BIOS 6312: Modern Regression Analysis

Andrew J. Spieker, Ph.D.

Assistant Professor of Biostatistics
Vanderbilt University Medical Center

Set 2: Simple Linear Regression

Version: 01/01/2023

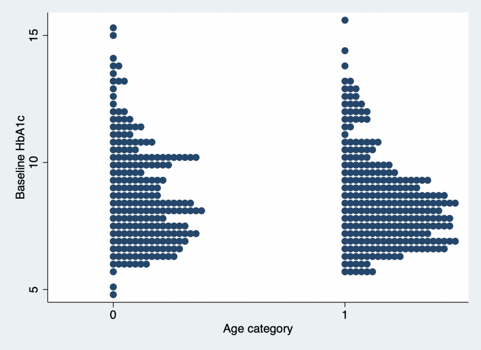
Table of Contents

1. Introduction to simple linear regression
2. Estimation by ordinary least squares
3. Subgroup-specific confidence intervals for mean outcomes
4. Prediction intervals
5. Diagnostic measures and tools

Simple linear regression

Re-framing: A simple linear model

- Suppose we seek to use the baseline data from the REACH study to learn about the association between age and mean HbA1c.
- Simple approach: categorize age (> 56 years) and conduct a t-test.



Simple linear regression

Re-framing: A simple linear model

- Now, we introduce some regression notation.
- Model: $\text{E}[Y|X=x] = \beta_0 + \beta_1 x$.
  - Recall: $\text{E}[Y|X=x]$ denotes mean (average) value of Y among the subgroup $X=x$.
- Note: $\text{E}[Y|X=0] = \beta_0 + \beta_1 \times 0 = \beta_0$.
  - Therefore, how do we interpret the "intercept," β_0?
- Further, note that $\text{E}[Y|X=1] = \beta_0 + \beta_1$.
- Therefore, it follows that:

$$\begin{aligned} \text{E}[Y|X=1] - \text{E}[Y|X=0] &= (\beta_0 + \beta_1) - (\beta_0) \\ &= \beta_1. \end{aligned}$$

- How do we interpret the "slope," β_1?

Simple linear regression

Re-framing: A simple linear model

- In the age-HbA1c example from REACH:
  - $\beta_0 = \mathrm{E}[Y|X = 0]$: mean HbA1c among younger group (≤ 56 years old).
  - $\beta_1 = \mathrm{E}[Y|X = 1] - \mathrm{E}[Y|X = 0]$: difference in mean HbA1c between age groups (specifically, older relative to younger).
- Note: Manipulation of regression models to obtain interpretations for their parameters is *highly* emphasized in this course. For this reason, I will spend a lot of time working through complete procedures even in the most basic settings (you'll thank me later—and if you don't believe that, ask students who have previously taken this course :)).

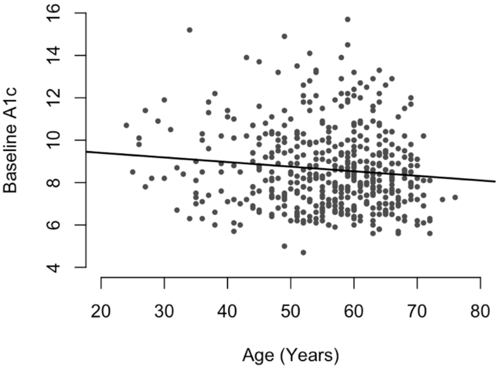
Simple linear regression

Re-framing: A simple linear model

- Previous slide is an example of a linear model—albeit a basic one.
  - In this very simple case of a binary predictor, the model is not really encoding any assumptions other than the existence of finite means in each group (an assumption we will always make in this course when it comes to real-world data, even if only implicitly).
- How does the model (previous slide) relate to the *t*-test?
- What are some limitations (conceptual or mathematical) of dichotomizing age?

Simple linear regression

Scatterplot: Baseline HbA1c and (continuous) age



(Noteworthy questions: How do we interpret/estimate the above line?)

Simple linear regression

Setup: Continuous predictor

- Model for the mean: $\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x$.
- Y: Outcome, response, ~~dependent variable~~.
- X: Exposure, predictor, explanatory variable, ~~independent variable~~.
- β_0: Mean value of Y among subgroup for which $X = 0$.
- β_1: Difference in mean value of Y comparing subgroups differing in their value of X by a single unit.
- This is called a *linear regression model*.

Simple linear regression

Continuous age:

- Mean HbA1c as a function of continuous age.
  - X: age (years).
  - Y: baseline HbA1c.
- Model: $\mathrm{E}[Y|X=x] = \beta_0 + \beta_1 x$.
  - Model encodes assumption: linearity of $\mathrm{E}[Y|X=x]$ in x.
  - May or may not hold (will later discuss evaluation of linearity).
- $\mathrm{E}[Y|X=0] = \beta_0 + \beta_1 \times 0 = \beta_0$
- $$\begin{aligned}\mathrm{E}[Y|X=x+1] - \mathrm{E}[Y|X=x] &= (\beta_0 + \beta_1(x+1)) - (\beta_0 + \beta_1 x)\\ &= \beta_1\end{aligned}$$
- The literal interpretations of β_0 and β_1 are as follows:
  - β_0: mean HbA1c among newborns.
  - β_1: difference in means between subgroups differing in age by one year.

Simple linear regression

A brief aside: Interpreting the intercept

- β_0 corresponds to a theoretical subgroup that is *not* well represented by the REACH study.
  - Recall: The age range in this study is 24 to 76 years.
  - To hold any stock in an estimate of β_0 would require extrapolation far beyond the range of our data.
  - An estimate of β_0 would be valid if the assumption of linearity held over the range 0 to 76 years, but we have no data to either graphically or analytically assess such an assumption.

Simple linear regression

Bizarre intercepts: Cringe-worthy interpretations!

- An even more extreme example: If the predictor of interest, X, were something like *height*, then the intercept (β_0) would mark the mean HbA1c among those of height zero.
  - Here, β_0 does not even carry a real-world interpretation, let alone the fact that our data cannot reasonably be used to reliably estimate it.

Simple linear regression

Discussion point:

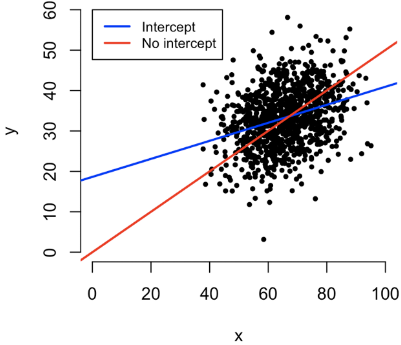
- Should still include an intercept in a model even when it cannot reliably be estimated or it does not carry a real-world interpretation?
  - To put it another way, why not reduce the model to the following form:

    $$\mathrm{E}[Y|X = x] = \beta x,$$

    the idea being that this model doesn't have any non-interpretable parameters?
- It turns out that this is a spectacularly bad idea, as by removing the intercept from the model we've actually made a very huge, untestable, and very often untrue assumption that the line goes through the origin! That is, by the above model, $\mathrm{E}[Y|X = 0] = 0$ necessarily.

Simple linear regression

Comparison: Which provides a better fit to the data?



- What would happen to each fitted line if I shifted everybody's value of X to the left, say, 20 units? How about to the right?

Simple linear regression

Discussion point: Include or exclude the intercept?

- To exclude an intercept is to anchor it to the origin.
  - If $\beta_0 \neq 0$, the slope of the fitted line in the no-intercept model depends upon where you center your values of X.
  - Is this true in the model that *includes* an intercept?
- Intercept allows the line to better fit the data over the range available, even if its clinical interpretability is dubious.
- It is important to know the coefficients...
  - ...that correspond to the clinical question!
  - ...that carry a real-world interpretation!
  - ...about which the data provide information!
- That's all I'll say about the intercept for now! Our focus is generally going to be on β_1.

Simple linear regression

Setup:

- Model for the mean: $\mathrm{E}[Y|X=x]=\beta_0+\beta_1 x$.
- Recall interpretation for β_1.
- Analogously,

$$\begin{aligned} \mathrm{E}[Y|X=x+c]-\mathrm{E}[Y|X=x] &= (\beta_0+\beta_1\times(x+c))-(\beta_0+\beta_1 x) \\ &= c\beta_1. \end{aligned}$$

- That is: $c\beta_1$ denotes difference in mean Y between subgroups differing in X by c units.
  - This is absolutely worth committing to memory. Do not compute subgroup-specific means and then subtract when you can use this very simple shortcut.

Simple linear regression

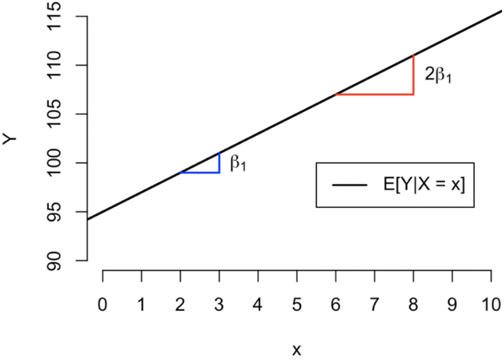


Table of Contents

1. Introduction to simple linear regression
2. Estimation by ordinary least squares
3. Subgroup-specific confidence intervals for mean outcomes
4. Prediction intervals
5. Diagnostic measures and tools

Simple linear regression

Notation: Setting the stage for estimation

- Model for the mean: $E[Y|X = x] = \beta_0 + \beta_1 x$.
- Re-expressed: $Y = \beta_0 + \beta_1 X + \epsilon,$ with $E[\epsilon|X = x] = 0$.
- We refer to ϵ as the *error* term.
  - "Error" is a misnomer in that it doesn't mean *mistake* in this context, but is better understood as "the part of Y that is not explained by X."
- Recognize that the model encodes a key assumption:
  - $E[Y|X = x]$ is linear in x.
  - This is sometimes written as $E[\epsilon|X = x] = 0$ for all x.
- Note: $E[\epsilon] = 0$ is a *convention*, but $E[\epsilon|X = x] = 0$ is an *assumption*.
  - Why?
- The way we estimate β_0 and β_1 will make use of this key assumption.

Simple linear regression

Notation: Setting the stage for estimation

- Let $\mathbf{x}$ denote the vector of exposure values for each observation.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}.$$

- But, confusingly, let $\mathbf{x}_i$ denote the covariate vector for subject i.

$$\mathbf{x}_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix}.$$

- Let $\mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{x} \end{pmatrix}$ denote the design matrix.
- Let $\mathbf{y}$ denote the outcome vector.

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}.$$

Simple linear regression

Point estimation: Least squares

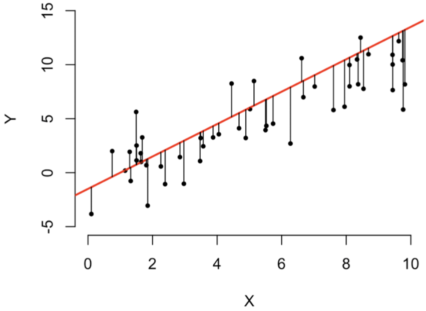
- One way to estimate $\boldsymbol{\beta}$ is to solve for the values that minimize the sum-of-squares expression:

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = \sum_{i=1}^{N}(y_i - \mathrm{E}[Y|X = x_i])^2 = \sum_{i=1}^{N}(y_i - (\beta_0 + \beta_1 x_i))^2.$$

- In the next set of notes, we will briefly discuss a geometric interpretation for this optimization problem that will justify its use.

Simple linear regression

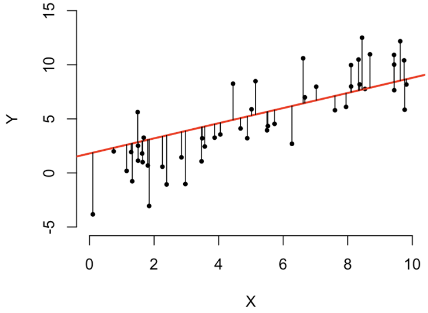
Least squares:



Sum of squared distances: 375.33 (Too high!)

Simple linear regression

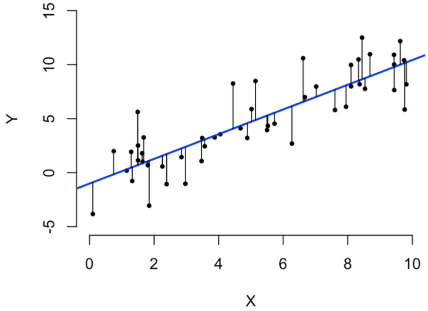
Least squares:



Sum of squared distances: 334.39 (Closer!)

Simple linear regression

Least squares:



Sum of squared distances: 228.62 (Just right!)

Simple linear regression

Point estimation: Least squares

- We solve this using matrix calculus (here for your reference, but *not* to appear on Type A HW problems or on exams).

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}}||\mathbf{y}-\mathbf{X}\boldsymbol{\beta}||^2 &= \mathbf{0}\\ \Rightarrow \mathbf{X}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) &= \mathbf{0}\\ \Rightarrow \mathbf{X}^T\mathbf{y}-\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} &= \mathbf{0}\\ \Rightarrow \widehat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.\end{aligned}$$

Rewriting,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^N (x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^N (x_i-\overline{x})^2} \quad \text{and} \quad \widehat{\beta}_0 = \overline{y}-\widehat{\beta}_1\overline{x}.$$

Simple linear regression

Point estimation: Least squares (bias and consistency)

- Assuming $\mathrm{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}$ is unbiased:

$$\begin{aligned}
\mathrm{E}[\widehat{\boldsymbol{\beta}}] &= \mathrm{E}[\mathrm{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] = \mathrm{E}[\mathrm{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}|\mathbf{X}]] \\
&= \mathrm{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{E}[\mathbf{y}|\mathbf{X}]] = \mathrm{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta})] \\
&= \mathrm{E}[(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta}] = \mathrm{E}[\boldsymbol{\beta}] = \boldsymbol{\beta}.
\end{aligned}$$

- $\widehat{\boldsymbol{\beta}}$ is consistent: ($\widehat{\boldsymbol{\beta}} \longrightarrow_p \boldsymbol{\beta}$). In words, $\widehat{\boldsymbol{\beta}}$ tends toward $\boldsymbol{\beta}$ as N grows with probability one.

Simple linear regression

Estimation: Least squares

- Focusing on $\widehat{\beta}_1$:

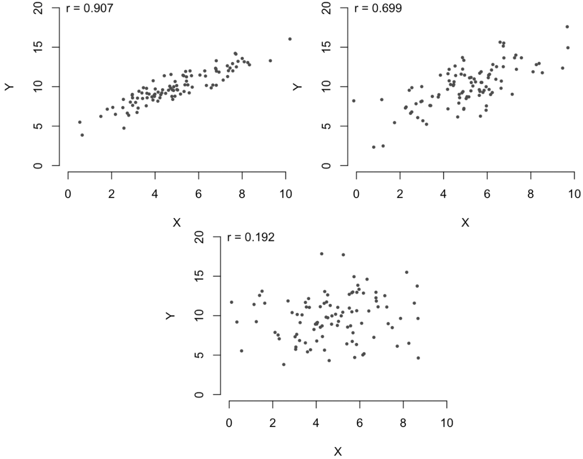
$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} = \frac{\widehat{\text{Cov}}(X,Y)}{\widehat{\text{Var}}[X]} = \frac{s_{xy}}{s_x^2}.$$

- Related to the correlation coefficient:

$$r = \widehat{\text{Corr}}(X,Y) = \frac{s_{xy}}{s_x s_y}.$$

- r bounded between -1 and 1, with $|r| = 1$ indicating perfect correlation (either positive or negative).
- Both $\widehat{\beta}_1$ and r must share the same sign (+/-/0).

Simple linear regression



- Guessing correlations is hard!
- Don't believe me? `http://guessthecorrelation.com`.

Simple linear regression

More definitions:

- We use the term *predicted value* as shorthand for "the estimated mean of Y among the subgroup having predictor value x."

$$\widehat{y}_i = \widehat{Y}(x_i) \;=\; \widehat{\mathrm{E}}[Y|X = x_i] = \widehat{\beta}_0 + \widehat{\beta}_1 x_i.$$

- The *residual*:

$$\widehat{\epsilon}_i = y_i - \widehat{y}_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i).$$

- The *residual sum-of-squares* (RSS):

$$\mathrm{RSS} \;=\; \sum_{i=1}^{N} \widehat{\epsilon}_i^2.$$

- Do not confuse the residual, $\widehat{\epsilon}$, with the *error*, ϵ.

Simple linear regression

$\mathbf{R^2}$: Proportion of variance explained

- Correlation coefficient:

$$r \quad = \quad \frac{s_{xy}}{s_x s_y},$$

  related to *coefficient of determination*:

$$R^2 \quad = \quad \frac{s_{xy}^2}{s_x^2 s_y^2} \overset{\text{(boring math)}}{=} \cdots = 1 - \frac{\text{RSS}}{\sum_{i=1}^{N}(y_i - \overline{y})^2}.$$

- R^2: the proportion of variance in Y explained by X.

Simple linear regression

Interesting fact: The regression line passes through the point of means

- The *true* line passes through the point $(\mathsf{E}[X], \mathsf{E}[Y])$:

$$\mathsf{E}[Y] = \mathsf{E}_X[\mathsf{E}[Y|X]] = \mathsf{E}_X[\beta_0 + \beta_1 X] = \beta_0 + \beta_1 \mathsf{E}[X].$$

- Recall from a previous slide that $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$.
- Therefore, it follows that:

$$\begin{aligned} \widehat{Y}(\overline{x}) &= \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x} \\ &= (\overline{y} - \widehat{\beta}_1 \overline{x}) + \widehat{\beta}_1 \overline{x} = \overline{y}. \end{aligned}$$

- The fitted regression line will always pass through the point $(\overline{x}, \overline{y})$.

Simple linear regression

Variance of $\widehat{\boldsymbol{\beta}}$:

- We now turn our attention to understanding the *sampling* distribution of $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ (in particular, its variance—at least for starters).
- The regression model makes the linearity assumption quite explicit:
  - $\mathrm{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$.
- To derive an expression for $\mathrm{Var}[\widehat{\boldsymbol{\beta}}]$ requires *additional* assumptions. To *start* with, we will assume:
  - ϵ_i's are pairwise independent*
  - $\mathrm{Var}[Y|X = x] = \sigma^2$ for all x (constant variance)
- **Fair warning**: We will—as soon as possible—eliminate the second assumption, known as *homoscedasticity*.

* Fine detail: Pairwise independent variables with finite variance are *uncorrelated*.

Simple linear regression

Variance of $\widehat{\boldsymbol{\beta}}$: Derivation under assumption of homoscedasticity

- We first develop the theory under the assumption of homoscedasticity:

$$\begin{aligned}
\mathrm{Var}[\widehat{\boldsymbol{\beta}}] &= \mathrm{Var}[\mathrm{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] + \mathrm{E}[\mathrm{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] \\
&= \mathrm{Var}[\boldsymbol{\beta}] + \mathrm{E}[\mathrm{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] = \mathrm{E}[\mathrm{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] \\
&= \mathrm{E}[\mathrm{Var}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}|\mathbf{X}]] \\
&= \mathrm{E}[((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathrm{Var}[\mathbf{y}|\mathbf{X}]((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T] \\
&= \mathrm{E}[((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\sigma^2\mathbf{I})((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T] \\
&= \mathrm{E}[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}].
\end{aligned}$$

- On a surface level, this may look like your ordinary biostatistics proof, but this slide contains some hidden nuggets that we'll try to unpack one by one.

Simple linear regression

Variance of $\widehat{\boldsymbol{\beta}}$: Derivation under assumption of homoscedasticity

- Derivation:

$$\begin{aligned}
\text{Var}[\widehat{\boldsymbol{\beta}}] &= \text{Var}[\text{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] + \text{E}[\text{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] \\
&= \text{Var}[\boldsymbol{\beta}] + \text{E}[\text{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] = \text{E}[\text{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] \\
&= \text{E}[\text{Var}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}|\mathbf{X}]] \\
&= \text{E}[((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\text{Var}[\mathbf{y}|\mathbf{X}]((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T] \\
&= \text{E}[((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\sigma^2\mathbf{I})((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T] \\
&= \text{E}[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}].
\end{aligned}$$

- Three assumptions are invoked in this particular derivation:
  - Linearity ($\text{E}[Y|X = x] = \beta_0 + \beta_1 x$).
  - Pairwise independent (uncorrelated) errors.
  - Homoscedasticity.
- Can you identify where each assumption was invoked?

Simple linear regression

Variance of $\widehat{\boldsymbol{\beta}}$: Derivation under assumption of homoscedasticity

- Note that if $\mathbf{X}$ is *fixed* by design (as per, for instance, a randomized trial), we can push this derivation forward an additional step:

$$
\begin{aligned}
\text{Var}[\widehat{\boldsymbol{\beta}}] &= \text{Var}[\text{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] + \text{E}[\text{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] \\
&= \vdots \\
&= \text{E}[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}]. \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.
\end{aligned}
$$

- A valid estimator of $\text{Var}[\widehat{\boldsymbol{\beta}}]$ requires a valid estimate of σ^2 (we will present one shortly).
  - When I refer to a *valid* variance estimator I mean there is theoretical justification for its use (e.g., unbiasedness).
  - I may sometimes speak of variance estimators that are *approximately* valid under certain conditions (e.g., sufficiently large samples).

Simple linear regression

Variance of $\widehat{\boldsymbol{\beta}}$: Derivation under assumption of homoscedasticity

- If $\mathbf{X}$ is random, we need to invoke an additional piece of information:

$$\begin{aligned}\mathrm{Var}[\widehat{\boldsymbol{\beta}}] &= \mathrm{Var}[\mathrm{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] + \mathrm{E}[\mathrm{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] \\ &= \vdots \\ &= \mathrm{E}[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}].\end{aligned}$$

- Key fact: If $\mathbf{X}$ is random, $\frac{1}{N}(\mathbf{X}^T\mathbf{X}) \longrightarrow_p \mathrm{E}\left[\mathbf{x}\mathbf{x}^T\right]$.
- Therefore, $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ can serve as a representation of $\mathrm{Var}[\widehat{\boldsymbol{\beta}}]$ even if $\mathbf{X}$ is random.
- Now, the key to estimating $\mathrm{Var}[\widehat{\boldsymbol{\beta}}]$ lies in finding an estimator of σ^2.

Simple linear regression

Variance estimation: Assuming homoscedasticity

- Having said that, how do we estimate σ^2?
- It turns out that $\mathrm{E}[\mathrm{RSS}] = \overset{\text{(boring math)}}{\cdots} = (N-2)\sigma^2$, so:

$$\widehat{\sigma}^2 = \frac{1}{N-2}\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2 = \text{mean squared error (MSE)}.$$

  - $N-2$: sample size *minus* number of model parameters.
- $\widehat{\sigma}^2$: mean squared error (MSE)
  - Estimates variance of Y among specific subgroups of X.
  - If homoscedasticity does not hold, this interpretation is **invalid**.
  - You'll also see $\widehat{\sigma} = \sqrt{\widehat{\sigma}^2}$, the root mean squared error (RMSE).
- Hence, we may estimate $\mathrm{Var}[\widehat{\boldsymbol{\beta}}]$ as follows:

$$\widehat{\mathrm{Var}}[\widehat{\boldsymbol{\beta}}] = \widehat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1};\ \text{in particular, } \widehat{\mathrm{Var}}[\widehat{\beta}_1] = \frac{\widehat{\sigma}^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2}.$$

Simple linear regression

Interval estimation and inference: Assuming homoscedasticity

- Asymptotic result:

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{\text{SE}}[\widehat{\beta}_1]} = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\widehat{\sigma}^2 / \sum_{i=1}^N (x_i - \overline{x})^2}} \dot{\sim} t_{N-2}$$

- Null vs. alternative hypothesis: $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.
- Test statistic:

$$t = \frac{\widehat{\beta}_1}{\widehat{\text{SE}}[\widehat{\beta}_1]} = \frac{\widehat{\beta}_1}{\sqrt{\widehat{\sigma}^2 / \sum_{i=1}^N (x_i - \overline{x})^2}}$$

- Under H_0, has an approximate t_{N-2} distribution (exact if the errors are normally distributed); its absolute value may be compared to the $(1 - \alpha/2)$-quantile.
- Two-sided p-value: $p = 2 \times \text{P}(T > |t|)$.

Simple linear regression

Interval estimation and inference: Assuming homoscedasticity

- $100(1-\alpha)\%$ CI for β_1 can be obtained:

$$\widehat{\beta}_1 \pm t_{1-\alpha/2,N-2}\widehat{\mathrm{SE}}[\widehat{\beta}_1] = \widehat{\beta}_1 \pm t_{1-\alpha/2,N-2}\sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^N (x_i - \overline{x})^2}}.$$

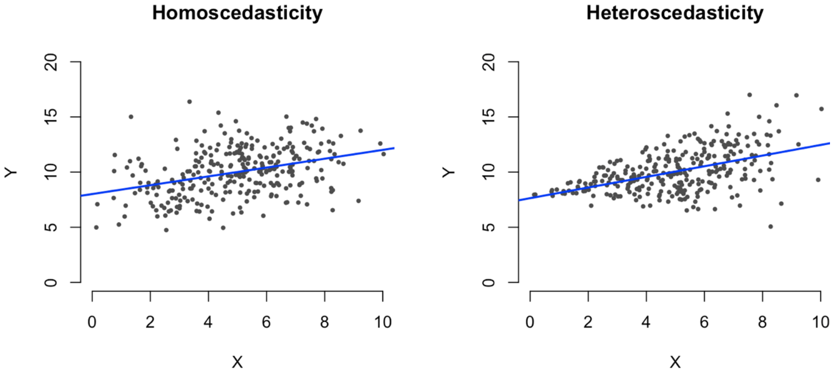
- The confidence interval inverts the test. Therefore, we may interpret confidence intervals in the "more interesting" way (i.e., as the set of all null hypotheses that cannot be ruled out by the data).

Simple linear regression

Homoscedasticity assumption: Do we need it?

- Recall the two flavors of the *t*-test (assuming equal variances and allowing unequal variances). Surprise! We're going to end up following the same pattern here.
  - Allowing unequal variances had almost no downside.
  - In this course, robust approaches are favored.
- If $\text{Var}[Y|X = x]$ depends upon x, we say there is *heteroscedasticity*.
- Our goal is to present and justify a variance estimator that is valid under the weaker condition of $\text{Var}[Y|X = x] < \infty$ for all x.

Simple linear regression



- Left: No obvious association between $\text{Var}[Y|X = x]$ and x.
- Right: $\text{Var}[Y|X = x]$ is larger for higher values of x.

Variance of $\widehat{\boldsymbol{\beta}}$: Derivation if heteroscedasticity allowed

- How far do we get in the derivation if we're not willing to assume homoscedasticity?

$$\begin{aligned}
\text{Var}[\widehat{\boldsymbol{\beta}}] &= \text{Var}[\text{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] + \text{E}[\text{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] \\
&= \text{Var}[\boldsymbol{\beta}] + \text{E}[\text{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] = \text{E}[\text{Var}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]] \\
&= \text{E}[\text{Var}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}|\mathbf{X}]] \\
&= \text{E}[((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\text{Var}[\mathbf{y}|\mathbf{X}]((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T] \\
&= \text{E}[(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\text{Var}[\mathbf{y}|\mathbf{X}]\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}].
\end{aligned}$$

- And that's about as far as we get before we're stuck.

Simple linear regression

Variance estimation: Allowing heteroscedasticity

- New line of attack: consider the following notation:

$$\mathbf{A} = \mathrm{E}[\mathbf{x}\mathbf{x}^T] \text{ and } \mathbf{B} = \mathrm{E}[\mathbf{x}(Y - \mathbf{x}^T\boldsymbol{\beta})(Y - \mathbf{x}^T\boldsymbol{\beta})^T\mathbf{x}^T].$$

- By a combination of the central limit theorem, the delta method, the law of large numbers, and Slutsky's theorem, it turns out *in large samples* that if the errors are uncorrelated,

$$\mathrm{Var}[\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \longrightarrow_p (\mathbf{A}^{-1})\mathbf{B}(\mathbf{A}^{-1})$$

- These can each be estimated:

$$\widehat{\mathbf{A}} = \frac{1}{N}(\mathbf{X}^T\mathbf{X}) \text{ and } \widehat{\mathbf{B}} = \frac{1}{N}(\mathbf{X}^T\mathrm{diag}(\widehat{\epsilon_i}^2)\mathbf{X})$$

Simple linear regression

Variance estimation: Allowing heteroscedasticity

- This suggests the following estimator:

$$\widehat{\text{Var}}[\widehat{\boldsymbol{\beta}}] \;=\; \frac{1}{N}(\widehat{\mathbf{A}}^{-1})\widehat{\mathbf{B}}(\widehat{\mathbf{A}}^{-1}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \text{diag}(\widehat{\epsilon}_i^2)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}.$$

- Known as the Huber-White, heteroscedasticity-consistent, robust, or sandwich variance estimator.
- There are other variants on this formula.
  - Some use a degrees of freedom correction (e.g., $N-1$ in the denominator).
  - Some weight the residuals according to leverage (we haven't talked about this yet).
  - In R, the "sandwich" package (and the `vcovHC`) are useful for this—be mindful of the types (HC0, HC1, etc.).

Simple linear regression

Homoscedasticity vs. heteroscedasticity:

- May replace SE that assumes homoscedasticity with the *robust* standard error to form CIs and conduct tests.
- If there is heteroscedasticity, then the MSE can instead be interpreted as estimating the *average* within-group variance (i.e., averaged over the values of X).
  - This interpretation is valid even when homoscedasticity holds, under which within-group variances would be constant (the average of something that is constant is the constant itself).
- **Note**: Choice of a standard error does not impact the value of $\widehat{\boldsymbol{\beta}}$.

Variance and interval estimation: Important notes

- Stata will accommodate heteroscedasticity in simple linear regression with the option `robust`.
- We have *not* assumed errors to follow a normal distribution.
  - Many textbooks state normality as an assumption of linear regression, but this is not sufficiently nuanced.
  - As sample size grows, the sampling distribution of $\widehat{\boldsymbol{\beta}}$ tends to a normal distribution by Lindeberg-Feller CLT, regardless of whether the errors are themselves normally distributed.
  - The question should not be about the error distribution, but about whether the sample size is sufficiently large for the normal approximation for $\widehat{\boldsymbol{\beta}}$ to be valid.
- We often use the *t*-distribution instead even when it is not exact, the idea being that it reflects the unknown error variance. This distinction matters less in large samples.

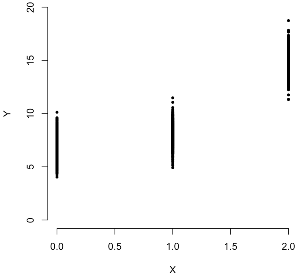
Simple linear regression

Variance of $\widehat{\boldsymbol{\beta}}$: If linearity is not satisfied…

- Interestingly, the robust variance estimator is still valid in large samples even when the linearity assumption is *not* satisfied, **as long as X is random and not fixed by design**.
- If **X** is fixed by design (as in a randomized experiment) and linearity is not correct, none of the variance estimation procedures we've learned so far will be valid.
  - To get a valid variance estimate, we would need something called the *conditional bootstrap*, which is something we'll cover toward the end of the semester. Hang tight!
- In the interest of time, I'm glossing over the theoretical details of why this is true, but I do want you to be aware of this important stipulation.

Simple linear regression

Variance of $\widehat{\boldsymbol{\beta}}$: If linearity is not satisfied and $\mathbf{X}$ is fixed ...



- **Nothing we have learned so far will get you the right answer!**
- The theoretical reason lies in estimation of the $\mathbf{B}$ matrix.
- Intuition: robust variance assumes $\mathbf{X}$ changes across study replicates; if it *doesn't*, you will tend to *overstate* the variance.
- The reason there is no problem with a fixed $\mathbf{X}$ when the model is correct is that $\mathrm{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]$ does not change with $\mathbf{X}$ anyway.

Simple linear regression

Understanding assumptions:

- Here is a summary of what has been specifically required for valid estimation of $\boldsymbol{\beta}$ and its variance so far ($\widehat{\text{Var}}_{\text{NR}}[\widehat{\boldsymbol{\beta}}]$ is the non-robust variance and $\widehat{\text{Var}}_{\text{R}}[\widehat{\boldsymbol{\beta}}]$ is the robust variance).

Assumption holds? (✓/✗)			**Valid? (Yes/No/Maybe)**		
Linearity	Homosc.	Normality	$\widehat{\beta}_1$	$\widehat{\text{Var}}_{\text{NR}}[\widehat{\beta}_1]$	$\widehat{\text{Var}}_{\text{R}}[\widehat{\beta}_1]$
✓	✓	✓	**Y**	**Y**	**Y**
✓	✓	✗	**Y**	M^2	M^2
✓	✗	✓	**Y**	**N**	**Y**
✓	✗	✗	**Y**	**N**	M^2
✗	✓	✓	$?^1$	**N**	M^3
✗	✓	✗	$?^1$	**N**	$M^{2,3}$
✗	✗	✓	$?^1$	**N**	M^3
✗	✗	✗	$?^1$	**N**	$M^{2,3}$

[1] - Not (yet) clear what β_1 actually means if linearity is violated.
[2] - Approximately valid under in large samples.
[3] - If **X** is random (but not if **X** is fixed).

- This table is, to me, as good of a justification I've got for use of the robust variance estimator.

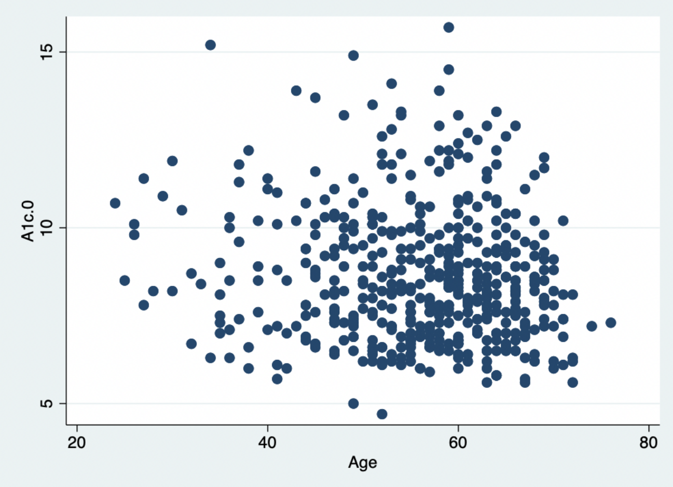
Simple linear regression

Example: REACH

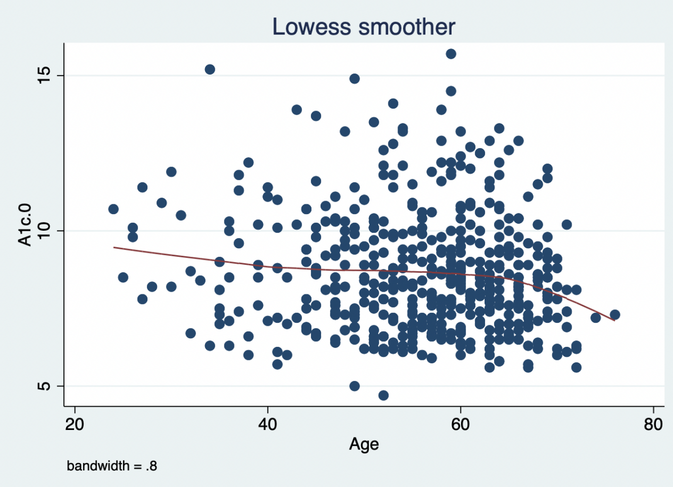
- Now that we've developed the theory, let's get a bit more practical and do what we've been longing to do: perform simple linear regression with *continuous* age as the predictor in the REACH example!
- Examining the data in a scatterplot can help orient you to the problem more easily:
  - Determine range of values for which model can be interpolated.
  - Detect serious violations to important assumptions.
  - Detect possible outliers.

Simple linear regression

Stata: Scatterplot (Stata: `scatter a1c0 age`)



Stata: LOWESS smoother (Stata: `lowess a1c0 age`)



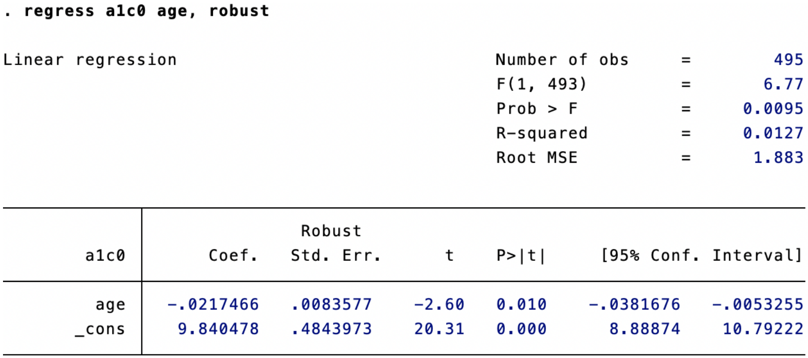
Simple linear regression

Example: REACH (Interpreting scatterplot)

- LOWESS (locally weighted scatterplot smoothing) provides insights into whether linearity approximately holds.
- Best not to worry too much about the curve at extreme values of X, in which the behavior can be unstable owing to sparse information.

Simple linear regression

Stata: REACH (continuous age)



```
. regress a1c0 age, robust

Linear regression                               Number of obs     =        495
                                                F(1, 493)         =       6.77
                                                Prob > F          =     0.0095
                                                R-squared         =     0.0127
                                                Root MSE          =      1.883

------------------------------------------------------------------------------
             |               Robust
        a1c0 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0217466   .0083577    -2.60   0.010    -.0381676   -.0053255
       _cons |   9.840478   .4843973    20.31   0.000      8.88874    10.79222
------------------------------------------------------------------------------
```

Simple linear regression

Example: REACH (Interpreting results)

- Based on the output, we estimate the mean HbA1c to be 9.84% in newborns (95% CI: [8.89, 10.8]). It is senseless to hold stock in this estimate given the minimum age in this data set is 24 years. Nevertheless, that is the literal interpretation on the basis of the model.
- When your target of inference is an *association*, you should spend little effort providing an in-depth interpretation for the intercept, particularly if it doesn't have real-world value.
- Instead, focus your efforts on a careful interpretation of the association you're trying to estimate (following slide).

Simple linear regression

Example: REACH (Interpreting results)

- This analysis provides evidence of an association between age and mean HbA1c ($p = 0.010$). We estimate the difference in mean baseline HbA1c to be 0.022% between subgroups differing in age by one year, with the older subgroup having the lower mean HbA1c. Based on a 95% CI, this estimate would not be deemed atypical if the true mean difference were between 0.0053% and 0.038%.

Simple linear regression

Example: REACH (Tables)

- When reporting results from a linear regression model in a manuscript, the following setup is recommended:

	Estimate	95% CI	p-value
Intercept	9.84	[8.89, 10.8]	< 0.001
REACH	-0.022	[-0.038, -0.0053]	0.010

- Note that we generally spend little time discussing the intercept.

Simple linear regression

Example: REACH (Further points)

- In this example, there is sufficient evidence of an association between age and mean HbA1c; not the case when we dichotomized age.
  - Dichotomizing inherently continuous variables often reduces power.
- Interestingly, association appears opposite of what we might expect, in the sense that HbA1c often increases as people age.
  - Recall: analysis is cross-sectional (observational) and is therefore subject to challenges of selection bias and confounding.

Simple linear regression

Example: REACH (Exercise)

- Suppose we seek to estimate difference in mean HbA1c comparing groups differing in age by five years?
- Simple solution:
  - Point estimate: $5\widehat{\beta}_1 = 5 \times (-0.0217466) = -0.109\%$.
  - 95% CI: multiply endpoints by 5:

$$[5 \times (-0.038168), 5 \times (-0.005326)] = [-0.191\%, 0.0266\%].$$

- Note that I am rounding to three significant digits only in the final stage of the computation.
  - This is the safest approach.

Table of Contents

1. Introduction to simple linear regression
2. Estimation by ordinary least squares
3. Subgroup-specific confidence intervals for mean outcomes
4. Prediction intervals
5. Diagnostic measures and tools

Learning about subgroups

Goal: Estimation of subgroup-specific means

- Suppose we seek to derive a point estimate and confidence interval for $\mathrm{E}[Y|X=x]$ (in the REACH example, this would be the mean HbA1c among the subgroups of "x").
- First must presume that linearity holds ($\mathrm{E}[\epsilon|X=x]=0$).
- If so, we can justify the following point estimate the point estimate.

$$\widehat{\mathrm{E}}[Y|X=x]=\widehat{Y}(x)=\widehat{\beta}_0+\widehat{\beta}_1 x.$$

- This is an unbiased estimator:

$$\mathrm{E}[\widehat{Y}(x)]=\mathrm{E}[\widehat{\beta}_0+\widehat{\beta}_1 x]=\beta_0+\beta_1 x=\mathrm{E}[Y|X=x]$$

- To form a confidence interval for $\mathrm{E}[Y|X=x]$, we need to be able to make statements about the sampling distribution of $\widehat{\mathrm{E}}[Y|X=x]$.

Learning about subgroups

Goal: Estimation of subgroup-specific means

- Assuming pairwise independence/homoscedasticity (constant error variance σ^2) makes it easier to determine $\text{Var}[\widehat{Y}(x)]$.

$$\text{Var}[\widehat{Y}(x)] = \overset{\text{(boring math)}}{\cdots} = \sigma^2 \left(\frac{1}{N} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{N} (x_i - \overline{x})^2} \right)$$

- Strictly speaking, this expression is only *technically* proper if X is fixed, but converges to $\text{E}[\text{Var}[\widehat{Y}(x)]]$ if X is random. In turn,

$$\widehat{\text{Var}}[\widehat{Y}(x)] = \frac{\text{RSS}}{N-2} \left(\frac{1}{N} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{N} (x_i - \overline{x})^2} \right).$$

Learning about subgroups

Goal: Estimation of subgroup-specific means

- Now, if we also assume that the errors are normally distributed, we have that

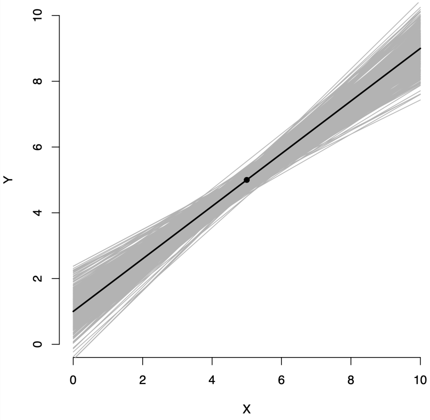
$$\frac{\widehat{Y}(x) - \mathrm{E}[Y|X = x]}{\sqrt{\widehat{\mathrm{Var}}[\widehat{Y}(x)]}} \sim t_{N-2}$$

- From this, we can formulate a (two-sided, symmetric) $100\times(1-\alpha)\%$ CI for $\mathrm{E}[Y|X = x]$:

$$\widehat{Y}(x) \pm t_{1-\alpha/2,N-2}\sqrt{\frac{\mathrm{RSS}}{N-2}\left(\frac{1}{N} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2}\right)}$$

Learning about subgroups

Intuition:



Estimated regression lines from 500 simulations. What stands out to you?

Learning about subgroups

Intuition:

- Notably, the width of a CI for $\widehat{Y}(x)$ does not depend on the relative concentration of values near $X = x$, but rather on the distance from $x = \overline{x}$ (this **is a consequence of linearity**). In fact, the CI for $\widehat{Y}(X)$ should be narrowest when $X = \overline{x}$ (see formula on previous slide).
- This can be explained heuristically:
  - We already know in advance that the regression line passes through the point of means.
  - An observation with a "central" value of X has information flowing towards it from *both* directions, whereas information closer to a boundary is only getting information from one direction.

Learning about subgroups

Goal: Estimation of subgroup-specific means

- Suppose we believe the necessary assumptions hold in the REACH example (we will discuss evaluation of assumptions, but for now we're focusing on procedures). Here are the components necessary to form a 95% CI for the mean HbA1c among the subgroup of age 30:
  - $N = 505$.
  - $\widehat{Y}(30) = \widehat{\beta}_0 + \widehat{\beta}_1 \times 30 = 9.8405 - 0.021747 \times 30 = 9.18808$.
  - $t_{1-\alpha/2,N-2} = t_{0.975,503} = 1.964691$.
    - In Stata: `disp invt(503, 0.975)`
  - $\widehat{\sigma}^2 = \text{MSE} = \text{RSS}/(N-2) = 1.883^2 = 3.545689$.
  - $(x - \overline{x})^2 = (30 - 55.89901)^2 = 670.7587$.
  - $\sum_{i=1}^{N}(x_i - \overline{x})^2 = (N-1)s_x^2 = 48165.851$.
- The 95% confidence interval is therefore given by

$$9.18809 \pm 1.96469 \times \sqrt{3.54569 \times \left(\frac{1}{505} + \frac{670.7587}{48165.851}\right)} = [8.72, 9.65].$$

Learning about subgroups

Comparison to naive approach:

- One simple (but naive) way that one could seek a confidence interval for $\widehat{Y}(30)$ would be to take observations in a small neighborhood of age 30 (say, ± one year) and form a CI for the mean using only those observations.
  - There are only four such observations!
  - Would you expect such a confidence interval to be *wider*, *narrower*, or *about the same width* as the confidence interval we formed on the prior slide?
- The model "borrows information," which allows us to learn about subgroups not specifically in our data set (as long as you're interpolating and not extrapolating).

Learning about subgroups

Comparison to naive approach:

- You should obtain a 95% CI of about [7.88, 12.9] using the naive approach (much wider).
    - Verify on your own =).
- A model can be used to "borrow information," across subgroups.
- Modeling assumptions (e.g., linearity) allow us to learn about a subgroup without requiring a large sample size *in that subgroup*.
    - As a caution: one should only interpolate, not extrapolate.

LEARNING ABOUT SUBGROUPS

Goal: Estimation of subgroup-specific means

- A more general expression for $\text{Var}[\widehat{Y}(x)]$ can be written down if you are unwilling to assume homoscedasticity. In general,

$$\text{Var}[\widehat{Y}(x)] \overset{\text{(boring math)}}{=\cdots=} \text{Var}[\widehat{\beta}_0] + (1+x)\text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) + x^2\text{Var}[\widehat{\beta}_1].$$

- To estimate this, let $\widehat{\text{Var}}[\widehat{\boldsymbol{\beta}}]$ be the sandwich estimator.

$$\frac{\widehat{Y}(x) - \text{E}[Y|X=x]}{\sqrt{\widehat{\text{Var}}[\widehat{Y}(x)]}} \dot{\sim} t_{N-2}$$

- An approximate $100\times(1-\alpha)\%$ CI can be derived in the usual way:

$$\widehat{Y}(x) \pm t_{1-\alpha/2,N-2}\sqrt{\widehat{\text{Var}}[\widehat{Y}(x)]}$$

Learning about subgroups

Stata: Estimation of mean HbA1c among subjects of age 30

- Though I want you to understand the assumptions and the intuition, I won't make you do these calculations "brute force."
- Instead, Stata's post-estimation command `lincom` can be used.
  - Stands for *linear combination*.
  - We will use this command for **many** purposes throughout the course.
- In fact, you can't actually do this calculation by brute force because Stata does not make the value of Cov$(\widehat{\beta}_0, \widehat{\beta}_1)$ apparent in the output.

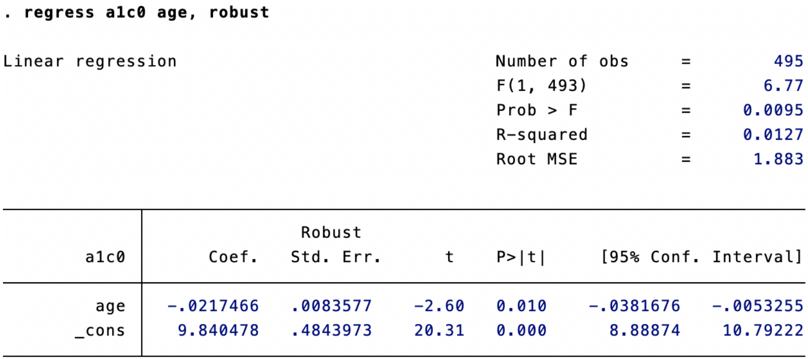
Learning about subgroups

Stata: Estimation of mean HbA1c among subjects of age 30

- Returning to the REACH example, suppose we seek to construct a 95% CI for mean HbA1c among those of age 30—but without having to assume error homoscedasticity.
  - Again, you can't form a 95% CI for $\beta_0 + 30 \times \beta_1$ without information on the covariance of $\widehat{\beta}_0$ and $\widehat{\beta}_1$.
  - Stata will do this for you when you use the `lincom` command.

Learning about subgroups

Stata: Estimation of mean HbA1c among subjects of age 30



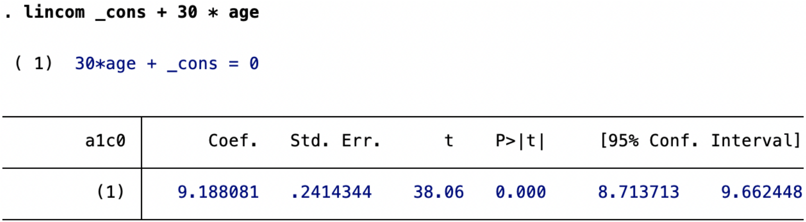
```
. regress a1c0 age, robust

Linear regression                               Number of obs     =        495
                                                F(1, 493)         =       6.77
                                                Prob > F          =     0.0095
                                                R-squared         =     0.0127
                                                Root MSE          =      1.883

------------------------------------------------------------------------------
             |               Robust
        a1c0 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0217466   .0083577    -2.60   0.010    -.0381676   -.0053255
       _cons |   9.840478   .4843973    20.31   0.000      8.88874    10.79222
------------------------------------------------------------------------------
```

Learning about subgroups

Stata: Estimation of mean HbA1c among subjects of age 30



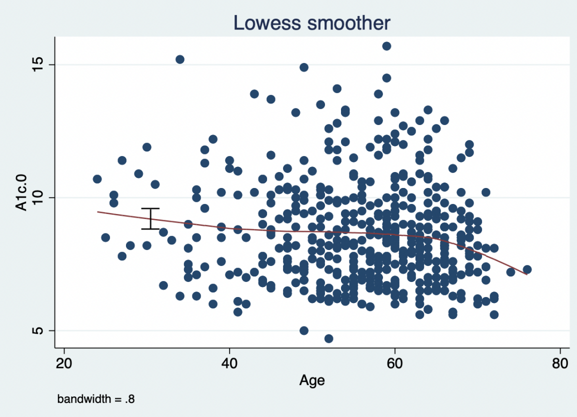
```
. lincom _cons + 30 * age

 ( 1)  30*age + _cons = 0

------------------------------------------------------------------------------
        a1c0 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   9.188081   .2414344    38.06   0.000     8.713713    9.662448
------------------------------------------------------------------------------
```

This provides a point estimate and (approximate) 95% CI for the mean HbA1c among those of age 30.

Learning about subgroups



Learning about subgroups

Stata: Estimation of mean HbA1c among subjects of age 30

- The difference between the “by-hand” calculation of the 95% CI and Stata’s calculation is very small (explainable by the use of robust standard errors).
- If using robust standard errors, we do not need homoscedastic, normally distributed errors; however, the linearity assumption is *key* to our ability to trust the subgroup-specific confidence interval for the mean.

Table of Contents

1. Introduction to simple linear regression
2. Estimation by ordinary least squares
3. Subgroup-specific confidence intervals for mean outcomes
4. Prediction intervals
5. Diagnostic measures and tools

Prediction intervals

Goal: Reference ranges and prediction intervals

- **Warning**: This next topic looks deceptively similar to the previous topic and it is therefore easy to confuse them.
- Now, suppose our goal is to develop an interval that encompasses range of plausible values for a *future* observations Y_{new} having exposure $X = x$.
- A naive $(1 - \alpha)$ prediction interval could be formed by treating the estimates as representations (approximations) of the truth:

$$\widehat{Y}(x) \pm z_{1-\alpha/2}\widehat{\sigma}$$

- If asked to construct a prediction interval "by hand" on an exam, this would be the most complicated method I would ask you to use, though it is limited in that it doesn't reflect the uncertainty around $\widehat{\boldsymbol{\beta}}$.

Prediction intervals

Goal: Reference ranges and prediction intervals

- Let $Y_{\text{new}}(x) - \widehat{Y}(x)$ denote the *prediction error*. Under pairwise independent, homoscedastic errors (common variance σ^2),

$$\text{Var}[Y_{\text{new}}(x) - \widehat{Y}(x)] = \sigma^2 + \text{Var}[\widehat{Y}(x)],$$

  which can be estimated as

$$\widehat{\text{Var}}[Y_{\text{new}}(x) - \widehat{Y}(x)] = \frac{\text{RSS}}{N-2}\left(1 + \frac{1}{N} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2}\right).$$

- If, further, the errors are normally distributed, it follows that

$$\frac{Y_{\text{new}}(x) - \widehat{Y}(x)}{\sqrt{\widehat{\text{Var}}[Y_{\text{new}}(x) - \widehat{Y}(x)]}} \sim t_{N-2}.$$

Prediction intervals

Goal: Reference ranges and prediction intervals

- An interval that encompasses $100\times(1-\alpha)\%$ of the distribution of $Y_{\text{new}}(x)$ can therefore be derived in the expected way:

$$\widehat{Y}(x) \pm t_{1-\alpha/2,N-2}\sqrt{\frac{\text{RSS}}{N-2}\left(1+\frac{1}{N}+\frac{(x-\overline{x})^2}{\sum_{i=1}^{N}(x_i-\overline{x})^2}\right)}.$$

Prediction intervals

Goal: A range of values for Y

- Suppose we believe the necessary assumptions hold in the REACH example (we will discuss evaluation of assumptions, but for now we're focusing on procedures). To form a 95% prediction interval for HbA1c among the subgroup of age 40 we need the following:
  - $N = 505$.
  - $\widehat{Y}(40) = \widehat{\beta}_0 + \widehat{\beta}_1 \times 40 = 9.8405 - 0.021747 \times 40 = 8.9706$.
  - $t_{1-\alpha/2,N-2} = t_{0.975,503} = 1.964691$.
  - $\widehat{\sigma}^2 = \text{MSE} = \text{RSS}/(N-2) = 1.883^2 = 3.545689$.
  - $(x - \overline{x})^2 = (40 - 55.89901)^2 = 252.77852$.
  - $\sum_{i=1}^{N}(x_i - \overline{x})^2 = (N-1)s_x^2 = 48165.851$.
- The 95% prediction interval is therefore given by

$$8.9706 \pm 1.96469 \times \sqrt{3.54569 \times \left(1 + \frac{1}{505} + \frac{252.7785}{48165.851}\right)} = [5.26, 12.7].$$

Prediction intervals

Goal: A range of values for Y

- Unlike the case in which we sought to derive subgroup-specific confidence intervals for the mean outcome, there is no straightforward generalization of the "prediction interval" methodology to accommodate the setting in which we are unwilling/unable to assume homoscedastic, normally distributed errors.
- In fact, Stata won't even *let* you estimate the prediction error variance as a post-estimation command if your regression model invoked the robust option.
- Suppose we want to use Stata to formulate the prediction interval on the previous slide.

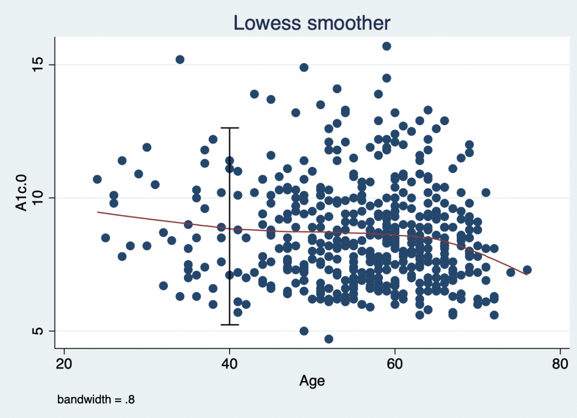
Prediction intervals

Stata: A range of values for HbA1c among subjects of age 40

```
regress a1c0 age
set obs `=_N+1'
replace id = _N if id == .
replace age = 40 if id == _N
predict prm if id == _N
predict prsd if id == _N, stdf
local tquant = invt(503, 0.975)
quietly: summarize prm
local a1c40m = r(mean)
quietly: summarize prsd
local a1c40s = r(mean)
display "[" `a1c40m' - `tquant' * `a1c40s' ", " `a1c40m' + `tquant' * `a1c40s' "]"
```

- `predict` features a number of options. If left blank, it assumes you're looking for the predicted mean. The `stdf` option is used, it will give you the prediction error variance for that observation.
- `local` stores a value outside of the data.
- The answer from Stata matches the "by-hand" calculations.
- This code does not explain itself; try each step yourself to see what I was doing and follow along in the logic.

Prediction intervals



Prediction intervals

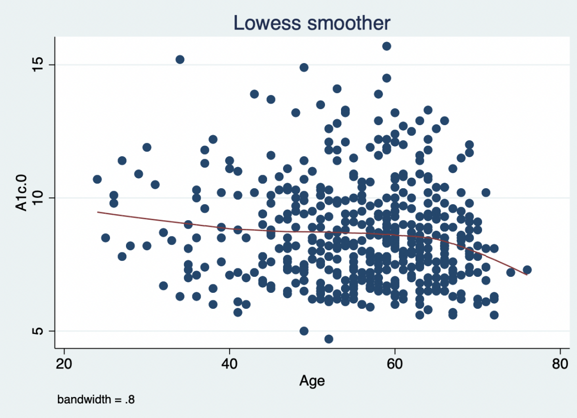
Defining subgroups: Do they exist in our data?

- Could we reasonably use this method to devise a prediction interval for individuals who are *exactly* 40.2845 years old (for example) even if no one in the data set is exactly that age?
- Predicted value of Y at a particular x borrows information from nearby values of x.
- If linearity holds in observed range of X, then the predicted value should be valid even for values of X that are not specifically observed in the data (interpolation). If our other assumptions hold, then the prediction interval will be valid.
- We get into trouble if we try to form prediction intervals for specific values of x that are *outside* the range of our study population (extrapolation).

Table of Contents

1. Introduction to simple linear regression
2. Estimation by ordinary least squares
3. Subgroup-specific confidence intervals for mean outcomes
4. Prediction intervals
5. Diagnostic measures and tools

Evaluation of assumptions



Exercise: Comment on the extent to which this figure provides insights on potential violations to assumptions of linearity, homoscedasticity, and error normality (Hint: This is hard).

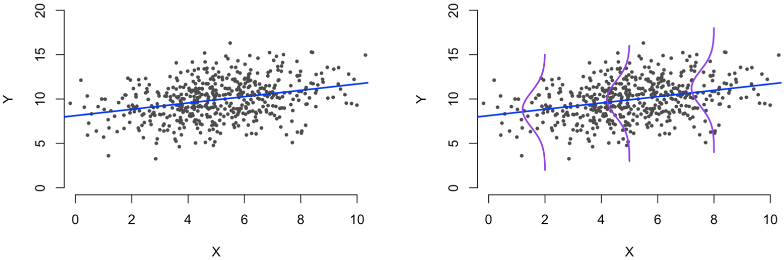
Evaluation of assumptions

Evaluating assumptions:

- In this section, we will discuss methods to evaluate regression modeling assumptions.
  - **Warning**: The assumptions required will depend upon the purpose for which the model is being used. Therefore, not all of the diagnostic methods will apply 100% of the time.
  - "Just because you can doesn't mean you should."
- The assumption of independent errors is one that can typically be justified from the study design (we will not use the data to validate or falsify the assumption of independent errors).
- Scatterplots can be used to detect moderate to severe departures from linearity, homoscedasticity, and normality.
  - Minor violations to homoscedasticity and normality of errors are very difficult to detect on a scatterplot.

Evaluation of assumptions

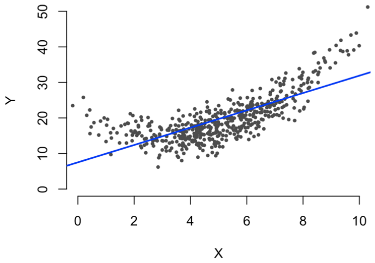
Example 1: All assumptions hold

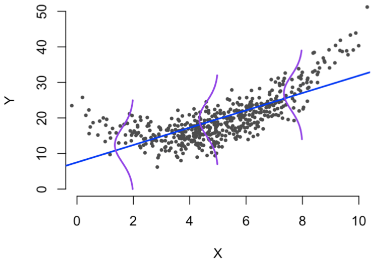


- Estimate of and CI for β_1 **valid**.
- Estimate of and CI for $\mathrm{E}[Y|X = x]$ **valid**.
- Prediction interval for $Y_{\text{new}}(x)$ **valid**.

Evaluation of assumptions

Example 2: Linearity violated; homoscedasticity/normality satisfied

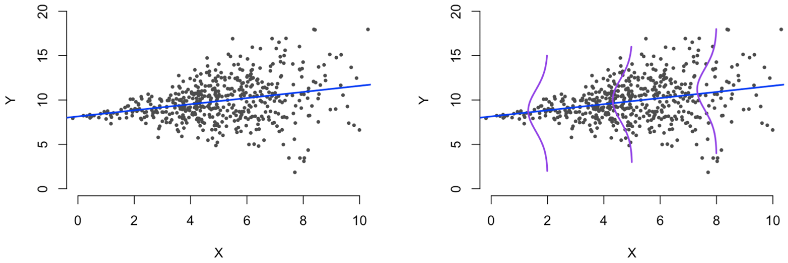




- Estimate of β_1 *harder* (but not actually impossible) to interpret.
- CI for β_1 **valid** if robust SE used, N sufficiently large, and $\mathbf{X}$ random.
- Estimate and CI for $\mathrm{E}[Y|X = x]$ **invalid**.
- Prediction interval for $Y_{\text{new}}(x)$ **invalid**.

Evaluation of assumptions

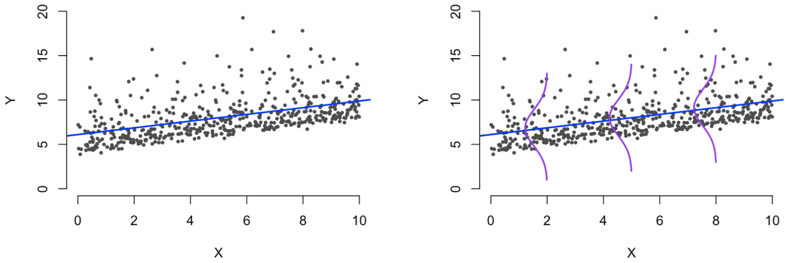
Example 3: Homoscedasticity violated; linearity/normality satisfied



- Estimate of β_1 **valid**.
- CI for β_1 **valid** if robust SE used and N sufficiently large.
- Estimate of $E[Y|X = x]$ **valid**.
- CI for $E[Y|X = x]$ **valid** if robust SE used.
- Prediction interval for $Y_{\text{new}}(x)$ **invalid**.

Evaluation of assumptions

Example 4: Normality violated; linearity/homoscedasticity satisfied



- Estimate of β_1 **valid**.
- CI for β_1 **valid** if N sufficiently large.
- Estimate of $E[Y|X = x]$ **valid**.
- CI for $E[Y|X = x]$ **valid** if N sufficiently large.
- Prediction interval for $Y_{\text{new}}(x)$ **invalid**.

Evaluation of assumptions

Notes:

- We can see, for instance, that prediction intervals suffer from some pretty serious flaws when assumptions are violated.
- It is challenging to use scatterplots to detect departures from assumptions when there is more than one violation at a time.
- This motivates the use of other diagnostic tools.

Diagnostics

Ideas:

- Regression diagnostics are a set of tools used to assess/inspect the validity of assumptions.
- There are *many* of visual and numeric aids, and I cannot possibly cover them all. The LOWESS smoother is one such visual aid. Others include:
  - Residual-versus-predictor plot (`rvpplot`).
  - Residual-versus-fitted plot (`rvfplot`).
  - Quantile-quantile plot (`qnorm`).
- These Stata commands can be entered directly after the regression model of interest has been fit. Personally, I don't like these commands because they don't use studentized residuals. They also don't automatically include LOWESS curves. We'll have to take the back roads. Sorry!

Diagnostics

Studentized residuals:

- Define the *hat* matrix as $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
  - The hat matrix is so named because $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$.
- If $\text{Var}[\boldsymbol{\epsilon}|\mathbf{X}] = \sigma^2\mathbf{I}$, then $\text{Var}[\widehat{\boldsymbol{\epsilon}}|\mathbf{X}] = \sigma^2(\mathbf{I} - \mathbf{H})$.
- In other words, if the *errors* are themselves homoscedastic and independent, the *residuals* will not be.
- We wish to graphically assess homoscedasticity. Therefore, define the studentized residual as:

  $$t_i = \frac{\widehat{\epsilon}_i}{\widehat{\sigma}\sqrt{1 - h_i}},$$

  where h_i is the i^{th} diagonal entry of $\mathbf{H}$.
- When ϵ_i is homoscedastic, so too is t_i. Therefore, we use t_i to evaluate homoscedasticity instead of $\widehat{\epsilon}_i$.

DIAGNOSTICS

Stata code: Prediction

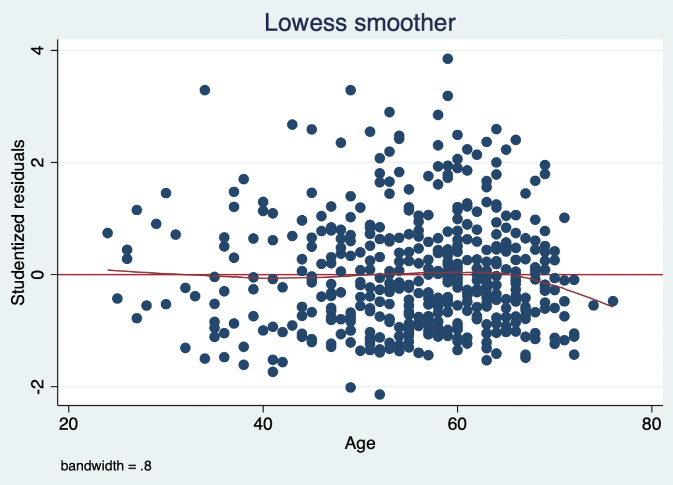
- Fitted values, $\widehat{Y}(x_i)$: `predict` *`fitted`*
    - *`fitted`* denotes the name of the variable you want to create that will contain the fitted (or "predicted") values.
    - You could have inputted *`pterodactyl`* instead of *`fitted`* and it would put in a column of fitted values with the name "pterodactyl."
- Studentized residuals: `predict` *`stres`*`, rstudent`
    - *`stres`* denotes the name of the variable you want to create that will contain the studentized residuals (you can choose it).
    - Stata only lets you do this following a regression command that assumes homoscedasticity (*without* `robust` option).

Diagnostics

Residual-versus-predictor plot:

- x-axis: predictor, X; y-axis: studentized residual.
- Code: `lowess` *`stres pred`*`, yline(0)`
  - *`stres`* denotes variable you created of studentized residuals.
  - *`pred`* denotes your predictor of interest.
- How to evaluate assumption violations:
  - Linearity: mean not zero (x-axis) across fitted values.
  - Homoscedasticity: variance non-constant across fitted values.
  - Normality: residual distribution asymmetric and/or more extreme values than expected.

Residual-versus-predictor plot:



Diagnostics

Residual-versus-predictor plot: Nuance in language!

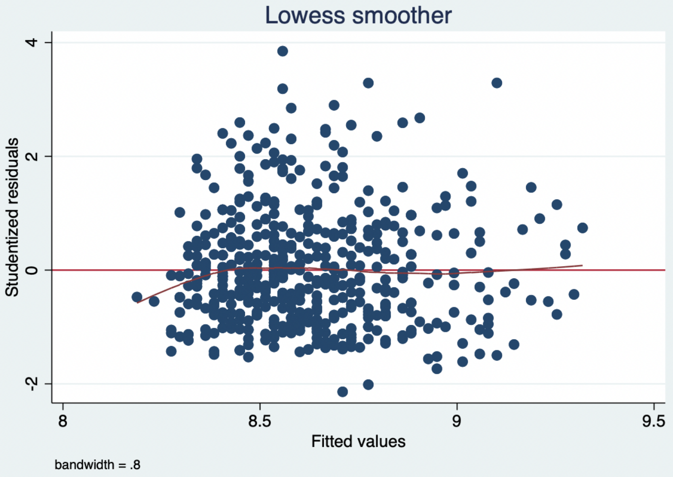
- Appropriate characterizations:
  - This plot does not provide compelling evidence of a serious departure to the assumption of linearity.
  - There is little clear graphical evidence of heteroscedasticity.
  - This plot suggests possible right-skewness of the errors.
- Inappropriate characterizations:
  - This plot shows nonlinearity based on the LOWESS curve's downward trend for higher values of age.
  - Based on the plot, the linearity assumption is satisfied.
  - This plot shows that the errors are homoscedastic.
  - The errors appear to be heteroscedastic because the errors have lower variance for lower values of age.
- Precise, careful, nuanced interpretation makes for a good collaborator—even when people *want* a clear yes or no, you can't always give it to them.

Diagnostics

Residual-versus-fitted plot:

- x-axis: fitted value, $\widehat{y}$; y-axis: studentized residual.
- Code: `lowess` *`stres fitted`*`, yline(0)`
  - *`stres`* denotes variable you created of studentized residuals.
  - *`fitted`* denotes variable you created of fitted values.
- How to evaluate assumptions:
  - Linearity: mean approximately zero across fitted values.
  - Homoscedasticity: variance approximately constant across fitted values.
  - Normality: symmetric; bell shape; few extreme values.
- In simple linear regression, this gives you essentially the same insights as the residual-versus-predictor plot. When we have multiple variables to consider, it can be helpful to summarize all the predictors into a fitted value.

Residual-versus-fitted plot:



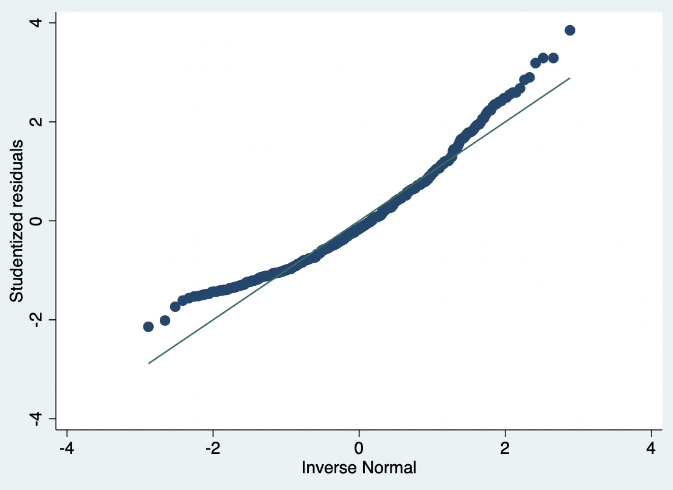
Diagnostics

Quantile-quantile (QQ) plot:

- *y*-axis: studentized residual.
- *x*-axis: corresponding quantile of a standard normal distribution.
- Code: `qnorm` *`stres`*.
  - *`stres`* denotes variable you created of studentized residuals.
- To generate QQ plot by hand:
  1. Obtain a percentile rank for each (studentized) residual.
  2. For each observation, determine the theoretical value of a standard normal distribution having that percentile rank (e.g., a residual 97.5% higher than other residuals would be assigned a value of 1.96).
  3. Plot the values in Step 2 on the *x*-axis against the studentized residuals themselves on the *y*-axis.
- If the errors are normally distributed, the studentized residuals should be very close to those corresponding normal quantiles.
- Hence, we look to see if the points are tracking along the line $y = x$.

Diagnostics

Quantile-quantile (QQ) plot:



Diagnostics

Quantile-quantile (QQ) plot:

- If the errors are normally distributed, then the QQ plot should appear closely aligned with the line $y = x$. This doesn't appear to be the case here. The QQ plot in the previous slide provides graphical evidence of a departure from normally distributed errors.
- In making this characterization, I'm not looking at the three trailing/extreme points on either side. I'm looking at what appears to be some real evidence of a "U"-type shape that encompasses a nontrivial amount of the data.
- In general, there may be a handful of "noisy" values on the sides that don't align. This isn't what you're looking for in finding evidence of a departure from normality.
- Instead, look for a clear, non-trivial departure from the line $y = x$.

Diagnostics

Diagnostics: Which ones are useful and when?

- If the goal is to estimate associations, key assumptions are independent errors, finite variance, and linearity.
  - We don't graphically evaluate independent errors. Instead, independent errors are ideally a feature of the study design.
  - We will generally always assume finite variance without too much hesitation.
- **Key point**: No need to cycle through each diagnostic plot every time you run a regression. They serve as visual tools to make you, the researcher, better informed.

Diagnostics

Diagnostics: Which ones are useful and when?

- If the goal is to form prediction intervals, further require errors to be homoscedastic and normally distributed.
  - In this setting, diagnostic plots are worth a closer look.

Diagnostics

Other diagnostics: Leverage

- Diagonal entries, h_i, of $\mathbf{H}$ are referred to as *leverage* of each observation i, and they can be realized as:

  $$h_i = \frac{\partial \widehat{Y}(x_i)}{\partial y_i},$$

  rate at which predicted mean changes with realized value of y.
- Fun facts about leverage:
  - $\text{Var}[\widehat{\epsilon}_i | X = x_i] = \sigma^2(1 - h_i)$.
  - $1/N \leq h_i \leq 1$.
  - $\sum_{i=1}^{N} h_i = K + 1$.
  - Leverage *only* depends upon X.
- Question: Do observations with *higher* leverage tend to be *more* or *less* concentrated around the fitted line?
- Question: What x-values are associated with higher leverage?

DIAGNOSTICS

Other diagnostics: Influence

- The *influence* of observation i is defined to be the amount $\widehat{\boldsymbol{\beta}}$ changes when the i^{th} observation is removed.
- The formula for the influence of an observation is given by:

$$\text{Influence}_i = \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i^T\widehat{\epsilon}_i}{1 - h_i}.$$

- Not interchangeable with leverage; an observation can have high leverage, high influence, both, or neither.

Other diagnostics: Outliers

- An outlier isn't necessarily high leverage or high influence.
- The decision to *remove* an outlier should generally not be made on the basis of significance, leverage, or influence.
- When an outlier is detected, it should be evaluated whether or not the value is plausible. Is it a data entry error? If so, either fix it (or if this is not possible, it is better to remove).
- When an outlier is not the result of an error, but its inclusion makes the difference between statistical significance, you should generally report the analysis *with* the outlier included, and show the results of the analysis *without* the outlier included as a sensitivity analysis.

Diagnostics

Linearity: A big deal?

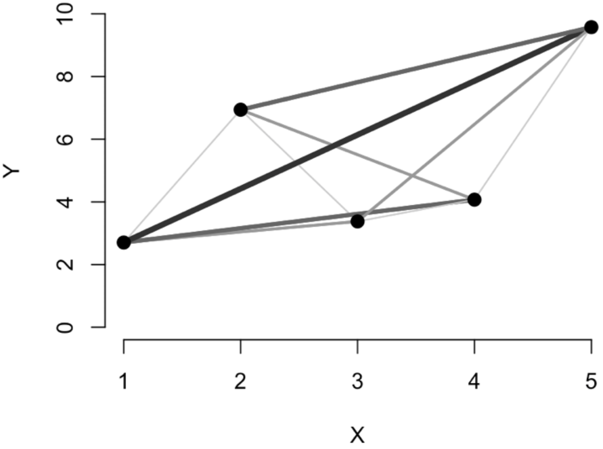
- Simple linear regression formally assumes that $\mathrm{E}[Y|X=x]$ and x are linearly related.
- However, $\widehat{\beta}_1$ can be expressed as follows ($1 \leq i, j \leq N$):

$$\widehat{\beta}_1 \quad = \quad \sum_{i,j=1}^{N} \left(\frac{(x_i - x_j)^2}{\sum_{k,l=1}^{N}(x_k - x_l)^2} \right) \left(\frac{y_i - y_j}{x_i - x_j} \right) = \sum_{i,j=1}^{N} w_{ij} \mathrm{slope}_{ij}.$$

- $\widehat{\beta}_1$: weighted sum of the slopes of lines connecting all pairs of points (weights are proportional to squared x-distances between pairs).
- When linearity is not exact, β_1 characterizes first-order trend. If the goal is simply to evaluate/test an overall "first-order" association, the threshold for defining a concerning departure from linearity may be a bit higher than if you were to use the model to learn about subgroups.

Diagnostics

Pairwise slopes:



Summary

Notes: Topics in this unit

- Interpretation.
- Assumptions.
- Estimation procedures.
- Uses.
- Most importantly: *how the above four points interact with each other*.

SUMMARY

Notes: Next unit

- Multiple linear regression.
- Multiple variables in the same model.
- Key transformations.
- Weighted least squares.