Here I work through two derivations for solving linear regession:

1. Ordinary least squares.
2. Maximum likelihood.

   - univariate linear regeression.
   - multivariate linear regression.

NOTE: LaTeX doesn't render properly in this notebook on github - see PDF instead.

Least squares derivation of linear regression

$$Lossfunction = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

Define loss function and expand:

$$Q = \sum_{i=1}^{n}(y_i - \hat{y_i})^2, \hat{y_i} = mx_i + b$$

$$Q(m, b) = \sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

$$Q(m, b) = \sum_{i=1}^{n}(y_i - mx_i - b)^2$$

$$Q(m, b) = \sum_{i=1}^{n}(y_i^2 - y_i mx_i - y_i b - y_i mx_i + m^2x_i^2 + mx_i b - y_i b + mx_i b + b^2)$$

$$Q(m, b) = \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i mx_i - \sum_{i=1}^{n} y_i b - \sum_{i=1}^{n} y_i mx_i + \sum_{i=1}^{n} m^2x_i^2 + \sum_{i=1}^{n} mx_i b - \sum_{i=1}^{n} y_i b + \sum_{i=1}^{n} mx_i b + \sum_{i=1}^{n} b^2$$

$$Q(m, b) = \sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} m^2x_i^2 + \sum_{i=1}^{n} b^2 - \sum_{i=1}^{n} 2y_i mx_i - \sum_{i=1}^{n} 2y_i b + \sum_{i=1}^{n} 2mx_i b$$

Differentiate with respect to **m** and **b**:

$$\frac{\partial Q}{\partial m} = 2m \sum x_i^2 - 2 \sum_{i=1}^{n} y_i x_i + 2b \sum_{i=1}^{n} x_i$$

$$\frac{\partial Q}{\partial b} = 2bN - 2 \sum y_i + 2m \sum x_i, N = number\ of\ points$$

We want to minimize m and b, so set both to 0:

$$2m \sum x_i^2 - 2 \sum y_i x_i + 2b \sum x_i = 0$$

$$2bN - 2 \sum y_i + 2m \sum x_i = 0$$

Solve for b first:

$$b = \frac{2 \sum y_i - 2m \sum x_i}{2N}$$

$$b = \frac{\sum y_i - m \sum x_i}{N}$$

Substitute b and solve for m:

$$2m \sum x_i^2 - 2 \sum y_i x_i + 2\frac{\sum y_i - m \sum x_i}{N} \sum x_i = 0$$

$$m \sum x_i^2 - \sum y_i x_i + \frac{\sum y_i - m \sum x_i}{N} \sum x_i = 0$$

$$m \sum x_i^2 - \sum y_i x_i + \frac{\sum y_i \sum x_i - m(\sum x_i)^2}{N} = 0$$

$$mN \sum x_i^2 - N \sum y_i x_i + \sum y_i \sum x_i - m(\sum x_i)^2 = 0$$

$$mN \sum x_i^2 - m(\sum x_i)^2 = N \sum y_i x_i - \sum y_i \sum x_i$$

$$m(N \sum x_i^2 - (\sum x_i)^2) = N \sum y_i x_i - \sum y_i \sum x_i$$

$$m = \frac{N \sum y_i x_i - \sum y_i \sum x_i}{N \sum x_i^2 - (\sum x_i)^2}$$

This solution assumes:

- Residuals are normally distributed.
- Residuals have an equal variance (no heteroscedasticity).
- The means of the residuals is 0

Least squares derivation of linear regression (abridged)

1. Define loss function and expand:

$$Q(m, b) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (mx_i + b))^2$$

...

$$Q(m, b) = \sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} m^2 x_i^2 + \sum_{i=1}^{n} b^2 - \sum_{i=1}^{n} 2y_i m x_i - \sum_{i=1}^{n} 2y_i b + \sum_{i=1}^{n} 2m x_i b$$

2. Differentiate with respect to **m** and **b** and set to 0 because we want to minimize error:

$$\frac{\partial Q}{\partial m} = 0 = 2m \sum x_i^2 - 2 \sum_{i=1}^{n} y_i x_i + 2b \sum_{i=1}^{n} x_i$$

$$\frac{\partial Q}{\partial b} = 0 = 2bN - 2 \sum y_i + 2m \sum x_i$$

$$N = number\ of\ points$$

3. Solve for **m** and **b**:

$$m = \frac{N \sum y_i x_i - \sum y_i \sum x_i}{N \sum x_i^2 - (\sum x_i)^2}$$

$$b = \frac{\sum y_i - m \sum x_i}{N}$$

Maximum likelihood derivation of linear regression

Model to be predicted:
$y = mx + b$

1. PDF for normally-distributed variable: $X \sim \mathcal{N}(\mu, \sigma^2)$
   $PDF(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
2. Treat the linear equation that we need to find as the mean of a normal distribution.
   $P(y|x; m, b, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-(mx+b))^2}{2\sigma^2}}$

   Why? When we draw a line through some points, the distance between the line and each point is a residual. We make three assumptions about the residuals:

   - Residuals are normally distributed.
   - Residuals have an equal variance (no heteroscedasticity).
   - The means of the residuals is 0.

1. Likelihood function for all observed points (x, y) is the product of the probability density for each point:
   $L(m, b, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \prod_{i=1}^{n} e^{\frac{-(y_i-(mx_i+b))^2}{2\sigma^2}}$
2. Goal: find parameters **b**, **m** and σ that **maximize *L*.**

1. Convert to log likelihood for easier math, **maximize *log(L)***:
   $log(L) = log[\frac{1}{\sqrt{2\pi\sigma^2}} \prod_{i=1}^{n} e^{\frac{-(y_i-(mx_i+b))^2}{2\sigma^2}}]$
   ...
   $log(L) = -\frac{1}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (mx_i + b))^2$
2. Or, **minimize** negative log likelihood:
   $-log(L) = \frac{1}{2} log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - (mx_i + b))^2$
3. Let's *imagine* that our variance term is a fixed constant.
   $-log(L) = \sum_{i=1}^{n} (y_i - (mx_i + b))^2$

Maximum likelihood derivation for multi-variate linear regression

Model to be predicted:
$y = \theta_0 + \theta_1 x_1 \; + \theta_1 x_2 \; + \theta_1 x_3 + \ldots \theta_n x_n$
$Y = X_{n\,x\,d} * \theta_{d\,x\,1}$

$y \sim \mathcal{N}(X\theta, \sigma^2)$

Given #1 and #2 above:

1. Likelihood function for all observed points (x, y) is the product of the probability density for each point:

$$L(\theta X, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \prod_{i=1}^{n} e^{\frac{-(y_i-(x_i\theta))^2}{2\sigma^2}}$$

   - Note technically $\frac{1}{\sqrt{2\pi\sigma^2}}$ can be to the n here, but this won't make a difference in the end.

2. Re-write with a sigma, then convert to matrix notation:

$$L(\theta X, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{\sum_{i=1}^{n} -(y_i-(x_i\theta))^2}{2\sigma^2}}$$

$$L(\theta X, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y-X\theta)^T(Y-X\theta)}{2\sigma^2}}$$

$$L(\theta X, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{|Y-X\theta|^2}{2\sigma^2}}$$

3. Maximizing likelihood is equivalent to minimizing sum of errors*. Partially differentiate with respect to theta and set to 0.

$$|Y - X\theta|^2 \; = Y^2 - 2X^T Y\theta + X^T X\theta^2$$
$$\frac{\partial |Y-X\theta|^2}{\partial\theta} = -2X^T Y + 2X^T X\theta$$
$$0 = -2X^T Y + 2X^T X\theta = -X^T Y + X^T X\theta$$
$$\theta = \frac{X^T Y}{X^T X} = (X^T Y)(X^T X)^-1$$

* https://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/section15.pdf (https://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/section15.pdf)