first five rows have been completed for you. In general, use R or a calculator to help with the graphing and numerical computations!

x = bill	**y = tip**	$x - \bar{x}$	$(x - \bar{x})^2$	$(x - \bar{x})y$	$y - \bar{y}$	$(y - \bar{y})^2$
41	8	41–62 = -21	$(-21)^2$ = 441	(-21)(8)= -168	8–10 = -2	$(-2)^2$ = 4
98	17	98–62 = 36	$(36)^2$ = 1296	(36)(17)= 612	17–10 = 7	$(7)^2$ = 49
25	4	25–62 = -37	$(-37)^2$ = 1369	(-37)(4)= -148	4–10 = -6	$(-6)^2$ = 36
85	12	85–62 = 23	$(23)^2$ = 529	(23)(12)= 276	12–10 = 2	$(2)^2$ = 4
50	5	50–62 = -12	$(-12)^2$ = 144	(-12)(5)= -60	5–10 = -5	$(-5)^2$ = 25
73	14					
372	**60**					

$$\bar{x} = \frac{372}{6} = \mathbf{62} \qquad \bar{y} = \frac{60}{6} = \mathbf{10}$$

Slope Estimate:

y-intercept Estimate:

Estimated Regression Line:

Predict the tip for a dinner guest who had a $50 bill.

Note: The 5th dinner guest in sample had a bill of $50 and the observed tip was $5.

Find the residual for the 5th observation.

Notation for a residual $= e_5 = y_5 - \hat{y}_5 =$

The residuals …

You found the residual for one observation. You could compute the residual for each observation. The following table shows each residual.

x = bill	**y = tip**	**predicted values** $\hat{y} = -0.5877 + 0.17077(x)$	**residuals** $e = y - \hat{y}$	**Squared residuals** $(e)^2 = (y - \hat{y})^2$
41	8	6.41	1.59	2.52
98	17	16.15	0.85	0.72
25	4	3.68	0.32	0.10
85	12	13.93	-1.93	3.73
50	5	7.95	-2.95	8.70
73	14	11.88	2.12	4.49
--	--	**--**		

SSE = sum of squared errors (or residuals) ≈

Measuring Strength and Direction of a Linear Relationship with Correlation

The ***correlation coefficient r*** is a measure of strength of the linear relationship between *y* and *x*.

Properties about the Correlation Coefficient *r*

1. r ranges from ...

2. Sign of r indicates ...

3. Magnitude of r indicates ...

   A “strong” *r* is discipline specific
   r = 0.8 might be an important (or strong) correlation in engineering
   r = 0.6 might be a strong correlation in psychology or medical research

4. r ONLY measures the strength of the _____________________ relationship.

Some pictures:

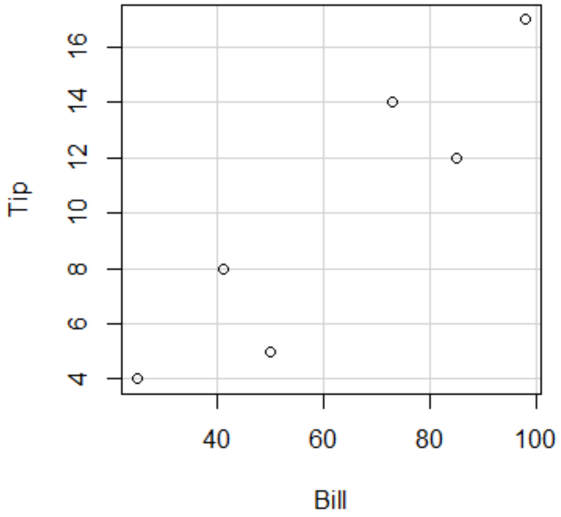
The **formula** for the correlation:
(but we will get it from computer output or from r^2)

$$r = \frac{1}{n-1}\sum_i \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

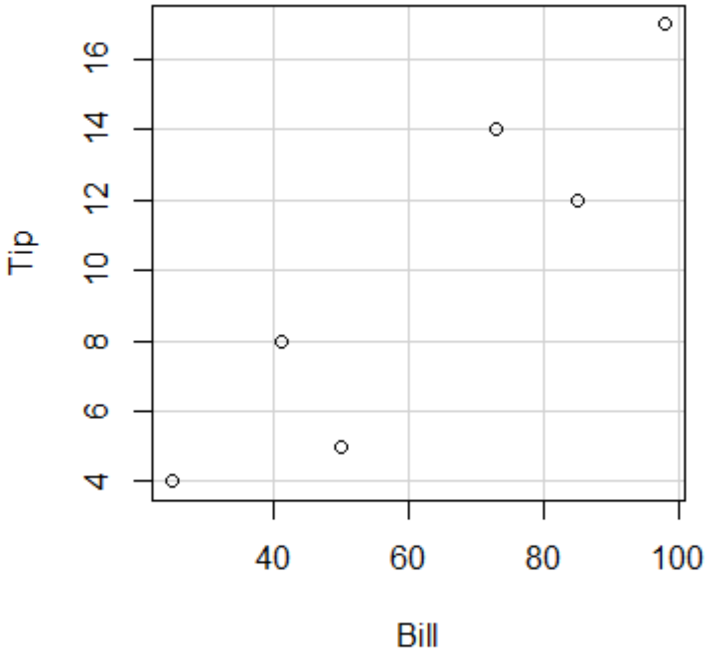
Tips Example:

r **= _____________**

Interpretation:



The square of the correlation r^2



The squared correlation coefficient r^2 always has a value between ____________ and is sometimes presented as a percent. It can be shown that the square of the correlation is related to the sums of squares that arise in regression.

The responses (the amount of tip) in data set are not all the same - they do vary. We would measure the ***total variation*** in these responses as $\text{SSTO} = \sum(y - \bar{y})^2$ (last column total in calculation table said we would use later).

Part of the reason why the amount of tip varies is because there is a linear relationship between amount of tip and amount of bill, and the study included different amounts of bill.

When we found the least squares regression line, there was still some small variation remaining of the responses from the line. This amount of **variation that is not accounted for by the linear relationship** is called the **SSE**.

The amount of **variation that is accounted for by the linear relationship** is called the sum of squares due to the model (or regression), denoted by **SSM** (or sometimes as SSR).

So we have: **SSTO = ____**

It can be shown that

r^2 =

= the proportion of total variability in the responses that can be explained by the linear relationship with the explanatory variable x.

Note: The value of r^2 and these sums of squares are summarized in an ***ANOVA table*** that is standard output from computer packages when doing regression.

Measuring Strength and Direction for Exam 2 vs Final

From our first calculation table we have:

SSTO = ____

From our residual calculation table we have:

SSE = _______________________________________

So the squared correlation coefficient for our exam scores regression is:

$$r^2 = \frac{SSTO - SSE}{SSTO} =$$

Interpretation:

We accounted for ______________ % of the variation in _________________________

by the linear regression on ____________________ .

The correlation coefficient is *r* = ____________________

A few more general notes:

- Nonlinear relationships
- Detecting Outliers and their influence on regression results.
- Dangers of Extrapolation (predicting outside the range of your data)
- Dangers of combining groups inappropriately (Simpson’s Paradox)
- Correlation does not prove causation

R Regression Analysis for Bill vs Tips

Let’s look at the R output for our Bill and Tip data.
We will see that much of the computations are done for us.

```
Call:
lm(formula = Tip ~ Bill, data = Tips)


Residuals:
      1       2       3       4       5       6
 1.5862  0.8523  0.3185 -1.9277 -2.9508  2.1215



Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.58769    2.41633  -0.243  0.81980
Bill         0.17077    0.03604   4.738  0.00905 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.251 on 4 degrees of freedom
Multiple R-squared:  0.8487,    Adjusted R-squared:  0.8109
F-statistic: 22.45 on 1 and 4 DF,  p-value: 0.009052
```

Correlation “Matrix”

```
          Bill       Tip
Bill 1.0000000 0.9212755
Tip  0.9212755 1.0000000
```

ANOVA Table

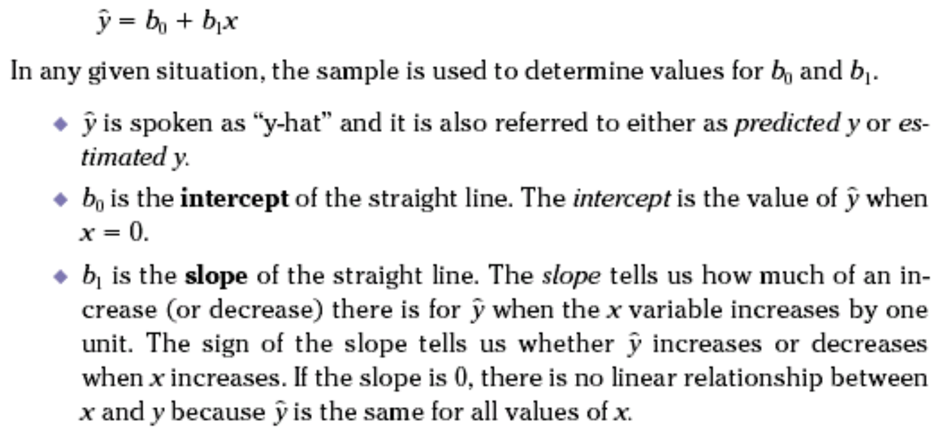
```
Response: Tip
          Df  Sum Sq Mean Sq F value   Pr(>F)
Bill       1 113.732 113.732  22.446 0.009052 **
Residuals  4  20.268   5.067
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Inference in Linear Regression Analysis

The material covered so far focuses on using the data for a **sample** to graph and describe the relationship. The slope and intercept values we have computed are statistics, they are estimates of the underlying true relationship for the larger population.

Next we turn to making inferences about the relationship for the larger **population**. Here is a nice summary to help us distinguish between the **regression line for the sample** and the **regression line for the population**.

Regression Line for the Sample

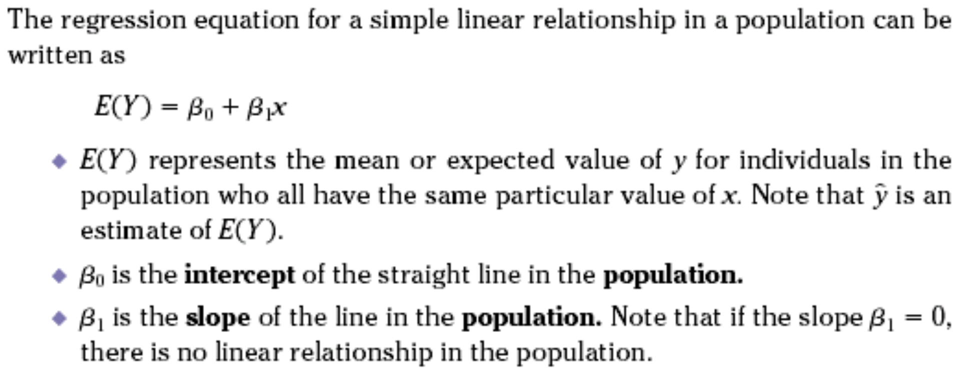


$$\hat{y} = b_0 + b_1 x$$

In any given situation, the sample is used to determine values for b_0 and b_1.

- $\hat{y}$ is spoken as "y-hat" and it is also referred to either as *predicted* y or *estimated* y.
- b_0 is the **intercept** of the straight line. The *intercept* is the value of $\hat{y}$ when $x = 0$.
- b_1 is the **slope** of the straight line. The *slope* tells us how much of an increase (or decrease) there is for $\hat{y}$ when the x variable increases by one unit. The sign of the slope tells us whether $\hat{y}$ increases or decreases when x increases. If the slope is 0, there is no linear relationship between x and y because $\hat{y}$ is the same for all values of x.

Regression Line for the Population



The regression equation for a simple linear relationship in a population can be written as

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

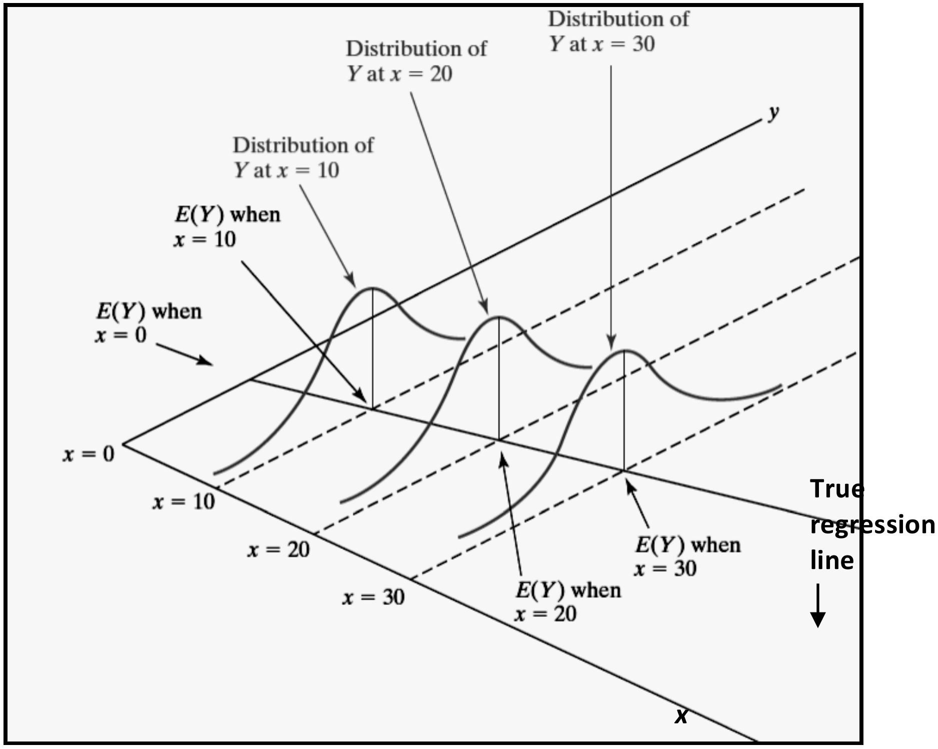
- $E(Y)$ represents the mean or expected value of y for individuals in the population who all have the same particular value of x. Note that $\hat{y}$ is an estimate of $E(Y)$.
- β_0 is the **intercept** of the straight line in the **population.**
- β_1 is the **slope** of the line in the **population.** Note that if the slope $\beta_1 = 0$, there is no linear relationship in the population.

To do formal inference, we think of our b_0 and b_1 as estimates of the unknown parameters β_0 and β_1 . Below we have the somewhat statistical way of expressing the underlying model that produces our data:

Linear Model: the response $y = [\beta_0 + \beta_1(x)] + \varepsilon$

$= [\text{Population relationship}] + \text{Randomness}$

This statistical model for simple linear regression assumes that for each value of *x* the observed values of the response (the population of *y* values) is **normally distributed**, varying around some true **mean (that may depend on *x* in a linear way)** and a **standard deviation σ that does not depend on *x***. This true mean is sometimes expressed as $E(Y) = \beta_0 + \beta_1(x)$. And the components and assumptions regarding this statistical model are show visually below.



The ε represents the *true error term.* These would be the deviations of a particular value of the response y from the *true regression line.* As these are the deviations from the mean, then these error terms should have a normal distribution with mean 0 and constant standard deviation σ.

Now, we cannot observe these ε's. However we will be able to use the estimated (observable) errors, namely the residuals, to come up with an estimate of the standard deviation and to check the conditions about the true errors.

So what have we done, and where are we going?

1. Estimate the regression line based on some data. **DONE!**
2. Measure the strength of the linear relationship with the correlation. **DONE!**
3. Use the estimated equation for predictions. **DONE!**
4. Assess if the linear relationship is statistically significant.
5. Provide interval estimates (confidence intervals) for our predictions.
6. Understand and check the assumptions of our model.

We have already discussed the descriptive goals of 1, 2, and 3. For the inferential goals of 4 and 5, we will need an estimate of the unknown standard deviation in regression σ.

Estimating the Standard Deviation for Regression

The standard deviation for regression can be thought of as measuring the **average size of the residuals**. A relatively small standard deviation from the regression line indicates that individual data points generally fall close to the line, so predictions based on the line will be close to the actual values.

It seems reasonable that our estimate of this average size of the residuals be based on the residuals using the sum of squared residuals and dividing by appropriate degrees of freedom. Our estimate of σ is given by:

$s =$

Note: Why $n - 2$?

Estimating the Standard Deviation: Bill vs Tip

Below are the portions of the R regression output that we could use to obtain the estimate of σ for our regression analysis.

From Summary:

```
Residual standard error: 2.251 on 4 degrees of freedom
Multiple R-squared:  0.8487,    Adjusted R-squared:  0.8109
F-statistic: 22.45 on 1 and 4 DF,  p-value: 0.009052
```

Or from ANOVA:

```
Response: Tip
          Df  Sum Sq Mean Sq F value   Pr(>F)
Bill       1 113.732 113.732  22.446 0.009052 **
Residuals  4  20.268   5.067
```

Significant Linear Relationship?

Consider the following hypotheses: $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$
What happens if the null hypothesis is true?

There are a number of ways to test this hypothesis. One way is through a t-test statistic (think about why it is a t and not a z test). The general form for a t test statistic is:

$$\text{t} = \frac{\text{sample statistic - null value}}{\text{standard error of the sample statistic}}$$

We have our sample estimate for β_1, it is b_1. And we have the null value of 0. So we need the standard error for b_1. We could "derive" it, using the idea of sampling distributions (think about the population of all possible b_1 values if we were to repeat this procedure over and over many times). Here is the result:

t-test for the population slope β_1

To test $H_0 : \beta_1 = 0$ we would use $t = \dfrac{b_1 - 0}{\text{s.e.}(b_1)}$

where $SE(b_1) = \dfrac{s}{\sqrt{\sum (x - \bar{x})^2}}$ and the degrees of freedom for the *t*-distribution are *n* – 2.

This t-statistic could be modified to test a variety of hypotheses about the population slope (different null values and various directions of extreme).

Try It!

Significant Relationship between Bill and Tip?

Is there a significant (non-zero) linear relationship between the total cost of a restaurant bill and the tip that is left? (is the bill a useful linear predictor for the tip?)
That is, test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ using a 5% level of significance.

Think about it:
Based on the results of the previous *t*-test conducted at the 5% significance level, do you think a 95% confidence interval for the true slope β_1 would contain the value of 0?

Confidence Interval for the population slope β_1

$b_1 \pm t^*\left[SE(b_1)\right]$ where df = *n* – 2 for the t^* value

Compute the interval and check your answer.

Could you interpret the 95% confidence level here?

Inference about the Population Slope using R

Below are the portions of the R regression output that we could use to perform the *t*-test and obtain the confidence interval for the population slope β_1.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.58769    2.41633  -0.243  0.81980
Bill         0.17077    0.03604   4.738  0.00905 **
```

Note: There is a third way to test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$.
It involves another *F*-test from an ANOVA for regression.

```
Response: Tip
          Df  Sum Sq Mean Sq F value   Pr(>F)
Bill       1 113.732 113.732  22.446 0.009052 **
Residuals  4  20.268   5.067
```

Predicting for Individuals versus Estimating the Mean

Consider the relationship between the bill and tip …

Least squares regression line (or estimated regression function):

$$\hat{y} =$$

We also have: $s =$

How would you predict the tip ***for Barb*** who had a $50 restaurant bill?

How would you estimate the mean tip ***for all customers*** who had a $50 restaurant bill?

So our estimate for **predicting a future observation** and for **estimating the mean response** are found using the same least squares regression equation. What about their standard errors? (We would need the standard errors to be able to produce an interval estimate.)

Idea: Consider a population of individuals and a population of means:

What is the standard deviation for a population of individuals?

What is the standard deviation for a population of means?
Which standard deviation is larger?

So a ***prediction interval for an individual response*** will be

(wider or narrower) than a ***confidence interval for a mean response***.

Here are the (somewhat messy) formulas:

Confidence interval for a mean response:

$$\hat{y} \pm t^{*}\text{s.e.(fit)}$$

where $$\text{s.e.(fit)} = s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i - \bar{x})^2}}$$ **df = *n* – 2**

Prediction interval for an individual response:

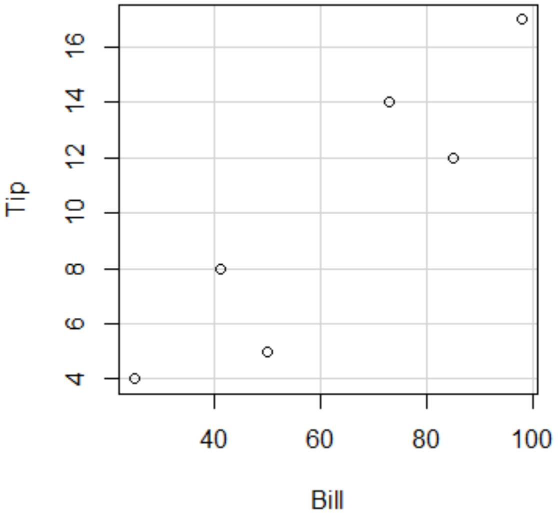
$$\hat{y} \pm t^{*}\text{s.e.(pred)}$$

where $$\text{s.e.(pred)} = \sqrt{s^2 + (\text{s.e.(fit)})^2}$$ **df = *n* – 2**

Try It! Bill vs Tip

Construct a 95% confidence interval for the mean tip given for all customers who had a $50 bill (x). Recall: n = 6, $\bar{x} = 62$, $\sum(x-\bar{x})^2 = S_{XX}$ = 3900, $\hat{y}$ = -0.58 +0.17(x), and s = 2.251.

Construct a 95% prediction interval for the tip from an individual customer who had a $50 bill (x).



Checking Assumptions in Regression

Let’s recall the statistical way of expressing the underlying model that produces our data:

Linear Model: the response $y = [\beta_0 + \beta_1(x)] + \varepsilon$
$= $ [Population relationship] + Randomness

where the ε‘s, the *true error terms* should be normally distributed
with mean 0 and constant standard deviation σ,
and this randomness is independent from one case to another.

Thus there are **four essential technical assumptions** required for inference in linear regression:

(1) Relationship is in fact linear.
(2) Errors should be normally distributed.
(3) Errors should have constant variance.
(4) Errors should not display obvious ‘patterns’.

Now, we cannot observe these ε ’s. However we will be able to use the estimated (observable) errors, namely the residuals, to come up with an estimate of the standard deviation and to check the conditions about the true errors.

So how can we check these assumptions with our data and estimated model?

(1) Relationship is in fact linear. →

(2) Errors should be normally distributed. →

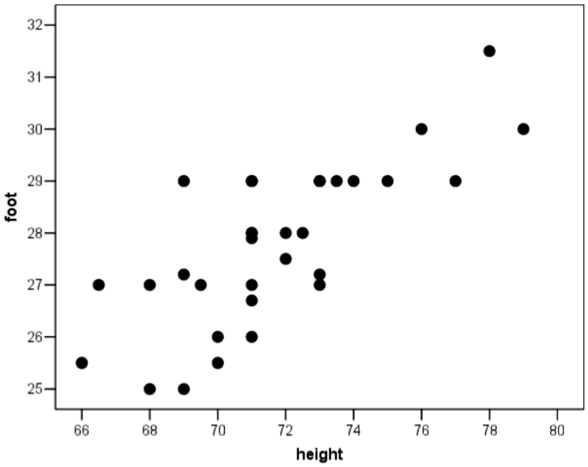
$\left.\begin{array}{l}\text{(3) Errors should have constant variance.}\\ \text{(4) Errors should not display obvious ‘patterns’.}\end{array}\right\}$ If we see …

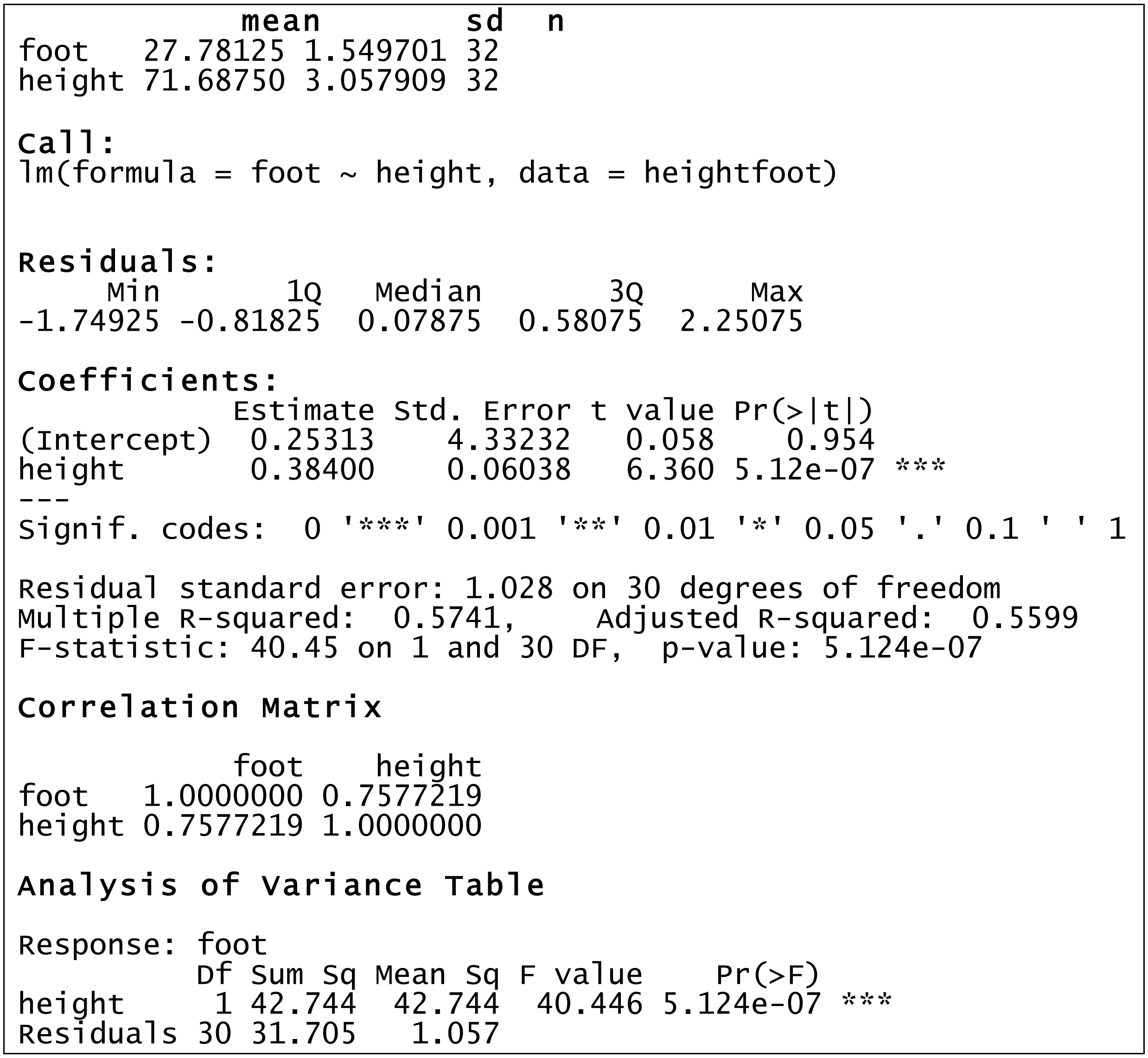
Now, if we saw …

Let's turn to one last full regression problem that includes checking assumptions.

Relationship between height and foot length for College Men

The heights (in inches) and foot lengths (in centimeters) of 32 college men were used to develop a model for the relationship between height and foot length. The scatterplot and R regression output are provided.





```
               mean        sd  n
foot   27.78125 1.549701 32
height 71.68750 3.057909 32

Call:
lm(formula = foot ~ height, data = heightfoot)


Residuals:
     Min       1Q   Median       3Q      Max
-1.74925 -0.81825  0.07875  0.58075  2.25075

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.25313    4.33232   0.058    0.954
height       0.38400    0.06038   6.360 5.12e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.028 on 30 degrees of freedom
Multiple R-squared:  0.5741,    Adjusted R-squared:  0.5599
F-statistic: 40.45 on 1 and 30 DF,  p-value: 5.124e-07

Correlation Matrix

            foot    height
foot   1.0000000 0.7577219
height 0.7577219 1.0000000

Analysis of Variance Table

Response: foot
          Df Sum Sq Mean Sq F value    Pr(>F)
height     1 42.744  42.744  40.446 5.124e-07 ***
Residuals 30 31.705   1.057
```

Also note that: $S_{XX} = \sum \left(x - \bar{x}\right)^2 = 289.87$

a. How much would you expect foot length to increase for each 1-inch increase in height? Include the units.

b. What is the correlation between height and foot length?

c. Give the equation of the least squares regression line for predicting foot length from height.

d. Suppose Max is 70 inches tall and has a foot length of 28.5 centimeters. Based on the least squares regression line, what is the value of the prediction error (residual) for Max? Show all work.

e. Use a 1% significance level to assess if there is a significant positive linear relationship between height and foot length. State the hypotheses to be tested, the observed value of the test statistic, the corresponding *p*-value, and your decision.

Hypotheses: H_0:______________________H_a:__________________________

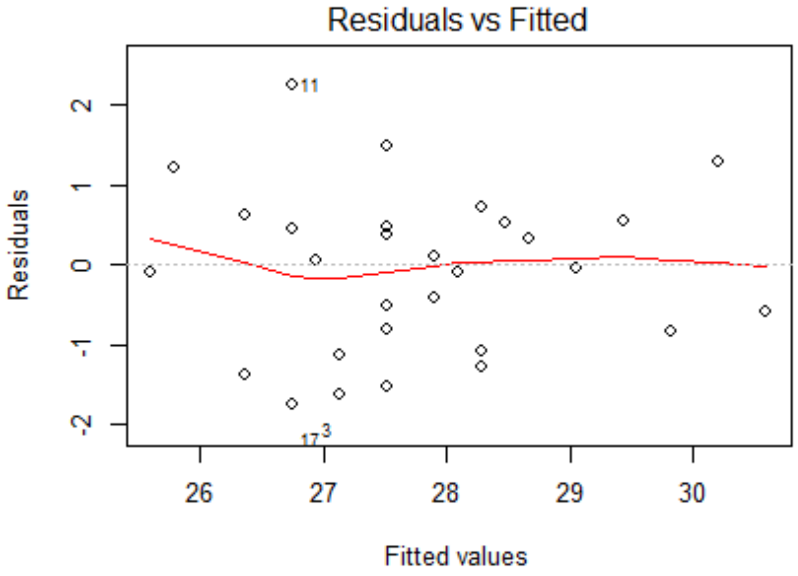
Test Statistic Value: ________________ *p*-value: ____________________

Decision: (circle) **Fail to reject H_0** **Reject H_0**

Conclusion:

f. Calculate a 95% confidence interval for the average foot length for all college men who are 70 inches tall. (Just clearly plug in all numerical values.)

g. Consider the residuals vs fitted plot shown.



Does this plot support the conclusion that the linear regression model is appropriate?

Yes ***No***

Explain:

Regression

Linear Regression Model

Population Version:

Mean: $\mu_Y(x) = E(Y) = \beta_0 + \beta_1 x$

Individual: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

where ε_i is $N(0, \sigma)$

Sample Version:

Mean: $\hat{y} = b_0 + b_1 x$

Individual: $y_i = b_0 + b_1 x_i + e_i$

Standard Error of the Sample Slope

$$\text{s.e.}(b_1) = \frac{s}{\sqrt{S_{XX}}} = \frac{s}{\sqrt{\sum (x - \bar{x})^2}}$$

Confidence Interval for β_1

$b_1 \pm t^* \text{s.e.}(b_1)$ df = *n* – 2

t-Test for β_1

To test $H_0 : \beta_1 = 0$ $t = \dfrac{b_1 - 0}{\text{s.e.}(b_1)}$

df = *n* – 2

or $F = \dfrac{MSREG}{MSE}$ df = 1, *n* – 2

Parameter Estimators

$$b_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum (x - \bar{x}) y}{\sum (x - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Confidence Interval for the Mean Response

$\hat{y} \pm t^* \text{s.e.}(\text{fit})$ df = *n* – 2

where $\text{s.e.}(\text{fit}) = s\sqrt{\dfrac{1}{n} + \dfrac{(x - \bar{x})^2}{S_{XX}}}$

Residuals

$e = y - \hat{y}$ = observed *y* – predicted *y*

Prediction Interval for an Individual Response

$\hat{y} \pm t^* \text{s.e.}(\text{pred})$ df = *n* – 2

where $\text{s.e.}(\text{pred}) = \sqrt{s^2 + (\text{s.e.}(\text{fit}))^2}$

Correlation and its square

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

$$r^2 = \frac{SSTO - SSE}{SSTO} = \frac{SSREG}{SSTO}$$

where $SSTO = S_{YY} = \sum (y - \bar{y})^2$

Estimate of σ

$s = \sqrt{MSE} = \sqrt{\dfrac{SSE}{n - 2}}$ where

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

Standard Error of the Sample Intercept

$$\text{s.e.}(b_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}$$

Confidence Interval for β_0

$b_0 \pm t^* \text{s.e.}(b_0)$ df = *n* – 2

t-Test for β_0

To test $H_0 : \beta_0 = 0$

$t = \dfrac{b_0 - 0}{\text{s.e.}(b_0)}$ df = *n* – 2

Additional Notes

A place to … jot down questions you may have and ask during office hours, take a few extra notes, write out an extra problem or summary completed in lecture, create your own summary about these concepts.