Inferences in Regression Analysis

Zhenisbek Assylbekov

Department of Mathematics

Regression Analysis

Normal Errors Regression

Throughout this chapter we assume

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

$$\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

Normal Errors Regression

Throughout this chapter we assume

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

$$\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

This is equivalent to

$$Y_i \overset{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$$

Inferences on β_1 and β_0

Inferences on $\mathrm{E}[Y]$ and $\hat{Y}$

Analysis of Variance Approach

Coefficient of Determination

Unbiasedness of b_1

Proposition. b_1 is an unbiased estimator of β_1.

Unbiasedness of b_1

Proposition. b_1 is an unbiased estimator of β_1.

Proof.
Recall that $b_1 = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2}$

Unbiasedness of b_1

Proposition. b_1 is an unbiased estimator of β_1.

Proof.
Recall that $b_1 = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2} \stackrel{?}{=} \frac{\sum(x_i - \bar{x})Y_i}{\sum(x_i - \bar{x})^2}$.

Unbiasedness of b_1

Proposition. b_1 is an unbiased estimator of β_1.

Proof.
Recall that $b_1 = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2} \stackrel{?}{=} \frac{\sum(x_i - \bar{x})Y_i}{\sum(x_i - \bar{x})^2}$.
Expected value of the numerator is

$$\mathrm{E}\left[\sum(x_i - \bar{x})Y_i\right] = \sum(x_i - \bar{x})\mathrm{E}[Y_i] = \sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i)$$
$$= \beta_0 \sum x_i - n\bar{x}\beta_0 + \beta_1 \sum x_i^2 - n\bar{x}^2\beta_1 = \beta_1 \left(\sum x_i^2 - n\bar{x}^2\right)$$

Unbiasedness of b_1

Proposition. b_1 is an unbiased estimator of β_1.

Proof.
Recall that $b_1 = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2} \stackrel{?}{=} \frac{\sum(x_i - \bar{x})Y_i}{\sum(x_i - \bar{x})^2}$.
Expected value of the numerator is

$$\mathrm{E}\left[\sum(x_i - \bar{x})Y_i\right] = \sum(x_i - \bar{x})\mathrm{E}[Y_i] = \sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i)$$
$$= \beta_0 \sum x_i - n\bar{x}\beta_0 + \beta_1 \sum x_i^2 - n\bar{x}^2\beta_1 = \beta_1 \left(\sum x_i^2 - n\bar{x}^2\right)$$

The denominator is

$$\sum(x_i - \bar{x})^2 = \sum x_i^2 - 2\sum x_i \cdot \bar{x} + \sum \bar{x}^2 = \sum x_i^2 - n\bar{x}^2$$

Unbiasedness of b_1

Proposition. b_1 is an unbiased estimator of β_1.

Proof.
Recall that $b_1 = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2} \stackrel{?}{=} \frac{\sum(x_i - \bar{x})Y_i}{\sum(x_i - \bar{x})^2}$.
Expected value of the numerator is

$$\mathrm{E}\left[\sum(x_i - \bar{x})Y_i\right] = \sum(x_i - \bar{x})\mathrm{E}[Y_i] = \sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i)$$
$$= \beta_0 \sum x_i - n\bar{x}\beta_0 + \beta_1 \sum x_i^2 - n\bar{x}^2\beta_1 = \beta_1 \left(\sum x_i^2 - n\bar{x}^2\right)$$

The denominator is

$$\sum(x_i - \bar{x})^2 = \sum x_i^2 - 2\sum x_i \cdot \bar{x} + \sum \bar{x}^2 = \sum x_i^2 - n\bar{x}^2$$

Hence, $\mathrm{E}[b_1] = \beta_1$. $\square$

Variance of b_1

Proposition. $\mathrm{Var}[b_1] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Variance of b_1

Proposition. $\mathrm{Var}[b_1] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Proof.

$$\mathrm{Var}[b_1] = \mathrm{Var}\left[\frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2}\right]$$

$\square$

Variance of b_1

Proposition. $\mathrm{Var}[b_1] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Proof.

$$\mathrm{Var}[b_1] = \mathrm{Var}\left[\frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2}\right] = \frac{\sum_i (x_i - \bar{x})^2 \mathrm{Var}[Y_i]}{\left[\sum_i (x_i - \bar{x})^2\right]^2}$$

$\square$

Variance of b_1

Proposition. $\mathrm{Var}[b_1] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Proof.

$$\begin{aligned}
\mathrm{Var}[b_1] = \mathrm{Var}\left[\frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2}\right] &= \frac{\sum_i (x_i - \bar{x})^2 \mathrm{Var}[Y_i]}{\left[\sum_i (x_i - \bar{x})^2\right]^2} \\
&= \frac{\sigma^2 \sum_i (x_i - \bar{x})}{\left[\sum_i (x_i - \bar{x})^2\right]^2}
\end{aligned}$$

$\square$

Variance of b_1

Proposition. $\mathrm{Var}[b_1] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Proof.

$$\mathrm{Var}[b_1] = \mathrm{Var}\left[\frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2}\right] = \frac{\sum_i (x_i - \bar{x})^2 \mathrm{Var}[Y_i]}{\left[\sum_i (x_i - \bar{x})^2\right]^2}$$

$$= \frac{\sigma^2 \sum_i (x_i - \bar{x})}{\left[\sum_i (x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

$\square$

Variance of b_1

Proposition. $\mathrm{Var}[b_1] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Proof.

$$\mathrm{Var}[b_1] = \mathrm{Var}\left[\frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2}\right] = \frac{\sum_i (x_i - \bar{x})^2 \mathrm{Var}[Y_i]}{\left[\sum_i (x_i - \bar{x})^2\right]^2}$$
$$= \frac{\sigma^2 \sum_i (x_i - \bar{x})}{\left[\sum_i (x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

$\square$

Remark.

- $\sum_i (x_i - \bar{x})^2 \uparrow \quad \Rightarrow \quad \mathrm{Var}[b_1] \downarrow$

Variance of b_1

Proposition. $\mathrm{Var}[b_1] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Proof.

$$\begin{aligned}
\mathrm{Var}[b_1] &= \mathrm{Var}\left[\frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2}\right] = \frac{\sum_i (x_i - \bar{x})^2 \mathrm{Var}[Y_i]}{\left[\sum_i (x_i - \bar{x})^2\right]^2} \\
&= \frac{\sigma^2 \sum_i (x_i - \bar{x})}{\left[\sum_i (x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.
\end{aligned}$$

$\square$

Remark.

- $\sum_i (x_i - \bar{x})^2 \uparrow \quad \Rightarrow \quad \mathrm{Var}[b_1] \downarrow$
- $n \uparrow \quad \Rightarrow \quad \mathrm{Var}[b_1] \downarrow$

Distribution of b_1

Rewrite b_1 as

$$b_1 = \frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2} = \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right] Y_i$$

Distribution of b_1

Rewrite b_1 as

$$b_1 = \frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2} = \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right] Y_i$$

Thus, b_1 is a linear combination of n independent normal random variables $Y_1, \ldots, Y_n$.

Distribution of b_1

Rewrite b_1 as

$$b_1 = \frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2} = \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right] Y_i$$

Thus, b_1 is a linear combination of n independent normal random variables $Y_1, \ldots, Y_n$. Therefore

$$b_1 \sim$$

Distribution of b_1

Rewrite b_1 as

$$b_1 = \frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2} = \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right] Y_i$$

Thus, b_1 is a linear combination of n independent normal random variables $Y_1, \ldots, Y_n$. Therefore

$$b_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right),$$

Distribution of b_1

Rewrite b_1 as

$$b_1 = \frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2} = \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right] Y_i$$

Thus, b_1 is a linear combination of n independent normal random variables $Y_1, \ldots, Y_n$. Therefore

$$b_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right),$$

and

$$\frac{b_1 - \beta_1}{\sqrt{\mathrm{Var}[b_1]}} \sim \mathcal{N}(0, 1).$$

Distribution of b_1

Rewrite b_1 as

$$b_1 = \frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2} = \sum_{i=1}^{n} \left[\frac{(x_i - \bar{x})}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right] Y_i$$

Thus, b_1 is a linear combination of n independent normal random variables $Y_1, \ldots, Y_n$. Therefore

$$b_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right),$$

and

$$\frac{b_1 - \beta_1}{\sqrt{\mathrm{Var}[b_1]}} \sim \mathcal{N}(0,1).$$

We never know σ^2, we estimate it by MSE $= \frac{1}{n-2} \sum_i (Y_i - \hat{Y}_i)^2$.

Distribution of $\frac{b_1 - \beta_1}{\mathrm{s}[b_1]}$

Denote $\mathrm{s}[b_1] = \sqrt{\frac{\mathsf{MSE}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$.

Distribution of $\frac{b_1-\beta_1}{\mathrm{s}[b_1]}$

Denote $\mathrm{s}[b_1] = \sqrt{\frac{\text{MSE}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$.

Theorem. $\frac{b_1-\beta_1}{\mathrm{s}[b_1]} \sim t_{n-2}$.

Proof.

$$\frac{b_1-\beta_1}{\mathrm{s}[b_1]} = \overbrace{\frac{b_1-\beta_1}{\sqrt{\mathrm{Var}[b_1]}}}^{Z} \cdot \frac{\sqrt{\mathrm{Var}[b_1]}}{\mathrm{s}[b_1]} = Z \cdot \sqrt{\frac{\sigma^2}{\frac{1}{n-2}\sum_i e_i^2}}$$

$$= \frac{Z}{\sqrt{\frac{\sum_i e_i^2/\sigma^2}{n-2}}} = \frac{Z}{\sqrt{\frac{\chi^2_{n-2}}{n-2}}} \sim t_{n-2},$$

where we used the fact that $\frac{\sum_i e_i^2}{\sigma^2} \sim \chi^2_{n-2}$ (Ch 5). $\square$

Confidence interval for β_1 and testing $\mathrm{H}_0 : \beta_1 = \beta_{10}$

A $(1-\alpha)\cdot 100\%$ CI for β_1 has endpoints

$$b_1 \pm t_{n-2,1-\alpha/2} \cdot \mathrm{s}[b_1].$$

Confidence interval for β_1 and testing $\mathrm{H}_0 : \beta_1 = \beta_{10}$

A $(1-\alpha)\cdot 100\%$ CI for β_1 has endpoints

$$b_1 \pm t_{n-2,1-\alpha/2} \cdot \mathrm{s}[b_1].$$

Under $\mathrm{H}_0 : \beta_1 = \beta_{1,0}$,

$$T = \frac{b_1 - \beta_{1,0}}{\mathrm{s}[b_1]} \sim t_{n-2}.$$

Confidence interval for β_1 and testing $\mathrm{H}_0 : \beta_1 = \beta_{10}$

A $(1-\alpha)\cdot 100\%$ CI for β_1 has endpoints

$$b_1 \pm t_{n-2,1-\alpha/2} \cdot \mathrm{s}[b_1].$$

Under $\mathrm{H}_0 : \beta_1 = \beta_{1,0}$,

$$T = \frac{b_1 - \beta_{1,0}}{\mathrm{s}[b_1]} \sim t_{n-2}.$$

P-values are computed as usually.

Confidence interval for β_1 and testing $\mathrm{H}_0 : \beta_1 = \beta_{10}$

A $(1-\alpha)\cdot 100\%$ CI for β_1 has endpoints

$$b_1 \pm t_{n-2,1-\alpha/2} \cdot \mathrm{s}[b_1].$$

Under $\mathrm{H}_0 : \beta_1 = \beta_{1,0}$,

$$T = \frac{b_1 - \beta_{1,0}}{\mathrm{s}[b_1]} \sim t_{n-2}.$$

P-values are computed as usually.

In simple linear regression

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

of particular interest is $\mathrm{H}_0 : \beta_1 = 0$, that Y_i and *does not depend on* x_i.

Expectation, Variance, and Distribution of b_0

Exercise. Show that

- $\mathrm{E}[b_0] = \beta_0$

Expectation, Variance, and Distribution of b_0

Exercise. Show that

- $\mathrm{E}[b_0] = \beta_0$
- $\mathrm{Var}[b_0] = \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right] \sigma^2$

Expectation, Variance, and Distribution of b_0

Exercise. Show that

- $\mathrm{E}[b_0] = \beta_0$
- $\mathrm{Var}[b_0] = \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right] \sigma^2$
- $b_0 \sim \mathcal{N}\left(\beta_0, \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right] \sigma^2\right)$

Expectation, Variance, and Distribution of b_0

Exercise. Show that

- $\mathrm{E}[b_0] = \beta_0$
- $\mathrm{Var}[b_0] = \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right] \sigma^2$
- $b_0 \sim \mathcal{N}\left(\beta_0, \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right] \sigma^2\right)$
- $\frac{b_0 - \beta_0}{\mathrm{s}[b_0]} \sim t_{n-2}$, where $\mathrm{s}[b_0]$ is obtained from $\sqrt{\mathrm{Var}[b_0]}$ when replacing σ^2 by MSE

Expectation, Variance, and Distribution of b_0

Exercise. Show that

- $\mathrm{E}[b_0] = \beta_0$
- $\mathrm{Var}[b_0] = \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right] \sigma^2$
- $b_0 \sim \mathcal{N}\left(\beta_0, \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right] \sigma^2\right)$
- $\frac{b_0 - \beta_0}{\mathrm{s}[b_0]} \sim t_{n-2}$, where $\mathrm{s}[b_0]$ is obtained from $\sqrt{\mathrm{Var}[b_0]}$ when replacing σ^2 by MSE

CI and Hypothesis Test for β_0 are as usual.

Table of regression coefficients

Regression output typically produces a table like:

Parameter	Estimate	Standard error	t^*	p-value
Intercept β_0	b_0	$\mathrm{s}[b_0]$	$t_0^* = \frac{b_0}{\mathrm{s}[b_0]}$	$\Pr(\lvert T \rvert > \lvert t_0^* \rvert)$
Slope β_1	b_1	$\mathrm{s}[b_1]$	$t_1^* = \frac{b_1}{\mathrm{s}[b_1]}$	$\Pr(\lvert T \rvert > \lvert t_1^* \rvert)$

Table of regression coefficients

Regression output typically produces a table like:

Parameter	Estimate	Standard error	t^*	p-value
Intercept β_0	b_0	$\mathrm{s}[b_0]$	$t_0^* = \frac{b_0}{\mathrm{s}[b_0]}$	$\Pr(\lvert T\rvert > \lvert t_0^*\rvert)$
Slope β_1	b_1	$\mathrm{s}[b_1]$	$t_1^* = \frac{b_1}{\mathrm{s}[b_1]}$	$\Pr(\lvert T\rvert > \lvert t_1^*\rvert)$

where $T \sim t_{n-p}$ and p is the number of parameters used to estimate the mean, here $p = 2$: β_0 and β_1. Later p will be the number of predictors in the model plus one.

Table of regression coefficients

Regression output typically produces a table like:

Parameter	Estimate	Standard error	t^*	p-value
Intercept β_0	b_0	$\mathrm{s}[b_0]$	$t_0^* = \frac{b_0}{\mathrm{s}[b_0]}$	$\Pr(\lvert T \rvert > \lvert t_0^* \rvert)$
Slope β_1	b_1	$\mathrm{s}[b_1]$	$t_1^* = \frac{b_1}{\mathrm{s}[b_1]}$	$\Pr(\lvert T \rvert > \lvert t_1^* \rvert)$

where $T \sim t_{n-p}$ and p is the number of parameters used to estimate the mean, here $p = 2$: β_0 and β_1. Later p will be the number of predictors in the model plus one.

The two p-values in the table test $\mathrm{H}_0 : \beta_0 = 0$ and $\mathrm{H}_1 : \beta_1 = 0$ respectively. The test for zero intercept is usually not of interest.

Regression Output in R: Poverty vs HS grad rate

https://raw.githubusercontent.com/zh3nis/MATH440/main/chp01/poverty.R

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.78097    6.80260   9.523 9.94e-13 ***
Graduates   -0.62122    0.07902  -7.862 3.11e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.082 on 49 degrees of freedom
```

Regression Output in R: Poverty vs HS grad rate

https://raw.githubusercontent.com/zh3nis/MATH440/main/chp01/poverty.R

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.78097    6.80260   9.523 9.94e-13 ***
Graduates   -0.62122    0.07902  -7.862 3.11e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.082 on 49 degrees of freedom
```

We reject $\mathrm{H}_0 : \beta_1 = 0$ at any reasonable significance level. There is a significant linear association between HS graduation and poverty rates.

Inferences on β_1 and β_0

Inferences on $\mathrm{E}[Y]$ and $\hat{Y}$

Analysis of Variance Approach

Coefficient of Determination

Inference on $\mathrm{E}[Y] = \beta_0 + \beta_1 x$

Let x be any value of the *predictor*; we want to estimate the mean of all responses in the *population* that correspond to x. This is given by

$$\mathrm{E}[Y] = \beta_0 + \beta_1 x.$$

Our estimator of $\mathrm{E}[Y]$ is

$$\begin{aligned}
\hat{Y} &= b_0 + b_1 x \\
&= \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} + \frac{(x_i - \bar{x})x}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \right] Y_i \\
&= \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x - \bar{x})(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \right] Y_i
\end{aligned}$$

Distribution of $\hat{Y}$

Again we have a linear combination of independent normals as our estimator. This leads, after some math, to

$$b_0 + b_1 x \sim \mathcal{N}\left(\beta_0 + \beta_1 x, \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]\right). \quad (1)$$

Distribution of $\hat{Y}$

Again we have a linear combination of independent normals as our estimator. This leads, after some math, to

$$b_0 + b_1 x \sim \mathcal{N}\left(\beta_0 + \beta_1 x, \sigma^2 \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]\right). \quad (1)$$

As before, this leads to a $(1-\alpha)\cdot 100\%$ CI for $\beta_0 + \beta_1 x$

$$b_0 + b_1 x \pm t_{n-2,1-\alpha/2} \cdot \mathrm{s}[b_0 + b_1 x],$$

Distribution of $\hat{Y}$

Again we have a linear combination of independent normals as our estimator. This leads, after some math, to

$$b_0 + b_1 x \sim \mathcal{N}\left(\beta_0 + \beta_1 x, \sigma^2 \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]\right). \quad (1)$$

As before, this leads to a $(1-\alpha)\cdot 100\%$ CI for $\beta_0 + \beta_1 x$

$$b_0 + b_1 x \pm t_{n-2,1-\alpha/2} \cdot \mathrm{s}[b_0 + b_1 x],$$

where $\mathrm{s}[b_0 + b_1 x] = \sqrt{\mathrm{MSE} \cdot \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]}$.

Distribution of $\hat{Y}$

Again we have a linear combination of independent normals as our estimator. This leads, after some math, to

$$b_0 + b_1 x \sim \mathcal{N}\left(\beta_0 + \beta_1 x, \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]\right). \quad (1)$$

As before, this leads to a $(1 - \alpha) \cdot 100\%$ CI for $\beta_0 + \beta_1 x$

$$b_0 + b_1 x \pm t_{n-2,1-\alpha/2} \cdot \mathrm{s}[b_0 + b_1 x],$$

where $\mathrm{s}[b_0 + b_1 x] = \sqrt{\mathrm{MSE} \cdot \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]}$.

For what value of x is the CI narrowest?

Distribution of $\hat{Y}$

Again we have a linear combination of independent normals as our estimator. This leads, after some math, to

$$b_0 + b_1 x \sim \mathcal{N}\left(\beta_0 + \beta_1 x, \sigma^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]\right). \tag{1}$$

As before, this leads to a $(1-\alpha)\cdot 100\%$ CI for $\beta_0 + \beta_1 x$

$$b_0 + b_1 x \pm t_{n-2,1-\alpha/2} \cdot \mathrm{s}[b_0 + b_1 x],$$

where $\mathrm{s}[b_0 + b_1 x] = \sqrt{\mathrm{MSE} \cdot \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]}$.

For what value of x is the CI narrowest? For $x = \bar{x}$.

Exercise. Prove (1).

Distribution of $\hat{Y}$

Again we have a linear combination of independent normals as our estimator. This leads, after some math, to

$$b_0 + b_1 x \sim \mathcal{N}\left(\beta_0 + \beta_1 x, \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]\right). \qquad (1)$$

As before, this leads to a $(1-\alpha)\cdot 100\%$ CI for $\beta_0 + \beta_1 x$

$$b_0 + b_1 x \pm t_{n-2,1-\alpha/2} \cdot \mathrm{s}[b_0 + b_1 x],$$

where $\mathrm{s}[b_0 + b_1 x] = \sqrt{\mathrm{MSE} \cdot \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}$.

For what value of x is the CI narrowest? For $x = \bar{x}$.

Exercise. Prove (1). Solution is on pp. 53–54 of the textbook.

Prediction intervals

- We discussed constructing a CI for the unknown $\mathrm{E}[Y]$ at x.

Prediction intervals

- We discussed constructing a CI for the unknown $\mathrm{E}[Y]$ at x.
- What if we want to find an interval that the actual *value* Y is in (versus only its mean) with fixed probability?

Prediction intervals

- We discussed constructing a CI for the unknown $\mathrm{E}[Y]$ at x.
- What if we want to find an interval that the actual *value* Y is in (versus only its mean) with fixed probability?
- If we knew β_0, β_1, and σ^2 this would be easy, because

$$Y = \beta_0 + \beta_1 x + \epsilon \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2),$$

and a $(1 - \alpha) \cdot 100\%$ CI for $\mathrm{E}[Y]$ would be

$$\beta_0 + \beta_1 x \pm z_{1-\alpha/2} \cdot \sigma.$$

Prediction intervals

- We discussed constructing a CI for the unknown $\mathrm{E}[Y]$ at x.
- What if we want to find an interval that the actual *value* Y is in (versus only its mean) with fixed probability?
- If we knew β_0, β_1, and σ^2 this would be easy, because

$$Y = \beta_0 + \beta_1 x + \epsilon \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2),$$

  and a $(1-\alpha) \cdot 100\%$ CI for $\mathrm{E}[Y]$ would be

$$\beta_0 + \beta_1 x \pm z_{1-\alpha/2} \cdot \sigma.$$

- Unfortunately, we don't know β_0, β_1, and σ, but we can estimate all three of these.

Variability of $b_0 + b_1 x + \epsilon$

An interval that contains Y with $(1 - \alpha)$ probability needs to account for

Variability of $b_0 + b_1 x + \epsilon$

An interval that contains Y with $(1 - \alpha)$ probability needs to account for

- the variability of the estimators b_0 and b_1

Variability of $b_0 + b_1 x + \epsilon$

An interval that contains Y with $(1 - \alpha)$ probability needs to account for

- the variability of the estimators b_0 and b_1
- the natural variability of response Y built into the model: $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Variability of $b_0 + b_1 x + \epsilon$

An interval that contains Y with $(1-\alpha)$ probability needs to account for

- the variability of the estimators b_0 and b_1
- the natural variability of response Y built into the model: $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We have

$$\begin{aligned}
\mathrm{Var}[b_0 + b_1 x + \epsilon] &= \mathrm{Var}[b_0 + b_1 x] + \mathrm{Var}[\epsilon] \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] + \sigma^2 \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + 1 \right]
\end{aligned}$$

Prediction interval

Estimating σ^2 by MSE we obtain that a $(1-\alpha)\cdot 100\%$ **prediction interval** (PI) for Y is

$$b_0 + b_1 x \pm t_{n-2,1-\alpha/2}\sqrt{\text{MSE}\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + 1\right]}.$$

Remark. As $n \to \infty$, we have $b_0 \xrightarrow{\text{P}} \beta_0$, $b_1 \xrightarrow{\text{P}} \beta_1$, $t_{n-2,1-\alpha/2} \to z_{1-\alpha/2}$, and $\text{MSE} \xrightarrow{\text{P}} \sigma^2$.

I.e., as the sample size grows, the PI converges to

$$\beta_0 + \beta_1 x \pm z_{1-\alpha/2}\cdot\sigma.$$

Example: Poverty vs HS Graduation data

- Find a 95% CI for the mean poverty rate $\mathrm{E}[Y]$ in a state with HS graduation rate $x = 80$.
- Find a 95% PI for the poverty rate Y in a state with HS graduation rate $x = 80$.
- `R` code follows. . .

R code

https://raw.githubusercontent.com/zh3nis/MATH440/main/chp02/pov_predict.R

```
> poverty = read.table("path/to/poverty.txt", h = T, sep = "\t")

> my_model = lm(Poverty ~ Graduates, data=poverty)

> new_x = data.frame(Graduates=80)

> predict.lm(my_model, new_x, interval="confidence", level=0.95)
       fit     lwr      upr
1 15.08363 13.9636 16.20365

> predict.lm(my_model, new_x, interval="prediction", level=0.95)
       fit     lwr      upr
1 15.08363 10.7527 19.41455
```

R code

https://raw.githubusercontent.com/zh3nis/MATH440/main/chp02/pov_predict.R

```
> poverty = read.table("path/to/poverty.txt", h = T, sep = "\t")

> my_model = lm(Poverty ~ Graduates, data=poverty)

> new_x = data.frame(Graduates=80)

> predict.lm(my_model, new_x, interval="confidence", level=0.95)
       fit     lwr      upr
1 15.08363 13.9636 16.20365

> predict.lm(my_model, new_x, interval="prediction", level=0.95)
       fit     lwr      upr
1 15.08363 10.7527 19.41455
```

A 95% CI for $\mathrm{E}[Y]$ given $x = 80$ is $[13.96, 16.20]$.

R code

https://raw.githubusercontent.com/zh3nis/MATH440/main/chp02/pov_predict.R

```
> poverty = read.table("path/to/poverty.txt", h = T, sep = "\t")

> my_model = lm(Poverty ~ Graduates, data=poverty)

> new_x = data.frame(Graduates=80)

> predict.lm(my_model, new_x, interval="confidence", level=0.95)
       fit     lwr      upr
1 15.08363 13.9636 16.20365

> predict.lm(my_model, new_x, interval="prediction", level=0.95)
       fit     lwr      upr
1 15.08363 10.7527 19.41455
```

A 95% CI for $\mathrm{E}[Y]$ given $x = 80$ is $[13.96, 16.20]$.
A 95% PI for Y given $x = 80$ is $[10.75, 19.41]$.

Confidence Band for Regression Line

Working-Hotelling
confidence band:

$$\hat{Y} \pm W \cdot \mathrm{s}[\hat{Y}],$$

where $W^2 = 2 \cdot F_{1-\alpha;2,n-2}$.

Confidence Band for Regression Line

Working-Hotelling confidence band:

$$\hat{Y} \pm W \cdot \mathrm{s}[\hat{Y}],$$

where $W^2 = 2 \cdot F_{1-\alpha;2,n-2}$. Gives region that *entire* regression line lies in with certain confidence.

Confidence Band for Regression Line

Working-Hotelling confidence band:

$$\hat{Y} \pm W \cdot \mathrm{s}[\hat{Y}],$$

where $W^2 = 2 \cdot F_{1-\alpha;2,n-2}$. Gives region that *entire* regression line lies in with certain confidence.

Confidence Band for Regression Line

Working-Hotelling confidence band:

$$\hat{Y} \pm W \cdot \mathrm{s}[\hat{Y}],$$

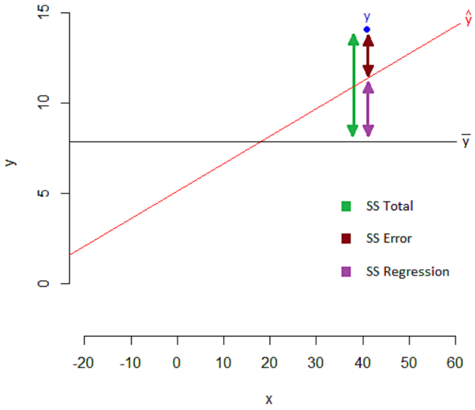
where $W^2 = 2 \cdot F_{1-\alpha;2,n-2}$. Gives region that *entire* regression line lies in with certain confidence.

https://raw.githubusercontent.com/zh3nis/MATH440/main/chp02/pov_cb.R

Decomposing $Y_i - \bar{Y}$

Notice that

$$\overbrace{Y_i - \bar{Y}}^{\text{Total}} = \overbrace{Y_i - \hat{Y}_i}^{\text{Error}} + \overbrace{\hat{Y}_i - \bar{Y}}^{\text{Regression}}$$



Partitioning SST

$$\overbrace{Y_i - \bar{Y}}^{\text{Total}} = \overbrace{Y_i - \hat{Y}_i}^{\text{Error}} + \overbrace{\hat{Y}_i - \bar{Y}}^{\text{Regression}} \tag{2}$$

Partitioning SST

$$\overbrace{Y_i - \bar{Y}}^{\text{Total}} = \overbrace{Y_i - \hat{Y}_i}^{\text{Error}} + \overbrace{\hat{Y}_i - \bar{Y}}^{\text{Regression}} \tag{2}$$

Squaring (2) and summing over i, we have

Partitioning SST

$$\overbrace{Y_i - \bar{Y}}^{\text{Total}} = \overbrace{Y_i - \hat{Y}_i}^{\text{Error}} + \overbrace{\hat{Y}_i - \bar{Y}}^{\text{Regression}} \tag{2}$$

Squaring (2) and summing over i, we have

$$\underbrace{\sum (Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum (Y_i - \hat{Y}_i)^2}_{\text{SSE}} + \underbrace{\sum (\hat{Y}_i - \bar{Y})^2}_{\text{SSR}}$$

Partitioning SST

$$\overbrace{Y_i - \bar{Y}}^{\text{Total}} = \overbrace{Y_i - \hat{Y}_i}^{\text{Error}} + \overbrace{\hat{Y}_i - \bar{Y}}^{\text{Regression}} \tag{2}$$

Squaring (2) and summing over i, we have

$$\underbrace{\sum (Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum (Y_i - \hat{Y}_i)^2}_{\text{SSE}} + \underbrace{\sum (\hat{Y}_i - \bar{Y})^2}_{\text{SSR}}$$

because (see Chapter 1)

$$\sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \sum e_i(\hat{Y}_i - \bar{Y}) = \sum e_i \hat{Y}_i - \bar{Y} \sum e_i = 0$$

Degrees of Freedom of SS

In Chapter 5 we will show that

- $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-2}$

Degrees of Freedom of SS

In Chapter 5 we will show that

- $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-2} \quad \Rightarrow \quad$ SSE has $n-2$ df

Degrees of Freedom of SS

In Chapter 5 we will show that

- $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-2} \quad \Rightarrow \quad \text{SSE has } n-2 \text{ df}$
- $\frac{\text{SSR}}{\sigma^2} \sim \chi^2_1$ (noncentral)

Degrees of Freedom of SS

In Chapter 5 we will show that

- $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-2} \quad \Rightarrow \quad$ SSE has $n-2$ df
- $\frac{\text{SSR}}{\sigma^2} \sim \chi^2_1$ (noncentral) $\quad \Rightarrow \quad$ SSR has 1 df
- SSE and SSR are indep.

Degrees of Freedom of SS

In Chapter 5 we will show that

- $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-2} \quad \Rightarrow \quad \text{SSE has } n-2 \text{ df}$
- $\frac{\text{SSR}}{\sigma^2} \sim \chi^2_1 \text{ (noncentral)} \quad \Rightarrow \quad \text{SSR has 1 df}$
- SSE and SSR are indep. $\quad \Rightarrow \quad \frac{\text{SST}}{\sigma^2} = \frac{\text{SSE}}{\sigma^2} + \frac{\text{SSR}}{\sigma^2} \sim \chi^2_{n-1}$

Degrees of Freedom of SS

In Chapter 5 we will show that

- $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-2} \quad \Rightarrow \quad$ SSE has $n-2$ df
- $\frac{\text{SSR}}{\sigma^2} \sim \chi^2_1$ (noncentral) $\quad \Rightarrow \quad$ SSR has 1 df
- SSE and SSR are indep. $\quad \Rightarrow \quad \frac{\text{SST}}{\sigma^2} = \frac{\text{SSE}}{\sigma^2} + \frac{\text{SSR}}{\sigma^2} \sim \chi^2_{n-1} \quad \Rightarrow$ SST has $n-1$ df

Expectations of SS

Proposition. $\mathrm{E}[\mathsf{MSE}] = \sigma^2$.

Expectations of SS

Proposition. $\mathrm{E}[\mathsf{MSE}] = \sigma^2$.

Proof.

$$\mathrm{E}[\mathsf{MSE}] = \mathrm{E}\left[\frac{\mathsf{SSE}}{n-2}\right] = \frac{\sigma^2}{n-2}\mathrm{E}\left[\frac{\mathsf{SSE}}{\sigma^2}\right] = \frac{\sigma^2}{n-2}(n-2) = \sigma^2. \qquad \square$$

Expectations of SS

Proposition. $\mathrm{E}[\mathsf{MSR}] = \sigma^2 + \beta_1^2 \sum_i (x_i - \bar{x})^2$.

Proof.

Expectations of SS

Proposition. $\mathrm{E}[\text{MSR}] = \sigma^2 + \beta_1^2 \sum_i (x_i - \bar{x})^2$.

Proof.

$$\text{SSR} = \sum_i (\hat{Y}_i - \bar{Y})^2 = \sum_i (b_0 + b_1 x_i - \bar{Y})^2$$

Expectations of SS

Proposition. $\mathrm{E}[\mathsf{MSR}] = \sigma^2 + \beta_1^2 \sum_i (x_i - \bar{x})^2$.

Proof.

$$\begin{aligned}
\mathsf{SSR} &= \sum_i (\hat{Y}_i - \bar{Y})^2 = \sum_i (b_0 + b_1 x_i - \bar{Y})^2 \\
&= \sum_i (\bar{Y} - b_1 \bar{x} + b_1 x_i - \bar{Y})^2 = b_1^2 \sum_i (x_i - \bar{x})^2.
\end{aligned}$$

Expectations of SS

Proposition. $\mathrm{E}[\mathsf{MSR}] = \sigma^2 + \beta_1^2 \sum_i (x_i - \bar{x})^2$.

Proof.

$$\begin{aligned} \mathsf{SSR} &= \sum_i (\hat{Y}_i - \bar{Y})^2 = \sum_i (b_0 + b_1 x_i - \bar{Y})^2 \\ &= \sum_i (\bar{Y} - b_1 \bar{x} + b_1 x_i - \bar{Y})^2 = b_1^2 \sum_i (x_i - \bar{x})^2. \end{aligned}$$

$$\mathrm{E}[\mathsf{MSR}] = \mathrm{E}\left[\frac{\mathsf{SSR}}{1}\right] = \mathrm{E}\left[b_1^2 \sum_i (x_i - \bar{x})^2\right]$$

Expectations of SS

Proposition. $\mathrm{E}[\mathsf{MSR}] = \sigma^2 + \beta_1^2 \sum_i (x_i - \bar{x})^2$.

Proof.

$$\begin{aligned}
\mathsf{SSR} &= \sum_i (\hat{Y}_i - \bar{Y})^2 = \sum_i (b_0 + b_1 x_i - \bar{Y})^2 \\
&= \sum_i (\bar{Y} - b_1 \bar{x} + b_1 x_i - \bar{Y})^2 = b_1^2 \sum_i (x_i - \bar{x})^2. \\
\mathrm{E}[\mathsf{MSR}] &= \mathrm{E}\left[\frac{\mathsf{SSR}}{1}\right] = \mathrm{E}\left[b_1^2 \sum_i (x_i - \bar{x})^2\right] \\
&= \sum_i (x_i - \bar{x})^2 (\mathrm{Var}[b_1] + (\mathrm{E}[b_1])^2)
\end{aligned}$$

Expectations of SS

Proposition. $\mathrm{E}[\mathsf{MSR}] = \sigma^2 + \beta_1^2 \sum_i (x_i - \bar{x})^2$.

Proof.

$$\begin{aligned}
\mathsf{SSR} &= \sum_i (\hat{Y}_i - \bar{Y})^2 = \sum_i (b_0 + b_1 x_i - \bar{Y})^2 \\
&= \sum_i (\bar{Y} - b_1 \bar{x} + b_1 x_i - \bar{Y})^2 = b_1^2 \sum_i (x_i - \bar{x})^2. \\
\mathrm{E}[\mathsf{MSR}] &= \mathrm{E}\left[\frac{\mathsf{SSR}}{1}\right] = \mathrm{E}\left[b_1^2 \sum_i (x_i - \bar{x})^2\right] \\
&= \sum_i (x_i - \bar{x})^2 (\mathrm{Var}[b_1] + (\mathrm{E}[b_1])^2) \\
&= \sigma^2 + \beta_1^2 \sum_i (x_i - \bar{x})^2.
\end{aligned}$$

Analysis of Variance (ANOVA) table

Source	SS	df	MS	$\mathrm{E}[\mathrm{MS}]$
Regression	$\mathrm{SSR} = \sum(\hat{Y}_i - \bar{Y})^2$	1	$\frac{\mathrm{SSR}}{1}$	$\sigma^2 + \beta_1^2 \sum(x_i - \bar{x})^2$
Error	$\mathrm{SSE} = \sum(Y_i - \hat{Y})^2$	$n-2$	$\frac{\mathrm{SSE}}{n-2}$	σ^2
Total	$\mathrm{SST} = \sum(Y_i - \bar{Y})^2$	$n-1$		

Analysis of Variance (ANOVA) table

Source	SS	df	MS	E[MS]
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$\frac{SSR}{1}$	$\sigma^2 + \beta_1^2 \sum(x_i - \bar{x})^2$
Error	$SSE = \sum(Y_i - \hat{Y})^2$	$n-2$	$\frac{SSE}{n-2}$	σ^2
Total	$SST = \sum(Y_i - \bar{Y})^2$	$n-1$		

Remark. $\frac{\mathrm{E}[MSR]}{\mathrm{E}[MSE]} = \begin{cases} 1 & \text{if } \beta_1 = 0 \\ > 1 & \text{if } \beta_1 \neq 0. \end{cases}$

Analysis of Variance (ANOVA) table

Source	SS	df	MS	E[MS]
Regression	$\text{SSR} = \sum(\hat{Y}_i - \bar{Y})^2$	1	$\frac{\text{SSR}}{1}$	$\sigma^2 + \beta_1^2 \sum(x_i - \bar{x})^2$
Error	$\text{SSE} = \sum(Y_i - \hat{Y})^2$	$n-2$	$\frac{\text{SSE}}{n-2}$	σ^2
Total	$\text{SST} = \sum(Y_i - \bar{Y})^2$	$n-1$		

Remark. $\frac{\mathrm{E}[\text{MSR}]}{\mathrm{E}[\text{MSE}]} = \begin{cases} 1 & \text{if } \beta_1 = 0 \\ > 1 & \text{if } \beta_1 \neq 0. \end{cases}$

Loosely, we expect MSR to be larger than MSE when $\beta_1 \neq 0$.

F-test for β_1

Proposition. Under $\mathrm{H}_0 : \beta_1 = 0$, we have $\frac{\mathrm{MSR}}{\mathrm{MSE}} \sim F_{1,n-2}$.

F-test for β_1

Proposition. Under $H_0 : \beta_1 = 0$, we have $\frac{\text{MSR}}{\text{MSE}} \sim F_{1,n-2}$.

Proof.

$$\frac{\text{MSR}}{\text{MSE}} = \frac{\frac{\text{SSR}}{\sigma^2}/1}{\frac{\text{SSE}}{\sigma^2}/(n-2)} = \frac{\chi_1^2/1}{\chi_{n-2}^2/(n-2)} \sim F_{1,n-2},$$

F-test for β_1

Proposition. Under $H_0 : \beta_1 = 0$, we have $\frac{\text{MSR}}{\text{MSE}} \sim F_{1,n-2}$.

Proof.

$$\frac{\text{MSR}}{\text{MSE}} = \frac{\frac{\text{SSR}}{\sigma^2}/1}{\frac{\text{SSE}}{\sigma^2}/(n-2)} = \frac{\chi_1^2/1}{\chi_{n-2}^2/(n-2)} \sim F_{1,n-2},$$

because SSR and SSE are statistically independent (Ch 5). $\square$

F-test for β_1

Proposition. Under $H_0 : \beta_1 = 0$, we have $\frac{\text{MSR}}{\text{MSE}} \sim F_{1,n-2}$.

Proof.

$$\frac{\text{MSR}}{\text{MSE}} = \frac{\frac{\text{SSR}}{\sigma^2}/1}{\frac{\text{SSE}}{\sigma^2}/(n-2)} = \frac{\chi_1^2/1}{\chi_{n-2}^2/(n-2)} \sim F_{1,n-2},$$

because SSR and SSE are statistically independent (Ch 5). $\square$

Remark. The F-test and t-test for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ are equivalent, since

F-test for β_1

Proposition. Under $H_0 : \beta_1 = 0$, we have $\frac{\text{MSR}}{\text{MSE}} \sim F_{1,n-2}$.

Proof.

$$\frac{\text{MSR}}{\text{MSE}} = \frac{\frac{\text{SSR}}{\sigma^2}/1}{\frac{\text{SSE}}{\sigma^2}/(n-2)} = \frac{\chi_1^2/1}{\chi_{n-2}^2/(n-2)} \sim F_{1,n-2},$$

because SSR and SSE are statistically independent (Ch 5). $\square$

Remark. The F-test and t-test for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ are equivalent, since

$$\frac{\text{MSR}}{\text{MSE}} = \frac{b_1^2 \sum (x_i - \bar{x})^2}{\text{MSE}} = \frac{b_1^2}{\text{MSE}/\sum (x_i - \bar{x})^2} = \frac{b_1^2}{\mathrm{s}^2[b_1]} = \left(\frac{b_1 - 0}{\mathrm{s}[b_1]}\right)^2.$$

F-test for β_1

Proposition. Under $H_0 : \beta_1 = 0$, we have $\frac{MSR}{MSE} \sim F_{1,n-2}$.

Proof.

$$\frac{MSR}{MSE} = \frac{\frac{SSR}{\sigma^2}/1}{\frac{SSE}{\sigma^2}/(n-2)} = \frac{\chi_1^2/1}{\chi_{n-2}^2/(n-2)} \sim F_{1,n-2},$$

because SSR and SSE are statistically independent (Ch 5). □

Remark. The F-test and t-test for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ are equivalent, since

$$\frac{MSR}{MSE} = \frac{b_1^2 \sum(x_i - \bar{x})^2}{MSE} = \frac{b_1^2}{MSE/\sum(x_i - \bar{x})^2} = \frac{b_1^2}{s^2[b_1]} = \left(\frac{b_1 - 0}{s[b_1]}\right)^2.$$

and both are generalized likelihood ratio tests (GLRT).

ANOVA F-test in R

```
> poverty = read.table("path/to/poverty.txt", h = T, sep = "\t")
> my_model = lm(Poverty ~ Graduates, data=poverty)
> anova(my_model)
Analysis of Variance Table

Response: Poverty
          Df Sum Sq Mean Sq F value    Pr(>F)
Graduates  1 267.88 267.881  61.809 3.109e-10 ***
Residuals 49 212.37   4.334

> summary(my_model)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.78097    6.80260   9.523 9.94e-13 ***
Graduates   -0.62122    0.07902  -7.862 3.11e-10 ***
```

ANOVA F-test in R

```
> poverty = read.table("path/to/poverty.txt", h = T, sep = "\t")
> my_model = lm(Poverty ~ Graduates, data=poverty)
> anova(my_model)
Analysis of Variance Table

Response: Poverty
          Df Sum Sq Mean Sq F value    Pr(>F)
Graduates  1 267.88 267.881  61.809 3.109e-10 ***
Residuals 49 212.37   4.334

> summary(my_model)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.78097    6.80260   9.523 9.94e-13 ***
Graduates   -0.62122    0.07902  -7.862 3.11e-10 ***
```

p-values of F-test and t-test for $H_0 : \beta_0 = 0$ are same.

Inferences on β_1 and β_0

Inferences on $\mathrm{E}[Y]$ and $\hat{Y}$

Analysis of Variance Approach

Coefficient of Determination

Coefficient of Determination

Definition. The **coefficient of determination** is

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}},$$

the proportion of total sample variation in Y that is explained by its linear relationship with x.

Coefficient of Determination

Definition. The **coefficient of determination** is

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}},$$

the proportion of total sample variation in Y that is explained by its linear relationship with x.

Note:

- $0 \leq R^2 \leq 1$.
- $R^2 = 1 \Rightarrow$ data perfectly linear.
- $R^2 = 0 \Rightarrow$ regression line horizontal ($b_1 = 0$).

Coefficient of Determination

Definition. The **coefficient of determination** is

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}},$$

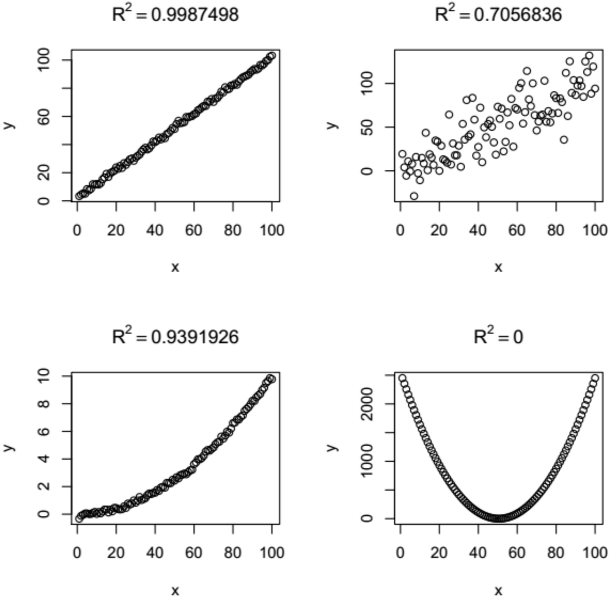
the proportion of total sample variation in Y that is explained by its linear relationship with x.

Note:

- $0 \leq R^2 \leq 1$.
- $R^2 = 1 \Rightarrow$ data perfectly linear.
- $R^2 = 0 \Rightarrow$ regression line horizontal ($b_1 = 0$).

The closer R^2 is to one, the greater the linear relationship between x and Y.

R^2 for different data sets



Sample correlation r

Let

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

be the sample correlation between x and Y.

Sample correlation r

Let

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

be the sample correlation between x and Y.

Exercise. Show that

- $R^2 = r^2$,
- $\mathrm{sgn}(r) = \mathrm{sgn}(b_1)$.

Sample correlation r

Let

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

be the sample correlation between x and Y.

Exercise. Show that

- $R^2 = r^2$,
- $\mathrm{sgn}(r) = \mathrm{sgn}(b_1)$.

Remark.

- $r \approx 0 \Rightarrow$ little linear association b/w x and Y
- $r \approx 1 \Rightarrow$ strong positive, linear association b/w x and Y
- $r \approx -1 \Rightarrow$ strong negative, linear association b/w x and Y.

Cautions about R^2 and r

It is possible that

- $R^2 \approx 1$, but the $\mathrm{E}[Y_i]$ may not lay on a line (Why?)

Cautions about R^2 and r

It is possible that

- $R^2 \approx 1$, but the $\mathrm{E}[Y_i]$ may not lay on a line (Why?)
- $R^2 \not\approx 1$, but a line is best for $\mathrm{E}[Y_i]$ (Why?)

Cautions about R^2 and r

It is possible that

- $R^2 \approx 1$, but the $\mathrm{E}[Y_i]$ may not lay on a line (Why?)
- $R^2 \not\approx 1$, but a line is best for $\mathrm{E}[Y_i]$ (Why?)
- $R^2 \approx 0$, but x and Y are highly related (Why?)

Cautions about R^2 and r

It is possible that

- $R^2 \approx 1$, but the $\mathrm{E}[Y_i]$ may not lay on a line (Why?)
- $R^2 \not\approx 1$, but a line is best for $\mathrm{E}[Y_i]$ (Why?)
- $R^2 \approx 0$, but x and Y are highly related (Why?)

Poverty vs HS Graduation data:

```
> summary(my_model)
...
Residual standard error: 2.082 on 49 degrees of freedom
Multiple R-squared:  0.5578, Adjusted R-squared:  0.5488
F-statistic: 61.81 on 1 and 49 DF,  p-value: 3.109e-10
```

Cautions about regression

- Concluding that x and Y are linearly related (that $\beta_1 \neq 0$) does not imply a causal relationship between x and Y.

Cautions about regression

- Concluding that x and Y are linearly related (that $\beta_1 \neq 0$) does not imply a causal relationship between x and Y. (Correlation does not imply causation!)

Cautions about regression

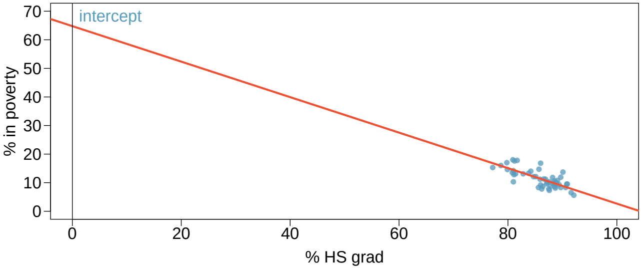
- Concluding that x and Y are linearly related (that $\beta_1 \neq 0$) does not imply a causal relationship between x and Y. (Correlation does not imply causation!)
- Beware of extrapolation: predicting Y for x far outside the range of x in the data. The relationship may not hold outside of the observed x-values.

Cautions about regression

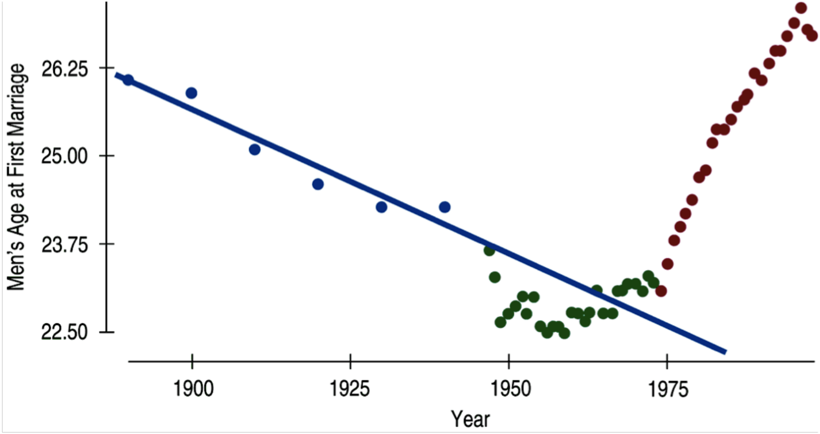
- Concluding that x and Y are linearly related (that $\beta_1 \neq 0$) does not imply a causal relationship between x and Y. (Correlation does not imply causation!)
- Beware of extrapolation: predicting Y for x far outside the range of x in the data. The relationship may not hold outside of the observed x-values.
  - Sometimes the intercept might be an extrapolation.

Cautions about regression

- Concluding that x and Y are linearly related (that $\beta_1 \neq 0$) does not imply a causal relationship between x and Y. (Correlation does not imply causation!)
- Beware of extrapolation: predicting Y for x far outside the range of x in the data. The relationship may not hold outside of the observed x-values.
  - Sometimes the intercept might be an extrapolation.



Examples of extrapolation



Examples of extrapolation

Momentous sprint at the 2156 Olympics?

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

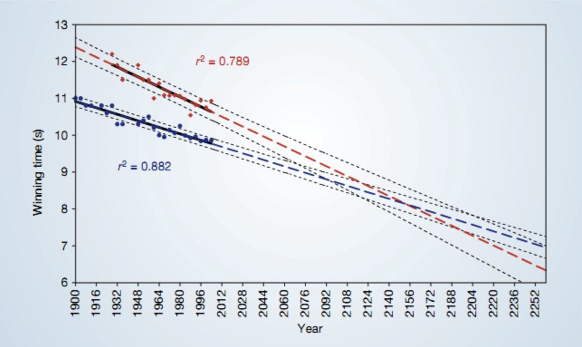


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) 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.