6.S085 Statistics for Research Projects **IAP 2014**

Lecture 3: January 23

Lecturer: Ramesh Sridharan and George Chen *Notes by: William Li*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

3.1 Linear Regression

Be sure to visualize! Recall Anscombe's Quartet.

Goal of linear regression: fit a line, $y = \beta_0 + \beta_1 x$ to data

y: dependent variable/response variable

β_0: intercept

β_1: slope (e.g. spring constant in Hooke's law)

x: independent variable/predictor variable

Slope close to 0: little/no relationship between x and y

3.2 Probabilistic Model

Observe $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$

Assume:

$y_i(x_i) = \beta_0 + \beta_1 x_i + \epsilon_i$

ϵ_i: normally distributed with mean 0, variance σ^2 ($N(0, \sigma^2)$)

β_0 and β_1 are fixed but unknown.

Under this model, there is a "good" way of to estimate β_0, β_1:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y - (\beta_0 + \beta_1 x_i))^2$$

(This is called least-squares linear regression)

Solution:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$$

Correlation coefficient: $r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$

where $s_x = \sqrt{\frac{1}{n-1}\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}$

r^2 is called the "coefficient of determination", $-1 \leq r \leq 1$

Correlation does not imply casuation, e.g. sun cycles was correlated with the number of Republicans in the Senate in the 1980s

3.3 Hypothesis Testing

Warning: Everything here assumes that the above probabilsitic model is true

3.3.1 Slope

$t_{\beta_1} = \frac{\hat{\beta}_1 - \beta_1}{s_{\beta_1}} \sim t_{n-2}$ (t distributed with $n-1$ degrees of freedom

$s_{beta_1} = \frac{\hat{\sigma}}{\sqrt{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}}$

$\hat{\sigma}^2 = \sum\limits_{i=1}^{n} \frac{(\hat{y_i} - y_i)^2)^2}{n-2}$

Let's break down these terms:

$\sqrt{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}$: How close together the x_i's are

Intuition:

When the x_i values are very close together, it's hard to fit a line. This makes s_{β_1} large. This makes t_{β_1} smaller (closer to 0). For a null hypothesis (H_0) of $\beta_1 = 0$, there will be more area outside of $[-t_{\beta_1}, t_{\beta_1}]$, which means that your statistical significance will go down. (Your p-value will be higher, or it will be harder to achieve your threshold of statistical significance).

When there is a large error in fit, the term $\hat{\sigma}^2$ will be large. This makes $s_{\beta_1} large...$ your statistical significance will go down (following the line of reasoning above).

3.3.2 Intercept

$t_{\beta_0} = \frac{\hat{\beta}_0 - \beta_0}{s_{\beta_0}} \sim t_{n-2}$

$s_{\beta_0} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}}$

If the x_i's are close together, then the s_{β_0} will be big. Then t_{beta_0} will be small. For a null hypothesis of $\beta_0 = 0$, there will be more area outside of $[-t_{\beta_0}, t_{\beta_0}]$, which means that your statistical significance will go down (your p-value will be higher).

Bonferroni correction: divide the p-value threshold by 2 to see if you have significance at p=0.05 (i.e. is p¡0.025 for both of them?)

3.3.3 Correlation Coefficient

$t_r = \sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$

3.3.4 Prediction

What can we say about a new point generated from same probabilistic model with x-value x^*?

$\hat{y}(x^*) = \beta_0 + \hat{\beta}_1 x^*$

$var[\hat{\beta}_0 + \hat{\beta}_1 x^* + \epsilon] = var[\hat{\beta}_0 + \hat{\beta}_1 x^*] + var[\epsilon]$

The first term is estimated with $\hat{\sigma}^2[\frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum(x_i-\bar{x})^2}$ — this is the variance in the mean estimate of the line at x^*

The second term is estimated with $\hat{\sigma}^2$ — it is variance introduced by extra noise

3.3.5 Multiple linear regression

$((x_1, x_2, ..., x_p), y)...$ — there are p predictors for each y

Probabilistic model:

$y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$

Why not use a very high degree polynomial? It will overfit.

3.3.6 Matrix Form

Solve:

$\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i \beta))^2$

Solution:

$\hat{\beta} = (X^T X)^{-1} X^T Y$

3.4 Model Evaluation

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

First term: difference explained by the model Second term: difference not explained by the model

Residual: $\hat{y}_i - y_i$: Visualizing this is important

With a bit of algebra:

$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$

SST = SSM + SSE (sum of squares total) = (sum of squares model) + sum of squares error

(sum of squares model) / (sum of squares total) is exactly r^2!

We want SSM/SSE to be large

MSM = SSM / $p - 1$

MSE = SSE / $n - p$

For technical reasons, instead look at: MSM/MSE $\sim F_{p-1,n-p}$