BIOSTAT 200B HW3

1. Questions for Lab3

(1) The intercept from such a regression of residuals on residuals will always be zero (disregarding rounding error). Why? (Hint: what is the sample mean of the residuals from a linear regression model? And what is the formula for the intercept in a simple linear regression model?)

The intercept in a simple regression model is $\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$. Since we know the sample mean of the residuals from a linear regression model will always be 0, we can get $\bar{X} = \bar{Y} = 0$ in this regression model, which regresses residuals on residuals. Therefore, $\widehat{\beta_0}$ will always be zero in this case.

(2) How do we interpret the parameter estimates for the coefficients for regnc, regs, regw? How do we interpret the intercept in this model?

The parameter estimates for the coefficients for **regnc** is -0.46696, which means when the hospital is located in North Central, then the infection risk would decrease by 0.46696. Similarly, the parameter estimates for the coefficients for **regs** and **regw** are -0.93369 and -0.47946, meaning that the infection risk would decrease by 0.93369 and 0.47946 when the hospital is located in South and West, respectively.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	13.99694	4.66565	2.71	0.0484
Error	109	187.38288	1.71911		
Corrected Total	112	201.37982			

Root MSE	1.31115	R-Square	0.0695
Dependent Mean	4.35487	Adj R-Sq	0.0439
Coeff Var	30.10765		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.86071	0.24778	19.62	<.0001
regnc	1	-0.46696	0.33929	-1.38	0.1716
regs	1	-0.93369	0.32842	-2.84	0.0053
regw	1	-0.47946	0.41090	-1.17	0.2458

The intercept in this model is 4.86071, and we can interpret it as the estimated infection risk when the hospital is located in North East since it is our reference category for these dummy variables.

(3) What region did we make the reference region by using the above code? Write code to fit a model using a different region as the reference group and run a regression model. Compare the output, especially the parameter estimates. Did R^2 change? Did the ANOVA table change?

As we stated in question (2), North East is the region where we made the reference region by using the code. We write another code, which lets North Central be the reference region.

```
data senic1; set senic;
      if region=1 then regne=1; else regne=0;
      if region=3 then regs=1; else regs=0;
      if region=4 then regw=1; else regw=0;
run;
proc reg data=senic1;
      model risk = regne regs regw;
run; quit;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	13.99694	4.66565	2.71	0.0484
Error	109	187.38288	1.71911		
Corrected Total	112	201.37982			

Root MSE	1.31115	R-Square	0.0695
Dependent Mean	4.35487	Adj R-Sq	0.0439
Coeff Var	30.10765		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.39375	0.23178	18.96	<.0001
regne	1	0.46696	0.33929	1.38	0.1716
regs	1	-0.46672	0.31652	-1.47	0.1432
regw	1	-0.01250	0.40146	-0.03	0.9752

The results of ANOVA table and the parameter estimates are shown beside.
We can find that the parameter estimates for the coefficients for **regs**, **regw**, and intercept all change. This is because we have changed our reference region form North East to North Central. Actually, the interpretation of these parameter estimates for the coefficients remain the same. In question (2), we knew that the infection risk would decrease by 0.46696 when the hospital is located in North Central, which means the estimated infection rate is 4.86071-0.46696=4.39375. In this question, we can also get the same coefficient for the intercept, as we can interpret it as the estimated infection risk when the hospital is located in North Central since it is our reference category in this case. While we comparing the R^2 and the ANOVA table, we can see that they are the same, because changing the reference category for dummy variables does not change the meaning of conducting these regression models.

(4) How do we interpret the p-values in the Parameter Estimates table, in terms of testing for differences in means by region? What means are being compared?

For the p-values in the Parameter Estimates table, if they are greater than the significance level α (let α be 0.05 in this question), than we cannot reject the null hypothesis, meaning that there is no statistically significant difference between the parameter estimates for the coefficients and 0. That is to say, we are comparing the mean infection risk of each

region and the reference region. For example, for the p-value of variable **regne** shown in question 3, p-value = 0.1716 > 0.05, and we can conclude that there is no statistically significant difference between the mean infection risk of North East and our reference category North Central.

(5) Write out the null and alternative hypotheses for the test conducted by "test_region". Give the distribution of the test statistic under the null, the value of the test statistic and the p-value. What do you conclude?

The null hypothesis H_0 is the parameter estimates for the coefficient of variable **regnc**, **regs**, and **regw** are all equal to 0. The alternative hypothesis H_1 is as least one of the parameter estimates for the coefficient of variable **regnc**, **regs**, and **regw** is not equal to 0.

Test test_region Results for Dependent Variable risk

Source	DF	Mean Square	F Value	Pr > F
Numerator	3	3.04987	2.50	0.0636
Denominator	107	1.22083		

The distribution of the test statistics is $F^* \sim F_{3,107}$, and the value of the test statistic F* is 2.50, with the p-value = 0.0636. Since the p-value = 0.0636 > 0.5, we cannot reject H_0 and can conclude that there is no significant difference between each of the parameter estimates for the coefficient of variable **regnc**, **regs**, **regw** and 0.

(6) Conduct the overall (omnibus) F test for the model risk = length census regnc regs regw. Write out the null and alternative hypotheses. Give the distribution of the test statistic under the null, the value of the test statistic and the p-value. What do you conclude?

The null hypothesis H_0 is the parameter estimates for the coefficient of variable **length**, **census**, **regnc**, **regs**, and **regw** are all equal to 0. The alternative hypothesis H_1 is as least one of the parameter estimates for the coefficient of variable **length**, **census**, **regnc**, **regs**, and **regw** is not equal to 0.

Analysis of Variance

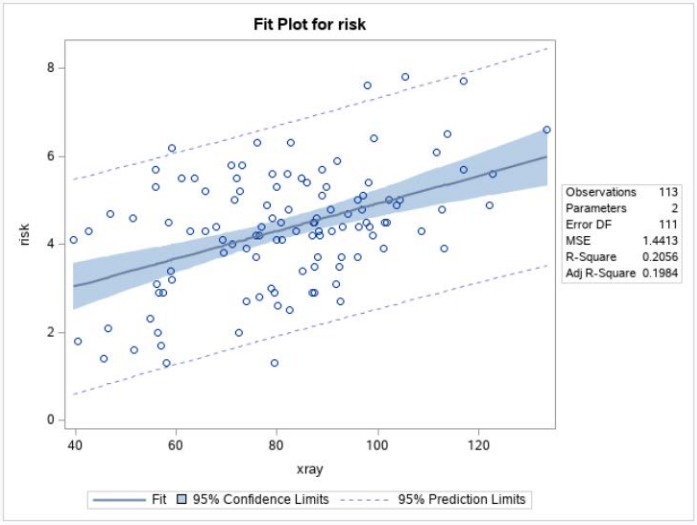
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	70.75096	14.15019	11.59	<.0001
Error	107	130.62886	1.22083		
Corrected Total	112	201.37982			

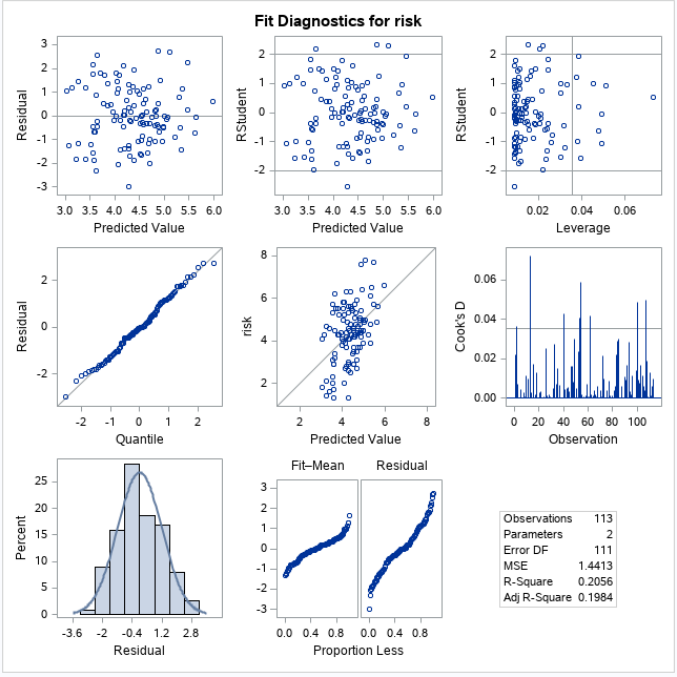
The distribution of the test statistics is $F^* \sim F_{5,107}$, and the value of the test statistic F* is

11.59, with the p-value < 0.0001. Since the p-value < 0.5, we can reject H_0 and can conclude that there is significant difference between at least one of the parameter estimates for the coefficient of variable **length**, **census**, **regnc**, **regs**, **regw** and 0.

(7) A regression of risk on xray shows a highly significant relationship. Fit the model and conduct model diagnostics (residuals analysis). Report this relationship. Provide an interpretation of the regression coefficients, using appropriate units.

```
proc reg data=senic;
      model risk = xray;
      run; quit;
```





The model fitting plot and residuals analysis are shown above. It looks like **xray** has a positive correlation with **risk**. Form residual analysis, we can find the regression function linear, and the error variances are constant; it also seems to meet the assumption of normality of residual according to Q-Q plot. We can interpret the regression coefficient of **xray** as the infection risk would increase by 0.0314 when the ratio of number of x-rays performed to number of patients without signs or symptoms of pneumonia X 100 increase 1 unit. The regression coefficient of intercept is 1.79202, which means that the estimated infection risk is 1.79202 when the ratio of number of x-rays is 0.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	41.39642	41.39642	28.72	<.0001
Error	111	159.98340	1.44129		
Corrected Total	112	201.37982			

Root MSE	1.20054	R-Square	0.2056
Dependent Mean	4.35487	Adj R-Sq	0.1984
Coeff Var	27.56773		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.79202	0.49136	3.65	0.0004
xray	1	0.03140	0.00586	5.36	<.0001

Investigators hypothesize that xray will be significantly related to risk after controlling for beds, nurses and svcs. Run the appropriate model and report the results. Provide an interpretation of each of the regression coefficients, using appropriate units. Also report the partial correlation of risk and xrays controlling for beds, nurses and svcs. Do the findings support their hypothesis or not?

```
proc reg data=senic;
      model risk = xray beds nurses svcs;
run; quit;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.86917	0.53682	1.62	0.1083
xray	1	0.02858	0.00542	5.27	<.0001
beds	1	-0.00033930	0.00141	-0.24	0.8106
nurses	1	0.00223	0.00191	1.17	0.2457
svcs	1	0.01976	0.01162	1.70	0.0918

The parameter estimates for the coefficients are shown beside in the table. With all other variables holding constant, we can interpret the regression coefficient of **xray** as the infection risk would increase by 0.02858 when the ratio of number of x-rays performed to number of patients without signs or symptoms of pneumonia X 100 increase 1 unit. Similarly, the interpretation of regression coefficient of **beds** is the infection risk would decrease by 0.00003393 when the average number of beds in hospital increase 1 unit; that of **nurses** is the infection risk would increase by 0.00223 when the average number of full-time equivalent nurses increase 1 unit; that of **svcs** is the infection risk would increase by 0.001976 when the percent of 35 potential facilities and services that are provided by the hospital increase 1 unit. The regression coefficient of intercept is 0.86917, which means that the estimated infection risk is 0.86917 when all the other variables are set as 0.

```
proc corr data = senic;
 var risk;
 with xray;
 partial beds nurses svcs;
run;
```

The CORR Procedure

3 Partial Variables:	beds nurses svcs
1 With Variables:	xray
1 Variables:	risk

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Partial Variance	Partial Std Dev
beds	113	252.16814	192.84269	28495	29.00000	835.00000		
nurses	113	173.24779	139.26539	19577	14.00000	656.00000		
svcs	113	43.15929	15.20086	4877	5.70000	80.00000		
xray	113	81.62832	19.36383	9224	39.60000	133.50000	377.06899	19.41826
risk	113	4.35487	1.34091	492.10000	1.30000	7.80000	1.50313	1.22602

Pearson Partial Correlation Coefficients, N = 113
Prob > |r| under H0: Partial Rho=0

	risk
xray	0.45260 <.0001

The result of the partial correlation of **risk** and **xrays** controlling for **beds**, **nurses** and **svcs** is shown beside. We can find that Pearson Partial Correlation Coefficients of **risk** and **xrays** is 0.4526, and the p-value < 0.001, meaning that these two variables are significantly related after controlling for **beds**, **nurses** and **svcs**. Therefore, this finding do support investigators' hypothesis.

Conduct a joint test of whether beds and nurses contribute to explaining variation in risk after controlling for svcs and xray.

```
proc reg data=senic;
      model risk = xray beds nurses svcs;
      join_test: test beds, nurses;
run; quit;
```

Test join_test Results for Dependent Variable risk

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	1.50050	1.24	0.2924
Denominator	108	1.20629		

The result of the joint test is shown above. In this test, our null hypothesis is the parameter estimates for the coefficient of variable **beds** and **nurses** are both equal to 0. From the table, we can find that the p-value is 0.2924, which is greater than our significant level 0.05, meaning that we cannot reject the null hypothesis and can conclude that there is no significant difference between each of the parameter estimates for the coefficient of variable **beds**, **nurses**, and 0.

2.(a)

```
proc reg data=d.spirometry;
    model fev1 = age weight;
run;
quit;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7.14313	3.57156	141.81	<.0001
Error	68	1.71267	0.02519		
Corrected Total	70	8.85579			

Root MSE	0.15870	R-Square	0.8066
Dependent Mean	0.82972	Adj R-Sq	0.8009
Coeff Var	19.12719		

The ANOVA table and the parameter estimates of the fitting model is shown beside. Form the table, we can obtain the value of R^2 is 0.8066.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-0.21321	0.07887	-2.70	0.0087
AGE	Age (Months)	1	0.01051	0.00168	6.27	<.0001
WEIGHT	Body Weight (kg)	1	0.02674	0.00732	3.65	0.0005

(b)

```
proc reg data=d.spirometry;
    model fev1 = age weight;
    output out=out_2 predicted=pre_value;
run;
quit;
```

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
pre_value	71	0.82972	0.31944	58.91000	0.48862	1.78346	Predicted Value of FEV1
FEV1	71	0.82972	0.35568	58.91000	0.39000	2.28000	Forced Expiratory Volume At 1 Sec (L)

```
proc corr data = out_2;
 var fev1;
 with pre_value;
run;
```

Pearson Correlation Coefficients, N = 71
Prob > |r| under H0: Rho=0

	FEV1
pre_value Predicted Value of FEV1	0.89811 <.0001

As the table we shown below, we obtain the predicted values for the model in question 2(a) and the Pearson correlation between the observed FEV1 values and the predicted values. The correlation between observed FEV1 values and the predicted one is 0.89811, and the square of it is $0.89811^2 = 0.8066$, which equals to the value of R^2 we obtained in previous question.