Probability and Mathematical Statistics in Data Science

Lecture 36: Section 12.3 Towards Multiple Regression

The Multiple RegressionModel

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

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_kx_k+\varepsilon$$

- 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.

Multiple Linear Regression

- With simple linear regression, we used a single explanatory variable to make predictions for the response variable

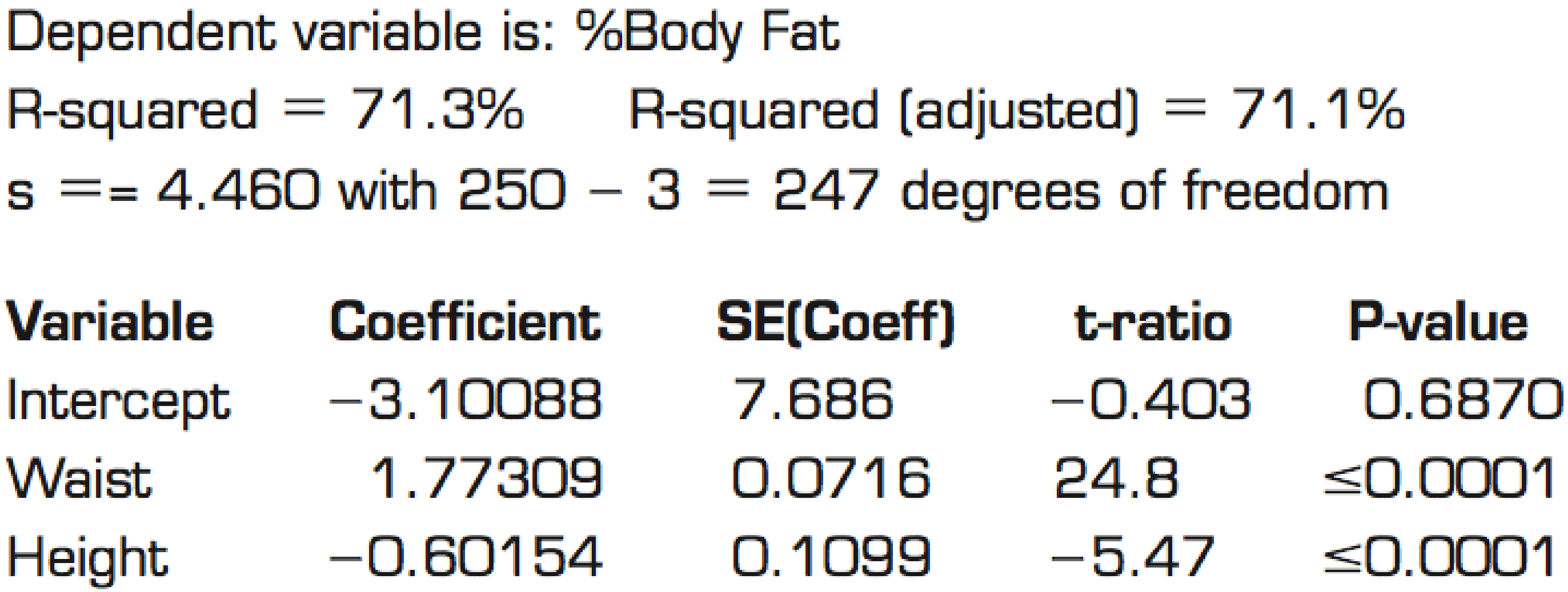
- With multiple linear regression, we can use multiple explanatory variables to make predictions for the response variable.

- Multiple linear regression also enables us to include controlling variables in our linear model

- It enables us to build statistical models that are a better reflection of how the world works and can lead to more accurate predictions for the response variable

Multiple Regression

- 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

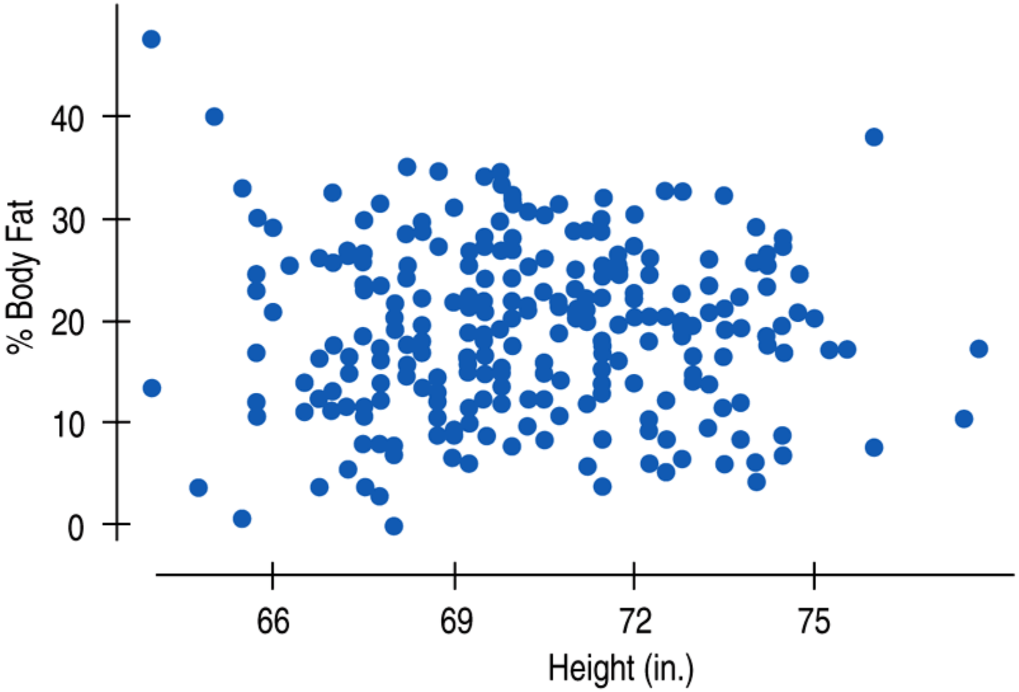
Variable	**Coefficient**	**SE(Coeff)**	**t-ratio**	**P-value**
Intercept	−3.10088	7.686	−0.403	0.6870
Waist	1.77309	0.0716	24.8	≤0.0001
Height	−0.60154	0.1099	−5.47	≤0.0001

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.

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

- It doesn’t look like *height* tells us much about *%body fat*. Or does it?

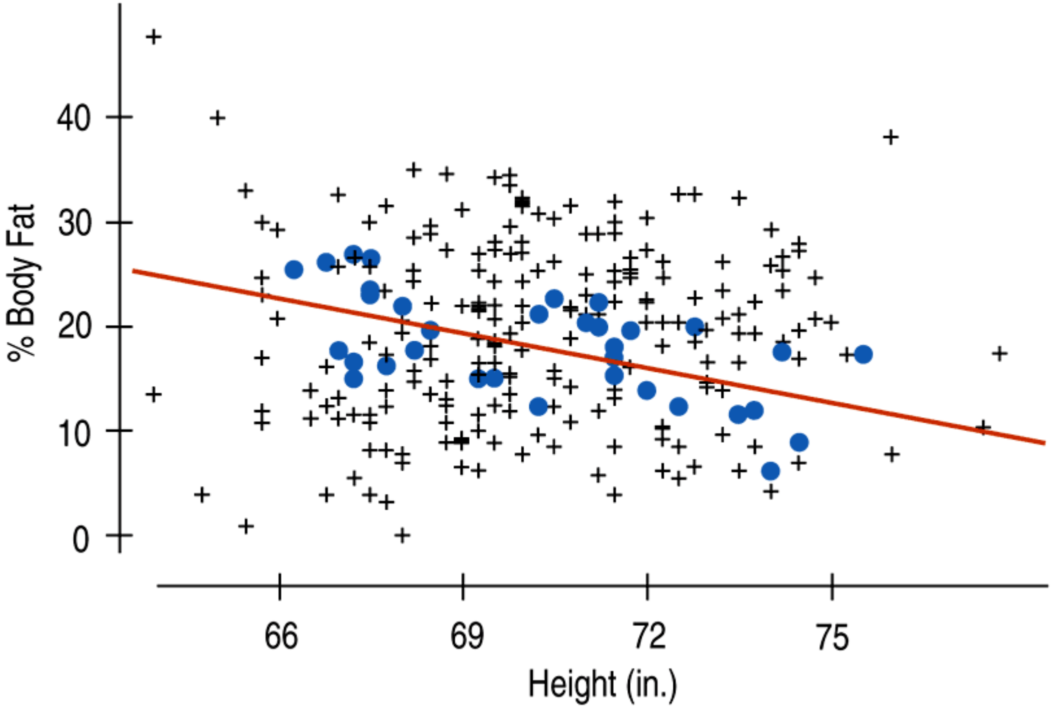
- The coefficient of *height* in the multiple regression model was statistically significant, so it *did* contribute to the *multiple* regression model.

- 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

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

Pred %Body Fat = -3.10 + 1.77(Waist) – 0.60(Height)



What Multiple Regression Coefficients Mean

- 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 waist size.
  - 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.

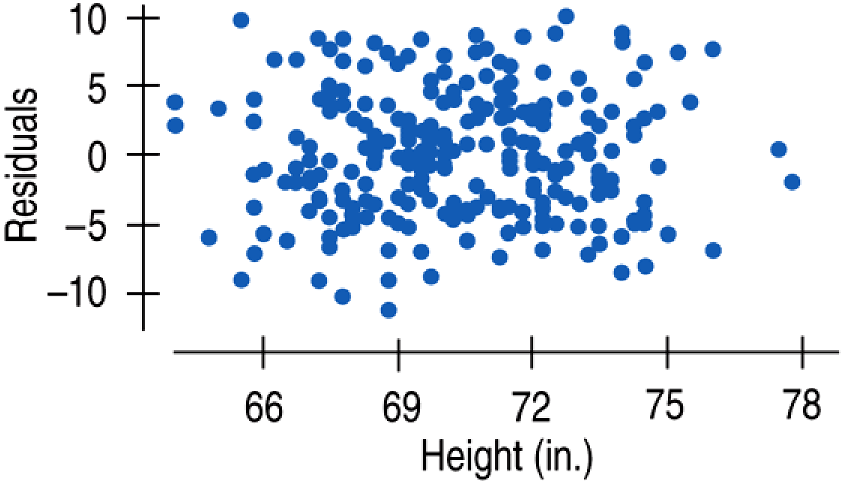
Assumptions and Conditions

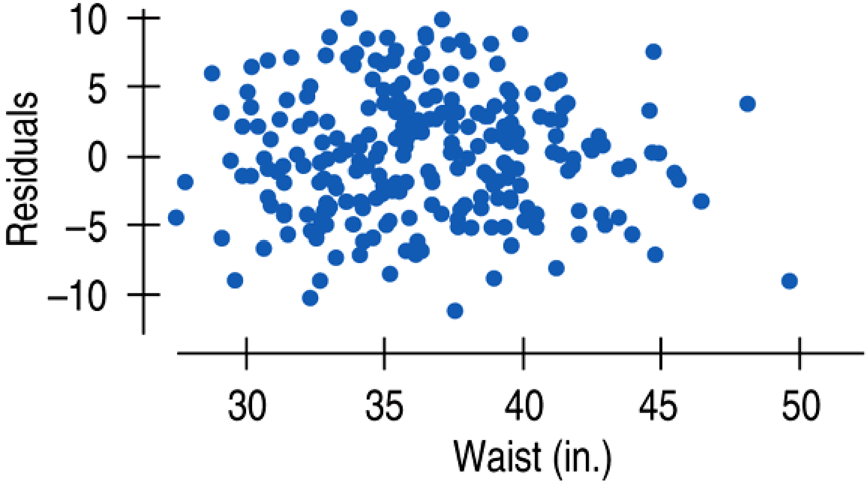
- **Linearity Assumption:**
  - **Straight Enough Condition**: Check the scatterplot for each candidate predictor variable—the shape must not be obviously curved or we can't consider that predictor in our multiple regression model.

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

Assumptions and Conditions

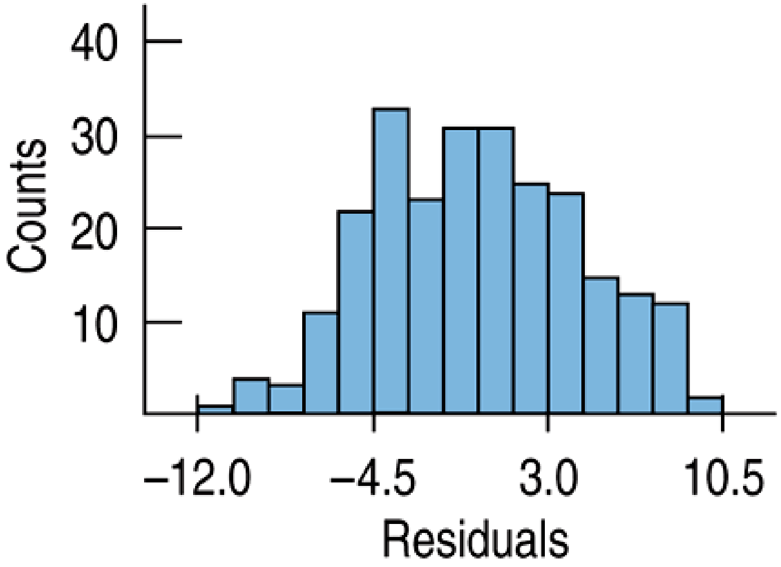
- **Equal Variance Assumption:**
  - **Does the Plot Thicken? Condition**: Check the residuals plot—the spread of the residuals should be uniform.





Assumptions and Conditions

- **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.



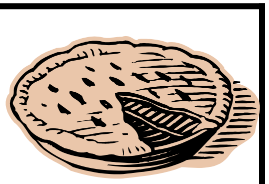
Multiple Regression Inference: I Thought I Saw an ANOVA Table…

- Now that we have more than one predictor, there's an overall test we should consider before we do more inference on the coefficients.
- We ask the global question "Is this multiple regression model any good at all?"
- We test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

- The *F*-statistic and associated P-value from the ANOVA table are used to answer our question.

Multiple Linear Regression



Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Multiple Regression Inference: Testing the Coefficients

- Once we check the *F*-test and reject the null hypothesis, we can move on to checking the test statistics for the individual coefficients.
- For each coefficient, we test

$$H_0 : \beta_j = 0$$

- If the assumptions and conditions are met (including the Nearly Normal Condition), these ratios follow a Student's *t*-distribution.

$$t_{n-k-1} = \frac{b_j - 0}{SE\left(b_j\right)}$$

Multiple Regression Inference II: Testing the Coefficients

- We can also find a confidence interval for each coefficient:

$$b_j \pm t^*_{n-k-1} \times SE\left(b_j\right)$$

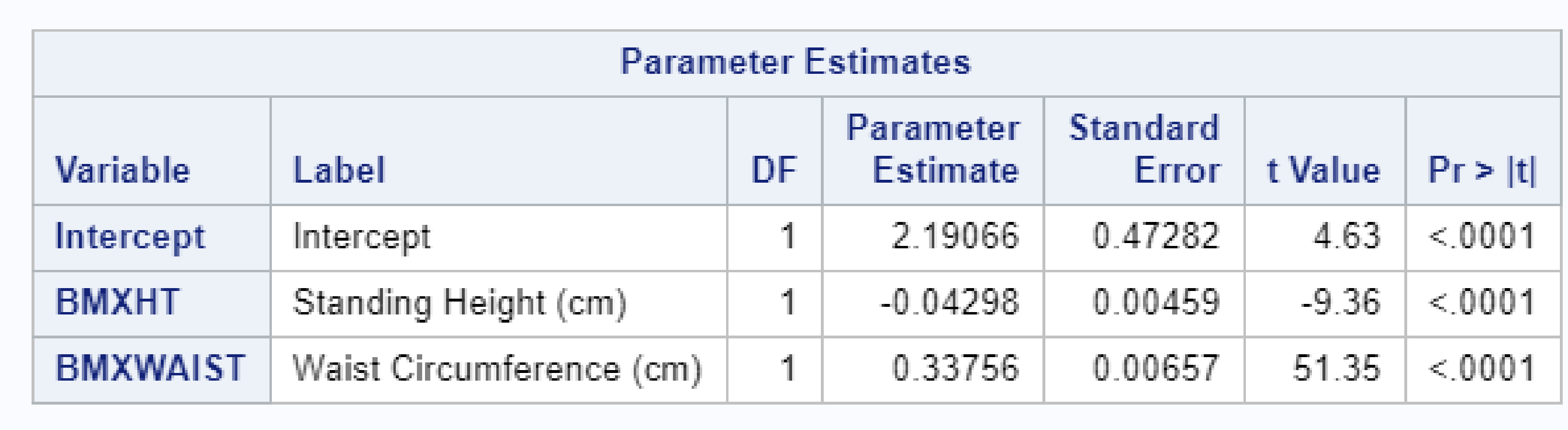
NHANES National Youth Fitness Survey

- To help us understand the statistical analysis from a multiple regression analysis, we will make use of the youth fitness survey data

- The response (predictor or dependent) variable will be BMXBMI and the explanatory variables with be BMXHT and BMXWAIST

- We are looking to see if the there is statistical evidence of a linear relationship between height, waist size and bmi.

Multiple Regression



Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	2.19066	0.47282	4.63	<.0001
BMXHT	Standing Height (cm)	1	-0.04298	0.00459	-9.36	<.0001
BMXWAIST	Waist Circumference (cm)	1	0.33756	0.00657	51.35	<.0001

- The last column in the output titled Pr > |t| contains the p-values

- Predicted BMI = 2.19 - 0.043 x Height + 0.34 x Waist

- For a multiple regression, there is a slope value for each of the explanatory variables which we call **coefficients**.

- However, the meaning of the coefficient for each explanatory variable has changed in a subtle but important to understand way!

Multiple Regression

- Predicted BMI = 2.19 - 0.043 x Height + 0.34 x Waist

- Now we can begin to understand the meaning of the coefficients in a multiple regression model

- For every centimeter change in height, the predicted value of BMI will change by -0.043 kg/m**2 **while holding the value of waist constant**

- For a child 140 cm tall with a 60 cm waist, then:

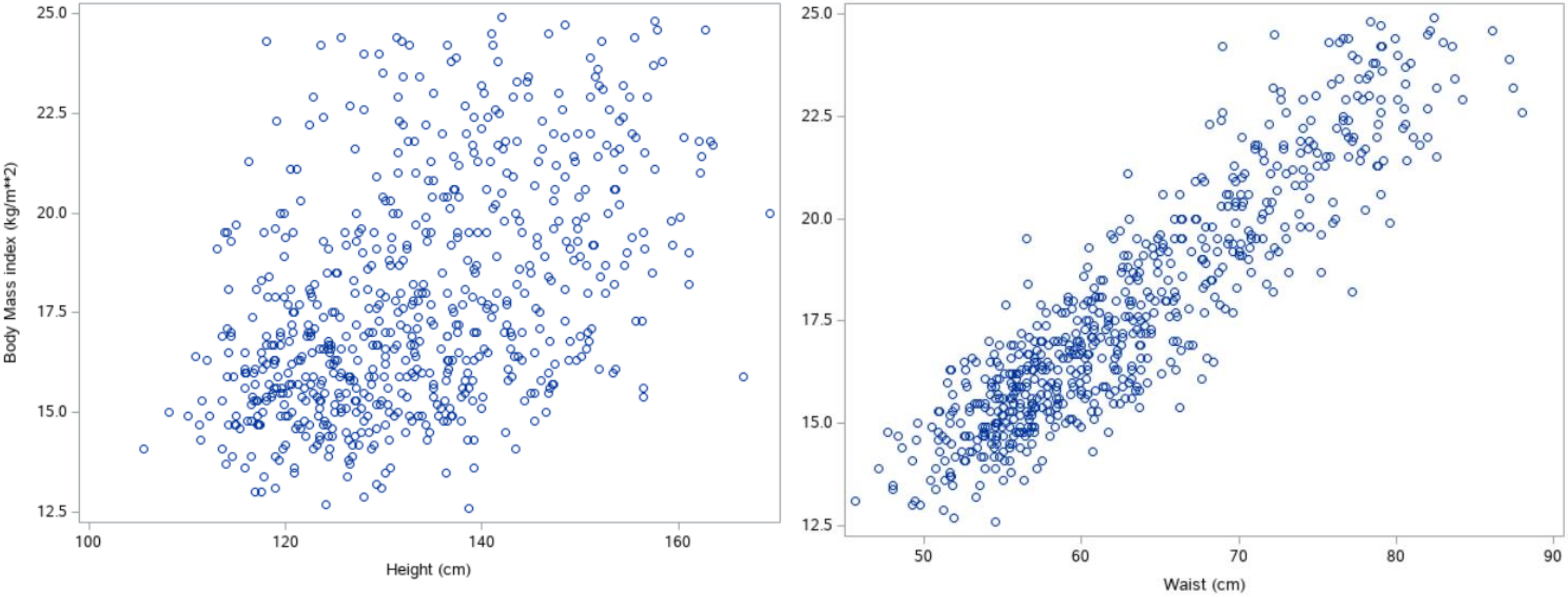
  Predicted BMI = 2.19 - 0.043 x (140) + 0.34 x (60) = 17.430

- For a child 141 cm tall with a 60 cm waist, then:

  Predicted BMI = 2.19 - 0.043 x (140) + 0.34 x (60) = 17.387

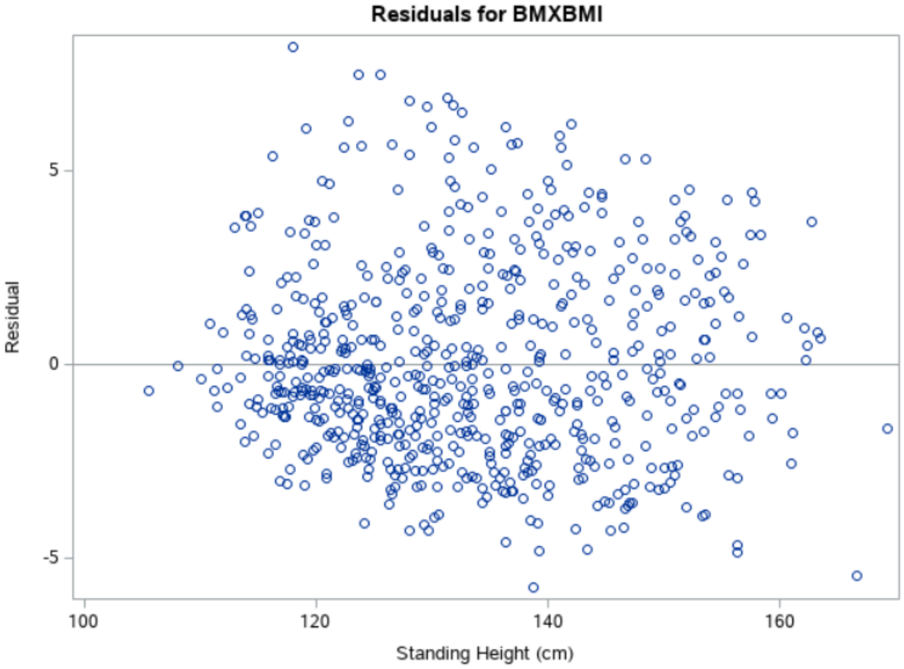
Checking the Assumptions and Conditions

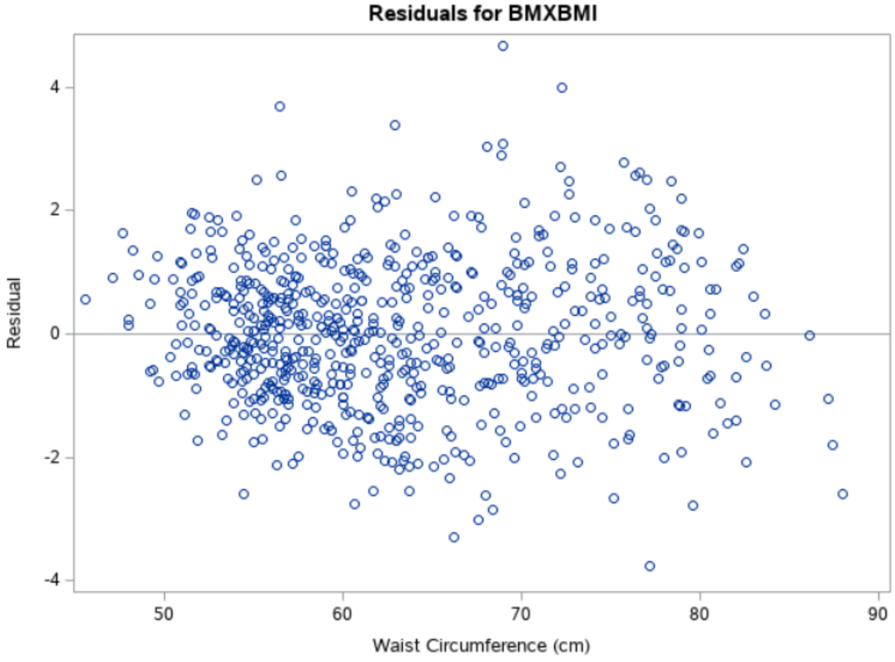
- **Linearity:** Check the scatterplots for each of the explanatory variables against the response variable. There should be no obvious curvature in the scatterplots



Checking the Assumptions and Conditions

- **Independence:** As discussed previously, the NNYFS collected nationally representative (random sample of) data. Residual plots should be a random scatter of points:

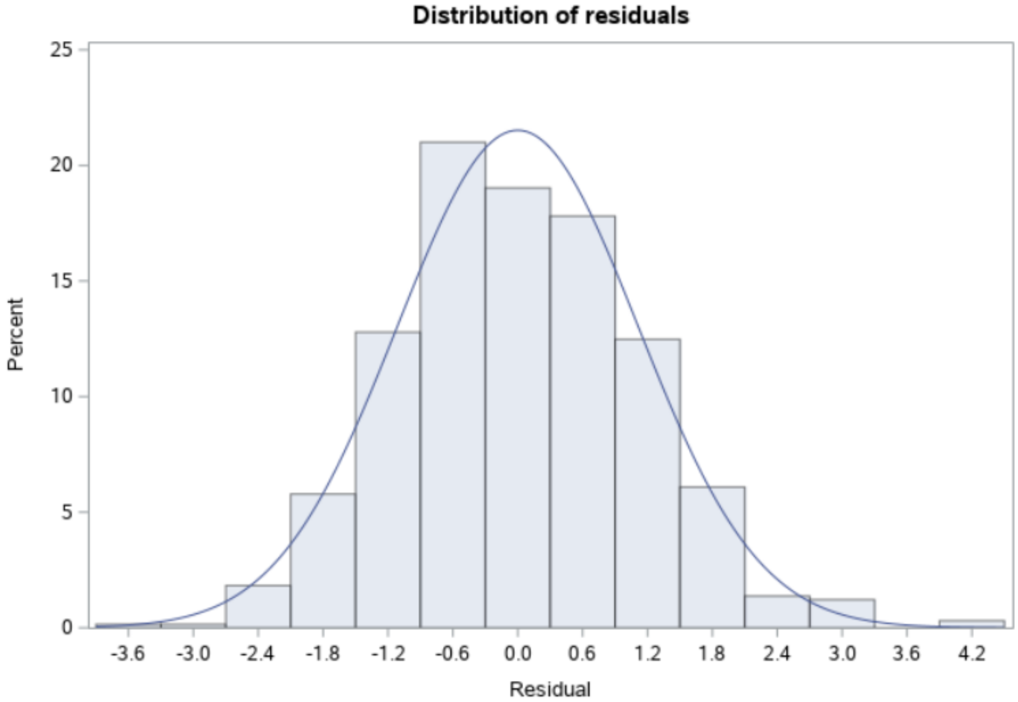




- **Equal Variance:** Residuals should be evenly spread across all values of each of the explanatory variables

Checking the Assumptions and Conditions

- **Normal Distribution of Residuals:** The residuals should be (approximately) normal



- If all the assumptions and conditions are met, we can say that the multiple regression analysis is valid

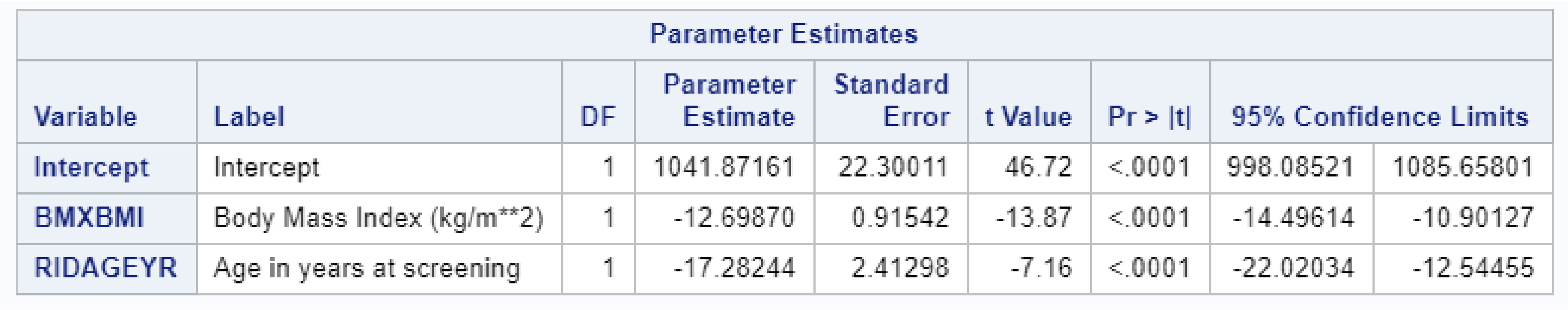
NHANES National Youth Fitness Survey

- In our simple linear regression model, we looked at the relationship between the explanatory variable Body Mass Index (BMI) and the response variable Maximum Endurance Time (MET) for the sample of children surveyed.

- We will now include another explanatory variable in our model.

- We will include the variable called RIDAGEYR - Age in years at screening. We want to analyze whether Age along with BMI are statistically significant predictors of MET

NHANES National Youth Fitness Survey



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	1041.87161	22.30011	46.72	<.0001	998.08521	1085.65801
BMXBMI	Body Mass Index (kg/m**2)	1	-12.69870	0.91542	-13.87	<.0001	-14.49614	-10.90127
RIDAGEYR	Age in years at screening	1	-17.28244	2.41298	-7.16	<.0001	-22.02034	-12.54455

- The p-values for both BMI and Age are highly statistically significant.
- This means that we have found statistical evidence in our data of a relationship between the two explanatory variables and the response variable MET
- The regression analysis also includes confidence intervals for the population coefficients

NHANES National Youth Fitness Survey

Predicted MET = 1041.87 – 12.70 x BMI - 17.28 x Age

- For every unit increase in BMI, the predicted value of MET will decrease by 12.70 while holding Age constant

- For every one year increase in Age the predicted value of MET will decrease by -17.28 while holding BMI constant

- For a 8 year child with a BMI equal to 17 then:

Predicted MET = 1041.87 – 12.70 x (17) - 17.28 x (8)
= 532.21 seconds

NHANES National Youth Fitness Survey

95% Confidence Intervals for the Population Coefficients

BMI: [-14.50, -10.90]

Q. What does the confidence interval mean?

AGE: [-22.02, -12.54]

Q. What does the confidence interval mean?

Measuring the Amount of Variation Explained by the Linear Regression Model

- When we were looking at the relationship between MET and the single explanatory variable BMI, **R-Square** was equal to 0.28 (-0.53 x -0.53). 28% of the variability in MET scores can be explained by a linear relationship with BMI

- When we were looking at the relationship between MET and the two explanatory variables BMI, and AGE, **R-Square** is equal to 0.34. 34% of the variability in MET scores can be explained by a linear relationship with BMI and AGE

- In other words, the variable AGE explained a further 6% of the variability in MET

Comparing Multiple Regression Models

- How do we know that some other choice of predictors might not provide a better model?

- What exactly *would* make an alternative model better?

- These questions are not easy—there's no simple measure of the success of a multiple regression model.

Comparing Multiple Regression Models

- Regression models should make sense.
  - Predictors that are easy to understand are usually better choices than obscure variables.
  - Similarly, if there is a known mechanism by which a predictor has an effect on the response variable, that predictor is usually a good choice for the regression model.
- The simple answer is that we can't know whether we have the best possible model.

Coefficient of Multiple Determination

- Reports the proportion of total variation in Y explained by all X variables taken together

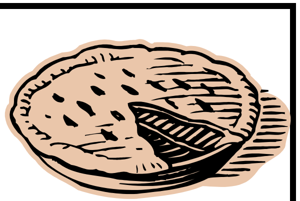
$$\boxed{R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}}$$

$$R^2 = 1 - \frac{SS_{Residual}}{SS_{Total}}$$

Multiple Coefficient of Determination

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

$$R^2 = \frac{SSR}{SST} = \frac{29460.0}{56493.3} = .52148$$



52.1% of the variation in pie sales is explained by the a linear relationship with price and advertising

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Adjusted R^2

- There is another statistic in the full regression table called the **adjusted R^2**.

- This statistic is a rough attempt to adjust for the simple fact that when we add another predictor to a multiple regression, the R^2 can’t go down and will most likely get larger.

- This fact makes it difficult to compare alternative regression models that have different numbers of predictors.

Adjusted R^2

- Shows the proportion of variation in Y explained by all X variables adjusted for the number of X variables used

Adjusted R_Sqr = 1- [(1-R_Sqr)(n-1)/(n-k-1)]

- n = sample size, k = no. of x variables used
  - Penalize excessive use of independent variables
  - Smaller than R^2
  - Useful in comparing among models

Adjusted R^2 in Excel

Regression Statistics	
Multiple R	0.72213
R Square	0.52148
Adjusted R Square	0.44172
Standard Error	47.46341
Observations	15

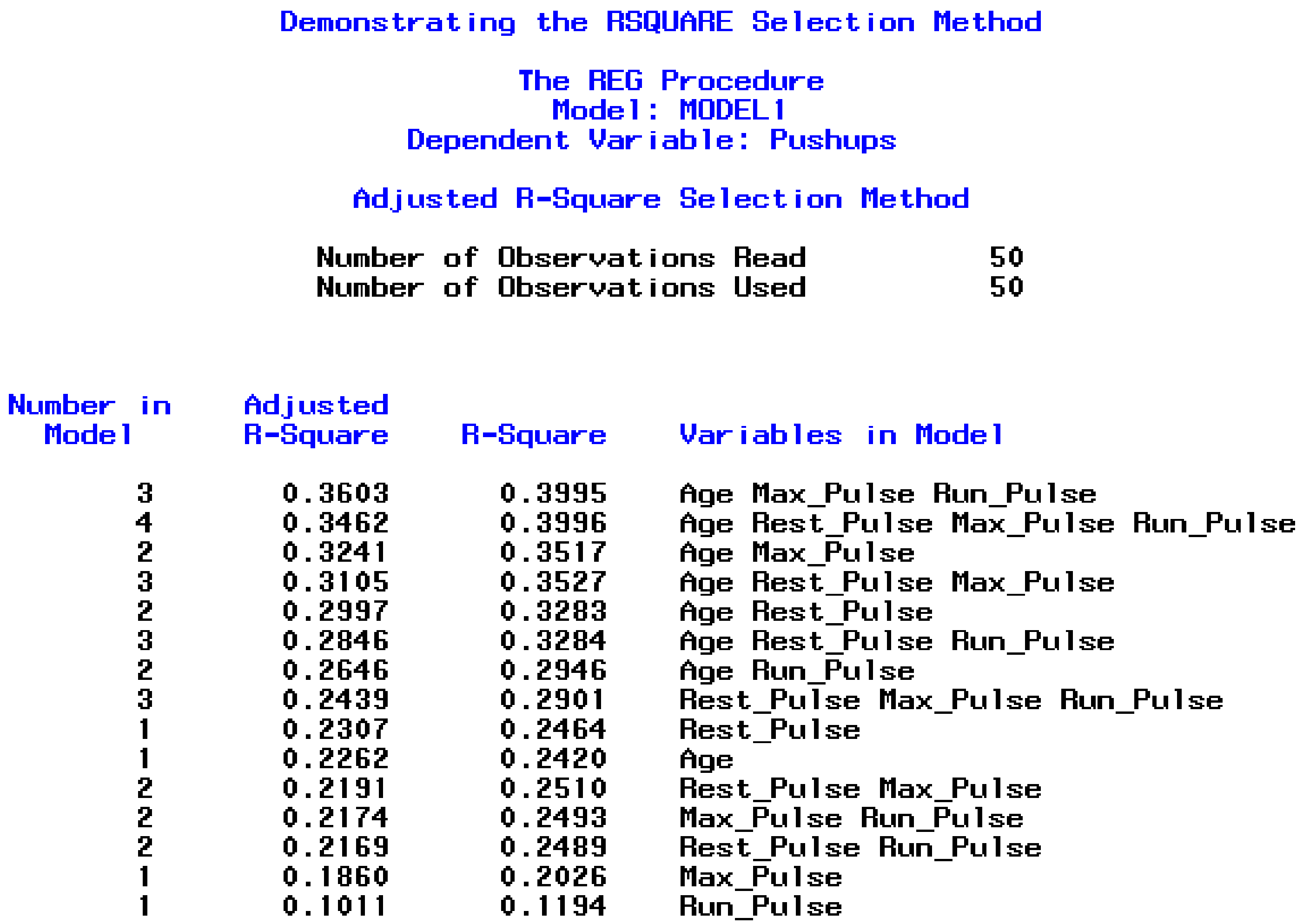
$$R^2_{adj} = .44172$$

44.2% of the variation in pie sales is explained by the variation in price and advertising, taking into account the sample size and number of independent variables

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	2	29460.027	14730.013	6.53861	0.01201
Residual	12	27033.306	2252.776		
Total	14	56493.333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	306.52619	114.25389	2.68285	0.01993	57.58835	555.46404
Price	-24.97509	10.83213	-2.30565	0.03979	-48.57626	-1.37392
Advertising	74.13096	25.96732	2.85478	0.01449	17.55303	130.70888

Adjusted RSQUARE Selection Method



```
                 Demonstrating the RSQUARE Selection Method

                          The REG Procedure
                           Model: MODEL1
                    Dependent Variable: Pushups

                Adjusted R-Square Selection Method

              Number of Observations Read          50
              Number of Observations Used          50


Number in     Adjusted
  Model       R-Square     R-Square     Variables in Model

      3         0.3603       0.3995     Age Max_Pulse Run_Pulse
      4         0.3462       0.3996     Age Rest_Pulse Max_Pulse Run_Pulse
      2         0.3241       0.3517     Age Max_Pulse
      3         0.3105       0.3527     Age Rest_Pulse Max_Pulse
      2         0.2997       0.3283     Age Rest_Pulse
      3         0.2846       0.3284     Age Rest_Pulse Run_Pulse
      2         0.2646       0.2946     Age Run_Pulse
      3         0.2439       0.2901     Rest_Pulse Max_Pulse Run_Pulse
      1         0.2307       0.2464     Rest_Pulse
      1         0.2262       0.2420     Age
      2         0.2191       0.2510     Rest_Pulse Max_Pulse
      2         0.2174       0.2493     Max_Pulse Run_Pulse
      2         0.2169       0.2489     Rest_Pulse Run_Pulse
      1         0.1860       0.2026     Max_Pulse
      1         0.1011       0.1194     Run_Pulse
```

The Best Multiple Regression Model

- The first and most important thing to realize is that often there is no such thing as the "best" regression model. (After all, all models are wrong.)

- Multiple regressions are subtle. The choice of which predictors to use determines almost everything about the regression.

The Best Multiple Regression Model (cont)

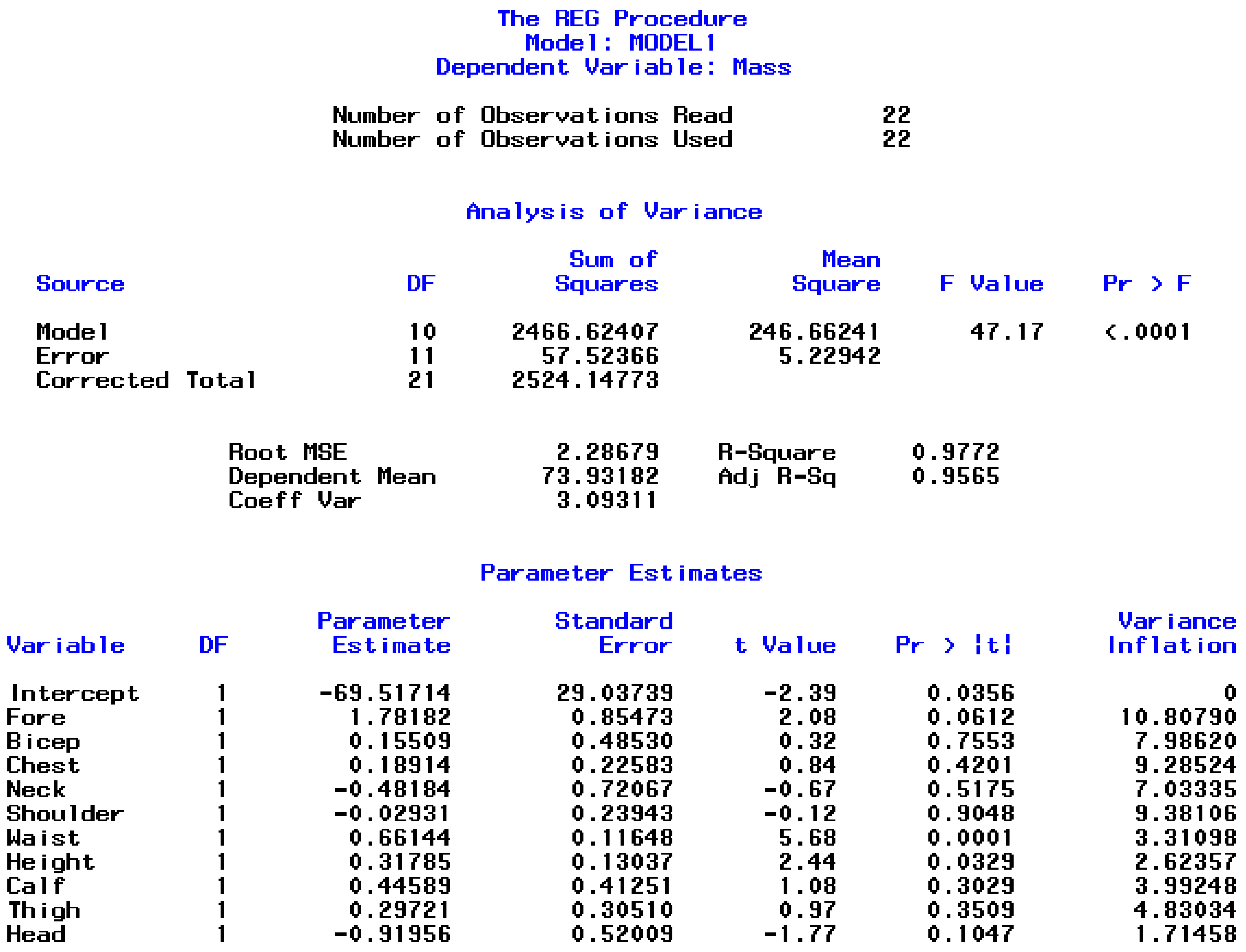
- The best regression models have:
  - Relatively few predictors.
  - A relatively high R^2.
  - A relatively small *s*, the standard deviation of the residuals.
  - Relatively small *P*-values for their *F*- and *t*-statistics.
  - No cases with extraordinarily high leverage.
  - No cases with extraordinarily large residuals;.
  - Predictors that are reliably measured and relatively unrelated to each other.

Collinearity

- When two or more predictors are linearly related, they are said to be **collinear**. The general problem of predictors with close (but perhaps not perfect) linear relationships is called the problem of **collinearity**.
- There is an easy way to assess collinearity. To measure how much one predictor is linearly related to the others, just find the regression of that predictor on the others and look at the R^2.
- That R^2 gives the fraction of the variability of the predictor in question that is accounted for by the other predictors.

Collinearity

- When a predictor is collinear with the other predictors in the model, two things can happen:
  - Its coefficient can be surprising, taking on an unanticipated sign or being unexpectedly large or small.
  - The standard error of its coefficient can be large, leading to a smaller *t*-statistic and correspondingly large P-value.
- If you have a collinear regression model, the simplest cure is to remove some of the predictors. Keep the predictors that are most reliably measured, least expensive to find, or even those that are politically important.



```
                           The REG Procedure
                             Model: MODEL1
                       Dependent Variable: Mass

               Number of Observations Read          22
               Number of Observations Used          22


                         Analysis of Variance

                                    Sum of           Mean
 Source                   DF       Squares         Square    F Value    Pr > F

 Model                    10    2466.62407      246.66241      47.17    <.0001
 Error                    11      57.52366        5.22942
 Corrected Total          21    2524.14773


           Root MSE              2.28679    R-Square     0.9772
           Dependent Mean       73.93182    Adj R-Sq     0.9565
           Coeff Var             3.09311


                         Parameter Estimates

                   Parameter       Standard                              Variance
Variable    DF      Estimate          Error    t Value    Pr > |t|      Inflation

 Intercept   1     -69.51714       29.03739      -2.39      0.0356              0
Fore         1       1.78182        0.85473       2.08      0.0612       10.80790
Bicep        1       0.15509        0.48530       0.32      0.7553        7.98620
Chest        1       0.18914        0.22583       0.84      0.4201        9.28524
Neck         1      -0.48184        0.72067      -0.67      0.5175        7.03335
Shoulder     1      -0.02931        0.23943      -0.12      0.9048        9.38106
Waist        1       0.66144        0.11648       5.68      0.0001        3.31098
Height       1       0.31785        0.13037       2.44      0.0329        2.62357
Calf         1       0.44589        0.41251       1.08      0.3029        3.99248
Thigh        1       0.29721        0.30510       0.97      0.3509        4.83034
Head         1      -0.91956        0.52009      -1.77      0.1047        1.71458
```