DS6371 FInal Exam

Travis Deason -12/9/2017

**10-1**

plastic <- read.csv('data/plastic\_fixed.csv')  
names(plastic) = sapply(names(plastic), tolower)  
names(plastic)

## [1] "x" "bacteria" "plastic" "region" "temperature"

plastic = plastic[,c("bacteria","plastic","region","temperature")]

### (12 pts.) Conduct a two-sample t-test to test the claim that the mean plastic level (in ounces) is greater in the West Pacific than in the East. Be sure to address the assumptions of the test. For the actual test, please provide only the conclusion with a p-value and a confidence interval, and finish with a scope of inference that addresses causality (if the result is significant) and the generalizability of the result.

east <- subset(plastic, region == 'EAST')  
west <- subset(plastic, region == 'WEST')  
east\_plastic = east$plastic  
west\_plastic = west$plastic  
mean(east\_plastic)

## [1] 2172.295

mean(west\_plastic)

## [1] 2123.008

hist(east\_plastic)

hist(west\_plastic)

qqnorm(east\_plastic)

qqnorm(west\_plastic)

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

t.test(east\_plastic, west\_plastic , 'greater')

##   
## Welch Two Sample t-test  
##   
## data: east\_plastic and west\_plastic  
## t = 0.41253, df = 102.98, p-value = 0.3404  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## -149.0116 Inf  
## sample estimates:  
## mean of x mean of y   
## 2172.295 2123.008

qt(.95, 102)

## [1] 1.65993

* The null hypothesis is that the east and west side of the pacific have the same amount of plastic
* The alternative hypothesis is that the west pacific has more plastic then the east pacific
* For this expired we will set the significance level to .95. Since the Data-set are large, but the variances are not equal we will use welchs' t-test
* The critical T-statistic at 102.98 degrees of freedom is 1.65993
* The confidence interval for our null hypothesis at .95 is 2172.295 to 2123.008
* Since the mean plastic observed in the west pacific is 2172; which is in our confidence interval, we fail to reject the null hypothesis with a p value of .3404

### # 10-2

### (12 pts.) Fit a simple linear regression model with the plastic level as the response and the Ideonella sakaiensis (bacteria) level as the explanatory variable. If the slope is significantly different from zero, provide a scatter plot with the regression line superimposed on the data. In addition, interpret the slope of the regression equation and include a confidence interval. Finally, interpret the intercept and include a confidence interval. Please address all relevant assumptions and include relevant plots to support those assumptions.

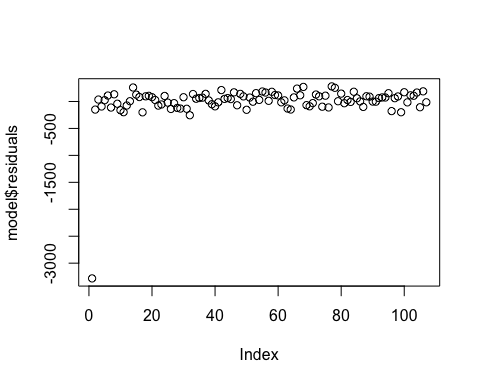
model <- lm(plastic~bacteria, data=plastic)  
#summary(model)  
min(plastic$plastic)

## [1] 107.5226

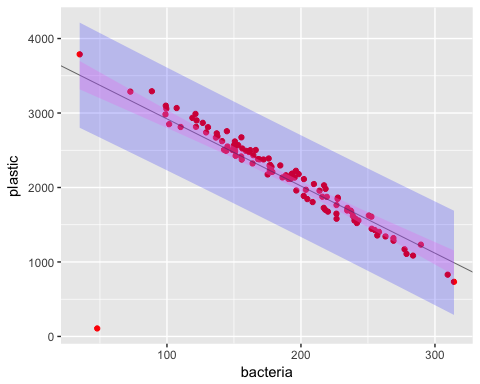
slope <- model$coefficients[2]  
intercept <- model$coefficients[1]  
pred.int = predict(model, interval="prediction")

## Warning in predict.lm(model, interval = "prediction"): predictions on current data refer to \_future\_ responses

#Confidence intervals  
conf.int = predict(model, interval="confidence")  
plastic$pred.lower <- pred.int[,2]  
plastic$pred.upper <- pred.int[,3]  
plastic$ci.upper <- conf.int[,2]  
plastic$ci.lower <- conf.int[,3]  
plot(model$residuals)



ggplot(data=plastic, aes(x=bacteria, y=plastic, main='Scatterplot of plastic vs Bacteria')) +   
 geom\_point(color= 'red') +  
 geom\_abline(intercept=intercept, slope=slope, color='black', size=.2) +  
 geom\_ribbon(data=plastic, aes(ymin= pred.lower, ymax= pred.upper), fill = "blue", alpha = 0.2) +  
 geom\_ribbon(data=plastic, aes(ymin= ci.lower, ymax= ci.upper), fill = "violet", alpha = 0.3)



## Removing the observation which seems like it may be too extreme to be a real value  
no\_outlier <- subset(plastic, plastic > 110)  
model2 <- lm(plastic~bacteria, data=no\_outlier)  
summary(model2)

##   
## Call:  
## lm(formula = plastic ~ bacteria, data = no\_outlier)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -212.499 -56.740 9.819 53.425 180.060   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4129.7828 30.4810 135.49 <2e-16 \*\*\*  
## bacteria -10.4975 0.1567 -67.01 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 87.06 on 104 degrees of freedom  
## Multiple R-squared: 0.9774, Adjusted R-squared: 0.9771   
## F-statistic: 4490 on 1 and 104 DF, p-value: < 2.2e-16

slope <- model2$coefficients[2]  
intercept <- model2$coefficients[1]  
t\_val <- qt(.975, 104)  
print(t\_val)

## [1] 1.983038

print('intercept confidence interval, min, max')

## [1] "intercept confidence interval, min, max"

print(intercept - t\_val \* 30.4810)

## (Intercept)   
## 4069.338

print(intercept + t\_val \* 30.4810)

## (Intercept)   
## 4190.228

print('slope confidence interval, min, max')

## [1] "slope confidence interval, min, max"

print(slope - t\_val \* 0.1567)

## bacteria   
## -10.80825

print(slope + t\_val \* 0.1567)

## bacteria   
## -10.18677

conf.int2 = predict(model2, interval="confidence")  
pred.int2 = predict(model2, interval="prediction")

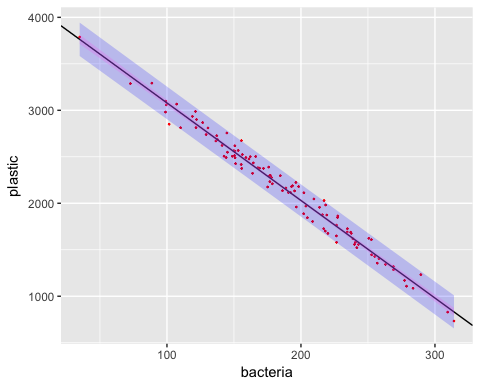
## Warning in predict.lm(model2, interval = "prediction"): predictions on current data refer to \_future\_ responses

no\_outlier$pred.lower <- pred.int2[,2]  
no\_outlier$pred.upper <- pred.int2[,3]  
no\_outlier$ci.upper <- conf.int2[,2]  
no\_outlier$ci.lower <- conf.int2[,3]  
qqnorm(model2$residuals)

plot(model2$residuals)

|  |  |
| --- | --- |
|  |  |

ggplot(data=no\_outlier, aes(x=bacteria, y=plastic, main='Scatterplot of plastic vs Bacteria')) +   
 geom\_point(color= 'red', size=.2) +  
 geom\_abline(intercept=intercept, slope=slope, color='black') +  
 geom\_ribbon(data=no\_outlier, aes(ymin= pred.lower, ymax= pred.upper), fill = "blue", alpha = 0.2) +  
 geom\_ribbon(data=no\_outlier, aes(ymin= ci.lower, ymax= ci.upper), fill = "violet", alpha = 0.3)



* In the linear model of Bacteria verses plastic there is one value for plastic 107.5526 which is so much lower then the remaining values that it is geometrically impossible. Using the models shown above, this point was removed from the data-set, as it caused model residuals to be non-randomly distributed. This point which more accurately shows the data. Since there is the possibility that the point was an actual observation, we will assume the scope of inference only covers plastic levels over 110 for this model. For the purposes of the remainder of this study, this observation will not be included.
* Model two meets all of the assumptions of linear regression. The residuals are normally distributed with equal variance, the
* With the new model (model2), the null hypothesis is that there is no relationship between the amount of plastic in the water and the bacteria level
* The alternative hypothesis is that there is a relationship between the amount of plastic in the water and the bacteria levels
* The significance level for this test is .95
* The critical t-value for this test is 1.983038
* Based on a t-test with 104 degrees of freedom, the confidence interval for the intercept is 4069.338 to 4190.228, and for the slope -10.80825 -10.18677.
* The p-value for this test is ~0 (too low to represent with double precision floating point)
* Based on the test values, we reject the null hypothesis that plastic and bacteria levels in the ocean are unrelated. There is strong evidence to suggest that an increase in bacteria in the ocean is correlated with an increase of plastic. Since this was an observational study, we cannot assume the relationship is causal, as there may be confounding variables.
* The critical values for our model are: The intercept is 4129, meaning if there was no remaining plastic in the ocean, we would expect to see a bacteria level per milliliter of 4129 The slope of the line is -10.49 Meaning that for each 10 additional bacteria, we would expect to see one less kilogram of plastic.

### # 10-3

### (10 pts.) Next, the researchers want to test the claim that the relationship (slope) between the Ideonella sakaiensis level and the plastic level were different between the East and the West Pacific. Add the necessary variable(s) to the model to test for this difference. Provide a copy of the parameter estimate table along with a short (1 or 3 sentences) conclusion about whether the slopes are significantly different for the East and the West, supported with a p-value and a confidence interval. You may assume the assumptions for your model are met.

no\_outlier$is\_west = no\_outlier$region == 'WEST'  
west\_model <- lm(plastic~(bacteria\*is\_west), data=no\_outlier)  
summary(west\_model)

##   
## Call:  
## lm(formula = plastic ~ (bacteria \* is\_west), data = no\_outlier)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -171.496 -64.393 -1.625 58.025 161.564   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4068.1350 41.8448 97.220 <2e-16 \*\*\*  
## bacteria -10.2944 0.2233 -46.105 <2e-16 \*\*\*  
## is\_westTRUE 139.9723 58.6492 2.387 0.0188 \*   
## bacteria:is\_westTRUE -0.4786 0.3024 -1.582 0.1166   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 83.06 on 102 degrees of freedom  
## Multiple R-squared: 0.9798, Adjusted R-squared: 0.9792   
## F-statistic: 1649 on 3 and 102 DF, p-value: < 2.2e-16

t.val = qt(.975, 102)  
print('interaction confidence interval, min, max')

## [1] "interaction confidence interval, min, max"

dslope <- west\_model$coefficients[4]  
print(dslope - t.val \* 0.3024)

## bacteria:is\_westTRUE   
## -1.078385

print(dslope + t.val \* 0.3024)

## bacteria:is\_westTRUE   
## 0.1212332

* The relationship between bacteria and plastic levels in the west pacific is different by -.4786 bacteria per kilogram of plastic. The response to bacteria by plastic appears to be stronger in the west then east pacific. This relationship is not strong enough to be considered statistically significant at a alpha of .05 as the confidence interval is from -1.078 to .1212, and the p-value is .1166

### # 10-4

### (6 pts.) Next, researchers want to revisit the claim that the mean plastic level is different in the West and East Pacific (similar to the claim tested in part (1), although this time they want to use additional information: the level of Ideonella sakaiensis in the water as well). Fit a model that regresses the plastic level (response) against the bacterial level (explanatory variable) and assume equal slopes for the East and the West to test this claim. Again, provide a copy of the parameter estimate table of the model used to answer this question as well as a scatter plot with the regression lines superimposed. Again, provide a short (1 to 3 sentence) conclusion providing guidance as to which mean is greater or if there is no evidence to suggest a difference. Your analysis should be supported by p-value(s) and confidence interval(s). You may assume the assumptions for your model are met.

t.crit = qt(.95, 102)  
print(t.crit)

## [1] 1.65993

regional\_model <- lm(plastic~bacteria + is\_west, data=no\_outlier)  
summary(regional\_model)

##   
## Call:  
## lm(formula = plastic ~ bacteria + is\_west, data = no\_outlier)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -192.023 -61.968 3.924 58.494 156.409   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4115.1723 29.6672 138.711 < 2e-16 \*\*\*  
## bacteria -10.5552 0.1517 -69.584 < 2e-16 \*\*\*  
## is\_westTRUE 50.7983 16.3758 3.102 0.00248 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 83.66 on 103 degrees of freedom  
## Multiple R-squared: 0.9793, Adjusted R-squared: 0.9789   
## F-statistic: 2436 on 2 and 103 DF, p-value: < 2.2e-16

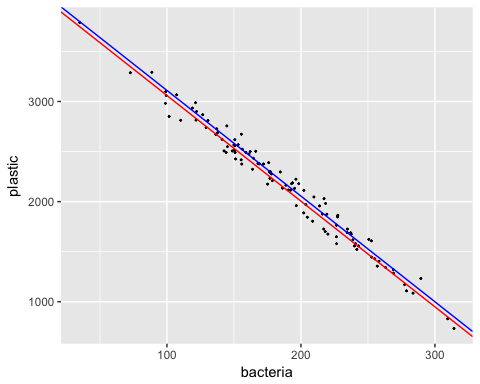
slope.r <- regional\_model$coefficients[2]  
intercept.r <- regional\_model$coefficients[1]  
offset.r <- regional\_model$coefficients[3]  
print(offset.r - t.val \* 16.3758)

## is\_westTRUE   
## 18.31702

print(offset.r + t.val \* 16.3758)

## is\_westTRUE   
## 83.27966

ggplot(data=no\_outlier, aes(x=bacteria, y=plastic, main='Scatterplot of plastic vs Bacteria')) +   
 geom\_point(color= 'black', size=.3) +  
 geom\_abline(intercept=intercept.r, slope=slope.r, color='red') +  
 geom\_abline(intercept=intercept.r + offset.r, slope=slope.r, color='blue')



* The null hypothesis is that there is no difference in the levels of plastic in the east and west pacific.
* The alternative hypothesis is that there is more plastic in the east pacific then the west pacific
* For this experiment, we will use a one sided t-test on a significance level of .95
* The critical t-value for this test is at 103 degrees of freedom 1.65993.
* The confidence interval for this test is 18.317 to 83.27966 kg of plastic per square kilometer
* The p-value for this test is .00248
* Based on the test data, we can reject the null hypothesis with a t-value of 3.102. There i evidence to suggest that there is a greater level of trash in the west pacific then the east pacific which is statistically significant at an alpha of .05. Since this study was observational and not experiential, we cannot infer the relationship is causal.

### # 10-5

### (10 pts.) Researchers have a special interest in a section of the Pacific Ocean about 20 miles off the coast of California (in the EAST Pacific Ocean). The researchers know that the level of Ideonella sakaiensis in the area is 170 bpml and would like to predict the level of the plastic in the East Pacific Ocean where the Ideonella sakaiensis level is 170 bpml. Write a short (1 to 3 sentence) response that includes an estimate of the plastic level for this individual sample with the appropriate interval (confidence or prediction, as appropriate).

vals = data.frame('is\_west' = TRUE, 'bacteria' = 170)  
predict(regional\_model, vals, interval="prediction", level=.95)

## fit lwr upr  
## 1 2371.582 2203.948 2539.216

* The predicted plastic level in the west Pacific ocean for a bacteria level of 170 bpml is 2371 +/- 168 kg/km\*km, or 2203 - 2539 at a .95 significance level.

## 

## BONUS

### (max 3 pts.) The researchers want to test if the findings of the test in part 4 were confounded with the temperature effect. Given the analysis thus far, fit a model that will test this claim. Again, provide a copy of the parameter estimate table of the model used to answer the question. Write a short (1 to 3 sentence) response supported by p-value(s) and confidence interval(s).

mean(west$temperature)

## [1] 60.96934

mean(east$temperature)

## [1] 60.64461

temp\_model <- lm(plastic~(temperature\*is\_west), data=no\_outlier)  
summary(temp\_model)

##   
## Call:  
## lm(formula = plastic ~ (temperature \* is\_west), data = no\_outlier)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1446.93 -420.72 49.96 385.82 1606.35   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2751.301 785.329 3.503 0.000684 \*\*\*  
## temperature -8.823 12.781 -0.690 0.491564   
## is\_westTRUE -1023.957 1068.654 -0.958 0.340240   
## temperature:is\_westTRUE 15.313 17.406 0.880 0.381064   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 580.3 on 102 degrees of freedom  
## Multiple R-squared: 0.01355, Adjusted R-squared: -0.01546   
## F-statistic: 0.467 on 3 and 102 DF, p-value: 0.7059

t.val = qt(.975, 102)  
print(t.val)

## [1] 1.983495

print('interaction confidence interval, min, max')

## [1] "interaction confidence interval, min, max"

dslope <- west\_model$coefficients[4]  
print(dslope - t.val \* 17.406)

## bacteria:is\_westTRUE   
## -35.00329

print(dslope + t.val \* 17.406)

## bacteria:is\_westTRUE   
## 34.04614

* The null hypothesis is that there is no regionally dependent relationship between temperature and levels of plastic in the ocean.
* The alternative hypothesis is that the relationship between temperature and plastic levels in the ocean is different between west and east regions of the pacific.
* The critical t-value for this test at 102 degrees of freedom is 1.9835
* The confidence interval we obtain for the slope of the interaction term between temperature and plastic levels for east and west pacific is -35.0033 and 34.04614
* the p-value for this test .880
* based on this test data, we fail to retest the null hypothesis at a significance level of .95 with a slope of 15.313 kg/mi\*mi and a p value of .888. There is not sufficient evidence to suggest that regional differences in temperature can explain the regional difference in plastic observed in the ocean.