Lesson 4
Simple Linear Regression

BIS 505b

Yale University
Department of Biostatistics

Date Modified: 3/3/2021

Goals for this Lesson

Addressing a Research Question

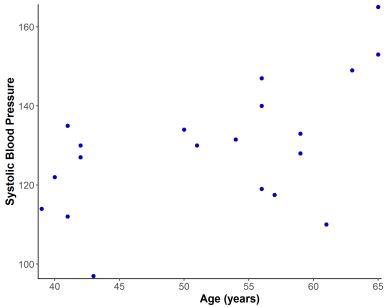
1. How to describe the linear relationship between continuous variables
2. How to estimate and predict a response for a given value of the explanatory variable
3. How to evaluate the fit and check the assumptions of a linear regression model

Contents

Progress this Unit

Relationship between Two Continuous Variables

- Graphically: Scatterplot



- Numerically: Strength and direction of the linear relationship. Estimating ρ using r.

Pearson Correlation

$$r = \frac{Cov(x, y)}{s_x \; s_y}$$

Correlation vs. Regression Analysis

- Correlation analysis is used when the goal is to assess if there is a **linear relationship** between two continuous variables (response/explanatory variable distinction not necessary)
  - Describe and estimate the direction and strength of the linear association
- Regression analysis is used when the goals are to (1) estimate the **relationship** between a predictor variable and the outcome and (2) **predict** the outcome using a set of predictors
- Must specify:
  - Response variable (y)
  - Explanatory variable(s) (x(s))

Linear Regression

Response variable		Explanatory variables
Outcome variable		Predictor variables
Dependent variable	←	Independent variables
Primary endpoint		(Exposure variable, confounders)
Y		X

Key Point

The appropriate regression model used to explore the relationship between explanatory and response variables is determined by the type of response variable analyzed

Continuous response variable → Linear regression

Population Regression Line

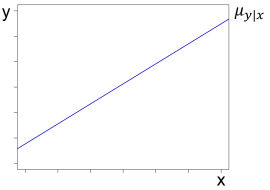
- We fit a linear regression model because we believe a line describes the relationship between x and y in the population

Population Regression Line

$$\mu_{y|x} = \alpha + \beta\, x$$

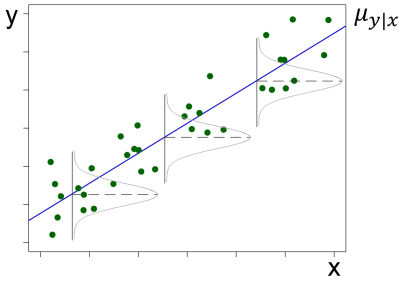
- $\mu_{y|x}$: Mean of y when independent variable $= x$; $E(Y|x)$
- x is the independent variable
- α: Mean of y when $x = 0$
- β: The slope of the line

} Parameters



Statistical Model for Linear Relationship

- Although we expect the relationship between the mean value of y and x to be a straight line, the relationship between individual values of y in the population and x will most likely not lie exactly on a straight line
  - At a particular value of x, the distribution of $Y \sim N(\mu_{y|x}, \sigma_{y|x})$



Statistical Model for Linear Relationship

- The linear relationship between y values and x values in the population is modeled as:

Simple Linear Regression Model

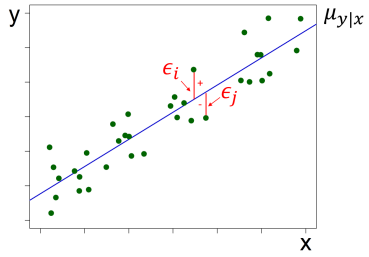
$$y = \alpha + \beta\, x + \epsilon$$

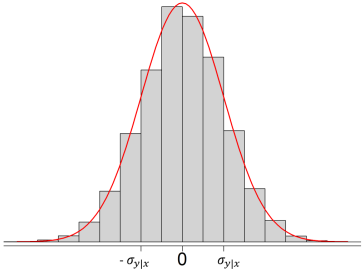
- y is the dependent variable
- x is the independent variable
- α is the y-intercept
- β is the slope

} Parameters

- ϵ is the random error $\sim N(0, \sigma_{y|x})$
- For an individual member of the population: $y_i = \alpha + \beta\, x_i + \epsilon_i$

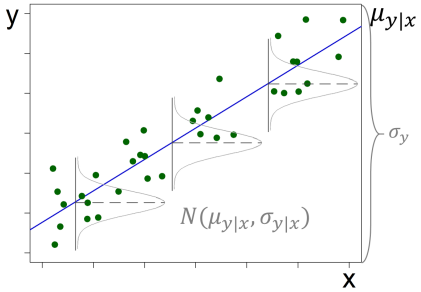
Random Error





- ϵ represent the errors (unexplained random variation)
- Distance the outcomes y lie from the population line, $\mu_{y|x}$
- Considered to be random and expect them to be independent from one another and $\epsilon \sim N(0, \sigma_{y|x})$

Variance from Regression, $\sigma^2_{y|x}$



Variance from Regression

$$\sigma^2_{y|x} = (1 - \rho^2)\sigma^2_y$$

- Standard deviation of y for a given x ($\sigma_{y|x}$) is constant
- $\sigma_{y|x} \leq \sigma_y$
  - Can make more accurate predictions of y using knowledge of x

Simple Linear Regression Assumptions

Assumptions of the Linear Regression Model

1. Linearity: The population regression line is a straight line, $\mu_{y|x} = \alpha + \beta\, x$
2. Independence: The outcomes y_i are independent
3. Approximate normality: For a given x, Y is approximately normally distributed with mean $\mu_{y|x}$ and standard deviation $\sigma_{y|x}$, $Y|x \sim N(\mu_{y|x}, \sigma_{y|x})$
4. Homoscedasticity: The standard deviation of Y given x ($\sigma_{y|x}$) is constant (the same) across values of x. Extension of ANOVA assumption.

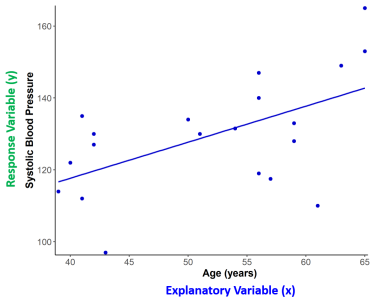
- Note: Simple linear regression (SLR): one predictor (x) variable
- Multiple linear regression (Lesson 5): more than one predictor variable

Simple Linear Regression

- The parameters of the population regression line (α and β) are estimated using a random sample of observation pairs (x_i, y_i), $i = 1, \ldots, n$

Subject i	x_i	y_i
1	x_1	y_1
2	x_2	y_2
$\vdots$		
n	x_n	y_n

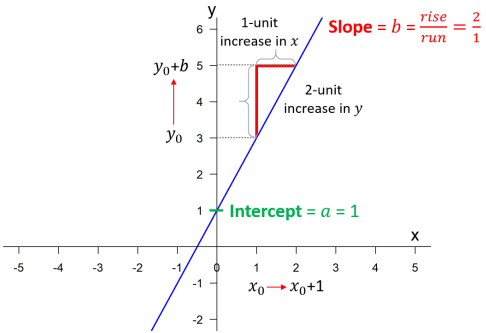
Figure: Simple random sample of $n = 20$



- **Analysis goal**: Estimate the equation of the line that best fits the data

Properties of a Line

- The estimated (fitted) regression line $\hat{y}$ has the form:



Estimated SLR Line

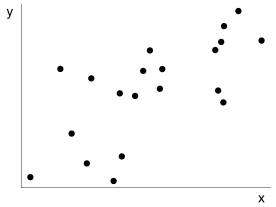
$$\hat{y} = \text{Intercept} + (\text{Slope} \times x)$$

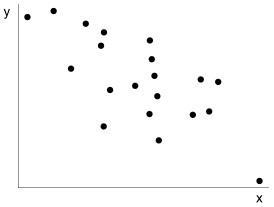
- **Intercept** (a): The estimated response when the explanatory variable = 0
- **Slope** (b): Expected change in the response when the explanatory variable increases by 1 unit

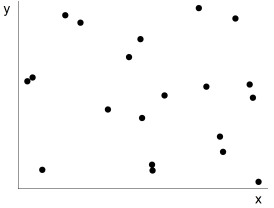
Slope

$$\hat{y} = \text{Intercept} + (\text{Slope} \times x)$$

- The slope is usually of greatest interest because it describes the relationship between the explanatory variable(s) and the response







- The slope is the expected change in y associated with a 1-unit increase in x

Slope: Linking Math to Interpretation

- When the covariate (e.g., age) equals any value (21 or 65 or x), the equation of the line is:

$$\begin{aligned}\hat{y}_0 &= \text{Intercept} + (\text{Slope} \times (x)) \\ &= a + b\,x\end{aligned}$$

- Increasing the covariate by 1 unit (22 or 66 or $x + 1$), the equation of the line is:

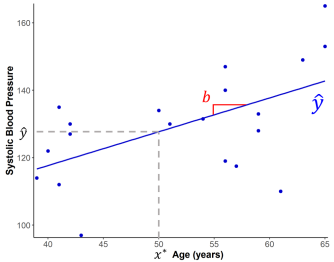
$$\begin{aligned}\hat{y}_1 &= \text{Intercept} + (\text{Slope} \times (x + 1)) \\ &= a + b\,(x + 1) \\ &= a + b\,x + b\end{aligned}$$

- The expected change in y, the response, that corresponds to a 1-unit increase in x is equal to the slope, b

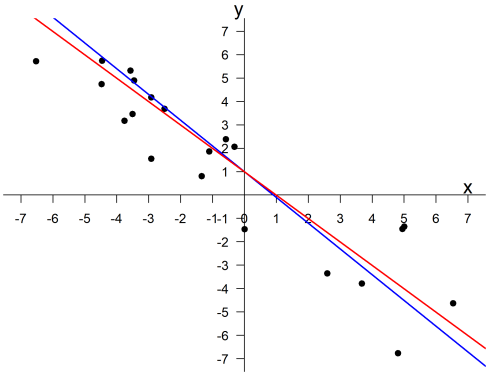
$$\hat{y}_1 - \hat{y}_0 = a + b\,x + b - (a + b\,x) = b$$

Goals of Simple Linear Regression

- Goals: Use the estimated regression line (line of best fit) to
  1. Investigate how the y variable changes in response to the x variable (slope, b)
  2. Predict or estimate the value of the response y that is associated with a fixed value of the explanatory variable $x^\star$, ($\hat{y}$)



Line of Best Fit



- There is one line that does the best job of fitting the data
- Mathematically determine the values of a and b that give the line of best fit
- Unique line that minimizes the sum of the squared vertical distances between the points and the line

Simple Linear Regression

- The method of least squares is used to find estimates of the intercept (a) and slope (b):

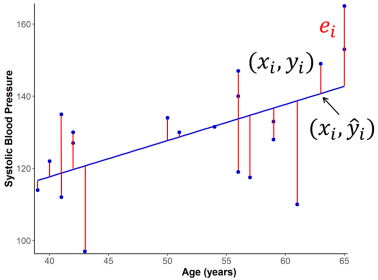
Fitted Least Squares Regression Line

$$\hat{y} = a + b\,x$$

  - a and b are statistics that estimate the population parameters (text notation: $\hat{\alpha}$ and $\hat{\beta}$)
  - $\hat{y}$ is the response that is estimated or predicted by the estimated (fitted) line

Residuals, e_i

- Even if the response and explanatory variable have a very strong relationship, do not expect to observe a *perfect* linear relationship in the sample data
  - Expect some random variation about the line

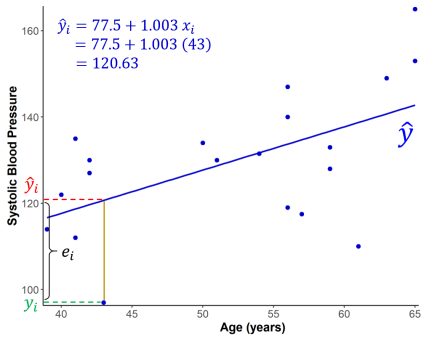


- Difference between observed and predicted responses = residuals

Residuals

$$e_i = \hat{\epsilon}_i = y_i - \hat{y}_i$$

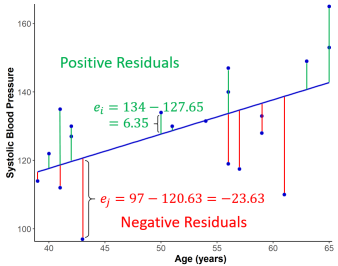
Residuals, e_i: Example



Age (x_i)	Observed (y_i)		Expected ($\hat{y}_i$)		Residual (e_i)
50	134	–	127.65	=	6.35
43	97	–	120.63	=	-23.63

- Residual, $e_i = y_i - \hat{y}_i$

The Principle of Least Squares

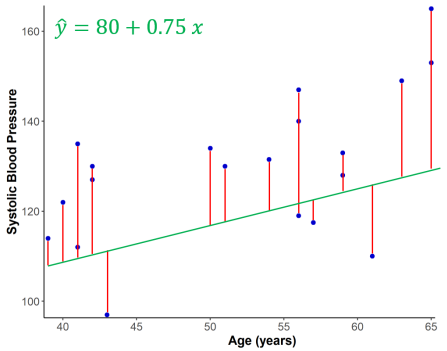


- The least squares line is the line that minimizes the sum of *squared* residuals, $e_i = y_i - \hat{y}_i$
- Error sum of squares is also known as residual sum of squares

Error Sum of Squares, SSE

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + \ldots + e_n^2$$

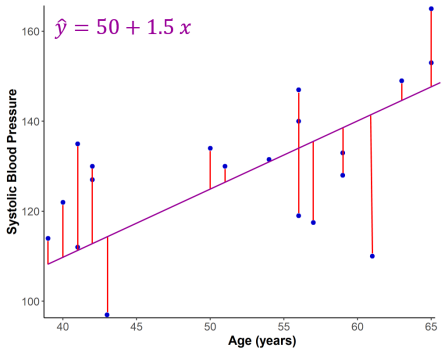
Minimizing SSE, Try 1: Example



$$SSE = \sum_{i=1}^{n} e_i^2 = 5802$$

x_i	y_i	$\hat{y}_i$	e_i	e_i^2
39	114	109.25	4.75	22.56
40	122	110	12	144
41	135	110.75	24.25	588.06
41	112	110.75	1.25	1.56
42	130	111.5	18.5	342.25
42	127	111.5	15.5	240.25
43	97	112.25	-15.25	232.56
50	134	117.5	16.5	272.25
51	130	118.25	11.75	138.06
54	131.5	120.5	11	121
56	140	122	18	324
56	147	122	25	625
56	119	122	-3	9
57	117.5	122.75	-5.25	27.56
59	128	124.25	3.75	14.06
59	133	124.25	8.75	76.56
61	110	125.75	-15.75	248.06
63	149	127.25	21.75	473.06
65	165	128.75	36.25	1314.06
65	153	128.75	24.25	588.06
				$\sum$ =5802

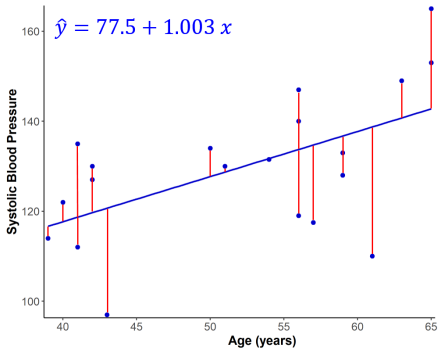
Minimizing SSE, Try 2: Example



$$SSE = \sum_{i=1}^{n} e_i^2 = 3855$$

x_i	y_i	$\hat{y}_i$	e_i	e_i^2
39	114	108.5	5.5	30.25
40	122	110	12	144
41	135	111.5	23.5	552.25
41	112	111.5	0.5	0.25
42	130	113	17	289
42	127	113	14	196
43	97	114.5	-17.5	306.25
50	134	125	9	81
51	130	126.5	3.5	12.25
54	131.5	131	0.5	0.25
56	140	134	6	36
56	147	134	13	169
56	119	134	-15	225
57	117.5	135.5	-18	324
59	128	138.5	-10.5	110.25
59	133	138.5	-5.5	30.25
61	110	141.5	-31.5	992.25
63	149	144.5	4.5	20.25
65	165	147.5	17.5	306.25
65	153	147.5	5.5	30.25
				$\sum$ =3855

Minimizing SSE, Try 3: Example



$$SSE = \sum_{i=1}^{n} e_i^2 = 3412.7$$

x_i	y_i	$\hat{y}_i$	e_i	e_i^2
39	114	116.62	-2.62	6.85
40	122	117.62	4.38	19.18
41	135	118.62	16.38	268.21
41	112	118.62	-6.62	43.86
42	130	119.63	10.37	107.62
42	127	119.63	7.37	54.38
43	97	120.63	-23.63	558.33
50	134	127.65	6.35	40.32
51	130	128.65	1.35	1.81
54	131.5	131.66	-0.16	0.03
56	140	133.67	6.33	40.09
56	147	133.67	13.33	177.74
56	119	133.67	-14.67	215.15
57	117.5	134.67	-17.17	294.84
59	128	136.68	-8.68	75.29
59	133	136.68	-3.68	13.52
61	110	138.68	-28.68	822.71
63	149	140.69	8.31	69.07
65	165	142.70	22.31	497.51
65	153	142.70	10.31	106.19
				$\sum$ =3412.7

Least Squares Slope, b

Fitted Least Squares Regression Line

$$\hat{y} = a + b\,x$$

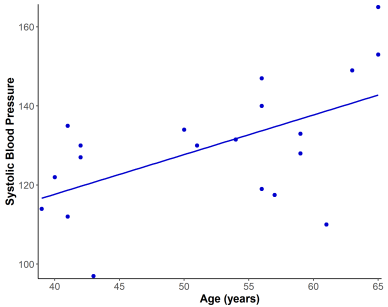
Least Squares Slope for SLR

$$b = r\left(\frac{s_y}{s_x}\right) = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Recall, $r = \dfrac{s_{xy}}{s_x\, s_y}$: Pearson correlation coefficient
- s_{xy}: Covariance between X and Y

Least Squares Slope: Example

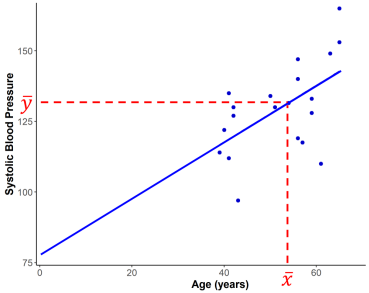
	Mean	SD (s)	Correlation (r)	Covariance (s_{xy})
y Systolic BP	129.7	16.19	0.56	82.13
x Age	52.0	9.05		



- $\hat{y} = a + b\,x$
- $b = r\left(\frac{s_y}{s_x}\right) = 0.56 \times \frac{16.19}{9.05} = 1.003$
- **Interpretation**: For every 1-year increase in age, expect systolic BP to be 1 mmHg higher, on average

Least Squares Intercept, a

- To find the intercept, need to know the slope of the line and a point that the line goes through
- The least squares line goes through the point $(\bar{x}, \bar{y})$



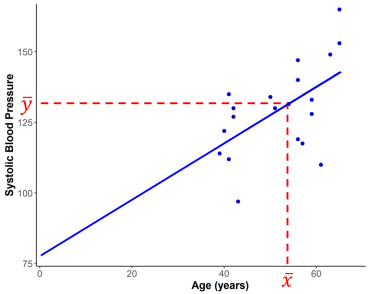
- $\hat{y} = a + b\,x \rightarrow \bar{y} = a + b\,\bar{x}$
- Solving for a:

Least Squares Intercept for SLR

$$a = \bar{y} - b\,\bar{x}$$

Least Squares Intercept: Example

	Mean	SD (s)	Correlation (r)	Covariance (s_{xy})
y Systolic BP	129.7	16.19	0.56	82.13
x Age	52.0	9.05		



- $\hat{y} = a + b\,x$
- $b = r\left(\dfrac{s_y}{s_x}\right) = 0.56 \times \dfrac{16.19}{9.05} = 1.003$
- $a = \bar{y} - b\,\bar{x} = 129.7 - 1.003(52) = 77.5$
- **Interpretation**: At age 0, expect systolic BP = 77.5 mmHg
- Caution with extrapolation

SLR: Example

R Code, SLR

```
# SLR
> mod.slr <- lm(SYSBP ~ AGE, data = fhssrs)

# Printing fitted model results
> summary(mod.slr)

Call:
lm(formula = SYSBP ~ AGE, data = fhssrs)

Residuals:
    Min      1Q  Median      3Q     Max
-28.726  -7.181   2.819   8.767  22.262

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  77.5496    18.4107   4.212 0.000524
AGE           1.0029     0.3491   2.873 0.010116

Residual standard error: 13.77 on 18 degrees of freedom
Multiple R-squared:  0.3144, Adjusted R-squared:  0.2763
F-statistic: 8.255 on 1 and 18 DF,  p-value: 0.01012
```

Estimating $\sigma^2_{y|x}$

- Assumption: Variance of y for a given x constant for all values of x
  - $\sigma^2_{y|x}$ is the residual variance in y after accounting for x

Figure: Population

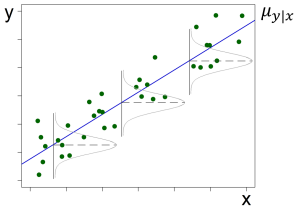
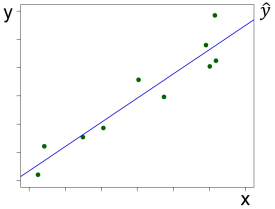
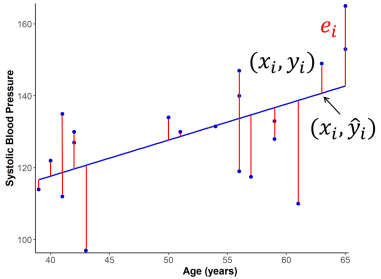


Figure: Sample



- Estimate $\sigma^2_{y|x}$, the variability of the population of responses about the population regression line, by $s^2_{y|x}$, the variability of the sample of responses y_i about the estimated regression line $\hat{y}$

SSE



SSE

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- SSE quantifies the residual variability not explained by the model (unexplained or attributed to error)
- The least squares procedure finds the line $\hat{y}$ that minimizes this error

Variance about Regression, $s^2_{y|x}$

- $\sigma^2_{y|x}$ is estimated by $s^2_{y|x}$

Variance about Regression, MSE

$$s^2_{y|x} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$$

  - Divide by SSE by (# observations) – (# parameters est) to get an estimate of $\sigma^2_{y|x}$

- $s_{y|x}$: "standard deviation from regression" or "root mean square error"

Variance about Regression: Example

R Code, ANOVA Table

```
# ANOVA table of fitted model
> anova(mod.slr)
Analysis of Variance Table

Response: SYSBP
          Df Sum Sq Mean Sq F value  Pr(>F)
AGE        1 1565.0 1565.01  8.2546 0.01012
Residuals 18 3412.7  189.59
```

- $s^2_{y|x} = \dfrac{SSE}{n-2}$
  $= \dfrac{3412.7}{18}$
  $= 189.59$ Mean Sq Residuals

- $s_{y|x} = \sqrt{\dfrac{SSE}{n-2}}$
  $= \sqrt{189.59}$
  $= 13.77$ Residual SE

Summary: Least Squares Estimation for SLR

Population

$$\mu_{y|x} = \alpha + \beta x \qquad Var(Y|x) = \sigma^2_{y|x}$$

Unknown parameters: α, β — Regression coefficients

$\sigma^2_{y|x}$ — Variance about regression

Estimates from Sample

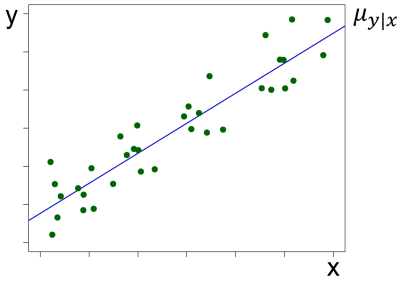
$$\hat{y} = a + b x \qquad \hat{Var}(Y|x) = s^2_{y|x} = \frac{SSE}{n-2}$$

Progress this Unit

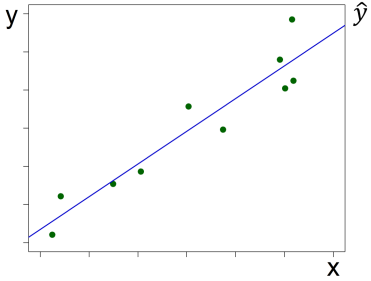
1. Simple Linear Regression
   - Defining the Linear Relationship
   - Estimating the Linear Relationship
2. Inference
   - Confidence Intervals and Hypothesis Tests
   - ANOVA Table
   - Prediction
3. Checking Assumptions
   - Diagnostics
   - Remedial Measures

Inference on the Parameters, α and β

- The least squares estimators a and b are point estimates of the regression parameters α and β
- With a different sample, the estimates would change
- a and b are random variables, each with a mean (expected value) and a variance



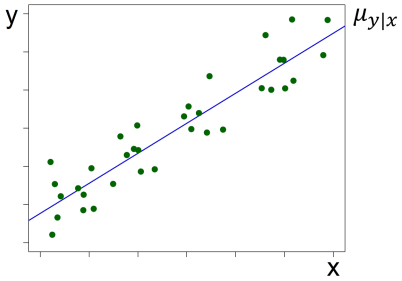
$\mu_{y|x} = 0.77 + 0.79\,x$



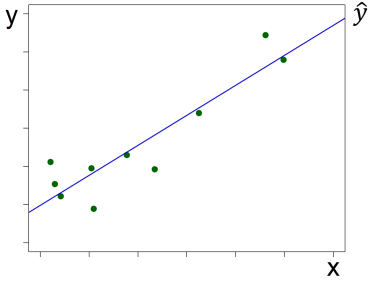
$\hat{y} = 0.35 + 0.86\,x$

Inference on the Parameters, α and β

- The least squares estimators a and b are point estimates of the regression parameters α and β
- With a different sample, the estimates would change
- a and b are random variables, each with a mean (expected value) and a variance



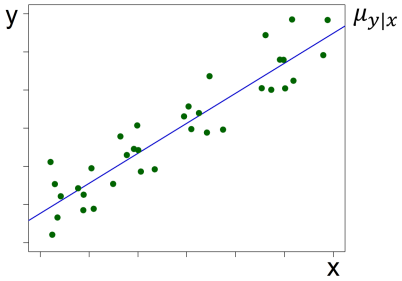
$\mu_{y|x} = 0.77 + 0.79\,x$



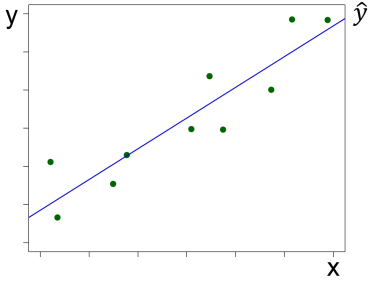
$\hat{y} = 0.98 + 0.79\,x$

Inference on the Parameters, α and β

- The least squares estimators a and b are point estimates of the regression parameters α and β
- With a different sample, the estimates would change
- a and b are random variables, each with a mean (expected value) and a variance



$\mu_{y|x} = 0.77 + 0.79\,x$



$\hat{y} = 0.85 + 0.81\,x$

Sampling Distribution of a and b

- Sampling distribution of a: $a \sim N(\alpha,\ \sigma_a)$
- Sampling distribution of b: $b \sim N(\beta,\ \sigma_b)$
  - σ_a is the standard error of the intercept a, estimated by s_a
  - σ_b is the standard error of the slope b, estimated by s_b

$$T = \frac{a - \alpha}{s_a} \sim t_{n-2} \qquad T = \frac{b - \beta}{s_b} \sim t_{n-2}$$

- Confidence intervals and hypothesis tests for the regression parameters use the t-distribution
- Text notation: $\hat{se}(a)$, $\hat{se}(b)$

Confidence Intervals for α and β

- In repeated sampling from the population, expect $100(1-\alpha)\%$ of estimated slopes to fall between:

$$\boxed{(b - t_{n-2,1-\frac{\alpha}{2}}\ s_b,\ b + t_{n-2,1-\frac{\alpha}{2}}\ s_b)}$$

Table: Confidence Intervals for SLR Parameters

Parameter	$100(1-\alpha)\%$ CI
α	$a \pm t_{n-2,1-\frac{\alpha}{2}}\ s_a$
β	$b \pm t_{n-2,1-\frac{\alpha}{2}}\ s_b$

Confidence Intervals for α and β: Example

Parameter	Estimate	Standard Error
Intercept (a)	77.5	18.41
Slope (b)	1.003	0.35
n	20	
$t_{n-2,1-\frac{\alpha}{2}} = t_{18,.975}$	2.101	

95% CI:

```
t975 = qt(1 - .05/2, df = 20 - 2)
```

- $a \pm t_{n-2,1-\frac{\alpha}{2}}\ s_a = 77.5 \pm 2.101\,(18.41) = (38.82, 116.18)$
- $b \pm t_{n-2,1-\frac{\alpha}{2}}\ s_b = 1.003 \pm 2.101\,(0.35) = (0.27, 1.74)$

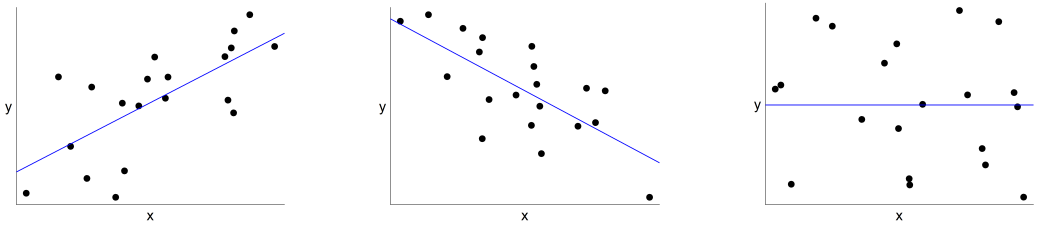
Confidence Intervals for α and β: Example

R Code, CI for Parameters

```
# 95% CIs for regression parameters
> confint(mod.slr)
                2.5 %     97.5 %
(Intercept) 38.8701842 116.229045
AGE          0.2695322   1.736252

# 90% CIs
> confint(mod.slr, level = 0.90)
                  5 %       95 %
(Intercept) 45.6243021 109.474927
AGE          0.3975899   1.608194
```

- 95% CI for α: $(38.87, 116.23)$
- 95% CI for β: $(0.27, 1.74)$

Hypothesis Test for β



- The goal of the analysis is to understand the relationship between x and y

General Hypothesis Tests for α and β

- **Hypothesis test** for the slope parameter
  - $H_0: \ \beta = 0$
  - $H_1: \ \beta \neq 0$ or $\beta > 0$ or $\beta < 0$

Test Statistic for Slope

$$t = \frac{b}{s_b}$$

- **Hypothesis test** for the intercept parameter
  - $H_0: \ \alpha = 0$
  - $H_1: \ \alpha \neq 0$ or $\alpha > 0$ or $\alpha < 0$

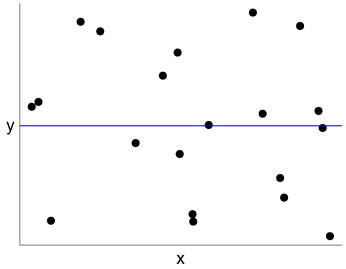
Test Statistic for Intercept

$$t = \frac{a}{s_a}$$

- Both test statistics are compared to a t-distribution with $n - 2$ degrees of freedom

Hypothesis Test for β

- Most frequently interested in tests of the slope



- Under H_0, there is no linear relationship between x and y
- $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$:
- Test statistic: $t = \dfrac{b}{s_b}$

Hypothesis Test for β: Example

- Example: Is there evidence of a significant linear relationship between systolic blood pressure and age?
- Step 1: State the hypotheses
  - $H_0 : \beta = 0$
  - $H_1 : \beta \neq 0$
- Step 2: Specify the significance level, $\alpha = 0.05$
- Step 3: Compute the appropriate test statistic
  - $t = \dfrac{b}{s_b} = \dfrac{1.003}{0.35} = 2.87$

$T \sim t_{20-2}$

Slope (b)	1.003
s_b	0.35
$t_{n-2,1-\frac{\alpha}{2}} = t_{18,.975}$	2.101

Hypothesis Test for β: Example

- Step 4: Generate the decision rule

```
t975 = qt(1 - .05/2, df = 20 - 2)
```

  - Given $\alpha = 0.05$ and a two-sided test is performed,
  - Reject H_0 if $|t| \geq t_{n-2,1-\frac{\alpha}{2}} = t_{18,.975} = t^* = 2.101$

- Step 5: Draw a conclusion about H_0

```
pval = 2 * (1 - pt(2.87, df = 20 - 2))
```

  - $t = 2.87$
  - $|t| \geq t^* \rightarrow$ Reject H_0
  - $p = 2 \times P(T \geq 2.87) = 0.0101$
  - $p \leq 0.05 \rightarrow$ Reject H_0

- Conclusion: The data provide evidence that there is a significant linear association between age and systolic blood pressure ($b = 1.003, p = 0.0101$)

Typical Presentation of Statistical Tests

- Regression output is typically presented in a table:

Parameter	Estimate	Standard Error	t-statistic	p-value
Intercept (α)	a	s_a	$t_0 = \frac{a}{s_a}$	$2 \times P(T > \lvert t_0 \rvert)$
Slope (β)	b	s_b	$t_1 = \frac{b}{s_b}$	$2 \times P(T > \lvert t_1 \rvert)$

- Where $T \sim t_{n-p}$ and p is the number of regression model parameters used to estimate the mean
- The two p-values in the table correspond to tests $H_0 : \alpha = 0$ and $H_0 : \beta = 0$, respectively

Equivalence between CI and 2-Sided Hypothesis Test for β

- In addition to giving a better estimate than a simple point estimate of the true parameter (β), the CI can be used to make inferences about the parameter
- The 95% confidence interval for β that was obtained from these data = (0.27, 1.74), representing a set of plausible values for the true slope
- CI does not include 0

Hypothesis Tests for α and β: Example

R Code, SLR

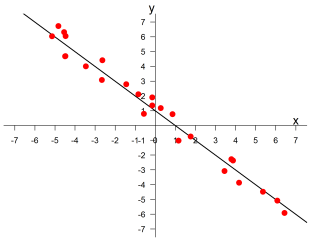
```
> mod.slr <- lm(SYSBP ~ AGE, data = fhssrs)
> summary(mod.slr)

Call:
lm(formula = SYSBP ~ AGE, data = fhssrs)

...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  77.5496    18.4107   4.212 0.000524
AGE           1.0029     0.3491   2.873 0.010116

...
> confint(mod.slr)
                2.5 %     97.5 %
(Intercept) 38.8701842 116.229045    # CI for alpha
AGE          0.2695322   1.736252    # CI for beta
```

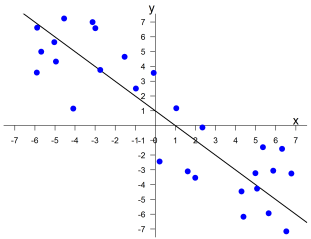
- Test of $H_0 : \alpha = 0$:
  $$t = \frac{77.55}{18.41} = 4.21$$
- Test of $H_0 : \beta = 0$:
  $$t = \frac{1.003}{0.35} = 2.87$$
- CIs do not include 0

ANOVA: Variability Explained by Linear Model

- When testing for differences between > 2 groups, F-statistic calculated from the ANOVA table
- ANOVA table is not limited to comparisons of groups
- Helpful in determining how well explanatory variable (x) accounts for variability in response variable (y) in the regression model

Variability Explained by Model





- A linear model that explains much of the variability in the response implies that the explanatory variable has a strong linear relationship with the response variable
- If a lot of the variability in the response is unexplained or attributed to error, then the model does not do an adequate job of explaining the different response values, and the explanatory variable has a weaker linear relationship with the response

Partitioning SST

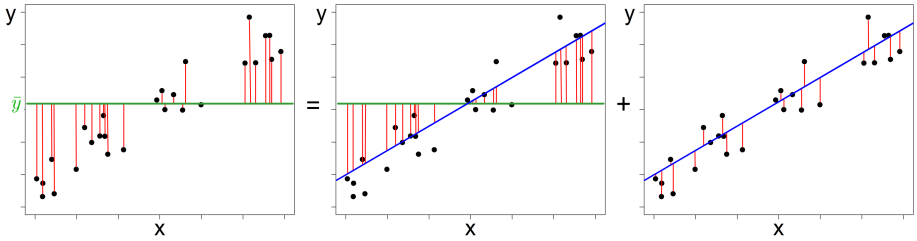
- Divide the variability of the outcome y (SST or total sum of squares) into:
  1. Variability that can be explained by the explanatory variable (i.e., the model) (SSM or model sum of squares)
  2. Variability that cannot be explained (SSE or error sum of squares)

$$SST = SSM + SSE$$

- Total variability
- Explained variability
- Unexplained variability

Partitioning SST

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad SSM = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



- Recall, $s_y^2 = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$

ANOVA Table for Simple Linear Regression

Table: ANOVA Table for Linear Regression with One Explanatory Variable

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Squares (MS)	F
Model	$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	1^*	$MSM = \frac{SSM}{df_1}$	$\frac{MSM}{MSE}$
Error	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n - 2^\dagger$	$MSE = \frac{SSE}{df_2}$	
Total	$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n - 1$		

$^*df_1$: Number of explanatory variables
$^\dagger df_2$: Total df $-$ Model df

- $s_y^2 = \frac{SST}{n-1}$

F-Test

- F-Test for Regression tests:
  - H_0: Model does not explain a significant amount of the variability in the response
  - H_1: Model explains enough variability in the response to give evidence that the explanatory variable is linearly related to the response

ANOVA F-Test Statistic

$$\mathsf{F} = \frac{MSM}{MSE}$$

- For SLR, $F \sim F(df_1 = 1, df_2 = n - 2)$ under H_0
- Reject H_0 if the test statistic F falls in the rejection region: $\mathsf{F} \geq F_{1-\alpha}(df_1, df_2)$; Rejection region always in upper tail

Another Test of $\beta = 0$

- ANOVA F-test in simple linear regression (i.e., testing whether the SLR model explains a significant amount of the variation in y) is equivalent to testing $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$
- $F(1, df_2) = (t_{df_2})^2$

ANOVA Table: Example

R Code, ANOVA Table

```
> summary(mod.slr)

Call:
lm(formula = SYSBP ~ AGE, data = fhssrs)  ...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  77.5496    18.4107   4.212 0.000524
AGE           1.0029     0.3491   2.873 0.010116
Residual standard error: 13.77 on 18 degrees of freedom
Multiple R-squared:  0.3144, Adjusted R-squared:  0.2763
F-statistic: 8.255 on 1 and 18 DF,  p-value: 0.01012

> anova(mod.slr)

Analysis of Variance Table
Response: SYSBP
          Df Sum Sq Mean Sq F value  Pr(>F)
AGE        1 1565.0 1565.01  8.2546 0.01012
Residuals 18 3412.7  189.59

> qf(1 - .05, df1 = 1, df2 = 20 - 2)   # f95
[1] 4.413873
> 1 - pf(8.25, 1, 18)                  # pval
[1] 0.01013366
```

$$F \sim F(1, 18)$$

- F $= \dfrac{MSM}{MSE} = \dfrac{1565.01}{189.59} = 8.25 > 4.41$
  - $p = P(F \geq 8.25) = 0.0101 < 0.05$
- Conclusion: We reject H_0 and conclude there is evidence that a significant amount of the total variability is explained by the model; there is a significant linear relationship between age and systolic BP
- $8.2546 = (2.873)^2$
- Test of $H_0 : \beta = 0$: $t = \dfrac{1.003}{0.35} = 2.87$

Coefficient of Determination

- Coefficient of Determination, R^2, gives the proportion of the total variation in the response variable that is explained by the model (explanatory variable)
- Assess effectiveness of using x to explain y by a linear relationship

Coefficient of Determination

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$$

- $R^2 = r^2$ in SLR, where $r =$ Pearson correlation coefficient
- $0 \leq R^2 \leq 1$
- If $R^2 = 1$, all points in sample lie on least squares line
- If $R^2 = 0$, there is no linear relationship between x and y

Coefficient of Determination: Example

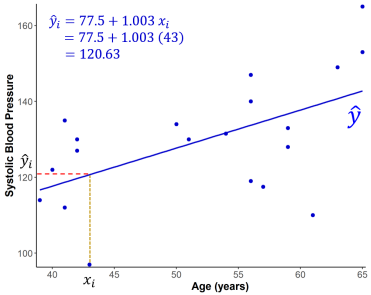
R Code, R^2

```
> summary(mod.slr)

... Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  77.5496    18.4107   4.212 0.000524
AGE           1.0029     0.3491   2.873 0.010116

Residual standard error: 13.77 on 18 degrees of freedom
Multiple R-squared:  0.3144, Adjusted R-squared:  0.2763
F-statistic: 8.255 on 1 and 18 DF,  p-value: 0.01012

> summary(mod.slr)$r.squared   # Extract R-squared alone
[1] 0.3144048

> anova(mod.slr)

Analysis of Variance Table
Response: SYSBP
          Df Sum Sq Mean Sq F value  Pr(>F)
AGE        1 1565.0 1565.01  8.2546 0.01012
Residuals 18 3412.7  189.59

> cor(fhssrs$SYSBP, fhssrs$AGE) # Pearson correlation
[1] 0.5607182
```

- $R^2 = \frac{SSM}{SST}$
  $= \frac{1565}{1565 + 3412.7}$
  $= \frac{1565}{4977.7}$
  $= 0.314 = (0.56^2)$
- **Interpretation**: The model explains 31.4% of the variability in the response variable (systolic blood pressure)

Prediction

- Beyond determining if there is a significant linear association between x and y, linear regression can be used to make predictions about y
- Using the fitted regression line, can estimate the expected or mean value of systolic BP ($\hat{y}$) for individuals with specific age values (x)

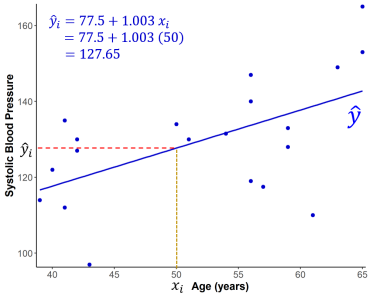


$$\hat{y} = 77.5 + 1.003\, x$$

- **Interpretation**:
  - For an individual whose age is 43, the expected systolic blood pressure is 120.63

Prediction

- Beyond determining if there is a significant linear association between x and y, linear regression can be used to make predictions about y
- Using the fitted regression line, can estimate the expected or mean value of systolic BP ($\hat{y}$) for individuals with specific age values (x)



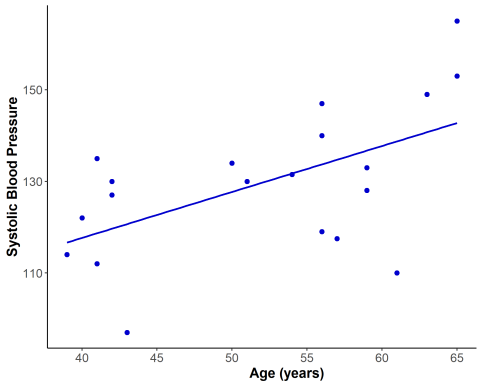
$$\hat{y} = 77.5 + 1.003\, x$$

- **Interpretation**:
  - For an individual whose age is 43, the expected systolic blood pressure is 120.63
  - For an individual whose age is 50, the expected systolic blood pressure is 127.65

Prediction: Example

R Code, Predicted Values

```
# Creating a subset of our analysis dataset
> dat <- select(fhssrs, c(RANDID, SYSBP, AGE))
# Appending column of yhat to dat
> dat$predicted <- predict(mod.slr)
> head(dat, n = 3)
   RANDID SYSBP AGE predicted
1 1609054 134.0  50  127.6942
2 1790448 130.0  42  119.6711
3 3239780 149.0  63  140.7318

# Scatterplot and fitted line
> ggplot(data = fhssrs, aes(x = AGE, y = SYSBP)) +
    geom_point() +
    geom_smooth(method = "lm", formula = y ~ x,
                se = FALSE) +
    labs(x = "Age (years)",
         y = "Systolic Blood Pressure") +
    theme_classic()  # white background (optional)
```



Confidence Intervals for Predicting a Subject's Systolic BP

- There are two types of intervals available:
  1. *Confidence interval* for the mean sytolic BP for a given age, $x^\star$, $\mu_{y|x^\star} = E(Y|x^\star)$

$$\hat{y} \pm t_{n-2,1-\frac{\alpha}{2}}\, s_{\hat{y}}$$

  2. *Prediction interval* for a single individual's systolic BP for an individual who is of a particular age, $x^\star$

$$\tilde{y} \pm t_{n-2,1-\frac{\alpha}{2}}\, s_{\tilde{y}}$$

Confidence Interval for Mean Predicted Value, $\mu_{y|x^\star}$

- A **confidence interval** for $\mu_{y|x^\star}$, the expected or mean value of y for a given $x^\star$, is centered at the point estimate $\hat{y}$

$$\hat{y} = a + b\,(x^\star)$$

- The estimated standard error of the predicted mean $\hat{y}$, $s_{\hat{y}}$, is needed to compute the margin of error of the CI

$$(\hat{y} - t_{n-2,1-\frac{\alpha}{2}}\ s_{\hat{y}},\ \hat{y} + t_{n-2,1-\frac{\alpha}{2}}\ s_{\hat{y}})$$

Table: Confidence Interval for Mean of y at $x^\star$

Parameter	$100(1-\alpha)\%$ CI
$\mu_{y\|x^\star}$	$\hat{y} \pm t_{n-2,1-\frac{\alpha}{2}}\ s_{y\|x} \sqrt{\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

Standard Error of $\hat{y}$

Standard Error of $\hat{y}$

$$s_{\hat{y}} = s_{y|x} \sqrt{\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- The term $(x^\star - \bar{x})^2$ in the numerator of the standard error:
- More confident about the mean value of y when we are closer to the mean value of x

Confidence Interval for Mean Predicted Value, $\mu_{y|x^\star}$: Example

- Example: Calculating a 95% confidence interval for the average systolic BP of individuals at age 40:
  - At $x^\star = 40$ years, $\hat{y} = 77.5 + 1.003(40) = 117.67$

n	20
$\bar{x}$	52.0
$\sum_{i=1}^{n}(x_i - \bar{x})^2$	1556
$s_{y\|x}$	13.77
$t_{n-2,1-\frac{\alpha}{2}} = t_{18,.975}$	2.101

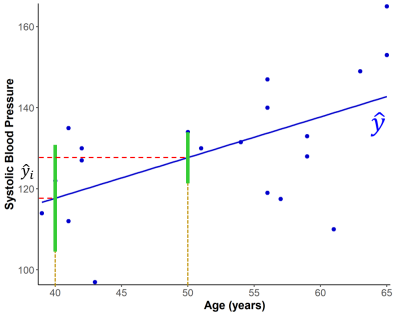
- $s_{\hat{y}} = s_{y|x}\sqrt{\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

$= 13.77\sqrt{\frac{1}{20} + \frac{(40-52)^2}{1556}}$

$= 5.198$

- $\hat{y} \pm t_{n-2,1-\frac{\alpha}{2}}\, s_{\hat{y}} = 117.67 \pm 2.101\,(5.198) = (106.75, 128.59)$

Confidence Interval for Mean Predicted Value, $\mu_{y|x^\star}$: Example

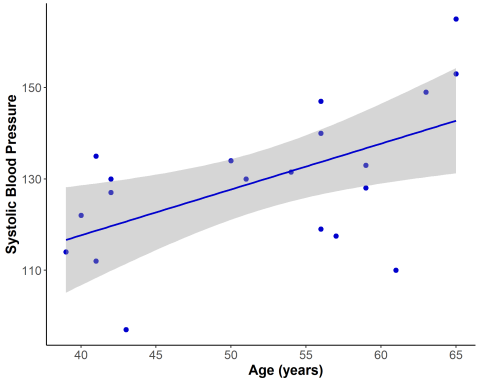


- At age $x^\star = 40$,
  $s_{\hat{y}} = 13.77\sqrt{\frac{1}{20} + \frac{(40-52)^2}{1556}} = 5.198$
  - Margin of Error $= 2.101 \times 5.198 = 10.921$
  - $(106.75, 128.59)$
- At age $x^\star = 50$,
  $s_{\hat{y}} = 13.77\sqrt{\frac{1}{20} + \frac{(50-52)^2}{1556}} = 3.157$
  - Margin of Error $= 2.101 \times 3.157 = 6.633$
  - $(121.06, 134.33)$

Confidence Interval for Mean Predicted Value, $\mu_{y|x^*}$: Example

R Code, Plotted CI for $\mu_{y|x^*}$

```
# Scatterplot, fitted line, CI for mean predicted
  values (shaded)
> ggplot(data = fhssrs, aes(x = AGE, y = SYSBP)) +
    geom_point() +
    geom_smooth(method = "lm", formula = y ~ x) +
    labs(x = "Age (years)",
         y = "Systolic Blood Pressure") +
    theme_classic()
```



Confidence Interval for Mean Predicted Value, $\mu_{y|x^*}$: Example

R Code, CI for $\mu_{y|x^*}$

```
# Specific values of x (AGE) of interest
> pred.x <- data.frame(AGE = c(40, 50))
> pred.x
  AGE
1  40
2  50

# Fitted value and lower and upper CI for mean at values of x in pred.x (default level = .95)
> predict(mod.slr, newdata = pred.x, interval = "confidence", level = 0.95)
       fit      lwr      upr   # "fit" = predicted value
1 117.6653 106.7434 128.5872
2 127.6942 121.0615 134.3270
```

Prediction Interval for Individual Observation

- A confidence interval for a mean provides information regarding the accuracy of the estimated mean value of y for a given $x^\star$
- Often, we are interested in how accurate our prediction would be for a single observation, not the mean of a group of observations
  - To answer this question, we create a confidence interval for an individual observation
  - This interval is called a prediction interval

Prediction Interval for Individual Observation

- A **prediction interval** for y for a given $x^\star$ is centered at the point estimate $\tilde{y} = \hat{y}$

$$\tilde{y} = a + b\,(x^\star) = \hat{y}$$

- The estimated standard error of $\tilde{y}$, $s_{\tilde{y}}$, is needed to compute the margin of error of the PI

$$(\tilde{y} - t_{n-2,1-\frac{\alpha}{2}}\ s_{\tilde{y}},\ \tilde{y} + t_{n-2,1-\frac{\alpha}{2}}\ s_{\tilde{y}})$$

Table: Prediction Interval for an Individual Outcome y at $x^\star$

$100(1-\alpha)\%$ CI
$\tilde{y} \pm t_{n-2,1-\frac{\alpha}{2}}\ s_{y\|x} \sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

Standard Error of $\tilde{y}$

- Variability in the prediction of a single observation contains two types of variability
  1. Variability of the estimate of the mean (confidence interval)
  2. Variability around the estimate of the mean (residual variability)

Standard Error of $\tilde{y}$

$$
\begin{aligned}
s_{\tilde{y}} &= \sqrt{s^2_{y|x} + s^2_{\hat{y}}} \\
&= s_{y|x} \sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}
\end{aligned}
$$

Prediction Interval for Individual Observation: Example

- Example: Suppose a new individual is selected from the underlying population of Framingham residents. We want to estimate his systolic BP. If his age is 40, his predicted systolic BP is:
  - At $x^\star = 40$ years, $\tilde{y} = 77.5 + 1.003(40) = 117.67$

n	20
$\bar{x}$	52.0
$\sum_{i=1}^{n}(x_i - \bar{x})^2$	1556
$s_{y\|x}$	13.77
$t_{n-2,1-\frac{\alpha}{2}} = t_{18,.975}$	2.101

- $$s_{\tilde{y}} = s_{y|x}\sqrt{1 + \frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$
$$= 13.77\sqrt{1 + \frac{1}{20} + \frac{(40-52)^2}{1556}}$$
$$= 14.72$$

- $\tilde{y} \pm t_{n-2,1-\frac{\alpha}{2}}\ s_{\tilde{y}} = 117.67 \pm 2.101\,(14.72) = (86.74, 148.59)$

Prediction Interval for Individual Observation: Example

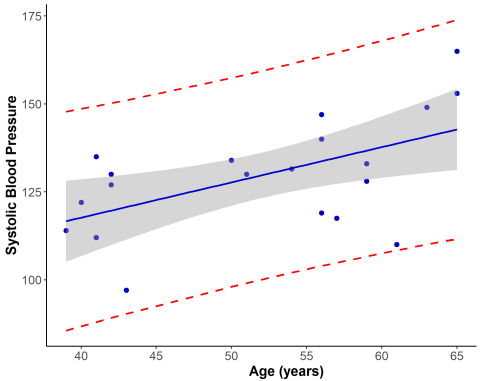
R Code, Prediction Interval

```
# Specific values of x (AGE) of interest
> pred.x <- data.frame(AGE = c(40, 50))

# Fitted value and lower and upper CI for individual prediction at values of x in pred.x
> predict(mod.slr, newdata = pred.x, interval = "prediction", level = 0.95)
       fit      lwr      upr   # "fit" = predicted value
1 117.6653 86.74394 148.5867
2 127.6942 98.01533 157.3731
```

CI for $\mu_{y|x^\star}$ vs. PI for y

- A prediction interval is similar in spirit to a confidence interval, except that
  - the prediction interval is designed to cover the random future value of y (moving target), while
  - the confidence interval is designed to cover the average (expected) value of y (fixed target) given $x^\star$
- Although both are centered at $\hat{y}$, the prediction interval is wider than the confidence interval, for a given $x^\star$ and confidence level

Prediction Interval for Individual Observation: Example



Progress this Unit

Assumptions of a Linear Regression Model

- Linear regression required several assumptions
  1. **Linearity** of the relationship between dependent and independent variables: $\mu_{y|x} = \alpha + \beta\, x$
  2. **Independence** of the errors
  3. **Errors**, ϵ, come from a Normal distribution that has
     - Mean 0
     - Constant variance $\sigma^2_{y|x}$
- If any of these assumptions is violated then inference, predictions and confidence intervals resulting from the regression model may be (at best) inefficient or (at worst) biased or misleading

Residuals

- Assumptions of the linear model can be checked by considering the residuals

$$\boxed{e_i = y_i - \hat{y}_i}$$

- Based on the assumptions of the linear regression model:
  - The mean of the residuals should equal 0
  - The histogram of the residuals should be symmetric and bell-shaped
  - The residuals should be uncorrelated, have constant variance, and exhibit a random pattern

Idea behind Residuals: Random Pattern

- When the response variable and explanatory variable have a linear relationship, the variability in the response can be partially explained by changes in the explanatory variable
- If the explanatory variable does a good job of explaining why one subject's response is different from another's, then the only reason the observed points do not fall exactly on the regression line is random variation
- Should be no left-over pattern in the residuals

Residual Plots

- To determine whether there are violations, examine **plots** of:
  1. residuals vs. predicted values $\hat{y}$, or
  2. residuals vs. x
- In SLR, pattern observed in the plot of residuals vs. $\hat{y}$ essentially the same as pattern observed in residuals vs. x

Residual Plots

R Code, Saving Residuals

```
> mod.slr <- lm(SYSBP ~ AGE, data = fhssrs)

# Selecting subset of variables
> dat <- select(fhssrs, c(RANDID, SYSBP, AGE))

# Adding predicted values (yhat) to dat
> dat$predicted <- predict(mod.slr)

# Adding residuals to dat
> dat$residuals <- resid(mod.slr)

> head(dat, n = 3)

   RANDID SYSBP AGE predicted residuals
1 1609054   134  50  127.6942  6.305784
2 1790448   130  42  119.6711 10.328920
3 3239780   149  63  140.7318  8.268188
```

Residual Plots

R Code, Residual Plots (2 different ways)

```
# Plot residuals vs. x
> plot(residuals ~ AGE, data = dat,
       ylab = "Residuals", xlab = "Age", pch = 19)
> abline(0, 0, col = "red")   # (intercept, slope of line)

# Plot residuals vs. yhat
> plot(dat$predicted, dat$residuals,
       ylab = "Residuals", xlab = "Fitted Value", pch = 19)
> abline(h = 0, col = "red")  # (h)orizontal line at 0
```

Figure: Residuals vs. x

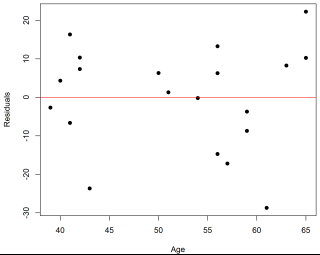
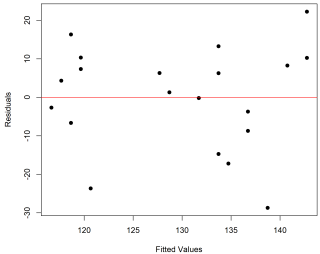


Figure: Residuals vs. $\hat{y}$



Residual Plots

- A plot of residuals helps us check for the following three model assumption violations:
  1. Violation of constant variance
  2. Violation of normality (including outliers)
  3. Violation of linearity

Violation of Constant Variance

- **Constant variance** (homoscedasticity) means that the standard deviation of the outcomes y ($\sigma_{y|x}$) is constant across all values of x
- Consequences:
  - Violations of homoscedasticity make it difficult to gauge the true standard deviation
  - Affects standard errors, hypothesis tests, CIs, and PIs
  - Does not result in biased parameter estimates
- Detection:
  1. Plot of residuals vs. predicted (fitted) values or residuals vs. independent variable
     - A fan shape indicates the variability of the residuals is not constant (heteroscedasticity)

Violation of Constant Variance

Figure: Relationship between high school GPA and university GPA

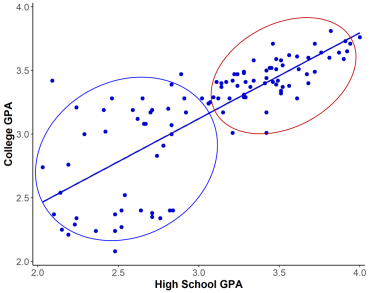
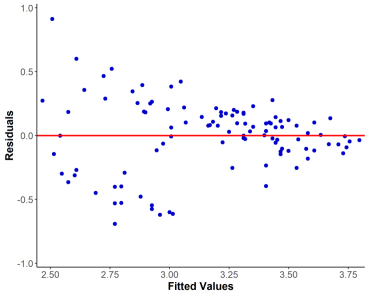


Figure: Residual plot showing violation of constant variance, GPA example



- Much more variability in university GPA in students with lower high school GPAs than in students with higher high school GPAs

Violation of Normality

- Regression requires that the errors are **normally distributed**
- Consequences:
  - Tests and CIs are robust against non-normality
  - Except for substantial non-normality that leads to outliers in the data, if the number of data points is not too small (> 10), then the linear regression will not be greatly affected even if the population distributions are skewed
- Detection:
  1. Histogram of residuals: Symmetric and bell-shaped
  2. Normal probability (Q-Q) plot: If distribution of residuals is normal, then points should align along diagonal reference line

Normality: Example

R Code, Histogram and Q-Q Plot

```
> summary(dat$residuals)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-28.726  -7.181   2.819   0.000   8.767  22.262
> ggplot(data = dat, aes(x = residuals)) +       # Histogram of residuals with overlayed density
    geom_histogram(aes(y = stat(density)),
                   breaks = seq(-30, 25, by = 5), col = "black", fill = "gray", closed = "left") +
    geom_density(alpha = 0.1, fill = "blue") + labs(x = "Residuals", y = "Density")
> ggplot(data = dat, aes(sample = residuals)) +  # Normal Q-Q plot of residuals
    stat_qq(size = 2, alpha = 0.5) + stat_qq_line(size = 0.75, color = "blue")
```

Figure: Histogram of residuals

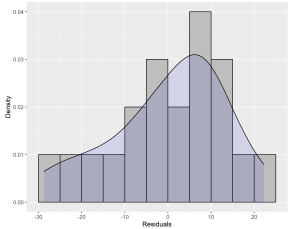
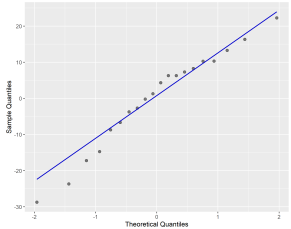
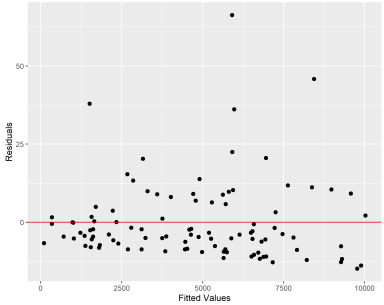


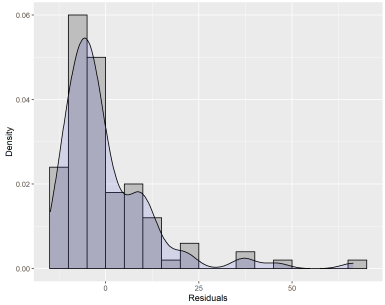
Figure: Normal Q-Q plot of residuals

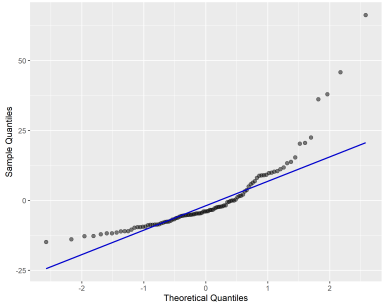


Violation of Normality

- Right skewed

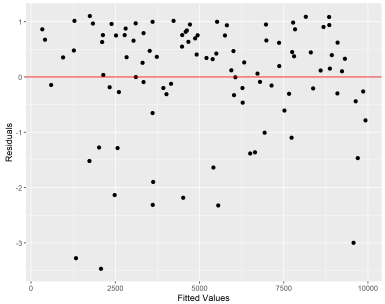


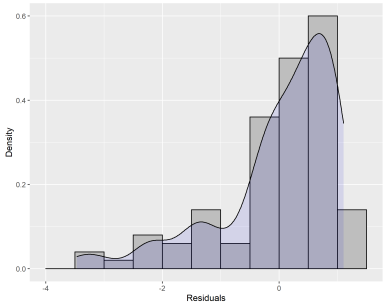


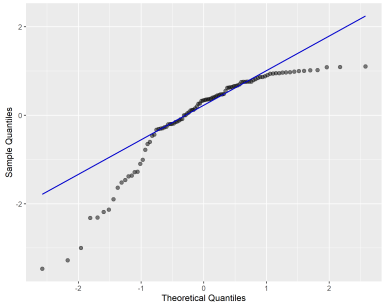


Violation of Normality

- Left skewed







Outliers

- Sometimes the error distribution is skewed by the presence of a few large **outliers**
- If a value is extreme in the vertical direction, residual will be extreme
- Consequences:
  - Since parameter estimation is based on the minimization of squared error, a few extreme observations can exert a disproportionate influence on parameter estimates
  - If a value is extreme in the x-direction, this value is said to have high leverage
  - Points with high leverage *could* have big impact on fitted model (high influence)
  - Outliers tend to increase MSE: more difficult to reject H_0
- Detection:
  1. Scatterplot of data
  2. Residual plots

Outliers

Figure: Outlying value seen in scatterplot

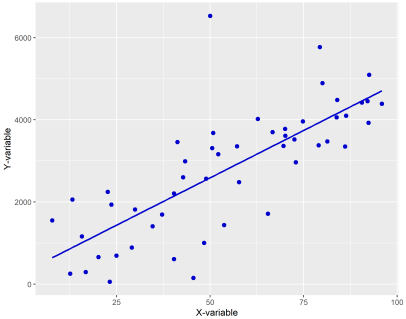
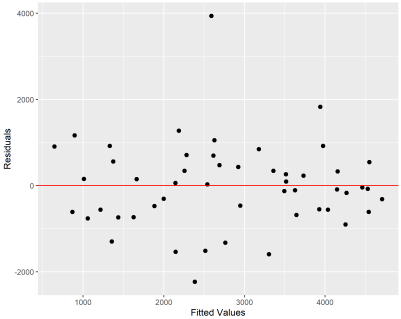


Figure: Outlying value seen in residual plot



High Influence Observation

Figure: Value with high leverage and influence

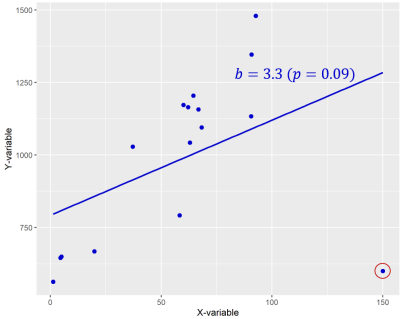
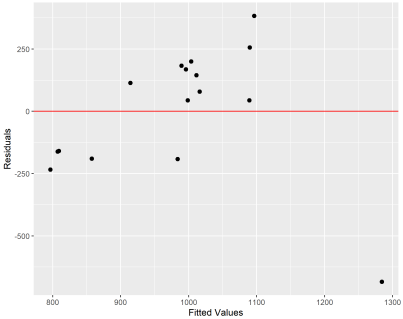


Figure: Outlying value seen in residual plot



High Influence Observation

Figure: Value with high leverage and influence

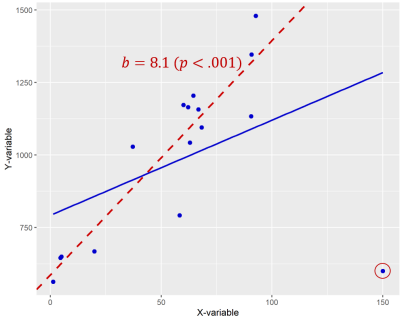
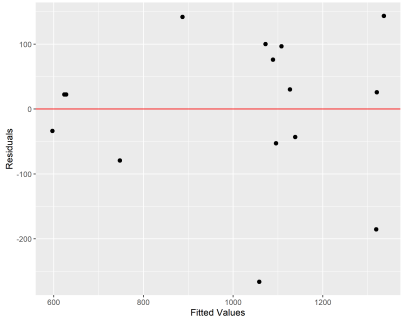


Figure: Value with high leverage and influence



Violation of Linearity

- Simple linear regression uses a line to summarize the relationship between y and x
- If a line does a poor job at describing the relationship, then the assumption of **linearity** has been violated
- Consequences:
  - If you fit a linear model to data which are nonlinearly related, your predictions are likely to be seriously in error, especially if you extrapolate beyond the range of the sample data
- Detection:
  1. Scatterplot of data
  2. Residual plots: Points should be symmetrically distributed around a horizontal line. Look for evidence of a bowed pattern.
  - Plot of residuals vs. x allow you to investigate each predictor separately (becomes important in multiple regression)

Violation of Linearity

Figure: Nonlinear relationship between x and y

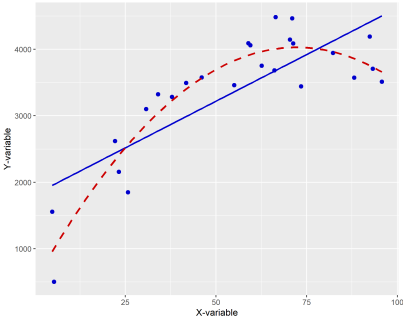
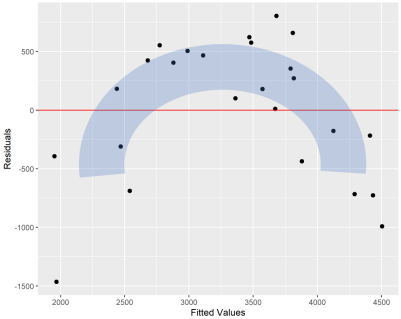


Figure: Curved pattern in residual plot



- Residual plot does not appear randomly scattered (curved)
- Shows points are below the fitted line at the ends and above the fitted line in the middle

Summary of Diagnostics: SLR

- The following **plots** are important tools for checking the assumptions of a linear regression model:

Summary of Diagnostic Plots

1. Scatterplot of y vs. x
   - Look for linearity/curvature, outliers, non-constant variance
2. Plot of residuals vs. $\hat{y}$ or x
   - Look for linearity/curvature, outliers, non-constant variance
3. Q-Q plot and/or histogram of residuals
   - Check for normality

Remedial Measures

- We have discussed how to detect departures from the important assumptions of our linear model
- Now to discuss some **remedial measures** to address these violations
- Focus on data transformations (y and/or x)

Transformations

Variable Transformations

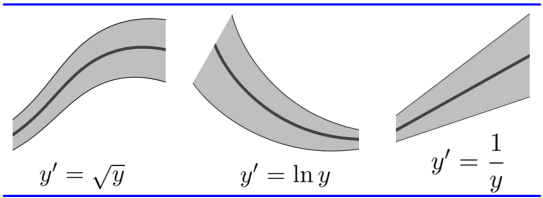
- Purpose of the transformations:
  1. Stabilize variance (reduce heteroscedasticity)
  2. Normalize residuals
  3. Linearize regression model
  4. Reduce the influence of unusual or extreme values
- In general, transforming the y variable corrects problems with error terms (and may help nonlinearity)
- Transforming the x variable primarily corrects nonlinearity

- Some transformations serve more than one purpose. For example, a transformation that linearizes the relationship may also normalize residuals.
- Once a transformation has been applied, re-check assumptions of linear model

Violation of Homoscedasticity

- **Violation of homoscedasticity (heteroscedasticity)**
  - A transformation of the y-variable may be used to stabilize the variance (reduce heteroscedasticity)
  - Most common are square root, log, and inverse transformations
  - A simultaneous transformation of x might also be required to maintain linearity
  - Advanced technique: Weighted least squares

Figure: Regression patterns in data suggesting unequal variance; transformations of y



Violation of Homoscedasticity: Example

- Example: Relationship between age (x) and plasma level of polyamine (y) for a sample of 25 healthy children, Fitted model: $\hat{\text{Plasma}} = 13.5 - 2.2\,\text{Age}$

Figure: Scatterplot

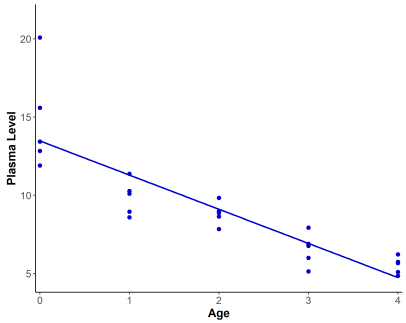
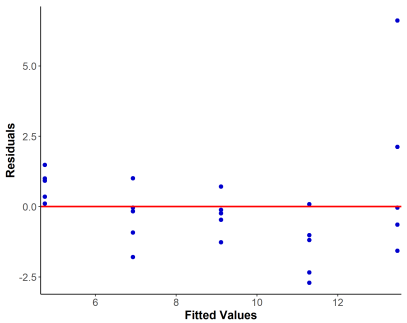


Figure: Heteroscedasticity in residuals



- Scatterplot suggests a log-transformation of y might help

Transformations: Example

R Code

```
> plasmadata <- read.csv("plasma_data.csv", header = TRUE)
> plasmadata$logplasma <- log(plasmadata$Plasma)
> mod.logplasma <- lm(logplasma ~ Age, data = plasmadata)
> mod.logplasma2 <- lm(log(Plasma) ~ Age, data = plasmadata)

# Example
> y <- c(1, 2, 3)
> y^2
[1] 1 4 9
> y**2
[1] 1 4 9
> sqrt(y)
[1] 1.000000 1.414214 1.732051
> log(y)           # log = ln
[1] 0.0000000 0.6931472 1.0986123
> log(2.718)       # e approx 2.718
[1] 0.9998963
```

- `log(y)` function is natural log
- $\sqrt{y}$: `sqrt(y)`
- y^2: `y^2` or `y**2`

Violation of Homoscedasticity: Example

- Improvement when modeling: $\log$ Plasma $= \alpha + \beta$ Age $+\epsilon$
  - Fitted model: $\log \hat{\text{Plasma}} = 2.6 - 0.24$ Age

Figure: Scatterplot using $\log$ Plasma

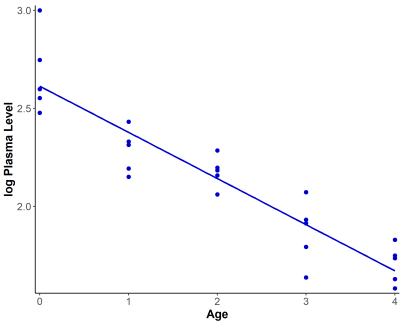
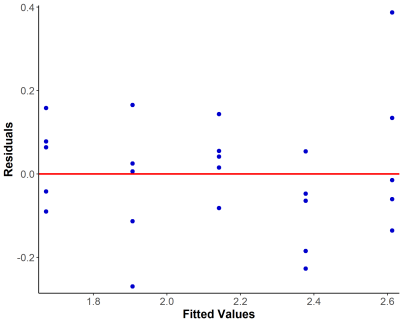


Figure: Improvement in variability



- One year increase in age associated with 0.24 reduction in **log** plasma levels

Violation of Normality

- **Violation of normality**
  - Transformations can often:
    - Reduce skew
    - Pull outliers in

Need to correct	Strength of transformation	Mathematical function
Positive Skew	Stronger	$\frac{1}{y}$
	Mild	$\log y$
	Mild	$\sqrt{y}$
Negative Skew	Mild	y^2
	Stronger	y^3

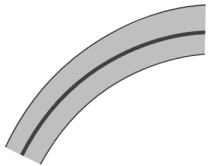
$\log y$ can only be applied to positive y

$\sqrt{y}$ can only be applied to non-negative y

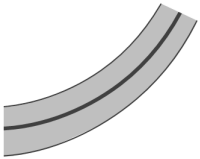
Violation of Linearity

- **Nonlinear relationships**
  - A transformation of the x-variable or the addition of a polynomial term (quadratic, cubic) may help restore a linear relationship between x and y
  - Linearity: Assumption that change in mean value of y associated with a 1-unit increase in x is the same regardless of level of x
  - Nonlinearity: The effect of x on y depends on the value of x

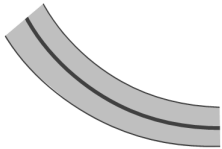
Figure: Regression patterns in data suggesting non-linear relationship; transformations of x



$x' = \ln x$ or $x' = \sqrt{x}$



$x' = x^2$ or $x' = \exp(x)$



$x' = \dfrac{1}{x}$ or $x' = \exp(-x)$

Violation of Linearity: Example

- Example: Relationship between birth rate per 1000 population and gross national income per capita for 203 countries, 2011

Figure: Nonlinear relationship between x and y

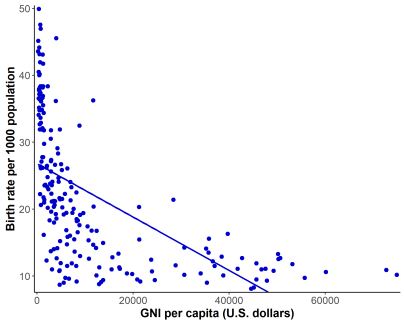
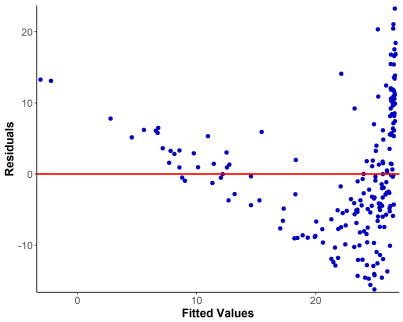


Figure: Pattern in residual plot



- Residual scatterplot shows a leftover pattern in the residuals

Transformation of x: Example

- The relationship between birth rate and the $\log$ of GNI appears much more linear
  - Fitted model: $\widehat{\text{BirthRate}} = 72.9 - 6.01 \log(\text{GNI})$

Figure: Relationship between $\log(x)$ and y

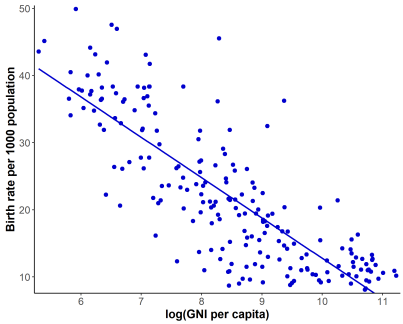
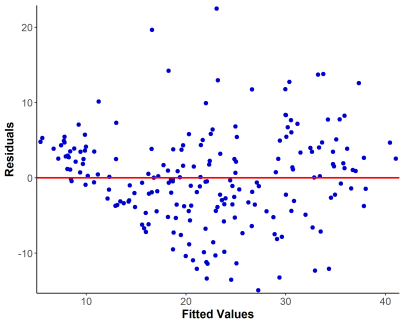


Figure: Residual plot



Summary of Transformation Remedial Measures

Summary of Steps

- If the residuals appear to be normal with constant variance, and the relationship is linear, then proceed with the regression model
- If the residuals appear to be normal with constant variance, but the relationship is non-linear, try transforming x to make the relationship between y and x linear
- If the residuals display non-constant error variance or lack of normality, try transforming y. If that stabilizes the variance but there is no longer linearity, try transforming x as well to restore linearity.

- Transformations might simultaneously fix problems with residuals and linearity (and normality)
- Once a transformation has been applied, re-check assumptions of the linear model

Drawbacks of Using Transformations

- Interpretation of regression model involves transformed variables and not the original variables themselves
- Relationship of the transformed variables to the original variables may be difficult to present/confusing
- Transformation may not be able to rectify all of the problems in the original data

Outliers

- **Outliers**
  - When the outlier is the result of an error in measuring or recording an observation, removal of the point improves the fit of the regression line
  - Removal of an outlier should only be done after after careful investigation, and perhaps discussion
  - Data point should not be discarded if it is not an error, or can't be shown that the data point belong to a different "population"
  - Care must be taken not to throw away points that are valid

Cautions about Regression

- Concluding that x and y are linearly related ($\beta \neq 0$) does not imply a cause and effect relationship between x and y
- When predicting future values, the conditions affecting y and x should remain similar for the prediction to be trustworthy
- Beware of extrapolation: Predicting y for $x^\star$ far outside the range of x in the data. The relationship may not hold outside of the observed x-values.

Journal Example: Extrapolation

Journal Example

NATURE | VOL 431 | 30 SEPTEMBER 2004 | www.nature.com/nature

Andrew J. Tatem*, Carlos A. Guerra*, Peter M. Atkinson†, Simon I. Hay*‡

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

The 2004 Olympic women's 100-metre sprint champion, Yuliya Nesterenko, is assured of fame and fortune. But we show here that — if current trends continue — it is the winner of the event in the 2156 Olympics whose name will be etched in sporting history forever, because this may be the first occasion on which the race is won in a faster time than the men's event.

The Athens Olympic Games could be viewed as another giant experiment in human athletic achievement. Are women narrowing the gap with men, or falling further behind? Some argue that the gains made by women in running events between the 1930s and the 1980s are decreasing as the women's achievements plateau[1]. Others contend that there is no evidence that athletes, male or female, are reaching the limits of their potential[1,2].

In a limited test, we plot the winning times of the men's and women's Olympic finals over the past 100 years

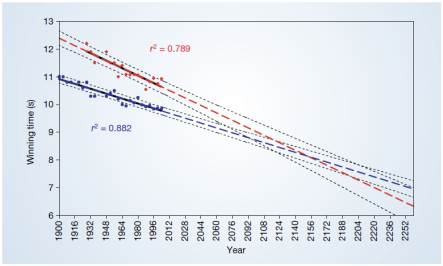


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

Link to article

Interpreting the Final Model

- Does your model make sense intuitively, based on what the data represent?
- Is the association positive or negative as expected?
- What is the interpretation of the estimated slope?
- What *other variables* may be involved?

Bonus Material: Logged Dependent Variable, Log-linear Model

- In the log-linear model, $\log \hat{y} = a + b\,x$, the interpretation of the estimated slope, b, is that a 1-unit increase in x results in an expected change in $\log y$ of b units

x:	$\log \hat{y}_0 = a + b\,x$	$\hat{y}_0 = e^{a+b\,x}$
$x+1$:	$\log \hat{y}_1 = a + b\,(x+1) = a + b\,x + b$	$\hat{y}_1 = e^{a+b\,x+b}$

$$\log \hat{y}_1 - \log \hat{y}_0 = \log\left(\frac{\hat{y}_1}{\hat{y}_0}\right) = a + b\,x + b - (a + b\,x) = b$$

- Through the nice properties of logs, can interpret coefficients in terms of percent change

$$\hat{y}_1 = e^b\,\hat{y}_0$$

  - $b > 0$: 1-unit increase in x associated with $100(e^b - 1)\%$ increase in y
  - $b < 0$: 1-unit increase in x associated with $100(1 - e^b)\%$ decrease in y
  - $100(1 - e^{-.24}) = 21\%$: One year increase in age associated with expected 21% reduction in plasma levels

Bonus Material: Logged Independent Variable

- In the model, $\hat{y} = a + b \log x$, the interpretation of the estimated slope, b, is that a 1-unit increase in $\log x$ results in an expected change in y of b units
- Instead, consider a 1% increase in x:

$$\begin{aligned} x = 1: \quad & \hat{y}_0 = a + b \log 1 \\ x = 1.01: \quad & \hat{y}_1 = a + b \log 1.01 \end{aligned}$$

$$\hat{y}_1 - \hat{y}_0 = a + b \log 1.01 - (a + b \log 1) = b \log \left(\frac{1.01}{1} \right) = b \log 1.01$$

- $\log 1.01 \approx \frac{1}{100} = 0.01$ $\qquad$ $\log 1.10 \approx \frac{1}{10} = 0.10$
  - $b > 0$: 1% increase in x associated with $b/100$ increase in y (absolute increase)
  - $b < 0$: 1% increase in x associated with $b/100$ decrease in y (absolute decrease)
  - $b = -6.01$: 1% increase in GNI is associated with a $6.01/100 = 0.06$-unit decrease in birth rate
    10% increase in GNI is associated with a $6.01/10 = 0.6$-unit decrease in birth rate