Regression and Correlation

Lin Cong

University of California, Riverside

gradquant@ucr.edu

October 14, 2019

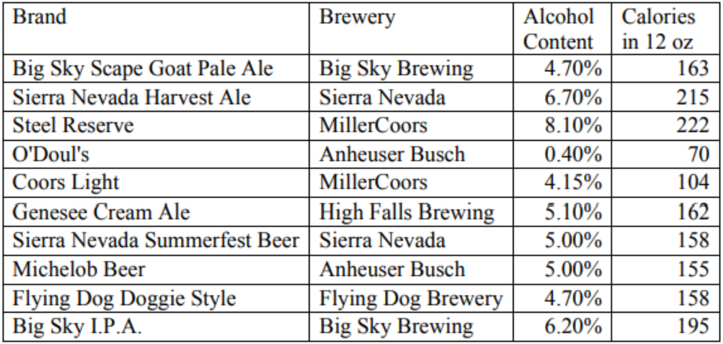
Contents

1. Simple Linear Regression
2. Potential problems
3. Correlation
4. Inference

Simple Linear Regression

Introduction

Beer data:
The dataset contains statistics about beers alcohol content and the number of calories in 12-ounce beer.

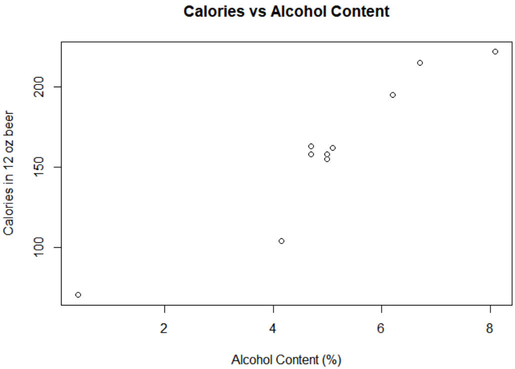


Brand	Brewery	Alcohol Content	Calories in 12 oz
Big Sky Scape Goat Pale Ale	Big Sky Brewing	4.70%	163
Sierra Nevada Harvest Ale	Sierra Nevada	6.70%	215
Steel Reserve	MillerCoors	8.10%	222
O'Doul's	Anheuser Busch	0.40%	70
Coors Light	MillerCoors	4.15%	104
Genesee Cream Ale	High Falls Brewing	5.10%	162
Sierra Nevada Summerfest Beer	Sierra Nevada	5.00%	158
Michelob Beer	Anheuser Busch	5.00%	155
Flying Dog Doggie Style	Flying Dog Brewery	4.70%	158
Big Sky I.P.A.	Big Sky Brewing	6.20%	195

Motivation

- Is there a relationship between beers alcohol content and calories?
- How accurately can we estimate the effect of beers alcohol content on calories?
- Is the relationship linear?

Motivation



Note: There is one beer in the list that is actually considered a non-alcoholic beer—O'Doul's. This could be a potential outlier, so we remove this data point.

Simple Linear Regression

- Given the independent data $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, the simple linear regression fits the data with the following model:

$$E(y) = \beta_0 + \beta_1 x$$

  where x is called independent/predictor/explanatory variable, y is called dependent/response variable.

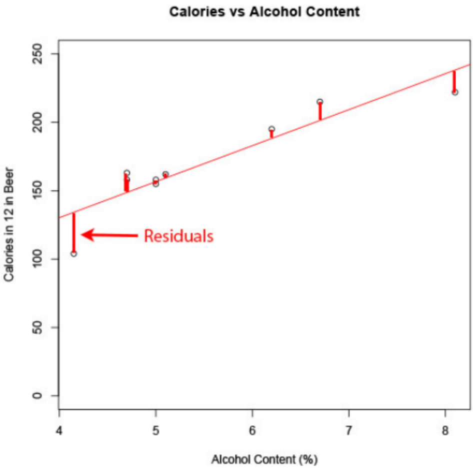
- The parameters β_0 and β_1 are estimated using the Least Square(LS) method, which is to minimize the sum of squares:

$$(y_1 - \beta_0 - \beta_1 x_1)^2 + ... + (y_n - \beta_0 - \beta_1 x_n)^2$$

Simple Linear Regression

- Based on the LSE $\hat{\beta}_0$ and $\hat{\beta}_1$, the prediction of y at $X = x_i$ can be derived as $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
- The residual for the ith data point is $e_i = y_i - \hat{y}_i$. So the minimized sum of squares is called the Residual Sum of Squares (RSS), which is:
$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + ... + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$
Then the model with smaller RSS will be the better fit.

Simple Linear Regression



Simple Linear Regression

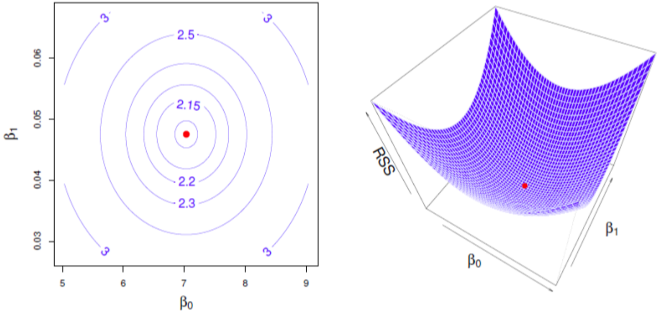


Figure: Example of contour and three-dimensional plots of the RSS.

Simple Linear Regression

By taking the derivative of RSS with respect to β_0 and β_1, the regression equation (also known as best fitting line or least squares line) can be derived with slope being $\hat{\beta}_1$ and the y-intercept being $\hat{\beta}_0$.

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_x}$$

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

where

$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

$$SS_x = \sum (x - \bar{x})^2$$

$$SS_y = \sum (y - \bar{y})^2$$

Simple Linear Regression

- Bias and Unbiasedness:
  The bias of an estimator means it might over or under estimate the truth averaging the corresponding estimates for a large number of data sets. An unbiased estimator does not systematically over- or under-estimate the true parameter. The property of unbiasedness holds for the least squares coefficient estimates.

- Standard error:
  The standard error of an estimator is standard deviation of the estimator, describing its variation due to repeated sampling. Denoted as $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$

- Confidence interval

Simple Linear Regression

Hypothesis testing — t-test

- Hypothesis:

$$H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$$

- Test statistic:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

- null distribution: Under H_0, the test statistic follows a t distribution with degrees of freedom $n - 2$.

Simple Linear Regression

- rejection rule:
  - critical value approach:
    If the test statistic derived from the observed data t is larger than $t_{critical}$, then the null hypothesis should be rejected.
  - P-value approach:
    p-value is the probability of observing any value equal to $|t|$ or larger, under the null hypothesis, which is $\beta_1 = 0$

    $$p - value_{two-sided} = 2 * P(T > |t||H_0)$$

    If p-value is less or equal to the predefined significance level, then the null hypothesis should be rejected.

Simple Linear Regression

Beer Example:

	Estimate	**Std. Error**	**t value**	**Pr($> \|t\|$)**
(Intercept)	25.031	24.999	1.001	0.350038
Beers alcohol content	26.319	4.432	5.938	0.000577***

How to interpret the results?

Simple Linear Regression

R implementation:

- lm() function can build the linear regression for you!
- Do you know how to access to the model fit?

Simple Linear Regression

Assessing the model fit(REVIEW):

- Partitioning Variation: Break down difference between observation and grand mean into two parts:

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \tag{1}$$

  - $Y_i - \bar{Y}$: Total deviation.
  - $\hat{Y}_i - \bar{Y}$: Deviation of fitted value around ground mean.
  - $Y_i - \hat{Y}_i$: Deviation around fitted value.

Simple Linear Regression

Sums of Squares:
Square both sides, and the cross-terms in $(\hat{Y}_i - \bar{Y}) * (Y_i - \hat{Y}_i)$ will cancel.

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2 \tag{2}$$

- $\sum_i (Y_i - \bar{Y})^2$: Sum of squares total.(SSTO)
- $\sum_i (\hat{Y}_i - \bar{Y})^2$: Sum of squares regression.(SSR)
- $\sum_i (Y_i - \hat{Y}_i)^2$: Residual sum of squares/RSS (Sum of squares error/SSE)

Simple Linear Regression

Assessing the model fit:

- R-square:
  R^2 measures the proportion of variability in Y that can be explained using X.
  $$R^2 = 1 - \frac{RSS}{SST}, SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
  $R^2 \in (0, 1)$.
  - An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression.
  - A number near 0 indicates that the regression did not explain much of the variability in the response; this might occur because the linear model is wrong, or the inherent error σ^2 is high, or both.
  - For simple linear regression, $R^2 = r^2$ where r is the Pearson correlation coefficient between X and Y .

Simple Linear Regression

Assessing the model fit:

- Residual standard error (RSE):
  Square root of the variance of the residuals, which is the average amount that the response will deviate from the true regression line.
  $$RSE = \sqrt{\frac{RSS}{n-2}}$$
  Lower values of RSE indicate better fit.
  Advantages:
  - Can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable.
  - RSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction.

Simple Linear Regression

Assessing the model fit:

- Beer Example:

Quantity	**Value**
Residual standard error(RSE)	15.64
R^2	0.8344

Disadvantage: R-square can only increase as predictors are added to the regression model. This increase is artificial when predictors are not actually improving the models fit.

Simple Linear Regression

Assessing the model fit

- 3. Adjusted R-square

$$\text{Adjusted } R^2 = 1 - \frac{RSS/df_{RSS}}{SST/df_{SST}}$$

Advantage: Adjusted R-squared will decrease as predictors are added if the increase in model fit does not make up for the loss of degrees of freedom. Likewise, it will increase as predictors are added if the increase in model fit is worthwhile.

Simple Linear Regression

Assessing the model fit

- 4. F-test
  - The F-test evaluates the null hypothesis that all regression coefficients are equal to zero versus the alternative that at least one is not.
  - F-test determines whether the proposed relationship between the response variable and the set of predictors is statistically reliable and can be useful when the research objective is either prediction or explanation.

Simple Linear Regression

Hypothesis test — F test:
We can test all the coefficients together.

- Hypothesis:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$
$$H_a\text{: Not all } \beta_j = 0, j = 1, ..., p$$

- Test statistic:

$$F = \frac{(SST - RSS)/p}{RSS/(n-p-1)}$$

- null distribution: Under H_0, the test statistic follows a F distribution with degrees of freedom p and $n - p - 1$.

Simple Linear Regression

Beer Example:

Quantity	Value
Adjusted R^2	0.8107
F-statistic	35.26
p-value	0.0005768

Simple Linear Regression

How can we use the estimated model—Prediction

- Find the number of calories when the alcohol content is 6.50%.

Simple Linear Regression

- Solution:
  $\hat{y} = 25.0 + 26.3 * (6.50) = 196$ calories
  If you are drinking a beer that is 6.50% alcohol content, then it is probably close to 196 calories. Notice, the mean number of calories is 170 calories. This value of 196 seems like a better estimate than the mean when looking at the original data. The regression equation is a better estimate than just the mean.

Simple Linear Regression

- Find the number of calories when the alcohol content is 2.00%.

Simple Linear Regression

- Solution:
  $\hat{y} = 25.0 + 26.3 * (2.00) = 78$ calories
  If you are drinking a beer that is 2.00% alcohol content, then it has probably close to 78 calories. This doesn't seem like a very good estimate. This estimate is what is called **extrapolation**. It is not a good idea to predict values that are far outside the range of the original data. This is because you can never be sure that the regression equation is valid for data outside the original data.

Question?

Potential problems

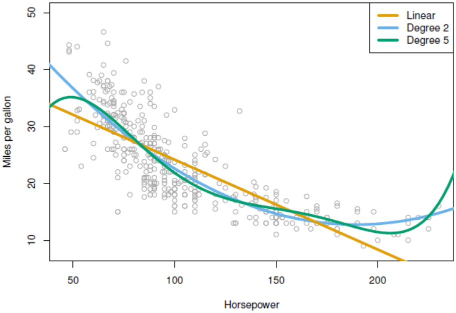
Simple Linear Regression

Assumptions:

- Linearity: There is linear relationship between beers alcohol content and calories.
- Independence: All calorie values are independent from each other.
- Homoscedasticity: The calorie values are homoscedastic.
- Normality: The distribution for each calorie value is normally distributed for every value of alcohol content in the beer.

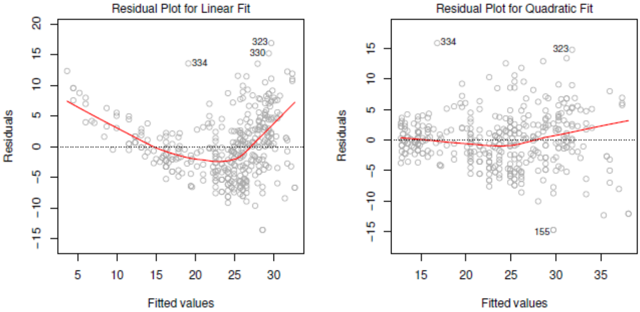
Non-linearity

Problem: Sometimes a linear relationship between response variables and predictors does not provide a good fit, even if large R^2 is achieved, such as following plot, then we need to introduce polynomial terms, such as quadratic term, cubic term, etc.



Non-linearity

Diagnosis: How to decide if we should include non-linear terms?
Residual Plot (observed versus predicted values/residuals versus predicted values)



Non-linearity

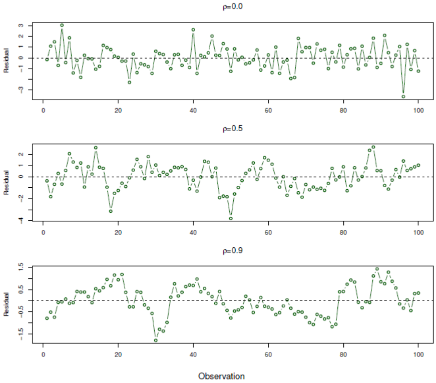
Solution:
If the residual plot indicates that there are non-linear associations in the data:

- A simple approach is to use non-linear transformations of the predictors, such as $log(x)$, $\sqrt{x}$, x^2 in the regression model.
- Add another regressor that is a nonlinear function of one of the other variables.
- Add some entirely different independent variable that explains or corrects for the nonlinear pattern or interactions among variables.

Note: Be careful for the over-fitting problem for the 2nd and 3rd solutions.

Correlation of error terms

Problem: Violations of independence are potentially very serious in time series regression models.



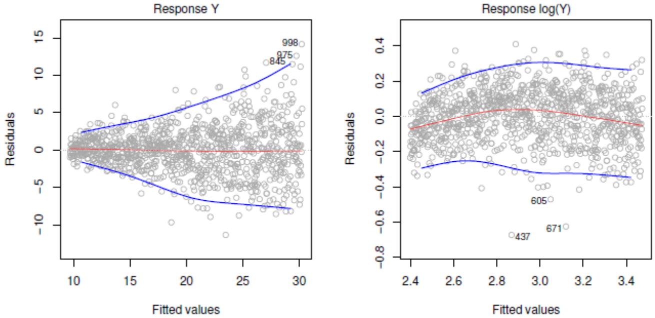
Plots of residuals from simulated time series data sets generated with different levels of correlation between error terms for adjacent time points.

Correlation of error terms

- Influence:
  The estimated standard errors will tend to underestimate the true standard errors. So the confidence and prediction intervals will be narrower.
- Diagnosis:
  Residual time series plot (residuals vs row number) and a table or plot of residual autocorrelations.
- Solution:
  Time series analysis, linear mixed effects models, etc.

Heteroscedasticity

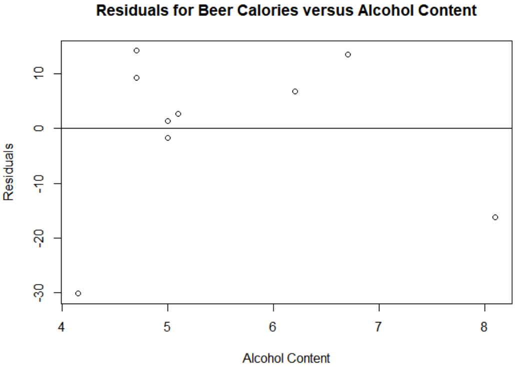
Problem: Non-constant variance of error terms:



Red line: smooth fit to the residuals to make it easier to identify a trend. Blue line: outer quantiles of the residuals, and emphasize patterns.

Heteroscedasticity

Beer example:

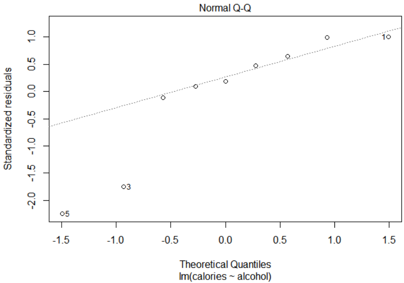


Heteroscedasticity

- Influence:
  Confidence intervals become too wide or too narrow; Giving too much weight to a small subset of the data where the error variance was largest when estimating coefficients.
- Diagnosis:
  Residuals versus predicted values plot.
- Solution—Transformation:
  Transform the response y using a concave function such as $log(y)$ or $\sqrt{y}$. Such a transformation results in a greater amount of shrinkage of the larger responses, leading to a reduction in heteroscedasticity.

Non-Normality

- Influence:
  Problematic for determining whether model coefficients are significantly different from zero and for calculating confidence intervals for forecasts.
- Diagnosis:
  QQ-plot(normal probability plot/normal quantile plot of the residuals):



Non-Normality

- Tests for normality:
  The Kolmogorov-Smirnov test, the Shapiro-Wilk test, the Jarque-Bera test, and the Anderson-Darling test.
- Potential reasons:
  Distributions of the dependent and/or independent variables are themselves significantly non-normal.
  The linearity assumption is violated.
- Solution:
  Transformation: Transform the response y using $log(y)$, $\sqrt{y}$, $\frac{1}{Y}$, or $2arcsine\sqrt{Y}$. Also some transformation method, such as Box-Cox transformation.

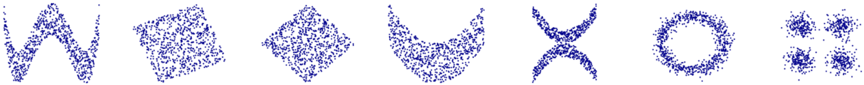
Question?

Correlation

Correlation

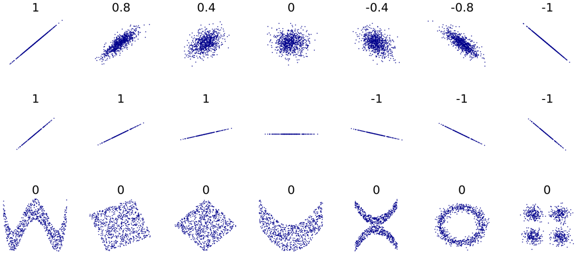
A **correlation** exists between two variables when the values of one variable are somehow associated with the values of the other variable.

When there is a pattern in the data, then there is a correlation in the data.



Correlation

- Linear correlation coefficient(Pearson correlation coefficient): Describes the strength of the linear relationship between the two variables.



- Notation:
  Population correlation coefficient:ρ
  Sample linear correlation coefficient: r

Correlation

- Expectation:
  $E(X) = \mu(X)$
- Variance:
  $Var(X) = \sigma(X)^2 = E[(X - \mu(X))^2] = E[X^2] - [E(X)]^2$
- Covariance:
  $Cov(X, Y) = E[(X - \mu(X))(Y - \mu(Y))] = E(XY) - E(X)E(Y)$
- Correlation coefficient:
  $\rho(X, Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$

Correlation

- Sample correlation coefficient:
  $r = \frac{Cov(X,Y)}{SS_{xy})\sqrt{SS_x SS_y}}$
  where
  $SS_x = \sum(x - \bar{x})^2$
  $SS_y = \sum(y - \bar{y})^2$
  $SS_{xy} = \sum(x - \bar{x})(y - \bar{y})$
- Pearson correlation coefficient is used when both variables X and Y are continuous.

Correlation

- Interpretation:
  $r \sim (-1, 1)$
  r = 1 means there is a perfect negative linear correlation.
  r = 1 means there is a perfect positive correlation.
  The closer r is to 1 or 1, the stronger the linear correlation. The closer r is to 0, the weaker the correlation.
  Note: r = 0 does not mean there is no correlation. It just means there is no linear correlation. There might be a very strong curved pattern.

Correlation

Beer example:

- X: alcohol content in the beer
- Y: calories in 12 ounce beer
- Pearson correlation coefficient: 0.913

Correlation

Causation:

- One common mistake people make is to assume that because there is a correlation, then one variable causes the other.
- For example, we can not say amount of alcohol in the beer causes it to have a certain number of calories. The fermentation of sugars is what causes the alcohol content.

Correlation

Example: A study showed a strong linear correlation between per capita beer consumption and teachers salaries. Does giving a teacher a raise cause people to buy more beer? Does buying more beer cause teachers to get a raise?

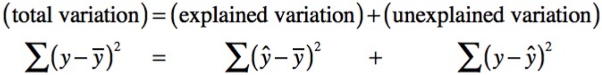
Correlation

Solution:
There is probably some other factor causing both of them to increase at the same time. Think about this: In a town where people have little extra money, they wont have money for beer and they wont give teachers raises. In another town where people have more extra money to spend it will be easier for them to buy more beer and they would be more willing to give teachers raises.

Correlation

Explained Variation:
Think of the beer example, Some of the variation in calories is due to alcohol content and some is due to other factors. How much of the variation in the calories is due to alcohol content?

Correlation



$$(\text{total variation}) = (\text{explained variation}) + (\text{unexplained variation})$$

$$\sum(y-\bar{y})^2 = \sum(\hat{y}-\bar{y})^2 + \sum(y-\hat{y})^2$$

Coefficient of determination:

$$r^2 = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2}$$

- 1. r^2 is the proportion of the variation that is explained by the model.
  2. For simple linear regression, $r^2 = (r)^2$

Correlation

Beer example:

$$r^2 = 0.8344$$
$$r = 0.913$$

Question?

Inference

Inference

Hypothesis Test for Correlation:

- 1. State the random variables:
  x = independent variable
  y = dependent variable
- 2. State the null and alternative hypotheses and the level of significance:
  $H_0 : \rho = 0$ (no correlation)
  $H_A : \rho \neq 0$ (correlation)
  or
  $H_A : \rho > 0$ (positive correlation)
  $H_A : \rho < 0$ (negative correlation)
  State significance level α.

Inference

- 3. State and check the assumptions for the hypothesis test.
  - Independence:
    All the (x, y) pairs are uncorrelated with each other.
  - Normality:
    (X, Y) follows bivariate normal distribution.
- 4. Define test statistic:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Under H_0, the test statistic has a $Student's t$-distribution with degree of freedom $n-2$. This holds approximately in case of non-normal observed values if sample sizes are large enough.

Inference

- 5. Decision rules:
  - critical value based:

$$r_{critical} = \frac{t}{n-2+t^2}$$

  - p-value based:

$$p - value_{two-sided} = 2 * P(T > |t||H_0)$$

- 6. Interpretation:
  The conclusion for a hypothesis test is that you either have enough evidence to show H_A is true, or you do not have enough evidence to show H_A is true.

Inference

Beer Example:

- t = 5.9384
- p-value = 0.0002884

Correlation and Regression Analysis Example

The following table contains randomly selected high temperatures at various cities on a single day and the elevation of the city.

Elevation (in feet)	7000	4000	6000	3000	7000	4500	5000
Temperature (F)	50	60	48	70	55	55	60

Inference

- 1. What are the random variables?
- 2. Find a regression equation for elevation and high temperature on a given day.
- 3. Find the residuals and create a residual plot.
- 4. Find the correlation coefficient and coefficient of determination and interpret both.
- 5. Is there enough evidence to show a negative correlation between elevation and high temperature? Test at the 5% level.

Question?

- For hands on help with your analyses, stop by our drop in hours or sign up for a consultation.

- Welcome to the workshops in the Fall quarter!

- If you have any workshop requests, now is the time to ask! We will be setting our fall schedule soon.

- For more details, visit our website:
  GradQuant.ucr.edu

Thank You
Welcome to GradQuant