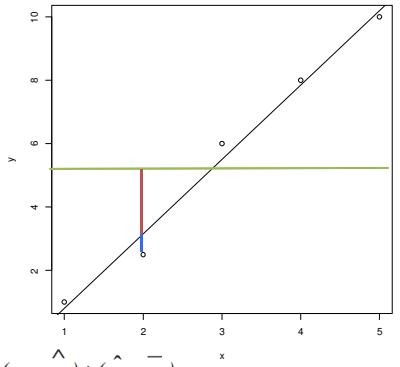
Chapter 7 Linear Regression Inference
Huamei Dong
04/07/2016

1. Correlation squared (R^2)
2. Inference for linear regression
3. Regression when predictors are categorical data

1. Correlation squared(R^2)



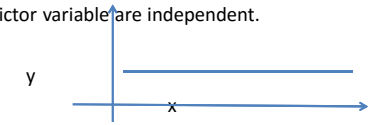
$(y_i - \overline{y}) = (y_i - \hat{y}) + (\hat{y} - \overline{y})$

The blue line represents $(y_i - \hat{y})$ which is the residual. The red line represent $(\hat{y} - \overline{y})$ Which can be explained by the linear model. The variation caused by the red part from all the points depends on correlation R. Actually it is just R^2.

2. Inference for linear regression

We have estimates b_0 b_1 for the linear model $y=\beta_0+\beta_1 x$ true parameters β_0 *and* β_1

Just like we did inference about a population mean and a population proportion ,e.g., Is there convincing evidence that the true mean or true proportion equals to 0 or some number, we wonder if there is convincing evidence that true linear model has slope 0, i.e., response variable and predictor variable are independent.



We did independent test for two way table . There we use Chi-square test. That is for categorical variables.

	Obama	Democrats	Republican	Total
Approval	842 (E=2119x1458/4223 =731.6)	736 (E=2119x1382/4223 =693.45)	541 (E=2119x1383/4223 =693.96)	2119
Disapprove	616 (E=2104x1458/4223 =726.4)	646 (E=2104x1382/4223 =688.55)	842 (E=2104x1383/4223 =689.04)	2104
total	1458	1382	1383	4223

In linear regression our response variable is continuous and our explanatory variable is continuous or categorical. We would like to see if the true linear model $y=\beta_0+\beta_1 x$ has slope $\beta_1=0$

H_0: $\beta_1=0$ verse H_A: $\beta_1 \neq 0$

Like all the hypothesis tests we did before for means and proportions , we need calculate the test statistics and then find p-value and draw a conclusion.

Here we use T test because if H_0 is true, the sampling distribution of $\hat{\beta}_1$ is T-distributed.

Actually $SE(\hat{\beta}_1) = \sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \sqrt{\dfrac{sum\ of\ residuals'\ squares}{(n-1)(n-2)(s_x)^2}}$

The calculation of SE is quite complicate. We use computer to calculate.

Example 1 Using the data (low_birth_weight_infants.txt):

(1) Find the least square regression line
(2) Test whether the slope is zero at significance level of 0.05

Answer:

```
>Beta1Estimate<-
sd(birth$headcirc)/sd(birth$gestage)*cor(birth$headcirc,birth$gestage)
>birthlm<-lm(birth$headcirc~birth$gestage)
>Beta1SE<-sqrt(sum(residuals(birthlm)^2)/(100-1)/100-2)/sd(birth$gestage)^2)
```

$$T=\frac{\hat{\beta}_1-0}{SE(\hat{\beta}_1)}=\frac{0.78}{0.063}=12.38$$

With degree of freedom n-2=98

```
Call:
lm(formula = birth$headcirc ~ birth$gestage)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5358 -0.8760 -0.1458  0.9041  6.9041

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.91426    1.82915    2.14   0.0348 *
birth$gestage 0.78005    0.06307   12.37   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.59 on 98 degrees of freedom
Multiple R-squared:  0.6095, Adjusted R-squared:  0.6055
F-statistic: 152.9 on 1 and 98 DF,  p-value: < 2.2e-
```

Conclusion: p-value is very small. So we reject the null hypothesis $\beta_1=0$

3. Regression when predictors are categorical data

Categorical variables are also useful in predicting outcomes. Here we consider the Following example.

Example 2 The data ("low_birth_weight_infants.txt") is a sample of 100 low birth infants born in Boston.

(1) Is there any correlation between toxemia and head circumference from a scatter plot? If there is, what is the correlation coefficient?
(2) Find least squares regression line using summary statistics.
(3) Find least squares regression line using R. Are they the same?
(4) Plot the residuals. Do the residuals satisfy the conditions: **linearity , nearly normal and constant variability?**
(5) Test whether toxemia and head circumference are independent (or the slope is 0) at significance level of 0.05.