

## Lab02: Data Transformations and Bivariate Regression Analysis

**Handed out:** Thursday, September 17, 2020

**Return date:** Thursday, October 1, 2020, by midnight at eLearning's **SubmitLab02**.

**Grading:** This lab counts 8 % towards your final grade

**Objectives:** This lab focuses on a review of basic bivariate regression analysis, confidence intervals, the Box-Cox transformation and bivariate regression to estimate the elasticity.

### Task 1: Confidence Intervals (2 points)

Open the **CONCORD1.SAV** file for this task as data-frame. To simplify things do not perform variable transformations.

**Task 1.1:** Run a bivariate regression model of **income** (dependent variable) on **education** (independent variable) and statistically interpret the model estimates for the intercept and slope as well as the  $R^2$ . (0.5 points)

```
library(foreign)
```

```
Concord <- read.spss('Concord1.sav', to.data.frame=TRUE)
```

```
summary(lm(income~educat, data=Concord))
```

```
lm(formula = income ~ educat, data = Concord)
```

**Residuals:**

Min	1Q	Median	3Q	Max
-26.848	-8.182	-0.997	6.471	74.003

**Coefficients:**

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.5893	2.5574	1.012	0.312
educat	1.4630	0.1783	8.203	2.04e-15 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.26 on 494 degrees of freedom

Multiple R-squared: 0.1199, Adjusted R-squared: 0.1181

F-statistic: 67.29 on 1 and 494 DF, p-value: 2.04e-15

Comment: The  $R^2$  is only 0.1199 which means only about 11.9% of the variation in the dependent variable **Income** can be explained by the independent variable **education**. The statistically significant slope ( $H_0: \beta_1 = 0$ ) is positive meaning with each additional year of education the income will increase the income by \$1.463. While the intercept is not statistically different from zero, it should be kept in

the model because [a] there is no logical reason why a person without education may have zero income, and [b] because otherwise some statistics of the OLS model, such as the  $R^2$  lose their properties.

**Task 1.2:** Calculate the 99 % confidence intervals around the estimated regression parameters. Can you draw the same conclusions as you did using the  $t$ -test in the **summary** output from task 1.1? (0.5 points)

```
reg <- lm(income~educat, data=Concord)
cbind("Coef"=coef(reg), confint(reg, level=0.99))
```

	Coef	0.5 %	99.5 %
(Intercept)	2.589322	-4.023659	9.202304
educat	1.462957	1.001811	1.924102

Comments: The  $t$ -test investigates the null hypotheses that the estimated regression parameters are zero. That is,  $H_0: \beta_0 = 0$  for the intercept and  $H_0: \beta_1 = 0$  for the slope. As long as the  $1 - \alpha$  confidence intervals cover the values under the null hypothesis, that is  $\beta_0 = 0$  and  $\beta_1 = 0$ , the null hypothesis cannot be rejected with an error probability of  $\alpha$ .

*Intercept:* 0 is inside the confidence interval so we fail to reject the null hypothesis.  $P$  value for the  $t$ -test in task 1.1 is larger than  $\alpha = 0.01$  so we fail to reject the null hypothesis. Both methods lead to the identical conclusions: intercept is not different from 0.

*Slope:* 0 is outside the confidence so we can reject the null hypothesis.  $P$  value for the  $t$ -test in task 1.1 is smaller than  $\alpha = 0.01$  so we can reject the null hypothesis. Both methods lead to identical conclusions that income is significantly influenced by education.

**Task 1.3:** Scatterplot both variables and add the predicted regression line as well as the lower and upper 90% confidence interval lines around the **point predictions**. (as known as prediction interval in Hamilton and **interval="prediction"** parameter in the **predict** function).

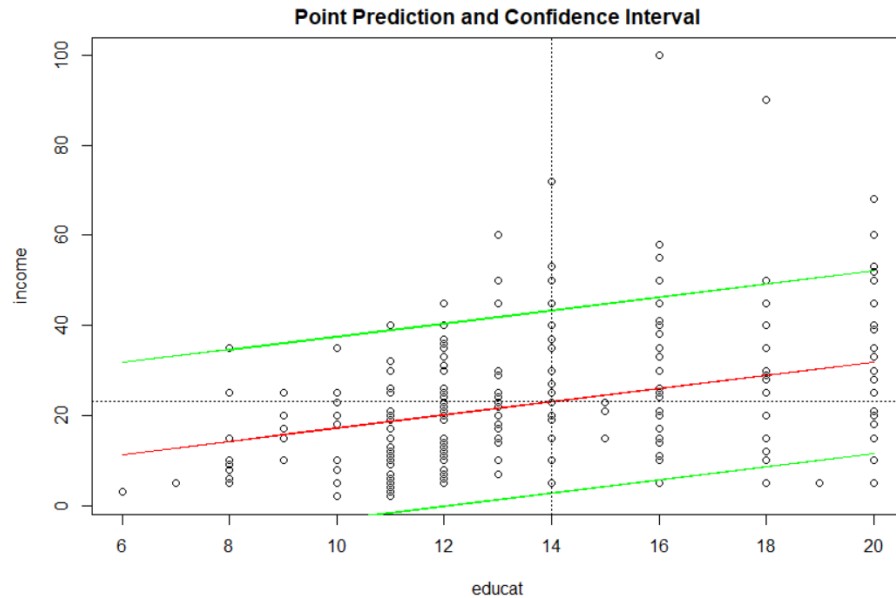
You should enhance your plot by adding lines for the means of education and income as well as adding a title. (1 point)

```
regPred <- predict(reg, interval="prediction", level = 0.90)
ConcordPred <- data.frame(Concord, regPred)

plot(income~educat, data=ConcordPred, main = "Point Prediction and Confidence Interval")

lines(ConcordPred$educat, ConcordPred$fit, col="red")
lines(ConcordPred$educat, ConcordPred$lwr, col="green")
lines(ConcordPred$educat, ConcordPred$upr, col="green")

abline(h=mean(ConcordPred$income), lty=3)
abline(v=mean(ConcordPred$educat), lty=3)
```



## Task 2. Univariate Variable Exploration and Transformations (3 points)

Use the **CPS1985** dataset in the library **AER** to explore the distribution of the respondents' **wage**.

Use the syntax `data(CPS1985, package="AER")` to read the dataframe.

**Task 2.1:** Find the best  $\lambda^{best}$ -value (see `summary(car::powerTransform(lm(varName~1)))`) for the Box-Cox transformation. Interpret the test whether the *log*-transformation (i.e.,  $\lambda = 0.0$ ) instead of  $\lambda^{best}$  could be used? Justify your answer. (1 point)

```
library(AER); library(car); library(e1071)
data(CPS1985)
summary(powerTransform(CPS1985$wage~1))
bcPower Transformation to Normality
Est Power Rounded Pwr Wald Lwr Bnd Wald Up Bnd
CPS1985$wage -0.0658 0 -0.1997 0.068
Likelihood ratio test that transformation parameter is equal to 0
(log transformation)
```

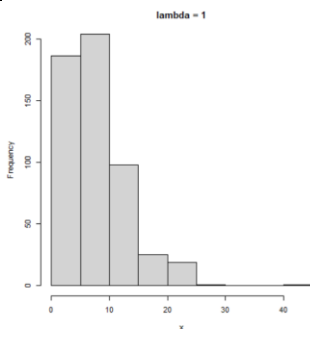
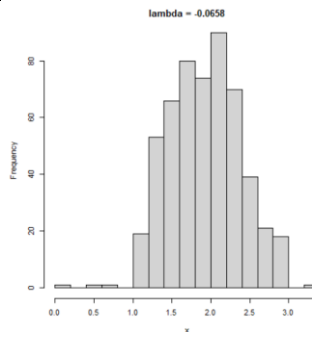
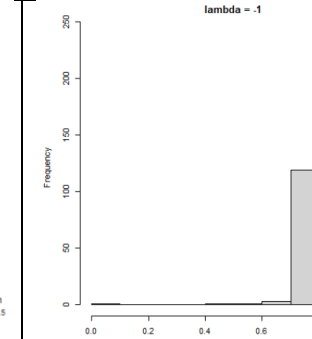
```
LRT df pval
LR test, lambda = (0) 0.9245 1 0.33628
Likelihood ratio test that no transformation is needed
LRT df pval
LR test, lambda = (1) 232.5055 1 < 2.22e-16
```

Comment: The estimated transformation power ( $\lambda = -0.0658$ ), which is very close to 0. The *p*-value of the likelihood ratio test ( $H_0: \lambda = 0$ ) is larger than 0.05, so we fail to reject the null hypothesis. Instead of using ( $\lambda = -0.0658$ ), the *log*-transformation should be used in this case.

**Task 2.2:** For the untransformed ( $\lambda = 1$ ), optimal ( $\lambda = \lambda^{best}$ ) and over-adjusted ( $\lambda = -1$ ) Box-Cox transformed **wage** variable repeat the following tasks and **comparatively interpret** the results. (2 points)

[a] Draw properly constructed histograms including a kernel density curve of all three transformed variables **wage**,

[b] evaluate the skewness (see `e1071::skewness( )`), and test whether the variables are approximately normal distributed (see `shapiro.test( )`).

	$\lambda = 1$	$\lambda^{best} = -0.0658$	$\lambda = -1$
<b>Histograms</b>			
<b>Skewness</b>	1.687762	0.001027028	-2.793214
<b>Shapiro test</b>	<pre>&gt; shapiro.test(wage) Shapiro-Wilk normality test data: wage W = 0.8673, p-value &lt; 2.2e-16</pre>	<pre>&gt; shapiro.test(wage.bc) Shapiro-Wilk normality test data: wage.bc W = 0.98923, p-value = 0.000586</pre>	<pre>&gt; shapiro.test(wage.bc.over) Shapiro-Wilk normality test data: wage.bc.over W = 0.83125, p-value &lt; 2.2e- 16</pre>

[c] Address the questions: Which transformed variable comes closest to the normal distribution? Is the transformation with  $\lambda = -1$  over-compensating the inherent positive skewness in the **wage** variable?

Comments: The untransformed **wage** variable has positive skewness with an outlier \$44500. The optimal transformation makes the transformed distribution almost symmetric with a tiny skewness value. However, the positive skewness is over-compensated when using ( $\lambda = -1$ ). This leads to substantial negative skewness.

The  $p$ -values of Shapiro-Wilk normality tests with  $H_0: X \sim N(\hat{\mu}, \hat{\sigma}^2)$  for the properly Box-Cox transformed data is much smaller than 0.05, therefore transformed **wage** still deviates from the normal distribution. However, this  $p$ -value is the largest one among all three scenarios. Therefore, we can conclude the optimal transformed variable becomes closest to the normal distribution.

### Task 3: Calibration and Prediction of a Bivariate Regression Model with Skewed Variables (3 points)

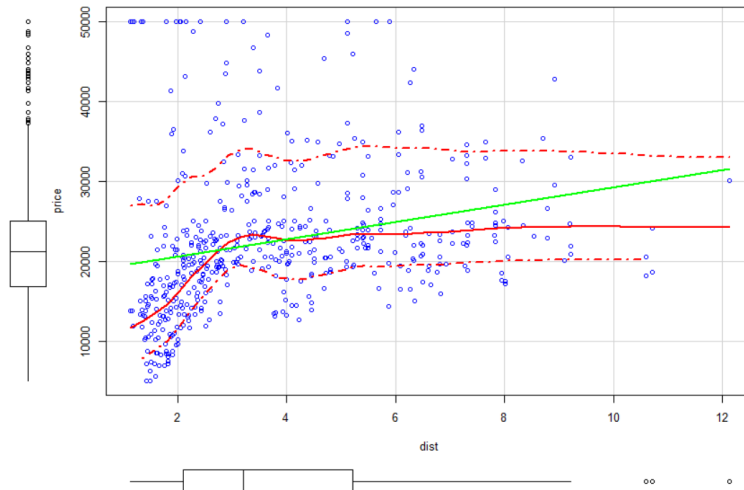
The STATA file **HPRICE2.DTA** has among other variables the **price** (home price in neighborhoods) and **dist** (weighted distance from 5 major employment centers).

Note: To import STATA files use the  function `foreign::read.dta( )`.

**Task 3.1:** Plot **price** in dependence of **dist** including their box-plots along the margins. By visual inspection of the marginal box-plots and the loess curve is a data transformation advisable? (0.5 points)

```
HPRICE2 <- foreign::read.dta('HPRICE2.DTA')
```

```
scatterplot(price ~ dist, data = HPRICE2, pch=1, smooth=list(span =
0.35, lty.smooth=1, col.smooth="red", col.var="red"), regLine=list(col="green"))
```



Comments: The dependent variable **dist** should be transformed because it is positively skewed. Furthermore, the residuals of the linear regression model are also positively skewed. To make sure the residuals satisfy the assumption of ordinary least squares, it is advisable that both variables are transformed.

**Task 3.2:** Find a proper transformation of both variables in a way that the independent variable **dist** is approximately symmetrically distributed and that the transformation of the dependent variable **price** leads to approximately symmetrically distributed regression residuals. (0.5 points)

**## Transformation of independent variable**

```
summary(powerTransform(lm(dist~1, data=HPRICE2)))
```

bcPower Transformation to Normality

	Est	Power	Rounded	Pwr	Wald	Lwr	Bnd	Wald	Upr	Bnd
Y1	-0.156		0		-0.3261			0.0141		

Likelihood ratio test that transformation parameter is equal to 0

(log transformation)

	LRT	df	pval
LR test, lambda = (0)	3.242834	1	0.071736

Likelihood ratio test that no transformation is needed

	LRT	df	pval
LR test, lambda = (1)	179.0603	1	< 2.22e-16

**## Transformation of dependent variable so residuals become approx. symmetric**

```
summary(powerTransform((lm(price~log(dist), data=HPRICE2))))
```

bcPower Transformation to Normality

	Est	Power	Rounded	Pwr	Wald	Lwr	Bnd	Wald	Upr	Bnd
--	-----	-------	---------	-----	------	-----	-----	------	-----	-----

```
Y1      0.0692      0      -0.0822      0.2206
```

Likelihood ratio test that transformation parameter is equal to 0

(log transformation)

```
LRT df      pval
```

```
LR test, lambda = (0) 0.8093437  1 0.36831
```

Likelihood ratio test that no transformation is needed

```
LRT df      pval
```

```
LR test, lambda = (1) 127.4888  1 < 2.22e-16
```

The suggested lambda parameters are  $\lambda = -0.156$  for the independent variable and  $\lambda = 0.0692$  for the regression model **lm(price~log(dist))** so that the residuals are approximately normal or at least symmetrically distributed.

**Task 3.3:** Test whether a **log**-transformation (i.e.,  $\lambda = 0$ ) is appropriate for the dependent variable and the independent variable. If the **log**-transformation is appropriate then **use it** because their relationship can be interpreted in terms of an elasticity. (0.5 points)

Comments: The likelihood ratio tests in task 3.2 suggests that the estimated lambda coefficients are not significantly different from zero. Therefore, both crime and police can be log-transformed

**Task 3.4:** Estimate the model in the transformed system and interpret the estimates. Also **test** if the elasticity (i.e., slope parameter) differs significantly from the neutral elasticity of 1, i.e.,  $H_0: \beta_1 = 1$ . This could be done manually by using the  $\beta_1$  standard error from the regression output. (1 point)

```
## Estimate the elasticity model
```

```
elast.lm <- lm(log(price)~log(dist), data=HPRICE2)
```

```
summary(elast.lm)
```

Call:

```
lm(formula = log(price) ~ log(dist), data = HPRICE2)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-1.18184 -0.21302 -0.02242  0.16747  1.20554
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.57679    0.04033  237.479  <2e-16 ***
log(dist)      0.30657    0.03091   9.919  <2e-16 ***
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.3747 on 504 degrees of freedom
```

```
Multiple R-squared:  0.1633,    Adjusted R-squared:  0.1617
```

F-statistic: 98.38 on 1 and 504 DF, p-value: < 2.2e-16

Comments: Since the bivariate regression model is specified in the log-log form, the results can be interpreted in terms of elasticity. One percent change in the number of crimes committed on university campuses will lead to 0.31 percent change in the size of the campuses police forces. The meaningful null hypothesis for elasticity is that 1% in the independent variable will lead to 1% change in the dependent variable.

```
## Test for H0: beta_log(dist)=1

slope <- coef(elast.lm)[2]

se <- summary(elast.lm)$coefficients[2, 2]

df <- nrow(HPRICE2) - 2

(t.value <- (slope-1)/se) # Note E(slope)=1 under H0

log(dist)

-22.43571

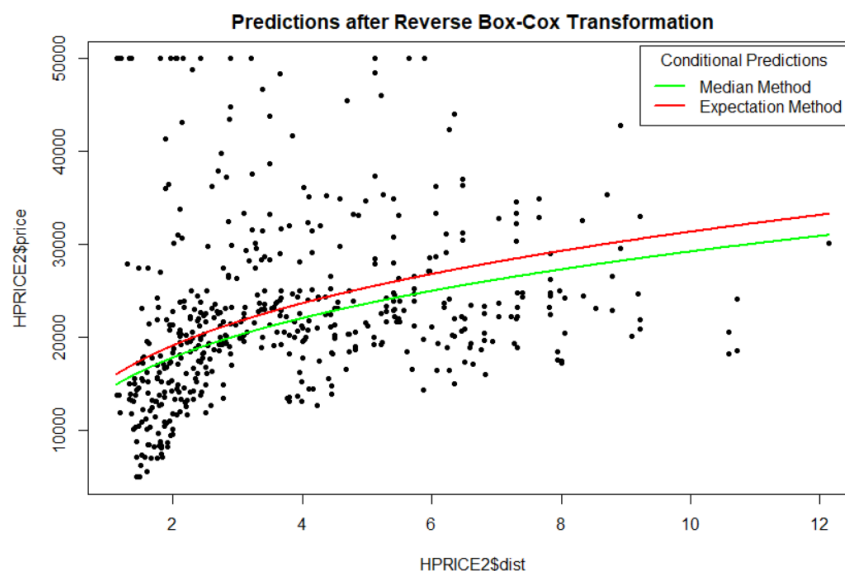
2*pt(-abs(t.value),df = df) # one-sided significance using cumulative distribution

log(dist)

8.130758e-78
```

Comments: The  $p$  value is virtually zero for this test, thus we can reject this null hypothesis  $H_0: \beta = 1$ . The elasticity 0.30657 of the regression model is significantly different from unity, which means the relative change of price is much less elastic than the relative change of dist. Ultimately it means that the relationship exhibits decreasing rates of law enforcement allocation.

**Task 3.5:** Perform a prediction in the original data units and plot the **median** and **expectation** curves. Why is the expected predicted value in this case larger than the median predicted value? (0.5 points)



Comments: For the predictions being mapped back into the original scale, the expected predicted value is larger than the median predicted value because mean is larger than median in the positively skewed distribution. This applies over the full data range of the independent variable.