hw3

AUTHOR
Hanbei Xiong

Question 1 (From Lab 3)

1.1 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?

Answer:

Since the formula for the intercept in a simple linear regression model is $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, and the sample mean of the residuals from a linear regression model is $\bar{e} = 0$ given by the property of residuals, so $\bar{Y} = \bar{X} = 0$ in regression of residuals on residuals. The previous statement is due to the fact that both Y and X are residuals derived from previous regressions. Their sample mean will follow the property of residual. Hence, the intercept will always be zero.

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

Answer:

Here are interpretation for each coefficient:

regnc: The expected risk is 0.46696 lower for hospital in north central region than expected risk for hospital in north east region.

regs: The expected risk is 0.93369 lower for hospital in south region than expected risk for hospital in north east region.

regw: The expected risk is 0.47946 lower for hospital in west region than expected risk for hospital in north east region.

The intercept is the expected risk for hospital in north east region.

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

1.3 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?

Answer: We used north east region as the reference region in above code. I fit the model using north central region as the reference region. The parameter for the two unchanged variables changed and the p value is different as well. The intercept did not change much. The R^2 and ANOVA table did not change.

Here is the anova table and parameters estimates using central region as a reference region.

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
regnw	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

Here is the original model with north east as the reference region.

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

1.4 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?

Answer: Each of these p-values tells whether the mean of the expected risk for hospital in that specific region is significantly different from the mean of the expected risk for hospital in reference region The means being compared are the mean of the expected risk for hospital in the reference region and the mean of the expected risk for hospital in the region of interest.

1.5 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?

Answer:

$H_0 : \beta_{regnc} = \beta_{regs} = \beta_{regw} = 0$

$H_A : \beta_{regnc}, \beta_{regs}, \beta_{regw}$ are not all equal to zero

$F^* \sim F_{3,107}$

$F^* = 2.50$

$p = 0.0636$

Conclusion: Since p-value is greater than 0.05, we fail to reject the null hypothesis. There is no evidence that the expected risk for hospital is associated with region, after controlling for length and census.

1.6 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?

Answer:

$H_0 : \beta_{length} = \beta_{census} = \beta_{regnc} = \beta_{regs} = \beta_{regw} = 0$

$H_A : \beta_{length}, \beta_{census}, \beta_{regnc}, \beta_{regs}, \beta_{regw}$ are not all equal to zero

$F^* \sim F_{5,107}$

$F^* = 11.59$

$p < 0.0001$

Conclusion: Since p-value is less than 0.05, we reject the null hypothesis. There is evidence that at least one of the variable is significant in associating with the expected risk for hospital.

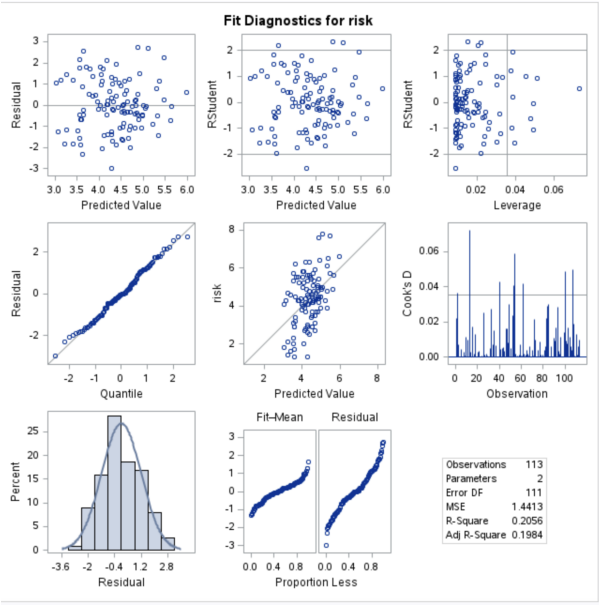
1.7 A regression of risk on xray shows a highly significant relationship. Fit the model and conduct model diagnostics (residuals analysis). Report this relationship. Provide an interpretation of the regression coefficients, using appropriate units.

Answer:

Simple Linear Regression Model:

$\hat{Y}_i = 1.79202 + 0.03140 x_i$

Here is model diagostics:



Thee assumptions of residuals are met after examining the residual plots. It shows a linear regression. For the coefficient, the intercept is 1.79202 which means the expected risk is 1.79202 percent when xray is 0. The slope is 0.03140 which means the expected risk increases 0.03140 percent when xray increases 1 percent.

Investigators hypothesize that xray will be significantly related to risk after controlling for beds, nurses and svcs. What model should you fit to test this hypothesis? Fit the model and report the results.

Given the new assumption, I should fit a multiple linear regression with risk regressed on variables beds, nurses, svcs and xray. The result is here:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	71.10022	17.77506	14.74	<.0001
Error	108	130.27960	1.20629		
Corrected Total	112	201.37982			

Root MSE	1.09831	R-Square	0.3531
Dependent Mean	4.35487	Adj R-Sq	0.3291
Coeff Var	25.22037		

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

Do the results support the hypothesis or not?

The result supports the hypothesis that xray is significantly related to risk since the p-value in partial F test for xray is less than 0.0001.

Provide an interpretation of each of the regression coefficients, using appropriate units. Also report the partial correlation of each predictor with risk.

Here are interpretations for each regression coefficient:

Intercept: The expected risk is 0.86917 percent when xray, beds, nurses, and svcs are 0.

xray: The expected risk increases 0.02858 percent for one ratio increase in the ratio of number of x-ray performed to number of patients without signs or symptoms of pneumonia, when holding other variables constant.

beds: The expected risk decreases 0.0003393 percent for one number increase in average number of beds in hospital during study period, when holding other variables constant.

nurses: The expected risk increases 0.00223 percent for one unit number in average number of full-time equivalent nurses during study period, when holding other variables constant.

svcs: The expected risk increases 0.01976 percent for one percent increase in percent of 35 potential facilities and services that are provided by the hospital, when holding other variables constant.

The partial correlation of each predictor with risk is:

xray: 0.45260

beds: -0.02311

nurses: 0.11162

svcs: 0.16153

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

$H_0 : \beta_{beds} = \beta_{nurses} = 0$

$H_A : \beta_{beds}, \beta_{nurses}$ are not all equal to zero

The REG Procedure
Model: MODEL1

Test test_region 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		

Conclusion: Since the pvalue is greater than 0.05, we fail to reject the null hypothesis. There is not enough evidence that beds and nurses contribute to explaining variation in risk after controlling for svcs and xray.

Question 2

Another interpretation of the coefficient of multiple determination, R^2: The value of R^2 is equal to the square of the Pearson correlation between the observed values Y_i and the predicted values $\hat{Y}_i$. In this sense, R^2 is a direct measure of how well the model fits the observed data. This can be proven mathematically, but we will forgo the proof and just show that this is true for an example:

a. For the spirometry/FEV1 data, fit the model regressing FEV1 on age and weight and obtain the value of R^2.
b. Obtain the predicted values for this model and the Pearson correlation between the observed FEV1 values and the predicted values. Confirm that the square of this correlation is equal to the value of R^2.

(a) Answer:

The fitted model is $\hat{Y} = -0.21321 + 0.01051\hat{\beta}_{AGE} + 0.02674\hat{\beta}_{WEIGHT}$

We obtain $R^2 = 0.8066$.

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

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) Answer:

The Pearson correlation is 0.89811. The square of pearson correlation is 0.8066 which equals to R^2 we derived in part a).

Pearson Correlation Coefficients, N = 71 Prob > |r| under H0: Rho=0		
	FEV1	predicted_value
FEV1 Forced Expiratory Volume At 1 Sec (L)	1.00000	0.89811 <.0001
predicted_value Predicted Value of FEV1	0.89811 <.0001	1.00000