

Linear Regression

Professor Jie Peng, PhD

Department of Statistics

University of California, Davis

Normal Error Model

Normal Error Model

Simple regression model + Normality assumption:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots, n,$$

where the error terms ε_i s are *independently and identically distributed (i.i.d.)* $N(0, \sigma^2)$ random variables.

MLE

Under the Normal error model:

- ▶ LS estimators $\hat{\beta}_0, \hat{\beta}_1$ are the *maximum likelihood estimator (MLE)* of β_0, β_1 , respectively.
- ▶ The MLE of σ^2 is SSE/n .

Sampling Distributions

Under the Normal error model:

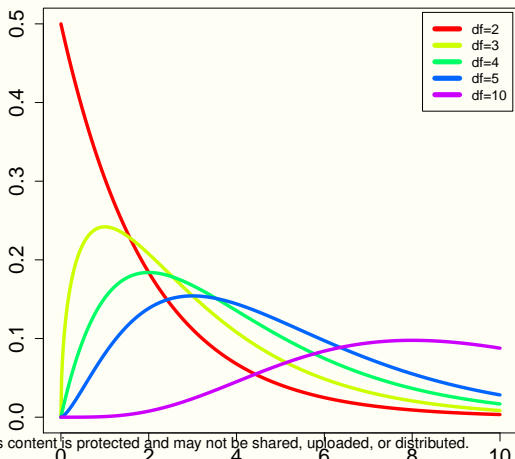
- ▶ $\hat{\beta}_0, \hat{\beta}_1$ are normally distributed:

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2\{\hat{\beta}_0\}), \quad \hat{\beta}_1 \sim N(\beta_1, \sigma^2\{\hat{\beta}_1\}).$$

- ▶ SSE/σ^2 follows a χ^2 distribution with $n - 2$ degrees of freedom, denoted by $\chi^2_{(n-2)}$.
- ▶ SSE is independent with both $\hat{\beta}_0$ and $\hat{\beta}_1$.

χ^2 Distributions

Figure: χ^2 distributions: probability density function



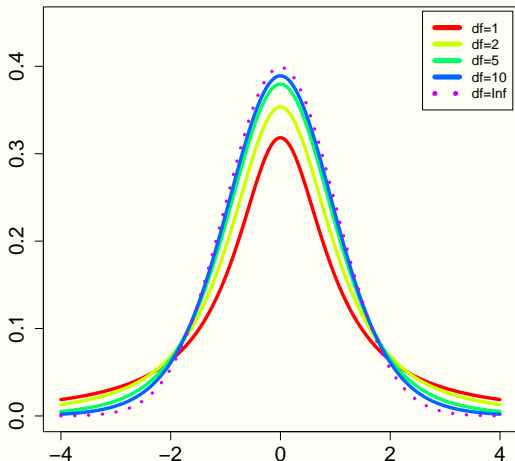
Confidence Intervals of Regression Coefficients

Pivotal Quantity

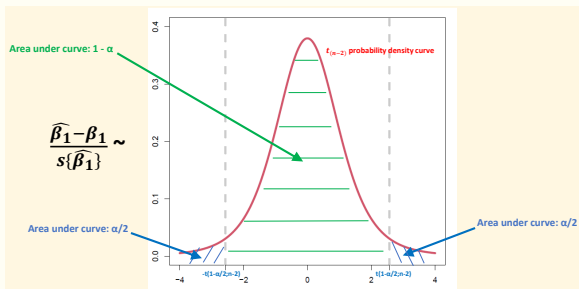
$$\frac{\hat{\beta}_1 - \beta_1}{s\{\hat{\beta}_1\}}$$

- ▶ The numerator is the difference between the LS estimator $\hat{\beta}_1$ and its mean β_1 .
- ▶ The denominator is the standard error of $\hat{\beta}_1$.
- ▶ This quantity follows a **known distribution**, $t_{(n-2)}$, t -distribution with $n - 2$ degrees of freedom.

Figure: t distributions: probability density function*



* t distribution with ∞ degrees of freedom is the standard normal $N(0, 1)$ distribution.



$$P\left(\left|\frac{\hat{\beta}_1 - \beta_1}{s\{\hat{\beta}_1\}}\right| \leq t(1 - \alpha/2; n - 2)\right) = 1 - \alpha \Rightarrow$$

$$P\left(\hat{\beta}_1 - t(1 - \alpha/2; n - 2)s\{\hat{\beta}_1\} \leq \beta_1 \leq \hat{\beta}_1 + t(1 - \alpha/2; n - 2)s\{\hat{\beta}_1\}\right) = 1 - \alpha$$

Confidence Interval

The $(1 - \alpha)100\%$ -confidence interval of β_1 :

$$\hat{\beta}_1 \pm t(1 - \alpha/2; n - 2)s\{\hat{\beta}_1\},$$

where $t(1 - \alpha/2; n - 2)$ is the $(1 - \alpha/2)100$ th percentile of $t_{(n-2)}$.

Confidence Coefficient: Accuracy

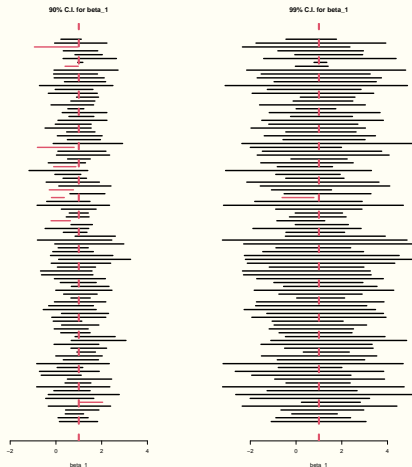
- ▶ $(1 - \alpha)100\%$ is called the *confidence coefficient* or the *confidence level*.
- ▶ Commonly used confidence coefficients are 95% ($\alpha = 0.05$), 90% ($\alpha = 0.1$), 99% ($\alpha = 0.01$).
- ▶ Confidence coefficient reflects **accuracy of the C.I.**: the larger (i.e., the smaller the α), the more accurate.

Confidence Interval Width: Precision

- ▶ The half-width: $t(1 - \alpha/2; n - 2)s\{\hat{\beta}_1\}$
- ▶ The width reflects **precision of the C.I.**: the narrower, the more precise
- ▶ Factors influencing the precision:
 - ▶ The larger the confidence coefficient (more accurate), the wider the C.I. (less precise)
 - ▶ The larger the sample size n (more data), the narrower the C.I. (more precise)
 - ▶ The larger the SE (more uncertainty), the wider the C.I. (less precise)

Simulation Experiment

Figure: C.I.s of β_1 : Left: 90% C.I.; Right: 99% C.I.



Heights

- ▶ $n = 928$, $\bar{X} = 68.316$, $\sum_{i=1}^n (X_i - \bar{X})^2 = 3038.761$, and

$$\hat{\beta}_0 = 24.54, \quad \hat{\beta}_1 = 0.637, \quad MSE = 5.031.$$

- ▶ $s\{\hat{\beta}_1\} = \sqrt{\frac{5.031}{3038.761}} = 0.0407.$

- ▶ 95%-confidence interval of β_1 :

$$\begin{aligned} 0.637 \pm t(0.975; 926) \times 0.0407 &= 0.637 \pm 1.963 \times 0.0407 \\ &= [0.557, 0.717]. \end{aligned}$$

- ▶ We are 95% confident that the regression slope is between 0.557 and 0.717.

T-test for β_1

- ▶ Null hypothesis: $H_0 : \beta_1 = \beta_1^{(0)}$, where $\beta_1^{(0)}$ is a given constant.
- ▶ **T-statistic:**

$$T^* = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{s\{\hat{\beta}_1\}}.$$

- ▶ **Null distribution:**

$$\text{Under } H_0 : \beta_1 = \beta_1^{(0)}, \quad T^* \sim t_{(n-2)}.$$

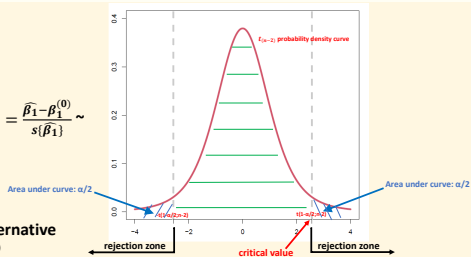
Decision Rules

At significance level α :

- ▶ *Two-sided alternative* $H_a : \beta_1 \neq \beta_1^{(0)}$: Reject H_0 if and only if $|T^*| > t(1 - \alpha/2; n - 2)$; Or equivalently, reject H_0 if and only if $\text{pvalue} := P(|t_{(n-2)}| > |T^*|) < \alpha$.
- ▶ *Left-sided alternative* $H_a : \beta_1 < \beta_1^{(0)}$: Reject H_0 if and only if $T^* < t(\alpha; n - 2)$; Or equivalently, reject H_0 if and only if $\text{pvalue} := P(t_{(n-2)} < T^*) < \alpha$.

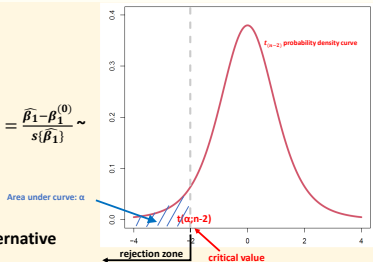
$$\text{under } H_0: T^* = \frac{\widehat{\beta}_1 - \beta_1^{(0)}}{s\{\widehat{\beta}_1\}} \sim$$

two-sided alternative
 $H_a: \beta_1 \neq \beta_1^{(0)}$



$$\text{under } H_0: T^* = \frac{\widehat{\beta}_1 - \beta_1^{(0)}}{s\{\widehat{\beta}_1\}} \sim$$

Left-sided alternative
 $H_a: \beta_1 < \beta_1^{(0)}$



Heights

Test whether there is a linear association between parent's height and child's height at significance level $\alpha = 0.01$.

- ▶ $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$.
- ▶ $T^* = \frac{\hat{\beta}_1 - 0}{s\{\hat{\beta}_1\}} = \frac{0.637}{0.0407} = 15.7$.
- ▶ **Critical value:** $t(1 - 0.01/2; 928 - 2) = 2.58$. Since the observed $|T^*| = |15.7| > 2.58$, reject the null hypothesis at level 0.01.
- ▶ **Pvalue:** $P(|t_{(926)}| > |15.7|) \approx 0$. Since $pvalue < \alpha = 0.01$, reject the null hypothesis at level 0.01.
- ▶ **Conclusion:** There is a significant association between parent's height and child's height at level 0.01.

Mean Response

Estimation of Mean Response

The mean response at $X = X_h$ is $E(Y_h) = \beta_0 + \beta_1 X_h$.

- ▶ An unbiased estimator of $E(Y_h)$:

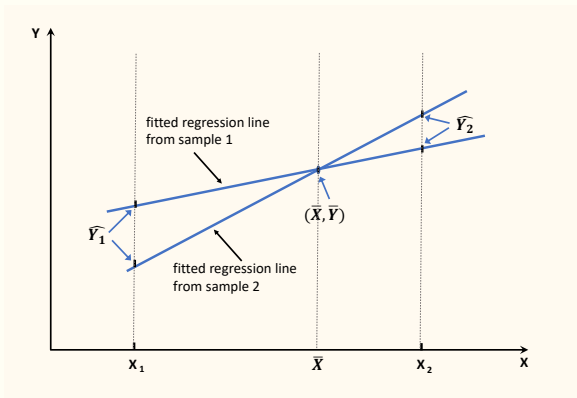
$$\widehat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h = \bar{Y} + \hat{\beta}_1 (X_h - \bar{X}).$$

- ▶ $\sigma^2\{\widehat{Y}_h\} = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right].$

- ▶ Standard error of \widehat{Y}_h :

$$s\{\widehat{Y}_h\} = \sqrt{MSE \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]}.$$

- ▶ The larger the sample size, or the larger the dispersion of X values, the smaller the SE of \widehat{Y}_h .
- ▶ The further X_h from \bar{X} , the larger the SE of \widehat{Y}_h .



Sampling Distribution of \widehat{Y}_h

Under the Normal error model:

- ▶ \widehat{Y}_h is normally distributed:

$$\widehat{Y}_h \sim \text{Normal}(E(Y_h), \sigma^2\{\widehat{Y}_h\})$$

- ▶ Pivotal quantity:

$$\frac{\widehat{Y}_h - E(Y_h)}{s(\widehat{Y}_h)} \sim t_{(n-2)}$$

Confidence Intervals of $E(Y_h)$

The $(1 - \alpha)100\%$ confidence interval of $E(Y_h)$:

$$\widehat{Y}_h \pm t(1 - \alpha/2; n - 2)s(\widehat{Y}_h)$$

Heights

What is the average height of children of 70in parents?

- ▶ $n = 928$, $\bar{X} = 68.316$, $\sum_{i=1}^n (X_i - \bar{X})^2 = 3038.761$ and
 $\hat{\beta}_0 = 24.54$, $\hat{\beta}_1 = 0.637$, $MSE = 5.031$
- ▶ $\hat{Y}_h = 24.54 + 0.637 \times 70 = 69.2$
- ▶ $s\{\hat{Y}_h\} = \sqrt{5.031 \times \left\{ \frac{1}{928} + \frac{(70 - 68.316)^2}{3038.761} \right\}} = 0.1$
- ▶ 95%-confidence interval: $69.2 \pm 1.963 \times 0.1 = [69, 69.40]$
- ▶ We are 95% confident that the average height of children of 70in parents is between [69in, 69.40in].

Prediction of New Outcome

Predict a **future outcome** at $X = X_h$:

$$Y_{h(new)} = \beta_0 + \beta_1 X_h + \epsilon_h$$

- ▶ Predict $Y_{h(new)}$ by the estimated mean response at $X = X_h$:

$$\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h = \bar{Y} + \hat{\beta}_1 (X_h - \bar{X})$$

- ▶ ϵ_h is assumed to be uncorrelated with ϵ_i s $\rightarrow Y_{h(new)}$ is uncorrelated with the observed Y_i s.

Pivotal Quantity

Under Normal error model:

- ▶ $\widehat{Y}_h - Y_{h(new)} \sim \text{Normal}(0, \sigma^2(pred_h))$, where

$$\begin{aligned}\sigma^2(pred_h) &:= \text{Var}(\widehat{Y}_h - Y_{h(new)}) = \sigma^2(\widehat{Y}_h) + \sigma^2(Y_{h(new)}) \\ &= \sigma^2(\widehat{Y}_h) + \sigma^2 = \sigma^2 \left[\mathbf{1} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]\end{aligned}$$

- ▶ Pivotal quantity: $\frac{\widehat{Y}_h - Y_{h(new)}}{s(pred_h)} \sim t_{(n-2)}$, where

$$s(pred_h) = \sqrt{MSE \left[\mathbf{1} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]}$$

Prediction Intervals

The $(1 - \alpha)100\%$ prediction interval of $Y_{h(new)}$:

$$\widehat{Y}_h \pm t(1 - \alpha/2; n - 2)s(pred_h)$$

Prediction vs. Estimation

- ▶ $Y_{h(new)}$ – a “moving target” (random variable) vs. $E(Y_h)$ – a fixed quantity (non-random).
- ▶ Two sources of variations in the prediction process: Variability from \hat{Y}_h and variability from the target $Y_{h(new)} \rightarrow s(pred_h) > s(\hat{Y}_h)$.
- ▶ At a given X value, the prediction interval of a new outcome is wider than the confidence interval of the mean response.

Heights

What would be the predicted height of the child of a 70in couple?

- ▶ $n = 928$, $\bar{X} = 68.316$, $\sum_{i=1}^n (X_i - \bar{X})^2 = 3038.761$, and
 $\hat{\beta}_0 = 24.54$, $\hat{\beta}_1 = 0.637$, $MSE = 5.031$

- ▶ Predicted height: $\hat{Y}_h = 24.54 + 0.637 \times 70 = 69.2$

- ▶ Standard error:

$$s\{pred_h\} = \sqrt{5.031 \times \left\{ 1 + \frac{1}{928} + \frac{(70 - 68.316)^2}{3038.761} \right\}} = 2.25$$

- ▶ 95% prediction interval: $69.2 \pm 1.8831 \times 2.25 = [64.75, 73.56]$
- ▶ We are 95% confident that the child's height will be between
[64.75in, 73.56in].

Extrapolation

Extrapolation occurs when predicting the outcome at an X value that lies outside of the observed data range.

- ▶ Every model has a **range of validity**.
- ▶ A model may be inappropriate when it is extended outside of the range of the observations upon which it was built.
- ▶ Extrapolation is less reliable than interpolation and need to be handled with caution.

Analysis of Variance

Analysis of Variance

- ▶ Basic idea: attributing variation in the data to different sources through **decomposition of the total variation**.
- ▶ In regression, the variation in the observations comes from:
 - ▶ variation in the error term
 - ▶ variation in X

Partition of Total Deviation

- ▶ **Total deviation:** difference between Y_i and the sample mean \bar{Y} :

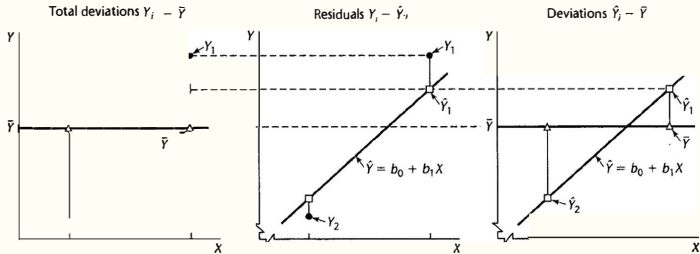
$$Y_i - \bar{Y}, \quad i = 1, \dots, n.$$

- ▶ Total deviation can be decomposed into the sum of two terms:

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}), \quad i = 1, \dots, n$$

- ▶ I.e., the *deviation of the observed value around the fitted regression line (residual)* and the *deviation of the fitted value from the sample mean*.

Figure: Partition of total deviation



Decomposition of Total Variation

- ▶ Taking sum of squares of the total deviations and noting that the sum of the cross product terms vanishes:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2.$$

- ▶ Decomposition of total variation:

$$SSTO = SSE + SSR$$

ANOVA: Sums of Squares

Total Sum of Squares (SSTO)

Quantify variation of the observations around the sample mean:

$$SSTO := \sum_{i=1}^n (Y_i - \bar{Y})^2, \quad d.f.(SSTO) = n - 1.$$

Error Sum of Squares (SSE)

Quantify variation of the observations around the fitted regression line:

$$SSE = \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2, \quad d.f.(SSE) = n - 2.$$

Regression Sum of Squares (SSR)

Quantify variation of the fitted values around the sample mean:

$$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2, \quad d.f.(SSR) = 1.$$

- ▶ $SSR = SSTO - SSE$: reduction of uncertainty in Y by utilizing the predictor X through a linear regression model
- ▶ The larger the fitted regression slope or the more the dispersion of X values, the larger SSR

Mean Squares

Sum of Squares divided by its degree of freedom:

$$MS = SS/d.f.(SS).$$

- ▶ Mean squared error:

$$MSE = \frac{SSE}{d.f.(SSE)} = \frac{SSE}{n - 2}$$

- ▶ Regression mean square:

$$MSR = \frac{SSR}{d.f.(SSR)} = \frac{SSR}{1}$$

ANOVA: F Tests

Expected Values of SS and MS

Under simple regression model:

- ▶ Expected values of SS:

$$E(SSE) = (n - 2)\sigma^2, \quad E(SSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2.$$

- ▶ Expected values of MS:

$$E(MSE) = \sigma^2, \quad E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2.$$

- ▶ $E(MSR) \geq E(MSE)$ and “=” holds iff $\beta_1 = 0$.

Sampling Distributions of SS

Under Normal error model:

- ▶ $SSE \sim \sigma^2 \chi^2_{(n-2)}$
- ▶ SSE and SSR are independent.

F Test

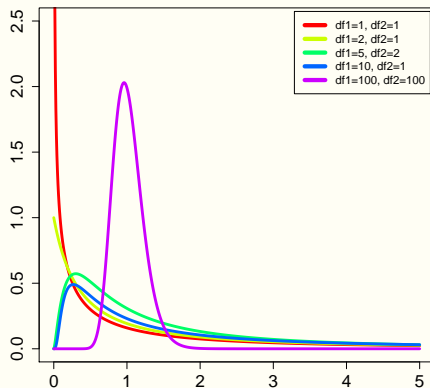
- ▶ $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$
- ▶ F ratio: $F^* = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}$
- ▶ Null distribution: $F^* \underset{H_0: \beta_1=0}{\sim} F_{1,n-2}$.
- ▶ Decision rule at the significance level α :

$$\text{reject } H_0 \text{ if } F^* > F(1 - \alpha; 1, n - 2),$$

where $F(1 - \alpha; 1, n - 2)$ is the $(1 - \alpha)$ 100th percentile of the $F_{1,n-2}$ distribution.

F Distributions

Figure: F distributions: probability density function



In simple linear regression, the F -test is equivalent to the two-sided t -test for testing $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$.

- ▶ $F^* = (T^*)^2$

- ▶ $F(1 - \alpha; 1, n - 2) = t^2(1 - \alpha/2; n - 2).$

ANOVA Table for Simple Regression

Source of Variation	SS	d.f.	MS=SS/d.f.	F^*
Regression	$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	1	$MSR = SSR/1$	MSR/MSE
Error	$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$n - 2$	$MSE = SSE/(n - 2)$	
Total	$SSTO = \sum_{i=1}^n (Y_i - \bar{Y})^2$	$n - 1$		

Heights

Source of Variation	SS	d.f.	MS=SS/d.f.	F^*
Regression	$SSR = 1234$	1	$MSR = 1234$	245
Error	$SSE = 4659$	926	$MSE = 5.03$	
Total	$SSTO = 5893$	927		

- ▶ Test whether there is a linear association between parent's height and child's height at significance level $\alpha = 0.01$.
- ▶ $F(0.99; 1, 926) = 6.66 < F^* = 245$, so reject $H_0 : \beta_1 = 0$ and conclude that there is a significant linear association between parent's height and child's height.

Coefficient of Determination

Coefficient of Determination R^2

A descriptive measure for **linear association** between X and Y :

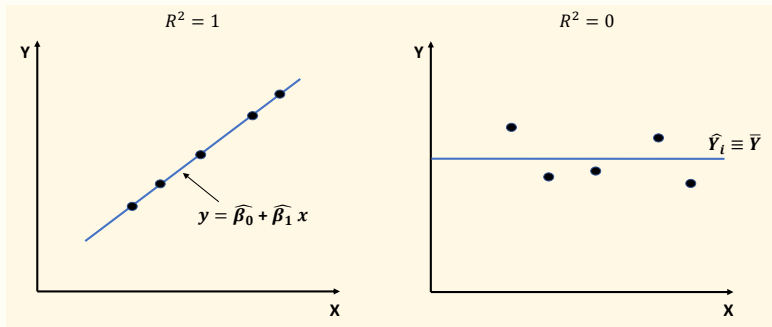
$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}.$$

- Heights: $R^2 = \frac{1234}{5893} = 0.209$. 20% of variation in child's height may be explained by the variation in parent's height.

Properties of R^2

- ▶ $0 \leq R^2 \leq 1$.
- ▶ If all observations fall on one straight line, then $R^2 = 1$.
 - ▶ X accounts for all variation in the observations.
- ▶ If the fitted regression line is horizontal, i.e., $\hat{\beta}_1 = 0$, then $R^2 = 0$.
 - ▶ X is of no use in explaining variation in the observations.
 - ▶ There is no evidence of linear association between X and Y in the data.

Figure:



Caution with Interpreting R^2

When the relationship between X and Y is nonlinear, R^2 is not a meaningful measure.

- ▶ *“A large R^2 means that the estimated regression line must be a good fit of the data”. Not necessarily!*
- ▶ *“A near zero R^2 means that X and Y are not related”. Not necessarily!*