Exercise Final

Final Task

15.11.2023

Contents

General Information

Name:

Student ID:

Study Program: 033.534 “Bachelorstudium Software & Information Engineering”

Task 0: Import:

```
chicken <- read.csv("Exercise_Final/chicken.csv")
set.seed(12119060)
id <- sample(1:nrow(chicken), floor(0.8 * nrow(chicken)))
#df <- df[id, ]

#chicken
```

Task a: Summary statistics

```
head(chicken)
```

```
##   state   consum    income pbeef pchick ppork
## 1    S1 12.33443  928.8349 124.1   52.1  95.4
## 2    S1 14.35851 1577.5363 165.5   63.7 130.9
## 3    S1 14.35090 1164.5042 142.9   58.3 123.5
## 4    S1 13.40811  766.3706 106.1   38.6  73.2
## 5    S1 12.06504  411.8078  79.2   38.1  52.0
## 6    S1 15.21261 2257.4348 221.6   66.4 141.0
```

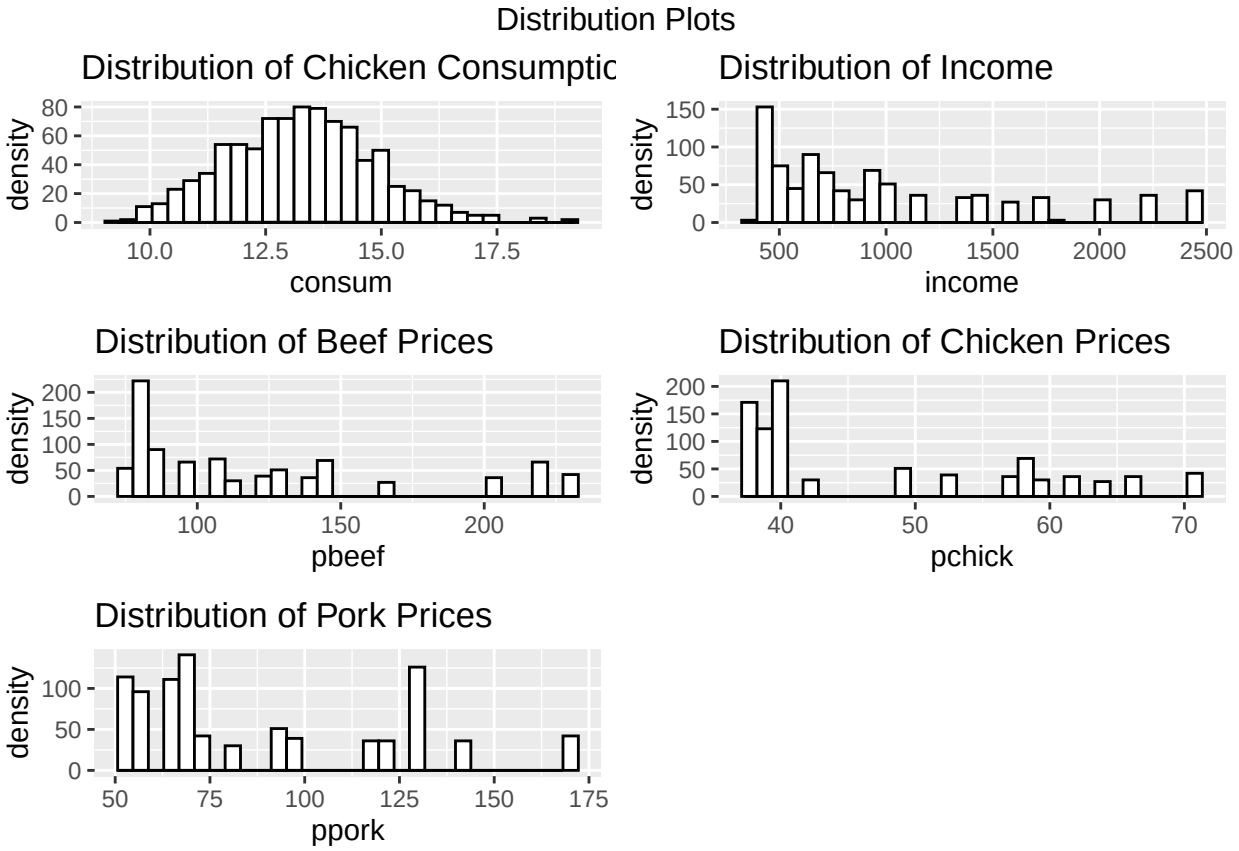
```
descriptive_stats <- summary(chicken)

t(descriptive_stats)
```

```
##
##    state   Length:900         Class :character   Mode  :character
##     consum Min.   : 9.277     1st Qu.:12.115     Median :13.281
##     income Min.   : 394.9     1st Qu.: 530.9     Median : 767.9
##     pbeef  Min.   : 77.4      1st Qu.: 80.2      Median :104.8
##     pchick Min.   :37.30      1st Qu.:38.60      Median :40.10
##     ppork  Min.   : 50.70     1st Qu.: 63.70     Median : 70.00
##
##    state
##     consum Mean   :13.259   3rd Qu.:14.287   Max.   :19.155
##     income Mean   : 998.1   3rd Qu.:1349.9   Max.   :2480.2
##     pbeef  Mean   :121.1    3rd Qu.:142.9    Max.   :232.6
##     pchick Mean   :47.22    3rd Qu.:57.90    Max.   :70.40
##     ppork  Mean   : 88.21   3rd Qu.:123.50   Max.   :168.20
```

Task b: Graphical description:

```
## Warning: Paket 'gridExtra' wurde unter R Version 4.3.2 erstellt
```



1. Distribution of Chicken Consumption (consum):

This histogram shows a roughly normal distribution with a slight right skew, indicating that most states have a chicken consumption around the average, with fewer states having significantly higher consumption.

2. Distribution of Income (income):

The income distribution is right-skewed, suggesting that most states have incomes below the average, with a few states having significantly higher incomes.

3. Distribution of Beef Prices (pbeef):

The distribution of beef prices shows a bimodal distribution, indicating two common price ranges for beef across states.

4. Distribution of Chicken Prices (pchick):

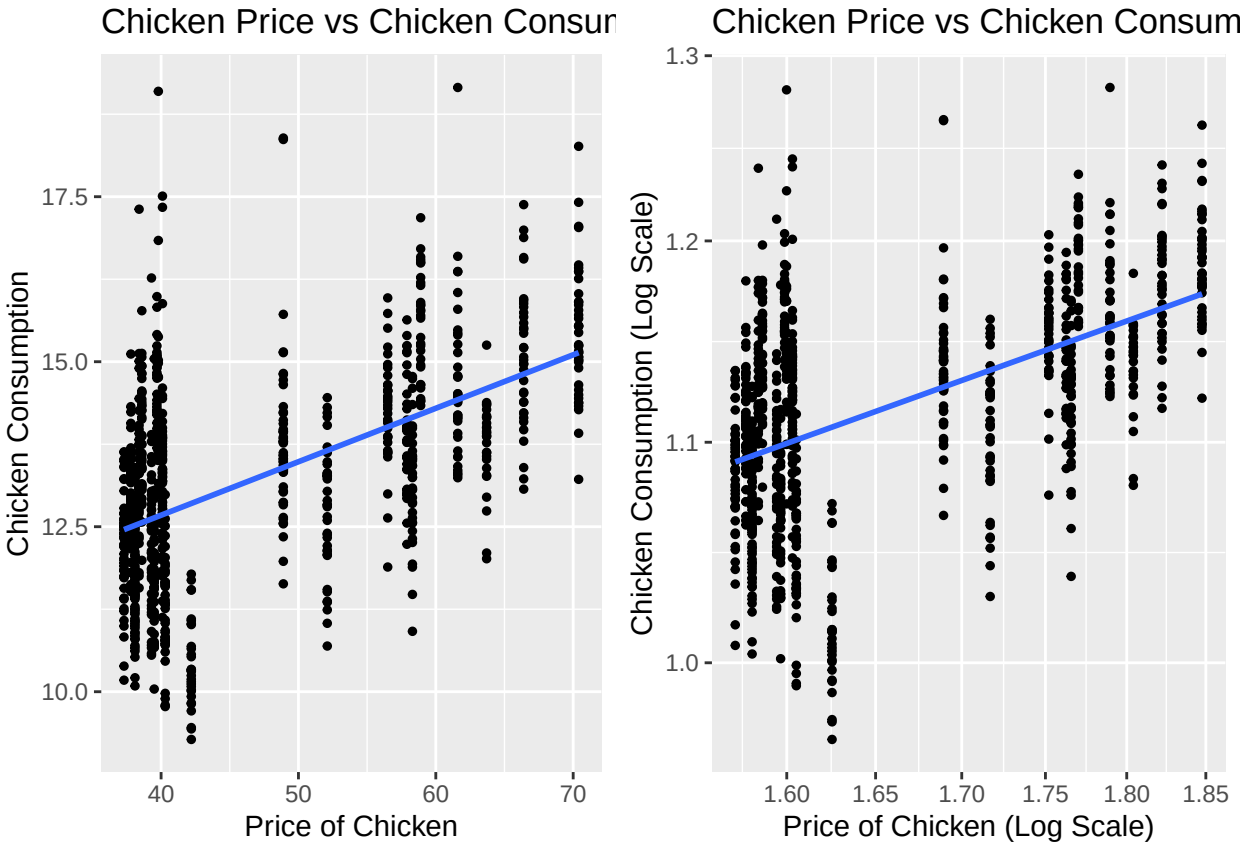
The chicken prices distribution appears to be bimodal as well, with one peak at a lower price range and another at a higher range.

5. Distribution of Pork Prices (ppork):

Similar to beef and chicken prices, the pork prices also show a bimodal distribution.

Task c: Graphical description:

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```




```
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
```

In the provided scatter plots, we see the relationship between the price of chicken and chicken consumption, both in a standard linear scale and in a logarithmic scale. Here's the observed relationship:

1. Standard Linear Scale (Left Plot):

The plot suggests a positive relationship between the price of chicken and chicken consumption, which is indicated by the upward slope of the regression line. This is counterintuitive to the basic law of demand, which states that demand typically decreases as prices increase.

2. Logarithmic Scale (Right Plot):

When both variables are transformed into a logarithmic scale, the positive relationship persists, as indicated by the upward slope of the regression line in the log-log plot. This suggests that the relationship could be modeled as a power function in the original scale, where consumption is a power function of price.

Analysis:

If we were to follow conventional economic theory, we would expect to see a negative relationship between price and demand, meaning that as the price of chicken rises, the demand for chicken would typically fall.

However, the observed positive relationship in both plots could be due to a variety of factors:

- **Substitution Effect:** If chicken is relatively cheaper compared to other meats (like beef and pork) even after the price rise, consumers might still consume more chicken as it becomes a more attractive option.
- **Income Effect:** If income levels have risen, the increase in demand could offset the effect of the price rise.
- **Quality Perception:** Consumers might associate higher prices with better quality, leading to increased demand.
- **Data Time Frame:** The data might be capturing a period where the demand for chicken is inelastic or less sensitive to price changes due to other prevailing market conditions.
- **Supply Shifts:** If there is a shift in the supply curve due to technological improvements or other factors, it could change the price-consumption relationship.

It is also possible that the dataset includes confounding variables that are not being controlled for in these simple scatter plots. A more in-depth analysis, possibly including additional data, would be necessary to draw more accurate conclusions about the relationship between chicken prices and consumption. (Next Task)

Task d: Estimation and interpretation:

```
## [1] "correlation_price_consum:"

## [1] 0.5447309

## [1] "regression summary:"

##
## Call:
## lm(formula = consum ~ pchick + pbeef + ppork + income, data = chicken_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0121 -0.6591 -0.0316  0.5535  5.7179
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.3212401  0.3095443  52.727  < 2e-16 ***
## pchick      -0.2444284  0.0139621 -17.507  < 2e-16 ***
## pbeef        0.0124936  0.0043099   2.899  0.00384 **
## ppork        0.0608102  0.0053952  11.271  < 2e-16 ***
## income       0.0016061  0.0004059   3.956 8.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9884 on 895 degrees of freedom
## Multiple R-squared:  0.6163, Adjusted R-squared:  0.6146
## F-statistic: 359.4 on 4 and 895 DF,  p-value: < 2.2e-16
```

The linear regression model based on Ordinary Least Squares (OLS) estimation for the relationship between chicken consumption (consum) and its predictors is as follows: (Note that I use B instead of beta and e instead of epsilon)

consum = β 0 + β 1 x pchick + β 2 x pbeef + β 3 x ppork + β 4 x income + ϵ

where:

`consum` is the chicken consumption.

`pchick` is the price of chicken. pbeef is the price of beef.

`ppork` is the price of pork. income is the average income.

`$\beta$0, $\beta$1, $\beta$2, $\beta$3, and $\beta$4` are the coefficients to be estimated.

`$\epsilon$` is the error term.

The OLS estimates from the regression for the whole dataset are as follows:

Intercept (β 0): 16.3212

Price of Chicken (β 1): -0.2444

Price of Beef (β 2): 0.0125

Price of Pork (β 3): 0.0608

Income (β 4): 0.0016

These coefficients represent the estimated change in chicken consumption (in the units measured) associated with a one-unit change in the respective predictor, holding all other predictors constant.

The correlation analysis between the price of chicken (pchick) and chicken consumption (consum) resulted in a Pearson correlation coefficient of approximately 0.545, which indicates a moderate positive correlation in the raw data without controlling for other factors.

However, when we control for other factors in the multiple regression model, we see that the price of chicken (pchick) has a negative coefficient (-0.2444), which aligns with the conventional law of demand—indicating that as the price of chicken increases, the demand for chicken decreases, all else being equal.

The model's R-squared value is 0.616, suggesting that approximately 61.6% of the variability in chicken consumption is explained by the model. The positive coefficients for pbeef, ppork, and income suggest that as the prices of beef and pork increase, or as income increases, chicken consumption also tends to increase, which may capture substitution effects or a general increase in meat consumption due to higher income.

The results from the OLS regression model provide several insights into the factors affecting chicken consumption:

Price of Chicken (pchick):

- The model suggests that the price of chicken has a negative impact on chicken consumption, with a coefficient of -0.2444. This means that for every one-unit increase in the price of chicken, we can expect chicken consumption to decrease by approximately 0.2444 units, all else being equal. This result aligns with the basic economic law of demand, which states that, typically, the quantity demanded of a good falls as the price rises, assuming other factors are held constant.

Price of Beef (pbeef):

- The price of beef has a positive coefficient (0.0125), suggesting that as beef becomes more expensive, chicken consumption slightly increases. This can be interpreted as a substitution effect; as beef becomes more expensive, consumers might switch to consuming more chicken.

Price of Pork (ppork):

- Similarly to beef, the positive coefficient for the price of pork (0.0608) indicates a substitution effect. An increase in pork prices is associated with an increase in chicken consumption, possibly because consumers view chicken as a substitute for pork.

Income (income):

- The positive coefficient (0.0016) indicates that higher income levels are associated with higher chicken consumption. This could be due to the income effect, where individuals with higher income consume more goods in general, or it might indicate that chicken is a normal good for which demand increases as consumer income rises.

These results contrast with the initial descriptive analysis, which showed a positive correlation between the price of chicken and chicken consumption. The regression analysis, which is more robust as it controls for other variables, reveals the negative impact of chicken price on demand, a more intuitive and expected economic relationship.

Overall, the regression analysis suggests that while there might be a positive univariate correlation between chicken prices and consumption in the raw data, once we account for the effects of other relevant economic variables, the expected negative relationship between price and demand is evident.

Task e: Interpretation of linear regression model coefficients:

The coefficient for disposable income in the regression model is 0.0016. This value quantifies the expected change in chicken consumption for every one-unit change in income, holding all other factors constant. However, to interpret this value correctly, we need to be aware of the units in which income and chicken consumption are measured.

Assuming that income is measured in dollars and chicken consumption is measured in some quantity unit (like pounds or kilograms), the coefficient of 0.0016 means that for every additional dollar of disposable income, the consumption of chicken is expected to increase by 0.0016 units (pounds or kilograms) on average.

For example, if an individual's disposable income increases by $1,000, the model predicts that their chicken consumption would increase by:

0.0016 units/dollar $\times$ 1.000 dollars = 1.6 units

So, in this case, an increase of $1,000 in disposable income is expected to increase chicken consumption by 1.6 units (pounds or kilograms), according to the model's predictions. This interpretation is essential for policymakers or businesses as it helps in understanding the sensitivity of chicken consumption to changes in consumer income, which can be useful for demand forecasting and pricing strategies.

Task f: Prediction for a new observation:

```
##        1
## 51.67611
```

The predicted demand for chicken, given a yearly disposable income of $22,000, a price of chicken at $0.50 per pound, a price of pork at $1.70 per pound, and a price of beef at $3.12 per pound, is approximately 51.68 pounds.

Task g: Sampling properties of OLS estimators

1. Find an estimate of the variance of the residuals.

```
## residual_variance:  0.9884301
```

2. Compute the value of the unbiased estimator of the error variance α^2

```
## error_variance_unbiased:  0.9715664
```

3. covariance matrix

```
## [1] "cov_matrix:"
```

```
##                     1        income         pbeef        pchick         ppork
## 1       9.581765e-02  4.765361e-05 -3.074516e-04 -3.183876e-03  5.133245e-04
## income  4.765361e-05  1.647951e-07 -1.521360e-06  1.575394e-06 -1.159615e-06
## pbeef  -3.074516e-04 -1.521360e-06  1.857556e-05 -2.100164e-05  6.440741e-06
## pchick -3.183876e-03  1.575394e-06 -2.100164e-05  1.949400e-04 -5.725354e-05
## ppork   5.133245e-04 -1.159615e-06  6.440741e-06 -5.725354e-05  2.910808e-05
```

```
## [1] "summary_model:"
```

```
##
## Call:
## lm(formula = consum ~ pchick + pbeef + ppork + income, data = chicken_data)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.0121 -0.6591 -0.0316  0.5535  5.7179
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.3212401  0.3095443  52.727  < 2e-16 ***
## pchick      -0.2444284  0.0139621 -17.507  < 2e-16 ***
## pbeef        0.0124936  0.0043099   2.899  0.00384 **
## ppork        0.0608102  0.0053952  11.271  < 2e-16 ***
## income       0.0016061  0.0004059   3.956 8.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9884 on 895 degrees of freedom
## Multiple R-squared:  0.6163, Adjusted R-squared:  0.6146
## F-statistic: 359.4 on 4 and 895 DF,  p-value: < 2.2e-16
```

```
## [1] "std errors:"
```

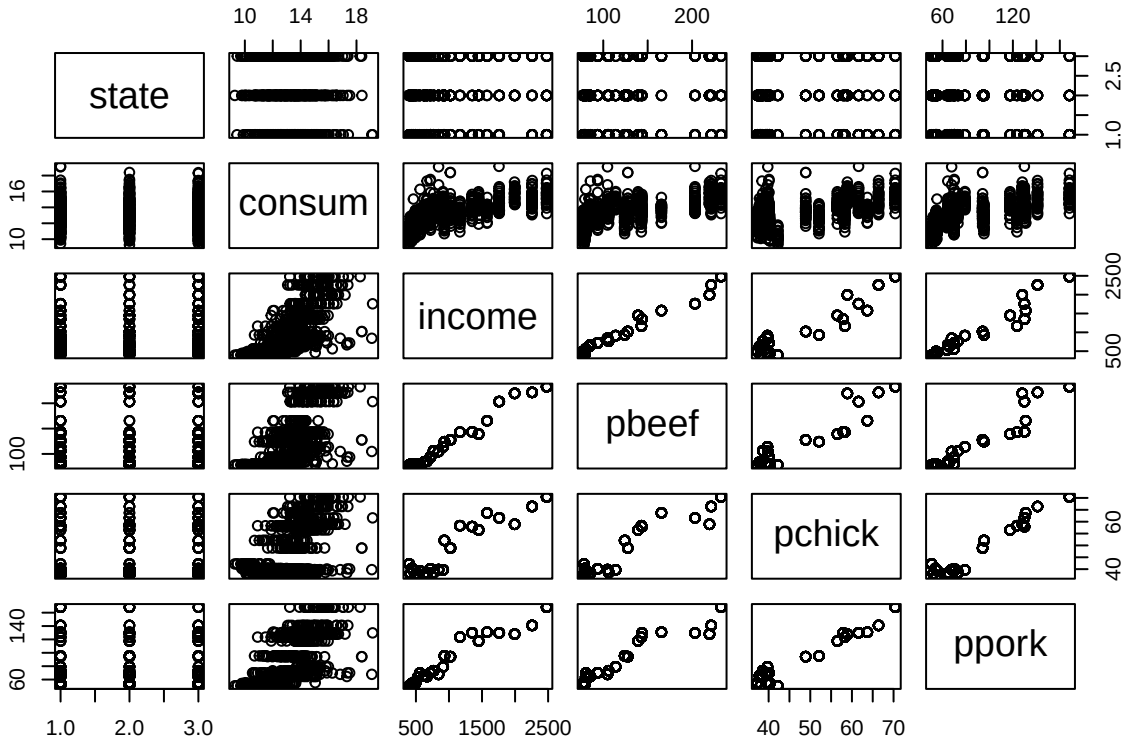
```
##  (Intercept)       pchick        pbeef        ppork       income
## 0.3095442674 0.0139620904 0.0043099379 0.0053951905 0.0004059497
```

To compare the covariance matrix with the standard errors from the summary of the linear model, we focus on the diagonal elements of the covariance matrix. The standard errors are the square roots of these diagonal elements.

1. **Intercept (`Intercept`)**:
   - Covariance Matrix: `9.581765e-02`
   - Summary Standard Error: `0.3095442674`
   - Square Root of Covariance Matrix Value: $\sqrt{9.581765e-02} \approx 0.3095$
2. **Income (`income`)**:
   - Covariance Matrix: `1.647951e-07`
   - Summary Standard Error: `0.0004059497`
   - Square Root of Covariance Matrix Value: $\sqrt{1.647951e-07} \approx 0.000406$
3. **Price Beef (`pbeef`)**:
   - Covariance Matrix: `1.857556e-05`
   - Summary Standard Error: `0.0043099379`
   - Square Root of Covariance Matrix Value: $\sqrt{1.857556e-05} \approx 0.00431$
4. **Price Chick (`pchick`)**:
   - Covariance Matrix: `1.949400e-04`
   - Summary Standard Error: `0.0139620904`
   - Square Root of Covariance Matrix Value: $\sqrt{1.949400e-04} \approx 0.01396$
5. **Price Pork (`ppork`)**:
   - Covariance Matrix: `2.910808e-05`
   - Summary Standard Error: `0.0053951905`
   - Square Root of Covariance Matrix Value: $\sqrt{2.910808e-05} \approx 0.00540$

The square roots of the diagonal elements of the covariance matrix closely match the standard errors in the summary of the linear model, indicating that the covariance matrix was correctly computed.

Task h: Model diagnostics



- **Linearity:**

The relationship between the predictors and the dependent variable (consum) should ideally show some linear pattern if a linear model is to work well. Non-linear relationships suggest that transformations of variables or non-linear models might be more appropriate.

- **Homoscedasticity:**

The spread of the residuals should be consistent across all levels of the predictors. If the spread (variance) of the points increases or decreases with the values of the predictors, this would indicate heteroscedasticity, which violates the OLS assumption of constant variance.

- **Outliers:**

Points that are far away from others can be potential outliers. These can have a disproportionate influence on the regression line, especially if the number of observations is not very large.

- **Multicollinearity:**

We look for patterns in the scatter plots between the independent variables. If there appears to be a linear relationship between two predictors, it may indicate multicollinearity, which can make the estimates of the regression coefficients unstable and their interpretation problematic.

Task i: Interactions:

```
## [1] "summary:"

##
## Call:
## lm(formula = formula, data = chicken_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8596 -0.6437 -0.0130  0.5496  5.6624
##
## Coefficients:
##                    Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)    16.31261518  0.31194303  52.294 < 0.0000000000000002 ***
## income          0.00133704  0.00069933   1.912               0.0562 .
## pbeef           0.01919326  0.00843270   2.276               0.0231 *
## pchick         -0.24389658  0.02028517 -12.023 < 0.0000000000000002 ***
## ppork           0.05443410  0.00878692   6.195       0.000000000892 ***
## income:stateS2 -0.00007446  0.00097227  -0.077               0.9390
## income:stateS3  0.00068935  0.00091812   0.751               0.4530
## pbeef:stateS2  -0.00657659  0.01094327  -0.601               0.5480
## pbeef:stateS3  -0.01223109  0.01081338  -1.131               0.2583
## pchick:stateS2 -0.00681585  0.02464815  -0.277               0.7822
## pchick:stateS3 -0.00087991  0.02318607  -0.038               0.9697
## ppork:stateS2   0.01380209  0.01361538   1.014               0.3110
## ppork:stateS3   0.00945143  0.01187739   0.796               0.4264
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9897 on 887 degrees of freedom
## Multiple R-squared:  0.6188, Adjusted R-squared:  0.6137
## F-statistic:   120 on 12 and 887 DF,  p-value: < 0.00000000000000022
```

In the regression model results, State 1 is not explicitly displayed because it serves as the reference or baseline category against which the other states (State 2, State 3, etc.) are compared.

Quote (From the Vorlesung):

> **$p < 0.05$:** This is the most commonly used threshold. If the p-value is less than 0.05, the result is typically considered statistically significant. This means there's less than a 5% probability that the observed results occurred by chance under the null hypothesis.
>
> **$p < 0.01$:** This is a more stringent threshold, indicating a stronger evidence against the null hypothesis. A p-value less than 0.01 suggests there's less than a 1% chance that the observed results are due to random chance under the null hypothesis.
>
> **$p < 0.001$:** This threshold is even more stringent and indicates very strong evidence against the null hypothesis. A p-value less than 0.001 means there's less than a 0.1% chance that the results occurred by chance under the null hypothesis.

From the regression results, we can comment on the significance of differences among states in terms of how various factors influence chicken consumption (consum). These factors include income, pbeef, pchick, and ppork. Interactions between the state and these variables in the model allow us to assess these differences.

Based on the regression, the differences among the states, as captured by the interaction terms between the state variable and the other predictors, are not statistically significant. This conclusion is drawn from the p-values associated with each of the interaction terms:

Summary

1. **Interaction Terms with State S2 and S3**:
   - Every interaction term involving `stateS2` (like `income:stateS2`, `pbeef:stateS2`, etc.) and `stateS3` has a p-value well above the conventional significance threshold of 0.05.
   - This indicates that the slopes (i.e., the relationships between the dependent variable `consum` and the independent variables `income`, `pbeef`, `pchick`, `ppork`) do not differ significantly between states S2 or S3 and the baseline state (presumably S1).
2. **Lack of Statistical Significance**:
   - The lack of statistical significance in these interaction terms suggests that the effect of the predictors on the dependent variable `consum` is consistent across the different states.
   - In other words, the model does not provide sufficient evidence to conclude that the states differ in how the independent variables influence `consum`.

Details:

Income:

- The interaction terms for income (C(state)[T.S2]:income and C(state)[T.S3]:income) are not statistically significant (p-values are 0.939 and 0.453, respectively). This suggests that the effect of income on chicken consumption does not differ significantly across states.

Price of Beef (pbeef):

- The interaction terms for the price of beef (C(state)[T.S2]:pbeef and C(state)[T.S3]:pbeef) also are not statistically significant (p-values are 0.548 and 0.258, respectively). This indicates that the impact of beef prices on chicken consumption is consistent across states.

Price of Chicken (pchick):

- The interaction terms for the price of chicken (C(state)[T.S2]:pchick and C(state)[T.S3]:pchick) are not significant either (p-values are 0.782 and 0.970). Therefore, the effect of chicken prices on its consumption is similar across the states.

Price of Pork (ppork):

- Similarly, for pork prices, the interaction terms (C(state)[T.S2]:ppork and C(state)[T.S3]:ppork) do not show significant differences among states (p-values are 0.311 and 0.426).

Overall, based on these results, we can conclude that the differences among the states are not statistically significant for the variables considered in this model. The impact of income, beef prices, chicken prices, and pork prices on chicken consumption is similar across the different states included in this analysis.

Task j: Negative log-likelihood

```
## Maximum Likelihood Estimates:  16.3192 0.001604746 0.01250888 -0.2443959 0.06081005
```

```
## Maximum Likelihood Estimates sigma squared:  0.971969
```

Result comparison:

MLE Estimates:

- **Intercept:** 16.319202621
- **Income:** 0.001604746
- **PBeef:** 0.012508878
- **PChick:** -0.244395931
- **PPork:** 0.060810050
- **Error Variance** (σ^2): 0.971969

OLS Estimates:

- **Intercept:** 16.0960317
- **Income:** 0.0012216
- **PBeef:** 0.0200197
- **PChick:** -0.2367273
- **PPork:** 0.0532103
- **Residual Standard Error:** (σ) 0.982
- **Residual Standard Error squared:** (σ^2) 0.964324

Comparison and Analysis:

Intercept: - MLE: 16.3192 - OLS: 16.0960 - Slightly higher in the MLE model.

Income: - MLE: 0.001605 - OLS: 0.001222 - The MLE estimate is marginally higher.

PBeef: - MLE: 0.012509 - OLS: 0.020020 - The OLS estimate is notably higher.

PChick: - MLE: -0.2444 - OLS: -0.2367 - Very close, with the MLE estimate being slightly more negative.

PPork: - MLE: 0.060810 - OLS: 0.053210 - The MLE estimate is higher.

Error Variance (σ^2): - MLE: 0.971969 - OLS: 0.964324 - Very close, with the MLE estimate being slightly higher.

Detailed Result comparison:

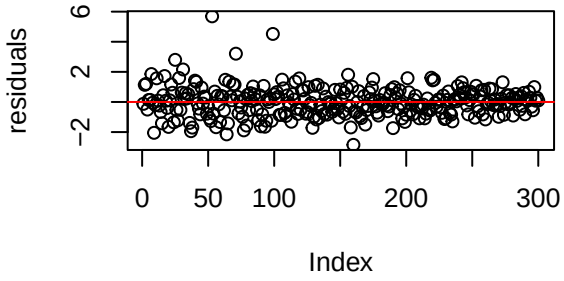
1. **Intercept**:
    - The intercept is slightly higher in the MLE model compared to the OLS model. This suggests that when all predictors are zero, the predicted value of the dependent variable (`consum`) is slightly higher under the MLE model.
2. **Income**:
    - The MLE estimate for income is marginally higher than the OLS estimate. While both indicate a positive relationship with `consum`, the MLE suggests a slightly stronger effect.
3. **PBeef**:
    - There is a notable difference in the estimates for `pbeef`. The OLS estimate is significantly higher, suggesting a stronger relationship with `consum` in the OLS model. This difference could be indicative of how the two methods handle the error structure or non-normality in the data.

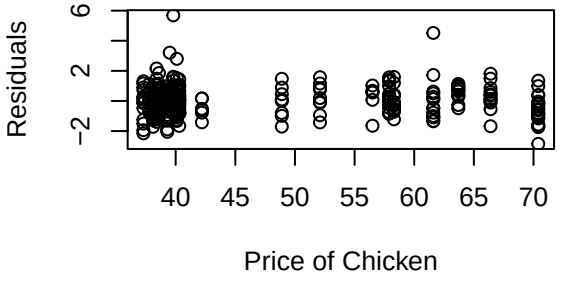
4. **PChick**:
    - Estimates for `pchick` are very close, though the MLE estimate is slightly more negative. Both models agree on a negative relationship between `pchick` and `consum`, but the strength of this relationship is marginally stronger in the MLE model.
5. **PPork**:
    - The MLE estimate for `ppork` is higher than the OLS estimate, indicating a stronger positive relationship with `consum` in the MLE model.
6. **Error Variance** (σ^2):
    - The error variance is slightly higher in the MLE model. This parameter reflects the variability in the data that is not explained by the model. A higher variance in the MLE model might suggest a slightly different interpretation of the error structure compared to the OLS model.

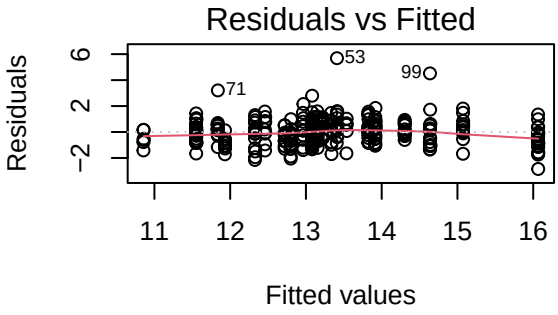
Implications

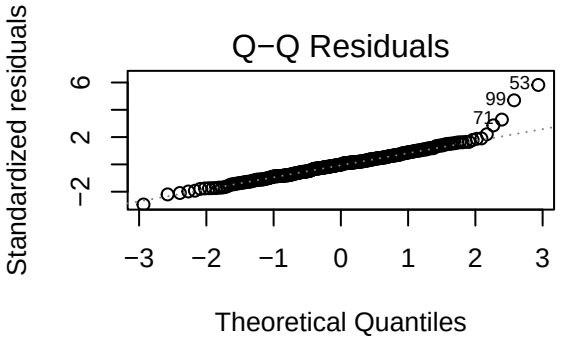
- The MLE and OLS estimates are quite similar, but there are noticeable differences. These differences can be attributed to the methods of estimation: OLS minimizes the sum of squared residuals, while MLE maximizes the likelihood under the assumption of normally distributed errors.
- The slight variations in estimates could also be influenced by the distribution and properties of the data, especially if there are deviations from the assumptions underlying the OLS method (like non-normality of errors).

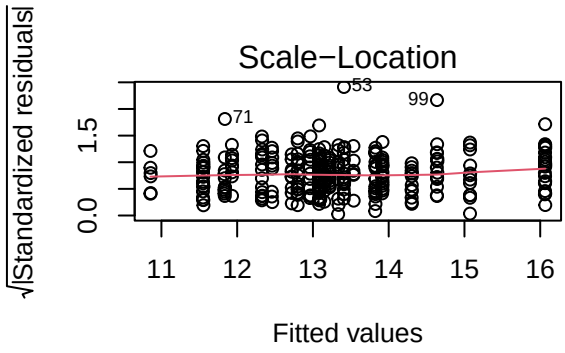
Task k: Randomized Test

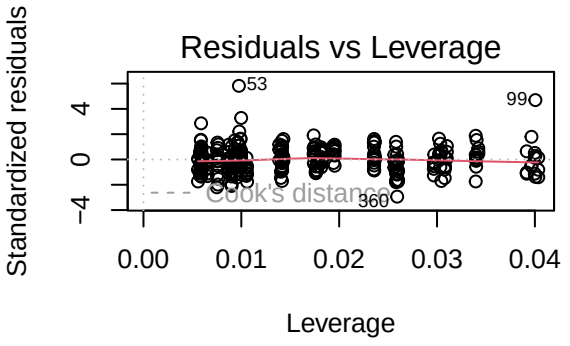












Residuals vs. Fitted Plot:

- This plot helps us check the assumption of linearity and homoscedasticity. We expect to see the residuals scattered randomly around the horizontal line at zero, with no clear pattern.
- If there's a pattern or a systematic structure in the plot, it suggests that the relationship is not linear, or there is some other specification error in the model.
- In the provided plot, residuals are scattered around the horizontal line at zero without a clear pattern, which is good. There's no obvious curvature or systematic pattern, suggesting that the linearity assumption is reasonable for this model. However, there are a few outliers visible that stand away from the zero line.

Normal Q-Q Plot:

- This plot is used to examine whether the residuals are approximately normally distributed. We expect the points to fall approximately along the reference line.
- A systematic deviation from the reference line suggests that the residuals have a distribution that is not normal.
- The Q-Q plot shows that the residuals roughly follow the reference line, but there are deviations on both ends of the plot. This suggests that while the residuals are approximately normal, there are some outliers, as evidenced by the points that stray from the line in the tails.

Scale-Location Plot (or Spread-Location Plot):

- This plot shows the spread of residuals versus the fitted values and is used to check homoscedasticity. We want to see a horizontal line with equally (randomly) spread points.
- A funnel shape would suggest heteroscedasticity, meaning the error variance changes with the level of the predictor variable.
- This plot shows that the residuals are spread fairly evenly across the range of fitted values, which is a good sign for equal variance (homoscedasticity). However, there seems to be a slight increase in spread for fitted values in the middle range, which could indicate potential issues with homoscedasticity.

Residuals vs. Leverage Plot:

- This plot helps us to find influential cases (if any), which are observations that have a large influence on the estimation of the regression coefficients.
- Observations with both high leverage and large residuals (outliers) are particularly influential to the regression line. Points outside the Cook's distance lines are potential influencers.
- There doesn't appear to be any points with high leverage and high residuals, which would be cause for concern. The Cook's distance lines do not seem to be exceeded, suggesting that there are no highly influential outliers in the model.

Histogram of Residuals:

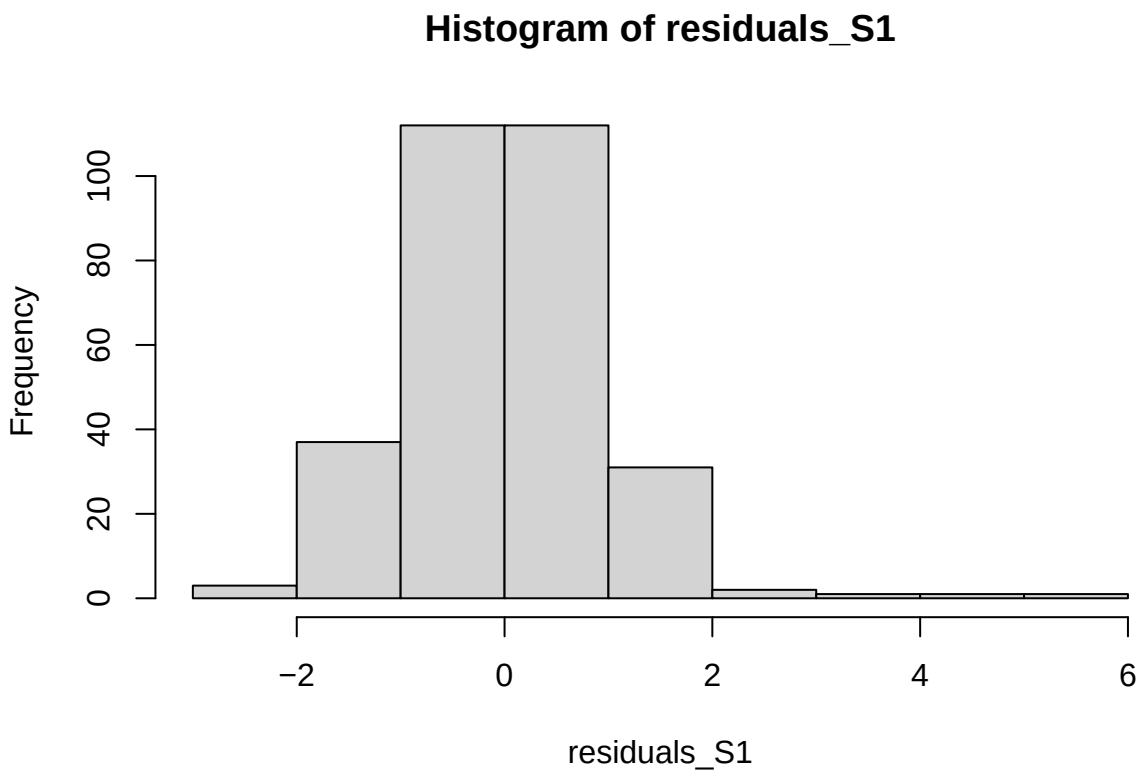
- This plot gives us a quick look at the distribution of residuals. We expect to see a bell-shaped distribution indicating normality.
- Any noticeable skewness or a distribution that is not symmetric around zero suggests the residuals are not normally distributed.
- The histogram shows that the residuals are not perfectly normally distributed. They are skewed right, with a few large residuals (indicating potential outliers), but the majority of the data is clustered around the center.

Residuals vs. an Independent Variable Plot:

- These plots can show whether there are any patterns in the residuals when plotted against individual predictors, which could indicate non-linearity.
- The scatter of residuals across different values of the price of chicken does not show any clear pattern or systematic structure, which is a positive indication. However, the concentration of data points at certain price levels suggests there might be specific price points where data is more or less dense.

---

review of Task d:




```
##
##  Shapiro-Wilk normality test
##
## data:  residuals_S1
## W = 0.94739, p-value = 0.000000007034
```

Shapiro-Wilk Normality Test

The Shapiro-Wilk test for normality on the residuals shows a p-value of 7.034×10^{-9}. This is significantly lower than the usual threshold of 0.05, indicating that the residuals do not follow a normal distribution. This result is crucial as the normality the normality assumtion is violated.

Randomization/Permutaion Test

```
## $standard_p_value
## [1] 0.00000001075972
##
```

```
## $mean_permuted_p_value
## [1] 0
```

Analysis of Linear Model for State S1

Model Summary

- **Coefficients**: The model shows significant coefficients for `income`, `pbeef`, `pchick`, and `ppork`, with `pchick` and `ppork` demonstrating particularly strong effects.
- **Residual Standard Error**: The value is 0.982, suggesting a good fit of the model to the data.
- **R-squared**: At 0.5979, this indicates that approximately 59.79% of the variability in the dependent variable is explained by the model.

Comparison of P-Values: Standard vs. Randomization Test

- **Standard P-Value for `ppork`**: 1.076×10^{-8}, showing strong statistical significance under normal theory-based testing.
- **Mean Permuted P-Value**: 0 from the randomization test (Kennedy method), indicating that under permutation of the dependent variable, such extreme test statistics as observed in the actual data are extremely rare or never observed in 1000 permutations.

Interpretation and Implications

1. **Violation of Normality Assumption**: This violation may undermine the reliability of the standard hypothesis tests in the OLS model. It is particularly relavant in smaller samples where the central limit theorem doesn't sufficiently ensure the normality of the sampling distribution.
2. **Randomization Test Result**: The stark contrast between the zero p-value from the randomization test and the standard p-value suggests that the significance of the `ppork` coefficient might not be as robust as indicated by the standard test.
3. **Model Reliability**: Despite the model explaining a substantial portion of the variability and the predictors appearing significant, the violation of normality and the results from the randomization test advise caution in over-relying on standard p-values for inference.
4. **Further Analysis and Validation**: Alternative approaches might be necessary, such as transforming variables, using robust regression techniques, or non-parametric methods, depending on the specific objectives and nature of the data.

Feedback

44.5/50

Subtask d: You cannot just take the variables from the non-linear multiplicative model and form a linear model with those variables. If you apply the log to the demand model, then we get
log(Y) = log(\beta_{0}^{*}) + \beta_1 * log(X_1) + \beta_2 * log(X_2) + \beta_3 * log(X_3) + \beta_4 * log(X_4) + u. This is the linear model we are looking for.
-2

Subtask e: Since we would consider the log transformed model, the interpretation would be: If we increase the yearly disposable income by x percent, we expect a rise in the yearly per capita chicken consumption in lbs by approximately \beta_1*x percent.
-2

Subtask g1: The variance of the residuals is given by $\mathbb{V}(r_i) = \sigma^{2}(1 - h_i)$ with h_i being the i-th diagonal entry of $H = X(X'X)^{-1}X'$ (this can be shown with an easy calculation). Since this was not covered in the lecture, I will not deduct points here.

Subtask g2 incorrect, not an unbiased estimator
-1

Subtask g3: You were supposed to calculate the variance-covariance matrix of the OLS estimator efficiently. -0.5

Good work!