Biost 517 / Biost 514
Applied Biostatistics II / Biostatistics II

Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 16:

Linear Regression Extension: Allowing for Heteroscedasticity

Linear Regression: Alternative Representation

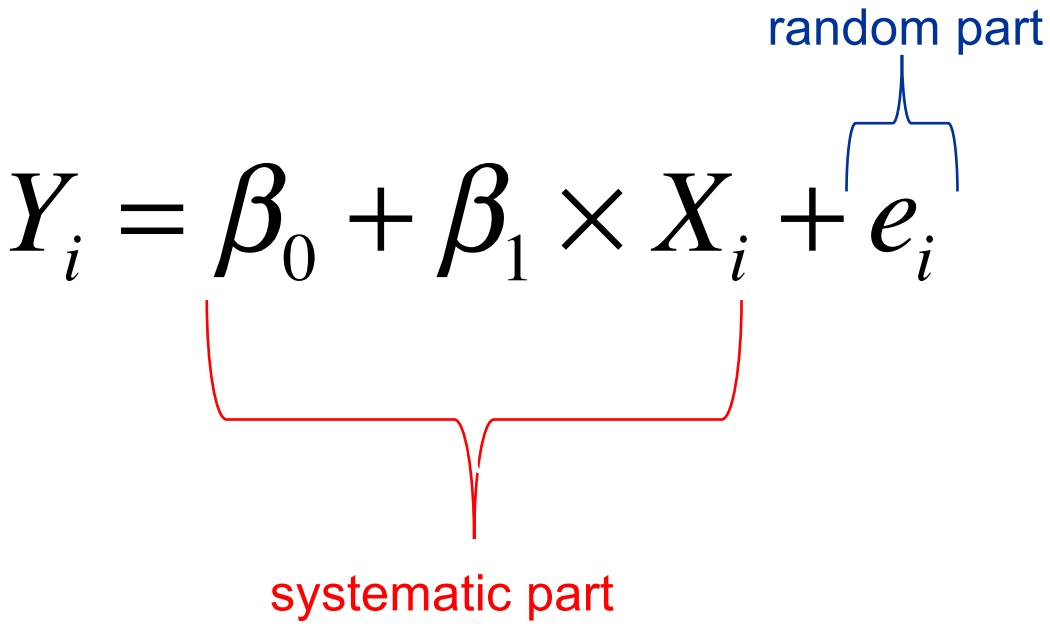
- In classical linear regression, for each individual i the response variable conditional on the predictor is assumed to have the following normal distribution: $Y_i \mid X_i \sim N(\theta_i, \sigma^2)$ where

$$\theta_i = E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$$

- Linear regression models are often expressed in terms of the response instead of the mean response

$$Y_i = \beta_0 + \beta_1 \times X_i + e_i$$

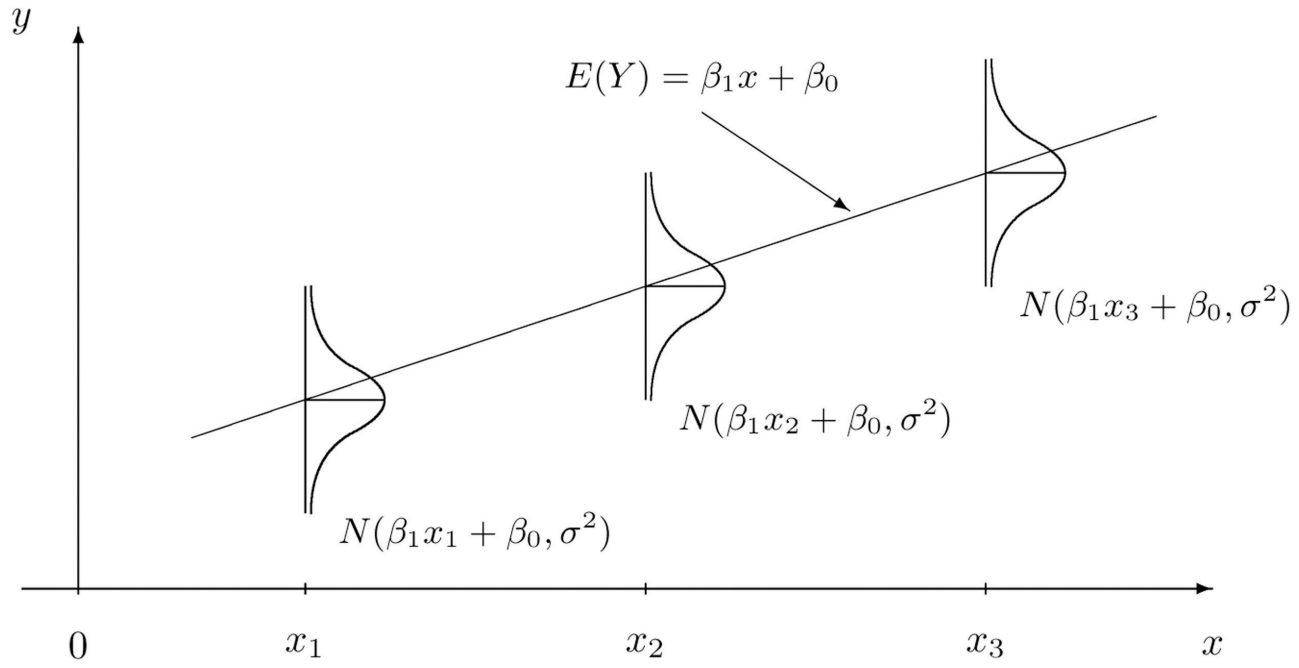
Alternative Representation



- Systematic part is "the signal": the mean
- The random part is "the noise": the residual is the error and is the difference between the observed value and the mean
- Classical linear regression assumes $e_i \sim N(0, \sigma_e^2)$ are independent and identically distributed (iid) for all individuals i in the sample.
- Note that $Var(Y_i) = Var(e_i) = \sigma_e^2$

Alternative Representation

- So e_i in classical linear regression assumes:
  - Normal distribution
  - Mean zero for every value of the predictor X
  - Constant variance σ_e^2 for every value of the predictor X
  - Values that are statistically independent
- Inference based on classical linear regression is based on an assumption of **homoscedasticity,** i.e., constant variance across groups



LR with Binary Predictor and T-test: Presuming Homoscedasticity

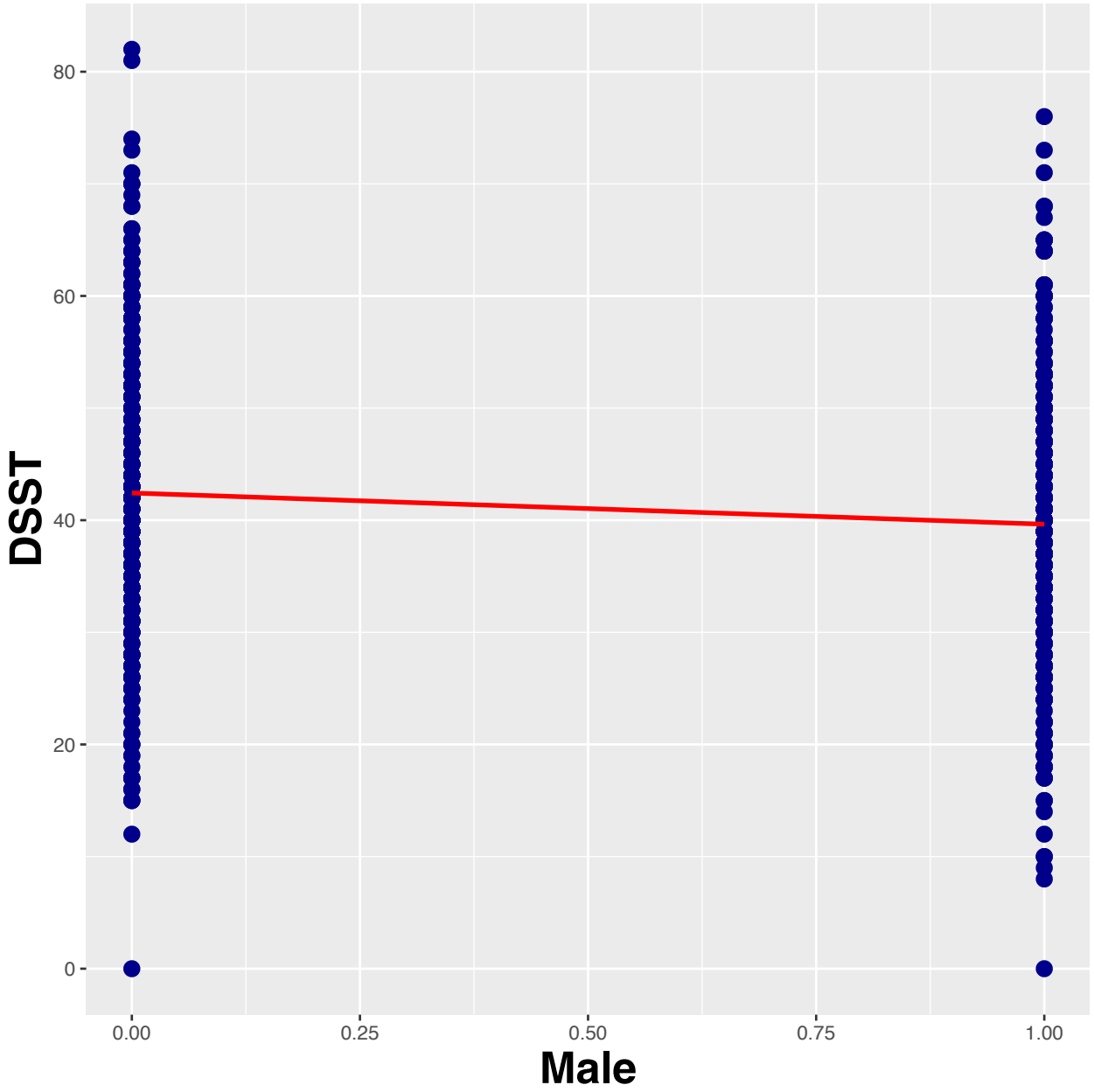
- Linear regression with a binary predictor
  - Classical LR: exactly the t test that presumes equal variances
- Example: Interested if there is an association between dsst scores and gender. Will compare mean dsst across groups to assess if there is an association.
- Male is a 0/1 variable (1 for males and 0 for females)
- What is the interpretation of the intercept and the slope in the linear regression model with a binary predictor for male?

Model	$E[Y_i \mid X_i] = \beta_0 + \beta_1 \times X_i$
$X_i = 0$	$E[Y_i \mid X_i = 0] = \beta_0$
$X_i = x$	$E[Y_i \mid X_i = x] = \beta_0 + \beta_1 \times x$
$X_i = x+1$	$E[Y_i \mid X_i = x+1] = \beta_0 + \beta_1 \times x + \beta_1$

Binary Predictor Example: Estimates

- A “saturated model”: Number of groups = number of parameters
  - The predictor variable used in the analysis only has two values
  - The regression model has two parameters
  - We are not borrowing information across the groups for the mean
  - Each group mean can be fit exactly
    - Intercept is the sample mean for females
    - Intercept plus slope is the sample mean for males
- We could of course reparameterize our model if we wanted
  - **`mridata$female <- 1 - mridata$male`**
  - **`lm(dsst~female,data=mridata)`**
    - Then intercept would be the sample mean for males
    - Intercept plus slope would be sample mean for females

Example: DSST by Sex (male reference)



Example: DSST by Sex (male reference)

```
> t.test(dsst~male,var.equal=TRUE,data=mridata)

        Two Sample t-test

data:  dsst by male
t = 2.9617, df = 721, p-value = 0.00316
alternative hypothesis: true difference in means is not equal to 0
> summary(lm(dsst~male,data=mridata))

Call:
lm(formula = dsst ~ male, data = mridata)

Residuals:
    Min      1Q  Median      3Q     Max
-42.428  -8.643  -0.428   8.572  39.572

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.4278     0.6597  64.311  < 2e-16 ***
male         -2.7845     0.9402  -2.962  0.00316 **
---
```

Classical LR and one-sample CI

- The CI for the intercept in classical linear regression is <u>not</u> the CI for the females for a one sample analysis, because in regression a pooled SD is used

One sample $$\bar{Y}_F \pm t_{.025,n_F-1} \times \frac{s_F}{\sqrt{n_F}}$$

Regression $$\hat{\beta}_0 \pm t_{.025,n_M+n_F-1} \times \frac{RMSE}{\sqrt{n_F}}$$

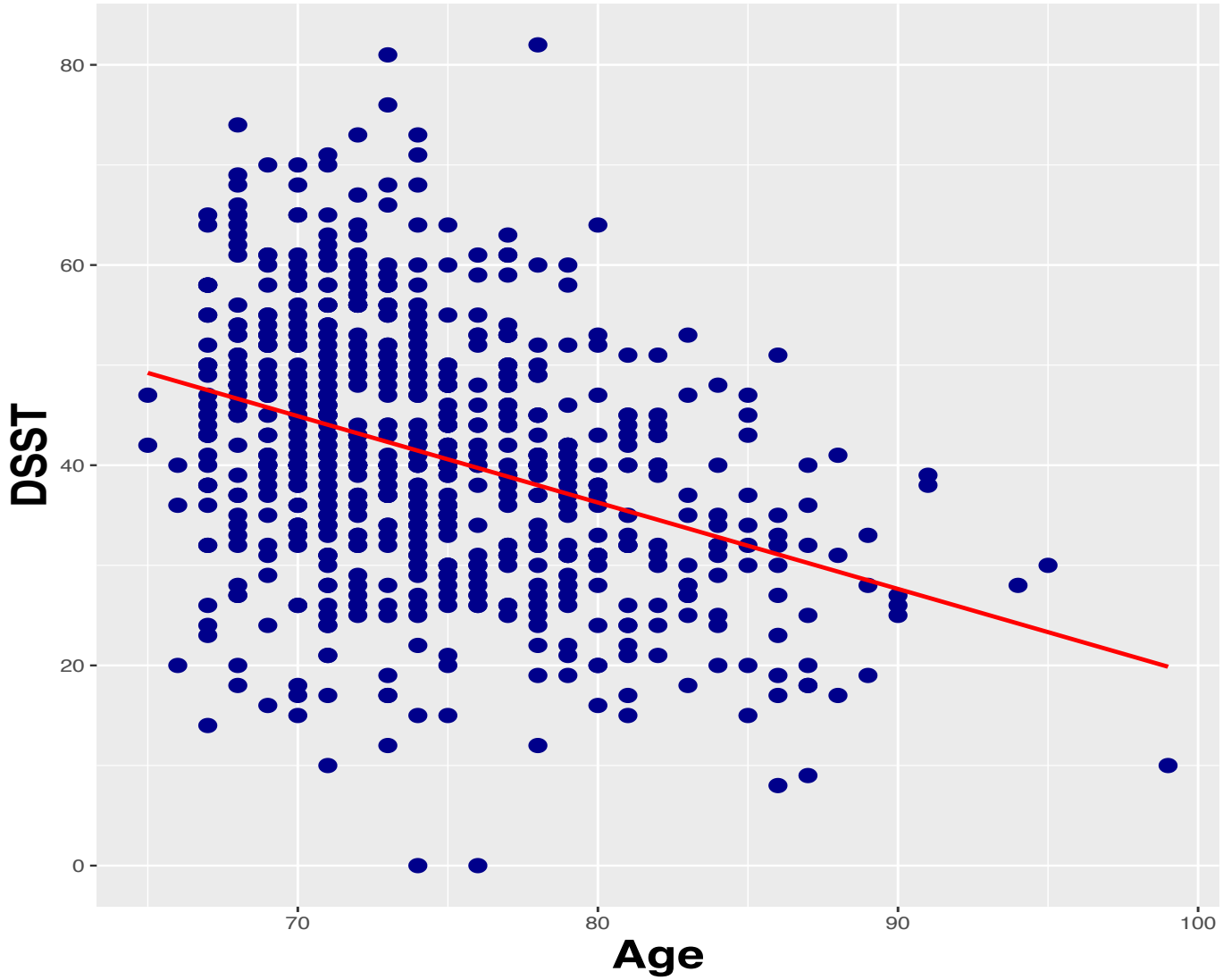
$$RMSE = \sqrt{s^2_{pool}} = \sqrt{\frac{(n_M-1)s^2_M + (n_F-1)s^2_F}{n_M+n_F-2}}$$

Violations of Constant Variance Assumption

- When the homoscedasticity assumption is violated, confidence intervals and p-values from classical linear regression may not be valid
- In particular, the significance of contrasts between groups can be misleading or incorrect if σ_e^2 differs substantially across the subgroups being compared, and subgroups differ in size
- Note that while violations of constant variance do not make the coefficient estimates biased, standard errors are incorrect, which affects precision and possibly inference.

Example:DSST and Age

- Subset of 723 individuals with mri measurements



Violations of Constant Variance Assumption

- How to detect violations of homoscedasticity, i.e., **heteroscedasticity**
- Most textbooks suggest plotting the residuals for violations of the constant variance assumptions
  - ❑ Residual versus predictors (RVP) plots
  - ❑ Residual versus fitted (RVF) plots
- If the constant variance assumption is met, then the vertical spread of the residuals should be similar across the ranges of the predictors and the fitted values.
- Heteroscedasticity is often identified when horizontal funnel shapes are observed in RVP or RVF plots

Classical Linear Regression: DSST on Age (subset of 723 individuals)

```
> model1<-lm(dsst~age,data=mridata)

> summary(model1)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 105.33955    6.15703   17.11   <2e-16 ***
age          -0.86333    0.08248  -10.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.85 on 721 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.1319,    Adjusted R-squared:  0.1307
F-statistic: 109.6 on 1 and 721 DF,  p-value: < 2.2e-16
```

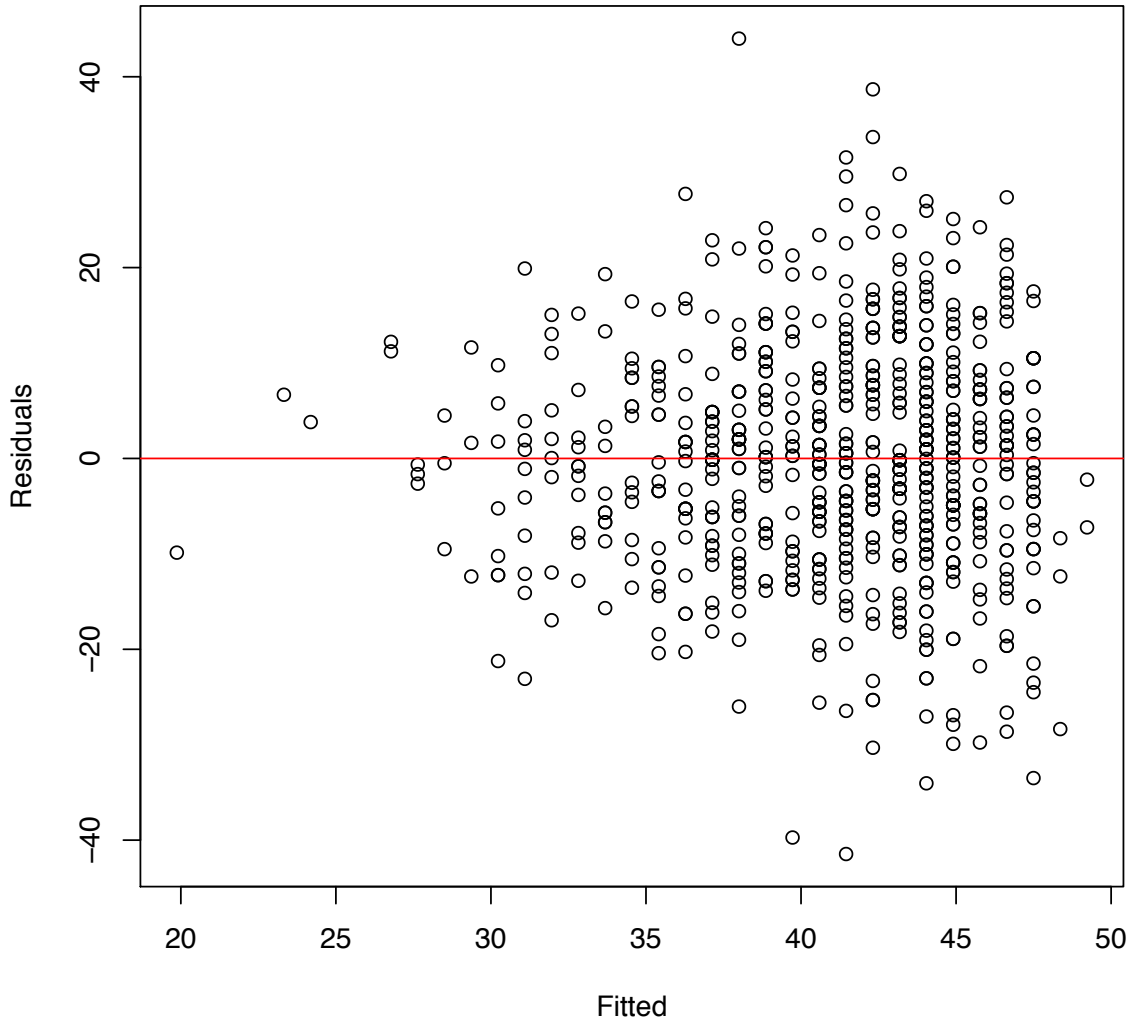
Confidence Intervals for Parameters

```
>confint.default(model1)

                 2.5 %       97.5 %
(Intercept) 93.271997 117.4070958
age         -1.024984  -0.7016759
```

Residuals vs. Fitted Values: DSST and Age

```
>plot(x=fitted(model1),y=resid(model1),ylab="Residuals", xlab="Fitted”)
>abline(h=0,col="red")
```



Violations of Constant Variance Assumption

- Nonconstant variance can sometimes be addressed using a variance-stabalizing transformation of the outcome (such as log or square root transformations), but often does not completely eliminate nonconstant variance
- A criticism with this approach is that we are now attempting to use multiple models to fit the data, which can lead to increased type-I error (multiple testing).
- Some statisticians would argue that a model should be chosen prior to seeing the data to answer the primary scientific question of interest.

Violations of Constant Variance Assumption

- Classical linear regression is based on a **Strong Null Hypothesis**
  - Distribution (mean and variance) of response identical in all groups
- Under assumptions of homoscedasticity
  - Estimated trend in mean DSST by age is an average difference of -.863 per one year differences in age (DSST lower in older)
  - 95% CI for trend: [-1.025, -0.701]

Example: Stratified Descriptives

- By scientifically relevant intervals

```
> descrip(mridata$dsst,strata=mridata$age5)
                                N     Msng  Mean     Std Dev   Min      25%      Mdn      75%      Max
mridata$dsst:  All             735    12   41.06     12.71    0.0000   32.00    40.00    50.00    82.00
mridata$dsst:    Str  (60,65]    2     0   44.50     3.536    42.00    43.25    44.50    45.75    47.00
mridata$dsst:    Str  (65,70]  175     2   45.23     12.29    14.00    38.00    46.00    53.00    74.00
mridata$dsst:    Str  (70,75]  298     1   43.00     12.45    0.0000   35.00    42.00    51.00    81.00
mridata$dsst:    Str  (75,80]  157     3   38.68     11.65    0.0000   31.00    39.00    45.75    82.00
mridata$dsst:    Str  (80,85]   67     2   33.54     9.584    15.00    26.00    32.00    42.00    53.00
mridata$dsst:    Str  (85,90]   30     3   26.67     9.767    8.000    19.00    27.00    32.50    51.00
mridata$dsst:    Str  (90,95]    5     1   33.75     5.560    28.00    29.50    34.00    38.25    39.00
mridata$dsst:    Str  (95,100]  1     0   10.00      NA      10.00    10.00    10.00    10.00    10.00
```

What if Heteroscedastic?

- What if the variances within each group are not equal?
- With t test we knew
  - Group with small sample size and higher variance ➔
    - t test that presumes equal variance is anti-conservative inference
      - Reported p values are too small
      - Reported CI is too narrow
  - Group with small sample size and lower variance ➔
    - t test that presumes equal variance is conservative inference
      - Reported p values are too high
      - Reported CI is too wide
- With linear regression similar findings for skewness of X
  - Anti-conservative inference if higher within group variance (Var[Y|X]) in outlying values of X
  - Conservative inference if lower within group variance (Var[Y|X]) in outlying values of X

Allowing for Heteroscedasticity : Robust Standard Errors.....

- Peter J. Huber, Friedhelm Eicker, and Halbert White made significant contributions to the problem of estimating standard errors of linear regression parameters when the residuals are heteroscedastic:

$$Var(e_i) \neq \sigma_e^2 \text{ for all } i$$

- Developed what is commonly referred to “robust [estimated] standard errors”, and “robust intervals”.
- In the literature, “Huber-White” or “Huber-White sandwich” estimates of the standard error correspond to ”robust estimated standard errors”
- Calculates a weighted average of standard errors across groups

Allowing for Heteroscedasticity : Robust Standard Errors

- "Robust" in the name suggests 'always good' properties, which is also misleading
  - for small sample sizes the behavior can be poor in some situations
- 'Model-agnostic' is a better name, but no-one uses it
- 'Robust' actually means minimal assumptions were made about the model, only assumptions about regularity conditions

Example: Robust Standard Errors

```
> library("sandwich")

> library(lmtest)

coeftest(model1, vcov=vcovHC(model1, "HC0"))
```

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 105.339546   5.700812  18.478 < 2.2e-16 ***
age          -0.863330   0.075405 -11.449 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Option "HC0" provides the simplest and bestknown robust standard errors, though the **sandwich** package authors didn't make it the default.

Example: Robust Standard Errors

-
  - Robust confidence intervals need to be calculated "manually"

```
> coef(model1) + sqrt(diag( vcovHC(model1, "HC0") ))%o% qnorm(c(0.025, 0.975))
                [,1]        [,2]
(Intercept) 94.16616 116.5129318
age         -1.01112  -0.7155391
```

Regression Coefficients: Estimates are the Same as Classical Regression

- Alternatively, can just use the R package **uwIntroStats** regression analysis with robust standard errors:

```
> library(uwIntroStats)
> robustmodel1 <- regress ("mean", dsst~age,data=mridata)
> robustmodel1
( 12  cases deleted due to missing values)


Call:
regress(fnctl = "mean", formula = dsst ~ age, data = mridata)

Residuals:
    Min      1Q  Median      3Q     Max
-41.453  -7.611  -0.136   7.547  44.000

Coefficients:
                Estimate  Naive SE  Robust SE    95%L      95%H         F stat    df Pr(>F)
[1] Intercept     105.3     6.157     5.709       94.13     116.5          340.49 1  < 0.00005
[2] age          -0.8633   0.08248   0.07551     -1.012    -0.7151        130.72 1  < 0.00005

Residual standard error: 11.85 on 721 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.1319,    Adjusted R-squared:  0.1307
F-statistic: 130.7 on 1 and 721 DF,  p-value: < 2.2e-16
```

Robust Parameter Estimates are the Same as Classical Regression

Robust:

```
Coefficients:
                Estimate  Naive SE  Robust SE    95%L      95%H        F stat    df Pr(>F)
[1] Intercept    105.3      6.157     5.709        94.13     116.5          340.49 1  < 0.00005
[2] age         -0.8633    0.08248   0.07551      -1.012    -0.7151        130.72 1  < 0.00005

Residual standard error: 11.85 on 721 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.1319,    Adjusted R-squared:  0.1307
F-statistic: 130.7 on 1 and 721 DF,  p-value: < 2.2e-16
```

Classical:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 105.33955    6.15703   17.11   <2e-16 ***
age          -0.86333    0.08248  -10.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.85 on 721 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.1319,    Adjusted R-squared:  0.1307
F-statistic: 109.6 on 1 and 721 DF,  p-value: < 2.2e-16
```

Regression Coefficients: Inference is different.....

Robust:

```
> coef(model1) + sqrt(diag( vcovHC(model1, "HC0") ))%o% qnorm(c(0.025, 0.975))
                [,1]        [,2]
(Intercept) 94.16616 116.5129318
age         -1.01112  -0.7155391
```

Classical:

```
> confint.default(model1)
                 2.5 %      97.5 %
(Intercept) 93.271997 117.4070958
age         -1.024984  -0.7016759
```

Robust Standard Errors

- Inference for association based on slope
  - Weak null based inference: No linear trend in summary measure across groups

- Estimated trend in mean DSST by age is an average difference of -.863 per one year differences in age (DSST lower in older)

- CI for trend: [-1.01, -0.715]

- P value < .0001 suggests mean DSST differs across age groups
  - T statistic: -11.43
  - Again, for simple linear regression $F = 130.72 = (-11.43)^2$

Example: Interpretation

"From linear regression analysis using Huber-White estimates of the standard error, we estimate that for each year difference in age between two populations, the difference in mean DSST is 0.863 points lower in the older population. A 95% CI suggests that this observation is not unusual if the true difference in mean DSST were between .715 and 1.01 points lower per year difference in age. Because the two sided P value is $P < .0005$, we reject the null hypothesis that there is no linear trend in the average DSST across age groups."

Choice of Inference

- Which inference is correct?
- Classical linear regression and robust standard error estimates differ in the strength of necessary assumptions
- If homoscedasticity holds, robust standard errors can be inefficient estimators of the standard errors of the regression parameters:
- One can also take a power hit, compared to the model-based approach;
- Under heteroscedasticity, regression analysis that assumes homoscedasticity is a biased (and inconsistent) estimate of the standard error – and gives invalid intervals, even asymptotically
- This is a typical 'bias-variance tradeoff'; here we might call it a 'robustness-efficiency' tradeoff

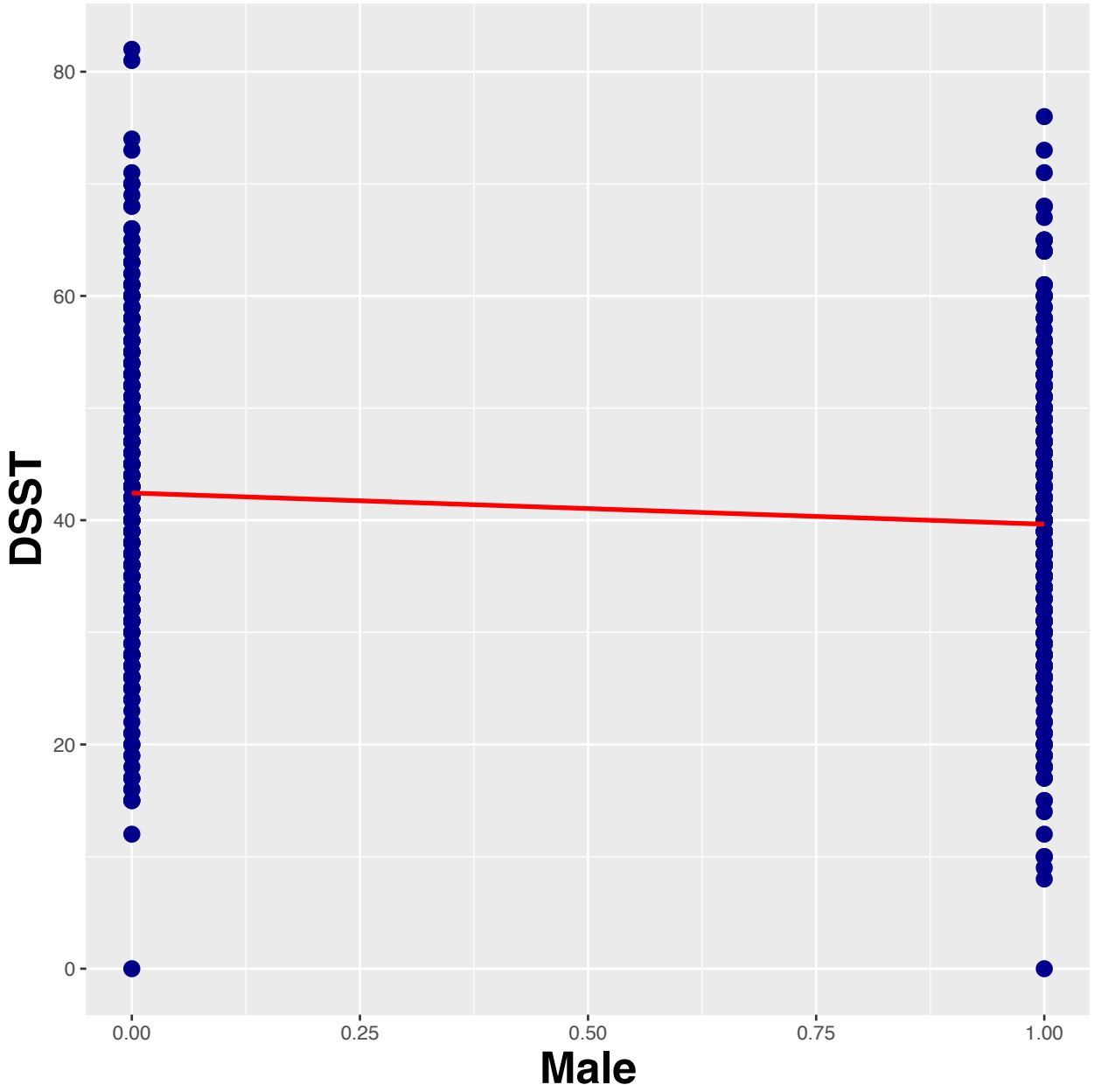
Choice of Inference

- As a rule, if all the assumptions of classical linear regression hold, it will be more precise
  - (Hence, we will have greatest precision to detect associations if the linear model is correct)
- The robust standard error estimates are, however, valid for detection of associations even in those instances

Binary Predctors: T test and with Robust SE

- Linear regression with a binary predictor
  - Classical LR: exactly the t test that presumes equal variances
  - Robust SE: approximates t test that allows unequal variances
- Consider again assessing if there is any association between mean dsst scores and gender.
- Will compare differences in mean dsst scores men and women
- Predictor variable is “male” which is an indicator is 0/1 variable (1 for males and 0 for females)
- Inference for an association will allow for heteroscedasticity

DSST by Sex (male reference): Robust



Example: DSST by Sex (male reference), ……Robust Analysis……

```
> t.test(dsst~male,data=mridata)

        Welch Two Sample t-test

data:  dsst by male
t = 2.9629, df = 720.98, p-value = 0.003148
alternative hypothesis: true difference in means is not equal to 0


 > model2<-lm(dsst~male,data=mridata)
 > coeftest(model2, vcov=vcovHC(model2, "HC0"))

 t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
 (Intercept) 42.42779    0.66715  63.596 < 2.2e-16 ***
 male        -2.78453    0.93850  -2.967  0.003107 **
```