Introduction to Econometrics
PS #4 - Introduction to Simple Regression Model

1 Population and sample regression function

Suppose you are interested in the relationship between class attendance (X) and test scores (Y) for your cohort.

1. Assume that you have data for the entire class (population). How would you best describe the average relationship denoted by $E(Y_i|X_i)$?
2. The $E(Y_i|X_i)$ (also called the CEF) does not fit all the data points perfectly. Write the equation for Y_i that accounts for this error.
3. Getting data on the entire population is often costly, suppose you draw a random sample instead. What is the best way to describe the relationship from your sample?

2 Derivation of OLS estimators

The true relationship in the population is: $Y_i = \beta_0 + \beta_1 X_i + u_i$. However, with access to only a random sample, we need the estimates $\hat{\beta_0}$ and $\hat{\beta_1}$.

1. Derive the OLS estimators by minimizing the sum of squared distances between each sample data point and the SRF.
2. Show that $\hat{\beta_1} = \dfrac{cov(X_i, Y_i)}{V(X_i)}$. Explain the intuition behind this estimator.

3 Re-scaling variables

Your boss asks you to estimate the relationship between the factory's output (Y_i) and the distance of the factory from the headquarters (D_i). Assume that the data on distance (D_i) is originally measured in Kilometers. However, your boss wants you to estimate this relationship using distance measured in Meters.

1. Specify the original regression model. Report the OLS estimators $\hat{\beta_0}$ and $\hat{\beta_1}$.
2. Specify the regression model of interest by scaling D_i appropriately. [Note: 1km = 1000m]
3. Find the OLS estimators $\tilde{\beta_0}$ and $\tilde{\beta_1}$.
4. Your boss is highly indecisive - he now asks for the relationship with distance measured in Kilometers. Find the resulting OLS estimate $\hat{\beta_1}$.

Suppose that you multiplied both the output (Y_i) and the distance (D_i) (measured in Kilometers) by 1000

5. Specify the regression model by scaling both variables.
6. Find the OLS estimators, are the results equivalent to $\mathcal{Q}3.1$?

4 Slight Detour: Law of Iterated Expectations

Assume that X and Y are two jointly distributed discrete random variables. Prove that $E(Y) = E\Big\{E(Y|X)\Big\}$

[**Note**: We have already seen LIE in action, in Problem Set #2 $\mathcal{Q}.3$]

5 OLS properties

Suppose that we are interested in the relationship between wages and the years of education for France. The simple regression model for the same is specified as:

$$w_i = \beta_0 + \beta_1 educ_i + u_i \tag{1}$$

1. Is the model linear in parameters?
2. Suppose that our sample is only drawn from the players of Paris Saint-Germain F.C. Would you consider this as a random sample in the context of the entire French labor market?
3. Assuming that $V(educ_i) = 0$, find the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. Do the estimators give any information on the relationship of interest?
4. Show that

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(educ_i - \overline{educ})u_i}{\sum_{i=1}^{n}(educ_i - \overline{educ})^2}$$

   Use the ZCM assumption *i.e.*, $E(u_i|educ_i) = 0$ to show that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators.
5. The errors of the regression are said to be homoskedastic when $V(u_i|educ_i) = \sigma^2$. From Figure 1 below, what can you conclude about the variance of u_i as e_i increases?

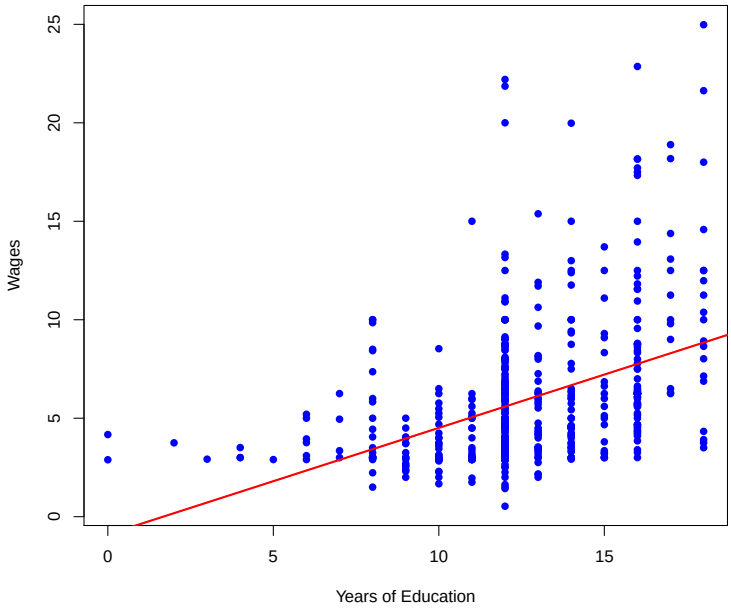


Figure 1: Scatterplot of wages and education + estimated regression line

(a) What are the consequences of heteroskedasticity?

(b) Suppose that the we know the functional form for heteroskedasticity: $V(u_i|educ_i) = \lambda g(educ_i)$. Use this specification to reach homoskedastic errors.

6 Measure of Fit and Prediction Accuracy

Assume the following univariate regression model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{2}$$

6.1 The R^2 and Standard Error of the Regression (SER)

1. Prove that $TSS = ESS + RSS$. Find the formula for R^2.
2. When is $R^2 = 0$, and $R^2 = 1$?
3. Report the formula for the standard error of the regression in equation (2).

6.2 The Standard Error of $\hat{\beta}_1$

1. Write $\hat{\beta}_1$ as a function of of X_i and u_i
2. Assuming homoskedasticity *i.e.*, $V(u_i|X_i) = \sigma^2$, show that $V(\hat{\beta}_1) = \sigma^2 / \sum(X_i - \bar{X})^2$
3. Report the sample estimator for σ^2, use this to find the standard error of $\hat{\beta}_1$, denoted as $SE(\hat{\beta}_1)$

7 Regression Analysis

Suppose that a researcher, using wage data on 250 randomly selected male workers and 280 female workers, obtains the following results of an OLS regression:

$$\widehat{\text{Wage}} = 12.68 + 2.79\,\text{Male}, \quad R^2 = 0.06$$
$$(0.18) \quad (0.84)$$

Where the dummy variable for Male is defined as:

$$\text{Male} = \begin{cases} 1, & \text{Male} \\ 0, & \text{Female} \end{cases}$$

1. What does the estimate of the intercept represent in this regression?
2. What is the estimated gender pay gap?
3. Test whether the estimated gender gap is significantly different than 0, with $\alpha = 0.05$
   (a) Use the rejection region approach (**Note**: t-distrib. converges to the Z distrib. for large samples)
   (b) Use the p-value approach, is the conclusion same as above?
4. Another researcher uses the same data, but instead decides to regress wages on a dummy for females

$$\text{Female} = \begin{cases} 1, & \text{Female} \\ 0, & \text{Male} \end{cases}$$

   Find the new regression estimates. Do you expect the R^2 to change?

8 R Exercise: Univariate Regression

This exercise is based on the 1976 Current Population Survey (U.S.A.). The dataset contains 526 observations data on 24 variables including wage, education, female (dummy) etc.

1. Import the dataset in RStudio.

```
# Import Data
data <- read.csv("Path to wage1.csv")
```

2. Regress wages on the female dummy variable. Report and interpret the regression results.

```
# Model
reg1 <- lm(wage ~ female, data = data)

# Print results
summary(reg1)
```

Table 1: Regression of Wages on Female

	Dependent variable:
	Wage
Female	-2.512^{***}
	(0.303)
Constant	7.099^{***}
	(0.210)
Observations	526
R^2	0.116
Adjusted R^2	0.114
Residual Std. Error	3.476 (df = 524)
F Statistic	68.537^{***} (df = 1; 524)
Note:	*p<0.1; **p<0.05; ***p<0.01

3. Plot wages as a function of the Female dummy variable, add the estimated regression line.

```
# Plot the wage data against female
plot(data$female, data$wage,
     xlab = "Female Dummy (0 = Male, 1 = Female)",
     ylab = "Wage",
     pch = 19, col = "blue")

# Add the regression line to the plot
abline(reg1, col = "red", lwd = 2)

# Optional: Add legend
legend("topright", legend = c("Data points", "Regression line"), col = c("blue",
    "red"), pch = 19, lwd = 2)
```

4. Regress wages on education. Report and interpret the coefficients.

```
# Model
reg2 <- lm(wage ~ educ, data = data)

# Print results
summary(reg2)
```

Table 2: Regression of Wages on Education

	Dependent variable:
	Wage
Education	0.541^{***}
	(0.053)
Constant	-0.905
	(0.685)
Observations	526
R^2	0.165
Adjusted R^2	0.163
Residual Std. Error	3.378 (df = 524)
F Statistic	103.363^{***} (df = 1; 524)
Note:	*p<0.1; **p<0.05; ***p<0.01

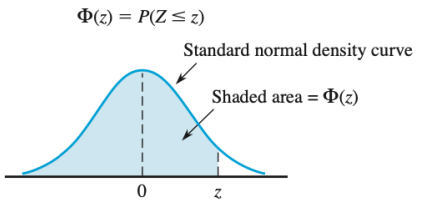
5. Plot wages as a function of education, add the estimated regression line.

```
# Plot the wage data against education
plot(data$educ, data$wage,
     xlab = "Education",
     ylab = "Wage",
     pch = 19, col = "blue")

# Add the regression line to the plot
abline(reg2, col = "red", lwd = 2)

# Optional: Add legend
legend("topright", legend = c("Data points", "Regression line"), col = c("blue",
    "red"), pch = 19, lwd = 2)
```

Table A.3 Standard Normal Curve Areas



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0038
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3482
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

(*continued*)

Table A.3 Standard Normal Curve Areas *(cont.)* $\Phi(z) = P(Z \leq z)$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Figure 2: Z Distribution Table