PUBH 7405 - Block 5 SLR

LAND ACKNOWLEDGEMENT

The School of Public Health at the University of Minnesota Twin Cities is built within the traditional homelands of the Dakota people. Minnesota comes from the Dakota name for this region, Mni Sóta Maḳoce, which loosely translates to the land where the waters reflect the skies.

It is important to acknowledge the peoples on whose land we live, learn, and work as we seek to improve and strengthen our relations with our tribal nations. We also acknowledge that words are not enough. We must ensure that our institution provides support, resources, and programs that increase access to all aspects of higher education for our American Indian students, staff, faculty, and community members.

Simple Linear Regression

Mean Function

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

where:

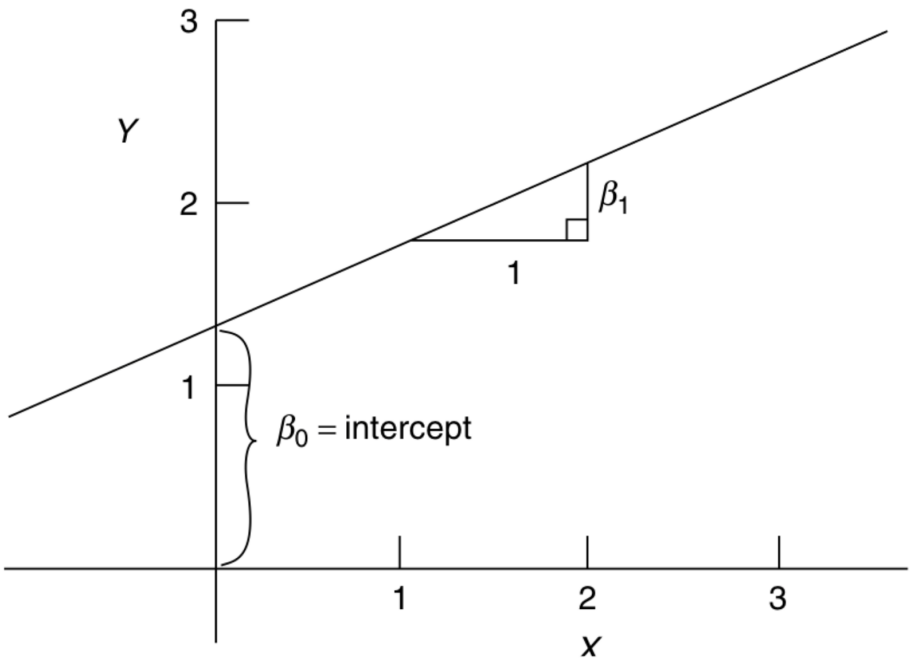
- β_0 is the intercept (value of $E(Y|X = x)$ when $x = 0$),
- β_1 is the slope (rate of change in $E(Y|X = x)$ for a unit change in x).

Variance Function

$$\text{Var}(Y|X = x) = \sigma^2 \quad (\sigma^2 > 0)$$

Note: β_0 is the value of $E(Y|X = x)$ when $x = 0$. β_1 is the rate of change in $E(Y|X = x)$ for a unit change in X.

(β_0, β_1) are unknown.



Something to Notice

Since $\sigma^2 > 0$, we have:

$$y_i \neq E(Y|X = x_i)$$

How to account for this difference?

We write:

$$y_i = E(Y|X = x_i) + e_i \quad \text{(statistical error)}$$

where:

$$e_i = y_i - E(Y|X = x_i)$$

Note: e_i depends on unknown parameters and is not observable.

Two Important Assumptions about Errors:

1) $E(e_i|X = x_i) = 0$

$\Rightarrow$ scatterplot (hypothetical) of e_i vs x_i has no patterns.

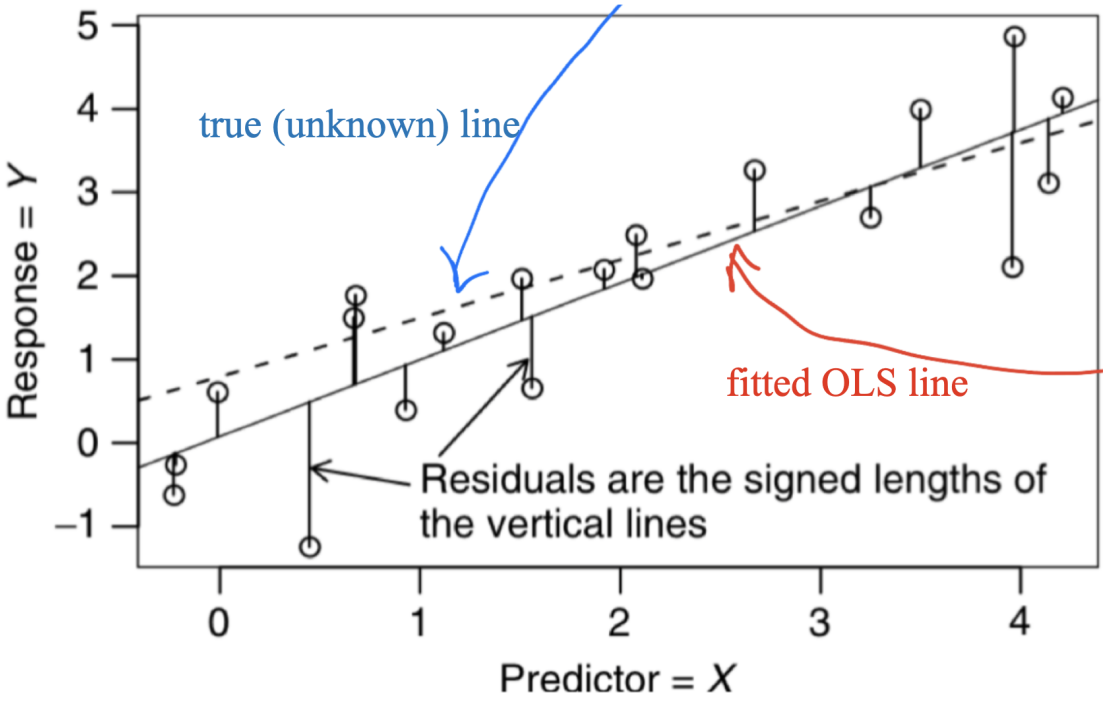
2) Errors are independent.

Constant variance assumption implicit from above.

Assuming the model is true and with a sample of data $\{(X_i, Y_i); i = 1, \ldots, n\}$ where the pairs of (X_i, Y_i) are independent.

Q: How to estimate β_0, β_1?

Method of Least Squares



Ordinary Least Squares Estimation

The residual sum of squares (RSS) is given by:

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Taking the partial derivatives with respect to β_0 and β_1:

$$\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

From this, we derive the normal equations:

$$\beta_0 n + \beta_1 \sum x_i = \sum y_i$$

$$\beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i$$

These are known as the "Normal Equations."

$\Rightarrow$ We can also express:

$$SXX = \sum_i (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$$

$$SXY = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$$

Table 2.1 Definitions of Symbols[a]

Quantity	Definition	Description
$\bar{x}$	$\sum x_i/n$	Sample average of x
$\bar{y}$	$\sum y_i/n$	Sample average of y
SXX	$\sum(x_i-\bar{x})^2 = \sum(x_i-\bar{x})x_i$	Sum of squares for the xs
SD_x^2	SXX$/(n-1)$	Sample variance of the xs
SD_x	$\sqrt{\text{SXX}/(n-1)}$	Sample standard deviation of the xs
SYY	$\sum(y_i-\bar{y})^2 = \sum(y_i-\bar{y})y_i$	Sum of squares for the ys
SD_y^2	SYY$/(n-1)$	Sample variance of the ys
SD_y	$\sqrt{\text{SYY}/(n-1)}$	Sample standard deviation of the ys
SXY	$\sum(x_i-\bar{x})(y_i-\bar{y}) = \sum(x_i-\bar{x})y_i$	Sum of cross-products
s_{xy}	SXY$/(n-1)$	Sample covariance
r_{xy}	$s_{xy}/(SD_x SD_y)$	Sample correlation

[a]In each equation, the symbol Σ means to add over all n values or pairs of values in the data.

$$\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad , \quad \hat{\beta}_1 = \frac{SXY}{SXX}$$

The fitted values are:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad \text{for } i = 1, \ldots, n$$

Note:

$$\hat{\beta}_1 = r_{xy}\frac{SD_y}{SD_x} = r_{xy}\left(\frac{SYY}{SXX}\right)^{\prime\prime 2}$$

Note: We are conditioning on X, so by convention:

$$\beta_0 = \beta_0|X \quad \text{and} \quad \beta_1 = \beta_1|X$$

$E(\hat{\beta}_1)$:

Rewrite:

$$\hat{\beta}_1 = \sum_i c_i y_i \quad \text{with } c_i = \frac{(x_i - \bar{x})}{SXX} \quad \text{(fixed)}$$

$$\Rightarrow E(\hat{\beta}_1) = E\left(\sum_i c_i y_i\right) = \sum_i c_i E(y_i) = \sum_i c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_i c_i + \beta_1 \sum_i c_i x_i$$

Note:

$$\sum_i c_i = 0 \quad \text{and} \quad \sum_i c_i x_i = 1 \Rightarrow E(\hat{\beta}_1) = \beta_1$$

$E(\hat{\beta}_0)$:

$$E(\hat{\beta}_0) = E(\bar{y}) - E(\hat{\beta}_1)\bar{x} = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$$

$\therefore \hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1 respectively.

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\sum_i c_i y_i\right) = \sum_i c_i^2 \mathrm{Var}(y_i) = \sigma^2 \sum_i c_i^2 = \frac{\sigma^2}{SXX}$$

$$\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(\bar{y} - \hat{\beta}_1\bar{x}) = \mathrm{Var}(\bar{y}) + \bar{x}^2\mathrm{Var}(\hat{\beta}_1) - 2\bar{x}\,\mathrm{Cov}(\bar{y}, \hat{\beta}_1)$$

$$\mathrm{Cov}(\bar{y}, \hat{\beta}_1) = \mathrm{Cov}\left(\frac{1}{n}\sum y_i, \sum c_i y_i\right) = \frac{1}{n}\sum c_i\,\mathrm{Cov}(y_i, y_i) = \frac{\sigma^2}{n}\sum c_i = 0$$

Since the y_i are independent by assumption and $\sum c_i = 0$,

$$\therefore \mathrm{Var}(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)$$

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \text{Cov}\left(\bar{y} - \hat{\beta}_1\bar{x}, \hat{\beta}_1\right) = \text{Cov}(\bar{y}, \hat{\beta}_1) - \bar{x}\,\text{Cov}(\hat{\beta}_1, \hat{\beta}_1)$$

$$= 0 - \sigma^2\frac{\bar{x}}{S_{XX}} = -\sigma^2\frac{\bar{x}}{S_{XX}}$$

We can use these results to get the variance of $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (fitted value):

$$\text{Var}(\hat{y})(\text{Var}(\hat{y}|X = x)) = \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x)$$

$$= \text{Var}(\hat{\beta}_0) + x^2\,\text{Var}(\hat{\beta}_1) + 2x\,\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$$

$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right) + \sigma^2\frac{x^2}{SXX} - 2\sigma^2\frac{x\bar{x}}{SXX}$$

$$= \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}\right)$$

Interesting Fact:

$$\hat{E}(Y|X = \bar{x}) = \bar{y} - \hat{\beta}_1\bar{x} = \bar{y}$$

$\therefore$ the fitted line passes through the point $(\bar{x}, \bar{y})$.

Example: Systolic Blood Pressure

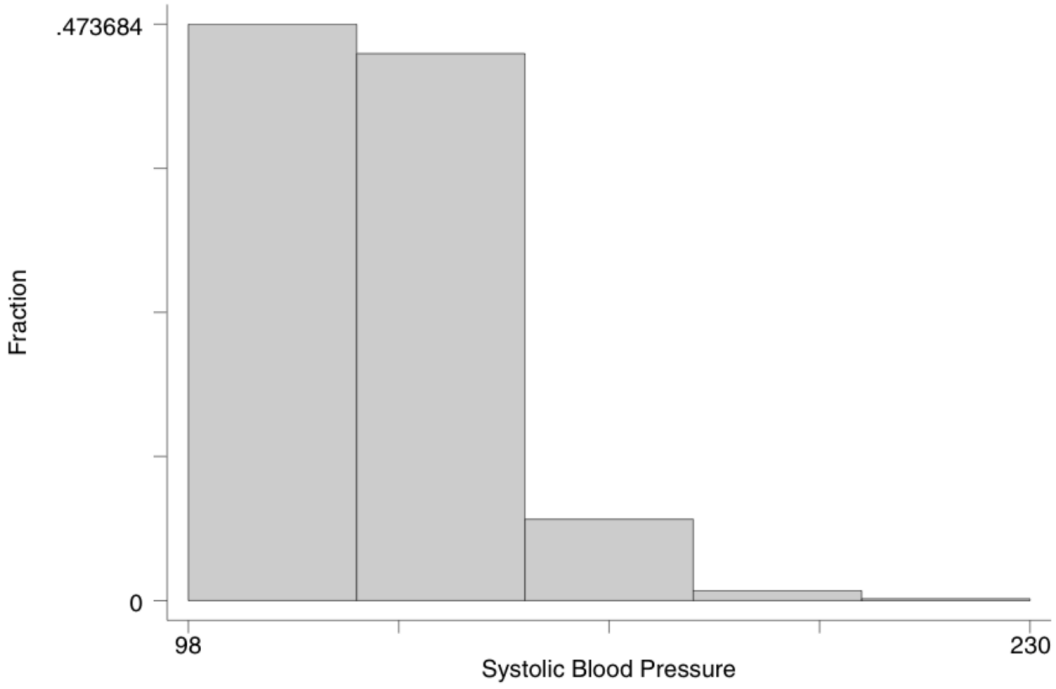
The Western Collaborative Group Study (WCGS) was a large epidemiological study designed to investigate the association between the "type A" behavior pattern and coronary heart disease (Rosenman *et al.*, 1964). We will revisit this study later in the book, focusing on the primary outcome, but for now we want to explore the distribution of systolic blood pressure (SBP).

Numerical Description

As a first step we obtain basic descriptive statistics for SBP. Table 2.1 gives detailed summary statistics for the systolic blood pressure variable, `sbp`. Several

Table 2.1. Numerical Description of Systolic Blood Pressure

```
. summarize sbp, detail

                        systolic BP
-------------------------------------------------------------
      Percentiles      Smallest
 1%           104            98
 5%           110           100
10%           112           100       Obs                3154
25%           120           100       Sum of Wgt.        3154

50%           126                     Mean           128.6328
                        Largest       Std. Dev.      15.11773
75%           136           210
90%           148           210       Variance       228.5458
95%           156           212       Skewness       1.204397
99%           176           230       Kurtosis       5.792465
```



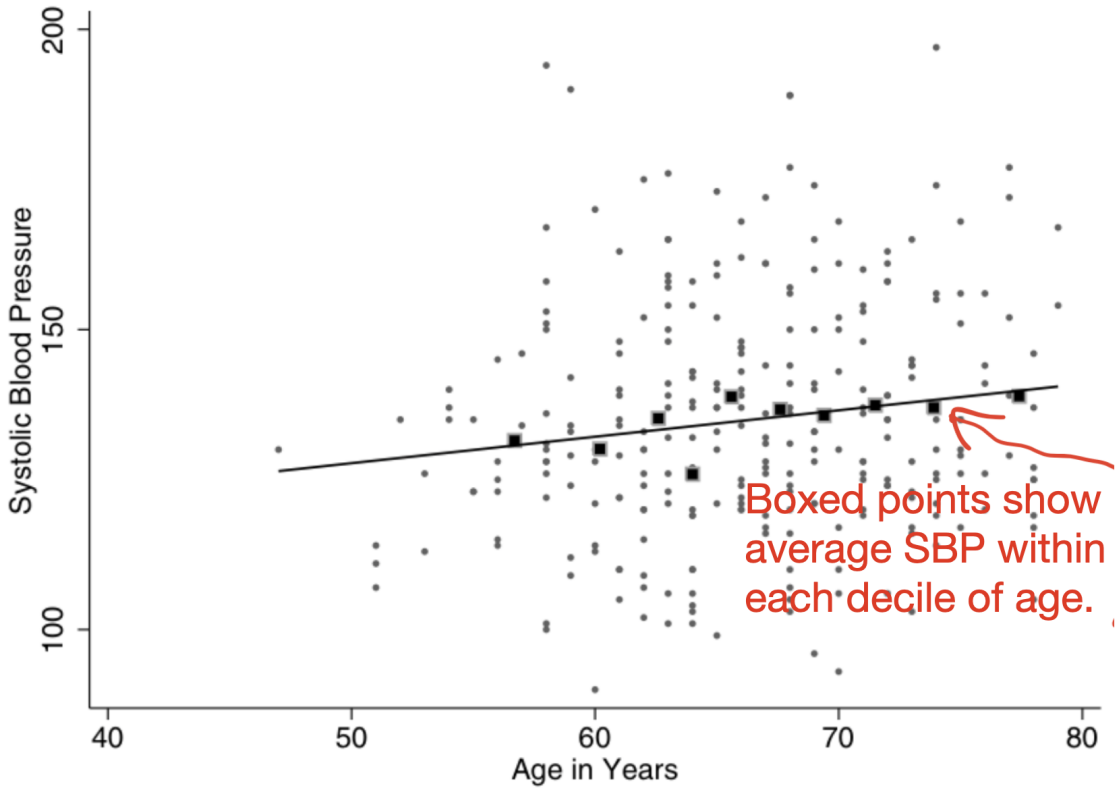


Fig. 3.1. Linear Regression Model for SBP and Age

Here:

$\hat{\beta}_0 = 105.7 \quad \Rightarrow \quad$ average SBP at age $= 0$ (not meaningful here)

$\hat{\beta}_1 = 0.44 \quad \Rightarrow \quad$ among women with heart disease, average SBP increases by 0.44 mmHg for each 1 year increase in age

Estimating the Variance σ^2

Logic: σ^2 is essentially the average squared size of the e_i^2.

$\Rightarrow \hat{\sigma}^2$ is obtained by averaging squared residuals.

Under the assumption that errors are uncorrelated random variables with mean 0 and variance σ^2, we have:

$$\Rightarrow \frac{RSS}{n} = \sum_i \frac{\hat{e}_i^2}{n} = \hat{\sigma}^2 \quad \text{(Candidate to estimate } \sigma^2\text{)}.$$

But:

$$E(\tilde{\sigma}^2) \neq \sigma^2 \quad \text{(biased)}.$$

We must use instead: $n - 2$:# parameters estimated (df).

$$\hat{\sigma}^2 = \frac{RSS}{n-2}$$

Now: $E(\hat{\sigma}^2) = \sigma^2$.

Confidence Intervals and t-Tests

Table 3.4. OLS Regression of SBP on Age

```
. reg SBP age

      Source |       SS       df       MS              Number of obs =     276
-------------+------------------------------           F(  1,   274) =    5.58
       Model |  2179.70702     1  2179.70702            Prob > F      =  0.0188
    Residual |  106991.347   274   390.47937            R-squared     =  0.0200
-------------+------------------------------           Adj R-squared =  0.0164
       Total |  109171.054   275  396.985652            Root MSE      =  19.761

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4405286    .186455     2.36   0.019     .0734621    .8075952
       _cons |    105.713   12.40238     8.52   0.000      81.2969     130.129
------------------------------------------------------------------------------
```

Point estimates SEs — Now turn our attention to these quantities

Confidence intervals result in interval estimates for true parameters.
Tests provide a methodology to make decisions concerning the value of a parameter or fitted value.

When errors are assumed to also be normally distributed:

$\Rightarrow$ Parameter estimates and fitted values will be normally distributed.

$\Rightarrow$ All linear combinations of the y_i and hence the e_i .

Confidence intervals (CIs) and tests are based on t-distribution theory:

- Appropriate with normal estimates but using $\hat{\sigma}^2$ to estimate σ^2.
- t-distribution is indexed by degrees of freedom (df) associated with $\hat{\sigma}^2$:

$$\Rightarrow t(\alpha/2, cl)$$

Quantile for $\alpha/2 \times 100\%$ in upper tail of the t-distribution, cl $\rightarrow$ df

The intercept First note:

$$SE(\hat{\beta}_0) = \sqrt{\text{Var}(\hat{\beta}_0)} = \hat{\sigma}\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right) \quad \text{(plugging in } \hat{\sigma} \text{ for } \sigma\text{)}$$

The $(1-\alpha) \times 100\%$ confidence interval (CI) for β_0 is:

$$\hat{\beta}_0 - t\,(\alpha/2, n-2)\, SE(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t\,(\alpha/2, n-2)\, SE(\hat{\beta}_0)$$

Table 3.4. OLS Regression of SBP on Age

```
. reg SBP age

      Source |       SS       df       MS              Number of obs =     276
-------------+------------------------------           F(  1,   274) =    5.58
       Model |  2179.70702     1  2179.70702            Prob > F      =  0.0188
    Residual |  106991.347   274   390.47937            R-squared     =  0.0200
-------------+------------------------------           Adj R-squared =  0.0164
       Total |  109171.054   275  396.985652            Root MSE      =  19.761

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4405286    .186455     2.36   0.019     .0734621    .8075952
       _cons |    105.713   12.40238     8.52   0.000      81.2969     130.129
------------------------------------------------------------------------------
```

test statistic for $H_0 : \beta_0 = 0$ 95% CI for β_0.

A hypothesis test of: (often interested in $\beta_0^* = 0$)

$$H_0 : \beta_0 = \beta_0^*, H_1 : \beta_0 \neq \beta_0^*$$

The test statistic is:

$$t = \frac{\beta_0 - \beta_0^*}{SE(\hat{\beta}_0)} \overset{H_0}{\sim} t(n-2)$$

$$\Rightarrow p\text{-value} \quad (\approx 0 \text{ in above Table 3.4}).$$

Note: Since H_1 is two-sided, the p-value corresponds to $P\left(t(n-2) > |t|\right)$.

The Slope

A $(1-\alpha) \times 100\%$ confidence interval (CI) for the slope is the set of β_1 such that: $\hat{\beta}_1 - t\left(\alpha/2, n-2\right) SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t\left(\alpha/2, n-2\right) SE(\hat{\beta}_1)$

$$\left[\text{recall } SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SXX}} = \sqrt{\text{Var}(\hat{\beta}_1)}\right]$$

Table 3.4. OLS Regression of SBP on Age

```
. reg SBP age

      Source |       SS       df       MS              Number of obs =     276
-------------+------------------------------           F(  1,   274) =    5.58
       Model |  2179.70702     1  2179.70702            Prob > F      =  0.0188
    Residual |  106991.347   274   390.47937            R-squared     =  0.0200
-------------+------------------------------           Adj R-squared =  0.0164
       Total |  109171.054   275  396.985652            Root MSE      =  19.761

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4405286    .186455     2.36   0.019     .0734621    .8075952
       _cons |    105.713   12.40238     8.52   0.000      81.2969     130.129
------------------------------------------------------------------------------
```

test statistic for $H_0 : \beta_1 = 0$ — 95% CI for β_1. (slope effect of age on SBP)

A Hypothesis Test of Particular Interest

A hypothesis test of:

$$H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$$

Similar to testing the intercept, the test statistic is:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \quad \overset{H_0}{\sim} \quad t(n-2)$$

The p-value (two-sided) is:

$$P\left(t(n-2) > |t|\right)$$

In above Table 3.4, p = 0.019, suggesting that SBP is linearly related to age.

Prediction

Estimated mean function $\hat{E}(Y|X)$ can be used to obtain values of response for given values of X.

Two important classes:

1. **Prediction** – new case (X^*) that was not used to estimate parameters.
2. **Fitted values** – want to estimate $E(Y|X = x^*)$.

Let's focus on prediction first. That is, we want to estimate Y^* at a value x^* (not yet observed).

Must assume the data used to estimate $E(Y|X)$ is relevant to the new case.

Given this, a point prediction of y^* is:

$$\hat{y^*} = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

This is a prediction of yet unobserved y^*.

Assuming the model is correct: $y^* = \beta_0 + \beta_1 x^* + e^*$

e^* : New random error with same assumptions.

Note: Even if β_0, β_1 were known, predictions would not match true values perfectly. They would be off by a random quantity with variance σ^2.

Two sources of variation:

1. Uncertainty in $\hat{\beta}_0, \hat{\beta}_1 \Rightarrow \text{var}(\hat{y})$
2. Variance of e^*

Combining these:

$$\text{Var}(\tilde{y}^*|X = x^*) = \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX} \right) + \sigma^2$$

This is the variance of $\hat{y} \uparrow$

$$\therefore SE(\tilde{y}^*|X = x^*) = \sqrt{\text{Var}(\tilde{y}^*|X = x^*)}$$

We can define a $(1 - \alpha) \times 100\%$ prediction interval for y^* (true value) as:

$$\tilde{y}^* \pm t\,(\alpha/2n - 2)\, SE(\tilde{y}^*)$$

$$\tilde{y}^* \equiv \hat{E}(Y|X = x^*) \quad SE(\tilde{y}^*) \equiv SE(\tilde{y}^*|X = x^*)$$

Using previous notation:

$$\tilde{y}^* - t\,(\alpha/2, n - 2)\, SE(\tilde{y}^*) \leq y^* \leq \tilde{y}^* + t\,(\alpha/2, n - 2)\, SE(\tilde{y}^*)$$

Now move to the fitted values. Here, we want to estimate:

$$E(Y|X = x^*)$$

Point estimate:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^* = \hat{E}(Y|X = x^*)$$

Standard error of the fitted value:

$$SE(\hat{y}|X = x^*) = \hat{\sigma}\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX}\right)^{\prime\prime 2}$$

$$\sqrt{\text{Var}(\hat{y}|X = x^*)} \uparrow$$

A $(1-\alpha) \times 100\%$ CI for $E(Y|X = x^*)$ is:

$$\hat{y} - t\,(\alpha/2, n-2)\,SE(\hat{y}) \leq E(Y|X = x^*) \leq \hat{y} + t\,(\alpha/2, n-2)\,SE(\hat{y})$$

$$\hat{y}^* \equiv \hat{E}(Y|X = x^*) \quad SE(\hat{y}^*) \equiv SE(\hat{y}^*|X = x^*)$$

Coefficient of Determination, R^2

Ignoring the predictor, the best prediction of y is $\bar{y}$.
The total sum of squares (SYY) represents the observed total variation, ignoring all predictors:

$$SYY = \sum_i (y_i - \bar{y})^2$$

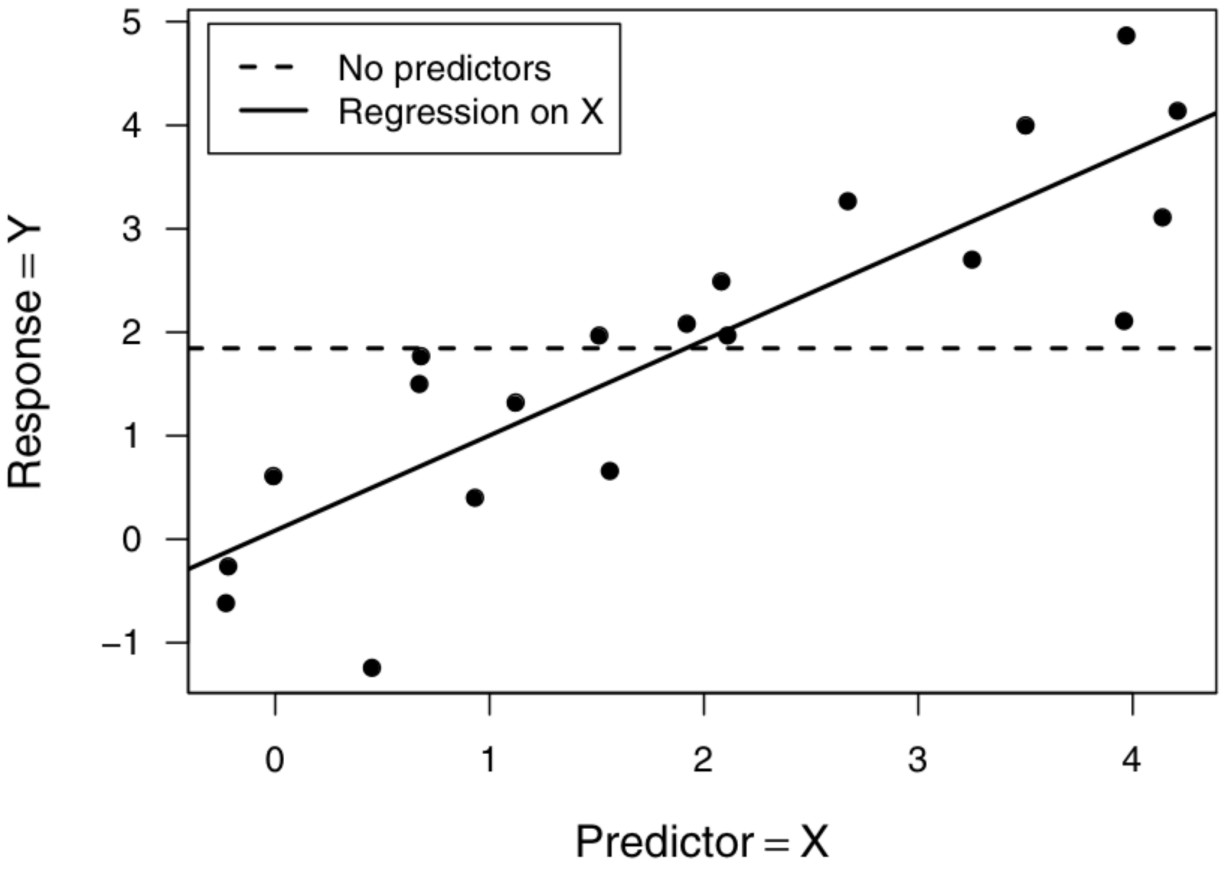
The sum of squares due to regression (SS_{reg}) is:

$$SS_{\text{reg}} = SYY - RSS$$

$$= SYY - \left(SYY - \frac{(SXY)^2}{SXX}\right)$$

$$= \frac{(SXY)^2}{SXX}$$

$$\Rightarrow \frac{SS_{\text{reg}}}{SYY} = 1 - \frac{RSS}{SYY}$$



LHS is the proportion of observed variability in response explained by regression on the predictor.

RHS is 1− remaining unexplained variability.

Then define:

$$R^2 = \frac{SS_{\text{reg}}}{SYY} = 1 - \frac{RSS}{SYY}$$

R^2 : coefficient of determination

- Scale-free
- One-number summary of the strength of the relationship of y on x.

Fact:

$$R^2 = \frac{SS_{\text{reg}}}{SYY} = \frac{(SXY)^2}{SXX \cdot SYY} = r_{xy}^2$$

r_{xy}^2: square of sample correlation

Adjusted R^2

$$R^2_{\text{adj}} = 1 - \frac{RSS/df}{SYY/(n-1)}$$

- Adds correction for degrees of freedom (df).
  - Can facilitate comparing models in multiple linear regression.

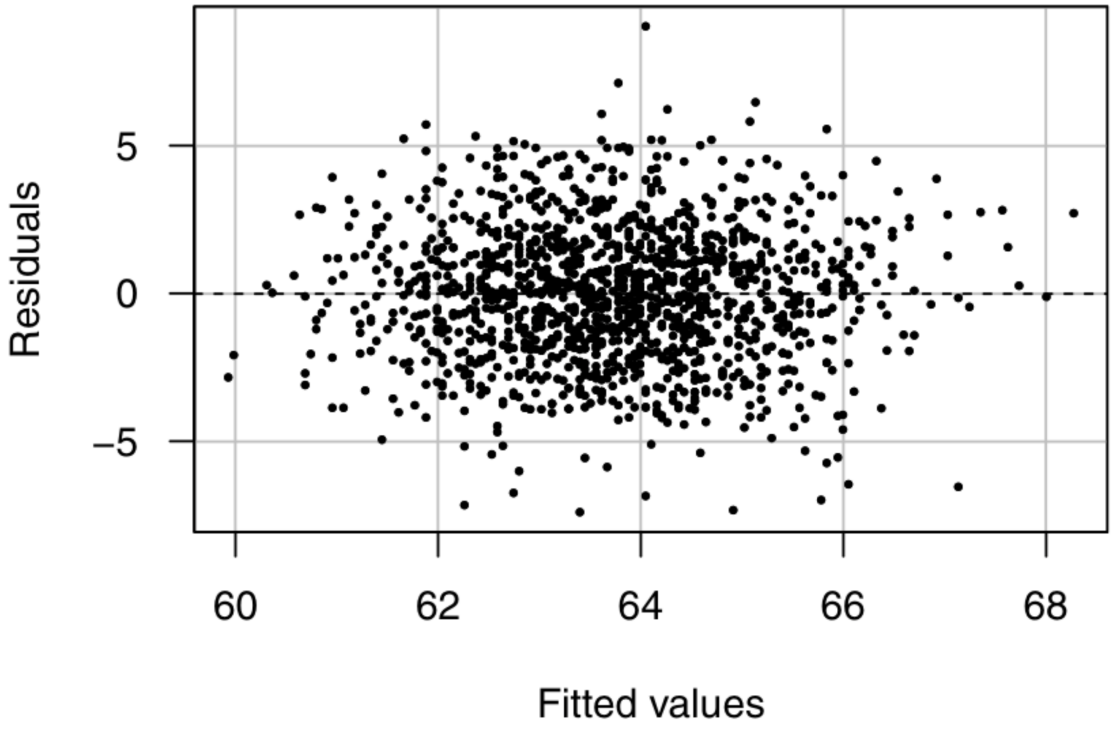
Table 3.4. OLS Regression of SBP on Age

```
. reg SBP age

      Source |       SS       df       MS              Number of obs =     276
-------------+------------------------------           F(  1,   274) =    5.58
       Model |  2179.70702     1  2179.70702            Prob > F      =  0.0188
    Residual |  106991.347   274   390.47937            R-squared     =  0.0200
-------------+------------------------------           Adj R-squared =  0.0164
       Total |  109171.054   275  396.985652            Root MSE      =  19.761

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .4405286    .186455     2.36   0.019      .0734621    .8075952
       _cons |    105.713   12.40238     8.52   0.000       81.2969     130.129
------------------------------------------------------------------------------
```

Residuals

Recall: $\hat{e}_i = y_i - \hat{y}_i$

Residuals are used to find failures of assumptions.



No pattern observed.

Residuals are small compared to fitted values.

(Will say more on residual diagnostics later.)