Solutions for Mid-Term Test for MAT3775 Fall 2013
Regression Analysis

[20] 1. Data is collected on wheat production, with yield measured in terms of soil moisture. The results from 30 independent fields are collected.

(a) Identify the response and explanatory variable.
Solution :

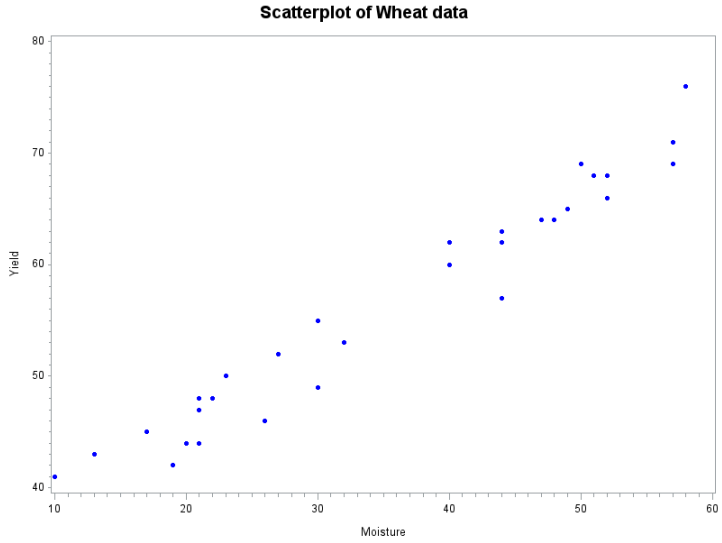
Response : yield (wheat production) (let's call it Y)
Explanatory : soil moisture (x)

(b) Write down the simple linear regression model to represent the data, and describe what each variable represent.
Solution :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, 30$$

where y_i is the yield of wheat of the i-th field, x_i the i-th field's moisture content of the soil and the ϵ_i's are independent uncorrelated errors with common variance σ^2 and mean 0. β_0 is the intercept of the model and β_1 the slope parameter. For inference purposes, the errors are assumed to be normally distributed.

(c) Take a look at the scatter plot below. Explain why the assumptions of linear regression appear reasonable for the data.



Solution :

The assumptions are that y and x have roughly a linear relationship and that the variance of the errors and thus of $y|x$ is constant, which seems to be the case here.

(d) Use the statistics below to compute the least squares estimates b_0 and b_1.

$$\sum_{i=1}^{30} x_i = 1065 \qquad \sum_{i=1}^{30} y_i = 1691$$
$$\sum_{i=1}^{30} x_i^2 = 44177 \qquad \sum_{i=1}^{30} y_i^2 = 98429$$
$$\sum_{i=1}^{30} (x_i - \bar{x})^2 = 6369.5 \qquad \sum_{i=1}^{30} (y_i - \bar{y})^2 = 3112.97$$
$$\sum_{i=1}^{30} (x_i - \bar{x})((y_i - \bar{y}) = 4349.5 \qquad \sum_{i=1}^{30} x_i y_i = 64380$$

Solution :

$$b_1 = \frac{S_{XY}}{S_{XX}} = \frac{4349.5}{6369.5} = 0.68286$$

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{1691}{30} - 0.68286 \times \frac{1065}{30} = 56.3667 - 0.68286 \times 35.5 = 32.125$$

(e) Write down the null hypothesis for testing for an effect of the explanatory variable on the response and compute the value of the test statistic.
Solution :

$H_0 : \beta_1 = 0$ Could use either the t test or the F test from the ANOVA table. Formula for the t test

$$t = \frac{b_1 - 0}{s/\sqrt{S_{XX}}}$$

where s is the square root of the mean square error given in question 2 as $MSE = SS_{Res}/(n-2) = 5.1$. Thus we have

$$t = \frac{0.68286 - 0}{\sqrt{5.1/6369.5}} = 24.13$$

Formula for the F test is

$$F = \frac{SS_{Reg}/1}{SS_{Res}/(n-2)} = \frac{SS_{Tot} - SS_{Res}}{5.1} = \frac{3112.97 - 5.1 \times 28}{5.1} = 582.17$$

(f) What are the degrees of freedom associated with the test from part 1 (e) ?
Solution :

For the t test, it is $n - 2 = 28$ degrees of freedom. For the F test, it's 1 (numerator) and 28 (denominator) df's.

(g) SAS reports that the p-value is <0.0001. The p-value is the probability of which event ?
Solution :

The p-value is the probability of observing a t (or F) statistic as extreme or more extreme than the one observed under the assumption that there is no relationship between the covariate and the response.

(h) Interpret b_0 and b_1 in the context of the study using one sentence for each.
Solution :

b_0 is the estimated yield of wheat of a field with a moisture content of 0 (thus a dry field is expected to produce 32.125 units of production of wheat, whatever those are).
b_1 is the slope of the model, and thus under the model, for each increase of one unit of moisture for a field, the yield is expected to increase on average by 0.68286 units of production.

(i) Compute de value of the coefficient of determination R^2 and interpret it within the context of this study.
Solution :

$$R^2 = 1 - \frac{SS_{Res}}{SS_{Tot}} = 1 - \frac{5.1 \times 28}{3112.97} = 1 - \frac{142.8}{3112.97} = 0.9541$$

So 95.41% of the variability in yield of wheat can be explained by the variability of the moisture content according to our linear model.

[10] 2. Consider the wheat study from Question 1. SAS reports that the Mean Squared Error is 5.10.

(a) Find the expected yield for a moisture content of 25.
Solution :

Fitted value is $\hat{y}(x_0) = b_0 + b_1(x_0) = 32.125 + 0.68286 \times 25 = 49.197$

(b) Write down the formula for a confidence interval for the mean.
Solution :

$$\hat{y}(x_0) \pm t_{\alpha/2}(n-2)s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}$$

(c) Using $t_{0.025}(28) = 2.0484$, give the 95% confidence interval for the mean yield at a moisture of 25.
Solution :

$$\begin{aligned} 49.197 \pm 2.0484\sqrt{5.1}\sqrt{\frac{1}{30} + \frac{(25-35.5)^2}{6369.5}} &= 49.197 \pm 2.0484 \times 2.2583 \times \sqrt{0.05064} \\ &= 49.197 \pm 1.041 = [48.156, 50.238] \end{aligned}$$

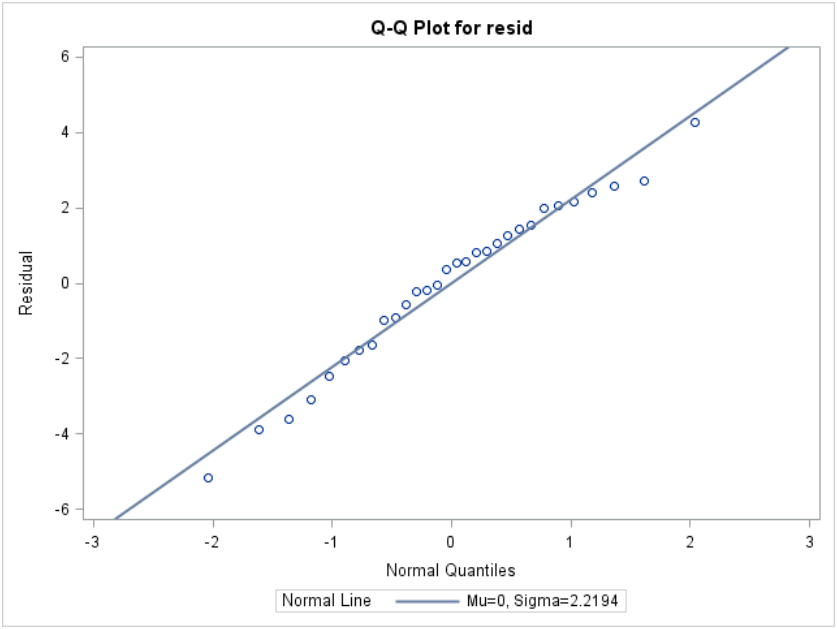
(d) Write down the formula for a prediction interval for the next observation of yield at the moisture of 25.
Solution :

$$\hat{y}(x_0) \pm t_{\alpha/2}(n-2)s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{XX}}}$$

(e) Compute the 95% prediction interval for yield at a moisture of 25.
Solution :

$$49.197 \pm 2.0484\sqrt{5.1}\sqrt{1+\frac{1}{30}+\frac{(25-35.5)^2}{6369.5}} = 49.197 \pm 2.0484 \times 2.2583 \times \sqrt{1.05064}$$
$$= 49.197 \pm 4.742 = [44.455, 53.939]$$

[10] 3. Consider the wheat study from Questions 1 and 2. Take a look at the residual Q-Q plot.



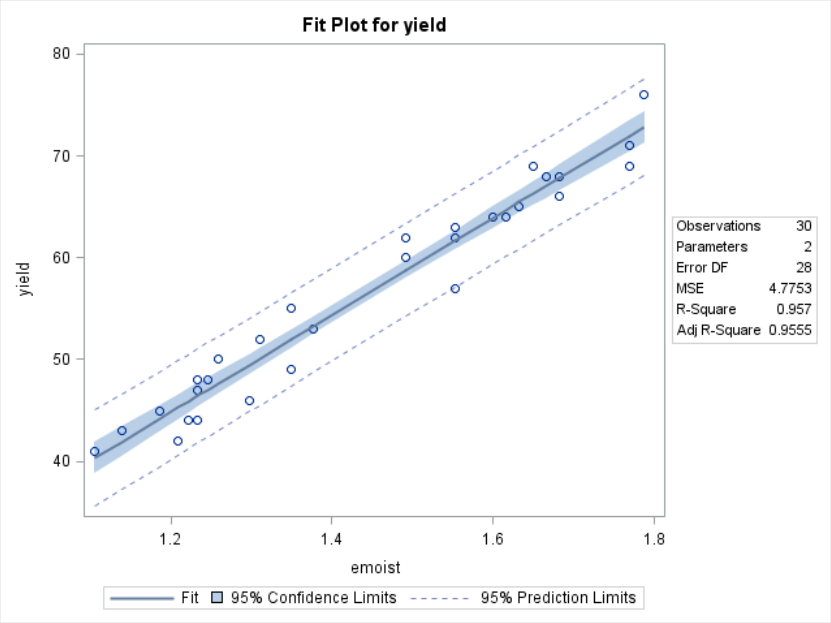
(a) There appears to be a bit of a pattern in this Q-Q plot. Interpret what this means with respect to the simple linear regression assumption.
Solution :

There appears to be a curve in the Q-Q plot, suggesting from its shape that the residual distribution is left-skewed and thus not the normal distribution.

(b) Assuming the assumption is violated, and considering the scatter plot of question 1(c), what transformation of the data could improve our model fit ?
Solution :

Assuming there is a slight upward curve in Y with respect to X, this would suggest either $X' = X^2$ or $X' = \exp(X)$ as the transformation, as the variance of $Y|X$ seems constant.

(c) Using the appropriate transformation and running the regression again, SAS has produced the following graphic :



Give two ways in which the transformation has improved the model.
Solution :

1. R^2 is slightly larger, 0.957 instead of 0.9541
2. Smaller mean square error, 4.77 instead of 5.1.

(d) Explain why the moist variable was transformed instead of the yield variable.
Solution :

Since the variance of the yield given moisture was already pretty constant to begin with, it would have been unwise to risk violating the constancy of variance assumption by transforming the response, thus the covariate was transformed.

[10] 4. Consider the wheat study from the previous question. There is another variable that was taken into account : variety of wheat. There were 5 different varieties, labeled 1 to 5.

(a) This is now a multiple regression problem, with yield as a function of moisture and variety. Consider variety as a single, quantitative variable. Write down the model

using matrix notation, and identify the dimensions clearly.
Solution :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Where $\mathbf{y}$ is the $n \times 1$ vector of wheat yield, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ is the 3×1 vector of regression coefficients, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is an $n \times 1$ vector of errors and $\mathbf{X}$ is the $n \times 3$ design matrix of the form

$$\begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix},$$

where $n = 30$, and the columns 2 and 3 of the matrix represent soil moisture and wheat variety respectively.

(b) What are the residual degrees of freedom of this model ?
Solution :

$n - 3 = 30 - 3 = 27$ degrees of freedom.

(c) SAS produced the following table.

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	30.40848	1.23258	24.67	<.0001	27.87943	32.93754
moist	1	0.67622	0.02631	25.70	<.0001	0.62224	0.73021
variety	1	0.65076	0.27107	2.40	0.0235	0.09456	1.20696

In light of this table, holding moisture constant, does variety have an influence on yield ? Explain why.
Solution :

Yes, as the p-value for the variety regression coefficient is smaller than 0.05 or, equivalently, the confidence interval for β_2 does not contain 0.

(d) Variety could be taken as a categorical variable. Explain why.
Solution :

Variety is a label, not intrisically a quantity measurable by a number. There is no inherent hierarchy of the varieties as far as the information given tells.

(e) If variety was an unordered categorical covariate, what would be the dimension of the design matrix to insure that it has full rank ? Explain why.
Solution :

Since the model already has an intercept and the covariates for categories take the form of indicator variables, i.e. take only values of 0 or 1, to have one indicator per variety would be linearly dependent with the intercept column of X, thus, not forgetting that we have soil moisture, the appropriate dimensions of the matrix are $n \times (1 + 1 + 5 - 1) = 30 \times 6$.