3 Simple Linear Regression

3.1 Introduction

Two Steps to Linear Regression

1. **Data:** summarize observed data, data fitting (estimation); no assumptions required
2. **Inference:** infer information (estimator properties, hypothesis testing, confidence intervals) about the regression function; assumptions required

Basic Model

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

where Y_i is a random (*independent* or *response*) variable, x_i are known or observable (*dependent* or *predictor*) variables, β_0, β_1 are fixed, unknown parameters, and $E[\epsilon_i] = 0$.

Given the last assumption, the *population regression function* is

$$E[Y_i \mid x_i] = \beta_0 + \beta_1 x_i$$

where we condition on x_i in the expectation to make the dependence on x_i explicit. (In some cases we'll consider later, x_i might be the realization of a random variable, in which case conditioning is as you learned in the probability introduction.)

Notice that so far there are two population parameters β_0 and β_1 and a population regression function $\beta_0 + \beta_1 x$. These together represent the population from which we sample Y_i and includes our assumption of a *linear model*. These population parameters and function are analogous to the population parameters μ and σ^2 for $Y_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ sampling models of basic statistical inference.

Linear Model

A linear model m eans that the conditional expectation is a linear function in the unknown parameters. The following are linear models (mix and match left and right sides as desired):

$$\begin{aligned} E[Y_i \mid x_i] &= \beta_0 + \beta_1 x_i \\ E[\log(Y_i) \mid x_i] &= \beta_0 + \beta_1 x_i^2 \\ E[\text{logit}(Y_i) \mid x_i] &= \beta_0 + \beta_1 / x_i \end{aligned}$$

The following are not valid right-hand-sides for the simple linear regression model:

$$\begin{aligned} \beta_0 + \beta_1^2 x_i \\ \beta_0 + \log(\beta_1) x_i \\ \text{...etc....} \end{aligned}$$

Is it reasonable to assume a linear model? No, probably not. The available theory may not support a lienar model, however it is very convenient mathematically and often does a remarkable job at approximating what is actually complex and nonlinear. *We assume a linear relationship adequately captures the true relationship.*

$$E[Y_i] \approx \beta_0 + \beta_1 x_i$$

Bivariate Normal Distribution

The bivariate normal distribution plays a prominent role in some forms of simple linear regression, so we shall describe it now.

Definition: bivariate normal distribution A length 2 vector of random variables has a bivariate normal distribution, and we write

$$(X, Y) \sim BivariateNormal(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho),$$

if the joint pdf is

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} exp\left\{\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\frac{x-\mu_X}{\sigma_X}\frac{y-\mu_Y}{\sigma_Y} + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\}$$

Notice that $X \in (-\infty, \infty), Y \in (-\infty, \infty), \mu_X \in (-\infty, \infty), \mu_Y \in (-\infty, \infty), \sigma_X \in (0, \infty), \sigma_Y \in (0, \infty)$. Also, one can derive the bivariate normal distribution by hypothesizing two independent standard normal Z_1, Z_2 variables and applying the following change-of-variable:

$$\begin{aligned} X &= \sqrt{\frac{1+\rho}{2}}\sigma_X Z_1 + \sqrt{\frac{1-\rho}{2}}\sigma_X Z_2 + \mu_X \\ Y &= \sqrt{\frac{1+\rho}{2}}\sigma_Y Z_1 - \sqrt{\frac{1-\rho}{2}}\sigma_Y Z_2 + \mu_Y \end{aligned}$$

Properties of Bivariate Normal

1. $X \sim N(\mu_X, \sigma_X^2)$
2. $Y \sim N(\mu_Y, \sigma_Y^2)$
3. Corr$(X, Y) = \rho = \frac{\text{Cov(X,Y)}}{\sigma_X\sigma_Y}$
4. $aX + bY \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y)$
5. $Y \mid X \sim N(\mu_X + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X), \sigma_Y^2(1-\rho^2))$

3.2 Least Squares Estimation

Problem Setup

We observe $(x_1, y_1), \ldots, (x_n, y_n)$ and hypothesize a linear relationship between X and Y. The points will not fall on a perfect straight line because of randomness, but we "summarize" the relationship as a line that is somehow best characterizing the relationship.

Example: You calibrate an instrument that measures the concentration of a compound with various mixtures of known concentration. For each known concentration, you use the instrument to measure the concentration several times to produce the data linked on the webpage.

To summarize the data, we can compute the summary statistics

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} y_i \qquad \text{sample means}$$

$$S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad S_{YY} = \textstyle\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad \text{sums of squares}$$

$$S_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad \text{sum of cross produces}$$

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \qquad \text{common estimates of population parameters } \beta_0, \beta_1$$

Least Squares Estimates of β_0, β_1

We seek a straight line that comes as close as possible to all points in some sense. We need not make any statistical assumptions to come up with this line. It is a purely mathematical argument.

Definition: (regression) residual

Let $\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ be the difference between the observed y_i and that predicted $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ by the estimated regession equation.

Definition: residual sum-of-squares

$$\text{RSS} = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \sum_{i=1}^{n}\hat{\epsilon}_i^2$$

Sometimes, statisticians write SSE (sum of squared error) for RRS.

Definition: least-squares estimators

The least squares estimators of β_0 and β_1 are those which minimize RSS

$$\hat{\beta}_0, \hat{\beta}_1 = \text{argmin}_{\beta_0,\beta_1} RSS$$

Theorem:

The least squares estimators are

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \qquad \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

Proof:

First consider the following Lemma.

Lemma:

Consider $x_1, \ldots, x_n$ with $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then, $\min_a \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Proof:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - a)^2 &= \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - a)^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - a) + \sum_{i=1}^{n}(\bar{x} - a)^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - a)^2
\end{aligned}$$

and clearly we can see that $a = \bar{x}$ minimizes this sum.

Continuing, it is clear that $\hat{\beta}_0 = \overline{y_i - \beta_1 x_i} = \bar{y} - \beta_1 \bar{x}$. To minimize over β_1, we are left with

$$\begin{aligned}\sum_{i=1}^{n} (y_i - \beta_1 x_i - \bar{y} - \beta_1 \bar{x})^2 &= \sum_{i=1}^{n} [(y_i - \bar{y}) - \beta_1 (x_i - \bar{x})]^2 \\ &= S_{YY} - 2\beta_1 S_{XY} + \beta_1^2 S_{YY}\end{aligned}$$

Take the derivative and set to 0,

$$\begin{aligned}\frac{\partial}{\partial \beta_1} \left(S_{YY} - 2\hat{\beta}_1 S_{XY} + \hat{\beta}_1^2 S_{XX} \right) &= 0 \\ 2\hat{\beta}_1 S_{XX} - 2S_{XY} &= 0 \\ \hat{\beta}_1 &= \frac{S_{XY}}{S_{XX}}\end{aligned}$$

Taking second derivatives verifies that this is a minimum.

Alternative Estimation

There are other ways to pick a best fitting line. Perhaps we could minimize the sum of squared horizontal distances between points and the line. Or, we could use another distance measure.

However, if x_i is predictor and y_i is the response, then it is reasonable to try to minimize the distance from our prediction (the estimated regression line) to the observed y_i.

How good is the fit?

Notice that the total variation in y_i can be partitioned.

$$\begin{aligned}\sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\ &= \text{RSS} + \text{regression sum of squares}\end{aligned}$$

where the first term is the residual sum of squares, which captures the extra variation observed in the data around the fitted line, and regression sum of squares, which represents the amount of variability in Y_i captured by the line. For example, if $\hat{y}_i = \bar{y}$ for all i, then the fitted line is flat and regression sum of squares is 0. In this case, the line does not account for any of the variability in Y. On the other hand if RSS$= 0$, then the points lie on a straight line and the line itself explains all of the variation in Y, i.e. knowing x_i, we can perfectly predict y_i.

Interpretting β_0 and β_1

Recall that $\beta_0 = E[Y \mid x = 0]$ is the expected observation when the predictor $x = 0$. This quantity is usually less meaningful than β_1, which is the slope of the line. The slope tells use how we expect Y to change given a particular change in the level x.

3.3 Properties of Least Squares Estimators

Now that we have an estimator, namely $\hat{\beta}_0 + \hat{\beta}_1 x$ for the population regression function, we seek to derive some of the statistical properties of this estimate. Now, we will be required to start making assumptions about the probability model for the sampled data $Y_1, \ldots, Y_n$.

Mean Squared Error

Definition: bias

The bias of an estimator W for population parameters θ is

$$\text{bias}_\theta(W) = E[W] - \theta$$

As discussed previously, *unbiased estimators* are those statistics W for which $\text{bias}_\theta(w) = 0$.

A good way to evaluate the accuracy of an estimator is via the MSE.

Definition: mean squared error

The mean square error for estimator W of population parameter θ is defined as

$$\text{MSE}(W) = \text{E}\left[(W-\theta)^2\right]$$

Properties: We will show one important property of the MSE.

$$\begin{aligned}
\text{E}\left[(W-\theta)^2\right] &= \text{E}\left[(W - E[W] + E[W] - \theta)^2\right] \\
&= \text{Var}(W) + (E[W]-\theta)^2 \\
&= \text{Var}(W) + \text{bias}_\theta(W) \\
&= \text{precision} + \text{accuracy}
\end{aligned}$$

Example: sample variance

Estimation of the sample variance provides a demonstration of the bias-variance trade-off.

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then we know the sample mean is an unbiased estimator for μ, i.e. $E[\bar{X}] = \mu$. It's MSE is the sample variance $MSE(\bar{X}) = \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$. The sample variance S^2 is also unbiased $E[S^2] = \sigma^2$. Let's consider its MSE.

$$\begin{aligned}
\text{MSE}(S^2) &= \text{Var}(S^2) \\
&= \frac{2\sigma^4}{n-1} \qquad \text{Can you derive this?}
\end{aligned}$$

Let's compare this sample variance estimator to another (the mle for σ^2)

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n-1}{n}S^2.$$

This estimator is biased, with bias

$$\text{bias}_{\sigma^2}(\hat{\sigma}^2) = E\left[\frac{n-1}{n}S^2\right] - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2$$

It has MSE

$$\begin{aligned}
\text{MSE}(\hat{\sigma}^2) &= \frac{2(n-1)\sigma^4}{n^2} + \left(\frac{-\sigma^2}{n}\right)^2 \\
&= \frac{(2n-1)\sigma^4}{n^2} < \frac{2\sigma^4}{n-1}
\end{aligned}$$

Therefore, the mle $\hat{\sigma}^2$ is biased, but has lower MSE than the sample variance S^2. We usually prize unbiased estimators highly (we prefer accuracy over precision) and accept the decrease in MSE.

The Bias-Variance Tradeoff

The S^2 vs. $\hat{\sigma}^2$ dilemma introduces the *bias-variance tradeoff*. In order to achieve unbiased estimates, one must sacrifice variance. To reduce variance, one must accept bias.

Linear Estimator

Definition: linear estimator

Consider a sample $Y_1, \ldots, Y_n$. A linear estimator is one of the form

$$W = \sum_{i=1}^{n} d_i Y_i$$

for some fixed and known constants d_i.

For example, the sample mean is a linear estimator.

Linear, Unbiased Estimators for β_0 and β_1

Consider linear estimators for our parameters, say

$$\begin{aligned} \hat{\beta}_1 &= \sum_{i=1}^{n} d_i Y_i \\ \hat{\beta}_0 &= \sum_{i=1}^{n} c_i Y_i \end{aligned}$$

Suppose we want these estimators to also be unbiased. Specifically $E[\hat{\beta}_1] = \beta_1$, so

$$\begin{aligned} E[\hat{\beta}_1] &= E\left[\sum_{i=1}^{n} d_i y_i\right] \\ &= \sum_{i=1}^{n} d_i E[Y_i] \\ &= \sum_{i=1}^{n} d_i \left(\beta_0 + \beta_1 x_i\right) \\ &= \beta_0 \sum_{i=1}^{n} d_i + \beta_1 \sum_{i=1}^{n} d_i x_i \end{aligned}$$

implies

$$\sum_{i=1}^{n} d_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} d_i x_i = 1$$

Similarly, linear estimator $\hat{\beta}_0$ is unbiased if

$$\sum_{i=1}^{n} c_i = 1 \qquad \text{and} \qquad \sum_{i=1}^{n} c_i x_i = 0$$

Variance of Linear Estimator

The variance of a linear estimator when the sample $Y_1, \ldots, Y_n$ is iid, with variance $\text{Var}(Y_i) = \sigma^2$, is very easy to compute:

$$\text{Var}\left(\sum_{i=1}^{n} d_i Y_i\right) = \sum_{i=1}^{n} d_i^2 \text{Var}(Y_i) = \sigma^2 \sum_{i=1}^{n} d_i^2$$

BLUE

Definition. BLUE

The BLUE is the Best Linear Unbiased Estimator and is the linear, unbiased estimator with smallest variance (and hence MSE).

BLUE for β_1

The BLUE for β_1 is the selection of d_i that minimizes

$$\min_{d_i, i=1,\ldots,n} \sum_{i=1}^{n} d_i^2$$

such that the following conditions are satisfied

$$\sum_{i=1}^{n} d_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} d_i x_i = 1$$

The solution (obtained with some effort) is

$$d_i = \frac{x_i - \bar{x}}{S_{xx}}$$

BLUE for β_0

Similarly, the BLUE for β_0 has

$$c_i = \frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{XX}}$$

Variance for $\hat{\beta}_0$ and $\hat{\beta}_1$

Using the variance formula for linear estimators, we have

$$\text{Var}(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{S_{XX}^2} = \frac{\sigma^2}{S_{XX}}$$

and

$$\text{Var}(\hat{\beta}_0) = \frac{\sigma^2}{nS_{XX}} \sum_{i=1}^{n} x_i^2$$

LS Estimators are BLUEs

One can show that the d_i and c_i given above yield the LS estimators we already derived.

3.4 Hypothesis Testing and Confidence Intervals

The Conditional Normal Model

Suppose

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

for all $i = 1, \ldots, n$ are independent.

Equivalently, we could state $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and impose restriction

$$\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

These models include the following hypotheses:

1. Independent observations
2. Common variance for all levels x_i
3. Normal distribution
4. Linear mean

Bivariate Normal Model

Suppose you observe an iid sample of bivariate normal random vectors $(X_i, Y_i) \overset{\text{iid}}{\sim} \text{BivN}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$. Notice that now the levels X_i are themselves random.

Then, we know

$$\begin{aligned} \text{E}[Y \mid X] &= \mu_y + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X) \\ \text{Var}(Y \mid X) &= \sigma_Y^2(1 - \rho^2), \end{aligned}$$

which implies

- The conditional expectation is linear with $\beta_0 = \mu_y - \frac{\rho\sigma_Y\mu_X}{\sigma_X}$ and $\beta_1 = \frac{\rho\sigma_Y}{\sigma_X}$.
- The variance is independent of level X.

For all intents and purposes, the bivariate normal model reduces to the Conditional Normal Model. Therefore, henceforth, we will focus on the conditional normal model and keep in mind that it also applies for the bivariate normal model.

Likelihood

Now that we have a fully specified probability model for our data $y = (y_1, \ldots, y_n)$ (now we use small y to indicate that these data are an observed realization of the random vector Y with elements Y_i), we can write down the likelihood

$$\begin{aligned} f(y \mid \beta_0, \beta_1, \sigma^2) &= f(y_1, \ldots, y_n \mid \beta_0, \beta_1, \sigma^2) \\ &= \textstyle\prod_{i=1}^n f(y_i \mid \beta_0, \beta_1, \sigma^2) && \text{by independence} \\ &= \textstyle\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right] && \text{normal distribution} \\ &= \textstyle\frac{1}{\sigma^2(2\pi)^{n/2}} \exp\left[-\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2/(2\sigma^2)\right] \end{aligned}$$

Maximum Likelihood Estimation

With a likelihood in hand, the MLEs become accessible to us. Recalling the procedure, we compute first the log likelihood.

$$\log f(y \mid \beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

To maximize the log likelihood we start by observing that only the last term involves β_0 and β_1, so to maximize with respect to β_0 and β_1 is equivalent to *minimizing* the last sum

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

Yet, we minimized this expression to obtain our LS estimates, thus the mles are

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{S_{XY}}{S_{XX}}\end{aligned}$$

To find the mle of σ^2, we take the partial derivative and set it equal to 0. The result is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Lemma: Covariance of Linear Combinations of Random Variables

Our next goal is to show that $\hat{\sigma}^2$ is biased.

Lemma:

Suppose $Y_1, \ldots, Y_n$ are uncorrelated with $\text{Var}(Y_i) = \sigma^2$ for all i. Suppose $c_1, \ldots, c_n$ and $d_1, \ldots, d_n$ are fixed constants. Then

$$\text{Cov}\left(\sum_{i=1}^{n} c_i Y_i, \sum_{j=1}^{n} d_j Y_j\right) = \sigma^2 \sum_{i=1}^{n} c_i d_i.$$

Proof:

You have come very close to proving this result in homework 2, except you did it for normally distributed Y_i.

$\hat{\sigma}^2$ is Biased

To show that $\hat{\sigma}^2$ is biased, first notice

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2$$

Further,

$$\text{E}[\hat{\epsilon}_i] = E[Y_i] - E[\hat{\beta}_0] - E[\hat{\beta}_1]x_i = \beta_0 + \beta_1 x_i - \beta_0 - \beta_1 x_i = 0$$

and (after some algebra)

$$\text{Var}(\hat{\epsilon}_i) = \text{E}[\hat{\epsilon}_i^2] = \sigma^2\left[\frac{n-2}{n} + \frac{1}{S_{XX}}\left(\frac{1}{n}\sigma_{j=1}^{n} x_j^2 + x_i^2 - 2(x_i - \bar{x})^2 - 2x_i\bar{x}\right)\right].$$

However, $S_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_i x_i^2 - 2\bar{x}\sum_i x_i + n(\bar{x})^2 = \sum_i x_i^2 - n\bar{x}^2$, so

$$\text{Var}(\hat{\sigma}^2) = \left(\frac{n-2}{n}\right)\sigma^2$$

is biased.

Unbiased Sample Variance

Valuing unbiased estimators over all else, we prefer

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Theorem: sampling distributions for conditional normal model

Theorem:

Assume the conditional normal distribution, then sampling distributions of $\hat{\beta}_0, \hat{\beta}_1$, and S^2 have the following properties:

1. $\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{nS_{XX}}\sum_{i=1}^{n} x_i^2\right)$
2. $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right)$
3. Cov)$(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \bar{X}}{S_{XX}}$
4. $(\hat{\beta}_0, \hat{\beta}_1)$ and S^2 are independent
5. $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

Proof (partial):

$\hat{\beta}_0$ and $\hat{\beta}_1$ are linear functions of independent normal random variables, therefore they are normally distributed. By construction, the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased, which yields the population means given above. We also computed their variance earlier, when introducing linear statistics.

To show independence, we notice

$$\hat{\epsilon}_i = \sum_{j=1}^{n} \left[\delta_{ij} - (c_j + d_j x_j)\right] Y_i$$

where $\delta_{ij} = 1$ iff $i = j$ and c_j and d_j are the fixed coefficients for the linear estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively.

To compute covariance, use the lemma for covariance of two linear combinations of random variables.

You can show, using the same lemma, that Cov$(\hat{\epsilon}_i, \hat{\beta}_0) =$ Cov$(\hat{\epsilon}_i, \hat{\beta}_1) = 0$, which by HW2#4 implies independence of $\hat{\epsilon}_i$ with $\hat{\beta}_0$ and $\hat{\beta}_1$. Therefore, S^2, which is a function of $\hat{\epsilon}_i$ is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Hypothesis Testing for $\hat{\beta}_0$

Because of the above theorem, we can immediately test

$$H_0 : \beta_0 = \beta$$

using statistic

$$t_0 = \frac{\hat{\beta}_0 - \beta}{S\sqrt{\sum x_i^2 / nS_{XX}}} \sim t_{n-2}$$

by independence of $\hat{\beta}_0 \sim N(\cdot, \cdot)$ and $S^2 \sim \chi^2_{n-2}$. The details follow exactly that of the derivation of Student's t-test.

Similarly, hypothesis

$$H_0 : \beta_1 = \beta$$

can be tested with statistic

$$t_1 = \frac{\hat{\beta}_0 - \beta}{S/\sqrt{S_{XX}}} \sim t_{n-2}$$

Generally, although $\hat{\beta}_0$ and $\hat{\beta}_1$ are not independent, they are tested separately against these marginal distributions. Together (t_0, t_1) follows a so-called bivariate t-distribution, but this distribution is not often used for inference.

Partitioning Total Variance

The total variability in the data Y, as measured by sums of squares can be partitioned into a contribution coming from the variation in the regression function itself and variation in the measurements around the regression function. The partitioning equation is

$$\text{total sum of squares} = \text{regression sum of squares} + \text{residual sum of squares}$$

where

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

and can also be shown to be

$$S_{YY} = \frac{S_{XY}^2}{S_{XX}} + \text{RSS}.$$

Coefficient of Determination

Definition: coefficient of determination

The coefficient of determination r^2 is

$$r^2 = \frac{\text{regression sum of squares}}{\text{total sum of squares}} = \frac{S_{XY}^2}{S_{XX}S_{YY}}$$

is a measure of the fraction of total variation in Y that can be explained by the regression function.

Notice that $r^2 \in [0, 1]$. $r^2 = 1$ when there is a perfect linear relationship between x and Y. $r^2 = 0$ when the best fitting regression line has $\hat{\beta}_1 = 0$, i.e. there is no relationship between x and Y, so that Y is independent of level x.

Confidence Intervals

Confidence intervals are trivial to compute. Suppose $t_{n-2,\alpha/2}$ is the critical value for the upper tail of a t-distribution with $n-2$ degrees of freedom. Then, a $100(1-\alpha)\%$ confidence interval for β_1 is

$$\hat{\beta}_1 - t_{n-2,\alpha/2}\frac{S}{\sqrt{S_{XX}}}, \hat{\beta}_1 + t_{n-2,\alpha/2}\frac{S}{\sqrt{S_{XX}}}$$

Estimating Population Mean at Unobserved Level x_0

Suppose we observe $Y_1, \ldots, Y_n$ for levels $x_1, \ldots, x_n$ and we want to predict the expected of Y at previously unobserved x_0. For example, we may wish to predict the expected score in STAT430 given the score in STAT330.

We know $E[Y \mid x_0] = \beta_0 + \beta_1 x_0$ is the population mean at x_0. A reasonable estimate is

$$\hat{\beta}_0 + \hat{\beta}_1 x_0$$

In fact, this estimate is unbiased

$$E[\hat{\beta}_0 + \hat{\beta}_1 x_0] = E[\hat{\beta}_0] + E[\hat{\beta}_1]x_0 = \beta_0 + \beta_1 x_0.$$

Furthermore,

$$\text{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \text{Var}(\hat{\beta}_0) + x_0^2 \text{Var}(\hat{\beta}_1) + 2x_0 \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \cdots = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right)$$

Finally, since $\hat{\beta}_0, \hat{\beta}_1$ are linear estimators, then reproductive property indicates

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N \left(\beta_0 + \beta_1 x_0, \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right] \right)$$

Because S^2 is independent of $(\hat{\beta}_0, \hat{\beta}_1)$, it is also independent of $\hat{\beta}_0 + \hat{\beta}_1 x_0$, so we can proceed to hypothesis testing, e.g. testing

$$H_0 : \beta_0 + \beta_1 x_0 = \beta$$

using

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_0 - \beta}{S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}} \sim t_{n-2}$$

or confidence interval construction

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{n-2,\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}, \hat{\beta}_0 + \hat{\beta}_1 x_0 + t_{n-2,\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}$$

Experimental Design in Estimation of $\beta_0 + \beta_1 x_0$

Notice that the width of the confidence interval depends on $x_1, \ldots, x_n$, i.e. the levels we selected in our experimental design at which to general sample data $Y_1, \ldots, Y_n$. This observation implies that we can modify the experimental design in order to get tighter confidence bounds on our estimate of the population mean

$$\beta_0 + \beta_1 x_0$$

at the untried value x_0.

In particular, we can reduce the variance by increasing our sample size n and by selecting $x_1, \ldots, x_n$ such that their average $\bar{x} = x_0$ or $\bar{x} \approx x_0$.

Predicting Observation at Unobserved Level x_0

Suppose we want to predict an unobserved random variable Y. For example, suppose we want to predict the performance in STAT430 of a student who received 92% in STAT330. Using the above procedure for estimating means, we can predict the mean performance of all STAT430 students with 92% in STAT330, but how do we predict the performance of a single such student?

Definition: prediction interval

A $100(1-\alpha)\%$ *prediction interval* for an unobserved random variable Y based on observed data X is a random interval $[L(X), U(X)]$ with

$$P(L \leq Y \leq U) \geq 1 - \alpha$$

Let us suppose we want to predict observation Y_0 at $x = x_0$ having observed already $(x_1, y_1), \ldots, (x_n, y_n)$. Note that Y_0 is independent of previous data and thus independent of the estimates $\hat{\beta}_0, \hat{\beta}_1, S^2$.

$Y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0$ has a normal distribution by the reproductive property, with

$$\begin{aligned} \mathrm{E}[Y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0] &= \beta_0 + \beta_1 x_0 - \beta_0 - \beta_1 x_0 = 0 \\ \mathrm{Var}(Y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0) &= \mathrm{Var}(Y_0) + \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x) \\ &= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}} \right) \end{aligned}$$

Because S^2 and $Y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0$ are independent, confidence intervals and t-tests can be constructed in the usual fashion. Note, the variance for predicting unobserved Y_0 is bigger than the variance for estimating the mean $\beta_0 + \beta_1 x_0$ at a novel level x_0. There is extra uncertainty in trying to preict a random variable such as Y as compared to a population parameter $\beta_0 + \beta_1 x_0$ that explains this difference.