Correlation and Simple Linear Regression

Last modified: 2026-11-12 22:02:32 (PST)

This page reviews two ways to relate two continuous variables: correlation coefficients, with tests of whether they differ from zero, and simple linear regression. It uses the t reference distribution defined on the Statistical Inference page. This page is adapted from Vittinghoff et al. [1], Chapter 3.

The HERS data

The examples on this page use the HERS data, which the Comparing Means page describes. The `rmb` R package includes the dataset; `haven::as_factor()` converts its Stata value labels to factors:

```
hers <- rmb::hers |> haven::as_factor()
hers |> head()
#> # A tibble: 6 x 37
#>   HT      age raceth nonwhite smoking drinkany exercise physact globrat poorfair
#>   <fct> <dbl> <fct>  <fct>    <fct>   <fct>    <fct>    <fct>   <fct>   <fct>
#> 1 plac~    70 Afric~ yes      no      no       no       much m~ good    no
#> 2 plac~    62 Afric~ yes      no      no       no       much l~ good    no
#> 3 horm~    69 White  no       no      no       no       about ~ good    no
#> 4 plac~    64 White  no       yes     yes      no       much l~ good    no
#> 5 plac~    65 White  no       no      no       no       somewh~ good    no
#> 6 horm~    68 Afric~ yes      no      yes      no       about ~ good    no
#> # i 27 more variables: medcond <dbl>, htnmeds <fct>, statins <fct>,
#> #   diabetes <fct>, dmpills <fct>, insulin <fct>, weight <dbl>, BMI <dbl>,
#> #   waist <dbl>, WHR <dbl>, glucose <dbl>, weight1 <dbl>, BMI1 <dbl>,
#> #   waist1 <dbl>, WHR1 <dbl>, glucose1 <dbl>, tchol <dbl>, LDL <dbl>,
#> #   HDL <dbl>, TG <dbl>, tchol1 <dbl>, LDL1 <dbl>, HDL1 <dbl>, TG1 <dbl>,
#> #   SBP <dbl>, DBP <dbl>, age10 <dbl>
```

Correlation

Testing the Pearson correlation

Definition 0.1 (Population correlation)**.** The **population correlation** of two random variables X and Y with positive variances is

$$\rho \stackrel{\text{def}}{=} \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}},$$

where $\operatorname{Cov}(X,Y)$ is their covariance.
Source: Wikipedia contributors [2].

ℹ Remark (The sample correlation estimates the population correlation)

Remark 0.1 (The sample correlation estimates the population correlation). The Pearson correlation coefficient r of a sample is the corresponding sample statistic, and estimates ρ.

Definition 0.2 (Test of zero correlation)**.** Let r be the Pearson correlation coefficient of $n \geq 3$ pairs, with $|r| < 1$. The **t-test of zero correlation** of $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$ (Definition 0.1) uses the statistic

$$t \stackrel{\text{def}}{=} \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Its p-value is $\Pr(|T| \geq |t|)$, where T has the t_{n-2} distribution (t-distribution).

Theorem 0.1 (Null distribution of the correlation t statistic)**.** *Let the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent, and let each Y_i, given $X_1, \ldots, X_n$, be Gaussian with a mean and variance that do not depend on the X values. Then the statistic t of Definition 0.2 has the t_{n-2} distribution [3, Sec. 9.6], pp. 472-473:*

$$t \sim t_{n-2}.$$

ℹ Remark (When the conditions hold)

Remark 0.2 (When the conditions hold). The conditions of Theorem 0.1 hold, for example, when the pairs are independent draws from a bivariate Gaussian distribution with $\rho = 0$.

Example 0.1 (Correlation between BMI and fasting glucose in HERS)**.** Figure 1 plots baseline fasting glucose against BMI.

```
hers |>
  dplyr::filter(!is.na(BMI)) |>
  ggplot2::ggplot() +
  ggplot2::aes(x = BMI, y = glucose) +
  ggplot2::geom_point(alpha = 0.3) +
  ggplot2::geom_smooth(method = "lm", formula = y ~ x) +
  ggplot2::labs(
    x = "BMI (kg/m^2)",
    y = "Fasting glucose (mg/dL)"
  )
```

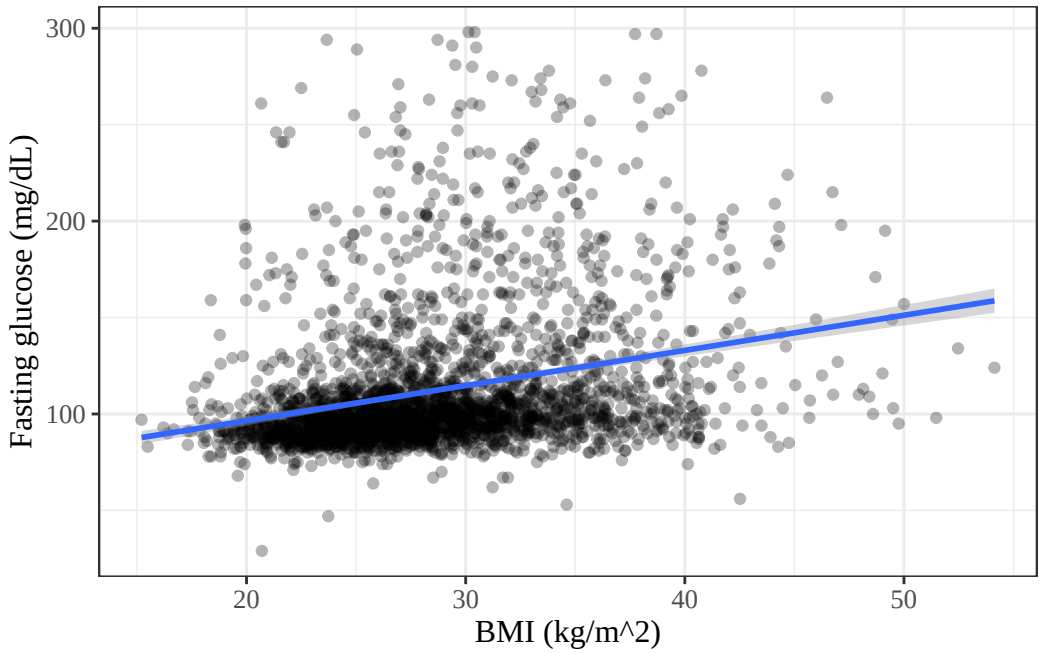


Figure 1: Baseline fasting glucose against BMI in HERS, with the least-squares line

The statistic of Definition 0.2, using the 2758 participants with a BMI measurement:

```r
hers_bmi <- hers |> dplyr::filter(!is.na(BMI))
r <- cor(hers_bmi$BMI, hers_bmi$glucose)
n <- nrow(hers_bmi)
t_cor <- r * sqrt(n - 2) / sqrt(1 - r^2)
c(r = r, t = t_cor, p_value = 2 * pt(-abs(t_cor), df = n - 2))
#>           r           t     p_value
#> 2.72664e-01 1.48779e+01 3.24201e-48
```

`cor.test()` reports the same values, and a confidence interval for ρ:

```r
cor.test(hers$BMI, hers$glucose, method = "pearson")
#>
#>  Pearson's product-moment correlation
#>
#> data:  hers$BMI and hers$glucose
#> t = 14.88, df = 2756, p-value <2e-16
#> alternative hypothesis: true correlation is not equal to 0
#> 95 percent confidence interval:
#>  0.237760 0.306865
#> sample estimates:
#>      cor
#> 0.272664
```

The correlation is positive but modest: glucose tends to be higher at higher BMI, with wide scatter around the line in Figure 1.

Read

- Hogg et al. [3, Sec. 9.6], pp. 471–476 covers tests concerning regression and correlation.
- Rice [4, Sec. 14.2.3], pp. 560–563 covers correlation and regression.

Spearman rank correlation

Definition 0.3 (Spearman rank correlation)**.** Replace each x_i by its rank u_i among $x_1, \ldots, x_n$, and each y_i by its rank v_i among $y_1, \ldots, y_n$, giving tied values the average of the ranks they span. The **Spearman rank correlation** r_S is the Pearson correlation coefficient evaluated at the rank pairs:

$$r_S \stackrel{\text{def}}{=} r((u_1, v_1), \ldots, (u_n, v_n)).$$

Source: Wikipedia contributors [5].

Remark (What the Spearman correlation measures)

Remark 0.3 (What the Spearman correlation measures). Because ranks depend only on the ordering of the values, r_S measures how close the association is to monotone, whether or not it is linear, and a single extreme value moves r_S less than it moves r.

Example 0.2 (Spearman correlation between BMI and fasting glucose in HERS)**.** The Pearson correlation of the ranks (Definition 0.3), and `cor.test()`'s Spearman test (`exact = FALSE`, because tied values rule out the exact p-value):

```
hers_bmi <- hers |> dplyr::filter(!is.na(BMI))
cor(rank(hers_bmi$BMI), rank(hers_bmi$glucose))
#> [1] 0.333751
cor.test(hers_bmi$BMI, hers_bmi$glucose, method = "spearman", exact = FALSE)
#>
#>  Spearman's rank correlation rho
#>
#> data:  hers_bmi$BMI and hers_bmi$glucose
#> S = 2.33e+09, p-value <2e-16
#> alternative hypothesis: true rho is not equal to 0
#> sample estimates:
#>      rho
#> 0.333751
```

r_S is larger than the Pearson r of Example 0.1: the association is closer to monotone than to linear.

Simple linear regression

Model specification

Definition 0.1 (Conditional Gaussian model)**.** A **conditional Gaussian model** for outcomes $Y_1, \ldots, Y_n$ with covariate values $x_1, \ldots, x_n$ (each a single value or a vector) says that, given the covariates, the outcomes are independent and Gaussian: outcome Y_i is

centered on a mean function $\mu(x)$ evaluated at x_i, and has its own variance σ_i^2:

$$Y_i \mid X_i = x_i \;\sim_{\perp\!\!\!\perp}\; \mathrm{N}(\mu_i, \sigma_i^2),$$
$$\mu_i \overset{\text{def}}{=} \mu(x_i).$$

Definition 0.2 (Deviation from the conditional mean)**.** In a conditional Gaussian model (Definition 0.1), the **deviation** $\mathbb{d}_i$ of outcome Y_i is its deviation from its population mean, taking the mean conditional on $X_i = x_i$:

$$\mathbb{d}_i \overset{\text{def}}{=} Y_i - \mu(x_i).$$

Exercise 0.1 (Decomposing an outcome)**.** Express each outcome Y_i in a conditional Gaussian model (Definition 0.1) in terms of its conditional mean μ_i and its deviation $\mathbb{d}_i$ (Definition 0.2).

Solution

Solution 0.1. Check that $Y_i = \mu_i + \mathbb{d}_i$, starting from the right-hand side:

$$\begin{aligned}
\mu_i + \mathbb{d}_i &= \mu(x_i) + \mathbb{d}_i && \text{(definition of } \mu_i\text{)} \\
&= \mu(x_i) + (Y_i - \mu(x_i)) && \text{(definition of } \mathbb{d}_i\text{)} \\
&= \mu(x_i) + Y_i - \mu(x_i) && \text{(remove the parentheses)} \\
&= Y_i + \mu(x_i) - \mu(x_i) && \text{(reorder the terms)} \\
&= Y_i + (\mu(x_i) - \mu(x_i)) && \text{(group the last two terms)} \\
&= Y_i + 0 && (a - a = 0) \\
&= Y_i && (a + 0 = a)
\end{aligned}$$

So $Y_i = \mu_i + \mathbb{d}_i$.
Alternatively, solve the definition of $\mathbb{d}_i$ for Y_i:

$$
\begin{aligned}
\mathbb{d}_i &= Y_i - \mu(x_i) && \text{(definition of } \mathbb{d}_i\text{)} \\
\mathbb{d}_i + \mu(x_i) &= Y_i - \mu(x_i) + \mu(x_i) && \text{(add } \mu(x_i) \text{ to both sides)} \\
\mathbb{d}_i + \mu(x_i) &= Y_i + (-\mu(x_i)) + \mu(x_i) && \text{(subtracting is adding the negative)} \\
\mathbb{d}_i + \mu(x_i) &= Y_i + (-\mu(x_i) + \mu(x_i)) && \text{(group the last two terms)} \\
\mathbb{d}_i + \mu(x_i) &= Y_i + 0 && (-a + a = 0) \\
\mathbb{d}_i + \mu(x_i) &= Y_i && (a + 0 = a) \\
Y_i &= \mathbb{d}_i + \mu(x_i) && \text{(swap the two sides)} \\
Y_i &= \mu(x_i) + \mathbb{d}_i && \text{(reorder the terms)} \\
Y_i &= \mu_i + \mathbb{d}_i && \text{(definition of } \mu_i\text{)}
\end{aligned}
$$

Theorem 0.1 (Additive form of the conditional Gaussian model)**.** *In a conditional Gaussian model (Definition 0.1), each outcome is its conditional mean plus its deviation (Definition 0.2):*

$$Y_i = \mu_i + \mathbb{d}_i.$$

Proof

Proof. This is the solution to Exercise 0.1. □

Exercise 0.2 (Distribution of the deviations)**.** Find the joint distribution of the deviations (Definition 0.2) in a conditional Gaussian model (Definition 0.1), given the covariate values.

Solution

Solution 0.2. Every probability, CDF and density in this solution is conditional on $X_i = x_i$; the condition is left out of the notation to keep the lines short. Write $F_{\mathbb{d}}$ and $f_{\mathbb{d}}$ for the CDF and density of $\mathbb{d}_i$, and F_Y and f_Y for those of Y_i.
First, the CDF of $\mathbb{d}_i$:

$$
\begin{aligned}
F_{\mathbb{d}}(t) &= \Pr(\mathbb{d}_i \le t) && \text{(definition of a CDF)} \\
&= \Pr(Y_i - \mu(x_i) \le t) && \text{(definition of } \mathbb{d}_i\text{)} \\
&= \Pr(Y_i - \mu_i \le t) && \text{(definition of } \mu_i\text{)} \\
&= \Pr(Y_i \le t + \mu_i) && \text{(add } \mu_i \text{ to both sides of the inequality)} \\
&= F_Y(t + \mu_i) && \text{(definition of a CDF)}
\end{aligned}
$$

Differentiating gives the density:

$$
\begin{aligned}
f_{\mathbb{d}}(t) &= \frac{\partial}{\partial t} F_{\mathbb{d}}(t) && \text{(a density is the derivative of its CDF)} \\
&= \frac{\partial}{\partial t} F_Y(t + \mu_i) && \text{(CDF of } \mathbb{d}_i\text{)} \\
&= f_Y(t + \mu_i) \frac{\partial}{\partial t}(t + \mu_i) && \text{(chain rule)}
\end{aligned}
$$

The inner derivative:

$$
\begin{aligned}
\frac{\partial}{\partial t}(t + \mu_i) &= \frac{\partial}{\partial t} t + \frac{\partial}{\partial t} \mu_i && \text{(derivative of a sum)} \\
&= 1 + 0 && (\mu_i \text{ does not depend on } t) \\
&= 1 && \text{(add)}
\end{aligned}
$$

Plugging the inner derivative back in, and using the Gaussian density of Y_i, which has mean $\mu_i = \mu(x_i)$ and variance σ_i^2 (Definition 0.1):

$$
\begin{aligned}
f_{\mathbb{d}}(t) &= f_Y(t + \mu_i) \cdot 1 && \text{(inner derivative is 1)} \\
&= f_Y(t + \mu_i) && \text{(multiply by 1)} \\
&= \frac{1}{\sigma_i \sqrt{2\pi}} \mathrm{e}^{-\frac{((t+\mu_i)-\mu_i)^2}{2\sigma_i^2}} && \text{(Gaussian density with mean } \mu_i\text{)} \\
&= \frac{1}{\sigma_i \sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2\sigma_i^2}} && \text{(subtract)}
\end{aligned}
$$

The last line is the density of $\mathrm{N}(0, \sigma_i^2)$, so $\mathbb{d}_i \mid X_i = x_i \sim \mathrm{N}(0, \sigma_i^2)$.
For independence, factor the joint CDF of $\mathbb{d}_1, \ldots, \mathbb{d}_n$, using the same CDF step as above for each i and the independence of the Y_i given the covariates (Definition 0.1):

$$
\begin{aligned}
\Pr(\mathbb{d}_1 \le t_1, \ldots, \mathbb{d}_n \le t_n) &= \Pr(Y_1 \le t_1 + \mu_1, \ldots, Y_n \le t_n + \mu_n) && \text{(add } \mu_i \text{ to both sides of each inequality)} \\
&= \prod_{i=1}^{n} \Pr(Y_i \le t_i + \mu_i) && \text{(the } Y_i \text{ are independent)} \\
&= \prod_{i=1}^{n} \Pr(\mathbb{d}_i \le t_i) && \text{(subtract } \mu_i \text{ from both sides of each inequality}
\end{aligned}
$$

The joint CDF is the product of the marginal CDFs, so $\mathbb{d}_1, \ldots, \mathbb{d}_n$ are independent. Given the covariate values, the deviations are therefore independent, with $\mathbb{d}_i \mid X_i = x_i \;\sim_{\perp\!\!\!\perp} \mathrm{N}(0, \sigma_i^2)$.

Theorem 0.2 (Distribution of the deviations)**.** *In a conditional Gaussian model (Definition 0.1), given the covariates, the deviations (Definition 0.2) are independent and Gaussian with mean 0 and variance σ_i^2:*

$$\mathbb{d}_i \mid X_i = x_i \;\sim_{\perp\!\!\!\perp}\; \mathrm{N}(0, \sigma_i^2).$$

Proof

Proof. This is the solution to Exercise 0.2. □

Definition 0.3 (Homoskedastic model)**.** A conditional Gaussian model (Definition 0.1) is **homoskedastic** if all outcomes share one variance σ^2, which does not depend on i:

$$\sigma_i^2 = \sigma^2 \text{ for all } i.$$

Source: Wikipedia contributors [6].

Corollary 0.1 (Deviations in a homoskedastic model)**.** *In a homoskedastic model (Definition 0.3), the deviations (Definition 0.2) all have the same variance σ^2:*

$$\mathbb{d}_i \mid X_i = x_i \;\sim_{\perp\!\!\!\perp}\; \mathrm{N}(0, \sigma^2).$$

Proof

Proof. By Theorem 0.2, $\mathbb{d}_i \mid X_i = x_i \;\sim_{\perp\!\!\!\perp}\; \mathrm{N}(0, \sigma_i^2)$, and Definition 0.3 sets $\sigma_i^2 = \sigma^2$ for all i. □

Definition 0.4 (Linear regression model)**.** A **linear regression model** is a homoskedastic (Definition 0.3) conditional Gaussian model (Definition 0.1) whose mean function is linear in p covariates, where observation i has covariate values $\underset{\sim}{x}_i = (x_{i1}, \ldots, x_{ip})$:

$$\mu(x_1, \ldots, x_p) \stackrel{\text{def}}{=} \beta_0 + \sum_{j=1}^{p} \beta_{x_j} x_j.$$

Source: Wikipedia contributors [7].

Definition 0.5 (Simple linear regression model)**.** A **simple linear regression** model is a linear regression model (Definition 0.4) with $p = 1$ covariate X, writing x_i for x_{i1}:

$$\mu(x) \stackrel{\text{def}}{=} \beta_0 + \beta_x x.$$

Source: Wikipedia contributors [8].

Definition 0.6 (Intercept)**.** In a simple linear regression model (Definition 0.5), the **intercept** is $\beta_0 = \mu(0)$: the mean of Y among observations with $X = 0$.

Definition 0.7 (Slope)**.** In a simple linear regression model (Definition 0.5), the **slope** is $\beta_x = \mu(x+1) - \mu(x)$: the difference in the mean of Y between two groups whose values of X differ by one unit.

ℹ Remark (Interpreting the variance parameter)

Remark 0.1 (Interpreting the variance parameter). σ^2 is the variance of Y around its mean at each value of X.

Ordinary least squares estimation

Definition 0.8 (Residual sum of squares)**.** For a regression model whose conditional mean $\mu(x; \underset{\sim}{\theta})$ depends on a parameter vector $\underset{\sim}{\theta}$, the **residual sum of squares** at $\underset{\sim}{\theta}$ is the sum of the squared residuals $\hat{\varepsilon}_i(\underset{\sim}{\theta}) \overset{\text{def}}{=} y_i - \mu(x_i; \underset{\sim}{\theta})$ that the model has when its parameters equal $\underset{\sim}{\theta}$:

$$\operatorname{RSS}(\underset{\sim}{\theta}) \overset{\text{def}}{=} \sum_{i=1}^{n} \hat{\varepsilon}_i(\underset{\sim}{\theta})^2.$$

At the estimate $\hat{\underset{\sim}{\theta}}$, $\hat{\varepsilon}_i(\hat{\underset{\sim}{\theta}})$ is the residual $\hat{\varepsilon}_i$ of the fitted model.
Source: Wikipedia contributors [9].

Example 0.1 (Residual sum of squares of a simple linear regression)**.** The mean function of a simple linear regression model has parameter vector $\underset{\sim}{\theta} = (\beta_0, \beta_x)$, so write $\operatorname{RSS}(\beta_0, \beta_x)$ for its residual sum of squares (Definition 0.8):

$$\begin{aligned}
\operatorname{RSS}(\beta_0, \beta_x) &= \sum_{i=1}^{n} \hat{\varepsilon}_i(\beta_0, \beta_x)^2 && \text{(definition of RSS)} \\
&= \sum_{i=1}^{n} (y_i - \mu(x_i; \beta_0, \beta_x))^2 && \text{(definition of } \hat{\varepsilon}_i(\underset{\sim}{\theta})\text{)} \\
&= \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_x x_i))^2 && \text{(definition of } \mu(x)\text{)} \\
&= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_x x_i)^2 && \text{(distribute the minus sign)}
\end{aligned}$$

Example 0.2 (Residual sum of squares of a line through three points)**.** For the points $(0, 1)$, $(1, 2)$, $(2, 2)$ and the line with $\beta_0 = 1$ and $\beta_x = 0.5$ (Example 0.1), the residuals are $1 - 1 = 0$, $2 - 1.5 = 0.5$, and $2 - 2 = 0$, so

$$\begin{aligned} \mathrm{RSS}(1, 0.5) &= 0^2 + 0.5^2 + 0^2 \\ &= 0.25 \end{aligned}$$

(Figure 2).

```r
rss_points <- tibble::tibble(x = c(0, 1, 2), y = c(1, 2, 2)) |>
  dplyr::mutate(fitted = 1 + 0.5 * x)

ggplot2::ggplot(rss_points, ggplot2::aes(x = x, y = y)) +
  ggplot2::geom_abline(intercept = 1, slope = 0.5) +
  ggplot2::geom_segment(
    ggplot2::aes(xend = x, yend = fitted),
    linetype = "dashed",
    color = "firebrick"
  ) +
  ggplot2::geom_point(size = 3) +
  ggplot2::coord_cartesian(xlim = c(-0.25, 2.25), ylim = c(0.75, 2.25))
```

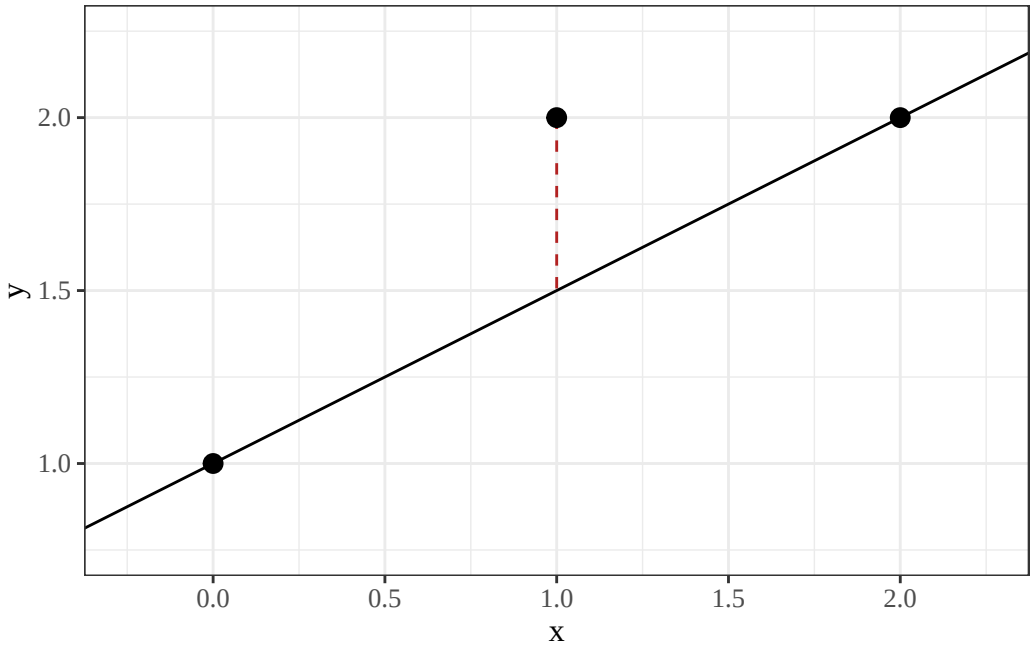


Figure 2: The three points, the line with intercept 1 and slope 0.5, and the one nonzero residual (dashed)

Exercise 0.3 (Sample mean squared error and residual sum of squares)**.** For a regression model with conditional mean $\mu(x;\underset{\sim}{\theta})$, let $\widehat{y}_i \overset{\text{def}}{=} \mu(x_i;\underset{\sim}{\theta})$ be its predictions of $n \geq 1$ outcomes $y_1, \ldots, y_n$, and let $\text{RSS}(\underset{\sim}{\theta})$ (Definition 0.8) be taken over those same outcomes. Express the sMSE of these predictions in terms of $\text{RSS}(\underset{\sim}{\theta})$, and $\text{RSS}(\underset{\sim}{\theta})$ in terms of that sMSE.

Solution

Solution 0.3. First, each squared prediction error equals the squared residual at $\underset{\sim}{\theta}$:

$$\begin{aligned}
{\varepsilon_{\text{pred}}}_i^2 &= (\hat{y}_i - y_i)^2 && \text{(definition of prediction error)} \\
&= (\mu(x_i; \underset{\sim}{\theta}) - y_i)^2 && \text{(definition of } \hat{y}_i\text{)} \\
&= (-(y_i - \mu(x_i; \underset{\sim}{\theta})))^2 && \text{(factor out } -1\text{)} \\
&= (-\hat{\varepsilon}_i(\underset{\sim}{\theta}))^2 && \text{(definition of } \hat{\varepsilon}_i(\underset{\sim}{\theta})\text{)} \\
&= \hat{\varepsilon}_i(\underset{\sim}{\theta})^2 && ((-a)^2 = a^2)
\end{aligned}$$

So the sMSE is the residual sum of squares divided by n:

$$\begin{aligned}
\text{sMSE}(\hat{y}) &= \frac{1}{n}\sum_{i=1}^{n} {\varepsilon_{\text{pred}}}_i^2 && \text{(definition of sMSE)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i(\underset{\sim}{\theta})^2 && ({\varepsilon_{\text{pred}}}_i^2 = \hat{\varepsilon}_i(\underset{\sim}{\theta})^2) \\
&= \frac{1}{n}\text{RSS}(\underset{\sim}{\theta}) && \text{(definition of RSS)}
\end{aligned}$$

Solving for $\text{RSS}(\underset{\sim}{\theta})$:

$$\begin{aligned}
\text{sMSE}(\hat{y}) &= \frac{1}{n}\text{RSS}(\underset{\sim}{\theta}) && \text{(result above)} \\
n\,\text{sMSE}(\hat{y}) &= \text{RSS}(\underset{\sim}{\theta}) && \text{(multiply both sides by } n\text{)} \\
\text{RSS}(\underset{\sim}{\theta}) &= n\,\text{sMSE}(\hat{y}) && \text{(swap the two sides)}
\end{aligned}$$

Theorem 0.3 (Sample mean squared error is the residual sum of squares divided by n)**.** *For a regression model with conditional mean $\mu(x; \underset{\sim}{\theta})$, the sMSE of its predictions $\hat{y}_i = \mu(x_i; \underset{\sim}{\theta})$ of $n \geq 1$ outcomes $y_1, \ldots, y_n$ and the residual sum of squares $RSS(\underset{\sim}{\theta})$ (Definition 0.8) over those same n outcomes satisfy*

$$\text{sMSE}(\hat{y}) = \frac{1}{n} RSS(\underset{\sim}{\theta}),$$

and equivalently

$$RSS(\underset{\sim}{\theta}) = n\,\text{sMSE}(\hat{y}).$$

Proof

Proof. Both equations are derived in Exercise 0.3. □

Remark (The same observations on both sides)

Remark 0.2 (The same observations on both sides). The identities in Theorem 0.3 need the sMSE and the residual sum of squares to be taken over the same n observations. The RSS minimized in fitting is taken over the fitting data, so dividing it by n gives the sMSE on the fitting data, not the sMSE of predictions of new outcomes.

Definition 0.9 (Ordinary least squares)**.** The **ordinary least squares (OLS) estimate** of a model's parameter vector $\underset{\sim}{\theta}$ is the value of $\underset{\sim}{\theta}$ that minimizes the residual sum of squares:

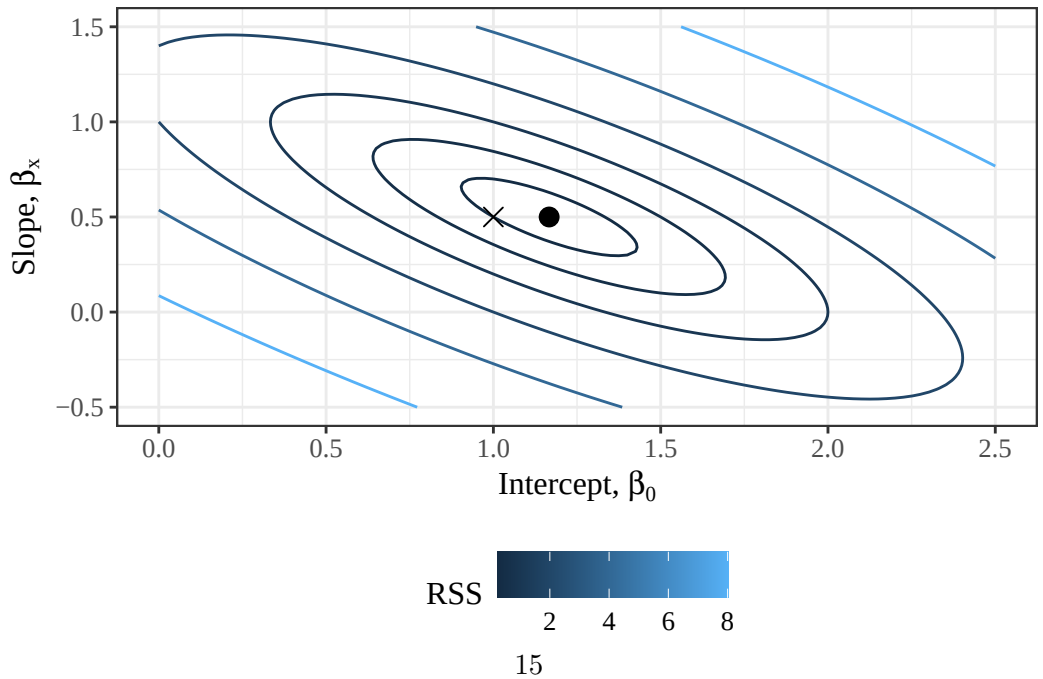
$$\hat{\underset{\sim}{\theta}} \stackrel{\text{def}}{=} \arg\min_{\underset{\sim}{\theta}} \text{RSS}(\underset{\sim}{\theta}).$$

Source: Wikipedia contributors [10].

Example 0.3 (OLS estimates of a simple linear regression)**.** For a simple linear regression, the OLS estimates (Definition 0.9) $\hat{\beta}_0$ and $\hat{\beta}_x$ are the values of β_0 and β_x that minimize $\text{RSS}(\beta_0, \beta_x)$ from Example 0.1. Figure 3 shows $\text{RSS}(\beta_0, \beta_x)$ for the three points of Example 0.2: the OLS estimates sit at the bottom of the bowl, and the line of Example 0.2 sits higher up.

```
rss_grid <- expand.grid(
  b0 = seq(0, 2.5, length.out = 101),
  bx = seq(-0.5, 1.5, length.out = 101)
)
rss_grid$rss <- mapply(
  function(b0, bx) sum((rss_points$y - b0 - bx * rss_points$x)^2),
  rss_grid$b0,
  rss_grid$bx
)
ols_coefs <- coef(lm(y ~ x, data = rss_points))

ggplot2::ggplot(rss_grid, ggplot2::aes(x = b0, y = bx, z = rss)) +
  ggplot2::geom_contour(
    ggplot2::aes(color = ggplot2::after_stat(level)),
    breaks = c(0.25, 0.5, 1, 2, 4, 8)
  ) +
  ggplot2::annotate(
    "point",
    x = ols_coefs[[1]], y = ols_coefs[[2]], size = 3
  ) +
  ggplot2::annotate("point", x = 1, y = 0.5, shape = 4, size = 3) +
  ggplot2::labs(
    x = expression("Intercept, " * beta[0]),
    y = expression("Slope, " * beta[x]),
    color = "RSS"
  )
```




```
ols_coefs
#> (Intercept)           x
#>     1.16667     0.50000
```

Figure 3: Contours of RSS(β_0, β_x) for the three points of Example 0.2. The dot marks

(This 3D surface is interactive; see the HTML or slide version of this page to view it.)

Figure 4: Surface of $\text{RSS}(\beta_0, \beta_x)$ for the three points of Example 0.2. The dot at the bottom of the bowl marks the OLS estimates; the cross marks the line of Example 0.2.

Definition 0.10 (Normal equations)**.** For a model fitted by least squares with parameter vector $\underset{\sim}{\theta}$, the **normal equations** set the gradient of the residual sum of squares to zero:

$$\frac{\partial}{\partial \underset{\sim}{\theta}} \text{RSS}(\underset{\sim}{\theta}) = \underset{\sim}{0}.$$

Example 0.4 (Normal equations of a simple linear regression)**.** For a simple linear regression, the parameter vector is $\underset{\sim}{\beta} \stackrel{\text{def}}{=} (\beta_0, \beta_x)^\top$, and $\text{RSS}(\underset{\sim}{\beta})$ is $\text{RSS}(\beta_0, \beta_x)$ from Example 0.1. Its gradient has one entry per coefficient, so the normal equations (Definition 0.10) are two scalar equations:

$$\frac{\partial \text{RSS}}{\partial \beta_0} = 0, \qquad \frac{\partial \text{RSS}}{\partial \beta_x} = 0.$$

Definition 0.11 (Centered sums of squares)**.** For data $(x_1, y_1), \ldots, (x_n, y_n)$, the **centered sums of squares** of the x_i and of the y_i are

$$S_{xx} \stackrel{\text{def}}{=} \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad S_{yy} \stackrel{\text{def}}{=} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Definition 0.12 (Centered sum of cross-products)**.** For the data of Definition 0.11, the **centered sum of cross-products** is

$$S_{xy} \stackrel{\text{def}}{=} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Exercise 0.4 (Summing the deviations from the sample mean)**.** Find $\sum_{i=1}^{n} (x_i - \bar{x})$.

Solution

Solution 0.4.

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x}) &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} && \text{(distribute the sum over the terms)} \\
&= \sum_{i=1}^{n} x_i - n\bar{x} && \text{(sum of a constant)} \\
&= n\bar{x} - n\bar{x} && (\sum_{i=1}^{n} x_i = n\bar{x}) \\
&= 0 && \text{(subtract)}
\end{aligned}$$

The same steps with y_i in place of x_i show that $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$.

Lemma 0.1 (Deviations from the sample mean sum to zero)**.** *For any data $z_1, \ldots, z_n$ with sample mean $\bar{z}$ (for example, the x_i or the y_i), the deviations from the sample mean sum to zero:*

$$\sum_{i=1}^{n}(z_i - \bar{z}) = 0.$$

Proof

Proof. This is the solution to Exercise 0.4. □

Exercise 0.5 (Expanding the centered sum of squares)**.** Expand the centered sum of squares S_{xx} (Definition 0.11) into an expression in terms of n, $\sum_{i=1}^{n} x_i^2$ and $\bar{x}$.

Solution

Solution 0.5.

$$
\begin{aligned}
S_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 && \text{(definition of } S_{xx}\text{)} \\
&= \sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) && \text{(expand the square)} \\
&= \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2\bar{x}x_i + \sum_{i=1}^{n} \bar{x}^2 && \text{(distribute the sum over the terms)} \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2 && \text{(constant factor out of the sum)} \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 && \text{(sum of a constant)} \\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\cdot n\bar{x} + n\bar{x}^2 && (\sum_{i=1}^{n} x_i = n\bar{x}) \\
&= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 && \text{(multiply)} \\
&= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 && \text{(collect the } \bar{x}^2 \text{ terms)}
\end{aligned}
$$

Lemma 0.2 (Expanded centered sum of squares)**.** *The centered sum of squares* S_{xx} *(Definition 0.11) equals*

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

Proof

Proof. This is the solution to Exercise 0.5. □

Exercise 0.6 (Expanding the centered sum of cross-products)**.** Expand the centered sum of cross-products S_{xy} (Definition 0.12) into an expression in terms of n, $\sum_{i=1}^{n} x_i y_i$, $\bar{x}$ and $\bar{y}$.

Solution

Solution 0.6.

$$
\begin{aligned}
S_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) && \text{(definition of } S_{xy}\text{)} \\
&= \sum_{i=1}^{n}(x_i y_i - \bar{y}x_i - \bar{x}y_i + \bar{x}\bar{y}) && \text{(expand the product)} \\
&= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \bar{y}x_i - \sum_{i=1}^{n} \bar{x}y_i + \sum_{i=1}^{n} \bar{x}\bar{y} && \text{(distribute the sum over the terms)} \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \bar{x}\bar{y} && \text{(constant factors out of the sum)} \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\bar{y} && \text{(sum of a constant)} \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y}\cdot n\bar{x} - \bar{x}\cdot n\bar{y} + n\bar{x}\bar{y} && (\sum_{i=1}^{n} x_i = n\bar{x},\ \sum_{i=1}^{n} y_i = n\bar{y}) \\
&= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} && \text{(multiply)} \\
&= \sum_{i=1}^{n} x_i y_i - 2n\bar{x}\bar{y} + n\bar{x}\bar{y} && \text{(add the two equal terms)} \\
&= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} && \text{(add the remaining term)}
\end{aligned}
$$

Lemma 0.3 (Expanded centered sum of cross-products)**.** *The centered sum of cross-products* S_{xy} *(Definition 0.12) equals*

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.$$

Proof

Proof. This is the solution to Exercise 0.6. □

Exercise 0.7 (Derivative of RSS with respect to the intercept)**.** Find the partial derivative of the simple linear regression residual sum of squares $\text{RSS}(\beta_0, \beta_x)$ (Example 0.1) with respect to the intercept, $\partial\text{RSS}/\partial\beta_0$.

Solution

Solution 0.7. Differentiate one operation at a time, starting from the definition (Example 0.1):

$$\begin{aligned}
\frac{\partial \text{RSS}}{\partial \beta_0} &= \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_x x_i)^2 && \text{(definition of RSS)} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_0} (y_i - \beta_0 - \beta_x x_i)^2 && \text{(derivative of a sum)} \\
&= \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_x x_i) \frac{\partial}{\partial \beta_0} (y_i - \beta_0 - \beta_x x_i) && \text{(chain rule)}
\end{aligned}$$

The inner derivative is:

$$\begin{aligned}
\frac{\partial}{\partial \beta_0} (y_i - \beta_0 - \beta_x x_i) &= \frac{\partial y_i}{\partial \beta_0} - \frac{\partial \beta_0}{\partial \beta_0} - \frac{\partial (\beta_x x_i)}{\partial \beta_0} && \text{(derivative of a sum)} \\
&= 0 - 1 - 0 && (y_i \text{ and } \beta_x x_i \text{ do not depend on } \beta_0) \\
&= -1 && \text{(subtract)}
\end{aligned}$$

Plugging the inner derivative back in gives:

$$\begin{aligned}
\frac{\partial \text{RSS}}{\partial \beta_0} &= \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_x x_i)(-1) && \text{(inner derivative is } -1) \\
&= \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_x x_i) && \text{(multiply 2 by } -1) \\
&= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_x x_i) && \text{(constant factor out of the sum)}
\end{aligned}$$

Exercise 0.8 (Derivative of RSS in terms of the means)**.** In the setting of Exercise 0.7, write the partial derivative in terms of $\bar{x}$ and $\bar{y}$.

Solution

Solution 0.8. Write the derivative from Exercise 0.7 in terms of $\bar{x}$ and $\bar{y}$:

$$
\begin{aligned}
\frac{\partial \text{RSS}}{\partial \beta_0} &= -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_x x_i) && \text{(derivative of the RSS)} \\
&= -2\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\beta_0 - \sum_{i=1}^{n}\beta_x x_i\right) && \text{(distribute the sum over the terms)} \\
&= -2\left(\sum_{i=1}^{n} y_i - n\beta_0 - \sum_{i=1}^{n}\beta_x x_i\right) && \text{(sum of a constant)} \\
&= -2\left(\sum_{i=1}^{n} y_i - n\beta_0 - \beta_x\sum_{i=1}^{n} x_i\right) && \text{(constant factor out of the sum)} \\
&= -2\left(n\bar{y} - n\beta_0 - \beta_x\sum_{i=1}^{n} x_i\right) && (\sum_{i=1}^{n} y_i = n\bar{y}) \\
&= -2(n\bar{y} - n\beta_0 - \beta_x n\bar{x}) && (\sum_{i=1}^{n} x_i = n\bar{x})
\end{aligned}
$$

Lemma 0.4 (Derivative of RSS with respect to the intercept)**.** *The partial derivative of the simple linear regression residual sum of squares RSS(β_0, β_x) (Example 0.1) with respect to the intercept is*

$$\frac{\partial RSS}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_x x_i) = -2(n\bar{y} - n\beta_0 - \beta_x n\bar{x}).$$

Proof

Proof. The derivative in sum form is the solution to Exercise 0.7. The form in terms of means is the solution to Exercise 0.8. □

Exercise 0.9 (Derivative of RSS with respect to the slope)**.** Find the partial derivative of the simple linear regression residual sum of squares RSS(β_0, β_x) (Example 0.1) with respect to the slope, $\partial \text{RSS}/\partial \beta_x$.

Solution

Solution 0.9. Differentiate one operation at a time, starting from the definition (Example 0.1):

$$\begin{aligned}
\frac{\partial \text{RSS}}{\partial \beta_x} &= \frac{\partial}{\partial \beta_x} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_x x_i)^2 && \text{(definition of RSS)} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_x} (y_i - \beta_0 - \beta_x x_i)^2 && \text{(derivative of a sum)} \\
&= \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_x x_i) \frac{\partial}{\partial \beta_x} (y_i - \beta_0 - \beta_x x_i) && \text{(chain rule)}
\end{aligned}$$

The inner derivative:

$$\begin{aligned}
\frac{\partial}{\partial \beta_x} (y_i - \beta_0 - \beta_x x_i) &= \frac{\partial y_i}{\partial \beta_x} - \frac{\partial \beta_0}{\partial \beta_x} - \frac{\partial (\beta_x x_i)}{\partial \beta_x} && \text{(derivative of a sum)} \\
&= 0 - 0 - x_i && (y_i \text{ and } \beta_0 \text{ do not depend on } \beta_x) \\
&= -x_i && \text{(subtract)}
\end{aligned}$$

Plugging the inner derivative back in:

$$\begin{aligned}
\frac{\partial \text{RSS}}{\partial \beta_x} &= \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_x x_i)(-x_i) && \text{(inner derivative is } -x_i) \\
&= \sum_{i=1}^{n} -2x_i(y_i - \beta_0 - \beta_x x_i) && \text{(rearrange the product)} \\
&= -2 \sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_x x_i) && \text{(constant factor out of the sum)}
\end{aligned}$$

Lemma 0.5 (Derivative of RSS with respect to the slope)**.** *The partial derivative of the simple linear regression residual sum of squares $RSS(\beta_0, \beta_x)$ (Example 0.1) with respect to the slope is*

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

Proof

Proof. This is the solution to Exercise 0.9. □

Exercise 0.10 (Solving the normal equations)**.** Using Exercise 0.7 and Exercise 0.9, solve the normal equations of Example 0.4 for β_0 and β_x, assuming $S_{xx} > 0$ (Definition 0.11).

Solution

Solution 0.10. **First normal equation.** Set the derivative from Exercise 0.7 to zero, then isolate β_0:

$$\begin{aligned}
0 &= -2(n\bar{y} - n\beta_0 - \beta_x n\bar{x}) && \text{(first normal equation)} \\
0 &= n\bar{y} - n\beta_0 - \beta_x n\bar{x} && \text{(divide both sides by } -2) \\
0 &= \bar{y} - \beta_0 - \beta_x \bar{x} && \text{(divide both sides by } n) \\
\beta_0 &= \bar{y} - \beta_x \bar{x} && \text{(add } \beta_0 \text{ to both sides)}
\end{aligned}$$

Second normal equation. Substitute that intercept into the derivative from Exercise 0.9:

$$\begin{aligned}
\frac{\partial \text{RSS}}{\partial \beta_x} &= -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_x x_i) && \text{(derivative with respect to } \beta_x) \\
&= -2\sum_{i=1}^{n} x_i(y_i - (\bar{y} - \beta_x \bar{x}) - \beta_x x_i) && \text{(substitute } \beta_0) \\
&= -2\sum_{i=1}^{n} x_i(y_i - \bar{y} + \beta_x \bar{x} - \beta_x x_i) && \text{(remove the inner parentheses)} \\
&= -2\sum_{i=1}^{n} x_i((y_i - \bar{y}) - (\beta_x x_i - \beta_x \bar{x})) && \text{(group terms)} \\
&= -2\sum_{i=1}^{n} x_i((y_i - \bar{y}) - \beta_x(x_i - \bar{x})) && \text{(factor out } \beta_x)
\end{aligned}$$

To shorten the next steps, write $d_i \overset{\text{def}}{=} (y_i - \bar{y}) - \beta_x(x_i - \bar{x})$, so that $\partial \text{RSS}/\partial \beta_x = -2\sum_{i=1}^{n} x_i d_i$. The d_i sum to zero, because deviations from a sample mean sum to zero (Exercise 0.4):

$$\begin{aligned}
\sum_{i=1}^{n} d_i &= \sum_{i=1}^{n}((y_i - \bar{y}) - \beta_x(x_i - \bar{x})) && \text{(definition of } d_i) \\
&= \sum_{i=1}^{n}(y_i - \bar{y}) - \sum_{i=1}^{n}\beta_x(x_i - \bar{x}) && \text{(distribute the sum over the terms)} \\
&= \sum_{i=1}^{n}(y_i - \bar{y}) - \beta_x \sum_{i=1}^{n}(x_i - \bar{x}) && \text{(constant factor out of the sum)} \\
&= 0 - \beta_x \cdot 0 && \text{(deviations sum to zero)} \\
&= 0 - 0 && \text{(multiply)} \\
&= 0 && \text{(subtract)}
\end{aligned}$$

So subtracting $\bar{x}\sum_{i=1}^{n} d_i$ from $\sum_{i=1}^{n} x_i d_i$ does not change it:

$$
\begin{aligned}
\sum_{i=1}^{n} x_i d_i &= \sum_{i=1}^{n} x_i d_i - \bar{x}\sum_{i=1}^{n} d_i && (\sum_{i=1}^{n} d_i = 0)\\
&= \sum_{i=1}^{n} x_i d_i - \sum_{i=1}^{n} \bar{x} d_i && \text{(constant factor into the sum)}\\
&= \sum_{i=1}^{n} (x_i - \bar{x}) d_i && \text{(combine the sums)}\\
&= \sum_{i=1}^{n} (x_i - \bar{x})((y_i - \bar{y}) - \beta_x(x_i - \bar{x})) && \text{(definition of } d_i)\\
&= \sum_{i=1}^{n} ((x_i - \bar{x})(y_i - \bar{y}) - \beta_x(x_i - \bar{x})^2) && \text{(distribute)}\\
&= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) - \sum_{i=1}^{n} \beta_x(x_i - \bar{x})^2 && \text{(distribute the sum over the terms)}\\
&= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) - \beta_x \sum_{i=1}^{n} (x_i - \bar{x})^2 && \text{(constant factor out of the sum)}\\
&= S_{xy} - \beta_x \sum_{i=1}^{n} (x_i - \bar{x})^2 && \text{(definition of } S_{xy})\\
&= S_{xy} - \beta_x S_{xx} && \text{(definition of } S_{xx})
\end{aligned}
$$

So $\partial \mathrm{RSS}/\partial \beta_x = -2\big(S_{xy} - \beta_x S_{xx}\big)$. Set it to zero and isolate β_x:

$$
\begin{aligned}
0 &= -2\big(S_{xy} - \beta_x S_{xx}\big) && \text{(second normal equation)}\\
0 &= S_{xy} - \beta_x S_{xx} && \text{(divide both sides by } -2)\\
\beta_x S_{xx} &= S_{xy} && \text{(add } \beta_x S_{xx} \text{ to both sides)}\\
\beta_x &= \frac{S_{xy}}{S_{xx}} && \text{(divide both sides by } S_{xx} > 0)
\end{aligned}
$$

So the normal equations have exactly one solution: $\beta_x = S_{xy}/S_{xx}$ and $\beta_0 = \bar{y} - \beta_x \bar{x}$.

Lemma 0.6 (Solution of the normal equations)**.** *If $S_{xx} > 0$ (Definition 0.11), the normal equations of Example 0.4 have exactly one solution:*

$$
\beta_x = \frac{S_{xy}}{S_{xx}}, \qquad \beta_0 = \bar{y} - \beta_x \bar{x}.
$$

Proof

Proof. This is the solution to Exercise 0.10. □

Exercise 0.11 (Second derivatives of RSS)**.** Find the matrix of second derivatives of $\text{RSS}(\beta_0, \beta_x)$ (Example 0.1) with respect to β_0 and β_x.

Solution

Solution 0.11. Differentiate each first derivative again.
Derivatives of ∂RSS$/\partial\beta_0$. By Exercise 0.7, $\partial\text{RSS}/\partial\beta_0 = -2(n\bar{y} - n\beta_0 - \beta_x n\bar{x})$. Differentiating with respect to β_0:

$$\begin{aligned}
\frac{\partial^2 \text{RSS}}{\partial \beta_0^2} &= -2\frac{\partial}{\partial \beta_0}(n\bar{y} - n\beta_0 - \beta_x n\bar{x}) && \text{(constant factor)} \\
&= -2\left(\frac{\partial(n\bar{y})}{\partial \beta_0} - \frac{\partial(n\beta_0)}{\partial \beta_0} - \frac{\partial(\beta_x n\bar{x})}{\partial \beta_0}\right) && \text{(derivative of a sum)} \\
&= -2(0 - n - 0) && \text{(differentiate each term)} \\
&= -2(-n) && \text{(drop the zeros)} \\
&= 2n && \text{(multiply)}
\end{aligned}$$

Differentiating with respect to β_x:

$$\begin{aligned}
\frac{\partial^2 \text{RSS}}{\partial \beta_x \partial \beta_0} &= -2\frac{\partial}{\partial \beta_x}(n\bar{y} - n\beta_0 - \beta_x n\bar{x}) && \text{(constant factor)} \\
&= -2\left(\frac{\partial(n\bar{y})}{\partial \beta_x} - \frac{\partial(n\beta_0)}{\partial \beta_x} - \frac{\partial(\beta_x n\bar{x})}{\partial \beta_x}\right) && \text{(derivative of a sum)} \\
&= -2(0 - 0 - n\bar{x}) && \text{(differentiate each term)} \\
&= -2(-n\bar{x}) && \text{(drop the zeros)} \\
&= 2n\bar{x} && \text{(multiply)}
\end{aligned}$$

Derivatives of ∂RSS$/\partial\beta_x$. By Exercise 0.9, $\partial\text{RSS}/\partial\beta_x = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_x x_i)$. Differentiating with respect to β_0, and reusing the inner derivative -1 from Exercise 0.7:

$$
\begin{aligned}
\frac{\partial^2 \text{RSS}}{\partial\beta_0\partial\beta_x} &= -2\frac{\partial}{\partial\beta_0}\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_x x_i) && \text{(constant factor)} \\
&= -2\sum_{i=1}^{n}\frac{\partial}{\partial\beta_0}(x_i(y_i - \beta_0 - \beta_x x_i)) && \text{(derivative of a sum)} \\
&= -2\sum_{i=1}^{n} x_i\frac{\partial}{\partial\beta_0}(y_i - \beta_0 - \beta_x x_i) && (x_i \text{ does not depend on } \beta_0) \\
&= -2\sum_{i=1}^{n} x_i(-1) && \text{(inner derivative is } -1) \\
&= -2\sum_{i=1}^{n}(-x_i) && \text{(multiply)} \\
&= -2\left(-\sum_{i=1}^{n} x_i\right) && \text{(constant factor out of the sum)} \\
&= 2\sum_{i=1}^{n} x_i && \text{(multiply)} \\
&= 2n\bar{x} && (\sum_{i=1}^{n} x_i = n\bar{x})
\end{aligned}
$$

Differentiating with respect to β_x, and reusing the inner derivative $-x_i$ from Exercise 0.9:

$$
\begin{aligned}
\frac{\partial^2 \text{RSS}}{\partial\beta_x^2} &= -2\frac{\partial}{\partial\beta_x}\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_x x_i) && \text{(constant factor)} \\
&= -2\sum_{i=1}^{n}\frac{\partial}{\partial\beta_x}(x_i(y_i - \beta_0 - \beta_x x_i)) && \text{(derivative of a sum)} \\
&= -2\sum_{i=1}^{n} x_i\frac{\partial}{\partial\beta_x}(y_i - \beta_0 - \beta_x x_i) && (x_i \text{ does not depend on } \beta_x) \\
&= -2\sum_{i=1}^{n} x_i(-x_i) && \text{(inner derivative is } -x_i) \\
&= -2\sum_{i=1}^{n}(-x_i^2) && \text{(multiply)} \\
&= -2\left(-\sum_{i=1}^{n} x_i^2\right) && \text{(constant factor out of the sum)} \\
&= 2\sum_{i=1}^{n} x_i^2 && \text{(multiply)}
\end{aligned}
$$

The matrix. Collecting the four second derivatives gives the matrix of second derivatives:

$$\begin{pmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum_{i=1}^{n} x_i^2 \end{pmatrix}.$$

Exercise 0.12 (Positive definiteness of the second-derivative matrix of RSS)**.** In the setting of Exercise 0.11, show that the matrix of second derivatives is positive definite when $S_{xx} > 0$ (Definition 0.11).

Solution

Solution 0.12. The top-left entry of the matrix of second derivatives from Exercise 0.11, $2n$, is positive. The determinant of this matrix, using the expanded form of S_{xx} from Exercise 0.5, is

$$\begin{aligned}
\det \begin{pmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum_{i=1}^{n} x_i^2 \end{pmatrix} &= (2n)\left(2\sum_{i=1}^{n} x_i^2\right) - (2n\bar{x})(2n\bar{x}) && (2 \times 2 \text{ determinant}) \\
&= 4n\sum_{i=1}^{n} x_i^2 - 4n^2\bar{x}^2 && (\text{multiply}) \\
&= 4n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) && (\text{factor out } 4n) \\
&= 4nS_{xx} && (\text{expanded form of } S_{xx})
\end{aligned}$$

which is positive when $S_{xx} > 0$. A symmetric 2×2 matrix with a positive top-left entry and a positive determinant is positive definite.

Lemma 0.7 (Second derivatives of RSS)**.** *The matrix of second derivatives of RSS(β_0, β_x) (Example 0.1) with respect to β_0 and β_x is positive definite at every (β_0, β_x) when $S_{xx} > 0$ (Definition 0.11), and equals*

$$\begin{pmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum_{i=1}^{n} x_i^2 \end{pmatrix}.$$

Proof

Proof. The matrix of second derivatives is the solution to Exercise 0.11. The positive definiteness is the solution to Exercise 0.12. □

Theorem 0.4 (Closed-form OLS estimates)**.** *Suppose $S_{xx} > 0$ (Definition 0.11), that is, not all x_i are equal. Then the OLS estimates (Example 0.3) are unique, and*

$$\hat{\beta}_x = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_x \bar{x}.$$

Proof

Proof. RSS is differentiable, so any point that minimizes it solves the normal equations. By Lemma 0.6, the normal equations have exactly one solution, $\beta_x = S_{xy}/S_{xx}$ and $\beta_0 = \bar{y} - \beta_x \bar{x}$. By Lemma 0.7, the matrix of second derivatives of RSS is positive definite at every point, so RSS is strictly convex and its only stationary point is its unique global minimum. So the OLS estimates exist, are unique, and equal that solution. □

The same estimates follow from a derivation in vector notation, which treats (β_0, β_x) as a single vector instead of differentiating with respect to each component separately.

Definition 0.13 (Covariate vector)**.** In simple linear regression (Definition 0.5), the **covariate vector** $\underset{\sim}{x}_i$ of observation i is its covariate value x_i, preceded by a 1 for the intercept:

$$\underset{\sim}{x}_i \overset{\text{def}}{=} \begin{pmatrix} 1 \\ x_i \end{pmatrix}.$$

Exercise 0.13 (The conditional mean and the covariate vector)**.** In simple linear regression (Definition 0.5), express the conditional mean $\mu(x_i)$ as a dot product of the covariate vector $\underset{\sim}{x}_i$ (Definition 0.13) and the coefficient vector $\underset{\sim}{\beta} = (\beta_0, \beta_x)^\top$ (Example 0.4).

Solution

Solution 0.13. Start from the dot product:

$$
\begin{aligned}
\underset{\sim}{x}_i \cdot \underset{\sim}{\beta} &= \begin{pmatrix} 1 \\ x_i \end{pmatrix} \cdot \underset{\sim}{\beta} && \text{(definition of } \underset{\sim}{x}_i\text{)} \\
&= \begin{pmatrix} 1 \\ x_i \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_x \end{pmatrix} && \text{(definition of } \underset{\sim}{\beta}\text{, written as a column)} \\
&= 1 \cdot \beta_0 + x_i \beta_x && \text{(definition of the dot product)} \\
&= \beta_0 + x_i \beta_x && (1 \cdot a = a) \\
&= \beta_0 + \beta_x x_i && \text{(reorder the factors)} \\
&= \mu(x_i) && \text{(definition of } \mu(x)\text{)}
\end{aligned}
$$

So $\mu(x_i) = \underset{\sim}{x}_i \cdot \underset{\sim}{\beta}$.

Lemma 0.8 (The conditional mean is a dot product)**.** *In simple linear regression (Definition 0.5), the conditional mean of observation i is the dot product of its covariate vector (Definition 0.13) and the coefficient vector $\underset{\sim}{\beta}$ (Example 0.4):*

$$\mu(x_i) = \underset{\sim}{x}_i \cdot \underset{\sim}{\beta}.$$

Proof

Proof. This is the solution to Exercise 0.13. □

Example 0.5 (Residual sum of squares in vector notation)**.** In simple linear regression, take $\underset{\sim}{\theta} = \underset{\sim}{\beta}$ and substitute $\mu(x_i) = \underset{\sim}{x}_i \cdot \underset{\sim}{\beta}$ (Lemma 0.8) into each residual of the residual sum of squares (Definition 0.8). The residual sum of squares is then a function of the coefficient vector $\underset{\sim}{\beta}$, the same function as $\text{RSS}(\beta_0, \beta_x)$ in Example 0.1:

$$\text{RSS}(\underset{\sim}{\beta}) = \sum_{i=1}^{n} (y_i - \underset{\sim}{x}_i \cdot \underset{\sim}{\beta})^2.$$

Exercise 0.14 (Gradient of RSS in vector notation)**.** Find the gradient of the residual sum of squares in vector notation (Example 0.5), $\frac{\partial}{\partial \underset{\sim}{\beta}} \text{RSS}(\underset{\sim}{\beta})$, by differentiating with respect to the vector $\underset{\sim}{\beta}$ directly, rather than one component at a time.

Solution

Solution 0.14. Differentiate one operation at a time, starting from $\text{RSS}(\underset{\sim}{\beta})$ in vector notation (Example 0.5):

$$
\begin{aligned}
\frac{\partial}{\partial \underset{\sim}{\beta}} \mathrm{RSS}(\underset{\sim}{\beta}) &= \frac{\partial}{\partial \underset{\sim}{\beta}} \sum_{i=1}^{n} (y_i - \underset{\sim}{x}_i \cdot \underset{\sim}{\beta})^2 && \text{(RSS in vector form)} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \underset{\sim}{\beta}} (y_i - \underset{\sim}{x}_i \cdot \underset{\sim}{\beta})^2 && \text{(derivative of a sum)} \\
&= \sum_{i=1}^{n} 2(y_i - \underset{\sim}{x}_i \cdot \underset{\sim}{\beta}) \frac{\partial}{\partial \underset{\sim}{\beta}} (y_i - \underset{\sim}{x}_i \cdot \underset{\sim}{\beta}) && \text{(chain rule)}
\end{aligned}
$$

For the inner derivative, we need the gradient of a linear function. For any constant vector $\underset{\sim}{a} = (a_0, a_x)^\top$:

$$
\begin{aligned}
\frac{\partial}{\partial \underset{\sim}{\beta}} \underset{\sim}{a} \cdot \underset{\sim}{\beta} &= \frac{\partial}{\partial \beta} (a_0 \beta_0 + a_x \beta_x) && \text{(write out the dot product)} \\
&= \begin{pmatrix} \frac{\partial}{\partial \beta_0} (a_0 \beta_0 + a_x \beta_x) \\ \frac{\partial}{\partial \beta_x} (a_0 \beta_0 + a_x \beta_x) \end{pmatrix} && \text{(definition of the gradient)} \\
&= \begin{pmatrix} a_0 + 0 \\ 0 + a_x \end{pmatrix} && \text{(derivative of each term)} \\
&= \begin{pmatrix} a_0 \\ a_x \end{pmatrix} && \text{(add zero)} \\
&= \underset{\sim}{a} && \text{(definition of } \underset{\sim}{a}\text{)}
\end{aligned}
$$

With $\underset{\sim}{a} = \underset{\sim}{x}_i$:

$$
\begin{aligned}
\frac{\partial}{\partial \underset{\sim}{\beta}} (y_i - \underset{\sim}{x}_i \cdot \underset{\sim}{\beta}) &= \frac{\partial}{\partial \underset{\sim}{\beta}} y_i - \frac{\partial}{\partial \underset{\sim}{\beta}} \underset{\sim}{x}_i \cdot \underset{\sim}{\beta} && \text{(derivative of a difference)} \\
&= \underset{\sim}{0} - \frac{\partial}{\partial \underset{\sim}{\beta}} \underset{\sim}{x}_i \cdot \underset{\sim}{\beta} && (y_i \text{ does not depend on } \underset{\sim}{\beta}) \\
&= \underset{\sim}{0} - \underset{\sim}{x}_i && \text{(gradient of a linear function)} \\
&= -\underset{\sim}{x}_i && \text{(subtract from zero)}
\end{aligned}
$$

Plugging the inner derivative back in:

$$
\begin{aligned}
\frac{\partial}{\partial \underset{\sim}{\beta}} \mathrm{RSS}(\underset{\sim}{\beta}) &= \sum_{i=1}^{n} 2(y_i - \underset{\sim}{x}_i \cdot \underset{\sim}{\beta})(-\underset{\sim}{x}_i) && \text{(inner derivative is } -\underset{\sim}{x}_i\text{)} \\
&= \sum_{i=1}^{n} 2(-\underset{\sim}{x}_i)(y_i - \underset{\sim}{x}_i \cdot \underset{\sim}{\beta}) && \text{(reorder the scalar and vector factors)} \\
&= \sum_{i=1}^{n} (-2)\underset{\sim}{x}_i (y_i - \underset{\sim}{x}_i \cdot \underset{\sim}{\beta}) && (2(-\underset{\sim}{x}_i) = -2\underset{\sim}{x}_i) \\
&= -2 \sum_{i=1}^{n} \underset{\sim}{x}_i (y_i - \underset{\sim}{x}_i \cdot \underset{\sim}{\beta}) && \text{(constant factor out of the sum)}
\end{aligned}
$$

Lemma 0.9 (Gradient of RSS in vector notation)**.** *The gradient of the residual sum of squares in vector notation (Example 0.5) is*

$$\frac{\partial}{\partial \underset{\sim}{\beta}} RSS(\underset{\sim}{\beta}) = -2\sum_{i=1}^{n} \underset{\sim}{x}_i(y_i - \underset{\sim}{x}_i \cdot \underset{\sim}{\beta}).$$

Proof

Proof. This is the solution to Exercise 0.14. □

Exercise 0.15 (Normal equations in vector notation)**.** Define the matrix and vector

$$A \overset{\text{def}}{=} \sum_{i=1}^{n} \underset{\sim}{x}_i \underset{\sim}{x}_i^{\top}, \qquad \underset{\sim}{c} \overset{\text{def}}{=} \sum_{i=1}^{n} \underset{\sim}{x}_i y_i.$$

Using Exercise 0.14, show that the normal equations can be written as $A\underset{\sim}{\beta} = \underset{\sim}{c}$.

Solution

Solution 0.15.
Distribute the sum in the gradient from Exercise 0.14:

$$\begin{aligned}
\nabla_{\underset{\sim}{\beta}} \mathrm{RSS}(\underset{\sim}{\beta}) &= -2\sum_{i=1}^{n} \underset{\sim}{x}_i(y_i - \underset{\sim}{x}_i \cdot \underset{\sim}{\beta}) && \text{(gradient of RSS)} \\
&= -2\sum_{i=1}^{n} (\underset{\sim}{x}_i y_i - \underset{\sim}{x}_i(\underset{\sim}{x}_i \cdot \underset{\sim}{\beta})) && \text{(distribute } \underset{\sim}{x}_i\text{)} \\
&= -2\left(\sum_{i=1}^{n} \underset{\sim}{x}_i y_i - \sum_{i=1}^{n} \underset{\sim}{x}_i(\underset{\sim}{x}_i \cdot \underset{\sim}{\beta})\right) && \text{(sum of differences)}
\end{aligned}$$

Each term of the second sum regroups as a matrix times $\underset{\sim}{\beta}$:

$$\begin{aligned}
\underset{\sim}{x}_i(\underset{\sim}{x}_i \cdot \underset{\sim}{\beta}) &= \underset{\sim}{x}_i(\underset{\sim}{x}_i^{\top}\underset{\sim}{\beta}) && \text{(dot product as a transpose product)} \\
&= (\underset{\sim}{x}_i \underset{\sim}{x}_i^{\top})\underset{\sim}{\beta} && \text{(matrix multiplication is associative)}
\end{aligned}$$

Substituting the regrouped sums gives

$$
\begin{aligned}
\nabla_{\underset{\sim}{\beta}} \mathrm{RSS}(\underset{\sim}{\beta}) &= -2\left(\underset{\sim}{c} - \sum_{i=1}^{n} \underset{\sim}{x}_i(\underset{\sim}{x}_i \cdot \underset{\sim}{\beta})\right) && \text{(definition of } \underset{\sim}{c}\text{)} \\
&= -2\left(\underset{\sim}{c} - \sum_{i=1}^{n} (\underset{\sim}{x}_i \underset{\sim}{x}_i^\top)\underset{\sim}{\beta}\right) && \text{(regroup each term)} \\
&= -2\left(\underset{\sim}{c} - \left(\sum_{i=1}^{n} \underset{\sim}{x}_i \underset{\sim}{x}_i^\top\right)\underset{\sim}{\beta}\right) && \text{(factor } \underset{\sim}{\beta} \text{ out of the sum)} \\
&= -2(\underset{\sim}{c} - A\underset{\sim}{\beta}) && \text{(definition of } A\text{)}
\end{aligned}
$$

The normal equations set this gradient equal to $\underset{\sim}{0}$. Solving for $A\underset{\sim}{\beta}$ one operation at a time:

$$
\begin{aligned}
-2(\underset{\sim}{c} - A\underset{\sim}{\beta}) &= \underset{\sim}{0} && \text{(normal equations)} \\
\underset{\sim}{c} - A\underset{\sim}{\beta} &= \underset{\sim}{0} && \text{(divide both sides by } -2\text{)} \\
\underset{\sim}{c} &= A\underset{\sim}{\beta} && \text{(add } A\underset{\sim}{\beta} \text{ to both sides)} \\
A\underset{\sim}{\beta} &= \underset{\sim}{c} && \text{(swap the sides)}
\end{aligned}
$$

Lemma 0.10 (Normal equations in vector notation)**.** *Let*

$$
A \overset{def}{=} \sum_{i=1}^{n} \underset{\sim}{x}_i \underset{\sim}{x}_i^\top, \qquad \underset{\sim}{c} \overset{def}{=} \sum_{i=1}^{n} \underset{\sim}{x}_i y_i.
$$

The normal equations of a simple linear regression are equivalent to

$$
A\underset{\sim}{\beta} = \underset{\sim}{c}.
$$

Proof

Proof. This is the solution to Exercise 0.15. □

Exercise 0.16 (Solving the normal equations in vector notation)**.** Assuming $S_{xx} > 0$ (Definition 0.11), show that A from Exercise 0.15 is invertible, solve $A\underset{\sim}{\beta} = \underset{\sim}{c}$ for $\underset{\sim}{\beta}$, and compare the result with Exercise 0.10.

Solution

Solution 0.16.
Write out the entries of each term of A and $\underset{\sim}{c}$, then add them up:

$$
\begin{aligned}
A &= \sum_{i=1}^{n} \underset{\sim}{x}_i \underset{\sim}{x}_i^{\top} && \text{(definition of } A\text{)} \\
&= \sum_{i=1}^{n} \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix} && \text{(outer product of } \underset{\sim}{x}_i\text{)} \\
&= \begin{pmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} && \text{(matrices add entrywise)} \\
&= \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} && \text{(sum of a constant)} \\
&= \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix} && (\sum_{i=1}^{n} x_i = n\bar{x})
\end{aligned}
$$

$$
\begin{aligned}
\underset{\sim}{c} &= \sum_{i=1}^{n} \underset{\sim}{x}_i y_i && \text{(definition of } \underset{\sim}{c}\text{)} \\
&= \sum_{i=1}^{n} \begin{pmatrix} y_i \\ x_i y_i \end{pmatrix} && \text{(scalar times a vector)} \\
&= \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix} && \text{(vectors add entrywise)} \\
&= \begin{pmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{pmatrix} && (\sum_{i=1}^{n} y_i = n\bar{y})
\end{aligned}
$$

The determinant of A is

$$
\begin{aligned}
\det A &= n \sum_{i=1}^{n} x_i^2 - (n\bar{x})(n\bar{x}) && \text{(determinant of a } 2 \times 2 \text{ matrix)} \\
&= n \sum_{i=1}^{n} x_i^2 - n^2 \bar{x}^2 && \text{(multiply)} \\
&= n \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) && \text{(factor out } n\text{)} \\
&= n S_{xx} && \text{(expansion of } S_{xx}\text{)}
\end{aligned}
$$

where the last step uses Exercise 0.5. Since $n > 0$ and $S_{xx} > 0$, $\det A > 0$, so A is invertible, and $A\beta = \underset{\sim}{c}$ has exactly one solution, $A^{-1}\underset{\sim}{c}$.

To simplify $A^{-1}\underset{\sim}{c}$, we need two expansions of uncentered sums. Adding $n\bar{x}^2$ to both sides of Exercise 0.5 gives $\sum_{i=1}^n x_i^2 = S_{xx} + n\bar{x}^2$. The cross-product sum expands the same way:

$$
\begin{aligned}
S_{xy} &= \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) && \text{(definition of } S_{xy}\text{)}\\
&= \sum_{i=1}^{n}(x_iy_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}) && \text{(expand the product)}\\
&= \sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}x_i\bar{y} - \sum_{i=1}^{n}\bar{x}y_i + \sum_{i=1}^{n}\bar{x}\bar{y} && \text{(sum of a sum)}\\
&= \sum_{i=1}^{n}x_iy_i - \bar{y}\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\bar{x}y_i + \sum_{i=1}^{n}\bar{x}\bar{y} && \text{(constant factor out of the sum)}\\
&= \sum_{i=1}^{n}x_iy_i - \bar{y}\sum_{i=1}^{n}x_i - \bar{x}\sum_{i=1}^{n}y_i + \sum_{i=1}^{n}\bar{x}\bar{y} && \text{(constant factor out of the sum)}\\
&= \sum_{i=1}^{n}x_iy_i - \bar{y}\sum_{i=1}^{n}x_i - \bar{x}\sum_{i=1}^{n}y_i + n\bar{x}\bar{y} && \text{(sum of a constant)}\\
&= \sum_{i=1}^{n}x_iy_i - \bar{y}(n\bar{x}) - \bar{x}\sum_{i=1}^{n}y_i + n\bar{x}\bar{y} && (\sum_{i=1}^{n}x_i = n\bar{x})\\
&= \sum_{i=1}^{n}x_iy_i - \bar{y}(n\bar{x}) - \bar{x}(n\bar{y}) + n\bar{x}\bar{y} && (\sum_{i=1}^{n}y_i = n\bar{y})\\
&= \sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} && \text{(reorder the factors)}\\
&= \sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y} && \text{(combine like terms)}
\end{aligned}
$$

Adding $n\bar{x}\bar{y}$ to both sides gives $\sum_{i=1}^{n}x_iy_i = S_{xy} + n\bar{x}\bar{y}$. Now compute the solution:

$$
\begin{aligned}
A^{-1}\underset{\sim}{c} &= \frac{1}{\det A}\begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}\underset{\sim}{c} && \text{(inverse of a } 2\times 2 \text{ matrix)} \\
&= \frac{1}{nS_{xx}}\begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}\underset{\sim}{c} && (\det A = nS_{xx}) \\
&= \frac{1}{nS_{xx}}\begin{pmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}\begin{pmatrix} n\bar{y} \\ \sum_{i=1}^n x_i y_i \end{pmatrix} && (\text{substitute } \underset{\sim}{c}) \\
&= \frac{1}{nS_{xx}}\begin{pmatrix} \sum_{i=1}^n x_i^2 \cdot n\bar{y} + (-n\bar{x})\sum_{i=1}^n x_i y_i \\ (-n\bar{x})\cdot n\bar{y} + n\sum_{i=1}^n x_i y_i \end{pmatrix} && \text{(matrix times a vector)} \\
&= \frac{1}{nS_{xx}}\begin{pmatrix} n\bar{y}\sum_{i=1}^n x_i^2 - n\bar{x}\sum_{i=1}^n x_i y_i \\ -n^2\bar{x}\bar{y} + n\sum_{i=1}^n x_i y_i \end{pmatrix} && \text{(multiply out the products)} \\
&= \frac{1}{nS_{xx}}\, n\begin{pmatrix} \bar{y}\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i y_i \\ -n\bar{x}\bar{y} + \sum_{i=1}^n x_i y_i \end{pmatrix} && \text{(factor } n \text{ out of the vector)} \\
&= \frac{1}{S_{xx}}\begin{pmatrix} \bar{y}\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i y_i \\ -n\bar{x}\bar{y} + \sum_{i=1}^n x_i y_i \end{pmatrix} && \text{(cancel } n\text{)} \\
&= \frac{1}{S_{xx}}\begin{pmatrix} \bar{y}\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \end{pmatrix} && \text{(reorder the terms of the second entry)} \\
&= \frac{1}{S_{xx}}\begin{pmatrix} \bar{y}(S_{xx} + n\bar{x}^2) - \bar{x}\sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \end{pmatrix} && (\sum_{i=1}^n x_i^2 = S_{xx} + n\bar{x}^2) \\
&= \frac{1}{S_{xx}}\begin{pmatrix} \bar{y}(S_{xx} + n\bar{x}^2) - \bar{x}\big(S_{xy} + n\bar{x}\bar{y}\big) \\ \big(S_{xy} + n\bar{x}\bar{y}\big) - n\bar{x}\bar{y} \end{pmatrix} && (\sum_{i=1}^n x_i y_i = S_{xy} + n\bar{x}\bar{y}) \\
&= \frac{1}{S_{xx}}\begin{pmatrix} \bar{y}S_{xx} + \bar{y}n\bar{x}^2 - \bar{x}S_{xy} - \bar{x}n\bar{x}\bar{y} \\ S_{xy} + n\bar{x}\bar{y} - n\bar{x}\bar{y} \end{pmatrix} && \text{(distribute)} \\
&= \frac{1}{S_{xx}}\begin{pmatrix} \bar{y}S_{xx} + n\bar{x}^2\bar{y} - \bar{x}S_{xy} - n\bar{x}\bar{x}\bar{y} \\ S_{xy} + n\bar{x}\bar{y} - n\bar{x}\bar{y} \end{pmatrix} && \text{(reorder the factors)} \\
&= \frac{1}{S_{xx}}\begin{pmatrix} \bar{y}S_{xx} + n\bar{x}^2\bar{y} - \bar{x}S_{xy} - n\bar{x}^2\bar{y} \\ S_{xy} + n\bar{x}\bar{y} - n\bar{x}\bar{y} \end{pmatrix} && (\bar{x}\bar{x} = \bar{x}^2) \\
&= \frac{1}{S_{xx}}\begin{pmatrix} \bar{y}S_{xx} - \bar{x}S_{xy} \\ S_{xy} + n\bar{x}\bar{y} - n\bar{x}\bar{y} \end{pmatrix} && \text{(first entry: cancel } \pm n\bar{x}^2\bar{y}\text{)} \\
&= \frac{1}{S_{xx}}\begin{pmatrix} \bar{y}S_{xx} - \bar{x}S_{xy} \\ S_{xy} \end{pmatrix} && \text{(second entry: cancel } \pm n\bar{x}\bar{y}\text{)} \\
&= \begin{pmatrix} \bar{y}S_{xx}/S_{xx} - \bar{x}\, S_{xy}/S_{xx} \\ S_{xy}/S_{xx} \end{pmatrix} && \text{(multiply each entry by } 1/S_{xx}\text{)} \\
&= \begin{pmatrix} \bar{y} - \bar{x}\, S_{xy}/S_{xx} \\ S_{xy}/S_{xx} \end{pmatrix} && \text{(cancel } S_{xx}\text{)}
\end{aligned}
$$

Its entries are $\beta_0 = \bar{y} - \beta_x \bar{x}$ and $\beta_x = S_{xy}/S_{xx}$, the same solution as Exercise 0.10.

Lemma 0.11 (Solution of the normal equations in vector notation)**.** *If $S_{xx} > 0$ (Definition 0.11), the matrix $A = \sum_{i=1}^{n} \underset{\sim}{x}_i \underset{\sim}{x}_i^\top$ is invertible, and the normal equations $A\underset{\sim}{\beta} = \underset{\sim}{c}$ with $\underset{\sim}{c} = \sum_{i=1}^{n} \underset{\sim}{x}_i y_i$ (Lemma 0.10) have exactly one solution,*

$$\underset{\sim}{\beta} = A^{-1}\underset{\sim}{c} = \begin{pmatrix} \bar{y} - \beta_x \bar{x} \\ S_{xy}/S_{xx} \end{pmatrix} \quad \text{with} \quad \beta_x = \frac{S_{xy}}{S_{xx}},$$

the same solution as Lemma 0.6.

Proof

Proof. This is the solution to Exercise 0.16. □

Exercise 0.17 (Hessian of RSS in vector notation)**.** Find the Hessian $\frac{\partial}{\partial \underset{\sim}{\beta}} \frac{\partial}{\partial \underset{\sim}{\beta}^\top} \mathrm{RSS}(\underset{\sim}{\beta})$.

Solution

Solution 0.17. Differentiate the gradient $-2(\underset{\sim}{c} - A\underset{\sim}{\beta})$ from Exercise 0.15 with respect to $\underset{\sim}{\beta}^\top$. $\underset{\sim}{c}$ does not depend on $\underset{\sim}{\beta}$. Each entry of $A\underset{\sim}{\beta}$ is the dot product of a row of A with $\underset{\sim}{\beta}$, so by the gradient of a linear function (Exercise 0.14), the derivative of $A\underset{\sim}{\beta}$ with respect to $\underset{\sim}{\beta}^\top$ is A. The Hessian is then:

$$\begin{aligned}
\frac{\partial}{\partial \underset{\sim}{\beta}} \frac{\partial}{\partial \underset{\sim}{\beta}^\top} \mathrm{RSS}(\underset{\sim}{\beta}) &= \frac{\partial}{\partial \underset{\sim}{\beta}^\top}(-2(\underset{\sim}{c} - A\underset{\sim}{\beta})) && \text{(differentiate the gradient)} \\
&= -2\frac{\partial}{\partial \underset{\sim}{\beta}^\top}(\underset{\sim}{c} - A\underset{\sim}{\beta}) && \text{(constant factor)} \\
&= -2\left(\frac{\partial}{\partial \underset{\sim}{\beta}^\top}\underset{\sim}{c} - \frac{\partial}{\partial \underset{\sim}{\beta}^\top} A\underset{\sim}{\beta}\right) && \text{(derivative of a difference)} \\
&= -2\left(\mathbf{0} - \frac{\partial}{\partial \underset{\sim}{\beta}^\top} A\underset{\sim}{\beta}\right) && (\underset{\sim}{c} \text{ does not depend on } \underset{\sim}{\beta}) \\
&= -2(\mathbf{0} - A) && \text{(derivative of a linear function)} \\
&= -2(-A) && \text{(subtract from zero)} \\
&= 2A && \text{(multiply)} \\
&= 2\sum_{i=1}^{n} \underset{\sim}{x}_i \underset{\sim}{x}_i^\top && \text{(definition of } A\text{)}
\end{aligned}$$

Exercise 0.18 (Positive definiteness of the Hessian of RSS)**.** In the setting of Exercise 0.17, show that the Hessian is positive definite when $S_{xx} > 0$ (Definition 0.11).

Solution

Solution 0.18. By Exercise 0.17, the Hessian is $2A$ with $A = \sum_{i=1}^{n} \underset{\sim}{x}_i \underset{\sim}{x}_i^\top$. To show that this Hessian is positive definite, take any vector $\underset{\sim}{v} = (v_0, v_x)^\top \neq \underset{\sim}{0}$:

$$
\begin{aligned}
\underset{\sim}{v}^\top (2A) \underset{\sim}{v} &= 2\, \underset{\sim}{v}^\top A \underset{\sim}{v} && \text{(constant factor)} \\
&= 2\, \underset{\sim}{v}^\top \left(\sum_{i=1}^{n} \underset{\sim}{x}_i \underset{\sim}{x}_i^\top \right) \underset{\sim}{v} && \text{(definition of } A\text{)} \\
&= 2 \sum_{i=1}^{n} \underset{\sim}{v}^\top (\underset{\sim}{x}_i \underset{\sim}{x}_i^\top) \underset{\sim}{v} && \text{(matrix multiplication distributes over the sum)} \\
&= 2 \sum_{i=1}^{n} (\underset{\sim}{v}^\top \underset{\sim}{x}_i)(\underset{\sim}{x}_i^\top \underset{\sim}{v}) && \text{(matrix multiplication is associative)} \\
&= 2 \sum_{i=1}^{n} (\underset{\sim}{v} \cdot \underset{\sim}{x}_i)(\underset{\sim}{x}_i \cdot \underset{\sim}{v}) && \text{(transpose products as dot products)} \\
&= 2 \sum_{i=1}^{n} (\underset{\sim}{v} \cdot \underset{\sim}{x}_i)^2 && \text{(dot product is symmetric)} \\
&= 2 \sum_{i=1}^{n} (v_0 + v_x x_i)^2 && \text{(write out the dot product)} \\
&\geq 0 && \text{(sum of squares)}
\end{aligned}
$$

Equality would need the condition $v_0 + v_x x_i = 0$ for every i. If $v_x \neq 0$, the condition would force every $x_i = -v_0/v_x$, contradicting $S_{xx} > 0$; if $v_x = 0$, the condition would force $v_0 = 0$, contradicting $\underset{\sim}{v} \neq \underset{\sim}{0}$. So the inequality is strict, and the Hessian is positive definite at every $\underset{\sim}{\beta}$.

Lemma 0.12 (Hessian of RSS in vector notation)**.** *The Hessian of the residual sum of squares is*

$$
\frac{\partial}{\partial \underset{\sim}{\beta}} \frac{\partial}{\partial \underset{\sim}{\beta}^\top} RSS(\underset{\sim}{\beta}) = 2 \sum_{i=1}^{n} \underset{\sim}{x}_i \underset{\sim}{x}_i^\top,
$$

and it is positive definite at every $\underset{\sim}{\beta}$ *when* $S_{xx} > 0$ *(Definition 0.11).*

Proof

Proof. The Hessian is the solution to Exercise 0.17. The positive definiteness is the solution to Exercise 0.18. □

Theorem 0.5 (OLS estimates in vector notation)**.** *Suppose $S_{xx} > 0$ (Definition 0.11). Then the matrix $A = \sum_{i=1}^{n} \underset{\sim}{x}_i \underset{\sim}{x}_i^{\top}$ is invertible, and the vector of OLS estimates (Definition 0.9) is*

$$\underset{\sim}{\hat{\beta}} = A^{-1} \underset{\sim}{c}, \qquad \underset{\sim}{c} = \sum_{i=1}^{n} \underset{\sim}{x}_i y_i.$$

Its entries are the estimates of Theorem 0.4.

Proof

Proof. RSS is differentiable, so any point that minimizes it solves the normal equations. By Lemma 0.10, the normal equations are $A\beta = \underset{\sim}{c}$. By Lemma 0.11, A is invertible, so their only solution is $A^{-1}\underset{\sim}{c}$, whose entries are those of Theorem 0.4. By Lemma 0.12, the Hessian of RSS is positive definite at every $\underset{\sim}{\beta}$, so RSS is strictly convex and $A^{-1}\underset{\sim}{c}$ is its unique global minimum. □

Exercise 0.19 (OLS slope and the correlation)**.** Assuming $S_{xx} > 0$ and $S_{yy} > 0$ (Definition 0.11), write the OLS slope $\hat{\beta}_x$ from Theorem 0.4 in terms of the Pearson correlation coefficient r and the sample standard deviations s_x and s_y of the x_i and the y_i.

Solution

Solution 0.19. In terms of the centered sums (Definition 0.11, Definition 0.12),

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}, \qquad s_x = \sqrt{\frac{S_{xx}}{n-1}}, \qquad s_y = \sqrt{\frac{S_{yy}}{n-1}}.$$

So:

$$
\begin{aligned}
r\,\frac{s_y}{s_x} &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \cdot \frac{\sqrt{S_{yy}/(n-1)}}{\sqrt{S_{xx}/(n-1)}} && \text{(substitute } r,\ s_x,\ s_y\text{)} \\
&= \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \cdot \frac{\sqrt{S_{yy}/(n-1)}}{\sqrt{S_{xx}/(n-1)}} && \text{(square root of a product)} \\
&= \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \cdot \frac{\sqrt{S_{yy}}/\sqrt{n-1}}{\sqrt{S_{xx}}/\sqrt{n-1}} && \text{(square root of a quotient)} \\
&= \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \cdot \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} && \text{(cancel the factors of } \sqrt{n-1}\text{)} \\
&= \frac{S_{xy}\sqrt{S_{yy}}}{\sqrt{S_{xx}}\sqrt{S_{yy}}\sqrt{S_{xx}}} && \text{(multiply the fractions)} \\
&= \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{xx}}} && \text{(cancel } \sqrt{S_{yy}}\text{, which is positive)} \\
&= \frac{S_{xy}}{S_{xx}} && (\sqrt{S_{xx}}\sqrt{S_{xx}} = S_{xx}) \\
&= \hat{\beta}_x && \text{(closed-form OLS slope)}
\end{aligned}
$$

Corollary 0.2 (OLS slope in terms of the correlation)**.** *If $S_{xx} > 0$ and $S_{yy} > 0$ (Definition 0.11), then*

$$\hat{\beta}_x = r\,\frac{s_y}{s_x},$$

where r is the Pearson correlation coefficient and s_x and s_y are the sample standard deviations of the x_i and the y_i.

Proof

Proof. This is the solution to Exercise 0.19. □

Video lecture

Hutchinson's Linear Regression (pt1) (17 min) covers fitting a linear regression by least squares, framed as machine learning [11]. The login for the video site is posted on Canvas.

Read

- Hogg et al. [3, Sec. 6.5], pp. 277–284 covers a simple regression problem.
- Casella and Berger [12, Sec. 11.3], pp. 539–562 covers simple linear regression.
- Rice [4, Sec. 14.2.1], pp. 547–549 covers slope and intercept properties.

Fitting a simple linear regression in R

Example 0.6 (Regression of fasting glucose on BMI in HERS)**.** The OLS estimates from Theorem 0.4, and the slope from Corollary 0.2, for the participants with a BMI measurement:

```
hers_bmi <- hers |> dplyr::filter(!is.na(BMI))
x <- hers_bmi$BMI
y <- hers_bmi$glucose
b1 <- sum((x - mean(x)) * (y - mean(y))) / sum((x - mean(x))^2)
c(
  intercept = mean(y) - b1 * mean(x),
  slope = b1,
  slope_from_r = cor(x, y) * sd(y) / sd(x)
)
#>    intercept        slope slope_from_r
#>     60.07405      1.82218      1.82218
```

`lm()` gives the same estimates:

```
slr_fit <- lm(glucose ~ BMI, data = hers)
parameters::parameters(slr_fit) |>
  parameters::print_md(digits = 3)
```

Parameter	Coefficient	SE	95% CI	t(2756)	p
(Intercept)	60.074	3.565	(53.084, 67.064)	16.852	$< .001$
BMI	1.822	0.122	(1.582, 2.062)	14.878	$< .001$

```r
performance::model_performance(slr_fit) |>
  insight::print_md(digits = 3)
```

Table 1: Indices of model performance

AIC	AICc	BIC	R2	R2 (adj.)	RMSE	Sigma
27517.9	27517.9	27535.7	0.074	0.074	35.471	35.484

Figure 5 shows the data with the fitted line.

```
ggplot2::ggplot(hers_bmi, ggplot2::aes(x = BMI, y = glucose)) +
  ggplot2::geom_point(alpha = 0.2) +
  ggplot2::geom_abline(
    intercept = coef(slr_fit)[[1]],
    slope = coef(slr_fit)[["BMI"]],
    color = "firebrick"
  ) +
  ggplot2::labs(x = "BMI (kg/m^2)", y = "Fasting glucose (mg/dL)")
```

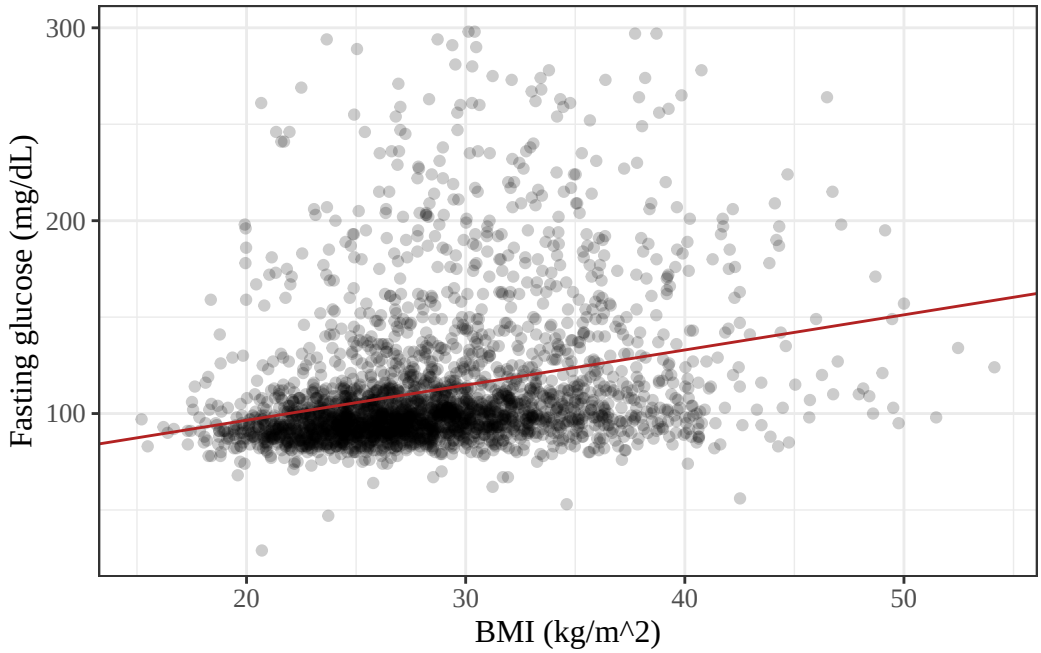


Figure 5: Fasting glucose against BMI in HERS, with the OLS line

The estimated slope is $\hat{\beta}_{\text{BMI}} = 1.82$ mg/dL per kg/m^2: mean fasting glucose is about 1.8 mg/dL higher among participants whose BMI is 1 kg/m^2 higher. The t statistic for the slope equals the correlation test statistic of Example 0.1.

The coefficient of determination

Definition 0.14 (Total sum of squares)**.** The **total sum of squares** of $y_1, \dots, y_n$ is

$$\mathrm{TSS} \stackrel{\text{def}}{=} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Source: Wikipedia contributors [13].

Example 0.7 (Total sum of squares of three values)**.** For $y = 1, 2, 2$, $\bar{y} = 5/3$, so

$$\begin{aligned} \mathrm{TSS} &= \left(1 - \tfrac{5}{3}\right)^2 + 2\left(2 - \tfrac{5}{3}\right)^2 && \text{(definition)} \\ &= \tfrac{4}{9} + \tfrac{2}{9} && \text{(square the deviations)} \\ &= \tfrac{2}{3} && \text{(arithmetic)} \end{aligned}$$

Definition 0.15 (Coefficient of determination)**.** The **coefficient of determination** of a model fitted with parameter estimate $\underset{\sim}{\hat{\theta}}$ is one minus the ratio of its residual sum of squares at $\underset{\sim}{\hat{\theta}}$ to the total sum of squares of its outcomes:

$$R^2 \stackrel{\text{def}}{=} 1 - \frac{\mathrm{RSS}(\underset{\sim}{\hat{\theta}})}{\mathrm{TSS}}.$$

Source: Wikipedia contributors [14].

Remark (Interpreting the coefficient of determination)

Remark 0.3 (Interpreting the coefficient of determination). R^2 is often described as the proportion of the variation in Y explained by the regression on X.

Exercise 0.20 (Residual sum of squares of the OLS fit)**.** Assuming $S_{xx} > 0$ (Definition 0.11), write the residual sum of squares at the OLS estimates (Theorem 0.4), $\mathrm{RSS}(\hat{\beta}_0, \hat{\beta}_x) = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$, in terms of S_{xx}, S_{yy} and S_{xy}.

Solution

Solution 0.20. The fitted values of the OLS fit are $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_x x_i$, so each residual is

$$
\begin{aligned}
\hat{\varepsilon}_i &= y_i - \hat{y}_i && \text{(definition of a residual)} \\
&= y_i - \left(\hat{\beta}_0 + \hat{\beta}_x x_i\right) && \text{(substitute } \hat{y}_i\text{)} \\
&= y_i - \left(\bar{y} - \hat{\beta}_x \bar{x} + \hat{\beta}_x x_i\right) && \text{(substitute } \hat{\beta}_0\text{)} \\
&= y_i - \bar{y} + \hat{\beta}_x \bar{x} - \hat{\beta}_x x_i && \text{(distribute the minus sign)} \\
&= (y_i - \bar{y}) - \left(\hat{\beta}_x x_i - \hat{\beta}_x \bar{x}\right) && \text{(group terms)} \\
&= (y_i - \bar{y}) - \hat{\beta}_x (x_i - \bar{x}) && \text{(factor out } \hat{\beta}_x\text{)}
\end{aligned}
$$

Squaring and summing:

$$
\begin{aligned}
\sum_{i=1}^{n} \hat{\varepsilon}_i^2 &= \sum_{i=1}^{n} \Big((y_i - \bar{y}) - \hat{\beta}_x (x_i - \bar{x})\Big)^2 && \text{(substitute } \hat{\varepsilon}_i\text{)} \\
&= \sum_{i=1}^{n} \Big((y_i - \bar{y})^2 - 2\hat{\beta}_x (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_x^2 (x_i - \bar{x})^2\Big) && \text{(expand the square)} \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} 2\hat{\beta}_x (x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=1}^{n} \hat{\beta}_x^2 (x_i - \bar{x})^2 && \text{(sum of a sum)} \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2\hat{\beta}_x \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=1}^{n} \hat{\beta}_x^2 (x_i - \bar{x})^2 && \text{(constant factor out of the sum)} \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2\hat{\beta}_x \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_x^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 && \text{(constant factor out of the sum)} \\
&= S_{yy} - 2\hat{\beta}_x \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_x^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 && \text{(definition of } S_{yy}\text{)} \\
&= S_{yy} - 2\hat{\beta}_x S_{xy} + \hat{\beta}_x^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 && \text{(definition of } S_{xy}\text{)} \\
&= S_{yy} - 2\hat{\beta}_x S_{xy} + \hat{\beta}_x^2 S_{xx} && \text{(definition of } S_{xx}\text{)} \\
&= S_{yy} - 2\frac{S_{xy}}{S_{xx}} S_{xy} + \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} && \text{(substitute } \hat{\beta}_x = S_{xy}/S_{xx}\text{)} \\
&= S_{yy} - 2\frac{S_{xy}^2}{S_{xx}} + \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} && \text{(multiply)} \\
&= S_{yy} - 2\frac{S_{xy}^2}{S_{xx}} + \frac{S_{xy}^2}{S_{xx}^2} S_{xx} && \text{(square of a quotient)} \\
&= S_{yy} - 2\frac{S_{xy}^2}{S_{xx}} + \frac{S_{xy}^2}{S_{xx}} && \text{(cancel one factor of } S_{xx}\text{)} \\
&= S_{yy} - \frac{S_{xy}^2}{S_{xx}} && \text{(combine like terms)}
\end{aligned}
$$

Lemma 0.13 (Residual sum of squares of the OLS fit)**.** *If $S_{xx} > 0$ (Definition 0.11), the residual sum of squares at the OLS estimates (Theorem 0.4) is*

$$RSS(\hat{\beta}_0, \hat{\beta}_x) = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.$$

Proof

Proof. This is the solution to Exercise 0.20. □

Exercise 0.21 (R^2 and the correlation)**.** Assuming $S_{xx} > 0$ and $S_{yy} > 0$ (Definition 0.11), use Exercise 0.20 to write the coefficient of determination R^2 of the OLS fit in terms of the Pearson correlation coefficient r, and show that $0 \leq R^2 \leq 1$.

Solution

Solution 0.21. The total sum of squares is S_{yy} (Definition 0.11), and Exercise 0.20 gives the residual sum of squares, so:

$$
\begin{aligned}
R^2 &= 1 - \frac{\mathrm{RSS}(\widehat{\theta})}{\mathrm{TSS}} && \text{(definition of } R^2\text{)} \\
&= 1 - \frac{\mathrm{RSS}(\hat{\beta}_0, \hat{\beta}_x)}{\mathrm{TSS}} && \text{(the OLS estimate is } (\hat{\beta}_0, \hat{\beta}_x)\text{)} \\
&= 1 - \frac{\mathrm{RSS}(\hat{\beta}_0, \hat{\beta}_x)}{S_{yy}} && \text{(TSS is } S_{yy}\text{)} \\
&= 1 - \frac{S_{yy} - S_{xy}^2/S_{xx}}{S_{yy}} && \text{(residual sum of squares of the OLS fit)} \\
&= 1 - \left(\frac{S_{yy}}{S_{yy}} - \frac{S_{xy}^2/S_{xx}}{S_{yy}} \right) && \text{(distribute the division over the difference)} \\
&= 1 - \left(1 - \frac{S_{xy}^2/S_{xx}}{S_{yy}} \right) && (S_{yy}/S_{yy} = 1) \\
&= 1 - 1 + \frac{S_{xy}^2/S_{xx}}{S_{yy}} && \text{(distribute the minus sign)} \\
&= \frac{S_{xy}^2/S_{xx}}{S_{yy}} && (1 - 1 = 0) \\
&= \frac{S_{xy}^2}{S_{xx} S_{yy}} && \text{(divide a fraction by } S_{yy}\text{)} \\
&= \left(\frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \right)^2 && \text{(write as a square)} \\
&= r^2 && (r = S_{xy}/\sqrt{S_{xx} S_{yy}})
\end{aligned}
$$

Since $-1 \leq r \leq 1$ (range of the correlation coefficient), $0 \leq r^2 \leq 1$.

Theorem 0.6 (R^2 of a simple linear regression)**.** *For the OLS fit of a simple linear regression, with $S_{xx} > 0$ and $S_{yy} > 0$ (Definition 0.11), R^2 is the square of the Pearson correlation coefficient r of the x_i and y_i, and so lies between 0 and 1:*

$$R^2 = r^2, \quad 0 \leq R^2 \leq 1.$$

Proof

Proof. This is the solution to Exercise 0.21, which uses the residual sum of squares from Lemma 0.13. □

Example 0.8 (R^2 for the regression of glucose on BMI in HERS)**.** For the fit in Example 0.6, R^2 computed from Definition 0.15, the square of the Pearson correlation, and `lm()`'s value agree:

```
fit <- lm(glucose ~ BMI, data = hers)
y <- fit$model$glucose
c(
  r_squared = 1 - sum(residuals(fit)^2) / sum((y - mean(y))^2),
  cor_squared = cor(fit$model$BMI, y)^2,
  lm = summary(fit)$r.squared
)
#>   r_squared cor_squared          lm
#>   0.0743455   0.0743455   0.0743455
```

BMI accounts for only about 7% of the variation in baseline fasting glucose.

Read

- Hogg et al. [3, Sec. 7.6], pp. 344–352 covers more regression.
- Rice [4, Sec. 14.2.2], pp. 550–559 covers assessing the fit of a regression line.

Comparing nested regression models

Definition 0.16 (Nested linear regression models)**.** Let a linear regression model have p features and an intercept. A second linear regression model is **nested** in it if it uses the same outcome and the same observations and is the special case in which q of the p coefficients equal 0. The first model is the *larger* model, and the second is the *smaller* model; the smaller model has $p - q$ features.
Source: Vittinghoff et al. [1, Sec. 4.3].

Example 0.9 (Nested models for fasting glucose in HERS)**.** In the HERS data, take the linear regression of `glucose` on `BMI` as the smaller model. The larger model adds `age` and `exercise`, so $q = 2$:

$$\begin{aligned} \text{smaller:} \quad & \mu(\text{glucose}) = \beta_0 + \beta_{\text{BMI}} \cdot \text{BMI}, \\ \text{larger:} \quad & \mu(\text{glucose}) = \beta_0 + \beta_{\text{BMI}} \cdot \text{BMI} + \beta_{\text{age}} \cdot \text{age} + \beta_{\text{exercise}} \cdot \text{exercise}. \end{aligned}$$

Both models must be fitted to the same HERS participants, so keep the participants with all four variables measured:

```r
hers_nested <- hers |>
  tidyr::drop_na(glucose, BMI, age, exercise)
fit_smaller <- lm(glucose ~ BMI, data = hers_nested)
fit_larger <- lm(glucose ~ BMI + age + exercise, data = hers_nested)
nrow(hers_nested)
#> [1] 2758
```

The smaller model sets $\beta_{\text{age}} = \beta_{\text{exercise}} = 0$ in the larger one.

Definition 0.17 (Partial F-test)**.** Let a smaller linear regression model be nested in a larger model with p features, so that they differ by q coefficients, and let both be fitted by ordinary least squares to n observations. Write RSS_0 and RSS_1 for the residual sums of squares of the smaller and the larger fit. The **partial F-test** of

$$H_0 : \text{the } q \text{ coefficients are all } 0$$

uses the statistic

$$F \overset{\text{def}}{=} \frac{(\text{RSS}_0 - \text{RSS}_1)/q}{\text{RSS}_1/(n-p-1)}.$$

Its p-value is $\Pr(F^* \geq F)$, where F^* has the $F_{q,\,n-p-1}$ distribution (F-distribution).
Source: Vittinghoff et al. [1, Sec. 4.3] and Dobson and Barnett [15, Sec. 6.3].

Theorem 0.7 (Null distribution of the partial F statistic)**.** *Let the errors of the larger model be independent and Gaussian with mean* 0 *and the same variance. If* H_0 *of Definition 0.17 holds, then* F *has the* $F_{q,\,n-p-1}$ *distribution for every number of observations* $n > p + 1$.
Source: Vittinghoff et al. [1, Sec. 4.3].

Proof

Proof. Under H_0, $(\text{RSS}_0 - \text{RSS}_1)/\sigma^2$ has the χ^2_q distribution and RSS_1/σ^2 has the χ^2_{n-p-1} distribution, where σ^2 is the error variance. The two are independent, because RSS_1 is the squared length of the part of the outcome vector orthogonal to the larger model's fitted values, while $\text{RSS}_0 - \text{RSS}_1$ is the squared length of the part in the larger model but orthogonal to the smaller one. Dividing each by its degrees of freedom, σ^2 cancels from the quotient, which is therefore $F_{q,\,n-p-1}$ by the definition of the F-distribution. □

Example 0.10 (Partial F-test of age and exercise in HERS)**.** For the models of Example 0.9, `anova()` gives the partial F-test:

```
anova(fit_smaller, fit_larger) |>
  parameters::parameters() |>
  parameters::print_md(digits = 3, footer = "")
```

Parameter	RSS	Sum_Squares	df	df (error)	Mean_Square	F	p
Model 1: glucose ~ BMI	3.470e+06			2756			
Model 2: glucose ~ BMI + age + exercise	3.461e+06	9188.413	2	2754	4594.207	3.656	0.026

The definition gives the same statistic:

```
rss0 <- sum(residuals(fit_smaller)^2)
rss1 <- sum(residuals(fit_larger)^2)
n <- nobs(fit_larger)
p <- length(coef(fit_larger)) - 1
q <- p - (length(coef(fit_smaller)) - 1)
f_stat <- ((rss0 - rss1) / q) / (rss1 / (n - p - 1))
c(
  f = f_stat,
  p_value = pf(f_stat, q, n - p - 1, lower.tail = FALSE)
)
#>         f   p_value
#> 3.6558211 0.0259658
```

With $q = 2$ and $n - p - 1 = 2754$, $F = 3.66$, so the data reject H_0 at the 5% level.

i Remark (The partial F-test and the likelihood ratio test)

Remark 0.4 (The partial F-test and the likelihood ratio test). Let the errors be Gaussian. The maximum likelihood estimate of the error variance of a model with residual sum of squares RSS is RSS/n, so its maximized log-likelihood is $-\frac{n}{2}\log(2\pi\,\mathrm{RSS}/n) - \frac{n}{2}$. The likelihood ratio statistic of the null hypothesis H_0 in the partial F-test is then

$$\begin{aligned}\Lambda &= 2\Big(\ell(\hat{\theta}_{\mathrm{ML}}) - \ell(\hat{\theta}_0)\Big) \\ &= n \log \frac{\mathrm{RSS}_0}{\mathrm{RSS}_1} \\ &= n \log\left(1 + \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1}\right).\end{aligned}$$

For large n under H_0, the quantity $(\mathrm{RSS}_0 - \mathrm{RSS}_1)/\mathrm{RSS}_1$ is near 0, so $\log(1+u) \approx u$ gives $\Lambda \approx n(\mathrm{RSS}_0 - \mathrm{RSS}_1)/\mathrm{RSS}_1$. Also $qF = (n-p-1)(\mathrm{RSS}_0 - \mathrm{RSS}_1)/\mathrm{RSS}_1$, which is close to the same value when n is much larger than p. Under H_0, Λ and qF therefore have approximately the χ^2_q distribution, but the F-test is exact for Gaussian errors at every n, while the likelihood ratio test holds only approximately.

Further reading

Linear Models Overview covers linear regression in depth, including inference for the coefficients and multiple predictors. Vittinghoff et al. [1] cover linear regression in Chapter 4.

References

[1] E. Vittinghoff, D. V. Glidden, S. C. Shiboski, and C. E. McCulloch, *Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models*, 2nd ed. Springer, 2012. doi: 10.1007/978-1-4614-1353-0.

[2] Wikipedia contributors, "Pearson correlation coefficient — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

[3] R. V. Hogg, E. A. Tanis, and D. L. Zimmerman, *Probability and statistical inference*, Tenth edition. Boston: Pearson, 2019.

[4] J. A. Rice, *Mathematical statistics and data analysis*, 3rd ed. Belmont, CA: Cengage Learning, 2007.

[5] Wikipedia contributors, "Spearman's rank correlation coefficient — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

[6] Wikipedia contributors, "Homoscedasticity and heteroscedasticity — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Homoscedasticity_and_heteroscedasticity

[7] Wikipedia contributors, "Linear regression — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Linear_regression

[8] Wikipedia contributors, "Simple linear regression — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Simple_linear_regression

[9] Wikipedia contributors, "Residual sum of squares — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Residual_sum_of_squares

[10] Wikipedia contributors, "Ordinary least squares — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Ordinary_least_squares

[11] B. Hutchinson, "DATA 471/571 (machine learning) and CSCI 481/581 (deep learning) video lectures." Western Washington University. Accessed: Sep. 28, 2026. [Online]. Available: https://facultyweb.cs.wwu.edu/~hutchib2/video_lectures/data371/

[12] G. Casella and R. Berger, *Statistical inference*, 2nd ed. Cengage Learning, 2002. Available: https://www.cengage.com/c/statistical-inference-2e-casella-berger/9780534243128/

[13] Wikipedia contributors, "Total sum of squares — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Total_sum_of_squares

[14] Wikipedia contributors, "Coefficient of determination — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Coefficient_of_determination

[15] A. J. Dobson and A. G. Barnett, *An introduction to generalized linear models*, 4th ed. CRC press, 2018. doi: 10.1201/9781315182780.