Case Study II

WS 2020/2021

Deadline: 8. November, 2020@ 23:55

1 Data Description

The dataset *chicken.wf1* (and *chicken.csv*) contains the per capita chicken consumption (in lb) (*consum*) in the USA from 1960 to 1982 (yearly data, sample size $n = 23$), the disposable income per capita (in \$) (*income*), as well as the real price of chicken (*pchick*), the real price of beef (*pbeef*) and the real price of pork (*ppork*) (all in cent / lb).

2 Model

We consider the following multiple log-log model:

$$Y = \beta_0^\star \cdot X_1^{\beta_1} \cdot X_2^{\beta_2} \cdot X_3^{\beta_3} \cdot X_4^{\beta_4} \cdot e^u, \quad \mathbb{E}[u|X_1, X_2, X_3, X_4] = 0, \quad \beta_0^\star > 0.$$

The response variable Y is the demand for chicken, the covariates are

- the disposable income (X_1),
- the price of chicken (X_2),
- the price of beef (X_3) and
- the price of pork (X_4).

Demand and all prices are positive.

3 Data Analysis

3.1

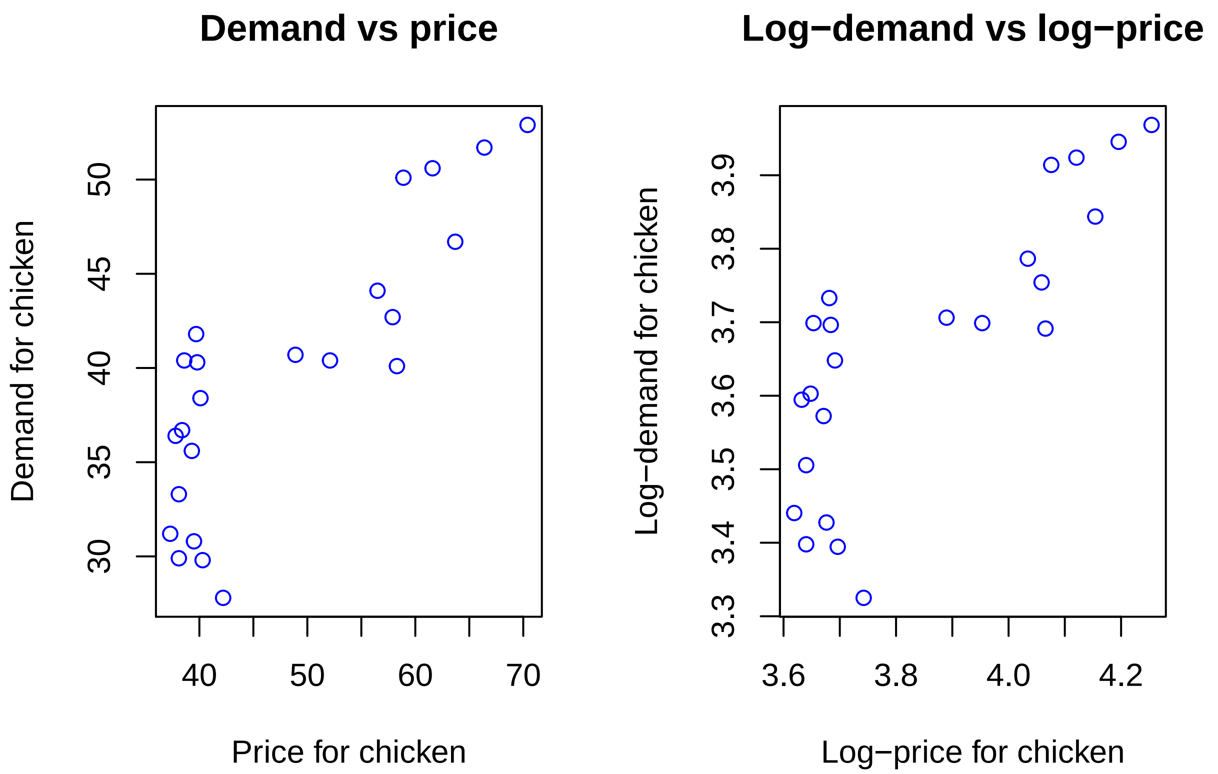
Create a scatter plot for the price of chicken versus the demand for chicken. What type of relation do you observe between the two variables? Is the relation observed compatible with what economic theory suggests?

```r
chicken <- read.csv("chicken.csv")

# Depict two pictures together
par(mfrow=c(1,2))

# Plot of chicken price against demand for chicken
plot(chicken$pchick, chicken$consum,
     main = "Demand vs price",
     ylab = "Demand for chicken",
     xlab = "Price for chicken",
     col = "blue")

# Plot of log-chicken price against log-demand for chicken
# (This is not asked in the question)
plot(log(chicken$pchick), log(chicken$consum),
     main = "Log-demand vs log-price",
     ylab = "Log-demand for chicken",
     xlab = "Log-price for chicken",
     col = "blue")
```



In the plot on the original scale, the demand for chicken varies between approximatelly 30 and 43 lb at prices of around 40 c/lb. However at higher prices (from around 47c/lb onwards), a strong positive relation can be observed where, as chicken prices rise, demand for chicken rises as well. This seems contrary to the law of demand. The result is mostly due to the fact that we have not taken demand shifters into account. When we add other variables we expect this relation to change.

3.2

Rewrite the model as

$$\log Y = \beta_0 + ...$$

Compute and report the OLS estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$. What does the result suggest regarding the relation between the demand for chicken and the price of chicken? Is the relation compatible with what economic theory suggests?

```
log_model <- lm(log(consum) ~ log(income) + log(pchick) + log(pbeef) + log(ppork), chick
summary(log_model)
```

```
##
## Call:
## lm(formula = log(consum) ~ log(income) + log(pchick) + log(pbeef) +
##     log(ppork), data = chicken)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.037877 -0.013242 -0.003969  0.010567  0.046636
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.18979    0.15571  14.063 3.77e-11 ***
## log(income)  0.34256    0.08327   4.114 0.000652 ***
## log(pchick) -0.50459    0.11089  -4.550 0.000248 ***
## log(pbeef)   0.09110    0.10072   0.905 0.377643
## log(ppork)   0.14855    0.09967   1.490 0.153452
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02759 on 18 degrees of freedom
## Multiple R-squared:  0.9823, Adjusted R-squared:  0.9784
## F-statistic: 249.9 on 4 and 18 DF,  p-value: 1.667e-15
```

We obtain $\hat{\beta}_0 = 2.18979$, $\hat{\beta}_1 = 0.34256$, $\hat{\beta}_2 = -0.50459$, $\hat{\beta}_3 = 0.09110$ and $\hat{\beta}_4 = 0.14855$. There is a negative relation between the price and demand and this is more in line with what economic theory suggests.

Extra question: What might be the motivation to consider a log-log model in the current context?

In economic theory, we are usually interested in the effects of small, *relative* changes in prices. This is due to the fact that prices may be crossly different for different goods. The log-log model allows for an interpretation of the effect of relative price changes.

3.3

Quantify how the change of disposable income affects the demand for chicken in expectation, while all other prices remain the same. The explanation should refer to not only the value, but the appropriate unit of change as well.

A 1% increase in income corresponds to approixmately 0.34% increase in demand for chcicken on average, keeping all other prices fixed.

3.4

How does a 2% increase in the price of pork affect the demand for chicken in expectation, while all other prices and disposable income remain the same?

An increase in the price of pork by 2% corresponds to approximately 0.30% increase in demand for chicken on average, keeping all other prices and income fixed.

3.5

Predict the log-demand for chicken when the disposable income is 2200$ per capita, the price of chicken 50 cent / lb, the price of pork 170 cent / lb, and the price of beef 312 cent / lb.

```
# %*% performs matrix multiplication
logpred <- c(1, log(2200), log(50), log(312), log(170)) %*% log_model$coef

# Alternatively
logpred_alt <- predict(log_model, data.frame(income = 2200, pchick = 50, ppork = 170, pb
```

$\widehat{\log y} = 4.14$

3.6

Compute and report the sample correlation (R) between the fitted values (i.e., $\widehat{\log y_i}$'s) and the observations (i.e., $\log y_i$'s). Now, calculate and report R^2 (called the *coefficient of determination*). What does this number tell us in general?

```
# The correlation of fitted values R with the observed log(y) is
Rr <- cor(fitted(log_model), log(chicken$consum))
Rr^2
```

```
## [1] 0.9823133
```

$R = 0.9911$ and $R^2 = 0.98231$. R^2 is a measure of goodness of fit. It gives the amount of variation of $\log(Y)$ which can be explained by variation in the regressors.

3.7

i) Find an estimate of the variance of the residuals.

ii) Compute the value of the unbiased estimator of the error variance σ^2.

iii) Estimate the covariance matrix of the OLS estimators of the parameters $\beta_1, \ldots, \beta_4$.

```
# The residuals are given via
resid(log_model)
```

```
##             1             2             3             4             5
## -0.0070641820 -0.0039685060 -0.0054192484 -0.0016912553 -0.0378768595
##             6             7             8             9            10
## -0.0118471721  0.0368267546  0.0068571886  0.0022901238  0.0234780257
##            11            12            13            14            15
##  0.0138048694  0.0073301305 -0.0146364830  0.0456356658 -0.0113458366
##            16            17            18            19            20
## -0.0333172061 -0.0320397885 -0.0189562069  0.0385915836  0.0466362065
##            21            22            23
## -0.0338330185  0.0003144222 -0.0097692079
```

```
# Therefore, the variance is
var(resid(log_model))
```

```
## [1] 0.0006228562
```

```
# To obtain an unbiased estimator for sigma^2,
# we need to adjust with the degrees of freedom
sigma2hat<- sum(resid(log_model)^2)/18
# Alternatively
sigma2hat_alt <- summary(log_model)$sigma^2

# Variance covarince matrix
vcm <- vcov(log_model)
covbeta <- matrix(vcm, nrow=5, ncol=5, byrow=TRUE)
covbeta
```

```
##              [,1]         [,2]         [,3]         [,4]         [,5]
## [1,]  0.024247119 -0.006559234 -0.015165454  0.007158685  0.010060694
## [2,] -0.006559234  0.006933278  0.005771445 -0.007055636 -0.006575017
## [3,] -0.015165454  0.005771445  0.012297538 -0.007322699 -0.008232352
## [4,]  0.007158685 -0.007055636 -0.007322699  0.010143805  0.004664088
## [5,]  0.010060694 -0.006575017 -0.008232352  0.004664088  0.009934626
```

i) $\hat{\sigma^2} = 0.0006228562$
ii) 0.0007612687
iii) covbeta

3.8

Fit two simple regression models:

$$Y = \alpha_0^\star \cdot X_1^{\alpha_1} \cdot e^u, \quad \text{(Model 2)}$$

and

$$Y = \alpha_0^\star \cdot X_2^{\alpha_2} \cdot e^u, \quad \text{(Model 3).}$$

Compare the estimated coefficients $\hat{\beta}_1$ and $\hat{\alpha}_1$ as well as $\hat{\beta}_2$ and $\hat{\alpha}_2$. What could be the possible reason for the differences? Explain.

```r
# First regression
log_model2 <- lm(log(consum) ~ log(income), chicken)
log_model2
```

```
##
## Call:
## lm(formula = log(consum) ~ log(income), data = chicken)
##
## Coefficients:
## (Intercept)  log(income)
##      1.4922       0.3201
```

```r
# Comparison
log_model$coef[2]
```

```
## log(income)
##   0.3425551
```

```r
# Second regression
log_model3 <- lm(log(consum) ~ log(pchick), chicken)
log_model3
```

```
##
## Call:
## lm(formula = log(consum) ~ log(pchick), data = chicken)
##
## Coefficients:
## (Intercept)  log(pchick)
##      1.0461       0.6805
```

```r
# Comparison
log_model$coef[3]
```

```
## log(pchick)
##  -0.5045924
```

The output of model 2 is close to the output of the main multiple regression model. When we compare the output of model 3 with the main model, the coefficient for the price of chicken in model 3 seems to imply a positive correlation between the price of chicken and demand, as seen in task 3.1. However, the coefficient for the price of chicken in the main model is vastly different, and has a different sign. The difference is attributable to a bias caused by ignoring the role of other variables (demand shifters) in the main model, more generally known as Omitted Variable Bias (OVB).

4 Theory

4.1

In a study relating college grade point average (GPA) to time spent in various activities, you distribute a survey to several students. The students are asked how many hours they spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student, the sum of hours in the four activities must be 168.

i) In the model

$$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + \beta_4 leisure + u$$

does it make sense to hold sleep, work, and leisure fixed, while changing study?

ii) Does the model violate any of the assumtions of multiple regression? If so explain why.

i) No. By definition, $study + sleep + work + leisure = 168$. Therefore, if we change study, we must change at least one of the other categories so that the sum is still 168.
ii) From part (i), we can write, say, study as a perfect linear function of the other independent variables: study = 168-sleep-work-leisure. This holds for every observation, so the 'no perfect multicollinearity' assumption is violated.

4.2

Suppose you are interested in estimating the ceteris paribus relationship between variables y and x_1 and for this purpose you collect data on two relevant variables x_2 and x_3. Let $\tilde{\beta}_1$ be the simple regression estimate from y on x_1 and let $\hat{\beta}_1$ be the multiple regression estimate from y on x_1, x_2, x_3.
If x_1 is almost uncorrelated with x_2 and x_3, and x_2 and x_3 have large partial effects on y, would you expect the standard error $se(\tilde{\beta}_1)$ or $se(\hat{\beta}_1)$ to be smaller? Explain.

In this case, adding x_2 and x_3 will decrease the residual variance without causing much collinearity (because x_1 is almost uncorrelated with x_2 and x_3), so we should have $se(\hat{\beta}_1)$ smaller than $se(\tilde{\beta}_1)$. The amount of correlation between x_2 and x_3 does not directly affect $se(\hat{\beta}_1)$.

For students who prefer mathematical explanation (although the above explanation is more than enough to earn you the full point if this was in your exam) remember that the standard error is an estimate of the standard deviation (not computable because we don't know the true σ^2). Therefore we have

$$se(\tilde{\beta}_1) \approx sd(\tilde{\beta}_1 | X = x_1)$$
$$se(\hat{\beta}_1) \approx sd(\hat{\beta}_1 | X = (x_1, x_2, x_3))$$

Where the approximation stems from the fact that we replace the true σ^2 with its estimate $\hat{\sigma}^2$. Now, to distinguish the true σ^2 of the SLR and MLR models, we denote the former as σ^2

and the latter as σ'^2 and similarly its esitimates as $\hat\sigma^2$ and $\hat\sigma'^2$.
With this, standard deviations above can be written as follows:

$$sd(\tilde\beta_1|X = x_1) = \frac{\sigma^2}{Ns_{x_1}^2} \quad \because \text{eq 22 in Eco I slides}$$
$$sd(\hat\beta_1|X = (x_1, x_2, x_3)) = \frac{\sigma'^2}{Ns_{x_1}^2(1 - R_1^2)} \quad \because \text{page 124}$$

here, you notice that terms N and $s_{x_1}^2$ are the same. Therefore, it is sufficient to just compare σ^2 with $\sigma'^2/(1 - R_1^2)$.
Rembmer that R_1 (defined in page 123) is the correlation between x_1 and $\hat{x_1}$ which we obtain from regressing x_1 on x_2 and x_3. Now, the problem specifically states that x_1 is almost uncorrelated with x_2 and x_3. Therefore, we can assume that $R_1 \approx 0$. Thus, it boils down to comparing σ^2 and σ'^2 if we were comparing standard deviations.
With this in mind, we know that the only term that is not observed in standard deviation is σ^2 which is substited by $\hat\sigma^2$ to obtain the standard error.
Therefore, the comparison between standard deviations we made above which ultimately amounted to the comparison between σ^2 and σ'^2 translates to the comparison between $\hat\sigma^2$ and $\hat\sigma'^2$ when comparing standard errors. These two quantities can be written down as follows:

$$\hat\sigma^2 = \frac{\sum_{i=1}^N (y_i - \tilde\beta_0 - \tilde\beta_1 x_1)^2}{N - 2}$$
$$\hat\sigma'^2 = \frac{\sum_{i=1}^N (y_i - \hat\beta_0 - \hat\beta_1 x_1 - \hat\beta_2 x_2 - \hat\beta_3 x_3)^2}{N - 4}$$

Although there is a slight difference in the denominator which inflates the SSR part (= numerator) of the first equation over the second, the problem also states that x_2 and x_3 have large partial effect on y_i. Thus, it is reasonable to assume that $\sum_{i=1}^N (y_i - \hat\beta_0 - \hat\beta_1 x_1 - \hat\beta_2 x_2 - \hat\beta_3 x_3)^2 \ll \sum_{i=1}^N (y_i - \tilde\beta_0 - \tilde\beta_1 x_1)^2$ which negates the slight difference in the denominator and thus results in $\hat{\sigma^2} > \hat\sigma'^2$. Therefore $se(\tilde\beta_1) > se(\hat\beta_1)$