

Introduction to Linear Regression

22S:138 Bayesian Statistics

Lecture 14
October 12, 2005

Kate Cowles, Ph.D.

Review of Frequentist Approach to Linear Regression Per Capita Health Spending and Per Capita Gross Domestic Product (GDP) in 24 OECD Countries, 1989

Schieber, Poullier, and Greenwald, Health Affairs, 1991

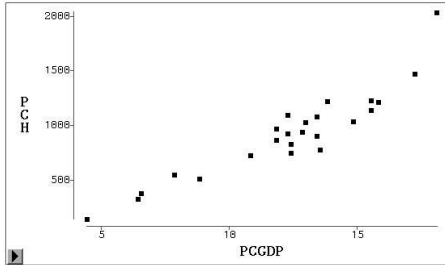
	Country	Per Cap Hlth	Per Cap GDP
1.	united states	2051	18.1429
2.	canada	1483	17.2857
3.	iceland	1241	15.5714
4.	sweden	1233	13.8571
5.	switzerland	1225	15.8571
6.	norway	1149	15.5714
7.	france	1105	12.2857
8.	germany	1093	13.4286
9.	luxemborg	1050	14.8571
10.	netherlands	1041	13.0000
11.	austria	982	11.8571
12.	finland	949	12.8571
13.	australia	939	12.2857
14.	japan	915	13.4286
15.	belgium	879	11.8571
16.	italy	841	12.4286
17.	denmark	792	13.5714
18.	united kingdom	758	12.4286
19.	new zealand	733	10.8571
20.	ireland	561	7.8571
21.	spain	521	8.8571
22.	portugal	386	6.5714
23.	greece	337	6.4286

24. turkey 148 4.4286

In regression analysis, we look at the conditional distribution of the **response variable** at different levels of a **predictor variable**

- Response variable
 - also called “dependent” or “outcome” variable
 - what we want to explain or predict
 - in simple linear regression, response variable is continuous
- Predictor variables
 - also called “independent” variables or “covariates”
 - in simple linear regression, predictor variable usually is also continuous
 - How we define which variable is response and which is predictor depends on our research question.

Per Capita Health Spending and Per Capita Gross Domestic Product (GDP) In 24 OECD Countries, 1989



Scatterplots

- response variable on Y axis
- predictor variable on X axis
- Relationship in this scatterplot looks roughly linear.
 - Makes sense to try to summarize the relationship between these two variables with a straight line.

Quick review of linear functions

$$Y = \beta_0 + \beta_1 X$$

- Y is a response variable that is a linear function of the predictor variable X
- β_0 : intercept; the value of Y when $X = 0$
- β_1 : slope; how much Y changes when X increases by 1 unit

Linear regression

- In linear regression analysis, $\beta_0 + \beta_1 X$ represents the **mean value** of all the Y's for a given value of X.

$$E(Y|X) = \beta_0 + \beta_1 X$$

- There is an entire distribution of Y values for each value of X (a conditional distribution)
 - Example: for any given value of per capita GDP, there is a distribution of values of per capita health spending among OECD countries
- We say the relationship between X and Y is **linear** if the means of the conditional distributions of $Y|X$ lie on a straight line.

Error terms

- In regression, we represent factors other than X_i that affect Y_i with an **error term**, ϵ_i .

- population model

$$\epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

$$\epsilon_i = Y_i - E[Y_i]$$

- or, equivalently,

$$Y_i = (\beta_0 + \beta_1 X_i) + \epsilon_i$$

$$Y_i = E[Y_i] + \epsilon_i$$

Notation

- population model

$$E[Y_i] = \beta_0 + \beta_1 X_i$$

- sample estimates

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

– \hat{Y}_i is the **fitted value** or **predicted value** of Y for case i

- residuals (e_i)

– sample estimates of error terms ϵ_i

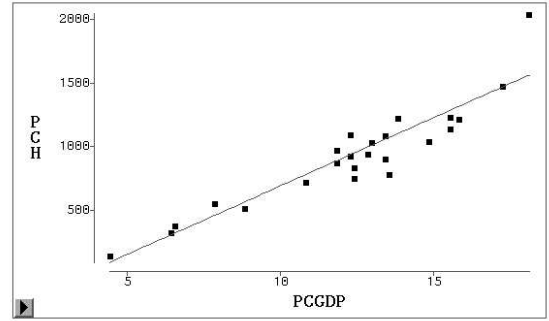
$$e_i = Y_i - \hat{Y}_i$$

- error sum of squares (SSE)

$$SSE = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

- OLS chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ so as to minimize SSE.
- n = number of observations in the dataset
- K = number of β s in the model (2 for SLR)

Determining the “best-fitting” line



Ordinary least squares method (OLS)

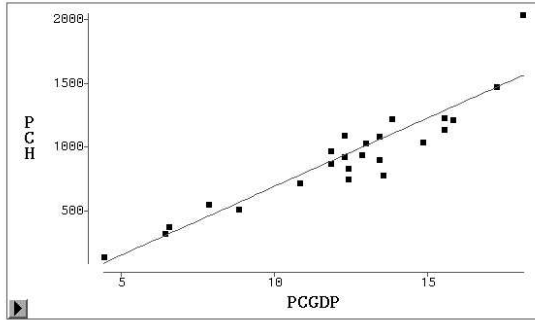
- computes the maximum likelihood estimates of the intercept and slope
- chooses the best-fitting line by minimizing the sum of the squared differences between each data point and the fitted line.

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

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

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	-387.163184	119.72572120	-3.234	0.0038
PCGDP	1	107.263249	9.37953624	11.436	0.0001

Calculating predicted values and residuals
Per capita health expenditures and per capita GDP

Obs	NAME	Dep Var PCH	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean
1	UnitedStates	2051.0	1558.9	63.075	1428.1	1689.7
2	Canada	1483.0	1467.0	56.251	1350.3	1583.6
3	Iceland	1241.0	1283.1	43.858	1192.1	1374.0
4	Sweden	1233.0	1099.2	34.641	1027.4	1171.0
5	Switzerland	1225.0	1313.7	45.765	1218.8	1408.6
6	Norway	1149.0	1283.1	43.858	1192.1	1374.0
7	France	1105.0	930.6	31.480	865.4	995.9
8	Germany	1093.0	1053.2	33.165	984.5	1122.0
9	Luxembourg	1050.0	1206.5	39.487	1124.6	1288.3
10	Netherlands	1041.0	1007.3	32.127	940.6	1073.9
11	Austria	982.0	884.7	31.771	818.8	950.6
12	Finland	949.0	991.9	31.886	925.8	1058.1
13	Australia	939.0	930.6	31.480	865.4	995.9
14	Japan	915.0	1053.2	33.165	984.5	1122.0
15	Belgium	879.0	884.7	31.771	818.8	950.6
16	Italy	841.0	946.0	31.497	880.6	1011.3
17	Denmark	792.0	1068.5	33.611	998.8	1138.3
18	UnitedKingdom	758.0	946.0	31.497	880.6	1011.3
19	NewZealand	733.0	777.4	34.322	706.2	848.6
20	Ireland	561.0	455.6	52.341	347.1	564.2
21	Spain	521.0	562.9	45.201	469.1	656.6
22	Portugal	386.0	317.7	62.399	188.3	447.1
23	Greece	337.0	302.4	63.559	170.6	434.2
24	Turkey	148.0	87.8628	80.394	-78.8646	254.6

Obs	NAME	Residual
1	UnitedStates	492.1
2	Canada	16.0428
3	Iceland	-42.0758
4	Sweden	133.8
5	Switzerland	-88.7209
6	Norway	-134.1
7	France	174.4
8	Germany	39.7679
9	Luxembourg	-156.5
10	Netherlands	33.7410
11	Austria	97.3321
12	Finland	-42.9311
13	Australia	8.3591
14	Japan	-138.2
15	Belgium	-5.6679
16	Italy	-105.0
17	Denmark	-276.5
18	UnitedKingdom	-188.0
19	NewZealand	-44.4046
20	Ireland	105.4
21	Spain	-41.8781
22	Portugal	68.2935
23	Greece	34.6107
24	Turkey	60.1372

Estimating the common variance

- One of the assumptions of linear regression is that the variance for each of the conditional distributions of $Y|X$ is the same at all values of X .

- The estimate of this common variance is

$$\hat{\sigma}^2 = \frac{SSE}{n - 2}$$

- analogous to estimate of variance in a normal sample
- $n - 2$ in denominator is “degrees of freedom”
 - number of observations minus number of estimated regression coefficients

Inferences for the Slope

- So far, we've been *describing* the relationship between two continuous variables.
- Now we want to perform a hypothesis test to determine whether there is a linear relationship between the two variables.
 - depends on assumptions of linear regression
- Question: Does the value of Y depend linearly on X?

$$E[Y_i] = \beta_0 + \beta_1 X_i$$

- Answer: Yes, unless $\beta_1 = 0$, in which case

$$E[Y_i] = \beta_0$$

- Hypotheses for test for linear relationship between Y and X

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

- Confidence interval for the slope:
 - A $(1 - \alpha)\%$ confidence interval for the true slope β_1 is given by:

$$\hat{\beta}_1 \pm (t_{1-(\alpha/2), df=n-2})(\hat{\sigma}_{\beta_1})$$
 - If this C.I. includes the value 0, we cannot reject the null hypothesis at significance level α .

- Test statistic

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}_{\beta_1}}$$

$$\text{where } \hat{\sigma}_{\beta_1} = \frac{\hat{\sigma}}{\sqrt{\sum(X_i - \bar{X})^2}}$$

- standard form of test statistic: estimate divided by its standard error
- standard error of $\hat{\beta}_1$ depends on
 - * variability of Ys
 - * how closely clustered the Xs are
- follows a t distribution with $n - 2$ degrees of freedom
 - * because we have to estimate 2 parameters (β_0 and β_1) to compute $\hat{\sigma}$
- p-value: the probability of obtaining a t statistic as extreme as, or more extreme than, what we got, if H_0 is true

Interpreting the test for zero slope

- Failure to reject $H_0 : \beta_1 = 0$
 - Type II error
 - X and Y related in a nonlinear way
 - X provides little help in predicting Y
- Rejecting $H_0 : \beta_1 = 0$
 - X provides significant information for predicting Y
 - Although the data fit a linear model, some nonlinear model may do even better
- Important caveat regarding inferences on β_1 : the best straight line may be terrible!

Inferences concerning the regression line

- Estimating the mean of the Y's for a particular value of X, say X_0
 - Example: what is the average per capita health spending for a country with per capita gross domestic product 10 PPP

$$E[Y|X_0] = \hat{Y}_{X_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

- estimated standard error of $E[Y|X_0]$

$$\hat{\sigma}_{\hat{Y}_{X_0}} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$$

- A $(1 - \alpha)\%$ confidence interval for $E[Y|X_0]$ is given by:

$$\hat{Y}_{X_0} \pm (t_{1-\alpha/2, df=n-2})(\hat{\sigma}_{\hat{Y}_{X_0}})$$

Obs	name	95% CL Mean		Residual
1	UnitedStates	1428	1690	492.0968
2	Canada	1350	1584	16.0428
3	Iceland	1192	1374	-42.0758
4	Sweden	1027	1171	133.8056
5	Switzerland	1219	1409	-88.7209
6	Norway	1192	1374	-134.0758
7	France	865.3552	995.9267	174.3591
8	Germany	984.4517	1122	39.7679
9	Luxemborg	1125	1288	-156.4576
10	Netherland	940.6317	1074	33.7410
11	Austria	818.7786	950.5571	97.3321
12	Finland	925.8032	1058	-42.9311
13	Australia	865.3552	995.9267	8.3591
14	Japan	984.4517	1122	-138.2321
15	Belgium	818.7786	950.5571	-5.6679
16	Italy	880.6486	1011	-104.9688
17	Denmark	998.8446	1138	-276.5493
18	UnitedKingdom	880.6486	1011	-187.9688
19	NewZealand	706.2243	848.5850	-44.4046
20	Ireland	347.0657	564.1641	105.3851
21	Spain	469.1369	656.6193	-41.8781
22	Portugal	188.2993	447.1138	68.2935
23	Greece	170.5763	434.2024	34.6107
24	Turkey	-78.8646	254.5903	60.1372

Prediction interval for a new individual's Y given that we know their value of X, say X_0

- Point estimate is the group mean

$$\hat{Y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

- Would you expect an individual's response to be more or less variable than the group's mean response?
- two sources of variability in prediction of individual Y
 - uncertainty of the group mean $E[Y|X_0]$
 - variability of individual responses around the group mean
 - * Example: The Netherlands has per capita PCGDP of 13, but is unlikely to have PCH of exactly the mean of all possible countries with the same GDP.

- estimated standard error of \hat{Y}_{new}

$$\hat{\sigma}_{\hat{Y}_{new}} = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$$

- A $(1 - \alpha)\%$ prediction interval for Y_{new} is given by:

$$\hat{Y}_{new} \pm (t_{1-\alpha/2, df=n-2})(\hat{\sigma}_{\hat{Y}_{new}})$$

Output Statistics

Obs	name	95% CL Predict	Residual
1	UnitedStates	1213	492.0968
2	Canada	1127	16.0428
3	Iceland	950.5722	1616
4	Sweden	771.4038	1427
5	Switzerland	980.1141	1647
6	Norway	950.5722	1616
7	France	604.2244	1257
8	Germany	726.0987	1380
9	Luxemborg	876.3187	1537
10	Netherland	680.5716	1334
11	Austria	558.1302	1211
12	Finland	665.3452	1319
13	Australia	604.2244	1257
14	Japan	726.0987	1380
15	Belgium	558.1302	1211
16	Italy	619.5454	1272
17	Denmark	741.2203	1396
18	UnitedKingdom	619.5454	1272
19	NewZealand	449.7583	1105
20	Ireland	117.8747	793.3551
21	Spain	229.6021	896.1541
22	Portugal	-27.3032	662.7163
23	Greece	-43.5300	648.3086
24	Turkey	-272.8081	448.5338

Plot showing 95% confidence limits and 95% prediction limits

