# STAT 332: 05-1 Regression Estimators

Tanvir Deol

Summary:

- Regression estimation of the mean $\hat{\mu}_{reg}$
  - Point estimate
  - Derive mean and variance of estimator
  - Calculate confidence intervals for the mean based on $\hat{\mu}_{reg}$
  - Compare $\hat{\mu}_{reg}$ to other estimators of the mean

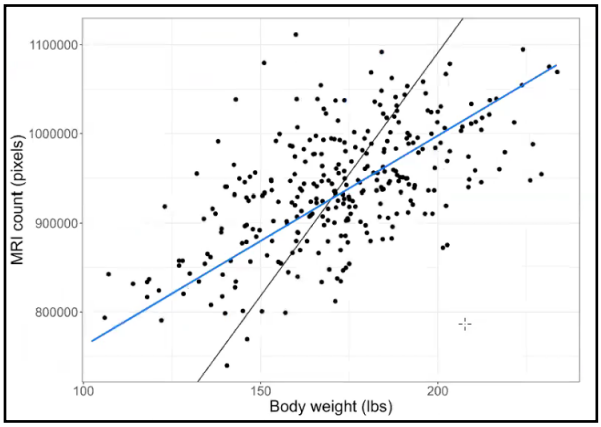
## Regression Estimation

Suppose relationship between $x$ and $y$ in population has a non-zero intercept. Like such:

$$y_i = \alpha + \beta x_i + R_i$$

$R_i$ : "noise" which has a mean of 0

- This is known as the residual
- It is the only random part of the equation
- Everything else is NOT random




This shows the difference between a line with and without an intercept.

- Blue line is the regression line (LOBF) w/ intercept
- Black line is the original ratio line (goes through 0) w/o intercept

## Regression Estimation of the Mean

Since $y_i \approx \alpha + \beta x_i$ we have: (we say approx because we left out the error term)

$$\begin{aligned}\frac{1}{N}\sum_{i\in U} y_i &\approx \frac{1}{N}\sum_{i\in U}[\alpha + \beta x_i] \\ &= \frac{1}{N}[N\alpha + \beta \sum_{i\in U} x_i] \\ &= \alpha + \beta\mu_x\end{aligned}$$

Which gives us that:

$$\mu_y \approx \alpha + \beta\mu_x$$

Here $U$ is universe and $\mu_x = \sum_{i \in U} x_i$ (the population mean for $X$).
Using the information above we can define our **regression estimate** of $\mu_y$:

$$\hat{\mu}_{reg} = \hat{\alpha} + \hat{\beta}\mu_x$$

Ordinary Least Squares (LS) allows us to estimate $\alpha$ and $\beta$:

$$\hat{\beta} = \frac{\sum_{i \in S}(x_i - \bar{x})y_i}{\sum_{i \in S}(x_i - \bar{x})^2}$$

*This is the estimated slope*

$$\hat{\alpha} = \bar{y} - \hat{b}\bar{x}$$

*This is the estimated intercept*

If we substitute $\hat{\alpha}$ in the $\hat{\mu}_{reg}$ expression then we can rewrite the regression estimate as

$$\begin{aligned}\hat{\mu}_{reg} &= \hat{\alpha} + \hat{\beta}\mu_x \\ &= (\bar{y} - \hat{\beta}\bar{x}) + \hat{\beta}\mu_x \\ &= (\hat{\mu}_y - \hat{\beta}\hat{\mu}_x) + \hat{\beta}\mu_x \\ &= \hat{\mu}_y + \hat{\beta}(\mu_x - \hat{\mu}_x)\end{aligned}$$

The $\hat{\beta}(\mu_x - \hat{\mu}_x)$ part of the equation is called the **additive corrective factor for** $\hat{\mu}_y = \bar{y}$

**Adjustment term/Additive Corrective Factor $\hat{\beta}[\mu_x - \hat{\mu}_x]$**

If $\hat{\beta} > 0$ (i.e larger values of x correspond to larger values of y) (Corr(x,y) $> 0$) then
- if $\mu_x > \hat{\mu}_x$ then the sample average for the y variate $\hat{\mu}_y$ is adjusted upward
- basically thats a positive correction
- if the slope is negative, then the correction happens in negative direction

Note:
The regression estimate "drags" $\bar{y}$ (sample mean of y aka $\hat{\mu}_y$) towards $\mu_y$ (the population mean)
Why is this important?
- Because our sample can be biased based on our sampling protocol
- $\bar{y}$ could be a lot higher or a lot lower than the "true" population mean
- that is where the correction factor saves the day

**Properties**

The estimator for the regression estimate is $\tilde{\mu}_{reg} = \tilde{\mu}_Y + \tilde{\beta}[\mu_X - \tilde{\mu}_X]$. We want to show this estimator is unbiased. By considering when n (sample size) gets very large and then taking the expectation of the estimator.

$$\tilde{\mu}_{reg} = \tilde{\mu}_Y + \tilde{\beta}[\mu_X - \tilde{\mu}_X]$$
$$\tilde{\mu}_{reg} - \mu_Y = \tilde{\mu}_Y - \mu_Y + [\tilde{\beta} - \beta + \beta][\mu_X - \tilde{\mu}_X]$$
$$\tilde{\mu}_{reg} - \mu_Y = \tilde{\mu}_Y - \mu_Y + \beta[\mu_X - \tilde{\mu}_X] + \underbrace{[\tilde{\beta} - \beta]}_{\text{A}} \underbrace{[\mu_X - \tilde{\mu}_X]}_{\text{B}}$$

Pay attention to A and B, as n becomes very large those parts go to 0.
So what we actually get for large n is:

$$\tilde{\mu}_{reg} - \mu_Y \approx [\tilde{\mu}_Y - \mu_Y] + \beta[\mu_X - \tilde{\mu}_X]$$

## Expectation

Now we take the expectation of the estimator to prove that it is unbiased.
Recall from SRSWOR that $E[\tilde{\mu}_Y] = \mu_Y$ and $E[\tilde{\mu}_X] = \mu_X$.

$$\begin{aligned} E[\tilde{\mu}_{reg} - \mu_Y] &\approx E\{[\tilde{\mu}_Y - \mu_Y] + \beta[\mu_X - \tilde{\mu}_X]\} \\ &= E[\tilde{\mu}_Y - \mu_Y] + \beta E[\mu_X - \tilde{\mu}_X] \\ &= 0 \end{aligned}$$

Note that $\mu_Y, \mu_X$ are constants so their expectation is themselves.
Given the above expectation, we can prove $\tilde{\mu}_{reg}$ is unbiased since:

$$E[\tilde{\mu}_{reg} - \mu_Y] \approx 0 \Rightarrow E[\tilde{\mu}_{reg}] - \mu_Y \approx 0 \Rightarrow E[\tilde{\mu}_{reg}] \approx \mu_Y$$

Therefore we've proven the estimator is approx unbiased under SRSWOR.

## Variance

Now we establish variance. Recall that $\tilde{\mu}_{reg} - \mu_Y \approx [\tilde{\mu}_Y - \mu_Y] + \beta[\mu_X - \tilde{\mu}_X]$ for large n. So we get:

$$\begin{aligned} Var(\tilde{\mu}_{reg}) &= Var(\tilde{\mu}_{reg} - \mu_Y) \\ &\approx Var(\tilde{\mu}_{reg} - \mu_Y + \beta[\mu_X - \tilde{\mu}_X]) \\ &= Var(\tilde{\mu}_e) \end{aligned}$$

Where $e_i$ is defined as $e_i = y_i - \mu_Y + \beta(\mu_X - x_i)$ and $\mu_e$ is defined as
$\mu_e = \frac{1}{N}\sum_{i \in U} e_i = \mu_Y - \mu_Y + \beta(\mu_X - \mu_X) = 0.$

## Sample Variance

By applying the results from SRSWOR to our new variable $e$ we get:

$$E[\tilde{\mu}_e] = \mu_e = 0$$

$$Var(\tilde{\mu}_e) = (1 - \frac{n}{N})\frac{\sigma_e^2}{n}$$

Where $\sigma_e^2$ s:

$$
\begin{aligned}
\sigma_e^2 &= \frac{1}{N-1} \sum_{i \in U} \underbrace{(e_i - \mu_e)^2}_{\mu_e = 0} \\
&= \frac{1}{N-1} \sum_{i \in U} e_i^2 \\
&= \frac{1}{N-1} \sum_{i \in U} [y_i - \mu_Y + \beta(\mu_X - x_i)]^2
\end{aligned}
$$

That was the **population variance** of the residuals $e_i$. Now we will get the **sample variance** of the estimated residuals (from a least squares fit of the sample data):

$$
\begin{aligned}
\hat{\sigma}_e^2 &= \frac{1}{n-1} \sum_{i \in s} [y_i - \bar{y} + \hat{\beta}(\bar{x} - x_i)]^2 \\
&= \frac{1}{n-1} \sum_{i \in s} [y_i - \bar{y} - \hat{\beta}(x_i - \bar{x})]^2 \\
&= \frac{1}{n-1} \sum_{i \in s} [y_i + (-\bar{y} + \hat{\beta}\bar{x}) - \hat{\beta} x_i]^2 \\
&= \frac{1}{n-1} \sum_{i \in s} [y_i - \hat{\alpha} - \hat{\beta} x_i]^2
\end{aligned}
$$