Introduction

May 7, 2014 1:46 PM

Notation

In the past:
μ - variable
$\tilde{\mu}$ - Random variable of estimate
$\hat{\mu}$ - Estimate

In this class, don't use $\tilde{\mu}$ - too much notation.

Instead of:
$Y = \alpha + \beta x + R$
(Y is random variable)

we use
$y = \beta_0 + \beta_1 x + \varepsilon$
because capitals will represent matrices.

Modeling

Data: $(x_1, y_1),\ (x_2, y_2), \ldots, (x_n, y_n)$
Aim: build a model for y conditional on x. x is known - not random.

Let's assume that $y \sim N(\mu(x), \sigma^2)$
can also write $y|x$ but we'll omit the condition for simplicit of notation.

Let's use maximum likelihood to estimate the model parameters.

$$L(\mu(x), \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\,\pi}} e^{-\frac{1}{2\sigma^2}\left(y_i - \mu(x_i)\right)^2} \propto \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu(x_i)\right)^2}$$

Suppose for now that σ^2 is known. We'll estimate it later.
Maximizing the likelihood is equivalent to minimizing

$$\sum_{i=1}^{n}\left(y_i - \mu(x_i)\right)^2$$

This is the least squares approach.

Intuitively, for some given function $\mu(x)$, this approach gives the parameters that provide the best fit of $\mu(x)$ to the data.

e.g. $\mu(x) = \beta_0 + \beta_1 x$ (linear function)
least squares approach: find estimates of β_0 and β_1 that minimize
$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$

Simple Linear Regression Model

May 9, 2014 1:04 PM

Simple Linear Regression Model

$y \sim N(\beta_0 + \beta_1 x, \sigma^2)$

$y = \underbrace{\beta_0 + \beta_1 x}_{\text{structural part}} + \underbrace{\epsilon_i}_{\text{random part}}, \quad \epsilon \sim N(0, \sigma^2)$

Assumptions

i) $E(\epsilon_i) = 0 \quad (\Rightarrow E(y_i) = \beta_0 + \beta_1 x)$
ii) $V(\epsilon_i) = \sigma^2 \quad (\Rightarrow V(y_i) = \sigma^2)$
iii) $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ are indepenent
iv) (Distributional Assumption): $\epsilon \overset{\text{i.i.d.}}{\sim} N(0, \sigma), \ i = 1, 2, \dots, n$

This assumption automatically accounts for assumptions i) - iii)

Independent and Identically Distributed

Denoted i.i.d., iid, or IID
Each random variable has the same probability distribution as the others and are all mutually independent.

Interpretation of Model Parameters

- Parameters: β_0, β_1, σ
- β_0 is the **mean** value of the response (y) when the explanatory (x) is zero.
- Interpretation of β_1

$E(y|x = c) = \beta_0 + \beta_1 c$
$E(y|x = c + 1) = \beta_0 + \beta_1(c + 1)$
$\Rightarrow E(y|x = c + 1) - E(y|x = c) = \beta_1$
$\Rightarrow \beta_1$ is the average change in y for a unit increase in x.

Least Squares Estimates of β_0 and β_1

Notation

$\theta =$ True (unknown) parameter
$\hat{\theta} =$ estimate of θ based on the sample data.
$V(\hat{\theta}) =$ Variance of the sampling distribution of θ (unknown, based on model parameters)
$\hat{V}(\hat{\theta}) =$ An estimate of $V(\hat{\theta})$
$se(\hat{\theta}) = \sqrt{\hat{V}(\hat{\theta})} =$ stardard error of $\hat{\theta}$

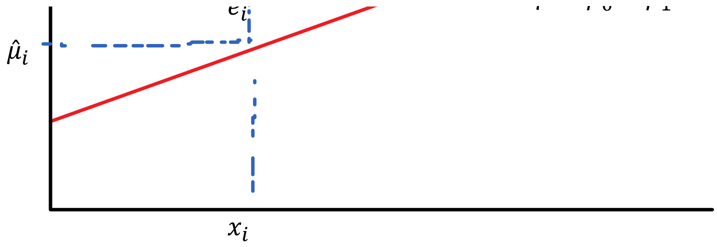
Example

$\theta = \mu$ (Simple response model $y_i \sim N(\mu, \sigma^2)$)

$\hat{\theta} = x, \quad V(\hat{\theta}) = V(x) = \frac{\sigma^2}{n}, \quad \hat{V}(\hat{\theta}) = \frac{\hat{\sigma}^2}{n}$

Finding the Least Squares Estimates (LSE)

β_0, β_1 : Unknown parameters
$\hat{\beta}_0, \hat{\beta}_1$: Estimates
True mean: $\mu_i = E(y_i) = \beta_0 + \beta_1 x_i$
Fitted values for y_i : $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
True (unknown) error: $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$
Residual (estimated) error: $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$



Minimization Procedure

Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

is minimized at $(\hat{\beta}_0, \hat{\beta}_1)$

Solve

$$i) \quad \frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$ii) \quad \frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\text{i)} \quad \Rightarrow \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$\overset{\div n}{\Rightarrow} \bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0 \Rightarrow \boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

$$\text{ii)} \quad \Rightarrow \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i \left((y_i - y) - \hat{\beta}_1 (x_i - x) \right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i (y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i (x_i - \bar{x})$$

$$\Rightarrow \boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} = \frac{S_{xy}}{S_{xx}}}$$

where

$$\boxed{S_{xy} = \sum_{i=1}^{n} x_i (y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x}) y_i}$$

$$\boxed{S_{xx} = \sum_{i=1}^{n} x_i (x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Aside:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = 0$$

Similarly,

$$\sum_{i=1}^{n} (y_i - y) = 0 \Rightarrow \sum_{i=1}^{n} (x_i - x)(y_i - y) = \sum_{i=1}^{n} x_i (y_i - y) - x \overbrace{\sum_{i=1}^{n} (y_i - y)}^{0}$$

Properties of the LES of β_0 and β_1

Expectation Value

$\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased for β_0 and β_1

$$\Rightarrow E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1$$

Proof

$E(\hat{\beta}_1) = E\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}} E(S_{xy})$

Since S_{xx} is constant but S_{xy} is random.

Use form $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}) y_i$

$$E(\hat{\beta}_1) = \frac{1}{S_{xx}} E\left(\sum_{i=1}^{n} (x_i - \bar{x}) y_i\right) = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \underbrace{E(y_i)}_{=\beta_0 + \beta_1 x_i} = \frac{1}{S_{xx}}\left(\beta_0 \sum_{i=1}^{n} (x_i - \bar{x}) + \beta_1 \underbrace{\sum_{i=1}^{n} x_i (x_i - \bar{x})}_{S_{xx}}\right) = \beta_1$$

Now,

$$E(\hat{\beta}_0) = E(\bar{y} - \hat{\beta}_1 \bar{x}) = E(\bar{y}) - \bar{x}\underbrace{E(\hat{\beta}_1)}_{\beta_1}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \beta_0 + \beta_1 \bar{x} + \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \Rightarrow E(\bar{y}) = \beta_0 + \beta_1 \bar{x}$$

$$\Rightarrow E(\hat{\beta}_0) = \beta_0 + \beta_1 \bar{x} - \bar{x}\beta_1 = \beta_0$$

Consequences

$\hat{\mu}$ is unbiased for μ. Recall $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$\mu = \beta_0 + \beta_1 x$$

$$E(\hat{\mu}) = E(\hat{\beta}_0) + E(\hat{\beta}_1) x = \beta_0 + \beta_1 x = \mu$$

Variance

Variance of $\hat{\beta}_0$ and $\hat{\beta}_1$

$$V(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

Proof

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i$$

$$V(\hat{\beta}_1) = V\left(\frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - x) y_i\right) = \frac{1}{S_{xx}^2} V\left(\sum_{i=1}^{n} (x_i - x) y_i\right) = \frac{1}{S_{xx}^2} \sum_{i=1}^{n} (x_i - x)^2 V(y_i)$$

Can bring variance into the sum because y_i are independent

$$= \frac{\sigma^2}{S_{xx}^2} S_{xx} = \frac{\sigma^2}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1}{n}\sum_{i=1}^{n} y_i - \left(\frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i\right)\bar{x} = \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right) y_i$$

Linear combination of independent y_i

$$V(\hat{\beta}_0) = \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right)^2 V(y_i) = \sigma^2 \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right)^2 = \sigma^2 \sum_{i=1}^{n}\left(\frac{1}{n^2} + \frac{(x_i - \bar{x})^2 \bar{x}^2}{S_{xx}^2} - \frac{2(x_i - \bar{x})\bar{x}}{n S_{xx}}\right)$$

$$= \sigma^2\left(\frac{n}{n^2} + \frac{S_{xx}\bar{x}^2}{S_{xx}^2} + 0\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

Consequence

$$V(\hat{\mu}_0) = \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\sigma^2$$

where μ_0 is the fitted values at $x = x_0$

Recall

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\Rightarrow \hat{\mu}_0 = \bar{y} + \hat{\beta}_1 (x_0 - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n} y_i + \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i$$

$$\Rightarrow \hat{\mu}_0 = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}} \right) y_i$$

$$\Rightarrow V(\hat{\mu}_0) = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}} \right)^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n^2} + \frac{(x_i - \bar{x})^2 (x_0 - \bar{x})^2}{S_{xx}^2} + \frac{2(x_i - \bar{x})(x_0 - \bar{x})}{nS_{xx}} \right) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

Least Squares Estimate

May 14, 2014 1:37 PM

Covariance

$\text{Cov}(U, V) = E[(U - \mu_U)(V - \mu_V)]$

Consequences of Least Squares Estimate

i) $\sum_{i=1}^{n} e_i = 0$

Residual Error $e_i = y_i - \hat{\mu}_i = y_i - \hat{\beta}_1 - \hat{\beta}_0 x$

$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} e_i = \bar{e} = 0$

ii) $\sum_{i=1}^{n} e_i x_i = 0$

$\Rightarrow \sum_{i=1}^{n}(e_i - \bar{e})x_i = 0 \Rightarrow \sum_{i=1}^{n}(e_i - \bar{e})(x_i - \bar{x}) = 0 \Rightarrow \text{Cov}(e, x) = 0$

$\Rightarrow$ Sample correlation between e and x is 0

Note: i) and ii) follow from the fact that $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize

$$S = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} e_i^2$$

$$\Rightarrow \frac{\partial S}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^{n}(y_0 - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \Rightarrow \sum_{i=1}^{n} e_i = 0$$

$$\Rightarrow \frac{\partial S}{\partial \beta_1} = 0 \Rightarrow \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \Rightarrow \sum_{i=1}^{n} e_i x_i = 0$$

iii) $\sum_{i=1}^{n} \hat{\mu}_i e_i = 0$

Since $\sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i)e_i = 0$

iv) $(\bar{x}, \bar{y})$ is always on the fitted regression line

$x = \bar{x}, \qquad \hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1\bar{x} = \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1\bar{x} = \bar{y}$

Estimate of σ^2

Recall $V(y_i) = V(\epsilon_i) = \sigma^2$

The LSE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Why $n - 2$?

n - Number of data points

2 - number of parameters estimated (excluding σ) in the structural part of the model.

Theoretically,

$$E\left(\sum_{i=1}^{n} e_i^2\right) = (n-2)\sigma^2 \Rightarrow E\left(\frac{1}{n-2}\sum_{i=1}^{n} e_i^2\right) = \sigma$$

$\Rightarrow E(\hat{\sigma}^2) = \sigma^2$

$\Rightarrow$ $\hat{\sigma}^2$ is an unbiased unbiased estimator of σ^2

Hypothesis Tests and Confidence Intervals

May 14, 2014 2:07 PM

β_1 is usually of interest
e.g. can test whether y is linearly related to x

Is $\beta_1 \neq 0$?
H_0: $\beta_1 = 0$, $H_A: \beta_1 \neq 0$

or

On average, does a unit increase in x result in a 5 unit increase in y?
$\Rightarrow$ Is $\beta_1 = 5$?
$H_0: \beta_1 = 5$, $H_A: \beta_1 \neq 5$

or

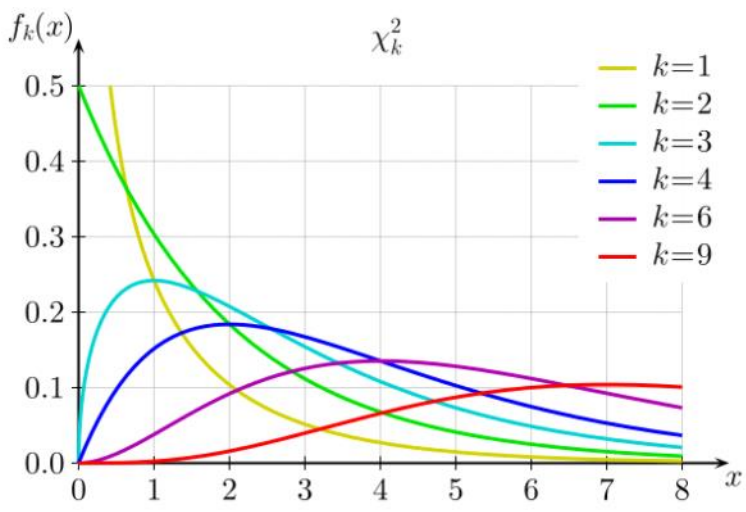
On average, does a unit increase in x result in a more than a 5 unit increase in y?
$\Rightarrow$ Is $\beta_1 > 5$?
$H_0: \beta_1 \leq 5$, $H_A: \beta_1 > 5$

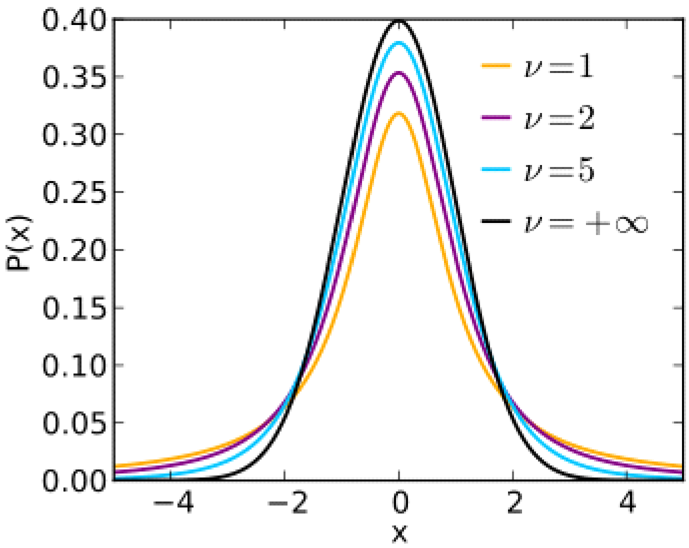
The main quantity (discrepancy measure) of interest is

$$\frac{\hat{\beta}_1 - \beta_1}{\mathrm{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}}$$

Is the number of standard deviations of the estimate from the assumed (true) value.
Before we continue, we need to determine its distribution.
Some sampling distributions

i) $X \sim N(\mu, \sigma^2)$, then $\dfrac{X-\mu}{\sigma} \sim N(0,1)$
ii) If $Z_1, \ldots, Z_n$ are i.i.d. $N(0,1)$ random variables, then $Z_i^2 \sim \chi^2(1)$, chi-squared distribution with 1 degree of freedom (d.f.)
$Z_1^2 + \cdots + Z_n^2 \sim \chi^2(n)$ chi-squared distribution with n d.f.
iii) If $Z \sim N(0,1)$ and $U \sim \chi^2(n)$ where Z is independent of U, then
$\dfrac{Z}{\sqrt{\dfrac{U}{n}}} \sim t(n)$ t distribution with n d.f.





iv) $\dfrac{\frac{\chi^2(m)}{m}}{\frac{\chi^2(n)}{n}} \sim F(m,n)$

F-distribution with m numerator d.f. and n denominator d.f.
Note: numerator and denominator are independent

Distribution of $\frac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{S_{xx}}}$

i) $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2) \Rightarrow y_i = \beta_0 + \beta_1 x + \epsilon_i \overset{\text{i.i.d.}}{\sim} N(\beta_0+\beta_1 x, \sigma^2)$

Now, $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{1}{S_{xx}}\displaystyle\sum_{i=1}^{n}(x_i-\bar{x})y_i$

$\Rightarrow \hat{\beta}_1 = \displaystyle\sum_{i=1}^{n} c_i y_i$

y_i is normal $\Rightarrow \hat{\beta}_1$ is normal

Since $E(\hat{\beta}_1) = \beta_1$ and $V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$ then $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$

$\Rightarrow \dfrac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$

ii) $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2) \Rightarrow \dfrac{\epsilon_i}{\sigma} \sim N(0,1), \qquad i = 1, 2, \ldots, n$

$\Rightarrow \left(\dfrac{\epsilon_i}{\sigma}\right)^2 \sim \chi^2(1) \Rightarrow \displaystyle\sum_{i=1}^{n}\left(\dfrac{\epsilon_i}{\sigma}\right)^2 \sim \chi^2(n)$

$\epsilon_i = y_{ii} - \beta_0 - \beta_1 x_i \Rightarrow \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2 \sim \chi^2(n)$

For every estimated parameter, we lose 1 degree of freedom.

$\Rightarrow \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}\Big(\overbrace{y_i-\hat{\beta}_0-\hat{\beta}_1 x_i}^{e_i}\Big)^2 \sim \chi^2(n-2)$

$\Rightarrow \dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n} e_i^2 \sim \chi(n-2) \Rightarrow \dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$

iii) Note: $\hat{\beta}_1$ is independent of $\hat{\sigma}^2$ (to be shown later)

$$\frac{\frac{\hat{\beta}_1-\beta_1}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\frac{\frac{(n-2)\hat{\sigma}^2}{\sigma^2}}{n-2}}} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-2)}{n-2}}} = t(n-2)$$

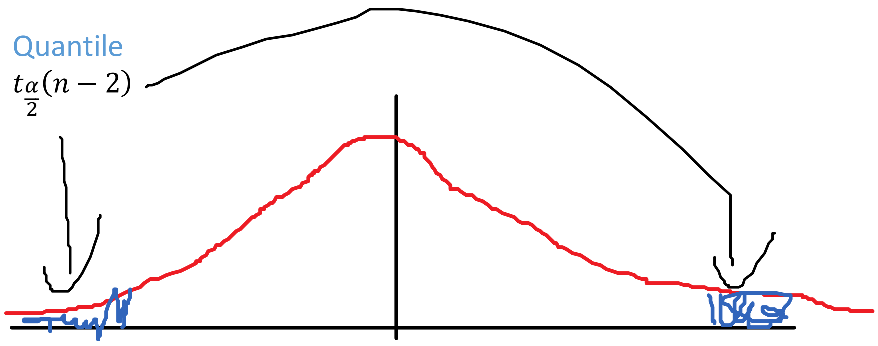
$$\Rightarrow \frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t(n-2)$$

Confidence Intervals

$100(1-\alpha)\%$ C.I. for β_1
e.g. $\alpha = 0.05$ or 0.01

Quantile

$t_{\frac{\alpha}{2}}(n-2)$



$$P\left(\left|\frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}/\sqrt{S_{xx}}}\right| < t_{\frac{\alpha}{2}}(n-2)\right) = 1-\alpha$$

$\Rightarrow$ Confidence interval for β_1 is

$$\boxed{\hat{\beta}_1 \pm t_{\frac{\alpha}{2}}(n-2)\mathrm{se}(\hat{\beta}_1)}$$

$$\mathrm{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

General Form

Estimate = (Critical Value)×(std. error)

Hypothesis Tests

May 16, 2014 1:51 PM

Two-Tailed Test

$H_0: \beta_1 = b$ v.s. $H_A: \beta_1 \neq b$
Under H_0,

$$\frac{\hat{\beta}_1 - b}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t(n-2)$$

Compute the test statistic (based on sample)

$$t^* = \frac{\hat{\beta}_1 - b}{\hat{\sigma}/\sqrt{S_{xx}}}$$

In general

$$t^* = \frac{\text{estimate} - \text{true value}}{\text{standard error}}$$

We will perform tests using significance levels (e.g. 5%, 1%, etc.)
Rule: at a $100\alpha\%$ significance level we reject H_0 if $t^* > t_{\frac{\alpha}{2}}(n-2)$

One-Tailed test

$H_0: \beta_1 \leq b$ vs $H_A: \beta_1 > b$
Test statistic under H_0 is

$$t^* = \frac{\hat{\beta}_1 - b}{\sigma/\sqrt{S_{xx}}}$$

Reject H_0 if $t^* > t_\alpha(n-2)$
Otherwise fail to reject H_0

Aside

Suppose we constructed a 95% confidence interval for β_1
If we test $H_0: \beta_1 = b$ vs $H_A: \beta_1 \neq b$ at a 5% significance level, then we reject H_0 iff b does not lie in the above confidence interval.

Predictions and Prediction Intervals

May 16, 2014 3:33 PM

Given $x = x_p$, what is the predicted y?

Notes:

i) Predicted value $\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$
ii) y_p is a random variable (future unknown value), independent of our sample.
iii) We **cannot** write $E(\hat{y}_p) = y_p$ (y_p is a R.V., not a value)
iv) The prediction error is $\hat{y}_p - y_p$ (main quantity of interest when forming prediction intervals)
v) $E(\hat{y}_p - y_p) = E(\hat{y}_p) - E(y_p) = E(\hat{\beta}_0 + \hat{\beta}_1 x_p) - E(\beta_0 + \beta_1 x_p + \epsilon_p)$
$= \beta_0 + \beta_1 x_p - (\beta_0 + \beta_1 x_p + 0) = 0$
Unbiased prediction
vi) $V(\hat{y}_p - y_p) = V(\hat{y}_p) + V(y_p)$ (y_p is independent of the sample)
$= V(\hat{y}_p) + \sigma^2 = V(\hat{\mu}_p) + \sigma^2$
$$V(\hat{\mu}_p) = \sigma^2\left(\frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right)$$
$$V(\hat{y}_p - y_p) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right)$$

Overall

$$E(\hat{y}_p - y_p) = 0$$
$$V(\hat{y}_p - y_p) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}\right)$$
$$\mathrm{se}(\hat{y}_p - y_p) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}}}$$
$\Rightarrow 100(1-\alpha)\%$ prediction interval (P.I.) for y_p is
$$P\left(\left|\frac{(\hat{y}_p - y_p) - 0}{\mathrm{se}(\hat{y}_p - y_p)}\right| < t_{\frac{\alpha}{2}}(n-2)\right) = 1 - \alpha$$
$$\Rightarrow \hat{y}_p \pm t_{\frac{\alpha}{2}}(n-2)\mathrm{se}(\hat{y}_p - y_p)$$

Analysis of Variance (ANOVA)

In the simple regression case, we use this to test $H_0: \beta_1 = 0$

Model:
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \dots, n$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$
$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
If $\beta_1 = 0$, then $y_i = \beta_0 + \epsilon_i$ and $\hat{\beta}_0 = \bar{y}$

The idea of ANOVA is to separate the total variability (SST) into two components:

i) Variability due to (or explained by) regression. (Sum of squared regression or SSR)
ii) Variability due to error (Sum of squared errors or SSE)

Write $y_i - \bar{y} = (y_i - \hat{\mu}_i) + (\hat{\mu}_i - \bar{y})$
$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
Decompose SST

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\left((y_i - \hat{\mu}_i) + (\hat{\mu}_i - \bar{y})\right)^2$$
$$= \sum_{i=1}^{n}\left((y_i - \hat{\mu}_i)^2 + (\hat{\mu}_i - \bar{y})^2 + 2(y_i - \hat{\mu}_i)(\hat{\mu}_i - \bar{y})\right)$$
$$= \underbrace{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{\mu}_i - \bar{y})^2}_{\text{SSR}} + \underbrace{2\sum_{i=1}^{n}(y_i - \hat{\mu}_i)(\hat{\mu}_i - \bar{y})}_{\text{Cross Term}}$$

i) $\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 = \sum_{i=1}^{n} e_i^2$

ii) $\text{SSR} = \sum_{i=1}^{n}(\hat{\mu}_i - \bar{y})^2 = \hat{\beta}_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2 = \hat{\beta}_1^2 S_{xx}$

iii) Cross Term

$$= 2\sum_{i=1}^{n}\underbrace{(y_i - \hat{\mu}_i)}_{e_i}\underbrace{(\hat{\mu}_i - \bar{y})}_{\hat{\beta}_1(x_i - \bar{x})} = 2\hat{\beta}_1\sum_{i=1}^{n} e_i(x_i - \bar{x}) = 2\hat{\beta}_1\left(\sum_{i=1}^{n} e_i x_i - \bar{x}\sum_{i=1}^{n} e_i\right) = 2\hat{\beta}_1(0 - 0)$$
$$= 0$$

$\Rightarrow \text{SST} = \text{SSR} + \text{SSE}$

Where $\text{SSR} = \hat{\beta}_1^2 S_{xx}$ and $\text{SSE} = \sum_{i=1}^{n} e_i^2$

We will now consider the ratio (dividing top/bottom by number of degrees of freedom)

$$\frac{\text{SSR}/1}{\text{SSE}/n-2}$$

Note: $\hat{\sigma}^2 = \dfrac{\text{SSE}}{n-2} = \dfrac{1}{n-2}\sum_{i=1}^{n} e_i^2$

Recall $E(\hat{\sigma}^2) = \sigma^2 \Rightarrow E\left(\frac{\text{SSE}}{n-2}\right) = \sigma^2$

$$E(\text{SSR}) = E\left(\hat{\beta}_1^2 S_{xx}\right) = S_{xx}E\left(\hat{\beta}_1^2\right) = S_{xx}\left(\text{Var}(\hat{\beta}_1) + E^2(\hat{\beta}_1)\right) = S_{xx}\left(\frac{\sigma^2}{S_{xx}} + \beta_1^2\right) = \sigma^2 + \beta_1^2 S_{xx}$$

So if $\beta_1 \neq 0$, the numerator will be greater than the denominator. Otherwise it will be close to 0

Distribution of the Ratio

$N^2(0,1) = \chi_1^2(1)$

$$\sum_{i=1}^{n}\chi_1^2(1) = \chi^2(n)$$

$$\frac{\chi^2(m)/m}{\chi^2(n)/n} = F(m,n)$$

Under the $H_0: \beta_1 = 0$,

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{V(\hat{\beta}_1)}} = \frac{\hat{\beta}_1}{\sqrt{V(\hat{\beta}_1)}} \sim N(0,1) \Rightarrow \frac{\hat{\beta}_1^2}{V(\hat{\beta}_1)} \sim \chi^2(1)$$

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$\Rightarrow \frac{\hat{\beta}_1^2 S_{xx}}{\sigma^2} \sim \chi^2(1) \Rightarrow \frac{\text{SSR}}{\sigma^2} \sim \chi^2(1)$$

Also,

$$\epsilon_i \sim N(0,\sigma) \Rightarrow \frac{\epsilon_i}{\sigma} \sim N(0,1) \Rightarrow \frac{\epsilon_i^2}{\sigma^2} \sim \chi^2(1) \Rightarrow \frac{\sum_{i=1}^{n}\epsilon_i^2}{\sigma^2} \sim \chi^2(n)$$

$$\Rightarrow \frac{\text{SSE}}{\sigma^2} = \frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \sim \chi^2(n-2)$$

Note: SSR is independent of SSE

$$\Rightarrow \frac{\frac{\text{SSR}}{\sigma^2}\Big/1}{\frac{\text{SSE}}{\sigma^2}\Big/n-2} = \frac{\text{SSR}/1}{\text{SSE}/n-2} \sim F(1, n-2)$$

Aside: Sometimes write

$$\text{MSR} = \frac{\text{SSR}}{1} = \text{Mean Squared Regression}$$

$$\text{MSF} = \frac{\text{SSF}}{n-2} = \text{Mean Squared Error}$$

so that

$$F = \frac{\text{MSR}}{\text{MSE}} \sim F(1, n-2)$$

Rule

Compute

$$F^2 = \frac{\text{SSR}/1}{\text{SSE}/n-2}$$

If $F^* > F_\alpha(1, n-2)$, reject H_0

Last Class:

ANOVA

$$\text{F-Statistic} = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/1}{\text{SSE}/n-2}$$

Coefficient of Determination

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}}$$

is a measure of goodness of fit.

Properties

i) $0 \le R^2 \le 1$

ii) $R^2 = 1$ iff $\text{SSE} = 0 \Leftrightarrow$ All $e_i = 0$ (perfect fit)

iii) $R^2 = 0$ iff $\text{SSR} = 0 \Rightarrow \hat{\mu}_i = \bar{y}$ (flat fitted line)

iv) $R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}} = \frac{\left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \left(\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\right)^2 = r^2$

$$\text{SSR} = \sum_{i=1}^{n} (\hat{\mu}_i - \bar{y})^2$$

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Back to Tutorial 1

Q8 Predicted Value

$= \hat{\beta}_0 + \hat{\beta}_1 120 = \cdots = 1637.687$

95% PI

$$\hat{y}_p \pm t_{0.025}(98)\sqrt{\hat{V}(\hat{y}_p - y_p)}$$

$$\hat{V}(\hat{y}_p - y_p) = \sigma^2\left(1 + \frac{1}{n} + \frac{(120-\bar{x})^2}{S_{xx}}\right)$$

$\Rightarrow 1637.687 \pm 1.9896 \times 428.365 \Rightarrow (787.587, 2487.767)$

Review of Matrix Algebra

May 28, 2014 1:41 PM

Important Results

Notation

$A_{m\times n} = \left(a_{ij}\right)_{m\times n}$ ← matrix of constants. m rows and n columns

$\left.\begin{array}{l} x = (x_1, \dots, x_n) \\ y = (y_1, \dots, y_n) \end{array}\right\}$ vectors of constants

i) $A_{n\times n}$ is symmetric if $A^T = A$

ii) Orthogonal vectors and matrices:

→ 2 vectors are orthogonal if $x^T y = 0$

→ Orthogonal Matrices: A is orthognal iff $A^T A = AA^T = I$

iii) Vectors $x_1, x_2, \dots, x_n$ are linearly independent iff $c_1 x_1 + c_2 x_2 + \cdots + c_n x_n = 0 \Rightarrow c_1 = c_2 = \cdots = c_n = 0$

iv) Rank of a matrix rank$(A_{m\times n})$ = max # of linearly independent columns

v) Trace tr$(A_{m\times m}) = \sum_{i=1}^{m} a_{ii}$

tr$(A_{m\times n} B_{n\times m}) = tr(B_{n\times m} A_{m\times n})$

If $c \in \mathbb{R}$, tr$(c) = c$

vi) Eigenvectors and Eigenvalues of a square matrix. A vector $v \neq 0$ is called an eigenvector of A if $\exists \lambda \ni Av = \lambda v$

vii) Spectral Decomposition: eigenvalue of a symmetric matrix.

For a symmetric matrix $A_{n\times n}$, the eigenvalues $\lambda_1, \dots, \lambda_n$ are real.

Furthermore, $\exists$ an orthogonal matrx $P \ni A = P\Lambda P^T$ where $\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$ and $P = [v_1, v_2, \dots, v_n]$

viii) Idempotent Matrix: $A_{n\times n}$ is idempotent if $A^2 = A \cdot A = A$

a. If A is idempotent then all eigenvalues are either zero or one.

Proof

$\lambda v = Av = AAv = A(\lambda v) = \lambda A v = \lambda^2 v$

$v \neq 0 \Rightarrow \lambda = \lambda^2 \Rightarrow \lambda \in \{0, 1\}$

b. If A idempotent, $\exists$ an orthogonal matrix $P \ni A = P\Lambda P^T$ where

$$\Lambda = \begin{bmatrix} 1 & & & & & & \\ & 1 & & & & 0 & \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & & 0 & & \\ & 0 & & & & \ddots & \\ & & & & & & 0 \\ & & & & & & & 0 \end{bmatrix}$$

$$\text{tr}(A) = \text{tr}(P\Lambda P^T) = \text{tr}(PP^T\Lambda) = \text{tr}(I\Lambda) = \text{tr}(\Lambda) = \sum_{i=1}^{n} \lambda_i = \#\text{ of 1 eigenvalues}$$

Random Vectors

May 28, 2014 2:11 PM

Random Vector

Suppose $y_1, \dots, y_n$ are random variables such that $E(y_i) = \mu_i, V(y_i) = \sigma_i^2$ and $\text{Cov}(y_i, y_j) = \sigma_{ij}$

$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = (y_1, \dots, y_n)^T$ is called a **random vector.**

Expected value of a random vector:

$$E(y) = \begin{bmatrix} E(y_1) \\ E(y_2) \\ \vdots \\ E(y_n) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

Variance-Covariance Matrix

$$V(y) = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22}^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn}^2 \end{bmatrix} = \{\text{Cov}(y_i, y_j)\}_{n \times n}$$

- It is a symmetric matrix
- If $y_1, \dots, y_n$ are uncorrelated then

  $$\text{Cov}(y_i, y_j) = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

  $$\Rightarrow V(y) = \begin{bmatrix} \sigma_{11}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{22}^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{nn}^2 \end{bmatrix}$$

  - If $\sigma_i^2 = \sigma^2$ then $V(y) = \sigma^2 I_{n \times n}$

Can write

$V(y) = E[(y-\mu)(y-\mu)^T]$

Results on E(y) and V(y)

Notation:

$y = (y_1, \dots, y_n) \leftarrow$ random vector

$A = \{a_{ij}\}_{p \times n} \leftarrow$ matrix of constants

$b = (b_1, \dots, b_p)^T \leftarrow$ vector of constants

$c = (c_1, \dots, c_n) \leftarrow$ vector of constants

Results

i) $E(A_{p \times n} y_{1 \times n} + b_{p \times 1}) = AE(y) + b$

ii) $\text{Var}(y + c) + \text{Var}(y)$

iii) $\text{Var}(Ay) = AV(y)A^T$

$\text{Var}(Ay) = E[(Ay - A\mu)(Ay - A\mu)^T] = E[A(y-\mu) \cdot [A(y-\mu)]^T] = AE((y-\mu)(y-\mu)^T)A^T = AV(y)A^T$

Aside - Question about homework

$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n-1}$

$E[\hat{\sigma}^2] = \sigma^2 \Leftrightarrow E\left[\sum_{i=1}^n (y_i - \hat{\beta}x_i)^2\right] = (n-1)\sigma^2$

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

Warning

y_i and $\bar{y}$ are not independent

Multivariate Normal

May 30, 2014 1:31 PM

Multivariate Normal Distribution

If $y = (y_1, \dots, y_n)^T$ follows a multivariate normal distribution then

$$f(y) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)}$$

where $\mu = E(y) =$ mean vector; and
$\Sigma = V(y) =$ Variance-Covariance matrix
We write $y \sim MVN(\mu, \Sigma)$

1) If $\boldsymbol{y} \sim MVN(\boldsymbol{\mu}, \Sigma)$, then
   $\boldsymbol{u} = A\boldsymbol{y} \sim MVN(A\boldsymbol{\mu}, A\Sigma A^T)$
2) If $\boldsymbol{y} \sim MVN(\boldsymbol{\mu}, \Sigma)$, zero correlation implies independence of $y_1, \dots, y_n$
3) If $\boldsymbol{y} \sim MVN(\boldsymbol{\mu}, \Sigma)$, and $\boldsymbol{u} = A\boldsymbol{y}$, $\boldsymbol{w} = B\boldsymbol{y}$, then $\boldsymbol{u}$ and $\boldsymbol{w}$ are independent iff $AV(y)B^T = 0$

Proof of 3)

$\text{Cov}(\boldsymbol{u}, \boldsymbol{w}) = E[(\boldsymbol{u} - A\boldsymbol{\mu})(\boldsymbol{w} - B\boldsymbol{\mu})^T] = AV(y)B^T$
So if Cov $= 0$ then we have independence

4) If $\boldsymbol{y} \sim MVN(\boldsymbol{0}, I)$, then
   i) $y_1, \dots, y_n \overset{\text{i.i.d.}}{\sim} N(0,1)$
   ii) $\boldsymbol{y}^T\boldsymbol{y} = \sum_{i=1}^{n} y_i^2 \sim \chi^2(n)$
   iii) If $\boldsymbol{z} = P\boldsymbol{y}$, where P is orthogonal $(PP^T = P^TP = I)$ then $\boldsymbol{z} \sim MVN(\boldsymbol{0}, I)$
   iv) If $\boldsymbol{y} \sim MVN(\boldsymbol{0}, \Sigma)$, where $\Sigma = P\Lambda P^T$ (Recall that Σ is symmetric $\Rightarrow P$ is orthogonal) then
   $\left(\Lambda^{-\frac{1}{2}}P^T\right)y \sim MVN(0, I)$

Matrix and Vector Differentiation

1) $f(y) = f(y_1, y_2, \dots, y_n)$

$$\Rightarrow \frac{d}{dy}f(y) = \begin{bmatrix} \frac{\partial}{\partial y_1}f(y) \\ \frac{\partial}{\partial y_2}f(y) \\ \vdots \\ \frac{\partial}{\partial y_n}f(y) \end{bmatrix}$$

2) $\boldsymbol{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}$

$$f(y) = c^T y = \sum_{i=1}^{n} c_i y_i$$

$$\frac{d}{dy}f(y) = \frac{d}{dy}\boldsymbol{c}^T\boldsymbol{y} = \boldsymbol{c}$$

3) $A = (a_{ij})_{n \times n}$

$$f(y) = y^T A y = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} y_i y_j$$

$$\frac{d}{dy}f(y) = \frac{d}{dy}y^T A y = 2Ay$$

Multiple Linear Regression

May 30, 2014 2:04 PM

Multiple linear regression (MLR) assumes that y is linearly related to a combination of x_i's

- $y =$ regression variate
- $x_1, \dots, x_p =$ predictive/explanatory variables
- Data $\{(y_i, x_{i1}, x_{i2}, \dots., x_{ip}), i = 1, \dots, n\}$

Model $y_i\beta_0 + \beta_1 x_i + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon_i$

Assumptions:

i) $E(\epsilon_i) = 0$

ii) $V(\epsilon_i) = \sigma^2$

iii) $\epsilon_1, \dots, \epsilon_n$ are independent of each other $\Rightarrow y_1, \dots, y_n$ are independent

iv) Stronger Distribution Assumption:

$\epsilon_1, \epsilon_2, \dots, \epsilon_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$

$\Rightarrow y_1, \dots, y_n$ are independently $N(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma^2)$

NB: If iv) is true then i), ii), and iii) are true.

Interpretation of Parameters

$$E[y_i | x_{i1}, x_{i2}, \dots, x_{ip}] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$
$$E[y_i | x_{i1} + c, x_{i2}, \dots, x_{ip}] = \beta_0 + \beta_1 (x_{i1} + c) + \cdots + \beta_p x_{ip}$$
$$\Rightarrow E[y_i | x_{i1} + c, x_{i2}, \dots, x_{ip}] - E[y_i | x_{i1}, x_{i2}, \dots, x_{ip}] = \beta_1 c$$

Let $c = 1 \Rightarrow \beta_1$ is the average change in the response when x_1 is increased by 1 unit and all other x's are held fixed.

Matrix Form

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

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times(n+1)}$$

Assumptions

i) $E(\boldsymbol{\epsilon}) = \boldsymbol{0} \Rightarrow E(\boldsymbol{y}) = X\boldsymbol{\beta}$

ii) $V(\boldsymbol{\epsilon}) = \sigma^2 I_{n\times n} \Rightarrow V(\boldsymbol{y}) = V(\boldsymbol{\epsilon}) = \sigma^2 I_{n\times n}$

iii) $\epsilon_1, \dots, \epsilon_n$ are iid random variables

iv) $\boldsymbol{\epsilon} \sim \text{MVN}(\boldsymbol{0}, \sigma^2 I)$

Parameter Estimation

$\rightarrow \beta_0, \beta_1, \dots, \beta_p$ unknown parameters

$\rightarrow \hat{\beta}_0, \hat{\beta}_1, , \dots, \hat{\beta}_p$ estimates (based on sample)

$\Rightarrow$ LSE of $\boldsymbol{\beta}$. Find $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ such that

$$S(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_p)^p$$

is minimal.

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta}) \\ &= (\boldsymbol{y}^T - \boldsymbol{\beta}^T X^T)(\boldsymbol{y} - X\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T X\boldsymbol{\beta} - \boldsymbol{\beta}^T X^T \boldsymbol{y} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta} \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T X\boldsymbol{\beta} - (\boldsymbol{\beta}^T X^T \boldsymbol{y})^T + \boldsymbol{\beta}^T X^T X\beta = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T X\boldsymbol{\beta} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta} \end{aligned}$$

Note:

$$\frac{d}{d\boldsymbol{\beta}}\boldsymbol{c}^T\boldsymbol{\beta} = c, \qquad \frac{d}{d\boldsymbol{\beta}}\boldsymbol{\beta}^T A\boldsymbol{\beta} = 2A\boldsymbol{\beta}$$

$$\frac{d}{d\boldsymbol{\beta}}S(\boldsymbol{\beta}) = 0 - 2(\boldsymbol{y}^T X)^T + 2(X^T X)\boldsymbol{\beta}$$

Set equal to zero to find $\widehat{\boldsymbol{\beta}}$

$\Rightarrow \ -2X^T\boldsymbol{y} + 2X^T X\widehat{\boldsymbol{\beta}} = 0$

$\Rightarrow X^T X\widehat{\boldsymbol{\beta}} = X^T\boldsymbol{y}$

Assume X has full column rank

$\Rightarrow rank(X) = p + 1$

Then $X^T X$ has full column rank $\Rightarrow rank(X^T X) = p + 1$

$\Rightarrow X^T X$ is invertible $\Rightarrow (X^T X)^{-1}$ exists

and so

$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1}X^T\boldsymbol{y}$

To find $\hat{\beta}_1, \dots, \hat{\beta}_p$, the the ith entry of $\widehat{\boldsymbol{\beta}}$

Properties of the LSE

1) Expected value of $\widehat{\boldsymbol{\beta}}$

   $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \Rightarrow \widehat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$.

   Proof:

   $E(\widehat{\boldsymbol{\beta}}) = E\left((X^T X)^{-1}X^T\boldsymbol{y}\right) = (X^T X)^{-1}X^T E(\boldsymbol{y}) = (X^T X)^{-1}(X^T X)\boldsymbol{\beta} = I\boldsymbol{\beta} = \boldsymbol{\beta}$

2) Variance of $\widehat{\boldsymbol{\beta}}$: $V(\widehat{\boldsymbol{\beta}}) = \sigma^2(X^T X)^{-1}$

   Proof:

   $V(\widehat{\boldsymbol{\beta}}) = V([(X^T X)^{-1}X^T]\boldsymbol{y}) = [(X^T X)^{-1}X^T]\underbrace{V(\boldsymbol{y})}_{\sigma^2 I}\underbrace{[(X^T X)^{-1}X^T]^T}_{X(X^T X)^{-1}}$

   $= \sigma^2[(X^T X)^{-1}X^T][X(X^T X)^{-1}] = \sigma^2(X^T X)^{-1}$

An estimate for σ^2

$$\sigma^2 = \frac{\sum_{i=1}^n e_i^2}{n - (p+1)} = \frac{\boldsymbol{e}^T\boldsymbol{e}}{n - p - 1}$$

Where $\boldsymbol{e} = \boldsymbol{y} - X\widehat{\boldsymbol{\beta}}$

An estimate for the variance of $\widehat{\boldsymbol{\beta}}$ is

$\hat{V}(\widehat{\boldsymbol{\beta}}) = \hat{\sigma}^2(X^T X)^{-1}$

Useful Results

Let $H = X(X^T X)^{-1}X^T$ (**hat matrix**)

Then

i) fitted values:

   $\hat{\mu} = X\hat{\beta} = X(X^T X)^{-1}X^T\boldsymbol{y} = H\boldsymbol{y}$

ii) residuals:

   $\boldsymbol{e} = \boldsymbol{y} - X\widehat{\boldsymbol{\beta}} = \boldsymbol{y} - \widehat{\boldsymbol{\mu}} = \boldsymbol{y} - H\boldsymbol{y} = (I - H)\boldsymbol{y}$

Note:

i) H is idempotent:

   $HH = X(X^T X)^{-1}X^T X(X^T X)^{-1}X^T = X(X^T X)^{-1}X^T = H$

ii) H is symmetric $\Rightarrow H^T = H$

iii) $(I - H)$ is idempotent

   $(I - H)(I - H) = I^2 - IH - HI + H^2 = I - 2H + H^2 = I - H$

Further Results and Consequences of Least Squares

i) Fitted Values:

   $\widehat{\boldsymbol{\mu}} = X\widehat{\boldsymbol{\beta}} = H\boldsymbol{y}$

   $E(\widehat{\boldsymbol{\mu}}) = E(X\widehat{\boldsymbol{\beta}}) = XE(\widehat{\boldsymbol{\beta}}) = X\boldsymbol{\beta} = \boldsymbol{\mu}$

   $V(\widehat{\boldsymbol{\mu}}) = V(H\boldsymbol{y}) = HV(\boldsymbol{y})H^T = \sigma^2 HH^T = \sigma^2 H^2 = \sigma^2 H$

ii) Residuals:

   $\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{\mu}} = (I - H)\boldsymbol{y}$

   $E(\boldsymbol{e}) = (I - H)E(\boldsymbol{y}) = X\boldsymbol{\beta} - HX\boldsymbol{\beta} = (X - X)\boldsymbol{\beta} = \boldsymbol{0}$

   $V(\boldsymbol{e}) = \left((I - H)\boldsymbol{y}\right) = (I - H)V(\boldsymbol{y})(I - H)^T = \sigma^2(I - H)^2 = \sigma^2(I - H)$

iii) $X^T\boldsymbol{e} = \mathbf{0}$ and $\hat{\boldsymbol{\mu}}\boldsymbol{e} = 0$

$X^T\boldsymbol{e} = X^T(I-H)\boldsymbol{y} = (X^T - X^TH)\boldsymbol{y} = (X^T - X^T)\boldsymbol{y} = \mathbf{0}$

$\hat{\boldsymbol{\mu}}\boldsymbol{e} = \left(X\hat{\boldsymbol{\beta}}\right)^T\boldsymbol{e} = X\left(\hat{\boldsymbol{\beta}}\boldsymbol{e}\right) = \mathbf{0}$

? I think should be $\hat{\boldsymbol{\mu}}^T\boldsymbol{e} = \left(X\hat{\boldsymbol{\beta}}\right)^T\boldsymbol{e} = \hat{\boldsymbol{\beta}}^TX^T\boldsymbol{e} = 0$

iv) Sampling distribution of $\hat{\boldsymbol{\beta}}$

Note: $\boldsymbol{y} \sim MVN(X\boldsymbol{\beta}, \sigma^2 I)$ from the 4th model assumption.

$\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{y} \sim MVN(\boldsymbol{\beta}, \sigma^2(X^TX)^{-1})\backslash$

v) $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{e}$ are independent of each other

$\boldsymbol{e} = (I-H)\boldsymbol{y} \sim MVN\left(0, \sigma^2(I-H)\right)$

Since $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{e}$ are MVN we need to show

$\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}, \boldsymbol{e}\right) = \mathbf{0}$ for independent

$\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}, \boldsymbol{e}\right) = \mathrm{Cov}\left((X^TX)^{-1}X^T\boldsymbol{y}, (I-H)\boldsymbol{y}\right) = (X^TX)^{-1}X^TV(\boldsymbol{y})(I-H)^T$

$= \sigma^2\left((X^TX)^{-1}X^T - (X^TX)^{-1}X^TH\right] = \mathbf{0}$

$\Rightarrow$ Independent

vi) $\hat{\sigma}^2 = \dfrac{1}{n-p-1}\displaystyle\sum_{i=1}^{n} e_i^2 = \dfrac{1}{n-p-1}\boldsymbol{e}^T\boldsymbol{e}$

Is unbiased for σ^2

Proof

Recall $V(\boldsymbol{e}) = \sigma^2(I-H)$

By definition, it is

$V(\boldsymbol{e}) = E\left[\left(e - E(e)\right)\left(e - E(e)\right)^T\right] = E(\boldsymbol{e}\boldsymbol{e}^T)$

Now, $E(\boldsymbol{e}^T\boldsymbol{e}) = E\left(\mathrm{tr}(\boldsymbol{e}^T\boldsymbol{e})\right) = E\left(\mathrm{tr}(\boldsymbol{e}\boldsymbol{e}^T)\right) = \mathrm{tr}\left(E(\boldsymbol{e}\boldsymbol{e}^T)\right) = \mathrm{tr}\left(V(\boldsymbol{e})\right) = \mathrm{tr}\left((I-H)\sigma^2\right) =$
$\sigma^2\mathrm{tr}(I-H) = \sigma^2\left(\mathrm{tr}(I_{n\times n}) - \mathrm{tr}(H)\right) = \sigma^2\left(n - \mathrm{tr}(H)\right) = \sigma^2\left(n - \mathrm{tr}(X(X^TX)^{-1}X^T)\right) =$
$\sigma^2\left(n - \mathrm{tr}\left((X^TX)^{-1}X^TX\right)\right) = \sigma^2\left(n - \mathrm{tr}\left(I_{(p-1)\times(p-1)}\right)\right) = \sigma^2(n-p-1)$

$\Rightarrow E(\boldsymbol{e}^T\boldsymbol{e}) = (n-p-1)\sigma^2$

$\Rightarrow E(\hat{\sigma}^2) = E\left(\dfrac{\boldsymbol{e}^T\boldsymbol{e}}{n-p-1}\right) = \sigma^2$

Note: $\hat{\sigma}^2$ is independent of $\hat{\boldsymbol{\beta}}$ since $\boldsymbol{e}$ is independent of $\hat{\boldsymbol{\beta}}$

vii) Sampling distribution of $\hat{\sigma}^2$

Result: $\dfrac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)$

Note

1) $I-H$ is symmetric
   $\Rightarrow$ (Spectral Decomposition) $(I-H) = P\Lambda P^T$ where P is orthogonal $(P^TP = PP^T = 1)$
2) $(I-H)$ is idempotent

   $$\Rightarrow \text{Eigenvalues are 0 or 1} \Rightarrow \Lambda = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}$$

   #1's is $\mathrm{tr}(I-H) = n-p-1$

Proof

$V(P^T\boldsymbol{e}) = P^T(I-H)\sigma^2P = P^TP\Lambda\sigma^2P^TP = \sigma^2\Lambda$

Let $\boldsymbol{u} = P^T\boldsymbol{e} \sim MVN(0, \sigma^2\Lambda)$

$\Rightarrow u_1, \dots, u_{n-p-1} \overset{\text{iid}}{\sim} N(0, \sigma^2),\ u_{n-p}, \dots, u_n = 0$

$\Rightarrow \displaystyle\sum_{i=1}^{n}\left(\frac{u_i}{\sigma}\right)^2 = \sum_{i=1}^{n-p-1}\left(\frac{u_i}{\sigma}\right)^2 = \chi^2(n-p-1)$

$\dfrac{(n-p-1)\hat{\sigma}^2}{\sigma^2} = \dfrac{\boldsymbol{e}^T\boldsymbol{e}}{\sigma^2} = \dfrac{(P^T\boldsymbol{u})^T(P\boldsymbol{u})}{\sigma^2} = \dfrac{1}{\sigma^2}\boldsymbol{u}^TP^TP\boldsymbol{u} = \dfrac{\boldsymbol{u}^T\boldsymbol{u}}{\sigma^2} = \dfrac{\sum_{i=1}^{n} u_i^2}{\sigma^2}$

$= \displaystyle\sum_{i=1}^{n}\left(\frac{u_i}{\sigma}\right)^2 \sim \chi^2(n-p-1)$

viii) Gauss-Markov Theorem

Recall $\widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{y}$. This is the best **linear unbiased** estimator of $\boldsymbol{\beta}$.

In other words, of all linear unbiased estimators of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}$ has the smallest variance.

Proof:

Consider another linear estimator $\widehat{\boldsymbol{\theta}} = M\boldsymbol{y}$, where $M = (X^TX)^{-1}X^T + A$

Note: $E(\widehat{\boldsymbol{\theta}}) = E(M\boldsymbol{y}) = (X^TX)^{-1}X^TX\boldsymbol{\beta} + AX\boldsymbol{\beta} = \boldsymbol{\beta} + AX\boldsymbol{\beta}$

So this is unbiased iff $AX = \mathbf{0}$

Now, $V(\widehat{\boldsymbol{\theta}}) = MV(\boldsymbol{y})M^T = \sigma^2 MM^T = \sigma^2\big((X^TX)^{-1}X^T + A\big)(X(X^TX)^{-1} + A^T)$

$= \sigma^2\left((X^TX)^{-1} + (X^TX)^{-1}\underbrace{X^TA^T}_{0} + \underbrace{AX}_{0}(X^TX)^{-1} + AA^T\right) = \sigma^2(X^TX)^{-1} + \sigma^2AA^T$

$= V(\widehat{\boldsymbol{\beta}}) + \sigma^2AA^T$

Consider a linear predictor $\boldsymbol{x}^T\widehat{\boldsymbol{\theta}}$

Now, $V(\boldsymbol{x}^T\widehat{\boldsymbol{\theta}}) = \boldsymbol{x}^TV(\widehat{\boldsymbol{\theta}})\boldsymbol{x} = \boldsymbol{x}^TV(\boldsymbol{\beta}^T)\boldsymbol{x} + \boldsymbol{x}^TAA^T\boldsymbol{x} = V(\boldsymbol{x}^T\widehat{\boldsymbol{\beta}}) + \underbrace{\sigma^2(A^T\boldsymbol{x})^T(A^T\boldsymbol{x})}_{\geq 0}$

$\Rightarrow V(\boldsymbol{x}^T\widehat{\boldsymbol{\theta}}) \geq V(\boldsymbol{x}^T\widehat{\boldsymbol{\beta}})$

$\Rightarrow \widehat{\boldsymbol{\beta}}$ produces the smallest variance

Notes on Handout

June 11, 2014 1:45 PM

$\hat{\beta}_i \sim N(\beta_i, \sigma^2 v_{ii})$

$\dfrac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{v_{ii}}} \sim N(0, 1)$

$t(k) = \dfrac{N(0,1)}{\sqrt{\dfrac{\chi^2(k)}{k}}}$

$$\sqrt{\frac{\left(\frac{(n-p-1)\hat{\sigma}^2}{\sigma}\right)}{n-p-1}} \sim \sqrt{\frac{\chi^2(n-p-1)}{n-p-1}}$$

$$\Rightarrow \frac{\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{v_{ii}}}}{\sqrt{\frac{\left(\frac{(n-p-1)\hat{\sigma}^2}{\sigma}\right)}{n-p-1}}} = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{v_{ii}}} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-p-1)}{n-p-1}}} = t(n-p-1)$$

$V(y_* - \hat{y}_*) = V(y_*) + V(\hat{y}_*) = \sigma^2 + V\left(\boldsymbol{x}_*^T\widehat{\boldsymbol{\beta}}\right) = \sigma^2 + \boldsymbol{x}_*^T V\left(\widehat{\boldsymbol{\beta}}\right)\boldsymbol{x}_* = \sigma^2(1 + \boldsymbol{x}_*^T (X^T X)^{-1}\boldsymbol{x}_*)$

Aside:

$\hat{\mu}_* = \hat{\beta}_0 + \hat{\beta}_1 x_* = \begin{bmatrix} 1 & x_* \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$

In matrix form,

$V\left(\widehat{\boldsymbol{\beta}}\right) = \begin{bmatrix} V\left(\hat{\beta}_0\right) & \mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) \\ \mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) & V\left(\hat{\beta}_1\right) \end{bmatrix}$

$\Rightarrow V(\hat{\mu}) = \boldsymbol{x}_*^T V\left(\widehat{\boldsymbol{\beta}}\right)\boldsymbol{x}_* = \begin{bmatrix} 1 & x_* \end{bmatrix} \begin{bmatrix} v_{00} & v_{01} \\ v_{10} & v_{11} \end{bmatrix} \begin{bmatrix} 1 \\ x_* \end{bmatrix} = v_{00} + v_{10}x_* + v_{01}x_* + v_{11}x_*^2$

$= V\left(\hat{\beta}_0\right) + 2\mathrm{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right)x_* + x_*^2 V\left(\hat{\beta}_1\right)$

Note

ANOVA has more power than doing 'p' individual t-tests (of $\beta_i = 0$ vs $\beta_i \neq 0$)

Multicollinearity

June 13, 2014 1:02 PM

Multicollinearity

We assume the columns of X were linearly independent $\Rightarrow (X^TX)^{-1}$ would exist $\Rightarrow \widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{y}$ where

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}_{n\times(p+1)} = [\boldsymbol{x}_0 \quad \boldsymbol{x}_1 \quad \cdots \quad \boldsymbol{x}_p]$$

Two Cases

1) Exact linear dependence.
   Suppose at least one of the $\boldsymbol{x}_j$'s is a linear combination of the other $\boldsymbol{x}_j$'s.
   $\Rightarrow |X^TX| = 0$ or ranke $< p+1$
   $\Rightarrow \widehat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{y}$ does not exist

   e.g. Suppose $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ and $x_1 = 5 + 3x_2$ (perfect linear dependence)
   $\Rightarrow y = \beta_0 + \beta_1(5+3x_2) + \beta_2 x_2 + \epsilon = \underbrace{(\beta_0 + 5\beta_1)}_{\beta_0^*} + \underbrace{(3\beta_1 + \beta_2)}_{\beta_1^*} x_2 + \epsilon$
   Remedy: Drop x_1 if x_2 is included since x_1 is redundant

2) Non-linear Dependence
   Suppose there exists a near (but not perfect) linear relationship between one $\boldsymbol{x}_j$ and the other $\boldsymbol{x}_j$'s.
   Then $(X^TX)^{-1}$ still exists. However, $|X^TX| \approx 0 \Rightarrow \frac{1}{|X^TX|}$ will bew very large.

   Consequences
   i) Numerically / Computationally unstable
   ii) Incorrect signs of $\hat{\beta}_j$'s (doesn't agree with what's plausible)
   iii) $\hat{\beta}_j$'s will be sensitive to small changes in the data
   iv) Implausible values/magnitudes for $\hat{\beta}_j$'s
   v) Since $V(\widehat{\boldsymbol{\beta}}) = \sigma^2(X^TX)^{-1}$, variance estimates tend to be inflated.
      a) Important predictors may show up insignificant.
      b) Confidence Intervals are very wide (which makes it useless for interpretation of β_j's

Remedies for Multicollinearity

i) Check for pairwise correlation of predictors.
   In R, cov(X) ← Variance-Covariance Matrix
   cor(X) ← Correlation Matrix
   plot(X) ← Pairwise scatter plots

ii) A better measure:
   Variance Inflation Factors (VIF)
   a) Treat $\boldsymbol{x}_j$ as the response
   b) Regress $\boldsymbol{x}_j$ on the other predictors
      $\Rightarrow$ Model $x_j = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{j-1}x_{j-1} + \alpha_{j+1}x_{j+1} + \cdots + \alpha_p x_p + \epsilon^*$
   c) Denote the model R-squared as R_j^2
      Compute $\mathrm{VIF}_j = \frac{1}{1-R_j^2}$, called the variance inflation factor
   d) Calculate VIF_j for $j = 1, \dots, p$

   <u>Rule of Thumb</u>
   $\max_{1\le j\le p}\{\mathrm{VIF}_j\} \ge 10$ evidence of multicollinearity

Note: It can be shown that $\hat{V}(\hat{\beta}_j)$ is proportional to $\frac{1}{1-R_j^2} = \text{VIF}_j$

i) $R_j^2 = 0 \Rightarrow$ no inflation $\Rightarrow \text{VIF}_j = 1 \Rightarrow x_j$ is linearly independent of the other predictors

ii) $R_j^2 > 0 \Rightarrow \text{VIF} > 1 \Rightarrow$ inflation in the variance estimate of $\hat{\beta}_j$

Dummy Variable Regression

June 20, 2014 1:09 PM

Idea: Extend the linear regression model to include categorical variables (factors) via indicator variables.

Example: Factors + Continuous Variables

$$\text{Data:}\begin{cases}\text{Fuel Consumption} \leftarrow \text{Response } (y) \\ \text{Engine Size} \leftarrow \text{Predictor } (x_1) \\ \text{Make} \leftarrow \text{Predictor, Categorical } (x_2)\end{cases}$$

Suppose $n = 10$ cases

Make #1: y_1, y_2, y_3, y_4 (BMW)

Make #2: $y_5, \dots, y_{10}$ (Audi)

Let $x_{i2} = \begin{cases} 0 & i = 1,2,3,4 \\ 1 & i = 5, \dots, 10 \end{cases}$

Case #1

Assume make has an effect on y, and it is the same regardless of engine size (the effect of enginge size of y doesn't depend on make)

$$E(y_i) = \begin{cases} \beta_0 + \beta_1 x_{i1} & , i = 1, 2, 3, 4 \\ \beta_0 + \beta_2 + \beta_1 x_{i1} & , i = 5, \dots, 10 \end{cases}$$

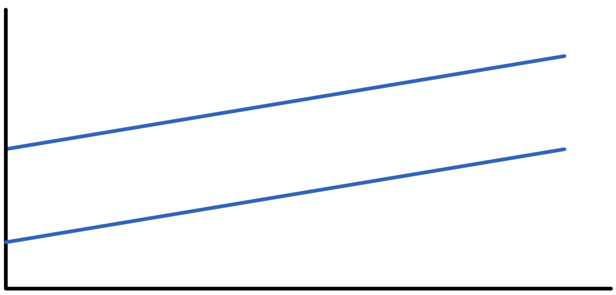
$\Rightarrow E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$

$\Rightarrow$ Model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, i = 1, \dots, 10$

Matrix Form: $\underline{y} + X\underline{\beta} + \underline{\epsilon}$

We can test for example:

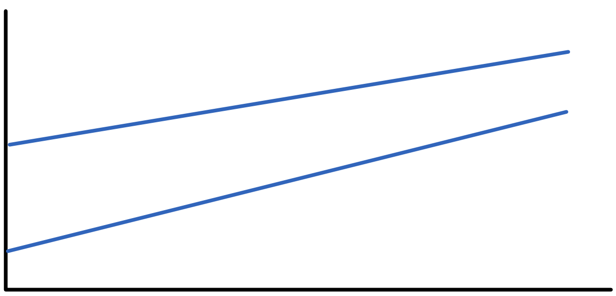
$H_0: \beta_2 = 0$ (does make have an affect on y?)



Case #2

Assume make has an effect on y, but this effect changes with engine size (called an **interaction**)

$$E(y_2) = \begin{cases} \beta_0 + \beta_1 x_{i1} & , i = 1,2,3,4 \\ \beta_0 + (\beta_1 + \beta_3) x_{i1} + \beta_2 x_{i2} & , i = 5, \dots, 10 \end{cases}$$



$E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{12} + \beta_3 x_{i1} x_{i2}$

2-way interaction
Matrix Form: $\underline{y} + X\underline{\beta} + \underline{\epsilon}$

$$\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \quad \underline{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{41} & 0 & 0 \\ 1 & x_{51} & 1 & x_{51} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{10,1} & 1 & x_{10,1} \end{bmatrix}$$

Can test for example $H_0\!: \beta_3 = 0$ (Does the effect of enginge size on y depend on make?)

Example: Comparing Several Groups

Data: $\begin{cases} \text{Diet} \leftarrow \text{predictor, categorical } (x) \\ \quad \text{Weight} \leftarrow \text{response } (y) \end{cases}$

Suppose $n = 10$ persons.
Diet #1: y_1, y_2, y_3
Diet #2: y_4, y_5, y_6
Diet #3: y_7, y_8, y_9, y_{10}
Question: Does diet affect weight gain?

Let $E(y_i) = \begin{cases} \mu_1 & i = 1,2,3 \\ \mu_2 & i = 4,5,6 \\ \mu_3 & i = 7,8,9,10 \end{cases}$

Can test $H_0\!: \mu_1 = \mu_2 = \mu_3$ (Does average weight gained depend on diet?)

Model:

Formulation #1

Let $x_{i1} = \begin{cases} 1 & i = 1,2,3 \\ 0 & \text{otherwise} \end{cases}$

$x_{i2} = \begin{cases} 1 & i = 4,5,6 \\ 0 & \text{otherwise} \end{cases}$

$x_{i3} = \begin{cases} 1 & i = 7,8,9,10 \\ 0 & \text{otherwise} \end{cases}$

$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$
(No intercept)

Matrix form: $\underline{y} + X\underline{\beta} + \underline{\epsilon}$

$$\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \quad \underline{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Aside: Had we included the intercept, we would have had a column of 1s in X. However, the sum of the current columns of X gives this columns of 1's $\Rightarrow$ Redundant information.

Note: $\mu_1 = \beta_1, \mu_2 = \beta_2, \mu_3 = \beta_3$
$\Rightarrow$ Testing $H_0 = \mu_1 = \mu_2 = \mu_3$ is equal to testing $H_0\!: \beta_1 = \beta_2 = \beta_3$

Formulation #2 (More Common)

Let $x_1 = \begin{cases} 1 & i = 1,2,3 \\ 0 & \text{otherwise} \end{cases}, \ x_2 = \begin{cases} 1 & i = 4,5,6 \\ 0 & \text{otherwise} \end{cases}$

Model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \ i = 1, 2, \dots, 10$
Matrix form: $\underline{y} = X\underline{\beta} + \underline{\epsilon}$

$$\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \qquad \underline{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

$\mu_1 = \beta_0 + \beta_1, \qquad \mu_2 = \beta_0 + \beta_2, \qquad \mu_3 = \beta_0$

Note: Diet x is the base case in this formulation. β_0 is the average weight gained under diet 3

$\beta_1 = \mu_1 - \mu_3$ is the excess average weight gained under diet 1 relative to diet 3.

$\beta_2 = \mu_2 - \mu_3$ is the excess average weight gained under diet 2 relative to diet 3

Testing $h_0: \mu_1 = \mu_2 = \mu_3$ is equivalent to testing $H_0: \beta_1 = \beta_2 = 0$

(ANOVA F-test provided in the R summary output)

Residual Analysis

June 25, 2014 1:16 PM

True error $\epsilon_i = y_i - \mu_i,\ \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$
Model assumption: $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$
Estimated Errors / Residuals $e_i = y_i - \hat{\mu}_i$
Properties:

1) $\sum_{i=1}^{n} e_i = 0 \Rightarrow \bar{e} = 0$
2) $\sum_{i=1}^{n} e_i x_{ik} = 0, \text{for } k = 1, 2, \dots, p$
   $\Rightarrow \text{Cor}(\underset{\sim}{e}, \underset{\sim}{x}_k) = 0$
3) $\sum_{i=1}^{n} e_i \hat{\mu}_i = 0$
   $\Rightarrow \text{Cor}(\underset{\sim}{e}, \underset{\sim}{\hat{\mu}}) = 0$
4) $\underset{\sim}{e} = \underset{\sim}{y} - \underset{\sim}{\hat{\mu}} = \underset{\sim}{y} - H\underset{\sim}{y} = (I - H)\underset{\sim}{y}$ where $H = X(X^T X)^{-1} X^T$
   $\Rightarrow \underset{\sim}{e} \sim MVN(0, (I - H)\sigma^2) \Rightarrow e_i \sim N(0, (1 - h_{ii})\sigma^2)$ and $\text{Cov}(e_i, e_j) = -h_{ij}\sigma^2$
5) **Studentized Residuals**
   $d_i = \dfrac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}, \qquad i = 1,2, \dots, n$
   Note: d_i's do note follow a t-distribution since the numerator and denominator are not independent of each other.
   $d_1, \dots, d_n \overset{\text{iid}}{\approx} N(0, 1)$ (approximate)
   It is useful to plot:
   i) the d_i's and if the assumptions are not violated, one should see random scatter
   ii) e_i's vs each predictor (should also see random scatter)
   iii) e_i's vs $\hat{\mu}_i$'s (should also see random scatter)

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon$
→ Estimate Residuals → e_i's
→ Consider another predictor x_3
To determine how useful the addition of x_3 would be to the model, plot e vs. x_3. If linear, consider adding x_3 to the model. (Conditional effect of x_3 on y given x_1 and x_2).

Added Variable Plots

Instead of plotting $\underset{\sim}{e}$ vs $\underset{\sim}{x}_3$, plot $\underset{\sim}{e}$ vs $\underset{\sim}{e}^*$ where e_i^* are the residuals of the model $x_3 = \beta_0^* + \beta_1^* x_1 + \beta_2^* x_2 + \epsilon^*$

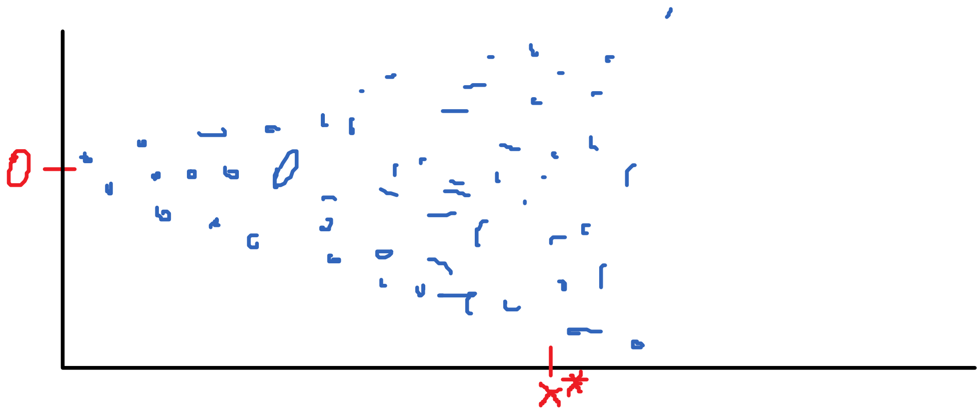
In R:
```
library(car)
avPlot(mod)
```

Variance Stabilizing Transformations

June 27, 2014 1:52 PM

Variance Stabilizing Transformation

Dealing with non-constant variance.



Model: $y_i = \beta_0 +' beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$

Suppose we want to predict the value of y at $x = x^*$. Would the width of the prediction interval be overestimated/underestimated based on the model above? Underestimated

$V(y_i) = \sigma^2 [h(\mu_i)]^2$

variance of y_i is a function of μ_i. Instead of using y, use some function of y, say $g(y)$

$g(y) \approx g(\mu) + g'(\mu)(y - \mu)$

$V\big(g(y)\big) = [g'(\mu)]^2 \sigma^2 [h(\mu)]^2$

Want $g'(\mu)h(\mu) = c$, a constnat $\Rightarrow g'(\mu) = \frac{c}{h(\mu)}$

$$\Rightarrow g(\mu) = \int \frac{c}{h(\mu)}$$

Examples

i) $h(\mu) = \mu \Rightarrow g(\mu) = \int \frac{c}{\mu}\, d\mu = c \ln(\mu)$

$\rightarrow$ Try $g(y) = \ln(y)$

ii) $h(\mu) = \sqrt{\mu} \Rightarrow g(\mu) = \int \frac{c}{\sqrt{\mu}}\, d\mu = 2c\sqrt{\mu}$

$\rightarrow$ Try $g(y) = \sqrt{y}$

Box Cox Transformations

Model $y_i = \mu_i + \epsilon_i$

Box Cox Transformation is a family of power transformations

$$g(y_i) = \frac{y_i^\lambda - 1}{\lambda} \text{ for some } \lambda \in \mathbb{R}$$

Choose λ such that $V\big(g(y_i)\big)$ is constants.

Notes

i) $\lambda = 1 \Rightarrow$ No transformation

ii) $\lambda = \frac{1}{2} \Rightarrow$ Square root transformation

iii) $\lambda = 0 \Rightarrow \lim_{\lambda \to 0} g(y_i) = \ln(y_i)$ By L'Hôpital's rule

Estimate λ by maximum likelihood

(MLE) Assume $g(y_i) \sim N\big(\mu_{i,\lambda}, \sigma_\lambda^2\big)$

The log-likelihood is

$$l(\lambda) = -\frac{1}{2\sigma_\lambda^2} \sum_{i=1}^{n} \big(g(y_i) - \mu_{i,\lambda}\big)^2 = -\frac{n}{2} \ln \sigma_\lambda^2 + (\lambda - 1) \sum_{i=1}^{n} \ln y_i$$

Maximizing $l(\lambda)$ with respect to $\lambda, \underline{\beta}_\lambda, \sigma_\lambda$

→ Not easy to do in practice

→ Use the profile likelihood instead

i) Consider a sequence of λ's, e.g. $\{-2, -1.9, \ldots, 1.9, 2\}$

ii) For each λ using $g(y_i)$ as the response, find the LSE of $\underline{\beta}_\lambda$ and σ_λ, Also compute $l(\lambda)$

iii) Select the value of λ which gies the largest $l(\lambda)$; denote by $\hat{\lambda}$

In R:

```
library(MASS)
boxcox(model)
```

Weighted Least Squares (WLS)

July 2, 2014 1:12 PM

Consider the model $y_i = \mu_i + \epsilon_i$
where $\epsilon_i \sim N(0, \sigma^2 v_i^2)$, non-constant variance.
and $\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$
Since $\text{Var}(\epsilon_i) = \sigma^2 v_i^2 \Rightarrow \text{Var}\left(\frac{\epsilon_i}{v_i}\right) = \sigma^2$
Re-write the model as
$$\frac{y_i}{v_i} = \frac{\beta_0}{v_i} + \beta_1\left(\frac{x_{i1}}{v_i}\right) + \cdots + \beta_p\left(\frac{x_{ip}}{v_i}\right) + \frac{\epsilon_i}{v_i}$$
Let $x_{i0}^w = \frac{1}{v_i},\ x_{ik}^w = \frac{x_{ik}}{v_i},\ k = 1, 2, \dots, p$
$$\Rightarrow y_i^w = \frac{y_i}{v_i} = \beta_0 x_{i0}^w + \cdots + \beta_p x_{ip}^w + \epsilon_i^w$$
$$\epsilon_i^w = \frac{\epsilon_i}{v_i} \sim N(0, \sigma^2)$$

WLS Estimates

$$S(\beta) = \sum_{i=1}^n \left(y_i^w - \beta_0 x_{i0}^w - \cdots - \beta_p x_{ip}^w\right)^2 = \sum_{i=1}^n \frac{1}{v_i^2}\left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}\right)^2$$
In matrix form:
$$\text{Let } W = \begin{bmatrix} w_1 & & & \\ & w_2 & & \\ & & \ddots & \\ & & & w_p \end{bmatrix}, \text{ where } w_i = \frac{1}{v_i^2}$$
$$\Rightarrow S(\underline{\beta}) = (\underline{y} - X\underline{\beta})^T W (\underline{y} - X\underline{\beta})$$
Results:

i) WLSE of $\underline{\beta}$ is $\hat{\underline{\beta}}_w = (X^T W X)^{-1} X^T W \underline{y}$
ii) $E(\hat{\underline{\beta}}_w) = \underline{\beta}$
iii) $V(\hat{\underline{\beta}}_w) = \sigma^2 (X^T W X)^{-1}$
$V(\underline{y}) = \sigma^2 W^{-1}$
iv) $\hat{\sigma}^2 = \dfrac{1}{n-p-1}\sum_{i=1}^n w_i e_i^2$

Residual Plots
Plot $e_i^w = \frac{e_i}{v_i}$ vs x_{ik} or $\hat{\mu}_i$ to see if the variance has stabilized.

Aside

$$S(\underline{\beta}) = \sum_{i=1}^n \left(\frac{y_i}{v_i} - \frac{\mu_i}{v_i}\right)^2$$
WLSE on $\hat{\beta}_1^w, \dots, \hat{\beta}_p^w$
$$\Rightarrow \hat{\sigma}^2 = \frac{S(\hat{\underline{\beta}}^w)}{n-p-1} = \frac{\sum\left(\frac{e_i}{v_i}\right)^2}{n-p-1} = \frac{\text{Sum of squared weighted residuals}}{n-p-1}$$

How to estimate v_i?

- Difficult in practice
- Construct a plot of e_i vs x_{ik}
- Try $v_i = x_{ik}^\gamma$ for some $k = 1, 2, \dots, p$ and some $\gamma \in \mathbb{R}$ (by trial and error)
- Reconstruct plots of $\frac{e_i}{v_i}$ vs $\hat{\mu}_i$ or x_{ik} until you have constant variance.

Outliers (Extraneous Observations)

July 2, 2014 2:02 PM

Cases:

i) Outliers due to misrecording
   Correct it or delete it.
      Can replace with the average of the other observations if it is a predictor.

ii) Outlier is a valid observation.

- Maybe a predictor is missing from the model which can explain this observation.
  - Can fit the model
    - With the outlier
    - Without the outlier
  - Keep the observation if conclusions (on fitted values, coefficients) do not change significantly.
  - If conclusions are changed greatly, we say the outlier is influential.
    - Remove it (or possibly correct it if you can)
    - Removal may lead to a redefinition of the population.

In R:
```
library(outliers)
```

Note

$$S(\hat{\beta}) = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 = \sum_{i=1}^{n} e_i^2$$

is minimized in the LS algorithm.
The LS algorithm tends to fit more towards outliers (especially with the squaring of the e_i)
An alternative algorithm: Minimize

$$\sum_{i=1}^{n}|e_i|$$

The effect of the outlier on the fitted line will not be as significant under this algorithm compared to the least squares algorithm.

Check for Outliers

1) Construct a residual plot.
   Recall studentized residuals:
   $$d_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$
   Construct a plot of $d_1, \ldots, d_n$. Note: $d_1, \ldots, d_n \approx N(0,1)$
   d_i's should be within (-3, 3)
   Also, 95% of d_i's should be within (-2, 2)
   In R, the plot code is as follows:
   ```
   library(MASS)
   plot(studres(model))
   ```
2) A formal test
   If $e_i \sim N(0, (1-h_{ii})\sigma^2)$ then $\frac{e_i}{\sigma\sqrt{1-h_{ii}}} \sim N(0,1)$
   Note e_i is not independent of $\hat{\sigma} = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2$
   $$\Rightarrow \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \nsim t(n-p-1)$$
   To check that the i-th data point is an outlier
   i) Delete the i-th observation and refit the model using n-1 observations.
      $e_1(-i), e_2(-i), \ldots, e_{i-1}(-1), e_{i+1}(-1), \ldots, e_n(-1)$ are the new residuals
   ii) $$\hat{\sigma}^2(-i) = \frac{1}{(n-1)-p-1}\sum_{\substack{j=1 \\ j\neq i}}^{n} e_j^2(-i)$$
      Can show that
      $$\frac{(n-p-2)\hat{\sigma}^2(-i)}{\sigma^2} \sim \chi^2(n-p-2)$$
   iii) Test statistic:
      $$t_i = \frac{e_i}{\hat{\sigma}(-i)\sqrt{1-h_{ii}}} \sim t(n-p-2)$$
      Rule: If $|t_i| > t_{\frac{\alpha}{2}}(n-p-2)$, then the i^{th} observation is an outlier.
      This approach is not very conservative.

      Another approach (Bonferroni Correction)
      Rule: If $|t_i| > t_{\frac{\alpha}{2n}}(n-p-2)$ then the ith observation is an outlier.
      Consider a single test $(n=1)$ and a significance level $\alpha = 0.05$
      - Without correction, $P(\text{false positive}) = P(\text{reject } H_0 | H_0 \text{ is true}) = 0.05$
      - With correction $P(\text{false positive}) = \frac{\alpha}{1} = \alpha = 0.05$

      There is no difference!
      Consider now $n > 1$ tests and assume significance level of $\alpha = 0.05$
      - Without correction, $P(\geq 1 \text{ false positive}) = 1 - P(0 \text{ false positives}) = 1-(1-\alpha)^n \to 1$ as $n \to \infty$
      - With correction, $P(\geq 1 \text{ false positive}) = 1 - P(0 \text{ false positives} = 1 - \left(1-\frac{x}{n}\right)^n \approx 1 - \left(1 - n\frac{\alpha}{n}\right) = \alpha$
        As $n \to \infty$, $P(\geq 1 \text{ false positive}) \to 1 - e^{-\alpha} \approx \alpha$

      In R
      ```
      library(outliers)
      ```

```
outlierTest(model)
```

Aside
1 - (1-x)^n = a
(1-x)^n = 1-a
(1-x) = (1-a)^(1/n)
x = 1 - (1-a)^(1/n) ≈ a/n

Influential Cases

Main problem: is the outlier influential? Does it affect our conclusions significantly when removed from the dataset?

Leverage

- Used to determine if an observation is an outlier in the x direction.
- A high leverage point is one which has a very large or small x-value relative to the other data points. (Far apart from the bulk of the data in the x direction).

Cases

i) Far in the x-direction ⇒ high leverage, far in the y direction ⇒ high influence
Consequence: Model coefficients and predictions are affected significantly.
ii) Low leverage (point lies around the average x) . Low influence since it will affect the model coefficients and predictions slightly.
iii) Point has high leverage but not an outlier in the y-direction. Not influential since changes in predictions and model coefficients are negligible.

Overall, high leverage is a prerequisite for making a case a high influence point, but not all high leverage points are highly influential.

Measure of Leverage

The hat matrix ? and leverage

- Recall $H = X(X^TX)^{-1}X^T = [h_{ij}]_{n\times n}$
- Also, $e_i \sim N(0, \sigma^2(1-h_{ii}))$

If $h_{ii} \approx 1$, then $e_i \approx 0 \Rightarrow y_i - \hat{\mu}_i \approx 0 \Rightarrow y_i \approx \hat{\mu}_i$
So the fitted line tends towards the ith observation.
In this case, the ith observation has high leverage ("pull") h_{ii} is called the leverage

Properties of Leverage

i) $H = X(X^TX)^{-1}X^T \leftarrow$ a function of the x's only (not y)
$\Rightarrow h_{ii}$ is a function of the x's
ii) $V(e_i) = \sigma^2(1-h_{ii}) \geq 0 \Rightarrow h_{ii} \leq 1$
We can further show that $\frac{1}{n} \leq h_{ii} \leq 1$
iii) $tr(H) = tr(X(X^TX)^{-1}X^T) = tr((X^TX)^{-1}X^TX) = tr(I_{(p+1)\times(p+1)}) = p+1$
$$\Rightarrow \sum_{i=1}^{n} h_{ii} = p+1$$
$\Rightarrow$ Average leverage $\bar{h} = \frac{p+1}{n}$
Rule of thumb: If $h_{ii} > 2\bar{h}$ then the i^{th} point is considered to be a high leverage point.
iv) h_{ii} is smallest when x_i is near $\bar{x}$.
SLR setting: Can show that $h_{ii} = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{S_{xx}}$ which is minimized at $x_i = \bar{x}$

Identifying highly influential cases

Recall $\underline{\hat{\beta}} = (X^TX)^{-1}X^T\underline{y} \sim MVN(\underline{\beta}, \sigma^2(X^TX)^{-1})$

$$\Rightarrow (X^TX)^{\frac{1}{2}}(\underline{\hat{\beta}} - \underline{\beta}) \sim MVN(0, \sigma^2 I)$$
$$\Rightarrow \frac{1}{\sigma}(X^TX)^{\frac{1}{2}}(\underline{\hat{\beta}} - \underline{\beta}) \sim MVN(0, I)$$
$$W = \left[\frac{1}{\sigma}(X^TX)^{\frac{1}{2}}(\underline{\hat{\beta}} - \underline{\beta})\right]^T \left[\frac{1}{\sigma}(X^TX)^{\frac{1}{2}}(\underline{\hat{\beta}} - \underline{\beta})\right] \sim \chi^2(p+1)$$
$$\Rightarrow W = \frac{(\underline{\hat{\beta}} - \underline{\beta})^T(X^TX)(\underline{\hat{\beta}} - \underline{\beta})}{\sigma^2} \sim \chi^2(p+1)$$
Also, $\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)$
$$\Rightarrow \frac{(\underline{\hat{\beta}} - \underline{\beta})^T(X^TX)(\underline{\hat{\beta}} - \underline{\beta})}{(p+1)\sigma^2} \sim F(p+1, n-p-1)$$

Cook's Distance

A measure of influence
$$D_i = \frac{(\underline{\hat{\beta}} - \underline{\hat{\beta}}_{-i})^T(X^TX)(\underline{\hat{\beta}} - \underline{\hat{\beta}}_i)}{(p+1)\hat{\sigma}^2}$$
where $\underline{\hat{\beta}}_{\mathrm{i}}$ is the vector of models coefficients when the ith data points is excluded.

Note

i) D_i does not have a F distribution but it may be compared to an $F(p+1, n-p-1)$ distirbution.
ii) Rule of thumb: If $D_i > 1$ (sometimes 0.5) then the i^{th} data point is influential.
iii) We can also write
$$D_i = \frac{(\underline{\hat{\mu}} - \underline{\hat{\mu}}_{-1})^T(\underline{\hat{\mu}} - \underline{\hat{\mu}}_{-1})}{(p+1)\hat{\sigma}^2}$$
Cook's distance is a measure of distance/influence on
a) model coefficients
b) fitted values
iv) Can also write
$$D_i = \frac{h_{ii}}{1-h_{ii}} \cdot \frac{d_i^2}{p+1}$$

d_i is the studentized residual

$$d_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

Model Selection (Ch. 7)

July 11, 2014 1:29 PM

Average prediction variance:

- Suppose we have p predictions
- Prediction model variance $= V(y_{\text{new}} - \hat{y}_{\text{new}}) = V(y_{\text{new}} - \hat{\mu}_{\text{new}}) = V(y_{\text{new}}) + V(\hat{\mu}_{\text{new}}) = \sigma^2 + V(\hat{\mu}_{\text{new}})$
- Do prediction for n points:
  Average prediction error variance
  $$= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n} V(\hat{\mu}_i) = \sigma^2 + \frac{1}{n}\text{tr}\left(V(\underline{\hat{\mu}})\right)$$
  $$V(\underline{\hat{\mu}}) = \sigma^2 H$$
  $$= \sigma^2 + \frac{1}{n}\text{tr}(\sigma^2 H) = \sigma^2 + \frac{\sigma^2}{n}(p+1) = \sigma^2\left(1 + \frac{p+1}{n}\right)$$
  If you start adding unnecessary predictors, $\hat{\sigma}^2$ may grow a bit larger and also 'p' increases.
  $\Rightarrow$ Average prediction variance increases with p.

Model Selection Handout

i) $R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$

$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

constant regardless of the models

To show that R^2 increases with p, need to show that SSE decrases with p

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$

e.g. Consider two models;

Model 1: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$

Model 2: $y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$

Want to show that $\text{SSE}(M_1) \le \text{SSE}(M_2)$

$$\text{SSE}(M_1) = \min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$

$$\text{SSE}(M_2) = \min_{\beta_0,\beta_1} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1})^2$$

Note: $\text{SSE}(M_2) = \text{SSE}(M_1 \text{ when } \beta_2 = 0)$

Remove the constraint on β_2 agove

$\Rightarrow \text{SSE}(M_2) \le \text{SSE}(M_1)$

General F-Test

July 18, 2014 1:44 PM

$y = \beta_1 \text{Trt} + \beta_2 \text{Ctrl} + \Sigma$, Test $\beta_1 = \beta_2 = \beta^*$
$y = \beta^* (\text{Trt} + \text{Ctrl}) + \Sigma$

$$\text{Trt} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \text{Ctrl} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Additional Sub Sequence Principle

- Useful for comparing nested models. Nested means one model is a special case of another.
- e.g. Model 1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$
  Model 2: $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \epsilon$

Is the reduced model adequate?

General F-Test

For testing linear set of hypotheses

$H_0: A\underline{\beta} = \underline{c}$, A is $r \times (p+1)$

Example

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$

1) $H_0: \beta_2 = \beta_3 = 0$
   $H_0: A\underline{\beta} = \underline{c}$

$$\underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \underline{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

   OR

$$A = \begin{bmatrix} 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

2) $H_0: \beta_2 = 0, \beta_1 = \beta_3$

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}, \quad \underline{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

The test statistic under $\mathrm{H}_0: \mathrm{A}\underline{\beta} = \underline{\mathrm{c}}$

$\Rightarrow$ Let $\underline{\theta} = A\underline{\beta} \Rightarrow H_0: \underline{\theta} = \underline{c}$

Estimate $\hat{\underline{\theta}} = \mathrm{A}\hat{\underline{\beta}}$
Note:
$\hat{\beta} \sim \text{MVN}(\underline{\beta}, \sigma^T (X^T X)^T)$
$\hat{\underline{\theta}}$ is a linear form $\Rightarrow \hat{\underline{\theta}} \sim MVN$
$F(\hat{\theta}) = F(A\hat{\beta}) = A\underline{B} = \underline{\theta}$
$V(\hat{\theta}) = V(A\hat{\beta}) = AV(\hat{\underline{\beta}})A^T = A(X^T X)^{-1} A^T$
$\Rightarrow \hat{\underline{\theta}} = MVN(\underline{\theta}, A(X^T X)^{-1} A^T \sigma^2)$

Want
$P(\hat{\underline{\theta}} - \underline{\theta}) \sim MVN(0, \sigma^2 I)$
$V\left(P(\hat{\underline{\theta}} - \underline{\theta})\right) = \sigma^2 I$
$PV(\hat{\theta})P^T = I\sigma^2$

$PA(X^TX)^{-1}A^TP^T\sigma^2 = \sigma^2 I$

$P[A(X^TX)^{-1}A^T]^{\frac{1}{2}}[A(X^TX)^{-1}A^T]^{\frac{1}{2}}P^T = I$

$Let\ P = (A(X^TX)^{-1}A^T)^{-\frac{1}{2}}$

Back to test statistic

$P(\hat{\underline{\theta}} - \underline{\theta}) \sim MVN(0, \sigma^2 I)$

$\Rightarrow \frac{1}{\sigma}P(\hat{\underline{\theta}} - \underline{\theta}) \sim MVN(0, I)$

$\Rightarrow \left[\frac{1}{\sigma}P(\hat{\underline{\theta}} - \underline{\theta})\right]^T \left[\frac{1}{\sigma}P(\hat{\underline{\theta}} - \underline{\theta})\right] \sim \chi^2(r)$

$\Rightarrow \frac{1}{\sigma^2}(\hat{\underline{\theta}} - \underline{\theta})^T P^TP(\hat{\underline{\theta}} - \underline{\theta}) \sim \chi^2(r)$

Know that

$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)$

$\Rightarrow \frac{(\hat{\underline{\theta}} - \underline{\theta})^T P^TP(\hat{\underline{\theta}} - \theta)}{r \cdot \hat{\sigma}^2} \sim F(r, n\ - p\ - 1)$

$\Rightarrow \frac{(\hat{\underline{\theta}} - \underline{\theta})^T (A(X^TX)^{-1}A^T)^{-1}(\hat{\underline{\theta}} - \underline{\theta})}{r \cdot \hat{\sigma}^2} \sim F(r, n-p-1)$

Rule: Reject if

$F^* > F_\alpha(r, n\ - p\ - 1)$

Too difficult to compute on an exam.

Alternative (Equivalent) Statistic

$F^* = \frac{[SSE_{\text{Reduced}} - \text{SSE}_{\text{Full}}]/_r}{\text{SSE}_{\text{Full}}/_{n-p-1}}$

General F-Test Examples

July 23, 2014 1:26 PM

General F-Test (Test a linear set of hypotheses)

Compare two models

i) Full Model $\text{SSE}_f, \text{SSR}_f, \text{SST}_f, \hat{\sigma}_f^2$

ii) Reduced Model $\text{SSE}_{re}, \text{SSR}_{re}, \text{SST}_{re}, \hat{\sigma}_{re}^2$

Statistic:

$$F^* = \frac{(\text{SSE}_{re} - \text{SSE}_f)/r}{\frac{\text{SSE}_f}{\text{df}_f}}$$

$r = $ # of restrictions

$\text{df}_f = $ degrees of freedom under full

If $F^* > F_\alpha(r, \text{df}_f)$ reject the null hypothesis

Example 1

Recall Chapter 5 example (comparing several groups)

$$\text{Data}\begin{cases}\text{Diet} & \leftarrow \text{Categorical}\\ \text{Weight} & \leftarrow \text{Response}\end{cases}$$

$n = 10$ people

Diet #1: y_1, y_2, y_3

Diet #2: y_4, y_5, y_6

Diet #3: y_7, y_8, y_9, y_{10}

Model: $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$

$$x_{1j} = \begin{cases}1 & j = 1,2,3\\ 0 & \text{otherwise}\end{cases}, \quad x_{2j} = \begin{cases}1 & j = 4,5,6\\ 0 & \text{otherwise}\end{cases}, \quad x_{3j} = \begin{cases}1 & j = 7,8,9,10\\ 0 & \text{otherwise}\end{cases}$$

Question: Does weight gained depend on diet?

$H_0: \beta_1 = \beta_2 = \beta_3$

(Cannot use the regular ANOVA F-Test here)

General F-test: $H_0: A\underline{\beta} = \underline{c}, \ \underline{\beta} = \begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix}$

$$A = \begin{bmatrix}0 & 1 & -1\\ 1 & -1 & 0\end{bmatrix}, \qquad \underline{c} = \begin{bmatrix}0\\ 0\end{bmatrix}$$

$r = $ # of rows $= 2$

$$F^* = \frac{(\text{SSE}_{\text{re}} - \text{SSE}_f)/2}{\hat{\sigma}_f^2}$$

Compare to $F_\alpha(2, 10 - 3 = 7)$

FEV Example

Model: $\text{Weight} = \beta_1 \cdot \text{Trt} + \beta_2 \cdot \text{Ctrl}$

$A\underline{\beta} = \underline{c}$

$$\underline{\beta} = \begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}, \qquad A = [1, -1], \qquad \underline{c} = [0], \qquad \Rightarrow r = 1$$

$$F^* = \frac{(\text{SSE}_{re} - \text{SSE}_f)/1}{\hat{\sigma}_f^2}$$

Full:

```
mod1 = lm(weight ~ group - 1)
```

$\text{SSE}_f = (0.6964)^2(18), \qquad \hat{\sigma}_f^2 = 0.6964^2$

Reduced Model:
```
modr = lm(weight ~ 1)
```
$\text{SSE}_{re} = (0.704)^2(19)$

$F^* = 1.416967$
$F_{0.95}(1,18) = qf(0.95,1,18) = 4.413873$
$\Rightarrow$ Conclusion. Since $F^* < 4.41$ we conclude that weight does not depend on diet.

Example: Show that the ANOVA F-test is a special case of the general F-test

$H_0 = \beta_1 = \beta_2 = \cdots = \beta_p = 0$

$$A = \begin{bmatrix} 1 & -1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

$\Rightarrow r = p$

$$F^* = \frac{(\text{SSE}_{re} - \text{SSE}_f)/p}{\frac{\text{SSE}_f}{n-p-1}}$$

Now, $\text{SSE}_{re} = \text{SSE}(y = \beta_0 + \epsilon) = \text{SST}$

$$\Rightarrow F^* = \frac{(\text{SST} - \text{SSE}_f)/p}{\frac{\text{SSE}_f}{n-p-1}} = \frac{\text{SSR}_f/p}{\text{SSE}_f/n-p-1}$$

Logistic Regression

July 25, 2014 1:07 PM

Earlier lectures: y_i was continuous
What if y_i is binary (an indicator)?

Example

$$y_i = \begin{cases} 1 & i^{\text{th}} \text{ row is a bad buy} \\ 0 & \text{otherwise} \end{cases}$$

$i = 1, 2, \dots, n$
Explanatory variables:
x_{i1} = Vehicle Age
x_{i2} = Milage
x_{i3} = Nationality
x_{i4} = Online Sale?

Aim: predict whether a car is a bad buy?
Previous Model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$
Can we use this? For now, let's ignore $x_2, x_3, x_4 \Rightarrow y \sim x_1$

Model for a binary response
Regression would have modelled $E(y|x)$
If $y_i \sim \text{Bernoulli}(\pi_i)$ then

y_i	0	1
$f(y_i)$	$1 - \pi_1$	π_1

$\Rightarrow E(y_i|x_i) = \pi_i$

Question: Can π_i be explained by our x_i's? Can we somehow use linear regression as before, perhaps by slightly changing the form of the model?
To answer this, consider the model $\eta_i = \beta_0 + \beta_1 x_{i1}$
Range of η_i is $(-\infty, \infty)$ we want to relate η_i to π_i where $\pi_i \in [0, 1]$
"Trick" Use a transformation on π_i say $g(\pi_i)$ so that for $\pi_i \in [0, 1]$, $g(\pi_i) \in (-\infty, \infty)$

Structural Part of Model

$$g(\pi_i) = \eta_i = \beta_0 + \beta_1 x_{i1}$$

Common Forms of $g(\pi_i)$

1) Logistic. $g(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$
   Interpretation: $\frac{\pi_i}{1-\pi_i}$ is the odds of success. We can think of logistic regression as modelling the log odds.
2) Probit Link
   Let $\Phi(z) = P(N(0,1) = z)$
   Then $g(\pi_i) = \Phi^{-1}(\pi_i)$
3) Complementary log-log
   $g(\pi_i) = \log\left(-(\log \pi_i)\right)$

Logistic Regression Model

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = X\beta$$

Note: $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i \Rightarrow \pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{e^{X\underline{\beta}}}{1+e^{X\underline{\beta}}}$
Interpretation of β_j: β_j is the increase or decrease in the log odds of success, when x_j is increased by 1 unit, and all other x's are held constant. Can show β_j = log odds ratio =

$\ln\left(\frac{\left(\frac{\pi_i}{1-\pi_i}\right)}{\left(\frac{\pi_i^*}{1-\pi_i^*}\right)}\right)$

Want to estimate β_j

Recall: $y_i \sim \text{Bernoulli}(\pi_i)$

$\Rightarrow P(y_i = 1) = \pi_i, P(y_i = 0) = \pi_i$

$P(y_i = j) = \pi_j^j (1-\pi_i)^{1-j}, \qquad j = 0, 1$

Likelihood function:

$$L(\underline{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

Log likelihood

$$l(\underline{\beta}) = \sum_{i=1}^{n}(y_i \log \pi_i + (1-y_i)\log(1-\pi_i))$$

Find $\underline{\beta} \ni \frac{dl}{d\underline{\beta}} = \underline{0}$

The resulting $\underline{\beta}$ is the maximum likelihood estimated, denoted $\underline{\hat{\beta}}$

$$\frac{dl}{d\beta_j} = \frac{\partial l}{\partial \pi_i} \cdot \frac{\partial \pi_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = \left[\frac{y_i}{\pi_i} - \frac{1-y_i}{1-\pi_i}\right]\frac{\partial \pi_i}{\partial \eta_i} x_{ij} = \left[\frac{y_i}{\pi_i} - \frac{1-y_i}{1-\pi_i}\right]\pi_i(1-\pi_i)x_{ij}$$

$$= [y_i(1-\pi_i) - (1-y_i)\pi_i]x_{ij} = [y_i - \pi_i]x_{ij}$$