Biost 517 / Biost 514
Applied Biostatistics II / Biostatistics II

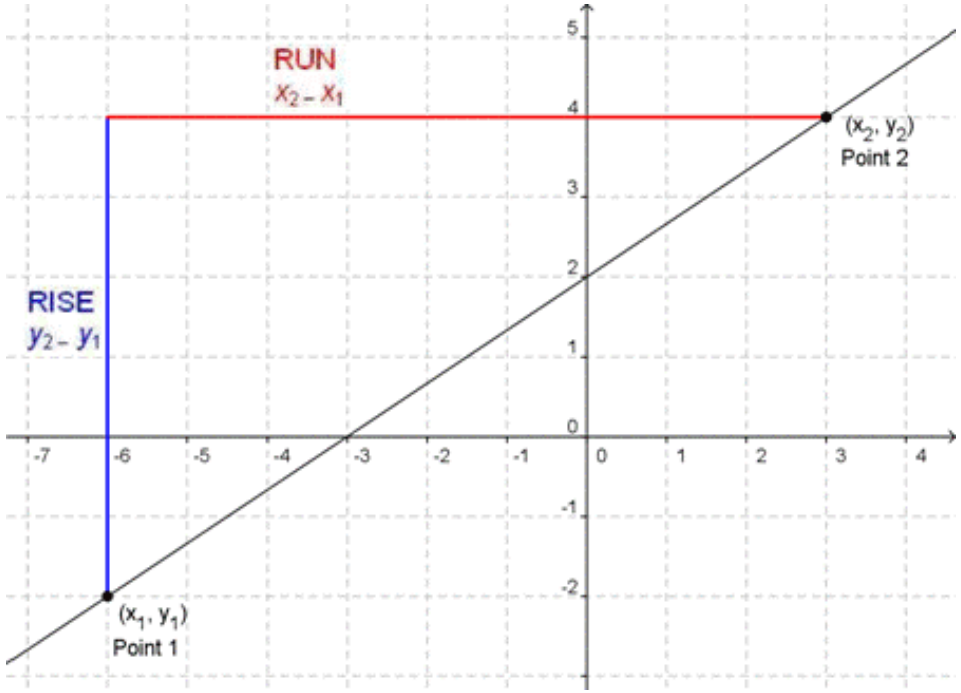
Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 17:

Linear Regression Coefficients: Interpretation and Reparameterization

Use of Straight Line Relationship

- Algebra: A line is of form $y = mx + b$
  - With no variation in the data, each value of y would lie exactly on a straight line
  - Intercept b is value of y when $x=0$
  - Slope m is difference in y per unit difference in x:
    - *slope = "rise over run"*



Regression Model Ingredients: Interpretation

- In the real world:
  - Response both within and between groups is variable
    - "Hidden variables"
    - Inherent randomness
  - The regression line describes the central tendency of the data of a scatterplot of the response versus the predictor
- Expected value (mean) of Y for a particular value of X is

$$E(Y \mid X) = \beta_0 + \beta_1 \times X$$

- Interpretation of "regression parameters"
  - Intercept β_0: Mean Y for a group with X=0
    - Often not of scientific interest or not scientifically meaningful
  - Slope β_1: A measure of linear association between Y and X
    - Difference in mean of Y across groups differing in X by 1 unit

Derivation of Interpretation

- Simple linear regression of response Y on predictor X
  - Mean for an arbitrary group derived from model
  - Interpretation of parameters by considering special cases

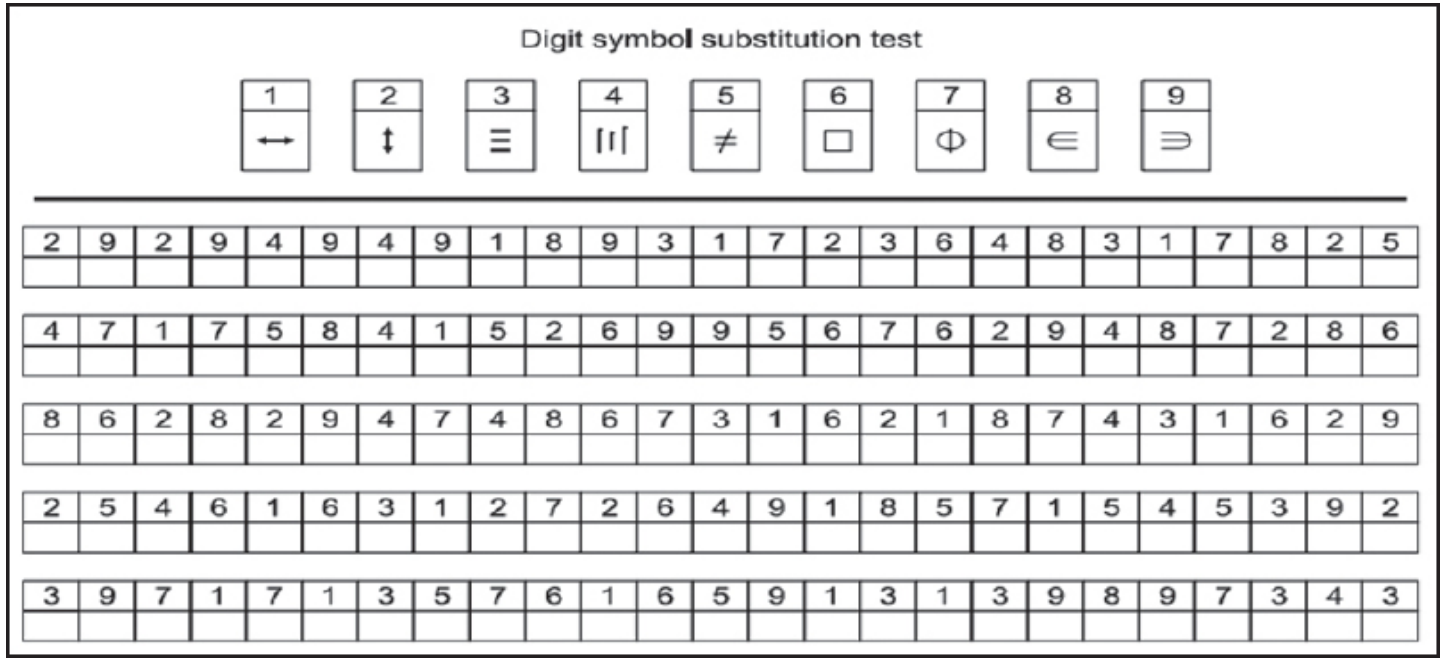
Model	$E[Y_i \mid X_i] = \beta_0 + \beta_1 \times X_i$
$X_i = 0$	$E[Y_i \mid X_i = 0] = \beta_0$
$X_i = x$	$E[Y_i \mid X_i = x] = \beta_0 + \beta_1 \times x$
$X_i = x+1$	$E[Y_i \mid X_i = x+1] = \beta_0 + \beta_1 \times x + \beta_1$

Example: Association between Cognitive Function and Age….

- Cardiovascular Health Study of elderly adults aged 65 years and older
- Data was collected at baseline on study participants for various behavioral (e.g., smoking, alcohol consumption), and functional (e.g., ability to perform routine tasks) measures
- Question: Is there an association between cognitive function, as measured by the digit symbol substitution test (DSST – a test of attention), and aging.
- Scientific question:
  - Does aging affect cognitive function?
- Statistical question: Does the distribution of cognitive function, as measured by DSST, differ across age groups?
  - What is the response? What is the predictor?

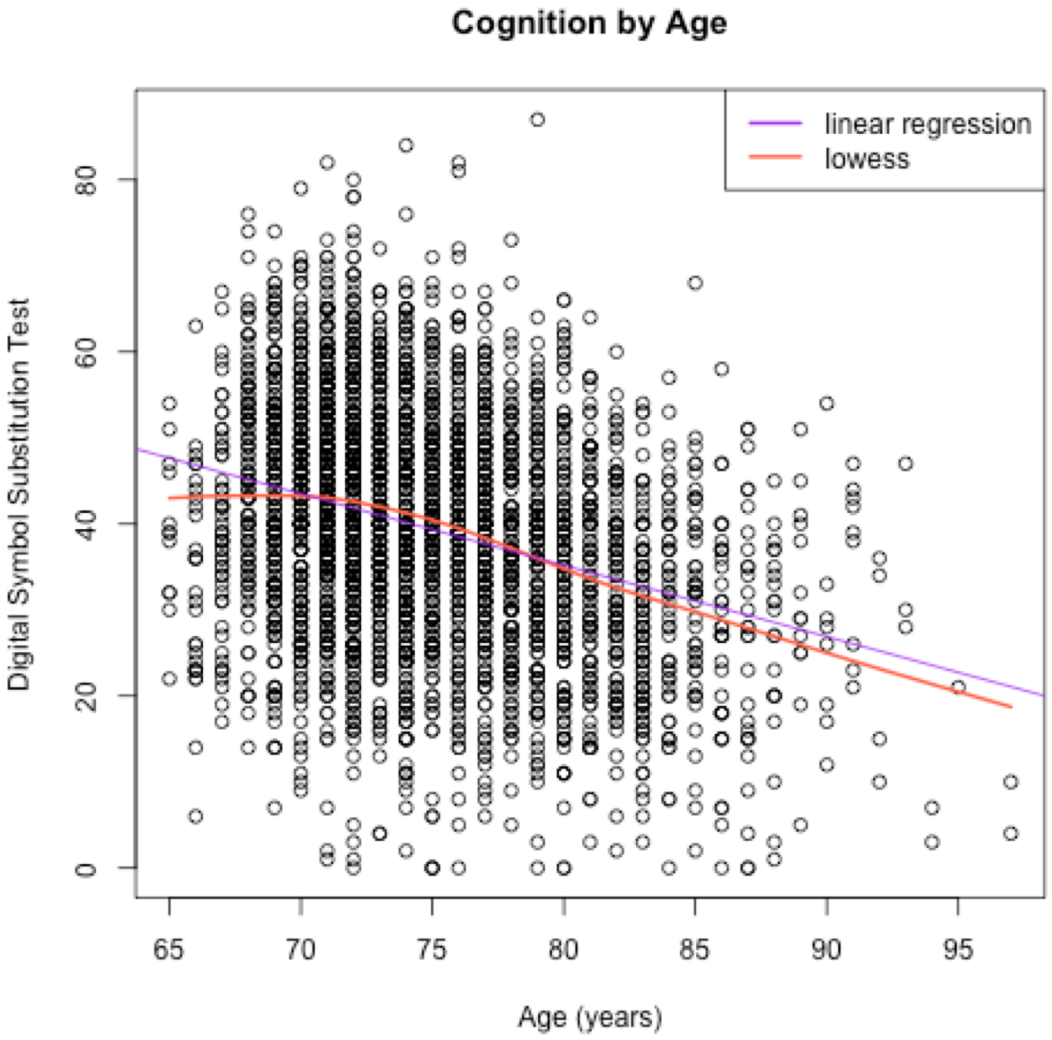
Example: Mental Function by Age

- DSST measured at baseline for 3,542 subjects
- DSST is a neuropsychological test
  - Consists of a list of digit-symbol pairs.
  - Under each digit the subject should write down the corresponding symbol as fast as possible.
  - The number of correct symbols within the allowed time (e.g. 2 minutes) is measured.



Example: Lowess, LS Line

- Locally weighted scatterplot smoothing (LOWESS) calculates a smooth curve that fits the relationship between y and x locally.



Linear Regression: presuming Homoscedasticity for DSST on Age

```
> regmodel=lm(dsst~age,data=dsstdata)
> summary(regmodel)

Call:
lm(formula = dsst ~ age, data = dsstdata)

Residuals:
    Min      1Q  Median      3Q     Max
-41.829  -8.513   0.171   8.676  50.992

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.70753    3.19394   31.84   <2e-16 ***
age          -0.83164    0.04251  -19.56   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.92 on 3540 degrees of freedom
Multiple R-squared:  0.09757,   Adjusted R-squared:  0.09732
F-statistic: 382.8 on 1 and 3540 DF,  p-value: < 2.2e-16
```

Presuming homoscedasticity: SD within each group

- Estimates of within group standard deviation
  - Within group SD is labeled “Residual standard error”
    - Estimated within group SD: 12.92
  - This presumes constant variance in age groups
    - If variance is not constant, this is an estimate of the average within group variance

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.70753    3.19394   31.84   <2e-16 ***
age          -0.83164    0.04251  -19.56   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.92 on 3540 degrees of freedom
Multiple R-squared:  0.09757,   Adjusted R-squared:  0.09732
F-statistic: 382.8 on 1 and 3540 DF,  p-value: < 2.2e-16
```

Linear Regression allowing for Heteroscedasticity: DSST on Age

```
> library(uwIntroStats)
> robustregmodel<-regress("mean",dsst~age,data=dsstdata)
> robustregmodel
( 118  cases deleted due to missing values)


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

Residuals:
    Min      1Q  Median      3Q     Max
-41.829  -8.513   0.171   8.676  50.992

Coefficients:
                 Estimate  Naive SE  Robust SE    95%L      95%H         F stat    df Pr(>F)
[1] Intercept      101.7     3.194     3.193        95.45     108.0          1014.77 1  < 0.00005
[2] age           -0.8316   0.04251   0.04241      -0.9148   -0.7485          384.60 1  < 0.00005

Residual standard error: 12.92 on 3540 degrees of freedom
  (118 observations deleted due to missingness)
Multiple R-squared:  0.09757,   Adjusted R-squared:  0.09732
F-statistic: 384.6 on 1 and 3540 DF,  p-value: < 2.2e-16
```

Deciphering Linear Regression R Output

- Estimates of within group means
  - Intercept
    - Estimated intercept: 101.71
  - Slope is labeled by variable name: “age”
    - Estimated slope: -.832
  - Estimated linear relationship:
    - Average DSST by age given by

$$E[DSST_i \mid Age_i] = 101.71 - 0.832 \times Age_i$$

  - Example: Fitted value for 70 year olds;

$$E[DSST_i \mid Age_i = 70] = 101.71 - 0.832 \times 70 = 43.47$$

Deciphering Linear Regression R Output: Interpretation of Intercept

$$E[DSST_i \mid Age_i] = 101.71 - 0.832 \times Age_i$$

- Estimated mean DSST for newborns is 101.71
  - Ridiculous estimate
    - We never sampled anyone less than 65
    - Maximum value for DSST is 100
    - Newborns would in fact (rather deterministically) score 0

- In this problem, the intercept is just a mathematical construct to fit a line over the range of our data

Linear Regression in R: Slope and Difference between Group Means

- Estimated difference in mean DSST for two groups differing by one year in age is -0.832, with older group having an average lower score
  - For 5 year age difference: 5 x -0.832 = - 4.16
  - For 10 year age difference: - 8.32
- If a straight line relationship of mean DSST and age is true, we interpret the slope as the difference in mean DSST per one year difference in age (i.e., between two groups who differ by one year in age) .
- If a straight line relationship is not true, we interpret the slope as the **average difference** in mean DSST per one year difference in age.

Example: Inference for DSST and Age Associaton

"From the linear regression analysis using Huber-White estimates of the standard error, we estimate that for each one year difference in age between two populations, the difference in mean DSST score is -0.83, with the older population group having a lower mean DSST score. A 95% CI suggests that this observation is not unusual if the true difference in mean DSST score per one year difference in age were between -0.91 and -0.75. Because the two sided P value for the mean difference is $P < .00005$, we reject the null hypothesis that there is no linear trend in average DSST across age groups."

Comments on Slope Interpretation

- I express this as a difference between group means rather than a change with aging
  - We did not do a longitudinal study

- To the extent that the true group means have a linear relationship, this interpretation applies exactly

- If the true relationship is non-linear
  - The slope estimates the “first order trend” for the sampled age distribution
  - We should not regard the estimates of individual group means as accurate for non-linear relationships

Reparameterization: Location Change

- It is possible to reparameterize our model in order to make the intercept more interpretable
  - Two models are the same if they have the same fitted values

Original model: $E[Y \mid X] = \beta_0 + \beta_1 \times X$

Recenter X: $X^* = X - 65$

Reparameterization
$$E[Y \mid X^*] = \beta_0^* + \beta_1^* \times X^*$$
$$= \beta_0^* + \beta_1^* \times (X - 65)$$
$$= (\beta_0^* - \beta_1^* \times 65) + \beta_1^* \times X$$
$$= \beta_0 + \beta_1 \times X$$
$$= E[Y \mid X]$$

Original Model: DSST on Age

```
> regmodel<-lm(dsst~age,data=dsstdata)

> summary(regmodel)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.70753    3.19394   31.84   <2e-16 ***
age          -0.83164    0.04251  -19.56   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.92 on 3540 degrees of freedom
Multiple R-squared:  0.09757,   Adjusted R-squared:  0.09732
F-statistic: 382.8 on 1 and 3540 DF,  p-value: < 2.2e-16
```

Intercept Changes with Location Change

```
> dsstdata$yrsabove65<-dsstdata$age-65
> regmodel2<-lm(dsst~yrsabove65,data=dsstdata)
> summary(regmodel2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 47.65088    0.47591  100.13   <2e-16 ***
yrsabove65  -0.83164    0.04251  -19.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.92 on 3540 degrees of freedom
Multiple R-squared:  0.09757,   Adjusted R-squared:  0.09732
F-statistic: 382.8 on 1 and 3540 DF,  p-value: < 2.2e-16
```

Interpretation of New Intercept

$$E[DSST_i \mid YrsAbove65_i] = 47.65 - 0.832 \times YrsAbove65_i$$

- Estimated mean DSST for 65 year olds is 47.65

- In this parameterization, the intercept has more relevance to our sampling scheme from an elderly population
  - But it is still not all that relevant to our question about associations between DSST and age

Reparameterization: Scale Change

- We can also reparameterize our model when the predictor is rescaled
  - Two models are the same if they have the same fitted values

Original model: $E[Y \mid X] = \beta_0 + \beta_1 \times X$

Rescale X to decades: $X^* = X/10$

Reparameterization $E[Y \mid X^*] = \beta_0^* + \beta_1^* \times X^*$

$$= \beta_0^* + \beta_1^* \times (X/10)$$

$$= \beta_0^* + (\beta_1^*/10) \times X$$

$$= \beta_0 + \beta_1 \times X$$

$$= E[Y \mid X]$$

Original Model: DSST on Age

```
> regmodel<-lm(dsst~age,data=dsstdata)

> summary(regmodel)


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.70753    3.19394   31.84   <2e-16 ***
age          -0.83164    0.04251  -19.56   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.92 on 3540 degrees of freedom
Multiple R-squared:  0.09757,   Adjusted R-squared:  0.09732
F-statistic: 382.8 on 1 and 3540 DF,  p-value: < 2.2e-16
```

Reparameterization: Slope rescaled by a factor of 10

```
> dsstdata$ageD<-dsstdata$age/10

> regmodel3<-lm(dsst~ageD,data=dsstdata)

> summary(regmodel3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.7075     3.1939   31.84   <2e-16 ***
ageD         -8.3164     0.4251  -19.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.92 on 3540 degrees of freedom
Multiple R-squared:  0.09757,   Adjusted R-squared:  0.09732
F-statistic: 382.8 on 1 and 3540 DF,  p-value: < 2.2e-16
```

Interpretation of New Slope

- Slope of the regression line: -8.32.
- One unit increase in reparameterized age now corresponds to a decade
- “We estimate that for each 10 year difference in age between two groups, the difference in mean DSST score is -8.32, with the older group having a lower mean score.”