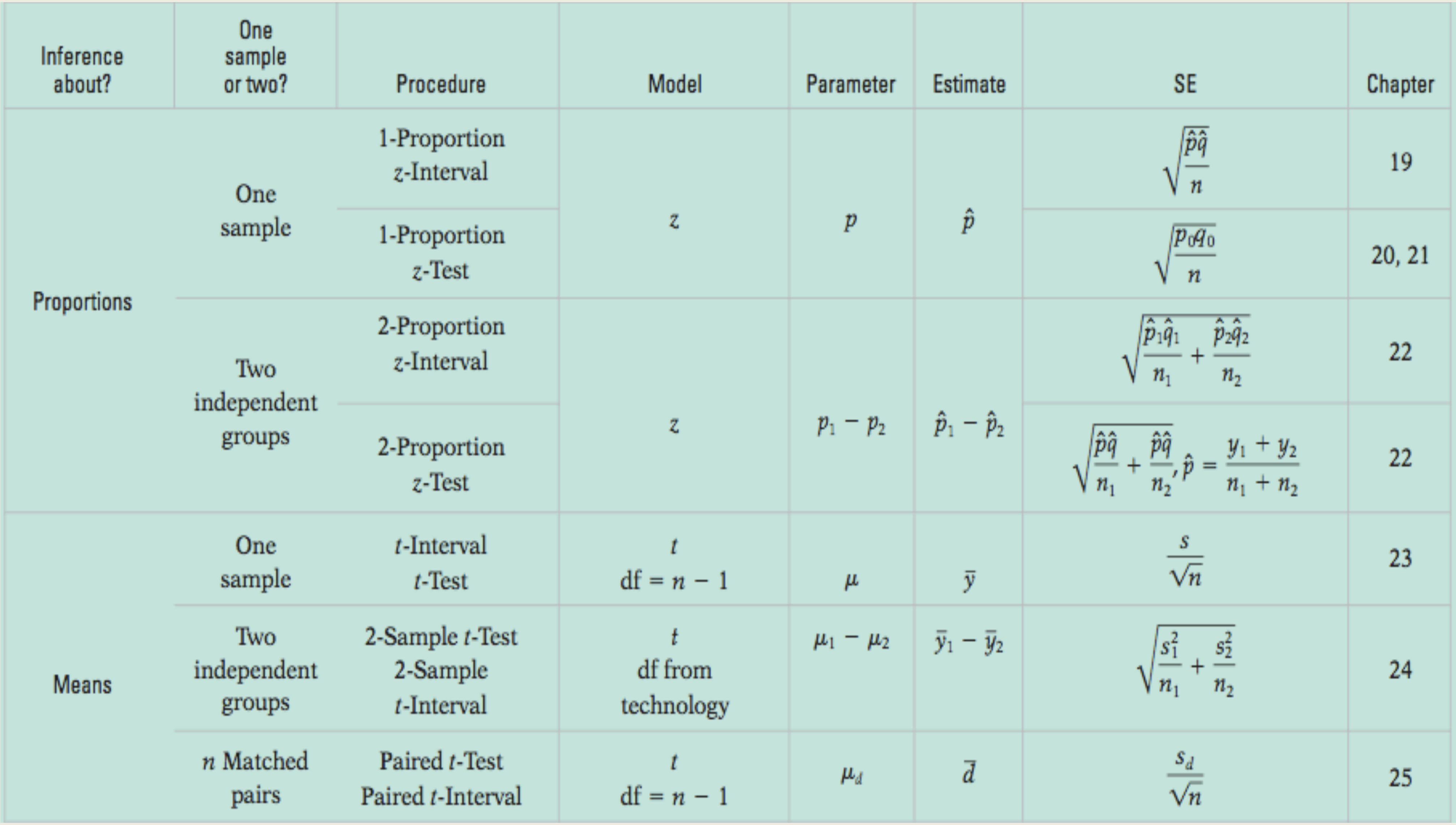
Quantitative Methods

Serena DeStefani – Lecture 19 –8/5/2020

Announcements: Final Exam

- Non-cumulative, 2&1/2 hours long, as the midterm
- Same format as HWs, focus on inference (no CH 15/16)
- Problems (can be both tests and CIs) on inference for means, proportions, regression (I will give you the table), one chi-square test, and questions about Analysis of Variance (I will give you the ANOVA table)

Review



Inference about?	One sample or two?	Procedure	Model	Parameter	Estimate	SE	Chapter
Proportions	One sample	1-Proportion z-Interval	z	p	$\hat{p}$	$\sqrt{\frac{\hat{p}\hat{q}}{n}}$	19
		1-Proportion z-Test				$\sqrt{\frac{p_0q_0}{n}}$	20, 21
	Two independent groups	2-Proportion z-Interval	z	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$	22
		2-Proportion z-Test				$\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}, \hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$	22
Means	One sample	t-Interval t-Test	t df $= n - 1$	μ	$\bar{y}$	$\frac{s}{\sqrt{n}}$	23
	Two independent groups	2-Sample t-Test 2-Sample t-Interval	t df from technology	$\mu_1 - \mu_2$	$\bar{y}_1 - \bar{y}_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	24
	n Matched pairs	Paired t-Test Paired t-Interval	t df $= n - 1$	μ_d	$\bar{d}$	$\frac{s_d}{\sqrt{n}}$	25

Would being part of a support group that meets regularly help people who are wearing the nicotine patch actually quit smoking? A county health department tries an experiment using several hundred volunteers who were planning to use the patch to help them quit smoking. The subjects were randomly divided into two groups. People in Group 1 were given the patch and attended a weekly discussion meeting with counselors and others trying to quit. People in Group 2 also used the patch but did not participate in the counseling groups. After six months 46 of the 143 smokers in Group 1 and 30 of 151 smokers in Group 2 had successfully stopped smoking.

Do these results suggest that such support groups could be an effective way to help people stop smoking?

Inference about?
One sample or two?
Procedure?
Model?
Parameter?
Estimate?
SE?

Would being part of a support group that meets regularly help people who are wearing the nicotine patch actually quit smoking? A county health department tries an experiment using several hundred volunteers who were planning to use the patch to help them quit smoking. The subjects were randomly divided into two groups. People in Group 1 were given the patch and attended a weekly discussion meeting with counselors and others trying to quit. People in Group 2 also used the patch but did not participate in the counseling groups. After six months 46 of the 143 smokers in Group 1 and 30 of 151 smokers in Group 2 had successfully stopped smoking.

Do these results suggest that such support groups could be an effective way to help people stop smoking?

Now that we’ve concluded the support program is beneficial, can we convince the government to fund it? That might depend on *how* effective it is.

Inference about?
One sample or two?
Procedure?
Model?
Parameter?
Estimate?
SE?

Chapter 25

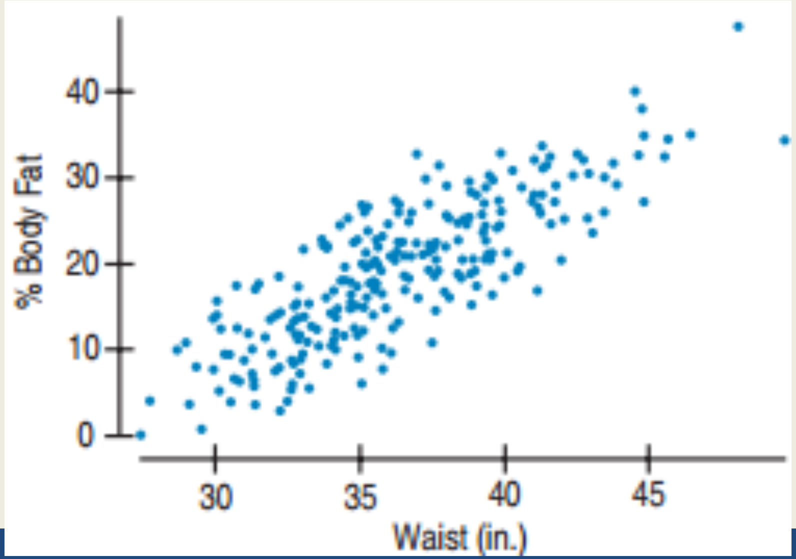
Inferences on Regression

25.1

The Population and the Sample

Waist size and %Body Fat

- Can we **predict** the amount of %Body-Fat from Waist size?
- We can take a sample of men, measure %Body-Fat and Waist size for each, and run a regression
- For each value of Waist size I will have maybe one value, maybe one or two values of %Body-Fat

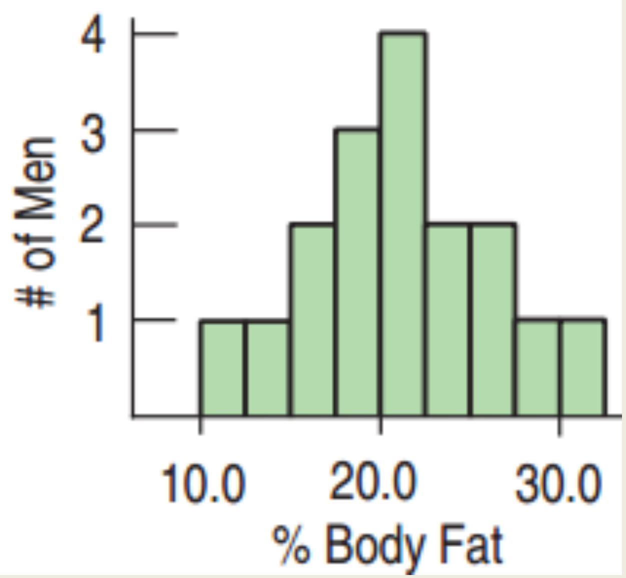


Waist size and %Body Fat

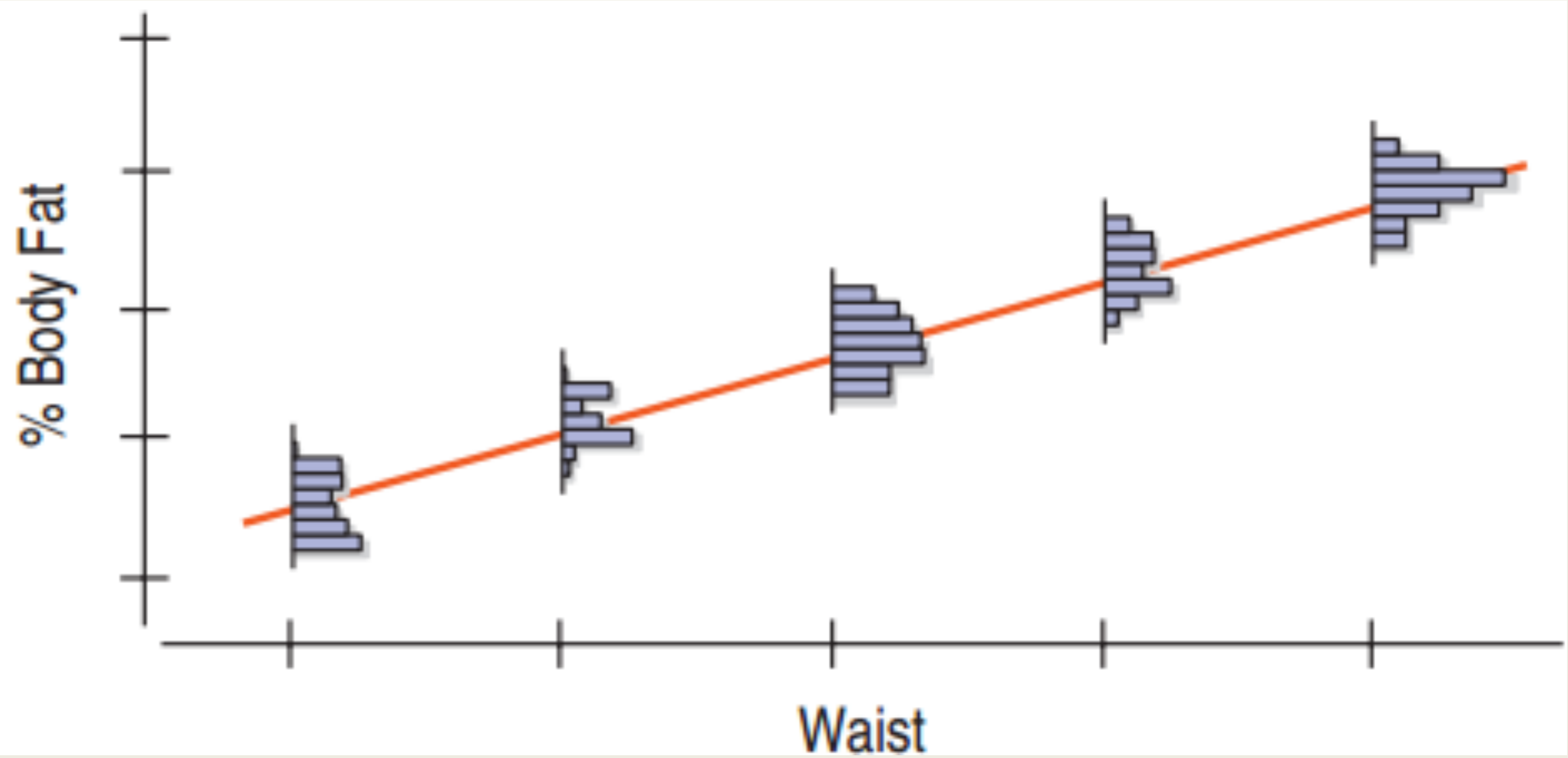
- For each value of Waist size I will have maybe one value, maybe one or two values of %Body-Fat
- But what happens if I think about the **whole population** of men?
- For each value of Waist Size I will have many different values of %Body-Fat !
- How they will be distributed?

Waist size and %Body Fat

- For each value of Waist Size I will have many different values of %Body-Fat !
- How they will be distributed?
- For **one** value of waist size:
- And for many values?

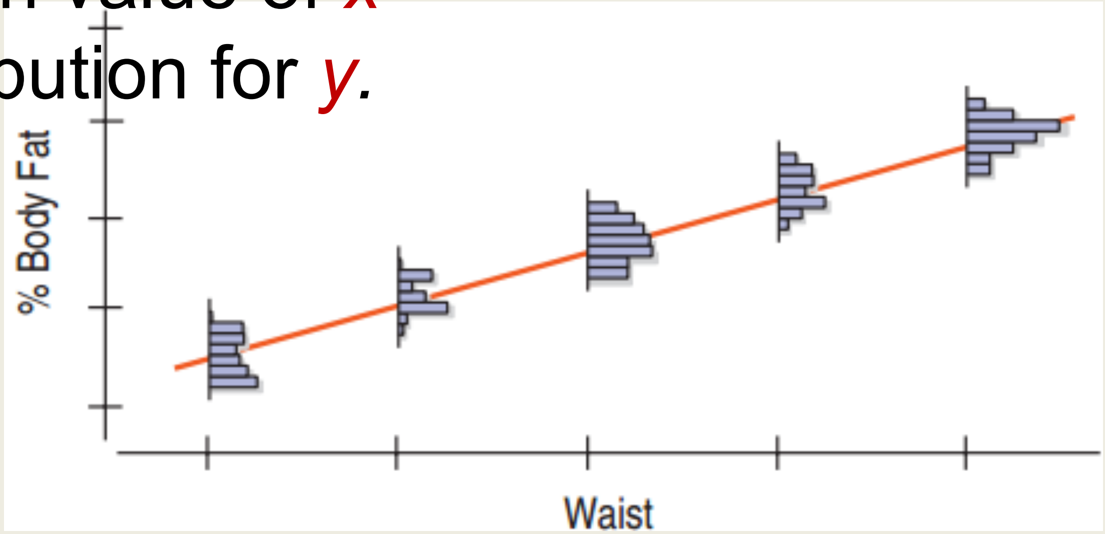


Many Distributions for Regression



Many Distributions for Regression

- We saw how to build a single sampling distribution for means and proportions.
- How do we translate this concept to regression?
- If we think about the population, for each value of waist there there are different values of %Body-Fat
- With regression, for each value of *x* there is a different distribution for *y*.
- This model assumes that for each *x*, the mean of the *y*'s is on the regression line.

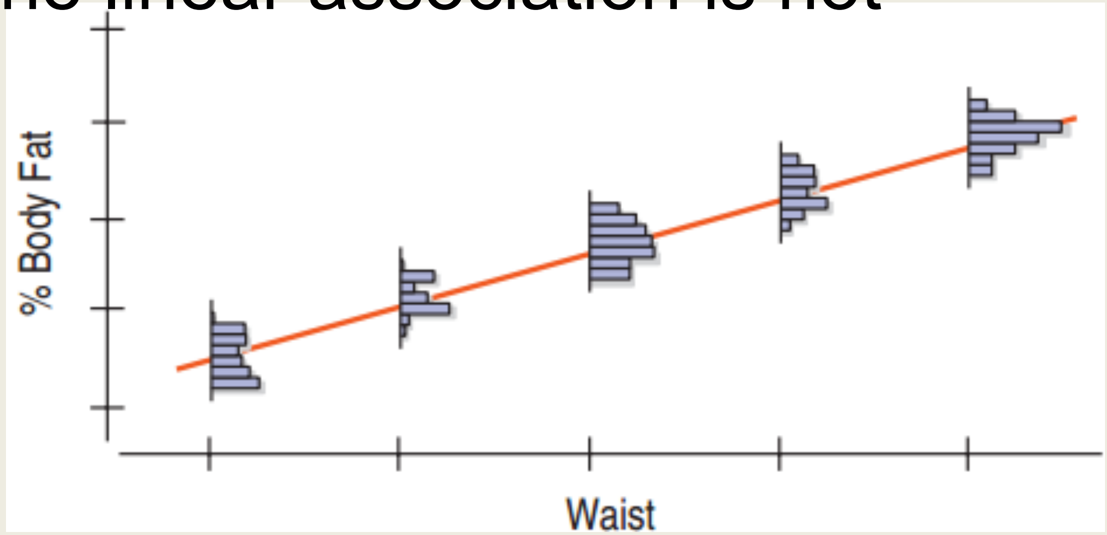


Many Distributions for Regression

- This **model** *assumes* that for each *x*, the mean of the *y*'s is on the regression line.
- Being a model, it's an abstraction. It assumes a perfect linear association between the variables. But…

1. We cannot test the whole population.
2. Of course, in reality, the linear association is not perfect.

- We will use the line from one sample as an **estimate**.



Sample vs. Model Regression

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

- This gives a prediction for *y* based on the sample.

Model: $\mu_y = \beta_0 + \beta_1 x$

- β_0 = *y*-intercept for the model
- β_1 = slope for the model
- The model assumes that for every value of *x*, the mean of all the *y*'s lies on the line.

Error

$$\mu_y = \beta_0 + \beta_1 x$$

- The model predicts the mean of y for each x, but misses the actual (observed) individual values of y.
- The error, ε, is the amount the line misses the value of y.
- ε is analogous to e, the residual: $e = y - \hat{y}$
- We want to define a new equation that incorporates the error:

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

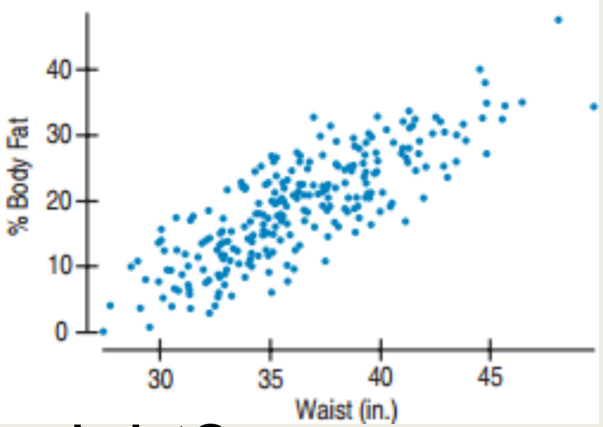
- This new equation gives the exact value of each of the observed y's.

How Good is the Model?

- The least squares regression line $\hat{y} = b_0 + b_1 x$ obtained from the sample gives estimates for the model.
  - b_0 is an estimate for β_0.
  - b_1 is an estimate for β_1.

- Challenge: How good are these estimates?

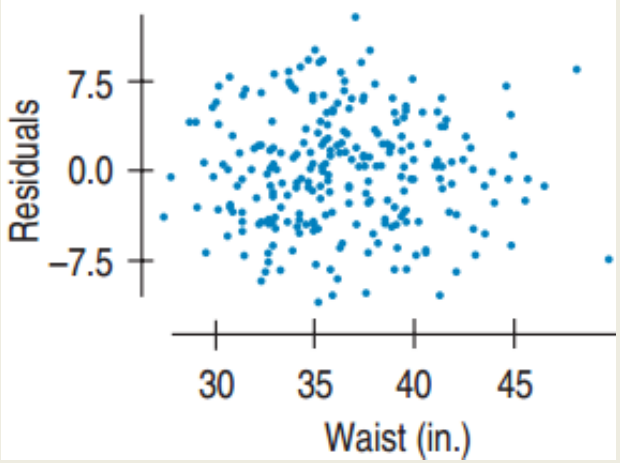
1. Linearity Assumption



Straight Enough Condition

- Does the scatterplot look relatively straight?
- Don’t draw the line. It can fool you.
- Look at scatterplot of the residuals.
  - Should have horizontal direction
  - Should not have a pattern
- If straight enough, check the other assumptions.
- If not straight, stop or re-express.

2. Independence Assumption



Errors (ε's) must be independent of each other.

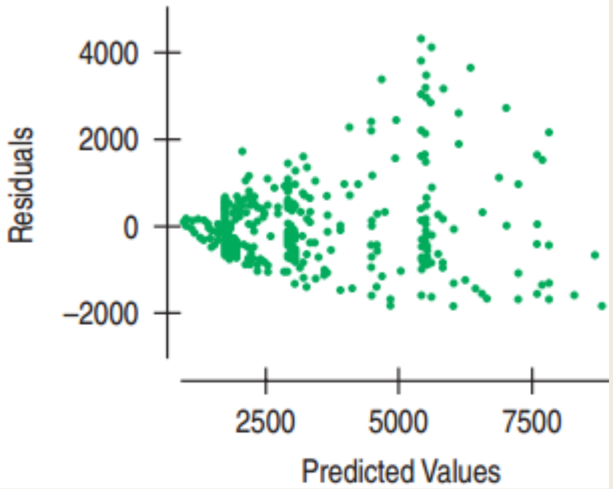
- Check residuals plot.
  - Shouldn't have clumps, trends, or a pattern.
  - Should look very random.
- For x = time, plot residuals vs. residuals one step later.
  - Should look very random.

- To make inferences about the population, the sample must be representative.

3. Equal Variance Assumption

Variability of *y* same for all *x*.

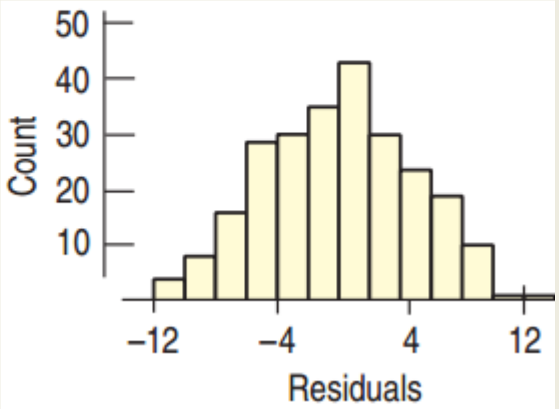
- Does the Plot Thicken? Condition: Spread along the line should be nearly constant.
- "Fan Shape" is bad →
- Standard Deviation of Residuals, s_e, will be used for CI and Hypothesis Tests. This requires same variance for each *x*.

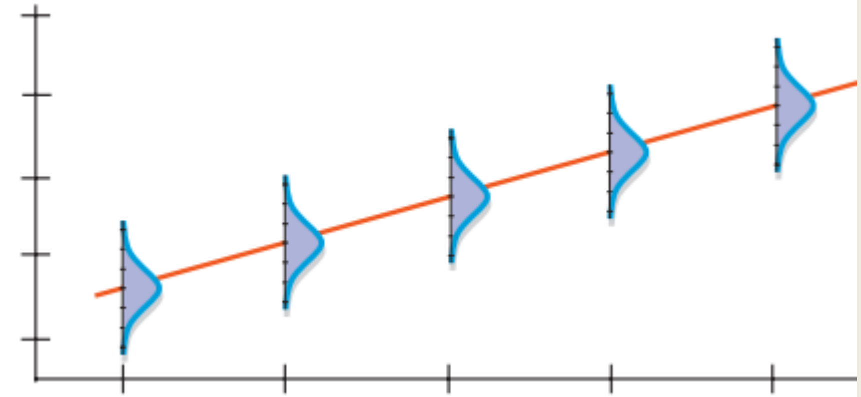


4. Normal Population Assumption

The errors for each fixed *x* must follow the Normal model.

- Good enough to use the Nearly Normal Condition – but check for Outliers. Look at the histogram.
- With large sample sizes, the Nearly Normal Condition is usually satisfied.

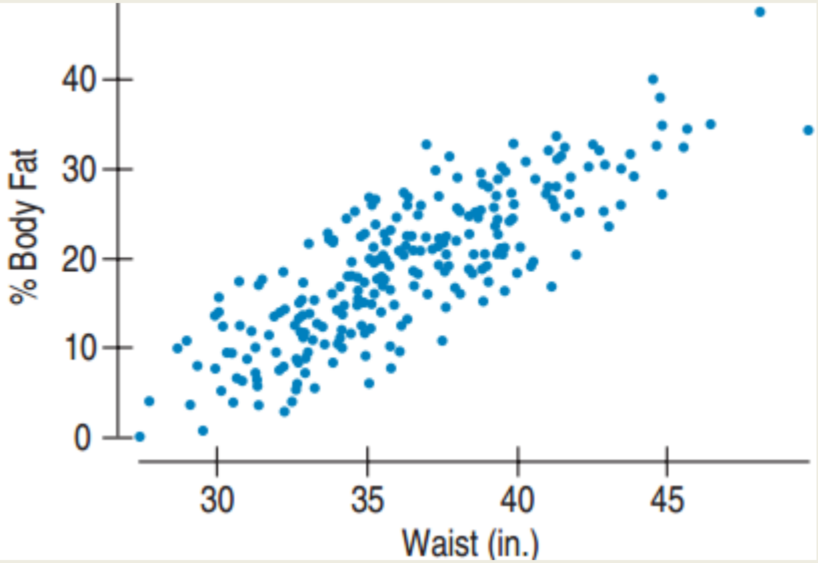




Body Fat and Waist Size

What is the relationship between body fat and waist size in men?

- **Plan:** I have measurements on 250 adult males.
- **Model:**
  - ✓ **Straight Enough Condition:** The scatterplot looks very straight.

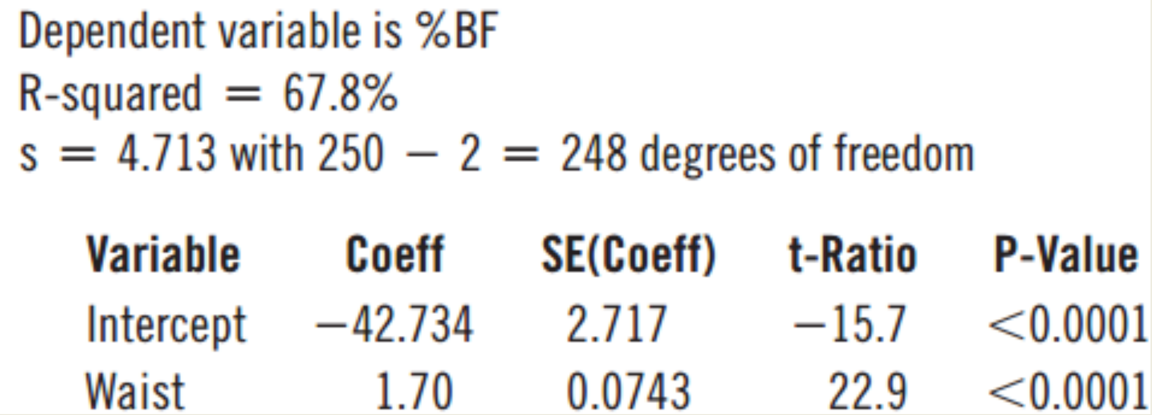


Body Fat and Waist Size

- **Model:**
  - ✓ **Independence Assumption:** No reason to think that one man’s fat influences another’s.
  - ✓ **Does the Plot Thicken? Condition:** Neither the original scatterplot nor the residuals show changing variances.

Body Fat and Waist Size

- **Mechanics:** Computer Output:



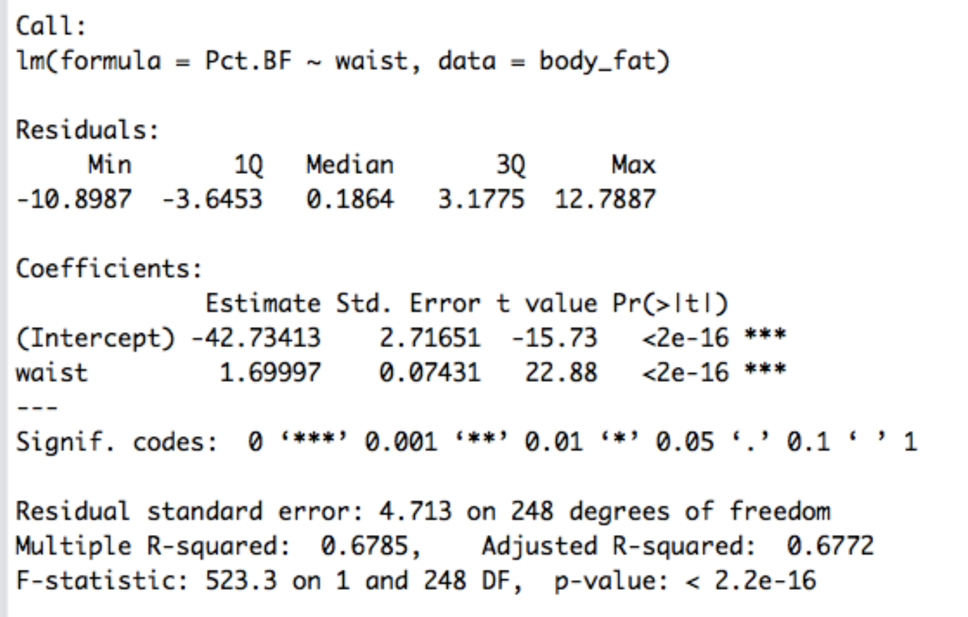
```
Dependent variable is %BF
R-squared = 67.8%
s = 4.713 with 250 − 2 = 248 degrees of freedom

    Variable     Coeff     SE(Coeff)   t-Ratio   P-Value
    Intercept   −42.734    2.717       −15.7     <0.0001
    Waist         1.70     0.0743       22.9     <0.0001
```

- The estimated regression equation is:

$$\widehat{\%Body\ Fat} = -42.73 + 1.70\ Waist$$

Body Fat and Waist Size: R code

```
>body_fat <- read.csv("Body_fat_complete.csv")
>reg_bf <- lm(Pct.BF~ waist,body_fat)
>summary(reg_bf)
```



```
Call:
lm(formula = Pct.BF ~ waist, data = body_fat)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.8987  -3.6453   0.1864   3.1775  12.7887 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -42.73413    2.71651  -15.73   <2e-16 ***
waist         1.69997    0.07431   22.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.713 on 248 degrees of freedom
Multiple R-squared:  0.6785,	Adjusted R-squared:  0.6772 
F-statistic: 523.3 on 1 and 248 DF,  p-value: < 2.2e-16
```

- The estimated regression equation is:

$$\widehat{\%Body\ Fat} = -42.73 + 1.70\ Waist$$

Body Fat and Waist Size

- **Conclusion:**
  - R^2 = 67.8%. Waist size accounts for about 2/3 of the variation in *%Body Fat*.
  - Slope = 1.7. *%Body Fat* increases about 1.7% for each inch increase in body fat, on average.
  - Standard Error for slope = 0.07, smaller than the slope. The estimate is reasonably precise.

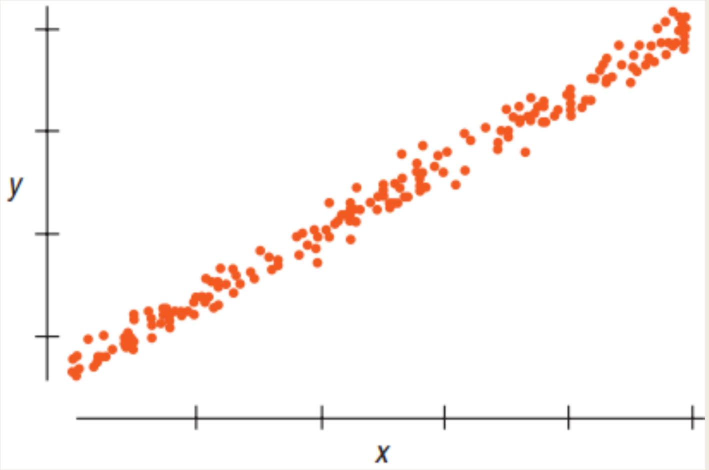
25.3

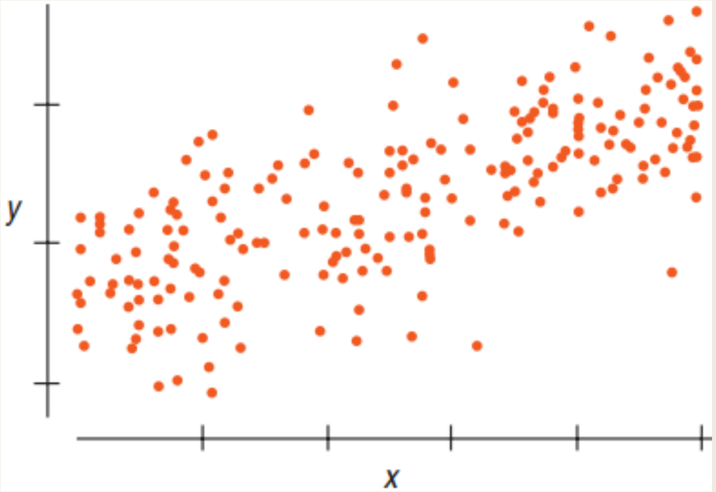
Intuition About Regression Inference

Sample-to-Sample Variation of Slope and Intercept

- Each sample will produce it’s own slope (b_1) and intercept (b_0).
- The expected value of b_1 should be β_1.
- What is the standard deviation of all possible b_1’s?
- What factors influence this standard deviation?
- What factors influence the standard deviation of b_0?

Spread Around the Line

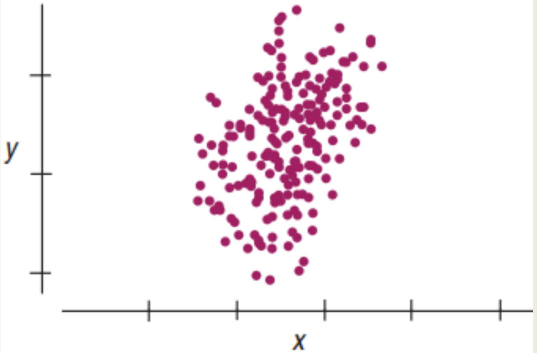


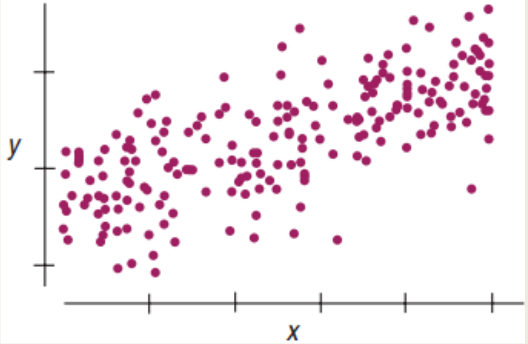


- Less scatter along line → Slope more consistent
- Residual Standard Deviation s_e measures this scatter.

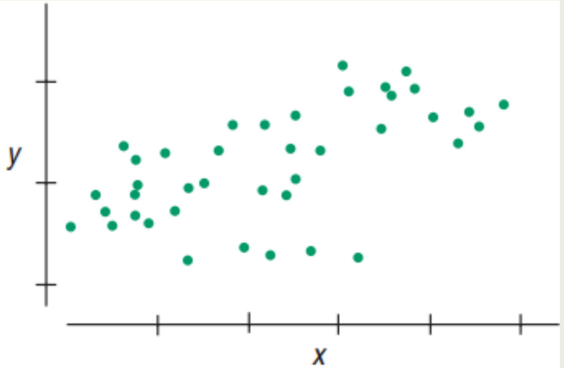
$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$$

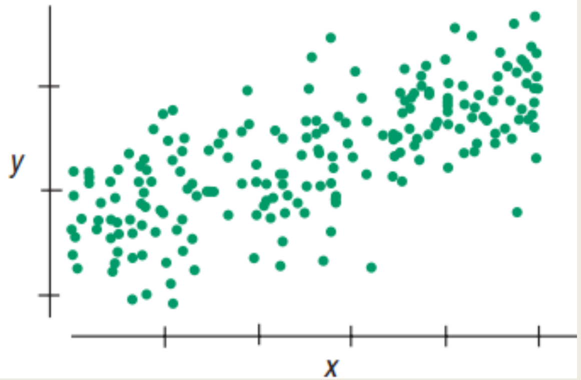
Spread of the *x*’s, Sample Size





- Larger s_x (SD in *x*) → More stable regression





- Larger Sample Size → More stable regression

Standard Error for the Slope

$$SE(b_1) = \frac{s_e}{\sqrt{n-1}\, s_x}$$

- From the formula, $SE(b_1)$ increases with s_e (SD of residuals) and decreases with s_x (SD of x)

- If we subtract β_1 from b_1 and divide by $SE(b_1)$, the result is a Student's *t*-model with df = $n - 2$.

$$\frac{b_1 - \beta_1}{SE(b_1)} \sim t_{n-2}$$

Sampling Distribution for Regression Slopes

- When the conditions are met, $t = \dfrac{b_1 - \beta_1}{SE(b_1)}$

  follows Student's t-model with df = n – 2.

- Estimate of the standard error:

$$SE(b_1) = \frac{s_e}{\sqrt{n-1}\, s_x}, \quad s_e\sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$$

What About the Intercept?

$$\frac{b_0 - \beta_0}{SE\left(b_0\right)} \sim t_{n-2}$$

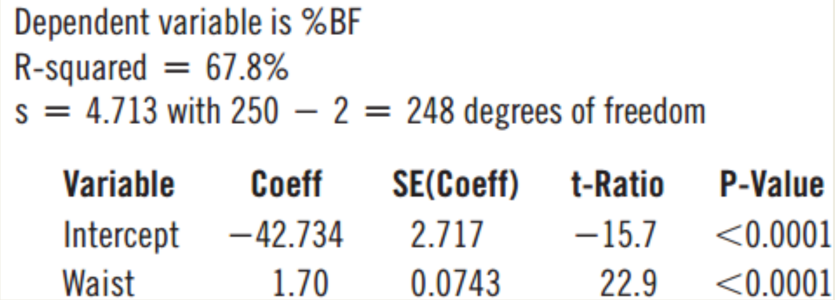
- The intercept is rarely of interest.
- Hypotheses and CI are usually about the slope only.

25.4

Regression Inference

Testing for β_1

- If no linear association, $\beta_1 = 0$
- H_0: $\beta_1 = 0$
- $$t_{n-2} = \frac{b_1 - 0}{SE(b_1)}$$



Dependent variable is %BF
R-squared = 67.8%
s = 4.713 with 250 − 2 = 248 degrees of freedom

Variable	Coeff	SE(Coeff)	t-Ratio	P-Value
Intercept	−42.734	2.717	−15.7	<0.0001
Waist	1.70	0.0743	22.9	<0.0001

- For the *%Body Fat* and *Waist* data:

$$\frac{1.7 - 0}{0.0743} \approx 22.9$$

$$\text{P-value} < 0.0001$$

- Very unlikely to have such a high b_1 if $\beta_1 = 0$
- IF no association in population, sample is unlikely

Confidence Interval for β_1

- The hypothesis test for *%Body Fat* and *Weight* told us what we already know.

- A confidence interval is needed.

$$b_1 \pm t^{*}_{n-2} \times SE(b_1)$$

- For %Body Fat and Weight:

  1.7 ± 1.97 × 0.074 = (1.55%, 1.85%)

- With 95% confidence the slope of the line for *%Body Fat* and *Weight* is between 1.55% and 1.85%.

25.5

Standard Errors for Predicted Values

Two Prediction Questions

- Let’s say we have a new value: 38 in for waist size
- There are two questions we can consider

1. Predict the *%Body Fat* of a man with waist 38 in.
   - Notice the prediction is for a single man.

2. Predict the mean *%Body Fat* for all men with waist 38 in.
   - This prediction is for a mean.

- Both have the same prediction: $\hat{y}_\nu = b_0 + b_1 x_\nu$
- The confidence intervals: $\hat{y}_\nu \pm t^*_{n-2} \times SE$
- The standard errors (*SE*) will differ.

Standard Error for the Mean

$$SE(\hat{\mu}_\nu) = \sqrt{SE^2(b_1) \times (x_\nu - \bar{x})^2 + \frac{s_e^2}{n}}$$

- $SE(\hat{\mu}_\nu)$ increases as $SE(b_1)$ increases.
- $SE(\hat{\mu}_\nu)$ increases as x_ν strays from the mean of the x's.
- $SE(\hat{\mu}_\nu)$ increases as s_e increases.
- $SE(\hat{\mu}_\nu)$ decreases as n increases.

Standard Error for a Single Prediction

$$SE(\hat{y}_{\nu}) = \sqrt{SE^2(b_1) \times \left(x_{\nu} - \bar{x}\right)^2 + \frac{s_e^2}{n} + s_e^2}$$

- Individual values vary more than means.

- $SE(\hat{y}_{\nu})$ has the extra positive term s_e^2.

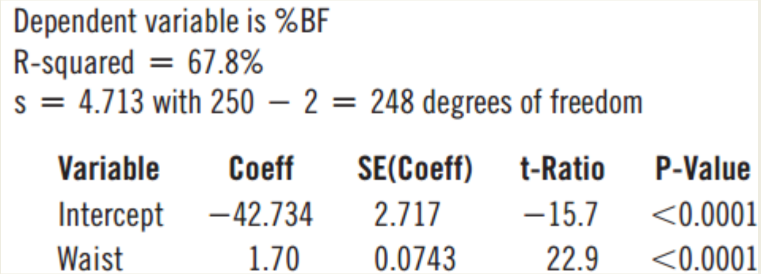
- When looking at a computer output, remember the smaller SE is for the predicted *mean* value and the larger is for the predicted *individual* value.

Confidence Interval for the Mean

Find the 95% CI for the **mean *%Body Fat* of all men** who have 38 inch waists.

- $s_e = 4.713,\ n = 250,\ SE(b_1) = 0.074,\ \bar{x} = 36.3,\ x_\nu = 38$

$\hat{y}_\nu = b_0 + b_1 x_\nu$ $SE(\hat{\mu}_\nu) = \sqrt{SE^2(b_1) \times (x_\nu - \bar{x})^2 + \frac{s_e^2}{n}}$



Dependent variable is %BF
R-squared = 67.8%
s = 4.713 with 250 − 2 = 248 degrees of freedom

Variable	Coeff	SE(Coeff)	t-Ratio	P-Value
Intercept	−42.734	2.717	−15.7	<0.0001
Waist	1.70	0.0743	22.9	<0.0001

$$\hat{y}_\nu = -42.7 + 1.7(38) = 21.9\%$$

$$SE(\hat{\mu}_\mu) = \sqrt{0.074^2 \times (38 - 36.3)^2 + \frac{4.713^2}{250}} = 0.32\%$$

Confidence Interval for the Mean

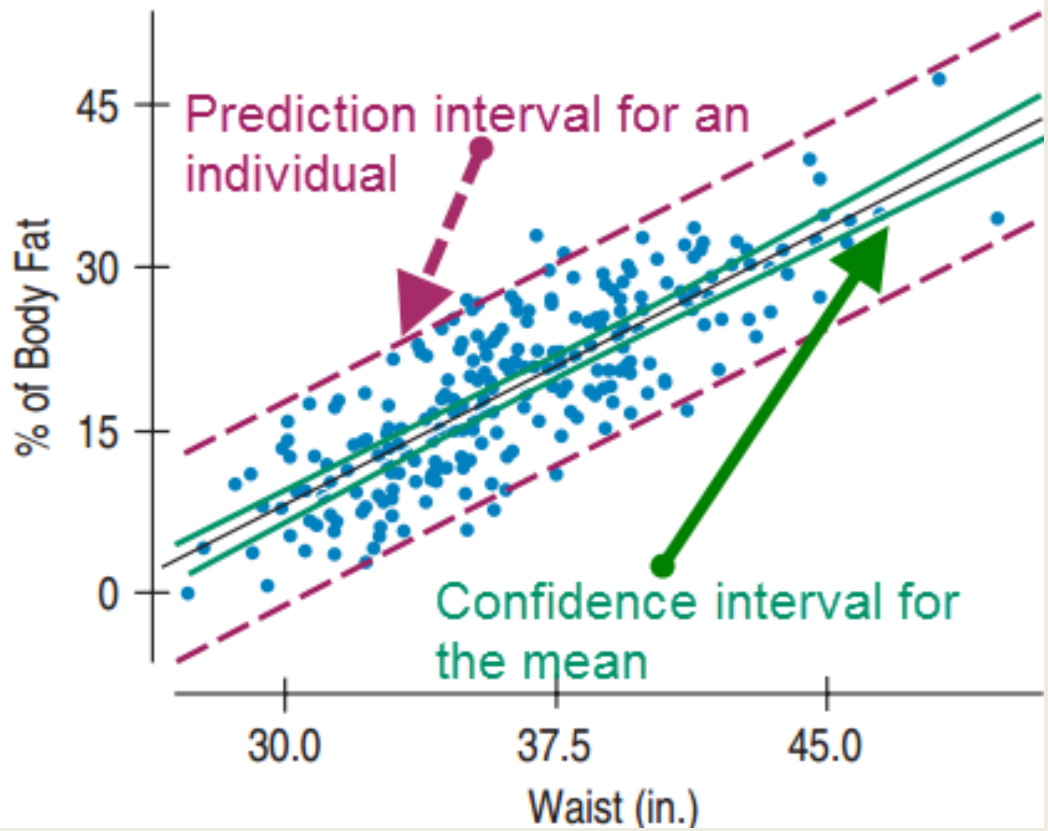
- $t^*_{248} = 1.97$
- $ME = 1.97 \times 0.32 = 0.63\%$
- CI: 21.9% ± 0.63%
- I am 95% confident that the mean *%Body Fat* for men with a 38 inch waist is 21.9% ± 0.63%.

Prediction Interval for an Individual

Find a prediction interval for the **%Body Fat for an individual man** with a 38-inch Waist.

- $SE(\hat{y}_{\mu}) = \sqrt{0.074^2 \times (38 - 36.3)^2 + \frac{4.713^2}{250} + 4.713^2} = 4.72\%$
- ME = 1.97 × 4.72 = 9.30%
- The prediction interval is: 21.9% ± 9.30%.
- There is 95% chance that this interval captures the true *%Body Fat* of a randomly selected man with a 38-inch waist.

Visualizing the Two Intervals



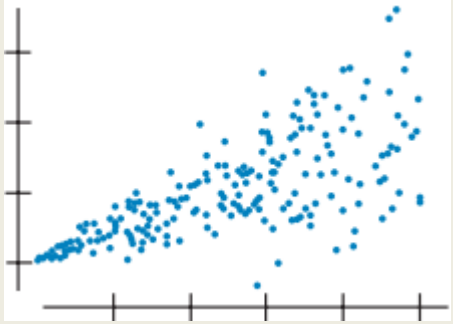
- The prediction interval for the individual is much wider than the confidence interval for the mean.

What Can Go Wrong?

Don’t fit a linear regression to data that aren’t straight.

- Stop here or try re-expressing if not straight.

Watch out for the plot thickening.



- If the points fan out, then the standard deviations are not constant. Don’t perform regression analysis.

Make sure the errors are Normal.

- Check the histogram and normal probability plot. Need Normal to invoke the CLT.

What Can Go Wrong?

Watch out for extrapolation.

- The model can fail for *x*-values that are far from the mean of the *x*’s.

Watch out for influential points and outliers.

- Like other analyses, regression can be strongly influenced by outliers.

Watch out for one-tailed tests.

- Software conducts two-tailed tests for regression. If you need one tail, divide the P-value by 2.

Chapter 28

Multiple Regression

28.1

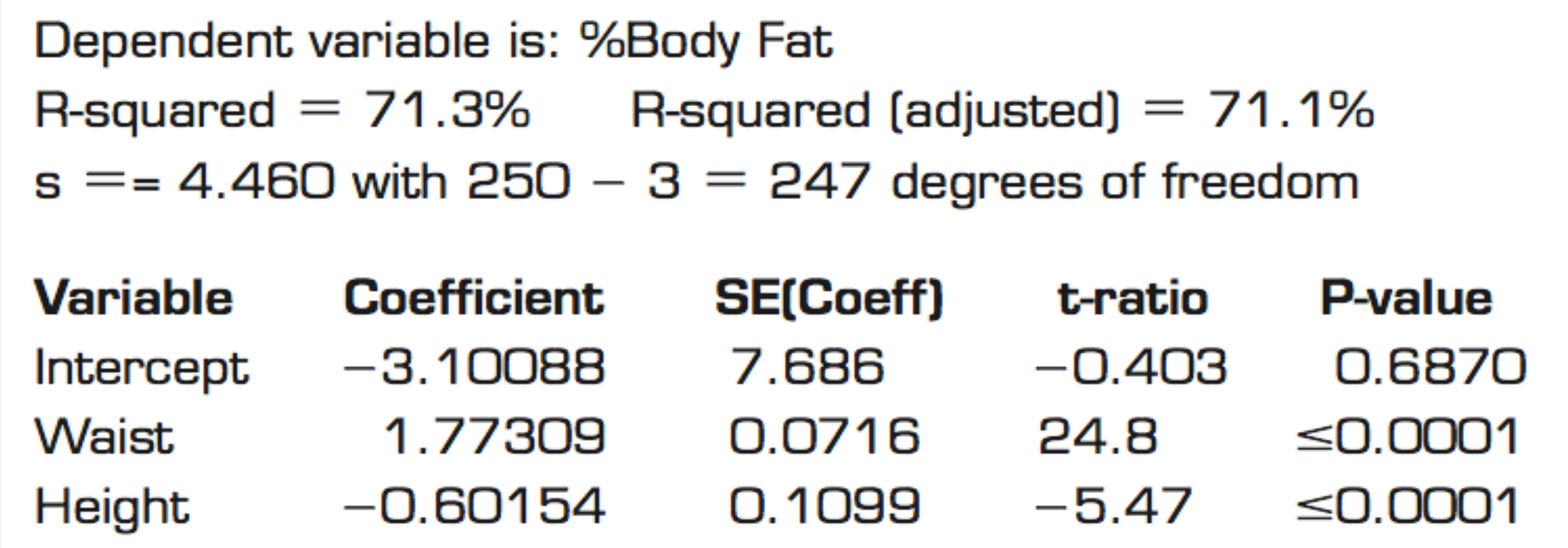
What is Multiple Regression?

Just Do It

- The **method of least squares** can be **expanded** to include more than one predictor. The method is known as **multiple regression**.
- For simple regression we found the Least Squares solution, the one whose coefficients made the sum of the squared residuals as small as possible.
- For multiple regression, we'll do the same thing, but this time with more coefficients.

Just Do It (cont.)

You should recognize most of the numbers in the following example (*%body fat*) of a multiple regression table. Most of them mean what you expect them to.



Dependent variable is: %Body Fat

R-squared = 71.3%   R-squared (adjusted) = 71.1%

s == 4.460 with 250 − 3 = 247 degrees of freedom

Variable	**Coefficient**	**SE(Coeff)**	**t-ratio**	**P-value**
Intercept	−3.10088	7.686	−0.403	0.6870
Waist	1.77309	0.0716	24.8	≤0.0001
Height	−0.60154	0.1099	−5.47	≤0.0001

So What’s New?

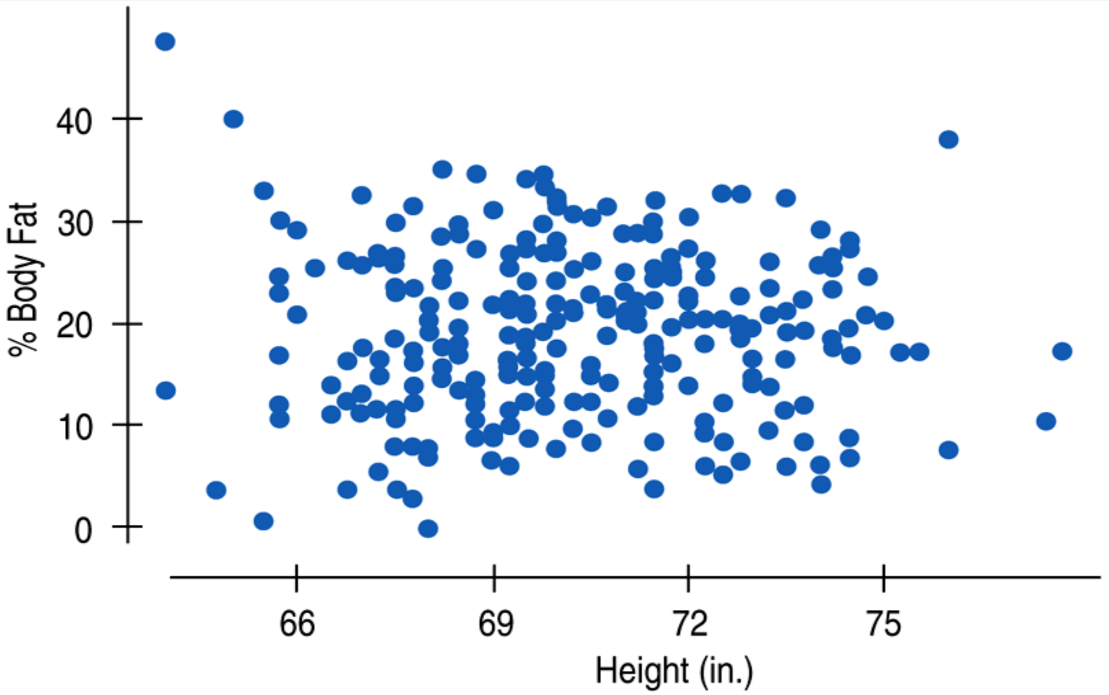
- The *meaning* of the coefficients in the regression model has changed in a subtle but important way.
- Multiple regression is an extraordinarily versatile calculation, underlying many widely used Statistics methods.
- Multiple regression offers our first glimpse into statistical methods that use more than two quantitative variables.

28.2

Interpreting Multiple Regression Coefficients

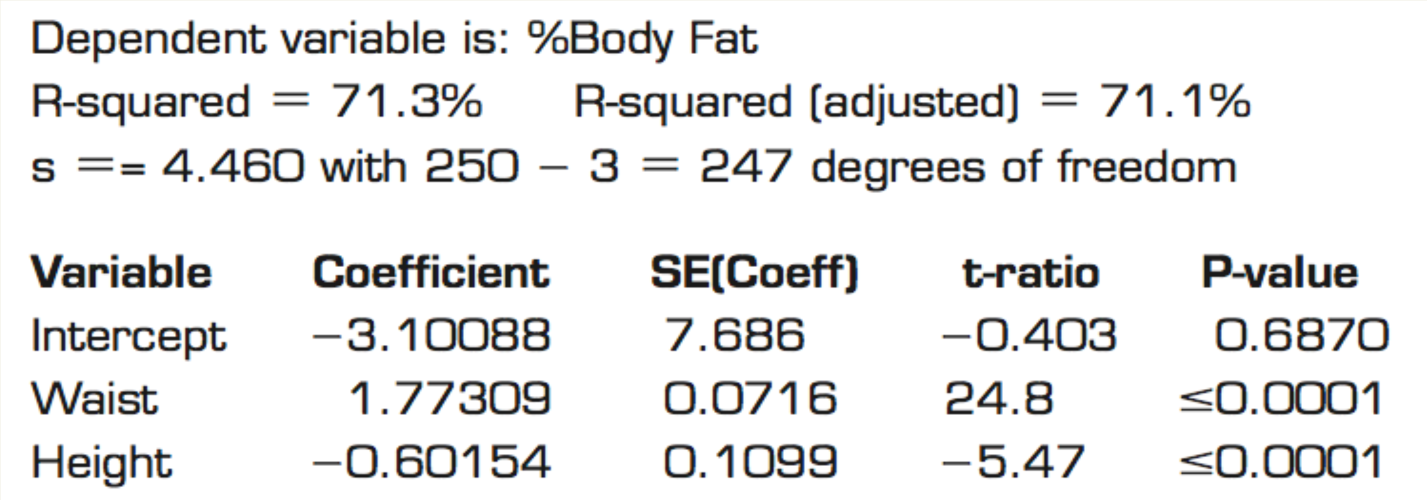
What Multiple Regression Coefficients Mean

- We said that **height** might be important in predicting body fat in men.
- What's the relationship between *%body fat* and *height* in men? Here's the scatterplot:



What Multiple Regression Coefficients Mean (cont.)

- It doesn’t look like *height* tells us much about *%body fat*. Or does it?
- The coefficient of *height* in the multiple regression model was statistically significant, so it *did* contribute to the *multiple* regression model.



Dependent variable is: %Body Fat

R-squared = 71.3% R-squared (adjusted) = 71.1%

s == 4.460 with 250 − 3 = 247 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−3.10088	7.686	−0.403	0.6870
Waist	1.77309	0.0716	24.8	≤0.0001
Height	−0.60154	0.1099	−5.47	≤0.0001

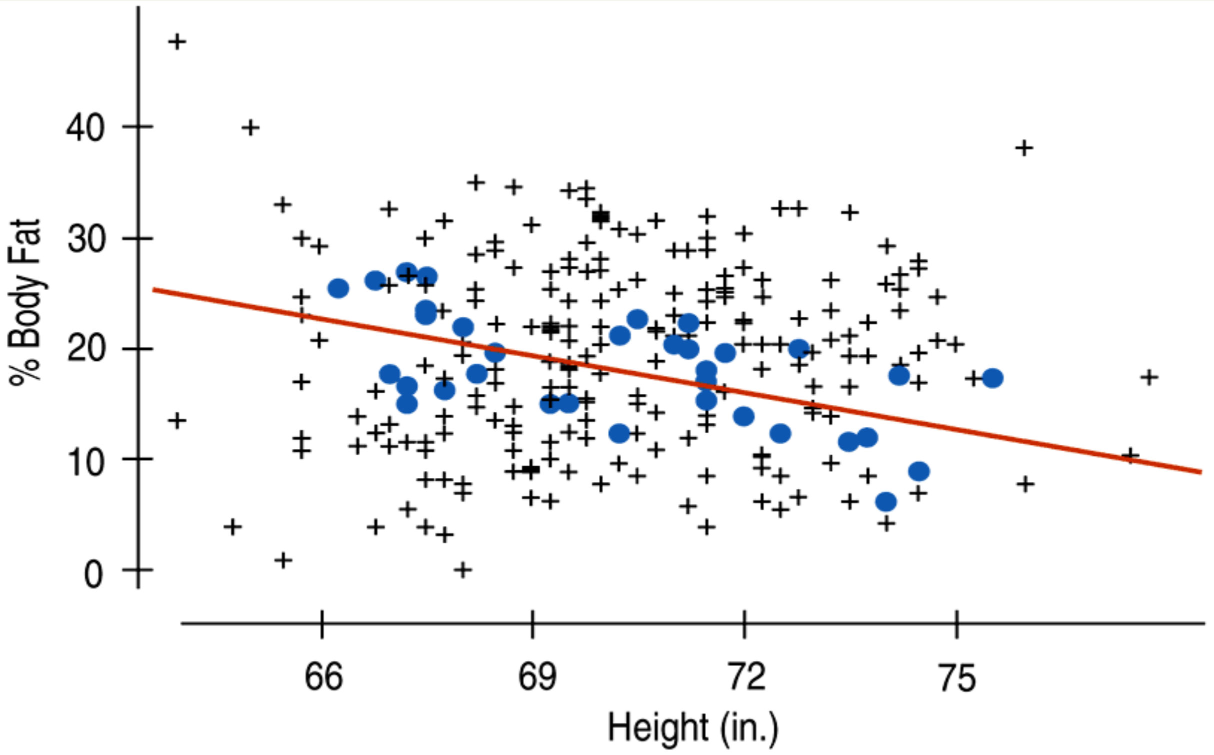
What Multiple Regression Coefficients Mean (cont.)

- It doesn’t look like *height* tells us much about *%body fat*. Or does it?
- The coefficient of *height* in the multiple regression model was statistically significant, so it *did* contribute to the *multiple* regression model.
- How can this be?

  The multiple regression coefficient of *height* takes account of the other predictor (*waist size*) in the regression model.

What Multiple Regression Coefficients Mean (cont.)

For example, when we restrict our attention to men with waist sizes between 36 and 38 inches (points in blue), we can see a relationship between *%body fat* and *height*:



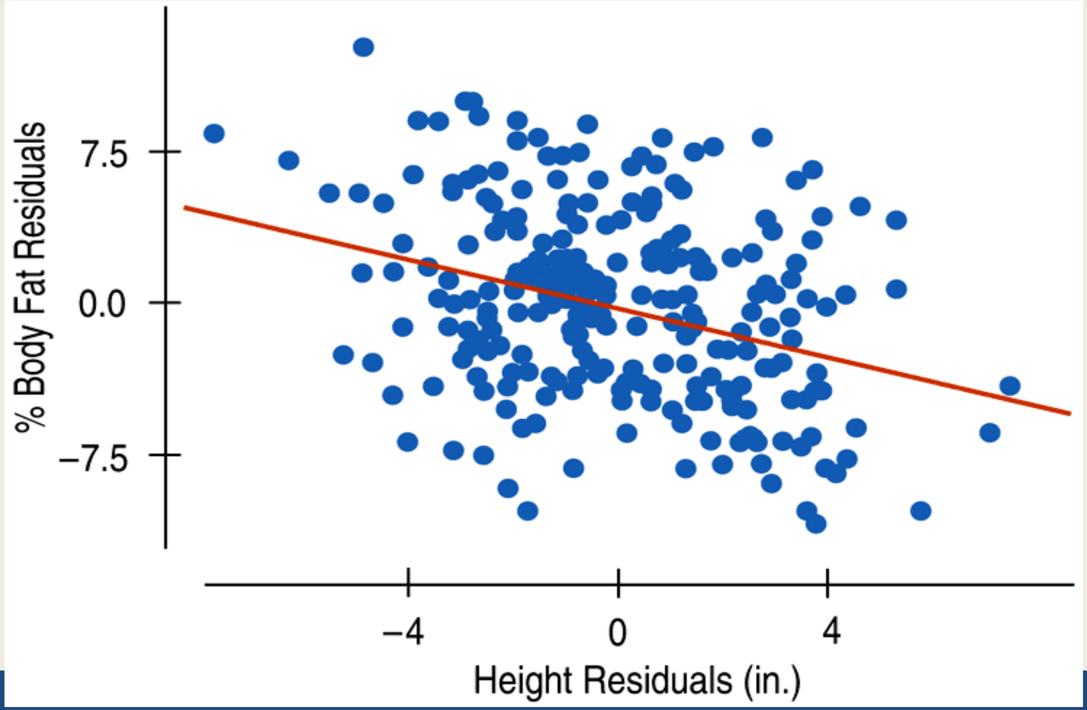
What Multiple Regression Coefficients Mean (cont.)

So, overall there’s little relationship between *%body fat* and *height*, but when we focus on *particular* waist sizes there is a relationship.

- This relationship is **conditional** because we’ve restricted our set to only those men with a certain range of waist sizes.
- For men with that waist size, an extra inch of height is associated with a decrease of about 0.60% in body fat.
- If that relationship is consistent for each *waist* size, then the multiple regression coefficient will estimate it.

What Multiple Regression Coefficients Mean (cont.)

The following partial regression plot based on residuals shows the coefficient of *height* in the regression model has a slope equal to the coefficient value in the multiple regression model:



28.3

The Multiple Regression Model-Assumptions and Conditions

The Simple Regression Model

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

The Multiple Regression Model

For a multiple regression with *k* predictors, the model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

The assumptions and conditions for the multiple regression model sound nearly the same as for simple regression, but with more variables in the model, we’ll have to make a few changes.

Assumptions and Conditions

Linearity Assumption:

- Straight Enough Condition: Check the **partial regression scatterplot** for **each candidate predictor variable**—the shape must not be obviously curved or we can't consider that predictor in our multiple regression model.

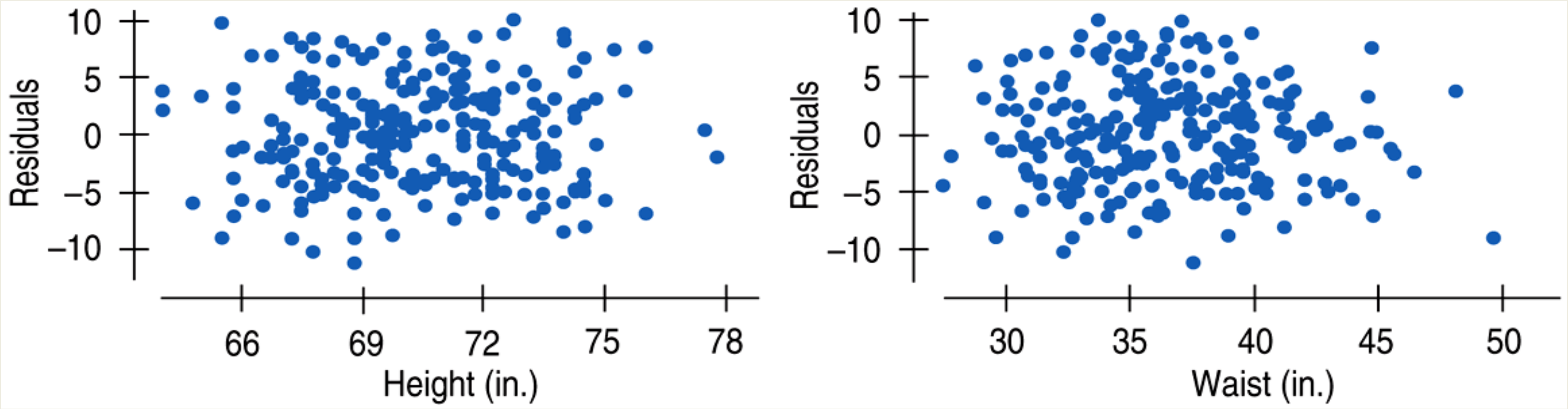
Independence Assumption:

- Randomization Condition: The data should arise from a random sample or randomized experiment. Also, check the residuals plot - the residuals should appear to be randomly scattered.

Assumptions and Conditions (cont.)

Equal Variance Assumption:

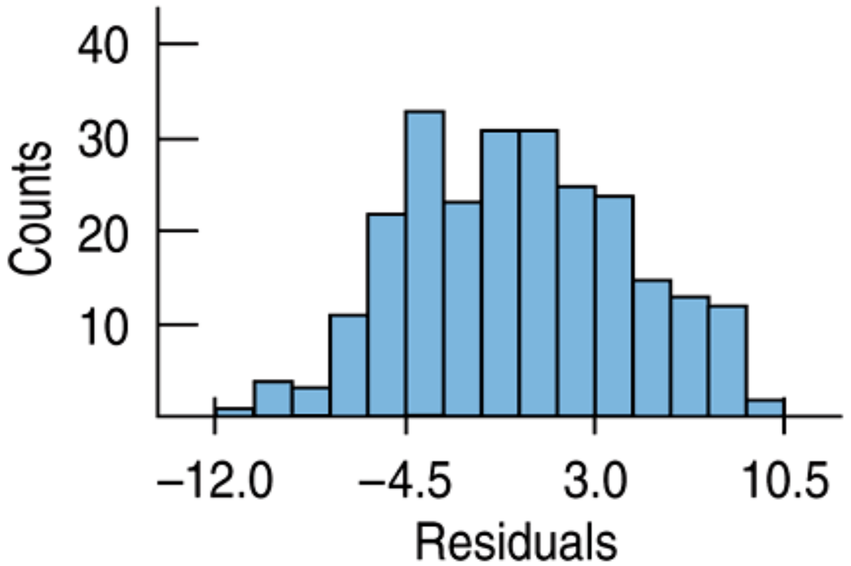
- Does the Plot Thicken? Condition: Check the residuals plot—the spread of the residuals should be uniform.

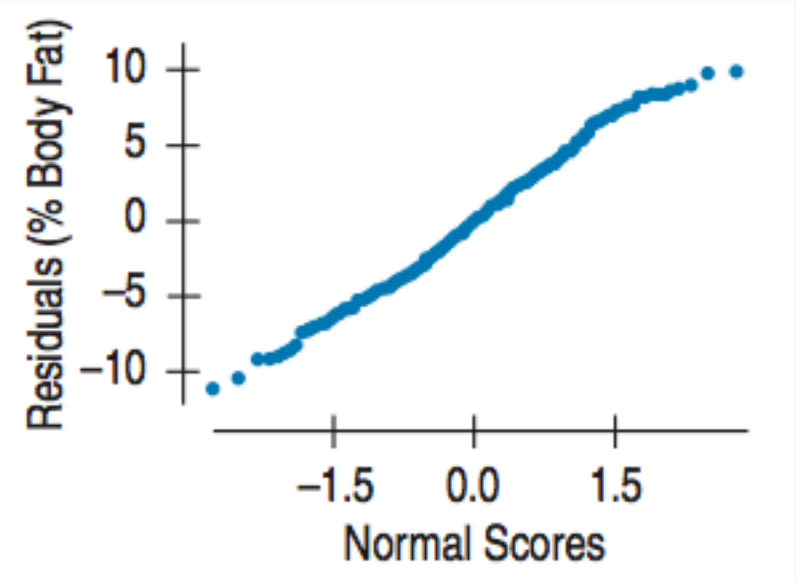


Assumptions and Conditions (cont.)

Normality Assumption:

- Nearly Normal Condition: Check a histogram of the residuals—the distribution of the residuals should be unimodal and symmetric, and the Normal probability plot should be straight.





Assumptions and Conditions (cont.)

Summary of the checks of conditions in order:

1. Check the Straight Enough Condition with **scatterplots of the *y*-variable against each *x*-variable**.
2. If the scatterplots are straight enough, fit a multiple regression model to the data.
3. Find the residuals and predicted values.
4. Make and check a scatterplot of the residuals against the predicted values. This plot should look patternless.

Assumptions and Conditions (cont.)

Summary of the checks of conditions in order:

5. Think about how the data were collected. Randomization? Representative? Plot residuals against time - patterns?
6. If the conditions check out this far, feel free to interpret the regression model and use it for prediction.
7. If you wish to test hypotheses about the coefficients or about the overall regression, then make a histogram and Normal probability plot of the residuals to check the Nearly Normal Condition.

Example: Multiple Regression

Step-By-Step

How should we model %*Body Fat* in terms of *Height* AND *Waist* size?

Example: Multiple Regression

Step-By-Step

Variables Name the variables, report the W’s, and specify the questions of interest:

Have body measurements on 250 adult males from BYU Human Performance Research Center.
Want to understand relationship between %Body Fat, Height, and Waist size.

Example: Multiple Regression

Step-By-Step

Plan

Think about the assumptions and check the conditions.

Straight Enough Condition:

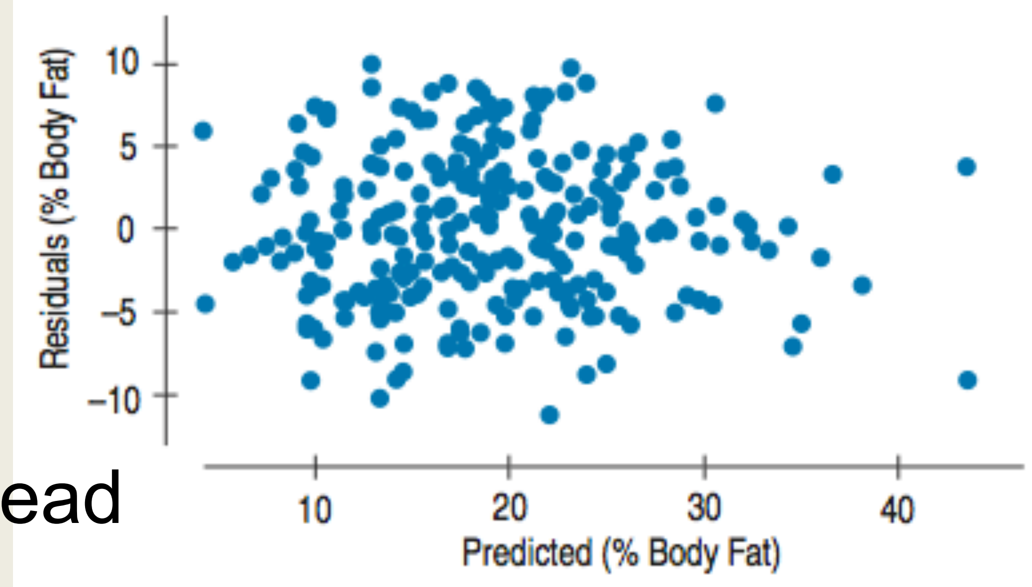
no obvious bend, scatterplot residuals

Independence Assumption:

data presented as representative of male population in U.S.

Does the Plot Thicken?

no obvious changes in the spread



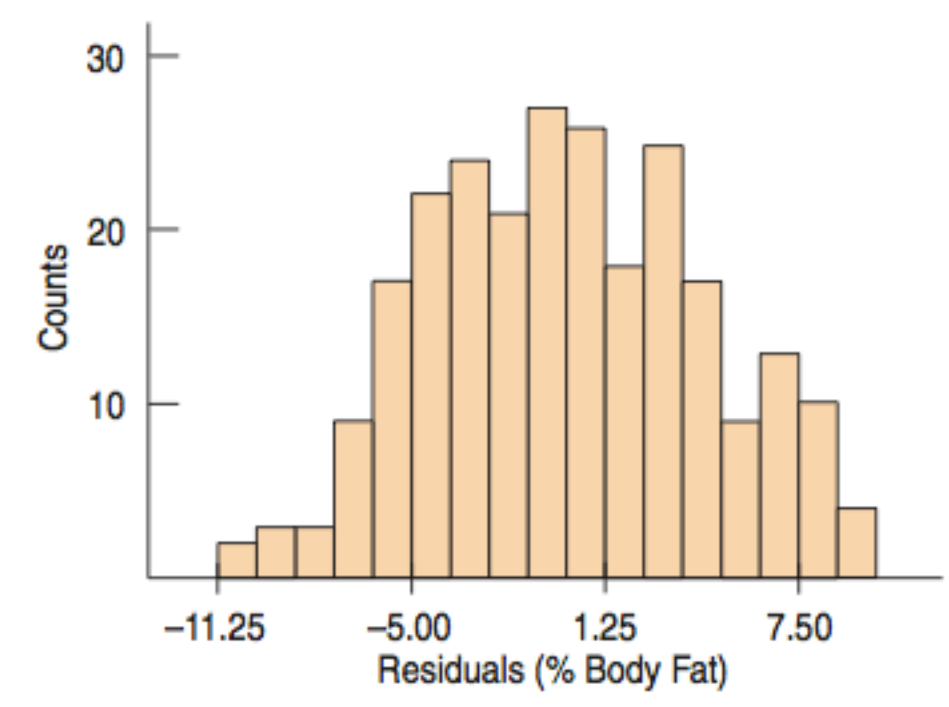
Example: Multiple Regression

Step-By-Step

Plan Think about the assumptions and check the conditions.

Nearly Normal Condition, Outlier Condition:

A histogram of the residuals is unimodal and symmetric.



Example: Multiple Regression

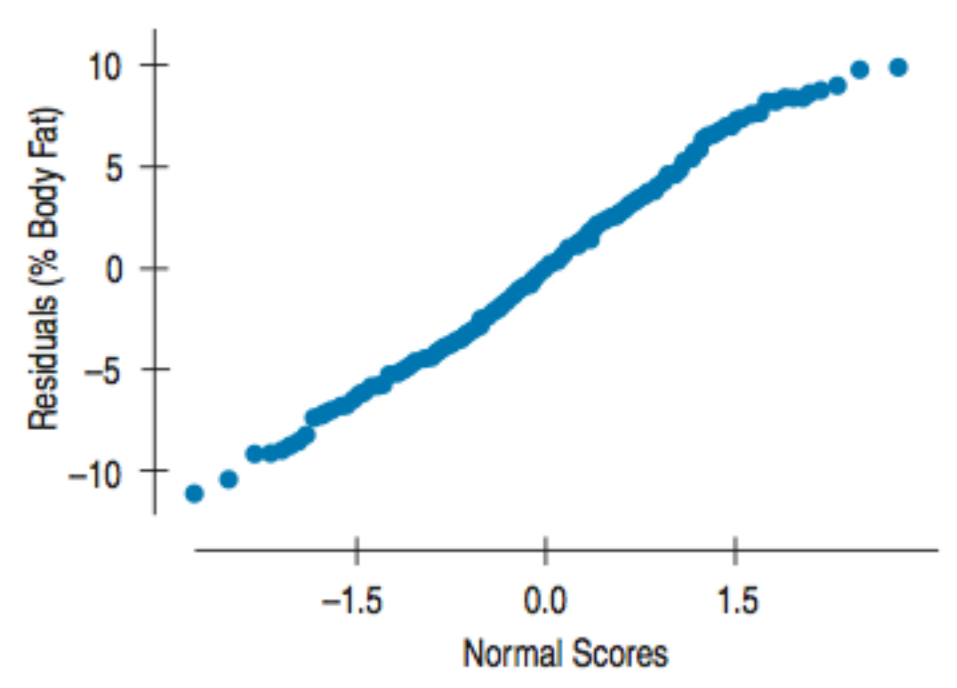
Step-By-Step

Plan

Think about the assumptions and check the conditions.

Nearly Normal Condition, Outlier Condition:

The Normal probability plot of the residuals is reasonably straight:

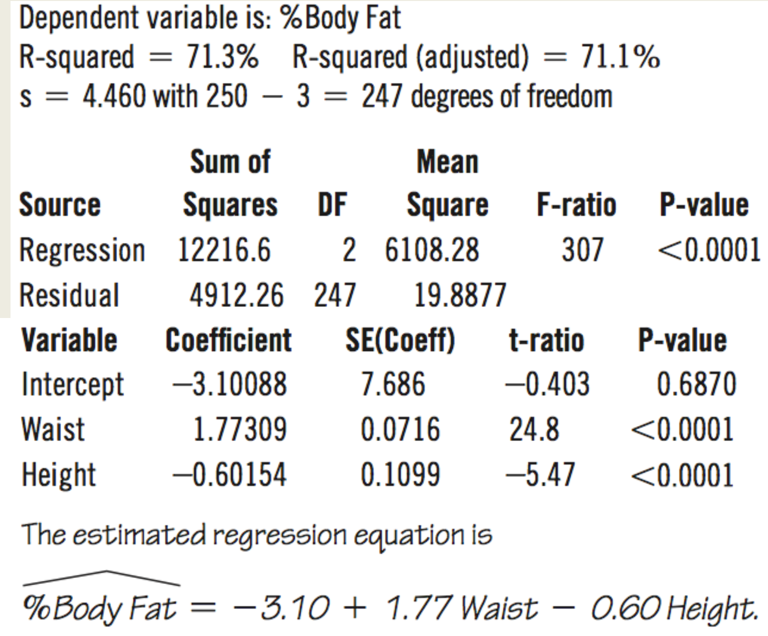


Under these conditions a full multiple regression analysis is appropriate.

Example: Multiple Regression

Step-By-Step

Mechanics Computer output:



```
Dependent variable is: %Body Fat
R-squared = 71.3%    R-squared (adjusted) = 71.1%
s = 4.460 with 250 − 3 = 247 degrees of freedom

               Sum of              Mean
Source         Squares    DF      Square     F-ratio    P-value
Regression    12216.6      2    6108.28        307      <0.0001
Residual       4912.26   247      19.8877
Variable     Coefficient    SE(Coeff)    t-ratio     P-value
Intercept     −3.10088        7.686       −0.403      0.6870
Waist          1.77309        0.0716      24.8       <0.0001
Height        −0.60154        0.1099      −5.47      <0.0001
```

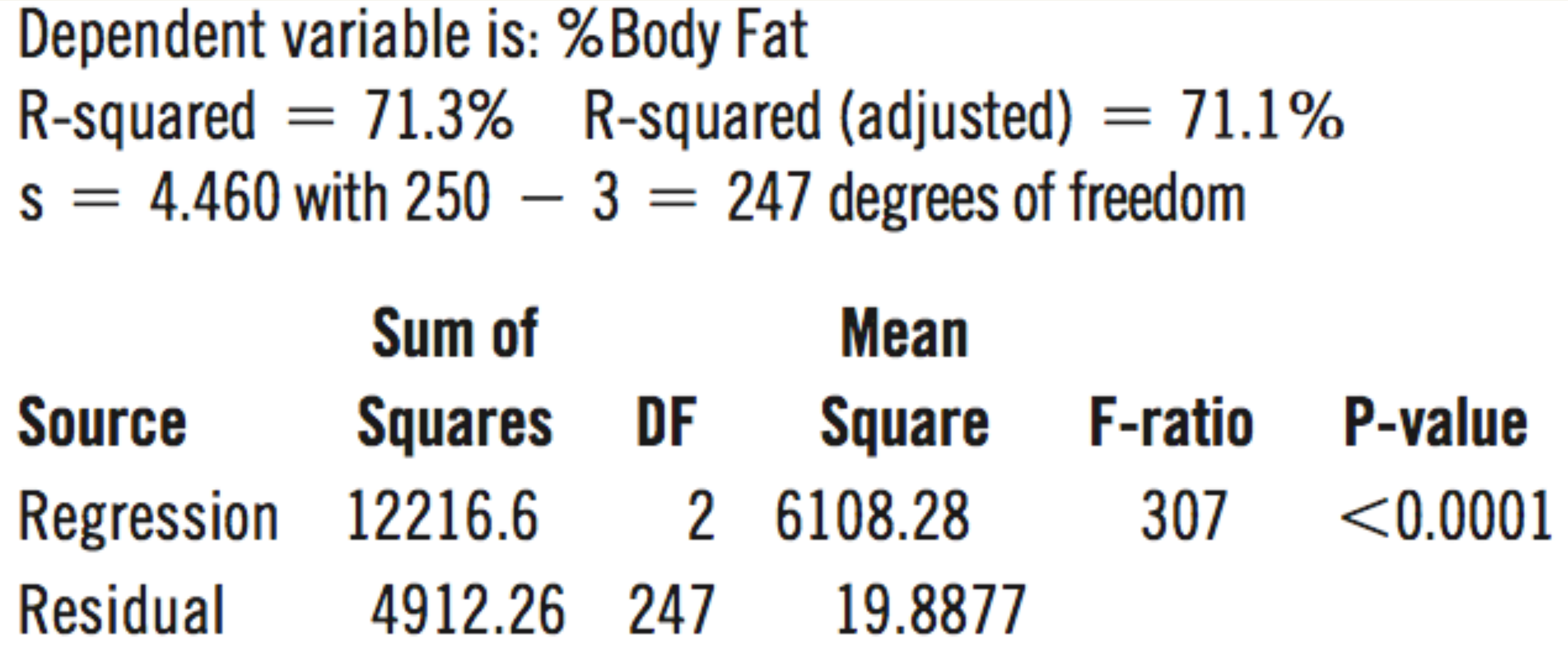
The estimated regression equation is

$$\widehat{\%Body\ Fat} = -3.10 + 1.77\,Waist - 0.60\,Height.$$

Example: Multiple Regression

Step-By-Step

Mechanics Computer output:

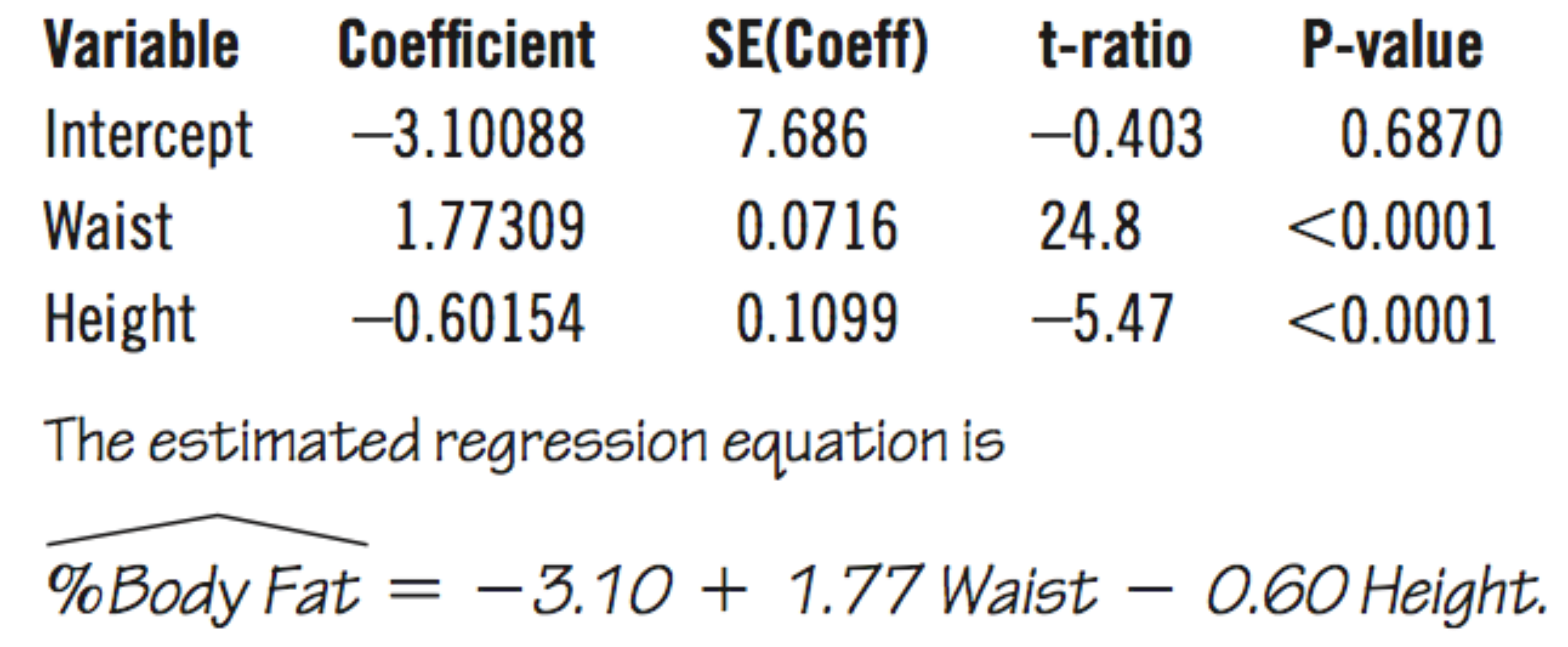


```
Dependent variable is: %Body Fat
R-squared = 71.3%    R-squared (adjusted) = 71.1%
s = 4.460 with 250 − 3 = 247 degrees of freedom

              Sum of            Mean
Source        Squares    DF     Square     F-ratio   P-value
Regression    12216.6      2    6108.28      307     <0.0001
Residual       4912.26   247      19.8877
```

Example: Multiple Regression

Step-By-Step

Mechanics



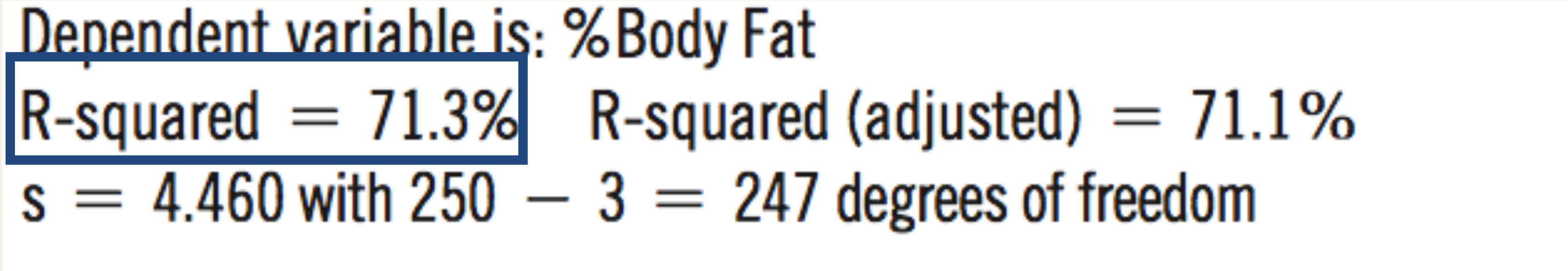
Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	–3.10088	7.686	–0.403	0.6870
Waist	1.77309	0.0716	24.8	<0.0001
Height	–0.60154	0.1099	–5.47	<0.0001

The estimated regression equation is

$$\widehat{\%\text{Body Fat}} = -3.10 + 1.77\,\text{Waist} - 0.60\,\text{Height}.$$

Example: Multiple Regression Step-By-Step

Interpretation



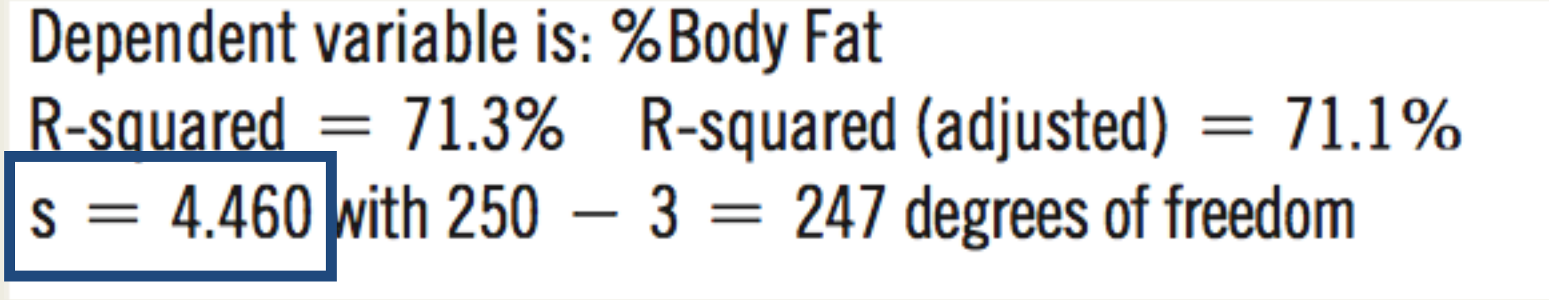
```
Dependent variable is: %Body Fat
R-squared = 71.3%    R-squared (adjusted) = 71.1%
s = 4.460 with 250 − 3 = 247 degrees of freedom
```

The R^2 for the regression is 71.3%.

Waist size and *Height* together account for about 71% of the variation.

Example: Multiple Regression Step-By-Step

Interpretation

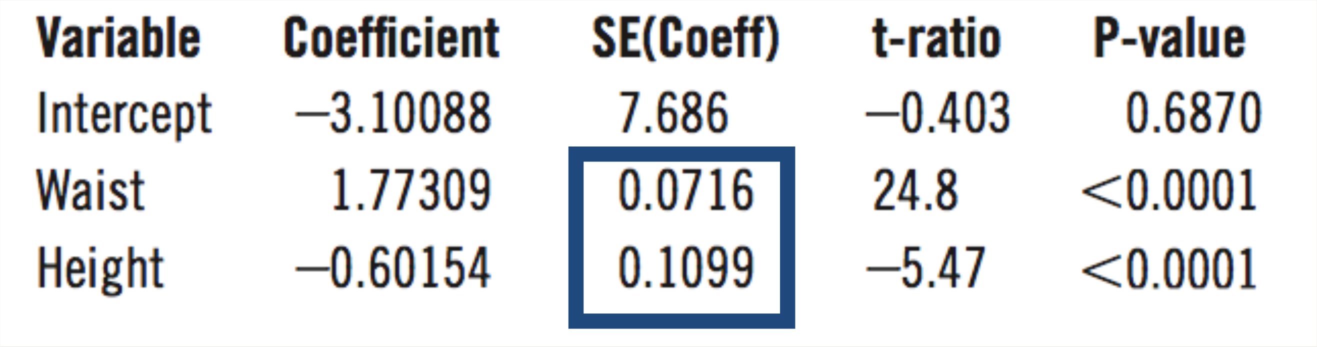


```
Dependent variable is: %Body Fat
R-squared = 71.3%    R-squared (adjusted) = 71.1%
s = 4.460 with 250 − 3 = 247 degrees of freedom
```

The residuals have a standard deviation of 4.46%, which gives an indication of how precisely we can predict *%Body Fat* with this model.

Example: Multiple Regression Step-By-Step

Interpretation



Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−3.10088	7.686	−0.403	0.6870
Waist	1.77309	0.0716	24.8	<0.0001
Height	−0.60154	0.1099	−5.47	<0.0001

The standard errors for the slopes of 0.07 (*Waist*) and 0.11 (*Height*) are both small compared with the slopes themselves

→ looks like the coefficient estimates are fairly precise.

Example: Multiple Regression Step-By-Step

Interpretation

$$\widehat{\%Body\ Fat} = -3.10 + 1.77\boxed{Waist} - 0.60\boxed{Height.}$$

Each inch in *Waist* size is associated with about a 1.77 increase in *%Body Fat* among men who are of a particular *Height*

Each inch of *Height* is associated with a decrease in %Body Fat of about 0.60 among men with a particular *Waist* size.