Example: Cats Data

- Goal: describe the relationship between Y (e.g., heart weight) and X (e.g., body weight). As a starting point, we assume the relationship is linear.
- Data $(y_i, x_i)_{i=1}^n$, where $y_i, x_i \in \mathbb{R}$.
- Apparently the data won't be able to fit on a straight line. Assume

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

$$\begin{aligned} (\beta_0, \beta_1) &: \quad \text{unknown regression coefficients,} \\ e_i's &: \quad \text{often assume to have mean } 0 \text{ and variance } \sigma^2 \end{aligned}$$

Overview for SLR (I)

- How to use LS to estimate (β_0, β_1)? We can obtain an explicit expression for $(\hat{\beta}_0, \hat{\beta}_1)$. There is a nice connection between the LS estimate of the slope, $\hat{\beta}_1$, and sample correlation/variance of X and Y, which will help you to remember the expression.

- Some jargons: fitted value, residual, RSS, R-square (used to access the overall model fit).

- How would the LS fitting/inference be affected if the data, X and/or Y, are shifted and/or scaled (i.e., linear transformed)?

- *SLR without the intercept*: fit a regression line passing the origin.

Parameter Estimation by Least Squares

We would like to choose a line which is close to the data points. We measure the closeness by squared errors[a].

Least Squares Estimation: find $(\hat{\beta}_0, \hat{\beta}_1)$ that minimize the residual sum of squares (RSS)

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

To find the solution, we have

$$\begin{aligned} \frac{\partial \mathsf{RSS}}{\partial \beta_0} &= -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) = 0, \\ \frac{\partial \mathsf{RSS}}{\partial \beta_1} &= -2 \sum_i x_i (y_i - \beta_0 - \beta_1 x_i) = 0. \end{aligned}$$

[a] Why squared error? Why not absolute error?

Re-arrange the equations,

$$\begin{aligned} \beta_0 n + \beta_1 \sum x_i &= \sum y_i, & (1) \\ \beta_0 \sum x_i + \beta_1 \sum x_i^2 &= \sum x_i y_i. & (2) \end{aligned}$$

From (1), we have

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Plug it back to (2),

$$\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i$$

$$\beta_1 \left(\sum x_i^2 - \sum x_i \bar{x} \right) = \sum x_i y_i - \sum x_i \bar{y}$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \sum x_i \bar{y}}{\sum x_i^2 - \sum x_i \bar{x}} = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})}.$$

Some equalities (basically centering one side is the same as centering both sides for cross-products):

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i(y_i - \bar{y}) = \sum_i (x_i - \bar{x}) y_i.$$

So the LS estimates of (β_0, β_1) can be expressed as

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \\ \hat{\beta}_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})(x_i - \bar{x})} = \frac{\mathsf{Sxy}}{\mathsf{Sxx}} = r_{\mathsf{XY}} \frac{\sqrt{\mathsf{Syy}}}{\sqrt{\mathsf{Sxx}}}, \end{aligned}$$

where

$$\begin{aligned} \mathsf{Sxy} &= \sum (x_i - \bar{x})(y_i - \bar{y}), \\ \mathsf{Sxx} &= \sum (x_i - \bar{x})^2, \quad \mathsf{Syy} = \sum (y_i - \bar{y})^2, \\ r_{\mathsf{XY}} &= \frac{\mathsf{Sxy}}{\sqrt{(\mathsf{Sxx})(\mathsf{Syy})}} \quad \text{(the sample correlation).} \end{aligned}$$

$$\hat{\beta}_1 = r_{\mathsf{XY}} \frac{\sqrt{\mathsf{Syy}}}{\sqrt{\mathsf{Sxx}}},$$

It is not surprising that the LS estimate of the coefficient is related to the sample correlation between X and Y. Recall that SLR assumes the dependence between X and Y is linear. Correlation is exactly the measure used to quantify the linear dependence between two variables[a].

[a]It is easy to construct an example, where Y depends on X via a nonlinear function and their correlation is zero.

Suppose we know the mean, variance of X and Y, and their correlation r. What is your guess of y given x? It seems reasonable to guess the "unit-free, location/scale invariant" version of Y by r times the "unit-free, location/scale invariant" version of X, i.e.,

$$\frac{y - \mu_y}{\sigma_y} \approx r_{\mathsf{xy}} \frac{x - \mu_x}{\sigma_x}.$$

Replace the mean, variance and correlation by the corresponding sample version:

$$\begin{aligned}
\frac{y - \bar{y}}{\sqrt{\mathsf{Syy}}} \approx r_{\mathsf{xy}} \frac{x - \bar{x}}{\sqrt{\mathsf{Sxx}}} \quad &\Longrightarrow \quad y - \bar{y} \approx r_{\mathsf{xy}} \sqrt{\frac{\mathsf{Syy}}{\mathsf{Sxx}}} (x - \bar{x}) \\
&\Longrightarrow \quad y \approx \left(\bar{y} - r_{\mathsf{xy}} \sqrt{\frac{\mathsf{Syy}}{\mathsf{Sxx}}} \bar{x} \right) + \left(r_{\mathsf{xy}} \sqrt{\frac{\mathsf{Syy}}{\mathsf{Sxx}}} \right) x
\end{aligned}$$

Some jargons:

- Fitted value at x_i or the prediction of y_i: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

- Residual at x_i: $r_i = y_i - \hat{y}_i$. Note that the two equations on p6 imply that

$$\sum_i r_i = 0, \quad \sum_i r_i x_i = 0.$$ [a]

- RSS $= \sum_{i=1}^n r_i^2$.

- The error variance is estimated by

$$\hat{\sigma}^2 = \frac{1}{n-2}\text{RSS} = \frac{1}{n-2}\sum_{i=1}^n r_i^2.$$

  The degree of freedom (df) of the residuals is $n-2$. In general

$$df\,(\text{residuals}) = \text{sample-size} - \text{number-of-parameters}.$$

[a] $\sum_i r_i = 0$ implies that the sample mean of $\hat{y}_i$ is just $\bar{y}$.

Goodness of Fit: R-square

Note the total variation (TSS) in y can be decomposed into the summation of RSS and the total variation in the fitted value $\hat{y}$ (FSS):

$$\begin{aligned}
\sum_i (y_i - \bar{y})^2 &= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_i (r_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_i r_i^2 + \sum_i (\hat{y}_i - \bar{y})^2 \qquad (3) \\
&= \mathsf{RSS} + \mathsf{FSS},
\end{aligned}$$

where the cross-product

$$\sum_i r_i(\hat{y}_i - \bar{y}) = \hat{\beta}_0 \sum_i r_i + \hat{\beta}_1 \sum_i r_i x_i - \bar{y} \sum_i r_i = 0.$$

Also note that the average of $\hat{y}_i$'s is the same as the average of y_i; this is true when intercept is included in the model.

A common measure on how well the model fits the data is the so-called coefficient of determination or simply R-square:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\mathsf{FSS}}{\mathsf{TSS}} = \frac{\mathsf{TSS} - \mathsf{RSS}}{\mathsf{TSS}} = 1 - \frac{\mathsf{RSS}}{\mathsf{TSS}}.$$

For a given data set where TSS is fixed, so smaller the RSS, larger the R^2.

We can also show that $R^2 = r^2_{\mathsf{XY}}$.

$R^2 = \frac{\mathsf{Var}(\hat{y})}{\mathsf{Var}(y)}$ measures how much variation in the original data y_i's is explained or reduced by the LS fitting. If Y and X are strongly linear dependent, a linear function of X can help to reduce the uncertainty (i.e., variation) of Y.

How Affine Transformations on the Data Affect Regression?

Suppose we have run a SLR model of Y on X.

- If we rescale the data y_i by $\tilde{y}_i = ay_i + b$, and then regress $\tilde{y}_i$ on x_i. How would the LS estimates and R^2 be affected?

- If we rescale the covariates x_i by $\tilde{x}_i = ax_i + b$, and then regress y_i on $\tilde{x}_i$. How would the LS estimates and R^2 be affected?

- If we regression X on Y instead, will the LS line be the same? How about R^2?

Regression Through the Origin

Sometimes we want to fit a line with no intercept (regression through the origin): $y_i \approx \beta_1 x_i$. For example, x_i denotes the intensity level of various exercises and y_i denotes the additional calories you burn with those exercises.

We can estimate β_1 using the LS principle

$$\min_{\beta_1} \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2 \Longrightarrow \hat{\beta}_1 = \frac{\sum_i x_i y_i}{\sum_i x_i^2}.$$

The ordinary definition of R-square is no longer meaningful; you could have RSS bigger than TSS, and therefore have a negative R-square, if you use formula $R^2 = 1 - \text{RSS}/\text{TSS}$.

The ordinary R-square measures the effect of X after removing the effect of the intercept by centering both y_i's and $\hat{y}_i$'s. For regression models with no intercept, we shouldn't do the centering when computing R-square.

Let's look at the following decomposition (slightly different from (3))

$$\sum_i y_i^2 = \sum_i (y_i - \hat{y}_i + \hat{y}_i)^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i \hat{y}_i^2.$$

Then define R-square for regression with no intercept as

$$\tilde{R}^2 = \frac{\sum_i \hat{y}_i^2}{\sum_i y_i^2} = 1 - \frac{\mathsf{RSS}}{\sum_i y_i^2}.$$

Remarks

- I want to emphasize here that $(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)$ are not the values of the true parameters $(\beta_0, \beta_1, \sigma^2)$, but estimates/estimators. This is why we put a hat on those symbols. If we happen to collect another data set, their values would be different; they are functions of the data, and therefore they are random variables.

- Next we'll *1)* check the statistical properties (such as unbiasedness or MSE) of those estimates, and *2)* do some statistical inference under the normal assumption.

Overview for SLR (II)

- Regarding the statistical properties of the LS estimates, we first check the properties of $(\hat{\beta}_0, \hat{\beta}_1)$ as an estimate of the true coefficient vector (β_0, β_1).
- We can compute their mean, variance and covariance. We can show that they are unbiased.
- We can also show that they achieve the smallest MSE among all unbiased estimators; this result holds general for MLR.
- Till this point, we only need to assume the 1st and 2nd moments of e_i's, i.e., $\mathbb{E}e_i = 0$, $\mathsf{Var}(e_i) = \sigma^2$, $\mathsf{Cov}(e_i, e_j) = 0, \ i \neq j$.

- For hypothesis testing and construct confidence/prediction intervals, we need to derive the distribution of $(\hat{\beta}_0, \hat{\beta}_1)$.

- We can make iid normal assumptions on e_i's; then use t-dist in testing and interval estimation.

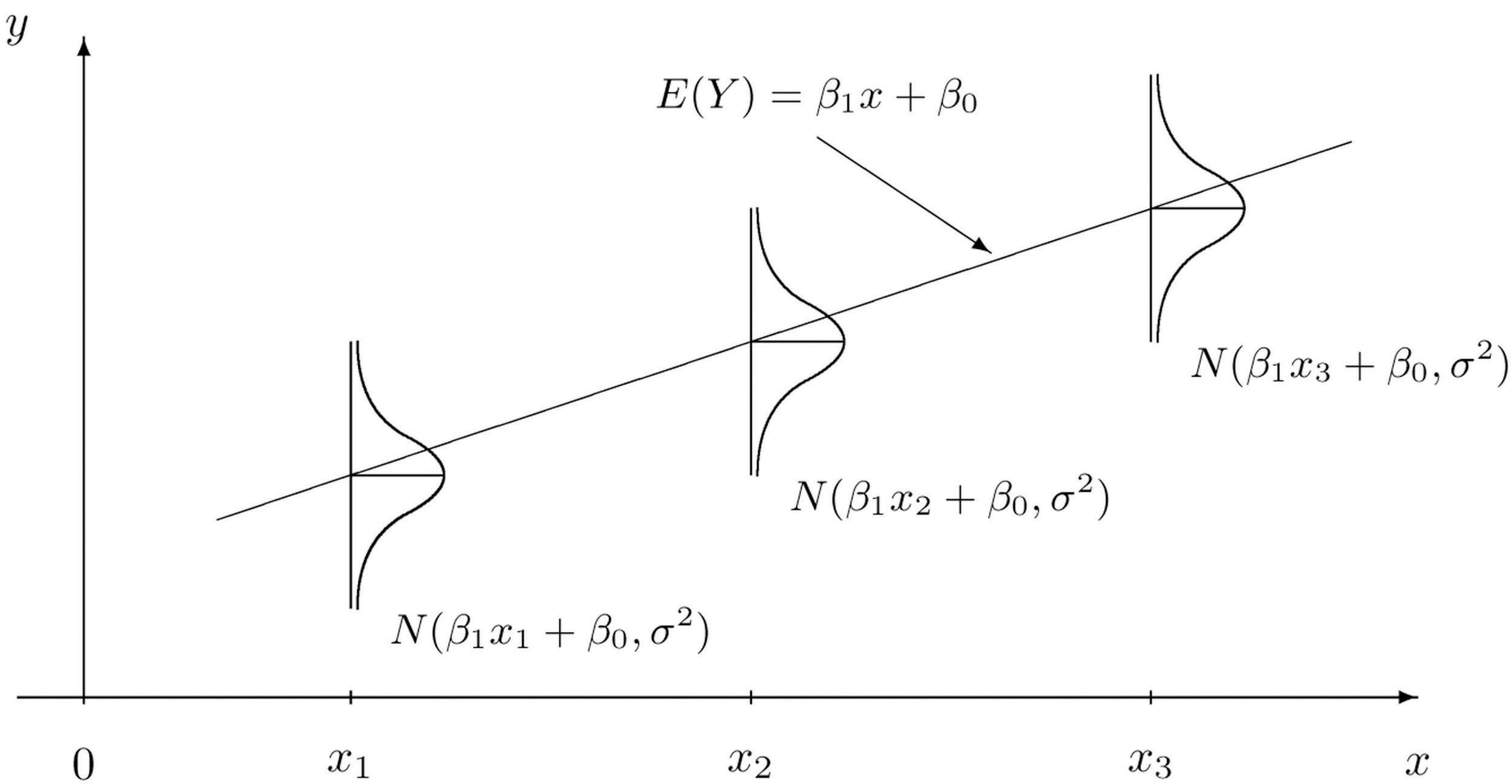
- OR, we can stick to the original weaker assumption on just the 1st and 2nd moments, and then call CLT to approximate the distribution of $(\hat{\beta}_0, \hat{\beta}_1)$, as well as some test statistics, by normals, when the sample size n is large enough.

Normal Assumptions

Assume: $y_i = \beta_0 + \beta_1 x_i + e_i$, and

$$e_i \text{ iid } \sim \mathsf{N}(0, \sigma^2), \text{ or equivalently, } y_i \text{ indep.} \sim \mathsf{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$

- The mean function is linear: $\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i$.
- Errors e_i's are independent; data y_i's are independent.
- Errors e_i's have homogeneous variance: $\mathsf{Var}(e_i) = \sigma^2$, and so are data y_i's.
- Each e_i is normally distributed and each y_i is normally distributed.
- Note that each e_i is normal + independence, so they are jointly normal. Consequently y_i's are jointly normal, and so are any linear combinations of y_i's, which is an important result that will be used later in our inference.



y
$E(Y) = \beta_1 x + \beta_0$
$N(\beta_1 x_3 + \beta_0, \sigma^2)$
$N(\beta_1 x_2 + \beta_0, \sigma^2)$
$N(\beta_1 x_1 + \beta_0, \sigma^2)$
0
x_1
x_2
x_3
x

Distributions of the LS estimates

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are jointly normally distributed with

$$\mathbb{E}\hat{\beta}_1 = \beta_1, \qquad \mathsf{Var}(\hat{\beta}_1) = \sigma^2 \frac{1}{\mathsf{Sxx}}$$

$$\mathbb{E}\hat{\beta}_0 = \beta_0, \qquad \mathsf{Var}(\hat{\beta}_0) = \sigma^2 \Big(\frac{1}{n} + \frac{\bar{x}^2}{\mathsf{Sxx}} \Big)$$

$$\mathsf{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{\mathsf{Sxx}}.$$

- RSS $\sim \sigma^2 \chi^2_{n-2}$ and therefore

$$\mathbb{E}\hat{\sigma}^2 = \frac{\mathbb{E}\ \mathsf{RSS}}{n-2} = \sigma^2.$$

- $(\hat{\beta}_0, \hat{\beta}_1)$ and RSS are independent (which will be proved for MLR later).

Hypothesis Testing

- Test $H_0 : \beta_1 = c$ versus $H_a : \beta_1 \neq c$
- The test statistic

$$t = \frac{\hat{\beta}_1 - c}{\text{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - c}{\hat{\sigma}/\sqrt{\text{Sxx}}} \sim T_{n-2} \text{ under } H_0.$$

- p-value $= 2 \times$ the area under the T_{n-2} dist more extreme than the observed statistic t.
- The p-value returned by the R command lm is for the test with $H_0 : \beta_1 = 0$.

F-test and ANOVA

An alternative way to test $\beta_1 = 0$ is based on the F-test. It can shown that t-test is equivalent to an F-test.