Exam Revision 3

In Son Zeng

University of Michigan

insonz@umich.edu

December 18, 2018

Overview

Office Hours Arrangement

1. My office hours for Monday (17th December) and Tuesday (18th December) are: Monday 16:00 - 17:00, Tuesday 15:30 - 18:00.
2. My office hours will be held at the same place: **USB 2165**. Please come with me with any test concerns and conceptual questions.
3. I will be wearing master's graduation gown on Tuesday before and after proctoring the exam 3. I welcome everyone to take graduation photographs with me before and after the exam 3.
4. There will be an exit survey to collect your constructive feedback for our class. Please spend some time before this Thursday night to filling this survey. Hope you enjoy the class and will love statistics in the future.

Confidence Interval for population mean (One sample)

Steps of constructing confidence interval for population mean:

1. The first step is the define the parameter correctly! The parameter should be the **population/true mean** of the subject mentioned in the question.
2. The second step is to check **randomness** and Underlying population distribution (approximately) normal (**UPDN**) of the sample. First, we check whether the sample is random. If it is told, great! Just move on! If not, we can explain (most of the time) by your own reasoning why you think the collect sample is random or not.
3. Then, we check normality. If you are told UPDN, just proceed to next step. If not, then we check whether the sample size is large enough $n \geq 30$ to employ the CLT, which claims that the sampling distribution of the sample mean of the measurements is approximately normal. If the sample size is small $n < 30$, then we rely on the robustness of the t procedures against violations of normality.

Confidence Interval for population mean (One sample)

Steps of constructing confidence interval for population mean:

1. If σ is given, skip this part. Otherwise, compute $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$, which is the square root of the sample variance.
2. **Computation** Find the t-score or z-score by referencing the table according to the specified significance level. For two-tails (keyword: **between**,$\pm$) , we find $t_{\frac{\alpha}{2}}$ or $z_{\frac{\alpha}{2}}$; for one-tail (keyword: **no greater than, no less than**), we derive the confidence upper bound (CUB) or confidence lower bound (CLB) by finding t_α or z_α
3. **Conclusion** We are approximately $100(1-\alpha)\%$ confident that the population mean of .....what question say..... is (between /no greater than/no smaller than) ...the confidence interval....

Hypothesis Testing for population mean (One sample)

Steps of performing hypothesis testing for population mean:

1. The first step is the define the parameter correctly! The parameter should be the **population/true mean** of the subject.
2. The second step is the set up null hypothesis and alternative hypothesis. For two-tails, $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ and for one-tail, we either
3. The second step is to check **randomness** and Underlying population distribution (approximately) normal (**UPDN**) of the sample.

Hypothesis Testing for population mean (One sample)

Steps of performing hypothesis testing for population mean:

1. We specify the significance level α, the values are usually 0.05 or 0.01.
2. We calculate the t-score with df $n-1$ by $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$
3. We check the t-table for the corresponding p-value given t-score you derived. If the question is asking two-tails, you need to multiply the p-value by 2. (No multiplication for one-tail)
4. Conclusion: With a p-value lower/greater than the significant level (such as 0.05, 0.01), we reject/fail to reject H_0. There is/is not sufficient evidence to suggest that the population mean of ...what question say... is (different from/greater than/smaller than) ...the number....

Important Comparison

Comparison of **difference between population means** and **population mean difference**:

1. Parameter: The parameter $\mu_X - \mu_Y$ is for difference between population means, while the parameter μ_D is for population mean difference.
2. Hypothesis: $H_0 : \mu_X - \mu_Y = \mu_0$ *vs* $H_1 : \mu_X - \mu_Y \neq \mu_0$ (or one-tail hypothesis) is for difference between population means, while $H_0 : \mu_D = \mu_0$ *vs* $H_1 : \mu_D \neq \mu_0$ is for population mean difference.
3. Theoretical Constructions: For **difference between population means**, we are taking the mean of two separate samples (independent of each other), and taking the difference between them. For **population mean difference**, we have a paired sample, for each observation in the sample, we take the difference between two measures, and then take the mean.

Important Comparison

Comparison of **difference between population means** and **population mean difference**:

1. More Theoretical Constructions: To estimate the **difference between population means**, we have $\bar{X}$ and $\bar{Y}$ for the two separate samples, so the sample statistics is $\bar{X} - \bar{Y}$. Therefore, the parameter is $\mu_X - \mu_Y$.
2. To estimate the **population mean difference**, we take $d_i = X_i - Y_i$ for each observation in the paired sample, then take the mean of them, which is $\bar{D} = \frac{\sum_{i=1}^{n} d_i}{n}$. This is our sample statistics called sample mean difference. Therefore, the parameter is $\mu_D = \mu_{X-Y}$.
3. Now, are you confident to say you are clear about the difference? Some students ask this problem in Piazza, could any of you rephrase my comparison in Piazza?

Hypothesis Testing for Difference between two population means

Steps of performing hypothesis testing for the difference between two population means:

1. The first step is the define the parameter correctly! The parameter should be the $\mu_X - \mu_Y$, the **difference between the population/true mean of A and the population/true mean of B**.
2. The second step is the set up null hypothesis and alternative hypothesis. For two-tails, $H_0 : \mu_X - \mu_Y = \mu_0$ vs $H_1 : \mu_X - \mu_Y \neq \mu_0$ and for one-tail, replace the equal sign with $\leq$ for upper-tail and $\geq$ for lower-tail.
3. The second step is to check **randomness** and Underlying population distribution (approximately) normal (**UPDN**) of the sample.

Hypothesis Testing for Difference between two population means

Steps of performing hypothesis testing for the difference between two population means:

1. If the two separate samples are given random, and the two samples are independent of each other, you are good to go! If not, we need to assume independence within the observations in two samples and between samples to proceed.
2. If the underlying distribution is given normal, you are good to go! If not, we first check the sample size of two samples. If the sample size is large, then we can use the CLT to say that the sampling distribution for the difference between sample means is approximately normal. If one of the sample has small sample size, we cannot use CLT and need to rely on the robustness of t-distribution and perform the two sample t-test.

Hypothesis Testing for Difference between two population means

Steps of performing hypothesis testing for the difference between two population means:

1. Sample statistics: First we compute the degree of freedom for two sample t-test without assuming equal variances:

$$v = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{1}{n_X - 1} \cdot \left(\frac{s_X^2}{n_X}\right)^2 + \frac{1}{n_Y - 1} \cdot \left(\frac{s_Y^2}{n_Y}\right)^2} \tag{1}$$

2. We typically round the degree of freedom down to the nearest integer.
3. Sample statistics: We derive the t-statistics by:

$$t = \frac{(\bar{x} - \bar{y}) - \mu_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \tag{2}$$

Hypothesis Testing for Difference between two population means

Steps of performing hypothesis testing for the difference between two population means:

1. Upon finding the t-value we refer back to the t-table and find the range of the corresponding p-value. If the hypothesis testing is one-tail, we just use that p-value; if the hypothesis testing is two-tail, we need to multiple the p-value by 2.
2. Conclusion for two-sample hypothesis testing: With a p-value lower than the significant level (such as 0.05, 0.01), we reject H_0. There is sufficient evidence to suggest that the **population mean** of A is (different from/greater than/smaller than) the **population mean** of B.
3. With a p-value greater than the significant level (such as 0.05, 0.01), we fail to reject H_0. There is not sufficient evidence to suggest that the **population mean of A** is (different from/greater than/smaller than) the **population mean of B**.

Confidence Interval for Difference between two population means

Steps of constructing confidence interval for the difference between two population means (same defining parameter and checking condition as above):

1. Find the t-score or z-score by referencing the table, given the specified significance level. For two-tails (keyword: **between**,$\pm$) , we find $t_{v,\frac{\alpha}{2}}$ or $z_{\frac{\alpha}{2}}$, and the confidence interval is (most often)

$$\boxed{(\bar{x}-\bar{y}) \pm t_{v,\frac{\alpha}{2}} \cdot \sqrt{\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}}} \tag{3}$$

Confidence Interval for Difference between two population means

Steps of constructing confidence interval for the difference between two population means (same defining parameter and checking condition as above):

1. For one-tail (keyword: **no greater than, no less than**), we find the confidence upper bound (CUB) or confidence lower bound (CLB) by finding $t_{v,\alpha}$ or z_α. The confidence upper bound is $(\bar{x}-\bar{y})+t_{v,\alpha}\cdot\sqrt{\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}}$ and the confidence lower bound is $(\bar{x}-\bar{y})-t_{v,\alpha}\cdot\sqrt{\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}}$
2. Conclusion for confidence interval: We are approximately (95%/99%) confident that the **difference between population means** of ....what question say..... is (between /no greater than/no smaller than) ...the result....

Hypothesis Testing for Population Mean Difference

Steps of performing hypothesis testing for population mean difference:

1. We specify our parameter of interest as μ_D, which is the **population mean difference between ..... (a paired sample)**
2. Hypothesis: In most cases we construct the hypothesis as the follows: $H_0 : \mu_D = \mu_0$ vs $H_1 : \mu_D \neq \mu_0$ (two tails). For one-tail cases, if we want to test (alternative hypothesis) whether the population mean difference of A and B is greater than 0, for example, then our hypothesis becomes $H_0 : \mu_D \leq 0$ vs $H_1 : \mu_D > 0$
3. Checking Conditions: If the paired samples are given random, you can proceed to check normality! If not, we can argue by reasoning that the paired samples are impacting each other (not independent) or not impacting each other (independent). Most of the time, we can draw a scatterplot to check randomness because the samples are given. Checking randomness is important because we need randomness of the paired samples to perform either t-test or z-test.

Hypothesis Testing for Population Mean Difference

Steps of performing hypothesis testing for population mean difference:

1. Checking Conditions: If the underlying distribution of the two paired samples are given normal, you can proceed to derive the test statistics! If not, we first need to check the sample size of the two paired samples (they should have the same sample size n). If n is large, i.e., $n \geq 30$, we can employ the CLT to say that the sampling distribution for the population mean difference is approximately normal.
2. • If n is small, then we cannot use CLT. Instead, we may use QQ-plot (or perform normality test) based on the data to see if the paired samples seem to be approximately normal in distribution (most of the time the paired data are given). If the samples do not show terrible outliers (check class note 20 for detail), then we rely on the robustness of the student-t distribution (and/or proceed with caution) against the mild violations of normality.

Hypothesis Testing for Population Mean Difference

Steps of performing hypothesis testing for population mean difference:

1. Sample statistics: First we specify the significance level α based on the quantity given in the question, which is often 0.05, 0.02, or 0.01. If we perform the t-test, the degree of freedom for population mean difference is $n-1$.
2. Then, we compute the sample standard deviation of the differences s_D. If we are required to compute by hand, we first compute the difference between two paired samples $(d_1, ..., d_n)$ and take the mean $\bar{D}$. Then we compute the sample standard deviation of the differences s_D as follows:

$$\boxed{s_D = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{D})^2}{n-1}}} \tag{4}$$

3. Most of the time we look at the sd for the population mean difference in the outputs.

Hypothesis Testing for Population Mean Difference

Steps of performing hypothesis testing for population mean difference:

1. Sample statistics: Then, we derive the t-statistics by:

$$\boxed{t = \frac{\bar{D} - \mu_0}{s_D/\sqrt{n}}} \tag{5}$$

2. Upon finding the t-value we refer back to the t-table and find the range of the corresponding p-value. Since the standard student-t distribution is symmetric (taught in chapter 4), if the hypothesis testing is one-tail, we just use that p-value; if the hypothesis testing is two-tail, we need to multiple the p-value by 2.
3. Conclusion: With a p-value lower/greater than the significant level (such as 0.05, 0.01), we reject/fail to reject H_0, the null hypothesis. Therefore, we (have/do not have) sufficient evidence to suggest that the **population mean difference of A and B** is (different from/greater than/smaller than) μ_0.

Confidence Interval for Population Mean Difference

Steps of constructing confidence interval for population mean difference:

1. We specify our parameter of interest as μ_D, which is the **population mean difference between ..... (a paired sample)**
2. Checking Conditions: If the paired samples are given random, you can proceed to check normality! If not, we can argue by reasoning that the paired samples are impacting each other (not independent) or not impacting each other (independent). Most of the time, we can draw a scatterplot to check randomness because the samples are given. Checking randomness is important because we need randomness of the paired samples to perform either t-test or z-test.

Confidence Interval for Population Mean Difference

Steps of constructing confidence interval for population mean difference:

1. Checking Conditions: If the underlying distribution of the two paired samples are given normal, you can proceed to derive the test statistics! If not, we first need to check the sample size of the two paired samples (they should have the same sample size n). If n is large, i.e., $n \geq 30$, we can employ the CLT to say that the sampling distribution for the population mean difference is approximately normal.
2. • If n is small, then we cannot use CLT. Instead, we may use QQ-plot (or perform normality test) based on the data to see if the paired samples seem to be approximately normal in distribution (most of the time the paired data are given). If the samples do not show terrible outliers (check class note 20 for detail), then we rely on the robustness of the student-t distribution (and/or proceed with caution) against the mild violations of normality.

Confidence Interval for Population Mean Difference

Steps of constructing confidence interval for population mean difference:

1. Sample statistics: If we perform the t-test, the degree of freedom for population mean difference is $n - 1$.
2. Then, we compute the sample standard deviation of the differences s_D. If we are required to compute by hand, we first compute the difference between two paired samples $(d_1, ..., d_n)$ and take the mean $\bar{D}$. Then we compute the sample standard deviation of the differences s_D as follows:

$$\boxed{s_D = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{D})^2}{n-1}}} \tag{6}$$

3. Most of the time we look at the sd for the population mean difference in the outputs.

Confidence Interval for Population Mean Difference

Steps of constructing confidence interval for population mean difference:

1. Sample statistics: The significance level α should be 1 minor the percentage given in the question. We refer to the t-table to search for the corresponding value for computation.
2. For two-tails (keyword: **between**) , we find $t_{n-1,\frac{\alpha}{2}}$ and the confidence interval is:

$$\boxed{\bar{D} \pm t_{n-1,\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}}} \tag{7}$$

3. For one-tail (keyword: **no greater than, no less than**), we find $t_{n-1,\alpha}$. Then, the confidence upper bound (CUB) and the confidence lower bound (CLB) are respectively:

$$\boxed{\bar{D} + t_{n-1,\alpha} \cdot \frac{s_d}{\sqrt{n}} \quad , \quad \bar{D} - t_{n-1,\alpha} \cdot \frac{s_d}{\sqrt{n}}} \tag{8}$$

Confidence Interval for Population Mean Difference

Steps of constructing confidence interval for population mean difference:

1. Conclusion: • We are approximately (95 percent/ 98 percent / 99 percent) confident that the **population mean difference** in ....what question say..... is (between /no greater than/no smaller than) ...the result....

Regression

1. Least-square Line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
2. Coefficient of Determination Interpretation: If $R^2 = 0.75$, we claim that 75% of the observed variation in the model can be explained by the simple linear regression relationship between (the independent variable) and (the dependent variable).
3. Correlation Coefficient: We have $R^2 = r^2 \rightarrow r = \sqrt{R^2}$. However, we need to be careful here. If the slope is positive/negative, then we take the positive/negative square root. If the slope is not available, you may look at the scatterplot to determine whether r is positive or negative.
4. Remember, both R^2 and r only work for evaluating linear relationships.

Regression

Simple Linear Model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ *or* $Y = \beta_0 + \beta_1 X + \epsilon$
Normal Assumption for Linear Model:

1. We need the observations are independent.
2. We need the errors $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are random and independent. In particular, the magnitude of any error term does not influence the value of the other error terms. Again, if the error terms appear to be monotone increasing or decreasing, U-shape or inverted-U-shape, the assumption for linear models is violated.
3. We need the error terms to be normally distributed, with mean 0 and with same variance (homoscedasticity). Notationally speaking, we need $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$
4. If the normal assumption is fulfilled, we can further have the response variable y_i follows normal distribution. Notationally speaking, $y_i \overset{\text{i.i.d.}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$, for $i = 1, 2, ..., n$.

Regression

Intercept and Slope for regression:

1. Slope: $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
2. Intercept: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$
3. Standard Error for slope:
   $SE(\hat{\beta}_1) = \sqrt{Var(\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2})} = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$
4. Standard Error for intercept:
   $SE(\hat{\beta}_0) = \sqrt{Var(\bar{y} - \hat{\beta}_1\bar{x})} = \sigma \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$
5. Estimate of variance: $\hat{\sigma}^2 = s^2 = MSE = (RMSE)^2 = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$

Regression

Hypothesis Tests and Confidence Intervals for the Intercept:

1. For Intercept, our hypothesis is $H_0 : \beta_0 = 0$ *vs* $H_1 : \beta_0 \neq 0$.
2. We perform standardization for $\beta_0 \rightarrow \frac{Observed - Mean}{SD} = \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)}$. This value should follow the t-distribution with $n-2$ degree of freedom, because the degree of freedom is computed by (sample size - number of coefficients estimated β_0, β_1), which is $n-2$.
3. If the p-value is smaller than the significance level α, we reject the null hypothesis and we have sufficient evidence to conclude that the intercept is significantly different from 0.
4. The $100(1-\alpha)\%$ confidence interval (you know how to derive the CUB and CLB already) of intercept, we have: $\hat{\beta}_0 \pm t_{n-2,\frac{\alpha}{2}} \cdot SE(\hat{\beta}_0)$. Then, we are **approximately** $100(1-\alpha)\%$ confident that the intercept value falls (between/below/above) .....

Regression

Hypothesis Tests and Confidence Intervals for the Slope:

1. For Slope, our hypothesis is $H_0 : \beta_1 = 0$ *vs* $H_1 : \beta_1 \neq 0$.
2. Again, we perform standardization for $\beta_1 \rightarrow \frac{Observed - Mean}{SD} = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$. This value also follow the t-distribution with $n - 2$ degree of freedom.
3. If the p-value is smaller than α, we reject H_0 and we have sufficient evidence to conclude that the slope is significantly different from 0. In this case, the predictor is significant in our linear model.
4. On the other hand, if the p-value is greater than the α, we fail to reject H_0 and we do not have sufficient evidence to conclude that the slope is significantly different from 0. In this case, the predictor can be dropped in our linear model.
5. The $100(1 - \alpha)\%$ confidence interval of slope, we have: $\hat{\beta}_1 \pm t_{n-2,\frac{\alpha}{2}} \cdot SE(\hat{\beta}_1)$. Then, we are **approximately** $100(1 - \alpha)\%$ confident that the slope value falls (between/below/above) .....

Regression

Inference for Mean Response:

1. Recap: $Y|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, $Y|x^* \sim N(\beta_0 + \beta_1 x^*, \sigma^2)$.
2. $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^* \rightarrow E[\hat{Y}] = \beta_0 + \beta_1 x^*$, $Var(\hat{Y}) = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]$
3. Now for the **estimated standard deviation** of $\hat{Y}$, we have $s_{\hat{Y}} = s \cdot \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$, $s = RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{(1-r^2)\sum_{i=1}^n (y_i - \bar{y})^2}{n-2}}$
4. Therefore, for hypothesis testing, we use the test statistics $\frac{\hat{Y} - (\beta_0 + \beta_1 x^*)}{s_{\hat{Y}}} \sim t_{n-2}$, remember we have degree of freedom $n-2$ because we are estimating intercept β_0 and slope β_1.
5. To construct $100(1-\alpha)\%$ confidence interval for mean response, we use $(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2}, n-2} \cdot s_{\hat{Y}} =$ $(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2}, n-2} \cdot \sqrt{MSE \cdot [\frac{1}{n} + \frac{(x^* - \bar{x})^2}{Sum\ of\ Square}]}$

Regression

Inference for Future Observations (Prediction Interval):

1. Error in prediction is calculated by
   $y - \hat{y} = (\beta_0 + \beta_1 x^* + \epsilon) - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$.
2. Since the future observation $Y_{new} = \beta_0 + \beta_1 x^* + \epsilon$ is assumed to be independent of the observed values, we have $cov(Y_{new}, \hat{Y}) = 0$.
   Therefore, the error of prediction has variance:
   $Var(Y_{new} - \hat{Y}) = \sigma^2 \cdot [1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}]$
3. Therefore, the **estimated standard deviation** of Y_{new} is
   $s_{Y_{new}} = \sqrt{s_{\hat{Y}}^2 + s^2} = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} =$
   $(RMSE) \cdot \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{Sum\ of\ Square}}$

Regression

Inference for Future Observations (Prediction Interval):

1. Therefore, for hypothesis testing, we use the test statistics $\frac{Y_{new}-(\hat{\beta}_0+\hat{\beta}_1 x^*)}{s_{Y_{new}}} \sim t_{n-2}$, remember we have degree of freedom $n-2$ because we are estimating intercept β_0 and slope β_1.
2. To construct $100(1-\alpha)\%$ prediction interval for future observations, we use

$$(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2},n-2} \cdot s_{Y_{new}}$$

$$= (\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2},n-2} \cdot \sqrt{MSE \cdot [1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{Sum\ of\ Square}]}$$

Useful Links

Let us watch some videos to revise the assumptions for various statistical concepts before the exam:

1. Important comparisons: https://youtu.be/k30AT-csyf0
2. Regression part 1: https://youtu.be/2ShMzFyInW0
3. Regression part 2: https://youtu.be/5H66cNRQVZk
4. Regression part 3: https://youtu.be/RSZbqHQQhIo
5. Regression Diagnostics:

Extra Questions

Question 1

Suppose the nutrition label of apple cider says that apple cider contains an average concentration of 90 grams of sugar per liter, with standard deviation of 10 grams of sugar per liter. Now we want to check whether the claim is true. Therefore, we bought 50 liters of apple cider and found that 5050g sugar are contained in the 50 liters of apple cider in total.

Based on this information, could we perform a hypothesis testing to check whether the claim in the nutrition label is true, at significance level $\alpha = 0.05$? If yes, please perform the hypothesis testing based on the steps taught in class. If not , please provide reason detailing why we could not perform such a hypothesis testing.

Extra Questions

Question 1 Solution

1. Let μ be the population mean concentration of sugar contained in a liter of apple cider.
2. The null hypothesis is $H_0 : \mu = \mu_0 = 90$ grams versus $H_1 \neq 90$ grams.
3. Check randomness: the sample is not given random, but we believe that each liter of apple cider will not affect the other samples (independent).
4. Check UPDN: UPDN is not given. We check the sample size: $n = 50 \geq 30$, so we can use CLT and claim that the sample is approximately normal. We still use the t-test for the calculation given the degree of freedom $n - 1 = 49$.

Extra Questions

Question 1 Solution

1. We specify the significance level $\alpha = 0.05$. We also have known sample mean $\bar{X} = \frac{5050}{50} = 101$ grams/liter
2. The t-statistics is: $t = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{101-90}{10/\sqrt{50}} = 7.778$
3. By $1 - pt(7.778, 49)$ in R (Check the table during the test), and multiply the p-value by 2, we get the p-value 4.2×10^{-10}
4. Conclusion: Since the p-value is lower than 0.05, we reject the null hypothesis. There is sufficient evidence to suggest that the population mean concentration of sugar contained in a liter of apple cider is different from 90 grams/liter.

Extra Questions

Question 2

Suppose the nutrition label of apple cider says that apple cider contains an average concentration of 90 grams of sugar per liter, but the standard deviation is not given. Now we bought 10 liters of apple ciders instead, and found the concentrations of sugar for each liter of apple cider as follows: 95.3, 101.2, 92.3, 90.1, 96.4, 99.2, 110.3, 103.4, 91.2, 89.9.

Based on this information, could we perform a hypothesis testing to check whether the claim in the nutrition label is true, at significance level $\alpha = 0.05$? If yes, please perform the hypothesis testing based on the steps taught in class. If not , please provide reason detailing why we could not perform such a hypothesis testing.

Extra Questions

Question 2 Solution

1. Again, let μ be the population mean concentration of sugar contained in a liter of apple cider.
2. The null hypothesis is $H_0 : \mu = \mu_0 = 90$ grams versus $H_1 \neq 90$ grams.
3. Check randomness: the sample is not given random, but we believe that each liter of apple cider will not affect the other samples (independent).
4. Check UPDN: UPDN is not given. We check the sample size: $n = 10 \leq 30$, so we cannot use CLT and claim that the sample is approximately normal. We need to rely on the robustness of t-test for the calculation given the degree of freedom $n - 1 = 9$.

Extra Questions

Question 2 Solution

1. We specify the significance level $\alpha = 0.05$. We also have known sample mean $\bar{X} = \frac{95.3+...+89.9}{10} = 96.93$ grams/liter and sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} = 7.00$
2. The t-statistics is: $t = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{96.93-90}{7.00/\sqrt{10}} = 3.13$
3. By $1 - pt(3.13, 9)$ in R (Check the table during the test), and multiply the p-value by 2, we get the p-value 0.0121.
4. Conclusion: Since the p-value is lower than 0.05, we reject the null hypothesis. There is sufficient evidence to suggest that the population mean concentration of sugar contained in a liter of apple cider is different from 90 grams/liter.

Extra Questions

Question 3

Many people are concerned that the average grade for STATS 250 in Winter 2018 semester was significantly lower than the average grade in previous semesters. Particularly, the grade distributions (assume that the first column is the Winter 2018 data, and the second column is the sample mean over 30 semesters) are as follows:

Grade	Observed Counts	Expected Counts	X^2 Contribution
As	468.52	684.76	68.28631579
Bs	720.8	728.008	0.071366337
Cs	468.52	275.706	134.8437778
Ds	108.12	75.684	13.90114286
Es	36.04	36.04	0
Total	1802	1800.2	**217.1026028**

Extra Questions

Question 3 (Continue)

a) If the standard deviation of the number of students getting A's is 220, what is the 99 percent confidence interval of the true mean of the student getting the A's? (Since we have data from 30 semesters, the sample size should be 30 here)
b) What is the p-value of the 468.52 students getting A's in the Winter 2018 semester? What is the interpretation of such a p-value?
c) If the standard deviation of the number of students getting B's is 130, what is the 95 percent confidence interval of the true mean of the student getting the B's?
d) What is the p-value of the 720.8 students getting B's in the Winter 2018 semester? What is the interpretation of such a p-value?
e) Could we use the average grade of STATS 250 to make inference to the average grade of all classes in the University of Michigan? Why or why not?

Extra Questions

Question 3 Solution:

a) If the standard deviation of the number of students getting A's is 220, what is the 99 percent confidence interval of the true mean of the student getting the A's? (Since we have data from 30 semesters, the sample size should be 30 here)
b) What is the p-value of the 468.52 students getting A's in the Winter 2018 semester? What is the interpretation of such a p-value?

1. Let μ be the population mean number of students getting A's.
2. Check randomness: the sample is not given random, but we believe that each semester grade distribution will not affect the other samples (independent).
3. Check UPDN: UPDN is not given. We check the sample size: $n = 30 \geq 30$, so we can use CLT and claim that the sample is approximately normal. We still use the t-test for the calculation given the degree of freedom $n - 1 = 29$.

Extra Questions

Question 3 Solution:

1. From the technology $qt(0.995, 29) = 2.756$, we have the confidence interval
$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} = 684.76 \pm 2.756 \cdot \frac{220}{\sqrt{30}} = 684.76 \pm 110.71 = (574.05, 795.47)$
2. Conclusion: Hence, we are approximately 99% confident that the population mean number of students getting A's is between 574.05 and 795.47 (approximately between 574 to 795 students).

Extra Questions

Question 3 Solution:

1. With the same setting, $t = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{468.52-684.76}{220/\sqrt{30}} = -5.384$.
2. From technology, the corresponding p-value is 4.37×10^{-6} (lower-tail), which means that the probability of observing a semester with 468.52 or even less students getting A's in STATS 250 is 4.37×10^{-6}.
3. Alternatively, we can use two-tail, which gives p-value 8.74×10^{-6}, which means that the probability of observing a semester as extreme as, or more extreme than, 468.52 students getting A's in STATS 250 is 8.74×10^{-6}.

Extra Questions

Question 3 Solution:

c) If the standard deviation of the number of students getting B's is 130, what is the 95 percent confidence interval of the true mean of the student getting the B's?
d) What is the p-value of the 720.8 students getting B's in the Winter 2018 semester? What is the interpretation of such a p-value?

1. Let μ be the population mean number of students getting B's.
2. Check randomness: the sample is not given random, but we believe that each semester grade distribution will not affect the other samples (independent).
3. Check UPDN: UPDN is not given. We check the sample size: $n = 30 \geq 30$, so we can use CLT and claim that the sample is approximately normal. We still use the t-test for the calculation given the degree of freedom $n - 1 = 29$.

Extra Questions

Question 3 Solution:

1. From the technology $qt(0.975, 29) = 2.045$, we have the confidence interval $\bar{x} \pm t_{\frac{\alpha}{2}, n-1} = 728.008 \pm 2.045 \cdot \frac{130}{\sqrt{30}} = 728.008 \pm 48.537 = (679.47, 776.55)$
2. Conclusion: Hence, we are approximately 95% confident that the population mean number of students getting B's is between 679.47 and 776.55 (approximately between 679 to 777 students).

Extra Questions

Question 3 Solution:

1. With the same setting, $t = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{720.8-728.008}{130/\sqrt{30}} = -0.304$.
2. From technology, the corresponding p-value is 0.382 (lower-tail), which means that the probability of observing a semester with 728.008 or even less students getting B's in STATS 250 is 0.382.
3. Alternatively, we can use two-tail, which gives p-value 0.764, which means that the probability of observing a semester as extreme as, or more extreme than, 728.008 students getting B's in STATS 250 is 0.764.

Extra Questions

Question 3 Solution:

e) Could we use the average grade of STATS 250 to make inference to the average grade of all classes in the University of Michigan? Why or why not?

1. The students studying STATS 250 only represents a subset of population, which is all students in University of Michigan. Therefore, we cannot use the average grade of STATS 250 to make inference to the population mean grade of all class in the University of Michigan.

Extra Questions

Question 4 (Data)

The following table shows the exam 1 and exam 2 evaluation (in 100 point scale) for our 189 students. Please answer the following questions:

	Min	quan25	Median	Mean	quan75	Max	Sample SD
Exam 1	0.00	74.67	82.0	80.588942	89.33	100.00	12.37404
Exam 2	44.17	80.00	90.0	86.754921	95.83	100.00	11.51831
Exam 2 – Exam 1	-27.34	-0.50	6.5	6.165979	12.17	94.17	13.68002

a) What are the sample sizes for population A and population B?
b) What are the conditions our data should fulfill to allow us to draw inference for the population mean difference?
c) What are the conditions our data should fulfill to allow us to draw inference for the difference between two population means?

Extra Questions

Question 4 (Questions)

Using these data, answer the following questions:

1. d) Is there evidence at the 0.05 and 0.01 significance level that the population mean difference in the test score of exam 2 over exam 1 is greater than 0?
2. e) Construct a 95% and 99% confidence interval for the population mean difference between exam 1 and exam 2.
3. f) Is there evidence at the 0.05 and 0.01 significance level that the mean test score of exam 2 is different from that of exam 1?
4. g) Construct a 99% confidence interval and a 99% confidence upper bound and confidence lower bound for the increase of mean test score in exam 2, compared to exam 1.
5. h) Compare the two 99% confidence intervals, which one is wider? What are possible reasons why one of the confidence interval is wider?

Extra Questions

Question 4 (Solution)

a) The sample sizes for population A and population B are both 189.

b) To draw inference for the population mean difference, we need to have a paired sample (we have), we need to check randomness: every individual in our sample should be independent of each other. (Note that normality condition can be fulfilled by sample size here)

c) To draw inference for the difference between two population means, we need to have the individuals in Exam 1 sample are independent of each other, individuals in Exam 2 sample are independent of each other, and two samples are independent. (Note that normality condition can be fulfilled by sample size here)

Extra Questions

Question 4 (Solution)

d) Perform hypothesis testing for population mean difference:

1. We specify our parameter of interest as μ_D, which is the **population mean difference** in the test score of exam 2 over exam 1.
2. Hypothesis: We construct the hypothesis as the follows: $H_0 : \mu_D \leq 0$ vs $H_1 : \mu_D > 0$
3. Checking Randomness Conditions: The paired sample is not given random but we can assume that the change of test score of exam 2 over exam 1 for one individual will not affect that for other individuals. As such, our paired sample should be independent.
4. Checking Normal Conditions: The underlying distribution of the paired samples are not given normal. However, since $n = 189 > 30$ is large, we can employ the CLT to say that the sampling distribution for the population mean difference is approximately normal.

Extra Questions

Question 4 (Solution)

d) Perform hypothesis testing for population mean difference:

1. To be safe, we still rely on the robustness of the student-t distribution (and/or proceed with caution) against the mild violations of normality.
2. Sample statistics: Then, we derive the t-statistics by:

$$\boxed{t = \frac{\bar{D} - 0}{s_D/\sqrt{n}} = \frac{6.5}{13.68} \cdot \sqrt{189} = 6.532} \tag{9}$$

3. We refer to the t-table and find the p-value $P(t_{188} > 6.532) = 2.92 * 10^{-10}$. We just use that p-value.
4. Conclusion: With a p-value lower than both the significant levels 0.05 and 0.01, we reject H_0. Therefore, we have sufficient evidence to suggest that the **population mean difference** in the test score of exam 2 over exam 1 is different from 0.

Extra Questions

Question 4 (Solution)

e) Construct confidence interval for population mean difference:

1. For two-tails (keyword: **between**) , we find $t_{n-1,\frac{\alpha}{2}} = t_{188,0.025} = 1.973$ and $t_{188,0.005} = 2.602$. Therefore the 95% and 99% confidence intervals for the population mean difference between exam 1 and exam 2 are:

$$\bar{D} \pm t_{n-1,\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}} = 6.5 \pm 1.973 \cdot \frac{13.68}{\sqrt{189}} = (4.536, 8.463)$$

$$\bar{D} \pm t_{n-1,\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}} = 6.5 \pm 2.602 \cdot \frac{13.68}{\sqrt{189}} = (3.911, 9.089)$$

2. Conclusion: We are approximately 95% and 99% confident that the population mean difference for exam 1 and exam 2 falls between (4.536, 8.463) and (3.911, 9.089), respectively.

Extra Questions

Question 4 (Solution)

f) Perform hypothesis testing for difference between population means:

1. We specify our parameter of interest as $\mu_X - \mu_Y$, the **difference between population means** of the test score of exam 2 and that of exam 1.
2. Hypothesis: We construct the hypothesis as the follows:
   $H_0 : \mu_X - \mu_Y = 0$ vs $H_1 : \mu_X - \mu_Y \neq 0$
3. Checking Randomness Conditions: The two separate samples are not given random conditions. However, we can assume that the test score of exam 1 and the test score of exam 2 for one individual will not affect that for other individuals. Also, we assume that the test score of exam 1 score will not affect the test score of exam 2. (Seems awkward, but we need to check these conditions) As such, our two samples should be independent.

Extra Questions

Question 4 (Solution)

f) Perform hypothesis testing for difference between population means:

1. Checking Normal Conditions: The underlying distribution of the two samples are not given normal. However, since $n = 189 > 30$ for both samples are large, we can employ the CLT to say that the sampling distribution for the difference between population means is approximately normal.
2. To be safe, we still rely on the robustness of the student-t distribution (and/or proceed with caution) against the mild violations of normality.

Extra Questions

Question 4 (Solution)

f) Perform hypothesis testing for difference between population means:

1. Sample statistics: We derive the df without assuming equal variances:

$$v = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{1}{n_X - 1} \cdot \left(\frac{s_X^2}{n_X}\right)^2 + \frac{1}{n_Y - 1} \cdot \left(\frac{s_Y^2}{n_Y}\right)^2} = \frac{\left(\frac{11.52^2}{189} + \frac{12.37^2}{189}\right)^2}{\frac{1}{188} \cdot \left(\frac{11.51^2}{189}\right)^2 + \frac{1}{188} \cdot \left(\frac{12.37^2}{189}\right)^2}$$

$$= 374.6682$$

2. We round the degree of freedom down to 374.
3. Sample statistics: We derive the t-statistics by:

$$t = \frac{(\bar{x} - \bar{y}) - \mu_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} = \frac{(86.755 - 80.589) - 0}{\sqrt{\frac{11.52^2}{189} + \frac{12.37^2}{189}}} = 5.015$$

Extra Questions

Question 4 (Solution)

f) Perform hypothesis testing for difference between population means:

1. We refer to the t-table and find the probability $P(t_{374} > 5.015) = 4.11 * 10^{-7}$. We multiply that probability by 2 for two-tails, obtaining the p-value $8.21 * 10^{-7}$.
2. Conclusion: With a p-value lower than both the significant levels 0.05 and 0.01, we reject H_0. Therefore, we have sufficient evidence to suggest that the **population means of exam 1 and exam 2** are different.

Extra Questions

Question 4 (Solution)

g) Construct confidence interval for difference between population means:

1. For two-tails (keyword: **between**) , we find $t_{v,\frac{\alpha}{2}} = t_{374,0.005} = 2.589$. Therefore the 99% confidence intervals for the population mean difference between exam 1 and exam 2 is:

$$(\bar{x}-\bar{y}) \pm t_{v,\frac{\alpha}{2}} \cdot \sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}} = (86.76-80.59) \pm 2.589 \cdot \sqrt{\frac{11.52^2}{189} + \frac{12.37^2}{189}}$$

$$= (2.983,\ 9.349)$$

2. Conclusion: We are approximately 99% confident that the difference between population mean score of exam 2 over exam 1 falls between (2.983, 9.349)

Extra Questions

Question 4 (Solution)

g) Construct confidence interval for difference between population means:

1. For one-tail, we find $t_{374,0.01} = 2.336$. Therefore the 99% confidence upper and lower bound for the population mean difference between exam 1 and exam 2 are:

$$(\bar{x}-\bar{y})+t_{v,\alpha}\cdot\sqrt{\frac{s_X^2}{n_X}+\frac{s_Y^2}{n_Y}} = (86.76-80.59)+-2.336\cdot\sqrt{\frac{11.52^2}{189}+\frac{12.37^2}{189}}$$

$= 9.038$ and $= 3.294$

2. Conclusion: We are approximately 99% confident that the difference between population mean score of exam 2 over exam 1 should fall below 9.038 and fall above 3.294, respectively.

Extra Questions

Question 4 (Solution)

h) Comparison between confidence interval for difference between population means and population mean difference:

1. Since we are using the same data, we are good to compare the two methods.
2. The 99% confidence interval obtained from **difference between population means** is (2.983, 9.349), while the 99% confidence interval obtained from **population mean difference** is (3.911, 9.089).
3. Conclusion: Therefore, the confidence interval for **difference between population means** is wider.
4. Taking population mean difference for paired sample usually results in less variability than taking the difference between population means.

Extra Questions

Question 5 (Data 1)

The following two graphs show the outputs for simple linear regression for 2 high school swimmers to swim 50 yards, based on the week of training.

Swimmer1 Data:

Swim1

week	time
1	23.1
3	23.2
5	22.9
7	22.9
9	22.8
11	22.7
13	22.6
15	22.7

Output:

```
Call:
lm(formula = time ~ week, data = Swim1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.07857 -0.05208 -0.02857  0.02619  0.14405

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.172024   0.063130  367.05 2.76e-14 ***
week        -0.038690   0.006847   -5.65  0.00132 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08875 on 6 degrees of freedom
Multiple R-squared:  0.8418,    Adjusted R-squared:  0.8154
F-statistic: 31.93 on 1 and 6 DF,  p-value: 0.001318
```

a) Compute the least-squares line for predicting swimming time from week.
b) Find a 95% confidence interval for the slope (week effect).

Extra Questions

Question 5 (Data 2)

Swimmer2 Data:

Swim2

week	time
1	22.7
3	22.6
5	22.8
7	22.8
9	22.9
11	22.8
13	22.9
15	22.8

Output:

```
lm(formula = time ~ week, data = Swim2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.11905 -0.04226  0.01726  0.04643  0.09881

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.67798    0.05606 404.569 1.54e-14 ***
week         0.01369    0.00608   2.252   0.0653 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07881 on 6 degrees of freedom
Multiple R-squared:  0.458,     Adjusted R-squared:  0.3677
F-statistic:  5.07 on 1 and 6 DF,  p-value: 0.06529
```

c) Again, compute the least-squares line for predicting swimming time from week.

d) Again, find a 95% confidence interval for the slope (week effect).

Extra Questions

Question 5

Using these two data, answer the following questions:
e) If the regression can predict the future swimming time, which swimmer will swim faster in week 16, and how much faster?
f) Interpret the R^2 values from both models. Which model better explains the observed variation in the model?
g) From the R^2 values from from both models, compute the correlation coefficient of the two models.
h) Which model(s) have significant week effect to the swimmer's swim time if $\alpha = 0.05, 0.01$? In which case we can drop the predictor?

Extra Questions

Question 5

Min	25% Quantile	Median	Mean	75% Quantile	Max	Variance	Corrected SS
1	4.5	8	8	11.5	15	24	168

Using the two data and the above table, answer the following questions:
i) Now we know that both swimmers had a training at week 10, but the records were lost. Could you compute the 95 % confidence interval for mean response and the prediction interval for the extra observation for both swimmers?
j) Which swimmer has a wider prediction interval? Which value in the summary table suggests which prediction interval is wider or narrower?

Extra Questions

Question 5 (Solution)

a) The least-square line for predicting swimming time from week for swimmer 1 is $\hat{y} = \hat{\beta}_0 + \beta_1 \cdot x = 23.172 - 0.039x$

b) From the R output, we have the 95% confidence interval for slope is $\hat{\beta}_1 \pm t_{n-2,\frac{\alpha}{2}} \cdot SE(\hat{\beta}_1) \to -0.0387 \pm t_{6,0.025} \cdot 0.0068 \to -0.0387 \pm 2.4469 \cdot 0.0068 = (-0.0554, -0.0219)$. Thus, we are **approximately** 95% confident that the slope for predicting swimming time from week for swimmer 1 falls between (-0.0554, -0.0219).

c) The least-square line for predicting swimming time from week for swimmer 2 is $\hat{y} = \hat{\beta}_0 + \beta_1 \cdot x = 22.678 + 0.014x$.

d) Similarly, we have the 95% confidence interval for slope: $\hat{\beta}_1 \pm t_{n-2,\frac{\alpha}{2}} \cdot SE(\hat{\beta}_1) \to 0.014 \pm 2.4469 \cdot 0.0061 = (-0.0012, 0.0286)$. Thus, we are **approximately** 95% confident that the slope for predicting swimming time from week for swimmer 1 falls between (-0.0012, 0.0286).

Extra Questions

Question 5 (Solution)

e) If the regression can predict the future swimming time, at week 16, the predicted swimming time for swimmer 1 and swimmer 2 are $\hat{y} = \hat{\beta}_0 + \beta_1 \cdot x = 23.172 - 0.039(16) = 22.548$ (swimmer 1), and $\hat{y} = \hat{\beta}_0 + \beta_1 \cdot x = 22.678 + 16 * 0.014 = 22.897$ (swimmer 2). Therefore, based on the simple linear regression relationship, swimmer 1 is expected to swim faster at week 16.

f) The two data give $R^2 = 0.8418$ for swimmer 1 and $R^2 = 0.458$, respectively. Therefore, we claim that approximately 84.2% and 45.8% of the observed variations, in the swimmer 1 and swimmer 2 models, can be explained by the simple linear regression relationship between training week and swimming time. Obviously, swimmer 1 model better explains the observed variation since the R^2 value is higher.

Extra Questions

Question 5 (Solution)

g) Judging from the R^2 in both models and the **slope coefficients** which tells positive/negative linear relationship between predictor and response, we have the correlation for swimmer 1 model is $r = -\sqrt{R^2} = -\sqrt{0.8418} = -0.9175$ (slope from the swimmer 1 output is negative, so negatively correlated) and the correlation for swimmer 2 model is $r = \sqrt{R^2} = \sqrt{0.458} = 0.6768$ (slope from the swimmer 2 output is positive, so positively correlated)

h) Here we need to perform hypothesis testing for the slope coefficients for both models. For both models, we take $H_0 : \beta_1 = 0$ *vs* $H_i : \beta_1 \neq 0$. We also assume that in both models, all the observations are independent of each other, and the error terms are random and independent, and normally distributed with same variance, so as to fulfill the normal assumptions for linear model.

Extra Questions

Question 5 (Solution)

h) (Continue) Test statistics: $T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{-0.0387 - 0}{0.00685} = -5.65 \sim t_6$. Here we have $P(T_6 < -5.65) = 0.00066$, so we multiply it by 2 to get the p-value for the slope is 0.00132. Therefore, this p-value is smaller than both the significance levels 0.05 and 0.01, we reject H_0 for both cases. We have sufficient evidence to claim that the week training is statistically significant in predicting swimmer 1's swimming time and should be included in the model.

h) (Continue) Test statistics: $T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{0.01369 - 0}{0.00608} = 2.252 \sim t_6$. Here we have $P(T_6 > 2.252) = 0.03263$, so we multiply it by 2 to get the p-value for the slope is 0.0653. Therefore, this p-value is greater than both the significance levels 0.05 and 0.01, we fail to reject H_0 for both cases. We do not have sufficient evidence to claim that the week training is statistically significant in predicting swimmer 2's swimming time and should be dropped in the model.

Extra Questions

Question 5 (Solution)

i) For swimmer 1, the 95% confidence interval for mean response for $x^* = 10$ is $(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2},n-2} \cdot s_{\hat{Y}} =$

$$(23.172 - 0.0387 * 10) \pm t_{0.025,6} \cdot \sqrt{MSE \cdot [\frac{1}{n} + \frac{(x^* - \bar{x})^2}{Sum\ \ of\ \ Square}]} =$$

$$22.785 \pm 2.447 \cdot 0.08875 \cdot \sqrt{\frac{1}{8} + \frac{(10-8)^2}{168}} = (22.701, 22.869).$$

i) For swimmer 2, the 95% confidence interval for mean response for $x^* = 10$ is $(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2},n-2} \cdot s_{\hat{Y}} =$

$$(22.678 + 0.0137 * 10) \pm t_{0.025,6} \cdot \sqrt{MSE \cdot [\frac{1}{n} + \frac{(x^* - \bar{x})^2}{Sum\ \ of\ \ Square}]} =$$

$$22.815 \pm 2.447 \cdot 0.07881 \cdot \sqrt{\frac{1}{8} + \frac{(10-8)^2}{168}} = (22.741, 22.889).$$

Conclusion: We are approximately 95% confident that the confidence intervals for true mean response for swimmer 1 and swimmer 2 fall between (22.701, 22.869), and between (22.741, 22.889), respectively.

Extra Questions

Question 5 (Solution)

i) For swimmer 1, the 95% prediction interval for future observation for $x^* = 10$ is $(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2},n-2} \cdot s_{Y_{new}} =$

$$(23.172 - 0.0387 * 10) \pm t_{0.025,6} \cdot \sqrt{MSE \cdot [1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{Sum\ of\ Square}]} =$$

$$22.785 \pm 2.447 \cdot 0.08875 \cdot \sqrt{1 + \frac{1}{8} + \frac{(10-8)^2}{168}} = (22.552, 23.018).$$

i) For swimmer 2, the 95% prediction interval for future observation for $x^* = 10$ is $(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2},n-2} \cdot s_{Y_{new}} =$

$$(22.678 + 0.0137 * 10) \pm t_{0.025,6} \cdot \sqrt{MSE \cdot [1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{Sum\ of\ Square}]} =$$

$$22.815 \pm 2.447 \cdot 0.07881 \cdot \sqrt{1 + \frac{1}{8} + \frac{(10-8)^2}{168}} = (22.608, 23.022).$$

Conclusion: We are approximately 95% confident that the prediction intervals for true future observation of swimming time for swimmer 1 and swimmer 2 fall between (22.552, 23.018), and between (22.608, 23.022).

Extra Questions

Question 5 (Solution)

j) The lengths of the 95% prediction interval for swimmer 1 and swimmer 2 are $23.018 - 22.552 = 0.466$ and $23.022 - 22.608 = 0.414$, respectively. Obviously, the 95% prediction interval for swimmer 1 is wider. In the summary table, we should look closely to the **Residual Standard Error**, which is the **RMSE**. Since we have same sample size 8 for both models, and have the same new observation $x^* = 10$, the greater the RMSE is, the wider the prediction interval will be.

Extra Questions

Question 6

The following graph shows the outputs for simple linear regression for the intensity of ultraviolet, based on the temperature in Celsius.

Temperature	Intensity
31.1	1.78
25.3	1.58
23.5	1.45
22.4	1.13
21.7	0.96
14.9	0.99
11.3	1.05
15.0	0.82
8.7	0.68
8.2	0.56

```
lm(formula = Frequency ~ Temperature)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30126 -0.11812  0.03119  0.10027  0.26929

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.258563   0.152243   1.698  0.12787
Temperature 0.046207   0.007759   5.955  0.00034 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1793 on 8 degrees of freedom
Multiple R-squared:  0.8159,    Adjusted R-squared:  0.7929
F-statistic: 35.46 on 1 and 8 DF,  p-value: 0.00034
```

a) Compute the least-squares line for predicting intensity from temperature.

Extra Questions

Question 6

b) If the temperature decreases by 2 Celsius, by how much would we expect for the change of intensity?
c) For what temperature would you predict for the intensity of ultraviolet being 1.5?
d) Without doing calculation, could you learn from the summary table to tell whether temperature is a statistically significant predictor for the intensity of ultraviolet at significance level 0.01? Then, verify your answer by performing hypothesis testing.

Extra Questions

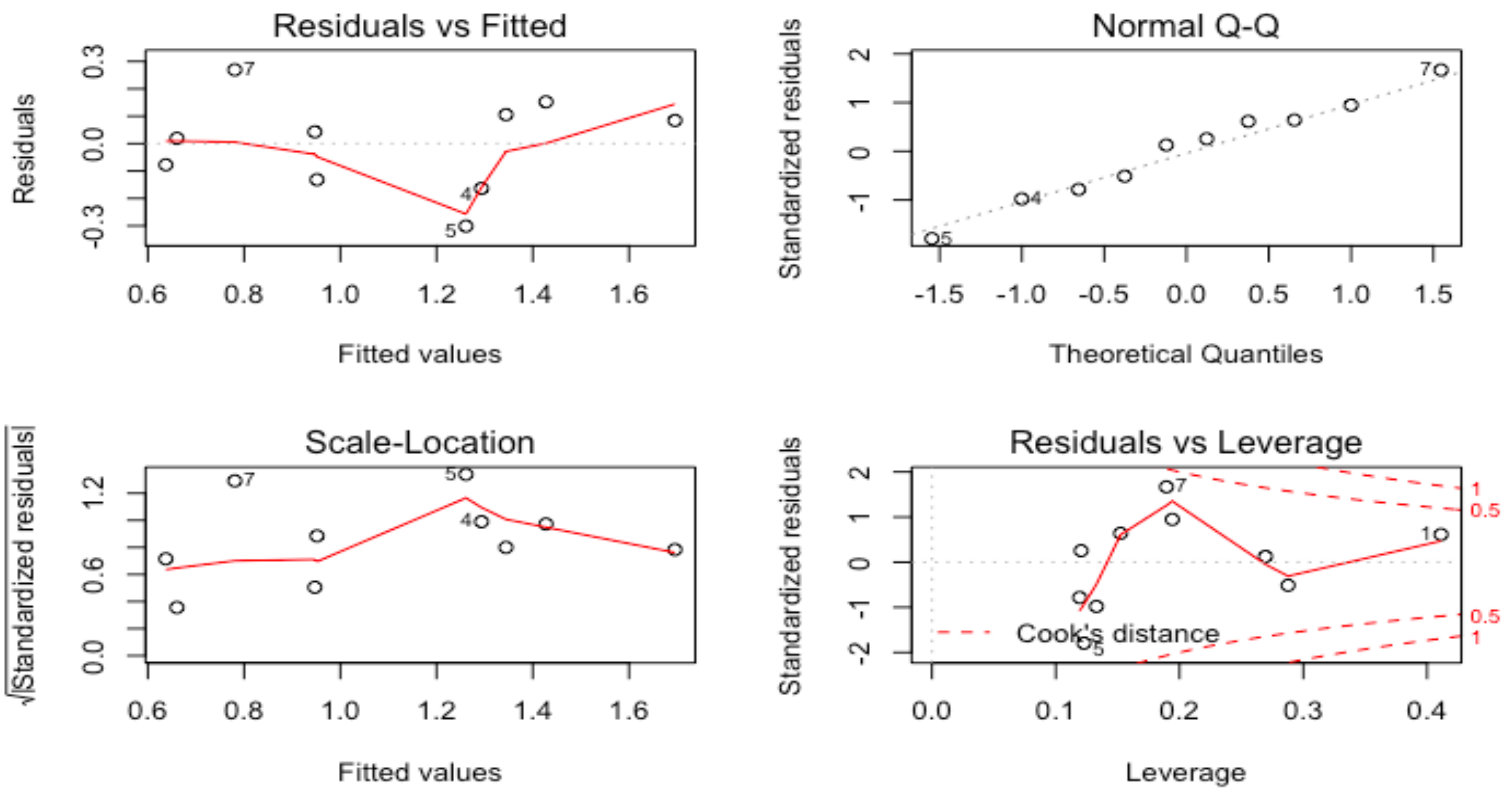
Question 6

Min	25% Quantile	Median	Mean	75% Quantile	Max	Variance	Corrected SS
8.2	12.2	18.35	18.21	23.225	31.1	59.3099	533.789

The forecast says that tomorrow's temperature at noon will be 30 Celsius.
e) Can you compute the 95 % prediction interval for the ultraviolet intensity tomorrow at noon?
f) Can you conclude that the mean ultraviolet intensity with a temperature of 30 Celsius is less than 1.8 units, for significance level 0.01?

Extra Questions

Question 6



g) What inferences can you draw from the scatterplot, Residual vs Fitted plot and normal Q-Q plot?

Extra Questions

Question 6 (Solution)

a) The least-square line for predicting intensity of ultraviolet from temperature is $\hat{y} = \hat{\beta}_0 + \beta_1 \cdot x = 0.2586 + 0.0462x$

b) If the temperature decreases by 2 Celsius, we anticipate the change of intensity be $0.0462 \cdot (-2) = -0.0924$?

c) When the intensity of ultraviolet is 1.5, we have $1.5 = 0.2586 + 0.0462x \rightarrow x = \frac{1.5-0.2586}{0.0462} = 26.87$. Therefore, we predict the temperature will be 26.87 Celsius degree (or approximately 27 Celsius degree).

d) Learning from the summary table, we can see in the row containing Temperature that the test-statistics is $T = 5.955 \sim t_8$, which corresponding to the two-tails p-value $0.00034 < 0.01$. Therefore, temperature is a statistically significant predictor for the intensity of ultraviolet at $\alpha = 0.01$.

Extra Questions

Question 6 (Solution)

d) (Continue) To test whether temperature is a statistically significant predictor for ultraviolet intensity at $\alpha = 0.01$, we take $H_0 : \beta_1 = 0$ *vs* $H_1 : \beta_1 \neq 0$. We also assume that the model fulfills the normal assumptions for linear model. Test statistics: $T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{0.0462 - 0}{0.0078} = 5.955 \sim t_6$. Here we have $P(T_8 < -5.65) = 0.00017$, so we multiply it by 2 to get the p-value for the slope is 0.00034. Therefore, this p-value is smaller than both the significance levels 0.01, we reject H_0 for both cases. We have sufficient evidence to claim that the temperature is statistically significant in predicting the intensity of ultraviolet and should be included in the model.

Extra Questions

Question 6 (Solution)

e) The 95% prediction interval for future observation for $x^* = 30$ is $(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\frac{\alpha}{2},n-2} \cdot s_{Y_{new}} =$

$$(0.2586 + 0.0462 * 30) \pm t_{0.025,8} \cdot \sqrt{MSE \cdot [1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{Sum\ \ of\ \ Square}]} =$$

$$1.6446 \pm 2.306 \cdot 0.1793 \cdot \sqrt{1 + \frac{1}{10} + \frac{(30-18.21)^2}{533.789}} = (1.1623, 2.1269).$$

Conclusion: We are approximately 95% confident that the prediction intervals for true intensity of ultraviolet tomorrow at noon will fall between (1.1623, 2.1269).

Extra Questions

Question 6 (Solution)

f) To test whether the mean ultraviolet intensity with a temperature of 30 Celsius is less than 1.8, at $\alpha = 0.01$, we take $H_0 : \beta_0 + \beta_1 x^* \geq 1.8$ *vs* $H_1 : \beta_0 + \beta_1 x^* < 1.8$. We also assume that the model fulfills the normal assumptions for linear model. Test statistics: $T = \frac{\hat{Y} - (\beta_0 + \beta_1 x^*)}{s_{\hat{Y}}} = \frac{(0.2586 + 0.0462 * 30) - 1.8}{\sqrt{MSE \cdot [\frac{1}{n} + \frac{(x^* - \bar{x})^2}{Sum\ \ of\ \ Square}]}} = \frac{-0.1554}{0.1793 \cdot \sqrt{\frac{1}{10} + \frac{(30 - 18.21)^2}{533.789}}} = -1.444 \sim t_8$. Here we have $P(T_8 < -1.444) = 0.0934$. Since this is one-tail hypothesis, the p-value is 0.0934.

Conclusion: Since the p-value is greater than $\alpha = 0.01$, we fail to reject H_0 and we do not have sufficient evidence to conclude that the true mean ultraviolet intensity when temperature of 30 Celsius is significantly less than 1.8.

Extra Questions

Question 6 (Solution)

g) We hope to see the following in a residual plot: 1) Random scatter, 2) No pattern or trend, 3) Constant variance and 4) No outliers. In this residual plot, although we see a slight curve, the residuals are randomly scattered above or below the 0 horizontal line. There are two potential outliers with labels, but their residual values are not too extreme. In addition, the variability seems constant (homoscedastic), so the constant variance assumption seems satisfied.

g) In the normal Q-Q plot, we would like to see the points follow the line fairly well. Here the points follow the line very closely, suggesting that the sample are approximately normal in distribution.

g) Taken together, we can tolerate the mild violation of normality appeared above and say that the normal assumptions for linear model are fulfilled. That is, the simple linear regression taking temperature as predictor and intensity of ultraviolet as response is well-fitted.

Go Wolverines!

New Fight Song

Go Wolverines

In Son Zeng

Go Wolverines, Go Wolverines,
Contend for victories through all means
Go Wolverines, Go Wolverines,
We are Michigan Wolverines, be proud our teens

Go Wolverines, Go Wolverines,
Pursue the brilliance against routines
Go Wolverines, Go Wolverines,
We are Michigan Wolverines, Go Blue our teens

Go Wolverines, Go Wolverines,
Be faithful one day fulfil our dreams
Go Wolverines, Go Wolverines,
We are Michigan Wolverines, Go Blue our teens

The End