MATH 503: Mathematical Statistics

Lecture 10: Linear Regression

Reading: C&B Sections 11.3,12.1-12.2.4

Kimberly F. Sellers

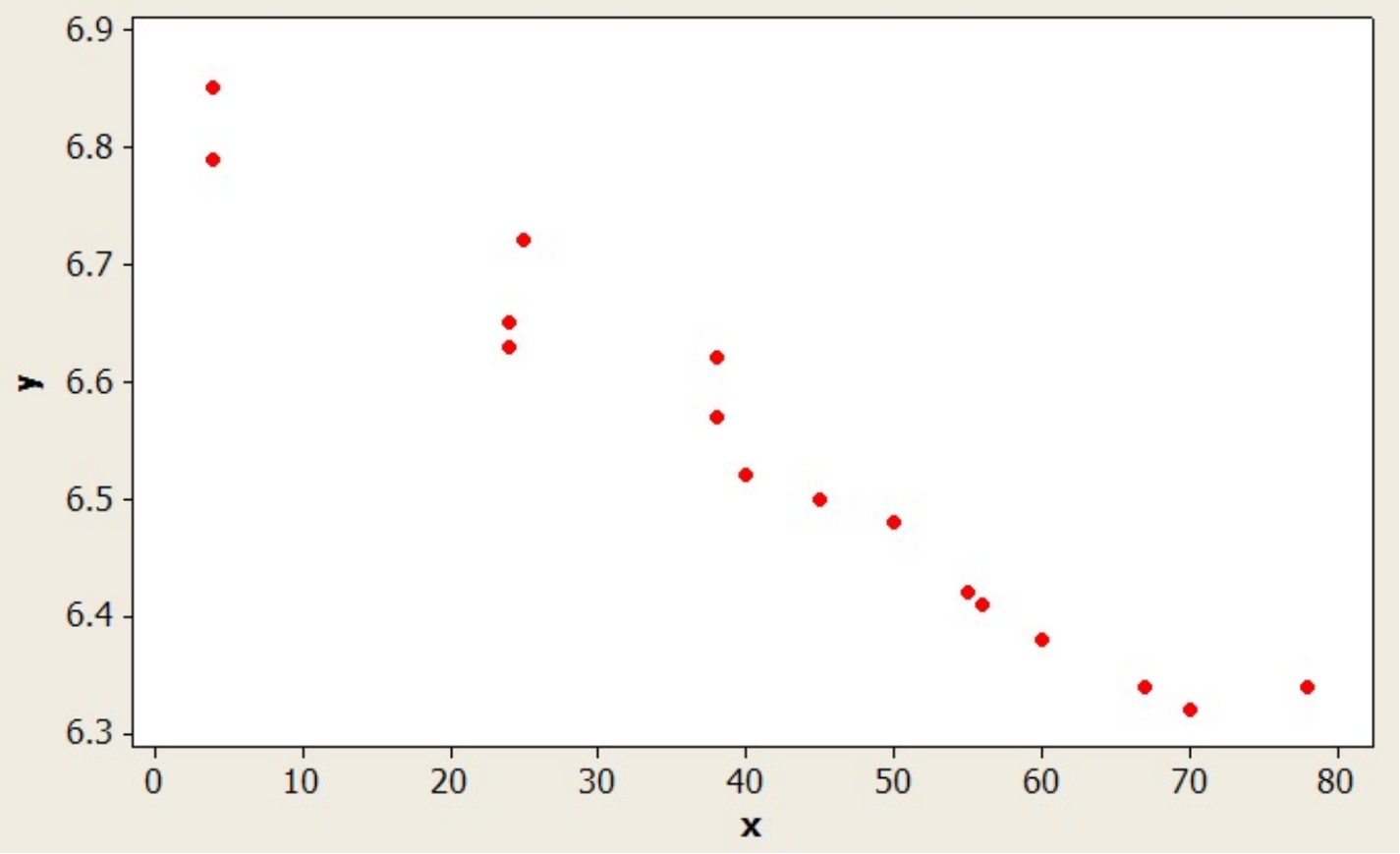
Department of Mathematics & Statistics

Today's Topics

- What's the point?
- Method of least squares
- Best linear unbiased estimators (BLUEs)
- Simple regression model assumptions
- Point estimation
- Sampling distributions
- Inference and testing

What's the point?

Given the values (*x*,*y*), we want to see if there is a relationship between *X* and *Y*.

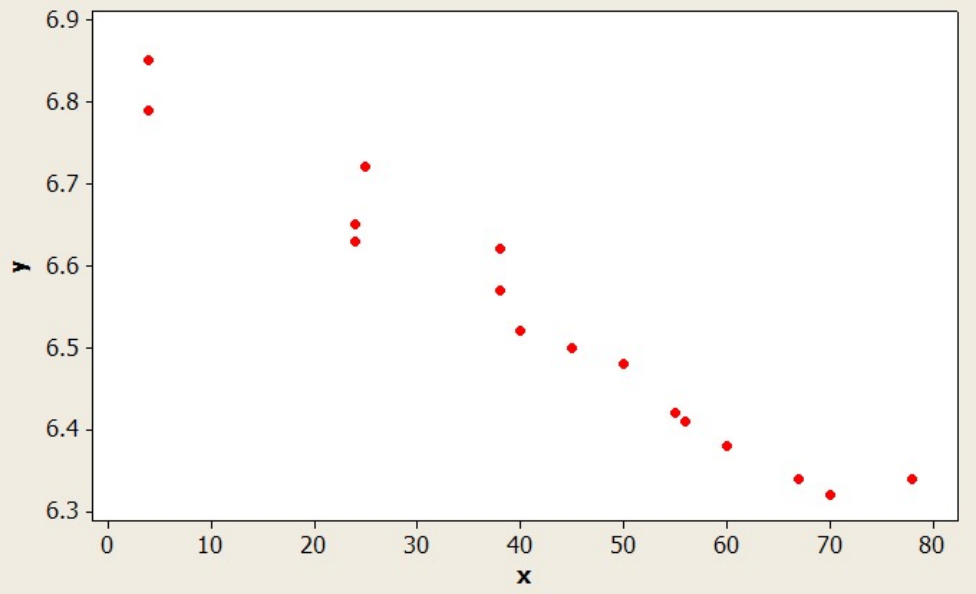


What's the point? (cont.)

- Simple (linear) regression refers to regression with one predictor variable
- "Linear" regression ⇒ linear in the parameters

- Which of the following are linear models?
  - $Y_i = \alpha + \beta x_i + \epsilon_i$
  - $\log(Y_i) = \alpha + \beta x_i^2 + \epsilon_i$
  - $Y_i = \alpha + \beta^2 x_i + \epsilon_i$

What's the point?



- For simple regression, we want to find a line $\hat{Y}_i = \hat{\alpha} + \hat{\beta} x_i$ that best describes the relationship displayed in the scatterplot.
- We may think of the value $\hat{Y}_i = \hat{\alpha} + \hat{\beta} x_i$ as predicting Y_i, and then define the ith residual as $r_i = Y_i - \hat{Y}_i = Y_i - (\hat{\alpha} + \hat{\beta} x_i)$. To judge the quality of the fit of the line, examine the r_i's.

Notation

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2$$

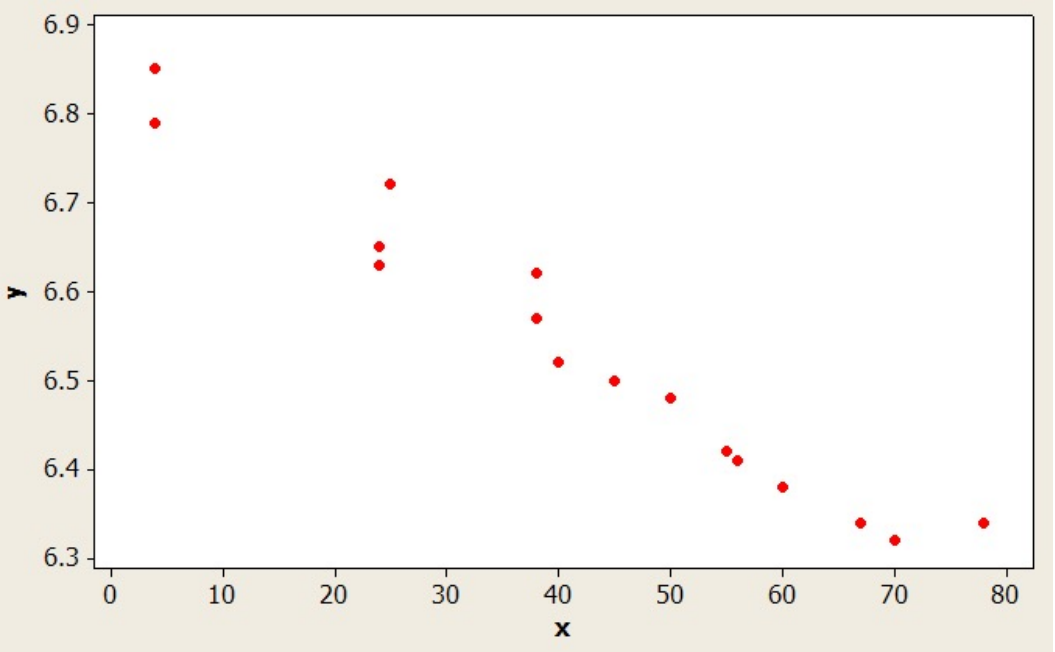
$$S_{yy} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$S_{xy} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

Method of Least Squares

The method of least squares chooses the line that has the smallest residual sum of squares, $RSS = \sum_{i=1}^{n} r_i^2$

Why is the least squares approach reasonable?



- Least squares is only one way to fit lines, and it has good and bad properties
  - Good: easily computable and have some nice mathematical properties
  - Bad: heavily influenced by outliers

Another Reasonable Approach

- Use horizontal distances instead of vertical distances

- The resulting line would be

$$x^* = a^* + b^* y$$

where $b^* = \frac{S_{xy}}{S_{yy}}$ and $a^* = \bar{x} - b^* \bar{y}$

- Re-expressing the line as a function of y on x implies

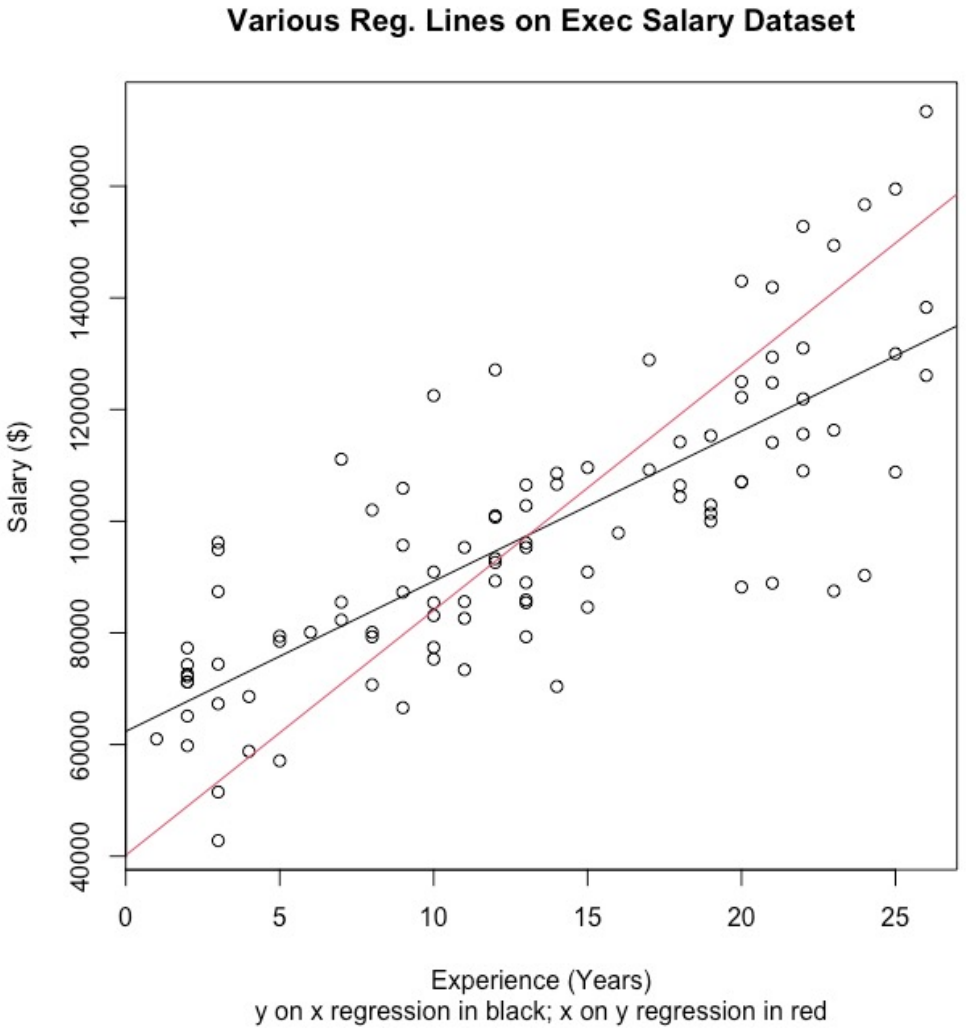
$$\hat{y} = \frac{-a^*}{b^*} + \frac{1}{b^*} x$$

What's the difference?

- If the two lines were the same, then the slopes would be equal, i.e. $b/_{(1/b^*)} = 1$
- In actuality,

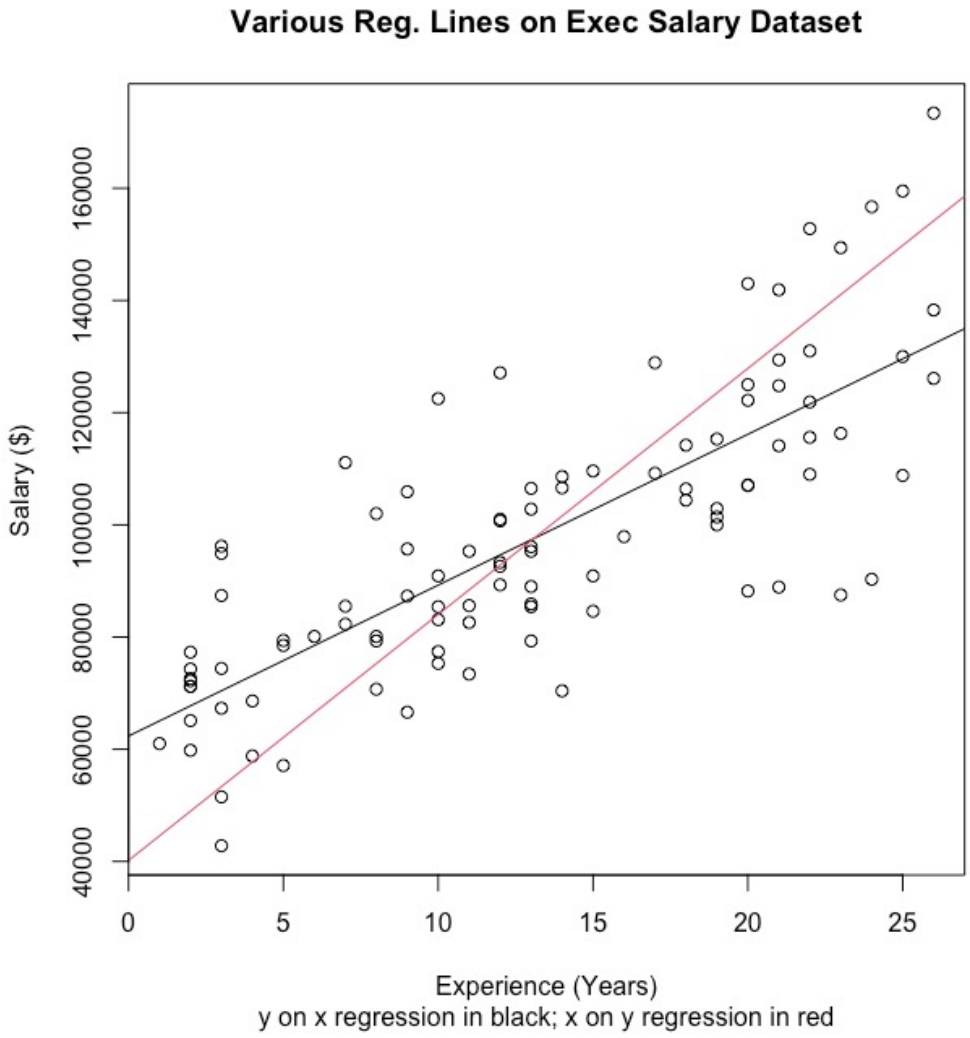
$$b/_{(1/b^*)} = bb^* = \frac{(S_{xy})^2}{S_{xx}S_{yy}} \leq 1$$

- Problem when there is no distinction between predictor and response variables



R code (if you were wondering)

```
> exec <- read.table("MATH 651/data/
    ExecSalary.txt")
> View(exec)
> plot(exec$Experience,
    exec$Salary,xlab="Experience
    (Years)", ylab="Salary ($)")
> exec.lm1 <- lm(Salary~Experience,
    data=exec)
> exec.lm2 <- lm(Experience~Salary,
    data=exec)
> abline(exec.lm1)
> abline(-exec.lm2$coefficients[1]/
    exec.lm2$coefficients[2],
    1/exec.lm2$coefficients[2],col=2)
> title(main="Various Reg. Lines on
    Exec Salary Dataset", sub="y on x
    regression in black; x on y
    regression in red")
```



Best Linear Unbiased Estimators (BLUEs)

- Setup:
  - Assume x_i's known & fixed
  - y_i's observed values from uncorrelated rv's Y_i's
  - Consider model $Y_i = \alpha + \beta x_i + \epsilon_i$, where ϵ_i's uncorrelated rv's with E(ϵ_i)=0 and Var(ϵ_i)=σ^2 unknown
  - Goal: determine estimates for α, β
- Restrict choice of estimators to class of linear estimators (i.e. of the form $\sum_{i=1}^{n} d_i Y_i$ where d_i's known & fixed)

  "Unbiased" is self-explanatory

  "Best" refers to estimator with smallest variance

Example

What specifications must be in place to satisfy a BLUE of β?

Result (Casella & Berger, Lemma 11.2.7)

Let $(v_1, \ldots, v_k)$ be constants and let $(c_1, \ldots, c_k)$ be positive constants. Then, for

$$A = \{\boldsymbol{a} = (a_1, \ldots, a_k): \textstyle\sum_{i=1}^{k} a_i = 0\},$$

$$\max_{\boldsymbol{a} \in A} \left\{ \frac{\left(\sum_{i=1}^{k} a_i v_i\right)^2}{\sum_{i=1}^{k} a_i^2 / c_i} \right\} = \sum_{i=1}^{k} c_i (v_i - \bar{v}_c)^2 ,$$

where $\bar{v}_c = \frac{\sum_{i=1}^{k} c_i v_i}{\sum_{i=1}^{k} c_i}$. The maximum is attained at any $\boldsymbol{a}$ of the form $a_i = K c_i (v_i - \bar{v}_c)$ where K is a nonzero constant.

What is the BLUE of β?

Using Lemma 11.2.7 ($k = n, v_i = x_i, c_i = 1, a_i = d_i$), d_i's maximize

$$\frac{(\sum_{i=1}^{n} d_i x_i)^2}{\sum_{i=1}^{n} d_i^2} = \frac{1}{\sum_{i=1}^{n} d_i^2}$$

Among all d_i's that satisfy $\sum_{i=1}^{n} d_i = 0$, assuming d_i has the form

$$d_i = K c_i (v_i - \bar{v}_c) = K(x_i - \bar{x}), \qquad i = 1, \ldots, n$$

Thus,

$$\sum_{i=1}^{n} d_i x_i = \sum_{i=1}^{n} K(x_i - \bar{x}) x_i = K S_{xx}$$

BLUE Results

- $b = \frac{S_{xy}}{S_{xx}}$ is the BLUE of β.
- $\mathrm{Var}(b) = \sigma^2 \sum_{i=1}^{n} d_i^2 = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- Similar analysis used to determine BLUE for α
- Constants $d_1, \ldots, d_n$ must satisfy

$$\sum_{i=1}^{n} d_i = 1 \quad \text{and} \quad \sum_{i=1}^{n} d_i x_i = 0$$

Model & Distribution Assumptions

- Conditional normal model:
  1. x_is known and fixed; y_is observed from Y_is
  2. $Y_i = \alpha + \beta x_i + \epsilon_i$; $i = 1, \ldots, n$ holds (linearity of the model)
  3. $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ iid

Model & Distribution Assumptions (cont.)

- Bivariate normal model:
  1. x_is can be observed from X_is; y_is observed from Y_is
  2. $(X_i, Y_i) \sim \text{BivariateNormal}\left(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho\right)$
  3. $$E(Y \mid x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_x}(x - \mu_x)$$
     $$= \left(\mu_Y - \rho \frac{\sigma_Y}{\sigma_x}\mu_x\right) + \left(\rho \frac{\sigma_Y}{\sigma_x}\right) x$$
  4. $\text{Var}(Y \mid x) = \sigma_Y^2(1 - \rho^2)$

Point Estimation

- Inference based on point estimators, intervals, tests same for both models
- Determine MLEs for α, β, σ^2 under conditional normal model:

$$Y_i \sim \mathrm{N}(\alpha + \beta x_i, \sigma^2), \qquad i = 1, \ldots, n,$$

i.e.

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, \ldots, n$$

where $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$

Point Estimation (cont.)

- $\hat{\alpha}, \hat{\beta}$ BLUEs for $\alpha, \beta \Rightarrow$ both are unbiased
- $\widehat{\sigma^2} = \frac{1}{n} RSS$ biased for σ^2 because

$$E\left(\widehat{\sigma^2}\right) = \frac{n-2}{n}\sigma^2$$

- What is an unbiased estimator for σ^2?

Summarizing the extent to which the line fits the data: s

- Error standard deviation, σ, represents average size of the error
- σ tells how far off, on average, we expect line to be in predicting a value y at any given x_i
- Estimated by $s = \sqrt{s^2}$ where

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - (\hat{\alpha} + \hat{\beta}x_i)\right)^2$$

  called the "residual mean squared error"
- Thought of as the standard deviation of the residuals
- Provides summary of the average deviation of Y_i values from the corresponding values predicted by the line
- Has the same units as Y

Sampling Distributions Theorem

Under conditional normal regression model, sampling distributions of $\hat{\alpha}, \hat{\beta}$, and S^2 are

$$\hat{\alpha} \sim \mathrm{N}\left(\alpha, \frac{\sigma^2}{nS_{xx}} \sum_{i=1}^{n} x_i^2\right)$$

$$\hat{\beta} \sim \mathrm{N}\left(\beta, \frac{\sigma^2}{S_{xx}}\right)$$

with $\mathrm{Cov}\left(\hat{\alpha}, \hat{\beta}\right) = \frac{-\sigma^2 \bar{x}}{S_{xx}}$. Further, $\left(\hat{\alpha}, \hat{\beta}\right)$ and S^2 are independent and $\frac{(\mathrm{n}-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$.

Inference Results

$$\frac{\hat{\alpha} - \alpha}{S\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)/(nS_{xx})}} \sim t_{n-2}$$

and

$$\frac{\hat{\beta} - \beta}{S/\sqrt{S_{xx}}} \sim t_{n-2}$$

This serves as the basis for determining CIs, decision rules for hypothesis tests!

Confidence Intervals for Slope

- To compute the $100(1-\alpha)\%$ CI, use

$$\hat{\beta} \pm z_{\alpha/2} \cdot \frac{S}{\sqrt{S_{xx}}}$$

- For small samples, substitute $t_{\alpha/2,n-2}$ for $z_{\alpha/2}$. Thus, we use

$$\hat{\beta} \pm t_{\alpha/2,n-2} \cdot \frac{S}{\sqrt{S_{xx}}}$$

as $100(1-\alpha)\%$ CI for β.

Model Utility Test (t-test)

- Understanding the association (increasing or decreasing tendency) between two variables can be essential in analyses
  - Assume that *y* is approximately linear in x
  - Consider the possibility that the slope of the line is zero, i.e. $H_0: \beta = 0$ vs. $H_1: \beta \neq 0$

Testing Approaches

- There are three approaches to solve this hypothesis test:
  - Find $100(1-\alpha)\%$ CI for β: $\hat{\beta} \pm t_{\alpha/2,n-2} \cdot \frac{S}{\sqrt{S_{xx}}}$
  - Using p-value or rejection region method associated with t-statistic,

    $$t = \frac{\hat{\beta}}{S/\sqrt{S_{xx}}}$$

    and t-distribution with $n-2$ degrees of freedom
  - Use ANOVA with $F_{1,n-2}(\alpha)$: $\left(\frac{\widehat{\beta}}{S/\sqrt{S_{xx}}}\right)^2 = \frac{\widehat{\beta}^2}{S^2/S_{xx}} > F_{1,n-2}(\alpha)$

Simple Regression ANOVA Table

Source	df	Sum of Squares	Mean square	F statistic
Regression (slope)	1	$\mathrm{SS(Reg)} = S_{xy}^2/S_{xx}$	$\mathrm{MS(Reg)} = S_{xy}^2/S_{xx}$	$\mathrm{F} = \frac{\mathrm{MS(Reg)}}{\mathrm{MSE}}$
Residual	$n-2$	$\mathrm{SSE} = \sum_{i=1}^{n} \hat{\epsilon}^2$	$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-2}$	
Total	$n-1$	$\mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$		

Summarizing the extent to which the line fits the data: R^2

- R^2 interpreted as fraction of variability in Y attributable to the regression (i.e. proportion of variability in Y explained by X);

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SSReg}}{\text{SST}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

  where $\text{SSE} =$ "sum of squares due to error" $= s^2$, and $\text{SST} =$ "total sum of squares" $= \sum_{i=1}^{n}(y_i - \bar{y})^2$

- $\frac{\text{SSE}}{\text{SST}}$ is proportion of variability in Y attributable to error
- Interpreted as "proportion of variability of Y explained by X"

Coefficient of Determination (cont.)

- $0 \leq R^2 \leq 1$
- R^2 is dimensionless (no reference units)
- No universal rule as to what constitutes a “large R^2”

Example (and R code)

The prevalence of respiratory symptoms was recorded for 9 groups of subjects exposed to differing levels of dust in their work environment. Dust exposure was measured as particules/ft^3/year scaled by 10^6. The direct outcome variable is "relative risk", the ratio of symptom prevalence at a given exposure level to symptom prevalence in the absence of workplace dust.

```
> dust <-
  data.frame(exposure=c(75,100,150,350,600,900,1300,1650,2250),
  RR=c(1.10,1.05,0.97,1.9,1.83,2.45,3.70,3.52,4.16))
> summary(lm(RR ~ exposure,data=dust))
```

R Output

```
Call:
lm(formula = RR ~ exposure, data = dust)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34055 -0.13997 -0.05667  0.02818  0.66226

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  1.0359939  0.1688447   6.136    0.000474 ***
exposure     0.0015398  0.0001541   9.993    2.15e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3363 on 7 degrees of freedom
Multiple R-Squared: 0.9345,     Adjusted R-squared: 0.9251
F-statistic: 99.85 on 1 and 7 DF,  p-value: 2.150e-05
```

SAS Code

```
data symptoms;
  input exposure RR;
  cards;
75 1.10
100 1.05
150 0.97
350 1.9
600 1.83
900 2.45
1300 3.70
1650 3.52
2250 4.16
;
proc print data=symptoms; run;

proc glm data=symptoms;
 model RR=exposure;
 ;
 run;
```

SAS Output

The GLM Procedure
Dependent Variable: RR

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	11.29121174	11.29121174	99.85	<.0001
Error	7	0.79154382	0.11307769		
Corrected Total	8	12.08275556			

R-Square	Coeff Var	Root MSE	RR Mean
0.934490	14.63459	0.336270	2.297778

Source	DF	Type I SS	Mean Square	F Value	Pr > F
exposure	1	11.29121174	11.29121174	99.85	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
exposure	1	11.29121174	11.29121174	99.85	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1.035993934	0.16884466	6.14	0.0005
exposure	0.001539804	0.00015409	9.99	<.0001