simon-5502-02-slides

Topics to be covered

- Add a topic next year on leverage, studentized residuals, Cook’s distance
- What you will learn
  - Three models for predicting a continuous outcome
  - Analysis of variance table
  - Variable selection
  - Stepwise regression
  - Residual analysis
  - Collinearity
  - Mediation

Three models

- All use a continuous dependent (outcome) variables
- All include multiple independent variables
- Multiple linear regression (Week 02)
  - All independent variables are continuous
- Analysis of covariance (Week 03)
  - Mix of continuous and categorical independent variables
- Multi-factor analysis of variance (Week 04)
  - All independent variables are categorical

Speaker notes

Over the next three weeks, you will learn about three different models for predicting a continuous outcome. All of these models will incorporate multiple independent variables.

Why three models?

- Historical precedents
- Different issues
  - Multicollinearity
  - Mediator variable
  - Risk adjustment
  - Moderator variable
  - Interactions

Speaker notes

If it seems arbitrary to have such different names, you are right. Some of this is historical. If you invent something, you get the right to name it, and different inventors for multiple linear regression, analysis of covariance, and multifactor analysis of variance just decided to use different names.

There are also issues that arise far more often in one model than the other. Multicollinearity, for example, is reserved primarily for multiple linear regression. You'll hear discussion about multicollinearity this week as well as mediator variables. Risk adjustment is a topic for next week under analysis of covariance, and you will see discussion of moderator variables and interactions two weeks from now under multifactor analysis of variance. In theory, these issues are not exclusive to one model or another. It's just easier to talk about them within a particular context.

The general linear model

- Single model that unites all three models.
- Use of indicator variables for categorical data
- Not the same as general **IZED** linear model
  - SAS: proc glm versus proc glim
  - R: lm() versus glm()

Speaker notes

It didn’t take long for researchers to discover a common linkage between multiple linear regression, analysis of covariance, and multifactor analysis of variance. In particular, you can treat categorical variables as if they were continuous through the use of indicator variables. You’ll see more discussion about this later.

One bad thing about statistics are all the terms that sound almost the same but mean something quite different. You may have noticed “moderator” and “mediator” from an earlier slide. I can never remember which is which. You’ll also learn about sensitivity and specificity in week 7 (Diagnostic testing) and these are also so easily confused.

But the worst, by far, are the general linear model and the generalized linear model. The IZED added to the end of “general” makes it an entirely different model.

You may see a bit about the generalized linear model in week 7 (logistic regression).

In any case, I have really preferred the use of a single model, the general linear model, in place of multiple linear regression, analysis of covariance, and multifactor analysis of variance. I am a “lumper” and not a “splitter”. For what it is worth, SPSS forces you to use the individual approaches at times instead of the single general linear model.

Arguments for the lm() function

- formula = $dependent - variable$ ~ $independent - variables$
  - $independent - variables$ can be numeric, factors, or strings
- data =
- subset =
- na.action =
  - na.fail
  - na.omit
  - na.exclude
- other arguments

Speaker notes

This week I will show how you fit a multiple linear regression model using the lm function. In the following two weeks, I will show how to use the lm function to compute analysis of covariance and multi-factor analysis of variance.

Live demo #1, ANOVA and R-squared

Refer to the demonstration program for module02.

Break #1

- What you have learned
  - Three models for predicting a continuous outcome
- What’s coming next
  - Analysis of variance table

Data for multiple linear regression

- Dependent variable: $Y_1, Y_2, \ldots, Y_n$
- First independent variable: $X_{11}, X_{12}, \ldots, X_{1n}$
- Second independent variable: $X_{21}, X_{22}, \ldots, X_{2n}$
- ...
- kth independent variable: $X_{k1}, X_{k2}, \ldots, X_{kn}$

Speaker notes

For a multiple linear regression model, you need a dependent or outcome variable, traditionally designated by the letter “Y”. The are n values of Y, one for each patient in your research study.

Then you have k independent variables. The first one, designated by X_1 also has n values. The second independent variable, X_2 works similarly. There are k total independent variables.

For now, let’s assume that every variable, the one dependent variable and the k independent variables, is continuous.

Matrix notation, 1 of 3

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_j \\ \vdots \\ Y_n \end{bmatrix}$$

Speaker notes

Often you will see multiple linear regression defined with a single column for the dependent variable Y …

Matrix notation, 2 of 3

$$X = \begin{bmatrix} X_{11} & X_{21} & \dots & X_{i1} & \dots & X_{k1} \\ X_{12} & X_{22} & \dots & X_{i2} & \dots & X_{k2} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{1j} & X_{2j} & \dots & X_{ij} & \dots & X_{kj} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{1n} & X_{2n} & \dots & X_{in} & \dots & X_{kn} \end{bmatrix}$$

Speaker notes

… and a matrix representing the independent variables. Each column in the matrix represents values of a particular independent variable for every subject and each row represents all of the independent variables for a particular subject.

Matrix notation, 3 of 3

$$X = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{i1} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{i2} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ 1 & X_{1j} & X_{2j} & \dots & X_{ij} & \dots & X_{kj} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{in} & \dots & X_{kn} \end{bmatrix}$$

Speaker notes

You may see a similar matrix, but with an extra column at the beginning with the value of 1 repeated down that column.

There are lots of theoretical reasons why the dependent variable and the independent variables are laid out in this format. I will not talk about this in detail. I just wanted to show you this, so you would not be alarmed if you see it in some of the future readings you might have on multiple linear regression.

Why use multiple linear regression

- Two competing purposes
  - Mechanisms
    - What variables have an impact on your outcome?
  - Prediction
    - What outcome will you see on tomorrow's patient?
    - "It is difficult to make predictions, especially about the future." Niels Bohr

Speaker notes

It is important to note that there are two competing reasons for using multiple linear regression (and other statistical models as well). The first is to try to understand the underlying mechanisms that explain the “how”. You are looking to understand how a disease harms or how a treatment helps. This could be called a “white box” feature that you get with multiple linear regression. The size and signs of the regression coefficients help you understand what factors contribute to the “how”.

You might think of a mechanistic goal as being able to peek under the hood. Some people are perfectly happy driving a car without having any idea of how the engine works. This is fine, and you are treating the automobile as a black box. You don’t know what goes on underneath the hood of the car. Other people want to understand how a car works and they believe (properly in my opinion) that this understanding makes them a better driver.

Alternatively, you can use a multiple linear regression model to predict the future. You want to tell a patient about his or her prognosis, you want to estimate the staffing needs in a hospital, or you want to estimate the cost savings associated changes in health care delivery.

Predicting the future is a perilous process. Pretty much every investment opportunity carries a warning “Past performance is no guarantee of future results

Multiple linear regression model

- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i$
  - Assumptions: ϵ_i are
    - independent,
    - same variance,
    - normally distributed.
  - Independent variables must be on interval or ratio scale
    - Nominal/ordinal scales require some care

Speaker notes

The multiple linear regression model says that Y_i is a linear function of X_{1i}, X_{2i}, …, X_{ki} plus an error term. The error term accounts for the fact that even the best set of independent variables will still fall a bit short.

The ϵ_i have to meet three assumptions: independence, equal variances, and normally distributed.

It's often not stated directly, but your independent variables must be on an interval or ratio scale. There are ways accommodate independent variables on an ordinal or interval scale, but these methods require a bit of care.

Estimates for the multiple linear regression model

- Use least squares to estimate $\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \ldots,\ \hat{\beta}_k$
  - $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki}$
  - $e_i = Y_i - \hat{Y}_i$

Speaker notes

To estimate the β values, use the principle of least squares. In other words find the β values that make the predicted values, $\hat{Y}_i$ as close as possible to the original Y_i.

The residual, e_i, is the deviation between the actual data value and what you predict based on the k independent variables.

Sum of squares

- “Bad” regression results.
  - All $\hat{Y}_i$ close to each other
  - Few $\hat{Y}_i$ close to Y_i
- “Good” regression results.
  - The $\hat{Y}_i$ are spread out
  - Many $\hat{Y}_i$ close to Y_i

Speaker notes

Let’s talk about a bad regression result. The word “bad” is a bit judgmental. You might substitute “disappointing” but that is also troublesome. Maybe “not too useful” is what I really mean.

Let me give a hypothetical example. Suppose you paid a statistician a fortune to produce a linear regression model that predicted the average length of stay for heart attack patients of a particular age with particular systolic and diastolic blood pressure.

An 84 year old patient with 130/90 blood pressure? The model predicts 2.7 days length of stay.

A 35 year old patient with 120/80 blood pressure? The model predicts 2.4 days length of stay.

A 69 year old patient with 190/140 blood pressure? The model predicts 2.6 days length of stay.

You ask yourself, why are we bothering collecting information about age and blood pressure. The predictions are pretty much all the same.

Another way of looking at this is that a “bad” or “not too useful” regression model is one where the predictions are not very close to the actual data values.

In contrast a “good” or “useful” regression model is one where the predictions change a lot for patients with different characteristics. A wide spread of the predicted values means that you have information that can help you. The first patient’s predicted length of stay is 6.8, the second is 1.5 and the third is 3.1? That’s useful information.

Another way of thinking about a “good” regression model is that it’s predictions are generally pretty close to the actual values.

The analysis of variance table helps you identify “good” and “bad” regression models.

Analysis of variance table for multiple linear regression, 1 of 3

- SSR or $SS_{regression} = \Sigma(\hat{Y}_i - \bar{Y})^2$
- SSE or SS_{error} or $SS_{residual} = \Sigma(Y_i - \hat{Y}_i)^2 = \Sigma\, e_i^2$
- SST or $SS_{total} = \Sigma(Y_i - \bar{Y})^2$

Speaker notes

The first column in the analysis of variance table is the sum of squares. The regression sum of squares is a measure of how spread out the predicted values ($\hat{Y}_i$) are.The error sum of squares is a measure of how close the predicted values are to the data. The total sum of squares is the variation of the data itself.

Analysis of variance table for multiple linear regression, 2 of 3

- $df_{regression} = k$
- $df_{error} = n - k - 1$
- $df_{total} = n - 1$
  - $MS = SS/df$

Speaker notes

The degrees of freedom for regression is k, the number of independent variables. The degrees of freedom for error and total are n-k-1 and n-1 respectively. You divide the sum of squares by the degrees of freedom to get the mean squares.

Analysis of variance table for multiple linear regression, 3 of 3

- $F = MSR/MSE$
- This tests the hypotheses
  - $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$
  - $H_1 : \beta_j \neq 0$ for at least one j
  - Accept H_0 is F is close to 1
  - Reject H_0 is F is much larger than 1

Speaker notes

You compute the F ratio by dividing the mean square for regression into the mean square for error. Think of the numerator as signal and the denominator as noise. If this ratio is close to 1, then you should accept the null hypothesis.

Example using fat data

```
Analysis of Variance Table

Model 1: fat_b ~ 1
Model 2: fat_b ~ chest + abdomen + hip
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    251 15079.0
2    248  4530.5  3     10548 192.47 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- SSR = 10,548
- SSE = 4,530.5
- SST = 15,079
- F = 192.47
- You should report the p-value as $p < 0.001$

Speaker notes

Here is an example using the dataset with body fat measurements and circumference measurements at the hip, chest, and abdomen. The sum of squares are 10,500 for regression, 4,500 for residual or error and 15,000 for total. The respective degrees of freedom are 3 (for the three independent variables), 248 (252-3-1) for error and 251 (252-1) for total. The mean squares are quite different 3,500 versus 18. The F-ratio, 192, is a lot larger than 1 and the small p-value indicates that you should reject the null hypothesis and conclude that at least one of the three regression slopes is not equal to zero.

Note that this test does not tell you which variable or variables are needed to predict body fat. It is an overall test of the combined effect of all three variables.

R-squared

- $R^2 = SSR/SST$ or $1 - (SSE/SST)$
  - Proportion of explained variation
  - 1 - proportion of unexplained variation

Speaker notes

R-squared is defined pretty much the same way as in simple linear regression (regression with only one independent variable). It is the ratio of explained variation ($SS_{regression}$) to total variation (SS_{total}) or 1 minus the ratio of unexplained variation (SS_{error}) to total variation. This is a descriptive measure and there is no rule of thumb, but you are happier with the quality of the regression predictions if R-squared is closer to 1.

Example using the fat data

```
  r.squared adj.r.squared    sigma statistic      p.value df    logLik      AIC
1 0.6995469     0.6959124 4.274143  192.4734 1.853251e-64  3 -721.6075 1453.215
       BIC deviance df.residual nobs
1 1470.862 4530.537         248  252
```

- R^2 = 10,548 / 15,079 = 0.6995469
  - Alternately R^2 = 1- (4,530.5 / 15,079)
- Round and express as a percentage
  - R^2 = 70%

Speaker notes

You can compute R-squared from the analysis of variance table. The sum of squares for regression is 10,548.480 and the sum of squares total is 15,079.017. When you are calculating, please use the maximum precision for any intermediate calculation, but then round aggressively with the final result.

Often you will see R-squared displayed as a percentage instead of a proportion (e.g., 70% instead of 0.70).

Adjusted R^2

- $1 - \frac{MSE}{MST}$ or
- $1 - \frac{SSE/(n-k)}{SST/(n-1)}$
  - Field textbook suggests a more complex formula
  - Penalizes for model complexity (but not enough)

Speaker notes

Many researchers propose an alternative measure, adjusted R-squared. This statistic modifies R-squared slightly downward. You can view this as a penalty for model complexity, but unfortunately, the penalty is small and does help enough to discourage the use of needlessly complex regression models.

Live demo, part 1

Live demo #2, ANOVA and R-squared

Break #2

- What you have learned
  - Analysis of variance table
- What’s coming next
  - Variable selection

Avoid needlessly complex regression models

- “Everything should be made as simple as possible, but not simpler” Albert Einstein (?)
- “If you have two competing ideas to explain the same phenomenon, you should prefer the simpler one.” Occam’s Razor
- “The parsimony principle for a statistical model states that: a simpler model with fewer parameters is favored over more complex models with more parameters, provided the models fit the data similarly well.” - ClubVITA

Speaker notes

Here are a series of quotes that all advise against needlessly complex models. There is some empirical evidence that complex models do not extrapolate well, but these exhortations for simplicity are largely based on subjective opinions. Even so, this is an attitude that I endorse heartily (with one reservation).

Choosing independent variables is a balancing act

Speaker notes

I have heard many times that Statistics is as much of an art as it is of a science, and I largely agree with this perspective. The selection of independent variables in a multiple linear regression model is an example of the artistic side of statistics.

It is a balancing act. You want a model that is as simple as possible, but not so simple as to oversimplify. This is a very subjective criteria, but it should not be ignored.

Selecting too many independent variables is definitely bad. It can be expensive. Looking at the data on body fat measurements, do you really need to measure circumferences at 10 different parts of the body to get a good handle on how much body fat someone has? Maybe you could do almost as well with only measurements at 3 locations.

The other problem with overly complex models is that they tend to do a poor job of replication. You might do really well with your own dataset, but when others collect similar data, the overly complex models generally don't look quite as good. You may even find that new data collected at your own location does work so well with a really complex model.

A model that is too simple might not explain enough of the variation in the data. It may also miss out on important features.

Deciding where to go between the simplest models and the most complex models is a tricky choice. It may depend on the goals of your research. If your goal is to try to understand the mechanisms behind a disease or behind a health care delivery process, you might gravitate towards the simpler models. Including too many variables just makes it harder to understand what is going on "under the hood". Too many variables is less of a problem if your main goal is prediction. You can tolerate a few extra unneeded variables. The complexity doesn't hurt you as much if you are only interested in prediction.

I'll return to this topic in the next section.

Counterpoint on complexity

- Machine learning algorithms
- Risk adjustment

Speaker notes

There are many who advocate that machine learning algorithms, which automatically choose among some very complex models, are preferred in many settings. In particular, if you are interested in prediction but not mechanisms, then complexity can be your friend, as long as you are careful about this.

A second area where a more complex model is called for is in risk adjustment. If you have an observational study and you want to control for confounding, it is often best to adjust for every medically plausible variable that could influence your outcome. A simple model may lead to residual confounding, a lingering bias that remains after you try to adjust with a simple model.

Rule of 15

- Developed by Frank Harrell in a different context
  - Ratio of observations to independent variables
    - $n/k > 15$
    - Some use 10 instead of 15
  - Smaller ratios imply poor replicability
- Not a replacement for a power calculation
- Some researchers have argued against this rule

Speaker notes

A commonly cited rule about model complexity is the rule of 15. It is a reminder that the more complex your statistical model, the more data you need.

This rule was originally developed by Frank Harrell in a slightly different context, logistic regression. It is commonly applied in a multiple linear regression context, and it seems reasonable enough.

The rule is that the number of observations in your dataset need to be a lot larger than the number of independent variables. Fifteen times larger, as a matter of fact. So if you have 20 independent variables, then you need at least 300 observations to insure a ratio of 15 to 1. If you have 6 independent variables, you need at least 90 observations.

Now there is nothing stopping you from running a regression model with fewer observations. If you have 50 observations and you want to run a model with 10 independent variables (a 5 to 1 ratio), there is nothing to stop you.

The rule of 15 is a rule of thumb and no one ever got thrown in jail for violating a rule of thumb.

The concern with running a really complex model with a small number of observations is that these models do not replicate well.

Now some researchers argue that a 10 to 1 ratio of observations to independent variables is fine.

Other researchers have argued that this rule of thumb is not valid and that it is okay to run complex models on small datasets. This is a minority viewpoint, however, and running a 10 variable model on just 50 observations is likely to get you in trouble.

An important point to remember here is that k is the number of candidate variables that are considered at some time during the data analysis. It is NOT the number of variables in the final model.

Testing the impact of chest and hip controlling for abdomen

```
Analysis of Variance Table

Model 1: fat_b ~ abdomen
Model 2: fat_b ~ chest + abdomen + hip
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    250 5094.9
2    248 4530.5  2    564.39 15.447 4.758e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Speaker notes

What is the impact of chest circumference in a model that already includes abdomen and hip circumference? This table shows the regression coefficients and the standard errors for a multiple linear regression model with all three variables. The regression coefficient for chest is -0.186 and the standard error is 0.081. The test statistic is -2.304 which is large negative and the p-value is small. You should reject H0 and conclude that there is a negative relationship between chest circumference and body fat, after holding hip and abdomen circumference constant.

Change in R-squared

- Partial R^2 = 0.6995469 - 0.6621178 = 0.0374291

Speaker notes

The regression model with chest, abdomen, and hip accounts for 70% of the variation, but a model with just abdomen does almost as well, accounting for 63% of the variation. The difference, 7% is the amount of additional variation accounted for when you add chest to a model that already included abdomen and hip.

That seems to contradict the earlier finding with the large negative test statistic and the small p-value. But actually, what it is saying is that although there is sufficient statistical evidence to conclude that chest circumference has a real impact, that impact is small. You will see this a lot, especially with datasets with very large sample sizes. You can often achieve statistical significance for an individual variable, but the practical impact can still be negligible.

Live demo, part 2

Break #3

- What you have learned
  - Variable selection
- What’s coming next
  - Stepwise regression

What is a good model

- “Everything but the kitchen sink” model
- Stepwise models
- Hierarchical models
- “Maximimum adjusted R-squared” model
- “Out of sample error” model
- “Use your brain” model
- Some combination of the above?

Speaker notes

Finding a good regression model is not easy. There are several criteria that have been used.

“Everything but the kitchen sink” model

- Include anything that seems remotely plausible
- Advantages
  - Simple computationally
- Disadvantages
  - Does not identify mechanisms
  - Can be expensive
  - Fails if k is very large

Speaker notes

“Everything but the kitchen sink” is an American idiom for “everything imaginable”. In this approach, you brainstorm a list of variables that might be associated with your outcome. Use your knowledge of past research in the area, your understanding of the medical and scientific processes involved, and just your plain old intuition. Anything that seems remotely possible to be related to the outcome variable is included in the regression model.

Kitchen sink model for body fat

```
Analysis of Variance Table

Model 1: fat_b ~ 1
Model 2: fat_b ~ neck + chest + abdomen + hip + thigh + knee + ankle +
    biceps + forearm + wrist
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    251 15079.0
2    241  3994.7 10     11084 66.871 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Speaker notes

Here’s a model with all the circumference measurements included.

Stepwise model, 1 of 2

- Forward selection
- Backward elimination
- Composite of both

Speaker notes

A stepwise model uses p-values to identify which variables to add to the model (forward selection) or which variables to remove from the model (backward elimination). You can combine these two approaches.

Forward selection starts with nothing. It looks at the p-values for every single variable model and chooses the variable with the smallest p-value. Then it looks at every possible two variable model that includes the first selected variable and chooses the second variable based on the smallest p-value. Then it looks at every possible third variable in a model that includes the first two variables. And so forth.

Backward elimination starts with everything. It looks at all the p-values for the k independent variables in the model and eliminates the variable with the largest p-value. Then it looks at all the p-values in the model with k-1 independent variables and removes the independent variable with the smallest p-value.

Both approaches take a “one step at a time” approach. You might be tempted in the backward elimination step to toss out every independent variable with a large p-value. But the p-values do change as you start removing variables. One of your independent variables might have a large p-value when you have all k variables in the model, but that p-value might shrink as other variables are tossed out.

A general approach is to start with a forward selection approach, but each time a new variable is added, you check to see if one of the variables which had a promising p-value earlier now looks not so hot. So each forward step is followed with a backward elimination check.

Stepwise model, 2 of 2

- Advantages
  - Automated in SPSS and other packages
- Disadvantages
  - Does not incorporate medical knowledge
  - Does not control Type I error rate

Speaker notes

The forward selection and backward elimination approaches are easy to program and are available in SPSS and most other statistical software packages.

Stepwise regression, however, does not incorporate medical knowledge into the process. You might have two variables that both have large p-values in a backward elimination approach. One is well known to be important by anyone with a medical background, but if it’s p-value is a bit larger than the second one that is a bit of a stretch, you will make the wrong choice, at least from a medical perspective.

The other criticism of stepwise models is that they do not control the Type I error rate. You use repeated hypothesis tests and the chances of making a Type I error and including a variable that does not really belong gets inflated. The number of tests that stepwise regression requires is very large, even if you only have a have only a handful of variables.

Hierarchical models

- Sometimes variables fall into natural groups.
  - Demographic features
  - Patient frailty
  - Environmental factors
  - Current treatments
- Enter data in blocks

Speaker notes

Be careful with the term “hierarchical” as it means many things. In the context of multiple linear regression, it means that variables fall into natural groups or blocks. Some examples are variables that represent demographic features (such as age, race, and gender) versus variables that represent patient frailty (stage of cancer, severity of illness, previous history of health issues) versus environmental factors (exposure to allergens, second hand smoke) versus current treatments (number of medications, dosages).

In a hierarchical multiple linear regression model, you enter all variables in a particular group together at once and then add a second group, third group, etc.

“Maximimum adjusted R-squared” model

- Fit all possible models ($= 2^k$)
  - Select model with largest adjusted R-squared
  - Alternative criteria
    - Mallows Cp
    - AIC, AICc
    - LASSO

Speaker notes

If the number of independent variables is not too big, you can look at every possible combination of independent variables. For the body fat example, there are ten measures and this means examining 1,024 different models. Not easy for you maybe, but your computer can do this in the blink of an eye.

Be careful, though. You may notice that 2^k, the number of models, increases (literally) exponentially. With 20 independent variables, you would need to fit over a million models. With 30, it would be over a billion.

You cannot choose the model that maximizes R-squared. R-squared never decreases as the model becomes more complex. It may only increase by a trivial amount, but because it never decreases. Choosing the model with the largest R-squared value will always select the most complex model, the one with every independent variable.

Adjusted R-squared can go down, because it adjusts for the degrees of freedom. In general, though, adjusted R-squared doesn't do enough adjusting and will tend to select needlessly complex models. There are other criteria: Mallows Cp, Akaike Information Criteria (AIC), a corrected version of AIC (AICc), and LASSO. These all do a better job of avoiding overly complex models.

“Out of sample error” model

- Split data into training/test sets
  - Use training data to build the model
  - Use test dataset to evaluate fit

“Use your brain” model

- Previous research
- Knowledge of medicine/science
- Superior to “brainless” approaches
  - “Ginny!” said Mr. Weasley, flabbergasted. “Haven’t I taught you anything? What have I always told you? Never trust anything that can think for itself if you can’t see where it keeps its brain?” from Harry Potter and the Chamber of Secrets, J.K. Rowling.

Speaker notes

There is a tendency to forget that you yourself have a lot to contribute to the process of selecting the right independent variables. You are familiar with previous research and you have insights into the medicine and science that underlies the model you are trying to build.

Relying solely on automated procedures is often a mistake because the automated procedures do not use anything beyond the numbers themselves.

If there is a problem with these approaches and more generally with many machine learning models, it is highlighted by the quote from J.K. Rowling. She was not thinking about machine learning when she crafted this quote, but it certainly applies.

Use a mixture

- Use a mixture of science/medicine with automated approaches?
  - Story about an industrial process

Speaker notes

At a minimum, use your knowledge as you evaluate these models. Avoid the tendency to think of a stepwise approach or any other automated approach as the final word. Think of it as an intelligent assistant, who might save you time, but you still have to carefully check everything they produce.

There’s an interesting anecdote that I heard many years ago (so long ago that I can’t remember the source). A researcher was examining an industrial process and was rating the importance of various factors in controlling the output. The statistician ranked all the variables using some sort of stepwise algorithm. They listed the most important variable, second most important, etc. and conclude with “and the least important variable is the amount of water present.”

At the point the whole audience erupted in laughter. They knew that water was very important because even a small amount of water in the system would cause a terrible explosion. In fact, the factory made every attempt to tightly control the amount of water. This control led to a very small standard deviation and very little power to detect a significant effect. The moral is to never blindly trust a model built solely on the data with no consideration of previous research and the underlying scientific and medical mechanisms.

Break #4

- What you have learned
  - Stepwise regression
- What’s coming next
  - Residual analysis

Using residuals to check for assumption violations

- Non-normality
  - QQ plot
- Lack of independence
  - Time sequence plot, Durbin-Watson statistic
    - Only for time-ordered data
- Unequal variances
  - Sactterplot of residuals versus predicted values
- Non-linearity
  - Scatterplot of residuals versus each independent variable
  - Scatterplot of residuals versus predicted values

Speaker notes

There are three important assumptions in multiple linear regression. You check the normality assumption with a Q-Q plot of the residuals. Nte that it is the residuals and not the dependent variable itself that you use here.

If your data has a natural ordering over time, then a time sequence plot and the Durbin-Watson statistic can help detect some violations of the independence assumption. You often do not have a time ordering for your data, and there may be other types of problems with independence associated with factors other than time.

The assumption of equal variances is sometimes violated when larger predicted values have greater variation. This occurs in quite a few setting.

Consider a mutliple linear regression model with housing prices as the dependent variable. You go to some older neighborhoods in the Kansas City area and the houses are mostly smaller and packed closely together on tiny lots. These houses might vary in price from 80 to 120 thousand dollars. Other neighborhoods have houses that are not so old, a bit larger, and a bit more space for your yard. These houses might sell across a broader range, maybe from 200 thousand to 400 thousand dollars. Then you have the neighborhoods with newer houses, lots of room both inside and outside. These houses might sell for 800 thousand dollars to 2 million dollars.

Notice how the variation marches in lock step with the prices. Small average prices have relatively little variation and large average prices have relatively large variation.

This also occurs with many other outcome variables that are bounded below by zero. The bound is quite constraining when you are on the low end, but less constraining on the high end.

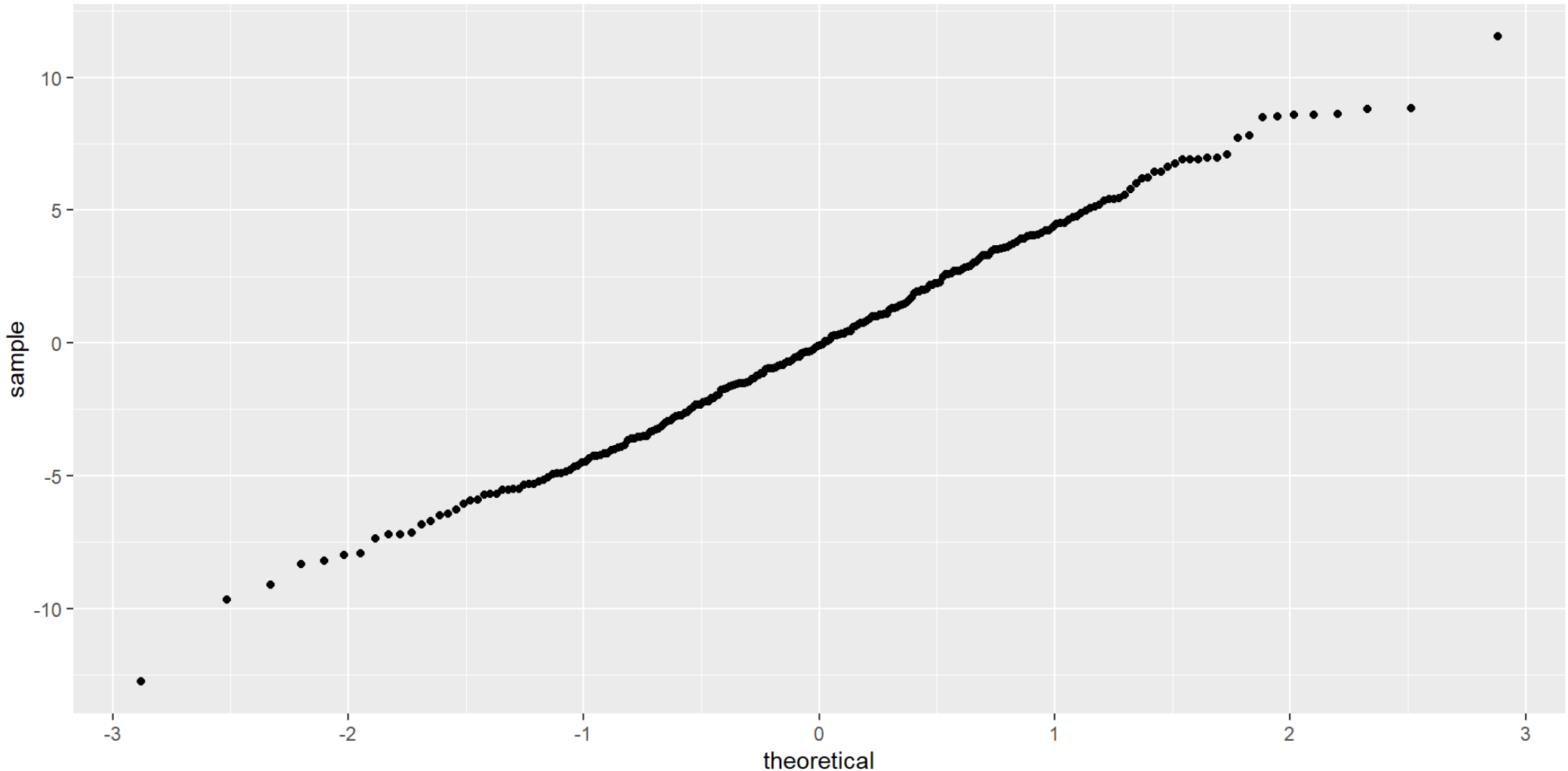
The way to look for the pattern is to plot the residuals versus the predicted values. A fanning out pattern, with more variation with residuals associated with larger predicted values, is a violation of the assumption of equal variances.

Linearity is an implicit assumption of multiple linear regression. Sometimes, however, the relationship is nonlinear between one of your independent variables and your dependent or outcome variable variable. You might, for example, see an increasing relationship, but the increase tends to slow down at the high end. This is sort of like diminishing returns. Or perhaps the opposite occurs, when your independent variable is on the high end, the outcome takes off almost exponentially. Sometimes you can have too much of a good thing. A moderate value for your independent variable leads to a large outcome, but larger values backfire and bring your outcome back down to earth.

A plot of the residuals versus each independent variable is helpful here. A random scatter of points is a sign that the linearity holds. Any trend indicates that there is a more complex relationship between your independent variable and your outcome.

If you have more than a few independent variables, you can substitute a single plot the plot of residuals versus the predicted value. Think of the predicted value as a composite or linear combination of all your independent variables. If there is no obvious pattern or trend in this plot, then there is a good chance that you wouldn’t see any obvious pattern or trend in a plot of any of your independent variables and your residuals.

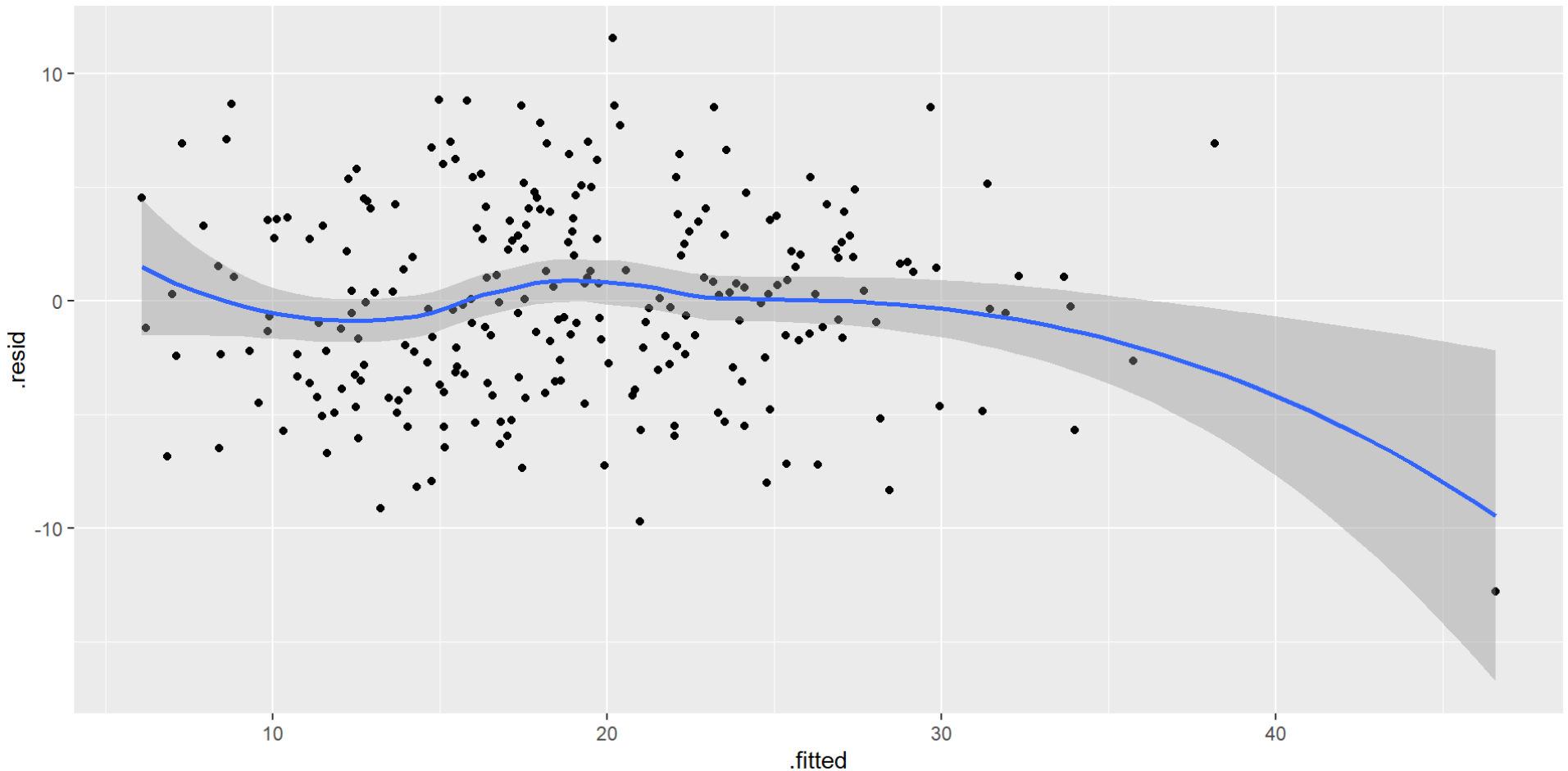
Q-Q plot of residuals



Speaker notes

This is the QQ plot of residuals. It looks like a straight line, indicating that the normality assumption is reasonable.

Scatterplot of residuals and predicted values



Speaker notes

While three variables is not too many, here is what the plot of residuals versus predicted values looks like. Notice that it has pretty much the same features as the three individual plots.

Measures of influence

- Same as in biostats-1
  - Leverage > 3(p+1)/n
  - Studentized residual > +/-3
  - Cook’s distance > 1

Live demo, part 3

Live demo #3, residual plots

Break #5

- What you have learned
  - Residual analysis
- What’s coming next
  - Collinearity

What is collinearity?

- Strong interrelationship among the independent variables
  - Also known as
    - multi-collinearity
    - near collinearity
    - ill-conditioning
- Interrelationship could be just two variables
  - Also could be three or more interrelated variables

Speaker notes

Collinearity is a strong interrelationship among two or more independent variables. There are alternate terms, collinearity, near collinearity, and ill-conditioning. People with a more theoretical background will quibble about these terms but you should consider them more or less interchageable.

The interrelationship could just be between two independent variables. These are relatively easy to spot. Sometimes the interrelationship is among three or more independent variables. These can sometimes be tricky to spot.

Examples of collinearity

- Birthweight and gestational age predicting length of stay
- Size of the home and size of the lot predicting sales price
- Calories from fat, from protein, and from carbohydrates predicting weight gain

Speaker notes

If you have to predict length of stay in the birth hospital using birthweight and gestational age, there is a relationship between the two variables. Lower birthweights are associated with earlier gestational age and higher birthweights are associated with later gestational ages.

If you are trying to estimate what a house will sell for, the size of the house itself and the size of lot it sits on are both important. Larger houses tend to be found on larger lots and smaller houses tend to be found on smaller lots. There are some exceptions, of course, but the interrelationship between house size and lot size is strong.

A more complex relationship involves calories consumed from three sources: fat, protein, and carbohydrates. If you get a lot of calories from fat, there is less room in your diet for calories from protein and carbohydrates. You can say a similar thing for calories from protein or calories from carbohydrates. Now some people will just eat a lot of everything, but most people gravitate to one source of calories or another.

Now none of these interrelationships are perfect. Sometimes you will see a slightly underweight baby in a full term birth, for example. But there is a general relationship between the two in spite of a few exceptional babies.

What problems are caused by collinearity?

- Difficulty in variable selection
- Loss of precision
  - wider confidence intervals
- Loss of power
  - Need for larger sample sizes
- Not a violation of assumptions
- Not a problem if you are only interested in prediction

Speaker notes

When two or more independent variables are interrelated, it becomes difficult to decide where one variable or the other or both are needed to provide good predictions.

You will see a loss in precision. Your confidence intervals will be wider.

You also have less power. The sample size needed will have to be larger.

Think of a table with four legs. Normally you would put the legs in all four corners. This provides stability. But if the legs go along the diagonal fron one corner to the opposite corner, you don’t have good support in the remaining two corners. The table wobbles and is unstable.

Now the presence of multicollinearity does not make a multiple linear regression model invalid. You’re not violating any of the assumptions the you need. Instead think of collinearity as a data insufficiency problem. You have data on small babies with early gestational age, large babies with later gestational age, but almost no data on large babies with early gestational age and almost no data on small babies with late gestational age.

Fixing collinearity

- Collect more data
- Oversample “rare” corners
- Prune your variables

Measures of collinearity

- Correlation matrix
- Tolerance
  - $Tol_i = 1 - R_i^2$
    - R_i^2 for predicting i^{th} independent variable from remaining independent variables
- Variance inflation factor
  - $VIF_i = \frac{1}{Tol_i}$
  - Increase in $Var(\hat{\beta}_i)$ due to collinearity

Speaker notes

A simple way to check for multicollinearity is to examine the correlations among the independent variables. This will catch some of the simpler forms of multicollinearity, especially when it is an interrelationship among only two independent variables.

Tolerance is defined as 1 minus the R-squared value in a model predicting the i^{th} independent variable using the remaining independent variables. A small value for at least one tolerance is an indication of trouble with collinearity. If you can account for most of the variation in one independent variable using all the other independent variables, that implies a strong interrelationship.

How small is small? I would say 0.1 or less, but others have suggested the 0.4 or smaller is an indication of collinearity.

The variance inflation factor is a comparable statistic. It is just the reciprocal of tolerance.

The variance inflation factor is the reciprocal of tolerance. It has a direct interpretation. It represents how much larger the estimated variance is for an estimated slope in multiple linear regression due to collinearity compared to the ideal case where there would be no correlation between any of the independent variables.

I view a variance inflation factor of 10 or greater to be worthy of concern.

Collinearity statistics for the fat dataset

```
   chest  abdomen      hip
6.340574 8.383840 4.332385
```

What is perfect collinearity?

- Exact relationship among independent variables
- Impossible to estimate regression coefficients
- Examples
  - Measuring temperature in both Fahrenheit and Centigrade
  - Thee percentages adding up to exactly 100%
- Only solution: drop one or more variables

Speaker notes

Although most collinearities represent only approximate interrelationships, sometimes you might encounter a perfect collinearity. If there is a precise relationship rather than an approximate one, then you cannot estimate the regression coefficients at all.

Some examples of perfect collinearities are when you have the same measurement, just with different units, such as temperature in Fahrenheit and temperature in Centigrade. Or three variables representing percentages of a total might be perfectly collinear if they total up to exactly 100%. This is a perfect collinearity because once you know two of the percentages, you can estimate the third one precisely.

The only solution to a perfect collinearity is to drop one or more independent variables. Keep temperature in Fahrenheit or Centigrade but not both. Drop one of the three percentages that add up to exactly 100%.

Live demo, part 4

Live demo #4, multicollinearity

Break #6

- What you have learned
  - Collinearity
- What’s coming next
  - Mediation

What is mediation?

Speaker notes

The image and quote from your Andy Field textbook is about as good as anything. A mediator is a third variable which partially or totally explains the relationship between two variables. You'll see more of this next week's module, but I wanted to introduce it here, because it is a very important concept.

Why do you need to understand mediation?

- Mostly irrelevant for prediction
- Very important for understanding mechanisms
  - Is there a direct relationship?

Speaker notes

An informal assessment of mediation

Model without the mediator

```
# A tibble: 2 × 5
  term        estimate std.error statistic       p.value
  <chr>          <dbl>     <dbl>     <dbl>         <dbl>
1 (Intercept)   11.0      1.74       6.31 0.00000000125
2 age            0.178    0.0372     4.78 0.00000304
```

Model with the mediator

```
# A tibble: 3 × 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept) -36.5       2.47      -14.8  6.23e-36
2 age           0.0660    0.0229      2.88 4.27e- 3
3 abdomen       0.567     0.0268     21.2  1.19e-57
```

Speaker notes

There’s a relationship between age and percentage body fat. It is a bit easier to follow if you convert age in years to age in decades. The relationship between decades of age and body fat is that you add about 1.8% to your body fat every decade on average.

The question is, how does this happen? Does some of the muscle turn directly into fat, or do you add fat through an expanding girth?

When you look at a model with both decade of age and abdomen circumference, the effect of age is cut markedly. Instead of adding 1.8% body fat on average, you add about 0.7% body fat on average. The rest of the body fat comes from the “love handles” that you develop as you age.

So gaining some fat is inevitable, even if you can still fit into those size 32 jeans that you wore as a young man. But much of the gain in percentage body fat comes from moving from size 32 over time to size 42.

Live demo, part 5

Live demo #5, mediation

Summary

- What you have learned
  - Three models for predicting a continuous outcome
  - Analysis of variance table
  - Variable selection
  - Stepwise regression
  - Residual analysis
  - Collinearity
  - Mediation